Vector-Valued Monte Carlo Integration Using Ratio Control Variates: Supplementary Material

HAOLIN LU, University of California San Diego, USA and Max Planck Institute for Informatics, Germany
DELIO VICINI, Google, Switzerland
WESLEY CHANG, University of California San Diego, USA
TZU-MAO LI, University of California San Diego, USA

1 OVERVIEW

In Section 2, we discuss the properties of *vector-valued importance sampling* and *entry-wise stratification* estimators. These properties reveal their inherent challenges in handling vector-valued integration, which motivates the development of ratio control variates. Section 2.3 introduces two additional variants of importance sampling for vector-valued problems, which are proved to be inferior to *vector-valued importance sampling*.

In Section 3, we briefly review difference control variates with multiple auxiliaries and how to get the optimal regression constants. This provides a foundation for the subsequent discussion on multi-auxiliary RCV.

In Section 4.1, we show how to approximate the mean squared error (MSE) and bias for the basic ratio estimator. In Section 4.2, we introduce several bias-reduced and debiased variants of ratio estimators and analyze their MSE and bias. In Section 4.3, we discuss the pitfalls of using one-sample RCV at each path vertex, which leads to an inconsistent estimator. This motivates the use of path-space RCV. In Section 4.4, we present a method to estimate the confidence in the accuracy of the auxiliary and improve RCV performance. In Section 4.5, We demonstrate three strategies for integrating multiple auxiliaries into a single RCV estimator.

For clarity, all lengthy proofs are deferred to the Appendix A.

2 IMPORTANCE SAMPLING

2.1 Vector-valued importance sampling

We formally define the vector-valued importance sampling estimator as follows:

Estimator 1. Vector-valued importance sampling. We draw N samples X_i, $i = 1, \ldots, N$ from a distribution $g(x)$. For each sample X_i, we always compute all function values $f_j(X_i)$, for every $j = 1, \ldots, M$, contributing to their respective estimations:

$$\hat{\mathbf{F}}^{IS} = \left\{ \hat{f}_j^{\,IS} = \frac{1}{N} \sum_{i=1}^{N} \frac{f_j(X_i)}{g(X_i)} \right., \text{ with } j = 1, \ldots, M. \tag{1}$$

Lemma 1. The MSE of the vector-valued importance sampling estimator (Estimator 1) is minimized when the PDF g(x) is proportional to the L2-norm of integrand $\mathbf{f}$: $g^*(x) \propto \sqrt{\sum_{j=1}^{M} f_j^2(x)}$. (proof: Appendix A.1)

However, even if we use optimal g^*, Estimator 1 generally can not achieve zero-variance:

Authors' addresses: Haolin Lu, University of California San Diego, USA and Max Planck Institute for Informatics, Germany; Delio Vicini, vicini@google.com, Google, Switzerland; Wesley Chang, University of California San Diego, USA; Tzu-Mao Li, University of California San Diego, USA.

Lemma 2. The vector-valued importance sampling estimator (Estimator 1) generally cannot achieve zero-variance, even if we use optimal $g^*(x)$ and all $\mathbf{M}$ integrands are non-negative. (proof: Appendix A.2)

Hereby we realize the limitation of the vector-valued importance sampling estimator: despite reusing all samples across integrands, the estimator loses the potential to use different importance sampling distributions for different integrand f_j. This motivates us to investigate the entry-wise stratification estimator.

2.2 Entry-wise stratification

Unlike in the main paper, we define entry-wise stratification here in a more detailed and general way:

Estimator 2. Entry-Wise Stratification. We partition M functions into K strata, $\mathcal{S}_k$, where $k = 1, \cdots, K$, which are mutually exclusive and collectively exhaustive. Then, we use K different importance sampling distributions, g_k, to estimate the integrals in each strata independently, with a deterministic number of samples N_k:

$$\hat{\mathbf{F}}^{EWS} = \left\{ \hat{f}_j^{\,EWS} = \frac{1}{N_{k(j)}} \sum_{j}^{N_{k(j)}} \frac{f_j(X_i)}{g_{k(j)}(X_i)} \right. , \text{ with } j = 1, \cdots, M. \tag{2}$$

where $k(j)$denotes the strata the function f_k belongs to. We restrict $\sum_k N_k = N$ to maintain an equal sample budget.

Lemma 3. The entry-wise stratification estimator (Estimator 2) may achieve lower variance than the vector-valued importance sampling estimator (Estimator 1), but none of them dominates (proved in Appendix A.3).

Therefore, whether entry-wise stratification or vector-valued importance sampling is better is really case by case. It could be challenging to rigorously determine which one is better in practice.

2.3 Other variants of importance sampling

We also enumerate two more variants of importance sampling, which appear to fall between Estimator 1 and Estimator 2. However, they are later shown to be always equal to or worse than Estimator 1 in optimal scenarios.

Estimator 3. Random Mixture Importance Sampling. To draw each sample, we first choose one out of M function to evaluate, with index k_i, from a discrete probability $p(k)$. Then, we further draw the actual sample from a conditional PDF $g(x|k)$, and will have:

$$\hat{\mathbf{F}}^{IS}_{Scalar} = \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{f_j(X_i)\mathbf{1}_{k=j}}{g(X_i|k_i)p(k_i)} \right. , \text{ with } j = 1, \cdots, M, \tag{3}$$

where $\mathbf{1}_{k=j}$ is the indicator function that equals 1 if $k = j$ and 0 otherwise. The estimator essentially uses different importance functions $g(x|j)$ to evaluate each function f_j independently. and uses a random probability $p(k)$ to allocate the samples for each estimation. It is analogous to the random mixture in the MIS literature.

Lemma 4. The variance of the random mixture importance sampling estimator (Estimator 3) is minimized when $g(x|k)p(k) \propto |f_k(x)|$ (proved in Appendix A.4).

Lemma 5. When using optimal distributions for both vector-valued importance sampling (Estimator 1) and random-mixture importance sampling (Estimator 3), the vector-valued importance sampling estimator always provides lower or equal variance than the random-mixture version, under equal samples comparison. (proved in Appendix A.5)

Estimator 4. Partial Vector-Valued Importance Sampling. Alternatively, we start with vector-valued importance sampling. To adapt more for each integrand $f_j(x)$, we use a different acceptance rate $p(x|j)$, and probabilistically discard the sample using p. Therefore, each integrand effectively receives a different importance sampling distribution $p(x|j)g(x)$. It gives an unbiased estimator:

$$\hat{\mathbf{F}}^{IS}_{Partial} = \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{f_j(X_i)}{g(X_i)p(X_i|j)} \right. , \text{ with } j = 1, \cdots, M. \tag{4}$$

Lemma 6. When using arbitrary but same distributions $g(x)$ for both vector-valued importance sampling (Estimator 1) and the partial vector-valued version (Estimator 4), the vector-valued importance sampling estimator always provides lower or equal variance, under equal samples comparison (proved in Appendix A.6).

Therefore, these two variants are less attractive than those mentioned in our main paper.

3 DIFFERENCE CONTROL VARIATES

3.1 Multiple regression control variates

To prepare for the subsequent discussions on ratio control variates with multiple auxiliaries, we briefly recap the *multiple regression control variates estimator* (Chapter 8.9, [Owen 2013]):

$$\hat{f} = \bar{f} + \sum_{k=1}^{K} \beta_k (H_k - \overline{h_k}), \tag{5}$$

where $h_k(x)$, $k = 1, \cdots, K$ are auxiliary functions with known integral values $H_k = \int h_k(x)\mathbf{d}x$, and β_k are constant weights to rescale each auxiliary function h_k.

Lemma 7. The optimal β can be estimated by:

$$\beta^* = \mathrm{Var}(h(x))^{-1}\mathrm{Cov}(h(x), f(x)), \tag{6}$$

(see Owen's book or Appendix A.7)

4 RATIO ESTIMATORS

4.1 Approximate MSE and Bias Ratio Estimator

4.1.1 Approximate MSE of Ratio Estimator. We can approximate the MSE of a basic ratio estimator $\bar{f}/\bar{h}$:

Lemma 8. The MSE of the Ratio Estimator can be approximated as

$$\mathrm{MSE}\left[\hat{F}^{Ratio}\right] \approx \frac{1}{N}\int \frac{(f(x) - F \cdot h(x))^2}{g(x)}\mathbf{d}x. \tag{7}$$

This is proved in Appendix A.8. The derivation is similar to that of the MSE for WIS.

4.1.2 Exact Bias of Ratio Estimator. Hartley and Ross [Hartley and Ross 1954] gave an exact expression of the bias for the basic ratio estimator, using the property of covariance:

$$\mathrm{Cov}(\frac{\bar{f}}{\bar{h}}, \bar{h}) = \mathbb{E}\left[\bar{f}\right] - \mathbb{E}\left[\frac{\bar{f}}{\bar{h}}\right]\mathbb{E}\left[\bar{h}\right] \tag{8}$$

$$\mathbb{E}\left[\frac{\bar{f}}{\bar{h}}\right] = \frac{\mathbb{E}\left[\bar{f}\right]}{\mathbb{E}\left[\bar{h}\right]} - \frac{\mathrm{Cov}(\frac{\bar{f}}{\bar{h}}, \bar{h})}{\mathbb{E}\left[\bar{h}\right]}, \text{ therefore:} \tag{9}$$

Lemma 9. The exact bias of the basic ratio estimator is given by

$$\mathrm{Bias}\left[\frac{\bar{f}}{\bar{h}}\right] = \mathbb{E}\left[\frac{\bar{f}}{\bar{h}}\right] - \frac{\mathbb{E}\left[\bar{f}\right]}{\mathbb{E}\left[\bar{h}\right]} = -\frac{\mathrm{Cov}(\frac{\bar{f}}{\bar{h}}, \bar{h})}{H}. \tag{10}$$

Notice, although Eq. (11) is an exact expression, the covariance term involves $\frac{\bar{f}}{\bar{h}}$, which makes the term intractable in practice. Therefore, directly debiasing the ratio of mean estimator is challenging.

4.1.3 Approximate Bias of Ratio Estimator with Taylor Expansion. Sukhatme et al. [Sukhatme et al. 1984] approximate the bias by applying Taylor expansion and retaining up to $O(\frac{1}{N})$ terms. Similarly,

Lemma 10. The approximation bias of the basic ratio control variates estimator is given by

$$Bias\left[\frac{\bar{f}}{\bar{h}} \cdot H\right] \approx \frac{F}{N}\left[\frac{\mathrm{Var}\left[h(x)\right]}{H^2} - \frac{\mathrm{Cov}\left[h(x), f(x)\right]}{FH}\right]. \tag{11}$$

We have rederived the conclusion in the context of Monte Carlo integration in Appendix A.9.

4.1.4 Numerical Validation of MSE and Bias Approximation. In Fig. 2, we show the MSE and bias evaluated with numerical simulation and the approximation derived in Section 4.1.1 and Section 4.1.3.

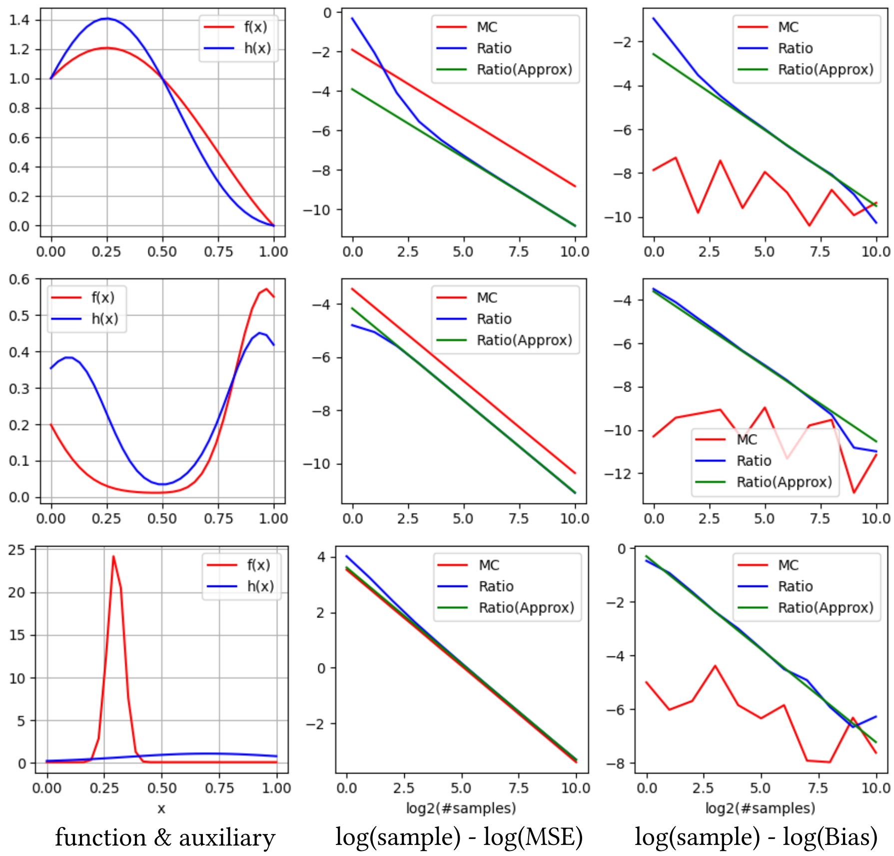


Fig. 1. We compare the numerical MSE and bias estimation with the estimated expressions derived in Section 4.1.1 and Section 4.1.3. We can observe that the approximation fits the simulation well in large sample counts. Our approximation works well especially for large enough sample counts. As we numerically evaluate the reference and all expectations, the bias estimation becomes inaccurate when it is very close to 0, leading to the waving curves, but this doesn't affect our conclusion.

4.2 Bias-reduced and debiased ratio estimators.

4.2.1 Bias and Variance of Almost-Unbiased Ratio Estimators. By approximate bias analysis as we do in Section 4.1.3, people can design estimators that remove the low-order terms of bias, namely *almost-unbiased ratio estimators*. We list some of these estimators' approximate bias and MSE in Table 1. We adapt the notation and results from Tin's work [Tin 1965] for the table, where:

$$\bar{f} = \frac{1}{N}\sum_{i=1}^{N} f(X_i); \quad \bar{h} = \frac{1}{N}\sum_{i=1}^{N} h(X_i); \quad \bar{r} = \frac{1}{N}\sum_{i=1}^{N} \frac{f(X_i)}{h(X_i)}; \quad C_{ij} = \frac{K_{ij}}{H^i F^j}$$

$$S_{fh} = \int (f(x) - F)(h(x) - H)\mathbf{d}x; \quad S_{hh} = \int (h(x) - H)(h(x) - H)\mathbf{d}x$$

In C_{ij}, the K_{ij} denotes the (i, j)th cumulant of h and f. As here all i and j are less than 4, it is similar to their moments, i.e.

$$K_{ij} = \mathbb{E}\left[(h(x) - H)^i (f(x) - F)^j\right] \tag{12}$$

Table 1. We list the vanilla ratio estimator (Ratio of Mean), Quenouille's estimator [Quenouille 1956], Beale's estimator [EML 1956], Tin's estimator [Tin 1965], Mean of Ratio estimator, Hartley Ross's Estimator [Hartley and Ross 1954], with their (approximated) mean squared error and bias. (*: For the average of ratio, both MSE and Bias are exact expressions instead of approximations.)

Estimator	Estimator Formula	MSE (w/ Approximation)	Bias (w/ Approximation)
Vanilla	$\hat{R} = \frac{\bar{f}}{\bar{h}}$	$\frac{1}{N}\frac{F^2}{H^2}(C_{11} + C_{00} - 2C_{01}) + O\left(\frac{1}{N^2}\right)$	$\frac{1}{N}\frac{F}{H}(C_{20} - C_{11})$ $+\frac{1}{N^2}\frac{F}{H}(C_{21} - C_{30} + 3C_{20}(C_{20} - C_{11}))$
Quenouille's	$\hat{R}_1 = \left[2 \cdot \frac{\bar{f}}{\bar{h}} - \frac{1}{2}(\frac{\bar{f}_1}{\bar{h}_1} + \frac{\bar{f}_2}{\bar{h}_2})\right]$	$\frac{1}{N}\frac{F^2}{H^2}(C_{11} + C_{00} - 2C_{01}) + O\left(\frac{1}{N^2}\right)$	$\frac{1}{N^2}\frac{F}{H}[6(C_{20}(C_{20} - C_{11})) - 2(C_{21} - C_{30})]$
Beale's	$\hat{R}_2 = \frac{\bar{f}}{\bar{h}} \cdot \frac{1+S_{fh}/(N \cdot \bar{f}\bar{h})}{1+S_{hh}/(N \cdot \bar{h}\bar{h})}$	$\frac{1}{N}\frac{F^2}{H^2}(C_{11} + C_{00} - 2C_{01}) + O\left(\frac{1}{N^2}\right)$	$\frac{1}{N^2}\frac{F}{H}[-2(C_{21} - C_{30}) - 2(C_{20}(C_{20} - C_{11}))]$
Tin's	$\hat{R}_3 = \frac{\bar{f}}{\bar{h}} \cdot \left[1 + \frac{1}{N}\left(\frac{S_{fh}}{\bar{f}\bar{h}} - \frac{S_{hh}}{\bar{h}\bar{h}}\right)\right]$	$\frac{1}{N}\frac{F^2}{H^2}(C_{11} + C_{00} - 2C_{01}) + O\left(\frac{1}{N^2}\right)$	$\frac{1}{N^2}\frac{F}{H}[-2(C_{21} - C_{30}) - 3(C_{20}(C_{20} - C_{11}))]$
Average of Ratio*	$\hat{R}_4 = \bar{r}$	$\frac{1}{N}\mathrm{Var}\left[\frac{f}{h}\right] + \left(\mathrm{Cov}\left(\frac{f}{h}, h\right)\right)^2$	$\mathrm{Cov}\left(\frac{f}{h}, h\right)$
Hartley Ross's	$\hat{R}_5 = \bar{r} + \frac{N}{N-1}\frac{\bar{f} - \bar{r}\bar{h}}{H}$	$\frac{1}{N}\frac{F^2}{H^2}(C_{11} + C_{00} - 2C_{01}) + O\left(\frac{1}{N^2}\right)$	0

4.2.2 Unbiasedness of Hartley Ross's Estimator. As discussed in the main paper, Hartley Ross's estimator is a fully debiased variant of the ratio estimator. The proof of its unbiasedness is included in Appendix A.10.

4.2.3 1D Example Considering Bias and MSE. In Fig. 2, we depicted the mean squared error (MSE) and bias for each estimator listed in Table 1. We numerically measure reference integral and all expectations, leading to an unstable bias estimation near convergence.

Generally, under high sample counts, all techniques have almost similar convergence speeds and usually close MSE values, similar to our MSE analysis in Table 1. Under lower sample counts, different estimators may behave quite differently, but there is no general rule to say which is always better. As we can observe, Hartley Rose usually has a higher MSE when the auxiliary is well correlated, but when bias is so huge that it dominates the MSE, HR may also be the best. In terms of bias, the vanilla ratio estimator is almost always the most biased. However, other estimators do not necessarily lead to lower MSE despite having a lower bias.

Specifically, we can observe that in the second example, the Hartley Ross's Estimator has significantly higher MSE than other techniques. This violates our MSE analysis in Table 1, which indicates with high sample counts,

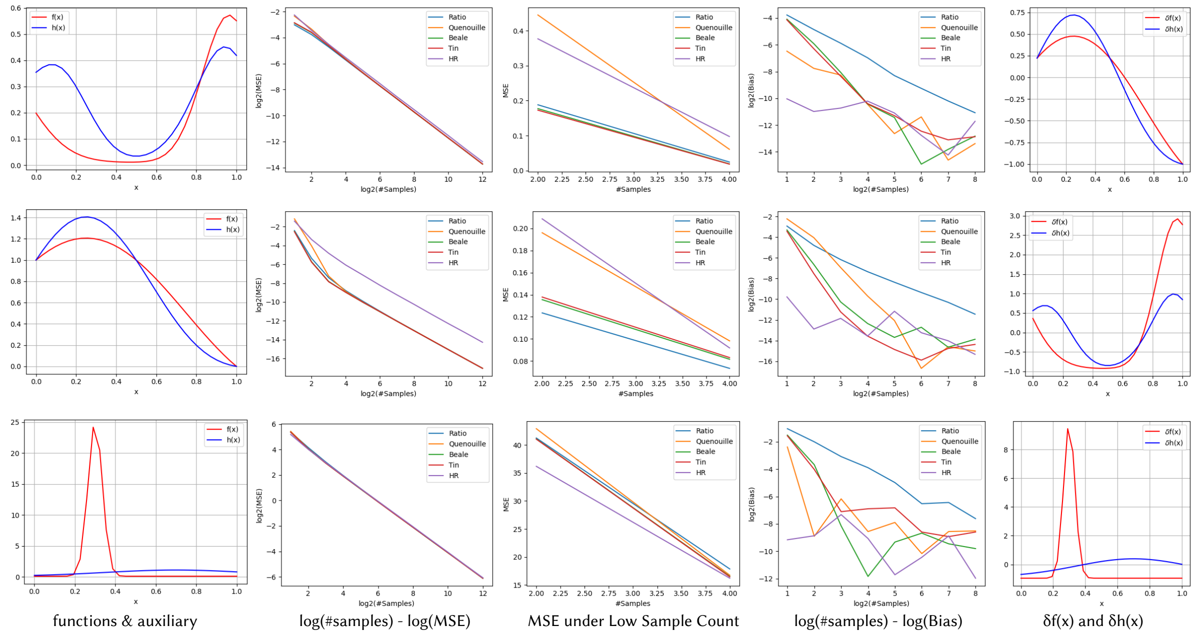


Fig. 2. We show a few 1D curve examples, measuring their mean squared error (MSE), low-sample MSE, and bias under different SPPs.

all methods should have very similar MSE. The reason is very likely because, in this example, the auxiliary $h(x)$, unfortunately, violates the assumption in the Taylor expansion approximation of the MSE for Hartley Ross's Estimator.

The variance approximation of Hartley Ross's Estimator [Goodman and Hartley 1958] highly depend on the assumption that:

$$\delta h(x) = \left|\frac{h(x) - H}{H}\right| \ll 1 \quad \text{and} \quad \delta f(x) = \left|\frac{f(x) - F}{F}\right| \ll 1. \tag{13}$$

Such an assumption is quite restrictive in practice, and we can observe in the rightmost column of Fig. 2 that both the second and third examples break the assumption. Hence, the approximation variance is far less legitimate. On the contrary, for the vanilla ratio estimator and its almost-unbiased variants, the corresponding assumption is:

$$\delta \bar{h}(x) = \left|\frac{\bar{h}(x) - H}{H}\right| \ll 1 \quad \text{and} \quad \delta \bar{f}(x) = \left|\frac{\bar{f}(x) - F}{F}\right| \ll 1. \tag{14}$$

As we know, the average of samples will ultimately converge to its expectation; with large sample counts, such an assumption is very loose. As a result, we may expect the actual MSE of Hartley Ross's estimator to potentially perform worse than other biased variants, especially in large sample counts.

4.3 Ratio Estimator with Nested Integration

In path tracing, multi-bounce path integrals are crucial for global illumination effects. This involves recursive integration estimation, which can be quite non-trivial for many non-nested estimators to extend to. For simplicity, we only discuss the case involving one nested integration, but all discussions can be trivially extended.

A 2-bounce path tracing problem can be formulated as:

$$F = \int f_1(x) F_2(x) \mathbf{d}x = \int f_1(x) \left(\int f_2(x, y) \mathbf{d}y \right) \mathbf{d}x, \tag{15}$$

where x and y are hemispherical solid angles on the first and second bounce, respectively. The internal integrand f_2 is conditional on x, as it varies with shading points determined by x.

A Naïve but Worrisome Adaptation. One naïve way to extend any non-nested estimator $\hat{F} = \mathrm{Est}(\cdot)$ to a nested problem is to apply it recursively at each level of integration:

$$\hat{F} = \mathrm{Est}(X_i \cdot \mathrm{Est}(X_i, Y_i)). \tag{16}$$

The fundamental path-tracing estimator we use follows such patterns, where $\mathrm{Est}(\cdot)$ used is a one-sample standard Monte Carlo estimator. However, for more general Est, using this formulation involves several critical considerations:

(1) Unbiasedness: If $\mathrm{Est}(\cdot)$ is an unbiased estimator, then $\mathrm{Est}(X_i \cdot \mathrm{Est}(X_i, Y_i))$ is also unbiased, thanks to the linearity.
(2) Consistency: However, if $\mathrm{Est}(\cdot)$ is a consistent estimator, then $\mathrm{Est}(X_i \cdot \mathrm{Est}(X_i, Y_i))$ is not necessarily consistent. As a counterexample, we show when Est is the ratio control variates estimator, the entire estimator may no longer keep consistency.
(3) Branching: We may apply debiasing techniques for inner estimations to avoid inconsistency. However, if the debiased estimator requires more than one sample, the branching may cause exponential evaluation increase when depth gets larger [West and Mukherjee 2024]. Unfortunately, all debias techniques discussed in Section 4.2 require at least two samples.

Path-Space Ratio Estimator. Although we cannot simultaneously avoid branching and inconsistency under Eq. (16), there is another way out. Specifically, we claim the following estimator is consistent:

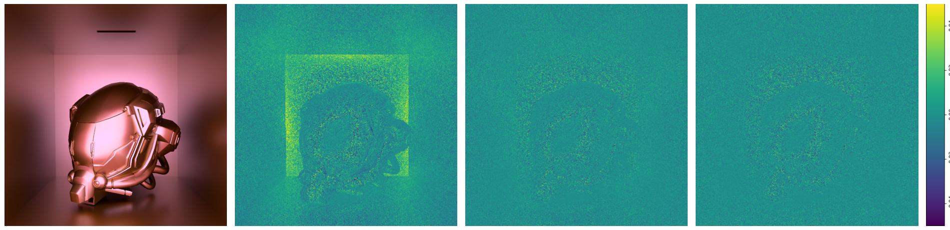
$$\hat{F}^{Ratio}_{Path} = \frac{\sum_{n=1}^{N} f_1(X_i) f_2(X_i, Y_i)}{\sum_{n=1}^{N} \frac{h_1(X_i)}{H_1} \frac{h_2(X_i, Y_i)}{H_2(x)}}. \tag{17}$$

It is easy to identify the nominator as an unbiased estimator of target integral cause this is exactly the basic path-tracing estimator. Therefore, once the denominator converges to one, the entire estimator will be consistent, and it can be proved as follows:

$$\mathbb{E}\left[\frac{h_1(X_i)}{H_1} \frac{h_2(X_i, Y_i)}{H_2(x)} \right] = \int \frac{h_1(x)}{H_1} \left(\int \frac{h_2(x, y)}{H_2(x)} \mathbf{d}y \right) \mathbf{d}x \tag{18}$$

$$= \int \frac{h_1(x)}{H_1} \cdot 1 \mathbf{d}x = 1. \tag{19}$$

One principal way to look at this is regarding the path formulation of the rendering equation, where we eventually formulate it as one global joint integral of all vertices, instead of some conditional nested integrals. We use the ratio control variates on this joint integral, and the critical point is we can construct such global auxiliary function



(a) (b) (c) (d)

Fig. 3. (a) Reference rendering of the scene with 3-bounce path tracing, with chromatic BRDF. (b) Bias of Naïve nested ratio estimator, N=20. (c) Bias of path-space ratio estimator, N=20. (d) Bias of Hartley-Ross estimator on top of path-space ratio estimator.

as the product of local normalized auxiliaries:

$$h(x, y, \cdots) = \frac{h_1(x)}{H_1} \cdot \frac{h_2(x, y)}{H_2(x)} \cdots . \tag{20}$$

As all random vairables $x, y, \cdots$ are indepdently drawn,

$$\mathbb{E}\left[h(x, y, \cdots)\right] = \mathbb{E}\left[\frac{h_1(x)}{H_1}\right] \cdot \mathbb{E}\left[\frac{h_2(x, y)}{H_2(x)}\right] \cdots = 1. \tag{21}$$

The general principle behind is $\mathbb{E}[A \cdot B] = \mathbb{E}[A] \cdot \mathbb{E}[B]$ when A and B are independent.

We can also easily build a debiased estimator on top of it without exponential branching. We simply substitute all f and h in Hartley Ross estimator with $f = f_1(X_i) f_2(X_i, Y_i)$ and $h = \frac{h_1(X_i)}{H_1} \cdot \frac{h_2(X_i, Y_i)}{H_2(X_i)}$.

Consistency of the path space ratio estimator and the unbiasedness of its Hartley-Ross variant are illustrated in Fig. 3.

4.4 Regression Ratio Estimator and Online Learning

For most variance reduction techniques, like importance sampling, control variates, and ratio control variates, the variance reduction may be less effective if the auxiliary function does not closely match the actual integrand. In the worst case, the variance could increase dramatically. We discuss how multiple auxiliaries are used in difference control variates with some confidence constants β in Section 3.1.

Inspired by that, we introduce its counterpart for regression control variates:

$$\hat{F}^{Ratio}_{Reg} = \left[\frac{1}{N}\sum_{i=1}^{N} f(X_i)\right]\left(\frac{H}{\frac{1}{N}\sum_{i=1}^{N} h(X_i)}\right)^{\alpha} \tag{22}$$

where α controls our belief in auxiliary $h(x)$. When $\alpha = 0$, it becomes a normal MC estimator; when $\alpha < 0$, it essentially becomes a product estimator. The estimator is always consistent for $\forall \alpha \in \mathbb{R}$. The same estimator has already been widely disccused in ratio estimator context [Naik and Gupta 1996].

Later in Section 4.5, we will show that the approximately optimal α can be obtained through regression, similar to the optimal β in difference control variates. We explored learning this parameter online, as shown in Fig. 4

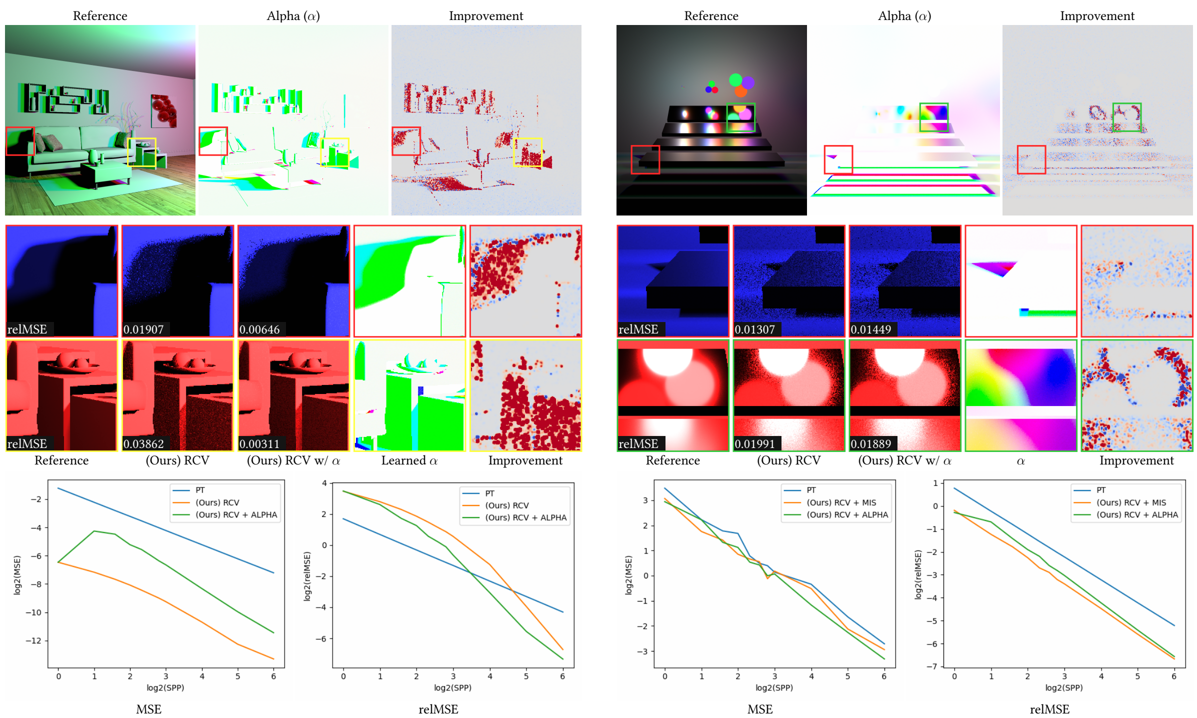


Fig. 4. We use NEE with RCV to handle chromatic lights and employ online regression to estimate the optimal α for Eq. (22). In the Bedroom example, we observe that in the shadow regions, the corresponding auxiliary performs poorly as it does not account for visibility, causing α to no longer be white. Using online learning for α significantly improves the results in these regions, as illustrated in the *improvement* figure, where red indicates improvement by using α.

4.5 Ratio Estimators with Multiple Auxiliary Variables

Multiple Importance Sampling (MIS) is widely used in rendering to combine multiple sampling techniques to reduce variance. We can also analogously combine multiple auxiliary functions for ratio estimators. Specifically, we introduce three approaches here.

Multiple Regression Ratio Estimator. Based on the fact that the $H/\left(\frac{1}{N}\sum h(X_i)\right)$ term converges to one when N approaches infinity, we can further see: 1) If we multiply multiple such terms, their product also converges to 1; 2) If we add arbitrary power on top of each term, it still converges to 1. Thus, we can get the estimator:

$$\hat{F}^{Ratio}_{MReg} = \left[\frac{1}{N}\sum_{i=1}^{N} f(X_i)\right]\left(\frac{H_1}{\frac{1}{N}\sum_{i=1}^{N} h_1(X_i)}\right)^{\alpha_1}\left(\frac{H_2}{\frac{1}{N}\sum_{i=1}^{N} h_2(X_i)}\right)^{\alpha_2} \tag{23}$$

The estimator has a strong connection with regression control variates with multiple auxiliaries [Owen 2013].

Lemma 11. The optimal MSE of Eq. (23) can be achieved when $\alpha_k^* = \beta_k^* \frac{\overline{h_k}}{\bar{f}}$, where β_k^* is the optimal regression constants for multiple regression difference control variates Section 3.1. (proved in Appendix A.11)

Random Mixture Multiple Ratio Estimator. As integration is a linear operation, $\int [\alpha_1 h_1(x) + \alpha_2 h_2(x)]\,\mathbf{d}x = \alpha_1 H_1 + \alpha_2 H_2$, which leads to an estimator analogous to random mixture [Owen 2013]:

$$\hat{F}^{Ratio}_{RMix} = \frac{\left[\sum_{i=1}^{N} f(X_i)\right][\alpha_1 H_1 + \alpha_2 H_2]}{\sum_{i=1}^{N}[\alpha_1 h_1(X_i) + \alpha_2 h_2(X_i)]} \tag{24}$$

Lemma 12. The optimal MSE of Eq. (24) can be achieved by solving a least square $A = B \cdot \alpha$, where

$$A_i = \int \frac{1}{g(x)}\,[h_i(x) - h_1(x)]\,[f(x) - F \cdot h_1(x)]\,\mathbf{d}x \tag{25}$$

$$B_{ik} = \int \frac{F}{g(x)}\,[h_i(x) - h_1(x)]\,[h_k(x) - h_1(x)]\,\mathbf{d}x. \tag{26}$$

(proved in Appendix A.12)

Deterministic Mixture Multiple Ratio Estimator. Moreover, we can also adapt the structure of deterministic mixture [Owen 2013], where we essentially decompose the target integrand $f(x)$ into 2 terms, $w_1(x)f(x)$ and $w_2(x)f(x)$, where $w_1(x) + w_2(x) = 1$ should hold for all x that $f(x) \neq 0$ just like MIS. Then, we evaluate each term separately and use different auxiliaries:

$$\hat{F}^{Ratio}_{DMix} = \frac{\sum_{i=1}^{N} w_1(X_i) f(X_i) H_1}{\sum_{i=1}^{N} h_1(X_i)} + \frac{\sum_{i=1}^{N} w_2(X_i) f(X_i) H_2}{\sum_{i=1}^{N} h_2(X_i)} \tag{27}$$

This is how we adapt for MIS in the main paper, but optimal weighting functions w_i seem very hard to write explicitly, as we discuss in Appendix A.13. In practice, we simply use either balance heuristic or power heuristic weights for simplicity.

REFERENCES

George E. Andrews. 1998. The Geometric Series in Calculus. The American Mathematical Monthly 105, 1 (1998), 36–40. http://www.jstor.org/stable/2589524

Beale EML. 1956. Some use of computers in operational research. 31 (1956), 27–28.

Leo A. Goodman and H. O. Hartley. 1958. The Precision of Unbiased Ratio-Type Estimators. J. Amer. Statist. Assoc. 53, 282 (1958), 491–508. http://www.jstor.org/stable/2281870

H. O. Hartley and A. Ross. 1954. Unbiased Ratio Estimators. Nature 174, 4423 (01 Aug 1954), 270–271. https://doi.org/10.1038/174270a0

Hee-Jin Hwang Jungtaek Oh and Key-Il Shin. 2021. A study on the generalized ratio-type estimator based on the multiple regression estimator. Communications in Statistics - Theory and Methods 50, 24 (2021), 6151–6166. https://doi.org/10.1080/03610926.2020.1740270 arXiv:https://doi.org/10.1080/03610926.2020.1740270

V. D. Naik and P. C. Gupta. 1996. A Note on Estimation of Mean with Known Population of Auxiliary Character. Journal of the Indian Society of Agricultural Statistics 48 (1996), 151–158.

Art B. Owen. 2013. Monte Carlo theory, methods and examples. https://artowen.su.domains/mc/.

Michael JD Powell and J Swann. 1966. Weighted uniform sampling—a Monte Carlo technique for reducing variance. IMA Journal of Applied Mathematics 2, 3 (1966), 228–236.

M. H. Quenouille. 1956. Notes on Bias in Estimation. Biometrika 43, 3/4 (1956), 353–360. http://www.jstor.org/stable/2332914

P V Sukhatme, B V Sukhatme, S Sukhatme, and C Asok. 1984. Sampling Theory of Surveys with Applications (third ed.). Indian Society of Agricultural Statistics. http://gen.lib.rus.ec/book/index.php?md5=D6771334D555A7A71B6E968CF57A70B0

Myint Tin. 1965. Comparison of Some Ratio Estimators. J. Amer. Statist. Assoc. 60, 309 (1965), 294–307. http://www.jstor.org/stable/2283154

Rex West and Sayan Mukherjee. 2024. Stylized Rendering as a Function of Expectation. ACM Trans. Graph. 43, 4, Article 96 (jul 2024), 19 pages. https://doi.org/10.1145/3658161

A PROOFS

A.1 Proof of Theorem 1

Theorem 1 The MSE of the vector-valued importance sampling estimator is minimized when the PDF g(x) is proportional to the L2-norm of integrand **f**: $g^*(x) \propto \sqrt{\sum_{k=1}^{K} f_k^2(x)}$.

Proof. The variance sum of the vector-valued importance sampling estimator is:

$$\mathrm{Var}\left[\hat{\mathbf{F}}^{IS}\right] = \sum_{j=1}^{M} \mathrm{Var}\left[\hat{F}_j^{\;IS}\right] = \frac{1}{N}\sum_{j=1}^{M}\left(\int_{\Omega}\frac{f_j^2(x)}{g(x)}\mathbf{d}x - F_j^2\right). \tag{28}$$

The minimizing PDF $g(x)$ can be found via Lagrange multipliers with functional derivative:

$$g^*(x) = \underset{g(x)}{\arg\min}\left[\sum_{j=1}^{M}\left(\int_{\Omega}\frac{f_j^2(x)}{g(x)}\mathbf{d}x - F_j^2\right) + \lambda\left(\int g(x)\mathbf{d}x - 1\right)\right] \tag{29}$$

$$\frac{\partial}{\partial g^*(x)}\left[\sum_{j=1}^{K}\left(\int_{\Omega}\frac{f_j^2(x)}{g^*(x)}\mathbf{d}x - F_j^2\right)\right] = 0 + \lambda \tag{30}$$

$$-\sum_{j=1}^{K}\frac{f_j^2(x)}{g^2(x)} + \lambda = 0 \tag{31}$$

$$\tag{32}$$

As a result, we can get the optimal $g^*(x)$:

$$g^*(x) = \sqrt{\frac{\sum_{j=1}^{K} f_j^2(x)}{\lambda}}. \tag{33}$$

On the other hand, we take the constrain $\int g(x)\mathbf{d}x = 1$ back to Eq. (33) to get complete $g^*(x)$:

$$\int_\Omega g^*(x) = \frac{\int_\Omega \sqrt{\sum_{k=1}^K f_k^2(x)}}{\sqrt{\lambda}} = 1 \tag{34}$$

$$\sqrt{\lambda} = \int_\Omega \sqrt{\sum_{k=1}^K f_k^2(x)} \tag{35}$$

$$g^*(x) = \frac{\sqrt{\sum_{k=1}^K f_k^2(x)}}{\int_\Omega \sqrt{\sum_{k=1}^K f_k^2(x)}} \tag{36}$$

We would see the minimum variance is achieved when we take $g^*(x) \propto \sqrt{\sum_{k=1}^K f_k^2(x)}$. As a degenerate case, when $K = 1$, we should simply sample proportional to the absolute value of $f(x)$, which is consistent with existing conclusion. □

A.2 Proof of Theorem 2

Theorem 2: The vector-valued importance sampling estimator is very likely to be incapable of achieving zero-variance, even if we use optimal $g^*(x)$ and all K integrands are non-negative.

Proof. By combining Eq. (28) and Eq. (36), we could see:

$$\mathrm{Var}\left[\langle \mathcal{F}_{N=1} \rangle^{\mathrm{LS}}\right]^* = \sum_{k=1}^K \left(\int_\Omega \frac{f_k^2(x)}{g^*(x)} \mathbf{d}x - \mu_k^2 \right) \tag{37}$$

$$= \int_\Omega \frac{\sum_{k=1}^K f_k^2(x)}{g^*(x)} \mathbf{d}x - \sum_{k=1}^K \mu_k^2 \tag{38}$$

$$= \int_\Omega \frac{\sum_{k=1}^K f_k^2(x)}{\frac{\sqrt{\sum_{k=1}^K f_k^2(x)}}{\mu_{L2}}} - \sum_{k=1}^K \mu_k^2 \tag{39}$$

$$= \mu_{L2} \int_\Omega \sqrt{\sum_{k=1}^K f_k^2(x)} - \sum_{k=1}^K \mu_k^2 \tag{40}$$

$$= \left(\int_\Omega \sqrt{\sum_{k=1}^K f_k^2(x) \mathbf{d}x} \right)^2 - \sum_{k=1}^K \left(\int_\Omega f_k(x) \mathbf{d}x \right)^2 \tag{41}$$

where we denote the normalizing term $\mu_{L2} = \int_\Omega \sqrt{\sum_{k=1}^K f_k^2(x)}$.

Despite the clean form of Eq. (41), we look back to Eq. (39) to step further.

$$\mathrm{Var}\left[\langle\mathcal{F}_{N=1}\rangle^{\mathrm{LS}}\right]^{*} = \int_{\Omega} \frac{\sum_{k=1}^{K} f_k^2(x)}{\underbrace{\sqrt{\sum_{k=1}^{K} f_k^2(x)}}_{\mu_{L2}}} - \sum_{k=1}^{K} \mu_k^2 \tag{42}$$

$$= \sum_{k=1}^{K}\left(\int_{\Omega} \frac{f_k^2(x)}{\sqrt{\sum_{k=1}^{K} f_k^2(x)}} \mathbf{d}x \mu_{L2} - \mu_k^2\right) \tag{43}$$

$$= \sum_{k=1}^{K}\left(\int_{\Omega} \frac{f_k^2(x)}{\sqrt{\sum_{k=1}^{K} f_k^2(x)}} \mathbf{d}x \int_{\Omega} \sqrt{\sum_{k=1}^{K} f_k^2(x)} \mathbf{d}x - \left(\int_{\Omega} f_k(x)\mathbf{d}x\right)^2\right) \tag{44}$$

$$\tag{45}$$

At this point, we use *Cauchy-Schwarz Inequality for Integrals* that:

$$\int a^2(x)\mathbf{d}x \int b^2(x)\mathbf{d}x \geq \left(\int a(x)b(x)\mathbf{d}x\right)^2. \tag{46}$$

When we let $a(x) = \frac{f_k(x)}{\left(\sum_{k=1}^{K} f_k^2(x)\right)^{\frac{1}{4}}}$, and $b(x) = \left(\sum_{k=1}^{K} f_k^2(x)\right)^{\frac{1}{4}}$, we can get:

$$\int_{\Omega} \frac{f_k^2(x)}{\sqrt{\sum_{k=1}^{K} f_k^2(x)}} \mathbf{d}x \int_{\Omega} \sqrt{\sum_{k=1}^{K} f_k^2(x)} \mathbf{d}x \geq \left(\int_{\Omega} f_k(x)\mathbf{d}x\right)^2 \tag{47}$$

Therefore, $\mathrm{Var}\left[\langle\mathcal{F}_{N=1}\rangle^{\mathrm{LS}}\right]^{*} \geq 0$, which is obvious as we would never get negative variance. More importantly, the equality holds iff all $a(x) = \frac{f_k(x)}{\left(\sum_{k=1}^{K} f_k^2(x)\right)^{\frac{1}{4}}}$ and $b(x) = \left(\sum_{k=1}^{K} f_k^2(x)\right)^{\frac{1}{4}}$ are linearly dependent, which leads to:

$$f_1(x) \propto f_2(x) \propto \cdots \propto f_K(x) \propto \sqrt{\sum_{k=1}^{K} f_k^2(x)}. \tag{48}$$

This is possible only when all function $f_K(x)$ have the same shape, and each of theorem need to be postivized as the last term is always non-negative. □

A.3 Proof of Theorem 3

Theorem 3: When using optimal distributions for both vector-valued importance sampling and scalar-valued importance sampling, the vector-valued importance sampling estimator always provides lower or equal variance than the scalar-valued version, under equal samples comparison.

Proof. From Eq. (70) and Eq. (41), we can see:

$$\mathrm{Var}\left[\langle \mathcal{F}_N \rangle^{\mathrm{SR}}\right]^* = \left(\sum_{k=1}^{K} \int_\Omega |f_k| \mathbf{d}x\right)^2 - \sum_{k=1}^{K} \left(\int_\Omega f_k(x) \mathbf{d}x\right)^2 \tag{49}$$

$$\mathrm{Var}\left[\langle \mathcal{F}_{N=1} \rangle^{\mathrm{LS}}\right]^* = \left(\int_\Omega \sqrt{\sum_{k=1}^{K} f_k^2(x)} \mathbf{d}x\right)^2 - \sum_{k=1}^{K} \left(\int_\Omega f_k(x) \mathbf{d}x\right)^2 \tag{50}$$

$$\mathrm{Var}\left[\langle \mathcal{F}_N \rangle^{\mathrm{SR}}\right]^* - \mathrm{Var}\left[\langle \mathcal{F}_{N=1} \rangle^{\mathrm{LS}}\right]^* = \left(\int_\Omega \sum_{k=1}^{K} |f_k| \mathbf{d}x\right)^2 - \left(\int_\Omega \sqrt{\sum_{k=1}^{K} f_k^2(x) \mathbf{d}x}\right)^2 \tag{51}$$

We can simply compare the left and right term by noticing the fact that the left term is a l1-norm of all f_k, while the right term is a l2-norm of all f_k. We can see the left term is always greater or equal to the right term, which means the lockstep estimator always provides equal or better variance than the single-response estimator.

To proof, we just need to notice the fact that:

$$(\sum_k |f_k(x)|)^2 = \sum_k |f_k(x)| \sum_k |f_k(x)| = \sum_k f_k(x)^2 + \sum_{k \neq j} |f_k(x)||f_j(x)| \geq \sum_k f_k(x)^2. \tag{52}$$

And by adding square root, integral, and square, we can get the inequality we need to proof:

$$\left(\int_\Omega \sum_{k=1}^{K} |f_k(x)| \mathbf{d}x\right)^2 \geq \left(\int_\Omega \sqrt{\sum_{k=1}^{K} f_k^2(x) \mathbf{d}x}\right)^2. \tag{53}$$

The equality holds iff for all x, at most one $f_k(x)$ in all K functions has non-zero value. This is consistent with our intuition because in this case, the lockstep estimator would only contribute to at most one function, making it equivalent to the single-response estimator. □

A.4 Proof of Theorem 4

Theorem 4: The variance of the random mixture importance sampling estimator is minimized when $g(x|k)p(k) \propto |f_k(x)|$.

Proof.

$$\mathrm{Var}\left[\langle \mathcal{F}_N \rangle^{\mathrm{SR}}\right] = \sum_{k=1}^{K} \mathrm{Var}\left[\left\langle F_{k,N} \right\rangle^{\mathrm{SR}}\right] \tag{54}$$

$$= \frac{1}{N} \sum_{k=1}^{K} \left[\int_{\Omega_k} \frac{f_k^2(x) \mathbf{1}_{k=k}(x)}{g(x|k)p(k)} \mathbf{d}x - \mu_k^2\right] \tag{55}$$

The minimizing PDF $g^*(x|k)$ can be found via Lagrange multipliers:

$$g^*(x|k) = arg\,min_{g(x|k)} \left[\sum_{k=1}^{K} \left(\int_\Omega \frac{f_k^2(x)\mathbf{1}_{k=k}(x)}{g(x|k)p(k)} \mathbf{d}x - \mu_k^2 \right) + \lambda \left(\int_{\Omega_k} g(x|k)\mathbf{d}x - 1 \right) \right] \tag{56}$$

$$0 = \frac{\partial}{\partial g^*(x|k)} \left[\int_\Omega \frac{f_k^2(x)\mathbf{1}_{k=k}(x)}{g^*(x|k)} \mathbf{d}x - \mu_k^2 \right] + \lambda \tag{57}$$

$$0 = -\frac{f_k^2(x)\mathbf{1}_{k=k}(x)}{g^{*2}(x|k)} + \lambda \tag{58}$$

$$g^*(x|k) = \sqrt{\frac{f_k^2(x)}{\lambda}} = \frac{|f_k(x)|}{\int |f_k(x)|\mathbf{d}x}. \tag{59}$$

Similarly, the minimizing $p^*(k)$ can be found by:

$$p^*(k) = arg\,min_{p(k)} \left[\sum_{k=1}^{K} \left(\int_\Omega \frac{f_k^2(x)\mathbf{1}_{k=k}(x)}{g(x|k)p(k)} \mathbf{d}x - \mu_k^2 \right) + \lambda \left(\sum_{k=1}^{K} p(k) - 1 \right) \right] \tag{60}$$

$$0 = \frac{\partial}{\partial p^*(k)} \left[\sum_{k=1}^{K} \left(\int_\Omega \frac{f_k^2(x)\mathbf{1}_{k=k}(x)}{g(x|k)p(k)} \mathbf{d}x - \mu_k^2 \right) + \lambda \left(\sum_{k=1}^{K} p(k) - 1 \right) \right] \tag{61}$$

$$0 = -\int_\Omega \frac{f_k^2(x)\mathbf{1}_{k=k}(x)}{g(x|k)p^{*2}(k)} \mathbf{d}x + \lambda \tag{62}$$

$$p^{*2}(k) = \int_\Omega \frac{f_k^2(x)\mathbf{1}_{k=k}(x)}{g(x|k)\lambda} \mathbf{d}x \tag{63}$$

$$p^{*2}(k) = \frac{1}{\lambda} \left(\int_\Omega |f_k|\mathbf{d}x \right)^2 \tag{64}$$

$$p^*(k) = \frac{\int_\Omega |f_k|\mathbf{d}x}{\sqrt{\lambda}} \tag{65}$$

$$p^*(k) = \frac{\int_\Omega |f_k|\mathbf{d}x}{\sum_{k=1}^{K} \int_\Omega |f_k|\mathbf{d}x} \tag{66}$$

Therefore, the variance is minimized when each sample $X_{j,k}$ is sampled proportional to $|f_k(x)|$:

$$g^*(x|k) \cdot p^*(k) = \frac{|f_k(x)|}{\int |f_k(x)|\mathbf{d}x} \frac{\int_\Omega |f_k|\mathbf{d}x}{\sum_{k=1}^{K} \int_\Omega |f_k|\mathbf{d}x} = \frac{|f_k(x)|}{\sum_{k=1}^{K} \int_\Omega |f_k|\mathbf{d}x}, \tag{67}$$

where we will have the minimum variance of the estimator:

$$\mathrm{Var}\left[\langle\mathcal{F}_N\rangle^{\mathrm{SR}}\right]^* = \sum_{k=1}^{K}\int_{\Omega_k}\frac{f_k^2(x)\mathbf{1}_{k=k}(x)}{g^*(x|k)p^*(k)}\mathbf{d}x - \mu_k^2 \tag{68}$$

$$= \sum_{k=1}^{K}\int_{\Omega_k}\frac{f_k^2(x)\mathbf{1}_{k=k}(x)}{\frac{|f_k(x)|}{\sum_{k=1}^{K}\int_{\Omega}|f_k|\mathbf{d}x}}\mathbf{d}x - \mu_k^2 \tag{69}$$

$$= \left(\sum_{k=1}^{K}\int_{\Omega}|f_k|\mathbf{d}x\right)^2 - \sum_{k=1}^{K}\left(\int_{\Omega}f_k(x)\mathbf{d}x\right)^2 \tag{70}$$

□

A.5 Proof of Theorem 5

Theorem 5: When using optimal distributions for both vector-valued importance sampling (Estimator 1) and random-mixture importance sampling (Estimator 3), the vector-valued importance sampling estimator always provides lower or equal variance than the random-mixture version, under equal samples comparison.

Proof. From Eq. (70) and Eq. (41), we can see:

$$\mathrm{Var}\left[\langle\mathcal{F}_N\rangle^{\mathrm{SR}}\right]^* = \left(\sum_{k=1}^{K}\int_{\Omega}|f_k|\mathbf{d}x\right)^2 - \sum_{k=1}^{K}\left(\int_{\Omega}f_k(x)\mathbf{d}x\right)^2 \tag{71}$$

$$\mathrm{Var}\left[\langle\mathcal{F}_{N=1}\rangle^{\mathrm{LS}}\right]^* = \left(\int_{\Omega}\sqrt{\sum_{k=1}^{K}f_k^2(x)\mathbf{d}x}\right)^2 - \sum_{k=1}^{K}\left(\int_{\Omega}f_k(x)\mathbf{d}x\right)^2 \tag{72}$$

$$\mathrm{Var}\left[\langle\mathcal{F}_N\rangle^{\mathrm{SR}}\right]^* - \mathrm{Var}\left[\langle\mathcal{F}_{N=1}\rangle^{\mathrm{LS}}\right]^* = \left(\int_{\Omega}\sum_{k=1}^{K}|f_k|\mathbf{d}x\right)^2 - \left(\int_{\Omega}\sqrt{\sum_{k=1}^{K}f_k^2(x)\mathbf{d}x}\right)^2 \tag{73}$$

We can simply compare the left and right term by noticing the fact that the left term is a l1-norm of all f_k, while the right term is a l2-norm of all f_k. We can see the left term is always greater or equal to the right term, which means the lockstep estimator always provides equal or better variance than the single-response estimator.

To proof, we just need to notice the fact that:

$$(\sum_k |f_k(x)|)^2 = \sum_k |f_k(x)| \sum_k |f_k(x)| = \sum_k f_k(x)^2 + \sum_{k\neq j}|f_k(x)||f_j(x)| \geq \sum_k f_k(x)^2. \tag{74}$$

And by adding square root, integral, and square, we can get the inequality we need to proof:

$$\left(\int_{\Omega}\sum_{k=1}^{K}|f_k(x)|\mathbf{d}x\right)^2 \geq \left(\int_{\Omega}\sqrt{\sum_{k=1}^{K}f_k^2(x)\mathbf{d}x}\right)^2. \tag{75}$$

The equality holds iff for all x, at most one $f_k(x)$ in all K functions has non-zero value. This is consistent with our intuition because in this case, the lockstep estimator would only contribute to at most one function, making it equivalent to the single-response estimator. □

A.6 Proof of Theorem 6

Theorem 6: When using arbitrary but same distributions $g(x)$ for both vector-valued importance sampling (Estimator 1) and the partial vector-valued version (Estimator 4), the vector-valued importance sampling estimator always provides lower or equal variance, under equal samples comparison.

PROOF.

$$\mathrm{Var}\left[\langle \mathcal{F}_N \rangle^{\text{Rej-LS}}\right] = \sum_k \left(\int_\Omega \frac{f^2(x)}{g(x) p_{acc}(x|k)} \mathbf{d}\mu(x) - \mu_k^2 \right) \tag{76}$$

$$\geq \sum_k \left(\int_\Omega \frac{f^2(x)}{g(x)} \mathbf{d}\mu(x) - \mu_k^2 \right) \tag{77}$$

$$= \mathrm{Var}\left[\langle \mathcal{F}_N \rangle^{\text{LS}}\right]. \tag{78}$$

Given the fact that $p_{acc}(x|k) \leq 1$, we can observe the variance of the rejectable lockstep estimator is always greater or equal to its counterpart without rejection. The equality holds iff the rejection only happens when $f_k(x) = 0$. Therefore rejection itself would not help to reduce the variance of the estimator, intuitively because of the rejection merely wastes more samples, while importance sampling expects us to move those reduced samples to the high-value regions. □

A.7 Proof of Theorem 7

PROOF. As an unbiased estimator, the MSE equals to its variance.

$$\hat{f} = \bar{f} - \sum_{k=1}^{K} \beta_k \overline{h_k} + \sum_{k=1}^{K} \beta_k H_k \tag{79}$$

$$\sigma^2 = \mathbb{E}\left[\left(f(x) - F - \sum_{k=1}^{K} \beta_k \left(h_k(x) - H_k \right) \right)^2 \right] \tag{80}$$

$$= \mathrm{Var}\left[f(x) \right] + \mathrm{Var}\left[\sum_{k=1}^{K} \beta_k \left(h_k(x) - H_k \right) \right] \tag{81}$$

$$- 2\mathrm{Cov}\left[f(x), \sum_{k=1}^{K} \beta_k \left(h_k(x) - H_k \right) \right] \tag{82}$$

$$= \mathrm{Var}\left[f(x) \right] - 2 \sum_{k=1}^{K} \beta_k \mathrm{Cov}\left[f(x), (h_k(x) - H_k) \right] \tag{83}$$

$$+ 2 \sum_i \sum_j \beta_i \beta_j \mathrm{Cov}((h_i(x) - H_i), (h_j(x) - H_j)) \tag{84}$$

By differentiating, we will see

$$\beta^* = \mathrm{Var}(h(x))^{-1} \mathrm{Cov}(h(x), f(x)) \tag{85}$$

□

A.8 Proof of Theorem 8

For the ratio estimator, the sample summation happens in both the numerator and the denominator, therefore we cannot directly simplify it to $N = 1$ case. Therefore, we will directly evaluate the MSE in high-dimensional space Y, where the element is a list of samples, $Y_i = (X_1, X_2, \cdots, X_N)$. Therefore, for a given sample count N, we can have:

$$\mathrm{MSE}\left[\hat{F}^{Ratio}\right] = \int \left[\frac{\sum_{j=1}^{N} \frac{f(X_j)}{g(X_j)}}{\sum_{j=1}^{N} \frac{h(X_j)}{g(X_j)}} - \int f(x)\mathbf{d}x\right]^2 \prod_{j=1}^{N} g(X_j)\mathbf{d}x \tag{86}$$

$$= \int \left[\frac{\sum_{j=1}^{N} \frac{f(X_j)-F\cdot h(x)}{g(X_j)}}{\sum_{j=1}^{N} \frac{h(X_j)}{g(X_j)}}\right]^2 \prod_{j=1}^{N} g(X_j)\mathbf{d}x \tag{87}$$

$$= \int \left[\frac{\sum_{j=1}^{N} \frac{f_{cv}(X_j)}{g(X_j)}}{\sum_{j=1}^{N} \frac{h(X_j)}{g(X_j)}}\right]^2 \prod_{j=1}^{N} g(X_j)\mathbf{d}x, \tag{88}$$

where we denote $f_{cv}(x) = f(x) - F \cdot h(x)$ for simplicity. For further simplification, we extend Powell's[Powell and Swann 1966] proof for the vanilla version of the WIS estimator, where $g(x) = 1$. It start with splitting the domain Y into two parts, A_N and B_N, defined as:

$$A_N = \left\{y : \left|\sum_{j=1}^{N} \frac{h(X_i)}{g(X_i)} - N\right| < N^{0.8}\right\} \tag{89}$$

$$B_N = \left\{y : \left|\sum_{j=1}^{N} \frac{h(X_i)}{g(X_i)} - N\right| \geq N^{0.8}\right\}. \tag{90}$$

Then we can split the integral Eq. (88) into two parts:

$$\mathrm{MSE}\left[\hat{F}^{Ratio}\right] = \int_{A_N} \left[\frac{\sum_{j=1}^{N} \frac{f_{cv}(X_j)}{g(X_j)}}{\sum_{j=1}^{N} \frac{h(X_j)}{g(X_j)}}\right]^2 \prod_{j=1}^{N} g(X_j)\mathbf{d}x \tag{91}$$

$$+ \int_{B_N} \left[\frac{\sum_{j=1}^{N} \frac{f_{cv}(X_j)}{g(X_j)}}{\sum_{j=1}^{N} \frac{h(X_j)}{g(X_j)}}\right]^2 \prod_{j=1}^{N} g(X_j)\mathbf{d}x. \tag{92}$$

Now we first try to simplify Eq. (92), by noticing the subdomain A_N is a small interval around N:

$$N(1 - N^{-0.2}) < \sum_{j=1}^{N} \frac{h(X_i)}{g(X_i)} < N(1 + N^{-0.2}). \tag{93}$$

When N is large enough, we can use the approximation $\sum_{j=1}^{N} \frac{h(X_i)}{g(X_i)} \approx N$, and the approximation deviation is up to a constant factor $(1 - N^{-0.2})$.

According to the domain division, we can then transform the first term in Eq. (92) as follows:

$$\int_{A_N} \left[\frac{\sum_{j=1}^{N} \frac{f'(x_i)}{g(x_i)}}{\sum_{j=1}^{N} \frac{h(x_j)}{g(x_j)}} \right]^2 \prod_{j=1}^{N} g(X_j) \mathbf{d}y \approx \frac{1}{N^2} \int_{A_N} \left[\sum_{j=1}^{N} \frac{f'(x_i)}{g(x_i)} \right]^2 \prod_{j=1}^{N} g(X_j) \mathbf{d}y \tag{94}$$

$$= \frac{1}{N^2} \left(\int_J \left[\sum_{j=1}^{N} \frac{f'(x_i)}{g(x_i)} \right]^2 \prod_{j=1}^{N} g(X_j) \mathbf{d}y - \int_{B_N} \left[\sum_{j=1}^{N} \frac{f'(x_i)}{g(x_i)} \right]^2 \prod_{j=1}^{N} g(X_j) \mathbf{d}y \right) \tag{95}$$

We will simplify the first term in Eq. (95) further, but firstly, we need to realize the fact:

$$\int f_{\mathrm{cv}}(x) \mathbf{d}x = \int f(x) - \mathbf{F} \cdot h(x) \mathbf{d}x = \int f(x) \mathbf{d}x - \mathbf{F} \cdot \int h(x) \mathbf{d}x = \mathbf{F} - \mathbf{F} \cdot 1 = 0. \tag{96}$$

Then we can simplify the first term in Eq. (95) as follows:

$$\int_J \left[\sum_{i=1}^{N} \frac{f_{cv}(x_i)}{g(x_i)} \right]^2 \prod_{k=1}^{N} g(X_k) \mathbf{d}y \tag{97}$$

$$= \int_I \cdots \int_I \left[\sum_{i=1}^{N} \frac{f_{cv}(x_i)}{g(x_i)} \right]^2 \prod_{k=1}^{N} g(X_k) \mathbf{d}X_1 \cdots \mathbf{d}X_N \tag{98}$$

$$= \int_I \cdots \int_I \left[\sum_{i=1}^{N} \sum_{j=1}^{N} \frac{f_{cv}(x_i)}{g(x_i)} \frac{f_{cv}(x_j)}{g(x_j)} \right] \prod_{k=1}^{N} g(X_k) \mathbf{d}X_1 \cdots \mathbf{d}X_N \tag{99}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} \int_I \cdots \int_I \left[\frac{f_{cv}(x_i)}{g(x_i)} \frac{f_{cv}(x_j)}{g(x_j)} \right] \prod_{k=1}^{N} g(X_k) \mathbf{d}X_1 \cdots \mathbf{d}X_N \tag{100}$$

$$= \sum_{i=1}^{N} \int_I \cdots \int_I \frac{[f_{cv}(x_i)]^2}{g(x_i)} \prod_{k=1, k \neq i}^{N} g(X_k) \mathbf{d}X_1 \cdots \mathbf{d}X_N \tag{101}$$

$$+ \sum_{i=1}^{N} \sum_{j=1}^{N} \int_I \cdots \int_I \left[f_{cv}(x_i) f_{cv}(x_j) \right] \prod_{k=1, k \neq i,j}^{N} g(X_k) \mathbf{d}X_1 \cdots \mathbf{d}X_N \tag{102}$$

$$= N \int_I \frac{[f_{cv}(x)]^2}{g(x)} \mathbf{d}x \tag{103}$$

Noticing that Eq. (102) is 0 because of Eq. (96). Intuitively, it is exactly the same as how the variance of the Monte Carlo estimator is reduced with $\frac{1}{N}$ factor. Then we can simplify the first term of Eq. (95) as:

$$\frac{1}{N^2} \int_J \left[\sum_{j=1}^{N} \frac{f_{cv}(x_i)}{g(x_i)} \right]^2 \prod_{j=1}^{N} g(X_j) \mathbf{d}y = \frac{1}{N} \int_I \frac{[f_{cv}(x)]^2}{g(x)} \mathbf{d}x \tag{104}$$

Then we move to the second term of Eq. (92), where we will proof it combined with the second term of Eq. (95) is a higher-order infinitesimal compared to the first term, i.e.:

$$\int_{B_N} \left(\left[\frac{\sum_{j=1}^{N} \frac{f_{cv}(X_j)}{g(X_j)}}{\sum_{j=1}^{N} \frac{h(X_j)}{g(X_j)}} \right]^2 - \frac{1}{N^2} \left[\sum_{j=1}^{N} \frac{f_{cv}(x_i)}{g(x_i)} \right]^2 \right) \prod_{j=1}^{N} g(X_j) \mathbf{d}x \approx o(\frac{1}{N}). \tag{105}$$

With valid setting of $f(x)$, $h(x)$ and $g(x)$, we should ensure the $\left(\left[\frac{\sum_{j=1}^{N}\frac{f_{cv}(X_j)}{g(X_j)}}{\sum_{j=1}^{N}\frac{h(X_j)}{g(X_j)}}\right]^2 - \frac{1}{N^2}\left[\sum_{j=1}^{N}\frac{f_{cv}(x_i)}{g(x_i)}\right]^2\right)$ term should be bounded by a constant. Therefore, we only need to proof the remaining integral be higher-order infinitesimal: $\int_{B_N}\prod_{j=1}^{N} g(X_j)\mathbf{d}x = o(\frac{1}{N})$.

To proof this, we first look at the forth moment of the function $\left(\frac{h(x)}{g(x)} - 1\right)$:

$$M_4 = \int_J \left[\sum_{i=1}^{N}\frac{h(X_i)}{g(X_i)} - N\right]^4 \prod_{m=1}^{N} g(X_m)\mathbf{d}y \tag{106}$$

$$= \int_J \left[\sum_{i=1}^{N}\left(\frac{h(X_i)}{g(X_i)} - 1\right)\right]^4 \prod_{m=1}^{N} g(X_m)\mathbf{d}y \tag{107}$$

$$= \sum_i\sum_j\sum_k\sum_l \int_J \left(\frac{h(X_i)}{g(X_i)} - 1\right)\left(\frac{h(X_j)}{g(X_j)} - 1\right) \tag{108}$$

$$\left(\frac{h(X_k)}{g(X_k)} - 1\right)\left(\frac{h(X_l)}{g(X_l)} - 1\right)\prod_{m=1}^{N} g(X_m)\mathbf{d}y \tag{109}$$

$$= N\int_I \left(\frac{h(x)}{g(x)} - 1\right)^4 g(x)\mathbf{d}x \tag{110}$$

$$+ 3N(N-1)\left[\int_I \left(\frac{h(x)}{g(x)} - 1\right)^2 g(x)\mathbf{d}x\right]^2 \tag{111}$$

$$= O(N^2). \tag{112}$$

By the definition of B_N,

$$M \geq \int_{B_N}\left[\sum_{i=1}^{N}\frac{h(X_i)}{g(X_i)} - N\right]\prod_{m=1}^{N} g(X_m)\mathbf{d}y \tag{113}$$

$$\geq \int_{B_N}(N^{0.8})^4\prod_{m=1}^{N} g(X_m)\mathbf{d}y \tag{114}$$

$$= N^{3.2}\int_{B_N}\prod_{m=1}^{N} g(X_m)\mathbf{d}y \tag{115}$$

$$O(N^2) \geq N^{3.2}\int_{B_N}\prod_{m=1}^{N} g(X_m)\mathbf{d}y \tag{116}$$

$$\int_{B_N}\prod_{m=1}^{N} g(X_m)\mathbf{d}y \leq O(N^{-1.2}) \tag{117}$$

$$\int_{B_N}\prod_{m=1}^{N} g(X_m)\mathbf{d}y = o(\frac{1}{N}). \tag{118}$$

Therefore, if we discard the higher-order infinitesimal, we can approximate the MSE with only the first term of Eq. (95), and by Eq. (104), we finally have:

$$\mathrm{MSE}\left[\hat{F}^{Ratio}\right] \approx \frac{1}{N}\int \frac{(f(x) - F \cdot h(x))^2}{g(x)}\mathbf{d}x. \tag{119}$$

A.9 Proof of Theorem 10

To start with, we define $f = F + \epsilon$ and $h = H + \xi$, thus:

$$\mathbb{E}\left[\bar{\epsilon}_N\right] = \mathbb{E}\left[\sum_i^N f(X_i) - F\right] = 0 \tag{120}$$

$$\mathbb{E}\left[\bar{\xi}_N\right] = \mathbb{E}\left[\sum_i^N h(X_i) - F\right] = 0 \tag{121}$$

$$\mathbb{E}\left[\bar{\xi}_N^2\right] = \mathbb{E}\left[\left(\sum_i^N h(X_i) - H\right)^2\right] = \frac{1}{N}\mathrm{Var}\left[h(x)\right] \tag{122}$$

Then, we rewrite the expectation of the ratio estimator with ϵ and ξ definition:

$$\mathbb{E}\left[\frac{\bar{f}}{\bar{h}}\right] = \mathbb{E}\left[\frac{F\left(1 + \frac{\bar{\epsilon}_N}{F}\right)}{H\left(1 + \frac{\bar{\xi}_N}{H}\right)}\right] \tag{123}$$

To further simplify, we consider the **geometric series**, where it is well-known [Andrews 1998] that: $\sum_{i=0}^{\infty} ax^i = \frac{a}{1-x}$ holds for $|x| < 1$. By letting $a = 1$ and $x = -t$, we can have:

$$(1+t)^{-1} = \sum_{i=0}^{\infty}(-t)^i = 1 - t + t^2 - t^3 + \cdots \tag{124}$$

As $\bar{\xi}_N$ have expectation 0 and standard Monte Carlo convergence rate, it is fair to assume $\frac{\bar{\xi}_N}{H} \leq 1$ when N is relatively large. Thus by substitute Eq. (124) into Eq. (123), we have:

$$\mathbb{E}\left[\frac{\bar{f}}{\bar{h}}\right] = \mathbb{E}\left[\frac{F\left(1 + \frac{\bar{\epsilon}_N}{F}\right)}{H}\left[1 - \frac{\bar{\xi}_N}{H} + \left(\frac{\bar{\xi}_N}{H}\right)^2 - \left(\frac{\bar{\xi}_N}{H}\right)^3 + \cdots\right]\right] \tag{125}$$

$$= \frac{F}{H}\mathbb{E}\left[1 + \frac{\bar{\epsilon}_N}{F} - \frac{\bar{\xi}_N}{H} - \frac{\bar{\epsilon}_N}{F}\frac{\bar{\xi}_N}{H} + \left(\frac{\bar{\xi}_N}{H}\right)^2 + \cdots\right] \tag{126}$$

$$\approx \frac{F}{H}\mathbb{E}\left[1 + \frac{\bar{\epsilon}_N}{F} - \frac{\bar{\xi}_N}{H} - \frac{\bar{\epsilon}_N}{F}\frac{\bar{\xi}_N}{H} + \left(\frac{\bar{\xi}_N}{H}\right)^2\right] \tag{127}$$

by neglecting the higher than 2-order terms. Because of Eq. (120) to Eq. (122),the only unknown term is $\mathbb{E}\left[\bar{\epsilon}_N\bar{\xi}_N\right]$, which we derive as:

$$\mathbb{E}\left[\bar{\epsilon}_N\bar{\xi}_N\right] = \frac{1}{N^2}\mathbb{E}\left[\sum_i^N \epsilon(X_i)\sum_i^N \xi(X_i)\right] \tag{128}$$

$$= \frac{1}{N^2}\mathbb{E}\left[\sum_{i=1}^N \epsilon(X_i)\xi(X_i) + \sum_{i\neq j}\epsilon(X_i)\xi(X_j)\right] \tag{129}$$

$$= \frac{1}{N}\int (f(x)-F)(h(x)-H)\mathbf{d}x \tag{130}$$

Also combining Eq. (120) to Eq. (122), we can see:

$$\mathbb{E}\left[\frac{\bar{f}}{\bar{h}}\right] \approx \frac{F}{H}\left(1+\frac{1}{N}\left[\frac{\mathrm{Var}\left[h(x)\right]}{H^2} - \frac{\mathrm{Cov}\left[h(x),f(x)\right]}{FH}\right]\right) \tag{131}$$

Thus, the bias can be approximated as:

$$Bias\left[\frac{\bar{f}}{\bar{h}}\cdot H\right] \approx \frac{F}{N}\left[\frac{\mathrm{Var}\left[h(x)\right]}{H^2} - \frac{\mathrm{Cov}\left[h(x),f(x)\right]}{FH}\right] \tag{132}$$

A.10 Proof of the unbiasedness of Hartley Ross

To proof the unbiasedness of Hartley Ross's Estimator [Hartley and Ross 1954], we start from $\mathrm{Cov}(A,B) = \mathbb{E}\left[AB\right] - \mathbb{E}\left[A\right]\mathbb{E}\left[B\right]$, and by taking $A = \frac{f(X_i)}{h(X_i)}$ and $B = h(X_i)$, we have: $\mathrm{Cov}(\frac{f}{h},h) = \mathbb{E}\left[f\right] - \mathbb{E}\left[\frac{f}{h}\right]\mathbb{E}\left[h\right]$. Therefore:

$$\frac{\mathbb{E}\left[f\right]}{\mathbb{E}\left[h\right]} = \mathbb{E}\left[\frac{f}{h}\right] + \frac{\mathrm{Cov}(\frac{f}{h},h)}{\mathbb{E}\left[h\right]} = \frac{1}{N}\sum_{i=1}^N\frac{f(X_i)}{h(X_i)} \tag{133}$$

$$+\frac{N\sum_{i=1}^N\left(\frac{f(X_i)}{h(X_i)} - \frac{1}{N}\sum_j\frac{f(X_j)}{h(X_j)}\right)\left(h(X_i)-\frac{1}{N}\sum_j h(X_j)\right)}{(N-1)H} \tag{134}$$

$$= \frac{1}{N}\sum_{i=1}^N\frac{f(X_i)}{h(X_i)} + \frac{N}{N-1}\frac{\sum_{i=1}^N f(X_i) - \sum_{i=1}^N\frac{f(X_i)}{h(X_i)}\sum_{i=1}^N h(X_i)}{H}. \tag{135}$$

A.11 Proof of Theorem 11

Proof. For the multiple regression ratio estimator:

$$\hat{F}_{MR}^{Ratio} = \left[\frac{1}{N}\sum_{i=1}^N f(X_i)\right]\prod_{k=1}^K\left(\frac{H_1}{\frac{1}{N}\sum_{i=1}^N h_k(X_i)}\right)^{\alpha_k} \tag{136}$$

We use first-order Taylor expansion [Jungtaek Oh and Shin 2021] to approximate the estimator:

$$\hat{F}_{MR}^{Ratio} = \left[\frac{1}{N}\sum_{i=1}^{N} f(X_i)\right] \prod_{k=1}^{K} \left(\frac{H_k}{\frac{1}{N}\sum_{i=1}^{N} h_k(X_i)}\right)^{\alpha_k} \tag{137}$$

$$\approx \bar{f}\left[1 - \sum_{k=1}^{K} \alpha_k \left(1 - \frac{\overline{H_k}}{\overline{h_k}}\right)\right] \tag{138}$$

$$= \bar{f}(1 - \sum_{k=1}^{K} \alpha_k) + \sum_{k=1}^{K} \alpha_k \frac{\bar{f}}{\overline{h_k}} H_k \tag{139}$$

To minimize the approximate MSE, we first define $\beta_k = \alpha_k \frac{\bar{f}}{\overline{h_k}}$, thus:

$$\hat{I} = \bar{f}(1 - \sum_{k=1}^{K} \alpha_k) + \sum_{k=1}^{K} \alpha_k \frac{\bar{f}}{\overline{h_k}} H_k \tag{140}$$

$$= \bar{f}(1 - \sum_{k=1}^{K} \beta_k \frac{\overline{h_k}}{\bar{f}}) + \sum_{k=1}^{K} \beta_k H_k \tag{141}$$

$$= \bar{f} - \sum_{k=1}^{K} \beta_k \overline{h_k} + \sum_{k=1}^{K} \beta_k H_k. \tag{142}$$

We can observe Eq. (142) is the same as the Multiple Regression Control Variates estimator in Eq. (79). Conceptually, it means the first-order approximate MSE of this estimator is equivalent to the Multiple Regression Control Variates estimator, with an extra $\frac{\bar{f}}{\overline{h_k}} \approx \frac{F}{H_k}$ scaling factor for each auxiliary variable h_k.

Therefore, minimizing the approximate MSE, we can start with solving the optimal β_k^* for the equivalent Multiple Regression Control Variates estimator, and simply get the optimal $\alpha_k^* = \beta_k^* \frac{\overline{h_k}}{\bar{f}}$. □

A.12 Proof of Theorem 12

Proof.

$$\hat{F}_{RM}^{Ratio} = \frac{\left[\sum_{i=1}^{N} f(X_i)\right]\left[\sum_{k=1}^{K} \alpha_k H_k\right]}{\sum_{i=1}^{N}\left[\sum_{k=1}^{K} \alpha_k h_k(X_i)\right]}, \tag{143}$$

where $\alpha_k, k = 1, \cdots, K$ are constants with constraint $\sum_{k=1}^{K} \alpha_k = 1$.

With the aforementioned approximation:

$$\mathrm{MSE}(\hat{F}_{RM}^{Ratio}) \approx \int \frac{\left(f(x) - F \cdot (\sum_{k=1}^{K} \alpha_k h_k(x))\right)^2}{g(x)} \mathbf{d}x \tag{144}$$

The minimizing parameters α_k can be found via minimizing Lagrangian:

$$\mathbf{L} = \mathrm{MSE}(\hat{F}_{RM}^{Ratio}) - \lambda(\sum_{k=1}^{K} \alpha_k - 1) \tag{145}$$

$$\frac{\partial \mathbf{L}}{\partial \alpha_i} = \int \frac{2\,(-F \cdot h_i(x))}{g(x)} \left(f(x) - F \cdot (\sum_{k=1}^{K} \alpha_k h_k(x))\right) \mathbf{d}x - \lambda \tag{146}$$

As we know $\frac{\partial \mathbf{L}}{\partial \alpha_i} = 0$ for all $i = 1, \cdots, K$, we first cancel out the constant λ by subtracting $\frac{\partial \mathbf{L}}{\partial \alpha_i} - \frac{\partial \mathbf{L}}{\partial \alpha_1} = 0 =$

$$\int \frac{2\left(-F \cdot (h_i(x) - h_1(x))\right)}{g(x)} \left(f(x) - F \cdot (\sum_{k=1}^{K} \alpha_k h_k(x)) \right) \mathbf{d}x \tag{147}$$

Then, we remove all α_1 by substitute $\alpha_1 = 1 - \sum_{k=2}^{K} \alpha_k$, thus:

$$\sum_{k=1}^{K} \alpha_k h_k(x) = h_1(x) + \sum_{k=2}^{K} \alpha_k (h_k(x) - h_1(x)). \tag{148}$$

To further simplify the form, we denote:

$$A_i = \int \frac{1}{g(x)} \left[h_i(x) - h_1(x)\right] \left[f(x) - F \cdot h_1(x)\right] \mathbf{d}x \tag{149}$$

$$B_{ik} = \int \frac{F}{g(x)} \left[h_i(x) - h_1(x)\right] \left[h_k(x) - h_1(x)\right] \mathbf{d}x. \tag{150}$$

And by combining the above equations back to Eq. (147), we will see:

$$A_i = \sum_{k=2}^{K} \alpha_k B_{ik}, \tag{151}$$

holds for all $i = 2, \cdots, K$. Therefore we can see its matrix form:

$$\begin{bmatrix} A_2 \\ A_3 \\ \vdots \\ A_K \end{bmatrix} = \begin{bmatrix} B_{22} & B_{23} & \cdots & B_{2K} \\ B_{32} & B_{33} & \cdots & B_{3K} \\ \vdots & \vdots & \ddots & \vdots \\ B_{K2} & B_{K3} & \cdots & B_{KK} \end{bmatrix} \begin{bmatrix} \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_K \end{bmatrix}, \tag{152}$$

which can be solved by the least square method. Then we can compute the canceled out term $\alpha_1 = 1 - \sum_{k=2}^{K} \alpha_k$.

For the special case of two auxiliary variables, we can directly solve the optimality by:

$$\alpha_2^* = \frac{A_2}{B_{22}} = \frac{\int \frac{1}{g(x)} \left[h_2(x) - h_1(x)\right] \left[f(x) - F \cdot h_1(x)\right] \mathbf{d}x}{\int \frac{F}{g(x)} \left[h_2(x) - h_1(x)\right] \left[h_2(x) - h_1(x)\right] \mathbf{d}x} \tag{153}$$

$$\alpha_1^* = 1 - \alpha_2^* = \frac{\int \frac{1}{g(x)} \left[h_2(x) - h_1(x)\right] \left[F \cdot h_2(x) - f(x)\right] \mathbf{d}x}{\int \frac{F}{g(x)} \left[h_2(x) - h_1(x)\right] \left[h_2(x) - h_1(x)\right] \mathbf{d}x} \tag{154}$$

□

A.13 Approx Optimal Deterministic Mixture Multiple Ratio Estimator

Proof. For the Deterministic Mixture Multiple Ratio Estimator

$$\sum_{k=1}^{K} \frac{\sum_{i=1}^{N} w_k(X_i) f(X_i) H_1}{\sum_{i=1}^{N} h_k(X_i)} \tag{155}$$

The Approximate MSE. With approximation mentioned above:

$$\mathrm{MSE}(\hat{F}_{DM}^{Ratio}) \approx \sum_{k=1}^{K} \int \frac{\left(w_k(x) f(x) - h_k(x) \cdot \int w_k(x) f(x) \mathbf{d}x\right)^2}{g(x)} \mathbf{d}x \tag{156}$$

Optimal $w_k(x)$ *minimizing Approximate MSE.* The minimizing weighting functions $w_k(x)$ can be found via minimizing Lagrangian:

$$\mathbf{L} = \mathrm{MSE}(\hat{F}_{DM}^{Ratio}) - \int \lambda(x) \left(\sum_{k=1}^{K} w_k(x) - 1 \right) \mathbf{d}x. \tag{157}$$

Taking partial functional derivatives to zero:

$$\frac{2f^2(x)w_i(x)}{g(x)} + \frac{2f(x)h_k^2(x) \int w_i(x)f(x)\mathbf{d}x}{g(x)} \tag{158}$$

$$- 2\left(f(x)h_k(x) \int w_i(x)f(x)\mathbf{d}x \right) \tag{159}$$

$$- 2f(x) \int w_i(x)f(x)h_i(x)\mathbf{d}x - \lambda = 0 \tag{160}$$

However, unfortunately, we find that solving this expression symbolically is challenging.

□