

Generalizing self-normalized importance sampling with couplings

Nicola Branchini

and

Víctor Elvira

School of Mathematics, University of Edinburgh, UK

June 28, 2024

Abstract

An essential problem in statistics and machine learning is the estimation of expectations involving probability density functions (PDFs) with *intractable normalizing constants*. The self-normalized importance sampling (SNIS) estimator, which normalizes the importance sampling (IS) weights, has become the standard approach due to its simplicity. However, the SNIS has been shown to exhibit high variance in challenging estimation problems, e.g, involving rare events or posterior predictive distributions in Bayesian statistics. Further, most of the state-of-the-art adaptive importance sampling (AIS) methods adapt the proposal as if the weights had not been normalized.

In this paper, we propose a framework that considers the original task as estimation of a ratio of two integrals. In our new formulation, we obtain samples from a joint proposal distribution in an extended space, with two of its marginals playing the role of proposals used to estimate each integral. Importantly, the framework allows us to induce and control a dependency between both estimators. We propose a construction of the joint proposal that decomposes in two (multivariate) marginals and a coupling. This leads to a two-stage framework suitable to be integrated with existing or new AIS and/or variational inference (VI) algorithms. The marginals are adapted in the first stage, while the coupling can be chosen and adapted in the second stage. We show in several realistic examples the benefits of the proposed methodology, including an application to Bayesian prediction with misspecified models.

Keywords: Importance sampling, Monte Carlo, Bayesian inference, coupling

1 Introduction

Importance sampling (IS) is a methodological framework that generalizes the Monte Carlo (MC) principle to approximate integrals [32, 1, 29]. These integrals often take the form of an expectation under a target probability density function (PDF). In IS, integrals are approximated by weighted sums using samples generated by a so-called proposal PDF. Within statistics, IS has found numerous applications to Bayesian computation where it has demonstrated to achieve state-of-the-art performance in combination with Markov chain Monte Carlo (MCMC) and/or neural networks [55, 44, 59, 84, 81] for adapting the proposal or in combination with annealing [60, 25].

In several applications, such as the estimation of Bayesian posterior predictives [14, 83], conformal Bayesian inference in probabilistic programming [71], rare event estimation problems [68] (among others) plain MC estimators (without IS) result in unacceptably large variance and variance reduction is required. The de facto standard approach in these problems is to use the self-normalized importance sampling (SNIS) estimator, which normalizes the IS weights.¹ Currently, the most widely used adaptive importance sampling (AIS) schemes adapt the proposals without considering that the SNIS estimator will be used, e.g., this is the case in population Monte Carlo algorithms [8, 30], adaptive multiple IS (AMIS) or moment-matching schemes [17, 26], and other optimization-based methods [75, 2, 28] (see [7] for a review). More precisely, these AIS algorithms do not take into account the normalization of the weights and design proposals that are close to the target PDF. As we will explain, this is a limitation when the SNIS estimator is the only alternative.

In this work, we develop a new framework that generalizes the celebrated SNIS estimator. The framework provides guidance, in a very general setting, for the minimization of the asymptotic variance of an estimator of expectations involving an unnormalized PDFs. We exploit the interpretation of the estimation problem as *estimating a ratio of two integrals*. The new framework

¹Other names for the SNIS estimator include ratio estimator [46, 63, 20], autonormalized estimator [1], or Hajek estimator [49].

contains as particular cases several previous works that use two different proposals to estimate the two integrals [42, 41, 52, 69, 66, 18]. Also, new valid alternatives appear naturally. A more detailed account of our contributions is given below.

Contributions.

- We show that the asymptotic variance of the SNIS estimator admits an interpretable decomposition with three terms by exploiting a perspective that views the estimator as implicitly estimating a ratio of two integrals. The third term, which has received less attention, can be exploited for variance reduction according to the dependence between the numerator and denominator estimators.
- We exploit the decomposition above to develop a generalization of the SNIS estimator that uses a joint distribution in an extended space to parameterize the dependence between the numerator and denominator estimators. The formulation puts under a unifying framework and considerably generalizes works using two proposals for ratio estimation using IS, which either (i) did not introduce dependence, (ii) were restricted to certain parametric families of marginals, or (iii) used restricted types of dependence.
- We develop a construction of the joint distribution by decomposing it into two multivariate marginals, which play the role of interpreted proposals, and a coupling, defined in this paper as a certain generalization of the notion of copula [19]. The generality of our approach is supported by the theory of vector copulas [33]. We open up a new framework to design proposal distributions in IS, where adaptation can be carried out both for marginals and for the coupling.
- We demonstrate the usefulness of our framework on synthetic examples where the asymptotic variance is closed form and in the estimation of the posterior predictive of Bayesian models when outliers or misspecification arise at test time.

Structure of the paper. Our methodological contribution starts in §3. In §2, we explain the perspective on the problem as ratio estimation and the implicit assumptions behind the SNIS estimator with their consequences, then summarize its limitations. We present in §3.1 our framework for understanding and reducing the asymptotic variance of $\hat{\mu}_{\text{SNIS}}^N$. We derive a decomposition explaining its limitations and introduce a new framework with samples from a joint distribution in an extended space. We define optimal joint distributions and propose a two-stage optimization strategy to adapt the marginals using any existing AIS method. We propose expressive joint distributions with different parametric marginals (§3.2) and propose parameterization and optimization methods §3.4. We develop theoretical guarantees (§3.6), and we unify previous methods as special cases of our framework (§3.5), connecting with related literature (§3.7). Finally, in §4 we demonstrate the framework on a synthetic example with closed form asymptotic variance and an application to misspecified Bayesian prediction. **Notation and standing assumptions.** The considered probability measures are assumed to be absolutely continuous w.r.t. the Lebesgue measure on \mathbb{R}^{d_x} with smooth and strictly positive PDFs, unless otherwise stated. We denote the set of such PDFs over \mathbb{R}^{d_x} with $\mathcal{P}(\mathbb{R}^{d_x})$, while the set of distributions as $\mathbb{P}(\mathbb{R}^{d_x})$. We denote probability measures (e.g., \mathbb{Q}) with blackboard bold, its PDFs (e.g., q) with lowercase letters, and its cumulative distribution function (CDF) (e.g., Q) with uppercase letters. Pushforwards measures use the notation $\mathbf{T}\sharp \mathbb{Q} \in \mathbb{P}(\mathbb{R}^{d_x})$ where \mathbf{T} is the map applied on the measurable (Borel) sets of \mathbb{Q} .

2 Estimating expectations with unnormalized PDFs

In this section, we present background on the SNIS estimator and discuss its interpretation as a ratio estimator (§2.1); we survey existing theoretical analyses of its mean squared error (MSE) (§2.2) and summarize its limitations (§2.3).

2.1 Unnormalized and self-normalized importance sampling

We are interested in the general problem of approximating an expectation w.r.t. a PDF $p(\mathbf{x})$ on $\mathbf{x} \in \mathbb{R}^{d_x}$, $d_x \in \mathbb{N}$ of a measurable, possibly unbounded test function $f : \mathbb{R}^{d_x} \rightarrow \mathbb{R}_{\geq 0}$ such that $\int f(\mathbf{x})^2 d\mathbf{x} < \infty$. We denote the value of the expectation as $\mu \in \mathbb{R}_{\geq 0}$, that is

$$\mu \stackrel{\text{def}}{=} \mathbb{E}_p[f(\mathbf{x})] = \int f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (1)$$

The integral can be approximated by the MC estimator

$$\hat{\mu}_{\text{MC}}^N \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^N f(\mathbf{x}^{(n)}), \quad \{\mathbf{x}^{(n)}\}_{n=1}^N \stackrel{\text{i.i.d.}}{\sim} p(\mathbf{x}), \quad (2)$$

which is unbiased, consistent, asymptotically normal, and asymptotic confidence intervals can be constructed easily. We evaluate performance by the MSE, i.e., $\mathbb{E}_p[(\mu - \hat{\mu}_{\text{MC}}^N)^2]$, as it is standard in the MC literature [63]. Next, we review the unnormalized IS estimator, which, importantly, can only be used when the normalizing constant of p is known.

Unnormalized Importance Sampling. Importance sampling generalizes the principle of MC integration, allowing us to sample from an arbitrary PDF $q(\mathbf{x})$, called proposal. The unnormalized importance sampling (UIS) estimator can be seen as a generalization of the MC estimator $\hat{\mu}_{\text{MC}}^N$, and is given by

$$\hat{\mu}_{\text{UIS}}^N \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^N \underbrace{\frac{p(\mathbf{x}^{(n)})}{q(\mathbf{x}^{(n)})}}_{w^{(n)}} f(\mathbf{x}^{(n)}), \quad \{\mathbf{x}^{(n)}\}_{n=1}^N \stackrel{\text{i.i.d.}}{\sim} q(\mathbf{x}), \quad (3)$$

where $w^{(n)}$ are known as importance weights (or ratios). The MC estimator $\hat{\mu}_{\text{MC}}^N$ is a particular case with $q(\mathbf{x}) = p(\mathbf{x})$. The MSE of the UIS estimator takes an interpretable form, as

$$\text{MSE}_q(\mu, \hat{\mu}_{\text{UIS}}^N) \stackrel{\text{def}}{=} \mathbb{E}_q[(\mu - \hat{\mu}_{\text{UIS}}^N)^2] = \mathbb{V}_q[\hat{\mu}_{\text{UIS}}^N] = \frac{\mu^2}{N} \chi^2(\mu^{-1} p(\mathbf{x}) f(\mathbf{x}) \parallel q(\mathbf{x})), \quad (4)$$

where $\chi^2(p_1(\mathbf{x}) \parallel p_2(\mathbf{x})) = \int p_1(\mathbf{x})^2 / p_2(\mathbf{x}) d\mathbf{x} - 1$ denotes the Pearson chi-squared divergence [76]. An immediate consequence of Eq. (4) is that the UIS estimator improves over the MC estimator, i.e., $\text{MSE}(\mu, \hat{\mu}_{\text{UIS}}^N) < \text{MSE}(\mu, \hat{\mu}_{\text{MC}}^N)$ exactly when $\chi^2(\mu^{-1} p(\mathbf{x}) f(\mathbf{x}) \parallel q(\mathbf{x})) < \chi^2(\mu^{-1} p(\mathbf{x}) f(\mathbf{x}) \parallel p(\mathbf{x}))$.

For example, when $p(\mathbf{x})$ is a Bayesian posterior, this implies that sampling from the true posterior does not minimize MSE of posterior estimates. The relationship between the MSE of the estimator and a statistical divergence involving the proposal q is recognized in the IS literature as a useful guide for the adaptation of proposals [61, 3, 43] with clear connections to variational inference (VI). A reformulation of Eq. (4) as an optimization problem over $\mathcal{P}(\mathbb{R}^{d_{\mathbf{x}}})$ leads to a natural definition of optimal proposal.

Definition 1 (Optimal UIS proposal). *The optimal UIS proposal PDF is defined as the solution to the following optimization problem in the infinite-dimensional space of PDFs,*

$$q_{UIS}^{\star}(\mathbf{x}) = \arg \min_{q \in \mathcal{P}(\mathbb{R}^{d_{\mathbf{x}}})} \text{MSE}_q(\mu, \hat{\mu}_{UIS}^N) = \arg \min_{q \in \mathcal{P}(\mathbb{R}^{d_{\mathbf{x}}})} \chi^2(\mu^{-1}p(\mathbf{x})f(\mathbf{x}) \parallel q(\mathbf{x})) = \mu^{-1}p(\mathbf{x})f(\mathbf{x}).$$

While $\text{MSE}_{q_{UIS}^{\star}}(\mu, \hat{\mu}_{UIS}^N) = 0$, we can never use q_{UIS}^{\star} , because it requires to use μ , which is unknown by definition. The most popular IS algorithms choose a parametric family of PDFs $\mathcal{Q}_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \{q(\mathbf{x}; \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \Theta} \subset \mathcal{P}(\mathbb{R}^{d_{\mathbf{x}}})$ indexed by parameters $\boldsymbol{\theta}$ for one or several proposals and adapt them iteratively with AIS. See [7] and [63, Chapter 10] for thorough overviews of AIS and [27] for a framework with multiple proposals.

Therefore, a more realistic notion of optimal proposal is as follows.

Definition 2 (Parametric optimal UIS proposal). *The parametric optimal UIS proposal PDF for a parametric class of proposals indexed by $\boldsymbol{\theta}$, $\mathcal{Q}_{\boldsymbol{\theta}} = \{q(\mathbf{x}; \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \Theta} \subset \mathcal{P}(\mathbb{R}^{d_{\mathbf{x}}})$, is defined as*

$$q_{UIS}^{\boldsymbol{\theta}^{\star}}(\mathbf{x}) = \arg \min_{q \in \mathcal{Q}_{\boldsymbol{\theta}}} \chi^2(q_{UIS}^{\star}(\mathbf{x}) \parallel q(\mathbf{x}; \boldsymbol{\theta})) \iff \boldsymbol{\theta}^{\star} = \arg \min_{\boldsymbol{\theta} \in \Theta} \chi^2(q_{UIS}^{\star}(\mathbf{x}) \parallel q(\mathbf{x}; \boldsymbol{\theta})).$$

Importantly, if $q_{UIS}^{\star} \notin \mathcal{Q}_{\boldsymbol{\theta}}$, i.e., there is no value of $\boldsymbol{\theta}$ for which $q(\mathbf{x}; \boldsymbol{\theta}) = q_{UIS}^{\star}(\mathbf{x})$, then $\text{MSE}_{q_{UIS}^{\boldsymbol{\theta}^{\star}}}(\mu, \hat{\mu}_{UIS}^N) > 0 \Rightarrow \chi^2(\mu^{-1}p(\mathbf{x})f(\mathbf{x}) \parallel q(\mathbf{x}; \boldsymbol{\theta}^{\star})) > 0$. As we see in §3, this fact will be an important part of the motivation behind our generalization of $\hat{\mu}_{SNIS}^N$.

Self-normalized importance sampling. We are interested in problems where the normalizing constant of $p(\mathbf{x})$ is unknown so we cannot use the UIS estimator $\hat{\mu}_{UIS}^N$. In these cases, the SNIS

estimator is the standard approach. We now present SNIS as it is commonly derived and highlight the implicit assumptions behind its construction. The SNIS estimator can be obtained by replacing the IS weights in the UIS estimator from Eq. (3) with their normalized counterparts, $\bar{w}^{(n)} \stackrel{\text{def}}{=} \frac{w^{(n)}}{\sum_{i=1}^N w^{(i)}}, n = 1, \dots, N$, leading to the common expression

$$\hat{\mu}_{\text{SNIS}}^N \stackrel{\text{def}}{=} \sum_{n=1}^N \bar{w}^{(n)} f(\mathbf{x}^{(n)}), \quad \{\mathbf{x}^{(n)}\}_{n=1}^N \stackrel{\text{i.i.d.}}{\sim} q(\mathbf{x}). \quad (5)$$

While $\hat{\mu}_{\text{SNIS}}^N$ is consistent just as $\hat{\mu}_{\text{UIS}}^N$ and asymptotically normal [38], characterizing its MSE is more complicated, as we see next.

Implicit construction behind the SNIS estimator. While the formulation of Eq. (5) is widespread and clearly explains how to implement the estimator, it hides that $\hat{\mu}_{\text{SNIS}}^N$ is implicitly estimating two integrals and combining their estimates in a ratio. Indeed, by defining the normalizing constant of p as $Z_p \stackrel{\text{def}}{=} \int \tilde{p}(\mathbf{x}) d\mathbf{x}$ and so $p(\mathbf{x}) = \tilde{p}(\mathbf{x})/Z_p$, we see that μ can be equivalently expressed as a ratio of two integrals

$$\mu = \int f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \frac{\int f(\mathbf{x}) \tilde{p}(\mathbf{x}) d\mathbf{x}}{\int \tilde{p}(\mathbf{x}) d\mathbf{x}} = \frac{I}{Z_p}, \quad (6)$$

where we have denoted the integral in the numerator as I . The observation leads to a more informative expression for $\hat{\mu}_{\text{SNIS}}^N$. By defining two UIS estimators constructed with a shared set of samples $\{\mathbf{x}^{(n)}\}_{n=1}^N \stackrel{\text{i.i.d.}}{\sim} q(\mathbf{x})$, i.e., $\hat{I}^N \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^N \frac{f(\mathbf{x}^{(n)}) \tilde{p}(\mathbf{x}^{(n)})}{q(\mathbf{x}^{(n)})}$ and $\hat{Z}_p^N \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^N \frac{\tilde{p}(\mathbf{x}^{(n)})}{q(\mathbf{x}^{(n)})}$, we can express the SNIS estimator as a ratio of the two UIS estimators as

$$\hat{\mu}_{\text{SNIS}}^N = \frac{\hat{I}^N}{\hat{Z}_p^N} = \frac{\frac{1}{Z_p} \frac{1}{N} \sum_{n=1}^N \frac{f(\mathbf{x}^{(n)}) \tilde{p}(\mathbf{x}^{(n)})}{q(\mathbf{x}^{(n)})}}{\frac{1}{Z_p} \frac{1}{N} \sum_{n=1}^N \frac{\tilde{p}(\mathbf{x}^{(n)})}{q(\mathbf{x}^{(n)})}} = \frac{\sum_{n=1}^N \frac{f(\mathbf{x}^{(n)}) \tilde{p}(\mathbf{x}^{(n)})}{q(\mathbf{x}^{(n)})}}{\sum_{n=1}^N \frac{\tilde{p}(\mathbf{x}^{(n)})}{q(\mathbf{x}^{(n)})}}. \quad (7)$$

Notice that \hat{Z}_p^N is the denominator of $\bar{w}^{(n)}$, except for the factor $1/N$ which cancels in the normalization. While Eq. (7) is mathematically equivalent to Eq. (5), it shows explicitly the ratio of the two UIS estimators, \hat{I}^N and \hat{Z}_p^N , which share the same set of samples $\{\mathbf{x}^{(n)}\}_{n=1}^N$. As in any IS method, the key challenge lies in selecting a good proposal q . Here we take the MSE as the main metric of interest instead other approaches that consider the effective sample size (ESS) (see [31] for a discussion). With the UIS estimator, it is at least theoretically possible to obtain zero

MSE as a consequence of Eq. (4) if $q_{\text{UIS}}^{\star} \in \mathcal{Q}_{\theta}$. Unfortunately, it is not possible to define a similar optimal proposal for the SNIS estimator to achieve zero MSE, as we see next.

2.2 Analysis of the MSE and optimal SNIS proposal

We summarize the difficulties in studying the MSE of $\hat{\mu}_{\text{SNIS}}^N$, which translates into the impossibility of deriving an exact expression for the proposal q that minimizes the MSE and the need to resort to asymptotic variance. Indeed, since $\hat{\mu}_{\text{SNIS}}^N$ is a ratio, the MSE involves terms such as $\mathbb{E}_q \left[\frac{\hat{I}^N}{\hat{Z}_p^N} \right]$. In general, $\hat{\mu}_{\text{SNIS}}^N$ is not unbiased since $\mathbb{E}_q \left[\frac{\hat{I}^N}{\hat{Z}_p^N} \right] \neq \frac{\mathbb{E}_q[\hat{I}^N]}{\mathbb{E}_q[\hat{Z}_p^N]} = \mu$. The variance term is similarly problematic to analyse.

In this work, we follow the standard practice of previous works that studied or used SNIS [46, 63, 16, 24, 78] who consider the asymptotic variance, defined by

$$\mathbb{V}_q^{\infty}[\hat{\mu}_{\text{SNIS}}^N] \stackrel{\text{def}}{=} \lim_{N \rightarrow +\infty} N \cdot \mathbb{V}_q[\hat{\mu}_{\text{SNIS}}^N], \quad (8)$$

as the proxy for the performance of $\hat{\mu}_{\text{SNIS}}^N$. Hesterberg [46, Chapter 2, Sec. 9] first showed the analytic expression for the proposal PDF that minimizes $\mathbb{V}_q^{\infty}[\hat{\mu}_{\text{SNIS}}^N]$, as follows.

Definition 3 (Asymptotically optimal SNIS proposal). *The asymptotically optimal proposal PDF for $\hat{\mu}_{\text{SNIS}}^N$ has an analytic solution and is defined as the minimizer (when it exists) of the asymptotic variance, as*

$$q_{\text{SNIS}}^{\star}(\mathbf{x}) \stackrel{\text{def}}{=} \arg \min_{q \in \mathcal{P}(\mathbb{R}^{d_x})} \mathbb{V}_q^{\infty}[\hat{\mu}_{\text{SNIS}}^N] = \frac{p(\mathbf{x})|f(\mathbf{x}) - \mu|}{\int p(\mathbf{x})|f(\mathbf{x}) - \mu| d\mathbf{x}}. \quad (9)$$

As we shall see in §2.3, $q_{\text{SNIS}}^{\star}(\mathbf{x})$ does not lead to zero variance, in sharp contrast with the optimal UIS proposal of Definition 1.

2.3 Limitations of the SNIS estimator

We start our discussion of the fundamental limitations of $\hat{\mu}_{\text{SNIS}}^N$ with a motivating example.

Motivating example (Out-of-sample predictive accuracy of Bayesian high-dimensional models with outliers/misspecification). Let $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{d_{\boldsymbol{\theta}}}$ be the parameter of interest of a statistical model (family of PDFs) $\{g(y; \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \Theta}$ with $y \in \mathbb{R}$, likelihood function $\mathcal{L}(y_{1:S}; \boldsymbol{\theta})$ and prior PDF $\pi(\boldsymbol{\theta})$, so that the posterior distribution has PDF $\pi(\boldsymbol{\theta}|y_{1:S}) = Z_{\pi}(y_{1:S})^{-1}\tilde{\pi}(\boldsymbol{\theta}|y_{1:S})$ where $\tilde{\pi}(\boldsymbol{\theta}|y_{1:S}) = \mathcal{L}(y_{1:S}; \boldsymbol{\theta})\pi(\boldsymbol{\theta})$. Suppose that the true parameter is $\boldsymbol{\theta}_\star$, i.e., $y_1, y_2, \dots, y_S \stackrel{i.i.d.}{\sim} g(y; \boldsymbol{\theta}_\star)$.

In practice, to evaluate the predictive performance, test data is often generated from a different distribution. This can be due to outliers [66] and/or model misspecification, i.e., $y_{S+1} \sim g(y; \boldsymbol{\theta}_{\text{wrong}})$, $\boldsymbol{\theta}_{\text{wrong}} \neq \boldsymbol{\theta}_\star$,² or $y_{S+1} \sim h(y; \boldsymbol{\gamma})$ for a different model $\{h(y; \boldsymbol{\gamma})\}_{\boldsymbol{\gamma} \in \Gamma}$ where there is no $\boldsymbol{\gamma} : h(y; \boldsymbol{\gamma}) = g(y; \boldsymbol{\theta}_\star)$. In both cases, for a fixed y_{S+1} one evaluates the out-of-sample performance by estimating the posterior predictive integral

$$\mu = \mathbb{E}_{\pi(\boldsymbol{\theta}|y_{1:S})}[g(y_{S+1}|\boldsymbol{\theta})] = \int g(y_{S+1}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|y_{1:S})d\boldsymbol{\theta} = \frac{\int g(y_{S+1}|\boldsymbol{\theta})\tilde{\pi}(\boldsymbol{\theta}|y_{1:S})d\boldsymbol{\theta}}{\int \tilde{\pi}(\boldsymbol{\theta}|y_{1:S})d\boldsymbol{\theta}}. \quad (10)$$

Such a quantity is of interest in many cases, such as in conformal Bayesian computation [34]. A possible option for estimation is to sample from $\pi(\boldsymbol{\theta}|y_{1:S})$ (e.g., with MCMC),³ but is inefficient since the variance of $\hat{\mu}_{MC}^N$ is $\propto \chi^2(\pi(\boldsymbol{\theta}|y_{1:S+1})||\pi(\boldsymbol{\theta}|y_{1:S}))$ (discussed in §2.1). This variance is large since outliers lead to large divergences [67]. We may instead turn to SNIS with a proposal $q(\boldsymbol{\theta})$ [83], whose asymptotic variance is $\propto \chi^2(\pi(\boldsymbol{\theta}|y_{1:S+1})||q) + \chi^2(\pi(\boldsymbol{\theta}|y_{1:S})||q) - 2 \cdot \mathcal{C}(q)$ (as we show in §3). In this context, we want to choose a q that is close to the posteriors $\pi(\boldsymbol{\theta}|y_{1:S+1})$ and $\pi(\boldsymbol{\theta}|y_{1:S})$, and that also maximizes the third term, $\mathcal{C}(q)$. For realistic posterior distributions, finding such a q is hard. We will propose a general methodology that allows us to construct a proposal that can match the two target posteriors and, at the same time, control the third term, significantly generalizing several previous works (see §3.5 and §3.7).

Asymptotic variance lower bound. As noted in [65, 63], the asymptotic variance obtained by $q_{\text{SNIS}}^\star(\boldsymbol{x})$ from Eq. (9) is positive and its expression can be derived analytically. For any $q \in \mathcal{P}(\mathbb{R}^{d_{\boldsymbol{x}}})$,

²The following discussion can be easily generalized for multiple test observations y_{S+1}, y_{S+2}, \dots , and multivariate observations.

³The above problem was considered also in the related work of target-aware Bayesian inference (TABI) [69]

it holds

$$\mathbb{V}_q^\infty[\hat{\mu}_{\text{SNIS}}^N] \geq \mathbb{V}_{q_{\text{SNIS}}^\star}^\infty[\hat{\mu}_{\text{SNIS}}^N] = (\mathbb{E}_p[|f(\mathbf{x}) - \mu|])^2 > 0 \quad (11)$$

with equality being possible only if $q = q_{\text{SNIS}}^\star$. Rainforth et al. [69], who also proposes an estimator to improve over $\hat{\mu}_{\text{SNIS}}^N$ (see §3.5), finds empirically that this lower bound leads to high variance. Further, [64] proved that the bound prevents a certain notion of exponential convergence in adaptive IS. Therefore, while with the UIS estimator, we can drive the MSE to zero with ever better choices of q , the performance of $\hat{\mu}_{\text{SNIS}}^N$ has fundamental theoretical limitations. In practice, the situation is even more challenging. Indeed, even attempting to simulate with MCMC from q_{SNIS}^\star would require further research, as the unnormalized target PDF involves μ . We now introduce our framework which will allow us to better understand the celebrated SNIS estimator, generalize it to reduce variance, and guide the design of new AIS algorithms.

3 Generalized SNIS

We now present our framework, which allows us to understand when $\hat{\mu}_{\text{SNIS}}^N$ has large asymptotic variance, and how to reduce it. First, we derive a decomposition of the asymptotic variance that intuitively explains the limitations of $\hat{\mu}_{\text{SNIS}}^N$. In §3.1 we introduce our new framework, where samples are obtained from a joint distribution in an extended space. We propose suitable definitions of optimal joint distributions and a two-stage optimization strategy that allows us to combine any existing AIS method to adapt the marginals of the joint. In §3.2 we show how expressive joint distributions, possibly semi-parametric, with marginals of different parametric families can be used. Then, in §3.3 and §3.4 we propose concrete ways to parameterize and optimize the joint. Finally, we develop theoretical guarantees in §3.6 and in §3.5 we unify some previous methods as special cases of our framework, also connecting more broadly with related literature in §3.7.

To introduce our decomposition of the asymptotic variance, recall from §2.1 that $\hat{\mu}_{\text{SNIS}}^N$ is a ratio of two UIS estimators.

Proposition 1. We let the optimal UIS proposals for numerator and denominator of $\mu = I/Z_p$, respectively as $q_{1,UIS}^\star(\mathbf{x}) = I^{-1}\tilde{p}(\mathbf{x}) \cdot f(\mathbf{x})$ and $q_{2,UIS}^\star(\mathbf{x}) = p(\mathbf{x})$. The relative asymptotic variance $\mathbb{V}_q^\infty[\hat{\mu}_{SNIS}^N]$ is given by

$$\mu^{-2} \cdot \mathbb{V}_q^\infty[\hat{\mu}_{SNIS}^N] = \chi^2(q_1^\star || q) + \chi^2(q_2^\star || q) - 2 \left(\overbrace{\mathbb{E}_q \left[\frac{q_1^\star(\mathbf{x})q_2^\star(\mathbf{x})}{q(\mathbf{x})^2} \right]}^{\mathcal{C}(q)} - 1 \right), \quad (12)$$

where we omit from now on the UIS subscripts. We remark that q_1^\star includes the test function $f(\mathbf{x})$.

Eq. (12) provides new intuition for how q should be chosen for the problem targeted by SNIS. The first two terms force q to be close to the two optimal UIS proposals q_1^\star, q_2^\star , where the first one depends on the test function $f(\mathbf{x})$. The third term $\mathcal{C}(q)$ has a direct correspondence with $\text{Cov}_q[\hat{I}^N, \hat{Z}_p^N]$ (as we discuss in Appendix F) and can affect the asymptotic variance significantly. For instance, overall zero asymptotic variance is theoretically possible when $\mathcal{C}(q)$ cancels the other two terms (see §3.6). While a few previous works in ratio estimation do consider the use of two proposals q_1, q_2 [42, 52, 41, 66] to (implicitly) minimize the first two terms, they do not take $\mathcal{C}(q)$ explicitly into consideration in the construction of the estimates. In §3.5, we review these works showing their estimators and $\hat{\mu}_{SNIS}^N$ are special cases of our generic framework introduced next. We will show that our new estimator's variance decomposes like Eq. (12) and it will allow us to explicitly control a similar term to $\mathcal{C}(q)$ to reduce variance.

3.1 Introducing an extended space joint proposal

We construct the new estimator by obtaining samples from a joint distribution in an extended space, $\mathbb{Q}_{1:2}(d\mathbf{x}_1, d\mathbf{x}_2) \in \mathbb{P}(\mathbb{R}^{2d_x})$ with marginals $\mathbb{Q}_1(d\mathbf{x}_1), \mathbb{Q}_2(d\mathbf{x}_2)$, as

$$\hat{\mu}_{\text{GenSNIS}}^N \stackrel{\text{def}}{=} \frac{\hat{I}^N}{\hat{Z}_p^N} = \frac{\frac{1}{N} \sum_{n=1}^N \frac{f(\mathbf{x}_1^{(n)})\tilde{p}(\mathbf{x}_1^{(n)})}{q_1(\mathbf{x}_1^{(n)})}}{\frac{1}{N} \sum_{n=1}^N \frac{\tilde{p}(\mathbf{x}_2^{(n)})}{q_2(\mathbf{x}_2^{(n)})}}, \quad [\mathbf{x}_1^{(n)}, \mathbf{x}_2^{(n)}]^\top \stackrel{\text{i.i.d.}}{\sim} \mathbb{Q}_{1:2}(d\mathbf{x}_1, d\mathbf{x}_2), n = 1, \dots, N. \quad (13)$$

The key element of our new methodology is the use of a generic joint distribution $\mathbb{Q}_{1:2}$, and its construction (which we detail in §3.2). Marginals of $\mathbb{Q}_{1:2}$ need to admit PDFs q_1 and q_2 to compute

the estimator. Yet, since the joint $\mathbb{Q}_{1:2}$ does not appear in the expression of $\hat{\mu}_{\text{GenSNIS}}^N$, we do not require its PDF to exist. Exploiting this, we will formalize a semi-parametric model for the joint, which allows us to interpret $\hat{\mu}_{\text{SNIS}}^N$ more formally as a special case, as well as other recent methods using two proposals (see §3.5).

Now we turn to the key question of how to design $\mathbb{Q}_{1:2}$. We extend the asymptotic variance decomposition of Eq. (12) to $\hat{\mu}_{\text{GenSNIS}}^N$ to obtain an objective function for the definition of an optimal joint in terms of asymptotic variance.

Proposition 2 (Asymptotic variance of $\hat{\mu}_{\text{GenSNIS}}^N$). *The asymptotic variance of $\hat{\mu}_{\text{GenSNIS}}^N$ satisfies (proof in Appendix G)*

$$\mu^{-2} \cdot \mathbb{V}_{\mathbb{Q}_{1:2}}^\infty [\hat{\mu}_{\text{GenSNIS}}^N] = \chi^2(q_1^\star || q_1) + \chi^2(q_2^\star || q_2) - 2 \left(\overbrace{\mathbb{E}_{\mathbb{Q}_{1:2}} \left[\frac{q_1^\star(\mathbf{x}_1)q_2^\star(\mathbf{x}_2)}{q_1(\mathbf{x}_1)q_2(\mathbf{x}_2)} \right]}^{\mathcal{C}(\mathbb{Q}_{1:2})} - 1 \right). \quad (14)$$

The decomposition in Eq. (14) allows us the following strategy: we design a joint such that its marginal q_1 minimizes the first term $\chi^2(q_1^\star || q_1)$, its marginal q_2 minimizes the second term $\chi^2(q_2^\star || q_2)$, and the third term $\mathcal{C}(\mathbb{Q}_{1:2})$ is maximized.

Remark 1 (A general framework for combining adaptive IS or VI algorithms). *By separating the adaptation of the marginals q_1 , q_2 from that of the full joint $\mathbb{Q}_{1:2}$, we allow a new framework for adaptive IS algorithms where a first run uses an arbitrary adaptive IS algorithm, or VI, to obtain q_1, q_2 .⁴ Then, one optimizes the joint to minimize asymptotic variance by maximizing $\mathcal{C}(\mathbb{Q}_{1:2})$. We will show in §3.2 that expressive families of joints and marginals can be used in practice within the framework, such as mixtures of Gaussians or Student-t distributions.*

Towards defining an optimal joint. Before optimizing $\mathbb{Q}_{1:2}$, we need to define an asymptotically optimal joint, similarly to Definition 3. To do so, we define a suitable set of joint distributions to search over.

⁴Any algorithm that is optimized, implicitly or explicitly, to minimize a χ^2 divergence to solve the problem Definition 1 and returns PDFs is suitable.

We define the set of all joint distributions with fixed marginals $\mathbb{Q}_1 \in \mathcal{Q}_{\theta_1}$, $\mathbb{Q}_2 \in \mathcal{Q}_{\theta_2}$, as

$$\Pi(\mathcal{Q}_{\theta_1}, \mathcal{Q}_{\theta_2}) \stackrel{\text{def}}{=} \left\{ \mathbb{Q}_{1:2} \in \mathbb{P}(\mathbb{R}^{2d_x}) \mid \int_{\mathbf{x}_2} \mathbb{Q}_{1:2}(d\mathbf{x}_{1:2}) = \mathbb{Q}_1, \int_{\mathbf{x}_1} \mathbb{Q}_{1:2}(d\mathbf{x}_{1:2}) = \mathbb{Q}_2 \right\}, \quad (15)$$

where $\mathbf{x}_{1:2} = [\mathbf{x}_1, \mathbf{x}_2]^\top$. Eq. (15) specifies a *semi-parametric* model, as we have not yet specified a parameterization for the full joint $\mathbb{Q}_{1:2}$, only for two of its multivariate marginals, leading to the following definition.

Definition 4 (Asymptotically optimal joint for $\hat{\mu}_{\text{GenSNIS}}^N$ for two given marginals). *Let q_1 and q_2 be any two given marginal PDFs, $q_1 \neq q_1^*$, $q_2 \neq q_2^*$. We define the asymptotically optimal joint proposal $\mathbb{Q}_{1:2}^*(d\mathbf{x}_{1:2})$ for $\hat{\mu}_{\text{GenSNIS}}^N$ as the solution to the following infinite-dimensional constrained optimization problem that maximizes $\mathcal{C}(\mathbb{Q}_{1:2})$, as*

$$\mathbb{Q}_{1:2}^*(d\mathbf{x}_{1:2}) \stackrel{\text{def}}{=} \arg \max_{\mathbb{Q}_{1:2} \in \Pi(\mathcal{Q}_{\theta_1}, \mathcal{Q}_{\theta_2})} \mathbb{E}_{\mathbb{Q}_{1:2}} \left[\frac{q_1^*(\mathbf{x}_1)q_2^*(\mathbf{x}_2)}{q_1(\mathbf{x}_1)q_2(\mathbf{x}_2)} \right]. \quad (16)$$

Note that if we used q_1^* and q_2^* instead of q_1 and q_2 , there would be no benefit in introducing a joint, since for any $\mathbb{Q}_{1:2}$ it holds that $\hat{\mu}_{\text{GenSNIS}}^N = \mu$ almost surely. As discussed in §2.1, we can never use q_1^* and q_2^* in practice.

Eq. (16) defines a challenging optimization problem with constraints that are difficult to enforce, in general. There is a clear connection to an unusual optimal transport problem, which could lead to interesting future research; see Appendix F. Next, we consider restricting the search over parametric joint families. Consider the following parameterized subset of $\Pi(\mathcal{Q}_{\theta_1}, \mathcal{Q}_{\theta_2})$ as defined in Eq. (15), indexed by $\boldsymbol{\theta} \in \Theta$, as

$$\Pi_{\boldsymbol{\theta}}(\mathcal{Q}_{\theta_1}, \mathcal{Q}_{\theta_2}) \stackrel{\text{def}}{=} \left\{ \mathbb{Q}_{1:2}(d\mathbf{x}_{1:2}; \boldsymbol{\theta}) \mid \int_{\mathbf{x}_2} \mathbb{Q}_{1:2}(d\mathbf{x}_{1:2}; \boldsymbol{\theta}) = \mathbb{Q}_1, \int_{\mathbf{x}_1} \mathbb{Q}_{1:2}(d\mathbf{x}_{1:2}; \boldsymbol{\theta}) = \mathbb{Q}_2 \right\}_{\boldsymbol{\theta} \in \Theta}. \quad (17)$$

The difference with Eq. (15) is that Eq. (17) is a parametric model instead of a semi-parametric one, with parameters $\boldsymbol{\theta}$. Now, the parameters $\boldsymbol{\theta}$ include the parameters of the marginals $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, but also additional parameters in general.

Definition 5 (Asymptotically optimal parametric joint for $\hat{\mu}_{\text{GenSNIS}}^N$). *We define the asymptotically optimal parametric joint, with parameter space $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_c]^\top \in \Theta = \Theta_1 \cup \Theta_2 \cup \Theta_c$, as the joint $\mathbb{Q}_{1:2}(d\mathbf{x}_1, d\mathbf{x}_2; \boldsymbol{\theta}^\star) \in \Pi_{\boldsymbol{\theta}}(\mathcal{Q}_{\boldsymbol{\theta}_1}, \mathcal{Q}_{\boldsymbol{\theta}_2})$ whose parameters are given by the solution to the following constrained optimization problem*

$$\boldsymbol{\theta}^\star \stackrel{\text{def}}{=} \arg \max_{\substack{\boldsymbol{\theta} \in \Theta \\ \mathbb{Q}_{1:2} \in \Pi_{\boldsymbol{\theta}}(\mathcal{Q}_{\boldsymbol{\theta}_1}, \mathcal{Q}_{\boldsymbol{\theta}_2})}} \mathbb{E}_{\mathbb{Q}_{1:2}^{\boldsymbol{\theta}}} \left[\frac{q_1^\star(\mathbf{x}_1) q_2^\star(\mathbf{x}_2)}{q_1(\mathbf{x}_1) q_2(\mathbf{x}_2)} \right]. \quad (18)$$

Similarly to the parametric optimal UIS proposal of Definition 1, in general, the above joint will differ from that of Definition 4. Yet, it is a step towards a practical optimization problem. A remaining challenge, which we address next, is to define suitable parameterizations for the joint which allow us to change the parameters $\boldsymbol{\theta}$ without changing those of the marginals.

Remark 2. Choosing typical families used in IS like multivariate normals (MVNs) imposes considerable restrictions on marginal families. For instance, if we were to consider $\mathbb{Q}_{1:2}(d\mathbf{x}_1, d\mathbf{x}_2; \boldsymbol{\theta})$ to be MVN with $\boldsymbol{\theta} = [\boldsymbol{\mu}, \boldsymbol{\Sigma}]$, i.e., with moment parameters $\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2] \in \mathbb{R}^{2d_{\mathbf{x}}}$, $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_{1 \times 2}^\top \\ \boldsymbol{\Sigma}_{1 \times 2} & \boldsymbol{\Sigma}_2 \end{bmatrix} \in \mathcal{S}_{++}^{2d_{\mathbf{x}} \times 2d_{\mathbf{x}}}$, then \mathbb{Q}_1 and \mathbb{Q}_2 are restricted to be MVNs by definition. The same restriction applies to the multivariate Student-t family [35]. The need for two different parametric families can easily arise in practice. Consider our motivating example from §2.3: if test data is sampled from a Student-t model, with the posterior conditional on previous data being MVN, then the two marginals of the joint should be a heavier-tailed distribution like Student-t and MVN.

Motivated by these limitations, we develop a general method to sample from a joint distribution with two given marginals of (i) possibly different parametric families, and (ii) such that the parametric family (of the joint) is not in a known or named family.

3.2 Constructing the joint with couplings and Brenier maps

Our proposed construction of the joint $\mathbb{Q}_{1:2}$ decomposes it into a coupling (defined next) and its marginals. In §3.6, we develop theoretical support for this construction.

Definition 6 (Coupling). A coupling is a joint probability distribution $\mathbb{C}_{1:2}(d\mathbf{u}_1, d\mathbf{u}_2)$ for random vectors $\mathbf{u}_1, \mathbf{u}_2 \in [0, 1]^{d_x}$ with the constraint that the marginals of \mathbf{u}_1 and \mathbf{u}_2 are uniform, i.e., $\mathbf{u}_1 \sim \mathbb{U}_{[0,1]^{d_x}}(d\mathbf{u}_1), \mathbf{u}_2 \sim \mathbb{U}_{[0,1]^{d_x}}(d\mathbf{u}_2)$. We denote its corresponding PDF, when it exists, as $c_{1:2}(\mathbf{u}_1, \mathbf{u}_2)$.

For $d_x = 1$, the above definition amounts to a copula [19]. Importantly, the notion of copula is *not* sufficient for our purposes, as it provably cannot model multivariate marginals [37].

Sampling from a joint with fixed marginals. We now discuss how to obtain samples from a joint with given multivariate marginals, a nontrivial task in general [48]. Assume that **(i)** we know how to sample from a coupling $\mathbb{C}_{1:2}(d\mathbf{u}_1, d\mathbf{u}_2)$ and **(ii)** have access to measurable functions $\mathbf{T}_i : [0, 1]^{d_x} \rightarrow \mathbb{R}^{d_x}$ that transport the uniform distribution to the desired marginals, i.e., $\mathbf{T}_i \sharp \mathbb{U}_{[0,1]^{d_x}}(d\mathbf{u}_i) = \mathbb{Q}_i(dx_i)$ for $i = 1, 2$. Then, we can obtain a sample $[\mathbf{x}_1, \mathbf{x}_2]^\top$ with the following two steps.

- **Step 1 (Coupling sampling):** Obtain a sample $[\mathbf{u}_1, \mathbf{u}_2] \sim \mathbb{C}_{1:2}(d\mathbf{u}_1, d\mathbf{u}_2)$.
- **Step 2 (Marginal transformation):** Apply the maps \mathbf{T}_i to \mathbf{u}_i to obtain $\mathbf{x}_i = \mathbf{T}_i(\mathbf{u}_i)$, for $i = 1, 2$. Denote the joint distribution of $\mathbf{x}_1, \mathbf{x}_2$ as $\mathbb{Q}_{1:2}(d\mathbf{x}_1, d\mathbf{x}_2)$.

The intuition is that we explore the space of $\mathbb{Q}_{1:2}$ by selecting different couplings and fixing the marginals by use of \mathbf{T}_i . We now detail the two steps. **Step 1 (Coupling sampling).** As a possible way to implement step (1), we exploit the use of an easy-to-deal-with *auxiliary* joint distribution $\mathbb{G}_{1:2}(d\mathbf{x}'_1, d\mathbf{x}'_2)$, e.g., a MVN or a Student-t distribution, by sampling from it $[\mathbf{x}'_1, \mathbf{x}'_2] \sim \mathbb{G}_{1:2}(d\mathbf{x}'_1, d\mathbf{x}'_2)$ and transforming the sample into one from a coupling by applying the inverse maps $\mathbf{T}_i^{-1} : \mathbb{R}^{d_x} \rightarrow [0, 1]^{d_x}$ to transform the marginals $\mathbf{T}_i^{-1} \sharp \mathbb{G}_i(d\mathbf{x}_i) = \mathbb{U}_{[0,1]^{d_x}}(d\mathbf{u}_i)$ for $i = 1, 2$ as before. Concretely, for $d_x = 1$, \mathbf{T}_i can be chosen as the inverse CDFs (quantile) of the marginals. In general, for $d_x > 1$, we propose the use of vector quantiles (VQs) introduced in [15]. Interestingly, VQs coincide exactly the celebrated optimal transport Brenier maps, as noted in [10].⁵ **Step 2**

⁵We note that our coupling sampling method is a generalization of the method described in [63, Chapter 5, Section

(marginal transformation). Brenier maps are defined as any $\mathbf{T}_i \sharp \mathbb{U}_{[0,1]^{d_{\mathbf{x}}}}(d\mathbf{u}_i) = \mathbb{Q}(d\mathbf{x}_i)$ that can be expressed as the gradient of a convex function, i.e., there exists a convex $\psi(\mathbf{u}_i) : [0, 1]^{d_{\mathbf{x}}} \rightarrow \mathbb{R} \cup \{+\infty\}$ such that $\mathbf{T}_i \stackrel{\text{def}}{=} \nabla_{\mathbf{u}_i} \psi(\mathbf{u}_i)$. While the definition is abstract, [33] shows that they can be constructively found for interesting distributions commonly used in IS, e.g., for MVN, Student-t marginals, and more. Notably, Brenier maps are diffeomorphisms, i.e., continuously differentiable with a continuously differentiable inverse, so that we have the (required) PDFs for $\mathbb{Q}_1, \mathbb{Q}_2$.

We give an illustration of the process of sampling from a joint with fixed marginals in Fig. 1. If the coupling also admits a PDF $c_{1:2}(\mathbf{u}_1, \mathbf{u}_2)$ (e.g., when using the method described in step (1)), then the joint PDF of $\mathbf{x}_1, \mathbf{x}_2$ is given by

$$q_{1:2}(\mathbf{x}_1, \mathbf{x}_2; \boldsymbol{\theta}) = q_1(\mathbf{x}_1; \boldsymbol{\theta}_1)q_2(\mathbf{x}_2; \boldsymbol{\theta}_2) c_{1:2}(\mathbf{T}_1^{-1}(\mathbf{x}_1), \mathbf{T}_2^{-1}(\mathbf{x}_2); \boldsymbol{\theta}_c). \quad (19)$$

We stress that in general, to compute the estimator $\hat{\mu}_{\text{GenSNIS}}^N$, the coupling (hence, the joint distribution) does not itself require a PDF. Indeed, some relevant couplings that we incorporate in our framework do not admit a PDF.

Remark 3. Our framework also applies for couplings that do not admit PDFs, which are common in the variance reduction literature (see, e.g., [63, Chapter 5]. We expand this in §3.5.

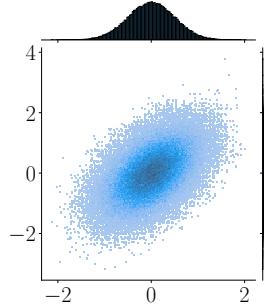
3.3 From optimal joint to optimal coupling

We are now equipped to rewrite Eq. (18) to obtain an optimization problem directly in terms of the coupling $\mathbb{C}_{1:2}$, taking advantage of the joint construction just developed. This allows us to make the optimization problem of Eq. (16) easier, as it will be constrained by construction.

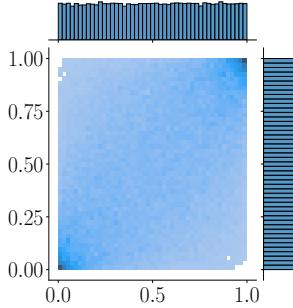
Rewriting the objective in Eq. (16) by applying the change of variables $\mathbf{x}_i \rightarrow \mathbf{T}_i^{-1}(\mathbf{x}_i) = \mathbf{u}_i$ for $i = 1, 2$ leads to

$$\mathbb{E}_{\mathbb{Q}_{1:2}} \left[\frac{q_1^*(\mathbf{x}_1) q_2^*(\mathbf{x}_2)}{q_1(\mathbf{x}_1) q_2(\mathbf{x}_2)} \right] = \mathbb{E}_{\mathbb{C}_{1:2}} [\omega(\mathbf{u}_1, \mathbf{u}_2)], \quad (20)$$

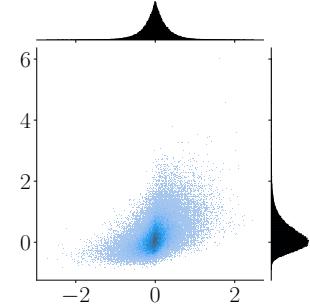
[6].



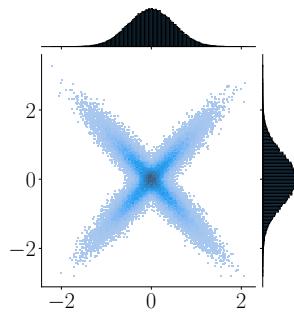
(a) $\mathbb{G}_{1:2}$ is MVN with $\rho = 0.5, \mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1$.



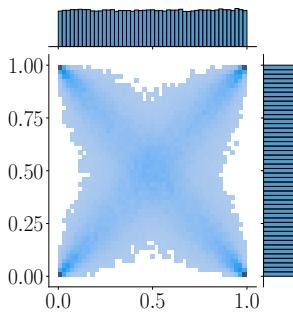
(b) Coupling PDF $c_{1:2}(u_1, u_2)$ associated to the MVN in Fig. 1a.



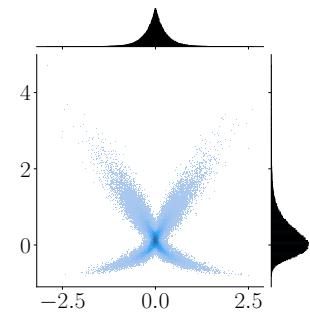
(c) $\mathbb{Q}_{1:2}$ by combining the coupling in Fig. 1b with Laplace and Gumbel marginals.



(d) $\mathbb{G}_{1:2}$ is a mixture of two MVNs with correlations $\rho_1 = -0.9$ and $\rho_2 = 0.9$. Marginals as in Fig. 1a.



(e) Coupling PDF $c_{1:2}(u_1, u_2)$ associated to the mixture of Fig. 1d.



(f) $\mathbb{Q}_{1:2}$ by combining the coupling in Fig. 1e with Laplace and Gumbel marginals.

Figure 1: In the first row, we combine the coupling of a MVN with $\rho = 0.5$ with a Laplace and a Gumbel marginal. In the second row, we instead use a coupling from a mixture of MVNs. The middle column displays the coupling PDFs, which is obtained by passing samples from $\mathbb{G}_{1:2}$ through their CDF. Note: ignoring the middle column, vertically marginals are the same, horizontally couplings are the same.

where the term inside the expectation is a kind of joint-space importance weight in the hypercube $[0, 1]^{2d_x}$, given by

$$\omega(\mathbf{u}_1, \mathbf{u}_2) \stackrel{\text{def}}{=} \omega_2(\mathbf{u}_1) \cdot \omega_2(\mathbf{u}_2), \text{ where } \omega_i(\mathbf{u}_i) \stackrel{\text{def}}{=} \frac{q_i^\star(\mathbf{T}_i(\mathbf{u}_i))}{q_i(\mathbf{T}_i(\mathbf{u}_i))}, i = 1, 2. \quad (21)$$

Notice that no assumptions about independence of the joint proposal $\mathbb{Q}_{1:2}$ is made in the above

derivation, and Eq. (21) follows by change of variable. Provided a coupling as per Definition 6 is used, we have now eliminated the challenging constraint of Eq. (16). Using the new objective, we can now straightforwardly provide definitions for asymptotically optimal coupling (and parametric) mirroring Definition 4 and Definition 5, which we write in Appendix F. As in our framework, the marginals are fixed, we can equivalently talk about asymptotically optimal joints or asymptotically optimal couplings.

3.4 Selection and adaptation of the coupling

We now propose concrete coupling choices that can be used in our framework. We will also propose a method for coupling adaptation to maximize Eq. (16), i.e., explicitly targeting an asymptotically optimal coupling $\mathbb{C}_{1:2}^*$.

To begin with, let us first summarize our overall two-stage framework for combining AIS or with coupling adaptation in the following:

- Inputs to the methods are the marginals $q_1(\mathbf{x}_1), q_2(\mathbf{x}_2)$, their corresponding maps \mathbf{T}_1 and \mathbf{T}_2 , their inverses; a coupling $\mathbb{C}_{1:2}(d\mathbf{u}_1, d\mathbf{u}_2; \boldsymbol{\theta}_c)$, possibly admitting PDF $c_{1:2}(\mathbf{u}_1, \mathbf{u}_2; \boldsymbol{\theta}_c)$.
- **Stage (1): marginals adaptation.** Use any AIS or VI algorithm to obtain marginals.
- **Stage (2): coupling selection/adaptation.** Select a coupling to maximize Eq. (16), possibly with an iterative process.

We start by describing coupling adaptation. A natural approach for stage (2) is first-order stochastic optimization, known for its scalability to large-scale problems [6]. Yet, we stress that in our generic framework, other optimization approaches (e.g., second-order, or sample average approximation (SAA)) could be used. We estimate the gradient of $\mathcal{C}(\boldsymbol{\theta}_c)$, i.e.,

$$\nabla_{\boldsymbol{\theta}_c} \mathcal{C}(\boldsymbol{\theta}_c) = \nabla_{\boldsymbol{\theta}_c} \mathbb{E}_{\mathbb{C}_{1:2}(d\mathbf{u}_1, \mathbf{u}_2; \boldsymbol{\theta}_c)} [\omega(\mathbf{u}_1, \mathbf{u}_2)]. \quad (22)$$

We can estimate the gradient Eq. (22) with either the pathwise or the score function estima-

tor [58], depending on the differentiability of $f(\mathbf{x})$. Stochastic gradient ascent (SGA) starts from initial value $\boldsymbol{\theta}_c^{(0)}$ and iteratively (over $t = 1, \dots$) maximizes $\mathcal{C}(\boldsymbol{\theta}_c)$ using the update rule $\boldsymbol{\theta}_c^{(t)} = \boldsymbol{\theta}_c^{(t-1)} + \eta_t \hat{\nabla}_{\boldsymbol{\theta}_c}^M \mathcal{C}(\boldsymbol{\theta}_c^{(t-1)})$, with $\hat{\nabla}_{\boldsymbol{\theta}_c}^M \mathcal{C}(\boldsymbol{\theta}_c^{(t-1)})$ an unbiased estimator of Eq. (22), stepsize sequence $\{\eta_t\}_{t=1}^T \geq 0$ satisfying $\sum_{t=1}^{\infty} \eta_t = \infty, \sum_{t=1}^{\infty} \eta_t^2 < \infty$ [72].⁶ SGA converges almost surely to a stationary point under weak assumptions even for non-convex smooth functions [4]. Next, we describe possible implementations of the coupling.

Multivariate Normal and Student-t couplings. We start considering the coupling that arises as a consequence of selecting the auxiliary distribution described in §3.2 as MVN. A MVN coupling $\mathbb{C}_{1:2}^{\text{MVN}}(\mathbf{u}_1, \mathbf{u}_2; \mathbf{S}_c)$ is the distribution of $\mathbf{u}_1, \mathbf{u}_2$ obtained by sampling $[\mathbf{z}_1, \mathbf{z}_2] \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_c)$ where $\boldsymbol{\Sigma}_c = \begin{bmatrix} \mathbf{I}_{d_{\mathbf{x}}} & \mathbf{S}_c^\top \\ \mathbf{S}_c & \mathbf{I}_{d_{\mathbf{x}}} \end{bmatrix}$ and setting $\mathbf{u}_i = \Phi(\mathbf{z}_i)$, where Φ is the CDF of the standard normal and is applied element-wise. In particular, to sample from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_c)$ one can use a matrix square root $\mathbf{L} : \mathbf{L}\mathbf{L}^\top = \boldsymbol{\Sigma}_c$, which will have the structure $\mathbf{L} = \begin{bmatrix} \mathbf{I}_{d_{\mathbf{x}}} & \mathbf{0} \\ \mathbf{S}_c & \mathbf{M} \end{bmatrix}$ where \mathbf{M} is a matrix such that $\mathbf{M}\mathbf{M}^\top = \mathbf{I}_{d_{\mathbf{x}}} - \mathbf{S}_c\mathbf{S}_c^\top$. Again, it is worth keeping in mind that using $\mathbb{C}_{1:2}^{\text{MVN}}(d\mathbf{u}_1, d\mathbf{u}_2; \mathbf{S}_c)$ does *not* mean restricting q_1, q_2 be MVNs. Extending to a Student-t coupling is straightforward.

Mixture couplings. A mixture of couplings is a valid coupling, leading to straightforward extensions of the MVN/Student-t couplings (Appendix F). Mixtures may allow us to find a coupling closer to $\mathbb{C}_{1:2}^*$, at the expense of a larger number of parameters to be optimized.⁷

Common random number couplings as special cases and initialization. We show how to incorporate in our framework couplings commonly used in the variance reduction literature for other tasks [63]. We discuss formally in §3.5 more formally their interpretation in terms of the implied joint distribution. These couplings simply set $\mathbf{u}_1 = \mathbf{u}_2$ or simple variations such as $\mathbf{u}_1 = 1 - \mathbf{u}_2$. We now generalize these couplings and interpret them as special cases of sampling from the MVN coupling. As a consequence, we will be able to initialise the parameter \mathbf{S}_c of a MVN coupling at the value that corresponds to using a common random number (CRN) coupling. To start with, we

⁶A special case of the more general stochastic approximation (SA) framework [23].

⁷Differentiation of mixture parameters can be done simply via the application of deterministic mixture sampling techniques, [63, Chapter 9], as done, e.g., in [51].

will represent \mathbf{S}_c via its singular value decomposition (SVD), $\mathbf{S}_c = \mathbf{U}_c \operatorname{diag}(\sigma_1, \dots, \sigma_{d_x}) \mathbf{V}_c^\top$ with \mathbf{U}_c and \mathbf{V}_c orthogonal matrices and singular values constrained to $(-1, 1)^{d_x}$ to ensure positive-semidefiniteness of Σ_c .⁸ **Case $\mathbf{u}_1 = \mathbf{u}_2$:** With MVNs marginals, this coupling is equivalent to setting $\mathbf{z}_1 = \mathbf{z}_2$ as discussed above on how to sample from $\mathbb{C}_{1:2}^{\text{MVN}}$. Therefore, by simply setting $\mathbf{S}_c = \mathbf{I}_{d_x}$, this coupling is recovered as a special case, in which case $\mathbf{M}\mathbf{M}^\top = \mathbf{I}_{d_x} - \mathbf{S}_c\mathbf{S}_c^\top = \mathbf{0}$. **Case $\mathbf{u}_1 = 1 - \mathbf{u}_2$:** this is a straightforward extension of the previous case. Indeed, setting \mathbf{S}_c to an orthogonal matrix generalizes both these cases and includes other new variants. In [45], this technique was briefly mentioned but not elaborated.

3.5 Special cases falling within our framework

We now show that the estimator $\hat{\mu}_{\text{SNIS}}^N$ is special case of $\hat{\mu}_{\text{GenSNIS}}^N$, formalizing the implicit assumption of using the same samples to estimate the integrals I and Z_p as discussed in §2.1. We also show how several existing methods can be viewed as special cases, and that the flexibility of our joint construction allows one to combine existing couplings with more generic marginal families, with theoretical guarantees which we expand on in §3.6.

Couplings based on common random numbers. Within our framework, sharing the same samples in both integrals, as done by $\hat{\mu}_{\text{SNIS}}^N$, corresponds to choosing a joint singular probability measure $\mathbb{Q}_{1:2}^{\text{SNIS}}(d\mathbf{x}_1, d\mathbf{x}_2) = \delta_{\mathbf{x}_1}(d\mathbf{x}_2) \cdot \mathbb{Q}_1(d\mathbf{x}_1) = \delta_{\mathbf{x}_2}(d\mathbf{x}_1) \cdot \mathbb{Q}_2(d\mathbf{x}_2)$ where **(i)** for almost every \mathbf{x}_1 (respectively, \mathbf{x}_2) the regular conditional distribution of one vector given the other is a Dirac measure and **(ii)** the marginals are equal, $\mathbb{Q}_1 = \mathbb{Q}_2$.⁹ The corresponding coupling is then also a singular probability measure; by a change of variable we get

$$\mathbb{C}_{1:2}^{\text{CRN}}(d\mathbf{u}_1, d\mathbf{u}_2) = \mathbf{T}_{1:2} \# \mathbb{Q}_{1:2}^{\text{SNIS}}(d\mathbf{x}_1, d\mathbf{x}_2) = \delta_{\mathbf{u}_1}(d\mathbf{u}_2) \cdot \mathbb{U}_{[0,1]^{d_x}}(d\mathbf{u}_1) = \delta_{\mathbf{u}_2}(d\mathbf{u}_1) \cdot \mathbb{U}_{[0,1]^{d_x}}(d\mathbf{u}_2), \quad (23)$$

where CRN stands for “common random numbers” and we use the VQ transformations $\mathbf{T}_{1:2} \stackrel{\text{def}}{=} \mathbf{T}_{1:2} \# \mathbb{Q}_{1:2}^{\text{SNIS}}$

⁸Note that \mathbf{S}_c is the off-diagonal block of Σ_c , and it is not required to be neither symmetric nor positive-semidefinite.

⁹A singular probability measure does not admit a density w.r.t. the Lebesgue measure.

$[\mathbf{T}_1(dx_1), \mathbf{T}_2(dx_2)]^\top$ applied to the product space of measurable subsets of \mathbb{R}^{d_x} . [33] notice that this coupling is a generalization of the so-called comonotonic copula [63, Chapter 5]. Similarly, we can of course have $\delta_{1-u_1}(du_2) \cdot \mathbb{U}_{[0,1]^{d_x}}(du_1)$ as another coupling, corresponding to the counter-monotonic copula.¹⁰ The above discussion formalizes and generalizes the method to sample from $\mathbb{C}_{1:2}^{\text{CRN}}$ (and variants) which we described at the end of §3.4.

Independent coupling. Another important coupling corresponds to generating two independent sets of samples, as done in the recent target-aware Bayesian inference (TABI) [69], given by

$$\mathbb{C}_{1:2}^{\text{INDEP}}(du_1, du_2) = \mathbb{U}_{[0,1]^{d_x}}(du_1) \cdot \mathbb{U}_{[0,1]^{d_x}}(du_2). \quad (24)$$

Such a coupling, differently from SNIS, allows one to have two different distributions. Crucially, however, it cannot exploit the third term of the asymptotic variance for variance reduction. Two different marginals were also used by [52], who linearly transform a Normal r.v. \mathbf{x}_1 into \mathbf{x}_2 . Their implicit coupling is the one in Eq. (23) for $d_x = 1$, while their method does not extend easily to $d_x > 1$, and they only considered Normal marginals. It is worth noting that a paper proposing the use of two different distributions q_1, q_2 for numerator and denominator first appeared in [42], which is mentioned in [46]. These previous works do not try to study the influence of the coupling on the asymptotic variance. A full overview of the methods falling within our framework is given in Table 1. Our unifying framework, along with the interpretable variance decomposition in Eq. (19) allows us to study which choices of couplings can be appropriate in which settings, and how to iteratively improve over them with adaptation. In §3.7, we will connect with existing literature on ratio estimation more broadly, that use an estimator fundamentally different from Eq. (13).

3.6 Theoretical properties of the framework

Our framework allows us to understand the maximum reduction in asymptotic variance for estimating μ , explicitly linking it to the quality of the joint $\mathbb{Q}_{1:2}$ and therefore of the coupling $\mathbb{C}_{1:2}$.

¹⁰Mixing the two would be possible, setting some components as $u_{1,1} = u_{2,1}$ and others as $u_{1,3} = 1 - u_{2,3}$.

Table 1: We show how our framework generalizes many previous methods in the literature for the problem of estimating μ with an intractable normalizing constant with IS. $\mathbb{C}_{1:2}^{\text{CRN}}$ is defined in Eq. (23), $\mathbb{C}_{1:2}^{\text{INDEP}}$ is defined in Eq. (24).

Special case of $\hat{\mu}_{\text{GenSNIS}}^N$	Choice of marginals	Choice of coupling
SNIS	$q_1(\mathbf{x}_1) = q_2(\mathbf{x}_2)$	$\mathbb{C}_{1:2} = \mathbb{C}_{1:2}^{\text{CRN}}$
UIS	Any $q_1(\mathbf{x}_1), q_2(\mathbf{x}_2) = p(\mathbf{x}_2)$	$\mathbb{C}_{1:2} = \mathbb{C}_{1:2}^{\text{CRN}}$
[52]	$q_1(\mathbf{x}_1) \neq q_2(\mathbf{x}_2)$ (Normal joint, $d_{\mathbf{x}} = 1$)	$\mathbb{C}_{1:2} = \mathbb{C}_{1:2}^{\text{CRN}}$
$[42], [41, 69]$ (TABI)	$q_1(\mathbf{x}_1) \neq q_2(\mathbf{x}_2)$	$\mathbb{C}_{1:2} = \mathbb{C}_{1:2}^{\text{INDEP}}$
Equal marginals, dependent (new)	$q_1(\mathbf{x}_1) = q_2(\mathbf{x}_2)$	Any $\mathbb{C}_{1:2}$
Equal marginals, independent (new)	$q_1(\mathbf{x}_1) = q_2(\mathbf{x}_2)$	$\mathbb{C}_{1:2} = \mathbb{C}_{1:2}^{\text{INDEP}}$

While identifying which coupling lead to maximum variance reduction in general is intractable, we identify a few special cases and prove their optimality. We then show that our construction of a joint via couplings and marginals enjoys theoretical guarantees of expressiveness thanks to a recent generalisation of the celebrated Sklar's theorem for copulas.

Maximum asymptotic variance reduction. The maximum asymptotic variance reduction is achieved when using the optimal joint distribution $\mathbb{Q}_{1:2}^{\star}(d\mathbf{x}_1, d\mathbf{x}_2)$ defined in Definition 4 or, equivalently, the optimal coupling $\mathbb{C}_{1:2}^{\star}$, leading to the following proposition.

Proposition 3. *The asymptotic variance of $\hat{\mu}_{\text{GenSNIS}}^N$ with two given marginals q_1, q_2 is lower bounded as*

$$\mu^{-2} \cdot \mathbb{V}_{\mathbb{Q}_{1:2}}^{\infty} [\hat{\mu}_{\text{GenSNIS}}^N] \geq \mu^{-2} \cdot \mathbb{V}_{\mathbb{Q}_{1:2}^{\star}}^{\infty} [\hat{\mu}_{\text{GenSNIS}}^N] = \left(\sqrt{\chi^2(q_2^{\star} || q_2)} - \sqrt{\chi^2(q_1^{\star} || q_1)} \right)^2. \quad (25)$$

The above follows by applying the Cauchy-Swartz inequality to the three terms in Eq. (19)(see

Appendix G for the proof). The joint, which appears in $\mathbb{V}_{\mathbb{Q}_{1:2}}^\infty[\hat{\mu}_{\text{GenSNIS}}^N]$, has disappeared in the lower bound, which only depends on the marginals. An interesting implication of Eq. (25) is that it is possible for $\hat{\mu}_{\text{GenSNIS}}^N$ to have low variance even if both marginals are far from their respective q_1^\star, q_2^\star , as long as the divergences $\chi^2(q_1^\star || q_1)$ and $\chi^2(q_2^\star || q_2)$ obtain a similar value.

Theory for asymptotically optimal couplings. In general, there is no systematic way to identify which coupling $\mathbb{C}_{1:2}^\star$ corresponds to $\mathbb{Q}_{1:2}^\star$. This motivates adaptation methods such as the one we proposed in §3.4. However, interestingly, we can identify a few particular cases where the optimal coupling is known.

Proposition 4. *Let the marginals be product PDFs, i.e., $q_1(\mathbf{x}_1) = \prod_{d=1}^{d_{\mathbf{x}}} q_1^{(d)}(x_{1,d})$, $q_2(\mathbf{x}_2) = \prod_{d=1}^{d_{\mathbf{x}}} q_2^{(d)}(x_{2,d})$. Then, let the functions $\mathbf{x}_1 \rightarrow f(\mathbf{x}_1)\tilde{p}(\mathbf{x}_1)/q_1(\mathbf{x}_1)$ and $\mathbf{x}_2 \rightarrow \tilde{p}(\mathbf{x}_2)/q_2(\mathbf{x}_2)$ be (i) uniformly continuous $\mathbb{Q}_{1:2}$ -almost surely and (ii) monotonic in the same direction w.r.t. the d -th component $x_{1,d}$ (respectively, $x_{2,d}$), for $d = 1, \dots, d_{\mathbf{x}}$. In this case, the CRN coupling is asymptotically optimal, i.e., $\mathbb{C}_{1:2}^\star = \arg \max_{\mathbb{C}_{1:2}} \mathcal{C}(\mathbb{C}_{1:2}) = \mathbb{C}_{1:2}^{\text{CRN}}$.*

The above follows by noticing that Proposition 1 in the Appendix of [74] applies with our expectation in Eq. (16). This result generalizes a simpler one for the difference of expectations holding for $d_{\mathbf{x}} = 1$ and replacing VQs for quantiles that can be found in textbooks [22].

In light of realistic estimation problems, (i) proposals q_1, q_2 which are factorized are highly restrictive and (ii) it is impractical/impossible to check exact monotonicity for all components. This also suggests that the choice of optimal coupling is significantly more difficult for $d_{\mathbf{x}} > 1$.

Next, we see that the joint construction proposed in §3.2 enjoys considerable generality in terms of the families of joint distributions that can be expressed.

Proposition 5 (Vector Sklar's theorem, [33] Theorem 1). *Let $\mathbb{Q}_{1:2}$ be any joint distribution on $\mathbb{R}^{d_{\mathbf{x}}} \times \mathbb{R}^{d_{\mathbf{x}}}$ with marginals $\mathbb{Q}_1, \mathbb{Q}_2$ on $\mathbb{R}^{d_{\mathbf{x}}}$, and let $\psi(\mathbf{u}_i) : [0, 1]^{d_{\mathbf{x}}} \rightarrow \mathbb{R} \cup \{+\infty\}$ be the convex function associated to the vector quantile, \mathbf{T}_i , for each marginal, $i = 1, 2$. Then, there exists a*

vector copula $\mathbb{C}_{1:2}(d\mathbf{u}_1, d\mathbf{u}_2)$ (coupling, in our terminology) such that the following properties hold.

- If $\mathbb{Q}_1, \mathbb{Q}_2$ are absolutely continuous on \mathbb{R}^{d_x} with support in a convex set, then $\mathbb{C}_{1:2}$ is the unique vector copula, such that for all Borel sets B_1, B_2 in $[0, 1]^{d_x} \times [0, 1]^{d_x}$, it holds $\mathbb{C}_{1:2}(B_1, B_2) = \mathbb{Q}_{1:2}(\nabla\psi_1(B_1), \nabla\psi_2(B_2))$.
- For any Borel subsets A_1, A_2 of \mathbb{R}^{d_x} , it holds $\mathbb{Q}_{1:2}(A_1, A_2) = \mathbb{C}_{1:2}(\partial\psi_1^*(A_1), \partial\psi_2^*(A_2))$, where $\partial\psi_i^*$ denotes the subdifferential of the convex conjugate of ψ_i .¹¹

The above result is well-suited for the IS setting, as the user designs and simulates from distributions, making it easy to satisfy the assumptions. The result applies, in particular, to both the semi-parametric and the parametric model of joints described in §3.

Asymptotic normality, almost sure convergence. Finally, $\hat{\mu}_{\text{GenSNIS}}^N$ enjoys the same theoretical guarantees as $\hat{\mu}_{\text{SNIS}}^N$ (see, e.g., [16]) in terms of asymptotic normality and consistency (almost sure convergence), since the particular choice of joint $\mathbb{Q}_{1:2}$ leaves these properties unchanged.

3.7 Further connections with the literature

Even as far back as [50], it was noticed that most quantities of interest arise as ratios of integrals and formulas for the asymptotic variance of the ratio estimator presented. However, previous works do not exploit what we call “the third term” to construct the proposal, especially not using a generic joint in extended space. Previous theoretical analyses covering SNIS asymptotics appear in [47, 38]. A more recent set of works [1, 76, 77] focusses on worst-case bounds on the MSE over classes of test functions $f(\mathbf{x})$ or proposes to use concentration inequalities [11]. Recent work [78] notes the inefficiency of SNIS for problems like our motivating example but with a different focus. We focus on cases where the new test point is highly informative, while they focus on improving stability (finiteness of the asymptotic variance) of the estimates when there are many test points, using a technique that appears to be similar to the defensive mixture sampling of [62] and the

¹¹Theorem 1 of [33] lists other properties of $\mathbb{C}_{1:2}$.

mixture sampling method of [13]. **Two proposal estimators and ratio estimation.** The use of two IS proposals in the context of ratio estimation has been proposed first, to the best of our knowledge, in [42]. Further similar ideas appear naturally in the context of bridge sampling [57, 56] (see also [63, Chapter 10] for an overview) and “ratio IS” [12], where the aim is to estimate a ratio of normalizing constants Z_0/Z_1 for two distributions, p_0 and p_1 , with different supports. When $f(\mathbf{x}) > 0$ (as assumed in the present paper) then $p(\mathbf{x})f(\mathbf{x})$ could be seen as p_0 , and $p(\mathbf{x})$ as p_1 ; the supports are generally the same, unless $f(\mathbf{x})$ is an indicator function. In bridge sampling, and related methods [82, 80] including the “reverse logistic regression” of Geyer [39], it is assumed that one can sample from either p_0 or p_1 , which is in contrast to most applications where SNIS is used [83]. The TABI work [69] also has a discussion on the disadvantages of bridge sampling for the type of problems we target. Further, some often use more complicated estimators which recycle samples (see [54] for an overview of these) which would render theoretical characterization of the optimality of proposal marginals and joint more difficult. However, for an algorithm, the techniques of these bridge sampling-related estimators could be combined with our framework in an extension. [52] and [41] (later extended in a journal version named “target-aware Bayesian inference” (TABI) [69]) are the two more recent works using two proposals for problems typically addressed by SNIS. Table 1 shows how these estimators can be seen as particular cases of $\hat{\mu}_{\text{GenSNIS}}^N$. [69] in particular briefly mentions the sub-optimality of having independent samples in the two integrals of the ratio and mentions the possibility of addressing this with standard common random numbers. [53] proposes an extension of [69] but still not addressing the third term of the asymptotic variance in $\hat{\mu}_{\text{SNIS}}^N$. [21] also introduces a correlation to improve a ratio estimator in the context of estimating the likelihood within pseudomarginal algorithms. **Adaptive antithetic sampling.** Our coupling adaptation can be viewed as related to an adaptive antithetic sampling technique. The most relevant previous work is [45]. There, the authors optimize a joint with two fixed (equal) marginals via the use of a Gaussian coupling for the optimization of a *difference* of expectations, being the first work we

are aware of to do so.¹² Their framework only covers Gaussian couplings and does not consider modern scalable stochastic optimization methods. **SNIS bias reduction.** [79, 9] (and references therein) focus on bias reduction. Finally, a recent related work also addresses expectations with unnormalized PDFs, but using a different set of techniques [36]. Yet another is [18], where the authors also address ratio estimation. Differently from us, they are (i) specialized for rare events (ii) they do not sample from a generic joint in an extended space (iii) their framework can be interpreted as a special case while using a specific parametric coupling, while ours is supported by the generality of vector copula theory (see §3.6), and importantly (iv) their framework to sample from (what we interpret as) marginals depends on the variables ordering [5], to an extent that is not well understood.

4 Experiments

We demonstrate that significantly more accurate estimates of expectations are achieved with a judicious choice of coupling, as informed by our general framework. In particular, we show, in all cases we tried, substantial improvements against SNIS with either of the two marginals as proposal, and better or competitive performance with the independent coupling, which corresponds to the recent state-of-the-art framework TABI [69]. We follow a similar initialisation strategy to [45], as discussed in §3.4; we consider the CRN couplings $\mathbf{u}_1 = \mathbf{u}_2$ and $\mathbf{u}_1 = 1 - \mathbf{u}_2$ as well as the independent coupling as starting points for initialisation; if no better coupling is found, we take the better one.

In §4.1, we start with a toy example where the asymptotic variance is closed form. Then we consider more realistic experiments with Bayesian logistic regression. The experiments target the problem given by our motivating example from §2.3. We demonstrate the combination with existing state-of-the-art AIS algorithms by adapting the marginals q_1, q_2 as heavy-tailed multivariate Student-t

¹²A similar coupling perspective on variance reduction in *difference of expectations* can be found in https://hackmd.io/@sp-monte-carlo/rJm9r_WeY

distributions with the AHTIS algorithm [43]. We generate corrupted test points by sampling from a Student-t distribution with $\nu = 3$ with mean equal to the empirical average of the observed data and $\Sigma = 10 \cdot \mathbf{I}$.¹³

4.1 Illustrative example: Bayesian linear regression

We now consider a Bayesian linear regression model where the asymptotic variance of $\hat{\mu}_{\text{GenSNIS}}^N$ can be computed in closed form. The main aim of this first example is to show that we need, simultaneously, different marginals and a judicious choice of coupling, even in such a relatively simple scenario, and that the choice of coupling is generally not obvious.

Denote with $\mathbf{X} \in \mathbb{R}^{N \times D}$ a matrix of N, D dimensional covariates/designs. We focus on a setting with fixed (non-random) designs. In this model, scalar observations $\{y^{(n)}\}_{n=1}^N$ have PDF $g(y^{(n)}|\mathbf{x}^{(n)}, \boldsymbol{\theta}) = \mathcal{N}_{y^{(n)}}(\mathbf{x}^{(n)\top} \boldsymbol{\theta}, \sigma^2)$, $n = 1, \dots, N$, where the variance σ^2 is known. Assuming a MVN prior $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}) = \mathcal{N}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Sigma_0)$, the posterior given N observations is given by $\pi(\boldsymbol{\theta}|\mathbf{X}, y_{1:N}) = \mathcal{N}_{\boldsymbol{\theta}}(\boldsymbol{\mu}_N, \Sigma_N)$. The quantity to be estimated is then $\mu = \mathbb{E}_{\pi(\boldsymbol{\theta}|\mathbf{X}, y_{1:N})}[g(y_{N+1}|\mathbf{x}_{N+1}, \mathbf{X}, y_{1:N})]$ for a test pair $(y_{N+1}, \mathbf{x}_{N+1})$. Since the true optimal marginals to estimate μ are MVN, $q_1^\star(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{X}, \mathbf{x}_{S+1}, y_{1:S+1})$ and $q_2^\star(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{X}, y_{1:S})$, we consider a MVN joint proposal $q_{1:2}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \mathcal{N}_{\boldsymbol{\theta}}\left(\boldsymbol{\mu}_{q_{1:2}} = \begin{bmatrix} \boldsymbol{\mu}_{q_1} \\ \boldsymbol{\mu}_{q_2} \end{bmatrix}, \Sigma_{q_{1:2}} = \begin{bmatrix} \Sigma_{q_1} & \mathbf{L}_{q_2} \mathbf{S}_c^\top \mathbf{L}_{q_1}^\top \\ \mathbf{L}_{q_1} \mathbf{S}_c \mathbf{L}_{q_2}^\top & \Sigma_{q_2} \end{bmatrix}\right)$, where $\mathbf{L}_{q_1}, \mathbf{L}_{q_2}$ are matrix square roots of, respectively, the marginal covariances $\Sigma_{q_1}, \Sigma_{q_2}$.¹⁴ Then, the coupling parameter to be chosen is \mathbf{S}_c . Recall from §3.5 that we parameterize \mathbf{S}_c via its SVD, i.e., $\mathbf{S}_c = \mathbf{U} \text{diag}(\lambda_1, \dots, \lambda_{d_{\mathbf{x}}}) \mathbf{V}^\top$ and when \mathbf{S}_c is orthogonal, so that its eigenvalues are ± 1 , the MVN coupling generalizes simple couplings such as $\mathbf{u}_1 = \mathbf{u}_2$ or $\mathbf{u}_1 = 1 - \mathbf{u}_2$.

We will study how the relative asymptotic variance changes as a function of the eigenvalues of \mathbf{S}_c .

¹³Python source code will be released post-review.

¹⁴This corresponds to a MVN coupling and MVN marginals.

Following our decomposition of Eq. (14), the relative asymptotic variance is

$$\chi^2(\pi(\boldsymbol{\theta}_1|\mathbf{X}, \mathbf{x}_{N+1}, y_{1:N+1})||q_2(\boldsymbol{\theta}_2)) + \chi^2(\pi(\boldsymbol{\theta}_2|\mathbf{X}, y_{1:N})||q_2(\boldsymbol{\theta}_2)) - 2 \left(\overbrace{\mathbb{E}_{q_{1:2}} \left[\frac{\pi(\boldsymbol{\theta}_1|\mathbf{X}, \mathbf{x}_{N+1}, y_{1:N+1})\pi(\boldsymbol{\theta}_2|\mathbf{X}, y_{1:N})}{q_1(\boldsymbol{\theta}_1)q_2(\boldsymbol{\theta}_2)} \right]}^{C(q_{1:2})} - 1 \right). \quad (26)$$

Only the third term depends on \mathbf{S}_c . As we derive in Appendix F, Eq. (26) leads to $C(q_{1:2}) = C_1 \cdot C_2 \cdot C_3$, where C_1 and C_2 are constants that result from multiplying and dividing multivariate normal PDFs, and $C_3 = \mathcal{N}_{\mathbf{r}}(\boldsymbol{\mu}_{q_{1:2}}, \mathbf{R} + \boldsymbol{\Sigma}_{q_{1:2}})$, where \mathbf{r} and \mathbf{R} are defined as

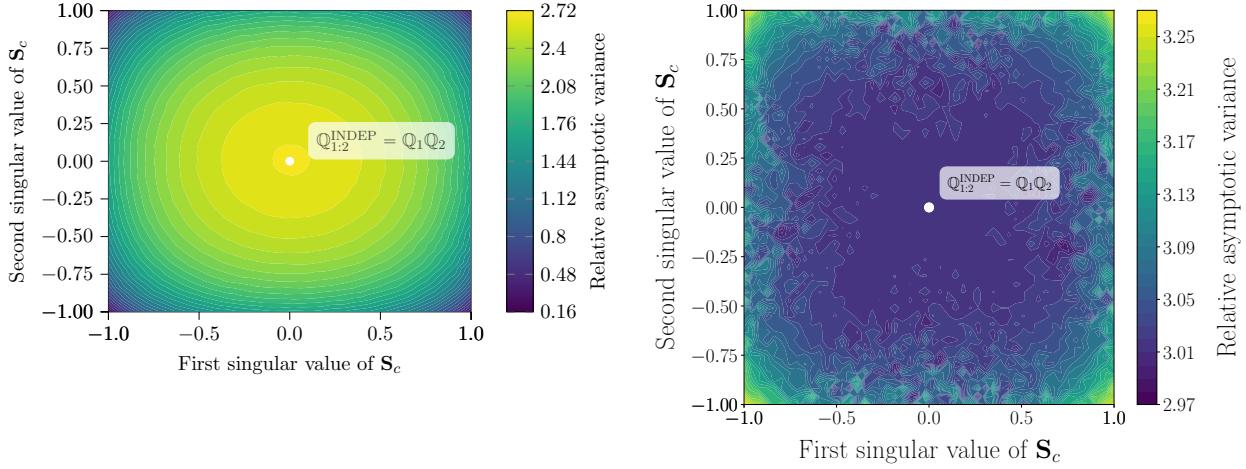
$$\begin{cases} \mathbf{R}_1 = (\boldsymbol{\Sigma}_{q_1}^{-1} - \boldsymbol{\Sigma}_{q_1}^{-1})^{-1}, & \mathbf{r}_1 = \mathbf{R}_1(\boldsymbol{\Sigma}_{q_1}^{-1}\boldsymbol{\mu}_{q_1}^* - \boldsymbol{\Sigma}_{q_1}^{-1}\boldsymbol{\mu}_{q_1}) \\ \mathbf{R}_2 = (\boldsymbol{\Sigma}_{q_2}^{-1} - \boldsymbol{\Sigma}_{q_2}^{-1})^{-1}, & \mathbf{r}_2 = \mathbf{R}_2(\boldsymbol{\Sigma}_{q_2}^{-1}\boldsymbol{\mu}_{q_2}^* - \boldsymbol{\Sigma}_{q_2}^{-1}\boldsymbol{\mu}_{q_2}) \end{cases}. \quad (27)$$

Since we can only visualize the variance as a function of the eigenvalues (for $D = 2$), plots show averages over 50 independent runs of simulating random orthogonal matrices for \mathbf{U}, \mathbf{V} from the uniform distribution over orthogonal matrices.¹⁵ We considered both possible cases of posteriors with independent dimensions, and factorised. In practically all settings we tried, the optimal solution is at the corners of the parameter space, as §4.1 shows. Yet, it is possible (although much less frequent) for the independent choice to be the minimum, as in §4.1. Interestingly, this can be even if the posterior has dependent dimensions.

4.2 Bayesian logistic regression

Next, we test how performance in estimation of μ in terms of MSE in practice can be achieved by choosing an appropriate coupling. As before, the quantity to be estimated is $\mu = \mathbb{E}_{\pi(\boldsymbol{\theta}|\mathbf{X}, y_{1:N})}[g(y_{N+1}|\mathbf{x}_{N+1}, \mathbf{X}, y_{1:N})]$ for test data $(y_{N+1}, \mathbf{x}_{N+1})$. We used 10 test points, $(y_{N+1}, \mathbf{x}_{N+1}), (y_{N+2}, \mathbf{x}_{N+2}, \dots, y_{N+10})$, so that the optimal marginals $\pi(\boldsymbol{\theta}|\mathbf{X}, y_{1:N})$ and $\pi(\boldsymbol{\theta}|\mathbf{X}, y_{1:N+10})$ are very different from each other. We find that there is consistent benefit in using $\hat{\mu}_{\text{GenSNIS}}^N$ with any coupling instead of $\hat{\mu}_{\text{SNIS}}^N$. Moreover, we find that in most settings it is also beneficial to find a better coupling than TABI [69], i.e., the independent coupling.

¹⁵See <https://www.ams.org/notices/200705/fea-mezzadri-web.pdf>.



- (a) For this example, we set $\mu_{q_1^*} = [-1/4, 1/4]$ and $\Sigma_{q_1^*} = 1/4\mathbf{I}$. $\mu_{q_2^*} = [1/4, -1/4]$, $\Sigma_{q_2^*} = \mathbf{I}$. The marginals have $\mu_{q_1} = [1/4, -1/4]$, $\mu_{q_2} = [-1/4, 1/4]$, $\Sigma_{q_1} = 4 \cdot \Sigma_{q_1^*}$, $\Sigma_{q_2} = 4 \cdot \Sigma_{q_2^*}$.
- (b) (a) For this example, we set $\mu_{q_1^*} = [-0.5, 0]$, $\Sigma_{q_1^*} = \begin{bmatrix} 0.4 & -0.5\sqrt{0.4} \\ -0.5\sqrt{0.4} & 1 \end{bmatrix}$. Similarly, $\mu_{q_2^*} = [0, -0.5]$, $\Sigma_{q_2^*} = \begin{bmatrix} 1 & 0.5\sqrt{0.4} \\ 0.5\sqrt{0.4} & 0.4 \end{bmatrix}$. For the proposals, $\mu_{q_1} = [0.5, 0]$, $\Sigma_{q_1} = \begin{bmatrix} 2.5 \cdot 0.4 & -0.4\sqrt{2.5^2 \cdot 0.4} \\ -0.4\sqrt{2.5^2 \cdot 0.4} & 2.5 \end{bmatrix}$. Similarly, $\mu_{q_2} = [0, -0.5]$, $\Sigma_{q_2} = \begin{bmatrix} 2.5 & 0.5\sqrt{2.5^2 \cdot 0.4} \\ 0.5\sqrt{2.5^2 \cdot 0.4} & 2.5 \cdot 0.4 \end{bmatrix}$.

Figure 2: **Bayesian linear regression example.** In this example, we show how the closed form asymptotic variance changes as a function of the 2 singular values of \mathbf{S}_c . Note that the corners of the parameter space correspond to a coupling given by an orthogonal transformation of \mathbf{x}_1 into \mathbf{x}_2 , as discussed in §3.4. With many settings tried, we found that most runs lead to a result as in §4.1. This shows that while different marginals are clearly needed to better match q_1^*, q_2^* , a better coupling is needed to minimize variance. In very few cases, the result is as in §4.1

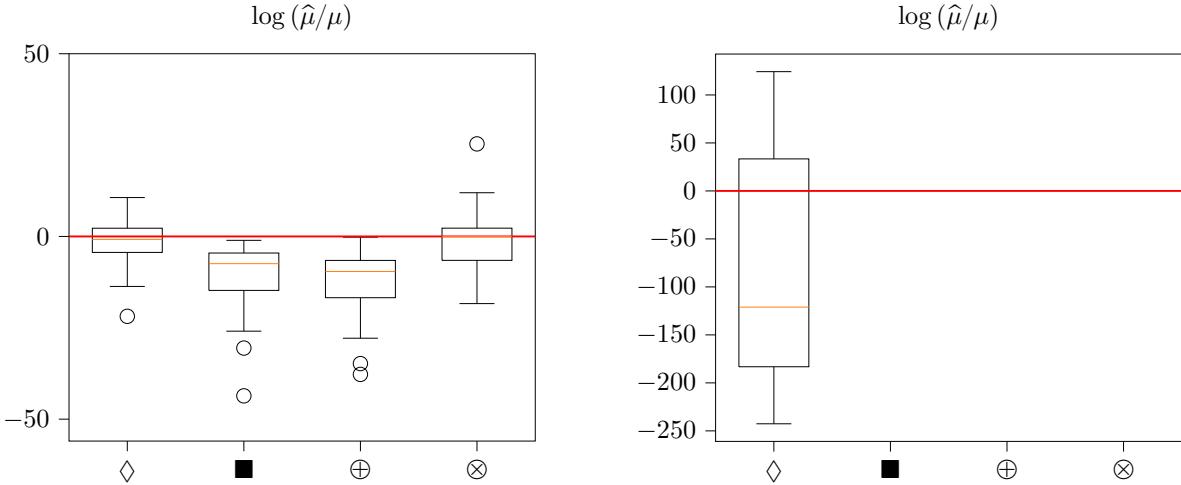
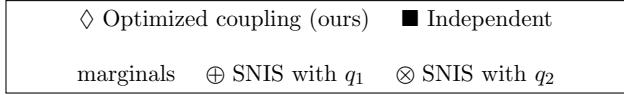
We will use the MVN coupling described in §3.4, which amounts to selecting and optimizing the matrix parameter \mathbf{S}_c , with its special cases as described before. To find the best coupling, we follow the strategy described in §3.4 of initialising a stochastic optimization procedure at $\mathbf{S}_c = \mathbf{I}$, $\mathbf{S}_c = -\mathbf{I}$ and $\mathbf{S}_c = \mathbf{0}$ (corresponding to TABI [69]) and taking the best. For stochastic optimization, we used the amsgrad optimizer [70] from the Python JAX library, and a pathwise gradient estimate, running for 500 iterations with a linear learning rate schedule from 0.05 to 0.005.

We consider a Bayesian logistic regression model which we briefly describe next. Covariates are denoted as $\mathbf{X} \in \mathbb{R}^{N \times D+1}$, with an additional column of 1's for the intercept. We denote binary responses with $y \in \{0, 1\}$ and coefficients with $\boldsymbol{\theta} \in \mathbb{R}^{D+1}$. We generate the covariates independently from a standard MVN. The standard logistic regression posterior is given as $\pi(\boldsymbol{\theta}|\mathbf{X}, y) \propto \prod_n g(y|\mathbf{x}_n, \boldsymbol{\theta})\pi(\boldsymbol{\theta})$ with $g(y|\mathbf{x}_n, \boldsymbol{\theta})$ a Poisson likelihood with sigmoid transformation and $\pi(\boldsymbol{\theta})$ a standard MVN.

We tested the estimator on a dataset with $N = 10$ and $D \in \{10, 40\}$ and replicate all experiments for 50 simulations. We adapted proposals with the state-of-the-art AHTIS algorithm [43] which can handle heavy tailed targets. We let M_{adapt} to be the number of samples used for adapting the marginals, and M_{eval} the number of samples used to evaluate the estimators. Fig. 3 shows the results. In all settings, choosing our optimization converges to the best coupling possible, and outperforms TABI and SNIS.

5 Conclusions

We have proposed a methodological framework to understand variance reduction for estimating expectations that involve intractable normalizing constants, exploiting the view of the problem as estimating a ratio of integrals. Within our framework, samples are obtained from a generic joint distribution in an extended space, while two of its (multivariate) marginals are used to construct a ratio of two estimators. Our methodology allows one to take the output of any two adaptive



(a) **Dimension 10.** Here, $M_{\text{adapt}} = 1500$ and $M_{\text{eval}} = 200$. In this example, we observed that the optimized coupling converges to $\mathbf{u}_1 = \mathbf{u}_2$, which gives the best performance and what is shown ($\mathbf{u}_1 = 1 - \mathbf{u}_2$ obtains similar performance). The plot shows significant improvement over the independent solution and SNIS even at low sample sizes such as 200. Note that the independent solution further has an outlier at $\log(\hat{\mu}/\mu) \approx +200$ (not shown), which denotes a larger instability. We obtained very similar results for other values of M_{eval} , up to 5000.

(b) **Dimension 40.** Note that the independent solution and SNIS are not visible as their bad performance leads to NaNs and $+\infty$. Here, $M_{\text{adapt}} = 5000$ and $M_{\text{eval}} = 3000$. In this example, where proposal marginals are clearly very mismatched, the optimized coupling converged to a very similar performance than $\mathbf{u}_1 = \mathbf{u}_2$, but interestingly with eigenvalues different than 1. This shows that there may be multiple good solutions. We obtained similar results for other (larger) values of M_{eval} .

Figure 3: Boxplots comparing performances of different options for the estimation of μ , the posterior predictive integral given $N = 10$ test points for a model of dimension $D = 10$ (§4.2) and $D = 40$ (§4.2). All options use the same fixed marginals q_1, q_2 , which were trained with AIS in a first stage. The best value is at 0, when $\hat{\mu} = \mu$. The ground truth μ was computed with $M_{\text{evaluate}} = 8 \cdot 10^5$ samples, from marginals that were adapted with $M_{\text{adapt}} = 5 \cdot 10^5$. For $D = 40$, $M_{\text{evaluate}} = M_{\text{adapt}} = 8 \cdot 10^5$.

IS (or VI) algorithms and further reduce estimation variance by coupling them through a joint distribution. We have shown how the widely used SNIS estimator and several other ratio estimators can be viewed as special cases that make an implicit choice of this joint. The framework highlights the limitations of existing approaches, allows us to develop new methodology, for which we provide concrete examples here, and paves the way for further variance reduction techniques.

6 Acknowledgements

We would like to thank several people for their feedback and useful discussions on this work and some of its earlier versions. In chronological order, we thank Simo Sarkka for interesting discussions on the functional optimization problem for the joint at BayesComp 2023; to Art Owen for generous comments and feedback at MCM 2023; to Christophe Andrieu for valuable suggestions on how the theory could be extended and relevant related work at the SMC workshop in 2024.

References

- [1] Sergios Agapiou, Omiros Papaspiliopoulos, Daniel Sanz-Alonso, and Andrew M Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 2017.
- [2] Omer Deniz Akyildiz. Global convergence of optimized adaptive importance samplers. *Foundations of Data Science*, 2024.
- [3] Ömer Deniz Akyildiz and Joaquín Míguez. Convergence rates for optimised adaptive importance samplers. *Statistics and Computing*, 31(2), 2021.
- [4] Dimitri P Bertsekas and John N Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3), 2000.
- [5] Zdravko I Botev. The normal law under linear restrictions: simulation and estimation via

- minimax tilting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(1):125–148, 2017.
- [6] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2), 2018.
- [7] Monica F Bugallo, Victor Elvira, Luca Martino, David Luengo, Joaquin Miguez, and Petar M Djuric. Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79, 2017.
- [8] Olivier Cappé, Arnaud Guillin, Jean-Michel Marin, and Christian P Robert. Population monte carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929, 2004.
- [9] Gabriel Cardoso et al. Br-snls: Bias reduced self-normalized importance sampling. *Advances in Neural Information Processing Systems*, 2022.
- [10] Guillaume Carlier, Victor Chernozhukov, Gwendoline De Bie, and Alfred Galichon. Vector quantile regression and optimal transport, from theory to numerics. *Empirical Economics*, pages 1–28, 2020.
- [11] Sourav Chatterjee and Persi Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2), 2018.
- [12] Ming-Hui Chen and Qi-Man Shao. On monte carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics*, 25(4), 1997.
- [13] Ming-Hui Chen and Qi-Man Shao. Monte carlo methods for bayesian analysis of constrained parameter problems. *Biometrika*, 1998.
- [14] Ming-Hui Chen and Qi-Man Shao. Monte carlo estimation of bayesian credible and hpd intervals. *Journal of computational and Graphical Statistics*, 8(1), 1999.

- [15] Victor Chernozhukov, Alfred Galichon, Marc Hallin, and Marc Henry. Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1), 2017.
- [16] Nicolas Chopin, Omiros Papaspiliopoulos, et al. *An introduction to sequential Monte Carlo*. Springer, 2020.
- [17] Jean-Marie Cornuet, Jean-Michel Marin, Antonietta Mira, and Christian P. Robert. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812, 2012.
- [18] Tiangang Cui, Sergey Dolgov, and Robert Scheichl. Deep importance sampling using tensor trains with application to a priori and a posteriori rare events. *SIAM Journal on Scientific Computing*, 46(1):C1–C29, 2024.
- [19] Claudia Czado. Analyzing dependent data with vine copulas. *Lecture Notes in Statistics, Springer*, 222, 2019.
- [20] Jyotishka Datta and Nicholas Polson. Inverse probability weighting: the missing link between survey sampling and evidence estimation. *arXiv preprint arXiv:2204.14121*, 2022.
- [21] George Deligiannidis, Arnaud Doucet, and Michael K Pitt. The correlated pseudomarginal method. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(5), 2018.
- [22] Luc Devroye. Nonuniform random variate generation. *Handbooks in operations research and management science*, 13:83–121, 2006.
- [23] Aymeric Dieuleveut, Gersende Fort, Eric Moulines, and Hoi-To Wai. Stochastic approximation beyond gradient for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 2023.
- [24] Randal Douc, Arnaud Guillin, J-M Marin, and Christian P Robert. Minimum variance importance sampling via population monte carlo. *ESAIM: Probability and Statistics*, 11:427–447, 2007.

- [25] Arnaud Doucet, Will Grathwohl, Alexander G Matthews, and Heiko Strathmann. Score-based diffusion meets annealed importance sampling. *Advances in Neural Information Processing Systems*, 35, 2022.
- [26] Yousef El-Laham, Víctor Elvira, and Mónica Bugallo. Recursive shrinkage covariance learning in adaptive importance sampling. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 624–628. IEEE, 2019.
- [27] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Generalized multiple importance sampling. *Statistical Science*, 34(1):129–155, 2019.
- [28] Víctor Elvira and Emilie Chouzenoux. Optimized population monte carlo. *IEEE Transactions on Signal Processing*, 70:2489–2501, 2022.
- [29] Víctor Elvira and Luca Martino. Advances in importance sampling. *arXiv preprint arXiv:2102.05407*, 2021.
- [30] Víctor Elvira, Luca Martino, David Luengo, and Mónica F Bugallo. Improving population monte carlo: Alternative weighting and resampling schemes. *Signal Processing*, 131:77–91, 2017.
- [31] Víctor Elvira, Luca Martino, and Christian P Robert. Rethinking the effective sample size. *International Statistical Review*, 90(3):525–550, 2022.
- [32] Michael Evans and Tim Swartz. Methods for approximating integrals in statistics with special emphasis on bayesian integration problems. *Statistical science*, 1995.
- [33] Yanqin Fan and Marc Henry. Vector copulas. *Journal of Econometrics*, 234(1), 2023.
- [34] Edwin Fong and Chris C Holmes. Conformal bayesian computation. *Advances in Neural Information Processing Systems*, 34:18268–18279, 2021.

- [35] Gabriel Frahm. *Generalized elliptical distributions: theory and applications*. PhD thesis, Universität zu Köln, 2004.
- [36] Lea Friedli, David Ginsbourger, Arnaud Doucet, and Niklas Linde. An energy-based model approach to rare event probability estimation. *arXiv preprint arXiv:2310.04082*, 2023.
- [37] Christian Genest, JJ Quesada Molina, and JA Rodríguez Lallena. De l'impossibilité de construire des lois à marges multidimensionnelles données à partir de copules. *Comptes rendus de l'Académie des sciences. Série 1, Mathématique*, 320(6), 1995.
- [38] John Geweke. Bayesian inference in econometric models using monte carlo integration. *Econometrica: Journal of the Econometric Society*, 1989.
- [39] Charles J Geyer. Estimating normalizing constants and reweighting mixtures. 1994.
- [40] Paul Glasserman. Performance continuity and differentiability in monte carlo optimization. In *Proceedings of the 20th conference on Winter simulation*, 1988.
- [41] Adam Golinski, Yee Whye Teh, Frank Wood, and Tom Rainforth. Amortized monte carlo integration. UAI Workshop on Uncertainty in Deep Learning, August 2018.
- [42] Ambuj Goyal, Philip Heidelberger, and Perwez Shahabuddin. Measure specific dynamic importance sampling for availability simulations. In *Proceedings of the 19th conference on Winter simulation*, 1987.
- [43] Thomas Guilmeau, Nicola Branchini, Emilie Chouzenoux, and Víctor Elvira. Adaptive importance sampling for heavy-tailed distributions via alpha-divergence minimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3871–3879. PMLR, 2024.
- [44] Jun Han and Qiang Liu. Stein variational adaptive importance sampling. In Gal Elidan, Kristian Kersting, and Alexander Ihler, editors, *Proceedings of the Thirty-Third Conference*

on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017.
AUAI Press, 2017.

- [45] Shane G Henderson and Samuel MT Ehrlichman. Sharpening comparisons via gaussian copulas and semidefinite programming. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 22(4), 2012.
- [46] Timothy Classen Hesterberg. *Advances in importance sampling*. PhD thesis, Stanford University, 1988.
- [47] Edward L Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2), 2008.
- [48] Nabil Kazi-Tani and Didier Rullière. On a construction of multivariate distributions given some multidimensional marginals. *Advances in Applied Probability*, 2019.
- [49] Samir Khan and Johan Ugander. Adaptive normalization for ipw estimation. *Journal of Causal Inference*, 11(1), 2023.
- [50] Teun Kloek and Herman K Van Dijk. Bayesian estimates of equation system parameters: an application of integration by monte carlo. *Econometrica: Journal of the Econometric Society*, 1978.
- [51] Oskar Kviman, Harald Melin, Hazal Koptagel, Victor Elvira, and Jens Lagergren. Multiple importance sampling elbo and deep ensembles of variational approximations. In *International Conference on Artificial Intelligence and Statistics*, pages 10687–10702. PMLR, 2022.
- [52] Roland Lamberti, Yohan Petetin, François Septier, and François Desbouvries. A double proposal normalized importance sampling estimator. In *2018 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, 2018.

- [53] F Llorente, L Martino, and D Delgado. Target-aware bayesian inference via generalized thermodynamic integration. *Computational Statistics*, 2023.
- [54] Fernando Llorente, Luca Martino, David Delgado, and Javier Lopez-Santiago. Marginal likelihood computation for model selection and hypothesis testing: an extensive review. *SIAM Review*, 65(1), 2023.
- [55] Luca Martino, Victor Elvira, David Luengo, and Jukka Corander. Layered adaptive importance sampling. *Statistics and Computing*, 27, 2017.
- [56] Xiao-Li Meng and Stephen Schilling. Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3), 2002.
- [57] Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 1996.
- [58] Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020.
- [59] Thomas Müller, Brian McWilliams, Fabrice Rousselle, Markus Gross, and Jan Novák. Neural importance sampling. *ACM Transactions on Graphics (ToG)*, 38(5):1–19, 2019.
- [60] Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11, 2001.
- [61] Geoffrey C. Orsak and Behnaam Aazhang. Constrained solutions in importance via robust statistics. *IEEE Transactions on Information Theory*, 37(2), 1991.
- [62] Art Owen and Yi Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449), 2000.
- [63] Art B. Owen. *Monte Carlo theory, methods and examples*. 2013.
- [64] Art B Owen and Yi Zhou. The square root rule for adaptive importance sampling. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 30(2), 2020.

- [65] Art B Owen et al. Importance sampling the union of rare events with an application to power systems analysis. *Electronic Journal of Statistics*, 2019.
- [66] Topi Paananen et al. Implicitly adaptive importance sampling. *Statistics and Computing*, 31(2), 2021.
- [67] Fengchun Peng and Dipak K Dey. Bayesian analysis of outlier problems using divergence measures. *Canadian Journal of Statistics*, 23(2):199–213, 1995.
- [68] Rick Picard and Brian Williams. Rare event estimation for computer models. *The American Statistician*, 67(1), 2013.
- [69] Tom Rainforth et al. Target-aware bayesian inference: how to beat optimal conventional estimators. *Journal of Machine Learning Research*, 2020.
- [70] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.
- [71] Tim Reichelt et al. Rethinking variational inference for probabilistic programs with stochastic support. In *Advances in Neural Information Processing Systems*, 2022.
- [72] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, 1951.
- [73] Reuven Y Rubinstein. Sensitivity analysis of discrete event systems by the “push out” method. *Annals of Operations Research*, 39(1), 1992.
- [74] RY Rubinstein and G Samorodnitsky. Variance reduction by the use of common and antithetic random variables. *Journal of Statistical Computation and Simulation*, 22(2):161–180, 1985.
- [75] Ernest K Ryu and Stephen P Boyd. Adaptive importance sampling via stochastic convex programming. *arXiv preprint arXiv:1412.4845*, 2014.

- [76] Daniel Sanz-Alonso. Importance sampling and necessary sample size: an information theory approach. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2), 2018.
- [77] Daniel Sanz-Alonso and Zijian Wang. Bayesian update with importance sampling: Required sample size. *Entropy*, 23(1), 2020.
- [78] Luca Alessandro Silva and Giacomo Zanella. Robust leave-one-out cross-validation for high-dimensional bayesian models. *Journal of the American Statistical Association*, pages 1–13, 2023.
- [79] Øivind Skare, Erik Bølviken, and Lars Holden. Improved sampling-importance resampling and reduced bias importance sampling. *Scandinavian Journal of Statistics*, 2003.
- [80] Zhiqiang Tan. On a likelihood approach for monte carlo integration. *Journal of the American Statistical Association*, 99(468):1027–1036, 2004.
- [81] Achille Thin, Yazid Janati El Idrissi, Sylvain Le Corff, Charles Ollion, Eric Moulines, Arnaud Doucet, Alain Durmus, and Christian X Robert. Neo: non equilibrium sampling on the orbits of a deterministic transform. *Advances in Neural Information Processing Systems*, 34, 2021.
- [82] Yehuda Vardi. Empirical distributions in selection bias models. *The Annals of Statistics*, pages 178–203, 1985.
- [83] Aki Vehtari et al. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 2017.
- [84] Giacomo Zanella and Gareth Roberts. Scalable importance tempering and bayesian variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(3), 2019.

Acronyms

AIS adaptive importance sampling. 1, 2, 4, 6, 10, 18, 26, 31

CDF cumulative distribution function. 4, 15, 17, 19

CRN common random number. 19

ESS effective sample size. 7

IS importance sampling. 1–3, 5–7, 14, 16, 22, 24, 25

MC Monte Carlo. 2, 5

MCMC Markov chain Monte Carlo. 2, 9, 10, 54

MSE mean squared error. 4–8, 10, 24, 54

MVN multivariate normal. 14–17, 19, 20, 27, 30, 44

PDF probability density function. 1, 2, 4–6, 8–13, 15–18, 23, 26–28, 44, 45, 50

SNIS self-normalized importance sampling. 1–4, 6–11, 21, 24–26, 30–32, 50

SVD singular value decomposition. 20, 27, 42

TABI target-aware Bayesian inference. 9, 22, 25, 26, 28, 30, 55

UIS unnormalized importance sampling. 5–8, 10, 11, 42

VI variational inference. 1, 6, 12, 18, 32, 42

VQ vector quantile. 15, 20, 23

A Unconstrained optimization of the coupling

We describe here details about how to perform unconstrained optimization for the parameter \mathbf{S}_c .

As discussed in §3.4, we parameterize \mathbf{S}_c via its SVD; so the parameters to be optimized are the orthogonal matrices \mathbf{U}, \mathbf{V} and the singular values, which are constrained in $[-1, 1]$. To reparameterize an orthogonal matrix, note that the matrix exponential of a skew-symmetric matrix is an orthogonal matrix. Therefore, we can take $\mathbf{U} = \exp_m(\mathbf{A} - \mathbf{A}^\top)$ where \mathbf{A} is a lower-diagonal matrix (with zeros on the diagonal) of unconstrained values. For the eigenvalues, we simply optimize an unconstrained vector $\mathbf{v} \in \mathbb{R}^{d_x}$ and obtain $\sigma = \tanh(\mathbf{v})$, which is differentiable.

Next, we briefly discuss gradient estimation. **Pathwise gradient.** We write the expectation in Eq. (22) as an expectation under a distribution independent of $\boldsymbol{\theta}_c$ using the push-out method/infinitesimal perturbation analysis/pathwise/reparameterization trick [73]. This involves finding noise variables $\mathbf{n}_1, \mathbf{n}_2 \sim \nu(\mathbf{n}_1, \mathbf{n}_2)$ such that $[\mathbf{u}_1, \mathbf{u}_2]^\top = [\mathbf{R}_1(\mathbf{n}_1; \boldsymbol{\theta}_c), \mathbf{R}_2(\mathbf{n}_2; \boldsymbol{\theta}_c)]^\top \sim c(\mathbf{u}_1, \mathbf{u}_2; \boldsymbol{\theta}_c)$, where $\mathbf{R}_i : \mathbb{R}^{d_x} \rightarrow [0, 1]^{d_x}$ parameterized by $\boldsymbol{\theta}_c$. Thus, we reformulate the objective as

$$\mathcal{C}(\boldsymbol{\theta}_c) = \mathbb{E}_{\nu(\mathbf{n}_1, \mathbf{n}_2)} [\log \omega_1(\mathbf{R}_1(\mathbf{n}_1; \boldsymbol{\theta}_c)) + \log \omega_2(\mathbf{R}_2(\mathbf{n}_2; \boldsymbol{\theta}_c))]. \quad (28)$$

(after a log-transformation that gives a lower bound, such as in VI) which allows one to push $\nabla_{\boldsymbol{\theta}_c}$ inside under a change of differentiation and expectation, holding when the expression in Eq. (28) satisfies technical conditions [40, Lemma 1], so that the gradient of $\mathcal{C}(\boldsymbol{\theta}_c)$ can be expressed as an expectation under $\nu(\mathbf{n}_1, \mathbf{n}_2)$ (see Appendix F for details). The pathwise gradient estimator is then

$$\hat{\nabla}_{\boldsymbol{\theta}_c}^{M, \text{Path}} \mathcal{C}(\boldsymbol{\theta}_c) \stackrel{\text{def}}{=} \frac{1}{M} \sum_{m=1}^M \nabla_{\boldsymbol{\theta}_c} \left[\log \omega_1(\mathbf{R}_1(\mathbf{n}_1^{(m)}; \boldsymbol{\theta}_c)) + \log \omega_2(\mathbf{R}_2(\mathbf{n}_2^{(m)}; \boldsymbol{\theta}_c)) \right], \quad (29)$$

where $[\mathbf{n}_1^{(m)}, \mathbf{n}_2^{(m)}]^\top \stackrel{\text{i.i.d.}}{\sim} \nu(\mathbf{n}_1, \mathbf{n}_2)$, $m = 1, \dots, M$.

Score function gradient. This approach corresponds to pushing the gradient inside the expectation in Eq. (22) and, since the resulting integral is not an expectation, applying UIS with proposal

$c_{1:2}$. The score function estimator is then

$$\hat{\nabla}_{\boldsymbol{\theta}_c}^{M,\text{Score}} \mathcal{C}(\boldsymbol{\theta}_c) \stackrel{\text{def}}{=} \frac{1}{M} \sum_{m=1}^M \log \omega(\mathbf{u}_1^{(m)}, \mathbf{u}_2^{(m)}) \nabla_{\boldsymbol{\theta}_c} \log c_{1:2}(\mathbf{u}_1^{(m)}, \mathbf{u}_2^{(m)}; \boldsymbol{\theta}_c), \quad (30)$$

where $[\mathbf{u}_1^{(m)}, \mathbf{u}_2^{(m)}]^\top \sim c_{1:2}(\mathbf{u}_1, \mathbf{u}_2; \boldsymbol{\theta}_c)$, $m = 1, \dots, M$.

B Connection with a continuous optimal transport problem

We now note that a simple rewriting of the relative asymptotic variance takes more clearly the form of an optimal transport problem.

Remark 4 (Connection with continuous optimal transport). *We can equivalently reformulate the task of finding an optimal joint as a Kantorovich optimal transport problem. Indeed, it holds*

$$\mu^{-2} \mathbb{V}_{\mathbb{Q}_{1:2}}^\infty [\hat{\mu}_{\text{GenSNIS}}^N] = \mathbb{E}_{\mathbb{Q}_{1:2}} \left[\left(\frac{q_1^\star(\mathbf{x}_1)}{q_1(\mathbf{x}_1)} - \frac{q_2^\star(\mathbf{x}_2)}{q_2(\mathbf{x}_2)} \right)^2 \right], \quad (31)$$

which is a Wasserstein distance minimization problem with the typical cost, squared Euclidean distance, but between the one-dimensional random variables $w_1(\mathbf{x}_1) = q_1^\star(\mathbf{x}_1)/q_1(\mathbf{x}_1)$, $w_2(\mathbf{x}_2) = q_2^\star(\mathbf{x}_2)/q_2(\mathbf{x}_2)$ instead of directly between $\mathbf{x}_1, \mathbf{x}_2$. The expectation can be written w.r.t. $\mathbb{Q}_{1:2}$, avoiding the need to consider the joint distribution of $w_1(\mathbf{x}_1), w_2(\mathbf{x}_2)$, by change of variable (in Statistics often referred to as “law of the unconscious statistician”).

C Closed form computation of the asymptotic variance

The asymptotic variance of $\hat{\mu}_{\text{GenSNIS}}^N$ can be computed in closed form for the Bayesian linear regression model. We have

$$\chi^2(\pi(\boldsymbol{\theta}_1 | \mathbf{X}, y_{1:S}, \mathbf{x}_{N+1}, y_{N+1}) || q_2(\boldsymbol{\theta}_2)) + \chi^2(\pi(\boldsymbol{\theta}_2 | \mathbf{X}, y_{1:S}) || q_2(\boldsymbol{\theta}_2)) - 2 \left(\mathbb{E}_{q_{1:2}} \left[\frac{\pi(\boldsymbol{\theta}_1 | \mathbf{X}, y_{1:S}, \mathbf{x}_{N+1}, y_{N+1}) \pi(\boldsymbol{\theta}_2 | \mathbf{X}, y_{1:S})}{q_1(\boldsymbol{\theta}_1) q_2(\boldsymbol{\theta}_2)} \right] - 1 \right) \quad (32)$$

Recalling that

$$\chi^2(\mathcal{N}(\boldsymbol{\mu}_A, \Sigma_A) || \mathcal{N}(\boldsymbol{\mu}_B, \Sigma_B)) = \frac{|\Sigma_B|}{\sqrt{|2\Sigma_B - \Sigma_A| \cdot |\Sigma_A|}} \exp \{ (\boldsymbol{\mu}_A - \boldsymbol{\mu}_B)^\top (2\Sigma_B - \Sigma_A)^{-1} (\boldsymbol{\mu}_A - \boldsymbol{\mu}_B) \} - 1 \quad (33)$$

allows us to compute the first two terms. We notice that in the third term, we have a product of two ratios between multivariate normal PDFs,

$$\mathbb{E}_{q_{1:2}} \left[\frac{\pi(\boldsymbol{\theta}_1 | \mathbf{X}, y_{1:S}, \mathbf{x}_{N+1}, y_{N+1}) \pi(\boldsymbol{\theta}_2 | \mathbf{X}, y_{1:S})}{q_1(\boldsymbol{\theta}_1) q_2(\boldsymbol{\theta}_2)} \right] = \mathbb{E}_{q_{1:2}} [r_1(\boldsymbol{\theta}_1) r_2(\boldsymbol{\theta}_2)]. \quad (34)$$

A ratio of two MVNs is proportional to a MVN, with constants given as per Appendix E. Denote these as C_1, C_2 . Let the PDFs in the numerator be $\mathcal{N}_{\boldsymbol{\theta}_1}(\boldsymbol{\mu}_{q_1\star}, \boldsymbol{\Sigma}_{q_1\star})$ and $\mathcal{N}_{\boldsymbol{\theta}_2}(\boldsymbol{\mu}_{q_2\star}, \boldsymbol{\Sigma}_{q_2\star})$. Then we have

$$\mathbb{E}_{q_{1:2}} [r_1(\boldsymbol{\theta}_1) r_2(\boldsymbol{\theta}_2)] = C_1 \cdot C_2 \cdot \mathbb{E}_{q_{1:2}} [\mathcal{N}_{\boldsymbol{\theta}_1}(\mathbf{r}_1, \mathbf{R}_1) \mathcal{N}_{\boldsymbol{\theta}_2}(\mathbf{r}_2, \mathbf{R}_2)] \quad (35)$$

where

$$\begin{cases} \mathbf{R}_1 = (\boldsymbol{\Sigma}_{q_1\star}^{-1} - \boldsymbol{\Sigma}_{q_1}^{-1})^{-1}, & \mathbf{r}_1 = \mathbf{R}_1 (\boldsymbol{\Sigma}_{q_1\star}^{-1} \boldsymbol{\mu}_{q_1\star} - \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_{q_1}) \\ \mathbf{R}_2 = (\boldsymbol{\Sigma}_{q_2\star}^{-1} - \boldsymbol{\Sigma}_{q_2}^{-1})^{-1}, & \mathbf{r}_2 = \mathbf{R}_2 (\boldsymbol{\Sigma}_{q_2\star}^{-1} \boldsymbol{\mu}_{q_2\star} - \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_{q_2}) \end{cases} \quad (36)$$

Let $\mathbf{r} = [\mathbf{r}_1, \mathbf{r}_2]$ and \mathbf{R} as

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_2 \end{bmatrix}. \quad (37)$$

Then we can show

$$\mathbb{E}_{q_{1:2}} [\mathcal{N}_{\boldsymbol{\theta}_1}(\mathbf{r}_1, \mathbf{R}_1) \mathcal{N}_{\boldsymbol{\theta}_2}(\mathbf{r}_2, \mathbf{R}_2)] = \int \mathcal{N}_{\boldsymbol{\theta}}([\mathbf{r}_1, \mathbf{r}_2]) \mathcal{N}_{\boldsymbol{\theta}}([\boldsymbol{\mu}_1, \boldsymbol{\mu}_2])^\top, \begin{bmatrix} \boldsymbol{\Sigma}_{q_1} & \mathbf{L}_{q_2} \mathbf{S}_c \mathbf{L}_{q_1}^\top \\ \mathbf{S}_c^\top & \mathbf{L}_{q_1} \boldsymbol{\Sigma}_{q_2} \mathbf{L}_{q_2}^\top \end{bmatrix}) d\boldsymbol{\theta} \quad (38)$$

$$= C_3 = \mathcal{N}_{\mathbf{r}}(\boldsymbol{\mu}_{q_{1:2}}, \mathbf{R} + \boldsymbol{\Sigma}_{q_{1:2}}). \quad (39)$$

D Product of multivariate normal PDFs

Let $\mathcal{N}_x(\mathbf{m}, \Sigma)$ denote a density of \mathbf{x} , then

$$\mathcal{N}_x(\mathbf{m}_1, \Sigma_1) \cdot \mathcal{N}_x(\mathbf{m}_2, \Sigma_2) = c_c \mathcal{N}_x(\mathbf{m}_c, \Sigma_c) \quad (40)$$

$$c_c = \mathcal{N}_{\mathbf{m}_1}(\mathbf{m}_2, (\Sigma_1 + \Sigma_2)) \quad (41)$$

$$= \frac{1}{\sqrt{\det(2\pi(\Sigma_1 + \Sigma_2))}} \exp \left[-\frac{1}{2} (\mathbf{m}_1 - \mathbf{m}_2)^T (\Sigma_1 + \Sigma_2)^{-1} (\mathbf{m}_1 - \mathbf{m}_2) \right] \quad (42)$$

$$\mathbf{m}_c = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1} (\Sigma_1^{-1} \mathbf{m}_1 + \Sigma_2^{-1} \mathbf{m}_2) \quad (43)$$

$$\Sigma_c = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}. \quad (44)$$

E Ratio of multivariate normal PDFs

Ratio of $\mathcal{N}_x(\boldsymbol{\mu}_1, \Sigma_1)$ and $\mathcal{N}_x(\boldsymbol{\mu}_2, \Sigma_2)$

$$\frac{\mathcal{N}_x(\boldsymbol{\mu}_1, \Sigma_1)}{\mathcal{N}_x(\boldsymbol{\mu}_2, \Sigma_2)} = \frac{\det \Sigma_1^{1/2}}{\det \Sigma_2^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_1)^\top \Sigma_1^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) + \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_2)^\top \Sigma_2^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) \right\} \quad (45)$$

$$= \frac{\det \Sigma_2^{1/2}}{\det \Sigma_1^{1/2}} \exp \left\{ -\frac{1}{2} \left(\mathbf{x}^\top (\Sigma_1^{-1} - \Sigma_2) \mathbf{x} - 2\mathbf{x}^\top (\Sigma_1^{-1} \boldsymbol{\mu}_1 - \Sigma_2^{-1} \boldsymbol{\mu}_2) + \underbrace{\boldsymbol{\mu}_1^\top \Sigma_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top \Sigma_2^{-1} \boldsymbol{\mu}_2}_{\text{constant terms}} \right) \right\} \quad (46)$$

Note the signs in the expression above. The aim is to relate the above to $\mathcal{N}_x(\boldsymbol{\mu}_3, \Sigma_3)$ where

$\Sigma_3 = (\Sigma_1^{-1} - \Sigma_2^{-1})^{-1}$, $\boldsymbol{\mu}_3 = \Sigma_3(\Sigma_1^{-1} \boldsymbol{\mu}_1 - \Sigma_2^{-1} \boldsymbol{\mu}_2)$. Let us expand $\mathcal{N}_x(\boldsymbol{\mu}_3, \Sigma_3)$ first, which gives

$$(2\pi)^{-d_x/2} \det(\Sigma_3)^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_3)^\top \Sigma_3^{-1} (\mathbf{x} - \boldsymbol{\mu}_3) \right\} \quad (47)$$

$$= (2\pi)^{-d_x/2} \det(\Sigma_3)^{-1/2} \exp \left\{ -\frac{1}{2} \left(\mathbf{x}^\top (\Sigma_1^{-1} - \Sigma_2^{-1}) \mathbf{x} - 2\mathbf{x}^\top (\Sigma_1^{-1} \boldsymbol{\mu}_1 - \Sigma_2^{-1} \boldsymbol{\mu}_2) + \boldsymbol{\mu}_3^\top \Sigma_3^{-1} \boldsymbol{\mu}_3 \right) \right\} \quad (48)$$

$$- 2\mathbf{x}^\top (\Sigma_1^{-1} - \Sigma_2^{-1})(\Sigma_1^{-1} - \Sigma_2^{-1})^{-1} (\Sigma_1^{-1} \boldsymbol{\mu}_1 - \Sigma_2^{-1} \boldsymbol{\mu}_2) \quad (49)$$

$$+ \boldsymbol{\mu}_3^\top \Sigma_3^{-1} \boldsymbol{\mu}_3 \right\} \quad (50)$$

where we already see that the first two terms inside \exp correspond to the first two terms of Eq. (45). Expanding the last term gives the following

$$\boldsymbol{\mu}_3^\top \boldsymbol{\Sigma}_3^{-1} \boldsymbol{\mu}_3 = (\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_3^\top \boldsymbol{\Sigma}_3^{-1} \boldsymbol{\Sigma}_3 (\boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2) \quad (51)$$

$$= \underbrace{\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_3 \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1}_{\textcircled{1}} - \underbrace{2\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_3 \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2}_{\textcircled{3}} + \underbrace{\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_3 \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2}_{\textcircled{2}} \quad (52)$$

Before proceeding recall the Woodbury matrix identity, that for any invertible A, C states the following useful identities

$$(A + C)^{-1} = A^{-1} - A^{-1}(C^{-1} + A^{-1})^{-1}A^{-1} \quad (53)$$

$$= (C + A)^{-1} = C^{-1} - C^{-1}(A^{-1} + C^{-1})C^{-1} \quad (54)$$

$$(55)$$

We deduce that

$$\boldsymbol{\Sigma}_3 = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)^{-1}\boldsymbol{\Sigma}_1 \quad (56)$$

Then we can write

$$\textcircled{1} = \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)^{-1}) \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 \quad (57)$$

$$= \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^\top (\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)^{-1} \boldsymbol{\mu}_1 \quad (58)$$

and analogously

$$\textcircled{2} = -\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2 + \boldsymbol{\mu}_2^\top (\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)^{-1} \boldsymbol{\mu}_2 \quad (59)$$

Note that the first terms of the above match the constants of Eq. (45).

$$\textcircled{3} = -2\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})^{-1} \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2 \quad (60)$$

$$= -2\boldsymbol{\mu}_1^\top (\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_2)^{-1} \boldsymbol{\mu}_2 \quad (61)$$

$$\mathcal{N}_{\mu_1}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1) = (2\pi)^{-d_{\boldsymbol{x}}/2} \det(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)^{-1/2} \quad (62)$$

$$\times \exp \left\{ -\frac{1}{2} \boldsymbol{\mu}_1^\top (\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)^{-1} \boldsymbol{\mu}_1 - 2 \boldsymbol{\mu}_1^\top (\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)^{-1} \boldsymbol{\mu}_2 + \boldsymbol{\mu}_2^\top (\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)^{-1} \boldsymbol{\mu}_2 \right\} \quad (63)$$

$$\frac{\mathcal{N}_{\boldsymbol{x}}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)}{\mathcal{N}_{\boldsymbol{x}}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)} = \sqrt{\frac{\det \boldsymbol{\Sigma}_2}{\det \boldsymbol{\Sigma}_1} \cdot \frac{\det \boldsymbol{\Sigma}_3}{\det(\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)}} \cdot \frac{\mathcal{N}_{\boldsymbol{x}}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)}{\mathcal{N}_{\boldsymbol{\mu}_1}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1)} \quad (64)$$

F Derivation of variance of GenSNIS

From Appendix H, we know

$$\mathbb{V}_{\mathbb{Q}_{1:2}}^{\delta}[\hat{\mu}_{\text{GenSNIS}}^N] = \frac{1}{Z_p^2} \mathbb{V}_{q_1}[\hat{I}^N] + \frac{\mu^2}{Z_p^2} \mathbb{V}_{q_2}[\hat{Z}_p^N] - \frac{2\mu}{Z_p^2} \cdot \text{Cov}_{\mathbb{Q}_{1:2}}[\hat{I}^N, \hat{Z}_p^N]. \quad (65)$$

We will expand each of the three terms. First we consider the case where $f(\boldsymbol{x}) \geq 0$.

$$\begin{aligned} \frac{1}{Z_p^2} \mathbb{V}_{q_1}[\hat{I}^N] &= \frac{1}{Z_p^2} \mathbb{V} \left[\frac{1}{N} \sum_{n=1}^N \frac{f(\boldsymbol{x}_1^{(n)}) \tilde{p}(\boldsymbol{x}_1^{(n)})}{q_1(\boldsymbol{x}_1^{(n)})} \right] = \frac{1}{Z_p^2 N} \left(\int \frac{(f(\boldsymbol{x}_1) \tilde{p}(\boldsymbol{x}_1))^2}{q_1(\boldsymbol{x}_1)} d\boldsymbol{x}_1 - I^2 \right) \\ &= \frac{1}{N} \left(\int \frac{(f(\boldsymbol{x}_1) p(\boldsymbol{x}_1))^2}{q_1(\boldsymbol{x}_1)} d\boldsymbol{x}_1 - \mu^2 \right) = \frac{1}{N} \left(\mu^2 \int \frac{(f(\boldsymbol{x}_1) p(\boldsymbol{x}_1))^2}{\mu^2 q_1(\boldsymbol{x}_1)} d\boldsymbol{x}_1 - \mu^2 \right) \\ &= \frac{\mu^2}{N} \chi^2(p \cdot f || q_1) \end{aligned}$$

Next, the case where f takes positive and negative values.

$$\begin{aligned} \frac{1}{Z_p^2} \mathbb{V}_{q_1}[\hat{I}^N] &= \frac{1}{Z_p^2} \mathbb{V} \left[\frac{1}{N} \sum_{n=1}^N \frac{f(\boldsymbol{x}_1^{(n)}) \tilde{p}(\boldsymbol{x}_1^{(n)})}{q_1(\boldsymbol{x}_1^{(n)})} \right] = \frac{1}{Z_p^2 N} \left(\int \frac{(f(\boldsymbol{x}_1) \tilde{p}(\boldsymbol{x}_1))^2}{q_1(\boldsymbol{x}_1)} d\boldsymbol{x}_1 - I^2 \right) \\ &= \frac{1}{N} \left(\int \frac{(f(\boldsymbol{x}_1) p(\boldsymbol{x}_1))^2}{q_1(\boldsymbol{x}_1)} d\boldsymbol{x}_1 - \mu^2 \right) = \frac{1}{N} \left(\int \frac{[(f_+(\boldsymbol{x}_1) - f_-(\boldsymbol{x}_1)) p(\boldsymbol{x}_1)]^2}{q_1(\boldsymbol{x}_1)} d\boldsymbol{x}_1 - \mu^2 \right) \\ &= \frac{1}{N} \left(\int \frac{(f_+^2(\boldsymbol{x}_1) + f_-^2(\boldsymbol{x}_1) - 2f_+(\boldsymbol{x}_1)f_-(\boldsymbol{x}_1)) p^2(\boldsymbol{x}_1)}{q_1(\boldsymbol{x}_1)} d\boldsymbol{x}_1 - \mu \right) \\ &= \frac{1}{N} \left(\int \frac{(f_+(\boldsymbol{x}_1) p(\boldsymbol{x}_1))^2}{q_1(\boldsymbol{x}_1)} d\boldsymbol{x}_1 + \int \frac{(f_-(\boldsymbol{x}_1) p(\boldsymbol{x}_1))^2}{q_1(\boldsymbol{x}_1)} d\boldsymbol{x}_1 - 2 \int \frac{f_+(\boldsymbol{x}_1) f_-(\boldsymbol{x}_1) p^2(\boldsymbol{x}_1)}{q_1(\boldsymbol{x}_1)} d\boldsymbol{x}_1 - \mu \right) \\ &= \frac{1}{N} (\mu_+^2 \cdot \chi^2(p \cdot f_+ || q_1) + \mu_+^2 + \mu_-^2 \cdot \chi^2(p \cdot f_- || q_1) + \mu_-^2 - \mu) \end{aligned}$$

The second term:

$$\begin{aligned} \frac{\mu^2}{Z_p^2} \mathbb{V}_{q_2}[\hat{Z}_p^N] &= \frac{\mu^2}{Z_p^2} \mathbb{V} \left[\frac{1}{N} \sum_{n=1}^N \frac{\tilde{p}(\boldsymbol{x}_2^{(n)})}{q_2(\boldsymbol{x}_2^{(n)})} \right] = \mu^2 \mathbb{V} \left[\frac{1}{N} \sum_{n=1}^N \frac{p(\boldsymbol{x}_2^{(n)})}{q_2(\boldsymbol{x}_2^{(n)})} \right] \\ &= \frac{\mu^2}{N} (\chi^2(p || q_2)) \end{aligned}$$

The third term (recalling $\mathbb{E}_{\mathbb{Q}_{1:2}}[\hat{I}^N] = \mathbb{E}_{q_1}[\hat{I}^N] = I$ and $\mathbb{E}_{\mathbb{Q}_{1:2}}[\hat{Z}_p^N] = \mathbb{E}_{q_2}[\hat{Z}_p^N] = Z_p$):

$$\begin{aligned} \frac{2\mu}{Z_p^2} \cdot \text{Cov}_{\mathbb{Q}_{1:2}}[\hat{I}^N, \hat{Z}_p^N] &= \frac{2\mu}{Z_p^2} \left(\mathbb{E}_{\mathbb{Q}_{1:2}}[\hat{I}^N \cdot \hat{Z}_p^N] - \mathbb{E}_{\mathbb{Q}_{1:2}}[\hat{I}^N \cdot Z_p] - \mathbb{E}_{\mathbb{Q}_{1:2}}[I \cdot \hat{Z}_p^N] + IZ_p \right) \\ &= \underbrace{\frac{2\mu}{Z_p^2} \cdot \mathbb{E}_{\mathbb{Q}_{1:2}}[\hat{I}^N \cdot \hat{Z}_p^N]}_{\stackrel{\text{def}}{=} (\text{T})} - 2\mu^2 \end{aligned}$$

Expanding (T) :

$$\begin{aligned} (\text{T}) &= \frac{2\mu}{Z_p^2} \cdot \left(\mathbb{E}_{\mathbb{Q}_{1:2}} \left[\left(\frac{1}{N} \sum_{n=1}^N \frac{f(\mathbf{x}_1) \tilde{p}(\mathbf{x}_1^{(n)})}{q_1(\mathbf{x}_1^{(n)})} \right) \left(\frac{1}{N} \sum_{n'=1}^N \frac{\tilde{p}(\mathbf{x}_2^{(n')})}{q_2(\mathbf{x}_2^{(n')})} \right) \right] \right) \\ &= \frac{2\mu}{N^2} \mathbb{E}_{\mathbb{Q}_{1:2}} \left[\left(\sum_{n=1}^N \frac{f(\mathbf{x}_1^{(n)}) p(\mathbf{x}_1^{(n)})}{q_1(\mathbf{x}_1^{(n)})} \right) \left(\sum_{n'=1}^N \frac{p(\mathbf{x}_2^{(n')})}{q_2(\mathbf{x}_2^{(n')})} \right) \right] \\ &= \frac{2\mu}{N^2} \sum_{n=1}^N \sum_{n'=1}^N \mathbb{E}_{\mathbb{Q}_{1:2}} \left[\frac{f(\mathbf{x}_1^{(n)}) p(\mathbf{x}_1^{(n)})}{q_1(\mathbf{x}_1^{(n)})} \cdot \frac{p(\mathbf{x}_2^{(n')})}{q_2(\mathbf{x}_2^{(n')})} \right] \\ &= \frac{2\mu}{N^2} \left(\sum_{n=n'} \mathbb{E}_{\mathbb{Q}_{1:2}} \left[\frac{f(\mathbf{x}_1^{(n)}) p(\mathbf{x}_1^{(n)})}{q_1(\mathbf{x}_1^{(n)})} \cdot \frac{p(\mathbf{x}_2^{(n')})}{q_2(\mathbf{x}_2^{(n')})} \right] + \underbrace{\sum_{n \neq n'} \mathbb{E}_{q_1} \left[\frac{f(\mathbf{x}_1^{(n)}) p(\mathbf{x}_1^{(n)})}{q_1(\mathbf{x}_1^{(n)})} \right]}_{=\mu} \underbrace{\mathbb{E}_{q_2} \left[\frac{p(\mathbf{x}_2^{(n')})}{q_2(\mathbf{x}_2^{(n')})} \right]}_{=1} \right) \\ &= \frac{2\mu}{N^2} \left(N \cdot \mathbb{E}_{\mathbb{Q}_{1:2}} \left[\frac{f(\mathbf{x}_1) p(\mathbf{x}_1)}{q_1(\mathbf{x}_1)} \cdot \frac{p(\mathbf{x}_2)}{q_2(\mathbf{x}_2)} \right] + N(N-1) \cdot \mu \right) \\ &= \frac{2\mu}{N} \cdot \mathbb{E}_{\mathbb{Q}_{1:2}} \left[\frac{f(\mathbf{x}_1) p(\mathbf{x}_1)}{q_1(\mathbf{x}_1)} \cdot \frac{p(\mathbf{x}_2)}{q_2(\mathbf{x}_2)} \right] + \frac{N-1}{N} \cdot 2\mu^2 \end{aligned}$$

Therefore, the third term becomes

$$\frac{2\mu}{Z_p^2} \cdot \text{Cov}_{\mathbb{Q}_{1:2}}[\hat{I}^N, \hat{Z}_p^N] = \frac{2\mu}{N} \cdot \mathbb{E}_{\mathbb{Q}_{1:2}} \left[\frac{f(\mathbf{x}_1) p(\mathbf{x}_1)}{q_1(\mathbf{x}_1)} \cdot \frac{p(\mathbf{x}_2)}{q_2(\mathbf{x}_2)} \right] + \frac{N-1}{N} \cdot 2\mu^2 - 2\mu^2 \quad (66)$$

$$= \frac{2\mu}{N} \cdot \mathbb{E}_{\mathbb{Q}_{1:2}} \left[\frac{f(\mathbf{x}_1) p(\mathbf{x}_1)}{q_1(\mathbf{x}_1)} \cdot \frac{p(\mathbf{x}_2)}{q_2(\mathbf{x}_2)} \right] - \frac{2\mu^2}{N} \quad (67)$$

$$= \frac{2\mu}{N} \left(\underbrace{\mathbb{E}_{\mathbb{Q}_{1:2}} \left[\frac{f(\mathbf{x}_1) p(\mathbf{x}_1)}{q_1(\mathbf{x}_1)} \cdot \frac{p(\mathbf{x}_2)}{q_2(\mathbf{x}_2)} \right]}_{(\text{H})} - \mu \right) \quad (68)$$

For the case where f can take negative values, we can now expand (H) as:

$$\textcircled{(H)} = \mathbb{E}_{\mathbb{Q}_{1:2}} \left[\frac{(f_+(\mathbf{x}_1) - f_-(\mathbf{x}_1))p(\mathbf{x}_1)}{q_1(\mathbf{x}_1)} \cdot \frac{p(\mathbf{x}_2)}{q_2(\mathbf{x}_2)} \right] \quad (69)$$

$$= \underbrace{\mathbb{E}_{\mathbb{Q}_{1:2}} \left[\frac{f_+(\mathbf{x}_1)p(\mathbf{x}_1)}{q_1(\mathbf{x}_1)} \cdot \frac{p(\mathbf{x}_2)}{q_2(\mathbf{x}_2)} \right]}_{\textcircled{H1}} - \underbrace{\mathbb{E}_{\mathbb{Q}_{1:2}} \left[\frac{f_-(\mathbf{x}_1)p(\mathbf{x}_1)}{q_1(\mathbf{x}_1)} \cdot \frac{p(\mathbf{x}_2)}{q_2(\mathbf{x}_2)} \right]}_{\textcircled{H2}} \quad (70)$$

G Derivation of q_{SNIS}^*

For convenience, let us restate

$$\hat{\mu}_{\text{SNIS}}^N \stackrel{\text{def}}{=} \frac{\sum_{n=1}^N \frac{f(\mathbf{x}^{(n)})\tilde{p}(\mathbf{x}^{(n)})}{q(\mathbf{x}^{(n)})}}{\sum_{n=1}^N \frac{\tilde{p}(\mathbf{x}^{(n)})}{q(\mathbf{x}^{(n)})}}, \quad (71)$$

and:

$$\mathbb{V}_q[\hat{\mu}_{\text{SNIS}}^N] \stackrel{\text{def}}{=} \frac{1}{Z_p^2} \mathbb{V}_q[\hat{I}^N] + \frac{\mu^2}{Z_p^2} \mathbb{V}_q[\hat{Z}_p^N] - \frac{2\mu}{Z_p^2} \cdot \text{Cov}_q[\hat{I}^N, \hat{Z}_p^N]. \quad (72)$$

The aim is to find the function $q(\mathbf{x})$ that minimizes the expression above, which is a functional of q , with the constraint that it is a valid pdf. To do this, we will take the functional derivative of Eq. (72) and equate to zero.

First, recalling that $\hat{I}^N = \frac{1}{N} \sum_{n=1}^N \frac{f(\mathbf{x}^{(n)})\tilde{p}(\mathbf{x}^{(n)})}{q(\mathbf{x}^{(n)})}$ and $\hat{Z}_p^N = \frac{1}{N} \sum_{n=1}^N \frac{\tilde{p}(\mathbf{x}^{(n)})}{q(\mathbf{x}^{(n)})}$, let us rewrite Eq. (72) more explicitly as:

$$\begin{aligned} \mathcal{F}[q] &\stackrel{\text{def}}{=} \frac{1}{N} \underbrace{\left(\int \frac{(f(\mathbf{x})p(\mathbf{x}))^2}{q(\mathbf{x})} d\mathbf{x} - \mu^2 \right)}_{\mathcal{F}_1[q]} + \frac{\mu^2}{N} \underbrace{\left(\int \frac{p(\mathbf{x})^2}{q(\mathbf{x})} d\mathbf{x} - 1 \right)}_{\mathcal{F}_2[q]} \\ &\quad - 2\mu \underbrace{\left(\frac{1}{N} \int \frac{p(\mathbf{x})^2}{q(\mathbf{x})} \cdot f(\mathbf{x}) d\mathbf{x} - \frac{\mu}{N-1} \right)}_{\mathcal{F}_3[q]}. \end{aligned} \quad (73)$$

We now notice that Eq. (73) is a sum of three terms, each of which is a functional of q in a so-called *integral form*. When looking at an infinitesimal change in the functional $\mathcal{F}[q]$, by adding a multiple of some arbitrary test function $\nu(\mathbf{x})$ to q , the functional derivative (first variation) $\frac{\delta \mathcal{F}[q]}{\delta q}$ is defined as:

$$\frac{\delta \mathcal{F}[q]}{\delta q} \stackrel{\text{def}}{=} \lim_{\epsilon \rightarrow 0} \frac{\mathcal{F}[q + \epsilon \nu] - \mathcal{F}[q]}{\epsilon}, \quad (74)$$

which exists under some regularity conditions. It can be shown that many standard rules for derivatives from calculus in \mathbb{R} hold for functional derivatives also. In particular, our functionals $\mathcal{F}_1[q], \mathcal{F}_2[q], \mathcal{F}_3[q]$ obtain a particularly tractable form, that is they are of integral form. Also let $\mathcal{L}[q] = \lambda \cdot (\int q(\mathbf{x}) d\mathbf{x} - 1)$. The functional derivative of our functional of interest is then:

$$\frac{\delta \mathcal{F}[q]}{\delta q} = \frac{\delta \mathcal{F}_1[q]}{\delta q} + \frac{\delta \mathcal{F}_2[q]}{\delta q} + \frac{\delta \mathcal{F}_3[q]}{\delta q} + \frac{\delta \mathcal{L}[q]}{\delta q} \quad (75)$$

$$= -\frac{1}{N} \frac{(f(\mathbf{x})p(\mathbf{x}))^2}{q(\mathbf{x})^2} - \frac{1}{N} \cdot \mu^2 \cdot \frac{p(\mathbf{x})^2}{q(\mathbf{x})^2} + \frac{1}{N} \cdot 2\mu \cdot \frac{p(\mathbf{x})^2 f(\mathbf{x})}{q(\mathbf{x})^2} + \lambda, \quad (76)$$

where we also added a positive Lagrangian multiplier condition to ensure that $q(\mathbf{x})$ integrates to 1 (for $\lambda > 0$). Now equating the above to 0, cancelling all $1/N$ and changing sign gives:

$$\frac{\delta \mathcal{F}[q]}{\delta q} = 0 \quad (77)$$

$$\Rightarrow \frac{p(\mathbf{x})^2(f(\mathbf{x})^2 + \mu^2 - 2\mu f(\mathbf{x}))}{q(\mathbf{x})^2} = \lambda \quad (78)$$

$$\Rightarrow \lambda \cdot q(\mathbf{x})^2 = p(\mathbf{x})^2(f(\mathbf{x}) - \mu)^2 \quad (79)$$

$$\Rightarrow q(\mathbf{x}) = \sqrt{\frac{p(\mathbf{x})^2(f(\mathbf{x}) - \mu)^2}{\lambda}} \quad (80)$$

which gives the desired $q_{\text{SNIS}}^\star \propto p(\mathbf{x})|f(\mathbf{x}) - \mu|$ (note that the solution is always a positive function).

Rather than turn to complicated second order conditions, we can verify that this is the optimal solution via the standard Cauchy-Swartz inequality. However, with this derivation we at least did not have to “guess” the solution beforehand entirely. This derivation extends that of the optimal q for the UIS estimator in [63, Chapter 9] to SNIS.

G.1 Proof of SNIS asymptotic variance lower bound

Here we give a proof of the lower bound on the asymptotic variance of SNIS given in §2.3. Recall $q_{\text{SNIS}}^\star \propto p(\mathbf{x})|f(\mathbf{x}) - \mu|$. For any PDF $q(\mathbf{x})$, recall that the asymptotic variance of $\hat{\mu}_{\text{SNIS}}^N$ can be written as:

$$\mathbb{V}_q^\infty[\hat{\mu}_{\text{SNIS}}^N] = \int \frac{p(\mathbf{x})^2(f(\mathbf{x}) - \mu)^2}{q(\mathbf{x})} d\mathbf{x}. \quad (81)$$

We showed in the previous subsection that q_{SNIS}^* is the minimizer of the asymptotic variance.

Plugging in q_{SNIS}^* into the above gives:

$$\mathbb{V}_{q_{\text{SNIS}}^*}^\infty [\hat{\mu}_{\text{SNIS}}^N] = \int \frac{p(\mathbf{x})^2(f(\mathbf{x}) - \mu)^2}{p(\mathbf{x})|f(\mathbf{x}) - \mu|} \cdot \left(\int p(\mathbf{x})|f(\mathbf{x}) - \mu|d\mathbf{x} \right) d\mathbf{x} \quad (82)$$

$$= \int p(\mathbf{x})|f(\mathbf{x}) - \mu|d\mathbf{x} \cdot \int p(\mathbf{x})|f(\mathbf{x}) - \mu|d\mathbf{x} \quad (83)$$

$$= \left(\int p(\mathbf{x})|f(\mathbf{x}) - \mu|d\mathbf{x} \right)^2. \quad \square \quad (84)$$

H Delta method MSE of $\hat{\mu}_{\text{SNIS}}^N$ (and $\hat{\mu}_{\text{GenSNIS}}^N$)

The MSE of $\hat{\mu}_{\text{SNIS}}$ can be analysed via the delta method, which obtains a Taylor approximation of 2nd order for the bias, and a 1st order one for the variance. It is possible to prove that the 1st order Taylor approximation of the variance coincides with the asymptotic variance, which is what we use in the main paper.

Recall $\hat{I}^N \stackrel{\text{def}}{=} \frac{1}{N} \sum_n \frac{f(\mathbf{x}^{(n)})\tilde{p}(\mathbf{x}^{(n)})}{q(\mathbf{x}^{(n)})}$, $\hat{Z}_p^N \stackrel{\text{def}}{=} \frac{1}{N} \sum_n \frac{\tilde{p}(\mathbf{x}^{(n)})}{q(\mathbf{x}^{(n)})}$ and that these are unbiased estimators. Note that $\hat{\mu}_{\text{SNIS}}^N = g(\hat{I}^N, \hat{Z}_p^N)$ with g being the ratio function, \hat{I}^N/\hat{Z}_p^N . The delta method defines first a (here, second-order) Taylor expansion of $g(\hat{I}, \hat{Z}_p)$ around the two-dimensional $[I, Z_p]^\top \in \mathbb{R}_{\geq 0}^2$:

$$g(\hat{I}^N, \hat{Z}_p^N) \approx g(I, Z_p) + \frac{\partial g(I, Z_p)}{\partial \hat{I}^N} (\hat{I}^N - I) + \frac{\partial g(I, Z_p)}{\partial \hat{Z}_p^N} (\hat{Z}_p^N - Z_p) \quad (85)$$

$$+ \frac{1}{2} \left(\frac{\partial g(I, Z_p)}{\partial \hat{I}^N \partial \hat{I}^N} (\hat{I}^N - I)^2 + 2 \cdot \frac{\partial g(I, Z_p)}{\partial \hat{I}^N \partial \hat{Z}_p^N} (\hat{I}^N - I)(\hat{Z}_p^N - Z_p) + \frac{\partial g(I, Z_p)}{\partial \hat{Z}_p^N \partial \hat{Z}_p^N} (\hat{Z}_p^N - Z_p)^2 \right) \quad (86)$$

$$\stackrel{\text{def}}{=} \hat{g}_2(I, Z_p) \quad (87)$$

where the 2 in the subscript of \hat{g}_2 denotes the order of the Taylor approximation. Here, $\frac{\partial g(I, Z_p)}{\partial \hat{I}^N}$ denotes derivative of $g(\hat{I}^N, \hat{Z}_p^N)$ w.r.t \hat{I}^N evaluated at I and Z_p .

Notice that:

$$g(\hat{I}, \hat{Z}_p) = \hat{g}_2(I, Z_p) + \mathcal{O}((\hat{I}^N - I)^3) + \mathcal{O}((\hat{Z}_p^N - Z_p)^3) \quad (88)$$

$$g(\hat{I}, \hat{Z}_p) = \hat{g}_1(I, Z_p) + \mathcal{O}((\hat{I}^N - I)^2) + \mathcal{O}((\hat{Z}_p^N - Z_p)^2). \quad (89)$$

The big-Oh notation ignores factors that are constant in the differences $(\hat{I}^N - I)$ and $(\hat{Z}_p^N - Z_p)$.

Also remember that estimators \hat{I}^N and \hat{Z}_p^N depend on N and share the same samples $\{\mathbf{x}^{(n)}\}_{n=1}^N$.

Recall that here I and Z_p are constants and only \hat{I}^N and \hat{Z}_p^N are r.v.s.

It is now feasible to take expectations to derive bias and variance of $\hat{\mu}_{\text{SNIS}}^N$. For the bias, we will use \hat{g}_2 , while for the variance we use \hat{g}_1 . To start with the bias, we look at the expected value

$$\begin{aligned}\mathbb{E}_q[\hat{g}_2(I, Z_p)] &= \underbrace{\frac{I}{Z_p}}_{\mu} + \frac{1}{Z_p} \cdot 0 - \frac{I}{Z_p^2} \cdot 0 - \frac{1}{Z_p^2} \text{Cov}_q[\hat{I}^N, \hat{Z}_p^N] + \frac{I}{Z_p^3} \mathbb{V}_q[\hat{Z}_p^N] \\ &= \mu + \frac{1}{Z_p^2} \left(\frac{I}{Z_p} \cdot \mathbb{V}_q[\hat{Z}_p^N] - \text{Cov}_q[\hat{I}^N, \hat{Z}_p^N] \right) \\ &= \mu + \frac{1}{Z_p^2} \left(\mu \cdot \mathbb{V}_q[\hat{Z}_p^N] - \varrho_{\hat{I}^N, \hat{Z}_p^N} \cdot \sqrt{\mathbb{V}_q[\hat{I}^N] \cdot \mathbb{V}_q[\hat{Z}_p^N]} \right)\end{aligned}\quad (90)$$

where $\varrho_{\hat{I}^N, \hat{Z}_p^N} \stackrel{\text{def}}{=} \text{Corr}_q[\hat{I}^N, \hat{Z}_p^N]$. Note that the expectations become variances here only because the estimators of I and Z_p are unbiased, and would be MSE's otherwise. A nice feature of this expression is that it has μ in front, so the bias of $\hat{\mu}_{\text{SNIS}}^N$ is immediately seen as the remaining part.

$$\text{Bias}_q^\delta[\hat{\mu}_{\text{SNIS}}^N] = \frac{1}{Z_p^2} \left(\mu \cdot \mathbb{V}_q[\hat{Z}_p^N] - \varrho_{\hat{I}^N, \hat{Z}_p^N} \cdot \sqrt{\mathbb{V}_q[\hat{I}^N] \cdot \mathbb{V}_q[\hat{Z}_p^N]} \right) \quad (91)$$

What the delta method does is approximating g first by a Taylor approximation, and then get the *exact* expectation of that approximation.

Variance We obtain the variance of $\hat{g}_1(I, Z_p)$, i.e.,

$$\mathbb{V}_q[\hat{\mu}_{\text{SNIS}}^N] \approx \mathbb{V}_q[\hat{g}_1(I, Z_p)] \quad (92)$$

$$= \mathbb{V}_q[\hat{g}_1(I, Z_p)] = \mathbb{V}_q \left[\frac{I}{Z_p} + \frac{1}{Z_p} (\hat{I}^N - I) - \frac{I}{Z_p^2} (\hat{Z}_p^N - Z_p) \right] \quad (93)$$

Since \hat{I}^N and \hat{Z}_p^N are correlated, $\frac{I}{Z_p}$ is a constant, using $\mathbb{V}[X - Y] = \mathbb{V}[X] + \mathbb{V}[Y] - 2 \text{Cov}[X, Y]$ we get

$$\mathbb{V}_q^\delta[\hat{\mu}_{\text{SNIS}}^N] = \frac{1}{Z_p^2} \mathbb{V}_q[\hat{I}^N] + \frac{I^2}{Z_p^4} \mathbb{V}_q[\hat{Z}_p^N] - \frac{2I}{Z_p^3} \cdot \text{Cov}_q[\hat{I}^N, \hat{Z}_p^N] \quad (94)$$

Therefore we can get the delta method MSE as the sum

$$\text{MSE}_\delta[\hat{\mu}_{\text{SNIS}}^N] = (\text{Bias}_q^\delta[\hat{\mu}_{\text{SNIS}}^N])^2 + \mathbb{V}_q^\delta[\hat{\mu}_{\text{SNIS}}^N] \quad (95)$$

$$= \frac{1}{Z_p^4} \left(\mu^2 \cdot \mathbb{V}_q[\hat{Z}_p^N] + \varrho_{\hat{I}^N, \hat{Z}_p^N}^2 \mathbb{V}_q[\hat{I}^N] \cdot \mathbb{V}_q[\hat{Z}_p^N] - 2\varrho_{\hat{I}^N, \hat{Z}_p^N} \mu \mathbb{V}_q[\hat{Z}_p^N] \sqrt{\mathbb{V}_q[\hat{I}^N] \cdot \mathbb{V}_q[\hat{Z}_p^N]} \right) \quad (96)$$

$$+ \frac{1}{Z_p^2} \mathbb{V}_q[\hat{I}^N] + \frac{I}{Z_p^4} \mathbb{V}_q[\hat{Z}_p^N] - \frac{2I}{Z_p^3} \cdot \varrho_{\hat{I}^N, \hat{Z}_p^N} \sqrt{\mathbb{V}_q[\hat{I}^N] \cdot \mathbb{V}_q[\hat{Z}_p^N]} \quad (97)$$

I Extension to bridge sampling-type estimators and connections

Here, we connect our method to bridge sampling, a methodology for estimation of the ratio of normalizing constants [57] (see an excellent explanation in [63, Chapter 10, Section 9]. Recall notation $\mu = \mathbb{E}_p[f(\mathbf{x})] = I/Z_p$. Bridge sampling is a ratio estimator that also involves sampling from two different distribution, although there is no joint in an extended space. In fact, bridge sampling requires us to sample exactly from $q_2^\star = p(\mathbf{x}) = \tilde{p}(\mathbf{x})/Z_p$ and $q_1^\star = f(\mathbf{x})p(\mathbf{x})/\mu$. Obtaining samples $\{\mathbf{x}^{(n_1)}\}_{n_1=1}^{N_1} \stackrel{\text{i.i.d.}}{\sim} q_1^\star(\mathbf{x})$ and $\{\mathbf{x}^{(n_2)}\}_{n_2=1}^{N_2} \stackrel{\text{i.i.d.}}{\sim} q_2^\star(\mathbf{x})$, the estimator is constructed as

$$\hat{\mu}_{\text{Simple-bridge}} \stackrel{\text{def}}{=} \frac{\frac{1}{N_2} \sum_{n_2=1}^{N_2} f(\mathbf{x}^{(n_2)})p(\mathbf{x}^{(n_2)})}{\frac{1}{N_1} \sum_{n_1=1}^{N_1} \tilde{p}(\mathbf{x}^{(n_1)})} = \frac{\frac{1}{N_2} \sum_{n_2=1}^{N_2} \tilde{q}_1^\star(\mathbf{x}^{(n_2)})}{\frac{1}{N_1} \sum_{n_1=1}^{N_1} \tilde{q}_2^\star(\mathbf{x}^{(n_1)})}, \quad (98)$$

where we used notation for the unnormalized targets $\tilde{q}_1^\star(\mathbf{x}) = f(\mathbf{x})p(\mathbf{x})$ and $\tilde{q}_2^\star(\mathbf{x}) = \tilde{p}(\mathbf{x})$. Notice the use of samples from q_2^\star in the numerator, while the samples from q_1^\star in the denominator, which may be at first counter-intuitive. Taking expectations of numerator and denominator of Eq. (98) gives

$$\frac{\frac{1}{N_2} \mathbb{E}_{q_2^\star} [\sum_{n_2=1}^{N_2} \tilde{q}_1^\star(\mathbf{x}^{(n_2)})]}{\frac{1}{N_1} \mathbb{E}_{q_1^\star} [\sum_{n_1=1}^{N_1} \tilde{q}_2^\star(\mathbf{x}^{(n_1)})]} = \frac{Z_p^{-1} \int \tilde{q}_1^\star(\mathbf{x}) \tilde{q}_2^\star(\mathbf{x}) d\mathbf{x}}{I^{-1} \int \tilde{q}_2^\star(\mathbf{x}) \tilde{q}_1^\star(\mathbf{x}) d\mathbf{x}} = \frac{Z_p^{-1} \cdot B}{I^{-1} \cdot B} = \frac{I}{Z_p} = \mu, \quad (99)$$

where the constant B cancels. Differently to $\hat{\mu}_{\text{GenSNIS}}^N$, this is a ratio of two unbiased estimators (almost, they are scaled by the constant B) of $1/I$ and $1/Z_p$, instead of I and Z_p . The asymptotic variance can also be obtained with the delta method as for $\hat{\mu}_{\text{GenSNIS}}^N$. Before we do so, in fact the more general version of the estimator is

$$\hat{\mu}_{\text{Bridge}} \stackrel{\text{def}}{=} \frac{\frac{1}{N_2} \sum_{n_2=1}^{N_2} \tilde{q}_1^\star(\mathbf{x}^{(n_2)}) \alpha(\mathbf{x}^{(n_2)})}{\frac{1}{N_1} \sum_{n_1=1}^{N_1} \tilde{q}_2^\star(\mathbf{x}^{(n_1)}) \alpha(\mathbf{x}^{(n_1)})} = \frac{\hat{\eta}_1}{\hat{\eta}_2} \quad (100)$$

for some function $\alpha(\mathbf{x})$ which can be chosen to minimize asymptotic variance.

Remark 5. Keys differences from $\hat{\mu}_{GenSNIS}^N$ are

- The need to sample exactly from q_1^\star and q_2^\star , which is a known disadvantage of bridge sampling that makes it difficult to apply to certain problems. Quoting [63, Chapter 10, Section 9]: “In the problems where normalization ratios are most needed, we may not be able to sample from either p_0 or p_1 ” (the equivalent of our q_1^\star and q_2^\star).¹⁶
- The MSE is lower bounded as in $\hat{\mu}_{SNIS}^N$, i.e., it is not theoretically possible to achieve zero MSE.

Therefore the two methodologies, $\hat{\mu}_{GenSNIS}^N$ and bridge sampling, can be seen as alternatives depending on the requirements of the estimation problem.

Nonetheless, the core underlying idea behind $\hat{\mu}_{GenSNIS}^N$ of improving variance by correlating the estimates can be applied to give a generalization of bridge sampling.

Delta method for bridge sampling. To do so, let us first derive the asymptotic variance of bridge sampling. We find the expectations of numerator and denominator as

$$\mathbb{E}_{q_2^\star}[\hat{\eta}_1] = B \cdot Z_p^{-1} \stackrel{\text{def}}{=} \eta_1, \quad \mathbb{E}_{q_1^\star}[\hat{\eta}_2] = B \cdot I^{-1} \stackrel{\text{def}}{=} \eta_2. \quad (101)$$

Let the ratio of the two estimators $g(\hat{\eta}_1, \hat{\eta}_2) = \hat{\eta}_1/\hat{\eta}_2$ be approximated by 1st-order Taylor approximation \hat{g}_1 and evaluated at the true quantities η_1, η_2 (keep in mind, $\mu = \eta_1/\eta_2$). The variance of the approximation is

$$\mathbb{V}_{\mathbb{Q}_{1:2}^\star}[\hat{g}_1(\eta_1, \eta_2)] = \mathbb{V}_{\mathbb{Q}_{1:2}^\star}\left[\frac{\eta_1}{\eta_2} + \frac{1}{\eta_2} \left(\hat{\eta}_1^N - \eta_1\right) - \frac{\eta_1}{\eta_2^2} (\hat{\eta}_2^N - \eta_2)\right], \quad (102)$$

and since this variance is the asymptotic variance, we have

$$\mathbb{V}_{\mathbb{Q}_{1:2}^\star}^{\infty}[\hat{\mu}_{Bridge}^N] = \frac{1}{\eta_2^2} \mathbb{V}_{q_2^\star}[\hat{\eta}_1] + \frac{\eta_1^2}{\eta_2^4} \mathbb{V}_{q_1^\star}[\hat{\eta}_2] - \frac{2\eta_1}{\eta_2^3} \cdot \text{Cov}_q[\hat{\eta}_1, \hat{\eta}_2], \quad (103)$$

¹⁶It is possible to use MCMC for generating such samples, but the analysis of the asymptotic variance and practical considerations complicate significantly.

(the coefficient of the third term can also be written as $-2\mu/\eta_2^2$) where

$$\mathbb{V}_{q_2^\star}[\hat{\eta}_1] = \frac{1}{N_2} \left(\mathbb{E}_{q_2^\star}[\tilde{q}_1^\star(\mathbf{x})^2 \alpha(\mathbf{x})^2] - \left(\int \tilde{q}_1^\star(\mathbf{x}) \tilde{q}_2^\star(\mathbf{x}) \alpha(\mathbf{x}) \right)^2 \right) \quad (104)$$

$$\mathbb{V}_{q_1^\star}[\hat{\eta}_2] = \frac{1}{N_1} \left(\mathbb{E}_{q_1^\star}[\tilde{q}_2^\star(\mathbf{x})^2 \alpha(\mathbf{x})^2] - \left(\int \tilde{q}_1^\star(\mathbf{x}) \tilde{q}_2^\star(\mathbf{x}) \alpha(\mathbf{x}) \right)^2 \right). \quad (105)$$

Generalized version of bridge sampling based on couplings. In bridge sampling, the covariance term of Eq. (103) is zero, as the two sets of samples are independent. A $\hat{\mu}_{\text{GenSNIS}}^N$ inspired version of bridge sampling would sample from a joint in extended space to obtain

$$\hat{\mu}_{\text{GenBridge}} \stackrel{\text{def}}{=} \frac{\frac{1}{N} \sum_{n=1}^N \tilde{q}_1^\star(\mathbf{x}_2^{(n)}) \alpha(\mathbf{x}_2^{(n)})}{\frac{1}{N} \sum_{n=1}^N \tilde{q}_2^\star(\mathbf{x}_1^{(n)}) \alpha(\mathbf{x}_1^{(n)})} = \frac{\hat{\eta}_1}{\hat{\eta}_2}, \quad [\mathbf{x}_1^{(n)}, \mathbf{x}_2^{(n)}] \stackrel{\text{i.i.d.}}{\sim} \mathbb{Q}_{1:2}^\star(d\mathbf{x}_1, d\mathbf{x}_2) \quad (106)$$

where $\mathbb{Q}_{1:2}^\star(d\mathbf{x}_1, d\mathbf{x}_2)$ is a joint with marginals $q_1^\star(\mathbf{x}_1)$ and $q_2^\star(\mathbf{x}_2)$. With this estimator, the covariance term in the asymptotic variance can be controlled as in $\hat{\mu}_{\text{GenSNIS}}^N$. In $\hat{\mu}_{\text{GenBridge}}$, we can control (i) the joint $\mathbb{Q}_{1:2}^\star(d\mathbf{x}_1, d\mathbf{x}_2)$ with fixed marginals and (ii) the function $\alpha(\mathbf{x})$ used in the estimator.

J Sample recycling

A straightforward way to use all the samples in the estimation is by having two UIS estimators in both numerator and denominator, as

$$\hat{\mu}_{\text{GenSNIS-Recycle}} = \frac{\sum_{n=1}^N \frac{f(\mathbf{x}_1^{(n)}) \cdot \tilde{p}(\mathbf{x}_1^{(n)})}{q_1(\mathbf{x}_1^{(n)})} + \sum_{n=1}^N \frac{f(\mathbf{x}_2^{(n)}) \cdot \tilde{p}(\mathbf{x}_2^{(n)})}{q_2(\mathbf{x}_2^{(n)})}}{\sum_{n=1}^N \frac{\tilde{p}(\mathbf{x}_2^{(n)})}{q_2(\mathbf{x}_2^{(n)})} + \sum_{n=1}^N \frac{\tilde{p}(\mathbf{x}_1^{(n)})}{q_1(\mathbf{x}_1^{(n)})}}, \quad [\mathbf{x}_1^{(n)}, \mathbf{x}_2^{(n)}]^\top \sim \mathbb{Q}_{1:2}(\mathbf{x}_1, \mathbf{x}_2). \quad (107)$$

In fact, all of the sample recycling variants proposed in TABI [69] can also be used straightforwardly for $\hat{\mu}_{\text{GenSNIS}}^N$. While such estimators have the advantage of using all the samples in both integrals, the variance analysis is more complicated, and sample sharing in numerator/denominator suffers from the same drawback of $\hat{\mu}_{\text{SNIS}}^N$, in the sense that the implicit CRN coupling implied by sample sharing is not guaranteed to reduce variance (and could increase it).