Importance Sampling for Reducing the Variance of SGD

Jingfeng Wu

Created: November 2019
Last updated: June 27, 2020

This is a brief note on importance sampling [1, 2] and its application on reducing SGD variance.

1 Importance sampling

Let i be the index random variable sampled from $[n] = \{1, 2, \dots, n\}$, $i \sim p(i)$, $\sum_{i=1}^{n} p(i) = 1,\ p(i) \in [0, 1]$. Let $X(i)$ be a sequence of random variables drawn with respect to index i.

In the context of importance sampling, we suppose different indexes have different "importance", and we would like to see important examples more frequently. Concisely, given an sequence of importance variable $w(i) \in (0, +\infty)$. For technical reason, assume $\mathbb{E}[w(i)] = \sum_{i=1}^{k} p(i)w(i) = 1$. Then we can re-weight index random variable as $i \sim p^w(i) = w(i)p(i)$. Note that $\sum_{i=1}^{n} w(i)p(i) = 1$, thus $p^w(i) = w(i)p(i)$ is a well defined distribution. For simplicity, we use $\mathbb{E}$ to represent taking expectation with respect to $i \sim p(i)$, and $\mathbb{E}^w$ to denote taking expectation with respect to $i \sim p^w(i)$.

Now consider a random variable $X(i)$. If we re-weight the distribution of index i, we will end up with a biased estimator of $X(i)$ in general. This is because generally

$$\mathbb{E}^w[X(i)] = \sum_{i=1}^{n} w(i)p(i)X(i) \neq \sum_{i=1}^{n} p(i)X(i) = \mathbb{E}[X(i)]. \tag{1}$$

Not too surprisingly, we can correct this issue by re-weighting $X(i)$.

Lemma 1. *If $X(i)$ is re-weighted as $\frac{1}{w(i)}X(i)$, then importance sampling generate an unbiased estimator, i.e.,*

$$\mathbb{E}^w \frac{1}{w(i)} X(i) = \mathbb{E}X(i). \tag{2}$$

Proof. By definition. □

The intuition of Lemma 1 is that we can estimate a random variable by seeing important examples more often, but use them less.

2 Variance reduction

Importance sampling keeps the first order moment. However it leads to different high order moments. For example, we now explore how to reduce variance of SGD by importance sampling.

Let $g(i)$ be the gradient of loss with respect to the i-th example. By lemma 1 we know that applying importance sampling would not change the expectation of SGD, i.e., $\mathbb{E}^w[\frac{g(i)}{w(i)}] = \mathbb{E}[g(i)]$. We turn to calculate its variance.

$$\operatorname{Tr}\operatorname{Var}^w[g(i)] = \mathbb{E}^w\left[\frac{g(i)^T g(i)}{w(i)^2}\right] - (\mathbb{E}^w[g(i)])^T(\mathbb{E}^w[g(i)]) = \mathbb{E}^w\left[\frac{g(i)^T g(i)}{w(i)^2}\right] - (\mathbb{E}[g(i)])^T(\mathbb{E}[g(i)]). \tag{3}$$

By Jesen's inequality, we have

$$\begin{aligned} \mathbb{E}^w\left[\frac{g(i)^T g(i)}{w(i)^2}\right] &= \sum_{i=1}^{n} w(i)p(i)\frac{g(i)^T g(i)}{w(i)^2} \\ &\geq \left(\sum_{i=1}^{n} w(i)p(i)\frac{\|g(i)\|_2}{w(i)}\right)^2 = \left(\sum_{i=1}^{n} p(i)\,\|g(i)\|_2\right)^2 = (\mathbb{E}[\|g(i)\|_2])^2\,. \end{aligned} \tag{4}$$

The inequality becomes equality if and only if $w(i) \propto \|g(i)\|_2$, i.e.

$$w(i) = \frac{\|g(i)\|_2}{\sum_{i=1}^{n} p(i) \|g(i)\|_2} = \frac{\|g(i)\|_2}{\mathbb{E}[\|g(i)\|_2]}. \tag{5}$$

When this “optimal” importance sampling happens, we reduce the variance of vanilla SGD in the scale of

$$\mathbb{E}[g(i)^T g(i)] - (\mathbb{E}[\|g(i)\|_2])^2 \geq 0. \tag{6}$$

For an example, if $p(i) = \frac{1}{n}$, and the norm of gradients vary quite a lot, then importance sampling indeed reduces the variance of SGD significantly.

However, one should notice that even with importance sampling, the variance of SGD is at least

$$(\mathbb{E}[\|g(i)\|_2])^2 - (\mathbb{E}[g(i)])^T (\mathbb{E}[g(i)]) \geq 0. \tag{7}$$

In the end, the optimal importance sampling for reducing SGD variance requires the information of the gradient norm of all samples, which is often impractical. There are some nice solutions for this problem, e.g., using LSH [2].

References

[1] Guillaume Alain, Alex Lamb, Chinnadhurai Sankar, Aaron Courville, and Yoshua Bengio. Variance reduction in sgd by distributed importance sampling. *arXiv preprint arXiv:1511.06481*, 2015.

[2] Beidi Chen, Yingchen Xu, and Anshumali Shrivastava. Fast and accurate stochastic gradient estimation. In *Advances in Neural Information Processing Systems*, pages 12339–12349, 2019.