Importance Sampling.
The research of optimization problems

Alena Shilova

August 2017

Strong convexity

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 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

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

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 $\mathcal{F}$.

In order to find the optimal distribution for the importance sampling estimator, in the beginning, we will try to minimize the variance of the estimator:

$$\min_{\theta\in\Theta} V(\theta) = \min_{\theta\in\Theta} \int \frac{\varphi^2(x) f^2(x)}{f_\theta(x)}\, dx - I^2.$$

As it was proved in [1], the problem will be convex if the exponential family of distributions is considered. Now it is interesting to know under which circumstances the function to minimize will be strongly convex and with which constant.

Let's narrow our problem 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 optimisation problem

Having our problem formulated, let's find out whether the objective function is strongly convex.

To start with, let's have a closer look on a function

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

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 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, $e^{-C} = \lambda_{min}^{\frac{n}{2}}$. Now we need to find β.

It is quite obvious that $\frac{1}{2}x^T Sx - m^T x$ 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

$$\frac{1}{2}\operatorname{Tr}\Big(\Big(S_1^{-1}m_1m_1^T S_1^{-1} - S_2^{-1}m_2m_2^T S_2^{-1}\Big)^T\Big(S_2 - S_1\Big)\Big) +$$
$$+\Big(S_1^{-1}m_1 - S_2^{-1}m_2\Big)^T\Big(m_1 - m_2\Big) = \frac{1}{2}(x_1 - x_2)^T(S_1 + S_2)(x_1 - x_2) > 0$$

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

Thus, we have found $\beta = \frac{\lambda_{min}^2}{2}$ as well. Finally, the coefficient of strong convexity of the function $\frac{1}{f_\theta(x)}$ is $\alpha = \beta e^{-C} = \frac{\lambda_{min}^{\frac{n}{2}+2}}{2}$.

Alternatively, one can achieve the similar estimation by exploiting some features of dual functions and the results of the Lemma of Juditsky and Nemirovski in [2], as it is described in [3]. 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 $\alpha = \frac{\lambda_{min}^{\frac{n}{2}+2}}{4}$ for the objective function. The coefficient isn't noticeably different from the previous one, i.e. the estimation wasn't improved.

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:

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

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 $\int\limits_X f^2(x)\phi^2(x)\,dx$.

Rényi divergence

References

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

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

[3] 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.