Importance Sampling.
The research of optimization problems

Alena Shilova

August 2017

Introduction

Let's consider the problem of estimating the value of the following integral:

$$I = \mathbb{E}\varphi(X) = \int \varphi(x)f(x)\,dx,$$

where $X \sim f$ is a random variable on $\mathbb{R}^k$ and $\varphi : \mathbb{R}^k \to \mathbb{R}$.

It can be used the traditional unbiased Monte-Carlo estimator to approximate this value, but it may demand a large amount of points to generate to get a relatively close value to one we want. That's why one can use importance sampling estimator instead.

To apply it, firstly, it is necessary to choose a sampling (importance) distribution $\tilde{f}$ satisfying $\tilde{f}(x) > 0$ whenever $\varphi(x)f(x) \neq 0$, take IID samples $X_1, X_2, \ldots, X_n \sim \tilde{f}$ (as opposed to sampling from f, the nominal distribution) and use

$$\hat{I}_n^{IS} = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i)\frac{f(X_i)}{\tilde{f}(X_i)}.$$

Again, the estimator is unbiased, i.e., $\mathbb{E}\hat{I}_n^{IS} = I$. When $\tilde{f} = f$, importance sampling reduces to standard Monte Carlo. Choosing $\tilde{f}$ wisely can reduce the variance and accelerate the convergence, but this can be difficult in general. One of the approaches is to automate the process of finding the sampling distribution through adaptive importance sampling.

In adaptive importance sampling, one adaptively improves the sampling distribution while simultaneously accumulating the estimate for I. A particular form of importance sampling generates a sequence of sampling distributions $\tilde{f}_1, \tilde{f}_2, \ldots$ and a series of samples $X_1 \sim \tilde{f}_1$, $X_2 \sim \tilde{f}_2$, $\ldots$ and forms the estimate

$$\hat{I}_n^{AIS} = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i)\frac{f(X_i)}{\tilde{f}_i(X_i)}.$$

At each iteration n, the sampling distribution $\tilde{f}_n$, which is itself random, is adaptively determined based on the past data, $\tilde{f}_1, \ldots, \tilde{f}_{n-1}$ and $X_1, \ldots, X_{n-1}$. Again, $\hat{I}_n^{AIS}$ is unbiased, i.e. $\mathbb{E}\hat{I}_n^{AIS} = I$, and

$$\mathbb{V}(\hat{I}_n^{AIS}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}_{\tilde{f}_i}\mathbb{V}_{X_i\sim\tilde{f}_i}\frac{\varphi(X_i)f(X_i)}{\tilde{f}_i(X_i)} = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}_{\tilde{f}_i}\left(\int \frac{\varphi^2(x)f^2(x)}{\tilde{f}_i(x)}\,dx - I^2\right),$$

where $\mathbb{E}_{\tilde{f}_i}$ denotes the expectation over the random sampling distribution $\tilde{f}_i$. When $\tilde{f}_i = \tilde{f}$ for all i, adaptive importance sampling reduces to standard (non-adaptive) importance sampling. Now determining how to choose $\tilde{f}_n$ at each iteration fully specifies the method.

For simplicity further we are going to consider the distributions from the exponential family of distributions $\mathcal{F}$. Define $T : \mathbb{R}^k \to \mathbb{R}^p$ and $h : \mathbb{R}^k \to \mathbb{R}^+$. Then our density function is

$$f_\theta(x) = \exp\left[\theta^T T(x) - A(\theta)\right] h(x),$$

where $A : \mathbb{R}^p \to \mathbb{R} \cup \infty$, defined as

$$A(\theta) = \log \int \exp(\theta^T T(x)) h(x)\, dx,$$

serves as a normalizing factor. (When $A(\theta) = \infty$, we define $f_\theta = 0$ and remember that this does not define a distribution.) Finally, let $\Theta \subseteq \mathbb{R}^p$ be a convex set, and our exponential family is $\mathcal{F} = \left\{ f_\theta \;\middle|\; \theta \in \Theta \right\}$, where θ is called the natural parameter of $\mathcal{F}$. Note that the choice of T, h, and Θ fully specifies our family $\mathcal{F}$.

Optimization problems

In order to find the optimal distribution for the importance sampling estimator, one can try to minimize Rényi generalized divergence, which can be defined as the following:

$$D_\alpha(g; f) = \begin{cases} \int \ln \frac{g(x)}{f(x)} g(x)\, dx & \alpha = 1 \\ \frac{1}{\alpha - 1} \ln \left(\int \left[\frac{g(x)}{f(x)} \right]^{\alpha - 1} g(x)\, dx \right) & \alpha > 0; \alpha \neq 1 \end{cases}$$

Further we will consider two special cases of it. Firstly, if we take $\alpha = 1$, we will get the task of minimizing Kullback–Leibler divergence, which can be easily shown that it is the same as solving the following problem

$$\max_\theta \int g(x) \ln f_\theta(x)\, dx.$$

Secondly, another case emerges while trying minimizing Rényi divergence with $\alpha = 2$, which is equivalent to minimization of the per-sample variance of the importance sampled estimator with sampling distribution f_θ as it was shown in [1].

Hence, there are two optimization problems:

$$\max_{\theta \in \Theta} \frac{1}{I} \int \varphi(x) f(x) \left[\theta^T T(x) - A(\theta)\right]\, dx,$$

or

$$\min_{\theta \in \Theta} \frac{1}{I} \int \varphi(x) f(x) \left[A(\theta) - \theta^T T(x)\right]\, dx,$$

and

$$\min_{\theta \in \Theta} \int \frac{\varphi^2(x) f^2(x)}{\exp(\theta^T T(x) - A(\theta)) h(x)}\, dx$$

or

$$\min_{\theta \in \Theta} \int \varphi^2(x) f^2(x) \exp(A(\theta) - \theta^T T(x)) h(x)\, dx$$

Both problems are convex because as it was shown in [2] $A(\theta)$ is a convex function, which implies convexity of the objectives mentioned above.

Let's narrow our problems one more time and now consider only the family of normal distributions $\mathcal{N}(\mu, \Sigma)$ with parameters (μ, Σ) and the eigenvalues of the covariance matrix Σ lie within some compact $[\lambda_{min}, \lambda_{max}]$. Its density function looks like the following:

$$f_{m,S}(x) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right).$$

These distributions can be represented as ones from the exponential family by performing change of variables: $S = \Sigma^{-1}$ and $m = \Sigma^{-1}\mu$. Then the density of distribution will look like the following:

$$g_{m,S}(x) = \frac{1}{\sqrt{(2\pi)^n}} \exp\Big(m^T x - \frac{1}{2}\operatorname{Tr}(Sxx^T)\Big) \exp\Big(-\frac{1}{2}\big(m^T S^{-1} m - \log|S|\big)\Big)$$

and in this case $A(m,S) = \frac{1}{2}m^T S^{-1} m - \frac{1}{2}\log|S|$, $T(x) = (x, -\frac{1}{2}xx^T)$ and $h(x) = \frac{1}{\sqrt{(2\pi)^n}}$.

Study of optimization problem

Minimization of Kullback-Leibler divergence

From this point, let's assume that $\exists \lambda_{min}, \lambda_{max} > 0$, s.t. $\lambda(\Sigma) \subseteq [\lambda_{min}, \lambda_{max}]$. This means that $\lambda(S) \subseteq [\frac{1}{\lambda_{max}}, \frac{1}{\lambda_{min}}]$. Taking all of this into account, let's find out if the problems formulated above strongly convex.

To start with, we are going to explore the first optimization problem. Taking the family of normal distributions into consideration, we've got the following look of the problem

$$\min_{m,S} \int \varphi(x) f(x) \left[\frac{1}{2}x^T S x - m^T x + \frac{1}{2}m^T S^{-1} m - \frac{1}{2}\log|S|\right] dx.$$

It is quite obvious that $\frac{1}{2}x^T S x - m^T x$ is convex w.r.t. distribution parameters m, S. One can easily check the convexity of $f(x) = \frac{1}{2}m^T S^{-1} m$ using the following criterion:

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq 0$$

Indeed, here $\nabla f(x) = (S^{-1}m, -\frac{1}{2}S^{-1}mm^T S^{-1})$, so

$$\frac{1}{2}\operatorname{Tr}\Big(\Big(S_1^{-1}m_1 m_1^T S_1^{-1} - S_2^{-1}m_2 m_2^T S_2^{-1}\Big)^T \Big(S_2 - S_1\Big)\Big) + \\ + \Big(S_1^{-1}m_1 - S_2^{-1}m_2\Big)^T \Big(m_1 - m_2\Big) = \frac{1}{2}(x_1 - x_2)^T (S_1 + S_2)(x_1 - x_2) > 0$$

where $x_i = S_i^{-1}m_i$ and last inequality holds due to S_i is positive-definite for any $i \in \{1, 2\}$

Let's prove the strong convexity of the function $f(S) = -\frac{1}{2}\log|S|$, using the criterion:

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \beta\|x - y\|_2^2$$

Here, as $\nabla f(S) = -\frac{1}{2}S^{-1}$, then

$$\frac{1}{2}\operatorname{Tr}((S_1^{-1} - S_2^{-1})^T(S_2 - S_1)) = \frac{1}{2}\operatorname{Tr}(S_1^{-1}(S_2 - S_1)^T S_2^{-1}(S_2 - S_1)) \geq$$
$$\geq \frac{\lambda_{min}^2}{2}\operatorname{Tr}((S_2 - S_1)^T(S_2 - S_1)).$$

In the latter expression we used the fact $\lambda(S_i^{-1}) = \lambda(\Sigma_i) \in [\lambda_{min}, \lambda_{max}]$.

Thus, we have found $\beta = \frac{\lambda_{min}^2}{2}$. Finally, the coefficient of strong convexity of the initial function with respect to the variable S is $\alpha = \frac{\lambda_{min}^2}{2}\int \varphi(x)f(x)\,dx$.

The result was obtained while considering the matrix S in a vectorized way in the context of Euclidean norm or Frobenius matrix norm. Taking into account norm equivalency relations, in particular,

$$\begin{aligned} \|x\|_2 &\leq \|x\|_1 \leq \sqrt{n}\,\|x\|_2 \\ \|x\|_\infty &\leq \|x\|_2 \leq \sqrt{n}\,\|x\|_\infty \\ \|x\|_\infty &\leq \|x\|_1 \leq n\,\|x\|_\infty\,, \end{aligned}$$

we can receive the coefficients of strong convexity in different norms. Indeed, using the expression written above:

$$\frac{1}{2}\operatorname{Tr}((S_1^{-1} - S_2^{-1})^T(S_2 - S_1)) \geq \frac{\lambda_{min}^2}{2}\|S_2 - S_1\|_F^2 = \frac{\lambda_{min}^2}{2}\|\operatorname{vec}(S_2 - S_1)\|_2^2 \geq$$
$$\geq \frac{\lambda_{min}^2}{2}\|\operatorname{vec}(S_2 - S_1)\|_\infty^2;$$

$$\frac{1}{2}\operatorname{Tr}((S_1^{-1} - S_2^{-1})^T(S_2 - S_1)) \geq \frac{\lambda_{min}^2}{2}\|S_2 - S_1\|_F^2 = \frac{\lambda_{min}^2}{2}\|\operatorname{vec}(S_2 - S_1)\|_2^2 \geq$$
$$\geq \frac{\lambda_{min}^2}{2\sqrt{n}}\|\operatorname{vec}(S_2 - S_1)\|_1^2;$$

$$\frac{1}{2}\operatorname{Tr}((S_1^{-1} - S_2^{-1})^T(S_2 - S_1)) \geq \frac{\lambda_{min}^2}{2}\|S_2 - S_1\|_F^2 = \frac{\lambda_{min}^2}{2}\sum_{i=1}^{n}\lambda_i^2(S_2 - S_1) \geq$$
$$\geq \frac{\lambda_{min}^2}{2}\max_i \lambda_i^2(S_2 - S_1) = \frac{\lambda_{min}^2}{2}\|S_2 - S_1\|_2^2;$$

$$\frac{1}{2}\operatorname{Tr}((S_1^{-1} - S_2^{-1})^T(S_2 - S_1)) \geq \frac{\lambda_{min}^2}{2}\|S_2 - S_1\|_F^2 = \frac{\lambda_{min}^2}{2}\sum_{i=1}^{n}\lambda_i^2(S_2 - S_1) \geq$$
$$\geq \frac{\lambda_{min}^2}{2\sqrt{n}}\left(\sum_{i=1}^{n}|\lambda_i(S_2 - S_1)|\right)^2 = \frac{\lambda_{min}^2}{2\sqrt{n}}\|S_2 - S_1\|_*^2;$$

Thus, we also have a strong convexity with respect to the vectorized variable S in the norm $\|\cdot\|_\infty$ with the coefficient $\alpha = \frac{\lambda_{min}^2}{2}\int \varphi(x)f(x)\,dx$; in the norm $\|\cdot\|_1$ with the coefficient $\alpha =$

$\frac{\lambda_{min}^2}{2\sqrt{n}} \int \varphi(x) f(x)\, dx$; moreover, there is a strong convexity with respect to the matrix variable S in the spectral norm with the coefficient $\alpha = \frac{\lambda_{min}^2}{2} \int \varphi(x) f(x)\, dx$; in the nuclear norm with the coefficient $\alpha = \frac{\lambda_{min}^2}{2\sqrt{n}} \int \varphi(x) f(x)\, dx$.

Alternatively, one can achieve the similar estimations by exploiting some features of dual functions and the results of the Lemma of Juditsky and Nemirovski in [3], as it is described in [4]. It is also worth mentioning the lemma itself here.

Lemma 1 (Juditsky and Nemirovski) *Let δ be an open interval. Suppose $\phi : \delta \to \mathbb{R}_*$ is a twice differentiable convex function such that ϕ'' is monotonically non-decreasing. Let $\mathbb{S}_n(\delta)$ be the set of all symmetric $n \times n$ matrices with eigenvalues in δ. Define the function $F : \mathbb{S}_n(\delta) \to \mathbb{R}_*$*

$$F(X) = \sum_{i=1}^{n} \phi(\lambda_i(X))$$

and let

$$f(t) = F(X + tH)$$

for some $X \in \mathbb{S}_n(\delta)$, $H \in \mathbb{S}_n$. Then, we have,

$$f''(0) \leq 2 \sum_{i=1}^{n} \phi''(\lambda_i(X)) \lambda_i^2(H)$$

Moreover, the following property of the dual functions is going to be used.

Theorem 2 *$f(S)$ is β-strongly convex w.r.t. a norm $\|\cdot\|$ if and only if f^* is $\frac{1}{\beta}$-strongly smooth w.r.t. the dual norm $\|\cdot\|_*$*

Here the strong smoothness with the coefficient β of some function means that for a differentiable function $f : X \to \mathbb{R}$ and for any x, y the following expression is true

$$f(x + y) \leq f(x) + \langle \nabla f(x), y \rangle + \frac{1}{2} \beta \|y\|^2$$

Considering $f(S) = -\frac{1}{2} \log |S|$ again, one can easily calculate its dual function, which is $f^*(S) = -2n + \frac{1}{2} \ln 2 - \frac{1}{2} \sum_{i=1}^{n} \ln |\lambda_i(S)|$, where S is a negative-definite matrix with eigenvalues lying in the interval $\left[-\frac{\lambda_{max}}{2}, -\frac{\lambda_{min}}{2}\right]$, i.e. has only negative eigenvalues. λ_{max} and λ_{min} are respectively maximal and minimal eigenvalues of the covariance matrix of the considered normal distribution.

Let's use Lemma 1 mentioned above and consider $\tilde{f}(t) = \frac{1}{2} \ln 2 - 2n - \frac{1}{2} \sum_{i=1}^{n} \ln |\lambda_i(X + tH)|$ for some fixed X, H. Applying lemma on it:

$$\tilde{f}''(0) \leq 2 \sum_{i=1}^{n} \frac{\lambda_i^2(H)}{2\lambda_i^2(X)} \leq \frac{4}{\lambda_{min}^2} \sum_{i=1}^{n} \lambda_i^2(H) = \frac{4}{\lambda_{min}^2} \|H\|_F^2$$

It means that f(S) is $\frac{\lambda_{min}^2}{4}-$strong convex, therefore $\alpha = \frac{\lambda_{min}^2}{4} \int \varphi(x) f(x)\, dx$ for the objective function. The coefficient isn't noticeably different from the previous one, i.e. the estimation wasn't improved.

Minimization of estimator's variance

Now it is time to pay attention to the second problem. In the case of normal distributions the task transforms to the following

$$\min_{m,S} \int \varphi^2(x) f^2(x) \sqrt{(2\pi)^n} \exp\left[\frac{1}{2}x^T Sx - m^T x + \frac{1}{2}m^T S^{-1} m - \frac{1}{2}\log|S|\right] dx.$$

It is worth exploring separately the function $\frac{1}{f_\theta(x)}$ as the parameters are entirely represented in it:

$$\frac{1}{f_\theta(x)} = \sqrt{(2\pi)^n} \exp\left(\frac{1}{2}\operatorname{Tr}(Sxx^T) - m^T x + \frac{1}{2}\operatorname{Tr}(S^{-1}mm^T) - \frac{1}{2}\log|S|\right).$$

It is an exponential function and there is a fact that if some arbitrary function $f(x)$ is strongly convex with a coefficient β in some norm $\|\cdot\|$, then $e^{f(x)}$ is a strongly convex function on some compact, i.e. let $\operatorname{dom} f = X$ and if $f(X) \subseteq [-C, +\infty]$ for some $C > 0$, then the coefficient of strong convexity is equal to βe^{-C}. Indeed, as we deal with the differentiable functions, then

$$\langle \exp f(x)\nabla f(x) - \exp f(y)\nabla f(y), x - y\rangle \geq e^{-C}\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq e^{-C}\beta\|x - y\|^2$$

where the last inequality coincides with the definition of strong convexity for the differentiable functions.

In our case as $\lambda(\Sigma) \in [\lambda_{min}, \lambda_{max}]$, then

$$\frac{1}{2}x^T Sx - m^T x + \frac{1}{2}m^T S^{-1}m - \frac{1}{2}\log|S| \geq -\frac{1}{2}\log|S| \geq \frac{n}{2}\log\lambda_{min}$$

Therefore, $e^{-C} = \lambda_{min}^{\frac{n}{2}}$. From what was found earlier it is obvious that $\beta = \frac{\lambda_{min}^2}{2}$. Finally, the coefficient of strong convexity with respact to the variable S of the function $\frac{1}{f_\theta(x)}$ is $\alpha = \beta e^{-C} = \frac{\lambda_{min}^{\frac{n}{2}+2}}{2}$. In the similar way as we did for the first optimization problem, we can get the coefficients in other norms. We have a strong convexity with respect to the vectorized variable S in the norm $\|\cdot\|_\infty$ with the coefficient $\alpha = \frac{\lambda_{min}^{\frac{n}{2}+2}}{2}\int \varphi(x)f(x)\,dx$; in the norm $\|\cdot\|_1$ with the coefficient $\alpha = \frac{\lambda_{min}^{\frac{n}{2}+2}}{2\sqrt{n}}\int \varphi(x)f(x)\,dx$; in addition, one can spot a strong convexity with respect to the matrix variable S in the spectral norm with the coefficient $\alpha = \frac{\lambda_{min}^{\frac{n}{2}+2}}{2}\int \varphi(x)f(x)\,dx$; in the nuclear norm with the coefficient $\alpha = \frac{\lambda_{min}^{\frac{n}{2}+2}}{2\sqrt{n}}\int \varphi(x)f(x)\,dx$.

Let's return to our initial objective function $V(\theta)$, where $\theta = (m, S)$. If we have a proper $\alpha(x)$ (now we will take $\alpha(x) = \alpha = \frac{\lambda_{min}^{\frac{n}{2}+2}}{2}$) and a set Θ, then for all $\theta_1, \theta_2 \in \Theta$ we will have:

$$\begin{aligned}
&\int \exp(A(t(\theta_1) + (1-t)\theta_2) - (t\theta_1 + (1-t)\theta_2)^T T(x)) f^2(x)\phi^2(x)h(x)\,dx \leq \\
\leq t &\int \exp(A(\theta_1) - \theta_1^T T(x)) f^2(x)\phi^2(x)h(x)\,dx + \\
+(1-t) &\int \exp(A(\theta_2) - \theta_2^T T(x)) f^2(x)\phi^2(x)h(x)\,dx - \\
-t(1-t)&\frac{\lambda_{min}^{\frac{n}{2}+2}}{2\sqrt{(2\pi)^n}} \int f^2(x)\phi^2(x)\,dx\|\theta_1 - \theta_2\|_2^2
\end{aligned}$$

From this expression above it became clear that the coefficient of strong convexity for the objective function is equal to $\frac{\lambda_{min}^{\frac{n}{2}+2}}{2\sqrt{(2\pi)^n}} \int f^2(x)\phi^2(x)\,dx$, that means that the coefficient depends on the value of $\int\limits_{\mathbb{R}^n} f^2(x)\phi^2(x)\,dx$. Although, it is worth mentioning that it is possible to take $\alpha(x) = \frac{\lambda_{min}^{\frac{n}{2}+2}}{2}\mathbb{I}(x \in X)$ for some X and then it is enough to know the value $\int\limits_{X} f^2(x)\phi^2(x)\,dx$.

References

[1] D. L. McLeish, “Bounded relative error importance sampling and rare event simulation,” *Department of Statistics, U. Waterloo and ETH Zürich*, 2008.

[2] E. K. Ryu and S. P. Boyd, “Adaptive importance sampling via stochastic convex programming,” *Institute for Computational and Mathematical Engineering, Stanford University*, 2015.

[3] A. Juditsky and A. Nemirovski., “Large deviations of vector-valued martingales in 2-smooth normed spaces.,” *Annals of Probability*, 2008.

[4] S. M. Kakade, S. Shalev-Shwartz, and A. Tewari, “On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization,” *Toyota Technological Institute—Chicago, USA*, 2009.