ECE 273 Convex Optimization and Applications — Instructor: Jun-Kun Wang
Scribe: Zhuqing Li, Mainak Sarkar, Ya-Chi Liao — April 23, 2024
Editor/TA: Marialena Sfyraki

Lecture 7: Introduction to Stochastic Optimization

1 Review: Projected Gradient Descent and Frank-Wolfe Method

We begin by reviewing some results concerning Projected Gradient Descent and Frank-Wolfe method. Below is a formal statement of the Frank-Wolfe method algorithm.

Algorithm 1 Frank-Wolfe method

1: Initialize $\mathbf{x}_1 \in C$.
2: **for** $k = 1, 2, \ldots$ **do**
3: $\quad \mathbf{v}_k = \arg\min_{\mathbf{v} \in C} \langle \mathbf{v}, \nabla f(\mathbf{x}_k) \rangle$ (linear optimization)
4: $\quad \mathbf{x}_{k+1} = (1 - \eta_k)\mathbf{x}_k + \eta_k \mathbf{v}_k$, where $\eta_k \in [0, 1]$.
5: **end for**

Theorem 1. *Assume $f(\cdot)$ is a L-smooth convex function. Denote $D := \max_{\mathbf{x},\mathbf{y} \in C} \|\mathbf{x} - \mathbf{y}\|_2$, $\forall \mathbf{x}, \mathbf{y} \in C$ as the diameter of the set C. Let $\eta_k = \min\{1, \frac{2}{k}\} \in [0, 1]$. Then, Frank-Wolfe has:*

$$f(x_K) - f(x_*) \leq \frac{2LD^2}{K}.$$

Recall that PGD and GD share the same convergence rate. Specifically, if PGD is to achieve an ϵ-optimality gap in a constrained optimization problem: $f(\mathbf{x}_k) - \min_{\mathbf{x} \in C} f(\mathbf{x}) \leq \epsilon$, or if GD is to achieve an ϵ-optimality gap in an unconstrained optimization problem: $f(\mathbf{x}_k) - \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \leq \epsilon$, then the table below illustrates the convergence rates for both PGD and GD.

Convergence Rate	PGD	GD
L-smooth convex	$O\left(\frac{L}{k}\right)$	$O\left(\frac{L}{k}\right)$
L-smooth and μ-strongly convex	$O\left(\exp\left(-\frac{\mu}{L}k\right)\right)$	$O\left(\exp\left(-\frac{\mu}{L}k\right)\right)$

A natural question is: Can we achieve a faster convergence rate than $O\left(\frac{1}{K}\right)$ using Frank-Wolfe method when f is assumed to be smooth and strongly convex? The answer to this crucially relies on the regularity of the constrained set. We present two examples (see [1] and [2]) in the following.

Example 1 (A negative result). *If C is a probability simplex, i.e., $C := \{x \in \mathbb{R}^d : \sum_{i=1}^d x[i] \leq 1, x[i] \geq 0\}$, then $K = \Omega\left(\max\left(\frac{L}{\epsilon}, \frac{d}{2}\right)\right)$.*

Example 2 (A positive result). *Frank-Wolfe method gives a faster convergence rate when C is a μ-strongly convex set w.r.t. a norm $\|\cdot\|$, i.e., $x, z \in C$ implies that a ball centered at $\alpha x + (1-\alpha)z$ with a radius in $\alpha(1-\alpha)\frac{\mu}{2}\|x-z\|^2$ is in C, where $\alpha \in [0,1]$. In particular, any l_p norm with $p \in (1,2]$ satisfies the requirement.*

We refer to [3] and [4] for further details.

2 Introduction to Stochastic Optimization

Consider the following problem:

$$\min_{x \in \mathbb{R}^d} F(x), \text{ where } F(x) := \mathbb{E}_z[f(x; z)].$$

Here z denotes the randomness of this problem. Below is a formal statement of the SGD algorithm.

Algorithm 2 Stochastic Gradient Descent (SGD)

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**

Example 3 (Finite-sum problem). *Let $F(x) := \frac{1}{n}\sum_{i=1}^n f_i(x) = \mathbb{E}_i[f_i(x)]$. Then the SGD algorithm for finite-sum problem can be explicitly stated as follows:*

Algorithm 3 Stochastic Gradient Descent (SGD) for finite-sum problem

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

Notice that,

$$\mathbb{E}[g_k] = \sum_{i=1}^n P(i_k = i)\; \nabla f_i(x_k) = \sum_{i=1}^n \frac{1}{n} \nabla f_i(x_k) = \nabla F(x_k).$$

The convergence rates for SGD and GD are presented in the following table.

Convergence Rate	**SGD**	**GD**
L-smooth convex	$O\left(\frac{1}{\sqrt{K}}\right)$	$O\left(\frac{1}{K}\right)$
L-smooth and μ-strongly convex	$O\left(\frac{1}{K}\right)$	$O\left(\exp(-K)\right)$

For SGD, we compute

$$\epsilon = \frac{1}{\sqrt{K}} \Leftrightarrow K = \frac{1}{\epsilon^2},$$

so we asymptotically need $\frac{1}{\epsilon^2}$ iterations to reach an ϵ-optimality gap. For GD, we compute

$$\epsilon = \frac{1}{K} \Leftrightarrow K = \frac{1}{\epsilon},$$

so we asymptotically need $\frac{1}{\epsilon}$ iterations to reach an ϵ-optimality gap. Therefore, considering the practical scenario where $0 < \epsilon << 1$, SGD generally requires more iterations than GD to reach the given optimality gap. However, in terms of running time, formally we have

$$\frac{\text{running time of SGD}}{\text{running time of GD}} = \frac{\text{\# iterations of SGD}}{\text{\# iterations of GD}} \times \frac{\text{cost per step SGD}}{\text{cost per step GD}}.$$

Comparing the convergence rate of SGD and GD, we have (see [5]) $\frac{\text{\# iterations of SGD}}{\text{\# iterations of GD}} = \frac{1}{\epsilon^2}/\frac{1}{\epsilon}$. In each iteration of SGD, we only need to compute the gradient for one random case among n possibilities, so $\frac{\text{cost per step SGD}}{\text{cost per step GD}} = \frac{1}{n}$. Therefore,

$$\frac{\text{running time of SGD}}{\text{running time of GD}} = \frac{1}{\epsilon n}.$$

In order for SGD to have a better performance than GD, i.e., $\frac{\text{running time of SGD}}{\text{running time of GD}} << 1$, we need to ensure $\frac{1}{\epsilon} << n$. This condition is usually fulfilled when we have a very large sample size/data set. For example, consider the empirical risk minimization problem. Let $\{(y_i, z_i),\ i \in [n]\}$ be the data set and the function $F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$, where $f_i(x) = \frac{1}{2}(y_i - z_i^T x)^2$. Since, in practical applications, we may have a very large n and expect a relatively large ϵ considering the possible overshooting effect, it is reasonable to assume that $\frac{1}{\epsilon} << n$.

3 Iteration Complexity of SGD:

We now present a theorem that provides an upper-bound for the optimality gap obtained through SGD.

Theorem 2. *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* $x_* = \arg\min F(x)$ *exists and the initial distance is bounded, i.e.,* $\|x_1 - x_*\| \leq D$*. Then,*

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

Proof. One way to gauge the progress of an iterative optimization algorithm is through the distance metric d_k, which calculates the Euclidean distance between the updated point in the k-th iteration, denoted as x_{k+1}, and the global minima point x_*. This can be expressed as $d_k = ||x_{k+1} - x_*||_2$.

Now, when we apply Stochastic Gradient Descent (SGD) to the convex function $F(x) = \mathbb{E}_z[f(x;z)] : \mathbb{R}^d \to \mathbb{R}$, it is clear from the nature of the SGD algorithm (refer to Algorithm 2) that the distance metric $d_k = ||x_{k+1} - x_*||_2$, will be a random variable since x_{k+1} is a random variable. Therefore, our first step in this proof will be to consider the expected value of the squared distance metric d_k, that is $\mathbb{E}[||x_{k+1} - x_*||_2^2]$.

We have the following,

$$\begin{aligned}
\mathbb{E}[\,||x_{k+1} - x_*||_2^2\,] \;&=\; \mathbb{E}[\,||x_k - \eta g_k - x_*||_2^2\,] \quad \text{(SGD's update, see Algorithm 2)} \\
&=\; \mathbb{E}[\,||x_k - x_*||_2^2 \;-\; 2\eta\langle g_k,\, x_k - x_*\rangle \;+\; \eta^2||g_k||_2^2\,]
\end{aligned}$$

Rearranging the above equation we get,

$$\begin{aligned}
2\eta \cdot \mathbb{E}[\,\langle g_k,\, x_k - x_*\rangle\,] \;&=\; \mathbb{E}[\,||x_k - x_*||_2^2 - ||x_{k+1} - x_*||_2^2\,] \;+\; \eta^2\mathbb{E}[\,||g_k||_2^2\,] \\
\Leftrightarrow\; \mathbb{E}[\,\langle g_k,\, x_k - x_*\rangle\,] \;&=\; \frac{\mathbb{E}[\,||x_k - x_*||_2^2 - ||x_{k+1} - x_*||_2^2\,]}{2\eta} \;+\; \frac{\eta}{2}\mathbb{E}[\,||g_k||_2^2\,]
\end{aligned} \tag{2}$$

Now, notice the term inside the expectation operator on the L.H.S of the equation

(2), i.e. $\langle g_k,\, x_k - x_* \rangle$. In this term, it is evident that g_k and x_k are random variables. Let us understand how.

In line 2 of the SGD Algorithm (refer to 2), it states: "Compute a stochastic gradient g_k such that $\mathbb{E}_z[g_k] = \nabla F(x_k)$". This definition of the stochastic gradient g_k indicates that g_k is influenced by both the current point x_k and the randomness associated with the k-th iteration, denoted as z_k (which corresponds to the sampled index number in the finite-sum problem). Consequently, considering the update expression of SGD as $x_{k+1} = x_k - \eta g_k$, it follows that the next update x_{k+1} also depends on the current point x_k and the randomness of the k-th iteration z_k.

Let's begin by assuming that the initial point x_1 is known. Consequently, the randomness affecting g_1 (and therefore x_2) is entirely based on z_1. In other words, if z_1 is known, then g_1 (and thus x_2) can be computed deterministically.

Now, let's delve into the second iteration step. We understand that the randomness influencing g_2 (and hence x_3) relies on both z_2 and the current point x_2, whose randomness, in turn, depends on z_1. This implies that if both z_1 and z_2 are known, then g_2 (and hence x_3) can be determined with certainty. Consequently, through this analysis and the application of mathematical induction, we can infer that g_k is influenced by $z_1,\, z_2, \ldots, z_k$, while x_k depends on $z_1,\, z_2, \ldots, z_{k-1}$.

So, with the analysis that we performed above, and denoting $z_{1:k} = z_1,\, z_2, \ldots, z_k$, we can write that,

$$\begin{aligned}
\mathbb{E}[\, \langle g_k,\, x_k - x_* \rangle \,] &= \mathbb{E}_{z_{1:k}}[\, \langle g_k,\, x_k - x_* \rangle \,] \\
&= \sum_{\circ} \Pr(z_{1:k-1} = \circ)\, \mathbb{E}_{z_k}[\langle g_k,\, x_k - x_* \rangle \mid z_{1:k-1} = \circ] \qquad \text{(Law of Total Expectation)} \\
&= \mathbb{E}_{z_{1:k-1}} \left[\, \mathbb{E}_{z_k}[\, \langle g_k,\, x_k - x_* \rangle \mid z_{1:k-1}] \,\right] \\
&= \mathbb{E}_{z_{1:k-1}} \left[\, \mathbb{E}_{z_k}[\, \langle g_k,\, x_k - x_* \rangle \mid x_k] \,\right] \qquad \text{(Since, } x_k \text{ is determined by } z_{1:k-1}\text{)}
\end{aligned} \tag{3}$$

Now, let us suppose that we are dealing with the finite-sum problem, where we attempt to minimize $F(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) = \mathbb{E}_i[f_i(x)]$. From Algorithm 3, it is evident that when SGD is applied for solving this problem, the index i_k is the source of randomness associated with each iteration. Furthermore, after randomly sampling

i_k, the stochastic gradient is computed as: $g_k = \nabla f_{i_k}(x_k)$. So, the inner expectation term $\mathbb{E}_{z_k}[\,\langle g_k,\ x_k - x_*\rangle \mid x_k]$ in Equation (3) can be written as follows,

$$
\begin{aligned}
\mathbb{E}_{z_k}[\,\langle g_k,\ x_k - x_*\rangle \mid x_k] \;&=\; \mathbb{E}_{i_k}[\,\langle \nabla f_{i_k}(x_k),\ x_k - x_*\rangle \mid x_k] \\
&=\; \sum_{i=1}^{n} \Pr(i_k = i)\,\langle \nabla f_i(x_k),\ x_k - x_*\rangle \\
&=\; \sum_{i=1}^{n} \frac{1}{n}\,\langle \nabla f_i(x_k),\ x_k - x_*\rangle \qquad (4)\\
&=\; \langle \sum_{i=1}^{n} \frac{1}{n} \nabla f_i(x_k),\ x_k - x_*\rangle \\
&=\; \langle\, \nabla F(x_k),\ x_k - x_*\rangle
\end{aligned}
$$

The final equation follows from the fact that

$$
F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \;\Rightarrow\; \nabla F(x_k) = \sum_{i=1}^{n} \frac{1}{n} \nabla f_i(x_k).
$$

Using the above expression for $\mathbb{E}_{z_k}[\,\langle g_k,\ x_k - x_*\rangle \mid x_k]$ in the R.H.S of Equation (3), we can write the following,

$$
\begin{aligned}
\mathbb{E}[\,\langle g_k,\ x_k - x_*\rangle\,] \;&=\; \mathbb{E}_{\,z_{1:k-1}}\left[\,\langle \nabla F(x_k),\ x_k - x_*\,\rangle\,\right] \\
&=\; \mathbb{E}_{\,z_{1:k}}\left[\,\langle \nabla F(x_k),\ x_k - x_*\,\rangle\,\right] \quad \text{(Since, } x_k \text{ is independent of } z_k) \qquad (5)
\end{aligned}
$$

Since it is given that $F(x) = \mathbb{E}_z[f(x;z)] : \mathbb{R}^d \to \mathbb{R}$ is a convex function, therefore from the first-order characterization of convexity, we can say that the following will be true,

$$F(x_*) \; \geq \; F(x_k) + \langle \, \nabla F(x_k), \; x_* - x_k \, \rangle$$

$$\Leftrightarrow F(x_k) \; - \; F(x_*) \; \leq \; \langle \, \nabla F(x_k), \; x_k - x_* \, \rangle \tag{6}$$

$$\Rightarrow \mathbb{E}_{\, z_{1:k}}[F(x_k) \; - \; F(x_*)] \; \leq \; \mathbb{E}_{\, z_{1:k}}[\, \langle \nabla F(x_k), \; x_k - x_* \, \rangle \,].$$

Using the above inequality in the last line of Equation (5) we get the following,

$$\mathbb{E}_{\, z_{1:k}}[F(x_k) \; - \; F(x_*)] \; \leq \; \mathbb{E}[\, \langle g_k, \; x_k - x_* \rangle \,]. \tag{7}$$

Now, if we further combine the inequality result that we obtained above with the final equation of 2 we get that,

$$\mathbb{E}[F(x_k) \; - \; F(x_*)] \; \leq \; \frac{\mathbb{E}[\, ||x_k - x_*||_2^2 - ||x_{k+1} - x_*||_2^2 \,]}{2\eta} \; + \; \frac{\eta}{2}\mathbb{E}[\, ||g_k||_2^2 \,]. \tag{8}$$

Summing up the inequality of (8) from $k = 1$ to K we get the following,

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

$$= \; \frac{||x_1 - x_*||_2^2}{2\eta} - \frac{\mathbb{E}[\, ||x_{K+1} - x_*||_2^2 \,]}{2\eta} \; + \; \frac{\eta}{2} \left(\sum_{k=1}^{K} \mathbb{E} \left[||g_k||_2^2 \right] \right) \qquad (x_1 \text{ is known})$$

$$\leq \; \frac{||x_1 - x_*||_2^2}{2\eta} \; + \; \frac{\eta}{2} \left(\sum_{k=1}^{K} \mathbb{E} \left[||g_k||_2^2 \right] \right) \qquad \left(\text{Since, } \frac{\mathbb{E}[\, ||x_{K+1} - x_*||_2^2 \,]}{2\eta} \; \geq \; 0 \right). \tag{9}$$

Using the Linearity of Expectation on the LHS of the inequality (9) we get the following,

$$\mathbb{E}\left[\sum_{k=1}^{K}(F(x_k) - F(x_*)\,)\right] \;\leq\; \frac{||x_1 - x_*||_2^2}{2\eta} \;+\; \frac{\eta}{2}\left(\sum_{k=1}^{K}\mathbb{E}\left[\|g_k\|_2^2\right]\right). \tag{10}$$

Now, if we divide both sides of the above inequality by K, we finally obtain the following guarantee inequation for the application of SGD to convex functions of the form $F(x) = \mathbb{E}_z[f(x;z)] : \mathbb{R}^d \to \mathbb{R}$,

$$\frac{1}{K}\mathbb{E}\left[\sum_{k=1}^{K}(F(x_k) - F(x_*)\,)\right] \leq \frac{\eta}{2K}\left(\sum_{k=1}^{K}\mathbb{E}\left[\|g_k\|_2^2\right]\right) + \frac{\|x_1 - x_*\|_2^2}{2\eta K}. \tag{11}$$

□

Corollary 1. Let us denote $\bar{x}_K := \frac{1}{K}\sum_{k=1}^{K} x_k$. Then,

$$\mathbb{E}\left[F(\bar{x}_K) - F(x_*)\right] \;\leq\; \frac{1}{K}\mathbb{E}\left[\sum_{k=1}^{K}(F(x_k) - F(x_*)\,)\right]. \tag{12}$$

Proof. Jensen's Inequality states the following: If $g(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is a convex function, and D is any discrete distribution over $x_1,\ x_2,\ .....,\ x_n\ \in\ \mathbb{R}^d$. Then the following inequality holds,

$$g\left(\sum_{i=1}^{n} p_i x_i\right) \;\leq\; \sum_{i=1}^{n} p_i \cdot g(x_i), \tag{13}$$

where $p_i \geq 0, \forall i \in [n]$, and $\sum_{i\in[n]} p_i \;=\; 1$.

Therefore, knowing that $F(x) = \mathbb{E}_z[f(x;z)] : \mathbb{R}^d \to \mathbb{R}$ is a convex function, and considering D to be the uniform discrete distribution over $x_1,\ x_2,\ x_3,\ .....,\ x_K$ (i.e. $p_k = \frac{1}{K},\ \ \forall k \in [K]$), we can write the following

$$
\begin{aligned}
& F\left(\frac{1}{K}\sum_{k=1}^{K} x_k\right) \;\leq\; \frac{1}{K}\sum_{k=1}^{K} F(x_k) \\
& \Leftrightarrow F\left(\frac{1}{K}\sum_{k=1}^{K} x_k\right) - F(x_*) \;\leq\; \left(\frac{1}{K}\sum_{k=1}^{K} F(x_k)\right) - F(x_*) \\
& \Leftrightarrow F(\bar{x}_K) - F(x_*) \;\leq\; \frac{1}{K}\sum_{k=1}^{K}\big(F(x_k) - F(x_*)\,\big) \qquad \left(\text{Denoting } \bar{x}_K := \frac{1}{K}\sum_{k=1}^{K} x_k\right) \\
& \Rightarrow \mathbb{E}\left[F(\bar{x}_K) - F(x_*)\right] \;\leq\; \frac{1}{K}\mathbb{E}\left[\sum_{k=1}^{K}\big(F(x_k) - F(x_*)\,\big)\right].
\end{aligned} \tag{14}
$$

□

Lemma 1. *In addition, let us make another assumption, that the expectation of the squared ℓ_2-norm of stochastic gradient is bounded, i.e. $\mathbb{E}\left[\,||g_k||_2^2\,\right] \leq G^2$. Then we shall have,*

$$
\mathbb{E}\left[F(\bar{x}_K) - F(x_*)\right] \;\leq\; \frac{\eta}{2}G^2 + \frac{D^2}{2\eta K}. \tag{15}
$$

(For a concrete understanding of the conditions under which this assumption shall hold, refer to Lemma 2.6 in [7])

Proof. By taking into account the guarantee of the SGD algorithm (as provided by **Theorem 2** in (1)), and the inequality identity that is provided by **Corollary 1** in (12), we can write the following,

$$
\mathbb{E}\left[F(\bar{x}_K) - F(x_*)\right] \;\leq\; \frac{\eta}{2K}\left(\sum_{k=1}^{K}\mathbb{E}\left[\|g_k\|_2^2\right]\right) + \frac{\|x_1 - x_*\|^2}{2\eta K}. \tag{16}
$$

Under the assumption that $\mathbb{E}\left[\,||g_k||_2^2\,\right] \leq G^2$, and using the fact that the initial distance is bounded, i.e., $\|x_1 - x_*\| \leq D$, we can re-write the final inequality expression of (16) as follows,

$$\mathbb{E}\left[F(\bar{x}_K) - F(x_*)\right] \;\le\; \frac{\eta}{2K}\left(\sum_{k=1}^{K}\mathbb{E}\left[\|g_k\|_2^2\right]\right) + \frac{\|x_1 - x_*\|^2}{2\eta K} \tag{17}$$

$$\le\; \frac{\eta}{2}G^2 + \frac{D^2}{2\eta K}.$$

□

Remark 1: The inequality of (17), thus gives an upper-bound for the expectation of the optimality gap, when the function $F(\cdot)$ is evaluated at the point formed by average of the updates $\left(\bar{x}_K = \frac{1}{K}\sum_{k=1}^{K} x_k\right)$.

Remark 2: Knowing that the step-size η is always positive, the upper-bound expression can be thought of as a convex function, $g(\eta) = \left(\frac{\eta}{2}G^2 + \frac{D^2}{2\eta K}\right) : \mathbb{R} \to \mathbb{R}$, since the second-order derivative $g^{"}(\eta) = \frac{D^2}{\eta^3 K}$ is always positive. Furthermore, it can be analytically determined that the minimum value of the upper-bound $g(\cdot)$ is $\frac{DG}{\sqrt{K}}$, which is obtained at $\eta^* = \frac{D}{G\sqrt{K}}$.

4 SGD for Non-Convex Functions:

Theorem 3. *Assume that the variance of the stochastic gradient $\nabla f(x; z)$ is at most σ^2 for all x, i.e., $\mathbb{E}_z\left[\|\nabla f(x; z) - \nabla F(x)\|_2^2\right] \le \sigma^2$. Suppose $F(\cdot)$ is L-smooth. Then, SGD with the step size $\eta \le \frac{1}{L}$ has the following guarantee,*

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

Remark 1: If $\eta = \min\left(\frac{1}{L}, \frac{\sqrt{F(x_1)-F_*}}{\sigma\sqrt{LK}}\right)$, then

$$\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla F(x_k)\|_2^2\right] \le 2\left(F(x_1) - F_*\right)L + 3\sigma\sqrt{\left(F(x_1) - F_*\right)LK}. \tag{18}$$

Remark 2: Furthermore, with $\eta = \min\left(\frac{1}{L}, \frac{\sqrt{F(x_1)-F_*}}{\sigma\sqrt{LK}}\right)$, if $\hat{x}$ is selected uniformly at random from $x_1, \ldots, x_K$, then we have,

$$\mathbb{E}\left[\|\nabla F(\hat{x})\|\right] \leq \frac{\sqrt{2\left(F(x_1) - F_*\right)L}}{\sqrt{K}} + \frac{\sqrt{3\sigma\sqrt{\left(F(x_1) - F_*\right)L}}}{K^{1/4}}. \tag{19}$$

5 Mini-batch SGD for Non-Convex Functions:

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

Below is a formal statement of the Mini-Batch SGD algorithm.

Algorithm 4 MINI-BATCH STOCHASTIC GRADIENT DESCENT(MINI-BATCH SGD)

```
for k = 1 to K do
    for i = 1 to B do
        g_{k,i} = ∇f(x_k; z_{(k-1)B+i})
    end for
    g_k = (1/B) Σ_{i=1}^{B} g_{k,i}
    x_{k+1} = x_k − η g_k
end for
```

Remark: The parameter B is called the *batch size* . When B=1, we have vanilla SGD.

5.1 The variance is $\frac{\sigma^2}{B}$

Lemma 2. *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,*

$$\mathbb{E}_z\left[\|g_k - \nabla F(x_k)\|_2^2\right] \leq \frac{\sigma^2}{B} \tag{20}$$

Proof. Given that $\mathbb{E}_z\left[\|\nabla f(x;z) - \nabla F(x)\|_2^2\right] \leq \sigma^2$. From the mini-batch SGD algorithm, as shown in 4, we know that,

$$g_k = \frac{1}{B}\sum_{i=1}^{B} g_{k,i} = \frac{1}{B}\sum_{i=1}^{B} \nabla f(x_k; z_{(k-1)B+i}) \tag{21}$$

The above formulation shows that, in order to compute g_k at each iteration step, the random variable z is sampled B times.

Let us denote, $z_{k,\,i} := z_{(k-1)B\,+\,i}$. We can make the generic assumption that in an iteration step k, each random variable $z_{k,1}, z_{k,2}, ..., z_{k,B}$ are independent of each other.

So, we have the following,

$$\mathbb{E}_z\left[\|g_k - \nabla F(x_k)\|_2^2\right] = \mathbb{E}_{z_{\,k,1:B}}\left[\left\|\frac{1}{B}\sum_{i=1}^{B}(g_{k,i} - \nabla F(x_k))\right\|_2^2\right] \qquad \text{(Using (21))}$$

$$= \frac{1}{B^2}\mathbb{E}_{z_{\,k,1:B}}\left[\left\|\sum_{i=1}^{B}(g_{k,i} - \nabla F(x_k))\right\|_2^2\right]$$

$$\leq \frac{1}{B^2}\mathbb{E}_{z_{\,k,1:B}}\left[\left(\sum_{i=1}^{B}\left\|g_{k,i} - \nabla F(x_k)\right\|_2\right)^2\right] \qquad \text{(By Triangular Inequality)}$$

$$= \frac{1}{B^2}\mathbb{E}_{z_{\,k,1:B}}\left[\left(\sum_{i=1}^{B}\left\|\nabla f(x_k; z_{k,i}) - \nabla F(x_k)\right\|_2\right)^2\right] \qquad \text{(From 4, } g_{k,i} = \nabla f(x_k; z_{k,i})\text{)}$$

$$= \frac{1}{B^2}\sum_{i=1}^{B}\mathbb{E}_{z_{\,k,i}}\left[\left\|\nabla f(x_k; z_{k,i}) - \nabla F(x_k)\right\|_2^2\right]$$

$$+ \frac{1}{B^2}\sum_{i=1}^{B-1}\sum_{i<j\leq B} 2\mathbb{E}_{z_{\,k,i}\,,\; z_{\,k,j}}\left[\;\left\|\nabla f(x_k; z_{k,i}) - \nabla F(x_k)\right\|_2\;\left\|\nabla f(x_k; z_{k,j}) - \nabla F(x_k)\right\|_2\;\right] \tag{22}$$

Consider the second term on the right-hand side of inequality (22). The term $\left\|\nabla f(x_k; z_{k,i}) - \nabla F(x_k))\right\|_2$ depends on $z_{k,i}$, while the term $\left\|\nabla f(x_k; z_{k,j}) - \nabla F(x_k))\right\|_2$ depends on $z_{k,j}$. When $i \neq j$, since $z_{k,i}$ is independent of $z_{k,j}$, it implies that $\left\|\nabla f(x_k; z_{k,i}) - \nabla F(x_k))\right\|_2$ is independent of $\left\|\nabla f(x_k; z_{k,j}) - \nabla F(x_k))\right\|_2$.

Since covariance of independent terms is 0, therefore $\forall i, j$ with $i \neq j$, we can write that,

$$\mathbb{E}_{z_{k,i}\,,\; z_{k,j}}\left[\left\|\nabla f(x_k; z_{k,i}) - \nabla F(x_k))\right\|_2 \left\|\nabla f(x_k; z_{k,j}) - \nabla F(x_k))\right\|_2\right] = 0.$$

Therefore, the final inequality expression of (22) becomes as follows,

$$\mathbb{E}_z\left[\|g_k - \nabla F(x_k)\|_2^2\right] \;\le\; \frac{1}{B^2}\sum_{i=1}^{B} \mathbb{E}_{z_{k,i}}\left[\left\|\nabla f(x_k; z_{k,i}) - \nabla F(x_k))\right\|_2^2\right]. \tag{23}$$

Now, incorporating our initial assumption that $\mathbb{E}_z\left[\|\nabla f(x; z) - \nabla F(x)\|_2^2\right] \le \sigma^2$ in the above inequality we get,

$$\mathbb{E}_z\left[\|g_k - \nabla F(x_k)\|_2^2\right] \;\le\; \frac{1}{B^2}\sum_{i=1}^{B} \mathbb{E}_{z_{k,i}}\left[\left\|\nabla f(x_k; z_{k,i}) - \nabla F(x_k))\right\|_2^2\right] \le \frac{1}{B^2}\sigma^2 B = \frac{\sigma^2}{B}. \tag{24}$$

Thus, we have shown that: $\mathbb{E}_z\left[\|g_k - \nabla F(x_k)\|_2^2\right] \le \frac{\sigma^2}{B}$.

□

5.2 Iteration complexity of Mini-Batch SGD

Recall **Remark 2** of when SGD is applied to non-convex smooth functions (see 19): with $\eta = \min\left(\frac{1}{L}, \frac{\sqrt{F(x_1) - F_*}}{\sigma\sqrt{LK}}\right)$, if $\hat{x}$ is selected uniformly at random from $x_1, \ldots, x_K$, then we have,

$$\mathbb{E}\left[\|\nabla F(\hat{x})\|\right] \le \frac{\sqrt{2(F(x_1) - F_*)L}}{\sqrt{K}} + \frac{\sqrt{3\sigma(\sqrt{F(x_1) - F_*)L}}}{K^{1/4}} \tag{25}$$

So, if we assume that the variance of the stochastic gradient $\nabla f(x; z)$ is at most σ^2 and set $\eta = \min\left(\frac{1}{L}, \sqrt{\frac{F(x_1) - F^*}{\left(\frac{\sigma}{\sqrt{B}}\right)\sqrt{LK}}}\right)$, then we could obtain the following guarantee for Mini-batch SGD by substituting $\sigma \leftarrow \frac{\sigma}{\sqrt{B}}$ in 25,

$$\mathbb{E}\left[\|\nabla F(\hat{x})\|\right] \le \frac{\sqrt{2(F(x_1) - F_*)L}}{\sqrt{K}} + \frac{\sqrt{3\sigma\sqrt{(F(x_1) - F_*)L}}}{(BK)^{1/4}} \tag{26}$$

5.3 Comparison between SGD and Mini-Batch SGD

	Vanilla SGD	**Mini-batch SGD**
Convergence Rate	$\frac{1}{K^{1/4}}$	$\frac{1}{(BK)^{1/4}}$
number of Stochastic Gradients per Iteration	1	B
total number of stochastic gradients over K	K	BK
Convergence Rate	$\frac{1}{(total\ \#\ of\ sg)^{1/4}}$	$\frac{1}{(total\ \#\ of\ sg)^{1/4}}$

Bibliographic notes

More information on Stochastic Optimization can be found in [6] and [5].

References

[1] G. Lan, The Complexity of Large-scale Convex Programming under a Linear Optimization Oracle, arXiv.org, 2014, doi: 10.48550/arxiv.1309.5550.

[2] J.-K. Wang, J. Abernethy, and K. Y. Levy, No-regret dynamics in the Fenchel game: a unified framework for algorithmic convex optimization, Mathematical programming, vol. 205, no. 12, pp. 203268, 2024, doi: 10.1007/s10107-023-01976-y.

[3] Elad. Hazan, Introduction to Online Convex Optimization, Second Edition., 1st ed. Cambridge: MIT Press, 2022.

[4] C. W. Combettes and S. Pokutta, Complexity of linear minimization and projection on some sets, Operations research letters, vol. 49, no. 4, pp. 565571, 2021, doi: 10.1016/j.orl.2021.06.005.

[5] A. Rakhlin, O. Shamir, and K. Sridharan, Making gradient descent optimal for strongly convex stochastic optimization, arXiv.org, 2012, doi: 10.48550/arXiv.1109.5647.

[6] John Duchi, Introductory Lectures on Stochastic Convex Optimization, Park City Mathematics Institute, Graduate Summer School Lectures, 2016.

[7] Shai Shalev-Shwartz Online learning and online convex optimization, Foundations and Trends® in Machine Learning, 2012.