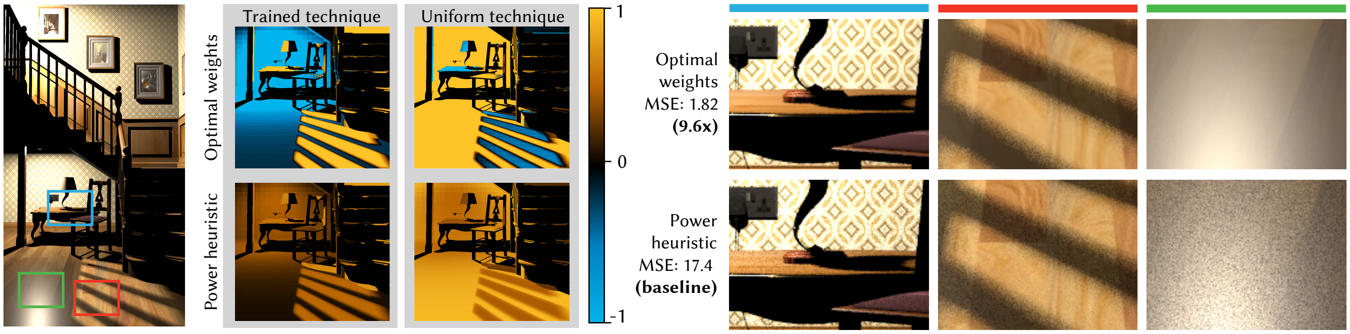
Optimal Multiple Importance Sampling

IVO KONDAPANENI*, Charles University, Prague
PETR VÉVODA*, Charles University, Prague and Render Legion, a. s.
PASCAL GRITTMANN, Saarland University
TOMÁŠ SKŘIVAN, IST Austria
PHILIPP SLUSALLEK, Saarland University and DFKI
JAROSLAV KŘIVÁNEK, Charles University, Prague and Render Legion, a. s.


Trained technique
Uniform technique
Optimal weights
Power heuristic
1
0
-1
Optimal weights MSE: 1.82 **(9.6x)**
Power heuristic MSE: 17.4 **(baseline)**

a) Reference b) MIS weights c) Equal-sample comparison

Fig. 1. Equal-sample comparison (20 per technique per pixel) of direct illumination estimated by an MIS combination of two light sampling techniques (*Trained* and *Uniform*, see Sec. 8.2 for details) with our optimal weights (top row) and the power heuristic (bottom row). The false-color images b) show per-pixel average MIS weight values as determined by the two weighting strategies. Unlike any of the existing MIS weighting heuristics, the optimal weights can have *negative* values, which provides additional opportunity for variance reduction, leading to an overall 9.6 times lower error per sample taken than the power heuristic in this scene.

Multiple Importance Sampling (MIS) is a key technique for achieving robustness of Monte Carlo estimators in computer graphics and other fields. We derive optimal weighting functions for MIS that provably minimize the variance of an MIS estimator, given a set of sampling techniques. We show that the resulting variance reduction over the balance heuristic can be higher than predicted by the variance bounds derived by Veach and Guibas, who assumed only non-negative weights in their proof. We theoretically analyze the variance of the optimal MIS weights and show the relation to the variance of the balance heuristic. Furthermore, we establish a connection between the new weighting functions and control variates as previously applied to mixture sampling. We apply the new optimal weights to integration problems in light transport and show that they allow for new design considerations when choosing the appropriate sampling techniques for a given integration problem.

CCS Concepts: • **Mathematics of computing → Probability and statistics**; • **Computing methodologies → Rendering**.

Additional Key Words and Phrases: Monte Carlo integration, Multiple Importance Sampling, combined estimators

ACM Reference Format:
Ivo Kondapaneni, Petr Vévoda, Pascal Grittmann, Tomáš Skřivan, Philipp Slusallek, and Jaroslav Křivánek. 2019. Optimal Multiple Importance Sampling. *ACM Trans. Graph.* 38, 4, Article 37 (July 2019), 14 pages. https://doi.org/10.1145/3306346.3323009

1 INTRODUCTION

Monte Carlo (MC) integration is an essential tool in light transport simulation [Pharr et al. 2016; Veach 1997] and other fields of science and engineering [Kalos and Whitlock 2008]. An inherent problem of MC integration is its slow convergence, which is why numerous variance reduction schemes have been proposed, notably importance sampling. Its extension, known as *multiple importance sampling* (MIS) [Veach and Guibas 1995], is particularly versatile as it enables combining different sampling techniques in a robust way to form better MC estimators.

In the context of light transport simulation, MIS has served as a cornerstone for robust bidirectional path sampling [Georgiev et al. 2012a; Hachisuka et al. 2012; Křivánek et al. 2014; Popov et al. 2015; Veach and Guibas 1995], Markov chain Monte Carlo light transport [Gruson et al. 2016; Hachisuka et al. 2014; Šik et al. 2016], adaptive path sampling (path guiding) [Herholz et al. 2016; Müller et al. 2017; Vorba et al. 2014], or in isolated integration problems such as direct

Authors' addresses: Ivo Kondapaneni, Charles University, Prague; Petr Vévoda, Charles University, Prague, Render Legion, a. s. Prague; Pascal Grittmann, Saarland University, Saarbrücken; Tomáš Skřivan, IST Austria, Vienna; Philipp Slusallek, Saarland University, Saarbrücken, DFKI, Saarbrücken; Jaroslav Křivánek, Charles University, Prague, Render Legion, a. s. Prague.
*Ivo Kondapaneni and Petr Vévoda share the first authorship of this work.

0730-0301/2019/7-ART37 $15.00
https://doi.org/10.1145/3306346.3323009

illumination estimation [Georgiev et al. 2012b; Veach and Guibas 1995; Vévoda et al. 2018].

The key to the efficiency of MIS are the weighting functions used to combine samples from different sampling techniques. A set of weighting functions known as the balance heuristic has been suggested as a *de facto* universal solution, as no other weights can yield substantially lower variance [Veach and Guibas 1995] (we show that this claim does not generally hold). Since the balance heuristic variance bounds can be fairly loose, alternative weights have been proposed to address shortcomings in some specific cases. The power, cut-off, or maximum heuristics can reduce variance for low-variance problems, but this comes at the expense of an overall variance increase [Veach and Guibas 1995]. The α-max heuristic incorporates prior assumptions to avoid assigning too high weights to poorly performing sampling techniques [Georgiev et al. 2012b]. However, the performance of different weighting heuristics is problem-specific and the existing work fails to provide a clear answer as to which weighting functions to use in which situation.

Our work focuses on weighting functions for MIS. We derive a set of weighting functions that *provably minimize the variance of the MIS estimator* for a given set of sampling techniques and a fixed number of samples. The resulting optimal weights *may be negative*, and this additional flexibility enables substantial variance reduction over the existing weighting heuristics. In fact, we show that the optimal weights can result in *variance lower than the balance heuristic bounds* derived by Veach and Guibas [1995], as non-negativity of the weights was a silent assumption made in their derivation.

We provide further theoretical insights into the new optimal weights: We establish a connection between MIS with our optimal weights and another common variance reduction scheme – control variates. Moreover, we relate the variance of the optimal weights and the balance heuristic. The derivation of the optimal MIS weights and their analysis comprise our main theoretical contribution.

Our practical contribution consists in proof-of-concept applications of the optimal weighting scheme in light transport, specifically in direct illumination calculation. Apart from the variance reduction afforded by using the optimal weights in an existing sampling setup, we show that the optimal weights allow for an additional flexibility in designing the sampling techniques themselves. More specifically, variance properties of the optimal weights directly motivate new sampling techniques that – while performing poorly with balance and power heuristics – provide a substantial speedup with our optimal weights.

2 PREVIOUS WORK

MIS in light transport. Multiple importance sampling (MIS) [Veach and Guibas 1995] offers a flexible way to combine a set of Monte Carlo integral estimators, so as to achieve reasonable performance in a wide range of scenarios – a property referred to as *robustness*. It has been one of the keys behind the recent success of physically-based light transport in VFX and computer animation [Keller et al. 2015]. MIS is typically used to combine a set of sampling techniques, each of which matches different features of the integrand, but none of which is a particularly good match across the entire domain. A prime example is direct illumination estimation [Veach and Guibas 1995], where MIS is used to mix BRDF- and light-sampling techniques. Likewise, bidirectional path tracing [Veach and Guibas 1995] and algorithms built upon it [Georgiev et al. 2012a; Hachisuka et al. 2012; Křivánek et al. 2014; Popov et al. 2015] rely on MIS to combine different techniques to sample entire light transport paths. In Markov chain Monte Carlo methods, MIS has been used to combine contributions from different chains [Kelemen et al. 2002; Šik et al. 2016] and to mix different target functions [Gruson et al. 2016].

Another important use-case for MIS is defensive sampling: An adaptively trained sampling distribution is combined with a defensive strategy to ensure robustness to over-fitting. In path guiding, adaptively constructed guiding distributions are typically mixed with BRDF sampling [Herholz et al. 2016; Müller et al. 2017; Vorba et al. 2014]. Similarly, in adaptive direct illumination sampling, MIS is used to combine learned light selection distributions with other, more defensive strategies [Georgiev et al. 2012b; Vévoda et al. 2018].

MIS estimator design. MIS represents a wide family of estimators parameterized by the combined sampling techniques, number of samples taken from each technique, and the weighting functions used to combine the samples. The choice of sampling techniques is application-dependent and we are not aware of any work addressing the sampling technique design specifically in the context of MIS. Another degree of freedom is the sample allocation. While Veach [1997] argues that “no strategy is much better than that of simply setting all [sample counts] equal”, the fixed sample allocation has its shortcomings. For instance, if one technique is particularly good, samples from other techniques only serve to incur overhead and increase variance. To determine the sample allocation among BSDF, light, and photon map-based sampling, Pajot et al. [2011] introduce the notion of “representativity” – a measure of how well each technique samples a given integrand. Similarly, Lu et al. [2013] optimize sample allocation among BSDF and environment-map sampling by approximately minimizing the MIS estimator variance in terms of the sample counts. Havran and Sbert [2014] and Sbert et al. [2016] show that the optimal sample allocation must equalize the second moment of the weighted estimates corresponding to the individual sampling techniques. Sbert and Havran [2017] use the above result to design an approximate sample allocation solution and Sbert et al. [2018] introduce new balance heuristic estimators better than the balance heuristic with equal sample count per technique. Finally, Cappé et al. [2008] apply population Monte Carlo to optimize sampling from mixture densities.

Alternative weighting heuristics. In our work, we assume the sample counts to be given and we focus on designing the optimal MIS *weighting functions* – a problem setup shared with several previous works. In the context of many-light sampling, Georgiev et al. [2012b] point out that the balance, power, and maximum heuristics perform poorly, and they introduce the α-max heuristic with the aim to achieve better stratification among the sampling techniques. Popov et al. [2015] introduce a new weighting heuristic accounting for correlations between paths in bidirectional path tracing. Elvira et al. [2015; 2016] propose clustering of sampling techniques to cut the overhead introduced by evaluating the balance heuristic when the number of sampling techniques is high. While these works design new weighting heuristic for some specific cases, our goal is more ambitious: The provably optimal MIS weighting functions (for a given set of sampling techniques and fixed sample allocation).

Control variates and mixture sampling. We show in Sec. 6 that our optimal weights are equivalent to optimal control variates (CV) [Lavenberg et al. 1982; Rubinstein and Marcus 1985; Venkatraman and Wilson 1986]. These were also studied by Owen and Zhou [2000], who realize CV by a mixture of sampling densities, and approximate the optimal CV coefficients through multiple linear regression over a set of observed estimates. We discuss the relation to their work in more detail in Sec. 7 and in the Supplemental material. Fan et al. [2006] then applied Owen and Zhou's approach in rendering, and we compare to their approach in Sec. 8.4. In the follow-up work [He and Owen 2014], the authors jointly optimize the CV coefficients and the sample allocation. They show that the MIS estimator variance is jointly convex in the above quantities and these can be found by convex optimization.

3 MULTIPLE IMPORTANCE SAMPLING

In this section, we review Monte Carlo (MC) integration, variance reduction via importance sampling, and multiple importance sampling (MIS), as first described by Veach and Guibas [1995].

Monte Carlo integration. Let $F = \int_D f(x)\,\mathrm{d}x$ be the integral of a function $f : D \to \mathbb{R}$ over the domain D, and let there be a *sampling technique* for generating random samples from D following the probability density p such that $f(x) \neq 0 \Rightarrow p(x) \neq 0$. Then the importance sampling estimator $\langle F \rangle = f(X)/p(X)$, where the random variable X is distributed according to p, is unbiased, i.e., its expected value $\mathrm{E}[F]$ equals to F. The shape of p has a dramatic impact on the estimator's variance $\mathrm{V}[\langle F \rangle]$: the closer p is to being proportional to the integrand f, the lower the variance.

Multiple importance sampling. The idea of MIS is to improve the robustness of MC integration by incorporating N sampling techniques with probability densities p_i, $i = 1, \ldots, N$, each of which could be a good match to a different feature of the integrand. An MIS estimator of the integral F is then defined as:

$$\langle F \rangle^* = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{w_i(X_{ij}) f(X_{ij})}{n_i p_i(X_{ij})}, \tag{1}$$

where $X_{ij} \in D$ is a random variable representing the j-th sample out of n_i samples generated by the i-th sampling technique, and $w_i(x)$ are *weighting functions*. All X_{ij} are independent. To keep the MIS estimator (1) unbiased, the weighting functions must satisfy:

$$f(x) \neq 0 \Rightarrow \textstyle\sum_{i=1}^{N} w_i(x) = 1, \tag{2}$$

$$p_i(x) = 0 \Rightarrow w_i(x) = 0, \tag{3}$$

i.e., they must sum up to 1 whenever $f(x)$ is nonzero, and each weight $w_i(x)$ must be zero whenever $p_i(x)$ is zero. A particular set of weighting functions is referred to as a *combination strategy*.

The above formulation of MIS, where a pre-determined number of samples are taken from each sampling technique, is known as the *multi-sample model*. On the other hand, the *one-sample model*

$$\langle F \rangle^{*1} = \frac{w_i(X_i) f(X_i)}{c_i p_i(X_i)}, \tag{4}$$

is evaluated by first selecting one sampling technique p_i at random with probability c_i, and then generating a sample X_i from it.

Balance and power heuristics. All combination strategies yield unbiased estimators, but they can differ in their variance. The two most commonly used combination strategies are the *balance* and *power* heuristics, sharing the common form

$$w_i^p(x) = \frac{[n_i p_i(x)]^\beta}{\sum_{k=1}^{N} [n_k p_k(x)]^\beta}. \tag{5}$$

For the *balance heuristic*, we have $\beta = 1$. Veach and Guibas [1995] showed that no other combination strategy can have significantly lower variance than the balance heuristic; we revisit this near-optimality claim below. The *power heuristic*, for $\beta > 1$, is a strategy better suited for low-variance problems, i.e., those where one p_i closely matches the integrand [Veach and Guibas 1995, Sec. 4.1]. We set $\beta = 2$, the choice that Veach and Guibas considered the best.

The same authors have additionally proposed the *cutoff* and *maximum* heuristics, but since these are used less frequently in practice and we do not consider them here further.

4 REVISITING BALANCE AND POWER HEURISTICS

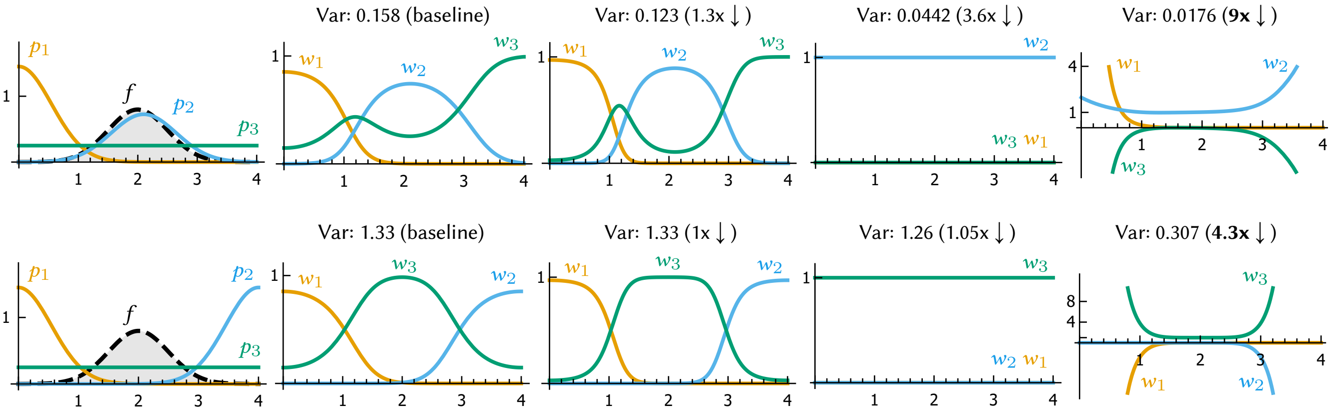
In this section, we first illustrate sub-optimal performance of the balance and power heuristics, we then revisit the balance heuristic variance bounds, and show that allowing for negative weights may yield far lower variance than predicted by the bounds.

4.1 Motivation

The balance and power heuristics enable combining sampling techniques in a robust way, so that the presence of a bad technique does not ruin the combined estimator's performance. But the robustness comes at the expense of decreased overall efficiency; the MIS combination can be far from optimal and sometimes significantly better results may be achieved by ignoring all samples but the ones taken from the single best technique.

Let us illustrate this observation on a simple 1D example shown in Fig. 2. Column a) depicts an integration problem where the integral of a function f is estimated via MIS. Three sampling techniques, p_1, p_2, and p_3, are used, and one sample is taken from each. The two rows differ solely in the sampling technique p_2: while p_2 closely matches f in the first row, in the second row it is fairly different. Columns b) and c) plot, respectively, the balance and the power heuristic weights. We additionally define the *best-technique heuristic*, depicted in column d), as the combination strategy assigning unit weight to the single technique yielding the lowest variance and zero to the others. We can now compare the variance of the balance, power, and best-technique heuristics.

While in the second row the variance of all the three strategies is similar, there is a significant difference in the first row. The power heuristic achieves somewhat lower variance (~0.123) than the balance heuristic (~0.158), as this case is an instance of the low-variance problem due to p_2 being a good match to the integrand. Nonetheless, the best-technique heuristic has by far the lowest variance (~0.0442), almost 3x lower than the power heuristic. This is an inherent problem of the balance and power heuristics; they are not optimal and sometimes much worse than using the best technique alone.



a) integrand and sampling techniques

b) balance heuristic weights

c) power heuristic weights

d) best-technique heuristic weights

e) optimal weights (unconstrained sign)

Fig. 2. a) The integrand f along with three sampling techniques p_1, p_2 and p_3. b-d) The weighting functions associated with the balance, power, and best-technique heuristic, respectively. e) Optimal weights (unconstrained sign). The two rows differ by the sampling technique p_2. See the Supplemental for additional results for the maximum and cutoff heuristics (slightly worse than the balance heuristic) and a Mathematica notebook used to produce this figure.

4.2 Balance heuristic variance bounds: Are they valid?

The balance heuristic is widely used for its robustness and because it is provably good: Veach [1997] has shown that a) for the multi-sample model, no other combination strategy can improve the variance beyond certain bounds, and b) it is optimal for the one-sample model. While the optimality proof for the one-sample model is valid in general, the proof of the variance bounds for the multi-sample model assumes non-negative weights that results in an entire class of combination strategies being omitted.

We now revisit the proof for the multi-sample model and point out that allowing negative weights (affine combinations rather than convex) can improve the variance beyond the bounds derived by Veach. To simplify the notation, we denote the inner product of two functions a and b defined over the domain D as $\langle a, b\rangle = \int_D a(x)b(x)\,\mathrm{d}x$.

According to Veach, the variance of a multi-sample MIS estimator utilizing the balance heuristic is no larger than the variance of *any* other MIS estimator plus some fraction of F^2, more precisely:

$$\mathrm{V}[\langle F\rangle^b] - \mathrm{V}[\langle F\rangle^*] \le \left(\frac{1}{\min_i n_i} - \frac{1}{\sum_{i=1}^N n_i}\right) F^2. \tag{6}$$

In the proof [Veach 1997, p. 288], the variance of an MIS estimator:

$$\mathrm{V}[\langle F\rangle^*] = \underbrace{\sum_{i=1}^N \int_D \frac{w_i(x)^2 f(x)^2}{n_i p_i(x)}\,\mathrm{d}x}_{\text{first term}} - \underbrace{\sum_{i=1}^N \frac{1}{n_i}\langle w_i, f\rangle^2}_{\text{second term}} \tag{7}$$

was inspected. While the balance heuristic was the result of the minimization of the first term (giving the optimum for the one-sample model), the variance bound $(1/_{\min_i n_i} - 1/_{\sum_{i=1}^N n_i})F^2$ was established as the difference of the upper and the lower bound of the second term in (7). The lower bound derivation did not rely on any specific assumption, but in the upper bound derivation:

$$\begin{aligned}\sum_{i=1}^N \frac{1}{n_i}\langle w_i, f\rangle^2 &\le \frac{1}{\min_i n_i}\sum_{i=1}^N \langle w_i, f\rangle^2 \\ &\overset{\star}{\le} \frac{1}{\min_i n_i}\left(\sum_{i=1}^N \langle w_i, f\rangle\right)^2 = \frac{1}{\min_i n_i} F^2,\end{aligned} \tag{8}$$

the second inequality ⋆ holds only if

$$\langle w_i, f\rangle \ge 0, \tag{9}$$

that is, in the context of rendering where the integrand is non-negative, only when $w_i(x) \ge 0$.[1] For $\langle w_i, f\rangle < 0$ the upper bound on the variance of the balance heuristic can in fact be *larger* than what Veach's result suggests.

To the best of our knowledge, this fact has not been previously recognized; the weighting functions are usually designed to be non-negative everywhere and for such the bounds are valid.

In what follows, we show that the non-negativity assumption is not necessary for an MIS estimator to remain unbiased. In fact, there are many cases where a combination strategy with $\langle w_i, f\rangle < 0$ produces an MIS estimator with variance lower than predicted by the bounds, and it can be significantly better than any other combination strategy considered by Veach [1997].

4.3 Weights with unconstrained sign: An example

Suppose we define weights allowing negative values for our integration problems from Sec. 4.1. One example of such weights is shown in Fig. 2e), along with the variance of the resulting estimators. They yield estimators with far lower variance than estimators utilizing any of the three heuristics discussed in Sec. 4.1.

For the integration problem in the second row, the MIS estimator using these weights has variance even *lower* than dictated by the variance bounds for the balance heuristic: the balance heuristic

[1] To be precise, the condition is slightly weaker, because a weighting function w_i negative in a part of the domain may still yield $\langle w_i, f\rangle \ge 0$.

variance is ∼1.3 and the bounds are ∼0.5, meaning that any other MIS estimator $\langle F\rangle^*$ with only positive weights should have variance above 0.8 (according to (6)). But the MIS estimator with the negative weights has variance ∼0.3, which is well below this threshold.

In the next section, we derive weighting functions that *provably minimize the variance of the MIS estimator*, should there be no constraint on the weights' sign. In fact, the weights used in Fig. 2e) resulted from that derivation.

5 OPTIMAL MIS WEIGHTS

We now derive optimal weights for MIS by directly minimizing the variance $\mathrm{V}[\langle F\rangle^*]$ of the combined estimator (1), without imposing any restrictions other than those necessary to obtain an unbiased estimator. More formally:

Problem 1. *Given the MIS estimator* (1), *minimize the functional* $V[w_1,\dots,w_N] = \mathrm{V}[\langle F\rangle^*]$ *in terms of weights* w_i*, while maintaining the constraints* $\sum_{i=1}^N w_i(x) = 1$ *and* $p_i(x) = 0 \Rightarrow w_i(x) = 0$*, and keeping the number of samples* n_i *and probability densities* p_i *fixed.*

To describe the solution let us first define some terms:

Definition 5.1. Let $f : D \to \mathbb{R}$ be a function to integrate, $p_i(x)$, $i = 1,\dots,N$ be a set of probability densities on D, and let n_i denote the number of samples taken from p_i. We define the technique matrix $\mathbf{A} = (a_{ik})$ as a symmetric $N \times N$ matrix with elements given by

$$a_{ik} = \left\langle p_i, p_k/(\textstyle\sum_{j=1}^N n_j p_j)\right\rangle, \tag{10}$$

and the contribution vector $\mathbf{b} = (b_1,\dots,b_N)^\intercal$ as a column vector of length N composed of

$$b_i = \left\langle f, p_i/(\textstyle\sum_{j=1}^N n_j p_j)\right\rangle. \tag{11}$$

The technique matrix is independent of the integrand f and it is composed of the inner products between all the probability densities normalized by the factor $(\sum_{i=1}^N n_i p_i)^{-1}$. Elements of the contribution vector represent contributions to the final $F = \int_D f(x)\,\mathrm{d}x$, because the dot product $(n_1,\dots,n_N)\cdot\mathbf{b}$ equals to the integral F.

The solution to Problem 1 can now be summarized as follows:

Theorem 5.2. *Let the column vector* $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_N)^\intercal$ *satisfy the system of linear equations*

$$\mathbf{A}\boldsymbol{\alpha} = \mathbf{b}, \tag{12}$$

where $\mathbf{A}$ *and* $\mathbf{b}$ *are the technique matrix and the contribution vector, respectively. Then the weighting functions*

$$w_i^o(x) = \alpha_i \frac{p_i(x)}{f(x)} + \frac{n_i p_i(x)}{\sum_{j=1}^N n_j p_j(x)}\left(1 - \frac{\sum_{j=1}^N \alpha_j p_j(x)}{f(x)}\right) \tag{13}$$

minimize the functional $V[w_1,\dots,w_N]$.

An MIS estimator using the weights $w_i^o(x)$ will be denoted $\langle F\rangle^o$. The proof of Theorem 5.2, given in Appendix B, employs the calculus of variations (Appendix A) to directly minimize the variance functional. It does not rely on any other assumptions than those necessary to ensure unbiasedness, and therefore the solution is indeed optimal in the MIS estimator family, i.e., *no other MIS combination strategy can result in a lower variance.*[2]

Due to the negative term in (13), the weights can be *negative*; the example in Sec. 4.3 shows that this indeed happens in practice.

Appendix C provides a discussion of the the existence and uniqueness of the optimal weights.

6 OPTIMAL WEIGHTS AS CONTROL VARIATES

In this section, we show that the optimal weights from Theorem 5.2 can be interpreted as control variates [Glasserman 2003]. Based on that we provide some intuition on the integration problems for which the optimal weights will yield the highest variance reduction.

6.1 Background: Control Variates

Consider an MC estimator $\langle F\rangle$ for the integral $F = \int f(x)\,\mathrm{d}x$. Take a set of K other estimators $\langle G_i\rangle$ with expected values G_i, $i = 1,\dots,K$, called control variates. Rewriting the original estimator $\langle F\rangle$ as

$$\begin{aligned}\langle F\rangle^{\mathrm{CV}} &= \langle F\rangle + \textstyle\sum_{i=1}^K \gamma_i (G_i - \langle G_i\rangle)\\ &= \textstyle\sum_{i=1}^K \gamma_i G_i + \langle F\rangle - \sum_{i=1}^K \gamma_i \langle G_i\rangle\end{aligned} \tag{14}$$

can reduce variance when some $\langle G_i\rangle$ is correlated with $\langle F\rangle$ and $\boldsymbol{\gamma} = (\gamma_1,\dots,\gamma_K)^\intercal$ is chosen appropriately. Variance is minimized for $\boldsymbol{\gamma}$ solving the system $\Sigma\boldsymbol{\gamma} = \boldsymbol{\sigma}$, where $\Sigma = (\sigma_{ik})$ is a $K \times K$ covariance matrix, and $\boldsymbol{\sigma} = (\sigma_1,\dots,\sigma_K)^\intercal$ is a covariance vector, with their elements defined as

$$\sigma_{ik} = \mathrm{Cov}[\langle G_i\rangle, \langle G_k\rangle], \qquad \sigma_i = \mathrm{Cov}[\langle G_i\rangle, \langle F\rangle]. \tag{15}$$

This is a well-known form [Lavenberg et al. 1982; Rubinstein and Marcus 1985; Venkatraman and Wilson 1986]. In the case of a single control variate ($K = 1$) variance is minimized for $\gamma_1 = \mathrm{Cov}[\langle G_1\rangle, \langle F\rangle]/\mathrm{V}[\langle G_1\rangle]$.

6.2 Optimal weights as control variates

Let us plug the optimal weights from (13) into the multi-sample MIS estimator in (1). Denoting $M = \sum_{i=1}^N n_i$, $c_i = n_i/M$, and $p_\mathbf{c}(x) = \sum_{i=1}^N c_i p_i(x)$, we obtain the optimal MIS estimator $\langle F\rangle^o$ in the form

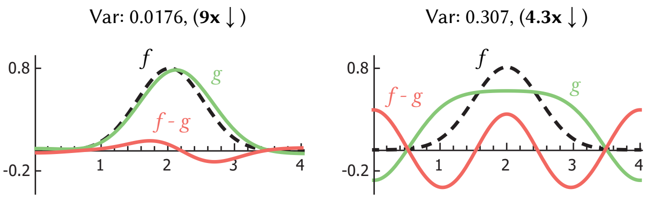
$$\langle F\rangle^o = \sum_{i=1}^N \alpha_i + \frac{1}{M}\sum_{i=1}^N\sum_{j=1}^{n_i}\left(\frac{f(X_{ij})}{p_\mathbf{c}(X_{ij})} - \frac{\sum_{k=1}^N \alpha_k p_k(X_{ij})}{p_\mathbf{c}(X_{ij})}\right). \tag{16}$$

The above form can be interpreted as the control variate estimator (14) utilizing one or N control variates. Here, for the purpose of further analysis, we interpret it as the former: Using $g(x) = \sum_{k=1}^N \alpha_k p_k(x)$, the above form is equivalent to (14) with $K = 1$, where

$$\langle F\rangle = \frac{1}{M}\sum_{i=1}^N\sum_{j=1}^{n_i}\frac{f(X_{ij})}{p_\mathbf{c}(X_{ij})}, \qquad \langle G_1\rangle = \frac{1}{M}\sum_{i=1}^N\sum_{j=1}^{n_i}\frac{g(X_{ij})}{p_\mathbf{c}(X_{ij})}, \tag{17}$$

the expected value $G_1 = \int \sum_{k=1}^N \alpha_k p_k(x)\,\mathrm{d}x = \sum_{k=1}^N \alpha_k$, and the parameter $\gamma_1 = 1$. The estimator $\langle F\rangle$ above is a multi-sample MIS estimator of F utilizing the balance heuristic, further denoted $\langle F\rangle^b$.

[2] Applies to combination strategies in MIS framework (1) as defined by Veach and Guibas [1995]. Other ways of combining samples (e.g. nonlinear ones) may perform still better, but these do not belong to the MIS family.



a) first row

b) second row

Fig. 3. Illustration of the difference $f - g$ for the first (a) and second (b) row of the integration problem from Fig. 2a along with the variance of MIS using the optimal weights, and the variance reduction with respect to MIS using the balance heuristic. Note, that the flatter the difference the higher the variance reduction.

Similarly, the $\langle G_1 \rangle$ estimator above is an MIS estimator of $\int_D g(x)\,\mathrm{d}x$, and we denote it $\langle G \rangle^b$.

6.3 Variance considerations

The $\boldsymbol{\alpha}$ vector from Theorem 5.2 yields an optimal control variate of the general form (16), minimizing its variance.[3] The variance is then equal to the variance of the balance heuristic MIS estimator of $\int_D f(x) - g(x)\,\mathrm{d}x$, and as such it depends on the magnitude of $f - g$ as well as its proportionality to $p_\mathbf{c}$. Intuitively, the "closer" the function g is in its shape to the integrand f, the higher the variance reduction due to the optimal weights compared to the balance heuristic. Moreover, the variance of $\langle F \rangle^o$ becomes zero for $f = g$, that is, when the integrand f can be written as a linear combination of the sampling pdfs p_k.

In Fig. 3 we plot the difference $f - g$ for the two integration problems from Sec. 4.3, where g is computed using the vector $\boldsymbol{\alpha}$ for the respective optimal weights. The overall amplitude of the difference is smaller for the first example and larger for the second, which is in line with the higher variance reduction for the former case. We build on these observations in Sec. 8.3 to design new sampling techniques specifically aiming at variance reduction with the optimal weights.

Relation to the balance heuristic. The optimal estimator $\langle F \rangle^o$ is given by the sum $\sum_{i=1}^{N} \alpha_i$ (no variance) plus the difference of two correlated MIS estimators $\langle F \rangle^b$ and $\langle G \rangle^b$, given by (17). The variance of $\langle F \rangle^o$ is therefore equal to the variance of that difference, i.e., $\mathrm{V}[\langle F \rangle^o] = \mathrm{V}[\langle F \rangle^b - \langle G \rangle^b]$. In Appendix D we prove that

$$\mathrm{V}[\langle F \rangle^o] = \mathrm{V}[\langle F \rangle^b] - \mathrm{V}[\langle G \rangle^b]. \tag{18}$$

This result confirms the expected: the optimal estimator's variance is less than or equal to the balance heuristic variance. More importantly, it shows that the balance heuristic is optimal whenever $\mathrm{V}[\langle G \rangle^b] = 0$. This occurs when $\boldsymbol{\alpha}$ is collinear with the vector $\mathbf{n} = (n_1, \ldots, n_N)^\mathsf{T}$, that is, when the elements of the vector $\boldsymbol{\alpha}$ are proportional to the number of samples from the individual sampling techniques. This result can be used to detect the achievable variance improvement over the balance heuristic.

Covariance vector and matrices. Interpreting (16) as a form utilizing N control variates

$$\langle G_k \rangle = \frac{1}{M} \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{p_k(X_{ij})}{p_\mathbf{c}(X_{ij})}, \quad k = 1, \ldots, N, \tag{19}$$

with expected values $G_k = 1$, we can verify that $\boldsymbol{\alpha}$ indeed represents the optimal parameters $\boldsymbol{\gamma}$. The technique matrix $\mathbf{A}$ and contribution vector $\mathbf{b}$ in Theorem 5.2 are related to their covariance counterparts (defined by (15)) by

$$\Sigma = (\mathbf{I} - \mathbf{A}\mathbf{N})\mathbf{A}, \qquad \boldsymbol{\sigma} = (\mathbf{I} - \mathbf{A}\mathbf{N})\mathbf{b}, \tag{20}$$

where $\mathbf{N}$ is a diagonal $N \times N$ matrix with the sample count n_i along the diagonal. The above relation emerges if we obtain the covariances σ_{ik} and σ_i in a similar way we obtained the covariance (38) in Appendix D. It follows that the full solution for alphas (see Appendix C) solves the system $\Sigma\boldsymbol{\gamma} = \boldsymbol{\sigma}$.

7 OPTIMAL WEIGHTS IN PRACTICE

An MIS estimator with the optimal weights (13) cannot be evaluated directly since the inner products forming the technique matrix $\mathbf{A}$ and contribution vector $\mathbf{b}$ from *Definition 5.1* generally do not have a closed form solution. Our implementation therefore follows three steps: 1) estimation of the technique matrix $\mathbf{A}$ and contribution vector $\mathbf{b}$; 2) estimation of the vector $\boldsymbol{\alpha}$ using the estimated $\mathbf{A}$ and $\mathbf{b}$; and 3) realization of an approximate optimal estimator $\langle F \rangle^o$ using the estimated $\boldsymbol{\alpha}$. We now elaborate on the individual steps.

7.1 Estimating A and b

The elements of the technique matrix $\mathbf{A}$ and the contribution vector $\mathbf{b}$ are given by the integrals (10) and (11), respectively. We estimate these integrals using MIS with the balance heuristic, and denote the result $\langle \mathbf{A} \rangle$ and $\langle \mathbf{b} \rangle$.[4] In the matrix form, the estimators $\langle \mathbf{A} \rangle$ and $\langle \mathbf{b} \rangle$ can be expressed as follows:

$$\langle \mathbf{A} \rangle = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \mathbf{W}_{ij} \mathbf{W}_{ij}^\mathsf{T}, \quad \langle \mathbf{b} \rangle = \sum_{i=1}^{N} \sum_{j=1}^{n_i} f(X_{ij})\, S_{ij}\, \mathbf{W}_{ij}, \tag{21}$$

where $S_{ij} = \left(\sum_{k=1}^{N} n_k p_k(X_{ij}) \right)^{-1}$ and $\mathbf{W}_{ij}$ is the column vector of all sampling techniques evaluated at X_{ij} and scaled by S_{ij},

$$\mathbf{W}_{ij} = S_{ij} \left(p_1(X_{ij}), \ldots, p_N(X_{ij}) \right)^\mathsf{T}. \tag{22}$$

Recall from (1) that X_{ij} denotes the j-th sample from p_i.

7.2 Estimating the vector $\boldsymbol{\alpha}$

The vector $\boldsymbol{\alpha}$ is given by the linear system (12). We estimate $\langle \boldsymbol{\alpha} \rangle$ by least squares minimization, because the *estimated* system $\langle \mathbf{A} \rangle \langle \boldsymbol{\alpha} \rangle = \langle \mathbf{b} \rangle$ may be (close to) singular, especially when the estimates $\langle \mathbf{A} \rangle$ and $\langle \mathbf{b} \rangle$ are based on just a few samples. While the $\langle \mathbf{A} \rangle$ and $\langle \mathbf{b} \rangle$ estimates are unbiased, the estimate $\langle \boldsymbol{\alpha} \rangle$ is generally biased, because the matrix inversion involved in solving the linear system does not preserve expectation, i.e. $(\mathrm{E}[\langle \mathbf{A} \rangle] = \mathbf{A}) \not\Rightarrow (\mathrm{E}[\langle \mathbf{A} \rangle^{-1}] = \mathbf{A}^{-1})$. Nonetheless, we can see from (16) that the resulting MIS estimator will be unbiased for any value of $\boldsymbol{\alpha}$. The difference between the true

[3] If it was not the optimum, then other weights better than $w_i^o(x)$ would exist, which is a contradiction.

[4] The power heuristic is less appropriate, as the integrals (10) and (11) are not low-variance, i.e., no sampling strategy is a particularly good match for any of the integrands.

```
ALGORITHM 1: Progressive estimator
$\langle \mathbf{A} \rangle \leftarrow 0^{N\times N}$; $\langle \mathbf{b} \rangle \leftarrow 0^{N\times 1}$; $\langle \boldsymbol{\alpha} \rangle \leftarrow 0^{N\times 1}$; $result \leftarrow 0$;
for $iteration \leftarrow 0$ to $maxIterations - 1$ do
    for $i \leftarrow 1$ to $N$ do
        $\{X_{ij}\}_{j=1}^{n_i} \leftarrow$ draw $n_i$ samples from technique $p_i$;
    end
    if $(iteration \geq 1)$ and $(iteration \bmod U) = 0$ then
        $\langle \boldsymbol{\alpha} \rangle \leftarrow$ solve linear system $\langle \mathbf{A} \rangle \langle \boldsymbol{\alpha} \rangle = \langle \mathbf{b} \rangle$;
    end
    $estimate \leftarrow$ evaluate $\langle F \rangle^o$ using $\langle \boldsymbol{\alpha} \rangle$;        // (16)
    $result \leftarrow result + estimate$;
    $\langle \mathbf{A} \rangle \leftarrow \langle \mathbf{A} \rangle + \sum_{i=1}^{N} \sum_{j=1}^{n_i} \mathbf{W}_{ij} \mathbf{W}_{ij}^{\mathsf{T}}$;        // (21)
    $\langle \mathbf{b} \rangle \leftarrow \langle \mathbf{b} \rangle + \sum_{i=1}^{N} \sum_{j=1}^{n_i} f(X_{ij})\, S_{ij}\, \mathbf{W}_{ij}$;        // (21)
end

return $result / maxIterations$
```

```
ALGORITHM 2: Direct estimator
$\langle \mathbf{A} \rangle \leftarrow 0^{N\times N}$; $\langle \mathbf{b} \rangle \leftarrow 0^{N\times 1}$;
for $iteration \leftarrow 0$ to $maxIterations - 1$ do
    for $i \leftarrow 1$ to $N$ do
        $\{X_{ij}\}_{j=1}^{n_i} \leftarrow$ draw $n_i$ samples from technique $p_i$;
    end





    $\langle \mathbf{A} \rangle \leftarrow \langle \mathbf{A} \rangle + \sum_{i=1}^{N} \sum_{j=1}^{n_i} \mathbf{W}_{ij} \mathbf{W}_{ij}^{\mathsf{T}}$;        // (21)
    $\langle \mathbf{b} \rangle \leftarrow \langle \mathbf{b} \rangle + \sum_{i=1}^{N} \sum_{j=1}^{n_i} f(X_{ij})\, S_{ij}\, \mathbf{W}_{ij}$;        // (21)
end
$\langle \boldsymbol{\alpha} \rangle \leftarrow$ solve linear system $\langle \mathbf{A} \rangle \langle \boldsymbol{\alpha} \rangle = \langle \mathbf{b} \rangle$;
return $\sum_{i=1}^{N} \langle \alpha_i \rangle$
```

Fig. 4. Pseudocode for two estimators with the approximated optimal MIS weights: the Progressive and Direct estimators (see Sec. 7.3). Differences are highlighted in red.

$\boldsymbol{\alpha}$ and its particular estimate $\langle \boldsymbol{\alpha} \rangle$ introduces extra variance in the final estimator $\langle F \rangle^o$. The extra variance diminishes thanks to the $\langle \boldsymbol{\alpha} \rangle$ estimate being *consistent*; this follows from $\langle \mathbf{A} \rangle^{-1}$ approaching $\mathbf{A}^{-1}$ with the increasing sample count in the $\langle \mathbf{A} \rangle$ estimate.

7.3 Approximate optimal estimator $\langle F \rangle^o$

We have various options to approximate the optimal estimator $\langle F \rangle^o$. For instance, we could estimate $\langle \boldsymbol{\alpha} \rangle$ from an initial batch of samples, hold it fixed, and use it to evaluate the optimal weights (13) for all subsequent samples. This approach would be suboptimal, however, as the estimated alphas would not evolve over time.

Progressive estimator. A more efficient option is to realize the approximate estimator in a progressive manner. The computation is performed in iterations. In each iteration, we first draw n_i samples from each sampling technique p_i, $i = 1, \ldots, N$. We then compute $\langle \boldsymbol{\alpha} \rangle$ based on the $\langle \mathbf{A} \rangle$ and $\langle \mathbf{b} \rangle$ estimates from the previous iterations. We plug it in formula (16) of the MIS estimator $\langle F \rangle^o$ to compute the integral estimate from the current samples and accumulate it. Note that for the first iteration we set $\langle \boldsymbol{\alpha} \rangle$ to zero which is equivalent to evaluating an MIS estimator with the balance heuristic. Finally, we update $\langle \mathbf{A} \rangle$ and $\langle \mathbf{b} \rangle$ using the current samples according to (21) and proceed to the next iteration. This procedure yields an unbiased MIS estimator, the efficiency of which improves over time as the estimates $\langle \boldsymbol{\alpha} \rangle$ converge to the true value. Since recomputing $\langle \boldsymbol{\alpha} \rangle$ every iteration may be time-consuming, we also allow for performing the recomputation only after every U updates to $\langle \mathbf{A} \rangle$ and $\langle \mathbf{b} \rangle$. We will call U the update step. See Algorithm 1 in Fig. 4 for a pseudocode.

An important note: Despite the division by the integrand $f(x)$ in the optimal weights (13), the MIS estimator $\langle F \rangle^o$ in the form (16) exists even for $f(x) = 0$. In contrast to previous MIS weighting heuristics, samples X with $f(X) = 0$ *must not be discarded*, because they generally have a non-zero contribution to the estimator.

Direct estimator. By definition (see (29) in Appendix B), each α_i is equal to the integral of f weighted by the optimal weight w_i^o:

$$\alpha_i = \int_D f(x) w_i^o(x)\, \mathrm{d}x. \tag{23}$$

Because the weighting functions sum up to one for all $x \in D$, we can express the integral of f as

$$\int_D f(x)\, \mathrm{d}x = \int_D f(x) \left(\textstyle\sum_{i=1}^{N} w_i^o(x) \right) \mathrm{d}x = \textstyle\sum_{i=1}^{N} \alpha_i. \tag{24}$$

We can therefore obtain an estimator $\langle F \rangle$ by summing the elements of $\langle \boldsymbol{\alpha} \rangle$. Such a *Direct estimator* will be biased, but consistent as follows from biasedness and consistency of $\langle \boldsymbol{\alpha} \rangle$, discussed in Sec. 7.2.

The Direct estimator is simpler and more efficient than the Progressive one: in each iteration, it only updates the $\langle \mathbf{A} \rangle$ and $\langle \mathbf{b} \rangle$ estimates, and the linear system is solved for $\langle \boldsymbol{\alpha} \rangle$ only once after all iterations have been processed. See Algorithm 2 in Fig. 4.

7.4 Empirical tests

Fig. 5 illustrates the behavior of the Progressive and Direct estimators, described above, on the example integration problem from Sec. 4.1 (depicted in Fig. 2a). The MSE of different estimators as a function of the number of iterations is shown in Fig. 5a. The *uncorrelated* version uses two independent sets of samples to estimate the technique matrix $\langle \mathbf{A} \rangle$ and the contribution vector $\langle \mathbf{b} \rangle$, respectively. The *correlated* version uses a single sample set for both.

In the correlated case (solid lines), both Progressive (cyan) and Direct (orange) estimators have similar performance, almost as good as the reference optimal estimator with a known $\boldsymbol{\alpha}$ vector (solid black). Interestingly, the behavior in the uncorrelated case (dashed lines) is vastly different, as both estimators perform much worse than in the correlated case. We hypothesize that the correlation between $\langle \mathbf{A} \rangle$ and $\langle \mathbf{b} \rangle$ is the key to a good performance of both estimators, though a full understanding of this effect remains for future work.

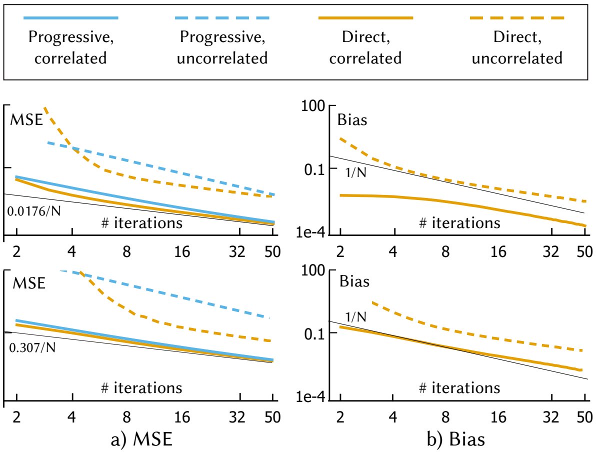


Fig. 5. a) MSE of the Progressive and Direct estimators versus the overall number of iterations plotted on the log-log scale, when used to estimate the first (top row) and second (bottom row) integration problem from Fig. 2a. The black line represents the analytically computed variance of MIS estimator with the optimal weights divided by N iterations. b) Bias of the Direct estimator on the log-log scale. The black line corresponds to $1/N$, where N is a number of iterations on the horizontal axis. For both (a) and (b) cases we show the correlated and uncorrelated estimator variants.

The Direct estimator is biased. In Fig. 5b, we can observe that both the correlated and uncorrelated versions are consistent, with the bias diminishing roughly at a $O(N^{-1})$ rate with the total number of iterations.[5] Similarly to the MSE, the bias is much lower in the correlated case. As discussed above, the Progressive estimator is unbiased, which we have verified experimentally.

7.5 Discussion of related work

Interestingly, the optimal estimator (16) has the same form as the control variate estimator analyzed by Owen and Zhou [2000]. They start off by postulating this form, using the mixture of sampling pdfs as a control variate, and then they estimate the optimal mixing parameters $\boldsymbol{\alpha}$ for this stated estimator form. We, on the other hand, show that both the form and the parameters of this estimator naturally emerge by direct minimization of the MIS estimator's variance, and that it provides the optimal solution in the MIS family.

Owen and Zhou estimate $\boldsymbol{\alpha}$ using linear regression on observed samples. For that they have to solve a (singular) linear system, but they also propose solving an equivalent (regular) truncated system, obtained by skipping some regressors. Though derived in a different way, their proposed $\boldsymbol{\alpha}$ estimator (denoted as $\hat{\boldsymbol{\beta}}$ in their Sec. 3), even in its truncated form, is in fact equivalent to our $\langle\boldsymbol{\alpha}\rangle$, *provided that the components of our technique matrix* $\mathbf{A}$ *and the contribution vector* $\mathbf{b}$ *are estimated with the balance heuristic as described in Sec. 7.1.* Hence, their approach can be seen as one particular way of approximating the optimal solution given by Theorem 5.2. Our result is more general as it is amenable to alternative strategies to approximate the optimal $\mathbf{A}$, $\mathbf{b}$, and $\boldsymbol{\alpha}$. See Sec. 3 in the Supplemental for details.

8 APPLICATIONS AND RESULTS

In this section we apply the optimal weights to light transport, specifically to direct illumination estimation. We show that they perform particularly well when used for defensive sampling. Subsequently, we introduce new sampling techniques that further increase the efficiency when mixed by the optimal weights. Furthermore, we compare the performance of the Progressive and Direct estimators.

8.1 Implementation

Our applications are implemented in PBRT [Pharr et al. 2016], and the implementation source code is provided in Sec. 5 in the Supplemental material. All scenes were rendered on a machine with an Intel Core i7-5820K CPU (6 cores, 12 threads) and 64GB of RAM.

We implement the Progressive and Direct estimators as defined in Sec. 7. Calculation proceeds pixel-by-pixel, in each pixel the respective algorithm from Fig. 4 is called and its output is stored in the pixel. We take one sample per techniques per iteration, i.e., $n_i = 1, i = 1, \ldots, N, N = 2$ and set $maxIterations$ to the target number of samples per technique per pixel. For an equal-time comparison we set $maxIterations$ individually for each estimator so they all render for roughly the same time.

In Sec. 8.2 and Sec. 8.3 we compare our Direct estimator to the balance and power heuristic combinations for two different applications. In Sec. 8.4 we compare the Direct and Progressive approaches.

8.2 Application I: Defensive sampling

One application where the optimal MIS weights have a particularly strong impact is defensive sampling. It is typically employed by adaptive approaches that construct sampling distributions based on previous samples [Georgiev et al. 2012b; Herholz et al. 2016]. The trained sampling technique is then mixed with one or more defensive techniques (e.g., uniform) to prevent bias and artifacts due to noise from the previous samples. Ideally, the trained technique has low variance across the majority of the domain, which is likely to trigger the low-variance problem discussed by Veach and Guibas [1995]. However, the power, maximum, and cutoff heuristics, proposed to address this case, still underperform [Georgiev et al. 2012b]. While the heuristics improve robustness, they also increase variance where the trained technique works well.

Our optimal MIS weights are particularly effective at solving this issue: the optimal combination of multiple sampling techniques can never be worse than a single technique on its own.[6] Therefore, no ad-hoc solutions are required and combinations with any number of defensive techniques is straightforward. We demonstrate this on a synthetic example as well as on a practical problem of light selection in direct illumination computation.

Synthetic example. Our simple example in the first row in Fig. 2 shows a combination of the almost ideal technique p_2 with defensive techniques p_1 and p_3. We can see that while the balance and power heuristic combinations produce more variance than the p_2 technique alone, with the optimal weights the variance is actually decreased.

[5] Bias is computed as the average absolute error of 1000 independent estimator realizations, each obtained using the number of samples on the horizontal axis.

[6] Using a single technique on its own is identical to a weighting strategy assigning unit weight to that technique and zero to all other techniques.

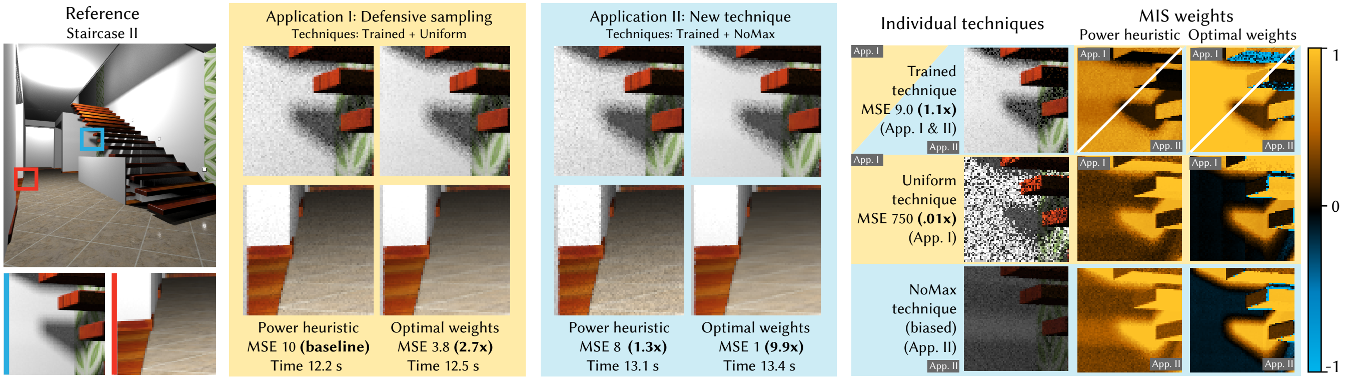


Fig. 6. Equal-sample comparison (20 per technique per pixel) of different combination strategies for a trained light selection technique (*Trained*) and defensive techniques (*Uniform*, *NoMax*). In contrast to the power heuristic, the optimal MIS weights are never worse than any of the techniques alone. The false color insets correspond to average weights per pixel for the three techniques. The MSE improvement in parentheses is with respect to the power heuristic combination of the *Trained* and *Uniform* techniques. See the Supplemental material for full-size images.

Light selection. MC estimation of the direct illumination often contributes a significant amount of noise to the image [Vévoda et al. 2018]. Direct illumination is computed as an integral $F_{\mathrm{DI}} = \int_A L_{\mathrm{e}} BVG \,\mathrm{d}y$ (we omitted arguments for brevity), where L_{e} is the emitted radiance, B the BRDF, V the visibility, G the geometry factor, and the domain A is the set of all emissive surfaces. A standard approach to design a direct illumination estimator is to first randomly select one light according to a light selection distribution and then sample a point on the selected light. A good light selection technique would select a light proportionally to its actual contribution to the integral. Unfortunately, this quantity cannot be computed analytically, especially because of the possibly complex visibility factor. It can, however, be estimated. While the resulting light selection technique is often close to ideal, error in the estimation may significantly increase variance or introduce bias. To prevent this, any such technique has to be combined with a defensive one (e.g., uniform light selection).

We demonstrate this approach on a particular light selection technique implemented in PBRT [Pharr et al. 2016]. It divides the scene using a regular grid, estimates the unoccluded contribution of all lights in each of its cells, and then uses these estimates as the light selection probabilities. We call this technique *Trained*. It is close to optimal on unoccluded surfaces but causes significant noise in shadows and must be combined with a defensive *Uniform* light selection technique.

Fig. 6 shows the results in the *Staircase II* scene lit by several small area light sources. All images using the optimal weights were rendered by the Direct estimator (Sec. 7) using the same number of samples per technique per pixel (20 samples). The *Trained* technique performs well on unoccluded surfaces but produces more noise than the *Uniform* technique in shadows. Intuitively, we would like to combine both techniques in the shadows and use the *Trained* technique alone on the unoccluded surfaces. However, the false color insets show that the power heuristic gives the uniform technique a positive weight everywhere, improving the performance in the shadows, and degrading the quality on the unoccluded surfaces. On the other hand, the optimal weights are zero or even negative on the unoccluded surfaces. As a result, the optimal weights maintain the good properties of both techniques everywhere and thus achieve 2.7× lower mean-squared error per sample. See the Supplemental material for complete results including the balance heuristic which performs 1.2× worse than the power heuristic.

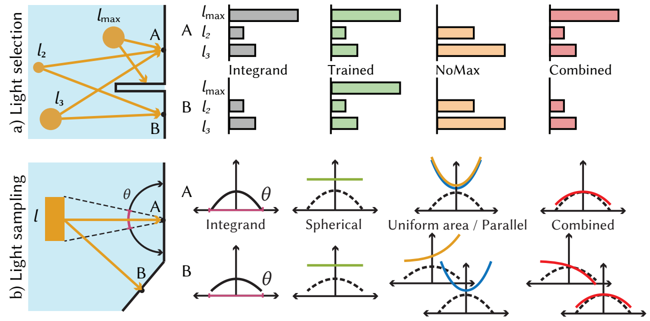


Fig. 7. a) Schematics illustrating the optimal combination of techniques *Trained* and *NoMax* for the light selection application, which can well approximate the integrand at both points A and B. b) Schematics for the light sampling application illustrating the optimal combination of techniques *Spherical*, *Uniform area*, and *Parallel*. At point B, where the surface is not parallel to the light, the optimal combination *Spherical* + *Uniform area* approximates the integrand much worse, while the optimal combination *Spherical* + *Parallel* is still good. The displayed quantities are in the solid angle measure, their derivations can be found in Sec. 2 in the Supplemental.

8.3 Application II: Design of new sampling techniques

As discussed in Sec. 6, the optimal weights form a control variate as a linear combination of the sampling pdfs i.e., as $\sum_{i=1}^{N} \alpha_i p_i$. We have shown that the closer the control variate approximates the integrand, the lower the variance. Introducing a new, properly designed technique (even a biased one!) can vastly expand the space of possibilities for the optimal weights to form a control variate closer to an integrand, and therefore can greatly improve the performance.

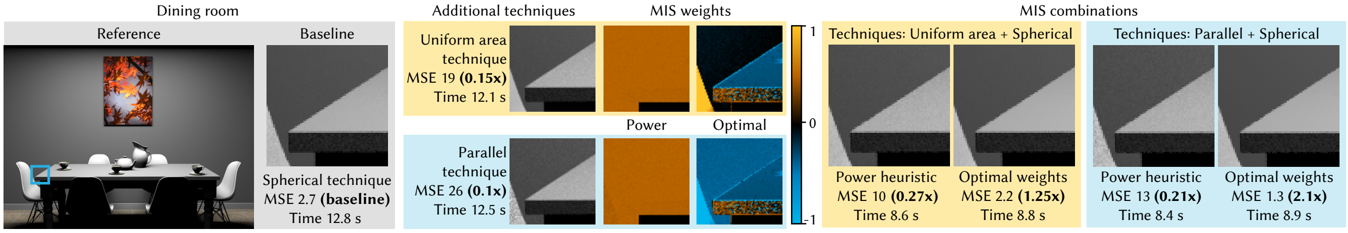


Fig. 8. Equal-sample comparison (40 per pixel in total) of combinations of standard light sampling techniques (*Uniform area*, *Spherical*) and a new one (*Parallel*) motivated by properties of the optimal MIS weights. The combination with the new technique using the optimal weights performs best. The MSE improvement in parentheses is with respect to 40 samples from the *Spherical* technique alone. The false color insets show weights of the *Uniform area* and *Parallel* techniques. See the Supplemental material for full size images.

We first revisit the light selection problem for direct illumination computation from Sec. 8.2 and introduce a new technique that substantially lowers the variance. Then, we show new techniques that improve sampling of a single light source.

New technique for light selection. The *Trained* light selection technique from Sec. 8.2 neglects visibility. In shadows, the technique's pdf does not match the integrand well, and variance goes up.

We illustrate that in Fig. 7a. For the point A, the *Trained* technique (green) is a good fit to the integrand (gray), and performs well. For the point B, however, the actual integrand has no contribution from the closest light due to occlusion, and there is a mismatch between the pdf of the *Trained* technique and the integrand itself.

To solve the issue at the point B, we construct a new technique with a pdf that matches the integrand well specifically for that case. Then we leave it up to the optimal weights for a particular image pixel to decide which of the two cases has occurred (A or B), and to form the optimal control variate from pdfs of both techniques. It is easy to construct such a technique from the pdf of the *Trained* technique: it is the same except it samples the strongest light with a zero probability. We call this technique *NoMax* (orange in Fig. 7a).

We demonstrate that in the *Staircase II* scene (Fig. 6). All images with optimal weights were rendered by the Direct estimator (Sec. 7) using the same number of samples per technique per pixel (20 samples). We see that using the *NoMax* technique alone causes a significant bias. But when optimally weighted with the *Trained* technique, it is is much better than any other result in Fig. 6. Note that the power heuristic is unable to create such a combination: It improves in shadows, but increases variance in the rest of the scene in comparison to *Trained* as well as to the power heuristic combination of *Trained* and *Uniform*. That gives the optimal weights 9.9× lower MSE per sample. Moreover, the optimal combination of the *Trained* and *NoMax* techniques improves 3.7× over the optimal combination of *Trained* and *Uniform*.

One special case, when the combination of the *Trained* technique and the *Uniform* technique works particularly well is when we have *exactly* two lights in the scene. We illustrate that on *Staircase I* scene in Fig. 1. In that case a linear combination of the *Trained* and *Uniform* techniques can approximate virtually any distribution, which results in 9.6× lower MSE per sample than the power heuristic.

The Supplemental provides complete results including the balance heuristic, which performs similarly to the power heuristic.

New techniques for light area sampling. While light selection contributes most direct illumination variance in scenes with many small lights, careful sampling of the point on the light source becomes important in the presence of larger light sources. Fig. 7b shows a schematic of a scene where a lambertian area light source illuminates a point on a diffuse surface. The figure plots the sampling densities of various techniques over the part of the hemisphere that receives illumination, as well as the integrand itself (in black), which in this case becomes $L_\mathrm{e}G$, where L_e is the emitted radiance and G the geometry term. A typical technique is the uniform sampling of the light surface, we denote it *Uniform area* (Fig. 7b, orange), but it is not a good approximation to the integrand as it neglects G. A better idea is to uniformly sample the light projection onto the unit sphere around the illuminated point [Arvo 1995], and we call this technique *Spherical* (Fig. 7b, green). But even better performs a linear combination of the *Uniform area* and *Spherical* techniques (shown in red), found by the optimal weights. That is, as long as the light is parallel to the illuminated surface.

If the light is *not* parallel, the shape of the *Uniform area* technique deforms (see the point B in Fig. 7b) and even the combination found by the optimal weights is worse than *Spherical* alone. We now replace the *Uniform area* technique with a new one: uniform sampling of the light projection onto *a plane parallel to the surface*, denoted *Parallel* (Fig. 7b, blue). Its pdf is similar to that of *Uniform area*, but does not depend on the light orientation.

We demonstrate these techniques in the *Dining room* scene (Fig. 8) lit by one large area light from above. All images were rendered by the Direct estimator (Sec. 7) using the same *total* number of samples per pixel (40 samples). As expected, the *Spherical* technique alone generally performs better than the other two. Therefore, their combination using the power heuristic will always be worse than relying only on the samples from this technique. However, if combined using the optimal weights the result is much better. While the combination with *Uniform area* decreases variance mainly on the table, the combination with *Parallel* further improves the result also on surfaces not parallel to the light (e.g., the wall) and provides 2.1× lower MSE than the *Spherical* technique alone. Note the negative value of the optimal weights of the *Uniform area* and *Parallel* techniques in the improved regions.

Let us underline that the methods introduced in Sec. 8.3 are not meant to be ready for production use. They serve as a proof of concept showing that this approach to construction of sampling techniques has an interesting potential.

Table 1. Performance statistics of the Direct and Progressive estimators, the latter with different values of the update step U (Sec. 7.3). Speedup and equal-sample improvement are ratios of the mean-squared error. The overhead is the relative increase of the rendering time with the same total number of samples. The baseline for these values is the power heuristic combination, except for the *Dining room* which also compares to using the spherical projection sampling alone. See the Supplemental for the corresponding images.

	Staircase I Techniques: *Train* + *Uni* Baseline: Power *Train* + *Uni*			**Staircase II** Techniques: *Train* + *M* Baseline: Power *Train* + *Uni* / Power *Train* + *M*			**Dining room** Techniques: *Par* + *Sp* Baseline: *Sp* / Power *Par* + *Sp*			**Veach** Techniques: *BSDF* + *Light* Baseline: Power *BSDF* + *Light*		
	Equal-time speedup	Equal-sample improvement	Overhead	Equal-time speedup	Equal-sample improvement	Overhead	Equal-time speedup	Equal-sample improvement	Overhead	Equal-time speedup	Equal-sample improvement	Overhead
Direct	8.89	9.56	6.20%	8.86 / 7.53	9.90 / 7.83	9.93% / 2.54%	3.40 / 9.99	2.12 / 10.05	-30.53% / 5.94%	1.02	1.02	5.02%
Progressive $U = 1$	3.01	4.37	33.02%	5.25 / 4.46	6.68 / 5.29	35.32% / 26.23%	1.87 / 5.48	1.27 / 6.00	-12.17% / 33.92%	0.77	1.03	38.24%
Progressive $U = 2$	2.76	3.42	19.32%	4.81 / 4.09	5.35 / 4.23	24.07% / 15.73%	1.87 / 4.92	1.03 / 4.88	-20.43% / 21.33%	0.86	1.04	20.88%
Progressive $U = 4$	2.03	2.33	12.44%	3.82 / 3.25	3.90 / 3.09	17.64% / 9.73%	1.50 / 4.40	0.74 / 3.50	-26.09% / 12.70%	0.94	1.03	14.71%

Legend: *Train* = *Trained*, *Uni* = *Uniform*, *M* = *NoMax*, *Par* = *Parallel*, and *Sp* = *Spherical*

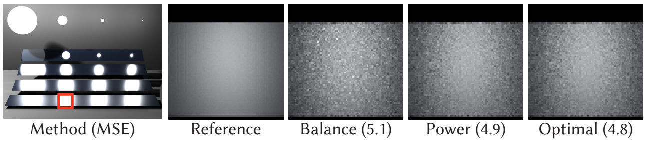


Fig. 9. Equal-sample comparison of the optimal MIS weights with the balance and power heuristics in the classic light vs. BSDF sampling scenario in the Veach's scene. The MSE values (in parantheses) are computed after 10 samples per light per technique per pixel. See Sec. 8.4 for details and the Supplemental material for full size images.

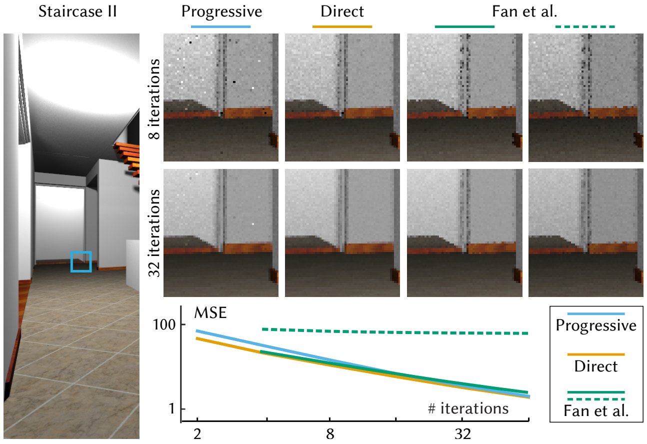


Fig. 10. The insets along with MSE plots for the *Staircase II* scene rendered with an increasing number of samples with the Direct and Progressive estimators and the method of Fan et al. with either the *Uniform* (solid) or the *Trained* (dashed) technique skipped. See Sec. 8.4 for details.

8.4 Additional results

Optimal weights for BSDF and light techniques. We investigated the behavior of the optimal weights for an MIS combination of the light area and BSDF sampling techniques. For that we rendered the classic Veach's scene [Veach and Guibas 1995]. Following Veach, we estimate illumination from individual lights separately, combining light area and BSDF sampling, and we add the contributions together. We combine the samples using the optimal weights and compare the result with the balance and power heuristics in Fig. 9. In this setting, the power heuristic appears to be close to the optimum, but the optimal weights still slightly improve the result.

Overhead. We have so far focused on equal-sample comparisons to clearly show the effect of the combination strategies unaffected by the implementation. For the sake of completeness, equal-time comparisons are provided in the Supplemental material and summarized in Table 1. The overhead of the Direct estimator (caused mainly by the $\langle \mathbf{A} \rangle$, $\langle \mathbf{b} \rangle$ updates) is at most 10%, making the equal-sample MSE improvement close to the equal-time speedup. Note that when comparing to the *Spherical* technique in the *Dining room* scene the overhead is negative; sampling the perfect spherical projection is considerably more expensive than the other techniques.

Regarding memory overhead, we need to store estimates for the technique matrix for each pixel and estimates for the contribution vector for each pixel and color channel, which in our cases meant storing $2^2 + 3 \cdot 2 = 10$ floats per pixel. When rendering the image by blocks, one pixel in a block at a time, the memory overhead is practically negligible.

Direct vs. Progressive estimators. All our results shown in Sec. 8.2 and Sec. 8.3 were obtained by the Direct estimator. Its bias and variance with respect to the Progressive estimator could be a concern. We provide both their equal-time and equal-sample comparisons in the Supplemental material with a summary in Table 1. In agreement with our synthetic tests from Sec. 7, the equal-sample MSE improvement of the Progressive estimator is always smaller (about 30%-40%), except for the Veach's scene, where both estimators perform equally. In Fig. 10, we show insets and MSE plots of the renderings using an increasing number of samples per technique (from 2 to 64) in the *Staircase II* scene. The Progressive estimator (blue) is unbiased but gains a spiky noise in the initial iterations, from which it takes long to recover. The Direct estimator (yellow) is biased only for a low number of samples (<16) and practically zero afterward, which is also in line with our synthetic tests.

As expected, the overhead of the Progressive estimator is higher than the Direct one because of the repeated solving of the linear system. As the update step U increases (Sec. 7.3), the overhead decreases from almost 40% for $U = 1$ to 15% for $U = 4$. But since the equal-sample MSE improvement also decreases, the equal-time speedup is actually worse as well. The best compromise seems to be using $U = 2$, yielding up to 5× speedup in our scenes.

Comparison to Fan et al. In Fig. 10 we compare our approach to Fan et al. [2006], who adopted the approach by Owen and Zhou for rendering. They estimate $\boldsymbol{\alpha}$ by solving a truncated system obtained by skipping regressors corresponding to a particular sampling technique from the data matrix. For a particular skipped technique their method is the same as our biased Direct estimator, except for two differences: First, they do not perform the estimation per pixel but by averaging per point estimates computed from fixed-sized batches, which makes their method not consistent. Second, they introduce a regularization strategy which can decrease variance at the cost of increased bias. For clarity, we provide pseudocode of our adaptation of their method in Sec. 4 in the Supplemental.

We set the batch size in their method to 8 samples (the same total number of samples as 4 iterations of our method) and rendered the *Staircase II* scene with an increasing number of batches. The green lines in the plot show their method when skipping the *Uniform* (solid) and *Trained* (dashed) technique, respectively. When the *Uniform* technique is skipped, their method behaves similarly to ours, and their regularization slightly reduces the noise in some parts of the image. When the *Trained* technique is skipped, the substantial bias of their method due to the computation in batches is further amplified by their regularization approach, resulting in a visibly darker image. As the performance of their method depends on a skipped technique, it might be difficult to predict the optimal technique for skipping for a given integration problem. Without the regularization, their method produces identical results to our Direct estimator for any technique skipped, but only for the first batch (with increasing number of batches the bias in their method does not diminish). A more in-depth discussion is provided in the Supplemental.

9 LIMITATIONS AND FUTURE WORK

Applications. While we believe that a derivation of optimal MIS weights is an important theoretical result, their application in practice is more complicated than for the traditional balance or power heuristics. Estimation and solution of the linear system results in computational overhead that grows super-linearly with the number of combined techniques. While the overhead in our tests was almost negligible, especially for the Direct estimator, this could become an issue as the number of sampling techniques increases.

Our rendering applications provide a proof of concept, but are far from being production-ready and leave space for further investigation. An obvious next step would be to integrate the optimal weights into a full global illumination solution. One interesting direction is the optimal combination of sampling techniques in bidirectional path tracing and derived methods [Georgiev et al. 2012a; Hachisuka et al. 2012; Veach and Guibas 1995], though handling the relatively high number of available sampling techniques could be challenging. Another class of algorithms that could greatly benefit from the optimal MIS weights is path guiding [Herholz et al. 2016; Müller et al. 2017; Vorba et al. 2014], where the necessity for defensive sampling limits the achievable improvements.

The MIS framework. A serious limitation of the MIS framework itself is its somewhat wasteful approach: samples are first taken but the contribution of many of them may be weighted almost to zero. Our optimal weights do not address this issue. More work is needed on optimizing the sample counts for different techniques (and whether or not some techniques should be included in the mix at all), while maintaining the estimator's robustness. Furthermore, we have shown that the optimal weights motivate the design of new sampling techniques, and this is is another direction that may benefit from further investigation.

10 CONCLUSION

We have presented optimal weighting functions for the multi-sample model of multiple importance sampling. In deriving the optimal weights, we have pointed out, for the first time, an unnecessary assumption on the non-negativity of weighting functions underpinning the previous claims concerning variance bounds for the balance heuristic. We have shown that this assumption effectively prohibited exploration of an entire class of efficient combination strategies, among which the optimal one.

We have shown the connection of the optimal weights to control variates, yielding interesting observations on the relation of variance of the optimal weights and balance heuristic. In particular, the optimal weights are a good choice for defensive sampling, where the balance heuristic is particularly inefficient. Our proof of concept applications in direct illumination estimation have shown that new sampling strategies motivated by the variance properties of the optimal weights yield further benefits. We believe that our work opens up new directions for improving efficiency of combined estimators.

ACKNOWLEDGMENTS

Many thanks to Benedikt Bitterli [2016] for the PBRT test scenes. The work was supported by the Charles University Grant Agency project GAUK 996218, by the grant SVV-2017-260452, and by the Czech Science Foundation grants 16-18964S and 19-07626S. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642841 (DISTRO).

REFERENCES

James Arvo. 1995. Stratified Sampling of Spherical Triangles. In *Proc. SIGGRAPH 1995*. ACM, New York, NY, USA, 437–438.

Gilles Aubert and Pierre Kornprobst. 2006. *Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations* (2nd ed.). Springer.

Benedikt Bitterli. 2016. Rendering resources. https://benedikt-bitterli.me/resources/.

Olivier Cappé, Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P. Robert. 2008. Adaptive importance sampling in general mixture classes. *Statistics and Computing* 18, 4 (01 Dec 2008), 447–459.

Víctor Elvira, Luca Martino, David Luengo, and Mónica F. Bugallo. 2015. Generalized multiple importance sampling. arXiv:1511.03095.

Víctor Elvira, Luca Martino, David Luengo, and Mónica F. Bugallo. 2016. Heretical multiple importance sampling. *IEEE Signal Processing Letters* 23, 10 (Oct 2016).

Shaohua Fan, Stephen Chenney, Bo Hu, Kam Wah Tsui, and Yu Chi Lai. 2006. Optimizing control variate estimators for rendering. *Comput. Graph. Forum (EUROGRAPHICS 2006)* 25, 3 (2006), 351–357.

Iliyan Georgiev, Jaroslav Křivánek, Tomáš Davidovič, and Philipp Slusallek. 2012a. Light Transport Simulation with Vertex Connection and Merging. *ACM Trans. Graph. (SIGGRAPH Asia 2012)* 31, 6, Article 192 (Nov. 2012), 10 pages.

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.

Paul Glasserman. 2003. *Monte Carlo method in financial engineering*. Springer-Verlag, New York, USA.

Adrien Gruson, Mickaël Ribardière, Martin Šik, Jiří Vorba, Rémi Cozot, Kadi Bouatouch, and Jaroslav Křivánek. 2016. A Spatial Target Function for Metropolis Photon Tracing. *ACM Trans. Graph.* 36, 4, Article 75a (Nov. 2016).

Toshiya Hachisuka, Anton S. Kaplanyan, and Carsten Dachsbacher. 2014. Multiplexed Metropolis Light Transport. *ACM Trans. Graph. (SIGGRAPH 2014)* 33, 4 (2014).

Toshiya Hachisuka, Jacopo Pantaleoni, and Henrik Wann Jensen. 2012. A Path Space Extension for Robust Light Transport Simulation. *ACM Trans. Graph. (SIGGRAPH Asia 2012)* 31, 6, Article 191 (Nov. 2012), 10 pages.

Vlastimil Havran and Mateu Sbert. 2014. Optimal Combination of Techniques in Multiple Importance Sampling. In *Proc. VRCAI '14*. ACM, New York, NY, 141–150.

Hera Y. He and Art B. Owen. 2014. Optimal mixture weights in multiple importance sampling. (2014), 1–22. arXiv:1411.3954

Sebastian Herholz, Oskar Elek, Jiří Vorba, Hendrik Lensch, and Jaroslav Křivánek. 2016. Product Importance Sampling for Light Transport Path Guiding. *Comput. Graph. Forum (EGSR 2016)* 35, 4 (2016), 67–77.

Malvin H. Kalos and Paula A. Whitlock. 2008. *Monte Carlo Methods* (2nd ed.). Wiley-VCH.

Csaba Kelemen, László Szirmay-Kalos, György Antal, and Ferenc Csonka. 2002. A Simple and Robust Mutation Strategy for the Metropolis Light Transport Algorithm. *Computer Graphics Forum* 21, 3 (2002), 531–540.

A. Keller, L. Fascione, M. Fajardo, I. Georgiev, P. Christensen, J. Hanika, C. Eisenacher, and G. Nichols. 2015. The Path Tracing Revolution in the Movie Industry. In *ACM SIGGRAPH 2015 Courses.* Article 24.

Jaroslav Křivánek, Iliyan Georgiev, Toshiya Hachisuka, Petr Vévoda, Martin Šik, Derek Nowrouzezahrai, and Wojciech Jarosz. 2014. Unifying Points, Beams, and Paths in Volumetric Light Transport Simulation. *ACM Trans. Graph. (SIGGRAPH 2014)* 33, 4, Article 103 (July 2014), 13 pages.

Stephen S. Lavenberg, Thomas L. Moeller, and Peter D. Welch. 1982. Statistical Results on Control Variables with Application to Queueing Network Simulation. *Operations Research* 30, 1 (1982), 182–202.

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.

Thomas Müller, Markus Gross, and Jan Novák. 2017. Practical Path Guiding for Efficient Light-Transport Simulation. *Comput. Graph. Forum (EGSR 2017)* 36, 4 (2017), 91–100.

Art Owen and Yi Zhou. 2000. Safe and Effective Importance Sampling. *J. Amer. Statist. Assoc.* 95, 449 (2000), 135–143.

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.

Matt Pharr, Wenzel Jakob, and Greg Humphreys. 2016. *Physically Based Rendering: From Theory to Implementation* (3rd ed.). Morgan Kaufmann.

Stefan Popov, Ravi Ramamoorthi, Fredo Durand, and George Drettakis. 2015. Probabilistic Connections for Bidirectional Path Tracing. *Comput. Graph. Forum (EGSR 2015)* 34, 4 (July 2015), 75–86.

Reuven Y. Rubinstein and Ruth Marcus. 1985. Efficiency of Multivariate Control Variates in Monte Carlo Simulation. *Operations Research* 33, 3 (1985), 661–677.

Mateu Sbert and Vlastimil Havran. 2017. Adaptive Multiple Importance Sampling for General Functions. *Vis. Comput.* 33, 6-8 (June 2017), 845–855.

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.

Mateu Sbert, Vlastimil Havran, and Laszlo Szirmay-Kalos. 2018. Multiple importance sampling revisited: breaking the bounds. *EURASIP Journal on Advances in Signal Processing* 2018, 1 (27 Feb 2018), 15.

Eric Veach. 1997. *Robust Monte Carlo methods for light transport simulation.* Ph.D. Dissertation. Stanford University.

Eric Veach and Leonidas J. Guibas. 1995. Optimally Combining Sampling Techniques for Monte Carlo Rendering. *Proc. SIGGRAPH '95*, 419–428.

Sekhar Venkatraman and James R. Wilson. 1986. The efficiency of control variates in multiresponse simulation. *Operations Research Letters* 5, 1 (1986), 37–42.

Petr Vévoda, Ivo Kondapaneni, and Jaroslav Křivánek. 2018. Bayesian Online Regression for Adaptive Direct Illumination Sampling. *ACM Trans. Graph. (SIGGRAPH 2018)* 37, 4, Article 125 (July 2018), 12 pages.

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. (SIGGRAPH 2014)* 33, 4, Article 101 (July 2014), 11 pages.

Martin Šik, Hisanari Otsu, Toshiya Hachisuka, and Jaroslav Křivánek. 2016. Robust Light Transport Simulation via Metropolised Bidirectional Estimators. *ACM Trans. Graph. (SIGGRAPH Asia 2016)* 35, 6, Article 245 (Nov. 2016), 12 pages.

A CALCULUS OF VARIATIONS

Our derivation of the optimal weights relies on the *calculus of variations* [Aubert and Kornprobst 2006], the basic elements of which we now informally review. It is typically used to find extrema of a *functional* – a mapping from some space of functions Ω onto real numbers. In our case, the functional of interest – the variance – conforms to a general form $\mathbf{F}(h) = \int \hat{F}(h(x))\,\mathrm{d}x$, where $h \in \Omega$ is a function (in our case the weights) and $\hat{F}$ is some operation on h.

A basic tool used to locate extrema of a functional $\mathbf{F}$ is its *functional derivative* $\frac{\partial \mathbf{F}}{\partial h}$, i.e., the rate of change of $\mathbf{F}$ with infinitesimally small perturbations of the function h. Similar to classic calculus, the extrema are given by the function(s) h for which the functional derivative equals to zero i.e., $\partial\mathbf{F}/\partial h(x) = 0$.

Calculation of the functional derivative can be transformed to classic differentiation from 'ordinary' calculus using the relation

$$\left\langle \frac{\partial \mathbf{F}}{\partial h}, \delta \right\rangle = \frac{d}{d\varepsilon}\Big|_{\varepsilon=0} \mathbf{F}(h + \varepsilon\delta), \tag{25}$$

where $\delta \in \Omega$ is a variation (a function), while $\varepsilon \in \mathbb{R}$ is a number. To obtain the functional derivative, we 1) replace any occurrence of h in the functional by $h + \varepsilon\delta$, 2) take derivative wrt ε, 3) set $\varepsilon = 0$. This yields an expression that is, by the relation (25), equal to the inner product of the variation δ and the functional derivative $\partial\mathbf{F}/\partial h$ that we seek to find, i.e., to the integral $\int_D \frac{\partial \mathbf{F}}{\partial h}\,\delta\,\mathrm{d}x$. The last step is therefore to extract the part of the expression corresponding to the functional derivative.

As in classic calculus, Lagrange multipliers can be used to handle *constraints.* To find extrema of $\mathbf{F}$ satisfying a constraint $g(h(x)) = 0$, we formulate a constraint functional $\mathbf{G}(h) = \int \lambda(x) g(h(x))\,\mathrm{d}x$, where $\lambda \in \Omega$ is the Lagrange multiplier. We then locate extrema of the *Lagrangian* $\mathbf{L}(h, \lambda) = \mathbf{F}(h) - \mathbf{G}(h, \lambda)$ both in terms of h and λ.

B PROOF OF THEOREM 5.2: OPTIMAL WEIGHTS

We prove Theorem 5.2 by construction. To do that, we seek weighting functions w_i, $i = 1, \ldots, N$ that minimize the variance functional $V[w_1, \ldots, w_N]$, given by Eq. (7), constrained by $\sum_{i=1}^{N} w_i(x) = 1$ and $p_i(x) = 0 \Rightarrow w_i(x) = 0$. To simplify the derivation, we leave out the latter constraint, and verify it at the end. Dropping the function arguments, the solution is given by the minimum of the Lagrangian

$$\mathbf{L} = V[w_1, \ldots, w_N] - \int_D \lambda \left(\textstyle\sum_{i=1}^{N} w_i \ - 1 \right)\,\mathrm{d}x, \tag{26}$$

in terms of the weights w_i and the Lagrange multiplier $\lambda : D \longrightarrow \mathbb{R}$. To find the minimum, we set all the partial functional derivatives $\partial\mathbf{L}/\partial w_i$ and $\partial\mathbf{L}/\partial\lambda$ to zero. Using the relation (25), we find $\partial\mathbf{L}/\partial w_i$ as

$$\begin{aligned}
\frac{\partial}{\partial\varepsilon}\Big|_{\varepsilon=0} \mathbf{L}(\ldots, w_i + \varepsilon\delta_i, \ldots) = \Big|_{\varepsilon=0} \Bigg[\frac{2}{n_i} \int_D \frac{(w_i + \varepsilon\delta_i) f^2 \delta_i}{p_i}\,\mathrm{d}x - \\
\frac{2}{n_i} \int_D (w_i + \varepsilon\delta_i) f\,\mathrm{d}x \int_D \delta_i f\,\mathrm{d}x - \int_D \lambda\delta_i\,\mathrm{d}x \Bigg] \\
= \int_D \underbrace{\left(\frac{2 w_i f^2}{p_i n_i} - \frac{2f}{n_i} \int_D w_i f\,\mathrm{d}x - \lambda \right)}_{\partial\mathbf{L}/\partial w_i} \delta_i\,\mathrm{d}x.
\end{aligned} \tag{27}$$

We proceed in a similar way to find $\partial\mathbf{L}/\partial\lambda$. This gives us a set of equations for w_i and λ:

$$w_i - \frac{p_i}{f} \int_D w_i f\,\mathrm{d}x = \frac{n_i}{2} \lambda \frac{p_i}{f^2}, \qquad \textstyle\sum_{i=1}^{N} w_i = 1 \tag{28}$$

The equation on the left can be rewritten as

$$w_i = \alpha_i \frac{p_i}{f} + \frac{n_i}{2}\lambda \frac{p_i}{f^2}, \quad \text{with} \quad \alpha_i = \int_D w_i f \, \mathrm{d}x. \tag{29}$$

Plugging the above equation for w_i into the constraint $\sum_{i=1}^N w_i = 1$, (i.e., $\partial \mathrm{L}/\partial\lambda = 0$), we can solve for the multiplier λ:

$$\lambda = 2\, \frac{f^2 - f \sum_{i=1}^N \alpha_i p_i}{\sum_{i=1}^N n_i p_i}. \tag{30}$$

The final form of the optimal weights $w_i^o(x)$, given by Eq. (13), is now obtained by plugging (30) back into (29), left.

Our next step is to find the α_i, $i = 1, \dots, N$. Plugging the optimal weights (13) into (29), right, we obtain a set of equations for α_j

$$\int_D n_i p_i \frac{f - \sum_{j=1}^N \alpha_j p_j}{\sum_{k=1}^N n_k p_k} \, \mathrm{d}x = 0, \quad i = 1 \dots N, \tag{31}$$

which can be rearranged into

$$\sum_{j=1}^N \alpha_j \int_D \frac{p_i p_j}{\sum_{k=1}^N n_k p_k} \, \mathrm{d}x = \int_D \frac{p_i f}{\sum_{k=1}^N n_k p_k} \, \mathrm{d}x. \tag{32}$$

This can be written in a matrix form as $\mathbf{A}\boldsymbol{\alpha} = \mathbf{b}$, where $\mathbf{A}$ and $\mathbf{b}$ are the technique matrix and contribution vector from *Definition* 5.1 and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_N)^\mathsf{T}$.

From (29) we can see that whenever $p_i(x) = 0$, we get $w_i(x) = 0$, which validates out second constraint. This completes the proof.

C SOLUTION EXISTENCE AND UNIQUENESS

Here we discuss the existence and uniqueness of the optimal weights from Theorem 5.2, and show that there are infinitely many values of $\boldsymbol{\alpha}$ yielding the same optimal weights.

Existence and uniqueness. The optimal weights exist whenever the linear system (12) is consistent. To prove the consistency, we would need to show that if two rows i, j of $\mathbf{A}$ are the same, then also $b_i = b_j$, which we have not yet been able to do.

Nonetheless, it holds that whenever one sampling strategy is a convex combination of other strategies, i.e., $p_i = \sum_{j \neq i} c_j p_j$, then the i-th row of $\mathbf{A}$ becomes the same linear combination of the other rows, and $b_i = \sum_{j \neq i} c_j b_j$. In such cases the linear system becomes singular (but remains consistent) and there are infinitely many solutions for $\boldsymbol{\alpha}$, each yielding possibly *different* MIS weights, but producing an MIS estimator *with the same variance.* This is because $\boldsymbol{\alpha} \in \{\mathbf{u} + \mathbf{v} | \mathbf{A}\mathbf{u} = \mathbf{b} \wedge \mathbf{v} \in \mathrm{Null}(\mathbf{A})\}$, and (42) is the same for all such $\boldsymbol{\alpha}$. If the linear system is non-singular, the $\boldsymbol{\alpha}$ vector and the resulting weights are unique.

Full solution for $\boldsymbol{\alpha}$. Adding a term $s\mathbf{n}$, where $s \in \mathbb{R}$ and $\mathbf{n} = (n_1, \dots, n_N)^\mathsf{T}$, to $\boldsymbol{\alpha}$ produces the same weights, despite the modified vector $\boldsymbol{\alpha}$ not being a solution to the system (12). This is because the offset $s\mathbf{n}$ cancels out when the modified $\boldsymbol{\alpha}$ is plugged into the weights (13). Therefore all $\widetilde{\boldsymbol{\alpha}} = \mathbf{A}^{-1}\mathbf{b} + s\mathbf{n}$ yield the same optimal weights and we refer to $\widetilde{\boldsymbol{\alpha}}$ as to the *full solution* for $\boldsymbol{\alpha}$.

D PROOF OF THE RELATIONSHIP (18)

The variance of the optimal estimator (16) can be expanded as

$$\mathrm{V}[\langle F \rangle^o] = \mathrm{V}[\langle F \rangle^b] + \mathrm{V}[\langle G \rangle^b] - 2\mathrm{Cov}[\langle F \rangle^b, \langle G \rangle^b]. \tag{33}$$

We now express the variance $\mathrm{V}[\langle G \rangle^b]$ and covariance $\mathrm{Cov}[\langle F \rangle^b, \langle G \rangle^b]$ from (33) in terms of the technique matrix, contribution vector and $\boldsymbol{\alpha}$. Using the shorthand notation $q = (\sum_{i=1}^N n_i p_i)^{-1}$ and dropping the function arguments, we obtain

$$\mathrm{V}[\langle G \rangle^b] = \int_D q \left(\sum_{i=1}^N \alpha_i p_i \right)^2 \mathrm{d}x - \sum_{i=1}^N n_i \left(\int_D q \, p_i \sum_{j=1}^N \alpha_j p_j \, \mathrm{d}x \right)^2, \tag{34}$$

Because the elements of $\mathbf{A}$ are given by $a_{ik} = \langle p_i, \; p_k \, q \rangle$, we can rewrite the first term in (34) as:

$$\textstyle\sum_{i=1}^N \sum_{k=1}^N \alpha_i a_{ik} \alpha_k = \boldsymbol{\alpha}^\mathsf{T} \mathbf{A} \boldsymbol{\alpha}. \tag{35}$$

The second term in (34) can be transformed in a similar fashion:

$$\sum_{j=1}^N n_j \left(\textstyle\sum_{i=1}^N \alpha_i a_{ij} \right) \left(\textstyle\sum_{k=1}^N a_{jk} \alpha_k \right) = \boldsymbol{\alpha}^\mathsf{T} \mathbf{A}\mathbf{N}\mathbf{A} \boldsymbol{\alpha}, \tag{36}$$

with $\mathbf{N}$ being a diagonal $N \times N$ matrix with the sample count n_i along the diagonal. Putting together (35), (36), and factoring out $\boldsymbol{\alpha}$, we obtain

$$\mathrm{V}[\langle G \rangle^b] = \boldsymbol{\alpha}^\mathsf{T} \left(\mathbf{A} - \mathbf{A}\mathbf{N}\mathbf{A} \right) \boldsymbol{\alpha}. \tag{37}$$

Now, we express the covariance $\mathrm{Cov}[\langle F \rangle^b, \langle G \rangle^b]$. Denoting $\langle F \rangle_{ij}^b$ and $\langle G \rangle_{ij}^b$ the parts of the MIS estimators for i-th technique and j-th independent sample, the covariance becomes

$$\mathrm{Cov}[\langle F \rangle^b, \langle G \rangle^b] = \sum_{i=1}^N n_i \mathrm{Cov}[\langle F \rangle_{i1}^b, \langle G \rangle_{i1}^b]. \tag{38}$$

That is because $\langle F \rangle_{ij}^b$ and $\langle G \rangle_{kl}^b$ are independent whenever $i \neq k$ and $j \neq l$, and thus their covariance is zero. Again, using $q = (\sum_{i=1}^N n_i p_i)^{-1}$, the relation (38) can be further expanded

$$\begin{aligned} \sum_{i=1}^N n_i \mathrm{Cov}[\langle F \rangle_{i1}^b, \langle G \rangle_{i1}^b] = \int_D q \, f \textstyle\sum_{i=1}^N \alpha_i p_i \, \mathrm{d}x \; - \\ - \sum_{i=1}^N n_i \left(\int_D q \, p_i f \, \mathrm{d}x \right) \left(\int_D q \, p_i \textstyle\sum_{j=1}^N \alpha_j p_j \, \mathrm{d}x \right). \end{aligned} \tag{39}$$

The first term in (39) equals to $\mathbf{b}^\mathsf{T}\boldsymbol{\alpha}$ where $\mathbf{b}$ is the contribution vector. The second term could be expanded as:

$$\sum_{i=1}^N n_i b_i \left(\textstyle\sum_{k=1}^N a_{ik} \alpha_k \right) = \mathbf{b}^\mathsf{T} \mathbf{N} \mathbf{A} \boldsymbol{\alpha}. \tag{40}$$

Subtracting (40) from $\mathbf{b}^\mathsf{T}\boldsymbol{\alpha}$ yields the desired relation for the covariance:

$$\mathrm{Cov}[\langle F \rangle^b, \langle G \rangle^b] = \mathbf{b}^\mathsf{T} (\mathbf{I} - \mathbf{N}\mathbf{A}) \boldsymbol{\alpha}. \tag{41}$$

Finally, expanding (33) using the relationships (37) and (41), we obtain:

$$\mathrm{V}[\langle R \rangle^b] = \mathrm{V}[\langle F \rangle^b] + \boldsymbol{\alpha}^\mathsf{T} \left(\mathbf{A} - \mathbf{A}\mathbf{N}\mathbf{A} \right) \boldsymbol{\alpha} - 2\mathbf{b}^\mathsf{T} (\mathbf{I} - \mathbf{N}\mathbf{A}) \boldsymbol{\alpha}. \tag{42}$$

By using $\mathbf{b}^\mathsf{T} = \boldsymbol{\alpha}^\mathsf{T}\mathbf{A}$ and simplifying, we obtain the desired relationship (18).