Importance Sampling: Computational Complexity and Intrinsic Dimension

S. Agapiou*, O. Papaspiliopoulos† D. Sanz-Alonso‡ and A. M. Stuart‡

agapiou.sergios@ucy.ac.cy, omiros.papaspiliopoulos@upf.edu, d.sanz-alonso@warwick.ac.uk, a.m.stuart@warwick.ac.uk

Abstract: The basic idea of importance sampling is to use independent samples from one measure in order to approximate expectations with respect to another measure. Understanding how many samples are needed is key to understanding the computational complexity of the method, and hence to understanding when it will be effective and when it will not. It is intuitive that the size of the difference between the measure which is sampled, and the measure against which expectations are to be computed, is key to the computational complexity. An implicit challenge in many of the published works in this area is to find useful quantities which measure this difference in terms of parameters which are pertinent for the practitioner. The subject has attracted substantial interest recently from within a variety of communities. The objective of this paper is to overview and unify the resulting literature in the area by creating an overarching framework. The general setting is studied in some detail, followed by deeper development in the context of Bayesian inverse problems and filtering.

1. Introduction

1.1. Our Purpose

Our purpose in this paper is to overview various ways of measuring the computational complexity of importance sampling, to link them to one another through transparent mathematical reasoning, and to create cohesion in the vast published literature on this subject. In addressing these issues we will study importance sampling in a general abstract setting, and then in the particular cases of Bayesian inversion and filtering. These two application settings are particularly important as there are many pressing scientific, technological and societal problems which can be formulated via inversion or filtering. An example of such an inverse problem is the determination of subsurface properties of the Earth from surface measurements; an example of a filtering problem is assimilation of atmospheric measurements into numerical weather forecasts.

The general abstract setting in which we work is as follows. We let μ and π be two probability measures on a measurable space $(\mathcal{X}, \mathcal{F})$ related via the

* Department of Mathematics and Statistics, University of Cyprus, 1 University Avenue, 2109 Nicosia, Cyprus.

†ICREA & Department of Economics, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain.

‡ Mathematics Institute, University of Warwick, Coventry CV4 7AL, United Kingdom.

expression

$$\frac{d\mu}{d\pi}(u) := g(u) \Big/ \int_{\mathcal{X}} g(u)\pi(du). \tag{1.1}$$

Here, g is the unnormalised *density (or Radon-Nikodym derivative)* of μ with respect to π. Note that the very existence of the density implies that the target is *absolutely continuous* with respect to the proposal; absolute continuity will play an important role in our subsequent developments of this subject.

Importance sampling is a method for using independent samples from the *proposal* π to approximately compute expectations with respect to the *target* μ. The computational complexity is measured by the number of samples required to control the worst error made when approximating expectations within a class of test functions. Intuitively it is clear that the computational complexity of importance sampling is related to how far the target measure is from the proposal measure. With this in mind, a key quantity in what follows is the second moment, under the proposal, of $d\mu/d\pi$, which throughout the paper is denoted by ρ. As we observe below, it is simply obtained as $\rho = \pi(g^2)/\pi(g)^2$.

The first application of this setting that we study is the linear inverse problem to determine $u \in \mathcal{X}$ from y where

$$y = Ku + \eta, \quad \eta \sim N(0, \Gamma). \tag{1.2}$$

We adopt a Bayesian approach in which we place a prior $u \sim \mathbb{P}_u = N(0, \Sigma)$, assume that η is independent of u, and seek the posterior $u|y \sim \mathbb{P}_{u|y}$. We study importance sampling with $\mathbb{P}_{u|y}$ being the target μ and the prior $\mathbb{P}_u$ being the proposal π.

The second application is the linear filtering problem of sequentially updating the distribution of $v_j \in \mathcal{X}$ given $\{y_i\}_{i=1}^j$ where

$$\begin{aligned} v_{j+1} &= Mv_j + \xi_j, \quad \xi_j \sim N(0, Q), \quad j \geq 0, \\ y_{j+1} &= Hv_{j+1} + \zeta_{j+1}, \quad \zeta_{j+1} \sim N(0, R), \quad j \geq 0. \end{aligned} \tag{1.3}$$

We assume that the problem has a Markov structure. We study the approximation of one step of the filtering update by means of particles, building on the study of importance sampling for the linear inverse problem. To this end it is expedient to work on the product space $\mathcal{X} \times \mathcal{X}$, and consider importance sampling for $(v_j, v_{j+1}) \in \mathcal{X} \times \mathcal{X}$. It then transpires that, for two different proposals, which are commonly termed the *standard proposal* and the *optimal proposal*, the complexity of one step of particle filtering may be understood by the study of a linear inverse problem on $\mathcal{X}$; we show this for both proposals, and then use the link to an inverse problem to derive results about the complexity of particle filters based on these two proposals.

For the abstract importance sampling problem we will relate ρ to a number of other natural quantities. These include the *effective sample size* ess, used heuristically in many application domains, and a variety of *distance metrics* between π and μ. Since the existence of a density between target and proposal is central in this discussion, we will also discuss what happens as this absolute continuity property breaks down. We study this first in *high dimensional*

problems, and second in *singular parameter limits* (by which we mean limits in which important parameters defining the problem tend to zero). The motivation for studying high dimensional problems can be appreciated by considering the two examples mentioned at the start of the introduction: inverse problems from the Earth's subsurface, and filtering for numerical weather prediction. In both cases the unknown which we are trying to determine from data is best thought of as a spatially varying field for subsurface properties such as permeability, or atmospheric properties, such as temperature. In practice the field will be discretized and represented as a high dimensional vector, for computational purposes, but for these types of application the state dimension can be of order 10^9. Furthermore as computer power advances there is pressure to resolve more physics, and hence for the state dimension to increase. Thus, it is important to understand infinite dimensional problems, and sequences of approximating finite dimensional problems which approach the infinite dimensional limit. A motivation for studying singular parameter limits arises, for example, from problems in which the noise is small and the relevant log-likelihoods scale inversely with the noise variance. Breakdown of absolute continuity will be related to limits in which the target and proposal become increasingly close to being *mutually singular*.

We will highlight a variety of notions of *intrinsic dimension* that have been introduced in the inverse problem literature; these may differ substantially from the dimensions of the spaces where the unknown u and the data y live. We then go on to show how these intrinsic dimensions relate to the parameter ρ, previously demonstrated to be central to computational complexity. We do so in various limits arising from large dimension of u and y, and/or small observational noise. We also link these concepts to breakdown of absolute continuity. Finally we apply our understanding of linear inverse problems to particle filters, translating the results from one to the other via the correspondence between the two problems, for both standard and optimal proposals, as described above.

It is often claimed that importance sampling suffers from the curse of dimensionality. Whilst there is some empirical truth in this fact, there is a great deal of confusion in the literature about what exactly makes importance sampling hard. In fact such a statement about the role of dimension is vacuous unless "dimension" is defined precisely. Throughout this paper we use the following convention:

- State space dimension is the dimension of the measurable space where the measures μ and π are defined. We will be mostly interested in the case where the measurable space $\mathcal{X}$ is a separable Hilbert space, in which case the state space dimension is the cardinality of an orthonormal basis of the space. In the context of inverse problems and filtering, the state space dimension is the dimension of the unknown.
- Data space dimension is the dimension of the space where the data lives.
- Nominal dimension is the minimum of the state space dimension and the data state dimension.
- Intrinsic dimension: we will use two notions of intrinsic dimension for in-

verse problems, denoted by efd and τ. These combine state/data dimension and small noise parameters. They can be interpreted as a measure of how informative the data is relative to the prior.

Our presentation shows how the intrinsic dimensions are natural when studying computational complexity of importance sampling. Furthermore we relate these quantities to the second moment of the Radon-Nikodym derivative between proposal and target, ρ, which will also be shown to arise naturally in the same context. In studying these quantities, and their inter-relations, we aim to achieve the purpose set out at the start of this subsection. Furthermore, a bibliography subsection, within each section, will link our overarching mathematical framework to the published literature in this area.

1.2. Organization of the Paper and Notation

Section 2 describes importance sampling in abstract form. In sections 3 and 4 the linear Gaussian inverse problem and the linear Gaussian filtering problem are studied. Our aim is to provide a digestible narrative and hence all proofs are left to an Appendix in section 6. Furthermore, as we study the inverse and filtering problems in both finite dimensional Euclidean space and infinite dimensional Hilbert space, there are some technical matters related to Gaussian measures in infinite dimensional spaces that we also detail in the Appendix, subsection 6.1, in order not to distract from the narrative flow.

Given a probability measure ν on a measurable space $(\mathcal{X}, \mathcal{F})$ expectations of a function $\phi : \mathcal{X} \to \mathbb{R}$ with respect to ν will be written as both $\nu(\phi)$ and $\mathbb{E}_\nu[\phi]$. When it is clear which measure is being used we may drop the suffix ν and write simply $\mathbb{E}[\phi]$. Similarly, the variance will be written as $\mathrm{Var}_\nu(\phi)$ and again we may drop the suffix when no confusion arises from doing so.

We will be interested in sequences of measures indexed by time, by the state space dimension or by a tempering scheme. These are denoted with a subscript, e.g. ν_t, ν_d or ν_i. Anything to do with samples from a measure is denoted with a superscript: N for the number of samples, and n for the indices of the samples. The i-th coordinate of a vector u is denoted by $u(i)$. Thus, $u_t^n(i)$ denotes i-th coordinate of the n-th sample from the measure of interest at time t. Finally, the law of a random variable v will be denoted by $\mathbb{P}_v$.

1.3. Literature Review

Some early developments of importance sampling as a method to reduce the variance in Monte Carlo estimation date back to the early 1950's [45], [44]. In particular the paper [45] demonstrates how to optimally choose the proposal density for given test function ϕ and target density. A modern view of importance sampling in the general framework (1.1) is given in [21]. A comprehensive description of Bayesian inverse problems in finite state/data space dimensions can be found in [46], and its formulation in infinite dimensional spaces in

[28, 58, 59, 60, 86]. Text books overviewing the subject of filtering and particle filters include [29, 8], and the article [25] provides a readable introduction to the area. For an up-to-date and in-depth survey of nonlinear filtering see [26]. The linear Gaussian inverse problem and the linear Gaussian filtering problem have been extensively studied because they arise naturally in many applications, lead to considerable algorithmic tractability, and provide theoretical insight. For references concerning linear Gaussian inverse problems see [35, 67, 62, 51]. The linear Gaussian filter –the Kalman filter– was introduced in [48]; see [57] for further analysis. The inverse problem of determining subsurface properties of the Earth from surface measurements is discussed in [73], while the filtering problem of assimilating atmospheric measurements for numerical weather prediction is discussed in [49].

The key role of ρ, the second moment of the Radon-Nikodym derivative between the target and the proposal, has long been acknowledged [65], [74]. The value of ρ is indeed known to be asymptotically linked to the effective sample size [54], [55], [65]. We will provide a further nonasymptotic justification of the relevance of ρ through its appearance in error bounds on the error in importance sampling; in this context it is of relevance to highlight the paper [24] which proved non-asymptotic bounds on the error in the importance-sampling based particle filter algorithm. In this paper we will also bound the importance sampling error in terms of different notions of distance between the target and the proposal measures; a useful overview of the subject of distances between probability measures is [39].

We formulate problems in both finite dimensional and infinite dimensional state spaces. We refer to [47] for a modern presentation of probability appropriate for understanding the material in this article. Some of our results are built on the rich area of Gaussian measures in Hilbert space; we include all the required background on this material in the Appendix subsection 6.1, and references are included there. However we emphasize that the presentation in the main body of the text is designed to keep technical material to a minimum and to be accessible to readers who are not versed in the theory of probability in infinite dimensional spaces. Absolute continuity of the target with respect to the proposal – or the existence of a density of the target with respect to the proposal – is central to our developments. This concept also plays a pivotal role in the understanding of Markov chain Monte Carlo (MCMC) methods in high and infinite dimensional spaces [87]. A key idea in MCMC is that breakdown of absolute continuity on sequences of problems of increasing state space dimension is responsible for poor algorithmic performance with respect to increasing dimension; this should be avoided if possible, such as for problems with a well-defined infinite dimensional limit [23]. Similar ideas will come in to play in this paper.

As well as the breakdown of absolute continuity through increase in dimension, small noise limits can also lead to sequences of proposal/target measures which are increasingly close to mutually singular and for which absolute continuity breaks down. Small noise regimes are of theoretical and computational interest for both inverse problems and filtering. For instance, in inverse problems

there is a growing interest in the study of the concentration rate of the posterior in the small observational noise limit, see [52], [4], [53], [6], [75], [91], [51]. In filtering and multiscale diffusions, the analysis and development of improved proposals in small noise limits is an active research area [90], [94], [33], [85] [88].

In order to quantify the computational complexity of a problem, a recurrent concept is that of intrinsic dimension. Several notions of intrinsic dimension have been used in different fields, including dimension of learning problems [14], [92], [93], of statistical inverse problems [66], of functions in the context of quasi Monte Carlo (QMC) integration in finance applications [17], [71], [56], and of data assimilation problems [22]. The underlying theme is that in many application areas where models are formulated in high dimensional state spaces, there is often a small subspace which captures most of the features of the system. It is the dimension of this subspace that effects the complexity of the problem. In the context of inverse problems the paper [10] proposed a notion of intrinsic dimension, which was shown to have a direct connection with the performance of importance sampling. We introduce a further notion of intrinsic dimension for Bayesian inverse problems which agrees with the notion of effective number of parameters used in machine learning and statistics [14]. We also establish that this notion of dimension and the one in [10] are finite, or otherwise, at the same time. Both intrinsic dimensions account for three key features of the complexity of the inverse problem: the nominal dimension (i.e. the minimum of the dimension of the state space and the data), the size of the observational noise and the regularity of the prior relative to the observation noise. Varying the parameters related to these three features may cause a break-down of absolute continuity. The deterioration of importance sampling in large nominal dimensional limits has been widely investigated [10], [13], [82], [81], [80], [79]. In particular, the key role of the intrinsic dimension, rather than the nominal one, in explaining this deterioration was studied in [10]. Here we study the different behaviour of importance sampling as absolute continuity is broken in the three regimes above, and we investigate whether, in all these regimes, the deterioration of importance sampling may be quantified by the various intrinsic dimensions that we introduce.

2. Importance Sampling

In subsection 2.1 we define importance sampling and in subsection 2.2 we demonstrate the role of the second moment of the target-proposal density, ρ; we prove two non-asymptotic theorems showing $\mathcal{O}\big((\rho/N)^{\frac{1}{2}}\big)$ convergence rate of importance sampling with respect to the number N of particles. Then in subsection 2.3 we show how ρ relates to the effective sample size ess as often defined by practitioners, whilst in subsection 2.4 we link ρ to various distances between probability measures. In subsection 2.5 we highlight the role of the breakdown of absolute continuity in the growth of ρ, as the dimension of the space $\mathcal{X}$ grows. Subsection 2.6 follows with a similar discussion relating to singular limits of the density between target and proposal. Subsection 2.7 contains a literature review and, in particular, sources for all the material in this section.

2.1. General Setting

We consider target μ and proposal π, both probability measures on the measurable space $(\mathcal{X}, \mathcal{F})$, related by (1.1). In many statistical applications interest lies in estimating expectations under μ, for a collection of test functions, using samples from π. For a test function $\phi : \mathcal{X} \to \mathbb{R}$ such that $\mu(|\phi|) < \infty$, the identity

$$\mu(\phi) = \frac{\pi(\phi g)}{\pi(g)},$$

leads to the *autonormalized importance sampling* estimator:

$$\begin{aligned} \mu^N(\phi) &:= \frac{\frac{1}{N}\sum_{n=1}^N \phi(u^n) g(u^n)}{\frac{1}{N}\sum_{m=1}^N g(u^m)}, \quad u^n \sim \pi \quad \text{i.i.d.} \\ &= \sum_{n=1}^N w^n \phi(u^n), \qquad w^n := \frac{g(u^n)}{\sum_{m=1}^N g(u^m)}; \end{aligned} \tag{2.1}$$

here the w^n's are called the *normalized weights*. As suggested by the notation, it is useful to view (2.1) as integrating a function ϕ with respect to the random probability measure $\mu^N := \sum_{n=1}^N w^n \delta_{u^n}$. Under this perspective, importance sampling consists of approximating the target μ by the measure μ^N, which is typically called the *particle approximation of* μ. Note that, while μ^N depends on the proposal π, we suppress this dependence for economy of notation. Our aim is to understand the quality of the approximation μ^N of μ. In particular we would like to know how large to choose N in order to obtain small error. This will quantify the computational complexity of importance sampling.

2.2. The Second Moment of the Target-Proposal Density

A fundamental quantity in addressing this issue is ρ, defined by

$$\rho := \frac{\pi(g^2)}{\pi(g)^2}\,. \tag{2.2}$$

Thus ρ is the second moment of the Radon-Nikodym derivative of the target with respect to the proposal. The Cauchy-Schwarz inequality shows that $\pi(g)^2 \leq \pi(g^2)$ and hence that $\rho \geq 1$. Our first non-asymptotic result shows that, for bounded test functions ϕ, both the bias and the mean square error (MSE) of the autonormalized importance sampling estimator are $\mathcal{O}(N^{-1})$ with constant of proportionality linear in ρ. The proof is in the Appendix, subsubsection 6.2.1.

Theorem 2.1. *Assume that μ is absolutely continuous with respect to π, with square-integrable density g, that is, $\pi(g^2) < \infty$. The bias and MSE of importance sampling over bounded test functions may be characterized as follows:*

$$\sup_{|\phi|\leq 1} \left| \mathbb{E}\big[\mu^N(\phi) - \mu(\phi)\big] \right| \leq \frac{12}{N}\rho\,,$$

and

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

Remark 2.2. *For a bounded test function $|\phi| \leq 1$, we trivially get $|\mu^N(\phi) - \mu(\phi)| \leq 2$; hence the bounds on bias and MSE provided in Theorem 2.1 are useful only when they are smaller than 2 and 4, respectively. The result is strongly suggestive that it is necessary to keep ρ/N small in order to obtain good importance sampling approximations. This heuristic dominates the developments in the remainder of the paper, and in particular our wish to study the behaviour of ρ in various limits.*

It is interesting to contrast Theorem 2.1 to a well-known elementary asymptotic result. First, note that

$$\mu^N(\phi)-\mu(\phi) = \frac{N^{-1}\sum_{n=1}^N \frac{g(u^n)}{\pi(g)}\left[\phi(u^n)-\mu(\phi)\right]}{N^{-1}\sum_{n=1}^N \frac{g(u^n)}{\pi(g)}}.$$

Therefore, under the condition $\pi\left(g^2\right) < \infty$, and provided additionally that $\pi\left(g^2\phi^2\right) < \infty$, an application of the Slutsky lemmas gives that

$$\sqrt{N}\left(\mu^N(\phi)-\mu(\phi)\right) \Longrightarrow N\left(0, \frac{\pi\left(g^2\overline{\phi}^2\right)}{\pi(g)^2}\right), \quad \text{where } \overline{\phi} := \phi - \mu(\phi)\,. \tag{2.3}$$

For bounded $|\phi| \leq 1$, the only condition needed for appealing to the asymptotic result is $\pi\left(g^2\right) < \infty$. Then (2.3) gives that, for large N and since $|\overline{\phi}| \leq 2$,

$$\mathbb{E}\left[\left(\mu^N(\phi)-\mu(\phi)\right)^2\right] \lessapprox \frac{4}{N}\rho\,,$$

which is in precise agreement with our non-asymptotic bound.

In comparison with the asymptotic result (2.3), our non-asymptotic theorem makes an identical assumption on the importance weights, that is $\pi\left(g^2\right) < \infty$, but stronger assumptions on the test functions. We can obtain non-asymptotic bounds on the MSE and bias for much larger classes of test functions but at the expense of more assumptions on the importance weights. The next theorem addresses the issue of relaxing the class of test functions, whilst still deriving nonasymptotic bounds; the proof can be found in the Appendix, subsubsection 6.2.2. To simplify the statement we first introduce the following notation. We write $m_t[h]$ for the t-th *central moment* with respect to π of a function $h : \mathcal{X} \to \mathbb{R}$. That is,

$$m_t[h] := \pi(|h(u) - \pi(h)|^t).$$

We also define, as above, $\overline{\phi} := \phi - \mu(\phi)$.

Theorem 2.3. *Suppose that ϕ and g are such that C_{MSE} defined below is finite:*

$$C_{\mathrm{MSE}} := \frac{3}{\pi(g)^2} m_2[\phi g] + \frac{3}{\pi(g)^4}\pi(|\phi g|^{2d})^{\frac{1}{d}} C_{2e}^{\frac{1}{e}} m_{2e}[g]^{\frac{1}{e}} + \frac{3}{\pi(g)^{2(1+\frac{1}{p})}}\pi(|\phi|^{2p})^{\frac{1}{p}} C_{2q(1+\frac{1}{p})}^{\frac{1}{q}} m_{2q(1+\frac{1}{p})}[g]^{\frac{1}{q}}.$$

Then the bias and MSE of importance sampling when applied to approximate $\mu(\phi)$ may be characterized as follows:

$$\left|\mathbb{E}\left[\mu^N(\phi) - \mu(\phi)\right]\right| \leq \frac{1}{N}\left(\frac{2}{\pi(g)^2} m_2[g]^{\frac{1}{2}} m_2[\overline{\phi}g]^{\frac{1}{2}} + 2C_{\mathrm{MSE}}^{\frac{1}{2}} \frac{\pi(g^2)^{\frac{1}{2}}}{\pi(g)}\right)$$

and

$$\mathbb{E}\left[\left(\mu^N(\phi) - \mu(\phi)\right)^2\right] \leq \frac{1}{N} C_{\mathrm{MSE}}.$$

The constants $C_t > 0, t \geq 2$, satisfy $C_t^{\frac{1}{t}} \leq t - 1$ and the two pairs of parameters d, e, and p, q are conjugate indices.

Remark 2.4. *In Bayesian inverse problems $\pi(g) < \infty$ often implies that $\pi\left(g^s\right) < \infty$ for any positive s; we will demonstrate this in a particular case in section 3. In such a case, Theorem 2.3 combined with Hölder's inequality shows that importance sampling converges at rate N^{-1} for any test function ϕ satisfying $\pi\left(|\phi|^{2+\epsilon}\right) < \infty$ for some $\epsilon > 0$.*

2.3. Effective Sample Size

Many practitioners define the *effective sample size* by the formula

$$\mathsf{ess} := \left(\sum_{n=1}^N (w^n)^2\right)^{-1} = \frac{\left(\sum_{n=1}^N g(u^n)\right)^2}{\sum_{n=1}^N g(u^n)^2} = N \frac{\pi_{\mathrm{MC}}^N(g)^2}{\pi_{\mathrm{MC}}^N\left(g^2\right)},$$

where π_{MC}^N is the empirical Monte Carlo random measure

$$\pi_{\mathrm{MC}}^N := \frac{1}{N}\sum_{n=1}^N \delta_{u^n}, \quad u^n \sim \pi.$$

By the Cauchy-Schwarz inequality it follows that $\mathsf{ess} \leq N$. Furthermore, since the weights lie in $[0, 1]$, we have

$$\sum_{n=1}^N (w^n)^2 \leq \sum_{n=1}^N w^n = 1$$

so that $\mathsf{ess} \geq 1$. These upper and lower bounds may be attained as follows. If all the weights are equal, and hence take value N^{-1}, then $\mathsf{ess} = N$, the optimal situation. On the other hand if exactly k weights take the same value, with the remainder then zero, $\mathsf{ess} = k$; in particular the lower bound of 1 is attained if precisely one weight takes the value 1 and all others are zero.

For large enough N, and provided $\pi\left(g^2\right) < \infty$, the strong law of large numbers gives

$$\mathsf{ess} \approx N/\rho\,.$$

Recalling that $\rho \geq 1$ we see that ρ^{-1} quantifies the proportion of particles that effectively characterize the sample size, in the large particle size asymptotic. Furthermore, by Theorem 2.1, we have that, for large N,

$$\sup_{|\phi|\leq 1} \mathbb{E}\left[\left(\mu^N(\phi) - \mu(\phi)\right)^2\right] \lessapprox \frac{4}{\mathsf{ess}}\,.$$

This provides a further justification for the use of ess as an effective sample size, in the large N asymptotic regime.

2.4. Probability Metrics

Intuition tells us that importance sampling will perform well when the distance between proposal π and target μ is not too large. Furthermore we have shown the role of ρ in measuring the rate of convergence of importance sampling. It is hence of interest to explicitly link ρ to distance metrics between π and μ. In fact we consider asymmetric divergences as distance measures; these are not strictly metrics, but certainly represent useful distance measures in many contexts in probability. First consider the χ^2 divergence, which satisfies

$$D_{\chi^2}(\mu\|\pi) := \pi\left(\left[\frac{g}{\pi(g)} - 1\right]^2\right) = \rho - 1\,. \tag{2.4}$$

The Kullback-Leibler divergence is given by

$$D_{\mathrm{KL}}(\mu\|\pi) := \pi\left(\frac{g}{\pi(g)}\log\frac{g}{\pi(g)}\right)\,,$$

and may be shown to satisfy

$$\rho \geq e^{D_{\mathrm{KL}}(\mu\|\pi)}\,. \tag{2.5}$$

Thus Theorem 2.1 suggests that the number of particles required for accurate importance sampling scales exponentially with the Kullback-Leibler divergence between proposal and target and linearly with the χ^2 divergence.

2.5. High State Space Dimension and Absolute Continuity

The preceding three subsections have demonstrated how, when the target is absolutely continuous with respect to the proposal, importance sampling converges as the square root of ρ/N. It is thus natural to ask if, and how, this desirable convergence breaks down for sequences of target and proposal measures which become increasingly close to singular. To this end, suppose that the underlying space is the Cartesian product $\mathbb{R}^d$ equipped with the corresponding product σ-algebra, the proposal is a product measure and the un-normalized weight function also has a product form, as follows:

$$\pi_d(du) = \prod_{i=1}^{d}\pi_1(du(i)), \quad \mu_d(du) = \prod_{i=1}^{d}\mu_1(du(i)), \quad g_d(u) = \exp\left\{-\sum_{i=1}^{d}h(u(i))\right\},$$

for probability measures π_1, μ_1 on $\mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}^+$ (and we assume it is not constant to remove the trivial case $\mu_1 = \pi_1$). We index the proposal, target, density and ρ with respect to d since interest here lies in the limiting behaviour as d increases. In the setting of (1.1) we now have

$$\mu_d(du) \propto g_d(u)\pi_d(du).$$

By construction g_d has all polynomial moments under π_d and importance sampling for each d has the good properties developed in the previous sections. It is also fairly straightforward to see that μ_∞ and π_∞ are mutually singular when h is not constant: one way to see this is to note that

$$\frac{1}{d}\sum_{i=1}^{d} u(i)$$

has a different almost sure limit under μ_∞ and π_∞. Two measures cannot be absolutely continuous unless they share the same almost sure properties. Therefore μ_∞ is not absolutely continuous with respect to π_∞ and importance sampling is undefined in the limit $d = \infty$. As a consequence we should expect to see a degradation in its performance for large state space dimension d.

To illustrate this degradation, assume that $\pi_1(h^2) < \infty$. Under the product structure (2.7), we have $\rho_d = (\rho_1)^d$. Furthermore $\rho_1 > 1$ (since h is not constant). Thus ρ_d grows exponentially with the state space dimension suggesting, when combined with Theorem 2.1, that exponentially many particles are required, with respect to dimension, to make importance sampling accurate.

A useful perspective on the preceding, which links to our discussion of the small noise limit in the next subsection, is as follows. By the central limit theorem we have that, for large d,

$$g_d(u) \approx c' \exp\bigl(-\sqrt{d}cz\bigr), \quad z \sim N(0,1), \tag{2.6}$$

where $c, c' > 0$ are constants with respect to z; in addition c is independent of dimension d, whilst c' may depend on d. From this it follows that (noting that any constant scaling, such as c', disappears from the definition of ρ_d)

$$\rho_d = \frac{\pi_d(g_d^2)}{\pi_d(g_d)^2} \approx \frac{\mathbb{E}\exp(-2\sqrt{d}cz)}{\bigl(\mathbb{E}\exp(-\sqrt{d}cz)\bigr)^2}, \tag{2.7}$$

where, here, $\mathbb{E}$ denotes expectation with respect to $z \sim N(0,1)$. Using the fact that $\mathbb{E}e^{-az} = e^{a^2/2}$ we see that $\rho_d \approx e^{c^2 d}$.

It is important to realise that it is not the product structure *per se* that leads to the collapse, rather the lack of absolute continuity in the limit of infinite state space dimension. Thinking about the role of high dimensions in this way is very instructive in our understanding of high dimensional problems, but is very much related to the setting in which all the coordinates of the problem play a similar role. This does not happen in many application areas. Often there is a diminishing response of the likelihood to perturbations in growing coordinate

index. When this is the case, increasing the state space dimension has only a mild effect in the complexity of the problem, and it is possible to have well-behaved infinite dimensional limits; we will see this perspective in subsections 3.1, 3.2 and 3.3 for inverse problems, and subsections 4.1, 4.2 and 4.3 for filtering.

2.6. Singular Limits

In the previous subsection we saw an example where for high dimensional state spaces the target and proposal became increasingly close to being mutually singular, resulting in ρ which grows exponentially with the state space dimension. In this subsection we observe that mutual singularity can also occur because of small parameters in the unnormalized density g appearing in (1.1), even in problems of fixed dimension; this will lead to ρ which grows algebraically with respect to the small parameter. To understand this situation let $\mathcal{X} = \mathbb{R}$ and consider (1.1) in the setting where

$$g(u) = \exp\bigl(-\epsilon^{-1}h(u)\bigr)$$

where $h : \mathbb{R} \to \mathbb{R}^+$. Furthermore assume, for simplicity, that h is twice differentiable and has a unique minimum at $u^\star$, and that $h''(u^\star) > 0$. Assume, in addition, that π has a Lebesgue density with bounded first derivative. Then the Laplace method shows that

$$\mathbb{E}\exp\bigl(-2\epsilon^{-1}h(u)\bigr) \approx \exp\bigl(-2\epsilon^{-1}h(u^\star)\bigr)\sqrt{\frac{2\pi\epsilon}{2h''(u^\star)}}$$

and that

$$\mathbb{E}\exp\bigl(-\epsilon^{-1}h(u)\bigr) \approx \exp\bigl(-\epsilon^{-1}h(u^\star)\bigr)\sqrt{\frac{2\pi\epsilon}{h''(u^\star)}}.$$

It follows that

$$\rho \approx \sqrt{\frac{h''(u^\star)}{4\pi\epsilon}}.$$

Thus Theorem 2.1 indicates that the number of particles required for importance sampling to be accurate should grow at least as fast as $\epsilon^{-\frac{1}{2}}$.

2.7. Literature Review

In subsection 2.1 we introduced the importance sampling approximation of a target μ using a proposal π, both related by (1.1). The resulting particle approximation measure μ^N is random because it is based on samples from π. Hence $\mu^N(\phi)$ is a *random* estimator of $\mu(\phi)$. This estimator is in general biased, and therefore a reasonable metric for its quality is the MSE

$$\mathbb{E}\Bigl[\bigl(\mu^N(\phi) - \mu(\phi)\bigr)^2\Bigr],$$

where the expectation is with respect to the randomness in the measure μ^N. We bound the MSE over the class of bounded test functions in Theorem 2.1. In fact we may view this theorem as giving a bound on a distance between the measure μ and its approximation μ^N. To this end let ν and μ denote mappings from an underlying probability space (which for us will be that associated with π) into the space of probability measures on $(\mathcal{X}, \mathcal{F})$; in the following, expectation $\mathbb{E}$ is with respect to this underlying probability space. In [76] a distance $d(\cdot,\cdot)$ beween such random measures is defined by

$$d(\nu,\mu)^2 = \sup_{|\phi|\leq 1} \mathbb{E}\Big(\big(\nu(\phi)-\mu(\phi)\big)^2\Big). \tag{2.8}$$

The paper [76] used this distance to study the convergence of particle filters. Note that if the measures are not random the distance reduces to total variation. Using this distance, together with the discussion in subsection 2.4 linking ρ to the χ^2 divergence, we see that Theorem 2.1 states that

$$d(\mu^N,\mu)^2 \leq \frac{4}{N}\big(1+D_{\chi^2}(\mu\|\pi)\big).$$

In subsection 2.4 we also link ρ to the Kullback-Leibler divergence; the bound (2.5) can be found in Theorem 4.19 of [15].

In subsections 2.5 and 2.6 we studied how limits in which the target and proposal become closer and closer

As was already noted, this suggests the need to increase the number of particles linearly with $D_{\chi^2}(\mu\|\pi)$ or exponentially with $D_{\mathrm{KL}}(\mu\|\pi)$. Provided that $\log\big(\frac{g(u)}{\pi(g)}\big)$, $u\sim\mu$, is concentrated around its expected value, as often happens in large dimensional and singular limits, it has recently been shown [19] that using a sample size of approximately $\exp\big(D_{\mathrm{KL}}(\mu\|\pi)\big)$ is both necessary and sufficient in order to control the L^1 error of the importance sampling estimator $\mu^N(\phi)$. Theorem 2.1 is similar to [29, Theorem 7.4.3]. However the later result uses a metric defined over subclasses of bounded functions. The resulting constants in their bounds rely on covering numbers, which are often intractable. In contrast, the constant ρ in Theorem 2.1 is more amenable to analysis and has several meaningful interpretations that will be explored in the remainder of the paper, including the one resulting in the preceding display. The central limit result in equation (2.3) shows that for large N the upper bound in Theorem 2.1 is sharp. Equation (2.3) can be seen as a trivial application of deeper central limit theorems for particle filters, see [20]. The constants $C_t > 0$, $t \geq 2$ in Theorem 2.3 are determined by the Marcinkiewicz-Zygmund inequality [78]. The proof of Theorem 2.3, provided in subsection 6.2.2 of the Appendix, follows the approach of [31] for evaluating moments of ratios. Further importance sampling results have been proved within the study of convergence properties of various versions of the particle filter as a numerical method for the approximation of the true filtering/smoothing distribution. These results are often formulated in finite dimensional state spaces, under bounded likelihood assumptions and for bounded test functions, see [24], [30], [25], [70], [2]. Generalizations for continuous time filtering can be found in [8] and [42].

The effective sample size ess, introduced in subsection 2.3, is a standard statistic used to assess and monitor particle approximation errors in importance sampling [54], [55]. The effective sample size ess does not depend on any specific test function, but is rather a particular function of the normalized weights which quantifies their variability. So does ρ, and as we show in subsection 2.3 there is an asymptotic connection between both. When interested in assessing the quality of the estimator $\mu^N(\phi)$ for a particular test function, a common diagnosis is the empirical variance of such estimator. In [19], the authors study the limitations of such a diagnosis by showing that in the non-asymptotic regime it fails to capture the distance between the target and the proposal; they also propose a new diagnosis. Our discussion of ess relies on the condition $\pi\big(g^2\big) < \infty$. Intuitively, the particle approximation will be rather poor when this condition is not met. Extreme value theory provides some clues about the asymptotic particle approximation error. First it may be shown that, regardless of whether $\pi\big(g^2\big)$ is finite or not, but simply on the basis that $\pi(g) < \infty$, the largest normalised weight, $w^{(N)}$, will converge to 0 as $N \to \infty$; see for example section 3 of [32] for a review of related results. On the other hand, [69] shows that, for large N,

$$\mathbb{E}\left[\frac{N}{\mathsf{ess}}\right] \approx \int_0^N \gamma S(\gamma) d\gamma,$$

where $S(\gamma)$ is the survival function of the distribution of the un-normalized weights, $\gamma := g(u)$ for $u \sim \pi$. For instance, if the weights have density proportional to γ^{-a-1}, for $1 < a < 2$, then $\pi\big(g^2\big) = \infty$ and, for large enough N and constant C,

$$\mathbb{E}\left[\frac{N}{\mathsf{ess}}\right] \approx C\, N^{-a+2}\,.$$

Thus, in contrast to the situation where $\pi\big(g^2\big) < \infty$, in this setting the effective sample size does not grow linearly with N.

In subsections 2.5 and 2.6 we studied how limits in which the target and proposal become closer and closer to being mutually singular (breakdown of absolute continuity) lead to problems for importance sampling. In subsection 2.5 we studied high dimensional problems, using analysis of problems with product structure to enable analytical tractability of the calculations. This use of product structure was pioneered for MCMC methods in [38]. The product structure was then used in a number of recent papers concerning the behaviour of importance sampling in high nominal dimensions, starting with the seminal paper [10], and leading on to others such as [11], [12], [13], [82], [80], [79], and [81].

In [10, Section 3.2] it is shown that, using (2.6), the maximum normalised importance sampling weight can be approximately written as

$$w^{(N)} \approx \frac{1}{1 + \sum_{n>1} \exp\big\{-\sqrt{d}c(z^{(n)} - z^{(1)})\big\}},$$

where $\{z^n\}_{n=1}^N$ are samples from $N(0,1)$ and the $z^{(n)}$ are the ordered statistics. In [13] a direct but non-trivial calculation shows that if N does not grow exponentially with d, the sum in the denominator converges to 0 in probability and

as a result the maximum weight to 1. Of course this means that all other weights are converging to zero, and that the effective sample size is 1. It chimes with the heuristic derived in subsection 2.5 where we show that ρ grows exponentially with d and that choosing N to grow exponentially is thus necessary to keep the upper bound in Theorem 2.1 small. The phenomenon is an instance of what is sometimes termed *collapse of importance sampling* in high dimensions. This type of behaviour can be obtained for other classes of targets and proposals; see [10], [82].

Within the product setting it may be possible, for some limited classes of problems, to avoid degeneracy of importance sampling-based algorithms for large d at polynomial cost. The idea is to use *tempering*, that is, to introduce a sequence of intermediate distributions $\{\mu_{d,i}\}_{i=1}^{p}$, with $p \geq 1$ depending on d, that 'bridge' the target and proposal measures

$$\frac{d\mu_d}{d\pi_d}(u) = \prod_{i=0}^{p} \frac{d\mu_{d,i+1}}{d\mu_{d,i}}(u),$$

where we have set $\mu_{d,0} := \pi_d$ and $\mu_{d,p+1} := \mu_d$. The distributions $\{\mu_{d,i}\}_{i=1}^{p+1}$ are targeted sequentially using some form of particle filter. A natural way to define the intermediate distributions is by

$$\frac{d\mu_{i+1}}{d\mu_i}(u) = g_d(u)^{a_i}, \quad 0 \leq i \leq p, \tag{2.9}$$

where the temperatures $0 < a_i < 1$ satisfy $\sum_{i=0}^{p} a_i = 1$, and have the effect of 'flattening' the change of measure g_d. The main idea underlying [11] is that using $p = d$ bridging distributions in $\mathbb{R}^d$ and $a_i = 1/d$ leads to d importance sampling steps with $\rho = \mathcal{O}(1)$. On the other hand, not using tempering leads to one importance sampling step with $\rho = \mathcal{O}(e^d)$. Therefore, *as long as one can guarantee that by solving d problems sequentially the errors do not grow exponentially with d*, tempering is advantageous. In this scenario, in order to avoid degradation of importance sampling without tempering, the number of particles needs to grow exponentially with d; with tempering one can hope to avoid collapse with computational cost $\mathcal{O}(Nd^2)$ under the stated assumption about growth of errors. On a related note, [36] proposed a method to combine the ensemble Kalman filter and particle filters. They introduced $p = 1$ bridging distributions, and used an ensemble Kalman filter approximation of $\mu_{d,1}$ to build a weighted particle approximation of μ.

Finally, in subsection 2.6 we use the Laplace method. This is a classical methodology for approximating integrals against near singular integrands, and can be found in many textbooks; see for instance [9]. The interested reader may compare the calculation in subsection 2.5, using the Gaussian approximation, with that arising in subsection 2.6, where the small noise limit is studied. At first glance they are similar in form, but the former calculation leads to exponential behaviour in dimension (since it results from different exponents) whilst the latter leads to algebraic behaviour in small noise (since it results from different normalizing constants).

3. Importance Sampling and Inverse Problems

The previous section showed that the distance between the proposal and the target is key in understanding the computational complexity of importance sampling and the central role played by ρ. In this section we study the computational complexity of importance sampling applied in the context of Bayesian inverse problems. In doing so we introduce a notion of intrinsic dimension.

The Bayesian approach to inverse problems consists of updating incomplete knowledge concerning a variable u, encoded in a prior probability distribution $\mathbb{P}_u$, based on some noisy observations of u, denoted by y. The updated knowledge is encoded in a posterior probability distribution $\mathbb{P}_{u|y}$. We study importance sampling with target $\mu := \mathbb{P}_{u|y}$ and proposal $\pi := \mathbb{P}_u$. To make the analysis tractable we consider linear Gaussian inverse problems.

In subsection 3.1 we describe the setting of the problem, working in a general Hilbert space, but developing finite dimensional intuition in parallel to aid the reader who is not familiar with the theory of Gaussian measures in Hilbert space; furthermore, we include subsection 6.1 in the Appendix which gives background on this theory. Subsection 3.2 introduces various notions of "intrinsic dimension" associated with this problem; a key point to appreciate in the sequel is that this dimension can be finite even when the problem is posed in an infinite dimensional Hilbert space.

We highlight that a useful notion of intrinsic dimension for an inverse problem summarizes how much information is contained in the data – relative to the prior – rather than the dimensions of the unknown u (the state space dimension) or the data y (the data space dimension). We show, in subsection 3.3, that when these latter dimensions are infinite then it is crucial that the posterior is absolutely continuous with respect to the prior in order for the intrinsic dimension to be finite; we also link absolute continuity and finite intrinsic dimension with boundedness of the second moment, ρ, of the Radon-Nikodym derivative of posterior with respect to prior. We then investigate, in subsection 3.4, the behaviour of the intrinsic dimension of the inverse problem as the measures μ and π approach mutual singularity; we study both high nominal dimensional limits and small noise limits. We conclude the section with a literature review in subsection 3.5, containing sources for all the material in this section.

3.1. General Setting

We study the inverse problem of finding u from y where

$$y = Ku + \eta. \tag{3.1}$$

In particular we work in the setting where u is an element of the (potentially infinite dimensional) separable Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle, \|\cdot\|)$. Two cases will help guide the reader:

Example 3.1 (Linear Regression Model)**.** *In the context of the linear regression model, $u \in \mathbb{R}^{d_u}$ is the regression parameter vector, $y \in \mathbb{R}^{d_y}$ is a vector of*

training outputs and $K \in \mathbb{R}^{d_y \times d_u}$ *is the so-called design matrix whose column space is used to construct a linear predictor for the scalar output. In this setting,* $d_u, d_y < \infty$*, although in modern applications they might be both very large, and the case* $d_u \gg d_y$ *is the so-called "large p (here* d_u*) small N (here* d_y*)" problem.*

Example 3.2 (Deconvolution Problem)**.** *In the context of signal deconvolution,* $u \in L^2(0,1)$ *is a square integrable unknown signal on the unit interval,* $K : L^2(0,1) \to L^2(0,1)$ *is a convolution operator* $Ku(x) = (\phi \star u)(x) = \int_0^1 \phi(x - z)u(z)dz$*, and* $y = Ku + \eta$ *is the noisy observation of the convoluted signal where* η *is observational noise. The convolution kernel* ϕ *might be, for example, a Gaussian kernel* $\phi(x) = e^{-\delta x^2}$*. Note also that discretization of the deconvolution problem will lead to a family of instances of the preceding linear regression model, parametrised by the dimension of the discretization space.*

The infinite dimensional setting does require some technical background, and this is outlined in the first subsection of the Appendix. Nevertheless, the reader versed only in finite dimensional Gaussian concepts will readily make sense of the notions of intrinsic dimension described in subsection 3.2 simply by thinking of (potentially infinite dimensional) matrix representations of covariances. In particular, the adjoint, denoted $\cdot^*$, can be thought of as generalization of the concept of transpose, and self-adjoint operators as symmetric matrices. However, to fully appreciate the links made in subsection 3.3, the infinite dimensional setting and the background material from Appendix subsection 6.1 will be helpful.

In equation (3.1) the data y is comprised of the image of the unknown u under a linear map K, with added observational noise η. Here K can be formally thought of as being a bounded linear operator in $\mathcal{H}$, which is ill-posed in the sense that if we attempt to invert the data using the (generalized) inverse of K, we get amplification of small errors η in the observation to large errors in the reconstruction of u. In such situations, we need to use regularization techniques in order to stably reconstruct of the unknown u, from the noisy data y.

We assume Gaussian observation noise $\eta \sim \mathbb{P}_\eta := N(0, \Gamma)$ and adopt a Bayesian approach by putting a prior on the unknown $u \sim \mathbb{P}_u = N(0, \Sigma)$, where $\Gamma : \mathcal{H} \to \mathcal{H}$ and $\Sigma : \mathcal{H} \to \mathcal{H}$ are bounded, self-adjoint, positive-definite linear operators. As discussed in subsection 6.1, if covariance Γ (respectively Σ) is trace class then $\eta \sim \mathbb{P}_\eta$ (respectively $u \sim \mathbb{P}_u$) is almost surely in $\mathcal{H}$. On the other hand, as also discussed in subsection 6.1, when covariance Γ (respectively Σ) is not trace-class we have that $\eta \notin \mathcal{H}$ but $\eta \in \mathcal{Y}$ $\mathbb{P}_\eta$-almost surely (respectively $u \notin \mathcal{H}$ but $u \in \mathcal{X}$ $\mathbb{P}_u$-almost surely) where $\mathcal{Y}$ (respectively $\mathcal{X}$) strictly contains $\mathcal{H}$; indeed $\mathcal{H}$ is compactly embedded into $\mathcal{X}, \mathcal{Y}$.

In this setting the prior $\mathbb{P}_u$ and posterior $\mathbb{P}_{u|y}$ are Gaussian conjugate and $\mathbb{P}_{u|y} = N(m, C)$, with mean and covariance given, under appropriate conditions detailed in the literature review subsection 3.5, by

$$m = \Sigma K^*(K\Sigma K^* + \Gamma)^{-1} y, \tag{3.2}$$

$$C = \Sigma - \Sigma K^*(K\Sigma K^* + \Gamma)^{-1} K\Sigma. \tag{3.3}$$

The reader wishing to derive these formulae using finite dimensional intuition

may note that, using Bayes' rule and completion of the square, the posterior mean and covariance can be expressed via precision matrices as

$$C^{-1} = \Sigma^{-1} + K^*\Gamma^{-1}K, \tag{3.4}$$

$$C^{-1}m = K^*\Gamma^{-1}y. \tag{3.5}$$

Use of the Schur complement yields (3.2).

We tacitly assume that K can be extended to act on elements in $\mathcal{X}$ and that the sum of Ku and η makes sense in $\mathcal{Y}$. In the setting outlined above we assume that the prior acts as a regularization for the inversion of the data y. This is encoded in the following assumption on the relationship between the operators K, Σ and Γ.

Assumption 3.3. *Define $S = \Gamma^{-\frac{1}{2}}K\Sigma^{\frac{1}{2}}$, $A = S^*S$ and assume that A, viewed as a linear operator in $\mathcal{H}$, is bounded. Furthermore, assume that the spectrum of A consists of a countable number of eigenvalues, sorted without loss of generality in a non-increasing way*

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_j \geq \cdots \geq 0.$$

In section 3.5 we give an intuitive explanation for the centrality of A and S, and discuss the role of the assumption in the context of inverse problems.

3.2. Intrinsic Dimension

The operator A defined in Assumption 3.3 plays an important role in what follows because it measures the size of the difference between the prior and posterior covariances Σ and C. The developments in section 2 indicate that a key measure determining the computational complexity of importance sampling is the distance between the target (here the posterior) and the proposal (here the prior). In the Gaussian setting considered in this section the differences between posterior and prior covariances will contribute to this distance and we now develop this idea. Note, however, that we say nothing here about the differences between prior and posterior means.

We illustrate the ideas in finite state/data space dimensions in the first instance, a setting in which we have the following result, proved in the Appendix, subsubsection 6.3.1. For extensions to Hilbert spaces, see the discussion in the literature review section 3.5.

Proposition 3.4. *In the finite dimensional setting, and under the assumption that both Σ and C are invertible,*

$$\mathrm{Tr}\big((C^{-1} - \Sigma^{-1})\Sigma\big) = \mathrm{Tr}(A), \quad \mathrm{Tr}\big((\Sigma - C)\Sigma^{-1}\big) = \mathrm{Tr}\big((I + A)^{-1}A\big).$$

Thus the traces of A and of $(I + A)^{-1}A$ measure the relative differences between the posterior and prior precision and covariance operators, respectively, relative to their prior values. For this reason they provide useful measures of

the computational compexity of importance sampling, motivating the following definitions:

$$\tau := \mathrm{Tr}(A), \qquad \mathsf{efd} := \mathrm{Tr}\big((I+A)^{-1}A\big)\,. \tag{3.6}$$

Note that the trace calculates the sum of the eigenvalues and is well-defined, although may be infinite, in the Hilbert space setting. We refer to efd as effective dimension; both τ and efd are measures of the intrinsic dimension of the inverse problem at hand. Remaining for the moment in the finite dimensional setting, we have the next result. The proof is given in the Appendix, subsubsection 6.3.1:

Proposition 3.5. *Let S and A be defined as in Assumption 3.3, and consider the finite dimensional setting.*

1. *The matrices $\Gamma^{1/2}S(I+A)^{-1}S^*\Gamma^{-1/2} \in \mathbb{R}^{d_y\times d_y}$, $S(I+A)^{-1}S^* \in \mathbb{R}^{d_y\times d_y}$ and $(I+A)^{-1}A \in \mathbb{R}^{d_u\times d_u}$ have the same non-zero eigenvalues and hence the same trace.*
2. *If $\lambda_i > 0$ is a non-zero eigenvalue of A then these three matrices have corresponding eigenvalue $\lambda_i(1+\lambda_i)^{-1} < 1$, and*

$$\mathsf{efd} = \sum_i \frac{\lambda_i}{1+\lambda_i} \le d = \min\{d_u, d_y\}\,.$$

Here, recall, $d = \min\{d_u, d_y\}$ is referred to as the nominal dimension of the problem. Part 2. of the preceding result demonstrates the connection between efd and the physical dimensions of the unknown and observation spaces, whilst part 1. demonstrates the equivalence between the traces of a variety of operators, all of which are used in the literature; this is discussed in greater detail in the literature review of subsection 3.5. In the Hilbert space setting, recall, the intrinsic dimensions efd and τ can be infinite. It is important to note, however, that this cannot happen if the rank of K is finite. That is, the intrinsic dimension efd is finite whenever the unknown u or the data y live in a finite dimensional subspace of $\mathcal{H}$. The following result, proved in subsubsection 6.3.1 of the Appendix, relates efd and τ. It shows in particular that they are finite, or otherwise, at the same time. It holds in the infinite dimensional setting.

Lemma 3.6. *Let Assumption 3.3 hold. Then A is trace class if and only if $(I+A)^{-1}A$ is trace class, and the following inequalities hold*

$$\frac{1}{\|I+A\|}\mathrm{Tr}(A) \le \mathrm{Tr}\left((I+A)^{-1}A\right) \le \mathrm{Tr}(A).$$

As a consequence

$$\frac{1}{\|I+A\|}\tau \le \mathsf{efd} \le \tau. \tag{3.7}$$

We are now ready to study the performance of importance sampling with posterior as target and prior as proposal. In subsection 3.3 we identify conditions under which we can guarantee that ρ in Theorem 2.1 is finite and absolute continuity holds. In subsection 3.4 we then study the growth of ρ as mutual singularity is approached in different regimes. The intrinsic dimensions τ and efd will be woven into these developments.

3.3. Absolute Continuity

In the finite dimensional setting, when both covariance matrices Σ and Γ are strictly positive-definite, the Gaussian proposal and target distributions have densities with respect to the Lebesgue measure. They are hence mutually absolutely continuous and it is hence straightforward to find the Radon-Nikodym derivative of the target with respect to the proposal by taking the ratio of the respective Lebesgue densities once the posterior is identified via Bayes' theorem; this gives:

$$\frac{d\mu}{d\pi}(u) = \frac{d\mathbb{P}_{u|y}}{d\mathbb{P}_u}(u;y) \propto \exp\left(-\frac{1}{2}u^*K^*\Gamma^{-1}Ku + u^*K^*\Gamma^{-1}y\right) =: g(u;y). \quad (3.8)$$

Direct calculation shows that, for $d_u, d_y < \infty$ and Γ invertible, the ratio ρ defined in (2.2) is finite, and indeed that g admits all polynomial moments, all of which are positive. In this subsection we study ρ in the Hilbert space setting. In general there is no guarantee that the posterior is absolutely continuous with respect to the prior; when it is not, g, and hence ρ, are not defined. We thus seek conditions under which such absolute continuity may be established.

To this end, we define the likelihood measure $y|u \sim \mathbb{P}_{y|u} := N(Ku, \Gamma)$, and the joint distribution of (u, y) under the model $\nu(du, dy) := \mathbb{P}_{y|u}(dy|u)\mathbb{P}_u(du)$, recalling that $\mathbb{P}_u = N(0, \Sigma)$. We also define the marginal distribution of the data under the joint distribution, $\nu_y(dy) = \mathbb{P}_y(dy)$. We have the following result, proved in subsubsection 6.3.2 of the Appendix:

Theorem 3.7. *Let Assumption 3.3 hold and let $\mu = \mathbb{P}_{u|y}$ and $\pi = \mathbb{P}_u$. The following are equivalent:*

i) $\mathsf{efd} < \infty$;
ii) $\tau < \infty$;
iii) $\Gamma^{-1/2}Ku \in \mathcal{H}$, *$\pi$-almost surely;*
iv) *for ν_y-almost all y, the posterior μ is well defined as a measure in $\mathcal{X}$ and is absolutely continuous with respect to the prior with*

$$\frac{d\mu}{d\pi}(u) \propto \exp\left(-\frac{1}{2}\left\|\Gamma^{-1/2}Ku\right\|^2 + \frac{1}{2}\langle\Gamma^{-1/2}y, \Gamma^{-1/2}Ku\rangle\right) =: g(u;y), \quad (3.9)$$

where $0 < \pi\big(g(\cdot;y)\big) < \infty$.

Remark 3.8. *Due to the exponential structure of g, we have that assertion (iv) of the last theorem is immediately equivalent to g being ν-almost surely positive and finite and for ν_y-almost all y the second moment of the target-proposal density is finite:*

$$\rho = \frac{\pi\left(g(\cdot;y)^2\right)}{\pi\big(g(\cdot;y)\big)^2} < \infty.$$

Note that item (iii) can also be interpreted as quantifying the dimension of the problem, since it is a requirement on the regularity of the forward image of

the unknown, relative to the noise; such regularity condition typically relates to smoothness of the underlying field, and thus to intrinsic dimension, as we show here.

We have established something very interesting: there are meaningful notions of intrinsic dimension for inverse problems formulated in infinite state/data state dimensions and, when the intrinsic dimension is finite, importance sampling may be possible as there is absolute continuity; moreover, in such situation ρ is finite. Thus, under any of the equivalent conditions i)-iv), Theorem 2.1 can be used to provide bounds on the effective sample size ess, defined in subsection 2.3; indeed the effective sample size is then proportional to N.

It is now of interest to understand how ρ, and the intrinsic dimensions τ and efd, depend on various parameters arising in the problem, such as small noise or the dimension of finite dimensional approximations of the inverse problem. Such questions are studied in the next subsection.

3.4. Singular Limits

The parameter ρ is a complicated nonlinear function of the eigenvalues of A and the data y. However, there are some situations in which we can lower bound ρ in terms of the intrinsic dimensions τ, efd and the size of the eigenvalues of A. We present two classes of examples of this type. The first is a simple but insightful example in which the eigenvalues cluster into a finite dimensional set of large eigenvalues and a set of small remaining eigenvalues. The second involves asymptotic considerations in a simultaneously diagonalizable setting.

3.4.1. Spectral Jump

Consider the setting where u and y both live in finite dimensional spaces of dimensions d_u and d_y respectively. Suppose that A has eigenvalues $\{\lambda_i\}_{i=1}^{d_u}$ with $\lambda_i = C \gg 1$ for $1 \le i \le k$, and $\lambda_i \ll 1$ for $k+1 \le i \le d_u$; indeed we assume that

$$\sum_{i=k+1}^{d_u} \lambda_i \ll 1.$$

Then $\tau(A) \approx Ck$, whilst the effective dimension satisfies efd $\approx k$. Using the identity

$$2D_{\mathrm{KL}}(\mathbb{P}_{u|y}\|\mathbb{P}_u) = \log\Big(\det(I+A)\Big) - \mathrm{Tr}\Big((I+A)^{-1}A\Big) + m^*\Sigma^{-1}m\,.$$

and studying the asymptotics for fixed m, with k and C large, we obtain

$$D_{\mathrm{KL}}(\mathbb{P}_{u|y}||\mathbb{P}_u) \approx \frac{\mathsf{efd}}{2}\log(C)\,.$$

Therefore, using (2.5),

$$\rho \gtrapprox C^{\frac{\mathsf{efd}}{2}}\,.$$

This suggests that ρ grows exponentially with the *number* of large eigenvalues, whereas it has an algebraic dependence on the *size* of the eigenvalues. Theorem 2.1 then suggests that the number of particles required for accurate importance sampling will grow exponentially with the number of large eigenvalues, and algebraically with the size of the eigenvalues. A similar distinction may be found by comparing the behaviour of ρ in large state space dimension in subsection 2.5 (exponential) and with respect to small scaling parameter in subsection 2.6 (algebraic).

3.4.2. Spectral Cascade

We now introduce a three-parameter family of inverse problems, defined through the eigenvalues of A. These three parameters represent the regularity of the prior and the forward map, the size of the observational noise, and the number of positive eigenvalues of A, which corresponds to the nominal dimension. We are interested in investigating the performance of importance sampling, as quantified by ρ, in different regimes for these parameters. We work in the framework of Assumption 3.3, and under the following additional assumption:

Assumption 3.9. *Within the framework of Assumption 3.3, we assume that* $\Gamma = \gamma I$ *and that* A *has eigenvalues* $\left\{\frac{j^{-\beta}}{\gamma}\right\}_{j=1}^{\infty}$ *with* $\gamma > 0$, *and* $\beta \geq 0$. *We consider a truncated sequence of problems with* $A(\beta, \gamma, d)$, *with eigenvalues* $\left\{\frac{j^{-\beta}}{\gamma}\right\}_{j=1}^{d}$, $d \in \mathbb{N} \cup \{\infty\}$. *Finally, we assume that the data is generated from a fixed underlying infinite dimensional truth* $u^\dagger$,

$$y = Ku^\dagger + \eta, \quad Ku^\dagger \in \mathcal{H},$$

and for the truncated problems the data is given by projecting y *onto the first* d *eigenfunctions of* A.

Note that d in the previous assumption is the data space dimension, which agrees here with the nominal dimension. The setting of the previous assumption arises, for example, when d is finite, from discretizing the data of an inverse problem formulated in an infinite dimensional state space. Provided that the forward map K and the prior covariance Σ commute, our analysis extends to the case where both the unknown and the data are discretized in the common eigenbasis. In all these cases, interest lies in understanding how the complexity of importance sampling depends on the level of the discretizations. The parameter γ may arise as an observational noise scaling, and it is hence of interest to study the complexity of importance sampling when γ is small. And finally the parameter β reflects regularity of the problem, as determined by the prior and noise covariances, and the forward map; critical phase transitions occur in computational complexity as this parameter is varied, as we will show.

The intrinsic dimensions $\tau = \tau(\beta, \gamma, d)$ and $\mathsf{efd} = \mathsf{efd}(\beta, \gamma, d)$ read

$$\tau = \frac{1}{\gamma} \sum_{j=1}^{d} j^{-\beta}, \qquad \mathsf{efd} = \sum_{j=1}^{d} \frac{j^{-\beta}}{\gamma + j^{-\beta}}. \tag{3.10}$$

Table 1 shows the scalings of the effective dimensions efd and τ with the model parameters. It also shows how ρ behaves under these scalings and hence gives, by Theorem 2.1, an indication of the number of particles required for accurate importance sampling in a given regime. In all the scaling limits where ρ grows to infinity the posterior and prior are approaching mutual singularity; we can then apply Theorem 2.1 to get an indication of how importance sampling deteriorates in these limits.

Note that by Theorem 3.7 we have $\tau(\beta, \gamma, d) < \infty$ if and only if $\mathsf{efd}(\beta, \gamma, d) < \infty$. It is clear from (3.10) that $\tau = \infty$ if and only if $\{d = \infty, \beta \leq 1\}$. By Theorem 3.7 again, this implies, in particular, that absolute continuity is lost in the limit as $d \to \infty$ when $\beta \leq 1$, and as $\beta \searrow 1$ when $d = \infty$. Absolute continuity is also lost in the limit $\gamma \to 0$, in which the posterior is fully concentrated around the data (at least in those directions in which the data live). In this limit we always have $\tau = \infty$, whereas $\mathsf{efd} < \infty$ in the case where $d < \infty$ and $\mathsf{efd} = \infty$ when $d = \infty$. Note that in the limit $\gamma = 0$ Assumption 3.3 does not hold, which explains why τ and efd are not finite simultaneously. Indeed, as was noted before, efd is always bounded by the nominal dimension d irrespective of the size γ of the noise.

Some important remarks on Table 1 are:

- ρ grows *algebraically* in the small noise limit ($\gamma \to 0$) if the nominal dimension d is finite.
- ρ grows *exponentially* in τ or efd as the nominal dimension grows ($d \to \infty$), or as the prior becomes rougher ($\beta \searrow 1$).
- ρ grows *factorially* in the small noise limit ($\gamma \to 0$) if $d = \infty$, and in the joint limit $\gamma = d^{-\alpha}, d \to \infty$. The exponent in the rates relates naturally to efd.

The scalings of τ and efd can be readily deduced by comparing the sums defining τ and efd with integrals. The analysis of the sensitivity of ρ to the model parameters relies on an explicit expression for this quantity. The details are in the Appendix, subsubsection 6.3.3.

3.5. Literature Review

Some more examples of linear inverse problems in both finite and infinite dimensions include the Radon Inversion used for X-ray imaging, the determination of the initial temperature from later measurements and the inversion of the Laplace transform. Many case studies as well as more elaborate nonlinear inverse problems can be found for example in [46], [86] which adopt a Bayesian approach to their solution, and [34], [72] which adopt a classical approach. The Bayesian

Regime	Parameters	efd	τ	ρ
Small noise	$\gamma \to 0$, $d < \infty$	d	γ^{-1}	$\mathcal{O}\left(\gamma^{-d/2}\right)$
	$\gamma \to 0$, $d = \infty$, $\beta > 1$	$\gamma^{-1/\beta}$	γ^{-1}	$\mathcal{O}\left(\gamma^{-\epsilon\beta(\gamma^{-1/\beta-\epsilon})/2}\right)$
Large d	$d \to \infty$, $\beta < 1$	$d^{1-\beta}$	$d^{1-\beta}$	$\mathcal{O}_P\left(\exp(d^{1-\beta})\right)$
Small noise	$\gamma = d^{-\alpha}$, $d \to \infty, \beta > 1, \alpha > \beta$	d	d^{α}	$\mathcal{O}\left(d^{(\alpha-\beta)d}\right)$
and large d	$\gamma = d^{-\alpha}$, $d \to \infty, \beta > 1, \alpha < \beta$	$d^{\alpha/\beta}$	d^{α}	$\mathcal{O}\left(d^{\epsilon d^{\alpha/\beta-\epsilon}}\right)$
	$\gamma = d^{-\alpha}$, $d \to \infty, \beta < 1, \alpha > \beta$	d	$d^{1+\alpha-\beta}$	$\mathcal{O}\left(d^{(\alpha-\beta)d}\right)$
	$\gamma = d^{-\alpha}$, $d \to \infty, \beta < 1, \alpha < \beta$	$d^{1+\alpha-\beta}$	$d^{1+\alpha-\beta}$	$\mathcal{O}\left(d^{\epsilon d^{\alpha/\beta-\epsilon}}\right)$
Regularity	$d = \infty, \beta \searrow 1$	$\frac{1}{\beta-1}$	$\frac{1}{\beta-1}$	$\mathcal{O}_P\left(\exp(\frac{1}{\beta-1})\right)$

TABLE 1

Scalings of efd, τ *and* ρ *with model parameters. The scalings for* ρ *are for almost all realizations of the data* y *when* $\gamma \to 0$*, and in probability for those regimes where* γ *is fixed.*

approach we undertake, in the example of linear regression (Example 3.1) becomes the Gaussian conjugate Bayesian analysis of linear regression models, as in [64].

Formulae (3.4), (3.5) for the mean and covariance expressed via precisions in the finite dimensional setting may be found in [64]. In fact sense can be given to these formulae in the infinite dimensional setting as well; see [4, Section 5]. Formulae (3.2), (3.3) in the infinite dimensional setting are derived in [67], [62]; in the specific case of inverting for the initial condition in the heat equation they were derived in [35]. The Appendix, subsection 6.1, has a discussion of Gaussian measures in Hilbert spaces and contains further background references.

As mentioned above, we tacitly assume that K can be extended to act on elements in $\mathcal{X}$ and that the sum of Ku and η makes sense in $\mathcal{Y}$. This assumption holds trivially if the three operators K, Σ, Γ are simultaneously diagonalizable. It also holds in non-diagonal settings, in which it is possible to link the domains of powers of the three operators by appropriate embeddings; for some examples see [4, Section 7].

The assumption that the spectrum of A introduced in Assumption 3.3 consists of a countable number of eigenvalues, means that the operator A can be thought of as an infinitely large diagonal matrix. It holds if A is compact [61, Theorem 3, Chapter 28], but is in fact more general since it covers, for example, the non-compact case $A = I$.

In the finite dimensional setting the assumption that A is bounded holds automatically if the noise covariance is invertible. The centrality of $S = \Gamma^{-\frac{1}{2}} K \Sigma^{\frac{1}{2}}$ may then be understood as follows. Under the prior and noise models we may write $u = \Sigma^{\frac{1}{2}} u_0$ and $\eta = \Gamma^{\frac{1}{2}} \eta_0$ where u_0 and η_0 are independent centred Gaussians with identity covariance operators (white noises). Under the assumption that Γ is invertible we then find that we may write (3.1), for $y_0 = \Gamma^{-\frac{1}{2}} y$, as

$$y_0 = S u_0 + \eta_0. \tag{3.11}$$

Thus all results may be derived for this inverse problem, and translated back to the original setting. The role of S, and hence A, is thus clear in the finite dimensional setting. This intuition carries over to infinite dimensions.

We note here that the inverse problem

$$y_0 = w_0 + \eta_0 \tag{3.12}$$

with η_0 a white noise and $w_0 \sim N(0, SS^*)$ is equivalent to (3.11), but formulated in terms of unknown $w_0 = Au_0$, rather than unknown u_0. In this picture the key operator is SS^* rather than $A = S^*S$. Note that by Lemma 6.5 $\mathrm{Tr}(S^*S) = \mathrm{Tr}(SS^*)$. Furthermore, if S is compact the operators SS^* and S^*S have the same nonzero eigenvalues [34, Section 2.2], thus $\mathrm{Tr}((I + SS^*)^{-1}SS^*) = \mathrm{Tr}((I + S^*S)^{-1}S^*S)$. The last equality holds even if S is non-compact, since then Lemma 6.5 together with Lemma 3.6 imply that both sides are infinite. Combining, we see that the intrinsic dimension (τ or efd) is the same regardless of whether we view w_0 or u_0 as the unknown. In particular, the assumption that A is bounded is equivalent to assuming that the operators S, S^* or SS^* are bounded [61, Theorem 14, Chapter 19]. For the equivalent formulation (3.12), the posterior mean equation (3.2) is

$$m = SS^*(SS^* + I)^{-1}y.$$

If SS^* is compact, that is, if its nonzero eigenvalues λ_i go to 0, then m is a regularized approximation of w_0, since the components of the data corresponding to small eigenvalues λ_i are shrunk towards zero. On the other hand, if SS^* is unbounded, that is, if its nonzero eigenvalues λ_i go to infinity, then there is no regularization and high frequency components in the data remain almost unaffected by SS^* in m. Therefore, the case SS^* is bounded is the borderline case for having that the prior has a regularizing effect in the inversion of the data.

In subsection 3.2 we study notions of dimension for Bayesian inverse problems. In the Bayesian setting, the prior infuses information and correlations on the components of the unknown u, reducing the number of parameters that are estimated. In the context of Bayesian or penalized likelihood frameworks, this has led to the notion of *effective number of parameters*, defined as

$$\mathrm{Tr}\Big(\Gamma^{1/2}S(I + S^*S)^{-1}S^*\Gamma^{-1/2}\Big).$$

This quantity agrees with efd by Proposition 3.5 and has been used extensively in Statistics and Machine Learning, see for example [84], and section 3.5.3 of [14] and references therein. One motivation for this definition is based on a Bayesian version of the "hat matrix", see for example [84]. However, in this article we provide a different motivation that is more relevant to our aims. Moreover, rather than as an effective number of parameters, we interpret efd as the effective dimension of the Bayesian linear model. Similar forms of effective dimension have been used for learning problems in [92], [93], [18] and for statistical inverse problems in [66]. In all of these contexts the size of the operator A quantifies how informative the data is; see the discussion below. The paper [13] introduced the notion of $\tau = \mathrm{Tr}(A)$ as an effective dimension for importance sampling within linear inverse problems and filtering. In that paper several transformations of the inverse problem are performed before doing the analysis. We undo these

transformations. The role of τ in the performance of the Ensemble Kalman filter had been previously studied in [37].

The operator A has played an important role in the study of linear inverse problems. First, it has been used for obtaining posterior contraction rates in the small noise limit, see the operator B^*B in [63], [5]. Its use was motivated by techniques for analyzing classical regularization methods, in particular regularization in Hilbert scales see [34, Chapter 8]. Furthermore, its eigenvalues and eigendirections can be used to determine (optimal) low-rank approximations of the posterior covariance [16], [83, Theorem 2.3]. The analogue of A in nonlinear Bayesian inverse problems is the so-called prior-preconditioned data-misfit Hessian, which has been used in [68] to design Metropolis Hastings proposals.

Proposition 3.5 shows that efd is at most as large as the nominal dimension, in finite dimensional settings. The difference between both is a measure of the effect the prior has on the inference relative to the maximum likelihood solution. This difference increases as the size of Σ increases, or as the correlation among the vectors that form the columns of K increases, while the difference decreases as the size of Γ decreases or as the correlations in Γ increase. Note also that in finite dimensional settings, Proposition 3.4 shows that efd quantifies how much change there is in going from the posterior to the prior, measured in terms of change in the covariance, in units of the prior; and τ plays a similar role expressed in terms of change in the precisions, again in units of the prior. By the cyclic property of the trace, Lemma 6.5(ii), and by Proposition 3.4, τ and efd may also be characterized as follows:

$$\begin{aligned}\tau &= \mathrm{Tr}\big((C^{-1}-\Sigma^{-1})\Sigma\big) = \mathrm{Tr}\big((\Sigma - C)C^{-1}\big),\\ \mathsf{efd} &= \mathrm{Tr}\big((\Sigma - C)\Sigma^{-1}\big) = \mathrm{Tr}\big((C^{-1}-\Sigma^{-1})C\big).\end{aligned}$$

Thus we may also view efd as measuring the change in the precision, measured in units given by the posterior precision; whilst τ measures the change in the covariance, measured in units given by the posterior covariance.

Note that Proposition 3.4 also holds in the general Hilbert space setting, provided formula (3.4) for the posterior precision operator can be justified; see Remark 6.6 in the Appendix. The above alternative identities for τ and efd can also be justified in those settings, using analogous techniques. We hence have that the interpretations of τ and efd discussed in the previous paragraph, carry over to such infinite dimensional settings.

In many applications, the unknown $u \in \mathbb{R}^{d_u}$ and often the data $y \in \mathbb{R}^{d_y}$ correspond to discretizations of continuum functions living in Hilbert spaces. The canonical illustration arises from discretizting Example 3.2 to obtain Example 3.1. In such situations the three matrices K, Γ, Σ defining the Bayesian inverse problem also correspond to discretizations of infinite dimensional linear operators. It is of interest to understand the performance of importance sampling as the discretization level increases in order to decide how to distribute the available budget between using more particles or investing in higher discretization levels. A deep analysis of importance sampling in the large d limit can be found in [10]. The authors show that, if $\beta \leq 1$ and $d \to \infty$, the maximum importance

sampling weight converges to 1 in probability, unless the number of particles grows super-exponentially with, essentially, $\tau(d)$. Here we show that $\rho(d)$ grows exponentially with $\tau(d)$ (and $\mathsf{efd}(d)$), which together with Theorem 2.1 suggests also the need to increase the number of samples exponentially with dimension.

It is straightforward to check that since $Ku^\dagger \in \mathcal{H}$, the probability measure of the data in Assumption 3.9 is equivalent to the marginal probability measure of the data under the model, $\nu_y(dy)$. Hence for data of the form of Assumption 3.9, Theorem 3.7 implies that the posterior is absolutely continuous with respect to the prior, almost surely with respect to the noise distribution.

The deviance information criterion introduced in [84], is based on a notion of effective number of parameters that generalises the one we discuss in this paper to more general Bayesian hierarchical models.

In the context of inverse problems, by (3.9), the tempered un-normalized likelihood $g(u;y)^a$ takes the form

$$g(u;y)^a = \exp\left(-\frac{a}{2\gamma}\left\|\Gamma^{-1/2}Ku\right\|^2 + \frac{a}{\gamma}\langle\Gamma^{-1/2}y, \Gamma^{-1/2}Ku\rangle\right). \tag{3.13}$$

This corresponds to the likelihood of our standard inverse problem, but where Γ is replaced by Γ/a and hence A in Assumption 3.3 is scaled by a. In particular, in the context of an inverse problem in the Euclidean space $\mathbb{R}^d$, if $a = \frac{1}{d}$ and $A(d)$ is a discretization of an operator A with eigenvalues bounded by $\lambda_{\max}$ we easily deduce that the tempered problem has intrinsic dimensions $\mathsf{efd}, \tau \leq \lambda_{\max}$, bounded independently of d. Applying this sequentially then leads to the sequence of measures $\mu_{d,i}$ as explained at the end of subsection 2.7. We remark that under the tempering approach d of these problems with bounded effective dimension would need to be solved sequentially; a careful study of the propagation of errors of such sequential scheme would be necessary to understand its complexity, but is beyond the scope of our work. In practice this issue can be ameliorated by including approriate mixing kernels, invariant with respect to $\mu_{d,i}$ for each i, as demonstrated in [50].

4. Importance Sampling and Filtering

In section 2 we introduced importance sampling, and studied its computational complexity. We highlighted the role of the density of the target with respect to the proposal. We also studied the behaviour of importance sampling when approaching loss of absolute continuity between target and proposal. In particular we studied the effect of various singular limits (large nominal dimension, small parameters) in this breakdown. Section 3 studied these issues for Bayesian linear inverse problems. Here we study them for the filtering problem, using the relationship between Bayesian inversion and filtering outlined in the introductory section, and detailed here. In subsection 4.1 we set-up the problem and derive a link between importance sampling based particle filters and the inverse problem. In subsections 4.2 and 4.3 respectively we use this connection to study the intrinsic dimension of filtering, and the connection to absolute continuity

between proposal and target, and in doing so make comparisons between the standard and optimal proposals. Subsection 4.4 contains some explicit computations which enable comparison of the complexity of the two proposals in various singular limits relating to high dimension or small observational noise. We conclude with the literature review subsection 4.5 which overviews the sources for the material herein.

The component of particle filtering which we analyse in this section is only that related to sequential importance sampling; we do not discuss the interaction between the simulated particles which arises via resampling schemes. Such interaction would not typically be very relevant in the two time-unit dynamical systems we study here, but would be necessary to get reasonable numerical schemes when assimilating data over many time units. We comment further on this, and the choice of the assimilation problem we study, in the literature review.

4.1. General Setting

We simplify the notation by setting $j = 0$ in (1.3) to obtain

$$\begin{aligned} v_1 &= Mv_0 + \xi, \quad v_0 \sim N(0, P), \quad \xi \sim N(0, Q), \\ y_1 &= Hv_1 + \zeta, \quad \zeta \sim N(0, R). \end{aligned} \tag{4.1}$$

Note that we have also imposed a Gaussian assumption on v_0. Because of the Markov assumption on the dynamics for $\{v_j\}$, we have that v_0 and ξ are independent. As in section 3 we set-up the problem in a separable Hilbert space $\mathcal{H}$, although the reader versed only in finite dimensional Gaussian measures should have no trouble following the developments, simply by thinking of the covariance operators as (possibly infinite) matrices. We assume throughout that the covariance operators $P, Q, R : \mathcal{H} \to \mathcal{H}$ are bounded, self-adjoint, positive linear operators, but not necessarily trace-class (see the discussion on this trace-class issue in section 3). We also assume that the operators $M, H : \mathcal{H} \to \mathcal{H}$ that describe, respectively, the unconditioned signal dynamics and the observation operator, can be extended to larger spaces if necessary; see the Appendix subsection 6.1 for further details on these technical issues.

Our goal in this section is to study the complexity of importance sampling within the context of both the standard and optimal proposals for particle filtering. For both these proposals we show that there is an inverse problem embedded within the particle filtering method, and compute the proposal covariance, the observation operator and the observational noise covariance. We may then use the material from the previous section, concerning inverse problems, to make direct conclusions about the complexity of importance sampling for particle filters.

The aim of one step of filtering may be expressed as sampling from the target $\mathbb{P}_{v_1, v_0|y_1}$. Particle filters do this by importance sampling, with this measure on the product space $\mathcal{X} \times \mathcal{X}$ as the target. We wish to compare two ways of doing this, one by using the proposal distribution $\mathbb{P}_{v_1|v_0}\mathbb{P}_{v_0}$ and the second by using as

proposal distribution $\mathbb{P}_{v_1|v_0,y_1}\mathbb{P}_{v_0}$. The first is known as the *standard proposal*, and the second as the *optimal proposal.* We now connect each of these proposals to a different inverse problem.

4.1.1. Standard Proposal

For the standard proposal we note that, using Bayes' theorem, conditioning, and that the observation y_1 does not depend on v_0 explicitly,

$$\begin{aligned}\mathbb{P}_{v_1,v_0|y_1} &\propto \mathbb{P}_{y_1|v_1,v_0}\mathbb{P}_{v_1,v_0}\\ &= \mathbb{P}_{y_1|v_1,v_0}\mathbb{P}_{v_1|v_0}\mathbb{P}_{v_0}\\ &= \mathbb{P}_{y_1|v_1}\mathbb{P}_{v_1|v_0}\mathbb{P}_{v_0}.\end{aligned}$$

Thus the density of the target $\mathbb{P}_{v_1,v_0|y_1}$ with respect to the proposal $\mathbb{P}_{v_1|v_0}\mathbb{P}_{v_0}$ is proportional to $\mathbb{P}_{y_1|v_1}$. Although this density concerns a proposal on the joint space of (v_0, v_1), since it involves only v_1 we may consider the related inverse problem of finding v_1, given y_1, and ignore v_0.

In this picture filtering via the standard proposal proceeds as follows:

$$\mathbb{P}_{v_0} \mapsto \mathbb{P}_{v_1} \mapsto \mathbb{P}_{v_1|y_1}.$$

Here the first step involves propagation of proability measures under the dynamics. This provides the proposal $\pi = \mathbb{P}_{v_1}$ used for importance sampling to determine the target $\mu = \mathbb{P}_{v_1|y_1}$. The situation is illustrated in the upper branch of Figure 1. Since

$$\mathbb{E}(v_1 v_1^*) = \mathbb{E}(Mv_0 + \xi)(Mv_0 + \xi)^*,$$

and v_0 and ξ are independent under the Markov assumption, the proposal distribution is readily seen to be a centred Gaussian with covariance $\Sigma = MPM^* + Q$. The observation operator is $K = H$ and the noise covariance $\Gamma = R$. We have established a direct connection between the particle filter, with standard proposal, and the inverse problem of the previous section. We will use this connection to study the complexity of the particle filter, with standard proposal, in what follows.

4.1.2. Optimal Proposal

For the optimal proposal we note that, by conditioning on v_0,

$$\begin{aligned}\mathbb{P}_{v_1,v_0|y_1} &= \mathbb{P}_{v_1|v_0,y_1}\mathbb{P}_{v_0|y_1}\\ &= \mathbb{P}_{v_1|v_0,y_1}\mathbb{P}_{v_0}\frac{\mathbb{P}_{v_0|y_1}}{\mathbb{P}_{v_0}}.\end{aligned}$$

Thus the density of the target $\mathbb{P}_{v_1,v_0|y_1}$ with respect to the proposal $\mathbb{P}_{v_1|v_0,y_1}\mathbb{P}_{v_0}$ is the same as the density of $\mathbb{P}_{v_0|y_1}$ with respect to $\mathbb{P}_{v_0}$. As a consequence, although this density concerns a proposal on the joint space of (v_0, v_1), it is

equivalent to an inverse problem involving only v_0. We may thus consider the related inverse problem of finding v_0 given y_1, and ignore v_1.

In this picture filtering via the optimal proposal proceeds as follows:

$$\mathbb{P}_{v_0} \mapsto \mathbb{P}_{v_0|y_1} \mapsto \mathbb{P}_{v_1|y_1}.$$

Here the first step involves importance sampling with proposal $\pi = \mathbb{P}_{v_0}$ and target $\mu = \mathbb{P}_{v_0|y_1}$. This target measure is then propagated under the conditioned dynamics to find $\mathbb{P}_{v_1|y_1}$; the underlying assumption of the optimal proposal is that $\mathbb{P}_{v_1|v_0,y_1}$ can be sampled so that this conditioned dynamics can be implemented particle by particle. The situation is illustrated in the lower branch of Figure 1. Since

$$y_1 = HMv_0 + H\xi + \zeta$$

the proposal distribution is readily seen to be a centred Gaussian with covariance $\Sigma = P$, the observation operator $K = HM$ and the noise covariance given by the covariance of $H\xi + \zeta$, namely $\Gamma = HQH^* + R$. Again we have established a direct connection between the particle filter, with optimal proposal, and the inverse problem of the previous section. We will use this connection to study the complexity of the particle filter, with optimal proposal, in what follows.

A key assumption of the optimal proposal is the second step: the ability to sample from the conditioned dynamics $\mathbb{P}_{v_1|v_0,y_1}$ and we make a few comments on this before returning to our main purpose, namely to study complexity of particle filtering via the connection to an inverse problem. The first comment is to note that since we are in a purely Gaussian setting, this conditioned dynamics is itself determined by a Gaussian and so may in principle be performed in a straightforward fashion. In fact the conditioned dynamics remains Gaussian even if the forward model Mv_0 is replaced by a nonlinear map $f(v_0)$, so that the optimal proposal has wider applicability than might at first be appreciated. Secondly we comment that the Gaussian arising in the conditioned dynamics has mean m and variance Ξ given by the formulae

$$\begin{aligned}\Xi &= Q - QH^*(HQH^* + R)^{-1}HQ,\\ m &= Mv_0 + QH^*(HQH^* + R)^{-1}(y_1 - HMv_0).\end{aligned}$$

It is a tacit assumption in what follows that the operators defining the filtering problem are such that $\Xi : \mathcal{H} \to \mathcal{H}$ is well-defined and that $m \in \mathcal{H}$ is well-defined. More can be said about these points, but doing so will add further technicalities without contributing to the main goals of this paper.

4.2. Intrinsic Dimension

Using the inverse problems that arise for the standard proposal and for the optimal proposal, and employing them within the definition of A from Assumption 3.3, we find the two operators A arising for these two different proposals:

$$A := A_{st} := (MPM^* + Q)^{1/2}H^*R^{-1}H(MPM^* + Q)^{1/2}$$

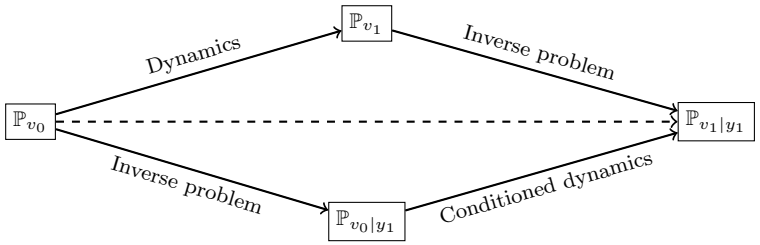

FIG 1. *Filtering step decomposed in two different ways. The upper path first pushes forward the measure $\mathbb{P}_{v_0}$ using the signal dynamics, and then incorporates the observation y_1. The lower path assimilates the observation y_1 first, and then propagates the conditioned measure using the signal dynamics. The standard proposal corresponds to the upper decomposition and the optimal one to the lower decomposition.*

	Standard Proposal	Optimal proposal
Proposal	$\mathbb{P}_{v_0}(dv_0)\mathbb{P}_{v_1\|v_0}(dv_1)$	$\mathbb{P}_{v_0}(dv_0)\mathbb{P}_{v_1\|v_0,y_1}(dv_1)$
BIP	$y_1 = Hv_1 + \eta_{st}$	$y_1 = HMv_0 + \eta_{op}$
Prior Cov.	$MPM^* + Q$	P
Data Cov.	R	$R + HQH^*$
$\log g(u; y_1)$	$-\frac{1}{2}\|Hv_1\|_R^2 + \langle y_1, Hv_1\rangle_R$	$-\frac{1}{2}\|HMv_0\|_{R+HQH^*}^2 + \langle y_1, HMv_0\rangle_{R+HQH^*}$

TABLE 2

for the standard proposal, and

$$A := A_{op} := P^{\frac{1}{2}} M^* H^* (R + HQH^*)^{-1} HMP^{1/2}$$

for the optimal proposal. Again here it is assumed that these operators are bounded in $\mathcal{H}$:

Assumption 4.1. *The operators A_{st} and A_{op}, viewed as linear operators in $\mathcal{H}$, are bounded. Furthermore, assume that the spectra of both A_{st} and A_{op} consist of a countable number of eigenvalues.*

Using these definitions of A_{st} and A_{op} we may define, from (3.6), the intrinsic dimensions τ_{st}, efd_{st} for the standard proposal and τ_{op}, efd_{op} for the optimal one in the following way

$$\tau_{st} = \mathrm{Tr}(A_{st}), \qquad \mathsf{efd}_{st} = \mathrm{Tr}\big((I + A_{st})^{-1} A_{st}\big)$$

and

$$\tau_{op} = \mathrm{Tr}(A_{op}), \qquad \mathsf{efd}_{op} = \mathrm{Tr}\big((I + A_{op})^{-1} A_{op}\big).$$

4.3. Absolute Continuity

The following two theorems are a straightforward application of Theorem 3.7, using the connections between filtering and inverse problems made above. The contents of the two theorems are summarized in Table 2.

Theorem 4.2. *Consider one-step of particle filtering for (4.1). Let $\mu = \mathbb{P}_{v_1|y_1}$ and $\pi = \mathbb{P}_{v_1} = N(0, Q + MPM^*)$. Then the following are equivalent:*

i) $\mathsf{efd}_{st} < \infty$;
ii) $\tau_{st} < \infty$;
iii) $R^{-1/2}Hv_1 \in \mathcal{H}$, *$\pi$-almost surely;*
iv) for ν_y-almost all y, the target distribution μ is well defined as a measure in $\mathcal{X}$ and is absolutely continuous with respect to the proposal with

$$\frac{d\mu}{d\pi}(v_1) \propto \exp\left(-\frac{1}{2}\left\|R^{-1/2}Hv_1\right\|^2 + \frac{1}{2}\langle R^{-1/2}y_1, R^{-1/2}Hv_1\rangle\right) =: g_{st}(v_1; y_1), \tag{4.2}$$

where $0 < \pi\big(g_{st}(\cdot; y_1)\big) < \infty$.

Theorem 4.3. *Consider one-step of particle filtering for (4.1). Let $\mu = \mathbb{P}_{v_0|y_1}$ and $\pi = \mathbb{P}_{v_0} = N(0, Q)$. Then, for $R_{op} = R + HQH^*$, the following are equivalent:*

i) $\mathsf{efd}_{op} < \infty$;
ii) $\tau_{op} < \infty$;
iii) $R_{op}^{-1/2}HMv_0 \in \mathcal{H}$, *$\pi$-almost surely;*
iv) for ν_y-almost all y, the target distribution μ is well defined as a measure in $\mathcal{X}$ and is absolutely continuous with respect to the proposal with

$$\frac{d\mu}{d\pi}(v_0) \propto \exp\left(-\frac{1}{2}\left\|R_{op}^{-1/2}HMv_0\right\|^2 + \frac{1}{2}\langle R_{op}^{-1/2}y_1, R_{op}^{-1/2}HMv_0\rangle\right) =: g_{op}(v_0; y_1), \tag{4.3}$$

where $0 < \pi\big(g_{op}(\cdot; y_1)\big) < \infty$.

Remark 4.4. *Because of the exponential structure of g_{st} and g_{op}, the assertion (iv) in the preceding two theorems is equivalent to g_{st} and g_{op} being ν-almost surely positive and finite and for almost all y_1 the second moment of the target-proposal density is finite. This second moment is given, for the standard and optimal proposals, by*

$$\rho_{st} = \frac{\pi\left(g_{st}(\cdot; y)^2\right)}{\pi\big(g_{st}(\cdot; y)\big)^2} < \infty$$

and

$$\rho_{op} = \frac{\pi\left(g_{op}(\cdot; y)^2\right)}{\pi\big(g_{op}(\cdot; y)\big)^2} < \infty$$

respectively. The relative sizes of ρ_{st} and ρ_{op} determine the relative efficiency of the standard and optimal proposal versions of filtering.

The following theorem shows that there is loss of absolute continuity for the standard proposal whenever there is for the optimal one. The result is formulated in terms of the intrinsic dimension τ, and we show that $\tau_{op} = \infty$ implies $\tau_{st} = \infty$. By Theorem 3.7, this implies the result concerning absolute continuity. Recalling that poor behaviour of importance sampling is intimately related to

such breakdown, this suggests that the optimal proposal is always at least as good as the standard one. The following theorem also gives a condition on the operators H, Q and R under which collapse for both proposals occurs at the same time, irrespective of the regularity of the operators M and P. Roughly speaking this simultaneous collapse result states that if R is large compared to Q then absolute continuity for both proposals is equivalent; and hence collapse of importance sampling happens under one proposal if and only if it happens under the other. Intuitively the advantages of the optimal proposal stem from the noise in the dynamics; they disappear completely if the dynamics is deterministic. The theorem quantifies this idea. Finally, an example demonstrates that there are situations where τ_{op} is finite, so that optimal proposal based importance sampling works well for finite dimensional approximations of an infinite dimensional problem, whilst τ_{st} is infinite, so that standard proposal based importance sampling works poorly for finite dimensional approximations. The proof of the theorem is given in the Appendix, subsection 6.4.

Theorem 4.5. *Suppose that Assumption 4.1 holds. Then,*

$$\tau_{op} \leq \tau_{st}. \tag{4.4}$$

Moreover, if $\mathrm{Tr}\left(HQH^*R^{-1}\right) < \infty$, *then*

$$\tau_{st} < \infty \iff \tau_{op} < \infty.$$

We remark that, under additional simplifying assumptions, we can obtain bounds of the form (4.4) for efd and ρ. We chose to formulate the result in terms of τ since we can prove the bound (4.4) in full generality. Moreover, by Theorem 3.7 the bound in terms of τ suffices in order to understand the different collapse properties of both proposals.

The following example demonstrates that it is possible that $\tau_{op} < \infty$ while $\tau_{st} = \infty$; in this situation filtering via the optimal proposal is well-defined, whilst using the standard proposal it is not. Loosely speaking, this happens if y_1 provides more information on v_1 than v_0.

Example 4.6. *Suppose that*

$$H = Q = R = M = I, \qquad \mathrm{Tr}(P) < \infty.$$

Then, it is straightforward from the definitions that $A_{st} = P + I$ *and* $A_{op} = P/2$. *In an infinite dimensional Hilbert the identity operator has infinite trace,* $\mathrm{Tr}(I) = \infty$, *and so*

$$\tau_{st} = \mathrm{Tr}(A_{st}) = \mathrm{Tr}(P + I) = \infty, \qquad \tau_{op} = \mathrm{Tr}(A_{op}) = \mathrm{Tr}(P/2) < \infty.$$

We have thus established an example of a filtering model for which $\tau_{st} = \infty$ *and* $\tau_{op} < \infty$. *We note that by Theorem 4.5, any such example satisfies the condition* $\mathrm{Tr}(HQH^*R^{-1}) = \infty$. *When this condition is met, automatically* $\tau_{st} = \infty$ *(see*

Regime	Param.	eig(A_{st})	eig(A_{op})	eig(P_∞)	ρ_{st}	ρ_{op}
Small obs. noise	$r \to 0$	r^{-1}	r	r	$\mathcal{O}(r^{-d/2})$	$\mathcal{O}(1)$
	$r = q \to 0$	1	1	$r(=q)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$
Large d	$d \to \infty$	1	1	1	$\mathcal{O}(\exp(d))$	$\mathcal{O}(\exp(d))$

TABLE 3

Scalings of the standard and optimal proposals in small noise and large d regimes for one filter step initialized from stationarity ($P = P_\infty$).

the proof of the Theorem 4.5 in the Appendix, subsection 6.4). However, τ_{op} can still be finite. Indeed, within the proof of that theorem we show that the inequality

$$\tau_{op} \leq \mathrm{Tr}(R^{-1}HMPM^*H^*)$$

always holds. The right-hand side may be finite provided that the eigenvalues of P decay fast enough. A simple example of this situation is where HM is a bounded operator and all the relevant operators have eigenvalues. In this case the Rayleigh-Courant-Fisher theorem – see the Appendix, subsection 6.3 for a reference – guarantees that the eigenvalues of $HMPM^*H^*$ can be bounded in terms of those of P. Again by the Rayleigh-Courant-Fisher theorem, since we are always assuming that the covariance R is bounded, it is possible to bound the eigenvalues of $R^{-1}HMPM^*H^*$ in terms of those of $HMPM^*H^*$. This provides a wider range of examples where $\tau_{st} = \infty$ while $\tau_{op} < \infty$.

4.4. Singular Limits

We are interested in the computational complexity of particle filtering. As stated in Remark 4.4 the values of the second moment of the target-proposal density, ρ_{st} and ρ_{op}, characterize the performance of particle filtering using impotance sampling with the standard and optimal proposals respectively. By comparing the values of ρ_{st} and ρ_{op} we can ascertain situations in which the optimal proposal has significant advantage over the standard proposal. We also recall, from section 3, the role of the intrinsic dimensions in determining the scaling of the second moment of the target-proposal density.

The following example will illustrate a number of interesting phenomena in this regard. In the setting of fixed finite state/data state dimension it will illustrate how the scalings of the various covariances entering the problem effect computational complexity. In the setting of increasing nominal dimension d, when the limiting target is singular with respect to the proposal, it will illustrate how computational complexity scales with d. And finally we will contrast the complexity of the filters in two differing initialization scenarios: (i) from an arbitrary initial covariance P, and from a steady state covariance P_∞. Such a steady state covariance is a fixed point of the covariance update map for the Kalman filter defined by (1.3).

Example 4.7. *Suppose that $M = H = I \in \mathbb{R}^{d\times d}$, and $R = rI$, $Q = qI$, with*

Regime	Param.	eig(A_{st})	eig(A_{op})	ρ_{st}	ρ_{op}
Small obs. noise	$r \to 0$	r^{-1}	1	$\mathcal{O}(r^{-d/2})$	$\mathcal{O}(1)$
	$r = q \to 0$	r^{-1}	r^{-1}	$\mathcal{O}(r^{-d/2})$	$\mathcal{O}(r^{-d/2})$
Large d	$d \to \infty$	1	1	$\mathcal{O}(\exp(d))$	$\mathcal{O}(\exp(d))$

TABLE 4

Scalings of the standard and optimal proposals in small noise and large d regimes for one filter step initialized from $P = pI$.

$r, q > 0$. *A simple calculation shows that the steady state covariance is given by*

$$P_\infty = \frac{\sqrt{q^2 + 4qr} - q}{2} I,$$

and that the operators A_{st} and A_{op} when $P = P_\infty$ are

$$A_{st} = \frac{\sqrt{q^2 + 4qr} + q}{2r} I, \qquad A_{op} = \frac{\sqrt{q^2 + 4qr} - q}{2(q + r)} I.$$

Note that A_{st} and A_{op} are a function of q/r, whereas P_∞ is not.

If the filtering step is initialized outside stationarity at $P = pI$, with $p > 0$, then

$$A_{st} = \frac{p + q}{r} I, \qquad A_{op} = \frac{p}{q + r} I.$$

Both the size and number of the eigenvalues of A_{op}/A_{st} play a role in determining the size of ρ, the second moment of the target-proposal variance. It is thus interesting to study how ρ scales in both the small observational noise regime $r \ll 1$ and the high dimensional regime $d \gg 1$. The results are summarized in Tables 3 and 4. Some conclusions from these tables are:

- *The standard proposal degenerates at an algebraic rate as $r \to 0$, for fixed dimenson d, for both initializations of P.*
- *The optimal proposal is not sensitive to the small observation limit $r \to 0$ if the size of the signal noise, q, is fixed. If started outside stationarity, the optimal proposal degenerates algebraically if $q \propto r \to 0$. However, even in this situation the optimal proposal scales well if initialized in the stationary regime.*
- *In this example the limiting problem with $d = \infty$ has infinite intrinsic dimension for both proposals, because the target and the proposal are mutually singular. As a result, ρ grows exponentially in the large d limit.*
- *Example 4.6 suggests that there are cases where ρ_{st} grows exponentially in the large dimensional limit $d \to \infty$ but ρ_{op} converges to a finite value. This may happen if $\mathrm{Tr}\left(HQH^* R^{-1}\right) < \infty$, but the prior covariance P is sufficiently smooth.*

4.5. Literature Review

In subsection 4.1 we follow [10], [13], [82], [80], [79], [81] and consider one step of the filtering model (1.3). There are two main motivations for studying one step

of the filter. Firstly, if keeping the filter error small is prohibitively costly for one step, then there is no hope that an online particle filter will be successful [10]. Secondly, it can provide insight for filters initialized close to stationarity [22]. As in [80], [79], [81] we cast the analysis of importance sampling in joint space and consider as target $\mu := \mathbb{P}_{u|y_1}$, with $u := (v_0, v_1)$ and with the standard and optimal proposals defined in subsection 4.1.

In general nonlinear, non-Gaussian problems the optimal proposal is usually not implementable, since it is not possible to evaluate the corresponding weights, or to sample from the distribution $\mathbb{P}_{v_1|v_0,y_1}$. However, the optimal proposal is implementable in our framework (see for example [1]) and understanding its behaviour is important in order to build and analyse improved and computable proposals which are informed by the data [88], [40], [89]. It is worth making the point that the so-called "optimal proposal" is really only locally optimal. In particular, this choice is optimal in minimizing the variance of the weights at the given step given that all previous proposals have been already chosen. This choice does not minimize the Monte Carlo variance for some time horizon for some family of test functions. A different optimality criterion is obtained by trying to simultaneously minimize the variance of weights at times $t \leq r \leq t + m$, for some $m \geq 1$, or minimize some function of these variances, say their sum or their maximum. Such look ahead procedures might not be feasible in practice. Surprisingly, examples exist where the standard proposal leads to smaller variance of weights some steps ahead relative to the locally optimally tuned particle fiter; see for example section 3 of [43], and the discussion in [21, Chapter 10]. Still, such examples are quite contrived and experience suggests that local adaptation is useful in practice.

Similarly as for inverse problems, the values of ρ_{st} and ρ_{op} determine the performance of importance sampling for the filtering model with the standard and optimal proposals. These depend in a nonlinear fashion on the eigenvalues of A_{st} and A_{op}. In subsection 4.3 we show that the conditions of collapse for the standard and optimal proposals (found in [80] and [13], respectively) correspond to any of the equivalent conditions of finite dimension or finite ρ described in Theorems 4.2 and 4.3.

In subsection 4.4 we study singular limits in the framework of [22]. Thus, we consider a diagonal filtering setting in the Euclidean space $\mathbb{R}^d$, and assume that all coordinates of the problem play the same role, which corresponds to the extreme case $\beta = 0$ in subsection 3.4. The paper [22] introduced a notion of effective dimension for detectable and stabilizable linear Gaussian data assimilation problems as the Frobenius norm of the steady state covariance of the filtering distribution. It is well known that the detectability and stabilizability conditions ensure the existence of such steady state covariance [57]. This notion of dimension quantifies the success of data assimilation in having reduced uncertainty on the unknown once the data has been assimilated. Therefore the definition of dimension given in [22] is at odds with both τ and efd : it does not quantify how much is learned from the data in one step, but instead how concentrated the filtering distribution is in the time asymptotic regime when the filter is in steady state. Our calculations demonstrate differences which can

occur in the computational complexity of filtering, depending on whether it is initialized in this statistical steady state, or at an arbitrary point.

5. Conclusions

The main motivation for this article is the study of computational complexity of importance sampling, and in particular provision of a framework which unifies the multitude of publications with bearing on this question. We study inverse problems and particle filters in Bayesian models that involve high and infinite state space and data dimensions.

Our study has required revisiting the fundamental structure of importance sampling on general state spaces. We have derived non-asymptotic concentration inequalities for the particle approximation error and related what turns out to be the key parameter of performance, the second moment of the density between the target and proposal, to many different importance sampling input and output quantities.

As a reasonable compromise between mathematical tractability and practical relevance we have focused on Bayesian linear models for regression and statistical inversion of ill-posed inverse problems. We have studied the efficiency of sampling-based posterior inference in these contexts carried out by importance sampling using the prior as proposal. We have demonstrated that performance is controlled by an intrinsic dimension, as opposed to the state space or data dimensions, and we have discussed and related two different measures of this dimension. It is important to emphasise that the intrinsic dimension is really a measure of relative strength between the prior and the likelihood in forming the posterior, as opposed to a measure of "degrees of freedom" in the prior. In other words, infinite-dimensional Bayesian linear models with finite intrinsic dimension are not identified with models for which the prior for the unknown is concentrated on a finite-dimensional manifold of the infinite-dimensional state space.

A similar consideration of balancing tractability and practical relevance has dictated the choice not to study interacting particles typically used for filtering, but rather to focus on one-step filtering using importance sampling. For such problems we introduce appropriate notions of intrinsic dimension and compare the relative merits of popular alternative schemes.

The most pressing topic for future research stemming from this article is the development of concrete recommendations for algorithmic design within classes of Bayesian models used in practice. Within the model structure we have studied here, practically relevant and important extensions include models with non-Gaussian priors on the unknown, nonlinear operators that link the unknown to the data, and unknown hyperparameters involved in the model specification. Linearisation of a nonlinear model around some reasonable value for the unknown (e.g. the posterior mean) is one way to extend our measures of intrinsic dimension in such frameworks. We can expect the subject area to see considerable development in the coming decade.

6. Appendix

6.1. Gaussian Measures in Hilbert Space

In section 3 we study Bayesian inverse problems in the Hilbert space setting. This enables us to talk about infinite dimensional limits of sequences of high dimensional inverse problems and is hence useful when studying the complexity of importance sampling in high dimensions. Here we provide some background on Gaussian measures in Hilbert space. We start by describing how to construct a random draw from a Gaussian measure on an infinite dimensional separable Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle, \|\cdot\|)$. Let $\mathcal{C} : \mathcal{H} \to \mathcal{H}$ be a self-adjoint, positive-definite and trace class operator. It then holds that $\mathcal{C}$ has a countable set of eigenvalues $\{\kappa_j\}_{j\in\mathbb{N}}$, with corresponding normalized eigenfunctions $\{e_j\}_{j\in\mathbb{N}}$ which form a complete orthonormal basis in $\mathcal{H}$.

Example 6.1. *We use as a running example the case where $\mathcal{H}$ is the space of square integrable real-valued functions on the unit interval, $\mathcal{H} = L^2(0,1)$ and where the Gaussian measure of interest is a unit centred Brownian bridge on the interval $(0,1)$. Then $m = 0$ and $\mathcal{C}$ is the inverse of the negative Laplacian on $(0,1)$ with homogeneous Dirichlet boundary conditions. The eigenfunctions and eigenvalues of $\mathcal{C}$ are given by*

$$e_j(t) = \sqrt{2}\sin(j\pi t), \quad \kappa_j = (j\pi)^{-2}.$$

The eigenvalues are summable and hence the operator $\mathcal{C}$ is trace class. For further details see [86].

For any $m \in \mathcal{H}$, we can write a draw $x \sim N(m, \mathcal{C})$ as

$$x = m + \sum_{j=1}^{\infty} \sqrt{\kappa_j}\zeta_j e_j,$$

where ζ_j are independent standard normal random variables in $\mathbb{R}$; this is the Karhunen-Loeve expansion [3, Chapter III.3]. The trace class assumption on the operator $\mathcal{C}$, ensures that $x \in \mathcal{H}$ with probability 1, see Lemma 6.2 in subsection 6.3. The particular rate of decay of the eigenvalues $\{\kappa_j\}$ determines the almost sure regularity properties of x. The idea is that the quicker the decay, the smoother x is, in a sense which depends on the basis $\{e_j\}$. For example if $\{e_j\}$ is the Fourier basis, which is the case if $\mathcal{C}$ is a function of the Laplacian on a tours, then a quicker decay of the eigenvalues of $\mathcal{C}$ means a higher Hölder and Sobolev regularity (see [86, Lemmas 6.25 & 6.27] and [28, Section 2.4]). For the Brownian bridge Example 6.1 above, draws are almost surely in spaces of both Hölder and Sobolev regularity upto (but not including) one half.

The above considerations suggest that we can work entirely in the "frequency" domain, namely the space of coefficients of the element of $\mathcal{H}$ in the eigenbasis of the covariance, the sequence space ℓ^2. Indeed, we can identify the Gaussian measure $N(m, \mathcal{C})$ with the independent product measure $\bigotimes_{j=1}^{\infty} N(m_j, \kappa_j)$,

where $m_j = \langle m, e_j \rangle$. Using this identification, we can define a sequence of Gaussian measures in $\mathbb{R}^d$ which converge to $N(m, \mathcal{C})$ as $d \to \infty$, by truncating the product measure to the first d terms. Even though in $\mathbb{R}^d$ any two Gaussian measures with strictly positive covariances are absolutely continuous with respect to each other (that is, equivalent as measures), in the infinite-dimensional limit two Gaussian measures can be mutually singular, and indeed are unless very stringent conditions are satisfied.

For $N(m, \mathcal{C})$ in $\mathcal{H}$, we define its Cameron-Martin space $E := \mathcal{D}(\mathcal{C}^{-\frac{1}{2}})$, which is characterized as the space of all the shifts in the mean which result in an equivalent Gaussian measure. Since $\mathcal{C}$ is a trace class operator, its inverse (hence also its square root) is an unbounded operator, therefore E is a compact subset of $\mathcal{H}$. In fact E has zero measure under $N(0, \mathcal{C})$. For example, if $\mathcal{C}$ is given by the Brownian bridge Example 6.1, then the Cameron-Martin space E is the Sobolev space of functions which vanish on the boundary and whose first derivative is in $\mathcal{H}$; as mentioned above, draws from this measure only have upto half a derivative in the Sobolev sense. The equivalence or singularity of two Gaussian measures with different covariance operators and different means depends on the compatibility of both their means and covariances, as expressed in the three conditions of the Feldman-Hajek theorem. For more details on the equivalence and singularity of Gaussian measures see [27].

The Karhunen-Loeve expansion makes sense even if $\mathcal{C}$ is not trace class, in which case it defines a Gaussian measure in a space $\mathcal{X} \supset \mathcal{H}$ with a modified covariance operator which *is* trace class. Indeed, let $D : \mathcal{H} \to \mathcal{H}$ be any injective bounded self-adjoint operator such that: a) D is diagonalizable in $\{e_j\}_{j\in\mathbb{N}}$, with (positive) eigenvalues $\{d_j\}_{j\in\mathbb{N}}$; b) the operator $D\mathcal{C}D$ is trace class, that is, $\{\kappa_j d_j^2\}_{j\in\mathbb{N}}$ is summable. Define the weighted inner product $\langle \cdot, \cdot \rangle_{D^{-2}} := \langle D\cdot, D\cdot \rangle$, the weighted norm $\|\cdot\|_{D^{-2}} = \|D\cdot\|$ and the space

$$\mathcal{X} := \overline{\text{span}\{e_j : j \in \mathbb{N}\}}^{\|\cdot\|_{D^{-2}}}.$$

Then the functions $\psi_j = d_j^{-1} e_j,\ j \in \mathbb{N}$, form a complete orthonormal basis in the Hilbert space $(\mathcal{X}, \langle \cdot, \cdot \rangle_{D^{-2}}, \|\cdot\|_{D^{-2}})$. The Karhunen-Loeve expansion can then be written as

$$x = m + \sum_{j=1}^{\infty} \sqrt{\kappa_j}\zeta_j e_j = m + \sum_{j=1}^{\infty} \sqrt{\kappa_j} d_j \zeta_j \psi_j,$$

so that we can view x as drawn from the Gaussian measure $N(m, D\mathcal{C}D)$ in $\mathcal{X}$, where $D\mathcal{C}D$ is trace class by assumption. For example, the case $\mathcal{H} = L^2(0,1)$ and $\mathcal{C} = I$, corresponding to Gaussian white noise for functions on the interval $(0,1)$, can be made sense of in negative Sobolev-Hilbert spaces with $-1/2 - \epsilon$ derivatives, for any $\epsilon > 0$. Finally, we stress that absolute continuity in general and the Cameron-Martin space in particular, are concepts which are independent of the space in which we make sense of the measure. In the Gaussian white noise example, we hence have that the Cameron-Martin space is $E = \mathcal{H}$.

The following lemma is similar to numerous results concerning Gaussian measures in function spaces. Because the precise form which we use is not in the literature, we provide a direct proof.

Lemma 6.2. *Let $\mathcal{X}$ be a separable Hilbert space with orthonormal basis $\{\varphi_j\}_{j\in\mathbb{N}}$. Define the Gaussian measure γ through the Karhunen-Loeve expansion*

$$\gamma := \mathcal{L}\Big(\sum_{j=1}^{\infty} \sqrt{\lambda_j}\xi_j\varphi_j\Big),$$

where λ_j is a sequence of positive numbers and where ξ_j are i.i.d. standard normal. Then draws from γ are in $\mathcal{X}$ almost surely if and only if $\sum_{j=1}^{\infty}\lambda_j < \infty$.

Proof. If $\sum_{j=1}^{\infty}\lambda_j < \infty$, then

$$\mathbb{E}_\gamma \|x\|_{\mathcal{X}}^2 = \mathbb{E}\sum_{j=1}^{\infty}\lambda_j\xi_j^2 = \sum_{j=1}^{\infty}\lambda_j < \infty,$$

hence $x \sim \gamma$ is in $\mathcal{X}$ almost surely.

For the converse, suppose that $x \sim \gamma$ is in $\mathcal{X}$ almost surely. Then

$$\|x\|_{\mathcal{X}}^2 = \sum_{j=1}^{\infty}\lambda_j\xi_j^2 < \infty, \quad \text{a.s.}$$

Note that this implies that $\lambda_j \to 0$, and so in particular $\lambda_\infty := \sup_j \lambda_j < \infty$.

By [47, Theorem 3.17], since $\sqrt{\lambda_j}\xi_j \sim N(0,\lambda_j)$ are independent and symmetric random variables, we get that

$$\sum_{j=1}^{\infty}\mathbb{E}[\lambda_j\xi_j^2 \wedge 1] < \infty.$$

A change of variable gives

$$\begin{aligned}\mathbb{E}[\lambda_j\xi_j^2 \wedge 1] &= \frac{2}{\sqrt{2\pi\lambda_j}}\int_0^1 y^2 e^{-\frac{y^2}{2\lambda_j}}dy \\ &= \frac{2\lambda_j^{\frac{3}{2}}}{\sqrt{2\pi\lambda_j}}\int_0^{1/\sqrt{\lambda_j}} z^2 e^{-\frac{z^2}{2}}dz = \frac{2\lambda_j}{\sqrt{2\pi}}\int_0^{1/\sqrt{\lambda_j}} z^2 e^{-\frac{z^2}{2}}dz.\end{aligned}$$

Thus, for every $j \in \mathbb{N}$,

$$\mathbb{E}[\lambda_j\xi_j^2 \wedge 1] \geq \frac{2\lambda_j}{\sqrt{2\pi}}\int_0^{1/\sqrt{\lambda_\infty}} z^2 e^{-\frac{z^2}{2}}dz.$$

Since the left hand side is summable, we conclude that

$$\sum_{j=1}^{\infty}\lambda_j < \infty.$$

□

6.2. Proofs Section 2

Throughout we denote by π^N_{MC} the empirical random measure

$$\pi^N_{\mathrm{MC}} := \frac{1}{N}\sum_{n=1}^{N} \delta_{u^n}, \quad u^n \sim \pi.$$

We recall that μ^N denotes the particle approximation of μ based on sampling from the proposal π.

6.2.1. Proof of Theorem 2.1

Proof of Theorem 2.1. For the bias we write

$$\begin{aligned}\mu^N(\phi) - \mu(\phi) &= \frac{1}{\pi^N_{\mathrm{MC}}(g)}\pi^N_{\mathrm{MC}}(\phi g) - \mu(\phi) \\ &= \frac{1}{\pi^N_{\mathrm{MC}}(g)}\pi^N_{\mathrm{MC}}\Big(\big(\phi - \mu(\phi)\big)g\Big).\end{aligned}$$

Then, letting $\overline{\phi} := \phi - \mu(\phi)$ and noting that

$$\pi(\overline{\phi} g) = 0$$

we can rewrite

$$\mu^N(\phi) - \mu(\phi) = \frac{1}{\pi^N_{\mathrm{MC}}(g)}\Big(\pi^N_{\mathrm{MC}}(\overline{\phi} g) - \pi(\overline{\phi} g)\Big).$$

The first of the terms in brackets is an unbiased estimator of the second one, and so

$$\begin{aligned}\mathbb{E}\big[\mu^N(\phi) - \mu(\phi)\big] &= \mathbb{E}\left[\Big(\frac{1}{\pi^N_{\mathrm{MC}}(g)} - \frac{1}{\pi(g)}\Big)\Big(\pi^N_{\mathrm{MC}}(\overline{\phi} g) - \pi(\overline{\phi} g)\Big)\right] \\ &= \mathbb{E}\left[\frac{1}{\pi^N_{\mathrm{MC}}(g)\pi(g)}\Big(\pi(g) - \pi^N_{\mathrm{MC}}(g)\Big)\Big(\pi^N_{\mathrm{MC}}(\overline{\phi} g) - \pi(\overline{\phi} g)\Big)\right].\end{aligned}$$

Therefore,

$$\begin{aligned}&\Big|\mathbb{E}\big[\mu^N(\phi) - \mu(\phi)\big]\Big| \\ &\le \Big|\mathbb{E}\Big[\big(\mu^N(\phi) - \mu(\phi)\big)1_{\{2\pi^N_{\mathrm{MC}}(g) > \pi(g)\}}\Big]\Big| + \Big|\mathbb{E}\Big[\big(\mu^N(\phi) - \mu(\phi)\big)1_{\{2\pi^N_{\mathrm{MC}}(g) \le \pi(g)\}}\Big]\Big| \\ &\le \frac{2}{\pi(g)^2}\mathbb{E}\Big[\big|\pi(g) - \pi^N_{\mathrm{MC}}(g)\big|\big|\pi^N_{\mathrm{MC}}(\overline{\phi} g) - \pi(\overline{\phi} g)\big|\Big] + 2\mathbb{P}\Big(2\pi^N_{\mathrm{MC}}(g) \le \pi(g)\Big) \\ &\le \frac{2}{\pi(g)^2}\frac{1}{\sqrt{N}}\pi\big(g^2\big)^{1/2}\frac{2}{\sqrt{N}}\pi\big(g^2\big)^{1/2} + 2\mathbb{P}\Big(2\pi^N_{\mathrm{MC}}(g) \le \pi(g)\Big),\end{aligned}$$

where in the second and third inequality we used that $|\phi| \leq 1$. Now note that

$$\mathbb{P}\Big(2\pi^N_{\mathrm{MC}}(g) \leq \pi(g)\Big) = \mathbb{P}\Big(2(\pi^N_{\mathrm{MC}}(g) - \pi(g)) \leq -\pi(g)\Big) \leq \mathbb{P}\Big(2|\pi^N_{\mathrm{MC}}(g) - \pi(g)| \geq \pi(g)\Big).$$

By the Markov inequality $\mathbb{P}\Big(2\pi^N_{\mathrm{MC}}(g) \leq \pi(g)\Big) \leq \frac{4}{N}\frac{\pi(g^2)}{\pi(g)^2}$, and so

$$\sup_{|\phi|\leq 1} \Big|\mathbb{E}\big[\mu^N(\phi) - \mu(\phi)\big]\Big| \leq \frac{12}{N}\frac{\pi(g^2)}{\pi(g)^2}.$$

This completes the proof of the result for the bias. For the MSE

$$\begin{aligned}
\mu^N(\phi) - \mu(\phi) &= \frac{1}{\pi^N_{\mathrm{MC}}(g)}\pi^N_{\mathrm{MC}}(\phi g) - \frac{1}{\pi(g)}\pi(\phi g) \\
&= \left(\frac{1}{\pi^N_{\mathrm{MC}}(g)} - \frac{1}{\pi(g)}\right)\pi^N_{\mathrm{MC}}(\phi g) - \frac{1}{\pi(g)}\Big(\pi(\phi g) - \pi^N_{\mathrm{MC}}(\phi g)\Big) \\
&= \frac{1}{\pi(g)}\Big(\pi(g) - \pi^N_{\mathrm{MC}}(g)\Big)\mu^N(\phi) - \frac{1}{\pi(g)}\Big(\pi(\phi g) - \pi^N_{\mathrm{MC}}(\phi g)\Big),
\end{aligned} \tag{6.1}$$

and so using the inequality $(a+b)^2 \leq 2(a^2+b^2)$ we obtain

$$\big(\mu^N(\phi) - \mu(\phi)\big)^2 \leq \frac{2}{\pi(g)^2}\left\{\Big(\pi(g) - \pi^N_{\mathrm{MC}}(g)\Big)^2\mu^N(\phi)^2 + \Big(\pi(\phi g) - \pi^N_{\mathrm{MC}}(\phi g)\Big)^2\right\}.$$

Therefore, for $|\phi| \leq 1$,

$$\begin{aligned}
\mathbb{E}\left[\big(\mu^N(\phi) - \mu(\phi)\big)^2\right] &\leq \frac{2}{\pi(g)^2}\left\{\mathbb{E}\left[\Big(\pi(g) - \pi^N_{\mathrm{MC}}(g)\Big)^2\right] + \mathbb{E}\left[\Big(\pi(\phi g) - \pi^N_{\mathrm{MC}}(\phi g)\Big)^2\right]\right\} \\
&= \frac{2}{\pi(g)^2}\Big\{\mathrm{Var}_\pi\big(\pi^N_{\mathrm{MC}}(g)\big) + \mathrm{Var}_\pi\big(\pi^N_{\mathrm{MC}}(\phi g)\big)\Big\} \\
&\leq \frac{2}{N\pi(g)^2}\Big\{\pi\big(g^2\big) + \pi\big(\phi^2 g^2\big)\Big\} \\
&\leq \frac{4}{N}\frac{\pi\big(g^2\big)}{\pi(g)^2},
\end{aligned}$$

and the proof is complete. □

Remark 6.3. *The constant* 12 *for the bias can be somewhat reduced by using in the proof the indicator* $1_{\{a\pi^N_{\mathrm{MC}}(g)\leq\pi(g)\}}$ *instead of* $1_{\{2\pi^N_{\mathrm{MC}}(g)\leq\pi(g)\}}$ *and optimizing over* $a > 0$. *Doing this yields the constant* $C \approx 10.42$ *rather than* $C = 12$.

6.2.2. Proof of Theorem 2.3

The proof of the MSE part of Theorem 2.3 uses the approach of [31] for calculating moments of ratios of estimators. The proof of the bias part is very similar to the proof of the bias part of Theorem 2.1.

In order to estimate the MSE, we use [31, Lemma 2] which in our setting becomes:

Lemma 6.4. *For $0 < \alpha < 1$, it holds*

$$\left|\mu^N(\phi) - \mu(\phi)\right| \leq \frac{|\pi^N_{\mathrm{MC}}(\phi g) - \pi(\phi g)|}{\pi(g)} + \frac{|\pi^N_{\mathrm{MC}}(\phi g)|}{\pi(g)^2}|\pi^N_{\mathrm{MC}}(g) - \pi(g)| + \max_{1\leq n\leq N}|\phi(u^n)|\frac{|\pi^N_{\mathrm{MC}}(g) - \pi(g)|^{1+\theta}}{\pi(g)^{1+\theta}}.$$

The main novelty of the above lemma compared to the bounds we used in the proof of Theorem 2.1, is not the bound on ϕ using the maximum, but rather the introduction of $\theta \in (0,1)$. This will be apparent in the proof of Theorem 2.3 below.

We also repeatedly use Hölder's inequality in the form

$$\mathbb{E}\left[|uv|^s\right] \leq \mathbb{E}\left[|u|^{sa}\right]^{\frac{1}{a}}\mathbb{E}\left[|v|^{sb}\right]^{\frac{1}{b}},$$

for any $s > 0$ and for $a, b > 1$ such that $\frac{1}{a} + \frac{1}{b} = 1$, as well as the Marcinkiewicz-Zygmund inequality [78], which for centered i.i.d. random variables X_n gives

$$\mathbb{E}\left[\left|\sum_{n=1}^N X_n\right|^t\right] \leq C_t N^{\frac{t}{2}}\mathbb{E}\left[|X_1|^t\right], \quad \forall t \geq 2.$$

There are known bounds on the constants, namely $C_t^{\frac{1}{t}} \leq t - 1$, [78]. We apply this inequality in several occasions with $X_n = h(u^n) - \pi(h)$ for different functions h, in which case we get

$$\mathbb{E}\left[\left|\pi^N_{\mathrm{MC}}(h) - \pi(h)\right|^t\right] \leq C_t\mathbb{E}\left[\left|h(u^1) - \pi(h)\right|^t\right]N^{-\frac{t}{2}}, \quad \forall t \geq 2. \tag{6.2}$$

We are now ready to prove Theorem 2.3.

Proof of Theorem 2.3. We first prove the MSE part. By Lemma 6.4 we have that

$$\mathbb{E}\left[\left(\mu^N(\phi) - \mu(\phi)\right)^2\right] \leq 3A_1 + 3A_2 + 3A_3,$$

where A_1, A_2, A_3 correspond to the second moments of the three terms respectively.

1. For the first term we have

$$A_1 = \frac{1}{\pi(g)^2}\mathbb{E}\left[\left(\pi^N_{\mathrm{MC}}(\phi g) - \pi(\phi g)\right)^2\right] \leq \frac{1}{\pi(g)^2}\mathbb{E}\left[\left(\phi(u^1)g(u^1) - \pi(\phi g)\right)^2\right]N^{-1}.$$

2. For the second term, Hölder's inequality gives

$$\begin{aligned} A_2 &= \frac{1}{\pi(g)^4}\mathbb{E}\left[\left|\pi^N_{\mathrm{MC}}(\phi g)\left(\pi^N_{\mathrm{MC}}(g) - \pi(g)\right)\right|^2\right] \\ &\leq \frac{1}{\pi(g)^4}\mathbb{E}\left[\left|\pi^N_{\mathrm{MC}}(\phi g)\right|^{2d}\right]^{\frac{1}{d}}\mathbb{E}\left[\left|\pi^N_{\mathrm{MC}}(g) - \pi(g)\right|^{2e}\right]^{\frac{1}{e}}, \end{aligned}$$

where $\frac{1}{d}+\frac{1}{e}=1$. Use of the triangle inequality yields

$$\mathbb{E}\Big[\big|\pi^N_{\mathrm{MC}}(\phi g)\big|^{2d}\Big]^{\frac{1}{d}} = \frac{1}{N^2}\mathbb{E}\Big[\Big|\sum_{n=1}^N \phi(u^n)g(u^n)\Big|^{2d}\Big]^{\frac{1}{d}} \leq \pi\big(|\phi g|^{2d}\big)^{\frac{1}{d}}.$$

Combining with (6.2) (note that $t=2e>2$) we get

$$A_2 \leq \frac{1}{\pi(g)^4}\pi\big(|\phi g|^{2d}\big)^{\frac{1}{d}} C_{2e}^{\frac{1}{e}}\mathbb{E}\Big[\big|g(u_1)-\pi(g)\big|^{2e}\Big]^{\frac{1}{e}} N^{-1}.$$

3. By Hölder we have

$$A_3 = \frac{1}{\pi(g)^{2(1+\theta)}}\mathbb{E}\left[\max_{1\leq n\leq N}|\phi(u^n)|^2\big|\pi(g)-\pi^N_{\mathrm{MC}}(g)\big|^{2(1+\theta)}\right] \leq \frac{1}{\pi(g)^{2(1+\theta)}}\mathbb{E}\left[\max_{1\leq n\leq N}\big|\phi(u^n)\big|^{2p}\right]^{\frac{1}{p}}\mathbb{E}\left[\big|\pi(g)-\pi^N_{\mathrm{MC}}(g)\big|^{2q(1+\theta)}\right]^{\frac{1}{q}},$$

where $\frac{1}{p}+\frac{1}{q}=1$. Note that

$$\mathbb{E}\left[\max_{1\leq n\leq N}\big|\phi(u^n)\big|^{2p}\right]^{\frac{1}{p}} \leq \mathbb{E}\left[\sum_{n=1}^N\big|\phi(u^n)\big|^{2p}\right]^{\frac{1}{p}} = N^{\frac{1}{p}}\pi\big(|\phi|^{2p}\big)^{\frac{1}{p}}.$$

Combining with (6.2), with $t_\theta = 2q(1+\theta)>2$, we get

$$A_3 \leq \frac{1}{\pi(g)^{2(1+\theta)}}N^{\frac{1}{p}}\pi\big(|\phi|^{2p}\big)^{\frac{1}{p}}C_{t_\theta}^{\frac{1}{q}}\mathbb{E}\Big[\big|g(u^1)-\pi(g)\big|^{t_\theta}\Big]^{\frac{1}{q}}N^{-1-\theta}.$$

Now choosing $\theta=\frac{1}{p}\in(0,1)$ gives the desired order of convergence

$$A_3 \leq \frac{1}{\pi(g)^{2(1+\frac{1}{p})}}\pi(|\phi|^{2p})^{\frac{1}{p}}C_{2q(1+\frac{1}{p})}^{\frac{1}{q}}\mathbb{E}\Big[\big|g-\pi(g)\big|^{2q(1+\frac{1}{p})}\Big]^{\frac{1}{q}}N^{-1}.$$

This completes the proof of the MSE part. For the bias, as in the proof of Theorem 2.1 we have

$$\Big|\mathbb{E}\big[\mu^N(\phi)-\mu(\phi)\big]\Big| \leq \frac{2}{\pi(g)^2}\mathbb{E}\Big[\Big|\pi(g)-\pi^N_{\mathrm{MC}}(g)\Big|\Big|\pi^N_{\mathrm{MC}}(\overline{\phi}g)-\pi(\overline{\phi}g)\Big|\Big] + \Big|\mathbb{E}\Big[\big(\mu^N(\phi)-\mu(\phi)\big)1_{\{2\pi^N_{\mathrm{MC}}(g)\leq\pi(g)\}}\Big]\Big|,$$

where $\overline{\phi} = \phi - \mu(\phi)$. Using the Cauchy-Schwarz inequality we obtain

$$\begin{aligned}
&\Big|\mathbb{E}\big[\mu^N(\phi) - \mu(\phi)\big]\Big| \\
&\quad \le \frac{2}{\pi(g)^2}\mathbb{E}\Big[\big|\pi(g) - \pi^N_{\mathrm{MC}}(g)\big|^2\Big]^{\frac{1}{2}}\mathbb{E}\Big[\big|\pi^N_{\mathrm{MC}}(\overline{\phi}g) - \pi(\overline{\phi}g)\big|^2\Big]^{\frac{1}{2}} \\
&\qquad\qquad + \mathbb{E}\Big[\big(\mu^N(\phi) - \mu(\phi)\big)^2\Big]^{\frac{1}{2}}\mathbb{P}\Big(2\pi^N_{\mathrm{MC}}(g) \le \pi(g)\Big)^{\frac{1}{2}} \\
&\quad \le \frac{2}{\pi(g)^2}\frac{1}{N}\mathbb{E}\Big[\big|g(u^1) - \pi(g)\big|^2\Big]^{\frac{1}{2}}\mathbb{E}\Big[\big|\overline{\phi}(u^1)g(u^1) - \pi(\overline{\phi}g)\big|^2\Big]^{\frac{1}{2}} + \frac{C_{\mathrm{MSE}}^{\frac{1}{2}}}{N^{\frac{1}{2}}}\frac{2}{N^{\frac{1}{2}}}\frac{\pi(g^2)^{\frac{1}{2}}}{\pi(g)},
\end{aligned}$$

where to bound the probability of $2\pi^N_{\mathrm{MC}}(g) \le \pi(g)$ we use the Markov inequality similarly as in the analogous part of the proof of Theorem 2.1. □

6.3. Proofs Section 3

We next state a lemma collecting several useful properties of the trace of linear operators. A compact linear operator T is said to belong in the trace class family, if its singular values $\{\sigma_i\}_{i=1}^{\infty}$ are summable. In this case we write $\mathrm{Tr}(T) = \sum_{i=1}^{\infty}\sigma_i$, while for notational convenience we define the trace even for non-trace class operators, with infinite value. T is said to belong in the Hilbert-Schmidt family, if its singular values are square summable (equivalently if T^*T is Hilbert-Schmidt).

Lemma 6.5. *Let T be an operator on a Hilbert space $\mathcal{H}$. Suppose for the next three items that T is trace class. Then*

i) $\mathrm{Tr}(T^*) = \overline{\mathrm{Tr}(T)}$. *In particular, if the eigenvalues of T are real then* $\mathrm{Tr}(T^*) = \mathrm{Tr}(T)$;

ii) for any bounded operator B in $\mathcal{H}$, $\mathrm{Tr}(TB) = \mathrm{Tr}(BT)$. *This assertion also holds if T and B are Hilbert-Schmidt;*

iii) for any bounded operator B in $\mathcal{H}$, $\mathrm{Tr}(TB) = \mathrm{Tr}(BT) \le \|B\|\,\mathrm{Tr}(T)$.

For any bounded linear operator T, it holds that

iv) $\mathrm{Tr}(T^*T) = \mathrm{Tr}(TT^*)$,

where if T (equivalently T^*) is not Hilbert-Schmidt, we define the trace to be $+\infty$.

If T is a linear operator and P is bounded and positive definite, such that TP^{-1} (equivalently $P^{-\frac{1}{2}}TP^{-\frac{1}{2}}$ or $P^{-1}T$) is bounded, it holds that

v) $\mathrm{Tr}(TP) = \mathrm{Tr}(P^{\frac{1}{2}}TP^{\frac{1}{2}}) = \mathrm{Tr}(PT)$,

where as in (iv) we allow infinite values of the trace.

Finally, suppose that D_1 is positive definite and D_2 is positive semi definite, and that T is self adjoint and bounded in $\mathcal{H}$. Furthermore, assume that $D_1^{-1}T$ and $(D_1 + D_2)^{-1}T$ have eigenvalues. Then

vi) $\mathrm{Tr}(D_1^{-1}T) \geq \mathrm{Tr}\big((D_1+D_2)^{-1}T\big)$.

Proof. The proofs of parts (i)-(iii) can be found in [61, Section 30.2], while (iv) is an exercise in [61, Section 30.8]. Part (v) can be shown using the infinite-dimensional analogue of matrix similarity, see [7, Section 2]. In particular, if we multiply TP to the left by $P^{1/2}$ and to the right by $P^{-1/2}$, we do not change its eigenvalues hence neither its trace, so $\mathrm{Tr}(TP) = \mathrm{Tr}(P^{\frac{1}{2}}TP^{\frac{1}{2}})$. Similarly, if we multiply TP to the left by P and to the right by P^{-1}, we get $\mathrm{Tr}(TP) = \mathrm{Tr}(PT)$. Part (vi) follows from the stronger fact that the ordered eigenvalues of $D_1^{-1}T$ are one by one bounded by the ordered eigenvalues of $(D_1+D_2)^{-1}T$. This in turn can be established using that the eigenvalues of these operators are determined by the generalized eigenvalue problem $Tv = \lambda D_1 v$ and $Tv = \lambda(D_1+D_2)v$, with associated Rayleigh quotients

$$\frac{\langle x, Tx\rangle}{\langle x, D_1 x\rangle} \geq \frac{\langle x, Tx\rangle}{\langle x, (D_1+D_2)x\rangle}, \tag{6.3}$$

and an application of the Rayleigh-Courant-Fisher theorem (see [61] and [77]). □

6.3.1. Proofs of subsection 3.2

Proof of Proposition 3.4. Under the given assumptions, expression (3.4) for C^{-1} is well-defined and gives

$$\Sigma^{\frac{1}{2}} C^{-1} \Sigma^{\frac{1}{2}} = I + A. \tag{6.4}$$

Thus

$$\begin{aligned}
\mathrm{Tr}(A) &= \mathrm{Tr}(C^{\frac{1}{2}} C^{-1} \Sigma^{\frac{1}{2}} - I) \\
&= \mathrm{Tr}\big(C^{\frac{1}{2}}(C^{-1} - \Sigma^{-1})\Sigma^{\frac{1}{2}}\big) \\
&= \mathrm{Tr}\big((C^{-1} - \Sigma^{-1})\Sigma\big),
\end{aligned}$$

where the last equality is justified using the cyclic property of the trace, Lemma 6.5(ii). For the second identity, since $(I+A)^{-1}A = I - (I+A)^{-1}$, we have again by (6.4)

$$\begin{aligned}
\mathrm{Tr}((I+A)^{-1}A) &= \mathrm{Tr}\Big(I - (I+A)^{-1}\Big) \\
&= \mathrm{Tr}\Big(I - \Sigma^{-1/2} C \Sigma^{-1/2}\Big) \\
&= \mathrm{Tr}\Big(\Sigma^{-1/2}(\Sigma - C)\Sigma^{-1/2}\Big) \\
&= \mathrm{Tr}\Big((\Sigma - C)\Sigma^{-1}\Big),
\end{aligned}$$

where the last equality is again justified via the cyclic property of the trace. □

Remark 6.6. *Proposition 3.4 also holds in the general separable Hilbert space setting, provided that formula (3.4) for the precision operator of the posterior is justified, see [4, Section 5]. Indeed, the proofs of the two identities are almost identical to the finite dimensional case, the only difference being in the justification of the last equalities in the two sequences of equalities above. In this case the two trace-commutativity equalities have to be justified using Lemma 6.5(v) rather than Lemma 6.5(ii). In the first case, Lemma 6.5(v) can be applied, since* $A = \Sigma^{\frac{1}{2}}(C^{-1} - \Sigma^{-1})\Sigma^{\frac{1}{2}}$ *is bounded by Assumption 3.3, and* Σ *is assumed to be positive definite and bounded. In the second case, Lemma 6.5(v) can be applied, since by Assumption 3.3 the operator* $(I + A)^{-1}A$ *is bounded, and* Σ *is bounded and positive definite.*

Proof of Proposition 3.5.

1. We have that (v_i, μ_i) is an eigenvector/value pair of the first matrix if and only if $(\Gamma^{-1/2}v_i, \mu_i)$ is of the second. It is also immediate that (v_i, μ_i) is a pair for the second if and only if (S^*v_i, μ_i) is for $A(I+A)^{-1}$. However, it is also easy to check that $A(I+A)^{-1} = (I+A)^{-1}A$.
2. In view of the above, note that (v_i, μ_i) is a pair for $(I + A)^{-1}A$ if an only if $(v_i, \mu_i/(1 - \mu_i))$ is for A. Hence, if λ_i is an eigenvalue of A, $\lambda_i/(1 + \lambda_i)$ is one for the other matrices. Given that this is always less or equal to 1 and the efd is a trace of either $d_y \times d_y$ or $d_u \times d_u$ matrices, the inequality follows immediately.

□

Proof of Lemma 3.6. If A is trace class then it is compact and since it is also self-adjoint and nonnegative it can be shown (for example using the spectral representation of A) that $\left\|(I + A)^{-1}\right\| \leq 1$. Then Lemma 6.5(iii) implies that

$$\mathrm{Tr}\big((I + A)^{-1}A\big) \leq \mathrm{Tr}(A).$$

Assume now that $(I + A)^{-1}A$ is trace class. Then A is too since it is the product of the bounded operator $I + A$ and the trace class operator $(I + A)^{-1}A$, see again Lemma 6.5(iii). In particular,

$$\mathrm{Tr}(A) \leq \|I + A\|\mathrm{Tr}\big((I + A)^{-1}A\big).$$

□

6.3.2. Proofs of subsection 3.3

Proof of Theorem 3.7. *i)* $\Leftrightarrow$ *ii)* is immediate from Lemma 3.6.
ii) $\Leftrightarrow$ *iii)* It holds that $\Gamma^{-\frac{1}{2}}Ku \sim N(0, \Gamma^{-\frac{1}{2}}K\Sigma K^*\Gamma^{-\frac{1}{2}})$ since $\Gamma^{-\frac{1}{2}}Ku$ is a linear transformation of the Gaussian $u \sim \mathbb{P}_u = N(0, \Sigma)$. By Lemma 6.2 and since A has eigenvalues, we hence have that $\Gamma^{-\frac{1}{2}}Ku \in \mathcal{H}$ if and only if $\mathrm{Tr}(\Gamma^{-\frac{1}{2}}K\Sigma K^*\Gamma^{-\frac{1}{2}}) < \infty$.

iii) $\Rightarrow$ *iv)* According to the discussion in subsection 6.1 on the absolute continuity of two Gaussian measures with the same covariance but different

means, the Gaussian likelihood measure $\mathbb{P}_{y|u} = N(Ku, \Gamma)$ and the Gaussian noise measure $\mathbb{P}_\eta = N(0, \Gamma)$ are equivalent if and only if $\Gamma^{-\frac{1}{2}} Ku \in \mathcal{H}$. Under *iii*), we hence have that $\mathbb{P}_{y|u}$ and $\mathbb{P}_\eta$ are equivalent for π-almost all u and under the Cameron-Martin formula [27] for π-almost all u we have

$$\frac{d\mathbb{P}_{y|u}}{d\mathbb{P}_\eta}(y) = \exp\left(-\frac{1}{2}\left\|\Gamma^{-1/2}Ku\right\|^2 + \left\langle \Gamma^{-1/2}y, \Gamma^{-1/2}Ku\right\rangle\right) =: g(u;y).$$

Defining the measure $\nu_0(u,y) := \pi(u) \times \mathbb{P}_\eta(y)$ in $\mathcal{X} \times \mathcal{Y}$, we then immediately have that

$$\frac{d\nu}{d\nu_0}(u,y) = g(u;y),$$

where ν is the joint distribution of (u,y) under the model $y = Ku + \eta$ with u and η independent Gaussians $N(0,\Sigma)$ and $N(0,\Gamma)$ respectively.

We next show that $\pi(g(\cdot;y)) > 0$ for $\mathbb{P}_\eta$-almost all y, which will in turn enable us to use a standard conditioning result to get that the posterior is well defined and absolutely continuous with respect to the prior. Indeed, it suffices to show that $g(u;y) > 0$ ν_0-almost surely. Fix $u \sim \pi$. Then, as a function of $y \sim \mathbb{P}_\eta$ the negative exponent of g is distributed as $N(\frac{1}{2}\|\Gamma^{-\frac{1}{2}}Ku\|^2, \|\Gamma^{-\frac{1}{2}}Ku\|^2)$ where $\|\Gamma^{-\frac{1}{2}}Ku\|^2 < \infty$ with π probability 1. Therefore, for ν_0-almost all (u,y) the exponent is finite and thus g is ν_0-almost surely positive implying that $\pi(g(\cdot;y)) > 0$ for $\mathbb{P}_\eta$-almost all y. Noticing that the equivalence of ν and ν_0 implies the equivalence of the marginal distribution of the data under the model, ν_y, with the noise distribution $\mathbb{P}_\eta$, we get that $\pi(g(\cdot;y)) > 0$ for ν_y-almost all y. Hence, we can apply Lemma 5.3 of [41], to get that the posterior measure $\mathbb{P}_{u|y}(\cdot) = \nu(\cdot|y)$ exists ν_y-almost surely and is given by

$$\frac{d\mu}{d\pi}(u) = \frac{1}{\pi(g)}\exp\left(-\frac{1}{2\gamma}\left\|\Gamma^{-1/2}Ku\right\|^2 + \frac{1}{\gamma}\left\langle \Gamma^{-1/2}y, \Gamma^{-1/2}Ku\right\rangle\right).$$

Finally, we note that since $\frac{d\nu}{d\nu_0} = g$, we have that $\int_{\mathcal{X}\times\mathcal{Y}} g\, d\nu_0(u,y) = 1$. Thus the Fubini-Tonelli theorem implies that $\pi(g(\cdot;y)) < \infty$ for $\mathbb{P}_\eta$-almost all y and hence also for ν_y-almost all y.

iv) $\Rightarrow$ *ii*) Under *iv*) we have that the posterior measure μ which, as discussed in subsection 3.1, is Gaussian with mean and covariance given by (3.2) and (3.3), is y-almost surely absolutely continuous with respect to the prior $\pi = N(0,\Sigma)$. By the Feldman-Hajek theorem [27], we hence have that y-almost surely the posterior mean lives in the common Cameron-Martin space of the two measures. This common Cameron-Martin space is the image space of $\Sigma^{\frac{1}{2}}$ in $\mathcal{H}$. Thus we deduce that $w := \Sigma^{-\frac{1}{2}}\Sigma K^*(K\Sigma K^* + \Gamma)^{-1}y \in \mathcal{H}$ almost surely. We next observe that, under ν, $\Gamma^{-\frac{1}{2}}y \sim N(0, SS^* + I)$. Furthermore

$$w = S^*(SS^* + I)^{-1}\Gamma^{-\frac{1}{2}}y,$$

thus under ν, $w \sim N(0, S^*(SS^* + I)^{-1}S)$ where S is defined in Assumption 3.3. Using Lemma 6.2, we thus get that *iv*) implies that $S^*(SS^* + I)^{-1}S$ is

trace class. Using Lemma 6.5(iv) with $T = (SS^* + I)^{-\frac{1}{2}} S$, we then also get that $(SS^* + I)^{-\frac{1}{2}} SS^* (SS^* + I)^{-\frac{1}{2}}$ is trace class. Since $(SS^* + I)^{\frac{1}{2}}$ is bounded, using Lemma 6.5(iii) twice we get that SS^* is trace class. Finally, again using Lemma 6.5(iv) we get that S^*S is trace class, thus ii) holds. □

6.3.3. Proofs of subsection 3.4

The scalings of τ and efd can be readily deduced by comparing the sums defining τ and efd with integrals:

$$\tau(\beta, \gamma, d) \approx \frac{1}{\gamma} \int_1^d \frac{1}{x^\beta}\, dx, \qquad \mathsf{efd} \approx \int_1^d \frac{1}{1 + \gamma x^\beta} = \gamma^{-1/\beta} \int_\gamma^{d\gamma^{1/\beta}} \frac{1}{1 + y^\beta}\, dy.$$

Our analysis of the sensitivity of $\rho = \rho(\beta, \gamma, d)$ to the model parameters relies in the following expression for ρ, which is valid unless the effective dimension is infinite, i.e. unless $d = \infty$, $\beta \leq 1$.

In the next result, and in the analysis that follows, we ease the notation by using subscripts to denote the coordinate of a vector. Thus we write, for instance, y_j rather than $y(j)$.

Lemma 6.7. *Under Assumption 3.9*

$$\rho = \rho(\beta, \gamma, d) := \prod_{j=1}^{d} \frac{\frac{j^{-\beta}}{\gamma} + 1}{\sqrt{2\frac{j^{-\beta}}{\gamma} + 1}} \exp\left(\sum_{j=1}^{d} \left(\frac{2}{2 + \gamma j^\beta} - \frac{1}{1 + \gamma j^\beta} \right) \frac{y_j^2}{\gamma} \right), \tag{6.5}$$

which is finite for ν_y-almost all y.

Proof of Lemma 6.7. We rewrite the expectation with respect to π as an expectation with respect to the law of Ku as follows. Note that here u_j is a dummy integration variable, which represents the j-th corrdinate of Ku, rather than

that of u. Precisely, we have:

$$\begin{aligned}
\pi\big(g(\cdot,y)\big) &= \int_{\mathcal{X}} g(u,y)d\pi(u) \\
&= \int_{\mathcal{X}} \exp\left(-\frac{1}{2\gamma}\sum_{j=1}^{\infty} u_j^2 + \frac{1}{\gamma}\sum_{j=1}^{d} y_j u_j\right) d\left(\bigotimes_{j=1}^{d} N\left(0, j^{-\beta}\right)(u_j)\right) \\
&= \prod_{j=1}^{d} \int_{\mathbb{R}} \exp\left(-\frac{1}{2\gamma}u_j^2 + \frac{1}{\gamma}y_j u_j\right) \frac{\exp\left(-\frac{j^\beta u_j^2}{2}\right)}{\sqrt{2\pi j^{-\beta}}} du_j \\
&= \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi j^{-\beta}}} \int_{\mathbb{R}} \exp\left(-(\gamma^{-1}+j^\beta)\frac{u_j^2}{2} + \frac{1}{\gamma}y_j u_j\right) du_j \\
&= \prod_{j=1}^{d} \frac{\exp\left(\frac{\gamma^{-2}y_j^2}{2(\gamma^{-1}+j^\beta)}\right)}{\sqrt{2\pi j^{-\beta}}} \int_{\mathbb{R}} \exp\left(-(\gamma^{-1}+j^\beta)\frac{\left(u_j - \frac{\gamma^{-1}y_j}{\gamma^{-1}+j^\beta}\right)^2}{2}\right) du_j \\
&= \prod_{j=1}^{d} \sqrt{\frac{j^\beta}{\gamma^{-1}+j^\beta}} \exp\left(\frac{\gamma^{-2}y_j^2}{2(\gamma^{-1}+j^\beta)}\right) \\
&= \prod_{j=1}^{d} \sqrt{\frac{\gamma j^\beta}{1+\gamma j^\beta}} \exp\left(\frac{\gamma^{-1}y_j^2}{2(1+\gamma j^\beta)}\right).
\end{aligned}$$

Thus,

$$\pi\big(g(\cdot,y)\big)^2 = \prod_{j=1}^{d} \frac{\gamma j^\beta}{1+\gamma j^\beta} \exp\left(\frac{\gamma^{-1}y_j^2}{1+\gamma j^\beta}\right)$$

and

$$\pi\big(g(\cdot,y)^2\big) = \prod_{j=1}^{d} \sqrt{\frac{\gamma j^\beta}{2+\gamma j^\beta}} \exp\left(\frac{2\gamma^{-1}y_j^2}{2+\gamma j^\beta}\right),$$

Taking the corresponding ratio gives the expression for ρ. □

Analysis of scalings of ρ. Here we show how to obtain the scalings in Table 1. Taking logarithms in (6.5)

$$\log(\rho) = \sum_{j=1}^{d} \log\left(\frac{\frac{j^{-\beta}}{\gamma}+1}{\sqrt{2\frac{j^{-\beta}}{\gamma}+1}}\right) + \sum_{j=1}^{d}\left(\frac{2}{2+\gamma j^\beta} - \frac{1}{1+\gamma j^\beta}\right)\gamma^{-1}y_j^2. \tag{6.6}$$

Note that every term of both sums is positive. In the small noise regimes the first sum dominates, whereas in the large d, $\beta \searrow 1$ the second does. We show here how to find the scaling of $\gamma \to 0$ when $d = \infty$.

We have that

$$\log(\rho) \geq \sum_{j=1}^{\infty} \log\left(\frac{\frac{j^{-\beta}}{\gamma}+1}{\sqrt{2\frac{j^{-\beta}}{\gamma}+1}}\right)$$
$$\approx \int_1^{f(\gamma)} \log\left(\frac{\frac{x^{-\beta}}{\gamma}+1}{\sqrt{2\frac{x^{-\beta}}{\gamma}+1}}\right) dx + \int_{f(\gamma)}^{\infty} \log\left(\frac{\frac{x^{-\beta}}{\gamma}+1}{\sqrt{2\frac{x^{-\beta}}{\gamma}+1}}\right) dx$$

where $f(\gamma)$ is a function of γ that we are free to choose. Choosing $f(\gamma) = \gamma^{-1/\beta-\epsilon}$ (ϵ small) the first integral dominates the second one and, for small γ, $\log(\rho) \geq \gamma^{-1/\beta-\epsilon}\log(\gamma^{-\epsilon\beta/2})$ from where the result in Table 1 follows. The joint large d, small γ scalings can be established similarly.

When the second sum in (6.6) dominates, the scalings hold in probability. To illustrate this, we study here how to derive the large d limit with $\beta < 1$. Without loss of generality we can assume in what follows that each y_j is centered, i.e. $y_j \sim N(0,\gamma)$ instead of $y_j \sim N\big((Ku)_j^\dagger, \gamma\big)$. This is justified since, for any $c > 0$,

$$\mathbb{P}(y_j^2 \geq c) = \mathbb{P}(|y_j| \geq c^{1/2}) \geq \mathbb{P}(|y_j - (Ku)_j^\dagger| \geq c^{1/2}).$$

Neglecting the first sum in (6.6), which can be shown to be of lower order in d, we get

$$\sum_{j=1}^{d}\left(\frac{2}{2+\gamma j^\beta} - \frac{1}{1+\gamma j^\beta}\right)\gamma^{-1}y_j^2 = S(y,d).$$

Using that $\mathbb{E}y_j^2 = \gamma$,

$$\mathbb{E}\log(\rho) \geq \sum_{j=1}^{d}\left(\frac{2}{2+\gamma j^\beta} - \frac{1}{1+\gamma j^\beta}\right)$$
$$\approx \int_1^d \left(\frac{2}{2+\gamma x^\beta} - \frac{1}{1+\gamma x^\beta}\right) dx \approx d^{1-\beta} =: m(d).$$

Also, since $\mathrm{Var}(y_j^2) = 3\gamma^2$,

$$\mathrm{Var}\log(\rho) \geq \sum_{j=1}^{d}\left(\frac{2}{2+\gamma j^\beta} - \frac{1}{1+\gamma j^\beta}\right)^2\gamma^2$$
$$\approx \int_1^d \left(\frac{2}{2+\gamma x^\beta} - \frac{1}{1+\gamma x^\beta}\right)^2 dx \approx d^{1-2\beta} =: c(d).$$

Thus we have

$$\begin{aligned}\mathbb{P}\Big(\log(\rho) \geq m(d)/2\Big) &\geq \mathbb{P}\Big(S(y,d) \geq m(d)/2\Big)\\ &\geq \mathbb{P}\Big(S(y,d) \geq \mathbb{E}S(y,d)/2\Big)\\ &\geq \mathbb{P}\Big(|S(y,d) - \mathbb{E}S(y,d)| \leq \mathbb{E}S(y,d)/2\Big)\\ &= 1 - \mathbb{P}\Big(|S(y,d) - \mathbb{E}S(y,d)| \geq \mathbb{E}S(y,d)/2\Big)\\ &\geq 1 - \mathbb{P}\Big(|S(y,d) - \mathbb{E}S(y,d)| \geq m(d)/2\Big)\\ &\geq 1 - 4\frac{c(d)}{m(d)^2} \to 1.\end{aligned}$$

□

6.4. Proofs Section 4

The following lemma will be used in the proof of Theorem 4.5. It justifies the use of the cyclic property in calculating certain traces in the infinite dimensional setting.

Lemma 6.8. *Suppose that $A = S^*S$, where $S = \Gamma^{-1/2}K\Sigma^{1/2}$ as in Assumption 3.3 is bounded. Then*

$$\tau = \mathrm{Tr}(A) = \mathrm{Tr}(\Gamma^{-1}K\Sigma K^*).$$

Therefore, using the equivalence in Table 2 we have that τ_{st} and τ_{op} admit the following equivalent expressions:

$$\tau_{st} = \mathrm{Tr}\big(R^{-1}H(MPM^* + Q)H^*\big) \tag{6.7}$$

and

$$\tau_{op} = \mathrm{Tr}\big((R + HQH^*)^{-1}HMPM^*H^*\big). \tag{6.8}$$

Proof. Using Lemma 6.5(iv) we have that $\tau = \mathrm{Tr}(S^*S) = \mathrm{Tr}(SS^*)$. Now note that $SS^* = \Gamma^{-1/2}K\Sigma K^*\Gamma^{-1/2}$ is bounded since A is, and that $\Gamma^{1/2}$ is also bounded, hence we can use Lemma 6.5(v) to get the desired result. □

Proof of Theorem 4.5. Using the previous lemma,

$$\begin{aligned}\tau_{st} &= \mathrm{Tr}\Big(R^{-1}HMPM^*H^*\Big) + \mathrm{Tr}\Big(R^{-1}HQH^*\Big)\\ &\geq \mathrm{Tr}\Big(R^{-1}HMPM^*H^*\Big)\\ &\geq \mathrm{Tr}\Big((R + HQH^*)^{-1}HMPM^*H^*\Big) = \tau_{op},\end{aligned}$$

where the first inequality holds because R is positive-definite and HQH^* is positive semi definite, and the second one follows from Lemma 6.5(vi).

If $\mathrm{Tr}(HQH^*R^{-1}) < \infty$ then there is $c > 0$ such that, for all x, $\|HQH^*x\| \leq c\|Rx\|$. Hence applying again Lemma 6.5(vi) for both directions of the equivalence, we obtain that

$$\tau_{op} = \mathrm{Tr}\Big((R + HQH^*)^{-1}HMPM^*H^*\Big) < \infty \iff \mathrm{Tr}\Big(R^{-1}HMPM^*H^*\Big) < \infty$$
$$\iff \tau_{st} < \infty.$$

□

References

[1] A. Doucet, and S. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3):197–208, 2000.

[2] K. Achutegui, D. Crisan, J. Miguez, and G. Rios. A simple scheme for the parallelization of particle filters and its application to the tracking of complex stochastic systems. *arXiv preprint arXiv:1407.8071*, 2014.

[3] R. J. Adler. An introduction to continuity, extrema, and related topics for general Gaussian processes. *Lecture Notes-Monograph Series*, 1990.

[4] S. Agapiou, S. Larsson, and A. M. Stuart. Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Processes and their Applications*, 123(10):3828–3860, 2013.

[5] S. Agapiou and P. Mathé. Preconditioning the prior to overcome saturation in Bayesian inverse problems. *arXiv preprint arXiv:1409.6496*, 2014.

[6] S. Agapiou, A. M. Stuart, and Y-X Zhang. Bayesian posterior contraction rates for linear severely ill-posed inverse problems. *Journal of Inverse and Ill-posed Problems*, 22(3):297–321, 2014.

[7] C. Apostol, D. A. Herrero, and D. Voiculescu. The closure of the similarity orbit of a Hilbert space operator. *Bulletin of the American Mathematical Society*, 6(3):421–426, 1982.

[8] A. Bain and D. Crisan. *Fundamentals of stochastic filtering*, volume 3. Springer, 2009.

[9] C. M. Bender and S. A. Orszag. *Advanced Mathematical Methods for Scientists and Engineers I*. Springer Science & Business Media, 1999.

[10] T. Bengtsson, P. Bickel, B. Li, et al. 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*, pages 316–334. Institute of Mathematical Statistics, 2008.

[11] A. Beskos, D. Crisan, A. Jasra, et al. On the stability of sequential Monte Carlo methods in high dimensions. *The Annals of Applied Probability*, 24(4):1396–1445, 2014.

[12] A. Beskos, D. Crisan, A. Jasra, K. Kamatani, and Y. Zhou. A stable particle filter in high-dimensions. *arXiv preprint arXiv:1412.3501*, 2014.

[13] P. Bickel, B. Li, T. Bengtsson, et al. 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, pages 318–329. Institute of Mathematical Statistics, 2008.

[14] C. M. Bishop. *Pattern recognition and machine learning*. Springer New York, 2006.

[15] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013.

[16] T. Bui-Thanh, O. Ghattas, J. Martin, and G. Stadler. A computational framework for infinite-dimensional Bayesian inverse problems part i: The linearized case, with application to global seismic inversion. *SIAM Journal on Scientific Computing*, 35(6):A2494–A2523, 2013.

[17] R. E. Caflisch, W. J. Morokoff, and A. B. Owen. *Valuation of mortgage backed securities using Brownian bridges to reduce effective dimension*. Department of Mathematics, University of California, Los Angeles, 1997.

[18] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[19] S. Chatterjee and P. Diaconis. The sample size required in importance sampling. *arXiv preprint arXiv:1511.01437*, 2015.

[20] N. Chopin. Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Annals of statistics*, pages 2385–2411, 2004.

[21] N. Chopin and O. Papaspiliopoulos. *A concise introduction to sequential Monte Carlo*. 2016.

[22] A. J. Chorin and M. Morzfeld. Conditions for successful data assimilation. *Journal of Geophysical Research: Atmospheres*, 118(20):11–522, 2013.

[23] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28(3):424–446, 2013.

[24] D. Crisan, P. Del Moral, and T. Lyons. *Discrete filtering using branching and interacting particle systems*. Citeseer, 1998.

[25] D. Crisan and A. Doucet. A survey of convergence results on particle filtering methods for practitioners. *Signal Processing, IEEE Transactions on*, 50(3):736–746, 2002.

[26] D. Crisan and B. Rozovskii. *The Oxford Handbook of Nonlinear Filtering*. Oxford University Press, 2011.

[27] G. Da Prato and J. Zabczyk. *Stochastic equations in infinite dimensions*. Cambridge university press, 1992.

[28] M. Dashti and A. M. Stuart. The Bayesian approach to inverse problems. *Handbook of Uncertainty Quantification, Editors R. Ghanem, D. Higdon and H. Owhadi*. `http://arxiv.org/abs/1302.6989`, 2016.

[29] P. Del Moral. *Feynman-Kac Formulae*. Springer, 2004.

[30] P. Del Moral and L. Miclo. *Branching and interacting particle systems approximations of Feynman-Kac formulae with applications to non-linear filtering*. Springer, 2000.

[31] P. Doukhan and G. Lang. Evaluation for moments of a ratio with application to regression estimation. *Bernoulli*, 15(4):1259–1286, 2009.

[32] P. J. Downey and P. E. Wright. The ratio of the extreme to the sum in a random sequence. *Extremes*, 10(4):249–266, 2007.
[33] P. Dupuis, K. Spiliopoulos, and H. Wang. Importance sampling for multiscale diffusions. *Multiscale Modeling & Simulation*, 10(1):1–27, 2012.
[34] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.
[35] J. N. Franklin. Well-posed stochastic extensions of ill-posed linear problems. *Journal of Mathematical Analysis and Applications*, 31(3):682–716, 1970.
[36] M. Frei and H. R. Künsch. Bridging the ensemble kalman and particle filters. *Biometrika*, 100(4):781–800, 2013.
[37] R. Furrer and T. Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255, 2007.
[38] A. Gelman, G. O. Roberts, and W. Gilks. Efficient metropolis jumping hules. *Bayesian statistics*, 5(599-608):42, 1996.
[39] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.
[40] J. Goodman, K. K. Lin, and M. Morzfeld. Small-noise analysis and symmetrization of implicit Monte Carlo samplers. *arXiv preprint arXiv:1410.6151*, 2014.
[41] M. Hairer, A. M. Stuart, J. Voss, et al. Analysis of spdes arising in path sampling part ii: the nonlinear case. *The Annals of Applied Probability*, 17(5/6):1657–1706, 2007.
[42] W. Han. *On the Numerical Solution of the Filtering Problem*. PhD thesis, Ph. D. Thesis. Department of Mathematics, Imperial College London, 2013.
[43] A.M. Johansen and A. Doucet. A note on auxiliary particle filters. *Statist. Probab. Lett.*, 78(12):1498–1504, 2008.
[44] H. Kahn. *Use of different Monte Carlo sampling techniques*. Rand Corporation, 1955.
[45] H. Kahn and A. W. Marshall. Methods of reducing sample size in Monte Carlo computations. *Journal of the Operations Research Society of America*, 1(5):263–278, 1953.
[46] J. P. Kaipio and E. Somersalo. *Statistical and computational inverse problems*, volume 160. Springer, 2005.
[47] O. Kallenberg. *Foundations of Modern Probability*. Probability and Its Applications. Springer Science, 2nd edition edition, 2002.
[48] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45, 1960.
[49] E. Kalnay. *Atmospheric Modeling, Data Assimilation, and Predictability*. Cambridge university press, 2003.
[50] N. Kantas, A. Beskos, and A. Jasra. Sequential monte carlo methods for high-dimensional inverse problems: a case study for the navier-stokes equations. *arXiv:1307.6127*.
[51] H. Kekkonen, M. Lassas, and Siltanen S. Posterior consistency and con-

vergence rates for bayesian inversion with hypoelliptic operators. *arXiv preprint arXiv:1507.01772*, 2015.

[52] B. T. Knapik, A. W. van Der Vaart, and J. H. van Zanten. Bayesian inverse problems with Gaussian priors. *The Annals of Statistics*, 39(5):2626–2657, 2011.

[53] B. T. Knapik, A. W. van der Vaart, and J. H. van Zanten. Bayesian recovery of the initial condition for the heat equation. *Communications in Statistics-Theory and Methods*, 42(7):1294–1313, 2013.

[54] A. Kong. A note on importance sampling using standardized weights. *University of Chicago, Dept. of Statistics, Tech. Rep*, 348, 1992.

[55] A. Kong, J. S. Liu, and W. H. Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American statistical association*, 89(425):278–288, 1994.

[56] F. Y. Kuo and I. H. Sloan. Lifting the curse of dimensionality. *Notices of the AMS*, 52(11):1320–1328, 2005.

[57] P. Lancaster and L. Rodman. *Algebraic Riccati Equations*. Oxford University Press, 1995.

[58] S. Lasanen. Measurements and infinite-dimensional statistical inverse theory. *PAMM*, 7(1):1080101–1080102, 2007.

[59] S. Lasanen. Non-Gaussian statistical inverse problems. Part I: Posterior distributions. *Inverse Problems and Imaging*, 6(2):215–266, 2012.

[60] S. Lasanen. Non-Gaussian statistical inverse problems. Part II: Posterior convergence for approximated unknowns. *Inverse Problems and Imaging*, 6(2):267–287, 2012.

[61] P. D. Lax. *Functional analysis*. Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York, 2002.

[62] M. S. Lehtinen, L. Paivarinta, and E. Somersalo. Linear inverse problems for generalised random variables. *Inverse Problems*, 5(4):599, 1989.

[63] K. Lin, S. Lu, and P. Mathé. Oracle-type posterior contraction rates in Bayesian inverse problems. *Inverse Problems & Imaging*, 9(3), 2015.

[64] D. V. Lindley and A. F. M. Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–41, 1972.

[65] J. S. Liu. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, 6(2):113–119, 1996.

[66] S. Lu and P. Mathé. Discrepancy based model selection in statistical inverse problems. *Journal of Complexity*, 30(3):290–308, 2014.

[67] A. Mandelbaum. Linear estimators and measurable linear transformations on a Hilbert space. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65(3):385–397, 1984.

[68] J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3):A1460–A1487, 2012.

[69] D. L. McLeish and G. L. O'Brien. The expected ratio of the sum of squares

to the square of the sum. *Ann. Probab.*, 10(4):1019–1028, 1982.

[70] J. Míguez, D. Crisan, and P. M. Djurić. On the convergence of two sequential Monte Carlo methods for maximum a posteriori sequence estimation and stochastic global optimization. *Statistics and Computing*, 23(1):91–107, 2013.

[71] B. Moskowitz and R. E. Caflisch. Smoothness and dimension reduction in quasi-Monte Carlo methods. *Mathematical and Computer Modelling*, 23(8):37–54, 1996.

[72] Jennifer L Mueller and Samuli Siltanen. *Linear and nonlinear inverse problems with practical applications*, volume 10. Siam, 2012.

[73] D. S. Oliver, A. C. Reynolds, and N. Liu. *Inverse Theory for Petroleum Reservoir Characterization and History Matching*. Cambridge University Press, 2008.

[74] M. K. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American statistical association*, 94(446):590–599, 1999.

[75] K. Ray. Bayesian inverse problems with non-conjugate priors. *Electronic Journal of Statistics*, 7:2516–2549, 2013.

[76] P. Rebeschini and R. van Handel. Can local particle filters beat the curse of dimensionality? *arXiv preprint arXiv:1301.6585*, 2013.

[77] M. Reed and B. Simon. *Analysis of Operators, Vol. IV of Methods of Modern Mathematical Physics*. New York, Academic Press, 1978.

[78] Y-F Ren and H-Y Liang. On the best constant in Marcinkiewicz–Zygmund inequality. *Statistics and probability letters*, 53(3):227–233, 2001.

[79] L. Slivinski and C. Snyder. Practical estimates of the ensemble size necessary for particle filters.

[80] C. Snyder. Particle filters, the "optimal" proposal and high-dimensional systems. In *Proceedings of the ECMWF Seminar on Data Assimilation for Atmosphere and Ocean*, 2011.

[81] C. Snyder and T. Bengtsson. Performance bounds for particle filters using the optimal proposal.

[82] C. Snyder, T. Bengtsson, P. Bickel, and J. Anderson. Obstacles to high-dimensional particle filtering. *Monthly Weather Review*, 136(12):4629–4640, 2008.

[83] A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio, and Y. Marzouk. Optimal low-rank approximations of Bayesian linear inverse problems. *arXiv preprint arXiv:1407.3463*, 2014.

[84] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(4):583–639, 2002.

[85] K. Spiliopoulos. Large deviations and importance sampling for systems of slow-fast motion. *Applied Mathematics & Optimization*, 67(1):123–161, 2013.

[86] A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

[87] L. Tierney. A note on Metropolis-Hastings kernels for general state spaces.

Ann. Appl. Probab., 8(1):1–9, 1998.

[88] X. Tu, M. Morzfeld, J. Wilkening, and A. J. Chorin. Implicit sampling for an elliptic inverse problem in underground hydrodynamics. *arXiv preprint arXiv:1308.4640*, 2013.

[89] P. J. van Leeuwen. Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Quarterly Journal of the Royal Meteorological Society*, 136(653):1991–1999, 2010.

[90] E. Vanden-Eijnden and J. Weare. Data assimilation in the low noise, accurate observation regime with application to the Kuroshio current. *Monthly Weather Review*, 141(arXiv: 1202.4952):1, 2012.

[91] S. J. Vollmer. Posterior consistency for Bayesian inverse problems through stability and regression results. *Inverse Problems*, 29(12):125011, 2013.

[92] T. Zhang. Effective dimension and generalization of kernel learning. In *Advances in Neural Information Processing Systems*, pages 454–461, 2002.

[93] T. Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005.

[94] W. Zhang, C. Hartmann, M. Weber, and C. Schütte. Importance sampling in path space for diffusion processes. *Multiscale Model. Sim*, 2013.