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
```

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. **?@sec-proof** presents the proof of our theoretical result; **?@sec-num-results-framework** introduces the numerical framework that we have adopted, and **?@sec-test-cases** 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):

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

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 $(\mathbf{d}_{k+1},\ldots,\mathbf{d}_n)$ complete $(\mathbf{d}_1,\ldots,\mathbf{d}_k)$ into an orthonormal basis. In particular, ν_i is the

variance in the direction of $\mathbf{d}_i$, i.e., $v_i = \mathbf{d}_i^\top \Sigma \mathbf{d}_i$. 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 **?@sec-mm**, we actually use the techniques of the present paper to provide a stronger theoretical justification of this choice, see **?@thm-thm2** 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()
```

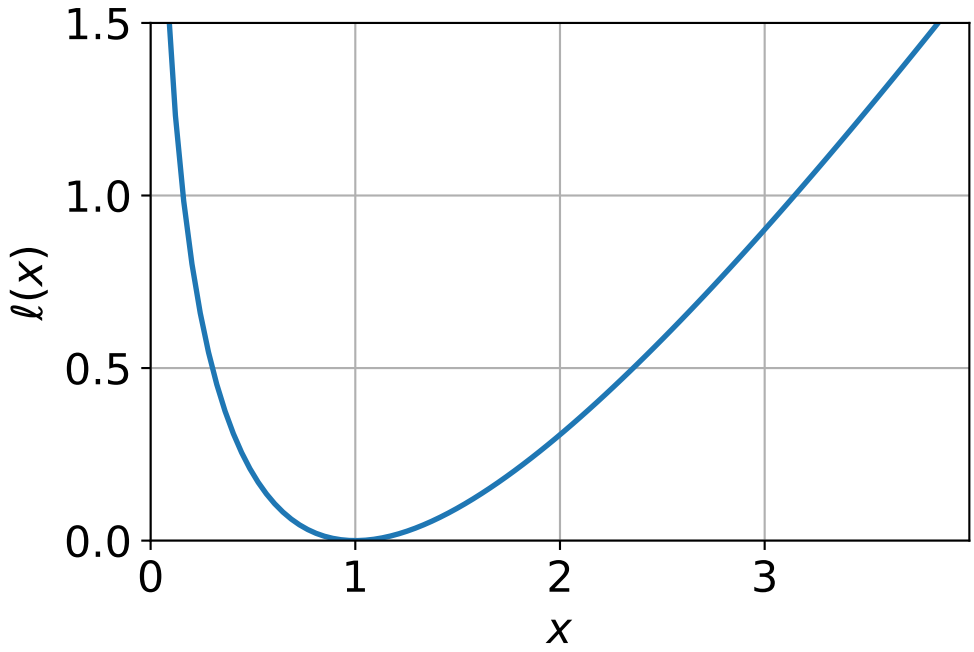

Figure 1: Plot of the function $\ell = -\log(x) + x - 1$ given by Equation 6

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.

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.

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.

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.

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.