COMPUTO

ISSN 2824-7795

Optimal projection for parametric importance sampling in high dimension

Maxime El Masri ONERA/DTIS, ISAE-SUPAERO, Université de Toulouse
Jérôme Morio ONERA/DTIS, Université de Toulouse
Florian Simatos ISAE-SUPAERO, Université de Toulouse

Abstract

In this paper we propose a dimension-reduction strategy in order to improve the performance of importance sampling in high dimension. The idea is to estimate variance terms in a small number of suitably chosen directions. We first prove that the optimal directions, i.e., the ones that minimize the Kullback–Leibler divergence with the optimal auxiliary density, are the eigenvectors associated to extreme (small or large) eigenvalues of the optimal covariance matrix. We then perform extensive numerical experiments that show that as dimension increases, these directions give estimations which are very close to optimal. Moreover, we show that the estimation remains accurate even when a simple empirical estimator of the covariance matrix is used to estimate these directions. These theoretical and numerical results open the way for different generalizations, in particular the incorporation of such ideas in adaptive importance sampling schemes.

Keywords: Importance sampling, High dimension, Gaussian covariance matrix, Kullback-Leibler divergence, Projection

Contents

```
import numpy as np
import scipy as sp
import scipy.stats
import matplotlib.pyplot as plt
from CEIS_vMFNM import *
from IPython.display import display, Math, Latex
from IPython.display import Markdown
from tabulate import tabulate
np.random.seed(10)
```

1 Introduction

Importance Sampling (IS) is a widely considered stochastic method to estimate integrals of the form $E = \int \phi f$ with a black-box function ϕ and a probability density function (pdf) f. It rests upon the choice of an auxiliary density which, when suitably chosen, can significantly improve the situation compared to the naive Monte Carlo (MC) method (Agapiou et al. 2017), (Owen and Zhou 2000). The theoretical optimal IS density, also called zero-variance density, is defined by $\phi f/E$ when ϕ is a positive function. This density is not available in practice as it involves the unknown integral E, but a classical strategy consists in searching an optimal approximation in a parametric family of densities. By minimising a ‘’distance” with the optimal IS density, such as the Kullback–Leibler divergence, one can find optimal parameters in this family to get an efficient sampling pdf. Adaptive Importance Sampling (AIS) algorithms, such as the Mixture Population Monte Carlo method (Cappé et al. 2008), the Adaptive Multiple Importance Sampling method (Cornuet et al. 2012), or the Cross Entropy method (Reuven Y. Rubinstein and Kroese 2011), estimate the optimal parameters adaptively by updating intermediate parameters (Bugallo et al. 2017).

An intense research activity has made these techniques work very well, but only for moderate dimensions. In high dimension, most of these techniques actually fail to give efficient parameters for two reasons. The first one is the so-called weight degeneracy problem, which is that in high dimension, the weights appearing in the IS densities (which are self-normalized likelihood ratios) degenerate. More precisely, the largest weight takes all the mass, while all other weights are negligible so that the final estimation essentially uses only one sample, see for instance (Bengtsson, Bickel, and Li 2008) for a theoretical analysis in the related context of particle filtering. But even without likelihood ratios, such techniques may fail if they need to estimate high-dimensional parameters such as covariance matrices, whose size increases quadratically in the dimension (Ashurbekova et al. 2020), (Ledoit and Wolf 2004). The conditions under which importance sampling is applicable in high dimension are notably investigated in a reliability context in (Au and Beck 2003): it is remarked that the optimal covariance matrix should not

deviate significantly from the identity matrix. (El-Laham, Elvira, and Bugallo 2019) tackle the weight degeneracy problem by applying a recursive shrinkage of the covariance matrix, which is constructed iteratively with a weighted sum of the sample covariance estimator and a biased, but more stable, estimator. Concerning the second problem of having to estimate high-dimensional parameters, the idea was recently put forth to reduce the effective dimension by only estimating these parameters (in particular the covariance matrix) in suitable directions (El Masri, Morio, and Simatos 2021), (Uribe et al. 2021). In this paper we seek to delve deeper into this idea. The main contribution of the present paper is to identify the optimal directions in the fundamental case when the parametric family is Gaussian, and perform numerical simulations in order to understand how they behave in practice. In particular, we propose directions which, in contrast to the recent paper (Uribe et al. 2021), does not require the objective function to be differentiable, and moreover optimizes the Kullback–Leibler distance with the optimal density instead of simply an upper bound on it, as in (Uribe et al. 2021). In Section 3.1 we elaborate in more details on the differences between the two approaches.

The paper is organised as follows: in Section 2 we recall the foundations of IS. In Section 3, we state our main theoretical result and we compare it with the current state-of-the-art. Section 4 presents the proof of our theoretical result; Section 5 introduces the numerical framework that we have adopted, and Section 6 presents the numerical results obtained on five different test cases to assess the efficiency of the directions that we propose. We conclude in **?@sec-Ccl** with a summary and research perspectives.

2 Importance Sampling

We consider the problem of estimating the following integral:

$$E = \mathbb{E}_f(\phi(\mathbf{X})) = \int \phi(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x},$$

where $\mathbf{X}$ is a random vector in $\mathbb{R}^n$ with Gaussian standard pdf f, and $\phi : \mathbb{R}^n \to \mathbb{R}_+$ is a real-valued, non-negative function. If one were to relax this Gaussian standard assumption, one would need to look for covariance matrices in a different auxiliary set. The function ϕ is considered as a black-box function which is potentially expensive to evaluate, which means the number of calls to ϕ should be limited.

IS is a widely considered approach to reduce the variance of the classical Monte Carlo estimator of E. The idea of IS is to generate a random sample $\mathbf{X}_1, \dots, \mathbf{X}_N$ from an auxiliary density g, instead of f, and to compute the following estimator:

$$\hat{E}_N = \frac{1}{N} \sum_{i=1}^{N} \phi(\mathbf{X}_i) L(\mathbf{X}_i), \tag{1}$$

with $L = f/g$ the likelihood ratio, or importance weight, and the density g, called importance sampling density, is such that $g(\mathbf{x}) = 0$ implies $\phi(\mathbf{x}) f(\mathbf{x}) = 0$ for every $\mathbf{x}$ (which makes the product ϕL well-defined). This estimator is consistent and unbiased but its accuracy strongly

depends on the choice of the auxiliary density g. It is well known that the optimal choice for g is (Bucklew 2013)

$$g^*(\mathbf{x}) = \frac{\phi(\mathbf{x})f(\mathbf{x})}{E}, \ \mathbf{x} \in \mathbb{R}^n.$$

Indeed, for this choice we have $\phi L = E$ and so $\hat{E}_N$ is actually the deterministic estimator E. For this reason, g^* is sometimes called zero-variance density, a terminology that we will adopt here. Of course, g^* is only of theoretical interest as it depends on the unknown integral E. However, it gives an idea of good choices for the auxiliary density g, and we will seek to approximate g^* by an auxiliary density that minimizes a distance between g^* and a given parametric family of densities.

In this paper, the parametric family of densities is the Gaussian family $\{g_{\mathbf{m},\Sigma} : \mathbf{m} \in \mathbb{R}^n, \Sigma \in \mathcal{S}_n^+\}$, where $g_{\mathbf{m},\Sigma}$ denotes the Gaussian density with mean $\mathbf{m} \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathcal{S}_n^+$ with $\mathcal{S}_n^+ \subset \mathbb{R}^{n\times n}$ the set of symmetric, positive-definite matrices:

$$g_{\mathbf{m},\Sigma}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^\top \Sigma^{-1}(\mathbf{x}-\mathbf{m})\right), \ \mathbf{x} \in \mathbb{R}^n.$$

with $|\Sigma|$ the determinant of Σ. Moreover, we will consider the Kullback–Leibler (KL) divergence to measure a ''distance" between g^* and $g_{\mathbf{m},\Sigma}$. Recall that for two densities f and h, with f absolutely continuous with respect to h, the KL divergence $D(f,h)$ between f and h is defined by:

$$D(f,h) = \mathbb{E}_f\left[\log\left(\frac{f(\mathbf{X})}{h(\mathbf{X})}\right)\right] = \int \log\left(\frac{f(\mathbf{x})}{h(\mathbf{x})}\right) f(\mathbf{x})\mathrm{d}\mathbf{x}.$$

Thus, our goal is to approximate g^* by $g_{\mathbf{m}^*,\Sigma^*}$ with the optimal mean vector $\mathbf{m}^*$ and the optimal covariance matrix Σ^* given by:

$$(\mathbf{m}^*,\Sigma^*) = \arg\min\left\{D(g^*, g_{\mathbf{m},\Sigma}) : \mathbf{m} \in \mathbb{R}^n, \Sigma \in \mathcal{S}_n^+\right\}. \tag{2}$$

In the Gaussian case of the present setting, it is well-known that $\mathbf{m}^*$ and Σ^* are simply the mean and variance of the zero-variance density (Reuven Y. Rubinstein and Kroese 2011), (Reuven Y. Rubinstein and Kroese 2017):

$$\mathbf{m}^* = \mathbb{E}_{g^*}(\mathbf{X}) \quad \text{and} \quad \Sigma^* = \mathrm{Var}_{g^*}(\mathbf{X}). \tag{3}$$

3 Main result and positioning of the paper

3.1 Projecting on a low dimensional subspace

As g^* is unknown (although, as will be considered below, we can in principle sample from it since it is known up to a multiplicative constant), the optimal parameters $\mathbf{m}^*$ and Σ^* given by Equation 3 are not directly computable. Therefore, usual estimation schemes start with estimating $\mathbf{m}^*$ and Σ^*, say through $\widehat{\mathbf{m}}^*$ and $\widehat{\Sigma}^*$, respectively, and then use these approximations to estimate E through Equation 1 with the auxiliary density $g_{\widehat{\mathbf{m}}^*,\widehat{\Sigma}^*}$. Although the estimation of E with the auxiliary density $g_{\mathbf{m}^*,\Sigma^*}$ usually provides very good results, it is well-known that in

high dimension, the additional error induced by the estimations of $\mathbf{m}^*$ and Σ^* severely degrades the accuracy of the final estimation (Papaioannou, Geyer, and Straub 2019), (Uribe et al. 2021). The main problem lies in the estimation of Σ^* which, in dimension n, involves the estimation of a quadratic (in the dimension) number of terms, namely $n(n+1)/2$. Recently, the idea to overcome this problem by only evaluating variance terms in a small number of influential directions was explored in (El Masri, Morio, and Simatos 2021) and (Uribe et al. 2021). In these two papers, the auxiliary covariance matrix Σ is modeled in the form

$$\Sigma = \sum_{i=1}^{k} (\nu_i - 1)\mathbf{d}_i\mathbf{d}_i^\top + I_n \tag{4}$$

where the $\mathbf{d}_i$'s are the k orthonormal directions which are deemed influential. It is easy to check that Σ is the covariance matrix of the Gaussian vector

$$\nu_1^{1/2} Y_1 \mathbf{d}_1 + \cdots + \nu_k^{1/2} Y_k \mathbf{d}_k + Y_{k+1}\mathbf{d}_{k+1} + \cdots + Y_n \mathbf{d}_n$$

where the Y_i's are i.i.d. standard normal random variables (one-dimensional), and the $n-k$ vectors $(\mathbf{d}_{k+1}, \dots, \mathbf{d}_n)$ complete $(\mathbf{d}_1, \dots, \mathbf{d}_k)$ into an orthonormal basis. In particular, ν_i is the variance in the direction of $\mathbf{d}_i$, i.e., $\nu_i = \mathbf{d}_i^\top \Sigma \mathbf{d}_i$. In Equation 4, k can be considered as the effective dimension in which variance terms are estimated. In other words, in (El Masri, Morio, and Simatos 2021) and (Uribe et al. 2021), the optimal variance parameter is not sought in $\mathcal{S}_n^+$ as in Equation 2}, but rather in the subset of matrices of the form

$$\mathscr{L}_{n,k} = \left\{ \sum_{i=1}^{k} (\alpha_i - 1) \frac{\mathbf{d}_i\mathbf{d}_i^\top}{\|\mathbf{d}_i\|^2} + I_n \,:\, \alpha_1, \dots, \alpha_k > 0 \text{ and the } \mathbf{d}_i\text{'s are orthogonal} \right\}.$$

The relevant minimization problem thus becomes

$$(\mathbf{m}_k^*, \Sigma_k^*) = \arg\min \left\{ D(g^*, g_{\mathbf{m},\Sigma}) \,:\, \mathbf{m} \in \mathbb{R}^n,\ \Sigma \in \mathscr{L}_{n,k} \right\} \tag{5}$$

instead of Equation 2, with the effective dimension k being allowed to be adjusted dynamically. By restricting the space in which the variance is looked up, one seeks to limit the number of variance terms to be estimated. The idea is that if the directions are suitably chosen, then the improvement of the accuracy due to the smaller error in estimating the variance terms will compensate the fact that we consider less candidates for the covariance matrix. In (El Masri, Morio, and Simatos 2021), the authors consider $k = 1$ and $\mathbf{d}_1 = \mathbf{m}^*/\|\mathbf{m}^*\|$. When f is Gaussian, this choice is motivated by the fact that, due to the light tail of the Gaussian random variable and the reliability context, the variance should vary significantly in the direction of $\mathbf{m}^*$ and so estimating the variance in this direction can bring information. (In Section 3.4, we actually use the techniques of the present paper to provide a stronger theoretical justification of this choice, see Theorem 3.2 and the discussion following it). The method in (Uribe et al. 2021) is more involved: k is adjusted dynamically, while the directions $\mathbf{d}_i$ are the eigenvectors associated to the largest eigenvalues of a certain matrix. They span a low-dimensional subspace called Failure-Informed Subspace, and the authors in (Uribe et al. 2021) prove that this choice minimizes an upper bound on the minimal KL divergence. In practice, this algorithm yields very accurate

results. However, we will not consider it further in the present paper for two reasons. First, this algorithm is tailored for the reliability case where $\phi = \mathbb{I}_{\{\varphi \geq 0\}}$, with a function $\varphi : \mathbb{R}^n \to \mathbb{R}$, whereas our method is more general and applies to the general problem of estimating an integral (see for instance our test case of **?@sec-sub:payoff**). Second, the algorithm in (Uribe et al. 2021) requires the evaluation of the gradient of the function φ. However, this gradient is not always known and can be expensive to evaluate in high dimension; in some cases, the function φ is even not differentiable, as will be the case in our numerical example in **?@sec-sub:portfolio**. In contrast, our method makes no assumption on the form or smoothness of ϕ: it does not need to assume that it is of the form $\mathbb{I}_{\{\varphi \geq 0\}}$, or to assume that $\nabla\varphi$ is tractable. For completeness, whenever the algorithm of (Uribe et al. 2021) was applicable and computing the gradient of φ did not require any additional simulation budget, we have run it on the test cases considered here and found that it outperformed our algorithm. In more realistic settings, computing $\nabla\varphi$ would likely increase the simulation budget, and it would be interesting to compare the two algorithms in more details to understand when this extra computation cost is worthwhile. We reserve such a question for future research and will not consider the algorithm of (Uribe et al. 2021) further, as our aim in this paper is to establish benchmark results for a general algorithm which works for any function ϕ.

3.2 Main result of the paper

The main result of the present paper is to actually compute the exact value for Σ_k^* in Equation 5, which therefore paves the way for efficient high-dimensional estimation schemes. The statement of our result involves the following function ℓ, which is represented in Figure 1:

$$\ell : x \in (0, \infty) \mapsto -\log(x) + x - 1. \tag{6}$$

In the following, $(\lambda, \mathbf{d}) \in \mathbb{R} \times \mathbb{R}^n$ is an eigenpair of a matrix A if $A\mathbf{d} = \lambda\mathbf{d}$ and $\|\mathbf{d}\| = 1$. A diagonalizable matrix has n distinct eigenpairs, say $((\lambda_i^*, \mathbf{d}_i^*), i = 1, \dots, n)$, and we say that these eigenpairs are ranked in decreasing ℓ-order if $\ell(\lambda_1^*) \geq \cdots \geq \ell(\lambda_n^*)$.

```
##############################################################################################
# Figure 1. Plot of the function "l"
##############################################################################################

x = np.linspace(np.finfo(float).eps,4.0,100)
y = -np.log(x) + x -1

# plot
fig, ax = plt.subplots()

ax.plot(x, y, linewidth=2.0)

ax.set(xlim=(0, 4), xticks=[0,1,2,3],
       ylim=(0, 0.5), yticks=[0,0.5,1,1.5])
```

```
plt.grid()
plt.xlabel(r"$x$",fontsize=16)
plt.ylabel(r"$\ell(x)$",fontsize=16)
for tickLabel in plt.gca().get_xticklabels() + plt.gca().get_yticklabels():
    tickLabel.set_fontsize(16)
plt.show()
```

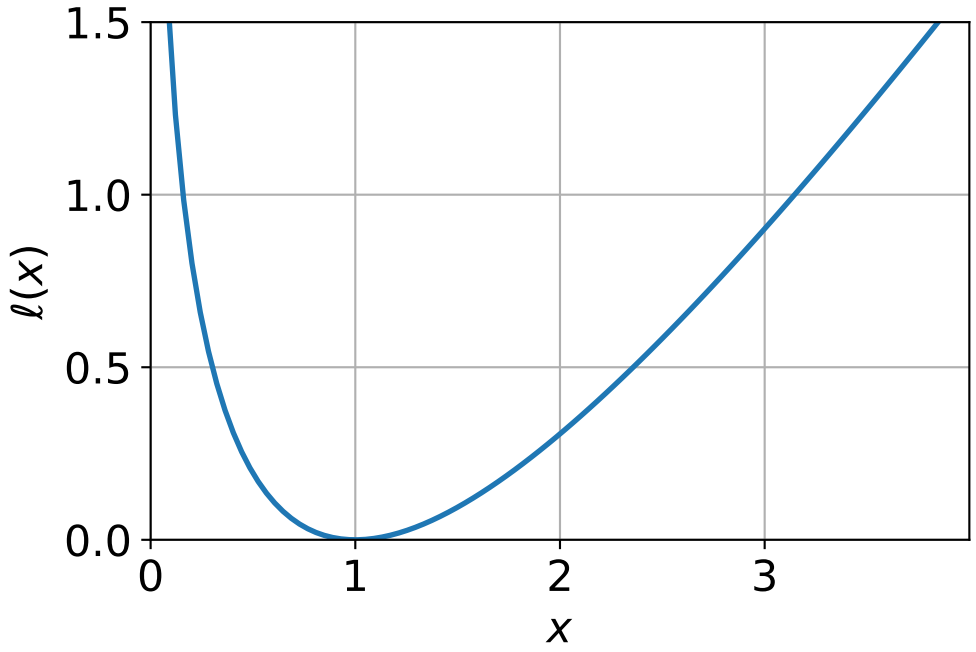


Figure 1: Plot of the function $\ell = -\log(x) + x - 1$ given by Equation 6

Theorem 3.1. *Let $(\lambda_i^*, \mathbf{d}_i^*)$ be the eigenpairs of Σ^* ranked in decreasing ℓ-order. Then for $1 \leq k \leq n$, the solution $(\mathbf{m}_k^*, \Sigma_k^*)$ to Equation 5 is given by*

$$\mathbf{m}_k^* = \mathbf{m}^* \quad \textit{and} \quad \Sigma_k^* = I_n + \sum_{i=1}^{k} \left(\lambda_i^* - 1\right) \mathbf{d}_i^* (\mathbf{d}_i^*)^\top. \tag{7}$$

For $k = 1$ for instance, the shape of the function ℓ depicted in Figure 1 implies that $\Sigma_1^* = I_n + (\lambda^* - 1)\mathbf{d}^*(\mathbf{d}^*)^\top$ with $(\lambda^*, \mathbf{d}^*)$ the eigenpair of Σ^* with λ^* either the largest or the smallest eigenvalue of Σ^*, depending on which one maximizes ℓ.

This theoretical result therefore suggests to reduce dimension by estimating eigenpairs of Σ^*, rank them in decreasing ℓ-order and then use the k first eigenpairs $((\hat{\lambda}_i^*, \hat{\mathbf{d}}_i^*), i = 1, \ldots, k)$ to build

the covariance matrix $\widehat{\Sigma}_k^* = \sum_{i=1}^k (\hat{\lambda}_i^* - 1)\hat{\mathbf{d}}_i^* (\hat{\mathbf{d}}_i^*)^\top + I_n$ and the corresponding auxiliary density. This scheme is summarized in Algorithm 1. The effective dimension k is obtained by Algorithm 2, see Section 3.3 below.

Algorithm 1 Algorithm suggested by Theorem 1.

Data: Sample sizes N and M
Result: Estimation $\hat{E}_N$ of integral E
Generate a sample $\mathbf{X}_1, \dots, \mathbf{X}_M$ on $\mathbb{R}^n$ independently according to g^*
Estimate $\widehat{\mathbf{m}}^*$ and $\widehat{\Sigma}^*$ defined in Equation 10 and Equation 11 with this sample
Compute the eigenpairs $(\hat{\lambda}_i^*, \hat{\mathbf{d}}_i^*)$ of $\widehat{\Sigma}^*$ ranked in decreasing ℓ-order
Compute the matrix $\widehat{\Sigma}_k^* = \sum_{i=1}^k (\hat{\lambda}_i^* - 1)\hat{\mathbf{d}}_i^* (\hat{\mathbf{d}}_i^*)^\top + I_n$ with k obtained by applying Algorithm 2 with input $(\hat{\lambda}_1^*, \dots, \hat{\lambda}_n^*)$
Generate a new sample $\mathbf{X}_1, \dots, \mathbf{X}_N$ independently from $g' = g_{\widehat{\mathbf{m}}^*, \widehat{\Sigma}_k^*}$
Return $\hat{E}_N = \dfrac{1}{N}\displaystyle\sum_{i=1}^N \phi(\mathbf{X}_i)\frac{f(\mathbf{X}_i)}{g'(\mathbf{X}_i)}$

Remark. The value 1 plays a particular role in Theorem 3.1, in that, as ℓ is minimized in 1, eigenvectors with eigenvalues 1 will only be selected once all other eigenvalues will have been picked: in other words, if $\lambda_i^* = 1$ then $\lambda_j^* = 1$ for all $j \geq i$. The reason why 1 plays this special role is due to the form of the covariance matrix that we impose. More precisely, looking for covariance matrices in the set $\mathscr{L}_{n,k}$ amounts to looking for covariance matrices which, once diagonalized, have one's on the diagonal except possibly for k values (the α_i's). As k will be small, typically $k = 1$ or 2, this amounts to looking for covariance matrices which are perturbation of the identity. This is particularly relevant as we assume f is a standard Gaussian density. What Theorem 3.1 tells is that, when trying to approximate Σ^* by such matrices, we should first consider eigenvectors with eigenvalues as different as possible from 1, the ''distance" to 1 being measured by ℓ. If one was imposing a different form on Σ_k^* (which can be interesting if the distribution f is not standardized), then a different result would arise. For instance, if one was looking for matrices where the ''default" choice would be some $\lambda > 0$ for the diagonal entries that are not estimated, i.e., a matrix of the form $\sum_i (\alpha_i - \lambda)\mathbf{d}_i \mathbf{d}_i^\top + \lambda I_n$, then eigenpairs would be ranked according to the function $\ell(\cdot/\lambda)$, meaning that one would look for eigenvectors associated to eigenvalues as different as possible from λ.

As mentioned above, we assume in the first step of Algorithm 1 that we can sample according to g^*. Since g^* is known up to a multiplicative constant, this is a reasonable assumption as classical techniques such as importance sampling with self-normalized weights or Markov Chain Monte–Carlo (MCMC) can be applied in this case (see for instance (Chan and Kroese 2012), (Grace, Kroese, and Sandmann 2014)). In this paper, we choose to apply a basic rejection method that yields perfect independent samples from g^*, possibly at the price of a high computational cost. As the primary goal of this paper is to understand whether the $\mathbf{d}_i^*$'s are indeed good projection directions, this computational cost to generate from g^* is not relevant for us and therefore not taken into account. Possible improvements to relax this assumption are discussed in the conclusion of the paper.

3.3 Choice of the number of dimensions k

The choice of the effective dimension k, i.e., the number of projection directions considered, is important. If it is close to n, then the matrix $\widehat{\Sigma}_k^*$ will be close to $\widehat{\Sigma}^*$ which is the situation we want to avoid in the first place. On the other hand, setting $k = 1$ in all cases may be too simple and lead to suboptimal results. In practice however, this is often a good choice. In order to adapt k dynamically, we consider a simple method based on the value of the KL divergence. Given the eigenvalues $\lambda_1, \dots, \lambda_n$ ranked in decreasing ℓ-order, we look for the maximal gap in the sequence $(\ell(\lambda_1), \dots, \ell(\lambda_n))$. This allows to choose k such that $\sum_{i=1}^{k} \ell(\lambda_i)$ is close to $\sum_{i=1}^{n} \ell(\lambda_i)$ which is equal, up to an additive constant, to the minimal KL divergence (see Equation 9 below). The precise method is described in Algorithm 2.

Algorithm 2 Choice of the number of dimensions

Data: Sequence of positive numbers $\lambda_1, \dots, \lambda_n$ in decreasing ℓ-order
Result: Number of selected dimensions k
Compute the increments $\delta_i = \ell(\lambda_{i+1}) - \ell(\lambda_i)$ for $i = 1 \dots n-1$
Return $k = \arg\max \delta_i$, the index of the maximum of the differences.

3.4 Theoretical result concerning the projection on $\mathbf{m}^*$

In (El Masri, Morio, and Simatos 2021), the authors propose to project on the mean $\mathbf{m}^*$ of the optimal auxiliary density g^*. Numerically, this algorithm is shown to perform well, but only a very heuristic explanation based on the light tail of the Gaussian distribution is provided to motivate this choice. It turns out that the techniques used in the proof of Theorem 3.1 can shed light on why projecting on $\mathbf{m}^*$ may indeed be a good idea. Let us first state our theoretical result, and then explain why it justifies the idea of projecting on $\mathbf{m}^*$.

Theorem 3.2. *Consider $\Sigma \in \mathscr{L}_{n,1}$ of the form $\Sigma = I_n + (\alpha - 1)\mathbf{d}\mathbf{d}^\top$ with $\alpha > 0$ and $\|\mathbf{d}\| = 1$. Then the minimizer in $(\alpha, \mathbf{d})$ of the KL divergence between f and $g_{\mathbf{m}^*,\Sigma}$ is $(1 + \|m^*\|^2, \mathbf{m}^*/\|\mathbf{m}^*\|)$:*

$$\left(1 + \|\mathbf{m}^*\|^2, \mathbf{m}^*/\|\mathbf{m}^*\|\right) = \arg\min_{\alpha,\mathbf{d}} \left\{ D(f, g_{\mathbf{m}^*, I_n + (\alpha-1)\mathbf{d}\mathbf{d}^\top}) \,:\, \alpha > 0,\ \|\mathbf{d}\| = 1 \right\}.$$

In other words, $\mathbf{m}^*$ appears as an optimal projection direction when one seeks to minimize the KL divergence between f and the Gaussian density with mean $\mathbf{m}^*$ and covariance of the form $I_n + (\alpha - 1)\mathbf{d}\mathbf{d}^\top$. Let us now explain why this minimization problem is indeed relevant, and why choosing an auxiliary density which minimizes this KL divergence may indeed lead to an accurate estimation. The justification deeply relies on the recent results by Chatterjee and Diaconis (Chatterjee and Diaconis 2018).

As mentioned above, in a reliability context where one seeks to estimate a small probability $p = \mathbb{P}(\mathbf{X} \in A)$, Theorem 1.3 in (Chatterjee and Diaconis 2018) shows that $D(g^*, g)$ governs the sample size required for an accurate estimation of p: more precisely, the estimation is accurate

if the sample size is larger than $e^{D(g^*,g)}$, and inaccurate otherwise. This motivates the rationale for minimizing the KL divergence with g^*.

However, in high dimension, importance sampling is known to fail because of the weight degeneracy problem whereby $\max_i L_i / \sum_i L_i \approx 1$, with the L_i's the unnormalized importance weights, or likelihood ratios: $L_i = f(\mathbf{X}_i)/g(\mathbf{X}_i)$ with the $\mathbf{X}_i$'s i.i.d. drawn according to g. Theorem 2.3 in (Chatterjee and Diaconis 2018) shows that the weight degeneracy problem is avoided if the empirical mean of the likelihood ratios is close to 1, and for this, Theorem 1.1 in (Chatterjee and Diaconis 2018) shows that the sample size should be larger than $e^{D(f,g)}$. In other words, these results suggest that the KL divergence with g^* governs the sample size for an accurate estimation of p, while the KL divergence with f governs the weight degeneracy problem.

In light of these results, it becomes natural to consider the KL divergence with f and not only g^*. Of course, minimizing $D(f, g_{\mathbf{m},\Sigma})$ without constraints on $\mathbf{m}$ and Σ is trivial since $g_{\mathbf{m},\Sigma} = f$ for $\mathbf{m} = 0$ and $\Sigma = I_n$. However, these choices are the ones we want to avoid in the first place, and so it makes sense to impose some constraints on $\mathbf{m}$ and Σ. If one keeps in mind the other objective of getting close to g^*, then the choice $\mathbf{m} = \mathbf{m}^*$ becomes very natural, and we are led, when $\Sigma \in \mathscr{S}_{n,1}$ is sought as a rank-1 perturbation of the identity, to considering the optimization problem of Theorem 3.2.

4 Proof of Theorems Theorem 3.1 and Theorem 3.2

We begin with a preliminary lemma.

Lemma 4.1. *Let f be the density of the standard Gaussian vector in dimension n, $\phi : \mathbb{R}^n \to \mathbb{R}_+$ and $g_* = f\phi/E$ with $E = \int f\phi$. Then for any $\mathbf{m}$ and any Σ of the form $\Sigma = I_n + \sum_i (\alpha_i - 1)\mathbf{d}_i\mathbf{d}_i^\top$ with $\alpha_i > 0$ and the $\mathbf{d}_i$'s orthonormal, we have*

$$D(g^*, g_{\mathbf{m},\Sigma}) = \frac{1}{2}\sum_i \left(\log \alpha_i - \left(1 - \frac{1}{\alpha_i}\right)\mathbf{d}_i^\top \Sigma^* \mathbf{d}_i\right) + \frac{1}{2}(\mathbf{m} - \mathbf{m}^*)^\top \Sigma^{-1}(\mathbf{m} - \mathbf{m}^*) - \frac{1}{2}\|\mathbf{m}^*\|^2 - \log E + \mathbb{E}_{g^*}(\log \phi(\mathbf{X})). \tag{8}$$

of Lemma 4.1. For any $\mathbf{m} \in \mathbb{R}^n$ and $\Sigma \in \mathscr{S}_n^+$, we have by definition

$$D(g^*, g_{\mathbf{m},\Sigma}) = \mathbb{E}_{g^*}\left(\log\left(\frac{g^*(\mathbf{X})}{g_{\mathbf{m},\Sigma}(\mathbf{X})}\right)\right) = \mathbb{E}_{g^*}\left(\log\left(\frac{\frac{\phi(\mathbf{X})e^{-\frac{1}{2}\|\mathbf{X}\|^2}}{E(2\pi)^{d/2}}}{\frac{e^{-\frac{1}{2}(\mathbf{X}-\mathbf{m})^\top\Sigma^{-1}(\mathbf{X}-\mathbf{m})}}{(2\pi)^{d/2}|\Sigma|^{1/2}}}\right)\right)$$

and so

$$D(g^*, g_{\mathbf{m},\Sigma}) = -\frac{1}{2}\mathbb{E}_{g^*}(\|\mathbf{X}\|^2) + \frac{1}{2}\mathbb{E}_{g^*}\left((\mathbf{X} - \mathbf{m})^\top\Sigma^{-1}(\mathbf{X} - \mathbf{m})\right) + \frac{1}{2}\log|\Sigma| - \log E + \mathbb{E}_{g^*}(\log \phi(\mathbf{X})).$$

Because $\mathbb{E}_{g^*}(\mathbf{X}) = \mathbf{m}^*$, we have $\mathbb{E}_{g^*}(\|\mathbf{X}\|^2) = \mathbb{E}_{g^*}(\|\mathbf{X} - \mathbf{m}^*\|^2) + \|\mathbf{m}^*\|^2$ and

$$\mathbb{E}_{g^*}\left((\mathbf{X}-\mathbf{m})^\top\Sigma^{-1}(\mathbf{X}-\mathbf{m})\right) = \mathbb{E}_{g^*}\left((\mathbf{X}-\mathbf{m}^*)^\top\Sigma^{-1}(\mathbf{X}-\mathbf{m}^*)\right) + (\mathbf{m}-\mathbf{m}^*)^\top\Sigma^{-1}(\mathbf{m}-\mathbf{m}^*).$$

In the following derivations, we use the linearity of the trace and of the expectation, which make these two operators commute, as well as the identity $a^\top b = \mathrm{tr}(ab^\top)$ for any two vectors a and b. With this caveat, we obtain

$$\mathbb{E}_{g^*}\left[\|\mathbf{X}-\mathbf{m}^*\|^2\right] = \mathbb{E}_{g^*}\left[\mathrm{tr}((\mathbf{X}-\mathbf{m}^*)(\mathbf{X}-\mathbf{m}^*)^\top)\right] = \mathrm{tr}(\Sigma^*)$$

and we obtain with similar arguments $\mathbb{E}_{g^*}((\mathbf{X}-\mathbf{m}^*)^\top\Sigma^{-1}(\mathbf{X}-\mathbf{m}^*)) = \mathrm{tr}(\Sigma^{-1}\Sigma^*)$. Consider now $\Sigma = I_n + \sum_i(\alpha_i - 1)\mathbf{d}_i\mathbf{d}_i^\top$ with $\alpha_i > 0$ and the $\mathbf{d}_i$'s orthonormal. Then the eigenvalues of Σ potentially different from 1 are the α_i's (α_i is the eigenvalue associated with $\mathbf{d}_i$), so that

$$\log|\Sigma| = \sum_i \log\alpha_i.$$

Moreover, we have $\Sigma^{-1} = I_n - \sum_i \beta_i\mathbf{d}_i\mathbf{d}_i^\top$ with $\beta_i = 1 - 1/\alpha_i$ and so

$$\mathrm{tr}(\Sigma^{-1}\Sigma^*) = \mathrm{tr}(\Sigma^*) - \sum_i \beta_i\mathbf{d}_i^\top\Sigma^*\mathbf{d}_i.$$

Gathering the previous relation, we finally obtain the desired result.

□

of Theorem 3.1. From Equation 8 we see that the only dependency of $D(g^*, g_{\mathbf{m},\Sigma})$ in $\mathbf{m}$ is in the quadratic term $(\mathbf{m}-\mathbf{m}^*)^\top\Sigma^{-1}(\mathbf{m}-\mathbf{m}^*)$. As Σ is definite positive, this term is ≥ 0, and so it is minimized for $\mathbf{m} = \mathbf{m}^*$. Next, we see that the derivative in α_i is given by (here and in the sequel, we see $D(g^*, g_{\mathbf{m},\Sigma})$ as a function of $\mathbf{v} = (\alpha_i)_i$ and $\mathbf{d} = (\mathbf{d}_i)_i$)

$$\frac{\partial D}{\partial\alpha_i}(\mathbf{v},\mathbf{d}) = \frac{1}{\alpha_i} - \frac{1}{\alpha_i^2}\mathbf{d}_i^\top\Sigma^*\mathbf{d}_i = \frac{1}{\alpha_i^2}\left(\alpha_i - \mathbf{d}_i^\top\Sigma^*\mathbf{d}_i\right).$$

Thus, for fixed $\mathbf{d}$, D is decreasing in α_i for $\alpha_i < \mathbf{d}_i^\top\Sigma^*\mathbf{d}_i$ and then increasing for $\alpha_i > \mathbf{d}_i^\top\Sigma^*\mathbf{d}_i$, which shows that, for fixed $\mathbf{d}$, it is minimized for $\alpha_i = \mathbf{d}_i^\top\Sigma^*\mathbf{d}_i$. For this value (and $\mathbf{m} = \mathbf{m}^*$) we have

$$D(g^*, g_{\mathbf{m}^*,\Sigma}) = \sum_{i=1}^k\left[\log(\mathbf{d}_i^\top\Sigma^*\mathbf{d}_i) + 1 - \mathbf{d}_i^\top\Sigma^*\mathbf{d}_i\right] + C = -\sum_{i=1}^k \ell(\mathbf{d}_i^\top\Sigma^*\mathbf{d}_i) + C \tag{9}$$

with $C = -\frac{1}{2}\|\mathbf{m}^*\|^2 - \log E + \mathbb{E}_{g^*}(\log\phi(\mathbf{X}))$ independent from the $\mathbf{d}_i$'s. Since ℓ is decreasing and then increasing, it is clear from this expression that in order to minimize D, one must choose the $\mathbf{d}_i$'s in order to either maximize or minimize $\mathbf{d}_i^\top\Sigma^*\mathbf{d}_i$, whichever maximizes ℓ. Since the variational characterization of eigenvalues shows that eigenvectors precisely solve this problem, we get the desired result.

□

of Theorem 3.2. In Equation 8, the $\mathbf{m}^*$ and the Σ^* that appear in the right-hand side are the mean and variance of the density g^* considered in the first argument of the Kullback–Leibler divergence. In particular, if we apply Equation 8 with $\phi \equiv 1$, we have $g^* = f$, and the $\mathbf{m}^*$ and Σ^* of the right-hand side become 0 and I_n, respectively, so that

$$D(f, g_{\mathbf{m},\Sigma}) = \frac{1}{2}\sum_i \left(\log \alpha_i - \left(1 - \frac{1}{\alpha_i}\right)\right) + \frac{1}{2}\mathbf{m}^\top \Sigma^{-1}\mathbf{m}.$$

Now, if we consider $\mathbf{m} = \mathbf{m}^*$ and $\Sigma = I + (\alpha - 1)\mathbf{d}\mathbf{d}^\top$, we obtain (using $\Sigma^{-1} = I - (1 - 1/\alpha)\mathbf{d}\mathbf{d}^\top$ as mentioned in the proof of Lemma 4.1)

$$D(f, g_{\mathbf{m}^*,\Sigma}) = \frac{1}{2}\left(\log \alpha - \left(1 - \frac{1}{\alpha}\right)\left(1 + (\mathbf{d}^\top \mathbf{m}^*)^2\right)\right) + \frac{1}{2}\|\mathbf{m}^*\|^2.$$

We have seen in the proof of Theorem 3.1 that the function $x \mapsto \log x + (1/x - 1)\gamma$ is minimized for $x = \gamma$ where it takes the value $-\ell(\gamma)$: $D(f, g_{\mathbf{m}^*,\Sigma})$ is therefore minimized for $\alpha = 1 + (\mathbf{d}^\top \mathbf{m}^*)^2$ and for this value, we have

$$D(f, g_{\mathbf{m}^*,\Sigma}) = -\frac{1}{2}\ell(1 + (\mathbf{d}^\top \mathbf{m}^*)^2) + \frac{1}{2}\|\mathbf{m}^*\|^2.$$

As ℓ is increasing in $[1, \infty)$, this last quantity is minimized by maximizing $(\mathbf{d}^\top \mathbf{m}^*)^2$, which is obtained for $\mathbf{d} = \mathbf{m}^*/\|\mathbf{m}^*\|$. The result is proved.

□

5 Framework for the numerical results

5.1 General framework

The objective of the numerical simulations is to evaluate the impact of the choice of the covariance matrix on the estimation accuracy of a high dimensional integral E. We compare in this section the estimation results for different choices of the auxiliary covariance matrix when the IS auxiliary density is Gaussian. To extend this comparison, we also compute the results when the IS auxiliary density is chosen with the von Mises–Fisher– Nakagami (vMFN) model recently proposed in (Papaioannou, Geyer, and Straub 2019) for high dimensional probability estimation.

In the following section we test these different models of auxiliary densities on five test cases, where f is a standard Gaussian density. This choice is not a theoretical limitation as we can in principle always come back to this case by transforming the vector $\mathbf{X}$ with isoprobabilistic transformations (see for instance (Hohenbichler and Rackwitz 1981), (Liu and Der Kiureghian 1986)).

The precise numerical framework that we will consider to assess the efficiency of the different auxiliary models is as follows. We assume first that M i.i.d. random samples $\mathbf{X}_1, \ldots, \mathbf{X}_M$ distributed from g^* are available from rejection sampling. From these samples, the parameters of the Gaussian and of the vMFN auxiliary density are computed to get an auxiliary density g'.

Finally, N samples are generated from g' to provide an estimation of E with IS. This procedure is summarized by the following stages:

1. Generate a sample $\mathbf{X}_1, \dots, \mathbf{X}_M$ independently according to g^*;
2. From $\mathbf{X}_1, \dots, \mathbf{X}_M$, compute the parameters of the auxiliary parametric density g';
3. Generate a new sample $\mathbf{X}_1, \dots, \mathbf{X}_N$ independently from g';
4. Estimate E with $\hat{E}_N = \frac{1}{N}\sum_{i=1}^{N} \phi(\mathbf{X}_i)\frac{f(\mathbf{X}_i)}{g'(\mathbf{X}_i)}$.

The number of samples M and N are respectively set to $M = 500$ and $N = 2000$. This procedure is then repeated 50 times to provide a mean estimation $\hat{E}$ of E. In the result tables, for each auxiliary density g' we report the corresponding value for the relative error $\hat{E}/E - 1$ and the coefficient of variation of the 50 iterations (the empirical standard deviation divided by E). As was established in the proof of Theorem 3.1, the KL divergence is, up to an additive constant, equal to $D'(\Sigma) = \log|\Sigma| + \mathrm{tr}(\Sigma^*\Sigma^{-1})$ which we will refer to as partial KL divergence. In the result tables, we also report thus the mean value of $D'(\Sigma)$ to analyse the relevancy of the auxiliary density $g_{\widehat{\mathbf{m}}^*,\Sigma}$ for these six choices of covariance matrix Σ. The next sections specify the different parameters of g' for the Gaussian model and for the vMFN model we have considered in the simulations.

5.2 Choice of the auxiliary density g' for the Gaussian model

The goal is to get benchmark results to assess whether one can improve estimations of Gaussian IS auxiliary density by projecting the covariance matrix Σ^* in the proposed directions $\mathbf{d}_i^*$. The algorithm that we study here (Algorithms 1+2) aims more precisely at understanding whether:

- projecting can improve the situation with respect to the empirical covariance matrix;
- the $\mathbf{d}_i^*$'s are good candidates, in particular compared to the choice $\mathbf{m}^*$ suggested in (El Masri, Morio, and Simatos 2021);
- what is the impact in making errors in estimating the eigenpairs $(\lambda_i^*, \mathbf{d}_i^*)$.

Let us define the estimate $\widehat{\mathbf{m}}^*$ of $\mathbf{m}^*$ from the M i.i.d. random samples $\mathbf{X}_1, \dots, \mathbf{X}_M$ distributed from g^* with

$$\widehat{\mathbf{m}}^* = \frac{1}{M}\sum_{i=1}^{M} \mathbf{X}_i. \tag{10}$$

In our numerical test cases, we will compare six different choices of Gaussian auxiliary distributions g' with mean $\widehat{\mathbf{m}}^*$ and the following covariance matrices (see Table 1):

1. Σ^*: the optimal covariance matrix given by Equation 3;
2. $\widehat{\Sigma}^*$: the empirical estimation of Σ^* given by

$$\widehat{\Sigma}^* = \frac{1}{M}\sum_{i=1}^{M} (\mathbf{X}_i - \widehat{\mathbf{m}}^*)(\mathbf{X}_i - \widehat{\mathbf{m}}^*)^\top. \tag{11}$$

The four other covariance matrices considered in the numerical simulations are of the form

$\sum_{i=1}^{k}(\nu_i - 1)\mathbf{d}_i\mathbf{d}_i^\top + I_n$ where ν_i is the variance of $\widehat{\Sigma}^*$ in the direction $\mathbf{d}_i$, $\nu_i = \mathbf{d}_i^\top \widehat{\Sigma}^* \mathbf{d}_i$. The considered choice of k and $\mathbf{d}_i$ gives the following covariance matrices:

3. $\widehat{\Sigma}_{\text{opt}}$ is obtained by choosing $\mathbf{d}_i = \mathbf{d}_i^*$ of Theorem 3.1, which is supposed to be perfectly known from Σ^* and k is computed with Algorithm 2;
4. $\widehat{\Sigma}_{\text{opt}}^{+\text{d}}$ is obtained by choosing $\mathbf{d}_i = \hat{\mathbf{d}}_i^*$ the i-th eigenvector of $\widehat{\Sigma}^*$ (in ℓ-order), which is an estimation of $\mathbf{d}_i^*$, and k is computed with Algorithm 2;
5. $\widehat{\Sigma}_{\text{mean}}$ is obtained by choosing $k = 1$ and $\mathbf{d}_1 = \mathbf{m}^*/\|\mathbf{m}^*\|$;
6. $\widehat{\Sigma}_{\text{mean}}^{+\text{d}}$ is obtained by choosing $k = 1$ and $\mathbf{d}_1 = \widehat{\mathbf{m}}^*/\|\widehat{\mathbf{m}}^*\|$, where $\widehat{\mathbf{m}}^*$ given by Equation 10.

The matrices $\widehat{\Sigma}_{\text{opt}}$ and $\widehat{\Sigma}_{\text{mean}}$ use the estimation $\widehat{\Sigma}^*$ but the actual directions $\mathbf{d}_i^*$ or $\mathbf{m}^*$, while the matrices $\widehat{\Sigma}_{\text{opt}}^{+\text{d}}$ and $\widehat{\Sigma}_{\text{mean}}^{+\text{d}}$ involve an additional estimation of the directions. By definition, Σ^* will give optimal results, while results for $\widehat{\Sigma}^*$ will deteriorate as the dimension increases, which is the well-known behavior which we try to improve. Moreover, for $\widehat{\Sigma}_{\text{mean}}$ and $\widehat{\Sigma}_{\text{opt}}$, the projection directions, if not known analytically, are obtained by a brute force Monte Carlo scheme with a very high simulation budget. Finally, we emphasize that Algorithm 1 corresponds to estimating and projecting on the $\mathbf{d}_i^*$'s, and so the matrix $\widehat{\Sigma}_k^*$ of Algorithm 1 is equal to the matrix $\widehat{\Sigma}_{\text{opt}}^{+\text{d}}$, i.e., $\widehat{\Sigma}_k^* = \widehat{\Sigma}_{\text{opt}}^{+\text{d}}$.

Table 1: Presentation of the six covariance matrices considered in the numerical examples. Except Σ^*, the five other matrices involve one or two estimations: $\widehat{\Sigma}^*$ is the empirical estimation of Σ^* given by Equation 11. The four others are obtained by projecting $\widehat{\Sigma}^*$ on: (i) the optimal directions $\mathbf{d}_i^*$ for $\widehat{\Sigma}_{\text{opt}}$; (ii) estimations $\hat{\mathbf{d}}_i^*$ of the optimal directions $\mathbf{d}_i^*$ for $\widehat{\Sigma}_{\text{opt}}^{+\text{d}}$; (iii) $\mathbf{m}^*$ for $\widehat{\Sigma}_{\text{mean}}$; (iv) the estimation $\widehat{\mathbf{m}}^*$ in Equation 3 of $\mathbf{m}^*$ for $\widehat{\Sigma}_{\text{mean}}^{+\text{d}}$. The subscript therefore indicates the choice for the projection direction, while the superscript +d indicates whether these directions are estimated or not.

	Σ^*	$\widehat{\Sigma}^*$	$\widehat{\Sigma}_{\text{opt}}$	$\widehat{\Sigma}_{\text{mean}}$	$\widehat{\Sigma}_{\text{opt}}^{+\text{d}}$	$\widehat{\Sigma}_{\text{mean}}^{+\text{d}}$
Initial covariance matrix	Σ^*	$\widehat{\Sigma}^*$	$\widehat{\Sigma}^*$	$\widehat{\Sigma}^*$	$\widehat{\Sigma}^*$	$\widehat{\Sigma}^*$
Projection directions (exact or estimated)	-	-	Exact	Exact	Esti-mated	Esti-mated
Choice for the projection direction	None	None	Opt	Mean	Opt	Mean

5.3 Choice of the auxiliary density g' for the von Mises–Fisher–Nakagami model

Von Mises–Fisher–Nakagami (vMFN) distributions were proposed in (Papaioannou, Geyer, and Straub 2019) as an alternative to the Gaussian parametric family to perform IS for high dimensional probability estimation. A random vector $\mathbf{X}$ drawn according to the vMFN distribution can be written as $\mathbf{X} = R\mathbf{A}$ where $\mathbf{A} = \frac{\mathbf{X}}{\|\mathbf{X}\|}$ is a unit random vector following the von Mises-Fisher distribution, and $R = \|\mathbf{X}\|$ is a positive random variable with a Nakagami distribution; further, R and $\mathbf{A}$ are independent. The vMFN pdf can be written as

$$g_{\text{vMFN}}(\mathbf{x}) = g_{\text{N}}(\|\mathbf{x}\|, p, \omega) \times g_{\text{vMF}}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}, \boldsymbol{\mu}, \kappa\right). \tag{12}$$

The density $g_{\text{N}}(\|\mathbf{x}\|, p, \omega)$ is the Nakagami distribution with shape parameter $p \geq 0.5$ and a spread parameter $\omega > 0$ defined by

$$g_{\text{N}}(\|\mathbf{x}\|, p, \omega) = \frac{2p^p}{\Gamma(p)\omega^p}\|\mathbf{x}\|^{2p-1}\exp\left(-\frac{p}{\omega}\|\mathbf{x}\|^2\right)$$

and the density $g_{\text{vMF}}(\frac{\mathbf{x}}{\|\mathbf{x}\|}, \boldsymbol{\mu}, \kappa)$ is the von Mises-Fisher distribution, given by

$$g_{\text{vMF}}\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}, \boldsymbol{\mu}, \kappa\right) = C_n(\kappa)\exp\left(\kappa\boldsymbol{\mu}^T\frac{\mathbf{x}}{\|\mathbf{x}\|}\right),$$

where $C_n(\kappa)$ is a normalizing constant, $\boldsymbol{\mu}$ is a mean direction $\boldsymbol{\mu}$ (with $\|\boldsymbol{\mu}\| = 1$) and $\kappa > 0$ is a concentration parameter $\kappa > 0$.

Choosing a vMFN distribution therefore amounts to choosing the parameters $p, \omega, \boldsymbol{\mu}$, and κ. There are therefore $n + 3$ parameters to estimate, which is a significant reduction compared to the $\frac{n(n+3)}{2}$ required parameters of the Gaussian model with full covariance matrix.

Following (Papaioannou, Geyer, and Straub 2019), given a sample $\mathbf{X}_1, \ldots, \mathbf{X}_M$ distributed from g^*, the parameters ω, p, $\boldsymbol{\mu}$ and κ are set in the following way in order to define g':

$$\widehat{\omega} = \frac{1}{M}\sum_{i=1}^{M}\|\mathbf{X}_i\|^2 \text{ and } \hat{p} = \frac{\widehat{\omega}^2}{\hat{\tau} - \widehat{\omega}^2} \text{ with } \hat{\tau} = \frac{1}{M}\sum_{i=1}^{M}\|\mathbf{X}_i\|^4$$

and

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{M}\frac{\mathbf{X}_i}{\|\mathbf{X}_i\|}}{\|\sum_{i=1}^{M}\frac{\mathbf{X}_i}{\|\mathbf{X}_i\|}\|} \text{ and } \hat{\kappa} = \frac{n\hat{\chi} - \hat{\chi}^3}{1 - \hat{\chi}^2} \text{ with } \hat{\chi} = \min\left(\left\|\frac{1}{M}\sum_{i=1}^{M}\frac{\mathbf{X}_i}{\|\mathbf{X}_i\|}\right\|, 0.95\right).$$

6 Numerical results on five test cases

The proposed numerical framework is applied on three examples that are often considered to assess the performance of importance sampling algorithms and also two test cases from the area of financial mathematics. Extended simulation results are given in the supplementary material associated with this article.

6.1 Test case 1: one-dimensional optimal projection

We consider a test case where all computations can be made exactly. This is a classical example of rare event probability estimation, often used to test the robustness of a method in high dimension. It is given by $\phi(\mathbf{x}) = \mathbb{I}_{\{\varphi(\mathbf{x})\geq 0\}}$ with φ the following affine function:

$$\varphi : \mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n \mapsto \sum_{j=1}^{n} x_j - 3\sqrt{n}. \tag{13}$$

The quantity of interest E is defined as $E = \int_{\mathbb{R}^n} \phi(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x} = \mathbb{P}_f(\varphi(\mathbf{X}) \geq 0) \simeq 1.35 \cdot 10^{-3}$ for all n where the density f is the standard n-dimensional Gaussian distribution. Here, the zero-variance density is $g^*(\mathbf{x}) = \dfrac{f(\mathbf{x})\mathbb{I}_{\{\varphi(\mathbf{x})\geq 0\}}}{E}$, and the optimal parameters $\mathbf{m}^*$ and Σ^* in Equation 3 can be computed exactly, namely $\mathbf{m}^* = \alpha\mathbf{1}$ with $\alpha = e^{-9/2}/(E(2\pi)^{1/2})$ and $\mathbf{1} = \frac{1}{\sqrt{n}}(1, \dots, 1) \in \mathbb{R}^n$ the normalized constant vector, and $\Sigma^* = (\nu - 1)\mathbf{1}\mathbf{1}^\top + I_n$ with $\nu = 3\alpha - \alpha^2 + 1$.

6.1.1 Evolution of the partial KL divergence and spectrum

Figure 2a represents the evolution as the dimension varies between 5 and 100 of the partial KL divergence D' for three different choices of covariance matrix: the optimal matrix Σ^*, its empirical estimation $\widehat{\Sigma}^*$ and the estimation $\widehat{\Sigma}^*_k$ of the optimal lower-dimensional covariance matrix. We can notice that the partial KL divergence for $\widehat{\Sigma}^*$ grows much faster than the other two, and that the partial KL divergence for $\widehat{\Sigma}^*_k$ remains very close to the optimal value $D'(\Sigma^*)$. As the KL divergence is a proxy for the efficiency of the auxiliary density (it is for instance closely related to the number of samples required for a given precision (Chatterjee and Diaconis 2018)), this suggests that using $\widehat{\Sigma}^*_k$ will provide results close to optimal.

We now check this claim. As $\Sigma^* = (\nu - 1)\mathbf{1}\mathbf{1}^\top + I_n$, its eigenpairs are $(\nu, \mathbf{1})$ and $(1, \mathbf{d}_i)$ where the $\mathbf{d}_i$'s form an orthonormal basis of the space orthogonal to the space spanned by $\mathbf{1}$. In particular, $(\nu, \mathbf{1})$ is the largest (in ℓ-order) eigenpair of Σ^* and $\Sigma^*_k = \Sigma^*$ for any $k \geq 1$.

In practice, we do not use this theoretical knowledge and Σ^*, Σ^*_k and the eigenpairs are estimated. The six covariance matrices introduced in Section 5.2 and in which we are interested are as follows:

- $\Sigma^* = (\nu - 1)\mathbf{1}\mathbf{1}^\top + I_n$;
- $\widehat{\Sigma}^*$ given by Equation 11;
- $\widehat{\Sigma}_{\text{opt}}$ and $\widehat{\Sigma}_{\text{mean}}$ are equal and given by $(\hat{\lambda} - 1)\mathbf{1}\mathbf{1}^\top + I_n$ with $\hat{\lambda} = \mathbf{1}^\top\widehat{\Sigma}^*\mathbf{1}$. This amounts to assuming that the projection direction $\mathbf{1}$ is perfectly known, whereas the variance in this direction is estimated;
- $\widehat{\Sigma}^{+\text{d}}_{\text{opt}} = (\hat{\lambda} - 1)\hat{\mathbf{d}}\hat{\mathbf{d}}^\top + I_n$ with $(\hat{\lambda}, \hat{\mathbf{d}})$ the smallest eigenpair of $\widehat{\Sigma}^*$. The difference with the previous case is that we do not assume anymore that the optimal projection direction $\mathbf{1}$ is known, and so it needs to be estimated;
- $\widehat{\Sigma}^{+\text{d}}_{\text{mean}} = (\hat{\lambda} - 1)\frac{\widehat{\mathbf{m}}^*(\widehat{\mathbf{m}}^*)^\top}{\|\widehat{\mathbf{m}}^*\|^2} + I_n$ with $\widehat{\mathbf{m}}^*$ given by Equation 10 and $\hat{\lambda} = \frac{(\widehat{\mathbf{m}}^*)^\top\widehat{\Sigma}^*\widehat{\mathbf{m}}^*}{\|\widehat{\mathbf{m}}^*\|^2}$. Here we assume that $\mathbf{m}^*$ is a good projection direction, but is unknown and therefore needs to be estimated.

Note that in the particularly simple case considered here, both $\widehat{\mathbf{m}}^*/\|\widehat{\mathbf{m}}^*\|$ and $\hat{\mathbf{d}}$ are estimators of $\mathbf{1}$ but they are obtained by different methods. In the next example we will consider a case where $\mathbf{m}^*$ is not an optimal projection direction as given by Theorem 3.1.

Figure 2b represents the images by ℓ of the eigenvalues of Σ^* and $\widehat{\Sigma}^*$. This picture carries a very important insight. We notice that the estimation of most eigenvalues is poor: indeed, all the blue crosses except the leftmost one are meant to be estimator of 1, whereas we see that they are more or less uniformly spread between 0.4 and 1.8. This means that the variance terms in the corresponding directions are poorly estimated, which could be the explanation on why the use of $\widehat{\Sigma}^*$ gives an inaccurate estimation. But what we remark also is that the function ℓ is quite flat around one: as a consequence, although the eigenvalues offer significant variability, this variability is smoothed by the action of ℓ. Indeed, the images of the eigenvalues by ℓ take values between 0 and 0.4 and have smaller variability. Moreover, $\ell(x)$ increases sharply as x approaches 0 and thus efficiently distinguishes between the two leftmost estimated eigenvalues and is able to separate them.

```
##########################################################################################################
# Figure 2. Evolution of the partial KL divergence and spectrum of the eigenvalues for the tes
##########################################################################################################

def Somme(x):
    n=np.shape(x)[1]
    return(np.sum(x,axis=1)-3*np.sqrt(n))

n=100          # dimension
phi=Somme
E=sp.stats.norm.cdf(-3)    # exact value of the integral

DKL=np.zeros(20)
DKLp=np.zeros(20)
DKLm=np.zeros(20)
DKLstar=np.zeros(20)

M=300

for d in range(5,n+1,5):
    # Mstar
    alpha=np.exp(-3**2/2)/(E*np.sqrt(2*np.pi))
    Mstar=alpha*np.ones(d)/np.sqrt(d)
    # Sigmastar
    vstar=3*alpha-alpha**2+1
    Sigstar= (vstar-1)*np.ones((d,d))/d+np.eye(d)
```

```
    ## g*-sample
    VA0=sp.stats.multivariate_normal(mean=np.zeros(d),cov=np.eye(d))
    X0=VA0.rvs(size=M*1000)

    ind=(phi(X0)>0)
    X=X0[ind,:]
    X=X[:M,:]            # g*-sample of size M

    ## estimated mean and covariance
    mm=np.mean(X,axis=0)

    Xc=(X-mm).T
    sigma =Xc @ Xc.T/np.shape(Xc)[1]

    ## projection with the eigenvalues of sigma
    Eig=np.linalg.eigh(sigma)
    logeig=np.sort(np.log(Eig[0])-Eig[0])
    delta=np.zeros(len(logeig)-1)
    for j in range(len(logeig)-1):
        delta[j]=abs(logeig[j]-logeig[j+1])

    k=np.argmax(delta)+1         # biggest gap between the l(lambda_i)

    indi=[]
    for l in range(k):
        indi.append(np.where(np.log(Eig[0])-Eig[0]==logeig[l])[0][0])

    P1=np.array(Eig[1][:,indi[0]],ndmin=2).T
    for l in range(1,k):
        P1=np.concatenate((P1,np.array(Eig[1][:,indi[l]],ndmin=2).T),axis=1)   # matrix of ini

    diagsi=np.diag(Eig[0][indi])
    sig_opt_d=P1.dot((diagsi-np.eye(k))).dot(P1.T)+np.eye(d)

    DKL[int((d-5)/5)]=np.log(np.linalg.det(sigma))+np.sum(np.diag(Sigstar.dot(np.linalg.inv(si
    DKLp[int((d-5)/5)]=np.log(np.linalg.det(sig_opt_d))+np.sum(np.diag(Sigstar.dot(np.linalg.i
    DKLstar[int((d-5)/5)]=np.log(np.linalg.det(Sigstar))+d

#### plot of partial KL divergence
plt.plot(range(5,105,5),DKL,'rs',label=r"$D'(\hat{\Sigma}^*)$")
plt.plot(range(5,105,5),DKLp,'k^',label=r"$D'(\hat{\Sigma}^*_k)$")
plt.plot(range(5,105,5),DKLstar,'bo',label=r"$D'(\Sigma^*)$")
```

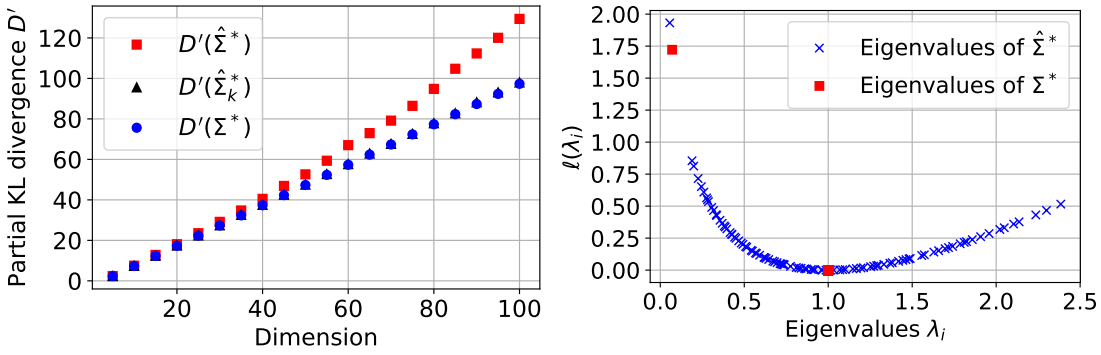
```
plt.grid()
plt.xlabel('Dimension',fontsize=16)
plt.ylabel(r"Partial KL divergence $D'$",fontsize=16)
plt.legend(fontsize=16)
for tickLabel in plt.gca().get_xticklabels() + plt.gca().get_yticklabels():
    tickLabel.set_fontsize(16)
plt.show()

#### plot of the eigenvalues
Eig1=np.linalg.eigh(sigma)
logeig1=np.log(Eig1[0])-Eig1[0]+1
Table_eigv=np.zeros((n,2))
Table_eigv[:,0]=Eig1[0]
Table_eigv[:,1]=-logeig1

Eigst=np.linalg.eigh(Sigstar)
logeigst=np.log(Eigst[0])-Eigst[0]+1
Table_eigv_st=np.zeros((n,2))
Table_eigv_st[:,0]=Eigst[0]
Table_eigv_st[:,1]=-logeigst

plt.grid()
plt.xlabel(r"Eigenvalues $\lambda_i$",fontsize=16)
plt.ylabel(r"$\ell(\lambda_i)$",fontsize=16)
for tickLabel in plt.gca().get_xticklabels() + plt.gca().get_yticklabels():
    tickLabel.set_fontsize(16)

plt.plot(Table_eigv[:,0],Table_eigv[:,1],'bx',label=r"Eigenvalues of $\hat{\Sigma}^*$")
plt.plot(Table_eigv_st[:,0],Table_eigv_st[:,1],'rs',label=r"Eigenvalues of $\Sigma^*$")
plt.legend(fontsize=16)
plt.show()
```


Agapiou, S., O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. 2017. "Importance Sampling : Intrinsic Dimension and Computational Cost." *Statistical Science, Volume 32, P405-431.* https://doi.org/10.1214/17-STS611.

Ashurbekova, Karina, Antoine Usseglio-Carleve, Florence Forbes, and Sophie Achard. 2020. "Optimal Shrinkage for Robust Covariance Matrix Estimators in a Small Sample Size Setting."

Au, S. K., and J. L. Beck. 2003. "Important Sampling in High Dimensions." *Structural Safety* 25 (2): 139–63. https://doi.org/10.1016/S0167-4730(02)00047-4.

Bengtsson, Thomas, Peter Bickel, and Bo Li. 2008. "Curse-of-Dimensionality Revisited: Collapse of the Particle Filter in Very Large Scale Systems." In *Institute of Mathematical Statistics Collections*, 316–34. Beachwood, Ohio, USA: Institute of Mathematical Statistics. https://doi.org/10.1214/193940307000000518.



(a) Evolution of the partial KL divergence as the dimension increases, with the optimal covariance matrix Σ^* (blue circles), the sample covariance $\widehat{\Sigma}^*$ (red squares), and the projected covariance $\widehat{\Sigma}_k^*$ (black triangles).

(b) Computation of $\ell(\lambda_i)$ for the eigenvalues of Σ^* (red squares) and $\widehat{\Sigma}^*$ (blue crosses) in dimension $n = 100$.

Figure 2: Partial KL divergence and spectrum for the function $\phi = \mathbb{I}_{\varphi \geq 0}$ with φ the linear function given by Equation 13.

Bucklew, James. 2013. *Introduction to Rare Event Simulation.* Springer Science & Business Media. https://doi.org/10.1007/978-1-4757-4078-3.

Bugallo, Monica F., Victor Elvira, Luca Martino, David Luengo, Joaquin Miguez, and Petar M. Djuric. 2017. "Adaptive Importance Sampling: The Past, the Present, and the Future." *IEEE Signal Processing Magazine* 34 (4): 60–79. https://doi.org/10.1109/MSP.2017.2699226.

Cappé, Olivier, Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P. Robert. 2008. "Adaptive Importance Sampling in General Mixture Classes." *Statistics and Computing* 18 (4): 447–59. https://doi.org/10.1007/s11222-008-9059-x.

Chan, Joshua C. C., and Dirk P. Kroese. 2012. "Improved Cross-Entropy Method for Estimation." *Statistics and Computing* 22 (5): 1031–40. https://doi.org/10.1007/s11222-011-9275-7.

Chatterjee, Sourav, and Persi Diaconis. 2018. "The Sample Size Required in Importance Sampling." *Ann. Appl. Probab.* 28 (2): 1099–1135. https://doi.org/10.1214/17-AAP1326.

Cornuet, Jean-Marie, Jean-Michel Marin, Antonietta Mira, and Christian P. Robert. 2012. "Adaptive Multiple Importance Sampling: *Adaptive Multiple Importance Sampling*." *Scandinavian Journal of Statistics* 39 (4): 798–812. https://doi.org/10.1111/j.1467-9469.2011.00756.x.

El Masri, Maxime, Jérôme Morio, and Florian Simatos. 2021. "Improvement of the Cross-Entropy Method in High Dimension for Failure Probability Estimation Through a One-Dimensional Projection Without Gradient Estimation." *Reliability Engineering & System Safety* 216: 107991. https://doi.org/10.1016/j.ress.2021.107991.

El-Laham, Yousef, Víctor Elvira, and Mónica Bugallo. 2019. "Recursive Shrinkage Covariance Learning in Adaptive Importance Sampling." In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 624–28. IEEE. https://doi.org/10.1109/CAMSAP45676.2019.9022450.

Grace, Adam W., Dirk P. Kroese, and Werner Sandmann. 2014. "Automated State-Dependent

Importance Sampling for Markov Jump Processes via Sampling from the Zero-Variance Distribution." *Journal of Applied Probability* 51 (3): 741–55. https://doi.org/10.1239/jap/1409932671.

Hohenbichler, Michael, and Rüdiger Rackwitz. 1981. "Non-Normal Dependent Vectors in Structural Safety." *Journal of the Engineering Mechanics Division* 107 (6): 1227–38. https://doi.org/10.1061/JMCEA3.0002777.

Ledoit, Olivier, and Michael Wolf. 2004. "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices." *Journal of Multivariate Analysis* 88 (2): 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4.

Liu, Pei-Ling, and Armen Der Kiureghian. 1986. "Multivariate Distribution Models with Prescribed Marginals and Covariances." *Probabilistic Engineering Mechanics* 1 (2): 105–12. https://doi.org/10.1016/0266-8920(86)90033-0.

Owen, Art, and Yi Zhou. 2000. "Safe and Effective Importance Sampling." *Journal of the American Statistical Association* 95 (449): 135–43. https://doi.org/10.1080/01621459.2000.10473909.

Papaioannou, Iason, Sebastian Geyer, and Daniel Straub. 2019. "Improved Cross Entropy-Based Importance Sampling with a Flexible Mixture Model." *Reliability Engineering & System Safety* 191 (November): 106564. https://doi.org/10.1016/j.ress.2019.106564.

Rubinstein, Reuven Y., and Dirk P. Kroese. 2017. *Simulation and the Monte Carlo Method*. Third edition. Wiley Series in Probability and Statistics. Hoboken, New Jersey: Wiley. https://doi.org/10.1002/9781118631980.

Rubinstein, Reuven Y, and Dirk P Kroese. 2011. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. New York; London: Springer. https://doi.org/10.1007/978-1-4757-4321-0.

Uribe, Felipe, Iason Papaioannou, Youssef M. Marzouk, and Daniel Straub. 2021. "Cross-Entropy-Based Importance Sampling with Failure-Informed Dimension Reduction for Rare Event Simulation." *SIAM/ASA Journal on Uncertainty Quantification* 9 (2): 818–47. https://doi.org/10.1137/20M1344585.