# Chapter 5
# Advances: Finite-sum Coupled Compositional Optimization

**Abstract** In this chapter, we study a novel family of stochastic compositional optimization problems namely **finite-sum coupled compositional optimization (FCCO)**, and introduce algorithms for solving them. These algorithms have direct applications in addressing the empirical X-risk minimization challenges discussed in Chapter 2. To ensure broad applicability, we examine various settings of this problem, characterized by different properties of outer and inner functions, including smooth and non-smooth cases, as well as convex, weakly convex, and non-convex scenarios. The results presented here also significantly extend and complement those discussed in Chapter 4. We also discuss how to efficiently optimize compositional optimized certainty equivalent risks, especially compositional entropic risk.

*Coupling reveals depth where composition meets reality!*

## Contents

## 5.1 Finite-sum Coupled Compositional Optimization

Specifically, we focus on the following optimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} f_i\big(\mathbb{E}_{\zeta\sim\mathbb{P}_i} g_i(\mathbf{w};\zeta)\big), \tag{5.1}$$

where $g_i(\cdot;\zeta) : \mathbb{R}^d \to \mathbb{R}^{d'}$ is a stochastic mapping, $f_i(\cdot) : \mathbb{R}^{d'} \to \mathbb{R}$ is a deterministic function, and $\mathbb{P}_i$ denotes the distribution of the random variable $\zeta$.

We refer to this problem as **finite-sum coupled compositional optimization (FCCO)**. If we interpret $i$ as an outer random variable, a distinctive feature that sets FCCO apart from standard stochastic compositional optimization (SCO) is that each inner stochastic function $g_i(\mathbf{w};\zeta)$ depends on both an inner random variable $\zeta$ and an outer index $i$, giving rise to the term *coupled*. While this problem can be cast as a special case of SCO by defining $f(\mathbf{g}) = \frac{1}{n}\sum_{i=1}^{n} f_i(g_i)$ and $\mathbf{g}(\mathbf{w}) = [g_1(\mathbf{w}), \ldots, g_n(\mathbf{w})]$, the high dimensionality of $\mathbf{g}$ due to large $n$, along with its stochastic components, significantly complicates the construction of unbiased estimators and theoretical analysis. Therefore, FCCO warrants the development of specialized optimization methods.

Below, we revisit several applications of FCCO in ML and discuss the properties of $f_i$ and $g_i$.

### Group DRO

In Section 2.2.3, we have formulated the CVaR divergence regularized group DRO as

$$\min_{\mathbf{w},\nu} \frac{1}{K\alpha}\sum_{i=1}^{K}[L_i(\mathbf{w}) - \nu]_+ + \nu, \tag{5.2}$$

where $\alpha \in (0,1)$, $L_i(\mathbf{w}) = \frac{1}{n_k}\sum_{j=1}^{n_k} \ell(\mathbf{w};\mathbf{x}_j^i, y_j^i)$ denotes the average loss over data from the $i$-th group. The first term above is an instance of the FCCO objective, where the outer function $f(g) = ([g]_1 - [g]_2)_+$ is a **convex but non-smooth** function of $g$, and each inner function $g_i(\mathbf{w},\nu) = [L_i(\mathbf{w}), \nu]^\top$ could be convex or non-convex, smooth or non-smooth depending on applications.

### AP Maximization

In Section 2.3.2, the AP maximization has been formulated as the following problem:

$$\min_{\mathbf{w}} \frac{1}{n_+}\sum_{\mathbf{x}_i\in\mathcal{S}_+} f(g_i(\mathbf{w})), \tag{5.3}$$

where $\mathcal{S}_+$ is the set of $n_+$ positive examples, $g_i(\mathbf{w}) = [g_1(\mathbf{w}; \mathbf{x}_i, \mathcal{S}), g_2(\mathbf{w}; \mathbf{x}_i, \mathcal{S})]^\top$ is a vector mapping with two components:

$$g_1(\mathbf{w}; \mathbf{x}_i, \mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x}_j \in \mathcal{S}} \mathbb{I}(y_j = 1) \ell(h(\mathbf{w}; \mathbf{x}_j) - h(\mathbf{w}; \mathbf{x}_i))$$
$$g_2(\mathbf{w}; \mathbf{x}_i, \mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x}_j \in \mathcal{S}} \ell(h(\mathbf{w}; \mathbf{x}_j) - h(\mathbf{w}; \mathbf{x}_i)),$$

and $f(\mathbf{g}) = -\frac{[\mathbf{g}]_1}{[\mathbf{g}]_2}$ is simple function. We can see that $f$ is **non-convex and smooth** if the loss value is upper bounded and $\ell(0)$ is lower bounded. The inner mapping $g_i(\mathbf{w})$ could be convex (e.g., a linear model) or non-convex (e.g., a deep model), smooth or non-smooth depending on applications.

#### Contrastive Representation Learning

The contrastive objective of self-supervised representation learning presented in (2.50), is the following:

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \tau \log \left( \varepsilon + \frac{1}{|\mathcal{S}_i^-|} \sum_{\mathbf{y} \in \mathcal{S}_i^-} \exp((s(\mathbf{w}; \mathbf{x}_i, \mathbf{y}) - s(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_i^+))/\tau) \right).$$

The outer function $f(g) = \tau \log(\varepsilon + g)$ is a non-convex function and smooth when $\varepsilon$ is lower bounded. Each inner function $g_i$ is a non-convex function of $\mathbf{w}$ in general.

## 5.2 Smooth Functions

In this section, we consider a non-convex but smooth objective function $F(\mathbf{w})$ with smooth outer functions. In addition, we assume the inner stochastic functions satisfy the following conditions throughout this section.

**Assumption 5.1.** *We assume that*

*(i)* $\mathbb{E}_{\zeta \sim \mathbb{P}_i}[\|g_i(\mathbf{w}; \zeta) - g_i(\mathbf{w})\|_2^2] \leq \sigma_0^2$.
*(ii)* $\mathbb{E}_{\zeta \sim \mathbb{P}_i}[\|\nabla g_i(\mathbf{w}; \zeta) - \nabla g_i(\mathbf{w})\|_2^2] \leq \sigma_2^2$.
*(iii)* $\mathbb{E}_{\zeta \sim \mathbb{P}_i}[\|\nabla g_i(\mathbf{w}; \zeta)\|_2^2] \leq G_2^2$.

### 5.2.1 The SOX Algorithm

The first algorithm for solving FCCO is called SOX, named by **S**tochastic **O**ptimization of **X**-risks. Owing to its ease of implementation and favorable practical performance, this algorithm is commonly adopted for addressing FCCO. Below, we outline the assumptions necessary for its analysis.

**Assumption 5.2.** *There exist* $G_1, L_1, L_F > 0$ *such that*

*(i)* $f_i : \mathbb{R}^{d'} \mapsto \mathbb{R}$ *is* $G_1$*-Lipschitz continuous and* $L_1$*-smooth;*
*(ii)* $F : \mathbb{R}^d \mapsto \mathbb{R}$ *is* $L_F$*-smooth;*
*(iii)* $F_* = \min_{\mathbf{w}} F(\mathbf{w}) \geq > -\infty$.

Similar to that for SCO, we also need to track and estimate the inner functions. However, the difference is that we need to maintain and update $n$ estimators for the $n$ inner functions $g_i(\mathbf{w}), i \in [n]$.

To this end, we maintain $n$ sequence of estimators $\{\mathbf{u}_{i,t}, t \in [T]\}_{i=1}^n$. At the $t$-th iteration, we draw a set of $B$ random indices $\mathcal{B}_t \subset [n]$ with $|\mathcal{B}_t| = B$. We update $\mathbf{u}_{i,t}, i \in [n]$ by the following:

$$\mathbf{u}_{i,t} = \begin{cases} (1-\gamma_t)\mathbf{u}_{i,t-1} + \gamma_t g_i(\mathbf{w}_t; \zeta_{i,t}), & i \in \mathcal{B}_t \\ \mathbf{u}_{i,t-1}, & \text{o.w.} \end{cases}, t = 1, \ldots, T, \tag{5.4}$$

where $\zeta_{i,t} \sim \mathbb{P}_i$ is a random variable. We refer to the above estimator as coordinate moving average estimator. Then, similar to SCMA, a moving average estimator of the gradient is computed by:

$$\begin{aligned} &\mathbf{v}_t = (1-\beta_t)\mathbf{v}_{t-1} + \beta_t \mathbf{z}_t, \\ &\text{where} \quad \mathbf{z}_t = \frac{1}{|\mathcal{B}_t|} \sum_{i \in \mathcal{B}_t} \nabla g_i(\mathbf{w}_t; \zeta'_{i,t}) \nabla f_i(\mathbf{u}_{i,t}). \end{aligned}$$

Then, the model parameters are updated by:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{v}_t.$$

The detailed steps are presented in Algorithm 14.

#### Convergence Analysis

Let us first define two notations:

$$\Delta_t = \|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2, \tag{5.5}$$

$$\delta_t = \frac{1}{n} \sum_{i=1}^n \left\|\mathbf{u}_{i,t} - g_i(\mathbf{w}_t)\right\|_2^2. \tag{5.6}$$

**Algorithm 14** SOX

**Input:** learning rate schedules $\{\eta_t\}_{t=1}^T$, $\{\gamma_t\}_{t=1}^T$, $\{\beta_t\}_{t=1}^T$; starting points $\mathbf{w}_0$, $\mathbf{u}_0$, $\mathbf{v}_0$
Let $\mathbf{w}_1 = \mathbf{w}_0 - \eta_0 \mathbf{v}_0$
**for** $t = 1, \ldots, T$ **do**
  Draw a batch of samples $\mathcal{B}_t \subset [n]$
  **for** $i \in \mathcal{B}_t$ **do**
    Draw two samples $\zeta_{i,t}, \zeta'_{i,t} \sim \mathbb{P}_i$
    Update the inner function value estimators

$$\mathbf{u}_{i,t} = (1 - \gamma_t)\mathbf{u}_{i,t-1} + \gamma_t g_i(\mathbf{w}_t; \zeta_{i,t}),$$

  **end for**
  Set $\mathbf{u}_{i,t} = \mathbf{u}_{i,t-1}, i \notin \mathcal{B}_t$
  Compute the vanilla gradient estimator $\mathbf{z}_t = \frac{1}{|\mathcal{B}_t|} \sum_{i \in \mathcal{B}_t} \nabla g_i(\mathbf{w}_t; \zeta'_{i,t}) \nabla f_i(\mathbf{u}_{i,t})$
  Update the MA gradient estimator $\mathbf{v}_t = (1 - \beta_t)\mathbf{v}_{t-1} + \beta_t \mathbf{z}_t$
  Update the model $\mathbf{w}$ by $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{v}_t$
**end for**

The descent lemma (Lemma 4.9) remains valid. Next, we analyze the recursion of $\Delta_t$ and $\delta_t$. One point of deviation is that only some randomly selected coordinates of $\mathbf{u}$ are updated and used for computing the gradient estimator $\mathbf{z}_t$. To facilitate the proof, we introduce a virtual sequence:

$$\bar{\mathbf{u}}_{i,t} = (1 - \gamma_t)\mathbf{u}_{i,t-1} + \gamma_t g_i(\mathbf{w}_t; \zeta_{i,t}), \forall i = 1, \ldots, n. \tag{5.7}$$

This is similar to that is done in the analysis of stochastic coordinate descent method in Section 3.3. Then, we have

$$\mathcal{M}_t = \mathbb{E}_{\mathcal{B}_t, \zeta'_t}[\mathbf{z}_t] = \frac{1}{n} \sum_{i=1}^n \nabla g_i(\mathbf{w}_t) \nabla f_i(\bar{\mathbf{u}}_{i,t}).$$

**Critical:** Since $\mathbf{u}_t$ is a random variable that depends on $\mathcal{B}_t$, hence

$$\mathbb{E}_{\mathcal{B}_t, \zeta'_t}[\mathbf{z}_t] \neq \frac{1}{n} \sum_{i=1}^n \nabla g_i(\mathbf{w}_t) \nabla f_i(\mathbf{u}_{i,t}).$$

We first bound the error recursion of $\delta_t$.

**Lemma 5.1** *Consider the* $\mathbf{u}_t$ *updates in Algorithm 14. Under Assumption 5.1, if* $\gamma_t \leq 1$*, then*

$$\mathbb{E}[\delta_t] \leq \left(1 - \frac{B\gamma_t}{2n}\right) \mathbb{E}[\delta_{t-1}] + \frac{2nG_2^2}{B\gamma_t} \mathbb{E}\left[\|\mathbf{w}_{t-1} - \mathbf{w}_t\|_2^2\right] + \frac{B\gamma_t^2 \sigma_0^2}{n}.$$

*Proof.* Since $\bar{\mathbf{u}}_{i,t}$ is updated using MA, then similar to (4.6), for all $i \in [n]$ we have

$$\mathbb{E}_{\zeta_{i,t}}[\|\bar{\mathbf{u}}_{i,t} - g_i(\mathbf{w}_t)\|_2^2] \le (1-\gamma_t)^2\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_t)\|_2^2 + \gamma_t^2\sigma_0^2.$$

Given $i \in [n]$, with a probability of $B/n$ that $i \in \mathcal{B}_t$, we have $\mathbf{u}_{i,t} = \bar{\mathbf{u}}_{i,t}$; otherwise, $\mathbf{u}_{i,t} = \mathbf{u}_{i,t-1}$. Hence,

$$\begin{aligned}
&\mathbb{E}_{\zeta_{i,t}}\mathbb{E}_{\mathcal{B}_t}[\|\mathbf{u}_{i,t} - g_i(\mathbf{w}_t)\|_2^2] \\
&= \frac{B}{n}\mathbb{E}_{\zeta_{i,t}}[\|\bar{\mathbf{u}}_{i,t} - g_i(\mathbf{w}_t)\|_2^2] + (1-\frac{B}{n})\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_t)\|_2^2 \\
&\le \frac{B}{n}(1-\gamma_t)^2\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_t)\|_2^2 + \frac{B\gamma_t^2\sigma_0^2}{n} + (1-\frac{B}{n})\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_t)\|_2^2 \\
&\le (1-\frac{B\gamma_t}{2n})^2\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_t)\|_2^2 + \frac{B\gamma_t^2\sigma_0^2}{n},
\end{aligned}$$

where we use the fact $\frac{B}{n}(1-\gamma_t)^2 + (1-\frac{B}{n}) \le (1-\frac{\gamma_t B}{2n})^2$. Then, taking expectation over all randomness on both sides yields

$$\mathbb{E}\left[\|\mathbf{u}_{i,t} - g_i(\mathbf{w}_t)\|_2^2\right] \le (1-\frac{B\gamma_t}{2n})^2\mathbb{E}\left[\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_t)\|_2^2\right] + \frac{B\gamma_t^2\sigma_0^2}{n}.$$

Then using the Young's inequality similar to the proof of Lemma 4.1, we have

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{u}_{i,t} - g_i(\mathbf{w}_t)\|_2^2\right] &\le (1+\frac{B\gamma_t}{2n})(1-\frac{B\gamma_t}{2n})^2\mathbb{E}\left[\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1})\|_2^2\right] \\
&+ (1+\frac{2n}{B\gamma_t})(1-\frac{B\gamma_t}{2n})^2\mathbb{E}\left[\|g_i(\mathbf{w}_{t-1}) - g_i(\mathbf{w}_t)\|_2^2\right] + \frac{B\gamma_t^2\sigma_0^2}{n} \\
&\le (1-\frac{B\gamma_t}{2n})\mathbb{E}\left[\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1})\|_2^2\right] + \frac{2nG_2^2}{B\gamma_t}\mathbb{E}\left[\|\mathbf{w}_{t-1} - \mathbf{w}_t\|_2^2\right] + \frac{B\gamma_t^2\sigma_0^2}{n},
\end{aligned}$$

where we use $\gamma_t \le 1 < \frac{2n}{B}$. The desired result follows by taking average over $i = 1, \ldots, n$ on both sides.

□

**Lemma 5.2 (Variance of $\mathbf{z}_t$)** *Let* $\sigma^2 = \frac{G_1^2\sigma_2^2}{B} + \frac{G_1^2G_2^2}{B}\frac{n-B}{n-1}$. *We have*

$$\mathbb{E}_t\left[\|\mathbf{z}_t - \mathcal{M}_t\|_2^2\right] \le \sigma^2.$$

*Proof.* First, using the variance bound of the average of $B$ independent zero-mean random variables gives

$$A_1 = \mathbb{E}_t\left[\left\|\frac{1}{B}\sum_{i\in\mathcal{B}_t}\nabla g_i(\mathbf{w}_t;\zeta'_{i,t})\nabla f_i(\bar{\mathbf{u}}_{i,t}) - \frac{1}{B}\sum_{i\in\mathcal{B}_t}\nabla g_i(\mathbf{w}_t)\nabla f_i(\bar{\mathbf{u}}_{i,t})\right\|_2^2\right] \le \frac{G_1^2\sigma_2^2}{B},$$

and using the variance bound of $B$ random variables without replacement yields

$$A_2 = \mathbb{E}_t \left[ \left\| \frac{1}{B} \sum_{i \in \mathcal{B}_t} \nabla g_i(\mathbf{w}_t) \nabla f_i(\bar{\mathbf{u}}_{i,t}) - \frac{1}{n} \sum_{i=1}^{n} \nabla g_i(\mathbf{w}_t) \nabla f_i(\bar{\mathbf{u}}_{i,t}) \right\|_2^2 \right] \leq \frac{G_1^2 G_2^2}{B} \frac{n-B}{n-1}.$$

As a result,

$$\begin{aligned}
&\mathbb{E}_t \left[ \|\mathbf{z}_t - \mathcal{M}_t\|_2^2 \right] \\
&= \mathbb{E}_t \left[ \left\| \frac{1}{B} \sum_{i \in \mathcal{B}_t} \nabla g_i(\mathbf{w}_t; \zeta'_{i,t}) \nabla f_i(\bar{\mathbf{u}}_{i,t}) - \frac{1}{n} \sum_{i=1}^{n} \nabla g_i(\mathbf{w}_t) \nabla f_i(\bar{\mathbf{u}}_{i,t}) \right\|_2^2 \right] \\
&= A_1 + A_2 \leq \frac{G_1^2 \sigma_2^2}{B} + \frac{G_1^2 G_2^2}{B} \frac{n-B}{n-1} := \sigma^2.
\end{aligned}$$

□

**Lemma 5.3** *Under Assumptions 5.1 and 5.2, if $\beta_t \leq 1$, the gradient estimation error $\Delta_t$ can be bounded as*

$$\begin{aligned}
\mathbb{E}[\Delta_t] \leq (1-\beta_t)\mathbb{E}\left[\Delta_{t-1}\right] + \frac{2L_F^2 + 8\beta_t^2 G_2^4 L_1^2}{\beta_t} \mathbb{E}\left[\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2\right] + 8\beta_t L_1^2 G_2^2 \mathbb{E}\left[\delta_{t-1}\right] \\
+ \beta_t^2 \sigma^2 + 4G_2^2 L_1^2 \beta_t \gamma_t^2 \sigma_0^2,
\end{aligned}$$

*where* $\sigma^2 = \frac{G_1^2 \sigma_2^2}{B} + \frac{G_1^2 G_2^2}{B} \frac{n-B}{n-1}$.

*Proof.* Since $\mathbf{v}_t$ is updated using MA, we apply Lemma 4.7 in light of Lemma 5.2, yielding

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2\right] \leq (1-\beta_t)\mathbb{E}[\|\mathbf{v}_{t-1} - \nabla F(\mathbf{w}_{t-1})\|_2^2] \\
+ \frac{2L_F^2}{\beta_t}\mathbb{E}[\|\mathbf{w}_{t-1} - \mathbf{w}_t\|_2^2] + 4\beta_t \mathbb{E}[\|\mathcal{M}_t - \nabla F(\mathbf{w}_t)\|_2^2] + \beta_t^2 \sigma^2.
\end{aligned} \tag{5.8}$$

Next, we bound $\mathbb{E}[\|\mathcal{M}_t - \nabla F(\mathbf{w}_t)\|_2^2]$.

$$\begin{aligned}
\|\mathcal{M}_t - \nabla F(\mathbf{w}_t)\|_2^2 &= \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla g_i(\mathbf{w}_t) \nabla f(\bar{\mathbf{u}}_{i,t}) - \frac{1}{n} \sum_{i=1}^{n} \nabla g_i(\mathbf{w}_t) \nabla f(g_i(\mathbf{w}_t)) \right\|_2^2 \\
&\leq G_2^2 L_1^2 \frac{1}{n} \sum_{i=1}^{n} \|\bar{\mathbf{u}}_{i,t} - g_i(\mathbf{w}_t)\|_2^2.
\end{aligned}$$

From Lemma 5.1, we have

$$\mathbb{E}_{\zeta_{i,t}}[\|\bar{\mathbf{u}}_{i,t} - g_i(\mathbf{w}_t)\|_2^2] \leq (1-\gamma_t)^2 \|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_t)\|_2^2 + \gamma_t^2 \sigma_0^2, \forall i.$$

Hence

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\|\bar{\mathbf{u}}_{i,t}-g_i(\mathbf{w}_t)\|_2^2\right] \le (1-\gamma_t)^2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{u}_{i,t-1}-g_i(\mathbf{w}_t)\|_2^2\right]+\gamma_t^2\sigma_0^2$$
$$\le (1-\gamma_t)^2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(2\|g_i(\mathbf{w}_t)-g_i(\mathbf{w}_{t-1})\|_2^2+2\|\mathbf{u}_{i,t-1}-g_i(\mathbf{w}_{t-1})\|_2^2)\right]+\gamma_t^2\sigma_0^2$$
$$\le 2\mathbb{E}[\delta_{t-1}]+\gamma_t^2\sigma_0^2+\mathbb{E}\left[2G_2^2\|\mathbf{w}_{t-1}-\mathbf{w}_t\|_2^2\right].$$

As a result,

$$\mathbb{E}[\|\mathcal{M}_t-\nabla F(\mathbf{w}_t)\|_2^2] \le 2G_2^2L_1^2\mathbb{E}[\delta_{t-1}]+G_2^2L_1^2\gamma_t^2\sigma_0^2+\mathbb{E}\left[2G_2^4L_1^2\|\mathbf{w}_{t-1}-\mathbf{w}_t\|_2^2\right].$$

Plugging the above results into (5.8) we finish the proof.

□

For combining the descent lemma and the above lemmas, we present a result similar to Lemma 4.10, with differences highlighted in boxes.

**Lemma 5.4** *If $\eta_t \le 1/L$, assume that there exist non-negative sequences $A_t, B_t, \Gamma_t, \Delta_t, \delta_t, t \ge 0$ satisfying:*

$$(*)A_{t+1} \le A_t+\eta_t\Delta_t-\eta_tB_t-\eta_t\Gamma_t$$
$$(\sharp)\Delta_{t+1} \le (1-\beta_{t+1})\Delta_t+C_1\beta_{t+1}\boxed{\delta_t}+\frac{C_2\eta_t^2}{\beta_{t+1}}\Gamma_t+\beta_{t+1}^2\sigma^2+\boxed{\beta_{t+1}\gamma_{t+1}^2\sigma'^2},$$
$$(\diamond)\delta_{t+1} \le (1-\gamma_{t+1})\delta_t+\frac{C_3\eta_t^2}{\gamma_{t+1}}\Gamma_t+\gamma_{t+1}^2\sigma'^2.$$

*If $\beta = \frac{\epsilon^2}{4\sigma^2}, \gamma = \min(\frac{\epsilon^2}{8C_1\sigma'^2}, \frac{\epsilon}{2\sigma''}), \eta = \min(\frac{1}{L}, \frac{\beta}{\sqrt{4C_2}}, \frac{\gamma}{\sqrt{8C_1C_3}})$, then in order to guarantee*

$$\sum_{t=0}^{T-1}\frac{1}{T}(B_t+\frac{1}{2}\Gamma_t) \le \epsilon^2.$$

*the iteration complexity is in the order of*

$$T = O\left(\max\left\{\frac{C_\Upsilon L_F}{\epsilon^2}, \frac{C_\Upsilon\sqrt{C_1C_3}\sigma''}{\epsilon^3}, \frac{C_\Upsilon\sigma^2\sqrt{C_2}}{\epsilon^4}, \frac{C_\Upsilon\sqrt{C_1C_3}C_1\sigma'^2}{\epsilon^4}\right\}\right).$$

*where $C_\Upsilon \le A_0-\min_t A_t+\frac{1}{2\sqrt{C_2}}\Delta_0+\sqrt{\frac{C_1}{2C_3}}\delta_0$.*

*Proof.* Following similar analysis to Lemma 4.10, we have

$$A_{t+1} + \frac{\eta_t}{\beta_{t+1}}\Delta_{t+1} + (C_1\eta_t\frac{1+\gamma_{t+1}}{\gamma_{t+1}} - C_1\eta_t)\delta_{t+1} \le A_t - \eta_t B_t - \eta_t\Gamma_t$$
$$+ \left(\eta_t + \frac{\eta_t}{\beta_{t+1}}(1-\beta_{t+1})\right)\Delta_t + \frac{C_2\eta_t^3}{\beta_{t+1}^2}\Gamma_t + \eta_t(\beta_{t+1}\sigma^2 + \boxed{\gamma_{t+1}^2\sigma''^2}) + \boxed{C_1\eta_t(\delta_t - \delta_{t+1})}$$
$$+ C_1\eta_t\frac{1+\gamma_{t+1}}{\gamma_{t+1}}(1-\gamma_{t+1})\delta_t + \frac{C_3C_1\eta_t^3(1+\gamma_{t+1})}{\gamma_{t+1}^2}\Gamma_t + C_1\eta_t(1+\gamma_{t+1})\gamma_{t+1}\sigma'^2.$$

where the terms in the box highlight the difference due to the slight difference in the recursion of $\Delta_t$. Under similar conditions of $\beta_{t+1}, \gamma_{t+1}, \eta_t$ and similar analysis, we get

$$A_{t+1} + \frac{\eta_t}{\beta_{t+1}}\Delta_{t+1} + \frac{C_1\eta_t}{\gamma_{t+1}}\delta_{t+1} \le A_t + \frac{\eta_{t-1}}{\beta_t}\Delta_t + \frac{C_1\eta_{t-1}}{\gamma_t}\delta_t + \boxed{C_1\eta_t(\delta_t - \delta_{t+1})}$$
$$- \eta_t B_t - \frac{1}{2}\eta_t\Gamma_t + \eta_t(\beta_{t+1}\sigma^2 + \boxed{\gamma_{t+1}^2\sigma''^2}) + 2C_1\eta_t\gamma_{t+1}\sigma'^2.$$

Since $\eta_{t+1} \le \eta_t$, we have

$$A_{t+1} + \frac{\eta_t}{\beta_{t+1}}\Delta_{t+1} + \frac{C_1\eta_t}{\gamma_{t+1}}\delta_{t+1} + \boxed{C_1\eta_{t+1}\delta_{t+1}} \le A_t + \frac{\eta_{t-1}}{\beta_t}\Delta_t + \frac{C_1\eta_{t-1}}{\gamma_t}\delta_t + \boxed{C_1\eta_t\delta_t}$$
$$- \eta_t B_t - \frac{1}{2}\eta_t\Gamma_t + \eta_t(\beta_{t+1}\sigma^2 + \boxed{\gamma_{t+1}^2\sigma''^2}) + 2C_1\eta_t\gamma_{t+1}\sigma'^2.$$

Define $\Upsilon_{t+1} = A_{t+1} + \frac{\eta_t}{\beta_{t+1}}\Delta_{t+1} + \frac{C_1\eta_t}{\gamma_{t+1}}\delta_{t+1} + \boxed{C_1\eta_{t+1}\delta_{t+1}}$, we have

$$\eta_t B_t + \frac{1}{2}\eta_t\Gamma_t \le \Upsilon_t - \Upsilon_{t+1} + \eta_t(\beta_{t+1}\sigma^2 + \boxed{\gamma_{t+1}^2\sigma''^2}) + 2C_1\eta_t\gamma_{t+1}\sigma'^2.$$

Hence

$$\sum_{t=0}^{T-1}(\eta_t B_t + \frac{1}{2}\eta_t\Gamma_t) \le \Upsilon_0 - A_* + \sum_{t=0}^{T-1}\left(\eta_t\beta_{t+1}\sigma^2 + 2C_1\eta_t\gamma_{t+1}\sigma'^2 + \eta_t\gamma_{t+1}^2\sigma''^2\right).$$

Next, let us consider $\eta_t = \eta, \beta_t = \beta, \gamma_t = \gamma$. Then we have

$$\sum_{t=0}^{T-1}\frac{1}{T}(B_t + \frac{1}{2}\Gamma_t) \le \frac{C_\Upsilon}{T} + \left(\beta\sigma^2 + 2\gamma C_1\sigma'^2 + \gamma^2\sigma''^2)\right).$$

Since $\eta_t = \eta, \gamma_t = \gamma, \beta_t = \beta$, in order to ensure the RHS is less than $\epsilon^2$, it suffices to have

$$\beta = \frac{\epsilon^2}{4\sigma^2}, \quad \gamma = \min(\frac{\epsilon^2}{8C_1\sigma'^2}, \frac{\epsilon}{2\sigma''}), \quad T \ge \frac{C_\Upsilon}{4\epsilon^2\eta}.$$

Since

$$\eta = \min(\frac{1}{L}, \frac{\beta}{\sqrt{4C_2}}, \frac{\gamma}{\sqrt{8C_1C_3}}).$$

Thus the order of $T$ becomes

$$\begin{aligned} T &= O\left(\max\left\{\frac{C_\Upsilon L_F}{\epsilon^2}, \frac{C_\Upsilon\sqrt{C_2}}{\epsilon^2\beta}, \frac{C_\Upsilon\sqrt{C_1C_3}}{\gamma\epsilon^2}\right\}\right) \\ &= O\left(\max\left\{\frac{C_\Upsilon L_F}{\epsilon^2}, \frac{C_\Upsilon\sqrt{C_1C_3}\sigma''}{\epsilon^3}, \frac{C_\Upsilon\sigma^2\sqrt{C_2}}{\epsilon^4}, \frac{C_\Upsilon\sqrt{C_1C_3}C_1\sigma'^2}{\epsilon^4}\right\}\right) \end{aligned}$$

where

$$C_\Upsilon = A_0 - A_* + \frac{\eta}{\beta}\Delta_0 + \frac{C_1\eta}{\gamma}\delta_0 + C_1\eta\delta_0 \leq A_0 - A_* + \frac{1}{2\sqrt{C_2}}\Delta_0 + 2\frac{\sqrt{C_1}}{\sqrt{8C_3}}\delta_0.$$

□

Finally, we state the convergence of SOX.

**Theorem 5.1** *Under Assumption 5.1 and 5.2, SOX with* $\beta = \frac{\epsilon^2}{4\sigma^2} < \frac{1}{4L_1G_2}, \gamma = \min(\frac{\epsilon^2}{64G_2^2L_1^2{\sigma_0}^2}, \frac{n}{2BG_1L_1\sigma_0}), \eta = \min(\frac{1}{2L_F}, \frac{\beta}{2\sqrt{C_2}}, \frac{B\gamma}{n\sqrt{32C_1C_3}})$*, can find* $\mathbf{w}_\tau$ *with* $\tau$ *randomly sampled from* $\{1, \ldots, T\}$ *so that* $\mathbb{E}\left[\|\mathbf{v}_\tau\|_2^2 + \|\nabla F(\mathbf{w}_\tau)\|_2^2\right] \leq \epsilon^2$ *with an iteration complexity of*

$$T = O\left(\max\left\{\frac{C_\Upsilon L_F}{\epsilon^2}, \frac{C_\Upsilon L_1^2\sigma_0}{\epsilon^3}, \frac{C_\Upsilon L_F\sigma^2}{\epsilon^4}, \frac{C_\Upsilon L_1^3 n\sigma_0^2}{\epsilon^4 B}\right\}\right),$$

*where* $C_1 = 8G_2^2L_1, C_2 = 4L_F^2 + 2, C_3 = 2G_2^2, \sigma^2 = \frac{G_1^2\sigma_2^2}{B} + \frac{G_1^2G_2^2}{B}\frac{n-B}{n-1}$*, and* $C_\Upsilon = O(F(\mathbf{w}_0) - F_* + \frac{1}{L_F}\|\mathbf{v}_0 - \nabla F(\mathbf{w}_0)\|_2^2 + L_1\frac{1}{n}\|\mathbf{u}_0 - g(\mathbf{w}_0)\|_2^2)$.

**Why it matters**

Theorem 5.1 shows that SOX achieves a complexity dominated by $O\left(\frac{C_\Upsilon L_1^3 n\sigma_0^2}{\epsilon^4 B}\right)$, which is comparable to that of SCMA for finding an $\epsilon$-stationary solution. The key difference is that the complexity of SOX is scaled by a factor of $n/B$, since it must track and estimate $n$ inner functions.

*Proof.* Assume that $\epsilon$ is sufficiently small such that $8\beta^2G_2^2L_1^2 \leq 1$. We have established the following three inequalities:

$$(*)\ \mathbb{E}\left[F(\mathbf{w}_{t+1})\right] \le \mathbb{E}\left[F(\mathbf{w}_t)\right] + \frac{\eta}{2}\mathbb{E}[\Delta_t] - \frac{\eta}{2}\mathbb{E}\left[\|\nabla F(\mathbf{w}_t)\|_2^2\right] - \frac{\eta}{4}\mathbb{E}\left[\|\mathbf{v}_t\|_2^2\right],$$

$$(\sharp)\ \mathbb{E}[\Delta_{t+1}] \le (1-\beta)\mathbb{E}\left[\Delta_t\right] + \frac{2L_F^2+1}{\beta}\eta^2\mathbb{E}\left[\|\mathbf{v}_t\|_2^2\right] + 8\beta L_1^2 G_2^2\mathbb{E}\left[\delta_t\right] + \beta^2\sigma^2 + 4G_2^2L_1^2\beta\gamma^2\sigma_0^2,$$

$$(\diamond)\ \mathbb{E}\left[\delta_{t+1}\right] \le \left(1 - \frac{B\gamma}{2n}\right)\mathbb{E}\left[\delta_t\right] + \frac{2nG_2^2\eta^2}{B\gamma}\mathbb{E}\left[\|\mathbf{v}_t\|_2^2\right] + \frac{B\gamma^2\sigma_0^2}{n}.$$

Let us define $\bar{\gamma} = \frac{B\gamma}{2n}$, the last inequality becomes

$$(\diamond)\ \mathbb{E}\left[\delta_{t+1}\right] \le (1-\bar{\gamma})\,\mathbb{E}\left[\delta_t\right] + \frac{G_2^2\eta^2}{\bar{\gamma}}\mathbb{E}\left[\|\mathbf{v}_t\|_2^2\right] + \frac{4n\bar{\gamma}^2\sigma_0^2}{B}.$$

Define $A_t = 2(F(\mathbf{w}_t) - F(\mathbf{w}_*))$ and $B_t = \|\nabla F(\mathbf{w}_t)\|_2^2$, $\Gamma_t = \|\mathbf{v}_t\|_2^2/2$, $\Delta_t = \|\nabla F(\mathbf{w}_t) - \mathbf{v}_t\|_2^2$, $\delta_t = \frac{1}{n}\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2$, and $\Upsilon_t = A_t + \frac{\eta_{t-1}}{\beta_t}\Delta_t + \frac{C_1\eta_{t-1}}{\bar{\gamma}_t}\delta_t$.

Then the three inequalities satisfy that in Lemma 4.10 with $C_1 = 8G_2^2L_1^2, C_2 = 2(2L_F^2+1), C_3 = 2G_2^2, \sigma^2 = \frac{G_1^2\sigma_2^2}{B} + \frac{G_1^2G_2^2}{B}\frac{n-B}{n-1}, \sigma'^2 = \frac{4n\sigma_0^2}{B}, \sigma''^2 = 4G_2^2L_1^2\sigma_0^2$. Then $\eta, \beta, \bar{\gamma}$ satisfy

$$\beta = \frac{\epsilon^2}{4\sigma^2}, \quad \bar{\gamma} = \min\left(\frac{\epsilon^2}{8C_1\sigma'^2}, \frac{\epsilon}{2\sigma''}\right) = \min\left(\frac{\epsilon^2 B}{128G_2^2L_1^2n{\sigma_0}^2}, \frac{\epsilon}{4G_2L_1\sigma_0}\right),$$
$$\eta = \min\left(\frac{1}{2L_F}, \frac{\beta}{\sqrt{4C_2}}, \frac{\bar{\gamma}}{\sqrt{8C_1C_3}}\right).$$

Thus the order of $T$ becomes

$$T = O\left(\max\left\{\frac{C_\Upsilon L_F}{\epsilon^2}, \frac{C_\Upsilon\sqrt{C_1C_3}\sigma''}{\epsilon^3}, \frac{C_\Upsilon\sigma^2\sqrt{C_2}}{\epsilon^4}, \frac{C_\Upsilon\sqrt{C_1C_3}C_1\sigma'^2}{\epsilon^4}\right\}\right)$$
$$= O\left(\max\left\{\frac{C_\Upsilon L_F}{\epsilon^2}, \frac{C_\Upsilon L_1^2\sigma_0}{\epsilon^3}, \frac{C_\Upsilon L_F\sigma^2}{\epsilon^4}, \frac{C_\Upsilon L_1^3n\sigma_0^2}{\epsilon^4 B}\right\}\right),$$

where

$$C_\Upsilon$$
$$\le 2(F(\mathbf{w}_0) - F(\mathbf{w}_*)) + \frac{1}{2\sqrt{C_2}}\|\mathbf{v}_0 - \nabla F(\mathbf{w}_0)\|_2^2 + \frac{\sqrt{C_1}}{\sqrt{2C_3}}\frac{1}{n}\|\mathbf{u}_0 - g(\mathbf{w}_0)\|_2^2$$
$$= 2(F(\mathbf{w}_0) - F(\mathbf{w}_*)) + O\left(\frac{1}{L_F}\right)\|\mathbf{v}_0 - \nabla F(\mathbf{w}_0)\|_2^2 + O(L_1)\frac{1}{n}\|\mathbf{u}_0 - g(\mathbf{w}_0)\|_2^2.$$

□

### 5.2.2 Multi-block Single-Probe Variance Reduction

In this subsection, we present a second algorithm for solving FCCO with an improved complexity than that of SOX under a stronger condition on $g_i$. We replace Assumption 5.2 by the following:

**Assumption 5.3.** *There exist $G_1, L_1, L_2 > 0$ such that*

*(i)* $f_i : \mathbb{R}^{d'} \mapsto \mathbb{R}$ *is $G_1$-Lipschitz continuous and $L_1$-smooth;*
*(ii)* $\nabla g_i(\cdot, \zeta) : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$ *is mean-squared Lipschitz continuous, i.e.,*

$$\mathbb{E}_\zeta[\|\nabla g_i(\mathbf{w}, \zeta) - \nabla g_i(\mathbf{w}', \zeta)\|_2^2] \leq L_2^2 \|\mathbf{w} - \mathbf{w}'\|_2^2, \forall \mathbf{w}, \mathbf{w}';$$

*(iii)* $F_* = \min_{\mathbf{w}} F(\mathbf{w}) \geq > -\infty$.

The idea is to leverage advanced variance reduction for tracking both the inner functions and the gradient. A straightforward approach is to change the update of $\mathbf{u}_{i,t-1}$ by using the STORM estimator and do similarly for the gradient estimator. In particular, one may change the update for $\mathbf{u}_{i,t}$ according to STORM:

$$\mathbf{u}_{i,t} = \begin{cases} (1-\gamma_t)\mathbf{u}_{i,t-1} + \gamma_t g_i\left(\mathbf{w}_t; \zeta_{i,t}\right) + \underbrace{(1-\gamma_t)(g_i\left(\mathbf{w}_t; \zeta_{i,t}\right) - g_i\left(\mathbf{w}_{t-1}; \zeta_{i,t}\right))}_{\text{error correction}} & i \in \mathcal{B}_t \\ \mathbf{u}_{i,t-1} & i \notin \mathcal{B}_t \end{cases}. \tag{5.9}$$

However, this naive approach does not work as the standard error correction term marked above only accounts for the randomness in $g_i(\mathbf{w}_t; \zeta_{i,t})$ but not in the randomness caused by sampling $i \in \mathcal{B}_t$.

In order to tackle this challenge, we introduce the following estimator termed multi-block single-probe variance reduction estimator (MSVR):

$$\mathbf{u}_{i,t} = \begin{cases} (1-\gamma_t)\mathbf{u}_{i,t-1} + \gamma_t g_i\left(\mathbf{w}_t; \zeta_{i,t}\right) + \gamma_t'\left(g_i\left(\mathbf{w}_t; \zeta_{i,t}\right) - g_i\left(\mathbf{w}_{t-1}; \zeta_{i,t}\right)\right) & i \in \mathcal{B}_t \\ \mathbf{u}_{i,t-1} & i \notin \mathcal{B}_t \end{cases}. \tag{5.10}$$

The difference from (5.9) lies at the value of $\gamma_t'$, which is set as $\frac{n-B}{B(1-\gamma_t)} + (1-\gamma_t)$ with $B = |\mathcal{B}_t|$. The MSVR estimator can track multiple functional mappings $(g_1, g_2, \cdots, g_n)$, simultaneously, while the number of sampled blocks $B_1$ for probing can be as small as one. It is notable that when $B = n$, i.e., all blocks are probed at each iteration, $\gamma_t' = 1 - \gamma_t$ and MSVR reduces to STORM applied to $\mathbf{g}(\mathbf{w})$. The additional factor in $\gamma_t'$, i.e., $\alpha_t = \frac{n-B}{B(1-\gamma_t)}$ is to account for the randomness in the sampled blocks and noise in those blocks that are not updated.

With $\mathbf{u}_t$, we compute a vanilla gradient estimator by

$$\mathbf{z}_t = \frac{1}{B}\sum_{i\in\mathcal{B}'_t} \nabla g_i(\mathbf{w}_t;\zeta'_{i,t})\nabla f_i(\mathbf{u}_{i,t}),$$

where $\mathcal{B}'_t \subset [n]$ is a mini-batch of $B$ indices independent of $\mathcal{B}_t$.

Similar to SCST, we apply another STORM estimator to estimate

$$\mathcal{M}_t = \mathbb{E}_{\mathcal{B}'_t,\zeta'_t}[\mathbf{z}_t] = \frac{1}{n}\sum_{i=1}^{n} \nabla g_i(\mathbf{w}_t)\nabla f_i(\mathbf{u}_{i,t}),$$

with an extra vanilla gradient estimator at previous iteration:

$$\tilde{\mathbf{z}}_{t-1} = \frac{1}{B}\sum_{i\in\mathcal{B}'_t} \nabla g_i(\mathbf{w}_{t-1};\zeta'_{i,t})\nabla f_i(\mathbf{u}_{i,t-1}).$$

This is computed by the following sequence:

$$\mathbf{v}_t = (1-\beta_t)\mathbf{v}_{t-1} + \beta_t\mathbf{z}_t + (1-\beta_t)(\mathbf{z}_t - \tilde{\mathbf{z}}_{t-1}). \tag{5.11}$$

Then we use $\mathbf{v}_t$ to update the model parameter. The full steps are presented in Algorithm 15.

**Critical:** We use an independent batch $\mathcal{B}'_t$ because $\mathbf{z}_t$ depends on $\mathbf{u}_t$, which depends on $\mathcal{B}_t$. If we use the same batch $\mathcal{B}_t$ to compute $\mathbf{z}_t$, then

$$\begin{aligned}\mathcal{M}_t &= \mathbb{E}_{\mathcal{B}_t,\zeta'_t}\left[\frac{1}{B}\sum_{i\in\mathcal{B}_t} \nabla g_i(\mathbf{w}_t;\zeta'_{i,t})\nabla f_i(\mathbf{u}_{i,t})\right] \\ &= \mathbb{E}_{\mathcal{B}_t,\zeta'_t}\left[\frac{1}{B}\sum_{i\in\mathcal{B}_t} \nabla g_i(\mathbf{w}_t;\zeta'_{i,t})\nabla f_i(\bar{\mathbf{u}}_{i,t})\right] = \frac{1}{n}\sum_{i=1}^{n} \nabla g_i(\mathbf{w}_t)\nabla f_i(\bar{\mathbf{u}}_{i,t}).\end{aligned}$$

where $\bar{\mathbf{u}}_t$ independent of $\mathcal{B}_t$ is defined in (5.12). However, we cannot construct an unbiased estimator of $\mathcal{M}_{t-1}$ since $\bar{\mathbf{u}}_{t-1}$ is not available in the algorithm.

An alternative approach is that we use $\mathbf{u}_{t-1}$ and $\mathbf{u}_{t-2}$ to compute $\mathbf{z}_t$ and $\tilde{\mathbf{z}}_{t-1}$, respectively, with $\mathcal{B}_t$, i.e.,

$$\begin{aligned}\mathbf{z}_t &= \frac{1}{B}\sum_{i\in\mathcal{B}_t} \nabla g_i(\mathbf{w}_t;\zeta'_{i,t})\nabla f_i(\mathbf{u}_{i,t-1}) \\ \tilde{\mathbf{z}}_{t-1} &= \frac{1}{B}\sum_{i\in\mathcal{B}_t} \nabla g_i(\mathbf{w}_{t-1};\zeta'_{i,t})\nabla f_i(\mathbf{u}_{i,t-2}),\end{aligned}$$

and compute $\mathbf{v}_t$ by

$$\mathbf{v}_t = (1-\beta_t)\mathbf{v}_t + \beta_t\mathbf{z}_t + (1-\beta_t)(\mathbf{z}_t - \tilde{\mathbf{z}}_{t-1}).$$

**Algorithm 15** MSVR

**Input:** learning rate schedules $\{\eta_t\}_{t=1}^T$, $\{\gamma_t\}_{t=1}^T$, $\{\beta_t\}_{t=1}^T$; starting points $\mathbf{w}_0$, $\mathbf{u}_0$, $\mathbf{v}_0$
Let $\mathbf{w}_1 = \mathbf{w}_0 - \eta_0 \mathbf{v}_0$
**for** $t = 1, \ldots, T$ **do**
  Draw two batches of samples $\mathcal{B}_t, \mathcal{B}_t' \subset [n]$
  **for** $i \in \mathcal{B}_t$ **do**
    Draw two samples $\zeta_{i,t}, \zeta_{i,t}' \sim \mathbb{P}_i$
    Update the inner function value estimators

$$\mathbf{u}_{i,t} = (1-\gamma_t)\mathbf{u}_{i,t-1} + \gamma_t g_i\left(\mathbf{w}_t; \zeta_{i,t}\right) + \gamma_t'\left(g_i\left(\mathbf{w}_t; \zeta_{i,t}\right) - g_i\left(\mathbf{w}_{t-1}; \zeta_{i,t}\right)\right)$$

  **end for**
  Set $\mathbf{u}_{i,t} = \mathbf{u}_{i,t-1}, i \notin \mathcal{B}_t$
  Compute the vanilla gradient estimator $\mathbf{z}_t = \frac{1}{B}\sum_{i\in\mathcal{B}_t'} \nabla g_i(\mathbf{w}_t; \zeta_{i,t}')\nabla f_i(\mathbf{u}_{i,t})$
  Compute the extra vanilla gradient estimator $\tilde{\mathbf{z}}_{t-1} = \frac{1}{B}\sum_{i\in\mathcal{B}_t'} \nabla g_i(\mathbf{w}_{t-1}; \zeta_{i,t}')\nabla f_i(\mathbf{u}_{i,t-1})$
  Update the STORM gradient estimator $\mathbf{v}_t = (1-\beta_t)\mathbf{v}_{t-1} + \beta_t \mathbf{z}_t + (1-\beta_t)(\mathbf{z}_t - \tilde{\mathbf{z}}_{t-1})$
  Update the model $\mathbf{w}$ by $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{v}_t$
**end for**

The converge analysis can be performed similarly with slight modifications.

### Convergence Analysis

We first analyze the error recursion of

$$\delta_t = \frac{1}{n}\|\mathbf{u}_t - \mathbf{g}(\mathbf{w}_t)\|_2^2 := \frac{1}{n}\sum_{i=1}^n \|\mathbf{u}_{i,t} - g_i(\mathbf{w}_t)\|_2^2.$$

Similar to the analysis of SOX, we introduce virtual sequences $\bar{\mathbf{u}}_{i,t}, \forall i$ :

$$\bar{\mathbf{u}}_{i,t} = (1-\gamma_t)\mathbf{u}_{i,t-1} + \gamma_t g_i\left(\mathbf{w}_t; \zeta_{i,t}\right) + \gamma_t'\left(g_i\left(\mathbf{w}_t; \zeta_{i,t}\right) - g_i\left(\mathbf{w}_{t-1}; \zeta_{i,t}\right)\right), \forall i. \quad (5.12)$$

**Lemma 5.5** *Consider the $\mathbf{u}_t$ updates in Algorithm 15. Under Assumption 5.1 and 5.3 (ii), by setting $\gamma_t' = \frac{n-B}{B(1-\gamma_t)} + (1-\gamma_t)$, for $\gamma_t \le \frac{1}{2}$, we have:*

$$\mathbb{E}\left[\delta_t\right] \le \left(1 - \frac{B\gamma_t}{n}\right)\mathbb{E}\left[\delta_{t-1}\right] + \frac{2B}{n}\gamma_t^2\sigma_0^2 + \frac{12nG_2^2}{B}\mathbb{E}\left[\|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2\right].$$

*Proof.* Let us consider a fixed $i \in [n]$. With a probability $B/n$ that $i \in \mathcal{B}_t$, we have $\mathbf{u}_{i,t} = \bar{\mathbf{u}}_{i,t}$; otherwise $\mathbf{u}_{i,t} = \mathbf{u}_{i,t-1}$. Hence,

$$\mathbb{E}\left[\|\mathbf{u}_{i,t} - g_i(\mathbf{w}_t)\|_2^2\right] = \frac{B}{n}\underbrace{\mathbb{E}\left[\|\bar{\mathbf{u}}_{i,t} - g_i(\mathbf{w}_t)\|_2^2\right]}_{A_1} + (1-\frac{B}{n})\underbrace{\mathbb{E}\left[\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_t)\|_2^2\right]}_{A_2}.$$

Note that the first term $A_1$ in the R.H.B. can be bounded similarly as in Lemma 4.12 for using the STORM estimator by building a recursion with $\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1})\|_2^2$. However, there exists the second term due to the randomness of $\mathcal{B}_t$, which can be decomposed as

$$\begin{aligned}
A_2 &= \mathbb{E}[\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1}) + g_i(\mathbf{w}_{t-1}) - g_i(\mathbf{w}_t)\|_2^2] \\
&= \underbrace{\mathbb{E}[\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1})\|_2^2]}_{A_{21}} + \underbrace{\mathbb{E}[\|g_i(\mathbf{w}_{t-1}) - g_i(\mathbf{w}_t)\|_2^2]}_{A_{22}} \\
&+ \underbrace{\mathbb{E}[2(\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1}))^\top (g_i(\mathbf{w}_{t-1}) - g_i(\mathbf{w}_t))]}_{A_{23}}.
\end{aligned}$$

The first two terms in RHS ($A_{21}$ and $A_{22}$) can be easily handled. The difficulty comes from the third term, which cannot be simply bounded by using Young's inequality. If doing so, it will end up with a non-diminishing error of $\mathbf{u}_{i,t}$. To combat this difficulty, we use the additional factor brought by $\gamma_t'(g_i(\mathbf{w}_t; \xi_t^i) - g_i(\mathbf{w}_{t-1}; \xi_t^i))$ in $A_1$ to cancel $A_{23}$. This is more clear by the following decomposition of $A_1$.

$$\begin{aligned}
A_1 = \mathbb{E}[\| &\underbrace{(1-\gamma_t)(\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1}))}_{A_{11}} + \underbrace{\alpha_t(g_i(\mathbf{w}_t) - g_i(\mathbf{w}_{t-1}))}_{A_{12}} \\
&+ \underbrace{\gamma_t(g_i(\mathbf{w}_t; \zeta_{i,t}) - g_i(\mathbf{w}_t))}_{A_{13}} \\
&+ \underbrace{\gamma_t'(g_i(\mathbf{w}_t; \zeta_{i,t}) - g_i(\mathbf{w}_{t-1}; \xi_{i,t}) - g_i(\mathbf{w}_t) + g_i(\mathbf{w}_{t-1}))}_{A_{14}} \|_2^2],
\end{aligned}$$

where $\alpha_t = \gamma_t' + \gamma_t - 1$. Since $\mathbb{E}_t[A_{13}] = 0, \mathbb{E}_t[A_{14}] = 0$, then we have

$$A_1 \leq \mathbb{E}[\|A_{11} + A_{12}\|_2^2] + \mathbb{E}\left[\|A_{13} + A_{14}\|_2^2\right].$$

In light of the above decomposition, we can bound $\mathbb{E}[\|A_{11} + A_{12}\|_2^2] \leq \mathbb{E}[\|A_{11}\|_2^2 + \|A_{12}\|_2^2 + 2A_{11}^\top A_{12}]$ and $\mathbb{E}[\|A_{13} + A_{14}\|_2^2] \leq 2\mathbb{E}[\|A_{13}\|_2^2] + 2\mathbb{E}[\|A_{14}\|_2^2]$. The resulting term $\mathbb{E}[2A_{11}^\top A_{12}]$ has a negative sign as $A_{23}$. Hence, by carefully choosing $\gamma_t'$, we can cancel both terms. Specifically, we have

$$
\begin{aligned}
\frac{B}{n}A_1 &\leq \frac{B}{n}\left(\mathbb{E}[\|A_{11}\|_2^2 + \|A_{12}\|_2^2 + 2A_{11}^\top A_{12}] + 2\mathbb{E}[\|A_{13}\|_2^2] + 2\mathbb{E}[\|A_{14}\|_2^2]\right) \\
&= \mathbb{E}\left[\frac{B}{n}(1-\gamma_t)^2\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1})\|_2^2\right] + \mathbb{E}\left[\frac{B}{n}\alpha_t^2\,\|g_i(\mathbf{w}_t) - g_i(\mathbf{w}_{t-1})\|_2^2\right] \\
&+ \mathbb{E}\left[\frac{B}{n}2\alpha_t(1-\gamma_t)(g_i(\mathbf{w}_t) - g_i(\mathbf{w}_{t-1}))^\top(\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1}))\right] \\
&+ \mathbb{E}\left[\frac{B}{n}2\gamma_t^2\left\|g_i(\mathbf{w}_t;\zeta_{i,t}) - g_i(\mathbf{w}_t)\right\|_2^2\right] \\
&+ \mathbb{E}\left[\frac{B}{n}2\gamma_t'^2\left\|(g_i(\mathbf{w}_t;\zeta_{i,t}) - g_i(\mathbf{w}_{t-1};\zeta_{i,t}) - g_i(\mathbf{w}_t) + g_i(\mathbf{w}_{t-1}))\right\|_2^2\right].
\end{aligned}
$$

Combining the upper bounds of $A_1$ and $A_2$, we have

$$
\begin{aligned}
&\frac{B}{n}A_1 + \frac{n-B}{n}A_2 \\
&\leq \mathbb{E}[\frac{B}{n}(1-\gamma_t)^2\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1})\|_2^2 + \frac{B}{n}\alpha_t^2\,\|g_i(\mathbf{w}_t) - g_i(\mathbf{w}_{t-1})\|_2^2] \\
&+ \mathbb{E}\left[\frac{B}{n}2\alpha_t(1-\gamma_t)(g_i(\mathbf{w}_t) - g_i(\mathbf{w}_{t-1}))^\top(\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1}))\right] \\
&+ \mathbb{E}\left[\frac{B}{n}2\gamma_t^2\left\|g_i(\mathbf{w}_t;\zeta_{i,t}) - g_i(\mathbf{w}_t)\right\|_2^2\right] \\
&+ \mathbb{E}\left[\frac{B}{n}2(\gamma_t')^2\left\|(g_i(\mathbf{w}_t;\zeta_{i,t}) - g_i(\mathbf{w}_{t-1};\zeta_{i,t}) - g_i(\mathbf{w}_t) + g_i(\mathbf{w}_{t-1}))\right\|_2^2\right] \\
&+ \mathbb{E}\left[\frac{n-B}{n}\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1})\|_2^2] + \mathbb{E}[\frac{n-B}{n}\|g_i(\mathbf{w}_{t-1}) - g_i(\mathbf{w}_t)\|_2^2\right] \\
&+ \mathbb{E}\left[\frac{n-B}{n}2(\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1}))^\top(g_i(\mathbf{w}_{t-1}) - g_i(\mathbf{w}_t))\right].
\end{aligned}
$$

Since $\frac{B}{n}2\alpha_t(1-\gamma_t) = 2\frac{B}{n}\frac{(n-B)}{B(1-\gamma_t)}(1-\gamma_t) = 2\frac{n-B}{n}$, then cross terms will cancel out. The remaining terms can be merged and handled separately. First,

$$
\begin{aligned}
&\mathbb{E}\left[\frac{B}{n}(1-\gamma_t)^2\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1})\|_2^2 + \frac{n-B}{n}\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1})\|_2^2\right] \\
&\leq (1 - \frac{B}{n}\gamma_t)\mathbb{E}[\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1})\|_2^2],
\end{aligned}
$$

where we use $\frac{B}{n}(1-\gamma_t)^2 + \frac{n-B}{n} \leq 1 - \frac{2B}{n}\gamma_t + \frac{B}{n}\gamma_t^2 \leq 1 - \frac{B}{n}\gamma_t$ due to $\gamma_t < 1$. Second

$$
\begin{aligned}
&\frac{B}{n}\alpha_t^2\,\|g_i(\mathbf{w}_t) - g_i(\mathbf{w}_{t-1})\|_2^2] + \frac{n-B}{n}\|g_i(\mathbf{w}_{t-1}) - g_i(\mathbf{w}_t)\|_2^2 \\
&\leq \left(\frac{B}{n}\frac{(n-B)^2}{B^2(1-\gamma_t)^2} + \frac{n-B}{n}\right)G_2^2\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2 \leq \frac{4n-4B}{B}G_2^2\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2,
\end{aligned}
$$

where we use $\frac{B}{n}\frac{(n-B)^2}{B^2(1-\gamma_t)^2}+\frac{n-B}{n} \leq \frac{n-B}{n}\left(\frac{(n-B)}{B(1-\gamma_t)^2}+1\right) \leq \frac{n-B}{n}\left(\frac{4(n-B)}{B}+4\right) = \frac{4n-4B}{B}$ due to $\gamma_t \leq 1/2$. Third,

$$\begin{aligned}
&\mathbb{E}\left[\frac{B}{n}2\gamma_t'^2\left\|(g_i(\mathbf{w}_t;\zeta_{i,t}) - g_i(\mathbf{w}_{t-1};\zeta_{i,t}) - g_i(\mathbf{w}_t) + g_i(\mathbf{w}_{t-1}))\right\|_2^2\right] \\
&\leq \frac{B}{n}2\gamma_t'^2\mathbb{E}\left[\left\|(g_i(\mathbf{w}_t;\zeta_{i,t}) - g_i(\mathbf{w}_{t-1};\zeta_{i,t})\right\|_2^2\right] \\
&\leq \frac{B}{n}2\left(\frac{n-B}{B(1-\gamma_t)}+1-\gamma_t\right)^2 G_2^2\|\mathbf{w}_t-\mathbf{w}_{t-1}\|_2^2 \leq \frac{8n-4B}{B}G_2^2\|\mathbf{w}_t-\mathbf{w}_{t-1}\|_2^2,
\end{aligned}$$

where we use $\frac{B}{n}2(\frac{n-B}{B(1-\gamma_t)}+1-\gamma_t)^2 \leq \frac{B}{n}2(\frac{2(n-B)}{B}+1)^2 \leq \frac{B}{n}2(\frac{2n-B}{B})^2 = \frac{2(2n-B)(2n-B)}{nB} \leq \frac{8n-4B}{B}$.

Combining the above results, we have

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{u}_{i,t}-g_i(\mathbf{w}_t)\|_2^2\right] &\leq (1-\frac{B}{n}\gamma_t)\mathbb{E}[\|\mathbf{u}_{i,t-1}-g_i(\mathbf{w}_{t-1})\|_2^2] \\
&+ \frac{12n-8B}{B}G_2^2\|\mathbf{w}_t-\mathbf{w}_{t-1}\|_2^2 + \frac{B}{n}2\gamma_t^2\sigma_0^2.
\end{aligned}$$

Averaging over $i = 1,\ldots,n$ concludes the proof. □

**Lemma 5.6** *Consider the $\mathbf{u}_t$ updates in Algorithm 15. Suppose that Assumption 5.1 and 5.3 hold. With $\gamma_t \leq \frac{1}{2}$ and $\gamma_t' = \frac{n-B}{B(1-\gamma_t)} + (1-\gamma_t)$, we have*

$$\mathbb{E}\left[\|\mathbf{u}_t-\mathbf{u}_{t-1}\|_2^2\right] \leq 6B\gamma_t^2\sigma_0^2 + 6B\gamma_t^2\mathbb{E}[\delta_{t-1}] + \frac{10n^2G_2^2}{B}\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}_{t-1}\|_2^2\right].$$

*Proof.* Since $\|\mathbf{u}_t-\mathbf{u}_{t-1}\|_2^2 = \sum_{i=1}^n\|\mathbf{u}_{i,t}-\mathbf{u}_{i,t-1}\|_2^2$, with a probability $B/n$ we have $\mathbf{u}_{i,t} = \bar{\mathbf{u}}_{i,t}$ and a probability $1-B/n$ we have $\mathbf{u}_{i,t} = \mathbf{u}_{i,t-1}$, then

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{u}_t-\mathbf{u}_{t-1}\|_2^2\right] \\
&= \frac{B}{n}\sum_{i=1}^n\mathbb{E}\left[\left\|-\gamma_t\mathbf{u}_{i,t-1}+\gamma_t g_i\left(\mathbf{w}_t;\zeta_{i,t}\right)+\gamma_t'\left(g_i\left(\mathbf{w}_t;\zeta_{i,t}\right)-g_i\left(\mathbf{w}_{t-1};\zeta_{i,t}\right)\right)\right\|_2^2\right] \\
&\leq \frac{B}{n}\sum_{i=1}^n\mathbb{E}\left[2\gamma_t^2\left\|g_i\left(\mathbf{w}_t;\zeta_{i,t}\right)-\mathbf{u}_{i,t-1}\right\|_2^2+2(\gamma_t')^2\left\|g_i\left(\mathbf{w}_t;\zeta_{i,t}\right)-g_i\left(\mathbf{w}_{t-1};\zeta_{i,t}\right)\right\|_2^2\right] \\
&\leq \frac{B}{n}\sum_{i=1}^n\mathbb{E}\left[2\gamma_t^2\left\|g_i\left(\mathbf{w}_t;\zeta_{i,t}\right)-\mathbf{u}_{i,t-1}\right\|_2^2\right]+2B(\gamma_t')^2G_2^2\left\|\mathbf{w}_t-\mathbf{w}_{t-1}\right\|_2^2.
\end{aligned}$$

To the first term on the RHS, we use the Young's inequality and Lipschitz continuity of $g_i$:

$$\begin{aligned}
&\mathbb{E}\left[\left\|g_i\left(\mathbf{w}_t;\zeta_{i,t}\right)-\mathbf{u}_{i,t-1}\right\|_2^2\right] \leq 3\mathbb{E}\left[\left\|g_i\left(\mathbf{w}_t;\zeta_{i,t}\right)-g_i\left(\mathbf{w}_t\right)\right\|_2^2\right] \\
&+3\mathbb{E}\left[\left\|g_i\left(\mathbf{w}_t\right)-g_i\left(\mathbf{w}_{t-1}\right)\right\|_2^2+3\left\|g_i\left(\mathbf{w}_{t-1}\right)-\mathbf{u}_{i,t-1}\right\|_2^2\right] \\
&\leq 3\sigma_0^2+3G_2^2\mathbb{E}\left[\left\|\mathbf{w}_t-\mathbf{w}_{t-1}\right\|_2^2+3\left\|g_i\left(\mathbf{w}_{t-1}\right)-\mathbf{u}_{i,t-1}\right\|_2^2\right].
\end{aligned}$$

Combining the above results, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{u}_t-\mathbf{u}_{t-1}\right\|_2^2\right] \\
&\leq 6B\gamma_t^2\sigma_0^2+6B\gamma_t^2\mathbb{E}\left[\delta_{t-1}\right]+2BG_2^2(3\gamma_t^2+(\gamma_t')^2)\mathbb{E}\left[\left\|\mathbf{w}_t-\mathbf{w}_{t-1}\right\|_2^2\right].
\end{aligned}$$

With $\gamma_t \leq \frac{1}{2}$, we have $\gamma_t' \leq \frac{2n}{B}$, which yields $(3\gamma_t^2+(\gamma_t')^2) \leq \frac{5n^2}{B^2}$. □

Next, we analyze error recursion of $\Delta_t := \|\mathbf{v}_t - \mathcal{M}_t\|_2^2$.

**Lemma 5.7** *Consider the $\mathbf{v}_t$ updates in Algorithm 15 and suppose that Assumption 5.1 and 5.3 hold. Then we have*

$$\begin{aligned}
\mathbb{E}[\Delta_t] &\leq (1-\beta_t)\mathbb{E}[\Delta_{t-1}]+\frac{24G_2^2L_1^2B\gamma_t^2}{n}\mathbb{E}[\delta_{t-1}] \\
&+\left(4L_2^2G_1^2+\frac{40G_2^4L_1^2n}{B}\right)\mathbb{E}\left[\left\|\mathbf{w}_{t-1}-\mathbf{w}_t\right\|_2^2\right]+2\beta_t^2\sigma^2+\frac{24G_2^2L_1^2B}{n}\gamma_t^2\sigma_0^2,
\end{aligned}$$

*where* $\sigma^2 = \frac{G_1^2\sigma_2^2}{B}+\frac{G_1^2G_2^2}{B}\frac{n-B}{n-1}$.

*Proof.* Similar to Lemma 5.2, we have $\mathbb{E}_t[\|\mathbf{z}_t - \mathcal{M}_t\|_2^2] \leq \sigma^2$. Since $\mathbf{v}_t = (1-\beta_t)\mathbf{v}_{t-1}+\beta_t\mathbf{z}_t+(1-\beta_t)(\mathbf{z}_t-\tilde{\mathbf{z}}_{t-1})$, applying Lemma 4.11, we have

$$\mathbb{E}_t\left[\left\|\mathbf{v}_t-\mathcal{M}_t\right\|_2^2\right] \leq (1-\beta_t)\left\|\mathbf{v}_{t-1}-\mathcal{M}_{t-1}\right\|_2^2+\mathbb{E}_t[2\|\mathbf{z}_t-\tilde{\mathbf{z}}_{t-1}\|_2^2]+2\beta_t^2\sigma^2.$$

To bound $\mathbb{E}_t[\|\mathbf{z}_t - \tilde{\mathbf{z}}_{t-1}\|_2^2]$, we have

$$\begin{aligned}
&\mathbb{E}_t[\|\mathbf{z}_t-\tilde{\mathbf{z}}_{t-1}\|_2^2] \\
&\leq 2\mathbb{E}_t\left[\frac{1}{B}\sum_{i\in\mathcal{B}_t'}\left\|\nabla g_i(\mathbf{w}_t;\zeta_{i,t}')\nabla f_i(\mathbf{u}_{i,t})-\nabla g_i(\mathbf{w}_t;\zeta_{i,t}')\nabla f_i(\mathbf{u}_{i,t-1})\right\|_2^2\right] \\
&+2\mathbb{E}_t\left[\frac{1}{B}\sum_{i\in\mathcal{B}_t'}\left\|\nabla g_i(\mathbf{w}_t;\zeta_{i,t}')\nabla f_i(\mathbf{u}_{i,t-1})-\nabla g_i(\mathbf{w}_{t-1};\zeta_{i,t}')\nabla f_i(\mathbf{u}_{i,t-1})\right\|_2^2\right] \\
&\leq 2G_2^2L_1^2\mathbb{E}_t\left[\frac{1}{B}\sum_{i\in\mathcal{B}_t'}\|\mathbf{u}_{i,t}-\mathbf{u}_{i,t-1}\|_2^2\right]+2L_2^2G_1^2\|\mathbf{w}_t-\mathbf{w}_{t-1}\|_2^2 \\
&= 2G_2^2L_1^2\mathbb{E}_t\left[\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{u}_{i,t}-\mathbf{u}_{i,t-1}\|_2^2\right]+2L_2^2G_1^2\|\mathbf{w}_t-\mathbf{w}_{t-1}\|_2^2,
\end{aligned}$$

where the last inequality follows the Assumption 5.3.

As a result, we have

$$\mathbb{E}[\Delta_t] \leq (1-\beta_t)\mathbb{E}[\Delta_{t-1}] + \frac{4G_2^2L_1^2}{n}\mathbb{E}\left[\|\mathbf{u}_t - \mathbf{u}_{t-1}\|_2^2\right] + 4L_2^2G_1^2\mathbb{E}\left[\|\mathbf{w}_{t-1} - \mathbf{w}_t\|_2^2\right] + 2\beta_t^2\sigma^2.$$

Combining with the result in Lemma 5.6, i.e.,

$$\mathbb{E}\left[\|\mathbf{u}_t - \mathbf{u}_{t-1}\|_2^2\right] \leq 6B\gamma_t^2\sigma_0^2 + 6B\gamma_t^2\mathbb{E}[\delta_{t-1}] + \frac{10n^2G_2^2}{B}\mathbb{E}\left[\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2\right].$$

we have

$$\begin{aligned}\mathbb{E}[\Delta_t] \leq\ & (1-\beta_t)\mathbb{E}[\Delta_{t-1}] + \frac{24BG_2^2L_1^2}{n}\gamma_t^2\mathbb{E}[\delta_{t-1}] \\ & + \left(4L_2^2G_1^2 + \frac{40G_2^4L_1^2n}{B}\right)\mathbb{E}\left[\|\mathbf{w}_{t-1} - \mathbf{w}_t\|_2^2\right] + 2\beta_t^2\sigma^2 + \frac{24BG_2^2L_1^2\gamma_t^2\sigma_0^2}{n},\end{aligned}$$

which completes the proof. □

**Lemma 5.8** *For the update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\mathbf{v}_t, t \geq 0$, if $\eta_t \leq 1/(2L_F)$ we have*

$$F(\mathbf{w}_{t+1}) \leq F(\mathbf{w}_t) + G_2^2L_1^2\eta_t\delta_t + \eta_t\Delta_t - \frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t)\|_2^2 - \frac{1}{4\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2. \tag{5.13}$$

*Proof.* It follows directly from Lemma 4.9 by noting that

$$\begin{aligned}&\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2 = \|\mathbf{v}_t - \mathcal{M}_t + \mathcal{M}_t - \nabla F(\mathbf{w}_t)\|_2^2 \\ &\leq 2\Delta_t + 2\left\|\frac{1}{n}\sum_{i=1}^n \nabla g_i(\mathbf{w}_t)\nabla f_i(\mathbf{u}_{i,t}) - \frac{1}{n}\sum_{i=1}^n \nabla g_i(\mathbf{w}_t)\nabla f_i(g_i(\mathbf{w}_t))\right\|_2^2 \\ &\leq 2\Delta_t + \frac{2G_2^2L_1^2}{n}\sum_{i=1}^n \|\mathbf{u}_{i,t} - g_i(\mathbf{w}_t)\|_2^2.\end{aligned}$$

Taking expectation over all randomness on both sides yields the desired result. □

Now we state the convergence theorem for MSVR.

**Theorem 5.2** *Suppose that Assumption 5.1 and 5.3 hold. Let $\beta = O(\frac{\epsilon\eta L_1\sqrt{n}}{\sigma\sqrt{B}})$, $\gamma = \min\left(\frac{\epsilon\eta L_1 n}{\sigma_0 B}, 1\right)$, $\eta = \min\left(\frac{1}{2L_F}, O(\frac{\epsilon\sqrt{B}}{L_1\sigma\sqrt{n}}), O(\frac{\epsilon B}{L_1^2\sigma_0 n}), O(\frac{B}{nL_1})\right)$. Then MSVR can find $\mathbf{w}_\tau$ that is sampled randomly from $\{0, \ldots, T-1\}$ satisfying*

$$\mathbb{E}\left[\|\mathbf{v}_\tau\|_2^2 + \|\nabla F(\mathbf{w}_\tau)\|_2^2\right] \leq O(\epsilon).$$

*with an iteration complexity of*

$$T = O\left(\max\left\{\frac{C_\Upsilon L_F}{\epsilon^2}, \frac{C_\Upsilon L_1 n}{\epsilon^2 B}, \frac{C_\Upsilon L_1 \sigma\sqrt{n}}{\epsilon^3\sqrt{B}}, \frac{C_\Upsilon L_1^2 \sigma_0 n}{\epsilon^3 B}\right\}\right).$$

*where* $\sigma^2 = \frac{G_1^2\sigma_2^2}{B} + \frac{G_1^2 G_2^2(n-B)}{B(n-1)}$, $C_\Upsilon = O(F(\mathbf{w}_0) - F_* + \frac{B}{nL_1^2\eta}\Delta_0 + \frac{B}{nL_1^2\eta}\delta_0)$.

**Why it matters**

Theorem 5.2 indicates that when the initial estimators $\mathbf{u}_0$ and $\mathbf{v}_0$ have an estimation error in the order of $O(\epsilon)$ such that $C_\Upsilon$ is $O(1)$, MSVR attains a better complexity than SOX for finding an $\epsilon$-stationary solution under stronger assumptions of the mean-Lipschitz continuity of $g$ and $\nabla g$. Its complexity is comparable to that of SCST in Theorem 4.4, up to a factor of $n/B$.

*Proof.* We have established the following:

$$(*)\ F(\mathbf{w}_{t+1}) \leq F(\mathbf{w}_t) + G_2^2 L_1^2 \eta_t \delta_t + \eta_t \Delta_t - \frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t)\|_2^2 - \frac{1}{4\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2$$

$$(\sharp)\mathbb{E}[\Delta_t] \leq (1-\beta_t)\mathbb{E}[\Delta_{t-1}] + \frac{24BG_2^2L_1^2}{n}\gamma_t^2\mathbb{E}[\delta_{t-1}]$$
$$+ \left(4L_2^2G_1^2 + \frac{40G_2^4L_1^2 n}{B}\right)\mathbb{E}\left[\|\mathbf{w}_{t-1} - \mathbf{w}_t\|_2^2\right] + 2\beta_t^2\sigma^2 + \frac{24BG_2^2L_1^2\sigma_0^2}{n}\gamma_t^2,$$

$$(\diamond)\mathbb{E}\left[\delta_t\right] \leq \left(1 - \frac{B\gamma_t}{n}\right)\mathbb{E}\left[\delta_{t-1}\right] + \frac{2B}{n}\gamma_t^2\sigma_0^2 + \frac{12nG_2^2}{B}\mathbb{E}\left[\|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2\right].$$

In order to apply Lemma 4.15, we let $A_t = F(\mathbf{w}_t) - F_*$, $B_t = \|\nabla F(\mathbf{w}_t)\|_2^2/2$, $\Gamma_t = \|\mathbf{v}_t\|_2^2/4$, $\bar{\delta}_t = L_1^2G_2^2\delta_t$, $\bar{\gamma}_t = \frac{B\gamma_t}{n}$. Then the following three inequalities

$$(*)\mathbb{E}[A_{t+1}] \leq \mathbb{E}[A_t + \eta_t\Delta_t + \eta_t\bar{\delta}_t - \eta_t B_t - \eta_t\Gamma_t]$$
$$(\sharp)\mathbb{E}\left[\Delta_{t+1}\right] \leq \mathbb{E}[(1-\beta_{t+1})\Delta_t + C_1\bar{\gamma}_{t+1}^2\bar{\delta}_t + C_2\eta_t^2\Gamma_t + \beta_{t+1}^2\sigma^2 + \bar{\gamma}_{t+1}^2\sigma'^2],$$
$$(\diamond)\mathbb{E}\left[\bar{\delta}_{t+1}\right] \leq \mathbb{E}[(1-\bar{\gamma}_{t+1})\bar{\delta}_t + C_3\eta_t^2\Gamma_t + \bar{\gamma}_{t+1}^2\sigma''^2].$$

hold with $C_1 = O(n/B), C_2 = O(L_1^2 n/B + L_2^2), C_3 = O(L_1^2 n/B), \sigma^2 = \frac{G_1^2\sigma_2^2}{B} + \frac{G_1^2G_2^2(n-B)}{B(n-1)}), \sigma'^2 = O(L_1^2\sigma_0^2 n/B), \sigma''^2 = O(L_1^2\sigma_0^2 n/B)$. Following the settings in Lemma 4.15, we can finish the proof with

| | $f_i$ | | | $g_i$ | | | $F$ |
|---|---|---|---|---|---|---|---|
| | Lipschitz continuity | Weak convexity | Monotonicity | Lipschitz continuity | Weak convexity | Smoothness | Weak convexity ($\rho$) |
| 5.5(i) | $G_1$ | $\rho_1$ | $\partial f \geq 0$ | $G_2$ | $\rho_2$ | - | $G_1\rho_2\sqrt{d'} + \rho_1 G_2^2$ |
| 5.5(ii) | $G_1$ | $\rho_1$ | $\partial f \geq 0$ or $\partial f \leq 0$ | $G_2$ | - | $L_2$ | $G_1 L_2\sqrt{d'} + \rho_1 G_2^2$ |

Table 5.1: Conditions of $f_i$ and $g_i$ to make $F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n f_i(g_i(\mathbf{w}))$ weakly convex, where $g_i : \mathbb{R}^d \to \mathbb{R}^{d'}$ and $f_i : \mathbb{R}^{d'} \to \mathbb{R}$.

$$
\begin{aligned}
\eta &= \min\left(\frac{1}{L}, \frac{\epsilon}{4\sqrt{C_2}\sigma}, \frac{\epsilon\sqrt{C_2}}{8C_3\sigma'}, \frac{\epsilon}{8\sqrt{C_3}\sigma''}, \frac{\sqrt{C_2}}{4C_3\sqrt{C_1}}\right) \\
&= \min\left(\frac{1}{2L_F}, O\left(\frac{\epsilon}{L_1\sigma}\sqrt{\frac{B}{n}}\right), O\left(\frac{\epsilon B}{L_1^2\sigma_0 n}\right), O\left(\frac{B}{nL_1}\right)\right), \\
\beta &= \frac{\epsilon\eta\sqrt{2C_2}}{2\sigma} = O\left(\frac{\epsilon\eta L_1}{2\sigma}\sqrt{\frac{n}{B}}\right), \\
\tilde{\gamma} &= \min\left(\frac{\epsilon\eta\sqrt{C_2}}{\sigma'}, \frac{\epsilon\eta\sqrt{C_3}}{\sigma''}, \frac{C_2}{2C_3C_1}\right) = \min\left(O\left(\frac{\epsilon\eta}{\sigma_0}\right), O\left(\frac{B}{n}\right)\right), \\
T &= O\left(\max\left\{\frac{C_\Upsilon L_F}{\epsilon^2}, \frac{C_\Upsilon L_1 n}{\epsilon^2 B}, \frac{C_\Upsilon L_1\sigma\sqrt{n}}{\epsilon^3\sqrt{B}}, \frac{C_\Upsilon L_1^2\sigma_0 n}{\epsilon^3 B}\right\}\right).
\end{aligned}
$$

where $C_\Upsilon = F(\mathbf{w}_0) - F_* + \frac{1}{4C_2\eta}\Delta_0 + \frac{1}{4C_3\eta}\delta_0$.

□

## 5.3 Non-Smooth Weakly Convex Functions

In this section, we consider non-smooth weakly convex functions, where either the outer function or the inner function are non-smooth. The group DRO objective (5.2) falls into this category. Another instance is the two-way partial AUC maximization problem as discussed in Section 6.4.3.

**Assumption 5.4.** *We assume that*

*(i)* $\mathbb{E}_{\zeta\sim\mathbb{P}_i}[\|g_i(\mathbf{w};\zeta) - g_i(\mathbf{w})\|_2^2] \leq \sigma_0^2$.

*(ii)* $\mathbb{E}_{\zeta\sim\mathbb{P}_i}[\|\mathcal{G}_i(\mathbf{w};\zeta)\|_2^2] \leq G_2^2$ *for any* $\mathcal{G}_i(\mathbf{w};\zeta) \in \partial g_i(\mathbf{w};\zeta)$.

The second condition above implies that $g_i$ is $G_2$-Lipschitz continuous.

**Assumption 5.5.** *We assume either of the following conditions holds:*

*(i)* $f_i$ *is* $\rho_1$*-weakly convex,* $G_1$*-Lipschitz continuous, and* $\partial f_i(g) \geq 0\ \forall g$*;* $g_i$ *is* $\rho_2$*-weakly convex.*

*(ii) $f_i$ is $\rho_1$-weakly convex, $G_1$-Lipschitz continuous, and $\partial f_i(g) \geq 0$ or $\partial f_i(g) \leq 0$ $\forall g$; and $g_i$ is $L_2$-smooth.*

We first characterize the conditions on $f_i$ and $g_i$ to induce weak convexity of $F$.

**Lemma 5.9** *Under Assumption 5.4 and 5.5, the objective function $F$ is $\rho$-weakly convex for some $\rho > 0$. If Assumption 5.5(i) holds, then $\rho = G_1\rho_2\sqrt{d'} + \rho_1 G_2^2$ and if Assumption 5.5(ii) holds, then $\rho = G_1 L_2\sqrt{d'} + \rho_1 G_2^2$.*

*Proof.* The weak convexity of $f_i$ implies that for any $\mathbf{v}_i \in \partial f_i(g_i(\mathbf{w}))$:

$$\begin{aligned} f_i(g_i(\mathbf{w}')) &\geq f_i(g_i(\mathbf{w})) + \mathbf{v}_i^\top(g_i(\mathbf{w}') - g_i(\mathbf{w})) - \frac{\rho_1}{2}\|g_i(\mathbf{w}') - g_i(\mathbf{w})\|_2^2 \\ &\geq f_i(g_i(\mathbf{w})) + \mathbf{v}_i^\top(g_i(\mathbf{w}') - g_i(\mathbf{w})) - \frac{\rho_1 G_2^2}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2. \end{aligned}$$

Let us first prove the weak convexity under Assumption 5.5(i). Since $g_i$ is $\rho_2$-weakly convex, we have for any $U_i \in \partial g_i(\mathbf{w})$

$$g_i(\mathbf{w}') - g_i(\mathbf{w}) \geq U_i^\top(\mathbf{w}' - \mathbf{w}) - \frac{\rho_2}{2}\|\mathbf{w}' - \mathbf{w}\|_2^2\mathbf{1}. \tag{5.14}$$

where $\mathbf{1}$ denotes a vector of all ones. Since $\mathbf{v}_i \geq 0$, we have

$$\begin{aligned} f_i(g_i(\mathbf{w}')) &\geq f_i(g_i(\mathbf{w})) + \mathbf{v}_i^\top(U_i^\top(\mathbf{w}' - \mathbf{w}) - \frac{\rho_2}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2\mathbf{1}) - \frac{\rho_1 G_2^2}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2 \\ &\geq f_i(g_i(\mathbf{w})) + (U_i\mathbf{v}_i)^\top(\mathbf{w}' - \mathbf{w}) - \frac{G_1\sqrt{d'}\rho_2 + \rho_1 G_2^2}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2 \end{aligned}$$

Since $U_i\mathbf{v}_i \in \partial g_i(\mathbf{w})\partial f_i(g_i(\mathbf{w}))$, the above inequality indicates that $f_i(g_i(\mathbf{w}))$ is $\rho$-weakly convex, where $\rho = G_1\sqrt{d'}\rho_2 + \rho_1 G_2^2$. As a result, $F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n f_i(g_i(\mathbf{w}))$ is $\rho$-weakly convex.

Next, we prove the weak convexity of $F$ under Assumption 5.5(ii). Due to the smoothness of $g(\cdot)$ we have

$$\begin{aligned} g(\mathbf{w}) - g(\mathbf{w}') &\leq \nabla g(\mathbf{w}')^\top(\mathbf{w} - \mathbf{w}') + \frac{L_2}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2\mathbf{1}, \\ g(\mathbf{w}) - g(\mathbf{w}') &\geq \nabla g(\mathbf{w}')^\top(\mathbf{w} - \mathbf{w}') - \frac{L_2}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2\mathbf{1}. \end{aligned} \tag{5.15}$$

If $\partial f_i(g_i(\mathbf{w})) \geq 0$, we use the second inequity above and follow the same steps as before to prove the $\rho$-weak convexity of $F$ with $\rho = G_1\sqrt{d'}L_2 + \rho_1 G_2^2$. If $\partial f_i(g_i(\mathbf{w})) \leq 0$, we will use the first inequality above to get:

**Algorithm 16** SONX

**Input:** learning rate schedules $\{\eta_t\}_{t=1}^T$, $\{\gamma_t\}_{t=1}^T$, $\{\beta_t\}_{t=1}^T$; starting points $\mathbf{w}_0$, $\mathbf{u}_0$
$\mathbf{w}_1 = \mathbf{w}_0$
**for** $t = 1, \ldots, T$ **do**
  Draw a batch of samples $\mathcal{B}_t \subset [n]$
  **for** $i \in \mathcal{B}_t$ **do**
    Draw two samples $\zeta_{i,t} \sim \mathbb{P}_i$
    Update the inner function value estimators by

$$\text{v1:} \quad \mathbf{u}_{i,t} = (1-\gamma_t)\mathbf{u}_{i,t-1} + \gamma_t g_i(\mathbf{w}_t; \zeta_{i,t})$$
$$\text{v2:} \quad \mathbf{u}_{i,t} = (1-\gamma_t)\mathbf{u}_{i,t-1} + \gamma_t g_i\left(\mathbf{w}_t; \zeta_{i,t}\right) + \gamma_t' \left(g_i\left(\mathbf{w}_t; \zeta_{i,t}\right) - g_i\left(\mathbf{w}_{t-1}; \zeta_{i,t}\right)\right)$$

  **end for**
  Set $\mathbf{u}_{i,t} = \mathbf{u}_{i,t-1}$, $i \notin \mathcal{B}_t$
  Compute $\mathbf{z}_t = \frac{1}{B}\sum_{i\in\mathcal{B}_t'} \partial g_i(\mathbf{w}_t; \zeta_{i,t}')\partial f_i(\mathbf{u}_{i,t})$ ◇ check text for discussion
  Update the model $\mathbf{w}$ by $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{z}_t$
**end for**

$$\begin{aligned}
f_i(g_i(\mathbf{w}')) &\geq f_i(g_i(\mathbf{w})) + \mathbf{v}_i^\top (g_i(\mathbf{w}') - g_i(\mathbf{w})) - \frac{\rho_1}{2}\|g_i(\mathbf{w}') - g_i(\mathbf{w})\|_2^2 \\
&\geq f_i(g_i(\mathbf{w})) + \mathbf{v}_i^\top (\nabla g_i(\mathbf{w})^\top (\mathbf{w}' - \mathbf{w}) + \frac{L_2}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2 \mathbf{1}) - \frac{\rho_1 G_2^2}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2 \\
&\geq f_i(g_i(\mathbf{w})) + (\nabla g_i(\mathbf{w})\mathbf{v}_i)^\top (\mathbf{w}' - \mathbf{w}) - \frac{G_1\sqrt{d'}L_2 + \rho_1 G_2^2}{2}\|\mathbf{w} - \mathbf{w}'\|_2^2.
\end{aligned}$$

This concludes the proof. □

### 5.3.1 SONX for Non-smooth Inner Functions

Since we do not assume smoothness for the overall objective function, the key difference from the previous two sections is that we no longer have the descent lemma in Lemma 4.9, hence cannot leverage the MA or STORM gradient estimators. Consequently, we employ the vanilla gradient estimator $\mathbf{z}_t$ to update the model parameter $\mathbf{w}_{t+1}$. The updating steps are summarized in Algorithm 16, referred to as **SONX**. The two options correspond to different strategies for updating the inner function value estimators: v1 uses a coordinate MA estimator, while v2 adopts the MSVR estimator.

For ease of presentation, we compute the vanilla gradient estimator $\mathbf{z}_t$ using a batch $\mathcal{B}_t'$ independent from $\mathcal{B}_t$:

$$\mathbf{z}_t = \frac{1}{B}\sum_{i\in\mathcal{B}_t'} \partial g_i(\mathbf{w}_t; \zeta_{i,t}')\partial f_i(\mathbf{u}_{i,t}).$$

However, for SONX-v1 with MA estimator, we can indeed use the same vanilla gradient estimator $\mathbf{z}_t$ as in SOX:

$$\mathbf{z}_t = \frac{1}{B}\sum_{i\in\mathcal{B}_t} \partial g_i(\mathbf{w}_t;\zeta'_{i,t})\partial f_i(\mathbf{u}_{i,t}).$$

An alternative method for using both options is to compute $\mathbf{z}_t$ by

$$\mathbf{z}_t = \frac{1}{B}\sum_{i\in\mathcal{B}_t} \partial g_i(\mathbf{w}_t;\zeta'_{i,t})\partial f_i(\mathbf{u}_{i,t-1}).$$

### Convergence Analysis

Similar to Section 3.1.4, we state the convergence using the Moreau envelope of $F$:

$$F_\lambda(\mathbf{w}) := \min_{\mathbf{u}} F(\mathbf{u}) + \frac{1}{2\lambda}\|\mathbf{u}-\mathbf{w}\|_2^2.$$

Recall the definition:

$$\mathrm{prox}_{\lambda F}(\mathbf{w}) = \arg\min_{\mathbf{u}} F(\mathbf{u}) + \frac{1}{2\lambda}\|\mathbf{u}-\mathbf{w}\|_2^2.$$

We first present a result similar to Lemma 3.5 for standard SGD to account for the bias of $\mathbf{z}_t$.

**Lemma 5.10** *Suppose Assumption 5.4 and 5.5 hold. Let $\bar\rho = \rho + \rho_2 G_1 + 2\rho_1 G_2^2$. Consider the step update of SONX, we have*

$$\mathbb{E}_{\zeta'_t,\mathcal{B}'_t}[F_{1/\bar\rho}(\mathbf{w}_{t+1})] \le F_{1/\bar\rho}(\mathbf{w}_t) + \frac{\eta_t^2\bar\rho G^2}{2} - \frac{\eta_t}{2}\|\nabla F_{1/\bar\rho}(\mathbf{w}_t)\|_2^2$$
$$+ \frac{\bar\rho\eta_t}{n}\sum_{i=1}^n\left[2G_1\|g_i(\mathbf{w}_t)-\mathbf{u}_{i,t}\|_2 + \rho_1\|g_i(\mathbf{w}_t)-\mathbf{u}_{i,t}\|_2^2\right].$$

*If $f_i$ is further $L_1$-smooth, then*

$$\mathbb{E}_{\zeta'_t,\mathcal{B}'_t}[F_{1/\bar\rho}(\mathbf{w}_{t+1})] \le F_{1/\bar\rho}(\mathbf{w}_t) + \frac{\eta_t^2\bar\rho G^2}{2} - \frac{\eta_t}{2}\|\nabla F_{1/\bar\rho}(\mathbf{w}_t)\|_2^2$$
$$+ \frac{\bar\rho\eta_t}{n}\sum_{i=1}^n\left[\frac{L_1}{2}\|g_i(\mathbf{w}_t)-\mathbf{u}_{i,t}\|_2^2 + \rho_1\|g_i(\mathbf{w}_t)-\mathbf{u}_{i,t}\|_2^2\right].$$

*where $G^2 = G_1^2G_2^2$.*

If $\mathbf{u}_{i,t} = g_i(\mathbf{w}_t)$, i.e., there is no bias in $\mathbf{z}_t$, then the terms in the square bracket are gone, the above lemma reduces to Lemma 3.4.

*Proof.* Define $\hat{\mathbf{w}}_t := \mathrm{prox}_{F/\bar\rho}(\mathbf{w}_t)$ and $\mathbb{E}_t[\cdot] = \mathbb{E}_{\zeta'_t,\mathcal{B}'_t}[\cdot]$. First,

$$\mathbb{E}_t[\|\mathbf{z}_t\|_2^2] \le \mathbb{E}_t\left[\frac{1}{B}\sum_{i\in\mathcal{B}_t}\|\partial g_i(\mathbf{w}_t,\zeta'_{i,t})\partial f_i(\mathbf{u}_{i,t})\|_2^2\right]$$
$$\le \mathbb{E}_t\left[\frac{1}{B}\sum_{i\in\mathcal{B}_t}\|\partial g_i(\mathbf{w}_t,\zeta'_{i,t})\|_2^2 G_1^2\right] \le G_2^2 G_1^2 = G^2.$$

Following Lemma 3.4, we have

$$\mathbb{E}_t[F_{1/\bar\rho}(\mathbf{w}_{t+1})] \le F_{1/\bar\rho}(\mathbf{w}_t) + \bar\rho\eta_t(\mathbb{E}_t[\mathbf{z}_t])^\top(\hat{\mathbf{w}}_t - \mathbf{w}_t) + \frac{\eta_t^2\bar\rho G^2}{2}. \tag{5.16}$$

Next we bound the term $\mathbb{E}_t[\mathbf{z}_t]^\top(\hat{\mathbf{w}}_t - \mathbf{w}_t)$ on the RHS of (5.16). Note that $\mathbb{E}_t[\mathbf{z}_t] = \frac{1}{n}\sum_{i=1}^n \partial g_i(\mathbf{w}_t)\partial f_i(\mathbf{u}_{i,t})$. For a given $i \in [n]$, we have

$$\begin{aligned}
&f_i(g_i(\hat{\mathbf{w}}_t)) - f_i(\mathbf{u}_{i,t})\\
&\overset{(a)}{\ge} \partial f_i(\mathbf{u}_{i,t})^\top(g_i(\hat{\mathbf{w}}_t) - \mathbf{u}_{i,t}) - \frac{\rho_1}{2}\|g_i(\hat{\mathbf{w}}_t) - \mathbf{u}_{i,t}\|_2^2\\
&\ge \partial f_i(\mathbf{u}_{i,t})^\top(g_i(\hat{\mathbf{w}}_t) - \mathbf{u}_{i,t}) - \rho_1\|g_i(\hat{\mathbf{w}}_t) - g_i(\mathbf{w}_t)\|_2^2 - \rho_1\|g_i(\mathbf{w}_t) - \mathbf{u}_{i,t}\|_2^2\\
&\ge \partial f_i(\mathbf{u}_{i,t})^\top(g_i(\hat{\mathbf{w}}_t) - \mathbf{u}_{i,t}) - \rho_1 G_2^2\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|_2^2 - \rho_1\|g_i(\mathbf{w}_t) - \mathbf{u}_{i,t}\|_2^2\\
&\overset{(b)}{\ge} \partial f_i(\mathbf{u}_{i,t})^\top\left[g_i(\mathbf{w}_t) - \mathbf{u}_{i,t} + \partial g_i(\mathbf{w}_t)^\top(\hat{\mathbf{w}}_t - \mathbf{w}_t) - \frac{\rho_2}{2}\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|_2^2\right]\\
&\quad - \rho_1 G_2^2\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|_2^2 - \rho_1\|g_i(\mathbf{w}_t) - \mathbf{u}_{i,t}\|_2^2\\
&\overset{(c)}{\ge} \partial f_i(\mathbf{u}_{i,t})^\top(g_i(\mathbf{w}_t) - \mathbf{u}_{i,t}) + \partial f_i(\mathbf{u}_{i,t})^\top\partial g_i(\mathbf{w}_t)^\top(\hat{\mathbf{w}}_t - \mathbf{w}_t)\\
&\quad - (\frac{\rho_2 G_1}{2} + \rho_1 G_2^2)\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|_2^2 - \rho_1\|g_i(\mathbf{w}_t) - \mathbf{u}_{i,t}\|_2^2,
\end{aligned}$$

where (a) follows from the $\rho_1$-weak-convexity of $f_i$, (b) follows from that $\partial f_i(\cdot) \ge 0$ and the weak convexity of $g_i$, (c) is due to $\|\partial f_i(\mathbf{u}_{i,t})\|_2 \le G_1$. When $\partial f_i(\cdot) \le 0$ and $g_i$ is smooth, we can bound similarly with $\rho_2$ in the last inequality replaced by $L_2$.

Then rearranging the above inequality and averaging over $i$ yields

$$\begin{aligned}
&\mathbb{E}_t[\mathbf{z}_t]^\top(\hat{\mathbf{w}}_t - \mathbf{w}_t) = \frac{1}{n}\sum_{i=1}^n \partial f_i(\mathbf{u}_{i,t})^\top\partial g_i(\mathbf{w}_t)^\top(\hat{\mathbf{w}}_t - \mathbf{w}_t)\\
&\le \frac{1}{n}\sum_{i=1}^n\Big[f_i(g_i(\hat{\mathbf{w}}_t)) - f_i(g_i(\mathbf{w}_t)) + f_i(g_i(\mathbf{w}_t)) - f_i(\mathbf{u}_{i,t})\\
&\quad - \partial f_i(\mathbf{u}_{i,t})^\top(g_i(\mathbf{w}_t) - \mathbf{u}_{i,t}) + (\frac{\rho_2 G_1}{2} + \rho_1 G_2^2)\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|_2^2 + \rho_1\|g_i(\mathbf{w}_t) - \mathbf{u}_{i,t}\|_2^2\Big].
\end{aligned} \tag{5.17}$$

Due to the $\rho$-weak convexity of $F(\mathbf{w})$, we have that $F(\mathbf{w}) + \frac{\bar\rho}{2}\|\mathbf{w}_t - \mathbf{w}\|_2^2$ is $(\bar\rho - \rho)$-strongly convex. Then $\left[F(\mathbf{w}_t) + \frac{\bar\rho}{2}\|\mathbf{w}_t - \mathbf{w}_t\|_2^2\right] - \left[F(\hat{\mathbf{w}}_t) + \frac{\bar\rho}{2}\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_2^2\right] \ge \frac{\bar\rho-\rho}{2}\|\hat{\mathbf{w}}_t - \mathbf{w}_t\|_2^2$. It follows that:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\Big[f_i(g_i(\hat{\mathbf{w}}_t)) - f_i(g_i(\mathbf{w}_t))\Big] = F(\hat{\mathbf{w}}_t) - F(\mathbf{w}_t) \\
&= \Big[F(\hat{\mathbf{w}}_t) + \frac{\bar{\rho}}{2}\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_2^2\Big] - \Big[F(\mathbf{w}_t) + \frac{\bar{\rho}}{2}\|\mathbf{w}_t - \mathbf{w}_t\|_2^2\Big] - \frac{\bar{\rho}}{2}\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_2^2 \qquad (5.18)\\
&\leq (\frac{\rho}{2} - \bar{\rho})\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_2^2
\end{aligned}$$

Combining inequality (5.17), (5.16) and (5.18) yields

$$\begin{aligned}
\mathbb{E}_t[F_{1/\bar{\rho}}(\mathbf{w}_{t+1})] \leq F_{1/\bar{\rho}}(\mathbf{w}_t) + \frac{\eta_t^2\bar{\rho}G^2}{2} - \frac{\bar{\rho}^2\eta_t}{2}\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_2^2 \\
+ \frac{\bar{\rho}\eta_t}{n}\sum_{i=1}^{n}\Big[f_i(g_i(\mathbf{w}_t)) - f_i(\mathbf{u}_{i,t}) - \partial f_i(\mathbf{u}_{i,t})^\top(g_i(\mathbf{w}_t) - \mathbf{u}_{i,t}) \\
+ \rho_1\|g_i(\mathbf{w}_t) - \mathbf{u}_{i,t}\|_2^2\Big].
\end{aligned}$$

We finish the proof by noting that $\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t)\|_2 = \bar{\rho}\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|_2$, using

$$f_i(g_i(\mathbf{w}_t)) - f_i(\mathbf{u}_{i,t}) - \partial f_i(\mathbf{u}_{i,t})^\top(g_i(\mathbf{w}_t) - \mathbf{u}_{i,t}) \leq 2G_1\|g_i(\mathbf{w}_t) - \mathbf{u}_{i,t}\|_2,$$

if $f_i$ is $G_1$-Lipschitz continuous, or using

$$f_i(g_i(\mathbf{w}_t)) - f_i(\mathbf{u}_{i,t}) - \partial f_i(\mathbf{u}_{i,t})^\top(g_i(\mathbf{w}_t) - \mathbf{u}_{i,t}) \leq \frac{L_1}{2}\|g_i(\mathbf{w}_t) - \mathbf{u}_{i,t}\|_2^2,$$

if $f_i$ is $L_1$-smooth. □

**Convergence of SONX-v1**

Recall the definition:

$$\delta_t = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{u}_{i,t} - g_i(\mathbf{w}_t)\|_2^2.$$

Let us also define:

$$\delta_t' = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{u}_{i,t} - g_i(\mathbf{w}_t)\|_2.$$

From Lemma 5.10, the key is to bound $\delta_t$ and $\delta_t'$.

**Lemma 5.11** *Consider the update of SONX-v1, under Assumptions 5.4 and 5.5, with constant parameters $\gamma_t = \gamma \leq 1$ and $\eta_t = \eta$, we have*

$$\mathbb{E}\left[\delta_t\right] \leq \left(1-\frac{B\gamma}{4n}\right)^{2t}\delta_0 + \frac{8n^2G_2^4G_1^2\eta^2}{B^2\gamma^2} + 4\gamma\sigma_0^2.$$
$$\mathbb{E}\left[\delta_t'\right] \leq \left(1-\frac{B\gamma}{4n}\right)^{t}\delta_0' + \frac{4nG_2^2G_1\eta}{B\gamma} + 2\sqrt{\gamma}\sigma_0.$$

*Proof.* From the proof of Lemma 5.1, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{u}_{i,t} - g_i(\mathbf{w}_t)\right\|_2^2\right] \\
&\leq \left(1-\frac{B\gamma_t}{2n}\right)\mathbb{E}\left[\left\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1})\right\|_2^2\right] + \frac{2nG_2^2}{B\gamma_t}\mathbb{E}\left[\|\mathbf{w}_{t-1}-\mathbf{w}_t\|_2^2\right] + \frac{B\gamma_t^2\sigma_0^2}{n} \\
&\leq \left(1-\frac{B\gamma_t}{2n}\right)\mathbb{E}\left[\left\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1})\right\|_2^2\right] + \frac{2nG_2^2\eta_{t-1}^2}{B\gamma_t}\mathbb{E}\left[\|\mathbf{z}_{t-1}\|_2^2\right] + \frac{B\gamma_t^2\sigma_0^2}{n} \\
&\leq \left(1-\frac{B\gamma_t}{4n}\right)^2\mathbb{E}\left[\left\|\mathbf{u}_{i,t-1} - g_i(\mathbf{w}_{t-1})\right\|_2^2\right] + \frac{2nG_2^4G_1^2\eta_{t-1}^2}{B\gamma_t} + \frac{B\gamma_t^2\sigma_0^2}{n}.
\end{aligned}$$

Applying the above inequality recursively for $\gamma_t = \gamma$ and $\eta_t = \eta$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{u}_{i,t} - g_i(\mathbf{w}_t)\right\|_2^2\right] \\
&\leq \left(1-\frac{B\gamma}{4n}\right)^{2t}\left\|\mathbf{u}_{i,0} - g_i(\mathbf{w}_0)\right\|_2^2 + \sum_{j=0}^{t-1}\left(1-\frac{B\gamma}{4n}\right)^{2j}\left(\frac{2nG_2^4G_1^2\eta^2}{B\gamma} + \frac{B\gamma^2\sigma_0^2}{n}\right) \\
&\leq \left(1-\frac{B\gamma}{4n}\right)^{2t}\left\|\mathbf{u}_{i,0} - g_i(\mathbf{w}_0)\right\|_2^2 + \frac{8n^2G_2^4G_1^2\eta^2}{B^2\gamma^2} + 4\gamma\sigma_0^2,
\end{aligned}$$

where we use

$$\sum_{j=0}^{t-1}(1-\alpha)^{2j} \leq \sum_{j=0}^{\infty}(1-\alpha)^{2j} = \frac{1}{1-(1-\alpha)^2} = \frac{1}{\alpha(2-\alpha)} \leq \frac{1}{\alpha}, \forall\alpha\in(0,1).$$

Averaging the above inequality over $i$, we prove the first result in the lemma.

It follows

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{u}_{i,t} - g_i(\mathbf{w}_t)\right\|_2\right] &\leq \sqrt{\mathbb{E}\left[\left\|\mathbf{u}_{i,t} - g_i(\mathbf{w}_t)\right\|_2^2\right]} \\
&\leq \sqrt{\left(1-\frac{B\gamma}{4n}\right)^{2t}\left\|\mathbf{u}_{i,0} - g_i(\mathbf{w}_0)\right\|_2^2 + \frac{8n^2G_2^4G_1^2\eta^2}{B^2\gamma^2} + 4\gamma\sigma_0^2} \\
&\leq \left(1-\frac{B\gamma}{4n}\right)^{t}\left\|\mathbf{u}_{i,0} - g_i(\mathbf{w}_0)\right\|_2 + \frac{4nG_2^2G_1\eta}{B\gamma} + 2\gamma^{1/2}\sigma_0.
\end{aligned}$$

Averaging the above result, we prove the second result. □

**Theorem 5.3 (Convergence of SONX-v1 with Lipschitz $f_i$)** *Consider SONX-v1, and suppose Assumption 5.4 and 5.5 hold and $f_i$ is $G_1$-Lipschitz continuous. Let $\eta_t = \eta = O(\frac{B\epsilon^6}{n\sigma_0^2})$, $\gamma_t = \gamma = O(\frac{\epsilon^4}{\sigma_0^2})$. Then after $T = O(\frac{n\sigma_0^2}{B\epsilon^8})$ iterations, we have $\mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t)\|_2^2] \leq O(\epsilon^2)$.*

*Proof.* From Lemma 5.10, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t)\|_2^2\right] \leq \mathbb{E}\left[\frac{2\sum_{t=1}^{T}(F_{1/\bar{\rho}}(\mathbf{w}_t) - F_{1/\bar{\rho}}(\mathbf{w}_{t+1}))}{\eta T}\right] + \eta\bar{\rho}G^2$$
$$+ 4\bar{\rho}G_1\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\delta_t'\right] + 2\bar{\rho}\rho_1\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\delta_t\right].$$

Next, we bound the last two terms. From Lemma 5.11, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\delta_t\right] \leq \frac{1}{T}\sum_{t=1}^{T}\left(1 - \frac{B\gamma}{4n}\right)^{2t}\delta_0 + \frac{8n^2G_2^4G_1^2\eta^2}{B^2\gamma^2} + 4\gamma\sigma_0^2.$$
$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\delta_t'\right] \leq \frac{1}{T}\sum_{t=1}^{T}\left(1 - \frac{B\gamma}{4n}\right)^{t}\delta_0' + \frac{4nG_2^2G_1\eta}{B\gamma} + 2\sqrt{\gamma}\sigma_0.$$

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

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\delta_t\right] \leq \frac{4n\delta_0}{B\gamma T} + \frac{8n^2G_2^4G_1^2\eta^2}{B^2\gamma^2} + 4\gamma\sigma_0^2.$$
$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\delta_t'\right] \leq \frac{4n\delta_0'}{B\gamma T} + \frac{4nG_2^2G_1\eta}{B\gamma} + 2\sqrt{\gamma}\sigma_0.$$

From Proposition 3.2, we have

$$\sum_{t=1}^{T}(F_{1/\bar{\rho}}(\mathbf{w}_t) - F_{1/\bar{\rho}}(\mathbf{w}_{t+1})) = F_{1/\bar{\rho}}(\mathbf{w}_1) - F_{1/\bar{\rho}}(\mathbf{w}_{T+1}) \leq F(\mathbf{w}_1) - F(\mathbf{w}_*).$$

Combining the above results, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t)\|_2^2\right] \leq \mathbb{E}\left[\frac{2(F(\mathbf{w}_1) - F_*)}{\eta T}\right] + \eta\bar{\rho}G^2$$
$$+ 4\bar{\rho}G_1\left(\frac{4n\delta_0'}{B\gamma T} + \frac{4nG_2^2G_1\eta}{B\gamma} + 2\sqrt{\gamma}\sigma_0\right) + 2\bar{\rho}\rho_1\left(\frac{4n\delta_0}{B\gamma T} + \frac{8n^2G_2^4G_1^2\eta^2}{B^2\gamma^2} + 4\gamma\sigma_0^2\right).$$

Plugging the order of $\eta, \gamma$, we finish the proof.

□

**Theorem 5.4 (Convergence of SONX-v1 with smooth $f_i$)** *Consider SONX-v1, and suppose Assumption 5.1 and 5.5 hold and $f_i$ is $L_1$-smooth. Let $\eta_t = \eta = O(\frac{B\epsilon^3}{n\sigma_0^2})$, $\gamma_t = \gamma = O(\frac{\epsilon^2}{\sigma_0^2})$, then after $T = O(\frac{n\sigma_0^2}{B\epsilon^5})$ iterations, we have $\mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t)\|_2^2] \leq O(\epsilon^2)$.*

*Proof.* By using the result for smooth $f_i$ in Lemma 5.10, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t)\|_2^2\right] \leq \mathbb{E}\left[\frac{2\sum_{t=1}^{T}(F_{1/\bar{\rho}}(\mathbf{w}_t) - F_{1/\bar{\rho}}(\mathbf{w}_{t+1}))}{\eta T}\right] + \eta\bar{\rho}G^2$$
$$+ \bar{\rho}(L_1 + 2\rho_1)\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\delta_t\right].$$

Plugging the bounds for the first and last term in the RHS, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_{1/\bar{\rho}}(\mathbf{w}_t)\|_2^2\right] \leq \mathbb{E}\left[\frac{2(F(\mathbf{w}_1) - F_*)}{\eta T}\right] + \eta\bar{\rho}G^2$$
$$+ \bar{\rho}(L_1 + 2\rho_1)\left(\frac{4n\delta_0}{B\gamma T} + \frac{8n^2G_2^4G_1^2\eta^2}{B^2\gamma^2} + 4\gamma\sigma_0^2\right).$$

Plugging the order of $\eta, \gamma$, we finish the proof.

□

### Convergence of SONX-v2

Similar to the first option, we need to bound $\delta_t, \delta_t'$ first.

**Lemma 5.12** *Under Assumption 5.4, 5.5, by setting $\gamma_t = \gamma \leq \frac{1}{2}$, $\eta_t = \eta$, $\gamma_t' = \frac{n-B}{B(1-\gamma)} + (1-\gamma)$, we have:*

$$\mathbb{E}\left[\delta_t\right] \leq \left(1 - \frac{B\gamma}{2n}\right)^{2t}\delta_0 + 4\gamma\sigma_0^2 + \frac{24n^2G_2^4G_1^2\eta^2}{B^2\gamma},$$
$$\mathbb{E}\left[\delta_t'\right] \leq \left(1 - \frac{B\gamma}{2n}\right)^{t}\delta_0' + 2\gamma^{1/2}\sigma_0 + \frac{5nG_2^2G_1\eta}{B\gamma^{1/2}}.$$

Proof is omitted as it is similar to that of Lemma 5.11 but based on Lemma 5.5.

**Theorem 5.5 (Convergence of SONX-v2)** *Consider SONX-v2, and suppose Assumption 5.4, and 5.5 hold.*

- *If $f_i$ is $G_1$-Lipschitz continuous, by setting $\eta = O(\frac{B\epsilon^4}{n\sigma_0})$, $\gamma = O(\frac{\epsilon^4}{\sigma_0^2})$, then after $T = O(\frac{n\sigma_0}{B\epsilon^6})$ iterations, we have $\mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{x}_t)\|_2^2] \leq \epsilon^2$.*

- *If $f_i$ is further $L_1$-smooth, by setting $\eta = O(\frac{B\epsilon^2}{n\sigma_0})$, $\gamma = O(\frac{\epsilon^2}{\sigma_0^2})$, then the complexity reduces to $T = O(\frac{n\sigma_0}{B\epsilon^4})$.*

The proof follows similarly to that of Theorem 5.3 and Theorem 5.4 and is left as an exercise for interested readers.

### 5.3.2 SONEX for Non-smooth Outer functions

When $f_i$ is Lipschitz continuous and non-smooth, the best complexity derived in last subsection is $O(n/(B\epsilon^6))$. Can we further improve the complexity when the inner functions are smooth? We present a method and its analysis in this section.

Let us make the following assumptions.

**Assumption 5.6.** *We assume that*

*(i)* $\mathbb{E}_{\zeta\sim\mathbb{P}_i}[\|g_i(\mathbf{w};\zeta) - g_i(\mathbf{w})\|_2^2] \leq \sigma_0^2$
*(ii)* $\mathbb{E}_{\zeta\sim\mathbb{P}_i}[\|\nabla g_i(\mathbf{w};\zeta) - \nabla g_i(\mathbf{w})\|_2^2] \leq \sigma_2^2$
*(iii)* $\mathbb{E}_{\zeta\sim\mathbb{P}_i}[\|\nabla g_i(\mathbf{w};\zeta)\|_2^2] \leq G_2^2$.

**Assumption 5.7.** *The following conditions hold:*

*(i) $f_i$ is $\rho_1$-weakly convex, $G_1$-Lipschitz continuous,*
*(ii) $g_i$ is $L_2$-smooth and $G_2$-Lipschitz continuous.*

#### Moreau Envelope Smoothing of the outer function

A classical approach of improving the convergence for non-smooth functions in convex optimization is smoothing, i.e., first smoothing the function and then using an optimizer for solving the resulting smoothed function. We define the Moreau envelope smoothing of $f_i$ as follows:

$$\bar{f}_i(g) = \min_{\mathbf{u}\in\mathbb{R}^{d'}} \frac{\bar{\rho}_1}{2}\|\mathbf{u} - g\|_2^2 + f_i(\mathbf{u}), \tag{5.19}$$

where $\bar{\rho}_1 > \rho_1$. We present a lemma below regarding $\bar{f}_i$.

**Lemma 5.13** *If $f_i$ is $G_1$-Lipschitz continuous and $\rho_1$-weakly convex, then $\bar{f}_i$ is $\bar{L}_1$-smooth and $G_1$ Lipschitz continuous, where $\bar{L}_1 = \frac{\bar{\rho}_1(2\bar{\rho}_1-\rho_1)}{(\bar{\rho}_1-\rho_1)}$.*

*Proof.* Define $\text{prox}_{f_i/\bar{\rho}_1}(g) = \arg\min_{\mathbf{u}\in\mathbb{R}^{d'}} \frac{\bar{\rho}_1}{2}\|\mathbf{u} - g\|_2^2 + f_i(\mathbf{u})$. We have

$$\nabla\bar{f}_i(g) = \bar{\rho}_1(g - \text{prox}_{f_i/\bar{\rho}_1}(g)).$$

Due to the optimality condition of $\text{prox}_{f_i/\bar{\rho}_1}(g)$, we have

$$\bar{\rho}_1(g - \text{prox}_{f_i/\bar{\rho}_1}(g)) \in \partial f_i(\text{prox}_{f_i/\bar{\rho}_1}(g)).$$

Hence, $\nabla \bar{f}_i(g) \in \partial f_i(\text{prox}_{f_i/\bar{\rho}_1}(g))$, which implies $\|\nabla \bar{f}_i(g)\| \leq G_1$. The smoothness of $\bar{f}_i$ follows from Proposition 3.1. □

### Relationship with Nesterov Smoothing

When $f_i$ is a convex function, its Moreau envelope smoothing is also equivalent to the well-known Nesterov smoothing. To see this, let $f_i^*$ denote the convex conjugate of $f_i$, i.e., $f_i^*(\mathbf{u}) = \max_{g \in \mathbb{R}^{d'}} \mathbf{u}^\top g - f_i(g)$. Since $f_i$ is convex, we have $f_i(g) = \max_{\mathbf{u} \in \mathcal{U}} \mathbf{u}^\top g - f_i^*(\mathbf{u})$, where $\mathcal{U} = \text{dom}(f_i^*)$ is bounded as $\|\partial f_i(g)\| \leq G_1$. As a result,

$$\bar{f}_i(g) = \min_{\mathbf{u} \in \mathbb{R}^{d'}} \frac{\bar{\rho}_1}{2}\|\mathbf{u} - g\|_2^2 + f_i(\mathbf{u}) = \min_{\mathbf{u} \in \mathbb{R}^{d'}} \frac{\bar{\rho}_1}{2}\|\mathbf{u} - g\|_2^2 + \max_{\mathbf{u}' \in \mathcal{U}} \mathbf{u}'^\top \mathbf{u} - f_i^*(\mathbf{u}').$$

By Sion's minimax theorem, we can switch the min and max. Hence,

$$\bar{f}_i(g) = \max_{\mathbf{u}' \in \mathcal{U}} \min_{\mathbf{u} \in \mathbb{R}^{d'}} \frac{\bar{\rho}_1}{2}\|\mathbf{u} - g\|_2^2 + \mathbf{u}'^\top \mathbf{u} - f_i^*(\mathbf{u}').$$

By solving the minimization over $\mathbf{u}$ and plugging the optimal solution into the expression, we get

$$\bar{f}_i(g) = \max_{\mathbf{u}' \in \mathcal{U}} g^\top \mathbf{u}' - f_i^*(\mathbf{u}') - \frac{1}{2\bar{\rho}_1}\|\mathbf{u}'\|_2^2. \tag{5.20}$$

This is known as Nesterov smoothing of the function $f_i(g)$. When $\bar{\rho}_1$ is sufficiently large, we can prove that $\bar{f}_i$ is sufficiently close to $f_i$.

**Example**

**Example 5.1.** *Let us consider the Nesterov smoothing of the hinge function $f(x) = [x]_+$. Let $\bar{\rho}_1 = 1/\varepsilon$ for some small $\varepsilon \ll 1$. Then, the Nesterov smoothing of the hinge function is*

$$\bar{f}(x) = \max_{u \in [0,1]} ux - \frac{\varepsilon}{2}u^2 = \begin{cases} x - \frac{\varepsilon}{2} & \text{if } x \geq \varepsilon \\ \frac{x^2}{2\varepsilon} & \text{if } 0 < x < \varepsilon \\ 0 & \text{o.w.} \end{cases}.$$

*This is also known as the smoothed hinge function.*

**Solving the smoothed problem**

With a smoothed outer function $\bar{f}_i$, we consider optimizing the following problem with some proper value of $\bar{\rho}_1$:

$$\min_{\mathbf{w}} \bar{F}(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} \bar{f}_i(g_i(\mathbf{w})). \tag{5.21}$$

Following Lemma 4.3, $\bar{F}(\cdot)$ is smooth with a smoothness parameter $\bar{L}_F = G_1 L_2 + G_2^2 \bar{L}_1$.

The key concern is how the convergence of solving the above problem translates to the convergence of solving the original problem (5.1). To address this question, we introduce a new convergence measure, named approximate $\epsilon$-stationarity.

**Definition 5.1 (Approximate $\epsilon$-stationary solution)** A point $\mathbf{w}$ is an approximate $\epsilon$-stationary solution to the original problem (5.1), if there exists $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ and $\lambda_i \in \partial f(\mathbf{u}_i), \forall i$ such that

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \nabla g_i(\mathbf{w})\lambda_i \right\|_2 \leq \epsilon, \tag{5.22}$$

$$\|\mathbf{u}_i - g_i(\mathbf{w})\|_2 \leq O(\epsilon), \forall i. \tag{5.23}$$

We note that this condition is closely related to the KKT condition of the following equivalent constrained formulation of the original problem (5.1):

$$\min_{\mathbf{w},\mathbf{u}} \quad \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{u}_i) \tag{5.24}$$

$$\text{s.t.} \quad g_i(\mathbf{w}) = \mathbf{u}_i, \forall i. \tag{5.25}$$

The Lagrangian function of this constrained formulation is given by

$$F(\mathbf{w}, \mathbf{u}, \lambda) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{u}_i) + \sum_{i=1}^{n} \lambda_i^\top (g_i(\mathbf{w}) - \mathbf{u}_i).$$

A solution $(\mathbf{w}, \mathbf{u}, \lambda)$ satisfies the KKT condition, if

$$\frac{1}{n}\sum_{i=1}^{n} \nabla g_i(\mathbf{w})\lambda_i = 0, \quad \lambda_i \in \partial f_i(\mathbf{u}_i)$$

$$\mathbf{u}_i = g_i(\mathbf{w}).$$

Hence, an approximate $\epsilon$-stationary solution satisfies the KKT condition approximately when $\epsilon \ll 1$.

If $f_i$ is $L_1$-smooth, an approximate $\epsilon$-stationary solution is also a standard $O(\epsilon)$-stationary solution. To see this, we have

**Algorithm 17** SONEX

**Input:** learning rate schedules $\{\eta_t\}_{t=1}^T$, $\{\gamma_t\}_{t=1}^T$, $\{\beta_t\}_{t=1}^T$; starting points $\mathbf{w}_0$, $\mathbf{v}_0$, $\mathbf{u}_0$
$\mathbf{w}_1 = \mathbf{w}_0 - \eta_0 \mathbf{v}_0$
**for** $t = 1, \ldots, T$ **do**
  Draw a batch of samples $\mathcal{B}_t \subset [n]$
  **for** $i \in \mathcal{B}_t$ **do**
    Draw two samples $\zeta_{i,t} \sim \mathbb{P}_i$
    Update the inner function value estimators by

$$\text{v1:} \quad \mathbf{u}_{i,t} = (1-\gamma_t)\mathbf{u}_{i,t-1} + \gamma_t g_i(\mathbf{w}_t; \zeta_{i,t})$$
$$\text{v2:} \quad \mathbf{u}_{i,t} = (1-\gamma_t)\mathbf{u}_{i,t-1} + \gamma_t g_i\left(\mathbf{w}_t; \zeta_{i,t}\right) + \gamma_t' \left(g_i\left(\mathbf{w}_t; \zeta_{i,t}\right) - g_i\left(\mathbf{w}_{t-1}; \zeta_{i,t}\right)\right)$$

  **end for**
  Set $\mathbf{u}_{i,t} = \mathbf{u}_{i,t-1}, i \notin \mathcal{B}_t$
  Compute the vanilla gradient estimator $\mathbf{z}_t = \frac{1}{B}\sum_{i\in\mathcal{B}_t'} \nabla g_i(\mathbf{w}_t; \zeta_{i,t}')\nabla \bar{f}_i(\mathbf{u}_{i,t})$
  Update the MA gradient estimator $\mathbf{v}_t = (1-\beta_t)\mathbf{v}_{t-1} + \beta_t \mathbf{z}_t$
  Update the model $\mathbf{w}$ by $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{v}_t$
**end for**

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^n \nabla g_i(\mathbf{w})\nabla f_i(g_i(\mathbf{w}))\right\|_2 \\
&= \left\|\frac{1}{n}\sum_{i=1}^n \nabla g_i(\mathbf{w})\nabla f_i(g_i(\mathbf{w})) - \frac{1}{n}\sum_{i=1}^n \nabla g_i(\mathbf{w})\nabla f_i(\mathbf{u}_i) + \frac{1}{n}\sum_{i=1}^n \nabla g_i(\mathbf{w})\nabla f_i(\mathbf{u}_i)\right\|_2 \\
&\leq \frac{1}{n}\sum_{i=1}^n G_2 L_1 \|\mathbf{u}_i - g_i(\mathbf{w})\|_2 + \epsilon \leq O(\epsilon).
\end{aligned}$$

The following proposition states that an $\epsilon$-stationary solution to the smoothed problem (5.21) is an approximate $\epsilon$-stationary solution to the original problem when $\bar{\rho}_1$ is sufficiently large.

**Proposition 5.1** *Let* $\mathbf{w}$ *be an* $\epsilon$*-stationary solution to (5.21), when* $\bar{\rho}_1 = 1/\epsilon$*, then* $\mathbf{w}$ *is also an approximate* $\epsilon$*-stationary solution to (5.1).*

*Proof.* Given that $\mathbf{w}$ be an $\epsilon$-stationary solution to (5.21), we have

$$\left\|\frac{1}{n}\sum_{i=1}^n \nabla g_i(\mathbf{w})\nabla \bar{f}_i(g_i(\mathbf{w}))\right\|_2 \leq \epsilon.$$

We define $\mathbf{u}_i = \text{prox}_{f_i/\bar{\rho}_1}(g_i(\mathbf{w})) = \arg\min_{\mathbf{u}} f_i(\mathbf{u}) + \frac{\bar{\rho}_1}{2}\|\mathbf{u} - g_i(\mathbf{w})\|_2^2$ and $\lambda_i = \nabla \bar{f}_i(g_i(\mathbf{w}))$. Since $\nabla \bar{f}_i(g_i(\mathbf{w})) \in \partial f_i(\text{prox}_{f_i/\bar{\rho}_1}(g_i(\mathbf{w}))) = \partial f_i(\mathbf{u}_i)$. As a result, we have $\lambda_i \in \partial f_i(\mathbf{u}_i)$ and $\left\|\frac{1}{n}\sum_{i=1}^n \nabla g_i(\mathbf{w})\lambda_i\right\|_2 \leq \epsilon$.

Due to the optimality condition of $\mathbf{u}_i$, we have $g_i(\mathbf{w}) - \mathbf{u}_i \in \partial f_i(\mathbf{u}_i)/\bar{\rho}_1$. Since $f_i$ is $G_1$-Lipschitz continuous and $\bar{\rho}_1 \geq 1/\epsilon$, hence, $\|\mathbf{u}_i - g_i(\mathbf{w})\|_2 \leq O(\epsilon)$. □

Next, we discuss algorithms and complexities for solving the smoothed problem when $\bar{\rho}_1 = 1/\epsilon$. Since both inner and outer functions of the smoothed problem are smooth, we can leverage the moving average gradient estimators. We present detailed steps for solving the smoothed problem in Algorithm 17, which is referred to as SONEX.

A step in implementing SONEX for solving the smoothed problem (5.21) is the calculation of $\nabla \bar{f}_i(\mathbf{u}_{i,t})$, which amounts to solving a proximal mapping of $f_i$, i.e.,

$$\text{prox}_{f_i/\bar{\rho}_1}(\mathbf{u}_{i,t}) = \arg\min_{\mathbf{u}\in\mathbb{R}^{d'}} \frac{\bar{\rho}_1}{2}\|\mathbf{u} - \mathbf{u}_{i,t}\|_2^2 + f_i(\mathbf{u}).$$

In fact, $\nabla \bar{f}_i(\mathbf{u}_{i,t}) = \bar{\rho}_1(\mathbf{u}_{i,t} - \text{prox}_{f_i/\bar{\rho}_1}(\mathbf{u}_{i,t}))$.

### Convergence of SONEX-v1

Finally, we present the complexity of SONEX-v1 for finding an $\epsilon$-stationary solution to the smoothed problem when $\bar{\rho}_1 = 1/\epsilon$.

**Corollary 5.1 (Convergence of SONEX-v1)** *Under Assumptions 5.6 and 5.7, if we set* $\mathbf{u}_0$ *such that* $\frac{1}{n}\mathbb{E}[\sum_{i=1}^n \|\mathbf{u}_{i,0} - g_i(\mathbf{w}_0)\|_2^2] \leq O(\epsilon)$, $\beta = O(\frac{\epsilon^2}{\sigma^2})$, $\gamma = O(\frac{\epsilon^4}{\sigma_0^2})$, $\eta = \min(\epsilon, O(\beta\epsilon), O(\frac{B\epsilon\gamma}{n}))$, $\bar{\rho}_1 = 1/\epsilon > \rho_1$, *then SONEX-v1 finds an approximate* $O(\epsilon)$*-stationary solution to the original problem (5.1) with a complexity of* $O(\frac{n\sigma_0^2}{B\epsilon^7})$.

*Proof.* The proof can be completed by using the convergence result of SOX with noting the order of $\bar{L}_1 = O(\bar{\rho}_1) = O(1/\epsilon)$ and $L_F = O(\bar{L}_1) = O(1/\epsilon)$. □

### Convergence of SONEX-v2

SONEX-v2 is a combination of SOX and MSVR, i.e., with $\mathbf{u}_t$ sequence from MSVR and $\mathbf{v}_t$ from SOX.

**Theorem 5.6 (Convergence of SONEX-v2)** *Under Assumptions 5.6 and 5.7, if we set* $\mathbf{u}_1$ *such that* $\frac{1}{n}\mathbb{E}[\sum_{i=1}^n \|\mathbf{u}_{i,0} - g_i(\mathbf{w}_0)\|_2^2] \leq O(\epsilon^3/\sigma_0)$, $\beta = O(\frac{\epsilon^2}{\sigma^2})$, $\gamma = O(\frac{\epsilon^2}{\sigma_0^2})$, $\eta = \min(O(\epsilon), O(\beta\epsilon), O(\frac{B\sqrt{\gamma}\epsilon}{n}))$ *and* $\bar{\rho}_1 = \frac{1}{\epsilon} > \rho_1$, *then SONEX-v2 finds an approximate* $\epsilon$*-stationary solution to the original problem (5.1) with a complexity of*

$$T = O\left(\max\left\{\frac{1}{\epsilon^3}, \frac{\sigma^2}{\epsilon^5}, \frac{n\sigma_0}{B\epsilon^5}\right\}\right),$$

where $\sigma^2 = \frac{G_1^2\sigma_2^2}{B} + \frac{G_1^2 G_2^2(n-B)}{B(n-1)}$.

*Proof.* The proof is similar to that of Theorem 4.3 except that the $\diamond$ inequality in Lemma 4.10 is replaced by the following for using MSVR estimators (see Lemma 5.5):

$$(\diamond) \quad \mathbb{E}\left[\delta_{t+1}\right] \leq \mathbb{E}[(1-\bar{\gamma})\delta_t + C_3\eta^2\Gamma_t + \bar{\gamma}^2\sigma'^2],$$

where $\bar{\gamma}_{=}\frac{B\gamma}{n}$, $\sigma'^2 = \frac{2n\sigma_0^2}{B}$, $C_3 = O(n/B)$

We only highlight the changes below and leave details as an exercise. First, the condition on $\eta$ in Lemma 4.10 is changed to

$$\eta \leq O\left(\frac{1}{L}, \frac{\beta}{\sqrt{C_2}}, \sqrt{\frac{\bar{\gamma}}{C_1 C_3}}\right).$$

The settings on $\beta, \bar{\gamma}$ remain the same as $\beta = O(\frac{\epsilon^2}{\sigma^2})$, $\bar{\gamma} = O(\frac{\epsilon^2}{C_1\sigma'^2})$. The iteration complexity becomes:

$$\begin{aligned} T &= O\left(\max\left\{\frac{C_\Upsilon L_F}{\epsilon^2}, \frac{C_\Upsilon\sqrt{C_2}}{\epsilon^2\beta}, \frac{C_\Upsilon\sqrt{C_1 C_3}}{\sqrt{\bar{\gamma}}\epsilon^2}\right\}\right) \\ &= O\left(\max\left\{\frac{C_\Upsilon L_F}{\epsilon^2}, \frac{C_\Upsilon\sigma^2\sqrt{C_2}}{\epsilon^4}, \frac{C_\Upsilon\sqrt{C_3}C_1\sigma'}{\epsilon^3}\right\}\right). \end{aligned}$$

and $C_\Upsilon$ is changed to

$$\begin{aligned} C_\Upsilon &= A_0 - A_* + \frac{\eta}{\beta}\Delta_0 + \frac{C_1\eta}{\bar{\gamma}}\delta_0 \leq A_0 - A_* + O\left(\frac{1}{\sqrt{C_2}}\right)\Delta_0 + O\left(\frac{\sqrt{C_1}}{\sqrt{C_3\bar{\gamma}}}\right)\delta_0 \\ &= A_0 - A_* + O\left(\frac{1}{\sqrt{C_2}}\right)\Delta_0 + O\left(\frac{C_1\sigma'}{\sqrt{8C_3}\epsilon}\right)\delta_0. \end{aligned}$$

Then, as in the proof of Theorem 5.1, we substitute $C_1 = O(\bar{L}_1^2)$, $C_2 = O(\bar{L}_F^2)$, $C_3 = O(n/B)$, $\sigma^2 = \frac{G_1^2\sigma_2^2}{B} + \frac{G_1^2G_2^2(n-B)}{B(n-1)}$, and $\sigma'^2 = O(n\sigma_0^2/B)$ into the above complexity expression and $C_\Upsilon$, and obtain

$$\begin{aligned} T &= O\left(\max\left\{\frac{C_\Upsilon\bar{L}_F}{\epsilon^2}, \frac{C_\Upsilon\bar{L}_F\sigma^2}{\epsilon^4}, \frac{C_\Upsilon n\bar{L}_1^2\sigma_0}{B\epsilon^3}\right\}\right), \\ C_\Upsilon &\leq O(F(\mathbf{w}_0) - \bar{F}_*) + O\left(\frac{1}{\bar{L}_F}\right)\Delta_0 + O\left(\frac{\bar{L}_1^2\sigma_0}{\epsilon}\right)\delta_0. \end{aligned}$$

We finish the proof by noting that $\bar{L}_1 = O(1/\epsilon)$ and $\bar{L}_F = O(1/\epsilon)$ and $C_\Upsilon = O(1)$ if $\delta_0 \leq O(\epsilon^3/\sigma_0)$.

□

## 5.4 Convex inner and outer functions

In Chapter 3, we discussed standard stochastic convex optimization and established the iteration complexities of various algorithms. For general convex problems,

**Algorithm 18** ALEXR

**Input:** learning rate schedules $\{\eta_t\}_{t=1}^T$, $\{\alpha_t\}_{t=1}^T$, $\theta \in [0,1]$; starting points $\mathbf{w}_0$, $\mathbf{y}_1 \in \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n$
Let $\mathbf{w}_1 = \mathbf{w}_0$
**for** $t = 1, \ldots, T-1$ **do**
  Sample a batch $\mathcal{B}_t \subset \{1, \ldots, n\}$, $|\mathcal{B}_t| = B$
  **for** each $i \in \mathcal{S}_t$ **do**
    Draw a sample $\zeta_{i,t}, \zeta'_{i,t} \sim \mathbb{P}_i$
    Compute $\tilde{g}_{i,t} = g_i(\mathbf{w}_t; \zeta_{i,t}) + \theta(g_i(\mathbf{w}_t; \zeta_{i,t}) - g_i(\mathbf{w}_{t-1}; \zeta_{i,t}))$
    Update $y_{i,t+1} = \arg\max_{y_i \in \mathcal{Y}_i} \left\{ y_i^\top \tilde{g}_{i,t} - f_i^*(y_i) - \frac{1}{\alpha_t} D_{\psi_i}(y_i, y_{i,t}) \right\}$
  **end for**
  For each $i \notin \mathcal{B}_t$, $y_{i,t+1} = y_{i,t}$
  Compute the vanilla gradient estimator $\mathbf{z}_t = \frac{1}{B} \sum_{i \in \mathcal{B}_t} [\partial g_i(\mathbf{w}_t; \zeta'_{i,t})]^\top y_{i,t+1}$
  Update $\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \left\{ \mathbf{z}_t^\top \mathbf{w} + \frac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_t\|_2^2 + r(\mathbf{w}) \right\}$
**end for**

stochastic gradient descent (SGD) achieves a complexity of $O(1/\epsilon^2)$, while for $\mu$-strongly convex problems, its complexity improves to $O(1/(\mu\epsilon))$. These analyses rely on the assumption of unbiased stochastic gradient estimators, which does not hold for convex compositional optimization problems.

In this section, we introduce stochastic algorithms for a family of convex FCCO problems, where both the inner and outer functions are convex. We establish that these algorithms attain the same order of iteration complexities as SGD in standard stochastic convex optimization. In particular, let us consider a regularized convex FCCO:

$$\min_{\mathbf{w} \in \mathbb{R}^d} F(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^n f_i(g_i(\mathbf{w})) + r(\mathbf{w}), \tag{5.26}$$

where $g_i(\mathbf{w}) = \mathbb{E}_{\zeta \sim \mathbb{P}_i}[g_i(\mathbf{w}; \zeta)]$, the outer and inner functions satisfy the following assumption.

**Assumption 5.8.** *The following conditions hold:*

*(i) $f_i$ is convex, $G_1$-Lipschitz continuous, and $\partial f_i(\cdot) \geq 0$.*
*(ii) $g_i$ is convex and $G_2$-Lipschitz continuous.*
*(iii) $r$ is $\mu$-strongly convex for some $\mu \geq 0$.*

Group DRO (5.2) could satisfy the above assumption when the individual loss function is convex and Lipschitz with respect to the model parameter. Two-way partial AUC maximization considered in Section 6.4.3 is another example satisfying the above assumption when the loss function is convex and Lipschitz continuous.

Let $f_i^*$ denote the convex conjugate of $f_i$. We can write $f_i(g_i(\mathbf{w}))$ as

$$f_i(g_i(\mathbf{w})) = \max_{y_i \in \mathcal{Y}_i} (y_i^\top g_i(\mathbf{w}) - f_i^*(y_i)),$$

where $\mathcal{Y}_i = \mathrm{dom}(f_1^*)$. Since $0 \leq \partial f_i(\cdot)$ and $\|\partial f_i(\cdot)\| \leq G_1$, hence $\mathcal{Y}_i$ is a compact set following from Lemma 1.8.

Then, we can convert (5.26) into an equivalent minimax optimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \max_{\mathbf{y}\in\mathcal{Y}} \frac{1}{n}\sum_{i=1}^{n}(y_i^\top g_i(\mathbf{w}) - f_i^*(y_i)) + r(\mathbf{w}), \tag{5.27}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \mathcal{Y}_n$. Thus, the above problem is convex-concave problem under Assumption 5.8.

We introduce a method to optimize the above minimax problem. However, there are several unique challenges: (i) updating all coordinates of $\mathbf{y}$ is difficult because it is computationally prohibitive to traverse all data points $i = 1, \ldots, n$ at each iteration; (ii) we only have access to stochastic evaluations of the functions $g_i(\mathbf{w}; \zeta)$, which limits our ability to update both the corresponding coordinate of $\mathbf{y}$ and the parameter $\mathbf{w}$.

### 5.4.1 The ALEXR Algorithm

To present the algorithm, we assume a strongly convex prox-function $\psi_i$ for the $i$-th coordinate and impose the following conditions.

**Assumption 5.9.** *Suppose $\psi_i$ is differentiable and obeys the following conditions*

*(i) $\psi_i$ is $\mu_\psi$-strongly convex with respect to $\|\cdot\|_2$, i.e., $\psi_i(y) \geq \psi_i(y') + \nabla\psi_i(y')^\top(y - y') + \frac{\mu_\psi}{2}\|y - y'\|_2^2$.*
*(ii) $D_{f_i^*}(y, y') \geq \rho D_{\psi_i}(y, y')$ for some $\rho \geq 0$.*
*(iii) The following proximal mapping can be easily computed:*

$$y_{i,t+1} = \arg\max_{y_i\in\mathcal{Y}_i} \left\{ y_i^\top \tilde{g}_{i,t} - f_i^*(y_i) - \frac{1}{\alpha_t} D_{\psi_i}(y_i, y_{i,t}) \right\}.$$

A meta-algorithm, termed ALEXR, is presented in Algorithm 18. ALEXR employs stochastic block-coordinate proximal mirror ascent to update $\mathbf{y}$, using the prox-function $\psi_i$ for the $i$-th coordinate, and applies stochastic proximal gradient descent to update $\mathbf{w}$. Below, we consider different choices of the prox-functions $\psi_i$ and the corresponding updates for $y_{i,t+1}$.

#### ALEXR-v1 for smooth $f_i$: using $\psi_i = f_i^*$

When $f_i$ is $L_1$-smooth, its convex conjugate $f_i^*$ is $1/L_1$-strongly convex. We can use $\psi_i = f_i^*$ to define a Bregman divergence $D_{\psi_i}(y, y') = D_{f_i^*}(y, y')$.

**Critical:** In this case, Assumption 5.9 (i) and (ii) hold with $\mu_\psi = 1/L_1$, and $\rho = 1$.

Let us consider the update of $y_{i,t+1}$, which becomes:

$$y_{i,t+1} = \arg\max_{y_i \in \mathcal{Y}_i} \left\{ y_i^\top \tilde{g}_{i,t} - f_i^*(y_i) - \frac{1}{\alpha_t} D_{f_i^*}(y_i, y_{i,t}) \right\}, \forall i \in \mathcal{B}_t. \tag{5.28}$$

The following lemma provides an efficient way to compute $y_{i,t+1}$, which also builds the connection to the sequence of $\mathbf{u}_{i,t}$ in SOX and MSVR.

**Lemma 5.14** *Let $\mathbf{u}_{i,t-1} \in \partial f_i^*(y_{i,t})$. Then for $i \in \mathcal{B}_t$ we have $y_{i,t+1} = \nabla f_i(\mathbf{u}_{i,t})$, where $\mathbf{u}_{i,t} = \frac{1}{1+\alpha_t}\mathbf{u}_{i,t-1} + \frac{\alpha_t}{1+\alpha_t}\tilde{g}_{i,t}$.*

*Proof.* For the problem (5.28), we have

$$\begin{aligned}
&y_i^\top \tilde{g}_{i,t} - f_i^*(y_i) - \frac{1}{\alpha_t} D_{f_i^*}(y_i, y_{i,t}) \\
&= y_i^\top \tilde{g}_{i,t} - f_i^*(y_i) - \frac{1}{\alpha_t}(f_i^*(y_i) - \partial f_i^*(y_{i,t})^\top (y_i - y_{i,t}) - f_i^*(y_{i,t})) \\
&= y_i^\top (\tilde{g}_{i,t} + \frac{1}{\alpha_t}\partial f_i^*(y_{i,t})) - (1 + \frac{1}{\alpha_t}) f_i^*(y_i) - \frac{1}{\alpha_t}\partial f_i^*(y_{i,t})^\top y_{i,t} + \frac{1}{\alpha_t} f_i^*(y_{i,t}).
\end{aligned}$$

Hence $y_{i,t+1} \in \arg\max_{y_i \in \mathcal{Y}_i} y_i^\top (\frac{\alpha_t}{1+\alpha_t}\tilde{g}_{i,t} + \frac{1}{1+\alpha_t}\partial f_i^*(y_{i,t})) - f_i^*(y_i)$. If we define $\mathbf{u}_{i,t} = \frac{\alpha_t}{1+\alpha_t}\tilde{g}_{i,t} + \frac{1}{1+\alpha_t}\partial f_i^*(y_{i,t})$, we have

$$f(\mathbf{u}_{i,t}) = \max_{y_i \in \mathcal{Y}_i} y_i^\top \mathbf{u}_{i,t} - f_i^*(y_i) = y_{i,t+1}^\top \mathbf{u}_{i,t} - f_i^*(y_{i,t+1}).$$

Hence, $\mathbf{u}_{i,t} \in \arg\max_{\mathbf{u}} y_{i,t+1}^\top \mathbf{u} - f_i(\mathbf{u})$ and therefore $y_{i,t+1} = \nabla f_i(\mathbf{u}_{i,t})$. □

If $f_i$ is a Legendre function such that $\nabla f_i^{-1} = \nabla f_i^*$ (see Lemma 1.8). Then, we can derive the following equivalent update of $\mathbf{u}$ sequence such that $y_{i,t} = \nabla f_i(\mathbf{u}_{i,t-1})$.

$$\mathbf{u}_{i,t} = \begin{cases} \frac{1}{1+\alpha_t}\mathbf{u}_{i,t-1} + \frac{\alpha_t}{1+\alpha_t}\tilde{g}_{i,t}, & \text{if } i \in \mathcal{B}_t \\ \mathbf{u}_{i,t-1} & \text{o.w.} \end{cases}. \tag{5.29}$$

When $\theta = 0$, the equivalent $\mathbf{u}$ update (5.64) becomes:

$$\mathbf{u}_{i,t} = (1 - \frac{\alpha_t}{1+\alpha_t})\mathbf{u}_{i,t-1} + \frac{\alpha_t}{1+\alpha_t} g_i(\mathbf{w}_t; \zeta_{i,t}), \forall i \in \mathcal{B}_t. \tag{5.30}$$

This is the same as the moving average estimator in SOX with $\gamma_t = \alpha_t/(1+\alpha_t)$. Using the equivalent $\mathbf{u}$ sequence, the stochastic gradient estimator becomes $\mathbf{z}_t = \frac{1}{B}\sum_{i \in \mathcal{B}_t} [\partial g_i(x_t; \zeta'_{i,t})]^\top \nabla f_i(\mathbf{u}_{i,t})$. If the regularizer $r$ is not present, the update of the model parameter becomes $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{z}_t$. In this case, ALEXR with $\theta = 0$ is the same as SOX with $\beta_t = 1$. We will prove its convergence for convex and strongly convex regularizer $r$.

When $\theta > 0$, the equivalent $\mathbf{u}$ update (5.64) becomes:

$$\mathbf{u}_{i,t} = (1 - \frac{\alpha_t}{1+\alpha_t})\mathbf{u}_{i,t-1} + \frac{\alpha_t}{1+\alpha_t} g_i(\mathbf{w}_t; \zeta_{i,t}) + \frac{\theta\alpha_t}{1+\alpha_t}(g_i(\mathbf{w}_t; \zeta_t) - g_i(\mathbf{w}_{t-1}; \zeta_t)). \tag{5.31}$$

This is similar to the MSVR estimator with $\gamma_t = \frac{\alpha_t}{1+\alpha_t}$ and $\gamma_t' = \frac{\theta\alpha_t}{1+\alpha_t}$. However, the key difference is that $\gamma_t'$ in MSVR is larger than 1, while it is smaller than 1 in ALEXR for convex problems. In practice, setting $\gamma_t' < 1$ is a better choice. We will prove a better convergence of ALEXR with $\theta \in (0, 1)$ for a strongly convex $r$.

#### ALEXR-v2 for non-smooth $f_i$: using a quadratic function $\psi_i(\cdot)$

When $f_i$ is non-smooth, we cannot use $f_i^*$ as the prox function. In this case, we will use a smooth and strongly convex $\psi_i$. a quadratic function $\psi_i(y) = \frac{1}{2}\|y\|_2^2$.

**Critical:** In this case, Assumption 5.9 (i) holds with $\mu_\psi = 1$, and Assumption 5.9 (ii) holds with $\rho = 0$.

**Example**

**Example 5.2.** *For the update of $y_{i,t+1}$, consider the example $f_i(\cdot) = [\,\cdot\,]_+$, as used in GDRO and TPAUC maximization. In this case, the conjugate $f_i^*(y)$ is the indicator function of the interval $[0, 1]$. Consequently, $y_{i,t+1}$ can be computed as:*

$$y_{i,t+1} = \arg\max_{y_i \in [0,1]} \left\{ y_i^\top \tilde{g}_{i,t} - \frac{1}{2\alpha_t}(y_i - y_{i,t})^2 \right\} = \Pi_{[0,1]}(y_{i,t} - \alpha_t \tilde{g}_{i,t}), \forall i \in \mathcal{B}_t,$$

*where $\Pi_{[0,1]}(\cdot)$ projects the input into the range of $[0, 1]$.*

### 5.4.2 Technical Lemmas

To facilitate the analysis, we define $(\mathbf{w}_*, \mathbf{y}_*)$ as the saddle point to the minimax problem and

$$
\begin{aligned}
&F(\mathbf{w},\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} y_i^\top g_i(\mathbf{w}) - f_i^*(y_i) + r(\mathbf{w}),\\
&\tilde{g}_t = (\tilde{g}_{1,t},\dots,\tilde{g}_{n,t})^\top,\\
&\bar{y}_{i,t+1} = \arg\max_{y_i\in\mathcal{Y}_i}\left\{y_i^\top \tilde{g}_{i,t} - f_i^*(y_i) - \frac{1}{\alpha_t}D_{\psi_i}(y_i,y_{i,t})\right\}, \forall i\in[n]\\
&D_\psi(\mathbf{y},\mathbf{y}') = \sum_{i=1}^{n} D_{\psi_i}(y_i,y_i').
\end{aligned}
$$

Note that $\bar{\mathbf{y}}_{t+1}$ is a virtual sequence, which is updated for all coordinates from $\mathbf{y}_t$ making it independent of $\mathcal{B}_t$.

We make the following assumption regarding the stochastic estimators.

**Assumption 5.10.** *We assume that*

*(i)* $\mathbb{E}_{\zeta\sim\mathbb{P}_i}[\|g_i(\mathbf{w};\zeta) - g_i(\mathbf{w})\|_2^2] \le \sigma_0^2$.

*(ii)* $\mathbb{E}_{\zeta\sim\mathbb{P}_i}[\|\nabla g_i(\mathbf{w};\zeta) - \nabla g_i(\mathbf{w})\|_2^2] \le \sigma_2^2$.

*(iii)* $\mathbb{E}_{i\sim\mathbb{U}_n}\left[\left\|y_i\nabla g_i(\mathbf{w}) - \frac{1}{n}\sum_{i=1}^n y_i\nabla g_i(\mathbf{w})\right\|_2^2\right] \le \delta^2$ *for any fixed* $\mathbf{y}$*, where* $\mathbb{U}_n$ *denotes a uniform distribution.*

**Lemma 5.15** *The following holds for any* $\mathbf{w},\mathbf{y}\in\mathcal{Y}$ *after the* $t$*-th iteration of Algorithm 18.*

$$
\begin{aligned}
&F(\mathbf{w}_{t+1},\mathbf{y}) - F(\mathbf{w},\bar{\mathbf{y}}_{t+1}) \qquad (5.32)\\
&\le \frac{1}{2\eta_t}\|\mathbf{w}-\mathbf{w}_t\|_2^2 - (\frac{1}{2\eta_t}+\frac{\mu}{2})\|\mathbf{w}-\mathbf{w}_{t+1}\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2\\
&+ A_t(\mathbf{y}) + B_t(\mathbf{y}) + C_t(\mathbf{w}),
\end{aligned}
$$

*where*

$$
\begin{aligned}
A_t(\mathbf{y}) &= \frac{1}{n\alpha_t}D_\psi(\mathbf{y},\mathbf{y}_t) - (\frac{1}{n\alpha_t}+\frac{\rho}{n})D_\psi(\mathbf{y},\bar{\mathbf{y}}_{t+1}) - \frac{1}{n\alpha_t}D_\psi(\bar{\mathbf{y}}_{t+1},\mathbf{y}_t)\\
B_t(\mathbf{y}) &= \frac{1}{n}\sum_{i=1}^{n}(g_i(\mathbf{w}_{t+1}) - \tilde{g}_{i,t})^\top(y_i - \bar{y}_{i,t+1})\\
C_t(\mathbf{w}) &= \frac{1}{n}\sum_{i=1}^{n}(g_i(\mathbf{w}_{t+1}) - g_i(\mathbf{w}))^\top\bar{y}_{i,t+1} - \mathbf{z}_t^\top(\mathbf{w}_{t+1}-\mathbf{w}).
\end{aligned}
$$

*Proof.* Following Lemma 3.10, for all $i\in[n]$ the dual update rule implies that for any $y\in\mathcal{Y}$ it holds

$$
\begin{aligned}
&\tilde{g}_{i,t}^\top(y_i - \bar{y}_{i,t+1}) + f_i^*(\bar{y}_{i,t+1}) - f_i^*(y_i)\\
&\le \frac{1}{\alpha_t}D_{\psi_i}(y_i,y_{i,t}) - (\frac{1}{\alpha_t}+\rho)D_{\psi_i}(y_i,\bar{y}_{i,t+1}) - \frac{1}{\alpha_t}D_{\psi_i}(\bar{y}_{i,t+1},y_{i,t}).
\end{aligned}
$$

Averaging this inequality over $i = 1,\dots,n$.

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{g}_{i,t}^{\top}(y_{i,t}-\bar{y}_{i,t+1})+\frac{1}{n}\sum_{i=1}^{n}f_i^*(\bar{y}_{i,t+1})-\frac{1}{n}\sum_{i=1}^{n}f_i^*(y_i) \tag{5.33}$$
$$\leq \frac{1}{n\alpha_t}D_\psi(\mathbf{y},\mathbf{y}_t)-(\frac{1}{n\alpha_t}+\frac{\rho}{n})D_\psi(\mathbf{y},\bar{\mathbf{y}}_{t+1})-\frac{1}{n\alpha_t}D_\psi(\bar{\mathbf{y}}_{t+1},\mathbf{y}_t).$$

According to Lemma 3.6, the primal update rule implies that

$$\mathbf{z}_t^{\top}(\mathbf{w}_{t+1}-\mathbf{w})+r(\mathbf{w}_{t+1})-r(\mathbf{w}) \tag{5.34}$$
$$\leq \frac{1}{2\eta_t}\|\mathbf{w}-\mathbf{w}_t\|_2^2-(\frac{1}{2\eta_t}+\frac{\mu}{2})\|\mathbf{w}-\mathbf{w}_{t+1}\|_2^2-\frac{1}{2\eta_t}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2.$$

By the definition of $F(\mathbf{w},\mathbf{y})$, we have

$$\begin{aligned}
&F(\mathbf{w}_{t+1},\mathbf{y})-F(\mathbf{w},\bar{\mathbf{y}}_{t+1})\\
&=\frac{1}{n}\sum_{i=1}^{n}y_i^{\top}g_i(\mathbf{w}_{t+1})-\frac{1}{n}\sum_{i=1}^{n}f_i^*(y_i)+r(\mathbf{w}_{t+1})-\frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i,t+1}^{\top}g_i(\mathbf{w})\\
&+\frac{1}{n}\sum_{i=1}^{n}f_i^*(\bar{y}_{i,t+1})-r(\mathbf{w})\\
&=\frac{1}{n}\sum_{i=1}^{n}g_i(\mathbf{w}_{t+1})^{\top}(y_i-\bar{y}_{i,t+1})+\frac{1}{n}\sum_{i=1}^{n}f_i^*(\bar{y}_{i,t+1})-\frac{1}{n}\sum_{i=1}^{n}f_i^*(y_i)\\
&+\frac{1}{n}\sum_{i=1}^{n}(g_i(\mathbf{w}_{t+1})-g_i(\mathbf{w}))^{\top}\bar{y}_{i,t+1}+r(\mathbf{w}_{t+1})-r(\mathbf{w}).
\end{aligned}$$

Combining the equation above with (5.34) and (5.33), we can finish the proof. □

Next, we bound the three terms $A_t(\mathbf{y}), B_t(\mathbf{y}), C_t(\mathbf{w})$ separately.

**Lemma 5.16** *Let $\tau_t = 1/\alpha_t$. For $\mathbf{y}$ that possibly depends on all randomness in the algorithm and any $\lambda_0 > 0$, we have*

$$\mathbb{E}[A_t(\mathbf{y})]=\mathbb{E}\left[\frac{\tau_t}{n}D_\psi(\mathbf{y},\mathbf{y}_t)-\frac{\tau_t+\rho}{n}D_\psi(\mathbf{y},\bar{\mathbf{y}}_{t+1})-\frac{\tau_t}{n}D_\psi(\bar{\mathbf{y}}_{t+1},\mathbf{y}_t)\right] \tag{5.35}$$
$$\leq \mathbb{E}\left[\frac{\tau_t+\rho\left(1-\frac{B}{n}\right)}{B}D_\psi(\mathbf{y},\mathbf{y}_t)-\frac{\tau_t+\rho}{B}D_\psi(\mathbf{y},\mathbf{y}_{t+1})\right]-\frac{\tau_t}{n}\mathbb{E}\left[D_\psi(\bar{\mathbf{y}}_{t+1},\mathbf{y}_t)\right]$$
$$+\mathbb{E}\left[\frac{\lambda_0(\tau_t+\rho)}{n}(D_\psi(\mathbf{y},\hat{\mathbf{y}}_t)-D_\psi(\mathbf{y},\hat{\mathbf{y}}_{t+1}))\right]$$
$$+\frac{(n-B)(\tau_t+\rho)}{2\mu_\psi\lambda_0 nB}\mathbb{E}\left[\sum_{i=1}^{n}\left\|\nabla\psi_i(\bar{y}_{i,t+1})-\nabla\psi_i(y_{i,t})\right\|_2^2\right],$$

*where the sequence $\{\hat{\mathbf{y}}_t\}_t, \hat{\mathbf{y}}_t \in \mathcal{Y}$ is virtual. In addition, for $\mathbf{y}_*$, we have*

$$\mathbb{E}[A_t(\mathbf{y}_*)] = \mathbb{E}\left[\frac{\tau_t}{n}D_\psi(\mathbf{y}_*, \mathbf{y}_t) - \frac{\tau_t+\rho}{n}D_\psi(\mathbf{y}_*, \bar{\mathbf{y}}_{t+1}) - \frac{\tau_t}{n}D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t)\right] \tag{5.36}$$
$$\leq \mathbb{E}\left[\frac{\tau_t+\rho\left(1-\frac{B}{n}\right)}{B}D_\psi(\mathbf{y}_*, \mathbf{y}_t) - \frac{\tau_t+\rho}{B}D_\psi(\mathbf{y}_*, \mathbf{y}_{t+1})\right] - \frac{\tau_t}{n}\mathbb{E}\left[D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t)\right].$$

*Proof.*

$$\begin{aligned}
&\frac{\tau_t}{n}D_\psi(\mathbf{y}, \mathbf{y}_t) - \frac{\tau_t+\rho}{n}D_\psi(\mathbf{y}, \bar{\mathbf{y}}_{t+1}) - \frac{\tau_t}{n}D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t) \qquad (5.37)\\
&= \frac{\tau_t+\rho\left(1-\frac{B}{n}\right)}{B}D_\psi(\mathbf{y}, \mathbf{y}_t) - \frac{\tau_t+\rho}{B}D_\psi(\mathbf{y}, \mathbf{y}_{t+1}) - \frac{\tau_t}{n}D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t)\\
&+ \left(\frac{\tau_t+\rho}{B}D_\psi(\mathbf{y}, \mathbf{y}_{t+1}) - \frac{\tau_t+\rho}{n}D_\psi(\mathbf{y}, \bar{\mathbf{y}}_{t+1}) + \frac{(B-n)(\tau_t+\rho)}{nB}D_\psi(\mathbf{y}, \mathbf{y}_t)\right).
\end{aligned}$$

For bounding the last three terms, we consider the following:

$$\begin{aligned}
&\frac{1}{B}D_{\psi_i}(y_i, y_{i,t+1}) - \frac{1}{n}D_{\psi_i}(y_i, \bar{y}_{i,t+1}) + \frac{(B-n)}{nB}D_{\psi_i}(y_i, y_{i,t}) \qquad (5.38)\\
&= \frac{1}{B}\left(\psi_i(y_i) - \psi_i(y_{i,t+1}) - \nabla\psi_i(y_{i,t+1})^\top(y_i - y_{i,t+1})\right)\\
&- \frac{1}{n}\left(\psi_i(y_i) - \psi_i(\bar{y}_{i,t+1}) - \nabla\psi_i(\bar{y}_{i,t+1})^\top(y_i - \bar{y}_{i,t+1})\right)\\
&+ \frac{(B-n)}{nB}\left(\psi_i(y_i) - \psi_i(y_{i,t}) - \nabla\psi_i(y_{i,t})^\top(y_i - y_{i,t})\right)\\
&= \left[\frac{1}{n}\left(\psi_i(\bar{y}_{i,t+1}) - \frac{n}{B}\psi_i(y_{i,t+1}) + \frac{n-B}{B}\psi_i(y_{i,t})\right)\right]\\
&+ \left[\frac{1}{B}\nabla\psi_i(y_{i,t+1})^\top y_{i,t+1} - \frac{1}{n}\nabla\psi_i(\bar{y}_{i,t+1})^\top\bar{y}_{i,t+1} + \frac{(B-n)}{nB}\nabla\psi_i(y_{i,t})^\top y_{i,t}\right]\\
&+ \underbrace{\frac{1}{n}(-\frac{n}{B}\nabla\psi_i(y_{i,t+1}) + \nabla\psi_i(\bar{y}_{i,t+1}) + \frac{n-B}{B}\nabla\psi_i(y_{i,t}))^\top y_i}_{\sharp}.
\end{aligned}$$

Taking expectation over $\mathcal{B}_t$ for the first two terms in the brackets of the above bound will give zeros. This is because that both $\bar{y}_{i,t+1}$ and $y_{i,t}$ are independent of $\mathcal{B}_t$ such that

$$\begin{aligned}
\mathbb{E}_{\mathcal{B}_t}[\psi_i(y_{i,t+1})] &= \frac{B}{n}\psi_i(\bar{y}_{i,t+1}) + \frac{n-B}{n}\psi_i(y_{i,t}),\\
\mathbb{E}_{\mathcal{B}_t}\left[\nabla\psi_i(y_{i,t+1})^\top y_{i,t+1}\right] &= \frac{B}{n}\nabla\psi_i(\bar{y}_{i,t+1})^\top\bar{y}_{i,t+1} + \frac{n-B}{n}\nabla\psi_i(y_{i,t})^\top y_{i,t},\\
\mathbb{E}_{\mathcal{B}_t}\left[\nabla\psi_i(y_{i,t+1})\right] &= \frac{B}{n}\nabla\psi_i(\bar{y}_{i,t+1}) + \frac{n-B}{n}\nabla\psi_i(y_{i,t}).
\end{aligned}$$

Next, we bound the $\sharp$ term. When $\mathbf{y} = \mathbf{y}_*$, expectation of $\sharp$ is also zero which proves (5.36).

When $\mathbf{y}$ is possibly random, let us apply Lemma 3.13 to the update $\hat{y}_{i,t+1} = \arg\min_v -\Delta_{i,t}^\top v + \lambda_0 D_{\psi_i}(v, \hat{y}_{i,t}), \forall i$ ($\lambda_0$ to be determined), where

$$\Delta_{i,t} \coloneqq -\frac{n}{B}\nabla\psi_i(y_{i,t+1}) + \nabla\psi_i(\bar{y}_{i,t+1}) + \frac{n-B}{B}\nabla\psi_i(y_{i,t})$$

is a martingale sequence due to

$$\mathbb{E}_{\mathcal{B}_t}\left[(\nabla\psi_i(y_{i,t+1}) - \nabla\psi_i(y_{i,t}))\right] = \frac{B}{n}(\nabla\psi_i(\bar{y}_{i,t+1}) - \nabla\psi_i(y_{i,t})).$$

We have

$$\mathbb{E}[\sharp] \le \mathbb{E}\left[\frac{\lambda_0}{n}(D_{\psi_i}(y_i, \hat{y}_{i,t}) - D_{\psi_i}(y_i, \hat{y}_{i,t+1}))\right] + \frac{1}{2n\mu_\psi\lambda_0}\mathbb{E}\left[\left\|\Delta_{i,t}\right\|_2^2\right].$$

Note that $\mathbb{E}_{\mathcal{B}_t}\left[(\nabla\psi_i(y_{i,t+1}) - \nabla\psi_i(y_{i,t}))\right] = \frac{B}{n}(\nabla\psi_i(\bar{y}_{i,t+1}) - \nabla\psi_i(y_{i,t}))$ such that

$$\begin{aligned}
\mathbb{E}_{\mathcal{B}_t}\left[\left\|\Delta_{i,t}\right\|_2^2\right] &= \mathbb{E}_{\mathcal{B}_t}\left\|(\nabla\psi_i(\bar{y}_{i,t+1}) - \nabla\psi_i(y_{i,t})) - \frac{n}{B}(\nabla\psi_i(y_{i,t+1}) - \nabla\psi_i(y_{i,t})\right\|_2^2 \\
&\le \frac{n^2}{B^2}\mathbb{E}_{\mathcal{B}_t}\left\|\nabla\psi_i(y_{i,t+1}) - \nabla\psi_i(y_{i,t})\right\|_2^2 - \left\|(\nabla\psi_i(\bar{y}_{i,t+1}) - \nabla\psi_i(y_{i,t}))\right\|_2^2 \\
&\le \frac{n}{B}\left\|\nabla\psi_i(\bar{y}_{i,t+1}) - \nabla\psi_i(y_{i,t})\right\|_2^2 - \left\|(\nabla\psi_i(\bar{y}_{i,t+1}) - \nabla\psi_i(y_{i,t}))\right\|_2^2.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\mathbb{E}[\sharp] \le{}& \mathbb{E}\left[\frac{\lambda_0}{n}(D_{\psi_i}(y_i, \hat{y}_{i,t}) - D_{\psi_i}(y_i, \hat{y}_{i,t+1}))\right] \\
&+ \frac{n-B}{2\mu_\psi\lambda_0 nB}\mathbb{E}\left[\left\|\nabla\psi_i(\bar{y}_{i,t+1}) - \nabla\psi_i(y_{i,t})\right\|_2^2\right].
\end{aligned}$$

Averaging (5.38) multiplied by $\tau_t + \rho$ and combining (5.37) finishes the proof. □

**Lemma 5.17** *Suppose $\psi_i$ is $\mu_\psi$-strongly convex. For any $\lambda_2, \lambda_3, \lambda_4, \lambda_5 > 0$ and $\mathbf{y}$ that possibly depends on all randomness in the algorithm, we have*

$$\begin{aligned}
\mathbb{E}[B_t(\mathbf{y})] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(g_i(\mathbf{w}_{t+1}) - \tilde{g}_{i,t})^\top(y_i - \bar{y}_{i,t+1}) &\le \mathbb{E}[\Gamma_{t+1} - \theta\Gamma_t] \qquad (5.39)\\
+ \frac{(\lambda_3 + \lambda_4\theta)\mathbb{E}[D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t)]}{\mu_\psi n} &+ \frac{G_2^2\mathbb{E}\left\|\mathbf{w}_{t+1} - \mathbf{w}_t\right\|_2^2}{2\lambda_3} + \frac{G_2^2\theta\mathbb{E}\left\|\mathbf{w}_t - \mathbf{w}_{t-1}\right\|_2^2}{2\lambda_4} \\
+ \frac{(1 + 3.5\theta + 3.5\theta^2)\sigma_0^2\alpha_t}{\mu_\psi} &+ \frac{(1+\theta)\lambda_2\sigma_0^2}{2\mu_\psi} + \frac{\theta\sigma_0^2\lambda_5}{2\mu_\psi} \\
+ \frac{1+\theta}{n\lambda_2}\mathbb{E}[D_\psi(\mathbf{y}, \tilde{\mathbf{y}}_t) - D_\psi(\mathbf{y}, \tilde{\mathbf{y}}_{t+1})] &+ \frac{\theta}{n\lambda_5}\mathbb{E}[D_\psi(\mathbf{y}, \check{\mathbf{y}}_t) - D_\psi(\mathbf{y}, \check{\mathbf{y}}_{t+1})],
\end{aligned}$$

*where* $\Gamma_t \coloneqq \frac{1}{n}\sum_{i=1}^n (g_i(\mathbf{w}_t) - g_i(\mathbf{w}_{t-1}))^\top (y_i - y_{i,t})$ *and* $\check{\mathbf{y}}_t, \tilde{\mathbf{y}}_t$ *are some virtual sequences. In addition, we have*

$$\mathbb{E}[B_t(\mathbf{y}_*)] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(g_i(\mathbf{w}_{t+1}) - \tilde{g}_{i,t})^\top (y_{i,*} - \bar{y}_{i,t+1}) \le \mathbb{E}[\Gamma^*_{t+1} - \theta\Gamma^*_t] \tag{5.40}$$
$$+ \frac{(\lambda_3 + \lambda_4\theta)\mathbb{E}[D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t)]}{\mu_\psi n} + \frac{G_2^2 \mathbb{E}\,\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2}{2\lambda_3} + \frac{G_2^2\theta\mathbb{E}\,\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2}{2\lambda_4}$$
$$+ \frac{(1 + 3.5\theta + 3.5\theta^2)\sigma_0^2\alpha_t}{\mu_\psi},$$

*where* $\Gamma^*_t \coloneqq \frac{1}{n}\sum_{i=1}^n (g_i(\mathbf{w}_t) - g_i(\mathbf{w}_{t-1}))^\top (y_{i,*} - y_{i,t})$.

*Proof.* Since

$$\tilde{g}_{i,t} = g_i(\mathbf{w}_t; \zeta_{i,t}) + \theta(g_i(\mathbf{w}_t; \zeta_{i,t}) - g_i(\mathbf{w}_{t-1}; \zeta_{i,t})),$$

we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n (g_i(\mathbf{w}_{t+1}) - \tilde{g}_{i,t})^\top (y_i - \bar{y}_{i,t+1}) \\
&= \underbrace{\frac{1+\theta}{n}\sum_{i=1}^n (g_i(\mathbf{w}_t) - g_i(\mathbf{w}_t; \zeta_{i,t}))^\top (y_i - \bar{y}_{i,t+1})}_{\text{I}} \\
&+ \underbrace{\frac{\theta}{n}\sum_{i=1}^n (g_i(\mathbf{w}_{t-1}; \zeta_{i,t}) - g_i(\mathbf{w}_{t-1}))^\top (y_i - \bar{y}_{i,t+1})}_{\text{II}} \\
&+ \underbrace{\frac{1}{n}\sum_{i=1}^n (g_i(\mathbf{w}_{t+1}) - g_i(\mathbf{w}_t)) + \theta(g_i(\mathbf{w}_{t-1}) - g_i(\mathbf{w}_t))^\top (y_i - \bar{y}_{i,t+1})}_{\text{III}}.
\end{aligned} \tag{5.41}$$

Define

$$\dot{y}_{i,t+1} \coloneqq \arg\max_{v \in \mathcal{Y}_i} \{v^\top((1+\theta)g_i(\mathbf{w}_t) - \theta g_i(\mathbf{w}_{t-1})) - f_i^*(v) - \frac{1}{\alpha_t} D_{\psi_i}(v, y_{i,t})\}, \forall i \in [n].$$

This update differs from that of $\bar{y}_{i,t+1}$ in that it uses full gradients instead of stochastic gradients. We decompose the I term in (5.41) as

$$
\begin{aligned}
\mathrm{I} &= \frac{1+\theta}{n}\sum_{i=1}^{n}(g_i(\mathbf{w}_t) - g_i(\mathbf{w}_t;\zeta_{i,t}))^\top(y_i - \bar{y}_{i,t+1}) \\
&= \underbrace{\frac{1+\theta}{n}\sum_{i=1}^{n}(g_i(\mathbf{w}_t) - g_i(\mathbf{w}_t;\zeta_{i,t}))^\top(\dot{y}_{i,t+1} - \bar{y}_{i,t+1})}_{\mathrm{I}_1} \\
&+ \underbrace{\frac{1+\theta}{n}\sum_{i=1}^{n}(g_i(\mathbf{w}_t) - g_i(\mathbf{w}_t;\zeta_{i,t}))^\top y_i}_{\mathrm{I}_2} - \underbrace{\frac{1+\theta}{n}\sum_{i=1}^{n}(g_i(\mathbf{w}_t) - g_i(\mathbf{w}_t;\zeta_{i,t}))^\top \dot{y}_{i,t+1}}_{\mathrm{I}_3}.
\end{aligned}
$$

Taking expectation over $\zeta_{i,t}, \forall i$ will make $\mathbb{E}_{\zeta_t}[\mathrm{I}_3] = 0$. Below, we will bound $\mathrm{I}_1$ and $\mathrm{I}_2$.

$$
\mathrm{I}_1 \le \frac{1+\theta}{n}\sum_{i=1}^{n}\left\|g_i(\mathbf{w}_t) - g_i(\mathbf{w}_t;\zeta_{i,t})\right\|_2 \left\|\dot{y}_{i,t+1} - \bar{y}_{i,t+1}\right\|_2.
$$

Since $D_{\psi_i}(y_i, y_{i,t})$ is $\mu_\psi$-strongly convex, Lemma 3.8 implies that

$$
\begin{aligned}
&\|\dot{y}_{i,t+1} - \bar{y}_{i,t+1}\|_2 \\
&\le \frac{\alpha_t}{\mu_\psi}\left((1+\theta)\left\|g_i(\mathbf{w}_t) - g_i(\mathbf{w}_t;\zeta_{i,t})\right\|_2 + \theta\left\|g_i(\mathbf{w}_{t-1}) - g_i(\mathbf{w}_{t-1};\zeta_{i,t})\right\|_2\right)
\end{aligned}
$$

Hence

$$
\begin{aligned}
&\mathbb{E}_{\zeta_t}[\mathrm{I}_1] \le \frac{(1+\theta)\alpha_t}{n\mu_\psi}\cdot \\
&\sum_{i=1}^{n}\mathbb{E}\left[(1+1.5\theta)\left\|g_i(\mathbf{w}_t) - g_i(\mathbf{w}_t;\zeta_{i,t})\right\|_2^2 + 0.5\theta\left\|g_i(\mathbf{w}_{t-1}) - g_i(\mathbf{w}_{t-1};\zeta_{i,t})\right\|_2^2\right] \\
&\le \frac{(1+\theta)(1+2\theta)\sigma_0^2\alpha_t}{\mu_\psi}. \qquad (5.42)
\end{aligned}
$$

Next, let us handle $\mathrm{I}_2$. Let us define an auxiliary sequence $\{\tilde{\mathbf{y}}_t\}_{t\ge 1}$,

$$
\tilde{y}_{i,t+1} = \arg\min_{v\in\mathcal{Y}_i}\{(g_i(\mathbf{w}_t;\zeta_{i,t}) - g_i(\mathbf{w}_t))^\top v + \frac{1}{\lambda_2}D_{\psi_i}(v, \tilde{y}_{i,t})\},
$$

where $\lambda_2 > 0$. Lemma 3.13 implies that

$$
(g_i(\mathbf{w}_t) - g_i(\mathbf{w}_t;\zeta_{i,t}))^\top y_i \le \frac{1}{\lambda_2}\mathbb{E}[D_{\psi_i}(y_i, \tilde{y}_{i,t}) - D_{\psi_i}(y_i, \tilde{y}_{i,t+1})] + \frac{\lambda_2\sigma_0^2}{2\mu_\psi}.
$$

Averaging over $i = 1, \ldots, n$ and multiplying $(1+\theta)$ yields a bound of $\mathrm{I}_2$:

$$\mathbb{E}[\mathrm{I}_2] \leq \frac{1+\theta}{n\lambda_2}\mathbb{E}[D_\psi(\mathbf{y}, \tilde{\mathbf{y}}_t) - D_\psi(\mathbf{y}, \tilde{\mathbf{y}}_{t+1})] + \frac{(1+\theta)\lambda_2\sigma_0^2}{2\mu_\psi}.$$

As a result, the I term in (5.41) can be bounded as

$$\mathbb{E}[\mathrm{I}] \leq \frac{1+\theta}{n\lambda_2}\mathbb{E}[D_\psi(\mathbf{y}, \tilde{\mathbf{y}}_t) - D_\psi(\mathbf{y}, \tilde{\mathbf{y}}_{t+1})] + \frac{(1+\theta)\lambda_2\sigma_0^2}{2\mu_\psi} + \frac{(1+\theta)(1+2\theta)\sigma_0^2\alpha_t}{\mu_\psi}. \tag{5.43}$$

Similarly, the II term in (5.41) can be bounded as

$$\mathbb{E}[\mathrm{II}] \leq \frac{\theta}{n\lambda_5}\mathbb{E}[D_\psi(\mathbf{y}, \check{\mathbf{y}}_t) - D_\psi(\mathbf{y}, \check{\mathbf{y}}_{t+1})] + \frac{\theta\lambda_5\sigma_0^2}{2\mu_\psi} + \frac{\theta(0.5+1.5\theta)\sigma_0^2\alpha_t}{\mu_\psi}. \tag{5.44}$$

where

$$\check{y}_{i,t+1} = \arg\min_{v\in\mathcal{Y}_i}\{(g_i(\mathbf{w}_{t-1}) - g_i(\mathbf{w}_{t-1}; \zeta_{i,t}))^\top v + \lambda_5 D_{\psi_i}(v, \check{y}_{i,t})\}, \forall i.$$

Next, let us bound III. Recall $\Gamma_t \coloneqq \frac{1}{n}\sum_{i=1}^n (g_i(\mathbf{w}_t) - g_i(\mathbf{w}_{t-1}))^\top (y_i - y_{i,t})$. For any $\lambda_3, \lambda_4 > 0$, III can be rewritten as

$$\begin{aligned}
\mathrm{III} &= \Gamma_{t+1} - \theta\Gamma_t + \frac{1}{n}\sum_{i=1}^n (g_i(\mathbf{w}_{t+1}) - g_i(\mathbf{w}_t))^\top (y_{i,t+1} - \bar{y}_{i,t+1}) \\
&- \frac{\theta}{n}\sum_{i=1}^n (g_i(\mathbf{w}_t) - g_i(\mathbf{w}_{t-1}))^\top (y_{i,t} - \bar{y}_{i,t+1}) \\
&\leq \Gamma_{t+1} - \theta\Gamma_t + \frac{G_2^2\left\|\mathbf{w}_{t+1} - \mathbf{w}_t\right\|_2^2}{2\lambda_3} + \frac{\lambda_3\left\|\mathbf{y}_{t+1} - \bar{\mathbf{y}}_{t+1}\right\|_2^2}{2n} \\
&+ \frac{G_2^2\theta\left\|\mathbf{w}_t - \mathbf{w}_{t-1}\right\|_2^2}{2\lambda_4} + \frac{\lambda_4\theta\left\|\mathbf{y}_t - \bar{\mathbf{y}}_{t+1}\right\|_2^2}{2n}.
\end{aligned}$$

Note that $y_{i,t+1} = \bar{y}_{i,t+1}$ if $i \in \mathcal{B}_t$ and $y_{i,t+1} = y_{i,t}$ otherwise. Then, $\left\|\mathbf{y}_{t+1} - \bar{\mathbf{y}}_{t+1}\right\|_2^2 \leq \left\|\mathbf{y}_t - \bar{\mathbf{y}}_{t+1}\right\|_2^2$ such that

$$\begin{aligned}
\mathrm{III} \leq \Gamma_{t+1} - \theta\Gamma_t + \frac{G_2^2\left\|\mathbf{w}_{t+1} - \mathbf{w}_t\right\|_2^2}{2\lambda_3} + \frac{G_2^2\theta\left\|\mathbf{w}_t - \mathbf{w}_{t-1}\right\|_2^2}{2\lambda_4} \\
+ \frac{(\lambda_3 + \lambda_4\theta)D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t)}{\mu_\psi n}.
\end{aligned} \tag{5.45}$$

Combining (5.43), (5.45), (5.44), we have

$$
\begin{aligned}
&\mathbb{E}[B_t(\mathbf{y})] \le \mathbb{E}[\Gamma_{t+1} - \theta\Gamma_t] \\
&+ \frac{1+\theta}{n\lambda_2}\mathbb{E}[D_\psi(\mathbf{y}, \tilde{\mathbf{y}}_t) - D_\psi(\mathbf{y}, \tilde{\mathbf{y}}_{t+1})] + \frac{\theta}{n\lambda_5}\mathbb{E}[D_\psi(\mathbf{y}, \check{\mathbf{y}}_t) - D_\psi(\mathbf{y}, \check{\mathbf{y}}_{t+1})] \\
&+ \frac{(\lambda_3 + \lambda_4\theta)\mathbb{E}[D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t)]}{\mu_\psi n} + \frac{G_2^2\mathbb{E}\,\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2}{2\lambda_3} + \frac{G_2^2\theta\mathbb{E}\,\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2}{2\lambda_4} \\
&+ \frac{(1+\theta)\lambda_2\sigma_0^2}{2\mu_\psi} + \frac{\theta\sigma_0^2\lambda_5}{2\mu_\psi} + \frac{(1 + 3.5\theta + 3.5\theta^2)\sigma_0^2\alpha_t}{\mu_\psi}.
\end{aligned}
$$

□

**Lemma 5.18** *When $\theta = 0$, for any $\lambda_2, \lambda_4 \ge 0$ and $\mathbf{y}$ that possibly depends on all randomness in the algorithm, we have*

$$
\mathbb{E}[B_t(\mathbf{y})] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(g_i(\mathbf{w}_{t+1}) - \tilde{g}_t)^\top(y_i - \bar{y}_{i,t+1}) \le \frac{G_2^2\mathbb{E}\left[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2\right]}{4\lambda_4} + 4\lambda_4 G_1^2
$$

$$
+ \frac{1}{n\lambda_2}\mathbb{E}[D_\psi(\mathbf{y}, \tilde{\mathbf{y}}_t) - D_\psi(\mathbf{y}, \tilde{\mathbf{y}}_{t+1})] + \frac{\lambda_2\sigma_0^2}{2\mu_\psi} + \frac{\sigma_0^2\alpha_t}{\mu_\psi}. \tag{5.46}
$$

*Proof.* For ALEXR with $\theta = 0$, we have $\tilde{g}_{i,t} = g_i(\mathbf{w}_t; \zeta_{i,t})$. Then, for any $\lambda_4 > 0$ we have

$$
\frac{1}{n}\sum_{i=1}^n \mathbb{E}(g_i(\mathbf{w}_{t+1}) - \tilde{g}_t)^\top(y_i - \bar{y}_{i,t+1}) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[(g_i(\mathbf{w}_{t+1}) - g_i(\mathbf{w}_t))^\top(y_i - \bar{y}_{i,t+1})\right]
$$

$$
+ \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[(g_i(\mathbf{w}_t) - g_i(\mathbf{w}_t; \zeta_{i,t}))^\top(y_i - \bar{y}_{i,t+1})\right]. \tag{5.47}
$$

We bound the first term on the RHS by Young's inequality:

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[(g_i(\mathbf{w}_{t+1}) - g_i(\mathbf{w}_t))^\top(y_i - \bar{y}_{i,t+1})\right] \\
&\le \frac{1}{n}\sum_{i=1}^n \left(\frac{G_2^2\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2}{4\lambda_4} + \lambda_4\left\|y_i - \bar{y}_{i,t+1}\right\|_2^2\right) \le \frac{G_2^2\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2}{4\lambda_4} + 4G_1^2.
\end{aligned}
$$

The second term in (5.47) can be bounded similarly as (5.43) by:

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[(g_i(\mathbf{w}_t) - g_i(\mathbf{w}_t; \zeta_{i,t}))^\top(y_i - \bar{y}_{i,t+1})\right] \\
&\le \frac{1}{n\lambda_2}\mathbb{E}[D_\psi(\mathbf{y}, \tilde{\mathbf{y}}_t) - D_\psi(\mathbf{y}, \tilde{\mathbf{y}}_{t+1})] + \frac{\lambda_2\sigma_0^2}{2\mu_\psi} + \frac{\sigma_0^2\alpha_t}{\mu_\psi}.
\end{aligned}
$$

Combining the above inequalities together, we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(g_i(\mathbf{w}_{t+1})-\tilde{g}_t)^\top(y_i-\bar{y}_{i,t+1}) \le \frac{G_2^2\mathbb{E}\left[\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2\right]}{4\lambda_4}+4\lambda_4 G_1^2$$
$$+\frac{1}{n\lambda_2}\mathbb{E}[D_\psi(\mathbf{y},\tilde{\mathbf{y}}_t)-D_\psi(\mathbf{y},\tilde{\mathbf{y}}_{t+1})]+\frac{\lambda_2\sigma_0^2}{2\mu_\psi}+\frac{\sigma_0^2\alpha_t}{\mu_\psi}.$$

□

**Lemma 5.19** *If $g_i$ is $L_2$-smooth and $\eta \le \frac{1}{2G_1L_2}$, then*

$$\mathbb{E}[C_t(\mathbf{w}_*)] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(g_i(\mathbf{w}_{t+1})-g_i(\mathbf{w}_*))^\top\bar{y}_{i,t+1}-\mathbf{z}_t^\top(\mathbf{w}_{t+1}-\mathbf{w}_*)\right] \tag{5.48}$$
$$\le \eta\sigma^2+\frac{1}{4\eta}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2.$$

*If $g_i$ is $G_2$-Lipschitz continuous, then*

$$\mathbb{E}[C_t(\mathbf{w}_*)] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(g_i(\mathbf{w}_{t+1})-g_i(\mathbf{w}_*))^\top\bar{y}_{i,t+1}-\mathbf{z}_t^\top(\mathbf{w}_{t+1}-\mathbf{w}_*)\right] \tag{5.49}$$
$$\le \eta(\sigma^2+4G_1^2G_2^2)+\frac{1}{4\eta}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2.$$

*where $\sigma^2 = \frac{G_1^2\sigma_2^2}{B}+\frac{G_1^2G_2^2(n-B)}{B(n-1)}$.*

*Proof.* We define $\Delta_t \coloneqq \frac{1}{B}\sum_{i\in\mathcal{B}_t}[\partial g_i(\mathbf{w}_t;\zeta'_{i,t})]^\top y_{i,t+1}-\frac{1}{n}\sum_{i=1}^{n}[\partial g_i(\mathbf{w}_t)]^\top\bar{y}_{i,t+1}$. Similar to Lemma 5.2, we have $\mathbb{E}_t[\|\Delta_t\|_2^2]\le\sigma^2$. To proceed, we have

$$\frac{1}{n}\sum_{i=1}^{n}(g_i(\mathbf{w}_{t+1})-g_i(\mathbf{w}_*))^\top\bar{y}_{i,t+1}-\mathbf{z}_t^\top(\mathbf{w}_{t+1}-\mathbf{w}_*)$$
$$=\frac{1}{n}\sum_{i=1}^{n}(g_i(\mathbf{w}_{t+1})-g_i(\mathbf{w}_t))^\top\bar{y}_{i,t+1}+\frac{1}{n}\sum_{i=1}^{n}(g_i(\mathbf{w}_t)-g_i(\mathbf{w}_*))^\top\bar{y}_{i,t+1}$$
$$+\frac{1}{n}\sum_{i=1}^{n}([\partial g_i(\mathbf{w}_t)]^\top\bar{y}_{i,t+1}+\Delta_t)^\top(\mathbf{w}_*-\mathbf{w}_{t+1}).$$

Since $g_i$ is convex and $\mathcal{Y}_i\subset\mathbb{R}_+^n$ as $\partial f_i\ge 0$, we have

$$\frac{1}{n}\sum_{i=1}^{n}(g_i(\mathbf{w}_t)-g_i(\mathbf{w}_*))^\top\bar{y}_{i,t+1}\le\frac{1}{n}\sum_{i=1}^{n}[\nabla g_i(\mathbf{w}_t)]^\top(\mathbf{w}_t-\mathbf{w}_*)^\top\bar{y}_{i,t+1}.$$

Adding the above two inequalities, we have

$$\frac{1}{n}\sum_{i=1}^{n}(g_i(\mathbf{w}_{t+1}) - g_i(\mathbf{w}_*))^\top \bar{y}_{i,t+1} - \mathbf{z}_t^\top(\mathbf{w}_{t+1} - \mathbf{w}_*) \tag{5.50}$$
$$\leq \frac{1}{n}\sum_{i=1}^{n}(g_i(\mathbf{w}_{t+1}) - g_i(\mathbf{w}_t) - \nabla g_i(\mathbf{w}_t)^\top(\mathbf{w}_{t+1} - \mathbf{w}_t))^\top \bar{y}_{i,t+1} + \frac{1}{n}\sum_{i=1}^{n}\Delta_t^\top(\mathbf{w}_* - \mathbf{w}_{t+1}).$$

If $g_i$ is $L_2$-smooth, the first term in (5.50) can be bounded by

$$\frac{1}{n}\sum_{i=1}^{n}(g_i(\mathbf{w}_{t+1}) - g_i(\mathbf{w}_t) - \nabla g_i(\mathbf{w}_t)^\top(\mathbf{w}_{t+1} - \mathbf{w}_t))^\top \bar{y}_{i,t+1}$$
$$\leq \frac{G_1}{n}\sum_{i=1}^{n}\left\|g_i(\mathbf{w}_{t+1}) - g_i(\mathbf{w}_t) - \nabla g_i(\mathbf{w}_t)^\top(\mathbf{w}_{t+1} - \mathbf{w}_t)\right\|_2 \leq \frac{G_1 L_2}{2}\left\|\mathbf{w}_{t+1} - \mathbf{w}_t\right\|_2^2. \tag{5.51}$$

To bound the second term in (5.50), we note that $\mathbb{E}_{\mathcal{B}_t,\tilde{\zeta}_t}[\Delta_t] = 0$. Let us define $\hat{\mathbf{w}}_{t+1} = \arg\min_{\mathbf{w}} \mathbf{w}^\top \frac{1}{n}\sum_{i=1}^{n}[\nabla g_i(\mathbf{w}_t)]^\top \bar{y}_{i,t+1} + \frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}_t\|_2^2 + r(\mathbf{w})$. Then we have

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\Delta_t^\top(\mathbf{w}_* - \mathbf{w}_{t+1})\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\Delta_t^\top(\mathbf{w}_* - \hat{\mathbf{w}}_{t+1} + \hat{\mathbf{w}}_{t+1} - \mathbf{w}_{t+1})\right]$$
$$= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\Delta_t^\top(\hat{\mathbf{w}}_{t+1} - \mathbf{w}_{t+1})\right],$$

where we use the fact that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\Delta_t^\top(\mathbf{w}_* - \hat{\mathbf{w}}_{t+1})\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{B}_t,\zeta_t'}[\Delta_t]^\top(\mathbf{w}_* - \hat{\mathbf{w}}_{t+1})\right] = 0.$$

According to Lemma 1.7 we have

$$\mathbb{E}[\Delta_t^\top(\hat{\mathbf{w}}_{t+1} - \mathbf{w}_{t+1})] \leq \frac{\eta}{1+\mu\eta}\mathbb{E}\left\|\Delta_t\right\|_2^2 \leq \frac{\eta\sigma^2}{1+\mu\eta}. \tag{5.52}$$

Then, combining (5.50), (5.51) and (5.52) leads to

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(g_i(\mathbf{w}_{t+1}) - g_i(\mathbf{w}_*))^\top \bar{y}_{i,t+1}] - \mathbb{E}\mathbf{z}_t^\top(\mathbf{w}_{t+1} - \mathbf{w}_*)$$
$$\leq \frac{\eta\sigma^2}{1+\mu\eta} + \frac{L_2 G_1}{2}\left\|\mathbf{w}_{t+1} - \mathbf{w}_t\right\|_2^2,$$

which finishes the first part by noting the condition on $\eta$.

If $g_i$ is $G_2$-Lipschitz continuous, we have

$$\frac{G_1}{n}\sum_{i=1}^{n}\|g_i(\mathbf{w}_{t+1})-g_i(\mathbf{w}_t)-\partial g_i(\mathbf{w}_t)(\mathbf{w}_{t+1}-\mathbf{w}_t)\|_2 \tag{5.53}$$
$$\leq 2G_1G_2\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2 \leq \eta 4G_1^2G_2^2+\frac{1}{4\eta}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2 .$$

Combining (5.50), (5.52), and (5.53), we get

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[g_i(\mathbf{w}_{t+1})-g_i(\mathbf{w}_*)^\top\bar{y}_{i,t+1}]-\mathbb{E}\mathbf{z}_t^\top(\mathbf{w}_{t+1}-\mathbf{w}_*)$$
$$\leq \eta(\sigma^2+4G_1^2G_2^2)+\frac{1}{4\eta}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2 .$$

□

### 5.4.3 Strongly convex objectives

In this section, we derive a complexity of $O(1/\epsilon)$ under the the following condition.

**Assumption 5.11.** *We assume that the function $r$ is $\mu$-strongly convex ($\mu > 0$) and each $f_i$ is $L_1$-smooth, both with respect to the Euclidean norm $\|\cdot\|_2$.*

With this assumption, the minimax problem becomes strongly convex and strongly concave since the dual $f_i^*$ is $1/L_1$-strongly convex with respect to $\|\cdot\|_2$. In this case, we will establish the convergence of $\mu\|\mathbf{w}-\mathbf{w}_*\|_2^2$.

**Critical:** Under Assumption 5.11, parts (i) and (ii) of Assumption 5.9 hold for both variants of ALEXR. For ALEXR-v1, we have $\mu_\psi = 1/L_1$ and $\rho = 1$, whereas for ALEXR-v2, we have $\mu_\psi = 1$ and $\rho = 1/L_1$. Hence, the following theorem holds for both variants of ALEXR.

Let us introduce a few notations:

$$a=\frac{\epsilon\mu_\psi\rho}{24\sigma_0^2},\quad b_1=3(\sigma^2+4G_1^2G_2^2),\quad b_2=3\sigma^2.$$

**Theorem 5.7** *Suppose Assumptions 5.8, 5.10 and 5.11 hold.*

- *If $g_i$ is $G_2$-Lipschitz continuous, by setting $\alpha=\frac{1-\theta}{\rho(\theta-(1-B/n))}$, $\eta=\frac{1-\theta}{\theta\mu}$ and*

$$\theta=\max\left\{1-\frac{a\frac{B}{n}}{1+a},\ 1-\frac{\mu\epsilon}{b_1+\mu\epsilon}\right\}.$$

*ALEXR finds a solution $\mathbf{w}_{T+1}$ such that $\mathbb{E}[\mu\|\mathbf{w}_{T+1}-\mathbf{w}_*\|_2^2]\leq\epsilon$ with an iteration complexity of*

$$T = O\left(\frac{1}{1-\theta}\log(3\Upsilon/\epsilon)\right) = \tilde{O}\left(\max\left(\frac{n}{B}, \frac{(\sigma^2 + G_1^2 G_2^2)}{\mu\epsilon}, \frac{n\sigma_0^2}{B\epsilon\mu_\psi\rho}\right)\right).$$

- *If $g_i$ is further $L_2$-smooth, by setting $\alpha = \frac{1-\theta}{\rho(\theta-(1-B/n))}$, $\eta = \frac{1-\theta}{\theta\mu}$ and*

$$\theta = \max\left\{1 - \frac{a\frac{B}{n}}{1+a}, 1 - \frac{\mu\epsilon}{b_2+\mu\epsilon}, 1 - \frac{\mu}{2G_1L_2+\mu}\right\},$$

*for sufficiently small $\epsilon$, ALEXR finds a solution $\mathbf{w}_{T+1}$ such that $\mathbb{E}[\mu\|\mathbf{w}_{T+1} - \mathbf{w}_*\|_2^2] \leq \epsilon$ with an iteration complexity of*

$$T = O\left(\frac{1}{1-\theta}\log(3\Upsilon/\epsilon)\right) = \tilde{O}\left(\max\left(\frac{G_1L_2}{\mu}, \frac{n}{B}, \frac{\sigma^2}{\mu\epsilon}, \frac{n\sigma_0^2}{B\epsilon\mu_\psi\rho}\right)\right).$$

*where $\Upsilon = \frac{\mu}{2}\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2 + \frac{2\rho}{B}D_\psi(\mathbf{y}_*, \mathbf{y}_1)$ and $\sigma^2 = \frac{G_1^2\sigma_2^2}{B} + \frac{G_1^2G_2^2(n-B)}{B(n-1)}$.*

**Why it matters**

For smooth functions $g_i$, the iteration complexity is improved in the sense that the $O(1/\epsilon)$ dependence is scaled by the variance of the stochastic estimators. In contrast, for non-smooth $g_i$, the complexity always has a term $\frac{G_1^2G_2^2}{\mu\epsilon}$ independent of variance.

*Proof.* We first consider non-smooth $g_i$. Combining (5.32), (5.36) for $A_t(\mathbf{y}_*)$, (5.40) for $B_t(\mathbf{y}_*)$, (5.49) for $C_t(\mathbf{w}_*)$ together we have

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_{t+1}, \mathbf{y}_*) - F(\mathbf{w}_*, \bar{\mathbf{y}}_{t+1})] \\
&\leq \frac{1}{2\eta_t}\|\mathbf{w}_* - \mathbf{w}_t\|_2^2 - (\frac{1}{2\eta_t} + \frac{\mu}{2})\|\mathbf{w}_* - \mathbf{w}_{t+1}\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&+ \mathbb{E}\left[\frac{\tau_t + \rho\left(1 - \frac{B}{n}\right)}{B}D_\psi(\mathbf{y}_*, \mathbf{y}_t) - \frac{\tau_t + \rho}{B}D_\psi(\mathbf{y}_*, \mathbf{y}_{t+1})\right] - \frac{\tau_t}{n}\mathbb{E}\left[D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t)\right] \\
&+ \mathbb{E}[\Gamma_{t+1}^* - \theta\Gamma_t^*] + \frac{(\lambda_3 + \lambda_4\theta)\mathbb{E}[D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t)]}{\mu_\psi n} \\
&+ \frac{G_2^2\mathbb{E}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2}{2\lambda_3} + \frac{G_2^2\theta\mathbb{E}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2}{2\lambda_4} + \frac{(1 + 3.5\theta + 3.5\theta^2)\sigma_0^2\alpha_t}{\mu_\psi} \\
&+ \eta_t(\sigma^2 + 4G_1^2G_2^2) + \frac{1}{4\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2.
\end{aligned}$$

Define $\Upsilon_{1,t} \coloneqq \frac{1}{2}\|\mathbf{w}_* - \mathbf{w}_t\|_2^2$ and $\Upsilon_{2,t} = \frac{1}{B}D_\psi(\mathbf{y}_*, \mathbf{y}_t)$. Since

$$F(\mathbf{w}_{t+1}, \mathbf{y}_*) - F(\mathbf{w}_*, \bar{\mathbf{y}}_{t+1}) \geq F(\mathbf{w}_{t+1}, \mathbf{y}_*) - F(\mathbf{w}_*, \mathbf{y}_*) + F(\mathbf{w}_*, \mathbf{y}_*) - F(\mathbf{w}_*, \bar{\mathbf{y}}_{t+1}) \geq 0,$$

multiplying the above inequality by $\theta^{-t}$ on both sides, we have

$$
\begin{aligned}
0 \leq\ & \theta^{-t}\mathbb{E}\left[\frac{1}{\eta_t}\Upsilon_{1,t} + (\tau_t + \rho(1-\frac{B}{n})))\Upsilon_{2,t} - \theta\Gamma_t^*\right] \\
& - \theta^{-t}\mathbb{E}\left[(\frac{1}{\eta_t}+\mu)\Upsilon_{1,t+1} + (\tau_t+\rho)\Upsilon_{2,t+1} - \Gamma_{t+1}^*\right] \\
& - \theta^{-t}(\frac{1}{2\eta_t}\mathbb{E}\left\|\mathbf{w}_{t+1}-\mathbf{w}_t\right\|_2^2 + \frac{\tau_t}{n}\mathbb{E}\left[D_\psi(\bar{\mathbf{y}}_{t+1},\mathbf{y}_t)\right]) + \theta^{-t}\frac{(\lambda_3+\lambda_4\theta)\mathbb{E}[D_\psi(\bar{\mathbf{y}}_{t+1},\mathbf{y}_t)]}{\mu_\psi n} \\
& + \theta^{-t}\frac{G_2^2\mathbb{E}\left\|\mathbf{w}_{t+1}-\mathbf{w}_t\right\|_2^2}{2\lambda_3} + \theta^{-t}\frac{G_2^2\theta\mathbb{E}\left\|\mathbf{w}_t-\mathbf{w}_{t-1}\right\|_2^2}{2\lambda_4} + \theta^{-t}\frac{1}{4\eta_t}\mathbb{E}\left\|\mathbf{w}_{t+1}-\mathbf{w}_t\right\|_2^2 \\
& + \theta^{-t}\left(\frac{(1+3.5\theta+3.5\theta^2)\sigma_0^2\alpha_t}{\mu_\psi} + \eta_t(\sigma^2+4G_1^2G_2^2)\right).
\end{aligned} \tag{5.54}
$$

Let

$$
\frac{1}{\eta_{t-1}} + \mu = \frac{1}{\eta_t\theta}, \quad (\tau_{t-1}+\rho) = \frac{1}{\theta}(\tau_t + \rho(1-\frac{B}{n})). \tag{5.55}
$$

Hence,

$$
\begin{aligned}
& \sum_{t=1}^{T}\left\{\theta^{-t}\left[\frac{1}{\eta_t}\Upsilon_{1,t} + (\tau_t+\rho(1-\frac{B}{n})))\Upsilon_{2,t} - \theta\Gamma_t^*\right]\right. \\
& \left. -\theta^{-t}\left[(\frac{1}{\eta_t}+\mu)\Upsilon_{1,t+1} + (\tau_t+\rho)\Upsilon_{2,t+1} - \Gamma_{t+1}^*\right]\right\} \\
& \leq \sum_{t=1}^{T}\left\{\theta^{-(t-1)}\left[(\frac{1}{\eta_{t-1}}+\mu)\Upsilon_{1,t} + (\tau_{t-1}+\rho)\Upsilon_{2,t} - \Gamma_t^*\right]\right. \\
& \left. -\theta^{-t}\left[(\frac{1}{\eta_t}+\mu)\Upsilon_{1,t+1} + (\tau_t+\rho)\Upsilon_{2,t+1} - \Gamma_{t+1}^*\right]\right\} \\
& = \left[(\frac{1}{\eta_0}+\mu)\Upsilon_{1,1} + (\tau_0+\rho)\Upsilon_{2,1} - \Gamma_1\right] \\
& - \theta^{-T}\left[(\frac{1}{\eta_T}+\mu)\Upsilon_{1,T+1} + (\tau_T+\rho)\Upsilon_{2,T+1} - \Gamma_{T+1}\right].
\end{aligned}
$$

Since

$$
\begin{aligned}
& -\Gamma_{T+1} \geq -\frac{1}{n}\sum_{i=1}^{n} G_2\|\mathbf{w}_{T+1}-\mathbf{w}_T\|_2\|y_{i,*}-y_{i,T+1}\|_2 \\
& \geq -\frac{1}{n}\sum_{i=1}^{n}(\frac{G_2^2 B}{n(\rho+\tau_T)\mu_\psi}\|\mathbf{w}_{T+1}-\mathbf{w}_T\|_2^2 + \frac{n\mu_\psi(\rho+\tau_T)}{4B}\|y_{i,*}-y_{i,T+1}\|_2^2) \\
& \geq -(\frac{G_2^2 B}{2n(\rho+\tau_T)\mu_\psi}\|\mathbf{w}_{T+1}-\mathbf{w}_T\|_2^2 + \frac{\rho+\tau_T}{2}\Upsilon_{2,T+1}).
\end{aligned} \tag{5.56}
$$

Summing (5.54) over $t = 1, \ldots, T$ and utilizing the above two inequalities, we have

$$\begin{aligned}
&\theta^{-T}\mathbb{E}\left[(\frac{1}{\eta_T}+\mu)\Upsilon_{1,T+1}+\frac{\rho+\tau_T}{2}\Upsilon_{2,T+1}\right]\\
&\leq\left[(\frac{1}{\eta_0}+\mu)\Upsilon_{1,1}+(\tau_0+\rho)\Upsilon_{2,1}-\Gamma_1\right]+\\
&\frac{\theta^{-T}G_2^2B}{2n(\rho+\tau_T)\mu_\psi}\mathbb{E}\|\mathbf{w}_{T+1}-\mathbf{w}_T\|_2^2-\mathbb{E}\left[\sum_{t=1}^{T}\frac{\theta^{-t}}{2}(\frac{1}{\eta_t}-\frac{G_2^2}{\lambda_3}-\frac{G_2^2}{\lambda_4}-\frac{1}{2\eta_t})\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2\right]\\
&-\mathbb{E}\left[\sum_{t=1}^{T}\frac{\theta^{-t}}{n}(\frac{1}{\alpha_t}-\frac{\lambda_3+\lambda_4\theta}{\mu_\psi})D_\psi(\bar{\mathbf{y}}_{t+1},\mathbf{y}_t)\right]\\
&+\sum_{t=1}^{T}\theta^{-t}\left(\frac{(1+3.5\theta+3.5\theta^2)\sigma_0^2\alpha_t}{\mu_\psi}+\eta_t(\sigma^2+4G_1^2G_2^2)\right),
\end{aligned}$$

where we use the fact $\sum_{t=1}^{T}\|\mathbf{w}_t-\mathbf{w}_{t-1}\|_2^2\leq\sum_{t=1}^{T+1}\|\mathbf{w}_t-\mathbf{w}_{t-1}\|_2^2=\sum_{t=1}^{T}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2$.

Let $\eta_t=\eta,\alpha_t=\frac{1}{\tau_t}=\alpha,\lambda_3=\lambda_4=8\eta G_2^2$. If $\alpha\leq\frac{\mu_\psi}{16\eta G_2^2}$ (to be verfied later), we have $\tau\geq\frac{4\eta G_2^2B}{n\mu_\psi}$. As a result, $\frac{G_2^2B}{2n(\rho+\tau_t)\mu_\psi}\leq\frac{1}{8\eta}$ and $\frac{1}{\alpha_t}\geq\frac{16\eta G_2^2}{\mu_\psi}$. Then the terms related to $\|\mathbf{w}_{t+1}-\mathbf{w}_t\|$ and $D_\psi(\bar{\mathbf{y}}_{t+1},\mathbf{y}_t)$ is less than zero. As a result,

$$\begin{aligned}
&\left[(\frac{1}{\eta}+\mu)\Upsilon_{1,T+1}+(\frac{\rho}{2}+\frac{1}{2\alpha})\Upsilon_{2,T+1}\right]\\
&\leq\theta^T\left[(\frac{1}{\eta}+\mu)\Upsilon_{1,1}+(\frac{1}{\alpha}+\rho)\Upsilon_{2,1}\right]+\sum_{t=1}^{T}\theta^{T-t}\left(\frac{8\sigma_0^2\alpha}{\mu_\psi}+\eta(\sigma^2+4G_1^2G_2^2)\right)\\
&\leq\theta^T\left[(\frac{1}{\eta}+\mu)\Upsilon_{1,1}+(\frac{1}{\alpha}+\rho)\Upsilon_{2,1}\right]+\frac{1}{1-\theta}\left(\frac{8\sigma_0^2\alpha}{\mu_\psi}+\eta(\sigma^2+4G_1^2G_2^2)\right).
\end{aligned}$$

Due to the relationship between $\eta,\alpha$ and $\theta$ in (5.55), we have

$$\begin{aligned}
\theta&=\frac{1}{1+\mu\eta}=\frac{1+\alpha\rho(1-B/n)}{1+\alpha\rho}\geq\frac{1}{1+\alpha\rho}\\
\alpha&=\frac{1-\theta}{\rho(\theta-(1-B/n))},\quad\eta=\frac{1-\theta}{\theta\mu}.
\end{aligned}$$

Then, we have

$$
\begin{aligned}
\left[\mu\Upsilon_{1,T+1}\right] &= \frac{\mu\eta}{1+\eta\mu}\left(\frac{1}{\eta}+\mu\right)\Upsilon_{1,T+1} \\
&\le \theta^T\mu\left[\Upsilon_{1,1}+\frac{(1+\alpha\rho)\eta}{\alpha(1+\eta\mu)}\Upsilon_{2,1}\right]+\frac{1}{1-\theta}\frac{\eta\mu}{1+\eta\mu}\left(\frac{8\sigma_0^2\alpha}{\mu_\psi}+\eta(\sigma^2+4G_1^2G_2^2)\right) \\
&= \theta^T\Upsilon+\frac{8\sigma_0^2\alpha}{\mu_\psi}+\eta(\sigma^2+4G_1^2G_2^2) \\
&\le \theta^T\Upsilon+\frac{1-\theta}{\rho(\theta-(1-B/n))}\frac{8\sigma_0^2}{\mu_\psi}+\frac{1-\theta}{\theta\mu}(\sigma^2+4G_1^2G_2^2),
\end{aligned}
$$

where $\Upsilon = \mu\Upsilon_{1,1}+\mu\frac{(1+\alpha\rho)\eta}{\alpha(1+\eta\mu)}\Upsilon_{2,1}$.

To let the RHS be less than $\epsilon$, it is sufficient to have

$$
\begin{aligned}
T &\ge \frac{1}{1-\theta}\log(3\Upsilon/\epsilon) \ge \frac{-1}{\log(\theta)}\log(3\Upsilon/\epsilon) \Rightarrow \theta^T\Upsilon \le \epsilon/3, \\
\theta &\ge 1-\frac{\epsilon\mu_\psi\rho B/(24\sigma_0^2 n)}{1+\epsilon\mu_\psi\rho/(24\sigma_0^2)} \Rightarrow \frac{1-\theta}{\rho(\theta-(1-B/n))}\frac{8\sigma_0^2}{\mu_\psi} \le \epsilon/3, \\
\theta &\ge \frac{1}{1+\mu\epsilon/(3(\sigma^2+4G_1^2G_2^2))} \Rightarrow \frac{1-\theta}{\theta\mu}(\sigma^2+4G_1^2G_2^2) \le \frac{\epsilon}{3}.
\end{aligned}
$$

As a result,

$$
T = O\left(\frac{1}{1-\theta}\log(3\Upsilon/\epsilon)\right) = \tilde{O}\left(\max\left(\frac{(\sigma^2+G_1^2G_2^2)}{\mu\epsilon}, \frac{n}{B}, \frac{n\sigma_0^2}{B\epsilon\mu_\psi\rho}\right)\right).
$$

Finally, we verify that if $\epsilon^2 \le \frac{9(\sigma^2+4G_1^2G_2^2)\sigma_0^2}{2G_2^2}$, then it holds that

$$
\alpha \le \frac{\mu_\psi\epsilon}{24\sigma_0^2} = \frac{\mu_\psi\epsilon^2}{24\sigma_0^2\epsilon} \le \frac{\mu_\psi 3(\sigma^2+4G_1^2G_2^2)}{16G_2^2\epsilon} \le \frac{\mu_\psi\theta\mu}{16G_2^2(1-\theta)} = \frac{\mu_\psi}{16\eta G_2^2}.
$$

Since $\alpha\rho \le O(1)$, we have

$$
\frac{(1+\alpha\rho)\eta}{(1+\mu\eta)\alpha} \le 2\frac{\eta}{\alpha} \le 2\frac{\rho}{\mu},
$$

thus $\Upsilon_{1,1}+\frac{(1+\alpha\rho)\eta}{\alpha(1+\mu\eta)}\Upsilon_{2,1} \le \Upsilon_{1,1}+\frac{2\rho}{\mu}\Upsilon_{2,1}$. Thus, $\Upsilon \le \mu\Upsilon_{1,1}+2\rho\Upsilon_{2,1}$.

For smooth $g_i$, the proof is similar by using (5.48) instead of using (5.49). Hence, $\eta_t(\sigma^2+4G_1^2G_2^2)$ becomes $\eta_t(\sigma^2)$ and there is additional condition $\eta_t \le \frac{1}{2G_1L_2}$, which transfers to a condition on $\theta$. □

### 5.4.4 Convex objectives with non-smooth outer functions

In this section, we only consider ALEXR-v2 for solving convex objectives with non-smooth $f_i$. For ALEXR-v2, we have that $\psi$ is 1-smooth and 1-strongly convex. Hence, we have

$$\frac{(n-B)(\tau+\rho)}{2\mu_\psi\lambda_0 nB}\mathbb{E}\left[\sum_{i=1}^n \left\|\nabla\psi_i(\bar{y}_{i,t+1}) - \nabla\psi_i(y_{i,t})\right\|_2^2\right] \tag{5.57}$$

$$\leq \frac{(n-B)(\tau+\rho)}{2\lambda_0 nB}\mathbb{E}\left[\sum_{i=1}^n \left\|\bar{y}_{i,t+1} - y_{i,t}\right\|_2^2\right] \leq \frac{(n-B)(\tau+\rho)}{\lambda_0 nB}\mathbb{E}\left[D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t)\right].$$

**Theorem 5.8** *Suppose Assumption 5.9 holds with $\rho = 0, \mu_\psi = 1$, and Assumptions 5.8, 5.10 hold. If $g_i$ is $G_2$-Lipschitz continuous, setting $\theta = 0$ and*

$$\alpha = \frac{\epsilon}{6\sigma_0^2}, \quad \eta = \frac{\epsilon}{6(\sigma^2 + 8G_1^2G_2^2)},$$

*ALEXR-v2 returns an $\epsilon$-optimal solution $\bar{\mathbf{w}}_T = \sum_{t=1}^T \mathbf{w}_t/T$ in expectation with a complexity of*

$$T = O\left(\frac{\sigma^2 + G_1^2G_2^2}{\epsilon^2}, \frac{\Omega\sigma_0^2}{B\epsilon^2}, \frac{\Omega\sigma_0^2}{n\epsilon^2}\right).$$

*where $\Omega$ is a constant such that $\mathbb{E}[D_\psi(\mathbf{y}_T^*, \mathbf{y}_1)] \leq \Omega \leq O(G_1^2 n)$, and $\mathbf{y}_T^* = \arg\max_{\mathbf{y}\in\mathcal{Y}_1\times\cdots\times\mathcal{Y}_n} F(\bar{\mathbf{w}}_T, \mathbf{y})$.*

**Why it matters**

In the worst case, the complexity is $O\left(\frac{G_1^2G_2^2}{\epsilon^2} + \frac{nG_1^2\sigma_0^2}{B\epsilon^2}\right)$. This will match the lower bounds established in next section.

*Proof.* Combining (5.35) with (5.57) yields

$$\mathbb{E}\left[\sum_{t=1}^T A_t(\mathbf{y})\right] \leq \frac{\tau}{B}\mathbb{E}[D_\psi(\mathbf{y}, \mathbf{y}_1)] - \frac{\tau}{n}\mathbb{E}\left[\sum_{t=1}^T D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t)\right] + \frac{\lambda_0\tau}{n}\mathbb{E}D_\psi(\mathbf{y}, \hat{\mathbf{y}}_1) \tag{5.58}$$

$$+ \frac{(n-B)\tau}{\lambda_0 nB}\mathbb{E}\left[\sum_{t=1}^T D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t)\right] \tag{5.59}$$

Adding this inequality with (5.32), (5.46), and (5.49) over $t = 1, \ldots, T$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} F(\mathbf{w}_{t+1}, \mathbf{y}) - F(\mathbf{w}_*, \bar{\mathbf{y}}_{t+1})\right] \leq \frac{1}{2\eta}\|\mathbf{w}_* - \mathbf{w}_1\|_2^2 - \frac{1}{2\eta}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2]$$
$$+ \frac{\tau}{B}\mathbb{E}[D_\psi(\mathbf{y}, \mathbf{y}_1)] - \frac{\tau}{n}\mathbb{E}\left[\sum_{t=1}^{T} D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t)\right] + \frac{\lambda_0\tau}{n}\mathbb{E}D_\psi(\mathbf{y}, \hat{\mathbf{y}}_1)$$
$$+ \frac{(n-B)\tau}{\lambda_0 nB}\mathbb{E}\left[\sum_{t=1}^{T} D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t)\right],$$
$$+ \frac{G_2^2}{4\lambda_4}\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2\right] + 4\lambda_4 T G_1^2 + \frac{1}{n\lambda_2}\mathbb{E}[D_\psi(\mathbf{y}, \tilde{\mathbf{y}}_1)]$$
$$+ \frac{\lambda_2\sigma_0^2}{2}T + \sigma_0^2\alpha T,$$
$$+ \eta T(\sigma^2 + 4G_1^2G_2^2) + \frac{1}{4\eta}\sum_{t=1}^{T}\mathbb{E}\,\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2\,.$$

If we set $\lambda_0 = \frac{n-B}{B}$ and $\frac{G_2^2}{4\lambda_4} = \frac{1}{4\eta}$, we observe that the terms involving $\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2\right]$ and $\mathbb{E}\left[\sum_{t=1}^{T} D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t)\right]$ cancel out, leaving us with the following:

$$\mathbb{E}\left[\sum_{t=1}^{T} F(\mathbf{w}_{t+1}, \mathbf{y}) - F(\mathbf{w}_*, \bar{\mathbf{y}}_{t+1})\right]$$
$$\leq \frac{1}{2\eta}\|\mathbf{w}_* - \mathbf{w}_1\|_2^2 + \left(\frac{\tau(1+\lambda_0 B/n)}{B} + \frac{1}{n\lambda_2}\right)\mathbb{E}D_\psi(\mathbf{y}, \mathbf{y}_1)$$
$$+ \eta T(\sigma^2 + 8G_1^2G_2^2) + \frac{\lambda_2\sigma_0^2}{2}T + \sigma_0^2\alpha T.$$

Let $\mathbf{y} = \mathbf{y}_T^* = \arg\max F(\bar{\mathbf{w}}_T, \mathbf{y})$. Since $\frac{1}{T}\sum_{t=1}^{T} F(\mathbf{w}_{t+1}, \mathbf{y}) \geq F(\bar{\mathbf{w}}_T, \mathbf{y}_T^*) = F(\bar{\mathbf{w}}_T)$ and $F(\mathbf{w}_*, \bar{\mathbf{y}}_{t+1}) \leq F(\mathbf{w}_*, \mathbf{y}_*)$, we have

$$\mathbb{E}\left[F(\bar{\mathbf{w}}_T) - F(\mathbf{w}_*)\right] \leq \frac{1}{2\eta T}\|\mathbf{w}_* - \mathbf{w}_1\|_2^2 + \frac{1}{T}\left(\frac{\tau(1+\lambda_0 B/n)}{B} + \frac{1}{n\lambda_2}\right)\Omega$$
$$+ \eta(\sigma^2 + 8G_1^2G_2^2) + \frac{\lambda_2\sigma_0^2}{2} + \sigma_0^2\alpha. \tag{5.60}$$

Let

$$\alpha = \frac{\epsilon}{6\sigma_0^2}, \quad \lambda_2 = \frac{\epsilon}{3\sigma_0^2}, \quad \eta = \frac{\epsilon}{6(\sigma^2 + 8G_1^2G_2^2)},$$
$$T \geq O\left(\max\left(\frac{\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2}{12\eta\epsilon}, \frac{\Omega(1+\lambda_0 B/n)}{6B\epsilon\alpha}, \frac{\Omega}{6n\lambda_2\epsilon}\right)\right).$$

Then, the RHS of (5.60) is less than $\epsilon$. As a result, the complexity is in the order of

$$O\left(\max\left(\frac{\sigma^2+G_1^2G_2^2}{\epsilon^2},\frac{\Omega\sigma_0^2}{B\epsilon^2},\frac{\Omega\sigma_0^2}{n\epsilon^2}\right)\right).$$

□

**Theorem 5.9** *Suppose Assumption 5.9 holds with $\rho=0,\mu_\psi=1$, Assumptions 5.8, 5.10 hold. If $g_i$ is $G_2$-Lipschitz continuous and $L_2$-smooth, for sufficiently small $\epsilon$, setting $\theta=1$ and*

$$\alpha=\frac{\epsilon}{64\sigma_0^2},\quad \eta=\min\left(\frac{\epsilon}{8\sigma^2},\frac{1}{2G_1L_2}\right)$$

*ALEXR-v2 returns an $\epsilon$-optimal solution $\bar{\mathbf{w}}_T=\sum_{t=1}^T\mathbf{w}_t/T$ in expectation with a complexity of*

$$T=O\left(\frac{G_1L_2}{\epsilon},\frac{\sigma^2}{\epsilon^2},\frac{\Omega\sigma_0^2}{B\epsilon^2},\frac{\Omega\sigma_0^2}{n\epsilon^2}\right).$$

*where $\Omega$ and $\mathbf{y}_T^*$ are defined similarly as in last theorem.*

**Why it matters**

For smooth functions $g_i$, the iteration complexity is improved in the sense that the $O(1/\epsilon^2)$ dependence is scaled by the variance of the stochastic estimators. In contrast, for non-smooth $g_i$, the complexity always includes a term $\frac{G_1^2G_2^2}{\epsilon^2}$, regardless of the variance.

*Proof.* The proof is similar to that of previous theorem except that we use (5.39) instead of (5.46), and using (5.48) instead of using (5.49). Additionally, we use

$$\begin{aligned}\sum_{t=1}^T(\Gamma_{t+1}-\Gamma_t)&=\Gamma_{T+1}-\Gamma_1\le\frac{1}{n}\sum_{i=1}^n G_2\|\mathbf{w}_{T+1}-\mathbf{w}_T\|_2\|y_i-y_{i,T+1}\|_2 \qquad (5.61)\\ &\le\frac{G_2^2B}{2n\tau}\|\mathbf{w}_{T+1}-\mathbf{w}_T\|_2^2+\frac{\tau n/B}{n}D_\psi(\mathbf{y},\mathbf{y}_{T+1}).\end{aligned}$$

Combining this with (5.39), we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^T B_t(\mathbf{y})\right]\le\frac{G_2^2B}{2n\tau}\mathbb{E}[\|\mathbf{w}_{T+1}-\mathbf{w}_T\|_2^2]+\frac{\tau}{B}\mathbb{E}[D_\psi(\mathbf{y},\mathbf{y}_{T+1})] \qquad (5.62)\\ &+\frac{2}{n\lambda_2}\mathbb{E}[D_\psi(\mathbf{y},\tilde{\mathbf{y}}_1)]+\frac{(\lambda_3+\lambda_4)}{n}\sum_{t=1}^T\mathbb{E}\left[D_\psi(\bar{\mathbf{y}}_{t+1},\mathbf{y}_t)\right]+\frac{G_2^2}{2\lambda_3}\sum_{t=1}^T\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2]\\ &+\frac{G_2^2}{2\lambda_4}\sum_{t=1}^T\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_{t-1}\|_2^2]+8\sigma_0^2\alpha T+\lambda_2\sigma_0^2T+\frac{\sigma_0^2\lambda_5}{2}T+\frac{1}{n\lambda_5}\mathbb{E}[D_\psi(\mathbf{y},\check{\mathbf{y}}_1)].\end{aligned}$$

Summing the inequalities in (5.32), (5.58), (5.62), and (5.49) over $t=1,\ldots,T$, we have

$$\mathbb{E}\Big[\sum_{t=1}^{T} F(\mathbf{w}_{t+1}, \mathbf{y}) - F(\mathbf{w}_*, \bar{\mathbf{y}}_{t+1})\Big] \leq \frac{1}{2\eta}\|\mathbf{w}_* - \mathbf{w}_1\|_2^2 - \frac{1}{2\eta}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2]$$
$$+ \frac{\tau}{B}\mathbb{E}[D_\psi(\mathbf{y}, \mathbf{y}_1) - D_\psi(\mathbf{y}, \mathbf{y}_{T+1})] - \frac{\tau}{n}\sum_{t=1}^{T}\mathbb{E}\left[D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t)\right]$$
$$+ \frac{\lambda_0\tau}{n}\mathbb{E}D_\psi(\mathbf{y}, \hat{\mathbf{y}}_1) + \frac{(n-B)\tau}{\lambda_0 nB}\sum_{t=1}^{T}\mathbb{E}\left[D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t)\right],$$
$$+ \frac{G_2^2 B}{2n\tau}\mathbb{E}[\|\mathbf{w}_{T+1} - \mathbf{w}_T\|_2^2] + \frac{\tau}{B}\mathbb{E}[D_\psi(\mathbf{y}, \mathbf{y}_{T+1})]$$
$$+ \frac{2}{n\lambda_2}\mathbb{E}[D_\psi(\mathbf{y}, \tilde{\mathbf{y}}_1)] + \frac{(\lambda_3+\lambda_4)}{n}\sum_{t=1}^{T}\mathbb{E}\Big[D_\psi(\bar{\mathbf{y}}_{t+1}, \mathbf{y}_t)\Big] + \frac{G_2^2}{2\lambda_3}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2]$$
$$+ \frac{G_2^2}{2\lambda_4}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2] + 8\sigma_0^2\alpha T + \lambda_2\sigma_0^2 T + \frac{\sigma_0^2\lambda_5}{2}T + \frac{1}{n\lambda_5}\mathbb{E}[D_\psi(\mathbf{y}, \check{\mathbf{y}}_1)],$$
$$+ \frac{1}{4\eta}\sum_{t=1}^{T}\mathbb{E}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 + \eta T\sigma^2.$$

Similarly as before, if we let $\lambda_0 = \frac{2(n-B)}{B}$, $\frac{G_2^2}{2\lambda_3} = \frac{G_2^2}{2\lambda_4} = \frac{1}{16\eta}$, $\lambda_3 + \lambda_4 = 16\eta G_2^2 \leq \tau/2$, and $\frac{G_2^2 B}{2n\tau} \leq \frac{1}{8\eta}$, we observe that all the cumulated terms cancel out, leaving us the following:

$$\mathbb{E}\Big[\sum_{t=1}^{T} F(\mathbf{w}_{t+1}, \mathbf{y}) - F(\mathbf{w}_*, \bar{\mathbf{y}}_{t+1})\Big] \leq$$
$$\frac{1}{2\eta}\|\mathbf{w}_* - \mathbf{w}_1\|_2^2 + \left(\frac{\tau(1+\lambda_0 B/n)}{B} + \frac{2}{n\lambda_2} + \frac{1}{n\lambda_5}\right)\mathbb{E}D_\psi(\mathbf{y}, \mathbf{y}_1)$$
$$+ \eta T\sigma^2 + 8\sigma_0^2\alpha T + \lambda_2\sigma_0^2 T + \frac{\sigma_0^2\lambda_5}{2}T.$$

Let $\mathbf{y} = \mathbf{y}_T^* = \arg\max F(\bar{\mathbf{w}}_T, \mathbf{y})$. Since $\frac{1}{T}\sum_{t=1}^{T} F(\mathbf{w}_{t+1}, \mathbf{y}) \geq F(\bar{\mathbf{w}}_T, \mathbf{y}_T^*) = F(\bar{\mathbf{w}}_T)$ and $F(\mathbf{w}_*, \bar{\mathbf{y}}_{t+1}) \leq F(\mathbf{w}_*, \mathbf{y}_*)$, we have

$$\mathbb{E}\Big[F(\bar{\mathbf{w}}_T) - F(\mathbf{w}_*)\Big] \leq \frac{1}{2\eta T}\|\mathbf{w}_* - \mathbf{w}_1\|_2^2 + \frac{1}{T}\left(\frac{\tau(1+\lambda_0 B/n)}{B} + \frac{2}{n\lambda_2} + \frac{1}{n\lambda_5}\right)\Omega$$
$$+ \eta(\sigma^2) + 8\sigma_0^2\alpha + \lambda_2\sigma_0^2 + \frac{\sigma_0^2\lambda_5}{2}. \tag{5.63}$$

Let

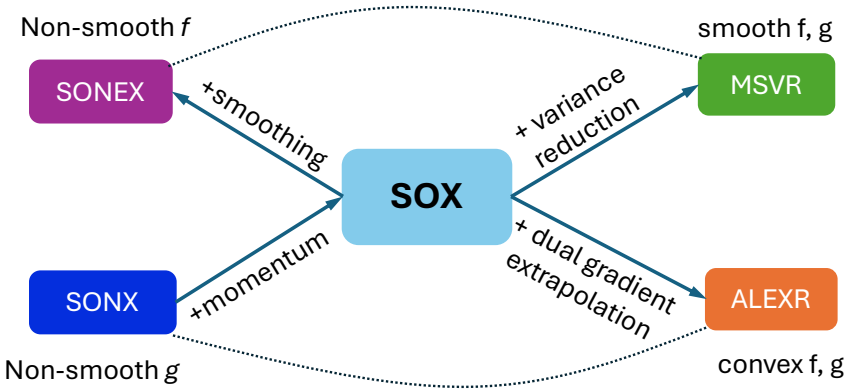


Fig. 5.1: Relationship between different algorithms for FCCO.

$$\alpha = \frac{\epsilon}{64\sigma_0^2}, \quad \lambda_2 = \frac{\epsilon}{8\sigma_0^2}, \quad \lambda_5 = \frac{\epsilon}{4\sigma_0^2}, \quad \eta = \min\left(\frac{\epsilon}{8\sigma^2}, \frac{1}{2G_1L_2}\right)$$
$$T \geq O\left(\max\left(\frac{\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2}{32\eta\epsilon}, \frac{\Omega(1+\lambda_0 B/n)}{8B\epsilon\alpha}, \frac{\Omega}{4n\lambda_2\epsilon}, \frac{\Omega}{8n\lambda_5\epsilon}\right)\right).$$

Then the conditions $16\eta G_2^2 \leq \tau/2$, $\frac{G_2^2 B}{2n\tau} \leq \frac{1}{8\eta}$ hold for sufficiently small $\epsilon$, and the RHS of (5.63) is less than $\epsilon$. As a result, the complexity is in the order of

$$O\left(\max\left(\frac{G_1L_2}{\epsilon}, \frac{\sigma^2}{\epsilon^2}, \frac{\Omega\sigma_0^2}{B\epsilon^2}, \frac{\sigma_0^2\Omega}{n\epsilon^2}\right)\right).$$

□

**Critical:** The convergence results above remain valid for ALEXR-v2 even when the outer functions $f_i$ are smooth. If $f_i$ is a smooth Legendre function, ALEXR-v1 can also be applied and its convergence can be established. The key is to note that

$$\left\|\nabla\psi_i(\bar{y}_{i,t+1}) - \nabla\psi_i(y_{i,t})\right\|_2^2 = \left\|\nabla f_i^*(\bar{y}_{i,t+1}) - \nabla f_i^*(y_{i,t})\right\|_2^2 = \left\|\bar{\mathbf{u}}_{i,t} - \mathbf{u}_{i,t-1}\right\|_2^2,$$

where $\mathbf{u}_{i,t-1}$ is defined in Lemma 5.14 and $\bar{\mathbf{u}}_{i,t}$ is a virtual sequence similar to $\mathbf{u}_{i,t}$ (5.64) except that all coordinates are updated by:

$$\bar{\mathbf{u}}_{i,t} = \frac{1}{1+\alpha_t}\mathbf{u}_{i,t-1} + \frac{\alpha_t}{1+\alpha_t}\tilde{g}_{i,t}, \forall i. \tag{5.64}$$

Then, similar to the analysis of SOX, we can establish a bound of $\sum_{t=1}^{T}\sum_{i=1}^{n}\mathbb{E}[\left\|\bar{\mathbf{u}}_{i,t} - \mathbf{u}_{i,t-1}\right\|_2^2]$ and use it to prove the convergence of ALEXR-v1. However, it remains unclear whether ALEXR-v1 provides any convergence advantage over ALEXR-v2 when $f_i$ are smooth.

### 5.4.5 Double-loop ALEXR for weakly convex inner functions

ALEXR can be also useful for solving non-convex FCCO with convex outer functions and weakly convex inner functions. In particular, we consider the following non-convex problem:

$$\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n} f_i(g_i(\mathbf{w})) + r(\mathbf{w}),$$

where $g_i : \mathbb{R}^d \to \mathbb{R}^{d'}$ and $f_i : \mathbb{R}^{d'} \to \mathbb{R}$ satisfy the following conditions:

**Assumption 5.12.** *Assume*

*(i) $f_i$ is convex, $G_1$-Lipschitz continuous and $\partial f(g) \geq 0$.*
*(ii) each dimension of $g_i$ is $\rho_2$-weakly convex and $G_2$-Lipschitz continuous.*
*(iii) $r(\mathbf{w})$ is a convex function.*

The key idea is to solve the following quadratic problem sequentially:

$$\mathbf{w}_{t+1} \approx \arg\min \bar{F}(\mathbf{w}, \mathbf{w}_t) := \frac{1}{n}\sum_{i=1}^{n} f_i(g_i(\mathbf{w})) + \frac{\bar{\rho}}{2}\|\mathbf{w} - \mathbf{w}_t\|_2^2,$$

where $\bar{\rho} > \rho$, with $\rho$ being the weak-convexity parameter of $F(\mathbf{w})$. We can employ ALEXR to solve $\min_{\mathbf{w}} \bar{F}(\mathbf{w}, \mathbf{w}_t)$ approximately up to an $\epsilon$-level. This yields a double-loop scheme.

| $f_i$ | $g_i$ | $r$ | F | Algorithm | Convergence Measure | Complexity | Theorem |
|---|---|---|---|---|---|---|---|
| sm | - | 0 | ncx, sm | SOX | Stationary | $O\left(\frac{n\sigma_0^2}{B\epsilon^4}\right)$ | Thm. 5.1 |
| sm | mss | 0 | ncx | MSVR | Stationary | $O\left(\frac{n\sigma_0}{B\epsilon^3}\right)$ | Thm. 5.2 |
| sm | - | pm | ncx, sm | SOX | Stationary | $O\left(\frac{n\sigma_0^2}{B\epsilon^4}\right)$ | Thm. 5.1 |
| sm | mss | pm | ncx | MSVR | Stationary | $O\left(\frac{n\sigma_0}{B\epsilon^3}\right)$ | Thm. 5.2 |
| wc, nd | wc | 0 | ncx | SONX (v1) | Nearly Stationary | $O\left(\frac{n\sigma_0^2}{B\epsilon^8}\right)$ | Thm. 5.3 |
| sm, nd | wc | 0 | ncx | SONX (v1) | Nearly Stationary | $O\left(\frac{n\sigma_0^2}{B\epsilon^5}\right)$ | Thm. 5.4 |
| wc, nd | wc | 0 | ncx | SONX (v2) | Nearly Stationary | $O\left(\frac{n\sigma_0}{B\epsilon^6}\right)$ | Thm. 5.5 |
| sm, nd | wc | 0 | ncx | SONX (v2) | Nearly Stationary | $O\left(\frac{n\sigma_0}{B\epsilon^4}\right)$ | Thm. 5.5 |
| wc, pm | sm | 0 | ncx | SONEX (v1) | Approx. Stationary | $O\left(\frac{n\sigma_0^2}{B\epsilon^7}\right)$ | Cor. 5.1 |
| wc, pm | sm | 0 | ncx | SONEX (v2) | Approx. Stationary | $O\left(\frac{n\sigma_0}{B\epsilon^5}\right)$ | Thm. 5.6 |
| nd, cvx, $f_i^*$ pm | sm, cvx | cvx, pm | cx | ALEXR (v2) | Obj. Gap | $O\left(\max\left(\frac{\sigma^2}{\epsilon^2}, \frac{n\sigma_0^2}{B\epsilon^2}\right)\right)$ | Thm. 5.9 |
| nd, cvx, $f_i^*$ pm | cvx | cvx, pm | cx | ALEXR (v2) | Obj. Gap | $O\left(\max\left(\frac{1}{\epsilon^2}, \frac{n\sigma_0^2}{B\epsilon^2}\right)\right)$ | Thm. 5.8 |
| sm, nd, cvx | cvx | scx, pm | cx | ALEXR | Dist. Gap | $O\left(\max\left(\frac{1}{\mu\epsilon}, \frac{n\sigma_0^2}{B\epsilon}\right)\right)$ | Thm.5.7 |
| sm, nd, cvx | sm, cvx | scx, pm | cx | ALEXR | Dist. Gap | $O\left(\max\left(\frac{\sigma^2}{\mu\epsilon}, \frac{n\sigma_0^2}{B\epsilon}\right)\right)$ | Thm.5.7 |
| sm, nd, cvx, $f_i^*$ pm | wc | cx, pm | ncx | ALEXR-DL | Nearly Stationary | $O\left(\max\left(\frac{1}{\epsilon^4}, \frac{n\sigma_0^2}{B\epsilon^4}\right)\right)$ | - |
| nd, cvx, $f_i^*$ pm | wc | cx, pm | ncx | ALEXR-DL | Approx. Stationary | $O\left(\max\left(\frac{1}{\epsilon^5}, \frac{n\sigma_0^2}{B\epsilon^5}\right)\right)$ | - |

Table 5.2: Complexity of solving FCCO $F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n f_i(g_i(\mathbf{w})) + r(\mathbf{w})$ under different conditions of $f_i$ and $g_i$, where $f_i$ is a deterministic Lipschitz continuous and $g_i$ is mean Lipschitz continuous. pms means "proximal mapping is simple to compute", mss mean "mean squared smoothness", and ALEXR-DL denotes a double-loop method that employs ALEXR in the inner loop.

We highlight the key results as follows. If each $f_i$ is non-smooth, the double loop method achieves a sample complexity of $O\left(\frac{n\sigma_0^2}{B\epsilon^6}\right)$ for finding a nearly $\epsilon$-stationary solution. The analysis can be found in (Zhou et al., 2025).

If each $f_i$ is $L_1$-smooth, the sample complexity improves to $O\left(\frac{nL_1\sigma_0^2}{B\epsilon^4}\right)$ for obtaining a nearly $\epsilon$-stationary solution. This result further implies that, for non-smooth $f_i$, we may apply the Nesterov smoothing $\bar{f}_i$ in (5.20) with $\bar{\rho}_1 = 1/\epsilon$, so that $\bar{f}_i$ becomes $L_1 = \bar{\rho}_1$-smooth. Hence, Proposition 5.1 implies that the double-loop ALEXR algorithm can find an approximate $\epsilon$-stationary stationary solution of $F(\mathbf{w})$ with a sample complexity $O\left(\frac{nL_1\sigma_0^2}{B\epsilon^4}\right) = O\left(\frac{n\sigma_0^2}{B\epsilon^5}\right)$. The analysis can be found in (Chen et al., 2025b).

Finally, we summarize the sample complexities of all methods introduced in this chapter in Table 5.2, and illustrate the relationship between different methods in Figure 5.1.

**Algorithm 19** Abstract Stochastic Update Scheme for Convex FCCO

Initialize affine subspaces $\mathfrak{X}_0, \mathfrak{Y}_0, \mathfrak{g}_0, \mathfrak{G}_0$
**for** $t = 0, 1, \ldots, T-1$ **do**
  Sample a batch $\mathcal{B}_t \subset \{1, \ldots, n\}$, $|\mathcal{B}_t| = B$
  **for** each $i \in \mathcal{B}_t$ **do**
    Sample $\zeta_{i,t}, \tilde{\zeta}_{i,t}$ from $\mathbb{P}_i$
    $\mathfrak{g}_{t+1}^{(i)} = \mathfrak{g}_t^{(i)} + \operatorname{span}\{g_i(\hat{x}; \zeta_{i,t}) \mid \hat{x} \in \mathfrak{X}_t\}$

$$\mathfrak{Y}_{t+1}^{(i)} = \mathfrak{Y}_t^{(i)} + \operatorname{span}\left\{\arg\max_{y_i}\left\{y_i \hat{g}^{(i)} - f_i^*(y_i) - \frac{1}{\alpha} D_{\psi_i}(y_i, \hat{y}^{(i)})\right\} \mid \hat{g}^{(i)} \in \mathfrak{g}_{t+1}^{(i)}, \hat{y}^{(i)} \in \mathfrak{Y}_t^{(i)}\right\}$$

  **end for**
  For each $i \notin \mathcal{S}_t$, $\mathfrak{g}_{t+1}^{(i)} = \mathfrak{g}_t^{(i)}$, $\mathfrak{Y}_{t+1}^{(i)} = \mathfrak{Y}_t^{(i)}$
  $\mathfrak{G}_{t+1} = \mathfrak{G}_t + \operatorname{span}\left\{\frac{1}{B}\sum_{i \in \mathcal{B}_t} \hat{y}^{(i)} \nabla g_i(\hat{x}; \tilde{\zeta}_{i,t}) \mid \hat{x} \in \mathfrak{X}_t, \hat{y} \in \mathfrak{Y}_{t+1}\right\}$
  $\mathfrak{X}_{t+1} = \mathfrak{X}_t + \operatorname{span}\left\{\arg\min_x \left\{\hat{G}^\top x + r(x) + \frac{1}{2\eta}\|x - \hat{x}\|_2^2\right\} \mid \hat{x} \in \mathfrak{X}_t, \hat{G} \in \mathfrak{G}_{t+1}\right\}$
**end for**

### 5.4.6 Lower Bounds

In this section, we prove that the complexities of ALEXR for strongly convex and convex FCCO problems are nearly optimal by establishing the matching *lower* bounds.

**What is a lower bound?**

A lower bound states: for any algorithm in a certain class, there exists a "hard" optimization problem such that the algorithm cannot converge faster than a specified rate.

Lower bounds for convex optimization are typically derived under the standard oracle model, where the algorithm has access only to first-order information—either exact gradients in the deterministic setting or unbiased stochastic gradients in the stochastic setting. In the latter case, a classical result by Nemirovski and Yudin establishes that no stochastic algorithm using unbiased gradient oracles can achieve a convergence rate faster than $O(1/\sqrt{T})$ in terms of the objective gap after $T$ iterations. For strongly convex problems, this lower bound improves to $O(1/T)$. Nevertheless, these lower bounds do not apply to convex FCCO problems or to ALEXR, because the algorithm does not have access to unbiased stochastic gradients.

Below, we establish lower bounds for an abstract stochastic update scheme described in Algorithm 19, where the symbol "+" denotes Minkowski addition. We consider an oracle model that, upon receiving a query point, returns unbiased stochastic function values and stochastic gradients of the inner functions $g_i$, as well as the solution to the proximal mirror-descent update of $f_i^*$ with respect to a proximal function $\psi_i$. Since there are $n$ inner functions in total, we assume that at each iteration the algorithm is allowed to access information from only $B$ randomly selected in-

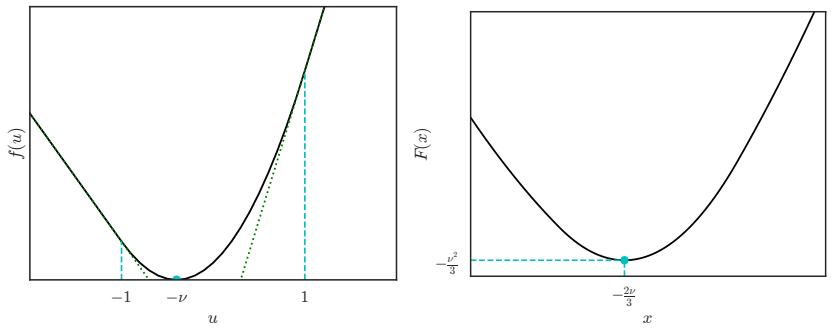


Fig. 5.2: Visualization of $f$ (left) and $F$ (right) in (5.65).

ner functions. Algorithm 19 is sufficiently general to encompass ALEXR, as well as SOX and MSVR.

**Theorem 5.10** *Consider the abstract scheme (Algorithm 19) with an initialization* $\mathfrak{X}_0^{(i)} = \{0\}$, $\mathfrak{Y}_0^{(i)} = \{0\}$, $\mathfrak{g}_0^{(i)} = \emptyset$, $\mathfrak{G}_0^{(i)} = \emptyset$.

- *There exists a convex FCCO problem (5.26) with smooth $f_i$ and $\mu$-strongly convex $r$ such that any algorithm in the abstract scheme requires at least $T = \Omega\left(\frac{n\sigma_0^2}{B\epsilon}\right)$ iterations to find an $\bar{x}$ that satisfies $\mathbb{E}\left[\frac{\mu}{2}\|\bar{x} - x_*\|_2^2\right] \leq \epsilon$ or $\mathbb{E}[F(\bar{x}) - F(x_*)] \leq \epsilon$.*
- *There exists a convex FCCO problem (5.26) with non-smooth $f_i$ such that any algorithm in the abstract scheme requires at least $T = \Omega\left(\frac{n\sigma_0^2}{B\epsilon^2}\right)$ iterations to find an $\bar{x}$ that satisfies $\mathbb{E}[F(\bar{x}) - F(x_*)] \leq \epsilon$.*

**Why it matters**

In light of this theorem, we see that ALEXR (v1/v2) attains a nearly optimal complexity up to a logarithmic factor for solving strongly convex FCCO problems, as its upper bounds are $\tilde{O}\left(\max\left(\frac{1}{\mu\epsilon}, \frac{n\sigma_0^2}{B\epsilon}\right)\right)$. Moreover, ALEXR-v2 achieves the optimal complexity for solving convex FCCO problems with non-smooth outer functions.

*Proof.* We construct the hard problems for (i) smooth $f_i$; and (ii) non-smooth $f_i$ separately.

**(i) Smooth $f_i$ and strongly convex $r$:** Consider the following strongly convex FCCO problem

$$\min_{x\in\mathcal{X}} \quad F(x) = \frac{1}{n}\sum_{i=1}^{n} f(g_i(x)) + r(x),$$
$$f(u) = \begin{cases} (\nu-1)u + \frac{1}{2}(\nu-1)^2 + \nu - 1 - \frac{\nu^2}{2}, & u \in (-\infty, -1) \\ \frac{1}{2}(u+\nu)^2 - \frac{\nu^2}{2}, & u \in [-1, 1] \\ (1+\nu)u + \frac{1}{2}(1+\nu)^2 - 1 - \nu - \frac{\nu^2}{2}, & u \in (1, \infty) \end{cases}, \quad r(x) = \frac{1}{4n}\|x\|_2^2, \tag{5.65}$$

where $\mathcal{X} = [-1, 1]^n$, the outer function $f : \mathbb{R} \to \mathbb{R}$ is smooth and Lipschitz continuous for some $\nu \in (0, 1/2)$. Besides, the inner function $g_i : \mathbb{R}^n \to \mathbb{R}$ is $g_i(x) = \mathbb{E}_{\zeta\sim\mathbb{P}}[g_i(x;\zeta)]$ and $g_i(x;\zeta) = x_i + \zeta$, where $\zeta$ follows a distribution $\mathbb{P}$ defined below:

$$\mathbb{P} : \begin{cases} \Pr(\zeta = -\nu) = 1 - p, \\ \Pr(\zeta = \nu(1-p)/p) = p \end{cases}, \text{ where } p \coloneqq \frac{\nu^2}{\sigma_0^2} < 1.$$

We will determine the values of $\nu$ later. We can verify that

$$\mathbb{E}_\zeta[|g_i(x;\zeta) - g_i(x)|^2] = \mathbb{E}_\zeta[\zeta^2] = \nu^2(1-p) + \frac{\nu^2(1-p)^2}{p} = \frac{\nu^2(1-p)}{p} \le \sigma_0^2.$$

By the definition of convex conjugate, for any $y_i \in \mathbb{R}$ we have

$$\begin{aligned} f^*(y_i) = \max\Bigg\{ & \sup_{u<-1}\left\{uy_i - \left((\nu-1)u + \frac{1}{2}(\nu-1)^2 + \nu - 1 - \frac{\nu^2}{2}\right)\right\}, \\ & \sup_{-1\le u\le 1}\left\{uy_i - \frac{1}{2}(u+\nu)^2 + \frac{\nu^2}{2}\right\}, \\ & \sup_{u>1}\left\{uy_i - \left((1+\nu)u + \frac{1}{2}(1+\nu)^2 - 1 - \nu - \frac{\nu^2}{2}\right)\right\}\Bigg\} \end{aligned} \tag{5.66}$$

$$= \begin{cases} +\infty, & y_i \in (-\infty, \nu-1) \cup (\nu+1, \infty) \\ \frac{1}{2}(y_i - \nu)^2, & y_i \in [\nu-1, \nu+1]. \end{cases} \tag{5.67}$$

We define $F_i(x_i) \coloneqq f(g_i(x)) + \frac{1}{4}[x_i]^2$ such that $F(x) = \frac{1}{n}\sum_{i=1}^n F_i(x_i)$. Let $x_* = \arg\min_{x\in\mathcal{X}} F(x)$. Since the problem is separable over the coordinates, we have $x_{i,*} = \arg\min_{x\in[-1,1]} F_i(x_i)$. Thus, we have $x_{i,*} = -\frac{2\nu}{3}$ and $F_i(x_{i,*}) = -\frac{\nu^2}{3}$.

Since $\mathbb{P}_i = \mathbb{P}$ in the "hard" problem (5.65), the abstract scheme (Algorithm 19) only needs to sample shared $\zeta_t, \tilde{\zeta}_t \sim \mathbb{P}$ for all coordinates $i \in \mathcal{S}_t$ in the $t$-th iteration. For any $i \in [n]$, suppose that $\mathfrak{g}_\tau^{(i)} = \emptyset$ or $\{-\nu\}$, $\mathfrak{y}_\tau^{(i)} = \{0\}$, $\mathfrak{x}_\tau^{(i)} = \{0\}$ for all $\tau \le t$. Note that when $\mathfrak{g}_\tau^{(i)} = \emptyset$, it means that the corresponding $y^{(i)}$ will not be updated. Then,

- If $i \notin \mathcal{B}_t$, the abstract scheme (Algorithm 19) leads to

$$\mathfrak{g}_{t+1}^{(i)} = \emptyset \text{ or } \{-\nu\}, \quad \mathfrak{y}_{t+1}^{(i)} = \{0\}, \quad \mathfrak{x}_{t+1}^{(i)} = \{0\}.$$

- If $i \in \mathcal{B}_t$ and $\zeta_t = -\nu$, the abstract scheme (Algorithm 19) proceeds as

$$\begin{aligned}
\mathfrak{g}_{t+1}^{(i)} &= \mathfrak{g}_t^{(i)} + \operatorname{span}\left\{\hat{x}_i + \zeta_t \mid \hat{x}_i \in \mathfrak{X}_t^{(i)}\right\}, \\
\mathfrak{Y}_{t+1}^{(i)} &= \mathfrak{Y}_t^{(i)} \\
&+ \operatorname{span}\left\{\operatorname*{arg\,max}_{y_i \in [\nu-1, \nu+1]} \left\{y_i \hat{g}_i - \frac{1}{2}(y_i - \nu)^2 - \frac{1}{\alpha} D_{\psi_i}(y_i, \hat{y}_i)\right\} \mid \hat{g}_i \in \mathfrak{g}_{t+1}^{(i)}, \hat{y}_i \in \mathfrak{Y}_t^{(i)}\right\}, \\
\mathfrak{X}_{t+1}^{(i)} &= \mathfrak{X}_t^{(i)} \\
&+ \operatorname{span}\left\{\operatorname*{arg\,min}_{x_i \in [-1,1]} \left\{\frac{1}{B}\hat{y}_i x_i + \frac{1}{4n}[x_i]^2 + \frac{1}{2\eta}(x_i - \hat{x}_i)^2\right\} \mid \hat{y}_i \in \mathfrak{Y}_{t+1}^{(i)}, \hat{x}_i \in \mathfrak{X}_t^{(i)}\right\}.
\end{aligned}$$

Then, we can derive that $\mathfrak{g}_{t+1}^{(i)} = \emptyset$ or $\{-\nu\}$, $\mathfrak{Y}_{t+1}^{(i)} = \{0\}$, and $\mathfrak{X}_{t+1}^{(i)} = \{0\}$.

To sum up, given the event $\bigcap_{\tau=1}^{t}\{\mathfrak{g}_\tau^{(i)} = \emptyset \text{ or } \{-\nu\}, \mathfrak{Y}_\tau^{(i)} = \{0\}, \mathfrak{X}_\tau^{(i)} = \{0\}\}$, we can make sure that $\{\mathfrak{g}_{t+1}^{(i)} = \emptyset \text{ or } \{-\nu\} \wedge \mathfrak{Y}_{t+1}^{(i)} = \{0\} \wedge \mathfrak{X}_{t+1}^{(i)} = \{0\}\}$ for the abstract scheme in Algorithm 19 when one of the following mutually exclusive events happens:

- Event I: $i \notin \mathcal{B}_t$;
- Event II: $i \in \mathcal{B}_t$ and $\zeta_t = -\nu$.

Note that the random variable $\zeta_t$ is independent of $\mathcal{B}_t$. Thus, the probability of the event $E_{t+1}^{(i)} \coloneqq \{\mathfrak{g}_{t+1}^{(i)} = \emptyset \text{ or } \{-\nu\} \wedge \mathfrak{Y}_{t+1}^{(i)} = \{0\} \wedge \mathfrak{X}_{t+1}^{(i)} = \{0\}\}$ conditioned on $\bigcap_{\tau=1}^{t} E_\tau^{(i)}$ can be bounded as

$$\begin{aligned}
\Pr\left[E_{t+1}^{(i)} \mid \bigcap_{\tau=1}^{t} E_\tau^{(i)}\right] &= \mathbb{P}\left[\left\{\mathfrak{g}_{t+1}^{(i)} = \emptyset \text{ or } \{-\nu\} \wedge \mathfrak{Y}_{t+1}^{(i)} = \{0\} \wedge \mathfrak{X}_{t+1}^{(i)} = \{0\}\right\} \mid \bigcap_{\tau=1}^{t} E_\tau^{(i)}\right] \\
&\geq \mathbb{P}\left[\{i \notin \mathcal{B}_t\}\right] + \mathbb{P}\left[\{\{i \in \mathcal{B}_t\} \wedge \{\zeta_t = -\nu\}\}\right] \\
&= \mathbb{P}\left[\{i \notin \mathcal{B}_t\}\right] + \mathbb{P}\left[\{i \in \mathcal{B}_t\}\right]\mathbb{P}\left[\{\zeta_t = -\nu\}\right] \\
&= \left(1 - \frac{B}{n}\right) + \frac{B}{n}(1-p) = 1 - \frac{Bp}{n}.
\end{aligned}$$

Since $\mathcal{B}_t$ and $\zeta_t$ in different iterations $t$ are mutually independent, we have

$$\Pr\left[E_T^{(i)}\right] \geq \mathbb{P}\left[\bigcap_{t=0}^{T-1} E_{t+1}^{(i)}\right] = \prod_{t=0}^{T-1} \mathbb{P}\left[E_{t+1}^{(i)} \mid \bigcap_{t=1}^{t} E_t^{(i)}\right] = \left(1 - \frac{Bp}{n}\right)^T > 3/4 - \frac{TBp}{n},$$

where the last inequality is due to the Bernoulli inequality $(1+x)^r \geq 1 + rx$ for every integer $r \geq 1$ and $x \geq -1$.

Thus, if $T < \frac{n}{4Bp}$ we have $\Pr\left[E_T^{(i)}\right] > \frac{1}{2}$. Let us set $\nu = 3\sqrt{2\epsilon}$ such that $p = \frac{\nu^2}{\sigma_0^2} = \frac{18\epsilon}{\sigma_0^2}$. For any $i \in [n]$ and any output $\bar{x}_i \in \mathfrak{X}_T^{(i)}$, we have

$$\begin{aligned}\mathbb{E}\left[\left(\bar{x}_i - x_{i,*}\right)^2\right] &= \mathbb{E}\left[\mathbb{I}_{E_T^{(i)}}\left(\bar{x}_i - x_{i,*}\right)^2 + \mathbb{I}_{\overline{E_T^{(i)}}}\left(\bar{x}^{(i)} - x_{i,*}\right)^2\right] \\ &\geq \mathbb{E}\left[\mathbb{I}_{E_T^{(i)}}\left(\bar{x}_i - x_{i,*}\right)^2\right] \\ &= \mathbb{E}\left[\mathbb{I}_{E_T^{(i)}}\left(x_{i,*}\right)^2\right] = \Pr\left[E_T^{(i)}\right]\left(x_{i,*}\right)^2 > \frac{2\nu^2}{9} = 4\epsilon,\end{aligned}$$

where $\mathbb{I}_E$ denotes the indicator function of an event $E$. Moreover, we have

$$\begin{aligned}\mathbb{E}[F_i(\bar{x}_i) - F_i(x_{i,*})] &= \mathbb{E}\left[\mathbb{I}_{E_T^{(i)}}\left(F_i(\bar{x}_i) - F_i(x_{i,*})\right) + \mathbb{I}_{\overline{E_T^{(i)}}}\left(F_i(\bar{x}_i) - F_i(x_{i,*})\right)\right] \\ &\geq \mathbb{E}\left[\mathbb{I}_{E_T^{(i)}}\left(F_i(\bar{x}_i) - F_i(x_{i,*})\right)\right] \\ &= \mathbb{E}\left[\mathbb{I}_{E_T^{(i)}}\left(F_i(0) - F_i(x_{i,*})\right)\right] = \Pr[E_T^{(i)}]\left(F_i(0) - F_i(x_{i,*})\right) \\ &> \frac{\nu^2}{6} > \epsilon.\end{aligned}$$

Since the derivations above hold for arbitrary $i \in [n]$ and the $r(x)$ in (5.65) is $\frac{1}{2n}$-strongly convex ($\mu = \frac{1}{2n}$), we can derive that

$$\mathbb{E}\left[\frac{\mu}{2}\left\|\bar{x} - x_*\right\|_2^2\right] = \mathbb{E}\left[\frac{1}{4n}\left\|\bar{x} - x_*\right\|_2^2\right] = \frac{1}{4n}\sum_{i=1}^{n}\mathbb{E}\left[\left(\bar{x}_i - x_{i,*}\right)^2\right] > \epsilon,$$

$$\mathbb{E}\left[F(\bar{x}) - F(x_*)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[F_i(\bar{x}_i) - F_i(x_{i,*})\right] > \epsilon.$$

Thus, to find an output $\bar{x}$ that satisfies $\mathbb{E}\left[\frac{\mu}{2}\left\|\bar{x} - x_*\right\|_2^2\right] \leq \epsilon$ or $\mathbb{E}\left[F(\bar{x}) - F(x_*)\right] \leq \epsilon$, the abstract scheme requires at least $T \geq \frac{n}{4Bp} = \frac{n\sigma_0^2}{72B\epsilon}$ iterations.

**(ii) Non-smooth $f_i$:** Let $g_i(x) = \mathbb{E}_\zeta[x_i + \zeta] = x_i$ be defined the same as in the smooth case. Let $F_i(x_i) \coloneqq f(g_i(x)) + \frac{\alpha}{2}[x_i]^2 = \beta\max\{x_i, -\nu\} + \frac{\alpha}{2}[x_i]^2$ such that $F(x) = \frac{1}{n}\sum_{i=1}^{n}F_i(x_i)$, where $\alpha, \beta > 0$. Let the domain $\mathcal{X}$ be $[-2\nu, 2\nu]^n$. Hence, $f$ is $\beta$-Lipschitz continuous and $F$ is $\alpha$-strongly convex. By the definition of convex conjugate, we have $f(\hat{g}_i) = \max_{y_i \in [0,\beta]}\left\{y_i\hat{g}_i - \nu(\beta - y_i)\right\}$.

Since the problem is separable over the coordinates, we have

$$x_{i,*} = \operatorname*{arg\,min}_{x \in [-2\nu, 2\nu]} F_i(x_i) = \operatorname*{arg\,min}_{x_i \in [-2\nu, 2\nu]}\left\{\beta\max\{x_i, -\nu\} + \frac{\alpha}{2}[x_i]^2\right\}.$$

Considering

$$F_i(x_i) = \begin{cases}\beta x_i + \frac{\alpha}{2}[x_i]^2 & x_i \geq -\nu \\ -\beta\nu + \frac{\alpha}{2}[x_i]^2 & x_i < -\nu\end{cases},$$

we have

$$x_{i,*} = \begin{cases} -\beta/\alpha & \text{if } \alpha > \beta/\nu \\ -\nu & \text{if } \alpha \in \frac{\beta}{\nu}[0,1] \end{cases}, \quad F_i(x_{i,*}) \le \begin{cases} -\beta^2/(2\alpha) & \text{if } \alpha > \beta/\nu \\ -\beta\nu/2 & \text{if } \alpha \in \frac{\beta}{\nu}[0,1]. \end{cases}$$

Since $F_i(0) = 0$, we can derive that $F_i(0) - F_i(x_{i,*}) \ge \frac{1}{2}\min\{\beta\nu, \beta^2/\alpha\}$. Consider an arbitrary $i \in [n]$. Suppose that $\mathfrak{g}_\tau^{(i)} = \emptyset$ or $\{-\nu\}$, $\mathfrak{X}_\tau^{(i)} = \{0\}$, $\mathfrak{Y}_\tau^{(i)} = \{0\}$ for all $\tau \le t$.

- If $i \notin \mathcal{B}_t$, the abstract scheme (Algorithm 19) leads to

$$\mathfrak{g}_{t+1}^{(i)} = \emptyset \text{ or } \{-\nu\}, \quad \mathfrak{Y}_{t+1}^{(i)} = \{0\}, \quad \mathfrak{X}_{t+1}^{(i)} = \{0\}.$$

- If $i \in \mathcal{B}_t$, the abstract scheme (Algorithm 19) proceeds as

$$\begin{aligned}
&\mathfrak{g}_{t+1}^{(i)} = \mathfrak{g}_t^{(i)} + \operatorname{span}\left\{\hat{x}_i + \zeta_t \mid \hat{x}_i \in \mathfrak{X}_t^{(i)}\right\}, \\
&\mathfrak{Y}_{t+1}^{(i)} = \mathfrak{Y}_t^{(i)} \\
&+ \operatorname{span}\left\{\operatorname*{arg\,max}_{y_i \in [0,\beta]} \left\{y_i \hat{g}_i - \nu(\beta - y_i) - \frac{1}{\alpha} D_\psi(y_i, \hat{y}_i)\right\} \mid \hat{g}_i \in \mathfrak{g}_{t+1}^{(i)}, \hat{y}_i \in \mathfrak{Y}_t^{(i)}\right\}, \\
&\mathfrak{X}_{t+1}^{(i)} = \mathfrak{X}_t^{(i)} \\
&+ \operatorname{span}\left\{\operatorname*{arg\,min}_{x_i \in [-2\nu, 2\nu]} \left\{\frac{1}{B}\hat{y}_i x_i + \frac{1}{n}[x_i]^2 + \frac{1}{2\eta}(x_i - \hat{x}_i)^2\right\} \mid \hat{y}_i \in \mathfrak{Y}_{t+1}^{(i)}, \hat{x}_i \in \mathfrak{X}_t^{(i)}\right\}.
\end{aligned}$$

Due to the same reason as in the smooth $f_i$ case, the probability of the event $E_T^{(i)} \coloneqq \{\mathfrak{g}_T^{(i)} = \emptyset \text{ or } \{-\nu\} \wedge \mathfrak{Y}_T^{(i)} = \{0\} \wedge \mathfrak{X}_T^{(i)} = \{0\}\}$ can be bounded as

$$\Pr\left[E_T^{(i)}\right] \ge \mathbb{P}\left[\bigcap_{t=0}^{T-1} E_{t+1}^{(i)}\right] = \prod_{t=0}^{T-1} \mathbb{P}\left[E_{t+1}^{(i)} \mid \bigcap_{t=1}^{t} E_t^{(i)}\right] = \left(1 - \frac{Bp}{n}\right)^T > 3/4 - \frac{TBp}{n}.$$

Thus, if $T < \frac{n}{4Bp}$ we have $\mathbb{P}\left[E_T^{(i)}\right] > \frac{1}{2}$. Let us set $\beta = G_1$, $\nu = \frac{4\epsilon}{G_1}$ such that $p \coloneqq \frac{\nu^2}{\sigma_0^2} = \frac{16\epsilon^2}{G_1^2\sigma_0^2}$. For any $i \in [n]$ and any output $\bar{x}_i \in \mathfrak{X}_T^{(i)}$, we have

$$\begin{aligned}
\mathbb{E}[F_i(\bar{x}_i) - F_i(x_{i,*})] &= \mathbb{E}\left[\mathbb{I}_{E_T^{(i)}}\left(F_i(\bar{x}_i) - F_i(x_{i,*})\right) + \mathbb{I}_{\overline{E_T^{(i)}}}\left(F_i(\bar{x}_i) - F_i(x_{i,*})\right)\right] \\
&\ge \mathbb{E}\left[\mathbb{I}_{E_T^{(i)}}\left(F_i(\bar{x}_i) - F_i(x_{i,*})\right)\right] \\
&= \mathbb{E}\left[\mathbb{I}_{E_T^{(i)}}\left(F_i(0) - F_i(x_{i,*})\right)\right] \\
&= \Pr[E_T^{(i)}]\left(F_i(0) - F_i(x_{i,*})\right) > \min\{\beta\nu, \beta^2/\alpha\}/4 = \epsilon.
\end{aligned}$$

Since the derivations above hold for arbitrary $i \in [n]$, we can derive that

$$\mathbb{E}[F(\bar{x}) - F(x_*)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[F_i(\bar{x}_i) - F_i(x_{i,*})] > \epsilon.$$

Thus, to find an output $\bar{x}$ that satisfies $\mathbb{E}[F(\bar{x}) - F(x_*)] \leq \epsilon$, the abstract scheme requires at least $T \geq \frac{n}{4Bp} = \frac{nG_1^2\sigma_0^2}{64B\epsilon^2}$ iterations. □

**Critical:** From the proof of the non-smooth case, we can see that even when the overall objective is strongly convex, the lower bound complexity is still $T = \Omega\left(\frac{n\sigma_0^2}{B\epsilon^2}\right)$ as long as $f_i$ is non-smooth. This behavior contrasts with standard strongly stochastic optimization with an optimal complexity of $O(1/\epsilon)$ and highlights a fundamental challenge in solving compositional problems.

## 5.5 Stochastic Optimization of Compositional OCE

The goal of this section is to present and analyze stochastic algorithms for solving compositional OCE (COCE) risk minimization as introduced in Chapter 3. In particular, we consider the following abstract problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d, \boldsymbol{\nu}\in\mathbb{R}^n} F(\mathbf{w}, \boldsymbol{\nu}) := \frac{1}{n}\sum_{i=1}^{n} F_i(\mathbf{w}, \nu_i), \tag{5.68}$$

where

$$F_i(\mathbf{w}, \nu_i) = \mathbb{E}_{\zeta\sim\mathbb{P}_i}[\Phi_i(\mathbf{w}, \nu_i; \zeta)], \quad \Phi_i(\mathbf{w}, \nu_i; \zeta) = \tau\phi^*\left(\frac{s_i(\mathbf{w}; \zeta) - \nu_i}{\tau}\right) + \nu_i,$$

where $\tau > 0$ is a constant.

In the special case when $\phi^*(\cdot) = [\cdot]_+/\alpha$ for some $\alpha \in (0, 1)$, the general COCE minimization problem reduces to

$$\min_{\mathbf{w},\boldsymbol{\nu}} F(\mathbf{w}, \boldsymbol{\nu}) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\zeta\sim\mathbb{P}_i} \frac{[s_i(\mathbf{w}; \zeta) - \nu_i]_+}{\alpha} + \nu_i. \tag{5.69}$$

We refer to this problem as the **compositional CVaR minimization (CCVaR)** problem. The direct one-way partial AUC optimization problem (2.39) can be reformulated as an instance of CCVaR minimization as shown in (6.26).

In the special case when $\phi^*(\cdot) = \exp(\cdot) - 1$, the problem (5.68) reduces to

$$\min_{\mathbf{w}} F(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} \tau \log\left(\mathbb{E}_{\zeta\sim\mathbb{P}_i} \exp\left(\frac{s_i(\mathbf{w}; \zeta)}{\tau}\right)\right). \tag{5.70}$$

**Algorithm 20** The ASGD Algorithm for solving (5.68)

Initialize $\mathbf{w}_0, \boldsymbol{\nu}_0$, step sizes $\eta_t$ and $\gamma_t$
**for** $t = 0, 1 \ldots, T-1$ **do**
  Sample $\mathcal{B}_t \subset \{1, \ldots, n\}$ and $|\mathcal{B}_t| = B$
  **for** each $i \in \mathcal{B}_t$ **do**
    Update $\boldsymbol{\nu}_{i,t+1} = \boldsymbol{\nu}_{i,t} - \gamma_t \partial_2 \Phi_i(\mathbf{w}_t, \boldsymbol{\nu}_{i,t}; \zeta_{i,t})$
  **end for**
  Compute $\mathbf{z}_t = \frac{1}{B} \sum_{i \in \mathcal{B}_t} \partial_1 \Phi_i(\mathbf{w}_t, \boldsymbol{\nu}_{i,t}; \zeta_{i,t})$
  Update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{z}_t$
**end for**

We refer to this problem as the **compositional entropic risk minimization (CERM)** problem. The cross-entropy loss for multi-class classification, the listwise cross-entropy loss for ranking, the indirect one-way partial AUC loss for imbalanced classification, and the contrastive losses for representation learning discussed in Chapter 2 are all instances of the CERM problem. In particular, for cross-entropy loss minimization, the proposed framework becomes especially relevant when the number of classes is very large, so that the normalization term in the loss cannot be computed efficiently. This setting naturally motivates the stochastic algorithms developed in this section.

Although we can cast the CERM problem into a special instance of FCCO, there remain some gaps to be filled. (i) For the convex CERM problem with a convex loss function $s_i(\cdot; \zeta)$, the ALEXR algorithm and its convergence analysis are not directly applicable, since the outer function $f(\cdot) = \tau \log(\cdot)$ is *not* convex, as required by ALEXR. Consequently, a convergence rate of $O(1/\epsilon^2)$ for solving convex CERM remains to be developed. (ii) For the CCVaR problem, the optimal solution of $\boldsymbol{\nu}$ given $\mathbf{w}$ is typically difficult to derive in closed form, and hence the problem cannot be reduced to an instance of FCCO. As a result, previous analyses for FCCO do not directly apply. We address these gaps in this section.

### 5.5.1 A Basic Algorithm

For optimizing the general COCE minimization problem, we present a basic stochastic algorithm in Algorithm 20. It alternates the stochastic block-coordinate update for $\boldsymbol{\nu}$ and a SGD update for $\mathbf{w}$, which is referred to as ASGD. Below, we present its convergence analysis for both convex and non-convex loss functions.

#### 5.5.1.1 Convex loss

For notational simplicity, we set $\tau = 1$ throughout the analysis.

**Assumption 5.13.** *$s_i(\cdot, \zeta)$ is a convex function.*

**Lemma 5.20** *$F(\mathbf{w}, \boldsymbol{\nu})$ is jointly convex in terms of $(\mathbf{w}^\top, \boldsymbol{\nu}^\top)^\top$ if $s_i(\cdot; \zeta)$ is convex.*

*Proof.* We prove that $\Phi_i(\mathbf{w}, \nu_i; \zeta)$ is jointly convex in terms of $(\mathbf{w}^\top, \nu_i)^\top$. Then the convexity of $F(\mathbf{w}, \boldsymbol{\nu})$ follows. Let $\mathbf{u} = (\mathbf{w}^\top, \nu)^\top$. Consider $\mathbf{u}_1, \mathbf{u}_2$, $\alpha \in [0, 1]$, and $\bar{\mathbf{u}} = \alpha\mathbf{u}_1 + (1-\alpha)\mathbf{u}_2$. Then

$$\Phi_i(\bar{\mathbf{u}}; \zeta) = \phi^*(s_i(\bar{\mathbf{w}}; \zeta) - \bar{\nu}) + \bar{\nu}.$$

If $s_i(\cdot; \zeta)$ is convex, we have $s_i(\bar{\mathbf{w}}; \zeta) \leq_i (\mathbf{w}_1; \zeta) + (1-\alpha)s_i(\mathbf{w}_2; \zeta)$. Since $\phi^*(\cdot)$ is non-decreasing (cf. Lemma 2.3), we have

$$\phi^*(s_i(\bar{\mathbf{w}}; \zeta) - \bar{\nu}) \leq \phi^*(\alpha(s_i(\mathbf{w}_1; \zeta) - \nu_1) + (1-\alpha)(s_i(\mathbf{w}_2; \zeta) - \nu_2)).$$

Since $\phi^*(\cdot)$ is convex, we further have

$$\begin{aligned}&\phi^*(\alpha(s_i(\mathbf{w}_1; \zeta) - \nu_1) + (1-\alpha)(s_i(\mathbf{w}_2; \zeta) - \nu_2))\\ &\leq \alpha\phi^*(s_i(\mathbf{w}_1; \zeta) - \nu_1) + (1-\alpha)\phi^*(s_i(\mathbf{w}_2; \zeta) - \nu_2).\end{aligned}$$

As a result,

$$\Phi_i(\bar{\mathbf{u}}; \zeta) \leq \alpha\Phi_i(\mathbf{u}_1; \zeta) + (1-\alpha)\Phi_i(\mathbf{u}_2; \zeta),$$

which proves the convexity of $\Phi_i(\mathbf{u}; \zeta)$.

□

**Assumption 5.14.** *Assume that either of the following conditions hold:*

- *(i) $F(\mathbf{w}, \boldsymbol{\nu})$ is smooth satisfying:*

$$\|\nabla_1 F(\mathbf{w}, \boldsymbol{\nu})\|_2^2 + \|\nabla_2 F(\mathbf{w}, \boldsymbol{\nu})\|_2^2 \leq 2L_F(F(\mathbf{w}, \boldsymbol{\nu}) - F(\mathbf{w}_*, \boldsymbol{\nu}_*)),$$

- *(ii) $F(\mathbf{w}, \boldsymbol{\nu})$ non-smooth such that for any $\mathbf{v}_1 \in \partial_1 F(\mathbf{w}, \boldsymbol{\nu}), \mathbf{v}_{2,i} \in \partial_2 F_i(\mathbf{w}, \nu_i)$ it holds*

$$\|\mathbf{v}_1\|_2^2 \leq G_1^2, \quad |\mathbf{v}_{2,i}|^2 \leq G_2^2,$$

*where $\mathbf{w}_*, \boldsymbol{\nu}_*$ denotes an optimal solution to (5.68), and $\nabla_1 F(\mathbf{w}, \boldsymbol{\nu})(\partial_1 F(\mathbf{w}, \boldsymbol{\nu}))$, and $\nabla_2 F(\mathbf{w}, \boldsymbol{\nu})(\partial_2 F(\mathbf{w}, \boldsymbol{\nu}))$ denote (partial) gradients with respect to $\mathbf{w}, \boldsymbol{\nu}$, respectively.*

**Critical:** For CERM, the smoothness assumption is satisfied when $s_i(\mathbf{w}; \zeta)$ is bounded, Lipschitz, and smooth. For CCVaR, the non-smoothness assumption is satisfied when $s_i(\mathbf{w}; \zeta)$ is bounded and Lipschitz.

**Assumption 5.15** (Bounded Variance)**.** *There exist $\sigma_1^2, \sigma_2^2, \delta^2$ such that*

$$\mathbb{E}_\zeta \|\nabla_1 \Phi_i(\mathbf{w}, \nu_i; \zeta) - \nabla_1 F_i(\mathbf{w}, \nu_i)\|_2^2 \le \sigma_1^2, \quad \forall i \in [n],$$
$$\mathbb{E}_\zeta \|\nabla_2 \Phi_i(\mathbf{w}, \nu_i; \zeta) - \nabla_2 F_i(\mathbf{w}, \nu_i)\|_2^2 \le \sigma_2^2, \quad \forall i \in [n],$$
$$\frac{1}{n}\sum_{i=1}^n \|\nabla_1 F_i(\mathbf{w}, \nu_i) - \nabla_1 F(\mathbf{w}, \boldsymbol{\nu})\|_2^2 \le \delta^2.$$

In the non-smooth case, the gradients above are replaced by subgradients. The subsequent analysis proceeds analogously.

**Lemma 5.21** *Let* $D_{\mathbf{w},0}^2 \coloneqq \mathbb{E}\|\mathbf{w}_0 - \mathbf{w}_*\|_2^2$ *and* $\eta_t = \eta$*, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}(2\mathbb{E}[\nabla_1 F(\mathbf{w}_t, \boldsymbol{\nu}_t)^\top(\mathbf{w}_t - \mathbf{w}_*)] - \eta\mathbb{E}\|\nabla_1 F(\mathbf{w}_t, \boldsymbol{\nu}_t)\|_2^2) \le \frac{D_{\mathbf{w},0}^2}{\eta T} + \eta\sigma^2.$$

*where* $\boldsymbol{\nu}_t = (\nu_{1,t}, \ldots, \nu_{n,t})^\top$ *and* $\sigma^2 = \frac{\sigma_1^2}{B} + \frac{\delta^2(n-B)}{B(n-1)}$.

*Proof.* Let $\mathbb{E}_t$ denote the expectation over the random samples in the $t$-th iteration. First, we note that $\mathbb{E}_t[\mathbf{z}_t] = \nabla_1 F(\mathbf{w}_t, \boldsymbol{\nu}_t)$. Similar to Lemma 5.2, we have

$$\begin{aligned}
&\mathbb{E}_t\|\mathbf{z}_t - \nabla_1 F(\mathbf{w}_t, \boldsymbol{\nu}_t)\|_2^2 \\
&= \mathbb{E}_t\left[\left\|\mathbf{z}_t - \frac{1}{B}\sum_{i\in\mathcal{B}_t}\partial_1 F_i(\mathbf{w}_t, \nu_{i,t}) + \frac{1}{B}\sum_{i\in\mathcal{B}_t}\partial_1 F_i(\mathbf{w}_t, \nu_{i,t}) - \nabla_1 F(\mathbf{w}_t, \boldsymbol{\nu}_t)\right\|_2^2\right] \\
&= \mathbb{E}_t\left[\left\|\mathbf{z}_t - \frac{1}{B}\sum_{i\in\mathcal{B}_t}\partial_1 F_i(\mathbf{w}_t, \nu_{i,t})\right\|_2^2\right] + \mathbb{E}_t\left[\left\|\frac{1}{B}\sum_{i\in\mathcal{B}_t}\partial_1 F_i(\mathbf{w}_t, \nu_{i,t}) - \nabla_1 F(\mathbf{w}_t, \boldsymbol{\nu}_t)\right\|_2^2\right] \\
&\le \frac{\sigma_1^2}{B} + \frac{\delta^2(n-B)}{B(n-1)} \coloneqq \sigma^2.
\end{aligned}$$

Due to the update of $\mathbf{w}$, we have

$$\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2 = \|\mathbf{w}_t - \mathbf{w}_*\|_2^2 - 2\eta\mathbf{z}_t^\top(\mathbf{w}_t - \mathbf{w}_*) + \eta^2\|\mathbf{z}_t\|_2^2.$$

Then,

$$\begin{aligned}
&\mathbb{E}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2 \le \mathbb{E}\|\mathbf{w}_t - \mathbf{w}_*\|_2^2 - 2\eta\mathbb{E}[\nabla_1 F(\mathbf{w}_t, \boldsymbol{\nu}_t)^\top(\mathbf{w}_t - \mathbf{w}_*)] \\
&+ \eta^2\mathbb{E}\|\nabla_1 F(\mathbf{w}_t, \boldsymbol{\nu}_t)\|_2^2 + \eta^2\sigma^2.
\end{aligned} \tag{5.71}$$

Summing over $t = 0, \ldots, T-1$ and rearranging it finishes the proof. □

**Lemma 5.22** *Let* $D_{\nu,0}^2 \coloneqq \mathbb{E}\|\boldsymbol{\nu}_0 - \boldsymbol{\nu}_*\|_2^2$ *and* $\gamma_t = \gamma$*, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}(2\mathbb{E}[\nabla_2 F(\mathbf{w}_t, \boldsymbol{\nu}_t)^\top(\boldsymbol{\nu}_t - \boldsymbol{\nu}_*)] - \gamma n\mathbb{E}\|\nabla_2 F(\mathbf{w}_t, \boldsymbol{\nu}_t)\|_2^2) \le \frac{D_{\nu,0}^2}{\gamma BT} + \gamma\sigma_2^2.$$

*Proof.* Let $\mathbb{E}_t$ denote the expectation over the random samples in the $t$-th iteration. Note that $\mathbb{E}_t[\nabla_2\Phi_i(\mathbf{w}_t, \nu_{i,t}; \zeta_{i,t})] = \nabla_2 F_i(\mathbf{w}_t, \nu_{i,t})$ and $\mathbb{E}_t\|\nabla_2\Phi_i(\mathbf{w}_t, \nu_{i,t}; \zeta_{i,t}) - \nabla_2 F_i(\mathbf{w}_t, \nu_{i,t})\|_2^2 \leq \sigma_0^2$ for each $i \in [n]$ (For those $i \notin \mathcal{B}_t$, $\nabla_2\Phi_i(\mathbf{w}_t, \nu_{i,t}; \zeta_{i,t})$ are not explicitly computed). For each $i \in [n]$, we have

$$\begin{aligned}
&\mathbb{E}\|\nu_{i,t+1} - \nu_{i,*}\|_2^2 \\
&= (1 - \frac{B}{n})\mathbb{E}\|\nu_{i,t} - \nu_{i,*}\|_2^2 + \frac{B}{n}\mathbb{E}\|\nu_{i,t} - \gamma\nabla_2\Phi_i(\mathbf{w}_t, \nu_{i,t}; \zeta_{i,t}) - \nu_{i,*}\|_2^2 \\
&\leq \mathbb{E}\|\nu_{i,t} - \nu_{i,*}\|_2^2 - \frac{2\gamma B}{n}\mathbb{E}[\nabla_2 F_i(\mathbf{w}_t, \nu_{i,t})^\top(\nu_{i,t} - \nu_{i,*})] + \frac{\gamma^2 B}{n}\mathbb{E}\|\nabla_2 F_i(\mathbf{w}_t, \nu_{i,t})\|_2^2 \\
&+ \frac{\gamma^2\sigma_2^2 B}{n}.
\end{aligned}$$

Summing over $i \in [n]$ leads to

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\nu}_{t+1} - \boldsymbol{\nu}_*\|_2^2 = \mathbb{E}\|\boldsymbol{\nu}_t - \boldsymbol{\nu}_*\|_2^2 - \frac{2\gamma B}{n}\mathbb{E}\Big[\sum_{i=1}^n \nabla_2 F_i(\mathbf{w}_t, \nu_{i,t})^\top(\nu_{i,t} - \nu_{i,*})\Big] \\
+ \frac{\gamma^2 B}{n}\mathbb{E}\Big[\sum_{i=1}^n \|\nabla_2 F_i(\mathbf{w}_t, \boldsymbol{\nu}_{i,t})\|_2^2\Big] + \gamma^2\sigma_2^2 B. \qquad (5.72)
\end{aligned}$$

Since

$$\begin{aligned}
\nabla_2 F(\mathbf{w}_t, \boldsymbol{\nu}_t)^\top(\boldsymbol{\nu}_t - \boldsymbol{\nu}_*) &= \frac{1}{n}\sum_{i=1}^n \nabla_2 F_i(\mathbf{w}_t, \nu_{i,t})(\nu_{i,t} - \nu_{i,*}) \\
\|\nabla_2 F(\mathbf{w}_t, \boldsymbol{\nu}_t)\|_2^2 &= \frac{1}{n^2}\sum_{i=1}^n \|\nabla_2 F_i(\mathbf{w}_t, \nu_{i,t})\|_2^2,
\end{aligned}$$

plugging these into (5.73) we have

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\nu}_{t+1} - \boldsymbol{\nu}_*\|_2^2 \leq \mathbb{E}\|\boldsymbol{\nu}_t - \boldsymbol{\nu}_*\|_2^2 - 2\gamma B\mathbb{E}\Big[\nabla_2 F(\mathbf{w}_t, \boldsymbol{\nu}_t)^\top(\boldsymbol{\nu}_t - \boldsymbol{\nu}_*)\Big] \\
+ \gamma^2 nB\mathbb{E}\Big[\|\nabla_2 F(\mathbf{w}_t, \boldsymbol{\nu}_t)\|_2^2\Big] + \gamma^2\sigma_2^2 B. \qquad (5.73)
\end{aligned}$$

Summing over $t = 0, \ldots, T-1$ and rearranging it finishes the proof. □

**Theorem 5.11 (Smooth case)** *Suppose Assumption 5.13, 5.14(i) and 5.15 hold. If we set* $\gamma = \min\{\frac{1}{2nL_F}, \frac{\epsilon}{2\sigma_2^2}\}$, $\eta = \min\{\frac{1}{2L_F}, \frac{\epsilon}{2\sigma^2}\}$ *and* $T = \max(\frac{2D_{\mathbf{w},0}^2}{\eta\epsilon}, \frac{2D_{\nu,0}^2}{\gamma B\epsilon})$, *then ASGD guarantees that*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}(F(\mathbf{w}_t, \boldsymbol{\nu}_t) - F(\mathbf{w}_*, \boldsymbol{\nu}_*))\right] \leq \epsilon.$$

*The iteration complexity is*

$$T = O\left(\max\left\{\frac{D_{\mathbf{w},0}^2 L_F}{\epsilon}, \frac{nD_{\nu,0}^2 L_F}{B\epsilon}, \frac{D_{\mathbf{w},0}^2 \sigma_1^2}{\epsilon^2}, \frac{D_{\nu,0}^2 \sigma_2^2}{B\epsilon^2}\right\}\right),$$

*where* $\sigma^2 = \frac{\sigma_1^2}{B} + \frac{\delta^2(n-B)}{B(n-1)}$.

*Proof.* From Lemma 5.21 and Lemma 5.22, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}(2\mathbb{E}\nabla_1 F(\mathbf{w}_t, \nu_t)^\top(\mathbf{w}_t - \mathbf{w}_*) - \eta\mathbb{E}\|\nabla_1 F(\mathbf{w}_t, \nu_t)\|_2^2) \le \frac{D_{\mathbf{w},0}^2}{\eta T} + \eta\sigma^2,$$
$$\frac{1}{T}\sum_{t=0}^{T-1}(2\mathbb{E}\nabla_2 F(\mathbf{w}_t, \nu_t)^\top(\nu_t - \nu_*) - \gamma n\mathbb{E}\|\nabla_2 F(\mathbf{w}_t, \nu_t)\|_2^2) \le \frac{D_{\nu,0}^2}{\gamma BT} + \gamma\sigma_2^2.$$

If $F$ is smooth and $\eta \le \frac{1}{2L_F}$ and $\gamma n \le \frac{1}{2L_F}$,

$$\begin{aligned}\eta\|\nabla_1 F(\mathbf{w}_t, \nu_t)\|_2^2 + \gamma n\|\nabla_2 F(\mathbf{w}_t, \nu_t)\|_2^2 &\le \frac{1}{2L_F}\left(\|\nabla_1 F(\mathbf{w}, \nu)\|_2^2 + \|\nabla_2 F(\mathbf{w}, \nu)\|_2^2\right)\\ &\le F(\mathbf{w}_t, \nu_t) - F(\mathbf{w}_*, \nu_*),\end{aligned}$$

where the last inequality uses the Lemma 1.5(b).

On the other hand, the joint convexity of $F$ implies

$$F(\mathbf{w}_t, \nu_t) - F(\mathbf{w}_*, \nu_*) \le \nabla_1 F(\mathbf{w}_t, \nu_t)^\top(\mathbf{w}_t - \mathbf{w}_*) + \nabla_2 F(\mathbf{w}_t, \nu_t)^\top(\nu_t - \nu_*).$$

Then combining the above inequalities, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}[F(\mathbf{w}_t, \nu_t) - F(\mathbf{w}_*, \nu_*)]\right] \le \frac{D_{\mathbf{w},0}^2}{2\eta T} + \frac{\eta\sigma^2}{2} + \frac{D_{\nu,0}^2}{2\gamma BT} + \frac{\gamma\sigma_2^2}{2}.$$

In order to let the RHS above be less than $\epsilon$, we set $\gamma = \min\{\frac{1}{2nL_F}, \frac{\epsilon}{2\sigma_2^2}\}$ and $\eta = \min\{\frac{1}{2L_F}, \frac{\epsilon}{2\sigma^2}\}$, and $T \ge \max(\frac{2D_{\mathbf{w},0}^2}{\eta\epsilon}, \frac{2D_{\nu,0}^2}{\gamma B\epsilon})$. As a result, the complexity is the in the order of

$$T = O\left(\max\left\{\frac{D_{\mathbf{w},0}^2 L_F}{\epsilon}, \frac{nD_{\nu,0}^2 L_F}{B\epsilon}, \frac{D_{\mathbf{w},0}^2 \sigma^2}{\epsilon^2}, \frac{D_{\nu,0}^2 \sigma_2^2}{B\epsilon^2}\right\}\right).$$

□

**Theorem 5.12 (Non-smooth case)** *Suppose Assumption 5.13, 5.14(ii) and 5.15 hold. If we set* $\gamma = \frac{\epsilon}{2(G_2^2+\sigma_2^2)}$, $\eta = \frac{\epsilon}{2(G_1^2+\sigma^2)}$ *and* $T = \max(\frac{2D_{\mathbf{w},0}^2}{\eta\epsilon}, \frac{2D_{\nu,0}^2}{\gamma B\epsilon})$, *then ASGD guarantees that*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}(F(\mathbf{w}_t, \nu_t) - F(\mathbf{w}_*, \nu_*))\right] \le \epsilon.$$

*The iteration complexity is*

$$T = O\left(\max\left\{\frac{D_{\mathbf{w},0}^2(G_1^2+\sigma^2)}{\epsilon^2}, \frac{D_{\nu,0}^2(G_2^2+\sigma_2^2)}{B\epsilon^2}\right\}\right).$$

We leave the proof as an exercise for the reader.

**Why it matters**

Since $F(\mathbf{w}, \nu)$ is jointly convex in $(\mathbf{w}, \nu)$, the above two theorems imply convergence of the objective with respect to the primary variable $\mathbf{w}$, i.e., $F_1(\mathbf{w}) = \min_\nu F(\mathbf{w}, \nu)$. In particular, if we define the averaged iterate $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{w}_t$, we have

$$\begin{aligned}\mathbb{E}[F_1(\bar{\mathbf{w}}_T) - F_1(\mathbf{w}_*)] &\leq \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}(F_1(\mathbf{w}_t) - F_1(\mathbf{w}_*))\right] \\ &\leq \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}(F(\mathbf{w}_t, \nu_t) - F(\mathbf{w}_*, \nu_*))\right] \leq \epsilon.\end{aligned}$$

#### 5.5.1.2 Non-convex loss

If $s_i(\mathbf{w}, \zeta)$ is non-convex, we consider two different cases: (1) smooth case and (2) non-smooth weakly convex case. If $F(\mathbf{w}, \nu)$ is smooth in terms of $\mathbf{w}$, $\nu$ and is strongly convex in terms of $\nu$ (e.g, compositional entropic risk or COCE with $\chi^2$ divergence for $\phi(\cdot)$), we can follow the analysis in Chapter 4 [Section 4.5] to design an algorithm and an analysis to prove the convergence for finding an $\epsilon$-stationary point of $F_1(\mathbf{w}) = \min_\nu F(\mathbf{w}; \nu)$. We leave this as an exercise for the reader.

Below, we analyze the convergence of ASGD for non-smooth weakly convex losses. We also assume $\phi^*$ is non-smooth such that it covers the CCVaR minimization.

**Assumption 5.16.** *Suppose the following conditions hold:*

- $s_i(\mathbf{w}; \zeta)$ *is* $\rho_0$*-weakly convex with respect to* $\mathbf{w}$*, and* $\mathbb{E}_\zeta[\|\partial s_i(\mathbf{w}; \zeta)\|_2^2] \leq G_\ell^2$*;*
- *Assume* $|\frac{\partial \phi^*(q)}{\partial q}| \leq G_0$ *for any* $q = s_i(\mathbf{w}, \zeta) - \nu_i$*.*

**Lemma 5.23** $F(\mathbf{w}, \nu)$ *is* $\rho$*-weakly convex with respect to* $(\mathbf{w}, \nu)$*, where* $\rho = \rho_0 G_0$*.*

*Proof.* We first prove that $\phi^*(s_i(\mathbf{w}; \zeta) - \nu_i)$ is weakly convex in terms of $(\mathbf{w}, \nu_i)$, i.e. there exists $\rho > 0$ such that $\phi^*(s_i(\mathbf{w}; \zeta) - \nu_i) + \frac{\rho}{2}\|\mathbf{w}\|_2^2 + \frac{\rho}{2}\nu_i^2$ is jointly convex in terms of $\mathbf{w}, \nu_i$.

Since $s_i(\mathbf{w}; \zeta)$ is $\rho_0$-weakly convex, we have that $q(\mathbf{w}, \nu_i, \zeta) = s_i(\mathbf{w}, \zeta) - \nu_i$ is $\rho_0$-weakly convex in terms of $\mathbf{v}_i = (\mathbf{w}, \nu_i)$:

$$q(\mathbf{v}_i, \zeta) \geq q(\mathbf{v}_i', \zeta) + \partial q(\mathbf{v}_i', \zeta)^\top(\mathbf{v}_i - \mathbf{v}_i') - \frac{\rho_0}{2}\|\mathbf{v}_i' - \mathbf{v}_i\|_2^2, \forall \mathbf{v}_i, \mathbf{v}_i'.$$

For any $\zeta$, we abbreviate $q(\mathbf{v}_i;\zeta)$ as $q(\mathbf{v}_i)$. Since $\phi^*$ is convex and monotonically non-decreasing, for any $\omega \in \partial\phi^*(q(\mathbf{v}_i')) \in [0, G_0]$ we have

$$\begin{aligned}
&\phi^*(q(\mathbf{v}_i)) \geq \phi^*(q(\mathbf{v}_i')) + \omega(q(\mathbf{v}_i) - q(\mathbf{v}_i')) \\
&\geq \phi^*(q(\mathbf{v}_i')) + \omega(\partial q(\mathbf{v}_i')^\top(\mathbf{v}_i - \mathbf{v}_i') - \frac{\rho_0}{2}\|\mathbf{v}_i - \mathbf{v}_i'\|_2^2) \\
&\geq \phi^*(q(\mathbf{v}_i')) + \partial\phi^*(q(\mathbf{v}_i'))^\top(\mathbf{v}_i - \mathbf{v}_i') - \frac{G_0\rho_0}{2}\|\mathbf{v}_i - \mathbf{v}_i'\|_2^2.
\end{aligned}$$

The above inequality implies that $\phi^*(s_i(\mathbf{w};\zeta) - \nu_i)$ is $\rho = G_0\rho_0$-weakly convex in terms of $(\mathbf{w}, \nu_i)$, i.e., $\mathbb{E}_\zeta\phi^*(s_i(\mathbf{w};\zeta) - \nu_i) + \frac{\rho}{2}(\|\mathbf{w}\|_2^2 + |\nu_i|^2)$ is convex. As a result $F(\mathbf{w}, \boldsymbol{\nu}) + \frac{\rho}{2}\|\mathbf{w}\|_2^2 + \frac{\rho}{2}\|\boldsymbol{\nu}\|_2^2$ is jointly convex in terms of $(\mathbf{w}, \boldsymbol{\nu})$. □

Similar to the SGD for weakly convex objectives in Chapter 3[Section 3.1.4], we use the Moreau envelope of $F(\mathbf{w};\boldsymbol{\nu})$. In particular, let $\mathbf{v} = (\mathbf{w}^\top, \boldsymbol{\nu}^\top)^\top$ and consider some $\bar{\rho} > \rho$, we define:

$$F_{1/\bar{\rho}}(\mathbf{v}) = \min_{\mathbf{u}} F(\mathbf{u}) + \frac{\bar{\rho}}{2}\|\mathbf{u} - \mathbf{v}\|_2^2, \tag{5.74}$$

$$\text{prox}_{F/\bar{\rho}}(\mathbf{v}) := \arg\min_{\mathbf{u}} F(\mathbf{u}) + \frac{\bar{\rho}}{2}\|\mathbf{u} - \mathbf{v}\|_2^2. \tag{5.75}$$

Convergence Analysis

**Lemma 5.24** *Under Assumption 5.16, we have*

$$\mathbb{E}_t[\|\mathbf{z}_t\|_2^2] \leq G_1^2, \quad |\partial_2 F_i(\mathbf{w}, \nu_i)|^2 \leq G_2^2,$$

*where* $G_1^2 = G_0^2 G_\ell^2$, *and* $G_2^2 = (1 + G_0)^2$.

*Proof.* For the first part,

$$\mathbb{E}_t[\|\mathbf{z}_t\|_2^2] = \mathbb{E}_t\left[\left\|\frac{1}{B}\sum_{i\in\mathcal{B}_t}\partial_1\Phi_i(\mathbf{w}_t, \nu_{i,t}; \zeta_{i,t})\right\|_2^2\right] \leq G_0^2 G_\ell^2.$$

For the second part,

$$|\partial_2 F_i(\mathbf{w}, \nu_i)|^2 = \left|\mathbb{E}_\zeta\left[-\frac{\partial\phi^*(q(\mathbf{w}, \nu_i; \zeta))}{\partial q} + 1\right]\right|^2 \leq (1 + G_0)^2.$$

□

**Lemma 5.25** *Under Assumption (5.16), let* $\mathbf{v}_t = (\mathbf{w}_t^\top, \boldsymbol{\nu}_t^\top)^\top$, *for one iteration of ASGD, we have*

$$\mathbb{E}_t[F_{1/\bar{\rho}}(\mathbf{v}_{t+1})] \le F_{1/\bar{\rho}}(\mathbf{v}_t) + \bar{\rho}\eta_t(F(\bar{\mathbf{v}}_t) - F(\mathbf{v}_t) + \frac{\rho}{2}\|\mathbf{v}_t - \bar{\mathbf{v}}_t\|_2^2) + \frac{\bar{\rho}\eta_t^2(G_1^2 + G_2^2/B)}{2},$$

*where* $\bar{\mathbf{v}}_t = prox_{F/\bar{\rho}}(\mathbf{v}_t)$.

*Proof.* Let $\mathbb{E}_t$ denote the expectation over the random samples at the $t$-th iteration conditioned on that in all previous iterations.

$$\begin{aligned}
&\mathbb{E}_t[F_{1/\bar{\rho}}(\mathbf{v}_{t+1})] \le \mathbb{E}_t\left[F(\bar{\mathbf{v}}_t) + \frac{\bar{\rho}}{2}\|\mathbf{v}_{t+1} - \bar{\mathbf{v}}_t\|_2^2\right] \\
&\le F(\bar{\mathbf{v}}_t) + \frac{\bar{\rho}}{2}\mathbb{E}_t[\|\mathbf{w}_t - \eta_t\mathbf{z}_t - \bar{\mathbf{w}}_t\|_2^2 + \|\boldsymbol{\nu}_{t+1} - \bar{\boldsymbol{\nu}}_t\|_2^2] \\
&\le F(\bar{\mathbf{v}}_t) + \frac{\bar{\rho}}{2}\mathbb{E}_t[\|\mathbf{w}_t - \eta_t\mathbf{z}_t - \bar{\mathbf{w}}_t\|_2^2] + \frac{\bar{\rho}}{2}\mathbb{E}_t[\|\boldsymbol{\nu}_{t+1} - \bar{\boldsymbol{\nu}}_t\|_2^2] \\
&\le F(\bar{\mathbf{v}}_t) + \frac{\bar{\rho}}{2}\|\mathbf{w}_t - \bar{\mathbf{w}}_t\|_2^2 + \bar{\rho}\eta_t\mathbb{E}_t[(\bar{\mathbf{w}}_t - \mathbf{w}_t)^\top\partial_1 F(\mathbf{w}_t, \boldsymbol{\nu}_t)] + \frac{\bar{\rho}\eta_t^2 G_1^2}{2} \\
&+ \frac{\bar{\rho}}{2}\mathbb{E}_t[\|\boldsymbol{\nu}_{t+1} - \bar{\boldsymbol{\nu}}_t\|_2^2]
\end{aligned}$$

where the last step uses $\mathbb{E}_t[\mathbf{z}_t] = \partial_1 F(\mathbf{w}_t, \boldsymbol{\nu}_t)$ and $\mathbb{E}[\|\mathbf{z}_t\|_2^2] \le G_1^2$.

Similar to (5.73), we can prove that

$$\mathbb{E}_t\|\boldsymbol{\nu}_{t+1} - \bar{\boldsymbol{\nu}}_t\|_2^2 = \|\boldsymbol{\nu}_t - \bar{\boldsymbol{\nu}}_t\|_2^2 - 2\gamma_t B\partial_2 F(\mathbf{w}_t, \boldsymbol{\nu}_t)^\top(\boldsymbol{\nu}_t - \bar{\boldsymbol{\nu}}_t) + \gamma_t^2 G_2^2 B.$$

Let $\gamma_t B = \eta_t$, combining the above we have

$$\begin{aligned}
&\mathbb{E}_t[F_{1/\bar{\rho}}(\mathbf{v}_{t+1})] \\
&\le F(\bar{\mathbf{v}}_t) + \frac{\bar{\rho}}{2}\|\mathbf{v}_t - \bar{\mathbf{v}}_t\|_2^2 + \bar{\rho}\eta_t\mathbb{E}_t[(\bar{\mathbf{v}}_t - \mathbf{v}_t)^\top\partial F(\mathbf{v}_t)] + \frac{\bar{\rho}\eta_t^2(G_1^2 + G_2^2/B)}{2} \\
&\le F(\bar{\mathbf{v}}_t) + \frac{\bar{\rho}}{2}\|\mathbf{v}_t - \bar{\mathbf{v}}_t\|_2^2 + \bar{\rho}\eta_t\mathbb{E}_t[(\bar{\mathbf{v}}_t - \mathbf{v}_t)^\top\partial F(\mathbf{v}_t)] + \frac{\bar{\rho}\eta_t^2(G_1^2 + G_2^2/B)}{2} \\
&\le F_{1/\bar{\rho}}(\mathbf{v}_t) + \bar{\rho}\eta_t(F(\bar{\mathbf{v}}_t) - F(\mathbf{v}_t) + \frac{\rho}{2}\|\mathbf{v}_t - \bar{\mathbf{v}}_t\|_2^2) + \frac{\bar{\rho}\eta_t^2(G_1^2 + G_2^2/B)}{2}.
\end{aligned}$$

where the last step uses the definition of $F_{1/\bar{\rho}}(\mathbf{v}_t)$ and the $\rho$-weak convexity of $F$. Rearranging this inequality finishes the proof.

□

**Theorem 5.13** *Suppose Assumption (5.16) holds and* $F_* = \inf F(\mathbf{w}, \boldsymbol{\nu}) \ge \infty$, *by setting* $\bar{\rho} = 2\rho$, $\eta = \epsilon^2/(2\bar{\rho}(G_1^2 + G_2^2/B))$, $\gamma = \eta/B$ *and* $T \ge \frac{4(F(\mathbf{w}_0, \boldsymbol{\nu}_0) - F_*)}{\epsilon^2\eta}$, *ASGD guarantees that*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla F_{1/\bar{\rho}}(\mathbf{v}_t)\|_2^2\right] \le \epsilon^2$$

*with a complexity of* $T = O\left(\frac{\rho(G_1^2+G_2^2/B)}{\epsilon^4}\right)$.

*Proof.* Since $F(\mathbf{v}) + \frac{\bar{\rho}}{2}\|\mathbf{v} - \mathbf{v}_t\|_2^2$ is $(\bar{\rho} - \rho)$-strongly convex and have a minimum solution at $\bar{\mathbf{v}}_t$, then we have

$$\begin{aligned}
&F(\mathbf{v}_t) - F(\bar{\mathbf{v}}_t) - \frac{\rho}{2}\|\mathbf{v}_t - \bar{\mathbf{v}}_t\|_2^2 \\
&= (F(\mathbf{v}_t) + \frac{\bar{\rho}}{2}\|\mathbf{v}_t - \mathbf{v}_t\|_2^2) - (F(\bar{\mathbf{v}}_t) + \frac{\bar{\rho}}{2}\|\bar{\mathbf{v}}_t - \mathbf{v}_t\|_2^2) + (\frac{\bar{\rho}}{2} - \frac{\rho}{2})\|\mathbf{v}_t - \bar{\mathbf{v}}_t\|_2^2 \\
&\geq \frac{(\bar{\rho} - \rho)}{2}\|\mathbf{v}_t - \bar{\mathbf{v}}_t\|_2^2 + \frac{(\bar{\rho} - \rho)}{2}\|\mathbf{v}_t - \bar{\mathbf{v}}_t\|_2^2 = (\bar{\rho} - \rho)\|\mathbf{v}_t - \bar{\mathbf{v}}_t\|_2^2 \\
&= \frac{\bar{\rho} - \rho}{\bar{\rho}^2}\|\nabla F_{1/\bar{\rho}}(\mathbf{v}_t)\|_2^2.
\end{aligned}$$

Combining this result with that in Lemma 5.25 and noting that $\bar{\rho} = 2\rho, \eta_t = \eta$, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F_{1/\bar{\rho}}(\mathbf{v}_t)\|_2^2\right] &\leq \frac{2(F_{1/\bar{\rho}}(\mathbf{v}_0) - F_*)}{\eta T} + \bar{\rho}\eta(G_1^2 + G_2^2/B) \\
&\leq \frac{2(F(\mathbf{v}_0) - F_*)}{\eta T} + \bar{\rho}\eta(G_1^2 + G_2^2/B)
\end{aligned}$$

By setting $\eta = \epsilon^2/(2\bar{\rho}(G_1^2+G_2^2/B))$ and $T \geq \frac{4(F(\mathbf{v}_0)-F_*)}{\epsilon^2\eta}$, we have $\mathbb{E}[\|\nabla F_{1/\bar{\rho}}(\mathbf{v}_\tau)\|_2^2] \leq \epsilon^2$ for a randomly selected $\tau \in \{0, \ldots, T-1\}$. □

### 5.5.2 A Geometry-aware Algorithm for Entropic Risk

Although last section presents a general algorithm for solving COCE risk minimization, it may exhibits numerical instability issue and slow convergence when solving compositional entropic risk minimization:

$$\begin{aligned}
&\min_{\mathbf{w}}\min_{\nu}\left[F(\mathbf{w}, \nu) = \frac{1}{n}\sum_{i=1}^{n}\{\mathbb{E}_\zeta \exp(s_i(\mathbf{w}; \zeta) - \nu_i) - 1 + \nu_i\}\right] \\
&= \min_{\mathbf{w}}\frac{1}{n}\sum_{i=1}^{n}\log\left(\mathbb{E}_\zeta \exp(s_i(\mathbf{w}; \zeta))\right).
\end{aligned}$$

The numerical instability issue is caused by dealing with exponential functions, e.g., $\exp(s_i(\mathbf{w}; \zeta) - \nu_i)$, in calculation of stochastic gradients of $\nu_i$. The slow convergence arises because the standard SGD update for $\nu_i$ fails to exploit the geometric structure of the problem.

#### 5.5.2.1 Stochastic Optimization of Log-E-Exp

We first consider a simplified problem where there is only one component $n = 1$, i.e.,

$$\min_{\mathbf{w}} F_1(\mathbf{w}) := \log\left(\mathbb{E}_\zeta \exp(s(\mathbf{w};\zeta))\right). \tag{5.76}$$

The KL-regularized DRO problem (2.14) is a special case. It is also known as log-E-Exp, a more general form of the log-Sum-Exp function, where the middle "E" denotes an expectation and highlights the associated computational challenges.

**Application of SCGD**

At the beginning of Section 4.1, we treat this problem as a special case of stochastic compositional optimization (SCO), where the outer function is $f(\cdot) = \log(\cdot)$ and the inner function is $g(\mathbf{w}) = \mathbb{E}_\zeta[\exp(s(\mathbf{w};\zeta))]$. Let us first apply the SCGD algorithm. The key updates are presented below:

$$\begin{aligned} u_t &= (1-\gamma_t)u_{t-1} + \gamma_t \exp(s(\mathbf{w}_t;\zeta_t)), \\ \mathbf{z}_t &= \frac{1}{u_t}\exp(s(\mathbf{w}_t;\zeta_t'))\nabla s(\mathbf{w}_t;\zeta_t'), \\ \mathbf{w}_{t+1} &= \mathbf{w}_t - \eta_t \mathbf{z}_t, \end{aligned} \tag{5.77}$$

where $u_t$ is an estimator of the inner function value $g(\mathbf{w}_t)$ and $\mathbf{z}_t = \nabla f(u_t)\nabla g(\mathbf{w}_t;\zeta_t')$ is a gradient estimator of $\mathbf{w}_t$.

From a practitioner's perspective, the algorithm can be readily implemented and applied to real applications. However, from a theoretical perspective, several open problems remain. In particular: (1) Can we establish an $O(1/\epsilon^2)$ convergence rate for this algorithm to find an $\epsilon$-optimal solution when $s(\mathbf{w};\zeta)$ is convex? (2) If yes, what are the practical advantages of this algorithm compared with the ASGD method presented in the previous section?

Wait! Shouldn't we established the convergence rate of SCGD in Chapter 4? It is true that we presented a convergence analysis of the above algorithm for non-convex problems under proper conditions, however, it remains an open problem to establish the complexity of $O(1/\epsilon^2)$ for finding an $\epsilon$-optimal solution under the convexity of $s(\mathbf{w};\zeta)$. A naive analysis of SCGD for convex problems yields a complexity of $O(1/\epsilon^4)$ (see Wang et al. (2017a)).

**A Novel Algorithm**

To address these open questions, we present a novel algorithm based on the min-min reformulation of log-E-exp, i.e.,

$$\min_{\mathbf{w}} \min_{\nu} F(\mathbf{w}, \nu) := \mathbb{E}_{\zeta} \exp(s(\mathbf{w}; \zeta) - \nu) + \nu. \tag{5.78}$$

where we ignored the constant $-1$ in the objective. As proved in Lemma 5.20, $F(\mathbf{w}; \nu)$ is jointly convex in terms of $\mathbf{w}, \nu$ when $s(\mathbf{w}; \zeta)$ is convex.

*Motivation*

The key novelty of our design is a **geometry-aware algorithm** for solving the equivalent min-min optimization (5.78). Let us first discuss the motivation. One challenge for solving the min-min optimization problem is that the objective function $F(\mathbf{w}, \nu)$ could have exponentially large smoothness constant in terms of $\nu$. We will formally analyze this phenomenon in next section. Hence, a vanilla gradient method that uses the firs-order approximation of $F$ will inevitably impacted by the large smoothness parameter.

To mitigate the adverse effects of a large smoothness parameter with respect to $\nu$, we resort to the classical approach of employing a proximal mapping. Proximal mappings have been widely used to handle a non-smooth function in composite objectives consisting of a smooth loss and a non-smooth regularizer. This approach enables optimization algorithms to retain the favorable convergence properties of smooth optimization and often leads to faster convergence despite the presence of non-smooth terms. Analogously, even when a function is smooth but characterized by a very large smoothness parameter, applying its proximal mapping can effectively alleviate the negative impact of this large smoothness constant.

However, there is an important distinction from classical proximal methods, which typically rely on direct access to the function of interest for computing the proximal mapping. In our setting, we cannot directly apply the proximal mapping of $F(\mathbf{w}, \nu)$. Instead, we only have access to a stochastic estimator

$$\Phi(\mathbf{w}, \nu; \zeta) = e^{s(\mathbf{w};\zeta)-\nu} + \nu,$$

defined for a random sample $\zeta$. As a result, it becomes necessary to explicitly account for the noise introduced by this stochastic approximation.

*Algorithm*

To account for the stochastic noise, we introduce a Bregman divergence $D_{\varphi}(\cdot, \cdot)$ and update $\nu_t$ according to the following scheme:

$$\nu_t = \arg\min_{\nu} \Phi(\mathbf{w}_t, \nu; \zeta_t) + \frac{1}{\alpha_t} D_{\varphi}(\nu, \nu_{t-1}), \tag{5.79}$$

where $\zeta_t \sim \mathbb{P}$ is a random sample and $\alpha_t > 0$ is a step size parameter. We refer to this step as **stochastic proximal mirror descent (SPMD)** update. To respect the geometry of the stochastic objective $\Phi(\mathbf{w}_t, \nu; \zeta_t)$, we construct a tailored Bregman divergence induced by the function $\varphi(\nu) = e^{-\nu}$, namely,

**Algorithm 21** The SCENT Algorithm for solving Log-E-Exp (5.76)

1: Initialize $\mathbf{w}_1$, $\nu_0$, step sizes $\eta_t$ and $\alpha_t$, $\varphi(\nu) = e^{-\nu}$.
2: **for** $t = 1 \ldots, T-1$ **do**
3: Sample $\zeta_t, \zeta_t'$
4: Update $\nu_t = \arg\min_\nu \exp(s(\mathbf{w}_t; \zeta_t) - \nu) + \nu + \frac{1}{\alpha_t} D_\varphi(\nu, \nu_{t-1})$
5: Compute $\mathbf{z}_t = \exp(s(\mathbf{w}_t; \zeta_t') - \nu_t) \nabla s(\mathbf{w}_t; \zeta_t')$
6: Compute $\mathbf{v}_t = (1 - \beta_t)\mathbf{v}_{t-1} + \beta_t \mathbf{z}_t$
7: Update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{v}_t$
8: **end for**

$$D_\varphi(\nu, \nu_{t-1}) = e^{-\nu} - e^{-\nu_{t-1}} + e^{-\nu_{t-1}}(\nu - \nu_{t-1}). \tag{5.80}$$

Once we have $\nu_t$, we compute a vanilla gradient estimator by

$$\mathbf{z}_t = \exp(s(\mathbf{w}_t; \zeta_t') - \nu_t) \nabla s(\mathbf{w}_t; \zeta_t'). \tag{5.81}$$

If the problem is non-convex, we compute a moving-average estimator $\mathbf{v}_t = (1 - \beta_t)\mathbf{v}_{t-1} + \beta_t \mathbf{z}_t$ and then update the model parameter $\mathbf{w}_{t+1}$. We present the full steps in Algorithm 21, which is referred to SCENT.

**SCGD is just a special case of SCENT**

To see the connection with SCGD, we present the following lemma.

**Lemma 5.26** *The update of $\nu_t$ defined by (5.79) can be computed by*

$$e^{\nu_t} = \frac{1}{1 + \alpha_t e^{\nu_{t-1}}} e^{\nu_{t-1}} + \frac{\alpha_t e^{\nu_{t-1}}}{1 + \alpha_t e^{\nu_{t-1}}} \exp(s(\mathbf{w}_t; \zeta_t)). \tag{5.82}$$

*If $y_t = e^{-\nu_t}$, we have*

$$y_t = \frac{y_{t-1} + \alpha_t}{1 + \alpha_t e^{s(\mathbf{w}_t; \zeta_t)}}.$$

*Proof.* We compute the gradient of the problem (5.79) and set it to zero for computing $\nu_t$, i.e.,

$$-\exp(s(\mathbf{w}_t; \zeta_t) - \nu_t) + 1 + \frac{1}{\alpha_t}(-e^{-\nu_t} + e^{-\nu_{t-1}}) = 0.$$

Solving this equation finishes the proof. □

If we define $u_t = e^{\nu_t}$ and $\gamma_t' = \frac{\alpha_t e^{\nu_{t-1}}}{1 + \alpha_t e^{\nu_{t-1}}}$, then the updates of SCENT ($\beta_t = 1$) are equivalent to

$$
\begin{aligned}
u_t &= (1-\gamma_t')u_{t-1} + \gamma_t' \exp(s(\mathbf{w}_t;\zeta_t)) \\
\mathbf{z}_t &= \frac{1}{u_t}\exp(s(\mathbf{w}_t;\zeta_t'))\nabla s(\mathbf{w}_t;\zeta_t'), \\
\mathbf{w}_{t+1} &= \mathbf{w}_t - \eta_t \mathbf{z}_t.
\end{aligned} \tag{5.83}
$$

Comparing this update with that of SCGD (5.77), the key difference lies in the choice of the moving-average parameter: SCENT adopts an adaptive parameter $\gamma_t' = \frac{\alpha_t e^{\nu_{t-1}}}{1+\alpha_t e^{\nu_{t-1}}}$, whereas SCGD uses a non-adaptive $\gamma_t$. If we set $\alpha_t = \frac{\gamma_t}{1-\gamma_t} e^{-\nu_{t-1}}$, then the updates of SCENT reduce to that of SCGD.

### Convergence analysis for convex problems

Since $\mathbf{z}_t = \nabla_{\mathbf{w}} \exp(s(\mathbf{w}_t;\zeta_t') - \nu_t)$, we have

$$
\mathbb{E}_{\zeta_t'}[\mathbf{z}_t] = \nabla_{\mathbf{w}} \mathbb{E}_{\zeta_t'}[\exp(s(\mathbf{w}_t;\zeta_t') - \nu_t)] = \nabla_1 F(\mathbf{w}_t, \nu_t).
$$

Let $\mathbf{w}_*, \nu_*$ be the optimal solution:

$$
(\mathbf{w}_*, \nu_*) = \arg\min_{\mathbf{w},\nu} F(\mathbf{w}, \nu).
$$

It is straightforward to derive $\nu_* = \log[\mathbb{E}\exp(s(\mathbf{w}_*;\zeta))]$.

**Assumption 5.17.** *Assume that the following conditions hold:*

*(i) $s(\mathbf{w};\zeta)$ is convex;*
*(ii) the loss function is bounded such that $s(\mathbf{w};\zeta) \in [c_0, c_1], \forall \mathbf{w}, \zeta$.*
*(iii) there exists $G$ such that $\mathbb{E}_\zeta \|\nabla s(\mathbf{w}_t, \zeta)\|_2^2] \le G^2, \forall t$.*

**Critical:** To relax the second assumption, we can assume that $\mathbf{w}$ is restricted to a bounded domain $\mathcal{W}$ and $s(\mathbf{w};\zeta)$ is regular. In practice, we always enforce the boundness of $\mathbf{w}_t$ through either projection onto $\mathcal{W}$ or using a regularizer $r(\mathbf{w})$. The update of $\mathbf{w}_{t+1}$ can be modified as the SPGD update:

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

The analysis can be performed similarly.

**Lemma 5.27** *Under Assumption 5.17(ii), $\nu_* \in [c_0, c_1]$ and if $\nu_0 \in [c_0, c_1]$ then $\nu_t \in [c_0, c_1], \forall t$.*

*Proof.* $\nu_* \in [c_0, c_1]$ can be seen from $\nu_* = \log[\mathbb{E}\exp(s(\mathbf{w}_*;\zeta))]$. The second result can be easily seen from the update of $e^{\nu_t}$ as in (5.82) by induction. □

For the ease of analysis, we define two quantities to capture the variance terms caused by using stochastic estimators.

$$\sigma_t^2 := \mathbb{E}_{\zeta'}\|\exp(s(\mathbf{w}_t;\zeta_t') - \nu_t)\nabla s(\mathbf{w}_t;\zeta_t')\|_2^2],$$
$$\delta_t^2 := \mathbb{E}_{\zeta_t}[e^{-\nu_{t-1}}|e^{s(\mathbf{w}_t;\zeta_t)} - \mathbb{E}_{\zeta}[e^{s(\mathbf{w}_t;\zeta)}]|^2].$$

Under Assumption 5.17 (ii) and (iii), $\sigma_t, \delta_t$ are bounded because $e^{\nu_t}, e^{\nu_{t-1}}$ and $e^{s(\mathbf{w}_t;\zeta_t)}$ is upper and lower bounded.

**Critical:** These two quantities are related to the variance of stochastic estimators in terms of $\mathbf{w}_t$ and $\nu_t$, respectively. Both quantities have a normalization term $e^{-\nu_t}$ or $e^{-\nu_{t-1}}$.

**Lemma 5.28** *Under Assumption 5.17 and $\beta_t = 1$, we have*

$$\mathbb{E}[\eta_t \nabla_1 F(\mathbf{w}_t, \nu_t)^\top (\mathbf{w}_t - \mathbf{w}_*)] \leq \mathbb{E}\left[\frac{1}{2}\|\mathbf{w}_t - \mathbf{w}_*\|_2^2 - \frac{1}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2\right] + \frac{\eta_t^2\sigma_t^2}{2}.$$

*Proof.* The proof is a simple application of Lemma 3.3. □

If the SPGD update is used, we can use Lemma 3.6 giving us

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

Then,

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

Taking expectation on both sides, we have

$$\mathbb{E}[\eta_t \nabla_1 F(\mathbf{w}_t, \nu_t)^\top (\mathbf{w}_t - \mathbf{w}_*)] + \eta_t(r(\mathbf{w}_t) - r(\mathbf{w}_*))$$
$$\leq \mathbb{E}\left[\left(\eta_t r(\mathbf{w}_t) + \frac{1}{2}\|\mathbf{w}_t - \mathbf{w}_*\|_2^2\right) - \left(\eta_t r(\mathbf{w}_{t+1}) + \frac{1}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2\right)\right] + \frac{\eta_t^2\sigma_t^2}{2}.$$

If $\eta_{t+1} \leq \eta_t$ and $r(\mathbf{w}) \geq 0$, then $\eta_t r(\mathbf{w}_{t+1}) \leq \eta_{t+1} r(\mathbf{w}_{t+1})$, then the terms in the square bracket will form a telescoping series over $t = 1, \ldots, T$. As a result, the following analysis will proceed similarly.

**Lemma 5.29** *Under Assumption 5.17 (ii), we have*

$$\alpha_t \nabla_2 \Phi(\mathbf{w}_t, \nu_t; \zeta_t)^\top (\nu_t - \nu_*) \le D_\varphi(\nu_*, \nu_{t-1}) - D_\varphi(\nu_*, \nu_t) - D_\varphi(\nu_t, \nu_{t-1}).$$

*Proof.* Recall the definition

$$\begin{aligned}
&\Phi(\mathbf{w}_t, \nu; \zeta_t) = \exp(s(\mathbf{w}_t; \zeta_t) - \nu) + \nu \\
&\varphi(\nu) = e^{-\nu}, \quad D_\varphi(a, b) = \varphi(a) - \varphi(b) - \langle \nabla\varphi(b), a - b\rangle,
\end{aligned}$$

and the update of $\nu_t$:

$$\nu_t = \arg\min_{\nu} \alpha_t \Phi(\mathbf{w}_t, \nu; \zeta_t) + D_\varphi(\nu, \nu_{t-1}).$$

The first-order optimality gives

$$\alpha_t \nabla_2 \Phi(\mathbf{w}_t, \nu_t; \zeta_t) + \nabla\varphi(\nu_t) - \nabla\varphi(\nu_{t-1}) = 0.$$

Taking inner product with $(\nu_t - \nu_*)$ and rearranging gives

$$\begin{aligned}
\alpha_t \langle \nabla_2 \Phi(\mathbf{w}_t, \nu_t; \zeta_t), \nu_t - \nu_* \rangle &= \langle \nabla\varphi(\nu_{t-1}) - \nabla\varphi(\nu_t), \nu_t - \nu_* \rangle \\
&= D_\varphi(\nu_*, \nu_{t-1}) - D_\varphi(\nu_*, \nu_t) - D_\varphi(\nu_t, \nu_{t-1})
\end{aligned}$$

where the last equality holds by three-point identity as in Lemma 3.9. □

**Critical:** To proceed the analysis, we need to bound $\mathbb{E}[\alpha_t \nabla_2 F(\mathbf{w}_t, \nu_t)^\top (\nu_t - \nu_*)]$. In light of the above lemma, we will bound the following difference in expectation:

$$\mathbb{E}[(\nabla_2 F(\mathbf{w}_t, \nu_t)^\top (\nu_t - \nu_*) - \nabla_2 \Phi(\mathbf{w}_t, \nu_t; \zeta_t)^\top (\nu_t - \nu_*)].$$

The challenge lies at $\nu_t$ depends on $\zeta_t$, making the above expectation not equal to zero.

**Lemma 5.30** *Assume $\alpha_t \le \rho e^{-\nu_{t-1}}$ for any constant $\rho > 0$, then we have*

$$|\mathbb{E}[(\nabla_2 F(\mathbf{w}_t, \nu_t)^\top (\nu_t - \nu_*) - \nabla_2 \Phi(\mathbf{w}_t, \nu_t; \zeta_t)^\top (\nu_t - \nu_*)]| \le \alpha_t \delta_t^2 C. \tag{5.84}$$

*where $C = (1 + \rho)(1 + c_1 - c_0)$.*

*Proof.* In the following proof, we let $\mathcal{F}_{t-1}$ denote the filtration (the "information available") up to time $t - 1$.

Let us define $z_t = e^{s(\mathbf{w}_t; \zeta_t)}$, $m_t = \mathbb{E}_\zeta[e^{s(\mathbf{w}_t; \zeta)} | \mathcal{F}_{t-1}]$, and $y_t = e^{-\nu_t}$. Let $z$ and $z'$ two independent variables so that $\mathbb{E}[z|\mathcal{F}_{t-1}] = \mathbb{E}[z'|\mathcal{F}_{t-1}] = m_t$ . Since $\nu_t$ depends on $z_t$, let us define random functions:

$$y_t(z) = \frac{y_{t-1} + \alpha_t}{\alpha_t z + 1}, \quad \nu_t(z) = -\log y_t(z)$$
$$h_t(z) = e^{-\nu_t(z)}\big(\nu_t(z) - \nu_*\big) = y_t(z)\big(\nu_t(z) - \nu_*\big).$$

According to the update of $\nu_t$, we have $y_t = y_t(z_t), \nu_t = \nu_t(z)$. For the target, we have

$$\begin{aligned}
&\mathbb{E}[(\nabla_2\Phi(\mathbf{w}_t, \nu_t; \zeta_t) - \nabla_2 F(\mathbf{w}_t, \nu_t))^\top (\nu_t - \nu_*) \mid \mathcal{F}_{t-1}] \\
&= \mathbb{E}[\mathbb{E}_\zeta[e^{s(\mathbf{w}_t;\zeta)}] - e^{s(\mathbf{w}_t;\zeta_t)})e^{-\nu_t}\big(\nu_t - \nu_*\big) \mid \mathcal{F}_{t-1}] \\
&= \mathbb{E}[(m_t - z_t)h_t(z_t) \mid \mathcal{F}_{t-1}] = \mathbb{E}_z[(m_t - z)h_t(z)|\mathcal{F}_{t-1}].
\end{aligned} \tag{5.85}$$

Since $z'$ is an i.i.d. copy of $z$ and independent of $z$ given $\mathcal{F}_{t-1}$,

$$m_t = \mathbb{E}[z \mid \mathcal{F}_{t-1}] = \mathbb{E}[z' \mid \mathcal{F}_{t-1}].$$

Using the conditional independence,

$$\mathbb{E}\big[(m_t - z)h_t(z) \mid \mathcal{F}_{t-1}\big] = \mathbb{E}\big[(z' - z)h_t(z) \mid \mathcal{F}_{t-1}\big].$$

By exchangeability of $(z, z')$ conditional on $\mathcal{F}_{t-1}$,

$$\mathbb{E}\big[(z' - z)h_t(z') \mid \mathcal{F}_{t-1}\big] = -\mathbb{E}\big[(z' - z)h_t(z) \mid \mathcal{F}_{t-1}\big].$$

Averaging the last two displays gives the standard symmetrization:

$$\mathbb{E}\big[(m_t - z)h_t(z) \mid \mathcal{F}_{t-1}\big] = \frac{1}{2}\mathbb{E}\big[(z' - z)\big(h_t(z) - h_t(z')\big) \mid \mathcal{F}_{t-1}\big]. \tag{5.86}$$

Next, we show that $h(z)$ is Lipschitz continuous. By definition,

$$y_t(z) = \frac{y_{t-1} + \alpha_t}{\alpha_t z + 1}, \quad h_t(z) = y_t(z)\big(\nu_t(z) - \nu_*\big).$$

Differentiate with respect to $z$:

$$\frac{dy_t(z)}{dz} = (y_{t-1} + \alpha_t)\,\frac{d}{dz}\big((\alpha_t z + 1)^{-1}\big) = -\frac{\alpha_t(y_{t-1} + \alpha_t)}{(\alpha_t z + 1)^2}.$$

Using $y_t(z)(\alpha_t z + 1) = y_{t-1} + \alpha_t$, we can rewrite this as

$$\frac{dy_t(z)}{dz} = -\frac{\alpha_t y_t(z)}{\alpha_t z + 1}.$$

Since $\nu_t(z) = -\log y_t(z)$, we have

$$\frac{d\nu_t(z)}{dz} = -\frac{1}{y_t(z)}\frac{dy_t(z)}{dz} = \frac{\alpha_t}{\alpha_t z + 1}.$$

As a result,

$$\frac{dh_t(z)}{dz} = \frac{dy_t(z)}{dz}\big(\nu_t(z) - \nu_*\big) + y_t(z)\frac{d\nu_t(z)}{dz} = \frac{\alpha_t y_t(z)}{\alpha_t z + 1}\big(1 - (\nu_t(z) - \nu_*)\big).$$

Since $\nu_t(z), \nu_* \in [c_0, c_1]$, then

$$\big|1 - (\nu_t(z) - \nu_*)\big| \le 1 + c_1 - c_0,$$

and since $y_t(z) = \frac{y_{t-1}+\alpha_t}{\alpha_t z+1} \le y_{t-1} + \alpha_t \le (1+\rho)y_{t-1}$, we have

$$\left|\frac{dh_t}{dz}\right| \le \alpha_t y_{t-1}(1+\rho)(1 + c_1 - c_0),$$

which means i.e. $h_t$ is $L_t$-Lipschitz with

$$L_t \le \alpha_t y_{t-1} C.$$

Then, it holds

$$\big|(z' - z)\big(h_t(z) - h_t(z')\big)\big| \le L_t\,(z' - z)^2 \le C\alpha_t y_{t-1}(z' - z)^2.$$

Thus,

$$\begin{aligned}&\mathbb{E}\Big[\big|(z' - z)\big(h_t(z) - h_t(z')\big) \mid \mathcal{F}_{t-1}\Big] \le C\alpha_t \mathbb{E}[y_{t-1}(z' - z)^2) \mid \mathcal{F}_{t-1}] \\ &= C\alpha_t \cdot 2\mathbb{E}[y_{t-1}(z - \mathbb{E}[z])^2 \mid \mathcal{F}_{t-1}] \le 2C\alpha_t \delta_t^2,\end{aligned}$$

where the last step uses the definition of $\delta_t^2$. Applying this result to (5.86), we have

$$\Big|\mathbb{E}\big[(\mu_t - z)h_t(z) \mid \mathcal{F}_{t-1}\big]\Big| \le \frac{1}{2}\mathbb{E}\Big[\big|(z' - z)\big(h_t(z) - h_t(z')\big)\big| \mid \mathcal{F}_{t-1}\Big] \le C\alpha_t \delta_t^2.$$

By noting (5.85), we finish the proof. □

Combining Lemma 5.29 and Lemma 5.30, we have the following lemma for one-step analysis of the $\nu$-update.

**Lemma 5.31** *Under Assumption (5.17) (ii), we have*

$$\mathbb{E}[\alpha_t \nabla_2 F(\mathbf{w}_t, \nu_t)^\top (\nu_t - \nu_*)] \le \mathbb{E}[D_\varphi(\nu_*, \nu_{t-1}) - D_\varphi(\nu_*, \nu_t) + C\alpha_t^2 \delta_t^2]. \quad (5.87)$$

Finally, we state the convergence result of SCENT in the following theorem.

**Theorem 5.14** *Suppose Assumption 5.17 holds. Let $\beta_t = 1, \eta_t = \eta\alpha_t$, $\alpha_t < \rho e^{-v_{t-1}}$, then SCENT guarantees that*

$$\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t(F(\mathbf{w}_t,\nu_t)-F(\mathbf{w}_*,\nu_*))\right]$$
$$\leq \frac{1}{2\eta}\|\mathbf{w}_1-\mathbf{w}\|_2^2+D_\varphi(\nu_*,\nu_0)+\mathbb{E}\left[\sum_{t=1}^{T}\frac{\eta\alpha_t^2\sigma_t^2}{2}+\sum_{t=1}^{T}C\alpha_t^2\delta_t^2\right].$$

*Proof.* Since $\eta_t=\eta\alpha_t$, from Lemma 5.28, we obtain

$$\mathbb{E}[\alpha_t\nabla_1 F(\mathbf{w}_t,\nu_t)^\top(\mathbf{w}_t-\mathbf{w}_*)]\leq\mathbb{E}\left[\frac{1}{2\eta}\|\mathbf{w}_t-\mathbf{w}\|_2^2-\frac{1}{2\eta}\|\mathbf{w}_{t+1}-\mathbf{w}_*\|_2^2+\frac{\eta\alpha_t^2\sigma_t^2}{2}\right].$$

Combining this with Lemma 5.31, we have

$$\mathbb{E}[\alpha_t(\nabla_1 F(\mathbf{w}_t,\nu_t)^\top(\mathbf{w}_t-\mathbf{w}_*)+\nabla_2 F(\mathbf{w}_t,\nu_t)^\top(\nu_t-\nu_*))]$$
$$\leq\mathbb{E}\left[\frac{1}{2\eta}\|\mathbf{w}_t-\mathbf{w}\|_2^2-\frac{1}{2\eta}\|\mathbf{w}_{t+1}-\mathbf{w}_*\|_2^2+D_\varphi(\nu_*,\nu_{t-1})-D_\varphi(\nu_*,\nu_t)\right]$$
$$+\mathbb{E}\left[\frac{\eta\alpha_t^2\sigma_t^2}{2}+C\alpha_t^2\delta_t^2\right].$$

By the joint convexity of $F(\mathbf{w},\nu)$, we have

$$\alpha_t(F(\mathbf{w}_t,\nu_t)-F(\mathbf{w}_*,\nu_*)\leq\alpha_t(\nabla_1 F(\mathbf{w}_t,\nu_t)^\top(\mathbf{w}_t-\mathbf{w}_*)+\nabla_2 F(\mathbf{w}_t,\nu_t)^\top(\nu_t-\nu_*)).$$

Combining the last two inequalities and summing over $t=1,\ldots,T$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t(F(\mathbf{w}_t,\nu_t)-F(\mathbf{w}_*,\nu_*))\right]$$
$$\leq\frac{1}{2\eta}\|\mathbf{w}_1-\mathbf{w}\|_2^2+D_\varphi(\nu_*,\nu_0)+\mathbb{E}\left[\sum_{t=1}^{T}\frac{\eta\alpha_t^2\sigma_t^2}{2}+\sum_{t=1}^{T}C\alpha_t^2\delta_t^2\right].$$

□

We present two corollaries of the above theorem.

**Corollary 5.2** *Suppose Assumption 5.17 holds. Let $\beta_t=1,\eta_t=\eta\alpha_t$, $\alpha_t=\frac{\alpha}{\sqrt{T}}<\rho e^{-v_{t-1}}$ for some constant $\rho>0$, then SCENT guarantees that*

$$\mathbb{E}\left[(F_1(\bar{\mathbf{w}}_T)-F_1(\mathbf{w}_*))\right]\leq\frac{D_0}{\alpha\sqrt{T}}+\frac{\alpha V}{\sqrt{T}}.$$

*where $\bar{\mathbf{w}}_T=\frac{\sum_{t=1}^{T}\mathbf{w}_t}{T}$, $D_0=\frac{1}{2\eta}\|\mathbf{w}_1-\mathbf{w}_*\|_2^2+D_\varphi(\nu_*,\nu_0)$ and*

$$V=\mathbb{E}\left[\frac{\eta\sum_{t=1}^{T}\sigma_t^2}{2T}+\frac{\sum_{t=1}^{T}C\delta_t^2}{T}\right].$$

*Proof.* Plugging $\alpha_t = \alpha/\sqrt{T}$ into Theorem 5.14, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(F(\mathbf{w}_t, \nu_t) - F(\mathbf{w}_*, \nu_*))\right] \leq \frac{D_0}{\alpha\sqrt{T}} + \frac{\alpha V}{\sqrt{T}}.$$

Using $F_1(\mathbf{w}) = \min_\nu F(\mathbf{w}, \nu), F_1(\mathbf{w}_*) = F(\mathbf{w}_*, \nu_*)$ and the Jensen inequality, we can finish the proof. □

**Why it matters**

Since $\delta_t, \sigma_t$ are finite, the above result implies a convergence rate of $O(1/\sqrt{T})$ for SCENT.

**Corollary 5.3** *Suppose Assumption 5.17 holds. Let* $\beta_t = 1, \eta_t = \eta\alpha_t$, $\alpha_t = \frac{\alpha e^{-\nu_{t-1}}}{\sqrt{T}}$, *if* $\frac{1}{T}\sum_{t=1}^{T} e^{-\nu_{t-1}} \geq S$ *almost surely, then SCENT guarantees that*

$$\mathbb{E}\left[F_1(\hat{\mathbf{w}}_T) - F_1(\mathbf{w}_*)\right] \leq \frac{D_0}{\alpha\sqrt{T}S} + \frac{\alpha\bar{V}}{\sqrt{T}S}.$$

*where* $\hat{\mathbf{w}}_T = \frac{\sum_t \alpha_t \mathbf{w}_t}{\sum_{t=1}^{T}\alpha_t}$ *and*

$$\bar{V} = \mathbb{E}\left[\frac{\eta\sum_{t=1}^{T} e^{-2\nu_{t-1}}\sigma_t^2}{2T} + \frac{\sum_{t=1}^{T} C e^{-2\nu_{t-1}}\delta_t^2}{T}\right].$$

*Proof.* Let $\hat{\alpha}_t = \frac{\alpha_t}{\sum_{t=1}^{T}\alpha_t}$. From Theorem 5.14, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\sum_{t=1}^{T}\hat{\alpha}_t(F(\mathbf{w}_t, \nu_t) - F(\mathbf{w}_*, \nu_*))\right] \\
&\leq \frac{1}{2\eta}\|\mathbf{w}_1 - \mathbf{w}\|_2^2 + D_\varphi(\nu_*, \nu_0) + \mathbb{E}\left[\sum_{t=1}^{T}\frac{\eta\alpha_t^2\sigma_t^2}{2} + \sum_{t=1}^{T} C\alpha_t^2\delta_t^2\right].
\end{aligned}$$

Since $\sum_{t=1}^{T}\alpha_t = \sum_{t=1}^{T}\frac{\alpha e^{-\nu_{t-1}}}{\sqrt{T}} \geq \alpha\sqrt{T}S$, then

$$\mathbb{E}\left[\sum_{t=1}^{T}\hat{\alpha}_t(F(\mathbf{w}_t, \nu_t) - F(\mathbf{w}_*, \nu_*))\right] \leq \frac{\frac{1}{2\eta}\|\mathbf{w}_1 - \mathbf{w}\|_2^2 + D_\varphi(\nu_*, \nu_0)}{\alpha\sqrt{T}S} + \frac{\alpha\bar{V}}{\sqrt{T}S}.$$

Applying the joint convexity of $F(\mathbf{w}, \nu)$ and $F_1 = \min_\nu F(\mathbf{w}, \nu)$, we can finish the proof. □

**Why it matters**

Under the stated setting, SCENT reduces to SCGD with $\gamma_t = \frac{\alpha}{\sqrt{T}+\alpha}$. Since $S$ can be lower bounded by a constant, the above corollary implies $O(1/\sqrt{T})$ convergence rate for SCGD to minimize log-E-Exp.

### Analysis of the Variance Terms

Since the final convergence bound depends on the variance terms $\sigma_t^2, \delta_t^2$, we would like to provide further analysis on them.

Let us introduce some notations:

$$z(\mathbf{w};\zeta) = e^{s(\mathbf{w};\zeta)},\ \mu(\mathbf{w}) = \log \mathbb{E}_\zeta e^{s(\mathbf{w};\zeta)}, \tag{5.88}$$

$$m_t = \mathbb{E}_\zeta e^{s(\mathbf{w}_t;\zeta)},\ \mu_t = \mu(\mathbf{w}_t) = \log m_t. \tag{5.89}$$

For the analysis, we make two reasonable assumptions.

**Assumption 5.18.** *Assume there exist constants $\kappa, \sigma'^2$ such that (i) $\mathbb{E}\left[\frac{\mathbb{E}[z(\mathbf{w};\zeta)^2]}{(\mathbb{E}[z(\mathbf{w};\zeta)])^2}\right] \leq \kappa$ for all $\mathbf{w}$; (ii) $\mathbb{E}\|e^{s(\mathbf{w}_t;\zeta')-\mu_t}\nabla s(\mathbf{w}_t;\zeta')\|^2 \leq \sigma'^2$ for all $t$;*

**Critical:** These assumptions are necessary. In next section, we show that the dependence on $\kappa$ is unavoidable. The second assumption is the standard bounded stochastic gradient assumption for optimizing $F_1(\mathbf{w})$.

**Lemma 5.32 (Dual Variance Term)** *Under Assumption 5.18, we have*

$$\delta_t^2 \leq 2(\kappa-1)m_t\Big(F(\mathbf{w}_t,\nu_{t-1}) - F(\mathbf{w}_*,\nu_*) + 1\Big). \tag{5.90}$$

**Why it matters**

When $F(\mathbf{w}_t,\nu_{t-1}) - F(\mathbf{w}_*,\nu_*) \to 0$, the variance term in the convergence bound caused by the stochastic update of $\nu_t$ will be dominated by $2(\kappa-1)m_t$. Large $m_t$ can be mitigated by choosing small $\alpha_t$.

*Proof.* Recall that

$$\delta_t^2 = \mathbb{E}_{\zeta_t}\left[e^{-\nu_{t-1}}\big(z(\mathbf{w}_t;\zeta_t) - m_t\big)^2\right]$$

By Assumption 5.18(i),

$$\mathrm{Var}(z(\mathbf{w}_t;\zeta)) \leq (\kappa-1)m_t^2.$$

Hence

$$\delta_t^2 = e^{-\nu_{t-1}} \operatorname{Var}(z(\mathbf{w}_t; \zeta)) \leq (\kappa - 1) e^{-\nu_{t-1}} m_t^2 = (\kappa - 1) m_t \cdot (m_t e^{-\nu_{t-1}}).$$

Let $\tilde{r}_{t-1} \coloneqq m_t e^{-\nu_{t-1}}$. By the definition:

$$F(\mathbf{w}_t, \nu_{t-1}) = \mathbb{E} e^{s(\mathbf{w}_t; \zeta) - \nu_{t-1}} + \nu_{t-1} = \tilde{r}_{t-1} + \nu_{t-1}.$$

Since $\tilde{r}_{t-1} = e^{\log m_t - \nu_{t-1}}$, we have

$$F(\mathbf{w}_t, \nu_{t-1}) - (1 + \mu_t) = \tilde{r}_{t-1} + \nu_{t-1} - (1 + \log m_t) = \tilde{r}_{t-1} - \log \tilde{r}_{t-1} - 1.$$

Using $r \leq 2(r - \log r)$ for all $r > 0$ yields

$$\tilde{r}_{t-1} \leq 2\big(F(\mathbf{w}_t, \nu_{t-1}) - (1 + \mu_t) + 1\big).$$

Since $\mathbf{w}_*$ minimizes $\mu(\mathbf{w})$, we have $\mu_t = \mu(\mathbf{w}_t) \geq \mu(\mathbf{w}_*)$ and thus $(1 + \mu_t) \geq (1 + \mu(\mathbf{w}_*)) = F(\mathbf{w}_*, \nu_*)$, implying

$$F(\mathbf{w}_t, \nu_{t-1}) - (1 + \mu_t) \leq F(\mathbf{w}_t, \nu_{t-1}) - F(\mathbf{w}_*, \nu_*).$$

As a result, we have

$$\tilde{r}_{t-1} \leq 2\big(F(\mathbf{w}_t, \nu_{t-1}) - F(\mathbf{w}_*, \nu_*) + 1\big). \tag{5.91}$$

Combining this with the bound of $\delta_t^2$, we complete the proof. □

**Lemma 5.33 (Primal Variance Term)** *Under Assumption 5.18, we have*

$$\sigma_t^2 \leq 4\sigma'^2 \big(F(\mathbf{w}_t, \nu_t) - F(\mathbf{w}_*, \nu_*) + 1\big)^2.$$

**Why it matters**

When $F(\mathbf{w}_t, \nu_t) - F(\mathbf{w}_*, \nu_*) \to 0$, the variance term in the convergence bound caused by the stochastic update of $\mathbf{w}_t$ will be dominated by $O(\sigma'^2)$.

*Proof.*

$$\begin{aligned}
\sigma_t^2 &= \mathbb{E}_{\zeta_t'} \| \exp(s(\mathbf{w}_t; \zeta_t') - \nu_t) \nabla s(\mathbf{w}_t; \zeta_t') \|_2^2], \\
&= \mathbb{E}_{\zeta_t'} [e^{2(\mu_t - \nu_t)} \| \exp(s(\mathbf{w}_t; \zeta_t') - \mu_t) \nabla s(\mathbf{w}_t; \zeta_t') \|_2^2] \leq r_t^2 \sigma'^2,
\end{aligned}$$

where $r_t = e^{\mu_t - \nu_t}$. Similar to (5.91), we have show that

$$r_t \leq 2\big(F(\mathbf{w}_t, \nu_t) - F(\mathbf{w}_*, \nu_*) + 1\big).$$

Hence,

$$\sigma_t^2 \leq 4\sigma'^2 \big(F(\mathbf{w}_t, \nu_t) - F(\mathbf{w}_*, \nu_*) + 1\big)^2.$$

□

**Algorithm 22** The SCENT Algorithm for solving CERM

1: Initialize $\mathbf{w}_1, \boldsymbol{\nu}_0$, step sizes $\eta_t$ and $\alpha_t$, $\varphi(\nu) = e^{-\nu}$.
2: **for** $t = 1 \ldots, T-1$ **do**
3: Sample $\mathcal{B}_t \subset \{1, \ldots, n\}$ with $|\mathcal{B}_t| = B$
4: **for** each $i \in \mathcal{B}_t$ **do**
5: Sample $\zeta_{i,t}, \zeta'_{i,t} \sim \mathbb{P}_i$
6: Update $\nu_{i,t} = \arg\min_\nu \exp(s_i(\mathbf{w}_t; \zeta_{i,t}) - \nu) + \nu + \frac{1}{\alpha_t} D_\varphi(\nu, \nu_{i,t-1})$
7: **end for**
8: Compute $\mathbf{z}_t = \frac{1}{B} \sum_{i \in \mathcal{B}_t} \exp(s_i(\mathbf{w}_t; \zeta'_{i,t}) - \nu_{i,t}) \nabla s_i(\mathbf{w}_t; \zeta'_{i,t})$
9: Compute $\mathbf{v}_t = (1 - \beta_t)\mathbf{v}_{t-1} + \beta_t \mathbf{z}_t$
10: Update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{v}_t$
11: **end for**

#### 5.5.2.2 Compositional Entropic Risk Minimization

In this section, we extend the results to solving compositional entropic risk minimization (CERM):

$$\min_{\mathbf{w}} F_1(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^{n} \log\left(\mathbb{E}_{\zeta \sim \mathbb{P}_i} \exp(s_i(\mathbf{w}; \zeta))\right)$$

via its equivalent min-min formulation:

$$\min_{\mathbf{w}} \min_{\boldsymbol{\nu}} F(\mathbf{w}, \boldsymbol{\nu}) := \frac{1}{n} \sum_{i=1}^{n} \{\mathbb{E}_{\zeta \sim \mathbb{P}_i} \exp(s_i(\mathbf{w}; \zeta) - \nu_i) + \nu_i\}.$$

The difference from Log-E-Exp is that there are multiple $\nu_i, i = 1, \ldots, n$, which needs to be updated using stochastic block coordinate method. The technique has been used in algorithms presented in previous sections of this chapter.

We present an extension of SCENT to solving CERM in Algorithm 22. The major change lies at the stochastic block coordinate update of $\boldsymbol{\nu}$ in Step 5. This extension is analogous to SOX for FCCO, employing stochastic block-coordinate updates for the inner estimators. Indeed, SOX applied to CERM can be recovered as a special case of SCENT by choosing the coordinate-wise step size $\alpha_{t,i} = \frac{\gamma_t}{1-\gamma_t} e^{-\nu_{i,t-1}}$, using an argument similar to (5.83).

#### Convergence analysis for convex problems

Let us define some notations:

$$\Phi_i(\mathbf{w}_t, \nu_i; \zeta) = \exp(s_i(\mathbf{w}_t; \zeta_t) - \nu_i) + \nu_i$$
$$F_i(\mathbf{w}_t, \nu_i) = \mathbb{E}_{\zeta \sim \mathbb{P}_i}[\Phi_i(\mathbf{w}_t, \nu_i; \zeta)]$$
$$(\mathbf{w}_*, \boldsymbol{\nu}_*) = \arg\min_{\mathbf{w}, \boldsymbol{\nu}} F(\mathbf{w}, \boldsymbol{\nu}).$$

Similar as before, $\nu_{i,*} = \log[\mathbb{E}_{\zeta\sim\mathbb{P}_i}\exp(s_i(\mathbf{w}_*;\zeta))]$. Since we deal with stochastic block coordinate update, we introduce a virtual sequence $\bar{\boldsymbol{\nu}}_t$, where

$$\bar{\nu}_{i,t} = \arg\min_{\nu} \exp(s_i(\mathbf{w}_t;\zeta_{i,t}) - \nu) + \nu + \frac{1}{\alpha_t} D_\varphi(\nu, \nu_{i,t-1}), \forall i$$

Following Lemma 5.26, we have

$$e^{\bar{\nu}_{i,t}} = \frac{1}{1+\alpha_t e^{\nu_{i,t-1}}} e^{\nu_{i,t-1}} + \frac{\alpha_t e^{\nu_{i,t-1}}}{1+\alpha_t e^{\nu_{i,t-1}}} \exp(s_i(\mathbf{w}_t;\zeta_t)), \forall i.$$

**Assumption 5.19.** *Assume that the following conditions hold:*

*(i)* $s_i(\mathbf{w};\zeta)$ *is convex;*
*(ii) the loss function is bounded such that* $s_i(\mathbf{w};\zeta) \in [c_0, c_1], \forall \mathbf{w}, \zeta, i.$
*(iii) there exists* $G$ *such that* $\mathbb{E}_\zeta \|\nabla s_i(\mathbf{w}_t,\zeta)\|_2^2] \le G^2, \forall t, i$

Define $\sigma_{i,t}, \delta_{i,t}$ as

$$\sigma_{i,t}^2 := \mathbb{E}_{\zeta'_{i,t}\sim\mathbb{P}_i} \| \exp(s_i(\mathbf{w}_t;\zeta'_{i,t}) - \nu_{i,t}) \nabla s_i(\mathbf{w}_t;\zeta'_{i,t})\|_2^2], \forall i, t,$$
$$\delta_{i,t}^2 := \mathbb{E}_{\zeta_{i,t}\sim\mathbb{P}_i}[e^{-\nu_{i,t-1}} | e^{s_i(\mathbf{w}_t;\zeta_{i,t})} - \mathbb{E}_{\zeta_i\sim\mathbb{P}_i}[e^{s_i(\mathbf{w}_t;\zeta_i)}]|^2], \forall i, t.$$

Similar to Lemma 5.27, the following lemma can be proved.

**Lemma 5.34** *Under Assumption 5.19, if* $\boldsymbol{\nu}_0 \in [c_0, c_1]$ *then* $\boldsymbol{\nu}_t \in [c_0, c_1], \forall t.$

Similar to Lemma 5.28, we have the following lemma regarding one-step update of $\mathbf{w}_t$.

**Lemma 5.35** *Under Assumption (5.19) and* $\beta_t = 1$*, we have*

$$\mathbb{E}[\eta_t \nabla_1 F(\mathbf{w}_t, \bar{\boldsymbol{\nu}}_t)^\top (\mathbf{w}_t - \mathbf{w}_*)] \le \mathbb{E}\left[\frac{1}{2}\|\mathbf{w}_t - \mathbf{w}_*\|_2^2 - \frac{1}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2\right] + \frac{\eta_t^2 \sigma_t^2}{2},$$

*where* $\sigma_t^2 = \frac{1}{n}\sum_{i=1}^n \sigma_{i,t}^2$.

*Proof.* We first bound $\mathbb{E}_t[\|\mathbf{z}_t\|_2^2 \mid \mathcal{F}_{t-1}]$, where $\mathbb{E}_t$ denotes the expectation over randomness in $t$-th iteration given $\mathbf{w}_t, \nu_{t-1}$.

$$\mathbb{E}_t[\|\mathbf{z}_t\|_2^2] = \mathbb{E}_t\left[\left\|\frac{1}{B}\sum_{i\in\mathcal{B}_t} \exp(s_i(\mathbf{w}_t;\zeta'_{i,t}) - \nu_{i,t})\nabla s_i(\mathbf{w}_t;\zeta'_{i,t})\right\|_2^2\right]$$
$$= \mathbb{E}_{\mathcal{B}_t,\zeta_t}\mathbb{E}_{\zeta'_t|\mathcal{B}_t,\zeta_t}\left[\left\|\frac{1}{B}\sum_{i\in\mathcal{B}_t} \exp(s_i(\mathbf{w}_t;\zeta'_{i,t}) - \nu_{i,t})\nabla s_i(\mathbf{w}_t;\zeta'_{i,t})\right\|_2^2\right]$$
$$\le \mathbb{E}_{\mathcal{B}_t,\zeta_t}\left[\frac{1}{B}\sum_{i\in\mathcal{B}_t}\sigma_{i,t}^2\right] = \frac{1}{n}\sum_{i=1}^n \sigma_{i,t}^2.$$

Since $\bar{\nu}_{i,t} = \nu_{i,t}, \forall i \in \mathcal{B}_t$, we have

$$\mathbb{E}_t[\mathbf{z}_t] = \mathbb{E}_{\zeta'_t,\zeta_t,\mathcal{B}_t}\left[\frac{1}{B}\sum_{i\in\mathcal{B}_t}\nabla\Phi_i(\mathbf{w}_t,\bar{\nu}_{i,t};\zeta'_{i,t})\right] = \nabla_1 F(\mathbf{w}_t,\bar{\boldsymbol{\nu}}_t).$$

Then following Lemma 3.3, we can finish the proof. □

Next, we analyze the update of $\bar{\nu}_t$.

**Lemma 5.36** *Under Assumption (5.19) (ii) and* $\alpha_t \le \min_i \rho e^{-\nu_{i,t-1}}$*, we have*

$$\mathbb{E}[\alpha_t\nabla_2 F(\mathbf{w}_t,\bar{\boldsymbol{\nu}}_t)^\top(\bar{\boldsymbol{\nu}}_t-\boldsymbol{\nu}_*)] \le \frac{1}{B}\mathbb{E}\left[D_\varphi(\boldsymbol{\nu}_*,\boldsymbol{\nu}_{t-1}) - D_\varphi(\boldsymbol{\nu}_*,\boldsymbol{\nu}_t)\right] + C\alpha_t^2\delta_t^2.$$

*where* $D_\varphi(\boldsymbol{\nu}_*,\boldsymbol{\nu}_t) = \sum_{i=1}^n D_\varphi(\nu_{i,*},\nu_{i,t})$ *and* $\delta_t^2 = \frac{1}{n}\sum_{i=1}^n \delta_{i,t}^2$.

*Proof.* By applying Lemma 5.30 and Lemma 5.29 for each coordinate of $\bar{\nu}_{i,t}$, we have

$$\mathbb{E}[\alpha_t\nabla_2 F_i(\mathbf{w}_t,\bar{\nu}_{i,t})^\top(\bar{\nu}_{i,t}-\nu_{i,*})] \le D_\varphi(\nu_{i,*},\nu_{i,t-1}) - D_\varphi(\nu_{i,*},\bar{\nu}_{i,t}) + C\alpha_t^2\delta_{i,t}^2, \forall i.$$

Averaging the above inequality over $i = 1,\ldots,n$, we have

$$\mathbb{E}[\alpha_t\nabla_2 F(\mathbf{w}_t,\bar{\boldsymbol{\nu}}_t)^\top(\bar{\boldsymbol{\nu}}_t-\boldsymbol{\nu}_*)] \le \frac{1}{n}\sum_{i=1}^n\left(D_\varphi(\nu_{i,*},\nu_{i,t-1}) - D_\varphi(\nu_{i,*},\bar{\nu}_{i,t})\right) + C\alpha_t\delta_t^2. \tag{5.92}$$

Due to the randomness of $\mathcal{B}_t$, we have

$$\mathbb{E}[D_\varphi(\nu_{i,*},\nu_{i,t})] = \mathbb{E}\left[(1-\frac{B}{n})D_\varphi(\nu_{i,*},\nu_{i,t-1}) + \frac{B}{n}D_\varphi(\nu_{i,*},\bar{\nu}_{i,t})\right], \forall i.$$

Hence

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left(D_\varphi(\nu_{i,*},\nu_{i,t-1}) - D_\varphi(\nu_{i,*},\bar{\nu}_{i,t})\right)\right]\\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left(D_\varphi(\nu_{i,*},\nu_{i,t-1}) - \frac{n}{B}D_\varphi(\nu_{i,*},\nu_{i,t}) + (\frac{n}{B}-1)D_\varphi(\nu_{i,*},\nu_{i,t-1})\right)\right]\\
&= \frac{1}{B}\mathbb{E}\left[\sum_{i=1}^n(D_\varphi(\nu_{i,*},\nu_{i,t-1}) - D_\varphi(\nu_{i,*},\nu_{i,t}))\right].
\end{aligned}$$

Combining this with (5.92), we finish the proof. □

Finally, we state the convergence result of SCENT in the following theorem.

**Theorem 5.15** *Suppose Assumption 5.19 holds. Let* $\beta_t = 1, \eta_t = \eta\alpha_t$*, and* $\alpha_t = \frac{\alpha}{\sqrt{T}} < \rho\min_i e^{-v_{i,t-1}}$*, then SCENT guarantees that*

$$\mathbb{E}\left[(F_1(\bar{\mathbf{w}}_T) - F_1(\mathbf{w}_*))\right] \leq \frac{1}{2\eta\alpha\sqrt{T}}\|\mathbf{w}_1 - \mathbf{w}_*\|_2^2 + \frac{D_\varphi(\boldsymbol{\nu}_*, \boldsymbol{\nu}_0)}{\alpha B\sqrt{T}} + \frac{\alpha V}{\sqrt{T}}.$$

*where* $\bar{\mathbf{w}}_T = \frac{\sum_{t=1}^T \mathbf{w}_t}{T}$, *and* $V = \mathbb{E}\left[\frac{\eta \sum_{t=1}^T \sigma_t^2}{2T} + \frac{\sum_{t=1}^T C\delta_t^2}{T}\right]$.

**Why it matters**

In order to achieve an $\epsilon$-optimal solution, the above convergence bound implies the following complexity:

$$T = O\left(\frac{\|\mathbf{w}_1 - \mathbf{w}_*\|_2^4}{\eta^2\alpha^2\epsilon^2} + \frac{D_\varphi(\boldsymbol{\nu}_*, \boldsymbol{\nu}_0)^2}{\alpha^2 B^2\epsilon^2} + \frac{\alpha^2 V^2}{\epsilon^2}\right).$$

For simplicity of discussion, let us consider a setting of $\eta$ such that the first term matches the second term. As a result, the complexity becomes:

$$T = O\left(\frac{D_\varphi(\boldsymbol{\nu}_*, \boldsymbol{\nu}_0)^2}{\alpha^2 B^2\epsilon^2} + \frac{\alpha^2 V^2}{\epsilon^2}\right).$$

**Insight 1:** Since $\sigma_t, \delta_t$ are finite, and $D_\varphi(\boldsymbol{\nu}_*, \boldsymbol{\nu}_0) = O(n)$, if $\alpha \propto \sqrt{n/B}$, the above result implies an iteration complexity of $O(\frac{n}{B\epsilon^2})$ for SCENT.
**Insight 2:** When the loss $s_i(\mathbf{w}_t; \zeta) \geq 0$ is large, the term $e^{-\nu_{i,t-1}}$ becomes very small, suggesting that the step size parameter $\alpha$ should be chosen small so as to mitigate the large variance term $\delta_t$. In contrast, when the loss $s_i(\mathbf{w}_t; \zeta) < 0$ is small, the term $e^{-\nu_{i,t-1}}$ can become large, allowing $\alpha$ to be set relatively larger, which helps offset the large distance measure $D_\varphi(\boldsymbol{\nu}_*, \boldsymbol{\nu}_0)$.

*Proof.* Since $\eta_t = \eta\alpha_t$, from Lemma 5.35, we obtain

$$\mathbb{E}[\alpha_t \nabla_1 F(\mathbf{w}_t, \bar{\boldsymbol{\nu}}_t)^\top (\mathbf{w}_t - \mathbf{w}_*)] \leq \mathbb{E}\left[\frac{1}{2\eta}\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \frac{1}{2\eta}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2 + \frac{\eta\alpha_t^2\sigma_t^2}{2}\right].$$

Adding this to the inequality in Lemma 5.36, we have

$$\begin{aligned}
&\mathbb{E}[\alpha_t(\nabla_1 F(\mathbf{w}_t, \bar{\boldsymbol{\nu}}_t)^\top (\mathbf{w}_t - \mathbf{w}_*) + \nabla_2 F(\mathbf{w}_t, \bar{\boldsymbol{\nu}}_t)^\top (\bar{\boldsymbol{\nu}}_t - \boldsymbol{\nu}_*))] \\
&\leq \mathbb{E}\left[\frac{1}{2\eta}\|\mathbf{w}_t - \mathbf{w}\|_2^2 - \frac{1}{2\eta}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|_2^2 + \frac{1}{B}D_\varphi(\boldsymbol{\nu}_*, \boldsymbol{\nu}_{t-1}) - \frac{1}{B}D_\varphi(\boldsymbol{\nu}_*, \boldsymbol{\nu}_t)\right] \\
&+ \mathbb{E}\left[\frac{\eta\alpha_t^2\sigma_t^2}{2} + C\alpha_t^2\delta_t^2\right].
\end{aligned}$$

By the joint convexity of $F(\mathbf{w}, \nu)$, we have

$$\alpha_t(F(\mathbf{w}_t, \bar{\boldsymbol{\nu}}_t) - F(\mathbf{w}_*, \boldsymbol{\nu}_*) \leq \alpha_t(\nabla_1 F(\mathbf{w}_t, \bar{\boldsymbol{\nu}}_t)^\top (\mathbf{w}_t - \mathbf{w}_*) + \nabla_2 F(\mathbf{w}_t, \bar{\boldsymbol{\nu}}_t)^\top (\bar{\boldsymbol{\nu}}_t - \boldsymbol{\nu}_*)).$$

Combining the last two inequalities and summing over $t = 1, \ldots, T$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t(F(\mathbf{w}_t,\bar{\boldsymbol{\nu}}_t)-F(\mathbf{w}_*,\boldsymbol{\nu}_*))\right]$$
$$\leq \frac{1}{2\eta}\|\mathbf{w}_1-\mathbf{w}\|_2^2+\frac{1}{B}D_\varphi(\boldsymbol{\nu}_*,\boldsymbol{\nu}_0)+\mathbb{E}\left[\sum_{t=1}^{T}\frac{\eta\alpha_t^2\sigma_t^2}{2}+\sum_{t=1}^{T}C\alpha_t^2\delta_t^2\right].$$

Since $F_1(\mathbf{w}_*)=F(\mathbf{w}_*,\boldsymbol{\nu}_*)$, and $F_1(\mathbf{w}_t)\leq F(\mathbf{w}_t,\bar{\boldsymbol{\nu}}_t)$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t(F_1(\mathbf{w}_t)-F_1(\mathbf{w}_*))\right]$$
$$\leq \frac{1}{2\eta}\|\mathbf{w}_1-\mathbf{w}\|_2^2+\frac{1}{B}D_\varphi(\boldsymbol{\nu}_*,\boldsymbol{\nu}_0)+\mathbb{E}\left[\sum_{t=1}^{T}\frac{\eta\alpha_t^2\sigma_t^2}{2}+\sum_{t=1}^{T}C\alpha_t^2\delta_t^2\right].$$

Plugging the value of $\alpha_t$, we finish the proof.

□

#### 5.5.2.3 Why SCENT is better than ASGD?

In this section, we provide theoretical insight into why SCENT outperforms ASGD for entropic risk minimization. The key distinction between the two methods lies in their updates of the dual variable $\nu$: SCENT employs a stochastic proximal mirror descent (SPMD) update, whereas ASGD relies on a standard SGD update. Accordingly, our analysis focuses exclusively on the $\nu$-update while keeping $\mathbf{w}$ fixed. In particular, we consider the following problem:

$$\min_{\nu} F(\nu) := \mathbb{E}_\zeta e^{s(\zeta)-\nu}+\nu, \tag{5.93}$$

where we omit $\mathbf{w}$ in $s(\zeta)$.

Recall the definitions $z \coloneqq e^{s(\zeta)}$, $m \coloneqq \mathbb{E}[z]$, $r(\nu) \coloneqq me^{-\nu}=e^{\nu_*-\nu}$ as used previously, and the facts $\nu_*=\arg\min_\nu F(\nu)=\log m, F(\nu_*)=me^{-\nu_*}+\nu_*=1+\nu_*$. Recall the SPMD update:

$$e^{\nu_t}=\frac{1}{1+\alpha_t e^{\nu_{t-1}}}e^{\nu_{t-1}}+\frac{\alpha_t e^{\nu_{t-1}}}{1+\alpha_t e^{\nu_{t-1}}}e^{s(\zeta_t)}.$$

Let us define an important quantity to characterize the difficulty of the problem:

$$\kappa=\frac{\mathbb{E}[z^2]}{(\mathbb{E}[z])^2},$$

which is known as second-order moment ratio. Larger $\kappa$ indicates heavier tails or higher variability relative to the mean.

### A Clean Bound of SPMD

The optimality gap can be written as

$$F(\nu) - F(\nu_*) = me^{-\nu} + \nu - (1 + \nu_*) = r(\nu) - \log r(\nu) - 1. \tag{5.94}$$

We assume $s(\zeta) \in [c_0, c_1]$ and without loss of generality we assume $c_1 \leq 0$. If not, we can define $s'(\zeta) = s(\zeta) - c_1, z' = e^{s'(\zeta)}$ and $F'(\nu') = \mathbb{E}[z'e^{-\nu'}] + \nu'$. Then $F(\nu) - F(\nu_*) = F'(\nu') - \min F'(\nu)$ if $\nu = \nu' - c_1$.

**Lemma 5.37 (Self-bounding inequality)** *For all $r > 0$,*

$$r \;\leq\; 2\,(r - \log r). \tag{5.95}$$

*Equivalently, for all $\nu \in \mathbb{R}$,*

$$r(\nu) \leq 2\big(F(\nu) - F(\nu_*) + 1\big). \tag{5.96}$$

*Proof.* If $0 < r \leq 2$, then $r \leq 2 \leq 2(r - \log r)$ since $r - \log r \geq 1$ for all $r > 0$. If $r \geq 2$, then $\log r \leq r/2$, hence $r - \log r \geq r/2$, i.e. $r \leq 2(r - \log r)$. Substituting $r = r(\nu)$ and using (5.94) yields (5.96). □

**Theorem 5.16** *Suppose $s(\zeta) \in [c_0, c_1] \leq 0$ holds. By setting $\alpha_t = \sqrt{\frac{D_\varphi(\nu_*, \nu_0)m}{2CT\mathrm{Var}(z)}} \leq \min(\frac{m}{4C\mathrm{Var}(z)}, \rho)$ for sufficiently large $T$, SPMD guarantees that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F(\nu_t) - F(\nu_*)] \leq 4\sqrt{2}\sqrt{\frac{C\,(\kappa - 1)\,\big(1 - r_0 + r_0 \log r_0\big)}{T}} + \frac{F(\nu_0) - F(\nu_*)}{T}. \tag{5.97}$$

*where $C = (1 + \rho)(1 + c_1 - c_0)$, and $r_0 = r(\nu_0) = e^{\nu_* - \nu_0}$.*

**Why it matters**

When $\nu_0 \gg \nu_*$ (over-estimation), then $1 - r_0 + r_0 \log r_0 = O(1)$, the dominating term becomes $O(\sqrt{\frac{\kappa}{T}})$. This upper bound characterizes the intrinsic complexity of SPMD, which depends on the second-order moment ratio $\kappa$. If $s(\zeta)) \sim \mathcal{N}(\mu_s, \sigma_s^2)$, then $\kappa = e^{\sigma_s^2}$, which does not depend on the exponential of the mean $\mu_s$ but rather $e^{\sigma_s^2}$.

*Proof.* From Lemma 5.31, we obtain the SPMD averaged bound

$$\bar{G}_T \;\coloneqq\; \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F(\nu_t) - F(\nu_*)] \;\leq\; \frac{D_\varphi(\nu_*, \nu_0)}{\alpha T} \;+\; C\,\alpha\,V, \tag{5.98}$$

where

$$V \coloneqq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\delta_t^2], \qquad \delta_t^2 = \mathbb{E}\big[e^{-\nu_{t-1}}(z_t - m)^2\big] = e^{-\nu_{t-1}}\operatorname{Var}(z).$$

Since $e^{-\nu_{t-1}} = r(\nu_{t-1})/m$, we can rewrite

$$V = \frac{\operatorname{Var}(z)}{m} \cdot \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[r(\nu_{t-1})]. \tag{5.99}$$

By Lemma 5.37,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[r(\nu_{t-1})] &\le \frac{2}{T}\sum_{t=1}^{T}\mathbb{E}[F(\nu_{t-1}) - F(\nu_*) + 1] \\ &= 2\left(1 + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F(\nu_{t-1}) - F(\nu_*)]\right).\end{aligned}$$

Next, observe the index shift:

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}[F(\nu_{t-1}) - F(\nu_*)] = \mathbb{E}[F(\nu_0) - F(\nu_*)] + \sum_{t=1}^{T-1}\mathbb{E}[F(\nu_t) - F(\nu_*)] \\ &\le \mathbb{E}[F(\nu_0) - F(\nu_*)] + \sum_{t=1}^{T}\mathbb{E}[F(\nu_t) - F(\nu_*)].\end{aligned}$$

Dividing by $T$ yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F(\nu_{t-1}) - F(\nu_*)] \le \frac{\mathbb{E}[F(\nu_0) - F(\nu_*)]}{T} + \bar{G}_T. \tag{5.100}$$

Combining this with (5.99) we have

$$V \le \frac{2\operatorname{Var}(z)}{m}\left(1 + \bar{G}_T + \frac{\mathbb{E}[F(\nu_0) - F(\nu_*)]}{T}\right). \tag{5.101}$$

Plugging (5.101) into (5.98) yields

$$\bar{G}_T \le \frac{D_\varphi(\nu_*, \nu_0)}{\alpha T} + \frac{2C\alpha\operatorname{Var}(z)}{m}\left(1 + \bar{G}_T + \frac{\mathbb{E}[F(\nu_0) - F(\nu_*)]}{T}\right).$$

If $\alpha \le \frac{m}{4C\operatorname{Var}(z)}$, then $\frac{2C\alpha\operatorname{Var}(z)}{m} \le \frac{1}{2}$, and therefore

$$\bar{G}_T \le \frac{2D_\varphi(\nu_*,\nu_0)}{\alpha T} + \frac{4C\alpha\,\mathrm{Var}(z)}{m}\left(1+\frac{\mathbb{E}[F(\nu_0)-F(\nu_*)]}{T}\right)$$
$$\le \frac{2D_\varphi(\nu_*,\nu_0)}{\alpha T} + \frac{4C\alpha\,\mathrm{Var}(z)}{m} + \frac{F(\nu_0)-F(\nu_*)}{T}.$$

Optimizing the right-hand side over $\alpha$ (assuming $T$ is large enough) gives:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F(\nu_t)-F(\nu_*)] \le 4\sqrt{2}\sqrt{\frac{C\,D_\varphi(\nu_*,\nu_0)\mathrm{Var}(z)}{mT}} + \frac{F(\nu_0)-F(\nu_*)}{T}.$$

With $r_0 \coloneqq r(\nu_0) = e^{\nu_*-\nu_0}$,

$$D_\varphi(\nu_*,\nu_0) = e^{-\nu_*} - e^{-\nu_0} + e^{-\nu_0}(\nu_*-\nu_0) = \frac{1}{m}\big(1-r_0+r_0\log r_0\big).$$

Since $\mathrm{Var}(z)/m^2 = \kappa - 1$, thus the convergence upper bound becomes

$$4\sqrt{2}\sqrt{\frac{C\,(\kappa-1)\big(1-r_0+r_0\log r_0\big)}{T}} + \frac{F(\nu_0)-F(\nu_*)}{T}.$$

□

**Comparison with SGD.**

*Benefit under the noise setting*

In order to control the variance, we consider projected SGD. Let $\Pi_{[c_0,c_1]}$ denote projection onto $[c_0,c_1]$. The projected SGD update is

$$\nu_{t+1} = \Pi_{[c_0,c_1]}\big(\nu_t - \alpha' g_t\big), \qquad g_t \coloneqq 1 - z_t e^{-\nu_t}, \tag{5.102}$$

where $\{z_t\}_{t\ge 0}$ are i.i.d. copies of $z$ and $\alpha' > 0$ is a constant step size. Note that $\mathbb{E}[g_t \mid \nu_t] = \nabla F(\nu_t) = 1 - me^{-\nu_t}$.

We present a corollary of Theorem 3.5 for SGD to minimize $F(\nu)$ below.

**Corollary 5.4** *Suppose $s(\zeta) \in [c_0,c_1]$ holds and $F(\cdot)$ is L-smooth in the range of $[c_0,c_1]$. Let $\{\nu_t\}$ follow (5.102). If $\eta \le \frac{1}{L}$, Then*

$$\bar{G}_T^{\mathrm{SGD}} \coloneqq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[F(\nu_t)-F(\nu_*)] \le \frac{(\nu_0-\nu_*)^2}{2\alpha' T} + \alpha' V'.$$

*where*

$$V' = \frac{\alpha'}{T}\sum_{t=0}^{T-1}(\delta_t')^2 = \frac{\mathrm{Var}(z)}{T}\sum_{t=0}^{T-1}\mathbb{E}[e^{-2\nu_t}].$$

We quantify the smoothness on the bounded domain of the objective, which introduces an exponential constant.

**Lemma 5.38** *On $[c_0, c_1]$, the function $F(\nu) = me^{-\nu} + \nu$ is $L$-smooth with*

$$L = \sup_{\nu \in [c_0, c_1]} F''(\nu) = \sup_{\nu \in [c_0, c_1]} me^{-\nu} = me^{-c_0} = e^{\nu_* - c_0}.$$

*Proof.* We have $F''(\nu) = me^{-\nu}$, which is decreasing in $\nu$, so the maximum over $[c_0, c_1]$ is attained at $c_0$. □

**Theorem 5.17** *By choosing the optimal $\alpha' = \frac{|\nu_0 - \nu_*| e^{c_0}}{\sqrt{2T\mathrm{Var}(z)}} \le \frac{1}{L} = \frac{e^{c_0}}{m}$, SGD's upper bound becomes*

$$\bar{G}_T^{\mathrm{SGD}} \le \sqrt{2} |\nu_0 - \nu_*| \, e^{\nu_* - c_0} \sqrt{\frac{\kappa - 1}{T}}. \tag{5.103}$$

*where $\kappa = \mathbb{E}[z^2]/(\mathbb{E}[z])^2$.*

*Proof.* The proof follows Corollary 5.4 by noting that $V' \le \mathrm{Var}(z) e^{-2c_0}$ and $\mathrm{Var}(z) = m^2(\kappa - 1) = e^{2\nu_*}(\kappa - 1)$. □

**Why it matters**

By comparing the convergence bound of SPMD with that of SGD, the resulting ratio is:

$$\frac{1}{|\nu_0 - \nu_*| e^{\nu_* - c_0}}.$$

Notably, this ratio becomes exponentially small in regimes where $\nu_* \gg c_0$, highlighting the superior efficiency of SPMD.

*Benefit under the noiseless setting*

We further show that, even in the noiseless setting, the dependence of the GD update on $|\nu_0 - \nu_*|$ is unavoidable, whereas the PMD update does not exhibit such dependence when $\nu_0 \gg \nu_*$.

In the noiseless setting, where $m = \mathbb{E}[e^{s(\zeta)}]$ is known, the gradient descent (GD) iteration becomes:

$$\nu_{t+1} = \nu_t - \alpha' \nabla F(\nu_t) = \nu_t - \alpha' \left(1 - me^{-\nu_t}\right), \qquad t \ge 0, \tag{5.104}$$

where $\alpha' > 0$ is a step size. For deterministic PMD, its update is equivalent to (cf. Lemma 5.26):

$$y_{t+1} = \frac{y_t + \alpha}{1 + \alpha m}, \tag{5.105}$$

where $y_t = e^{-\nu_t}$.

**Lemma 5.39 (GD vs PMD)** *Assume $\nu_0 \gg \nu_*$. Let $\{\nu_t\}_{t \ge 0}$ follow (5.104) with $\alpha' \le 1$. Then in order to have $|\nabla F(\nu_t)| \le \epsilon$, then we need at least*

$$t \geq \frac{\nu_0 - \nu_* - \log\left(\frac{1}{1-\epsilon}\right)}{\alpha'}. \tag{5.106}$$

*In contrast, for deterministic PMD update (5.105), in order to ensure $|\nabla F(\nu_t)| \leq \epsilon$. it suffices that*

$$t = \left\lceil \frac{\log\left(|1 - r_0|/\epsilon\right)}{\log(1 + \alpha m)} \right\rceil. \tag{5.107}$$

*Proof.* Recall the definition $r(\nu) \coloneqq me^{-\nu} = e^{\nu_* - \nu}$. We have $|\nabla F(\nu)| = |1 - r(\nu)|$. From (5.104),

$$\nu_{t+1} = \nu_t - \alpha'\left(1 - e^{\nu_* - \nu_t}\right).$$

If $\nu_t \geq \nu_*$, then $\nu_{t+1} - \nu_* = \nu_t - \nu_* - \alpha'\left(1 - e^{\nu_* - \nu_t}\right) \geq 0$ provided $\alpha' \leq 1$. Let $r_t = e^{\nu_* - \nu_t} > 0$. Then, from GD update we have

$$r_{t+1} = r_t e^{\alpha'(1 - r_t)} \leq r_t e^{\alpha'} \leq r_0 e^{\alpha'(t+1)}.$$

In order to have $\|\nabla F(\nu_t)\|_2^2 \leq \epsilon^2$, it is necessary to have $r_t \geq 1 - \epsilon$. Hence, we need at least $t \geq \frac{\log \frac{1-\epsilon}{r_0}}{\alpha'} = \frac{\nu_0 - \nu_* - \log\left(\frac{1}{1-\varepsilon}\right)}{\alpha'}$.

For deterministic PMD update (5.105), since $r_t = my_t$ we have

$$r_{t+1} - 1 = \frac{r_t - 1}{1 + \alpha m}.$$

Taking absolute value yields

$$|\nabla F(\nu_{t+1})| = \frac{|\nabla F(\nu_t)|}{(1 + \alpha m)}.$$

Solving $|\nabla F(\nu_t)| \leq |\nabla F(\nu_0)|/(1 + \alpha m)^t \leq \epsilon$ yields (5.107). □

**Why it matters**

Deterministic GD needs at least $\Omega((\nu_0 - \nu_*)/\alpha')$ steps to enter a constant-accuracy region, whereas PMD reduces $|\nabla F(\nu_t)|$ geometrically with rate $(1 + \alpha m)^{-1}$, yielding a complexity of order $O\left(\frac{1}{\log(1+\alpha m)} \log \frac{1}{\epsilon}\right)$, which does not scale with $\nu_0$ due to $|1 - r_0| = |1 - e^{\nu_* - \nu_0}| \leq 1$.
Indeed, in the noiseless setting for PMD, taking the formal limit $\alpha \to \infty$ yields $y_1 \to 1/m$ thus $\nu_1 \to \nu_*$. This highlights that the PMD update is an implicit, geometry-matched step.

#### 5.5.2.4 An Optimal bound for SPMD

In fact, we can improve the convergence rate of SPMD to $O\left(\frac{\kappa - 1}{T}\right)$, which matches a lower bound to be established. The key is just to use a specially designed learning

rate scheme $\alpha_t$. Recall the SPMD update:

$$y_t = \frac{y_{t-1} + \alpha_t}{1 + \alpha_t z_t}, \quad \forall t \geq 1, \tag{5.108}$$

where $y_{t-1} = e^{-\nu_{t-1}}, z_t = e^{s(\zeta_t)}$.

**Lemma 5.40** *Let $S_t := \sum_{i=1}^t z_i$ and $\bar{z}_t := S_t/t$. Initialize $y_1 = 1/z_1$ (or equivalently $\alpha_1 = \infty$) and for $t \geq 2$ choose*

$$\alpha_t := \frac{y_{t-1}}{t-1} = \frac{1}{S_{t-1}}. \tag{5.109}$$

*Then for all $t \geq 1$,*

$$y_t = \frac{t}{S_t}, \qquad \nu_t = -\log y_t = \log\Big(\frac{S_t}{t}\Big) = \log \bar{z}_t. \tag{5.110}$$

*In particular, $\nu_t$ is the exact minimizer of the empirical objective*

$$\widehat{F}_t(\nu) := \bar{z}_t e^{-\nu} + \nu \quad \textit{since} \quad \arg\min_{\nu} \widehat{F}_t(\nu) = \log \bar{z}_t.$$

*Proof.* We prove (5.110) by induction. For $t = 1$, $y_1 = 1/z_1 = 1/S_1$ holds by initialization. Assume $y_{t-1} = (t-1)/S_{t-1}$. Then (5.109) gives $\alpha_t = 1/S_{t-1}$, and the recursion (5.108) yields

$$y_t = \frac{\frac{t-1}{S_{t-1}} + \frac{1}{S_{t-1}}}{1 + \frac{z_t}{S_{t-1}}} = \frac{\frac{t}{S_{t-1}}}{\frac{S_{t-1}+z_t}{S_{t-1}}} = \frac{t}{S_{t-1} + z_t} = \frac{t}{S_t}.$$

Thus $y_t = t/S_t$ and $\nu_t = -\log y_t = \log(S_t/t) = \log \bar{z}_t$. □

**Assumption 5.20.** *Assume $s(\zeta)$ is $\sigma^2$-subgaussian, i.e.,*

$$\mathbb{E}\big[e^{\lambda(s(\zeta) - \mathbb{E}[s(\zeta)])}\big] \leq e^{\lambda^2\sigma^2/2} \quad \forall \lambda \in \mathbb{R}.$$

This includes Bernoulli distribution (indeed, if $s(\zeta) \in [c_0, c_1]$ a.s., then $s(\zeta) - \mathbb{E}[s(\zeta)]$ is $(c_1 - c_0)^2/4$-subgaussian by Hoeffding's lemma).

Since $\frac{\mathrm{Var}(z)}{(\mathbb{E}[z])^2} = \kappa - 1$, we have

$$\mathrm{Var}(\bar{z}_T) = \frac{\mathrm{Var}(z)}{T} = \frac{(\kappa-1)m^2}{T}.$$

Since Lemma 5.40 gives $\nu_T = \log \bar{z}_T$, in light of (5.94) we can write

$$F(\nu_T) - F(\nu_*) = \frac{m}{\bar{z}_T} - 1 + \log\Big(\frac{\bar{z}_T}{m}\Big) = \frac{1}{Q_T} + \log Q_T - 1, \qquad Q_T := \frac{\bar{z}_T}{m}. \tag{5.111}$$

Note that $\mathbb{E}[Q_T] = 1$ and $\mathrm{Var}(Q_T) = (\kappa-1)/T$.

Let $U_T := Q_T - 1 = (\bar{z}_T - m)/m$. Then $\mathbb{E}[U_T] = 0$ and $\mathbb{E}[U_T^2] = (\kappa - 1)/T$. Define

$$g(u) := \frac{1}{1+u} + \log(1+u) - 1, \forall u > -1$$

so that by (5.111) we have $F(\nu_T) - F(\nu_*) = g(U_T)$.

**Lemma 5.41** *For all $u \geq -\frac{1}{2}$,*

$$g(u) \leq 2u^2.$$

*Proof.* Define $h(u) := 2u^2 - g(u)$ for $u > -1$. Since $g'(u) = \frac{u}{(1+u)^2}$, we have

$$h'(u) = 4u - \frac{u}{(1+u)^2} = u\Big(4 - \frac{1}{(1+u)^2}\Big).$$

For $u \geq -\frac{1}{2}$, $(1+u)^2 \geq \frac{1}{4}$, hence $\frac{1}{(1+u)^2} \leq 4$. Therefore $h'(u) \leq 0$ for $u \in [-\frac{1}{2}, 0]$ and $h'(u) \geq 0$ for $u \geq 0$. Thus $h$ attains its minimum over $[-\frac{1}{2}, \infty)$ at $u = 0$, where $h(0) = 0$. Hence $h(u) \geq 0$ on $[-\frac{1}{2}, \infty)$, i.e., $g(u) \leq 2u^2$ there. □

**Lemma 5.42** *Let $z_i \geq 0$ i.i.d. with finite $\kappa$. Then*

$$\mathbb{P}(Q_T \leq 1/2) = \mathbb{P}(\bar{z}_T \leq m/2) \leq \exp\Big(-\frac{T}{8\kappa}\Big).$$

*Proof.* For any $\lambda > 0$, by Chernoff bound,

$$\mathbb{P}\Big(\sum_{i=1}^{T} z_i \leq \tfrac{Tm}{2}\Big) = \mathbb{P}\Big(e^{-\lambda \sum_{i=1}^{T} z_i} \geq e^{-\lambda Tm/2}\Big) \leq e^{\lambda Tm/2}\Big(\mathbb{E}[e^{-\lambda z}]\Big)^T.$$

Using $e^{-x} \leq 1 - x + x^2/2$ for $x \geq 0$,

$$\mathbb{E}[e^{-\lambda z}] \leq 1 - \lambda m + \frac{\lambda^2}{2}\mathbb{E}[z^2] \leq \exp\Big(-\lambda m + \frac{\lambda^2}{2}\mathbb{E}[z^2]\Big).$$

Therefore

$$\mathbb{P}(\bar{z}_T \leq m/2) \leq \exp\Big(T\Big(\lambda m/2 - \lambda m + \frac{\lambda^2}{2}\mathbb{E}[z^2]\Big)\Big) = \exp\Big(-T\Big(\frac{\lambda m}{2} - \frac{\lambda^2}{2}\mathbb{E}[z^2]\Big)\Big).$$

Choose $\lambda = m/(2\mathbb{E}[z^2])$ to get the exponent $-Tm^2/(8\mathbb{E}[z^2]) = -T/(8\kappa)$. □

**Lemma 5.43** *If $s$ is $\sigma^2$-subgaussian, then*

$$m^2\,\mathbb{E}[z^{-2}] \;=\; (\mathbb{E}[e^s])^2\,\mathbb{E}[e^{-2s}] \;\leq\; e^{3\sigma^2}.$$

*Proof.* Let $\mu = \mathbb{E}[s]$ and $X = s - \mu$. Then $\mathbb{E}[X] = 0$ and $z = e^s = e^\mu e^X$. Thus

$$m^2\mathbb{E}[z^{-2}] = \big(e^\mu \mathbb{E}[e^X]\big)^2 \cdot \big(e^{-2\mu}\mathbb{E}[e^{-2X}]\big) = \big(\mathbb{E}[e^X]\big)^2\,\mathbb{E}[e^{-2X}].$$

By subgaussianity,

$$\mathbb{E}[e^{X}] \le e^{\sigma^2/2}, \qquad \mathbb{E}[e^{-2X}] \le e^{(2^2)\sigma^2/2} = e^{2\sigma^2}.$$

Hence $m^2\mathbb{E}[z^{-2}] \le e^{\sigma^2} e^{2\sigma^2} = e^{3\sigma^2}$. □

**Theorem 5.18** *Under Assumption 5.20, the SPMD iterate $\nu_T$ produced by $\alpha_t = y_{t-1}/(t-1)$ satisfies*

$$\mathbb{E}\big[F(\nu_T) - F(\nu_*)\big] \;\le\; \frac{2(\kappa-1)}{T} \;+\; e^{\frac{3}{2}\sigma^2} \exp\Big(-\frac{T}{16\kappa}\Big). \tag{5.112}$$

*In particular, since the second term is exponentially small in $T/\kappa$,*

$$\mathbb{E}\big[F(\nu_T) - F(\nu_*)\big] = O(\kappa/T),$$

*for every $\sigma^2$-subgaussian $s(\zeta)$.*

*Proof.* Since $F(\nu_T) - F(\nu_*) = g(U_T)$, we split the expectation on the events $\{U_T \ge -1/2\}$ and $\{U_T < -1/2\}$:

$$\mathbb{E}[g(U_T)] = \mathbb{E}[g(U_T)\mathbf{1}\{U_T \ge -1/2\}] + \mathbb{E}[g(U_T)\mathbf{1}\{U_T < -1/2\}].$$

On $\{U_T \ge -1/2\}$, Lemma 5.41 yields

$$\mathbb{E}[g(U_T)\mathbf{1}\{U_T \ge -1/2\}] \le 2\,\mathbb{E}[U_T^2] = 2\,\mathrm{Var}(Q_T) = 2\,\frac{\mathrm{Var}(z)}{m^2 T} = \frac{2(\kappa-1)}{T}.$$

On $\{U_T < -1/2\}$ we have $Q_T \le 1/2$, and since $\log Q_T - 1 \le 0$,

$$g(U_T) = \frac{1}{Q_T} + \log Q_T - 1 \le \frac{1}{Q_T}.$$

Hence, by Cauchy–Schwarz,

$$\mathbb{E}[g(U_T)\mathbf{1}\{U_T < -1/2\}] \le \mathbb{E}[Q_T^{-1}\mathbf{1}\{Q_T \le 1/2\}] \le \big(\mathbb{E}[Q_T^{-2}]\big)^{1/2}\,\mathbb{P}(Q_T \le 1/2)^{1/2}.$$

By Jensen inequality and Lemma 5.43,

$$\mathbb{E}[Q_T^{-2}] = m^2\,\mathbb{E}[\bar{z}_T^{-2}] \le m^2\,\mathbb{E}[z^{-2}] \le e^{3\sigma^2}.$$

By Lemma 5.42, $\mathbb{P}(Q_T \le 1/2) \le \exp(-T/(8\kappa))$. Therefore,

$$\mathbb{E}[g(U_T)\mathbf{1}\{U_T < -1/2\}] \le e^{\frac{3}{2}\sigma^2} \exp\Big(-\frac{T}{16\kappa}\Big).$$

Combining the two pieces proves (5.112). □

### A Distribution-free lower bound

Indeed, we can show that $O\left(\frac{\kappa-1}{T}\right)$ is an optimal bound by establishing matching a lower bound for a black-box oracle model where the underlying distribution of $z$ is unknown and for any query $\nu$ the oracle returns

$$\Phi(\nu;\zeta) = ze^{-\nu} + \nu, \qquad g(\nu;\zeta) = \nabla_\nu \Phi(\nu;\zeta) = 1 - ze^{-\nu}.$$

Since

$$z(\zeta) = e^{\nu}(\Phi(\nu;\zeta) - \nu) = e^{\nu}(1 - g(\nu;\zeta)),$$

hence, any $T$-query algorithm can reconstruct $T$ i.i.d. samples $z_1, \dots, z_T$ from $P$. Thus, it suffices to prove the lower bound in the standard i.i.d. sampling model for $z$.

Let us define a distribution class. For $\kappa \geq 2$, define

$$\mathcal{P}_\kappa \coloneqq \left\{ P :\ z \geq 0,\ 0 < \mathbb{E}_P[z] < \infty,\ \frac{\mathbb{E}_P[z^2]}{(\mathbb{E}_P[z])^2} \leq \kappa \right\}.$$

Equivalently, $\mathrm{Var}_P(z)/(\mathbb{E}_P[z])^2 \leq \kappa - 1$. For $P \in \mathcal{P}_\kappa$ let $m(P) = \mathbb{E}_P[z]$ and $\nu_*(P) = \log m(P)$.

**Lemma 5.44** *Let $\phi(u) \coloneqq e^{-u} + u - 1$. Then $\phi(0) = \phi'(0) = 0$ and $\phi''(u) = e^{-u}$. In particular, for all $|u| \leq 1$,*

$$\phi(u) \ \geq\ \frac{e^{-1}}{2}\, u^2. \tag{5.113}$$

*Proof.* On the interval $[-1, 1]$, $\phi''(u) = e^{-u} \geq e^{-1}$, so $\phi$ is $e^{-1}$-strongly convex on $[-1, 1]$. Since $\phi(0) = \phi'(0) = 0$, strong convexity implies $\phi(u) \geq \frac{e^{-1}}{2} u^2$ for all $|u| \leq 1$. □

**Lemma 5.45** *Let $\phi(u) = e^{-u} + u - 1$. Fix $\nu_0 < \nu_1$ and let $\Delta \coloneqq \nu_1 - \nu_0$. Define*

$$H(\nu) \coloneqq \phi(\nu - \nu_0) + \phi(\nu - \nu_1).$$

*Then $H$ is strictly convex and its unique minimizer $\nu^\dagger$ lies in $(\nu_0, \nu_1)$. Moreover, if $\Delta \leq 1$, then*

$$\inf_{\nu \in \mathbb{R}} H(\nu) \ \geq\ \frac{e^{-1}}{4}\, \Delta^2. \tag{5.114}$$

*Proof.* We have $\phi'(u) = 1 - e^{-u}$ and $\phi''(u) = e^{-u} > 0$, hence $H$ is strictly convex with

$$H'(\nu) = \phi'(\nu - \nu_0) + \phi'(\nu - \nu_1) = 2 - e^{-(\nu - \nu_0)} - e^{-(\nu - \nu_1)}.$$

At the endpoints,

$$H'(\nu_0) = 2 - 1 - e^{-(\nu_0 - \nu_1)} = 1 - e^{\Delta} < 0, \qquad H'(\nu_1) = 2 - e^{-(\nu_1 - \nu_0)} - 1 = 1 - e^{-\Delta} > 0.$$

Since $H'$ is strictly increasing (because $H'' > 0$), there is a unique root $\nu^\dagger \in (\nu_0, \nu_1)$ and thus $\inf_{\nu\in\mathbb{R}} H(\nu) = \inf_{\nu\in[\nu_0,\nu_1]} H(\nu)$.

Assume $\Delta \le 1$. Then for all $\nu \in [\nu_0, \nu_1]$ we have $|\nu - \nu_0| \le \Delta \le 1$ and $|\nu - \nu_1| \le \Delta \le 1$. On $[-1, 1]$, $\phi''(u) = e^{-u} \ge e^{-1}$, so $\phi(u) \ge \frac{e^{-1}}{2}u^2$ for all $|u| \le 1$. Therefore, for all $\nu \in [\nu_0, \nu_1]$,

$$H(\nu) \ge \frac{e^{-1}}{2}\big((\nu - \nu_0)^2 + (\nu - \nu_1)^2\big).$$

Minimizing the RHS over $\nu$ yields $\inf_\nu\big((\nu - \nu_0)^2 + (\nu - \nu_1)^2\big) = \Delta^2/2$, hence $\inf_{\nu\in\mathbb{R}} H(\nu) \ge \frac{e^{-1}}{4}\Delta^2$. □

**Lemma 5.46 (Le Cam's Two-point Method)** *Let $P_0, P_1$ be two distributions and let $L_0(\cdot), L_1(\cdot)$ be nonnegative loss functions. For any estimator $\widehat{a}$ measurable w.r.t. the data,*

$$\max\{\mathbb{E}_{P_0}[L_0(\widehat{a})],\ \mathbb{E}_{P_1}[L_1(\widehat{a})]\} \ \ge\ \frac{1 - \mathrm{TV}(P_0, P_1)}{2}\ \inf_a\big(L_0(a) + L_1(a)\big). \quad (5.115)$$

*Proof.* Let $M \coloneqq (P_0 + P_1)/2$ and write $dP_0 = (1 + f)\,dM$, $dP_1 = (1 - f)\,dM$ where $|f| \le 1$ and $\int |f|\,dM = \mathrm{TV}(P_0, P_1)$. Then for any (possibly random) decision $A$,

$$\begin{aligned}
\mathbb{E}_{P_0}[L_0(A)] + \mathbb{E}_{P_1}[L_1(A)] &= \int \Big(L_0(A)(1 + f) + L_1(A)(1 - f)\Big)\,dM \\
&= \int \Big((L_0(A) + L_1(A)) + f(L_0(A) - L_1(A))\Big)\,dM \\
&\ge \int \Big((L_0(A) + L_1(A)) - |f|\,(L_0(A) + L_1(A))\Big)\,dM \\
&= \int (L_0(A) + L_1(A))(1 - |f|)\,dM \\
&\ge \inf_a(L_0(a) + L_1(a)) \int (1 - |f|)\,dM \\
&= (1 - \mathrm{TV}(P_0, P_1))\inf_a(L_0(a) + L_1(a)).
\end{aligned}$$

Taking half and using $\max\{x, y\} \ge (x + y)/2$ yields (5.115). □

The final distribution-free suboptimality lower bound is stated in the following theorem.

**Theorem 5.19** *Let $z = e^{s(\zeta)} \ge 0$ with $m(P) = \mathbb{E}_P[z]$ and $\nu_*(P) = \log m(P)$. For $\kappa \ge 2$, define*

$$\mathcal{P}_\kappa \coloneqq \left\{P :\ z \ge 0,\ 0 < \mathbb{E}_P[z] < \infty,\ \frac{\mathbb{E}_P[z^2]}{\mathbb{E}_P[z]^2} \le \kappa\right\}.$$

*Let $F_P(\nu) \coloneqq m(P)e^{-\nu} + \nu$ and $\nu_*(P) = \arg\min_\nu F_P(\nu)$. Then there exists an absolute constant $c > 0$ such that for all $T \ge \kappa$, any (possibly adaptive) algorithm using*

*$T$ value/gradient oracle calls and outputting $\widehat{\nu}$ satisfies*

$$\sup_{P\in\mathcal{P}_\kappa} \mathbb{E}_P[F_P(\widehat{\nu}) - F_P(\nu_*(P))] \;\geq\; c\,\frac{\kappa-1}{T}. \tag{5.116}$$

*Proof.* We construct two strictly positive hard instances in $\mathcal{P}_\kappa$. Fix $\varepsilon \in (0,1]$ and define two distributions supported on $\{\varepsilon, \kappa\}$:

$$P_i^\varepsilon: \quad \mathbb{P}(z=\kappa) = p_i, \qquad \mathbb{P}(z=\varepsilon) = 1-p_i, \qquad i\in\{0,1\},$$

where

$$p_0 \coloneqq \frac{1}{\kappa}, \qquad p_1 \coloneqq p_0 + h, \qquad h \coloneqq \frac{1}{8\sqrt{\kappa T}}.$$

Since $T \geq \kappa$, we have $h \leq \frac{1}{8\kappa}$ so $p_1 \in (0,1)$.

Next we show that $P_0^\varepsilon, P_1^\varepsilon \in \mathcal{P}_\kappa$. For a generic $p \in (0,1)$ and support $\{\varepsilon,\kappa\}$, define

$$R_\varepsilon(p) \coloneqq \frac{\mathbb{E}[z^2]}{\mathbb{E}[z]^2} = \frac{p\kappa^2 + (1-p)\varepsilon^2}{\big(p\kappa + (1-p)\varepsilon\big)^2}.$$

Let $u \coloneqq \varepsilon/\kappa \in (0, 1/\kappa] \subset (0,1]$. Then

$$R_\varepsilon(p) = \frac{p + (1-p)u^2}{\big(p + (1-p)u\big)^2}.$$

We claim $R_\varepsilon(p) \leq \frac{1}{p}$ for all $u \in [0,1]$. Indeed,

$$\begin{aligned}
&\big(p+(1-p)u\big)^2 - p\big(p+(1-p)u^2\big) \\
&= p^2 + 2p(1-p)u + (1-p)^2u^2 - p^2 - p(1-p)u^2 \\
&= (1-p)u\Big(2p + (1-2p)u\Big) \;\geq\; 0,
\end{aligned}$$

because $u \in [0,1]$ and $2p + (1-2p)u \geq \min\{2p, 1\} \geq 0$. Thus $R_\varepsilon(p) \leq 1/p$. Since $p_0 = 1/\kappa$ and $p_1 \geq p_0$, we have $1/p_i \leq \kappa$, hence $R_\varepsilon(p_i) \leq \kappa$ and therefore $P_0^\varepsilon, P_1^\varepsilon \in \mathcal{P}_\kappa$.

Next, we compute the separation $\Delta$ between $\nu_*$'s. Let $m_i^\varepsilon = \mathbb{E}_{P_i^\varepsilon}[z] = \varepsilon + p_i(\kappa-\varepsilon)$ and $\nu_i^\varepsilon = \log m_i^\varepsilon$. Then

$$m_1^\varepsilon - m_0^\varepsilon = h(\kappa-\varepsilon) \geq h(\kappa-1), \qquad m_0^\varepsilon = \varepsilon + p_0(\kappa-\varepsilon) = 1 + \Big(1-\frac{1}{\kappa}\Big)\varepsilon \in [1,2].$$

Hence

$$\Delta \coloneqq |\nu_1^\varepsilon - \nu_0^\varepsilon| = \log\!\left(1 + \frac{m_1^\varepsilon - m_0^\varepsilon}{m_0^\varepsilon}\right) \geq \frac{1}{2}\cdot\frac{h(\kappa-1)}{2} = \frac{\kappa-1}{32\sqrt{\kappa T}},$$

where we used $\log(1+x) \geq x/2$ for $x \in [0, 1/2]$ and the fact that $\frac{h(\kappa-\varepsilon)}{m_0^\varepsilon} \leq h\kappa \leq 1/8$. In particular, $\Delta \leq h\kappa \leq 1/8 < 1$.

Next, we show the lower bound of $\inf_\nu \Big( (F_0(\nu) - F_0(\nu_0^\varepsilon)) + (F_1(\nu) - F_1(\nu_1^\varepsilon)) \Big)$. Under $P_i^\varepsilon$ the objective is $F_i(\nu) = m_i^\varepsilon e^{-\nu} + \nu$ and the optimal value is $F_i(\nu_i^\varepsilon) = 1 + \nu_i^\varepsilon$. Thus the suboptimality can be written as

$$F_i(\nu) - F_i(\nu_i^\varepsilon) = e^{\nu_i^\varepsilon - \nu} + (\nu - \nu_i^\varepsilon) - 1 = \phi(\nu - \nu_i^\varepsilon), \qquad \phi(u) = e^{-u} + u - 1.$$

Let $\nu_0^\varepsilon < \nu_1^\varepsilon$ and set $u = \nu - \nu_0^\varepsilon$. Then

$$\phi(\nu - \nu_0^\varepsilon) + \phi(\nu - \nu_1^\varepsilon) = \phi(u) + \phi(u - \Delta).$$

The function $u \mapsto \phi(u) + \phi(u - \Delta)$ is convex and its minimizer lies in $[0, \Delta]$. Since $\Delta \leq 1$, applying Lemma 5.45 gives

$$\phi(u) + \phi(u - \Delta) \;\geq\; \frac{e^{-1}}{4}\Delta^2.$$

Therefore,

$$\inf_\nu \Big( (F_0(\nu) - F_0(\nu_0^\varepsilon)) + (F_1(\nu) - F_1(\nu_1^\varepsilon)) \Big) \;\geq\; \frac{e^{-1}}{4}\Delta^2. \tag{5.117}$$

Next, we show the total variation between $P_0^\varepsilon$, and $P_1^\varepsilon$ is bounded. Because the two distributions differ only in the Bernoulli parameter,

$$\mathrm{KL}(P_0^\varepsilon, P_1^\varepsilon) = p_0 \log \frac{p_0}{p_1} + (1 - p_0) \log \frac{1 - p_0}{1 - p_1}.$$

Using the bound $\mathrm{KL}(P, Q) \leq \chi^2(P, Q)$ and the fact that for Bernoulli measures $\chi^2(P_0^\varepsilon, P_1^\varepsilon) = \frac{h^2}{p_1(1-p_1)}$, we get

$$\mathrm{KL}(P_0^\varepsilon, P_1^\varepsilon) \leq \frac{h^2}{p_1(1 - p_1)}.$$

Since $h \leq \frac{1}{2\kappa}$, we have $p_1 \leq p_0 + h \leq \frac{3}{2\kappa} \leq \frac{3}{4}$, hence $1 - p_1 \geq 1/4$, and also $p_1 \geq p_0 = 1/\kappa$. Therefore $p_1(1 - p_1) \geq \frac{1}{4\kappa}$ and

$$\mathrm{KL}(P_0^\varepsilon, P_1^\varepsilon) \leq 4\kappa h^2.$$

For $T$ i.i.d. samples, this gives

$$\mathrm{KL}\big((P_0^\varepsilon)^{\otimes T}, (P_1^\varepsilon)^{\otimes T}\big) = T\,\mathrm{KL}(P_0^\varepsilon, P_1^\varepsilon) \leq 4\kappa T h^2 = \frac{1}{16}.$$

By Pinsker's inequality,

$$\mathrm{TV}\big((P_0^\varepsilon)^{\otimes T}, (P_1^\varepsilon)^{\otimes T}\big) \le \sqrt{\frac{1}{2}\mathrm{KL}\big((P_0^\varepsilon)^{\otimes T}, (P_1^\varepsilon)^{\otimes T}\big)} \le \sqrt{\frac{1}{32}} \le \frac{1}{4}.$$

Finally, we apply Lemma 5.46 to $P_0 = (P_0^\varepsilon)^{\otimes T}$, $P_1 = (P_1^\varepsilon)^{\otimes T}$ and losses

$$L_i(\nu) \coloneqq F_i(\nu) - F_i(\nu_i^\varepsilon) \ge 0.$$

Using (5.117) and $\mathrm{TV} \le 1/4$ yields for any estimator $\widehat{\nu}$,

$$\max_{i\in\{0,1\}} \mathbb{E}_{P_i^\varepsilon}\big[F_i(\widehat{\nu}) - F_i(\nu_i^\varepsilon)\big] \ge \frac{1-\mathrm{TV}}{2}\cdot\frac{e^{-1}}{4}\Delta^2 \ge \frac{3}{8}\cdot\frac{e^{-1}}{4}\Delta^2 = \frac{3e^{-1}}{32}\Delta^2.$$

Substituting $\Delta^2 \ge \frac{(\kappa-1)^2}{1024\,\kappa\,T} \ge \frac{\kappa-1}{2048\,T}$ (since $\kappa \ge 2$) gives

$$\max_{i\in\{0,1\}} \mathbb{E}_{P_i^\varepsilon}\big[F_i(\widehat{\nu}) - F_i(\nu_i^\varepsilon)\big] \;\ge\; \frac{3}{65536\,e}\cdot\frac{\kappa-1}{T}.$$

Since $P_0^\varepsilon, P_1^\varepsilon \in \mathcal{P}_\kappa$, this implies (5.116) with $c = \frac{3}{65536\,e}$. □

## 5.6 History and Notes

Finite-sum coupled compositional optimization (FCCO) was first formalized in our work (Qi et al., 2021c) for optimizing average precision, an empirical estimator of the area under the precision–recall curve. We proposed the SOAP algorithm for AP maximization and established the first complexity bound of $O\left(\frac{n}{\epsilon^5}\right)$ for finding an $\epsilon$-stationary solution. Their algorithm is closely related to SOX, but differs in that it does not employ a moving-average gradient estimator. The framework was demonstrated on applications including image classification and molecular property prediction for drug discovery. The analysis of SOAP draws inspiration from the original SCGD analysis Wang et al. (2017a), while significantly improving upon its $O(1/\epsilon^8)$ complexity with the a better hyper-parameter setting, leading to Theorem 4.1.

To accelerate convergence, we subsequently adopted the moving average gradient estimator for FCCO (Wang et al., 2022). While this approach achieves a complexity order of $O\left(\frac{n}{B\epsilon^4}\right)$, it does not benefit from the variance reduction gained by using mini-batches to estimate inner function values. The limitation arises because we treat all inner functions as a single vector variable and compute a sparse unbiased stochastic estimator for this vector; consequently, the estimator does not enjoy the advantages of inner mini-batching. This improved rate and analysis was inspired by the stochastic compositional momentum method (Ghadimi et al., 2020).

Subsequently, we proposed the SOX algorithm-a significant advancement for solving FCCO (Wang and Yang, 2022), encompassing new design, theoretical analysis, and practical applications. In that work, we established a complexity of $O\left(\frac{n\sigma_0^2}{B\epsilon^4}\right)$

for SOX to find an $\epsilon$-stationary solution in non-convex smooth FCCO problems. It integrates the analysis of stochastic block coordinate update of the $\mathbf{u}$ sequences with that of stochastic compositional momentum method.

Building on this, we developed a double-loop restarted algorithm that utilizes SOX in the inner loop to address non-convex problems under the $\mu$-PL (Polyak-Lojasiewicz) condition, i.e., $\|\nabla F(\mathbf{w})\|_2^2 \geq \mu(F(\mathbf{w}) - \min_{\mathbf{w}} F(\mathbf{w}))$. This approach yields an improved complexity of $O\left(\frac{n\sigma_0^2}{\mu^2 B\epsilon}\right)$ for finding an $\epsilon$-optimal solution. This result further implies a complexity of $O\left(\frac{n\sigma_0^2}{\mu^2 B\epsilon}\right)$ for strongly convex FCCO problems and $O\left(\frac{n\sigma_0^2}{B\epsilon^3}\right)$ for convex FCCO problems, requiring no assumptions on the individual convexity of inner and outer functions beyond the overall convexity of the objective. The improved convergence analysis under the PL condition for the double-loop restarted algorithm was inspired by our prior work on stochastic compositional optimization for distributionally robust learning (Qi et al., 2021b). A comparable complexity bound of $O(\frac{1}{\mu^2\epsilon})$ for a single-loop algorithm in the context of Stochastic Convex Optimization (SCO) under the PL condition was subsequently established in (Jiang et al., 2023), which considers the application of SCO in training energy-based models.

Furthermore, for convex FCCO instances where the outer function is both convex and monotonically non-decreasing and the inner functions are convex, (Wang and Yang, 2022) reformulated the problem as a convex-concave min-max optimization problem and established a complexity of $O\left(\frac{n\sigma_0^2}{B\epsilon^2}\right)$ under a weak duality convergence measure. Finally, when a $\mu$-strongly convex regularizer is present, the complexity is further refined to $O\left(\frac{n\sigma_0^2}{\mu^2 B\epsilon}\right)$ for finding an $\epsilon$-optimal solution in terms of Euclidean distance to the optimum. This analysis was mostly inspired by (Zhang and Lan, 2024), which is the first work that establishes the optimal complexity for solving convex SCO where the outer function is both convex and monotonically non-decreasing and the inner function is convex.

Later, Jiang et al. (2022) proposed the Multi-Block-Single-Probe Variance Reduction (MSVR) algorithm for FCCO, establishing improved complexity bounds over SOX by leveraging the mean squared smoothness of the inner functions. For non-convex smooth FCCO problems, MSVR improves the complexity to $O\left(\frac{n\sigma_0}{B\epsilon^3}\right)$ for identifying an $\epsilon$-stationary solution.

For objectives satisfying the $\mu$-PL condition, a double-loop restarted MSVR algorithm achieves an improved complexity of $O\left(\frac{n\sigma_0}{\mu B\epsilon}\right)$ to find an $\epsilon$-optimal solution. Consequently, this approach yields a complexity of $O\left(\frac{n\sigma_0}{\mu B\epsilon}\right)$ for strongly convex FCCO problems and $O\left(\frac{n\sigma_0}{B\epsilon^2}\right)$ for convex FCCO problems.

The analysis for non-smooth weakly convex FCCO and the SONX (v2) algorithm was studied in our work (Hu et al., 2024b). This work established a complexity of $O\left(\frac{n\sigma_0}{B\epsilon^6}\right)$ for finding a nearly $\epsilon$-stationary solution for weakly convex inner and outer

functions. A similar analysis for a special case of weakly-convex SCO was conducted in (Zhu et al., 2023c). When the outer function is smooth, the complexity is improved in this book to $O\left(\frac{n\sigma_0}{B\epsilon^4}\right)$. The SONEX algorithm for solving weakly convex FCCO with non-smooth outer functions was proposed in our work (Chen et al., 2025b).

The ALEXR algorithm and its analysis for convex FCCO instances appeared in our work (Wang and Yang, 2023), where the outer function is both convex and monotonically non-decreasing and the inner functions are convex. For the first time, we established a complexity of $O\left(\frac{n\sigma_0^2}{B\epsilon^2}\right)$ for finding an $\epsilon$-optimal solution of convex FCCO. Our analysis of the stochastic block coordinate update for the dual variables is primarily informed by the framework in Alacaoglu et al. (2025), which addresses convex-concave minimax problems with bilinear structures. The extrapolation for the gradient of the dual variable is inspired by (Zhang et al., 2021). It is worth mentioning that for strongly convex FCCO with smooth outer functions, we only established the convergence of ALEXR for the Euclidean distance to the optimum. However, it is possible to establish the convergence for the objective gap and even the duality gap following our work on strongly-convex strongly-concave min-max optimization (Yan et al., 2020b).

In (Wang and Yang, 2023), we also established the lower bounds for convex FCCO and strongly convex FCCO, which matches the upper bounds. Our derivation of the lower bound for convex FCCO with non-smooth outer functions builds upon the construction presented in (Zhang and Lan, 2024) for SCO.

The double-loop ALEXR was developed in Chen et al. (2025b), which was mostly inspired by a line of work on weakly-convex concave min-max problems (Rafique et al., 2018; Yan et al., 2020b; Zhang et al., 2022). (Rafique et al., 2018) is the first work that proves the convergence for weakly-convex (strongly)-concave problems. Yan et al. (2020b) simplified the algorithm for weakly-convex strongly-concave problems with $\mu$-strong concavity on the dual variable and established a complexity of $O(\frac{1}{\mu^2\epsilon^4})$ for finding an nearly $\epsilon$-stationary point. The later work (Zhang et al., 2022) improved the complexity to $O(\frac{1}{\mu\epsilon^4})$ with a simple change on the number of iteration for the inner loop.

The non-convex analysis of ASGD for compositional CVaR minimization first appeared in (Zhu et al., 2022b) for one-way partial AUC optimization. The geometric-aware algorithm SCENT for CERM and its analysis were developed in (Wei et al., 2026). It remains an interesting problem to conduct fine-grained analysis of SCENT for non-convex problems.

A more general framework than FCCO is the so-called conditional stochastic optimization (CSO), defined as:

$$\min_{\mathbf{w}} \mathbb{E}_{\xi}\left[f_{\xi}\left(\mathbb{E}_{\zeta|\xi}[g(\mathbf{w};\zeta,\xi)]\right)\right].$$

This paradigm was formally introduced by Hu et al. (2020), who analyzed a biased SGD (BSGD) algorithm employing a large inner mini-batch and a constant outer mini-batch. For non-convex smooth problems, using an inner batch size of $O(\epsilon^{-2})$ results in an iteration complexity of $O(\epsilon^{-4})$, which translates to a total sample com-

plexity of $O(\epsilon^{-6})$. This performance is inferior to that of SOX when $n/B < \epsilon^{-2}$. For convex and $\mu$-strongly convex CSO problems, an inner batch size of $O(\epsilon^{-1})$ yields iteration complexities of $O(\epsilon^{-2})$ and $O(\mu^{-2}\epsilon^{-1})$, respectively. Notably, the latter complexity is likewise worse than that of restarted SOX when $n/B < O(\epsilon^{-1})$.