DSC 211 Introduction to Optimization Winter 2024 Instructor: Jun-Kun Wang
Scribe: James Chen January 30, 2024
Editors/TAs: Merlin Chang, Marialena Sfyraki

Lecture 7: Introduction to stochastic optimization

1 Stochastic Optimization

Assume that a problem has an underlying structure. Our goal is still

$$\min_x F(x)$$

where $F(x) := \mathbb{E}_z[f(x; z)]$

Algorithm 1 Stochastic Gradient Descent

1: **for** $k = 1, 2, \ldots$ **do**
2: Compute a stochastic gradient g_k that satisfies $\mathbb{E}_z[g_k] = \nabla F(x_k)$
3: $x_{k+1} = x_k - \eta g_k$.
4: **end for**

For finite-sum problem, the algorithm becomes

Algorithm 2 Finite-sum problem

1: **for** $k = 1, 2, \ldots, K$ **do**
2: sample $i_k \in [n]$
3: $g_k \leftarrow \nabla f_{i_k}$
4: $x_{k+1} = x_k - \eta g_k$.
5: **end for**

For example, $F(x) \triangleq \frac{1}{n}\sum_{i=1}^n f_i(x)$, we have

$$\mathbb{E}_{i_k}[f_{i_k}(x)] = F(x)$$
$$i_k \in [n] \text{ when } \Pr(i_k = i) = \frac{1}{n}$$

1.1 Comparison between iteration complexity

Recall the comparison between Stochastic Gradient Descent and Gradient Descent

	SGD	GD
smooth convex	$O(\frac{1}{\sqrt{K}})$	$O(\frac{1}{K})$
strongly convex and smooth	$O(\frac{1}{K})$	$O(\exp(-K))$

Thus, in order to reach an ϵ-gap, we need to set K for **smooth** convex function to:

$$(SGD) \qquad \epsilon = O\left(\frac{1}{\sqrt{K}}\right) \Rightarrow K = O\left(\frac{1}{\epsilon^2}\right)$$
$$(GD) \qquad \epsilon = O\left(\frac{1}{K}\right) \Rightarrow K = O\left(\frac{1}{\epsilon}\right)$$

We can compute the ratio between the running time of SGD and GD,

$$\begin{aligned}\frac{\text{running time of SGD}}{\text{running time of GD}} &= \frac{\text{\# iterations of SGD}}{\text{\# iterations of GD}} \times \frac{\text{cost per step SGD}}{\text{cost per step GD}} \\ &= \frac{1/\epsilon^2}{1/\epsilon} \times \frac{1}{N} \\ &= \frac{1}{\epsilon N} \ll 1,\end{aligned}$$

where N is the number of dimensions or number of data points. Thus, the condition when SGD is faster than GD is

$$\frac{1}{\epsilon} \ll N.$$

Remark 1. *When the sample size N is much larger than the inverse of the targeted ϵ, SGD is faster than GD. In machine learning and data science, N is usually large and we do not need ϵ to be very small to fit the data.*

1.2 Iteration complexity of SGD

Theorem 1. *Let $F(x) = \mathbb{E}_z[f(x;z)] : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Consider the update*

$$x_{k+1} = x_k - \eta g_k,$$

where $\mathbb{E}_z[g_k] = \nabla F(x_k)$. Suppose $x^* = \arg\min_x F(x)$ exists and the initial distance is bounded, i.e., $\|x_1 - x^*\| \le D$. Then, the following inequality holds:

$$\frac{1}{K}\left(\sum_{k=1}^{K} F(x_k) - F(x^*)\right) \le \frac{\eta}{2K}\left(\sum_{k=1}^{K} \mathbb{E}\left[\|g_k\|^2\right]\right) + \frac{\|x_1 - x^*\|^2}{2\eta K}.$$

Also, let $\bar{x}_K = \frac{1}{K}\sum_{k=1}^{K} x_k$. Then, for the average of the optimality gap, the following inequality regarding the function value at $\bar{x}_K$ compared to the optimal value $F(x^*)$ holds:

$$F(\bar{x}_K) - F(x^*) \le \frac{1}{K}\left(\sum_{k=1}^{K} f(x_k) - F(x^*)\right). \tag{1}$$

Remark 2. *The result* (1) *is by an application of Jensen's inequality, which we will show in HW2. Jensen's inequality states that if f is a convex function, then for any $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ and any non-negative weights $a_1, a_2, \ldots, a_n$ such that $\sum_{i=1}^n a_i = 1, a_i \geq 0, \forall i$, the following inequality holds:*

$$f\left(\sum_{i=1}^n a_i x_i\right) \leq \sum_{i=1}^n a_i f(x_i).$$

The 0*-order characterization of convexity is equivalent to the base case of Jensen's inequality.*

$$\forall x_1, x_2 \in \mathbb{R}^n, \forall \alpha \in [0,1], \quad f((1-\alpha)x_1 + \alpha x_2) \leq (1-\alpha)f(x_1) + \alpha f(x_2).$$

Remark 3. *Given the additional assumption that $\mathbb{E}[\|g_k\|^2] \leq G^2$, we have the following inequality for the function value at the averaged iterates $\bar{x}_K$ compared to the optimal value $F(x^*)$:*

$$F(\bar{x}_K) - F(x^*) \leq \frac{\eta G^2}{2} + \frac{D^2}{2\eta K},$$

where $\bar{x}_K := \frac{1}{K}\sum_{k=1}^K x_k$. To minimize the upper bound, we choose η to satisfy

$$\frac{\eta G^2}{2} = \frac{D^2}{2\eta K}$$

From here we have

$$\eta^2 G^2 K = D^2 \Rightarrow \eta = \frac{D}{G\sqrt{K}}$$

Plugging it back to the bound and we get:

$$\frac{\eta G^2}{2} + \frac{D^2}{2\eta K} = \frac{DG}{\sqrt{K}},$$

It is noted that we need to identify a condition such that the assumption of the expected sqaured size of the stochastic gradient g_k does hold, i.e., $\mathbb{E}[\|g_k\|^2] \leq G^2$ holds. One of the remedy is by considering that the underlying function $F(\cdot)$ is Lipschitz and using Projected SGD. Specifically, recall that a function is G-Lipschitz over C with respect to the norm $\|\cdot\|$, if for any $x, y \in C$, $|f(x) - f(y)| \leq G\|x - y\|$, where $G > 0$. Then, the following theorem states that a Lipschitz function over a bounded domain has a bounded gradient.

Theorem 2. *Suppose $f(\cdot) : C \to \mathbb{R}$ is a convex function. Then $f(\cdot)$ is G-Lipschitz over C with respect to a norm $\|\cdot\|$ if and only if for any $x \in C$, its sub-gradients $g_x \in \partial f(x)$ satisfy*

$$\|g_x\|_* \leq G,$$

where G is a constant and $\|\cdot\|_*$ denotes the dual norm.

For the proof, see Lemma 2.6 in Online Learning and Online Convex Optimization by Shai Shalev-Shwartz [2].

From here we move to the next section on how to derive these results.

Proof. (of Theorem 2) Given the update step of SGD, the expectation of the squared norm between the next iterate and the optimum is expressed as:

$$\begin{aligned}\mathbb{E}\left[\|x_{k+1} - x_*\|^2\right] &= \mathbb{E}\left[\|x_k - \eta g_k - x_*\|^2\right] \\ &= \mathbb{E}\left[\|x_k - x_*\|^2 - 2\eta\langle g_k, x_k - x_*\rangle + \eta^2\|g_k\|^2\right].\end{aligned}$$

Rearranging and using linearity of expectation gives:

$$\begin{aligned}\mathbb{E}\left[\langle g_k, x_k - x_*\rangle\right] &= \frac{1}{2\eta}\mathbb{E}\left[\|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2\right] + \frac{\eta}{2}\mathbb{E}\left[\|g_k\|^2\right] \\ &= \mathbb{E}_{z_1 - z_k}\left[\langle g_k, x_k - x_*\rangle\right] \\ &\quad \text{(since the expected values of the first term is expectation over all the randomness in k iterations)} \\ &= \sum_* \mathbf{Pr}(z_{1:k-1} = *)\mathbb{E}_{z_k}\left[\langle g_k, x_k - x_*\rangle | z_{1:k-1} = *\right] \\ &\quad (z_{1:k-1} \text{ is defined in the form of the realization of the randomness of } z \text{ from 1 to } k-1) \\ &= \mathbb{E}_{z_1 - z_{k-1}}\left[\mathbb{E}_{z_k}\left[\langle g_k, x_k - x_*\rangle | z_{1:k-1}\right]\right] \\ &\quad (x_k := \text{ realization of } z_{1:k-1},\ x_k \text{ is determined by } z_{1:k-1}) \\ &= \mathbb{E}_{z_1 - z_{k-1}}\left[\mathbb{E}_{z_k}\left[\langle g_k, x_k - x_*\rangle | x_k\right]\right],\end{aligned}$$

Additionally, we have

$$\begin{aligned}\mathbb{E}_{z_k}\left[\langle g_k, x_k - x_*\rangle | x_k\right] &= \sum_{i=1}^{n} \mathbf{Pr}(i_k = i)\langle g_k, x_k - x_*\rangle \\ &= \langle \nabla F(x_k), x_k - x_*\rangle \ \textit{given}\ x_k\end{aligned}$$

In summary, we note that the expectation $\mathbb{E}[\langle g_k, x_k - x^*\rangle]$ can be expanded as:

$$\begin{aligned}\mathbb{E}_{z_1:k}[\langle g_k, x_k - x_*\rangle] &= \mathbb{E}_{z_1:k-1}[\mathbb{E}_{z_k}[\langle g_k, x_k - x_*\rangle | z_{1:k-1}]] \\ &= \mathbb{E}_{z_1:k-1}[\mathbb{E}_{z_k}[\langle g_k, x_k - x_*\rangle | x_k]] \\ &= \mathbb{E}_{z_1:k-1}[\langle \nabla F(x_k), x_k - x_*\rangle] \\ &= \mathbb{E}_{z_1:k}[\langle \nabla F(x_k), x_k - x_*\rangle] \\ &\geq \mathbb{E}_{z_1:k}[F(x_k) - F(x_*)] \text{ by first order convexity of } F\end{aligned}$$

By chaining the above inequalities, we have:

$$\begin{aligned}\mathbb{E}[F(x_k) - F(x_*)] &\leq \mathbb{E}_{z_{1:k}}[\langle g_k, x_k - x_*\rangle] \\ &= \frac{1}{2\eta}\mathbb{E}\left[\|x_k - x_*\|_2^2 - \|x_{k+1} - x_*\|_2^2\right] + \frac{\eta}{2}\mathbb{E}\left[\|g_k\|_2^2\right].\end{aligned}$$

Summing up over K iterations, we obtain:

$$\mathbb{E}\left[\sum_{k=1}^{K} F(x_k) - F(x_*)\right] \leq \frac{\|x_1 - x_*\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{k=1}^{K}\mathbb{E}\left[\|g_k\|_2^2\right].$$

□

2 Non-convex SGD

Theorem 3. *Assume that the variance of the stochastic gradient $\nabla f(x; z)$ is at most σ^2 for all x, i.e.,*

$$Var(\nabla f(x; z)) = \mathbb{E}_z\left[\|\nabla f(x; z) - \nabla F(x)\|_2^2\right] \leq \sigma^2.$$

Suppose $F(\cdot)$ is L-smooth. Then, SGD with the step size $\eta \leq \frac{1}{L}$ satisfies

$$\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla F(x_k)\|_2^2\right] \leq \frac{2(F(x_1) - F_*)}{\eta} + \eta L\sigma^2 K.$$

Remark 4: If the step size η is chosen as

$$\eta = \min\left(\frac{1}{L}, \sqrt{\frac{F(x_1) - F_*}{\sigma^2 LK}}\right),$$

then the sum of expected squared norms of the gradients over K iterations is bounded by

$$\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla F(x_k)\|_2^2\right] \leq 2(F(x_1) - F_*)L + 3\sigma\sqrt{(F(x_1) - F_*)LK}.$$

Remark 5: If $\hat{x}$ is selected uniformly at random from $\{x_1, \ldots, x_K\}$, then using Jensen we have

$$\mathbb{E}\left[\|\nabla F(\hat{x})\|\right] \leq \sqrt{\frac{2(F(x_1) - F_*)L}{K}} + \frac{3\sigma}{K^{1/4}}\sqrt{(F(x_1) - F_*)L}.$$

2.1 Mini-batch SGD

Algorithm 3 Minibatch SGD

1: **for** $k = 1, 2, \ldots, K$ **do**
2: **for** $i = 1, 2, \ldots, B$ **do**
3: $g_{k,i} = \nabla f(x_k; z_{(k-1)B+i})$.
4: **end for**
5: $g_k = \frac{1}{B}\sum_{i=1}^{B} g_{k,i}$, so here $\mathbb{E}[g_k] = \nabla F(x_k)$
6: $x_{k+1} = x_k - \eta g_k$.
7: **end for**

Remark 6. B is the batch size. In vanilla SGD we have $B = 1$.

Lemma 1. *Assume that the variance of the stochastic gradient $\nabla f(x; z)$ is at most σ^2 for all x, i.e.,*

$$\mathbb{E}_z\left[\|\nabla f(x; z) - \nabla F(x)\|_2^2\right] \leq \sigma^2.$$

Then, for the mini-batch gradient g_k, it holds that

$$\mathbb{E}_z\left[\|g_k - \nabla F(x_k)\|_2^2\right] \leq \frac{\sigma^2}{B},$$

where B is the mini-batch size.

Proof. The variance of mini-batch stochastic gradient is given by

$$g_k \triangleq \frac{1}{B}\sum_{i\in[B]} g_{k,i}$$

We have that

$$\begin{aligned}
\mathbb{E}_z\left[\|g_k - \nabla F(x_k)\|_2^2\right] &= \mathbb{E}_z\left[\left\|\frac{1}{B}\sum_{i\in[B]}(g_{k,i} - \nabla F(x_k))\right\|_2^2\right] \\
&= \mathbb{E}_z\left[\frac{1}{B^2}\sum_{i,j}\langle g_{k,i} - \nabla F(x_k), g_{k,j} - \nabla F(x_k)\rangle\right] \\
&= \mathbb{E}_z\left[\frac{1}{B^2}(\sum_i \left\|(g_{k,i} - \nabla F(x_k))\right\|_2^2) + \sum_{j\neq i}\langle g_{k,i} - \nabla F(x_k), g_{k,j} - \nabla F(x_k)\rangle\right]
\end{aligned}$$

The second term is 0 since for each element in the summation, the expection is 0:

$$\begin{aligned}
&\mathbb{E}\left[\langle g_{k,i} - \nabla F(x_k), g_{k,j} - \nabla F(x_k)\rangle\right] \\
=&\langle \mathbb{E}\left[g_{k,i} - \nabla F(x_k)\right], \mathbb{E}\left[g_{k,j} - \nabla F(x_k)\right]\rangle \\
=&\langle 0, 0\rangle \\
=&0
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{E}_z\left[\|g_k - \nabla F(x_k)\|_2^2\right] \\
=&\mathbb{E}_z\left[\frac{1}{B^2}\sum_{i,j}\langle g_{k,i} - \nabla F(x_k), g_{k,j} - \nabla F(x_k)\rangle\right] \\
=&\mathbb{E}_z\left[\frac{1}{B^2}(\sum_i \left\|(g_{k,i} - \nabla F(x_k))\right\|_2^2)\right] \\
\leq&\frac{1}{B^2}B\sigma^2 \\
=&\frac{\sigma^2}{B}
\end{aligned}$$

□

2.2 Iteration complexity of Mini-Batch SGD

Recall the result of the last theorem for a randomly selected $\hat{x}$ from $\{x_1, \ldots, x_K\}$, we have:

$$\mathbb{E}\left[\|\nabla F(\hat{x})\|\right] \leq \frac{\sqrt{2(F(x_1) - F_*)L}}{\sqrt{K}} + \frac{\sqrt{3\sigma\sqrt{(F(x_1) - F_*)L}}}{K^{1/4}}.$$

Now, let $\sigma \leftarrow \sigma\sqrt{B}$ for Mini-batch SGD.

Theorem 4. *Assume the variance of the stochastic gradient $\nabla f(x; z)$ is at most σ^2. Set*

$$\eta = \min\left(\frac{1}{L}, \frac{\sqrt{F(x_1) - F_*}}{(\sigma/\sqrt{B})\sqrt{LK}}\right),$$

then Mini-batch SGD achieves

$$\mathbb{E}\left[\|\nabla F(\hat{x})\|\right] \leq \frac{\sqrt{2(F(x_1) - F_*)L}}{\sqrt{K}} + \frac{\sqrt{3\sigma\sqrt{(F(x_1) - F_*)L}}}{(BK)^{1/4}}. \tag{2}$$

2.3 Comparison between vanilla SGD and mini-batch SGD

Now assume that the last term in (2) dominates (i.e., being the slow term). Then, we can compare vanilla SGD and mini-batch SGD as follows.

	mini-batch SGD	vanilla SGD
convergent rate	$\frac{1}{(BK)^{\frac{1}{4}}}$	$\frac{1}{K^{\frac{1}{4}}}$
cost per iteration	B	1
total cost over K	BK	K
convergent rate	$\frac{1}{(\text{total cost})^{\frac{1}{4}}}$	$\frac{1}{(\text{total cost})^{\frac{1}{4}}}$

In terms of the total cost, there is no difference of using different batch size. But for the mini-batch SGD, we can do the computation of stochastic gradients in a mini-batch in a parallel fashion. So in terms of the actual running time, mini-batch SGD would be faster comparing to SGD, provided by the aid of parallel computing.

It is noted that we cannot keep increase the batch size and available computing units to keep improving mini-batch SGD, since at some threshold, the first term $\frac{\sqrt{2(F(x_1)-F_*)L}}{\sqrt{K}}$ in (2) dominates and increasing the batch size after this threshold will just be harmful as it increases the cost. It is also noted that the $O(1/\sqrt{K})$ rate is the rate of GD.

Bibliographic notes

For more details about SGD, see e.g., [1]. The material for mini-bach SGD is based on [3].

References

[1] Alexander Rakhlin, Ohad Shamir, Karthik Sridharan. Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization ICML 2012

[2] Shai Shalev-Shwartz. Online Learning and Online Convex Optimization doi: 10.1561/2200000018

[3] Ashok Cutkosky. Lecture Note of Optimization for Machine Learning