Low-Cost Lipschitz-Independent Adaptive Importance Sampling of Stochastic Gradients

Xiaolu Wang

Presentation for ICPR 2020

Joint Work with: Huikang Liu, Jiajin Li, Anthony Man-Cho So

Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong

December 17, 2020
Consider the following empirical risk minimization (ERM) problem

\[
\min_{\mathbf{w} \in \mathbb{R}^d} \left\{ F(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{w}) \right\}
\]  

- It finds many applications in machine learning and pattern recognition;
- \( f_i(\mathbf{w}) = \ell(h(x_i; \mathbf{w}), y_i) \) where \((x_i, y_i)\) is the \(i\)-th training example;
  - \( h \) is the decision function parameterized by \( \mathbf{w} \);
  - \( \ell \) is the loss function.
- SGD plays a central role in solving optimization problem (1):
  \[
  \mathbf{w}_{k+1} = \mathbf{w}_k - \eta_k \nabla f_i(\mathbf{w}_k),
  \]
Adaptive Importance Sampling

To control the variance of the stochastic gradient, SGD with adaptive importance sampling is introduced:

$$w_{k+1} = w_k - \frac{\eta_k}{np_k} \nabla f_i(w_k), \quad (3)$$

where $p^k := (p_1^k, p_2^k, \ldots, p_n^k)^\top$ is the importance sampling distribution. A natural idea of choosing distribution $p^k$ is to minimize the variance

$$\min_{p^k} \quad \text{Var} \left[ \frac{1}{np^k_i} \nabla f_i(w_k) \right] \quad \text{s.t.} \quad \sum_{j=1}^n p_j^k = 1, \ p_i^k \geq 0, \ \forall i \in \{1, \ldots, n\}. \quad (4)$$

Problem (4) has a closed-form optimal solution, which is

$$\left( p_i^k \right)^* = \frac{\| \nabla f_i(w_k) \|^2_2}{\sum_{j=1}^n \| \nabla f_j(w_k) \|^2_2}, \ \forall i \in \{1, \ldots, n\}. \quad (5)$$

**Question:** How to compute the optimal sampling distribution?

**Key Idea:** Using the most recently evaluated gradient norms $\| \nabla f_i(w_k') \|^2_2$ to approximate $\| \nabla f_i(w_k) \|^2_2$. 
SGD-AIS Algorithm

Algorithm 1 SGD-AIS

1: **Input:** step sizes \( \{ \eta_k \} \), weights \( \alpha_k \in [\alpha, \overline{\alpha}] \subset (0, 1) \) for all \( k \in \mathbb{N} \).
2: **Initialize:** \( w_1, \pi_i = 1 \) for all \( i \in \{1, \ldots, n\} \)
3: **for** \( k = 1, 2 \ldots \) **do**
4: \( \text{Update the sampling probabilities for all } i \in \{1, \ldots, n\} \)
   \[ p_i = \alpha_k \frac{\pi_i}{\sum_{j=1}^{n} \pi_j} + (1 - \alpha_k) \frac{1}{n} \quad (6) \]
5: \( \text{Randomly pick } i_k \in [n] \text{ based on distribution } p \)
6: \( \text{Compute stochastic gradient } g_k = \frac{1}{n \pi_{i_k}} \nabla f_{i_k}(w_k) \)
7: \( \text{Set } \pi_{i_k} = \| \nabla f_{i_k}(w_k) \|_2 \)
8: \( \text{Set } w_{k+1} = w_k - \eta_k g_k \)
9: **end for**

**Complexity:** By resorting to a binary tree data structure, only additional \( O(\log n) \) per-iteration cost is needed to implement the adaptive sampling.
Applying AIS strategy to SGD with momentum, we just need $g_k$ to be

$$g_k = \theta g_{k-1} + (1 - \theta) \frac{1}{np_{i_k}} \nabla f_i(w_k).$$

(7)

Applying AIS strategy to ADAM, we just need $g_k$ to be

$$g_k = \frac{\hat{m}_k}{\sqrt{\hat{h}_k} + \varepsilon},$$

(8)

where

$$\hat{m}_k = \left( \theta_1 m_{k-1} + (1 - \theta_1) \frac{1}{np_{i_k}} g_k \right) \bigg/ (1 - \theta_1^k),$$

(9)

and

$$\hat{h}_k = \left( \theta_2 h_{k-1} + (1 - \theta_2) \frac{1}{np_{i_k}} g_k^2 \right) \bigg/ (1 - \theta_2^k).$$

(10)
## Theoretical Results

### Theorem

Under some mild assumptions, the sequence \( \{w_k\} \) generated by SGD-AIS with a fixed stepsize \( \eta_k = \eta \) for all \( k \in \mathbb{N} \) satisfying

\[
\mathbb{E}[F(w_k) - F^*] \leq \frac{\eta L (1 - \gamma) G^2}{4\sigma} + (1 - 2\eta\sigma)^{k-1} (F(w_1) - F^*)
\]

\[
\lim_{k \to \infty} \eta L (1 - \gamma) G^2
\]

\[
\frac{4\sigma}{4\sigma}.
\]

\[(11)\]

- Compared with vanilla SGD, the convergence bounds of SGD-AIS are improved by a factor of \( 1 - \gamma < 1 \);
- Similar improvement still holds if we choose diminishing stepsize;
- We also provide more convergence analysis under the nonconvex settings.
We implement three algorithms, which are SGD-AIS, SGD with uniform sampling (SGD-US), SGD with Lipschitz-based importance sampling (SGD-LIS) for performance comparison.
SGDm and ADAM for SVM

We conduct experiments on SVM with squared hinge loss to evaluate the performance of SGDm-AIS and ADAM-AIS.
SGDm and ADAM for Neural Networks

We further conduct simulation on MLP, CNN and LeNet-5, and use two common benchmark datasets, namely MNIST and CIFAR-10.

Figure: Column 1-3: MLP (MNIST), LeNet-5 (MNIST), CNN (Cifar-10); Row 1: SGD-US v.s. SGD-AIS; Row 2: ADAM-US v.s. ADAM-AIS
Thank You!