Optimality in Noisy Importance Sampling

Fernando Llorente*, Luca Martino⋆, Jesse Read†, David Delgado–Gómez*

* Universidad Carlos III de Madrid, Leganés, Spain.

⋆ Universidad Rey Juan Carlos, Fuenlabrada, Spain.

† École Polytechnique, Palaiseau, France.

Abstract

Many applications in signal processing and machine learning require the study of probability density functions (pdfs) that can only be accessed through noisy evaluations. In this work, we analyze the noisy importance sampling (IS), i.e., IS working with noisy evaluations of the target density. We present the general framework and derive optimal proposal densities for noisy IS estimators. The optimal proposals incorporate the information of the variance of the noisy realizations, proposing points in regions where the noise power is higher. We also compare the use of the optimal proposals with previous optimality approaches considered in a noisy IS framework.

Keywords: Bayesian Inference; Noisy Monte Carlo; Pseudo-marginal Metropolis-Hastings; Noisy IS.

1 Introduction

A wide range of modern applications, especially in Bayesian inference framework [1], require the study of probability density functions (pdfs) which can be evaluated stochastically, i.e., only noisy evaluations can be obtained [2, 3, 4, 5]. For instance, this is the case of the pseudo-marginal approaches and doubly intractable posteriors [6, 7], approximate Bayesian computation (ABC) and likelihood-free schemes [8, 9], where the target density cannot be computed in closed-form.

The noisy scenario also appears naturally when mini-batches of data are employed instead of considering the complete likelihood of huge amounts of data [10, 11]. More recently, the analysis of noisy functions of densities is required in reinforcement learning (RL), especially in direct policy search which is an important branch of RL, with applications in robotics [12, 13]. The topic of inference in noisy settings (or where a function is known with a certain degree of uncertainty) is also of interest in the inverse problem literature, such as in the calibration of expensive computer codes [14, 15]. This is also the case when the construction of an *emulator* is considered, as a surrogate model [4, 16, 17].

In this work, we study the importance sampling (IS) scheme under noisy evaluations of the target pdf. The noisy IS scenario has been already analyzed in the literature [2, 3, 18]. In the context of optimization, some theoretical results can be found [19]. In the sequential framework, IS schemes with random weights can be found and have been studied in different works [2, 20, 21, 22]. We provide the optimal proposal densities for different noisy scenarios, including also the case of integrals involving vector-valued functions. Moreover, we discuss the convergence and variance of the estimators in a general setting. We consider a different approach with respect to other studies in the literature [3, 23]. In those works, the authors analyzed the trade-off between decreasing the noise power (by increasing the number of auxiliary samples) and increasing the total number of samples in the IS estimators. Here, this information is encompassed within the optimal proposal density, which plays a similar role to an acquisition function in active learning [17, 16]. This is information is relevant, especially if the noisy evaluations are also costly to obtain.

2 Background

2.1 Bayesian inference

In many applications, we aim at inferring a variable of interest given a set of observations or measurements. Let us denote the variable of interest by $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^{d_x}$, and let $\mathbf{y} \in \mathbb{R}^{d_y}$ be the observed data. The posterior pdf is then

$$\bar{p}(\mathbf{x}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\mathbf{x})g(\mathbf{x})}{Z(\mathbf{y})}, \tag{1}$$

where $\ell(\mathbf{y}|\mathbf{x})$ is the likelihood function, $g(\mathbf{x})$ is the prior pdf, and $Z(\mathbf{y})$ is the model evidence (a.k.a. marginal likelihood) which is a useful quantity in model selection problems [24]. For simplicity, in the following, we skip the dependence on $\mathbf{y}$ in $\bar{p}(\mathbf{x}) = \bar{p}(\mathbf{x}|\mathbf{y})$ and $Z = Z(\mathbf{y})$. Generally, Z is unknown, so we are able to evaluate the unnormalized target function, i.e., the numerator on the right hand side of Eq. (1),

$$p(\mathbf{x}) = \ell(\mathbf{y}|\mathbf{x})g(\mathbf{x}). \tag{2}$$

The analytical study of the posterior density $\bar{p}(\mathbf{x}) \propto p(\mathbf{x})$ is unfeasible, so that numerical approximations are required [25, 1].

2.2 Noisy framework

Generally, we desire to approximate the unnormalized density $p(\mathbf{x})$, $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$, and the corresponding normalizing constant Z, using Monte Carlo methods. The unnormalized density $p(\mathbf{x})$ can represent a posterior density in a Bayesian inference problem, as described above. We assume that, for any $\mathbf{x}$, we cannot evaluate $p(\mathbf{x})$ exactly, but we only have access to a related noisy realization. Moreover, in many applications, obtaining such a noisy realization may be expensive. Hence, analyzing in which $\mathbf{x}$ we require a noisy realization of $p(\mathbf{x})$ is an important problem, which is related to the concept of *optimality* that we consider below.

In the following, we introduce a concise mathematical formalization of the noisy scenario. This simple framework contains real application scenarios, such as latent variable models [3] (see example 4 in Sect. 4.1), likelihood-free inference setting [8], doubly intractable posteriors [7], mini batch-based inference [10]. More specifically, we assume to have access to a noisy realization related to $p(\mathbf{x})$, i.e.,

$$\widetilde{m}(\mathbf{x}) = H(p(\mathbf{x}), \epsilon), \tag{3}$$

where H is a non-linear transformation involving $p(\mathbf{x})$ and ϵ, that is some noise perturbation. Thus, *for a fixed value* $\mathbf{x}$, $\widetilde{m}(\mathbf{x})$ is a random variable with

$$m(\mathbf{x}) = \mathbb{E}[\widetilde{m}(\mathbf{x})], \quad s(\mathbf{x})^2 = \mathrm{Var}[\widetilde{m}(\mathbf{x})], \tag{4}$$

for some *mean function*, $m(\mathbf{x})$, and *variance function*, $s(\mathbf{x})^2$. The assumption that $\widetilde{m}(\mathbf{x})$ must be strictly positive is important in practice [2, 23].

Noise power. In some applications, it is also possible to control the noise power $s(\mathbf{x})^2$, for instance by adding/removing data to the mini-batches (e.g., in the context of Big Data) [10], increasing the

number of auxiliary samples in latent variables models [6], or interacting with an environment over longer/shorter periods of time (e.g., in reinforcement learning) [12].

Unbiased scenario and related cases. The scenario where $m(\mathbf{x}) = p(\mathbf{x})$ appears naturally in some applications (such as in the estimation of latent variable and stochastic volatility models in statistics [3, 26]; or in the context of optimal filtering of partially observed stochastic processes [18]), or it is often assumed as a pre-established condition by the authors [3, 2]. In some other scenarios, the noisy realizations are known to be unbiased estimates of some transformation of $p(\mathbf{x})$, e.g., of $\log p(\mathbf{x})$ [27, 28]. This situation can be encompassed by the following special case. If we consider an additive perturbation,

$$\widetilde{m}(\mathbf{x}) = G\left(p(\mathbf{x})\right) + \epsilon, \qquad \text{with } \mathbb{E}[\epsilon] = 0, \tag{5}$$

we have $m(\mathbf{x}) = \mathbb{E}[\widetilde{m}(\mathbf{x})] = G\left(p(\mathbf{x})\right)$, where $G(\cdot) : \mathbb{R} \to \mathbb{R}$. If G is known and invertible, we have $p(\mathbf{x}) = G^{-1}\left(m(\mathbf{x})\right)$.

Generally, we can state that $m(\mathbf{x})$ always contains statistical information related to $p(\mathbf{x})$. The subsequent use of $m(\mathbf{x})$ depends on the specific application. Thus, we study the mean function $m(\mathbf{x})$. Hence, our goal is to approximate efficiently integrals involving $m(\mathbf{x})$, i.e.,

$$\mathbf{I} = \frac{1}{\bar{Z}} \int_{\mathcal{X}} \mathbf{f}(\mathbf{x}) m(\mathbf{x}) d\mathbf{x}, \quad \bar{Z} = \int_{\mathcal{X}} m(\mathbf{x}) d\mathbf{x}, \tag{6}$$

where $\mathbf{f}(\mathbf{x}) : \mathcal{X} \to \mathbb{R}^{d_f}$ and $\mathbf{I} = [I_1, \ldots, I_{d_f}]^\top \in \mathbb{R}^{d_f}$ denotes the vector of integrals of interest. Note that, in the unbiased case $m(\mathbf{x}) = p(\mathbf{x})$, we have $\bar{Z} = Z$. An integral involving $m(\mathbf{x})$ can be approximated employing a cloud of random samples using the noisy realizations $\widetilde{m}(\mathbf{x})$ via Monte Carlo methods.

3 Noisy Importance Sampling

In a non-noisy IS scheme, a set of samples is drawn from a proposal density $q(\mathbf{x})$. Then each sample is weighted according to the ratio $\frac{p(\mathbf{x})}{q(\mathbf{x})}$. A noisy version of importance sampling can be obtained when we substitute the evaluations of $p(\mathbf{x})$ with noisy realizations of $\widetilde{m}(\mathbf{x})$. See Table 1 and note that the importance weights w_n in Eq. (9) are computed using the noisy realizations. Below, we show that

$$\widehat{Z} = \frac{1}{N} \sum_{n=1}^{N} w_n, \tag{7}$$

is an unbiased estimator of $\bar{Z}$, and

$$\widehat{\mathbf{I}}_{\mathrm{std}} = \frac{1}{N\bar{Z}} \sum_{n=1}^{N} w_n \mathbf{f}(\mathbf{x}_n), \qquad \widehat{\mathbf{I}}_{\mathrm{self}} = \frac{1}{N\widehat{Z}} \sum_{n=1}^{N} w_n \mathbf{f}(\mathbf{x}_n), \tag{8}$$

are consistent estimators of $\mathbf{I}$. The estimator $\widehat{\mathbf{I}}_{\mathrm{std}}$ requires the knowledge of $\bar{Z}$, that is not needed in the so-called self-normalized estimator, $\widehat{\mathbf{I}}_{\mathrm{self}}$.

Table 1: Noisy importance sampling algorithm

1. **Inputs:** Proposal distribution $q(\mathbf{x})$.

2. For $n = 1, \ldots, N$:

   (a) Sample $\mathbf{x}_n \sim q(\mathbf{x})$ and obtain one realization $\widetilde{m}(\mathbf{x}_n)$.

   (b) Compute

$$w_n = \frac{\widetilde{m}(\mathbf{x}_n)}{q(\mathbf{x}_n)} \tag{9}$$

4 **Outputs:** the weighted samples $\{\mathbf{x}_n, w_n\}_{n=1}^N$.

Theorem. *The estimators above constructed from the output of noisy IS converge to expectations under $m(\mathbf{x})$. More specifically, we have $\widehat{Z}$ and $\widehat{\mathbf{I}}_{std}$ are unbiased estimators of $\bar{Z}$ and $\mathbf{I}$ respectively, and $\widehat{\mathbf{I}}_{self}$ is a consistent estimator of $\mathbf{I}$. Moreover, these estimators have higher variance than their non-noisy counterparts.*

Proof. Here, we provide a simple proof of convergence by applying iterated conditional expectations. Equivalently, the correctness of the approach can be proved by using an extended space view (see, e.g., [18, 3]).

Let $\mathbf{x}_{1:N} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ denote the N samples from q. By the law of total expectation, we have that $\mathbb{E}[\widehat{Z}] = \mathbb{E}\left[\mathbb{E}[\widehat{Z}|\mathbf{x}_{1:N}]\right]$. In the inner expectation, we use the fact the w_i's are i.i.d., hence

$$\mathbb{E}[\widehat{Z}|\mathbf{x}_{1:N}] = \frac{1}{N}\sum_{i=1}^N \mathbb{E}[w_i|\mathbf{x}_i] = \frac{1}{N}\sum_{i=1}^N \frac{1}{q(\mathbf{x}_i)}\mathbb{E}[\widetilde{m}(\mathbf{x}_i)|\mathbf{x}_i] = \frac{1}{N}\sum_{i=1}^N \frac{m(\mathbf{x}_i)}{q(\mathbf{x}_i)} = \widetilde{Z},$$

where $\widetilde{Z}$ is the non-noisy IS estimator of $\bar{Z}$, which is also unbiased, i.e.,

$$\mathbb{E}[\widehat{Z}] = \mathbb{E}\left[\mathbb{E}[\widehat{Z}|\mathbf{x}_{1:N}]\right] = \mathbb{E}[\widetilde{Z}] = \bar{Z}.$$

Therefore, $\widehat{Z}$ is an unbiased estimator of $\bar{Z} = \int_{\mathcal{X}} m(\mathbf{x})d\mathbf{x}$, i.e., $\mathbb{E}[\widehat{Z}] = \bar{Z}$. Moreover, we show below that $\mathrm{Var}[\widehat{Z}]$ decreases to zero as $N \to \infty$. Hence, $\widehat{Z}$ is a consistent estimator of $\bar{Z}$. Now, with the same arguments, we can prove that the estimator $\widehat{\mathbf{E}} = \frac{1}{N}\sum_{i=1}^N \frac{\widetilde{m}(\mathbf{x})\mathbf{f}(\mathbf{x})}{q(\mathbf{x})}$ is also unbiased and converges to $\mathbf{E} = \int_{\mathcal{X}} \mathbf{f}(\mathbf{x})m(\mathbf{x})d\mathbf{x}$. Thus, both the estimator $\widehat{\mathbf{I}}_{std}$, and the ratio

$$\widehat{\mathbf{I}} = \frac{1}{\widehat{Z}}\widehat{\mathbf{E}} = \frac{1}{\sum_{j=1}^N w_j}\sum_{i=1}^N w_i\mathbf{f}(\mathbf{x}_i),$$

which is the noisy self-normalized IS estimator $\widehat{\mathbf{I}}_{self}$ in Eq. (8), are consistent estimators of

$$\mathbf{I} = \frac{\int_{\mathcal{X}} \mathbf{f}(\mathbf{x})m(\mathbf{x})d\mathbf{x}}{\int_{\mathcal{X}} m(\mathbf{x})d\mathbf{x}} = \frac{1}{\bar{Z}}\int_{\mathcal{X}} \mathbf{f}(\mathbf{x})m(\mathbf{x})d\mathbf{x},$$

given in Eq. (6).

$\square$

Variance of $\widehat{Z}$. By the law of total variance, we have that

$$\text{Var}[\widehat{Z}] = \mathbb{E}\left[\text{Var}[\widehat{Z}|\mathbf{x}_{1:N}]\right] + \text{Var}\left[\mathbb{E}[\widehat{Z}|\mathbf{x}_{1:N}]\right].$$

In a non-noisy scenario, i.e., in a non-noisy IS setting, the first term is null. Using the fact that $\widehat{Z}$ is unbiased, we have that the second term is

$$\text{Var}\left[\mathbb{E}[\widehat{Z}|\mathbf{x}_{1:N}]\right] = \text{Var}[\widetilde{Z}] = \mathcal{O}\left(1/N\right).$$

Regarding the first term, we have

$$\text{Var}[\widehat{Z}|\mathbf{x}_{1:N}] = \frac{1}{N^2}\sum_{i=1}^{N}\text{Var}[w_i|\mathbf{x}_i] = \frac{1}{N^2}\sum_{i=1}^{N}\frac{1}{q(\mathbf{x}_i)^2}\text{Var}[\widetilde{m}(\mathbf{x}_i)|\mathbf{x}_i] = \frac{1}{N^2}\sum_{i=1}^{N}\frac{s(\mathbf{x}_i)^2}{q(\mathbf{x}_i)^2}.$$

Assuming that $\frac{s(\mathbf{x})^2}{q(\mathbf{x})^2} < \infty$ for all $\mathbf{x}$, we have that

$$\mathbb{E}\left[\text{Var}[\widehat{Z}|\mathbf{x}_{1:N}]\right] = \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\frac{s(\mathbf{x}_i)^2}{q(\mathbf{x}_i)^2}\right] = \frac{1}{N}\mathbb{E}\left[\frac{s(\mathbf{x})^2}{q(\mathbf{x})^2}\right], \quad \text{where} \quad \mathbf{x} \sim q(\mathbf{x}).$$

Hence, we finally have that

$$\text{Var}[\widehat{Z}] = \frac{1}{N}\mathbb{E}\left[\frac{s(\mathbf{x})^2}{q(\mathbf{x})^2}\right] + \text{Var}[\widetilde{Z}] \geq \text{Var}[\widetilde{Z}]. \tag{10}$$

Therefore, $\widehat{Z}$ has a greater variance than $\widetilde{Z}$, but the same convergence speed, i.e., its variance decreases at $\frac{1}{N}$ rate. Proving that $\widehat{\mathbf{E}}$ has greater variance than its non-noisy version is straightforward.

4 Optimal Proposal Density in Noisy IS

In this section, we derive the optimal proposals for the noisy IS estimators $\widehat{Z}$, $\widehat{\mathbf{I}}_{\text{std}}$ and $\widehat{\mathbf{I}}_{\text{self}}$.

4.1 Optimal proposal for $\widehat{Z}$

We can rewrite the variance of $\widehat{Z}$ in Eq. (10) as

$$\text{Var}[\widehat{Z}] = \frac{1}{N}\mathbb{E}\left[\frac{m(\mathbf{x})^2 + s(\mathbf{x})^2}{q(\mathbf{x})^2}\right] - \frac{1}{N}\bar{Z}^2.$$

By Jensen's inequality, the first term is bounded below by

$$\mathbb{E}\left[\frac{m(\mathbf{x})^2 + s(\mathbf{x})^2}{q(\mathbf{x})^2}\right] \geq \left(\mathbb{E}\left[\frac{\sqrt{m(\mathbf{x})^2 + s(\mathbf{x})^2}}{q(\mathbf{x})}\right]\right)^2.$$

The minimum variance $V_{min} = \min_q Var[\widehat{Z}]$ is thus attained at

$$q_{opt}(\mathbf{x}) \propto \sqrt{m(\mathbf{x})^2 + s(\mathbf{x})^2}, \tag{11}$$

Note that, for finite N, V_{min} is always greater than 0, specifically,

$$V_{min} = \frac{1}{N}\left[\int_{\mathcal{X}} \sqrt{m(\mathbf{x})^2 + s(\mathbf{x})^2}d\mathbf{x}\right]^2 - \frac{1}{N}\bar{Z}^2. \tag{12}$$

Hence, differently from the non-noisy setting, in noisy IS the optimal proposal does not provide an estimator with null variance. If $s(\mathbf{x}) = 0$ for all $\mathbf{x}$, then we come back to the non-noisy scenario and $V_{min} = \frac{1}{N}\left[\int_{\mathcal{X}} m(\mathbf{x})d\mathbf{x}\right]^2 - \frac{1}{N}\bar{Z}^2 = 0$. Note that the variance of using $q(\mathbf{x}) = \frac{1}{\bar{Z}}m(\mathbf{x})$ is

$$V_{sub\text{-}opt} = \frac{\bar{Z}}{N}\int_{\mathcal{X}} \frac{m(\mathbf{x})^2 + s(\mathbf{x})^2}{m(\mathbf{x})}d\mathbf{x} - \frac{1}{N}\bar{Z}^2 = \frac{\bar{Z}}{N}\int_{\mathcal{X}} \frac{s(\mathbf{x})^2}{m(\mathbf{x})}d\mathbf{x}. \tag{13}$$

In the following, we show several examples of noise models and their corresponding optimal proposal densities.

Example 1. Let us consider a Bernoulli-type noise where $\widetilde{m}(\mathbf{x}) = p_{max}\epsilon$, where $\epsilon \sim \text{Bernoulli}\left(\frac{p(\mathbf{x})}{p_{max}}\right)$, and $p_{max} = \max p(\mathbf{x})$. Then, we have

$$m(\mathbf{x}) = p(\mathbf{x}), \qquad s(\mathbf{x})^2 = p(\mathbf{x})[p_{max} - p(\mathbf{x})].$$

Replacing in Eq. (11), the optimal proposal density in this case is

$$q_{opt}(\mathbf{x}) \propto p(\mathbf{x})\sqrt{1 + [p_{max} - p(\mathbf{x})]^2}. \tag{14}$$

Example 2. Let us consider $\widetilde{m}(\mathbf{x}) = |p(\mathbf{x}) + \epsilon|$, with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. In this scenario, the random variable $\widetilde{m}(\mathbf{x})$ corresponds to a folded Gaussian random variable. We have

$$m(\mathbf{x}) = \sigma\sqrt{\frac{2}{\pi}}\exp\left(-p^2(\mathbf{x})/2\sigma^2\right) + p(\mathbf{x})[1 - 2\boldsymbol{\Phi}(-p(\mathbf{x})/\sigma)],$$
$$s(\mathbf{x})^2 = p(\mathbf{x})^2 + \sigma^2 - m(\mathbf{x})^2,$$

where $\boldsymbol{\Phi}(\mathbf{x})$ is the cumulative function of the standard Gaussian distribution. Then,

$$q_{opt}(\mathbf{x}) \propto \sqrt{p(\mathbf{x})^2 + \sigma^2}. \tag{15}$$

Example 3. Let us consider a multiplicative noise $\widetilde{m}(\mathbf{x}) = e^\epsilon p(\mathbf{x})$ with $\mathbb{E}[\epsilon] = 0$, hence

$$m(\mathbf{x}) = p(\mathbf{x})\mathbb{E}[e^\epsilon] \propto p(\mathbf{x}), \qquad s(\mathbf{x})^2 = p(\mathbf{x})^2 Var[e^\epsilon].$$

If we denote $A = \mathbb{E}[e^\epsilon]$ and $\sigma^2 = Var[e^\epsilon]$, then $m(\mathbf{x}) = Ap(\mathbf{x})$ and $s(\mathbf{x})^2 = \sigma^2 p^2(\mathbf{x})$. In this case, the optimal proposal coincides with the optimal one in the non-noisy setting, since

$$q_{opt}(\mathbf{x}) \propto \sqrt{A^2 p^2(\mathbf{x}) + \sigma^2 p^2(\mathbf{x})} = p(\mathbf{x})\sqrt{A^2 + \sigma^2} \propto p(\mathbf{x}). \tag{16}$$

Example 4. In latent variable models, the noisy realization corresponds to the product of d_y independent IS estimators, each built from R auxiliary samples. With d_y large enough, the distribution of this realization is approximately lognormal, i.e.,

$$\widetilde{m}(\mathbf{x}) \sim \log \mathcal{N}(\mu(\mathbf{x}), \sigma^2(\mathbf{x})),$$

where $\mu(\mathbf{x}) = \log p(\mathbf{x}) - \frac{\gamma^2(\mathbf{x})}{2R}$ and $\sigma^2(\mathbf{x}) = \frac{\gamma^2(\mathbf{x})}{R}$, for some function $\gamma^2(\mathbf{x})$ [3, 23]. Equivalently, they write $\widetilde{m}(\mathbf{x}) = p(\mathbf{x})e^\epsilon$, where $\epsilon \sim \mathcal{N}(\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$. Hence,

$$m(\mathbf{x}) = p(\mathbf{x}), \qquad s(\mathbf{x})^2 = (e^{\gamma^2(\mathbf{x})/R} - 1)p(\mathbf{x})^2,$$

and the optimal proposal is

$$q_{\mathrm{opt}}(\mathbf{x}) \propto p(\mathbf{x})e^{\frac{\gamma^2(\mathbf{x})}{2R}}. \tag{17}$$

This example is related with the cases studied in [3, 23].

4.2 Optimal proposal for $\widehat{\mathbf{I}}_{std}$

We have already seen that the optimal proposal that minimizes the variance of $\widehat{Z}$ is $q_{\mathrm{opt}}(\mathbf{x}) \propto \sqrt{m(\mathbf{x})^2 + s(\mathbf{x})^2}$. Let us consider now the estimator $\widehat{\mathbf{I}}_{std}$. Note that this estimator assumes we can evaluate $\bar{Z} = \int_{\mathcal{X}} m(\mathbf{x})d\mathbf{x}$. Since we are considering a vector-valued function, the estimator has d_f components $\widehat{\mathbf{I}}_{std} = [\widehat{I}_{std,1} \ldots \widehat{I}_{std,d_f}]^\top$, and $\mathrm{Var}[\widehat{\mathbf{I}}_{std}]$ corresponds to a $d_f \times d_f$ covariance matrix. We aim to find the proposal that minimizes the sum of diagonal variances. From the results of the previous section, it is straightforward to show that the variance of the p-th component is

$$\mathrm{Var}[\widehat{I}_{std,p}] = \mathrm{Var}[\widetilde{I}_{std,p}] + \frac{1}{N\bar{Z}^2}\mathbb{E}\left[\frac{f_p(\mathbf{x})^2 s(\mathbf{x})^2}{q(\mathbf{x})^2}\right]$$

$$= \frac{1}{N\bar{Z}^2}\mathbb{E}\left[\frac{f_p(\mathbf{x})^2(m(\mathbf{x})^2 + s(\mathbf{x})^2)}{q(\mathbf{x})^2}\right] - \frac{1}{N\bar{Z}^2}I_p^2,$$

where $f_p(\mathbf{x})$ and I_p are respectively the p-th components of $\mathbf{f}(\mathbf{x})$ and $\mathbf{I}$, and $\widetilde{I}_{std,p}$ denotes the non-noisy estimator (i.e. using $m(\mathbf{x})$ instead of $\widetilde{m}(\mathbf{x})$). Thus,

$$\sum_{p=1}^{d_f}\mathrm{Var}[\widehat{I}_{std,p}] = \frac{1}{N\bar{Z}^2}\mathbb{E}\left[\frac{\sum_{p=1}^{d_f} f_p(\mathbf{x})^2(m(\mathbf{x})^2 + s(\mathbf{x})^2)}{q(\mathbf{x})^2}\right] - \frac{1}{N\bar{Z}^2}\sum_{p=1}^{d_f}I_p^2.$$

By Jensen's inequality, we have

$$\mathbb{E}\left[\frac{\sum_{p=1}^{d_f} f_p(\mathbf{x})^2(m(\mathbf{x})^2 + s(\mathbf{x})^2)}{q(\mathbf{x})^2}\right] \geq \left(\mathbb{E}\left[\frac{\sqrt{m(\mathbf{x})^2 + s(\mathbf{x})^2}\,\|\mathbf{f}(\mathbf{x})\|_2}{q(\mathbf{x})}\right]\right)^2,$$

where $\|\mathbf{f}(\mathbf{x})\|_2$ denotes the euclidean norm. The equality holds if and only if $\frac{\sqrt{m(\mathbf{x})^2 + s(\mathbf{x})^2}\|\mathbf{f}(\mathbf{x})\|_2}{q(\mathbf{x})}$ is constant. Hence, the optimal proposal is

$$q_{\mathrm{opt}}(\mathbf{x}) \propto \|\mathbf{f}(\mathbf{x})\|_2 \sqrt{m(\mathbf{x})^2 + s(\mathbf{x})^2}. \tag{18}$$

4.3 Optimal proposal for $\widehat{\mathbf{I}}_{self}$

Let us consider the case of the self-normalized estimator $\widehat{\mathbf{I}}_{\text{self}}$. Recall that $\widehat{\mathbf{I}}_{\text{self}} = \frac{\widehat{\mathbf{E}}}{\widehat{Z}}$, where $\widehat{\mathbf{E}}$ denotes the noisy estimator of $\mathbf{E} = \int_{\mathcal{X}} \mathbf{f}(\mathbf{x})m(\mathbf{x})d\mathbf{x}$, so that we are considering ratios of estimators. Again, we aim to find the proposal that minimizes the variance of the vector-valued estimator $\widehat{\mathbf{I}}_{\text{self}}$. When N is large enough, the variance of p-th ratio is approximated as [29],

$$\text{Var}[\widehat{I}_{\text{self},p}] = \text{Var}\left[\frac{\widehat{E}_p}{\widehat{Z}}\right] \approx \frac{1}{\bar{Z}^2}\text{Var}[\widehat{E}_p] - 2\frac{E_p}{Z}\text{Cov}[\widehat{E}_p, \widehat{Z}] + \frac{E_p^2}{\bar{Z}^4}\text{Var}[\widehat{Z}],$$

where E_p is the p-th component of $\mathbf{E}$, and

$$\text{Var}[\widehat{E}_p] = \frac{1}{N}\mathbb{E}\left[\frac{f_p(\mathbf{x})^2(m(\mathbf{x})^2 + s(\mathbf{x})^2)}{q(\mathbf{x})^2}\right] - \frac{1}{N}E_p^2,$$

$$\text{Var}[\widehat{Z}] = \frac{1}{N}\mathbb{E}\left[\frac{m(\mathbf{x})^2 + s(\mathbf{x})^2}{q(\mathbf{x})^2}\right] - \frac{1}{N}\bar{Z}^2,$$

$$\text{Cov}[\widehat{E}_p, \widehat{Z}] = \frac{1}{N}\mathbb{E}\left[\frac{f_p(\mathbf{x})(m(\mathbf{x})^2 + s(\mathbf{x})^2)}{q(\mathbf{x})^2}\right] - \frac{1}{N}E_p\bar{Z}.$$

The first two results have been already obtained in the previous sections. The third result is given in Appendix A. The sum of the variances is thus

$$\sum_{p=1}^{d_f} \text{Var}[\widehat{I}_{\text{self},p}] \approx \frac{1}{N\bar{Z}^2}\mathbb{E}\left[\frac{(m(\mathbf{x})^2 + s(\mathbf{x})^2)\sum_{p=1}^{d_f}(f_p(\mathbf{x}) - I_p)^2}{q(\mathbf{x})^2}\right].$$

By Jensen's inequality, we can derive that the optimal proposal is

$$q_{\text{opt}}(\mathbf{x}) \propto \|\mathbf{f}(\mathbf{x}) - \mathbf{I}\|_2 \sqrt{m(\mathbf{x})^2 + s(\mathbf{x})^2}. \tag{19}$$

Relationship with active learning. The optimal density $q_{opt}(\mathbf{x})$ can be interpreted as an *acquisition density*, suggesting the regions of the space which require more number of acquisitions of the realizations $\widetilde{m}(\mathbf{x})$. Namely, $q_{opt}(\mathbf{x})$ plays a role similar to an acquisition function in active learning. This is information is relevant, especially if the noisy evaluations are also costly to obtain.

4.4 Connection with other types of optimality

Here, we discuss another approach for optimality in noisy IS and connect it with our work. Other related works, in Monte Carlo and noisy optimization literature, focus on the trade-off between accuracy/noisiness and computational cost [3, 23, 30]. In those settings, it is assumed that one can control the variance $s(\mathbf{x})^2$ of the noisy realizations $\widetilde{m}(\mathbf{x})$. Clearly, taking samples with higher accuracy, i.e. small variance $s(\mathbf{x})^2$, is beneficial since it decrease the magnitude of the terms $\mathbb{E}\left[\frac{s(\mathbf{x})^2}{m(\mathbf{x})^2}\right]$ and $\mathbb{E}\left[\frac{f_p(\mathbf{x})^2 s(\mathbf{x})^2}{m(\mathbf{x})^2}\right]$, which are responsible for the efficiency loss in the estimators, due to the presence of noise. However, taking accurate estimates implies increased computational cost, hence one must reduce the number of samples N, which affect the overall Monte Carlo variance. This trade-off have been investigated in both MCMC and IS frameworks [3, 23].

Let R denote the number of auxiliary samples employed to reduce the variance of the noisy realizations. Namely, greater R implies greater accuracy but also greater cost. Moreover, this number could depend on $\mathbf{x}$, i.e., $R(\mathbf{x}) : \mathcal{X} \to \mathbb{N}^+\backslash\{0\}$. Then, the goal is to obtain the optimal function $R(\mathbf{x})$ by balancing the decrease in variance with the extra computational cost (see, e.g., Sections 3.3, 3.4 and 5 of [3]). Namely, in this different approach, they try to reduce $s(\mathbf{x})^2$ at certain $\mathbf{x}$ increasing the value of $R(\mathbf{x})$, instead of using an optimal proposal pdf for the noisy scenario. On the contrary, in this work we have considered the use of optimal proposal pdfs and that $s(\mathbf{x})^2$ is not tuned by the user, which means that R is arbitrary and set constant for all $\mathbf{x}$.

5 Numerical experiments

In this section, we consider two illustrative numerical example where we clearly show the performance of the optimal proposal pdf in the noisy IS setting (showing the variance gains in estimation, with respect to the use the optimal proposal density from the non-noisy setting). For simplicity, we consider one-dimensional scenarios, and test the optimal proposal pdf with different densities $p(x)$ (uniform and Gaussian), and different types of variance behavior, $\sigma(x)$.

First experiment. Let $p(x) = \frac{1}{b-a}$ for $x \in [a, b]$, i.e., a uniform density in $[a, b]$ with $a = 0.1$ and $b = 10$. We set $\widetilde{m}(\mathbf{x}) = p(\mathbf{x})e^\epsilon$ with $\epsilon \sim \mathcal{N}(-\sigma^2/2, \sigma^2)$ so that $\mathbb{E}[e^\epsilon] = 1$, and we have $m(x) = \mathbb{E}[\widetilde{m}(\mathbf{x})] = p(x)$.
We consider the estimation of $\bar{Z} = 1$ using the optimal proposal pdf $q_{\text{opt}}(x)$ in Eq. (11), and the optimal proposal pdf in the non-noisy setting, i.e., $q_{\text{sub-opt}}(x) = p(x)$. More specifically, we consider

$$\sigma(x) = A|\log(x)|, \quad A > 0.$$

Hence,

$$s(x)^2 = \frac{e^{\sigma(x)^2} - 1}{(b-a)^2}, \quad \text{and} \quad q_{\text{opt}}(x) \propto \frac{1}{b-a}e^{\sigma(x)^2}.$$

Clearly, by changing A, we change the form of both $s(x)^2$ and $q_{\text{opt}}(x)$. For instance, increasing A also increases the magnitude of $s(x)^2$ and hence the mismatch between $q_{\text{sub-opt}}(x) = p(x)$ and $q_{\text{opt}}(x)$, as depicted in Figure 1. Indeed, for $A = 0.2$, $q_{\text{opt}}(x)$ is almost identical to $p(x)$ since the magnitude of $s(x)$ is small w.r.t. the values of $p(x)$. As A increases, $q_{\text{opt}}(x)$ deviates from $p(x)$, being in the middle between $p(x)$ and $s(x)$, and eventually would converge to $s(x)$ for $A \gg 1$. It is also interesting to note that $q_{\text{opt}}(x)$ with $A = 1.2$ has very little probability mass around $x = 1$, where the noise is zero, since it needs to concentrate probability mass in the extremes of the interval, where the noise power is huge.
Let also denote as $V_{\text{sub-opt}}$ the variance obtained using $q_{\text{sub-opt}}(x) = p(x)$ given in Eq. (12), and $V_{\text{opt}} = V_{\text{min}}$ the variance obtained using $q_{\text{opt}}(x)$ given in Eq. (13). In Figure 3(a), we show the ratio of variances $\frac{V_{\text{sub-opt}}}{V_{\text{opt}}}$ both theoretically and empirically, as a function of A, where $V_{\text{sub-opt}}$ and V_{opt} are the variances of $\widehat{Z}$ when using $p(x)$ and $q_{\text{opt}}(x)$ as proposals, respectively. We can observe the clear advantage of using the optimal proposal density $q_{\text{opt}}(x)$ in Eq. (11).

Second experiment. Let us consider now a Gaussian pdf, $p(x) = \mathcal{N}(x|0, 1)$, and the same

error model as in the previous example but considering

$$\sigma(x) = A|x|^{\frac{1}{2}}, \quad A > 0.$$

Figure 2 depicts the $q_{\text{opt}}(x)$ and $s(x)$, as a function of x, for several values of A. Note that, in this example, increasing A makes $q_{\text{opt}}(x)$ become bimodal. As in the previous example, as A increases, the optimal proposal $q_{\text{opt}}(x)$ will converge to $s(x)$. The theoretical and empirical curves of the ratio of variances, $\frac{V_{\text{sub-opt}}}{V_{\text{opt}}}$, in estimating Z are shown in Figure 3(b).
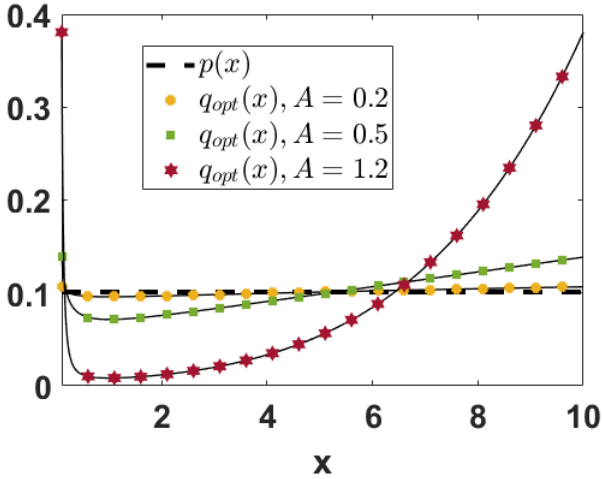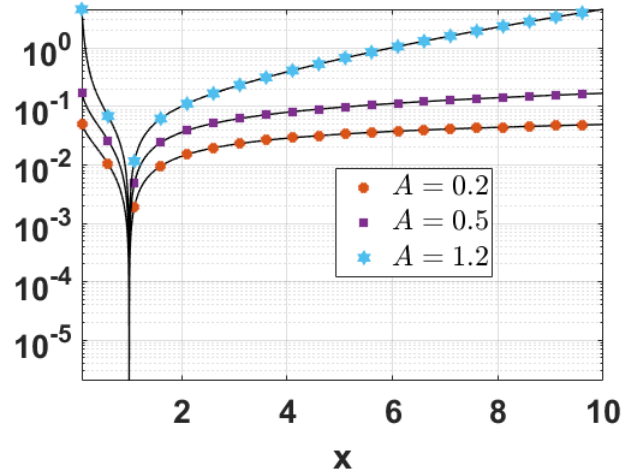
(a) $q_{\text{opt}}(x)$ for different values of A.   (b) $s(x)$ for different values of A in log-scale.
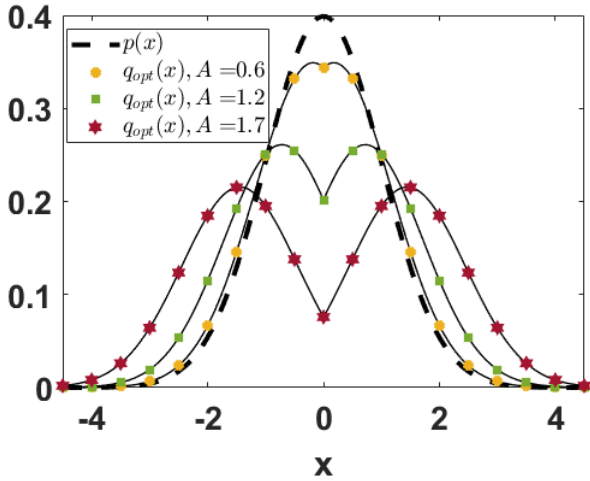
Figure 1: Uniform example. **(a)** Optimal proposals $q_{\text{opt}}(x)$ for different values of A, and the $q_{\text{sub-opt}}(x) = p(x)$ in dashed line; **(b)** The standard deviation $s(x)$ for different values of A.

6   Conclusions

Working with noisy evaluations of the target density is usual in Monte Carlo, especially in the last years. In this work, we have analyzed the use of optimal proposal densities in a noisy IS framework. Previous works have focused on the trade-off between accuracy in the evaluation and computational cost in order to form optimal estimators. In this work, we have considered a general setting and derived the optimal proposals for the noisy IS estimators. These optimal proposals incorporate the variance function of the noisy evaluation in order to propose samples in regions that are more affected by noise. In this sense, we can informally state that the optimal proposal densities play the role of an acquisition function that also take into account the noise power.

References

[1] D. Luengo, L. Martino, M. Bugallo, V. Elvira, and S. Särkkä, "A survey of Monte Carlo methods for parameter estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2020, pp. 1–62, 2020.
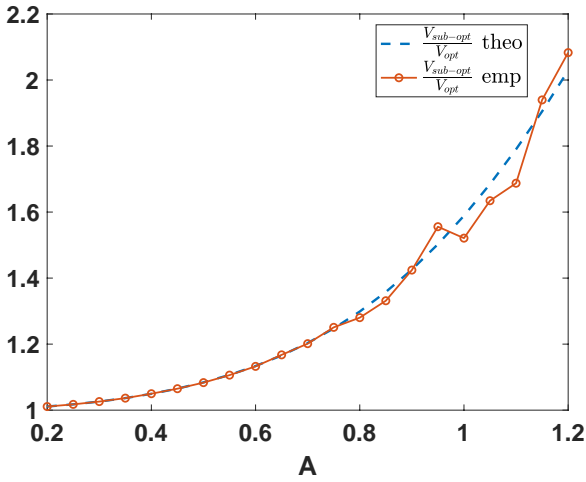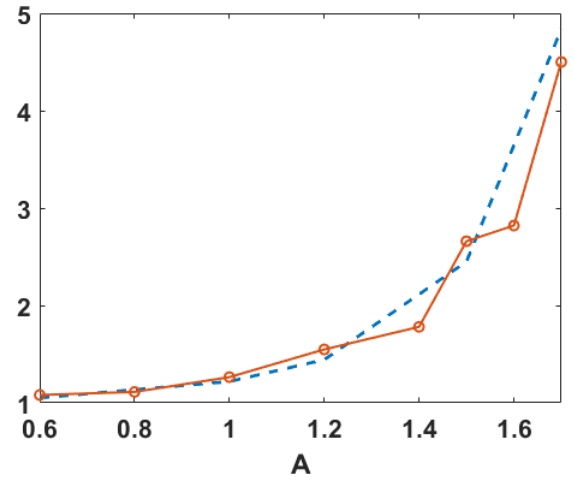
(a) $q_{\text{opt}}(x)$ for different values of A.
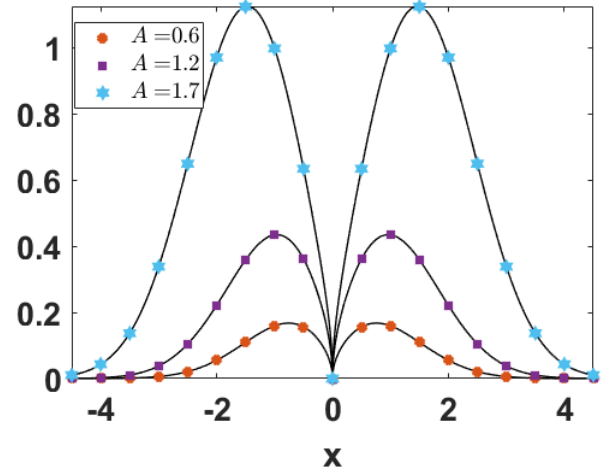
(b) $s(x)$ for different values of A.

Figure 2: Gaussian example. **(a)** Optimal proposals $q_{\text{opt}}(x)$ for different values of A, and the $q_{\text{sub-opt}}(x) = p(x)$ in dashed line; **(b)** The standard deviation $s(x)^2$ for different values of A.

(a) Uniform example.

(b) Gaussian example.

Figure 3: Theoretical and empirical ratio of variances $\frac{V_{\text{sub-opt}}}{V_{\text{opt}}}$ in the estimation of $\bar{Z} = 1$ for both experiments. The x-axis denotes the noise level (larger A means greater noise).

[2] P. Fearnhead, O. Papaspiliopoulos, G. O. Roberts, and A. Stuart, "Random-weight particle filtering of continuous time processes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 4, pp. 497–512, 2010.

[3] M.-N. Tran, M. Scharth, M. K. Pitt, and R. Kohn, "Importance sampling squared for Bayesian inference in latent variable models," *arXiv preprint arXiv:1309.3339*, 2013.

[4] L. Acerbi, "Variational Bayesian Monte Carlo with noisy likelihoods," *arXiv:2006.08655*, 2020.

[5] F. Llorente, L. Martino, J. Read, and D. Delgado, "A survey of Monte Carlo methods for noisy and costly densities with application to reinforcement learning," *arXiv:2108.00490*, 2021.

[6] C. Andrieu and G. O. Roberts, "The pseudo-marginal approach for efficient Monte Carlo computations," *The Annals of Statistics*, vol. 37, no. 2, pp. 697–725, 2009.

[7] J. Park and M. Haran, "Bayesian inference in the presence of intractable normalizing functions," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1372–1390, 2018.

[8] L. F. Price, C. C. Drovandi, A. Lee, and D. J. Nott, "Bayesian synthetic likelihood," *Journal of Computational and Graphical Statistics*, vol. 27, no. 1, pp. 1–11, 2018.

[9] M. U. Gutmann and J. Corander, "Bayesian optimization for likelihood–free inference of simulator–based statistical models," *Journal of Machine Learning Research*, 2016.

[10] R. Bardenet, A. Doucet, and C. Holmes, "On Markov chain Monte Carlo methods for tall data," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 1515–1557, 2017.

[11] M. Quiroz, R. Kohn, M. Villani, and M.-N. Tran, "Speeding up MCMC by efficient data subsampling," *Journal of the American Statistical Association*, 2018.

[12] M. P. Deisenroth, G. Neumann, and J. Peters, "A survey on policy search for robotics," *Foundations and trends in Robotics*, vol. 2, no. 1-2, pp. 388–403, 2013.

[13] K. Chatzilygeroudis, V. Vassiliades, F. Stulp, S. Calinon, and J.-B. Mouret, "A survey on policy search algorithms for learning robot controllers in a handful of trials," *IEEE Transactions on Robotics*, vol. 36, no. 2, pp. 328–347, 2019.

[14] A. B. Duncan, A. M. Stuart, and M.-T. Wolfram, "Ensemble inference methods for models with noisy and expensive likelihoods," *arXiv:2104.03384*, 2021.

[15] N. Bliznyuk, D. Ruppert, C. Shoemaker, R. Regis, S. Wild, and P. Mugunthan, "Bayesian calibration and uncertainty analysis for computationally expensive models using optimization and radial basis function approximation," *Journal of Computational and Graphical Statistics*, vol. 17, no. 2, pp. 270–294, 2008.

[16] D. H. Svendsen, L. Martino, and G. Camps-Valls, "Active emulation of computer codes with Gaussian processes–Application to remote sensing," *Pattern Recognition*, vol. 100, pp. 107103, 2020.

[17] F. Llorente, L. Martino, V. Elvira, D. Delgado, and J. Lopez-Santiago, "Adaptive quadrature schemes for Bayesian inference via active learning," *arXiv:2006.00535*, 2020.

[18] P. Fearnhead, O. Papaspiliopoulos, and G. O. Roberts, "Particle filters for partially observed diffusions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 4, pp. 755–777, 2008.

[19] O. D. Akyildiz, I. P. Marino, and J. Míguez, "Adaptive noisy importance sampling for stochastic optimization," in *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2017, pp. 1–5.

[20] D. Crisan and J. Miguez, "Nested particle filters for online parameter estimation in discrete-time state-space markov models," *Bernoulli*, vol. 24, no. 4A, pp. 3039–3086, 2018.

[21] N. Chopin, P.E. Jacob, and O. Papaspiliopoulos, "SMC2: an efficient algorithm for sequential analysis of state space models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 75, no. 3, pp. 397–426, 2013.

[22] L. Martino, J. Read, V. Elvira, and F. Louzada, "Cooperative parallel particle filters for online model selection and applications to urban mobility," *Digital Signal Processing*, vol. 60, pp. 172–185, 2017.

[23] A. Doucet, M. K Pitt, G. Deligiannidis, and R. Kohn, "Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator," *Biometrika*, vol. 102, no. 2, pp. 295–313, 2015.

[24] F. Llorente, L. Martino, D. Delgado, and J. Lopez-Santiago, "Marginal likelihood computation for model selection and hypothesis testing: an extensive review," *arXiv:2005.08334*, 2020.

[25] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, 2004.

[26] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 3, pp. 269–342, 2010.

[27] M. Järvenpää, M. U. Gutmann, A. Vehtari, and P. Marttinen, "Parallel Gaussian process surrogate Bayesian inference with noisy likelihood evaluations," *Bayesian Analysis*, vol. 16, no. 1, pp. 147–178, 2021.

[28] C. C. Drovandi, M. T. Moores, and R. J. Boys, "Accelerating pseudo-marginal MCMC using Gaussian processes," *Computational Statistics & Data Analysis*, vol. 118, pp. 1–17, 2018.

[29] A. Stuart and K. Ord, *Kendall's Advanced Theory of Statistics*, Arnold, London, 1998.

[30] D. V. Arnold, *Noisy optimization with evolution strategies*, vol. 8, Springer Science & Business Media, 2012.

Appendices

A   Covariance between $\widehat{E}_p$ and $\widehat{Z}$

We show that
$$\text{Cov}[\widehat{E}_p, \widehat{Z}] = \frac{1}{N}\mathbb{E}\left[\frac{f_p(\mathbf{x})(m(\mathbf{x})^2 + s(\mathbf{x})^2)}{q(\mathbf{x})^2}\right] - \frac{1}{N}E_p\bar{Z}.$$

First, recall that $\text{Cov}[\widehat{E}_p, \widehat{Z}] = \mathbb{E}[\widehat{E}_p\widehat{Z}] - E_p\bar{Z}$. By the law of iterated expectations,
$$\mathbb{E}[\widehat{E}_p\widehat{Z}] = \mathbb{E}[\mathbb{E}[\widehat{E}_p\widehat{Z}|\mathbf{x}_{1:N}]].$$

The inner expectation is

$$\mathbb{E}[\widehat{E}_p\widehat{Z}|\mathbf{x}_{1:N}] = \mathbb{E}\left[\frac{1}{N^2}\sum_{i=1}^{N}w_i^2 f_p(\mathbf{x}_i) + \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j>i}^{N}w_i w_j f_p(\mathbf{x}_i)\middle|\mathbf{x}_{1:N}\right]$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\frac{f_p(\mathbf{x}_i)(s(\mathbf{x}_i)^2 + m(\mathbf{x}_i)^2)}{q(\mathbf{x}_i)^2} + \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{m(\mathbf{x}_i)f(\mathbf{x}_i)}{q(\mathbf{x}_i)}\frac{m(\mathbf{x}_j)}{q(\mathbf{x}_j)}.$$

Hence, we obtain

$$\mathbb{E}\left[\mathbb{E}[\widehat{E}_p\widehat{Z}|\mathbf{x}_{1:N}]\right] = \frac{1}{N}\mathbb{E}\left[\frac{f(\mathbf{x})(s(\mathbf{x})^2 + m(\mathbf{x})^2)}{q(\mathbf{x})^2}\right]$$

$$+ \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\mathbb{E}\left[\frac{m(\mathbf{x}_i)f(\mathbf{x}_i)}{q(\mathbf{x}_i)}\right]\mathbb{E}\left[\frac{m(\mathbf{x}_j)}{q(\mathbf{x}_j)}\right]$$

$$= \frac{1}{N}\mathbb{E}\left[\frac{f(\mathbf{x})(s(\mathbf{x})^2 + m(\mathbf{x})^2)}{q(\mathbf{x})^2}\right] + \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j>i}^{N}E_p\bar{Z}$$

$$= \frac{1}{N}\mathbb{E}\left[\frac{f(\mathbf{x})(s(\mathbf{x})^2 + m(\mathbf{x})^2)}{q(\mathbf{x})^2}\right] + E_p\bar{Z}\left(1 - \frac{1}{N}\right).$$

Combining the results, we obtain the desired expression.