Chapter 3
Classic: Stochastic Optimization

Abstract In this chapter, we introduce standard stochastic optimization problems and present key stochastic optimization algorithms along with their complexity analysis. While many important stochastic algorithms have been proposed for solving stochastic optimization and empirical risk minimization problems, we focus on seven foundational methods that gained prominence before the deep learning era. These algorithms have had a profound impact on machine learning and provide essential insights for understanding more advanced methods presented in later chapters. The selected algorithms include stochastic gradient descent (SGD), stochastic proximal gradient descent, stochastic mirror descent, adaptive gradient methods, stochastic coordinate descent, stochastic gradient descent ascent, and stochastic optimistic mirror prox. We concentrate on the complexity analysis in the convex setting.

Stochastic optimization is classical wisdom in motion!

Contents

Algorithm 1 `SGD`

1: **Input:** learning rate schedule $\{\eta_t\}_{t=1}^T$, starting point $\mathbf{w}_0$
2: **for** $t = 1, \ldots, T$ **do**
3: Compute an unbiased gradient estimator $\mathbf{z}_t = \nabla g(\mathbf{w}_t; \zeta_t)$
4: Update the model $\mathbf{w}$ by $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{z}_t$
5: **end for**

3.1 Stochastic Gradient Descent

Let us consider the following standard stochastic optimization problem:

$$\min_{\mathbf{w}} g(\mathbf{w}) := \mathbb{E}_\zeta [g(\mathbf{w}; \zeta)]. \tag{3.1}$$

If g is differentiable, the stochastic gradient descent (SGD) method takes the following update:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla g(\mathbf{w}_t; \zeta_t). \tag{3.2}$$

If g is non-differential, we use a stochastic subgradient $\mathcal{G}(\mathbf{w}; \zeta)$ to update the model parameter:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathcal{G}(\mathbf{w}_t; \zeta_t). \tag{3.3}$$

The key assumption regarding the stochastic gradient or subgradient is the following.

Assumption 3.1. *For any* $\mathbf{w}$*, we have* $\mathbb{E}_\zeta [\nabla g(\mathbf{w}; \zeta)] = \nabla g(\mathbf{w})$ *or* $\mathbb{E}_\zeta [\mathcal{G}(\mathbf{w}; \zeta)] \in \partial g(\mathbf{w})$.

Explanation of SGD update

The update (3.2) is equivalent to:

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} g(\mathbf{w}_t; \zeta_t) + \nabla g(\mathbf{w}_t; \zeta_t)^\top (\mathbf{w} - \mathbf{w}_t) + \frac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_t\|_2^2. \tag{3.4}$$

The stochastic linear approximation $\tilde{g}(\mathbf{w}; \zeta_t) = g(\mathbf{w}_t; \zeta_t) + \nabla g(\mathbf{w}_t; \zeta_t)^\top (\mathbf{w} - \mathbf{w}_t)$ serves as a stochastic surrogate for $g(\mathbf{w})$. Since it is only an approximation, we avoid optimizing it directly; instead, we seek a solution close to $\mathbf{w}_t$ that minimizes this surrogate.

When SGD is applied to solving ERM (2.1), ζ_t could represent one randomly sampled data with an index from $\{1, \ldots, n\}$ or a mini-batch of random data.

Below, we present the convergence analysis for smooth and non-smooth, convex and non-convex objective functions.

3.1.1 Smooth Convex Functions

For a point $\mathbf{w}$, convergence is typically measured by the objective gap:

$$g(\mathbf{w}) - \min_{\mathbf{w}} g(\mathbf{w}) = g(\mathbf{w}) - g(\mathbf{w}_*),$$

where $\mathbf{w}_*$ denotes a global optimal solution. A convergence analysis aims to show that after T iterations of updates, we can obtain a solution $\hat{\mathbf{w}}_T$ such that the expected objective gap is bounded by

$$\mathbb{E}\left[g(\hat{\mathbf{w}}_T) - g(\mathbf{w}_*)\right] \leq O\left(\frac{1}{T^\alpha}\right), \tag{3.5}$$

for some $\alpha > 0$. The term $1/T^\alpha$ is referred to as the *convergence rate*. Accordingly, to guarantee a small objective gap $\mathbb{E}[g(\hat{\mathbf{w}}_T) - g(\mathbf{w}_*)] \leq \epsilon$ for some $\epsilon \ll 1$, the bound implies that $T = O\left(\frac{1}{\epsilon^{1/\alpha}}\right)$, which is known as the iteration complexity.

Let us first assume that g is smooth and its stochastic gradient $\nabla g(\mathbf{w}; \zeta)$ satisfies the following assumption.

Assumption 3.2. *(i) $g(\mathbf{w})$ is L-smooth and convex; (ii) For any $\mathbf{w}$, we have*

$$\mathbb{E}[\|\nabla g(\mathbf{w}; \zeta) - \nabla g(\mathbf{w})\|_2^2] \leq \sigma^2$$

for some $\sigma \geq 0$.

The following lemma is useful for convergence analysis.

Lemma 3.1 *Consider the update (3.2). For any $\mathbf{w}$ we have*

$$\nabla g(\mathbf{w}_t; \zeta_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}) \leq \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2.$$

Proof. Since the problem (3.4) is $1/\eta_t$ strongly convex and has an optimal solution $\mathbf{w}_{t+1}$, following (1.18) for any $\mathbf{w}$ we have

$$\begin{aligned}
&\nabla g(\mathbf{w}_t; \zeta_t)^\top (\mathbf{w} - \mathbf{w}_t) + \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|_2^2 \\
&\geq \nabla g(\mathbf{w}_t; \zeta_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 + \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2.
\end{aligned}$$

Re-arranging the inequality, we have

$$\nabla g(\mathbf{w}_t; \zeta_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}) \leq \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2.$$

□

The following lemma shows one-step objective gap bound.

Lemma 3.2 *Suppose Assumption 3.1 and 3.2 hold. For one step SGD update* $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla g(\mathbf{w}_t; \xi_t)$*, we have*

$$\mathbb{E}[g(\mathbf{w}_{t+1}) - g(\mathbf{w}_*)] \leq \mathbb{E}\left[\frac{1}{2\eta_t}\|\mathbf{w}_* - \mathbf{w}_t\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_* - \mathbf{w}_{t+1}\|_2^2\right] + \eta_t \sigma^2.$$

Proof. From Lemma 3.1, we have

$$\begin{aligned}\nabla g(\mathbf{w}_t)^\top(\mathbf{w}_{t+1} - \mathbf{w}) \leq &\frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\ &+ (\nabla g(\mathbf{w}_t) - \nabla g(\mathbf{w}_t; \zeta_t))^\top(\mathbf{w}_{t+1} - \mathbf{w}).\end{aligned} \tag{3.6}$$

By the smoothness and convexity of g, we have

$$\begin{aligned}g(\mathbf{w}_{t+1}) &\leq g(\mathbf{w}_t) + \nabla g(\mathbf{w}_t)^\top(\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\ &\leq g(\mathbf{w}) + \nabla g(\mathbf{w}_t)^\top(\mathbf{w}_t - \mathbf{w}) + \nabla g(\mathbf{w}_t)^\top(\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\ &\leq g(\mathbf{w}) + \nabla g(\mathbf{w}_t)^\top(\mathbf{w}_{t+1} - \mathbf{w}) + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2.\end{aligned} \tag{3.7}$$

Combining this with (3.6), we have

$$\begin{aligned}g(\mathbf{w}_{t+1}) - g(\mathbf{w}) \leq &\frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\ &+ \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 + (\nabla g(\mathbf{w}_t) - \nabla g(\mathbf{w}_t; \zeta_t))^\top(\mathbf{w}_{t+1} - \mathbf{w}).\end{aligned} \tag{3.8}$$

Then if $\eta_t \leq 1/L$ and plugging $\mathbf{w} = \mathbf{w}_*$, we have

$$\begin{aligned}g(\mathbf{w}_{t+1}) - g(\mathbf{w}_*) \leq &\frac{1}{2\eta_t}\|\mathbf{w}_* - \mathbf{w}_t\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_* - \mathbf{w}_{t+1}\|_2^2 \\ &+ (\nabla g(\mathbf{w}_t) - \nabla g(\mathbf{w}_t; \zeta_t))^\top(\mathbf{w}_{t+1} - \mathbf{w}_*).\end{aligned}$$

The challenge lies at handling the last term where $\mathbf{w}_{t+1}$ depends on ζ_t, hence its expectation is not equal to zero. To bound the last term, we introduce

$$\hat{\mathbf{w}}_{t+1} = \arg\min_{\mathbf{w}} \nabla g(\mathbf{w}_t)^\top(\mathbf{w} - \mathbf{w}_t) + \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|_2^2.$$

Note that $\hat{\mathbf{w}}_{t+1}$ is independent of ζ_t. Then $\mathbb{E}_{\zeta_t}[(\nabla g(\mathbf{w}_t) - \nabla g(\mathbf{w}_t; \zeta_t))^\top(\hat{\mathbf{w}}_{t+1} - \mathbf{w}_*)] = 0$. Thus, we have

$$\begin{aligned}\mathbb{E}[g(\mathbf{w}_{t+1}) - g(\mathbf{w}_*)] \leq &\mathbb{E}\left[\frac{1}{2\eta_t}\|\mathbf{w}_* - \mathbf{w}_t\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_* - \mathbf{w}_{t+1}\|_2^2\right] \\ &+ \mathbb{E}[(\nabla g(\mathbf{w}_t) - \nabla g(\mathbf{w}_t; \zeta_t))^\top(\mathbf{w}_{t+1} - \hat{\mathbf{w}}_{t+1})].\end{aligned}$$

Due to Lemma 1.7, we have $\|\mathbf{w}_{t+1} - \hat{\mathbf{w}}_{t+1}\|_2 \leq \eta_t \|\nabla g(\mathbf{w}_t) - \nabla g(\mathbf{w}_t; \zeta_t)\|_2$, thus

$$\mathbb{E}[g(\mathbf{w}_{t+1}) - g(\mathbf{w}_*)] \leq \mathbb{E}\left[\frac{1}{2\eta_t}\|\mathbf{w}_* - \mathbf{w}_t\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_* - \mathbf{w}_{t+1}\|_2^2\right] + \eta_t \sigma^2.$$

□

Theorem 3.1 *Suppose Assumption 3.1 and 3.2 hold. Let the learning rate* $\{\eta_t\}$ *be* $\eta_t = \eta \leq 1/L$ *and* $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_{t+1}$. *Then after* T *iterations of SGD update we have*

$$\mathbb{E}\left[g(\bar{\mathbf{w}}_T) - g(\mathbf{w}_*)\right] \leq \frac{\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2}{2\eta T} + \eta\sigma^2. \tag{3.9}$$

If $\eta = \min(\frac{1}{L}, \frac{\|\mathbf{w}_1 - \mathbf{w}_*\|_2}{\sqrt{2T}\sigma})$, *then*

$$\mathbb{E}\left[g(\bar{\mathbf{w}}_T) - g(\mathbf{w}_*)\right] \leq \frac{\sqrt{2}\sigma\|\mathbf{w}_1 - \mathbf{w}_*\|_2}{\sqrt{T}} + \frac{L\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2}{T}.$$

Why it matters:

In the convergence upper bound (3.9), the first term captures the optimization error due to the finite time horizon, while the second term represents the error induced by stochastic gradient noise.

If $\sigma = 0$ (no noise), SGD reduces to gradient descent, then a constant step size $\eta = 1/L$ can be used and the convergence rate becomes $O\left(\frac{L\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2}{T}\right)$. If $\sigma^2 > 0$ (there is noise in stochastic gradient), in order to guarantee convergence, we have to set $\eta_t \to 0$ or a small value to guarantee certain level of accuracy.

For a fixed number of iterations T, a smaller variance σ allows for faster convergence with a larger learning rate η (up to a certain limit).

The iteration complexity required to achieve $\mathbb{E}[g(\bar{\mathbf{w}}_T) - g(\mathbf{w}_*)] \leq \epsilon$ is

$$T = O\left(\max\left(\frac{\sigma^2\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2}{\epsilon^2}, \frac{L\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2}{\epsilon}\right)\right).$$

If a mini-batch of size B is used to compute the stochastic gradient at each iteration, the variance of the stochastic gradient decreases by a factor of B. This implies that increasing the batch size, up to a certain point, can reduce the number of iterations needed.

Finally, the result also highlights that the initial learning rate η cannot be too large; in practice, an excessively large initial learning rate may cause the algorithm to diverge.

Proof. If $\eta_t = \eta$, summing the inequalities in Lemma 3.2 over $t = 1, \ldots, T$, we have

$$\mathbb{E}\Big[\sum_{t=1}^{T}(g(\mathbf{w}_{t+1})-g(\mathbf{w}_*))\Big] \leq \mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{2\eta}\|\mathbf{w}_*-\mathbf{w}_t\|_2^2-\frac{1}{2\eta}\|\mathbf{w}_*-\mathbf{w}_{t+1}\|_2^2\right] + T\eta\sigma^2.$$

The first term in $[\cdot]$ is a telescoping series,

$$\sum_{t=1}^{T}\frac{1}{2\eta}\|\mathbf{w}_*-\mathbf{w}_t\|_2^2-\frac{1}{2\eta}\|\mathbf{w}_*-\mathbf{w}_{t+1}\|_2^2 \leq \frac{1}{2\eta}\|\mathbf{w}_*-\mathbf{w}_1\|_2^2-\frac{1}{2\eta}\|\mathbf{w}_*-\mathbf{w}_{T+1}\|_2^2 \leq \frac{1}{2\eta}\|\mathbf{w}_*-\mathbf{w}_1\|_2^2.$$

As a result,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(g(\mathbf{w}_{t+1})-g(\mathbf{w}_*))\right] \leq \frac{1}{2\eta T}\|\mathbf{w}_*-\mathbf{w}_1\|_2^2+\eta\sigma^2,$$

which concludes the proof of the first part of the theorem.

For the second part, optimizing the upper bound over η gives $\eta_* = \frac{\|\mathbf{w}_1-\mathbf{w}_*\|}{\sqrt{2T}\sigma}$. If $\eta_* \leq 1/L$, i.e., $T \geq \frac{\|\mathbf{w}_1-\mathbf{w}_*\|_2^2L^2}{2\sigma^2}$, we set $\eta=\eta_*$, then

$$\mathbb{E}\left[g(\bar{\mathbf{w}}_T)-g(\mathbf{w}_*)\right] \leq \frac{2\sigma\|\mathbf{w}_1-\mathbf{w}_*\|_2}{\sqrt{2T}}.$$

If $\eta_* > 1/L$, i.e., $\sigma^2 \leq \frac{\|\mathbf{w}_1-\mathbf{w}_*\|_2^2L^2}{2T}$, we set $\eta = 1/L$, then

$$\mathbb{E}\left[g(\bar{\mathbf{w}}_T)-g(\mathbf{w}_*)\right] \leq \frac{L\|\mathbf{w}_1-\mathbf{w}_*\|_2^2}{2T}+\frac{L\|\mathbf{w}_1-\mathbf{w}_*\|_2^2}{2T} = \frac{L\|\mathbf{w}_1-\mathbf{w}_*\|_2^2}{T}.$$

□

3.1.2 Non-smooth Convex Functions

Now, let us consider the SGD update (3.3) for non-smooth convex functions under the following assumption.

Assumption 3.3. *(i) $g(\mathbf{w})$ is convex; (ii) For any $\mathbf{w}$, we have $\mathbb{E}[\|\mathcal{G}(\mathbf{w};\zeta)\|_2^2] \leq G^2$.*

Lemma 3.3 *Suppose Assumption 3.1 and 3.3 hold. For one step SGD update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\mathcal{G}(\mathbf{w}_t;\xi_t)$, we have*

$$\mathbb{E}[g(\mathbf{w}_t)-g(\mathbf{w}_*)] \leq \mathbb{E}\left[\frac{1}{2\eta_t}\|\mathbf{w}_*-\mathbf{w}_t\|_2^2-\frac{1}{2\eta_t}\|\mathbf{w}-\mathbf{w}_{t+1}\|_2^2\right]+\frac{\eta_t}{2}G^2.$$

Proof. From Lemma 3.1, we have

$$\begin{aligned}\mathcal{G}(\mathbf{w}_t;\zeta_t)^\top(\mathbf{w}_t-\mathbf{w}_*) &\le \frac{1}{2\eta_t}\|\mathbf{w}_*-\mathbf{w}_t\|_2^2-\frac{1}{2\eta_t}\|\mathbf{w}_*-\mathbf{w}_{t+1}\|_2^2-\frac{1}{2\eta_t}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2\\ &\quad+\mathcal{G}(\mathbf{w}_t;\zeta_t)^\top(\mathbf{w}_{t+1}-\mathbf{w}_t)\\ &\le \frac{1}{2\eta_t}\|\mathbf{w}_*-\mathbf{w}_t\|_2^2-\frac{1}{2\eta_t}\|\mathbf{w}_*-\mathbf{w}_{t+1}\|_2^2-\frac{1}{2\eta_t}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2\\ &\quad+\frac{\eta_t}{2}\|\mathcal{G}(\mathbf{w}_t;\zeta_t)\|_2^2+\frac{1}{2\eta_t}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2,\end{aligned} \tag{3.10}$$

where the last inequality uses the Young's inequality. Taking expectation on both sides, we have

$$\mathbb{E}[\mathcal{G}(\mathbf{w}_t;\zeta_t)^\top(\mathbf{w}_t-\mathbf{w}_*)] \le \mathbb{E}\left[\frac{1}{2\eta_t}\|\mathbf{w}_*-\mathbf{w}_t\|_2^2-\frac{1}{2\eta_t}\|\mathbf{w}-\mathbf{w}_{t+1}\|_2^2\right]+\frac{\eta_t}{2}G^2. \tag{3.11}$$

Since $\mathbf{w}_t$ is independent of ζ_t, we have $\mathbb{E}[\mathcal{G}(\mathbf{w}_t;\zeta_t)^\top(\mathbf{w}_t-\mathbf{w}_*)] = \mathbb{E}[\mathbf{v}_t^\top(\mathbf{w}_t-\mathbf{w}_*)]$ for some $\mathbf{v}_t \in \partial g(\mathbf{w}_t)$. By the convexity of g, we have

$$\begin{aligned}&\mathbb{E}[g(\mathbf{w}_t)-g(\mathbf{w}_*)] \le \mathbb{E}[\mathbf{v}_t^\top(\mathbf{w}_t-\mathbf{w}_*)] = \mathbb{E}[\mathcal{G}(\mathbf{w}_t;\zeta_t)^\top(\mathbf{w}_t-\mathbf{w}_*)]\\ &\le \mathbb{E}\left[\frac{1}{2\eta_t}\|\mathbf{w}_*-\mathbf{w}_t\|_2^2-\frac{1}{2\eta_t}\|\mathbf{w}-\mathbf{w}_{t+1}\|_2^2\right]+\frac{\eta_t}{2}G^2.\end{aligned} \tag{3.12}$$

□

The theorem below establishes the convergence of SGD for non-smooth convex functions as measured by the objective gap.

Theorem 3.2 *Suppose Assumption 3.1 and 3.3 hold. Let the learning rate $\{\eta_t\}$ be $\eta_t = \eta$ and $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t$. Then after for T iterations of SGD update (3.3) we have*

$$\mathbb{E}\left[g(\bar{\mathbf{w}}_T)-g(\mathbf{w}_*)\right] \le \frac{\|\mathbf{w}_1-\mathbf{w}_*\|_2^2}{2\eta T}+\frac{\eta G^2}{2}.$$

If $\eta = \frac{\|\mathbf{w}_1-\mathbf{w}_*\|_2}{\sqrt{T}G}$, then

$$\mathbb{E}\left[g(\bar{\mathbf{w}}_T)-g(\mathbf{w}_*)\right] \le \frac{G\|\mathbf{w}_1-\mathbf{w}_*\|_2}{\sqrt{T}}.$$

Why it matters:

The above theorem exhibits the key difference in the convergence of SGD for smooth convex functions and non-smooth convex functions. Even with a zero variance for the stochastic subgradient , the convergence rate is still $O(1/\sqrt{T})$. The reason is that for smooth convex functions when $g(\mathbf{w}) \to g(\mathbf{w}_*)$, we have

$\nabla g(\mathbf{w}) \to 0$ (cf. Lemma 1.5(b)), which is not true for non-smooth convex functions.

Proof. The proof is similar to that in the smooth case. □

3.1.3 Smooth Non-Convex Functions

For a non-convex function, it is generally NP-hard to find a global optimal solution. Hence, our goal here is to establish the complexity of SGD for finding an ϵ-stationary solution with $\epsilon \ll 1$, as defined below.

Definition 3.1 (ϵ-stationary solution) $\mathbf{w}$ is an ϵ-stationary solution to $\min_{\mathbf{w}} g(\mathbf{w})$, if $\|\nabla g(\mathbf{w})\|_2 \leq \epsilon$.

Assumption 3.4. *(i) $g(\mathbf{w})$ is L-smooth and non-convex; (ii) For any $\mathbf{w}$, we have*

$$\mathbb{E}[\|\nabla g(\mathbf{w}; \zeta) - \nabla g(\mathbf{w})\|_2^2] \leq \sigma^2$$

for some $\sigma \geq 0$.

Based on the above assumptions, we establish the following convergence guarantee.

Theorem 3.3 *Suppose Assumption 3.1 and 3.4 hold. Let the learning rate $\{\eta_t\}$ be $\eta_t = \min\{\frac{1}{L}, \frac{D}{\sigma\sqrt{T}}\}$ for some constant $D > 0$. Let $\tau \in \{1, \ldots, T\}$ be a random sample following a distribution $\Pr(\tau = t) = \frac{1}{T}$. Then we have*

$$\mathbb{E}[\|\nabla g(\mathbf{w}_\tau)\|_2^2] \leq \frac{2L(g(\mathbf{w}_1) - g(\mathbf{w}_*)}{T} + \left(\frac{2(g(\mathbf{w}_1) - g(\mathbf{w}_*))}{D} + DL\right)\frac{\sigma}{\sqrt{T}}.$$

Proof. For brevity of notation, we let $\nabla g_t(\mathbf{w}_t) = \nabla g(\mathbf{w}_t; \zeta_t)$. Due to the L-smoothness of g, we have

$$
\begin{aligned}
&g(\mathbf{w}_{t+1}) \leq g(\mathbf{w}_t) + \nabla g(\mathbf{w}_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{L}{2} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&= g(\mathbf{w}_t) - \eta_t \nabla g(\mathbf{w}_t)^\top \nabla g_t(\mathbf{w}_t) + \frac{\eta_t^2 L}{2} \|\nabla g_t(\mathbf{w}_t)\|_2^2 \\
&= g(\mathbf{w}_t) - \eta_t \|\nabla g(\mathbf{w}_t)\|_2^2 + \eta_t \nabla g(\mathbf{w}_t)^\top (\nabla g(\mathbf{w}_t) - \nabla g_t(\mathbf{w}_t)) + \frac{\eta_t^2 L}{2} \|\nabla g_t(\mathbf{w}_t)\|_2^2 \\
&= g(\mathbf{w}_t) - \eta_t \|\nabla g(\mathbf{w}_t)\|_2^2 + \eta_t \nabla g(\mathbf{w}_t)^\top (\nabla g(\mathbf{w}_t) - \nabla g_t(\mathbf{w}_t)) \\
&+ \frac{\eta_t^2 L}{2} \|\nabla g_t(\mathbf{w}_t) - \nabla g(\mathbf{w}_t) + \nabla g(\mathbf{w}_t)\|_2^2 \\
&= g(\mathbf{w}_t) - (\eta_t - \frac{\eta_t^2 L}{2}) \|\nabla g(\mathbf{w}_t)\|_2^2 + (\eta_t - \eta_t^2 L) \nabla g(\mathbf{w}_t)^\top (\nabla g(\mathbf{w}_t) - \nabla g_t(\mathbf{w}_t)) \\
&+ \frac{\eta_t^2 L}{2} \|\nabla g_t(\mathbf{w}_t) - \nabla g(\mathbf{w}_t)\|_2^2 .
\end{aligned}
$$

Taking expectation over ζ_t given $\mathbf{w}_t$ on both sides, we have

$$
\mathbb{E}_t \left[g(\mathbf{w}_{t+1}) \right] \leq g(\mathbf{w}_t) - (\eta_t - \frac{\eta_t^2 L}{2}) \|\nabla g(\mathbf{w}_t)\|_2^2 + \frac{\eta_t^2 L}{2} \sigma^2. \tag{3.13}
$$

Telescoping this from $t = 1$ to T gives

$$
\mathbb{E} \left[\sum_{t=1}^{T} (\eta_t - \frac{\eta_t^2 L}{2}) \|\nabla g(\mathbf{w}_t)\|_2^2 \right] \leq (g(\mathbf{w}_1) - g(\mathbf{w}_*) + \sum_{t=1}^{T} \frac{\eta_t^2 L}{2} \sigma^2.
$$

As a result,

$$
\mathbb{E} \left[\|\nabla g(\mathbf{w}_\tau)\|_2^2 \right] \leq \frac{(g(\mathbf{w}_1) - g(\mathbf{w}_*))}{\sum_{t=1}^{T} (\eta_t - \frac{\eta_t^2 L}{2})} + \frac{\sum_{t=1}^{T} \eta_t^2 L}{2 \sum_{t=1}^{T} (\eta_t - \frac{\eta_t^2 L}{2})} \sigma^2.
$$

Plugging the value of $\eta_t = \min(\frac{1}{L}, \frac{D}{\sigma\sqrt{T}})$, we have

$$
\begin{aligned}
\mathbb{E} \left[\|\nabla g(\mathbf{w}_\tau)\|_2^2 \right] &\leq \frac{(g(\mathbf{w}_1) - g(\mathbf{w}_*))}{T(\eta_1 - \frac{\eta_1^2 L}{2})} + \frac{T \eta_1^2 L}{2T(\eta_1 - \frac{\eta_1^2 L}{2})} \sigma^2 \\
&\leq \frac{2(g(\mathbf{w}_1) - g(\mathbf{w}_*))}{T\eta_1} + \eta_1 L \sigma^2 \\
&\leq \max \left(\frac{2L(g(\mathbf{w}_1) - g(\mathbf{w}_*))}{T}, \frac{2(g(\mathbf{w}_1) - g(\mathbf{w}_*))\sigma}{D\sqrt{T}} \right) + \frac{D\sigma L}{\sqrt{T}} \\
&\leq \frac{2L(g(\mathbf{w}_1) - g(\mathbf{w}_*))}{T} + \left(\frac{2(g(\mathbf{w}_1) - g(\mathbf{w}_*))}{D} + DL \right) \frac{\sigma}{\sqrt{T}}.
\end{aligned}
$$

If we set $\eta_t = \min(\frac{1}{L}, \frac{D}{\sigma\sqrt{t}})$, then $\sum_{t=1}^{T} \eta_t \geq \Omega(\sqrt{T})$ and $\sum_{t=1}^{T} \eta_t^2 \leq O(\log(T))$, then $\mathbb{E} \left[\|\nabla g(\mathbf{w}_\tau)\|_2^2 \right] \leq O(\log T / T)$. □

3.1.4 Non-smooth Weakly Convex Functions

Next, let us extend the analysis to non-smooth non-convex functions. Consider a function $g : \mathbb{R}^d \mapsto \mathbb{R}$ and a point $\mathbf{w} \in \mathbb{R}^d$ with $g(\mathbf{w})$ finite. The Fréchet subdifferential of g at $\mathbf{w}$, denoted $\partial g(\mathbf{w})$, consists of all vectors $\mathbf{v}$ satisfying

$$g(\mathbf{w}) \geq g(\mathbf{w}') + \mathbf{v}^\top(\mathbf{w} - \mathbf{w}') + o(\|\mathbf{w} - \mathbf{w}'\|_2) \text{ as } \mathbf{w}' \to \mathbf{w}.$$

We consider a family of non-convex functions, namely weakly convex functions. A lower semi-continuous function g is called ρ-weakly, if there exists $\rho > 0$ such that:

$$g(\mathbf{w}) \geq g(\mathbf{w}') + \mathbf{v}^\top(\mathbf{w} - \mathbf{w}') - \frac{\rho}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2, \quad \forall \mathbf{w}, \mathbf{w}', \mathbf{v} \in \partial g(\mathbf{w}').$$

It is easy to show that if g is ρ-weakly convex, then $g(\mathbf{w}) + \frac{\rho}{2}\|\mathbf{w}\|_2^2$ is a convex function of $\mathbf{w}$. A smooth function is weakly convex, but the reverse is not necessarily true.

Example

Example 3.1 (Compositional functions). *Let $F(\mathbf{x}) = f(g(\mathbf{x}))$. If f convex and G_1-Lipschitz continuous and $g(\mathbf{x})$ is L_2-smooth, then F is ρ-weakly convex for some $\rho > 0$. We will prove this in Section 5.3. The OCE risk (2.22) is a special case when ϕ^* is non-smooth and the loss function $\ell(\mathbf{w}; \mathbf{z})$ is smooth non-convex.*

Example 3.2 (Compositional functions). *Let $F(\mathbf{x}) = f(g(\mathbf{x}))$. If f L_1-smooth and monotonically non-decreasing and $g(\mathbf{x})$ is non-smooth convex and G_2-Lipschitz continuous, then F is ρ-weakly convex for some $\rho > 0$.*
Let us prove it. Since $f(g)$ is L_1 smooth, i.e., for any $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d$, we have $f(g(\mathbf{v})) + f'(g(\mathbf{v}))(g(\mathbf{w}) - g(\mathbf{v})) - \frac{L_1}{2}|g(\mathbf{w}) - g(\mathbf{v})|^2 \leq f(g(\mathbf{w}))$. Since g is convex, i.e. for any $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d$, $g(\mathbf{w}) \geq g(\mathbf{v}) + \partial g(\mathbf{v})^\top(\mathbf{w} - \mathbf{v})$, then

$$\begin{aligned} f(g(\mathbf{w})) - f(g(\mathbf{v})) \geq & f'(g(\mathbf{v}))\partial g(\mathbf{v})^\top(\mathbf{w} - \mathbf{v}) - \frac{L_1}{2}|g(\mathbf{w}) - g(\mathbf{v})|^2 \\ \geq & f'(g(\mathbf{v}))\partial g(\mathbf{v})^\top(\mathbf{w} - \mathbf{v}) - \frac{G_2^2 L_1}{2}\|\mathbf{w} - \mathbf{v}\|_2^2, \end{aligned}$$

where the first inequality uses $f'(g(\mathbf{w})) \geq 0$; the second inequality uses the fact that $\|\partial g(\mathbf{w})\|_2 \leq G_2$. That is, $f(g(\mathbf{w}))$ is G^2L-weakly convex.
An important application of this function in machine learning is optimizing the truncation of a convex loss $g(\mathbf{w}) = \ell(\mathbf{w}; \mathbf{z}) \geq 0$ with a smooth truncation function $f(\ell(\mathbf{w}; \mathbf{z})) = \alpha \log(1 + \frac{\ell(\mathbf{w};\mathbf{z})}{\alpha})$ for some $\alpha > 0$, which is useful for tackling heavy-tailed data distribution.

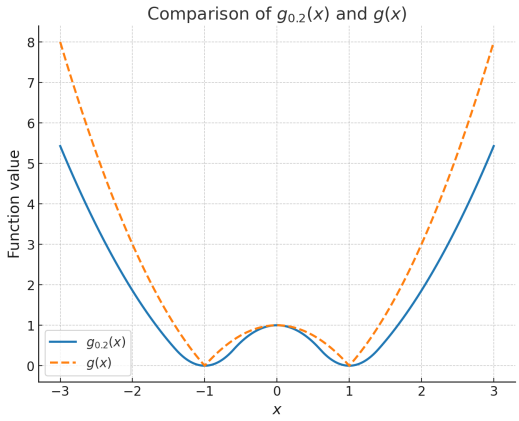

Fig. 3.1: Moreau envelope of $g(x) = |x^2 - 1|$ with $\lambda = 0.2$.

Nearly ϵ-stationary solution

When $g(\cdot)$ is non-smooth, finding an ϵ-stationary solution such that $\|\nabla g(\mathbf{w})\|_2 \leq \epsilon$ is difficult even for a convex function. Let us consider a simple example $\min_w |w|$. The only stationary point is the optimal solution $w_* = 0$, and any $w \neq 0$ is not an ϵ-stationary solution ($\epsilon < 1$) no matter how close w to 0. To address this issue, we introduce a weak notion of ϵ-stationary solution, termed nearly ϵ-stationary solution.

Definition 3.2 (Nearly ϵ-stationary solution) $\mathbf{w}$ is a nearly ϵ-stationary solution to $\min_{\mathbf{w}} g(\mathbf{w})$, if there exists $\hat{\mathbf{w}}$ such that $\|\mathbf{w} - \hat{\mathbf{w}}\| \leq O(\epsilon)$ and $\text{dist}(0, \partial g(\hat{\mathbf{w}})) \leq \epsilon$.

A useful tool for deriving a nearly ϵ-stationary solution is the Moreau envelope of g:

$$g_\lambda(\mathbf{w}) := \min_{\mathbf{u}} g(\mathbf{u}) + \frac{1}{2\lambda}\|\mathbf{u} - \mathbf{w}\|_2^2, \tag{3.14}$$

$$\text{prox}_{\lambda g}(\mathbf{w}) := \arg\min_{\mathbf{u}} g(\mathbf{u}) + \frac{1}{2\lambda}\|\mathbf{u} - \mathbf{w}\|_2^2. \tag{3.15}$$

When λ is sufficiently small, $g_\lambda(\cdot)$ is a smooth function. An example of a weakly convex function and its Moreau envelope is illustrated in Figure 3.1.

Proposition 3.1. *Consider a ρ-weakly convex function $g(\cdot)$. Then for any $\lambda \in (0, \rho^{-1})$, the Moreau envelope $g_\lambda(\cdot)$ is $\frac{2-\lambda\rho}{\lambda(1-\lambda\rho)}$-smooth, with gradient given by*

$$\nabla g_\lambda(\mathbf{w}) = \frac{1}{\lambda}(\mathbf{w} - prox_{\lambda g}(\mathbf{w})).$$

Proof. First, when $\lambda < \rho^{-1}$ we have $g(\mathbf{u}) + \frac{1}{2\lambda}\|\mathbf{u} - \mathbf{w}\|_2^2$ become $(\frac{1}{\lambda} - \rho)$ -strongly convex. Hence the solution $\text{prox}_{\lambda g}(\mathbf{w})$ is unique for a given $\mathbf{w}$. We can also write $\text{prox}_{\lambda g}(\mathbf{w})$ as

$$\begin{aligned}\text{prox}_{\lambda g}(\mathbf{w}) &:= \arg\min_{\mathbf{u}} g(\mathbf{u}) + \frac{1}{2\lambda}\|\mathbf{u}-\mathbf{w}\|_2^2 \\ &= \arg\min_{\mathbf{u}} \underbrace{g(\mathbf{u}) + \frac{\rho}{2}\|\mathbf{u}\|_2^2}_{r(\mathbf{u})} + \frac{1}{2}\left(\frac{1}{\lambda}-\rho\right)\left\|\mathbf{u} - \frac{1}{1-\lambda\rho}\mathbf{w}\right\|_2^2.\end{aligned}$$

Due to Lemma 1.7, we have $\|\text{prox}_{\lambda g}(\mathbf{w}) - \text{prox}_{\lambda g}(\mathbf{w}')\|_2 \le \frac{1}{1-\lambda\rho}\|\mathbf{w}-\mathbf{w}'\|_2$. Then

$$\begin{aligned}&\|\nabla g_\lambda(\mathbf{w}) - \nabla g_\lambda(\mathbf{w}')\|_2 = \frac{1}{\lambda}\|(\mathbf{w} - \text{prox}_{\lambda g}(\mathbf{w})) - (\mathbf{w}' - \text{prox}_{\lambda g}(\mathbf{w}'))\|_2 \\ &\le \frac{1}{\lambda}\left(\|\mathbf{w}-\mathbf{w}'\|_2 + \frac{1}{1-\lambda\rho}\|\mathbf{w}-\mathbf{w}'\|_2\right) = \frac{2-\lambda\rho}{\lambda(1-\lambda\rho)}\|\mathbf{w}-\mathbf{w}'\|_2.\end{aligned}$$

□

With the Moreau envelope, we can use the norm of its gradient to measure the convergence for optimizing the original function.

Proposition 3.2. *If $\lambda < \rho^{-1}$, we have*

$$g_\lambda(\mathbf{w}) \le g(\mathbf{w}), \quad \min_{\mathbf{w}} g_\lambda(\mathbf{w}) = \min_{\mathbf{w}} g(\mathbf{w}). \tag{3.16}$$

If $\|\nabla g_\lambda(\mathbf{w})\|_2 \le \epsilon$, then $\hat{\mathbf{w}} = prox_{\lambda g}(\mathbf{w})$ is a nearly ϵ-stationary solution. In particular,

$$\begin{aligned}&\|\hat{\mathbf{w}} - \mathbf{w}\|_2 = \lambda\|\nabla g_\lambda(\mathbf{w})\|_2 \le \lambda\epsilon, \\ &dist(0, \partial g(\hat{\mathbf{w}})) \le \|\nabla g_\lambda(\mathbf{w})\|_2 \le \epsilon.\end{aligned} \tag{3.17}$$

Proof. $g_\lambda(\mathbf{w}) \le g(\mathbf{w})$ follows the definition of $g_\lambda(\mathbf{w})$. Then $g_\lambda(\mathbf{w}_*) \le g(\mathbf{w}_*)$. To prove they are equal, we have

$$g_\lambda(\mathbf{w}) = \min_{\mathbf{u}} g(\mathbf{u}) + \frac{1}{2\lambda}\|\mathbf{u}-\mathbf{w}\|_2^2 \ge \min_{\mathbf{u}} g(\mathbf{u}) = g(\mathbf{w}_*).$$

Since $\nabla g_\lambda(\mathbf{w}) = \frac{1}{\lambda}(\mathbf{w}-\hat{\mathbf{w}})$, which implies the second inequality. The first inequality is due to the first-order optimality condition of $\min_{\mathbf{u}} g(\mathbf{u}) + \frac{1}{2\lambda}\|\mathbf{u}-\mathbf{w}\|_2^2$. □

Why it matters:

Proposition 3.2 shows that if we can make $\|\nabla g_\lambda(\mathbf{w})\|_2$ small, then $\mathbf{w}$ is close to an ϵ-stationary solution $\hat{\mathbf{w}}$ of the original function $g(\mathbf{w})$. The smaller the λ, the closer between $\mathbf{w}$ and $\hat{\mathbf{w}}$.

Convergence Analysis

Assumption 3.5. *(i) $g(\mathbf{w})$ is ρ-weakly convex; (ii) For any $\mathbf{w}$, $\mathbb{E}_\zeta[\|\mathcal{G}(\mathbf{w},\zeta)\|_2^2] \le G^2$ for some $G \ge 0$.*

Lemma 3.4 *Let us consider an update* $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{z}_t$. *If* $\mathbb{E}_t[\mathbf{z}_t] = \mathcal{M}_t$ *and* $\mathbb{E}_t[\|\mathbf{z}_t\|_2^2] \leq G^2$, *then we have*

$$\mathbb{E}_t[g_\lambda(\mathbf{w}_{t+1})] \leq g_\lambda(\mathbf{w}_t) + \frac{\eta_t}{\lambda}(\hat{\mathbf{w}}_t - \mathbf{w}_t)^\top \mathcal{M}_t + \frac{\eta_t^2 G^2}{2\lambda},$$

where $\hat{\mathbf{w}}_t = prox_{\lambda g}(\mathbf{w}_t)$.

Proof. We have

$$\begin{aligned} g_\lambda(\mathbf{w}_{t+1}) &= g(\hat{\mathbf{w}}_{t+1}) + \frac{1}{2\lambda}\|\hat{\mathbf{w}}_{t+1} - \mathbf{w}_{t+1}\|_2^2 \leq g(\hat{\mathbf{w}}_t) + \frac{1}{2\lambda}\|\hat{\mathbf{w}}_t - \mathbf{w}_{t+1}\|_2^2 \\ &= g(\hat{\mathbf{w}}_t) + \frac{1}{2\lambda}\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|_2^2 - \frac{1}{2\lambda}\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|_2^2 + \frac{1}{2\lambda}\|\hat{\mathbf{w}}_t - \mathbf{w}_{t+1}\|_2^2. \end{aligned}$$

Merging the first two terms we get $g_\lambda(\mathbf{w}_t)$, and using the three-point equality $2(a-b)(b-c) = \|a-c\|_2^2 - \|a-b\|_2^2 - \|b-c\|_2^2$ to merge the last two terms we get

$$\begin{aligned} g_\lambda(\mathbf{w}_{t+1}) &= g_\lambda(\mathbf{w}_t) + \frac{1}{\lambda}(\hat{\mathbf{w}}_t - \mathbf{w}_t)^\top(\mathbf{w}_t - \mathbf{w}_{t+1}) + \frac{1}{2\lambda}\|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2^2 \\ &= g_\lambda(\mathbf{w}_t) + \frac{1}{\lambda}(\hat{\mathbf{w}}_t - \mathbf{w}_t)^\top \eta_t \mathbf{z}_t + \frac{\eta_t^2}{2\lambda}\|\mathbf{z}_t\|_2^2. \end{aligned}$$

Taking expectation over ζ_t given $\mathbf{w}_t$ on both sides, we have

$$\mathbb{E}_t[g_\lambda(\mathbf{w}_{t+1})] \leq g_\lambda(\mathbf{w}_t) + \frac{1}{\lambda}(\hat{\mathbf{w}}_t - \mathbf{w}_t)^\top \eta_t \mathcal{M}_t + \frac{\eta_t^2 G^2}{2\lambda}.$$

□

Lemma 3.5 *Under the same setting of Lemma 3.4 we have*

$$\eta_t(1 - \lambda\rho)\|\nabla g_\lambda(\mathbf{w}_t)\|_2^2 \leq g_\lambda(\mathbf{w}_t) - \mathbb{E}_t[g_\lambda(\mathbf{w}_{t+1})] + \frac{\eta_t^2 G^2}{2\lambda}.$$

Proof. Due to the weak convexity of g, for any $\mathcal{M}_t \in \partial g(\mathbf{w}_t)$, we have

$$\begin{aligned} &\mathcal{M}_t^\top(\mathbf{w}_t - \hat{\mathbf{w}}_t) \geq g(\mathbf{w}_t) - g(\hat{\mathbf{w}}_t) - \frac{\rho}{2}\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|_2^2 \\ &= (g(\mathbf{w}_t) + \frac{1}{2\lambda}\|\mathbf{w}_t - \mathbf{w}_t\|_2^2) - (g(\hat{\mathbf{w}}_t) + \frac{1}{2\lambda}\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|_2^2) + (\frac{1}{2\lambda} - \frac{\rho}{2})\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|_2^2. \end{aligned}$$

Since $h(\mathbf{w}) = g(\mathbf{w}) + \frac{1}{2\lambda}\|\mathbf{w} - \mathbf{w}_t\|_2^2$ is $(1/\lambda - \rho)$-strongly convex and $\hat{\mathbf{w}}_t = \arg\min h(\mathbf{w})$, then applying Lemma 1.6(a), we get

$$(g(\mathbf{w}_t) + \frac{1}{2\lambda}\|\mathbf{w}_t - \mathbf{w}_t\|_2^2) - (g(\hat{\mathbf{w}}_t) + \frac{1}{2\lambda}\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|_2^2) \geq (\frac{1}{2\lambda} - \frac{\rho}{2})\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|_2^2.$$

Combining the above two inequalities we have

$$\mathcal{M}_t^\top(\mathbf{w}_t - \hat{\mathbf{w}}_t) \geq g(\mathbf{w}_t) - g(\hat{\mathbf{w}}_t) - \frac{\rho}{2}\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|_2^2$$
$$\geq (\frac{1}{2\lambda} - \frac{\rho}{2})\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|_2^2 + (\frac{1}{2\lambda} - \frac{\rho}{2})\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|_2^2 = (\lambda - \lambda^2\rho)\|\nabla g_\lambda(\mathbf{w}_t)\|_2^2.$$

Plugging this into the inequality in Lemma 3.4, we have

$$\eta_t(1 - \lambda\rho)\|\nabla g_\lambda(\mathbf{w}_t)\|_2^2 \leq g_\lambda(\mathbf{w}_t) - \mathbb{E}_t[g_\lambda(\mathbf{w}_{t+1})] + \frac{\eta_t^2 G^2}{2\lambda}.$$

□

Theorem 3.4 *Suppose the learning rate* $\{\eta_t\}$ *is set to* $\eta_t = \frac{C}{\sqrt{T}}$. *Let* $\tau \in \{1, \ldots, T\}$ *be a random sample following a distribution* $\Pr(\tau = t) = \frac{1}{T}$. *Then for any* $\lambda \in (0, \rho^{-1})$, *we have*

$$\mathbb{E}[\|\nabla g_\lambda(\mathbf{w}_\tau)\|_2^2] \leq \frac{g(\mathbf{w}_1) - g(\mathbf{w}_*)}{(1 - \lambda\rho)C\sqrt{T}} + \frac{CG^2}{2\lambda(1 - \lambda\rho)\sqrt{T}}.$$

Proof. Summing up the inequalities in Lemma 3.5 over $t = 1, \ldots, T$ and taking expectation over all randomness, we have

$$\mathbb{E}\left[\sum_{t=1}^T \eta_t(1 - \lambda\rho)\|\nabla g_\lambda(\mathbf{w}_t)\|_2^2\right] \leq g(\mathbf{w}_1) - g(\mathbf{w}_*) + \sum_{t=1}^T \frac{\eta_t^2 G^2}{2\lambda}.$$

where we have used $g_\lambda(\mathbf{w}) \leq g(\mathbf{w})$ and $\min g_\lambda(\mathbf{w}) = g(\mathbf{w}_*)$. Then

$$\mathbb{E}[\|\nabla g_\lambda(\mathbf{w}_\tau)\|_2^2] \leq \frac{\mathbb{E}[g(\mathbf{w}_1) - g(\mathbf{w}_*)]}{(1 - \lambda\rho)C\sqrt{T}} + \frac{CG^2}{2\lambda(1 - \lambda\rho)\sqrt{T}}.$$

□

3.2 Stochastic Proximal Gradient Descent

Let us consider the following stochastic composite optimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) := \mathbb{E}_\zeta[g(\mathbf{w}; \zeta)] + r(\mathbf{w}), \tag{3.18}$$

where $g(\mathbf{w}) = \mathbb{E}_\zeta[g(\mathbf{w}; \zeta)]$ is a smooth function and $r(\mathbf{w})$ is a possibly non-smooth function. In machine learning, r usually corresponds to some regularizer on the model parameter. We make the following assumption.

Assumption 3.6. *Suppose the following conditions hold:*

(i) $g(\mathbf{w})$ *is* L*-smooth and convex, and* $r(\mathbf{w})$ *is convex.*
(ii) There exists $\sigma > 0$ *such that* $\mathbb{E}_\zeta[\|\nabla g(\mathbf{w}; \zeta) - \nabla g(\mathbf{w})\|_2^2] \leq \sigma^2$ *for all* $\mathbf{w}$.

Algorithm 2 SPGD

Input: learning rate schedule $\{\eta_t\}_{t=1}^T$, starting point $\mathbf{w}_1$
for $t = 1, \ldots, T$ **do**
  Compute an unbiased gradient estimator $\mathbf{z}_t = \nabla g(\mathbf{w}_t; \zeta_t)$
  Update the model $\mathbf{w}$ by $\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \mathbf{z}_t^\top(\mathbf{w} - \mathbf{w}_t) + \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|_2^2 + r(\mathbf{w})$.
end for

If the regularizer r is non-smooth, the overall objective function is also non-smooth. Consequently, applying SGD directly cannot exploit the smoothness of g, which would otherwise enable faster convergence and reduce the variance of its stochastic gradients.

To address this challenge, we can employ the stochastic proximal gradient descent (SPGD) method:

$$\begin{aligned}\mathbf{w}_{t+1} &= \arg\min_{\mathbf{w}\in\mathbb{R}^d} \nabla g(\mathbf{w}_t; \zeta_t)^\top(\mathbf{w} - \mathbf{w}_t) + r(\mathbf{w}) + \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|_2^2 \\ &= \arg\min_{\mathbf{w}\in\mathbb{R}^d} r(\mathbf{w}) + \frac{1}{2\eta_t}\|\mathbf{w} - (\mathbf{w}_t - \eta_t \nabla g(\mathbf{w}_t; \zeta_t))\|_2^2.\end{aligned} \tag{3.19}$$

This is also known as forward-backward splitting, where $\widetilde{\mathbf{w}}_{t+1} = \mathbf{w}_t - \eta_t \nabla g(\mathbf{w}_t; \zeta_t)$ is the forward step and the proximal mapping of r is the backward step:

$$\mathbf{w}_{t+1} = \mathrm{prox}_{\eta_t r}(\widetilde{\mathbf{w}}_{t+1}) = \arg\min_{\mathbf{w}} r(\mathbf{w}) + \frac{1}{2\eta_t}\|\mathbf{w} - \widetilde{\mathbf{w}}_{t+1}\|_2^2.$$

When r is absent, the above update is equivalent to the SGD update. If $r(\mathbf{w})$ corresponds to a domain constraint $\mathbf{w} \in \mathcal{W}$, i.e., $r(\mathbf{w}) = \mathbb{I}_{0-\infty}(\mathbf{w} \in \mathcal{W})$, the above update becomes

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}[\widetilde{\mathbf{w}}_{t+1}] = \min_{\mathbf{w}\in\mathcal{W}} \|\mathbf{w} - \widetilde{\mathbf{w}}_{t+1}\|_2^2, \tag{3.20}$$

which is the projection of $\widetilde{\mathbf{w}}_{t+1} = \mathbf{w}_t - \eta_t \nabla g(\mathbf{w}_t, \zeta_t)$ onto the constrained domain $\mathcal{W}$. This is known as projected SGD method.

Explanation of SPGD update

The update (3.19) is equivalent to:

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} g(\mathbf{w}_t; \zeta_t) + \nabla g(\mathbf{w}_t; \zeta_t)^\top(\mathbf{w} - \mathbf{w}_t) + r(\mathbf{w}) + \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|_2^2.$$

Unlike SGD, SPGD uses a stochastic linear approximation $g(\mathbf{w}_t; \zeta_t) + \nabla g(\mathbf{w}_t; \zeta_t)^\top(\mathbf{w} - \mathbf{w}_t) + r(\mathbf{w})$ as a stochastic surrogate for $g(\mathbf{w}) + r(\mathbf{w})$.
Using the first-order optimality condition of (3.19), $\mathbf{w}_{t+1}$ satisfies

Regularization	$r(\cdot)$	$\text{prox}_{\eta r}(\bar{\mathbf{w}})$ or $\text{prox}_{\eta r}(\bar{W})$
Euclidean norm square	$\frac{\lambda}{2}\|\mathbf{w}\|_2^2$	$\frac{\bar{\mathbf{w}}}{1+\lambda\eta}$
Euclidean norm	$\lambda\|\mathbf{w}\|_2$	$\left(1-\frac{\lambda\eta}{\|\bar{\mathbf{w}}\|_2}\right)_+\bar{\mathbf{w}}$
Lasso	$\lambda\|\mathbf{w}\|_1$	$\text{sign}(\bar{\mathbf{w}})\odot\max\{\|\bar{\mathbf{w}}\|-\lambda\eta,\,0\}$
Group Lasso	$\lambda\sum_g\|\mathbf{w}_g\|_2$	$\left(1-\frac{\lambda\eta}{\|\bar{\mathbf{w}}_g\|_2}\right)_+\bar{\mathbf{w}}_g$ (for each group g)
Elastic Net	$\alpha\|\mathbf{w}\|_1+\frac{\beta}{2}\|\mathbf{w}\|_2^2$	$\frac{1}{1+\eta\beta}\left(\text{sign}(\bar{\mathbf{w}})\odot\max\{\|\bar{\mathbf{w}}\|-\eta\alpha,\,0\}\right)$
Trace norm (nuclear)	$\lambda\|W\|_*=\lambda\sum_i\sigma_i(W)$	$U\,\text{diag}\left((\sigma_i-\lambda\eta)_+\right)V^\top$ $(\bar{W}=U\,\text{diag}(\sigma_i)V^\top)$

Table 3.1: Examples of regularization functions $r(\cdot)$ and their proximal mappings, where σ_i denote the i-th singular value of a matrix.

$$\mathbf{w}_{t+1}=\mathbf{w}_t-\eta(\nabla g(\mathbf{w}_t;\zeta_t)+\partial r(\mathbf{w}_{t+1})). \tag{3.21}$$

It resembles SGD but differs in that it uses a stochastic gradient of g evaluated at $\mathbf{w}_t$ and a subgradient of r evaluated at $\mathbf{w}_{t+1}$.

In order to make the update efficient, the proximal mapping of r should be easily computable. Table 3.1 presents several examples of regularizers r and the corresponding solutions of their proximal mappings, followed by explanations below.We leave the detailed derivations of these proximal mappings to the reader as exercises.

Examples

Example 3.3 (Euclidean norm square)**.** *This is the most commonly used regularizer. Its proximal mapping shrinks the magnitude of the input vector $\bar{\mathbf{w}}$, effectively performing weight decay.*

Example 3.4 (ℓ_1 norm)**.** *The ℓ_1 norm regularizer $\lambda\|\mathbf{w}\|_1$ is used in the well-known Lasso method for linear regression. Its proximal mapping promotes sparsity in the solution by setting some entries to zero if the corresponding component of $\bar{\mathbf{w}}$ is smaller than $\eta\lambda$ in magnitude.*

Example 3.5 (Group Lasso)**.** *This is an extension of Lasso that groups features together and enforces group-wise sparsity. Specifically, if one weight within a group is set to zero, then all weights in that group are simultaneously set to zero.*

Example 3.6 (Trace norm)**.** *The trace norm regularizer for a matrix is analogous to the ℓ_1 norm for a vector, as it promotes low-rank structure. Its proximal mapping induces a low-rank solution by setting the singular values of the input matrix to zero whenever they are smaller than $\eta\lambda$.*

3.2.1 Convex Functions

Lemma 3.6 *Consider the update*

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \mathbf{z}_t^\top(\mathbf{w}-\mathbf{w}_t) + \frac{1}{2\eta_t}\|\mathbf{w}-\mathbf{w}_t\|_2^2 + r(\mathbf{w}). \tag{3.22}$$

If r is μ_r-strongly convex, for any $\mathbf{w}$ we have

$$\mathbf{z}_t^\top(\mathbf{w}_{t+1}-\mathbf{w}) + r(\mathbf{w}_{t+1}) - r(\mathbf{w}) \le \frac{1}{2\eta_t}\|\mathbf{w}_t-\mathbf{w}\|_2^2 - (\frac{1}{2\eta_t}+\frac{\mu_r}{2})\|\mathbf{w}_{t+1}-\mathbf{w}\|_2^2$$
$$-\frac{1}{2\eta_t}\|\mathbf{w}_t-\mathbf{w}_{t+1}\|_2^2.$$

Proof. By the first-order optimality condition of (3.22), for any $\mathbf{w}$ we have

$$(\mathbf{z}_t + \partial r(\mathbf{w}_{t+1}) + \frac{1}{\eta_t}(\mathbf{w}_{t+1}-\mathbf{w}_t))^\top(\mathbf{w}-\mathbf{w}_{t+1}) \ge 0. \tag{3.23}$$

By the strong convexity of r, we have

$$r(\mathbf{w}_{t+1}) \le r(\mathbf{w}) + \partial r(\mathbf{w}_{t+1})^\top(\mathbf{w}_{t+1}-\mathbf{w}) - \frac{\mu_r}{2}\|\mathbf{w}-\mathbf{w}_{t+1}\|_2^2.$$

Adding the above two inequalities, we have

$$\mathbf{z}_t^\top(\mathbf{w}_{t+1}-\mathbf{w}) + r(\mathbf{w}_{t+1}) - r(\mathbf{w}) \le \frac{1}{\eta_t}(\mathbf{w}_t-\mathbf{w}_{t+1})^\top(\mathbf{w}_{t+1}-\mathbf{w}) - \frac{\mu_r}{2}\|\mathbf{w}-\mathbf{w}_{t+1}\|_2^2$$
$$= \frac{1}{2\eta_t}(\|\mathbf{w}_t-\mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1}-\mathbf{w}\|_2^2 - \|\mathbf{w}_t-\mathbf{w}_{t+1}\|_2^2) - \frac{\mu_r}{2}\|\mathbf{w}-\mathbf{w}_{t+1}\|_2^2.$$

where the last equality uses the fact that $2(a-b)^\top(b-c) = \|a-c\|_2^2 - \|a-b\|_2^2 - \|b-c\|_2^2$. □

Theorem 3.5 *Suppose Assumption 3.6 holds. Let $\eta_t = \eta \le 1/L$ and $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_{t+1}$. Then after T iterations of SPGD update (3.19), we have*

$$\mathbb{E}[F(\bar{\mathbf{w}}_T) - F(\mathbf{w}_*)] \le \frac{\|\mathbf{w}_1-\mathbf{w}_*\|_2^2}{2\eta T} + \eta\sigma^2.$$

If $\eta = \min(\frac{1}{L}, \frac{\|\mathbf{w}_1-\mathbf{w}_*\|_2}{\sqrt{2T}\sigma})$, then

$$\mathbb{E}\left[F(\bar{\mathbf{w}}_T) - F(\mathbf{w}_*)\right] \le \frac{\sqrt{2}\sigma\|\mathbf{w}_1-\mathbf{w}_*\|_2}{\sqrt{T}} + \frac{L\|\mathbf{w}_1-\mathbf{w}_*\|_2^2}{T}.$$

Why it matters:

Insight 1: The theorem indicates that even if the objective has a non-smooth regularizer r, the convergence rate of SPGD still depends on the variance bound σ^2 instead of the Lipschitz constant of the objective function as in the analysis of SGD for non-smooth convex functions.
Insight 2: Employing the proximal mapping of r renders the convergence independent of the smoothness of r. Consequently, this approach is advantageous even when r is smooth, particularly if it possesses a large smoothness constant.

Proof. Without loss of generality, we assume g is μ-strongly convex with $\mu \geq 0$ and r is μ_r-strongly convex with $\mu_r \geq 0$ so that it covers both convex and strongly convex cases.

By Lemma 3.6, we have

$$\nabla g(\mathbf{w}_t, \zeta_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}) + r(\mathbf{w}_{t+1}) \leq r(\mathbf{w}) + \frac{1}{2\eta_t}(\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2) - \frac{\mu_r}{2}\|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2^2.$$

By the smoothness of g, we have

$$g(\mathbf{w}_{t+1}) \leq g(\mathbf{w}_t) + \nabla g(\mathbf{w}_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2.$$

By the strong convexity of g, we have

$$g(\mathbf{w}_t) \leq g(\mathbf{w}) + \nabla g(\mathbf{w}_t)^\top (\mathbf{w}_t - \mathbf{w}) - \frac{\mu}{2}\|\mathbf{w}_t - \mathbf{w}\|_2^2.$$

Adding the above two inequalities, we have

$$g(\mathbf{w}_{t+1}) \leq g(\mathbf{w}) + \nabla g(\mathbf{w}_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}) - \frac{\mu}{2}\|\mathbf{w}_t - \mathbf{w}\|_2^2 + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2.$$

Combining this with the first inequality for $\mathbf{w} = \mathbf{w}_*$, we have

$$\begin{aligned} F(\mathbf{w}_{t+1}) - F(\mathbf{w}_*) \leq{}& \frac{1}{2\eta_t}(\|\mathbf{w}_t - \mathbf{w}_*\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2 - \|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2^2) \\ &- \frac{\mu}{2}\|\mathbf{w}_t - \mathbf{w}_*\|_2^2 - \frac{\mu_r}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2 + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\ &+ (\nabla g(\mathbf{w}_t) - \nabla g(\mathbf{w}_t, \zeta_t))^\top (\mathbf{w}_{t+1} - \mathbf{w}_*). \end{aligned} \tag{3.24}$$

This is similar to (3.8) except for the two negative terms $-\frac{\mu}{2}\|\mathbf{w}_t - \mathbf{w}_*\|_2^2 - \frac{\mu_r}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2$, which are due to the μ_r-strong convexity of r and μ-strong convexity of g. If $\mu_r = \mu = 0$, the remaining proof is similar to that of Theorem 3.1 with the following definition of $\hat{\mathbf{w}}_{t+1}$:

$$\hat{\mathbf{w}}_{t+1} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2\eta_t}\|\mathbf{w} - (\mathbf{w}_t - \eta_t \nabla g(\mathbf{w}_t))\|_2^2 + r(\mathbf{w}).$$

It used to bound the expectation of last term in the RHS of (3.24):

$$\begin{aligned}&\mathbb{E}[(\nabla g(\mathbf{w}_t) - g(\mathbf{w}_t, \zeta_t))^\top (\mathbf{w}_{t+1} - \hat{\mathbf{w}}_{t+1} + \hat{\mathbf{w}}_{t+1} - \mathbf{w}_*)] \\ &= \mathbb{E}[(\nabla g(\mathbf{w}_t) - g(\mathbf{w}_t, \zeta_t))^\top (\mathbf{w}_{t+1} - \hat{\mathbf{w}}_{t+1})] \le \eta_t \mathbb{E}[\|(\nabla g(\mathbf{w}_t) - g(\mathbf{w}_t, \zeta_t)\|_2^2] = \eta_t \sigma^2,\end{aligned} \tag{3.25}$$

where the inequality is due to Lemma 1.7.

□

3.2.2 Strongly Convex Functions

We can prove a faster convergence when the loss function or the regularizer is strongly convex.

Theorem 3.6 *Suppose Assumption 3.6 holds and g is μ-strongly convex and r is μ_r-strongly convex. Let $\eta_t = 1/((\mu+\mu_r)t + L)$ and $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_{t+1}$. Then after T iterations of SPGD update (3.19), we have*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T (F(\mathbf{w}_{t+1}) - F(\mathbf{w}_*))\right] \le \frac{(L+\mu_r)\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2}{T} + \frac{(1+\log T)\sigma^2}{T(\mu+\mu_r)}.$$

Proof. Similar to the proof of Theorem 3.5, if $\eta_t \le \frac{1}{L}$ we have

$$\begin{aligned}&\mathbb{E}[(F(\mathbf{w}_{t+1}) - F(\mathbf{w}_*))] \\ &\le \mathbb{E}\left[\left(\frac{1}{2\eta_t}\|\mathbf{w}_t - \mathbf{w}_*\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2 - \frac{\mu}{2}\|\mathbf{w}_t - \mathbf{w}_*\|_2^2 - \frac{\mu_r}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2\right]\right. \\ &\left. + \eta_t\sigma^2.\right.\end{aligned}$$

Taking summation over $t = 1, \ldots, T$ we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^T (F(\mathbf{w}_{t+1}) - F(\mathbf{w}_*))\right] \\ &\le \mathbb{E}\left[\sum_{t=1}^T \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} - \frac{\mu+\mu_r}{2}\right)\|\mathbf{w}_t - \mathbf{w}_*\|_2^2 + \frac{1}{2\eta_0}\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2 + \frac{\mu_r}{2}\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2\right] \\ &+ \sum_{t=1}^T \eta_t\sigma^2.\end{aligned}$$

Let $\eta_t = \frac{1}{(\mu+\mu_r)t+L}$. Then $\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} - \frac{\mu+\mu_r}{2} = 0$ and we have

Algorithm 3 `Restarted SPGD`

1: **Input:** a learning schedule $\{\eta_k, T_K\}_{k=1}^T$, starting point $\mathbf{w}_1$
2: **for** $k = 1, \ldots, K$ **do**
3: run SPGD with a learning rate η_k for T_k iterations starting from $\mathbf{w}_k$
4: return an averaged solution $\mathbf{w}_{k+1}$
5: **end for**

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(F(\mathbf{w}_{t+1}) - F(\mathbf{w}_*))\right]$$
$$\leq \frac{L+\mu_r}{2T}\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2 + \frac{1}{T}\sum_{t=1}^{T}\frac{\sigma^2}{(\mu+\mu_r)t} \leq \frac{L+\mu_r}{2T}\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2 + \frac{(1+\log T)\sigma^2}{T(\mu+\mu_r)}.$$

□

A Restarted Approach

The $\log T$ factor in the convergence bound can be removed using a restarting scheme. It runs in multiple stages. At stage k, it start with a step size η_k and ran SGD with a number of iterations T_k and returns an averaged solution $\mathbf{w}_k$. By choosing η_k, T_k appropriately, after a logarithmic number of K stages, we will get a solution $\mathbf{w}_K$ satisfying $\mathbb{E}[F(\mathbf{w}_K) - F(\mathbf{w}_*)] \leq \epsilon$. The key motivation is coming from the one-stage convergence bound in Theorem 3.5:

$$\mathbb{E}[F(\bar{\mathbf{w}}_T) - F(\mathbf{w}_*)] \leq \frac{\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2}{2\eta T} + \eta\sigma^2. \tag{3.26}$$

Since the μ-strong convexity of F implies that $\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2 \leq \frac{2}{\mu}(F(\mathbf{w}_1) - F(\mathbf{w}_*))$, then we can establish a recursion of the objective gap in a stage-wise manner. From which, we can show the objective gap will decrease geometrically if η_k decreases geometrically and T_k increases accordingly. This is formally stated in the following theorem.

Theorem 3.7 *Suppose Assumption 3.6 holds, F is μ-strongly convex and there exists ϵ_1 such that $F(\mathbf{w}_1) - F(\mathbf{w}_*) \leq \epsilon_1$. Let $\eta_k = \min(\frac{1}{L}, \frac{\epsilon_1}{2^{k+1}\sigma^2})$ and $T_k = \frac{4}{\mu\eta_k}$. Then after $K = \lfloor \log_2(\epsilon_1/\epsilon) \rfloor$ stages of Restarted SPGD updates (Alg. 3), we have*

$$\mathbb{E}\left[F(\mathbf{w}_{K+1}) - F(\mathbf{w}_*)\right] \leq \epsilon.$$

The iteration complexity is $\sum_{k=1}^{K} T_k = O(\frac{\sigma^2}{\mu\epsilon} + \frac{L}{\mu}\log(\frac{\epsilon_1}{\epsilon}))$.

Proof. Let $\epsilon_k = \epsilon_1/2^k$. Then $\epsilon_{K+1} = \epsilon_1/2^{K+1} \leq \epsilon$ and $\epsilon_K \geq \epsilon$.

Applying the one-stage analysis of SPGD, we have

$$\mathbb{E}[F(\bar{\mathbf{w}}_{k+1}) - F(\mathbf{w}_*)] \leq \frac{\mathbb{E}[\|\mathbf{w}_k - \mathbf{w}_*\|_2^2]}{2\eta_k T_k} + \eta_k \sigma^2.$$
$$\leq \frac{\mathbb{E}[F(\mathbf{w}_k) - F(\mathbf{w}_*)]}{\mu \eta_k T_k} + \eta_k \sigma^2.$$

Then we prove by induction. Assume $\mathbb{E}[F(\mathbf{w}_k) - F(\mathbf{w}_*)] \leq \epsilon_k$, we prove $\mathbb{E}[F(\mathbf{w}_{k+1}) - F(\mathbf{w}_*)] \leq \epsilon_{k+1}$.

$$\mathbb{E}[F(\bar{\mathbf{w}}_{k+1}) - F(\mathbf{w}_*)] \leq \frac{\mathbb{E}[\|\mathbf{w}_k - \mathbf{w}_*\|_2^2]}{2\eta_k T_k} + \eta_k \sigma^2.$$
$$\leq \frac{\epsilon_k}{\mu \eta_k T_k} + \eta_k \sigma^2 \leq \frac{\epsilon_k}{\mu \eta_k T_k} + \frac{\epsilon_{k+1}}{2} \leq \frac{\epsilon_k}{4} + \frac{\epsilon_{k+1}}{2} = \epsilon_{k+1}.$$

Thus, $\mathbb{E}[F(\mathbf{w}_{K+1}) - F(\mathbf{w}_*)] \leq \epsilon_{K+1} \leq \epsilon$. The total number of iterations is

$$\sum_{k=1}^{K} T_k = \sum_{k=1}^{K} \frac{4}{\mu \eta_k} = \sum_{k=1}^{K} \max(\frac{4 \cdot 2^{k+1} \sigma^2}{\mu \epsilon_1}, \frac{4L}{\mu})$$
$$\leq \sum_{k=1}^{K} \max(\frac{8\sigma^2}{\mu \epsilon 2^{K-k}}, \frac{4L}{\mu}) = O\left(\frac{\sigma^2}{\mu \epsilon} + \frac{L}{\mu} \log\left(\frac{\epsilon_1}{\epsilon}\right)\right).$$

□

Last-iterate Convergence

Furthermore, if $g(\cdot)$ and/or r is strongly convex, we can also prove $\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2$ converges to zero.

Lemma 3.7 *If g is L-smooth and μ-strongly convex and r is μ_r-strongly convex, for the update (3.19) with $\eta_t \leq 2/L$ we have*

$$\mathbb{E}_{\zeta_t}[\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2] \leq \frac{(1 - (2\eta_t - \eta_t^2 L)\mu)\|\mathbf{w}_t - \mathbf{w}_*\|_2^2 + \eta_t^2 \sigma^2}{1 + \eta \mu_r}. \tag{3.27}$$

If g μ-strongly convex and $\|\partial g(\mathbf{w})\|_2 \leq G$ for $\mathbf{w} \in dom(r)$, for the update (3.19) we have

$$\mathbb{E}_{\zeta_t}[\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2] \leq \frac{(1 - 2\eta_t \mu)\|\mathbf{w}_t - \mathbf{w}_*\|_2^2 + \eta_t^2 (\sigma^2 + 4G^2)}{1 + \eta \mu_r}. \tag{3.28}$$

Proof. Let $\mathbb{E}_t = \mathbb{E}_{\zeta_t}$. Let us consider smooth case first. Due to the optimality condition of $\mathbf{w}_*$, we have

$$\begin{aligned}\mathbf{w}_* &= \arg\min_{\mathbf{w}\in\mathbb{R}^d} \nabla g(\mathbf{w}_*)^\top(\mathbf{w}-\mathbf{w}_*) + \frac{1}{2\eta_t}\|\mathbf{w}-\mathbf{w}_*\|_2^2 + r(\mathbf{w}) \\ &= \mathrm{prox}_{\eta_t r}(\mathbf{w}_* - \eta_t \nabla g(\mathbf{w}_*)).\end{aligned}$$

Due to the Lipschitz continuity of the prox operator (see Lemma 1.7), we have

$$\mathbb{E}_t\|\mathbf{w}_{t+1}-\mathbf{w}_*\|_2^2 \le \frac{1}{1+\eta\mu_r}\mathbb{E}_t\|\mathbf{w}_t - \eta_t\nabla g(\mathbf{w}_t;\zeta_t) - [\mathbf{w}_* - \eta_t\nabla g(\mathbf{w}_*)]\|_2^2. \quad (3.29)$$

Next, we bound

$$\begin{aligned}&\mathbb{E}_t\|\mathbf{w}_t - \eta_t\nabla g(\mathbf{w}_t;\zeta_t) - [\mathbf{w}_* - \eta_t\nabla g(\mathbf{w}_*)]\|_2^2 \\ &= \mathbb{E}_t\|[\mathbf{w}_t - \eta_t\nabla g(\mathbf{w}_t)] - [\mathbf{w}_* - \eta_t\nabla g(\mathbf{w}_*)] + \eta_t\nabla g(\mathbf{w}_t) - \eta_t\nabla g(\mathbf{w}_t,\zeta_t)\|_2^2 \\ &= \mathbb{E}_t\|[\mathbf{w}_t - \eta_t\nabla g(\mathbf{w}_t)] - [\mathbf{w}_* - \eta_t\nabla g(\mathbf{w}_*)]\|_2^2 + \eta_t^2\sigma^2,\end{aligned}$$

where the last inequality uses $\mathbb{E}_t[\nabla g(\mathbf{w}_t,\zeta_t) - \nabla g(\mathbf{w}_t)] = 0$ by expanding the RHS. Let us bound the first term below.

$$\begin{aligned}&\mathbb{E}_t\|[\mathbf{w}_t - \eta_t\nabla g(\mathbf{w}_t)] - [\mathbf{w}_* - \eta_t\nabla g(\mathbf{w}_*)]\|_2^2 \\ &= \mathbb{E}_t\|\mathbf{w}_t-\mathbf{w}_*\|_2^2 + \eta_t^2\mathbb{E}_t\|\nabla g(\mathbf{w}_t) - \nabla g(\mathbf{w}_*)\|_2^2 - 2\eta_t\mathbb{E}_t(\mathbf{w}_t-\mathbf{w}_*)^\top(\nabla g(\mathbf{w}_t) - \nabla g(\mathbf{w}_*)) \\ &\le \mathbb{E}_t\|\mathbf{w}_t-\mathbf{w}_*\|_2^2 + \eta_t^2 L\mathbb{E}_t(\mathbf{w}_t-\mathbf{w}_*)^\top(\nabla g(\mathbf{w}_t) - \nabla g(\mathbf{w}_*)) \\ &- 2\eta_t\mathbb{E}_t(\mathbf{w}_t-\mathbf{w}_*)^\top(\nabla g(\mathbf{w}_t) - \nabla g(\mathbf{w}_*)) \\ &= \mathbb{E}_t\|\mathbf{w}_t-\mathbf{w}_*\|_2^2 - (2\eta_t - \eta_t^2 L)\mathbb{E}_t(\mathbf{w}_t-\mathbf{w}_*)^\top(\nabla g(\mathbf{w}_t) - \nabla g(\mathbf{w}_*)) \\ &\le \mathbb{E}_t\|\mathbf{w}_t-\mathbf{w}_*\|_2^2 - (2\eta_t - \eta_t^2 L)\mu\mathbb{E}_t\|\mathbf{w}_t-\mathbf{w}_*\|_2^2 \\ &\le (1-(2\eta_t-\eta_t^2 L)\mu)\mathbb{E}_t\|\mathbf{w}_t-\mathbf{w}_*\|_2^2,\end{aligned}$$

where the first inequality uses Lemma 1.5(c) and the second inequality follows from Lemma 1.6(c).

If g is non-smooth, we bound $\mathbb{E}\|\nabla g(\mathbf{w}_t) - \nabla g(\mathbf{w}_*)\|_2^2 \le 4G^2$. Combining this with (3.29) concludes the proof.

□

Theorem 3.8 *Suppose Assumption 3.6 holds and g is μ-strongly convex and r is μ_r-strongly convex. Let $\eta_t = \eta \le \min(1/L, 1/\mu_r)$. Then after T iterations of SPGD (3.19) update, we have*

$$\mathbb{E}[\|\mathbf{w}_{T+1}-\mathbf{w}_*\|_2^2] \le e^{-\frac{\eta(\mu+\mu_r)T}{2}}\mathbb{E}[\|\mathbf{w}_1-\mathbf{w}_*\|_2^2] + \frac{\eta\sigma^2}{\mu+\mu_r}. \quad (3.30)$$

Why it matters:

This theorem indicates that if we set $\eta \le O((\mu+\mu_r)\epsilon/\sigma^2)$, then with $T = \tilde{O}\left(\frac{\sigma^2}{(\mu+\mu_r)^2\epsilon}\right)$ iterations, the algorithm finds an solution $\mathbf{w}_{T+1}$ that is ϵ-close to

the optimal solution $\mathbf{w}_*$ measured by $\mathbb{E}[\|\mathbf{w}_{T+1} - \mathbf{w}_*\|_2^2]$, where $\tilde{O}(\cdot)$ hides a logarithmic factor of $\log(1/\epsilon)$.

Proof. If $\eta \leq 1/L$, Lemma 3.7 implies that

$$\begin{aligned}\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2] &\leq \frac{(1-\eta\mu)\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|_2^2] + \eta^2\sigma^2}{1+\eta\mu_r} \\ &\leq \left(1 - \frac{\eta\mu_r}{2}\right)\{(1-\eta\mu)\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|_2^2] + \eta^2\sigma^2\} \\ &\leq \left(1 - \frac{\eta\mu_r}{2} - \eta\mu + \frac{\eta^2\mu\mu_r}{2}\right)\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|_2^2] + \eta^2\sigma^2,\end{aligned}$$

where the first inequality is due to $1 \leq (1+\eta\mu_r)(1-\frac{\eta\mu_r}{2}) = 1 + \frac{\eta\mu_r}{2} - \frac{\eta^2\mu_r^2}{2}$ as $\eta\mu_r \leq 1$. Then

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2] \leq (1 - \frac{\eta\mu_r}{2} - \frac{\eta\mu}{2})\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|_2^2] + \eta^2\sigma^2.$$

Unroll this inequality for $t = 1, \ldots, T$, we have

$$\mathbb{E}[\|\mathbf{w}_{T+1} - \mathbf{w}_*\|_2^2] \leq \left(1 - \frac{\eta(\mu+\mu_r)}{2}\right)\mathbb{E}\|\mathbf{w}_T - \mathbf{w}_*\|_2^2 + \eta^2\sigma^2.$$

Applying this inequality T times gives

$$\begin{aligned}&\mathbb{E}[\|\mathbf{w}_{T+1} - \mathbf{w}_*\|_2^2] \\ &\leq \left(1 - \frac{\eta(\mu+\mu_r)}{2}\right)^T \mathbb{E}\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2 + \sum_{t=0}^{T-1}\left(1 - \frac{\eta(\mu+\mu_r)}{2}\right)^t \eta^2\sigma^2.\end{aligned}$$

Since $(1-\alpha)^T \leq e^{-\alpha T}$ for $\alpha \in (0,1)$ and $\sum_{t=0}^{T-1}\alpha^t < \frac{1}{1-\alpha}$, we have

$$\begin{aligned}&\mathbb{E}[\|\mathbf{w}_{T+1} - \mathbf{w}_*\|_2^2] \leq e^{-\frac{\eta(\mu+\mu_r)T}{2}}\mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2] + \eta^2\sigma^2\frac{2}{\eta(\mu+\mu_r)} \\ &= e^{-\frac{\eta(\mu+\mu_r)T}{2}}\mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2] + \frac{2\eta\sigma^2}{\mu+\mu_r}.\end{aligned}$$

□

Corollary 3.1. *Under the setting of Theorem 3.8, if* $\frac{1}{\eta_t} = \frac{\bar{\mu}}{2} + \sqrt{(\frac{\bar{\mu}}{2})^2 + \frac{1}{\eta_{t-1}^2}}$ *with* $\eta_0 \leq \min(1/L, 1/\mu_r)$ *and* $\bar{\mu} = (\mu+\mu_r)/2$*, then we have*

$$\mathbb{E}[\|\mathbf{w}_{T+1} - \mathbf{w}_*\|_2^2] \leq \frac{4\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2}{\eta_0^2\bar{\mu}^2T^2} + \frac{2\sigma^2}{\bar{\mu}^2T}.$$

Proof. Let $\delta_t = \|\mathbf{w}_t - \mathbf{w}_*\|_2^2$. Due to the update of η_t, we have $1 - \bar{\mu}\eta_t = \frac{\eta_t^2}{\eta_{t-1}^2}$. Hence, we have:

$$\mathbb{E}[\ \delta_{T+1}] \leq \mathbb{E}[(1 - \bar{\mu}\eta_T)\delta_T] + \sigma^2\eta_T^2 \leq \mathbb{E}\left[\frac{\eta_T^2}{\eta_{T-1}^2}\delta_T\right] + \sigma^2\eta_T^2.$$

Unrolling this inequality for $t = 1, \ldots, T$, we have

$$\mathbb{E}[\ \delta_{T+1}] \leq \mathbb{E}\left[\frac{\eta_T^2}{\eta_{T-2}^2}\delta_{T-1}\right] + \sigma^2\eta_T^2 * 2 \leq \frac{\eta_T^2}{\eta_0^2}\delta_1 + \sigma^2\eta_T^2 * T.$$

Since $\frac{1}{\eta_t} = \frac{\bar{\mu}}{2} + \sqrt{(\frac{\bar{\mu}}{2})^2 + \frac{1}{\eta_{t-1}^2}}$. Then, we have $\frac{1}{\eta_t} \geq \frac{\bar{\mu}}{2} + \frac{1}{\eta_{t-1}}$. As a result, $\frac{1}{\eta_T} \geq \frac{\bar{\mu}T}{2} + \frac{1}{\eta_0} \geq \max(L, \mu_r)$, where $\eta_0 \leq \min(\frac{1}{L}, \frac{1}{\mu_r})$. Hence, $\eta_T \leq \frac{2}{\bar{\mu}T}$, and

$$\mathbb{E}[\delta_{T+1}] \leq \frac{4\delta_1}{\eta_0^2\bar{\mu}^2T^2} + \frac{2\sigma^2}{\bar{\mu}^2T}.$$

□

Why it matters:

This corollary shows that a decreasing learning rate schedule can be used without requiring prior knowledge of ϵ, in order to obtain a solution $\mathbf{w}_{T+1}$ that is ϵ-close to the optimum $\mathbf{w}_*$, measured by $\mathbb{E}[\|\mathbf{w}_{T+1} - \mathbf{w}_*\|_2^2]$. The iteration complexity is

$$T = O\left(\max\left\{\frac{1}{\bar{\mu}\eta_0\sqrt{\epsilon}}, \frac{\sigma^2}{\bar{\mu}^2\epsilon}\right\}\right).$$

3.3 Stochastic Coordinate Descent

In this section, we present stochastic coordinate descent (SCD) for solving the stochastic optimization:

$$\min_{\alpha \in \Omega \subseteq \mathbb{R}^n} f(\alpha) = \mathbb{E}[f(\alpha, \xi)]. \tag{3.31}$$

where $\Omega = \Omega_1 \times \Omega_2 \cdots \times \Omega_n$.

The key motivation is that if the dimensionality n of α is very large, then computing $\nabla f(\alpha, \xi)$ could be expensive at each iteration. However, if the function exhibits decomposable structure over dimensions of α, then we can sample a random coordinate of α and update it. To this end, we assume that $[\nabla f(\alpha, \xi)]_i, \forall i \in [n]$ is easy to compute. In machine learning applications, this is possible if $f(\alpha, \xi) = \alpha^\top \boldsymbol{g}(\xi)$ and

computing each coordinate of $\boldsymbol{g}(\xi)$ is much more cheaper than computing itself. An example is the COCE problem (2.62), which will be discussed in Section 5.5.

Let us consider a simple version of SCD. At each iteration t, a coordinate denoted by i_t is randomly sampled from $\{1, \ldots, n\}$ with uniform probabilities. Then we compute $\nabla_{i_t} f(\alpha_t, \xi_t) = [\nabla f(\alpha_t, \xi_t)]_{i_t}$ and update α by

$$\alpha_{t+1,i} = \begin{cases} \Pi_{\Omega_i}[\alpha_{t,i} - \eta \nabla_i f(\alpha_t, \xi_t)] & \text{if } i = i_t \\ \alpha_{t,i} & \text{o.w.} \end{cases}$$

Convergence Analysis

We make the following assumption.

Assumption 3.7. *The following conditions hold:*

(i) $f(\alpha)$ *is convex;*
(ii) For any α*, we have* $\mathbb{E}[\|\nabla_i f(\alpha; \zeta) - \nabla_i f(\alpha)\|_2^2] \le \sigma_i^2$ *for some* $\sigma_i \ge 0$*;*
(iii) ∇f *is* L_i*-Lipschitz continuous w.r.t to the i-th coordinate, i.e.,*

$$\|\nabla f(\alpha) - \nabla f(\alpha + \mathbf{e}_i \delta)\|_2 \le L_i |\delta|.$$

Theorem 3.9 *Let* $\bar{\alpha}_T = \frac{1}{T}\sum_{t=1}^T \alpha_{t+1}$, $\bar{L} = \max_i L_i$. *If* $\eta_t = \eta \le \frac{1}{\bar{L}}$*, after* T *iterations of SCD update we have*

$$\mathbb{E}\Big[f(\bar{\alpha}_t) - f(\alpha_*)\Big] \le \frac{(n-1)(f(\alpha_1) - f(\alpha_*))}{T} + \frac{n}{2\eta T}\|\alpha_1 - \alpha_*\|_2^2 + \sum_{i=1}^n \eta \sigma_i^2.$$

If $\|\alpha_1 - \alpha_*\|_2^2 \le D^2$, $\sum_{i=1}^n \sigma_i^2 \le \sigma^2$*, with* $\eta = O(\min(\frac{\sqrt{n}}{\sqrt{2T}\sigma}, 1/\bar{L}))$*, we have*

$$\mathbb{E}\Big[f(\bar{\alpha}_t) - f(\alpha_*)\Big] \le \frac{(n-1)(f(\alpha_1) - f(\alpha_*))}{T} + \frac{\sqrt{2n}D\sigma}{\sqrt{T}} + \frac{\bar{L}nD^2}{T}.$$

Why it matters:

According to the theorem, SCD's iteration complexity is $O(\frac{nD^2\sigma^2}{\epsilon^2})$. Although this is n times higher than that of SGD, it is offset by the fact that each individual iteration of SCD can be n times cheaper to compute.

Proof. To facilitate the analysis, we consider a virtual sequence $\{\tilde{\alpha}_t\}$ defined by

$$\tilde{\alpha}_{t+1} = \Pi_\Omega[\alpha_t - \eta_t \nabla f(\alpha_t, \xi_t)].$$

Due to the decomposability of $\Omega = \Omega_1 \times \cdots \Omega_n$, it implies that

$$\tilde{\alpha}_{t+1,i} = \Pi_{\Omega_i}[\alpha_{t,i} - \eta_t \nabla_i f(\alpha_t, \xi_t)], \forall i.$$

Algorithm 4 SCD

1: **Input:** learning rate schedule $\{\eta_t\}_{t=1}^T$, starting point α_1
2: **for** $t = 1, \ldots, T$ **do**
3: Sample a coordinate i_t uniformly
4: Compute an unbiased coordinate gradient estimator $\nabla_{i_t} f(\alpha_t, \xi_t)$
5: Update

$$\alpha_{t+1,i} = \begin{cases} \Pi_{\Omega_i}[\alpha_{t,i} - \eta_t \nabla_i f(\alpha_t, \xi_t)] & \text{if } i = i_t \\ \alpha_{t,i} & \text{o.w.} \end{cases}$$

6: **end for**

Applying Lemma 3.6 to each coordinate of $\tilde{\alpha}_{t+1}$ with $r(\alpha_i) = \mathbb{I}_{0-\infty}(\alpha_i \in \Omega_i)$, we have

$$\mathbb{E}[\nabla_i f(\alpha_t, \xi_t)^\top (\tilde{\alpha}_{t+1,i} - \alpha_{*,i})] \leq \frac{1}{2\eta_t}\mathbb{E}[\|\alpha_{t,i} - \alpha_{*,i}\|_2^2 - \frac{1}{2\eta_t}\|\tilde{\alpha}_{t+1,i} - \alpha_{*,i}\|_2^2] - \frac{1}{2\eta_t}\mathbb{E}[\|\alpha_{t,i} - \tilde{\alpha}_{t+1,i}\|_2^2].$$

Then,

$$\mathbb{E}[\nabla_i f(\alpha_t)^\top (\tilde{\alpha}_{t+1,i} - \alpha_{*,i})] \leq \frac{1}{2\eta_t}\mathbb{E}[\|\alpha_{t,i} - \alpha_{*,i}\|_2^2 - \frac{1}{2\eta_t}\|\tilde{\alpha}_{t+1,i} - \alpha_{*,i}\|_2^2] - \frac{1}{2\eta_t}\mathbb{E}[\|\alpha_{t,i} - \tilde{\alpha}_{t+1,i}\|_2^2] + \mathbb{E}[(\nabla_i f(\alpha_t) - \nabla_i f(\alpha_t, \xi_t))^\top (\tilde{\alpha}_{t+1,i} - \alpha_{*,i})].$$

Similar to (3.25), the last term in the RHS can be bounded by $\mathbb{E}[(\nabla_i f(\alpha_t) - \nabla_i f(\alpha_t, \xi_t))^\top (\tilde{\alpha}_{t+1,i} - \alpha_{*,i})] \leq \mathbb{E}(\nabla_i f(\alpha_t) - \nabla_i f(\alpha_t, \xi_t))^2 \leq \eta_t \sigma_i^2$. Then adding the above inequality over $i = 1, \ldots, n$, we have

$$\mathbb{E}[\nabla_i f(\alpha_t)^\top (\tilde{\alpha}_{t+1,i} - \alpha_{*,i})] \leq \frac{1}{2\eta_t}\mathbb{E}\left[\|\alpha_{t,i} - \alpha_{*,i}\|_2^2 - \frac{1}{2\eta_t}\|\tilde{\alpha}_{t+1,i} - \alpha_{*,i}\|_2^2\right] - \frac{1}{2\eta_t}\mathbb{E}[\|\alpha_{t,i} - \tilde{\alpha}_{t+1,i}\|_2^2] + \eta_t \sigma_i^2.$$

Due to the randomness of i_t, we have

$$\mathbb{E}[(\alpha_{t+1,i} - \alpha_{*,i})^2] = \frac{1}{n}\mathbb{E}[(\tilde{\alpha}_{t+1,i} - \alpha_{*,i})^2] + (1 - \frac{1}{n})\mathbb{E}[(\alpha_{t,i} - \alpha_{*,i})^2]$$

$$\mathbb{E}[\nabla_i f(\alpha_t)^\top (\alpha_{t+1,i} - \alpha_{*,i})] = \frac{1}{n}\mathbb{E}[\nabla_i f(\alpha_t)^\top (\tilde{\alpha}_{t+1,i} - \alpha_{*,i})] + (1 - \frac{1}{n})\mathbb{E}[\nabla_i f(\alpha_t)^\top (\alpha_{t,i} - \alpha_{*,i})]$$

$$\mathbb{E}[\|\alpha_{t,i} - \alpha_{t+1,i}\|_2^2] = \frac{1}{n}\mathbb{E}[\|\alpha_{t,i} - \tilde{\alpha}_{t+1,i}\|_2^2].$$

Combining the above, we have

$$\begin{aligned}
&\mathbb{E}[n\nabla_i f(\alpha_t)^\top(\alpha_{t+1,i}-\alpha_{*,i})-(n-1)\nabla_i f(\alpha_t)^\top(\alpha_{t,i}-\alpha_{*,i})]\\
&\le \frac{1}{2\eta_t}\mathbb{E}[\|\alpha_{t,i}-\alpha_{*,i}\|_2^2]-\frac{1}{2\eta_t}\mathbb{E}[(n\|\alpha_{t+1,i}-\alpha_{*,i}\|_2^2-(n-1)\|\alpha_{t,i}-\alpha_{*,i}\|_2^2)]\\
&-\frac{n}{2\eta_t}\mathbb{E}[\|\alpha_{t,i}-\alpha_{t+1,i}\|_2^2]+\eta_t\sigma_i^2.
\end{aligned}$$

Adding this over $i=1,\dots,n$, we have

$$\begin{aligned}
&\mathbb{E}\left[n\sum_{i=1}^n\nabla_i f(\alpha_t)^\top(\alpha_{t+1,i}-\alpha_{*,i})-\sum_{i=1}^n(n-1)\nabla_i f(\alpha_t)^\top(\alpha_{t,i}-\alpha_{*,i})\right]\\
&\le \frac{n}{2\eta_t}\mathbb{E}\left[\sum_{i=1}^n\|\alpha_{t,i}-\alpha_{*,i}\|_2^2\right]-\frac{n}{2\eta_t}\mathbb{E}\left[\sum_{i=1}^n\|\alpha_{t+1,i}-\alpha_{*,i}\|_2^2\right]\\
&-\frac{n}{2\eta_t}\mathbb{E}\left[\sum_{i=1}^n\|\alpha_{t,i}-\alpha_{t+1,i}\|_2^2\right]+\sum_{i=1}^n\eta_t\sigma_i^2.
\end{aligned}$$

For the LHS, we have

$$\begin{aligned}
&n\sum_{i=1}^n\nabla_i f(\alpha_t)^\top(\alpha_{t+1,i}-\alpha_{*,i})-\sum_{i=1}^n(n-1)\nabla_i f(\alpha_t)^\top(\alpha_{t,i}-\alpha_{*,i})\\
&=n\nabla_{i_t} f(\alpha_t)^\top(\alpha_{t+1,i_t}-\alpha_{*,i_t})+n\sum_{i\neq i_t}^n\nabla_i f(\alpha_t)^\top(\alpha_{t,i}-\alpha_{*,i})\\
&\quad-\sum_{i=1}^n(n-1)\nabla_i f(\alpha_t)^\top(\alpha_{t,i}-\alpha_{*,i})\\
&=n\nabla_{i_t} f(\alpha_t)^\top(\alpha_{t+1,i_t}-\alpha_{*,i_t})-n\nabla_{i_t} f(\alpha_t)^\top(\alpha_{t,i_t}-\alpha_{*,i_t})\\
&\quad+\sum_{i=1}^n\nabla_i f(\alpha_t)^\top(\alpha_{t,i}-\alpha_{*,i})\\
&=n\nabla_{i_t} f(\alpha_t)^\top(\alpha_{t+1,i_t}-\alpha_{t,i_t})+\nabla f(\alpha_t)^\top(\alpha_t-\alpha_*).
\end{aligned}$$

By the assumption, we have

$$\begin{aligned}
\nabla_{i_t} f(\alpha_t)^\top(\alpha_{t+1,i_t}-\alpha_{t,i_t})&=\nabla f(\alpha_t)^\top\mathbf{e}_{i_t}(\alpha_{t+1,i_t}-\alpha_{t,i_t})\\
&\ge f(\alpha_{t+1})-f(\alpha_t)-\frac{L_{i_t}}{2}\|(\alpha_{t+1,i_t}-\alpha_{t,i_t})\|_2^2\\
\nabla f(\alpha_t)^\top(\alpha_t-\alpha_*)&\ge f(\alpha_t)-f(\alpha_*).
\end{aligned}$$

Combining the above, we have

$$n\sum_{i=1}^{n}\nabla_i f(\alpha_t)^\top(\alpha_{t+1,i}-\alpha_{*,i})-\sum_{i=1}^{n}(n-1)\nabla_i f(\alpha_t)^\top(\alpha_{t,i}-\alpha_{*,i})$$
$$\geq n(f(\alpha_{t+1})-f(\alpha_t)-\frac{L_{i_t}}{2}\|\alpha_{t+1,i_t}-\alpha_{t,i_t}\|_2^2)+f(\alpha_t)-f(\alpha_*).$$

Thus, we have

$$\mathbb{E}[n(f(\alpha_{t+1})-f(\alpha_t)-\frac{L_{i_t}}{2}\|\alpha_{t+1,i_t}-\alpha_{t,i_t}\|_2^2)+f(\alpha_t)-f(\alpha_*)]$$
$$\leq\frac{n}{2\eta_t}\mathbb{E}\Big[\sum_{i=1}^{n}\|\alpha_{t,i}-\alpha_{*,i}\|_2^2\Big]-\frac{n}{2\eta_t}\mathbb{E}\Big[\sum_{i=1}^{n}\|\alpha_{t+1,i}-\alpha_{*,i}\|_2^2\Big]$$
$$-\frac{n}{2\eta_t}\mathbb{E}\Big[\sum_{i=1}^{n}\|\alpha_{t,i}-\alpha_{t+1,i}\|_2^2\Big]+\sum_{i=1}^{n}\eta_t\sigma_i^2.$$

Re-arranging this, we have

$$\mathbb{E}[n(f(\alpha_{t+1})-f(\alpha_*)-(n-1)(f(\alpha_t)-f(\alpha_*))]$$
$$\leq\frac{n}{2\eta_t}\mathbb{E}\Big[\sum_{i=1}^{n}\|\alpha_{t,i}-\alpha_{*,i}\|_2^2\Big]-\frac{n}{2\eta_t}\mathbb{E}\Big[\sum_{i=1}^{n}\|\alpha_{t+1,i}-\alpha_{*,i}\|_2^2\Big]+\sum_{i=1}^{n}\eta_t\sigma_i^2$$
$$+\mathbb{E}\Big[\frac{nL_{i_t}}{2}\|\alpha_{t+1,i_t}-\alpha_{t,i_t}\|_2^2\Big]-\frac{n}{2\eta_t}\mathbb{E}\Big[\sum_{i=1}^{n}\|\alpha_{t,i}-\alpha_{t+1,i}\|_2^2\Big]$$
$$=\frac{n}{2\eta_t}\mathbb{E}\Big[\sum_{i=1}^{n}\|\alpha_{t,i}-\alpha_{*,i}\|_2^2\Big]-\frac{n}{2\eta_t}\mathbb{E}\Big[\sum_{i=1}^{n}\|\alpha_{t+1,i}-\alpha_{*,i}\|_2^2\Big]+\sum_{i=1}^{n}\eta_t\sigma_i^2$$
$$+\mathbb{E}\Big[\sum_{i=1}^{n}\frac{nL_i}{2}\|\alpha_{t+1,i}-\alpha_{t,i}\|_2^2\Big]-\frac{n}{2\eta_t}\mathbb{E}\Big[\sum_{i=1}^{n}\|\alpha_{t,i}-\alpha_{t+1,i}\|_2^2\Big].$$

If $\eta_t\leq\frac{1}{L}$, the sum of the last two terms is less than 0, then we have

$$\mathbb{E}[f(\alpha_{t+1})-f(\alpha_*)]$$
$$\leq\mathbb{E}[(n-1)(f(\alpha_t)-f(\alpha_*))-(n-1)(f(\alpha_{t+1})-f(\alpha_*))]$$
$$+\frac{n}{2\eta_t}\mathbb{E}\Big[\sum_{i=1}^{n}\|\alpha_{t,i}-\alpha_{*,i}\|_2^2\Big]-\frac{n}{2\eta_t}\mathbb{E}\Big[\sum_{i=1}^{n}\|\alpha_{t+1,i}-\alpha_{*,i}\|_2^2\Big]+\sum_{i=1}^{n}\eta_t\sigma_i^2.$$

Averaging over $t=1,\ldots,T$, we have

$$\mathbb{E}\Big[\frac{1}{T}\sum_{t=1}^{T}f(\alpha_{t+1})-f(\alpha_*)\Big]\leq\frac{(n-1)(f(\alpha_1)-f(\alpha_*))}{T}+\frac{n}{2\eta T}\|\alpha_1-\alpha_*\|_2^2$$
$$+\sum_{i=1}^{n}\eta\sigma_i^2,$$

Algorithm 5 SMD

Input: learning rate schedule $\{\eta_t\}_{t=1}^T$, starting point $\mathbf{w}_1$
for $t = 1, \ldots, T$ **do**
  Compute an unbiased gradient estimator $\mathbf{z}_t = \nabla g(\mathbf{w}_t; \zeta_t)$
  Update the model $\mathbf{w}$ by $\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \mathbf{z}_t^\top(\mathbf{w} - \mathbf{w}_t) + \frac{1}{\eta_t} D_\varphi(\mathbf{w}, \mathbf{w}_t) + r(\mathbf{w})$.
end for

which concludes the proof.

□

3.4 Stochastic Mirror Descent

The SGD update (3.2) and the SPGD update (3.19) can be generalized using the Bregman divergence instead of the Euclidean distance. Let φ be an α-strongly convex function with respect to a general norm $\|\cdot\|$. Recall the definition of Bregman divergence $D_\varphi(\mathbf{w}, \mathbf{w}')$ in Definition 1.7 induced by φ. Due to the strong convexity of φ, we have,

$$D_\varphi(\mathbf{w}, \mathbf{w}') \geq \frac{\alpha}{2}\|\mathbf{w} - \mathbf{w}'\|^2. \tag{3.32}$$

The stochastic mirror descent (SMD) update applied to non-smooth convex optimization problem (3.1) is given by

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \mathcal{G}(\mathbf{w}_t; \zeta_t)^\top(\mathbf{w} - \mathbf{w}_t) + \frac{1}{\eta_t} D_\varphi(\mathbf{w}, \mathbf{w}_t). \tag{3.33}$$

The SMD update applied to composite optimization problem (3.18) is given by

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \nabla g(\mathbf{w}_t; \zeta_t)^\top(\mathbf{w} - \mathbf{w}_t) + \frac{1}{\eta_t} D_\varphi(\mathbf{w}, \mathbf{w}_t) + r(\mathbf{w}). \tag{3.34}$$

Examples

Example 3.7 (Euclidean distance)**.** *By choosing $\varphi(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, which is 1-strongly convex with respect to $\|\cdot\|_2$, the Bregman divergence reduces to the Euclidean distance, and the above updates simplify to SGD or SPGD.*

Example 3.8 (KL Divergence)**.** *Let us consider another example, where $r(\mathbf{w}) = \mathbb{I}_{0-\infty}(\mathbf{w} \in \Delta)$ and $\Delta = \{\mathbf{w} \in \mathbb{R}^d : \mathbf{w} \geq 0,\ \sum_{i=1}^d [\mathbf{w}]_i = 1\}$.*
By choosing $\varphi(\mathbf{w}) = \sum_{i=1}^d [\mathbf{w}]_i \log[\mathbf{w}]_i$, which is 1-strongly convex w.r.t $\|\cdot\|_1$ (cf. Lemma 1.10), the Bregman divergence reduces to the KL-divergence:

$$D_\varphi(\mathbf{w},\mathbf{u}) = \sum_{i=1}^{d} [\mathbf{w}]_i \log \frac{[\mathbf{w}]_i}{[\mathbf{u}]_i},$$

and the SMD update (3.34) simplifies to

$$[\mathbf{w}_{t+1}]_i = \frac{[\mathbf{w}_t]_i \exp(-\eta_t [\nabla g(\mathbf{w}_t;\xi_t)]_i)}{\sum_{j=1}^{d} [\mathbf{w}_t]_j \exp(-\eta_t [\nabla g(\mathbf{w}_t;\xi_t)]_j)},$$

which is also known as stochastic exponential gradient descent.

Convergence Analysis

The following lemma is similar to Lemma 1.7.

Lemma 3.8 *If $r(\cdot)$ is convex and φ is α-strongly convex w.r.t a norm $\|\cdot\|$, with*

$$\mathbf{z}_1 = \arg\min_{\mathbf{w}} \mathbf{w}^\top \mathbf{a} + r(\mathbf{w}) + \frac{1}{\eta} D_\varphi(\mathbf{w},\mathbf{z}),$$
$$\mathbf{z}_2 = \arg\min_{\mathbf{w}} \mathbf{w}^\top \mathbf{b} + r(\mathbf{w}) + \frac{1}{\eta} D_\varphi(\mathbf{w},\mathbf{z}),$$

we have $\|\mathbf{z}_1 - \mathbf{z}_2\| \le \frac{\eta}{\alpha}\|\mathbf{a}-\mathbf{b}\|_*$.

Proof. By the optimality of $\mathbf{z}_1$ and $\mathbf{z}_2$ we have

$$\mathbf{u} := \frac{\nabla\varphi(\mathbf{z}) - \nabla\varphi(\mathbf{z}_1)}{\eta} - \mathbf{a} \in \partial r(\mathbf{z}_1)$$
$$\mathbf{v} := \frac{\nabla\varphi(\mathbf{z}) - \nabla\varphi(\mathbf{z}_2)}{\eta} - \mathbf{b} \in \partial r(\mathbf{z}_2).$$

Since $r(\mathbf{x})$ is convex, we have

$$r(\mathbf{z}_1) \ge r(\mathbf{z}_2) + \mathbf{v}^\top(\mathbf{z}_1 - \mathbf{z}_2)$$
$$r(\mathbf{z}_2) \ge r(\mathbf{z}_1) + \mathbf{u}^\top(\mathbf{z}_2 - \mathbf{z}_1).$$

Adding them together, we have

$$0 \le (\mathbf{u}-\mathbf{v})^\top(\mathbf{z}_1 - \mathbf{z}_2) = \frac{1}{\eta}(\eta\mathbf{b} - \eta\mathbf{a} + \nabla\varphi(\mathbf{z}_2) - \nabla\varphi(\mathbf{z}_1))^\top(\mathbf{z}_1 - \mathbf{z}_2),$$

which implies

$$\frac{1}{\eta}(\nabla\varphi(\mathbf{z}_1) - \nabla\varphi(\mathbf{z}_2))^\top(\mathbf{z}_1 - \mathbf{z}_2) \le (\mathbf{b}-\mathbf{a})^\top(\mathbf{z}_1 - \mathbf{z}_2) \le \|\mathbf{b}-\mathbf{a}\|_*\|\mathbf{z}_1 - \mathbf{z}_2\|.$$

Since φ is α-strongly convex, similar to Lemma 1.6 (c) we have

$$(\nabla\varphi(\mathbf{z}_1) - \nabla\varphi(\mathbf{z}_2))^\top(\mathbf{z}_1 - \mathbf{z}_2) \geq \alpha\|\mathbf{z}_1 - \mathbf{z}_2\|^2.$$

Combining the above two inequalities, we have $\|\mathbf{z}_1 - \mathbf{z}_2\| \leq \frac{\eta}{\alpha}\|\mathbf{a} - \mathbf{b}\|_*$. □

Lemma 3.9 (Generalized Three-point Equality) *For any* $\mathbf{w}, \mathbf{w}_t, \mathbf{w}_{t+1}$*, we have*

$$(\nabla\varphi(\mathbf{w}_t) - \nabla\varphi(\mathbf{w}_{t+1}))^\top(\mathbf{w}_{t+1} - \mathbf{w}) = D_\varphi(\mathbf{w}, \mathbf{w}_t) - D_\varphi(\mathbf{w}, \mathbf{w}_{t+1}) - D_\varphi(\mathbf{w}_{t+1}, \mathbf{w}_t).$$

Proof.

$$\begin{aligned}
&D_\varphi(\mathbf{w}, \mathbf{w}_{t+1}) - D_\varphi(\mathbf{w}, \mathbf{w}_t) \\
&= -\varphi(\mathbf{w}_{t+1}) - \nabla\varphi(\mathbf{w}_{t+1})^\top(\mathbf{w} - \mathbf{w}_{t+1}) + \varphi(\mathbf{w}_t) + \nabla\varphi(\mathbf{w}_t)^\top(\mathbf{w} - \mathbf{w}_t) \\
&= (\nabla\varphi(\mathbf{w}_{t+1}) - \nabla\varphi(\mathbf{w}_t))^\top(\mathbf{w}_{t+1} - \mathbf{w}) - \varphi(\mathbf{w}_{t+1}) + \varphi(\mathbf{w}_t) + \nabla\varphi(\mathbf{w}_t)^\top(\mathbf{w}_{t+1} - \mathbf{w}_t) \\
&= (\nabla\varphi(\mathbf{w}_{t+1}) - \nabla\varphi(\mathbf{w}_t))^\top(\mathbf{w}_{t+1} - \mathbf{w}) - D_\varphi(\mathbf{w}_{t+1}, \mathbf{w}_t).
\end{aligned}$$

Rearranging this equality finishes the proof. □

The following lemma is similar to Lemma 3.6.

Lemma 3.10 *Consider the update*

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \mathbf{z}_t^\top(\mathbf{w} - \mathbf{w}_t) + \frac{1}{\eta_t}D_\varphi(\mathbf{w}, \mathbf{w}_t) + r(\mathbf{w}). \tag{3.35}$$

If $D_r(\mathbf{w}, \mathbf{w}') \geq \mu D_\varphi(\mathbf{w}, \mathbf{w}')$*, then for any* $\mathbf{w}$ *we have*

$$\begin{aligned}
\mathbf{z}_t^\top(\mathbf{w}_{t+1} - \mathbf{w}) + r(\mathbf{w}_{t+1}) - r(\mathbf{w}) \leq \frac{1}{\eta_t}D_\varphi(\mathbf{w}, \mathbf{w}_t) - (\frac{1}{\eta_t} + \mu)D_\varphi(\mathbf{w}, \mathbf{w}_{t+1}) \\
- \frac{1}{\eta_t}D_\varphi(\mathbf{w}_{t+1}, \mathbf{w}_t).
\end{aligned}$$

Proof. By the first-order optimality condition of (3.35), we have

$$(\mathbf{z}_t + \partial r(\mathbf{w}_{t+1}) + \frac{1}{\eta_t}(\nabla\varphi(\mathbf{w}_{t+1}) - \nabla\varphi(\mathbf{w}_t)))^\top(\mathbf{w} - \mathbf{w}_{t+1}) \geq 0. \tag{3.36}$$

By the assumption of r, we have

$$\mu D_\varphi(\mathbf{w}, \mathbf{w}_{t+1}) \leq r(\mathbf{w}) - r(\mathbf{w}_{t+1}) - \partial r(\mathbf{w}_{t+1})^\top(\mathbf{w} - \mathbf{w}_{t+1}).$$

Adding the above two inequalities, we have

$$
\begin{aligned}
&\mathbf{z}_t^\top(\mathbf{w}_{t+1}-\mathbf{w}) + r(\mathbf{w}_{t+1}) - r(\mathbf{w}) \\
&\le \frac{1}{\eta_t}(\nabla\varphi(\mathbf{w}_t) - \nabla\varphi(\mathbf{w}_{t+1}))^\top(\mathbf{w}_{t+1}-\mathbf{w}) - \mu D_\varphi(\mathbf{w},\mathbf{w}_{t+1}) \\
&= \frac{1}{\eta_t}D_\varphi(\mathbf{w},\mathbf{w}_t) - (\frac{1}{\eta_t}+\mu)D_\varphi(\mathbf{w},\mathbf{w}_{t+1}) - \frac{1}{\eta_t}D_\varphi(\mathbf{w}_{t+1},\mathbf{w}_t).
\end{aligned}
$$

where the last equality uses Lemma 3.9. □

3.4.1 Non-smooth Composite Problems

Let us first analyze SMD (3.34) for the composite problem (3.18) under a modified Assumption.

Assumption 3.8. *Suppose the following conditions hold:*

(i) g is convex and L-smooth with respect to the norm $\|\cdot\|$, and r is convex.
(ii) There exists $\sigma > 0$ such that $\mathbb{E}_\zeta[\nabla g(\mathbf{w};\zeta)] = \nabla g(\mathbf{w})$ and $\mathbb{E}_\zeta[\|\nabla g(\mathbf{w};\zeta) - \nabla g(\mathbf{w})\|_*^2] \le \sigma^2$ for all $\mathbf{w}$.

Theorem 3.10 *Suppose Assumption 3.8 holds. Let $\eta_t = \eta \le \alpha/L$ and $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_{t+1}$. After T iterations of SMD update (3.34) for the composite problem (3.18), we have*

$$
\mathbb{E}[F(\bar{\mathbf{w}}_T) - F(\mathbf{w}_*)] \le \frac{D_\varphi(\mathbf{w}_1,\mathbf{w}_*)}{\eta T} + \frac{\eta\sigma^2}{\alpha}.
$$

If $\eta = \min\left(\frac{\alpha}{L}, \frac{\sqrt{\alpha D_\varphi(\mathbf{w}_1,\mathbf{w}_*)}}{\sqrt{T}\sigma}\right)$, then

$$
\mathbb{E}\left[F(\bar{\mathbf{w}}_T) - F(\mathbf{w}_*)\right] \le \frac{2\sigma\sqrt{D_\varphi(\mathbf{w}_1,\mathbf{w}_*)}}{\sqrt{T\alpha}} + \frac{2LD_\varphi(\mathbf{w}_1,\mathbf{w}_*)}{T\alpha}.
$$

Why it matters

The key difference of the above result of SMD from that of SPGD in Theorem 3.5 lies in the divergence measure and the variance bound that is measured in the dual norm. Let us consider $r(\mathbf{w}) = \mathbb{I}_{0-\infty}(\mathbf{w}\in\Delta)$. With the Euclidean setup, the convergence upper bound is dominated by $O(\frac{\sigma_2\|\mathbf{w}_1-\mathbf{w}_*\|_2}{\sqrt{T}})$, where $\sigma_2^2 \ge \mathbb{E}\|\nabla g(\mathbf{w},\zeta) - \nabla g(\mathbf{w})\|_2^2$ for all $\mathbf{w},\zeta$.
In contrast, with the stochastic exponential gradient descent update, the convergence upper bound is dominated by $O(\frac{\sigma_\infty\sqrt{D_\varphi(\mathbf{w}_1,\mathbf{w}_*)}}{\sqrt{T}})$, where $\sigma_\infty^2 \ge \mathbb{E}\|\nabla g(\mathbf{w},\zeta) - \nabla g(\mathbf{w})\|_\infty^2$ for all $\mathbf{w},\zeta$. If we set $[\mathbf{w}_1]_i = \frac{1}{n}$ for all i, then we get $D_\varphi(\mathbf{w}_1,\mathbf{w}_*) \le \log d$ for all $\mathbf{w}_*\in\Delta$. In addition, $\|\mathbf{w}_1 - \mathbf{w}_*\|_2$ could be

$O(1)$. However, the constant σ_∞^2 can be smaller than σ_2^2 by a factor of d. Hence $\frac{\sigma_\infty\sqrt{D_\varphi(\mathbf{w}_1,\mathbf{w}_*)}}{\sigma_2\|\mathbf{w}_1-\mathbf{w}_*\|_2} = O(\frac{\log d}{\sqrt{d}})$, which indicates that stochastic exponential gradient descent may converge faster than SGD.

Proof. From Lemma 3.10, we have

$$\nabla g(\mathbf{w}_t,\zeta_t)^\top(\mathbf{w}_{t+1}-\mathbf{w}) + r(\mathbf{w}_{t+1}) - r(\mathbf{w}) \le \frac{1}{\eta_t}(D_\varphi(\mathbf{w},\mathbf{w}_t) - D_\varphi(\mathbf{w},\mathbf{w}_{t+1})) - \frac{1}{\eta_t}D_\varphi(\mathbf{w}_{t+1},\mathbf{w}_t).$$

Same as (3.7) we have

$$g(\mathbf{w}_{t+1}) \le g(\mathbf{w}) + \nabla g(\mathbf{w}_t)^\top(\mathbf{w}_t-\mathbf{w}) + \nabla g(\mathbf{w}_t)^\top(\mathbf{w}_{t+1}-\mathbf{w}_t) + \frac{L}{2}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2.$$

Adding the above two inequalities for $\mathbf{w} = \mathbf{w}_*$, we have

$$F(\mathbf{w}_{t+1}) - F(\mathbf{w}_*) \le \frac{1}{\eta_t}(D_\varphi(\mathbf{w},\mathbf{w}_t) - D_\varphi(\mathbf{w},\mathbf{w}_{t+1})) - \frac{1}{\eta_t}D_\varphi(\mathbf{w}_{t+1},\mathbf{w}_t) + \frac{L}{2}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2 + (\nabla g(\mathbf{w}_t) - g(\mathbf{w}_t,\zeta_t))^\top(\mathbf{w}_{t+1}-\mathbf{w}_*). \tag{3.37}$$

Similar to the analysis of SPGD, we define:

$$\hat{\mathbf{w}}_{t+1} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \nabla g(\mathbf{w}_t)^\top(\mathbf{w}-\mathbf{w}_t) + \frac{1}{\eta_t}D_\varphi(\mathbf{w},\mathbf{w}_t) + r(\mathbf{w}),$$

which uses the full gradient $\nabla g(\mathbf{w}_t)$, making it independent of ζ_t. Then we have

$$\begin{aligned}&(\nabla g(\mathbf{w}_t) - g(\mathbf{w}_t,\zeta_t))^\top(\mathbf{w}_{t+1}-\mathbf{w}_*) \\ &\le (\nabla g(\mathbf{w}_t) - g(\mathbf{w}_t,\zeta_t))^\top(\mathbf{w}_{t+1}-\hat{\mathbf{w}}_{t+1}) + (\nabla g(\mathbf{w}_t) - g(\mathbf{w}_t,\zeta_t))^\top(\hat{\mathbf{w}}_{t+1}-\mathbf{w}_*).\end{aligned} \tag{3.38}$$

In addition,

$$\begin{aligned}&(\nabla g(\mathbf{w}_t) - g(\mathbf{w}_t,\zeta_t))^\top(\mathbf{w}_{t+1}-\hat{\mathbf{w}}_{t+1}) \le \|\nabla g(\mathbf{w}_t) - g(\mathbf{w}_t,\zeta_t)\|_*\|\mathbf{w}_{t+1}-\hat{\mathbf{w}}_{t+1}\| \\ &\le \frac{\eta_t}{\alpha}\|\nabla g(\mathbf{w}_t) - g(\mathbf{w}_t,\zeta_t)\|_*^2,\end{aligned} \tag{3.39}$$

where the last inequality follows Lemma 3.8. Adding (3.37), (3.38) and (3.39) and using (3.32), we have

$$F(\mathbf{w}_{t+1}) - F(\mathbf{w}_*) \le \frac{1}{2\eta_t}(D_\varphi(\mathbf{w},\mathbf{w}_t) - D_\varphi(\mathbf{w},\mathbf{w}_{t+1})) - \left(\frac{\alpha}{2\eta_t} - \frac{L}{2}\right)\|\mathbf{w}_t-\mathbf{w}_{t+1}\|^2 + (\nabla g(\mathbf{w}_t) - g(\mathbf{w}_t,\zeta_t))^\top(\mathbf{w}_{t+1}-\hat{\mathbf{w}}_{t+1}) + \frac{\eta_t}{\alpha}\|\nabla g(\mathbf{w}_t) - g(\mathbf{w}_t,\zeta_t)\|_*^2.$$

Taking expectation over ζ_t on both sides, we have

$$\begin{aligned}&\mathbb{E}_{\zeta_t}[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_*)]\\ &\quad\le \mathbb{E}_{\zeta_t}\left[\frac{1}{\eta_t}(D_\varphi(\mathbf{w},\mathbf{w}_t) - D_\varphi(\mathbf{w},\mathbf{w}_{t+1})) - \left(\frac{\alpha}{2\eta_t} - \frac{L}{2}\right)\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2\right] + \frac{\eta_t}{\alpha}\sigma^2.\end{aligned}$$

If $\eta_t \le \frac{\alpha}{L}$, we have

$$\mathbb{E}_{\zeta_t}[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_*)] \le \mathbb{E}_{\zeta_t}\left[\frac{1}{\eta_t}(D_\varphi(\mathbf{w},\mathbf{w}_t) - D_\varphi(\mathbf{w},\mathbf{w}_{t+1}))\right] + \frac{\eta_t}{\alpha}\sigma^2.$$

Summing over $t = 1,\dots,T$, we have

$$\mathbb{E}\left[\frac{1}{\sum_{t=1}^T \eta_t}\sum_{t=1}^T \eta_t(F(\mathbf{w}_{t+1}) - F(\mathbf{w}_*))\right] \le \frac{D_\varphi(\mathbf{w}_1,\mathbf{w}_*)}{\sum_{t=1}^T \eta_t} + \frac{\sum_{t=1}^T \eta_t\sigma^2}{\alpha\sum_{t=1}^T \eta_t}.$$

Let $\eta_t = \eta$ and optimizing the upper bound over η finishes the proof. □

3.4.2 Non-smooth Problems

Next, we present the convergence analysis of SMD (3.33) for non-smooth convex objectives under the following assumption.

Assumption 3.9. *For any* $\mathbf{w}$*, we have* $\mathbb{E}_\zeta[\mathcal{G}(\mathbf{w};\zeta)] \in \partial g(\mathbf{w})$ *and* $\mathbb{E}[\|\mathcal{G}(\mathbf{w};\zeta)\|_*^2] \le G^2$.

Theorem 3.11 *Suppose Assumption 3.9 holds. Let the learning rate* $\{\eta_t\}$ *be* $\eta_t = \eta$ *and* $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t$*. After* T *iterations of SMD update (3.34), we have*

$$\mathbb{E}[g(\bar{\mathbf{w}}_T) - g(\mathbf{w}_*)] \le \frac{D_\varphi(\mathbf{w}_*,\mathbf{w}_1)}{\eta T} + \frac{\eta G^2}{2\alpha}.$$

If $\eta = \frac{\sqrt{2\alpha D_\varphi(\mathbf{w}_*,\mathbf{w}_1)}}{\sqrt{T}G}$*, then*

$$\mathbb{E}[g(\bar{\mathbf{w}}_T) - g(\mathbf{w}_*)] \le \frac{G\sqrt{2D_\varphi(\mathbf{w}_*,\mathbf{w}_1)}}{\sqrt{\alpha T}}.$$

Proof. From Lemma 3.10, we have

$$\mathcal{G}(\mathbf{w}_t,\zeta_t)^\top(\mathbf{w}_{t+1} - \mathbf{w}) \le \frac{1}{\eta_t}(D_\varphi(\mathbf{w},\mathbf{w}_t) - D_\varphi(\mathbf{w},\mathbf{w}_{t+1})) - \frac{1}{\eta_t}D_\varphi(\mathbf{w}_{t+1},\mathbf{w}_t).$$

Rearranging it, we get

$$
\begin{aligned}
&\eta_t \mathcal{G}(\mathbf{w}_t;\zeta_t)^\top(\mathbf{w}_t-\mathbf{w}) \\
&\quad \le D_\varphi(\mathbf{w},\mathbf{w}_t) - D_\varphi(\mathbf{w},\mathbf{w}_{t+1}) - D_\varphi(\mathbf{w}_{t+1},\mathbf{w}_t) + \eta_t \mathcal{G}(\mathbf{w}_t;\zeta_t)^\top(\mathbf{w}_t-\mathbf{w}_{t+1}) \\
&\quad \le D_\varphi(\mathbf{w},\mathbf{w}_t) - D_\varphi(\mathbf{w},\mathbf{w}_{t+1}) - D_\varphi(\mathbf{w}_{t+1},\mathbf{w}_t) \\
&\quad + \frac{\eta_t^2}{2\alpha}\|\mathcal{G}(\mathbf{w}_t;\zeta_t)\|_*^2 + \frac{\alpha}{2}\|\mathbf{w}_t-\mathbf{w}_{t+1}\|^2,
\end{aligned}
$$

where the last inequality uses the Cauchy-Schwarz inequality. Using (3.32), we have

$$
\eta_t \mathcal{G}(\mathbf{w}_t;\zeta_t)^\top(\mathbf{w}_t-\mathbf{w}) \le D_\varphi(\mathbf{w},\mathbf{w}_t) - D_\varphi(\mathbf{w},\mathbf{w}_{t+1}) + \frac{\eta_t^2}{2\alpha}\|\mathcal{G}(\mathbf{w}_t;\zeta_t)\|_*^2. \quad (3.40)
$$

The remaining proof is similar to that of Theorem 3.2. □

3.5 Adaptive Gradient Method (AdaGrad)

The stochastic algorithms discussed so far are fairly general and were originally developed to address a wide range of problems, extending beyond those encountered specifically in machine learning. Nevertheless, the ERM problem of machine learning may exhibit some unique properties dependent on data. How to leverage them to develop a stochastic algorithm that could be potentially faster in practice?

Below, we introduce Adaptive Gradient Method (AdaGrad), which employs an adaptive step size, which incorporates knowledge of the geometry of the data observed in earlier iterations to perform more informative gradient-based learning.

While AdaGrad was considered an important breakthrough in machine learning, it indeed evolves from SMD. We use the same language as SMD to present AdaGrad and its analysis. Let us consider the smooth problem (3.1) and recall the update of SMD:

$$
\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \nabla g(\mathbf{w}_t;\zeta_t)^\top \mathbf{w} + \frac{1}{\eta} D_\varphi(\mathbf{w},\mathbf{w}_t).
$$

The key design to AdaGrad is to use a time-varying proximal function φ_t that changes across iterations. A specific way to construction φ_t is the following.

Let $H_t = \mathrm{diag}(s_{t,1},\ldots,s_{t,d})$ be a diagonal positive matrix. Define $\varphi_t(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top H_t \mathbf{w}$ and a general norm $\|\mathbf{w}\|_H = \sqrt{\mathbf{w}^\top H \mathbf{w}}$. Then the Bregman divergence induced by φ_t becomes:

$$
D_{\varphi_t}(\mathbf{w},\mathbf{w}') = \frac{1}{2}(\mathbf{w}-\mathbf{w}')^\top H_t(\mathbf{w}-\mathbf{w}') = \frac{1}{2}\sum_{i=1}^d s_{t,i}(w_i - w_i')^2,
$$

which is 1-strongly convex w.r.t $\|\cdot\|_H$. The weights $\mathbf{s}_t = (s_{t,1},\ldots,s_{t,d})$ are updated according to the following:

Algorithm 6 AdaGrad

Input: learning rate parameter η, starting point $\mathbf{w}_1$
for $t = 1, \ldots, T$ **do**
  Compute an unbiased gradient estimator $\mathbf{z}_t = \nabla g(\mathbf{w}_t; \zeta_t)$
  Update $s_{t,i} = \sqrt{\sum_{\tau=1}^{t} \|[\nabla g(\mathbf{w}_\tau; \zeta_t)]_i\|_2^2}, \forall i.$
  Update the model $\mathbf{w}$ by $\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\eta}{\mathbf{s}_t} \circ \mathbf{z}_t$
end for

$$s_{t,i} = \sqrt{\sum_{\tau=1}^{t} [\nabla g(\mathbf{w}_\tau; \zeta_\tau)]_i^2}, \forall i, \tag{3.41}$$

which essentially measures the growth of stochastic gradients across all iterations before t.

Let $\mathbf{z}_t = \nabla g(\mathbf{w}_t, \zeta_t)$, and $\mathbf{m}_{1:t} = [\mathbf{z}_1, \ldots, \mathbf{z}_t]$, and $\mathbf{m}_{1:t,i}$ denotes its i-th row vector. Then $s_{t,i} = \|\mathbf{m}_{1:t,i}\|_2$. As a result, the updating step becomes

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta H_t^{-1} \nabla g(\mathbf{w}_t; \zeta_t) = \mathbf{w}_t - \frac{\eta}{\mathbf{s}_t} \circ \nabla g(\mathbf{w}_t; \zeta_t), \tag{3.42}$$

where $\circ$ denotes element-wise product. The full steps of AdaGrad are summarized in Algorithm 6.

Compared with SGD, there are two differences: (i) the effective step size $\frac{\eta}{\mathbf{s}_t}$ is adaptive that depends on the history of updates, hence depends on data sampled $\zeta_1, \ldots, \zeta_t$. This is the reason it is called adaptive step size; (ii) each coordinate of $\mathbf{w}$ will receive a different step size. This feature makes it useful to tackle deep neural networks as the parameters at each layer usually have different orders of gradient.

Convergence Analysis

Let the dual norm of $\|\cdot\|_H$ is given by $\|\mathbf{u}\|_{H^{-1}} = \sqrt{\mathbf{u}^\top H^{-1} \mathbf{u}}$. Then, φ_t is 1-strongly convex in terms of $\|\cdot\|_{H_t}$.

Lemma 3.11 *We have*

$$\sum_{t=1}^{T} \{D_{\varphi_t}(\mathbf{w}_*, \mathbf{w}_t) - D_{\varphi_t}(\mathbf{w}_*, \mathbf{w}_{t+1})\} \leq \frac{1}{2} \max_{t \leq T} \|\mathbf{w}_* - \mathbf{w}_t\|_\infty^2 \sum_{i=1}^{d} s_{T,i}.$$

Proof.

$$
\begin{aligned}
&\sum_{t=1}^{T}\{D_{\varphi_t}(\mathbf{w}_*,\mathbf{w}_t)-D_{\varphi_t}(\mathbf{w}_*,\mathbf{w}_{t+1})\}\\
&=\sum_{t=1}^{T}\{D_{\varphi_t}(\mathbf{w}_*,\mathbf{w}_t)-D_{\varphi_{t-1}}(\mathbf{w}_*,\mathbf{w}_t)+D_{\varphi_{t-1}}(\mathbf{w}_*,\mathbf{w}_t)-D_{\varphi_t}(\mathbf{w}_*,\mathbf{w}_{t+1})\}\\
&\le\sum_{t=1}^{T}\{D_{\varphi_t}(\mathbf{w}_*,\mathbf{w}_t)-D_{\varphi_{t-1}}(\mathbf{w}_*,\mathbf{w}_t)\}+D_{\varphi_0}(\mathbf{w}_*,\mathbf{w}_1)\\
&=D_{\varphi_0}(\mathbf{w}_*,\mathbf{w}_1)+\frac{1}{2}\sum_{t=1}^{T}(\mathbf{w}_*-\mathbf{w}_t)^\top(H_t-H_{t-1})(\mathbf{w}_*-\mathbf{w}_t).
\end{aligned}
$$

Since $\mathbf{s}_t \succeq \mathbf{s}_{t-1}$, we have

$$
\begin{aligned}
&\sum_{t=1}^{T}(\mathbf{w}_*-\mathbf{w}_t)^\top(H_t-H_{t-1})(\mathbf{w}_*-\mathbf{w}_t)=\sum_{t=1}^{T}\sum_{i=1}^{d}(s_{t,i}-s_{t-1,i})([\mathbf{w}_*]_i-[\mathbf{w}_t]_i)^2\\
&\le\max_{t\le T}\|\mathbf{w}_*-\mathbf{w}_t\|_\infty^2\sum_{t=1}^{T}\sum_{i=1}^{d}(s_{t,i}-s_{t-1,i})=\max_{t\le T}\|\mathbf{w}_*-\mathbf{w}_t\|_\infty^2\sum_{i=1}^{d}(s_{T,i}-s_{0,i}).
\end{aligned}
$$

Combining the above two inequalities, we have

$$
\begin{aligned}
&\sum_{t=1}^{T}D_{\varphi_t}(\mathbf{w}_*,\mathbf{w}_t)-D_{\varphi_t}(\mathbf{w}_*,\mathbf{w}_{t+1})\\
&\le D_{\varphi_0}(\mathbf{w}_*,\mathbf{w}_1)+\frac{1}{2}\max_{t\le T}\|\mathbf{w}_*-\mathbf{w}_{t+1}\|_\infty^2\sum_{i=1}^{d}(s_{T+1,i}-s_{1,i})\\
&\le\frac{1}{2}\|\mathbf{w}_1-\mathbf{w}_*\|_\infty^2\sum_{i=1}^{d}s_{0,i}+\frac{1}{2}\max_{t\le T}\|\mathbf{w}_*-\mathbf{w}_t\|_\infty^2\sum_{i=1}^{d}(s_{T,i}-s_{0,i})\\
&\le\frac{1}{2}\max_{t\le T}\|\mathbf{w}_*-\mathbf{w}_t\|_\infty^2\sum_{i=1}^{d}s_{T,i}.
\end{aligned}
$$

□

Lemma 3.12 *We have*

$$
\sum_{t=1}^{T}\|\nabla g(\mathbf{w}_t;\zeta_t)\|_{H_t^{-1}}^2\le 2\sum_{i=1}^{d}s_{T,i}.
$$

Proof. Let us first prove a general result in the following: for a general real-value sequence $\{a_t\}$, we have

$$\sum_{t=1}^{T} \frac{a_t^2}{\|a_{1:t}\|_2} \leq 2 \sum_{t=1}^{T} \|a_{1:T}\|_2, \tag{3.43}$$

where $a_{1:t} = (a_1, \ldots, a_t)$. We prove this by induction. First, it holds trivially for $t = 1$. Now, assume it holds for $T - 1$, we prove it holds for T.

$$\sum_{t=1}^{T} \frac{a_t^2}{\sqrt{\sum_{\tau=1}^{t} a_t^2}} = \sum_{t=1}^{T-1} \frac{a_t^2}{\sqrt{\sum_{\tau=1}^{t} a_t^2}} + \frac{a_T^2}{\|a_{1:T}\|_2} \leq 2 \sum_{t=1}^{T} \|a_{1:T-1}\|_2 + \frac{a_T^2}{\|a_{1:T}\|_2}.$$

Let $b_T = \sqrt{\sum_{t=1}^{T} a_t^2}$, then we have

$$2 \sum_{t=1}^{T} \|a_{1:T-1}\|_2 + \frac{a_T^2}{\|a_{1:T}\|_2} = 2\sqrt{b_T^2 - a_T^2} + \frac{a_T^2}{\sqrt{b_T^2}}.$$

Since $\sqrt{\cdot}$ is a concave function, applying $\sqrt{x + \delta} \leq \sqrt{x} + \delta \frac{1}{2\sqrt{x}}$ we have

$$\sqrt{b_T^2 - a_T^2} \leq \sqrt{b_T^2} - (a_T^2) \frac{1}{2\sqrt{b_T^2}}.$$

Hence, $2\sqrt{b_T^2 - a_T^2} + \frac{a_T^2}{\sqrt{b_T^2}} \leq 2\sqrt{b_T^2}$. Thus, we prove (3.43) for T.

Next, we apply this result to the following:

$$\begin{aligned}
&\sum_{t=1}^{T} \|\nabla g(\mathbf{w}_t; \zeta_t)\|_{H_t^{-1}}^2 = \sum_{t=1}^{T} \nabla g(\mathbf{w}_t; \zeta_t)^\top \mathrm{diag}(\mathbf{s}_t)^{-1} \nabla g(\mathbf{w}_t; \zeta_t) \\
&= \sum_{i=1}^{d} \frac{[\nabla g(\mathbf{w}_t; \zeta_t)]_i^2}{\sqrt{\sum_{\tau=1}^{t} [\nabla g(\mathbf{w}_\tau; \zeta_\tau)]_i^2}} \leq \sum_{i=1}^{d} 2\sqrt{\sum_{\tau=1}^{t} [\nabla g(\mathbf{w}_\tau; \zeta_\tau)]_i^2}.
\end{aligned}$$

□

Theorem 3.12 *Let* $\bar{\mathbf{w}}_T = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}_t$*, then AdaGrad guarantees that*

$$\begin{aligned}
\mathbb{E}[g(\bar{\mathbf{w}}_t) - g(\mathbf{w}_*)] \leq & \frac{\mathbb{E}\left[\max_{t \leq T} \|\mathbf{w}_* - \mathbf{w}_t\|_\infty^2 \sum_{i=1}^{d} \|\mathbf{m}_{1:T,i}\|_2\right]}{2\eta T} \\
& + \frac{\eta \mathbb{E}\left[\sum_{i=1}^{d} \|\mathbf{m}_{1:T,i}\|_2\right]}{T}.
\end{aligned}$$

If $\max_t \|\mathbf{w}_* - \mathbf{w}_t\|_\infty \leq D_\infty$ *and* $\eta = D_\infty / \sqrt{2}$*, we have*

$$\mathbb{E}[g(\bar{\mathbf{w}}_t) - g(\mathbf{w}_*)] \leq \frac{\sqrt{2} D_\infty \mathbb{E}\left[\sum_{i=1}^{d} \|\mathbf{m}_{1:T,i}\|_2\right]}{T}.$$

Why it matters

The above result shows the convergence rate depends on the growth rate of the cumulative stochastic gradient $\sum_{i=1}^{d} \|\mathbf{m}_{1:T,i}\|_2$. In the worst case, it grows at a rate of $O(\sqrt{T})$, inducing a convergence rate of $O(1/\sqrt{T})$, similar to SGD. However, when the cumulative stochastic gradient grows slower than $O(\sqrt{T})$, AdaGrad will enjoy a convergence rate of $o(1/\sqrt{T})$.

Let us consider the following linear model with sparse random data scenario, where $g(\mathbf{w}_t, \zeta_t) = [1 - \mathbf{w}_t^\top \zeta_t]_+$ and the data vectors $\zeta_t \in \{-1, 0, 1\}^d$. Assume that at in each round t, feature i appears with probability $p_i = \min\{1, ci^{-\alpha}\}$ for some $\alpha \in (1, \infty)$ and a dimension-independent constant c. Then we have

$$\mathbb{E}\left[\sum_{i=1}^{d} \|\mathbf{m}_{1:T,i}\|_2\right] = \mathbb{E}\left[\sum_{i=1}^{d} \sqrt{\left|t : \mathbf{z}_{t,i} = 1\right|}\right] \leq \sum_{i=1}^{d} \sqrt{\mathbb{E}\left[\left|t : \mathbf{z}_{t,i} = 1\right|\right]} = \sum_{i=1}^{d} \sqrt{p_i T}.$$

by Jensen's inequality. In the rightmost sum, we have $c\sum_{i=1}^{d} i^{-\alpha/2} = O(\log d)$ for $\alpha \geq 2$, and $\sum_{i=1}^{d} i^{-\alpha/2} = O(d^{1-\alpha/2})$ for $\alpha \in (1, 2)$. If $\mathbf{w}_t$ is restricted in a domain $\mathcal{W} = \{\mathbf{w} : \|\mathbf{w}\|_\infty \leq 1\}$, then $D_\infty = 2$, and the convergence rate of AdaGrad becomes $O(\max\{\log d, d^{1-\alpha/2}\}/\sqrt{T})$. For contrast, the convergence rate of SGD in Theorem 3.2 is $O(\sqrt{d/T})$. So we see that in this sparse yet heavy tailed feature setting, AdaGrad's convergence bound can be exponentially smaller in the dimension d than the non-adaptive bound of SGD.

Proof. Similar to (3.40) in the proof of Theorem 3.11, we have

$$\eta\langle \nabla g(\mathbf{w}_t; \zeta_t), \mathbf{w}_t - \mathbf{w}\rangle \leq D_{\varphi_t}(\mathbf{w}, \mathbf{w}_t) - D_{\varphi_t}(\mathbf{w}, \mathbf{w}_{t+1}) + \frac{\eta_t^2}{2}\|\nabla g(\mathbf{w}_t; \zeta_t)\|_{H_t^{-1}}^2. \quad (3.44)$$

Taking expectation and summation over $t = 1, \ldots, T$, we have

$$\sum_{t=1}^{T} \eta\mathbb{E}[g(\mathbf{w}_t) - g(\mathbf{w}_*)] \leq \mathbb{E}\left[\sum_{t=1}^{T} D_{\varphi_t}(\mathbf{w}, \mathbf{w}_t) - D_{\varphi_t}(\mathbf{w}, \mathbf{w}_{t+1})\right] + \mathbb{E}\left[\sum_{t=1}^{T} \frac{\eta^2}{2}\|\nabla g(\mathbf{w}_t; \zeta_t)\|_{H_t^{-1}}^2\right].$$

Using the results from the two lemmas above, we conclude the proof.

□

3.6 Stochastic Gradient Descent Ascent

In this section, we consider stochastic convex–concave min–max optimization problems:

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{u}\in\mathcal{U}} f(\mathbf{w},\mathbf{u}) := \mathbb{E}_{\zeta}\big[f(\mathbf{w},\mathbf{u};\zeta)\big].$$

This class of problems has two important applications in machine learning: (1) it serves as a foundation for directly formulating learning tasks (e.g., the DRO problem (2.11)); (2) it provides a tool for reformulating standard minimization problems to enable more efficient optimization.

A solution of interest is the so-called saddle point $(\mathbf{w}_*,\mathbf{u}_*)\in\mathcal{W}\times\mathcal{U}$ satisfying:

$$f(\mathbf{w}_*,\mathbf{u}) \le f(\mathbf{w}_*,\mathbf{u}_*) \le f(\mathbf{w},\mathbf{u}_*), \forall \mathbf{w}\in\mathcal{W}, \mathbf{u}\in\mathcal{U}.$$

In many machine learning applications, we may be only interested in finding a global optimal solution to the objective $F(\mathbf{w}) = \max_{\mathbf{u}\in\mathcal{U}} f(\mathbf{w},\mathbf{u})$. It is easy to see that if $(\mathbf{w}_*,\mathbf{u}_*)$ is a saddle point, then $\mathbf{w}_*$ is a global optimal solution to $F(\mathbf{w})$. This can be seen from

$$\max_{\mathbf{u}\in\mathcal{U}} f(\mathbf{w}_*,\mathbf{u}) \le f(\mathbf{w}_*,\mathbf{u}_*) \le f(\mathbf{w},\mathbf{u}_*) \le \max_{\mathbf{u}\in\mathcal{U}} f(\mathbf{w},\mathbf{u}).$$

For a point $(\mathbf{w},\mathbf{u})\in\mathcal{W}\times\mathcal{U}$, a convergence measure is defined by the duality gap:

$$\Delta(\mathbf{w},\mathbf{u}) = \max_{\mathbf{u}'\in\mathcal{U}} f(\mathbf{w},\mathbf{u}') - \min_{\mathbf{w}'\in\mathcal{W}} f(\mathbf{w}',\mathbf{u}).$$

A simple method for solving the convex-concave min-max problem is the stochastic gradient descent ascent (SGDA) algorithm, which is an extension of SGD. It employs two key updates:

$$\begin{aligned}
\mathbf{w}_{t+1} &= \arg\min_{\mathbf{w}\in\mathcal{W}} \partial_1 f(\mathbf{w}_t,\mathbf{u}_t;\zeta_t)^\top(\mathbf{w}-\mathbf{w}_t) + \frac{1}{2\eta_1}\|\mathbf{w}-\mathbf{w}_t\|_2^2 \\
\mathbf{u}_{t+1} &= \arg\min_{\mathbf{u}\in\mathcal{U}} -\partial_2 f(\mathbf{w}_t,\mathbf{u}_t;\zeta_t)^\top(\mathbf{u}-\mathbf{u}_t) + \frac{1}{2\eta_2}\|\mathbf{u}-\mathbf{u}_t\|_2^2,
\end{aligned} \tag{3.45}$$

where $\partial_1 f(\mathbf{w},\mathbf{u};\zeta)$ and $\partial_2 f(\mathbf{w},\mathbf{u};\zeta)$ denote the stochastic partial subgradients such that $\mathbb{E}_\zeta[\partial_1 f(\mathbf{w},\mathbf{u};\zeta)]\in\partial_1 f(\mathbf{w},\mathbf{u})$ and $\mathbb{E}_\zeta[\partial_2 f(\mathbf{w},\mathbf{u};\zeta)]\in\partial_2 f(\mathbf{w},\mathbf{u})$.

Convergence Analysis

Below, we analyze the convergence rate of SGDA under the following assumptions.

Assumption 3.10. *Suppose the following conditions hold:*

(i) $f(\mathbf{w},\mathbf{u})$ *is convex w.r.t* $\mathbf{w}$ *and concave w.r.t* $\mathbf{u}$.

Algorithm 7 SGDA

Input: learning rates $\{\eta_1, \eta_2\}$, starting points $\mathbf{w}_1, \mathbf{u}_1$
for $t = 1, \ldots, T$ **do**
  Compute unbiased gradient estimators $\mathbf{z}_{1,t} = \partial_1 f(\mathbf{w}_t; \zeta_t)$ and $\mathbf{z}_{2,t} = \partial_2 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t)$
  Update the primal variable $\mathbf{w}$ by $\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} \mathbf{z}_{1,t}^\top (\mathbf{w} - \mathbf{w}_t) + \frac{1}{2\eta_1}\|\mathbf{w} - \mathbf{w}_t\|_2^2$.
  Update the dual variable $\mathbf{u}$ by $\mathbf{u}_{t+1} = \arg\min_{\mathbf{u}\in\mathcal{U}} -\mathbf{z}_{2,t}^\top (\mathbf{u} - \mathbf{u}_t) + \frac{1}{2\eta_2}\|\mathbf{u} - \mathbf{u}_t\|_2^2$.
end for

(ii) There exist $G_1, G_2 > 0$ *such that*

$$\mathbb{E}_\zeta[\|\partial_1 f(\mathbf{w}, \mathbf{u}; \zeta)\|_2^2] \le G_1^2, \forall \mathbf{w} \in \mathcal{W}, \mathbf{u} \in \mathcal{U}, \tag{3.46}$$

$$\mathbb{E}_\zeta[\|\partial_2 f(\mathbf{w}, \mathbf{u}; \zeta)\|_2^2] \le G_2^2, \forall \mathbf{w} \in \mathcal{W}, \mathbf{u} \in \mathcal{U}. \tag{3.47}$$

(iii) $\max_{\mathbf{w}\in\mathcal{W}, \mathbf{w}'\in\mathcal{W}} \|\mathbf{w} - \mathbf{w}'\| \le D_1$ *and* $\max_{\mathbf{u}\in\mathcal{U}, \mathbf{u}'\in\mathcal{U}} \|\mathbf{u} - \mathbf{u}'\| \le D_2$.

Lemma 3.13 *Let us consider a martingale difference sequence* $\{\delta_t\}_{t\ge1}$ *and a sequence* $\{y_t\}_{t\ge1}$*:*

$$y_{t+1} = \arg\min_{v\in\mathcal{V}} \{-\delta_t^\top v + \alpha D_\psi(v, y_t)\}.$$

If ψ *is* μ_ψ*-strongly convex w.r.t.* $\|\cdot\|$ *(*$\mu_\psi > 0$*). For any* v *(that possibly depends on* $\{\delta_t\}$*) we have*

$$\mathbb{E}\left[\delta_t^\top v\right] \le \mathbb{E}\left[\alpha D_\psi(v, y_t) - \alpha D_\psi(v, y_{t+1}) + \frac{1}{2\alpha\mu_\psi}\|\delta_t\|_*^2\right].$$

Why it matters

In standard minimization problems, the convergence measure is usually defined with respect to the optimal solution $\mathbf{w}_*$, which is fixed and independent of the randomness introduced by the algorithm. In contrast, in stochastic min–max optimization we are concerned with the duality gap $\Delta(\mathbf{w}, \mathbf{u}) = \max_{\mathbf{u}'\in\mathcal{U}} f(\mathbf{w}, \mathbf{u}') - \min_{\mathbf{w}'\in\mathcal{W}} f(\mathbf{w}', \mathbf{u})$, where the optimal $\mathbf{w}'$ and $\mathbf{u}'$ depend on the current random iterates $(\mathbf{w}, \mathbf{u})$. This dependency introduces additional subtleties into the analysis.

The preceding lemma applies to any random variable v that may depend on the entire randomness of the algorithm, and will be useful for our analysis. Recall that a sequence $\{X_t\}$ is a *martingale difference sequence* if the conditional expectation of each variable given the past is zero, i.e., $\mathbb{E}[X_t \mid X_1, \ldots, X_{t-1}] = 0$.

Proof. Applying Lemma 3.10 to the update of y_{t+1}, we have

$$\mathbb{E}\left[-\delta_t^\top (y_{t+1} - v)\right] \le \mathbb{E}\left[\alpha D_\psi(y, y_t) - \alpha D_\psi(y, y_{t+1}) - \alpha D_\psi(y_{t+1}, y_t)\right].$$

Hence,

$$\begin{aligned}\mathbb{E}\left[\delta_t^\top (v - y_t)\right] &\leq \mathbb{E}\left[\alpha D_\psi(v, y_t) - \alpha D_\psi(v, y_{t+1}) - \alpha D_\psi(y_{t+1}, y_t)\right] \\ &\quad + \mathbb{E}[\delta_t^\top (y_{t+1} - y_t)] \\ &\leq \mathbb{E}\left[\alpha D_\psi(v, y_t) - \alpha D_\psi(v, y_{t+1})\right] \\ &- \mathbb{E}\left[\frac{\alpha\mu_\psi}{2}\|y_{t+1} - y_t\|^2 + \frac{\mu_\psi\alpha}{2}\|y_{t+1} - y_t\|^2 + \frac{1}{2\mu_\psi\alpha}\|\delta_t\|_*^2\right].\end{aligned}$$

Since $\mathbb{E}[\delta_t] = 0$ and y_t is independent of δ_t, we have $\mathbb{E}[\delta_t^\top y_t] = 0$. As a result,

$$\mathbb{E}[\delta_t^\top v] \leq \mathbb{E}\left[\alpha D_\psi(v, y_t) - \alpha D_\psi(v, y_{t+1})\right] + \frac{1}{2\mu_\psi\alpha}\mathbb{E}\left[\|\delta_t\|_*^2\right].$$

□

Theorem 3.13 *Let* $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t, \bar{\mathbf{u}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{u}_t$. *After T iterations, SGDA (3.45) guarantees that*

$$\mathbb{E}[\Delta(\bar{\mathbf{w}}_T, \bar{\mathbf{u}}_T)] \leq \frac{D_1^2}{\eta_1 T} + \frac{D_2^2}{\eta_2 T} + \frac{5\eta_1 G_1^2}{2} + \frac{5\eta_2 G_2^2}{2}.$$

If we set $\eta_1 = O(\frac{D_1}{G_1\sqrt{T}})$ *and* $\eta_2 = O(\frac{D_2}{G_2\sqrt{T}})$, *we have*

$$\mathbb{E}[\Delta(\bar{\mathbf{w}}_T, \bar{\mathbf{u}}_T)] \leq O\left(\frac{D_1 G_1}{\sqrt{T}} + \frac{D_2 G_2}{\sqrt{T}}\right).$$

Proof. Similar to (3.10), for the primal update and dual update for any $\mathbf{w} \in \mathcal{W}, \mathbf{u} \in \mathcal{U}$ we have

$$\begin{aligned}&\partial_1 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t)^\top (\mathbf{w}_t - \mathbf{w}) \leq \\ &\qquad \frac{1}{2\eta_1}\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \frac{1}{2\eta_1}\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 + \frac{1}{2}\eta_1\|\partial_1 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t)\|_2^2 \\ &-\partial_2 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t)^\top (\mathbf{u}_t - \mathbf{u}) \leq \\ &\qquad \frac{1}{2\eta_2}\|\mathbf{u}_t - \mathbf{u}\|_2^2 - \frac{1}{2\eta_2}\|\mathbf{u}_{t+1} - \mathbf{u}\|_2^2 + \frac{1}{2}\eta_2\|\partial_2 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t)\|_2^2.\end{aligned}$$

The difference from the SGD analysis is that we cannot fix $\mathbf{w}$ as $\mathbf{w}_*$ and fix $\mathbf{u}$ as $\mathbf{u}_*$, which will not yield the duality gap measure. Indeed, at the end we need to take max over $\mathbf{w} \in \mathcal{W}$ and min over $\mathbf{u} \in \mathcal{U}$ to obtain the duality gap, making them dependent on the randomness.

To proceed, we have

$$
\begin{aligned}
&\partial_1 f(\mathbf{w}_t, \mathbf{u}_t)^\top (\mathbf{w}_t - \mathbf{w}) \leq \frac{1}{2\eta_1}\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \frac{1}{2\eta_1}\|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2 \\
&+ \frac{1}{2}\eta_1 \|\partial_1 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t)\|_2^2 + (\partial_1 f(\mathbf{w}_t, \mathbf{u}_t) - \partial_1 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t))^\top (\mathbf{w}_t - \mathbf{w}) \\
&- \partial_2 f(\mathbf{w}_t, \mathbf{u}_t)^\top (\mathbf{u}_t - \mathbf{u}) \leq \frac{1}{2\eta_2}\|\mathbf{u}_t - \mathbf{u}\|_2^2 - \frac{1}{2\eta_2}\|\mathbf{u}_{t+1} - \mathbf{u}\|_2^2 \\
&+ \frac{1}{2}\eta_2 \|\partial_2 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t)\|_2^2 + (\partial_2 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t) - \partial_2 f(\mathbf{w}_t, \mathbf{u}_t))^\top (\mathbf{u}_t - \mathbf{u}).
\end{aligned}
$$

Adding these inequalities we have

$$
\begin{aligned}
&\partial_1 f(\mathbf{w}_t, \mathbf{u}_t)^\top (\mathbf{w}_t - \mathbf{w}) - \partial_2 f(\mathbf{w}_t, \mathbf{u}_t)^\top (\mathbf{u}_t - \mathbf{u}) \\
&\leq \frac{1}{2\eta_1}\left(\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2\right) + \frac{1}{2\eta_2}\left(\|\mathbf{u}_t - \mathbf{u}\|_2^2 - \|\mathbf{u}_{t+1} - \mathbf{u}\|_2^2\right) \\
&+ \frac{1}{2}\eta_1 \|\partial_1 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t)\|_2^2 + \frac{1}{2}\eta_2 \|\partial_2 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t)\|_2^2 \\
&+ (\partial_1 f(\mathbf{w}_t, \mathbf{u}_t) - \partial_1 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t))^\top (\mathbf{w}_t - \mathbf{w}) \\
&+ (\partial_2 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t) - \partial_2 f(\mathbf{w}_t, \mathbf{u}_t))^\top (\mathbf{u}_t - \mathbf{u}).
\end{aligned}
$$

Due to the convexity and concavity of $f(\mathbf{w}, \mathbf{u})$ in terms of $\mathbf{w}, \mathbf{u}$, respectively, we have

$$
\begin{aligned}
f(\mathbf{w}_t, \mathbf{u}_t) - f(\mathbf{w}, \mathbf{u}_t) &\leq \partial_1 f(\mathbf{w}_t, \mathbf{u}_t)^\top (\mathbf{w}_t - \mathbf{w}), \\
f(\mathbf{w}_t, \mathbf{u}) - f(\mathbf{w}_t, \mathbf{u}_t) &\leq -\partial_2 f(\mathbf{w}_t, \mathbf{u}_t)^\top (\mathbf{u}_t - \mathbf{u}).
\end{aligned}
$$

Adding these two equalities, we have

$$
f(\mathbf{w}_t, \mathbf{u}) - f(\mathbf{w}, \mathbf{u}_t) \leq \partial_1 f(\mathbf{w}_t, \mathbf{u}_t)^\top (\mathbf{w}_t - \mathbf{w}) - \partial_2 f(\mathbf{w}_t, \mathbf{u}_t)^\top (\mathbf{u}_t - \mathbf{u}).
$$

As a result, we have

$$
\begin{aligned}
&f(\mathbf{w}_t, \mathbf{u}) - f(\mathbf{w}, \mathbf{u}_t) \\
&\leq \frac{1}{2\eta_1}\left(\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \|\mathbf{w}_{t+1} - \mathbf{w}\|_2^2\right) + \frac{1}{2\eta_2}\left(\|\mathbf{u}_t - \mathbf{u}\|_2^2 - \|\mathbf{u}_{t+1} - \mathbf{u}\|_2^2\right) \\
&+ \frac{1}{2}\eta_1 \|\partial_1 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t)\|_2^2 + \frac{1}{2}\eta_2 \|\partial_2 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t)\|_2^2 \\
&+ (\partial_1 f(\mathbf{w}_t, \mathbf{u}_t) - \partial_1 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t))^\top (\mathbf{w}_t - \mathbf{w}) \\
&+ (\partial_2 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t) - \partial_2 f(\mathbf{w}_t, \mathbf{u}_t))^\top (\mathbf{u}_t - \mathbf{u}).
\end{aligned}
$$

Taking average over $t = 1, \ldots, T$, we have

$$
\begin{aligned}
f(\bar{\mathbf{w}}_T, \mathbf{u}) - f(\mathbf{w}, \bar{\mathbf{u}}_T) &\leq \frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{w}_t, \mathbf{u}) - f(\mathbf{w}, \mathbf{u}_t)\right) \\
&\leq \frac{1}{2\eta_1 T}\|\mathbf{w}_1 - \mathbf{w}\|_2^2 + \frac{1}{2\eta_2 T}\|\mathbf{u}_1 - \mathbf{u}\|_2^2 \\
&+ \frac{\eta_1}{2T}\sum_{t=1}^{T}\|\partial_1 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t)\|_2^2 + \frac{\eta_2}{2T}\sum_{t=1}^{T}\|\partial_2 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t)\|_2^2 \\
&+ \frac{1}{T}\sum_{t=1}^{T}(\partial_1 f(\mathbf{w}_t, \mathbf{u}_t) - \partial_1 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t))^\top(\mathbf{w}_t - \mathbf{w}) \\
&+ \frac{1}{T}\sum_{t=1}^{T}(\partial_2 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t) - \partial_2 f(\mathbf{w}_t, \mathbf{u}_t))^\top(\mathbf{u}_t - \mathbf{u}).
\end{aligned}
$$

Let $\mathbf{w}, \mathbf{u}$ be the solution to $\max_{\mathbf{w}\in\mathcal{W}, \mathbf{u}\in\mathcal{U}} f(\bar{\mathbf{w}}_T, \mathbf{u}) - f(\mathbf{w}, \bar{\mathbf{u}}_T)$, which are random variables. Taking expectation over both sides, we have

$$
\begin{aligned}
\mathbb{E}[\Delta(\bar{\mathbf{w}}_T, \bar{\mathbf{u}}_T)] \leq & \frac{1}{2\eta_1 T}\|\mathbf{w}_1 - \mathbf{w}\|_2^2 + \frac{1}{2\eta_2 T}\|\mathbf{u}_1 - \mathbf{u}\|_2^2 + \frac{\eta_1 G_1^2}{2} + \frac{\eta_2 G_2^2}{2} \\
&+ \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}(\partial_1 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t) - \partial_1 f(\mathbf{w}_t, \mathbf{u}_t))^\top \mathbf{w}\right] \\
&+ \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}(\partial_2 f(\mathbf{w}_t, \mathbf{u}_t) - \partial_2 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t))^\top \mathbf{u}\right].
\end{aligned} \tag{3.48}
$$

Next, we apply Lemma 3.13 to bound the last two terms. To this end, we introduce two virtual sequences with $\hat{\mathbf{w}}_1 = \mathbf{w}_1, \hat{\mathbf{u}}_1 = \mathbf{u}_1$:

$$
\begin{aligned}
\hat{\mathbf{w}}_{t+1} &= \arg\min_{\mathbf{w}\in\mathcal{W}} -(\partial_1 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t) - \partial_1 f(\mathbf{w}_t, \mathbf{u}_t))^\top \mathbf{w} + \frac{1}{2\eta_1}\|\mathbf{w} - \hat{\mathbf{w}}_t\|_2^2 \\
\hat{\mathbf{u}}_{t+1} &= \arg\min_{\mathbf{u}\in\mathcal{U}} (\partial_2 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t) - \partial_2 f(\mathbf{w}_t, \mathbf{u}_t))^\top \mathbf{u} + \frac{1}{2\eta_2}\|\mathbf{u} - \hat{\mathbf{u}}_t\|_2^2.
\end{aligned}
$$

Applying Lemma 3.13, we have

$$
\begin{aligned}
&\mathbb{E}\left[(\partial_1 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t) - \partial_1 f(\mathbf{w}_t, \mathbf{u}_t))^\top \mathbf{w}\right] \leq \frac{1}{2\eta_1}\left(\|\hat{\mathbf{w}}_t - \mathbf{w}\|_2^2 - \|\hat{\mathbf{w}}_{t+1} - \mathbf{w}\|_2^2\right) \\
&+ \frac{\eta_1}{2}\mathbb{E}[\|\partial_1 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t) - \partial_1 f(\mathbf{w}_t, \mathbf{u}_t)\|_2^2] \\
&\mathbb{E}\left[(\partial_2 f(\mathbf{w}_t, \mathbf{u}_t) - \partial_2 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t))^\top \mathbf{u}\right] \leq \frac{1}{2\eta_2}\left(\|\hat{\mathbf{u}}_t - \mathbf{u}\|_2^2 - \|\hat{\mathbf{u}}_{t+1} - \mathbf{u}\|_2^2\right) \\
&+ \frac{\eta_2}{2}\mathbb{E}[\|\partial_1 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t) - \partial_1 f(\mathbf{w}_t, \mathbf{u}_t)\|_2^2].
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}(\partial_1 f(\mathbf{w}_t,\mathbf{u}_t;\zeta_t)-\partial_1 f(\mathbf{w}_t,\mathbf{u}_t))^\top \mathbf{w}\right] \\
&+\mathbb{E}\left[\sum_{t=1}^{T}(\partial_2 f(\mathbf{w}_t,\mathbf{u}_t)-\partial_2 f(\mathbf{w}_t,\mathbf{u}_t;\zeta_t))^\top \mathbf{u}\right] \\
&\leq \frac{1}{2\eta_1}\|\hat{\mathbf{w}}_1-\mathbf{w}\|_2^2+\frac{1}{2\eta_2}\|\hat{\mathbf{u}}_t-\mathbf{u}\|_2^2+\frac{4\eta_1 G_1^2 T}{2}+\frac{4\eta_2 G_2^2 T}{2}.
\end{aligned} \tag{3.49}
$$

Combining (3.48) and (3.49), we have

$$
\mathbb{E}[\Delta(\bar{\mathbf{w}}_T,\bar{\mathbf{u}}_T)] \leq \frac{1}{\eta_1 T}\mathbb{E}[\|\mathbf{w}_1-\mathbf{w}\|_2^2]+\frac{1}{\eta_2 T}\mathbb{E}[\|\mathbf{u}_1-\mathbf{u}\|_2^2]+\frac{5\eta_1 G_1^2}{2}+\frac{5\eta_2 G_2^2}{2}.
$$

Hence, we conclude the proof. □

3.7 Stochastic Optimistic Mirror Prox

While simple in design, SGDA cannot enjoy a faster convergence when the function is smooth and the stochastic gradients have zero variance. A classical method to address this limitation is to use an extra-gradient. Let

$$
\mathbf{v}=\begin{bmatrix}\mathbf{w}\\ \mathbf{u}\end{bmatrix}, \quad \mathcal{M}(\mathbf{v})=\begin{bmatrix}\nabla_1 f(\mathbf{w},\mathbf{u})\\ -\nabla_2 f(\mathbf{w},\mathbf{u})\end{bmatrix}, \quad \mathcal{V}=\mathcal{W}\times\mathcal{U}.
$$

The extra-gradient method takes the following update with an initialization of $\mathbf{x}_1 \in \mathcal{V}$:

$$
\begin{aligned}
\mathbf{y}_t &= \arg\min_{\mathbf{v}\in\mathcal{V}} \mathcal{M}(\mathbf{x}_t)^\top \mathbf{v}+\frac{1}{2\eta}\|\mathbf{v}-\mathbf{x}_t\|_2^2 \\
\mathbf{x}_{t+1} &= \arg\min_{\mathbf{v}\in\mathcal{V}} \mathcal{M}(\mathbf{y}_t)^\top \mathbf{v}+\frac{1}{2\eta}\|\mathbf{v}-\mathbf{x}_t\|_2^2.
\end{aligned} \tag{3.50}
$$

The name "extragradient" comes from the fact that it uses two gradients $\mathcal{M}(\mathbf{x}_t)$ and $\mathcal{M}(\mathbf{y}_t)$ at each iteration.

The extragradient method can be generalized using the mirror descent steps with a Bregmand divergence $D_\varphi(\cdot,\cdot)$ defined by a strongly-convex function $\varphi : \mathcal{V} \to \mathbb{R}$:

$$
\begin{aligned}
\mathbf{y}_t &= \arg\min_{\mathbf{v}\in\mathcal{V}} \mathcal{M}(\mathbf{x}_t)^\top \mathbf{v}+\frac{1}{\eta}D_\varphi(\mathbf{v},\mathbf{x}_t) \\
\mathbf{x}_{t+1} &= \arg\min_{\mathbf{v}\in\mathcal{V}} \mathcal{M}(\mathbf{y}_t)^\top \mathbf{v}+\frac{1}{\eta}D_\varphi(\mathbf{v},\mathbf{x}_t).
\end{aligned} \tag{3.51}
$$

This method is called mirror prox.

Both methods can be extended to their stochastic versions. For example, the stochastic mirror prox method (SMP) uses the following update:

Algorithm 8 Stochastic Optimistic Mirror Prox (SOMP)

Input: learning rates η, starting points $\mathbf{x}_1 = \mathbf{y}_0 = (\mathbf{w}_1, \mathbf{u}_1)$
Compute $\mathbf{y}_1 = \arg\min_{\mathbf{v}\in\mathcal{V}} \mathcal{M}(\mathbf{y}_0;\zeta_0)^\top \mathbf{v} + \frac{1}{\eta} D_\varphi(\mathbf{v}, \mathbf{x}_1)$.
for $t = 1, \ldots, T$ **do**
    Compute unbiased gradient mapping $\mathcal{M}(\mathbf{y}_t;\zeta_t)$
    Update $\mathbf{x}_{t+1} = \arg\min_{\mathbf{v}\in\mathcal{V}} \mathcal{M}(\mathbf{y}_t;\zeta_t)^\top \mathbf{v} + \frac{1}{\eta} D_\varphi(\mathbf{v}, \mathbf{x}_t)$.
    Update $\mathbf{y}_{t+1} = \arg\min_{\mathbf{v}\in\mathcal{V}} \mathcal{M}(\mathbf{y}_t;\zeta_t)^\top \mathbf{v} + \frac{1}{\eta} D_\varphi(\mathbf{v}, \mathbf{x}_{t+1})$.
end for

$$\begin{aligned} \mathbf{y}_t &= \arg\min_{\mathbf{v}\in\mathcal{V}} \mathcal{M}(\mathbf{x}_t;\zeta_t')^\top \mathbf{v} + \frac{1}{\eta} D_\varphi(\mathbf{v}, \mathbf{x}_t) \\ \mathbf{x}_{t+1} &= \arg\min_{\mathbf{v}\in\mathcal{V}} \mathcal{M}(\mathbf{y}_t;\zeta_t)^\top \mathbf{v} + \frac{1}{\eta} D_\varphi(\mathbf{v}, \mathbf{x}_t), \end{aligned} \tag{3.52}$$

where $\mathbb{E}_\zeta[\mathcal{M}(\mathbf{x};\zeta)] = \mathcal{M}(\mathbf{x})$.

Stochastic Optimistic Mirror Prox: a variant with a Single Gradient Sequence

The updates of SMP (3.52) need to compute two stochastic gradient sequences $\{\mathcal{M}(\mathbf{x}_t, \zeta_t')\}$ and $\{\mathcal{M}(\mathbf{y}_t;\zeta_t)\}$, which double the costs of SGDA. A simple remedy is to use $\mathcal{M}(\mathbf{y}_{t-1};\zeta_{t-1})$ in the first update of $\mathbf{y}_t$, yielding

$$\begin{aligned} \mathbf{y}_t &= \arg\min_{\mathbf{v}\in\mathcal{V}} \mathcal{M}(\mathbf{y}_{t-1};\zeta_{t-1})^\top \mathbf{v} + \frac{1}{\eta} D_\varphi(\mathbf{v}, \mathbf{x}_t) \\ \mathbf{x}_{t+1} &= \arg\min_{\mathbf{v}\in\mathcal{V}} \mathcal{M}(\mathbf{y}_t;\zeta_t)^\top \mathbf{v} + \frac{1}{2\eta} D_\varphi(\mathbf{v}, \mathbf{x}_t). \end{aligned} \tag{3.53}$$

As a result, we only need to compute one sequence of stochastic gradients $\{\mathcal{M}(\mathbf{y}_t;\zeta_t)\}$. This method is known as stochastic optimistic mirror prox (SOMP).

Let us consider a special case when $\mathcal{V} = \mathbb{R}^d \times \mathbb{R}^{d'}$ and $D_\varphi(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$. The above update reduces to

$$\begin{aligned} \mathbf{y}_t &= \mathbf{x}_t - \eta \mathcal{M}(\mathbf{y}_{t-1};\zeta_{t-1}) \\ \mathbf{x}_{t+1} &= \mathbf{x}_t - \eta \mathcal{M}(\mathbf{y}_t;\zeta_t). \end{aligned} \tag{3.54}$$

This update can be re-written using one sequence of $\{\mathbf{y}_t\}$. By subtracting the second equation from the first one, we have

$$\mathbf{y}_t - \mathbf{x}_{t+1} = \eta \mathcal{M}(\mathbf{y}_t;\zeta_t) - \eta \mathcal{M}(\mathbf{y}_{t-1};\zeta_{t-1}). \tag{3.55}$$

As a result,

$$\begin{aligned} \mathbf{y}_t &= \mathbf{x}_{t+1} + \eta \mathcal{M}(\mathbf{y}_t;\zeta_t) - \eta \mathcal{M}(\mathbf{y}_{t-1};\zeta_{t-1}) \\ &= \mathbf{y}_{t+1} + \eta \mathcal{M}(\mathbf{y}_t;\zeta_t) + \eta \mathcal{M}(\mathbf{y}_t;\zeta_t) - \eta \mathcal{M}(\mathbf{y}_{t-1};\zeta_{t-1}). \end{aligned}$$

From this, we derive that

$$\mathbf{y}_{t+1} = \mathbf{y}_t - \eta(\mathcal{M}(\mathbf{y}_t;\zeta_t) + \mathcal{M}(\mathbf{y}_t;\zeta_t) - \mathcal{M}(\mathbf{y}_{t-1};\zeta_{t-1})). \tag{3.56}$$

This method applied to the min-max problem is known as stochastic optimistic gradient descent ascent (SOGDA), yielding the following primal and dual updates:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta(2\nabla_1 f(\mathbf{w}_t,\mathbf{u}_t;\zeta_t) - \nabla_1 f(\mathbf{w}_{t-1},\mathbf{u}_{t-1};\zeta_{t-1})) \tag{3.57}$$

$$\mathbf{u}_{t+1} = \mathbf{u}_t + \eta(2\nabla_2 f(\mathbf{w}_t,\mathbf{u}_t;\zeta_t) - \eta\nabla_2 f(\mathbf{w}_{t-1},\mathbf{u}_{t-1};\zeta_{t-1}))). \tag{3.58}$$

Convergence Analysis

We analyze the stochastic optimistic mirror prox method in Algorithm 8. We make the following assumption.

Assumption 3.11. *Suppose the following conditions hold:*

(i) $f(\mathbf{w},\mathbf{u})$ *is convex w.r.t* $\mathbf{w}$ *and concave w.r.t* $\mathbf{u}$.
(ii) Let $\varphi(\mathbf{z})$ *be a* α*-strongly convex function with respect to the norm* $\|\cdot\|$*, whose dual norm is denoted by* $\|\cdot\|_*$*,*
(ii) $\mathcal{M}(\mathbf{v})$ *is L-Lipschitz continuous such that*

$$\|\mathcal{M}(\mathbf{v}) - \mathcal{M}(\mathbf{v}')\|_*^2 \leq L^2\|\mathbf{v}-\mathbf{v}'\|^2.$$

(ii) There exist $\sigma_1,\sigma_2 > 0$ *such that*

$$\mathbb{E}_\zeta[\|\mathcal{M}(\mathbf{x};\zeta) - \mathcal{M}(\mathbf{x})\|_*^2] \leq \sigma^2, \forall \mathbf{x}\in\mathcal{V}.$$

(iii) $\max_{\mathbf{x}\in\mathcal{V},\mathbf{x}'\in\mathcal{V}} D_\varphi(\mathbf{x},\mathbf{x}') \leq D^2$.

Lemma 3.14 *Given* $\mathbf{x}$*, consider the updates:*

$$\begin{aligned}\mathbf{y} &= \arg\min_{\mathbf{v}\in\mathcal{V}} \gamma\mathcal{M}(\xi)^\top\mathbf{v} + D_\varphi(\mathbf{v},\mathbf{x}),\\ \mathbf{x}_+ &= \arg\min_{\mathbf{v}\in\mathcal{V}} \gamma\mathcal{M}(\zeta)^\top\mathbf{v} + D_\varphi(\mathbf{v},\mathbf{x}).\end{aligned} \tag{3.59}$$

For any $\mathbf{v}\in\mathcal{V}$*, we have*

$$\begin{aligned}\gamma\mathcal{M}(\zeta)^\top(\mathbf{y}-\mathbf{v}) \leq& D_\varphi(\mathbf{v},\mathbf{x}) - D_\varphi(\mathbf{v},\mathbf{x}_+) + \frac{\gamma^2}{\alpha}\|\mathcal{M}(\xi)-\mathcal{M}(\zeta)\|_*^2\\ &- \frac{\alpha}{2}[\|\mathbf{y}-\mathbf{x}\|^2 + \|\mathbf{y}-\mathbf{x}_+\|^2].\end{aligned} \tag{3.60}$$

Proof. First, by Lemma 3.8, we have

$$\|\mathbf{y}-\mathbf{x}_+\| \leq \frac{\gamma}{\alpha}\|\mathcal{M}(\zeta)-\mathcal{M}(\xi)\|_*. \tag{3.61}$$

Let $\phi(\mathbf{v}) = \gamma\mathcal{M}(\zeta)^\top(\mathbf{y}-\mathbf{v}) - D_\varphi(\mathbf{v},\mathbf{x}) + D_\varphi(\mathbf{v},\mathbf{x}_+)$. Given the optimality condition of $\mathbf{x}_+$, it is easy to verify that it also satisfies the optimality condition of $\max_{\mathbf{v}\in\mathcal{V}} \phi(\mathbf{v})$. As a result, $\phi(\mathbf{v}) \le \phi(\mathbf{x}_+), \forall \mathbf{v} \in \mathcal{V}$, i.e.,

$$\begin{aligned}
&\gamma\mathcal{M}(\zeta)^\top(\mathbf{y}-\mathbf{v}) - D_\varphi(\mathbf{v},\mathbf{x}) + D_\varphi(\mathbf{v},\mathbf{x}_+) \\
&\le \gamma\mathcal{M}(\zeta)^\top(\mathbf{y}-\mathbf{x}_+) - D_\varphi(\mathbf{x}_+,\mathbf{x}) \\
&= \gamma\mathcal{M}(\zeta)^\top(\mathbf{y}-\mathbf{x}_+) + \varphi(\mathbf{x}) + \nabla\varphi(\mathbf{x})^\top(\mathbf{x}_+-\mathbf{x}) - \varphi(\mathbf{x}_+) \\
&= \gamma(\mathcal{M}(\zeta)-\mathcal{M}(\xi))^\top(\mathbf{y}-\mathbf{x}_+) + \gamma\mathcal{M}(\xi)^\top(\mathbf{y}-\mathbf{x}_+) \\
&+ \varphi(\mathbf{x}) + \nabla\varphi(\mathbf{x})^\top(\mathbf{x}_+-\mathbf{x}) - \varphi(\mathbf{x}_+).
\end{aligned} \tag{3.62}$$

By the optimality condition of $\mathbf{y}$, for any $\mathbf{v} \in \mathcal{V}$ we have

$$(\gamma\mathcal{M}(\xi) + \nabla\varphi(\mathbf{y}) - \nabla\varphi(\mathbf{x}))^\top(\mathbf{y}-\mathbf{v}) \le 0$$

Plugging $\mathbf{v} = \mathbf{x}_+$ into the above inequality, we have

$$(\gamma\mathcal{M}(\xi) + \nabla\varphi(\mathbf{y}) - \nabla\varphi(\mathbf{x}))^\top(\mathbf{y}-\mathbf{x}_+) \le 0,$$

which implies that

$$\gamma\mathcal{M}(\xi)^\top(\mathbf{y}-\mathbf{x}_+) \le (\nabla\varphi(\mathbf{y}) - \nabla\varphi(\mathbf{x}))^\top(\mathbf{x}_+-\mathbf{y}).$$

Combining this with (3.62), we have

$$\begin{aligned}
&\gamma\mathcal{M}(\zeta)^\top(\mathbf{y}-\mathbf{v}) - D_\varphi(\mathbf{v},\mathbf{x}) + D_\varphi(\mathbf{v},\mathbf{x}_+) \le \gamma(\mathcal{M}(\zeta)-\mathcal{M}(\xi))^\top(\mathbf{y}-\mathbf{x}_+) \\
&+ (\nabla\varphi(\mathbf{y}) - \nabla\varphi(\mathbf{x}))^\top(\mathbf{x}_+-\mathbf{y}) + \varphi(\mathbf{x}) + \nabla\varphi(\mathbf{x})^\top(\mathbf{x}_+-\mathbf{x}) - \varphi(\mathbf{x}_+) \\
&= \gamma(\mathcal{M}(\zeta)-\mathcal{M}(\xi))^\top(\mathbf{y}-\mathbf{x}_+) \\
&+ \varphi(\mathbf{x}) + \nabla\varphi(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) - \varphi(\mathbf{x}_+) + (\nabla\varphi(\mathbf{y}))^\top(\mathbf{x}_+-\mathbf{y}) \\
&= \gamma(\mathcal{M}(\zeta)-\mathcal{M}(\xi))^\top(\mathbf{y}-\mathbf{x}_+) \\
&+ \varphi(\mathbf{x}) + \nabla\varphi(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) - \varphi(\mathbf{y}) + \varphi(\mathbf{y}) + (\nabla\varphi(\mathbf{y}))^\top(\mathbf{x}_+-\mathbf{y}) - \varphi(\mathbf{x}_+) \\
&= \gamma(\mathcal{M}(\zeta)-\mathcal{M}(\xi))^\top(\mathbf{y}-\mathbf{x}_+) - D_\varphi(\mathbf{y},\mathbf{x}) - D_\varphi(\mathbf{x}_+,\mathbf{y}) \\
&\le \frac{\gamma^2}{\alpha}\|\mathcal{M}(\zeta)-\mathcal{M}(\xi)\|_*^2 - \frac{\alpha}{2}\|\mathbf{y}-\mathbf{x}\|^2 - \frac{\alpha}{2}\|\mathbf{x}_+-\mathbf{y}\|^2,
\end{aligned}$$

where the last inequality uses (3.61) and the α-strong convexity of φ.

□

Theorem 3.14 *Let* $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t, \bar{\mathbf{u}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{u}_t$. *After* T *iterations, SOMP guarantees that*

$$\mathbb{E}[\Delta(\bar{\mathbf{w}}_T, \bar{\mathbf{u}}_T)] \le \frac{2D^2}{T\eta} + \frac{8\sigma^2\eta}{\alpha}.$$

If we set $\eta = \min(\frac{D}{2\sqrt{T}\sigma}, \frac{\alpha}{\sqrt{12}L})$, *we have*

$$\mathbb{E}[\Delta(\bar{\mathbf{w}}_T, \bar{\mathbf{u}}_T)] \leq O\left(\frac{LD^2}{T\alpha} + \frac{\sigma D}{\sqrt{T\alpha}}\right).$$

Why it matters

This result is consistent with the convergence of SGD for smooth convex minimization in Theorem 3.1. In particular, when $\sigma = 0$ (i.e., using the deterministic gradient), the convergence rate simplifies to $O(1/T)$.

Proof. Since the updates of $\mathbf{y}_t, \mathbf{x}_{t+1}$ follow that in (3.59), by applying Lemma 3.14, we have

$$\begin{aligned}
&\eta \mathcal{M}(\mathbf{y}_t, \zeta_t)^\top (\mathbf{y}_t - \mathbf{v}) \leq D_\varphi(\mathbf{v}, \mathbf{x}_t) - D_\varphi(\mathbf{v}, \mathbf{x}_{t+1}) \\
&+ \frac{\eta^2}{\alpha} \|\mathcal{M}(\mathbf{y}_t, \zeta_t) - \mathcal{M}(\mathbf{y}_{t-1}, \zeta_{t-1})\|_*^2 - \frac{\alpha}{2}[\|\mathbf{y}_t - \mathbf{x}_t\|^2 + \|\mathbf{y}_t - \mathbf{x}_{t+1}\|^2] \\
&\leq D_\varphi(\mathbf{v}, \mathbf{x}_t) - D_\varphi(\mathbf{v}, \mathbf{x}_{t+1}) \\
&+ \frac{\eta^2}{\alpha} \|\mathcal{M}(\mathbf{y}_t, \zeta_t) - \mathcal{M}(\mathbf{y}_{t-1}, \zeta_{t-1}) - \mathcal{M}(\mathbf{y}_t) + \mathcal{M}(\mathbf{y}_{t-1}) + (\mathcal{M}(\mathbf{y}_t) - \mathcal{M}(\mathbf{y}_{t-1}))\|_*^2 \\
&- \frac{\alpha}{2}[\|\mathbf{y}_t - \mathbf{x}_t\|^2 + \|\mathbf{y}_t - \mathbf{x}_{t+1}\|^2].
\end{aligned}$$

Let $\sigma_t^2 = \|\mathcal{M}(\mathbf{y}_t, \zeta_t) - \mathcal{M}(\mathbf{y}_t)\|_*^2$, then we have

$$\begin{aligned}
&\|\mathcal{M}(\mathbf{y}_t, \zeta_t) - \mathcal{M}(\mathbf{y}_{t-1}, \zeta_{t-1}) - \mathcal{M}(\mathbf{y}_t) + \mathcal{M}(\mathbf{y}_{t-1}) + (\mathcal{M}(\mathbf{y}_t) - \mathcal{M}(\mathbf{y}_{t-1})\|_*^2 \\
&\leq 3\|\mathcal{M}(\mathbf{y}_t, \zeta_t) - \mathcal{M}(\mathbf{y}_t)\|_*^2 + 3\|\mathcal{M}(\mathbf{y}_{t-1}, \zeta_{t-1}) - \mathcal{M}(\mathbf{y}_{t-1})\|_*^2 \\
&+ 3\|\mathcal{M}(\mathbf{y}_t) - \mathcal{M}(\mathbf{y}_{t-1})\|_*^2 \\
&\leq 3\sigma^2 + 3\sigma^2 + 3L^2\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2.
\end{aligned}$$

Combining the above two inequalities, we have

$$\begin{aligned}
&\eta \mathcal{M}(\mathbf{y}_t, \zeta_t)^\top (\mathbf{y}_t - \mathbf{v}) \leq D_\varphi(\mathbf{v}, \mathbf{x}_t) - D_\varphi(\mathbf{v}, \mathbf{x}_{t+1}) \\
&+ \frac{\eta^2}{\alpha}(6\sigma^2 + 3L^2\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2) - \frac{\alpha}{2}[\|\mathbf{y}_t - \mathbf{x}_t\|^2 + \|\mathbf{y}_t - \mathbf{x}_{t+1}\|^2].
\end{aligned}$$

Taking average over $t = 1, \ldots, T$, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T \mathcal{M}(\mathbf{y}_t)^\top (\mathbf{y}_t - \mathbf{v}) &\leq \frac{1}{T\eta} D_\varphi(\mathbf{v}, \mathbf{x}_1) \\
&+ \frac{\eta}{\alpha T}\sum_{t=1}^T (6\sigma^2 + 3L^2\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2) - \frac{\alpha}{2\eta T}\sum_{t=1}^T [\|\mathbf{y}_t - \mathbf{x}_t\|^2 + \|\mathbf{y}_t - \mathbf{x}_{t+1}\|^2] \\
&+ \frac{1}{T}\sum_{t=1}^T (\mathcal{M}(\mathbf{y}_t) - \mathcal{M}(\mathbf{y}_t, \zeta_t))^\top (\mathbf{y}_t - \mathbf{v}).
\end{aligned}$$

Let $\mathbf{y}_t = (\mathbf{w}_t, \mathbf{u}_t)$ and $\mathbf{v} = (\mathbf{w}, \mathbf{u}) = \arg\max_{\mathbf{w}\in\mathcal{W},\mathbf{u}\in\mathcal{U}} f(\bar{\mathbf{w}}_T, \mathbf{u}) - f(\mathbf{w}, \bar{\mathbf{u}}_T)$. We have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathcal{M}(\mathbf{y}_t)^\top(\mathbf{y}_t - \mathbf{v}) = \frac{1}{T}\sum_{t=1}^{T}(\nabla_1 f(\mathbf{w}_t, \mathbf{u}_t)^\top(\mathbf{w}_t - \mathbf{w}) - \nabla_2 f(\mathbf{w}_t, \mathbf{u}_t)^\top(\mathbf{u}_t - \mathbf{u}))\\
&\geq \frac{1}{T}\sum_{t=1}^{T}(f(\mathbf{w}_t, \mathbf{u}_t) - f(\mathbf{w}, \mathbf{u}_t) + f(\mathbf{w}_t, \mathbf{u}) - f(\mathbf{w}_t, \mathbf{u}_t))\\
&= \frac{1}{T}\sum_{t=1}^{T}(f(\mathbf{w}_t, \mathbf{u}) - f(\mathbf{w}, \mathbf{u}_t)) \geq f(\bar{\mathbf{w}}_T, \mathbf{u}) - f(\mathbf{w}, \bar{\mathbf{u}}_T).
\end{aligned}$$

As a result,

$$\begin{aligned}
&\Delta(\bar{\mathbf{w}}_T, \bar{\mathbf{u}}_T) \leq \frac{1}{T}\sum_{t=1}^{T}\mathcal{M}(\mathbf{y}_t)^\top(\mathbf{y}_t - \mathbf{v}) \leq \frac{1}{T\eta}D_\varphi(\mathbf{v}, \mathbf{x}_1)\\
&+ \frac{\eta}{\alpha T}\sum_{t=1}^{T}(6\sigma^2 + 3L^2\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2) - \frac{\alpha}{2\eta T}\sum_{t=1}^{T}[\|\mathbf{y}_t - \mathbf{x}_t\|^2 + \|\mathbf{y}_t - \mathbf{x}_{t+1}\|^2]\\
&+ \frac{1}{T}\sum_{t=1}^{T}(\mathcal{M}(\mathbf{y}_t) - \mathcal{M}(\mathbf{y}_t, \zeta_t))^\top(\mathbf{y}_t - \mathbf{v}).
\end{aligned}$$

The last term can be bounded by using Lemma 3.13. Define the virtual sequence with $\hat{\mathbf{y}}_1 = \mathbf{x}_1$:

$$\hat{\mathbf{y}}_{t+1} = \arg\min_{\mathbf{v}\in\mathcal{V}}(\mathcal{M}(\mathbf{y}_t) - \mathcal{M}(\mathbf{y}_t, \zeta_t))^\top\mathbf{v} + \frac{1}{\eta}D_\varphi(\mathbf{v}, \hat{\mathbf{y}}_t).$$

Then Lemma 3.13 implies that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\mathcal{M}(\mathbf{y}_t, \zeta_t) - \mathcal{M}(\mathbf{y}_t))^\top\mathbf{v}\right] \leq \mathbb{E}\left[\frac{1}{\eta T}D_\varphi(\mathbf{v}, \hat{\mathbf{y}}_1)\right]\\
&+ \mathbb{E}\left[\frac{\eta}{2\alpha T}\sum_{t=1}^{T}\|\mathcal{M}(\mathbf{y}_t) - \mathcal{M}(\mathbf{y}_t, \zeta_t)\|_*^2\right].
\end{aligned}$$

Combining the above results, we have

$$
\begin{aligned}
\mathbb{E}[\Delta(\bar{\mathbf{w}}_T, \bar{\mathbf{u}}_T)] &\leq \frac{2}{T\eta} D_\varphi(\mathbf{v}, \mathbf{x}_1) + \frac{8\sigma^2\eta}{\alpha} \\
&+ \mathbb{E}\left[\frac{3L^2\eta}{\alpha T}\sum_{t=1}^{T}\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2 - \frac{\alpha}{2\eta T}\sum_{t=1}^{T}[\|\mathbf{y}_t - \mathbf{x}_t\|^2 + \|\mathbf{y}_t - \mathbf{x}_{t+1}\|^2\right] \\
&\leq \frac{2}{T\eta} D_\varphi(\mathbf{v}, \mathbf{x}_1) + \frac{8\sigma^2\eta}{\alpha} + \mathbb{E}\left[\frac{3L^2\eta}{\alpha T}\sum_{t=1}^{T}[2\|\mathbf{y}_t - \mathbf{x}_t\|^2 + 2\|\mathbf{x}_t - \mathbf{y}_{t-1}\|^2]\right] \\
&- \mathbb{E}\left[\frac{\alpha}{2\eta T}\sum_{t=1}^{T}[\|\mathbf{y}_t - \mathbf{x}_t\|^2 + \|\mathbf{y}_t - \mathbf{x}_{t+1}\|^2\right].
\end{aligned}
$$

If $6L^2\frac{\eta}{\alpha} \leq \frac{\alpha}{2\eta}$, i.e., $\eta \leq \frac{\alpha}{\sqrt{12}L}$, the sum of the last two terms will be less than zero due to $\mathbf{x}_1 = \mathbf{y}_0$. Then, we have

$$
\mathbb{E}[\Delta(\bar{\mathbf{w}}_T, \bar{\mathbf{u}}_T)] \leq \frac{2}{T\eta} D_\varphi(\mathbf{v}, \mathbf{x}_1) + \frac{8\sigma^2\eta}{\alpha} \leq \frac{2D^2}{T\eta} + \frac{8\sigma^2\eta}{\alpha}.
$$

For the second part, optimizing the upper bound over η gives $\eta_* = \frac{D\sqrt{\alpha}}{2\sqrt{T}\sigma}$. If $\eta_* \leq \frac{\alpha}{\sqrt{12}L}$, i.e., $T \geq \frac{3D^2L^2}{\sigma^2\alpha}$, we set $\eta = \eta_*$, then

$$
\mathbb{E}\left[F(\bar{\mathbf{w}}_T) - F(\mathbf{w}_*)\right] \leq \frac{8\sigma D}{\sqrt{T\alpha}}.
$$

If $\eta_* > \frac{\alpha}{\sqrt{12}L}$, i.e., $\sigma^2 \leq \frac{3D^2L^2}{\alpha T}$, we set $\eta = \frac{\alpha}{\sqrt{12}L}$, then

$$
\mathbb{E}\left[\Delta(\bar{\mathbf{w}}_T, \bar{\mathbf{u}}_T)\right] \leq \frac{2\sqrt{12}LD^2}{T\alpha} + \frac{12LD^2}{\sqrt{3}T\alpha}.
$$

□

3.8 History and Notes

Stochastic Approximation and Mathematical Optimization

Stochastic approximation has a long history dating back to Robbins and Monro (1951) for solving a root finding problem $f(x) = \alpha$ using an iterative method $x_{t+1} = x_t - a_t(y_t - \alpha)$, where y_t is a stochastic variable such that $\mathbb{E}[y_t] = f(x_t)$. They studied the asymptotic convergence of $\lim_{t\to\infty}\mathbb{E}[(x_t - \theta)^2] = 0$ under some conditions, where θ is the solution to the root finding problem. It is notable that Herbert Robbins was regarded as one of the most influential mathematicians of the latter half of the 20th century, renowned for his seminal contributions to probability, algebra, and graph theory.

Inspired by Robbins and Monro (1951), Kiefer and Wolfowitz (1952) considered stochastic maximization of a regression function using a stochastic finite difference estimator of the gradient. Later, Chung (1954) established the convergence bound of Robbins-Monro's method under some conditions. Since then, the convergence of SGD has been extensively studied. Polyak and Juditsky (1992) analyzed the convergence of SGD with a simple averaging for stochastic optimization, which is sometimes referred to as Polyak-Juditsky averaging or Polyak averaging. This work assumes smoothness and strong convexity of the objective function and established a convergence rate of $O(1/T)$.

Nemirovski and Yudin (1978) is probably the first work that analyzes the non-asymptotic convergence of SGDA for general convex-concave min-max optimization without smoothness and strong convexity assumption. Their paper introduces the weighted averaging (weighted by the step size at each iteration) and establishes the convergence rate of $O(1/\sqrt{T})$. The optimal rate $O(1/T)$ for strongly-convex strongly-concave min-max problem was recently proved in Yan et al. (2020a).

The mirror descent method originates from Nemirovsky and Yudin (1983), which is also the work that is often cited for the lower bound of $O(1/\sqrt{T})$ for general convex problems. A more general form of SMD and its extension for convex-concave min-max problems using a Bregman divergence was later considered in (Nemirovski et al., 2009).

The non-asymptotic analysis of SGD for non-convex optimization was initiated by (Ghadimi and Lan, 2013). The non-asymptotic analysis of SGD for weakly convex optimization was developed by (Davis and Drusvyatskiy, 2019).

The proximal method dates back to the proximal point method proposed by Martinet (1972) and further developed in (Rockafellar, 1976). Lions and Mercier (1979) proposed a splitting method for finding a zero point of the sum of two maximal monotone operators. The forward backward splitting was first proposed by Pazy (1979) in the same context of finding a zero of sum of monotone operators. Its special instance for minimization problems known as projected gradient method was first studied by Goldstein (1964).

Coordinate descent has a long history in optimization, with its earliest roots traceable to the Gauss–Seidel iterations for solving linear systems in the 19th century. The method was later formalized and discussed in early optimization literature, including (Warga, 1963; Ortega and Rheinboldt, 1970; Luenberger, 1973). Rigorous analysis of convergence properties was developed in a sequence of influential works by Paul Tseng and others, including (Luo and Tseng, 1992; Tseng, 1990; Tseng and Bertsekas, 1987; Tseng, 2001). Recent developments of block coordinate descent including accelerated coordinate descent (Nesterov, 2012) and stochastic block coordinate descent (Dang and Lan, 2015).

The extragradient method was first proposed by Korpelevich (1976). The mirror prox method and its convergence rate $O(1/T)$ was proposed and established by Nemirovski (2004). The stochastic mirror prox method was analyzed in (Juditsky et al., 2011).

Optimization in machine learning

Frank Rosenblatt's pioneering work in the late 1950s introduced a learning rule for updating the Perceptron model (a single-layer neural network for binary classification) (Rosenblatt, 1962), a method that shares a conceptual foundation with modern stochastic gradient descent (SGD). The earliest instance of SGD for machine learning is perhaps the Widrow-Hoff algorithm (Widrow and Hoff, 1960) (also known as the least mean square' algorithm), which was used to train ADALINE - a single-layer neural network that produces a continuous output. Amari (1967) is the first work that applies SGD to optimize a neural network for binary and multi-class classification.

Starting in late 1980s, online prediction problem has attracted increasing attention in machine learning, whose developments have many parallels to stochastic optimization. Littlestone (1988) proposed the Winnow algorithm for learning Boolean functions. It was shown to be better than the earlier Perceptron learning algorithm in the sense that the number of mistakes grows only logarithmically with the number of irrelevant attributes in the examples. Later, it was generalized to weighted majority for learning with expert advice (Littlestone and Warmuth, 1994), and the exponentiated gradient method (Kivinen and Warmuth, 1997) for online learning with a decision variable from a simplex, which is a special case of SMD using the KL-divergence. It has impact on the development of Adaboost (Freund and Schapire, 1997).

During the first decade of the 21st century, online convex optimization emerged as a central topic in machine learning. A key focus was on regret bound analysis, which can be transferred into convergence guarantees for stochastic optimization via the online-to-batch conversion technique (Dekel and Singer, 2005). Regret bounds for online gradient descent were established for both convex loss functions (Zinkevich, 2003) and strongly convex loss functions (Hazan et al., 2007). The multi-epoch scheme for achieving an optimal rate of $O(1/T)$ for stochastic strongly convex optimization was established independently and concurrently in (Iouditski and Nesterov, 2010; Hazan and Kale, 2011; Ghadimi and Lan, 2012). Later, SGD has shown to be able to achieve the optimal rate for stochastic non-smooth strongly convex optimization using tail averaging (Rakhlin et al., 2012) or increased weighted averaging (Lacoste-Julien et al., 2012). The last iterate convergence of SGD for non-smooth convex optimization was established in (Shamir and Zhang, 2013).

The use of the ℓ_1 norm as a regularizer in the Lasso method was pioneered by Tibshirani (1996). The elastic net regularizer was later proposed in (Zou and Hastie, 2003), while the group lasso regularizer was introduced by (Yuan and Lin, 2006). More recently, the Piecewise Affine Regularizer (PAR) was proposed in (Jin et al., 2025). The nuclear norm minimization for promoting a low-rank matrix was first considered in (Fazel et al., 2001).

Pioneering works on the application of SGD to regularized empirical risk minimization in machine learning, including support vector machines, include (Zhang, 2004a; Shalev-Shwartz et al., 2007). The application of the proximal gradient method to ℓ_1 norm regularized problem was initiated by Daubechies et al. (2004), yielding an algorithm known as iterative thresholding. The application of SPGD to machine

learning with various regularization terms was studied in (Duchi and Singer, 2009). The application of SGD for optimizing truncated loss and its theory was studied in (Xu et al., 2019b).

The most famous application of coordinate descent methods in machine learning is the solver for support vector machine (Chang et al., 2008; Hsieh et al., 2008).

AdaGrad, proposed by Duchi et al. (2011), was a breakthrough in stochastic optimization for machine learning. It later inspired several popular stochastic algorithms for deep learning, including RMSprop (Hinton, 2018) and Adam (Kingma and Ba, 2015), which will be discussed in Chapter 6.

The first variant of stochastic optimistic mirror prox method appeared in the author's award-winning work (Chiang et al., 2012), inspired by Nemirovski's mirror prox method. It was introduced to address a long-standing challenge in online convex optimization for achieving variational regret bounds. This line of research later inspired the work of (Rakhlin and Sridharan, 2013), which formally coined the term optimistic mirror descent. More recently, stochastic optimistic mirror prox has been adopted for solving min–max problems in machine learning, including applications such as training generative adversarial networks (Daskalakis et al., 2018).

Discussion. The most important factor that affects the practical performance of SGD and other stochastic algorithms is the learning rate scheme η_t. In this chapter, we focus on a fixed learning rate $\eta_t = \eta$. However, it is usually not the best choice in practice. We can also develop theoretical analysis of these algorithms using decreasing learning rates, e.g., $\eta_t \propto 1/\sqrt{t}, 1/t$. However, these theoretical learning rate schemes are usually also not the best in practice. A practical approach is the step decay strategy as in Theorem 3.7, which gives a convergence that has only logarithmic dependence on the initial distance $\|\mathbf{w}_1 - \mathbf{w}_*\|_2$. This strategy also works for general stochastic convex optimization under generic error bound conditions in the form $\|\mathbf{w} - \mathbf{w}_*\|_2 \leq c(g(\mathbf{w}) - g(\mathbf{w}_*))^\theta$ with $\theta \in (0, 1]$ (Xu et al., 2017). Another issue of theoretical learning rates is that their best values that optimize the convergence bound may depend on some unknown parameters of the problem, e.g., $\mathbf{w}_*$, the smoothness constant, strong convexity modulus. This issue has triggered a line of research known as parameter-free algorithms (Orabona, 2019; Lan et al., 2023).

While this chapter focuses on classical stochastic methods that not only have important applications in machine learning but also significantly influence the approaches presented in later chapters, it does not cover several important algorithms, most notably accelerated gradient methods and their stochastic variants. These methods achieve optimal convergence rates for smooth convex objectives when the variance of stochastic gradients vanishes (Lan, 2012). For a comprehensive treatment of accelerated gradient methods, we refer to the textbook by Nesterov (2004), and for stochastic accelerated algorithms, we recommend Lan (2020). Variants of these methods will be introduced in Chapter 6.

Finally, I recommend the textbook (Recht and Wright, 2025), which provides a comprehensive treatment of convex optimization algorithms tailored for data analysis.