entropy

MDPI

Article

Bayesian Update with Importance Sampling: Required Sample Size

Daniel Sanz-Alonso * and Zijian Wang

Department of Statistics, University of Chicago, Chicago, IL 60637, USA; zijianwang@uchicago.edu
* Correspondence: sanzalonso@uchicago.edu

Abstract: Importance sampling is used to approximate Bayes' rule in many computational approaches to Bayesian inverse problems, data assimilation and machine learning. This paper reviews and further investigates the required sample size for importance sampling in terms of the χ^2-divergence between target and proposal. We illustrate through examples the roles that dimension, noise-level and other model parameters play in approximating the Bayesian update with importance sampling. Our examples also facilitate a new direct comparison of standard and optimal proposals for particle filtering.

Keywords: importance sampling; Bayesian update; sample size; singular limits; particle filtering

Citation: Sanz-Alonso, D.; Wang, Z. Bayesian Update with Importance Sampling: Required Sample Size. *Entropy* **2021**, *23*, 22. https://dx.doi.org/10.3390/e23010022

Received: 23 September 2020
Accepted: 23 December 2020
Published: 26 December 2020

1. Introduction

Importance sampling is a mechanism to approximate expectations with respect to a *target* distribution using independent weighted samples from a *proposal* distribution. The variance of the weights—quantified by the χ^2-divergence between target and proposal— gives both necessary and sufficient conditions on the sample size to achieve a desired worst-case error over large classes of test functions. This paper contributes to the understanding of importance sampling to approximate the Bayesian update, where the target is a posterior distribution obtained by conditioning the proposal to observed data. We consider illustrative examples where the χ^2-divergence between target and proposal admits a closed formula and it is hence possible to characterize explicitly the required sample size. These examples showcase the fundamental challenges that importance sampling encounters in high dimension and small noise regimes where target and proposal are far apart. They also facilitate a direct comparison of standard and optimal proposals for particle filtering.

We denote the target distribution by μ and the proposal by π and assume that both are probability distributions in Euclidean space $\mathbb{R}^d$. We further suppose that the target is absolutely continuous with respect to the proposal and denote by g the *un-normalized* density between target and proposal so that, for any suitable test function φ,

$$\int_{\mathbb{R}^d} \varphi(u)\mu(du) = \frac{\int_{\mathbb{R}^d} \varphi(u)g(u)\pi(du)}{\int_{\mathbb{R}^d} g(u)\pi(du)}. \tag{1}$$

We write this succinctly as $\mu(\varphi) = \pi(\varphi g)/\pi(g)$. For simplicity of exposition, we will assume that g is positive π-almost surely. Importance sampling approximates $\mu(\varphi)$ using independent samples $\{u^{(n)}\}_{n=1}^N$ from the proposal π, computing the numerator and denominator in (1) by Monte Carlo integration,

$$\begin{aligned} \mu(\varphi) &\approx \frac{\frac{1}{N}\sum_{n=1}^N \varphi(u^{(n)})g(u^{(n)})}{\frac{1}{N}\sum_{n=1}^N g(u^{(n)})} \\ &= \sum_{n=1}^N w^{(n)}\varphi(u^{(n)}), \qquad w^{(n)} := \frac{g(u^{(n)})}{\sum_{\ell=1}^N g(u^{(\ell)})}. \end{aligned} \tag{2}$$

The weights $w^{(n)}$—called *autonormalized* or self-normalized since they add up to one—can be computed as long as the un-normalized density g can be evaluated point-wise; knowledge of the normalizing constant $\pi(g)$ is not needed. We write (2) briefly as $\mu(\varphi) \approx \mu^N(\varphi)$, where μ^N is the *random* autonormalized particle approximation measure

$$\mu^N := \sum_{n=1}^{N} w^{(n)} \delta_{u^{(n)}}, \qquad u^{(n)} \overset{\text{i.i.d.}}{\sim} \pi. \tag{3}$$

This paper is concerned with the study of importance sampling in Bayesian formulations to inverse problems, data assimilation and machine learning tasks [1–5], where the relationship $\mu(du) \propto g(u)\pi(du)$ arises from application of Bayes' rule $\mathbb{P}(u|y) \propto \mathbb{P}(y|u)\,\mathbb{P}(u)$; we interpret $u \in \mathbb{R}^d$ as a parameter of interest, $\pi \equiv \mathbb{P}(u)$ as a prior distribution on u, $g(u) \equiv g(u;y) \equiv \mathbb{P}(y|u)$ as a likelihood function which tacitly depends on observed data $y \in \mathbb{R}^k$, and $\mu \equiv \mathbb{P}(u|y)$ as the posterior distribution of u given y. With this interpretation and terminology, the goal of importance sampling is to approximate posterior expectations using prior samples. Since the prior has fatter tails than the posterior, the Bayesian setting poses further structure into the analysis of importance sampling. In addition, there are several specific features of the application of importance sampling in Bayesian inverse problems, data assimilation and machine learning that shape our presentation and results.

First, Bayesian formulations have the potential to provide uncertainty quantification by computing *several* posterior quantiles. This motivates considering a worst-case error analysis [6] of importance sampling over large classes of test functions φ or, equivalently, bounding a certain distance between the random particle approximation measure μ^N and the target μ, see [1]. As we will review in Section 2, a key quantity in controlling the error of importance sampling with bounded test functions is the χ^2-divergence between target and proposal, given by

$$d_{\chi^2}(\mu\|\pi) = \frac{\pi(g^2)}{\pi(g)^2} - 1.$$

Second, importance sampling in inverse problems, data assimilation and machine learning applications is often used as a building block of more sophisticated computational methods, and in such a case there may be little or no freedom in the choice of proposal. For this reason, throughout this paper we view both target and proposal as given and we focus on investigating the required sample size for accurate importance sampling with bounded test functions, following a similar perspective as [1,7,8]. The complementary question of how to choose the proposal to achieve a small variance for a given test function is not considered here. This latter question is of central interest in the simulation of rare events [9] and has been widely studied since the introduction of importance sampling in [10,11], leading to a plethora of adaptive importance sampling schemes [12].

Third, high dimensional and small noise settings are standard in inverse problems, data assimilation and machine learning, and it is essential to understand the scalability of sampling algorithms in these challenging regimes. The curse of dimension of importance sampling has been extensively investigated [1,13–17]. The early works [13,14] demonstrated a *weight collapse* phenomenon, by which unless the number of samples is scaled exponentially with the dimension of the parameter, the maximum weight converges to one. The paper [1] also considered small noise limits and further emphasized the need to define precisely the dimension of learning problems. Indeed, while many inverse problems, data assimilation models and machine learning tasks are defined in terms of millions of parameters, their *intrinsic dimension* can be substantially smaller since (i) all parameters may not be equally important; (ii) a priori information about some parameters may be available; and (iii) the data may be lower dimensional than the parameter space. If the intrinsic dimension is still large, which occurs often in applications in geophysics and machine learning, it is essential to leverage the correlation structure of the parameters or the observations by performing localization [18–20]. Local particle filters are reviewed in [21] and their potential to beat the curse of dimension is investigated from a theoretical

viewpoint in [16]. Localization is popular in ensemble Kalman filters [20] and has been employed in Markov chain Monte Carlo [22,23]. Our focus in this paper is not on localization but rather on providing a unified and accessible understanding of the roles that dimension, noise-level and other model parameters play in approximating the Bayesian update. We will do so through examples where it is possible to compute explicitly the χ^2-divergence between target and proposal, and hence the required sample size.

Finally, in the Bayesian context the normalizing constant $\pi(g)$ represents the marginal likelihood and is often computationally intractable. This motivates our focus on the *autonormalized* importance sampling estimator in (2), which estimates *both* $\pi(g\varphi)$ and $\pi(g)$ using Monte Carlo integration, as opposed to un-normalized variants of importance sampling [8].

Main Goals, Specific Contributions and Outline

The main goal of this paper is to provide a rich and unified understanding of the use of importance sampling to approximate the Bayesian update, while keeping the presentation accessible to a large audience. In Section 2 we investigate the required sample size for importance sampling in terms of the χ^2-divergence between target and proposal. Section 3 builds on the results in Section 2 to illustrate through numerous examples the fundamental challenges that importance sampling encounters when approximating the Bayesian update in small noise and high dimensional settings. In Section 4 we show how our concrete examples facilitate a new direct comparison of standard and optimal proposals for particle filtering. These examples also allow us to identify model problems where the advantage of the optimal proposal over the standard one can be dramatic.

Next, we provide further details on the specific contributions of each section and link them to the literature. We refer to [1] for a more exhaustive literature review.

- Section 2 provides a unified perspective on the sufficiency and necessity of having a sample size of the order of the χ^2-divergence between target and proposal to guarantee accurate importance sampling with bounded test functions. Our analysis and presentation are informed by the specific features that shape the use of importance sampling to approximate Bayes' rule. The key role of the second moment of the χ^2-divergence has long been acknowledged [24,25], and it is intimately related to an effective sample size used by practitioners to monitor the performance of importance sampling [26,27]. A topic of recent interest is the development of adaptive importance sampling schemes where the proposal is chosen by minimizing—over some admissible family of distributions—the χ^2-divergence with respect to the target [28,29]. The main original contributions of Section 2 are Proposition 2 and Theorem 1, which demonstrate the *necessity* of suitably increasing the sample size with the χ^2-divergence along singular limit regimes. The idea of Proposition 2 is inspired by [7], but adapted here from relative entropy to χ^2-divergence. Our results complement sufficient conditions on the sample size derived in [1] and necessary conditions for *un-normalized* (as opposed to autonormalized) importance sampling in [8].
- In Section 3, Proposition 4 gives a closed formula for the χ^2-divergence between posterior and prior in a linear-Gaussian Bayesian inverse problem setting. This formula allows us to investigate the scaling of the χ^2-divergence (and thereby the rate at which the sample size needs to grow) in several singular limit regimes, including small observation noise, large prior covariance and large dimension. Numerical examples motivate and complement the theoretical results. Large dimension and small noise singular limits were studied in [1] in a *diagonal setting*. The results here are generalized to a nondiagonal setting, and the presentation is simplified by using the closed formula in Proposition 4. Moreover, we include singular limits arising from large prior covariance. In an infinite dimensional setting, Corollary 1 establishes an equivalence between absolute continuity, finite χ^2-divergence and finite intrinsic dimension. A similar result was proved in more generality in [1] using the advanced

theory of Gaussian measures in Hilbert space [30]; our presentation and proof here are elementary, while still giving the same degree of understanding.

- In Section 4 we follow [1,13–15,31] and investigate the use of importance sampling to approximate Bayes' rule within one filtering step in a linear-Gaussian setting. We build on the examples and results in Section 3 to identify model regimes where the performance of standard and optimal proposals can be dramatically different. We refer to [2,32] for an introduction to standard and optimal proposals for particle filtering and to [33] for a more advanced presentation. The main original contribution of this section is Theorem 2, which gives a direct comparison of the χ^2-divergence between target and standard/optimal proposals. This result improves on [1], where only a comparison between the intrinsic dimension was established.

2. Importance Sampling and χ^2-Divergence

The aim of this section is to demonstrate the central role of the χ^2-divergence between target and proposal in determining the accuracy of importance sampling. In Section 2.1 we show how the χ^2-divergence arises in both sufficient and necessary conditions on the sample size for accurate importance sampling with bounded test functions. Section 2.2 describes a well-known connection between the effective sample size and the χ^2-divergence. Our investigation of importance sampling to approximate the Bayesian update—developed in Sections 3 and 4—will make use of a closed formula for the χ^2-divergence between Gaussians, which we include in Section 2.3 for later reference.

2.1. Sufficient and Necessary Sample Size

Here we provide general sufficient and necessary conditions on the sample size in terms of

$$\rho := d_{\chi^2}(\mu \| \pi) + 1.$$

We first review upper-bounds on the worst-case bias and mean-squared error of importance sampling with bounded test functions, which imply that accurate importance sampling is guaranteed if $N \gg \rho$. The proof of the bound for the mean-squared error can be found in [1] and the bound for the bias in [2].

Proposition 1 (Sufficient Sample Size)**.** *It holds that*

$$\sup_{|\varphi|_\infty \leq 1} \left| \mathbb{E}\left[\mu^N(\varphi) - \mu(\varphi) \right] \right| \leq \frac{4}{N}\rho,$$
$$\sup_{|\varphi|_\infty \leq 1} \mathbb{E}\left[\left(\mu^N(\varphi) - \mu(\varphi) \right)^2 \right] \leq \frac{4}{N}\rho.$$

The next result shows the existence of bounded test functions for which the error may be large with a high probability if $N \ll \rho$. The idea is taken from [7], but we adapt it here to obtain a result in terms of the χ^2-divergence rather than relative entropy. We denote by $\mathrm{g} := g/\pi(g)$ the *normalized* density between μ and π, and note that $\rho = \pi(\mathrm{g}^2) = \mu(\mathrm{g})$.

Proposition 2 (Necessary Sample Size)**.** *Let $U \sim \mu$. For any $N \geq 1$ and $\alpha \in (0,1)$ there exists a test function φ with $|\varphi|_\infty \leq 1$ such that*

$$\mathbb{P}\Big(|\mu^N(\varphi) - \mu(\varphi)| = \mathbb{P}(\mathrm{g}(U) > \alpha\rho) \Big) \geq 1 - \frac{N}{\alpha\rho}. \tag{4}$$

Proof. Observe that for the test function $\varphi(u) := \mathbb{1}\{\mathrm{g}(u) \leq \alpha\rho\}$, we have $\mu(\varphi) = \mathbb{P}\big(\mathrm{g}(U) \leq \alpha\rho\big)$. On the other hand, $\mu^N(\varphi) = 1$ if and only if $\mathrm{g}(u^{(n)}) \leq \alpha\rho$ for all $1 \leq n \leq N$. This implies that

$$\mathbb{P}\Big(|\mu^N(\varphi) - \mu(\varphi)| = \mathbb{P}(\mathrm{g}(U) > \alpha\rho)\Big) \geq 1 - N\,\mathbb{P}(\mathrm{g}(u^{(1)}) > \alpha\rho) \geq 1 - \frac{N}{\alpha\rho}. \tag{5}$$

□

The power of Proposition 2 is due to the fact that in some singular limit regimes the distribution of $\mathrm{g}(U)$ concentrates around its expected value ρ. In such a case, for any fixed $\alpha \in (0,1)$ the probability of the event $\mathrm{g}(U) > \alpha\rho$ will not vanish as the singular limit is approached. This idea will become clear in the proof of Theorem 1 below.

In Sections 3 and 4 we will investigate the required sample size for importance sampling approximation of the Bayesian update in various singular limits, where target and proposal become further apart as a result of reducing the observation noise, increasing the prior uncertainty or increasing the dimension of the problem. To formalize the discussion in a general abstract setting, let $\{(\mu_\theta, \pi_\theta)\}_{\theta>0}$ be a family of targets and proposals such that $\rho_\theta := d_{\chi^2}(\mu_\theta \| \pi_\theta) \to \infty$ as $\theta \to \infty$. The parameter θ may represent for instance the size of the precision of the observation noise, the size of the prior covariance or a suitable notion of dimension. Our next result shows a clear dichotomy in the performance of importance sampling along the singular limit depending on whether the sample size grows sublinearly or superlinearly with ρ_θ.

Theorem 1. *Suppose that* $\rho_\theta \to \infty$ *and that* $\mathcal{V} := \sup_\theta \frac{\mathbb{V}[\mathrm{g}_\theta(U_\theta)]}{\rho_\theta^2} < 1$. *Let* $\delta > 0$.

i *If* $N_\theta = \rho_\theta^{1+\delta}$, *then*

$$\lim_{\theta\to\infty} \sup_{|\varphi|_\infty \leq 1} \mathbb{E}\big[\big(\mu_\theta^{N_\theta}(\varphi) - \mu_\theta(\varphi)\big)^2\big] = 0. \tag{6}$$

ii *If* $N_\theta = \rho_\theta^{1-\delta}$, *then there exists a fixed* $c \in (0,1)$ *such that*

$$\lim_{\theta\to\infty} \sup_{|\varphi|_\infty \leq 1} \mathbb{P}\Big(|\mu_\theta^{N_\theta}(\varphi) - \mu_\theta(\varphi)| > c\Big) = 1. \tag{7}$$

Proof. The proof of (i) follows directly from Proposition 1. For (ii) we fix $\alpha \in (0, 1-\mathcal{V})$ and $c \in \Big(0, 1 - \frac{\mathcal{V}}{(1-\alpha)^2}\Big)$. Let $\varphi_\theta(u) := \mathbb{1}(\mathrm{g}_\theta(u) \leq \alpha\rho_\theta)$ as in the proof of Proposition 2. Then,

$$\mathbb{P}\big(\mathrm{g}_\theta(U_\theta) > \alpha\rho_\theta\big) \geq 1 - \mathbb{P}\big(|\rho_\theta - \mathrm{g}_\theta(U_\theta)| \geq (1-\alpha)\rho_\theta\big) \geq 1 - \frac{\mathbb{V}[\mathrm{g}_\theta(U_\theta)]}{(1-\alpha)^2\rho_\theta^2} \geq 1 - \frac{\mathcal{V}}{(1-\alpha)^2} > c.$$

The bound in (5) implies that

$$\mathbb{P}\Big(|\mu_\theta^{N_\theta}(\varphi_\theta) - \mu_\theta(\varphi_\theta)| > c\Big) \geq \mathbb{P}\Big(|\mu_\theta^{N}(\varphi_\theta) - \mu_\theta(\varphi_\theta)| = \mathbb{P}(\mathrm{g}_\theta(U_\theta) > \alpha\rho_\theta)\Big) \geq 1 - \frac{N_\theta}{\alpha\rho_\theta}.$$

This completes the proof, since if $N_\theta = \rho_\theta^{1-\delta}$ the right-hand side goes to 1 as $\theta \to \infty$. □

Remark 1. *As noted in [1], the bound in Proposition 1 is sharp in the asymptotic limit* $N \to \infty$. *This implies that, for any* fixed θ, *the bound* $4\rho_\theta/N$ *becomes sharp as* $N \to \infty$. *We point out that this statement does not provide direct understanding of the joint limit* $\theta, N_\theta \to \infty$ *analyzed in Theorem 1.*

The assumption that $\mathcal{V} < 1$ can be verified for some singular limits of interest, in particular for small noise and large prior covariance limits studied in Sections 3 and 4; details

will be given in Example 1. While the assumption $\mathcal{V} < 1$ may fail to hold in high dimensional singular limit regimes, the works [13,14] and our numerical example in Section 4.4 provide compelling evidence of the need to suitably scale N with ρ along those singular limits in order to avoid a weigh-collapse phenomenon. Further theoretical evidence was given for un-normalized importance sampling in [8].

2.2. χ^2-Divergence and Effective Sample Size

The previous subsection provides theoretical nonasymptotic and asymptotic evidence that a sample size larger than ρ is necessary and sufficient for accurate importance sampling. Here we recall a well known connection between the χ^2-divergence and the effective sample size

$$\mathrm{ESS} := \frac{1}{\sum_{n=1}^N (w^{(n)})^2}, \tag{8}$$

widely used by practitioners to monitor the performance of importance sampling. Note that always $1 \leq \mathrm{ESS} \leq N$; it is intuitive that $\mathrm{ESS} = 1$ if the maximum weight is one and $\mathrm{ESS} = N$ if the maximum weight is $1/N$. To see the connection between ESS and ρ, note that

$$\frac{\mathrm{ESS}}{N} = \frac{1}{N\sum_{n=1}^N (w^{(n)})^2} = \frac{\left(\sum_{n=1}^N g(u^{(n)})\right)^2}{N\sum_{n=1}^N g(u^{(n)})^2} = \frac{\left(\frac{1}{N}\sum_{n=1}^N g(u^{(n)})\right)^2}{\frac{1}{N}\sum_{n=1}^N g(u^{(n)})^2} \approx \frac{\pi(g)^2}{\pi(g^2)}.$$

Therefore, $\mathrm{ESS} \approx N/\rho$: if the sample-based estimate of ρ is significantly larger than N, ESS will be small which gives a warning sign that a larger sample size N may be needed.

2.3. χ^2-Divergence between Gaussians

We conclude this section by recalling an analytical expression for the χ^2-divergence between Gaussians. In order to make our presentation self-contained, we include a proof in Appendix A.

Proposition 3. *Let $\mu = \mathcal{N}(m, C)$ and $\pi = \mathcal{N}(0, \Sigma)$. If $2\Sigma \succ C$, then*

$$\rho = \frac{|\Sigma|}{\sqrt{|2\Sigma - C||C|}} \exp\left(m'(2\Sigma - C)^{-1}m\right).$$

Otherwise, $\rho = \infty$.

It is important to note that nondegenerate Gaussians $\mu = \mathcal{N}(m, C)$ and $\pi = \mathcal{N}(0, \Sigma)$ in $\mathbb{R}^d$ are always equivalent. However, $\rho = \infty$ unless $2\Sigma \succ C$. In Sections 3 and 4 we will interpret μ as a posterior and π as a prior, in which case automatically $\Sigma \succ C$ and $\rho < \infty$.

3. Importance Sampling for Inverse Problems

In this section we study the use of importance sampling in a linear Bayesian inverse problem setting where the target and the proposal represent, respectively, the posterior and the prior distribution. In Section 3.1 we describe our setting and we also derive an explicit formula for the χ^2-divergence between the posterior and the prior. This explicit formula allows us to determine the scaling of the χ^2-divergence in small noise regimes (Section 3.2), in the limit of large prior covariance (Section 3.3) and in a high dimensional limit (Section 3.4). Our overarching goal is to show how the sample size for importance sampling needs to grow along these limiting regimes in order to maintain the same level of accuracy.

3.1. Inverse Problem Setting and χ^2-Divergence between Posterior and Prior

Let $A \in \mathbb{R}^{k\times d}$ be a given *design* matrix and consider the linear inverse problem of recovering $u \in \mathbb{R}^d$ from data $y \in \mathbb{R}^k$ related by

$$y = Au + \eta, \qquad \eta \sim \mathcal{N}(0, \Gamma), \tag{9}$$

where η represents measurement noise. We assume henceforth that we are in the under-determined case $k \leq d$, and that A is full rank. We follow a Bayesian perspective and set a Gaussian prior on u, $u \sim \pi = \mathcal{N}(0, \Sigma)$. We assume throughout that Σ and Γ are given symmetric positive definite matrices. The solution to the Bayesian formulation of the inverse problem is the posterior distribution μ of u given y. We are interested in studying the performance of importance sampling with proposal π (the prior) and target μ (the posterior). We recall that under this linear-Gaussian model the posterior distribution is Gaussian [2], and we denote it by $\mu = \mathcal{N}(m, C)$. In order to characterize the posterior mean m and covariance C, we introduce standard data assimilation notation

$$\begin{aligned} S &:= A\Sigma A' + \Gamma, \\ K &:= \Sigma A' S^{-1}, \end{aligned}$$

where K is the Kalman gain. Then we have

$$\begin{aligned} m &= Ky, \\ C &= (I - KA)\Sigma. \end{aligned} \tag{10}$$

Proposition 3 allows us to obtain a closed formula for the quantity $\rho = d_{\chi^2}(\mu\|\pi) + 1$, noting that (10) implies that

$$\begin{aligned} 2\Sigma - C &= (I + KA)\Sigma \\ &= \Sigma + \Sigma A' S^{-1} A\Sigma \succ 0. \end{aligned}$$

The proof of the following result is then immediate and therefore omitted.

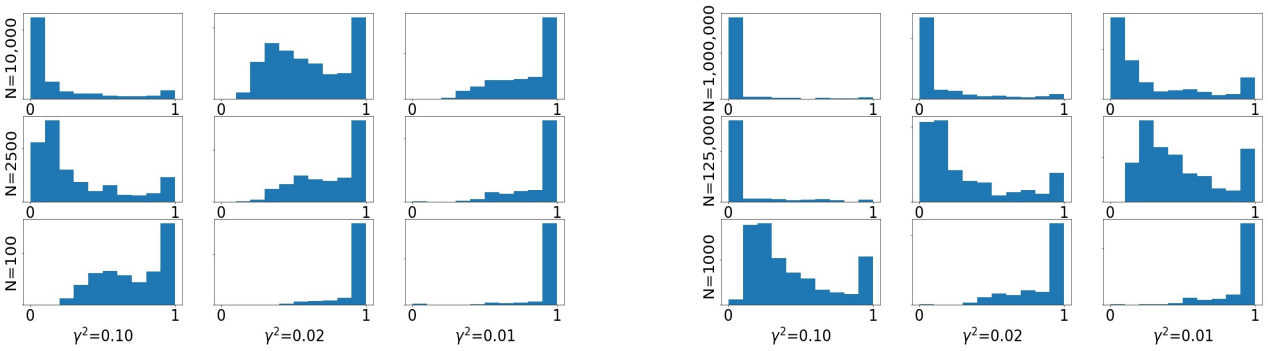
Proposition 4. *Consider the inverse problem* (9) *with prior* $u \sim \pi = \mathcal{N}(0, \Sigma)$ *and posterior* $\mu = \mathcal{N}(m, C)$ *with* m *and* C *defined in* (10)*. Then* $\rho = d_{\chi^2}(\mu\|\pi) + 1$ *admits the explicit characterization*

$$\rho = (|I + KA||I - KA|)^{-\frac{1}{2}} \exp\Big(y' K' [(I + KA)\Sigma]^{-1} Ky\Big).$$

In the following two subsections we employ this result to derive by direct calculation the rate at which the posterior and prior become further apart —in χ^2-divergence— in small noise and large prior regimes. To carry out the analysis we use parameters $\gamma^2, \sigma^2 > 0$ to scale the noise covariance, $\gamma^2\Gamma$, and the prior covariance, $\sigma^2\Sigma$.

3.2. Importance Sampling in Small Noise Regime

To illustrate the behavior of importance sampling in small noise regimes, we first introduce a motivating numerical study. A similar numerical setup was used in [13] to demonstrate the curse of dimension of importance sampling. We consider the inverse problem setting in Equation (9) with $d = k = 5$ and noise covariance $\gamma^2\Gamma$. We conduct 18 numerical experiments with a fixed data y. For each experiment, we perform importance sampling 400 times and report in Figure 1 a histogram with the largest autonormalized weight in each of the 400 realizations. The 18 experiments differ in the sample size N and the size of the observation noise γ^2. In both Figure 1a,b we consider three choices of N (rows) and three choices of γ^2 (columns). These choices are made so that in Figure 1a it holds that $N = \gamma^{-4}$ along the bottom-left to top-right diagonal, while in Figure 1b $N = \gamma^{-6}$ along the same diagonal.



(**a**) $N = \gamma^{-4}$. (**b**) $N = \gamma^{-6}$.

Figure 1. Noise scaling with $d = k = 5$. Each histogram represents the empirical distribution of the largest autonormalized weight of importance sampling with a given choice of sample size N and noise level γ^2. The empirical distribution is obtained using 400 sets of random weights and the histograms are arranged so that in (**a**) $N = \gamma^{-4}$ along the bottom-left to top-right diagonal, while in (**b**) $N = \gamma^{-6}$ along the same diagonal. With scaling γ^{-4} the distribution of the maximum weight concentrates around 1 along this diagonal, suggesting weight collapse. In contrast, with scaling γ^{-6} weight collapse is avoided with high probability.

We can see from Figure 1a that $N = \gamma^{-4}$ is not a fast enough growth of N to avoid weight collapse: the histograms skew to the right along the bottom-left to top-right diagonal, suggesting that weight collapse (i.e., one weight dominating the rest, and therefore the variance of the weights being large) is bound to occur in the joint limit $N \to \infty$, $\gamma \to 0$ with $N = \gamma^{-4}$. In contrast, the histograms in Figure 1b skew to the left along the same diagonal, suggesting that the probability of weight collapse is significantly reduced if $N = \gamma^{-6}$. We observe a similar behavior with other choices of dimension d by conducting experiments with sample sizes $N = \gamma^{-d+1}$ and $N = \gamma^{-d-1}$, and we include the histograms with $d = k = 4$ in Appendix C. Our next result shows that these empirical findings are in agreement with the scaling of the χ^2-divergence between target and proposal in the small noise limit.

Proposition 5. *Consider the inverse problem setting*

$$y = Au + \eta, \qquad \eta = \mathcal{N}(0, \gamma^2\Gamma), \qquad u \sim \pi = \mathcal{N}(0, \Sigma).$$

Let μ_γ denote the posterior and let $\rho_\gamma = d_{\chi^2}(\mu_\gamma \| \pi) + 1$. Then, for almost every y,

$$\rho_\gamma \sim \mathcal{O}(\gamma^{-k})$$

in the small noise limit $\gamma \to 0$.

Proof. Let $K_\gamma = \Sigma A'(A\Sigma A' + \gamma^2\Gamma)^{-1}$ denote the Kalman gain. We observe that $K_\gamma \to \Sigma A'(A\Sigma A')^{-1}$ under our standing assumption that A is full rank. Let $U'\Xi V$ be the singular value decompostion of $\Gamma^{-\frac{1}{2}} A \Sigma^{\frac{1}{2}}$ and $\{\xi_i\}_{i=1}^k$ be the singular values. Then we have

$$\begin{aligned} K_\gamma A &\sim \Sigma^{\frac{1}{2}} A' \Gamma^{-\frac{1}{2}} (\Gamma^{-\frac{1}{2}} A \Sigma A' \Gamma^{-\frac{1}{2}} + \gamma^2 I)^{-1} \Gamma^{-\frac{1}{2}} A \Sigma^{\frac{1}{2}} \\ &= V'\Xi' U (U'\Xi V V'\Xi' U + \gamma^2 I)^{-1} U'\Xi\, V \\ &\sim \Xi'(\Xi\Xi' + \gamma^2 I)^{-1}\Xi, \end{aligned}$$

where here "$\sim$" denotes matrix similarity. It follows that $I + K_\gamma A$ converges to a finite limit, and so does the exponent $y' K_\gamma' \Sigma^{-1} (I + K_\gamma A)^{-1} K_\gamma y$ in Proposition 4. On the other hand,

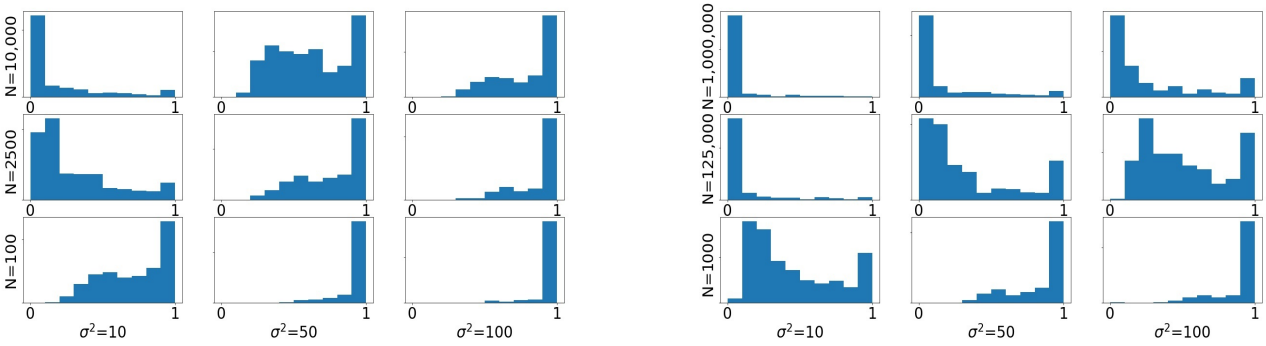
$$(|I + K_\gamma A||I - K_\gamma A|)^{-\frac{1}{2}} = \Big(\prod_{i=1}^{k} \frac{\gamma^2}{\xi_i^2 + \gamma^2}\Big)^{-\frac{1}{2}} \sim \mathcal{O}(\gamma^{-k})$$

as $\gamma \to 0$. The conclusion follows. □

Remark 2. *The scaling of ρ with γ^2 obtained in Proposition 5 agrees with the lower bound reported in Table 1 in [1], which was derived in a diagonalized setting.*

3.3. Importance Sampling and Prior Scaling

Here we illustrate the behavior of importance sampling in the limit of large prior covariance. We start again with a motivating numerical example, similar to the one reported in Figure 1. The behavior is analogous to the small noise regime, which is expected since the *ratio* of prior and noise covariances determines the closeness between target and proposal. Figure 2 shows that when $d = k = 5$ weight collapse is observed frequently when the sample size N grows as σ^4, but not so often with sample size $N = \sigma^6$. Similar histograms with $d = k = 4$ are included in Appendix C. These empirical results are in agreement with the theoretical growth rate of the χ^2-divergence between target and proposal in the limit of large prior covariance, as we prove next.



(**a**) $N = \sigma^4$. (**b**) $N = \sigma^6$.

Figure 2. Prior scaling with $d = k = 5$. The setting is similar to the one considered in Figure 1.

Proposition 6. *Consider the inverse problem setting*

$$y = Au + \eta, \qquad \eta \sim \mathcal{N}(0, \Gamma), \qquad u \sim \pi_\sigma = \mathcal{N}(0, \sigma^2 \Sigma).$$

Let μ_σ denote the posterior and $\rho_\sigma = d_{\chi^2}(\mu_\sigma \| \pi_\sigma) + 1$. Then, for almost every y,

$$\rho_\sigma \sim \mathcal{O}(\sigma^d)$$

in the large prior limit $\sigma \to \infty$.

Proof. Let $\Sigma_\sigma = \sigma^2 \Sigma$, let $K_\sigma = \Sigma_\sigma A' (A \Sigma_\sigma A' + \Gamma)^{-1}$ be the Kalman gain. Observing that $K_\sigma = K_{\gamma = \frac{1}{\sigma}}$, we apply Proposition 5 and deduce that when $\sigma \to \infty$:

1. $K_\sigma \to \Sigma A' (A \Sigma A' + \gamma^2 \Gamma)^{-1}$;
2. $I + K_\sigma A$ has a well-defined and invertible limit;
3. $|I - K_\sigma A|^{-\frac{1}{2}} \sim \mathcal{O}(\sigma^k)$.

On the other hand, we notice that the quadratic term

$$K_\sigma' \Sigma_\sigma^{-1} (I + K_\sigma A)^{-1} K_\sigma = \sigma^{-2} K_\sigma' \Sigma (I + K_\sigma A)^{-1} K_\sigma$$

vanishes in limit. The conclusion follows by Proposition 4. □

3.4. Importance Sampling in High Dimension

In this subsection we study importance sampling in high dimensional limits. To that end, we let $\{a_i\}_{i=1}^\infty$, $\{\gamma_i^2\}_{i=1}^\infty$ and $\{\sigma_i^2\}_{i=1}^\infty$ be infinite sequences and we define, for any $d \geq 1$,

$$\begin{aligned} A_{1:d} &:= \operatorname{diag}\left\{a_1, \ldots, a_d\right\} \in \mathbb{R}^{d\times d}, \\ \Gamma_{1:d} &:= \operatorname{diag}\left\{\gamma_1^2, \ldots, \gamma_d^2\right\} \in \mathbb{R}^{d\times d}, \\ \Sigma_{1:d} &:= \operatorname{diag}\left\{\sigma_1^2, \ldots, \sigma_d^2\right\} \in \mathbb{R}^{d\times d}. \end{aligned}$$

We then consider the inverse problem of reconstructing $u \in \mathbb{R}^d$ from data $y \in \mathbb{R}^d$ under the setting

$$y = A_{1:d} u + \eta, \qquad \eta \sim \mathcal{N}(0, \Gamma_{1:d}), \qquad u \sim \pi_{1:d} = \mathcal{N}(0, \Sigma_{1:d}). \tag{11}$$

We denote the corresponding posterior distribution by $\mu_{1:d}$, which is Gaussian with a diagonal covariance. Given observation y, we may find the posterior distribution μ_i of u_i by solving the one dimensional linear-Gaussian inverse problem

$$y_i = a_i u_i + \eta_i, \qquad \eta_i \sim \mathcal{N}(0, \gamma_i^2), \qquad 1 \leq i \leq d, \tag{12}$$

with prior $\pi_i = \mathcal{N}(0, \sigma_i^2)$. In this way we have defined, for each $d \in \mathbb{N} \cup \{\infty\}$, an inverse problem with prior and posterior

$$\pi_{1:d} = \prod_{i=1}^d \pi_i, \qquad \mu_{1:d} = \prod_{i=1}^d \mu_i. \tag{13}$$

In Section 3.4.1 we include an explicit calculation in the one dimensional inverse setting (12), which will be used in Section 4.4 to establish the rate of growth of $\rho_d = d_{\chi^2}(\mu_{1:d} \| \pi_{1:d})$ and thereby how the sample size needs to be scaled along the high dimensional limit $d \to \infty$ to maintain the same accuracy. Finally, in Section 3.4.3 we establish from first principles and our simple one dimensional calculation the equivalence between (i) certain notion of dimension being finite; (ii) $\rho_\infty < \infty$; and (iii) absolute continuity of $\mu_{1:\infty}$ with respect to $\pi_{1:\infty}$.

3.4.1. One Dimensional Setting

Let $a \in \mathbb{R}$ be given and consider the one dimensional inverse problem of reconstructing $u \in \mathbb{R}$ from data $y \in \mathbb{R}$, under the setting

$$y = au + \eta, \qquad \eta \sim \mathcal{N}(0, \gamma^2), \qquad u \sim \pi = \mathcal{N}(0, \sigma^2). \tag{14}$$

By defining

$$g(u) := \exp\Big(-\frac{a^2}{2\gamma^2} u^2 + \frac{ay}{\gamma^2} u\Big),$$

we can write the posterior density $\mu(du)$ as $\mu(du) \propto g(u)\pi(du)$. The next result gives a simplified closed formula for $\rho = d_{\chi^2}(\mu \| \pi) + 1$. In addition, it gives a closed formula for the Hellinger integral

$$\mathcal{H}(\mu, \pi) := \frac{\pi\big(g^{\frac{1}{2}}\big)}{\pi(g)^{\frac{1}{2}}},$$

which will facilitate the study of the case $d = \infty$ in Section 3.4.3.

Lemma 1. *Consider the inverse problem in (14). Let* $\lambda := a^2\sigma^2/\gamma^2$ *and* $z^2 := \frac{y^2}{a^2\sigma^2+\gamma^2}$. *Then, for any* $\ell > 0$,

$$\frac{\pi(g^\ell)}{\pi(g)^\ell} = \frac{(\lambda+1)^{\frac{\ell}{2}}}{\sqrt{\ell\lambda+1}} \exp\Big(\frac{(\ell^2-\ell)\lambda}{2(\ell\lambda+1)} z^2\Big). \tag{15}$$

In particular,

$$\rho = \frac{\lambda+1}{\sqrt{2\lambda+1}} \exp\Big(\frac{\lambda}{2\lambda+1} z^2\Big), \tag{16}$$

$$\mathcal{H}(\mu,\pi) = \sqrt{\frac{2\sqrt{\lambda+1}}{\lambda+2}} \exp\Big(-\frac{\lambda z^2}{4(\lambda+2)}\Big). \tag{17}$$

Proof. A direct calculation shows that

$$\pi(g) = \frac{1}{\sqrt{\lambda+1}} \exp\Big(\frac{1}{2}\frac{\lambda y^2}{a^2\sigma^2+\gamma^2}\Big).$$

The same calculation, but replacing γ^2 by γ^2/ℓ and λ by $\ell\lambda$, gives similar expressions for $\pi(g^\ell)$, which leads to (15). The other two equations follow by setting ℓ to be 2 and $\frac{1}{2}$. □

Lemma 1 will be used in the two following subsections to study high dimensional limits. Here we show how this lemma also allows us to verify directly that the assumption $\mathcal{V} < 1$ in Theorem 1 holds in small noise and large prior limits.

Example 1. *Consider a sequence of inverse problems of the form (14) with* $\lambda = a^2\sigma^2/\gamma^2$ *approaching infinity. Let* $\{(\mu_\lambda, \pi_\lambda)\}_{\lambda>0}$ *be the corresponding family of posteriors and priors and let* g_λ *be the normalized density. Lemma 1 implies that*

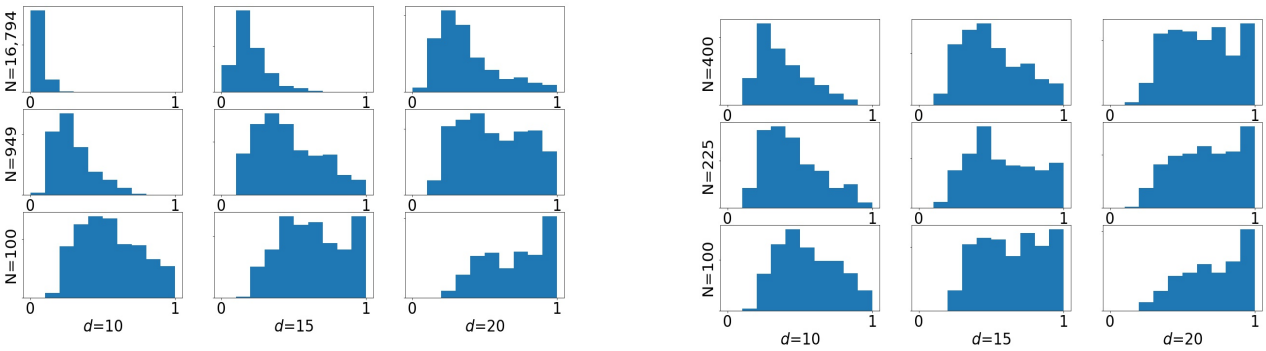
$$\frac{\pi_\lambda(\mathrm{g}_\lambda^3)}{\pi_\lambda(\mathrm{g}_\lambda^2)^2} = \frac{2\lambda+1}{\sqrt{(3\lambda+1)(\lambda+1)}} \exp\Big(\frac{\lambda}{(2\lambda+1)(3\lambda+1)} z^2\Big) \to \frac{2}{\sqrt{3}} < 2,$$

as $\lambda \to \infty$. *This implies that, for* λ *sufficiently large,*

$$\frac{\mathbb{V}[\mathrm{g}_\lambda(U_\lambda)]}{\rho_\lambda^2} = \frac{\pi_\lambda(\mathrm{g}_\lambda^3)}{\pi_\lambda(\mathrm{g}_\lambda^2)^2} - 1 < 1.$$

3.4.2. Large Dimensional Limit

Now we investigate the behavior of importance sampling in the limit of large dimension, in the inverse problem setting (11). We start with an example similar to the ones in Figures 1 and 2. Figure 3 shows that for $\lambda = 1.3$ fixed, weight collapse happens frequently when the sample size N grows polynomially as d^2 but not so often if N grows at rate $\mathcal{O}\left(\prod_{i=1}^d \left(\frac{\lambda+1}{\sqrt{2\lambda+1}} e^{\frac{\lambda z_i^2}{2\lambda+1}}\right)\right)$. These empirical results are in agreement with the growth rate of ρ_d in the large d limit.



(**a**) $N = \mathcal{O}\left(\prod_{i=1}^{d}\left(\frac{\lambda+1}{\sqrt{2\lambda+1}}e^{\frac{\lambda z_i^2}{2\lambda+1}}\right)\right)$. (**b**) $N = d^2$.

Figure 3. Dimensional scaling with $\lambda = 1.3$. The experimental setting is similar to those in Figures 1 and 2.

Proposition 7. *For any* $d \in \mathbb{N} \cup \{\infty\}$,

$$\rho_d = \prod_{i=1}^{d}\left(\frac{\lambda_i+1}{\sqrt{2\lambda_i+1}}e^{\frac{\lambda_i z_i^2}{2\lambda_i+1}}\right),$$
$$\mathbb{E}_{z_{1:d}}[\rho_d] = \prod_{i=1}^{d}(\lambda_i+1).$$

Proof. The formula for ρ_d is a direct consequence of Equation (16) and the product structure. Similarly, we have

$$\begin{aligned}\mathbb{E}_{z_i}\left[\frac{\lambda_i+1}{\sqrt{2\lambda_i+1}}e^{\frac{\lambda_i z_i^2}{2\lambda_i+1}}\right] &= \frac{\lambda_i+1}{\sqrt{2\lambda_i+1}}\int_{\mathbb{R}}\frac{1}{\sqrt{2\pi}}e^{-\frac{z_i^2}{2}+\frac{\lambda_i z_i^2}{2\lambda_i+1}}dz_i \\ &= \frac{\lambda_i+1}{\sqrt{2\lambda_i+1}}\int_{\mathbb{R}}\frac{1}{\sqrt{2\pi}}e^{-\frac{z_i^2}{2(2\lambda_i+1)}}dz_i \\ &= \lambda_i+1.\end{aligned}$$

□

Proposition 7 implies that, for $d \in \mathbb{N} \cup \{\infty\}$,

$$\sup_{|\varphi|_\infty \le 1}\mathbb{E}\left[\left(\mu^N_{1:d}(\varphi)-\mu_{1:d}(\varphi)\right)^2\right] \le \frac{4}{N}\prod_{i=1}^{d}\left(\frac{\lambda_i+1}{\sqrt{2\lambda_i+1}}e^{\frac{\lambda_i z_i^2}{2\lambda_i+1}}\right),$$
$$\mathbb{E}\left[\sup_{|\varphi|_\infty \le 1}\mathbb{E}\left[\left(\mu^N_{1:d}(\varphi)-\mu_{1:d}(\varphi)\right)^2\right]\right] \le \frac{4}{N}\prod_{i=1}^{d}(\lambda_i+1).$$

Note that the outer expected value in the latter equation averages over the data, while the inner one averages oversampling from the prior $\pi_{1:d}$. This suggests that

$$\log\mathbb{E}\left[\sup_{|\varphi|_\infty \le 1}\mathbb{E}\left[\left(\mu^N_{1:d}(\varphi)-\mu_{1:d}(\varphi)\right)^2\right]\right] \lesssim \sum_{i=1}^{d}\lambda_i - \log N.$$

The quantity $\tau := \sum_{i=1}^{d}\lambda_i$ had been used as an *intrinsic dimension* of the inverse problem (11). This simple heuristic together with Theorem 1 suggest that increasing N *exponentially* with τ is both necessary and sufficient to maintain accurate importance sampling along the high dimensional limit $d \to \infty$. In particular, if all coordinates of the problem play the same

role, this implies that N needs to grow exponentially with d, a manifestation of the curse of dimension of importance sampling [1,13,14].

3.4.3. Infinite Dimensional Singularity

Finally, we investigate the case $d = \infty$. Our goal in this subsection is to establish a connection between the effective dimension, the quantity ρ_∞ and absolute continuity. The main result, Corollary 1, had been proved in more generality in [1]. However, our proof and presentation here requires minimal technical background and is based on the explicit calculations obtained in the previous subsections and in the following lemma.

Lemma 2. *It holds that $\mu_{1:\infty}$ is absolutely continuous with respect to $\pi_{1:\infty}$ if and only if*

$$\mathcal{H}(\mu_{1:\infty}, \pi_{1:\infty}) = \prod_{i=1}^{\infty} \frac{\pi_i\left(g_i^{\frac{1}{2}}\right)}{\pi_i(g_i)^{\frac{1}{2}}} > 0, \tag{18}$$

where g_i is an un-normalized density between μ_i and π_i. Moreover, we have the following explicit characterizations of the Hellinger integral $\mathcal{H}(\mu_{1:\infty}, \pi_{1:\infty})$ and its average with respect to data realizations,

$$\mathcal{H}(\mu_{1:\infty}, \pi_{1:\infty}) = \prod_{i=1}^{\infty} \left(\sqrt{\frac{2\sqrt{\lambda_i + 1}}{\lambda_i + 2}} e^{-\frac{\lambda_i z_i^2}{4(\lambda_i+2)}} \right),$$

$$\mathbb{E}_{z_{1:\infty}}[\mathcal{H}(\mu_{1:\infty}, \pi_{1:\infty})] = \prod_{i=1}^{\infty} \frac{2(\lambda_i + 1)^{\frac{1}{4}}}{\sqrt{3\lambda_i + 4}}.$$

Proof. The formula for the Hellinger integral is a direct consequence of Equation (17) and the product structure. On the other hand,

$$\begin{aligned} \mathbb{E}_{z_i}\left[\sqrt{\frac{2\sqrt{\lambda_i + 1}}{\lambda_i + 2}} e^{-\frac{\lambda_i z_i^2}{4(\lambda_i+2)}} \right] &= \frac{\sqrt{2}(\lambda_i + 1)^{\frac{1}{4}}}{\sqrt{\lambda_i + 2}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{\lambda_i z_i^2}{4(\lambda_i+2)} - \frac{z_i^2}{2}} dz_i \\ &= \frac{2(\lambda_i + 1)^{\frac{1}{4}}}{\sqrt{3\lambda_i + 4}}. \end{aligned}$$

The proof of the equivalence between finite Hellinger integral and absolute continuity is given in Appendix B. □

Corollary 1. *The following statements are equivalent:*

i $\tau = \sum_{i=1}^{\infty} \lambda_i < \infty$;
ii $\rho_\infty < \infty$ *for almost every* y;
iii $\mu_{1:\infty} \ll \pi_{1:\infty}$ *for almost every* y.

Proof. Observe that $\lambda_i \to 0$ is a direct consequence of all three statements, so we will assume $\lambda_i \to 0$ from now on.
$(i) \Leftrightarrow (ii)$: By Proposition 7,

$$\log\left(\mathbb{E}_{z_{1:\infty}}[\rho_\infty]\right) = \sum_{i=1}^{\infty} \log(1 + \lambda_i) = \mathcal{O}(\sum_{i=1}^{\infty} \lambda_i),$$

since $\log(1+\lambda_i) \approx \lambda_i$ for large i.
$(i) \Leftrightarrow (iii)$: Similarly, we have

$$\begin{aligned}\log\Big(\mathbb{E}_{z_{1:\infty}}[\mathcal{H}(\mu_{1:\infty}, \pi_{1:\infty})]\Big) &= -\frac{1}{4}\sum_{i=1}^{\infty}\log\frac{(3\lambda_i+4)^2}{16(\lambda_i+1)} \\ &= -\frac{1}{4}\sum_{i=1}^{\infty}\log\left(1+\frac{9\lambda_i^2+8\lambda_i}{16\lambda_i+16}\right) \\ &= -\frac{1}{4}\mathcal{O}(\sum_{i=1}^{\infty}\lambda_i).\end{aligned}$$

The conclusion follows from Lemma 2. □

4. Importance Sampling for Data Assimilation

In this section, we study the use of importance sampling in a particle filtering setting. Following [13–15] we focus on one filtering step. Our goal is to provide a new and concrete comparison of two proposals, referred to as *standard* and *optimal* in the literature [1]. In Section 4.1 we introduce the setting and both proposals and show that the χ^2-divergence between target and standard proposal is larger than the χ^2-divergence between target and optimal proposal. Sections 4.3 and 4.4 identify small noise and large dimensional limiting regimes where the sample size for the standard proposal needs to grow unboundedly to maintain the same level of accuracy, but the required sample size for the optimal proposal remains bounded.

4.1. One-Step Filtering Setting

Let M and H be given matrices. We consider the one-step filtering problem of recovering v_0, v_1 from y, under the following setting

$$v_1 = Mv_0 + \xi, \qquad v_0 \sim \mathcal{N}(0,P), \quad \xi \sim \mathcal{N}(0,Q), \tag{19}$$
$$y = Hv_1 + \zeta, \qquad \zeta \sim \mathcal{N}(0,R). \tag{20}$$

Similar to the setting in Subsection 3.1, we assume that P, Q, R are symmetric positive definite and that M and H are full rank. From a Bayesian point of view, we would like to sample from the target distribution $\mathbb{P}_{v_0,v_1|y}$. To achieve this, we can either use $\pi_{\text{std}} = \mathbb{P}_{v_1|v_0}\,\mathbb{P}_{v_0}$ or $\pi_{\text{opt}} = \mathbb{P}_{v_1|v_0,y}\,\mathbb{P}_{v_0}$ as the proposal distribution.

The standard proposal π_{std} is the prior distribution of (v_0, v_1) determined by the prior $v_0 \sim \mathcal{N}(0,P)$ and the *signal dynamics* encoded in Equation (19). Then assimilating the observation y leads to an inverse problem [1,2] with design matrix, noise covariance and prior covariance given by

$$\begin{aligned}A_{\text{std}} &:= H, \\ \Gamma_{\text{std}} &:= R, \\ \Sigma_{\text{std}} &:= MPM' + Q.\end{aligned} \tag{21}$$

We denote $\pi_{\text{std}} = \mathcal{N}(0,\Sigma_{\text{std}})$ the prior distribution and by μ_{std} the corresponding posterior distribution.

The optimal proposal π_{opt} samples from v_0 and the conditional kernel $v_1|v_0, y$. Then assimilating y leads to the inverse problem [1,2]

$$y = HMv_0 + H\xi + \zeta,$$

where the design matrix, noise covariance and prior covariance are given by

$$\begin{aligned} A_{\text{opt}} &:= HM, \\ \Gamma_{\text{opt}} &:= HQH' + R, \\ \Sigma_{\text{opt}} &:= P. \end{aligned} \tag{22}$$

We denote $\pi_{\text{opt}} = \mathcal{N}(0, \Sigma_{\text{opt}})$ the prior distribution and μ_{std} the corresponding posterior distribution.

4.2. χ^2 -Divergence Comparison between Standard and Optimal Proposal

Here we show that

$$\rho_{\text{std}} := d_{\chi^2}(\mu_{\text{std}} \| \pi_{\text{std}}) + 1 > d_{\chi^2}(\mu_{\text{opt}} \| \pi_{\text{opt}}) + 1 =: \rho_{\text{opt}}.$$

The proof is a direct calculation using the explicit formula in Proposition 4. We introduce, as in Section 3, standard Kalman notation

$$\begin{aligned} K_{\text{std}} &:= \Sigma_{\text{std}} A'_{\text{std}} S_{\text{std}}^{-1}, & S_{\text{std}} &:= A_{\text{std}} \Sigma_{\text{std}} A'_{\text{std}} + \Gamma_{\text{std}}, \\ K_{\text{opt}} &:= \Sigma_{\text{opt}} A'_{\text{opt}} S_{\text{opt}}^{-1}, & S_{\text{opt}} &:= A_{\text{opt}} \Sigma_{\text{opt}} A'_{\text{opt}} + \Gamma_{\text{opt}}. \end{aligned}$$

It follows from the definitions in (21) and (22) that

$$\begin{aligned} S_{\text{std}} &= H(MPM' + Q)H + R \\ &= HMPM'H + HQH' + R \\ &= S_{\text{opt}}. \end{aligned}$$

Since $S_{\text{std}} = S_{\text{opt}}$ we drop the subscripts in what follows and denote both simply by S.

Theorem 2. *Consider the one-step filtering setting in Equations (19) and (20). If M and H are full rank and P, Q, R are symmetric positive definite, then, for almost every y,*

$$\rho_{\text{std}} > \rho_{\text{opt}}.$$

Proof. By Proposition 4 we have

$$\begin{aligned} \rho_{\text{std}} &= (|I - K_{\text{std}} A_{\text{std}}||I + K_{\text{std}} A_{\text{std}}|)^{-\frac{1}{2}} \exp\left(y' K'_{\text{std}} [(I + K_{\text{std}} A_{\text{std}})\Sigma_{\text{std}}]^{-1} K_{\text{std}}\, y\right), \\ \rho_{\text{opt}} &= (|I - K_{\text{opt}} A_{\text{opt}}||I + K_{\text{opt}} A_{\text{opt}}|)^{-\frac{1}{2}} \exp\left(y' K'_{\text{opt}} [(I + K_{\text{opt}} A_{\text{opt}})\Sigma_{\text{std}}]^{-1} K_{\text{opt}}\, y\right). \end{aligned}$$

Therefore, it suffices to prove the following two inequalities:

$$|I - K_{\text{std}} A_{\text{std}}||I + K_{\text{std}} A_{\text{std}}| < |I - K_{\text{opt}} A_{\text{opt}}||I + K_{\text{opt}} A_{\text{opt}}|, \tag{23}$$

$$K'_{\text{std}} [(I + K_{\text{std}} A_{\text{std}})\Sigma_{\text{std}}]^{-1} K_{\text{std}} \prec K'_{\text{opt}} [(I + K_{\text{opt}} A_{\text{opt}})\Sigma_{\text{std}}]^{-1} K_{\text{opt}}. \tag{24}$$

We start with inequality (24). Note that

$$\begin{aligned} (I + K_{\text{std}} A_{\text{std}})\Sigma_{\text{std}} &= \Sigma_{\text{std}} + \Sigma_{\text{std}} A'_{\text{std}} S^{-1} A_{\text{std}} \Sigma_{\text{std}}, \\ (I + K_{\text{opt}} A_{\text{opt}})\Sigma_{\text{opt}} &= \Sigma_{\text{opt}} + \Sigma_{\text{opt}} A'_{\text{opt}} S^{-1} A_{\text{opt}} \Sigma_{\text{opt}}. \end{aligned}$$

Using the definitions in (21) and (22) it follows that

$$\begin{aligned} K'_{\text{std}}\Sigma_{\text{std}}^{-1}(I + K_{\text{std}}A_{\text{std}})^{-1}K_{\text{std}} &= H\Big\{(MPM' + Q)^{-1} + H'SH\Big\}^{-1}H' \\ &\prec H\Big\{(MPM')^{-1} + H'SH\Big\}^{-1}H' \\ &= K'_{\text{opt}}\Sigma_{\text{opt}}^{-1}(I + K_{\text{opt}}A_{\text{opt}})^{-1}K_{\text{opt}}. \end{aligned}$$

For inequality (23), we notice that

$$\begin{aligned} K_{\text{std}}A_{\text{std}} &= (MPM' + Q)H'S^{-1}H = M\tilde{P}M'H'S^{-1}H \sim (H'S^{-1}H)^{\frac{1}{2}}M\tilde{P}M'(H'S^{-1}H)^{\frac{1}{2}}, \\ K_{\text{opt}}A_{\text{opt}} &= PM'H'S^{-1}HM \sim (H'S^{-1}H)^{\frac{1}{2}}MPM'(H'S^{-1}H)^{\frac{1}{2}}, \end{aligned}$$

where $\tilde{P} := P + M^{\dagger}QM'^{\dagger}$. Therefore

$$K_{\text{opt}}A_{\text{opt}} \prec K_{\text{std}}A_{\text{std}}$$

which, together with $K_{\text{std}}A_{\text{std}} \prec I$, implies that

$$|I - K_{\text{std}}A_{\text{std}}||I + K_{\text{std}}A_{\text{std}}| - |I - K_{\text{opt}}A_{\text{opt}}||I + K_{\text{opt}}A_{\text{opt}}| = |I - (K_{\text{std}}A_{\text{std}})^2| - |I - (K_{\text{opt}}A_{\text{opt}})^2| > 0,$$

as desired. □

Remark 3. *It is well known that if the signal dynamics are deterministic, i.e., if $Q = 0$ in (19), then the standard and optimal proposal agree, and therefore $\rho_{\text{opt}} = \rho_{\text{std}}$. Theorem 2 shows that $\rho_{\text{std}} > \rho_{\text{opt}}$ provided that Q is positive definite. Further works that have investigated theoretical and practical benefits of the optimal proposal over the standard proposal include [1,2,31,34]. In particular, [1] shows that use of the optimal proposal reduces the intrinsic dimension. Theorem 2 compares directly the χ^2-divergence, which is the key quantity that determines the performance of importance sampling.*

4.3. Standard and Optimal Proposal in Small Noise Regime

It is possible that along a certain limiting regime, ρ diverges for the standard proposal but not for the optimal proposal. This has been observed in previous work [1,31], and here we provide some concrete examples using the scaling results from Section 3. Precisely, consider the following one-step filtering setting

$$\begin{aligned} v_1 &= Mv_0 + \xi, \qquad v_0 \sim \mathcal{N}(0, P), \quad \xi \sim \mathcal{N}(0, Q), \\ y &= Hv_1 + \zeta, \qquad \zeta \sim \mathcal{N}(0, r^2R), \end{aligned}$$

where $r \to 0$. Let $\mu_{\text{opt}}^{(r)}, \mu_{\text{std}}^{(r)}$ be the optimal/standard targets and $\pi_{\text{opt}}^{(r)}, \pi_{\text{std}}^{(r)}$ be the optimal/standard proposals. We assume that $M \in \mathbb{R}^{d\times d}$ and $H \in \mathbb{R}^{k\times d}$ are full rank.

Proposition 8. *If $r \to 0$, then we have*

$$\begin{aligned} \rho_{\text{opt}}^{(r)} &< \infty, \\ \rho_{\text{std}}^{(r)} &\sim \mathcal{O}(r^{-k}). \end{aligned}$$

Proof. Consider the two inverse problems that correspond to $\mu^{(r)}_{\text{opt}}, \pi^{(r)}_{\text{opt}}$ and $\mu^{(r)}_{\text{std}}, \pi^{(r)}_{\text{std}}$. Note that the two problems have identical prior and design matrix. Let $\Gamma^{(r)}_{\text{opt}}$ and $\Gamma^{(r)}_{\text{std}}$ denote the noise in those two inverse problems. When r goes to 0, we observe that

$$\Gamma^{(r)}_{\text{opt}} = r^2 R + HQH' \to HQH',$$
$$\Gamma^{(r)}_{\text{std}} = r^2 R \to 0.$$

Therefore, the limit of $\rho^{(r)}_{\text{opt}}$ converges to a finite value, but Lemma 5 implies that $\rho^{(r)}_{\text{std}}$ diverges at rate $\mathcal{O}(r^{-k})$. □

4.4. Standard and Optimal Proposal in High Dimension

The previous subsection shows that the standard and optimal proposals can have dramatically different behavior in the small noise regime $r \to 0$. Here we show that both proposals can also lead to dramatically different behavior in high dimensional limits. Precisely, as a consequence of Corollary 1 we can easily identify the exact regimes where both proposals converge or diverge in limit. The notation is analogous to that in Section 4.4, and so we omit the details.

Proposition 9. *Consider the sequence of particle filters defined as above. We have the following convergence criteria:*

1. $\mu^{(1:\infty)}_{\text{opt}} \ll \pi^{(1:\infty)}_{\text{opt}}$ *and* $\rho_{\text{opt}} < \infty$ *if and only if* $\sum_{i=1}^{\infty} \frac{h_i^2 m_i^2 p_i^2}{h_i^2 q_i^2 + r_i^2} < \infty$,
2. $\mu^{(1:\infty)}_{\text{std}} \ll \pi^{(1:\infty)}_{\text{std}}$ *and* $\rho_{\text{std}} < \infty$ *if and only if* $\sum_{i=1}^{\infty} \frac{h_i^2 m_i^2 p_i^2}{r_i^2} < \infty$ *and* $\sum_{i=1}^{\infty} \frac{h_i^2 q_i^2}{r_i^2} < \infty$.

Proof. By direct computation, we have

$$\lambda^{(i)}_{\text{std}} = \frac{h_i^2 m_i^2 p_i^2 + h_i^2 q_i^2}{r_i^2} = \frac{h_i^2 m_i^2 p_i^2}{r_i^2} + \frac{h_i^2 q_i^2}{r_i^2},$$
$$\lambda^{(i)}_{\text{opt}} = \frac{h_i^2 m_i^2 p_i^2}{h_i^2 q_i^2 + r_i^2}.$$

Theorem 1 gives the desired result. □

Example 2. *As a simple example where absolute continuity holds for the optimal proposal but not for the standard one, let* $h_i = m_i = p_i = r_i = 1$. *Then* $\rho_{\text{std}} = \infty$, *but* $\rho_{\text{opt}} < \infty$ *provided that* $\sum_{i=1}^{\infty} \frac{1}{q_i^2+1} < \infty$.

Author Contributions: Funding acquisition, D.S.-A.; Investigation, Z.W.; Supervision, D.S.-A.; Writing—original draft, Z.W.; Writing—review & editing, D.S.-A. and Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: The work of DSA was supported by NSF and NGA through the grant DMS-2027056. DSA also acknowledges partial support from the NSF grant DMS-1912818/1912802.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. χ^2-Divergence between Gaussians

We recall that the distribution P_θ parameterized by θ belongs to the exponential family $\mathcal{E}_F(\Theta)$ over a natural parameter space Θ, if $\theta \in \Theta$ and P_θ has density of the form

$$f(u;\theta) = e^{\langle t(u),\theta\rangle - F(\theta) + k(u)},$$

where the natural parameter space is given by

$$\Theta = \left\{ \theta : \int e^{\langle t(u), \theta \rangle + k(u)} du < \infty \right\}.$$

The following result can be found in [35].

Lemma A1. *Suppose* $\theta_{1,2} \in \Theta$ *are parameters for probability densities* $f(u; \theta_{1,2}) = e^{\langle t(u), \theta_{1,2} \rangle - F(\theta_{1,2}) + k(u)}$ *with* $2\theta_1 - \theta_2 \in \Theta$. *Then,*

$$d_{\chi^2}\big(f(\cdot\,;\theta_1) \| f(\cdot\,;\theta_2)\big) = e^{F(2\theta_1 - \theta_2) - 2F(\theta_1) + F(\theta_2)} - 1.$$

Proof. By direct computation,

$$\begin{aligned} d_{\chi^2}\big(f(\cdot\,;\theta_1) \| f(\cdot\,;\theta_2)\big) + 1 &= \int f(u;\theta_1)^2 f(u;\theta_2)^{-1} du \\ &= \int e^{\langle t(u), 2\theta_1 - \theta_2 \rangle - (2F(\theta_1) - F(\theta_2)) + k(u)}\, du \\ &= e^{F(2\theta_1 - \theta_2) - 2F(\theta_1) + F(\theta_2)} \int f(u; 2\theta_1 - \theta_2)\, du \\ &= e^{F(2\theta_1 - \theta_2) - 2F(\theta_1) + F(\theta_2)}. \end{aligned}$$

Note that $\int f(u; 2\theta_1 - \theta_2)\, du = 1$ since $2\theta_1 - \theta_2 \in \Theta$ by assumption. □

Using Lemma A1 we can compute the χ^2-divergence between Gaussians. To do so, we note that $d-$dimensional Gaussians $\mathcal{N}(\mu, \Sigma)$ belong to the exponential family over the parameter space $\mathbb{R}^d \bigoplus \mathbb{R}^{d \times d}$ by letting $\theta = [\Sigma^{-1}\mu; -\frac{1}{2}\Sigma^{-1}]$ and $F(\theta) = \frac{1}{2}\mu'\Sigma^{-1}\mu + \frac{1}{2}\log|\Sigma|$. In the context of Gaussians, an exponential parameter $\theta = [\Sigma^{-1}\mu; -\frac{1}{2}\Sigma^{-1}]$ belongs to the natural parameter space Θ if and only if Σ is symmetric and positive definite. Indeed, the integral $\int \exp(-\frac{1}{2}(u-\mu)'\Sigma^{-1}(u-\mu))du$ is finite if and only if $\Sigma \succ 0$.

Proof of Proposition 3. Let θ_μ, θ_π be the exponential parameters of μ, π. Then $2\theta_\mu - \theta_\pi$ corresponds to a Gaussian with mean $(2C^{-1} - \Sigma^{-1})^{-1}(2C^{-1}m)$ and covariance $(2C^{-1} - \Sigma^{-1})^{-1}$. We have

$$\begin{aligned} F(2\theta_\mu - \theta_\pi) - 2F(\theta_\mu) + F(\theta_\pi) &= \frac{1}{2}\log|(2C^{-1} - \Sigma^{-1})^{-1}| - \log|C| + \frac{1}{2}\log|\Sigma| + \\ &\quad \frac{1}{2}(2C^{-1}m)'(2C^{-1} - \Sigma^{-1})^{-1}(2C^{-1}m) - m'C^{-1}m \\ &= \log\sqrt{\frac{|\Sigma|}{|2C^{-1} - \Sigma^{-1}||C|^2}} + m'(C^{-1}(2C^{-1} - \Sigma^{-1})^{-1}2C^{-1})m \\ &\quad - m'(C^{-1}(2C^{-1} - \Sigma^{-1})^{-1}(2C^{-1} - \Sigma^{-1}))m \\ &= \log\frac{|\Sigma|}{\sqrt{|2\Sigma - C||C|}} + m'(C^{-1}(2C^{-1} - \Sigma^{-1})^{-1}\Sigma^{-1})m \\ &= \log\frac{|\Sigma|}{\sqrt{|2\Sigma - C||C|}} + m'(2\Sigma - C)^{-1}m. \end{aligned}$$

Applying Lemma A1 gives

$$\begin{aligned} d_{\chi^2}(\mu \| \pi) &= \exp\Big(F(2\theta_\mu - \theta_\pi) - 2F(\theta_\mu) + F(\theta_\pi)\Big) - 1 \\ &= \frac{|\Sigma|}{\sqrt{|2\Sigma - C||C|}} \exp\Big(m'(2\Sigma - C)^{-1}m\Big) - 1, \end{aligned}$$

if $2\theta_\mu - \theta_\pi \in \Theta$. In other words, the corresponding covariance matrix $(2C^{-1} - \Sigma^{-1})^{-1}$ is positive definite. □

Remark A1. *By translation invariance of Lebesgue measure, we can obtain the more general formula for χ^2-divergence between two Gaussians with nonzero mean by replacing m with the difference between the two mean vectors:*

$$d_{\chi^2}\Big(\mathcal{N}(m_1, C)\|\,\mathcal{N}(m_2, \Sigma)\Big) = \frac{|\Sigma|}{\sqrt{|2\Sigma - C||C|}} e^{(m_1-m_2)'(2\Sigma-C)^{-1}(m_1-m_2)} - 1.$$

Appendix B. Proof of Lemma 2

Proof. Dividing g by its normalizing constant, we may assume without loss of generality that g is exactly the Radon–Nikodym derivative $\frac{d\mu}{d\pi}$ and $\mathcal{H}(\mu, \pi) = \pi_i(\sqrt{g})$.
If $\mu_{1:\infty} \ll \pi_{1:\infty}$, then the Radon-Nikodym derivative $g_{1:\infty}$ cannot be $\pi_{1:\infty}$ a.e. zero since $\pi_{1:\infty}$ and $\mu_{1:\infty}$ are probability measures. As a consequence, $\prod_{i=1}^{\infty} \pi_i\big(\sqrt{g_i}\big) = \pi_{1:\infty}\big(\sqrt{g_{1:\infty}}\big) > 0$ by the product structure of $\mu_{1:\infty}$ and $\pi_{1:\infty}$.
Now we assume $\prod_{i=1}^{\infty} \pi_i\big(\sqrt{g_i}\big) > 0$. It suffices to show that $g_{1:\infty}$ is well-defined, i.e., convergence of $\prod_{i=1}^{L} g_i$ in L^1_π as $L \to \infty$. It suffices to prove that the sequence is Cauchy, in other words

$$\lim_{L,\ell\to\infty} \pi_{1:\infty}(|g_{1:L+\ell} - g_{1:L}|) = 0.$$

We observe that

$$\begin{aligned}
\|g_{1:L+\ell} - g_{1:L}\|_1 &\le \|\sqrt{g_{1:L+\ell}} - \sqrt{g_{1:L}}\|_2 \|\sqrt{g_{1:L+\ell}} + \sqrt{g_{1:L}}\|_2 \\
&\le \|\sqrt{g_{1:L+\ell}} - \sqrt{g_{1:L}}\|_2 (\|\sqrt{g_{1:L+\ell}}\|_2 + \|\sqrt{g_{1:L}}\|_2) \\
&= 2\|\sqrt{g_{1:L+\ell}} - \sqrt{g_{1:L}}\|_2.
\end{aligned}$$

Expanding the square of the right-hand side gives

$$\begin{aligned}
\pi_{1:\infty}\Big(|\sqrt{g_{1:L+\ell}} - \sqrt{g_{1:L}}|^2\Big) &= \pi_{1:\infty}(g_{1:L+\ell} + g_{1:L} - 2\sqrt{g_{1:L+\ell} g_{1:L}}) \\
&= 2 - 2\pi_{1:L}(g_{1:L})\pi_{L+1:\infty}\left(\sqrt{\frac{g_{1:L+\ell}}{g_{1:L}}}\right) \\
&= 2\left(1 - \frac{\pi_{1:L+\ell}\big(\sqrt{g_{1:L+\ell}}\big)}{\pi_{1:L}\big(\sqrt{g_{1:L}}\big)}\right).
\end{aligned}$$

Therefore, it is enough to show

$$\lim_{L,\ell\to\infty} \frac{\pi_{1:L+\ell}\big(\sqrt{g_{1:L+\ell}}\big)}{\pi_{1:L}\big(\sqrt{g_{1:L}}\big)} = 1.$$

By Jensen's inequality, for any two probability measures $\mu \ll \pi$ with density g, we have

$$\pi(\sqrt{g}) \le \sqrt{\pi(g)} = 1. \tag{A1}$$

Combining with our assumption, we deduce that

$$0 < \prod_{i=1}^{\infty} \pi_i(\sqrt{g_i}) = \pi_{1:\infty}(\sqrt{g_{1:\infty}}) \le 1,$$

which is equivalent to

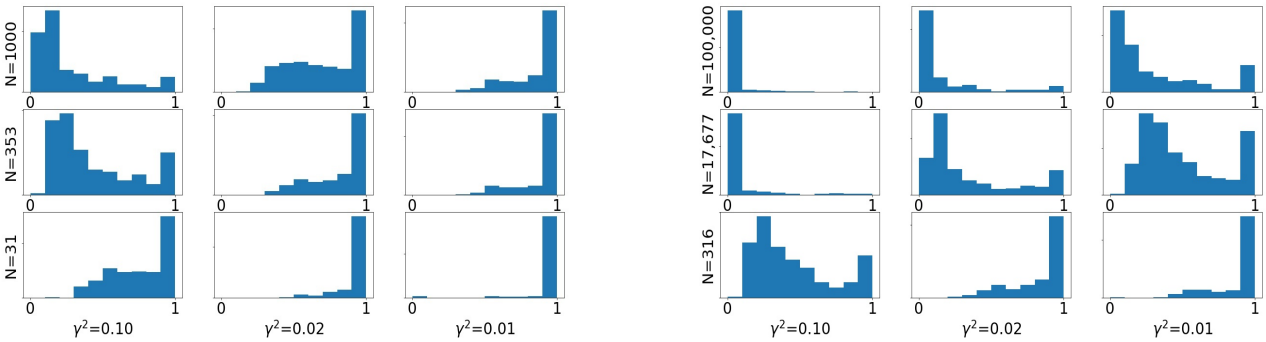
$$-\infty < \sum_{i=1}^{\infty} \log(\pi_i(\sqrt{g_i})) \le 0.$$

This series is monotonely decreasing by (A1) and bounded below, so it converges and satisfies that

$$\lim_{L,\ell\to\infty} \frac{\pi_{1:L+\ell}\left(\sqrt{g_{1:L+\ell}}\right)}{\pi_{1:L}\left(\sqrt{g_{1:L}}\right)} = \lim_{L,\ell\to\infty} e^{\sum_{i=L}^{L+\ell} \log\left(\pi_i\left(\sqrt{g_i}\right)\right)} = 1.$$

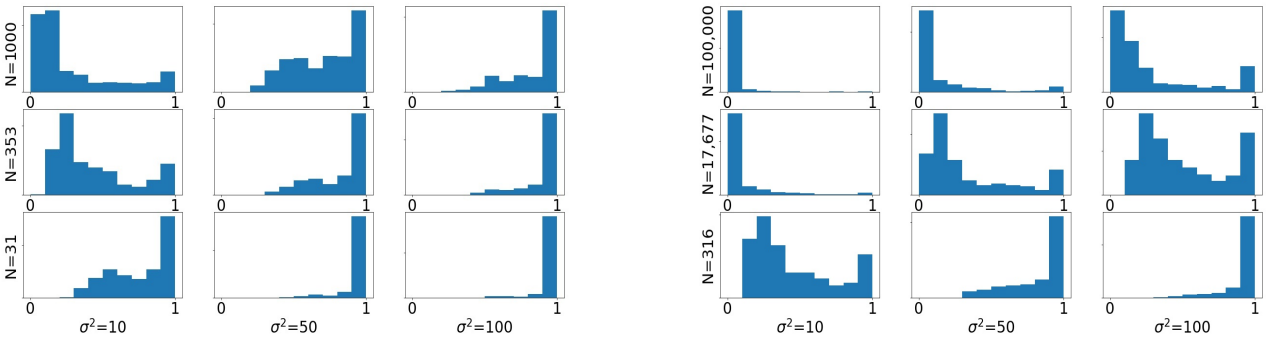
□

Appendix C. Additional Figures



(**a**) $N = \gamma^{-3}$. (**b**) $N = \gamma^{-5}$.

Figure A1. Noise scaling with $d = k = 4$.



(**a**) $N = \sigma^{-3}$. (**b**) $N = \sigma^{-5}$.

Figure A2. Prior scaling with $d = k = 4$.

References

1. Agapiou, S.; Papaspiliopoulos, O.; Sanz-Alonso, D.; Stuart, A.M. Importance sampling: Intrinsic dimension and computational cost. *Stat. Sci.* **2017**, *32*, 405–431. [CrossRef]
2. Sanz-Alonso, D.; Stuart, A.M.; Taeb, A. Inverse Problems and Data assimilation. *arXiv* **2018**, arXiv:1810.06191.
3. Barber, D. *Bayesian Reasoning and Machine Learning*; Cambridge University Press: Cambridge, UK, 2012.
4. Garcia Trillos, N.; Kaplan, Z.; Samakhoana, T.; Sanz-Alonso, D. On the consistency of graph-based Bayesian semi-supervised learning and the scalability of sampling algorithms. *J. Mach. Learn. Res.* **2020**, *21*, 1–47.
5. Garcia Trillos, N.; Sanz-Alonso, D. The Bayesian update: Variational formulations and gradient flows. *Bayesian Anal.* **2018**, *15*, 29–56. [CrossRef]
6. Dick, J.; Kuo, F.Y.; Sloan, I.H. High-dimensional integration: The quasi-Monte Carlo way. *Acta Numer.* **2013**, *22*, 133. [CrossRef]
7. Chatterjee, S.; Diaconis, P. The sample size required in importance sampling. *arXiv* **2015**, arXiv:1511.01437.
8. Sanz-Alonso, D. Importance sampling and necessary sample size: An information theory approach. *SIAM/ASA J. Uncertain. Quantif.* **2018**, *6*, 867–879. [CrossRef]
9. Rubino, G.; Tuffin, B. *Rare Event Simulation Using Monte Carlo Methods*; John Wiley & Sons: Hoboken, NJ, USA, 2009.
10. Kahn, H.; Marshall, A.W. Methods of reducing sample size in Monte Carlo computations. *J. Oper. Res. Soc. Am.* **1953**, *1*, 263–278. [CrossRef]
11. Kahn, H. *Use of different Monte Carlo Sampling Techniques*; Rand Corporation: Santa Monica, CA, USA, 1955.
12. Bugallo, M.F.; Elvira, V.; Martino, L.; Luengo, D.; Miguez, J.; Djuric, P.M. Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Process. Mag.* **2017**, *34*, 60–79. [CrossRef]

13. Bengtsson, T.; Bickel, P.; Li, B. Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In *Probability and Statistics: Essays in honor of David A. Freedman*; Institute of Mathematical Statistics: Beachwood, OH, USA, 2008; pp. 316–334.
14. Bickel, P.; Li, B.; Bengtsson, T. Sharp failure rates for the bootstrap particle filter in high dimensions. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*; Institute of Mathematical Statistics: Beachwood, OH, USA, 2008; pp. 318–329.
15. Snyder, C.; Bengtsson, T.; Bickel, P.; Anderson, J. Obstacles to high-dimensional particle filtering. *Mon. Weather Rev.* **2008**, *136*, 4629–4640. [CrossRef]
16. Rebeschini, P.; Van Handel, R. Can local particle filters beat the curse of dimensionality? *Ann. Appl. Probab.* **2015**, *25*, 2809–2866. [CrossRef]
17. Chorin, A.J.; Morzfeld, M. Conditions for successful data assimilation. *J. Geophys. Res. Atmos.* **2013**, *118*, 11–522. [CrossRef]
18. Houtekamer, P.L.; Mitchell, H.L. Data assimilation using an ensemble Kalman filter technique. *Mon. Weather Rev.* **1998**, *126*, 796–811. [CrossRef]
19. Hamill, T.M.; Whitaker, J.S.; Anderson, J.L.; Snyder, C. Comments on "Sigma-point Kalman filter data assimilation methods for strongly nonlinear systems". *J. Atmos. Sci.* **2009**, *66*, 3498–3500. [CrossRef]
20. Morzfeld, M.; Hodyss, D.; Snyder, C. What the collapse of the ensemble Kalman filter tells us about particle filters. *Tellus A Dyn. Meteorol. Oceanogr.* **2017**, *69*, 1283809. [CrossRef]
21. Farchi, A.; Bocquet, M. Comparison of local particle filters and new implementations. *Nonlinear Process. Geophys.* **2018**, *25*. [CrossRef]
22. Morzfeld, M.; Tong, X.T.; Marzouk, Y.M. Localization for MCMC: Sampling high-dimensional posterior distributions with local structure. *J. Comput. Phys.* **2019**, *380*, 1–28. [CrossRef]
23. Tong, X.T.; Morzfeld, M.; Marzouk, Y.M. MALA-within-Gibbs samplers for high-dimensional distributions with sparse conditional structure. *SIAM J. Sci. Comput.* **2020**, *42*, A1765–A1788. [CrossRef]
24. Liu, J.S. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Stat. Comput.* **1996**, *6*, 113–119. [CrossRef]
25. Pitt, M.K.; Shephard, N. Filtering via simulation: Auxiliary particle filters. *J. Am. Stat. Assoc.* **1999**, *94*, 590–599. [CrossRef]
26. Kong, A. A note on importance sampling using standardized weights. *Univ. Chicago, Dept. Stat. Tech. Rep* **1992**, *348*.
27. Kong, A.; Liu, J.S.; Wong, W.H. Sequential imputations and Bayesian missing data problems. *J. Am. Stat. Assoc.* **1994**, *89*, 278–288. [CrossRef]
28. Ryu, E.K.; Boyd, S.P. Adaptive importance sampling via stochastic convex programming. *arXiv* **2014**, arXiv:1412.4845.
29. Akyildiz, Ö.D.; Míguez, J. Convergence rates for optimised adaptive importance samplers. *arXiv* **2019**, arXiv:1903.12044.
30. Bogachev, V.I. *Gaussian Measures*; Number 62; American Mathematical Society: Providence, RI, USA, 1998.
31. Snyder, C.; Bengtsson, T.; Morzfeld, M. Performance bounds for particle filters using the optimal proposal. *Mon. Weather Rev.* **2015**, *143*, 4750–4761. [CrossRef]
32. Doucet, A.; De Freitas, N.; Gordon, N. An Introduction to Sequential Monte Carlo Methods. In *Sequential Monte Carlo Methods in Practice*; Springer: Berlin/Heisenberg, Germany, 2001; pp. 3–14.
33. Del Moral, P. *Feynman-Kac Formulae*; Springer: Berlin/Heisenberg, Germany, 2004.
34. Doucet, A.; Godsill, S.; Andrieu, C. On sequential Monte Carlo sampling methods for Bayesian filtering. *Stat. Comput.* **2000**, *10*, 197–208. [CrossRef]
35. Nielsen, F.; Nock, R. On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Process. Lett.* **2013**, *21*, 10–13. [CrossRef]