Importance Sampling.
Comparison of two methods and its realisation

Alena Shilova

12 February 2018

1 First optimization task

First we are going to use an estimate, which is based on mixture importance sampling without control variates.

$$\hat{\mu}_\alpha = \frac{1}{n}\sum_{i=1}^{n} \frac{p(x_i)H_{1:J}(x_i)}{\sum_{j=0}^{J}\alpha_j q_j(x_i)} = \frac{1}{n}\sum_{i=1}^{n}\frac{H_{1:J}(x_i)}{\sum_{j=0}^{J}\alpha_j H_j(x_i)P_j^{-1}}$$

This estimate is unbiased.

$$\mathbb{E}\hat{\mu}_\alpha = \mu$$

The corresponding optimization task is related to minimization of variance of the mentioned above estimation.

$$f(\alpha) = n\,\mathrm{Var}\,\hat{\mu}_\alpha = \left(\int \frac{H_{1:J}(x)p^2(x)}{\sum_{j=0}^{J}\alpha_j q_j(x)}\,dx - \mu^2\right) = \left(\int \frac{H_{1:J}(x)p(x)}{\sum_{j=0}^{J}\alpha_j H_j(x)P_j^{-1}}\,dx - \mu^2\right) \longrightarrow \min_{\alpha\in S} \quad (1)$$

Here S denotes simplex and $x_i \sim q_\alpha$.

1.1 Owen's realization

In the article https://arxiv.org/pdf/1411.3954.pdf it was suggested to replace the integral with its estimation. In this case, given $x_i \sim q_{\alpha'}$, the appropriate replacement is

$$g(\alpha) = \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{H_{1:J}(x_i)p^2(x_i)}{q_\alpha(x_i)q_{\alpha'}(x_i)} = \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{H_{1:J}(x_i)}{(\sum_{j=0}^{J}\alpha_j H_j(x_i)P_j^{-1})(\sum_{j=0}^{J}\alpha'_j H_j(x_i)P_j^{-1})} \longrightarrow \min_{\alpha\in S}.$$

Here α should satisfy several constraints, among which there are $\alpha_j \geq \epsilon$ for all j and $\sum_{j=0}^{J}\alpha_j < 1+\eta$ for $\eta \geq 0$. In the article the authors used damped Newton method for the unconstrained optimization where the objective consists of the initial objective function and log-barrier functions, formed by the constraints. So our task is

$$g(\alpha) + \rho\sum_{j=0}^{J}\log(\alpha_j - \epsilon) + \rho\log\left(1+\eta-\sum_{j=0}^{J}\alpha_j\right) \longrightarrow \min_{\alpha}$$

Having solved this problem, we get $\alpha(\rho)$. Then ρ should be decreased

$$\rho := \frac{\rho}{\kappa}$$

with $\kappa > 1$ and the problem above should be solved again providing us with a new $\alpha(\rho)$ and this iteration has to be repeated until $(J+1)\rho < \epsilon_1 g(\alpha)$ with ϵ_1 as tolerance rate.

1.2 Stochastic gradient descent

I've also tried the method proposed in the Boyd's article https://arxiv.org/pdf/1412.4845.pdf, that is to use stochastic gradient descent to minimize per-sample variance sampling every time from a new sample distribution depending on the updated parameter. The stochastic gradient has the following look

$$[\nabla f(\alpha)]_k = \mathbb{E}\left[-\frac{H_{1:J}(x_i)p^2(x_i)q_k(x_i)}{q_\alpha^3(x_i)}\right] = -\mathbb{E}\left[\frac{H_{1:J}(x_i)H_k(x_i)P_k^{-1}}{\left(\sum_{j=0}^J \alpha_j H_j(x_i)P_j^{-1}\right)^3}\right]$$

2 Second optimization task

Now, let's consider the second type of problems. Here we use control variates in our estimate. Let's suppose that we know some function $h(x)$ such that $\mathbb{E}h(x) = \int h(x)p(x)\,dx = \theta$, and θ is known as well. In this paper we are going to consider only $h(x) \subset \mathbb{R}^J$, which can be represented as

$$h_j(x) = \begin{cases} \frac{q_j(x)}{p(x)} & p(x) > 0 \\ 0 & \text{otherwise} \end{cases}$$

which implies $\theta = (1, \dots, 1)$. Furthermore, it is required to introduce $\beta, \gamma \in \mathbb{R}^J$. The second estimation which we are going to use is

$$\hat{\mu}_{\alpha,\beta,\gamma} = \frac{1}{n}\sum_{i=1}^n \frac{H_{1:J}(x_i) - \sum_{j=1}^J \beta_j H_j(x_i)P_j^{-1} + \sum_{j=1}^J \gamma_j}{\sum_{j=0}^J \alpha_j H_j(x_i)P_j^{-1}}.$$

Bias isn't equal to zero:

$$\mathbb{E}\hat{\mu}_{\alpha,\beta,\gamma} = \mathbb{E}\tilde{\mu}_{\alpha,\beta,\gamma} = \mu + \sum_{j=1}^J (\gamma_j - \beta_j).$$

The corresponding variance can be written as

$$\begin{aligned} &f(\alpha,\beta,\gamma) = n \operatorname{Var}\hat{\mu}_{\alpha,\beta,\gamma} = \\ &= \int\limits_{D_{q_\alpha}} \frac{\left(H_{1:J}(x) - \sum_{j=1}^J (\beta_j H_j(x)P_j^{-1} - \gamma_j)\right)^2 p(x)}{\sum_{j=0}^J \alpha_j H_j(x)P_j^{-1}}\,dx - \left(\mu + \sum_{j=1}^J (\gamma_j - \beta_j)\right)^2 \longrightarrow \min_{\alpha,\beta,\gamma}. \end{aligned} \tag{2}$$

Taking into account the fact that the estimation is biased, it isn't enough to minimize only the variance of the estimation, that's why

$$\operatorname{Var}\hat{\mu}_{\alpha,\beta,\gamma} + bias^2(\hat{\mu}_{\alpha,\beta,\gamma}) \longrightarrow \min_{\alpha,\beta,\gamma} \tag{3}$$

is more relevant optimization task in this case.

2.1 Owen's realization

As we did for (1), we can also rewrite (2) in the similar way, having $x_i \sim q_{\alpha'}$, but first let's denote $X_{ij} = H_j(x_i)P_j^{-1}$, $y_i = H_{1:J}(x_i) = y_i$ and $z_i = \sum_{j=0}^J \alpha'_j H_j(x_i)P_j^{-1}$, then

$$\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{(y_i - \sum_{j=1}^J X_{ij}\beta_j + \sum_{j=1}^J \gamma_j)^2}{z_i(\sum_{j=0}^J X_{ij}\alpha_j)} - 2\mu\sum_{j=1}^J(\gamma_j - \beta_j) - \mu^2 - \left(\sum_{j=1}^J(\gamma_j - \beta_j)\right)^2 \simeq$$

$$\simeq g(\alpha,\beta,\gamma) - \frac{2}{n_1}\sum_{i=1}^{n_1}\frac{y_i}{z_i}\sum_{j=1}^J(\gamma_j - \beta_j) - \mu^2 - \left(\sum_{j=1}^J(\gamma_j - \beta_j)\right)^2$$

Here, we used instead of μ its approximation $\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{y_i}{z_i}$. Therefore, (3) will be equivalent to

$$g(\alpha,\beta,\gamma) - \frac{2}{n_1}\sum_{i=1}^{n_1}\frac{y_i}{z_i}\left(\sum_{j=1}^J(\gamma_j - \beta_j)\right) + (n-1)\left(\sum_{j=1}^J(\gamma_j - \beta_j)\right)^2 \longrightarrow \min_{\alpha,\beta,\gamma}$$

After adding the constraints in the form of the log-barrier function we solve the problem as it was discussed earlier in the text with the help of damped Newton method.

2.2 Stochastic gradient descent

Like in the previous case, for the optimization we only need to calculate stochastic gradient. However, in this situation we minimize per sample variance plus squared bias and there are three times more parameters to optimize. So

$$[\nabla_\alpha f(\alpha,\beta,\gamma)]_k = -\mathbb{E}\left[\frac{\left(y_i - \sum_{j=1}^J(\beta_j X_{ij} - \gamma_j)\right)^2}{\left(\sum_{j=0}^J \alpha_j X_{ij}\right)^3} X_{ik}\right],$$

$$\left[\nabla_\beta\left\{f(\alpha,\beta,\gamma) + b^2(\beta,\gamma)\right\}\right]_k = -2\mathbb{E}\left[\frac{y_i - \sum_{j=1}^J(\beta_j X_{ij} - \gamma_j)}{\left(\sum_{j=0}^J \alpha_j X_{ij}\right)^2} X_{ik} - \frac{y_i}{\sum_{j=0}^J \alpha_j X_{ij}}\right],$$

$$\left[\nabla_\gamma\left\{f(\alpha,\beta,\gamma) + b^2(\beta,\gamma)\right\}\right]_k = 2\mathbb{E}\left[\frac{y_i - \sum_{j=1}^J(\beta_j X_{ij} - \gamma_j)}{\left(\sum_{j=0}^J \alpha_j X_{ij}\right)^2} - \frac{y_i}{\sum_{j=0}^J \alpha_j X_{ij}}\right].$$

Here X_{ij}, y_i are constructed from $x_i \sim q_\alpha$.

3 Results for simple simulations

I checked the work of all the mentioned methods on two examples where it was required to calculate the probability to get into the area located outside the allowed zone. This zone can be described with an intersection of H_j^c where $H_j = \left\{x \in \mathbb{R}^2 \;\middle|\; w_j^T x \geq \tau\right\}$. In the first example the half-spaces can be defined by $w_j^T = (\sin(2\pi j/J), \cos(2\pi j/J))$, for $j = 1, \ldots, J$ and a single τ, thus $H_j^c = \mathcal{P}(J, \tau)$. The second example differs from the first one by fixating $J = 360$ and $w_j^T = (\sin(2\pi p(j)/360), \cos(2\pi p(j)/360))$ where $p(j)$ is the j'th prime among integers up to 360.

For damped Newton method I took

$$\alpha_j^{(0)} = \epsilon + \frac{1 + \eta - J\epsilon}{J + 1}$$

for all $j \in \overline{0, J}$ and $\beta_j^{(0)} = \gamma_j^{(0)} = 0$ for all $j \in \overline{1, J}$. For stochastic gradient descent I tried $\alpha_j^{(0)} = 1/(J+1)$ and also $\alpha_0^{(0)} = 1 - \epsilon J$ and $\alpha_j^{(0)} = 1 - \epsilon J$ for all $j \neq 0$. One can notice that the optimal alpha should be equal to $\alpha_j^{(0)} = 1/(J+1)\ \forall j$ as $P(w_j^T x \geq \tau) = \Phi(-\tau)$ which is the same for all half-spaces. Hence, if we take such α which has all components being equal, then the methods with an estimation without control variates don't update the parameter at all. The same picture can be seen where we used damped Newton method with control variates.

When it was checked the work of stochastic gradient descent to optimize (1) with zero alpha containing one component different from others, very small changes to an initial α appeared. It can be explained by the fact that the value of the gradient is very small and the updated α after being projected onto simplex almost doesn't change compared with the previous parameter value.

Considering the task with control variates, there is a problem, when trying to implement simple stochastic gradient descent to the task. Starting from the zero values, β and γ tend to become infinitely large after a few iterations. In order to make the method work with the task, SGD with Nesterov momentum was implemented and applied to the task. Still, this method also diverge and can't find optimal parameters. Hence, another method is required for this type of the optimization problem.

4 Results for cases

Moreover, the method were applied to the cases gathered in MATPOWER 6.0. For now, only the case Winter peak 2383 was analysed. The constraints defining the critical zone outside some polytop were formed in the similar way as in https://arxiv.org/pdf/1710.06965.pdf.

It is worth mentioning that if μ is too small (rare events), then the methods won't update the initial parameters due to the fact that descent direction derived from the gradient (or stochastic gradient) can contain only very small components to affect the parameters. So, further results are given for $\mu \in [0.001, 0.015]$.

For the case introduced above it is easy to find $\underline{\mu} = \max_j P_j = 0.001284$ and $\overline{\mu} = \sum_{j=1}^J P_j = 0.010400$. ALORE after 10^5 iterations returned $\hat{\mu} = 0.010362$. Newton method for the estimation with and without control variates provided with $\hat{\mu}_\alpha = 0.010353$ and $\hat{\mu}_{\alpha,\beta,\gamma} = 0.010350$ respectively, which is very close to the estimation received after ALORE.

SGD without control variates gives us $\mu_\alpha = 0.010162$. This noticeable difference probably appeared because the parameter α almost didn't change throughout all 10^5 iterations. Compared to initial α, final α has one component which dominates the others. After the closer look one can see this component corresponds to one of the most probable constraints.

Analogously, SGD with control variates still can't converge for this case, too.