

Advances in Monte Carlo Rendering: The Legacy of Jaroslav Křivánek

Alexander Keller
NVIDIA

Pascal Gautron
NVIDIA

Jiří Vorba
Weta Digital

Iliyan Georgiev
Autodesk

Martin Šik
Chaos Czech

Eugene d'Eon
NVIDIA

Pascal Griftmann
Saarland University

Petr Vévoda
Charles University Prague

Ivo Kondapaneni
Charles University Prague

SIGGRAPH 2020 Course



Jaroslav Křivánek's research aimed at finding the one robust and efficient light transport simulation algorithm that would handle any given scene with any complexity of transport. He had a clear and unique vision of how to reach this ambitious goal. On his way, he created an impressive track of significant research contributions. In this course, his collaborators tell the story of Jaroslav's quest for that "one" algorithm and discuss his impact and legacy.

Contents

| | |
|--|------------|
| 1 Syllabus | 2 |
| 1.1 Introduction (Alexander Keller) | 2 |
| 1.2 In Memoriam of Jaroslav Křivánek (Pascal Gautron) | 2 |
| 1.3 Irradiance and radiance caching (Pascal Gautron) | 2 |
| 1.4 Sampling paths (Iliyan Georgiev) | 2 |
| 1.5 Zero-variance walks (Eugene d'Eon) | 2 |
| 1.6 Path guiding (Jiří Vorba) | 3 |
| 1.7 Direct lighting (Petr Vévoda) | 3 |
| 1.8 Multiple importance sampling (Pascal Grittmann, Ivo Kondapaneni) | 3 |
| 1.9 Markov chain methods (Martin Šik) | 3 |
| 2 Presenters | 4 |
| 2.1 Alexander Keller, NVIDIA | 4 |
| 2.2 Pascal Gautron, NVIDIA | 4 |
| 2.3 Jiří Vorba, Weta Digital | 4 |
| 2.4 Iliyan Georgiev, Autodesk | 4 |
| 2.5 Martin Šik, Chaos Czech | 5 |
| 2.6 Eugene d'Eon, NVIDIA | 5 |
| 2.7 Pascal Grittmann, Saarland University | 5 |
| 2.8 Petr Vévoda, Charles University Prague | 5 |
| 2.9 Ivo Kondapaneni, Charles University Prague | 6 |
| 3 The Legacy of Jaroslav Křivánek | 7 |
| 4 Irradiance and Radiance Caching | 19 |
| 5 Sampling Paths | 55 |
| 6 Zero-Variance Theory for Efficient Subsurface Scattering | 107 |
| 7 Path Guiding | 144 |
| 7.1 Introduction | 144 |
| 7.2 Define "Path Guiding" | 145 |
| 7.3 Previous Work | 146 |
| 7.4 On-line Learning of Incident Radiance | 147 |
| 7.5 Optimal Path Lengths: Guided Russian Roulette and Splitting | 148 |
| 7.6 Glossy BSDFs: Product Sampling | 151 |
| 7.7 Path Guiding in Volumes | 152 |
| 7.8 Variance-Aware Path Guiding | 152 |
| 7.9 Subsequent Research | 152 |
| 7.10 Industry Impact | 153 |
| 7.11 Future Works | 153 |
| 7.12 Conclusion | 154 |
| 8 Direct Lighting | 155 |
| 9 Multiple Importance Sampling | 212 |
| 10 Markov Chain Methods | 300 |
| References | 374 |

1 Syllabus

Jaroslav Křivánek has been an outstanding and highly respected researcher of the rendering community who passed away far ahead of time. Through his numerous contributions to light transport simulation he managed to profoundly influence an entire domain of academia and industry.

In this course, we recap many important contributions of Jaroslav's career, underlining their practicality and pointing out how they all were consequent steps to finding the "one" robust light transport simulation algorithm that would efficiently render any given scene. Rarely has a single person had such an impact, and the authors believe it is worth remembering and continuing his legacy.

1.1 Introduction (Alexander Keller)

Alex will provide a brief introduction to the course.

1.2 In Memoriam of Jaroslav Křivánek (Pascal Gautron)

Beyond the amazing scientist, Jaroslav was also famous for his humanity, kindness, and infectious smile that left a mark on each and every person he met. A tribute to a life of science, friendship, and fearlessness.

1.3 Irradiance and radiance caching (Pascal Gautron)

Irradiance Caching has been the solution of choice to amortize the computation of diffuse inter-reflections over entire regions in world space. This idea marked the beginning of Jaroslav's search for a generalized and efficient light transport solution, and he extended the principle to global illumination on glossy surfaces (KGPB05). While effective in principle, (ir)radiance caching has numerous caveats, such as the surface roughness range in which the algorithm is applicable, interpolation artifacts, and corner oversampling. We elaborate on solutions towards practical, robust (ir)radiance caching and its applications in production rendering.

1.4 Sampling paths (Iliyan Georgiev)

Monte Carlo rendering methods are based on sampling light transport paths that connect emitters and sensors. The key to achieving efficiency is to find those paths that bring significant amount of light to the camera. Jaroslav recognized that devising a single robust path sampling technique for all types of scenes is an elusive challenge. Instead, he focused on simple, specialized techniques for different illumination effects and on efficiently combining these techniques. This effort has pushed the state of the art in both surface (GKDS12) and volumetric (KGH⁺14, GKH⁺13, GMH⁺19) rendering. We review these advances and discuss the valuable insights they have provided.

1.5 Zero-variance walks (Eugene d'Eon)

For efficient Monte Carlo light transport simulation, it is vital to sample only light paths that contribute to the image to avoid wasting computations on sampling irrelevant paths. We show how the theory of zero-variance estimators from neutron transport can inform the design of low-variance estimators by making globally-informed (as opposed to purely local) importance sampling decisions at every scattering event in a medium

to guide paths towards light sources in a way that balances their final contributions back at the camera. We demonstrate the theory using a novel perfectly zero-variance estimator due to Jaroslav, and also review a practical variance reduction scheme for subsurface scattering (Kd14).

1.6 Path guiding (Jiří Vorba)

Traditional path sampling techniques are inefficient in scenes with complex geometric occlusion. This can be addressed by designing an estimator inspired by the zero-variance theory, which guide paths towards relevant regions of the scene. The work of Vorba and colleagues (VKŠ⁺14), under Jaroslav's supervision, has resumed the interest in such path guiding methods, showing their practical potential on scenes with complex visibility and as a complement to methods like VCM (GKDS12). More importantly, this was the first work to point out that path guiding can be viewed as learning uncertainty, and as such an abundant toolbox of machine learning techniques can be explored within the rendering context. We cover path guiding techniques explored by the team around Jaroslav (VK16, HEV⁺16, HZE⁺19a), show their connection to zero-variance theory and neutron transport, and discuss the impact of these works in research and industry.

1.7 Direct lighting (Petr Vévoda)

Direct and indirect illumination calculations are two important components of any physically based renderer. While the indirect component has been traditionally considered a more complex problem and has been studied in many research works, Jaroslav acknowledged that improving the efficiency of direct illumination could have a substantial impact on the overall rendering performance, especially with complex visibility and in the presence of many light sources. In this part, we cover direct illumination sampling based on online learning of light selection probability distributions. We show how to formulate the learning process as Bayesian regression to prevent over-fitting and ensure robustness even in the early stages of computation (VKK18).

1.8 Multiple importance sampling (Pascal Grittmann, Ivo Kondapaneni)

Efficiently combining various sampling techniques is vital in modern realistic rendering. For over a decade, the balance and power multiple importance sampling (MIS) heuristics have been universally accepted, and the problem was largely deemed solved by the community. Jaroslav's search for the "one" algorithm led him to challenge these widespread beliefs. We discuss how the optimal weights can be far better than the balance heuristic (KVG⁺19), and we show that injecting variance information can improve robustness (GGSK19). We further discuss how the theoretical insights around MIS have been used to make algorithms more lightweight, robust, and efficient (KŠV⁺19).

1.9 Markov chain methods (Martin Šik)

Jaroslav saw the Markov chain methods as an alternative way towards the "one" rendering algorithm. These methods generate sequences of correlated samples, which yield faster convergence than independent Monte Carlo sampling. However, that convergence can be irregular and unpredictable, which had been overlooked (KKG⁺14, ŠK19b). We discuss how to achieve more uniform convergence in Markov chain Monte Carlo (ŠK16, GRŠ⁺16, ŠOHK16) to improve its viability in practice (ŠK19a).

2 Presenters

2.1 Alexander Keller, NVIDIA



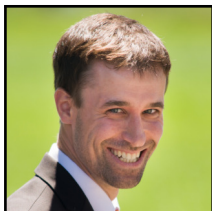
Alexander Keller is a Director of Research at NVIDIA. Before, he had been the Chief Scientist of mental images, where he had been responsible for research and the conception of future products and strategies including the design of the NVIDIA Iray light transport simulation and rendering system. Prior to industry, he worked as a full professor at Ulm University, where he co-founded the UZWR (Ulmer Zentrum für wissenschaftliches Rechnen) and received an award for excellence in teaching. Alexander Keller has more than 3 decades of experience in ray tracing and pioneered quasi-Monte Carlo methods for light transport simulation. His current interests include machine learning and wireless communication.

2.2 Pascal Gautron, NVIDIA



Pascal Gautron's work at NVIDIA is focused on designing and optimizing fast, high-quality rendering solutions. Over the last 15 years, he has gathered an academic and industrial background in computer graphics research, photorealistic image synthesis, real-time rendering, and movie post-production.

2.3 Jiří Vorba, Weta Digital



Jiří is a researcher and rendering software developer at Weta Digital. He has received his Ph.D. from Charles University in Prague in 2017. From September 2012 to January 2013, he undertook an internship at Max Planck Institute for Informatics in Saarbrücken under the supervision of Dr. Tobias Ritschel. In 2014, as part of his internship with Weta, he implemented research results on path guiding achieved during his PhD. into Manuka, Weta Digital's renderer.

2.4 Iliyan Georgiev, Autodesk



Iliyan is a researcher and principal software engineer at Autodesk. He holds a degree from Saarland University, Germany, for which he received the Eurographics PhD Thesis Award. His research is focused primarily on Monte Carlo methods for physically based light transport simulation. Iliyan publishes regularly at top-tier scientific journals and conferences, and his work has been incorporated into various production rendering systems.

2.5 Martin Šik, Chaos Czech



Martin Šik is a senior researcher and developer at Chaos Czech, a.s., where he helps to develop Corona Renderer. Martin received his Ph.D. from Charles University in 2019, where he studied under the supervision of Jaroslav Křivánek. His primary research interest is in realistic rendering with the focus on both Markov chain Monte Carlo and ordinary Monte Carlo methods for light transport simulation.

2.6 Eugene d'Eon, NVIDIA



Eugene d'Eon is a research scientist at NVIDIA working on realistic simulation of surface and volume scattering. He has published a number of papers on skin and hair rendering and has helped develop the Manuka renderer at Weta Digital and the Arnold renderer at Autodesk. Much of his last few years has been devoted to generalizing analytical and Monte Carlo methods in classical linear transport theory to support correlated random media.

2.7 Pascal Grittmann, Saarland University



Pascal Grittmann is a PhD student at Saarland University under the supervision of Prof. Philipp Slusallek. He obtained his BSc (2016) and MSc (2018) in computer science at Saarland University. In 2017, he spent three months at Charles University in Prague on an internship, working with Jaroslav Křivánek. His research focuses on Monte Carlo methods for light transport simulation, chasing the dream of the one algorithm to render them all.

2.8 Petr Vévoda, Charles University Prague



Petr Vévoda is a PhD student at Charles University and a researcher at Chaos Czech. He was supervised by Jaroslav Křivánek at both positions. His research is focused on realistic rendering algorithms and their application in production. He collaborated on research of a new algorithm for rendering participating media, a principled way of learning a distribution for many-lights sampling from previous observations, and a provably optimal MIS weights.

2.9 Ivo Kondapaneni, Charles University Prague



Ivo Kondapaneni is a PhD candidate at Charles university, and was working in Computer Graphics Group under supervision of Jaroslav Křivánek and currently under supervision of Alexander Wilkie. His research focuses on Monte Carlo methods, machine learning and statistical modeling for Light transport simulation.

3 The Legacy of Jaroslav Křivánek



ADVANCES IN MONTE CARLO RENDERING

Alexander Keller, Director of Research, NVIDIA

Welcome to the SIGGRAPH 2020 Course on Advances in Monte Carlo Rendering.

My name is Alexander Keller and I am a director of research at NVIDIA.



THE LEGACY OF JAROSLAV KŘIVÁNEK

This year's course is special, as it is in commemoration of Jaroslav Křivánek.

Jaroslav Křivánek has been an outstanding and highly respected researcher of the rendering community who passed away far ahead of his time. Through his numerous contributions to light transport simulation he managed to profoundly influence an entire domain of academia and industry.

In this course, we therefore will recap many important contributions of Jaroslav's career, underlining their practicality and pointing out how they all were consequent steps to finding the "one" robust light transport simulation algorithm that would efficiently render any given scene. Rarely has a single person had such an impact, and we believe it is worth remembering and continuing the legacy of Jaroslav Křivánek.

THE LEGACY OF JAROSLAV KŘIVÁNEK

The SIGGRAPH Courses

► Practical Global Illumination with Irradiance Caching (2007)

- reuse by similarity heuristics
- temporal coherence and stability

► Path Space Filtering (2016) with weights $w_{i,j}$ based on the same similarity heuristics

$$\bar{c}_i := \frac{\sum_{j=0}^{b^m-1} \chi_{\mathcal{B}(n)}(\mathbf{x}_{s_i+j} - \mathbf{x}_i) \cdot w_{i,j} \cdot c_{s_i+j}}{\sum_{j=0}^{b^m-1} \chi_{\mathcal{B}(n)}(\mathbf{x}_{s_i+j} - \mathbf{x}_i) \cdot w_{i,j}}$$

► path reuse from Continuous Multiple Importance Sampling (2020)

$$\langle I(\mathbf{y}) \rangle = \sum_{i=1}^n \frac{f_i(\mathbf{y}, z_i)}{\sum_{j=1}^n p(z_j | \mathbf{y}_j)}$$

3

Before we dive in, please note that we have a web page for this course. Its link will be provided on the last slide of the presentation. Also note that in the PDF of this presentation, all references are links that can be clicked.

Teaching is certainly one important part of Jaroslav's legacy, and in this introduction to the course, I will give some exemplar evidence of the impact of the SIGGRAPH courses that Jaroslav had been part of or that he was organizing.

His first course was on practical global illumination with irradiance caching, an algorithm that a large part of the rendering industry had been working on at that time. Irradiance caching provided a way to speed up global illumination computations by sharing irradiance computations when appropriate. The possibility of sharing had been determined by similarity heuristics.

Invented by Greg Ward, Jaroslav with Pascal Gautron investigated the algorithm, provided a profound understanding of its limitations, extended it accordingly, and made it more practical.

They also took the next step and investigated temporal coherence and image stability when computing multiple frames of an animation.

+

It were exactly these similarity heuristics that later were used in the weights of path space filtering, a technique that uses a weighted average to share the contributions c of light paths among multiple camera paths by averaging them. A limitation of that technique is the restriction of the heuristics to create only binary weights, as otherwise using the weighted average may increase variance.

+

At this year's SIGGRAPH this limitation of path reuse has been lifted by introducing continuous multiple importance sampling: Instead of a weighted average, the contributions of n light transport paths are shared in a query camera path \mathbf{y} by combining the contribution f_i of each light path segment z_i divided by the sum of the probabilities of having generated the light path segment, given that its camera path segment would have been \mathbf{y}_j .

This is a first example of Jaroslav's impact over multiple years.

THE LEGACY OF JAROSLAV KŘIVÁNEK

The SIGGRAPH Courses

► **Optimizing Realistic Rendering with Many-Light Methods (2012)**

- handling difficult light paths (virtual spherical lights and virtual point light distribution)
- scalability and real-time

► **Instant Radiosity (1997) and Illumination in the Presence of Weak Singularities (2004)**

$$L_r(\mathbf{x}, \omega_r) = \int_A f_r(\omega, \mathbf{x}, \omega_r) \min\{G(\mathbf{x}, y), b\} V(\mathbf{x}, y) L_e(y, -\omega) dy \\ + \int_{S^2} L_e(h(\mathbf{x}, \omega), -\omega) \frac{\max\{G(\mathbf{x}, h(\mathbf{x}, \omega)) - b, 0\}}{G(\mathbf{x}, h(\mathbf{x}, \omega))} f_r(\omega, \mathbf{x}, \omega_r) \cos^+ \theta_x d\omega$$

- MIS Compensation: Optimizing Sampling Techniques in Multiple Importance Sampling (2019)
- Optimal multiple importance sampling (2019)

► **Spatiotemporal Reservoir Resampling for Real-time Ray Tracing with Dynamic Direct Lighting (2020)**

4

For the second example, I like to go back to the 2012 course on “Optimizing Realistic Rendering with Many-Light Methods”. Many lights had been identified as a challenge, and numerous approaches to the problem were around. Jaroslav with his collaborators focused on the deep understanding and making the algorithms practical - already including the aspects of scalability and real-time rendering.

+

At that time Jaroslav invited me to talk about instant radiosity, a simple algorithm that traced a set of light transport paths to create what later has been named virtual point light sources. Using this point cloud, global illumination can be simply computed by adding up the contributions of all virtual point lights visible from the point to be shaded.

An issue is that the geometry term needed to be clipped or bounded by b in order to avoid overmodulation due to the inverse square distance in case the point to be shaded and the virtual point light are too close.

As it turns out, this limitation was easy to get rid of: Clipping the geometry term is compensated by + adding back the clipped part of the reflection integral, however, now integrated over the hemisphere to avoid the weak singularity. Simply speaking, the missing part is acquired by scattering a ray into the hemisphere. This in fact makes a heuristic for multiple importance sampling. A heuristic that instead of combining multiple sampling techniques just partitioned the domain of integration.

+

As you may already guess, Jaroslav with his coworkers extended multiple importance sampling their way: Going beyond what Veach introduced in 1996, they came up with foundational improvements to multiple importance sampling over 20 years after the original work.

+

At this year’s SIGGRAPH, a part of Jaroslav’s vision comes true: scalable real-time rendering of many lights is a reality in “Spatiotemporal Reservoir Resampling for Real-time Ray Tracing with Dynamic Direct Lighting”.

THE LEGACY OF JAROSLAV KŘIVÁNEK

The SIGGRAPH Courses

► Recent Advances in Light Transport Simulation: Some Theory and a Lot of Practice (2014)

- ways to formulate the radiance L_r reflected in a surface point x

$$\begin{aligned}
 L_r(x, \omega_r) &= \int_{S^2_-(x)} L_i(x, \omega) f_r(\omega_r, x, \omega) \cos \theta_x d\omega \\
 &= \int_{\partial V} V(x, y) L_i(x, \omega) f_r(\omega_r, x, \omega) \cos \theta_x \frac{\cos \theta_y}{|x - y|^2} dy \\
 &= \lim_{r(x) \rightarrow 0} \int_{\partial V} \int_{S^2_-(y)} \frac{\chi_B(x - h(y, \omega))}{\pi r(x)^2} L_i(h(y, \omega), \omega) f_r(\omega_r, h(y, \omega), \omega) \cos \theta_y d\omega dy \\
 &= \lim_{r(x) \rightarrow 0} \int_{S^2_-(x)} \frac{\int_{B(x)} w(x, x') L_i(x', \omega) f_r(\omega_r, x, \omega) \cos \theta_{x'} dx'}{\int_{B(x)} w(x, x') dx'} d\omega
 \end{aligned}$$

- push-button rendering: **Corona Renderer**
 - usability
 - robustness

5

Besides practice, Jaroslav very successfully drove rendering and Monte Carlo methods theory. In fact, production rendering companies were highly interested in Jaroslav's algorithms and he had been sharing his work all over the industry in addition to academia. Rendering algorithm research reached a new level by exploring more abstract approaches and unification. As a short example, let's take a look at the reflection integral formulated as integration over the hemisphere.

+

Its formulation over the surface ∂V has been known for long and is the basis of algorithms as the aforementioned multiple importance sampling.

+

New then was the formulation of photon mapping in integral form. Whenever a photon hit sufficiently close to a point of query x , it was included in the computation. To make this technique consistent, the radius r had to go to zero. In the limit, this amounts to the surface integral.

+

Path space filtering is a generalization of integrating over the solid angle. If a contribution of incident light in x' is sufficiently close to x , it may be shared. Again, in the limit, this technique is consistent.

+

Besides theory, half the course had been dedicated to practitioners, i.e. making the algorithms real. Besides Solid Angle, Pixar, and Next Limit Technologies, Ondrej Karlik presented the Corona Renderer, which is all about usability. Together with Jaroslav and Adam Hotovy the Corona renderer later on became commercial.

This was at a tipping point in rendering industry: The path tracing revolution in movie industry was about to happen, replacing complicated rendering algorithms with lots of parameters to optimize by the push-button paradigm. The new generation of renderers had only a minimal set of parameters which resulted in a much improved usability and robustness, not to speak of the much better image quality by physically based rendering.

THE LEGACY OF JAROSLAV KŘIVÁNEK

The SIGGRAPH Courses

- ▶ Machine Learning and Rendering (2018)
 - On-line learning of parametric mixture models for light transport simulation (2014)
- ▶ Learning Light Transport the Reinforced Way (2016)
 - identity of reinforcement learning and light transport simulation
- ▶ Neural Importance Sampling (2019) and Neural Control Variates (2020)

6

In 2014 Jaroslav's student Jiří Vorba and team presented the "On-line learning of parametric mixture models for light transport simulation". Core of the work was that light transport simulation could be made more efficient by learning which light transport paths were important. This seminal article caused an avalanche of articles on machine learning and rendering, which became subject of the 2018 SIGGRAPH Course.

+

Work that followed showed that in fact reinforcement learning and light transport simulation follow the same integral equation. All that needed to be done was guiding path tracing towards the light sources. This was as simple as learning where the light came from. It is worth a note that the data structures were very similar to irradiance caching. Even the similarity heuristics were similar. It actually was only a small tweak in the "old" algorithms to unleash a whole new level of improved performance.

+

All the related techniques were termed "path guiding". Besides the classic data structures, neural importance sampling released in the 2018 course showed that in fact neural networks were capable of efficient path guiding. They even can replace the classic data structures to approximate radiance, which in turn enabled neural control variates in light transport simulation.

THE LEGACY OF JAROSLAV KŘIVÁNEK

The SIGGRAPH Courses

- ▶ **The Path-Tracing Revolution in the Movie Industry (2015)**
 - ACM Transactions on Graphics [Special Issue on Production Rendering](#)
- ▶ **Realistic Rendering in Architecture and Product Visualization (2018)**
 - architectural, automotive, and product visualization
 - juxtapose this technology to rendering for motion pictures and point out the most significant differences
 - relatively little attention in the communication at SIGGRAPH

 - we planned to apply for an ACM Transactions on Graphics Special Issue...
- ▶ check Jaroslav's presentation [Open Problems and Research Directions \(2018\)](#)

7

As mentioned before, the path-tracing revolution changed the movie industry and following the 2015 SIGGRAPH course, the production rendering companies described their technologies in depth in a seminal Special Issue of the ACM Transactions on Graphics.

+

2018 Jaroslav called for a course on the other kind of industrial rendering, the Realistic Rendering in Architecture and Product Visualization. In fact, renderers for architectural, automotive, and product visualization are based on design decisions that are very different from production rendering. Jaroslav made the fact public by “juxtaposing this technology to rendering for motion pictures and pointing out the most significant differences” as described in the course abstract.

+

The abstract also states that “relatively little attention in the communication at SIGGRAPH” is dedicated to such rendering technologies and the presentations in that course made the case. Jaroslav and myself planned to propose an ACM Transactions on Graphics Special issue - complementary to the Production Rendering Special Issue...

+

In that sense, I like to recommend checking Jaroslav's presentation on Open Problems and Research Directions (2018) as it still serves as vision and guideline for future research.

ADVANCES IN MONTE CARLO RENDERING

The Legacy of Jaroslav Krivánek



Alex Keller
NVIDIA



Pascal Gautron
NVIDIA



Jiří Vorba
Weta Digital

8

Now, let me introduce the presenters of this course. Besides myself, Pascal Gautron and Jiří Vorba helped with the organization of the course and will present their research with Jaroslav.

ADVANCES IN MONTE CARLO RENDERING

The Legacy of Jaroslav Křivánek



Iliyan Georgiev
Autodesk



Martin Šik
Chaos Czech



Eugene d'Eon
NVIDIA

We then have Iliyan Georgiev, Martin Šik, and Eugene d'Eon who were close collaborators of Jaroslav, too.

ADVANCES IN MONTE CARLO RENDERING

The Legacy of Jaroslav Křivánek



Ivo Kondapaneni
Charles University Prague



Pascal Grittmann
Saarland University



Petr Vévoda
Charles University Prague

10

Last but not least, there are Ivo Kondapaneni, Pascal Grittmann, and Petr Vévoda to share their part of science with Jaroslav.

JAROSLAV KŘIVÁNEK

A brilliant mind

<https://cgg.mff.cuni.cz/~jaroslav/>

course web page

<https://sites.google.com/view/legacyofjaroslav/home>



Jaroslav's research aimed at finding the one robust and efficient light transport simulation algorithm that would handle any given scene with any complexity of transport. He had a clear and unique vision of how to reach this ambitious goal.

On his way he created an impressive track of significant research contributions that you will find on his web page.

4 Irradiance and Radiance Caching



SIGGRAPH THINK BEYOND
2020 19-23 JULY WASHINGTON DC

IRRADIANCE AND RADIANCE CACHING
Pascal Gautron - NVIDIA

© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

GLOBAL ILLUMINATION



Direct Lighting Only



Global Illumination

© PDI/Dreamworks



corona

rendered with Corona Renderer | www.corona-renderer.com

© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

2

Simulating the multiple bounces of light is crucial to achieve realistic appearance for virtual environments. Due to its inherent complexity, global illumination was originally approximated by crude methods such as a uniform ambient lighting or artist-positioned virtual light sources. This was later extended to interreflections on diffuse surfaces. Nowadays global illumination is a first-class citizen in any renderer. However, accurately simulating the behavior of light in all cases remains a challenge, that Jaroslav decided to take systematically, filling the gaps step by step. This chapter will focus on his early research work of extending the irradiance caching algorithm to compute global illumination on glossy surfaces.

32 YEARS AGO: IRRADIANCE CACHING



Direct+Indirect



Indirect Only

© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

Computer Graphics, Volume 22, Number 4, August 1988

A Ray Tracing Solution to Global Illumination

Gregory L. Heath
Robert C. Case
Luciano Bimonte
James R. Van Dam
IBM Research Division
Yorktown Heights, NY 10598

Abstract
An efficient ray tracing method is presented for calculating indirect lighting contributions with both diffuse and specular surfaces. A global illumination algorithm is used to generate indirect lighting. The method is based on the observation that indirect lighting varies slowly over surfaces. This is achieved by using a sparse cache of irradiance values and interpolating between these values to estimate the irradiance at any point in the scene. The method is based on the observation that indirect lighting varies slowly over surfaces. This is achieved by using a sparse cache of irradiance values and interpolating between these values to estimate the irradiance at any point in the scene. The method is based on the observation that indirect lighting varies slowly over surfaces. This is achieved by using a sparse cache of irradiance values and interpolating between these values to estimate the irradiance at any point in the scene.

1. Introduction
The indirect component of lighting is a significant factor in the overall appearance of a scene. It is the component of lighting that is most difficult to calculate and is the component that is most often neglected in ray tracing. The indirect component of lighting is a significant factor in the overall appearance of a scene. It is the component of lighting that is most difficult to calculate and is the component that is most often neglected in ray tracing.

2. Motivation
The indirect component of lighting is a significant factor in the overall appearance of a scene. It is the component of lighting that is most difficult to calculate and is the component that is most often neglected in ray tracing. The indirect component of lighting is a significant factor in the overall appearance of a scene. It is the component of lighting that is most difficult to calculate and is the component that is most often neglected in ray tracing.

3. Method
The indirect component of lighting is a significant factor in the overall appearance of a scene. It is the component of lighting that is most difficult to calculate and is the component that is most often neglected in ray tracing. The indirect component of lighting is a significant factor in the overall appearance of a scene. It is the component of lighting that is most difficult to calculate and is the component that is most often neglected in ray tracing.

4. Results
The indirect component of lighting is a significant factor in the overall appearance of a scene. It is the component of lighting that is most difficult to calculate and is the component that is most often neglected in ray tracing. The indirect component of lighting is a significant factor in the overall appearance of a scene. It is the component of lighting that is most difficult to calculate and is the component that is most often neglected in ray tracing.

5. Conclusion
The indirect component of lighting is a significant factor in the overall appearance of a scene. It is the component of lighting that is most difficult to calculate and is the component that is most often neglected in ray tracing. The indirect component of lighting is a significant factor in the overall appearance of a scene. It is the component of lighting that is most difficult to calculate and is the component that is most often neglected in ray tracing.

IRRADIANCE CACHING: FOUNDATIONS

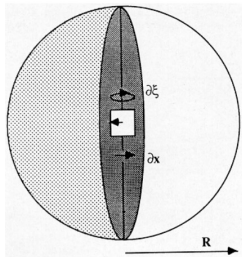


Figure 3: The split sphere model. A surface element is located at the center of a half-dark sphere.

© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

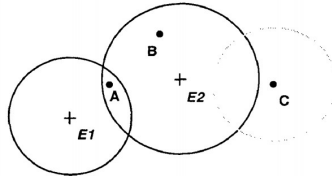


Figure 2: Illuminances $E1$ and $E2$ were calculated previously using the primary method. Test point A uses an average of $E1$ and $E2$. Point B uses $E2$. Point C results in a new indirect illuminance value at that location.

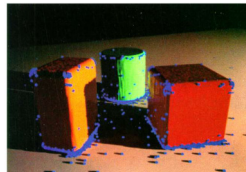


Figure 5: Blocks with illuminance value locations in blue.

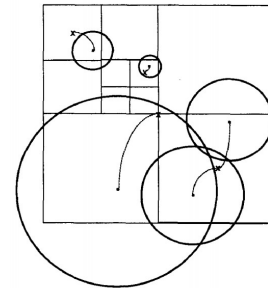


Figure 8: Five indirect illuminance values are shown with their respective domains (circles) linked by dotted lines to the appropriate nodes (squares).

4

The principle of interpolating indirect lighting is present in many approaches, such as the radiosity-based techniques. In comparison with radiosity the irradiance caching (IC) algorithm mainly differs by its simplicity and built-in adaptivity.

The core principle behind the adaptive nature of IC is the observation that while irradiance varies slowly over surfaces, the rate of variation is directly dependent on the distance to the surrounding objects. A worst-case scenario is defined by the “split-sphere” model (left) where, for a given point, half the hemisphere above is completely dark and one is completely bright. If we were to reuse the irradiance in the neighborhood of that point, what would the error be compared to the ground truth? This error is a direct consequence of the distance to the surrounding objects, which is obtained as a by-product of irradiance evaluation. Using an estimate of that error one can deduce an accuracy criterion that will control how far an irradiance value can be reused around a given point (middle).

The irradiance records are generated by hierarchically traversing the image, and looking up the existing records in an octree (right). If no record covers the considered point the algorithm creates a new record and stores it in the octree. This has the advantage of lazily computing global illumination only where needed.

IRRADIANCE GRADIENTS

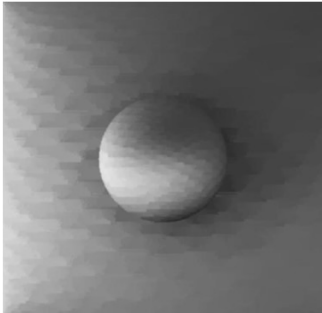


Figure 1a. Irradiance extrapolation without gradients.

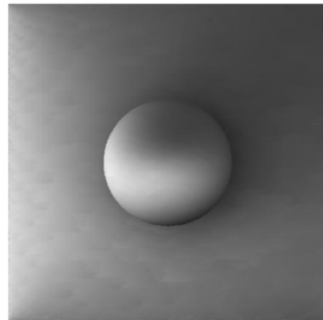


Figure 1b. Irradiance extrapolation with gradients.

© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

Irradiance Gradients
Gregory J. Ward*
LEDA-M*
École Polytechnique, Université de Lorraine
CR1003 Lorraine
Paul R. Heckbert*
Department of Technical Mathematics & Informatics
Duke University of Technology
Information 112
2820 Hill Road
North Carolina

ABSTRACT
A new method for improving the accuracy of a diffuse interreflection calculation is introduced in a real-time context. The information from a hierarchical sampling of the luminance environment is interpreted in a new way to produce the change in irradiance as a function of position and surface orientation. The additional computation involved is minimal and the results are demonstrable. An improved interpretation of irradiance resulting from the gradient calculation produces smoother, more accurate shadows. This result is achieved through better exploitation of ray samples rather than additional samples or alternate sampling strategies. Thus, the technique is applicable to a variety of global illumination algorithms that are heuristic or Monte Carlo sampling techniques.

1. Introduction
Global illumination can be simulated using both ray tracing and radiosity algorithms. Both approaches typically rely on extrapolations of patch irradiances which are used to revise other patch irradiances from BSSRDF or an incident radiance source. In most radiosity algorithms, patch irradiances are considered

IRRADIANCE GRADIENTS

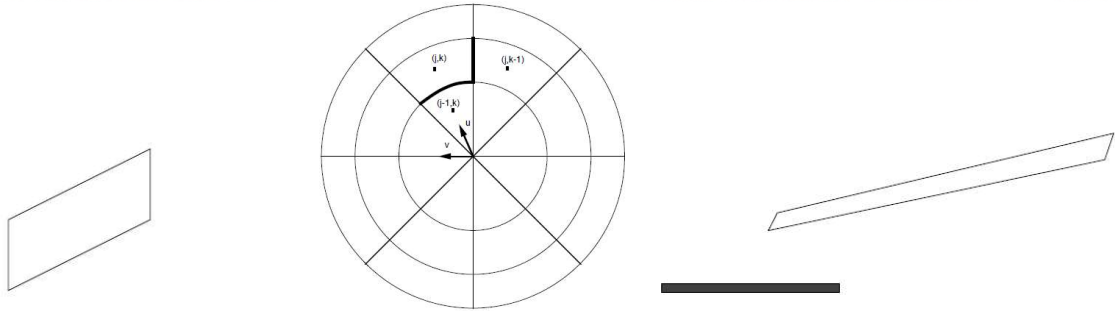


Figure 4. Cells of an example hemisphere sampling.

Figure 2a. As our point is rotated counter-clockwise, the surface's contribution increases.

Figure 2b. Translational Gradient. As our point moves to the right, irradiance increases.

The irradiance estimate is based on a stratified sampling of the hemisphere above each irradiance record. For each cell the irradiance gradients model how the 'walls' of the cell would move if the surface would be tilted or translated. While computing the rotation gradient is straightforward, the translation gradient needs to account for an estimate of the parallax effect.

CHALLENGES: GLOSSY SURFACES



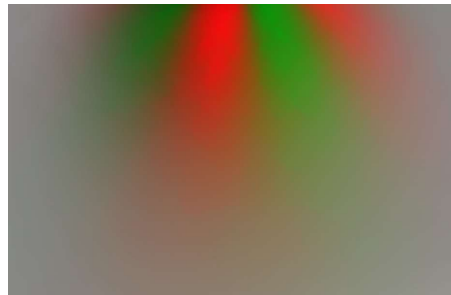
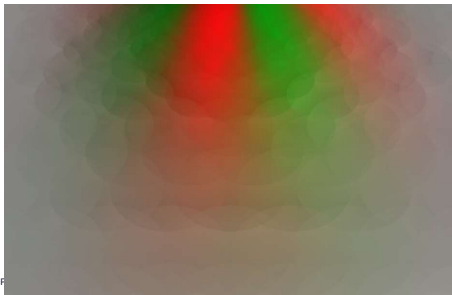
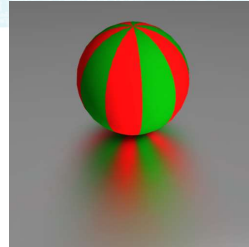
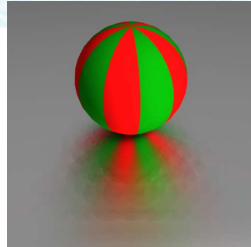
© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

7

The irradiance caching algorithm can provide accurate estimates of indirect diffuse lighting, and has been the core of the Radiance software, originally released in 1994 (<https://floyd.lbl.gov/radiance/>).

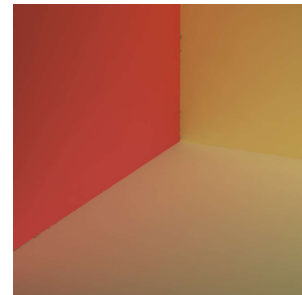
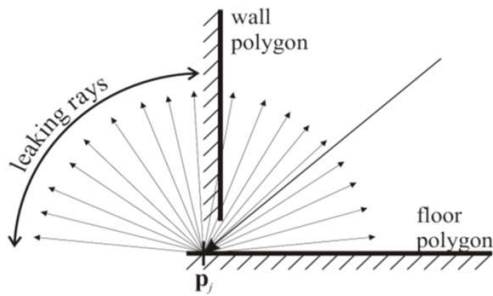
However its scope is limited to diffuse interreflections, and cannot simulate the view-dependent effects of glossy materials (left: diffuse-only GI, right: ground truth). As most real-life objects have a directional component, extending this approach to glossy surfaces was a natural step towards designing a single algorithm for general light transport.

CHALLENGES: GRADIENT ACCURACY



The irradiance gradients provide a limited estimate of the irradiance changes within the radius of an irradiance record, and are tied to one specific stratified sampling scheme. Improving the gradients is a crucial part of making this algorithm more robust.

CHALLENGES: SAMPLING AND LIGHT LEAKING



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

9

Irradiance caching is also very sensitive to ray leaking. The radius of a record is a direct function of the distance to its surrounding geometry. A consequence of this is when some rays 'leak' through imperfections in the geometry (left, typically due to floating-point errors) the radius is overestimated and the irradiance is underestimated, creating visible errors (middle).

RADIANCE CACHING: FILLING THE GAP



Irradiance
Caching

Path
Tracing

IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 11, NO. 5, SEPTEMBER/OCTOBER 2005

Radiance Caching for Efficient Global Illumination Computation

Jaroslav Křivánek, Pascal Caumont, Sumanta Pattanaik, Member, IEEE, and Kati Bouatouch, Member, IEEE

Abstract—In this paper, we present a fast tracing method for accurate global illumination computation in scenes with the increasing amount of geometry. The method is based on scene sampling, and it incorporates radiance and global illumination, as well as radiance caching. The method is able to handle scenes with a large number of objects, and it is able to handle scenes with a large number of objects. The method is able to handle scenes with a large number of objects. The method is able to handle scenes with a large number of objects.

Index Terms—Global illumination, ray tracing, hierarchical radiance, radiance caching, irradiance caching

1 INTRODUCTION

Monte Carlo ray tracing is the method of choice for generating images of complex environments with global illumination [2]. Even for the radiance method, high quality images are created by first gathering, then using Monte Carlo ray tracing [3].

Monte Carlo ray tracing is, however, expensive when it comes to computing indirect illumination on surfaces with low-frequency BRDFs. Bidirectional radiance distribution functions (BRDFs) may have to be traced to get a reasonably precise estimate of the incoming radiance at a point. Fortunately, a high degree of coherence in the incoming radiance field on these materials [1], [4], [5], [6] can be exploited by interpretation [1], [2] to obtain a significant performance gain.

Our goal is to accelerate Monte Carlo ray tracing-based global illumination computation in the presence of surfaces with low-frequency glossy BRDFs. We achieve it through scene sampling, caching, and interpolating radiance on these surfaces. In particular, we extend Ward et al.'s irradiance caching [1], [4] to glossy surfaces. Irradiance caching is based on the observation that reflected radiance on diffuse surfaces due to indirect illumination changes very slowly with position. However, the quality is reduced significantly by the glossy BRDFs. Additionally, this observation, we extend Ward et al.'s work to cache and interpolate the directional incident radiance instead of the radiance. This allows us to accelerate indirect lighting computation on surfaces with glossy BRDFs. We call the new method radiance caching.

The incoming radiance at a point is represented by a vector of coefficients with respect to spherical or hemispherical harmonics [9]. Due to the basic orthogonality, the illumination integral of radiance [2] reduces to dot product of the coefficients. Radiance interpolation is carried out by interpolating the coefficients. We enhance the interpolation quality by the use of translational gradients. We propose novel methods for sampling gradients that overcome the limitations of the method in [10] and [11].

Radiance caching shares all the advantages of Ward's novel method, but, importantly, gradients that are stored at a few track locations are computed on track points of the scene; no restrictions are imposed on the scene geometry, implementation is straightforward on track points. Our approach can be directly used with any BRDF represented by photophysical parameters, including near-transparent BRDFs.

This paper extends the initial description of radiance caching on the computation of radiance caching on glossy surfaces, an automatic method for reducing BRDFs suitable for radiance caching, new methods for computing translational radiance gradients, and integration of radiance caching in a ray tracing pipeline. The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 gives an overview of how radiance caching works and how it is integrated into the scene. Section 4 shows different types of radiance caching. Section 5 presents the results. Section 6 discusses various aspects not covered in the algorithm description. Section 7 concludes the paper and summarizes our ideas for future work.

2 RELATED WORK

2.1 Interpretation

Interpretation can be used to global illumination whenever there is a certain level of coherence in the quantity being computed. The radiance method was interpreted in the form of surface illumination, e.g., [12], [13], [14]. In the context of Monte Carlo ray tracing, approaches have been proposed to reduce variance [15].

Irradiance caching tends to converge rapidly on diffuse materials. Path tracing converges very quickly on highly specular surfaces. For other, moderately glossy materials, neither caching nor simple path tracing can provide fast and accurate results.

RADIANCE CACHING: FILLING THE GAP



IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 11, NO. 5, SEPTEMBER/OCTOBER 2005

Radiance Caching for Efficient Global Illumination Computation

Jaroslav Křivánek, Pascal Caumont, Sumanta Pattanaik, Member, IEEE, and Kati Bouatouch, Member, IEEE

Abstract—In this paper, we present a ray tracing based method for accelerating global illumination computation in scenes with the radiance cache (RC). The method is based on storing and retrieving radiance information at global illumination (GI) nodes, and is designed to be efficient in terms of memory and computation. The radiance cache is used to store and retrieve radiance information at global illumination nodes. The radiance cache is used to store and retrieve radiance information at global illumination nodes. The radiance cache is used to store and retrieve radiance information at global illumination nodes.

Index Terms—Global illumination, ray tracing, hemisphere normals, spherical harmonics, irradiance distribution



Irradiance
Caching

Radiance
Caching

Path
Tracing

1 INTRODUCTION

Monte Carlo ray tracing is the method of choice for simulating images of complex environments with global illumination [2]. Even for the radiance method, high quality images are created by first gathering, then using Monte Carlo ray tracing [3].

Monte Carlo ray tracing is, however, expensive when it comes to computing indirect illumination on surfaces with low-frequency BRDFs. Bidirectional radiance distribution functions (BRDFs) may have to be traced to get a reasonably precise estimate of the incoming radiance at a point. Fortunately, a high degree of coherence in the incoming radiance field on these surfaces [1], [4], [5], [6] can be exploited by interpretation [1], [2] to obtain a significant performance gain.

Our goal is to accelerate Monte Carlo ray tracing-based global illumination computation in the presence of surfaces with low-frequency BRDFs. We achieve it through sphere sampling, caching, and interpolating radiance on these surfaces. In particular, we extend Ward et al.'s irradiance caching [1], [4] to glossy surfaces. Irradiance caching is based on the observation that reflected radiance on glossy surfaces does not change very rapidly with position. However, the ability to capture color and polarization is essential. Additionally, this observation, we extend Ward et al.'s work to cache and interpolate the directional incident radiance instead of the irradiance. This allows us to accelerate indirect lighting computation on surfaces with glossy BRDFs. We call the new method radiance caching.

The incoming radiance at a point is represented by a vector of coefficients with respect to spherical or hemispherical harmonics [9]. Due to the basic orthogonality, the illumination integral of equation (2) reduces to a dot product of the coefficients. Radiance interpolation is carried out by interpolating the coefficients. We enhance the interpolation quality by the use of translational gradients. We propose novel methods for approximating gradients that overcome the limitations of Ward and Kajiya [9]. Radiance caching shows all the advantages of Ward's method, but with the added advantage of being able to be implemented in conjunction with radiance caching. Our approach can be directly used with any BRDF representation by photophysical functions, including near-normal BRDFs.

This paper extends the initial description of radiance caching on the occasion of radiance caching glossy surfaces, an automatic method for reducing BRDFs suitable for radiance caching, new methods for computing translational radiance gradients, and integration of radiance caching in a ray tracing system. The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 gives an overview of the radiance caching works and how it is integrated into the system. Section 4 details the details of radiance caching. Section 5 presents the results. Section 6 discusses various aspects not covered in the algorithm description. Section 7 concludes the paper and summarizes our ideas for future work.

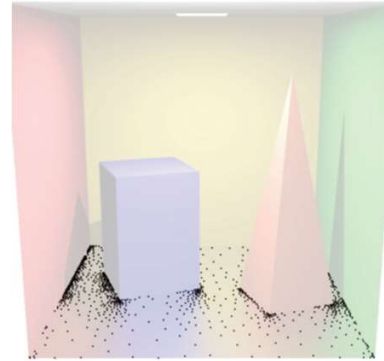
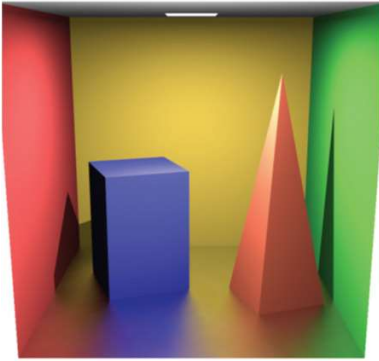
2 RELATED WORK

2.1 Interpretation

Interpretation can be used to speed up global illumination whenever there is a certain level of coherence in the quantity being computed. The radiance method uses interpretation in the form of surface interpolation, e.g., [10], [11], [12]. In the context of Monte Carlo ray tracing, approaches have been proposed to speed up rendering [13].

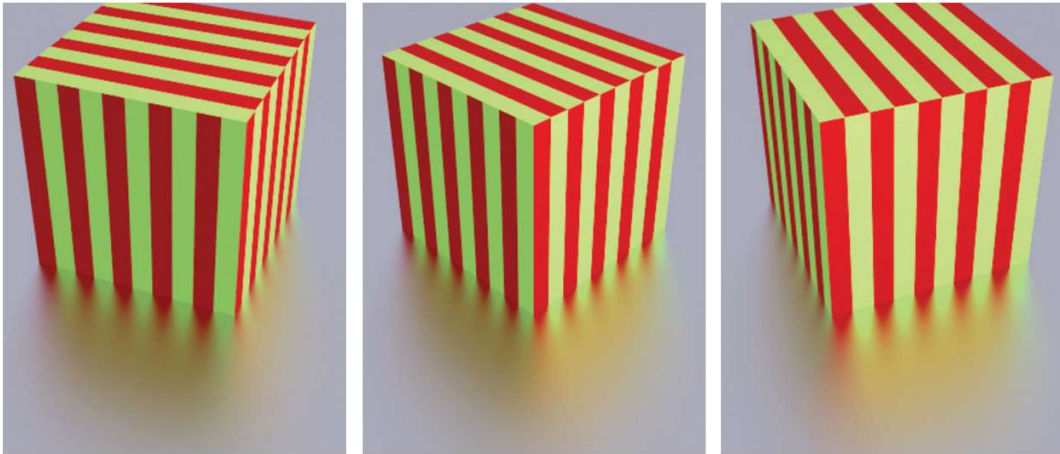
In the search of the 'one' algorithm, the development of Radiance Caching aimed at bridging the gap between caching techniques and brute force path tracing by providing a caching method suitable for storing directional radiance information.

OBSERVATIONS: SLOW SPATIAL VARIATIONS



When observing the behavior of glossy interreflections on rough surfaces, we can see the assumption of slow spatial variations still hold. We could therefore use the principles of the existing irradiance caching scheme on glossy surfaces.

OBSERVATIONS: VIEW DEPENDENCE

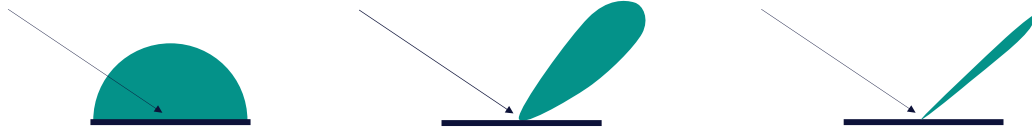
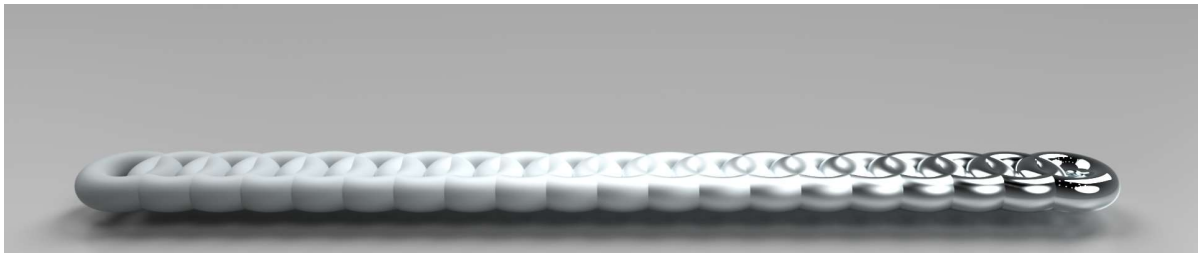


© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

13

The main difference between diffuse and glossy surfaces lies in their directional component: for a given incoming light direction, this light will be mostly reflected in a specific direction.

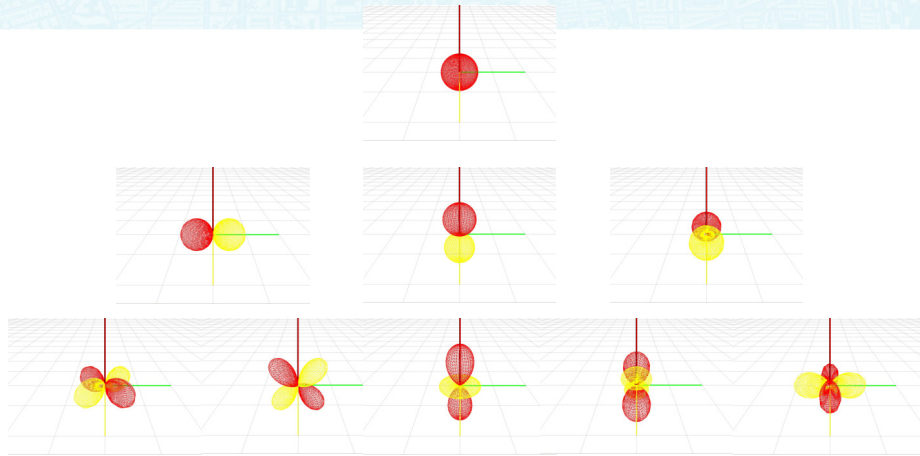
GLOSSINESS: HOW?



This directionality is described by the Bidirectional Reflectance Distribution Function, which dictates how light gets reflected on a surface. The more diffuse, the more uniform the BRDF (left). When surfaces are closer to mirrors (right), the BRDF tends to reflect most of the light in the mirror direction.

In order to define a caching scheme adapted to glossy surfaces we therefore need to cache some directional radiance information instead of just simple irradiance value.

SPHERICAL HARMONICS



Spherical Harmonics are functions defined on the sphere, which define an orthonormal functional basis.

SPHERICAL HARMONICS PROJECTION

$$\lambda = \left[\begin{array}{c} \text{teal lobe} \cdot \text{red dot} \\ \text{teal lobe} \cdot \text{red dot} \\ \text{teal lobe} \cdot \text{red dot} \\ \text{teal lobe} \cdot \text{red dot} \\ \text{teal lobe} \cdot \text{red dot} \\ \dots \end{array} \right] = \left[\begin{array}{c} \lambda_0^0 \\ \lambda_1^{-1} \\ \lambda_1^0 \\ \lambda_1^1 \\ \lambda_2^{-2} \\ \dots \end{array} \right]$$

© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

16

Any spherical function can be projected onto the spherical harmonics (SH) basis, creating a vector of projection coefficients. Each coefficient is the dot product of the function with a given spherical harmonic, ie. the integral of the product of the function with the SH. A perfect projection would require an infinity of coefficients in the general case. For practical purposes the number of coefficients is limited, for example to 100.

SPHERICAL HARMONICS PROJECTION

$$f(\omega) = \begin{bmatrix} \text{Lobe 1} \\ \text{Lobe 2} \\ \text{Lobe 3} \\ \text{Lobe 4} \\ \text{Lobe 5} \\ \dots \end{bmatrix} \cdot \begin{bmatrix} \text{SH}_1(\omega) \\ \text{SH}_2(\omega) \\ \text{SH}_3(\omega) \\ \text{SH}_4(\omega) \\ \text{SH}_5(\omega) \\ \dots \end{bmatrix} \cdot \begin{bmatrix} \text{Coeff}_1(\omega) \\ \text{Coeff}_2(\omega) \\ \text{Coeff}_3(\omega) \\ \text{Coeff}_4(\omega) \\ \text{Coeff}_5(\omega) \\ \dots \end{bmatrix}$$

© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

17

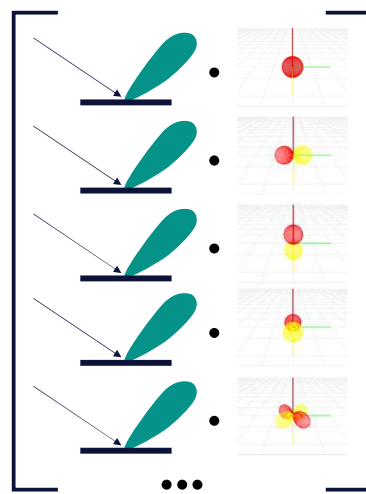
Evaluating the value of the projected function for a given direction is achieved by computing the dot product of the coefficient vector with a vector containing the SH functions evaluated for that direction.

SPHERICAL HARMONICS ROTATION

Fast Approximation to Spherical Harmonics Rotation
Srinivas Aravamudan, Jonathan R. Kwan, and Jonathan D. L. Durkin
Abstract
Spherical harmonics (SH) are a common representation for functions on the sphere. They are used in many applications, including computer graphics, geophysics, and astronomy. The SH coefficients are often stored in a compact form, and the rotation of the function is performed by multiplying the coefficients by a rotation matrix. This paper presents a fast approximation to the rotation matrix, which can be generated much faster than the exact matrix. The approximation is based on a Taylor series expansion of the rotation matrix, and is accurate to within a few percent for small rotations. The approximation is also accurate for large rotations, provided that the rotation axis is known. The approximation is implemented in a simple, efficient way, and is suitable for real-time applications.

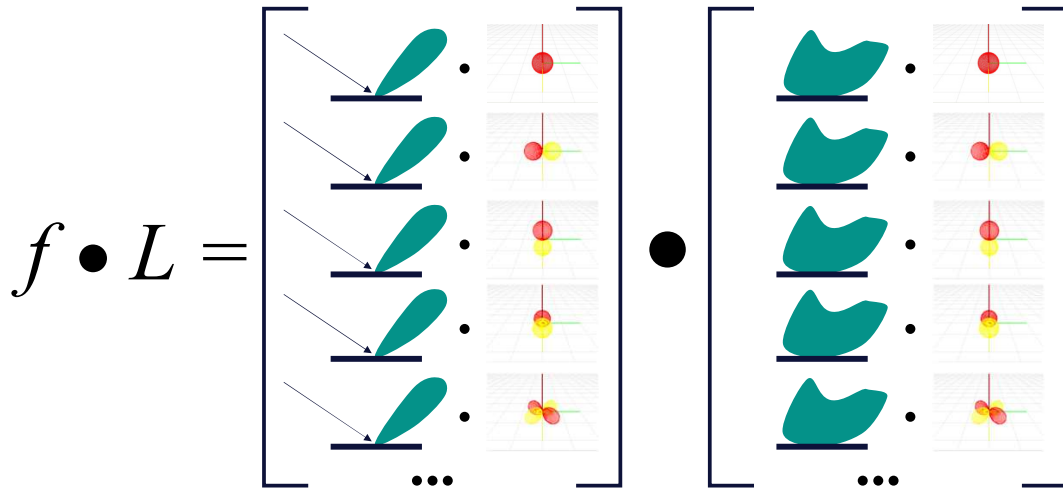
1. Introduction
Spherical harmonics (SH) are a common representation for functions on the sphere. They are used in many applications, including computer graphics, geophysics, and astronomy. The SH coefficients are often stored in a compact form, and the rotation of the function is performed by multiplying the coefficients by a rotation matrix. This paper presents a fast approximation to the rotation matrix, which can be generated much faster than the exact matrix. The approximation is based on a Taylor series expansion of the rotation matrix, and is accurate to within a few percent for small rotations. The approximation is also accurate for large rotations, provided that the rotation axis is known. The approximation is implemented in a simple, efficient way, and is suitable for real-time applications.

$$R_{\theta}(f) = R_{\theta}$$

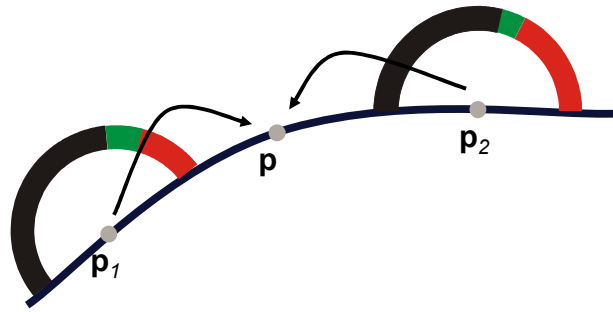


A projected function can be easily rotated using a rotation matrix, that can be efficiently generated. It is also possible to generate approximate rotation matrices much faster.

SPHERICAL HARMONICS DOT PRODUCT



A most important property of spherical harmonics (and other orthonormal bases) is the ability to compute the dot product of two projected functions. If the incoming radiance and the BRDF are projected into SH, the dot product of those functions is a simple dot product of the coefficient vectors.



$$\lambda_p = \alpha \lambda_{p1} + (1 - \alpha) \lambda_{p2}$$

Interpolating two projected functions can be efficiently achieved by a component-wise linear interpolation. When storing SH-projected incoming radiance functions at points p_1 and p_2 , one can rotate them to match their local frames, and interpolate them to obtain an approximate incoming radiance function at point p . The dot product with the BRDF at point p then yields the actual reflected radiance in a given viewing direction.

RADIANCE GRADIENTS

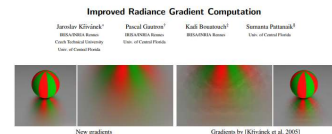
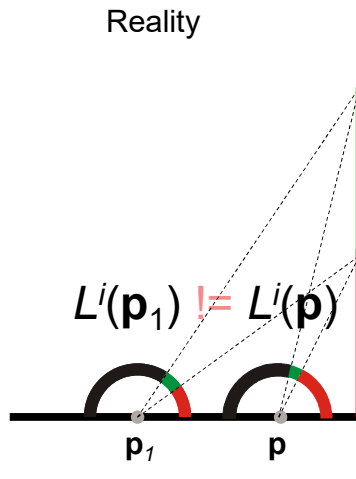
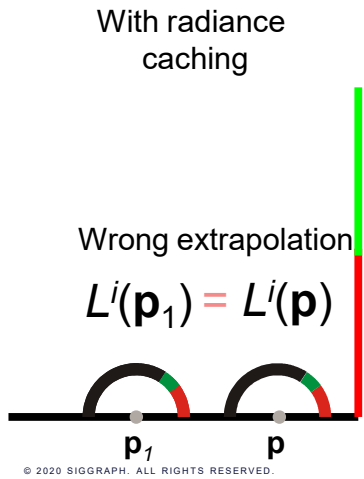


Figure 1: Right: The gradient computation proposed by Krivanek et al. (2005) does not properly handle significant change of occlusion in the sample distribution and leads to inaccurate radiance interpolation on the glossy floor. The two images in the middle are cut out from the two images on the very left and very right.

Abstract

We describe a new and accurate algorithm for computing transmittance gradients of rendering radiance in the context of ray tracing based global illumination solutions. The gradient characteristics have been investigated and radiance gradient changes with respect to the surface. We use the gradient for a radiance gradient extrapolation on the glossy surface in the framework of the radiance caching algorithm. The proposed algorithm overcomes the radiance gradient computation by Ward and Heitsch (1992) in the context of non-diffuse, glossy, media. Compared to the method for radiance gradient computation, the new algorithm yields better gradient estimates in the presence of significant occlusion changes in the sampled environment, allowing a smoother radiance interpolation.

CR Categories: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Shading, Shadowing

Keywords: radiance gradient, radiance caching, irradiance gradient, global illumination, ray tracing

Introduction and Previous Work

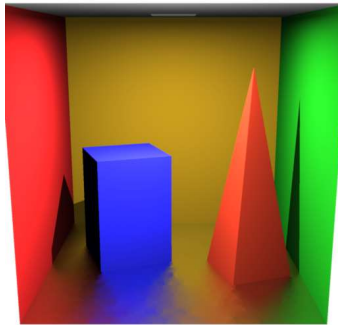
Radiance caching (Preater et al. 1998; Krivanek et al. 2005) is an efficient technique for computing indirect radiance. It involves sampling and storing the radiance of the environment over multiple bounces of light. The radiance is then used to estimate the radiance at a given point by extrapolating the stored radiance values. The proposed algorithm overcomes the radiance gradient computation by Ward and Heitsch (1992) in the context of non-diffuse, glossy, media. Compared to the method for radiance gradient computation, the new algorithm yields better gradient estimates in the presence of significant occlusion changes in the sampled environment, allowing a smoother radiance interpolation.

Irradiance caching required irradiance gradients to avoid artifacts due to the extrapolation of irradiance values in the validity radius of the samples. Those artifacts are even more visible on glossy surfaces. The original gradients formulation cannot easily adapt to any sample distribution, and is focused on irradiance only.

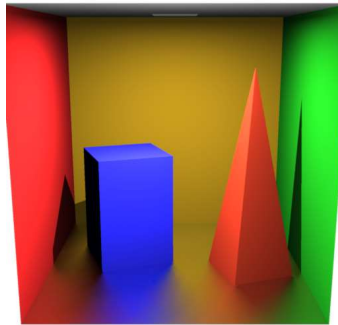
RADIANCE GRADIENTS



With radiance caching



Reality



Improved Radiance Gradient Computation

Jaroslav Křivánek¹, Pascal Gauthier², Karl Bressanelli³, Sébastien Patenaud³
¹Intel/Disney, ²Intel/Disney, ³Intel/Disney

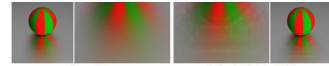


Figure 1: Right: The gradient computation proposed by Křivánek et al. (2005) does not properly handle significant change of occlusion in the scene (illumination and hence radiance distribution). Left: The modified gradient computation proposed in this paper handles such change and leads to a smoother radiance distribution interpolation on the glossy floor. The two spheres in the middle are cut out from the two images on the very left and very right.

Abstract

We describe a new and accurate algorithm for computing radiance gradients of incoming radiance to the context of ray tracing based global illumination solutions. The gradient characteristics have been investigated and radiance function change with respect to the surface. We use the gradient for a smoother radiance interpolation on the glossy surface in the framework of the radiance caching algorithm. The proposed algorithm operates the radiance gradient computation by Ward and Hecker (1992) in the context of non-diffuse, glossy, surfaces. Compared to previous methods for radiance gradient computation, the new algorithm yields better gradient estimates in the presence of significant occlusion changes in the sampled environment, allowing a smoother indirect illumination interpolation.

CR Categories: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Shading, Shadowing

Keywords: Global gradient, radiance caching, irradiance gradient, global illumination, ray tracing

¹jaroslav.krivanek@intel.com
²pascal.gauthier@intel.com
³karl.bressanelli@intel.com
⁴sebastien.patenaud@intel.com

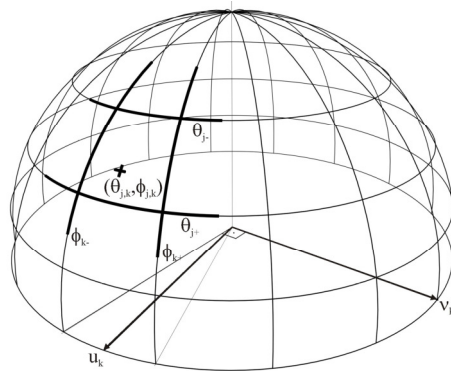
1 Introduction and Previous Work

Radiance caching [Preater et al. 1996; Křivánek et al. 2005] is an efficient means to deal with complex indirect illumination. Storing and reusing the indirect diffuse irradiance over surfaces, instead of recomputing it every time, is the key to its efficiency. It is a common practice to compute the radiance of incoming light over through a number of bounces (see, e.g., [Křivánek et al. 2005]) and then reuse it for a certain time, since it is a smooth function. Ward and Hecker (1992) found that the interpolation quality on diffuse surfaces can be significantly improved by the use of radiance and radiance gradient gradients. The radiance gradient characterizes the change of radiance with a small displacement on a surface and the radiance gradient describes the change with surface normals. The gradients are computed independently with the hemisphere sampling and stored in the cache to be used later. During the interpolation, a reflection ray is sampled from the interpolation point.

In our earlier work on radiance caching [Křivánek et al. 2005], we used the radiance gradient computation to significantly improve radiance interpolation on glossy surfaces. We achieved this by using an old directional sampling radiance function (proposed by [Lafortune et al. 1995]) instead of the current one. We used the gradient to compute the radiance of incoming light over through a number of bounces (see, e.g., [Křivánek et al. 2005]) and then reuse it for a certain time, since it is a smooth function. Ward and Hecker (1992) found that the interpolation quality on diffuse surfaces can be significantly improved by the use of radiance and radiance gradient gradients. The radiance gradient characterizes the change of radiance with a small displacement on a surface and the radiance gradient describes the change with surface normals. The gradients are computed independently with the hemisphere sampling and stored in the cache to be used later. During the interpolation, a reflection ray is sampled from the interpolation point.

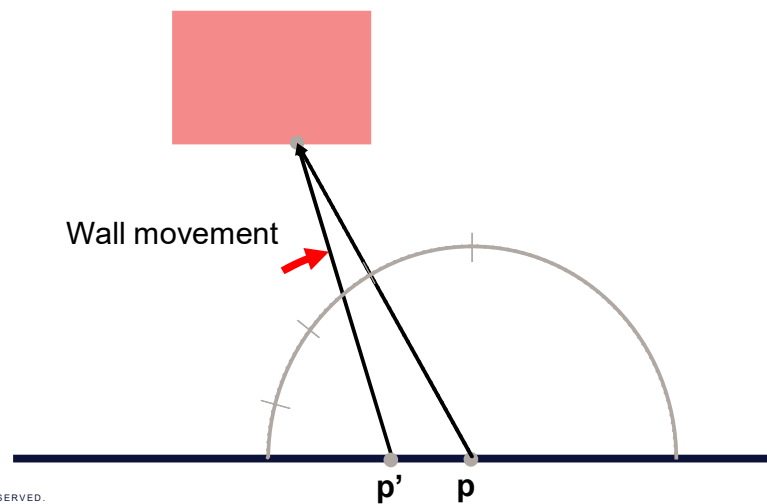
Filling in the gaps required the definition of a new way of extrapolation information from the incoming radiance function in the radiance records. First, thanks to the ability to rotate spherical harmonics, the rotation gradient becomes obsolete.

STRATIFIED SAMPLING



The translation gradient takes into account the actual stratification used in the sampling.

STRATIFIED SAMPLING



In a way similar to the original gradients, this approach attempts to compute the movement of the walls of the cells with small movements around the records location p . Intuitively, we aim at determining the energy transfer between the stratification cells due to that movement.

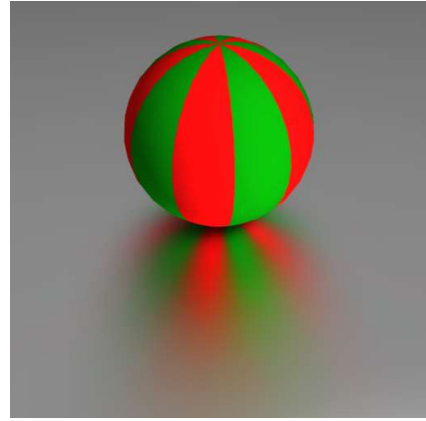
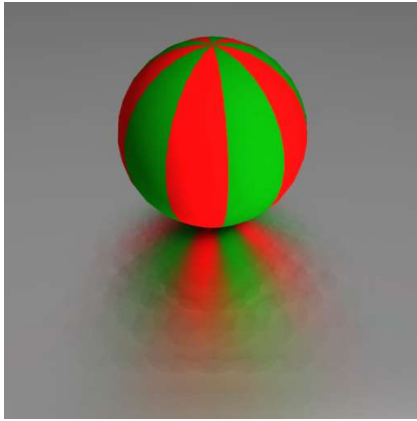
TRANSLATION GRADIENT FROM WALL MOVEMENTS

$$\vec{\nabla} \lambda_l^m = \sum_{k=0}^{N-1} \left[\hat{u}_k \frac{2\pi}{N} \sum_{j=1}^{M-1} \frac{\cos \theta_{j-} \sin \theta_{j-}}{\min\{r_{j,k}, r_{j-1,k}\}} (L_{j,k}^i - L_{j-1,k}^i) H_l^m(\theta_{j,k}, \phi_{j,k}) + \hat{v}_{k-} \frac{1}{M} \sum_{j=1}^{M-1} \frac{1}{\sin \theta_{j,k} \min\{r_{j,k}, r_{j,k-1}\}} (L_{j,k}^i - L_{j,k-1}^i) H_l^m(\theta_{j,k}, \phi_{j,k}) \right]$$

Sum together
Cell area change
Incoming radiance change
Weight by the basis function

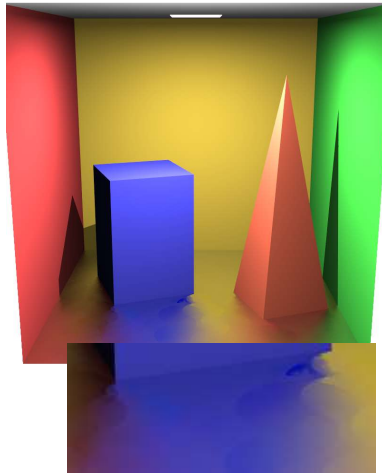
With a small movement on either axis of the tangent plane each cell would undergo a small change in its area, as well as an energy transfer with its neighboring cells. The gradient is then the product of those changes with the spherical harmonics functions (or any other (hemi)spherical basis).

RADIANCE GRADIENTS



Those SH-enabled gradients allow for a smoother interpolation of radiance functions, and hence drastically reduce the visual artifacts.

ADAPTIVE CACHING & NEIGHBOR CLAMPING



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

Proceedings of the ACM SIGGRAPH Conference on Computer Graphics and Interactive Techniques, August 2020

Making Radiance and Irradiance Caching Practical: Adaptive Caching and Neighbor Clamping

Jensel Křivánek¹ Kail Reinhard² Samanta Patra³ Jie Zou¹

¹Czech Technical University in Prague, Czech Republic ²IMAX - IMBA, Disney Pixar ³University of Central Florida, USA

Abstract

Radiance and irradiance caching are effective global illumination algorithms based on interpolating indirect illumination from a sparse set of cached values. In this paper we propose an adaptive algorithm for packing spatial density of the cached values to radiance and irradiance caching. The density is adapted to the rate of change of indirect illumination in order to avoid visible interpolation artifacts and produce smooth interpolated illumination. In addition, we discuss some practical problems caused by the implementation of radiance and irradiance caching and propose techniques for solving those problems. Finally, the neighbor clamping technique is proposed as an enhancement for detecting small corners of indirect illumination and for dealing with problems caused by ray leakage through small gaps between adjacent surfaces.

Categories and Subject Descriptors according to ACM CCS: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism - Rendering, Global Illumination

1. Introduction

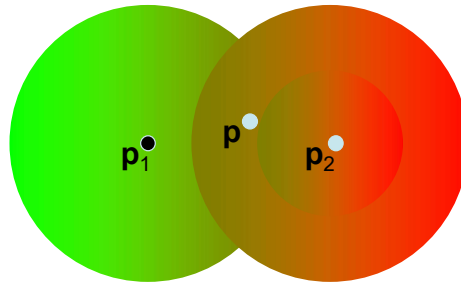
One of the practical and widely used algorithms for computing diffuse indirect illumination is irradiance caching [WATSON, WARD, 1995]. The irradiance cache is represented as a sparse set of points on surfaces, used as a cache, and then interpolated elsewhere. Radiance caching [DREYER, 2002] generalizes irradiance caching to direct surfaces with the required BRDFs. To faithfully approximate indirect illumination with only a sparse set of values, their density must be proportional to the rate of change of indirect illumination, otherwise, interpolation artifacts may appear. Irradiance caching, the algorithm used in this paper, is similar to irradiance caching, but the interpolation is based on the irradiance change is estimated based on the scene geometry. Even though the actual irradiance is constant, irradiance changes, and provides good image quality with a relatively low number of cached values.

In radiance caching, however, estimating the rate of change of indirect illumination is more difficult. On direct surfaces, not only the illumination characteristics, but also the shape of the surface BRDF and the varying distance influence the actual rate of change of indirect illumination. It would be complicated to design an interpolation error criterion that takes all these factors into account, and the resulting formula is likely to be quite computationally intensive. We solve this by using a different approach: we estimate the density of cached values based on a simple perceptual metric, a view-dependent metric.

In this paper we propose an adaptive algorithm for controlling density of cached values to radiance and irradiance caching that we refer to as adaptive caching. It starts with an initial set of cached indirect illumination values and then refines their density where necessary to eliminate interpolation artifacts. The main metric of quality in radiance and irradiance caching are visible discontinuities or boundaries of indirect lighting on visible discontinuities or boundaries of adjacent surfaces of cached values. The proposed error criterion is based on the visible discontinuities of indirect illumination. The adaptive refinement is designed to detect those discontinuities. The reference is then, perpendicular to the corner that is most not primarily and shows the physical corners of the scene, but about the visible surface discontinuities.

With adaptive caching, the record density is adapted to the actual illumination conditions better than with the original criterion used in irradiance caching (GIPDF). If indirect illumination is simple, without sudden changes, adaptive caching generates fewer records and maintains a faster rate of change of indirect illumination, adaptive caching can

ADAPT TO GRADIENT MAGNITUDE

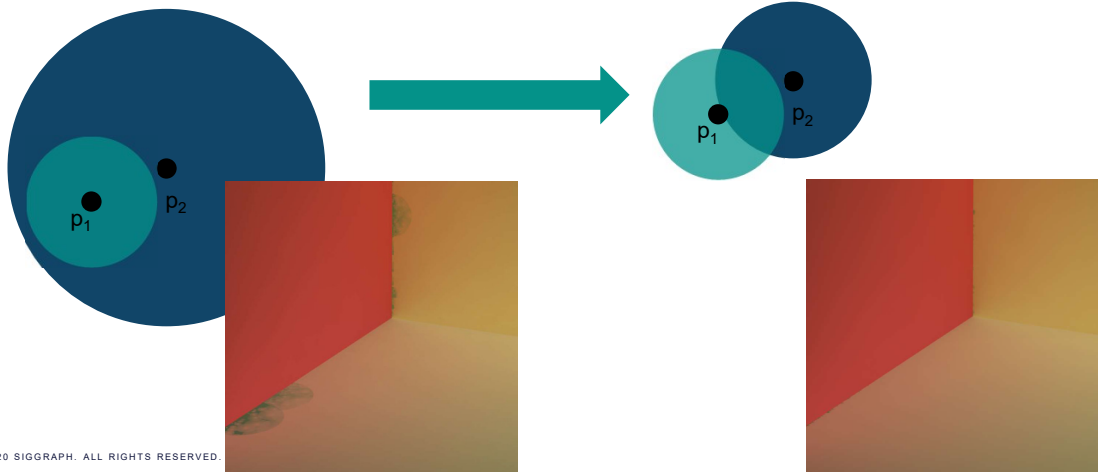


If $|L1(\mathbf{p}) - L2(\mathbf{p})| > \tau$
then decrease radius

Based on the Weber law

The artifacts in the reflections are due to the use of the “split sphere” model of irradiance caching to determine the validity radius of a radiance record. With glossy surfaces this model does not represent the worst-case scenario, resulting in a possible overestimation of the radius. Adaptive caching addresses this problem by comparing the radiance values provided by neighboring records. If the difference between the radiances extrapolated from point p_1 and p_2 is too high, then the radii of those records is reduced. This, in turn, allows the algorithm to add more records in that area and increase the fidelity of the reconstructed radiance.

NEIGHBOR CLAMPING

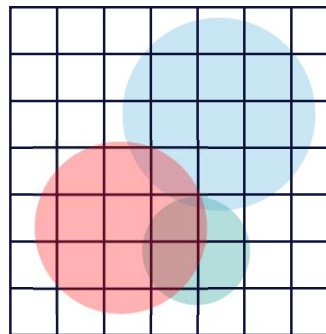
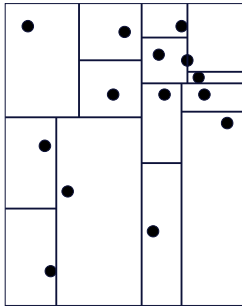


© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

29

Neighbors also provide valuable information to reduce the impact of outliers. By construction new records are only created in the scene areas where no existing record can contribute. Therefore, a record that encloses one of its neighbors is most likely an outlier with an excessive radius (left). From this observation the neighbor clamping algorithm reduces the radius of potential outliers so that no record overlaps the center of its neighbors (right).

RADIANCE CACHING ON GRAPHICS HARDWARE



Computer Graphics in Building (CGI)
Kurtz, 2019, 2020, 2021

Radiance Cache Splatting: A GPU-Friendly Global Illumination Algorithm

Pascal Giacomini¹, Jonathan Kivitski², Kai Brunnack³, Suresh Pattanaik⁴

Abstract
Fast global illumination computation is a challenge in several fields such as lighting simulation and computer-generated image effects for movies. In this work, the radiance caching algorithm is extended and improved to provide high-quality rendering in a reasonable time. However, this algorithm relies on a spatial data structure such as octrees to store and query records. We propose a novel approach to global illumination using radiance cache splatting. This method directly uses the processing constraints of graphics hardware to store a record for every cache location and algorithm. Moreover, the resulting quality is comparable to that of traditional radiance cache splatting. This method also uses an implementation of our algorithm to provide a GPU-friendly approach to local radiance caching.

Computer and Graphics Techniques in Building, in ACM SIGGRAPH 2020 Computer Graphics, Three-Dimensional Graphics and Visualization, Houston, Texas, August 2-6, 2020.

1. Introduction

The goal of global illumination computation is to simulate multiple bounces of light in a scene. As computers become more and more powerful, high-quality global illumination computation can be achieved. In the context of fields such as architectural design, cinema and video games, however, the computation is performed using ray tracing and Monte Carlo sampling, and is very costly. A number of methods have been proposed to reduce the cost of global illumination. Some approaches have been proposed to make globally illuminated scenes be rendered with an explicit GPU-friendly approach. However, interactive methods based on ray tracing rely on global processing using several computers to maintain a reasonable frame rate. An efficient approach to global illumination is to use a GPU-friendly approach to global illumination.



Figure 1: The Castle scene (CGI) rendered (obtained by ray-tracing) using the radiance cache splatting algorithm for global illumination in 2017.7 at resolution 1080x1080.

The radiance caching algorithm is based on the progressive refinement of a hierarchical data structure, and gathering operations using nearest-neighbor queries on this structure. Graphics hardware are more suited to simpler structures, and scattering operations are usually more efficient than gathering ones.

The irradiance and radiance caching algorithms are both based on a notion of validity radius around records, plus some requirements on the surface normals. Those algorithms can then be reversed: instead of finding nearby records in an octree, each record can be splatted on the image plane, with alpha blending taking care of the interpolation.

PRODUCTION RENDERING



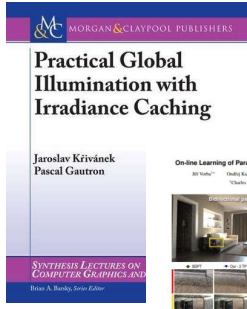
Irradiance Caching in Shrek 2
© PDI/Dreamworks



Radiosity Map for Ratatouille © Pixar

Irradiance caching has been used extensively in production rendering.

FURTHER READING



Spatial Directional Radiance Caching

Authors: Jaroslav Křivánek, Charles University Prague, Jaroslav Křivánek, Charles University Prague, Filip Dvořák, Masaryk University Brno, Filip Dvořák, Masaryk University Brno

Abstract: This paper presents a novel method for spatial directional radiance caching. It introduces a new data structure that stores directional radiance information for each point in the scene. This allows for more accurate and efficient global illumination simulations, particularly in scenes with complex geometry and materials.

Keywords: Global illumination, radiance caching, directional radiance, spatial data structures.

Light Transport Simulation with Vertex Connection and Merging

Authors: Filip Dvořák, Masaryk University Brno, Jaroslav Křivánek, Charles University Prague, Filip Dvořák, Masaryk University Brno, Filip Dvořák, Masaryk University Brno

Abstract: This paper introduces a new method for light transport simulation called Vertex Connection and Merging (VCM). It improves upon traditional path tracing by connecting vertices in the scene and merging paths, which leads to faster convergence and more accurate results, especially in scenes with high-frequency features.

Keywords: Light transport simulation, path tracing, vertex connection, merging, global illumination.

A Radiance Cache Method for Highly Glossy Surfaces

Authors: Jaroslav Křivánek, Charles University Prague, Jaroslav Křivánek, Charles University Prague, Filip Dvořák, Masaryk University Brno, Filip Dvořák, Masaryk University Brno

Abstract: This paper presents a specialized radiance cache method for highly glossy surfaces. It addresses the challenges of simulating specular reflection and refraction by using a different caching strategy that accounts for the high reflectivity of these materials.

Keywords: Radiance caching, glossy surfaces, specular reflection, refraction, global illumination.

Radiance Caching for Participating Media

Authors: Jaroslav Křivánek, Charles University Prague, Jaroslav Křivánek, Charles University Prague, Filip Dvořák, Masaryk University Brno, Filip Dvořák, Masaryk University Brno

Abstract: This paper extends radiance caching to scenes containing participating media, such as smoke, fog, and glass. It introduces a new method for handling light transport through these media, allowing for more realistic and efficient simulations.

Keywords: Radiance caching, participating media, light transport, global illumination.

Online Learning of Parametric Mixture Models for Light Transport Simulation

Authors: Jaroslav Křivánek, Charles University Prague, Jaroslav Křivánek, Charles University Prague, Filip Dvořák, Masaryk University Brno, Filip Dvořák, Masaryk University Brno

Abstract: This paper introduces an online learning method for parametric mixture models. It allows for the automatic adaptation of simulation parameters based on the current scene, leading to more efficient and accurate light transport simulations.

Keywords: Online learning, parametric mixture models, light transport simulation, global illumination.

Importance Caching for Complex Illumination

Authors: Jaroslav Křivánek, Charles University Prague, Jaroslav Křivánek, Charles University Prague, Filip Dvořák, Masaryk University Brno, Filip Dvořák, Masaryk University Brno

Abstract: This paper presents an importance caching method for complex illumination scenarios. It uses importance sampling to store and retrieve radiance information, which significantly improves the efficiency of the simulation in complex scenes.

Keywords: Importance caching, complex illumination, importance sampling, global illumination.

Massively Parallel Path Space Filtering

Authors: Jaroslav Křivánek, Charles University Prague, Jaroslav Křivánek, Charles University Prague, Filip Dvořák, Masaryk University Brno, Filip Dvořák, Masaryk University Brno

Abstract: This paper introduces a massively parallel path space filtering method. It leverages parallel processing to filter out unnecessary paths in the simulation, resulting in faster convergence and reduced memory usage.

Keywords: Massively parallel, path space filtering, global illumination, parallel processing.

Space Discretization by Plane Histogram Descriptors

Authors: Jaroslav Křivánek, Charles University Prague, Jaroslav Křivánek, Charles University Prague, Filip Dvořák, Masaryk University Brno, Filip Dvořák, Masaryk University Brno

Abstract: This paper presents a space discretization method using plane histogram descriptors. It provides a more efficient way to discretize the scene space, leading to improved performance in global illumination simulations.

Keywords: Space discretization, plane histogram descriptors, global illumination, efficiency.

© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

Jaroslav's research journey began with filling the gaps between irradiance caching and path tracing. The book 'Practical Global Illumination with Irradiance Caching' provides many solutions to improve the robustness of both irradiance and radiance caching. As this solution was only a step towards the search for a single algorithm for light transport, the shortcomings of radiance caching encouraged him to develop many significant contributions, such as the unification of points, rays and beams, and research on ray guiding that are described in the following sections of the course.

RADIANCE CACHING REFERENCES



- Křivánek, J. Gautron, P. - Practical Global Illumination with Irradiance Caching - Morgan and Claypool Publishers, March 2009
- Gautron, P., Křivánek, J., Pattanaik, S. N., Bouatouch, K. - A novel hemispherical basis for accurate and efficient rendering – Eurographics Symposium on Rendering 2004, 321–330
- Gautron, P., Křivánek, J., Bouatouch, K., Pattanaik, S. N. - Radiance cache splatting: A GPU-friendly global illumination algorithm - Eurographics Symposium on Rendering 2005 - 55–64
- Gautron, P., Bouatouch, K., Pattanaik, S. N. - Temporal radiance caching - IEEE Transactions on Visualization and Computer Graphics 13, 5 (September/October 2007)
- Jarosz, W., Donner, C., Zwicker, M., Jensen, H. W. - Radiance caching for participating media - ACM Trans. Graph. 27, 1 (March 2008)
- Jarosz, W., Zwicker, M., Jensen, H. W. - Irradiance gradients in the presence of participating media and occlusions - Eurographics Symposium on Rendering 2008, 27, 4

RADIANCE CACHING REFERENCES



- Křivánek, J. Gautron, P., Bouatouch, K., Pattanaik, S. - Improved radiance gradient computation - Spring Conference on Computer graphics 2005 - 155–159
- Křivánek, J. Gautron, P., Pattanaik, S., Bouatouch, K.- Radiance caching for efficient global illumination computation - IEEE Transactions on Visualization and Computer Graphics 11, 5 (September/October 2005)
- Křivánek, J., Konttinen, J., Bouatouch, K., Pattanaik, S., Žára, J. - Fast approximation to spherical harmonic rotation - Spring Conference on Computer graphics 2006
- Křivánek, J., Konttinen, J., Bouatouch, K., Pattanaik, S., Žára, J. - Making radiance and irradiance caching practical: Adaptive caching and neighbor clamping - Eurographics Symposium on Rendering 2006
- Křivánek, J.- Radiance Caching for Global Illumination Computation on Glossy Surfaces - PhD thesis, Université de Rennes 1 and Czech Technical University – 2005
- Larson, G. W., Shakespeare, R. - Rendering with Radiance, The Art and Science of Lighting Visualization - Morgan Kaufmann Publishers - 1998

RADIANCE CACHING REFERENCES



- Smyk, M., Ichi Kinuwaki, S., Durikovic, R., Myszkowski, K. 2005. Temporally coherent irradiance caching for high quality animation rendering - Proceedings of Eurographics 2005, 24, 3.
- Tabellion, E., And Lamorlette, A. - An approximate global illumination system for computer generated films - Proceedings of ACM SIGGRAPH 2004, 23, 3, 469–476
- Ward, G. J., Heckbert, P. S. - Irradiance gradients – Proceedings of Eurographics Workshop on Rendering 1992, 85–98
- Ward, G. J., Rubinstein, F. M., Clear, R. D. - A ray tracing solution for diffuse interreflection - Proceedings of ACM SIGGRAPH 1988, 85–92.
- Ward, G. J. - The Radiance lighting simulation and rendering system - Proceedings of ACM SIGGRAPH 1994, 459–472.

5 Sampling Paths

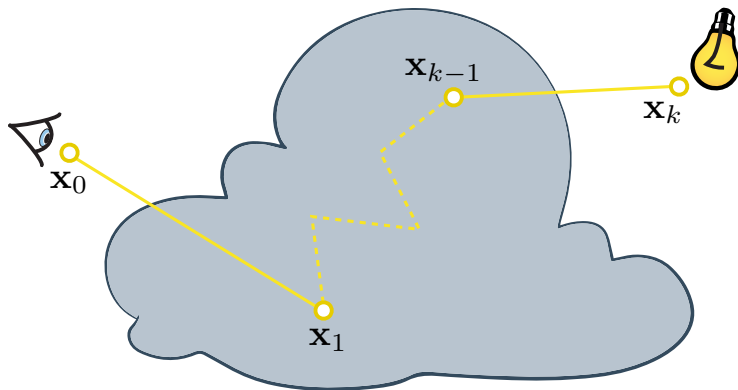
Sampling paths

Iliyan Georgiev

Autodesk

In this section we will view rendering as the problem of finding light trajectories that carry energy from the light sources to the camera. We will show how this formalism enables devising novel rendering methods as well as combining different methods in a way that preserves their individual strengths.

Path integral framework



Pixel value

$$I_j = \int_{\mathcal{P}} f_j(\bar{\mathbf{x}}) d\bar{\mathbf{x}}$$

Pixel estimator

$$\langle I_j \rangle = \frac{1}{N} \sum_{i=1}^N \frac{f_j(\bar{\mathbf{x}}_i)}{p(\bar{\mathbf{x}}_i)}$$

path contribution \leftarrow
path pdf \leftarrow

Path contribution

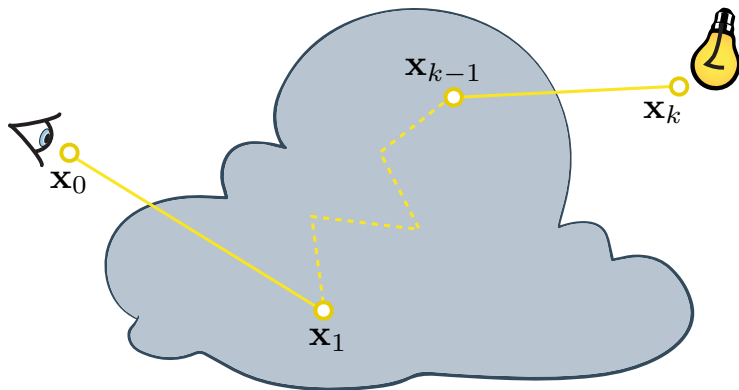
$$f_j(\bar{\mathbf{x}}) = W_j(\mathbf{x}_0, \mathbf{x}_1) \left[\prod_i f_s(\mathbf{x}_i) G(\mathbf{x}_i, \mathbf{x}_{i+1}) T(\mathbf{x}_i, \mathbf{x}_{i+1}) \right] L_e(\mathbf{x}_k, \mathbf{x}_{k-1})$$

camera response
BSDF/phase
geometry
transmittance
emitted radiance

In 1995, Eric Veach introduced the path integral formulation of light transport, which expresses the problem of computing the value of a pixel as a conceptually simple integral over the space of all trajectories, or paths, connecting the light source in the scene to the camera through an arbitrary number of bounces at surfaces or in media. The contribution of each such possible path is the product of terms, including the BSDF/phase function at each vertex, and the mutual position, orientation, and transmittance between subsequent vertex pairs.

This light transport integral can be estimated via ordinary Monte Carlo technique, i.e., by constructing a random path, evaluating its contribution and dividing by its sampling density. Note that if unbiased estimation is desired, the only degree of freedom here is in the choice of path sampling pdf. Different sampling methods, or *techniques*, can only differ in the pdf they use.

Path integral framework



Pixel value

$$I_j = \int_{\mathcal{P}} f_j(\bar{\mathbf{x}}) d\bar{\mathbf{x}}$$

Pixel estimator

$$\langle I_j \rangle = \frac{1}{N} \sum_{i=1}^N f_j(\bar{\mathbf{x}}_i) p(\bar{\mathbf{x}}_i)$$

ideally
proportional

\propto

Path contribution

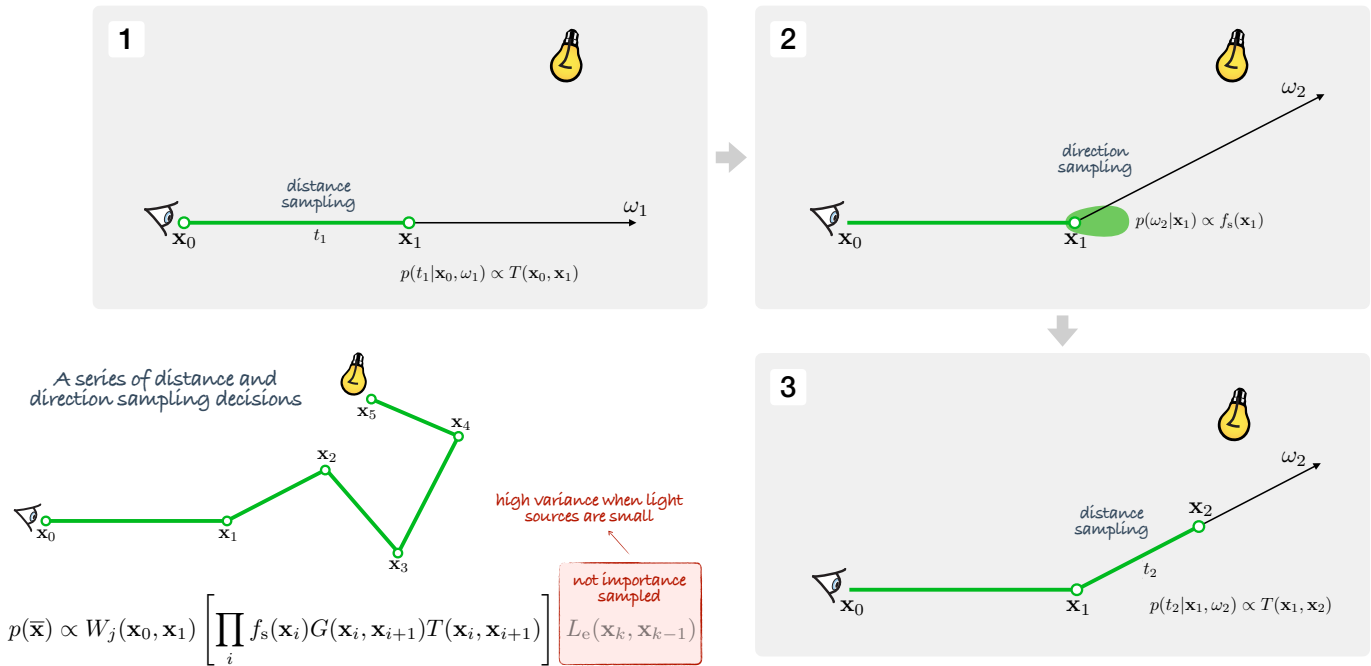
$$f_j(\bar{\mathbf{x}}) = W_j(\mathbf{x}_0, \mathbf{x}_1) \left[\prod_i f_s(\mathbf{x}_i) G(\mathbf{x}_i, \mathbf{x}_{i+1}) T(\mathbf{x}_i, \mathbf{x}_{i+1}) \right] L_e(\mathbf{x}_k, \mathbf{x}_{k-1})$$

camera response
BSDF/phase
geometry
transmittance
emitted radiance

Such sampling-based approximation introduces error. Averaging over multiple paths reduces this error. However, a much more efficient way to achieve that is to *importance sample* the paths, i.e., to use a sampling distribution that to each path assigns a density as closely proportional to its pixel contribution as possible. Unfortunately, this is a difficult task due to the complex shape of the contribution function, containing many discontinuities as well as singularities (which we will discuss below).

Importance sampling is still achievable in a localized manner, when sampling the vertices of paths in succession. Most practical methods employ such form of importance sampling as we describe next.

Unidirectional path sampling



A simple and widely used method is unidirectional path sampling. Given an initial vertex on the lens and ray through a pixel, the second vertex is determined by sampling a distance proportionally to the transmittance along the ray. In the absence of participating media this simplifies to (deterministically) finding the closest visible surface along the ray.

Once the second vertex is known, a new direction from it is sampled, typically proportionally to the local scattering distribution as given by the BSDF or the phase function at that point. A new distance is sampled along the resulting ray, and this process continues until a light source has been hit (or the path is terminated early, e.g. via Russian roulette.)

Interestingly, this process of iterative distance and direction sampling yields a path pdf that is proportional to all terms of the contribution function, with the exception of the emitted radiance at the last vertex. This technique can perform well in scenes filled with emissive surfaces. However, in many practical scenes the light sources are comparatively small and the chance of hitting them with random rays can be extremely small, resulting in substantial noise in the rendered image.

One could also perform the sampling in the opposite direction, starting from the light sources; however, the chance of randomly landing on the lens is even smaller.

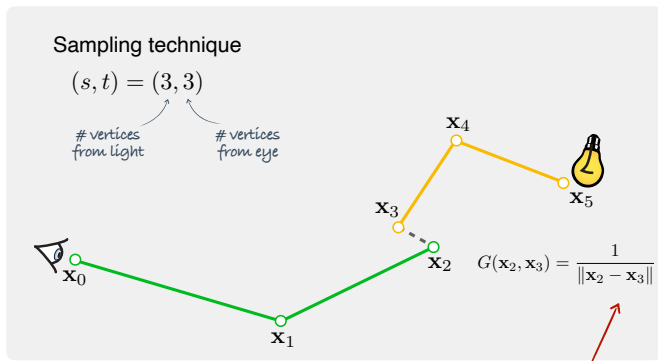
Bidirectional path sampling



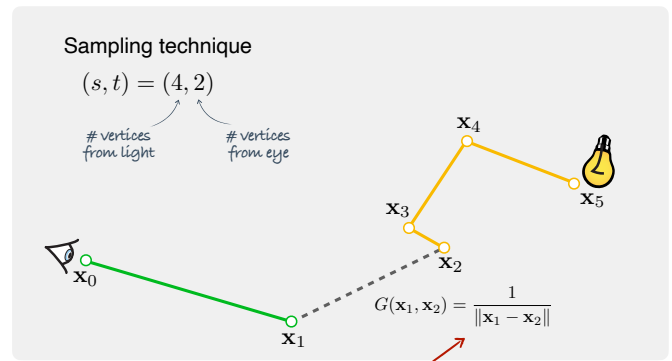
Instead of hoping the last path vertex to land on the light source by chance (1), one could directly sample that vertex on the light source (2). This technique is known as *next-event estimation*.

A subpath could also be started from that vertex and connected to the camera subpath (3). With this scheme, a path of length k edges (and $k+1$ vertices; here $k=2$) can be constructed in $k+2$ different ways (here $k+2=7$), by varying the edge along which the connection is performed (1-7). Each corresponds to a distinct sampling technique identified by the number of vertices s and $t=k+1-s$, sampled from the light and the eye respectively.

Geometric singularity



explodes when distance is small



this technique avoids the singularity

When a technique performs a connection between two vertices, it does not importance sample any of the path contribution terms associated with these vertices, since the light and eye subpaths are sampled independently from each other.

An issue arises when the two connected vertices are very close to each other in space, e.g. near a geometric corner. The geometry term associated with the edge explodes due to the inverse squared distance appearing in its denominator. The pixel estimate explodes too as the geometry term is not importance sampled. This is a well-known problem in some rendering methods based on bidirectional sampling, e.g. instant radiosity (and generally, many-light methods).

However, for each such case there are other techniques which can construct the same path by performing the connection along the other edges. The pixel estimates of these techniques have much lower magnitudes in this case. Notably, the sampling pdfs of these techniques are higher as they importance sample the high-magnitude geometry term.

So there are multiple techniques that can sample the same path, with different efficiency. Ideally we want to sample each path using the most efficient technique. However, the best technique for a given path can only be identified once it is constructed. An alternative is to use all techniques but weigh their estimates based on their efficiency.

Multiple importance sampling (MIS)

Given an integral and n estimation techniques

$$I = \int_{\Omega} f(x) d\mu(x) \quad \langle I \rangle_i = \frac{f(x)}{p_i(x)}$$

Weighted combination

$$\begin{aligned} \langle I \rangle_{\text{MIS}} &= \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} w(x_{i,j}) \langle I \rangle_{i,j} \\ &= \sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} w(x_{i,j}) \frac{f(x_{i,j})}{p_i(x_{i,j})} \end{aligned}$$

Balance heuristic

$$w_i(x) = \frac{p_i(x)}{p_1(x) + \dots + p_n(x)}$$

Multiple importance sampling (MIS) provides a way to achieve this. Given n techniques, with n_i samples $\{x_{i,j}\}_{j=1}^{n_i}$ for each, drawn from pdf p_i , the MIS estimator combines all estimates. A weight is applied to each estimate which is normalized for each sample independently, over all other techniques.

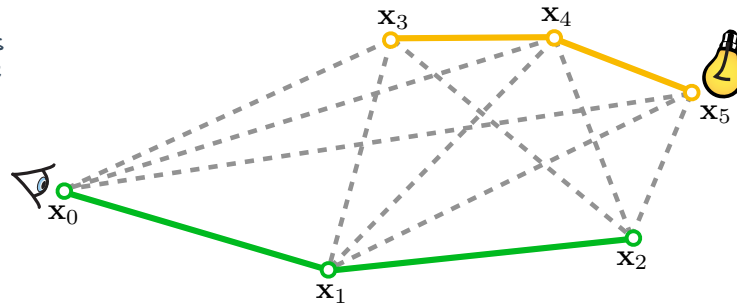
There is a lot of freedom in the choice of weighting function. The balance heuristic is a provably good choice, which weighs techniques proportionally to their sampling pdf. This provides a combination that preserves the qualities of the individual techniques and ameliorates their inefficiencies.

Bidirectional path tracing

Combined MIS pixel estimator:

$$\langle I_j \rangle = \sum_s \sum_t w_{s,t}(\bar{\mathbf{x}}_{i,j}) \frac{f_j(\bar{\mathbf{x}}_{i,j})}{p_{s,t}(\bar{\mathbf{x}}_{i,j})}$$

vertices from light \swarrow \nwarrow # vertices from eye



MIS is a general-purpose variance reduction technique for Monte Carlo integral estimation. Thanks to the formalizing light transport as a pure integration problem, MIS can be applied in this setting.

Bidirectional path tracing applies MIS to combine the pixel estimates of all possible vertex connection techniques. By weighing each technique proportionally to its sampling pdf, it gracefully handles geometric singularities – techniques that do not importance sample the high-magnitude geometry term are assigned proportionally lower weights.

Bidirectional path tracing



Here we compare the commonly used MIS combination between unidirectional sampling and next-event estimation against full bidirectional path tracing (BPT). Thanks to combining many more techniques, BPT handles the complex lighting in this scene a significantly more robustly, especially the caustics on the table.

Unfortunately, BPT is notoriously inefficient in rendering caustics that are seen through reflection or refraction. This is because none of the techniques it combines can sample (i.e. find) such specular-diffuse-specular paths with high enough probability. In literature, this issue has been referred to as “the problem of insufficient techniques”. Recent work has addressed this problem, as we will discuss next.

Combining bidirectional path tracing and photon mapping

In the previous section, we showed how the path integral formulation of light transport enables robust estimation by efficiently combining various sampling techniques. We will now discuss how we can leverage this framework to address the problem of insufficient techniques by combining photon mapping and bidirectional path tracing via MIS.



Bidirectional path tracing is one of the most versatile light transport simulation algorithms available. It can robustly handle a wide range of illumination and scene configurations, but is notoriously inefficient for specular-diffuse-specular light interactions. This is seen in the caustic reflections seen in the mirror and the window.



And here is the same scene rendered with progressive photon mapping. Photon mapping [Jensen 1997] is well known for its efficient handling of caustics, and this progressive variant [Hachisuka and Jensen 2009] converges to the correct result with a fixed memory footprint. It reproduces the reflected caustics in the scene well, but it has a hard time handling the glossy reflections on the table and the strong distant indirect illumination coming from the part of the scene behind the camera.

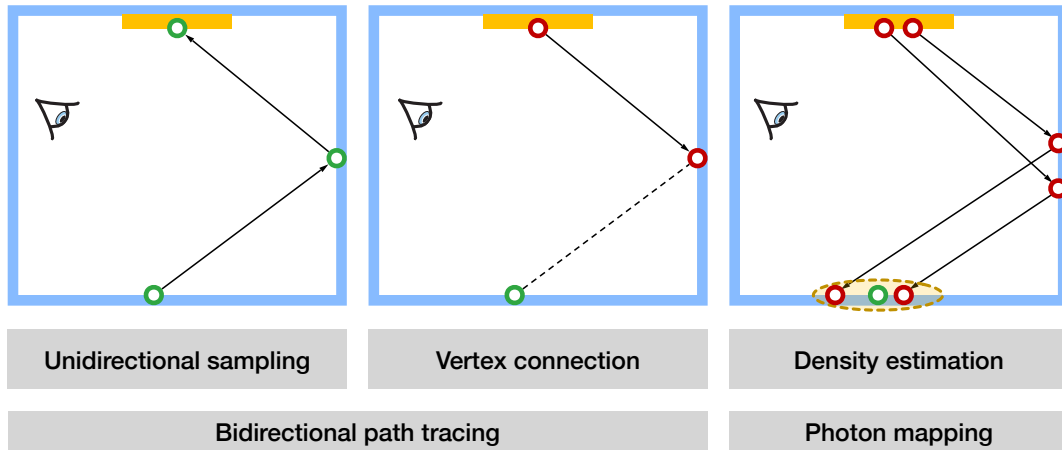
Henrik Wann Jensen. 2001. Realistic Image Synthesis Using Photon Mapping. A. K. Peters, Ltd., Natick, MA, USA.

Toshiya Hachisuka and Henrik Wann Jensen. 2009. Stochastic progressive photon mapping. ACM Trans. Graph. 28, 5. doi.org/fv7fmg



We will show how to combine estimators from bidirectional path tracing (BPT) and photon mapping (PM) to find a good mixture of techniques for each individual light transport path and to produce a clean image in the same amount of time.

Bidirectional path tracing vs photon mapping



Let us quickly review the techniques BPT and PM use to construct light transport paths connecting the eye and the light sources.

The BPT techniques can be roughly categorized to unidirectional sampling (US) and vertex connection (VC). US samples a path by starting either from a light source or the eye and performs a random walk until termination. On the other hand, VC traces one subpath from the eye and another subpath from a light source, connecting their endpoints.

In contrast, PM first traces a number of light subpaths and stores their hit points (a.k.a. photons). It then traces subpaths from the eye and employs photon density estimation to compute the outgoing radiance at the eye hit points.

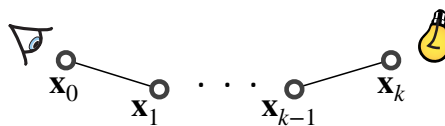
Problem & solution

⊗ Problem: different mathematical frameworks

- **BPT**: Monte Carlo integration
- **PM**: Density estimation

👉 **Key idea**: *Reformulate photon mapping as a path sampling technique*

- Formalize a path sampling technique



- Derive path pdf

$$p(\bar{\mathbf{x}}) = p(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{k-1}, \mathbf{x}_k)$$

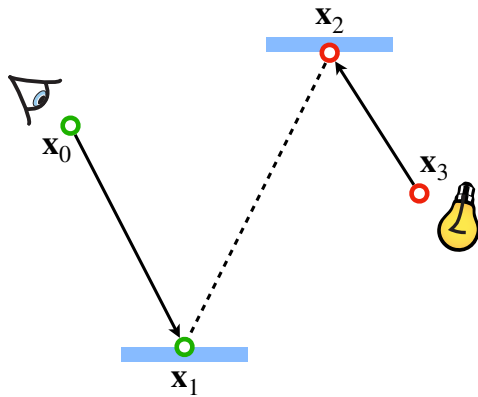
It has been long recognized that BPT and PM complement each other in terms of the light transport effects they can efficiently handle. However, even though both methods have been published more than 20 years ago, a rigorous analysis of their relative performance and their efficient combination have remained elusive until very recently. The reason for this is that BPT and PM have originally been formulated in different theoretical frameworks – BPT as a standard Monte Carlo estimator for the light transport integral, and PM as an outgoing radiance estimator based on photon density estimation.

The first step toward combining these two methods is to put them in the same mathematical framework. The path integral framework is a natural choice: it already subsumes the BPT techniques and also hosts MIS.

We need to do two things: (1) express PM as a sampling technique that constructs light transport paths connecting the light sources and the camera, and (2) derive the pdf for the paths sampled with that technique. This will give us a basis for reasoning about the relative efficiency of BPT and PM. And more importantly, it will lay the ground for combining their corresponding estimators via MIS.

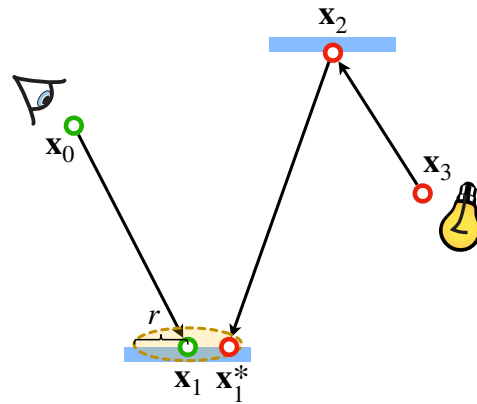
Bidirectional sampling

- Camera vertex
- Light vertex



Vertex connection

$$p_{VC}(\bar{\mathbf{x}}) = p(\mathbf{x}_0) p(\mathbf{x}_0 \rightarrow \mathbf{x}_1) \times p(\mathbf{x}_3) p(\mathbf{x}_3 \rightarrow \mathbf{x}_2)$$



Photon mapping

$$p_{PM}(\bar{\mathbf{x}}) = p(\mathbf{x}_0) p(\mathbf{x}_0 \rightarrow \mathbf{x}_1) \times p(\mathbf{x}_3) p(\mathbf{x}_3 \rightarrow \mathbf{x}_2) p(\mathbf{x}_2 \rightarrow \mathbf{x}_1^*)$$

Let us consider a simple length-3 path. We first trace one subpath from the camera and another one from a light source. Now let us see how we can complete a full path.

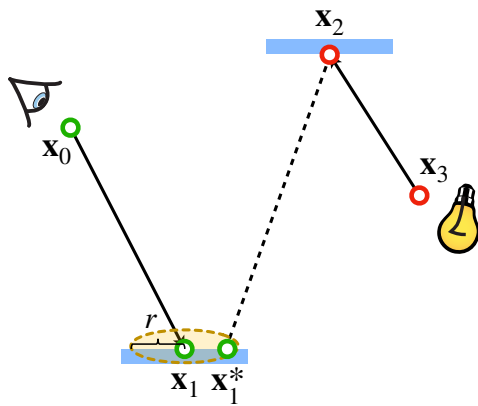
Bidirectional path tracing (left) connects the subpath endpoints deterministically. We call this technique vertex connection (VC). The sampling density of the resulting full path is simply the product of the densities of two independently sampled subpaths.

On the other hand, photon mapping (right) extends the light subpath with one more vertex. Pixel contribution is made if that “photon” lands within some distance r from the eye subpath end point. The joint pdf of all sampled vertices is derived similarly to VC, since again the subpaths are sampled independently.

However, this is not sufficient for applying MIS to combine with VC. The reason is that with this interpretation the two methods sample paths with different numbers of vertices, and consequently their pdfs have different units. Plugging these PDF into MIS wouldn't produce a meaningful result, because the heuristics expect all pdf to be expressed w.r.t. the same measure. A meaningful MIS combination needs the pdfs to have the same measure.

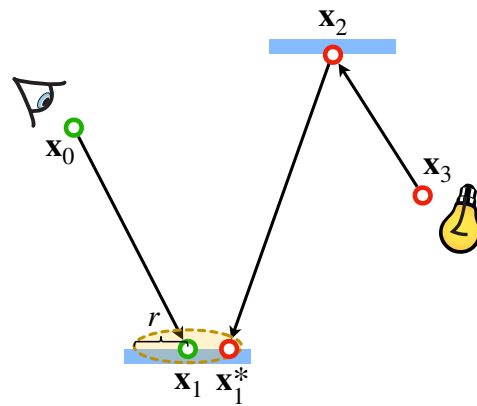
Extended path space formulation

- Camera vertex
- Light vertex



Extended vertex connection

$$p_{VC^*}(\bar{\mathbf{x}}) = p(\mathbf{x}_0) p(\mathbf{x}_0 \rightarrow \mathbf{x}_1) p(\mathbf{x}_1 \rightarrow \mathbf{x}_1^*) \times \underbrace{p(\mathbf{x}_3) p(\mathbf{x}_3 \rightarrow \mathbf{x}_2)}_{\frac{1}{\pi r^2}}$$



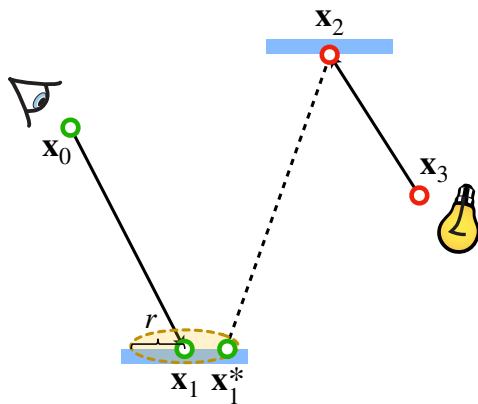
Photon mapping

$$p_{PM}(\bar{\mathbf{x}}) = p(\mathbf{x}_0) p(\mathbf{x}_0 \rightarrow \mathbf{x}_1) \times p(\mathbf{x}_3) p(\mathbf{x}_3 \rightarrow \mathbf{x}_2) p(\mathbf{x}_2 \rightarrow \mathbf{x}_1^*)$$

To address these issues, Hachisuka et al. [2012] express the vertex connection PDF in the higher-dimensional space of photon mapping. They consider an extension of vertex connection that samples a vertex \mathbf{x}_1^* by randomly perturbing the eye vertex \mathbf{x}_1 within an r -neighborhood. This new vertex corresponds to the photon vertex in PM and is connected to the last light subpath vertex. Assuming that the surface in this neighborhood is locally flat, i.e. that the region is a disk, the PDF of the new vertex is $\frac{1}{\pi r^2}$.

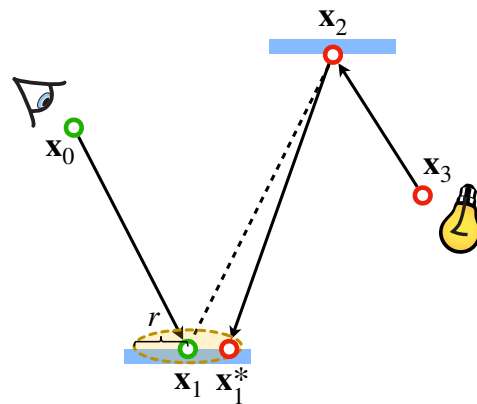
Vertex merging formulation

- Camera vertex
- Light vertex



Vertex connection

$$p_{VC^*}(\bar{\mathbf{x}}) = p(\mathbf{x}_0) p(\mathbf{x}_0 \rightarrow \mathbf{x}_1) \times p(\mathbf{x}_3) p(\mathbf{x}_3 \rightarrow \mathbf{x}_2)$$



Vertex merging

$$p_{PM}(\bar{\mathbf{x}}) = p(\mathbf{x}_0) p(\mathbf{x}_0 \rightarrow \mathbf{x}_1) \times p(\mathbf{x}_3) p(\mathbf{x}_3 \rightarrow \mathbf{x}_2) \underbrace{P\left(\|\mathbf{x}_1 - \mathbf{x}_1^*\| < r\right)}_{p(\mathbf{x}_2 \rightarrow \mathbf{x}_1^*) \pi r^2}$$

Alternatively, we can stick with the original VC technique and instead express PM as a technique in the lower-dimensional space of VC [Georgiev et al. 2012].

To that end, we can interpret the PM sampling process as establishing a regular vertex connection between the end points \mathbf{x}_1 and \mathbf{x}_2 , but conditioning its acceptance on the random event that the “photon” vertex \mathbf{x}_1^* sampled from \mathbf{x}_2 lands within a distance r to \mathbf{x}_1 . This probabilistic acceptance is simply a Russian roulette decision.

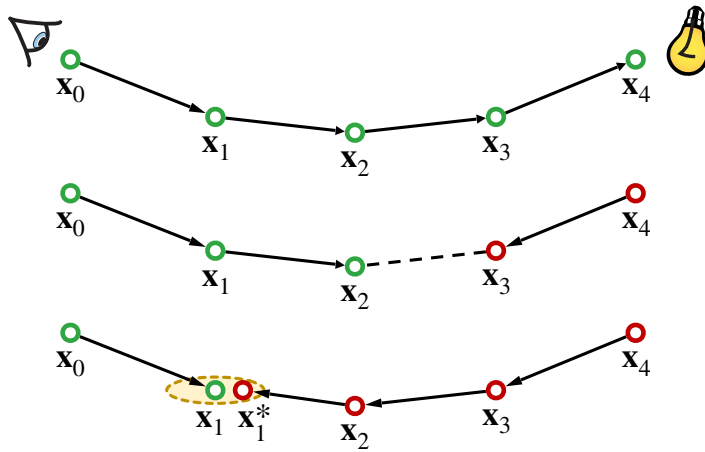
The full path pdf is then the product of the two VC subpath pdfs, as on the left, but in addition multiplied by the probability of sampling the photon vertex \mathbf{x}_1^* within in an r -neighborhood of \mathbf{x}_1 . This acceptance probability is the integral of the pdf of \mathbf{x}_1^* over that neighborhood. Assuming that this neighborhood is a disk again, and also that the density of the photon is constant inside this disc, the integral can be approximated by the pdf of the actually sampled point \mathbf{x}_1^* , multiplied by the disc area πr^2 .

We dub this technique *vertex merging* (VM), as it can be intuitively thought to weld the endpoints of the two subpaths if they lie close to each other.

Note that while in the interpretation of Hachisuka et al. [2012] we had πr^2 in the VC pdf denominator, in the VM interpretation, this term appears in the pdf numerator. Both interpretations result in the same MIS combination weights. In the remainder of the discussion we will use the VM interpretation, but the final combined algorithm I will present is identical with both interpretations.

Available sampling techniques

- Camera vertex
- Light vertex

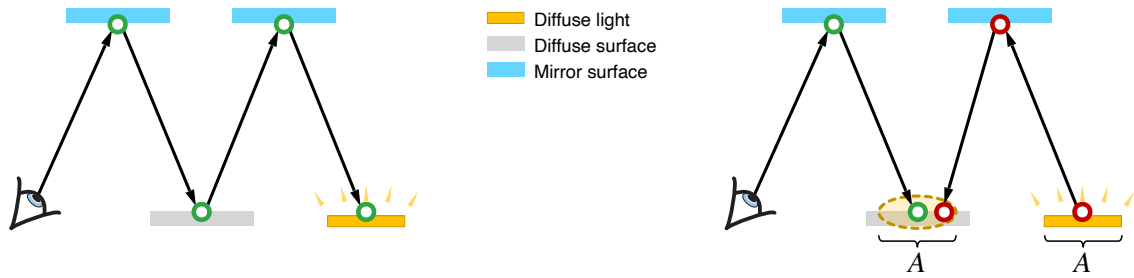


| | |
|-------------------|----------------|
| Unidirectional | 2 ways |
| Vertex connection | 4 ways |
| Vertex merging | 5 ways |
| Total | 11 ways |

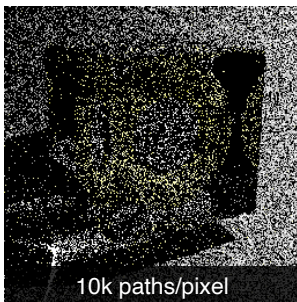
Having formulated the vertex merging path sampling technique, we can put it side by side with the already available techniques in BPT. There are two ways to sample a length-4 path unidirectionally, and four ways with vertex connection. Vertex merging adds five new ways to sample the path, corresponding to merging at the five individual path vertices. In practice, we can avoid merging at the light source and the camera.

But with so many ways to sample the same light transport path, one might ask: How efficiently do these different techniques handle various types of paths?

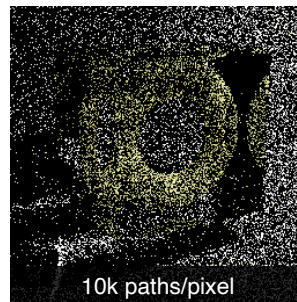
Caustic case: with path reuse



Unidirectional sampling



Vertex merging

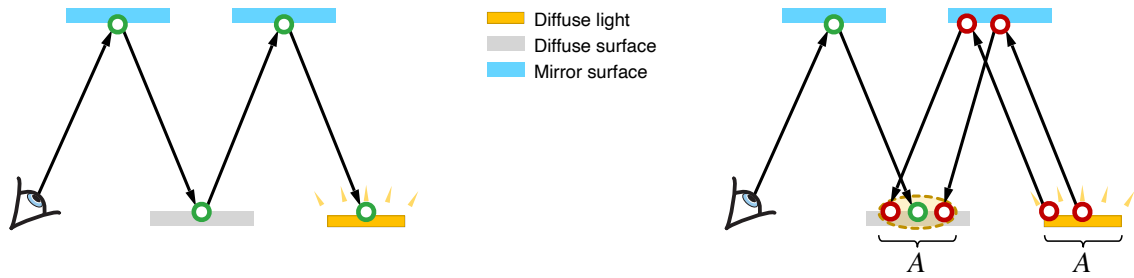


To answer this question, let us first take a look at specular-diffuse-specular (SDS) paths. Here, BPT can only rely on unidirectional sampling: it is forced to trace a path from the camera and hope it randomly hits the light source. With vertex merging, we can trace one light and one camera subpath, and merge their endpoints on the diffuse surface.

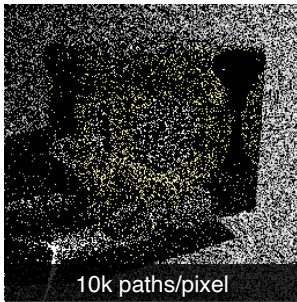
It can be shown that if the light source and the merging disk have the same area, then unidirectional sampling and vertex merging sample paths with roughly the same probability density. Perhaps surprisingly, this means that we should expect the two techniques to perform similarly in terms of rendering quality as they are equally likely to find such paths.

To verify this, we render these images with both the US and VM techniques progressively, sampling one full path per pixel per iteration: For US we trace paths from the camera until they hit the light or escape the scene. For VM, we trace subpaths from both ends, and merge their endpoints if they lie within a distance from each other. Both images look equally noisy, even after sampling 10,000 paths per pixel. This result confirms that vertex merging, and thus photon mapping, is not an intrinsically more efficient sampling technique for SDS paths than unidirectional sampling.

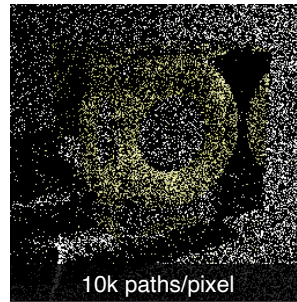
Caustic case: with path reuse



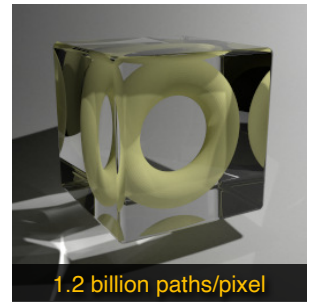
Unidirectional sampling



Vertex merging



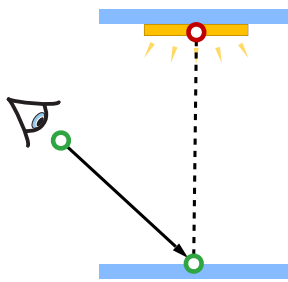
Vertex merging (reuse)



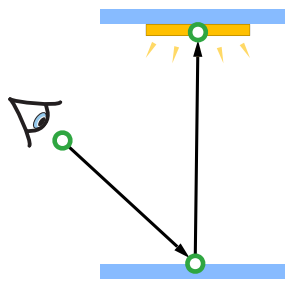
However, VM has one advantage – computational efficiency. For each pixel, we can very cheaply reuse light subpaths traced for all other pixels, at the cost of one range search query. This allows us to quickly construct orders of magnitude more light transport estimators from the same sampling data, and with minimal computational overhead, resulting in a substantial quality improvement.

For all three images above we have traced roughly the same number of rays. The only difference between the center one and the right one is that for the right one we have enabled path reuse: at every rendering iteration we store and look up the light-subpath vertices in a photon map. It is this efficient path reuse that makes PM better than BPT for SDS paths.

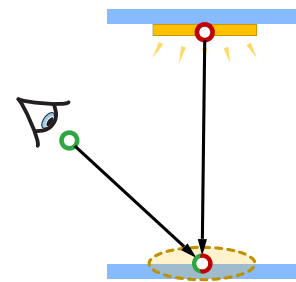
Diffuse case



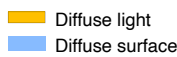
Vertex connection



Unidirectional sampling



Vertex merging



Roughly equal sampling densities

$$p_{US} = p_{VM} \approx \frac{p_{VC}}{100,000}$$

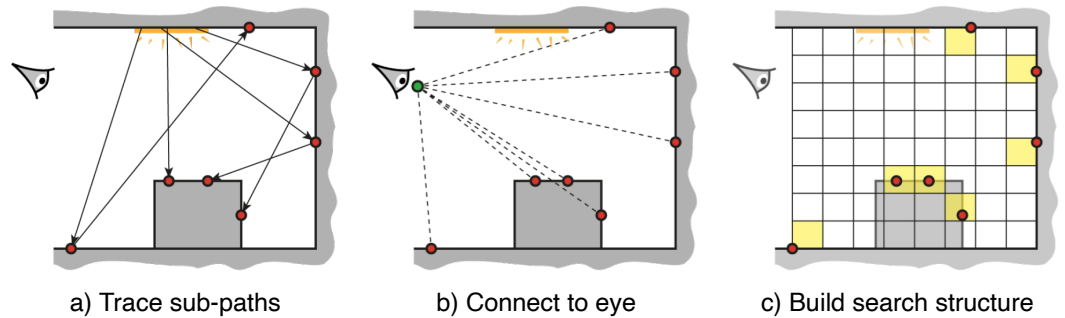
Let us also look at another extreme example – direct illumination on a diffuse surface. Here, VC can construct a connection edge to the vertex sampled on the emitter, while US and VM both rely on random direction sampling from or to the light.

Once again, it can be shown that if the emitter and the merging disk have the same area, then US and VM sample this path with roughly equal probability density. For the specific case shown here, this density is about 100,000 lower than that of VC. This demonstrates that VM is not intrinsically more efficient than VC either. This is not surprising if we recall the expression for the VM path pdf. That pdf can be at most equal to that of its VC counterpart, since the VM pdf additionally multiplies by an acceptance probability. Nevertheless, by reusing paths across pixels, VM, and thus photon mapping, gains a lot of efficiency over US.

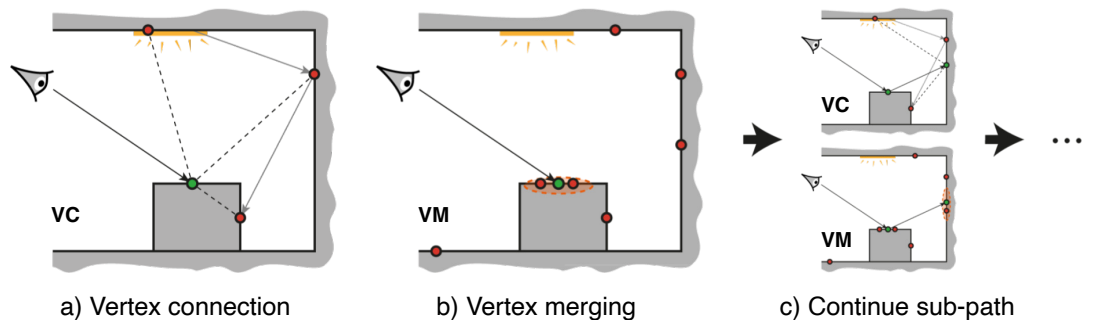
Note that all these insights emerge from the formulation of photon mapping as a path sampling technique.

A combined algorithm

Stage 1: Light sub-path sampling



Stage 2: Eye sub-path sampling

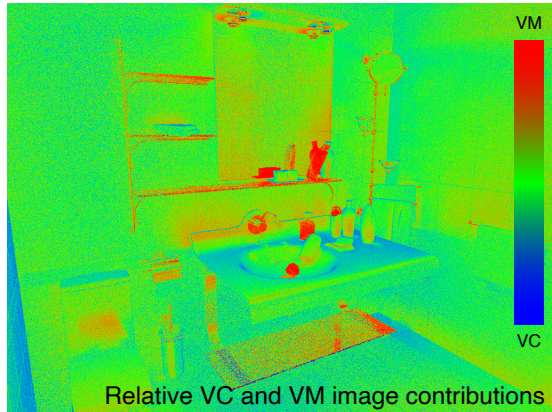


We now have the necessary ingredients to combine PM and BPT into one unified algorithm. The US, VC, and VM path pdf can be used to weight the techniques in MIS, and the insights from the previous two slides urges us to strive for path reuse.

The combined algorithm, which we call vertex connection and merging (VCM), operates in two stages: In the first stage, we trace the light subpaths for all pixels, connect them to the camera, and store them in a range search acceleration data structure (e.g. a kd-tree or a hashed grid). In the second stage, we trace an eye subpath for every pixel. Upon generating each eye subpath vertex, we (1) connect it to a light source, (2) connect it to the vertices of the light subpath paired with that pixel, and (3) merge it with the vertices of all light subpaths. We then sample the next eye subpath vertex and recurse.

In a progressive rendering setup, we can perform these two steps once per iteration and reduce the vertex merging radius r thereafter.

Note that while (P)PM performs merging at a single vertex along each eye subpath, VCM performs merging at *every* such vertex, thereby employing significantly more techniques.



This scene with various glossy and specular materials is especially difficult for both BPT and (P)PM. BPT under-samples reflected and refracted caustics, whereas PPM is known to handle glossy surfaces inefficiently. The combined VCM algorithm is equipped with more sampling techniques than the other two methods combined and extracts the best of each to produce an image free of fireflies.

We also visualize the relative contributions of VM and VC techniques to the VCM image. We can observe the regions in which VCM assigns more weight to each of the two family of techniques.



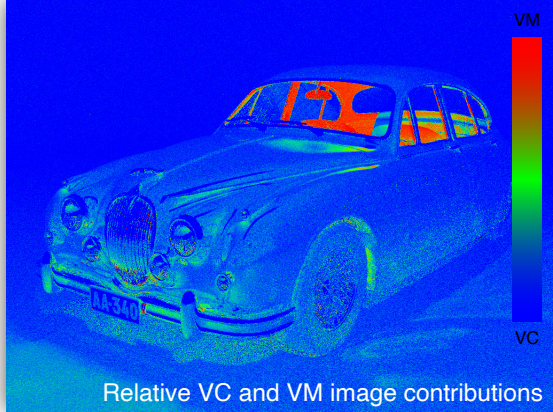
Bidirectional path tracing



Stochastic progressive photon mapping



Vertex connection and merging



Relative VC and VM image contributions

The results on this scene are very similar.

Summary

Reformulate photon mapping as a path sampling technique

Efficient MIS combination with bidirectional path tracing

- ▶ Improved convergence rate over progressive photon mapping

Reformulating PM as a sampling technique in the path integral framework allows us to augment BPT with techniques that efficiently handle the notoriously difficult specular-diffuse-specular light transport.

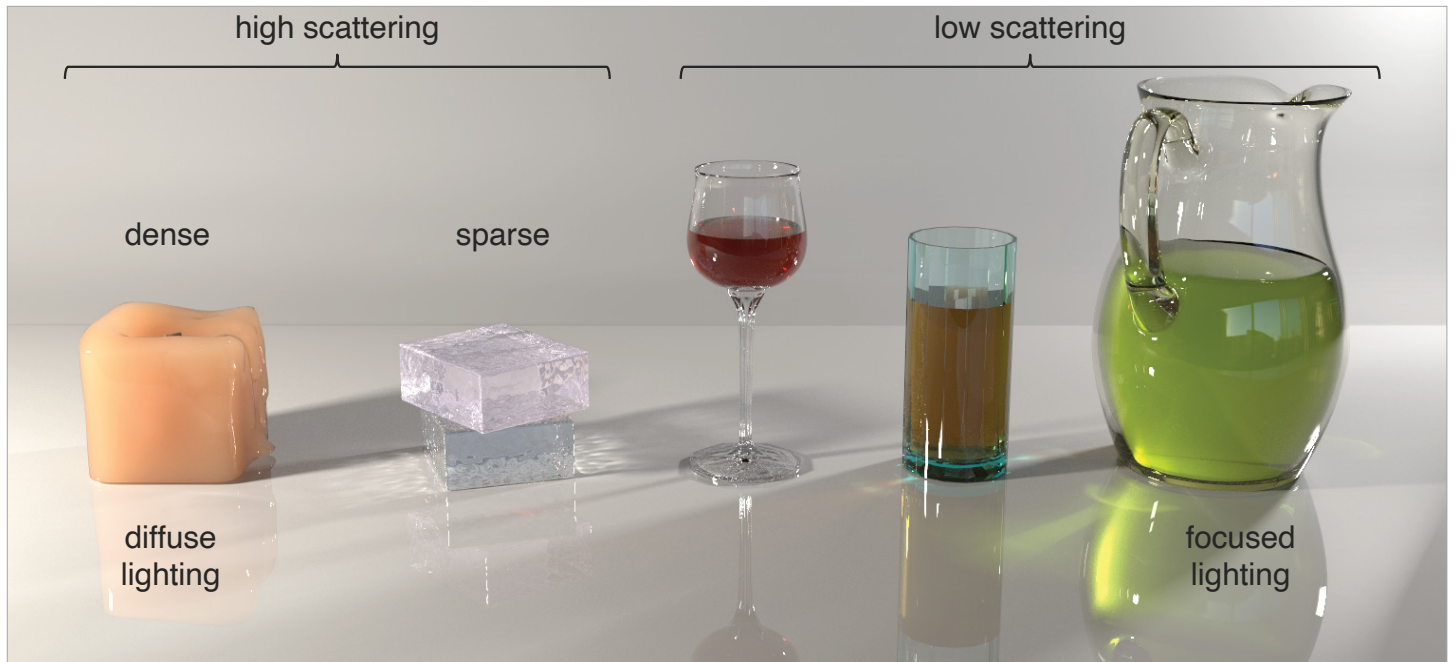
An important property of the combined VCM algorithm is that it retains the higher convergence rate of BPT. This means that it approaches the correct solution faster than PPM as the computational effort increases, i.e. as we sample more paths. In fact, VCM is asymptotically equivalent to BPT, since the MIS weight of VM techniques vanishes as the merging radius progressively goes to zero. However, the VM contributions bring a significant initial variance reduction.

Even though VCM has proven very useful in practice, it has some limitations. Most importantly, it does not improve over BPT and PM for light transport that is difficult to both BPT and PM. A prominent example are caustics falling on a glossy surface.

Combining points, beams, and paths in participating media

Given the success of VCM in handling surface scattering, a logical next step is to try transfer its ideas to improve rendering of participating media.

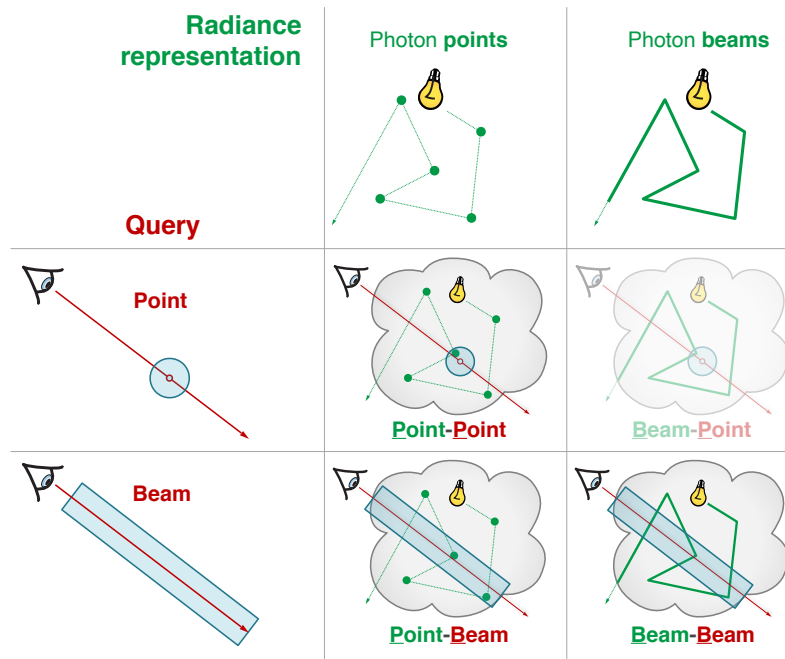
Goals



As before, we are looking for an algorithm that can render participating media in a manner that is robust to media properties and to lighting, as demanded by the scene shown here. We want to handle optically dense or rare media with high or low scattering albedo. We want to handle diffusive multiple scattering (as in subsurface scattering) or highly focused lighting (as in volumetric caustics).

The algorithm we will discuss has all these features and was actually used to render the above image.

Volumetric photon-based estimators

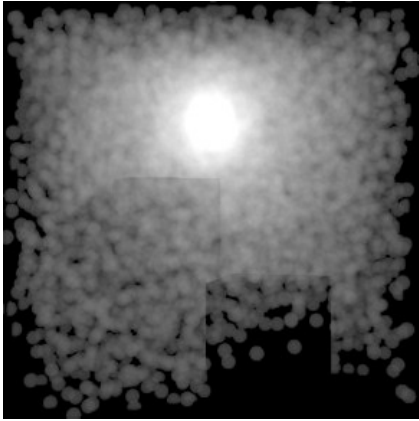


[Jarosz et al. 2011]

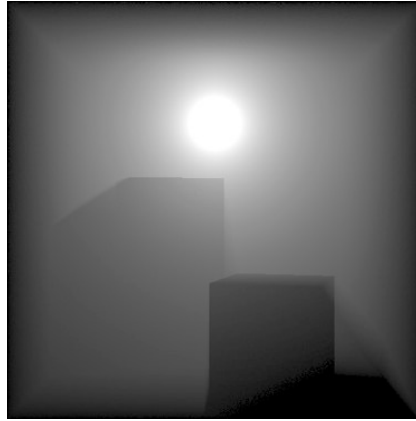
In addition to volumetric bidirectional path tracing (BPT), for media we also have techniques derived from photon density estimation, such as volumetric photon mapping (VPM), the beam radiance estimate (BRE), and photon beams (PB).

In media, we can represent radiance either by particles (i.e. photon points) or by particle tracks (i.e. photon beams). The radiance estimate can then be performed at one point or along an entire ray (i.e. a query beam). This gives us four basic types of estimators: point-point, beam-point, point-beam, and beam-beam. In practice, we do not use the beam-point estimator because it has similar properties to the point-beam one but with a much less efficient implementation. Even with that, a relevant question is: Does it make sense to combine all the estimators or do some always perform better than others?

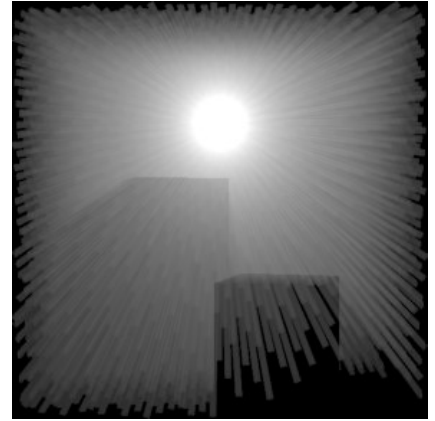
Points vs beams



100k photon **points**



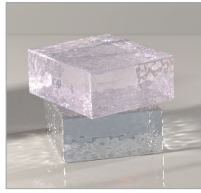
Reference



5k photon **beams**

Above we show images rendered with photon points (VPM) and photon beams, at intentionally low sample counts to illustrate the error they produce. Intuitively, one may expect that because the beams fill up the space so much better, they should always perform better than points. But in reality, while photon beams are very efficient in some types of media, they may be outperformed by points in other media.

Points vs beams



Sparse media



Dense media

Beams:

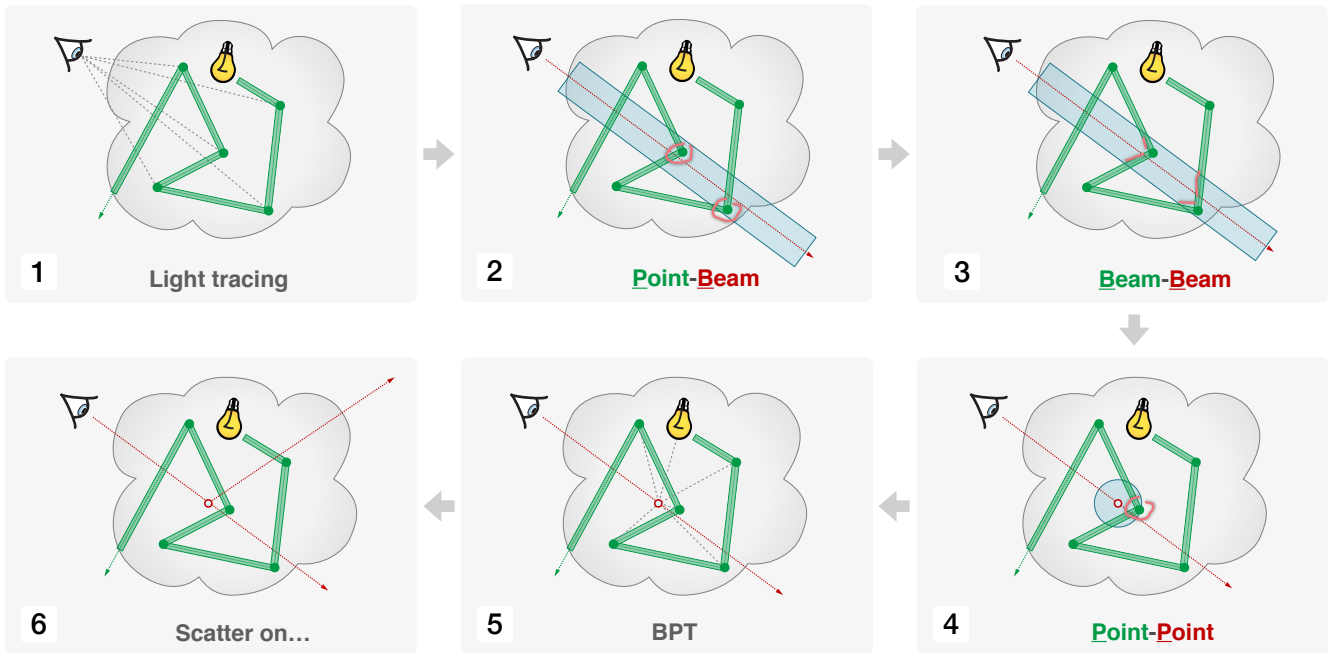


Points:



It turns out that beams are better in rare media, where the mean free path (MFP) is much longer than the density-estimation kernel size. On the other hand, in dense media, when the MFP is shorter than the kernel size, points perform better.

Combined algorithm (UPBP)



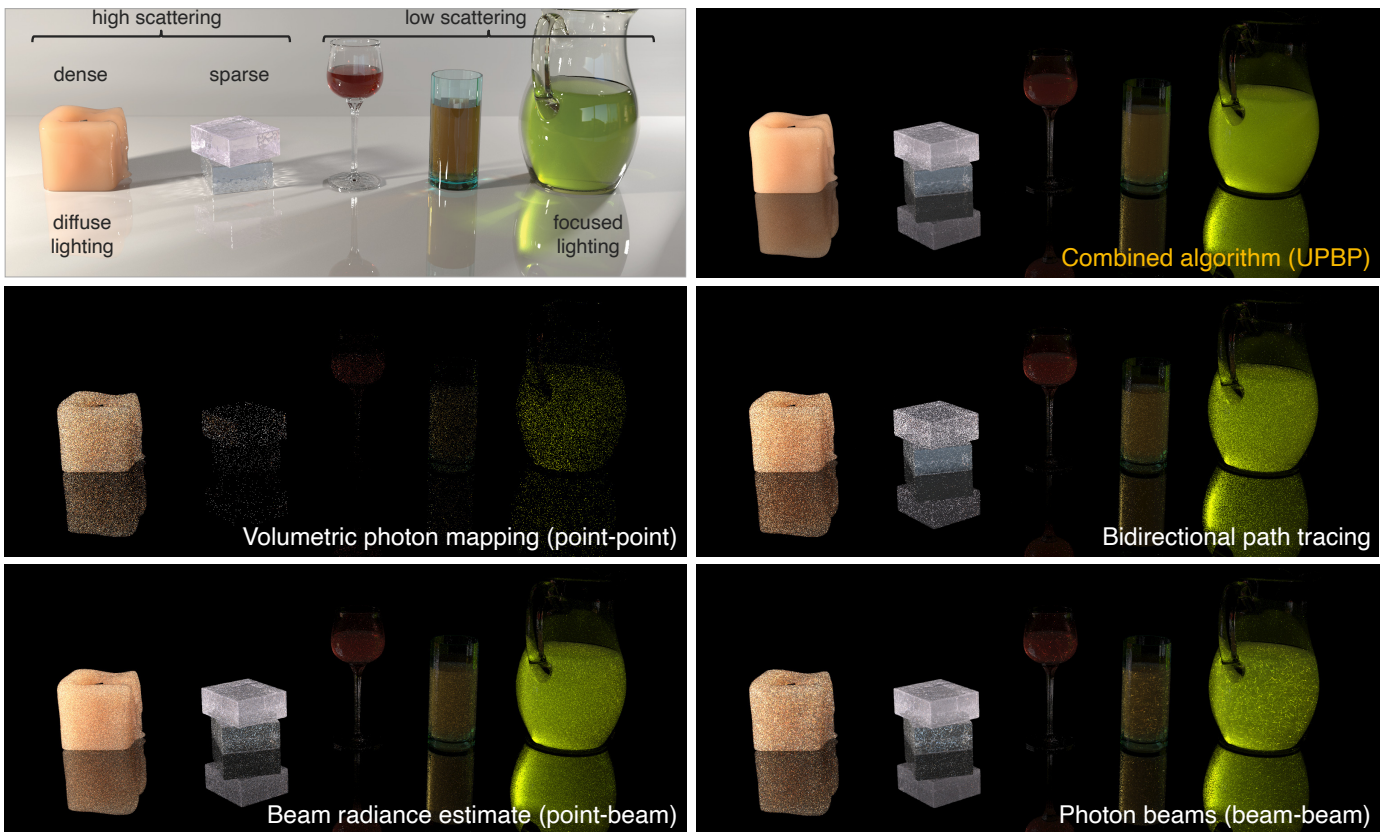
The derivation of the path pdfs of these techniques is similar in spirit to that of photon mapping discussed above. The algorithm that combines these techniques with BPT also proceeds similarly to VCM.

In each progressive rendering iteration, we start by tracing a number of paths from the light sources. We connect their vertices to the eye, which corresponds to light tracing (1). We store the vertices as photon points, and the path segments as photon beams.

Next, we trace eye subpaths through each pixel. For each segment of a subpath, we look up the photons and evaluate to the point-beam estimator (2). We then look up the beams, which is the beam-beam estimator (3).

Then we choose a scattering location long the query ray and look up the photons around that point to evaluate the point-point estimator (4). We also connect the scattering vertex to the vertices of the paired light subpath, which corresponds to BPT (5).

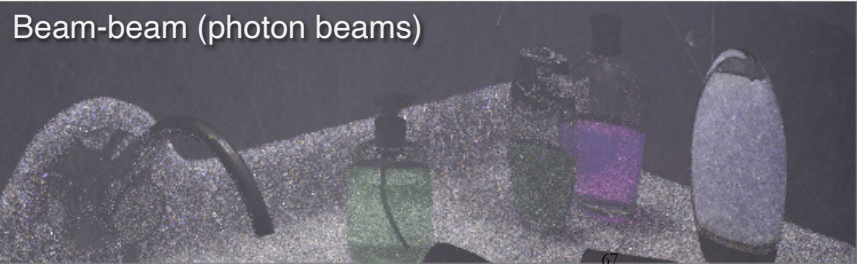
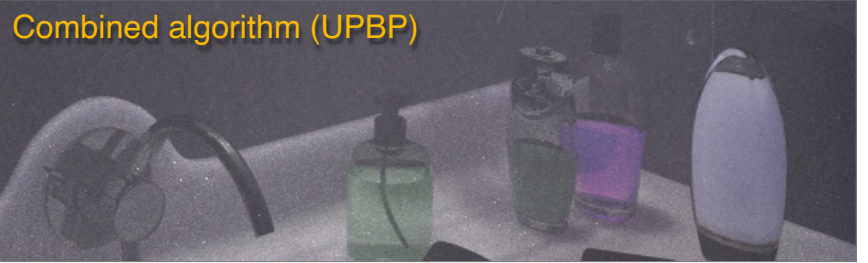
Finally, we extend the eye subpath and repeat (6).



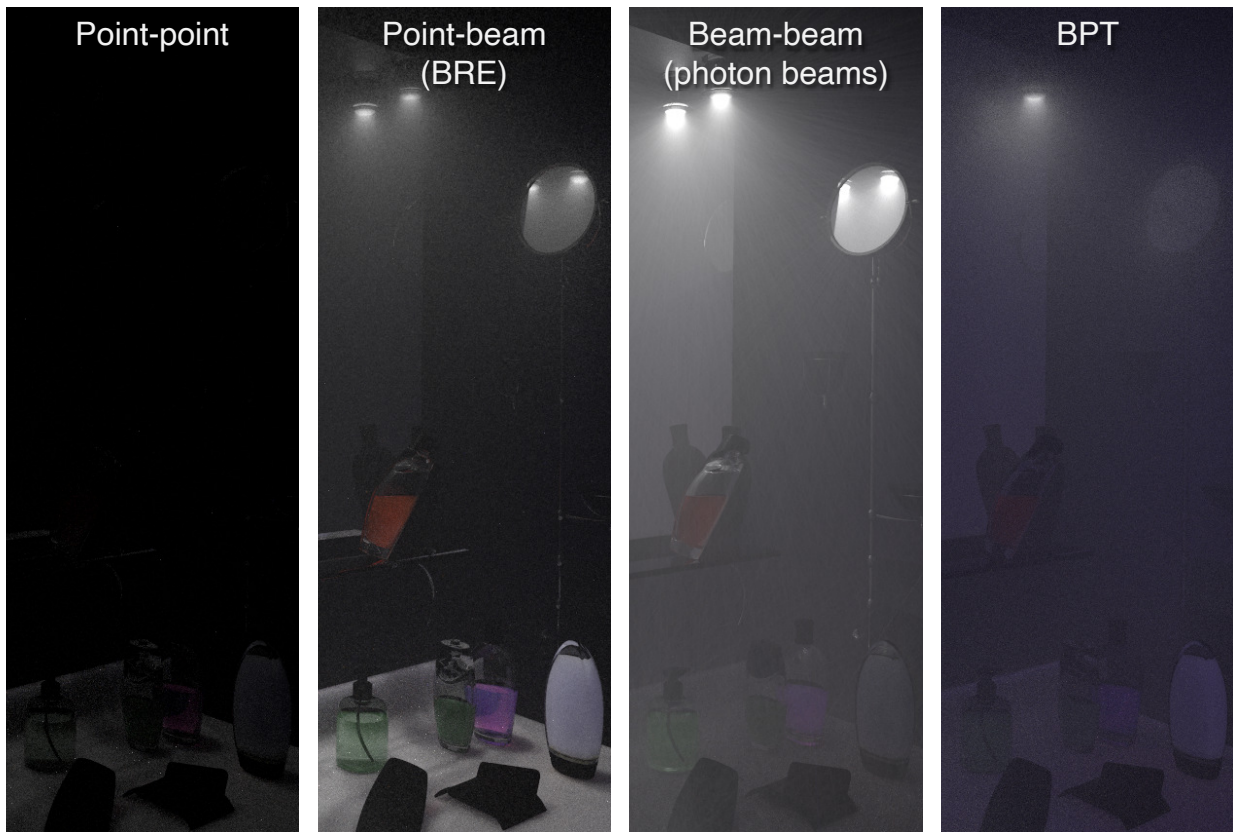
Let us now see how the various methods perform on the scene from earlier.

In this scene, the BPT image remains noisy even after an hour of rendering. Volumetric photon mapping (point-point) overall performs worst. Beam radiance estimate (point-beam) much better but still not great. Photon beams perform well only on the thin soap medium, producing conspicuous artifacts in the other media that resemble the beam shapes.

The UPBP algorithm is able to produce a much cleaner image in the same amount of time. It is worth pointing out that even though none of the previous algorithms handle this scene well, their combination is almost noise-free. This provides some evidence that the MIS-based combination is more robust than a heuristic combination that would be based on selecting a particular estimator for each medium.



Here, point-beam (top) handles the dense media much better than photon beams (bottom) and vice versa for the sparser fog. UPBP takes the best of both.



Similarly, when we look at the weighted contributions, dense media like the wash-basing are mostly covered by the point-beam (BRE) estimator, thinner media like the fog by the photon beams.

The fact that BPT is in charge of the surface-to-media transport is quite apparent here: it resolves the blue tint to the media due to reflections from the blue tiles on the walls.

Summary

Beams not always better than points

- ▶ Sparse media: beams
- ▶ Dense media: points

Efficient MIS combination

- ▶ But considers only variance

Available techniques are often *too many*

An important result of this work is that beams are not always better than points. Beams are better in sparse media, and dense media are better handled by points.

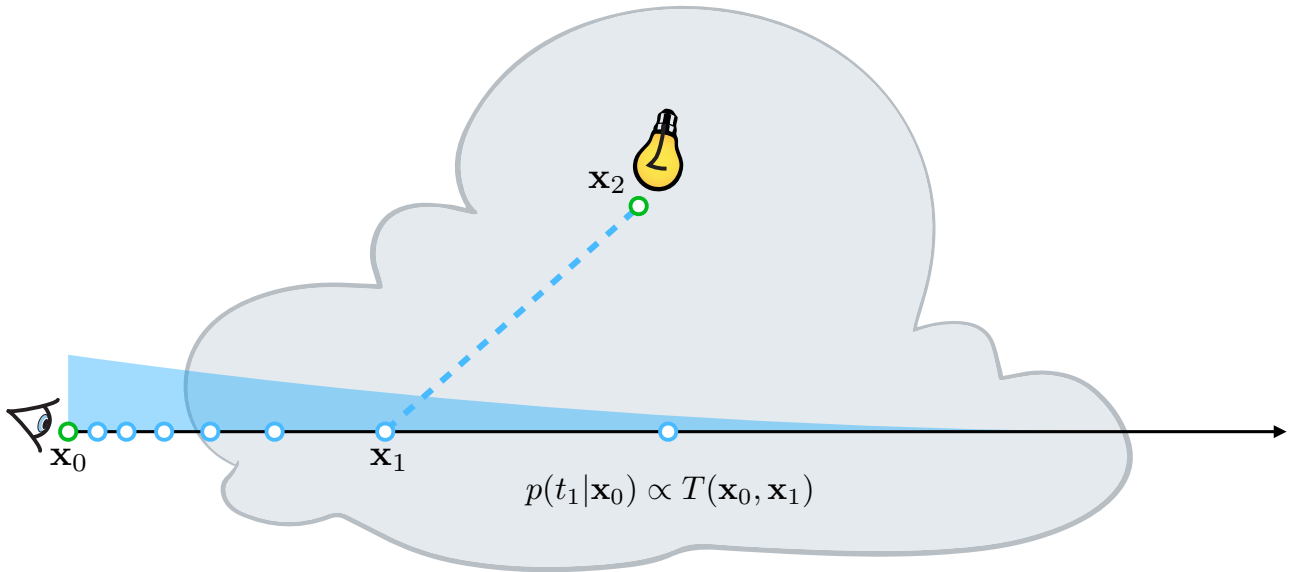
Formalizing volumetric photon density estimation methods as path sampling techniques allows combining them with traditional sampling techniques with MIS to more robustly handle a wide range of media. However the combination relies is based only on variance considerations; taking bias and efficiency into account could significantly improve the results.

We also have to pay a price for combining all the estimators: if one medium is best handled by just one estimator, running the other ones only incurs overhead. Having a solid theory that would indicate how many samples to take from each estimator would be extremely useful, especially in the cases where some estimators could be completely disabled.

Joint path sampling in participating media

The methods we have discussed thus far are all based on generating paths vertex by vertex, using only *local* information to determine the position of the next vertex. However, the path integral view of light transport allows for more flexibility in the sampling, namely coordinating the sampling decisions across vertices. We will now show how such *global sampling* can produce substantial noise reduction in participating media.

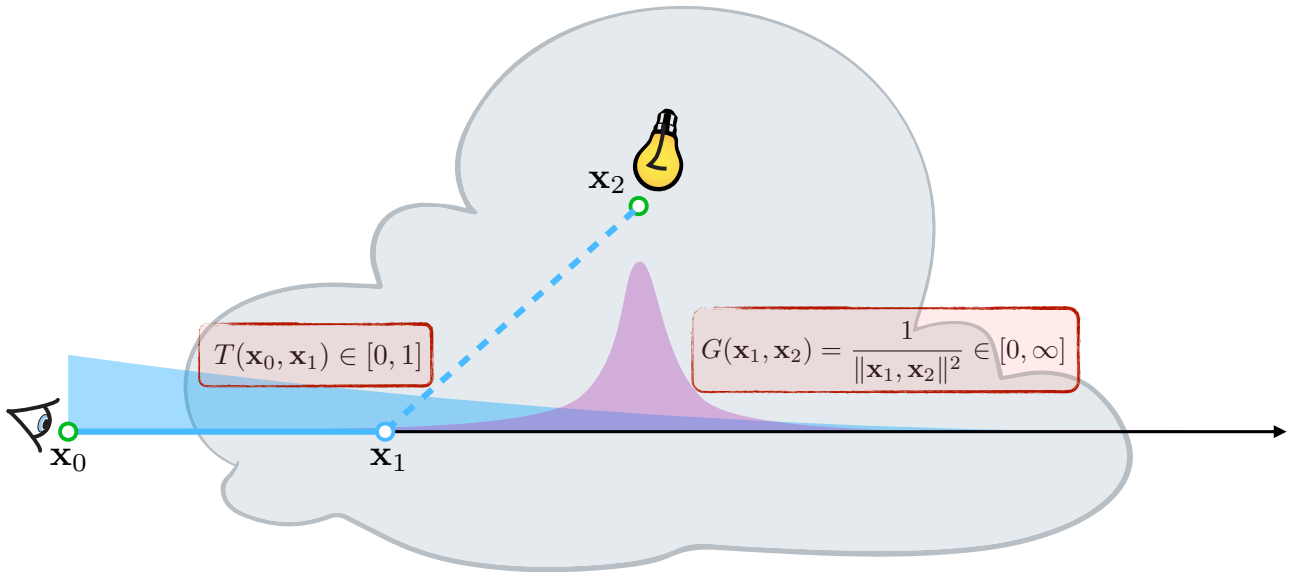
Explicit light sampling: Transmittance



Given a camera ray and a point on a light source, computing direct lighting (a.k.a. single scattering) in media involves sampling a distance along the ray which determines a scattering location that is finally connected to the light point.

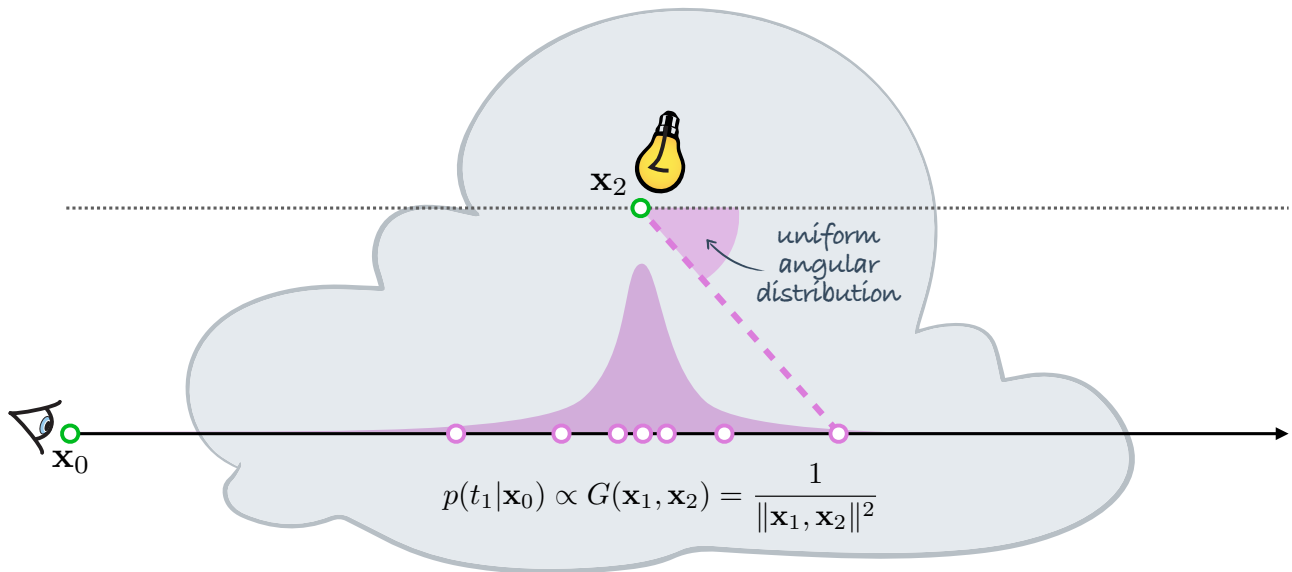
The traditional way of sampling that distance is with density proportional to the transmittance along the given camera ray. This importance sampling scheme ensures that the scattering location is more likely to occur toward the beginning of the ray where the transmittance is high.

Explicit light sampling: Transmittance

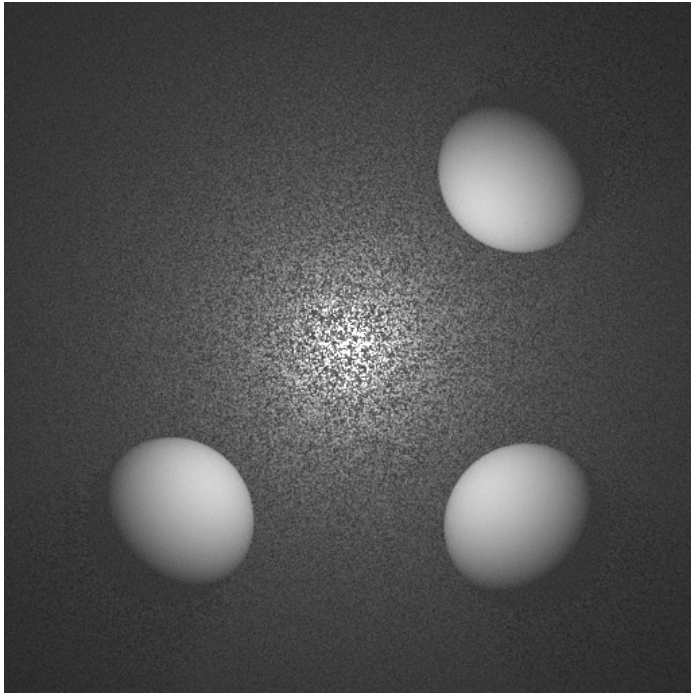


The contribution of the resulting length-2 path also includes the geometry term along the connection segment which, as we discussed previously, is not importance sampled with this scheme. This can be an issue when the light vertex is very close to the ray, creating extreme variation in the geometry term. In contrast, the transmittance along the ray is always bounded between zero and one.

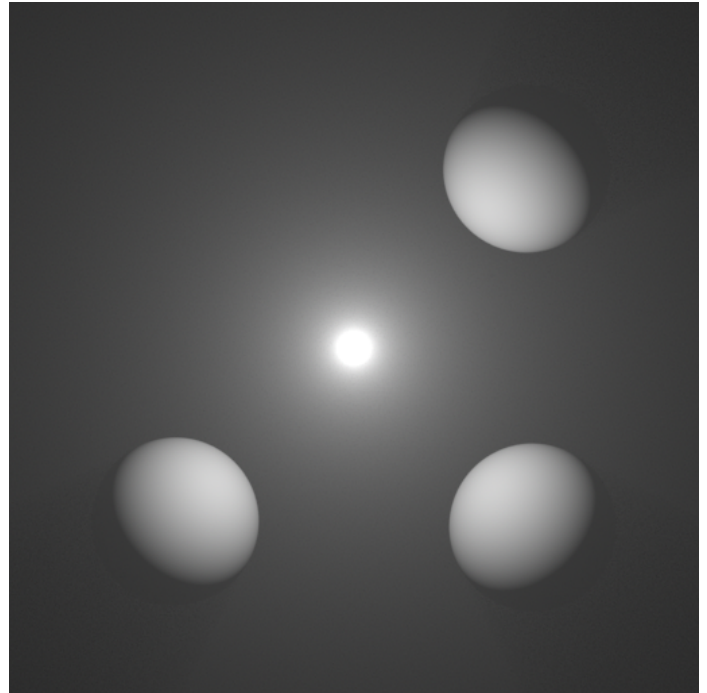
Explicit light sampling: Equiangular



Kulla and Fajardo [2012] showed that it is possible to instead sample the distance along the ray proportionally to that geometry term, thereby cancelling out its variation. The technique is very simple: it samples the angle between the ray and the connection segment uniformly. Hence, it's dubbed 'equiangular' sampling.



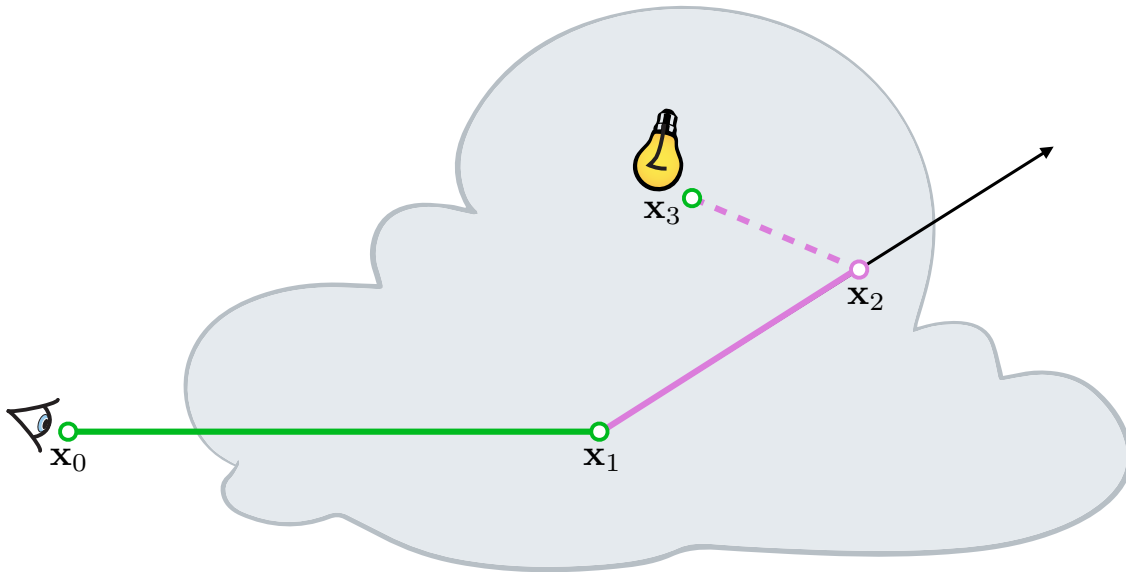
Transmittance sampling, 16 spp



Equiangular sampling, 16 spp

This can make for a substantial noise improvement, as we can see in the comparison above.

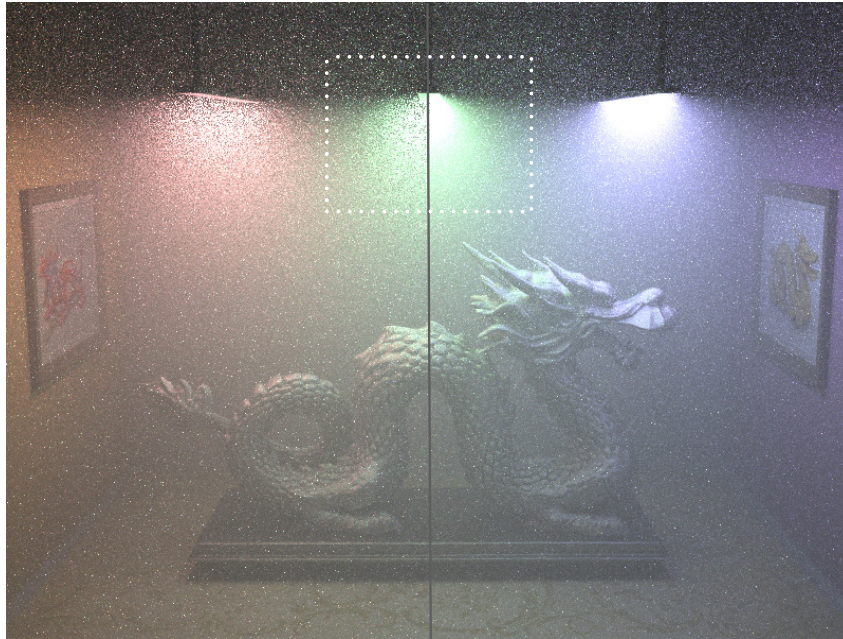
Explicit light sampling: Equiangular



To render higher-order scattering in a path tracer, we can apply equiangular sampling every time we sample a scattering direction.

Note that we now sample two points along each ray: one for the equiangular single scattering and one (transmittance-based) to generate the next path segment for higher-order scattering.

Explicit light sampling: Equiangular

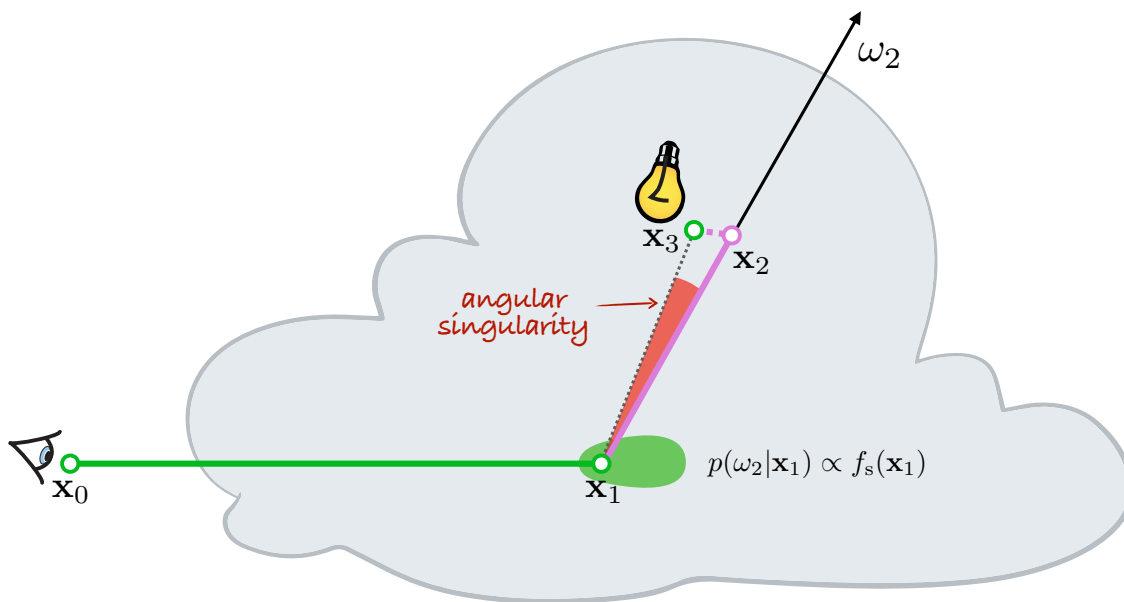


Transmittance connections

Equiangular connections

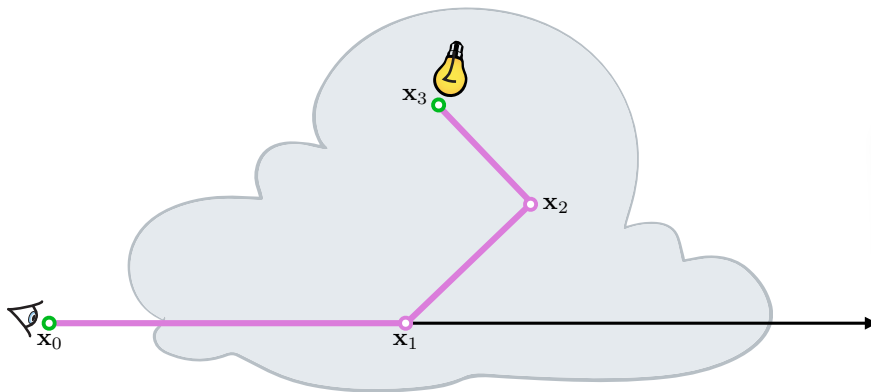
This provides good noise reduction over traditional, purely transmittance-based sampling, especially in the regions around the light source. However, a substantial amount of noise remains, with some pixel estimates having extreme magnitudes.

Unidirectional + explicit



The cause of these fireflies is a singularity in the orientation of the rays along which equiangular sampling is applied. The contribution becomes infinite along rays that align with the direction to the light vertex. However, the ray sampling densities are proportional to the local scattering distributions, disregarding the location of the light vertex.

Local vs joint sampling



Joint path sampling:

- 1) Prescribe joint pdf
- 2) Derive conditional pdfs via successive joint pdf marginalization
- 3) Conditionals are obtained in reverse order

TRADITIONAL: prescribes conditional pdfs, no explicit control over joint pdf

JOINT SAMPLING: prescribe joint pdf, conditional pdfs derived from it

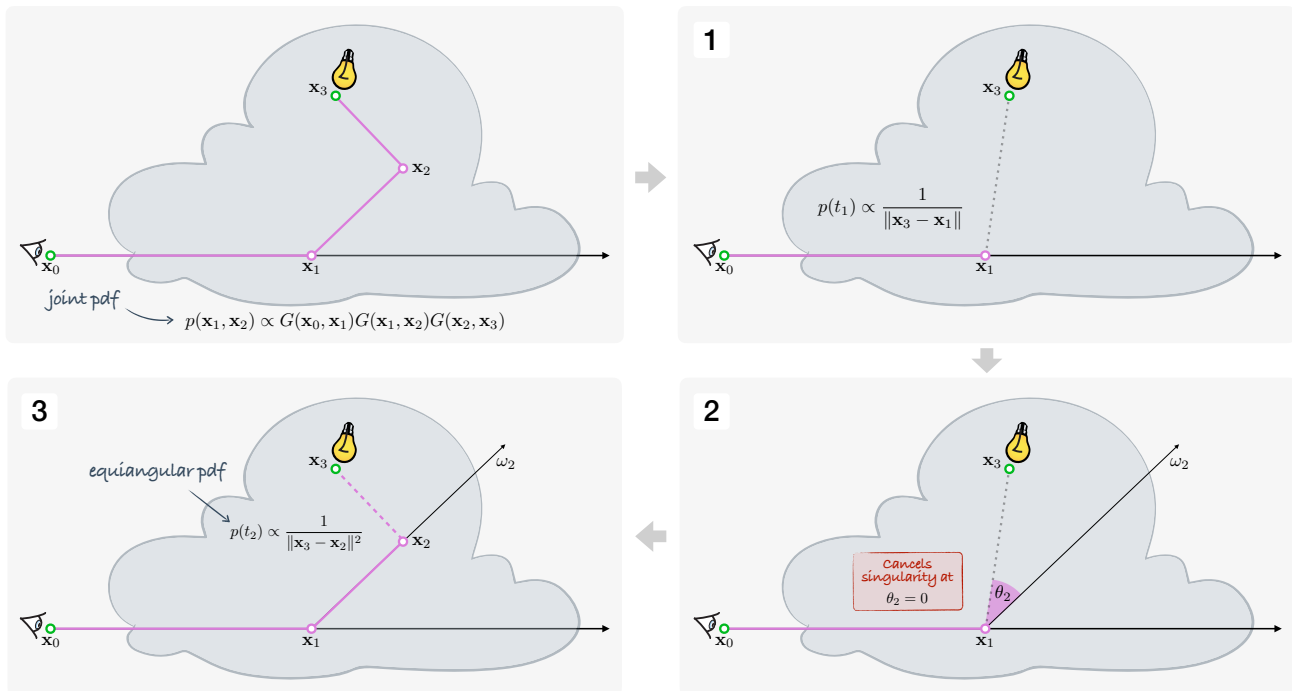
And this brings us back to the point that paths are traditionally constructed incrementally by importance sampling some of the contribution terms and only locally at each vertex. Directional distributions are proportional to the local phase function, and propagation distances along the resulting rays are proportional to the transmittance along them.

The resulting joint path density is then a consequence of these local decisions. We can only hope that it is somewhat proportional to the path contribution, as we ideally want, but we have no explicit control over this. Bidirectional path tracing, which samples paths from both ends, also suffers from this problem. (When constructing two paths independently, they can go in completely opposite directions.)

This answers why existing methods can produce so noisy images: they prescribe the local sampling decisions, and the final joint distribution is only a consequence of these decisions.

Importance sampling theory postulates that we should ideally do is the opposite: prescribe the joint distribution for the entire path, and then *derive* the vertex sampling decisions from that joint. Only this way can we make sure that the path density is indeed proportional to all contribution terms we want to importance sample.

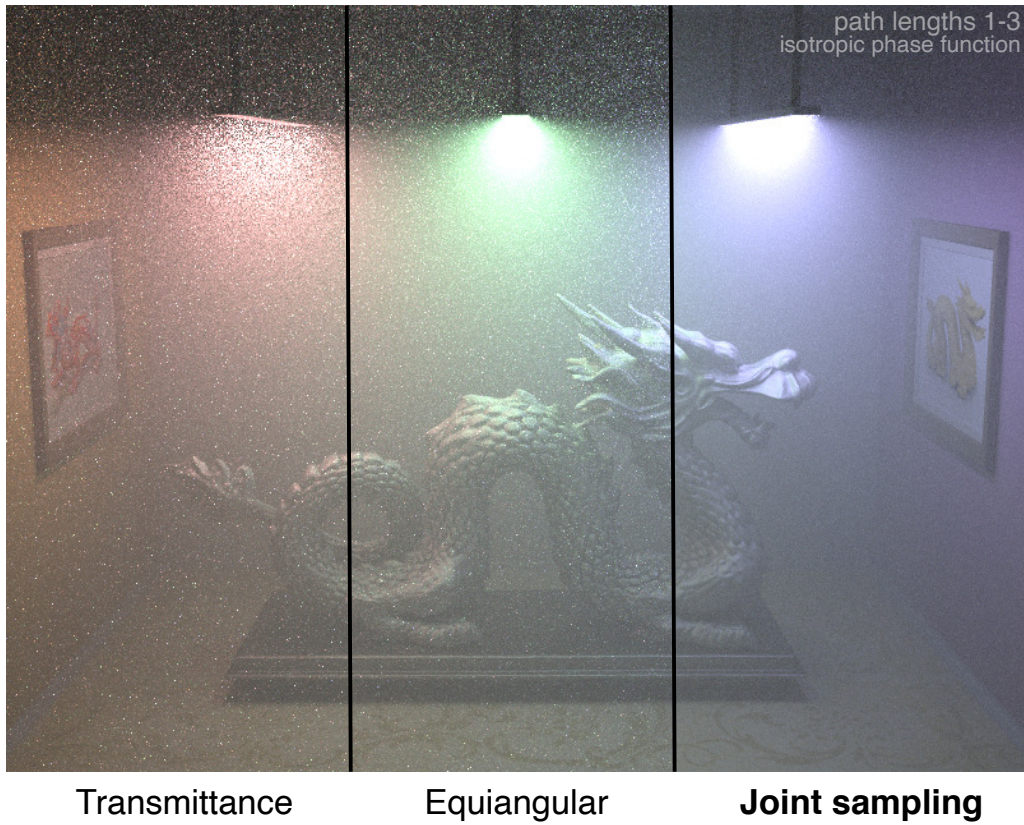
Joint path sampling



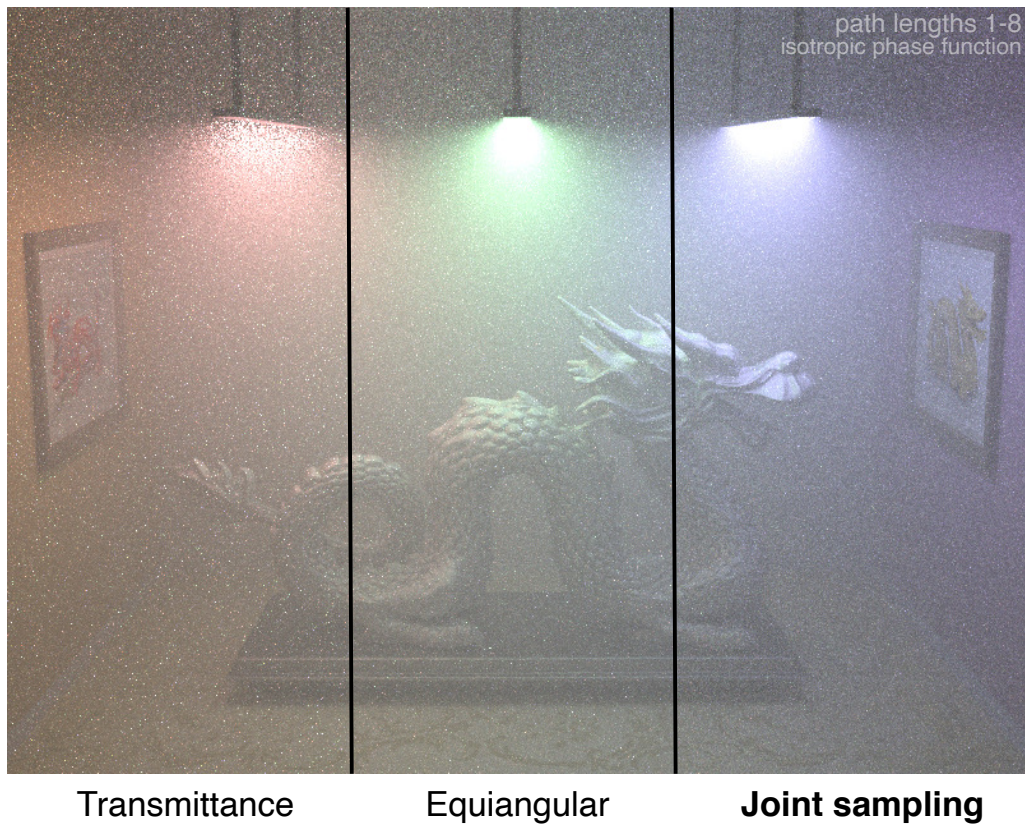
Here is a more specific problem statement. We are given a vertex on a camera subpath (here a point on the lens) and another one on a light subpath (here a point on an emitter). Our goal is to construct a subpath of length 3 (edges) connecting these two vertices, \mathbf{x}_0 and \mathbf{x}_3 , by sampling two new vertices, \mathbf{x}_1 and \mathbf{x}_2 , from a prescribed joint distribution.

We choose this joint distribution to be proportional to the product of the geometry and scattering terms on the connection subpath (the terms shown in blue), as these terms contribute most to the variation in the path contribution. In isotropically scattering media, the phase function is constant and the joint is only proportional to the geometry terms. (Anisotropic scattering can be handled via compact tabulation; we will show results below.)

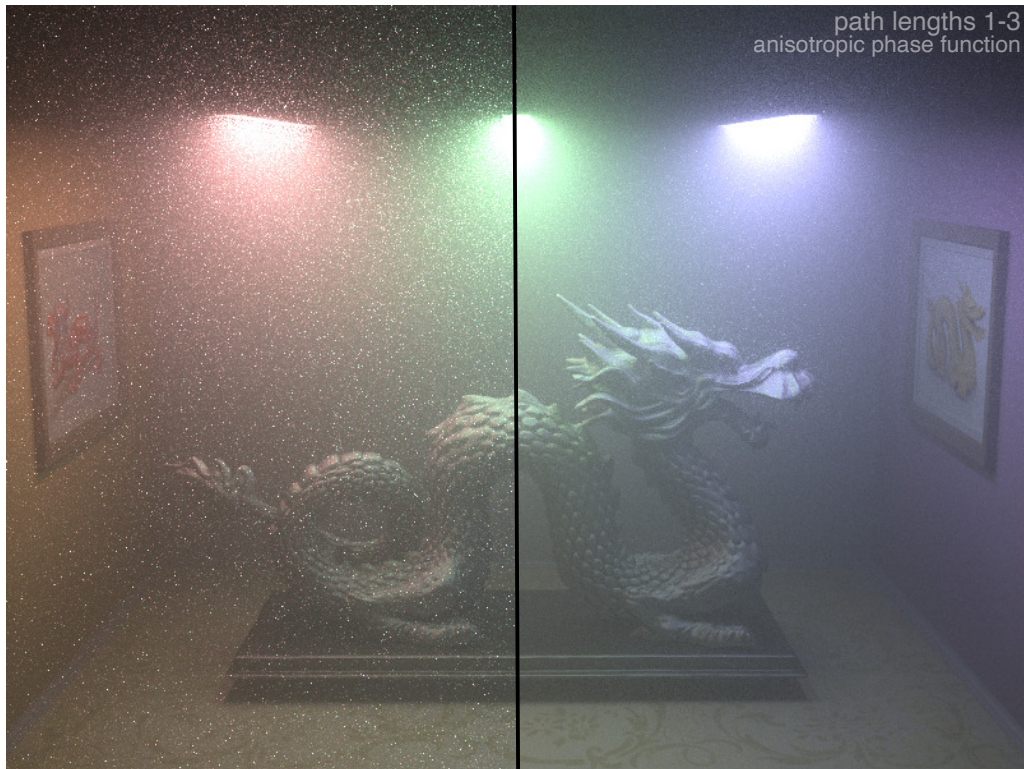
Having prescribed the joint distribution, the corresponding vertex sampling routines can be derived via successive marginalization of the joint. To make this problem tractable, we also assume that we are given a direction from the eye vertex \mathbf{x}_0 . The distance to \mathbf{x}_1 should then be sampled proportionally to the inverse distance between \mathbf{x}_1 and \mathbf{x}_3 (1). It is then particularly crucial to importance sample the subsequent direction ω_2 , due to the singularity at $\theta_2 = 0$. The derived importance sampling routine correctly cancels this singularity (2). The final distance needs to be sampled with density proportional to the inverse squared distance between the resulting vertex \mathbf{x}_2 and \mathbf{x}_3 , which is precisely what the equiangular sampling technique discussed before does.



As evident in these images, importance sampling all the geometry terms jointly cancelling out the singularity, producing 3 orders of magnitude reduction over classical transmittance based sampling and eliminating all fireflies. The sampling technique is analytic and only requires a few lines of code.



The method only constructs short subpath connections, but it can be applied to render higher-order scattering as well, similarly to how we extended classical path tracing with equiangular light connections earlier. That is, the input ray to the subpath connection construction can be on an arbitrary-length eye subpath. The variance reduction remains significant.



Transmittance connections

Joint tabulated path sampling

In the general case, where the phase functions are anisotropic, deriving analytic expressions is difficult. An alternative is to resort to numerical marginalization of the prescribed joint distribution via tabulation. For a given phase function, the tables can be pre-computed once before rendering. The curse of dimensionality can be avoided by exploiting symmetries in the geometric configuration, which significantly reduces the table dimensionality, keeping the full joint tabulation compact and practical. The visual improvement in this case is visually even more striking.



Transmittance connections

Joint tabulated path sampling

Again, even though the technique is designed for joint importance sampling of paths of up to length 3, using it as a general-purpose connection technique delivers significant variance reduction for higher-order scattering as well.

Summary

Importance sampling across light bounces

Substantial improvement in the presence of singularities

High-order scattering remains challenging

Ideally incorporate surface scattering

Traditional Monte Carlo path sampling techniques for participating media are a legacy from surface rendering. Attacking media directly allows taking advantage of the extra dimensionality and devising joint importance sampling of sequences of path vertices. This paves the way to thinking about light transport differently, by considering entire subpaths instead of individual points.

While the presented method can significantly outperform previous approaches, there is still a lot of room for improvement. For instance, since it only importance samples connections of up to length 3, the method is not as efficient for higher-order scattering as it is for single and double scattering.

In addition, the distributions do not take into account surface scattering, making the method sub-optimal even for double scattering paths that include a surface interaction. Incorporating surface scattering can lead to further improvements.

6 Zero-Variance Theory for Efficient Subsurface Scattering

6 Zero-Variance Theory for Efficient Subsurface Scattering

Eugene d’Eon and Jaroslav Krivánek¹

6.1 Introduction

The topic of this chapter is *zero-variance Monte Carlo schemes* and their use for improving the convergence rates of Monte Carlo subsurface scattering (SSS) calculations for image synthesis. We expand upon a previous work by the authors [Křivánek and d’Eon 2014] and include several new result such as

- Two new perfectly-zero-variance half-space escape schemes,
- Zero-variance theory for generalized radiative transfer (GRT) (non-exponential random media),
- An exit-resampling procedure for asymptotic/Dwivedi guiding that better accounts for the importance change near boundaries.

6.1.1 Brute Force Subsurface Scattering

Brute force Monte Carlo subsurface scattering is now commonplace in production rendering software [Chiang et al. 2016; Kulla et al. 2018; Fascione et al. 2018; Christensen et al. 2018; Georgiev et al. 2018]. This approach works by sampling random walks/flights inside a participating medium to connect illuminated surface points to nearby exit points (Figure 1). These random walks are unbiased Monte Carlo estimators of fully general *bidirectional scattering-surface reflectance-distribution functions* (BSSRDFs) [Nicodemus et al. 1977] and so are highly flexible and accurate. However, they can be considerably slower than methods that use approximate BSSRDFs. In this chapter we show how analytic importance functions can be used to guide the sampling of these random walks such that the efficiency of the method is improved without losing accuracy.

The BSSRDF is what gives rise to the characteristic bleeding of light that makes translucent materials like human skin appear soft. High quality predictive image synthesis requires that the BSSRDFs are accurately specified and sampled. However, in contrast to BRDFs that are typically known analytically, in any practical setting the BSSRDF is a high-dimensional and *unknown* function. This is because it follows from the solution to an integral equation for the collision density inside the material and that solution depends on the shape of the boundary. The boundary, and therefore the BSSRDF, might even change over time—the BSSRDF of your nose changes as you wiggle your toes (although not measurably). Even in idealized scenarios where exact solutions are known [Williams 2007; Machida et al. 2010; Liemert and Kienle 2013], they are only known in a semi-analytic form and exhibit no obvious importance sampling scheme for generating outgoing surface positions and directions in a single step. Approximate BSSRDFs can be sampled very efficiently, however, but at the cost of accuracy.

Most efficient SSS algorithms proposed in graphics [Jensen et al. 2001; Borshukov and Lewis 2003; Donner and Jensen 2005; d’Eon et al. 2007; Donner et al. 2008; D’Eon and Irving 2011; Christensen 2015] approximate the BSSRDF with a 2D lateral convolution of the incident light based on solutions of the transport equation in plane geometry and then impose diffusive angular shapes on the outgoing radiance. Later methods have improved upon the angular domain of this approach [Habel et al. 2013; d’Eon 2014; Frisvad et al. 2014; Frederickx and Dutré 2017], but lack the general accuracy and flexibility of the random walk approach in curved geometry. With increasing compute power the trend is more and more in favour of exact BSSRDFs that satisfy the equation of radiative transfer.

¹This chapter contains novel material by both authors that was regrettably not published before Jaroslav’s passing. As such, it is essential that any reference to this work includes attribution to Jaroslav.

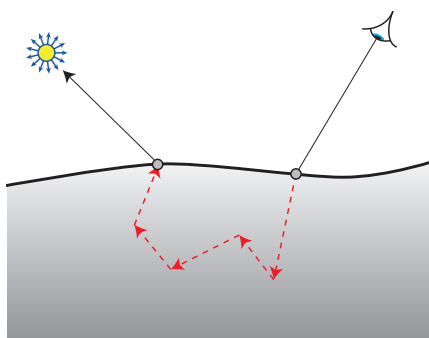


Figure 1: *Random-walk SSS can result in long complex paths (illustrated here as dashed lines) inside the material that transport light beneath the surface from a point of illumination to some nearby location. The standard methods for sampling these paths can result in high variance weights due to longer paths being absorbed more. This chapter discusses guided sampling techniques that reduce this variance, yielding faster convergence, shorter paths on average, and, therefore, shorter render times.*

Despite not knowing the exact BSSRDF itself, we always have a simple Monte Carlo procedure for importance sampling it: the random walk is generated from alternate sampling of the free-path-length distribution (moving the position of the walk) and phase function (changing the walk’s direction) until escape is sampled (with BSDF sampling at the boundary). This procedure follows directly from the integral equation that the collision rate density inside the material satisfies [Lafortune and Willemis 1996; Raab et al. 2008]. This flexible approach works regardless of the shape of the object, how scattering and absorption processes inside vary, or what BSDF is on the boundary.

In the absence of absorption inside the volume, the classical random walk method is already *zero variance*—every path is sampled in exact proportion to the BSSRDF with unit weight (assuming nothing inside or on the boundary absorbs light). For such impossibly-white materials, the content in this chapter has nothing to offer. When absorption is present, however, the weights of this sampling procedure vary with α^n where α is the *single-scattering albedo* at each collision event and n is the number of medium collisions along the path (the number of times the phase function is sampled). The main objective of applying zero-variance schemes to random walk SSS is to remove all variance in the path weight that is due to internal absorption. Because of the plane-parallel nature of the guiding, these methods also apply directly to stochastic methods for sampling layered materials using “position-free” walks [Hanrahan and Krueger 1993; Guo et al. 2018].

6.1.2 Terminology

Much of the zero-variance theory that we apply originates from the neutron transport literature [Kahn 1956; Coveyou et al. 1967; Hoogenboom 2008a]. In this literature, statistically unbiased estimators that converge to the correct answer are referred to as “fair games”, and an estimator “scores” a value (its final particle weight, usually). The term “analog” sampling refers to always *locally* sampling free-path distributions and phase functions directly from their given distributions in isolation, oblivious to where light sources or camera sensors lie in the scene. As such, the simulated particle does the physical analog of a real particle in a physical system [Spanier and Gelbard 1969]. In analog sampling, the particle weight is always 1 (continue to scatter) or 0 (death by absorption, terminating the walk).

In the context of rendering, analog sampling is only implicitly used for materials like glass and mirrors that do not lose any energy. The analog sampling of other BSDFs like a Lambertian reflector would sample outgoing directions and terminate the particle with a probability equal to one minus the diffuse

albedo (equivalent to Russian Roulette that always ensure unit particle weight). Instead, we almost always directly jump to using “Implicit capture”—a form of variance reduction that uses a statistical particle weight to account for absorption. In a participating medium, for example, this works by adjusting the particle weight by a factor of the single-scattering albedo α at every collision. When we refer to “classical sampling”, we mean analog sampling plus implicit capture, which is technique described in graphics text books [Pharr et al. 2016].

In the neutron transport literature, “biased” can refer to importance sampling anything other than the analog distributions. When this literature refers to, for example, “biased direction sampling” in the context of zero-variance theory, they are simply referring to drawing directions from a distribution other than the phase function and adopting the appropriate weight adjustment to ensure a “fair game”. We will instead use “guiding”, to avoid any confusion with “statistical bias”.

We will limit our attention to BSSRDF sampling alone and not to the challenging task of sampling the product of incident illumination with the BSSRDF. This is equivalent to assuming a uniform isotropic source everywhere on the boundary surface and we use the term “guiding-to-escape” for this class of problem. However, the same general theory applies (with higher-dimensional importance functions) to guide SSS random walks when the incident illumination at the boundary is known both in the angular and spatial domains. In this case, it is common to use a two-stage procedure where an approximate importance function is predetermined in the volume in some discrete form either using deterministic or Monte Carlo methods before random walks begin [Turner and Larsen 1997].

We will follow neutron transport and use “collision” to refer to interactions with the medium, which includes both absorbing and scattering collisions.

6.1.3 Outline

Our main goal in this course is to complement the theoretical literature on zero variance schemes by working through several examples that clearly illustrate how the theory is applied in practice. A secondary motivation is to show how the theory can be applied in random media (GRT). After reviewing related work in the next section we define and motivate GRT in Section 6.3. Several key differences between classical and non-exponential (non-Beerian) transport are discussed before defining the general framework of escaping a half space with isotropic scattering in GRT (Section 6.4). In Section 6.5 we derive two new exactly-zero-variance random walks, one for classical scattering in a rod and one for a closely related problem of Gamma-2 random flights in 3D. These examples not only demonstrate that exactly zero variance walks are possible, but also illustrate how such walks differ from classical unguided walks, and how the notion of adjoint importance (exact or approximate) is used to product sample free-path-length and angle sampling decisions to guide a random walk towards a zero-variance version. We review asymptotic (Dwivedi) guiding in Section 6.6 and discuss anisotropic scattering. We finish with some general tips (Section 6.7).

6.2 Related Work

We recommend Hoogenboom [2008a] for a thorough review of the history of zero variance theory including a complete treatment of last-event, collision and track-length estimators. We also recommend Turner and Larsen [1997] for additional details, but prefer the integral equation approach of Hoogenboom, not only because the integro-differential form gets messy, but mostly because of its natural fit for GRT. The related *contributon* theory is also worth noting [Williams 1991].

For a survey of methods that use deterministic importance functions for particle guiding, see [Haghighat and Wagner 2003].

Deep-Penetration Monte Carlo The primary motivation for analytical zero-variance estimators is for shielding calculations in particle transport where the variance reduction for guided vs unguided

walks is many orders of magnitude and the guiding is, on average, towards deeper locations in the material, as opposed to subsurface scattering, where we guide to escape the volume anywhere, although typically back towards the entry location. For a recent survey on variance reduction methods for deep-penetration neutron transport, see [Munk and Slaybaugh 2019].

Condensed History There are several other ways to improve the efficiency of the random walk approach to SSS. Similarity theory and condensed-history schemes can be used to progressively alter the analog sampling distributions as the walk is generated in order to simulate more than one propagation step at a time (for example, making the phase function more isotropic after some number of events, and adjusting the future mean free path to compensate). In doing this, the history of the particle is condensed into fewer individual steps. These methods often introduce small errors, but some aspects of these schemes can exactly maintain desired properties of the uncondensed transport. Condensed history schemes are highly effective in infinite media, but handling boundary crossing/escape without significant error is a major challenge.

A variety of condensed history schemes called *shell-tracing* (in computer graphics [Müller et al. 2016]) begins by finding the largest sphere around a previous collision such that the medium can be considered homogeneous inside that sphere. The particle is then teleported to that sphere’s surface with an appropriate weight adjustment [Fleck and Canfield 1984; Moon et al. 2007]. For some problems this can yield massive gains.

Both condensed history and similarity theory have the most to offer in weakly absorbing materials where thousands of collisions per walk are common, whereas zero-variance guiding-to-escape schemes provide more relative benefit when the material absorbs, making the two approaches complementary. They can be combined using the same steps outlined in this chapter by normalizing the appropriate product involving the importance function. For use of similarity theory in graphics see [Frisvad et al. 2007; Zhao et al. 2014]. For more on condensed history see [Bhan and Spanier 2007; d’Eon 2016].

Guiding and Importance in Graphics The zero-variance Monte Carlo theory is tightly coupled to the theory of adjoint estimators and importance. See Christensen [2003] for an excellent summary of the use of adjoint importance in graphics. We also note several works [Xu et al. 2001; Xu et al. 2006] that applied the zero variance theory explicitly for global illumination in scenes with no participating media.

Machine Learning Several recent works have used machine learning to directly importance sample BSSRDFs on curved domains [Vicini et al. 2019] and to accelerate subsurface transport using learned infinite medium Green’s functions [Deng et al. 2020]. Almost certainly we will see more applications involving machine learning to path guiding in volumes. We hope that some of the deterministic principles that we touch upon in this chapter will inform the design of these methods.

6.3 Generalized Radiative Transfer (GRT)

The transport of waves or particles in a random medium consisting of optically active particles/microstructure is sensitive to exactly how these particles are distributed. When particles in a region with a fixed number density are reconfigured to obey positive (clumpy) or negative (repelling) spatial correlation, this will give rise to different attenuation laws and bulk transport (Figure 2). This phenomena has been long recognized under a variety of names, such as the *sieve/package effect* [Rabinowitch 1951; Kirk 1975], the *channeling effect* [Burrus 1958; Burrus 1960], *distributional error* [Fukshansky 1987], or *large scale inhomogeneities, clumping, mixing-fraction variations, particle-self-shielding* [Randall 1962]. Particle reconfiguration can completely transform the properties of the material from transparent to opaque [Torquato 2016]. It is desirable to formulate transport theory machinery that can efficiently account for these effects in order to simulate the broadest class of materials.

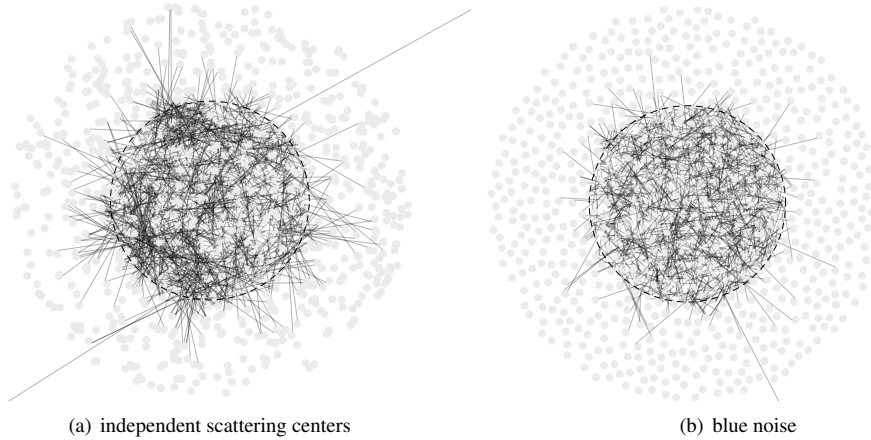


Figure 2: The same density of particles reconfigured from an independent distribution (left) to a configuration with negative (repelling) correlation yields shorter mean free paths. Here we show 2000 paths in each image with origins drawn uniformly from the dashed circle extended to their first collision.

The gold standard approach for including distributional effects in random media is with a stochastic transport equation that permits the material coefficients to become random variables. This approach was first applied in wave transport [Frisch 1968; Ishimaru 1978], and later to the scalar equation of radiative transfer [Anisimov and Fukshansky 1992]. Solving for the mean transport over all permissible random realizations of the system with an averaging step is a rigorous approach but it is challenging to derive exact solutions from this method without making additional assumptions about the magnitude of the correlations. The corresponding rigorous Monte Carlo approach is called *quenched disorder* [Larmier et al. 2017] and works by sampling a number of explicit random realizations for the medium and then performing classical (deterministic or Monte Carlo) transport calculations within each. The desired transport quantities are averaged over the simulated realizations. This approach is also prohibitively expensive and neither of these rigorous approaches is likely to be directly applied in computer graphics. However, both are important benchmarking tools that can be used to evaluate faster approximate methods.

One highly efficient approach to approximating the stochastic transfer equation is to adopt a short term memory and only remember enough of the past to exactly exhibit the free-path-length statistics between collisions [Randall 1964; Hoffman 1964; Audic and Frisch 1993; Moon et al. 2007; Larsen and Vasques 2011]. This is the foundation of what we will refer to as *generalized radiative transfer* (GRT) [d’Eon 2019a; Davis and Xu 2014]. This allows a new aspect of random media that classical transport theory lacks, which is that the distribution of *free-path-lengths between collisions* $p_c(s)$ can be non-exponential (and the attenuation law non-Beerian). This distribution can be measured from Monte Carlo simulation in quenched disorder [Audic and Frisch 1993; Moon et al. 2007; Larsen and Vasques 2011] or from analytical analysis of a given stochastic model for the random extinction coefficient μ_t [Davis and Mineev-Weinstein 2011]. The stochastic process of a particle moving through the system is then a continuous time random walk [Weiss 1983] or, from a time-independent viewpoint, simply a general random flight based on $p_c(s)$ [Dutka 1985].

To apply zero variance theory to GRT we need a transport equation. Two equivalent such equations are known. The integro-differential-like equation of GRT includes a time-like integration over a memory variable s —the distance since the previous medium or boundary interaction [Larsen and Vasques 2011].

This increases the phase space of transport with an extra dimension. This memory is required to exhibit the semi-Markov nature of the particle flight. From a discrete-time point of view (over collision order), the collision chain is fully Markovian, and the collision-rate density satisfies a generalized Peierl's integral equation [Grosjean 1951; d'Eon 2019a]. This is the simpler equation of transfer, closest to the classical form, where all memory is encoded in $p_c(s)$, and from this the zero variance theory is immediately applicable. These integral equations have been used to generalize the volume rendering equation in computer graphics [d'Eon 2013; Jarabo et al. 2018; Bitterli et al. 2018].

To summarize, GRT is a non-exponential random flight where intercollision free path lengths are drawn from $p_c(s)$, and absorption and scattering are non-stochastic (do not depend on s). The attenuation law when leaving a collision is then [Larsen and Vasques 2011]

$$X_c(s) = \int_s^\infty p_c(s') ds'. \quad (1)$$

We require a second set of statistics to apply GRT to bounded domains in a form that satisfies Helmholtz reciprocity. This follows from the need to distinguish between stochastic and deterministic origins in GRT [Audic and Frisch 1993]. Consider the mean chord length between particles in the medium over various realizations. We can only begin such paths from an origin where the last collision ended. Thus, we average over only those realizations with a particle at the origin. This origin is then *correlated* to the other particles in the volume and we use the label "c". In contrast, a deterministic location on a material boundary lies in all realizations of the system. The statistics for free-path length from the boundary must average over the full ensemble (these path-lengths are not chords [Lu and Torquato 1992]). This leads to a related distribution $p_u(s)$ for the *free-path-lengths to next collision from an uncorrelated origin*. The distribution $p_u(s)$ is used for any path leaving a boundary surface or emission from the volume in an uncorrelated manner. Otherwise $p_c(s)$ is used and the two distributions only align for the unique case of exponential random media $p_c(s) = p_u(s) = e^{-s/\ell}/\ell$, where ℓ is the *mean free path*. There is a related attenuation law from uncorrelated origins given by

$$X_u(s) = \int_s^\infty p_u(s') ds'. \quad (2)$$

For an example illustrating why the two distributions differ, see Figure 3.

If any one of $p_c(s)$, $p_u(s)$, $X_c(s)$, $X_u(s)$ are known, the other three are uniquely determined by simple relations [d'Eon 2018]. The distribution $p_u(s)$ is also known as the equilibrium distribution of free path lengths, and $X_c(s)$ and $p_u(s)$ are proportional [Feller 1971; Tunaley 1974; Tunaley 1976; Weiss 1983].

6.3.1 Radiance and Collision Density

An important distinction between two fundamental transport quantities arises in GRT due to the breaking of their classical local proportionality: radiance and collision rate density [d'Eon 2013]. Radiance $L(\mathbf{x}, \omega)$ describes the density of particles *in flight* at position \mathbf{x} in direction ω . This tells us what we would measure if we inserted a tiny camera sensor in the volume and let particles hit that detector. This measurement is of the particles in flight, not the scatterers in the medium. The collision rate density $C(\mathbf{x}, \omega)$ is defined such that $C(\mathbf{x}, \omega) d\omega d\mathbf{x}$ is the rate at which particles are entering collisions within positions $d\mathbf{x}$ about \mathbf{x} and confined to directions in $d\omega$ about ω . Measuring this quantity is to observe the medium itself: the scatterers. Only in classical exponential media do we find the *local* proportionality

$$C(\mathbf{x}, \omega) = \mu_t(\mathbf{x}, \omega) L(\mathbf{x}, \omega). \quad (3)$$

In GRT, the extinction coefficient $\mu_t(s) = p_c(s)/X_c(s)$ is not a locally defined quantity, and so no local conversion is possible. Volumes in GRT are therefore specified with $p_c(s)$, albedo α and phase function P , as opposed to absorption and scattering coefficients.

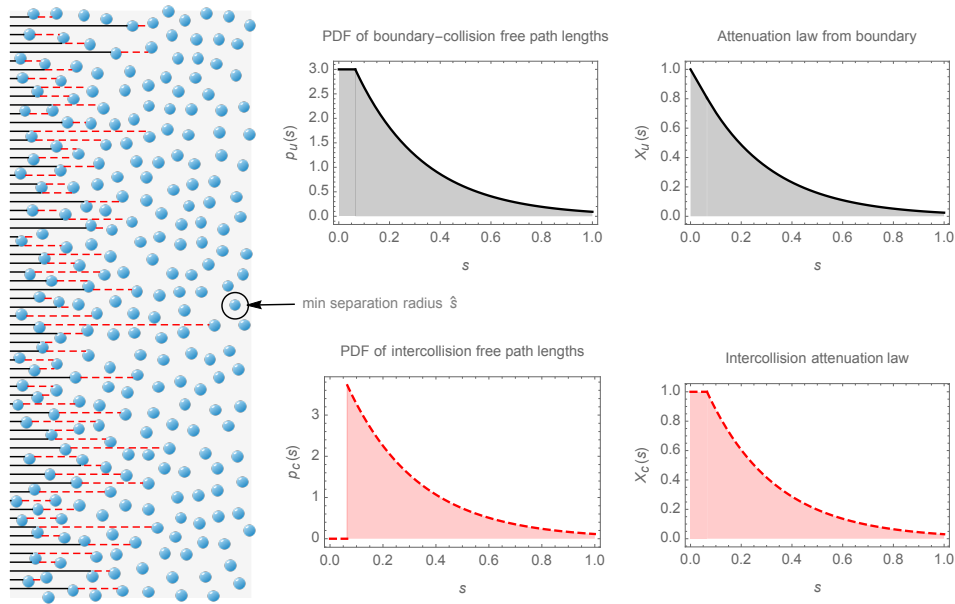


Figure 3: When scatterers in a random medium are spatially correlated, the free-path length statistics between collisions are necessarily distinct from those for paths beginning at a boundary interface. Here we illustrate the case of negatively-correlated convex scatterers separated by a minimum distance $\hat{s} = 0.065$. For paths beginning at the left boundary of a unit thickness slab (solid-black) collisions can occur arbitrarily close to the boundary and the related path length PDF $p_u(s)$ and attenuation law $X_u(s)$ reflect this. Continuing in the same direction from the first collision to the second collision (red-dashed), we find path lengths with a minimum length \hat{s} . The intercollision free-path distribution $p_c(s)$ is therefore identically zero for $s < \hat{s}$ due to the scatterers separation, and the attenuation law between collisions $X_c(s)$ is 1 for this initial distance. Note that $X_c(s)$ and $p_u(s)$ are always proportional.

Because of the new relationship between radiance and collision density in GRT (and their scalar counterparts, fluence and scalar collision density $C(\mathbf{x})$), generalization of classical methods require extra care. Each of these quantities has distinct collision and track-length estimators and diffusion approximations in GRT, where in the classical case there was effectively only one form of these tools [d'Eon 2019a]. This is important to keep in mind with respect to graphics literature where the integral equation inside of volumes is always written over radiance

$$L(\mathbf{x}, \omega) = \int_0^\infty X_c(s) \int_{4\pi} \mu_s P(\omega' \rightarrow \omega) L(\mathbf{x} - s\omega, \omega') d\omega' ds \quad (4)$$

probably for the reason that radiance is the quantity at the camera aperture that forms the final image. However, it is only the collisions in the volume that the camera sees, not all the particles in flight, and so the integral equation for collision density is more directly tied to what we integrate in volumetric path tracing

$$L(\mathbf{x}, \omega) = \int_0^\infty X_c(s) \int_{4\pi} \alpha P(\omega' \rightarrow \omega) C(\mathbf{x} - s\omega, \omega') d\omega' ds \quad (5)$$

For GRT this becomes a critical distinction: the importance functions needed to guide a random walk towards zero-variance satisfy the integral equation for collision density

$$C(\mathbf{x}, \omega) = \int_0^\infty p_c(s) \int_{4\pi} \alpha P(\omega' \rightarrow \omega) C(\mathbf{x} - s\omega, \omega') d\omega' ds. \quad (6)$$

The adjoint incoming radiance field in the scene [Novák et al. 2018] is of little use for path guiding.

6.4 Guiding-to-Escape in a Half Space

We turn now to a half space escape problem that will form the basis for much of the following sections. We assume a homogeneous semi-infinite 3D medium defined by $x > 0$ with a flat indexed-matched boundary and isotropic scattering and absorption in the interior (see Figure 4).

6.4.1 Sources and Detectors

A linear transport problem is defined by specifying a medium/scene, its properties and boundaries, and a set of light sources. We then define a detector sensitivity or measurement functional over the phase space (typically just a camera in rendering). In the general case, we seek a zero variance estimator that has particles leaving the sources and arriving at the detectors such that every simulated particle reaches a detector and reaches it such that the particle weight times the detector sensitivity at that position and direction is a constant. In this general case, the first step of the zero variance derivation is to determine the guided spatial and angular distributions from which to leave the sources [Hoogenboom 2008a]. In the case of BSSRDF sampling, however, our source is always a single element of phase space: an incident position and direction, which we always sample with weight $w = 1$. Since we assume a flat homogeneous geometry, it suffices to only know the incident cosine, and so we will derive 1D families of estimators over μ_i . Our detector sensitivity is defined as 1 for all positions and directions that escape the medium.

6.4.2 The Classical Estimator

We are given as a starting point that a particle arrives at the boundary entering the medium along a direction with a cosine to the inward normal of μ_i . The classical estimator proceeds with (see Figure 4)

1. Particle weight $w = 1$
2. Sample initial displacement s_1 from $p_u(s)$ and move particle
3. Absorb $w \rightarrow w * \alpha$

4. Sample direction ω from phase function P
5. Sample intercollision displacement s from $p_c(s)$ and move particle
6. if $x < 0$ return/score w at the exitant boundary position and direction and terminate the walk
7. goto 3.

6.4.3 The Guided Estimator

The key shortcoming of the classical estimator is that the sampling of $p_u(s)$, P and $p_c(s)$ are locally greedy—they are perfect estimators of these normalized distributions, but are ignorant of the end goal, like playing chess while only thinking one move ahead. We will derive the zero variance estimator for escaping the half space by guiding each of these three sampling decisions. The distributions that are required to achieve zero variance depend on the position and direction of the particle right before these sampling steps are performed and are uniquely determined from a value or importance function W that satisfies an adjoint integral transport equation for collision rate density inside the volume [Hoogenboom 2008a]. In the final step we will also adjust the escape scoring to use an expected value estimator.

Initial Free Flight: Our first step is to sample the initial free-flight distance s_1 from a guided distribution $p_1(s)$ that achieves the zero-variance goal. In sampling s_1 from $p_1(s)$ instead of $p_u(s)$, the particle must adopt a weight factor of $w_1(s_1) = p_u(s_1)/p_1(s_1)$. After traversing free-flight distance s_1 the particle enters a collision at depth $x_1 = \mu_i s_1$ with weight $w_1(s_1)$. Let $W(x, \mu)$ be the probability that a particle entering a collision at depth x along direction with cosine μ eventually escapes the medium. The expected contribution of our particle after initial displacement is therefore its current state times the expected total future state, $w_1(s_1)W(x_1, \mu_i)$. For the random walk to be zero variance this result must be a constant, and that constant must be equal to the diffuse albedo of the medium for incoming direction μ_i

$$w_1(s_1)W(x_1, \mu_i) = \frac{p_u(s_1)}{p_1(s_1)}W(\mu_i s_1, \mu_i) = R(\mu_i). \quad (7)$$

From this we see that $p_1(x)$ must be

$$p_1(s) = \frac{p_u(s)W(s\mu_i, \mu_i)}{R(\mu_i)}. \quad (8)$$

We see that the guided distribution is the product of the analog distribution and the importance function with a normalization factor. We don't need to know $R(\mu_i)$, because we can find it by simply requiring that $p_1(s)$ integrates to 1. For no absorption, we see $R = 1$, $W(x) = 1$ and analog sampling $p_1(s) = p_u(s)$, as desired.

The Full Guided Estimator: The previous example illustrates the key components of the general procedure for deriving each step of a random walk in order to achieve zero variance:

- The end state of a guided step will be the resulting particle weight and the particle position and direction
- The resulting particle weight will be the prior weight times the ratio of the analog and guided distributions at the sampled distance/direction
- There is a unique probability to escape the medium at the sampled particle position and direction (importance W). Care must be taken here to distinguish between entering and leaving collisions.
- The guided distribution must be the normalization of W times the analog distribution.

We will see the details of the remaining steps in the general procedure during the following examples.

To summarize the notation (also summarized in Table 1) of the upcoming guided distributions: absorption is handled identically to the classical random walk, applying implicit capture per collision with

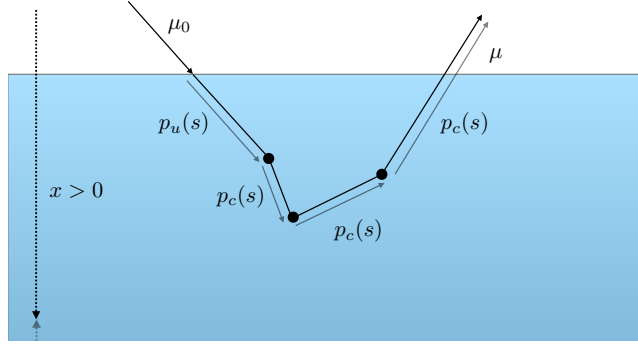


Figure 4: For guided BSSRDF sampling we consider the illustrated GRT random walk in a 3D half space. A particle arriving along a direction with cosine μ_0 enters the medium and collides after a distance drawn from $p_u(s)$. Absorption with probability $1 - \alpha$ occurs at each collision event. If not absorbed, the particle continues after phase function sampling until exit is sampled. Each zero-variance derivation assumes some importance function for escape $W(x)$ when entering collision at depth x and from this follows guided distributions $p_1(s)$ for the initial free-path lengths and related distributions for direction and intercollision length sampling. These distributions and their related weight adjustments are summarized in Table 1.

weight factor α . Phase function sampling requires polar μ and azimuthal ϕ angle decisions drawn from guided distributions $P^g(\mu; x)$ and a uniform azimuth distribution $1/2\pi$, respectively. The guided azimuthal sampling is identical to the analog case because of the plane symmetry of the medium and detector sensitivity, so the weight factor for azimuthal guiding is $w_\phi = (1/2\pi)/(1/2\pi) = 1$. For polar angle sampling, $w_\mu = (1/2)/P^g(\mu; x)$ accounts for guiding away from the uniform $(1/2)$ distribution of isotropic scattering. Weight factor $w_s = p_c(s)/p_c^g(s; x, \mu)$ accounts for the guided intercollision free-path length sampling from $p_c^g(s; x, \mu)$.

6.5 Two Exactly-Zero-Variance Walks

Achieving a perfectly zero-variance walk for a given problem is almost always more challenging than estimating the desired quantity, because the importance function is required everywhere in the scene. However, it can still be useful to apply the theory using an approximate importance $W(x)$, to reduce the absorption variance. This can improve upon classical sampling even if the geometry is curved, if the medium coefficients vary with position or if there is other geometry imbedded in the medium. Since there are several examples with isotropic scattering in half spaces where the exactly zero variance estimator is possible, we will review those in order to best demonstrate how zero variance walks are derived and how they differ from the classical estimators.

The first example we consider is for the reflection from a classical half rod with isotropic scattering: a simplified one-dimensional domain where particles can only move in one of two discrete directions, left ($-$) and right ($+$). Because the collision rate density in the half rod is a simple exponential, the guiding importance sampling decisions can be handled analytically and we avoid the complexity of the singular eigenfunctions of the related problem in a 3D half space.

| Sampling Decision | Analog | Guided | Weight Factor |
|--|------------|--------------------|--|
| Initial free-path length s_1 from the boundary | $p_u(s)$ | $p_1(s)$ | $w_1 = \frac{p_u(s_1)}{p_1(s_1)}$ |
| Intercollision free-path length s | $p_c(s)$ | $p_c^g(s; x, \mu)$ | $w_s = \frac{p_c(s)}{p_c^g(s; x, \mu)}$ |
| Direction cosine μ | $(1/2)$ | $P^g(\mu; x)$ | $w_\mu = \frac{(1/2)}{P^g(\mu; x)}$ |
| Direction azimuth ϕ | $(1/2\pi)$ | $(1/2\pi)$ | $w_\phi = \frac{(1/2\pi)}{(1/2\pi)} = 1$ |

Table 1: Summary of our notation for the analog and guided distributions for planar guiding to escape in a homogeneous GRT volume.



Figure 5: The Albedo problem for the half rod.

The second example we consider is a new derivation for GRT in a 3D half space where free-path lengths between collision are drawn from a Gamma/Erlang-2 distribution. This zero variance estimator shows how the zero-variance theory extends to easily handle GRT. We will also see that a projection of this random walk onto the depth axis is equivalent to our first example in the rod.

6.5.1 The Zero-Variance Walk in the Half Rod: Křivánek's Walk

We now consider the problem of external illumination reflecting from a one-dimensional absorbing and scattering half space with isotropic scattering and vacuum boundary conditions (Figure 5). We consider specifically the *rod model*²—a simplified one-dimensional domain in which particles can only flow right or left ([Wing 1962; Hoogenboom 2008b]). This problem corresponds to the classical albedo problem of linear transport theory [Chandrasekhar 1960], but in a 1D universe—the unique dimensionality for which the full solution both at the boundary and internally is known exactly in terms of simple explicit expressions [d'Eon and McCormick 2019].

While 1D rod transport has limited direct physical application [Zoia et al. 2011], study of this problem provides all of the essential ingredients for building a zero-variance half space walk, without the distraction of complex importance functions. The rod has been used several times to demonstrate zero-variance walks [Hoogenboom 1981; Hoogenboom 2008b]. However, to our knowledge, the zero variance walk we derive in this section is new³.

Let us define the half rod to occupy the positive axis $x > 0$ with direction $\omega = 1$ corresponding to flight deeper into the rod and $\omega = -1$ towards the boundary. The phase space for monoenergetic particles/photons is then $\mathbb{R} \times \{-1, 1\}$. Scattering is isotropic, where each collision draws a new direction ω from $\{-1, 1\}$ with equal probability, and the single-scattering albedo is α . This example assumes classical media with exponential free-path length distributions and attenuation laws $p_c(s) = p_u(s) = X_c(s) = e^{-s}$.

Our random walk begins entering the rod at the boundary $x = 0, \omega = 1$ and proceeds with an initial free-flight transition followed by a chain of collision and free-flight steps until the particle is either absorbed or escapes. The analog walk chooses between collision and absorption with a discrete binary decision and clearly leads to unresolvable variance, so the first step in guided sampling is to use implicit

²Also known as the two-directional or Fermi model

³This result was communicated to the first author by the second author on Nov 24, 2013 and has been named to reflect its origin.

capture, as is standard in volumetric light transport. This is accounted for by a particle weight w that begins the walk at 1 and is multiplied by the single-scattering albedo for every collision inside the rod.

Next, following Hoogenboom [2008a], we extend the rod to the full line, letting the exterior portion $x < 0$ be purely absorbing. This is a mathematical convenience that informs derivation of the importance function for the entire system that is used to guide the random walk. In this extended interpretation of the problem, any collision in $x < 0$ scores the current particle weight and terminates the walk. Any absorption inside the rod scores 0 and continues. This imparts a *last event collision estimator* interpretation on escaping the medium.

We now define an importance (or value) function $W(x, \omega)$ for the rod defined as follows: $W(x, \omega)$ is the probability that a particle *entering* a collision at position x moving in direction ω (before the collision) eventually escapes the rod. From the assumption of isotropic scattering we see immediately that the desired importance function is independent of direction ω . This is a hallmark of deriving zero variance walks for problems with isotropic scattering: the dimensionality of the importance function is greatly reduced.

We can find $W(x)$ from known solutions for the collision rate density inside a half rod due to external illumination. The two are directly related, by reciprocity. The specific solution follows from solving a Wiener-Hopf integral equation with the Picard/Lalesco kernel [Wing 1962; d'Eon and McCormick 2019] (more on this later). The result is

$$W(x) = \begin{cases} (1 - \sqrt{1 - \alpha}) e^{-\sqrt{1 - \alpha}x}, & x \geq 0 \\ 1 & x < 0 \end{cases} \quad (9)$$

where we have set the value to 1 for any position outside of the volume.

We note several important features of this result. For the conservative medium $\alpha = 1$, $W(x) = 1$ everywhere because entering a collision anywhere eventually leads to escape, which shows that the classical random walk estimation of the albedo R is already zero variance. Given W , we immediately have the final weight of our zero-variance walk, the escape probability, given by

$$R = \int_0^\infty p_u(s)W(s)ds = \int_0^\infty e^{-s}W(s)ds = \frac{2}{\alpha} (1 - \sqrt{1 - \alpha}) - 1. \quad (10)$$

Initial Free Flight: Following the arguments from the previous section we see that $p_1(x)$ must be

$$p_1(x) = \frac{p_u(x)W(x)}{R}. \quad (11)$$

We find that $p_1(x)$ simplifies to a simple exponential

$$p_1(x) = (\sqrt{1 - \alpha} + 1) e^{-(\sqrt{1 - \alpha} + 1)x}, \quad (12)$$

which we can easily importance sample by CDF inversion, giving

$$x_1 = -\frac{\log(1 - \xi)}{1 + \sqrt{1 - \alpha}} \quad (13)$$

where $\xi \in [0, 1)$ is a uniform random variate.

Direction Sampling: For each collision at depth $x > 0$ we need to sample an outgoing scattering direction ω such that the future contributions from subsequent collision and escape are perfectly balanced.

Let the guided distribution $p^+(x)$ be the probability that the positive direction is sampled after collision at depth x , and $p^-(x) = 1 - p^+(x)$ the probability of scattering towards the boundary. This is a discrete variant of $P^g(\mu; x)$ described in the previous section. For every collision, the particle enters with weight $w = R/W(x)$. Immediately following the collision the weight is adjusted by implicit capture to $w' = \alpha R/W(x)$. If the particle scatters positive, we have a further weight adjustment of $w_\omega = (1/2)/p^+(x)$ due to guiding away from the analog choice of equal probabilities for both directions. The expected score of the particle having gone right is then the total weight after scattering, $w'w_\omega$, multiplied by the expected final score over all possible free-flight distances s ,

$$w'w_\omega \int_0^\infty p_c(s)W(x+s)ds = R \quad (14)$$

Solving this equation for $p^+(x)$ we find simply

$$p^+(x) = \frac{1}{2}(1 - \sqrt{1 - \alpha}). \quad (15)$$

Remarkably, this result is invariant to depth—no matter where we collide in the rod, we need to sample away from the boundary with the same probability that depends only on the absorption level in the rod. As absorption increases and α decreases, we sample towards the boundary with increasing probability—paths are guided towards the exit. When there is no absorption ($\alpha = 1$) we recover the analog phase function sampling $p^+(x) = (1/2)$, as desired.

Direction ω is easily sampled from $\{p^+(x), p^-(x)\}$ using a single random number for the discrete choice. The weight factor due to this importance sampling simplifies to

$$w_\omega = \begin{cases} \frac{1}{1+\sqrt{1-\alpha}}, & \omega = -1 \\ \frac{1}{1-\sqrt{1-\alpha}}, & \omega = 1 \end{cases}. \quad (16)$$

General Free-Path Sampling: The final step in building the zero-variance walk for the rod is to determine the guided intercollision free-path length distribution $p_c^g(s; x, \omega)$ and to handle the case where the particle exits the volume. Here, $p_c^g(s; x, \omega)ds$ is the probability that we sample a guided distance-to-collision s falling in $[s, s + ds]$ when leaving a collision at x in direction ω .

In the case of moving in the positive direction, $\omega = 1$, we need to sample a intercollision distance s^+ from a distribution proportional to $p_c(s)W(x+s)$. This results in the same exponential distribution we saw above for the initial collision depth x_1 and so we have

$$p_c^g(s; x, 1) = p_1(s) \quad (17)$$

with sampling procedure given in Eq.(13). For free-flight distances s in the negative direction we again need to sample from the normalized distribution that is proportional to the product of $p_c(s)$ and the importance function $p_c(s)W(x-s)$. We find the normalization constant to be

$$\int_0^\infty p_c(s)W(x-s)ds = e^{\sqrt{1-\alpha}(-x)}, \quad (18)$$

resulting in

$$p_c^g(s; x, -1) = (1 - \sqrt{1 - \alpha}) \left(e^{-s(1-\sqrt{1-\alpha})} \right), \quad 0 < s < x. \quad (19)$$

Like the positive direction case, we again find a distribution that is translationally invariant. The shape of the PDF beyond the boundary $s > x$ is not important—we only need to observe that this distribution up

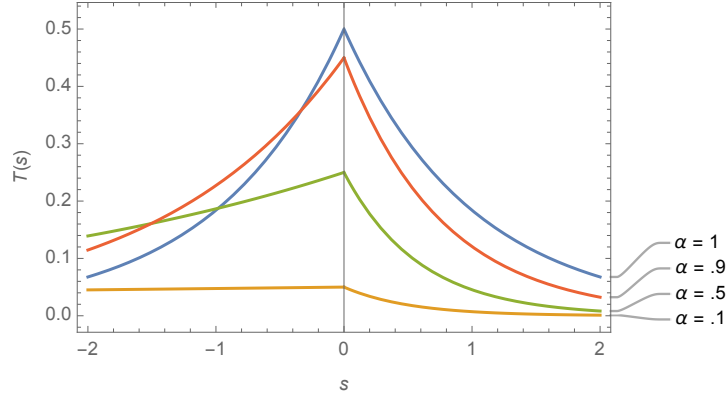


Figure 6: Guided transition kernels $T(s)$ (combining phase function and free-flight sampling) for the zero-variance walk in the half rod. With decreasing single-scattering albedo α negative displacements towards the boundary (escape) are increasingly preferred.

to the boundary is an exponential with a mean free path of $1/(1 - \sqrt{1 - \alpha})$ and sample that distribution. Any time a distance past the boundary is sampled, we apply a mean-value weight factor w_{esc} , which is the ratio of the analog probability for escape to the probably of escaping with the guided distribution

$$w_{\text{esc}} = \frac{X_c(x)}{e^{-x(1 - \sqrt{1 - \alpha})}}. \quad (20)$$

Finally, if we sample an interior collision $s < x$, we apply the weight factor for the guided free-path length

$$w \rightarrow w * \frac{p_c(s)}{p_c^g(s; x, -1)}. \quad (21)$$

This completes the derivation of the zero-variance walk. We include a Mathematica implementation of it in the supplemental material.

It is informative to look at combined transition kernel $T(s)$ that combines direction and displacement sampling together using a signed free-flight distance s where the sign indicates whether or not the depth of the next collision is closer to the boundary and farther into the rod. We find

$$T(s) = \begin{cases} \frac{\alpha}{2} e^{s(1 - \sqrt{1 - \alpha})}, & s < 0 \\ \frac{\alpha}{2} e^{-s(1 + \sqrt{1 - \alpha})}, & s > 0 \end{cases}. \quad (22)$$

These guided displacement kernels are plotted in Figure 6 for various absorption levels and show how increased absorption leads to increased preference for negative (towards the boundary) displacements in order to get the particle out before it is overly absorbed. Figure 7 shows the relative change in particle weight after a net positive or negative displacement in the rod with the zero-variance scheme. It is interesting that this shows no discontinuity at 0 displacement.

6.5.2 The Zero-Variance Walk in the Gamma-2 Half Space

In this section we derive the first perfectly-zero-variance walk for escaping an absorbing half space in 3D. To our knowledge, this is also the first zero-variance walk of any form derived for GRT.

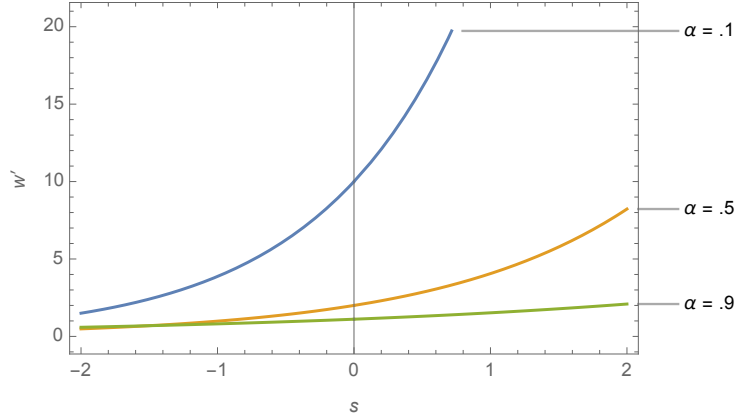


Figure 7: Relative change in particle weight w' in transitioning a relative distance s in the rod.

Specifically, we consider the 3D half space $x > 0$ with Gamma-2-distributed free-path lengths between collisions. Upon specifying $p_c(s)$, the following full set of GRT statistics follow [d'Eon 2018]

$$p_c(s) = e^{-s}s, \quad (23)$$

$$X_c(s) = e^{-s}(1 + s), \quad (24)$$

$$p_u(s) = \frac{e^{-s}(1 + s)}{2}. \quad (25)$$

Importance: As in any zero or near-zero variance random walk derivation, we begin with the importance function, which in the present case is the probability to eventually escape the medium (after any number of subsequent collisions) upon entering a collision at depth x . Because the scattering is isotropic, the importance function $W(x)$ is independent of the incoming direction of the particle.

We can derive or estimate $W(x)$ in a number of ways. We could tabulate a discrete numerical approximation of $W(x)$ for a given absorption level by taking the mean escape probability of some number of unguided random walks, each beginning in some narrow interval of depths $x_0 \in [x, x + dx]$. Alternatively, by reciprocity, we could sample a suitably weighted uniform surface source and tally collision densities in narrow depth intervals within the medium. We have chosen a problem which admits an exact and very simple importance function in order to clearly illustrate the subsequent steps in determining the full guided walk. However, all of the following principles apply to any approximate tabulated or fitted function $W(x)$.

We now derive the exact escape probability for our problem from the Wiener-Hopf integral equation that applies to the collision rate density inside the volume. The details of this derivation are not essential to the guiding sampling that follows, but we include these details for completeness. The Wiener-Hopf integral equation for the collision rate density $C(x)$ with a unit Dirac delta of initial collisions at depth x_0 is

$$C(x) = \delta(x - x_0) + \alpha \int_0^\infty C(x')K_C(x - x')dx'. \quad (26)$$

The displacement kernel K_C for Gamma-2 flights in 3D with isotropic scattering follows from [d'Eon

and McCormick 2019; d'Eon 2019b]

$$K_C(x) = \frac{1}{2} \int_0^1 p_c(|x|/\mu) \frac{1}{\mu} d\mu = \frac{1}{2} e^{-|x|}, \quad (27)$$

which is the Picard/Lalesco kernel [Picard 1911]. From the Fourier transform of the kernel

$$\tilde{K}_C(t) \equiv \int_{-\infty}^{\infty} K_C(x) e^{ixt} dx = \frac{1}{1+t^2} \quad (28)$$

we immediately have the Green's function (the solution to Eq.(26)) in terms of the Chandrasekhar H function for the problem. In general, H is given uniquely by [Ivanov 1994]

$$H(z) = \exp\left(\frac{z}{\pi} \int_0^{\infty} \frac{1}{1+z^2 t^2} \log\left[\frac{1}{1-\alpha \tilde{K}_C(t)}\right] dt\right), \quad \text{Re } z > 0. \quad (29)$$

For the Picard kernel we find [d'Eon and McCormick 2019]

$$H(\mu) = \frac{(1+\mu)}{(1+\mu/\nu_0)} \quad (30)$$

where ν_0 is the discrete eigenvalue of the transport operator, the unique positive solution of the dispersion equation,

$$1 - \alpha \tilde{K}_C(i/\nu_0) = 0, \quad \nu_0 = \frac{1}{\sqrt{1-\alpha}}. \quad (31)$$

If we define the Laplace transform

$$\mathcal{L}_x[f(x)](s) \equiv \int_0^{\infty} f(x) e^{-sx} dx, \quad (32)$$

then we have, from Ivanov ([1994], Eqs. (19) and (21)), that the double Laplace transform of the Green's function is

$$\bar{\mathbb{G}}(s, s_0) = \mathcal{L}_x[\mathcal{L}_{x_0}[\mathbb{G}(x, x_0)]](s, s_0) = \frac{H(1/s)H(1/s_0)}{s+s_0}. \quad (33)$$

Inverting both Laplace transforms gives the Green's function $\mathbb{G}(x, x_0)$, which is the rate density of collisions in the system at x due to the initial collision at depth x_0 . However, we only need to invert one of the Laplace transforms, because we want the total rate of collisions inside the entire half space, which is conveniently given when $s = 0$ in Eq.(33). To find the total collision rate $\langle C(x_0) \rangle$, we therefore take the inverse Laplace transform of $\bar{\mathbb{G}}(0, s_0)$ with respect to s_0 ,

$$\begin{aligned} \langle C(x_0) \rangle &= \mathcal{L}_{s_0}^{-1} \left[\frac{H(\infty)H(1/s_0)}{s_0} \right] (x_0) = \mathcal{L}_{s_0}^{-1} \left[\frac{(1+s_0)\nu_0^2}{s_0(s_0\nu_0+1)} \right] (x_0) \\ &= \nu_0 \left(\nu_0 - (\nu_0 - 1)e^{-\frac{x_0}{\nu_0}} \right) \end{aligned} \quad (34)$$

where here we have used $H(\infty) = 1/\sqrt{1-\alpha}$ [Ivanov 1994]. The mean absorption per collision is $1-\alpha$, and there are a mean number of collisions given by $\langle C(x_0) \rangle$, and so the mean energy not absorbed in the system is (and by normalization, the escape probability) is $1 - (1-\alpha)\langle C(x_0) \rangle$, giving our importance function for the problem,

$$W(x) = \begin{cases} \frac{(\nu_0-1)e^{-\frac{x}{\nu_0}}}{\nu_0}, & x \geq 0 \\ 1 & x < 0 \end{cases} \quad (35)$$

Eq.(35) is, in fact, the exact same importance function for the exponential half rod example above (Eq.(9)).

The last quantity we need for deriving the zero variance walk is the expected value of our estimator for a single particle arriving at the boundary along cosine $0 < \mu_i \leq 1$ to the x axis. The known albedo for the problem is [d'Eon 2019b]

$$R(\alpha, \mu_i) = \int_0^\infty p_u(s)W(s\mu_i)ds = \frac{\alpha(\sqrt{1-\alpha}\mu_i + 2)}{2(\sqrt{1-\alpha} + 1)(\sqrt{1-\alpha}\mu_i + 1)^2}. \quad (36)$$

Initial Free-Flight Distance: Guided sampling of the initial free-flight distance s_1 is found from normalizing the product of the uncorrelated-origin FPD and the importance function at depth $\mu_i s$ yielding

$$p_1(s, \mu_i) = \frac{p_u(s)W(\mu_i s)}{R(\alpha, \mu_i)} = e^{-s(\frac{\mu_i}{\nu_0} + 1)}(s + 1)\frac{(\frac{\mu_i}{\nu_0} + 1)^2}{\frac{\mu_i}{\nu_0} + 2} \quad (37)$$

Using three independent uniform random variates ξ_1, ξ_2, ξ_3 , we can sample this as a sum of an exponential and an Erlang-2 distribution,

$$s_1 = \begin{cases} -m(\mu_i) \log(\xi_2), & \xi_1 < \frac{1}{1+m(\mu_i)} \\ -m(\mu_i) \log(\xi_2 \xi_3), & \text{else} \end{cases} \quad (38)$$

where

$$m(\mu) = \frac{1}{1 + \frac{\mu}{\nu_0}} \quad (39)$$

is a path-length stretching factor.

Guided Direction Sampling: Let us define the new angular importance function

$$W_o(x, \mu) = \int_0^\infty W(x + \mu s)p_c(s)ds \quad (40)$$

for *leaving* a collision. This function takes the analog probability $p_c(s)ds$ that the next collision is within ds of s away from the starting position, and multiplies by the probability $W(x + \mu s)$ of escaping after collision there. Integration over all possible s then gives the mean probability of eventually escaping the medium when leaving a collision at depth x in direction μ . Zero-variance direction sampling then results from drawing outgoing direction cosines μ from the normalization of $P(\mu)W_o(x, \mu)$. This is the same general form we saw when deriving the initial path length but note here the different importance function W_o . It is essential that each step in the zero variance derivation carefully consider the escape probability immediately following the action that is being sampled, and to distinguish between pre/post absorption and collision, or for hitting or leaving a Fresnel boundary, etc.

The analog direction cosine phase function is isotropic $P(\mu) = (1/2)$. We seek a guided direction distribution $P^g(\mu; x) = aP(\mu)W_o(x, \mu)$ where constant a is chosen to achieve normalization

$$\int_{-1}^1 P^g(\mu; x)d\mu. \quad (41)$$

After some calculations in Mathematica, we find

$$P^g(\mu; x) = \begin{cases} \frac{\nu^2 - 1}{2(\mu + \nu_0)^2}, & \mu > 0 \\ \frac{(\nu_0 + 1) \left(e^{z(\frac{1}{\mu} + \frac{1}{\nu_0})} (\mu(\mu^2 + 2\mu\nu_0 + \nu_0) - (\mu + 1)z(\mu + \nu_0)) + \mu(\nu_0 - 1)\nu_0 \right)}{2\mu\nu_0(\mu + \nu_0)^2}, & \mu < 0. \end{cases}$$

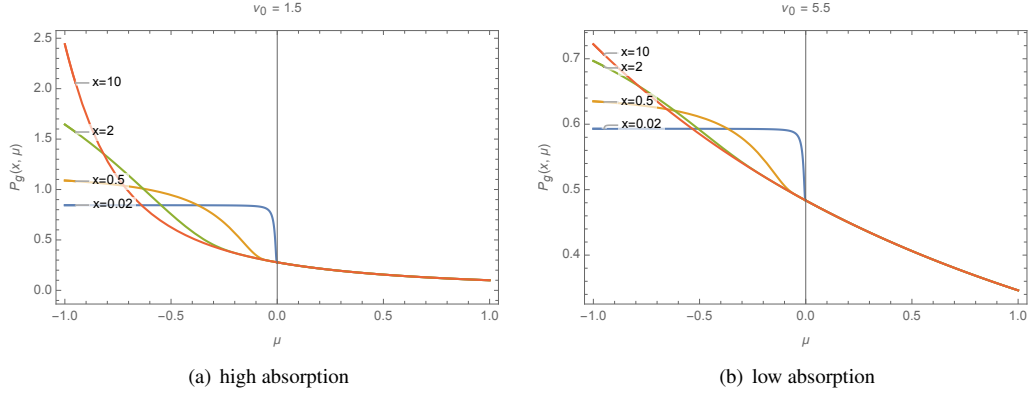


Figure 8: The zero-variance walk in 3D with Gamma-2 flights samples upwelling $\mu < 0$ collisions more often than downwelling ones. Nearer the boundary the upwelling distribution flattens into a uniform distribution because all directions lead to escape with negligible attenuation. The downwelling direction sampling is independent of depth x .

Remarkably, the angle selection in the downward hemisphere (away from the boundary $\mu > 0$) does not depend on the depth x of the particle. This is because the importance function is a pure exponential. Gamma-2 random flights are the unique distribution $p_c(s)$ that produce this result in 3D under isotropic scattering.

To sample this distribution over outgoing cosine $\mu \in [-1, 1]$ we split the sampling into the downwelling (+) and upwelling (-) hemispheres. Because the downwelling direction sampling is independent of depth, the total probability of choosing a downwelling direction must too be depth-independent and, indeed, we find

$$p^+ \equiv \int_0^1 P^g(\mu; x) d\mu = \frac{1}{2} (1 - \sqrt{1 - \alpha}). \quad (42)$$

Choosing a downwelling direction with probability p^+ we need to sample a direction cosine μ from

$$\frac{P^g(\mu; x)}{p^+} = \frac{\nu_0(\nu_0 + 1)}{(\mu + \nu_0)^2}, \quad 0 < \mu < 1. \quad (43)$$

From CDF inversion we find a downwelling cosine μ^+ is sampled using

$$\mu^+ = \frac{\nu_0 + \xi}{1 + \nu_0 + \xi} \quad (44)$$

where $0 < \xi < 1$ is a uniform random variate.

Sampling upwelling direction cosines is more challenging. We need to sample from

$$\frac{P^g(\mu; x)}{1 - p^+} = \frac{\mu(\nu_0 - 1)\nu_0 + e^{x(\frac{1}{\mu} + \frac{1}{\nu_0})} (\mu(\mu^2 + 2\mu\nu_0 + \nu_0) - (\mu + 1)x(\mu + \nu_0))}{\mu(\mu + \nu_0)^2}$$

with CDF

$$\int_{-1}^k \frac{P^g(\mu; x)}{1 - p^+} d\mu = \frac{(k + 1) \left(k e^{x(\frac{1}{k} + \frac{1}{\nu_0})} + \nu_0 \right)}{k + \nu_0}, \quad -1 < k < 0. \quad (45)$$

We did not find an exact sampling procedure for this distribution but found 3 iterations of Newton's method started at $\mu = -0.5$ very accurate for the limited testing we undertook.

General Free-Flight Sampling: For downwelling directions we find a simple guided free-path length distribution by normalizing $p_c(s)W(x + s\mu)$, similar to the initial free-path length procedure above, but with $p_c(s)$ instead of $p_u(s)$ because the particle is leaving a collision and not a deterministic location on the boundary. We find,

$$p_c^g(s; x, \mu) = \frac{se^{-\frac{s}{m(\mu)}}}{m(\mu)^2}, \quad 0 < \mu < 1, \quad (46)$$

which is a stretched Gamma-2 distribution with factor m given in Eq.(39) that is easily importance sampled via

$$s^+ = -m(\mu) \log(\xi_1 \xi_2). \quad (47)$$

Note how similar this is to Asymptotic/Dwivedi guiding in the classical 3D half space. This is a direct generalization of the exponential transform that was the original guiding tool of choice in neutron transport literature [Dwivedi 1982]. Here, we find an analogous stretching of the intercollision free-path distribution, the Gamma-2 transform, appearing in the exactly-zero-variance walk.

For the upwelling directions, we again find the guided free-path length distribution by normalizing $p_c(s)W(x + s\mu)$, but find

$$\begin{aligned} & \int_0^\infty p_c(s)W(x + s\mu)ds \\ &= \frac{e^{-\frac{x}{\nu_0}} \left(\mu(\nu_0 - 1)\nu_0 + e^{x\left(\frac{1}{\mu} + \frac{1}{\nu_0}\right)} (\mu(\mu^2 + 2\mu\nu_0 + \nu_0) - (\mu + 1)x(\mu + \nu_0)) \right)}{\mu(\mu + \nu_0)^2} \end{aligned}$$

Past $s = -x/\mu$ we will escape the boundary, so we only need to compute this probability and sample a continuous depth in the case that we do not escape. We find the escape probability

$$\begin{aligned} p_{esc}(x, \mu) &= \frac{\int_{-x/\mu}^\infty p_c(s)W(x + s\mu)ds}{\int_0^\infty p_c(s)W(x + s\mu)ds} \\ &= \frac{(\mu + \nu_0)^2(\mu - x)e^{x\left(\frac{1}{\mu} + \frac{1}{\nu_0}\right)}}{\mu(\nu_0 - 1)\nu_0 + e^{x\left(\frac{1}{\mu} + \frac{1}{\nu_0}\right)} (\mu(\mu^2 + 2\mu\nu_0 + \nu_0) - (\mu + 1)x(\mu + \nu_0))}. \quad (48) \end{aligned}$$

If we sample to stay inside the medium, using a random choice $\xi > p_{esc}(x, \mu)$ then we sample a free-path length distance s from

$$\begin{aligned} p_c^g(s; x, \mu) &= \frac{p_c(s)W(x + s\mu)}{\int_0^\infty p_c(s)W(x + s\mu)ds} \\ &= -\frac{\mu s(\mu + \nu_0)^2 e^{-\frac{s(\mu + \nu_0)}{\nu_0}}}{\nu_0 \left(e^{x\left(\frac{1}{\mu} + \frac{1}{\nu_0}\right)} (\mu\nu_0 - x(\mu + \nu_0)) - \mu\nu_0 \right)}, \quad -1 < \mu < 0. \quad (49) \end{aligned}$$

We can sample this by CDF inversion finding

$$s = \frac{\nu_0 \left(-W_{-1} \left(\xi \left(e^{x\left(\frac{1}{\mu} + \frac{1}{\nu_0}\right)} \right)^{-1} \left(x \left(\frac{1}{\mu} + \frac{1}{\nu_0} \right) - 1 \right) + \frac{1}{e} \right) - \frac{1}{e} \right)}{\mu + \nu_0} \quad (50)$$

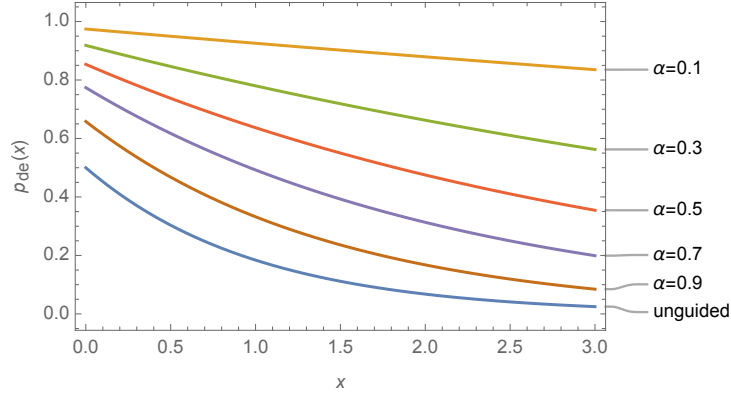


Figure 9: The probability of directly escaping the medium with no further collisions $p_{de}(x)$ when leaving a collision at a depth x in the half space. The guided walk prefers direct escape with increasing probability as the single-scattering albedo α reduces. For the classical unguided/analog walk, the direct escape probability is independent of α and equal to $1/2$ at the boundary $x = 0$.

where $W_{-1}(x)$ is a Product Log function.

If an escape is sampled, we incur one last weight factor using an expected value optimization, the ratio of the analog escape probability to the corresponding guiding escape probability. From depth x along direction $-1 < \mu < 0$ we escape along the final path length of $d = -x/\mu$. The analog escape probability leaving a correlated event (the last collision) is then $X_c(d)$. The weight factor for escape is thus

$$w_{esc} = \frac{X_c(d)}{p_{esc}(x, \mu)}. \quad (51)$$

This completes the derivation of the zero variance walk, and illustrates all of the essential steps in deriving an exact or near-zero-variance walk for escape a 3D half space with no Fresnel interactions at the boundary.

Translationally-Invariant Guiding with Exit Resampling: Our derivation above has taken a purely sequential approach for determining the guided walk: a complete free-path-length distribution is determined and sampled, and then a phase function distribution, and back and forth until escape. This has led to rather complicated distributions in the upwelling hemisphere due to the discontinuity in the importance function past the boundary. It turns out that many of these complexities can be avoided if we instead sample directions and displacements assuming a non-truncated exponential importance that now extends upward past 1 outside of the volume,

$$W(x) = e^{-x/\nu_0}. \quad (52)$$

Regardless of initial depth, the guided displacement and direction sampling steps using this importance function reduce to the downwelling equations above but for all directions $-1 < \mu < 1$. The angular distribution that we sample over the full sphere is now the generalized discrete Case eigenfunction for our Gamma-2 flight [d'Eon 2019b]

$$\phi(\mu, \nu_0) = \frac{c}{2} \left(\frac{1}{1 + \mu/\nu_0} \right)^2. \quad (53)$$

With CDF inversion we find sampling of outgoing polar angle μ from

$$-\frac{-2\nu_0\xi + \nu_0 + 1}{\nu_0 - 2\xi + 1}, \quad (54)$$

where $\xi \in [0, 1]$ is a uniform random variate. Given outgoing μ , displacement sampling follows from Eq.(46) for all $-1 < \mu < 1$. The probability that this procedure escapes the volume over all possible outgoing directions is (using Eq.(48))

$$p_{esc\phi}(x) = \int_{-1}^0 \phi(\mu, \nu_0) p_{esc}(x, \mu) d\mu = \frac{(\nu_0 + 1) e^{\left(\frac{1}{\nu_0} - 1\right)x}}{2\nu_0} \quad (55)$$

and it can be shown that this exactly matches the probability of the more complicated scheme above. The problem is, however, that the outgoing directions leaving the medium, when escape is sampled, are not the distribution required for zero variance because we messed with the importance function outside of the volume. However, we can compute the exitant cosine distribution that the zero-variance walk does produce when starting from x and leaving in a single step,

$$p_e(x, \mu) = \frac{\int_0^\infty p_c(s) \Theta(-x - s\mu) ds}{\int_{-1}^0 \int_0^\infty p_c(s) \Theta(-x - s\mu) ds d\mu} = \frac{e^{\frac{x}{\mu} + x} (\mu - x)}{\mu}, \quad (56)$$

where $\Theta(x)$ is the Heaviside Function. We can sample direction cosine μ from Eq.(56) using

$$\mu = -\frac{x}{W_{productlog}\left(-\frac{e^x x}{\xi - 1}\right)} \quad (57)$$

where $W_{productlog}$ is the product log function, typically written as W . Combining these two results, the walk proceeds with the unclamped distance and angle decisions until escape is sampled. Then we back up to the last collision prior to escape, resample an outgoing direction using Eq.(56) and jump to the boundary along that path. The expected-value weight calculation for this escape sampling is a ratio of angle pdfs times a ratio of escape pdfs,

$$w_{esc} = \frac{1/2}{p_e(x, \mu)} \frac{X_c(-x/\mu)}{p_{esc\phi}(x)}. \quad (58)$$

We will see in the next section that this modified scheme is closely related to asymptotic guiding in a classical 3D half space and that resampled escape can greatly reduce the variance relative to the method originally presented for rendering [Křivánek and d'Eon 2014].

It is also fascinating to note that we have just derived two new zero variance estimators for classical scattering in the half rod, our first example above. Observe that if we enter the Gamma-2 half space by sampling a uniform (Lambertian) surface source, that the expected analog distance of the first collision is the simple exponential

$$2 \int_0^1 p_u(x/\mu) d\mu = e^{-x}. \quad (59)$$

From here, all displacements in the 3D space when projected onto the x -axis exactly behave as the classical exponential walk in 1D. And the final albedo of the 3D Gamma-2 half space under diffuse uniform illumination is exactly the same as the 1D classical rod:

$$2 \int_0^1 R(\mu) \mu = \frac{2}{\alpha} (1 - \sqrt{1 - \alpha}) - 1 \quad (60)$$

in agreement with (10). We also see the same probabilities for upwelling and downwelling directions in all three walks. This is a great example of how an importance sampling process can be achieved in many different ways with auxiliary dummy variables that place the simulation in a higher dimension space.

Further Considerations: We hope that our zero-variance estimators for the Gamma-2 GRT can add value in traditional rendering of classical media, despite the different free-path statistics. This hunch is based on limited testing of rendering objects with the diffuse BRDF for Gamma-2 GRT and comparing to Chandrasekhar’s H -function BRDF for the classical medium. Both transport BRDFs exhibit a dusty appearance and significantly differ from the “CG” Lambertian appearance. We notice very similar appearance between the Gamma-2 and exponential BRDFs (Figure 10), suggesting that Gamma-2 may be a generally useful replacement for classical transport. There are several other reasons to consider this proposal. In addition to having an exact zero-variance estimator for thick flat geometry, the BRDF for Gamma-2 GRT also has a explicit expression, which we call the *diffusion transport* BRDF [d’Eon 2019b]

$$f_r(\theta_i, \theta_o) = \frac{\alpha}{4\pi} \left(\frac{H(\mu_i)H(\mu_o)}{\mu_i + \mu_o} \right)^2 \left(\frac{\mu_i^2 + 3\mu_i\mu_o + \mu_o^2}{\mu_i + \mu_o} - \frac{U_1}{2(1 + \mu_i)^2(1 + \mu_o)^2} \right) \quad (61)$$

where

$$U_1 = (1 - \sqrt{1 - \alpha}) (\mu^2 + 3\mu_i\mu_o + 2\mu_o) (\mu_o^2 + 3\mu_i\mu_o + 2\mu_i) + \frac{\alpha\mu_i\mu_o}{\mu_i + \mu_o} (\mu_i^3 + \mu_o^3 + \mu_i\mu_o (2(\mu_i^2 + \mu_o^2 + 1) + 6\mu_i\mu_o + 3(\mu_i + \mu_o))) \quad (62)$$

with the Picard H function given in Eq.(30), and $\mu_i = \cos \theta_i, \mu_o = \cos \theta_o$. This avoids the integrals required to evaluate the Milne H function in Chandrasekhar’s BRDF. Also, this BRDF admits a simple closed-form albedo mapping. The diffuse albedo R of the Gamma-2 halfspace under uniform illumination is

$$R = \frac{\alpha}{(\sqrt{1 - \alpha} + 1)^2} \quad (63)$$

which easily inverts to single scattering albedo α from diffuse albedo R ,

$$\alpha = \frac{4R}{(R + 1)^2}. \quad (64)$$

There may also be opportunity to apply some of the sampling distributions in this zero variance walk to different types of media with some appropriate fitting procedures.

6.6 Asymptotic (Dwivedi) Guiding

In the last two examples, we saw exact zero variance walks from absorbing half spaces with isotropic scattering. These were possible because the importance functions were known exactly and were simple expressions that admitted the required sampling manipulations. This is atypical of practical problems, even in plane geometry, so now we turn our attention to scenarios where we are forced to assume some approximate function for importance-to-escape; specifically, the approximation that results from taking the rigorous asymptotic diffusion term from the exact solution and discarding the transient portion. This method is highly effective for shielding calculations through optically thick shields because far from the boundaries, the transient terms in the exact importance function fall off and the resulting guiding becomes exact. In our previous work we attributed this method to Dwivedi [1982] but it appears that the original proposal of asymptotic guiding was earlier [Lanore 1971; Marchuk et al. 2013]. See also several more recent works on the topic [Meng et al. 2016; Medvedev and Mikhailov 2008].



Figure 10: Comparison of 3 diffuse BRDFs. Chandrasekhar’s BRDF and the new diffusion transport BRDF for Gamma-2 GRT look very similar, but the latter has a zero-variance random walk and simple albedo mapping.

Motivation Like the examples above, the asymptotic guiding zero-variance method begins by first trying to find an exact importance-to-escape function $W(x)$. For classical exponential transport in a 3D half space with isotropic scattering the Milne kernel arises and is singular. Here, the exact importance function for escape is not a simple exponential. Instead, we find Case’s exact solution involving a discrete asymptotic diffusion term (an exponential with a complicated constant) and a transient term that is an integral of exponentials [Case 1960; McCormick and Kuščer 1973; d’Eon 2016; d’Eon and McCormick 2019]. This relates to a rich set of results that began with observations by Davison [2000] and later expanded upon by Case [1960]. The importance function that results can also be equivalently found via the Wiener Hopf method. The final solution is expressed as a Fourier inversion, and via contour manipulation the discrete portion of the answer pops out as the residue of a pole, creating a diffusion result—but not the P_1 or “classical” diffusion result—the diffusion length is different. For anisotropic scattering the same things happens but more than one discrete diffusion term appear as the phase function gets increasingly peaked.

We now have the exact answer at hand, but an issue arises. The transient portion of the importance function involves integrals of eigenfunctions that are singular in direction⁴ and sometimes negative and so are not amenable to guiding. This has motivated the approximation of discarding the transient term and assuming the discrete term well approximates the full solution. For escaping a 3D half space, this becomes simply the translationally invariant $W(x) = e^{-x/\nu_0}$, where ν_0 is the discrete eigenvalue of the Milne kernel.

Discrete Eigenvalue Having made the approximation for $W(x)$ we proceed with the derivation analogous to the previous example for Gamma-2 GRT. The diffusion length we want follows from normalization of the guided angle sampling distribution

$$\phi(\mu, \nu_0) = \frac{\alpha}{2} \int_0^\infty p_c(s) e^{-s\mu/\nu_0} = \frac{c}{2} \left(\frac{1}{1 + \mu/\nu_0} \right). \quad (65)$$

Normalizing this polar angle distribution produces the dispersion equation

$$1 = \frac{\alpha\nu_0 \tanh^{-1} \left(\frac{\ell}{\nu_0} \right)}{\ell}. \quad (66)$$

⁴The eigen expansion of the angular collision rate and radiance inside the volume must include singularities and generalized distribution “functions” because of the reduced-intensity term from the source at the boundary, which is a delta in direction. In fact, even with a diffuse source at the boundary, the exact radiance in the volume at each depth is expressed as a superposition of the singular distributions even though the final result is smooth.

Our approximate importance $W(x)$ follows from finding the positive real root ν_0 of this equation. Eq.(66) is often called a transcendental equation but actually has a closed-form solution [Siewert 1980; d'Eon and McCormick 2019]. The exact solution is not numerically convenient, so we recommend the following approximation, with a relative error bounded by 0.0001

$$\nu_0 \approx \ell \frac{1}{\sqrt{1 - \alpha^{2.44294 - 0.0215813\alpha + \frac{0.578637}{\alpha}}}}. \quad (67)$$

Equation (67) is an order of magnitude more accurate than other piecewise approximations [Winslow 1968; Harel et al. 2020].

The remaining details of the asymptotic guiding scheme are found in several works [Dwivedi 1982; Křivánek and d'Eon 2014; Meng et al. 2016; Lanore 1971; Marchuk et al. 2013]. We will touch upon various select topics related to the method and refer the reader to these works for full details.

Weight Factor Simplification It is worth mentioning why this particular form of approximate importance function works so well and why, despite the approximation, undesired weight fluctuations that plagued earlier attempts to apply the exponential transform don't arise for this scheme. This happens because of a synergistic cancellation between weight factors in the direction and step length sampling steps [Dwivedi and Gupta 1986]. Referring now briefly to the notation in [Křivánek and d'Eon 2014], the weight adjustment when sampling stretched transition distance picks up a multiplicative weight correction of

$$w_s = \frac{e^{-s}}{\sigma'_t e^{-s\sigma'_t}}. \quad (68)$$

The angle selection incurs a multiplicative weight correction of

$$w_\mu = \frac{1}{2} \frac{1}{\frac{\alpha}{2} \frac{1}{1-\mu/\nu_0}}. \quad (69)$$

The eigenfunction $\phi(\mu, \nu_0)$ that appears in the denominator of w_μ mostly cancels with the σ'_t in w_s . When using fitted or tabulated distributions for angle and step lengths that do not exhibit this precise cancellation there can be low number of paths where significantly high particle weights arise.

We can further simplify the final weight w_o after angle selection, absorption and transition, expressed as a multiplication of the previous weight w_i before collision with the other weight adjustments, including the single-scattering albedo multiplier α , sees significant cancellation, giving simply

$$w_o = w_i * \alpha * w_s * w_\mu = w_i * \frac{\xi_s^{-1 + \frac{\nu_0}{\sigma_t(\nu_0 - \mu)}}}{\sigma_t} \quad (70)$$

where ξ_s was the random number used to sample displacement s .

Curved Geometry and General Lighting For general shapes, multiple importance sampling (MIS) can be used to combine analog/unguided sampling decisions with guided ones [Křivánek and d'Eon 2014]. This avoids the increased variance in regions with high curvature where particles exit the medium where the importance function was expected to be low. Figure 11 illustrates the impact of this combination of classical and guided estimators. Figure 12 shows the performance of the method under general lighting. Despite not sampling the product of BSSRDF and lighting, the reduction of the absorption variance is significant. Also, the average path length is reduced in guiding paths out of the medium more often than the classical walk. The histograms over path-length for both methods are compared in Figure 14 and examples scattering histories are shown in Figure 13 to more clearly illustrate how increased absorption alters the set of sampled paths. For expanded results on handling general light and geometry see [Meng et al. 2016].

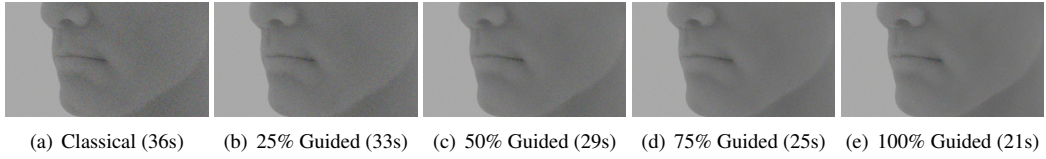


Figure 11: MIS between guided and classical sampling.

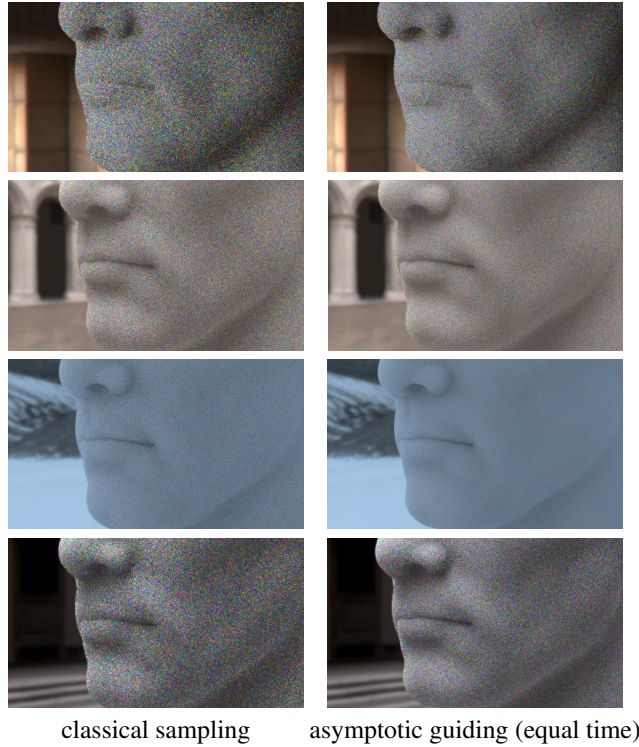


Figure 12: A gray material with isotropic scattering and single-scattering albedo of 0.943 under a variety of illumination conditions. The images rendered with classical sampling use 100 samples/pixel while with guided sampling can perform 50% more samples/pixel in the same time. The guided sampling assumes uniform hemispherical illumination everywhere on the surface and flat geometry yet still improves the convergence rate of random walk SSS for curved geometry under arbitrary illumination conditions.

6.6.1 Fresnel Boundaries

Largely missing from the zero-variance literature is the role of general BSDFs at medium boundaries and the impact of this on the zero-variance scheme. To see the influence of BSDF interactions on the derivation, consider the half space: in the downwelling directions the procedure is as before. We can think of upward angle selection as before but now the probability to leave a collision in an upwelling direction depends upon the more complicated result of importance from future collisions up to the boundary plus a new term that considers reflections back into the medium and the total escape probability, which is now a general BSDF integral over the exitant hemisphere,

$$\int_{4\pi} f_s(\omega_i, \omega) |\omega \cdot \vec{n}| W_b(\omega) d\omega \quad (71)$$

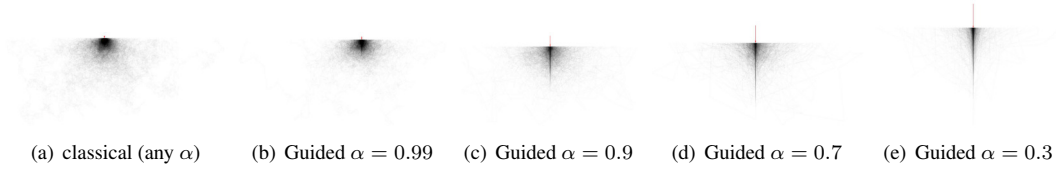


Figure 13: In each subfigure we show 2000 randomly sampled paths created using either classical volumetric sampling (a) or the Dwivedi sampling scheme (b-e). The figures have differing scales—the red arrow is one mean-free-path long and indicates the illumination position and direction. All paths continue inside the semi-infinite medium with isotropic scattering until an escape is sampled. Each path is rasterized with the same opacity, regardless of sample weight. Irrespective of absorption level (the value of α), the classical scheme samples the wide distribution of paths shown in (a), even though many of these paths are heavily absorbed and contribute negligible energy to the final result. Russian roulette helps avoid this wasteful sampling, but increases variance of each sample as a consequence. The Dwivedi sampling scheme we use adapts to the absorption levels of the medium and creates shorter, important paths more often, while simultaneously decreasing the variance of each sample.

where we define $W_b(\omega)$ as the importance function that is the probability that a particle leaving the boundary along direction ω eventually escapes, which is

$$W_b(\omega) = \begin{cases} \int_0^\infty p_u(s)W(s\mu)ds, & \omega \text{ is downwelling} \\ 1, & \omega \text{ is upwelling.} \end{cases} \quad (72)$$

We also have a new sampling decision to make upon jumping to the boundary during the walk, which is guided sampling of the BSDF. As in the derivation of the other steps, we start with the analog sampling distribution, the BSDF itself, and multiply it by the corresponding importance function $W_b(\omega)$ and normalize the result. Thus, having arrived at the boundary from inside *from* direction ω_i we must sample guided direction ω leaving the boundary from the normalization of

$$\int_{4\pi} f_s(\omega_i, \omega)W_b(\omega) |\cos \theta_o| d\omega. \quad (73)$$

For anything but a smooth Fresnel interface, this becomes a complicated problem to sample analytically. Novel methods will be required to efficiently perform this sampling for rough dielectric interfaces with multiple scattering [Dupuy et al. 2016; Heitz et al. 2016].

6.6.2 Asymptotic Guiding with Exit Resampling

We briefly tested the exit resampling approach from the Gamma-2 GRT estimator in the case of classical exponential transport in a 3D half space with isotropic scattering and indexed-matched smooth boundary. The approach uses the procedure described in prior work [Křivánek and d'Eon 2014; Meng et al. 2016] but when the translationally-invariant sampling produces escape, we backup and resample outgoing polar angle, now from

$$\frac{e^{\frac{x}{\mu}+x}}{1 - e^x x E_1(x)} \quad (74)$$

where $E_1(x)$ is the exponential integral function. We sampled this using naive rejection and performed some tests viewing flat patches of half space under uniform white illumination (Figure 15). We found the reduction in variance for resampled Dwivedi vs Dwivedi ranging from 10 times lower for $\alpha = 0.95$ to 45 times lower for $\alpha = 0.3$. We expect the additional sampling time is mostly due to the naive rejection sampling.

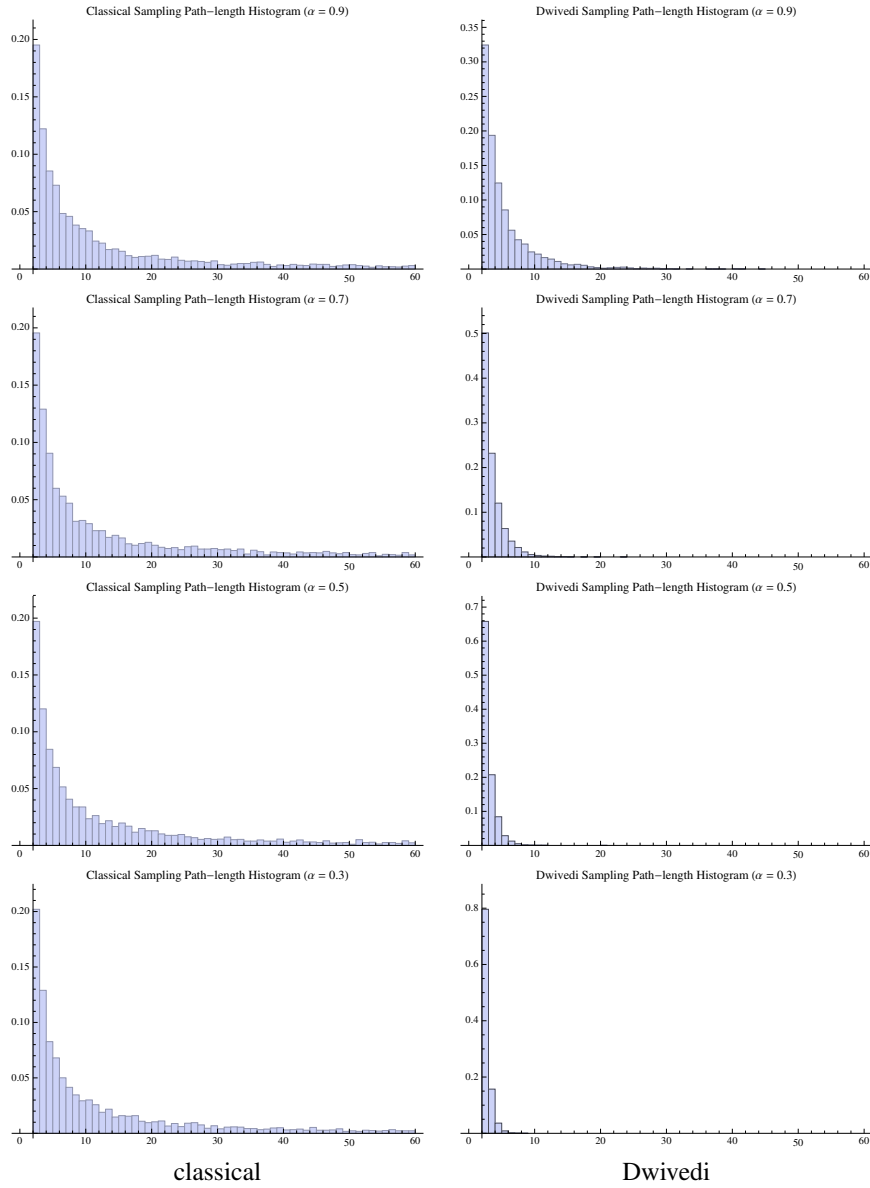


Figure 14: Comparison of the distributions of path lengths (in terms of path segment count) generated by classical sampling (without Russian roulette) and our application of Dwivedi sampling for the problem of reflection of normally-incident illumination from an isotropically-scattering semi-infinite medium. The zero-variance-based Dwivedi sampling scheme generates much shorter paths on average whilst simultaneously decreasing variance (as opposed to Russian roulette). The method automatically adapts to the single-scattering albedo α of the medium.

This suggests that much of the remaining variance in asymptotic guiding is not so much from errors in the importance function inside the medium but from not clamping it to 1 outside. While the exit resampling procedure would not be easy to apply in general curved geometry, this result suggests that finding a clamped exponential sampling scheme would be well worth the effort.

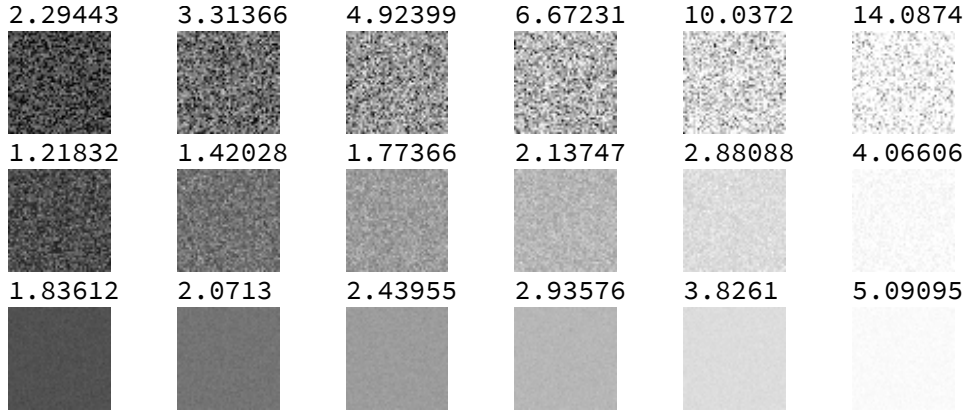


Figure 15: Normally-viewed patches of a 3D half space under uniform white illumination rendered with three estimators, classical (top), standard Dwivedi (middle), and Dwivedi with exit resampling (bottom). Single scattering albedo from left to right: 0.3, 0.5, 0.7, 0.8, 0.9, 0.95. Each patch is 50 by 50 pixels with 5 samples per pixel. Timings in seconds above each patch. Gamma correction of 2.0.

6.6.3 Asymptotic Guiding in GRT

In Section 6.5.2 we considered a form of GRT in 3D with Gamma random flights that admits an exactly zero-variance walk analytically. We also saw that asymptotic guiding was not a zero-variance walk, but could be corrected with exit resampling. We chose this form of GRT because of its mathematical properties. It is the unique form of GRT in 3D with isotropic scattering where the collision density inside the volume exactly satisfies a diffusion equation [d'Eon 2013]. Diffusion is *not an approximation* in Gamma-2 3D!. While helpful for illustrating how guided walks are derived, we are unaware of any specific microstructure that would motivate these exact free-flight statistics. It likely corresponds to a short-length negative correlation of some kind. For more general forms of GRT motivated by observed spatial variability in the volume coefficients, diffusion will not be an exact answer and asymptotic guiding or alternative approximate importance functions will be needed.

One popular [Davis 2006; Wrenninge et al. 2017; Jarabo et al. 2018; Bitterli et al. 2018] and practical GRT model for random media that includes long-range correlations and power-law asymptotics, while avoiding the more complex Mittag-Leffler functions that satisfy fractional diffusion equations [Liemert and Kienle 2018], derives from a continuum model of random scattering-particle number densities drawn from a Gamma distribution, producing [d'Eon 2018; Jarabo et al. 2018]

$$p_c(s) = a(a+1)\ell(al)^\alpha(al+s)^{-a-2}, \quad a > 0 \quad (75)$$

where the mean free path between collisions is ℓ and a shape parameter $a > 0$ adjusts the correlation between scattering events with classical exponential media recovered in the limit $a \rightarrow \infty$. The intercollisions distribution $p_c(s)$ does not decay exponentially due to the long-range nature of the correlations. For isotropic scattering in 3D plane geometry, this leads to a discrete Case eigenfunction derivation of [d'Eon 2019a]

$$\phi(\mu, \nu_0) = \frac{\alpha}{2} \int_0^\infty p_c(s) e^{-s\mu/\nu_0} = \frac{\alpha}{2} (a+1) e^{\frac{a\mu\ell}{\nu_0}} E_{a+2} \left(\frac{a\mu\ell}{\nu_0} \right), \quad 0 < \mu \leq 1. \quad (76)$$

The integral diverges, however, in the upwelling directions, so the exponential, and unbounded, importance function could only be used to guide downwelling direction sampling. This illustrates a failure

of the approach of Case that assumes exponentially-decaying kernels. For this class of flights, the dispersion equation admits a pair of complex roots, but no real ν_0 eigenvalue exists. It is an interesting open problem to investigate what asymptotic importance function might apply in this setting and if the Mittag-Leffler functions that generalize the exponential distribution make an appearance here.

6.6.4 Anisotropic Scattering

Including anisotropic scattering in guiding-to-escape walk derivations complicates things substantially. The importance function for escape upon entering a collision depends on the cosine μ as well as the position. The direction sampling is much more complicated, requiring importance to leave a collision $W_o(\mathbf{x}, \omega)$ in terms of general direction and to sample the product of this distribution with the phase function, for which the normalization factor is typically impossible to determine analytically. To address this issue Lanore [1971] offers some insight. We recommend Ueki and Larsen [1998] for more details on linearly and quadratically-anisotropic phase functions and procedures for sampling the product of the phase function and the importance function, and also [Marchuk et al. 2013].

6.7 General Tips

Validating the Walk When deriving analytic importance functions or fitting tabulated data from adjoint Monte Carlo simulation it can be helpful to ensure the correctness of these solutions using forward Monte Carlo simulation to simulate exactly the probability that is needed at a given sampling step. For example, if we require $W(x, \omega)$, the probability for a single particle to escape the medium upon entering a collision at x into direction ω , then we would start a Monte Carlo random walk at x that begins by sampling a collision right away, applying α and sampling the phase function with direction ω before stepping through the volume. Testing this for a variety of absorption levels α s, depths x and directions ω will validate any adjoint fittings or derivations. If W is off by even a small forgotten factor of α , the resultant walk will continue to show considerable variance.

Another debugging tool that we found helpful is to check at each collision entry that $wW(x, \omega) = R$. When an implementation that should be zero variance is not, this can help identify what step is causing the issue. This can also help identify what steps in an almost-zero-variance walk are causing the most variance.

Finally, the walk should always reduce to the classical method of analog sampling plus implicit capture when absorption is removed, $\alpha = 1$.

Russian Roulette Russian roulette is a common device for reducing the lengths of long random walks when the weight becomes low [Arvo and Kirk 1990]. However, if the importance function and its use to guide the random walk are both accurate, then it is most likely that Roulette will only increase variance and possibly reduce efficiency. Hero wavelength sampling and MIS complicate this conclusion, however. We recommend undertaking a thorough analysis for your particular problem to determine how and when to employ roulette with guided walks.

6.8 Acknowledgements

I am grateful to Jaroslav for travelling to New Zealand to work together with me on this topic in what would become one of the most rewarding and thrilling collaborations I've been lucky to enjoy. He approached the neutron transport literature with great excitement and fascination and with an unwavering determination to find an exactly zero-variance walk. When our sandbox experiments showed the potential of Dwivedi's asymptotic guiding for half space problems he saw exactly how to defeat the deficiencies of the method in curved geometry with MIS and had it working in Manuka within a week—a feat that would certainly have not happened without him. We lost a gifted researcher from whom we learned so much, and also a dear friend.

References

- ANISIMOV, O., AND FUKSHANSKY, L. 1992. Stochastic radiation in macroheterogeneous random optical media. *Journal of Quantitative Spectroscopy and Radiative Transfer* 48, 2, 169–186. [https://doi.org/10.1016/0022-4073\(92\)90087-K](https://doi.org/10.1016/0022-4073(92)90087-K).
- ARVO, J., AND KIRK, D. 1990. Particle transport and image synthesis. *ACM SIGGRAPH Computer Graphics* 24, 4, 63–66. <https://doi.org/10.1145/97879.97886>.
- AUDIC, S., AND FRISCH, H. 1993. Monte-Carlo simulation of a radiative transfer problem in a random medium: Application to a binary mixture. *Journal of Quantitative Spectroscopy and Radiative Transfer* 50, 2, 127–147. [https://doi.org/10.1016/0022-4073\(93\)90113-V](https://doi.org/10.1016/0022-4073(93)90113-V).
- BHAN, K., AND SPANIER, J. 2007. Condensed history monte carlo methods for photon transport problems. *Journal of computational physics* 225, 2, 1673–1694. <https://doi.org/10.1016/j.jcp.2007.02.012>.
- BITTERLI, B., RAVICHANDRAN, S., MÜLLER, T., WRENNINGE, M., NOVÁK, J., MARSCHNER, S., AND JAROSZ, W. 2018. A radiative transfer framework for non-exponential media. *ACM Transactions on Graphics* 37, 6. <https://doi.org/10.1145/3272127.3275103>.
- BORSHUKOV, G., AND LEWIS, J. P. 2003. Realistic human face rendering for “The Matrix Reloaded”. In *ACM SIGGRAPH Sketches and Applications*, ACM, 1. <https://doi.org/10.1145/1198555.1198593>.
- BURRUS, W. 1958. How channeling between chunks raises neutron transmission through boral. *Nucleonics (US) Ceased publication* 16.
- BURRUS, W. 1960. Radiation transmission through boral and similar heterogeneous materials consisting of randomly distributed absorbing chunks. Tech. rep., Oak Ridge National Lab., Tenn. <https://doi.org/10.2172/4196641>.
- CASE, K. 1960. Elementary solutions of the transport equation and their applications*. *Annals of Physics* 9, 1, 1–23. [https://doi.org/10.1016/0003-4916\(60\)90060-9](https://doi.org/10.1016/0003-4916(60)90060-9).
- CHANDRASEKHAR, S. 1960. *Radiative Transfer*. Dover.
- CHIANG, M. J.-Y., KUTZ, P., AND BURLEY, B. 2016. Practical and controllable subsurface scattering for production path tracing. In *ACM SIGGRAPH 2016 Talks*. 1–2. <https://doi.org/10.1145/2897839.2927433>.
- CHRISTENSEN, P., FONG, J., SHADE, J., WOOTEN, W., SCHUBERT, B., KENSLER, A., FRIEDMAN, S., KILPATRICK, C., RAMSHAW, C., BANNISTER, M., RAYNER, B., BROUILLAT, J., AND LIANI, M. 2018. Renderman: An advanced path-tracing architecture for movie rendering. *ACM Trans. Graph.* 37, 3 (Aug.). <https://doi.org/10.1145/3182162>.
- CHRISTENSEN, P. H. 2003. Adjoints and importance in rendering: An overview. *IEEE Transactions on Visualization and Computer Graphics* 9, 3, 329–340. <https://doi.org/10.1109/TVCG.2003.1207441>.
- CHRISTENSEN, P. H. 2015. An approximate reflectance profile for efficient subsurface scattering. In *ACM SIGGRAPH 2015 Talks*. 1–1. <https://doi.org/10.1145/2775280.2792555>.
- COVEYOU, R., CAIN, V., AND YOST, K. 1967. Adjoint and Importance in Monte Carlo Application. *Nucl. Sci. Eng.* 27, 1, 219–234. <https://doi.org/10.13182/NSE67-A18262>.

- DAVIS, A. B., AND MINEEV-WEINSTEIN, M. B. 2011. Radiation propagation in random media: From positive to negative correlations in high-frequency fluctuations. *Journal of Quantitative Spectroscopy and Radiative Transfer* 112, 4, 632–645. <https://doi.org/10.1016/j.jqsrt.2010.10.001>.
- DAVIS, A. B., AND XU, F. 2014. A generalized linear transport model for spatially correlated stochastic media. *Journal of Computational and Theoretical Transport* 43, 1-7, 474–514. <https://doi.org/10.1080/23324309.2014.978083>.
- DAVIS, A. B. 2006. Effective propagation kernels in structured media with broad spatial correlations, illustration with large-scale transport of solar photons through cloudy atmospheres. In *Computational Methods in Transport*. Springer, 85–140. https://doi.org/10.1007/3-540-28125-8_5.
- DAVISON, B. 2000. Angular distribution due to an isotropic point source and spherically symmetrical eigensolutions of the transport equation (MT-112). *Progress in Nuclear Energy* 36, 3, 323 – 365. Nuclear Reactor Theory in Canada 1943-1946. [https://doi.org/10.1016/S0149-1970\(00\)00012-3](https://doi.org/10.1016/S0149-1970(00)00012-3).
- DENG, H., WANG, B., WANG, R., AND HOLZSCHUCH, N. 2020. A practical path guiding method for participating media. *Computational Visual Media*, 1–15. <https://doi.org/10.1007/s41095-020-0160-1>.
- D'EON, E., AND IRVING, G. 2011. A quantized-diffusion model for rendering translucent materials. In *ACM Transactions on Graphics (TOG)*, vol. 30, ACM, 56. <https://doi.org/10.1145/2010324.1964951>.
- D'EON, E., AND MCCORMICK, N. J. 2019. Radiative transfer in half spaces of arbitrary dimension. *Journal of Computational and Theoretical Transport* 48, 7, 280–337. <https://doi.org/10.1080/23324309.2019.1696365>.
- D'EON, E., LUEBKE, D., AND ENDERTON, E. 2007. Efficient rendering of human skin. In *Rendering Techniques*, 147–157. <https://doi.org/10.5555/2383847.2383869>.
- D'EON, E. 2013. Rigorous Asymptotic and Moment-Preserving Diffusion Approximations for Generalized Linear Boltzmann Transport in Arbitrary Dimension. *Transport Theory and Statistical Physics* 42, 6-7, 237–297. <https://doi.org/10.1080/00411450.2014.910231>.
- D'EON, E. 2014. A Dual-beam 3D Searchlight BSSRDF. *ACM SIGGRAPH 2014 Talks* 65, 1, 1. <http://doi.acm.org/10.1145/2614106.2614140>.
- D'EON, E. 2016. A Hitchhiker's Guide to Multiple Scattering. (*self published*). <http://eugenedeon.com/hitchhikers>.
- D'EON, E. 2018. A reciprocal formulation of nonexponential radiative transfer. 1: Sketch and motivation. *Journal of Computational and Theoretical Transport*. <https://doi.org/10.1080/23324309.2018.1481433>.
- D'EON, E. 2019. A reciprocal formulation of nonexponential radiative transfer. 2: Monte-Carlo Estimation and Diffusion Approximation. *Journal of Computational and Theoretical Transport* 48, 6, 201–262. <https://doi.org/10.1080/23324309.2019.1677717>.
- D'EON, E. 2019. The Albedo Problem in Nonexponential Radiative Transfer. In *International Conference on Transport Theory (ICTT-26) - Abstracts*. <https://www.researchgate.net/publication/333325137>.
- DONNER, C., AND JENSEN, H. W. 2005. Light Diffusion in Multi-Layered Translucent Materials. *ACM Trans. Graphic.* 24, 3, 1032–1039. <https://doi.org/10.1145/1186822.1073308>.

- DONNER, C., WEYRICH, T., D'EON, E., RAMAMOORTHI, R., AND RUSINKIEWICZ, S. 2008. A layered, heterogeneous reflectance model for acquiring and rendering human skin. In *SIGGRAPH Asia '08: ACM SIGGRAPH Asia 2008 papers*, ACM, New York, NY, USA, 1–12. <https://doi.org/10.1145/1457515.1409093>.
- DUPUY, J., HEITZ, E., AND D'EON, E. 2016. Additional progress towards the unification of microfacet and microflake theories. In *EGSR (EI&I)*, 55–63. <https://doi.org/10.5555/3056507.3056519>.
- DUTKA, J. 1985. On the problem of random flights. *Archive for history of exact sciences* 32, 3, 351–375. <https://doi.org/10.1007/BF00348451>.
- DWIVEDI, S., AND GUPTA, H. 1986. Biasing parameter limits for synergistic monte carlo in deep-penetration calculations. *Nuclear Science and Engineering* 92, 4, 545–549. <https://doi.org/10.13182/NSE86-A18611>.
- DWIVEDI, S. 1982. A new importance biasing scheme for deep-penetration Monte Carlo. *Annals of Nuclear Energy* 9, 7, 359–368. [https://doi.org/10.1016/0306-4549\(82\)90038-X](https://doi.org/10.1016/0306-4549(82)90038-X).
- FASCIONE, L., HANIKA, J., LEONE, M., DROSKE, M., SCHWARZHaupt, J., DAVIDOVIČ, T., WEIDLICH, A., AND MENG, J. 2018. Manuka: A batch-shading architecture for spectral path tracing in movie production. *ACM Transactions on Graphics (TOG)* 37, 3, 1–18. <https://doi.org/10.1145/3182161>.
- FELLER, W. 1971. *An Introduction to Probability theory and its application Vol II*. John Wiley and Sons.
- FLECK, J., AND CANFIELD, E. 1984. A random walk procedure for improving the computational efficiency of the implicit monte carlo method for nonlinear radiation transport. *Journal of Computational Physics* 54, 3, 508–523. [https://doi.org/10.1016/0021-9991\(84\)90130-X](https://doi.org/10.1016/0021-9991(84)90130-X).
- FREDERICKX, R., AND DUTRÉ, P. 2017. A forward scattering dipole model from a functional integral approximation. *ACM Transactions on Graphics (TOG)* 36, 4, 109. <https://doi.org/10.1145/3072959.3073681>.
- FRISCH, U. 1968. Wave propagation in random media. In *Probabilistic methods in applied mathematics*, A. Bharucha-Reid, Ed. Academic Press, 75–198.
- FRISVAD, J. R., CHRISTENSEN, N. J., AND JENSEN, H. W. 2007. Computing the Scattering Properties of Participating Media Using Lorenz-Mie Theory. In *ACM SIGGRAPH 2007 Papers*, Association for Computing Machinery, New York, NY, USA, SIGGRAPH '07, 60–es. <https://doi.org/10.1145/1275808.1276452>.
- FRISVAD, J. R., HACHISUKA, T., AND KJELDSSEN, T. K. 2014. Directional dipole model for subsurface scattering. *ACM Transactions on Graphics (TOG)* 34, 1, 1–12. <https://doi.org/10.1145/2682629>.
- FUKSHANSKY, L. 1987. Absorption statistics in turbid media. *Journal of quantitative spectroscopy and radiative transfer* 38, 5, 389–406. [https://doi.org/10.1016/0022-4073\(87\)90033-1](https://doi.org/10.1016/0022-4073(87)90033-1).
- GEORGIEV, I., IZE, T., FARNSWORTH, M., MONTAYA-VOZMEDIANO, R., KING, A., LOMMEL, B. V., JIMENEZ, A., ANSON, O., OGAKI, S., JOHNSTON, E., HERUBEL, A., RUSSELL, D., SERVANT, F., AND FAJARDO, M. 2018. Arnold: A brute-force production path tracer. *ACM Trans. Graph.* 37, 3 (Aug.). <https://doi.org/10.1145/3182160>.
- GROSJEAN, C. 1951. The Exact Mathematical Theory of Multiple Scattering of Particles in an Infinite Medium. *Memoirs Kon. Vl. Ac. Wetensch.* 13, 36.

- GUO, Y., HAŠAN, M., AND ZHAO, S. 2018. Position-free monte carlo simulation for arbitrary layered bsdfs. *ACM Transactions on Graphics (TOG)* 37, 6, 1–14. <https://doi.org/10.1145/3272127.3275053>.
- HABEL, R., CHRISTENSEN, P. H., AND JAROSZ, W. 2013. Photon beam diffusion: a hybrid Monte Carlo method for subsurface scattering. In *Computer Graphics Forum*, vol. 32, Wiley Online Library, 27–37. <https://doi.org/10.1111/cgf.12148>.
- HAGHIGHAT, A., AND WAGNER, J. C. 2003. Monte Carlo variance reduction with deterministic importance functions. *Progress in Nuclear Energy* 42, 1, 25–53. [https://doi.org/10.1016/S0149-1970\(02\)00002-1](https://doi.org/10.1016/S0149-1970(02)00002-1).
- HANRAHAN, P., AND KRUEGER, W. 1993. Reflection from layered surfaces due to subsurface scattering. In *Proceedings of ACM SIGGRAPH 1993*, 164–174. <https://doi.org/10.1145/166117.166139>.
- HAREL, R., BUROV, S., AND HEIZLER, S. I. 2020. The Time-Dependent Asymptotic P_N Approximation for the Transport Equation. *arXiv preprint arXiv:2006.11784*. <https://arxiv.org/abs/2006.11784v1>.
- HEITZ, E., HANIKA, J., D'EON, E., AND DACHSBACHER, C. 2016. Multiple-scattering microfacet BSDFs with the Smith model. *ACM Transactions on Graphics (TOG)* 35, 4, 58. <https://doi.org/10.1145/2897824.2925943>.
- HOFFMAN, W. 1964. Wave propagation in a general random continuous medium. *Proc. Symp. Appl. Math.* 16, 117–144.
- HOOGENBOOM, J. 1981. A practical adjoint Monte Carlo technique for fixed-source and eigenfunction neutron transport problems. *Nuclear Science and Engineering* 79, 4, 357–373. <https://doi.org/10.13182/NSE81-A21387>.
- HOOGENBOOM, E. 2008. Zero-Variance Monte Carlo Schemes Revisited. *Nucl. Sci. Eng.* 160, 1, 1–22. <https://doi.org/10.13182/NSE160-01>.
- HOOGENBOOM, J. 2008. The Two-Direction Neutral-Particle Transport Model: A Useful Tool for Research and Education. *Transport Theory and Statistical Physics* 37, 1, 65–108. <https://doi.org/10.1080/00411450802271791>.
- ISHIMARU, A. 1978. *Wave Propagation and Scattering in Random Media*. Oxford University Press.
- IVANOV, V. 1994. Resolvent method: exact solutions of half-space transport problems by elementary means. *Astronomy and Astrophysics* 286, 328–337. <https://ui.adsabs.harvard.edu/abs/1994A&A...286..328I>.
- JARABO, A., ALIAGA, C., AND GUTIERREZ, D. 2018. A radiative transfer framework for spatially-correlated materials. *ACM Transactions on Graphics* 37, 4, 14. <https://doi.org/10.1145/3197517.3201282>.
- JENSEN, H. W., MARSCHNER, S. R., LEVOY, M., AND HANRAHAN, P. 2001. A practical model for subsurface light transport. In *Proceedings of ACM SIGGRAPH 2001*, 511–518. <https://doi.org/10.1145/383259.383319>.
- KAHN, H. 1956. Applications of monte carlo. Tech. Rep. RM-1237-AEC. https://www.rand.org/pubs/research_memoranda/RM1237.html.
- KIRK, J. 1975. A theoretical analysis of the contribution of algal cells to the attenuation of light within natural waters I. General treatment of suspensions of pigmented cells. *New phytologist* 75, 1, 11–20. <https://doi.org/10.1111/j.1469-8137.1975.tb01366.x>.

- KULLA, C., CONTY, A., STEIN, C., AND GRITZ, L. 2018. Sony Pictures Imageworks Arnold. *ACM Trans. Graph.* 37, 3 (Aug.). <https://doi.org/10.1145/3180495>.
- KŘIVÁNEK, J., AND D'EON, E. 2014. A Zero-variance-based sampling scheme for Monte Carlo sub-surface scattering. *ACM SIGGRAPH 2014 Talks* 66, 1, 1. <https://doi.org/10.1145/2614106.2614138>.
- LAFORTUNE, E. P., AND WILLEMS, Y. D. 1996. Rendering participating media with bidirectional path tracing. In *Rendering Techniques*, 91–100. https://doi.org/10.1007/978-3-7091-7484-5_10.
- LANORE, J.-M. 1971. Weighting and Biasing of a Monte Carlo Calculation for Very Deep Penetration of Radiation. *Nucl. Sci. Eng.* 45, 1, 66–72. <https://doi.org/10.13182/NSE71-A20346>.
- LARMIER, C., HUGOT, F.-X., MALVAGI, F., MAZZOLO, A., AND ZOIA, A. 2017. Benchmark solutions for transport in d -dimensional Markov binary mixtures. *Journal of Quantitative Spectroscopy and Radiative Transfer* 189, 133–148. <https://doi.org/10.1016/j.jqsrt.2016.11.015>.
- LARSEN, E. W., AND VASQUES, R. 2011. A generalized linear Boltzmann equation for non-classical particle transport. *Journal of Quantitative Spectroscopy and Radiative Transfer* 112, 4, 619–631. <https://doi.org/10.1016/j.jqsrt.2010.07.003>.
- LIEMERT, A., AND KIENLE, A. 2013. Exact and efficient solution of the radiative transport equation for the semi-infinite medium. *Scientific Reports* 3. <https://doi.org/10.1038/srep02018>.
- LIEMERT, A., AND KIENLE, A. 2018. Fractional radiative transport in the diffusion approximation. *Journal of Mathematical Chemistry* 56, 2, 317–335. <https://doi.org/10.1007/s10910-017-0792-2>.
- LU, B., AND TORQUATO, S. 1992. Lineal-path function for random heterogeneous materials. *Physical Review A* 45, 2, 922. <https://doi.org/10.1103/PhysRevA.45.922>.
- MACHIDA, M., PANASYUK, G., SCHOTLAND, J., AND MARKEL, V. 2010. The green's function for the radiative transport equation in the slab geometry. *Journal of Physics A: Mathematical and Theoretical* 43, 065402. <https://doi.org/10.1088/1751-8113/43/6/065402>.
- MARCHUK, G. I., MIKHAILOV, G. A., NAZARELIEV, M., DARBINJAN, R. A., KARGIN, B. A., AND ELEPOV, B. S. 2013. *The Monte Carlo methods in atmospheric optics*, vol. 12. Springer.
- MCCORMICK, N., AND KUŠČER, I. 1973. Singular eigenfunction expansions in neutron transport theory. In *Advances in Nuclear Science and Technology*, Academic Press, E. J. Henley and J. Lewins, Eds., vol. 7, 181–282. <https://doi.org/10.1016/B978-0-12-029307-0.50010-X>.
- MEDVEDEV, I., AND MIKHAILOV, G. 2008. A new criterion for finiteness of weight estimator variance in statistical simulation. In *Monte Carlo and Quasi-Monte Carlo Methods 2006*. Springer, 561–576. https://doi.org/10.1007/978-3-540-74496-2_33.
- MENG, J., HANIKA, J., AND DACHSBACHER, C. 2016. Improving the Dwivedi sampling scheme. In *Computer Graphics Forum*, vol. 35, Wiley Online Library, 37–44. <https://doi.org/10.1111/cgf.12947>.
- MOON, J., WALTER, B., AND MARSCHNER, S. 2007. Rendering discrete random media using precomputed scattering solutions. *Rendering Techniques 2007*, 231–242. <https://doi.org/10.2312/EGWR/EGSR07/231-242>.
- MÜLLER, T., PAPAS, M., GROSS, M., JAROSZ, W., AND NOVÁK, J. 2016. Efficient rendering of heterogeneous polydisperse granular media. *ACM Transactions on Graphics (TOG)* 35, 6, 1–14. <https://doi.org/10.1145/2980179.2982429>.

- MUNK, M., AND SLAYBAUGH, R. N. 2019. Review of hybrid methods for deep-penetration neutron transport. *Nuclear Science and Engineering*, 1–35. <https://doi.org/10.1080/00295639.2019.1586273>.
- NICODEMUS, F. E., RICHMOND, J. C., HSIA, J. J., GINSBERG, I. W., AND LIMPERIS, T. 1977. *Geometrical Considerations and Nomenclature for Reflectance*. National Bureau of Standards.
- NOVÁK, J., GEORGIEV, I., HANIKA, J., AND JAROSZ, W. 2018. Monte carlo methods for volumetric light transport simulation. In *Computer Graphics Forum*, vol. 37, Wiley Online Library, 551–576. <https://doi.org/10.1111/cgf.13383>.
- PHARR, M., JAKOB, W., AND HUMPHREYS, G. 2016. *Physically based rendering: From theory to implementation*. Morgan Kaufmann.
- PICARD, É. 1911. Sur un exemple simple d'une équation singulière de fredholm où la nature analytique de la solution dépend du second membre. In *Annales scientifiques de l'École Normale Supérieure*, vol. 28, 313–324. <https://eudml.org/doc/81303>.
- RAAB, M., SEIBERT, D., AND KELLER, A. 2008. Unbiased global illumination with participating media. In *Monte Carlo and Quasi-Monte Carlo Methods 2006*. Springer, 591–605. https://doi.org/10.1007/978-3-540-74496-2_35.
- RABINOWITCH, E. I. 1951. *Photosynthesis and related processes*, vol. 72. LWW.
- RANDALL, C. H. 1962. Stochastic modles for heterogeneous materials. I.(Large scale inhomogeneities and neutron transmission. Tech. Rep. KAPL-M-CHR-6, Knolls Atomic Power Lab., Schenectady, NY.
- RANDALL, C. 1964. Generalized treatment of particle self-shielding. In *The Naval Reactors Handbook Vol. 1: Selected Basic Techniques*, A. Radkowsky, Ed. United States Atomic Energy Comission, 553.
- SIEWERT, C. 1980. On computing eigenvalues in radiative transfer. *Journal of Mathematical Physics* 21, 9, 2468–2470. <https://doi.org/10.1063/1.524684>.
- SPANIER, J., AND GELBARD, E. 1969. *Monte Carlo principles and neutron transport problems*. Addison-Wesley Pub. Co.
- TORQUATO, S. 2016. Hyperuniformity and its generalizations. *Physical Review E* 94, 2, 022122. <https://doi.org/10.1103/PhysRevE.94.022122>.
- TUNALEY, J. 1974. Theory of ac conductivity based on random walks. *Physical Review Letters* 33, 17, 1037. <https://doi.org/10.1103/PhysRevLett.33.1037>.
- TUNALEY, J. 1976. Moments of the Montroll-Weiss continuous-time random walk for arbitrary starting time. *Journal of Statistical Physics* 14, 5, 461–463. <https://doi.org/10.1007/BF01040704>.
- TURNER, S. A., AND LARSEN, E. W. 1997. Automatic variance reduction for three-dimensional Monte Carlo simulations by the local importance function transform–1: Analysis. *Nucl. Sci. Eng.* 127, 1. <https://doi.org/10.13182/NSE127-22>.
- UEKI, T., AND LARSEN, E. W. 1998. A kinetic theory for nonanalog monte carlo particle transport algorithms: Exponential transform with angular biasing in planar-geometry anisotropically scattering media. *Journal of Computational Physics* 145, 1, 406–431. <https://doi.org/10.1006/jcph.1998.6039>.
- VICINI, D., KOLTUN, V., AND JAKOB, W. 2019. A learned shape-adaptive sub-surface scattering model. *ACM Transactions on Graphics (TOG)* 38, 4, 1–15. <https://doi.org/10.1145/3306346.3322974>.

- WEISS, G. H. 1983. Random walks and their applications: Widely used as mathematical models, random walks play an important role in several areas of physics, chemistry, and biology. *American Scientist* 71, 1, 65–71. <https://www.jstor.org/stable/27851819>.
- WILLIAMS, M. 1991. Generalized contribution response theory. *Nucl. Sci. Eng.* 108, 355–383. <https://doi.org/10.13182/NSE90-33>.
- WILLIAMS, M. M. R. 2007. The searchlight problem in radiative transfer with internal reflection. *Journal of Physics A: Mathematical and Theoretical* 40, 24, 6407. <https://doi.org/10.1088/1751-8113/40/24/009>.
- WING, G. 1962. *An introduction to transport theory*. Wiley.
- WINSLOW, A. M. 1968. Extensions of asymptotic neutron diffusion theory. *Nuclear Sci. and Eng.* 32, 101–110. <https://doi.org/10.13182/NSE68-A18829>.
- WRENNINGE, M., VILLEMEN, R., AND HERY, C. 2017. Path traced subsurface scattering using anisotropic phase functions and non-exponential free flights. Tech. Rep. 17-07, Pixar. <https://graphics.pixar.com/library/PathTracedSubsurface>.
- XU, Q., SUN, J., WEI, Z., SHU, Y., MESSELODI, S., AND CAI, J. 2001. Zero variance importance sampling driven potential tracing algorithm for global illumination.
- XU, Q., WANG, W., AND BAO, S. 2006. A new computational way to Monte Carlo global illumination. *International Journal of Image and Graphics* 6, 01, 23–34. <https://doi.org/10.1142/S0219467806002057>.
- ZHAO, S., RAMAMOORTHY, R., AND BALA, K. 2014. High-order similarity relations in radiative transfer. *ACM Transactions on Graphics (TOG)* 33, 4, 104. <https://doi.org/10.1145/2601097.2601104>.
- ZOIA, A., DUMONTEIL, E., AND MAZZOLO, A. 2011. Collision densities and mean residence times for d-dimensional exponential flights. *Physical Review E* 83, 4, 041137. <https://doi.org/10.1103/PhysRevE.83.041137>.

7 Path Guiding



Figure 2: Dr. Křivánek contributed to development of path guiding, a path sampling technique, which allows efficient rendering of notoriously difficult light transport conditions. It was adopted in production by Weta Digital already in 2014 and later used to render for example these shots from *Alita: Battle Angel* (VHH⁺19). It reduces render times of indirectly lit scenes including effects like specular-diffuse-specular caustics in eyes, god-rays in vast underwater scenes, or caustics on the lake bed. ©2018 Twentieth Century Fox Film Corporation. All rights reserved.

7.1 Introduction

I will dare to use this opportunity to tell my personal story about my professional relationship with Jaroslav while introducing topic of this chapter. I started my collaboration with Jaroslav in 2011 as his student while working on bidirectional photon mapping (Vor11) for my master thesis. We were looking for a robust algorithm to handle challenging light transport due to combinations of various glossy, specular, and diffuse materials in the scene without resorting to fragile heuristics. Jaroslav continued in this direction and, with his collaborators, they took this idea even further and combined the bidirectional photon mapping with path tracing, taking the best from both, and derived VCM¹ (GKDS12) path sampling framework (see Chapter ??). Already at that time, Jaroslav expressed his concerns that complex visibility will be probably prohibitive in many non-cornell-box-like scenes even for robust bidirectional estimators which are based on merging and connecting light sub-paths. Indeed, it turned out that without sampling paths in important regions there are almost no samples to be merged or connected and thus, in turn, efficiency of even advanced bidirectional estimators diminishes significantly.

We addressed this problem by importance sampling reflected rays based on incident radiance (VKŠ⁺14) and thus guiding them towards interesting regions which increases chance that sampled paths would transport significant amount of energy from lights to our pixels. People had been looking into this problem before us (see Sec.7.3) and it was just a good time to resume this research. It turned out to be relatively challenging task especially as we were just starting our PhD. However, Jaroslav's leadership and his contagious enthusiasm held our team together. For me, the most important lesson learned on this project was the strength and importance of collaboration and team work which Jaroslav always stressed and without which this project simply would not have worked out.

¹At the same time, the same algorithm known as UPS was independently discovered by Hachisuka and colleagues (HPJ12).

Important aspect of this 2014 path guiding paper is that it was probably the first work pointing out that path guiding can be formulated as learning uncertainty and as such, abundant toolbox of *machine learning* techniques opened up for exploration within path guiding context. The on-line learning approach we took enabled us to apply guiding in scenarios where only a handful of samples occur at early stages of light transport simulation. Further, this work immediately revealed the importance of guiding Russian roulette as it turned out that traditional albedo-driven path termination can work against the path guiding turning its advantages into pure overhead.

We addressed this problem by using learned approximations of light field also to terminate or split paths in a way that keeps their contributions oscillating around expected values of our pixels (VK16). Interestingly, we learned that similar techniques were already explored within neutron transport context many years ago and are essential part of actively used simulators like MCNP from Los Alamos National Laboratory (We17). Neutron transport frames the path guiding as we know it in computer graphics within zero-variance sampling theory which was subject of Chapter 6 and which is an invaluable tool for designing and reasoning about path guiding schemes.

Our work on path guiding revealed an interesting result with the respect of chasing the “one” sampling algorithm. We observed that guiding unidirectional path tracing with next event estimation within complex bidirectional estimators can make many merging and connecting sub-techniques almost redundant. This is still not true for extremely difficult conditions like caustics due to small light sources like for example realistic sun light (VHH⁺19). This observation mean, that sampling paths from some techniques can be expensive on time while their contribution is down-weighted within MIS framework and thus practically only extend the overall render time. This is one (but not the only) reason why such heavy techniques are not favored in production (VHH⁺19). Jaroslav continued to build upon this observation with his students and collaborators to make complex estimators more lightweight and practical (see Chapter 9).

In this chapter, we first define path guiding (Sec. 7.2) and acknowledge the previous and pioneering work in this area (Sec. 7.3). From Sec. 7.4 to Sec. 7.8 we go over Jaroslav’s research into path guiding. In Sec. 7.9, we briefly go over subsequent research in this area done by other researches and finally, in Sec. 7.10 and Sec. 7.11, we discuss industry impact of path guiding research on VFX industry and possible future avenues.

In this part, we cover path guiding techniques explored by the team around Jaroslav Křivánek, show their connection to zero-variance theory and neutron transport, and discuss the impact of these works to both current research and the industry.

7.2 Define “Path Guiding”

For efficient Monte-Carlo light transport simulation, it is vital to sample paths between camera and light sources which transfer the highest amount of energy while, at the same time, avoid wasting computational time on sampling irrelevant paths. In scenes with complex visibility, we could guide the path sampling and achieve low variance calculation if we knew the full transport in the scene, which however, is not known a-priori in practice. *Path guiding* refers broadly to techniques which use a global knowledge about the scene, more specifically approximation of radiance field in the scene and/or additional sampling statistics, to efficiently distribute transport paths (VHH⁺19) to sample both direct and indirect illumination.

Nevertheless, in this chapter, we aim specifically at sampling only indirect illumination. Jaroslav’s endeavor in guiding sampling of direct illumination is subject to Chapter 8.

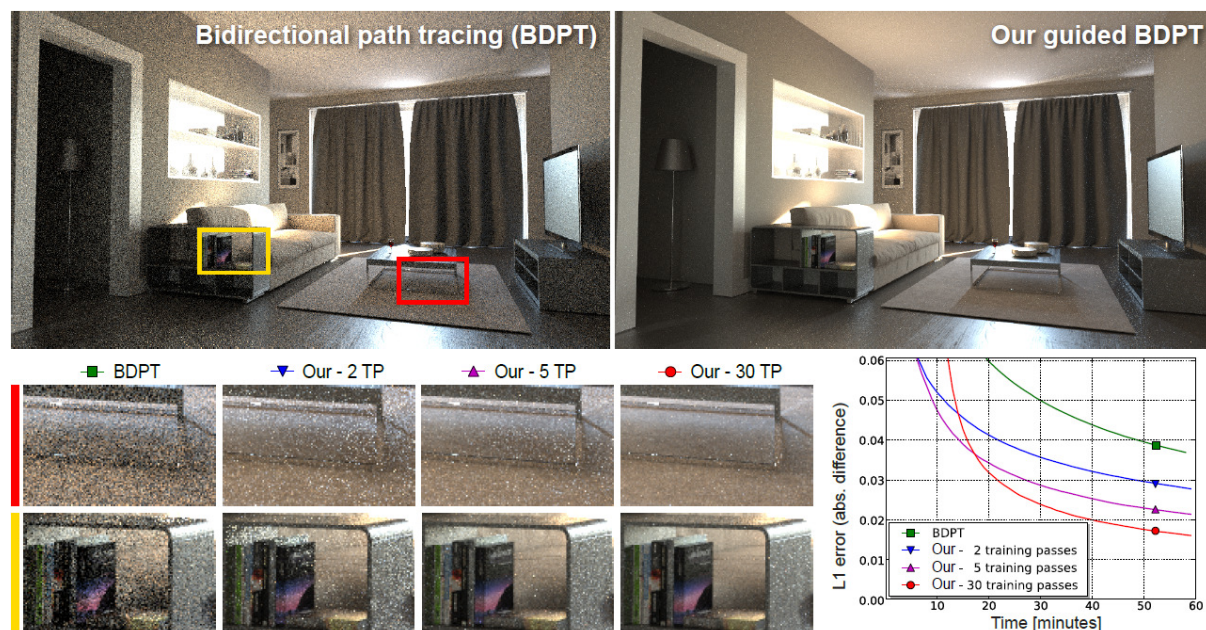


Figure 3: On-line path guiding progressively learns from previous samples on a scene rendered by bidirectional path tracing (BDPT). Light transport in this scene is difficult for sampling because sun light enters the room through a small gap. Path guiding significantly reduces noise (left) as opposed to traditional BDPT. The path guiding performance depends on the number of samples used for learning the global information about the transport in the scene as shown in the insets and plots. The illustration is borrowed from work of (VKŠ⁺14).

The approximation of radiance field in the scene is learned from previously sampled paths or in a pre-process step. The paths are samples of an estimator of the measurement equation shown in Sec. 6. The outgoing radiance L from a point \mathbf{x} to a direction ω_o is described by the rendering equation

$$L(\mathbf{x}, \omega_o) = L_e(\mathbf{x}, \omega_i) + \underbrace{\int_S L(\mathbf{x}, \omega_i) \rho_s(\mathbf{x}, \omega_o, \omega_i) d\omega_i^\perp}_{L_o^r}, \quad (1)$$

where we integrate product of L and bidirectional scattering distribution function ρ_s over a sphere of directions S to compute reflected radiance term L_o^r (we use projected solid angle measure $d\omega_i^\perp$). For simplicity, we now consider only surfaces. The extension of path guiding to volumes is described in Sec. 7.7.

As we incrementally construct paths vertex by vertex, we *bias* random decisions taken in the process to *guide* the paths towards important regions in the scene. In these regions, paths are very likely to make significant contributions to the image. We can *bias* (i.e. change probability of) multiple decisions along each path, like choosing direction after each scattering event, free path sampling in volumes, absorption or choosing a light source for connection. Note, that biasing in this sense does not introduce bias (systematic error) in the image but results in the expected (correct) solution.

7.3 Previous Work

In computer graphics, path guiding was pioneered by works of Jensen (Jen95) and Lafortune and Willems (LW95). Both works differ in representation used for the light

field approximation. While the former used regular histograms in the spherical domain reconstructed from photons, the latter applied 5D tree to discretize the light field. Another crucial difference is the transport direction of samples used for learning as described in (VHH⁺19), Sec. 7.10.

These were followed by works of Hey and Purgathofer (HP02) and Bashford-Rogers et al. (BRDC12) each using yet another representation namely hemispherical footprints (i.e. cones with varying radii centered around sample directions) and mixtures of cosine distributions respectively.

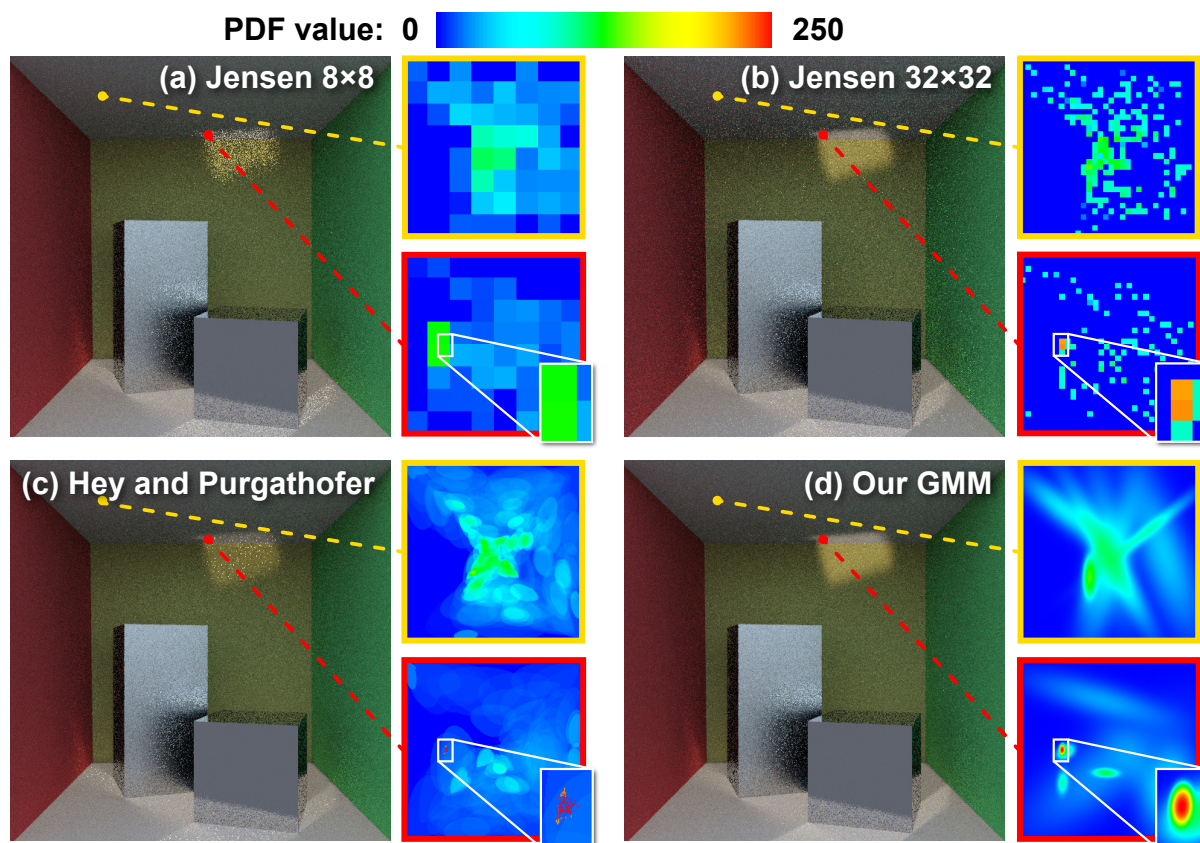


Figure 4: Demonstration of the superior flexibility of the parametric Gaussian mixture model (GMM) over previously used models. Four renderings of a Cornell box scene with diffuse walls and two glossy blocks lit by the sun are rendered by guided path tracing using Jensen’s (Jen95) method with the histogram resolution of 8×8 (a) and 32×32 (b), Hey and Purgathofer’s (HP02) hemispherical footprints (c), and with GMM (d). The distributions trained at two selected locations in the scene are also visualized. One distribution contains low-frequency illumination while the other contains a sharp directional peak caused by a reflection of the sun. The illustration is borrowed from work of (VKŠ⁺14).

7.4 On-line Learning of Incident Radiance

To guide paths efficiently in the production scenes we need to deal with high frequencies in the light field which are typically caused by combination of small light sources, complex geometry with small openings, and also by materials with various roughness ranging from almost smooth dielectrics and metals to almost diffuse materials like wood. Previous methods were not able to handle high frequencies and/or

suffered from high memory footprint that limited achievable precision and thus, in turn, efficiency of guiding (see Fig. 4).

To remedy this, we formulated guiding as learning uncertainty (VKŠ⁺14) which is a central problem of machine learning. Namely, we used Bayesian treatment where each observed sample is considered as evidence forming our prior beliefs about unknown distributions.

As a representation, we chose parametric mixtures of Gaussians representing angular pdfs proportional to incident radiance. For learning, we used combination of batch and so called stepwise Expectation-Maximization (EM) algorithms. The former for training initial pdfs if we had sufficient amount of samples which enables faster convergence. The latter allows progressive training from a stream of samples and thus avoiding problems with limited memory (Fig. 7.2). These mixtures are cached within the scene in the irradiance-caching-like lazy scheme. If a distribution is not available within certain distance from valid cache records, we train and insert a new one.

When sampling a new path, we need to sample a direction at every scattering event (that is interaction with surface or volume). We need to decide whether we sample according to BSDF or our pdfs proportional to incident radiance. We mix both using multiple importance sampling (MIS). Note that for Dirac or almost Dirac BSDFs, samples from our pdf will yield zero (or almost zero) contribution, thus we can keep sampling only BSDFs not to waste samples. If we need sample from our pdf, given the position x of the event, we search for the closest distribution in our normal aware the cache.

This work revealed an interesting result with the respect of chasing the “one” sampling algorithm. We observed that guiding unidirectional path tracing with next event estimation within complex bidirectional estimators can make many merging and connecting sub-techniques almost redundant within MIS estimator. Consequences of this observation are briefly discussed in Sec. 7.10 and also in the course on path guiding in production (VHH⁺19), Sec. 7.9.

For further details, we refer the reader to the work of Vorba et al. (VKŠ⁺14).

7.5 Optimal Path Lengths: Guided Russian Roulette and Splitting

Path sampling can benefit from directional guiding as long as we can efficiently decide whether it is worth to continue tracing the path or we should rather terminate it and start tracing a new one from the camera. Traditionally, this decision called Russian roulette has been driven by albedo of surfaces or volumes.

However, scattering light many times in the scene before reaching the camera² can result in premature termination of paths that would significantly contribute to the image otherwise. As a result, their contribution turns into strong noise. This issue arises also in scenes with fully path-traced sub-surface scattering when path emerges from the object and is terminated at the next vertex without being able to connect to a light. On the other hand, albedo driven termination may result into spending too much time on tracing reflections between materials with high albedo (white walls, snow, blond hair, white fur, etc.) without actually finding a light source and scoring a contribution.

To remedy this, we introduced *guided Russian roulette and splitting* (also known as adjoint-driven Russian roulette and splitting) (VK16) which allows to optimize path termination by using global knowledge about the scene learned from previous samples

²Probability of surviving the whole paths is multiplication of survival probabilities at each vertex and thus it typically drops very fast with the path length.

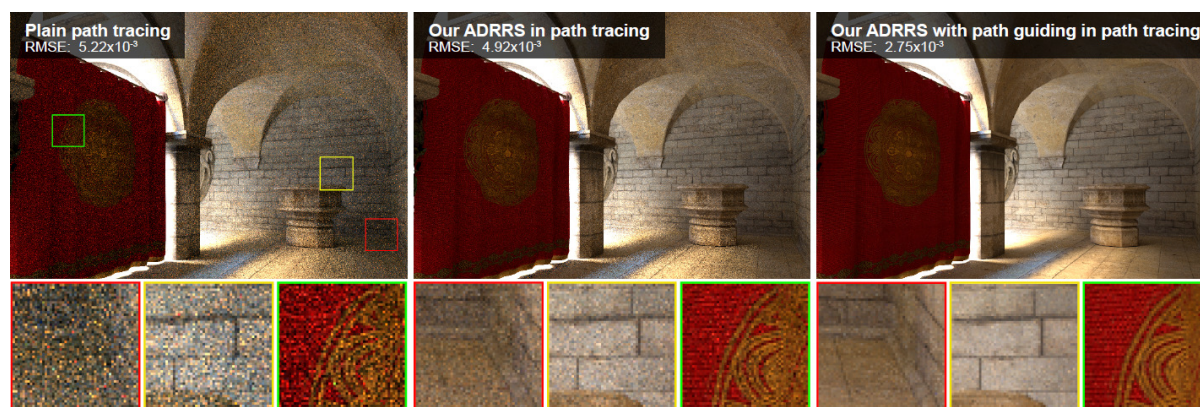


Figure 5: In scenes where light is scattered many times before reaching the camera, good importance sampling and thus noise reduction can be achieved by guided (adjoint-driven) Russian roulette and splitting (ADRRS). Using traditional albedo-driven Russian roulette in path tracing is sub-optimal under these conditions (left) because paths are either terminated too soon or time is wasted on sampling overly long paths. Using global knowledge about the scene clearly reduces noise in indirectly lit regions (middle). Directional path guiding can be naturally combined with ADRRS which results in synergic noise reduction (right). Image courtesy of (VK16).

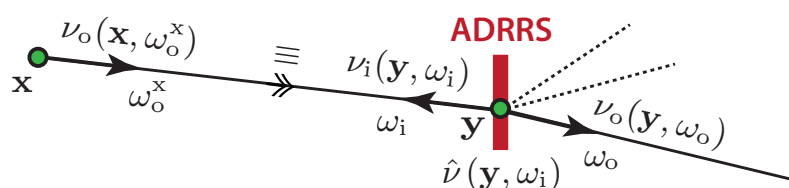


Figure 6: After we account for a particle's contribution from a collision at \mathbf{y} , we apply our adjoint-driven Russian roulette and splitting (ADRRS) aka GRRS to decide about the particle's termination/splitting. All potentially spawned particles at \mathbf{y} have weight $\hat{\nu}(\mathbf{y}, \omega_i)$ and are scattered and traced independently.

(see Fig. 7.5). Moreover, path guiding can use this knowledge to split the path in important regions which are, in turn, covered by more samples. Increased efficiency follows from the fact that path splitting amortizes the work spent on tracing the whole path up to the splitting point.

The more scattering events along the path the greater benefit the guided Russian roulette and splitting provides. This is a reason why it is so important for efficient volumetric transport where average path length is usually high. In Sec. 7.7, we describe this volumetric extension in more detail.

Termination and Splitting Rate. To determine survival probability/splitting rate

$$q(\mathbf{y}, \omega_i) = \frac{E[c(\mathbf{y}, \omega_i)]}{I} = \frac{\nu_i(\mathbf{y}, \omega_i)L_0^r(\mathbf{y}, \omega_i)}{I}, \quad (2)$$

guided Russian roulette and splitting (GRRS) compares expected contribution $E[c(\mathbf{y}, \omega_i)]$ of a current path at vertex \mathbf{y} coming from direction ω_i to the computed pixel value I (Fig. 6). The path contribution $c(\mathbf{y}, \omega_i)$ is a random variable associated with a path that has reached the point \mathbf{y} from the direction ω_i and has the weight $\nu_i(\mathbf{y}, \omega_i)$. The variable is distributed over all possible realizations of the path beyond \mathbf{y} , as shown in Fig. 7. For

example in path tracing, the outcome of c for one such specific realization is given by the path's contribution to the sum in the measurement estimator of Eq. (??). Note

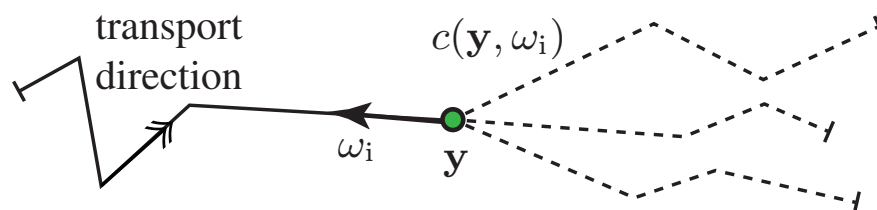


Figure 7: Realizations of the path contribution variable $c(\mathbf{y}, \omega_i)$ correspond to the different possible particle paths beyond \mathbf{y} .

that each particle sampled beyond \mathbf{y} is an unbiased estimate of reflected radiance $L_o^r(\mathbf{y}, \omega_i)$ which is defined by Eq. (1). Thus the expected contribution $E[c(\mathbf{y}, \omega_i)]$ is given by the product of the path weight ν_i and the outgoing reflected radiance L_o^r .

In practice, if we expect that the path could make the same or higher contribution compared to the current pixel value, we always keep tracing further. On the other hand, relatively low expectation results into high chance that the path will terminate. If q is higher than 1, we split the path into q independently traced paths.

To deal with non-integer values of q , we can either (1) take a ceiling of q and split into $n = \lceil q \rceil$ path or (2) use an *expected-value split* when we split into $n = \lfloor q \rfloor$ new particles with a probability $n + 1 - q$ or into $n + 1$ particles otherwise. In the latter, the new value of each split path is equal ν_i/q where ν_i is the MC value of path when it arrived at \mathbf{y} . Originally, we implemented the expected-value split (VK16), however, this approach needs a random number and is likely to introduce some variance. We haven't experimented with both approaches to compare them thoroughly because we supposed that the differences would be marginal overall.

Practical Consideration. Guided Russian roulette and splitting (GRRS) requires two kinds of estimate to work: (a) an estimate of the computed pixel value I and (b) estimates of incoming radiance at each scattering event (i.e. path vertex) along the sampled path. Essentially, these two quantities are compared at every scattering event to decide whether the path will be terminated, split, or will continue. It seems prohibitive that these quantities are not known up front, however, this technique can work with relatively crude estimates.

There are many options for computing the pixel estimates I . Originally, we used a pre-computation step to cache estimates of incident radiance in the scene and determined the pixel estimates I in a gathering step (VK16). However, this extends the time to first pixel and thus is not suitable for progressive rendering which immediately provides a preview of the computed image. As discussed in the course on path guiding in production (VK16), Sec. 7.10, the *forward* learning guiding methods (MGN17, SJHD18) are inherently progressive because they learn while the image is computed. To avoid increasing the time to first pixel due to GRRS in guiding methods without pre-computation, we use filtered current pixel estimates I which are updated on-line as the light transport simulation proceeds. Our filtering, which provides low-variance estimates, is implemented as a hierarchical sub-sampling of the image and the estimates are refined up to a pixel level as more paths are traced (VK16), Sec. 7.7.

Obviously one can use even more advanced de-noising methods and consider GPU support if this is available. It is only important to achieve the pixel estimates fast without claiming too much of computational resources needed for the rest of the light



Figure 8: In scenes with glossy surfaces, product importance sampling (HEV⁺16) (right) yields higher sample quality (512 samples) over pure radiance based path guiding (VKŠ⁺14) (left).

transport simulation. An important insight is, that pixel estimates I do not need to be absolutely precise, yet GRRS can provide significant time savings.

As we mentioned above, we also need to have estimates of incident radiance at every point in the scene. While these can also be only approximations of true values, practical implementation should consider to start using GRRS only when the estimates are based on sufficient number of samples and have variance below a reasonable threshold (VK16).

Importance Sampling and Zero-Variance. In this work, we also study relation of ADRRS to zero-variance sampling pdf and show, that it effectively works as rejection sampling/splitting that reacts at each vertex locally on previously sampled decisions (VK16), Sec. 7. In other words, it implicitly considers the true zero-variance pdf (given our approximations are perfect) and compares it to the current sampling pdf. The termination/splitting rate q is determined so that it compensates for differences in respective pdfs.

7.6 Glossy BSDFs: Product Sampling

Traditionally, practical simulators have used only local importance sampling techniques like for example BSDF sampling, which is often sub-optimal. Path guiding (VKŠ⁺14) (VK16) described so far is based on sampling proportionally to incident illumination and BSDFs separately while mixing two sets of samples by MIS. This approach importance sample only parts of the integrand in reflected radiance L_o^r (Eq. (1)) and thus is sub-optimal for non-diffuse materials. We explored importance sampling of the full product in the work of Herholz et al. (HEV⁺16).

Mixtures of Gaussian distributions used to represent incident illumination (VKŠ⁺14) allow analytical calculation of their product. To utilize this fact, we have to fit Gaussian mixtures (GMM) to BSDFs in our scene. Subsequently, when we need to sample a direction ω at the scattering event x , we look-up GMM distribution of incident illumination at that position and also a GMM fit to BSDFs at given position. These fits are pre-calculated for all BSDFs in the scene and stored in a cache for discretized view directions. Next we calculate the product and sample the direction ω from resulting GMM.

Since now our product distribution is also aware of BSDFs, it is worth noting, that now we can safely take 90% of all samples from this product and only 10% from BSDF for MIS combination. Note that before we had to stay rather conservative and took 50% from

incident illumination distribution and 50% from BSDF (VKŠ⁺14). This further increases efficiency of path guiding.

7.7 Path Guiding in Volumes

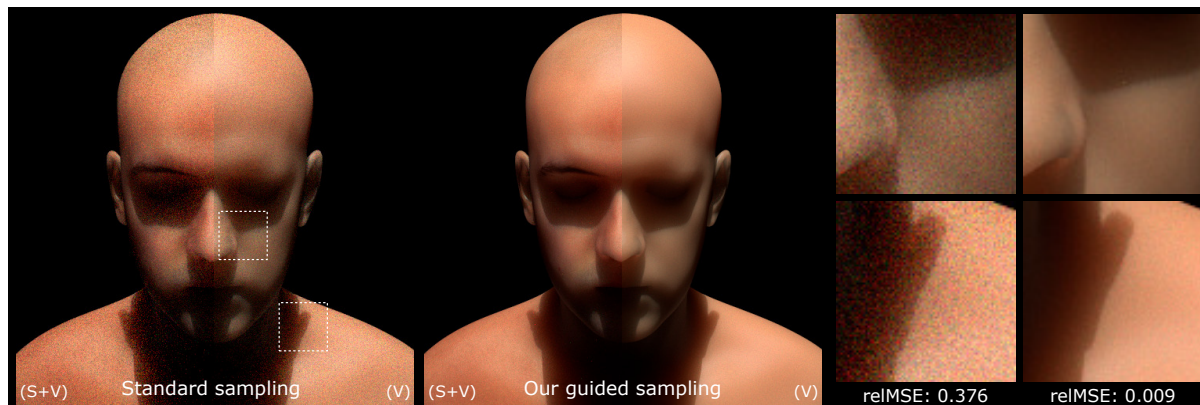


Figure 9: Path guiding in volumes (HZE⁺19a) importance sample all integrated terms based on zero-variance random walk theory. This provides superior sample quality over standard importance sampling of only transmittance and phase function. Images show a scene with optically dense medium rendered for 30 minutes where (V) is only volumetric transport and (S+V) is also with surfaces.

Jaroslav Křivánek also supervised extension of path guiding and adjoint-driven Russian roulette and splitting into participating media (HZE⁺19b). This work stresses the importance of guiding multiple decisions along the path based on the zero-variance theory applied to participating media. Such importance sampling of all terms of volumetric rendering equation is efficient even for honestly traced sub-surface scattering 9.

To this end, they apply mixtures of von Mises-Fisher distributions (vMFM) to represent incident and in-scattered radiance learned in a similar process to Vorba et al. (VKŠ⁺14). Von Mises-Fisher distribution is isotropic and mathematically very similar to Gaussian distribution but it is conveniently defined over sphere of directions. These properties allow analytic calculation of a product and convolution of a phase function and learned incident radiance cached within the scene which is essential for the efficient volume guiding. Note, that in practice, this requires fitting of phase functions by vMFM which can be pre-calculated.

7.8 Variance-Aware Path Guiding

Instead of pursuing pure zero-variance path guiding, that is learning pdfs proportional to radiance or its product with BSDFs, Rath et al. (RGH⁺20) explores learning of theoretically optimal densities that account for variance in nested estimators. This compensates for the fact that, in practice, we never arrive at the exact zero-variance pdfs due to used approximations or low numbers of observed samples or simply due to omitting some terms from importance sampling (e.g. when not using product sampling).

7.9 Subsequent Research

Jaroslav has supervised several papers on path guiding and his work has inspired others to further explore this avenue.

Müller et al. (MGN17) proposed using SD-trees to represent directional pdfs instead of parametric mixtures and also came up with an efficient lock-less caching scheme. They recently proposed several improvements (VHH⁺19), Sec. 10, including spatial and directional filtering of learned samples as well as optimized allocation of samples between BSDF/incident radiance sampling. Diolatzis et al. (DGJ⁺20) extends SD-trees by product sampling using linearly transformed cosines. This approach has been also extended to participating media (DWWH20).

Just recently, Ruppert et al. (RHL20) have reported significant improvements with parametric mixtures (namely with vMFM) thanks to optimized and improved learning algorithms and to exploiting advantage of parametric properties of vMFM enabling so called parallax-aware warping. Further, akin to Müller et al. (MGN17), they enable *forward* learning, that is learning from sampled camera paths rather than from photons, which makes it more practical for production (VHH⁺19). They even reported interesting improvements in quality on the whole range of scenes comparing their work to the SD-tree based approach.

Dahm and Keller (DK18) formulated path guiding as Q-learning rather than by means of zero-variance theory. Such reinforcement learning enables using cached approximations instead of contributions of full paths which increases the number of non-zero samples. This, in turn, improves learning at early stages. Pantaleoni (Pan20) proposed using GMMs in the context of Q-learning and path space filtering (KDB14, BFK18) while targeting real-time setting.

While so far described path guiding methods can be considered local, i.e. paths are guided by a chain of local decisions using marginalized pdfs at every vertex, Simon et al. (SJHD18) explored guiding with complete light transport paths. They retain a set of outlier guide paths for guiding subsequently sampled paths which includes all aspects of high-dimensional path space as opposed to local methods. This comes with both advantages and drawbacks (VHH⁺19), Sec. 12.

Path guiding has been explored even within the context of deep learning which enables harnessing dedicated GPU hardware (MMR⁺19). Recent works also explored the idea of using control variates next to path guiding in various contexts (Pan20, MRNK20).

7.10 Industry Impact

Path guiding has been implemented in production renders of several VFX studios. Namely at Weta Digital's Manuka and also Hyperion, the renderer used at Walt Disney Animation Studios (VHH⁺19). In Manuka, early implementation of path guiding has been available since 2014 and has been maintained and developed up until now. From production point of view, path guiding is relatively appealing since it minimizes the need of complex and rather cumbersome³ bidirectional estimators (VHH⁺19), Sec. 7.9. Also it can greatly reduce rendering times in many scenarios.

7.11 Future Works

Some opened problems of path guiding were listed in the course on production path guiding (VHH⁺19). It is worth noting, that some of them, like faster learning, parallax-problems, or second moment guiding, has already been addressed in recent works (RGH⁺20, RHL20). From practical point of view, it makes sense to further pursue the goal of seamless guiding implementation that would smoothly fit into ecosystem of path sampling

³It is not easy to maintain while adding new features into the renderer. Some production features that make for examples materials depend on the eye-path prefix are rather challenging to implement within bidirectional context.

techniques used in a production system (FHH⁺19) without potential risk of excessive overhead in simple scenes.

7.12 Conclusion

Jaroslav has left clearly visible footprint in the field of light transport simulation. Whichever direction the future research will go, it is certain that Jaroslav will be greatly missed on this path towards the “one” sampling light transport algorithm.

8 Direct Lighting



SIGGRAPH THINK
BEYOND

2020 19-23 JULY WASHINGTON DC

ADVANCES IN MONTE-CARLO
RENDERING:
THE LEGACY OF JAROSLAV KŘIVÁNEK

Direct Lighting

ADVANCES IN MONTE CARLO RENDERING: THE
LEGACY OF JAROSLAV KŘIVÁNEK

- Direct and indirect illumination calculations are two important components of any physically-based renderer. While the indirect component has been traditionally considered a more complex problem and has been studied in many research works, Jaroslav acknowledged that improving the efficiency of direct illumination could have a substantial impact on the overall rendering performance, especially with complex visibility and in the presence of large numbers of light sources.
- In this part, we will cover direct illumination sampling based on online learning of light selection probability distributions. We will show how to formulate the learning process as Bayesian regression to prevent over-fitting and ensure robustness even in the early stages of computation.

- Any guiding needs **radiance approximations**
- How to learn them **reliably?**
- Jaroslav's proposition:
(Online, Bayesian) **Machine learning**
[Vorba et al. 2014, Vévoda et al. 2018]

- However, we would like to go beyond the specific problem of direct illumination and really focus on the fundamental methodology we used to address this problem, since it is applicable in a wider context in guiding and adaptive sampling.
- As we have seen in the previous part of this course, any guided sampling requires some information about the radiance distribution in the scene. The radiance is not available upfront, so we must resort to approximations. The approximations can be obtained by learning from the Monte Carlo samples themselves – but these provide noisy and unreliable estimates, especially early in the computation. So the general question we are striving to address is how to obtain reliable information from unreliable, noisy samples.
- Jaroslav's proposition was that machine learning, and specifically Bayesian statistics, provides some excellent tools to tackle exactly this problem. This was the underlying idea in two guiding works Jaroslav supervised: the direction guiding [Vorba et al. 2014] mentioned earlier in this course, and the direct illumination guiding [Vévoda et al. 2018] that we will present now.

Bayesian online regression for adaptive direct illumination sampling

Petr Vévoda, Ivo Kondapaneni, and Jaroslav Křivánek

Chaos Czech, a.s.
Charles University, Prague



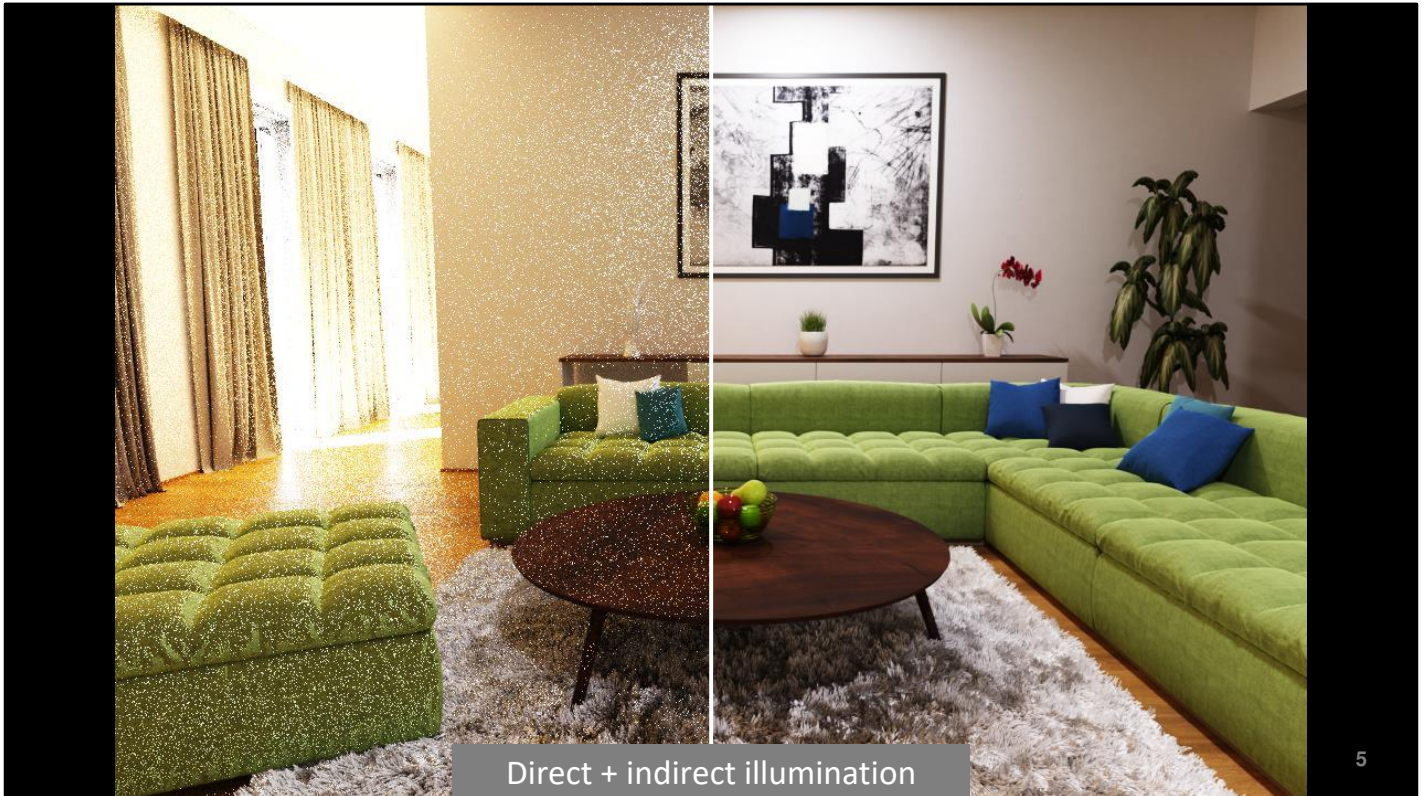
Computer
Graphics
Charles
University

- This method was first presented at Siggraph 2018 under the name Bayesian online regression for adaptive direct illumination sampling.
- It was a result of collaboration with Chaos Czech, the developer of Corona Renderer, and became the default solution for direct illumination calculation in Corona Renderer version 3.

MOTIVATION

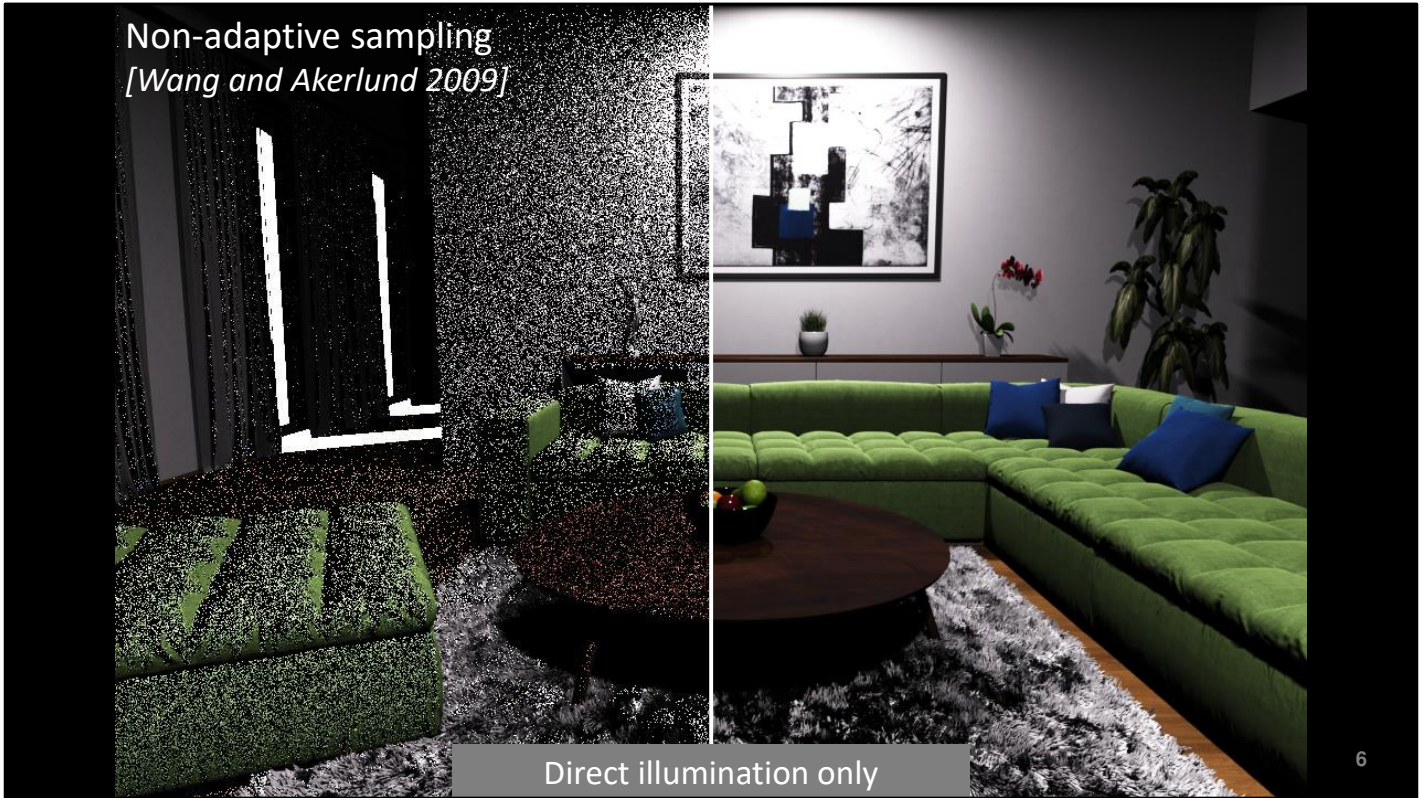
ADVANCES IN MONTE CARLO RENDERING: THE LEGACY OF JAROSLAV KŘIVÁNEK

- Let us first motivate this work.

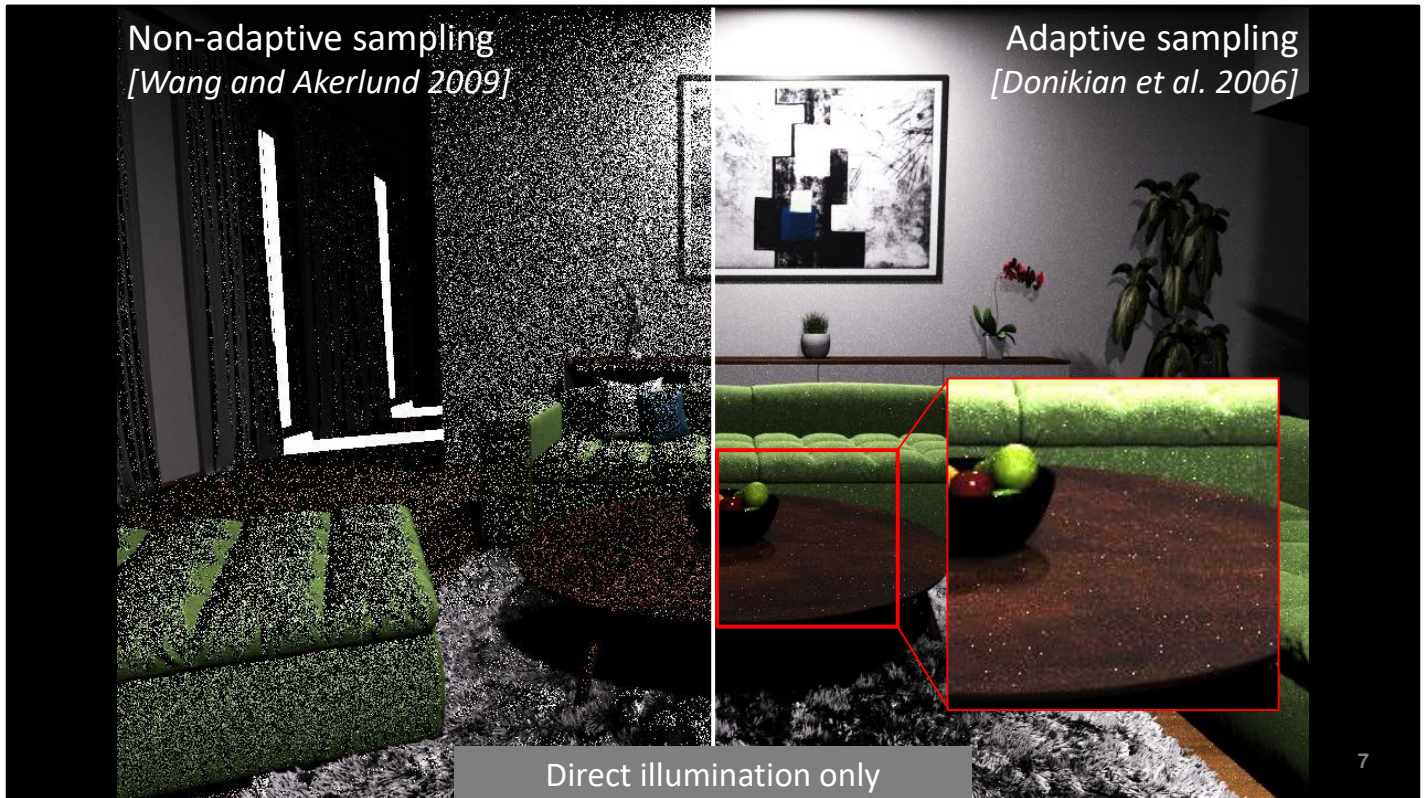


5

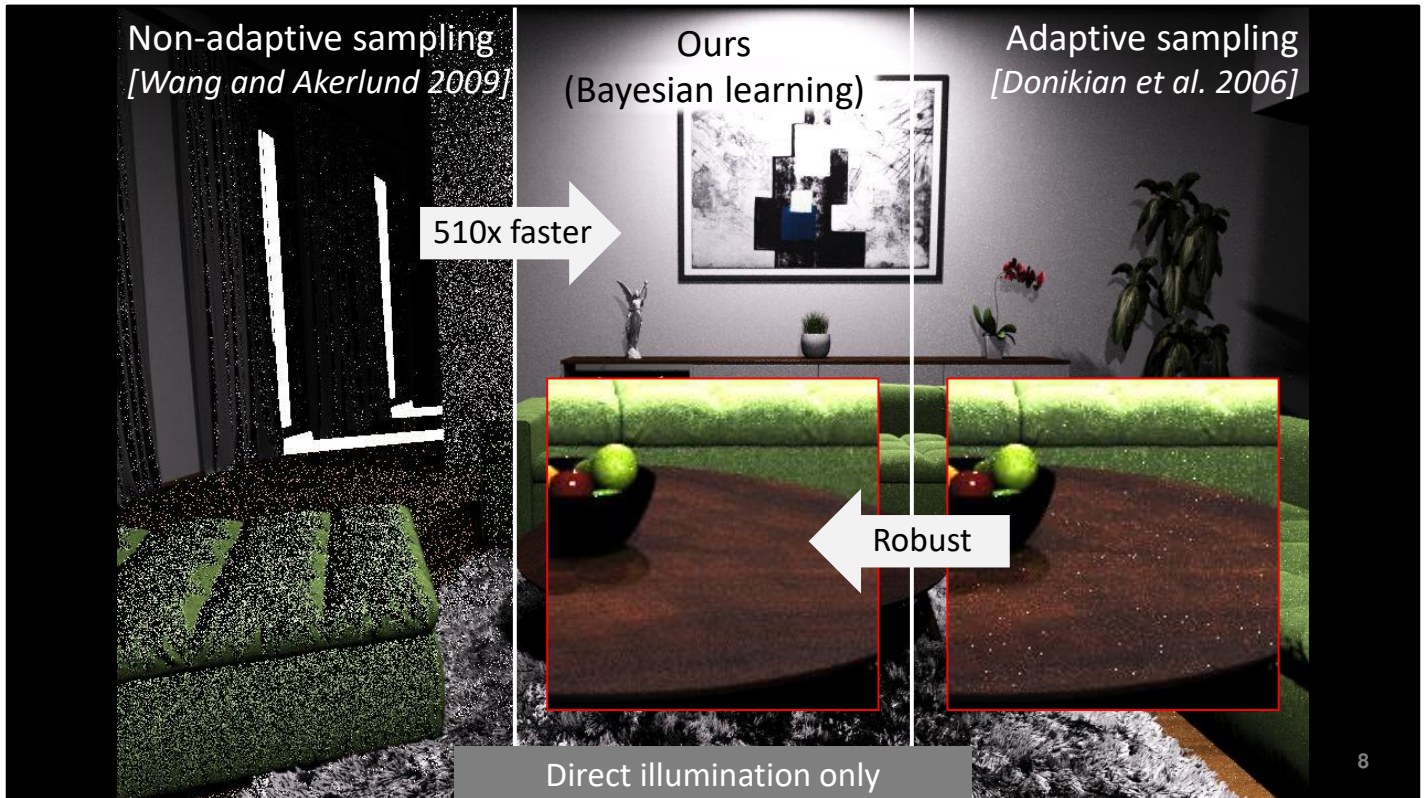
- Monte Carlo rendering algorithms are currently getting more and more popular, but they suffer from noise.
- This image shows an example of a scene rendered using Monte Carlo. During rendering the noise shown on the left slowly goes away until the noise-free image shown on the right is obtained. Generally, it can take up to several hours for the noise to disappear.
- Illumination of any point in a scene can be split into two components:
 - Direct = illumination coming directly from light sources
 - Indirect = illumination coming from light sources after at least one interaction with the scene, e.g. after reflection from a different part of the scene
- Traditionally, the indirect component has been considered as the main source of noise, and it has been subject to lot of research. But in this scene, as well as many other production scenes, it is the direct component which causes the trouble.



- We can clearly see the problem in this image showing direct illumination only.
- An example of a non-adaptive direct illumination sampling method is shown on the left. It samples lights proportionally to a conservative estimate of their unconcluded contribution. It struggles to work efficiently in this scene, because it wastes a lot of samples on the strong but almost completely occluded sun (sunlight enters the scene only through narrow gaps between the curtains).



- One possible solution is to use previous **adaptive** methods which try to improve sampling based on past samples. An example is shown on the right.
- However, while they can decrease the amount of noise significantly, they can also introduce various artifacts and spiky noise because they are based on adhoc solutions and they tend to overfit to the input noisy data.
- This lack of **robustness** is a consequence of adhoc solutions to crucial questions in adaptive sampling:
 - When is it safe to rely on the noisy samples?
 - How should the noisy samples be combined with any previous, a priori knowledge?



- In 2018 we proposed a solid theoretical framework based on Bayesian learning that addressed the above questions, and thanks to that enabled robust adaptive sampling in rendering.
- In this scene, our solution is more than 500 times faster than the non-adaptive solution and we can achieve much better robustness than the previous adaptive sampling method by Donikian et al. [2006].
- Moreover, the Bayesian framework we present here is not limited to adaptive direct illumination. Other guiding / adaptive sampling methods can benefit from it as well.



- In the context of Monte Carlo simulation there is a lot of work related to ours.

- General Monte Carlo
 - Vegas algorithm
 - *[Lepage 1980]*
 - Population MC
 - *[Cappé et al. 2004, ...]*

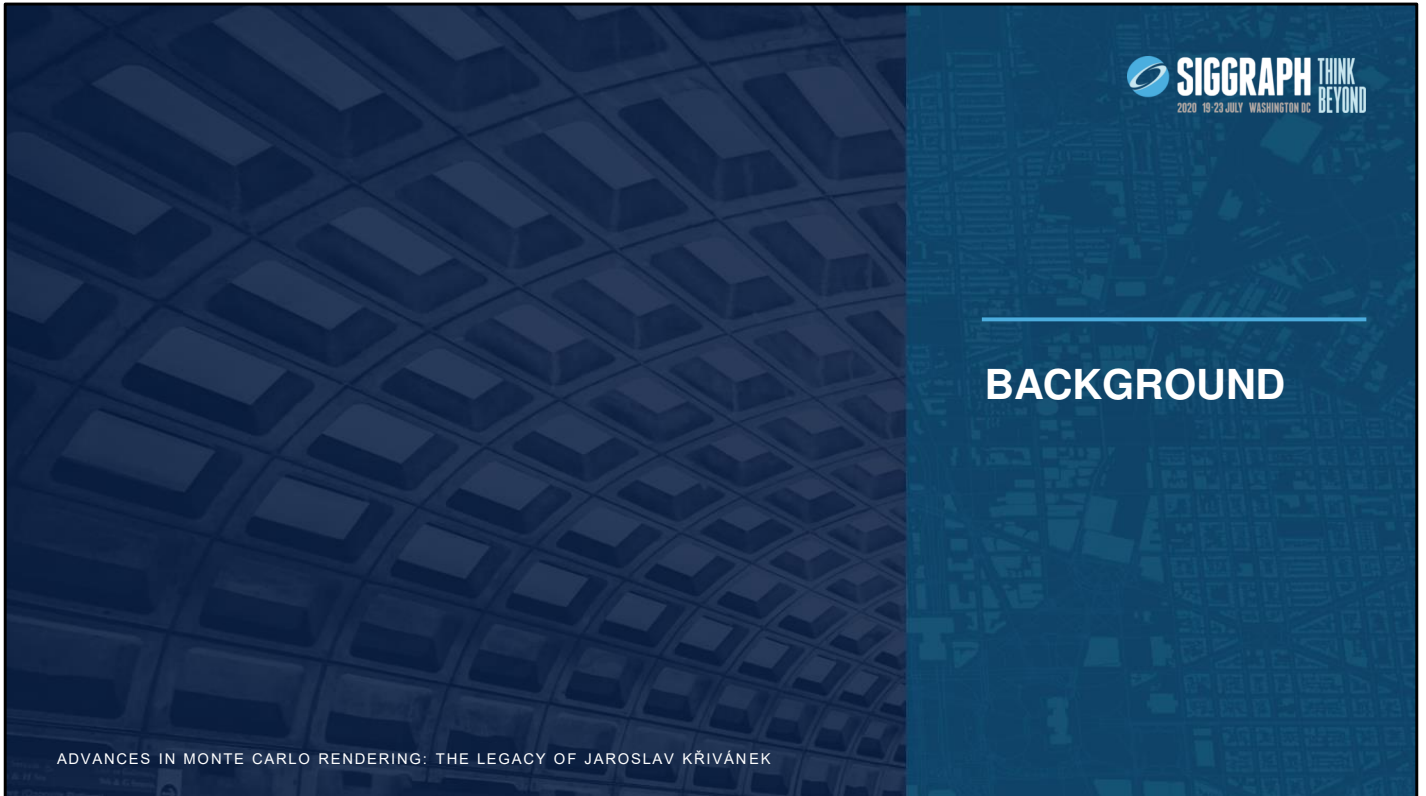
- Adaptivity in Monte Carlo simulation is not a new concept.
- A classic adaptive algorithm for general Monte Carlo estimation is the Vegas algorithm by Lepage [1980], which works by histogramming (stratifying) the integrand and using these histograms for better sampling in subsequent steps.
- Another example is Population Monte Carlo [Cappé et al. 2004], which uses a population of simulation particles and tracks how well they sample the integrand. Based on that, they keep the best individuals for subsequent sampling rounds.

- Rendering
 - Image sampling
 - *[Mitchell 1987, ...]*
 - Indirect illumination (path guiding)
 - *[Dutr  and Willems 1995, Jensen 1995, Lafortune et al. 1995, ...]*
 - *[Vorba et al. 2014, Muller et al. 2017]*
 - Direct illumination
 - *[Shirley et al. 1996, Donikian et al. 2006, Wang and Akerlund 2009]*

- A lot of adaptive sampling work exists in rendering.
- Of the many works, we mention Mitchell [1987] which deals with allocation of samples into image parts with high-frequency content.
- In adaptive indirect illumination computation, some of the early works were done by Dutr  and Willems [1995], where the authors adaptively traced particles from lights; Jensen [1995] who used photon maps to construct path sampling distributions (this is probably the first work in graphics on direction guiding); and Lafortune et al. [1995] who applied a Vegas-like approach in Monte Carlo light transport simulation.
- The ongoing renewed interest in path guiding / adaptive path space sampling could be attributed to work of Vorba et al. [2014] previously described in this course. The direction sampling algorithm by Muller et al. [2017] is a version of the Vegas-style algorithm by Lafortune from 1995.
- As for direct illumination, we mention the pioneering work by Shirley et al. [1996] who adaptively classified lights into important and unimportant ones.
- The most closely related is the work by Donikian et al. [2006] that we describe in more detail later.
- Wang et al. [2009] sample lights adaptively based on surface reflectance and estimates of lights' contributions.
- None of these works deal with a problem of determining when and how to incorporate all the information into a robust, reliable 'trained' sampling distributions.

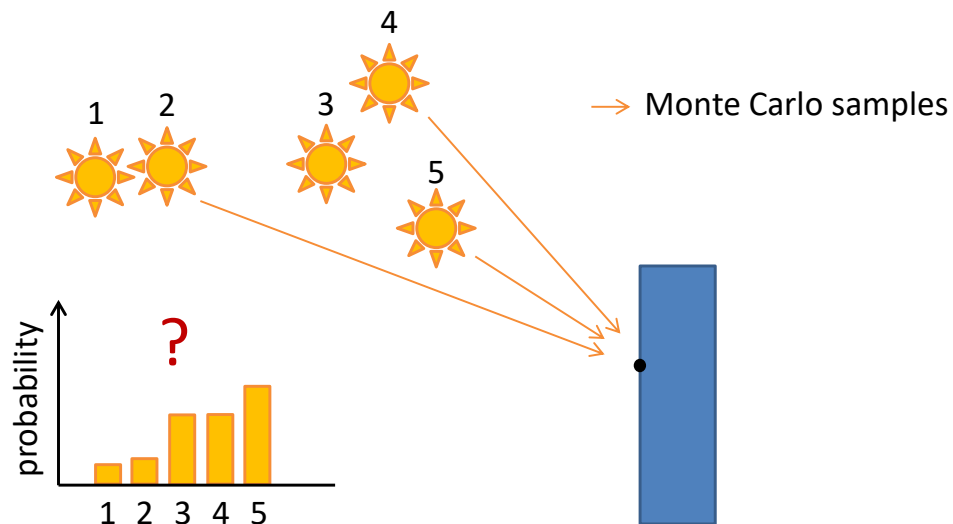
- Filtering
 - NonLocal Bayes
 - *[Boughida and Boubekeur 2017]*
- Global illumination
 - Bayesian Monte Carlo
 - *[Brouilat et al. 2009, Marques et al. 2013]*
 - Path guiding
 - *[Vorba et al. 2014]*

- On the other hand, Bayesian methodology has so far not been widely applied in rendering.
- This list mentions the few Bayesian methods in image filtering and global illumination.



- Now we give you some background related to the direct illumination problem.

DIRECT ILLUMINATION SAMPLING

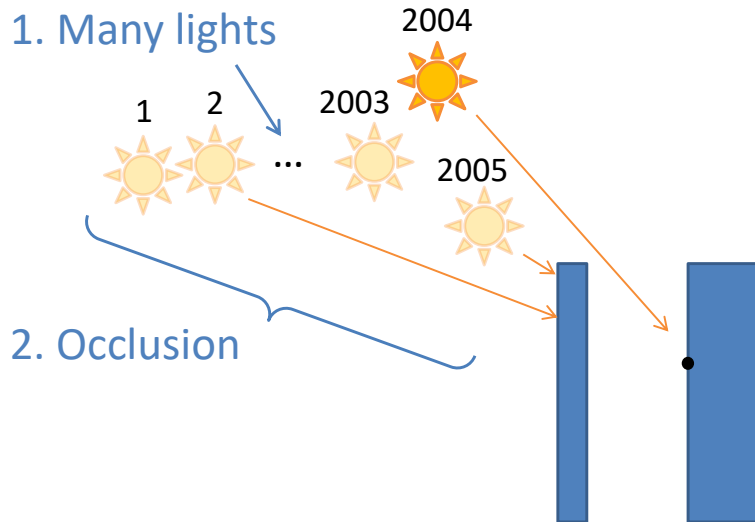


ADVANCES IN MONTE CARLO RENDERING: THE LEGACY OF JAROSLAV KŘIVÁNEK

14

- Consider a scene with several light sources (orange stars) and geometry (blue block). The goal is to calculate the **direct** contribution of all lights towards a given point in the scene (black point).
- In Monte Carlo rendering this is achieved by randomly sampling points on the surface of all lights and accumulating contributions from these samples (orange arrows). Efficiency of this approach depends mainly on the probability distribution used for drawing the samples: the closer the distribution matches the actual contribution from points on lights, the less noise in the resulting image.
- Direct illumination sampling and the corresponding distribution can be broken down to two stages:
 - A) pick a light at random according to a discrete distribution over lights
 - B) randomly sample a position on the chosen light according to a continuous distribution over the light surface
- While much previous work has addressed optimal choices in step B, we focus solely on step A: finding the best discrete distribution over lights (orange bars). That is a complex task mainly for two reasons.

PROBLEMS

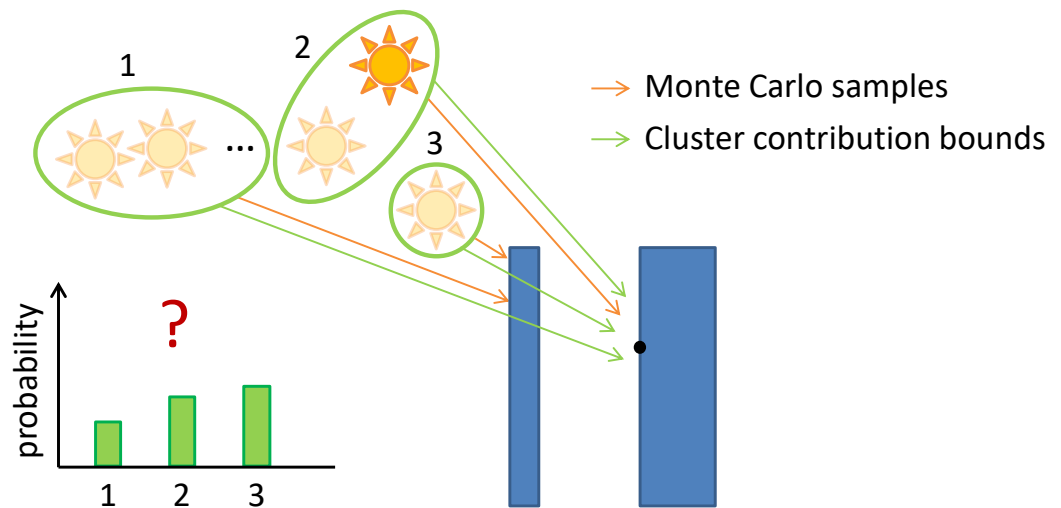


ADVANCES IN MONTE CARLO RENDERING: THE LEGACY OF JAROSLAV KŘIVÁNEK

15

- First, there is a problem with light count. As a construction of the sampling distribution scales linearly with the number of lights (for each light the probability has to be computed, the sampling distribution has to be constructed, turned to a cumulative distribution to facilitate the sampling, and properly normalized), it becomes computationally expensive in scenes with many light sources. This can pose a significant limitation as even thousands of lights are sometimes used in practice (for example in a city at night).
- Second problem is the highly uneven light contributions which are difficult to predict. In particular, occlusion of a light because of other scene geometry is usually not known in advance and can have substantial impact on the light contribution and thus on the sampling distribution we strive to find.

1. MANY LIGHTS → CLUSTERING

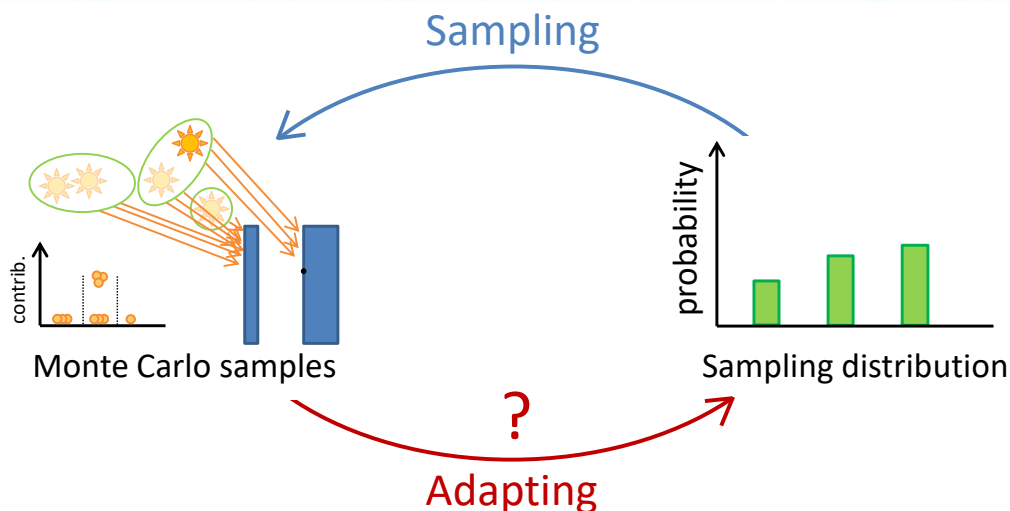


ADVANCES IN MONTE CARLO RENDERING: THE LEGACY OF JAROSLAV KŘIVÁNEK

16

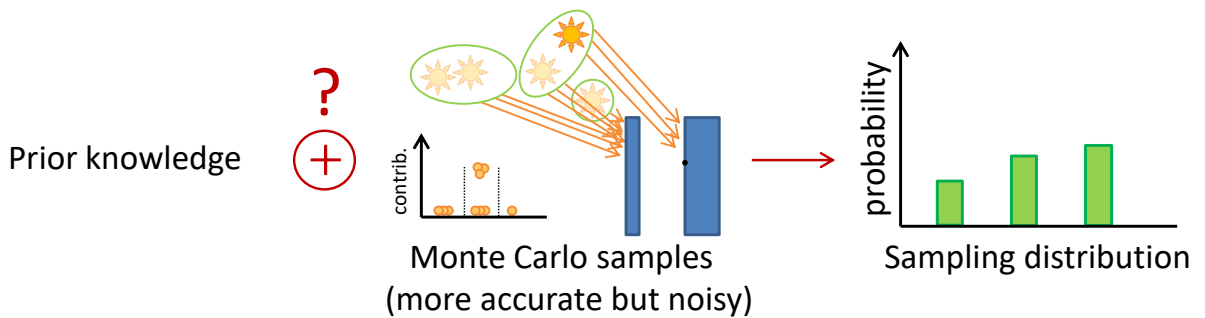
- In order to improve scalability of our method we incorporate clustering of lights using the Lightcuts method [Walter et al. 2005]. Before the rendering starts, we pre-cluster all lights. Then during rendering, we always choose clustering (green ellipses) according to conservative contribution bounds (green arrows) such that all lights in a cluster have similar (approximated, unoccluded) contribution to a given scene point.
- Similarly to work of Wang and Akerlund [2009] we then sample the clusters, i.e. instead of a discrete distribution over all lights we now seek the best discrete distribution over clusters (green bars). Lights within a cluster are then picked randomly based on their flux.
- This way we can greatly reduce the size of the constructed distribution, we can even limit its maximum size without omitting any light simply by making the clusters larger. While it does not solve the second problem (occlusion), without the clustering it would be pointless to even look for a solution as any non-trivial construction of the sampling distribution would become computationally infeasible in presence of many lights.

2. OCCLUSION → ADAPTIVITY



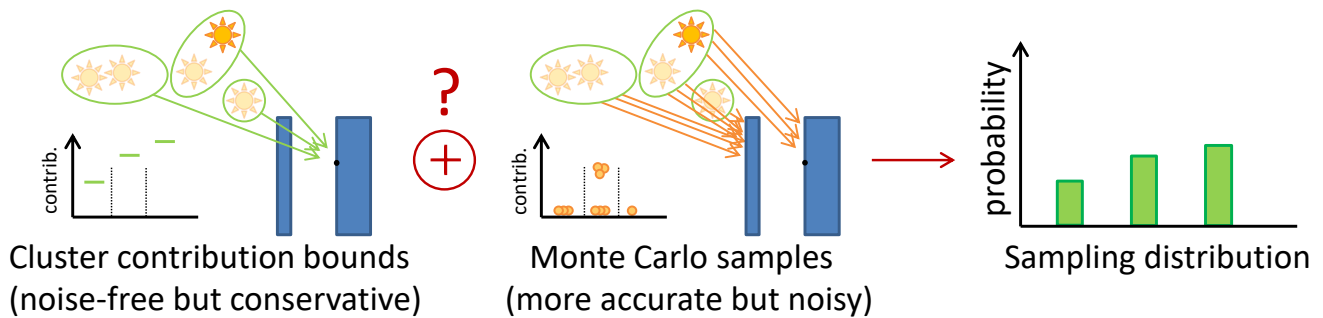
- Complex visibility in a scene can cause many light clusters to be fully or partially occluded. This is usually very hard to find out in advance, before samples are drawn. Instead, we can first draw the samples and adapt the sampling distribution afterwards. This is typically an iterative process. We start with some initial distribution and draw first samples (orange arrows). We compute their contributions (orange points) and recompute the sampling distribution accordingly. We then use this updated distribution for drawing next samples and repeat the process. This is the basic idea behind all adaptive methods.
- However, there is an open question – how exactly should be the sampling distribution computed based on the samples? The problem is that Monte Carlo samples are noisy, they provide the correct answer only after averaging many of them. Therefore, it is not safe to rely on them in early stages of rendering. Doing so may result in image artefacts.
- For example, one way of adapting the distribution is proposed by Donikian et al. [2006]. They gather statistics from samples about true cluster contributions in screen space for each pixel. But the per-pixel statistics are too noisy, so they additionally gather average statistics over entire blocks of pixels. Then they mix both the per-pixel and per-block statistics to obtain the final sampling distributions: early on, the blocks are given more weight and as more samples arrive, the per-pixel statistics are given more weight. But the mixing is done in an ad-hoc way, which often results in overfitting of the sampling distribution to the samples, and consequently produces image artifacts.

ADAPTIVITY WITH PRIOR



- An important part of an adaptive solution is incorporation of prior knowledge which expresses our initial belief about the sampling distribution. It serves as a starting point which gets updated as more Monte Carlo samples are gathered. However, how exactly should be the prior knowledge combined with samples is another open question.
- For example, Donikian et al. assume that all clusters are initially equally probable. Therefore, they additionally mix their distribution from per-pixel and per-block statistics with a uniform distribution in an ad-hoc way.
- In our case, we take advantage of the cluster contribution bounds provided by the clustering and use them as the prior knowledge. We seek a principled way how to combine them with the Monte Carlo samples.

PROBLEM SUMMARY



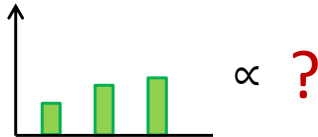
- Let us now summarize the problem at hand.
- Our goal is to compute direct illumination by means of Monte Carlo sampling, and for that we need to find **optimal** discrete sampling distribution over clusters.
- We choose to use **adaptive sampling** so that we can deal with (partial) light occlusion, but we strive for a **robust, 100% reliable solution**, even in complex edge cases.
- We have two kinds of information at our disposal:
 - First, the bounds of cluster contribution towards a point, which are conservative and noise-free. These can be computed on the fly, so they are available for the start. As such, they can serve as our prior belief about the clusters' real contribution to any given scene point.
 - Second, the Monte Carlo samples of direct illumination (i.e. clusters' real contribution). These are noisy at the start, but as more samples are taken, their average eventually converges to the correct answer - over time, they become more accurate than the conservative bounds.
- We propose a well-founded approach to combining these two sources of information in a robust way based on Bayesian inference, as we explain shortly.

OUR APPROACH

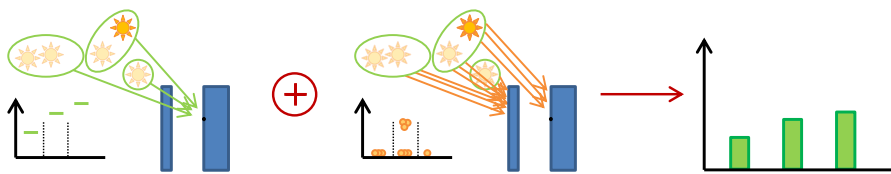
ADVANCES IN MONTE CARLO RENDERING: THE LEGACY OF JAROSLAV KŘIVÁNEK

- Let us now introduce our approach.

- **Optimal sampling distribution**



- **Adaptive sampling by Bayesian inference**

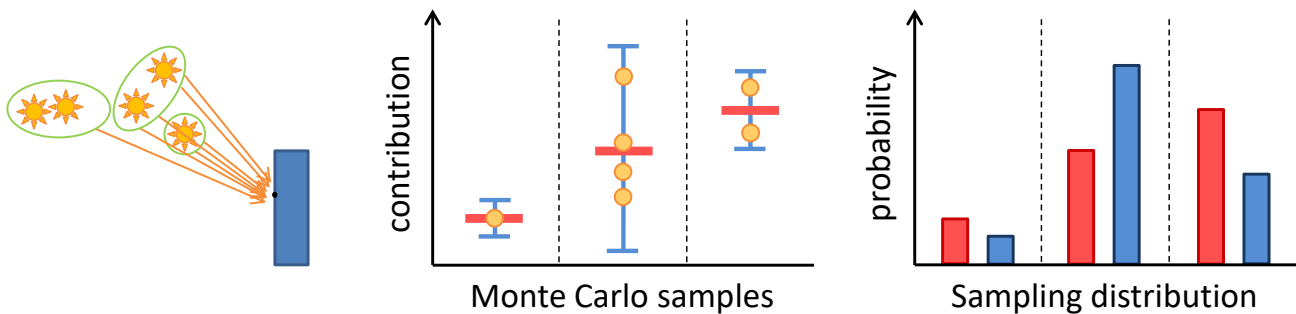


- We make two main contributions:
 - First, we show what should be the optimal discrete probability distribution for choosing the cluster, given the statistics of cluster contributions.
 - Our second and main contribution is the use of Bayesian inference to learn these sampling distributions. This gives us a robust solution and allows us to combine Monte Carlo samples with cluster contribution bounds in a principled manner.
- We start with the first point.

OPTIMAL SAMPLING DISTRIBUTION

usually: $P \propto \text{mean}$

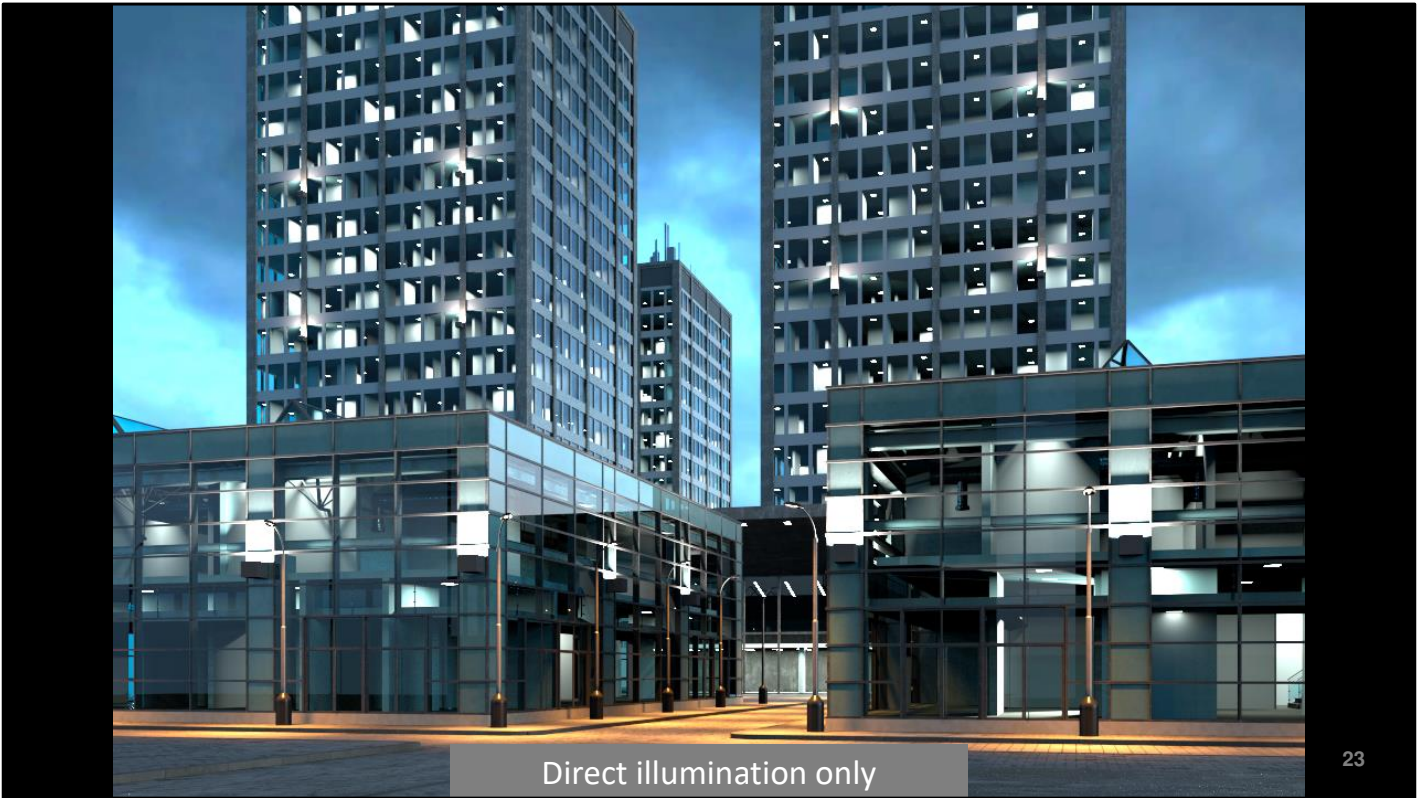
optimal: $P \propto \sqrt{\text{mean}^2 + \text{variance}}$



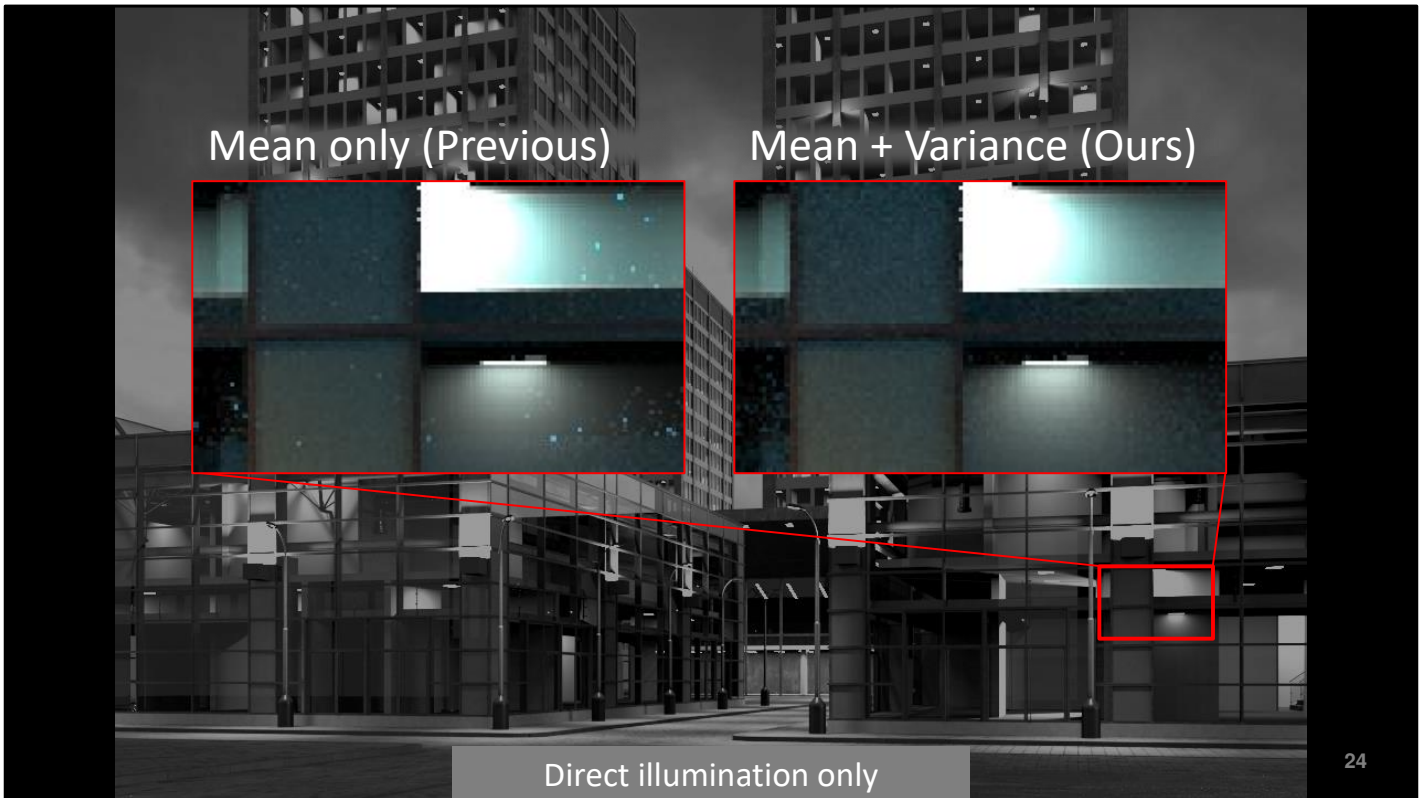
ADVANCES IN MONTE CARLO RENDERING: THE LEGACY OF JAROSLAV KŘIVÁNEK

22

- What is the discrete probability distribution for choosing a cluster at random that provides the lowest possible variance?
- The usual approach in Monte Carlo sampling would be to make the clusters' sampling probabilities proportional to their mean contribution to a given point. We showed in the paper that this traditional choice may in fact be far from optimal, and we derived the optimal solution.
- It is important to realize that once we pick a cluster, another random decision follows, which selects a particular light and a sample location on the light. The contributions from these sample locations will generally vary: usually the larger the cluster and/or the sharper the surface BRDF, the more variance in the individual sample contributions. Since this variance eventually creeps into the overall direct illumination estimator over all clusters, we need to reduce it - by allocating more samples to those clusters that yields highly varying contributions.
- The final optimal sampling should therefore take into account both the mean contribution but also the variance of the contributions for each cluster. The specific formula is given on the slide.
- The figure on the left shows an example of three clusters and their samples (orange arrows). Contribution of these samples is plotted in the middle figure (orange points) together with their mean (red thick line) and variance (blue interval). The usual sampling distribution (red bars) and the optimal one (blue bars) are shown on the right. Note that the second cluster gets much higher probability when its high variance is taken into account.



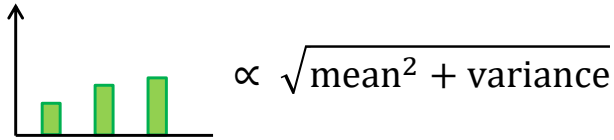
- Let us show you the practical example: this scene contains more than 5000 light sources so the clusters can be large and complicated.



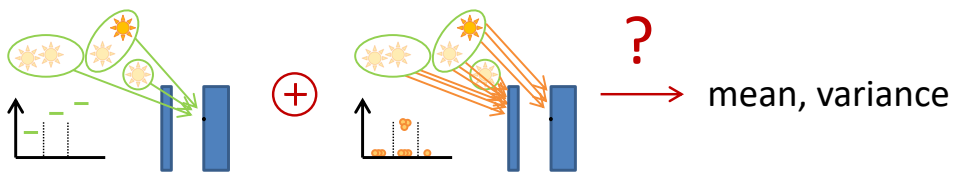
- On the left we can see an inset showing how sampling according to a mean only performs. It undersamples some tricky cluster which leads to spiky noise. And on the right we can see that sampling according to both the mean and the variance eliminates this issue.

CONTRIBUTIONS

- Optimal sampling distribution

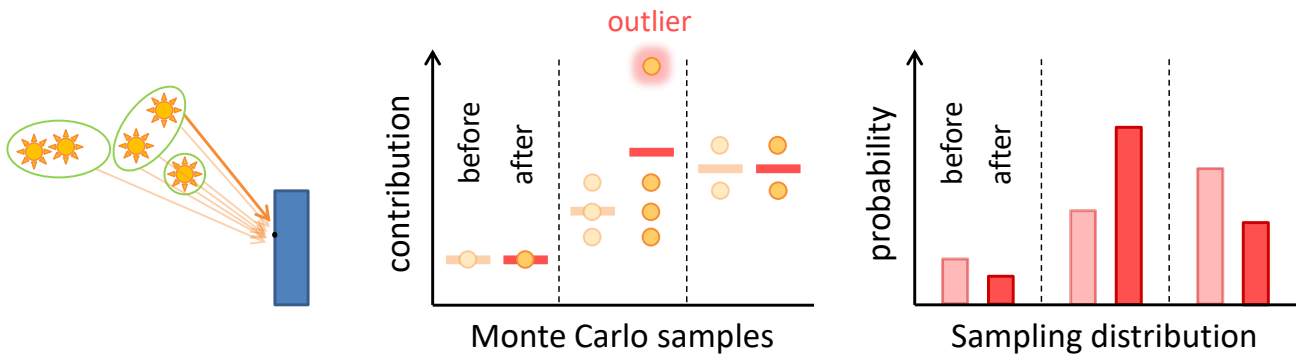


- Adaptive sampling by Bayesian inference



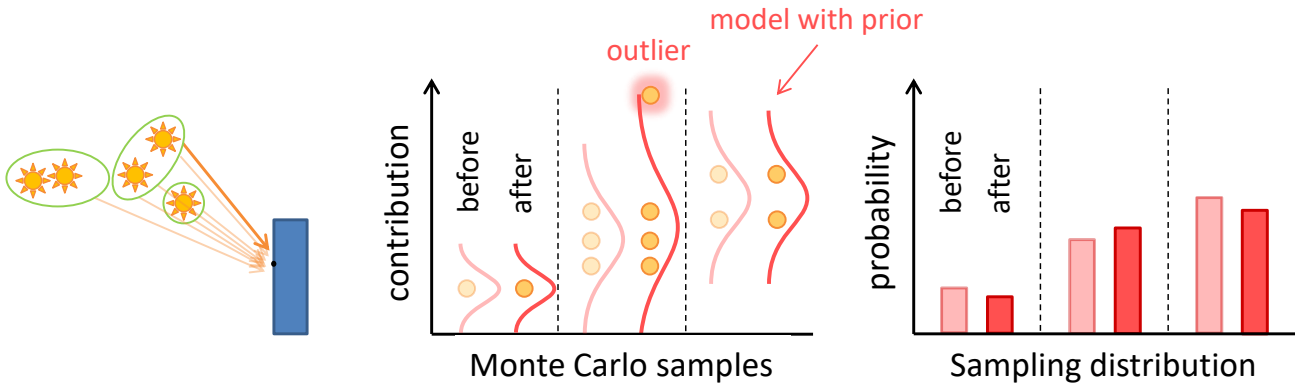
- Having explained the target optimal sampling distribution that we strive to achieve, we now show how to learn this desired distribution using Bayesian regression. The issue is that the mean and the variance needed for the optimal sampling are not known upfront and need to be learned during rendering.

Direct mean estimation is sensitive to outlier samples



- To motivate our approach, let us first show how a naive adaptive approach would behave.
- Suppose we have already taken some samples from clusters and obtained some cluster sampling distribution. Now suppose we have taken a direct illumination sample which happens to be an outlier. For instance, the sample may lie on a light extremely close to the illuminated point.
- Samples and probabilities for situation before and after the new sample are shown side by side in the same figures, the situation before uses less saturated colours.
- If we estimated the sample means (red thick lines) directly (i.e. as an average), the estimates would change abruptly due to the outlier sample. That would have a disproportionately strong effect on further cluster sampling: one cluster would get sampled very often at the expense of other clusters, leading to increased image noise or even strong fireflies.

Bayesian mean estimation is much more robust



- We propose to estimate the means (and variances) in a Bayesian manner: We model the distributions of Monte Carlo samples seen so far, while we also have some prior information about parameters of that distribution.
- As a result, when we get a new – possibly outlier – sample, our distribution is affected less abruptly and so do the cluster sampling probabilities derived from it.
- Resiliency to outlier samples without compromising the ability to learn from the new samples is the basis of our robust solution.

THE BAYESIAN WAY

1. Define **data**, their **model** with parameters θ and **prior** for θ
2. Express posterior probability of model parameters θ :

$$\text{Posterior}(\theta) \propto \prod_{x \in \text{Data}} \text{Model}(x|\theta) \times \text{Prior}(\theta)$$

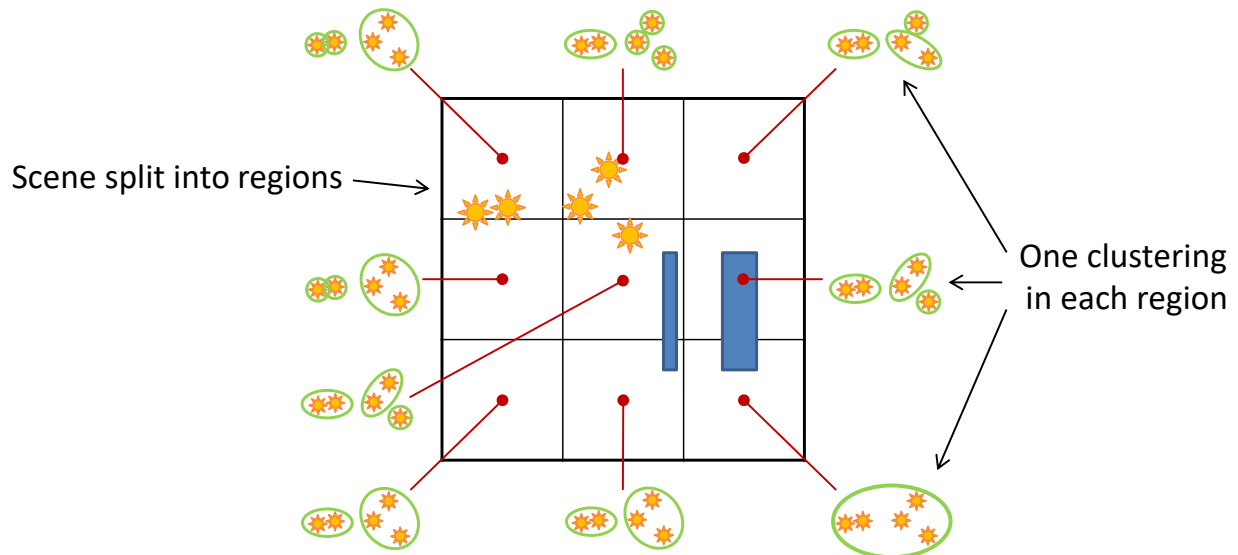
3. Find model parameters θ_{MAX} that maximize posterior:

$$\theta_{\text{MAX}} = \text{argmax}(\text{Posterior}(\theta))$$

4. Compute mean and variance of the model with θ_{MAX} :

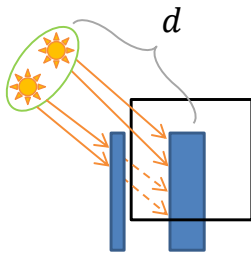
$$E[\text{Model}(x|\theta_{\text{MAX}})], \text{Var}[\text{Model}(x|\theta_{\text{MAX}})]$$

- So how do we estimate the mean and variance of the samples the Bayesian way?
 1. We first define what exactly our **data** are and how we gather them. Based on their shape we derive a parametric **model** – an analytic probability distribution that approximates the true unknown distribution of the data. Mean and variance of this model is what we are looking for. However, the model depends on a set of parameters θ whose values are not known in advance. Therefore, we define a **prior** distribution over θ that express our initial belief about their values, and we seek such values that would best explain the observed data given the prior.
 2. In order to find such values we need to express the posterior distribution – probability of θ values after observing the data. It is given by the well-known Bayesian formula: posterior \propto likelihood \times prior, where likelihood expresses a probability of observing the data, i.e. it is a product of the model over all data.
 3. We then maximize posterior w.r.t. θ which gives us the most probable θ values given the prior and the observed data. This is a so-called maximum a posteriori estimate. The direct mean estimation we showed earlier in the naive adaptive sampling corresponds to omitting the prior and maximizing directly the likelihood. This so-called maximum likelihood estimate is known to be prone to overfitting. More information on Bayesian treatment can be found in [Bishop 2006].
 4. Finally, we plug the obtained θ into the model and compute its mean and variance which we were looking for.
- Having described our entire Bayesian framework we now return to the first point and define the three key components: data, model and prior.

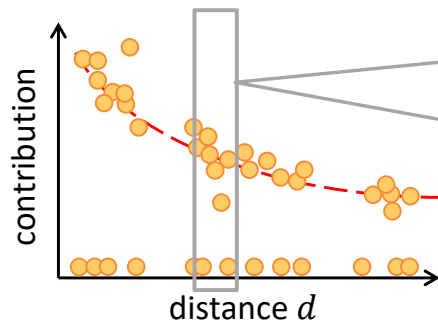


- We start by explaining our basic setup.
- We split the scene into fixed spatial **regions** and for each region we compute exactly one light clustering. Our method then operates independently on each cluster-region pair.
- Technical details:
 - Light clustering is computed on demand when a respective region is queried for the first time and is kept cached in that region for further use.
 - A regular grid is used for splitting the scene. However, as we discuss in results, the method is not sensitive to size of the regions, so any space subdivision could be used.

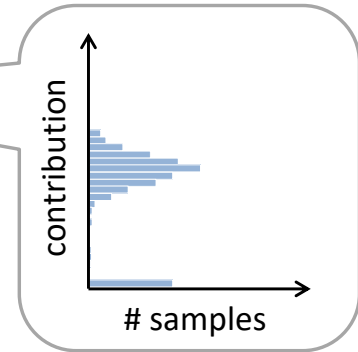
Cluster-region pair



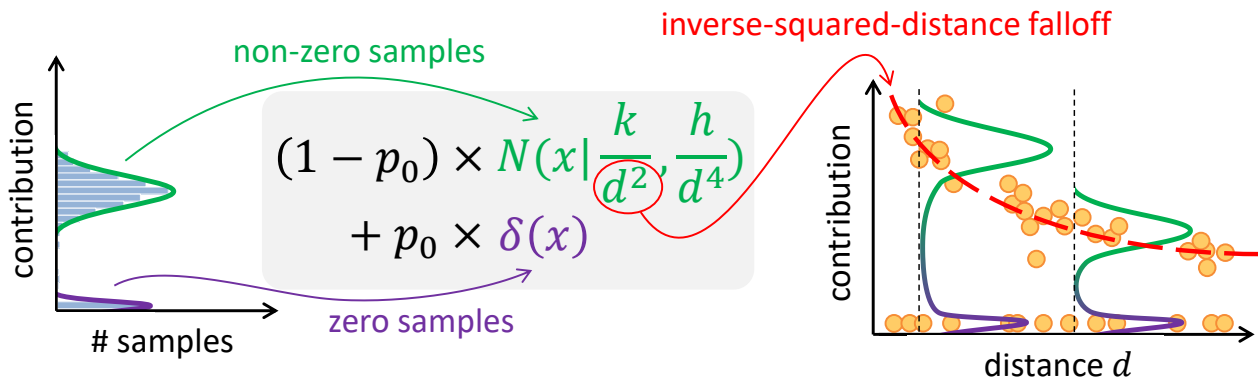
Gathered data



Histogram for one distance



- Let's now focus on one cluster-region pair data, i.e. on samples collected from **one cluster** in a **single scene region** (an example shown on the left).
- For each sample we keep track of its contribution (i.e. MC direct illumination estimate) and the distance d in the geometry factor.
- If we plot the gathered data w.r.t. to the distance, the plot (shown in the middle) reveals the nature of the relation of illumination estimates to the distance. One may observe the inverse-squared falloff with the distance (red dashed line), and a number of zero-valued (occluded) samples (yellow points at the bottom of the plot).
- In order to model these data, we take a closer look on how the data are distributed for a particular distance. From a histogram (shown on the right) we can see that the non-zero data follow a curve similar to a Gaussian and the rest forms a sharp peak at zero. This suggests how the data could be modeled.



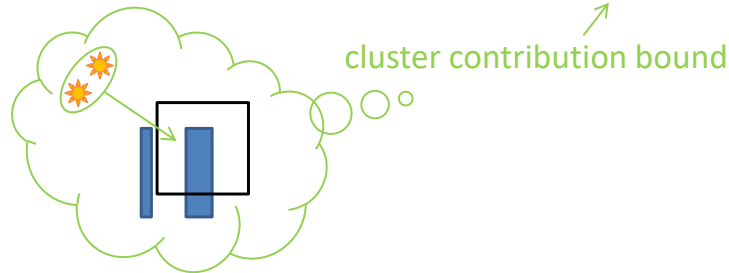
Model parameters θ : k, h - normal distr. parameters
 p_0 - probability of occlusion

- Our model (in grey) is therefore a **parametric regression model**, which - for a given distance d - yields a distribution of direct illumination sample values. We design our data model as follows:
 - Non-zero samples are modeled by a normal distribution (i.e. a Gaussian) with mean and variance being a function of the distance associated with the samples. This part of the model has two parameters, k and h .
 - The zero valued samples are incorporated by mixing the inverse-squared-distance falloff model with a delta function, and it is controlled by the parameter p_0 , which has the meaning of occlusion probability.
- Having designed our data model, we proceed to define the prior distribution for all three model parameters, k , h , and p_0 .

Conjugate prior distributions for model parameters θ :

$$p_0 \sim \text{Beta}(p_0 | \dots)$$

$$k, h \sim \text{Normal inverse gamma}(k, h | \mu_0, \dots)$$



- The model we have just defined has parameters k , h and p_0 . In the Bayesian treatment, these are viewed as random variables governed by a certain probability distribution.
- The “prior” is a specific probability distribution that we believe these parameters follow, without observing any data at all.
- A “conjugate prior” is a convenience construction: we choose the prior to have a functional form that will be preserved when multiplied by the likelihood of observed data.
- We showed in the paper that the conjugate prior in our case takes form of the Beta distribution for p_0 and the normal-inverse-gamma distribution for the parameters k and h .
- There are various **hyperparameters** in the equations, but one parameter which stands out is μ_0 , for which we use the conservative **cluster contribution bound**. This hyperparameter expresses our a priori belief about the mean of our data: That is, the cluster contribution bounds provide a prior information about the expected contribution each cluster will make to each scene region. This belief is then continually refined as we observe the actual direct illumination contributions made by sampling the clusters.

ALGORITHMIC SUMMARY

- During each direct illumination estimation
 - Obtain clustering for the current region
 - For each cluster in the clustering estimate mean and variance the Bayesian way (using statistics of all previous cluster-region samples)
 - Build sampling distribution over clusters (prop. to $\sqrt{\text{mean}^2 + \text{variance}}$)
 - Draw a new sample (i.e. choose a cluster from the distribution, then a light in the cluster, then a point on the light)
 - Update the cluster-region statistics with the new sample

- To wrap it up, our algorithm is as described on the slide.

RESULTS

ADVANCES IN MONTE CARLO RENDERING: THE LEGACY OF JAROSLAV KŘIVÁNEK

- We now demonstrate our solution in practice.

TESTS

- Performance

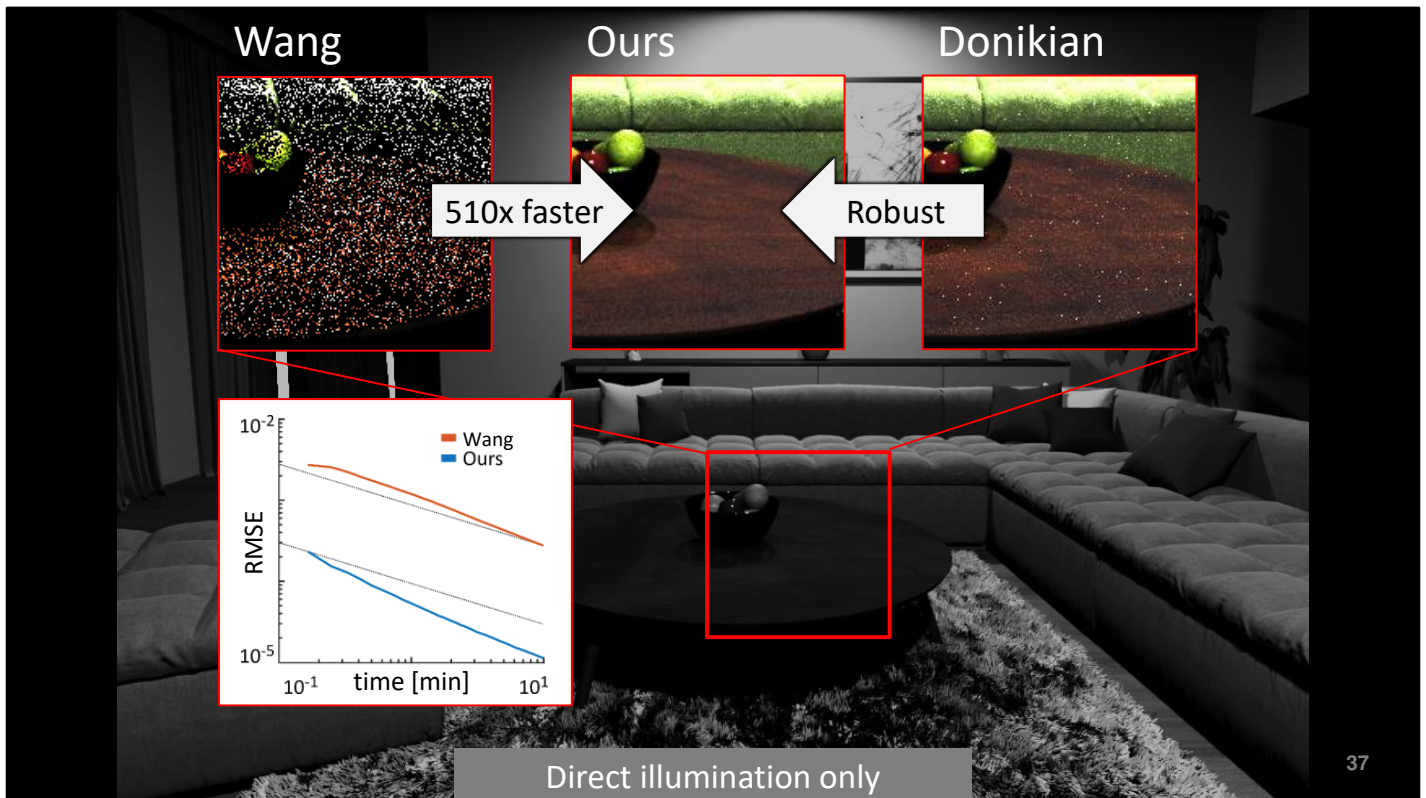
| | Direct only | Direct + indirect |
|-------------------|-------------|-------------------|
| Simple occlusion | | |
| Complex occlusion | | |

- Grid resolution

- This is a list of all comparisons we will present. We will start with performance testing in a scene with simple occlusion in direct illumination only setting.



- This is the living room scene from the beginning of our presentation. It is lit mostly by a few small area lights on the ceiling, only in the left part sunlight can enter the room through the windows.



- As you could already see, the non-adaptive sampling of Wang and Akerlund [2009] does not perform well in this scene. Why is that? The sun is much stronger than the ceiling lights and is therefore sampled much more often even though it is actually occluded - and so most of the samples are wasted.
- Donikian et al.'s algorithm [2006] improves the result significantly, as it quickly learns the sun occlusion. On the other hand, it struggles with the ceiling lights. They are covered by shades which block some of the samples. The method believes these lights are actually completely occluded, and consequently undersamples these lights and introduces spiky noise.
- Our method can also quickly learn the sun occlusion and converges more than 500× faster than the non-adaptive method of Wang and Akerlund. Interestingly, in the RMSE plot we can even observe a higher empirical convergence rate. At the same time, thanks to the Bayesian treatment, our method is robust, does not get confused by the occluded samples and avoids the spiky noise.

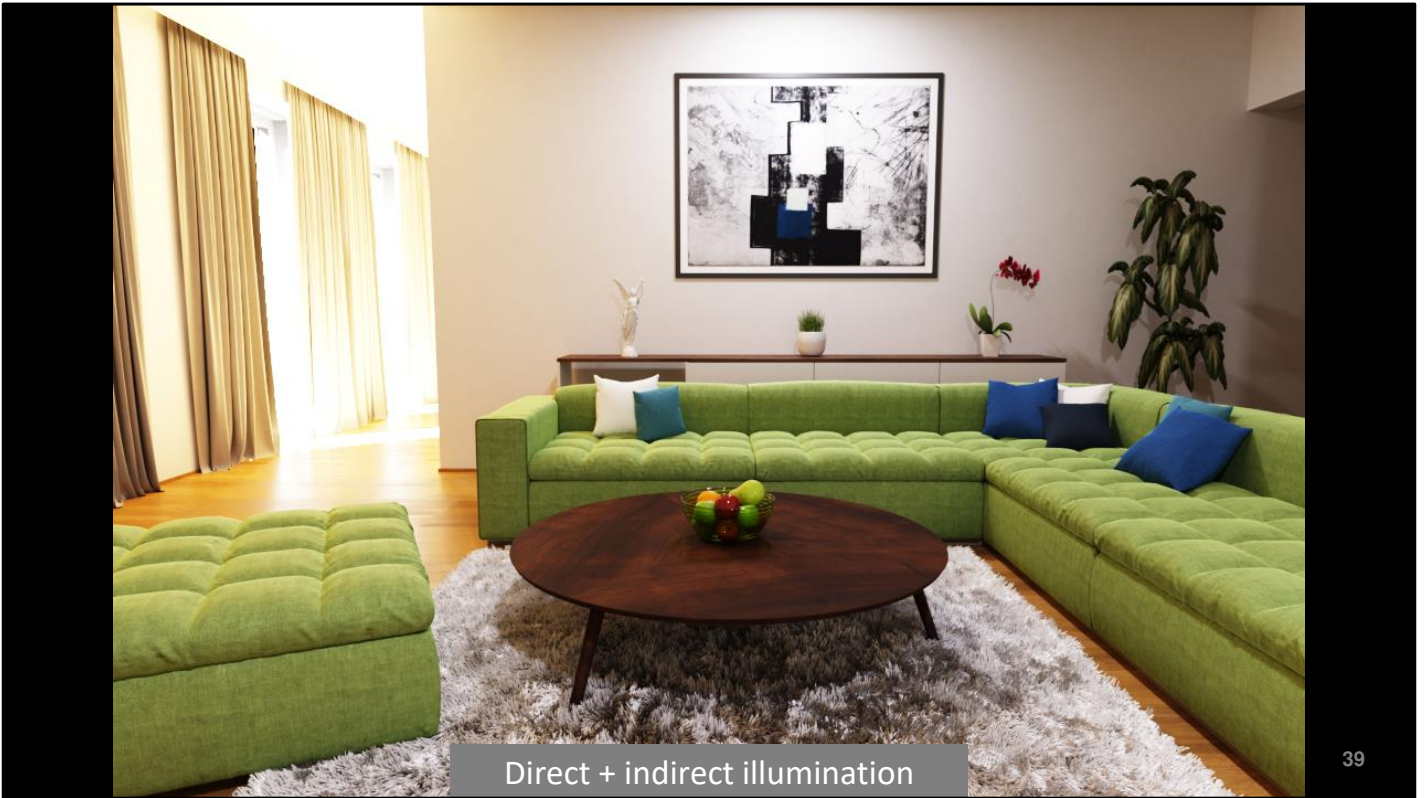
TESTS

- Performance

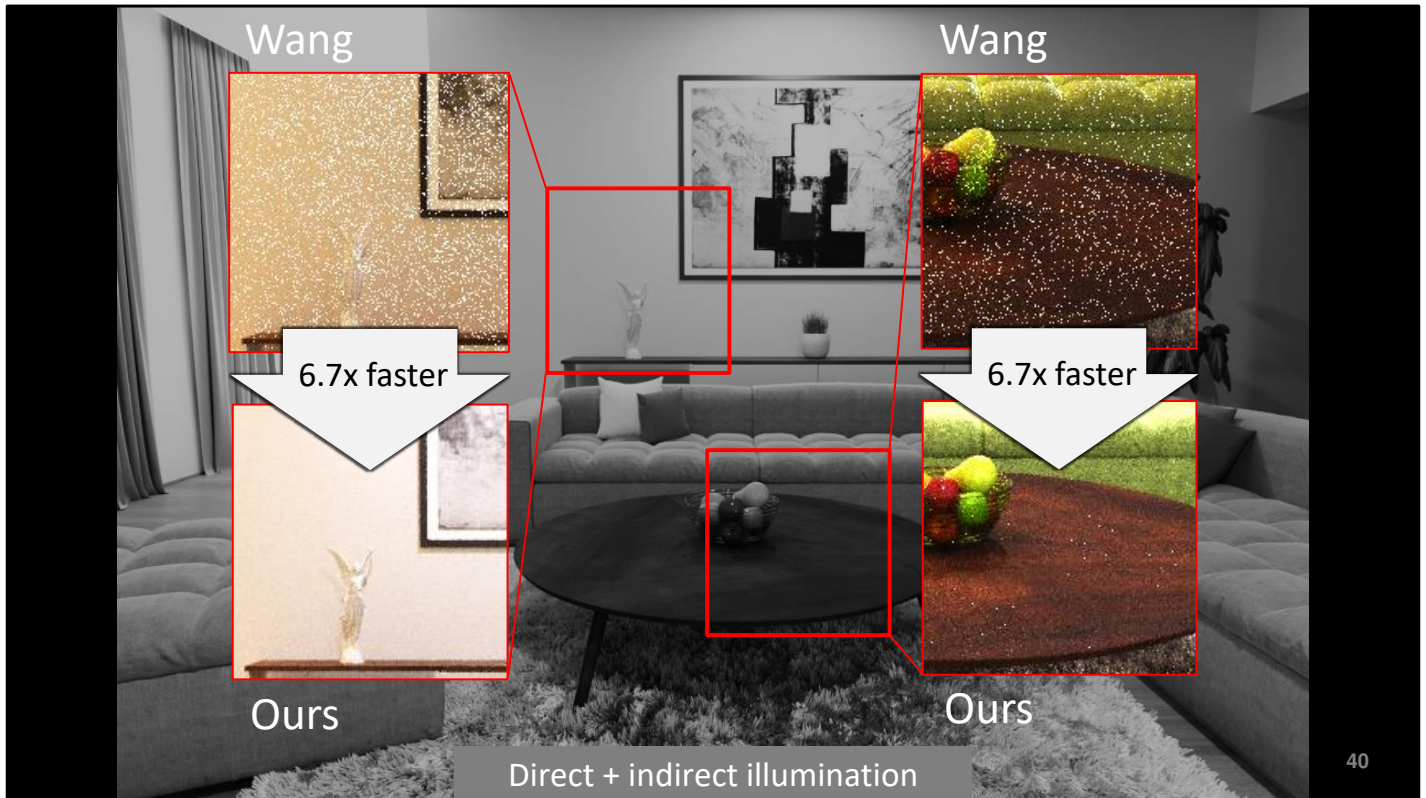
| | Direct only | Direct + indirect |
|-------------------|-------------|-------------------|
| Simple occlusion | ✓ | |
| Complex occlusion | | |

- Grid resolution

- So, that was the direct illumination. However, in practice one is usually more interested in images containing both direct and indirect components.



- This is the same scene but with the indirect component included.



- We can see that the strong direct illumination noise of Wang and Akerlund dominates also in the complete image. The direct component is definitely the main source of noise in this scene.
- By using our method in the next event estimation in path tracing, we are able to improve the light sampling on every path vertex and get more than 6x overall speedup.
- Note that the remaining noise at the bottom right of the image is caused solely by the indirect component and cannot be remedied by our method.

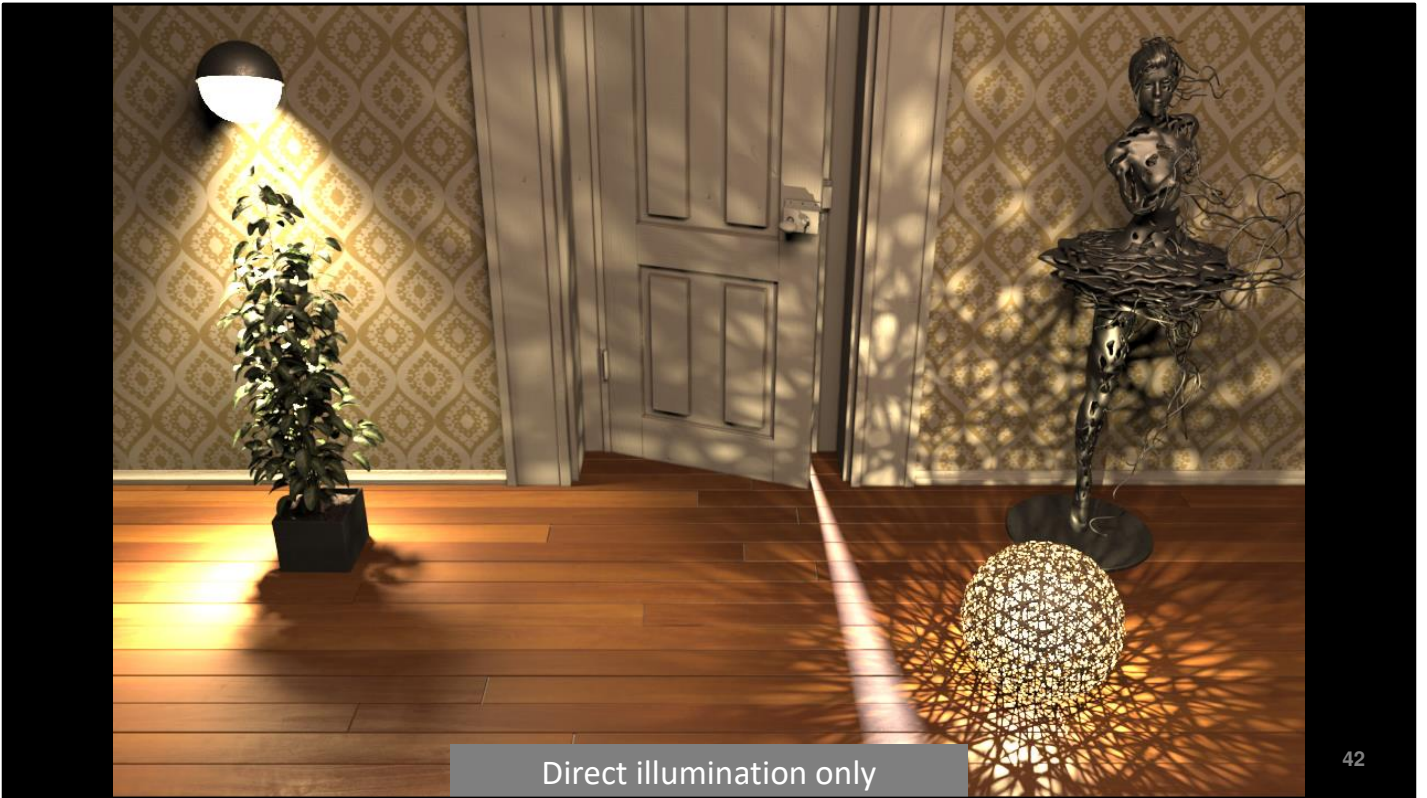
TESTS

- Performance

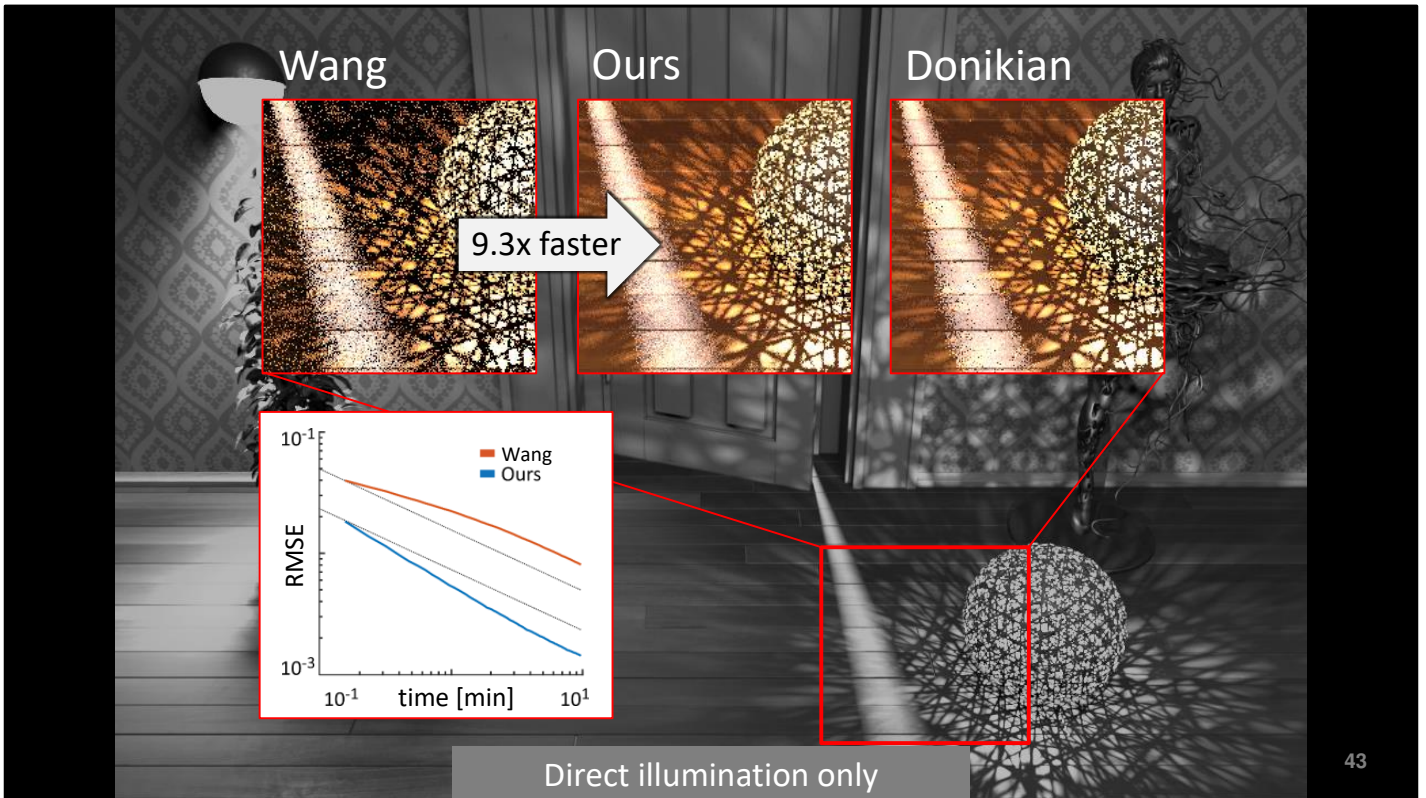
| | Direct only | Direct + indirect |
|-------------------|-------------|-------------------|
| Simple occlusion | ✓ | ✓ |
| Complex occlusion | | |

- Grid resolution

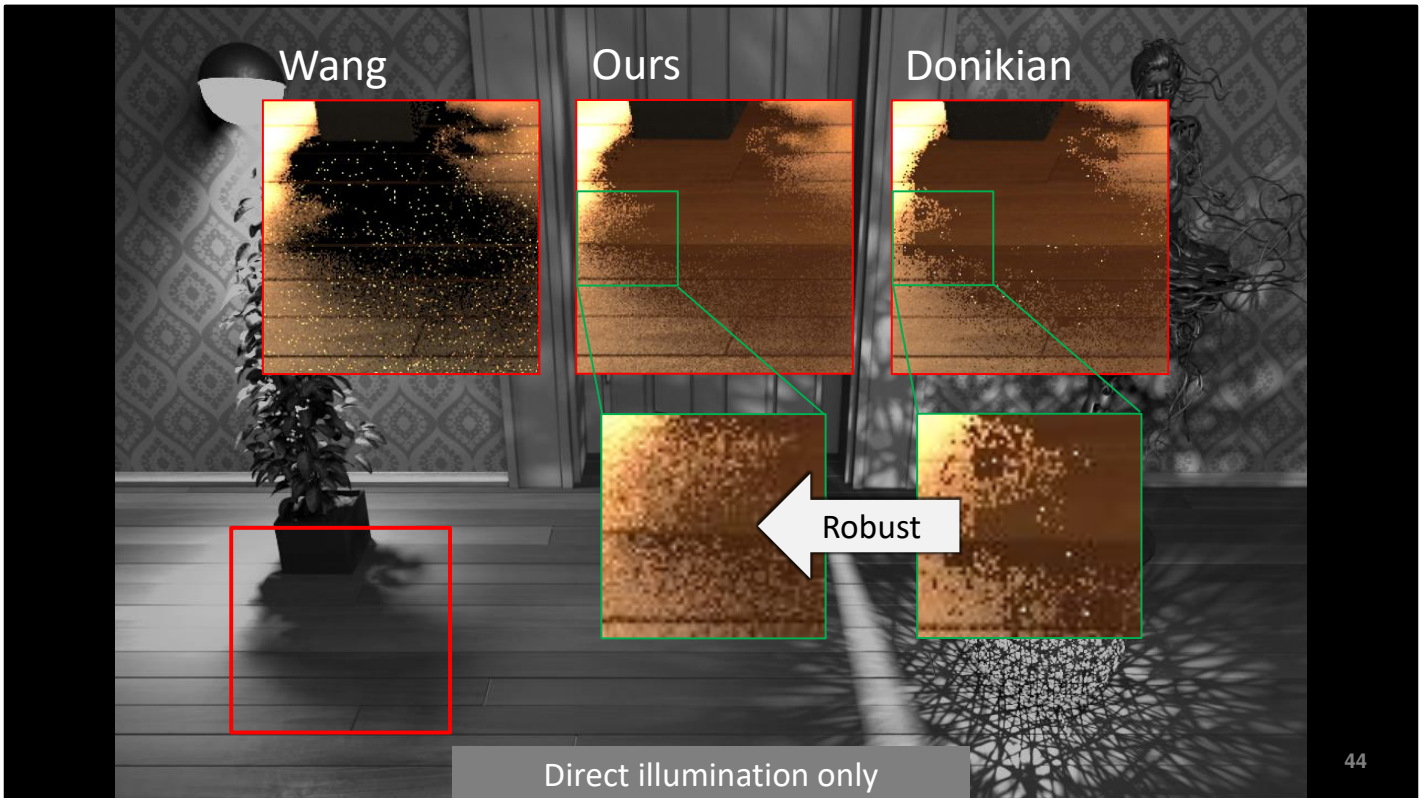
- Now we stress-test robustness of our method in a scene with complex occlusion.



- This scene presents a real challenge due to its highly structured illumination plus there are lights in the other room behind the door.



- In this part the method of Wang and Akerlund produces a lot of noise again, as it wastes samples on the lights behind the door.
- On the other hand, our method performs well. It is more than 9 times faster, and again we can observe higher empirical convergence rate. And all that without introducing any artifacts in such a complicated illumination setting.
- Donikian et al.'s method at first also seems to perform well but further inspection would discover small blocky artifacts in the shadows.



- In this inset, the non-adaptive sampling of Wang and Akerlund again does not perform well.
- But this time also the Donikian et al.'s method fails badly: The illumination coming through the plant leaves is too complex for the ad-hoc learning to handle it well, the method overfits and produces square-shaped artifacts.
- This is exactly the problem of essentially all previous adaptive methods: While they can sometime provide substantial speedup, **they do not fail gracefully**, and one cannot rely on them.
- Our Bayesian learning makes our method much more robust and artifact-free.

TESTS

- Performance

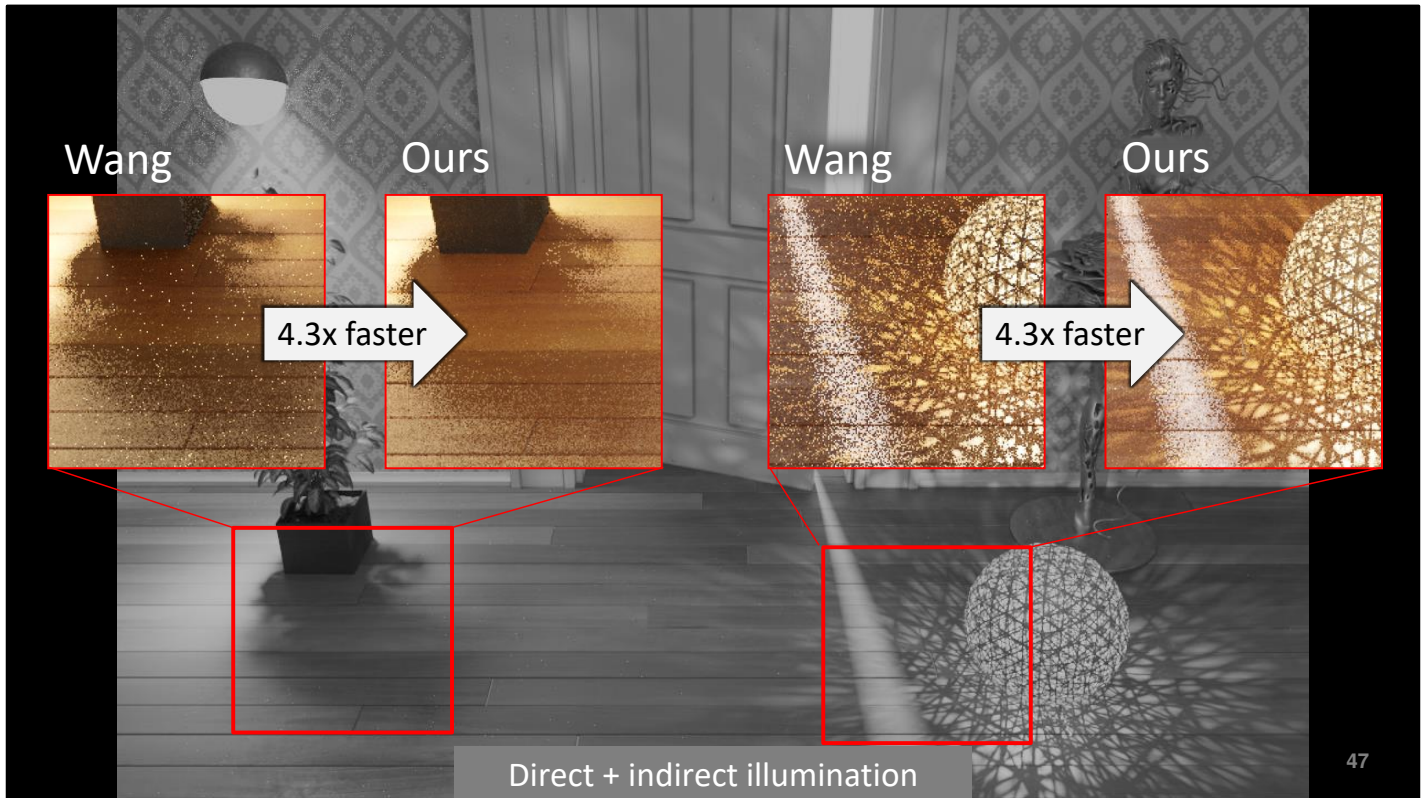
| | Direct only | Direct + indirect |
|-------------------|-------------|-------------------|
| Simple occlusion | ✓ | ✓ |
| Complex occlusion | ✓ | |

- Grid resolution

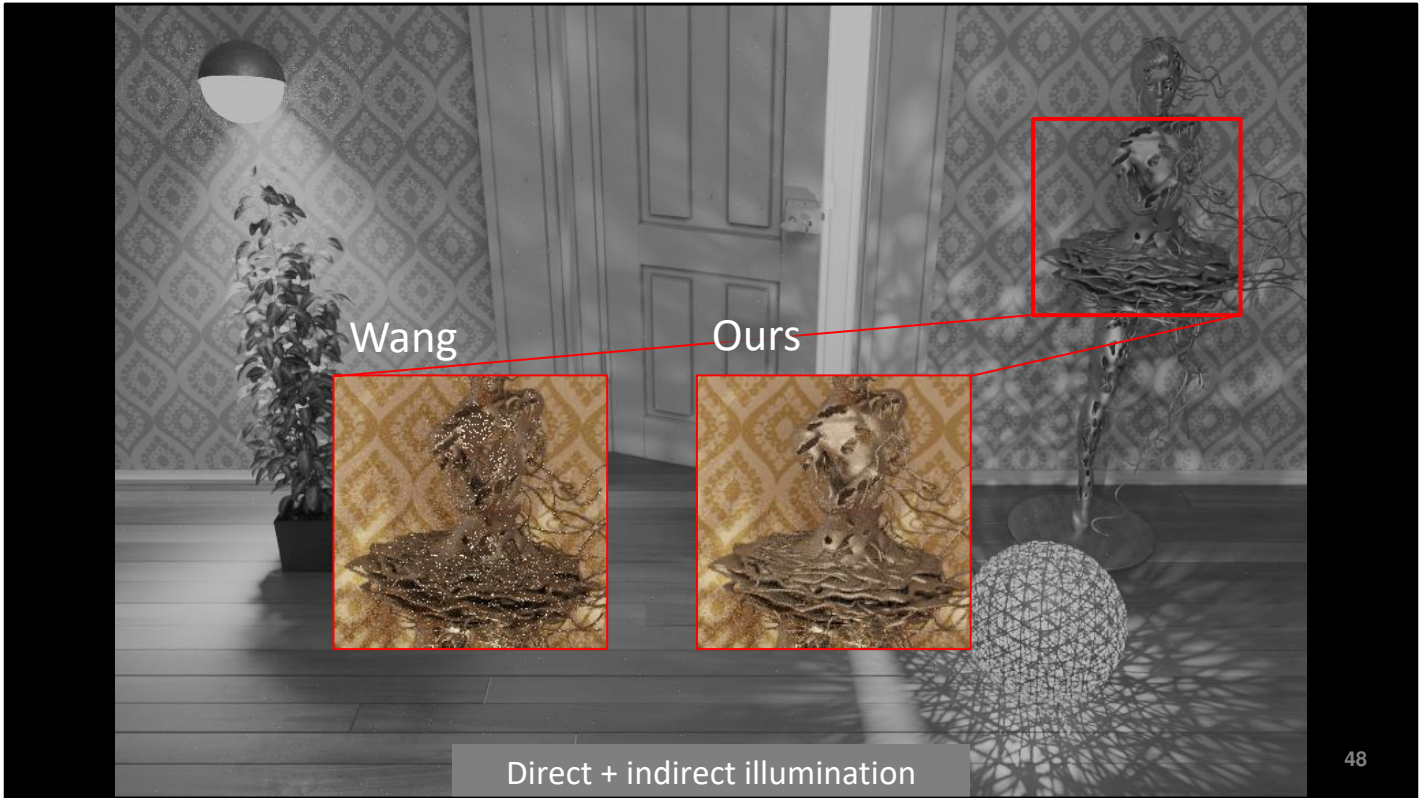
- Finally, let's test the complex occlusion also with the indirect component.



- We take a look at the same scene.



- We can see that the direct illumination noise again dominates the complete image when rendered using the method of Wang and Akerlund.
- Our method eliminates it and renders the complete image more than four times faster and without any artifacts.



- There is one more interesting area in this scene: The statue is made of glossy metal, and even though our method does not take the surface BRDF into account, it performs significantly better even there.

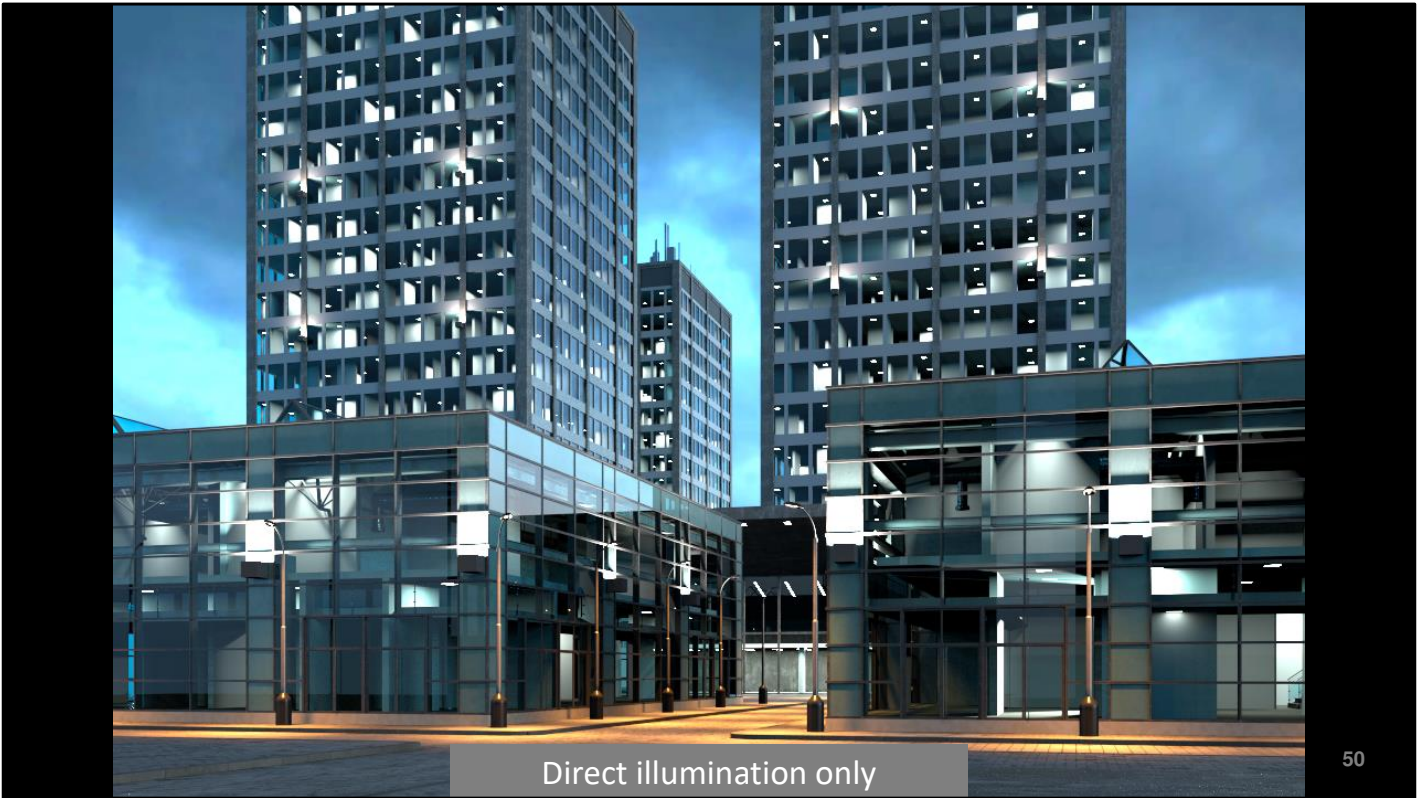
TESTS

- Performance ✓

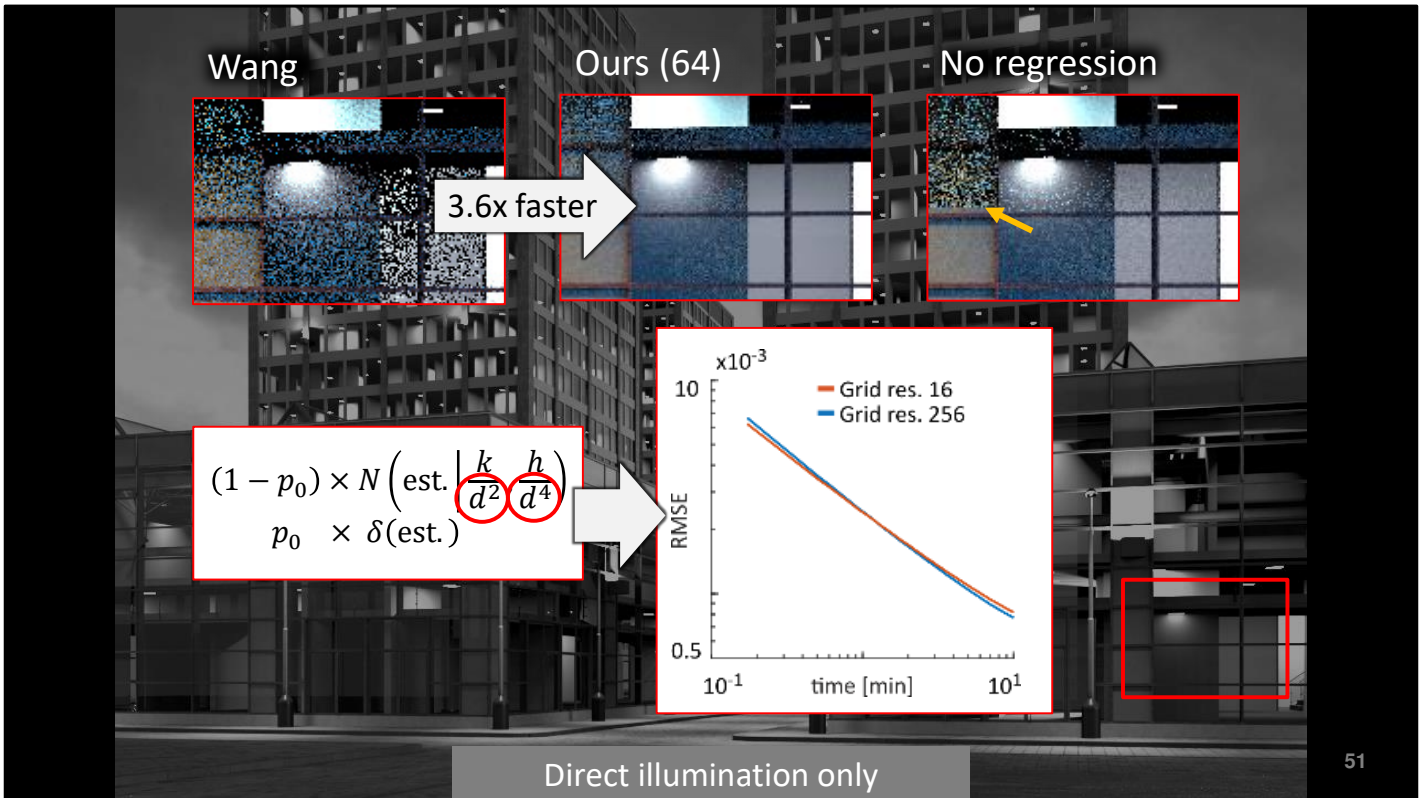
| | Direct only | Direct + indirect |
|-------------------|-------------|-------------------|
| Simple occlusion | ✓ | ✓ |
| Complex occlusion | ✓ | ✓ |

- Grid resolution

- Recall that we divide the scene into regions by a uniform grid of a fixed resolution. We now test how this resolution affects the algorithm's performance.



- For that purpose we use this relatively large scene containing many lights.



- With our default choice of 64 regions per the shortest grid dimension, our method performs more than three times faster than the method of Wang and Akerlund. So what about other grid resolutions? As it turns out, our method is rather insensitive to the actual grid resolution. And so even much smaller as well as much higher resolutions all perform roughly the same, as shown in the plot. This is due to the regression modeling of the distance falloff. Without the regression model we would have to use a much higher grid resolution, otherwise we would see sudden noise transitions between regions, as in the inset in the upper right corner.

CONCLUSION

ADVANCES IN MONTE CARLO RENDERING: THE LEGACY OF JAROSLAV KŘIVÁNEK

CONTRIBUTION

- Bayesian framework for robust adaptivity
- Optimal cluster sampling
- Algorithm for direct illumination
 - Unbiased, adaptive, robust
 - Easy to integrate into a path tracer

- To conclude, the main contribution of our work is a Bayesian framework for adaptive/guided Monte Carlo quadrature. It enables exploiting the large potential of the adaptive sampling approach in Monte Carlo methods, while avoiding the biggest weakness of previous attempts – the lack of robustness.
- We applied this framework on the problem of direct illumination sampling. In the process we derived the optimal sampling distribution, taking variance into account, and developed an unbiased adaptive direct illumination algorithm with online learning of light sampling distributions. It is easy to integrate into a path tracer and suitable for interactive rendering (the up front cost is minimized as all learning happens on the fly during rendering).
- Our new framework is not limited to the direct illumination though and we are certain that other applications of adaptive sampling will benefit from it as well and it opens the path for many other tools of statistical machine learning (such as full Bayes or variational Bayes).

Machine Learning | Bayesian modeling
=
Excellent framework for
guided/adaptive Monte Carlo

References

- [Bishop 2006] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer-Verlag New York Inc.
- [Boughida and Boubekeur 2017] Malik Boughida and Tamy Boubekeur. 2017. Bayesian collaborative denoising for Monte Carlo rendering. *Computer Graphics Forum* 36, 4 (2017), 137–153.
- [Brouillat et al. 2009] Jonathan Brouillat, Christian Bouville, Brad Loos, Charles Hansen, and Kadi Bouatouch. 2009. A Bayesian Monte Carlo approach to global illumination. *Computer Graphics Forum* 28, 8 (2009), 2315–2329.
- [Cappé et al. 2004] Olivier Cappé, Arnaud Guillin, Jean-Michel Marin, and Christian P. Robert. 2004. Population Monte Carlo. *Journal of Computational and Graphical Statistics* 13, 4 (2004), 907–929.
- [Donikian et al. 2006] Michael Donikian, Bruce Walter, Kavita Bala, Sebastian Fernandez, and Donald P. Greenberg. 2006. Accurate direct illumination using iterative adaptive sampling. *IEEE Transactions on Visualization and Computer Graphics* 12, 3 (2006), 353–363.
- [Dutré and Willems 1995] Philip Dutré and Yves D. Willems. 1995. Potential-driven Monte Carlo Particle Tracing for Diffuse Environments with Adaptive Probability Functions. *Rendering Techniques* 95 (1995).
- [Jensen 1995] Henrik Wann Jensen. 1995. Importance driven path tracing using the photon map. *Rendering Techniques* 95 (1995), 326–335.
- [Lafortune and Willems 1995] Eric P. Lafortune and Yves D. Willems. 1995. A 5D Tree to Reduce the Variance of Monte Carlo Ray Tracing. *Rendering Techniques* 95 (1995), 11–20.
- [Lepage 1980] Peter G. Lepage. 1980. VEGAS - an adaptive multi-dimensional integration program. *CLNS-447* (1980), 30 pages.
- [Marques et al. 2013] Ricardo Marques, Christian Bouville, Mickaël Ribardiere, Luís Paulo Santos, and Kadi Bouatouch. 2013. A spherical gaussian framework for Bayesian Monte Carlo rendering of glossy surfaces. *IEEE Trans. Vis. Comput. Graph.* 19, 10 (2013), 1619–1632.
- [Mitchell 1987] Don P. Mitchell. 1987. Generating Antialiased Images at Low Sampling Densities. *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques* (1987).
- [Müller et al. 2017] Thomas Müller, Markus Gross, and Jan Novák. 2017. Practical Path Guiding for Efficient Light-Transport Simulation. *Eurographics Symposium on Rendering* 36, 4 (2017).
- [Shirley et al. 1996] Peter Shirley, Changyaw Wang, and Kurt Zimmerman. 1996. Monte Carlo techniques for direct lighting calculations. *ACM Transactions on Graphics* 15, 1 (1996), 1–36.
- [Vévoda et al. 2018] Petr Vévoda, Ivo Kondapaneni, and Jaroslav Křivánek. 2018. Bayesian Online Regression for Adaptive Direct Illumination Sampling. *ACM Transactions on Graphics* 37, 4 (2018).
- [Vorba et al. 2014] Jiří Vorba, Ondřej Karlík, Martin Šik, Tobias Ritschel, and Jaroslav Křivánek. 2014. On-line Learning of Parametric Mixture Models for Light Transport Simulation. *ACM Transactions on Graphics* 33, 4 (2014), 101:1–101:11.

[Walter et al. 2005] Bruce Walter, Sebastian Fernandez, Adam Arbree, Kavita Bala, Michael Donikian, and Donald P Greenberg. 2005. Lightcuts: a scalable approach to illumination. *ACM Transactions on Graphics* 24, 3 (2005), 1098–1107.

[Wang and Akerlund 2009] Rui Wang and Oskar Akerlund. 2009. Bidirectional Importance Sampling for Unstructured Direct Illumination. *Computer Graphics Forum* 28, 2 (2009), 269–278.

9 Multiple Importance Sampling



SIGGRAPH THINK BEYOND
2020 19-23 JULY WASHINGTON DC

ADVANCES IN MONTE-CARLO
RENDERING:
THE LEGACY OF JAROSLAV KŘIVÁNEK

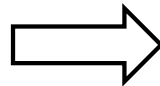
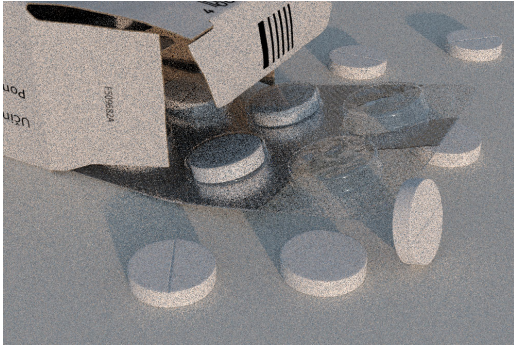
Multiple Importance Sampling



ADVANCES IN MONTE CARLO RENDERING: THE
LEGACY OF JAROSLAV KŘIVÁNEK

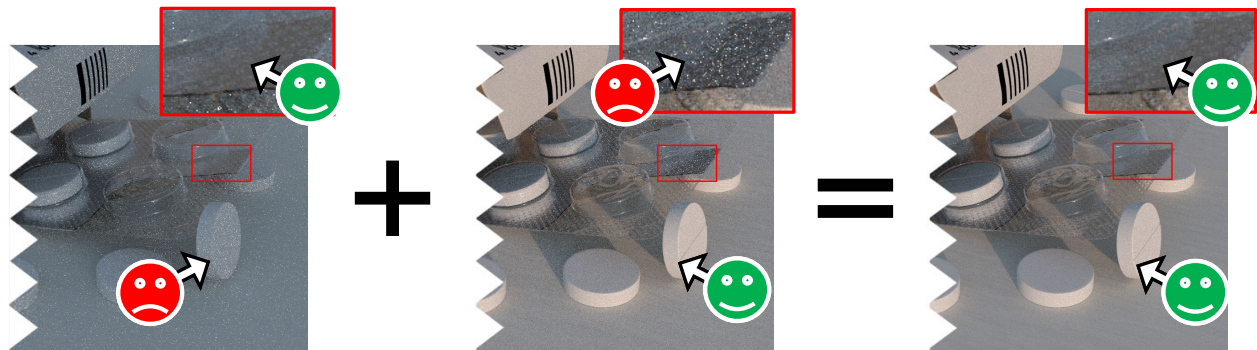
In this part of the course, we will discuss recent advances for Multiple Importance Sampling (MIS) – a technique to combine multiple rendering algorithms into a better one.

RENDERING GOAL: LOW NOISE IN SHORT TIME



In Monte Carlo rendering, images contain noise. Eventually, given enough time, that noise will disappear. The goal when developing rendering algorithms is to minimize the time it takes to obtain a noise-free image.

MULTIPLE IMPORTANCE SAMPLING (MIS)



Technique A

Technique B

Combined (MIS)

When designing the “one” algorithm that can render all scenes at the lowest level of noise possible, multiple importance sampling (MIS) plays an important role. It is unlikely that we will find a single sampling technique that is robust and efficient enough to form the “one” algorithm.

MIS allows us to combine multiple techniques into one algorithm, while retaining the advantages of every individual technique. Here, we see an example that combines two techniques “A” and “B”.

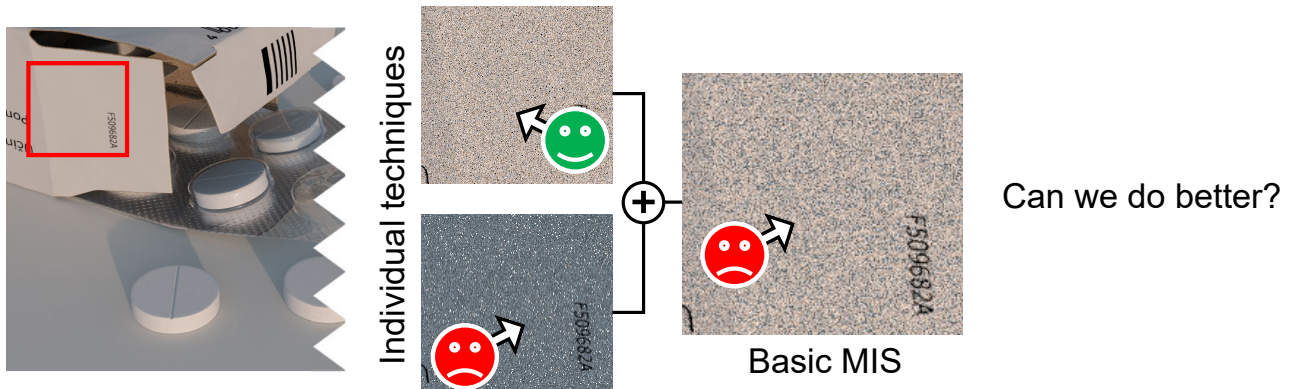
Technique “A” in this case performs well in the shadows but poorly for the direct illumination. Technique “B” is the exact opposite. When combining both via MIS, their benefits add up and we achieve a much nicer image overall.

Veach and Guibas: “**Optimally Combining Sampling Techniques for Monte Carlo Rendering.**”
SIGGRAPH '95.



MIS was invented in 1995 by Veach and Guibas. It had such a tremendous impact that Eric Veach was awarded the Scientific and Engineering Award in 2014.

MIS: INCREASES ROBUSTNESS, MAY REDUCE EFFICIENCY



While it is a great tool, MIS is not perfect. Let us look at a different part of the “pills” scene. Again we combine the same two techniques. On the packaging, one is almost perfect, the other performs poorly.

Unfortunately, in MIS the samples are distributed among the two techniques and thus we may get worse quality than with just one technique. So the question is: can we somehow do better than that?

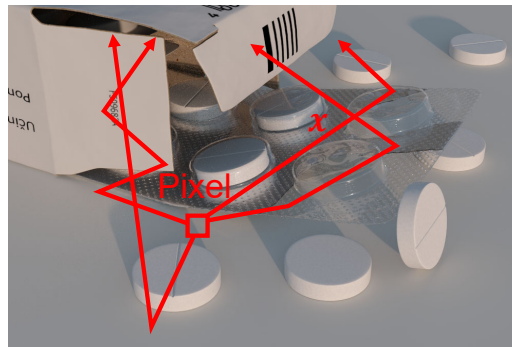
CAN WE DO BETTER?

- Yes, with ...
 - ... better weighting functions
 - ... better sampling densities
 - ... better sample allocation

Fortunately, there is three ways how to improve upon MIS: better weighting functions, better sampling densities, and better sample allocation. In the following, we will first briefly review some required background, then we will discuss these possible improvements in more detail.

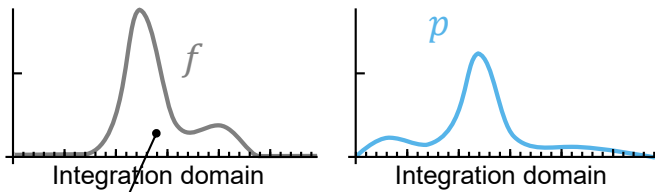
$$I_{\text{pixel}} = \int \underbrace{f(x)}_{\text{Contribution}} dx$$

Path x



When we perform light transport simulation, we need to integrate over all light paths that connect a pixel to a light source. For each such path, we compute its contribution to the image. Monte Carlo integration and importance sampling are used to estimate that integral efficiently.

IMPORTANCE SAMPLING



$$F = \int f(x) dx$$

$$X_i \sim p$$

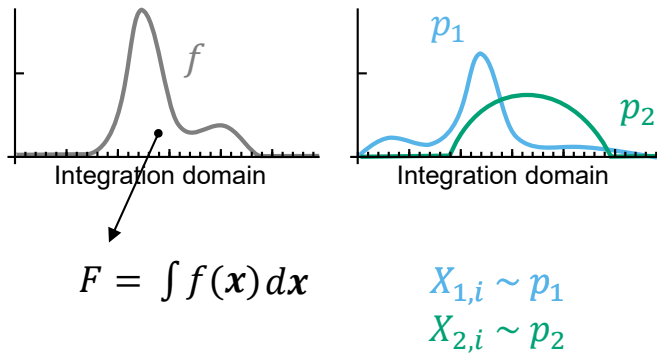
$$\langle F \rangle = \frac{1}{n} \sum \frac{f(X_i)}{p(X_i)}$$

Consider a simple integration problem, where we integrate a function f .

To estimate its integral, we use an importance sampling technique p and we draw n samples according to it. We obtain an estimate of the integral by combining all samples.

The better we sample important parts of the integrand, the lower the variance of the estimator. But what if we cannot find a single technique that would sample f well?

MULTIPLE IMPORTANCE SAMPLING



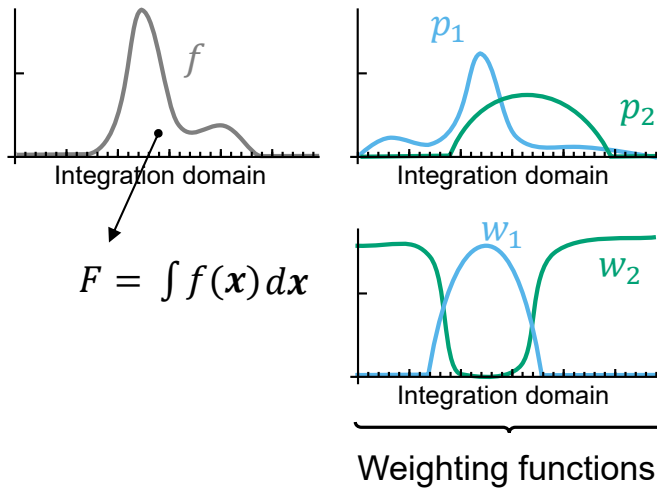
$$\langle F \rangle_1 = \frac{1}{n_1} \sum \frac{f(X_{1,i})}{p_1(X_{1,i})}$$
$$\langle F \rangle_2 = \frac{1}{n_2} \sum \frac{f(X_{2,i})}{p_2(X_{2,i})}$$

Veach and Guibas [1995]

Multiple importance sampling can help us. There might be another technique suited for sampling a different part of f .

To use MIS, we generate another set of samples, from a different density p_2 , forming another estimator for the same integral.

MULTIPLE IMPORTANCE SAMPLING



$$\langle F \rangle = \frac{1}{n_1} \sum \frac{f(X_{1,i})}{p_1(X_{1,i})} w_1(X_{1,i}) + \frac{1}{n_2} \sum \frac{f(X_{2,i})}{p_2(X_{2,i})} w_2(X_{2,i})$$

Veach and Guibas [1995]

Each sample from the two techniques is weighted by a weighting function w . If we sum up the weighted estimates, we achieve a new combined estimator, hopefully with lower variance.

When chosen well, the weighting functions assign higher weight to the regions that are sampled well by the corresponding density.

PREVIOUS ATTEMPTS TO IMPROVE MIS

- Sample allocation ○

- Pajot et al. [2011]
- Lu et al [2013]
- Havran and Sbert [2014], Sbert et al. [2016], Sbert and Havran [2017], ...

- Weighting functions ○

- Georgiev et al. [2012]
- Elvira et al. [2015; 2016]
- Kondapaneni et al. [2019]
- Grittmann et al. [2019]

- Sampling distributions ○

- Karlik et al. [2019]

$$\langle F \rangle = \frac{1}{n_1} \sum \frac{f(X_{1,i})}{p_1(X_{1,i})} w_1(X_{1,i}) + \frac{1}{n_2} \sum \frac{f(X_{2,i})}{p_2(X_{2,i})} w_2(X_{2,i})$$

This basic recipe can be improved upon in multiple ways.

Most previous work has tackled the question of better sample allocation. Distributing the available sample budget well across all techniques can significantly improve efficiency.

Less attention has been on finding better weighting functions. There are a few domain specific enhancements to the original heuristics proposed by Veach and Guibas [1995]. We will discuss how to derive the truly optimal weights, and how to enhance MIS with variance information to tackle hard cases.

A novel avenue from improvement is adapting the sampling densities themselves.

PREVIOUS ATTEMPTS TO IMPROVE MIS

- Sample allocation

- Pajot et al. [2011]
- Lu et al [2013]
- Havran and Sbert [2014], Sbert et al. [2016], Sbert and Havran [2017], ...

- Weighting functions

- Georgiev et al. [2012]
- Elvira et al. [2015: 2016]
- Kondapaneni et al. [2019]
- Grittmann et al. [2019]

- Sampling distributions

- Karlík et al. [2019]

$$\langle F \rangle = \frac{1}{n_1} \sum \frac{f(X_{1,i})}{p_1(X_{1,i})} w_1(X_{1,i}) + \frac{1}{n_2} \sum \frac{f(X_{2,i})}{p_2(X_{2,i})} w_2(X_{2,i})$$

We will discuss these

In the remainder of this part of the course, we will provide an overview over these three approaches. More details can be found in the respective papers:

[Kondapaneni et al. 2019]

Ivo Kondapaneni, Petr Vévoda, Pascal Grittmann, Tomáš Skřivan, Philipp Slusallek, Jaroslav Křivánek.

Optimal Multiple Importance Sampling.

[Grittmann et al. 2019]

Pascal Grittmann, Iliyan Georgiev, Philipp Slusallek, Jaroslav Křivánek.

Variance-Aware Multiple Importance Sampling.

[Karlík et al. 2019]

Ondřej Karlík, Martin Šik, Petr Vévoda, Tomáš Skřivan, Jaroslav Krivanek.

MIS compensation: optimizing sampling techniques in multiple importance sampling.

OPTIMAL MULTIPLE IMPORTANCE SAMPLING

[Kondapaneni et al. 2019]

ADVANCES IN MONTE CARLO RENDERING: THE LEGACY OF JAROSLAV KŘIVÁNEK

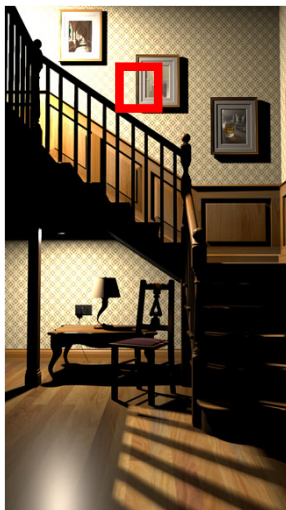
We start with our research on optimal MIS weights. First we will show, that the balance heuristic is further from the optimum than has been believed so far. Then we will show that provably optimal weights, minimizing the variance of an MIS estimator, exist and have a closed form solution. With that closed form solution, we will also show that the optimal weights are equivalent to a control variate. And lastly, we will show that the optimal weights are not a mere theoretical construct: They lend themselves to a practical implementation in light transport.

- Simple weighting functions: $w_i(x) = \frac{n_i p_i(x)}{\sum n_k p_k(x)}$
- Close to optimal
 - tight variance bounds by *Veach and Guibas [1995]*
 - no other strategy can do much better
- ⇒ A de facto universal solution

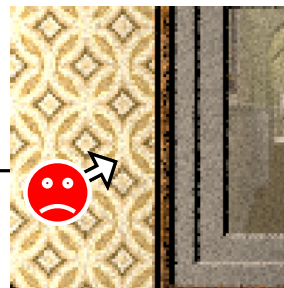
We start with the balance heuristic bounds. The balance heuristic was introduced by Veach and Guibas [1995] and together with the so called power heuristic it is very popular in Monte Carlo rendering. The balance heuristic weights are very easy to compute, as they are proportional to the sampling density times the number of samples. Apart from their simplicity, the authors of the original work were able to prove that these weights are also close to optimal. According to the tight variance bounds derived by Veach and Guibas, no other weights can achieve much lower variance.

For these reasons, the balance heuristic gained big popularity and has been used as a de facto universal solution.

MIS: INCREASES ROBUSTNESS, MAY REDUCE EFFICIENCY



+



MIS with the
balance heuristic



MIS with our
optimal weights

We illustrate the balance heuristics effect on a simple example. In the above staircase scene, we combine two techniques to render direct illumination. We can see that one is almost perfect on the wall above the stairs, while the other performs poorly there.

And in such settings, unfortunately, MIS with the balance heuristic keeps some of the error of the worse technique. So, in more general terms, the balance heuristic improves robustness, but can reduce efficiency. Our research question was: can we somehow do better than that?

The simple answer: yes, we can. We call our solution the optimal MIS weights, because they are provably optimal for a given allocation of samples to each technique (and assuming independency of samples). Before delving into the details, let us point out that in the above example, these optimal weights keep the lower error of the almost perfect technique. Also the image overall has a ten times lower level of noise.

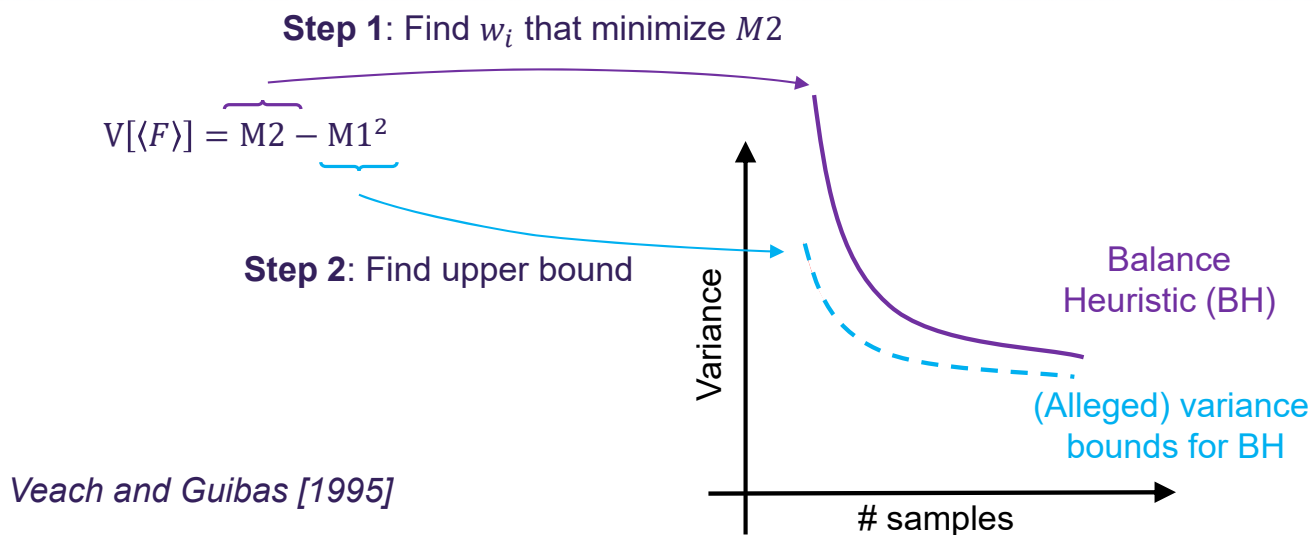
- Simple weighting functions: $w_i(x) = \frac{n_i p_i(x)}{\sum n_k p_k(x)}$
- Close to optimal
 - ~~tight variance bounds by Veach and Guibas [1995]~~
 - ~~no other strategy can do much better~~
- ⇒ A de facto universal solution

We show that it does not hold!

But what about our 10x speedup?

If we return to what has been said about the balance heuristic a bit earlier, we can wonder: How could we get this 10-times speedup when the balance heuristic was supposed to be almost optimal? The reason lies in the fact that the variance bounds do not hold! At least not in a fully general setting. This observation forms the first of our contributions.

VARIANCE BOUNDS DERIVATION

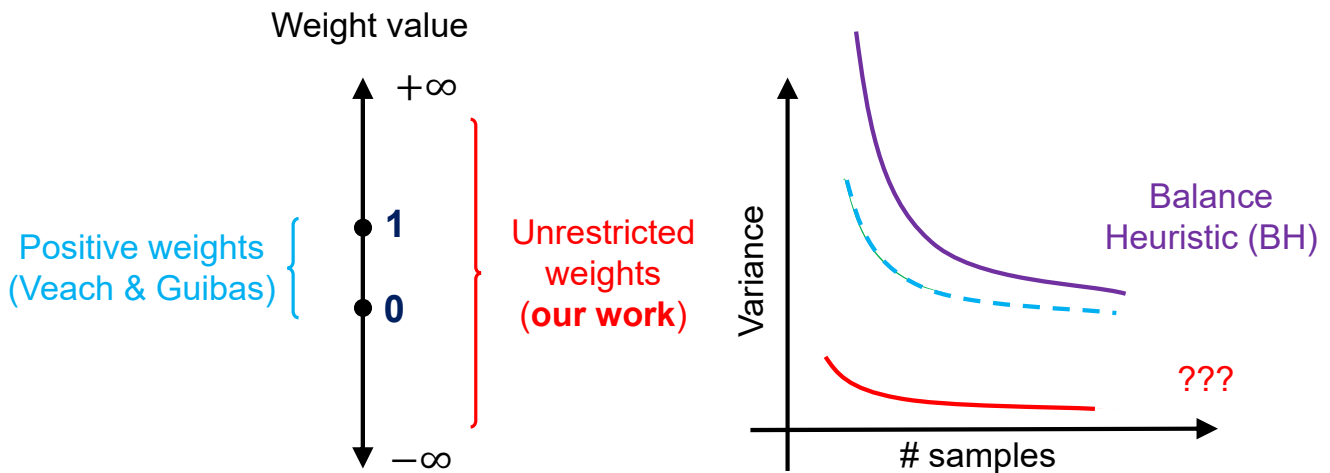


Veach and Guibas [1995]

When deriving the bounds Veach and Guibas considered the variance of an MIS estimator split into two terms, $M1$ representing the mean and $M2$ representing the second moment of an estimator. By minimizing the $M2$ part, they obtained the balance heuristic. We plot the variance of the resulting estimator vs. total number of samples on the right.

Then, the authors bounded the second term from above, which gave them a conservative estimate of how much better the 'best' possible weighting functions might be with respect to the balance heuristic. In other words no alternative weighting functions yield an estimator with variance below the dashed blue line in the graph.

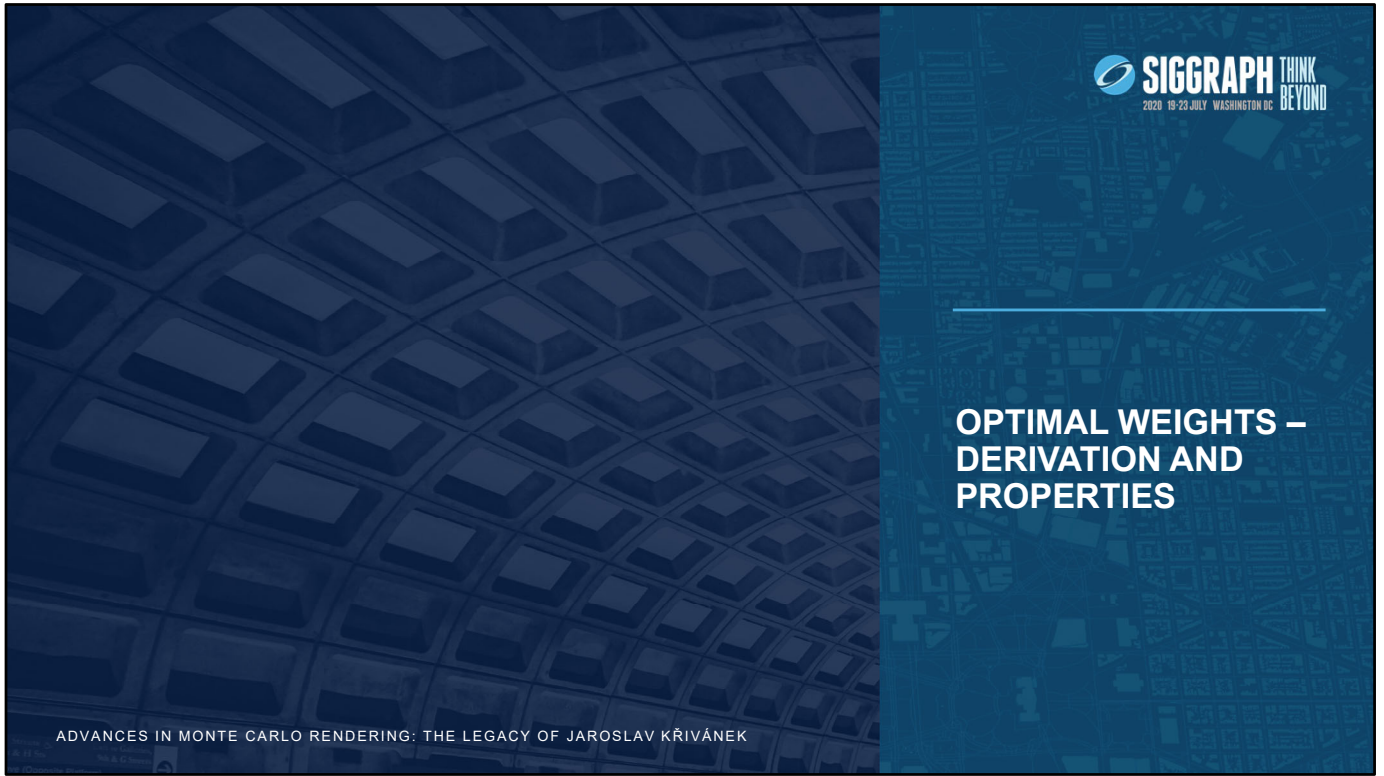
ASSUMPTION: POSITIVE WEIGHTS



However, we investigated their derivation further and realized that they assumed only positive weights are allowed, restricted to the interval 0 and 1. But the MIS framework allows for weights which are not restricted!

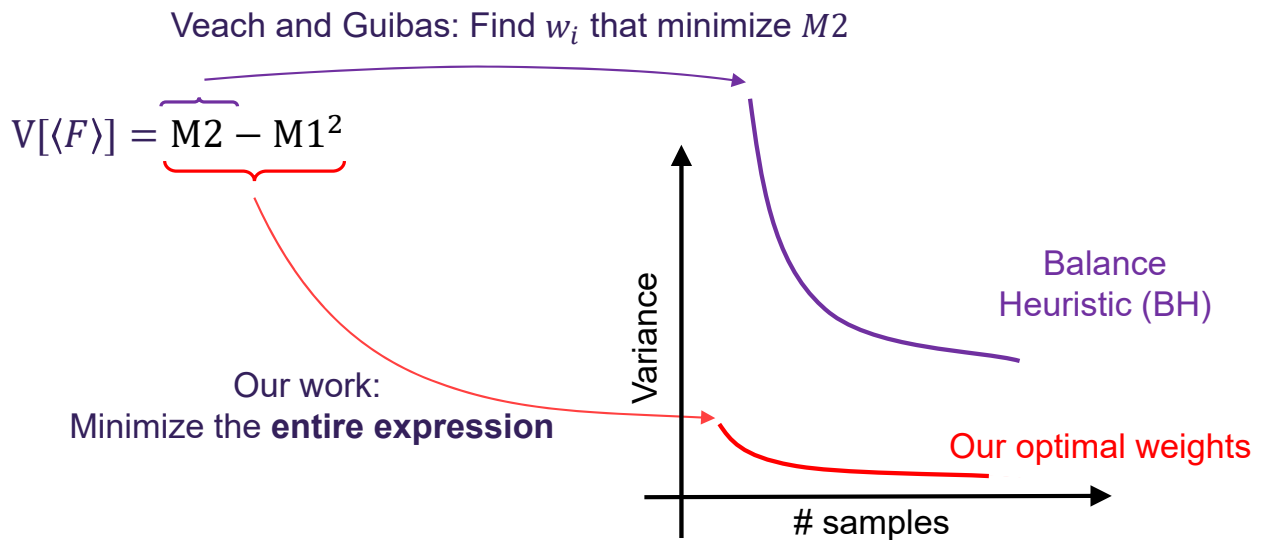
This simple fact has not been recognized up until now. Removing the restriction of positive weights invalidates Veach's bounds and it opens a possibility that the truly optimal MIS weights have much lower variance.

Now we know that the optimal solution can be much better than the bounds suggest. But how can we compute it?



In the following, we show how to derive the optimal weights.

OPTIMAL WEIGHTS DERIVATION



Our starting point is again the MIS variance formula. But instead of minimizing just the first part, we apply calculus of variations to minimize everything in terms of weighting functions.

That gives us provably optimal weights. These weights can have negative values, and in fact that happens in many cases. But they always sum up to one, which is the necessary condition to achieve unbiasedness within the MIS framework.

OPTIMAL WEIGHTS FORMULATION

$$w_i(x) = \frac{\alpha_i p_i(x)}{f(x)} + \frac{n_i p_i(x)}{\sum n_k p_k(x)} \left(1 - \frac{\sum \alpha_k p_k(x)}{f(x)} \right) \quad (1)$$

Balance heuristic

$$a_{ij} = \int \frac{p_i p_j}{\sum n_k p_k}$$

$$\underbrace{\begin{pmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NN} \end{pmatrix}}_A \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{pmatrix} = \underbrace{\begin{pmatrix} b_1 \\ \vdots \\ b_N \end{pmatrix}}_b \quad (2)$$

$$b_i = \int \frac{p_i f}{\sum n_k p_k}$$

$N = \#$ sampling techniques

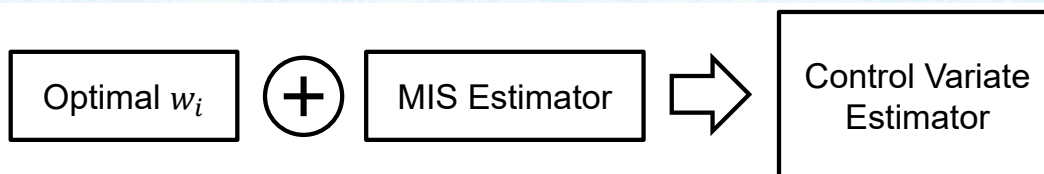
The optimal weights obtained from minimization then have the form of Eq. 1.

Note that they include the balance heuristic as part of their formulation. Also note that they include the integrand f itself in the denominators, which is very uncommon among combination strategies widely used. The formula also contains additional coefficients, which we denote alpha. There are as many of these coefficients as there are sampling techniques.

The alphas are the solution to the linear system in Eq 2. represented by a matrix A and a vector b , where the matrix A 's size is N -by- N and the vector b is a column vector of length N . The individual elements forming the matrix A resemble projections of sampling techniques onto themselves, and elements of the vector b resemble a projection of f into a system of sampling techniques.

To compute the optimal weights, we need to first compute and solve the linear system. The resulting alphas can then be used to compute the actual weights.

EQUIVALENCE TO OPTIMAL CONTROL VARIATES



$$\langle F \rangle = G + \langle F - G \rangle \quad (1)$$

The equation is accompanied by two graphs. The left graph shows a function f (dashed black line) and a control variate $g = \sum \alpha_i p_i$ (solid blue line) over a domain from 1 to 4. The right graph shows the residual function $f - g$ (solid red line) over the same domain. Arrows indicate the mapping from the terms in the equation to the corresponding parts of the graphs.

We also found that there is a relationship between the optimal MIS weights and optimal control variates.

Control variates are a variance reduction technique, where the control variate estimator $\langle F \rangle$ is obtained as a linear combination of the original estimator of F with a correlated estimator of G in such a way that the mean value does not change, as we can see in Eq. 1. We can also see it as estimating an integral of a linear combination of integrands f and g .

If we take the formula for our optimal weights and plug it into the formula for an MIS estimator, the resulting estimator also has the form of Eq. 1, which means that it is equivalent to a control variate. In our case, the function g is a linear combination of the sampling techniques, where the coefficients are the alphas.

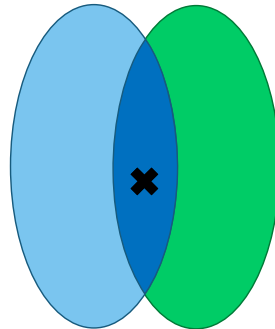
- **Owen and Zhou [2000]**

- **Setup**

- MIS + control variate
- Form $g = \sum \beta_i p_i$

- **Result**

- Optimal coeff. β_i



$$\beta_i = \alpha_i$$

- **Our approach**

- **Setup**

- MIS framework

- **Result**

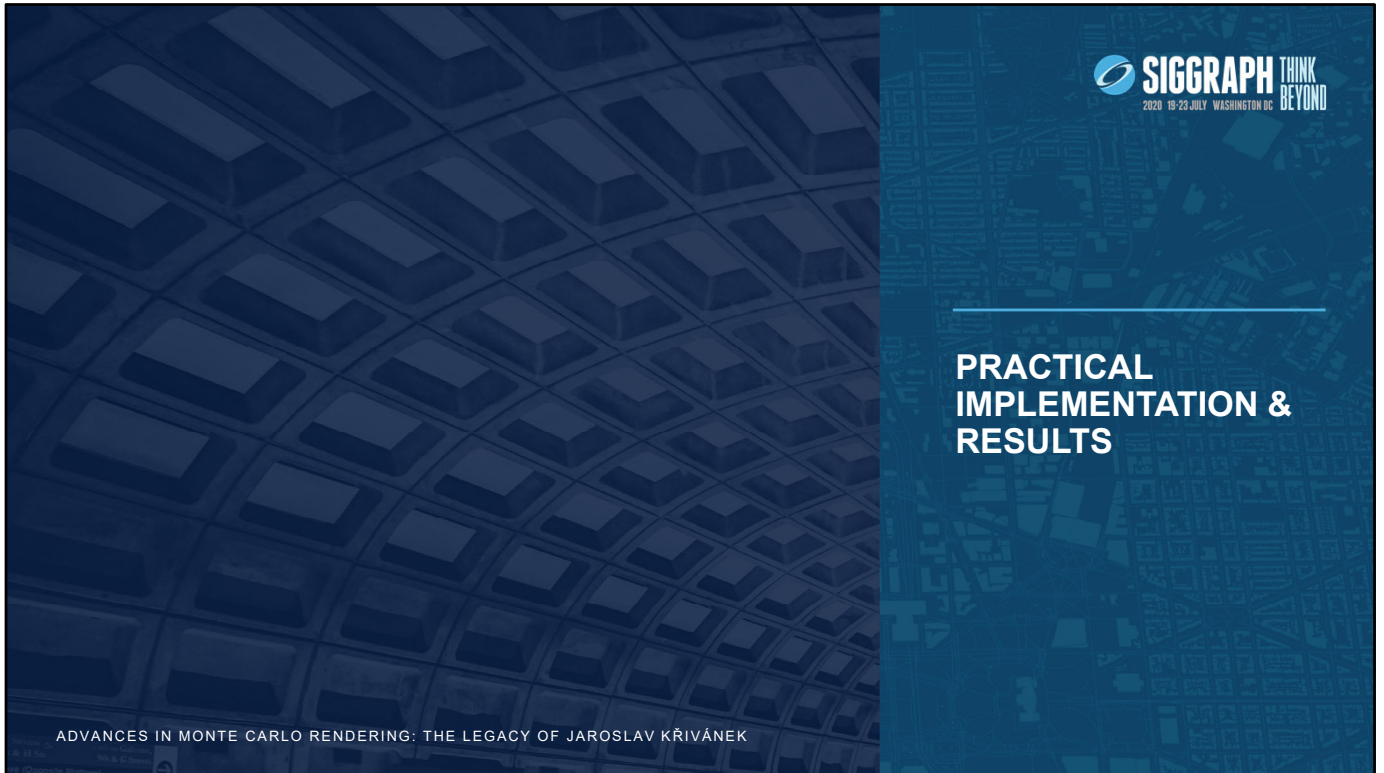
- Optimal MIS weights

Later, we found out that control variates of this form have been studied before, the most related work is by Owen and Zhou. And we found out that they reached the same result as we do, but from a different starting point (and with different implications).

They assumed the Balance heuristic used together with a Control variate (CV). Then they limited themselves to CVs formed as some linear combination of the sampling techniques and they found the optimal coefficients in this space. That by itself does not give any implications about MIS weights as such.

We, on the other hand, took the MIS framework and without any further assumptions we found the provably optimal MIS weights. Then we showed that all CVs of the linear combination form are equivalent to some MIS weights and that the optimal solutions to both problems are the same. Effectively we found the relation between two seemingly unrelated variance reduction techniques.

We believe this will allow borrowing theory from both ends to achieve further interesting insights, and that it will help steer the further investigation of MIS and Control variates.



Knowing the optimal solution in theory is one thing, practical use of that knowledge is another. Let us demonstrate that the optimal weights can be efficiently used in practice, in a light transport application.

HOW TO COMPUTE?

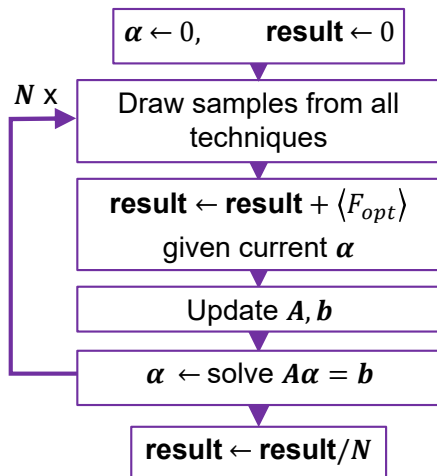
$$\begin{aligned} a_{ij} &= \int \frac{p_i p_j}{\sum n_k p_k} \\ b_i &= \int \frac{p_i f}{\sum n_k p_k} \end{aligned}$$
$$\underbrace{\begin{pmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NN} \end{pmatrix}}_A \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{pmatrix} = \underbrace{\begin{pmatrix} b_1 \\ \vdots \\ b_N \end{pmatrix}}_b$$

a_{ij}, b_i can be estimated from drawn samples

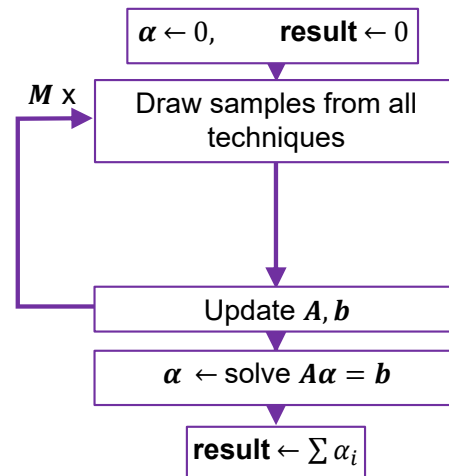
Recall the linear system we must solve to obtain the alpha coefficients in the optimal weights. The elements of A and b are defined as integrals, but they can be easily estimated from the samples we draw when computing the MIS estimator. For that we suggest two possible practical implementations.

HOW TO COMPUTE?

Progressive algorithm



Direct algorithm

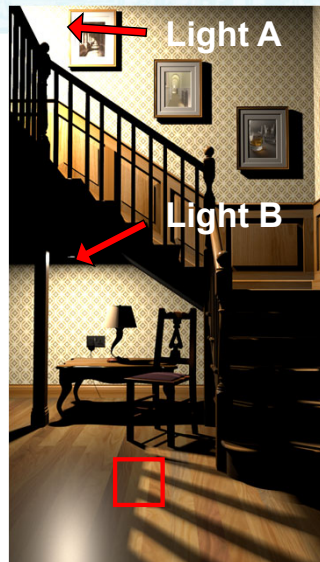


The first one is called Progressive. After the initialization, we first draw samples from all techniques. Then, we accumulate the MIS estimate using the optimal weights, computed with alphas estimated from all previously seen samples. We update the linear system and recompute the alphas. This is repeated several times, and finally, after leaving the loop, we return average of all the estimates.

The second approach how to implement the optimal weights is based on an observation that a sum of the estimated alphas also forms an estimator of the integral F . We call the resulting algorithm Direct. Here, instead of using the optimal weights formula for mixing the individual contributions, we just keep updating the linear system. After leaving the main loop, we solve the system for alphas and the result is then formed as their sum. While this algorithm is slightly biased, it is consistent and more efficient than the Progressive one.

We implemented and tested both algorithms and applied them to the problem of direct illumination. In practical terms, they mostly differ in a low sample count setting, and for high sample counts they have very similar properties. For that reason we will show only the results obtained by the direct algorithm (for further details see the original paper).

COMBINING STANDARD TECHNIQUES



Trained technique



Uniform technique



When using the optimal weights we explored two directions: applying them to standard techniques and designing new techniques.

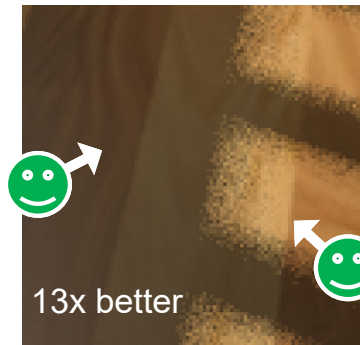
Let us start by showing the results for standard techniques. We will use the staircase scene from the very beginning, where the scene is illuminated by two lights, above and below the stairs. In a path tracer, whenever we are shading a point, we must randomly select one of the lights and compute its contribution. The light selection strategy has a strong impact on the result.

Suppose we have a technique which samples the lights according to their unoccluded contribution. We call this technique Trained. We can see it works nicely in places illuminated by both lights and much worse when light occlusion occurs. Occlusion dramatically influences the contribution of each light at the shading point.

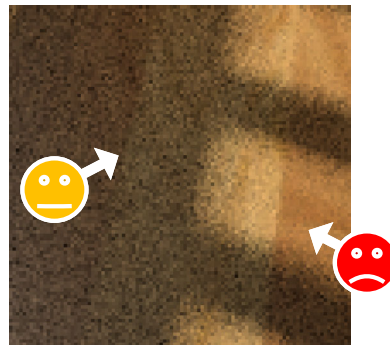
For the other technique, which distributes samples across the lights uniformly (and we call it Uniform), we see the opposite effect: it works much better in shadows.

COMBINING STANDARD TECHNIQUES

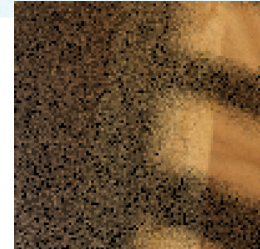
Our optimal weights



Balance heuristic



Trained technique



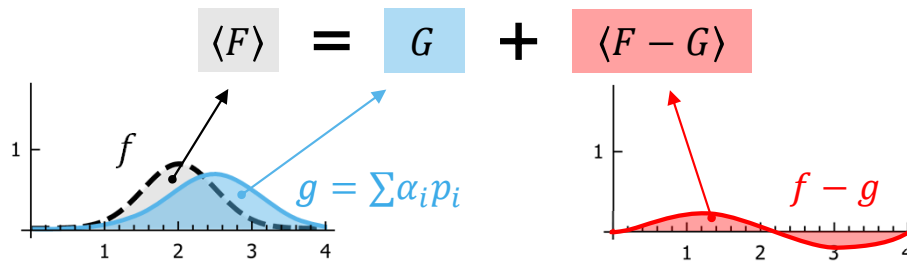
Uniform technique



When we combine the two respective techniques, using MIS with the balance heuristic, we obtain decent results, where we no longer see the excessive noise from the Trained technique in the shadow. But in unoccluded regions, where the Trained technique performed well alone, the result is now compromised by the uniform technique.

Using the optimal weights, we can see much better results. They even out-perform the individual sampling techniques where they performed well already. The reason behind this behavior is in the optimal weights acting as control variates.

REMINDER: OPTIMAL WEIGHTS ARE CONTROL VARIATES

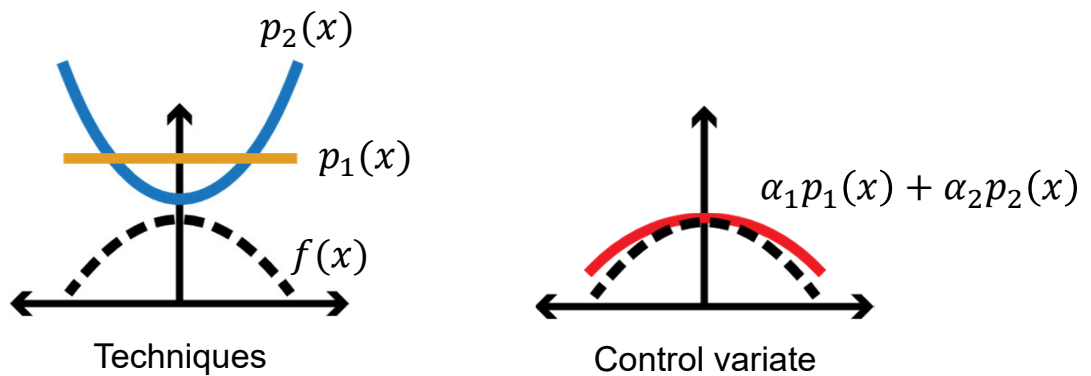


The closer $f - g$ to zero, the smaller the variance

The control variate functionality of the optimal weights then lead us to investigate usage of alternative sampling techniques, which nobody would normally think of in an MIS setting.

To recap, the optimal weights represent a control variate formulation, where a function g , formed by the linear combination of the sampling techniques, acts as a control variate for the integrand. And the closer the control variate g is to the integrand, the lower the variance will be.

INSIGHT: “BAD” TECHNIQUES CAN REDUCE VARIANCE

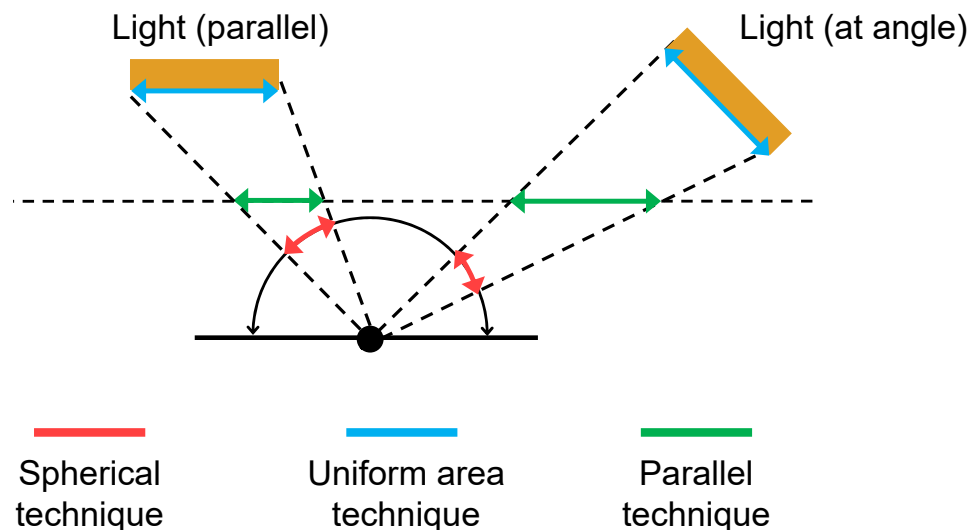


Therefore, we design sampling techniques in a way that improves their linear combination.

Consider this illustration where the dashed line is the function f that we want to integrate. For that, we have a uniform sampling technique, shown in orange. To improve the expressive power of the linear combination, we add the blue technique. For importance sampling, this technique would be a horrible choice.

For a control variate, however, we achieve a close to perfect linear combination, the red line on the right. Now we will apply this idea in a rendering context.

CAN WE DO EVEN BETTER?



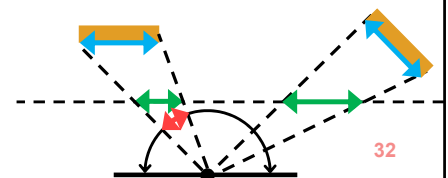
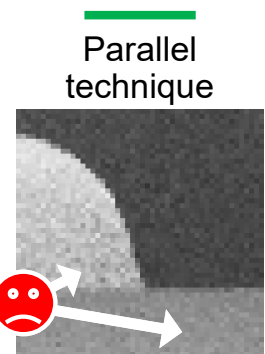
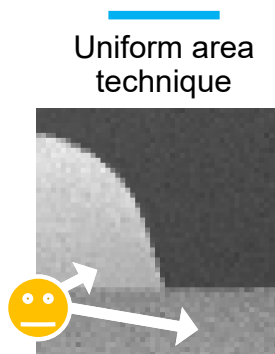
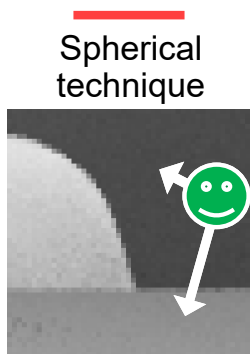
Consider the sketched example of two Lambertian area lights, positioned parallel to the surface or at an angle. We want to compute the direct illumination from those lights onto a diffuse surface.

Also consider the following sampling techniques:

- the spherical technique, which samples uniformly the light's projection onto the hemisphere,
- the uniform area technique, which samples uniformly the light's surface. Note that the uniform area technique is non-uniform when expressed on the spherical domain.

For a light which is parallel to the surface, a linear combination of the Uniform and Spherical techniques can compensate for the cosine geometry factor at the surface. This assumption breaks, when the light is at an angle. Then, the spherical and uniform area techniques will be similar, and their linear combination will have less expressive power. To achieve a similar effect as before, which would work regardless of light orientation, we introduce a completely new technique: sampling the *parallel projection* of the light source area. For a light which is parallel to the surface it behaves identically to the uniform-area technique and otherwise it tries to sample a proxy light parallel to the surface.

ALONE, THE PARALLEL TECHNIQUE IS EVEN WORSE

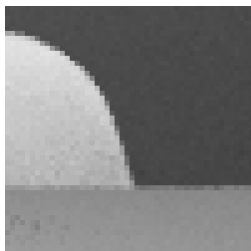


Now, let us compare these techniques in a simple scene, the dining room, which is illuminated by a single rectangular light above and parallel to the table.

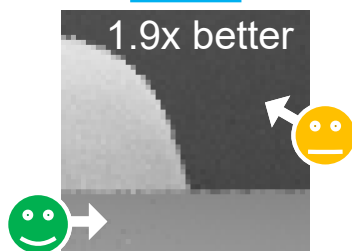
We can see that the spherical technique alone produces a nice image overall, while the uniform area sampling has a higher level of noise throughout. We can also see that the parallel projection is not a sensible technique on its own. The level of noise is significantly higher than with either of the other two techniques.

WITH OUR OPTIMAL WEIGHTS: EVEN BETTER

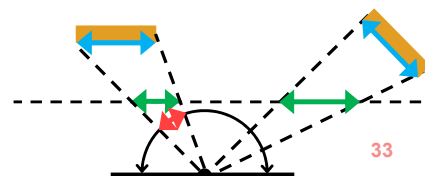
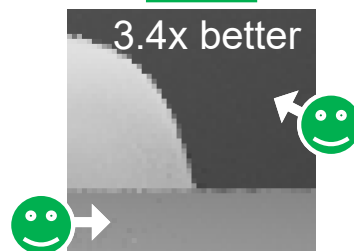
Spherical
alone
(**baseline**, 20spp)



Spherical
+
Uniform area



Spherical
+
Parallel



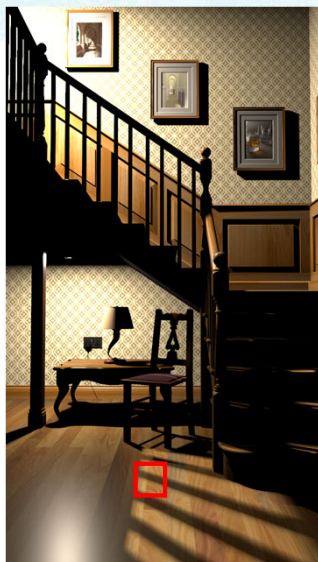
Now we show their combination using the optimal weights. Instead of taking 20 samples per pixel from the Spherical technique alone, we replace half of them by samples from either the Uniform or the Parallel technique. And, as expected, the combination of Spherical and Uniform yields better results than using Spherical technique alone on the surfaces parallel to the light. The combination of Spherical and Parallel is even better and improves also on surfaces which are at an angle with the light.

LIMITATIONS

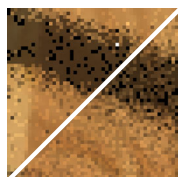
ADVANCES IN MONTE CARLO RENDERING: THE LEGACY OF JAROSLAV KŘIVÁNEK

The optimal weights have also their drawbacks and limitations.

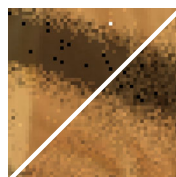
LIMITATIONS: SALT & PEPPER FOR FEW SAMPLES



Optimal weights / Balance heuristic



2 spp



4 spp



8 spp



16 spp

For example, when using either the progressive or direct version of our algorithm, we can observe salt and pepper noise for very low sample counts. That is caused by instability of the linear system we need to solve for the alphas. This type of noise can be easily denoised if such low sample counts are really needed.

LIMITATIONS: OVERHEAD

- Overhead for a large number N of techniques

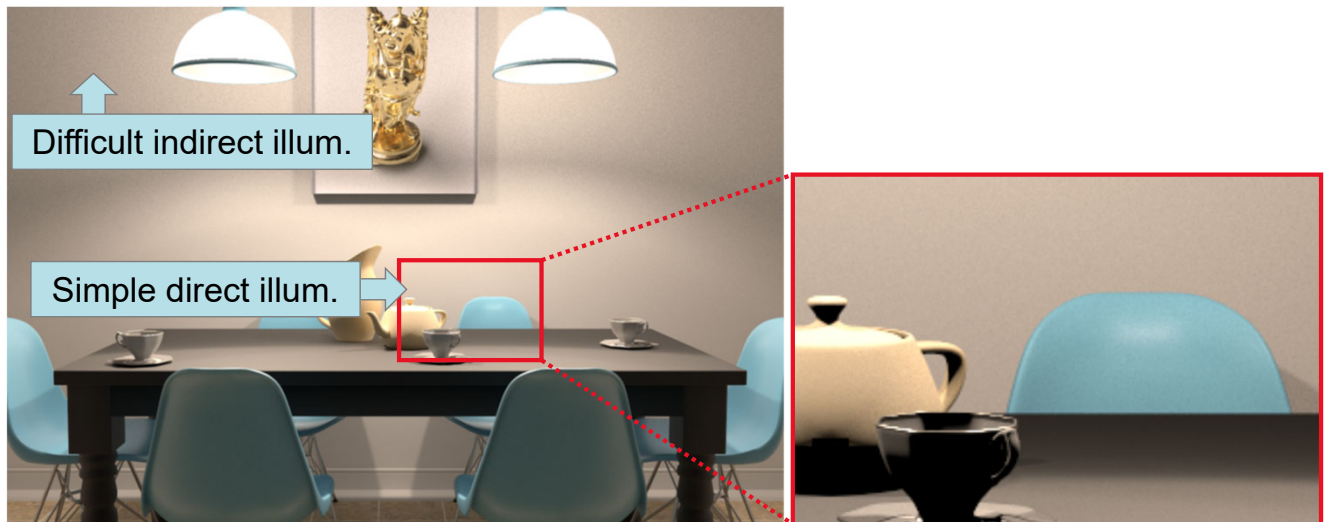
$$\underbrace{\begin{pmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NN} \end{pmatrix}}_{N \times N \text{ matrix}} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_N \end{pmatrix}$$

Another issue is the overhead when using the optimal weights, as the linear system complexity is quadratic with the number of sampling techniques used. This is relevant for example in a bi-directional path tracer, where for each path length we have corresponding number of techniques which need to be combined.



As a cheaper, albeit not optimal, alternative for more complex applications like bidirectional path tracing, we discuss a method that enhances the balance heuristic with variance estimates.

A USE-CASE: BIDIRECTIONAL PATH TRACING



ADVANCES IN MONTE CARLO RENDERING: THE LEGACY OF JAROSLAV KRIVÁNEK

38

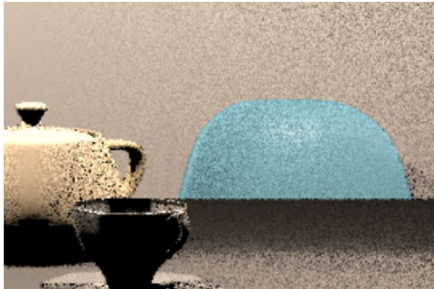
We will primarily focus our discussion on one common application of MIS: bidirectional estimators that combine paths from the camera with paths from the light sources.

The motivation to use such bidirectional methods, for example the classical Bidirectional Path tracer (BDPT) [Veach and Guibas 1995a] [Lafortune and Willems 1993], is visible in this scene.

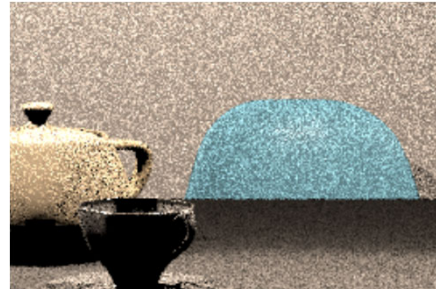
A path tracer, which starts from the camera, is quite good at rendering effects like the direct illumination on the table of this scene, but not so good at the difficult indirect illumination on the wall. Tracing paths from the light sources can help with the indirect illumination. Hence combining both should produce a robust algorithm.

This scene is a very good example for a scene where bidirectional path tracing is particularly useful.

A FAILURE CASE: BIDIRECTIONAL PATH TRACING



Path tracer



Path tracer + **Bidir.**
(Balance heuristic)

Unfortunately, the scene is also an example where MIS for bidirectional path tracing goes wrong. The two zoom-ins compare the rendered result of just the path tracer (left) and the combined method (right).

The balance heuristic combination contains the exact same samples that the path tracer is using, and some additional bidirectional samples. In other words, the samples on the left are taken and some additional work is done on top of them. The reward? Significantly higher levels of noise in the simple direct illumination.

OUR CONTRIBUTIONS



- Variance reduction effects ignored by MIS
 - Example: stratification

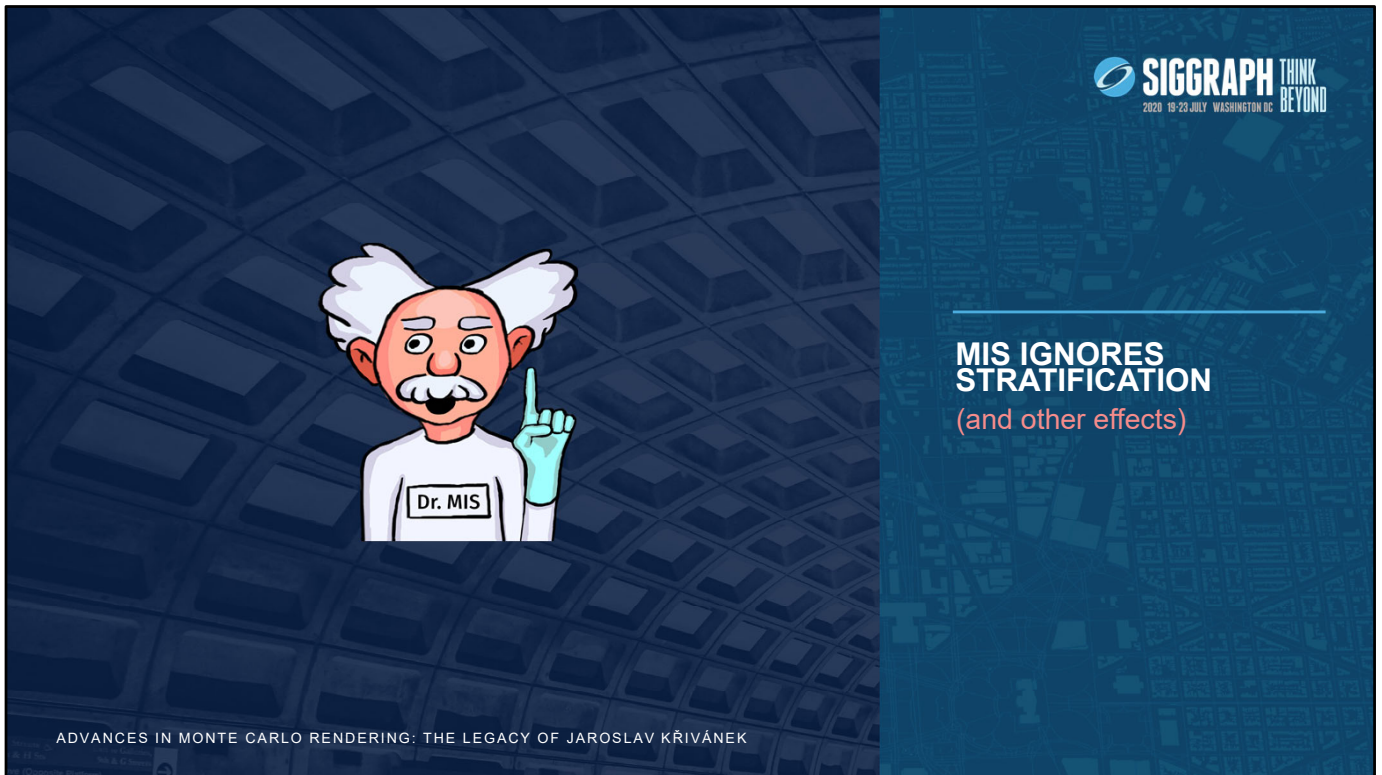


- A new heuristic: MIS + variance estimates

In the following, we discuss our approach to fix this regression.

First, we analyze what is causing the issue. As it turns out, there is a number of variance reduction techniques that are completely ignored by the balance heuristic. We show that the poor behavior in the previous example can be traced back to the balance heuristic's disregard for sample stratification.

We propose a simple yet effective trick to rectify the problem: Enhancing the balance heuristic, or any other MIS heuristic for that matter, with variance estimates.

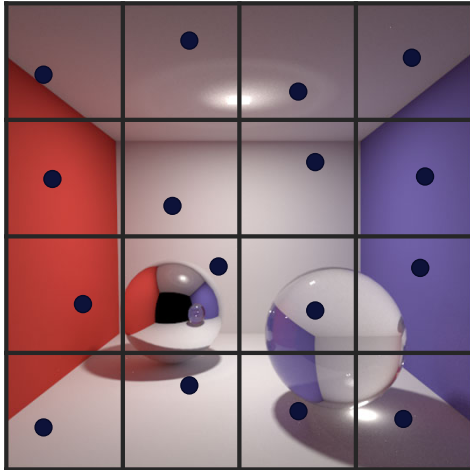


There are variance reduction methods which are ignored by the balance heuristic. This leads the balance heuristic to “believe” that one technique in the mix has a higher variance than it actually does. As a result, that technique receives a too low weight, harming the overall result if it is, in fact, the best technique for a certain effect. Just like the path tracer was the best technique for the direct illumination in the previous example, yet the weight it received by the balance heuristic was far too low.

Examples for such variance reduction methods are stratification and sample correlation, for example due to splitting. We will use stratification as an example in the following.

NO IMAGE PLANE STRATIFICATION (E.G. BDPT)

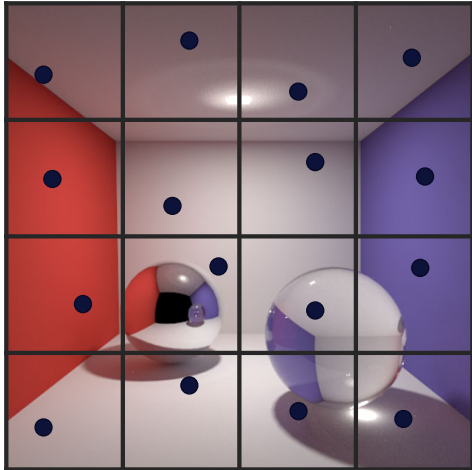
One sample per pixel



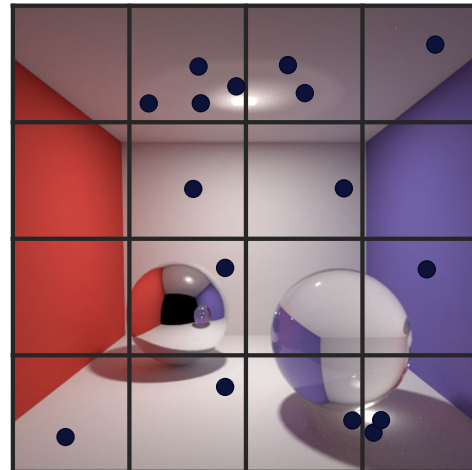
First, let us recap what sample stratification means. We consider the simple, yet very important, case of image plane stratification. If every pixel of an image receives exactly one sample, we say that we have achieved image plane stratification. This means that there is no randomness, and hence no variance, in whether a pixel receives any value at all. Almost all techniques that start paths from the camera can be naturally stratified on the image plane. This is one of the most important properties that makes those techniques so successful.

NO IMAGE PLANE STRATIFICATION (E.G. BDPT)

One sample per pixel



Sample over entire image

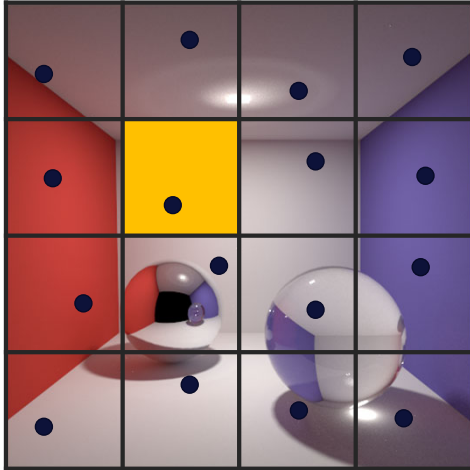


In contrast, if paths are traced bidirectionally, we often have no control over which pixels receive a value. We do not have image plane stratification. On the one hand, this additional uncertainty adds variance. On the other hand, distributing samples freely over the image can also reduce variance, because sampling can focus on caustics and other regions of focused illumination.

The lack of image plane stratification is the reason why bidirectional methods perform well for focused indirect illumination, like caustics. It is also the reason why they do not perform well for smoother illumination, like the direct illumination in the example shown in the beginning. It is a curse and a blessing.

EFFECTIVE DENSITY WITH STRATIFICATION

One sample per pixel



With uniform sampling
and equal pixel sizes

Effective density:

$$1 \times \frac{1}{\text{pixel_area}} = \frac{\text{num_pixels}}{\text{image_area}}$$

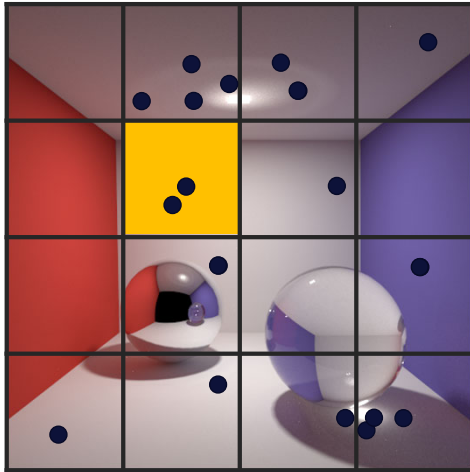
But why is (image plane) stratification a problem for the balance heuristic?

Let us consider a simple case, where samples are taken uniformly over the image, but stratified. One sample is taken in each pixel, uniformly over the surface area of the pixel.

The effective density, i.e., the term used by the balance heuristic, then boils down to the ratio of the number of pixels to the total surface area of the image.

EFFECTIVE DENSITY WITHOUT STRATIFICATION

Sample over entire image



With uniform sampling
and equal pixel sizes

Effective density:

$$\text{num_pixels} \times \frac{1}{\text{image_area}}$$

The same!

Now, let us pretend we had the exact same sample distribution, yet without the stratification. This can happen if the light tracer samples paths with a very similar distribution as the path tracer.

We still take the same number of samples as there are pixels in the image, but each is distributed uniformly over a larger domain, the entire area of the image. The resulting sampling density is the number of pixels, divided by the surface area of the image. The exact same result as the stratified version!

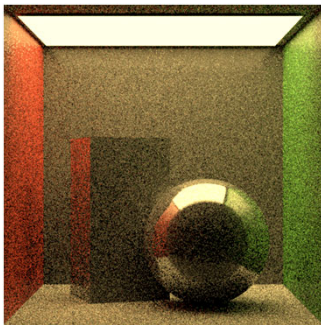
The balance heuristic uses the same values for both techniques when combining them. Unfortunately, the variance of the stratified version can be considerably lower than the unstratified one. The resulting combination can perform poorly, as we have observed in the example in the beginning.



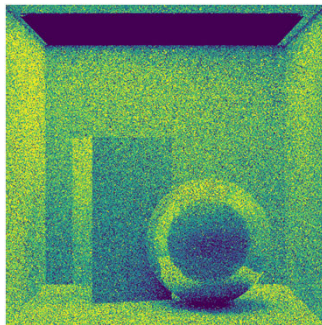
Before we discuss how to enhance the balance heuristic to better handle such combinations, we review an alternative approach to MIS: weighting each technique with a constant factor, namely its reciprocal variance.

COMBINING ESTIMATORS

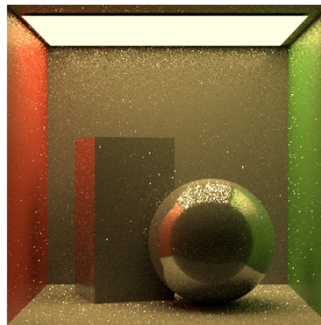
σ^2



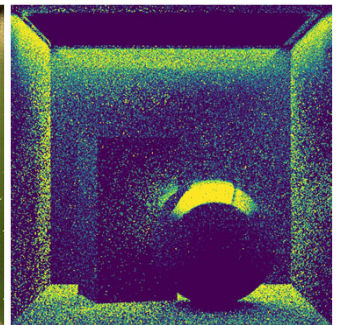
First technique



Estimated variance (σ^2)



Second technique



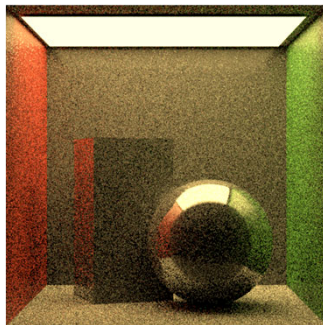
Estimated variance (σ^2)

Given two techniques, we can estimate their variances in a number of different ways. We might even have an analytical solution or approximation of the true variance.

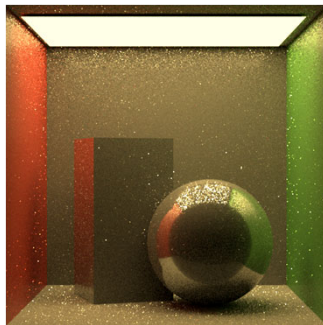
In this example scene, two techniques have very different variances in different regions of the image. Intuitively, it is apparent that we could combine the two, using the pixels from that technique that has lower variance locally.

Such combinations are, in fact, relatively common, especially in scenarios where MIS weighting is not possible. For example, because pdfs are unknown or storing individual samples is too costly.

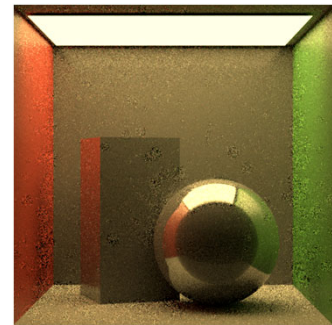
COMBINING ESTIMATORS



First technique



Second technique



Combined image

Indeed, using the variance based weighting approach in this simple example produces results that are on par with the balance heuristic combination.

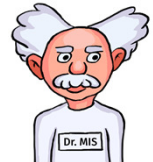
TWO SIMPLE HEURISTICS

Variance-based



- Coarse
- Relies on estimates
- + Accounts for all effects on the variance

Balance heuristic MIS



- + Fine grained
- + Uses exact values
- Ignores some variance altering effects

When comparing the variance-based weighting approach with MIS, there are three key differences.

Variance-based weighting is done globally, for example per pixel, whereas MIS can weight individual samples. This gives MIS an edge to further reduce variance in many cases.

Additionally, we often have to estimate variances. If the estimates are off, the resulting combination can be poor, or even show artefacts. The balance heuristic, on the other hand, relies solely on known exact quantities. The balance heuristic might not always produce the best image, but it will not produce artefacts either.

The important benefit of the variance-based approach, however, is that, by definition, it always considers all effects on the variance. It would not suffer the same problems as the balance heuristic when faced with, for example, differences in image plane stratification.



ADVANCES IN MONTE CARLO RENDERING: THE LEGACY OF JAROSLAV KŘIVÁNEK

So how can we achieve a better result? We use the same underlying idea of MIS: given two approaches that complement each other well, we combine them. In this case, we combine the balance heuristic with variance estimates. We do so in a fashion that retains the benefits of both methods.

VARIANCE AND THE BALANCE HEURISTIC

- The optimal weights minimize the variance of each technique t
- The balance heuristic ignores a residual r_t

Minimized by optimal weights

$$\sigma_t^2 = \int_{\Omega} \frac{f^2(x)}{n_t p_t(x)} dx - r_t$$

Minimized by balance

There is different ways to enhance the balance heuristic with variance estimates. To find the best one, we investigate the variance of a single technique t , σ_t^2 , as shown on this slide.

Optimally, we would like to find MIS weights that will minimize this full variance for every single technique. The balance heuristic only considers the integral term, the second moment of the primary estimator, divided by the number of samples. This term is an upper bound for the variance of any Monte Carlo estimator. There is, however, always some residual term r_t that will be ignored.

THE RESIDUAL DIFFERS BETWEEN TECHNIQUES

- Simplest case: $r_t = \frac{1}{n_t} \mu_t^2$
- With stratified samples: $r_t = \sum_{\text{strata}} \mu_{t,s}^2$
- With correlated samples: $r_t = \text{Cov} + \frac{1}{n_t} \mu_t^2$

Unfortunately, the ignored residual also differs between techniques. In the simple case of independent, unstratified samples, it is merely the squared ground truth divided by the number of samples. With stratification, it is the sum of the squared mean values of all strata. With correlated samples, it also contains some covariance term. If we had a combination that contained all three kinds of techniques, results are likely to be poor when using the balance heuristic.

This is the mathematical reason for the observation that we have made before. The balance heuristic ignores effects like image plane stratification, because these only change the residual term.

WHEN IS THE BALANCE HEURISTIC OPTIMAL?

- When the residual is tiny:

$$\int_{\Omega} \frac{f^2(x)}{n_t p_t(x)} dx \gg r_t$$

- i.e., **when variance is high**

There is also another interesting observation to be made here. Whenever the residual term is tiny in comparison, the balance heuristic will perform well. This happens whenever the variance of all techniques is relatively high.

In other words, if no technique in a combination clearly outperforms the others, the balance heuristic will perform reasonably well, maybe even optimally.

- We compute the ratio between the considered term and the full variance:

$$v_t = \frac{\int_{\Omega} \frac{f^2(x)}{n_t p_t(x)} dx}{\sigma_t^2}$$

- And modify the balance heuristic with it:

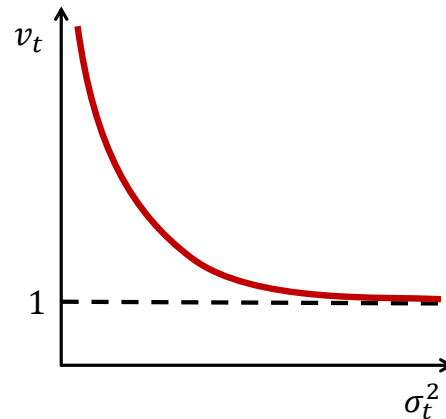
$$w_t(x) = \frac{v_t n_t p_t(x)}{\sum_k v_k n_k p_k(x)}$$

We now use these observations to derive a factor that we can plug into the balance heuristic weight. We compute the ratio of the term that the balance heuristic minimizes and the full variance. This ratio will be one if the balance heuristic minimizes the full variance. It will be very large if the balance heuristic considerably overestimates the actual variance.

More details on the derivation can be found in the paper [Grittmann et al. 2019].

BEHAVIOUR OF OUR WEIGHTS

- Low variance: increase weight
 - Where balance is known to perform poorly
- High variance: balance heuristic
 - Where balance works well



If the variance of a technique is high, our additional factor quickly decays to one. That is, we will produce the exact same weighting as the balance heuristic in such a case.

In contrast, if the variance is low, we considerably increase the weight assigned to that technique.

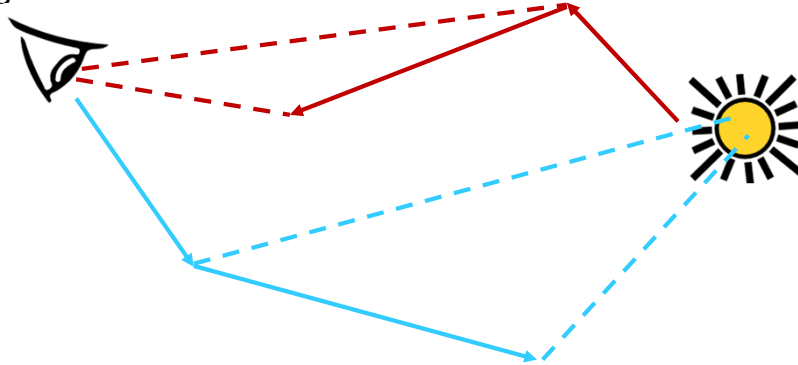
Previous work, even when first introducing the balance heuristic [Veach & Guibas 1995], has observed that the balance heuristic assigns too low weights for techniques that have low variance. Our method also rectifies that shortcoming.



Now, let us see how the resulting weights fare when used in practice. We have implemented the weights in a bidirectional path tracer. Variances are estimated from a single sample per pixel, at no measurable overhead. Details on the implementation can be found in the paper [Grittmann et al. 2019].

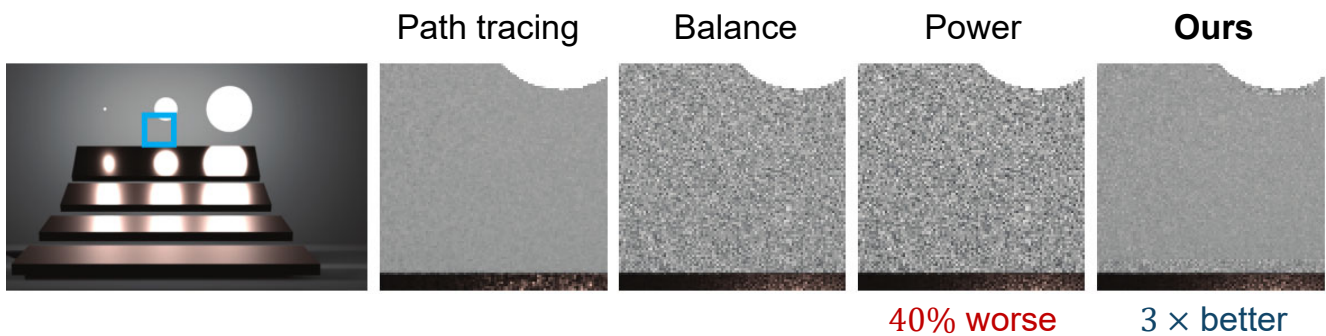
BIDIRECTIONAL PATH TRACING

- Path tracing is stratified on the image plane
- Light tracing is not



In a bidirectional path tracer, we are tracing paths from the camera, which are stratified on the image plane. These paths are combined with paths from the light sources. The paths from the lights might be stratified across the lights, but they are not stratified over the image plane: it cannot be guaranteed that every pixel receives one.

RESULTS: IMPROVEMENT IN FAILURE CASES



We can see what problems this might cause, in this classical example scene for MIS. The crops on the right compare results on the diffuse wall behind the lights.

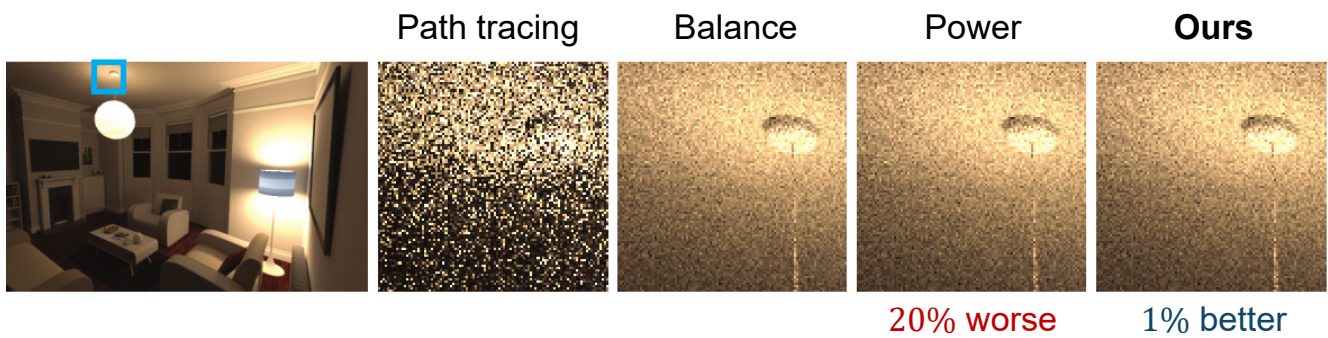
The path tracer samples the direct illumination there with almost no variance, thanks to the image plane stratification.

The bidirectional path tracer, using the balance heuristic, produces a significantly worse image by assigning too high weights to the light tracer.

One might be tempted to use the power heuristic instead. After all, the power heuristic was also introduced to fix cases where one technique has very low variance. Unfortunately, the power heuristic performs even worse than the balance heuristic here, by assigning even higher weights to the unstratified light tracer.

Our weights ensure robustness, producing the same result as the path tracer alone.

RESULTS: NO HARM DONE OTHERWISE

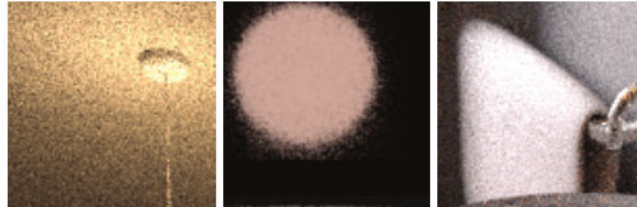


This scene features focused direct illumination, which the light tracer is quite good at. This time, the path tracer has higher variance and the balance heuristic performs quite well. Interestingly, the power heuristic yet again performs worse than the balance heuristic, this time by assigning too high weights to the path tracer.

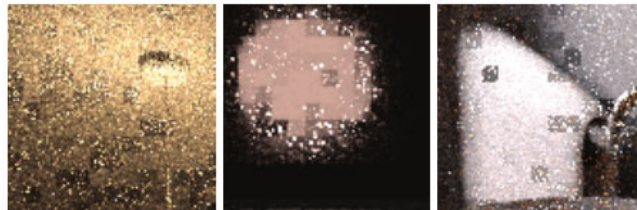
Our variance factors in this case are almost one, producing the same result as the balance heuristic with some marginal improvements.

RESULTS: MORE ROBUST THAN VARIANCE ALONE

Ours



Variance-only



Robustness is an important feature of our weights. Here, we compare results to a purely variance-based approach, where we are using variance estimates from a large number of samples per pixel.

Despite being based on the same variance estimates, our method shows none of the artefacts while also improving upon the balance heuristic.

SOURCE CODE

- <https://github.com/pgrit/var-aware-mis-pbrt>



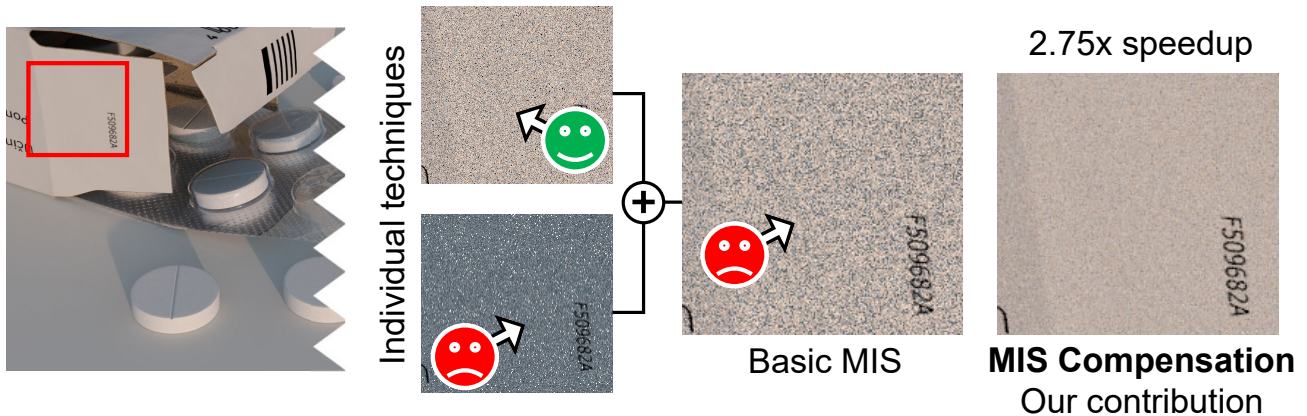
You can find our PBRT implementation by following the link or scanning the QR code.

MIS COMPENSATION

[Karlík et al. 2019]

ADVANCES IN MONTE CARLO RENDERING: THE LEGACY OF JAROSLAV KŘIVÁNEK

MIS: INCREASES ROBUSTNESS, MAY REDUCE EFFICIENCY



In the previous part of this course, we were optimizing the weights of a given, fixed combination of techniques.

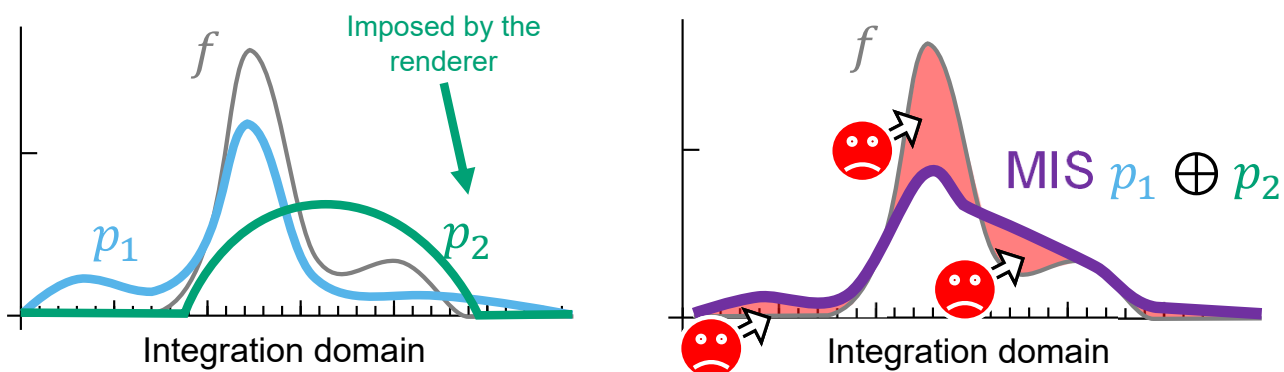
In this part of the course, we will tackle MIS improvement from a slightly different angle. Unlike the other approaches, that try to tweak how sampling techniques are combined, we directly modify one of the sampling techniques, so that the MIS combination performs better.

OUR APPROACH: 3 LINES OF CODE

```
void MIS_compensation()  
{  
    for (int i = 0; i < N; ++i) {  
        probability[i] = max(probability[i] - averageValue, 0.f);  
    }  
}
```

We will show that this not only allows us to significantly speed up rendering, but we will be able to achieve this with just three lines of a simple code.

OUR APPROACH: MOTIVATION

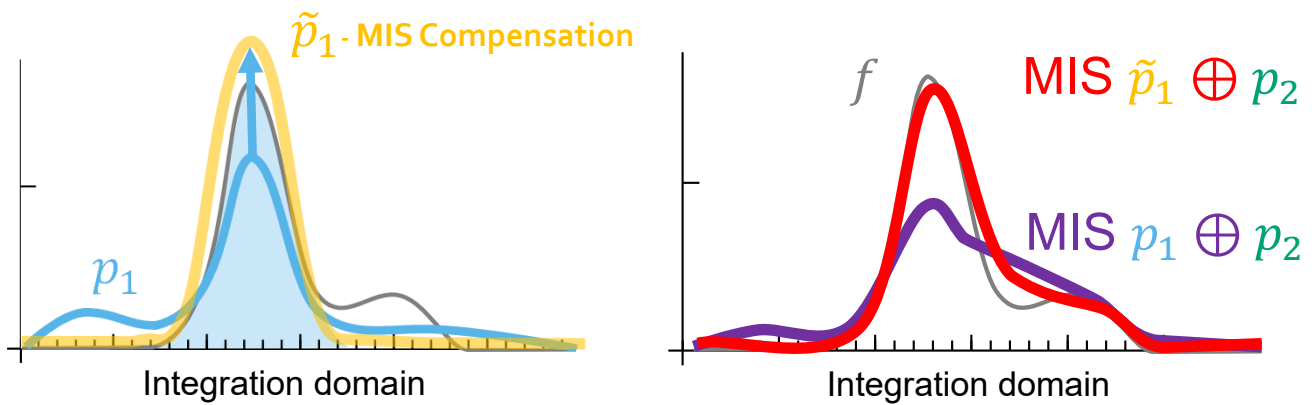


Let us illustrate our insight on a simplified integration problem depicted on the left, where we want to integrate f over an integration domain.

In rendering, it often happens that we have one technique p_1 , that is in many regions similar to the integrand, but under-samples the integrand in some other regions. We also have a second technique p_2 , imposed by the renderer, which is not possible to modify. As a practical example of p_2 we can take sampling according to the BRDF, as it is often used for more than one purpose, like computation of both direct and indirect illumination.

The MIS combination of these two techniques (depicted on the right), using the standard balance heuristic and the same number of samples for each technique, will give us a combined density shown in purple. And as we can see, the resulting combined pdf is far from perfect – it still oversamples some parts of the integrand and, consequently, under-samples others.

OUR APPROACH: MIS COMPENSATION

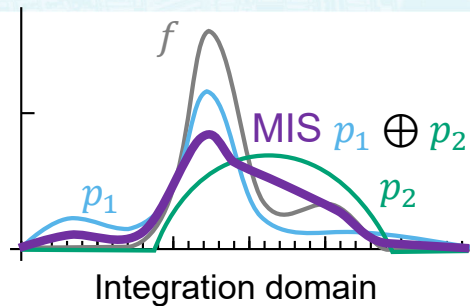


To improve this, we first realize that the technique p_2 , imposed by the renderer, already samples some parts of the integrand well. What we would like to do is to modify the sampling technique p_1 to focus more on parts that are under-sampled by the other technique, here highlighted in blue.

And that is exactly what our method, MIS compensation, does – it modifies one of the sampling techniques with respect to MIS. Thanks to MIS compensation, the resulting pdf, here drawn in red, is a much closer match to the integrand than when using the unmodified sampling techniques. Using MIS compensation thus decreases the estimator's variance.

OUR APPROACH: INTUITIVE FORMULA

1.
$$\langle F \rangle = \frac{f(x)}{\underbrace{\frac{1}{2}p_1(x) + \frac{1}{2}p_2(x)}_{\text{MIS}}}$$



2.
$$\int f(x) = F = \frac{f(x)}{\frac{1}{2}\tilde{p}_1(x) + \frac{1}{2}p_2(x)}$$

OPTIMIZE

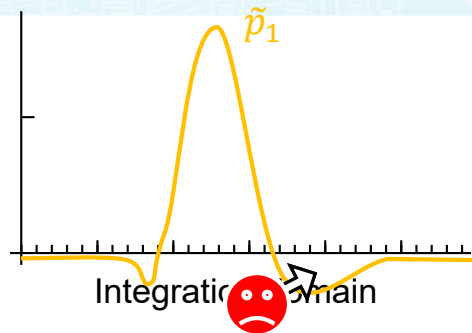
Now let us discuss how exactly we apply MIS compensation. The modification of the sampling technique can be derived intuitively. We formulate the problem as follows:

- We start with a one-sample MIS estimator, using the balance heuristic with an equal number of samples allocated to each technique. The denominator of the estimator then corresponds to the combined pdf given by the balance heuristic.
- Our goal is to, ideally, obtain a zero variance estimator – meaning that the estimate will be equal to the integral for any number of samples.
- We want to achieve that goal by making p_1 a free function to optimize.

What is the solution for \tilde{p}_1 , under this problem setup?

OUR APPROACH: INTUITIVE FORMULA

$$\tilde{p}_1(x) = 2 \underbrace{\frac{f(x)}{F}}_{\text{Ideal PDF}} - \underbrace{p_2(x)}_{\text{Compensation}}$$

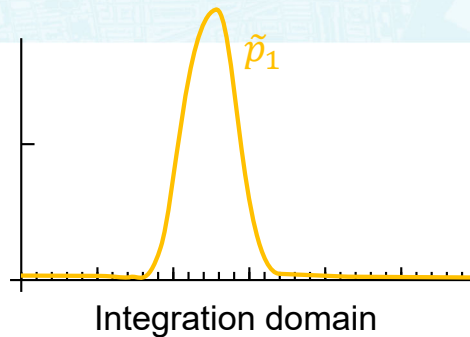


- Invalid PDF: unnormalized and/or negative

Simple algebra yields a formula for \tilde{p}_1 , where we subtract from the ideal zero-variance pdf the value of pdf of the fixed technique p_2 . This way we compensate for the MIS combination with p_2 . However, the formula can give us an invalid pdf – it can be unnormalized or even negative.

OUR APPROACH: INTUITIVE FORMULA

$$\tilde{p}_1(x) = \frac{1}{b} \max \left\{ 2 \frac{f(x)}{F} - p_2(x), 0 \right\}$$



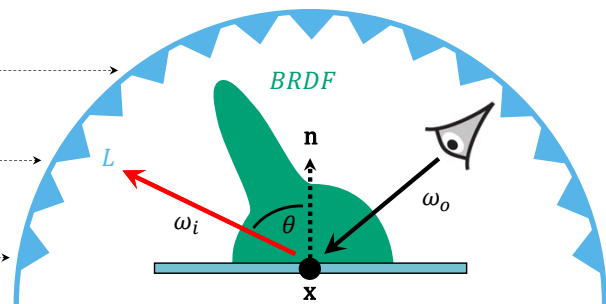
- Ensuring validity: **max** and normalization **b**
- Ad-hoc, but close to provable optimum

To ensure the resulting pdf is normalized and non-negative, we apply a max operator and renormalize, which gives us the final formula. While this last step in our derivation is ad-hoc, we show in our paper that this solution is close to the provable optimum.



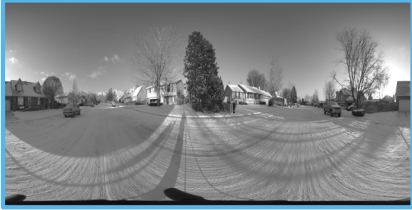
Let us illustrate one of the applications of our MIS compensation technique – Image based lighting.

$$F(\mathbf{x}, \omega_o) = \int_H L \cdot BRDF \cdot \cos\theta \cdot V d\omega_i$$



Realistic illumination can often be defined by an HDR texture that is spherically mapped around the scene. To compute the illumination at a point x as seen along the outgoing direction ω_o we take the emission from the HDR map coming from direction ω_i , multiply it by the BRDF, by the cosine at the surface, and by the visibility term. Then we integrate it over all incoming directions on the hemisphere.

$$F(\mathbf{x}, \omega_o) = \int_H L \cdot BRDF \cdot \cos\theta \cdot V d\omega_i$$



$p_1(\omega_i) \propto \text{HDR map emission } L$
(tabulated)



$p_2(\omega_i) \propto BRDF \cdot \cos\theta$
(analytical)

This integral is typically estimated using an MIS combination of two techniques for sampling the incoming direction:

- One is proportional to the HDR map emission and is usually implemented as a tabulated pdf.
- The second is proportional to the BRDF-cosine product, and its pdf is given by an analytical formula.

To apply our MIS compensation, we optimize one of these techniques. Since modifying a tabulated pdf is simple, we choose to optimize the first technique.

OUR APPROACH

$$F(\mathbf{x}, \omega_o) = \int_H L \cdot BRDF \cdot \cos\theta \cdot V d\omega_i$$
$$\tilde{p}_1(\omega_i) = \frac{1}{b} \max \left\{ 2 \frac{f}{F} - p_2, 0 \right\}$$
$$p_2(\omega_i) \propto BRDF \cdot \cos\theta \quad (\text{analytical})$$

A new pdf is computed according to our MIS compensation formula, where F is the full integral, f is the integrand, and p_2 is the fixed second technique. While this solution is close to optimal, it is not yet practical. It contains the target integral value and depends on both the surface position x and the outgoing direction. It is, therefore, inefficient.

3 lines of C++ code

Simplifying assumptions:

- Diffuse BRDF
- Average over normals
- Ignore visibility

```
void MIS_compensation()
{
    for (int i = 0; i < N; ++i) {
        probability[i] = max(probability[i] - averageValue, 0.f);
    }
}
```

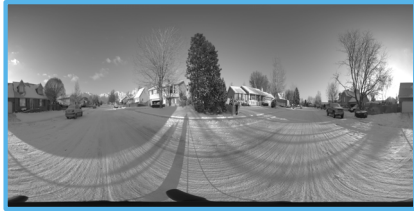
$$\tilde{p}_1(\omega_i) = \frac{1}{b} \max \{ L - \bar{L}, 0 \}$$

We arrive at a more efficient solution by adding three simplifying assumptions:

- We assume a perfectly diffuse BRDF,
- we average over all surface normals,
- and we ignore the visibility between the surface point and the HDR map.

This gives us a much simpler formula that corresponds to subtracting the average HDR map pixel value \bar{L} from each original pixel value L . This solution is not only efficient, but also trivial to implement.

OUR APPROACH

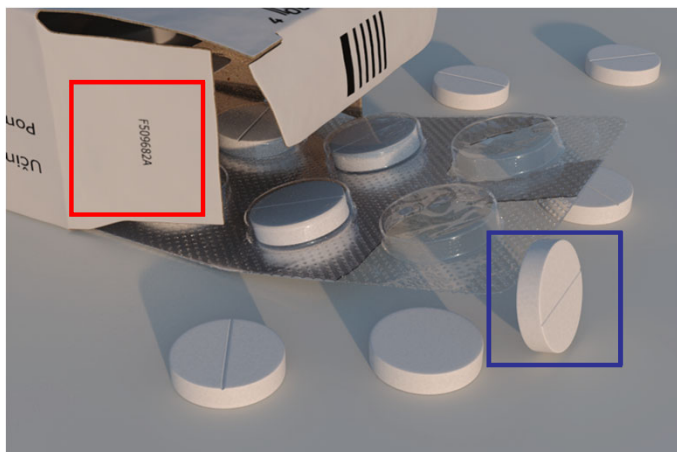


$$p_1(\omega_i) \propto \text{HDR map emission } L$$

$$\tilde{p}_1(\omega_i) = \frac{1}{b} \max \left\{ L - \bar{L}, 0 \right\}$$

Using this formula, the original pdf for sampling the HDR map is simply replaced by a pdf with much higher contrast. The previously bright parts have now even higher probability of being sampled and thus compensate for any under-sampling induced by MIS. Note that, due to this compensation, the resulting \tilde{p}_1 will often be further from the ideal sampling pdf than the original p_1 , when used on its own.

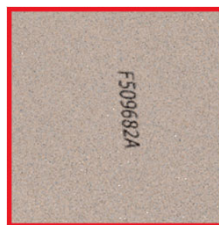
RESULTS: SIMPLE SCENE



Reference



Basic MIS
NMSE 3.38



Our method
NMSE 1.23 (2.75x)

pdf

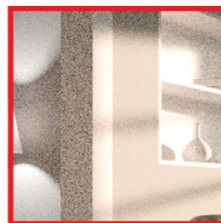
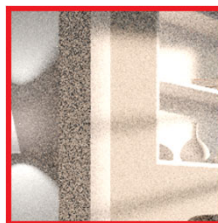
Equal-time comparison (5 s)

Let us show and discuss some of the results. Here, we have a simple scene that is lit by an HDR map that is mostly unoccluded. With our method, we can achieve a 2.75 speedup compared to the basic MIS combination with the original pdf.

RESULTS: OCCLUSION



Reference



Basic MIS
NMSE 1.05

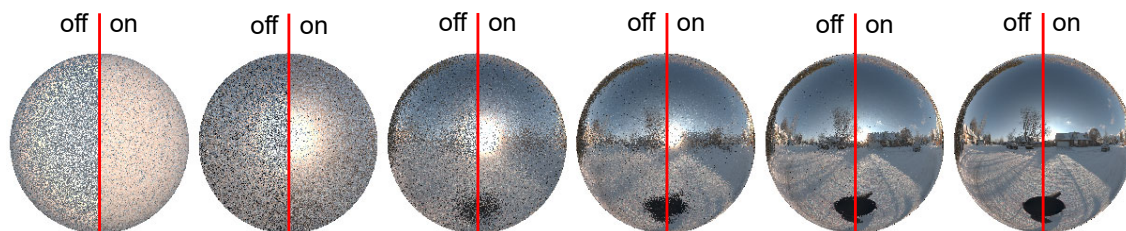
Our method
NMSE 0.6 (1.75x)

Equal-time comparison (50 s) pdf

In this interior scene, which is lit mainly through the window, the illuminating environment map is heavily occluded. Despite the occlusion, which we assumed to be not present, we can achieve a 1.75 speedup.

RESULTS: GLOSSINESS

Our method



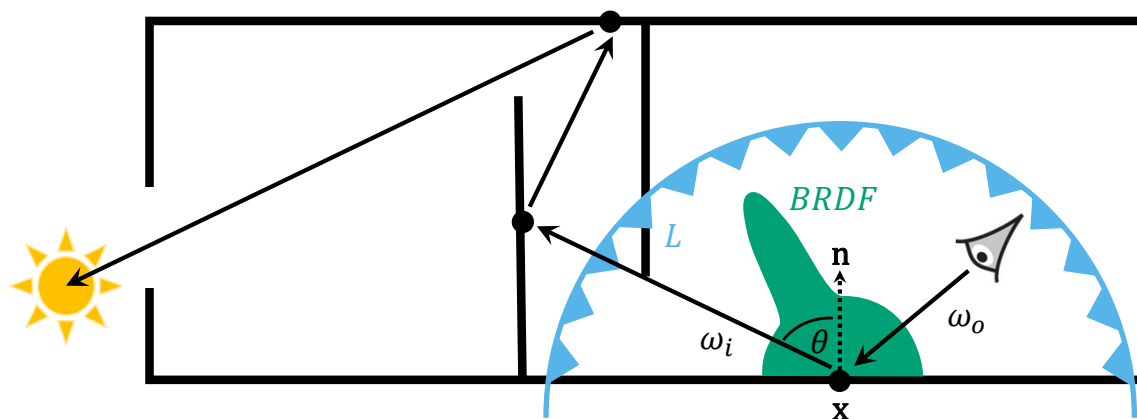
BRDF glossiness →

Another assumption during the derivation was that the surface was diffuse. The image above illustrates what happens when we break this assumption. We render a sphere illuminated by an HDR map and we modify the sphere's glossiness as we move to the right. We render half of the sphere using basic MIS (compensation is off), and half with our solution (compensation is on). As expected, the improvement provided by our method diminishes with increasing glossiness, but, and that is important, our method does not make the result worse.

In fact, in all our tests we have never encountered a case where our method would perform worse than basic MIS.



Let us now move on to the second application – path guiding.



Unlike image-based lighting that handles direct illumination only, path guiding concerns itself also with indirect illumination. That is technically the same as the direct one, except it is not coming from an HDR map, but it is reflected from all surrounding scene surfaces.

The indirect illumination at the point x is due to the direct and indirect illumination at all visible surfaces from that point.

MIS {

$$\tilde{p}_1(x, \omega_i) \propto L - \bar{L}$$

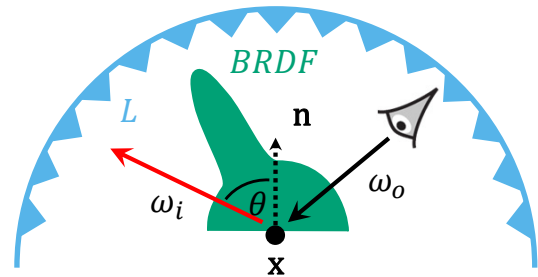
MIS compensation

$$p_2(x, \omega_i) \propto BRDF \cdot \cos\theta$$

analytical

$$p_1(x, \omega_i) \propto L$$

learned (tabulated) distribution
Müller et al. [2017]



To compute the indirect illumination, ordinary path tracing relies on just one sampling technique, which is proportional to the BRDF-cosine product.

Path guiding methods use an additional sampling technique. They sample proportionally to the illumination coming from the surrounding scene towards a given scene point. To achieve this, a path guiding method learns an approximation from the previous samples or from some preprocess. To ensure robust estimation, the two sampling techniques are usually combined using MIS. And this opens an opportunity to use MIS compensation.

To apply MIS compensation in this setting, we build up on a guided path tracer from the work of Müller et al. As they use tabulated pdfs to store illumination distribution, we have essentially the same setup as before in image-based lighting.

PATH GUIDING: RESULTS

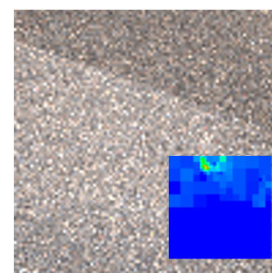
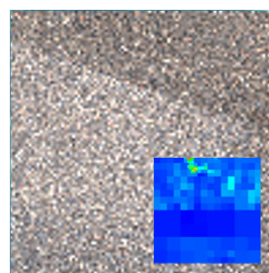


ADVANCES IN MONTE CARLO RENDERING: THE LEGACY OF JAROSLAV KRIVÁNEK

Equal – time (150 s)

Müller et al.
NMSE 3.07

Our method
NMSE 2.22 (**1.38x**)



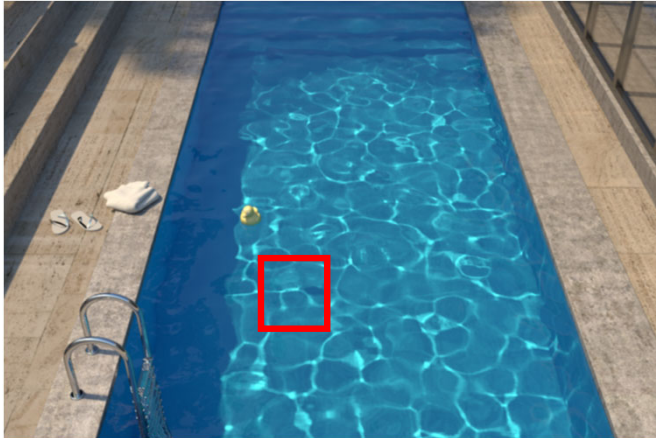
MIN

MAX

Now let's look at the results. First, we have an interior scene that features glossy materials and is lit only from the outside. We can see that the MIS compensation speeds up the rendering by a factor of 1.38 compared to the original path guiding approach with basic MIS.

We can also see the distribution of light in false color, as estimated at the middle of the inset region (left) and our MIS compensated version of that distribution. Blue color corresponds to the areas that are sampled with low probability, while green and red areas are sampled with higher probability. The MIS compensation modifies the light distribution to focus more on already highly sampled regions.

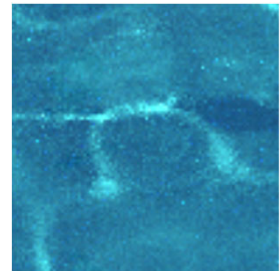
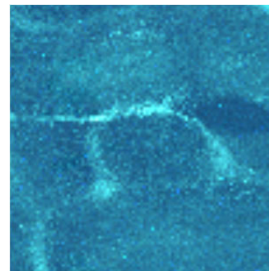
PATH GUIDING: RESULTS



Equal – time (150 s)

Müller et al.
NMSE 3.64

Our method
NMSE 2.27 (1.6x)



Here, we have another scene where guiding is used to generate caustics at the bottom of a pool. MIS compensation speeds up the rendering by a factor of 1.6. As in the previous application, we did not encounter any fail cases.

CONCLUSION & FUTURE WORK

ADVANCES IN MONTE CARLO RENDERING: THE LEGACY OF JAROSLAV KŘIVÁNEK

CONCLUSION

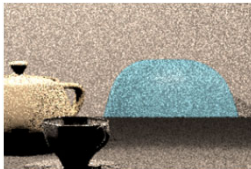
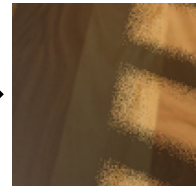
- MIS is **not** a solved problem: There is room for improvement!



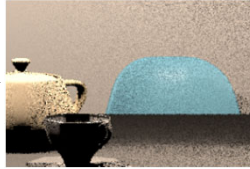
3x



13x



2x



MIS with the balance heuristic has been universally accepted as “the” solution, no questions asked. The three projects we have discussed in this part of the course demonstrate quite clearly that there is still a lot of room for improvements.

Firstly, the optimal weights can be negative and hence perform an order of magnitude better than the balance heuristic. Secondly, the balance heuristic can perform poorly in unexpected cases, like bidirectional methods. Enhancing it with variance estimates can help with those. Lastly, significant improvements can be achieved by modifying or designing sampling densities specifically to reap the most benefits in an MIS setting.

Robust and efficient algorithms, to our current knowledge, cannot be achieved without MIS. Investing into better MIS weights, sample allocations, or sampling techniques can yield great rewards.

References

- [Elvira et al. 2015] Víctor Elvira, Luca Martino, David Luengo, and Mónica F. Bugallo. 2015. Generalized multiple importance sampling. arXiv:1511.03095.
- [Elvira et al. 2016] Víctor Elvira, Luca Martino, David Luengo, and Mónica F. Bugallo. 2016. Heretical multiple importance sampling. *IEEE Signal Processing Letters* 23, 10 (Oct 2016).
- [Georgiev et al. 2012] Iliyan Georgiev, Jaroslav Křivánek, Stefan Popov, and Philipp Slusallek. 2012b. Importance Caching for Complex Illumination. *Comput. Graph. Forum (EUROGRAPHICS 2012)* 31, 2pt3 (May 2012), 701–710.
- [Grittmann et al. 2019] Pascal Grittmann, Iliyan Georgiev, Philipp Slusallek, Jaroslav Křivánek. Variance-Aware Multiple Importance Sampling. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2019)*, 38(6): 152:1-152:9, 2019.
- [Havran and Sbert 2014] Vlastimil Havran and Mateu Sbert. 2014. Optimal Combination of Techniques in Multiple Importance Sampling. In *Proc. VRCAI '14*. ACM, New York, NY, 141–150.
- [Jensen 1995] Henrik Wann Jensen. 1995. Importance Driven Path Tracing using the Photon Map. In *Rendering Techniques*.
- [Karlík et al. 2019] Ondřej Karlík, Martin Šik, Petr Vévoda, Tomáš Skřivan, Jaroslav Krivanek. MIS compensation: optimizing sampling techniques in multiple importance sampling. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2019)*, 38(6): 151:1-151:12, 2019.
- [Kondapaneni et al. 2019] Ivo Kondapaneni, Petr Vévoda, Pascal Grittmann, Tomáš Skřivan, Philipp Slusallek, Jaroslav Křivánek. Optimal Multiple Importance Sampling. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2019)*, 38(4): 37:1-37:14, 2019.
- [Lafortune and Willems 1993] Eric P Lafortune and Yves D Willems. 1993. Bi-directional Path Tracing.
- [Lu et al. 2013] H. Lu, R. Pacanowski, and X. Granier. 2013. Second-Order Approximation for Variance Reduction in Multiple Importance Sampling. *Comput. Graph. Forum (EGSR 2013)* 32, 7 (2013), 131–136.
- [Müller et al. 2017] Thomas Müller, Markus H. Gross, and Jan Novák. 2017. Practical Path Guiding for Efficient Light-Transport Simulation. *Comput. Graph. Forum* 36 (2017), 91–100.
- [Owen and Zhou 2000] Art Owen and Yi Zhou. 2000. Safe and Effective Importance Sampling. *J. Amer. Statist. Assoc.* 95, 449 (2000), 135–143.
- [Pajot et al. 2011] Anthony Pajot, Loic Barthe, Mathias Paulin, and Pierre Poulin. 2011. Representativity for Robust and Adaptive Multiple Importance Sampling. *IEEE Transactions on Visualization and Computer Graphics* 17, 8 (Aug. 2011), 1108–1121.
- [Sbert et al. 2016] Mateu Sbert, Vlastimil Havran, and Laszlo Szirmay-Kalos. 2016. Variance Analysis of Multi-sample and One-sample Multiple Importance Sampling. *Computer Graphics Forum* 35, 7 (2016), 451–460.

[Sbert and Havran 2017] Mateu Sbert and Vlastimil Havran. 2017. Adaptive Multiple Importance Sampling for General Functions. *Vis. Comput.* 33, 6-8 (June 2017), 845–855.

[Veach and Guibas 1995] Eric Veach, Leonidas Guibas. 1995. Optimally Combining Sampling Techniques for Monte Carlo Rendering. *SIGGRAPH 1995*.

[Veach and Guibas 1995a] Eric Veach and Leonidas Guibas. 1995. Bidirectional Estimators for Light Transport. In *Photorealistic Rendering Techniques*. Springer, 145–167.

[Vorba et al. 2014] Jiří Vorba, Ondřej Karlík, Martin Šik, Tobias Ritschel, and Jaroslav Křivánek. 2014. On-line Learning of Parametric Mixture Models for Light Transport Simulation. *ACM Trans. Graph.* (Proceedings of SIGGRAPH 2014) 33, 4 (2014).

10 Markov Chain Methods



In this part of the course, we will discuss Markov chain Monte Carlo - an alternative approach to the “one rendering algorithm” that can render any given scene efficiently.

MARKOV CHAIN MONTE CARLO



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

2

Markov chain Monte Carlo [*Metropolis et al. 1953*], or just MCMC, can be often more efficient than the ordinary Monte Carlo.

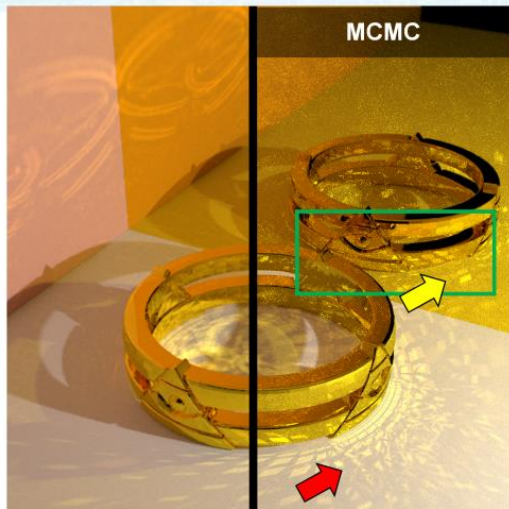
Consider this scene, which is lit from the outside and the light must travel through the window and the blinds into the classroom. On the left side we can see that the result of ordinary Monte Carlo is still noisy after 1 hour of rendering. On the right side we can see a result of an algorithm based on MCMC which is much cleaner.

MARKOV CHAIN MONTE CARLO

- Markov chain Monte Carlo rarely used in practice
- MCMC suffers from irregular/unpredictable convergence

Even though the MCMC algorithms often converge much faster than ordinary Monte Carlo methods, they are rarely used in practice. The reason behind this is that MCMC often suffers from irregular and unpredictable convergence.

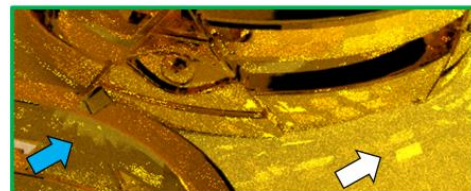
MARKOV CHAIN MONTE CARLO - ISSUES



Reference



MCMC



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

4

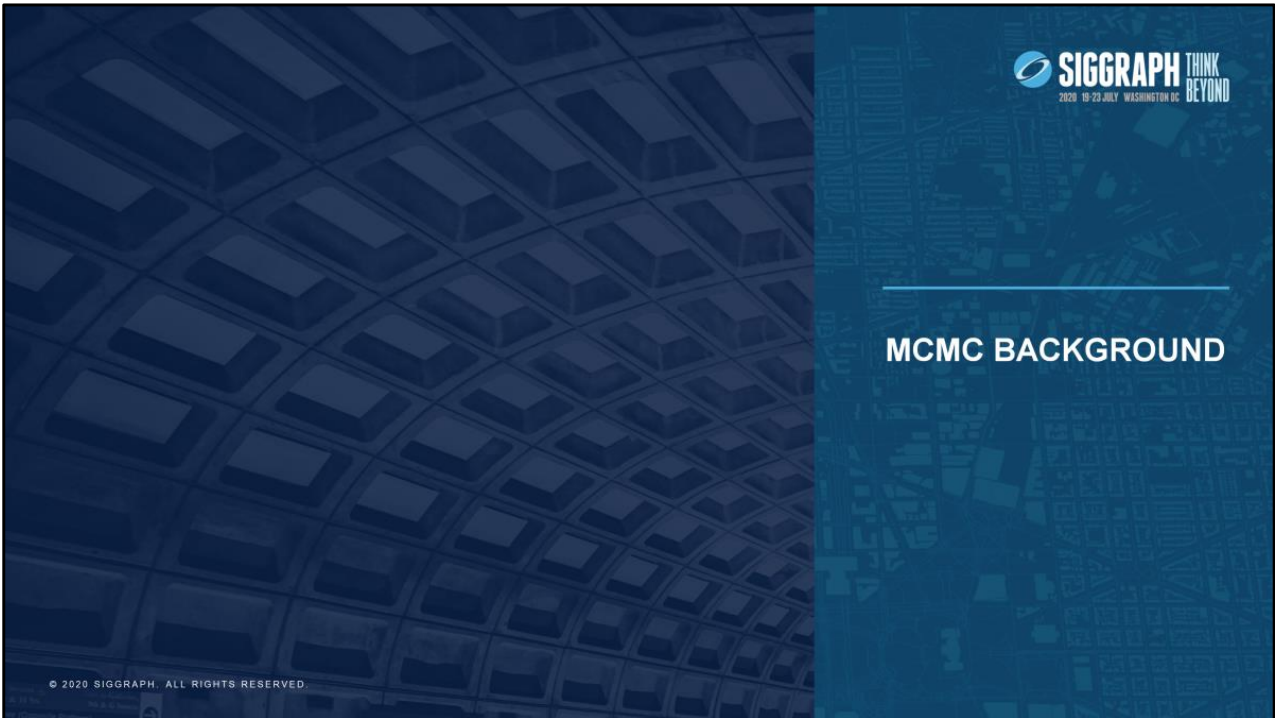
I demonstrate the issue of irregular convergence on this very simple scene that contains glossy and specular materials. Here we can see a glossy ring illuminated by a small light source, which creates caustics on the floor (red arrow). These caustics are reflected in the mirror behind the ring (yellow arrow).

In this scene, an MCMC algorithm oversamples some light transport in the scene (white arrow), while it fails to discover other (blue arrow). The irregular convergence of MCMC is especially visible in animations where it manifests as flickering (please visit the course webpage to view the animation).

MCMC GOAL

- New MCMC algorithm:
 - Predictable convergence
 - Faster convergence than ordinary Monte Carlo

The goal of our research was therefore to develop a new MCMC algorithm which would have more predictable convergence, while keeping MCMC's ability to converge faster than ordinary Monte Carlo in many difficult scenes.



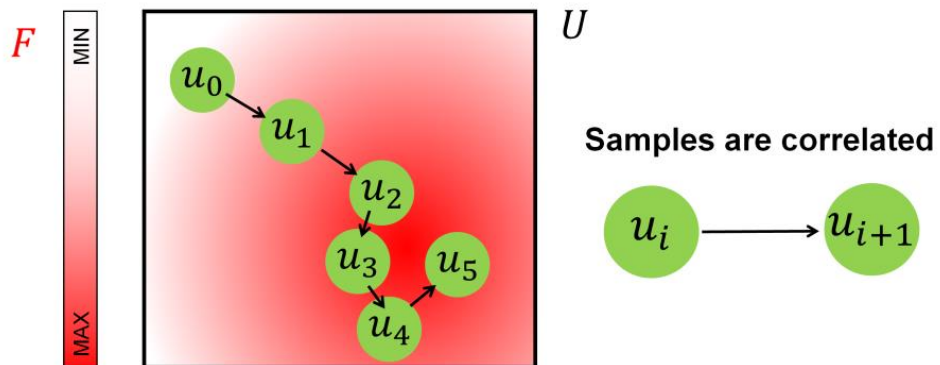
Before we discuss how we have approached this goal, let me give you some brief introduction to Markov chain Monte Carlo.

- Markov chain Monte Carlo [*Metropolis et al. 1953*]
 - General sampling technique
 - Sampling from unnormalized density F = target function

Markov chain Monte Carlo [*Metropolis et al. 1953*] is a general technique for generating samples from any unnormalized density F . To clarify, F can be any non-negative real function and is often called „target function“. The samples can then be used in Monte Carlo to estimate mean of a given function.

MCMC BACKGROUND

Given: State space U and Target function F
Samples = realizations of Markov chain states



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

8

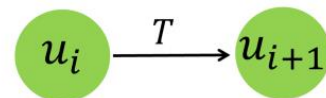
Given a state space U (the square), from which we want to draw samples, and the target function F (here represented by the white to red gradient), Markov chain Monte Carlo algorithm defines a Markov chain, whose states are the desired samples.

Two things are important to notice here:

1. A state of a Markov chain depends on the previous state, so the samples are correlated.
2. More samples are generated in the red area, where the target function has higher value.

MCMC BACKGROUND

- Samples converge to stationary distribution F^*
- F^* depends on the transition probability $T(u_i \rightarrow u_{i+1})$
- Metropolis-Hastings algorithm [Hastings 1970]
 - Ensures $F^* \sim F$



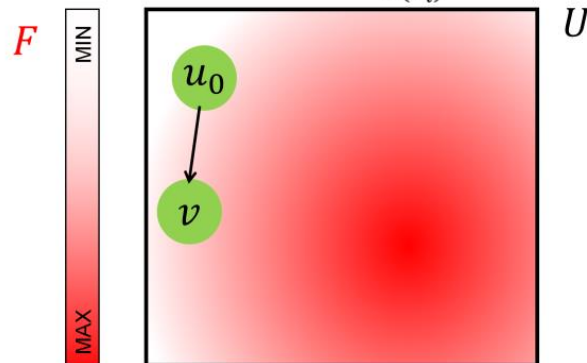
Given some conditions, the samples generated by the Markov chain will reach so-called stationary distribution F^* (please refer to [Šik and Křivánek 2018] for more details). This distribution depends on the transition probability $T(u_i \rightarrow u_{i+1})$ from one Markov chain state to the next state.

To ensure the stationary distribution F^* is proportional to the desired unnormalized density F , one can apply one of the MCMC algorithms. Probably the most famous one is Metropolis-Hastings [Hastings 1970].

METROPOLIS-HASTINGS

1) Generates proposal v

2) Proposal is probabilistically accepted $\sim \frac{F(v)}{F(u_i)}$



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

10

The Metropolis-Hastings algorithm works as follows:

1. First it generates a proposal v given an initial sample u_0 using a proposal distribution $Q(u_0 \rightarrow v)$
2. The proposal is then probabilistically accepted or rejected. The probability is based on a ratio of the target function values $\frac{F(v)}{F(u_i)}$

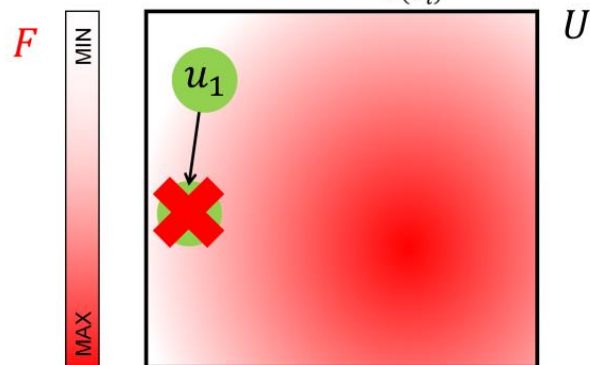
Note that for non-symmetric proposal distributions Q , ratio of proposal distributions $\frac{Q(v \rightarrow u_i)}{Q(u_i \rightarrow v)}$ must be also considered. The whole probability is then equal to $\min\left(1, \frac{F(v)}{F(u_i)} \frac{Q(v \rightarrow u_i)}{Q(u_i \rightarrow v)}\right)$.

The proposal distribution can be any distribution that depends on the current sample and allows the algorithm to sample the whole state space U .

METROPOLIS-HASTINGS

1) Generates proposal v

2) Proposal is probabilistically accepted $\sim \frac{F(v)}{F(u_i)}$

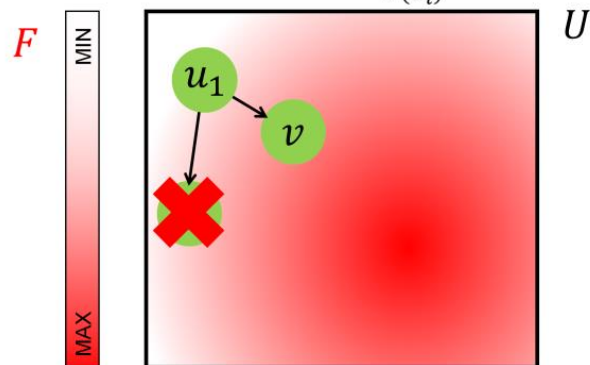


If the proposal v is rejected, a new sample u_1 is equal to the previous sample u_0 .

METROPOLIS-HASTINGS

1) Generates proposal v

2) Proposal is probabilistically accepted $\sim \frac{F(v)}{F(u_i)}$



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

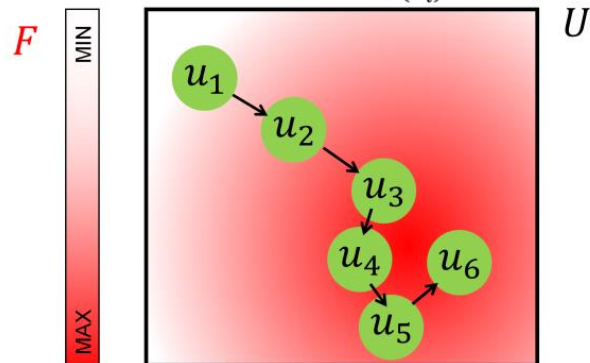
12

Another proposal v is generated using a proposal distribution that depends on u_1 .

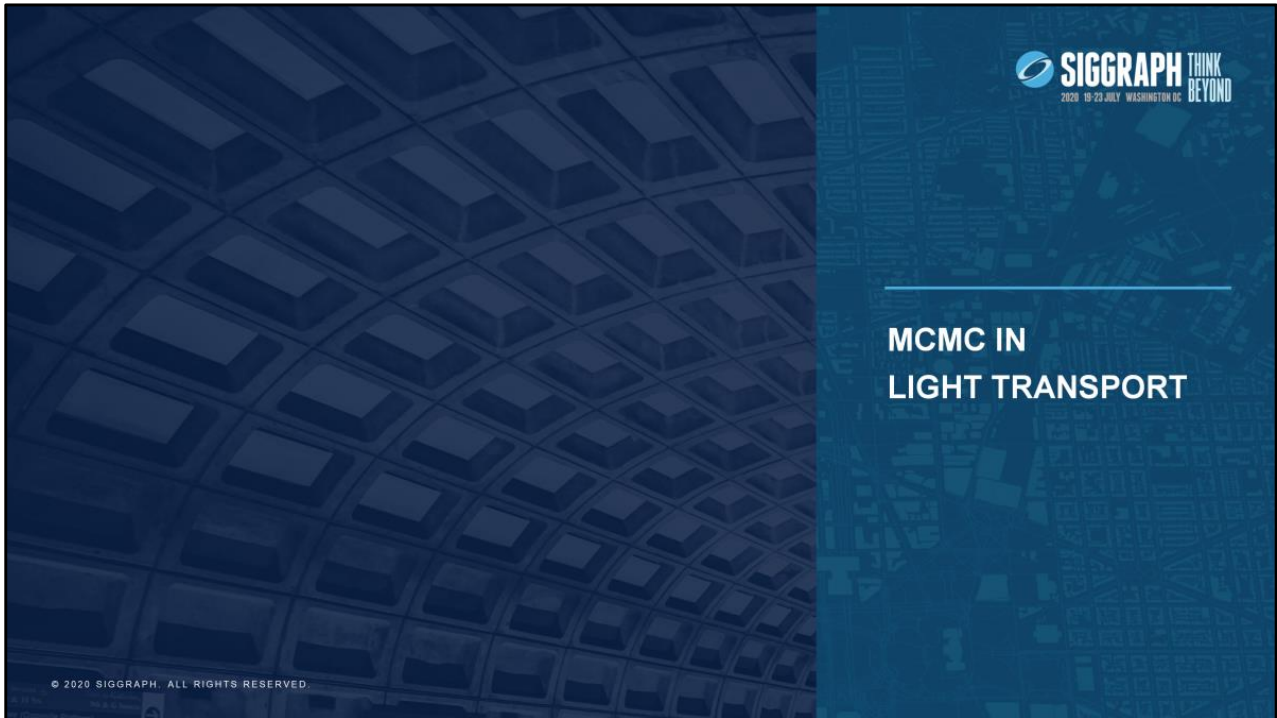
METROPOLIS-HASTINGS

1) Generates proposal v

2) Proposal is probabilistically accepted $\sim \frac{F(v)}{F(u_i)}$



In the case of acceptance, the new sample u_2 is set to be equal to the proposal v . This way we continue, until we have enough samples.



Let us now discuss how can we utilize MCMC in light transport simulation.

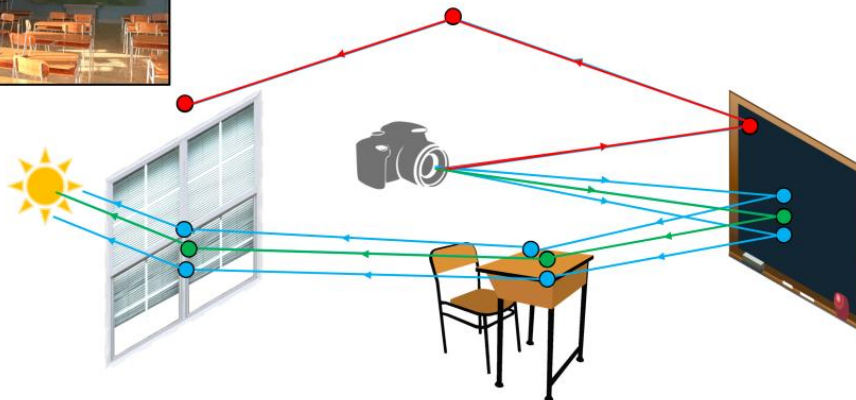
- Metropolis light transport [Veach and Guibas 1997]
 - State space = Path space
 - Samples = Paths
 - Target function = Contribution function
- Distribution of paths converges to the ideal (zero-variance) distribution

MCMC was introduced to light transport by Veach and Guibas in their algorithm Metropolis light transport. In this case the state space equals the path space and thus MCMC generates whole light transport paths. The target function is set to be equal to the path contribution function. This means that the paths are generated **almost** according to their contribution to the image.

The almost is important here, since the desired stationary distribution is only reached in infinity (e.g. infinitely many samples generated from the Markov chain will have the desired distribution). The distribution of the paths converges to the ideal distribution, which leads to Monte Carlo estimate with zero-variance. However, in practice we never reach the ideal distribution.



• Exploitation = Local exploration



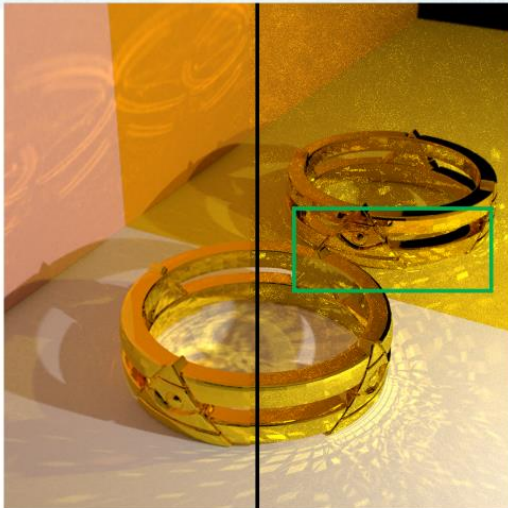
© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

16

Let us look how the actual algorithm works in the Classroom scene, which we have seen in the beginning.

MCMC will randomly generate proposal paths (red) until it finds and accepts one contributing path (green). Then it can utilize localized proposal distributions to generate more contributing paths (blue). Such localized proposals are often called (local) path mutations. Using these mutations we are effectively exploiting the original (green) path in order to locally explore an important region of the path space. This local exploration is behind the effectiveness of MCMC.

MCMC IN LIGHT TRANSPORT – ISSUES

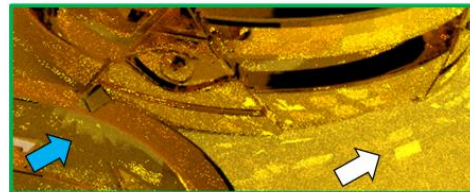


© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

Reference



MCMC

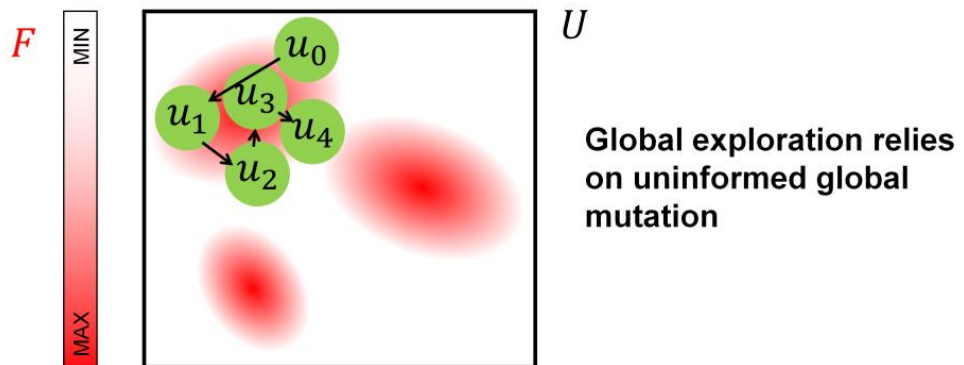


17

However, excessive local exploration can lead to convergence issues. This can be seen in this ring scene, which we have shown previously.

MCMC over-exploits some of the paths here which results in some parts of the image to be over-bright (white arrow), while other features are completely missing (blue arrow). This is especially visible in animations, where we can see random appearance of image features (e.g. the reflected caustics). Please see the course website for the animation.

- **Insufficient global exploration**
- Failure to discover and frequently sample important areas



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

18

The cause of this over-exploitation is insufficient global exploration. It is a failure to discover and frequently sample all the important areas in the scene. This is often the case with a target function that has high variation, such as the one equal to path contribution.

Let us now look again at the simple state space with a target function that has several separate modes. In this case the Markov chain may get „stuck“ in one of the modes – oversampling it – while not discovering the other modes. This is because the discovery of new important samples is usually done using some uninformed global mutation/proposal distribution. Proposals generated by such mutations are often rejected, since they will very likely have a low target function value.

GLOBAL EXPLORATION

Focus on local exploration/exploitation

- Manifold Exploration, [Jakob and Marschner 2012]
- Half-Vector Space Light Transport, [Kaplanyan et al. 2014]
- Multiplexed Metropolis Light Transport, [Hachisuka et al. 2014]
- Anisotropic Gaussian Mutations for Metropolis Light Transport through Hessian-Hamiltonian Dynamics, [Li et al. 2015]

Global exploration remained unaddressed!

In the past, most of the research works on MCMC in light transport simulation focused on local exploration. They present different mutations that allow effective exploitation of many types of paths. However, the major issue of global exploration that prevents adoption of MCMC into practice remained unaddressed.

OUR GOAL

- **Goal: Improve global exploration/uniformity of convergence**
- New more predictable MCMC algorithms
- Allow adoption of MCMC to practice

In our research we therefore focused on improving global exploration in Markov chain Monte Carlo algorithms. The main goal was to develop new MCMC algorithms which exhibit more uniform convergence. We believed that solving this issue would allow the algorithms to be adopted into practice.

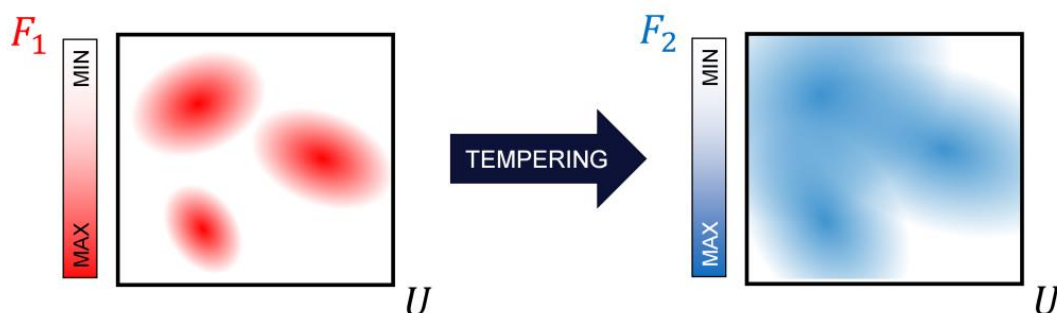
- **Improving Global Exploration of MCMC Light Transport Simulation**
Martin Šik and **Jaroslav Křivánek**, 2016
- **A Spatial Target Function for Metropolis Photon Tracing**
Adrien Gruson, Mickael Ribardiere, **Martin Šik**, Jiří Vorba, Rémy Cozot, Kadi Bouatouch, and **Jaroslav Křivánek**, 2016
- **Robust Light Transport Simulation via Metropolised Bidirectional Estimators**
Martin Šik, Hisanari Otsu, Toshiya Hachisuka, and **Jaroslav Křivánek**, 2016

In the following slides, I will discuss three of our works that tackle the issue of unpredictable MCMC convergence.

RESEARCH #1
Improving Global
Exploration of MCMC
Light Transport Simulation

REPLICA EXCHANGE - BACKGROUND

- Replica exchange, [Swendsen and Wang 1986]
- Combination of several Markov chains
- Different target functions

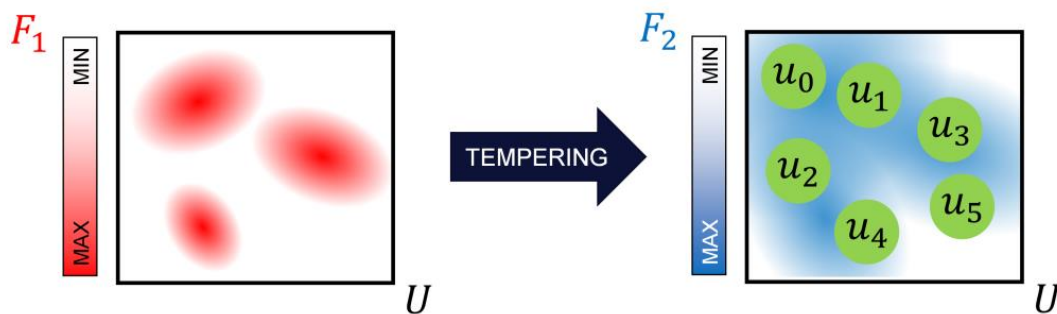


In our first research, we have approached the issue of insufficient global exploration by utilizing replica exchange and tempering [Swendsen and Wang 1986].

Replica exchange is a general technique that allows for combination of several Markov chains with different target functions. So we can have a chain with one target function F_1 equal to path contribution that allows for efficient local exploration. And another chain that uses less varying target function F_2 . The transition from the more varying target function to the smoother one is often called tempering.

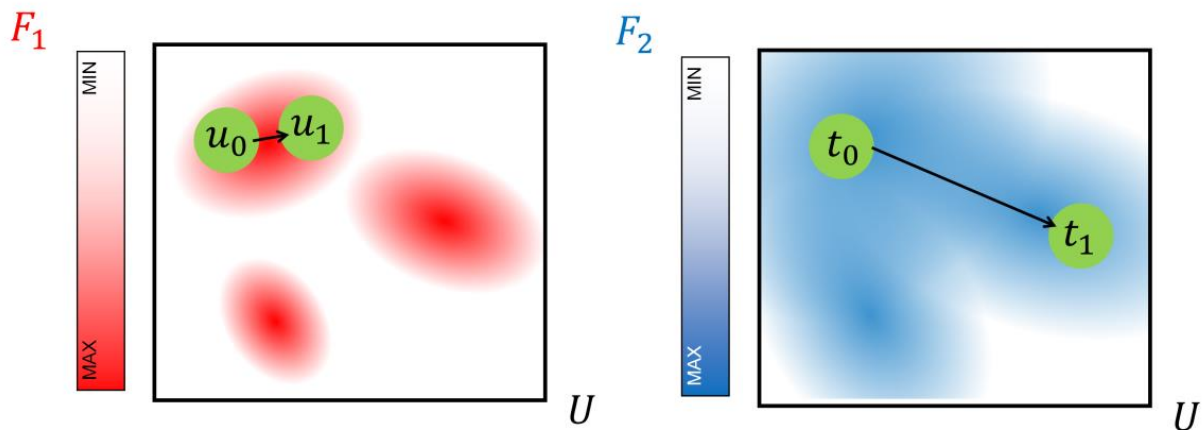
REPLICA EXCHANGE - BACKGROUND

- Replica exchange, [Swendsen and Wang 1986]
- Combination of several Markov chains
- Different target functions



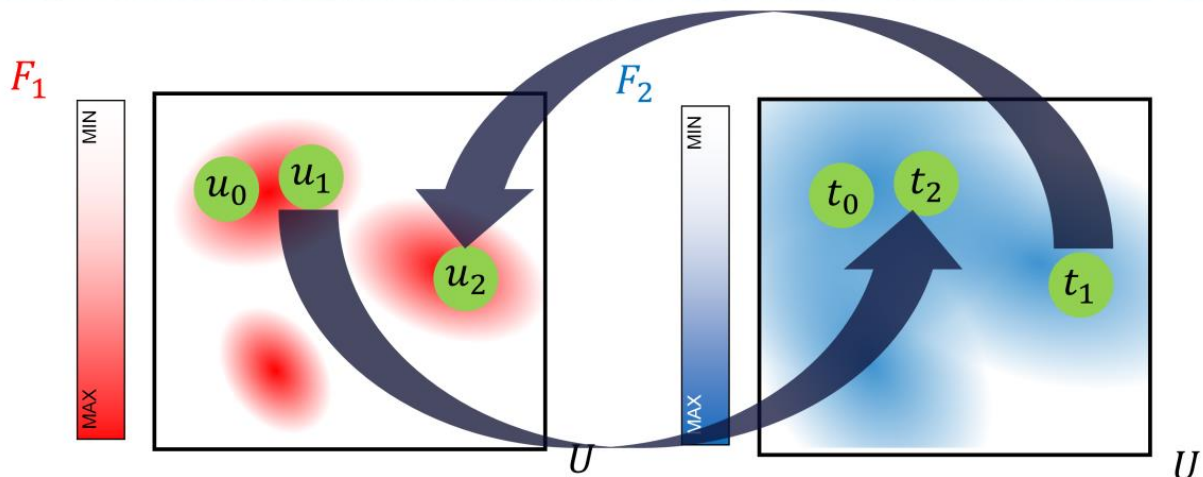
Such a less varying target function allows for easier global exploration of the whole state space.

REPLICA EXCHANGE - BACKGROUND



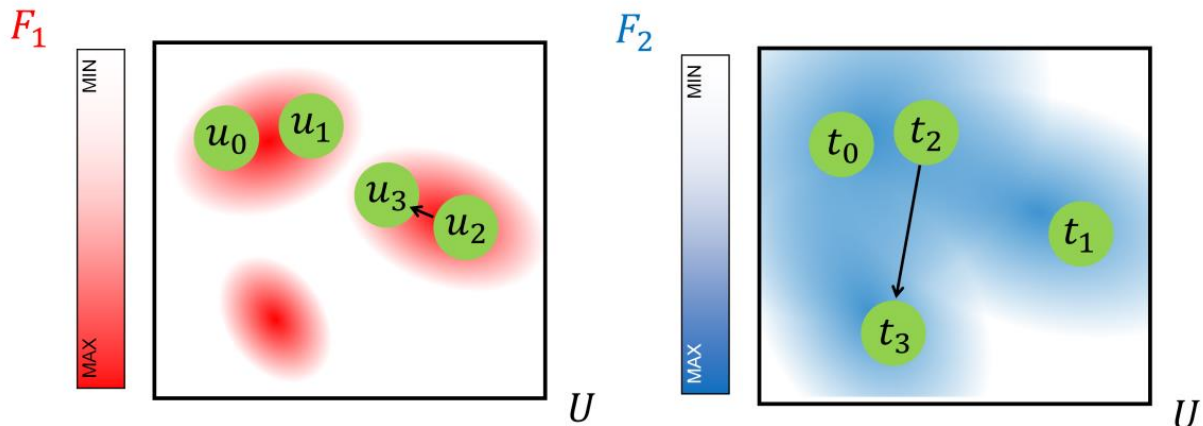
When utilizing more Markov chains, we can mutate them separately: $u_0 \rightarrow u_1$ and $t_0 \rightarrow t_1$.

REPLICA EXCHANGE - BACKGROUND



However, to enable benefits of both target functions, we exchange the current samples of the corresponding chains: $u_2 = t_1$ and $t_2 = u_1$. In this example, the chain with the target function F_1 discovered a new mode due to the exchange.

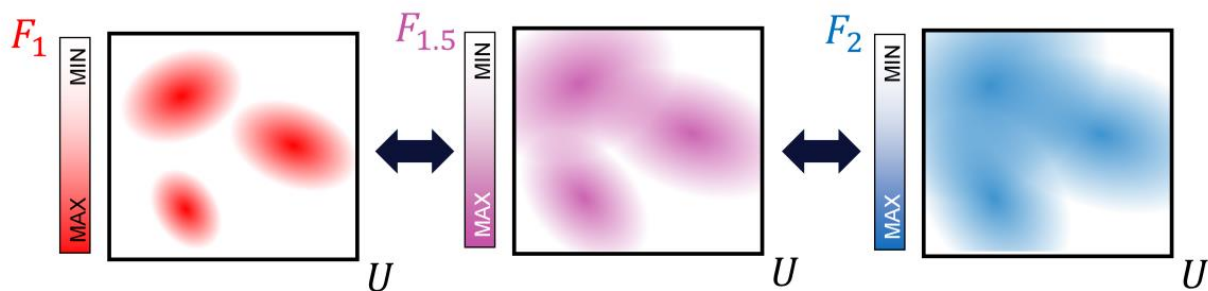
REPLICA EXCHANGE - BACKGROUND



After the exchange we continue mutating each chain separately ($u_2 \rightarrow u_3$ and $t_2 \rightarrow t_3$). The exchanges are performed every time after few separate mutations and they are accepted with a probability that depends on the values of their target functions: $\min\left(1, \frac{F_2(u_i) F_1(t_j)}{F_1(u_i) F_2(t_j)}\right)$.

REPLICA EXCHANGE - BACKGROUND

- We further optimize the exchange moves



To ensure that the current samples of two chains are exchanged with high probability, it is common to use more chains with increasingly more tempered target functions. The exchanges are then performed only between the neighboring chains. In our research we also use multiple chains, but we optimize the exchanging of their samples.

REPLICA EXCHANGE – PREVIOUS WORK

- Replica exchange light transport
[Kitaoka et al. 2009]
- Robust Adaptive Photon Tracing using Photon Path Visibility
[Hachisuka and Jensen 2011]

Replica exchange was used before in light transport simulation:

Replica exchange light transport *[Kitaoka et al. 2009]* used several different chains, where each chain was effective at exploiting a different type of paths. However, none of the chains had a target function that could efficiently explore the whole state space and thus global exploration was not improved in many cases (e.g. paths corresponding to reflected caustics could not be efficiently found nor exploited).

Robust Adaptive Photon Tracing using Photon Path Visibility *[Hachisuka and Jensen 2011]* used replica exchange between two chains, where one of them had a constant target function (i.e. it accepted every proposal). However, in this case using replica exchange did not bring any advantage compared to using uninformed global mutation for global exploration (both had the same probability of being accepted).

REPLICA EXCHANGE – OUR APPROACH

Focus on both key components of replica exchange:

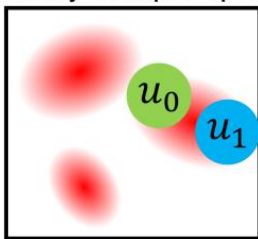
- Target function tempering
- Replica exchange moves

In our research, we have tried to maximize replica exchange potential by focusing on both of its key components: Target function tempering and replica exchange moves – the exchanging of the chains' samples.

PRIMARY SAMPLE SPACE

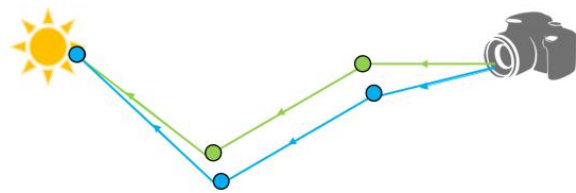
- Primary sample space Metropolis light transport, [Kelemen et al. 2002]

Primary sample space



PATH SAMPLING

Path space



Before I discuss how we approached tempering, note that our method is built on top of Primary sample space metropolis light transport (PSSMLT) [Kelemen et al. 2002]. PSSMLT does not directly mutate paths in the path space. Instead it mutates a sample (green point on the left) in so called primary sample space (hypercube) and this sample is then utilized as a random vector during path sampling, which constructs the path (green path) in the path space. Mutating the sample (green -> blue point) results in the path being mutated (green -> blue path).

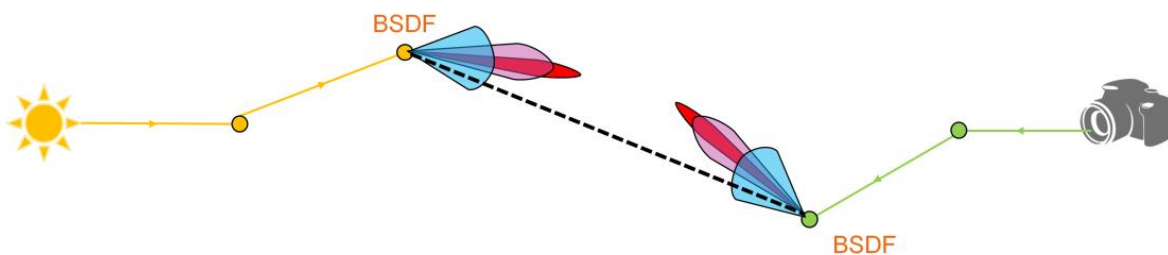
Such utilization of MCMC is advantageous in many aspects:

1. It greatly simplifies mutations, since they now occur in hypercube.
2. We can utilize path sampling techniques from existing algorithms (path tracing, bidirectional path tracing etc.) to construct paths from the primary sample space.
3. If the paths are efficiently sampled by the path sampling techniques, the target function will have lower variation (see [Kelemen et al. 2002] for details)

For the above reasons we always utilize primary sample space in our algorithms.

TEMPERING – OUR APPROACH

- Base algorithm: Bidirectional path tracing, [Veach and Guibas 1994]
- BSDFs at connecting vertices = major source of variation
- Widen lobes of BSDFs at connecting vertices



In our method, we utilize path sampling techniques from bidirectional path tracing [Veach and Guibas 1994] to construct paths from samples in primary sample space. Bidirectional path tracing creates paths by tracing a path from the camera (green) and a path from a light source (yellow). These paths are then connected (black dashed line). The amount of energy carried through the connection depends on the bidirectional scattering distribution function (BSDF) that defines the reflection profile of the material.

In the case of glossy materials, the BSDF will have a sharp lobe (red lobes) and will be a major source of target function variation. We therefore gradually widen BSDF lobes (red -> purple -> blue), which leads to a smoother target function.

REPLICA EXCHANGE MOVES – OUR APPROACH

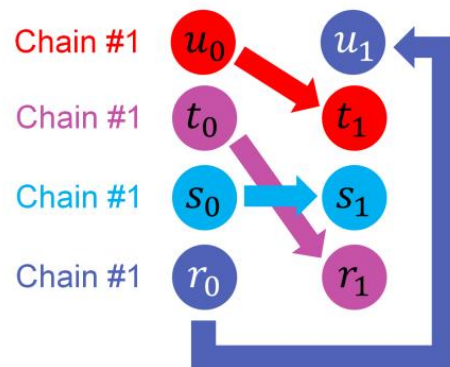
Tested replica exchange strategies:

- Neighbour swapping
- Equi-energy moves, [Baragatti et al. 2012]
- Equi-energy sampler, [Kou et al. 2006]
- Frequent equi-energy moves
- Importance-sampled permutations
- Importance-sampled swaps

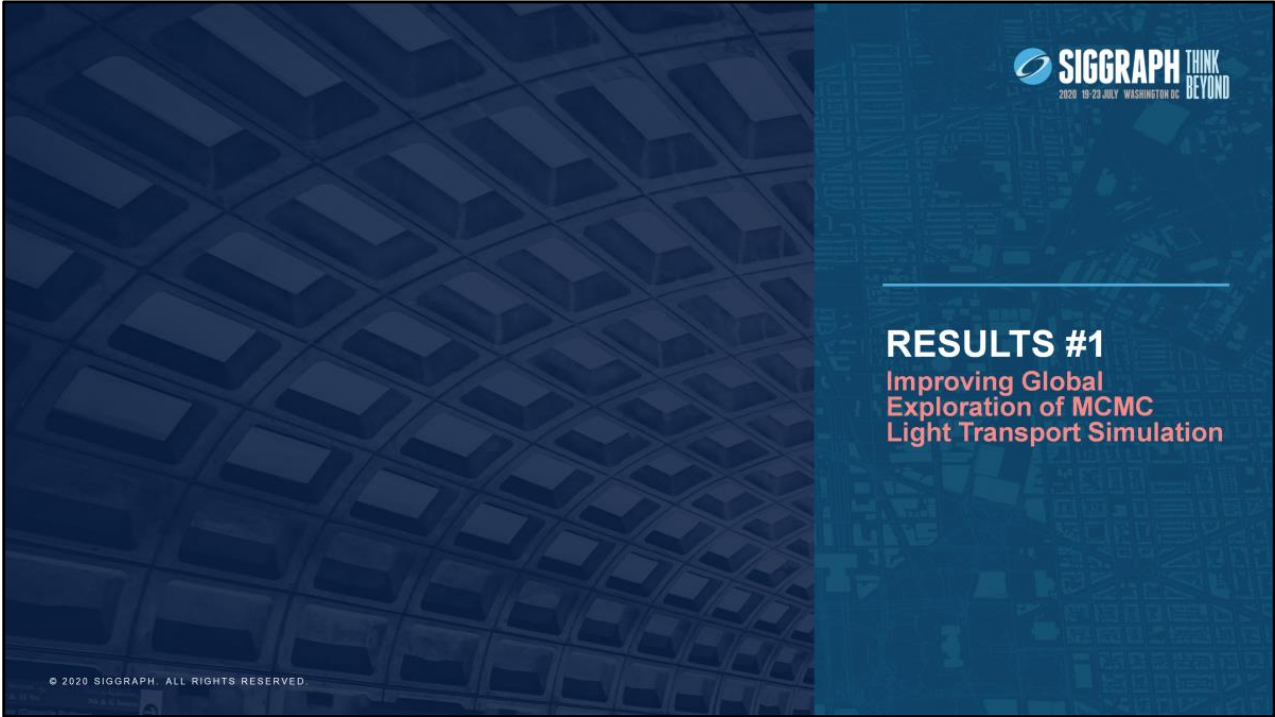
We wanted to find the best strategy for exchanging the samples of the chains. For that purpose, we have tested several strategies. These include commonly used strategies from general Markov chain Monte Carlo literature, but also exchange moves of our own original design (shown in red). For the description of all these techniques, please refer to the doctoral thesis: **Global exploration in Markov chain Monte Carlo methods for light transport simulation [Šik 2018]**

IMPORTANCE-SAMPLED PERMUTATIONS

- **Importance-sampled permutations** = the best strategy based on synthetic and rendering tests
- Permutes all chains at once with 100% probability

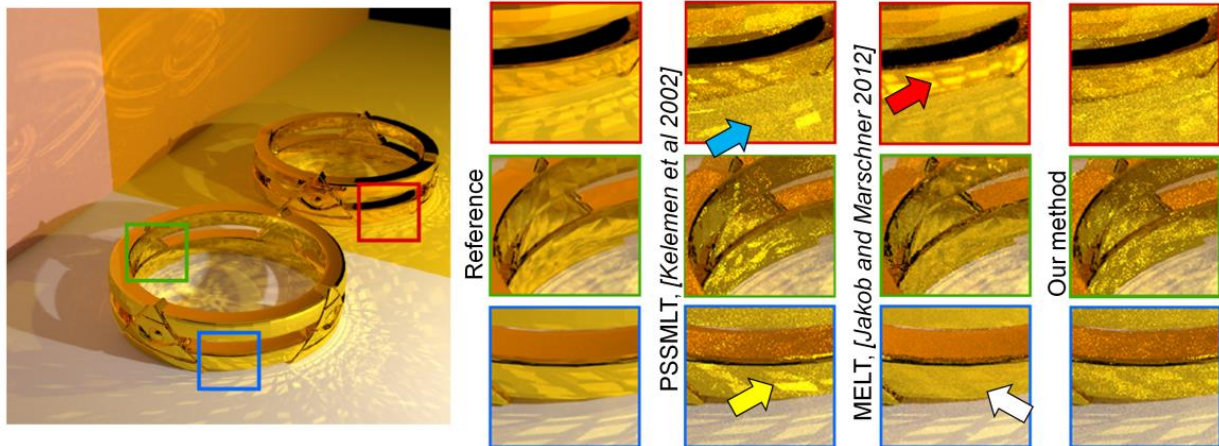


From synthetic and rendering tests we have determined that the best option is to use our own strategy called importance-sampled permutations. This strategy allows to permute the current samples of all the chains at once. The permutation is also sampled in such a way that it is always accepted and thus it efficiently combines the chains.



Now that we have covered the method, I can show you its results.

COMPARISON - RING



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

36

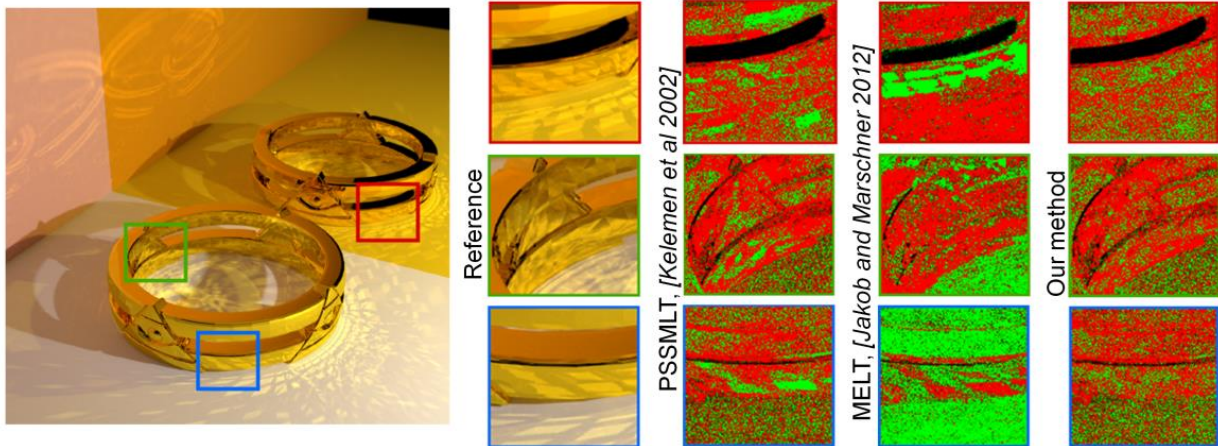
As I have shown before, the ring scene contains specular and glossy materials that lead to poor global exploration of the current methods.

We can see that after 15 minutes of rendering, an MCMC algorithm - Primary sample space metropolis light transport (PSSMLT) [Kelemen et al. 2002], failed to sufficiently sample some of the reflected caustics (blue arrow), while others are oversampled (yellow arrow).

We also show the result of Manifold exploration light transport (MELT) [Jakob and Marschner 2012]. Its result seems to be more converged due to its superior local exploration mutation, however again some of the transport is oversampled (red arrow), while other parts of the scene are under sampled (white arrow).

While the result of our method is quite noisy, it contains most of the specular/glossy transport due to its improved global exploration.

COMPARISON - RING



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

37

If we look at the positive-negative difference with the reference (red color = reference is brighter, green color = reference is darker), we can see that our method delivers more uniform red-green noise than the other methods due to its more regular convergence.

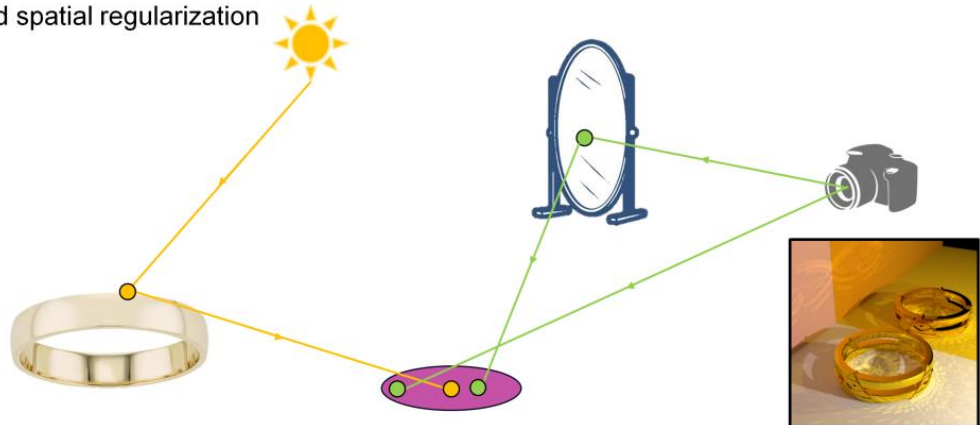
Please visit the webpage of this course to see an animation of convergence of different methods and a comparison of temporal coherence. These comparisons show that our method has more uniform convergence and higher temporal stability. However, the results are far from perfect, which motivated our further research.



In the second research, we have again focused on improving convergence uniformness in the context of Markov chain Monte Carlo. But this time we specialize at a specific light transport algorithm: Stochastic progressive photon mapping [Hachisuka and Jensen 2009].

STOCHASTIC PROGRESSIVE PHOTON MAPPING

- Stochastic progressive photon mapping, [Hachisuka and Jensen 2009]
- Path reuse and spatial regularization



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

39

Let us begin with a quick recapitulation of Stochastic progressive photon mapping (SPPM) [Hachisuka and Jensen 2009]. I demonstrate how the algorithm works on the schematic view of the ring scene.

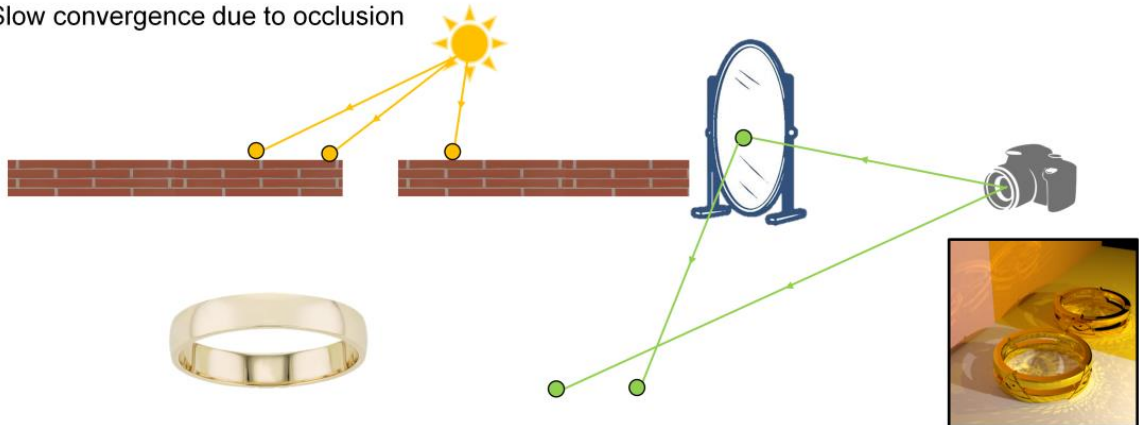
The algorithm starts by generating rays from the camera (green) and recording their hit points, so called **measurement points**. If the ray hits highly glossy surface, it bounces off it and we continue tracing until we find a diffuse enough surface, where the measurement point is stored.

Once it has recorded all the measurement points, it starts tracing paths from light sources (yellow). A light path bounces off the surfaces of the scene and on each diffuse bounce, it performs density estimate (purple disc). Each measurement point that falls into the density estimate radius records the light contribution and propagates it back to its origin (camera pixel).

The generation of camera rays and light paths is then interleaved during the algorithm. Due to path reuse and spatial regularization inherent to density estimation, the algorithm is very effective at handling effects like caustics or even reflected caustics.

STOCHASTIC PROGRESSIVE PHOTON MAPPING

- Stochastic progressive photon mapping, [Hachisuka and Jensen 2009]
- Slow convergence due to occlusion



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

40

However, the algorithm's efficiency will drop in scenes where it is difficult to find a contributing path. For example, if most of the light paths fail to reach the region visible by the camera due to occlusion (represented by a brick wall on the slide).

STOCHASTIC PROGRESSIVE PHOTON MAPPING

Equal-time comparison (10 min)



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

Reference



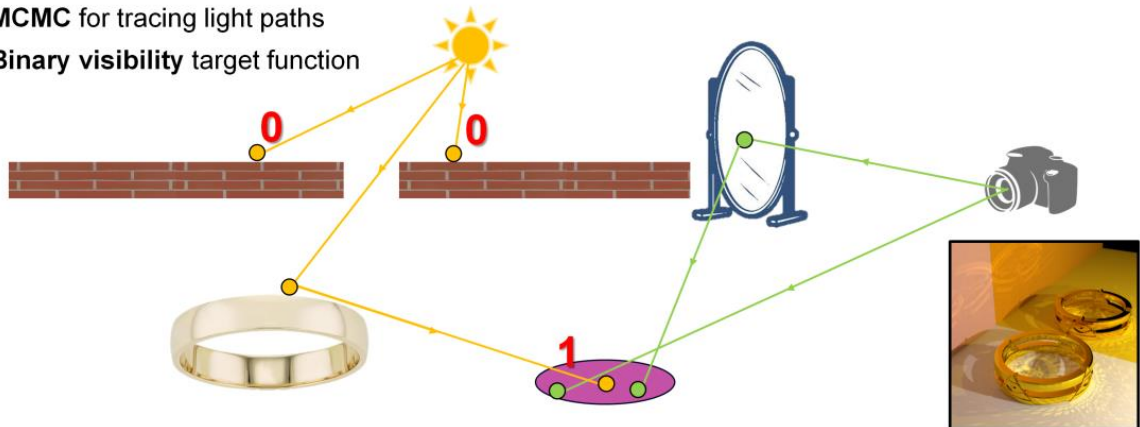
SPPM, [Hachisuka and Jensen 2009]

41

An example of such a scene is here. All the light is coming from the outside through the windows, so most of the light paths never find the interior. Especially the part of the scene, which is poorly lit (red inset), is very noisy in the stochastic progressive photon mapping (SPPM) result.

ROBUST ADAPTIVE PHOTON TRACING USING PHOTON PATH VISIBILITY

- Robust Adaptive Photon Tracing using Photon Path Visibility, [Hachisuka and Jensen 2011]
- **MCMC** for tracing light paths
- **Binary visibility** target function



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

42

To improve the algorithm performance in such scenes, Hachisuka and Jensen has utilized Markov chain Monte Carlo to trace the light paths towards the visible region. They choose a very simple target function, so called binary visibility. The function is simply zero for non-contributing paths and one for all contributing paths. Note that camera paths are traced using an ordinary independent sampler.

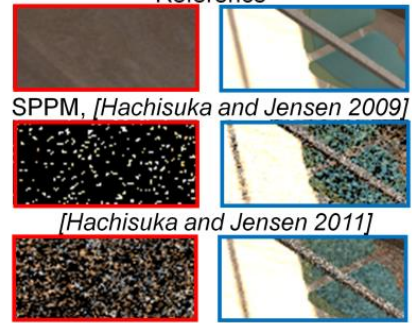
ROBUST ADAPTIVE PHOTON TRACING USING PHOTON PATH VISIBILITY

Equal-time comparison (10 min)



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

Reference



SPPM, [Hachisuka and Jensen 2009]

[Hachisuka and Jensen 2011]

By using such a target function they were able to deliver more light paths to the visible region and thus significantly improve the results. However, we can see that there is much more noise in red inset compared to the blue one. The algorithm still distributes more photons to the more lit regions, which leads to non-uniform convergence. This is an issue we have addressed in our research.

OUR METHOD - IDEA

Equal probability to hit any measurement point => uniform image error
Target function = inverse photon (light path vertex) density



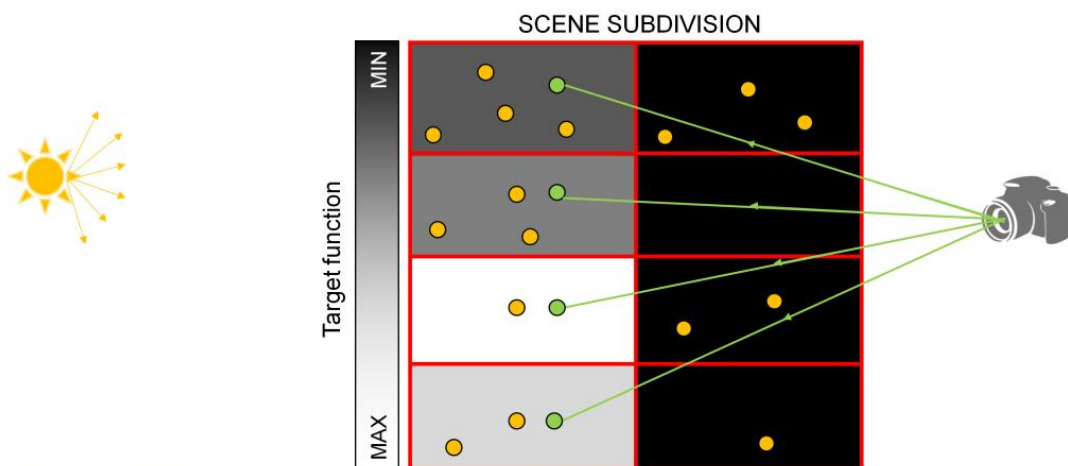
© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

44

It can be proven (see our paper [Gruson et al. 2016] for details) that an image has uniform relative error if each measurement point G has the same probability of receiving a non-zero contribution from any light path (under simplifying assumptions, such as a diffuse BRDF).

This is achieved if the target function value for a light path contributing to a measurement point G is equal to $1/P(G)$, where $P(G)$ is the probability of generating a light path by an ordinary Monte Carlo (i.e. uniform sampling of the primary sample space). To compute such target function, one can estimate photon (light path vertex) density in the scene and then set the target function to be inverse of this estimate.

TARGET FUNCTION COMPUTATION



In practice the target function is computed as follows:

1. Subdivide the scene into spatial regions (red grid).
2. Trace light paths and record how many photons (yellow points) land in each region (adjusted for the non-uniform light path tracing probability).
3. Update target function accordingly (white = maximum target function value, black = minimum value).
4. Return to step 2. and optionally subdivide some regions to improve the accuracy of the estimate.

Notice that the target function is set to zero in regions without any measurement points (green), since we don't need to trace light paths there.

- Main target function = inverse photon density \times the inverse squared distance to the camera
- Avoid poor global exploration => replica exchange with 4 chains:
 - Main target function
 - Inverse squared distance to the camera
 - Binary visibility [*Hachisuka and Jensen 2011*]
 - Uniform

Beside the main idea of the algorithm there are many fine details that make the algorithm improve upon the previous ones. To get more smoother target function in large spatial regions, the whole target function is multiplied by the inverse squared distance to the camera. This results in generating more photons closer to the camera, where they matter more.

Since the resulting target function can be quite spiky and thus lead to poor global exploration, we have utilized replica exchange as in the previous research. The algorithm uses four Markov chains with different target functions:

1. The main target function based on photon density
2. Target function equal to the inverse squared distance to the camera
3. Binary visibility target function (the original target function from [*Hachisuka and Jensen 2011*])
4. Uniform target function (Always equal to 1, always accepted)

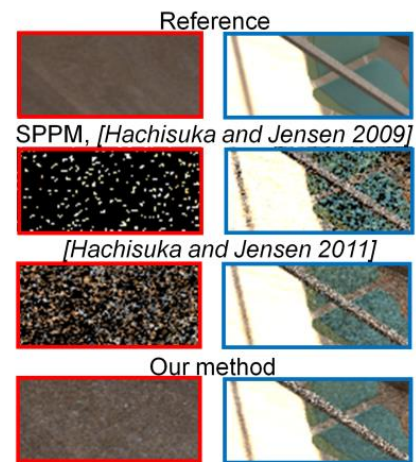
Note that the first 3 target functions are set to zero for non-contributing paths.



Now that we have covered the method, I can show you its results.

OUR METHOD #2 - RESULTS

Equal-time comparison (10 min)

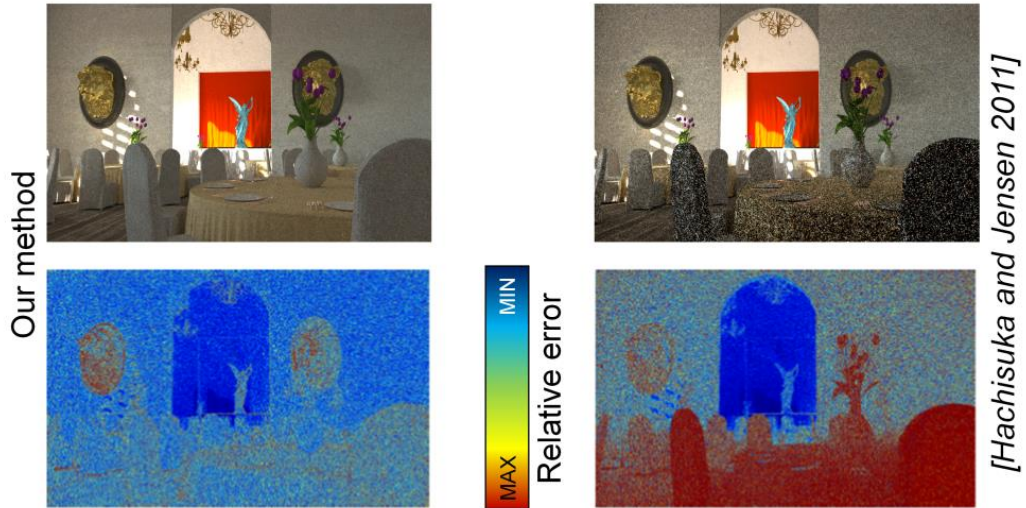


© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

48

As you can clearly see our method significantly reduces noise compared to the previous methods even in the dimly lit area of the scene (red inset).

OUR METHOD #2 – UNIFORM RELATIVE IMAGE ERROR



In this equal-time (30 min) comparison we show that our method has not only lower relative image error compared to the previous method [Hachisuka and Jensen 2011], but the error is also more uniform in the whole image.

OUR METHOD #2 – ISSUES

Reference



Our method



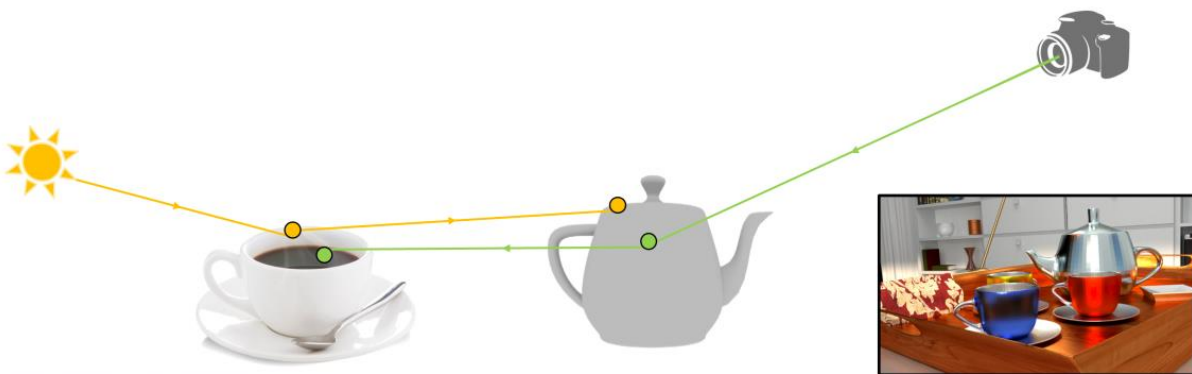
Unfortunately, since our method is based on stochastic progressive photon mapping, it can't efficiently handle many types of paths (e.g. glossy-glossy transport). On the slide you can see a scene containing glossy materials. After 1 hour of rendering with our method, the scene is still very noisy. Note that the previous method by Hachisuka and Jensen would not deal with this scene any better.



In the last research, I discuss here, we have therefore focused on developing a more robust light transport algorithm, that would utilize Markov chain Monte Carlo to handle all types of scenes, while not exhibiting any convergence issues.

VERTEX CONNECTION AND MERGING

- Vertex connection and merging, [Georgiev et al. 2012/ Hachisuka et al. 2012]



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

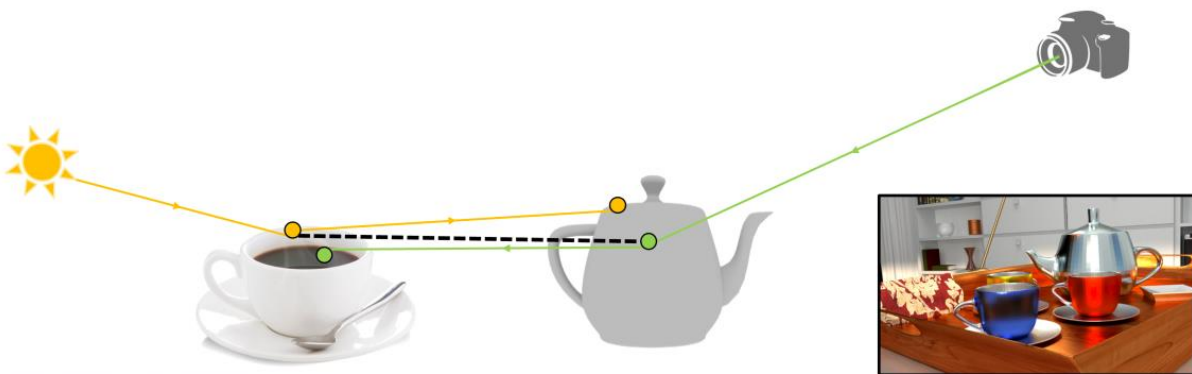
52

One of the light transport algorithms that can robustly handle many types of transport paths is vertex connection and merging [Georgiev et al. 2012/ Hachisuka et al. 2012]. Vertex connection and merging (VCM) is an ordinary Monte Carlo algorithm, which was already presented in this course and thus I will just quickly recapitulate its description.

I demonstrate how it works on the schematic view of the „tray“ scene, which caused issues to our previous method. VCM creates the paths by combining a path from the camera (green) and a path from a light source (yellow).

VERTEX CONNECTION AND MERGING

- Vertex connection and merging, [Georgiev et al. 2012/ Hachisuka et al. 2012]



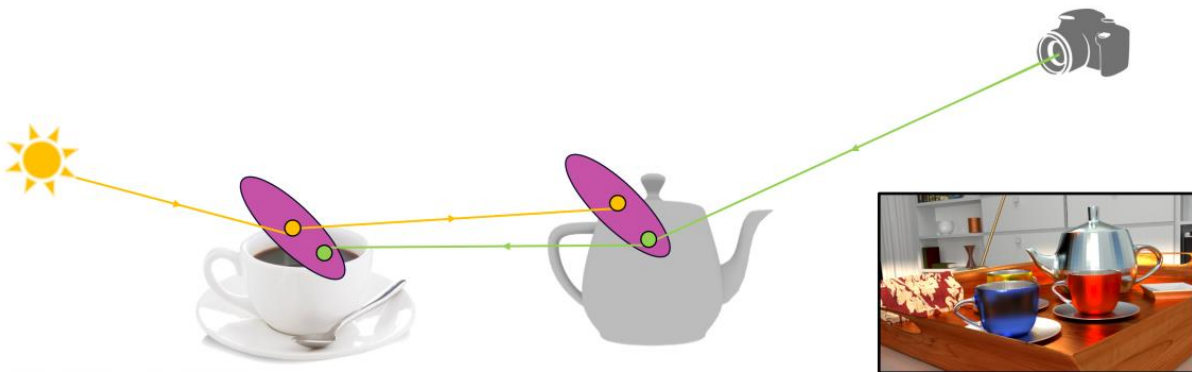
© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

53

It combines the paths using **connections**, where any of the light path vertices can be connected to any camera path vertex. One such connection (black dashed line) is shown on the slide.

VERTEX CONNECTION AND MERGING

- Vertex connection and merging, [Georgiev et al. 2012/ Hachisuka et al. 2012]
- Merging = path reuse + spatial regularization



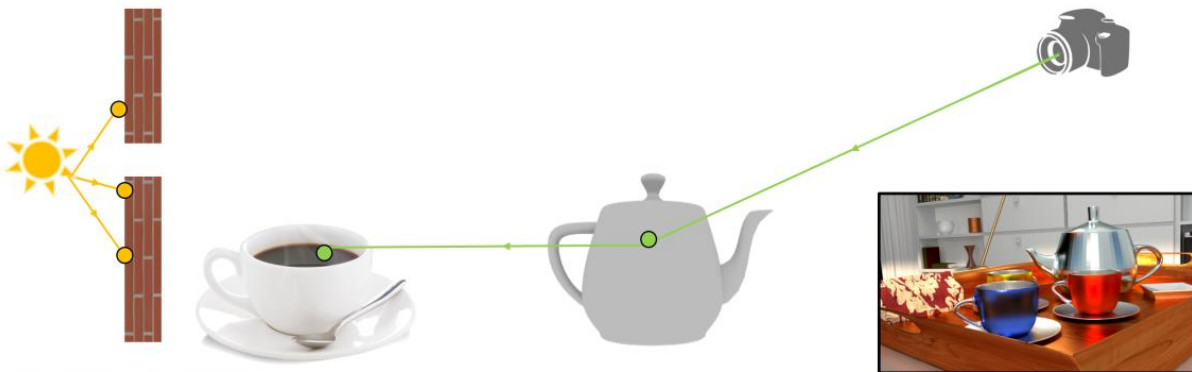
© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

54

It can also combine the paths using density estimation, in this case called **merging**. Note that due to the inherent spatial regularization and path reuse, merging is effective at handling caustics and reflected caustics. However, unlike the algorithms based on SPPM, VCM allows merging at any vertex. This increases robustness in the case of glossy-glossy transport.

VERTEX CONNECTION AND MERGING

- Vertex connection and merging, [Georgiev et al. 2012/ Hachisuka et al. 2012]
- Slow convergence due to occlusion



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

55

However, like SPPM, VCM becomes quite inefficient in scenes, where it is difficult to find a contributing path. For example, if most of the light paths fail to reach the region visible by the camera due to an occlusion (represented by a brick wall on the slide).

OUR METHOD

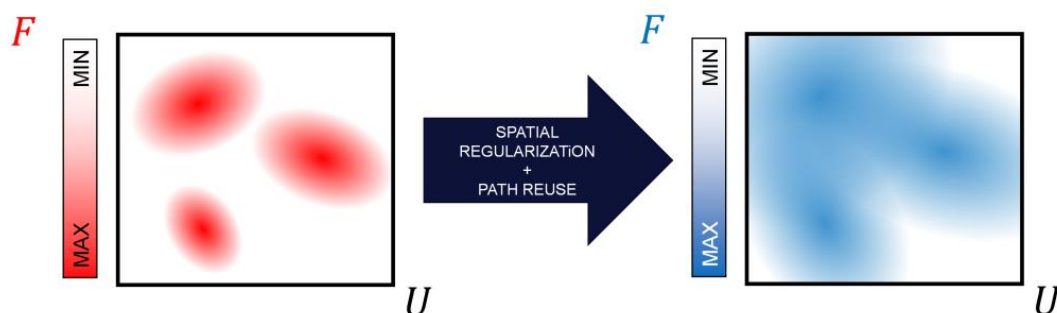
Robust combination:

- Markov chain Monte Carlo = exploitation
- Vertex connection and merging = glossy/specular transport

In our research we combine vertex connection and merging with MCMC in one robust algorithm. MCMC will enable efficient exploitation of paths, which will lead to generation of more contributing paths. While VCM techniques (connections and merging) will effectively handle glossy/specular transport.

COMBINING VCM AND MCMC

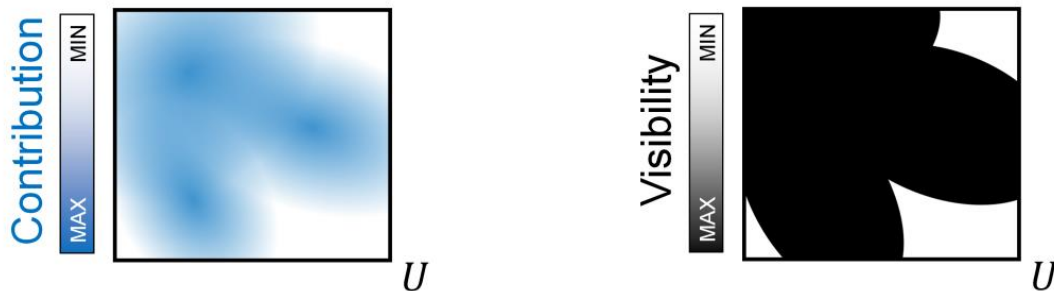
- Primary sample space Metropolis light transport, [Kelemen et al. 2002]
- Reduce target function variation in specular/glossy transport



As in the previous method we built on top of Primary sample space Metropolis light transport [Kelemen et al. 2002], but here we utilize all the techniques of VCM (merging and connections) and thus the target function F equal to path contribution has lower variation in scenes with specular and glossy transport. This will allow MCMC to more easily explore the whole state space and significantly reduce the issues connected to poor global exploration.

UTILIZING REPLICA EXCHANGE

- Replica exchange: two target functions
- Binary visibility target function, [Hachisuka and Jensen 2011]

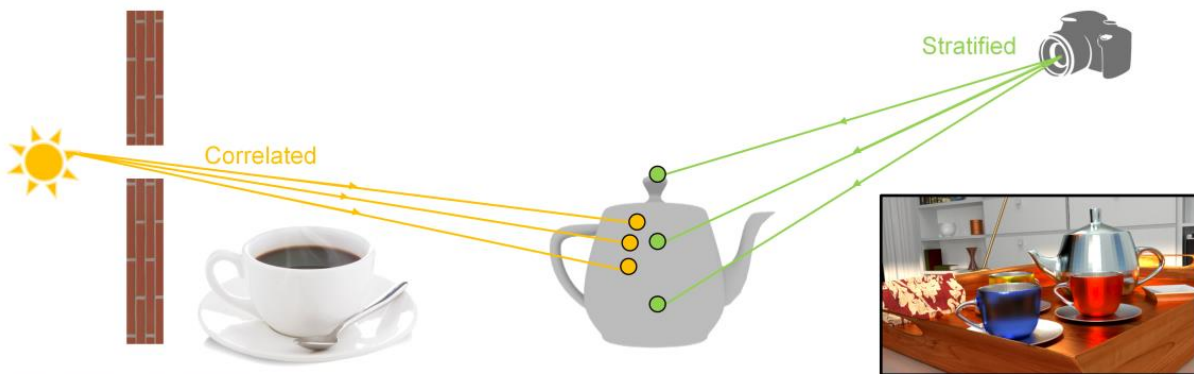


To further improve global exploration of our method, we apply replica exchange as in the previous methods. However, here it is sufficient to only use two target functions:

1. The main target function equal to path contribution – for better local exploration (exploitation).
2. The binary visibility target function that is non-zero for all contributing paths [Hachisuka and Jensen 2011].

STRATIFIED CAMERA PATHS

- Further reduce correlation of the samples
- MCMC used to exploit paths from light sources



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

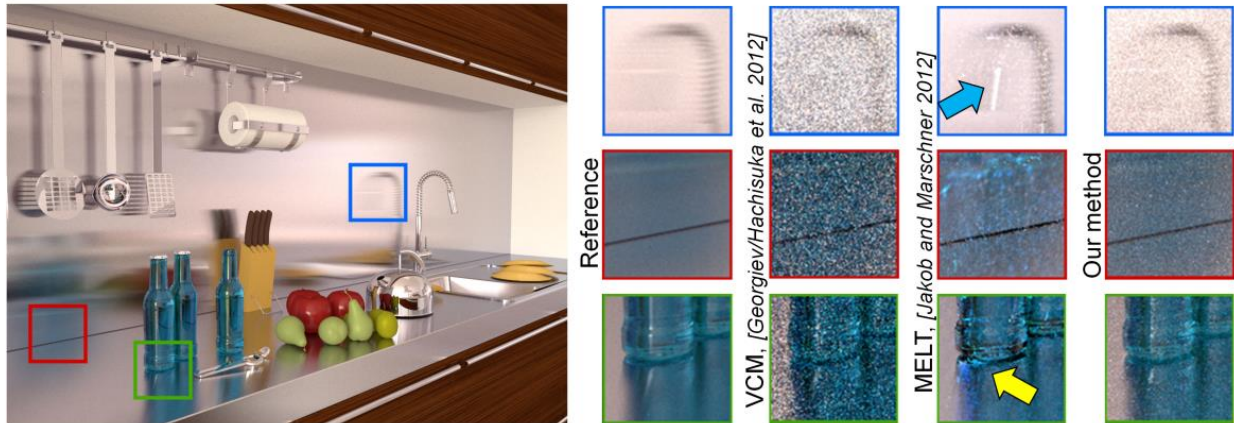
59

We can utilize MCMC to guide both light paths from the light sources (yellow) and camera paths from the camera (green). However, we have found out that using independent stratified sampler for camera paths leads to better results. And thus we only use MCMC to guide paths from light sources to ensure effective exploitation in many difficult scenes.



Now that we have covered the method, I can show you its results.

COMPARISON - KITCHEN



© 2020 SIGGRAPH. ALL RIGHTS RESERVED.

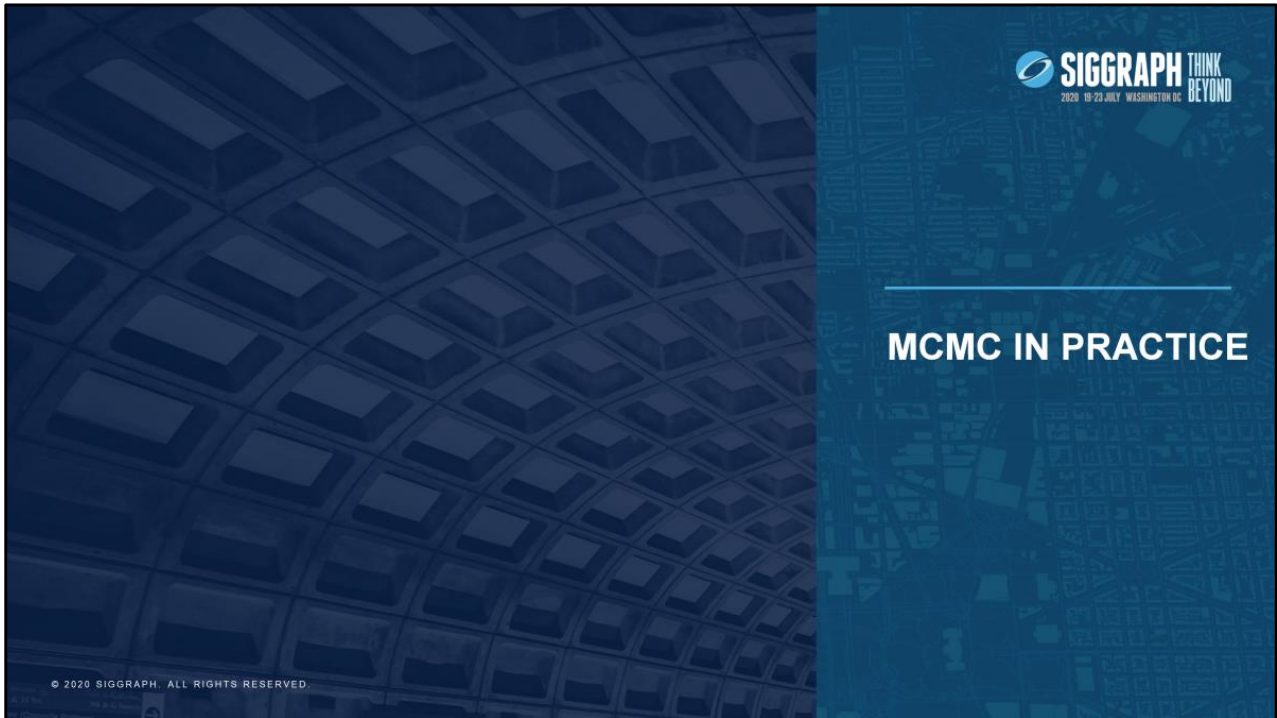
61

We demonstrate the results of our method in this Kitchen scene, which contains glossy and specular materials on the kitchen counter. The scene is only lit from the outside by the sun and sky. The light therefore needs to go through a window before it reaches the visible region.

The result of vertex connection and merging (VCM) [Georgiev et al. 2012/Hachisuka et al. 2012] is very noisy even after 1 hour of rendering. This is due to the algorithm's inability to deliver enough light paths to visible region.

While the MCMC algorithm Manifold exploration light transport (MELT) [Jakob and Marschner 2012] delivers more cleaner result due to local exploration, the resulting image contains many artifacts (blue and yellow arrows). These are caused by the algorithm's poor global exploration.

We can see that the result of our method only introduces some high-frequency noise comparable to ordinary Monte Carlo due to better global exploration. The noise level is also much lower compared to vertex connection and merging due to the ability of local exploration. Our algorithm has also better temporal coherence compared to previous MCMC methods. (Please visit the course webpage to view the animation)



During our research we believed that improving global exploration and removing convergence issues of MCMC algorithms will lead to their adoption in practice. So let us quickly look if that really happened.

- Sony Picture Imageworks.
 - Sony Pictures Imageworks Arnold, [*Kulla et al. 2018*]

- Corona renderer
 - Implementing One-Click Caustics in Corona Renderer, [*Šik and Křivánek 2019*]

To this day, we are aware of two MCMC-based light transport algorithms that are being used to render the most difficult scenes. Both of these solutions are similar to the last described algorithm. Since we are the authors of the second practical solution, I will shortly talk about it.

CORONA RENDERER – CAUSTIC SOLVER

- One click solution – no need to set anything
- Lightweight solution (inspired by Lightweight photon mapping, [Grittmann et al. 2018])
- Refined target function
- Adaptive selection of emitting light sources

MCMC is used in the Corona renderer (www.corona-renderer.com) to efficiently render caustics and their reflections. While the algorithm is based on our combination of vertex connection merging and MCMC, it differs in some way.

Our goal was to create a solver that would complement the main algorithm (path tracing) in difficult scenes while having easy setup and minimum overhead. To achieve this goal, we got an inspiration from Lightweight photon mapping [Grittmann et al. 2018] and we use only some of VCM techniques and only in places where they are necessary (e.g. where path tracing converges too slowly). We also automatically set number of light paths and all other parameters of VCM.

Furthermore, we have refined the MCMC target function to even better distribute the light paths. It was also necessary to devise an adaptive framework for selecting a light source during emission of light paths, since the real scenes can have thousands of light sources and not all of them generate difficult transport (e.g. caustics).

We refer the readers to our paper **Implementing One-Click Caustics in Corona Renderer** [Šik and Křivánek 2019] for more details.

CAUSTIC SOLVER - RESULTS



Here we show some results from the Corona renderer and its caustics solver. Before the solver was implemented, users had to use fakes to render glass in practical scenes (left image). After enabling the caustic solver, realistic caustics are computed using MCMC algorithm (right image), while the rendering is less than 2 times slower.

CAUSTIC SOLVER - RESULTS



The users can create scenes that contain both reflective and refractive caustics and the algorithm has enough temporal stability to enable rendering of animations. (Please visit the course webpage to view the animation)

**FURTHER READING
&
FUTURE WORK**

FURTHER READING

- **Survey of Markov Chain Monte Carlo Methods in Light Transport Simulation**
Martin Šik and Jaroslav Krivánek, 2018

The newest research in MCMC-based light transport:

- **Delayed Rejection Metropolis Light Transport**, [Rioux-Lavoie et al. 2020]
- **Stratified Markov chain Monte Carlo Light Transport**, [Gruson et al. 2020]
- **Langevin Monte Carlo Rendering with Gradient-based Adaptation**, [Luan et al. 2020]

For the readers interested about the recent development of MCMC techniques in light transport simulation, I recommend reading our survey, which covers most of these methods. The survey also points out directions for improving the existing MCMC algorithms. As for the newest development, I also recommend reading the papers shown on the slide.

FUTURE WORK

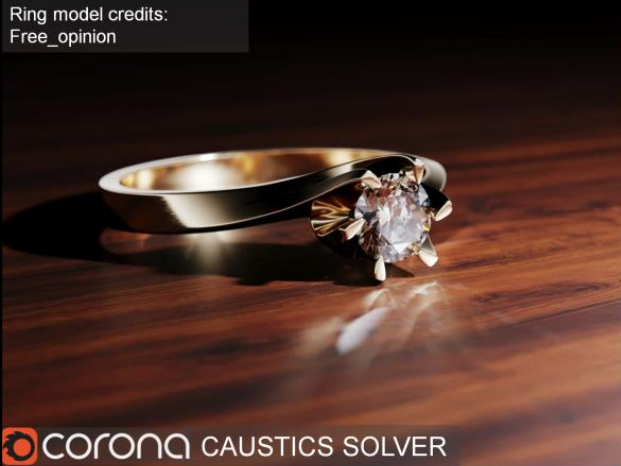
- One of the biggest issues: uninformed global mutation
- Informed mutations (preprocess or previous samples)
 - Adaptive MCMC [*Haario et al. 2001*]
- Possible combination: MCMC + path guiding


While our work has significantly improved global exploration, the issue is far from being solved. One of the biggest issues is the current use of an uninformed mutation to explore the whole state space. We believe that designing more informed “global” mutations could further improve global exploration. These mutations could learn the necessary information from some preprocess or during rendering from previous samples (utilizing so called adaptive Markov chain Monte Carlo [*Haario et al. 2001*]). There is a certain similarity with path guiding and thus an interesting avenue is the possible combination of path guiding and Markov chain Monte Carlo.

THANK YOU!



Ring model credits:
Free_opinion



 CORONA CAUSTICS SOLVER

THANK YOU FOR YOU ATTENTION!

Many thanks to all coauthors of the presented papers:

Adrien Gruson, Mickael Ribardiere, Jiří Vorba,
Rémy Cozot, Kadi Bouatouch, Hisanari Otsu,
Toshiya Hachisuka



Computer
Graphics
Charles
University

CHAOS
Czech

References

- [Baragatti et al. 2012] Meili Baragatti, Agnès Grimaud, and Denys Pommeret. Parallel tempering with equi-energy moves. *Statistics and Computing*, 23(3), 2012. ISSN 1573-1375.
- [Haario et al. 2001] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 04 2001.
- [Hachisuka and Jensen 2009] Toshiya Hachisuka and Henrik W. Jensen. Stochastic progressive photon mapping. *ACM Transactions on Graphics (SIGGRAPH Asia 2009)*, 28(5):141:1–141:8, 2009.
- [Hachisuka and Jensen 2011] Toshiya Hachisuka and Henrik W. Jensen. Robust adaptive photon tracing using photon path visibility. *ACM Transactions on Graphics*, 30(5):114:1–114:11, 2011. ISSN 0730-0301.
- [Hachisuka et al. 2012] Toshiya Hachisuka, Jacopo Pantaleoni, and Henrik W. Jensen. A path space extension for robust light transport simulation. *ACM Transactions on Graphics (SIGGRAPH Asia '12)*, 31(6):191:1–191:10, 2012. ISSN 0730-0301.
- [Hachisuka et al. 2014] Toshiya Hachisuka, Anton S. Kaplanyan, and Carsten Dachsbacher. Multiplexed Metropolis light transport. *ACM Transactions on Graphics*, 33(4): 100:1–100:10, July 2014. ISSN 0730-0301.
- [Hastings 1970] Wilfred K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. ISSN 00063444.
- [Georgiev et al. 2012] Iliyan Georgiev, Jaroslav Křivánek, Tomáš Davidovič, and Philipp Slusallek. Light transport simulation with vertex connection and merging. *ACM Transactions on Graphics (SIGGRAPH Asia '12)*, 31(6):192:1–192:10, 2012. ISSN 0730-0301.
- [Grittmann et al. 2018] Pascal Grittmann and Arsène Pérard-Gayot and Philipp Slusallek and Jaroslav Křivánek. Efficient Caustic Rendering with Lightweight Photon Mapping. *Computer Graphics Forum (Proceedings of the 29th Eurographics Symposium on Rendering)*, 2018.
- [Gruson et al. 2016] Adrien Gruson, Mickael Ribardiere, Martin Šik, Jiří Vorba, Rémy Cozot, Kadi Bouatouch, and Jaroslav Křivánek. A Spatial Target Function for Metropolis Photon Tracing. *ACM Trans. Graph.* 2016.
- [Gruson et al. 2020] Adrien Gruson, Rex West, and Toshiya Hachisuka. Stratified Markov chain Monte Carlo Light Transport, *Eurographics*, 2020.
- [Jakob and Marschner 2012] Wenzel Jakob and Steve Marschner. Manifold exploration: A Markov chain Monte Carlo technique for rendering scenes with difficult specular transport. *ACM Transactions on Graphics*, 31(4):58:1–58:13, 2012. ISSN 0730-0301.
- [Kaplanyan et al. 2014] Anton S. Kaplanyan, Johannes Hanika, and Carsten Dachsbacher. The natural constraint representation of the path space for efficient light transport simulation. *ACM Transactions on Graphics*, 33(4):102:1–102:13, July 2014. ISSN 0730-0301.

[Kelemen et al. 2002] Csaba Kelemen, László Szirmay-Kalos, Gyorgy Antal, and Ferenc Csonka. A simple and robust mutation strategy for the Metropolis light transport algorithm. *Computer Graphics Forum (Eurographics)*, 21(3):531–540, 2002. ISSN 1467-8659.

[Kitaoka et al. 2009] Shinya Kitaoka, Yoshifumi Kitamura, and Fumio Kishino. Replica exchange light transport. *Computer Graphics Forum*, 28(8):2330–2342, 2009. ISSN 1467-8659.

[Kou et al. 2006] S. C. Kou, Qing Zhou, and Wing H. Wong. Equi-energy sampler with applications in statistical inference and statistical mechanics. *The Annals of Statistics*, 34(4):1581–1619, 08 2006.

[Kulla et al. 2018] Christopher Kulla, Alejandro Conty, Clifford Stein, and Larry Israel. *ACM Transactions on Graphics* August 2018, 29, 2018.

[Li et al. 2015] Tzu-Mao Li, Jaakko Lehtinen, Ravi Ramamoorthi, Wenzel Jakob, and Frédo Durand. Anisotropic Gaussian mutations for Metropolis light transport through Hessian-Hamiltonian dynamics. *ACM Transactions on Graphics (SIGGRAPH Asia 2015)*, 34(6):209:1–209:13, 2015.

[Luan et al. 2020] Fujun Luan, Shuang Zhao, Kavita Bala, and Ioannis Gkioulekas. Langevin Monte Carlo Rendering with Gradient-based Adaptation. In *ACM Transactions on Graphics (To be presented at SIGGRAPH)*, 2020.

[Metropolis et al. 1953] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[Rioux-Lavoie et al. 2020] Damien Rioux-Lavoie, Joey Litalien, Adrien Gruson, Toshiya Hachisuka, and Derek Nowrouzezahrai. Delayed Rejection Metropolis Light Transport. In *ACM Transactions on Graphics (To be presented at SIGGRAPH)*, 2020.

[Swendsen and Wang 1986] Robert H. Swendsen and Jian-Sheng Wang. Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, 57:2607–2609, 1986.

[Šik 2018] Martin Šik. Global exploration in Markov chain Monte Carlo methods for light transport simulation, Doctoral thesis, 2018.

[Šik and Křivánek 2016] Martin Šik and Jaroslav Křivánek. Improving Global Exploration of MCMC Light Transport Simulation. *ACM SIGGRAPH 2016 Posters*.

[Šik and Křivánek 2018] Martin Šik and Jaroslav Křivánek. Survey of Markov Chain Monte Carlo Methods in Light Transport Simulation. *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[Šik and Křivánek 2019] Martin Šik and Jaroslav Křivánek. Implementing One-Click Caustics in Corona Renderer. *Eurographics Symposium on Rendering - Industry Track*, 2019.

[Šik et al. 2016] Martin Šik, Hisanari Otsu, Toshiya Hachisuka, and Jaroslav Křivánek. Robust Light Transport Simulation via Metropolised Bidirectional Estimators. *ACM Trans. Graph., SIGGRAPH Asia 2016*.

[Veach and Guibas 1994] Eric Veach and Leonidas Guibas. Bidirectional estimators for light transport. In Proc. Eurographics Rendering Workshop, pages 147–162, 1994.

[Veach and Guibas 1997] Eric Veach and Leonidas J. Guibas. Metropolis light transport. In Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '97, pages 65–76, 1997. ISBN 0-89791-896-7.

References

- (BFK18) Nikolaus Binder, Sascha Fricke, and Alexander Keller. Fast path space filtering by jittered spatial hashing. In *ACM SIGGRAPH 2018 Talks*, New York, NY, USA, 2018. Association for Computing Machinery.
- (BRDC12) Thomas Bashford-Rogers, Kurt Debattista, and Alan Chalmers. A significance cache for accelerating global illumination. *Comput. Graph. Forum*, 31(6):1837–1851, 2012.
- (DGJ+20) Stavros Diolatzis, Adrien Gruson, Wenzel Jakob, Derek Nowrouzezahrai, and George Drettakis. Practical Product Path Guiding Using Linearly Transformed Cosines. *Computer Graphics Forum*, 39(4), July 2020.
- (DK18) Ken Dahm and Alexander Keller. Learning light transport the reinforced way. In A. Owen and P. Glynn, editors, *Monte Carlo and Quasi-Monte Carlo Methods. MCQMC 2016. Proceedings in Mathematics & Statistics*, volume 241, pages 181–195. Springer, 2018.
- (DWWH20) Hong Deng, Beibei Wang, Rui Wang, and Nicolas Holzschuch. A Practical Path Guiding Method for Participating Media. *Computational Visual Media*, 6, 2020.
- (FHH+19) Luca Fascione, Johannes Hanika, Daniel Heckenberg, Christopher Kulla, Mark Droske, and Jorge Schwarzhaupt. Path tracing in production – part 1. In *SIGGRAPH Courses*, 2019.
- (GGSK19) Pascal Grittmann, Iliyan Georgiev, Philipp Slusallek, and Jaroslav Křivánek. Variance-aware multiple importance sampling. *ACM Trans. Graph.*, 38(6), November 2019.
- (GKDS12) Iliyan Georgiev, Jaroslav Křivánek, Tomáš Davidovič, and Philipp Slusallek. Light transport simulation with vertex connection and merging. *ACM Trans. Graph.*, 31(6), November 2012.
- (GKH+13) Iliyan Georgiev, Jaroslav Křivánek, Toshiya Hachisuka, Derek Nowrouzezahrai, and Wojciech Jarosz. Joint importance sampling of low-order volumetric scattering. *ACM Trans. Graph.*, 32(6), November 2013.
- (GMH+19) Iliyan Georgiev, Zackary Misso, Toshiya Hachisuka, Derek Nowrouzezahrai, Jaroslav Křivánek, and Wojciech Jarosz. Integral formulations of volumetric transmittance. *ACM Trans. Graph.*, 38(6), November 2019.
- (GRŠ+16) Adrien Gruson, Mickaël Ribardière, Martin Šik, Jiří Vorba, Rémi Cozot, Kadi Bouatouch, and Jaroslav Křivánek. A spatial target function for metropolis photon tracing. *ACM Trans. Graph.*, 36(1):4:1–4:13, November 2016.
- (HEV+16) Sebastian Herholz, Oskar Elek, Jiří Vorba, Hendrik Lensch, and Jaroslav Křivánek. Product importance sampling for light transport path guiding. *Computer Graphics Forum (Proceedings of Eurographics Symposium on Rendering)*, 35(4):67–77, 2016.
- (HP02) Heinrich Hey and Werner Purgathofer. Importance sampling with hemispherical particle footprints. In *Proceedings of the 18th Spring Conference on Computer Graphics*, pages 107–114, 2002.

- (HPJ12) Toshiya Hachisuka, Jacopo Pantaleoni, and Henrik W. Jensen. A path space extension for robust light transport simulation. *ACM Transactions on Graphics (SIGGRAPH Asia '12)*, 31(6):191:1–191:10, 2012.
- (HZE⁺19a) Sebastian Herholz, Yangyang Zhao, Oskar Elek, Derek Nowrouzezahrai, Hendrik P. A. Lensch, and Jaroslav Křivánek. Volume path guiding based on zero-variance random walk theory. *ACM Trans. Graph.*, 38(3), June 2019.
- (HZE⁺19b) Sebastian Herholz, Yangyang Zhao, Oskar Elek, Derek Nowrouzezahrai, Hendrik P. A. Lensch, and Jaroslav Křivánek. Volume path guiding based on zero-variance random walk theory. *ACM Trans. Graph.*, 38(3), June 2019.
- (Jen95) Henrik Wann Jensen. Importance driven path tracing using the photon map. pages 326–335, 1995.
- (Kd14) Jaroslav Křivánek and Eugene d’Eon. A zero-variance-based sampling scheme for monte carlo subsurface scattering. In *ACM SIGGRAPH 2014 Talks*, pages 1–1. 2014.
- (KDB14) Alexander Keller, Ken Dahm, and Nikolaus Binder. Path space filtering. In *ACM SIGGRAPH 2014 Talks*, 2014.
- (KGH⁺14) Jaroslav Křivánek, Iliyan Georgiev, Toshiya Hachisuka, Petr Vévoda, Martin Šik, Derek Nowrouzezahrai, and Wojciech Jarosz. Unifying points, beams, and paths in volumetric light transport simulation. *ACM Trans. Graph.*, 33(4), July 2014.
- (KGPB05) Jaroslav Křivánek, Pascal Gautron, Sumanta Pattanaik, and Kadi Bouatouch. Radiance caching for efficient global illumination computation. *Visualization and Computer Graphics, IEEE Transactions on*, 11(5):550–561, sept.-oct. 2005.
- (KKG⁺14) Jaroslav Křivánek, Alexander Keller, Iliyan Georgiev, Anton Kaplanyan, Marcos Fajardo, Mark Meyer, Jean-Daniel Nahmias, Ondřej Karlík, and Juan Canada. Recent advances in light transport simulation: Some theory and a lot of practice. In *ACM SIGGRAPH 2014 Courses, SIGGRAPH '14*, pages 17:1–17:6, New York, NY, USA, 2014. ACM.
- (KŠV⁺19) Ondřej Karlík, Martin Šik, Petr Vévoda, Tomáš Skřivan, and Jaroslav Křivánek. Mis compensation: Optimizing sampling techniques in multiple importance sampling. *ACM Trans. Graph.*, 38(6), November 2019.
- (KVG⁺19) Ivo Kondapaneni, Petr Vevoda, Pascal Grittmann, Tomáš Skřivan, Philipp Slusallek, and Jaroslav Křivánek. Optimal multiple importance sampling. *ACM Trans. Graph.*, 38(4), July 2019.
- (LW95) Eric P. Lafortune and Yves D. Willems. A 5d tree to reduce the variance of monte carlo ray tracing. In *Rendering Techniques '95 (Proc. of the 6th Eurographics Workshop on Rendering)*, pages 11–20, 1995.
- (MGN17) Thomas Müller, Markus Gross, and Jan Novák. Practical path guiding for efficient light-transport simulation. 36(4):91–100, June 2017.
- (MMR⁺19) Thomas Müller, Brian McWilliams, Fabrice Rousselle, Markus Gross, and Jan Novák. Neural importance sampling. *ACM Trans. Graph.*, 38(5), October 2019.

- (MRNK20) Thomas Müller, Fabrice Rousselle, Jan Novák, and Alexander Keller. Neural control variates, 2020.
- (Pan20) Jacopo Pantaleoni. Online path sampling control with progressive spatio-temporal filtering, 2020.
- (RGH⁺20) Alexander Rath, Pascal Grittmann, Sebastian Herholz, Petr Vévoda, Philipp Slusallek, and Jaroslav Křivánek. Variance-aware path guiding. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2020)*, 39(4), 2020.
- (RHL20) Lukas Ruppert, Sebastian Herholz, and Hendrik P. A. Lensch. Robust fitting of parallax-aware mixtures for path guiding. *ACM Trans. Graph.*, 39(4), 2020.
- (SJHD18) Florian Simon, Alisa Jung, Johannes Hanika, and Carsten Dachsbacher. Selective guided sampling with complete light transport paths. *Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 37(6), December 2018.
- (ŠK16) Martin Šik and Jaroslav Křivánek. Improving global exploration of MCMC light transport simulation. In *ACM SIGGRAPH 2016 Posters*, SIGGRAPH '16, pages 50:1–50:2, New York, NY, USA, 2016. ACM.
- (ŠK19a) Martin Šik and Jaroslav Křivánek. Implementing one-click caustics in corona renderer. In Tamy Boubekour and Pradeep Sen, editors, *Eurographics Symposium on Rendering - DL-only and Industry Track*, pages 61–67. The Eurographics Association, 2019.
- (ŠK19b) Martin Šik and Jaroslav Křivánek. Survey of Markov chain Monte Carlo methods in light transport simulation. *IEEE Transactions on Visualization and Computer Graphics*, 2019.
- (ŠOHK16) Martin Šik, Hisanari Otsu, Toshiya Hachisuka, and Jaroslav Křivánek. Robust light transport simulation via metropolised bidirectional estimators. *ACM Trans. Graph.*, 35(6):245:1–245:12, November 2016.
- (VHH⁺19) Jiří Vorba, Johannes Hanika, Sebastian Herholz, Thomas Müller, Jaroslav Křivánek, and Alexander Keller. Path guiding in production. In *ACM SIGGRAPH 2019 Courses*, SIGGRAPH '19, pages 18:1–18:77, New York, NY, USA, 2019. ACM.
- (VK16) Jiří Vorba and Jaroslav Křivánek. Adjoint-driven russian roulette and splitting in light transport simulation. *ACM Trans. Graph.*, 35(4), July 2016.
- (VKK18) Petr Vévoda, Ivo Kondapaneni, and Jaroslav Křivánek. Bayesian online regression for adaptive direct illumination sampling. *ACM Trans. Graph. (SIGGRAPH 2018)*, 37(4):125:1–125:12, July 2018.
- (VKŠ⁺14) Jiří Vorba, Ondřej Karlík, Martin Šik, Tobias Ritschel, and Jaroslav Křivánek. On-line learning of parametric mixture models for light transport simulation. *ACM Trans. on Graphics (Proc. SIGGRAPH)*, 33(4):101:1–101:11, August 2014.
- (Vor11) Jiří Vorba. Bidirectional photon mapping. CESC, 2011.
- (We17) C.J. Werner (editor). MCNP users manual - code version 6.2. *LA-UR-17-29981*, 2017.