Certified dimension reduction in nonlinear Bayesian inverse problems

Olivier Zahm, Tiangang Cui, Kody Law, Alessio Spantini, Youssef Marzouk

▶ **To cite this version:**

Olivier Zahm, Tiangang Cui, Kody Law, Alessio Spantini, Youssef Marzouk. Certified dimension reduction in nonlinear Bayesian inverse problems. Mathematics of Computation, 2022, 91, pp.1789-1835. 10.1090/mcom/3737 . hal-01834039v2

HAL Id: hal-01834039
https://hal.science/hal-01834039v2

Submitted on 9 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Certified dimension reduction in nonlinear Bayesian inverse problems

Olivier Zahm[*], Tiangang Cui[†], Kody Law[‡], Alessio Spantini[§], Youssef Marzouk[§]

March 9, 2021

Abstract

We propose a dimension reduction technique for Bayesian inverse problems with nonlinear forward operators, non-Gaussian priors, and non-Gaussian observation noise. The likelihood function is approximated by a ridge function, i.e., a map which depends non-trivially only on a few linear combinations of the parameters. We build this ridge approximation by minimizing an upper bound on the Kullback–Leibler divergence between the posterior distribution and its approximation. This bound, obtained via logarithmic Sobolev inequalities, allows one to certify the error of the posterior approximation. Computing the bound requires computing the second moment matrix of the gradient of the log-likelihood function. In practice, a sample-based approximation of the upper bound is then required. We provide an analysis that enables control of the posterior approximation error due to this sampling. Numerical and theoretical comparisons with existing methods illustrate the benefits of the proposed methodology.

Keywords dimension reduction, nonlinear Bayesian inverse problem, logarithmic Sobolev inequality, certified error bound, non-asymptotic analysis.

1 Introduction

Solving Bayesian inverse problems [24, 43] is a challenging task in many domains of application, due to the complexity of the posterior distribution. One of the primary sources of complexity is the dimension of the parameters to be inferred, which is often high or in principle infinite—e.g., when the posterior is a distribution over functions or their finite-dimensional discretization. High dimensionality presents difficulties for posterior sampling: care is required to design sampling algorithms that mix effectively while remaining robust under refinement of the discretization. High

[*]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France (Corresponding author olivier.zahm@inria.fr)

[†]School of Mathematical Sciences, Monash University, Victoria 3800, Australia.

[‡]School of Mathematics, University of Manchester, Manchester, M13 9PL, UK.

[§]Department of Aeronautics and Astronautics, MIT, Cambridge, MA 02139, USA.

dimensionality also raises significant hurdles to the use of model reduction or approximation schemes (e.g., [9, 13, 31, 40]) that attempt to reduce the cost of likelihood or forward model evaluations.

Successful strategies for high-dimensional Bayesian inverse problems typically exploit the presence of some low-dimensional structure. A common structure in inverse problems is that the posterior is a low-dimensional *update* of the prior, in the sense that change from prior to posterior is most prominent on a low-dimensional subspace of the parameter space. Put another way, the likelihood is influential, relative to the prior, only on a low-dimensional subspace (we will make these intuitive notions more precise later). Sources of such structure include the smoothing properties of the forward operator and limitations on the number or accuracy of the observations. In the linear–Gaussian case, this structure is already well understood. For instance, low-rank approximations of the (prior-preconditioned) Hessian of the log-likelihood have been used in [16] to approximate the posterior covariance. In [42], this approach is shown to yield *optimal* approximations of the posterior covariance and of the posterior mean. A heuristic extension of this approach to nonlinear forward models, known as the *likelihood-informed subspace* (LIS) method, is proposed in [12] and shown to perform well in many applications [4, 25]. A similar idea underlies the *active subspace* method applied to Bayesian problems [11]. Identifying the subspace on which changes from the prior to posterior are most prominent also provides a foundation for infinite-dimensional MCMC algorithms that "split" the parameter space, such as the *dimension-independent likelihood-informed* (DILI) MCMC samplers of [14]. In [15], LIS-based dimension reduction is shown to be an effective prelude to model order reduction, where the forward model is replaced with a computationally inexpensive approximation. Yet all of these dimension reduction approaches, outside of the linear–Gaussian setting, are essentially heuristics. Optimal low-dimensional approximations—or approximations endowed with *certified error bounds*—have not yet been developed for the nonlinear/non-Gaussian setting. Such approximations are the subject of this paper.

Other forms of dimension reduction for Bayesian inverse problems have also been proposed, besides the "update" form of dimension reduction described above. For instance, in [28, 32], a truncated Karhunen-Loève decomposition of the prior distribution is used to reduce the parameter dimension of the entire inverse problem. This approach exploits only the low-dimensional structure of the prior distribution, however, and does not take advantage of structure in the likelihood function or forward model. In [29], a greedy algorithm is used to identify a parameter subspace capable of reproducing the forward model. This approach, in contrast, does not take advantage of the prior correlation structure, and moreover can remove directions that are uninformed by the likelihood but that still retain large variation under the true posterior. More recent results on the intrinsic dimension of linear Bayesian inverse problems [1] reinforce the idea that the *update* from prior to posterior should instead be a central object, e.g., when characterizing the performance of importance sampling schemes.

1.1 Contribution

In this paper we propose a methodology to detect and exploit the low-dimensional structure of the update from prior to posterior. Our approach addresses Bayesian inverse problems with nonlinear forward operators, non-Gaussian priors, and non-Gaussian observation noise: the posterior ν is a measure on $\mathbb{R}^d$ given by $\mathrm{d}\nu \propto f\mathrm{d}\mu$ where μ is the prior measure and f is a any unnormalized likelihood function defined on $\mathbb{R}^d$. Our approach consists in approximating $x \mapsto f(x)$ by a *ridge function*, i.e., a function of the form $x \mapsto h(Ax)$ which depends non-trivially only on $r \ll d$ linear combinations of the parameters Ax, with $A \in \mathbb{R}^{r\times d}$.

More precisely, we seek a *controlled approximation* such that the Kullback–Leibler (KL) divergence $\mathrm{D}_{\mathrm{KL}}(\nu||\nu_r)$ from the resulting posterior approximation $\mathrm{d}\nu_r(x) \propto h(Ax)\mathrm{d}\mu(x)$ to the posterior

ν is below some user-defined threshold. As our main contribution, we derive in Theorem 2.9 and Corollary 2.10 an upper bound on the KL divergence. Minimizing this bound permits us to construct the ridge approximation so that the bound falls below the threshold.

This bound is obtained via *logarithmic Sobolev inequalities*, an important class of inequalities in measure theory [19, 20, 34] with many implications on concentration of measure phenomena [7, 26]. Using logarithmic Sobolev inequalities requires some assumptions on the prior distribution and on the regularity of the likelihood function. In particular, we need the gradient of the log-likelihood to be square-integrable over the posterior distribution. A similar methodology has been proposed in [48] to reduce the input dimension of multivariate functions. In that paper, a bound on the function approximation error in L^2 norm is obtained via Poincaré inequalities, another class of Sobolev inequality.

In the proposed dimension reduction method, the informed subspace is constructed as the dominant eigenspace of the matrix

$$H = \int (\nabla \log f)(\nabla \log f)^T \mathrm{d}\nu.$$

Once the informed subspace is identified, the posterior approximation is constructed by computing a conditional expectation of the likelihood function. We prove that the resulting divergence $\mathrm{D}_{\mathrm{KL}}(\nu||\nu_r)$ is bounded by the sum of the $d-r$ smallest eigenvalues of H: a quickly decaying spectrum in H thus reveals the low-dimensional structure in the posterior.

The proposed method requires computing (i) the matrix H and (ii) a conditional expectation of the likelihood function. In practice, we approximate both quantities with Monte Carlo estimates, requiring samples from the posterior for (i) and samples from the prior for (ii). Provided the sample sizes are sufficiently large, we prove that the resulting random posterior approximation is a quasi-optimal approximation compared to the theoretical posterior approximation obtained by exact integration. This quasi-optimality result is given in expectation for (ii) and in high probability for (i). In particular, we show that the number of posterior samples to approximate (i) should scale in proportion to the *rank* of the matrix to be estimated, which can be much smaller than the ambient dimension. Finally, even if the method only requires a limited number of posterior samples, it can be difficult to obtain the posterior samples in (i). We thus propose several alternatives for computing (i) that do not require sampling from the exact posterior. These alternatives include an iterative procedure that builds a sequence of low-dimensional posterior approximations in order to obtain an accurate final estimate.

The outline of the paper is as follows. In Section 2 we introduce the theoretical tools to derive and to minimize the error bound. We demonstrate the benefits of this method on an analytical example. In Section 3 we propose algorithms for the numerical construction of the low-dimensional posterior approximation. We give a theoretical analysis of the convergence of these sample-based estimators. In order to provide some context for our developments, we show in Section 4 how the proposed methodology compares with existing dimension reduction methods, including a comparison with a Karhunen-Loève decomposition. Finally, Section 5 illustrates the benefits of our approach on two numerical examples.

2 Dimension reduction for the approximation of high-dimensional distributions

Let μ be a probability distribution defined over the Borel sets $\mathcal{B}(\mathbb{R}^d)$ of $\mathbb{R}^d$. Given a measurable function $f : \mathbb{R}^d \to \mathbb{R}^+$ such that $\int f \, \mathrm{d}\mu < \infty$, let ν be the probability distribution such that

$$\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \propto f.$$

In the context of Bayesian inverse problems, one can view μ as the *prior distribution* and ν as the *posterior distribution*, while f represents, up to a multiplicative constant, the *likelihood function*. We consider the problem of approximating the posterior ν by a probability distribution ν_r such that

$$\frac{\mathrm{d}\nu_r}{\mathrm{d}\mu} \propto g \circ P_r, \tag{1}$$

where $P_r : \mathbb{R}^d \to \mathbb{R}^d$ is a linear projector with rank r and $g : \mathbb{R}^d \to \mathbb{R}^+$ is a Borel function called the *profile function*. Throughout this paper we identify the projector with its matrix representation $P_r \in \mathbb{R}^{d\times d}$. Notice that P_r is not restricted to be orthogonal: it can be any matrix $P_r \in \mathbb{R}^{d\times d}$ which satisfies $P_r^2 = P_r$ and $\mathrm{rank}(P_r) = r$, but not necessarily $P_r^T = P_r$. Any vector $x \in \mathbb{R}^d$ can be uniquely decomposed as

$$x = x_r + x_\perp \quad \text{with} \begin{cases} x_r = P_r x \\ x_\perp = (I_d - P_r)x \end{cases}$$

where $I_d \in \mathbb{R}^{d\times d}$ denotes the identity matrix. If $r \ll d$, the approximation of ν by ν_r consists essentially in replacing the high-dimensional likelihood f by a function of fewer variables. Indeed, $x \mapsto g(P_r x) = g(x_r)$ depends only on the variable $x_r \in \mathrm{Im}(P_r) \cong \mathbb{R}^r$ and is constant along $\mathrm{Ker}(P_r) \cong \mathbb{R}^{d-r}$. Yet ν_r cannot itself be considered a low-dimensional distribution, since its support could be the same as that of ν.

We use the Kullback–Leibler divergence $\mathrm{D_{KL}}(\cdot \,||\, \cdot)$ to measure the dissimilarity between probability distributions. It is defined as

$$\mathrm{D_{KL}}(\nu_1||\nu_2) = \int \log\left(\frac{\mathrm{d}\nu_1}{\mathrm{d}\nu_2}\right) \mathrm{d}\nu_1, \tag{2}$$

for any probability distributions ν_1 and ν_2 such that ν_1 is absolutely continuous with respect to ν_2, and $\mathrm{D_{KL}}(\nu_1||\nu_2) = \infty$ otherwise. Given a prescribed tolerance $\varepsilon \geq 0$, our goal is to build an approximation ν_r of ν of the form (1) such that

$$\mathrm{D_{KL}}(\nu||\nu_r) \leq \varepsilon. \tag{3}$$

Of course if P_r is the identity matrix and $g = f$, the distribution ν_r is exactly the posterior ν, so that (3) is trivially satisfied. In that case, the rank of $P_r = I_d$ is d and there is no dimension reduction. The goal of this section is to give sufficient conditions under which we can build an approximation ν_r such that (3) holds with a rank $r = r(\varepsilon)$ that is significantly smaller than d.

Remark 2.1. Instead of approximating ν by ν_r as given by (1), we could have considered an approximation $\widetilde{\nu}_r$ of the form

$$\frac{\mathrm{d}\widetilde{\nu}_r}{\mathrm{d}\mu} \propto h \circ A_r,$$

with $h : \mathbb{R}^r \to \mathbb{R}^+$ a Borel function and $A_r \in \mathbb{R}^{r\times d}$ a matrix of full row rank. Functions of the form $x \mapsto h(A_r x)$ are a particular kind of *ridge function* [36]. They are used in various domains to approximate multivariate functions; see [8, 17, 46], for example. However, any function of the form $h \circ A_r$ can be written as $g \circ P_r$ and vice-versa, where $g : \mathbb{R}^d \to \mathbb{R}^+$ is some Borel function and P_r is some rank-r projector. More precisely, for any $r \leq d$ we can show that

$$\left\{ h \circ A_r \,\middle|\, \begin{array}{l} h : \mathbb{R}^r \to \mathbb{R}^+, \text{Borel function} \\ A_r \in \mathbb{R}^{r\times d}, \text{rank-}r \text{ matrix} \end{array} \right\} = \left\{ g \circ P_r \,\middle|\, \begin{array}{l} g : \mathbb{R}^d \to \mathbb{R}^+, \text{Borel function} \\ P_r \in \mathbb{R}^{d\times d}, \text{rank-}r \text{ projector} \end{array} \right\}.$$

This means that approximating ν by ν_r or by $\widetilde{\nu}_r$ are the essentially the same problem: we do not gain approximation power in using $\widetilde{\nu}_r$ instead of ν_r.

2.1 Optimal profile function

We begin by characterizing the *optimal* profile function g^*, defined as a minimizer of $g \mapsto \mathrm{D_{KL}}(\nu||\nu_r)$ over the set of positive and measurable functions, where $\frac{\mathrm{d}\nu_r}{\mathrm{d}\mu} \propto g \circ P_r$. In this section, we assume that the projector P_r is given (fixed). We will address the problem of constructing P_r later in Section 2.2.

Let us denote by $\sigma(P_r)$ the σ-algebra generated by P_r. It is defined by $\sigma(P_r) = \{P_r^{-1}B \mid B \in \mathcal{B}(\mathbb{R}^d)\}$, where $P_r^{-1}B = \{x \in \mathbb{R}^d : P_r x \in B\}$ denotes the pre-image of $B \in \mathcal{B}(\mathbb{R}^d)$ under P_r. The following lemma corresponds to the Doob–Dynkin lemma; see, for example, Lemma 1.13 in [33]. It states that, for any projector P_r, the set of all functions of the form $g \circ P_r$ for some measurable function g is exactly the set of all $\sigma(P_r)$-measurable functions.

Lemma 2.2. Let $P_r \in \mathbb{R}^{d\times d}$ be a projector. Given a measurable function g defined on $\mathbb{R}^d$, the function $h = g \circ P_r$ is $\sigma(P_r)$-measurable. Conversely, given a $\sigma(P_r)$-measurable function h, there exists a measurable function g defined on $\mathbb{R}^d$ such that $h = g \circ P_r$.

We denote by $\mathbb{E}_\mu(f|\sigma(P_r)) : \mathbb{R}^d \to \mathbb{R}$ the conditional expectation of f given $\sigma(P_r)$ under the distribution μ. It is the unique $\sigma(P_r)$-measurable function that satisfies

$$\int \mathbb{E}_\mu(f|\sigma(P_r))\, h \,\mathrm{d}\mu = \int f\, h \,\mathrm{d}\mu, \tag{4}$$

for any $\sigma(P_r)$-measurable function $h : \mathbb{R}^d \to \mathbb{R}$. We consider the probability distribution ν_r^* defined by

$$\frac{\mathrm{d}\nu_r^*}{\mathrm{d}\mu} \propto \mathbb{E}_\mu(f|\sigma(P_r)).$$

This distribution is well defined because $\mathbb{E}_\mu(f|\sigma(P_r))$ is a positive function (the conditional expectation preserves the sign) and because $\int \mathbb{E}_\mu(f|\sigma(P_r))\,\mathrm{d}\mu = \int f \,\mathrm{d}\mu < \infty$ (by letting $h = 1$ in (4)). By definition of the Kullback–Leibler divergence (2), we can write

$$\begin{aligned} \mathrm{D_{KL}}(\nu||\nu_r) - \mathrm{D_{KL}}(\nu||\nu_r^*) &= \int \frac{f}{Z_1} \log \frac{(Z_1)^{-1}\mathbb{E}_\mu(f|\sigma(P_r))}{(Z_2)^{-1} g \circ P_r} \,\mathrm{d}\mu & (5)\\ &= \int \frac{\mathbb{E}_\mu(f|\sigma(P_r))}{Z_1} \log \frac{(Z_1)^{-1}\mathbb{E}_\mu(f|\sigma(P_r))}{(Z_2)^{-1} g \circ P_r} \,\mathrm{d}\mu & (6)\\ &= \mathrm{D_{KL}}(\nu_r^*||\nu_r) & (7) \end{aligned}$$

where Z_1 and Z_2 are two normalizing constants defined by $Z_1 = \int f \, \mathrm{d}\mu = \int \mathbb{E}_\mu(f|\sigma(P_r))\mathrm{d}\mu$ and $Z_2 = \int g \circ P_r \, \mathrm{d}\mu$. To go from (5) to (6), we used relation (4) with $h = \log \frac{(Z_1)^{-1}\mathbb{E}_\mu(f|\sigma(P_r))}{(Z_2)^{-1} g\circ P_r}$, which is a $\sigma(P_r)$-measurable function. By (7) we have

$$\mathrm{D_{KL}}(\nu||\nu_r) \geq \mathrm{D_{KL}}(\nu||\nu_r^*), \quad \text{with} \quad \begin{cases} \frac{\mathrm{d}\nu_r}{\mathrm{d}\mu} \propto g \circ P_r, \\ \frac{\mathrm{d}\nu_r^*}{\mathrm{d}\mu} \propto \mathbb{E}_\mu(f|\sigma(P_r)), \end{cases}$$

for any measurable function g. From this inequality we deduce that any function g^* which satisfies $g^* \circ P_r = \mathbb{E}_\mu(f|\sigma(P_r))$ is a minimizer of $g \mapsto \mathrm{D_{KL}}(\nu||\nu_r)$. By Lemma 2.2 such a function g^* exists because $\mathbb{E}_\mu(f|\sigma(P_r))$ is a $\sigma(P_r)$-measurable function.

Remark 2.3. The conditional expectation $\mathbb{E}_\mu(f|\sigma(P_r))$ is known to be the best approximation of a function f with respect to the L^2_μ-norm, meaning that it minimizes $h \mapsto \int (f-h)^2 \mathrm{d}\mu$ among all $\sigma(P_r)$-measurable functions h. Here we showed that it is also optimal with respect to the Kullback–Leibler divergence. As pointed out in [3], $\mathbb{E}_\mu(f|\sigma(P_r))$ is also an optimal approximation of f with respect to the class of *expected Bregman divergences*, which includes the Kullback–Leibler divergence and the L^2_μ-norm distance.

The following proposition recalls the well-known analytical expression for the conditional expectation. For the sake of completeness, the proof is given in Section A.

Proposition 2.4 (Explicit expression for the conditional expectation)**.** Let μ be a probability distribution on $\mathbb{R}^d$ which admits a Lebesgue density ρ. Given a rank-r projector P_r, we denote by $U_\perp \in \mathbb{R}^{d\times(d-r)}$ a matrix whose columns form a basis of $\mathrm{Ker}(P_r)$. Let $p_\perp(\,\cdot\,|P_r x)$ be the conditional probability density on $\mathbb{R}^{d-r}$ defined by

$$p_\perp(\xi_\perp|P_r x) = \frac{\rho(P_r x + U_\perp \xi_\perp)}{\int_{\mathbb{R}^{d-r}} \rho(P_r x + U_\perp \xi'_\perp)\mathrm{d}\xi'_\perp}, \tag{8}$$

for all $\xi_\perp \in \mathbb{R}^{d-r}$ and any $x \in \mathbb{R}^d$, with the convention $p_\perp(\xi_\perp|P_r x) = 0$ whenever the denominator of (8) is zero. Then, for any measurable function f, the conditional expectation $\mathbb{E}_\mu(f|\sigma(P_r))$ can be written as

$$\mathbb{E}_\mu(f|\sigma(P_r)) : x \mapsto \int_{\mathbb{R}^{d-r}} f(P_r x + U_\perp \xi_\perp)\; p_\perp(\xi_\perp|P_r x)\, \mathrm{d}\xi_\perp. \tag{9}$$

We conclude this section with a remarkable property of the conditional expectation $\mathbb{E}_\mu(f|\sigma(P_r))$. Let Q_r be any projector whose kernel is the same as that of P_r. Then the relations $P_r = P_r \circ Q_r$ and $Q_r = Q_r \circ P_r$ hold; see, for instance, the proof of Proposition 2.2 in [48]. Then Lemma 2.2 ensures that any $\sigma(P_r)$-measurable function is $\sigma(Q_r)$-measurable and vice-versa. In other words, being $\sigma(P_r)$-measurable or $\sigma(Q_r)$-measurable are equivalent such that, by definition of the conditional expectation (4), we have

$$\mathrm{Ker}(P_r) = \mathrm{Ker}(Q_r) \quad \Rightarrow \quad \mathbb{E}_\mu(f|\sigma(P_r)) = \mathbb{E}_\mu(f|\sigma(Q_r)).$$

In other words, the conditional expectation $\mathbb{E}_\mu(f|\sigma(P_r))$ is invariant with respect to the image of P_r—so that the error $\mathrm{D_{KL}}(\nu||\nu_r^*)$, seen as a function of the projector P_r, is actually only a function of $\mathrm{Ker}(P_r)$. In the context of Bayesian inference, this property means that when the optimal profile function g^* is used in (1), the important feature of P_r to be discovered is its kernel, and not its image. Thus reducing the dimension of a Bayesian inverse problem consists in identifying the subspace $\mathrm{Ker}(P_r)$ on which the data are *not* informative.

2.2 A controlled approximation

In this section we show how to build a projector P_r with a sufficiently large rank so that ν_r^*, defined by $\frac{\mathrm{d}\nu_r^*}{\mathrm{d}\mu} \propto \mathbb{E}_\mu(f|\sigma(P_r))$, satisfies $\mathrm{D_{KL}}(\nu||\nu_r^*) \leq \varepsilon$ for some prescribed tolerance ε. We argue that under some conditions on the likelihood f, the rank of P_r can be significantly smaller than d.

We make the following assumption on the prior distribution μ. Here, the notation $A \succeq B$ means that the matrix $A - B$ is positive semidefinite.

Assumption 2.5. The distribution μ is supported on a convex set $K = \mathrm{supp}(\mu) \subset \mathbb{R}^d$ and admits a Lebesgue density ρ such that $\rho \propto \exp(-V - \Psi)$, where V and Ψ are two functions defined on K which satisfy the following two properties.

(a) V is twice continuously differentiable and there exists a symmetric positive definite matrix $\Gamma \in \mathbb{R}^{d\times d}$ such that for all $x \in K$ we have

$$\nabla^2 V(x) \succeq \Gamma. \tag{10}$$

(b) Ψ is a bounded function over K such that

$$\exp(\sup \Psi - \inf \Psi) \leq \kappa, \tag{11}$$

for some $\kappa \geq 1$.

Let us make some comments on this assumption. First consider the case $\kappa = 1$, which means that Ψ is a constant function so that $\rho \propto \exp(-V - \Psi) \propto \exp(-V)$. Assumption 2.5(a) corresponds to a strong convexity property of the function V. Intuitively, it means that V is "at least quadratically convex." Now, when $\kappa > 1$, Assumption 2.5(b) means that $c\exp(-V(x)) \leq \rho(x) \leq C\exp(-V(x))$ holds for all $x \in K$, where $c > 0$ and $C < \infty$ are two constants that satisfy $C/c \leq \kappa < \infty$.

Example 2.6 (Gaussian distribution)**.** Any Gaussian prior $\mu = \mathcal{N}(m, \Sigma)$ with mean $m \in \mathbb{R}^d$ and invertible covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ satisfies Assumption 2.5 with $K = \mathbb{R}^d$, $\kappa = 1$ and $\Gamma = \Sigma^{-1}$. Indeed, the density ρ associated to $\mathcal{N}(m, \Sigma)$ is such that $\rho \propto \exp(-V)$ where the function $V : x \mapsto \frac{1}{2}(x - m)^T\Sigma^{-1}(x - m)$ satisfies $\nabla^2 V(x) = \Gamma$ for all $x \in K$.

Example 2.7 (Gaussian mixture)**.** Let $\mu \propto \sum_{i=1}^n \alpha_i \mathcal{N}(m_i, \Sigma_i)$ with $\alpha_1 > 0$ and $0 \prec \Sigma_1^{-1} \prec \Sigma_i^{-1}$ for all $1 < i \leq n$, where $A \prec B$ means that $B - A$ is a positive definite matrix. The density ρ of the Gaussian mixture μ can be written as follows,

$$\rho \propto \sum_{i=1}^n \alpha_i \frac{\exp(-V_i)}{Z_i} = \exp(-V_1)\Big(\frac{\alpha_1}{Z_1} + \sum_{i=2}^n \frac{\alpha_i}{Z_i}\exp(-V_i + V_1)\Big),$$

where $V_i : x \mapsto \frac{1}{2}(x - m_i)^T\Sigma_i^{-1}(x - m_i)$ and $Z_i = \int \exp(-V_i)\mathrm{d}x$. For any $1 < i \leq n$ the function $-V_i + V_1$ is quadratic and, by assumption, its Hessian satisfies $-\Sigma_i^{-1} + \Sigma_1^{-1} \prec 0$. This ensures that $-V_i + V_1$ is bounded from above, and so is the function $\Psi = \log(\frac{\alpha_1}{Z_1} + \sum_{i=2}^n \frac{\alpha_i}{Z_i}\exp(-V_i + V_1))$. Furthermore, the relation $\Psi(x) \geq \log(\frac{\alpha_1}{Z_1}) > -\infty$ holds for all $x \in \mathbb{R}^d$, which means that Ψ is bounded from below. As a consequence we have $\rho \propto \exp(-V_1 - \Psi)$ so that μ satisfies Assumption 2.5 with $\Gamma = \Sigma_1^{-1}$ and for some finite κ.

Example 2.8 (Uniform distribution)**.** The uniform measure μ on a convex and bounded domain $K \subset \mathbb{R}^d$ satisfies Assumption 2.5 with

$$\Gamma = \frac{4}{\operatorname{diam}(K)^2} I_d, \text{ and } \quad \kappa = \exp(1),$$

where $\operatorname{diam}(K) = \max\{\|x - y\|_2 : x, y \in K\}$. To show this, let us first denote by $\mathcal{B}(m, r)$ the minimum enclosing ball of K, that is, the ball with minimal radius $r \geq 0$ such that $K \subset \mathcal{B}(m, r)$. By Jung's theorem [23] we have $r \leq \operatorname{diam}(K)\sqrt{d/(2(d+1))}$ which yields $r \leq \operatorname{diam}(K)/\sqrt{2}$. Second, observe that the density $\rho \propto 1_K$ of μ can be written as $\rho \propto \exp(-V - \Psi)$ where $V(x) = \|x - m\|^2/r^2$ and $\Psi = -V$. To conclude, we recognize that $\nabla^2 V(x) \succeq \frac{4}{\operatorname{diam}(K)^2} I_d$ and $\exp(\sup \Psi - \inf \Psi) \leq \exp(\sup_{x \in \mathcal{B}(m,r)} \|x - m\|^2/r^2) = \exp(1)$.

Assumption 2.5 provides sufficient conditions for the prior distribution μ to satisfy the following *logarithmic Sobolev inequality:*

$$\int h^2 \log \frac{h^2}{\int h^2 \, \mathrm{d}\mu} \, \mathrm{d}\mu \leq 2\kappa \int \|\nabla h\|^2_{\Gamma^{-1}} \, \mathrm{d}\mu, \tag{12}$$

for any smooth enough function $h : \mathbb{R}^d \to \mathbb{R}$, where $\|\cdot\|_{\Gamma^{-1}}$ denotes the norm on $\mathbb{R}^d$ such that $\|x\|^2_{\Gamma^{-1}} = x^T \Gamma^{-1} x$ for all $x \in \mathbb{R}^d$. This result relies on Bakry–Émery theorem [2, 6, 34], which uses Assumption 2.5(a), and on Holley–Stroock perturbation lemma [22], which uses Assumption 2.5(b). More precisely, Proposition 3.1 in [6] states that in the case $\kappa = 1$ (i.e., Ψ is constant), Assumption 2.5(a) is sufficient to have (12). Then, following the lines of the original proof in [22] (see also the proof of Theorem 1.9 in [18], for instance), we have that (12) still holds when Ψ is such that $\kappa > 1$.

The following proposition shows that Assumption 2.5 also implies another class of inequalities, which we call the *subspace logarithmic Sobolev inequalities.*

Theorem 2.9 (Subspace logarithmic Sobolev inequality)**.** Let μ be a probability distribution that satisfies Assumption 2.5 for some $\Gamma \in \mathbb{R}^{d \times d}$ and $\kappa \geq 1$. Then the relation

$$\int h^2 \log \frac{h^2}{\mathbb{E}_\mu(h^2 | \sigma(P_r))} \, \mathrm{d}\mu \leq 2\kappa \int \|(I_d - P_r^T)\nabla h\|^2_{\Gamma^{-1}} \, \mathrm{d}\mu, \tag{13}$$

holds for any continuously differentiable function $h : \mathbb{R}^d \to \mathbb{R}$ such that $\int \|\nabla h\|^2_{\Gamma^{-1}} \mathrm{d}\mu < \infty$ and for any projector $P_r \in \mathbb{R}^{d \times d}$.

The proof is given in Section B. The subspace logarithmic Sobolev inequality (13) allows us to derive an upper bound for the Kullback–Leibler divergence $\mathrm{D}_{\mathrm{KL}}(\nu || \nu_r^*)$.

Corollary 2.10. Let μ be a distribution which satisfies the subspace logarithmic Sobolev inequality (13) for some $\Gamma \in \mathbb{R}^{d \times d}$ and $\kappa \geq 1$, and let ν be such that $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \propto f$ for some continuously differentiable function f such that $\int \|\nabla \log f\|^2_{\Gamma^{-1}} \mathrm{d}\mu < \infty$. Then for any projector $P_r \in \mathbb{R}^{d \times d}$, the distribution ν_r^* such that $\frac{\mathrm{d}\nu_r^*}{\mathrm{d}\mu} \propto \mathbb{E}_\mu(f | \sigma(P_r))$ satisfies

$$\mathrm{D}_{\mathrm{KL}}(\nu || \nu_r^*) \leq \frac{\kappa}{2} \, \mathcal{R}_\Gamma(P_r, H), \tag{14}$$

where $\mathcal{R}_\Gamma(P_r, H) = \operatorname{trace}(\Gamma^{-1}(I_d - P_r^T) H (I_d - P_r))$ and where $H \in \mathbb{R}^{d \times d}$ is the symmetric positive matrix defined by

$$H = \int (\nabla \log f)(\nabla \log f)^T \mathrm{d}\nu. \tag{15}$$

Proof. The proof consists in rewriting (13) with $h = (f/Z)^{1/2}$, where $Z = \int \mathbb{E}_\mu(f|\sigma(P_r))\,\mathrm{d}\mu = \int f\,\mathrm{d}\mu$. We have $\nabla h = \frac{1}{2}(f/Z)^{1/2}\nabla \log f$ so that inequality (13) becomes

$$\int \frac{f}{Z} \log \frac{f/Z}{\mathbb{E}_\mu(f|\sigma(P_r))/Z}\,\mathrm{d}\mu \le \frac{\kappa}{2}\int \|(I_d - P_r^T)\nabla \log f\|^2_{\Gamma^{-1}} \frac{f}{Z}\,\mathrm{d}\mu.$$

and yields $\mathrm{D_{KL}}(\nu||\nu_r^*) \le \frac{\kappa}{2}\int \|(I_d - P_r^T)\nabla \log f\|^2_{\Gamma^{-1}}\mathrm{d}\nu$. By the definition of H, we can write

$$\begin{aligned}\int \|(I_d - P_r^T)\nabla \log f\|^2_{\Gamma^{-1}}\mathrm{d}\nu &= \int (\nabla \log f)^T(I_d - P_r)\Gamma^{-1}(I_d - P_r^T)\nabla \log f\ \mathrm{d}\nu \\ &= \int \mathrm{trace}\left(\Gamma^{-1}(I_d - P_r^T)\nabla \log f(\nabla \log f)^T(I_d - P_r)\right)\mathrm{d}\nu \\ &\overset{(15)}{=} \mathrm{trace}\left(\Gamma^{-1}(I_d - P_r)^T H(I_d - P_r)\right) = \mathcal{R}_\Gamma(P_r, H),\end{aligned}$$

which yields (14) and concludes the proof. □

Corollary 2.10 provides an upper bound for $\mathrm{D_{KL}}(\nu||\nu_r^*)$ that is proportional to $\mathcal{R}_\Gamma(P_r, H)$, where H is given in (15). As noticed in the previous proof, the relation

$$\mathcal{R}_\Gamma(P_r, H) = \int \|\nabla \log f - P_r^T \nabla \log f\|^2_{\Gamma^{-1}}\,\mathrm{d}\nu, \tag{16}$$

holds and shows that $\mathcal{R}_\Gamma(P_r, H)$ is the mean squared error of the approximation of the random vector $\nabla \log f(X)$ by $P_r^T \nabla \log f(X)$, where $X \sim \nu$. In the context of principal component analysis, $\mathcal{R}_\Gamma(P_r, H)$ is usually called the *reconstruction error*. The following proposition gives a closed form expression for a minimizer of $P_r \mapsto \mathcal{R}_\Gamma(P_r, H)$ over the set of the rank-r projectors. It corresponds to Proposition 2.6 in [48] where Σ is replaced by Γ^{-1}.

Proposition 2.11. Let $\Gamma \in \mathbb{R}^{d\times d}$ be a symmetric positive definite matrix and $H \in \mathbb{R}^{d\times d}$ be a symmetric positive semidefinite matrix. Denote by $(\lambda_i, v_i) \in \mathbb{R}_{\ge 0} \times \mathbb{R}^d$ the i-th generalized eigenpair of the matrix pencil (H, Γ), meaning $Hv_i = \lambda_i \Gamma v_i$ with $\lambda_i \ge \lambda_{i+1}$ and $\|v_i\|_\Gamma = 1$, where $\|\cdot\|_\Gamma = \sqrt{(\cdot)^T\Gamma(\cdot)}$. For any $r \le d$ we have

$$\min_{\substack{P_r \in \mathbb{R}^{d\times d} \\ \text{rank-}r \text{ projector}}} \mathcal{R}_\Gamma(P_r, H) = \sum_{i=r+1}^{d} \lambda_i\,. \tag{17}$$

Furthermore, a solution to (17) is the following Γ-orthogonal projector (i.e., satisfying $x^T\Gamma x = \|x\|^2_\Gamma = \|P_r x\|^2_\Gamma + \|(I_d - P_r)x\|^2_\Gamma$ for all $x \in \mathbb{R}^d$) given by

$$P_r = \Big(\sum_{i=1}^{r} v_i v_i^T\Big)\Gamma\,. \tag{18}$$

Corollary 2.10 and Proposition 2.11 ensure that, provided μ satisfies the subspace logarithmic Sobolev inequality (13) and provided P_r is defined as in (18), the approximation ν_r^* of ν defined by $\frac{\mathrm{d}\nu_r^*}{\mathrm{d}\mu} \propto \mathbb{E}_\mu(f|\sigma(P_r))$ is such that

$$\mathrm{D_{KL}}(\nu||\nu_r^*) \le \frac{\kappa}{2}\sum_{i=r+1}^{d} \lambda_i,$$

where λ_i is the i-th generalized eigenvalue of $H = \int(\nabla \log f)(\nabla \log f)^T\mathrm{d}\nu$. This relation holds for any $r \le d$. Then for any $\varepsilon \ge 0$, the choice $r = r(\varepsilon) := \min\{r : \frac{\kappa}{2}\sum_{i=r+1}^{d}\lambda_i \le \varepsilon\}$ is sufficient to

obtain $\mathrm{D}_{\mathrm{KL}}(\nu||\nu_r^*) \leq \varepsilon$. Observe that a strong decay in the generalized eigenvalues of H ensures that $r(\varepsilon) \ll d$. In particular, if H is rank deficient, we have $\lambda_i = 0$ for all $i > \mathrm{rank}(H)$ so that $r \geq \mathrm{rank}(H)$ implies $\mathrm{D}_{\mathrm{KL}}(\nu||\nu_r^*) = 0$. The spectrum of the matrix H reveals the low effective dimensionality of the posterior distribution: a strong decay in the generalized spectrum of H, or certainly a rank deficiency of H, ensures that there exists an approximation ν_r^* of ν such that $\mathrm{D}_{\mathrm{KL}}(\nu||\nu_r^*) \leq \varepsilon$ with small $r(\varepsilon) \ll d$.

Remark 2.12 (Prior-based dimension reduction). For any projector P_r we can write

$$\begin{aligned}\mathcal{R}_\Gamma(P_r, H) &= \mathrm{trace}(H(I_d - P_r)\Gamma^{-1}(I_d - P_r^T)) \\ &\leq \|H\| \,\mathrm{trace}((I_d - P_r)\Gamma^{-1}(I_d - P_r^T)) \\ &= \|H\| \mathcal{R}_{I_d}(P_r^T, \Gamma^{-1}),\end{aligned}$$

where $\|H\|$ is the spectral norm of H. In the case of Gaussian prior $\mu = \mathcal{N}(m, \Sigma)$ we have $\Gamma = \Sigma^{-1}$, see Example 2.6, and so Proposition 2.11 ensures that the minimizer of $P_r \mapsto \mathcal{R}_{I_d}(P_r^T, \Gamma^{-1}) = \mathcal{R}_{I_d}(P_r^T, \Sigma)$ is the orthogonal projector onto the leading eigenspace of the prior covariance Σ. This projector corresponds to the truncated Karhunen–Loève decomposition; see Section 4.1 for a detailed discussion.

2.3 An illustrative example

To illustrate how sharp the bound given by Corollary 2.10 can be, we consider a simple example for which the Kullback–Leibler divergence $\mathrm{D}_{\mathrm{KL}}(\nu||\nu_r^*)$ is computable in closed form. This allows a comparison of the error $\mathrm{D}_{\mathrm{KL}}(\nu||\nu_r^*)$ and its upper bound $\frac{\kappa}{2}\mathcal{R}_\Gamma(P_r, H)$.

Assume that the prior μ is the standard Gaussian distribution $\mathcal{N}(0, I_d)$ and let the likelihood function be given by

$$f : x \mapsto \exp(-\frac{1}{2}x^T A x),$$

where $A \in \mathbb{R}^{d\times d}$ is a symmetric positive semidefinite matrix. The Lebesgue density of μ is $\rho(x) \propto \exp(-\frac{1}{2}x^T x)$, so that μ satisfies Assumption 2.5 with $\Gamma = I_d$ and $\kappa = 1$. The posterior ν defined by $\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \propto f$ is also Gaussian with zero mean and covariance $\Sigma = (I_d + A)^{-1}$. In this setting, the matrix H defined by (15) can be written as follows

$$H = \int (\nabla \log f)(\nabla \log f)^T \mathrm{d}\nu = \int (Ax)(Ax)^T \nu(\mathrm{d}x) = A(I_d + A)^{-1}A.$$

Consider the generalized eigenvalue problem $Hv_i = \lambda_i \Gamma v_i$ which, since $\Gamma = I_d$, is simply the standard eigenvalue problem $Hv_i = \lambda_i v_i$. Notice that H is a rational function in A, so that H and A have the same eigenvectors and their eigenvalues satisfy the relation

$$\lambda_i = \frac{\alpha_i^2}{1 + \alpha_i},$$

for all $1 \leq i \leq d$, where α_i is the i-th largest eigenvalue of A. According to Proposition 2.11, a projector minimizing $P_r \mapsto \mathcal{R}_\Gamma(P_r, H)$ over the set of rank-r projectors is $P_r = \sum_{i=1}^r v_i v_i^T$. Corollary 2.10 ensures that the distribution ν_r^* defined by $\frac{\mathrm{d}\nu_r^*}{\mathrm{d}\mu} \propto \mathbb{E}_\mu(f|\sigma(P_r))$ is such that

$$\mathrm{D}_{\mathrm{KL}}(\nu||\nu_r^*) \leq \frac{1}{2}\sum_{i=r+1}^d \lambda_i = \frac{1}{2}\sum_{i=r+1}^d \frac{\alpha_i^2}{1 + \alpha_i}. \tag{19}$$

To analyze the sharpness of this inequality, we now compute $\mathrm{D}_{\mathrm{KL}}(\nu||\nu_r^*)$. Using Proposition 2.4, we can express the conditional expectation $\mathbb{E}_\mu(f|\sigma(P_r))$ as follows:

$$\mathbb{E}_\mu(f|\sigma(P_r)) : x \mapsto (\det(\Sigma_r^{-1}\Sigma))^{1/2} \exp(-\frac{1}{2} x^T P_r A P_r x), \tag{20}$$

where $\Sigma_r = (I_d + P_r A P_r)^{-1}$. Then ν_r^* such that $\frac{\mathrm{d}\nu_r^*}{\mathrm{d}\mu} \propto \mathbb{E}_\mu(f|\sigma(P_r))$ is Gaussian with zero mean and covariance Σ_r. The Kullback–Leibler divergence from $\nu_r^* = \mathcal{N}(0, \Sigma_r)$ to $\nu = \mathcal{N}(0, \Sigma)$ admits the following closed form expression:

$$\mathrm{D}_{\mathrm{KL}}(\nu||\nu_r^*) = \frac{1}{2}\big(\operatorname{trace}(\Sigma_r^{-1}\Sigma) - \log(\det(\Sigma_r^{-1}\Sigma)) - d\big).$$

To continue the calculation, one needs the eigenvalues of $\Sigma_r^{-1}\Sigma$. Let U be the orthogonal matrix containing the eigenvectors of A and let $D = \operatorname{diag}(\alpha_1, \ldots, \alpha_d)$ so that $A = UDU^T$. By construction, $P_r A P_r = U D_r U^T$ where $D_r = \operatorname{diag}(\alpha_1, \ldots, \alpha_r, 0, \ldots, 0)$, so that $\Sigma_r^{-1}\Sigma = (I_d + P_r A P_r)(I_d + A)^{-1} = U(I_d + D_r)(I_d + D)^{-1}U$. From this relation, we deduce that the i-th eigenvalue of $\Sigma_r^{-1}\Sigma$ is 1 when $i \leq r$ and $(1+\alpha_i)^{-1}$ otherwise. Then we have

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(\nu||\nu_r^*) &= \frac{1}{2}\Big(r + \sum_{i=r+1}^{d} \frac{1}{1+\alpha_i} + \sum_{i=r+1}^{d} \log(1+\alpha_i) - d\Big) \\
&= \frac{1}{2}\sum_{i=r+1}^{d} \Big(\log(1+\alpha_i) - \frac{\alpha_i}{1+\alpha_i}\Big).
\end{aligned}$$

We now analyze the deficit in inequality (19). With a Taylor expansion as α_i goes to zero for all $i \geq r$, we can write

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}(\nu||\nu_r^*) &= \frac{1}{4}\sum_{i=r+1}^{d} \big(\alpha_i^2 + \mathcal{O}(\alpha_i^3)\big), \text{ and} \\
\frac{1}{2}\sum_{i=r+1}^{d} \frac{\alpha_i^2}{1+\alpha_i} &= \frac{1}{2}\sum_{i=r+1}^{d} \big(\alpha_i^2 + \mathcal{O}(\alpha_i^3)\big).
\end{aligned}$$

If the function f is nearly constant ($\alpha_i \approx 0$) along the subspace $\operatorname{span}\{v_{r+1}, \ldots, v_d\}$, then the upper bound (19) is close to $2\,\mathrm{D}_{\mathrm{KL}}(\nu||\nu_r^*)$. In that case, the upper bound is, up to a multiplicative factor of two, the same as the true error.

In this particular example, the projector obtained by minimizing the upper bound in fact yields the *optimal* approximation of ν in Kullback–Leibler divergence, for any given rank r. This can be shown by Theorem 2.3 in [42]. Of course, minimizing the upper bound does not produce the optimal projector in general.

3 Building the approximation

In this section we propose and analyze algorithms for the numerical construction of a low-rank projector P_r and of a profile function g, such that the distribution ν_r given by $\frac{\mathrm{d}\nu_r}{\mathrm{d}\mu} \propto g \circ P_r$ is a controlled approximation of the posterior distribution ν. Recall that in the previous section (see (7)) we obtained the following decomposition of the Kullback–Leibler divergence:

$$\mathrm{D}_{\mathrm{KL}}(\nu||\nu_r) = \mathrm{D}_{\mathrm{KL}}(\nu||\nu_r^*) + \mathrm{D}_{\mathrm{KL}}(\nu_r^*||\nu_r), \tag{21}$$

where ν_r^* is given by $\frac{\mathrm{d}\nu_r^*}{\mathrm{d}\mu} \propto \mathbb{E}_\mu(f|\sigma(P_r))$. The first term in this decomposition depends only on P_r, not on g; using the material presented in Section 2.2, one can build P_r to ensure that $\mathrm{D_{KL}}(\nu||\nu_r^*)$ is arbitrarily small. The second term $\mathrm{D_{KL}}(\nu_r^*||\nu_r)$ can be interpreted as a distance between $g \circ P_r$ and $\mathbb{E}_\mu(f|\sigma(P_r))$ so that the construction of g consists essentially in approximating a conditional expectation. The computational strategy we adopt here is to: (i) construct P_r in order to control the first term; and (ii) given such a projector, build the function g so that the second term is arbitrarily close to zero—or, at least, of the same order as the first term.

3.1 Construction of the projector

Let us recall some of the results of Section 2.2. Under Assumption 2.5 and provided that f is sufficiently regular, Corollary 2.10 provides an upper bound for the first term $\mathrm{D_{KL}}(\nu||\nu_r^*)$ in the decomposition (21) by means of the reconstruction error $\mathcal{R}_\Gamma(P_r, H) = \mathrm{trace}(\Gamma^{-1}(I_d - P_r^T)H(I_d - P_r))$ where $H = \int (\nabla \log f)(\nabla \log f)^T \mathrm{d}\nu$. This bound holds for any projector P_r. We denote by P_r^* a rank-r projector which minimizes the reconstruction error, meaning

$$\mathcal{R}_\Gamma(P_r^*, H) = \min_{\substack{P_r \in \mathbb{R}^{d\times d}, \\ \text{rank-}r \text{ projector}}} \mathcal{R}_\Gamma(P_r, H). \tag{22}$$

By Proposition 2.11, P_r^* can be obtained by means of the generalized eigendecomposition of H. In practice, however, the matrix H may be difficult to compute exactly because it requires computing a high-dimensional integral when $d \gg 1$. Instead, we consider the rank-r projector $\widehat{P}_r$ that is a minimizer of the *approximate* reconstruction error $\mathcal{R}_\Gamma(P_r, \widehat{H})$, i.e.,

$$\mathcal{R}_\Gamma(\widehat{P}_r, \widehat{H}) = \min_{\substack{P_r \in \mathbb{R}^{d\times d}, \\ \text{rank-}r \text{ projector}}} \mathcal{R}_\Gamma(P_r, \widehat{H}), \tag{23}$$

where $\widehat{H}$ is a Monte Carlo approximation of H defined by

$$\widehat{H} = \frac{1}{K} \sum_{i=1}^{K} \big(\nabla \log f(X_i)\big)\big(\nabla \log f(X_i)\big)^T. \tag{24}$$

Here $X_1, \ldots, X_K$ are independent samples from the posterior distribution ν. This corresponds to principal component analysis (PCA) of the random vector $\nabla \log f(X)$, $X \sim \nu$. An important question is how large should K be in order to control the reconstruction error $\mathcal{R}_\Gamma(\widehat{P}_r, H)$? We refer to [5, 37] for recent progress in this direction. The following proposition gives a new sufficient condition on K so that $\widehat{P}_r$ is a quasi-optimal solution to (22). This result relies on concentration properties of sub-Gaussian random vectors; see [7, 47]. The proof is given in Section C.

Proposition 3.1. Let μ, ν be two probability distributions and f be a sufficiently smooth function. Assume there exists a constant $L \geq 0$ such that the random vector $\nabla \log f(X)$ with $X \sim \nu$ satisfies

$$\|u^T \nabla \log f(X)\|_{\psi_2} \leq L \sqrt{u^T H u}, \tag{25}$$

for any $u \in \mathbb{R}^d$, where $H = \int (\nabla \log f)(\nabla \log f)^T \mathrm{d}\nu$. Here $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm, meaning $\|\xi\|_{\psi_2} = \sup_{k\geq 1} k^{-1/2} \mathbb{E}(|\xi|^k)^{1/k}$ for any real-valued random variable ξ. Let $\widehat{H}$ be a K-sample Monte Carlo approximation of H. For any $0 < \delta, \eta < 1$, the condition

$$K \geq \Omega \delta^{-2} L^4 \big(\sqrt{\mathrm{rank}(H)} + \sqrt{\log(2\eta^{-1})}\big)^2, \tag{26}$$

for some absolute (numerical) constant Ω, ensures that with probability at least $1-\eta$, the following relation holds for any projector P_r:

$$(1-\delta)\ \mathcal{R}_\Gamma(P_r, H) \le \mathcal{R}_\Gamma(P_r, \widehat{H}) \le (1+\delta)\ \mathcal{R}_\Gamma(P_r, H). \tag{27}$$

Assumption (25) means that $\nabla \log f(X)$ is a sub-Gaussian vector. Intuitively it says that the tails of the distribution of $\nabla \log f(X)$ are at most Gaussian. Together with the independence of $X_1, \dots, X_K$, the sub-Gaussian property is an essential ingredient in the proof of Proposition 3.1. We give now examples for which (25) is satisfied.

Example 3.2 (Bounded gradients)**.** Assume $x \mapsto \nabla \log f(x)$ is uniformly bounded and assume H is full rank. Let $L < \infty$ be such that $\|\nabla \log f(x)\|_{H^{-1}} \le L$ for all $x \in \mathbb{R}^d$. We have

$$\sup_{u \in \mathbb{R}^d \setminus \{0\}} \frac{|u^T \nabla \log f(x)|}{\sqrt{u^T H u}} = \|\nabla \log f(x)\|_{H^{-1}} \le L$$

for any $x \in \mathbb{R}^d$ so that $|u^T \nabla \log f(x)| \le L\sqrt{u^T H u}$ for all $x, u \in \mathbb{R}^d$. This means that $|u^T \nabla \log f(X)|$ is almost surely bounded by $L\sqrt{u^T H u}$. Since the ψ_2-norm is bounded by the L^∞-norm, we have $\|u^T \nabla \log f(X)\|_{\psi_2} \le L\sqrt{u^T H u}$. This shows that relation (25) holds true for all $u \in \mathbb{R}^d$.

Example 3.3 (Unbounded gradients)**.** We now consider an example where $\nabla \log f(X)$ satisfies (25) without being bounded. As in Section 2.3, let $\mu = \mathcal{N}(0, I_d)$ be the standard normal prior and let $f : x \mapsto \exp(-\frac{1}{2} x^T A x)$ for some symmetric matrix $A \succeq 0$. Notice that $\nabla \log f(x) = -Ax$ can be arbitrarily large when $\|x\| \to \infty$. We have $X \sim \nu = \mathcal{N}(0, (I_d + A)^{-1})$ and $H = A(I_d + A)^{-1} A$. For any $u \in \mathbb{R}^d$, the random variable $Z = u^T \nabla \log f(X) = -u^T A X$ is Gaussian with zero mean and variance $\sigma_Z^2 = u^T A (I_d + A)^{-1} A u = u^T H u$. Then Z is sub-Gaussian and $\|Z\|_{\psi_2} \le L \sigma_Z$ holds for some absolute constant L; see, for instance, Example 5.8 in [47]. This means that (25) holds for all $u \in \mathbb{R}^d$ with a constant L which does not depend on A.

Proposition 3.1 gives a sufficient condition for (27) to hold with high probability. This relation yields the following quasi-optimality result:

$$\mathcal{R}_\Gamma(\widehat{P}_r, H) \overset{(27)}{\le} \frac{1}{1-\delta} \mathcal{R}_\Gamma(\widehat{P}_r, \widehat{H}) \overset{(23)}{\le} \frac{1}{1-\delta} \mathcal{R}_\Gamma(P_r^*, \widehat{H}) \overset{(27)}{\le} \frac{1+\delta}{1-\delta} \mathcal{R}_\Gamma(P_r^*, H).$$

Then, even though $\widehat{P}_r$ is not a minimizer of $P_r \mapsto \mathcal{R}_\Gamma(P_r, H)$, the reconstruction error $\mathcal{R}_\Gamma(\widehat{P}_r, H)$ is no greater than an arbitrary multiplicative constant $\frac{1+\delta}{1-\delta} \ge 1$ times the minimum of the reconstruction error $\mathcal{R}_\Gamma(P_r^*, H)$. In particular $\mathcal{R}_\Gamma(P_r^*, H) = 0$ implies $\mathcal{R}_\Gamma(\widehat{P}_r, H) = 0$.

Now let λ_i and $\widehat{\lambda}_i$ be the i-th eigenvalues of the matrix pencils (H, Γ) and $(\widehat{H}, \Gamma)$, respectively. Together with Corollary 2.10 and Proposition 2.11, the above quasi-optimality result allows us to deduce that, with probability at least $1-\eta$, we have

$$\mathrm{D_{KL}}(\nu || \nu_r^*) \le \frac{\kappa}{2(1-\delta)} \sum_{i=r+1}^{d} \widehat{\lambda}_i \ \le\ \frac{\kappa(1+\delta)}{2(1-\delta)} \sum_{i=r+1}^{d} \lambda_i, \tag{28}$$

where $\frac{\mathrm{d}\nu_r^*}{\mathrm{d}\mu} \propto \mathbb{E}_\mu(f | \sigma(P_r))$ with $P_r = \widehat{P}_r$. The first inequality above shows that the eigenvalues $\widehat{\lambda}_1, \widehat{\lambda}_2, \dots$ can be used to control the error $\mathrm{D_{KL}}(\nu || \nu_r^*)$: a fast decay in the spectrum of $(\widehat{H}, \Gamma)$ reveals the low effective dimension of the posterior ν. The second part of this relation ensures that,

up to a multiplicative constant $\frac{1+\delta}{1-\delta}$, the error $\mathrm{D_{KL}}(\nu||\nu_r^*)$ decays with r at the same rate as if we had used the projector $P_r = P_r^*$ that minimizes the true reconstruction error.

We note that condition (26) requires K to be at least proportional to the rank of H. If H is full rank and if $d \gg 1$, then K ought to be large in order to satisfy (26). In practice, though, we observe that suitable projectors can be obtained by choosing K proportional to the rank r of $\widehat{P}_r$ (which is PCA's "rule of thumb"). Here there is a challenge in finding weaker conditions on K that nonetheless allow one to obtain a quasi-optimality result. As in [37], one should try to exploit some properties of H such as a rapid spectral decay or a large spectral gap.

3.2 Approximation of the conditional expectation

In this section we assume that a rank-r projector $P_r \in \mathbb{R}^{d\times d}$ is given. We consider the problem of approximating the conditional expectation $\mathbb{E}_\mu(f|\sigma(P_r))$. Recall that, using the notation of Proposition 2.4, we have

$$\mathbb{E}_\mu(f|\sigma(P_r)) : x \mapsto \int_{\mathbb{R}^{d-r}} f(P_r x + U_\perp \xi_\perp)\, p_\perp(\xi_\perp|P_r x)\, \mathrm{d}\xi_\perp. \tag{29}$$

For any $x \in \mathbb{R}^d$, an approximation of $\mathbb{E}_\mu(f|\sigma(P_r))(x)$ can be obtained via a Monte Carlo estimate of the form $\frac{1}{M}\sum_{i=1}^M f\big(P_r x + U_\perp \Xi_i\big)$, where $\Xi_1, \dots, \Xi_M \in \mathbb{R}^{d-r}$ are independent copies of the random vector $\Xi \sim p_\perp(\,\cdot\,|P_r x)$. In general, the law of Ξ depends on x and we should redraw the samples $\Xi_1, \dots, \Xi_M$ for each different value of x, which can be computationally demanding.

Instead, we will consider an approximation of $\mathbb{E}_\mu(f|\sigma(P_r))$ of the form

$$\widehat{F}_r : x \mapsto \frac{1}{M}\sum_{i=1}^M f(P_r x + (I_d - P_r)Y_i), \tag{30}$$

where $Y_1, \dots, Y_M$ are independent samples from the prior μ. From a computational perspective, the samples Y_i will be drawn *once* and reused for each evaluation of $x \mapsto \widehat{F}_r(x)$. We define $\widehat{\nu}_r$ to be the random probability distribution defined by

$$\frac{\mathrm{d}\widehat{\nu}_r}{\mathrm{d}\mu} \propto \widehat{F}_r. \tag{31}$$

Notice that the laws of $(I_d - P_r)Y_i$ and of $U_\perp \Xi_i$ are in general different, so that $\widehat{F}_r(x)$ is a biased estimate of $\mathbb{E}_\mu(f|\sigma(P_r))(x)$. In this case we cannot hope for $\mathrm{D_{KL}}(\nu_r^*||\widehat{\nu}_r)$ to go to zero with M. In order to analyze this bias, let us introduce the probability distribution μ' such that

$$\int h \,\mathrm{d}\mu' = \mathbb{E}\big(h(P_r X + (I_d - P_r)Y)\big), \tag{32}$$

for any Borel function h, where X and Y are independent random variables with distribution μ. Equation (32) is equivalent to saying that μ' is the probability distribution of the random vector $P_r X + (I_d - P_r)Y$. The following proposition is proven in Section D.

Proposition 3.4. The conditional expectation $\mathbb{E}_{\mu'}(f|\sigma(P_r))$ is such that

$$\mathbb{E}_{\mu'}(f|\sigma(P_r)) : x \mapsto \mathbb{E}\big(f(P_r x + (I_d - P_r)Y)\big).$$

By Proposition 3.4 we can write $\mathbb{E}(\widehat{F}_r) = \mathbb{E}_{\mu'}(f|\sigma(P_r))$ which allows us to interpret $\widehat{F}_r$ as a Monte Carlo estimate of $\mathbb{E}_{\mu'}(f|\sigma(P_r))$. The following proposition gives a bound for the expectation of the error $\mathrm{D_{KL}}(\nu_r^*||\widehat{\nu}_r)$. The proof is given in Section E.

Proposition 3.5. Let ν_r' be the distribution such that $\frac{\mathrm{d}\nu_r'}{\mathrm{d}\mu} \propto \mathbb{E}_{\mu'}(f|\sigma(P_r))$. We have

$$\mathbb{E}\big(\mathrm{D_{KL}}(\nu_r^*||\widehat{\nu}_r)\big) \leq \mathrm{D_{KL}}(\nu_r^*||\nu_r') + \frac{1}{2}\,\mathbb{E}\int \Big(\frac{\widehat{F}_r - \mathbb{E}_{\mu'}(f|\sigma(P_r))}{\mathbb{E}_{\mu'}(f|\sigma(P_r))}\Big)^2 \mathrm{d}\nu_r^* + \mathbb{E}\int \mathcal{O}\left(\Big(\frac{\widehat{F}_r(x) - \mathbb{E}_{\mu'}(f|\sigma(P_r))(x)}{\mathbb{E}_{\mu'}(f|\sigma(P_r))(x)}\Big)^3\right) \nu_r^*(\mathrm{d}x). \tag{33}$$

Proposition 3.5 shows that, up to a third order term, the expectation of the error $\mathrm{D_{KL}}(\nu_r^*||\widehat{\nu}_r)$ is bounded by the sum of two terms. The first term is the Kullback–Leibler divergence from ν_r' to ν_r^* and corresponds to a measure of the bias $\mathbb{E}(\widehat{F}_r) \neq \mathbb{E}_\mu(f|\sigma(P_r))$. The second term can be interpreted as a measure of the variance of $\widehat{F}_r$. Under some assumptions on the distribution μ, the following proposition provides an upper bound for those two terms.

Proposition 3.6. Assume μ admits a Lebesgue density $\rho \propto \exp(-V - \Psi)$, where Ψ is a bounded function such that $\exp(\sup\Psi - \inf\Psi) \leq \kappa$ and where $V : x \mapsto \frac{1}{2}\|x - m\|_\Gamma^2$ for some $m \in \mathbb{R}^d$ and for some symmetric positive-definite matrix $\Gamma \in \mathbb{R}^{d\times d}$. Then, for any $\|\cdot\|_\Gamma$-orthogonal projector P_r, we have

$$\mathrm{D_{KL}}(\nu_r^*||\nu_r') \leq \kappa^7(\kappa^2 - 1)\mathcal{E}_\Gamma(P_r, f), \tag{34}$$

and

$$\frac{1}{2}\,\mathbb{E}\int \Big(\frac{\widehat{F}_r - \mathbb{E}_{\mu'}(f|\sigma(P_r))}{\mathbb{E}_{\mu'}(f|\sigma(P_r))}\Big)^2 \mathrm{d}\nu_r^* \leq \frac{\kappa^7}{2M}\mathcal{E}_\Gamma(P_r, f), \tag{35}$$

where

$$\mathcal{E}_\Gamma(P_r, f) = \int \|(I_d - P_r^T)\nabla \log f\|_{\Gamma^{-1}}^2 \frac{f}{\mathbb{E}_\mu(f|\sigma(P_r))}\mathrm{d}\nu. \tag{36}$$

The proof is given is Section F. Proposition 3.6 requires μ to be a bounded perturbation of a Gaussian distribution $\mathcal{N}(m, \Gamma^{-1})$. This is a stronger assumption than Assumption 2.5. Neglecting the third order term in (33), Proposition 3.6 allows one to bound $\mathbb{E}(\mathrm{D_{KL}}(\nu_r^*||\widehat{\nu}_r))$ by $(C_1 + \frac{C_2}{M})\mathcal{E}_\Gamma(P_r, f)$ where C_1 and C_2 are two constants which depends only on κ. If μ is Gaussian ($\kappa = 1$) then $C_1 = 0$ and $C_2 = 1/2$. In that case, $\mathbb{E}(\mathrm{D_{KL}}(\nu_r^*||\widehat{\nu}_r))$ goes to zero with M. This is not surprising because $\mu' = \mu$ holds whenever μ is Gaussian,[1] so that $\nu_r^* = \nu_r'$ and hence $\mathrm{D_{KL}}(\nu_r^*||\nu_r') = 0$. In the general case $\kappa \neq 1$, (34) and (35) show that both the variance and the bias of $\widehat{F}_r$ are no greater than a constant independent of P_r times $\mathcal{E}_\Gamma(P_r, f)$. Note that the quantity $\mathcal{E}_\Gamma(P_r, f)$ differs from the reconstruction error $\mathcal{R}_\Gamma(P_r, H)$ only by the term $f\,\mathbb{E}_\mu(f|\sigma(P_r))^{-1}$; compare (16) and (36). Provided that this term is uniformly bounded from above, we can write

$$\mathcal{E}_\Gamma(P_r, f) \leq \Big(\sup \frac{f}{\mathbb{E}_\mu(f|\sigma(P_r))}\Big)\mathcal{R}_\Gamma(P_r, H).$$

The above relation shows that the error $\mathbb{E}(\mathrm{D_{KL}}(\nu_r^*||\widehat{\nu}_r))$ can be controlled by the reconstruction error $\mathcal{R}_\Gamma(P_r, H)$, provided the supremum of $f\,\mathbb{E}_\mu(f|\sigma(P_r))^{-1}$ is finite. Then a small number of samples M will be enough to have a guarantee that $\mathrm{D_{KL}}(\nu_r^*||\widehat{\nu}_r)$ is, in expectation, of the same order of magnitude as the first term $\mathrm{D_{KL}}(\nu||\nu_r^*)$ in the decomposition (21), which was our initial motivation. We now give an example for which the term $\sup f\,\mathbb{E}_\mu(f|\sigma(P_r))^{-1}$ decreases to one with the rank of the projector.

[1] Recall that P_r is an orthogonal projector with respect to the norm induced by the precision matrix of μ.

Example 3.7. As in Section 2.3, consider $\mu = \mathcal{N}(0, I_d)$ and $f : x \mapsto \exp(-\frac{1}{2}x^T A x)$ for some symmetric matrix $A \succeq 0$. Again, let $P_r = \sum_{i=1}^r v_i v_i^T$ where v_i is the i-th eigenvector of A and let $\Sigma = (I_d + A)^{-1}$ and $\Sigma_r = (I_d + P_r A P_r)^{-1}$. Using the closed-form expression (20) for $\mathbb{E}_\mu(f|\sigma(P_r))$ and since $A - P_r A P_r = (I_d - P_r)A(I_d - P_r) \succeq 0$, we have

$$\begin{aligned}\sup \frac{f}{\mathbb{E}_\mu(f|\sigma(P_r))} &= \sup_{x\in\mathbb{R}^d} \frac{\exp(-\frac{1}{2}x^T(A - P_r A P_r)x)}{\det(\Sigma_r^{-1}\Sigma)^{1/2}} \\ &= \frac{1}{\det(\Sigma_r^{-1}\Sigma)^{1/2}} = \prod_{i=r+1}^{d} (1+\alpha_i)^{1/2},\end{aligned}$$

where $\alpha_i \geq 0$ is the i-th eigenvalue of A. This shows that when P_r is a minimizer of the reconstruction error, the supremum of $f\,\mathbb{E}_\mu(f|\sigma(P_r))^{-1}$ goes monotonically to one with the rank r of P_r.

Remark 3.8. In the way they are presented, the error analyses of the projector and of the conditional expectation (i.e., the two terms on the right-hand side of (21)) are not unified: the first term is bounded in high probability (28), while the second is controlled in expectation (via Proposition 3.5 and subsequent discussion). One could unify these results, for instance by using a Markov inequality to control the probability that $\mathrm{D}_{\mathrm{KL}}(\nu_r^* || \widehat{\nu}_r)$ exceeds a certain value. However, we believe this step is not essential to the purpose of this section, which is to show that the associated sample approximations are feasible and sound.

3.3 Algorithms

3.3.1 Ideal algorithm

Algorithm 1 can be used to construct an approximation $\widehat{\nu}_r$ of the posterior distribution ν. It assumes that we can draw samples from the posterior distribution ν. Since this is typically not possible in actual practice, this algorithm is called "ideal".

Algorithm 1 Ideal algorithm

Require: Error threshold ε, sample sizes K and M.

1: Draw K independent samples $X_1, \ldots, X_K$ from ν
2: Compute $\nabla \log f(X_k)$ for $k = 1, \ldots, K$
3: Assemble the matrix $\widehat{H} = \frac{1}{K}\sum_{k=1}^K \big(\nabla \log f(X_k)\big)\big(\nabla \log f(X_k)\big)^T$
4: Compute a rank-r projector P_r with the smallest rank such that $\mathcal{R}_\Gamma(P_r, \widehat{H}) \leq \varepsilon$
5: Draw M samples $Y_1, \ldots, Y_M$ from μ
6: Return the approximate distribution $\widehat{\nu}_r$ defined by

$$\frac{\mathrm{d}\widehat{\nu}_r}{\mathrm{d}\mu} \propto \widehat{F}_r, \quad \text{where} \quad \widehat{F}_r = \frac{1}{M}\sum_{i=1}^M (P_r x + (I_d - P_r)Y_i)$$

3.3.2 Construction using approximate measures

Obtaining samples from the posterior distribution ν at step 1 of Algorithm 1 can be difficult in practice; indeed, this challenge is one of the motivations for the dimension reduction approach presented here. To alleviate this difficulty, we can construct $\widehat{H}$ using samples from another distribution $\widetilde{\nu}$ that can be directly simulated. By the definition (15) of H, we have

$$\begin{aligned}\mathcal{R}_\Gamma(P_r, H) &= \int \|(I_d - P_r^T)\nabla \log f\|^2_{\Gamma^{-1}} \frac{\mathrm{d}\nu}{\mathrm{d}\widetilde{\nu}} \mathrm{d}\widetilde{\nu} \\ &\leq \big(\sup \frac{\mathrm{d}\nu}{\mathrm{d}\widetilde{\nu}}\big) \int \|(I_d - P_r^T)\nabla \log f\|^2_{\Gamma^{-1}} \mathrm{d}\widetilde{\nu},\end{aligned}$$

so that, thanks to Corollary 2.10, we can write

$$\mathrm{D_{KL}}(\nu||\nu_r^*) \leq \big(\sup \frac{\mathrm{d}\nu}{\mathrm{d}\widetilde{\nu}}\big) \frac{\kappa}{2} \mathcal{R}_\Gamma(P_r, \widetilde{H}), \quad \text{where} \quad \widetilde{H} = \int (\nabla \log f)(\nabla \log f)^T \mathrm{d}\widetilde{\nu}, \tag{37}$$

for any projector P_r. Assuming $\widetilde{\nu}$ satisfies $\sup \frac{\mathrm{d}\nu}{\mathrm{d}\widetilde{\nu}} < +\infty$ (a common assumption in importance sampling [35, 38]), the above relation suggests that minimizing the approximate reconstruction error $P_r \mapsto \mathcal{R}_\Gamma(P_r, \widetilde{H})$ can yield suitable projectors. In particular if $\widetilde{H}$ is rank deficient, then $\mathcal{R}_\Gamma(P_r, \widetilde{H}) = 0$ and thus $\mathrm{D_{KL}}(\nu||\nu_r^*) = 0$ for a suitable projector P_r with $r = \mathrm{rank}(\widetilde{H})$. In general, however, the constant $\sup \frac{\mathrm{d}\nu}{\mathrm{d}\widetilde{\nu}}$ will be unknown in practice, and we will not be able to use (37) as a quantitative error bound for the error $\mathrm{D_{KL}}(\nu||\nu_r^*)$.

By drawing samples from $\widetilde{\nu}$ rather than from ν at step 1 of Algorithm 1, the matrix $\widehat{H}$ computed at step 2 is a Monte Carlo approximation of $\widetilde{H}$. Proposition 3.1 still applies when replacing ν by $\widetilde{\nu}$. Then for K sufficiently large ($K = \mathcal{O}(\mathrm{rank}(\widetilde{H}))$, for instance), it holds with high probability that any rank-r projector that minimizes $P_r \mapsto \mathcal{R}_\Gamma(P_r, \widehat{H})$ will be a quasi-optimal solution to the minimization problem of $P_r \mapsto \mathcal{R}_\Gamma(P_r, \widetilde{H})$; see the discussion of Section 3.1. Here we list two common choices of $\widetilde{\nu}$.

1. **Laplace approximation**. The Laplace approximation constructs a Gaussian distribution $\widetilde{\nu} = \mathcal{N}(\widetilde{m}, \widetilde{\Sigma})$, where $\widetilde{m}$ is the mode of ν and the matrix $\widetilde{\Sigma}^{-1}$ is obtained from Hessian of the negative log density of ν evaluated at $\widetilde{m}$. Note that if ν is Gaussian then $\widetilde{\nu}$ is exactly ν. Even though Laplace's method is a simple way to obtain a Gaussian approximation of ν, there is no guarantee that $\sup \frac{\mathrm{d}\nu}{\mathrm{d}\widetilde{\nu}}$ is finite in general.

2. **Prior distribution**. Sampling from the prior distribution μ is usually tractable. With the choice $\widetilde{\nu} = \mu$, we have $\frac{\mathrm{d}\nu}{\mathrm{d}\widetilde{\nu}} = \frac{\mathrm{d}\nu}{\mathrm{d}\mu} = \frac{f}{\int f \mathrm{d}\mu}$. For most applications, the likelihood function f is bounded so that $\sup \frac{\mathrm{d}\nu}{\mathrm{d}\widetilde{\nu}} < \infty$. Note that this choice has been considered in [11].

3.3.3 Iterative construction

As suggested by [15], rather than limiting ourselves to a fixed approximation $\widetilde{H}$, we can approximate the true H (15) using a sequential importance sampling framework. Let us consider a sequence of posterior approximations $\widetilde{\nu}_r^{(0)}, \ldots, \widetilde{\nu}_r^{(L)}$ where $\widetilde{\nu}_r^{(0)} = \mu$ and, for any $1 \leq l \leq L$, the distribution $\widetilde{\nu}_r^{(l)}$ is associated with a projector $P_r^{(l)}$ by

$$\frac{\mathrm{d}\widetilde{\nu}_r^{(l)}}{\mathrm{d}\mu} \propto \widehat{F}_r^{(l)}, \quad \text{where} \quad \widehat{F}_r^{(l)} : x \mapsto \frac{1}{M} \sum_{i=1}^M f\left(P_r^{(l)} x + (I_d - P_r^{(l)}) Y_i\right).$$

Notice that the same samples Y_i are used for every $1 \leq l \leq L$. The idea is to use $\widetilde{\nu}_r^{(l)}$ as a biasing distribution for the estimation of H. Let $X_1^{(l)}, \ldots, X_K^{(l)}$ be K independent samples from $\widetilde{\nu}_r^{(l)}$. We can write

$$H = \int \big(\nabla \log f(x)\big)\big(\nabla \log f(x)\big)^T \frac{\mathrm{d}\nu}{\mathrm{d}\widetilde{\nu}_r^{(l)}} \mathrm{d}\widetilde{\nu}_r^{(l)},$$

so that

$$\widehat{H}^{(l)} = \frac{1}{\sum_{k=1}^K w_k^{(l)}} \sum_{k=1}^K w_k^{(l)} \big(\nabla \log f(X_k^{(l)})\big)\big(\nabla \log f(X_k^{(l)})\big)^T, \quad w_k^{(l)} = \frac{f(X_k^{(l)})}{\widehat{F}_r^{(l)}(X_k^{(l)})}$$

is a self-normalized importance sampling estimator of H. Having computed $\widehat{H}^{(l)}$, the $(l+1)$-th projector $P_r^{(l+1)}$ is defined as a projector with minimal rank such that the approximate reconstruction error $\mathcal{R}_\Gamma(P_r^{(l+1)}, \widehat{H}^{(l)})$ is below some prescribed tolerance.

This iterative construction is detailed in Algorithm 2. Notice that at the first iteration, the importance weights $w_k^{(0)}$ are set to one. We make this choice in order to avoid the potential degeneracy (i.e., large variance of the weights) that might occur when μ is a poor approximation of ν. Also, our implementation includes a constraint on the rank of the projector, so that it cannot exceed a user-defined maximum rank $r_{\max}$. By doing so, we avoid any explosion of the rank in the earlier stages of the algorithm, i.e., when a poor posterior approximation might yield a crude approximation of H.

4 Alternative approaches to dimension reduction

4.1 Karhunen–Loève-based dimension reduction

The Karhunen–Loève decomposition is a simple and powerful tool for reducing the dimension of a given random vector $X \in \mathbb{R}^d$. Letting $m = \mathbb{E}(X)$, this method exploits the fact that $X - m$ may take values mostly on a low-dimensional subspace of $\mathbb{R}^d$, so that $X - m$ can be well approximated by $P_r(X - m)$ for some low-rank projector $P_r \in \mathbb{R}^{d \times d}$. The standard approach is to seek P_r such that the mean squared error $\mathbb{E}(\|(X - m) - P_r(X - m)\|_2^2)$ is below some prescribed tolerance. We can write

$$\mathbb{E}(\|(X - m) - P_r(X - m)\|_2^2) = \mathrm{trace}((I_d - P_r)\Sigma(I_d - P_r^T)),$$

where $\Sigma = \mathbb{E}((X - m)(X - m)^T)$ is the covariance matrix of X. Using Proposition 2.11, we have that the orthogonal projector onto the r-dimensional leading eigenspace of Σ is a minimizer of the mean squared error over the set of rank-r projectors. Furthermore we have

$$\min_{\substack{P_r \in \mathbb{R}^{d \times d}, \\ \text{rank-}r \text{ projector}}} \mathbb{E}(\|(X - m) - P_r(X - m)\|_2^2) = \sum_{i=r+1}^d \sigma_i^2,$$

where σ_i^2 is the i-th eigenvalue of Σ. This relation shows that a strong decay in the spectrum of Σ ensures the mean squared error can be arbitrarily small for some $r \ll d$. The eigenvectors of Σ are called the Karhunen–Loève modes of X and $m + P_r(X - m)$ corresponds to the truncated Karhunen–Loève decomposition of X.

This methodology can be used to approximate the prior distribution μ. Assuming $X \sim \mu$ has mean $m = \mathbb{E}(X)$ and covariance matrix Σ, let μ_r be the distribution of $m + P_r(X - m)$, where P_r

Algorithm 2 Iterative algorithm

Require: Threshold ε, sample sizes K and M, maximum number of iterations L, maximum rank $r_{\max}$

Draw M samples $Y_1, \dots, Y_M$ from μ
for $l = 0, \dots, L$ **do**
  if $l = 0$ **then**
    Draw K samples $X_1^{(l)}, \dots, X_K^{(l)}$ from μ
    Compute $\nabla \log f(X_k^{(l)})$ and set the weights $w_k^{(l)} = 1$ for $k = 1, \dots, K$
  else
    Draw K samples $X_1, \dots, X_K$ from $\widehat{\nu}_r^{(l)}$ using MCMC
    Compute $\nabla \log f(X_k^{(l)})$ and $w_k^{(l)} = \frac{f(X_k^{(l)})}{\widehat{F}_r^{(l)}(X_k^{(l)})}$ for $k = 1, \dots, K$
  end if
  Assemble the matrix

$$\widehat{H}^{(l)} = \frac{1}{\sum_{k=1}^K w_k^{(l)}} \sum_{k=1}^K w_k^{(l)} \big(\nabla \log f(X_k^{(l)})\big) \big(\nabla \log f(X_k^{(l)})\big)^T$$

  Compute the lowest rank r_ε such that $\mathcal{R}_\Gamma(P, \widehat{H}^{(l)}) \leq \varepsilon$ for some rank-r_ε projector P
  Put $r = \min(r_{\max}, r_\varepsilon)$ and form the rank-r projector $P_r^{(l+1)}$ which minimizes $P \mapsto \mathcal{R}_\Gamma(P, \widehat{H}^{(l)})$
  Define the approximate distribution $\widehat{\nu}_r^{(l+1)}$ as

$$\frac{\mathrm{d}\widehat{\nu}_r^{(l+1)}}{\mathrm{d}\mu} \propto \widehat{F}_r^{(l+1)}, \quad \text{where} \quad \widehat{F}_r^{(l+1)} : x \mapsto \frac{1}{M} \sum_{i=1}^M f\left(P_r^{(l+1)} x + (I_d - P_r^{(l+1)}) Y_i\right)$$

end for
Return the approximate distribution $\widehat{\nu}_r^{(l+1)}$.

is the rank-r orthogonal projector onto the dominant eigenspace of Σ. As proposed in [28, 32, 27], we can then introduce the approximate posterior $\tilde{\nu}_r$ such that

$$\frac{\mathrm{d}\tilde{\nu}_r}{\mathrm{d}\mu_r} \propto f.$$

This Karhunen–Loève-based dimension reduction and the dimension reduction method primarily considered in this paper yield *different approximation formats* for the posterior measure: the latter considers approximate measures that are fully supported on $\mathbb{R}^d$, while the former seeks an approximate measure that is defined on a low-dimensional affine subspace of $\mathbb{R}^d$ (i.e., $m + \mathrm{Im}(P_r)$). The difference in support makes the two approximations hard to compare since, by construction, the divergence $\mathrm{D}_{\mathrm{KL}}(\nu || \tilde{\nu}_r)$ is infinite. This is hardly surprising, as Karhunen–Loève dimension reduction does not try to minimize $\mathrm{D}_{\mathrm{KL}}(\nu || \tilde{\nu}_r)$, but rather focuses on the the mean squared error $\mathbb{E}(\|(X - m) - P_r(X - m)\|^2)$, with $X \sim \mu$.

From a computational perspective, Karhunen–Loève dimension reduction relies only on the prior measure μ: there is no need to compute expectations over the posterior measure ν or to evaluate gradients of the likelihood function, and hence the resulting approximation is easy to compute.

Yet these computational advantages highlight some fundamental limitations of the method. The efficiency of the reduction strategy is limited to cases where there is a sharp decay in the eigenvalues of the covariance matrix Σ; the method does not exploit any low-dimensional structure that the likelihood function might have.

Note that the orthogonal projector onto the leading eigenspace of the prior covariance was already considered in Section **??**, to construct a different kind of *a priori* dimension reduction—one yielding approximations that are again fully supported on $\mathbb{R}^d$. In that case, we obtained an upper bound (**??**) for $\mathrm{D}_{\mathrm{KL}}(\nu||\nu_r^*)$, where $\frac{\mathrm{d}\nu_r^*}{\mathrm{d}\mu} \propto \mathbb{E}_\mu(f|\sigma(P_r))$.

Remark 4.1. It is also possible to apply the Karhunen–Loève dimension reduction method to the posterior measure ν. In that case, we seek an approximation of ν defined as the distribution of $m + P_r(X - m)$ for some projector P_r, where $X \sim \nu$ follows the posterior distribution and $m = \mathbb{E}(X)$ is the posterior mean. As before, the projector can be defined as the minimizer of $P_r \mapsto \mathbb{E}(\|(X - m) - P_r(X - m)\|_2^2)$ for $X \sim \nu$, which turns out to be an orthogonal projector onto the leading eigenspace of the posterior covariance. This approach is no longer an *a priori* method since it requires computing an integral over the posterior distribution.

4.2 Likelihood-informed subspace

The likelihood-informed subspace (LIS) reduction method [12] leverages optimality results available for the Bayesian linear–Gaussian model [42] to propose a structure-exploiting approximation of the posterior distribution. The method assumes the prior to be Gaussian $\mu = \mathcal{N}(m, \Sigma)$ and the likelihood function f, up to a multiplicative constant, to have the following form:

$$f : x \mapsto \exp\Big(-\frac{1}{2}\|G(x) - y\|^2_{\Sigma_{\mathrm{obs}}^{-1}}\Big).$$

Here G is a suitably regular forward operator, y is the observed data vector, and Σ_{obs} is the covariance matrix of the observational noise which is assumed to be additive and Gaussian. The LIS reduction approximates the posterior ν by ν_r^{LIS}, defined as

$$\frac{\mathrm{d}\nu_r^{\mathrm{LIS}}}{\mathrm{d}\mu}(x) \propto f(P_r x + (I_d - P_r)m).$$

Here P_r is a projector that is self–adjoint with respect to the inner product induced by Σ^{-1} and whose range is spanned by the leading generalized eigenvectors of the matrix pencil $(H_{\mathrm{LIS}}, \Sigma^{-1})$, where

$$H_{\mathrm{LIS}} = \int \nabla G^T \, \Sigma_{\mathrm{obs}}^{-1} \, \nabla G \, \mathrm{d}\nu. \tag{38}$$

The matrix H_{LIS} is the expectation over the posterior ν of the Gauss–Newton Hessian of the log-likelihood $\nabla G^T \, \Sigma_{\mathrm{obs}}^{-1} \, \nabla G$. The work of [15] generalizes the construction of the matrix H_{LIS} to reference measures other than the posterior—for example, replacing ν with the prior measure or the Laplace approximation to the posterior.

The LIS methodology differs from the methodology considered in this paper in the following ways. First, the likelihood function is approximated by $x \mapsto f(P_r x + (I_d - P_r)m)$, rather than by the conditional expectation $\mathbb{E}_\mu(f|\sigma(P_r))$. This choice leads to a suboptimal approximation with respect to the Kullback–Leibler divergence; see Section 2.1. Notice however that this approximation choice is quite similar to the Monte Carlo approximation of the conditional expectation presented in Section 3.2. Indeed, using only $M = 1$ sample, the approximation of the conditional expectation defined by (30) is $\widehat{F}_r : x \mapsto f(P_r x + (I_d - P_r)Y)$, where Y is a sample drawn from the prior distribution μ.

The other important difference is in the definition of the projector. Recall that in Section 2.2, P_r is defined as the projector onto the leading eigenspace of the pencil (H, Σ^{-1}), where H is defined in (15). The matrices H and H_{LIS} are in general different, and so are the resulting projectors. The projector P_r in the present paper is defined as the minimizer of an upper bound on the Kullback–Leibler divergence between ν and its approximation (see Section 2.2). In contrast, the projector introduced in [12] is only justified by analogy with optimality results developed in the linear–Gaussian case [42]. As a consequence, the LIS projector does not come with certified error bounds on the resulting posterior approximation, while the strategy presented in this paper does. Yet both methodologies can perform remarkably well in applications, and can be even comparable; cf. Section 5.

4.3 Active subspace for Bayesian inverse problems

The active subspace (AS) method [10, 39] is a dimension reduction technique which addresses the approximation (in the L^2 sense) of a real-valued function by means of a ridge function. These directions span the so-called *active subspace* [41]. In [11], this methodology is used to approximate the log-likelihood function in a Bayesian inverse problem. Denoting the prior covariance by Σ, the posterior distribution ν is approximated by

$$\frac{\mathrm{d}\nu_r^{\mathrm{AS}}}{\mathrm{d}\mu} \propto \exp \mathbb{E}_\mu(\log f | \sigma(P_r)), \tag{39}$$

where P_r is defined as the Σ^{-1}-orthogonal projector onto the leading generalized eigenspace of the matrix pencil $(H_{\mathrm{AS}}, \Sigma^{-1})$, with

$$H_{\mathrm{AS}} = \int (\nabla \log f)(\nabla \log f)^T \, \mathrm{d}\mu. \tag{40}$$

Active subspace reduction differs from the methodology introduced in this paper in the following aspects. First, the conditional expectation in (39) applies to the log-likelihood function rather than to the likelihood function itself. This choice is motivated by the fact that $\mathbb{E}_\mu(\log f | \sigma(P_r))$ yields an optimal approximation of $\log f$ in the set $\{g \circ P_r,\, g : \mathbb{R}^d \to \mathbb{R}\}$ with respect to the L^2_μ-norm. However, the function $\exp \mathbb{E}_\mu(\log f | \sigma(P_r))$ is *not* optimal with respect to the Kullback-Leibler divergence; see Section 2.1

Another difference is in the choice of projector. Comparing (40) with (15), we see that the integral in (40) is taken over the prior μ rather than the posterior ν, and thus the matrices H_{AS} and H are in general different. Notice that H_{AS} corresponds to the matrix $\widetilde{H}$ defined in (37) with the choice $\widetilde{\nu} = \mu$.

Finally, we mention that the active subspace method comes with an upper bound on the Hellinger distance between ν and its approximation; see Theorem 3.1 in [11]. This analysis relies on a Poincaré inequality rather than on a logarithmic Sobolev inequality. Moreover, the Hellinger bound in [11] contains unknown (and uncontrolled) constants. In this paper, we provide sufficient conditions on the prior μ to control the constants associated with the logarithmic Sobolev inequality and hence with our Kullback–Leibler error bound. From a theoretical point of view, it is challenging to relate the two bounds, and we cannot make a definitive statement about their relative merit: comparing upper bounds of two different metrics is not really informative. Instead we will compare the two methodologies by means of numerical experiments in Section 5.

5 Numerical illustration

We use two Bayesian inverse problems to numerically demonstrate various theoretical aspects of the proposed dimension reduction method. In both examples we assume that the prior distribution is Gaussian, $\mu = \mathcal{N}(m, \Sigma)$. This choice of prior distribution satisfies Assumption 2.5 with $\Gamma = \Sigma^{-1}$ and $\kappa = 1$. In both examples, we suppose that the data y correspond to predictions of a nonlinear forward model $x \mapsto G(x)$ that are corrupted by measurement noise, where the latter is additive and normally distributed with zero mean and covariance Σ_{obs}. Thus we have a likelihood function, up to a multiplicative constant, of the form

$$f : x \mapsto \exp\left(-\frac{1}{2}\|G(x) - y\|^2_{\Sigma_{\text{obs}}^{-1}}\right). \tag{41}$$

5.1 Example 1: Atmospheric remote sensing

Our first example is a realistic atmospheric remote sensing problem, where satellite observations from the *Global Ozone MOnitoring System* (GOMOS) are used to estimate the concentration profiles of various gases in the atmosphere.

5.1.1 Problem setup

The GOMOS instrument repeatedly measures light intensities at different wavelengths λ and different altitudes "alt." The light transmissions $T_{\lambda,\text{alt}}$ are modeled using Beer's law:

$$T_{\lambda,\text{alt}} = T_{\lambda,\text{alt}}\left(\kappa^1, \ldots, \kappa^{N_{\text{gas}}}\right) = \exp\left(-\int_{\text{path(alt)}} \sum_{\text{gas}=1}^{N_{\text{gas}}} a_\lambda^{\text{gas}}(z(\zeta))\kappa^{\text{gas}}(z(\zeta))\,\mathrm{d}\zeta\right), \tag{42}$$

where the integral is taken along the ray path, and $z(\zeta)$ is the height in the atmosphere of a point ζ on the path. The curvature of the earth is taken into account; see the illustration in Figure 1. The quantity $a_\lambda^{\text{gas}}(z)$, known from laboratory measurements, is called the cross-section. It is a measure of how much a gas absorbs light of a given wavelength λ at a given height z. In this model there are N_{gas} gases, and for each we would like to infer the density profile $\kappa^{\text{gas}} : z \mapsto \kappa^{\text{gas}}(z)$. Each of the N_{gas} profiles are modeled as independent random processes with log-normal prior distributions. That is, $\log \kappa^{\text{gas}}(\cdot)$ follows a Gaussian distribution $\mathcal{N}(m^{\text{gas}}, \Sigma^{\text{gas}})$ where Σ^{gas} is the covariance operator associated with the squared exponential kernel

$$(z, z') \mapsto \sigma^2_{\text{gas}} \exp\left(-\frac{\|z - z'\|^2}{2z_0^2}\right). \tag{43}$$

These priors are chosen to promote smooth gas density profiles with large variations.

Henceforward, we consider a discretization of the vertical axis z into N_{alt} layers with piecewise constant densities. We denote by $x \in \mathbb{R}^{N_{\text{alt}} \cdot N_{\text{gas}}}$ the vector containing the logarithms of the (unknown) N_{gas} gas densities in the N_{alt} layers, i.e.,

$$x = \text{vec}\begin{pmatrix} \log(\kappa^1(z_1)) & \ldots & \log(\kappa^{N_{\text{gas}}}(z_1)) \\ \vdots & \ddots & \vdots \\ \log(\kappa^1(z_{N_{\text{alt}}})) & \ldots & \log(\kappa^{N_{\text{gas}}}(z_{N_{\text{alt}}})) \end{pmatrix},$$

where $z_1, \ldots, z_{N_{\text{alt}}}$ denote the N_{alt} altitudes and $\text{vec}(\cdot)$ the vectorization operator. Because κ^{gas} is log-normally distributed, the prior distribution $\mu = \mathcal{N}(m, \Sigma)$ of the parameter x is Gaussian, and its prior mean m and its prior covariance Σ are derived from the discretization of m^{gas} and Σ^{gas}.

The data y is a noisy measurement of the light transmissions $T_{\lambda,\text{alt}}$ at N_λ different wavelengths and at N_{alt} different altitudes. It is a vector $y \in \mathbb{R}^{N_\lambda N_{\text{alt}}}$ given by

$$y = G(x) + \varepsilon_{\text{obs}}$$

where the forward model $G(x)$ results from the discretization of Beer's law (42), and where $\varepsilon_{\text{obs}} \sim \mathcal{N}(0, \Sigma_{\text{obs}})$ is centered Gaussian noise with known observation covariance Σ_{obs}. With this model, the likelihood function f takes the form of (41).

Here we adopt the same model setup and synthetic data set used in [12, 15]. The atmosphere is discretized into $N_{\text{alt}} = 50$ layers and with $N_{\text{gas}} = 4$ profiles to infer so the total parameter dimension is $d = N_{\text{alt}} N_{\text{gas}} = 200$. We have observations at $N_\lambda = 1416$ wavelengths, and thus the dimension of the data is $N_{\text{alt}} N_\lambda = 70800$. We refer the readers to [44, 21] for a further description of the model setup, the data set, and the Bayesian treatment of this inverse problem.

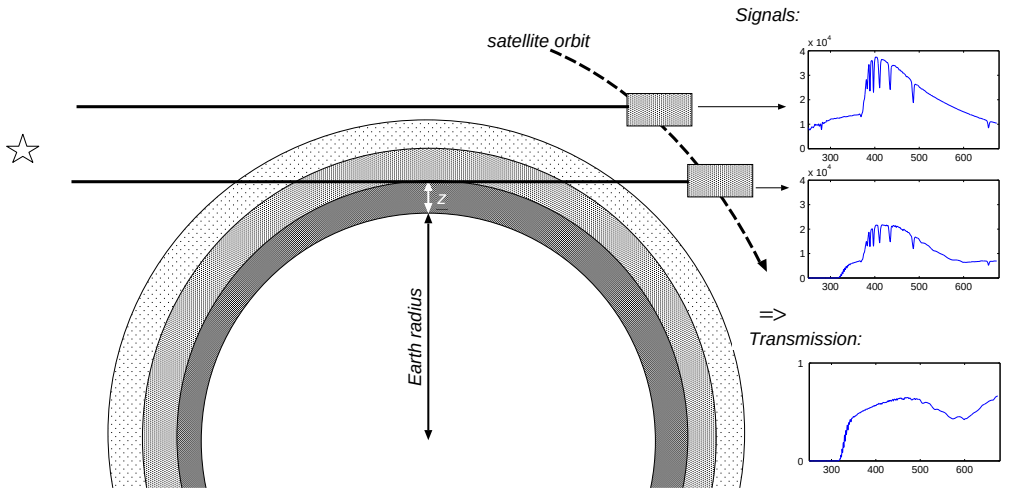

Figure 1: The principle of the GOMOS measurement. The atmosphere is represented as spherical layers around the Earth. Note that the thickness of the layers is much larger relative to the Earth in this figure than in reality. The figure is adopted from [21], with the permission of theauthors.

5.1.2 Comparison of approximations

We compare the posterior approximations obtained by the proposed method (see Sections 2 and 3) to those obtained by the existing methods described in Section 4. The comparison is summarized in Figures 2 and 3.

Except for the prior-based method described in Section **??**, which builds a projector only from the prior covariance matrix Σ, the projector P_r is built with a generalized eigendecomposition of

$$H_\rho^{(\nabla \log f)} = \int (\nabla \log f)(\nabla \log f)^T \, \mathrm{d}\rho \quad \text{or} \quad H_\rho^{(\nabla G)} = \int (\nabla G)^T \Sigma_{\text{obs}}^{-1} (\nabla G) \, \mathrm{d}\rho, \tag{44}$$

where ρ is either the prior μ, the posterior ν, or the Laplace approximation of the posterior ν. Using this new notations, we have $H = H_\nu^{(\nabla \log f)}$ with $\rho = \nu$. Each of the six possible combinations $(\{\nabla \log f; \nabla G\} \times \{\mu; \nu; \text{Laplace}(\nu)\})$ corresponds to a subplot in Figure 2. In all cases, we compute

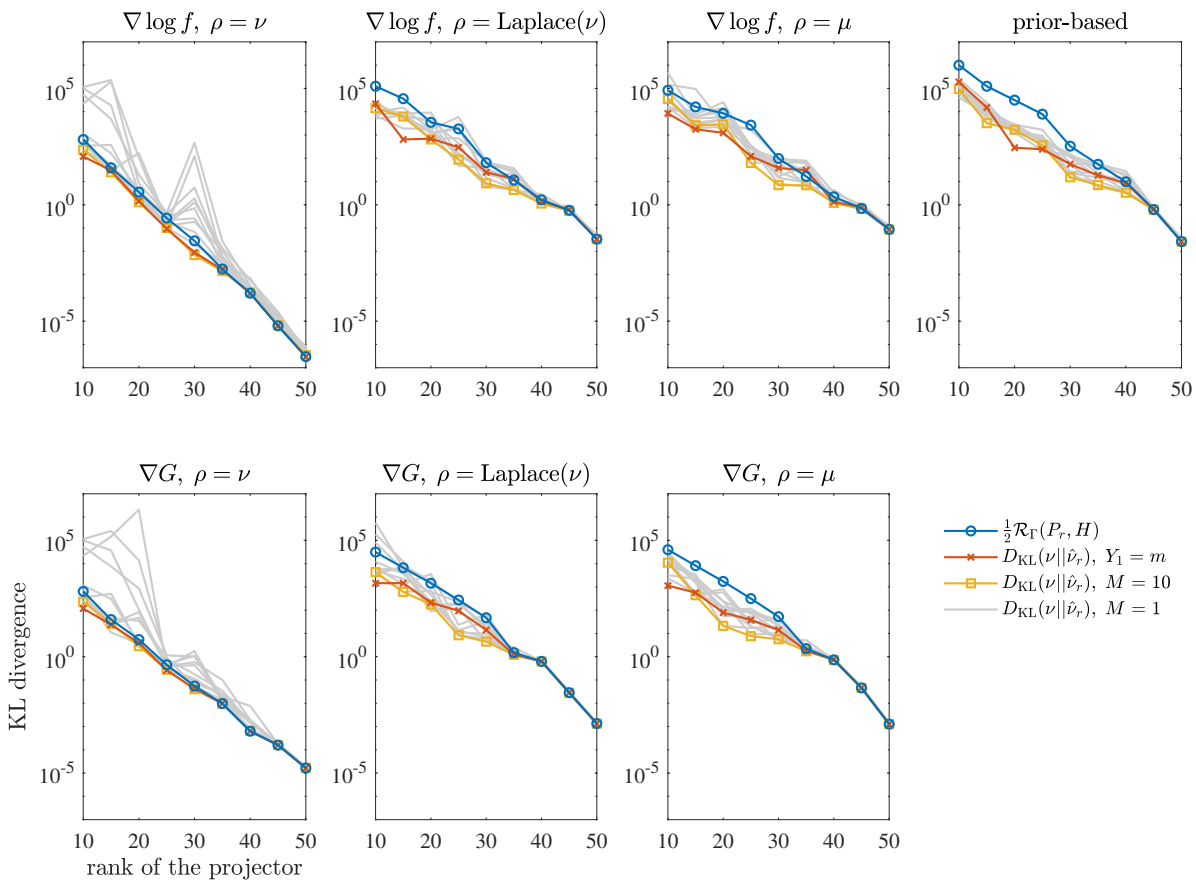

Figure 2: GOMOS example. KL divergence $D_{\mathrm{KL}}(\nu\|\widehat{\nu}_r)$ versus the rank of the projector P_r. The title of each plot (except the top right one) summarizes the combination of integrand and reference measure ρ used to build the projector, see Equation (44). For each projector, we consider three different approximations of the conditional expectation: the prior mean approximation (crosses, $Y_1 = m$), the sample average approximation (squares, $M = 10$), and the one-sample approximation (grey lines, $M = 1$). The reconstruction error $\frac{1}{2}\mathcal{R}_\Gamma(P_r, H)$ is illustrated by circles.

Monte Carlo estimates of the matrices $H_\rho^{(\nabla \log f)}$ or $H_\rho^{(\nabla G)}$ above using $K = 10^6$ samples, to minimize the impact of Monte Carlo error in this first comparison. When $\rho = \nu$, the DILI algorithm [14] is used to obtain samples from the posterior distribution.

Once we have estimated the matrix $H_\rho^{(\nabla \log f)}$ or $H_\rho^{(\nabla G)}$ for a $\rho \in \{\mu; \nu; \text{Laplace}(\nu)\}$, we compute its generalized eigendecomposition (i.e., $(H_\rho^{(\nabla \log f)}, \Sigma^{-1})$ or $(H_\rho^{(\nabla G)}, \Sigma^{-1})$) and assemble the rank-$r$ projector P_r onto the corresponding leading eigenspace for $r \in \{10; 15; \ldots; 50\}$. Then we approximate the conditional expectation $\mathbb{E}_\mu(f|\sigma(P_r))$ by

$$x \mapsto \frac{1}{M}\sum_{i=1}^{M} f(P_r x + (I_d - P_r)Y_i), \qquad \text{where } Y_1, \ldots, Y_M \overset{\text{iid}}{\sim} \mu,$$

with either $M = 10$ prior samples (yellow line with squares in Figure 2) or with $M = 1$ prior sample (solid grey lines). We also consider the deterministic approximation $x \mapsto f(P_r x + (I_d - P_r)m)$, which

corresponds to using $M = 1$ sample that is fixed to the prior mean, $Y_1 = m$ (red line with crosses). Finally, the blue line with circles represents the reconstruction error $\mathcal{R}_\Gamma(P_r, H)$, where P_r is the projector constructed either from $H_\rho^{(\nabla \log f)}$ or $H_\rho^{(\nabla G)}$, and where $H = \int (\nabla \log f)(\nabla \log f)^T \mathrm{d}\nu$.

Once both the projector and the approximation of the conditional expectation are defined, the KL divergence $D_{\mathrm{KL}}(\nu \| \widehat{\nu}_r)$ is approximated with 10^6 samples drawn from the posterior using the DILI algorithm [14]. This approximation is assumed to be sufficiently accurate for the purpose of our experiments.

Among the different ways of approximating the conditional expectation, the sample-based approach with $M = 10$ outperforms both the one-sample approach ($M = 1$) and the deterministic approach ($Y_1 = m$). As shown in the theoretical analysis of Section 3.2, the error due to approximating the conditional expectation is, in expectation, proportional to the reconstruction error; this is why we observe that the error decays quickly with the rank of the projector even when M is small. Note that the cost for evaluating the approximation of the conditional expectation is proportional to M: it is exactly M times the cost of evaluating the exact likelihood function f, which might be expensive when M is large. Therefore we would prefer using the deterministic prior mean approach ($Y_1 = m$) which, in addition to having a low evaluation cost, has an accuracy comparable to that of the sample-based approach with $M = 10$ and seems better than the one-sample approach ($M = 1$).

Figure 2 also shows that the reconstruction error $\mathcal{R}_\Gamma(P_r, H)$ provides a good estimate of the KL divergence, a quantity which is rarely available in practice. In theory, the reconstruction error gives a certified bound on the KL divergence for any projector, provided that the conditional expectation is computed exactly—which is not what we do here. In practice, however, we observe that the reconstruction error is in general a fairly good error indicator even when using a sample approximation of the conditional expectation. This might not be true when sample sizes are very small; see, for instance, the cases $\rho = \nu$ and $M = 1$ in Figure 2.

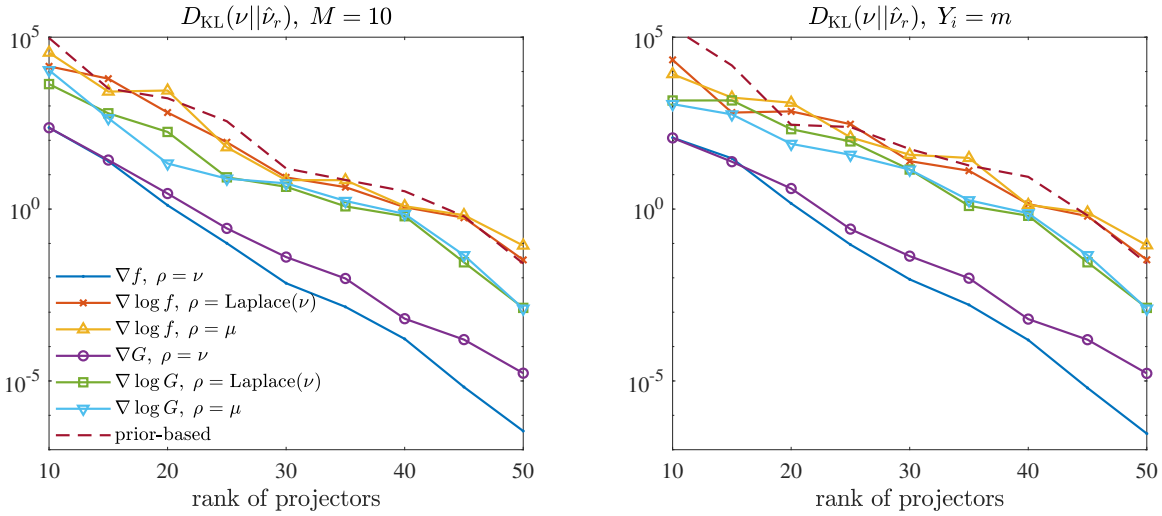

Figure 3: GOMOS example. KL divergence $D_{\mathrm{KL}}(\nu \| \widehat{\nu}_r)$ for posterior approximations $\widehat{\nu}_r$ obtained using various projectors P_r. The left plot shows results when a sample average ($M = 10$) is used to approximate the conditional expectation, while the right plot shows results using the prior mean approximation ($Y_1 = m$).

In Figure 3 we compare the performance of the different definitions of the projector. Here we only consider the KL divergence of posterior approximations obtained using the sample-based

approximation of the conditional expectation with $M = 10$ (left) and the prior mean approximation $Y_1 = m$ (right). The projector obtained from $H_\rho^{(\nabla \log f)}$ with $\rho = \nu$ achieves the best accuracy overall. The LIS method, obtained from $H_\rho^{(\nabla \log G)}$ with $\rho = \nu$, also performs reasonably well. We observe a performance gap between these two projectors and the other five: the latter lead to significantly less accurate posterior approximations, at any given rank.

5.1.3 Demonstration of Algorithm 2

Now we demonstrate the iterative procedure given in Algorithm 2 for constructing projectors and posterior approximations. In this exercise, we approximate the conditional expectation using the prior mean option ($Y_1 = m$). For each iteration l, the rank of the projector, the reconstruction error $\mathcal{R}_\Gamma(P_r^{(l)}, H)$, and the KL divergence from the posterior measure ν to the resulting approximation $\widehat{\nu}_r^{(l)}$, are shown in Figure 4. The left column of Figure 4 illustrates the iterative procedure where the error threshold $\varepsilon = 10^{-2}$ is used to determine the rank of the projector in each iteration. Here we also set the maximum rank of the projector to be $r_{\max} = 40$. The right column of Figure 4 shows a fixed-rank variant of the iterative procedure. In this variant, the rank of the projector is held constant at $r = 30$.

The iterative procedure appears to be effective in approximating the posterior. When the rank of the projector is dynamically adjusted (left column of Figure 4), the iterative procedure achieves the desired level of accuracy within the first two iterations; then the rank of the projector, the reconstruction error, and the KL divergence are stabilized in later iterations. When the rank of the projector is fixed (right column of Figure 4), the KL divergence decays significantly in the first two iterations, and then is stabilized in later iterations.

5.1.4 Impact of sample size

We now analyze the impact of the sample size K in the Monte Carlo approximation $\widehat{H}$ of H; see Section 3.1. Recall that $\widehat{H}$ is used to construct the projector $\widehat{P}_r$ that minimizes the *approximate* reconstruction error $P_r \mapsto \mathcal{R}_\Gamma(P_r, \widehat{H})$ over the set of all rank-r projectors. We will measure the impact of the sample size K via the following criteria:

1. Is $\mathcal{R}_\Gamma(\widehat{P}_r, H)$ close to the minimum of the true reconstruction error $\mathcal{R}_\Gamma(P_r^*, H)$? Since our primary goal is to minimize a bound on the KL divergence, it makes sense to measure the quality of $\widehat{P}_r$ in terms of its ability to minimize the reconstruction error. Thus we do not look at distances between $\widehat{P}_r$ and the minimizer P_r^* (such as the operator norm $\|\widehat{P}_r - P_r^*\|$ or the Frobenius norm $\|\widehat{P}_r - P_r^*\|_F$) because these distances are not directly related to the problem of posterior approximation.

2. Can we use $\mathcal{R}_\Gamma(\widehat{P}_r, \widehat{H})$ as an error indicator for the KL divergence? Since the quantity $\mathcal{R}_\Gamma(\widehat{P}_r, \widehat{H})$ is the only one that is accessible in practice, we would like to know whether it is safe to use as an error estimator.

Numerical results are summarized in the left plot of Figure 5 for different sample sizes $K \in \{50, 100, 200, 500, 1000, 2000\}$. To assess the variability of these results due to the randomness in $\widehat{H}$, we also repeat the previous experiment 10 times for $K \in \{50, 200, 1000\}$ (right plot of Figure 5).

Concerning the first criterion, we observe that the quality of $\widehat{P}_r$ depends not only on the sample size K, but also on the rank r of the projector. Indeed, $\mathcal{R}_\Gamma(\widehat{P}_r, H)$ (solid colored lines in Figure 5) is closest to the minimum of the reconstruction error (black line) for small r and large K. This

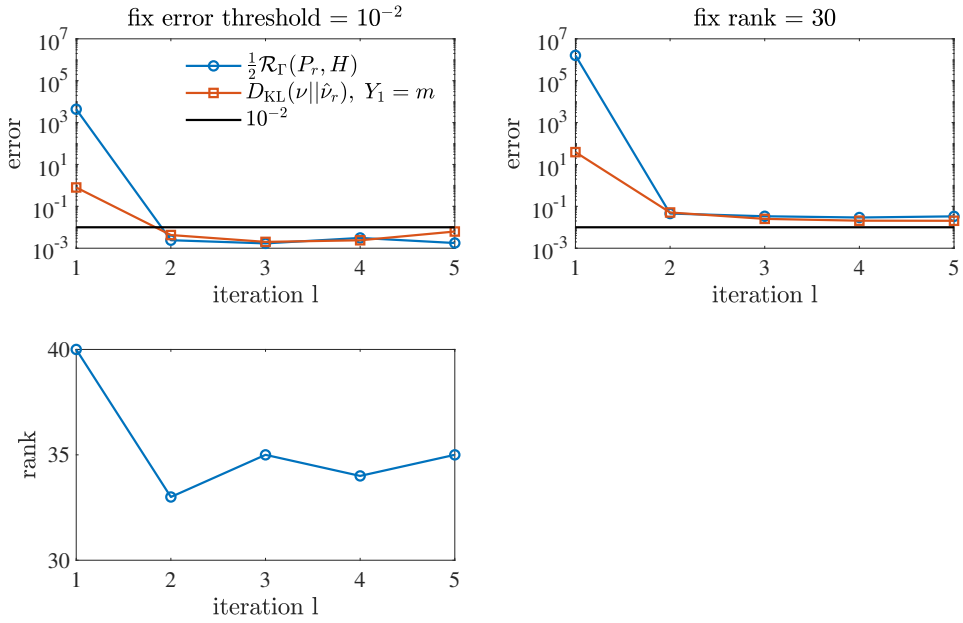

Figure 4: GOMOS example. Rank of the projector, reconstruction error $\mathcal{R}_\Gamma(P_r^{(l)}, H)$, and KL divergence from the exact posterior measure ν to the approximate measure $\widehat{\nu}_r^{(l)}$ produced by the iterative procedure (Algorithm 2). The left column shows results where a target target error tolerance $\varepsilon = 10^{-2}$ and $r_{\max} = 40$ is used to determine the rank of the projector. The right column shows a fixed-rank variant of the iterative procedure, where the rank is held at $r = 30$.

suggests that the sample size should be chosen larger when r is chosen to be large. Note that this fact is not revealed by the theoretical analysis of Section 3.1.

Concerning the second criterion, we observe that the approximate reconstruction error $\mathcal{R}_\Gamma(\widehat{P}_r, \widehat{H})$ (dashed colored lines) is always smaller than the true reconstruction error $\mathcal{R}_\Gamma(\widehat{P}_r, H)$ (solid colored lines). This relationship is particularly apparent when the sample size K is small and the rank r is large. This means that one should be careful when using $\mathcal{R}_\Gamma(\widehat{P}_r, \widehat{H})$ in place of the exact reconstruction error, as it tends to underestimate the error.

Overall, in this example, both criteria are validated for sufficiently large sample sizes. With $K \geq 1000$, the resulting projector $\widehat{P}_r$ has a reconstruction error comparable to that of the projector P_r^*, and the approximate reconstruction error $\mathcal{R}_\Gamma(\widehat{P}_r, \widehat{H})$ provides a fairly good estimate for $\mathcal{R}_\Gamma(\widehat{P}_r, H)$.

5.2 Example 2: Elliptic PDE

Our second example is an inverse problem aiming at estimating the spatially inhomogeneous coefficient of an elliptic PDE, adopted from [15]. In physical terms, our problem setup corresponds to inferring the transmissivity field of a two-dimensional groundwater aquifer from partial observations of the stationary drawdown field of the water table, measured from well bores.

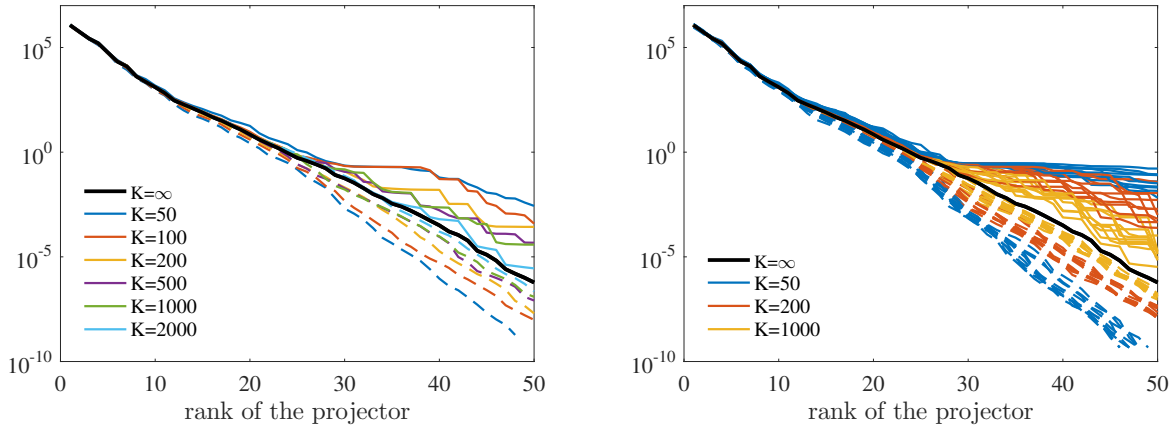

Figure 5: GOMOS example. The approximate reconstruction error $\mathcal{R}_\Gamma(\widehat{P}_r, \widehat{H})$ (dashed lines) and the true reconstruction error $\mathcal{R}_\Gamma(\widehat{P}_r, H)$ (solid lines) for projectors $\widehat{P}_r$ obtained using different sample sizes K. Different sample sizes are represented by lines marked with different colors. The solid black line represents the reconstruction error $\mathcal{R}_\Gamma(P_r^*, H)$. Left: errors obtained in single-trial experiments. Right: errors obtained in 10 repeated experiments. Here, the reference matrix H is computed using 10^6 posterior samples generated by the DILI algorithm.

5.2.1 Problem setup

Consider a three kilometer by one kilometer problem domain $\Omega = [0\,\mathrm{m}, 3000\,\mathrm{m}] \times [0\,\mathrm{m}, 1000\,\mathrm{m}]$, with boundary $\partial\Omega$. We denote the spatial coordinate by $\zeta \in \Omega$. Consider the transmissivity field $T(\zeta)$ (units [m^2/day]), the drawdown field $p(\zeta)$ (units [m]), and sink/source terms $q(\zeta)$ (units [m/day]). The drawdown field for a given transitivity and source/sink configuration is governed by the elliptic equation:

$$-\nabla \cdot (T(\zeta)\nabla p(\zeta)) = q(\zeta), \quad \zeta \in \Omega. \tag{45}$$

We prescribe the drawdown value to be zero on the boundary (i.e., a zero Dirichlet boundary condition), and define the source/sink term $q(\zeta)$ as the superposition of four weighted Gaussian plumes with standard width 50 meters. The plumes are centered near the four corners of the domain (at $[20\,\mathrm{m}, 20\,\mathrm{m}]$, $[2980\,\mathrm{m}, 20\,\mathrm{m}]$, $[2980\,\mathrm{m}, 980\,\mathrm{m}]$ and $[20\,\mathrm{m}, 980\,\mathrm{m}]$) with magnitudes of -3000, 2000, 4000, and -3000 [m/day], respectively. We solve (45) by the finite element method.

The discretized transitivity field $T(\zeta)$ is endowed with a log-normal prior distribution, i.e.,

$$T(\zeta) = \exp(x(\zeta)), \quad \text{where} \quad x \sim \mathcal{N}(m, \Sigma),$$

where the prior mean is set to $\log(1000\,[\mathrm{m/day}])$ and the inverse of the covariance matrix Σ^{-1} is defined through the discretization of an Laplace-like stochastic partial differential equation [30],

$$(-\nabla \cdot K\nabla + \kappa^2)x(\zeta) = \mathcal{W}(\zeta),$$

where $\mathcal{W}(\zeta)$ is white noise. In this example, we set the stationary, anisotropic correlation tensor K to

$$K = \begin{bmatrix} 0.55 & -0.45 \\ -0.45 & 0.55 \end{bmatrix},$$

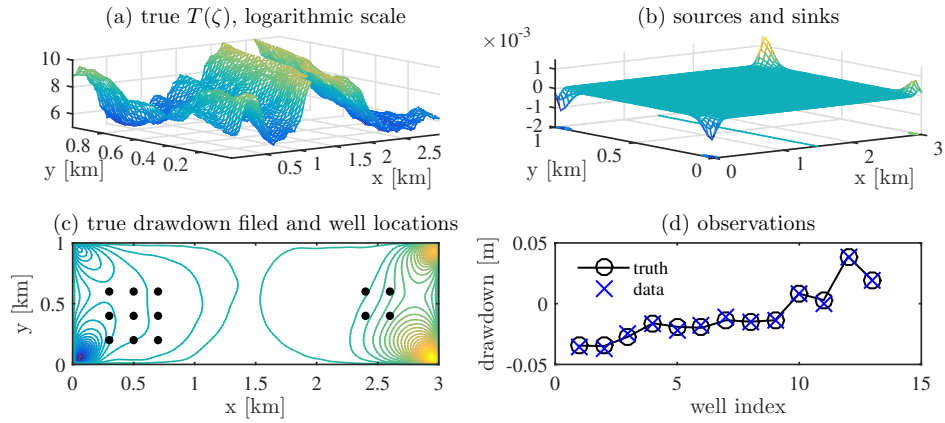

Figure 6: Setup of the groundwater inversion example. (a) "True" transmissivity field. (b) Sources and sinks. (c) Drawdown field resulting from the true transmissivity field, with observation wells indicated by black dots. (d) Data y; circles represent the noise-free drawdowns at each well, while crosses represent the observed drawdowns with measurement noise.

and put $\kappa = 50$. The "true" transmissivity field is a realization from the prior distribution. The true transmissivity field, sources/sinks, simulated drawdown field, and synthetic data are shown in Figure 6. Partial observations of the pressure field are collected at 13 sensors whose locations are depicted by black dots in Figure 6(c), with additive error $e \sim \mathcal{N}(0, \sigma^2 I_{13})$. The standard deviation σ of the measurement noise is prescribed so that the observations have signal-to-noise ratio 20. The noisy data are shown in Figure 6(d).

In this example, the finite element discretization of (45) uses 120×40 bilinear elements to represent the drawndown field $p(\zeta)$, while the transmissivity field $T(\zeta)$ is modeled as piecewise constant for each element. This yields the discretized forward model $G : \mathbb{R}^{4800} \to \mathbb{R}^{13}$, where the parameter x is of dimension $d = 4800$.

5.2.2 Numerical results

We proceed with the same comparison as in the GOMOS example. We compare the KL divergence $D_{\mathrm{KL}}(\nu \| \widehat{\nu}_r)$ for different posterior approximations $\widehat{\nu}_r$ obtained using the various projectors given in Sections 3 and 4. The projectors used in this comparison are built in the same way as in the GOMOS example. Also we consider the same three options ($M = 1$, $M = 10$, and $Y_1 = m$) for approximating the conditional expectation within $\widehat{\nu}_r$. Again, 10^6 samples drawn from the posterior using the DILI algorithm [14] are used to compute the KL divergence $D_{\mathrm{KL}}(\nu \| \widehat{\nu}_r)$. The comparison is summarized in Figures 7 and 8.

In Figure 7 we observe that the approximation of the conditional expectation with $M = 10$ samples outperforms the two other options, $M = 1$ and $Y_1 = m$. This holds true regardless of how the projector is constructed. We also notice that the prior-mean approximation $Y_1 = m$ in general performs better than the one-sample approximation $M = 1$. Similar to the GOMOS example, the reconstruction error $\mathcal{R}_\Gamma(P_r, H)$ provides an effective error indicator for posterior approximations that employ the $M = 10$-sample estimate of the conditional expectation, for any choice of projector P_r. Also, the KL divergence decays with r at the same rate as the reconstruction error, independently of the choice of approximation scheme for the conditional expectation.

A direct comparison of the posterior approximations defined by different projectors is shown in

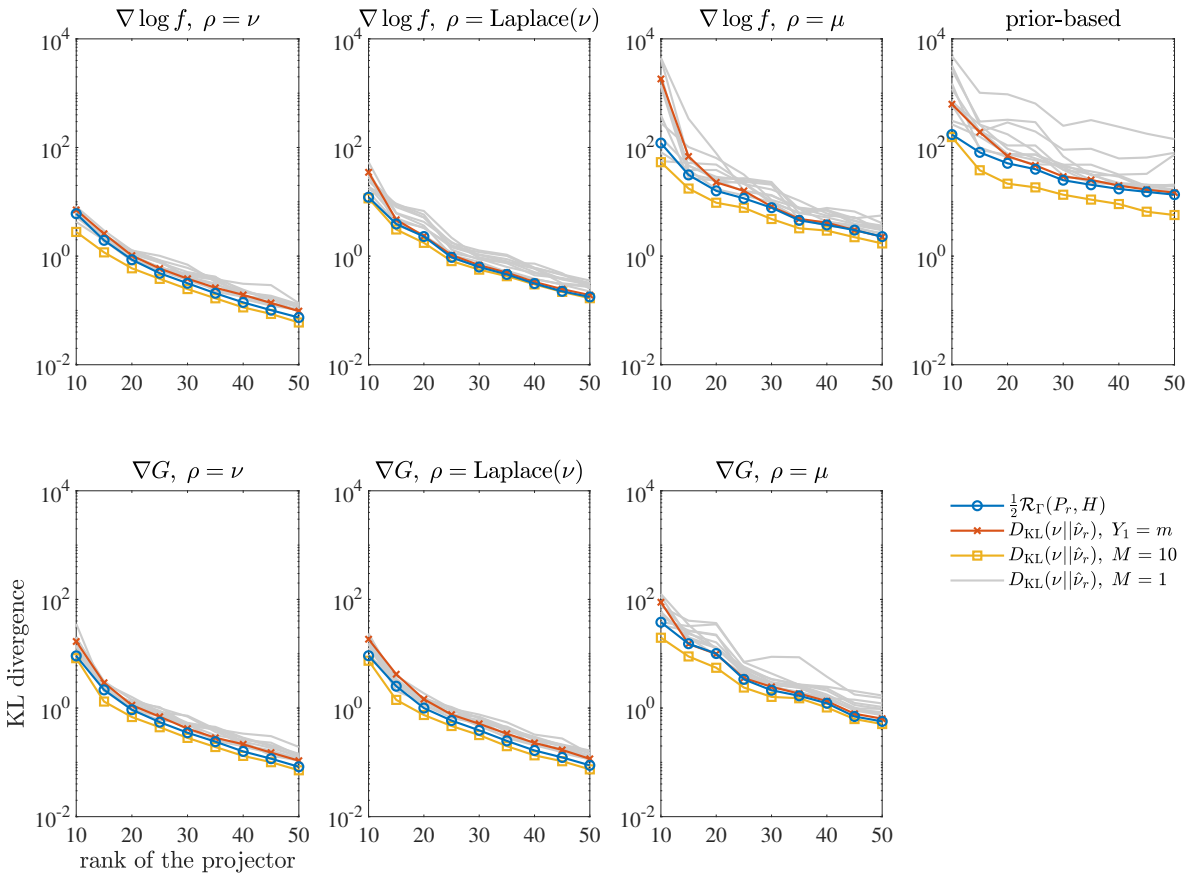

Figure 7: Same as Figure 2, but for the elliptic inverse problem.

Figure 8, where approximations using the sample estimate ($M = 10$) and prior mean approximation ($Y_1 = m$) of the conditional expectation are collected in the left and right plots, respectively. The projector obtained from $H_\rho^{(\nabla \log f)}$ with $\rho = \nu$ outperforms all the other methods. In other words—and as observed in the previous example—the method proposed in Section 2 provides more effective dimension reduction than the alternatives. Note also that the projectors obtained from $H_\rho^{(\nabla G)}$ with $\rho = \nu$ and with $\rho = \text{Laplace}(\nu)$ perform better than the projector obtained from $H_\rho^{(\nabla \log f)}$ with $\rho = \text{Laplace}(\nu)$, indicating that efficient dimension reduction can also be obtained from gradient of the forward model itself. All other projectors have a rather large accuracy gap relative to the abovementioned projectors. In particular, choosing $\rho = \mu$ in this example seems to be much less effective than averaging over the posterior distribution or its Laplace approximation.

6 Conclusion

We have addressed the problem of reducing the dimension of a Bayesian inverse problem, in the nonlinear/non-Gaussian setting. A Bayesian inverse problem has a low intrinsic dimension when the update from the prior distribution to the posterior distribution is essentially low-dimensional, meaning that the data only inform a few directions in the parameter space. We proposed a methodology that reveals and exploits such structure by seeking an approximation of the likelihood as a

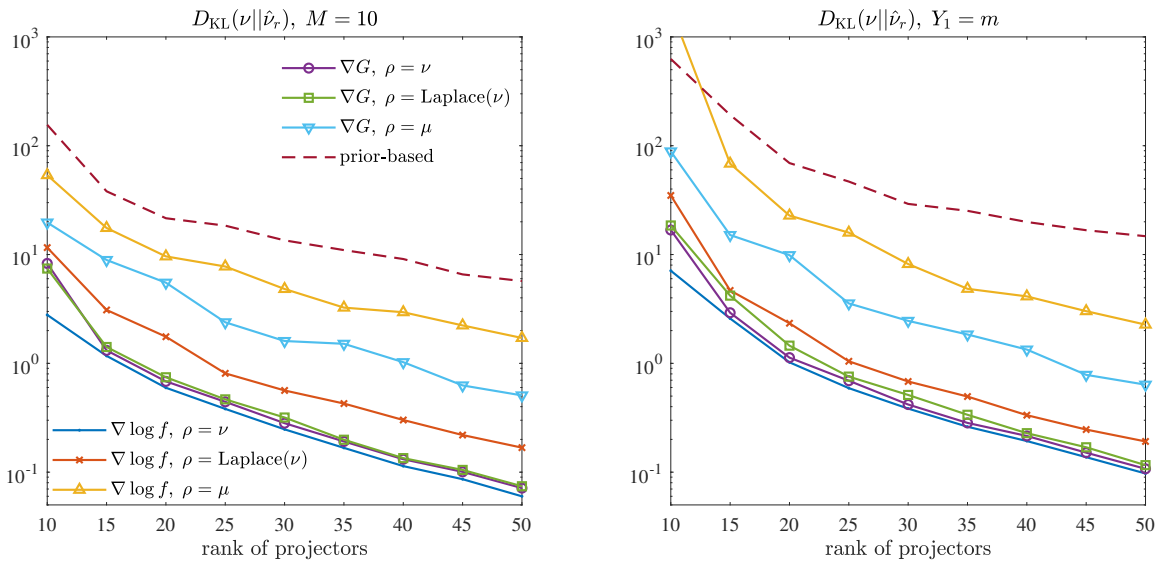

Figure 8: Same as Figure 3, but for the elliptic inverse problem.

ridge function, i.e., a function that varies only on a low-dimensional subspace of its input space. To obtain this approximation, we first identified the optimal profile function of the ridge approximation, meaning the profile that minimizes the Kullback–Leibler divergence from the posterior approximation to the exact posterior. Then, using logarithmic Sobolev inequalities, we derived an upper bound for the remaining error. This bound admits a simple form and can be easily minimized: this is how the informed directions are discovered. In addition, our methodology provides a certified estimator of the approximation error in the posterior distribution, measured with the Kullback–Leibler divergence, a quantity which is difficult to compute in practice.

Our method is fundamentally gradient-based and requires the second moment matrix of the gradient of log-likelihood function. Computing this matrix can be challenging in practice because it entails integrating over the posterior distribution. We thus propose several sample-based approximation schemes, including an iterative algorithm which employs a sequence of low-dimensional posterior approximations.

The use of logarithmic Sobolev inequalities requires some (rather strong) assumptions on the prior distribution, e.g., being Gaussian, being a Gaussian mixture, or being a bounded perturbation of a Gaussian. One open question left for future work is how to weaken the assumption on the prior which, for instance, does not encompass priors with heavy tails. Another possible improvement of the methodology is the extension to infinite-dimensional parameter spaces.

Analytical and numerical examples demonstrate good performance of the proposed method. In particular, we show that it outperforms other state-of-the-art gradient-based dimension reduction schemes. We also note that effective MCMC algorithms for large-scale Bayesian inverse problems, such as the DILI sampler of [14], fundamentally rely on the low-dimensional structure of the posterior distribution. In its original form, [14] uses the LIS method to discover this low-dimensional structure, a method which is less efficient than the one we propose. Incorporating our new developments in the DILI algorithm could thus yield better sampling performance.

7 Acknowledgements

O. Zahm, A. Spantini, Y. Marzouk, and K. Law acknowledge support from the DARPA Enabling Quantification of Uncertainty in Physical Systems (EQUiPS) program, F. Fahroo and J. Vandenbrande, program managers, and from the Inria Associate team UNQUESTIONNABLE. K. Law also acknowledges support from the Oak Ridge National Laboratory Directed Research and Development program and from The Alan Turing Institute under the EPSRC grant EP/N510129/1. O. Zahm and T. Cui acknowledge support from the Robert Bartnik Fellowship offered by the School of Mathematical Sciences at Monash University.

A Proof of Proposition 2.4

We denote by $U_r \in \mathbb{R}^{d\times r}$ a matrix whose columns form a basis of $\mathrm{Im}(P_r)$ and we let $U \in \mathbb{R}^{d\times d}$ be defined as the horizontal concatenation of U_r and $U_\perp$. We have $U\xi = U_r\xi_r + U_\perp\xi_\perp$ for any $\xi \in \mathbb{R}^d$, where $\xi_r \in \mathbb{R}^r$ and $\xi_\perp \in \mathbb{R}^{d-r}$ are the vectors containing respectively the r first and the $d-r$ last components of ξ. Let $F : x \mapsto \int_{\mathbb{R}^{d-r}} f(P_r x + U_\perp\xi_\perp)\ p_\perp(\xi_\perp|P_r x)\mathrm{d}\xi_\perp$. By definition of $p_\perp(\xi_\perp|P_r x)$ and since $P_r U_r = U_r$, this function satisfies

$$F(U_r\xi_r) = \frac{\int_{\mathbb{R}^{d-r}} f(U_r\xi_r + U_\perp\xi_\perp)\ \rho(U_r\xi_r + U_\perp\xi_\perp)\mathrm{d}\xi_\perp}{\int_{\mathbb{R}^{d-r}} \rho(U_r\xi_r + U_\perp\xi'_\perp)\mathrm{d}\xi'_\perp},$$

for any $\xi_r \in \mathbb{R}^r$. Let h be $\sigma(P_r)$-measurable function. By Corollary 2.2, h can be written as the composition of a function with P_r, so that $h(U_r\xi_r + U_\perp\xi_\perp) = h(U_r\xi_r)$ holds for any $\xi_r \in \mathbb{R}^r$ and $\xi_\perp \in \mathbb{R}^{d-r}$. Thus we have

$$\begin{aligned}
\int f\ h\ \mathrm{d}\mu &= \int_{\mathbb{R}^d} f(x)h(x)\rho(x)\mathrm{d}x = \int_{\mathbb{R}^d} f(U\xi)h(U\xi)\rho(U\xi)|U|\mathrm{d}\xi \\
&= \int_{\mathbb{R}^r}\int_{\mathbb{R}^{d-r}} f(U_r\xi_r + U_\perp\xi_\perp)h(U_r\xi_r)\rho(U_r\xi_r + U_\perp\xi_\perp)|U|\mathrm{d}\xi_\perp\mathrm{d}\xi_r \\
&= \int_{\mathbb{R}^r}\left(\int_{\mathbb{R}^{d-r}} f(U_r\xi_r + U_\perp\xi_\perp)\rho(U_r\xi_r + U_\perp\xi_\perp)\mathrm{d}\xi_\perp\right) h(U_r\xi_r)|U|\mathrm{d}\xi_r \\
&= \int_{\mathbb{R}^r}\left(F(U_r\xi_r)\int_{\mathbb{R}^{d-r}} \rho(U_r\xi_r + U_\perp\xi'_\perp)\mathrm{d}\xi'_\perp\right) h(U_r\xi_r)|U|\mathrm{d}\xi_r \\
&= \int_{\mathbb{R}^r}\int_{\mathbb{R}^{d-r}} F(U_r\xi_r)h(U_r\xi_r)\rho(U_r\xi_r + U_\perp\xi'_\perp)|U|\mathrm{d}\xi'_\perp\mathrm{d}\xi_r \\
&= \int_{\mathbb{R}^d} F(x)h(x)\rho(x)\mathrm{d}x = \int F\ h\ \mathrm{d}\mu,
\end{aligned}$$

where $|U|$ denotes the determinant of U. This shows that F satisfies (4). Since it is a $\sigma(P_r)$-measurable function (as the composition of a function with P_r), then it is the conditional expectation $\mathbb{E}_\mu(f|\sigma(P_r))$.

B Proof of Theorem 2.9

Assumption 2.5 implies $\frac{\mathrm{d}\mu}{\mathrm{d}x} = \rho \propto e^{-(V+\Psi)}$ where $V : K \mapsto \mathbb{R}$ and $\Psi : K \mapsto \mathbb{R}$ satisfy respectively (10) and (11). Let $P_r \in \mathbb{R}^{d\times d}$ be a rank-r projector and let $U_\perp \in \mathbb{R}^{d\times(d-r)}$ be a matrix whose columns form a basis of $\mathrm{Ker}(P_r)$. By convexity of K we have $P_r x \in K$ for any $x \in K$. For

any $x \in K$ we denote by $K_\perp(x) = \{\xi_\perp \in \mathbb{R}^{d-r} : U_\perp \xi_\perp + P_r x \in K\} \subset \mathbb{R}^{d-r}$ the pre-image of K under $\xi_\perp \mapsto U_\perp \xi_\perp + P_r x$. The conditional density $p_\perp(\cdot|P_r x)$ defined by (8) is such that $p_\perp(\cdot|P_r x) \propto e^{-(V_\perp + \Psi_\perp)}$ where $V_\perp$ and $\Psi_\perp$ are functions on $K_\perp(x)$ given by $V_\perp(\xi_\perp) = V(P_r x + U_\perp \xi_\perp)$ and $\Psi_\perp(\xi_\perp) = \Psi(P_r x + U_\perp \xi_\perp)$ for any $\xi_\perp \in K_\perp(x)$. The function $V_\perp$ inherits the convexity property of V: using chain rule derivative, we have $\nabla^2 V_\perp(\xi_\perp) = U_\perp^T \nabla^2 V(P_r x + U_\perp \xi_\perp) U_\perp$ so that

$$\nabla^2 V_\perp(\xi_\perp) \succeq U_\perp^T \Gamma U_\perp \succ 0,$$

for any $\xi_\perp \in K_\perp(x)$. Also, $\Psi_\perp$ is a bounded function which satisfies

$$e^{\sup \Psi_\perp - \inf \Psi_\perp} \leq e^{\sup \Psi - \inf \Psi} \leq \kappa.$$

Then the distribution on $K_\perp(x)$ with Lebesgue density $p_\perp(\cdot|P_r x)$ satisfies Assumption 2.5. Then it satisfies the logarithmic Sobolev inequality (12), meaning that

$$\begin{aligned} \int g(\xi_\perp)^2 \log \frac{g(\xi_\perp)^2}{\int g(\xi'_\perp)^2 p_\perp(\xi'_\perp|P_r x) \mathrm{d}\xi'_\perp} \, p_\perp(\xi_\perp|P_r x) \mathrm{d}\xi_\perp \\ \leq 2\kappa \int \|\nabla g(\xi_\perp)\|_*^2 \, p_\perp(\xi_\perp|P_r x) \mathrm{d}\xi_\perp, \end{aligned} \tag{46}$$

holds for any function g with sufficient regularity, where the norm $\|\cdot\|_*$ is defined by $\|v\|_*^2 = v^T (U_\perp^T \Gamma U_\perp)^{-1} v$ for any $v \in \mathbb{R}^{d-r}$. Inequality (46) holds in particular when the function g is defined by $g(\xi_\perp) = h(P_r x + U_\perp \xi_\perp)$ for any $\xi_\perp \in K_\perp(x)$, where h is any function on $\mathbb{R}^d$ with sufficient regularity. By chain rule derivative we have $\nabla g(\xi_\perp) = U_\perp^T \nabla h(P_r x + U_\perp \xi_\perp)$ so that (46) can be written as

$$\begin{aligned} \int h(P_r x + U_\perp \xi_\perp)^2 \log \Big(\frac{h(P_r x + U_\perp \xi_\perp)^2}{\int h(P_r x + U_\perp \xi'_\perp)^2 p_\perp(\xi'_\perp|P_r x) \mathrm{d}\xi'_\perp} \Big) p_\perp(\xi_\perp|P_r x) \mathrm{d}\xi_\perp \\ \leq 2\kappa \int \|U_\perp^T \nabla h(P_r x + U_\perp \xi_\perp)\|_*^2 \, p_\perp(\xi_\perp|P_r x) \mathrm{d}\xi_\perp. \end{aligned}$$

Note that the denominator in the logarithm is the conditional expectation $\mathbb{E}_\mu(h^2|\sigma(P_r))$ evaluated at x, see Proposition 2.4. Let $U_r \in \mathbb{R}^{d \times r}$ be a matrix whose columns form a basis of $\mathrm{Im}(P_r)$, and let $\xi_r \in K_r(x) = \{\xi_r \in \mathbb{R}^r : U_r \xi_r + (I_d - P_r)x \in K\} \subset \mathbb{R}^r$. Replacing x by $U_r \xi_r$ in the above inequality and by definition of the conditional density (8), we obtain

$$\begin{aligned} \int h(U_r \xi_r + U_\perp \xi_\perp)^2 \log \Big(\frac{h(U_r \xi_r + U_\perp \xi_\perp)^2}{\mathbb{E}_\mu(h^2|\sigma(P_r)(U_r \xi_r)} \Big) \rho(U_r \xi_r + U_\perp \xi_\perp) \mathrm{d}\xi_\perp \\ \leq 2\kappa \int \|U_\perp^T \nabla h(U_r \xi_r + U_\perp \xi_\perp)\|_*^2 \, \rho(U_r \xi_r + U_\perp \xi_\perp) \mathrm{d}\xi_\perp. \end{aligned}$$

By integrating over ξ_r we obtain

$$\int h^2 \log \frac{h^2}{\mathbb{E}_\mu(h^2|\sigma(P_r))} \, \mathrm{d}\mu \leq 2\kappa \int_{\mathbb{R}^d} \|U_\perp^T \nabla h\|_*^2 \, \mathrm{d}\mu.$$

To conclude the proof, it remains to show that $\|U_\perp^T \nabla h\|_*^2 \leq \|(I_d - P_r)^T \nabla h\|_{\Gamma^{-1}}^2$ holds. By definition of $U_\perp$ we have $(I_d - P_r) U_\perp = U_\perp$ so that $\|U_\perp^T \nabla h\|_* = \|U_\perp^T (I_d - P_r)^T \nabla h\|_*$. Also, note that $\|\cdot\|_*$

is the dual norm of the norm $\|\cdot\|_{U_\perp^T\Gamma U_\perp}$ defined by $\|\cdot\|^2_{U_\perp^T\Gamma U_\perp} = (\cdot)^T U_\perp^T \Gamma U_\perp(\cdot)$, meaning that $\|v\|_* = \sup\{|\xi_\perp^T v| : \xi_\perp \in \mathbb{R}^{d-r}, \|\xi_\perp\|_{U_\perp^T\Gamma U_\perp} \leq 1\}$ for any $v \in \mathbb{R}^{d-r}$. Thus we have

$$\begin{aligned}
\|U_\perp^T \nabla h\|_* = \|U_\perp^T (I_d - P_r)^T \nabla h\|_* &= \sup_{0\neq \xi_\perp \in \mathbb{R}^{d-r}} \frac{|(U_\perp \xi_\perp)^T (I_d - P_r)^T \nabla h|}{\|U_\perp \xi_\perp\|_\Gamma} \\
&\leq \sup_{0\neq x \in \mathbb{R}^d} \frac{|x^T (I_d - P_r)^T \nabla h|}{\|x\|_\Gamma} \overset{y=\Gamma^{1/2}x}{=} \sup_{0\neq y \in \mathbb{R}^d} \frac{|y^T \Gamma^{-1/2} (I_d - P_r)^T \nabla h|}{\|y\|} \\
&= \|\Gamma^{-1/2}(I_d - P_r)^T \nabla h\| = \|(I_d - P_r^T)\nabla h\|_{\Gamma^{-1}},
\end{aligned}$$

where $\|\cdot\| = \sqrt{(\cdot)^T(\cdot)}$ denotes the Euclidean norm of $\mathbb{R}^d$ and $\Gamma^{1/2}$ denotes a symmetric positive-definite square root of Γ. This concludes the proof.

C Proof of Proposition 3.1

The proof of Proposition 3.1 requires concentration properties of sub-gaussian random vectors. We will need the following lemma, which is essentially Theorem 5.39 in [47].

Lemma C.1. Let Z be an isotropic sub-gaussian random vector in $\mathbb{R}^n$, meaning that $\mathbb{E}(ZZ^T) = I_n$ and that there exists a constant L such that

$$\|w^T Z\|_{\psi_2} := \sup_{p\geq 1} p^{-1/2} \mathbb{E}(|w^T Z|^p)^{1/p} \leq L\|w\|, \tag{47}$$

for any $w \in \mathbb{R}^n$, where $\|\cdot\|$ denotes the Euclidean norm of $\mathbb{R}^n$. Let $\widehat{\Sigma} = \frac{1}{K}\sum_{k=1}^K Z_k Z_k^T$, where $Z_1, \ldots, Z_K$ are K independent copies of Z. Then for every $t \geq 0$, with probability at least $1 - 2\exp(-ct^2)$, we have

$$\|\widehat{\Sigma} - I_d\| \leq \max\{\tau, \tau^2\} \quad \text{where} \quad \tau = C\sqrt{\frac{n}{K}} + \frac{t}{\sqrt{K}}. \tag{48}$$

where $\|\widehat{\Sigma} - I_d\| = \max\{|\lambda| : \lambda \in \mathrm{Sp}(\widehat{\Sigma} - I_d)\}$ denotes the spectral norm of $\widehat{\Sigma} - I_d$. Here, $c = L^{-4}c_1$ and $C = L^2\sqrt{\log(9)/c_1}$, where c_1 is an absolute (numerical) constant.

of Lemma C.1. The proof is exactly the one of Theorem 5.39 in [47] with $A \in \mathbb{R}^{K\times d}$ the random matrix whose rows contain the vectors $Z_1, \ldots, Z_K$. By (47) the rows of A are independent sub-gaussian isotropic vectors with sub-gaussian norm $\|Z\|_{\psi_2}$ smaller than L. Then, following [47], we have that (48), which is nothing by Equation (5.23) in [47] with $\widehat{\Sigma} = \frac{1}{K}AA^T$, holds with probability greater than $1 - 2\exp(-c_1 t^2 \|Z\|_{\psi_2}^{-4})$. Here c_1 is an absolute constant coming from Corollary 5.17 in [47]. Since $\|Z\|_{\psi_2} \leq L$, Relation (48) holds with probability greater than $1 - 2\exp(-ct^2)$, where $c = L^{-4}c_1$. Finally, as mentioned at Step 3 of the proof in [47], any constant C such that $C \geq \|Z\|^2_{\psi_2}\sqrt{\log(9)/c_1}$ gives the result, so we can chose $C = L^2\sqrt{\log(9)/c_1}$. □

We now give the proof of Proposition 3.1.

Assume we are given an approximation $\widehat{H}$ of H such that

$$(1-\delta)H \preceq \widehat{H} \preceq (1+\delta)H, \tag{49}$$

for some $\delta < 1$. For any P_r we can write $\mathcal{R}_\Gamma(P_r, H) = \mathrm{trace}(\Gamma^{-1}(I_d - P_r^T)H(I_d - P_r)) = \mathrm{trace}(HB)$ with $B = (I_d - P_r)\Gamma^{-1}(I_d - P_r^T)$. Since B is a symmetric matrix with $B \succeq 0$, the relation

$\mathrm{trace}(H_1 B) \leq \mathrm{trace}(H_2 B)$ holds for any symmetric matrices H_1 and H_2 such that $H_1 \preceq H_2$. Then Relation (49) yields (27). It remains to show that, under the assumptions (25) and (26), the K-sample Monte Carlo estimate $\widehat{H}$ of H satisfies (49) with probability greater than $1-\eta$.

Let $n = \mathrm{rank}(H)$ and let $G \in \mathbb{R}^{d\times n}$ be a full column rank matrix[2] such that $H = GG^T$. Since $\mathrm{Im}(H) = \mathrm{Im}(G)$ and since $\nabla \log f(X) \in \mathrm{Im}(H)$ almost surely, where $X \sim \nu$, there exists a random vector Z in $\mathbb{R}^n$ such that $GZ = \nabla \log f(X)$. We can write

$$G^T G \mathbb{E}(ZZ^T) G^T G = G^T \mathbb{E}((GZ)(GZ)^T) G = G^T H G = (G^T G)^2,$$

so that, since $G^T G$ is invertible (G is full column rank), we have $\mathbb{E}(ZZ^T) = I_n$. Then Z is isotropic. For any $w \in \mathbb{R}^n$, Relation (25) allows writing

$$\|w^T Z\|_{\psi_2} = \|u^T \nabla \log f(X)\|_{\psi_2} \leq L\sqrt{u^T H u} = L\|w\|,$$

where $u \in \mathbb{R}^d$ is any vector such that $w = G^T u$. Then Z is sub-gaussian and satisfies (47) with the same L as in (25). Thus Lemma C.1 ensures that, for any $t \geq 0$ and with a probability greater than $1 - 2\exp(-ct^2)$, the matrix $\widehat{\Sigma} = \frac{1}{K}\sum_{k=1}^K Z_k Z_k^T$ satisfies (48). Let $\delta \in (0,1)$ and assume

$$K \geq \delta^{-2}(C\sqrt{n} + t)^2. \tag{50}$$

This is equivalent to $\tau \leq \delta$ where $\tau = C\sqrt{nK^{-1}} + t\sqrt{K^{-1}}$. Since $\delta < 1$ we have $\max\{\tau, \tau^2\} = \tau \leq \delta$ so that (48) yields $(1-\delta)I_n \preceq \widehat{\Sigma} \preceq (1+\delta)I_n$. By multiplying by G to the left and by G^T to the right, we obtain (49). Then for any $t \geq 0$, the condition (50) ensures that (49) holds with a probability at least $1 - 2\exp(-ct^2)$. Let $\eta \in (0,1)$ and chose t such that $\eta = 2\exp(-ct^2)$, meaning $t = \sqrt{c^{-1}\log(2\eta^{-1})}$. Since $c = L^{-4}c_1$ and $C = L^2\sqrt{\log(9)/c_1}$ for some absolute constant c_1, we can write

$$\begin{aligned}\delta^{-2}(C\sqrt{n} + t)^2 &= \delta^{-2}(C\sqrt{n} + \sqrt{c^{-1}\log(2\eta^{-1})})^2 \\ &\leq \Omega\delta^{-2}L^4(\sqrt{\mathrm{rank}(H)} + \sqrt{\log(2\eta^{-1})})^2\end{aligned}$$

where $\Omega = \log(9)/c_1$. Then (26) implies (50) so that (49) holds with a probability greater than $1-\eta$, which concludes the proof of Proposition 3.1.

D Proof of Proposition 3.4

Let $F : x \mapsto \mathbb{E}(f(P_r x + (I_d - P_r)Y))$ where $Y \sim \mu$ and let h be any $\sigma(P_r)$-measurable function. By Doob-Dynkin's lemma, see Corollary 2.2, h can be written as a composition of a function with P_r, so that $h(P_r x + (I_d - P_r)y) = h(x)$ holds for all $x, y \in \mathbb{R}^d$. F satisfies also the same property. Thus, and by definition of μ', we can write

$$\begin{aligned}\int F\, h \,\mathrm{d}\mu' &= \int\int F(P_r x + (I_d - P_r)y)\, h(P_r x + (I_d - P_r)y)\, \mu(\mathrm{d}x)\mu(\mathrm{d}y) \\ &= \int F(x)h(x)\,\mu(\mathrm{d}x) = \int \Big(\int f(P_r x + (I_d - P_r)y)\mu(\mathrm{d}y)\Big)h(x)\,\mu(\mathrm{d}x) \\ &= \int\int f(P_r x + (I_d - P_r)y)\, h(P_r x + (I_d - P_r)y)\, \mu(\mathrm{d}x)\mu(\mathrm{d}y) = \int f\, h \,\mathrm{d}\mu',\end{aligned}$$

which shows that F is the conditional expectation $\mathbb{E}_{\mu'}(f|\sigma(P_r))$.

[2]For instance $G = UD^{1/2}$ where $D \in \mathbb{R}^{n\times n}$ is the diagonal matrix containing the non-zeros eigenvalues of the SPD matrix H and where $U \in \mathbb{R}^{d\times n}$ is the matrix whose columns are the corresponding eigenvectors.

E Proof of Proposition 3.5

For any realization of $\widehat{F}_r$, we can decompose $\mathrm{D}_{\mathrm{KL}}(\nu_r^*||\widehat{\nu}_r)$ as follow

$$\mathrm{D}_{\mathrm{KL}}(\nu_r^*||\widehat{\nu}_r) = \mathrm{D}_{\mathrm{KL}}(\nu_r^*||\nu_r') + \int \frac{\mathbb{E}_\mu(f|\sigma(P_r))}{Z} \log\Big(\frac{\mathbb{E}_{\mu'}(f|\sigma(P_r))}{\widehat{F}_r}\Big)\mathrm{d}\mu + \log\Big(\frac{\widehat{Z}}{Z'}\Big),$$

where Z, Z' and $\widehat{Z}$ are the normalizing constants associated with $\mathbb{E}_\mu(f|\sigma(P_r))$, $\mathbb{E}_{\mu'}(f|\sigma(P_r))$ and $\widehat{F}_r$ respectively. Because $\mathbb{E}(\widehat{Z}) = \mathbb{E}(\int \widehat{F}_r \mathrm{d}\mu) = \int \mathbb{E}(\widehat{F}_r)\mathrm{d}\mu = \int \mathbb{E}_{\mu'}(f|\sigma(P_r))\mathrm{d}\mu = Z'$, Jensen's inequality allows writing $\mathbb{E}(\log(\widehat{Z})) \leq \log(\mathbb{E}(\widehat{Z})) = \log(Z')$. Then, taking the expectation in the previous relation yields

$$\mathbb{E}\big(\mathrm{D}_{\mathrm{KL}}(\nu_r^*||\widehat{\nu}_r)\big) \leq \mathrm{D}_{\mathrm{KL}}(\nu_r^*||\nu_r') + \mathbb{E}\int \log\Big(\frac{\mathbb{E}_{\mu'}(f|\sigma(P_r))}{\widehat{F}_r}\Big)\,\mathrm{d}\nu_r^*. \tag{51}$$

Using the second order Taylor expansion $\log(1+x) = x - \frac{1}{2}x^2 + \mathcal{O}(x^3)$ and since $\mathbb{E}(\widehat{F}_r - \mathbb{E}_{\mu'}(f|\sigma(P_r))) = 0$, we can write

$$\begin{aligned}
\mathbb{E}\int \log\Big(\frac{\mathbb{E}_{\mu'}(f|\sigma(P_r))}{\widehat{F}_r}\Big)\,\mathrm{d}\nu_r^* &= -\mathbb{E}\int \log\Big(1 + \frac{\widehat{F}_r - \mathbb{E}_{\mu'}(f|\sigma(P_r))}{\mathbb{E}_{\mu'}(f|\sigma(P_r))}\Big)\,\mathrm{d}\nu_r^* \\
&= \mathbb{E}\int \frac{1}{2}\Big(\frac{\widehat{F}_r - \mathbb{E}_{\mu'}(f|\sigma(P_r))}{\mathbb{E}_{\mu'}(f|\sigma(P_r))}\Big)^2 + \mathcal{O}\left(\Big(\frac{\widehat{F}_r - \mathbb{E}_{\mu'}(f|\sigma(P_r))}{\mathbb{E}_{\mu'}(f|\sigma(P_r))}\Big)^3\right)\,\mathrm{d}\nu_r^*,
\end{aligned}$$

which together with (51) conclude the proof.

F Proof of Proposition 3.6

To prove this proposition, we will need the following lemmas and corollaries.

Lemma F.1. Let ν_1 and ν_2 be two probability distributions such that $\frac{\mathrm{d}\nu_1}{\mathrm{d}\mu} \propto f_1$ and $\frac{\mathrm{d}\nu_2}{\mathrm{d}\mu} \propto f_2$. Then we have

$$\mathrm{D}_{\mathrm{KL}}(\nu_1||\nu_2) \leq \frac{Z_2}{Z_1^2}\int \frac{(f_1 - f_2)^2}{f_2}\mathrm{d}\mu$$

where $Z_1 = \int f_1 \,\mathrm{d}\mu$ and $Z_2 = \int f_2 \,\mathrm{d}\mu$ are normalizing constants.

Proof. Let $\chi^2(\nu_1||\nu_2) = \int \frac{(f_1/Z_1 - f_2/Z_2)^2}{f_2/Z_2}\mathrm{d}\mu$ be the χ^2-divergence from ν_2 to ν_1. Using Jensen's inequality, we have $\mathrm{D}_{\mathrm{KL}}(\nu_1||\nu_2) \leq \chi^2(\nu_1||\nu_2)$. For any $\alpha \in \mathbb{R}$ we can write

$$\int \frac{(f_1/Z_1 - \alpha f_2/Z_2)^2}{f_2/Z_2}\mathrm{d}\mu = \int \frac{(f_1/Z_1)^2}{f_2/Z_2}\mathrm{d}\mu - 2\alpha + \alpha^2 = \chi^2(\nu_1||\nu_2) + (1-\alpha)^2.$$

The choice $\alpha = Z_2/Z_1$ yields $\chi^2(\nu_1||\nu_2) \leq \int \frac{(f_1/Z_1 - f_2/Z_1)^2}{f_2/Z_2}\mathrm{d}\mu$ which gives the result. □

Lemma F.2. Let μ and μ' be two probability distributions with Lebesgue densities ρ and ρ'. Then for any measurable function h we have

$$\Big(\int h \,\mathrm{d}\mu - \int h \,\mathrm{d}\mu'\Big)^2 \leq \chi^2(\mu||\mu')\int \Big(h - \int h \mathrm{d}\mu'\Big)^2 \mathrm{d}\mu'$$

where $\chi^2(\mu||\mu') = \int \frac{(\rho - \rho')^2}{\rho'}\mathrm{d}x$ denotes the χ^2-divergence from μ' to μ.

Proof. Using Cauchy-Schwarz inequality we can write

$$\Big(\int h\ \mathrm{d}\mu - \int h\ \mathrm{d}\mu'\Big)^2 = \Big(\int (h\ \sqrt{\rho'})\Big(\frac{\rho-\rho'}{\sqrt{\rho'}}\Big)\mathrm{d}x\Big)^2$$
$$\leq \Big(\int h^2\rho'\ \mathrm{d}x\Big)\Big(\int \frac{(\rho-\rho')^2}{\rho'}\mathrm{d}x\Big)$$

Replacing h by $h - \int h\mathrm{d}\mu'$ gives the result. □

Lemma F.3. Assume $\rho(x) = \exp(-\frac{1}{2}\|x-m\|_\Gamma^2 - \Psi(x))$ where Ψ is a bounded function such that $\exp(\sup\Psi - \inf\Psi) \leq \kappa$ and $m \in \mathbb{R}^d$. Given an $\|\cdot\|_\Gamma$-orthogonal projector P_r, let $U_r \in \mathbb{R}^{d\times r}$ and $U_\perp \in \mathbb{R}^{d\times(d-r)}$ be two matrices whose columns form an basis of $\mathrm{Im}(P_r)$ and $\mathrm{Ker}(P_r)$ respectively. Let

$$p_\perp(\xi_\perp) = \frac{1}{C}\int_{\mathbb{R}^r} \rho(U_r\xi_r' + U_\perp\xi_\perp)\mathrm{d}\xi_r'$$
$$p_r(\xi_r) = \frac{1}{C}\int_{\mathbb{R}^{d-r}} \rho(U_r\xi_r + U_\perp\xi_\perp')\mathrm{d}\xi_\perp',$$

where $C = \int_{\mathbb{R}^r\times\mathbb{R}^{d-r}} \rho(U_r\xi_r' + U_\perp\xi_\perp')\mathrm{d}\xi_r'\mathrm{d}\xi_\perp'$. Then we have

$$\kappa^{-2}p_r(\xi_r)p_\perp(\xi_\perp) \leq \frac{1}{C}\ \rho(U_r\xi_r + U_\perp\xi_\perp) \leq \kappa^2 p_r(\xi_r)p_\perp(\xi_\perp). \tag{52}$$

Furthermore the relations $\kappa^{-2}p_\perp(\xi_\perp) \leq p_\perp(\xi_\perp|P_rx) \leq \kappa^2 p_\perp(\xi_\perp)$ and $\chi^2(p_\perp(\cdot|P_rx)||p_\perp) \leq \kappa^2 - 1$ holds for any $x \in \mathbb{R}^d$, where

$$p_\perp(\xi_\perp|P_rx) = \frac{\rho(P_rx + U_\perp\xi_\perp)}{\int_{\mathbb{R}^{d-r}}\rho(P_rx + U_\perp\xi_\perp')\mathrm{d}\xi_\perp'}.$$

Proof. We first note that $x \mapsto \rho(x)$ and $x \mapsto \rho(x+m)$ have the same structure so that, without loss of generality, we can assume $m = 0$. Let $\alpha = \exp(-\sup\Psi)$ and $\beta = \exp(-\inf\Psi)$ so that $\beta/\alpha \leq \kappa$. For any $\xi_r \in \mathbb{R}^r$, $\xi_\perp \in \mathbb{R}^{d-r}$ we have

$$\alpha\ \exp(-\frac{1}{2}\|U_r\xi_r + U_\perp\xi_\perp\|_\Gamma^2) \leq \rho(U_r\xi_r + U_\perp\xi_\perp) \leq \beta\ \exp(-\frac{1}{2}\|U_r\xi_r + U_\perp\xi_\perp\|_\Gamma^2). \tag{53}$$

Using the fact that $\|U_r\xi_r + U_\perp\xi_\perp\|_\Gamma^2 = \|U_r\xi_r\|_\Gamma^2 + \|U_\perp\xi_\perp\|_\Gamma^2$, and integrating over either ξ_r of $\xi_\perp$, Equation (53) yields

$$\alpha C_r \exp(-\frac{1}{2}\|U_\perp\xi_\perp\|_\Gamma^2) \leq \int_{\mathbb{R}^r}\rho(U_r\xi_r' + U_\perp\xi_\perp)\mathrm{d}\xi_r' \leq \beta C_r\exp(-\frac{1}{2}\|U_\perp\xi_\perp\|_\Gamma^2) \tag{54}$$

$$\alpha C_\perp \exp(-\frac{1}{2}\|U_r\xi_r\|_\Gamma^2) \leq \int_{\mathbb{R}^{d-r}}\rho(U_r\xi_r + U_\perp\xi_\perp')\mathrm{d}\xi_\perp' \leq \beta C_\perp\exp(-\frac{1}{2}\|U_r\xi_r\|_\Gamma^2), \tag{55}$$

where $C_r = \int_{\mathbb{R}^{d-r}}\exp(-\frac{1}{2}\|U_r\xi_r'\|_\Gamma^2)\mathrm{d}\xi_r'$ and $C_\perp = \int_{\mathbb{R}^{d-r}}\exp(-\frac{1}{2}\|U_\perp\xi_\perp'\|_\Gamma^2)\mathrm{d}\xi_\perp'$. By integrating (55) over ξ_r we get

$$\alpha C_rC_\perp \leq C = \int_{\mathbb{R}^r\times\mathbb{R}^{d-r}}\rho(U_r\xi_r' + U_\perp\xi_\perp')\mathrm{d}\xi_r'\mathrm{d}\xi_\perp' \leq \beta C_rC_\perp. \tag{56}$$

Combining (53), (54), (55), (56) we obtain (52). Finally, dividing (52) by $p_r(\xi_r)$ and letting ξ_r such that $U_r\xi_r = P_rx$ for some $x \in \mathbb{R}^d$, we can write $\kappa^{-2}p_\perp(\xi_\perp) \leq p_\perp(\xi_\perp|P_rx) \leq \kappa^2 p_\perp(\xi_\perp)$ and then

$$\chi^2(p_\perp(\cdot|P_rx)||p_\perp) = \int\frac{(p_\perp(\xi_\perp|P_rx) - p_\perp(\xi_\perp))^2}{p_\perp(\xi_\perp)}\mathrm{d}\xi_\perp = \int\frac{p_\perp(\xi_\perp|P_rx)^2}{p_\perp(\xi_\perp)}\mathrm{d}\xi_\perp - 1$$
$$= \int\frac{p_\perp(\xi_\perp|P_rx)}{p_\perp(\xi_\perp)}p_\perp(\xi_\perp|P_rx)\mathrm{d}\xi_\perp - 1 \leq \kappa^2 - 1.$$

This concludes the proof.

□

Corollary F.4 (Subspace Poincaré inequality)**.** Under Assumption 2.5, the probability distribution μ satisfies the following *subspace Poincaré inequality*:

$$\int \big(h - \mathbb{E}_\mu(h|\sigma(P_r))\big)^2 \mathrm{d}\mu \leq \kappa \int \|(I_d - P_r^T)\nabla h\|^2_{\Gamma^{-1}} \mathrm{d}\mu, \tag{57}$$

for any function h with sufficient regularity.

Proof. It is well known that a distribution which satisfies the logarithmic Sobolev inequality also satisfies the Poincaré inequality. Similarly, the subspace logarithmic Sobolev inequality (13) implies (57) by replacing h by $1 + \varepsilon h$ in (13) and using a Taylor expansion of the logarithm as ε goes to zero. □

We now have all the material to proof Proposition 3.6.

We first show that (34) holds. By Lemma F.1 we have

$$\mathrm{D}_{\mathrm{KL}}(\nu_r^* || \nu_r') \leq \frac{Z'}{Z^2} \int \frac{(\mathbb{E}_\mu(f|\sigma(P_r)) - \mathbb{E}_{\mu'}(f|\sigma(P_r)))^2}{\mathbb{E}_{\mu'}(f|\sigma(P_r))} \mathrm{d}\mu, \tag{58}$$

where Z and Z' are normalizing constants associated with $\mathbb{E}_\mu(f|\sigma(P_r))$ and $\mathbb{E}_{\mu'}(f|\sigma(P_r))$ respectively. Fix $x \in \mathbb{R}^d$. Using the notations of Lemma (F.3), and by Propositions 2.4 and 3.4 we can write

$$\mathbb{E}_\mu(f|\sigma(P_r))(x) = \int_{\mathbb{R}^{d-r}} f(P_r x + U_\perp \xi_\perp) p_\perp(\xi_\perp | P_r x) \mathrm{d}\xi_\perp \tag{59}$$

$$\mathbb{E}_{\mu'}(f|\sigma(P_r))(x) = \int_{\mathbb{R}^{d-r}} f(P_r x + U_\perp \xi_\perp) p_\perp(\xi_\perp) \mathrm{d}\xi_\perp, \tag{60}$$

By Lemma F.3, the relation

$$\kappa^{-2}\, \mathbb{E}_{\mu'}(f|\sigma(P_r))(x) \leq \mathbb{E}_\mu(f|\sigma(P_r))(x) \leq \kappa^2\, \mathbb{E}_{\mu'}(f|\sigma(P_r))(x), \tag{61}$$

holds. Also, notice that both $\mathbb{E}_\mu(f|\sigma(P_r))(x)$ and $\mathbb{E}_{\mu'}(f|\sigma(P_r))(x)$ can be written as an expectation of $\xi_\perp \mapsto f(P_r x + U_\perp \xi_\perp)$ over the densities $p_\perp(\cdot|P_r x)$ and $p_\perp$ respectively. Then, using Lemma F.2 and Lemma F.3, we have

$$\begin{aligned}
&\big(\mathbb{E}_\mu(f|\sigma(P_r))(x) - \mathbb{E}_{\mu'}(f|\sigma(P_r))(x)\big)^2 \\
&\quad \leq (\kappa^2 - 1) \int_{\mathbb{R}^{d-r}} \big(f(P_r x + U_\perp \xi_\perp) - \mathbb{E}_{\mu'}(f|\sigma(P_r))(x)\big)^2 p_\perp(\xi_\perp) \mathrm{d}\xi_\perp \\
&\quad = (\kappa^2 - 1) \int \big(f(P_r x + (I_d - P_r) y) - \mathbb{E}_{\mu'}(f|\sigma(P_r))(x)\big)^2 \mu(\mathrm{d}y).
\end{aligned}$$

Dividing by $\mathbb{E}_{\mu'}(f|\sigma(P_r))(x)$ and taking the expectation over $x \sim \mu$ we obtain

$$\begin{aligned}
&\int \frac{(\mathbb{E}_\mu(f|\sigma(P_r)) - \mathbb{E}_{\mu'}(f|\sigma(P_r)))^2}{\mathbb{E}_{\mu'}(f|\sigma(P_r))} \mathrm{d}\mu \\
&\quad \leq (\kappa^2 - 1) \int \int \frac{(f(P_r x + (I_d - P_r) y) - \mathbb{E}_{\mu'}(f|\sigma(P_r))(x))^2}{\mathbb{E}_{\mu'}(f|\sigma(P_r))(x)} \mu(\mathrm{d}y) \mu(\mathrm{d}x) &(62)\\
&\quad = (\kappa^2 - 1) \int \int \frac{(f(P_r x + (I_d - P_r) y) - \mathbb{E}_{\mu'}(f|\sigma(P_r))(P_r x + (I_d - P_r) y))^2}{\mathbb{E}_{\mu'}(f|\sigma(P_r))(P_r x + (I_d - P_r) y)} \mu(\mathrm{d}y) \mu(\mathrm{d}x) &(63)\\
&\quad = (\kappa^2 - 1) \int \frac{(f - \mathbb{E}_{\mu'}(f|\sigma(P_r)))^2}{\mathbb{E}_{\mu'}(f|\sigma(P_r))} \mathrm{d}\mu' &(64)
\end{aligned}$$

Going from (62) to (63), we used the fact that, since $\mathbb{E}_{\mu'}(f|\sigma(P_r))$ is a $\sigma(P_r)$-measurable function, it can be writen as the composition of some function with P_r, so that relation $\mathbb{E}_{\mu'}(f|\sigma(P_r))(x) = \mathbb{E}_{\mu'}(f|\sigma(P_r))(P_r x + (I_d - P_r)y)$ holds for any $x, y \in \mathbb{R}^d$. To obtain (64), we used the definition of μ', see Equation (32). The last term in (64) satisfies

$$\int \frac{(f - \mathbb{E}_{\mu'}(f|\sigma(P_r)))^2}{\mathbb{E}_{\mu'}(f|\sigma(P_r))} \mathrm{d}\mu' = \int \Big(\frac{f}{\mathbb{E}_{\mu'}(f|\sigma(P_r))^{1/2}} - \mathbb{E}_{\mu'}(f|\sigma(P_r))^{1/2} \Big)^2 \mathrm{d}\mu' \tag{65}$$

$$\leq \int \Big(\frac{f}{\mathbb{E}_{\mu'}(f|\sigma(P_r))^{1/2}} - \frac{\mathbb{E}_{\mu}(f|\sigma(P_r))}{\mathbb{E}_{\mu'}(f|\sigma(P_r))^{1/2}} \Big)^2 \mathrm{d}\mu' \tag{66}$$

$$\leq \kappa^4 \int \Big(\frac{f}{\mathbb{E}_{\mu}(f|\sigma(P_r))^{1/2}} - \frac{\mathbb{E}_{\mu}(f|\sigma(P_r))}{\mathbb{E}_{\mu}(f|\sigma(P_r))^{1/2}} \Big)^2 \mathrm{d}\mu \tag{67}$$

$$= \kappa^4 \int \Big(\frac{f}{\mathbb{E}_{\mu}(f|\sigma(P_r))^{1/2}} - \mathbb{E}_{\mu}(f|\sigma(P_r))^{1/2} \Big)^2 \mathrm{d}\mu \tag{68}$$

$$\leq \kappa^5 \int \|(I_d - P_r^T)\nabla \Big(\frac{f}{\mathbb{E}_{\mu}(f|\sigma(P_r))^{1/2}} \Big) \|^2_{\Gamma^{-1}} \mathrm{d}\mu \tag{69}$$

$$= \kappa^5 \int \|(I_d - P_r^T)\nabla \log f\|^2_{\Gamma^{-1}} \frac{f^2}{\mathbb{E}_{\mu}(f|\sigma(P_r))} \mathrm{d}\mu \tag{70}$$

Let us explain the previous steps.

- To go from (65) to (66) notice that $\mathbb{E}_{\mu'}(f|\sigma(P_r))^{1/2}$ is a $\sigma(P_r)$-measurable function which satisfies $\mathbb{E}_{\mu'}(f|\sigma(P_r))^{1/2} = \mathbb{E}_{\mu'}(f\mathbb{E}_{\mu'}(f|\sigma(P_r))^{-1/2}|\sigma(P_r))$. In particular, $\mathbb{E}_{\mu'}(f|\sigma(P_r))^{1/2}$ is a best approximation of $f\mathbb{E}_{\mu'}(f|\sigma(P_r))^{-1/2}$ among all $\sigma(P_r)$-measurable functions with respect to the $L^2_{\mu'}$-norm. Then (66) follows from the fact that $\frac{\mathbb{E}_{\mu}(f|\sigma(P_r))}{\mathbb{E}_{\mu'}(f|\sigma(P_r))^{1/2}}$ is $\sigma(P_r)$-measurable.

- Going from (66) to (67) is essentially a consequence of Lemma (F.3). First notice that (66) is an integral of a positive function over μ'. By (52) it can be bounded by κ^2 times the integral of the same function over μ. Also by (61), we can lower bound the denominators in (66) to get (67).

- From (68) to (69), we applied the subspace Poincaré inequality (57) with $h = f\mathbb{E}_{\mu}(f|\sigma(P_r)^{-1/2}$. Indeed, since $\mathbb{E}_{\mu}(h|\sigma(P_r)) = \mathbb{E}_{\mu}(f|\sigma(P_r))^{1/2}$, (68) can be written as $\kappa^4 \int (h - \mathbb{E}_{\mu}(h|\sigma(P_r)))^2 \mathrm{d}\mu$, which is κ^4 times the left hand side of (57).

Together with (58) and (64), (70) yields

$$\mathrm{D}_{\mathrm{KL}}(\nu_r^* || \widehat{\nu}_r) \leq \kappa^5(\kappa^2 - 1)\frac{Z'}{Z} \int \|(I_d - P_r^T)\nabla \log f\|^2_{\Gamma^{-1}} \frac{f}{\mathbb{E}_{\mu}(f|\sigma(P_r))} \frac{f}{Z} \mathrm{d}\mu. \tag{71}$$

Notice that by integrating (61) over $x \sim \mu$ we obtain $Z' \leq \kappa^2 Z$, so that (71) yields (34).

We now show that (35) holds. First note that, for any $x \in \mathbb{R}^d$, the independence of the Y_i's in the definition (30) of $\widehat{F}_r$ ensures that

$$\mathbb{E}\Big(\Big(\frac{\widehat{F}_r(x) - \mathbb{E}_{\mu'}(f|\sigma(P_r))(x)}{\mathbb{E}_{\mu'}(f|\sigma(P_r))(x)} \Big)^2 \Big)$$
$$= \frac{1}{M} \int \Big(\frac{f(P_r x + (I_d - P_r)y) - \mathbb{E}_{\mu'}(f|\sigma(P_r))(x)}{\mathbb{E}_{\mu'}(f|\sigma(P_r))(x)} \Big)^2 \mu(\mathrm{d}y).$$

Multiplying by $\mathbb{E}_\mu(f|\sigma(P_r))(x)/Z$ and taking the expectation with respect to $x \sim \mu$ we have

$$\begin{aligned}
&\int \frac{\mathbb{E}_\mu(f|\sigma(P_r))}{Z} \mathbb{E}\Big(\frac{\widehat{F}_r - \mathbb{E}_{\mu'}(f|\sigma(P_r))}{\mathbb{E}_{\mu'}(f|\sigma(P_r))}\Big)^2 \mathrm{d}\mu \\
&= \frac{1}{M} \int \frac{\mathbb{E}_\mu(f|\sigma(P_r))(x)}{Z} \int \Big(\frac{f(P_r x + (I_d - P_r)y) - \mathbb{E}_{\mu'}(f|\sigma(P_r))(x)}{\mathbb{E}_{\mu'}(f|\sigma(P_r))(x)}\Big)^2 \mu(\mathrm{d}y)\mu(\mathrm{d}x) \\
&= \frac{1}{M} \int \frac{\mathbb{E}_\mu(f|\sigma(P_r))}{Z} \Big(\frac{f - \mathbb{E}_{\mu'}(f|\sigma(P_r))}{\mathbb{E}_{\mu'}(f|\sigma(P_r))}\Big)^2 \mathrm{d}\mu' \qquad (72) \\
&\leq \frac{\kappa^2}{MZ} \int \frac{(f - \mathbb{E}_{\mu'}(f|\sigma(P_r)))^2}{\mathbb{E}_{\mu'}(f|\sigma(P_r))} \mathrm{d}\mu' \qquad (73) \\
&\leq \frac{\kappa^7}{MZ} \int \|(I_d - P_r^T)\nabla \log f\|^2_{\Gamma^{-1}} \frac{f^2}{\mathbb{E}_\mu(f|\sigma(P_r))} \mathrm{d}\mu \qquad (74)
\end{aligned}$$

Let us detail the previous steps. To get (72) we can use similar arguments as when going from (62) to (64), meaning exploiting the $\sigma(P_r)$-measurablility of $\mathbb{E}_{\mu'}(f|\sigma(P_r))$ and of $\mathbb{E}_\mu(f|\sigma(P_r))$ and the property (32) of μ'. Going from (72) to (73), we used relation (61). Using (70) we get (74) which yields (35) and concludes the proof of Proposition 3.6.

References

[1] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. Stuart, *Importance sampling: intrinsic dimension and computational cost*, Statistical Science, 32 (2017), pp. 405–431.

[2] D. Bakry, M. Émery, and M. I. É. Mery, *Diffusions hypercontractives*, in S{é}minaire de Probabilit{é}s XIX 1983/84, vol. 19, Springer, 1985, pp. 177–206.

[3] A. Banerjee, X. Guo, and H. Wang, *On the optimality of conditional expectation as a Bregman predictor*, IEEE Transactions on Information Theory, 51 (2005), pp. 2664–2669.

[4] A. Beskos, A. Jasra, K. Law, Y. Marzouk and Y. Zhou, *Multilevel sequential Monte Carlo with dimension-independent likelihood-informed proposals*, SIAM/ASA Journal on Uncertainty Quantification, 6 (2018), pp. 762–786.

[5] G. Blanchard, O. Bousquet, and L. Zwald, *Statistical properties of kernel principal component analysis*, Machine Learning, 66 (2007), pp. 259–294.

[6] S. G. Bobkov and M. Ledoux, *From Brunn-Minkowski to Brascamp-Lieb and to logarithmic sobolev inequalities*, Geometric & Functional Analysis GAFA, 10 (2000), pp. 1028–1052.

[7] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*, Oxford university press, 2013.

[8] A. Cohen, I. Daubechies, R. DeVore, G. Kerkyacharian, and D. Picard, *Capturing Ridge Functions in High Dimensions from Point Queries*, Constructive Approximation, 35 (2012), pp. 225–243.

[9] P. Conrad, Y. Marzouk, N. Pillai, and A. Smith, *Accelerating asymptotically exact MCMC for computationally intensive models via local approximations*, Journal of the American Statistical Association, 111 (2016), pp. 1591–1607.

[10] P. G. Constantine, E. Dow, and Q. Wang, *Active subspace methods in theory and practice: applications to Kriging surfaces*, SIAM Journal on Scientific Computing, 36 (2014), pp. A1500–A1524.

[11] P. G. Constantine, C. Kent, and T. Bui-Thanh, *Accelerating Markov chain Monte Carlo with active subspaces*, SIAM Journal on Scientific Computing, 38 (2016), pp. A2779–A2805.

[12] T. Cui, J. Martin, Y. M. Marzouk, A. Solonen, and A. Spantini, *Likelihood-informed dimension reduction for nonlinear inverse problems*, Inverse Problems, 29 (2014), pp. 114015.

[13] T. Cui, Y. Marzouk, and K. Willcox, *Data-driven model reduction for the Bayesian solution of inverse problems*, International Journal for Numerical Methods in Engineering, 102 (2015), pp. 966–990.

[14] T. Cui, K. Law, and Y. M. Marzouk, *Dimension-independent likelihood-informed MCMC*, Journal of Computational Physics, 304 (2016), pp. 109–137.

[15] T. Cui, Y. M. Marzouk, K. Willcox, *Scalable posterior approximations for large-scale Bayesian inverse problems via likelihood-informed parameter and state reduction*, Journal of Computational Physics, 315 (2016), pp. 363–387.

[16] H. P. Flath, L. C. Wilcox, V. Akçelik, J. Hill, B. van Bloemen Waanders and O. Ghattas, *Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations*, Journal of Computational SIAM Journal on Scientific Computing, 33 (2011), pp. 407–432.

[17] M. Fornasier, K. Schnass, and J. Vybiral, *Learning functions of few arbitrary linear parameters in high dimensions*, Foundations of Computational Mathematics, 12 (2012), pp. 229–262.

[18] N. Gozlan, C. Roberto, and P. M. Samson, *Characterization of Talagrand's transport-entropy inequalities in metric spaces*, Annals of Probability, 41 (2013), pp. 3112–3139.

[19] L. Gross, *Logarithmic sobolev inequalities*, American Journal of Mathematics, 97 (1975), pp. 1061–1083.

[20] A. Guionnet and B. Zegarlinksi, *Lectures on logarithmic Sobolev inequalities*, Séminaire de probabilités XXXVI (2003), pp. 1–134.

[21] H. Haario, M. Laine, M. Lehtinen, E. Saksman, and J. Tamminen. *Markov chain Monte Carlo methods for high dimensional inversion in remote sensing.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66 (2004), pp. 591–608.

[22] R. Holley and D. Stroock, *Logarithmic Sobolev inequalities and stochastic Ising models*, Journal of Statistical Physics, 46 (1987), pp. 1159–1194.

[23] Jung, Heinrich, *Ueber die kleinste Kugel, die eine räumliche Figur einschliesst.*, Journal für die reine und angewandte Mathematik, 123 (1901), pp, 241–257.

[24] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer (2005).

[25] O. Lamminpää, M. Laine, S. Tukiainen and J. Tamminen, *Likelihood informed dimension reduction for inverse problems in remote sensing of atmospheric constituent profiles*, arXiv preprint (2017) arXiv:1709.02611.

[26] M. Ledoux, *Concentration of measure and logarithmic Sobolev inequalities*, Séminaire de probabilités XXXIII (1999), pp. 120–216.

[27] J. Li, *A note on the Karhunen-Loève expansions for infinite-dimensional Bayesian inverse problems*, Statistics and Probability Letters, 106 (2015), pp. 1–4.

[28] W. Li and O. A. Cirpka, *Efficient geostatistical inverse methods for structured and unstructured grids*, Water Resources Research, 42 (2006), pp. 1–15.

[29] C. Lieberman, K. Willcox and O. Ghattas, *Parameter and state model reduction for large-scale statistical inverse problems*, SIAM Journal on Scientific Computing, 32 (2010), pp. 2523–2542.

[30] F. Lindgren, H. Rue, and J. Lindström, *An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach.*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73 (2011), pp. 423–498.

[31] A. Manzoni, S. Pagani and T. Lassila, *Accurate solution of Bayesian inverse uncertainty quantification problems combining reduced basis methods and reduction error models*, SIAM/ASA Journal on Uncertainty Quantification, 4 (2016), pp. 380–412.

[32] Y. M. Marzouk and H. N. Najm, *Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems*, Journal of Computational Physics, 228 (2009), pp. 1862–1902.

[33] T. Mikosch and O. Kallenberg, *Foundations of Modern Probability*, Journal of the American Statistical Association, 93 (1998), p. 1243.

[34] F. Otto and C. Villani, *Generalization of an Inequality By Talagrand, and Links With the Logarithmic Sobolev Inequality*, J. of Functional Analysis, 173 (1999), pp. 361–400.

[35] A. Owen, *Monte Carlo theory, methods and examples*, 2013.

[36] A. Pinkus, *Ridge Functions*, Cambridge Tracts in Mathematics, Cambridge University Press, 2015.

[37] M. Reiss and M. Wahl, *Non-asymptotic upper bounds for the reconstruction error of PCA*, (2016), pp. 1–43.

[38] Robert, C. P. and Casella, G., Monte Carlo statistical methods. Springer, NYC. 1999.

[39] T. Russi, *Uncertainty Quantification with Experimental Data and Complex System Models*, PhD thesis, UC Berkeley, 2010.

[40] P. B. Rubio, F. Louf and L. Chamoin, *Fast model updating coupling Bayesian inference and PGD model reduction*, Computational Mechanics (2018) pp. 1–25

[41] T. M. Russi, *Uncertainty quantification with experimental data and complex system models*, PhD thesis, 2010.

[42] A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio, and Y. Marzouk, *Optimal Low-rank Approximations of Bayesian Linear Inverse Problems*, SIAM Journal on Scientific Computing, 37 (2015), pp. A2451–A2487.

[43] A. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numerica, 19 (2010), pp. 451–559.

[44] J. Tamminen, *Adaptive Markov chain Monte Carlo algorithms with geophysical applications*, Ph.D. thesis, Finnish Meteorologocal Institute, 2004.

[45] Y. W. Teh, D. Newman, and M. Welling, *A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation*, in Advances in neural information processing systems, 2007, pp. 1353–1360.

[46] H. Tyagi and V. Cevher, *Learning non-parametric basis independent models from point queries via low-rank methods*, Applied and Computational Harmonic Analysis, 37 (2014), pp. 389–412.

[47] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, in Compressed Sensing: Theory and Applications, Y. C. Eldar and G. Kutyniok, eds., Cambridge University Press, 2012, pp. 210–268.

[48] O. Zahm, P. Constantine, C. Prieur and Y. Marzouk, *Gradient-based dimension reduction of multivariate vector-valued functions*, SIAM Journal on Scientific Computing, 42 (2020), pp. A534–A558.