Optimization for Deep Learning

Lecture 7-1: Sampling Strategies in SGD

Kun Yuan

Peking University

Main contents in this lecture

- Finite-sum minimization
- SGD with finite samples
- Importance sampling
- Random reshuffling

Stochastic optimization

- Consider the stochastic optimization problem:

$$\min_{x \in \mathbb{R}^d} \quad f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[F(x; \xi)] \tag{1}$$

    - ξ is a random variable indicating data samples
    - $\mathcal{D}$ is the data distribution; unknown in advance
    - $F(x; \xi)$ is differentiable in terms of x

- Many applications in signal processing and machine learning

Finite-sum minimization

- In real practice, we typically have finite data samples

$$\mathcal{M} = \{\xi_1, \xi_2, \cdots, \xi_N\}$$

where N is the sample size

- Suppose in distribution $\mathcal{D}$, each data will be sampled uniformly randomly, i.e.,

$$\mathbb{P}(\xi = \xi_i) = \frac{1}{N}, \quad \forall\ i,$$

Problem (1) becomes finite-sum minimization

$$\min_{x \in \mathbb{R}^d} \quad f(x) = \mathbb{E}_{\xi \sim \mathcal{D}}[F(x;\xi)] = \frac{1}{N}\sum_{i=1}^{N} F(x;\xi_i)$$

- Finite-sum minimization is a special example of stochastic optimization

Stochastic gradient descent with finite samples

- Applying SGD to finite-sum minimization, we achieve

$$\text{Sample } \xi_k \sim \{\xi_1, \cdots, \xi_N\} \textbf{ uniformly} \text{ and } \textbf{randomly}$$
$$x_{k+1} = x_k - \gamma \nabla F(x_k; \xi_k)$$

  which is referred to as SGD with finite samples.

- If we assume the stochastic gradient is unbiased and has bounded variance, all convergence theories in the last lecture apply to SGD with finite samples.

- However, SGD with finite samples has its own structures.

Stochastic gradient noise

Let $\mathcal{F}_k = \{x_k, \xi_{k-1}, x_{k-1}, \cdots, \xi_0\}$ be the filtration containing all historical variables at and before iteration k (except for ξ_k).

Lemma 1

Suppose $f(x)$ is L-smooth. Given the filtration $\mathcal{F}_k$, we have

$$\mathbb{E}[\nabla F(x_k;\xi_k)|\mathcal{F}_k] = \nabla f(x_k)$$
$$\mathbb{E}[\|\nabla F(x_k;\xi_k) - \nabla f(x_k)\|^2|\mathcal{F}_k] \le 2L^2\|x_k - x^\star\|^2 + \sigma^2$$

where $\sigma^2 = \frac{2}{N}\sum_{i=1}^{N}\|\nabla F(x^\star;\xi_i)\|^2$.

SGD with finite sample size can have unbounded variance

Stochastic gradient noise

Proof: The unbiased property is easy to verify due to

$$\mathbb{E}[\nabla F(x_k;\xi_k)|\mathcal{F}_k] = \frac{1}{N}\sum_{i=1}^{N}\nabla F(x_k;\xi_i) = \nabla f(x_k)$$

To examine the variance, we have

$$\begin{aligned}
&\mathbb{E}[\|\nabla F(x_k;\xi_k) - \nabla f(x_k)\|^2|\mathcal{F}_k] \\
\le\ & 2\mathbb{E}[\|\nabla F(x_k;\xi_k) - \nabla F(x^\star;\xi_k) - \nabla f(x_k)\|^2|\mathcal{F}_k] + 2\mathbb{E}[\|\nabla F(x^\star;\xi_k)\|^2|\mathcal{F}_k] \\
\le\ & 2\mathbb{E}[\|\nabla F(x_k;\xi_k) - \nabla F(x^\star;\xi_k)\|^2|\mathcal{F}_k] + 2\mathbb{E}[\|\nabla F(x^\star;\xi_k)\|^2|\mathcal{F}_k] \\
\le\ & 2L^2\|x_k - x^\star\|^2 + \frac{2}{N}\sum_{i=1}^{N}\|\nabla F(x^\star;\xi_i)\|^2
\end{aligned}$$

which concludes the proof.

Convergence

Theorem 1

Suppose $F(x;\xi_i)$ is L-smooth for any x and ξ_i, and $f(x)$ is strongly-convex. If $\gamma \le \frac{\mu(L-\mu)}{2L^2(L+\mu)}$, SGD with finite sample size will converge at the following rate

$$\mathbb{E}\|x_{k+1} - x^\star\|^2 \le (1-\gamma\mu)\mathbb{E}\|x_k - x^\star\|^2 + 2\gamma^2\sigma^2$$

where $\sigma^2 = \frac{2}{N}\sum_{i=1}^{N}\|\nabla F(x^\star;\xi_i)\|^2$. Keeping iterating the recursion, it holds that

$$\mathbb{E}\|x_k - x^\star\|^2 \le (1-\gamma\mu)^k\|x_0 - x^\star\|^2 + \frac{2\gamma\sigma^2}{\mu}$$

- SGD **can** converge even if with potentially unbounded variance
- The $O(\gamma\sigma^2)$ term dominates the convergence rate

Convergence

Proof: With SGD recursions, we have

$$
\begin{aligned}
& \mathbb{E}[\|x_{k+1} - x^\star\|^2 | \mathcal{F}_k] \\
= & \ \mathbb{E}[\|x_k - x^\star - \gamma \nabla F(x_k; \xi_k)\|^2 | \mathcal{F}_k] \\
= & \ \|x_k - x^\star - \gamma \nabla f(x_k)\|^2 + \gamma^2 \mathbb{E}[\|\nabla F(x_k; \xi_k) - \nabla f(x_k)\|^2 | \mathcal{F}_k] \\
\le & \ (1 + 2\gamma^2 L^2)\|x_k - x^\star\|^2 - 2\gamma \langle x_k - x^\star, \nabla f(x_k) - \nabla f(x^\star)\rangle + \gamma^2 \|\nabla f(x_k)\|^2 + 2\gamma^2 \sigma^2 \\
\le & \ (1 + 2\gamma^2 L^2 - \frac{2\gamma\mu L}{\mu + L})\|x_k - x^\star\|^2 - (\frac{2\gamma}{\mu + L} - \gamma^2)\|\nabla f(x_k)\|^2 + 2\gamma^2 \sigma^2 \\
\le & \ (1 - \gamma\mu)\|x_k - x^\star\|^2 + 2\gamma^2 \sigma^2
\end{aligned}
$$

where the last inequality holds when

$$\gamma \le \frac{\mu(L - \mu)}{2L^2(L + \mu)}$$

Taking expectations over the filtration $\mathcal{F}_k$, we achieve the result.

SGD with importance sampling

- Is there any sampling strategy **better** than uniform sampling? Yes, importance sampling (Zhao and Zhang, 2015; Yuan et al., 2016)!

- Assume each data is sampled from distribution $\mathcal{D}_p$, i.e.,

$$\mathbb{P}(\xi_k = \xi_i) = p_i, \quad \forall\ i$$

and define $F_p(x;\xi_i) = \frac{1}{Np_i}F(x;\xi_i)$, it is easy to verify that

$$\frac{1}{N}\sum_{i=1}^{N} F(x;\xi_i) = \mathbb{E}_{\xi\sim\mathcal{D}_p}[F_p(x;\xi)]$$

- In summary, finite-sum minimization is equivalent to

$$\min_{x\in\mathbb{R}^d} \frac{1}{N}\sum_{i=1}^{N} F(x;\xi_i) \quad \Longleftrightarrow \quad \min_{x\in\mathbb{R}^d} \mathbb{E}_{\xi\sim\mathcal{D}_p}[F_p(x;\xi)]$$

SGD with importance sampling

- Consider the stochastic optimization problem

$$\min_{x \in \mathbb{R}^d} \quad \mathbb{E}_{\xi \sim \mathcal{D}_p}[F_p(x;\xi)]$$

where $\mathbb{P}(\xi = \xi_i) = p_i$. We will decide $\{p_i\}_{i=1}^n$ later.

- Applying SGD to the above problem, we reach the following recursion

$$\text{Sample } \xi_k \sim \mathcal{D}_p \text{ with probability } \mathbb{P}(\xi_k = \xi_i) = p_i$$
$$x_{k+1} = x_k - \gamma \nabla F_p(x_k;\xi_k) = x_k - \frac{\gamma}{N p_i} \nabla F(x_k;\xi_k)$$

We refer the above algorithm as SGD with importance sampling

SGD with importance sampling

Now we examine the property of the stochastic gradient $\nabla F_p(x;\xi)$

- the stochastic gradient is unbiased

$$\mathbb{E}[\nabla F_p(x;\xi)] = \sum_{i=1}^{N} \frac{p_i}{Np_i} \nabla F(x;\xi_i) = \frac{1}{N} \sum_{i=1}^{N} \nabla F(x;\xi_i) = \nabla f(x)$$

- the variance is bounded by

$$\mathbb{E}\|\nabla F_p(x_k;\xi) - \nabla f(x_k)\|^2 \leq 2L_p^2 \|x_k - x^\star\|^2 + \sigma_p^2$$

where

$$L_p^2 = \sum_{i=1}^{N} \frac{L^2}{p_i N^2}, \quad \sigma_p^2 = \sum_{i=1}^{N} \frac{2}{p_i N^2} \|\nabla F(x^\star;\xi_i)\|^2$$

(We leave it as an exercise)

SGD with importance sampling

- Similar to Theorem 1, we can derive the convergence of SGD with importance sampling as follows (we leave it as an exercise)

$$\mathbb{E}\|x_k - x^\star\|^2 \le (1-\gamma\mu)^k \|x_0 - x^\star\|^2 + \frac{2\gamma\sigma_p^2}{\mu}$$

- Note that sampling probability influences σ_p^2

- We now determine the optimal sampling probability that minimizes σ_p^2

$$\begin{aligned} \min_{\{p_i\}_{i=1}^n} \quad & \sum_{i=1}^N \frac{1}{p_i} \|\nabla F(x^\star;\xi_i)\|^2 \\ \text{s.t.} \quad & \sum_{i=1}^N p_i = 1, \quad p_i \ge 0 \end{aligned} \tag{2}$$

SGD with importance sampling

- Solve problem (2), we achieve

$$p_i^\star = \frac{\|\nabla F(x^\star; \xi_i)\|}{\sum_{j=1}^{N} \|\nabla F(x^\star; \xi_j)\|}$$

- Substituting $p_i^\star$ into σ_p^2, we have

$$\sigma_p^2 = 2\Big(\frac{1}{N}\sum_{i=1}^{N} \|\nabla F(x^\star; \xi_i)\|\Big)^2$$

  which is equal to or less than $\sigma^2 = \frac{2}{N}\sum_{i=1}^{N} \|\nabla F(x^\star; \xi_i)\|^2$ achieved by uniform sampling

- Importance sampling achieves more accurate solution than uniform sampling

SGD with importance sampling

- However, the optimal sampling probability

$$p_i^\star = \frac{\|\nabla F(x^\star; \xi_i)\|}{\sum_{j=1}^N \|\nabla F(x^\star; \xi_j)\|}$$

  cannot be directly used due to the unknown $x^\star$.

- We thus approximate it by

$$p_i^k = \frac{\|\nabla F(x_k; \xi_i)\|}{\sum_{j=1}^N \|\nabla F(x_k; \xi_j)\|}$$

  and expect $p_i^k \to p_i^\star$.

- Very expensive due to the computation of $\{\|\nabla F(x_k; \xi_j)\|\}_{j=1}^N$ every iteration

SGD with importance sampling

- We introduce an auxiliary vector $\psi_k \in \mathbb{R}^N$

- Each entry $\psi_k(i)$ is to estimate $\|\nabla F(x_k; \xi_i)\|$ as follows

$$\psi_k(i) = \begin{cases} \eta\psi_{k-1}(i) + (1-\eta)\|\nabla F(x_k; \xi_i)\| & \text{if } \xi_i \text{ is sampled} \\ \psi_{k-1}(i) & \text{if } \xi_i \text{ is not sampled} \end{cases} \tag{3}$$

- Only one $\|\nabla F(x_k; \xi_i)\|$ is calculated per iteration, not all of them.

- We introduce θ_k to track $\sum_{j=1}^{N} \|\nabla F(x_k; \xi_j)\|$ as follows

$$\begin{aligned} \theta_k &= \sum_{j=1}^{N} \psi_k(j) = \sum_{j=1}^{N} \psi_{k-1}(j) + (\psi_k(i) - \psi_{k-1}(i)) \\ &= \theta_{k-1} + (1-\eta)(\|\nabla F(x_k; \xi_i)\| - \psi_{k-1}(i)) \end{aligned}$$

  Very efficient to update.

SGD with importance sampling

Update sample probability $p_k^i = \psi_{k-1}(i)/\theta_{k-1}$

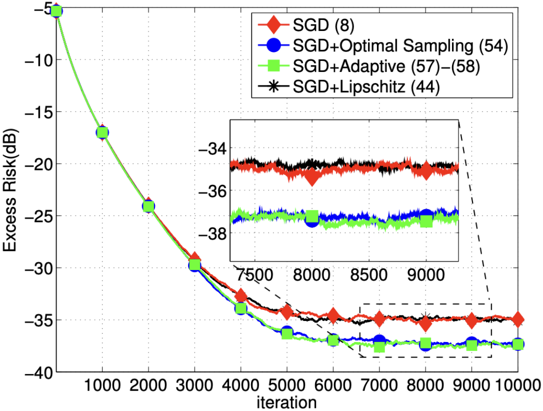
Sample $\xi_k \sim \mathcal{D}_p$ with probability $\mathbb{P}(\xi_k = \xi_i) = p_k^i$

$$x_{k+1} = x_k - \frac{\gamma}{N p_k^i} \nabla F_p(x_k; \xi_k)$$

Update $\psi_k(i)$ according to (3)

Update $\theta_k = \theta_{k-1} + (1-\eta)(\|\nabla F(x_k; \xi_i)\| - \psi_{k-1}(i))$

SGD with importance sampling



SGD with random reshuffling

- In SGD discussed above, we sample data **with** replacement
- In practice, we usually sample data **without** replacement

For $t = 1, \cdots, T$ **do**

Sample a permutation $\sigma(1), \cdots, \sigma(N)$

from $\{1, \cdots, N\}$ uniformly at random

For $k = 1, \cdots, N$ **do**

$$x_{k+1}^t = x_k^t - \gamma \nabla F(x_k^t; \xi_{\sigma(k)})$$

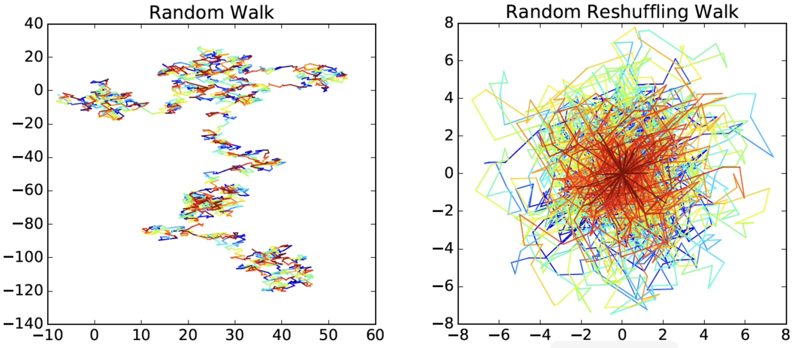
End For

$x_0^{t+1} = x_N^t$

End For

SGD with random reshuffling

Random reshuffling can reduce the variance of gradient noise



The scale of random walk is much larger than random reshuffling.

SGD with random reshuffling

- Standard SGD in the strongly convex and smooth scenario will converge as

$$\limsup_{k\to\infty} \mathbb{E}\|x_k - x^\star\|^2 = O(\gamma) \qquad \text{(constant learning rate)}$$

$$\mathbb{E}\|x_k - x^\star\|^2 = O(1/k) \qquad \text{(decay learning rate)}$$

- SGD with RR will converge as

Theorem 2

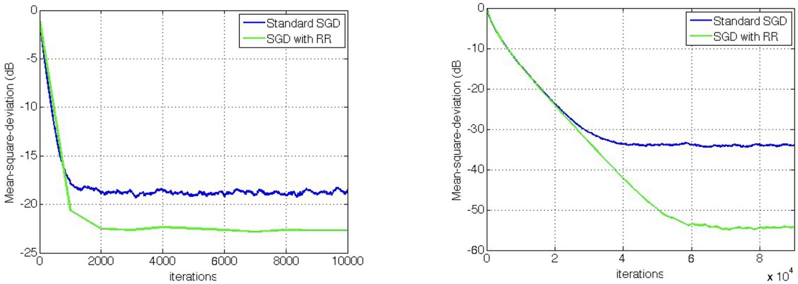
Suppose $f(x)$ is L-smooth and μ-strongly convex. SGD with random reshuffling will converge as

$$\limsup_{k\to\infty} \mathbb{E}\|x_k - x^\star\|^2 = O(\gamma^2) \qquad \textit{(constant learning rate)}$$

$$\mathbb{E}\|x_k - x^\star\|^2 = O(1/k^2) \qquad \textit{(decay learning rate)}$$

Random reshuffling improves the convergence rate of SGD.

SGD with random reshuffling



$\gamma = 0.003$ $\qquad$ $x\ (\text{dB}) = 10\log_{10}(x)$ $\qquad$ $\gamma = 0.0003$

Summary

- Finite-sum minimization is a special example of stochastic optimization
- SGD with finite sample size converges without bounded variance assumption
- Importance sampling improves SGD performance
- Random reshuffling improves SGD performance

References I

K. Yuan, B. Ying, S. Vlaski, and A. H. Sayed, "Stochastic gradient descent with finite samples sizes," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2016, pp. 1–6.

P. Zhao and T. Zhang, "Stochastic optimization with importance sampling for regularized loss minimization," in *international conference on machine learning*. PMLR, 2015, pp. 1–9.