Chapter 4
Foundations: Stochastic Compositional Optimization

Abstract In this chapter, we introduce stochastic compositional optimization problems and their optimization algorithms, including stochastic compositional gradient descent and stochastic compositional momentum metholds. We also consider extensions of these techniques to structured optimization with compositional gradient including non-convex regularized problems, min-max optimization, min-min optimization and bilevel optimization. We focus on the complexity of these metholds for non-convex optimization.

Moving average is the core ingredient!

Contents

4.1 Stochastic Compositional Optimization

We have seen several advanced machine learning frameworks in the Chapter 2, including DRO, GDRO, EXM, and COCE. Unfortunately, existing stochastic gradient metholds such as SGD are not directly applicable to these new problems. The reason will become clear shortly. To address this challenge, we need new optimization tools.

In this chapter, we will consider a family of stochastic optimization problems called **stochastic compositional optimization (SCO)**, whose objective is given by

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) := \mathbb{E}_\xi f(\mathbb{E}_\zeta g(\mathbf{w};\zeta);\xi) \tag{4.1}$$

where ξ and ζ are random variables, $g(\cdot;\zeta) : \mathbb{R}^d \to \mathbb{R}^{d'}$ is the inner random function, and $f(\cdot;\xi) : \mathbb{R}^{d'} \to \mathbb{R}$ is the outer random function. Let $f(\cdot) = \mathbb{E}_\xi f(\cdot;\xi)$ and $g(\cdot) = \mathbb{E}_\zeta g(\cdot;\zeta)$. Then the objective function $F(\mathbf{w}) = f(g(\mathbf{w}))$ is a composition of two functions.

Examples

Example 4.1. *The KL-regularized DRO (2.14) is a special case of SCO by setting* $f(\cdot) = \lambda\log(\cdot)$ *and* $g(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \exp(\ell(\mathbf{w};\mathbf{x}_i,y_i)/\lambda)$.

Example 4.2. *The KL-constrained DRO (2.19) is a special case of SCO by setting* $\bar{g} = (g_1, g_2)$ $f(\bar{g}) = g_1\log(g_2) + g_1\rho$ *and* $g_1(\mathbf{w},\lambda) = \lambda, g_2(\mathbf{w},\lambda) = \frac{1}{n}\sum_{i=1}^n \exp(\ell(\mathbf{w};\mathbf{x}_i,y_i)/\lambda)$.

Example 4.3. *The compositional objective for AUC maximization (2.32) has a compositional term of* $f(g(\mathbf{w}))$*, where* $g(\mathbf{w})$ *is a stochastic function and* f *is a deterministic function.*

Optimization Challenge

The challenge of solving SCO lies in how to estimate the gradient $\nabla F(\mathbf{w}) = \nabla g(\mathbf{w})\nabla f(g(\mathbf{w}))$, where $\nabla g(\mathbf{w}) \in \mathbb{R}^{d\times d'}$ denotes the transpose of the Jacobian matrix of g at $\mathbf{w}$ and $\nabla f(g) \in \mathbb{R}^{d'}$ is a gradient of f at g.

A simple way of estimating the gradient is by using stochastic samples, i.e., $G(\mathbf{w};\xi,\zeta,\zeta') = \nabla g(\mathbf{w};\zeta)\nabla f(g(\mathbf{w};\zeta');\xi)$, where ξ,ζ,ζ' are random samples. One can also use mini-batch of random samples to compute the estimator. However, the problem is that $G(\mathbf{w};\xi,\zeta,\zeta')$ is a biased estimator when f is non-linear, i.e., $\mathbb{E}_{\xi,\zeta,\zeta'}G(\mathbf{w};\xi,\zeta,\zeta') \neq \nabla F(\mathbf{w})$. This will break all assumptions made in the convergence analysis in Chapter 3. Directly using this estimator in SGD could result in non-convergence or it requires a large batch size for estimating $g(\mathbf{w})$.

Algorithm 9 SCGD

Input: learning rate schedules $\{\eta_t\}_{t=1}^T$, $\{\gamma_t\}_{t=1}^T$; starting points $\mathbf{w}_1$, $\mathbf{u}_0$
for $t = 1, \ldots, T$ **do**
  Sample ζ_t, ζ_t' and ξ_t
  Compute the inner function value estimator $\mathbf{u}_t = (1-\gamma_t)\mathbf{u}_{t-1} + \gamma_t g(\mathbf{w}_t; \zeta_t)$
  Compute the vanilla gradient estimator $\mathbf{z}_t = \nabla g(\mathbf{w}_t; \zeta_t')\nabla f(\mathbf{u}_t; \xi_t)$
  Update the model $\mathbf{w}$ by $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{z}_t$
end for

4.2 Stochastic Compositional Gradient Descent

We assume both f and g are differentiable. Next, we introduce stochastic compositional gradient descent (SCGD) as a solution method for SCO. The key to the design is to track the sequence of $\{g(\mathbf{w}_t), t = 1, \ldots, T\}$ by a sequence of estimators $\{\mathbf{u}_t, t = 1, \ldots, T\}$. Let us consider the following problem:

$$\min_u \frac{1}{2}\|\mathbf{u} - g(\mathbf{w}_t)\|_2^2. \tag{4.2}$$

We compute $\mathbf{u}_t$ by using the SGD update:

$$\mathbf{u}_t = \mathbf{u}_{t-1} - \gamma_t(\mathbf{u}_{t-1} - g(\mathbf{w}_t; \zeta_t)) = (1-\gamma_t)\mathbf{u}_{t-1} + \gamma_t g(\mathbf{w}_t; \zeta_t), t \in [T], \tag{4.3}$$

where $g(\mathbf{w}; \zeta)$ is stochastic estimator of $g(\mathbf{w})$ such that $\mathbb{E}_\zeta[g(\mathbf{w}; \zeta)] = g(\mathbf{w})$. The update is also known as moving average sequence of $g(\mathbf{w}_t)$.

The intuition behind this is that when $\mathbf{w}_t$ converges (i.e., $\mathbf{w}_t - \mathbf{w}_{t-1} \to 0$), $\mathbf{u}_t$ is a better estimator of $g(\mathbf{w}_t)$ than $g(\mathbf{w}_t; \zeta_t)$. With $\mathbf{u}_t$, the gradient estimator can be computed by

$$\mathbf{z}_t = \nabla g(\mathbf{w}_t; \zeta_t')\nabla f(\mathbf{u}_t; \xi_t), \tag{4.4}$$

where ζ_t' is another independent random variable. Then, we can use it for updating $\mathbf{w}_t$:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla g(\mathbf{w}_t; \zeta_t')\nabla f(\mathbf{u}_t; \xi_t).$$

The detailed steps are presented in Algorithm 9.

Critical: Using ζ_t' instead of ζ_t in computing $\nabla g(\mathbf{w}_t; \zeta_t')$ is for simplicity of analysis, which decouple the dependence between $\mathbf{u}_t$ and ζ_t' as $\mathbf{u}_t$ depends on ζ_t. However, this will increase the number of random samples per-iteration. For practical implementation, one may just use $\zeta_t' = \zeta_t$.

4.2.1 Convergence Analysis

We make the following assumptions regarding the SCO problem (4.1).

Assumption 4.1. *There exist* $L_1, G_1 > 0$ *such that*

(i) f *is* L_1*-smooth, i.e.,* $\|\nabla f(g) - \nabla f(g')\|_2 \leq L_1 \|g - g'\|_2, \forall g, g'$*;*
(ii) $\mathbb{E}[\|\nabla f(g;\xi)\|_2^2] \leq G_1^2, \forall g.$

Assumption 4.2. *There exist* $G_2 > 0$ *such that* $\mathbb{E}[\|\nabla g(\mathbf{w};\zeta)\|_2^2] \leq G_2^2, \forall \mathbf{w}.$

Due to Jensen's inequality, $\mathbb{E}[\|\nabla f(\cdot;\xi)\|_2^2] \leq G_1^2$, and $\mathbb{E}[\|\nabla g(\mathbf{w};\zeta)\|_2^2] \leq G_2^2$ indicate the G_1-Lipschitz condition of f and G_2-Lipschitz condition of g, respectively.

Assumption 4.3. *There exist* $\sigma_0, \sigma_1, \sigma_2 > 0$ *such that*

(i) $\mathbb{E}[\|g(\mathbf{w};\zeta) - g(\mathbf{w})\|_2^2] \leq \sigma_0^2, \forall \mathbf{w}$*;*
(ii) $\mathbb{E}[\|\nabla f(g;\xi) - \nabla f(g)\|_2^2] \leq \sigma_1^2, \quad \mathbb{E}[\|\nabla g(\mathbf{w};\zeta) - \nabla g(\mathbf{w})\|_2^2] \leq \sigma_2^2, \forall \mathbf{w}, g.$
(iii) $F_* = \min_{\mathbf{w}} F(\mathbf{w}) > -\infty.$

Assumption 4.4. F *is* L_F*-smooth, i.e., there exist* $L_F > 0$ *such that* $\nabla F(\cdot)$ *is* L_F*-Lipschitz continuous.*

It is notable that the smoothness of F does not necessarily imply that g is smooth. One example is that if $g(\mathbf{w}) = \|\mathbf{w}\|_2$ and $f(g) = g^2$, the overall function $F(\mathbf{w}) = \|\mathbf{w}\|_2^2$ is smooth but the inner function g is non-smooth.

Lemma 4.1 *Under Assumptions 4.2 and 4.3(i), the* $\{\mathbf{u}_t\}_{t\geq 1}$ *sequence (4.3) satisfies that*

$$\mathbb{E}_{\zeta_t}\left[\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2\right] \leq (1-\gamma_t)\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2^2 + \gamma_t^2\sigma_0^2 + \frac{G_2^2}{\gamma_t}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2. \tag{4.5}$$

where $\mathbb{E}_{\zeta_t}$ *denotes the expectation over* ζ_t *given all previous randomness.*

Why it matters

The lemma admits an intuitive interpretation. The first term shows that $\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2$ is bounded by a contracting sequence. The second term is due to the noise in $g(\mathbf{w}_t;\zeta_t)$ and the third term is caused by the drifting from $\mathbf{w}_{t-1}$ to $\mathbf{w}_t$, both of which decay to zero under the conditions $\gamma_t^2 \to 0$ and $\eta_t^2/\gamma_t \to 0$, respectively.

Proof. In the following proof, we abuse the notation $\mathbb{E}_t$ to denote $\mathbb{E}_{\zeta_t}$. According to the update formula $\mathbf{u}_t = (1-\gamma_t)\mathbf{u}_{t-1} + \gamma_t g(\mathbf{w}_t;\zeta_t)$ we have

$$\begin{aligned}\mathbb{E}_t\left[\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2\right] &= \mathbb{E}_t\left[\|(1-\gamma_t)\mathbf{u}_{t-1} + \gamma_t g(\mathbf{w}_t;\zeta_t) - g(\mathbf{w}_t)\|_2^2\right] \\ &= \mathbb{E}_t\left[\|(1-\gamma_t)(\mathbf{u}_{t-1} - g(\mathbf{w}_t)) + \gamma_t(g(\mathbf{w}_t;\zeta_t) - g(\mathbf{w}_t))\|_2^2\right].\end{aligned}$$

Note that $\mathbb{E}_t\left[(\mathbf{u}_{t-1} - g(\mathbf{w}_t))^\top (g(\mathbf{w}_t;\zeta_t) - g(\mathbf{w}_t))\right] = 0$. Thus,

$$\mathbb{E}_t\left[\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2\right] \le (1-\gamma_t)^2 \|\mathbf{u}_{t-1} - g(\mathbf{w}_t)\|_2^2 + \gamma_t^2\sigma_0^2. \tag{4.6}$$

This inequality is same as Lemma 3.7 when we consider $\mathbf{u}_t$ as the SGD update for (4.2).

Due to the Young's inequality of inner product, we have $\|\mathbf{u}_{t-1} - g(\mathbf{w}_t)\|_2^2 \le (1+\alpha)\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2^2 + (1+1/\alpha)\|g(\mathbf{w}_t) - g(\mathbf{w}_{t-1})\|_2^2$ for any $\alpha > 0$. Whence,

$$\begin{aligned}\mathbb{E}_t\left[\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2\right] \le& (1-\gamma_t)^2(1+\gamma_t)\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2^2 \\ &+ (1-\gamma_t)^2(1+1/\gamma_t)G_2^2\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2 + \gamma_t^2\sigma_0^2.\end{aligned}$$

The proof is completed by noticing $(1-\gamma_t)^2(1+\gamma_t) \le 1-\gamma_t$ and $(1-\gamma_t)^2(1+1/\gamma_t) \le \frac{1}{\gamma_t}$. □

Lemma 4.2 *Under Assumptions 4.1, 4.2, 4.3 and 4.4 , SCGD satisfies*

$$\begin{aligned}\mathbb{E}_{\zeta_t,\xi_t,\zeta_t'}\left[F(\mathbf{w}_{t+1})\right] \le& F(\mathbf{w}_t) - \frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t)\|_2^2 + \frac{\eta_t G_2^2 L_1^2}{2}\mathbb{E}_{\zeta_t}\left[\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2\right] \\ &+ \frac{\eta_t^2 L_F G_1^2 G_2^2}{2}.\end{aligned} \tag{4.7}$$

Proof. In the following proof, we abuse the notation $\mathbb{E}_t$ to denote $\mathbb{E}_{\zeta_t,\xi_t,\zeta_t'}$. According to L_F-smoothness of F, we have

$$\begin{aligned}&F(\mathbf{w}_{t+1}) \le F(\mathbf{w}_t) + \nabla F(\mathbf{w}_t)^\top(\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{L_F}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\ &= F(\mathbf{w}_t) - \eta_t \nabla F(\mathbf{w}_t)^\top \nabla g(\mathbf{w}_t;\zeta_t')\nabla f(\mathbf{u}_t;\xi_t) + \frac{\eta_t^2 L_F}{2}\left\|\nabla g(\mathbf{w}_t;\zeta_t')\nabla f(\mathbf{u}_t;\xi_t)\right\|_2^2.\end{aligned}$$

Then, we have

$$\begin{aligned}\mathbb{E}_t\left[F(\mathbf{w}_{t+1})\right] \le& F(\mathbf{w}_t) - \eta_t\|\nabla F(\mathbf{w}_t)\|_2^2 \\ &+ \eta_t\boxed{\mathbb{E}_t\left[\nabla F(\mathbf{w}_t)^\top(\nabla g(\mathbf{w}_t;\zeta_t')\nabla f(g(\mathbf{w}_t)) - \nabla g(\mathbf{w}_t;\zeta_t')\nabla f(\mathbf{u}_t))\right]} \\ &+ \frac{\eta_t^2 L_F}{2}\mathbb{E}_t\left[\left\|\nabla g(\mathbf{w}_t;\zeta_t')\nabla f(\mathbf{u}_t;\xi_t)\right\|_2^2\right],\end{aligned} \tag{4.8}$$

where we use the fact

$$\begin{aligned}&\mathbb{E}_{\zeta_t'}\left[\nabla g(\mathbf{w}_t;\zeta_t')\nabla f(g(\mathbf{w}_t))\right] = \nabla F(\mathbf{w}_t) \\ &\mathbb{E}_{\zeta_t,\zeta_t',\xi_t}\left[\nabla g(\mathbf{w}_t;\zeta_t')\nabla f(\mathbf{u}_t;\xi_t))\right] = \mathbb{E}_{\zeta_t,\zeta_t'}\left[\nabla g(\mathbf{w}_t;\zeta_t')\nabla f(\mathbf{u}_t))\right].\end{aligned}$$

Due to the Cauchy-Schwarz inequality and the Young's inequality of inner product, we have

$$\begin{aligned}
&\mathbb{E}_t\left[\nabla F(\mathbf{w}_t)^\top(\nabla g(\mathbf{w}_t;\zeta_t')\nabla f(g(\mathbf{w}_t)) - \nabla g(\mathbf{w}_t;\zeta_t')\nabla f(\mathbf{u}_t))\right] \\
&\leq \mathbb{E}_t\left[\frac{\|\nabla F(\mathbf{w}_t)\|_2^2 \left\|\nabla g(\mathbf{w}_t;\zeta_t')\right\|_2^2}{2G_2^2}\right] + \mathbb{E}_{\zeta_t}\left[\frac{G_2^2}{2}\left\|\nabla f(g(\mathbf{w}_t)) - \nabla f(\mathbf{u}_t)\right\|_2^2\right] \\
&\leq \frac{\|\nabla F(\mathbf{w}_t)\|_2^2}{2} + \frac{G_2^2 L_1^2}{2}\mathbb{E}_{\zeta_t}\left\|\mathbf{u}_t - g(\mathbf{w}_t)\right\|_2^2. \qquad (4.9)
\end{aligned}$$

For bounding the last term in (4.8), we proceed as follows:

$$\begin{aligned}
\boxed{\mathbb{E}_t\left[\left\|\nabla g(\mathbf{w}_t,\zeta_t')\nabla f(\mathbf{u}_t,\xi_t)\right\|_2^2\right]} &\leq \mathbb{E}_{\zeta_t,\zeta_t'}\left[\left\|\nabla g(\mathbf{w}_t;\zeta_t')\right\|_2^2 \mathbb{E}_{\xi_t|\zeta_t,\zeta_t'}\left\|\nabla f(\mathbf{u}_t;\xi_t)\right\|_2^2\right] \\
&\leq G_1^2 G_2^2. \qquad (4.10)
\end{aligned}$$

We finish the proof by plugging the last two inequalities into (4.8). □

Critical: We comment on the modifications required in the analysis when the same sample ζ_t is used to compute $\nabla g(\mathbf{w}_t;\zeta_t)$. In the original proof, there are two places highlighted in boxes, where we explicitly rely on the independence between $\mathbf{u}_t$ and ζ_t'. If instead we use the coupled estimator $\nabla g(\mathbf{w}_t;\zeta_t)\nabla f(\mathbf{u}_t;\xi_t)$, then the first term must be modified and bounded as follows:

$$\begin{aligned}
&\mathbb{E}_t\left[\nabla F(\mathbf{w}_t)^\top\left(\nabla g(\mathbf{w}_t;\zeta_t)\nabla f(g(\mathbf{w}_t);\xi_t) - \nabla g(\mathbf{w}_t;\zeta_t)\nabla f(\mathbf{u}_t;\xi_t)\right)\right] \\
&\leq \mathbb{E}_t\left[\frac{\|\nabla F(\mathbf{w}_t)\|^2\|\nabla g(\mathbf{w}_t;\zeta_t)\|^2}{2G_2^2}\right] + \mathbb{E}_t\left[\frac{G_2^2}{2}\|\nabla f(g(\mathbf{w}_t);\xi_t) - \nabla f(\mathbf{u}_t;\xi_t)\|^2\right].
\end{aligned}$$

To recover the same bound as in (4.9), we must impose a stronger regularity condition on f, namely,

$$\mathbb{E}_\xi\left[\|\nabla f(g;\xi) - \nabla f(g';\xi)\|^2\right] \leq L_1\|g - g'\|_2^2.$$

For the second boxed term, the corresponding expression becomes $\mathbb{E}_t\left[\|\nabla g(\mathbf{w}_t;\zeta_t)\nabla f(\mathbf{u}_t;\xi_t)\|^2\right]$, which in turn requires assuming that this quantity is uniformly bounded by a constant.

Combining Lemma 4.1 and Lemma 4.2, we can prove the following theorem of convergence for SCGD for a non-convex function.

Theorem 4.1 *Suppose Assumptions 4.1, 4.2, 4.3 and 4.4 hold. After T iterations of SCGD updates with parameters $\eta_t = \frac{\eta_1}{T^{3/5}}, \gamma_t = \frac{\gamma_1}{T^{2/5}}$, we have*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|_2^2\right] \leq \frac{2C_\Upsilon}{\eta_1 T^{2/5}} + \frac{L_1^2 G_1^2 G_2^6 \eta_1^2}{\gamma_1^2 T^{2/5}} + \frac{L_1^2 G_2^2 \sigma_0^2 \gamma_1}{T^{2/5}} + \frac{L_F G_1^2 G_2^2 \eta_1}{2T^{3/5}},$$

where $C_\Upsilon = F(\mathbf{w}_1) - F_* + \frac{L_1^2 C_2^2 \sigma_0^2}{2}\frac{\eta_1}{\gamma_1}$. *If* $\eta_t = \eta_1/t^{3/5}, \gamma_t = \gamma_1/t^{2/5}$, *then the convergence rate becomes* $O(\log T/T^{2/5})$.

Proof. Adding $\frac{L_1^2 G_2^2}{2}\frac{\eta_t}{\gamma_t}\mathbb{E}_t\left[\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2\right]$ on (4.7), we have

$$\begin{aligned}
&\mathbb{E}_t\left[F(\mathbf{w}_{t+1})\right] + \frac{L_1^2 G_2^2}{2}\frac{\eta_t}{\gamma_t}\mathbb{E}_t\left[\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2\right] \\
&\leq F(\mathbf{w}_t) - \frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t)\|_2^2 + (1+\gamma_t)\frac{\eta_t L_1^2 G_2^2}{2\gamma_t}\mathbb{E}_t\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2 + \frac{\eta_t^2 L_F G_1^2 G_2^2}{2}.
\end{aligned}$$

Applying Lemma 4.1 to bound the right hand side, we have

$$\begin{aligned}
&\mathbb{E}_t\left[F(\mathbf{w}_{t+1})\right] + \frac{L_1^2 G_2^2}{2}\frac{\eta_t}{\gamma_t}\mathbb{E}_t\left[\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2\right] \\
&\leq F(\mathbf{w}_t) - \frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t)\|_2^2 + (1-\gamma_t)(1+\gamma_t)\frac{L_1^2 G_2^2}{2}\frac{\eta_t}{\gamma_t}\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2^2 \\
&+ \frac{(1+\gamma_t)L_1^2 G_2^2 G_2^2 \eta_t}{2\gamma_t^2}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2 + \gamma_t\eta_t(1+\gamma_t)\frac{L_1^2 G_2^2\sigma_0^2}{2} + \frac{\eta_t^2 L_F G_1^2 G_2^2}{2} \\
&\overset{\gamma_t\leq 1}{\leq} F(\mathbf{w}_t) + \frac{L_1^2 G_2^2}{2}\frac{\eta_t}{\gamma_t}\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2^2 + \frac{\eta_t L_1^2 G_2^4}{\gamma_t^2}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2 \\
&+ \gamma_t\eta_t L_1^2 G_2^2\sigma_0^2 + \frac{\eta_t^2 L_F G_1^2 G_2^2}{2} - \frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t)\|_2^2.
\end{aligned}$$

We define the potential function $\Upsilon_t = F(\mathbf{w}_t) + \frac{L_1^2 G_2^2}{2}\frac{\eta_t}{\gamma_t}\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2^2$. By the setting, we have $\frac{\eta_{t+1}}{\gamma_{t+1}} \leq \frac{\eta_t}{\gamma_t}$, then

$$\Upsilon_{t+1} = F(\mathbf{w}_{t+1}) + \frac{L_1^2 G_2^2}{2}\frac{\eta_{t+1}}{\gamma_{t+1}}\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2 \leq F(\mathbf{w}_{t+1}) + \frac{L_1^2 G_2^2}{2}\frac{\eta_t}{\gamma_t}\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2.$$

Then,

$$\begin{aligned}
\mathbb{E}_t\left[\Upsilon_{t+1}\right] \leq \Upsilon_t &+ \frac{\eta_t L_1^2 G_2^4}{\gamma_t^2}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2 + \gamma_t\eta_t L_1^2 G_2^2\sigma_0^2 + \frac{\eta_t^2 L_F G_1^2 G_2^2}{2} \\
&- \frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t)\|_2^2.
\end{aligned}$$

Telescoping the above over $t = 1$ to T and use the tower property of conditional expectation.

$$\mathbb{E}\left[\sum_{t=1}^{T}\eta_t\left\|\nabla F(\mathbf{w}_t)\right\|_2^2\right] \leq 2\mathbb{E}\left[\Upsilon_1-\Upsilon_{T+1}\right]+2L_1^2G_2^4\sum_{t=1}^{T}\gamma_t^{-2}\eta_t\eta_{t-1}^2G_1^2G_2^2 + L_1^2G_2^2\sigma_0^2\sum_{t=1}^{T}\gamma_t\eta_t+\frac{L_FG_1^2G_2^2}{2}\sum_{t=1}^{T}\eta_t^2.$$

where we use the fact $\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_{t-1}\|_2^2] = \mathbb{E}[\eta_{t-1}^2\|\nabla g(\mathbf{w}_t;\zeta_t')\nabla f(\mathbf{u}_t;\xi_t)\|_2^2] \leq \eta_{t-1}^2G_1^2G_2^2$. Let $\mathbf{w}_0=\mathbf{w}_1$ and $\mathbf{u}_0=g(\mathbf{w}_0;\zeta_1)$. Then, we have

$$\begin{aligned}\mathbb{E}\left[\Upsilon_1-\Upsilon_{T+1}\right] &\leq \mathbb{E}\left[F(\mathbf{w}_1)+\frac{L_1^2C_2^2}{2}\frac{\eta_1}{\gamma_1}\left\|\mathbf{u}_0-g(\mathbf{w}_0)\right\|_2^2\right]-F_*\\ &\leq F(\mathbf{w}_1)-F_*+\frac{L_1^2G_2^2\sigma_0^2}{2}\frac{\eta_1}{\gamma_1}.\end{aligned}$$

We define $C_\Upsilon = F(\mathbf{w}_1)-F_*+\frac{L_1^2G_2^2\sigma_0^2}{2}\frac{\eta_1}{\gamma_1}$. Then we have

$$\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_\tau)\right\|_2^2\right] \leq \frac{2C_\Upsilon}{\sum_{t=1}^{T}\eta_t}+L_1^2G_2^6G_1^2\frac{\sum_{t=1}^{T}\gamma_t^{-2}\eta_t\eta_{t-1}^2}{\sum_{t=1}^{T}\eta_t} + L_1^2G_2^2\sigma_0^2\frac{\sum_{t=1}^{T}\gamma_t\eta_t}{\sum_{t=1}^{T}\eta_t}+\frac{L_FG_1^2G_2^2}{2}\frac{\sum_{t=1}^{T}\eta_t^2}{\sum_{t=1}^{T}\eta_t}.$$

Plugging the constant values of $\eta_t=\frac{\eta_1}{T^{3/5}}$ and $\gamma_t=\frac{\gamma_1}{T^{2/5}}$, we have

$$\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_\tau)\right\|_2^2\right] \leq \frac{2C_\Upsilon}{\eta_1T^{2/5}}+\frac{L_1^2G_1^2G_2^6\eta_1^2}{\gamma_1^2T^{2/5}}+\frac{L_1^2G_2^2\sigma_0^2\gamma_1}{T^{2/5}}+\frac{L_FG_1^2G_2^2\eta_1}{2T^{3/5}}.$$

If $\eta_t=O(1/t^{3/5})$, $\gamma_t=O(1/t^{2/5})$, $\frac{\eta_{t+1}}{\gamma_{t+1}}\leq\frac{\eta_t}{\gamma_t}$ is satisfied. Besides, we have $\sum_{t=1}^{T}\eta_t = O(T^{2/5})$, $\sum_{t=1}^{T}\eta_t^2=O(1)$, $\sum_{t=1}^{T}\gamma_t\eta_t=O(\log T)$, $\sum_{t=1}^{T}\gamma_t^{-2}\eta_t\eta_{t-1}^2=O(\log T)$. Then, we have $\mathbb{E}\left[\left\|\nabla F(\mathbf{w}_\tau)\right\|_2^2\right]\leq\tilde{O}(1/T^{2/5})$. □

4.2.2 An Improved Complexity with Smooth Inner Function

If we replace the smoothness assumption of F by the smoothness of g, we can establish a better complexity of SCGD.

Assumption 4.5. *g is L_2-smooth, i.e., there exist $L_2>0$ such that $\nabla g(\cdot)$ is L_2-Lipschitz continuous.*

Assumptions 4.1 and 4.5 ensures that F is smooth.

Lemma 4.3 *Under Assumptions 4.1 and 4.5, we have F is L_F-smooth, where $L_F = G_1L_2+G_2^2L_1$.*

Proof. Since $\nabla F(\mathbf{w}) = \nabla g(\mathbf{w})\nabla f(g(\mathbf{w}))$, we have

$$\begin{aligned}
&\|\nabla g(\mathbf{w}_1)\nabla f(g(\mathbf{w}_1)) - \nabla g(\mathbf{w}_2)\nabla f(g(\mathbf{w}_2))\|_2 \\
&= \|\nabla g(\mathbf{w}_1)\nabla f(g(\mathbf{w}_1)) - \nabla g(\mathbf{w}_1)\nabla f(g(\mathbf{w}_2)) \\
&\quad + \nabla g(\mathbf{w}_1)\nabla f(g(\mathbf{w}_2)) - \nabla g(\mathbf{w}_2)\nabla f(g(\mathbf{w}_2))\|_2 \\
&\leq G_2^2 L_1 \|\mathbf{w}_1 - \mathbf{w}_2\|_2 + G_1 L_2 \|\mathbf{w}_1 - \mathbf{w}_2\|_2.
\end{aligned}$$

□

Lemma 4.4 *Let* $\mathbf{z}_t = \nabla g(\mathbf{w}_t; \zeta_t')\nabla f(\mathbf{u}_t; \xi_t)$, $\mathcal{M}_t = \mathbb{E}_t[\mathbf{z}_t]$. *Then*

$$\begin{aligned}
&\mathbb{E}_t[\|\mathbf{z}_t - \mathcal{M}_t\|_2^2] \leq G_1^2\sigma_2^2 + G_2^2\sigma_1^2, \\
&\mathbb{E}_t[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2] \leq \eta_t^2 G_1^2 G_2^2, \\
&\mathbb{E}_t[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2] \leq \eta_t^2\|\mathcal{M}_t\|_2^2 + \eta_t^2(G_1^2\sigma_2^2 + G_2^2\sigma_1^2).
\end{aligned}$$

where $\mathbb{E}_t$ *denotes* $\mathbb{E}_{\zeta_t', \xi_t}$ *conditioned on* $\mathbf{w}_t, \mathbf{u}_t$.

Proof. First, we have

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{z}_t - \mathcal{M}_t\|_2^2] = \mathbb{E}[\left\|\nabla g(\mathbf{w}_t; \zeta_t')\nabla f(\mathbf{u}_t; \xi_t) - \nabla g(\mathbf{w}_t)\nabla f(\mathbf{u}_t)\right\|_2^2] \\
&= \mathbb{E}[\|\nabla g(\mathbf{w}_t)\nabla f(\mathbf{u}_t) - \nabla g(\mathbf{w}_t)\nabla f(\mathbf{u}_t; \xi_t) \\
&+ \nabla g(\mathbf{w}_t)\nabla f(\mathbf{u}_t; \xi_t) - \nabla g(\mathbf{w}_t; \zeta_t')\nabla f(\mathbf{u}_t; \xi_t)\|_2^2] \\
&\leq G_2^2\sigma_1^2 + G_1^2\sigma_2^2.
\end{aligned}$$

Next, due to Assumption 4.1, 4.2 we have

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2] = \mathbb{E}[\eta_t^2\|\nabla g(\mathbf{w}_t; \zeta_t')\nabla f(\mathbf{u}_t; \xi_t)\|_2^2] \leq \eta_t^2 G_1^2 G_2^2.$$

Second, we have

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2] = \mathbb{E}[\eta_t^2\|\mathbf{z}_t - \mathcal{M}_t + \mathcal{M}_t\|_2^2] = \mathbb{E}[\eta_t^2\|\mathbf{z}_t - \mathcal{M}_t\|_2^2] + \eta_t^2\|\mathcal{M}_t\|_2^2.$$

Plugging the first result into the above, we finish the proof. □

Next, we develop two lemmas similar to Lemma 4.1 and Lemma 4.2.

Lemma 4.5 *Under Assumptions 4.2, 4.3 and 4.5, if* $\eta_{t-1}^2 \leq \frac{\gamma_t}{L_2^2 G_1^2}$ *then the* $\{\mathbf{u}_t\}_{t\geq 1}$ *sequence (4.3) satisfies that*

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2\right] &\leq (1-\gamma_t)\mathbb{E}[\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2^2] + \frac{4\eta_{t-1}^2 G_2^2}{\gamma_t}\mathbb{E}[\|\mathcal{M}_{t-1}\|_2^2] \\
&\quad + \gamma_t^2\sigma_0^2 + \frac{3\eta_{t-1}^2 G_2^2}{2}(G_1^2\sigma_2^2 + G_2^2\sigma_1^2). \qquad (4.11)
\end{aligned}$$

Proof. Similar to the proof of Lemma 4.1, we have

$$\mathbb{E}_t\left[\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2\right] \leq (1-\gamma_t)^2\|\mathbf{u}_{t-1} - g(\mathbf{w}_t)\|_2^2 + \gamma_t^2\sigma_0^2. \tag{4.12}$$

Next, we will handle $\|\mathbf{u}_{t-1} - g(\mathbf{w}_t)\|_2^2$ differently by using the smoothness of g.

$$\begin{aligned}
\|\mathbf{u}_{t-1} - g(\mathbf{w}_t)\|_2^2 &= \|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1}) + g(\mathbf{w}_{t-1}) - g(\mathbf{w}_t)\|_2^2 \\
&= \|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2^2 + \|g(\mathbf{w}_{t-1}) - g(\mathbf{w}_t)\|_2^2 \\
&+ (\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1}))^\top(g(\mathbf{w}_{t-1}) - g(\mathbf{w}_t)) \\
&\leq \|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2^2 + G_2^2\|\mathbf{w}_{t-1} - \mathbf{w}_t\|_2^2 \\
&+ 2(\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1}))^\top(g(\mathbf{w}_{t-1}) - g(\mathbf{w}_t)).
\end{aligned}$$

Taking expectation on both sides and applying Lemma 4.4, we have

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{u}_{t-1} - g(\mathbf{w}_t)\|_2^2] \leq \mathbb{E}[\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2^2] + \eta_{t-1}^2 G_2^2\mathbb{E}[\|\mathcal{M}_{t-1}\|_2^2] \\
&+ \eta_{t-1}^2 G_2^2(G_2^2\sigma_1^2 + G_1\sigma_2^2) + \mathbb{E}[2(\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1}))^\top(g(\mathbf{w}_{t-1}) - g(\mathbf{w}_t))].
\end{aligned}$$

Instead of using the Young's inequality of inner product to bound the last term, we proceed as follows:

$$\begin{aligned}
&\mathbb{E}[(\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1}))^\top(g(\mathbf{w}_{t-1}) - g(\mathbf{w}_t))] \\
&= \underbrace{\mathbb{E}[(\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1}))^\top\nabla g(\mathbf{w}_{t-1})^\top(\mathbf{w}_{t-1} - \mathbf{w}_t)]}_{A} \\
&+ \underbrace{\mathbb{E}[(\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1}))^\top(g(\mathbf{w}_{t-1}) - g(\mathbf{w}_t) + \nabla g(\mathbf{w}_{t-1})^\top(\mathbf{w}_t - \mathbf{w}_{t-1}))]}_{B}.
\end{aligned}$$

To bound A, we have

$$\begin{aligned}
A &= \mathbb{E}[(\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1}))^\top\nabla g(\mathbf{w}_{t-1})^\top\eta_{t-1}\mathcal{M}_{t-1}] \\
&\leq \mathbb{E}[\alpha_t\|(\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1}))^\top\|^2 + \frac{\eta_{t-1}^2}{4\alpha_t}\|\nabla g(\mathbf{w}_{t-1})^\top\mathcal{M}_{t-1}\|_2^2] \\
&\leq \mathbb{E}[\alpha_t\|(\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1}))^\top\|^2 + \frac{\eta_{t-1}^2 G_2^2}{4\alpha_t}\|\mathcal{M}_{t-1}\|_2^2].
\end{aligned}$$

To bound B, we have

$$\begin{aligned}
B &\leq \mathbb{E}[\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2\|g(\mathbf{w}_{t-1}) - g(\mathbf{w}_t) + \nabla g(\mathbf{w}_{t-1})^\top(\mathbf{w}_t - \mathbf{w}_{t-1})\|_2] \\
&\leq \mathbb{E}[\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2\frac{L_2}{2}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2] \\
&\leq \frac{L_2^2}{4G_2^2}\mathbb{E}[\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2^2\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2] + \frac{G_2^2}{4}\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2]
\end{aligned}$$

where the first inequality uses the smoothness of g and the last inequality uses the Young's inequality. To proceed, we utilize the first bound of $\mathbb{E}_{t-1}[\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2]$

in lemma 4.4 to bound the first term, and utilize its second bound in lemma 4.4 to bound the second $\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2]$. Thus, we have

$$B \leq \frac{\eta_{t-1}^2 L_2^2 G_1^2}{4}\mathbb{E}[\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2^2] + \frac{\eta_{t-1}^2 G_2^2}{4}\mathbb{E}[\|\mathcal{M}_{t-1}\|_2^2] + \frac{\eta_{t-1}^2 G_2^2}{4}(G_1^2\sigma_2^2 + G_2^2\sigma_1^2)].$$

Combing the bounds for A and B, we have

$$\begin{aligned}&\mathbb{E}[(\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1}))^\top (g(\mathbf{w}_{t-1}) - g(\mathbf{w}_t))] \\ &= \left(\alpha_t + \frac{\eta_{t-1}^2 L_2^2 G_1^2}{4}\right)\mathbb{E}[\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2^2] + \left(\frac{\eta_{t-1}^2 G_2^2}{4\alpha_t} + \frac{\eta_{t-1}^2 G_2^2}{4}\right)\mathbb{E}[\|\mathcal{M}_{t-1}\|_2^2] \\ &+ \frac{\eta_{t-1}^2 G_2^2}{4}(G_1^2\sigma_2^2 + G_2^2\sigma_1^2).\end{aligned}$$

As a result,

$$\begin{aligned}\mathbb{E}[\|\mathbf{u}_{t-1} - g(\mathbf{w}_t)\|_2^2] &\leq \left(1 + 2\alpha_t + \frac{\eta_{t-1}^2 L_2^2 G_1^2}{2}\right)\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2^2 \\ &+ \left(\eta_{t-1}^2 G_2^2 + \frac{\eta_{t-1}^2 G_2^2}{2\alpha_t} + \frac{\eta_{t-1}^2 G_2^2}{2}\right)\mathbb{E}[\|\mathcal{M}_{t-1}\|_2^2] \\ &+ \left(\eta_{t-1}^2 G_2^2 + \frac{\eta_{t-1}^2 G_2^2}{2}\right)(G_1^2\sigma_2^2 + G_2^2\sigma_1^2).\end{aligned}$$

We let $\alpha_t = \frac{\gamma_t}{4} < 1$, $\frac{\eta_{t-1}^2 L_2^2 G_1^2}{2} \leq \frac{\gamma_t}{2}$. Combining the above inequality with (4.12), we can finish the proof.

□

Lemma 4.6 *Under Assumptions 4.1, 4.2, 4.3 and 4.5, if $\eta_t L_F \leq 1/4$ then SCGD satisfies*

$$\begin{aligned}\mathbb{E}[F(\mathbf{w}_{t+1})] \leq &\mathbb{E}\left[F(\mathbf{w}_t) - \frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t)\|_2^2 - \frac{\eta_t}{4}\|\mathcal{M}_t\|_2^2\right] \\ &+ \frac{\eta_t G_2^2 L_1^2}{2}\mathbb{E}[\|g(\mathbf{w}_t) - \mathbf{u}_t\|_2^2] + 2\eta_t^2 L_F(G_2^2\sigma_1^2 + G_1^2\sigma_2^2).\end{aligned} \tag{4.13}$$

Proof. According to Lemma 4.3 (L_F-smoothness of F), we have

$$\begin{aligned}&F(\mathbf{w}_{t+1}) \leq F(\mathbf{w}_t) + \nabla F(\mathbf{w}_t)^\top(\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{L_F}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\ =&F(\mathbf{w}_t) - \eta_t \nabla F(\mathbf{w}_t)^\top \nabla g(\mathbf{w}_t; \zeta_t')\nabla f(\mathbf{u}_t; \xi_t) + \frac{\eta_t^2 L_F}{2}\left\|\nabla g(\mathbf{w}_t; \zeta_t')\nabla f(\mathbf{u}_t; \xi_t)\right\|_2^2.\end{aligned}$$

Taking expectation on both sides, we have

$$\mathbb{E}[F(\mathbf{w}_{t+1})] \leq \mathbb{E}[F(\mathbf{w}_t)] - \eta_t \mathbb{E}[\nabla F(\mathbf{w}_t)^\top \mathcal{M}_t] + \frac{\eta_t^2 L_F}{2} \mathbb{E}[\|\mathbf{z}_t - \mathcal{M}_t + \mathcal{M}_t\|_2^2]$$
$$= \mathbb{E}[F(\mathbf{w}_t)] - \eta_t \mathbb{E}[\nabla F(\mathbf{w}_t)^\top \mathcal{M}_t] + \eta_t^2 L_F \mathbb{E}[\|\mathbf{z}_t - \mathcal{M}_t\|_2^2] + \eta_t^2 L_F \mathbb{E}[\|\mathcal{M}_t\|_2^2]$$

Using $-2\mathbf{a}^\top \mathbf{b} = \|\mathbf{a} - \mathbf{b}\|_2^2 - \|\mathbf{a}\|_2^2 - \|\mathbf{b}\|_2^2$, we have

$$\mathbb{E}[F(\mathbf{w}_{t+1})] \leq \mathbb{E}[F(\mathbf{w}_t) - \frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t)\|_2^2 - \frac{\eta_t}{2}\|\mathcal{M}_t\|_2^2]$$
$$+ \frac{\eta_t}{2}\mathbb{E}[\|\nabla F(\mathbf{w}_t) - \mathcal{M}_t\|_2^2] + \eta_t^2 L_F \mathbb{E}[\|\mathbf{z}_t - \mathcal{M}_t\|_2^2] + \eta_t^2 L_F \mathbb{E}[\|\mathcal{M}_t\|_2^2].$$

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

$$\mathbb{E}[\|\nabla F(\mathbf{w}_t) - \mathcal{M}_t\|_2^2] = \mathbb{E}[\|\nabla g(\mathbf{w}_t)\nabla f(g(\mathbf{w}_t)) - \nabla g(\mathbf{w}_t)\nabla f(\mathbf{u}_t)\|_2^2]$$
$$\leq G_2^2 L_1^2 \mathbb{E}[\|g(\mathbf{w}_t) - \mathbf{u}_t\|_2^2].$$

Combining the above inequalities, we have

$$\mathbb{E}[F(\mathbf{w}_{t+1})] \leq \mathbb{E}[F(\mathbf{w}_t) - \frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t)\|_2^2 - \frac{\eta_t}{2}\|\mathcal{M}_t\|_2^2]$$
$$+ \frac{\eta_t G_2^2 L_1^2}{2}\mathbb{E}[\|g(\mathbf{w}_t) - \mathbf{u}_t\|_2^2] + \eta_t^2 L_F (G_2^2\sigma_1^2 + G_1^2\sigma_2^2) + \eta_t^2 L_F \mathbb{E}[\|\mathcal{M}_t\|_2^2].$$

If $\eta_t L_F \leq 1/4$, we have $-\frac{\eta_t}{2}\|\mathcal{M}_t\|_2^2 + \eta_t^2 L_F \|\mathcal{M}_t\|_2^2 \leq \frac{\eta_t}{4}\|\mathcal{M}_t\|_2^2$, which concludes the proof. □

Finally, we establish the following convergence of SCGD under the smoothness condition of g.

Theorem 4.2 *Suppose Assumptions 4.1, 4.5 and 4.3 hold. Run SCGD with T iterations with parameters $\eta_t = \frac{\eta_1}{\sqrt{T}}, \gamma_t = \frac{\gamma_1}{\sqrt{T}}$, where $\eta_1 \leq \min(\frac{\gamma_1}{\sqrt{8G_2^2 L_1}}, \frac{\sqrt{2\gamma_1}}{L_2 G_1}, \frac{1}{4L_F})$. Then we have*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_t)\|_2^2\right] \leq O\left(\frac{C_\Upsilon}{\eta_1\sqrt{T}} + \frac{L_1\gamma_1^2\sigma_0^2}{\eta_1\sqrt{T}} + \frac{\eta_1(L_F + L_1G_2^2)(G_2^2\sigma_1^2 + G_1^2\sigma_2^2)}{\sqrt{T}}\right),$$

where $C_\Upsilon = F(\mathbf{w}_1) - F_* + \frac{L_1}{\sqrt{6}}\|\mathbf{u}_1 - g(\mathbf{w}_1)\|_2^2$.

Why it matters

From Theorem 4.2, we can derive that in order to find an ϵ-level stationary solution of a smooth non-convex compositional function (whose gradient norm is less than ϵ), SCGD needs a sample complexity of $O(\frac{L_1^4}{\epsilon^4})$. The order in terms of ϵ is the same order as that of SGD for solving non-convex ERM.

Proof. By Lemma 4.5, and Lemma 4.6, we have

$$\mathbb{E}[F(\mathbf{w}_{t+1})] \leq \mathbb{E}[F(\mathbf{w}_t) - \frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t)\|_2^2 - \frac{\eta_t}{4}\|\mathcal{M}_t\|_2^2]$$
$$+ \frac{\eta_t G_2^2 L_1^2}{2}\mathbb{E}[\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2] + \eta_t^2 L_F(G_2^2\sigma_1^2 + G_1^2\sigma_2^2),$$
$$\mathbb{E}\left[\|\mathbf{u}_{t+1} - g(\mathbf{w}_{t+1})\|_2^2\right] \leq (1-\gamma_{t+1})\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2 + \frac{4\eta_t^2 G_2^2}{\gamma_{t+1}}\mathbb{E}[\|\mathcal{M}_t\|_2^2]$$
$$+ \gamma_{t+1}^2\sigma_0^2 + \frac{3\eta_t^2 G_2^2}{2}(G_1^2\sigma_2^2 + G_2^2\sigma_1^2).$$

Multiplying the second inequality by $G_2^2 L_1^2\eta_t/(2\gamma_{t+1})$ and adding it to the first inequality, we have

$$\mathbb{E}\left[F(\mathbf{w}_{t+1}) + \frac{\eta_t G_2^2 L_1^2}{2\gamma_{t+1}}\|\mathbf{u}_{t+1} - g(\mathbf{w}_{t+1})\|_2^2\right] \leq \mathbb{E}\left[F(\mathbf{w}_t) - \frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t)\|_2^2 - \frac{\eta_t}{4}\|\mathcal{M}_t\|_2^2\right]$$
$$+ \frac{\eta_t G_2^2 L_1^2}{2\gamma_{t+1}}\mathbb{E}[\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2] + \frac{\eta_t G_2^2 L_1^2}{2\gamma_{t+1}}\frac{4\eta_t^2 G_2^2}{\gamma_{t+1}}\mathbb{E}[\|\mathcal{M}_t\|_2^2]$$
$$+ \eta_t^2 L_F(G_2^2\sigma_1^2 + G_1^2\sigma_2^2) + \frac{\eta_t G_2^2 L_1^2}{2\gamma_{t+1}}\gamma_{t+1}^2\sigma_0^2 + \frac{\eta_t G_2^2 L_1^2}{2\gamma_{t+1}}\frac{3\eta_t^2 G_2^2}{2}(G_1^2\sigma_2^2 + G_2^2\sigma_1^2).$$

Since $\frac{\eta_t G_2^2 L_1^2}{2\gamma_{t+1}}\frac{4\eta_t^2 G_2^2}{\gamma_{t+1}} \leq \frac{\eta_t}{4}$ due to $\eta_t \leq \frac{\gamma_{t+1}}{\sqrt{8}G_2^2 L_1}$, the term involving $\|\mathcal{M}_t\|_2^2$ will be less than zero. If $\frac{\eta_t}{\gamma_{t+1}} \leq \frac{\eta_{t-1}}{\gamma_t}$, we obtain

$$\mathbb{E}\left[F(\mathbf{w}_{t+1}) + \frac{\eta_t G_2^2 L_1^2}{2\gamma_{t+1}}\|\mathbf{u}_{t+1} - g(\mathbf{w}_{t+1})\|_2^2\right]$$
$$\leq \mathbb{E}\left[F(\mathbf{w}_t) + \frac{\eta_{t-1} G_2^2 L_1^2}{2\gamma_t}\mathbb{E}[\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2\right] - \frac{\eta_t}{2}\mathbb{E}[\|\nabla F(\mathbf{w}_t)\|_2^2] + \frac{\eta_t G_2^2 L_1^2}{2\gamma_{t+1}}\gamma_{t+1}^2\sigma_0^2$$
$$+ \eta_t^2 L_F(G_2^2\sigma_1^2 + G_1^2\sigma_2^2) + \frac{\eta_t G_2^2 L_1^2}{2\gamma_{t+1}}\frac{3\eta_t^2 G_2^2}{2}(G_1^2\sigma_2^2 + G_2^2\sigma_1^2).$$

Applying $\eta_t \leq \frac{\gamma_{t+1}}{\sqrt{8}G_2^2 L_1}$ to the R.H.S, we have

$$\mathbb{E}\left[F(\mathbf{w}_{t+1}) + \frac{\eta_t G_2^2 L_1^2}{2\gamma_{t+1}}\|\mathbf{u}_{t+1} - g(\mathbf{w}_{t+1})\|_2^2\right]$$
$$\leq \mathbb{E}\left[F(\mathbf{w}_t) + \frac{\eta_{t-1} G_2^2 L_1^2}{2\gamma_t}\mathbb{E}[\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2\right] - \frac{\eta_t}{2}\mathbb{E}[\|\nabla F(\mathbf{w}_t)\|_2^2]$$
$$+ \frac{L_1}{2\sqrt{8}}\gamma_{t+1}^2\sigma_0^2 + \eta_t^2(L_F + L_1 G_2^2)(G_2^2\sigma_1^2 + G_1^2\sigma_2^2).$$

Define $\Upsilon_t = F(\mathbf{w}_t) + \frac{\eta_{t-1} G_2^2 L_1^2}{2\gamma_t}\mathbb{E}[\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2$. Then we have $\sum_{t=1}^T(\Upsilon_t - \Upsilon_{t+1}) \leq C_\Upsilon := \Upsilon_1 - F_*$ and

4.2.3 A Straightforward Approach with a Large Mini-batch

Before ending this section, we compare the complexity of SCGD with a straightforward approach that uses a large batch size for estimating the gradient. In particular, we update the model parameter by the following:

$$\bar{\mathbf{u}}_t = \frac{1}{B}\sum_{j=1}^{B} g(\mathbf{w}_t; \zeta_{j,t}), \quad \bar{\mathbf{v}}_t = \frac{1}{B}\sum_{i=1}^{B} \nabla g(\mathbf{w}_t; \zeta'_{i,t}) \nabla f\left(\bar{\mathbf{u}}_t; \xi_{i,t}\right) \tag{4.14}$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \bar{\mathbf{v}}_t. \tag{4.15}$$

Then under Assumptions 4.1, 4.2, we have

$$\begin{aligned}
&\mathbb{E}[\|\bar{\mathbf{v}}_t - \nabla F(\mathbf{w}_t)\|_2^2] \\
&\leq \mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B} \nabla g(\mathbf{w}_t; \zeta'_{i,t}) \nabla f\left(\bar{\mathbf{u}}_t; \xi_{i,t}\right) - \nabla g(\mathbf{w}_t) \nabla f\left(\bar{\mathbf{u}}_t\right)\right.\right. \\
&\qquad \left.\left. + \nabla g(\mathbf{w}_t) \nabla f\left(\bar{\mathbf{u}}_t\right) - \nabla F(\mathbf{w}_t)\right\|_2^2\right].
\end{aligned}$$

Since

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B} \nabla g(\mathbf{w}_t; \zeta'_{i,t}) \nabla f\left(\bar{\mathbf{u}}_t; \xi_{i,t}\right) - \nabla g(\mathbf{w}_t) \nabla f\left(\bar{\mathbf{u}}_t\right)\right\|_2\right] \\
&\leq \frac{G_2^2\sigma_1^2 + G_1^2\sigma_2^2}{B}, \\
&\mathbb{E}\left[\left\|\nabla g(\mathbf{w}_t) \nabla f\left(\bar{\mathbf{u}}_t\right) - \nabla F(\mathbf{w}_t)\right\|_2^2\right] \leq \mathbb{E}[G_2^2 L_1^2 \|\bar{\mathbf{u}}_t - g(\mathbf{w}_t)\|_2^2] \leq \frac{G_2^2 L_1^2 \sigma_0^2}{B},
\end{aligned}$$

then, $\mathbb{E}[\|\bar{\mathbf{v}}_t - \nabla F(\mathbf{w}_t)\|_2^2] \leq O\left(\frac{L_1^2\sigma_0^2}{B} + \frac{\sigma_1^2+\sigma_2^2}{B}\right)$. Hence, if Assumption 4.4 holds and by setting $B = O(\max(L_1^2\sigma_0^2/\epsilon^2, (\sigma_1^2+\sigma_2^2)/\epsilon^2))$, $\eta = O(1/L_F)$ and $T = O(L_F/\epsilon^2)$, Lemma 4.9 will indicate that the naive approach can find an ϵ-stationary solution. Overall, it yields a sample complexity of

$$BT = O\left(\max\left(\frac{L_F L_1^2 \sigma_0^2}{\epsilon^4}, \frac{L_F(\sigma_1^2+\sigma_2^2)}{\epsilon^4}\right)\right).$$

Algorithm 10 SCMA

Input: learning rate schedules $\{\eta_t\}_{t=1}^T$, $\{\gamma_t\}_{t=1}^T$; 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**
    Sample ζ_t, ζ_t' and ξ_t
    Compute the inner function value estimator $\mathbf{u}_t = (1-\gamma_t)\mathbf{u}_{t-1} + \gamma_t g(\mathbf{w}_t; \zeta_t)$
    Compute the vanilla gradient estimator $\mathbf{z}_t = \nabla g(\mathbf{w}_t; \zeta_t')\nabla f(\mathbf{u}_t; \xi_t)$
    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

Critical: Compared with Theorem 4.1, the sample complexity of this naïve approach is improved by an order of magnitude. In comparison to Theorem 4.2, while the order of ϵ remains identical, the dependence on the Lipschitz constant L_1 is reduced. Specifically, SCGD exhibits a dependence of $O(L_1^4)$, whereas the large mini-batch approach achieves $O(L_1^3)$, assuming $L_F = O(L_1)$.

4.3 Stochastic Compositional Momentum Metholds

In this section, we present a method that matches the sample complexity of the large mini-batch approach without using large mini-batches under the smoothness conditions of f and F. The idea is to design a gradient estimator such that its error can be reduced gradually. It turns out this technique, related to the momentum metholds for standard stochastic optimization, is more widely applicable to other problems discussed later in this chapter. Furthermore, we introduce advanced metholds to further improve the complexity to $O(1/\epsilon^3)$ under stronger conditions.

It is worth noting that the results in this section apply to the standard stochastic optimization problem (3.1) under the smoothness assumption of $g(\mathbf{w})$ by setting $f_i(g) = g$ and $L_1 = 0$ in the complexity results and removing the $\mathbf{u}$ update in the algorithm.

4.3.1 Moving-Average Gradient Estimator

The first method is to use the following moving-average gradient estimator:

$$\mathbf{v}_t = (1-\beta_t)\mathbf{v}_{t-1} + \beta_t \nabla g(\mathbf{w}_t; \zeta_t')\nabla f(\mathbf{u}_t; \xi_t), \tag{4.16}$$

where $0 \leq \beta_t < 1$. With $\mathbf{v}_t$, the model parameter is updated by:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{v}_t. \tag{4.17}$$

We present the full steps in Algorithm 10 and refer to it as SCMA.

To understand this method, we can view $\mathbf{v}_t$ as a better estimator of the gradient, with its estimation error gradually decreasing over iterations—a property we will prove shortly. This yields an enhanced stability of momentum-based methods observed in practice.

Connection with Stochastic Momentum Methods

This method is analogous to applying the stochastic momentum method to the ERM problem, using the term $\nabla g(\mathbf{w}_t; \zeta_t')\nabla f(\mathbf{u}_t; \xi_t)$ as a surrogate for the true stochastic gradient. This connection is revealed by reformulating the update into a canonical momentum form:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t' \nabla g(\mathbf{w}_t; \zeta_t')\nabla f(\mathbf{u}_t; \xi_t) + \beta_t'(\mathbf{w}_t - \mathbf{w}_{t-1}), \tag{4.18}$$

where the effective step size and momentum parameters are $\eta_t' = \eta_t\beta_t$ and $\beta_t' = \eta_t(1-\beta_t)/\eta_{t-1}$, respectively. The term $\mathbf{w}_t - \mathbf{w}_{t-1}$ is the momentum term.

In the special case where f is the identity function, the update is identical to the classical stochastic momentum method (also known as stochastic heavy-ball method), renowned for its accelerated performance on quadratic functions relative to plain gradient descent. Hence, the convergence analysis presented below also applies to the stochastic momentum method for ERM by setting $L_1 = 0$.

Convergence Analysis

First, we prove a generic lemma that establishes the error recursion of $\mathbf{v}_t$.

Lemma 4.7 *Let* $\mathbf{v}_t = (1-\beta_t)\mathbf{v}_{t-1} + \beta_t \mathbf{z}_t$, *where* $\mathbb{E}_t[\mathbf{z}_t] = \mathcal{M}_t$. *If* $\mathbb{E}_t[\|\mathbf{z}_t - \mathcal{M}_t\|_2^2] \le \sigma^2$, *then we have*

$$\begin{aligned}\mathbb{E}_t\left[\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2\right] &\le (1-\beta_t)\|\mathbf{v}_{t-1} - \nabla F(\mathbf{w}_{t-1})\|_2^2 + \beta_t^2\sigma^2 \\ &+ \frac{2L_F^2}{\beta_t}\|\mathbf{w}_{t-1} - \mathbf{w}_t\|_2^2 + 4\beta_t\|\mathcal{M}_t - \nabla F(\mathbf{w}_t)\|_2^2.\end{aligned} \tag{4.19}$$

Proof. Due to the update formula $\mathbf{v}_t = (1-\beta_t)\mathbf{v}_{t-1} + \beta_t\mathbf{z}_t$, we have

$$
\begin{aligned}
&\mathbb{E}_t\left[\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2\right] \\
&= \mathbb{E}_t\left[\|(1-\beta_t)\mathbf{v}_{t-1} + \beta_t \mathbf{z}_t - \nabla F(\mathbf{w}_t)\|_2^2\right] \\
&= \mathbb{E}_t\Bigg[\| \underbrace{(1-\beta_t)\mathbf{v}_{t-1} - \nabla F(\mathbf{w}_t) + \beta_t \mathcal{M}_t}_{\mathbf{a}_t} + \underbrace{\beta_t(\mathbf{z}_t - \mathcal{M}_t)}_{\mathbf{b}_t} \|_2^2\Bigg].
\end{aligned}
$$

Note that $\mathbb{E}_t[\mathbf{a}_t^\top \mathbf{b}_t] = 0$. Besides, we have $\mathbb{E}_t[\|\mathbf{b}_t\|_2^2] \leq \beta_t^2\sigma^2$. Due to Young's inequality, we have $\|a+b\|_2^2 \leq (1+\alpha)\|a\|_2^2 + (1+1/\alpha)\|b\|_2^2$ for any $\alpha > 0$. Hence,

$$
\begin{aligned}
\|\mathbf{a}_t\|_2^2 &= \|(1-\beta_t)(\mathbf{v}_{t-1} - \nabla F(\mathbf{w}_{t-1})) + (1-\beta_t)(\nabla F(\mathbf{w}_{t-1}) - \nabla F(\mathbf{w}_t)) \\
&\quad + \beta_t(\mathcal{M}_t - \nabla F(\mathbf{w}_t))\|_2^2 \\
&\leq (1-\beta_t)^2(1+\beta_t)\|(\mathbf{v}_{t-1} - \nabla F(\mathbf{w}_{t-1}))\|_2^2 \\
&+ (1+\frac{1}{\beta_t})\|(1-\beta_t)(\nabla F(\mathbf{w}_{t-1}) - \nabla F(\mathbf{w}_t)) + \beta_t(\mathcal{M}_t - \nabla F(\mathbf{w}_t))\|_2^2 \\
&\leq (1-\beta_t)\,\|\mathbf{v}_{t-1} - \nabla F(\mathbf{w}_{t-1})\|_2^2 + \frac{2(1+\beta_t)(1-\beta_t)^2}{\beta_t}\,\|\nabla F(\mathbf{w}_{t-1}) - \nabla F(\mathbf{w}_t)\|_2^2 \\
&+ \frac{2(1+\beta_t)\beta_t^2}{\beta_t}\,\|\mathcal{M}_t - \nabla F(\mathbf{w}_t)\|_2^2 \\
&\leq (1-\beta_t)\,\|\mathbf{v}_{t-1} - \nabla F(\mathbf{w}_{t-1})\|_2^2 + \frac{2L_F^2}{\beta_t}\,\|\mathbf{w}_{t-1} - \mathbf{w}_t\|_2^2 + 4\beta_t\,\|\mathcal{M}_t - \nabla F(\mathbf{w}_t)\|_2^2\,.
\end{aligned}
$$

Combining the above results, we finish the proof. □

With the above lemma, we are able to establish the error recursion of $\mathbf{v}_t$ of SCMA.

Lemma 4.8 *Under Assumptions 4.1, 4.2 and 4.3, for $t \geq 1$ SCMA satisfies that*

$$
\begin{aligned}
\mathbb{E}_{\xi_t,\zeta_t'}\left[\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2\right] &\leq (1-\beta_t)\,\|\mathbf{v}_{t-1} - \nabla F(\mathbf{w}_{t-1})\|_2^2 \\
&+ \frac{2L_F^2}{\beta_t}\,\|\mathbf{w}_{t-1} - \mathbf{w}_t\|_2^2 + 4G_2^2L_1^2\beta_t\,\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2 + \beta_t^2\sigma^2.
\end{aligned} \tag{4.20}
$$

where $\sigma^2 = G_1^2\sigma_2^2 + G_2^2\sigma_1^2$.

Why it matters

The above lemma establishes the recursion of the error of stochastic gradient estimator $\mathbf{v}_t$. It is the key to show that the average of the estimator error of $\mathbf{v}_t$ will converge to zero.

Proof. We denote by $\mathbb{E}_t[\cdot] = \mathbb{E}_{\xi_t,\zeta_t'}[\cdot]$. Let $\mathbf{z}_t = \nabla g(\mathbf{w}_t;\zeta_t')\nabla f(\mathbf{u}_t;\xi_t)$ and $\mathcal{M}_t = \mathbb{E}_t[\mathbf{z}_t] = \nabla g(\mathbf{w}_t)\nabla f(\mathbf{u}_t)$. Lemma 4.4 proves that

$$
\mathbb{E}_t[\|\mathbf{z}_t - \mathcal{M}_t\|_2^2] \leq G_2^2\sigma_1^2 + G_1^2\sigma_2^2, \tag{4.21}
$$

and

$$\begin{aligned}\|\mathcal{M}_t - \nabla F(\mathbf{w}_t)\|_2^2 &= \|\nabla g(\mathbf{w}_t)\nabla f(\mathbf{u}_t) - \nabla g(\mathbf{w}_t)\nabla f(g(\mathbf{w}_t))\|_2^2 \\ &\leq G_2^2 L_1^2 \|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2.\end{aligned}$$

Plugging these two results into Lemma 4.7, we finish the proof. □

Critical: If we use the same random sample ζ_t to compute

$$\mathbf{z}_t = \nabla g(\mathbf{w}_t; \zeta_t)\nabla f(\mathbf{u}_t; \xi_t),$$

then $\mathcal{M}_t = \mathbb{E}_{\xi_t,\zeta_t}[\mathbf{z}_t]$ is not equal to $\nabla g(\mathbf{w}_t)\nabla f(\mathbf{u}_t)$. However, we just need to assume that $\mathbb{E}_{\zeta_t,\xi_t}[\|\mathbf{z}_t - \mathcal{M}_t\|_2^2]$ is bounded and $\|\nabla g(\mathbf{w}_t; \zeta_t)\|_2^2 \leq G_2$. Then

$$\begin{aligned}\|\mathcal{M}_t - \nabla F(\mathbf{w}_t)\|_2^2 &= \|\mathbb{E}_{\zeta_t}\nabla g(\mathbf{w}_t; \zeta_t)\nabla f(\mathbf{u}_t) - \mathbb{E}_{\zeta_t}\nabla g(\mathbf{w}_t; \zeta_t)\nabla f(g(\mathbf{w}_t))\|_2^2 \\ &\leq \mathbb{E}_{\zeta_t}\|\nabla g(\mathbf{w}_t; \zeta_t)\nabla f(\mathbf{u}_t) - \nabla g(\mathbf{w}_t; \zeta_t)\nabla f(g(\mathbf{w}_t))\|_2^2 \\ &\leq \mathbb{E}_{\zeta_t}[\|\nabla g(\mathbf{w}_t; \zeta_t)\|_2^2\|\nabla f(\mathbf{u}_t) - \nabla f(g(\mathbf{w}_t))\|_2^2] \\ &\leq \mathbb{E}_{\zeta_t}\left[G_2^2 L_1^2 \|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2\right].\end{aligned}$$

The following analysis will proceed in the same manner.

To enjoy the above recursion of the gradient estimator's error, we state the following lemma, which is a variant of the standard descent lemma of gradient descent.

Lemma 4.9 *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*

$$F(\mathbf{w}_{t+1}) \leq F(\mathbf{w}_t) + \frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t) - \mathbf{v}_t\|_2^2 - \frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t)\|_2^2 - \frac{1}{4\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2. \tag{4.22}$$

Why it matters

This lemma ensures that if the stochastic gradient error satisfies $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \|\nabla F(\mathbf{w}_t) - \mathbf{v}_t\|_2^2\right] \to 0$, then the convergence of $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \|\nabla F(\mathbf{w}_t)\|_2^2\right]$ to zero is guaranteed.

Proof. Due to the smoothness of F, we have

$$
\begin{aligned}
&F(\mathbf{w}_{t+1}) \leq F(\mathbf{w}_t) + \nabla F(\mathbf{w}_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{L_F}{2} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&= F(\mathbf{w}_t) + (\nabla F(\mathbf{w}_t) - \mathbf{v}_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}_t) + \mathbf{v}_t^\top (\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{L_F}{2} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&= F(\mathbf{w}_t) - \eta_t (\nabla F(\mathbf{w}_t) - \mathbf{v}_t)^\top \mathbf{v}_t - \left(\frac{1}{\eta_t} - \frac{L_F}{2}\right) \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&= F(\mathbf{w}_t) + \eta_t \|(\nabla F(\mathbf{w}_t) - \mathbf{v}_t)\|_2^2 - \eta_t (\nabla F(\mathbf{w}_t) - \mathbf{v}_t)^\top \nabla F(\mathbf{w}_t) \\
&- \left(\frac{1}{\eta_t} - \frac{L_F}{2}\right) \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 .
\end{aligned}
$$

Since $(\nabla F(\mathbf{w}_t) - \mathbf{v}_t)^\top \nabla F(\mathbf{w}_t) = \frac{1}{2}\left(\|\nabla F(\mathbf{w}_t) - \mathbf{v}_t\|_2^2 + \|\nabla F(\mathbf{w}_t)\|_2^2 - \|\mathbf{v}_t\|_2^2\right)$, then we have

$$
\begin{aligned}
&F(\mathbf{w}_{t+1}) \leq F(\mathbf{w}_t) + \eta_t \|\nabla F(\mathbf{w}_t) - \mathbf{v}_t\|_2^2 - \left(\frac{1}{\eta_t} - \frac{L_F}{2}\right) \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&- \frac{\eta_t}{2}\left(\|\nabla F(\mathbf{w}_t) - \mathbf{v}_t\|_2^2 + \|\nabla F(\mathbf{w}_t)\|_2^2 - \|\mathbf{v}_t\|_2^2\right) \\
&= F(\mathbf{w}_t) + \frac{\eta_t}{2} \|\nabla F(\mathbf{w}_t) - \mathbf{v}_t\|_2^2 - \frac{\eta_t}{2} \|\nabla F(\mathbf{w}_t)\|_2^2 - \left(\frac{1}{2\eta_t} - \frac{L_F}{2}\right) \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 .
\end{aligned}
$$

□

To prove the final convergence of SCMA, we present a useful lemma.

Lemma 4.10 *If $\eta_t \leq 1/L$, assume that there exist non-negative sequences $A_t, B_t, \Gamma_t, \Delta_t, \delta_t, t \geq 0$ satisfying:*

$$
\begin{aligned}
&(*) A_{t+1} \leq A_t + \eta_t \Delta_t - \eta_t B_t - \eta_t \Gamma_t \\
&(\sharp) \Delta_{t+1} \leq (1 - \beta_{t+1}) \Delta_t + C_1 \beta_{t+1} \delta_{t+1} + \frac{C_2 \eta_t^2}{\beta_{t+1}} \Gamma_t + \beta_{t+1}^2 \sigma^2, \\
&(\diamond) \delta_{t+1} \leq (1 - \gamma_{t+1}) \delta_t + \frac{C_3 \eta_t^2}{\gamma_{t+1}} \Gamma_t + \gamma_{t+1}^2 \sigma'^2.
\end{aligned}
$$

Let $\Upsilon_t = A_t + \frac{\eta_{t-1}}{\beta_t} \Delta_t + \frac{C_1 \eta_{t-1}}{\gamma_t} \delta_t$. If $\frac{\eta_t}{\beta_{t+1}} \leq \frac{\eta_{t-1}}{\beta_t}$, $\frac{\eta_t}{\gamma_{t+1}} \leq \frac{\eta_{t-1}}{\gamma_t}$, $\eta_t \leq \min(\frac{\beta_{t+1}}{\sqrt{4C_2}}, \frac{\gamma_{t+1}}{\sqrt{8C_1C_3}})$, and $\Upsilon_t \geq A_*$, then we have

$$
\sum_{t=0}^{T-1} \frac{1}{\sum_{t=0}^{T-1} \eta_t} (\eta_t B_t + \frac{1}{2} \eta_t \Gamma_t) \leq \frac{C_\Upsilon}{\sum_{t=0}^{T-1} \eta_t} + \frac{\sum_{t=0}^{T-1} \left(\eta_t \beta_{t+1} \sigma^2 + 2 C_1 \eta_t \gamma_{t+1} \sigma'^2\right)}{\sum_{t=0}^{T-1} \eta_t},
$$

where $C_\Upsilon = \Upsilon_0 - A_* \leq A_0 - A_* + \frac{1}{2\sqrt{C_2}} \Delta_0 + \sqrt{\frac{C_1}{8C_3}} \delta_0$.

If $\beta = \frac{\epsilon^2}{3\sigma^2}, \gamma = \frac{\epsilon^2}{6C_1 \sigma'^2}, \eta = \min(\frac{1}{L}, \frac{\beta}{\sqrt{4C_2}}, \frac{\gamma}{\sqrt{8C_1C_3}})$, then in order to guarantee

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

the iteration complexity is the in the order of

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

Critical: If $(*)$, $(\sharp)$, $(\diamond)$ hold in expectation, then the concluding inequalities also hold in expectation.

Proof. The proof is constructive. The idea is to construct a telescoping series of $A_t + a_t\Delta_t + b_t\delta_t$ with some appropriate sequences of a_t, b_t. First, we have

$$\begin{aligned}
&A_{t+1} + a_{t+1}\Delta_{t+1} + b_{t+1}\delta_{t+1} \leq A_t + \eta_t\Delta_t - \eta_t B_t - \eta_t\Gamma_t \\
&+ a_{t+1}(1-\beta_{t+1})\Delta_t + a_{t+1}C_1\beta_{t+1}\delta_{t+1} + a_{t+1}\frac{C_2\eta_t^2}{\beta_{t+1}}\Gamma_t + a_{t+1}\beta_{t+1}^2\sigma^2 \\
&+ b_{t+1}(1-\gamma_{t+1})\delta_t + b_{t+1}\frac{C_3\eta_t^2}{\gamma_{t+1}}\Gamma_t + b_{t+1}\gamma_{t+1}^2\sigma'^2.
\end{aligned}$$

Let $a_{t+1} = \eta_t/\beta_{t+1} \leq \eta_{t-1}/\beta_t$ and $b_{t+1} = C_1\eta_t(1+\gamma_{t+1})/\gamma_{t+1}$, we have

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

Thus,

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

Since $\eta_t/\beta_{t+1} \leq \eta_{t-1}/\beta_t$ and $\eta_t/\gamma_{t+1} \leq \eta_{t-1}/\gamma_t$ and $\gamma_{t+1} \leq 1$, we have

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

Since $C_2\eta_t^3/\beta_{t+1}^2 \le \eta_t/4$ (because $\eta_t \le \beta_{t+1}/\sqrt{4C_2}$) and $2C_3C_1\eta_t^3/\gamma_{t+1}^2 \le \eta_t/4$ (because $\eta_t \le \gamma_{t+1}/\sqrt{8C_1C_3}$), we have

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

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

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

Hence

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

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

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

In order to ensure the RHS is less than ϵ^2, it suffices to have

$$
\beta = \frac{\epsilon^2}{3\sigma^2}, \quad \gamma = \frac{\epsilon^2}{6C_1\sigma'^2}, \quad T = \frac{C_\Upsilon}{3\epsilon^2\eta}.
$$

Since

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

thus the order of T becomes

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

where

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

□

Finally, let us prove the convergence of SCMA.

Theorem 4.3 *Suppose Assumptions 4.1, 4.2 and 4.3 hold. For the SCMA algorithm, set the parameters as follows:* $\beta = \frac{\epsilon^2}{3\sigma^2}$, $\gamma = \frac{\epsilon^2}{6C_1\sigma_0^2}$, *and* $\eta = \min\left(\frac{1}{2L_F}, \frac{\beta}{\sqrt{4C_2}}, \frac{\gamma}{\sqrt{8C_1C_3}}\right)$, *where* $\sigma^2 = G_2^2\sigma_1^2 + G_1^2\sigma_2^2$, $C_1 = 4G_2^2L_1^2$, $C_2 = 4L_F^2$, $C_3 = 2G_2^2$. *Then, the following*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left\{\frac{1}{4}\|\mathbf{v}_t\|_2^2 + \|\nabla F(\mathbf{w}_t)\|_2^2\right\}\right] \leq \epsilon^2$$

holds, with an iteration complexity of

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

where $C_\Upsilon := 2\left(F(\mathbf{w}_0) - F_*\right) + \frac{1}{8L_F}\|\nabla F(\mathbf{w}_0) - \mathbf{v}_0\|_2^2 + \frac{L_1}{2}\|\mathbf{u}_0 - g(\mathbf{w}_0)\|_2^2$.

Why it matters

Insights 1: Theorem 4.3 indicates that SCMA enjoys the same complexity of $O(1/\epsilon^4)$ for finding an ϵ-stationary solution as SGD for ERM. In addition, the averaged estimation error of the moving-average gradient estimator $\mathbf{v}_t$, i.e., $\mathbb{E}[\frac{1}{T}\sum_{t=0}^{T-1}\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2]$, converges to zero as $T \to \infty$.
Insights 2: We can apply the above result to the Momentum method (6.2) for solving the standard stochastic optimization $\min_{\mathbf{w}} F(\mathbf{w}) = \mathbb{E}_\zeta[f(\mathbf{w};\zeta)]$ by setting $L_1 = 0$. The complexity of the Momentum method is

$$T = O\left(\max\left\{\frac{(F(\mathbf{w}_0) - F_*)L_F}{\epsilon^2}, \frac{(F(\mathbf{w}_0) - F_*)\sigma^2 L_F}{\epsilon^4}, \frac{\|\nabla F(\mathbf{w}_0) - \mathbf{v}_0\|_2^2\,\sigma^2}{\epsilon^4}\right\}\right),$$

which is no worse than that of SGD in Theorem 3.3. The key advantage of the Momentum method over SGD is that it ensures the averaged estimation error of the moving-average gradient estimator $\mathbf{v}_t$ converge to zero.
The convergence bound also suggests that it is better to initialize $\mathbf{v}_0$ in a way such that $\|\nabla F(\mathbf{w}_0) - \mathbf{v}_0\|_2^2$ is small, e.g., using the mini-batch gradient at $\mathbf{w}_0$ instead of initializing it to zero.

Proof. The three inequalities in Lemma 4.8, 4.9 and 4.1 that we have proved so far are

$$(*)F(\mathbf{w}_{t+1}) \leq F(\mathbf{w}_t) + \frac{\eta_t}{2}\left\|\nabla F(\mathbf{w}_t) - \mathbf{v}_t\right\|_2^2 - \frac{\eta_t}{2}\left\|\nabla F(\mathbf{w}_t)\right\|_2^2 - \frac{\eta_t}{4}\left\|\mathbf{v}_t\right\|_2^2, t \geq 0$$

$$(\sharp)\mathbb{E}\left[\left\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\right\|_2^2\right] \leq \mathbb{E}[(1-\beta_t)\left\|\mathbf{v}_{t-1} - \nabla F(\mathbf{w}_{t-1})\right\|_2^2]$$
$$+ \mathbb{E}\left[4G_2^2L_1^2\beta_t\left\|\mathbf{u}_t - g(\mathbf{w}_t)\right\|_2^2 + \frac{2L_F^2\eta_{t-1}^2}{\beta_t}\left\|\mathbf{v}_{t-1}\right\|_2^2 + \beta_t^2\sigma^2\right],$$

$$(\diamond)\mathbb{E}\left[\left\|\mathbf{u}_t - g(\mathbf{w}_t)\right\|_2^2\right] \leq \mathbb{E}[(1-\gamma_t)\left\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\right\|_2^2]$$
$$+ \mathbb{E}\left[\frac{G_2^2\eta_{t-1}^2}{\gamma_t}\left\|\mathbf{v}_{t-1}\right\|_2^2 + \gamma_t^2\sigma_0^2\right].$$

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

Then the three inequalities satisfy that in Lemma 4.10 with $C_1 = 4G_2^2L_1^2, C_2 = 4L_F^2, C_3 = 2G_2^2, \sigma^2 = G_1^2\sigma_2^2 + G_2^2\sigma_1^2, \sigma'^2 = \sigma_0^2$. Then $\eta_t, \beta_t, \gamma_t$ satisfy

$$\frac{\eta_t}{\beta_{t+1}} \leq \frac{\eta_{t-1}}{\beta_t}, \frac{\eta_t}{\gamma_{t+1}} \leq \frac{\eta_{t-1}}{\gamma_t}, \eta_t \leq \min(\frac{\beta_{t+1}}{\sqrt{4C_2}}, \frac{\gamma_{t+1}}{\sqrt{8C_1C_3}}).$$

Then we have

$$\mathbb{E}\left[\sum_{t=0}^{T-1}\frac{1}{\sum_{t=0}^{T-1}\eta_t}(\eta_t\|\nabla F(\mathbf{w}_t)\|_2^2 + \frac{\eta_t}{4}\left\|\mathbf{v}_t\right\|_2^2)\right]$$
$$\leq \frac{C_\Upsilon}{\sum_{t=1}\eta_t} + \frac{\sum_{t=0}^{T-1}\left(\eta_t\beta_{t+1}\sigma^2 + 2C_1\eta_t\gamma_{t+1}\sigma_0^2\right)}{\sum_{t=0}^{T-1}\eta_t}.$$

Since the setting of η, γ, β satisfy that in Lemma 4.10, the order of T becomes

$$T = O\left(\max\left\{\frac{C_\Upsilon L_F}{\epsilon^2}, \frac{C_\Upsilon\sigma^2\sqrt{C_2}}{\epsilon^4}, \frac{C_\Upsilon\sqrt{C_1C_3}C_1\sigma_0^2}{\epsilon^4}\right\}\right)$$
$$= O\left(\max\left\{\frac{C_\Upsilon L_F}{\epsilon^2}, \frac{C_\Upsilon\sigma^2 L_F}{\epsilon^4}, \frac{C_\Upsilon L_1^3\sigma_0^2}{\epsilon^4}\right\}\right),$$

where

$$C_\Upsilon = 2(F(\mathbf{w}_0) - F_*) + \frac{1}{2\sqrt{C_2}}\|\mathbf{v}_0 - \nabla F(\mathbf{w}_0)\|_2^2 + \frac{\sqrt{C_1}}{\sqrt{8C_3}}\|\mathbf{u}_0 - g(\mathbf{w}_0)\|_2^2$$
$$= 2(F(\mathbf{w}_0) - F_*) + \frac{1}{4L_F}\left\|\mathbf{v}_0 - \nabla F(\mathbf{w}_0)\right\|_2^2 + \frac{L_1}{2}\left\|\mathbf{u}_0 - g(\mathbf{w}_0)\right\|_2^2.$$

□

4.3.2 STORM Estimators

We can further reduce the error of the gradient estimator by using advanced variance reduction techniques under stronger assumptions. We make the following assumptions.

Assumption 4.6. *There exists* $L_1, G_1 > 0$ *such that*

(i) $\mathbb{E}[\|\nabla f(g;\xi) - \nabla f(g';\zeta)\|_2^2] \leq L_1^2\|g - g'\|_2^2, \forall g, g';$
(ii) $\mathbb{E}[\|\nabla f(g;\xi)\|_2^2] \leq G_1^2, \forall g.$

Assumption 4.7. *There exists* $L_2, G_2 > 0$ *such that*

(i) $\mathbb{E}[\|\nabla g(\mathbf{w};\zeta) - \nabla g(\mathbf{w}';\zeta)\|_2^2] \leq L_2^2\|\mathbf{w} - \mathbf{w}'\|_2^2, \forall \mathbf{w}, \mathbf{w}';$
(ii) $\mathbb{E}[\|\nabla g(\mathbf{w};\zeta)\|_2^2] \leq G_2^2, \forall \mathbf{w}.$

Due to Jensen's inequality, Assumption (4.6)(i) implies the Lipschitz continuity assumption of ∇f in Assumption (4.1)(i). Similarly, Assumption (4.7)(i) implies that in Assumption 4.2(i), respectively. Hence, Assumption (4.6)(i) and Assumption (4.7)(i) are stronger, which are referred to as mean-square smoothness condition of f and g.

The STORM estimator

Let us first discuss a generic STORM estimator, an improved variant of the moving average estimator. Without loss of generality, we consider estimating a sequence of mappings $\{\mathcal{M}(\mathbf{w}_t)\}_{t=1}^T$ through their stochastic values at each iteration $\{\mathcal{M}(\mathbf{w}_t;\zeta_t)\}_{t=1}^T$, where $\mathbb{E}_{\zeta_t}[\mathcal{M}(\mathbf{w}_t;\zeta_t)] = \mathcal{M}(\mathbf{w}_t) \in \mathbb{R}^{d'}$. We assume the mapping $\mathcal{M}$ satisfies:

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

The STORM estimator is give by a sequence of $\mathcal{U}_1, \ldots, \mathcal{U}_T$, where

$$\mathcal{U}_t = (1-\gamma_t)\mathcal{U}_{t-1} + \gamma_t\mathcal{M}(\mathbf{w}_t;\zeta_t) + (1-\gamma_t)(\mathcal{M}(\mathbf{w}_t;\zeta_t) - \mathcal{M}(\mathbf{w}_{t-1};\zeta_t)), \tag{4.23}$$

and $\gamma_t \in (0,1)$.
It augments the moving-average estimator by adding an extra term $(1-\gamma_t)(\mathcal{M}(\mathbf{w}_t;\zeta_t) - \mathcal{M}(\mathbf{w}_{t-1};\zeta_t))$, which can be viewed as an error correction term.

Applying the STORM estimator to estimating the sequence of $\{g(\mathbf{w}_t)\}_{t\geq 1}$, we have the following sequence:

$$\mathbf{u}_t = (1-\gamma_t)\mathbf{u}_{t-1} + \gamma_t g(\mathbf{w}_t;\zeta_t) + (1-\gamma_t)(g(\mathbf{w}_t;\zeta_t) - g(\mathbf{w}_{t-1};\zeta_t)). \tag{4.24}$$

Given $\mathbf{u}_t$, we can compute a moving-average gradient estimator (4.16) similar to SCMA. However, this will not yield an improved rate compared with SCMA. To

Algorithm 11 SCST

Input: learning rate schedules $\{\eta_t\}_{t=0}^T$, $\{\gamma_t\}_{t=1}^T$; 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**
  Sample ζ_t, ζ_t' and ξ_t
  Update the inner function value estimator

$$\mathbf{u}_t = (1-\gamma_t)\mathbf{u}_{t-1} + \gamma_t g(\mathbf{w}_t;\zeta_t) + (1-\gamma_t)(g(\mathbf{w}_t;\zeta_t) - g(\mathbf{w}_{t-1};\zeta_t))$$

  Compute the vanilla gradient estimator $\mathbf{z}_t = \nabla g(\mathbf{w}_t;\zeta_t')\nabla f(\mathbf{u}_t;\xi_t)$
  Compute $\tilde{\mathbf{z}}_{t-1} = \nabla g(\mathbf{w}_{t-1};\zeta_t')\nabla f(\mathbf{u}_{t-1};\xi_t)$
  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 by $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{v}_t$
end for

reduce the estimator error of the gradient, we apply another STORM estimator to estimate $\mathcal{M}_t = \nabla g(\mathbf{w}_t)\nabla f(\mathbf{u}_t)$. This is computed by the following sequence:

$$\begin{aligned}\mathbf{v}_t &= (1-\beta_t)\mathbf{v}_{t-1} + \beta_t \nabla g(\mathbf{w}_t;\zeta_t')\nabla f(\mathbf{u}_t;\xi_t) \\ &\quad + (1-\beta_t)(\nabla g(\mathbf{w}_t;\zeta_t')\nabla f(\mathbf{u}_t;\xi_t) - \nabla g(\mathbf{w}_{t-1};\zeta_t')\nabla f(\mathbf{u}_{t-1};\xi_t)).\end{aligned} \tag{4.25}$$

With $\mathbf{v}_t$, we update the model parameters by

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

The full steps of this method is presented in Algorithm 11, which is referred to as SCST.

Connection with Variance-reduced metholds for Non-convex optimization

In the special case where f is the identity function, the update is identical to the classical variance-reduced method (also known as STROM) for non-convex optimization $\min_{\mathbf{w}} \mathbb{E}_\zeta [g(\mathbf{w};\zeta)]$, i.e.,

$$\begin{aligned}&\mathbf{v}_t = (1-\beta_t)\mathbf{v}_{t-1} + \beta_t \nabla g(\mathbf{w}_t;\zeta_t') + (1-\beta_t)(\nabla g(\mathbf{w}_t;\zeta_t') - \nabla g(\mathbf{w}_{t-1};\zeta_t')), \\ &\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{v}_t.\end{aligned} \tag{4.26}$$

It is renowned for its improved complexity of $O(1/\epsilon^3)$ better than the complexity $O(1/\epsilon^4)$ of SGD for finding an ϵ-stationary solution.

Convergence Analysis

We first prove a general result of the STORM estimator that applies to both $\mathbf{u}_t$ and $\mathbf{v}_t$.

Lemma 4.11 *Consider* $\mathbf{v}_t = (1-\beta_t)\mathbf{v}_{t-1} + \beta_t \mathbf{z}_t + (1-\beta_t)(\mathbf{z}_t - \tilde{\mathbf{z}}_{t-1})$, *where* $\beta_t \in (0,1)$. *Let* $\mathbb{E}_t$ *denote the expectation over randomness associated with* $\mathbf{z}_t, \tilde{\mathbf{z}}_{t-1}$ *condition on the randomness before t-the iteration. If* $\mathbb{E}_t[\mathbf{z}_t] = \mathcal{M}_t$ *and* $\mathbb{E}_t[\tilde{\mathbf{z}}_{t-1}] = \mathcal{M}_{t-1}$. *If* $\mathbb{E}_t[\|\mathbf{z}_t - \mathcal{M}_t\|_2^2] \leq \sigma^2$, *then we have*

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

Proof.

$$\begin{aligned}
&\mathbb{E}_t\left[\|\mathbf{v}_t - \mathcal{M}_t\|_2^2\right] \\
&= \mathbb{E}_t\left[\|(1-\beta_t)\mathbf{v}_{t-1} - \mathcal{M}_t + \beta_t\mathbf{z}_t + (1-\beta_t)(\mathbf{z}_t - \tilde{\mathbf{z}}_{t-1})\|_2^2\right] \\
&= \mathbb{E}_t\left[\|(1-\beta_t)(\mathbf{v}_{t-1} - \mathcal{M}_{t-1}) + (1-\beta_t)((\mathbf{z}_t - \tilde{\mathbf{z}}_{t-1}) - (\mathcal{M}_t - \mathcal{M}_{t-1})) \right.\\
&\qquad \left. + \beta_t(\mathbf{z}_t - \mathcal{M}_t)\|_2^2\right].
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathbb{E}_t\left[\langle(1-\beta_t)(\mathbf{v}_{t-1} - \mathcal{M}_{t-1}),\right. \\
&\quad \left.(1-\beta_t)((\mathbf{z}_t - \tilde{\mathbf{z}}_{t-1}) - (\mathcal{M}_t - \mathcal{M}_{t-1})) + \beta_t(\mathbf{z}_t - \mathcal{M}_t)\rangle\right] = 0.
\end{aligned}$$

Then,

$$\begin{aligned}
&\mathbb{E}_t\left[\|\mathbf{v}_t - \mathcal{M}_t\|_2^2\right] \leq (1-\beta_t)^2\|\mathbf{v}_t - \mathcal{M}_{t-1}\|_2^2 \\
&+ \|(1-\beta_t)((\mathbf{z}_t - \tilde{\mathbf{z}}_{t-1}) - (\mathcal{M}_t - \mathcal{M}_{t-1})) + \beta_t(\mathbf{z}_t - \mathcal{M}_t)\|_2^2 \\
&\overset{(\diamond)}{\leq} (1-\beta_t)^2\|\mathbf{v}_t - \mathcal{M}_{t-1}\|_2^2 \\
&+ 2(1-\beta_t)^2\mathbb{E}_t[\|((\mathbf{z}_t - \tilde{\mathbf{z}}_{t-1}) - (\mathcal{M}_t - \mathcal{M}_{t-1}))\|_2^2] + 2\beta_t^2\mathbb{E}_t[\|\mathbf{z}_t - \mathcal{M}_t\|_2^2] \\
&\overset{(*)}{\leq} (1-\beta_t)^2\|\mathbf{v}_{t-1} - \mathcal{M}_{t-1}\|_2^2 + 2(1-\beta_t)^2\mathbb{E}_t[\|\mathbf{z}_t - \tilde{\mathbf{z}}_{t-1}\|_2^2] + 2\beta_t^2\sigma^2,
\end{aligned}$$

where $(\diamond)$ uses the Young's inequality, $(*)$ uses the fact that $\mathbb{E}[\|a - \mathbb{E}[a]\|_2^2] \leq \mathbb{E}[\|a\|_2^2]$, and $\mathbb{E}_t[\mathbf{z}_t - \tilde{\mathbf{z}}_{t-1}] = \mathcal{M}_t - \mathcal{M}_{t-1}$. □

Let us first prove an error recursion of $\mathbf{u}_t$ in the lemma below.

Lemma 4.12 *Under Assumption (4.7)(ii), we have:*

$$\begin{aligned}
&\mathbb{E}_{\zeta_t}\left[\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2\right] \leq (1-\gamma_t)\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2^2 + 2\gamma_t^2\sigma_0^2 + 2G_2^2\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2 \\
&\mathbb{E}_{\zeta_t}\left[\|\mathbf{u}_t - \mathbf{u}_{t-1}\|_2^2\right] \leq 2\gamma_t^2\sigma_0^2 + 4\gamma_t^2\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2^2 + 6G_2^2\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2.
\end{aligned}$$

Why it matters

Compared to the error recursion of $\mathbf{u}_t$ to that in Lemma 4.1, the improvement comes from the last term reducing from $\frac{2G_2^2\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2}{\gamma_t}$ to $2G_2^2\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2$.

Proof. The first part follows directly from Lemma 4.11 by noting the mean-Lipschitz continuity of $g(\mathbf{w};\zeta)$. To prove the second part, we proceed as follows:

$$\begin{aligned}
&\mathbb{E}_t\left[\|\mathbf{u}_t-\mathbf{u}_{t-1}\|_2^2\right]\\
=&\mathbb{E}_t\left[\|\gamma_t\left(g(\mathbf{w}_t;\zeta_t)-\mathbf{u}_{t-1}\right)+(1-\gamma_t)\left(g(\mathbf{w}_t;\zeta_t)-g(\mathbf{w}_{t-1};\zeta_t)\right)\|_2^2\right]\\
\le&\mathbb{E}_t\left[2\gamma_t^2\|(g(\mathbf{w}_t;\zeta_t)-\mathbf{u}_{t-1})\|_2^2+2(1-\gamma_t)^2\|g(\mathbf{w}_t;\zeta_t)-g(\mathbf{w}_{t-1};\zeta_t)\|_2^2\right]\\
\le&\mathbb{E}_t\left[2\gamma_t^2\|(g(\mathbf{w}_t;\zeta_t)-\mathbf{u}_{t-1})\|_2^2\right]+2(1-\gamma_t)^2G_2^2\|\mathbf{w}_t-\mathbf{w}_{t-1}\|_2^2.
\end{aligned}$$

Next, we bound the first term on the RHS as

$$\begin{aligned}
&\mathbb{E}_t\left[2\gamma_t^2\|(g(\mathbf{w}_t;\zeta_t)-\mathbf{u}_{t-1})\|_2^2\right]=\mathbb{E}_t\left[2\gamma_t^2\|(g(\mathbf{w}_t;\zeta_t)-g(\mathbf{w}_t)+g(\mathbf{w}_t)-\mathbf{u}_{t-1})\|_2^2\right]\\
&\le 2\gamma_t^2\sigma_0^2+2\gamma_t^2\|g(\mathbf{w}_t)-\mathbf{u}_{t-1}\|_2^2\\
&\le 2\gamma_t^2\sigma_0^2+2\gamma_t^2\|g(\mathbf{w}_t)-g(\mathbf{w}_{t-1})+g(\mathbf{w}_{t-1})-\mathbf{u}_{t-1}\|_2^2\\
&\le 2\gamma_t^2\sigma_0^2+4\gamma_t^2\|g(\mathbf{w}_{t-1})-\mathbf{u}_{t-1}\|_2^2+4\gamma_t^2G_2^2\|\mathbf{w}_t-\mathbf{w}_{t-1}\|_2^2,
\end{aligned}$$

where the first inequality uses the fact $\mathbb{E}\left[g(\mathbf{w}_t;\zeta_t)-g(\mathbf{w}_t)\right]=0$. Combining the above results, we finish the proof. □

Next, we build an error recursion of $\|\mathbf{v}_t-\mathcal{M}_t\|_2^2$.

Lemma 4.13 *Let $\sigma^2=G_2^2\sigma_1^2+G_1^2\sigma_2^2$. Under Assumptions (4.6) and Assumption (4.7), (4.25) satisfies that*

$$\begin{aligned}
&\mathbb{E}_{\zeta_t',\xi_t}\left[\|\mathbf{v}_t-\mathcal{M}_t\|_2^2\right]\le(1-\beta_t)\|\mathbf{v}_{t-1}-\mathcal{M}_{t-1}\|_2^2 \qquad (4.27)\\
&+16G_2^2L_1^2\gamma_t^2\|\mathbf{u}_{t-1}-g(\mathbf{w}_{t-1})\|_2^2+(24G_2^4L_1^2+4G_1^2L_2^2)\|\mathbf{w}_t-\mathbf{w}_{t-1}\|_2^2\\
&+2\beta_t^2\sigma^2+8G_2^2L_1^2\gamma_t^2\sigma_0^2.
\end{aligned}$$

Proof. First, (4.21) gives $\mathbb{E}_t[\|\mathbf{z}_t-\mathcal{M}_t\|_2^2]\le\sigma^2$. Second,

$$\begin{aligned}
&\mathbb{E}_t\left[\|\mathbf{z}_t-\tilde{\mathbf{z}}_{t-1}\|_2^2\right]\\
&=\mathbb{E}_t[\|\nabla g(\mathbf{w}_t;\zeta_t')\nabla f(\mathbf{u}_t;\xi_t)-\nabla g(\mathbf{w}_{t-1};\zeta_t')\nabla f(\mathbf{u}_{t-1};\xi_t)\|_2^2]\\
&=\mathbb{E}_t[\|\nabla g(\mathbf{w}_t;\zeta_t')\nabla f(\mathbf{u}_t;\xi_t)-\nabla g(\mathbf{w}_t;\zeta_t')\nabla f(\mathbf{u}_{t-1};\xi_t)\\
&\qquad+\nabla g(\mathbf{w}_t;\zeta_t')\nabla f(\mathbf{u}_{t-1};\xi_t)-\nabla g(\mathbf{w}_{t-1};\zeta_t')\nabla f(\mathbf{u}_{t-1};\xi_t)\|_2^2]\\
&\overset{(\triangle)}{\le}2G_2^2L_1^2\|\mathbf{u}_t-\mathbf{u}_{t-1}\|_2^2+2G_1^2L_2^2\|\mathbf{w}_t-\mathbf{w}_{t-1}\|_2^2,
\end{aligned}$$

where $(\triangle)$ uses the Assumption (4.6)(i) and Assumption (4.7)(i). It then follows:

$$\begin{aligned}
\mathbb{E}_t\left[\|\mathbf{v}_t-\mathcal{M}_t\|_2^2\right]\le&(1-\beta_t)^2\|\mathbf{v}_{t-1}-\mathcal{M}_{t-1}\|_2^2\\
&+4G_2^2L_1^2\|\mathbf{u}_t-\mathbf{u}_{t-1}\|_2^2+4G_1^2L_2^2\|\mathbf{w}_t-\mathbf{w}_{t-1}\|_2^2+2\beta_t^2\sigma^2.
\end{aligned}$$

By using the second inequality of Lemma **??**, i.e.,

$$\mathbb{E}_{\zeta_t}\left[\|\mathbf{u}_t-\mathbf{u}_{t-1}\|_2^2\right]\leq 2\gamma_t^2\sigma_0^2+4\gamma_t^2\|\mathbf{u}_{t-1}-g(\mathbf{w}_{t-1})\|_2^2+6G_2^2\|\mathbf{w}_t-\mathbf{w}_{t-1}\|_2^2,$$

we have

$$\begin{aligned}\mathbb{E}_t\left[\|\mathbf{v}_t-\mathcal{M}_t\|_2^2\right]&\leq(1-\beta_t)\|\mathbf{v}_{t-1}-\mathcal{M}_{t-1}\|_2^2+16G_2^2L_1^2\gamma_t^2\|\mathbf{u}_{t-1}-g(\mathbf{w}_{t-1})\|_2^2\\&+(24G_2^4L_1^2+4G_1^2L_2^2)\|\mathbf{w}_t-\mathbf{w}_{t-1}\|_2^2+2\beta_t^2\sigma^2+8G_2^2L_1^2\gamma_t^2\sigma_0^2.\end{aligned}$$

□

Similar to Lemma 4.9, we have the following descent lemma.

Lemma 4.14 *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*

$$\begin{aligned}F(\mathbf{w}_{t+1})\leq F(\mathbf{w}_t)&+\eta_tG_2^2L_1^2\|\mathbf{u}_t-g(\mathbf{w}_t)\|_2^2+\eta_t\|\mathbf{v}_t-H_t\|_2^2\\&-\frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t)\|_2^2-\frac{1}{4\eta_t}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2.\end{aligned}\tag{4.28}$$

This lemma can be proved following that of lemma 4.9 by bound $\|\mathbf{v}_t-\nabla F(\mathbf{w}_t)\|_2^2\leq 2\|\mathbf{v}_t-\mathcal{M}_t\|_2^2+2\|\mathcal{M}_t-\nabla F(\mathbf{w}_t)\|_2^2\leq 2\|\mathbf{v}_t-\mathcal{M}_t\|_2^2+2G_2^2L_1^2\|\mathbf{u}_t-g(\mathbf{w}_t)\|_2^2$.

Lemma 4.15 *For* $\eta_t\leq 1/L$, *the non-negative sequences* $A_t,B_t,\Gamma_t,\Delta_t,\delta_t,t\geq 0$ *satisfy:*

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

Let $\Upsilon_{t+1}=A_{t+1}+\frac{c}{\eta_t}\Delta_{t+1}+\frac{c'}{\eta_t}\delta_{t+1}\geq A_*$. *Suppose* $c,c',\eta_t,\gamma_t,\beta_t$ *satisfy:*

$$\begin{aligned}&C_2c+C_3c'\leq\frac{1}{2},\quad \eta_t+\frac{c}{\eta_t}(1-\beta_{t+1})\leq\frac{c}{\eta_{t-1}},\\&\eta_t+\frac{c}{\eta_t}C_1\gamma_{t+1}^2+\frac{c'}{\eta_t}(1-\gamma_{t+1})\leq\frac{c'}{\eta_{t-1}}.\end{aligned}\tag{4.29}$$

Then,

$$\sum_{t=0}^{T-1}(\eta_tB_t+\frac{1}{2}\eta_t\Gamma_t)\leq C_\Upsilon+\sum_{t=0}^{T-1}\left(\frac{c\beta_{t+1}^2}{\eta_t}\sigma^2+\frac{c\gamma_{t+1}^2}{\eta_t}\sigma'^2+\frac{c'\gamma_{t+1}^2}{\eta_t}\sigma''^2\right).\tag{4.30}$$

If we set $c=\frac{1}{4C_2},c'=\frac{1}{4C_3},\beta_t=\frac{\epsilon\eta\sqrt{C_2}}{\sigma}$, $\gamma_t=\min\left(\frac{\epsilon\eta\sqrt{C_2}}{\sigma'},\frac{\epsilon\eta\sqrt{C_3}}{\sigma''},\frac{C_2}{2C_3C_1}\right)$, *and* $\eta_t=\eta=\min\left(\frac{1}{L},\frac{\epsilon}{4\sqrt{C_2}\sigma},\frac{\epsilon\sqrt{C_2}}{8C_3\sigma'},\frac{\epsilon}{8\sqrt{C_3}\sigma''},\frac{\sqrt{C_2}}{4C_3\sqrt{C_1}}\right)$, *then in order to grantee*

$$\sum_{t=0}^{T-1} \frac{1}{T}(B_t + \frac{1}{2}\Gamma_t) \leq \epsilon^2, \tag{4.31}$$

the iteration complexity is in the order of

$$T = O\left(\max\left\{\frac{C_\Upsilon L}{\epsilon^2}, \frac{C_\Upsilon C_3\sqrt{C_1/C_2}}{\epsilon^2}, \frac{C_\Upsilon\sigma\sqrt{C_2}}{\epsilon^3}, \frac{C_\Upsilon C_3\sigma'}{\epsilon^3\sqrt{C_2}}, \frac{C_\Upsilon\sigma''\sqrt{C_3}}{\epsilon^3}\right\}\right)$$

where $C_\Upsilon = \Upsilon_0 - A_* = A_0 + \frac{1}{4C_2\eta}\Delta_0 + \frac{1}{4C_3\eta}\delta_0 - A_*$.

Critical: If $(*)$, $(\sharp)$, $(\diamond)$ hold in expectation, then the two inequalities in (4.30) and (4.31) hold in expectation.

Proof. The proof is constructive. The idea is to multiply the second inequality by a_{t+1} and the third inequality by b_{t+1} such that we can construct a telescoping series of $A_t + a_t\Delta_t + b_t\delta_t$. First, we have

$$\begin{aligned}
&A_{t+1} + a_{t+1}\Delta_{t+1} + b_{t+1}\delta_{t+1} \leq A_t + \eta_t\Delta_t + \eta_t\delta_t - \eta_t B_t - \eta_t\Gamma_t \\
&+ a_{t+1}(1-\beta_{t+1})\Delta_t + a_{t+1}C_1\gamma_{t+1}^2\delta_t + a_{t+1}C_2\eta_t^2\Gamma_t + a_{t+1}\beta_{t+1}^2\sigma^2 + a_{t+1}\gamma_{t+1}^2\sigma'^2. \\
&+ b_{t+1}(1-\gamma_{t+1})\delta_t + b_{t+1}C_3\eta_t^2\Gamma_t + b_{t+1}\gamma_{t+1}^2\sigma''^2.
\end{aligned}$$

Let $a_{t+1} = c/\eta_t$ and $b_{t+1} = c'/\eta_t$, we have

$$\begin{aligned}
&A_{t+1} + \frac{c}{\eta_t}\Delta_{t+1} + \frac{c'}{\eta_t}\delta_{t+1} \leq A_t - \eta_t B_t - \eta_t\Gamma_t \\
&+ \left(\eta_t + \frac{c}{\eta_t}(1-\beta_{t+1})\right)\Delta_t + C_2 c\eta_t\Gamma_t + \frac{c\beta_{t+1}^2}{\eta_t}\sigma^2 + \frac{c\gamma_{t+1}^2}{\eta_t}\sigma'^2 \\
&+ \left(\eta_t + \frac{c}{\eta_t}C_1\gamma_{t+1}^2 + \frac{c'}{\eta_t}(1-\gamma_{t+1})\right)\delta_t + C_3 c'\eta_t\Gamma_t + \frac{c'\gamma_{t+1}^2}{\eta_t}\sigma''^2.
\end{aligned}$$

With (4.29) we have

$$\begin{aligned}
&A_{t+1} + \frac{c}{\eta_t}\Delta_{t+1} + \frac{c'}{\eta_t}\delta_{t+1} \leq A_t + \frac{c}{\eta_{t-1}}\Delta_t + \frac{c'}{\eta_{t-1}}\delta_t - \eta_t B_t - \frac{1}{2}\eta_t\Gamma_t \\
&+ \frac{c\beta_{t+1}^2}{\eta_t}\sigma^2 + \frac{c\gamma_{t+1}^2}{\eta_t}\sigma'^2 + \frac{c'\gamma_{t+1}^2}{\eta_t}\sigma''^2
\end{aligned}$$

Define $\Upsilon_{t+1} = A_{t+1} + \frac{c}{\eta_t}\Delta_{t+1} + \frac{c'}{\eta_t}\delta_{t+1}$, we have

$$\eta_t B_t + \frac{1}{2}\eta_t\Gamma_t \leq \Upsilon_t - \Upsilon_{t+1} + \frac{c\beta_{t+1}^2}{\eta_t}\sigma^2 + \frac{c\gamma_{t+1}^2}{\eta_t}\sigma'^2 + \frac{c'\gamma_{t+1}^2}{\eta_t}\sigma''^2.$$

Hence

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

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

$$\sum_{t=0}^{T-1}\frac{1}{T}(B_t + \frac{1}{2}\Gamma_t) \leq \frac{\Upsilon_0 - A_*}{\eta T} + \left(\frac{c\beta^2}{\eta^2}\sigma^2 + \frac{c\gamma^2}{\eta^2}\sigma'^2 + \frac{c'\gamma^2}{\eta^2}\sigma''^2\right).$$

In order to ensure the RHS is less than ϵ^2, it suffices to have

$$\beta = \frac{\epsilon\eta}{2\sqrt{c}\sigma}, \quad \gamma = \min\left(\frac{\epsilon\eta}{2\sqrt{c}\sigma'}, \frac{\epsilon\eta}{2\sqrt{c'}\sigma''}\right), \quad T = \frac{C_\Upsilon}{4\epsilon^2\eta}.$$

To ensure (4.29), it suffices to have

$$\eta^2 \leq c\beta, \quad C_1 c\gamma \leq c'/2, \quad \eta^2 \leq c'\gamma/2, \quad c = \frac{1}{4C_2}, \quad c' = \frac{1}{4C_3}.$$

As a result, if we set

$$\begin{aligned}\eta &= \min\left(\frac{1}{L}, \frac{\epsilon\sqrt{c}}{2\sigma}, \frac{\epsilon c'}{4\sqrt{c}\sigma'}, \frac{\epsilon\sqrt{c'}}{4\sigma''}, \frac{c'}{2\sqrt{cC_1}}\right)\\ &= \min\left(\frac{1}{L}, \frac{\epsilon}{4\sqrt{C_2}\sigma}, \frac{\epsilon\sqrt{C_2}}{8C_3\sigma'}, \frac{\epsilon}{8\sqrt{C_3}\sigma''}, \frac{\sqrt{C_2}}{4C_3\sqrt{C_1}}\right)\\ \beta &= \frac{\epsilon\eta\sqrt{C_2}}{\sigma}, \quad \gamma = \min\left(\frac{\epsilon\eta\sqrt{C_2}}{\sigma'}, \frac{\epsilon\eta\sqrt{C_3}}{\sigma''}, \frac{C_2}{2C_3C_1}\right),\end{aligned}$$

we have

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

Plugging the values of η into the requirement of T yields the order of T.

□

Theorem 4.4 *Suppose that Assumptions 4.3, 4.6,and 4.7 hold. For SCST, in order to guarantee*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left\{\frac{1}{4}\|\mathbf{v}_t\|_2^2 + \|\nabla F(\mathbf{w}_t)\|_2^2\right\}\right] \leq \epsilon^2,$$

we can set the parameters as $\eta = \min\{O(\frac{1}{L_F}), O(\frac{\epsilon}{L_1\sigma}), O(\frac{\epsilon}{L_1^2\sigma_0})\}$, $\beta = O(\frac{\epsilon\eta L_1}{\sigma})$, *and* $\gamma = \min\{O(\frac{\epsilon\eta}{\sigma_0}, O(1)\}$, *and the iteration complexity is*

$$T = O\left(\max(\frac{C_\Upsilon L_1(\sigma_1+\sigma_2)}{\epsilon^3}, \frac{C_\Upsilon\sigma_0 L_1^2}{\epsilon^3}, \frac{C_\Upsilon L_F}{\epsilon^2})\right),$$

where $C_Y = O(F(\mathbf{w}_0) - F_* + \frac{1}{L_1^2\eta}\|\nabla g(\mathbf{w}_0)\nabla f(\mathbf{u}_0) - \mathbf{v}_0\|_2^2 + \frac{1}{L_1^2\eta}\|g(\mathbf{w}_0) - \mathbf{u}_0\|_2^2)$.

Why it matters

We only explicitly maintain the dependence on L_1, which will have implications when we handle non-smooth f in next Chapter.

The above theorem can help us establish an improved iteration complexity of $O(1/\epsilon^3)$. First, we need to ensure $C_Y = O(1)$, which can be satisfied by using a large initial batch size. In particular, we can set $\mathbf{u}_0 = \frac{1}{B_0}\sum_{i=1}^{B_0} g(\mathbf{w}_0;\zeta_i)$, $\mathbf{v}_0 = \frac{1}{B_0}\sum_{i=1}^{B_0}\nabla g(\mathbf{w}_0;\zeta_i')\nabla f(\mathbf{u}_0;\xi_i)$, where $\{\zeta_i,\zeta_i',\zeta_i\}_{i=1}^{B_0}$ are independent random variables. Thus, we have $\mathbb{E}[\|\mathbf{u}_0 - g(\mathbf{w}_0)\|_2^2] \leq O(\frac{1}{B_0})$ and $\mathbb{E}[\|\mathbf{v}_0 - \nabla g(\mathbf{w}_0)\nabla f(\mathbf{u}_0)\|_2^2] \leq O(\frac{1}{B_0})$. Hence, if we set $B_0 = O(\frac{\sigma}{L_1\epsilon}, \frac{\sigma_0}{\epsilon})$ we have $C_Y = O(1)$. This initial batch size requirement can be removed by using a decreasing parameters $\eta_t = O(1/t^{1/3}), \beta_t = O(1/t^{2/3}), \gamma_t = O(1/t^{2/3})$.

Compared to the result of SCMA in Theorem 4.3, SCST has a higher order of step size η and a smaller order of iteration complexity.

Proof. Let us recall the three inequalities in Lemma 4.14, 4.13 and 4.12:

$$(*)\ F(\mathbf{w}_{t+1}) \leq F(\mathbf{w}_t) + \eta_t G_2^2 L_1^2\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2 + \eta_t\|\mathbf{v}_t - \mathcal{M}_t\|_2^2 - \frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t)\|_2^2 - \frac{1}{4\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2,$$

$$(\sharp)\ \mathbb{E}\left[\|\mathbf{v}_t - \mathcal{M}_t\|_2^2\right] \leq \mathbb{E}[(1-\beta_t)\|\mathbf{v}_{t-1} - \mathcal{M}_{t-1}\|_2^2] + 16G_2^2L_1^2\gamma_t^2\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2^2] + \mathbb{E}[(24G_2^4L_1^2 + 4G_1^2L_2^2)\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2 + 2\beta_t^2\sigma^2 + 8G_2^2L_1^2\gamma_t^2\sigma_0^2],$$

$$(\diamond)\ \mathbb{E}_{\zeta_t}\left[\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2\right] \leq (1-\gamma_t)\|\mathbf{u}_{t-1} - g(\mathbf{w}_{t-1})\|_2^2 + \mathbb{E}[2G_2^2\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2 + 2\gamma_t^2\sigma_0^2].$$

Define

$$A_t = F(\mathbf{w}_t) - F_*, \quad B_t = \|\nabla F(\mathbf{w}_t)\|_2^2/2,$$
$$\Gamma_t = \|\mathbf{v}_t\|_2^2/4, \quad \Delta_t = \|\mathbf{v}_t - H_t\|_2^2, \quad \delta_t = L_1^2G_2^2\|\mathbf{u}_t - g(\mathbf{w}_t)\|_2^2.$$

They satisfy the three inequalities marked by $*, \sharp, \diamond$ in Lemma 4.15 with Then we have $C_1 = 16, C_2 = O(G_2^4L_1^2 + G_1^2L_2^2), C_3 = O(L_1^2G_2^2), \sigma^2 = O(G_2^2\sigma_1^2 + G_1^2\sigma_2^2), \sigma'^2 = O(L_1^2G_2^2\sigma_0^2), \sigma''^2 = O(L_1^2G_2^2\sigma_0^2)$. Plugging these into Lemma 4.15, we can finish the proof. □

4.4 Non-smooth (Non-convex) Regularized Problems

In this section, we consider the following regularized stochastic compositional optimization:

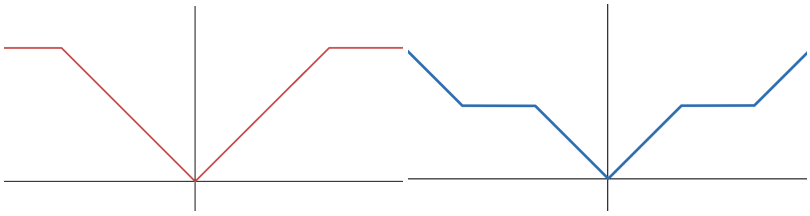

Fig. 4.1: Left: the capped ℓ_1-norm regularizer; Right: a non-convex PAR regularizer

$$\min_{\mathbf{w}\in\mathbb{R}^d} \bar{F}(\mathbf{w}) := \mathbb{E}_\xi f(\mathbb{E}_\zeta[g(\mathbf{w};\zeta)];\xi) + r(\mathbf{w}), \tag{4.32}$$

where r is a non-smooth regularizer, which is potentially non-convex. This includes constrained problems, where $r(\mathbf{w}) = \mathbb{I}(\mathbf{w} \in \mathcal{W})$. For example, the KL-constrained DRO (2.19) has a constraint $\lambda \geq 0$.

We extend the definition of ϵ-stationary solution of a smooth function to the non-smooth composite function by noting that $\partial(F + r)(\mathbf{w}) = \nabla F(\mathbf{w}) + \partial r(\mathbf{w})$.

Definition 4.1 (ϵ-stationary solution) A solution $\mathbf{w}$ is called an ϵ-stationary solution to $\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) + r(\mathbf{w})$ where F is smooth and r is non-differentiable, if $\text{dist}(0, \nabla F(\mathbf{w}) + \partial r(\mathbf{w})) \leq \epsilon$.

To handle non-smoothness or r, we assume the proximal mapping of r is simple to compute:

$$\text{prox}_r(\hat{\mathbf{w}}) = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{w} - \hat{\mathbf{w}}\|_2^2 + r(\mathbf{w}).$$

Below, we give some examples of non-convex regularizers and their proximal mappings, whose derivations are left as exercises for interested readers.

Examples

Example 4.4 (Capped ℓ_1-norm)**.** *It is defined as $r(\mathbf{w}) = \lambda \sum_{i=1}^d \psi(w_i)$, where $\psi(w_i) = \min(|w_i|, \theta)$ (cf. Figure (4.1)). It penalizes small coefficients heavily (encouraging sparsity) but stops penalizing once coefficients are large enough. It was shown to reduce the bias issue of LASSO, which cannot exactly recover the non-zero coefficients under some conditions. Its proximal mapping is given by*

$$prox_{\lambda\psi}(u) = \begin{cases} x_1 = \min(sign(u)(|u| - \lambda)_+, \theta) & \text{if } h(x_1;u) < h(x_2;u) \\ x_2 = \max(|u|, \theta) & \text{otherwise}, \end{cases} \tag{4.33}$$

where $h(x;u) = \frac{1}{2}(x-u)^2 + \lambda \min(|x|, \theta)$. *Similar non-convex sparse regularizers include minimax concave penalty (MCP) and Smoothly Clipped Absolute Deviation (SCAD).*

Example 4.5 (Nonconvex Piecewise Affine Regularization (PAR))**.** *A non-convex PAR is defined as* $r(\mathbf{w}) = \lambda \sum_{i=1}^{d} \psi(w_i)$ *(cf Figure (4.1)), where*

$$\psi(x) = \begin{cases} |x| - kq & \text{if } kq \le |x| \le \frac{2k+1}{2}q, \\ \frac{k+1}{2}q & \text{if } \frac{2k+1}{2}q \le |x| \le (k+1)q, \end{cases} \quad k = 0, 1, \ldots, \tag{4.34}$$

Its proximal mapping is defined as:

- *When the regularization strength* $\lambda \le q$*, we have*

$$prox_{\lambda\psi}(u) = \begin{cases} sign(u)kq & \text{if } kq \le |u| \le kq + \lambda, \\ sign(u)(|u| - \lambda) & \text{if } kq + \lambda \le |u| \le \frac{2k+1}{2}q + \frac{\lambda}{2}, \\ sign(u)|u| & \text{if } \frac{2k+1}{2}q + \frac{\lambda}{2} \le |u| \le (k+1)q. \end{cases} \tag{4.35}$$

- *When the regularization strength* $\lambda \ge q$*, we have*

$$prox_{\lambda\psi}(u) = sign(u) \left\lfloor \frac{|u| - \frac{\lambda}{2}}{q} \right\rceil q. \tag{4.36}$$

where $\lfloor \cdot \rceil$ *denotes the nearest integer. When* λ *exceeds a certain threshold (e.g.,* $\lambda \ge q$*), the proximal operator becomes a* ***hard quantizer****, mapping inputs exactly to discrete levels in a quantization set* $Q = \{0, \pm q, \pm 2q, \pm 3q, \ldots\}$.

Algorithms

We can easily extend SCMA and SCST to solving the non-smooth regularized SCO problems using the following update:

$$\mathbf{w}_{t+1} = \arg\min \frac{1}{2\eta_t} \|\mathbf{w} - (\mathbf{w}_t - \eta_t \mathbf{v}_t)\|_2^2 + r(\mathbf{w}), \tag{4.37}$$

where $\mathbf{v}_t$ is the MA or STORM gradient estimator as in SCMA or SCST.

Convergence Analysis

We first present a lemma similar to Lemma 4.9.

Lemma 4.16 *Consider the update in (4.37), if* $\eta_t \le \frac{1}{4L_F}$ *then we have*

$$\bar{F}(\mathbf{w}_t) - \bar{F}(\mathbf{w}_{t+1}) \leq \eta_t \|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2 - \frac{\eta_t}{10} dist(0, \partial \bar{F}(\mathbf{w}_{t+1}))^2 - \frac{1}{80\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2.$$

Proof. Recall the update of $\mathbf{w}_{t+1}$:

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

Then following variational analysis, we have

$$-\mathbf{v}_t - \frac{1}{\eta_t}(\mathbf{w}_{t+1} - \mathbf{w}_t) \in \partial r(\mathbf{w}_{t+1}),$$

which implies that

$$\nabla F(\mathbf{w}_{t+1}) - \mathbf{v}_t - \frac{1}{\eta_t}(\mathbf{w}_{t+1} - \mathbf{w}_t) \in \nabla F(\mathbf{w}_{t+1}) + \partial r(\mathbf{w}_{t+1}) = \partial \bar{F}(\mathbf{w}_{t+1}). \tag{4.38}$$

Hence, we have

$$\text{dist}(0, \partial \bar{F}(\mathbf{w}_{t+1}))^2 \leq \|\nabla F(\mathbf{w}_{t+1}) - \mathbf{v}_t - \frac{1}{\eta_t}(\mathbf{w}_{t+1} - \mathbf{w}_t)\|_2^2 \tag{4.39}$$

Due to the update of $\mathbf{w}_{t+1}$, we also have

$$r(\mathbf{w}_{t+1}) + \langle \mathbf{v}_t, \mathbf{w}_{t+1} - \mathbf{w}_t \rangle + \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \leq r(\mathbf{w}_t). \tag{4.40}$$

Since $F(\mathbf{w})$ is smooth with parameter L_F, then

$$F(\mathbf{w}_{t+1}) \leq F(\mathbf{w}_t) + \langle \nabla F(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t \rangle + \frac{L_F}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2. \tag{4.41}$$

Combining these two inequalities (4.40) and (4.41) we get

$$\bar{F}(\mathbf{w}_{t+1}) + \langle \mathbf{v}_t - \nabla F(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t \rangle \leq \bar{F}(\mathbf{w}_t) - (\frac{1}{2\eta_t} - \frac{L_F}{2})\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2.$$

From the above inequality, we obtain two results. The first result is

$$
\begin{aligned}
&\frac{2}{\eta_t}\langle \mathbf{v}_t - \nabla F(\mathbf{w}_{t+1}), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle \\
&\leq \frac{2(\bar{F}(\mathbf{w}_t) - \bar{F}(\mathbf{w}_{t+1}))}{\eta_t} - \frac{1}{\eta_t}(\frac{1}{\eta_t} - L_F)\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&+ \frac{2}{\eta_t}\langle \nabla F(\mathbf{w}_t) - \nabla F(\mathbf{w}_{t+1}), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle \\
&\leq \frac{2(\bar{F}(\mathbf{w}_t) - \bar{F}(\mathbf{w}_{t+1}))}{\eta_t} - \frac{1}{\eta_t}(\frac{1}{\eta_t} - 3L_F)\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2. \qquad (4.42)
\end{aligned}
$$

The second result is

$$
\begin{aligned}
&(\frac{1}{2\eta_t} - \frac{L_F}{2})\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \leq \bar{F}(\mathbf{w}_t) - \bar{F}(\mathbf{w}_{t+1}) + \langle \nabla F(\mathbf{w}_t) - \mathbf{v}_t, \mathbf{w}_{t+1} - \mathbf{w}_t\rangle \\
&= \bar{F}(\mathbf{w}_t) - \bar{F}(\mathbf{w}_{t+1}) + \eta_t\|\nabla F(\mathbf{w}_t) - \mathbf{v}_t\|_2^2 + \frac{1}{4\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2.
\end{aligned}
$$

If $\frac{L_F}{2} \leq \frac{1}{8\eta_t}$, i.e., $\eta_t \leq \frac{1}{4L_F}$, the above inequality indicates:

$$
\frac{1}{8\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \leq \bar{F}(\mathbf{w}_t) - \bar{F}(\mathbf{w}_{t+1}) + \eta_t\|\nabla F(\mathbf{w}_t) - \mathbf{v}_t\|_2^2. \qquad (4.43)
$$

To proceed, we have

$$
\begin{aligned}
&\|\mathbf{v}_t - \nabla F(\mathbf{w}_{t+1}) + \frac{1}{\eta_t}(\mathbf{w}_{t+1} - \mathbf{w}_t)\|_2^2 \\
&= 2\langle \mathbf{v}_t - \nabla F(\mathbf{w}_{t+1}), \frac{1}{\eta_t}(\mathbf{w}_{t+1} - \mathbf{w}_t)\rangle + \|\mathbf{v}_t - \nabla F(\mathbf{w}_{t+1})\|_2^2 + \frac{1}{\eta_t^2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2.
\end{aligned}
$$

Adding the above inequality to (4.42) we have

$$
\begin{aligned}
&\|\mathbf{v}_t - \nabla F(\mathbf{w}_{t+1}) + \frac{1}{\eta_t}(\mathbf{w}_{t+1} - \mathbf{w}_t)\|_2^2 \\
&\leq \frac{2(\bar{F}(\mathbf{w}_t) - \bar{F}(\mathbf{w}_{t+1}))}{\eta_t} - \frac{1}{\eta_t}(\frac{1}{\eta_t} - 3L_F)\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&+ \|\mathbf{v}_t - \nabla F(\mathbf{w}_{t+1})\|_2^2 + \frac{1}{\eta_t^2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&= \frac{2(\bar{F}(\mathbf{w}_t) - \bar{F}(\mathbf{w}_{t+1}))}{\eta_t} + \frac{3L_F}{\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 + \|\mathbf{v}_t - \nabla F(\mathbf{w}_{t+1})\|_2^2.
\end{aligned}
$$

Since

$$
\begin{aligned}
\|\mathbf{v}_t - \nabla F(\mathbf{w}_{t+1})\|_2^2 &= \|\mathbf{v}_t - \nabla F(\mathbf{w}_t) + \nabla F(\mathbf{w}_t) - \nabla F(\mathbf{w}_{t+1})\|_2^2 \\
&\leq 2\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2 + 2\|\nabla F(\mathbf{w}_t) - \nabla F(\mathbf{w}_{t+1})\|_2^2 \\
&\leq 2\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2 + 2L_F^2\|\mathbf{w}_t) - \mathbf{w}_{t+1}\|_2^2.
\end{aligned}
$$

Due to $2L_F^2 \leq \frac{L_F}{2\eta_t}$, we have

$$\begin{aligned}
&\|\mathbf{v}_t - \nabla F(\mathbf{w}_{t+1}) + \frac{1}{\eta_t}(\mathbf{w}_{t+1} - \mathbf{w}_t)\|_2^2 \\
&\leq \frac{2(\bar{F}(\mathbf{w}_t) - \bar{F}(\mathbf{w}_{t+1}))}{\eta_t} + \frac{3.5L_F}{\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 + 2\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2.
\end{aligned}$$

Multiplying both sides by η_t, we have

$$\begin{aligned}
&\eta_t\|\mathbf{v}_t - \nabla F(\mathbf{w}_{t+1}) + \frac{1}{\eta_t}(\mathbf{w}_{t+1} - \mathbf{w}_t)\|_2^2 \\
&\leq 2(\bar{F}(\mathbf{w}_t) - \bar{F}(\mathbf{w}_{t+1})) + 3.5L_F\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 + 2\eta_t\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2.
\end{aligned}$$

Adding this inequality to (4.43) gives

$$\begin{aligned}
&\eta_t\|\mathbf{v}_t - \nabla F(\mathbf{w}_{t+1}) + \frac{1}{\eta_t}(\mathbf{w}_{t+1} - \mathbf{w}_t)\|_2^2 + \frac{1}{8\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&\leq 3(\bar{F}(\mathbf{w}_t) - \bar{F}(\mathbf{w}_{t+1})) + 3\eta_t\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2 + 3.5L_F\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2.
\end{aligned}$$

Applying (4.43) again to the RHS, we have

$$\begin{aligned}
&\eta_t\|\mathbf{v}_t - \nabla F(\mathbf{w}_{t+1}) + \frac{1}{\eta_t}(\mathbf{w}_{t+1} - \mathbf{w}_t)\|_2^2 + \frac{1}{8\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\
&\leq (3 + 28L_F\eta_t)(\bar{F}(\mathbf{w}_t) - \bar{F}(\mathbf{w}_{t+1})) + (3\eta_t + 28\eta_t^2 L_F)\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2 \\
&\leq 10(\bar{F}(\mathbf{w}_t) - \bar{F}(\mathbf{w}_{t+1})) + 10\eta_t\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2.
\end{aligned}$$

Combining this with (4.39), we finish the proof. □

Since the above lemma resembles that in Lemma 4.9, hence, it remains a simple exercise to derive the complexity of using the MA estimator similar to Theorem 4.3 and of using the STORM estimator similar to Theorem 4.4.

Corollary 4.1 *Consider the method (4.37). Under the same assumptions and similar settings as in Theorem 4.3, the method finds an ϵ-stationary solution with a complexity of $O(1/\epsilon^4)$. Under the same assumptions and similar settings as in Theorem 4.4, the method finds an ϵ-stationary solution with a complexity of $O(1/\epsilon^3)$.*

Why it matters

Since standard regularized stochastic optimization $\mathbb{E}_\zeta[g(\mathbf{w};\zeta)] + r(\mathbf{w})$ is a special case, the above results directly apply. This corollary shows that regularized problems can be solved with the same complexities as unregularized ones by employing either the moving-average gradient estimator or the STORM gradient estimator. In contrast, without these estimators, solving non-convex regularized problems requires a large batch size at every iteration (Lan, 2020)[Section 6.2.3].

4.5 Structured Optimization with Compositional Gradient

In this section, we extend the compositional optimization technique to address other structured optimization problems, including min-max optimization, min-min optimization, and bilevel optimization. These problems share a common structure in the form of a compositional gradient, denoted by $\mathcal{M}(\mathbf{w}, \mathbf{u}^*(\mathbf{w}))$, where $\mathcal{M}$ is a mapping that is Lipschitz continuous with respect to its second argument, and $\mathbf{u}^*(\mathbf{w})$ is defined as the solution to a strongly convex optimization problem:

$$\mathbf{u}^*(\mathbf{w}) = \arg\min_{\mathbf{u}\in\mathcal{U}} h(\mathbf{w}, \mathbf{u}). \tag{4.44}$$

This structure generalizes the gradient of a compositional function $f(g(\mathbf{w}))$, whose gradient takes the form $\mathcal{M}(\mathbf{w}, \mathbf{u}^*(\mathbf{w})) = \nabla g(\mathbf{w})\nabla f(\mathbf{u}^*(\mathbf{w}))$ with

$$\mathbf{u}^*(\mathbf{w}) = \arg\min_{\mathbf{u}} \|\mathbf{u} - g(\mathbf{w})\|_2^2.$$

The high-level idea underlying the algorithms and analysis presented below is summarized as follows. To estimate $\mathcal{M}(\mathbf{w}, \mathbf{u}^*(\mathbf{w}))$ at $\mathbf{w}_t$, we use an auxiliary variable $\mathbf{u}_t$ to track the optimal solution $\mathbf{u}^*(\mathbf{w}_t)$, which is defined by solving (4.44) with one step update at $\mathbf{w}_t$. A key aspect of the analysis is that the error in the approximation of $\mathcal{M}(\mathbf{w}_t, \mathbf{u}_t)$ is controlled by the estimation error $\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2$, due to the Lipschitz continuity of $\mathcal{M}$:

$$\|\mathcal{M}(\mathbf{w}_t, \mathbf{u}_t) - \mathcal{M}(\mathbf{w}_t, \mathbf{u}^*(\mathbf{w}_t))\|_2^2 \leq O(\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2). \tag{4.45}$$

Moreover, since $\mathbf{u}^*(\mathbf{w})$ is the solution to a strongly convex problem and is Lipschitz continuous with respect to $\mathbf{w}$, we can construct a recursion for $\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2$ to effectively bound the cumulative error over iterations.

In cases where $\mathcal{M}(\mathbf{w}_t, \mathbf{u}_t)$ cannot be computed exactly and is instead approximated by a stochastic estimator $\mathcal{M}(\mathbf{w}_t, \mathbf{u}_t; \zeta_t)$, where ζ_t is a random variable, we employ a moving average (MA) estimator:

$$\mathbf{v}_t = (1 - \beta_t)\mathbf{v}_{t-1} + \beta_t \, \mathcal{M}(\mathbf{w}_t, \mathbf{u}_t; \zeta_t).$$

The model update is then performed using:

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

Alternatively, if $\mathcal{M}(\mathbf{w}_t, \mathbf{u}_t)$ is directly computable, the update simplifies to:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \, \mathcal{M}(\mathbf{w}_t, \mathbf{u}_t).$$

4.5.1 Non-convex Min-Max Optimization

We consider a non-convex min-max optimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \max_{\mathbf{u}\in\mathcal{U}} f(\mathbf{w},\mathbf{u}) := \mathbb{E}_\xi[f(\mathbf{w},\mathbf{u};\xi)], \tag{4.46}$$

where $f(\mathbf{w},\mathbf{u})$ is a continuous and differentiable and $\mathcal{U}$ is a closed convex set. Let $F(\mathbf{w}) = \max_{\mathbf{u}\in\mathcal{U}} f(\mathbf{w},\mathbf{u})$. Denote by $\nabla_1 f(\cdot,\cdot)$ and $\nabla_2 f(\cdot,\cdot)$ the partial gradients of the first and second variable, respectively.

We make the following assumptions.

Assumption 4.8. *Regarding the problem (4.46), the following conditions hold:*

(i) $f(\mathbf{w},\mathbf{u})$ *is* μ*-strongly concave in terms of* $\mathbf{u}$*, and* $\mathbf{u}^*(\mathbf{w}) = \arg\max_{\mathbf{u}\in\mathcal{U}} f(\mathbf{w},\mathbf{u})$ *exists for any* $\mathbf{w}$.

(ii) $\nabla_1 f(\mathbf{w},\mathbf{u})$ *is* L_1*-Lipschitz continuous such that*

$$\|\nabla_1 f(\mathbf{w},\mathbf{u}) - \nabla_1 f(\mathbf{w}',\mathbf{u}')\|_2 \leq L_1(\|\mathbf{w}-\mathbf{w}'\|_2 + \|\mathbf{u}-\mathbf{u}'\|_2). \tag{4.47}$$

(iii) $\nabla_2 f(\mathbf{w},\mathbf{u})$ *is* L_{21}*-Lipschitz continuous with respect to the first variable and is* L_{22}*-Lipschitz continuous with respect to the second variable*

$$\|\nabla_2 f(\mathbf{w},\mathbf{u}) - \nabla_2 f(\mathbf{w}',\mathbf{u}')\|_2 \leq L_{21}\|\mathbf{w}-\mathbf{w}'\|_2 + L_{22}\|\mathbf{u}-\mathbf{u}'\|_2. \tag{4.48}$$

(iv) there exist σ_1, σ_2 *such that*

$$\mathbb{E}[\|\nabla_1 f(\mathbf{w},\mathbf{u};\xi) - \nabla_1 f(\mathbf{w},\mathbf{u})\|_2^2] \leq \sigma_1^2, \tag{4.49}$$

$$\mathbb{E}[\|\nabla_2 f(\mathbf{w},\mathbf{u};\xi) - \nabla_2 f(\mathbf{w},\mathbf{u})\|_2^2] \leq \sigma_2^2. \tag{4.50}$$

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

4.5.1.1 A Double-loop Large mini-batch method

Let us first consider a straightforward approach that updates $\mathbf{w}_t$ using a large-batch gradient estimator

$$\mathbf{v}_t = \frac{1}{B}\sum_{i=1}^{B} \nabla_1 f(\mathbf{w}_t,\mathbf{u}_t;\zeta_{i,t}),$$

and computes $\mathbf{u}_t$ via an inner-loop SGD with K updates. It suffices to have $K = O(L_1^2\sigma_2^2/(\mu^2\epsilon^2))$ (by Lemma 3.8) such that

$$\mathbb{E}[\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2] \leq \frac{\epsilon^2}{L_1^2}.$$

If $B = O(\sigma_1^2/\epsilon^2)$, following the Lemma 4.18 below we have

Algorithm 12 SMDA

1: **Input:** learning rate schedules $\{\eta_t\}_{t=1}^T$, $\{\gamma_t\}_{t=1}^T$, $\{\beta_t\}_{t=1}^T$; starting points $\mathbf{w}_0, \mathbf{u}_1, \mathbf{v}_0$
2: $\mathbf{w}_1 = \mathbf{w}_0 - \eta_0 \mathbf{v}_0$
3: **for** $t = 1, \ldots, T$ **do**
4: Sample ζ_t
5: Update $\mathbf{u}_{t+1} = \Pi_{\mathcal{U}}[\mathbf{u}_t + \gamma_t \nabla_2 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t)]$
6: Compute the vanilla gradient estimator $\mathbf{z}_t = \nabla_1 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t)$
7: Update the MA gradient estimator $\mathbf{v}_t = (1 - \beta_t)\mathbf{v}_{t-1} + \beta_t \mathbf{z}_t$
8: Update the model by $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{v}_t$
9: **end for**

$$\mathbb{E}[\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2] \leq \mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B} \nabla_1 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_{i,t}) - \nabla_1 f(\mathbf{w}_t, \mathbf{u}^*(\mathbf{w}_t))\right\|_2^2\right]$$
$$\leq O\left(\frac{\sigma_1^2}{B} + L_1^2 \|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2\right) \leq \epsilon^2.$$

Combining this with Lemma 4.9, we can set the step size $\eta_t = O(1/L_F)$ and the number of iterations $T = O(L_F/\epsilon^2)$, yielding an overall sample complexity of

$$BT + KT = O\left(\frac{L_F \sigma_1^2}{\epsilon^4} + \frac{L_F L_1^2 \sigma_2^2}{\mu^2 \epsilon^4}\right).$$

4.5.1.2 A Stochastic Momentum Method

We present a solution method in Algorithm 12, referred to as **SMDA** (Stochastic Momentum Descent-Ascent). The method begins by updating the dual variable using stochastic gradient ascent (Step 4), then computes the moving average gradient estimator $\mathbf{v}_t$ for the primal variable (Step 6), and finally updates the primal variable using this estimator (Step 7). When $\beta_t = 1$, the method reduces to SGDA. However, setting $\beta_t < 1$ is crucial for achieving improved complexity. Conceptually, the method shares similarities with SCMA.

Convergence Analysis

We will prove the convergence of the gradient norm of $F(\mathbf{w})$. We first prove the following lemmas.

Lemma 4.17 *Let* $\mathbf{u}^*(\mathbf{w}) = \arg\max_{\mathbf{u}\in\mathcal{U}} f(\mathbf{w}, \mathbf{u})$. *Under Assumption 4.8(i), (iii),* $\mathbf{u}^*(\cdot)$ *is* κ*-Lipschitz continuous with* $\kappa = \frac{L_{21}}{\mu}$.

Proof. Let us consider $\mathbf{w}_1, \mathbf{w}_2$. By the optimality condition of $\mathbf{u}^*(\mathbf{w}_1)$ and $\mathbf{u}^*(\mathbf{w}_2)$ for a concave function, we have

$$\nabla_2 f(\mathbf{w}_1, \mathbf{u}^*(\mathbf{w}_1))^\top (\mathbf{u} - \mathbf{u}^*(\mathbf{w}_1)) \le 0, \quad \forall \mathbf{u} \in \mathcal{U}$$
$$\nabla_2 f(\mathbf{w}_2, \mathbf{u}^*(\mathbf{w}_2))^\top (\mathbf{u} - \mathbf{u}^*(\mathbf{w}_2)) \le 0, \quad \forall \mathbf{u} \in \mathcal{U}$$

Let $\mathbf{u} = \mathbf{u}^*(\mathbf{w}_2)$ in the first inequality and $\mathbf{u} = \mathbf{u}^*(\mathbf{w}_1)$ in the second equality and add them together we have

$$(\nabla_2 f(\mathbf{w}_1, \mathbf{u}^*(\mathbf{w}_1)) - \nabla_2 f(\mathbf{w}_2, \mathbf{u}^*(\mathbf{w}_2)))^\top (\mathbf{u}^*(\mathbf{w}_2) - \mathbf{u}^*(\mathbf{w}_1)) \le 0$$

Since $-f(\mathbf{w}_1, \cdot)$ is μ-strongly convex, due to Lemma 1.6, we have

$$\begin{aligned}(\nabla_2 f(\mathbf{w}_1, \mathbf{u}^*(\mathbf{w}_1)) - \nabla_2 f(\mathbf{w}_1, \mathbf{u}^*(\mathbf{w}_2)))^\top (\mathbf{u}^*(\mathbf{w}_2) - \mathbf{u}^*(\mathbf{w}_1)) \\ \ge \mu \|\mathbf{u}^*(\mathbf{w}_2) - \mathbf{u}^*(\mathbf{w}_1)\|_2^2.\end{aligned}$$

Combining these two inequalities we have

$$\begin{aligned}\mu \|\mathbf{u}^*(\mathbf{w}_2) - \mathbf{u}^*(\mathbf{w}_1)\|_2^2 &\le (\nabla_2 f(\mathbf{w}_2, \mathbf{u}^*(\mathbf{w}_2)) - \nabla_2 f(\mathbf{w}_1, \mathbf{u}^*(\mathbf{w}_2)))^\top (\mathbf{u}^*(\mathbf{w}_2) - \mathbf{u}^*(\mathbf{w}_1)) \\ &\le \|\nabla_2 f(\mathbf{w}_2, \mathbf{u}^*(\mathbf{w}_2)) - \nabla_2 f(\mathbf{w}_1, \mathbf{u}^*(\mathbf{w}_2))\|_2 \|\mathbf{u}^*(\mathbf{w}_2) - \mathbf{u}^*(\mathbf{w}_1)\|_2 \\ &\le L_{21} \|\mathbf{w}_2 - \mathbf{w}_1\|_2 \|\mathbf{u}^*(\mathbf{w}_2) - \mathbf{u}^*(\mathbf{w}_1)\|_2.\end{aligned}$$

Thus,

$$\|\mathbf{u}^*(\mathbf{w}_2) - \mathbf{u}^*(\mathbf{w}_1)\|_2 \le \frac{L_{21}}{\mu} \|\mathbf{w}_2 - \mathbf{w}_1\|_2.$$

□

Lemma 4.18 *Under Assumption 4.8(i) and (ii), $\nabla F(\mathbf{w}) = \nabla_1 f(\mathbf{w}, \mathbf{u}^*(\mathbf{w}))$, and it is is L_F-Lipschitz continuous with $L_F = L_1(1 + \kappa)$.*

Proof. If $\mathcal{U}$ is bounded, the Danskin's theorem implies that $\nabla F(\mathbf{w}) = \nabla_1 f(\mathbf{w}, \mathbf{u}^*(\mathbf{w}))$. If $\mathcal{U}$ is unbounded, we have

$$\nabla F(\mathbf{w}) = \nabla_1 f(\mathbf{w}, \mathbf{u}^*(\mathbf{w})) + \frac{\partial u^*(\mathbf{w})}{\partial \mathbf{w}}^\top \nabla_2 f(\mathbf{w}, \mathbf{u}^*(\mathbf{w})) = \nabla_1 f(\mathbf{w}, \mathbf{u}^*(\mathbf{w})) \quad (4.51)$$

where the last equality follows from $\nabla_2 f(\mathbf{w}, \mathbf{u}^*(\mathbf{w})) = 0$. To establish the Lipschitz continuity of $\nabla F(\mathbf{w})$, let us consider $\mathbf{w}_1$ and $\mathbf{w}_2$. We have

$$\begin{aligned}&\|\nabla F(\mathbf{w}_1) - \nabla F(\mathbf{w}_2)\|_2 = \|\nabla_1 f(\mathbf{w}_1, \mathbf{u}^*(\mathbf{w}_1)) - \nabla_1 f(\mathbf{w}_2, \mathbf{u}^*(\mathbf{w}_2))\|_2 \\ &\le L_1(\|\mathbf{w}_1 - \mathbf{w}_2\|_2 + \|\mathbf{u}^*(\mathbf{w}_1) - \mathbf{u}^*(\mathbf{w}_2)\|_2) \le L_1(1 + \kappa)\|\mathbf{w}_1 - \mathbf{w}_2\|_2.\end{aligned}$$

□

Next, we prove two lemmas similar to Lemma 4.8 and Lemma 4.1, regarding the recursion of gradient estimation error and the estimation error of $\mathbf{u}$, respectively. The descent lemma (Lemma 4.9) still holds.

Lemma 4.19 *It holds that*

$$\mathbb{E}_{\xi_t}\left[\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2\right] \le (1-\beta_t)\|\mathbf{v}_{t-1} - \nabla F(\mathbf{w}_{t-1})\|_2^2 + \frac{2L_F^2}{\beta_t}\|\mathbf{w}_{t-1} - \mathbf{w}_t\|_2^2 + 4L_1^2\beta_t\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2 + \beta_t^2\sigma_1^2.$$

Proof. Let $\mathbf{z}_t = \nabla_1 f(\mathbf{w}_t, \mathbf{u}_t; \xi_t)$ and $\mathcal{M}_t = \mathbb{E}_t[\mathbf{z}_t] = \nabla_1 f(\mathbf{w}_t, \mathbf{u}_t)$. Then $\mathbf{v}_t = (1-\beta_t)\mathbf{v}_{t-1} + \beta_t \mathbf{z}_t$. Noting that $\mathbb{E}_t[\|\mathbf{z}_t - \mathcal{M}_t\|_2^2] \le \sigma_1^2$ and $\|\mathcal{M}_t - \nabla F(\mathbf{w}_t)\|_2^2 \le L_1^2\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w})\|_2^2$. Plugging these into Lemma 4.7 finishes the proof. □

Lemma 4.20 *Suppose Assumption 4.8 (i), (iii), (iv) hold. Consider the update* $\mathbf{u}_t = \Pi_{\mathcal{U}}[\mathbf{u}_t + \gamma_t \nabla_2 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t)]$. *If* $\gamma_t < 1/L_{22} < 1/\mu$, *we have*

$$\mathbb{E}_t[\|\mathbf{u}_{t+1} - \mathbf{u}^*(\mathbf{w}_{t+1})\|_2^2] \le (1 - \frac{\gamma_t\mu}{2})\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2 + \frac{3\kappa^2}{\gamma_t\mu}\mathbb{E}_t[\|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2^2] + 2\gamma_t^2\sigma_2^2.$$

Proof. By Lemma 3.7, if $\gamma < 1/L_{22}$ we have

$$\mathbb{E}_t[\|\mathbf{u}_{t+1} - \mathbf{u}^*(\mathbf{w}_t)\|_2^2] \le (1-\gamma_t\mu)\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2 + \gamma_t^2\sigma_2^2. \tag{4.52}$$

Then,

$$\begin{aligned}
&\mathbb{E}_t[\|\mathbf{u}_{t+1} - \mathbf{u}^*(\mathbf{w}_{t+1})\|_2^2] \le (1 + \frac{\gamma_t\mu}{2})\mathbb{E}_t[\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2] \\
&+ (1 + \frac{2}{\gamma_t\mu})\mathbb{E}_t[\|\mathbf{u}^*(\mathbf{w}_t) - \mathbf{u}^*(\mathbf{w}_{t+1})\|_2^2] \\
&\le (1 + \frac{\gamma_t\mu}{2})(1-\gamma_t\mu)\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2 + (1 + \frac{\gamma_t\mu}{2})\gamma_t^2\sigma_2^2 \\
&+ \frac{2+\gamma_t\mu}{\gamma_t\mu}\kappa^2\mathbb{E}_t[\|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2^2] \\
&\le (1 - \frac{\gamma_t\mu}{2})\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2 + 2\gamma_t^2\sigma_2^2 + \frac{3\kappa^2}{\gamma_t\mu}\mathbb{E}_t[\|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2^2],
\end{aligned}$$

where the first inequality uses the Young's inequality, and the last inequality uses $\gamma\mu < 1$. □

Finally, we can prove the following theorem regarding the convergence of SMDA.

Theorem 4.5 *Suppose Assumption 4.8 holds. By setting* $\beta_t = \beta = \epsilon^2/(3\sigma_1^2)$, $\gamma_t = \gamma = \mu\epsilon^2/(96L_1^2\sigma_2^2)$ *and* $\eta_t = \eta = \min(\frac{\beta}{\sqrt{8}L_F}, \frac{\gamma\mu}{16\sqrt{3}L_1\kappa}, \frac{1}{2L_F})$ *in SMDA, then the following holds*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left\{\frac{1}{4}\|\mathbf{v}_t\|_2^2 + \|\nabla F(\mathbf{w}_t)\|_2^2\right\}\right] \le \epsilon^2, \tag{4.53}$$

with an iteration complexity of

$$T = O\left(\max\left\{\frac{C_\Upsilon L_F}{\epsilon^2}, \frac{C_\Upsilon \sigma_1^2 L_F}{\epsilon^4}, \frac{C_\Upsilon L_1^3 \kappa \sigma_2^2}{\epsilon^4 \mu^2}\right\}\right), \tag{4.54}$$

where $C_\Upsilon = 2(F(\mathbf{w}_0) - F_*) + \frac{1}{\sqrt{8}L_F}\|\mathbf{v}_0 - \nabla F(\mathbf{w}_0)\|_2^2 + \frac{L_1}{\sqrt{3}\kappa}\|\mathbf{u}_0 - \mathbf{u}^*(\mathbf{w}_0)\|_2^2$.

Why it matters

The MA gradient estimator in SMDA is critical to obtaining a complexity of $O(1/\epsilon^4)$. If we simply update the primal variable by SGD, the algorithm becomes SGDA. The convergence analysis of SGDA for non-convex minimax problems will suffer from a large batch size issue or slow convergence. In particular, SGDA with a batch size of $O(1/\epsilon^2)$ can find an ϵ-stationary solution in $O(1/\epsilon^2)$ iterations when the problem is smooth in terms of primal and dual variables and strongly-concave in terms of dual variable, yielding a sample complexity of $O(1/\epsilon^4)$. If using a constant batch size $O(1)$, SGDA may need $O(1/\epsilon^8)$ iterations for finding an ϵ-stationary solution (Lin et al., 2020).

Proof. The proof is similar to Theorem 4.3. Let us see the three inequalities in Lemma 4.9, Lemma 4.19, and 4.20 that we have proved so far:

$$(*) F(\mathbf{w}_{t+1}) \leq F(\mathbf{w}_t) + \frac{\eta}{2}\|\nabla F(\mathbf{w}_t) - \mathbf{v}_t\|_2^2 - \frac{\eta}{2}\|\nabla F(\mathbf{w}_t)\|_2^2 - \frac{\eta}{4}\|\mathbf{v}_t\|_2^2,$$

$$(\sharp)\mathbb{E}\left[\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2\right] \leq \mathbb{E}\left[(1-\beta)\|\mathbf{v}_{t-1} - \nabla F(\mathbf{w}_{t-1})\|_2^2 + \frac{2L_F^2\eta^2}{\beta}\|\mathbf{v}_{t-1}\|_2^2\right]$$
$$+ 4L_1^2\beta\mathbb{E}\left[\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2 + \beta^2\sigma_1^2\right]$$

$$(\diamond)\mathbb{E}\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2 \leq \mathbb{E}\left[(1 - \frac{\gamma\mu}{2})\|\mathbf{u}_{t-1} - \mathbf{u}^*(\mathbf{w}_{t-1})\|_2^2 + 2\gamma^2\sigma_2^2 + \frac{3\kappa^2\eta^2}{\gamma\mu}\|\mathbf{v}_{t-1}\|_2^2\right].$$

Let $\bar{\gamma} = \gamma\mu/2$, the last inequality becomes:

$$(\diamond)\mathbb{E}\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2 \leq \mathbb{E}\left[(1-\bar{\gamma})\|\mathbf{u}_{t-1} - \mathbf{u}^*(\mathbf{w}_{t-1})\|_2^2 + 8\bar{\gamma}^2\frac{\sigma_2^2}{\mu^2} + \frac{3\kappa^2\eta^2}{2\bar{\gamma}}\|\mathbf{v}_{t-1}\|_2^2\right].$$

Let us define $A_t = 2(F(\mathbf{w}_t) - F_*)$ and $B_t = \|\nabla F(\mathbf{w}_t)\|_2^2$, $\Gamma_t = \|\mathbf{v}_t\|_2^2/2$, $\Delta_t = \|\nabla F(\mathbf{w}_t) - \mathbf{v}_t\|_2^2$, $\delta_t = \|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2$. Then the three inequalities $(*)$, $(\sharp)$, $(\diamond)$ satisfy that in Lemma 4.10 with $C_1 = 4L_1^2, C_2 = 2L_F^2, C_3 = 3\kappa^2/2, \sigma^2 = \sigma_1^2, \sigma'^2 = 8\sigma_2^2/\mu^2$. If $\eta, \beta, \bar{\gamma}$ satisfy

$$\beta = \frac{\epsilon^2}{3\sigma^2} = \frac{\epsilon^2}{3\sigma_1^2}, \quad \bar{\gamma} = \frac{\epsilon^2}{6C_1\sigma'^2} = \frac{\epsilon^2\mu^2}{192L_1^2\sigma_2^2},$$
$$\eta = \min(\frac{1}{2L_F}, \frac{\beta}{\sqrt{4C_2}}, \frac{\bar{\gamma}}{\sqrt{8C_1C_3}}) = \min(\frac{1}{2L_F}, \frac{\beta}{\sqrt{8}L_F}, \frac{\bar{\gamma}}{\sqrt{48}L_1\kappa}),$$

then (4.89) holds, and the iteration complexity becomes

$$T = O\left(\max\left\{\frac{C_\Upsilon L_F}{\epsilon^2}, \frac{C_\Upsilon\sigma^2\sqrt{C_2}}{\epsilon^4}, \frac{C_\Upsilon\sqrt{C_1C_3}C_1\sigma'^2}{\epsilon^4}\right\}\right)$$
$$= O\left(\max\left\{\frac{C_\Upsilon L_F}{\epsilon^2}, \frac{C_\Upsilon\sigma_1^2 L_F}{\epsilon^4}, \frac{C_\Upsilon L_1^3\kappa\sigma_2^2}{\epsilon^4\mu^2}\right\}\right).$$

□

Critical: It is worth mentioning that an improved complexity of $O(1/\epsilon^3)$ can be achieved by employing the STORM gradient estimator for both the primal and dual variables under the mean-square smooth condition of the objective.

4.5.2 Non-convex Min-Min Optimization

We can extend SMDA to solving a non-convex strongly-convex min-min problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d}\min_{\mathbf{u}\in\mathcal{U}} f(\mathbf{w},\mathbf{u}) := \mathbb{E}_\xi[f(\mathbf{w},\mathbf{u};\xi)], \tag{4.55}$$

where $f(\mathbf{w},\mathbf{u})$ is smooth, non-convex in terms of $\mathbf{w}$ and strongly convex in terms of $\mathbf{u}$ and $\mathcal{U}$ is a closed convex set. If the $\mathbf{u}^*(\mathbf{w}) = \arg\min_{\mathbf{u}\in\mathcal{U}} f(\mathbf{w},\mathbf{u})$ exists and unique, then we have $\nabla F(\mathbf{w}) = \nabla_1 f(\mathbf{w},\mathbf{u}^*(\mathbf{w}))$. Hence, its gradient also exhibits a compositional structure, where the inner function $\mathbf{u}^*(\mathbf{w})$ is a solution to a strongly convex problem.

SMDA can be modified by replacing the $\mathbf{u}$ update with

$$\mathbf{u}_{t+1} = \Pi_{\mathcal{U}}[\mathbf{u}_t - \gamma_t\nabla_2 f(\mathbf{w}_t,\mathbf{u}_t;\zeta_t)].$$

Then, the same convergence result in the last subsection can be established for min-min problem, which is omitted here.

4.5.2.1 Application to weakly convex minimization

Next, we present an application to solving weakly convex minimization problems:

Algorithm 13 A novel method for weakly convex minimization

Input: learning rate schedules $\{\eta_t\}_{t=1}^T$, $\{\gamma_t\}_{t=1}^T$; starting points $\mathbf{w}_1, \mathbf{u}_1, \mathbf{v}_1$
for $t = 1, \ldots, T$ **do**
    Sample ζ_t and compute $\mathcal{G}(\mathbf{u}_t; \zeta_t) = \partial g(\mathbf{u}_t; \zeta_t)$
    Update $\mathbf{u}_{t+1} = \mathbf{u}_t - \gamma_t(\mathcal{G}(\mathbf{u}_t; \zeta_t) + \rho(\mathbf{u}_t - \mathbf{w}_t))$
    Update $\mathbf{w}_{t+1} = (1 - 2\eta_t\rho)\mathbf{w}_t + 2\eta_t\rho\mathbf{u}_t$
end for

$$\min_{\mathbf{w}} F(\mathbf{w}) := \mathbb{E}[g(\mathbf{w}; \zeta)], \tag{4.56}$$

where $F > -\infty$ is ρ-weakly convex, as discussed in Chapter 3.

As argued in Section 3.1.4, an ϵ-stationary solution of the Moreau envelope of $F(\mathbf{w})$ corresponds to a nearly ϵ-stationary solution of the original problem. Hence, we consider optimizing the Moreau envelope directly:

$$\min_{\mathbf{w}} F_\rho(\mathbf{w}) := \min_{\mathbf{u}} \mathbb{E}[g(\mathbf{u}; \zeta)] + \rho\|\mathbf{u} - \mathbf{w}\|_2^2. \tag{4.57}$$

Define $f(\mathbf{w}, \mathbf{u}) = \mathbb{E}[g(\mathbf{u}; \zeta)] + \rho\|\mathbf{u} - \mathbf{w}\|_2^2$. Then $f(\mathbf{w}, \mathbf{u})$ is ρ-strongly convex with respect to $\mathbf{u}$ due to the ρ-weak convexity of F.

For updating $\mathbf{u}$, we use the standard SGD:

$$\mathbf{u}_{t+1} = \mathbf{u}_t - \gamma_t(\mathcal{G}(\mathbf{u}_t; \zeta_t) + 2\rho(\mathbf{u}_t - \mathbf{w}_t)). \tag{4.58}$$

where $\mathcal{G}(\mathbf{u}_t; \zeta_t) \in \partial g(\mathbf{u}_t; \zeta_t)$. For updating $\mathbf{w}$, then we just apply GD with its gradient given by $\nabla_1 f(\mathbf{w}_t, \mathbf{u}_t) = 2\rho(\mathbf{w}_t - \mathbf{u}_t)$:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t 2\rho(\mathbf{w}_t - \mathbf{u}_t) = (1 - 2\eta_t\rho)\mathbf{w}_t + 2\eta_t\rho\mathbf{u}_t. \tag{4.59}$$

We present the updates in Algorithm 13. An interesting observation about this algorithm is that the $\mathbf{u}$ update is similar to the Momentum update (4.18) except that the momentum term $\mathbf{u}_t - \mathbf{u}_{t-1}$ is replaced by $\mathbf{u}_t - \mathbf{w}_t$, where $\mathbf{w}_t$ is a MA weight vector.

Convergence Analysis

Let us first prove the following lemma.

Lemma 4.21 *We have (i) F_ρ is L_F-smooth with $L_F = \frac{6}{\rho}$; (ii) $\nabla_1 f(\mathbf{w}, \mathbf{u})$ is Lipschitz continuous with $L_1 = 2\rho$, and (iii) $\mathbf{u}^*(\mathbf{w})$ is 1-Lipschitz continuous.*

Proof. The smoothness of F_ρ has been proved in Proposition 3.1 with $\lambda = \rho/2$. The Lipschitz continuity of $\nabla_1 f(\mathbf{w}, \mathbf{u}) = 2\rho(\mathbf{w} - \mathbf{u})$ is obvious. Next, let us prove the Lipschitz continuity of $\mathbf{u}^*(\mathbf{w})$. The proof is similar to that of Lemma 4.17.

Let us consider $\mathbf{w}_1, \mathbf{w}_2$. By the optimality condition of $\mathbf{u}^*(\mathbf{w}_1)$ and $\mathbf{u}^*(\mathbf{w}_2)$ for a concave function, there exists $\mathbf{v}(\mathbf{w}_1) \in \partial_2 f(\mathbf{w}_1, \mathbf{u}^*(\mathbf{w}_1)), \mathbf{v}(\mathbf{w}_2) \in \partial_2 f(\mathbf{w}_2, \mathbf{u}^*(\mathbf{w}_2))$

$$\mathbf{v}(\mathbf{w}_1)^\top(\mathbf{u} - \mathbf{u}^*(\mathbf{w}_1)) \le 0, \quad \forall \mathbf{u}$$
$$\mathbf{v}(\mathbf{w}_2)^\top(\mathbf{u} - \mathbf{u}^*(\mathbf{w}_2)) \le 0, \quad \forall \mathbf{u}$$

Let $\mathbf{u} = \mathbf{u}^*(\mathbf{w}_2)$ in the first inequality and $\mathbf{u} = \mathbf{u}^*(\mathbf{w}_1)$ in the second equality and add them together we have

$$(\mathbf{v}(\mathbf{w}_1) - \mathbf{v}(\mathbf{w}_2))^\top(\mathbf{u}^*(\mathbf{w}_2) - \mathbf{u}^*(\mathbf{w}_1)) \le 0.$$

Since $-f(\mathbf{w}_1, \cdot)$ is ρ-strongly convex, similar to Lemma 1.6, we have for any $\mathbf{v} \in \partial_2 f(\mathbf{w}_1, \mathbf{u}^*(\mathbf{w}_2))$,

$$(\mathbf{v}(\mathbf{w}_1) - \mathbf{v})^\top(\mathbf{u}^*(\mathbf{w}_2) - \mathbf{u}^*(\mathbf{w}_1)) \ge \rho\|\mathbf{u}^*(\mathbf{w}_2) - \mathbf{u}^*(\mathbf{w}_1)\|_2^2.$$

Combining these two inequalities we have

$$\begin{aligned}\rho\|\mathbf{u}^*(\mathbf{w}_2) - \mathbf{u}^*(\mathbf{w}_1)\|_2^2 &\le (\mathbf{v}(\mathbf{w}_2) - \mathbf{v})^\top(\mathbf{u}^*(\mathbf{w}_2) - \mathbf{u}^*(\mathbf{w}_1)) \\ &\le \|\mathbf{v}(\mathbf{w}_2) - \mathbf{v}\|_2\|\mathbf{u}^*(\mathbf{w}_2) - \mathbf{u}^*(\mathbf{w}_1)\|_2.\end{aligned}$$

Since there exists $\mathbf{v}' \in \partial g(\mathbf{u}^*(\mathbf{w}_2))$ such that $\mathbf{v}(\mathbf{w}_2) = \mathbf{v}' + \rho(\mathbf{u}^*(\mathbf{w}_2) - \mathbf{w}_2)$, we let $\mathbf{v} = \mathbf{v}' + \rho(\mathbf{u}^*(\mathbf{w}_2) - \mathbf{w}_1)$, then

$$\|\mathbf{u}^*(\mathbf{w}_2) - \mathbf{u}^*(\mathbf{w}_1)\|_2 \le \|\mathbf{w}_2 - \mathbf{w}_1\|_2.$$

□

Since $\partial_2 f(\mathbf{w}, \mathbf{u})$ is not Lipschitz continuous with respect to $\mathbf{u}$, lemma 4.20 is not directly applicable. We develop a similar one below.

Lemma 4.22 *Consider the following update:*

$$\mathbf{u}_{t+1} = \mathbf{u}_t - \gamma_t(\mathcal{G}(\mathbf{u}_t; \zeta_t) + 2\rho(\mathbf{u}_t - \mathbf{w}_t)).$$

If $\mathbb{E}_\zeta[\|\mathcal{G}(\mathbf{u}; \zeta)\|_2^2] \le G^2$ *and* $\gamma_t\rho < 1/8$*, then we have*

$$\begin{aligned}&\mathbb{E}_t\|\mathbf{u}_{t+1} - \mathbf{u}^*(\mathbf{w}_{t+1})\|_2^2 \\ &\le \left(1 - \frac{\gamma_t\rho}{2}\right)\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2 + 8\gamma_t^2G^2 + \frac{12}{\gamma_t\rho}\mathbb{E}_t\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2.\end{aligned}$$

Proof. Since $\mathbf{u}_{t+1}$ is one-step SGD update of $f(\mathbf{w}_t, \mathbf{u})$, the proof is similar to Lemma 3.7 for the non-smooth case.

$$\begin{aligned}&\|\mathbf{u}_{t+1} - \mathbf{u}^*(\mathbf{w}_t)\|_2^2 = \|\mathbf{u}_t - \gamma_t\left(\mathcal{G}(\mathbf{u}_t; \zeta_t) + 2\rho(\mathbf{u}_t - \mathbf{w}_t)\right) - \mathbf{u}^*(\mathbf{w}_t)\|_2^2 \qquad (4.60)\\ &= \|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2 + \gamma_t^2\|\mathcal{G}(\mathbf{u}_t; \zeta_t) + 2\rho(\mathbf{u}_t - \mathbf{w}_t)\|_2^2 \\ &- 2\gamma_t(\mathcal{G}(\mathbf{u}_t; \zeta_t) + 2\rho(\mathbf{u}_t - \mathbf{w}_t))^\top(\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)).\end{aligned}$$

Note that $0 \in \partial g(\mathbf{u}^*(\mathbf{w}_t)) + 2\rho(\mathbf{u}^*(\mathbf{w}_t) - \mathbf{w}_t)$. Thus, $\mathbf{v}_{t-1} = 2\rho(\mathbf{w}_t - \mathbf{u}^*(\mathbf{w}_t)) \in \partial g(\mathbf{u}^*(\mathbf{w}_t))$,

$$\begin{aligned}\mathbb{E}_t \left\|\mathcal{G}(\mathbf{u}_t;\zeta_t) + 2\rho(\mathbf{u}_t - \mathbf{w}_t)\right\|_2^2 &= \mathbb{E}_t \left\|\mathcal{G}(\mathbf{u}_t;\zeta_t) - \mathbf{v}_{t-1} + 2\rho(\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t))\right\|_2^2 \\ &\le 2(\mathbb{E}_t \left\|\mathcal{G}(\mathbf{u}_t;\zeta_t) + \mathbf{v}_{t-1}\right\|_2 + 8\rho^2 \left\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\right\|_2^2 \\ &\le 8G^2 + 8\rho^2 \left\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\right\|_2^2,\end{aligned}$$

where the last inequality uses $\|\mathbf{v}_{t-1}\|_2 \le G$. For the last term in (4.60), let $\mathbf{v}_{t-1} = \mathbb{E}[\mathcal{G}(\mathbf{u}_t;\zeta_t)] + 2\rho(\mathbf{u}_t - \mathbf{w}_t) \in \partial_2 f(\mathbf{w}_t, \mathbf{u}_t)$, then we have

$$\begin{aligned}&\mathbb{E}_t(\mathcal{G}(\mathbf{u}_t;\zeta_t) + 2\rho(\mathbf{u}_t - \mathbf{w}_t))^\top(\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)) = \mathbf{v}_{t-1}^\top(\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)) \\ &= (\mathbf{v}_{t-1} - \mathbf{v}(\mathbf{w}_t))^\top(\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)) \ge \rho \left\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\right\|_2^2.\end{aligned}$$

where $0 = \mathbf{v}(\mathbf{w}_t) \in \partial_2 f(\mathbf{w}_t, \mathbf{u}^*(\mathbf{w}_t))$ and the last inequality is due to the strong convexity of f in terms of $\mathbf{u}$. Combining the above inequalities we have

$$\begin{aligned}&\left\|\mathbf{u}_{t+1} - \mathbf{u}^*(\mathbf{w}_t)\right\|_2^2 = \left\|\mathbf{u}_t - \gamma_t\left(\partial g(\mathbf{u}_t;\zeta_t) + 2\rho(\mathbf{u}_t - \mathbf{w}_t)\right) - \mathbf{u}^*(\mathbf{w}_t)\right\|_2^2 \\ &\le \left\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\right\|_2^2 + \gamma_t^2(8G^2 + 8\rho^2\left\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\right\|_2^2) - 2\gamma_t\rho\left\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\right\|_2^2 \\ &= (1 - 2\gamma_t\rho + 8\gamma_t^2\rho^2)\left\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\right\|_2^2 + 4\gamma_t^2G^2 \\ &\le (1 - \gamma_t\rho)\left\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\right\|_2^2 + 8\gamma_t^2G^2\end{aligned}$$

where the last inequality uses $\gamma_t \le \frac{1}{8\rho}$. Since $\mathbf{u}^*(\mathbf{w})$ is 1-Lipschitz continuous, we have

$$\begin{aligned}&\mathbb{E}_t \left\|\mathbf{u}_{t+1} - \mathbf{u}^*(\mathbf{w}_{t+1})\right\|_2^2 \\ &\le \left(1 + \frac{\gamma_t\rho}{2}\right)\mathbb{E}_t \left\|\mathbf{u}_{t+1} - \mathbf{u}^*(\mathbf{w}_t)\right\|_2^2 + \left(1 + \frac{2}{\gamma_t\rho}\right)\left\|\mathbf{u}^*(\mathbf{w}_{t+1}) - \mathbf{u}^*(\mathbf{w}_t)\right\|_2^2 \\ &\le \left(1 - \frac{\gamma_t\rho}{2}\right)\left\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\right\|_2^2 + 8\gamma_t^2G^2 + \frac{3}{\gamma_t\rho}\mathbb{E}_t \left\|\mathbf{w}_{t+1} - \mathbf{w}_t\right\|_2^2.\end{aligned}$$

□

Lemma 4.23 *Let $\mathbf{z}_t = 2\rho(\mathbf{w}_t - \mathbf{u}_t)$. For the update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\mathbf{z}_t$, if $\eta_t \le 1/(2L_F)$, we have*

$$F_\rho(\mathbf{w}_{t+1}) \le F_\rho(\mathbf{w}_t) + \frac{\eta_t}{2}\left\|\nabla F_\rho(\mathbf{w}_t) - \mathbf{z}_t\right\|_2^2 - \frac{\eta_t}{2}\left\|\nabla F_\rho(\mathbf{w}_t)\right\|_2^2 - \frac{1}{4\eta_t}\left\|\mathbf{w}_{t+1} - \mathbf{w}_t\right\|_2^2,$$

where L_F is the smoothness parameter of $F_\rho(\cdot)$.

Since $\nabla F_\rho(\mathbf{w}_t) = 2\rho(\mathbf{w}_t - \mathbf{u}^*(\mathbf{w}_t))$, hence $\|F_\rho(\mathbf{w}_t) - \mathbf{z}_t\|_2^2 = 4\rho^2\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2$, whose recursion has been established in Lemma 4.22. We can combine these two lemmas and establish a complexity of $O(1/\epsilon^4)$ for Algorithm 13 in order to find an ϵ-stationary solution to $F_\rho(\cdot)$.

4.5.2.2 Application to weakly-convex strongly-concave min-max problems

The same technique can be applied to solving weakly-convex strongly-concave min-max problems $\min_{\mathbf{w}} \max_{\mathbf{u}\in\mathcal{U}} f(\mathbf{w},\mathbf{u})$ with a single loop algorithm. In subsection 4.5.1, we assume the partial gradient $\nabla_1 f(\mathbf{w},\mathbf{u})$ is Lipschitz continuous. We replace this assumption by an assumption that $f(\mathbf{w},\mathbf{u})$ is ρ-weakly convex in terms of $\mathbf{w}$ for any $\mathbf{u}\in\mathcal{U}$.

In this case, $F(\mathbf{w}) = \max_{\mathbf{u}\in\mathcal{U}} f(\mathbf{w},\mathbf{u})$ is not smooth but weakly convex. Let us consider its Moreau envelope:

$$\min_{\mathbf{w}} F_\rho(\mathbf{w}) = \min_{\mathbf{u}_1} F(\mathbf{u}_1) + \rho\|\mathbf{u}_1 - \mathbf{w}\|_2^2.$$

This problem is equivalent to

$$\min_{\mathbf{w},\mathbf{u}_1} \max_{\mathbf{u}_2\in\mathcal{U}} f(\mathbf{u}_1,\mathbf{u}_2) + \rho\|\mathbf{u}_1 - \mathbf{w}\|_2^2,$$

which is strongly convex in terms of $\mathbf{u}_1$ and strongly concave in terms of $\mathbf{u}_2$.

Compared to (4.57), this problem just adds another layer of inner maximization. However, it can be still mapped to the general framework as discussed at the beginning. The gradient of $F_\rho(\mathbf{w})$ is given by $\mathcal{M}(\mathbf{w},\mathbf{u}_1^*(\mathbf{w})) = \rho(\mathbf{w} - \mathbf{u}_1^*(\mathbf{w}))$. If we track $\mathbf{u}_1^*(\mathbf{w}_t)$ by $\mathbf{u}_{1,t}$ and its update relies on the gradient $\partial_1 f(\mathbf{u}_{1,t}, \mathbf{u}_2^*(\mathbf{u}_{1,t}))$. Hence, we just need another variable $\mathbf{u}_{2,t}$ to track $\mathbf{u}_2^*(\mathbf{u}_{1,t})$.

We can develop a similar algorithm. First, let us update $\mathbf{u}_1, \mathbf{u}_2$. Given $\mathbf{w}_t, \mathbf{u}_{1,t}, \mathbf{u}_{2,t}$, we update $\mathbf{u}_{1,t+1}, \mathbf{u}_{2,t+1}$ with SGD update by

$$\mathbf{u}_{2,t+1} = \Pi_{\mathcal{U}}[\mathbf{u}_{2,t} + \gamma_2 \partial_2 f(\mathbf{u}_{1,t}, \mathbf{u}_{2,t}; \zeta_t)] \tag{4.61}$$

$$\mathbf{u}_{1,t+1} = \mathbf{u}_{1,t} - \gamma_1(\partial_1 f(\mathbf{u}_{1,t}, \mathbf{u}_{2,t}; \zeta_t) + 2\rho(\mathbf{u}_{1,t} - \mathbf{w}_t)). \tag{4.62}$$

Then we update $\mathbf{w}_{t+1}$ with GD update by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta 2\rho(\mathbf{w}_t - \mathbf{u}_{1,t}) = (1 - 2\eta\rho)\mathbf{w}_t + 2\eta\rho\mathbf{u}_{1,t}. \tag{4.63}$$

This algorithm also enjoys a complexity of $O(1/\epsilon^4)$ for finding a nearly ϵ-stationary solution of $F(\mathbf{w})$. We refer the readers to (Hu et al., 2024a) for a convergence analysis of this algorithm.

4.5.2.3 Application to Compositional Optimization

We can apply a similar strategy to a compositional function $F(\mathbf{w}) = f_0(g(\mathbf{w}))$, where f_0 is smooth convex and g is weakly convex. With the conjugate of f_0, we can write

$$\min_{\mathbf{w}} f_0(g(\mathbf{w})) = \min_{\mathbf{w}} \max_{\mathbf{u}_2\in\mathcal{U}} f(\mathbf{w},\mathbf{u}) = \mathbf{u}_2^\top g(\mathbf{w}) - f_0^*(\mathbf{u}_2).$$

Since f_0 is smooth, then f_0^* is strongly convex. Then if g is weakly convex and $\mathcal{U}$ is bounded (i.e., f_0 is Lipschitz), then $f(\mathbf{w}, \mathbf{u})$ is weakly convex and strongly concave. Optimizing the Moreau envelope of $f_0(g(\mathbf{w}))$ yields:

$$\min_{\mathbf{w},\mathbf{u}_1} \max_{\mathbf{u}_2 \in \mathcal{U}} \mathbf{u}_2^\top g(\mathbf{u}_1) - f_0^*(\mathbf{u}_2) + \rho\|\mathbf{u}_1 - \mathbf{w}\|_2^2,$$

which is strongly convex in terms of $\mathbf{u}_1$ and strongly concave in terms of $\mathbf{u}_2$. We give an update below:

$$\begin{aligned}
\mathbf{u}_{2,t+1} &= \Pi_{\mathcal{U}}[\mathbf{u}_{2,t} + \gamma_2 g(\mathbf{u}_{1,t}; \zeta_t)] \\
\mathbf{u}_{1,t+1} &= \mathbf{u}_{1,t} - \gamma_1(\partial_1 g(\mathbf{u}_{1,t}; \zeta_t)\mathbf{u}_{2,t} + 2\rho(\mathbf{u}_{1,t} - \mathbf{w}_t)) \\
\mathbf{w}_{t+1} &= \mathbf{w}_t - \eta 2\rho(\mathbf{w}_t - \mathbf{u}_{1,t}) = (1 - 2\eta\rho)\mathbf{w}_t + 2\eta\rho\mathbf{u}_{1,t}.
\end{aligned}$$

Then similar convergence analysis can be developed with a complexity of $O(1/\epsilon^4)$ for finding a nearly ϵ-stationary solution to F.

4.5.3 Non-convex Bilevel Optimization

In this section, we discuss the application of the compositional gradient estimation technique to non-convex bilevel optimization defined by

$$\begin{aligned}
&\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}, \mathbf{u}^*(\mathbf{w})) \\
&\mathbf{u}^*(\mathbf{w}) = \arg\min_{\mathbf{u}\in\mathbb{R}^{d'}} g(\mathbf{w}, \mathbf{u}),
\end{aligned} \tag{4.64}$$

where g is twice differentiable and μ_g-strongly convex in terms of $\mathbf{u}$. Let $F(\mathbf{w}) = f(\mathbf{w}, \mathbf{u}^*(\mathbf{w}))$. The following lemma states the gradient of the objective $F(\mathbf{w})$.

Lemma 4.24 *We have*

$$\nabla F(\mathbf{w}) = \nabla_1 f(\mathbf{w}, \mathbf{u}^*(\mathbf{w})) - \nabla_{21} g(\mathbf{w}, \mathbf{u}^*(\mathbf{w}))^\top (\nabla_{22} g(\mathbf{w}, \mathbf{u}^*(\mathbf{w})))^{-1} \nabla_2 f(\mathbf{w}, \mathbf{u}^*(\mathbf{w})).$$

Proof. By the optimality condition of $\mathbf{u}^*(\mathbf{w})$, we have

$$\nabla_2 g(\mathbf{w}, \mathbf{u}^*(\mathbf{w})) = 0.$$

By taking derivative on both sides, using the chain rule, and the implicit function theorem, we obtain

$$\nabla_{21} g(\mathbf{w}, \mathbf{u}^*(\mathbf{w})) + \nabla_{22} g(\mathbf{w}, \mathbf{u}^*(\mathbf{w})) \frac{\partial \mathbf{u}^*(\mathbf{w})}{\partial \mathbf{w}} = 0.$$

Hence

$$\frac{\partial \mathbf{u}^*(\mathbf{w})}{\partial \mathbf{w}} = -(\nabla_{22} g(\mathbf{w}, \mathbf{u}^*(\mathbf{w})))^{-1} \nabla_{21} g(\mathbf{w}, \mathbf{u}^*(\mathbf{w})).$$

Thus,

$$\begin{aligned}\nabla F(\mathbf{w}) &= \nabla_1 f(\mathbf{w}, \mathbf{u}^*(\mathbf{w})) + \frac{\partial \mathbf{u}^*(\mathbf{w})}{\partial \mathbf{w}}^\top \nabla_2 f(\mathbf{w}, \mathbf{u}^*(\mathbf{w})) \\ &= \nabla_1 f(\mathbf{w}, \mathbf{u}^*(\mathbf{w})) - \nabla_{21} g(\mathbf{w}, \mathbf{u}^*(\mathbf{w}))^\top (\nabla_{22} g(\mathbf{w}, \mathbf{u}^*(\mathbf{w})))^{-1} \nabla_2 f(\mathbf{w}, \mathbf{u}^*(\mathbf{w})).\end{aligned}$$

□

Let us define

$$\begin{aligned}&\mathcal{M}(\mathbf{w}, \mathbf{u}^*(\mathbf{w})) = \\ &\nabla_1 f(\mathbf{w}, \mathbf{u}^*(\mathbf{w})) - \nabla_{21} g(\mathbf{w}, \mathbf{u}^*(\mathbf{w}))^\top (\nabla_{22} g(\mathbf{w}, \mathbf{u}^*(\mathbf{w})))^{-1} \nabla_2 f(\mathbf{w}, \mathbf{u}^*(\mathbf{w})).\end{aligned}$$

If we can establish the Lipschitz continuity of $\mathcal{M}(\mathbf{w}, \mathbf{u}^*(\mathbf{w}))$ in terms of the second argument and the Lipschitz continuity of $\mathbf{u}^*(\mathbf{w})$, then the similar technique can be leveraged. Let $\mathbf{u}^*(\mathbf{w}_t)$ be tracked by $\mathbf{u}_t$. It can be updated by SGD:

$$\mathbf{u}_{t+1} = \mathbf{u}_t - \gamma_t \nabla_2 g(\mathbf{w}_t, \mathbf{u}_t; \zeta_t). \tag{4.65}$$

With $\mathbf{u}_t$, the gradient at $\mathbf{w}_t$ can be estimated by

$$\mathcal{M}(\mathbf{w}_t, \mathbf{u}_t) = \nabla_1 f(\mathbf{w}_t, \mathbf{u}_t) + \nabla_{21} g(\mathbf{w}_t, \mathbf{u}_t)^\top (\nabla_{22} g(\mathbf{w}_t, \mathbf{u}_t))^{-1} \nabla_2 f(\mathbf{w}_t, \mathbf{u}_t). \tag{4.66}$$

However, another challenge is to handle the Hessian inverse $(\nabla_{22} g(\mathbf{w}_t, \mathbf{u}_t)^{-1}$, which itself is a compositional structure. We will discuss three different ways to tackle this challenge. If we have a stochastic estimator of $\mathcal{M}(\mathbf{w}_t, \mathbf{u}_t)$ denoted by $\mathbf{v}_t$, then we update the model parameter by:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{v}_t. \tag{4.67}$$

4.5.3.1 Approach 1: The MA Estimator

If the lower level problem is low-dimensional such that the inverse of the Hessian matrix can be efficiently computed, we can estimate $\nabla_{22} g(\mathbf{w}_t, \mathbf{u}_t)$ by a MA estimator:

$$H_{22,t} = S_{\mu_g}[(1-\beta) H_{22,t-1} + \beta \nabla_{22} g(\mathbf{w}_t, \mathbf{u}_t; \zeta_{2,t})].$$

where $S_{\mu_g}[\cdot]$ is a projection operator that projects a matrix into a matrix whose minimum eigen-value is lower bounded by μ_g, where μ_g is the lower bound of eigen-values of $\nabla_{22} g(\mathbf{w}, \mathbf{u})$. The projection ensures that $[H_{22,t}]^{-1}$ is Lipschitz continuous with respect to $H_{22,t}$.

The a vanilla stochastic gradient estimator of $\mathbf{w}_t$ and its MA estimator are computed by

$$\begin{aligned}\mathbf{z}_t &= \nabla_1 f(\mathbf{w}_t, \mathbf{u}_t; \xi_t) + \nabla_{21} g(\mathbf{w}_t, \mathbf{u}_t; \zeta'_{2,t})^\top (H_{22,t})^{-1} \nabla_2 f(\mathbf{w}_t, \mathbf{u}_t; \xi_t) \\ \mathbf{v}_t &= (1-\beta)\mathbf{v}_{t-1} + \beta \mathbf{z}_t.\end{aligned} \tag{4.68}$$

Convergence Analysis

The proof is largely similar to that of Theorem 4.3. We provide a sketch of proof below. Recall that

$$\mathcal{M}(\mathbf{w}_t, \mathbf{u}_t) = \nabla_1 f(\mathbf{w}_t, \mathbf{u}_t) + \nabla_{21} g(\mathbf{w}_t, \mathbf{u}_t)^\top (\nabla_{22} g(\mathbf{w}_t, \mathbf{u}_t))^{-1} \nabla_2 f(\mathbf{w}_t, \mathbf{u}_t).$$

Define:

$$\hat{\mathcal{M}}(\mathbf{w}_t, \mathbf{u}_t) = \nabla_1 f(\mathbf{w}_t, \mathbf{u}_t) + \nabla_{21} g(\mathbf{w}_t, \mathbf{u}_t)^\top H_{22,t}^{-1} \nabla_2 f(\mathbf{w}_t, \mathbf{u}_t).$$

First, similar to Lemma 4.9, we have the following:

$$F(\mathbf{w}_{t+1}) \leq F(\mathbf{w}_t) + \frac{\eta_t}{2}\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2 - \frac{\eta_t}{2}\|\nabla F(\mathbf{w}_t)\|_2^2 - \frac{1}{4\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2. \tag{4.69}$$

We establish a recursion of the error $\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2$ similar to Lemma 4.7 by noting that $\mathbb{E}_{\xi_t, \zeta'_{2,t}}[\mathbf{z}_t] = \hat{\mathcal{M}}(\mathbf{w}_t, \mathbf{u}_t)$ and there exists $\sigma > 0$ such that $\mathbb{E}_{\xi_t, \zeta'_{2,t}}[\|\mathbf{z}_t - \hat{\mathcal{M}}(\mathbf{w}_t, \mathbf{u}_t)\|_2^2] \leq \sigma^2$. Thus, Lemma 4.7 implies that

$$\begin{aligned}\mathbb{E}_{\xi_t, \zeta'_{2,t}}\left[\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2\right] &\leq (1-\beta_t)\|\mathbf{v}_{t-1} - \nabla F(\mathbf{w}_{t-1})\|_2^2 \\ &+ \frac{2L_F^2}{\beta_t}\|\mathbf{w}_{t-1} - \mathbf{w}_t\|_2^2 + 4\beta_t \left\|\hat{\mathcal{M}}(\mathbf{w}_t, \mathbf{u}_t) - \nabla F(\mathbf{w}_t)\right\|_2^2 + \beta_t^2 \sigma^2.\end{aligned} \tag{4.70}$$

Then, we bound $\|\hat{\mathcal{M}}(\mathbf{w}_t, \mathbf{u}_t) - \nabla F(\mathbf{w}_t)\|_2^2$ by

$$\begin{aligned}\|\hat{\mathcal{M}}(\mathbf{w}_t, \mathbf{u}_t) - \nabla F(\mathbf{w}_t)\|_2^2 &\leq 2\|\hat{\mathcal{M}}(\mathbf{w}_t, \mathbf{u}_t) - \mathcal{M}(\mathbf{w}_t, \mathbf{u}_t)\|_2^2 \\ &+ 2\|\mathcal{M}(\mathbf{w}_t, \mathbf{u}_t) - \nabla F(\mathbf{w}_t)\|_2^2 \\ &\leq O(\|H_{22,t} - \nabla_{22} g(\mathbf{w}_t, \mathbf{u}_t)\|_2^2) + O(\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2).\end{aligned}$$

As a result, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{v}_t - \nabla F(\mathbf{w}_t)\|_2^2] &\leq (1-\beta_t)\|\mathbf{v}_{t-1} - \nabla F(\mathbf{w}_{t-1})\|^2 + \frac{2L_F^2}{\beta_t}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2 \\ &+ \beta_t(O(\|H_{22,t} - \nabla_{22} g(\mathbf{w}_t, \mathbf{u}_t)\|_2^2) + O(\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2)) + \beta_t^2 O(\sigma^2).\end{aligned}$$

This result is similar to that in Lemma 4.8.

We can further build the error recursion of $\|H_{22,t} - \nabla_{22} g(\mathbf{w}_t, \mathbf{u}_t)\|_2^2$ similar to Lemma 4.1, and the error recursion of $\|\mathbf{u}_t - \mathbf{u}^*(\mathbf{w}_t)\|_2^2$ similar to Lemma 4.20.

Combining these results, we can establish a complexity of $O(1/\epsilon^4)$ for finding an ϵ-stationary solution of $F(\cdot)$ in expectation.

4.5.3.2 Approach 2: The Neumann Series (Matrix Taylor Approximation)

If the lower level problem is high-dimensional such that it is prohibited to compute the Hessian, one approach is to leverage the Neuman series:

$$A^{-1} = \sum_{i=0}^{\infty} (I - A)^i, \quad \text{if} \quad \|A\| \leq 1. \tag{4.71}$$

Hence, if $\|\nabla_{22} g(\mathbf{w}_t, \mathbf{u}_t)\| \leq L_{22}$, we estimate the inverse of $\frac{1}{L_{22}} \nabla_{22} g(\mathbf{w}_t, \mathbf{u}_t)$, yielding

$$(\nabla_{22} g(\mathbf{w}_t, \mathbf{u}_t))^{-1} \approx \frac{1}{L_{22}} \sum_{i=0}^{K-1} \left(I - \frac{1}{L_{22}} \nabla_{22} g(\mathbf{w}_t, \mathbf{u}_t) \right)^i. \tag{4.72}$$

This can be further estimated by a stochastic route, by sampling k from $\{0, \ldots, K-1\}$ randomly, then estimate the Hessian inverse by

$$Q_{22,t} = \begin{cases} \frac{K}{L_{22}} \prod_{i=1}^{k} \left(I - \frac{1}{L_{22}} \nabla_{22} g(\mathbf{w}_t, \mathbf{u}_t; \zeta_i) \right) & \text{if } k \geq 1 \\ \frac{K}{L_{22}} I & \text{if } k = 0 \end{cases}. \tag{4.73}$$

This is can be justified by

$$\begin{aligned} \mathbb{E}[Q_{22,t}] &= \frac{1}{K} \frac{K}{L_{22}} I + \frac{K-1}{K} \mathbb{E}_{k \sim \{1,\ldots,K-1\}} \left[\frac{K}{L_{22}} \prod_{i=1}^{k} \left(I - \frac{1}{L_{22}} \mathbb{E}[\nabla_{22} g(\mathbf{w}_t, \mathbf{u}_t; \zeta_i)] \right) \right] \\ &= \mathbb{E}_k \frac{K}{L_{22}} \left(I - \frac{1}{L_2} \nabla_{22} g(\mathbf{w}_t, \mathbf{u}_t) \right)^k = \sum_{k=0}^{K-1} \frac{1}{L_{22}} \left(I - \frac{1}{L_{22}} \nabla_{22} g(\mathbf{w}_t, \mathbf{u}_t) \right)^k. \end{aligned}$$

Then the vanilla gradient estimator of $\mathbf{w}_t$ and its MA estimator are computed by

$$\begin{aligned} \mathbf{z}_t &= \nabla_1 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_{1,t}) + \nabla_{21} g(\mathbf{w}_t, \mathbf{u}_t; \zeta'_{2,t})^\top Q_{22,t} \nabla_2 f(\mathbf{w}_t, \mathbf{u}_t; \zeta_{1,t}) \\ \mathbf{v}_t &= (1 - \beta) \mathbf{v}_{t-1} + \beta \mathbf{z}_t. \end{aligned} \tag{4.74}$$

Convergence Analysis

We provide a proof sketch below. We can understand that $\mathbf{z}_t$ is a unbiased stochastic estimator of

$$\hat{\mathcal{M}}(\mathbf{w}_t, \mathbf{u}_t) = \nabla_1 f(\mathbf{w}_t, \mathbf{u}_t) + \nabla_{21} g(\mathbf{w}_t, \mathbf{u}_t)^\top \mathbb{E}[Q_{22,t}] \nabla_2 f(\mathbf{w}_t, \mathbf{u}_t)$$

We decompose the estimation error of $\mathbf{v}_t$ similarly as in (4.70) and bound $\|\hat{\mathcal{M}}(\mathbf{w}_t, \mathbf{u}_t) - \nabla F(\mathbf{w}_t)\|_2^2$ by

$$\begin{aligned}\|\hat{\mathcal{M}}(\mathbf{w}_t, \mathbf{u}_t) - \nabla F(\mathbf{w}_t)\|_2^2 \leq\ & 2\|\mathcal{M}(\mathbf{w}_t, \mathbf{u}_t) - \nabla F(\mathbf{w}_t)\|_2^2 \\ & + 2\|\hat{\mathcal{M}}(\mathbf{w}_t, \mathbf{u}_t) - \mathcal{M}(\mathbf{w}_t, \mathbf{u}_t)\|_2^2.\end{aligned}$$

The error recursion of the first term on the right hand side can be similarly bounded as before. To bound the last error, since

$$\left[\nabla_{22}^2 g(\mathbf{w}, \mathbf{u})\right]^{-1} = \mathbb{E}[Q_{22}] + \frac{1}{L_{22}} \sum_{i=K}^{\infty} \left[I - \frac{1}{L_{22}} \nabla_{22}^2 g(\mathbf{w}, \mathbf{u})\right]^i,$$

we have

$$\begin{aligned}&\|\mathcal{M}(\mathbf{w}_t, \mathbf{u}_t) - \hat{\mathcal{M}}(\mathbf{w}_t, \mathbf{u}_t)\|_2^2 \leq O(\| \left[\nabla_{22}^2 g(\mathbf{w}, \mathbf{u})\right]^{-1} - \mathbb{E}[Q_{22}]\|_2^2) \\ &\left\| \left[\nabla_{22}^2 g(\mathbf{w}, \mathbf{u})\right]^{-1} - \mathbb{E}[Q_{22}] \right\|_2 \leq \frac{1}{L_{22}} \sum_{i=K}^{\infty} \left\| I - \frac{1}{L_{22}} \nabla_{22}^2 g(\mathbf{w}, \mathbf{u}) \right\|_2^i \leq \frac{1}{\mu_g} \left(1 - \frac{\mu_g}{L_{22}}\right)^K.\end{aligned}$$

As a result, if $K = O(\frac{L_{22}}{\mu_g} \log(1/(\mu_g \beta_t \sigma^2)))$, then $\|\mathcal{M}(\mathbf{w}_t, \mathbf{u}_t) - \mathcal{M}'(\mathbf{w}_t, \mathbf{u}_t)\|_2^2 \leq O(\beta_t \sigma^2)$. Then similar to the analysis of approach 1, we can establish a complexity of $O(1/\epsilon^4)$ for finding an ϵ-stationary solution of $F(\cdot)$ in expectation.

4.5.3.3 Approach 3: The penalty method

An alternative approach to avoid computing the Hessian inverse and Jacobian matrices is to reformulate the problem as a constrained optimization problem:

$$\begin{aligned}\min_{\mathbf{w},\mathbf{u}} \quad & f(\mathbf{w}, \mathbf{u}) \\ \text{s.t.} \quad & g(\mathbf{w}, \mathbf{u}) \leq \min_{\mathbf{u}} g(\mathbf{w}, \mathbf{u}).\end{aligned}$$

This constrained problem can be addressed using a penalty method (see Chapter 6.7):

$$\min_{\mathbf{w},\mathbf{u}} f(\mathbf{w}, \mathbf{u}) + \lambda \big(g(\mathbf{w}, \mathbf{u}) - \min_{\mathbf{y}} g(\mathbf{w}, \mathbf{y})\big)_+,$$

where $\lambda > 0$ is a penalty parameter and $(\cdot)_+$ denotes the positive part. Since $g(\mathbf{w}, \mathbf{u}) \geq \min_{\mathbf{y}} g(\mathbf{w}, \mathbf{y})$, the formulation simplifies to:

$$\min_{\mathbf{w},\mathbf{u}} f(\mathbf{w}, \mathbf{u}) + \lambda \left(g(\mathbf{w}, \mathbf{u}) - \min_{\mathbf{y}} g(\mathbf{w}, \mathbf{y})\right) \tag{4.75}$$

$$= \min_{\mathbf{w},\mathbf{u}} \max_{\mathbf{y}} f(\mathbf{w}, \mathbf{u}) + \lambda \left(g(\mathbf{w}, \mathbf{u}) - g(\mathbf{w}, \mathbf{y})\right). \tag{4.76}$$

If both f and g are smooth and g is strongly convex in its second argument, the resulting formulation becomes a *non-convex strongly-concave min-max problem*, which can be effectively addressed using the SMDA algorithm with the following update for $t \geq 1$:

$$\begin{aligned}
&\mathbf{y}_{t+1} = \mathbf{y}_t + \gamma_t \lambda \nabla_2 g(\mathbf{w}_t, \mathbf{y}_t; \xi_t), \\
&\mathbf{z}_t = \nabla f(\mathbf{w}_t, \mathbf{u}_t; \zeta_t) + \lambda \left(\nabla g(\mathbf{w}_t, \mathbf{u}_t; \xi_t) - \begin{bmatrix} \nabla_1 g(\mathbf{w}_t, \mathbf{y}_t; \xi_t) \\ 0 \end{bmatrix} \right), \\
&\mathbf{v}_t = (1 - \beta_t)\mathbf{v}_{t-1} + \beta_t \mathbf{z}_t, \\
&\begin{bmatrix} \mathbf{w}_{t+1} \\ \mathbf{u}_{t+1} \end{bmatrix} = \begin{bmatrix} \mathbf{w}_t \\ \mathbf{u}_t \end{bmatrix} - \eta_t \mathbf{v}_t.
\end{aligned} \tag{4.77}$$

Convergence Analysis

The convergence analysis of (4.77) for the min–max problem (4.75) follows a similar approach to that of Theorem 4.5 for SMDA. However, a remaining challenge lies in converting the convergence result for the min–max formulation into that of the original problem. To address this, we provide the detailed convergence analysis below. We begin by stating the following assumption.

Assumption 4.9. *Regarding the problem (4.64), the following conditions hold:*

(i) $g(\mathbf{w}, \mathbf{u})$ *is* μ*-strongly concave in terms of* $\mathbf{u}$.
(ii) $\nabla f(\mathbf{w}, \mathbf{u})$ *is* L_f*-Lipschitz continuous such that*

$$\|\nabla f(\mathbf{w}_1, \mathbf{u}_1) - \nabla f(\mathbf{w}_2, \mathbf{u}_2)\|_2 \leq L_f \left\| \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{u}_1 \end{pmatrix} - \begin{pmatrix} \mathbf{w}_2 \\ \mathbf{u}_2 \end{pmatrix} \right\|_2. \tag{4.78}$$

(iii) $\nabla g(\mathbf{w}, \mathbf{u})$ *is* L_g*-Lipschitz continuous such that*

$$\|\nabla g(\mathbf{w}_1, \mathbf{u}_1) - \nabla g(\mathbf{w}_2, \mathbf{u}_2)\|_2 \leq L_g \left\| \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{u}_1 \end{pmatrix} - \begin{pmatrix} \mathbf{w}_2 \\ \mathbf{u}_2 \end{pmatrix} \right\|_2. \tag{4.79}$$

(iv) there exist σ_f, σ_g *such that*

$$\mathbb{E}[\|\nabla f(\mathbf{w}, \mathbf{u}; \zeta) - \nabla f(\mathbf{w}, \mathbf{u})\|_2^2] \leq \sigma_f^2, \tag{4.80}$$

$$\mathbb{E}[\|\nabla g(\mathbf{w}, \mathbf{u}; \xi) - \nabla g(\mathbf{w}, \mathbf{u})\|_2^2] \leq \sigma_g^2. \tag{4.81}$$

(v) $\min_{\mathbf{w},\mathbf{u}} f(\mathbf{w}, \mathbf{u}) \geq -\infty$.

Let us define $\bar{\mathbf{w}} = (\mathbf{w}, \mathbf{u})$ and

$$\bar{f}(\bar{\mathbf{w}}, \mathbf{y}) = f(\mathbf{w}, \mathbf{u}) + \lambda \left(g(\mathbf{w}, \mathbf{u}) - g(\mathbf{w}, \mathbf{y}) \right) \tag{4.82}$$

$$\bar{F}(\bar{\mathbf{w}}) = \max_{\mathbf{y}} \bar{f}(\bar{\mathbf{w}}, \mathbf{y}). \tag{4.83}$$

Then

$$\begin{aligned}
\nabla_1 \bar{f}(\bar{\mathbf{w}}, \mathbf{y}) &= \nabla f(\mathbf{w}, \mathbf{u}) + \lambda \left(\nabla g(\mathbf{w}, \mathbf{u}) - \begin{bmatrix} \nabla_1 g(\mathbf{w}, \mathbf{y}) \\ 0 \end{bmatrix} \right), \\
\nabla_2 \bar{f}(\bar{\mathbf{w}}, \mathbf{y}) &= -\lambda \nabla_2 g(\mathbf{w}, \mathbf{y}), \\
\nabla_1 \bar{f}(\bar{\mathbf{w}}, \mathbf{y}; \varepsilon) &= \nabla f(\mathbf{w}, \mathbf{u}; \zeta) + \lambda \left(\nabla g(\mathbf{w}, \mathbf{u}; \xi) - \begin{bmatrix} \nabla_1 g(\mathbf{w}, \mathbf{y}; \xi) \\ 0 \end{bmatrix} \right), \\
\nabla_2 \bar{f}(\bar{\mathbf{w}}, \mathbf{y}; \xi) &= -\lambda \nabla_2 g(\mathbf{w}, \mathbf{y}; \xi).
\end{aligned}$$

where $\varepsilon = (\zeta, \xi)$. We first show $\bar{f}(\bar{\mathbf{w}}, \mathbf{y})$ satisfies the conditions in Assumption (4.8).

Lemma 4.25 *Under Assumption 4.9, we have*

(i) $\bar{f}(\bar{\mathbf{w}}, \mathbf{y})$ *is* $\mu\lambda$*-strongly concave in terms of* $\mathbf{u}$.
(ii) $\nabla_1 \bar{f}(\bar{\mathbf{w}}, \mathbf{y})$ *is Lipschitz continuous, i.e.,*

$$\|\nabla_1 \bar{f}(\bar{\mathbf{w}}_1, \mathbf{y}_1) - \nabla_1 \bar{f}(\bar{\mathbf{w}}_2, \mathbf{y}_2)\|_2 \le (L_f + 2L_g\lambda)(\|\bar{\mathbf{w}}_1 - \bar{\mathbf{w}}_2\|_2 + \|\mathbf{y}_1 - \mathbf{y}_2\|_2).$$

(iii) $\nabla_2 \bar{f}(\bar{\mathbf{w}}, \mathbf{y})$ *is Lipschitz continuous, i.e.,*

$$\|\nabla_2 \bar{f}(\bar{\mathbf{w}}_1, \mathbf{y}_1) - \nabla_2 f(\bar{\mathbf{w}}_2, \mathbf{y}_2)\|_2 \le L_g\lambda\|\bar{\mathbf{w}}_1 - \bar{\mathbf{w}}_2\|_2 + L_g\lambda\|\mathbf{y}_1 - \mathbf{y}_2\|_2.$$

(iv)

$$\begin{aligned}
\mathbb{E}[\|\nabla_1 \bar{f}(\bar{\mathbf{w}}, \mathbf{y}; \varepsilon) - \nabla_1 \bar{f}(\bar{\mathbf{w}}, \mathbf{y})\|_2^2] &\le 3\sigma_f^2 + 6\lambda^2\sigma_g^2, \\
\mathbb{E}[\|\nabla_2 \bar{f}(\bar{\mathbf{w}}, \mathbf{y}; \xi) - \nabla_2 f(\bar{\mathbf{w}}, \mathbf{y})\|_2^2] &\le \lambda^2\sigma_g^2.
\end{aligned}$$

(v) $\bar{F}(\bar{\mathbf{w}}) := \max_{\mathbf{y}} \bar{f}(\bar{\mathbf{w}}, \mathbf{y}) \ge -\infty$.

Proof. (i) is obvious. The Lipschitz continuity of $\nabla_1 \bar{f}(\bar{\mathbf{w}}, \mathbf{y})$ follows that of $\nabla f(\mathbf{w}, \mathbf{u})$ and $\nabla g(\mathbf{w}, \mathbf{u})$. For (iii), we have

$$\begin{aligned}
&\|\nabla_2 \bar{f}(\bar{\mathbf{w}}_1, \mathbf{y}_1) - \nabla_2 \bar{f}(\bar{\mathbf{w}}_2, \mathbf{y}_2)\|_2 = \lambda\|\nabla_2 g(\mathbf{w}_1, \mathbf{u}_1) - \nabla_2 g(\mathbf{w}_2, \mathbf{u}_2)\|_2 \\
&\le \lambda\|\nabla g(\mathbf{w}_1, \mathbf{u}_1) - \nabla g(\mathbf{w}_2, \mathbf{u}_2)\|_2 \le \lambda L_g \left\| \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{u}_1 \end{pmatrix} - \begin{pmatrix} \mathbf{w}_2 \\ \mathbf{u}_2 \end{pmatrix} \right\|_2 \\
&\le \lambda L_g(\|\mathbf{w}_1 - \mathbf{w}_2\|_2 + \|\mathbf{u}_1 - \mathbf{u}_2\|_2) \le \lambda L_g(\|\bar{\mathbf{w}}_1 - \bar{\mathbf{w}}_2\|_2 + \|\mathbf{u}_1 - \mathbf{u}_2\|_2).
\end{aligned}$$

It is trivial to prove (iv). The last result follows that $\max_{\mathbf{y}} \bar{f}(\bar{\mathbf{w}}, \mathbf{y}) \ge f(\mathbf{w}, \mathbf{u}) \ge \infty$. □

Theorem 4.6 *Suppose Assumption 4.9 hold. By setting*

$$\beta_t = \beta = \frac{\epsilon^2}{9\sigma_f^2 + 18\lambda^2\sigma_g^2},$$

$$\gamma_t = \gamma = \frac{\mu_g \epsilon^2}{96(L_f + 2L_g\lambda)^2\lambda\sigma_g^2},$$

$$\eta_t = \min\left\{\frac{\beta}{\sqrt{8}(L_f + 2L_g\lambda)(1+L_g)}, \frac{\gamma\mu_g\lambda}{16\sqrt{3}(L_f + 2L_g\lambda)L_g}, \frac{1}{2(L_f + 2L_g\lambda)(1+L_g)}\right\}$$

in (4.77), then the following holds

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left\{\frac{1}{4}\|\mathbf{v}_t\|_2^2 + \left\|\nabla\bar{F}(\bar{\mathbf{w}}_t)\right\|_2^2\right\}\right] \leq \epsilon^2, \tag{4.84}$$

with an iteration complexity of

$$T = O\left(\max\left\{\frac{C_\Upsilon\lambda}{\epsilon^2}, \frac{C_\Upsilon(\lambda\sigma_f^2 + \lambda^3\sigma_g^2)}{\epsilon^4}, \frac{C_\Upsilon\lambda^3\sigma_g^2}{\epsilon^4\mu_g^2}\right\}\right), \tag{4.85}$$

where $C_\Upsilon = 2(\bar{F}(\bar{\mathbf{w}}_0) - \min_{\bar{\mathbf{w}}}\bar{F}(\bar{\mathbf{w}})) + \frac{1}{\sqrt{8}L_F}\|\mathbf{v}_0 - \nabla\bar{F}(\bar{\mathbf{w}}_0)\|_2^2 + \frac{L_1}{\sqrt{3}\kappa}\|\mathbf{y}_0 - \mathbf{y}^*(\mathbf{w}_0)\|_2^2$.

Proof. We map the problem into the setting in Theorem 4.5 with $L_1 = L_f + 2L_g\lambda, L_{21} = L_g\lambda, L_2 = L_g\lambda, \mu = \mu_g\lambda, \kappa = L_{21}/(\mu_g\lambda) = L_g, L_F = L_1(1+\kappa) = (L_f + 2L_g\lambda)(1+L_g), \sigma_1^2 = 3\sigma_f^2 + 6\lambda^2\sigma_g^2, \sigma_2^2 = \lambda^2\sigma_g^2$. Then, plugging these values into the result in Theorem 4.5, we obtain the results. □

Convergence of the original function

Next, we derive the convergence of the original function in terms of $\|\nabla F(\mathbf{w})\|_2$. We need the following additional assumption.

Assumption 4.10. *(i) g is twice differentiable and $\nabla_{21}g(\mathbf{w},\mathbf{u})$ and $\nabla g_{22}(\mathbf{w},\mathbf{u})$ are L_{gg}-Lipschitz continuous; and (ii) $\|\nabla_2 f(\mathbf{w},\mathbf{u})\|_2 \leq G_f$.*

Lemma 4.26 *Let $\mathbf{u}_\lambda^*(\mathbf{w}) = \arg\min_{\mathbf{u}}\bar{F}(\mathbf{w},\mathbf{u}), \mathbf{u}^*(\mathbf{w}) = \arg\min_{\mathbf{u}} g(\mathbf{w},\mathbf{u})$. Under Assumption 4.10(i), we have*

$$\|\nabla F(\mathbf{w}) - \nabla_1\bar{F}(\mathbf{w},\mathbf{u}_\lambda^*(\mathbf{w}))\|_2 \leq L_f(1 + \frac{L_g}{\mu_g})\|\mathbf{u}_\lambda^*(\mathbf{w}) - \mathbf{u}^*(\mathbf{w})\|_2 + L_{gg}\lambda(1 + \frac{L_g}{\mu_g})\|\mathbf{u}_\lambda^*(\mathbf{w}) - \mathbf{u}^*(\mathbf{w})\|_2^2.$$

Proof. Let $\mathbf{u}^* = \mathbf{u}^*(\mathbf{w})$. Then,

$$\nabla_1 \bar{F}(\mathbf{w}, \mathbf{u}) = \nabla_1 f(\mathbf{w}, \mathbf{u}) + \lambda(\nabla_1 g(\mathbf{w}, \mathbf{u}) - \nabla_1 g(\mathbf{w}, \mathbf{u}^*))$$
$$\nabla_2 \bar{F}(\mathbf{w}, \mathbf{u}) = \nabla_2 f(\mathbf{w}, \mathbf{u}) + \lambda \nabla_2 g(\mathbf{w}, \mathbf{u}).$$

Due to Lemma 4.24, we have

$$\begin{aligned}\nabla F(\mathbf{w}) - \nabla_1 \bar{F}(\mathbf{w}, \mathbf{u}) &= \nabla_1 f(\mathbf{w}, \mathbf{u}^*) - \nabla_1 f(\mathbf{w}, \mathbf{u}) \\ &- \nabla_{12} g(\mathbf{w}, \mathbf{u}^*) \nabla_{22} g(\mathbf{w}, \mathbf{u}^*)^{-1} \nabla_2 f(\mathbf{w}, \mathbf{u}^*) - \lambda(\nabla_1 g(\mathbf{w}, \mathbf{u}) - \nabla_1 g(\mathbf{w}, \mathbf{u}^*)).\end{aligned} \tag{4.86}$$

We can rearrange terms for $(\nabla_1 g(\mathbf{w}, \mathbf{u}) - \nabla_1 g(\mathbf{w}, \mathbf{u}^*))$ as the following:

$$\begin{aligned}\nabla_1 g(\mathbf{w}, \mathbf{u}) - \nabla_1 g(\mathbf{w}, \mathbf{u}^*) = \nabla_1 g(\mathbf{w}, \mathbf{u}) - \nabla_1 g(\mathbf{w}, \mathbf{u}^*) - \nabla_{12} g(\mathbf{w}, \mathbf{u}^*)^\top (\mathbf{u} - \mathbf{u}^*) \\ + \nabla_{12} g(\mathbf{w}, \mathbf{u}^*)^\top (\mathbf{u} - \mathbf{u}^*).\end{aligned} \tag{4.87}$$

To continue, we have

$$\begin{aligned}\mathbf{u} - \mathbf{u}^* = &- \nabla_{22} g(\mathbf{w}, \mathbf{u}^*)^{-1} (\nabla_2 g(\mathbf{w}, \mathbf{u}) - \nabla_2 g(\mathbf{w}, \mathbf{u}^*) - \nabla_{22} g(\mathbf{w}, \mathbf{u}^*)(\mathbf{u} - \mathbf{u}^*)) \\ &+ \nabla_{22} g(\mathbf{w}, \mathbf{u}^*)^{-1} (\nabla_2 g(\mathbf{w}, \mathbf{u}) - \nabla_2 g(\mathbf{w}, \mathbf{u}^*)).\end{aligned}$$

By the optimality condition for $\mathbf{u}^*$, $\nabla_2 g(\mathbf{w}, \mathbf{u}^*) = 0$, and $\nabla_2 \bar{F}(\mathbf{w}, \mathbf{u}) = \nabla_2 f(\mathbf{w}, \mathbf{u}) + \lambda \nabla_2 g(\mathbf{w}, \mathbf{u})$, we can express $\mathbf{u} - \mathbf{u}^*$ as

$$\begin{aligned}\mathbf{u} - \mathbf{u}^* = &- \nabla_{22} g(\mathbf{w}, \mathbf{u}^*)^{-1} (\nabla_2 g(\mathbf{w}, \mathbf{u}) - \nabla_2 g(\mathbf{w}, \mathbf{u}^*) - \nabla_{22} g(\mathbf{w}, \mathbf{u}^*)(\mathbf{u} - \mathbf{u}^*)) \\ &+ \frac{1}{\lambda} \nabla_{22} g(\mathbf{w}, \mathbf{u}^*)^{-1} (\nabla_2 \bar{F}(\mathbf{w}, \mathbf{u}) - \nabla_2 f(\mathbf{w}, \mathbf{u})).\end{aligned} \tag{4.88}$$

Plugging (4.87) and (4.88) back to (4.86), we have

$$\begin{aligned}&\nabla F(\mathbf{w}) - \nabla_1 \bar{F}(\mathbf{w}, \mathbf{u}) = \nabla_1 f(\mathbf{w}, \mathbf{u}^*) - \nabla_1 f(\mathbf{w}, \mathbf{u}) \\ &- \nabla_{12} g(\mathbf{w}, \mathbf{u}^*) \nabla_{22} g(\mathbf{w}, \mathbf{u}^*)^{-1} \nabla_2 f(\mathbf{w}, \mathbf{u}^*) \\ &- \lambda(\nabla_1 g(\mathbf{w}, \mathbf{u}) - \nabla_1 g(\mathbf{w}, \mathbf{u}^*) - \nabla_{12} g(\mathbf{w}, \mathbf{u}^*)^\top (\mathbf{u} - \mathbf{u}^*)) \\ &+ \lambda \nabla_{12} g(\mathbf{w}, \mathbf{u}^*)^\top \nabla_{22} g(\mathbf{w}, \mathbf{u}^*)^{-1} (\nabla_2 g(\mathbf{w}, \mathbf{u}) - \nabla_2 g(\mathbf{w}, \mathbf{u}^*) - \nabla_{22} g(\mathbf{w}, \mathbf{u}^*)(\mathbf{u} - \mathbf{u}^*)) \\ &- \nabla_{12} g(\mathbf{w}, \mathbf{u}^*)^\top \nabla_{22} g(\mathbf{w}, \mathbf{u}^*)^{-1} (\nabla_2 \bar{F}(\mathbf{w}, \mathbf{u}) - \nabla_2 f(\mathbf{w}, \mathbf{u})).\end{aligned}$$

As a result, we have

$$\begin{aligned}&\nabla F(\mathbf{w}) - \nabla_1 \bar{F}(\mathbf{w}, \mathbf{u}) + \nabla_{12} g(\mathbf{w}, \mathbf{u}^*)^\top \nabla_{22} g(\mathbf{w}, \mathbf{u}^*)^{-1} \nabla_2 \bar{F}(\mathbf{w}, \mathbf{u}) \\ &= \nabla_1 f(\mathbf{w}, \mathbf{u}^*) - \nabla_1 f(\mathbf{w}, \mathbf{u}) \\ &- \nabla_{12} g(\mathbf{w}, \mathbf{u}^*) \nabla_{22} g(\mathbf{w}, \mathbf{u}^*)^{-1} (\nabla_2 f(\mathbf{w}, \mathbf{u}^*) - \nabla_2 f(\mathbf{w}, \mathbf{u})) \\ &- \lambda(\nabla_1 g(\mathbf{w}, \mathbf{u}) - \nabla_1 g(\mathbf{w}, \mathbf{u}^*) - \nabla_{12} g(\mathbf{w}, \mathbf{u}^*)^\top (\mathbf{u} - \mathbf{u}^*)) \\ &+ \lambda \nabla_{12} g(\mathbf{w}, \mathbf{u}^*)^\top \nabla_{22} g(\mathbf{w}, \mathbf{u}^*)^{-1} (\nabla_2 g(\mathbf{w}, \mathbf{u}) - \nabla_2 g(\mathbf{w}, \mathbf{u}^*) - \nabla_{22} g(\mathbf{w}, \mathbf{u}^*)(\mathbf{u} - \mathbf{u}^*)).\end{aligned}$$

By the Assumption 4.10 we have

$$\|\nabla_1 g(\mathbf{w},\mathbf{u}) - \nabla_1 g(\mathbf{w},\mathbf{u}^*) - \nabla_{12} g(\mathbf{w},\mathbf{u}^*)^\top(\mathbf{u}-\mathbf{u}^*)\|_2 \leq L_{gg}\|\mathbf{u}-\mathbf{u}^*\|_2^2,$$
$$\|\nabla_2 g(\mathbf{w},\mathbf{u}) - \nabla_2 g(\mathbf{w},\mathbf{u}^*) - \nabla_{22} g(\mathbf{w},\mathbf{u}^*)(\mathbf{u}-\mathbf{u}^*)\|_2 \leq L_{gg}\|\mathbf{u}-\mathbf{u}^*\|_2^2.$$

By the Assumption 4.9 we have

$$\|\nabla_1 f(\mathbf{w},\mathbf{u}^*) - \nabla_1 f(\mathbf{w},\mathbf{u})\|_2 \leq L_f\|\mathbf{u}^*-\mathbf{u}\|_2,$$
$$\|\nabla_2 f(\mathbf{w},\mathbf{u}^*) - \nabla_2 f(\mathbf{w},\mathbf{u})\|_2 \leq L_f\|\mathbf{u}^*-\mathbf{u}\|_2,$$
$$\|\nabla_{12} g(\mathbf{w},\mathbf{u}^*)\nabla_{22} g(\mathbf{w},\mathbf{u}^*)^{-1}\|_2 \leq \frac{L_g}{\mu_g}.$$

Thus, we have

$$\begin{aligned}&\|\nabla F(\mathbf{w}) - \nabla_1 \bar{F}(\mathbf{w},\mathbf{u}) + \nabla_{12} g(\mathbf{w},\mathbf{u}^*)^\top \nabla_{22} g(\mathbf{w},\mathbf{u}^*)^{-1}\nabla_2 \bar{F}(\mathbf{w},\mathbf{u})\|_2 \\ &\leq L_f(1+\frac{L_g}{\mu_g})\|\mathbf{u}-\mathbf{u}^*\|_2 + L_{gg}\lambda(1+\frac{L_g}{\mu_g})\|\mathbf{u}-\mathbf{u}^*\|_2^2.\end{aligned}$$

Plugging $\mathbf{u} = \mathbf{u}_\lambda^*(\mathbf{w}) = \min_{\mathbf{u}} \bar{F}(\mathbf{w},\mathbf{u})$, then $\nabla_2 \bar{F}(\mathbf{w},\mathbf{u}_\lambda^*(\mathbf{w})) = 0$ and then we have

$$\begin{aligned}&\|\nabla F(\mathbf{w}) - \nabla_1 \bar{F}(\mathbf{w},\mathbf{u}_\lambda^*(\mathbf{w}))\|_2 \\ &\leq L_f(1+\frac{L_g}{\mu_g})\|\mathbf{u}_\lambda^*(\mathbf{w})-\mathbf{u}^*\|_2 + L_{gg}\lambda(1+\frac{L_g}{\mu_g})\|\mathbf{u}_\lambda^*(\mathbf{w})-\mathbf{u}^*\|_2^2.\end{aligned}$$

□

Next, we bound $\|\mathbf{u}_\lambda^*(\mathbf{w}) - \mathbf{u}^*(\mathbf{w})\|_2$.

Lemma 4.27 *Under Assumption 4.10(ii), we have* $\|\mathbf{u}_\lambda^*(\mathbf{w}) - \mathbf{u}^*(\mathbf{w})\|_2 \leq \frac{G_f}{\lambda\mu_g}$.

Proof. By the definitions of $\mathbf{u}_\lambda^*(\mathbf{w}), \mathbf{u}^*(\mathbf{w})$, we have

$$\mathbf{u}_\lambda^*(\mathbf{w}) = \arg\min_{\mathbf{u}} \frac{1}{\lambda} f(\mathbf{w},\mathbf{u}) + g(\mathbf{w},\mathbf{u})$$
$$\mathbf{u}^*(\mathbf{w}) = \arg\min_{\mathbf{u}} g(\mathbf{w},\mathbf{u}).$$

By the optimality condition,

$$\frac{1}{\lambda}\nabla_2 f(\mathbf{w},\mathbf{u}_\lambda^*(\mathbf{w})) + \nabla_2 g(\mathbf{w},\mathbf{u}_\lambda^*(\mathbf{w})) = 0$$
$$\nabla_2 g(\mathbf{w},\mathbf{u}^*(\mathbf{w})) = 0.$$

Since $g(\mathbf{w},\mathbf{u})$ is μ_g-strongly convex w.r.t $\mathbf{u}$ for any $\mathbf{w}$, then we have

$$
\begin{aligned}
g(\mathbf{w}, \mathbf{u}_\lambda^*(\mathbf{w})) \geq & g(\mathbf{w}, \mathbf{u}^*(\mathbf{w})) + \nabla_2 g(\mathbf{w}, \mathbf{u}^*(\mathbf{w}))^\top (\mathbf{u}_\lambda^*(\mathbf{w}) - \mathbf{u}^*(\mathbf{w})) \\
& + \frac{\mu_g}{2} \|\mathbf{u}_\lambda^*(\mathbf{w}) - \mathbf{u}^*(\mathbf{w})\|_2^2 \\
g(\mathbf{w}, \mathbf{u}^*(\mathbf{w})) \geq & g(\mathbf{w}, \mathbf{u}_\lambda^*(\mathbf{w})) + \nabla_2 g(\mathbf{w}, \mathbf{u}_\lambda^*(\mathbf{w}))^\top (\mathbf{u}^*(\mathbf{w}) - \mathbf{u}_\lambda^*(\mathbf{w})) \\
& + \frac{\mu_g}{2} \|\mathbf{u}_\lambda^*(\mathbf{w}) - \mathbf{u}^*(\mathbf{w})\|_2^2.
\end{aligned}
$$

Adding these two inequalities yields:

$$
\begin{aligned}
\mu_g \|\mathbf{u}_\lambda^*(\mathbf{w}) - \mathbf{u}^*(\mathbf{w})\|_2^2 &\leq -\nabla_2 g(\mathbf{w}, \mathbf{u}_\lambda^*(\mathbf{w}))^\top (\mathbf{u}^*(\mathbf{w})) - \mathbf{u}_\lambda^*(\mathbf{w}))) \\
&= \frac{1}{\lambda} \nabla_2 f(\mathbf{w}, \mathbf{u}_\lambda^*(\mathbf{w}))^\top (\mathbf{u}^*(\mathbf{w})) - \mathbf{u}_\lambda^*(\mathbf{w}))) \\
&\leq \frac{1}{\lambda} \|\nabla_2 f(\mathbf{w}, \mathbf{u}_\lambda^*(\mathbf{w}))\|_2 \|(\mathbf{u}^*(\mathbf{w})) - \mathbf{u}_\lambda^*(\mathbf{w}))\|_2.
\end{aligned}
$$

Dividing both sides by $\|\mathbf{u}^*(\mathbf{w})) - \mathbf{u}_\lambda^*(\mathbf{w})\|_2$ and noting $\|\nabla_2 f(\mathbf{w}, \mathbf{u}_\lambda^*(\mathbf{w}))\|_2 \leq G_f$ concludes the proof.

□

Corollary 4.2 *Under the same setting as in Theorem 4.6 with $\lambda = O(\frac{1}{\epsilon}) > 2L_f/\mu_g$ and assume $\|\mathbf{y}_0 - \mathbf{y}^*(\mathbf{w}_0)\|_2^2 \leq O(\epsilon)$, then the following holds*

$$
\mathbb{E}\left[\|\nabla F(\mathbf{w}_\tau)\|_2\right] \leq O(\epsilon), \tag{4.89}
$$

with an iteration complexity of

$$
T = O\left(\max\left\{\frac{1}{\epsilon^3}, \frac{\sigma_f^2}{\epsilon^5}, \frac{\sigma_g^2}{\epsilon^7}\right\}\right), \tag{4.90}
$$

where $\tau \in \{0, \ldots, T-1\}$ is randomly sampled.

Proof. Combining Lemma 4.25 and Lemma 4.27, we have

$$
\begin{aligned}
\|\nabla F(\mathbf{w}_\tau)\|_2 &= \|\nabla F(\mathbf{w}_\tau) - \nabla_1 \bar{F}(\mathbf{w}_\tau, \mathbf{u}_\lambda^*(\mathbf{w}_\tau))\|_2 + \|\nabla_1 \bar{F}(\mathbf{w}_\tau, \mathbf{u}_\lambda^*(\mathbf{w}_\tau))\|_2 \\
&\leq L_f(1 + \frac{L_g}{\mu_g}) \frac{G_f}{\mu_g \lambda} + L_{gg} \lambda (1 + \frac{L_g}{\mu_g}) \frac{G_f^2}{\mu_g^2 \lambda^2} \\
&+ \|\nabla_1 \bar{F}(\mathbf{w}_\tau, \mathbf{u}_\lambda^*(\mathbf{w}_\tau)) - \nabla_1 \bar{F}(\mathbf{w}_\tau, \mathbf{u}_\tau)\|_2 + \|\nabla_1 \bar{F}(\mathbf{w}_\tau, \mathbf{u}_\tau)\|_2 \\
&\leq \frac{2L_f L_g G_f}{\mu_g^2 \lambda} + \frac{2L_{gg} L_g G_f^2}{\mu_g^3 \lambda} \\
&+ \|\nabla_1 \bar{F}(\mathbf{w}_\tau, \mathbf{u}_\lambda^*(\mathbf{w}_\tau)) - \nabla_1 \bar{F}(\mathbf{w}_\tau, \mathbf{u}_\tau)\|_2 + \|\nabla_1 \bar{F}(\mathbf{w}_\tau, \mathbf{u}_\tau)\|_2.
\end{aligned}
$$

Since $\bar{F}(\mathbf{w}, \mathbf{u})$ is $(\lambda\mu_g - L_f)$-strongly convex w.r.t $\mathbf{u}$, Lemma 1.6(c) implies that

$$(\lambda\mu_g - L_f)\|\mathbf{u}_\lambda^*(\mathbf{w}_\tau) - \mathbf{u}_\tau\|_2^2 \leq \frac{1}{(\lambda\mu_g - L_f)}\|\nabla_2\bar{F}(\mathbf{w}_\tau, \mathbf{u}_\tau) - \nabla_2\bar{F}(\mathbf{w}_\tau, \mathbf{u}_\lambda^*(\mathbf{w}_\tau))\|_2^2$$
$$= \frac{1}{(\lambda\mu_g - L_f)}\|\nabla_2\bar{F}(\mathbf{w}_\tau, \mathbf{u}_\tau)\|_2^2$$

Due to $\nabla_1\bar{F}(\mathbf{w}, \mathbf{u}) = \nabla_1 f(\mathbf{w}, \mathbf{u}) + \lambda(\nabla_1 g(\mathbf{w}, \mathbf{u}) - \nabla_1 g(\mathbf{w}, \mathbf{u}^*(\mathbf{w})))$, we have

$$\|\nabla_1\bar{F}(\mathbf{w}_\tau, \mathbf{u}_\lambda^*(\mathbf{w}_\tau)) - \nabla_1\bar{F}(\mathbf{w}_\tau, \mathbf{u}_\tau)\|_2 \leq (L_f + \lambda L_g)\|\mathbf{u}_\lambda^*(\mathbf{w}_\tau) - \mathbf{u}_\tau\|_2$$
$$\leq \frac{(L_f + \lambda L_g)}{(\lambda\mu_g - L_f)}\|\nabla_2\bar{F}(\mathbf{w}_\tau, \mathbf{u}_\tau)\|_2$$
$$\leq \frac{2(\lambda\mu_g/2 + \lambda L_g)}{\lambda\mu_g}\|\nabla_2\bar{F}(\mathbf{w}_\tau, \mathbf{u}_\tau)\|_2 = \frac{\mu_g + 2L_g}{\mu_g}\|\nabla_2\bar{F}(\mathbf{w}_\tau, \mathbf{u}_\tau)\|_2$$

where the last inequality uses $L_f \leq \lambda\mu_g/2$. Combining the above inequalities, we obtain

$$\|\nabla F(\mathbf{w}_\tau)\|_2 \leq \frac{2L_f L_g G_f}{\mu_g^2\lambda} + \frac{2L_{gg}L_g G_f^2}{\mu_g^3\lambda}$$
$$+ \frac{\mu_g + 2L_g}{\mu_g}\|\nabla_2\bar{F}(\mathbf{w}_\tau, \mathbf{u}_\tau)\|_2 + \|\nabla_1\bar{F}(\mathbf{w}_\tau, \mathbf{u}_\tau)\|_2.$$

From Theorem 4.6, we have

$$\mathbb{E}\big[\|\nabla_2\bar{F}(\mathbf{w}_\tau, \mathbf{u}_\tau)\|_2^2 + \|\nabla_1\bar{F}(\mathbf{w}_\tau, \mathbf{u}_\tau)\|_2^2\big] \leq \epsilon^2.$$

Hence, it follows that $\mathbb{E}[\|\nabla_2\bar{F}(\mathbf{w}_\tau, \mathbf{u}_\tau)\|_2] \leq \epsilon$ and $\mathbb{E}[\|\nabla_1\bar{F}(\mathbf{w}_\tau, \mathbf{u}_\tau)\|_2] \leq \epsilon$. If $\lambda = O(1/\epsilon)$, then $\mathbb{E}[\|\nabla F(\mathbf{w}_\tau)\|_2] \leq O(\epsilon)$. The iteration complexity can be established by substituting $\lambda = O(1/\epsilon)$ into Theorem 4.6 and noting that $C_\Upsilon = O(1)$ when $\|\mathbf{y}_0 - \mathbf{y}^*(\mathbf{w}_0)\|_2^2 \leq O(\epsilon)$.

□

Critical: The complexity of $O(1/\epsilon^7)$ is not the state-of-the-art sample complexity achievable under the same assumptions. Indeed, a double-loop large-batch method—similar to the one presented in Section 4.5.1.1 for solving the min-max problem $\min_{\bar{\mathbf{w}}} \max_{\mathbf{y}} \bar{f}(\bar{\mathbf{w}}, \mathbf{y})$—can yield a superior sample complexity of $O(1/\epsilon^6)$ for achieving the stationarity condition $\mathbb{E}[\|\nabla F(\mathbf{w})\|_2] \leq \epsilon^2$. To see this, we apply the results from Section 4.5.1.1, which indicates that a sample complexity for achieving $\mathbb{E}[\|\nabla \bar{F}(\bar{\mathbf{w}})\|_2^2] \leq \epsilon^2$ is $O\left(\frac{\bar{L}_F \bar{\sigma}_1^2}{\epsilon^4} + \frac{\bar{L}_F \bar{L}_1^2 \bar{\sigma}_2^2}{\bar{\mu}^2 \epsilon^4}\right)$. Here, $\bar{L}_F$ denotes the smoothness constant of the objective function $\bar{F}(\bar{\mathbf{w}}) = \max_{\mathbf{y}} \bar{f}(\bar{\mathbf{w}}, \mathbf{y})$. The remaining parameters are defined as follows:

- $\bar{L}_1 = O(\lambda)$ is the Lipschitz constant of $\nabla_1 \bar{f}(\cdot, \cdot)$;
- $\bar{\mu} = O(\lambda)$ is the strong concavity parameter of $\bar{f}(\cdot, \mathbf{y})$ with respect to $\mathbf{y}$;
- $\bar{\sigma}_2^2 = O(\lambda^2)$ represents the variance of the stochastic gradient with respect to $\mathbf{y}$;
- $\bar{\sigma}_1^2 = O(\lambda^2)$ is the variance of the stochastic gradient with respect to $\bar{\mathbf{w}} = (\mathbf{w}, \mathbf{u})$.

Given that we can establish $\bar{L}_F = O(1)$ independent of λ (Chen et al., 2025a, see Lemma B.7) and $\lambda = O(1/\epsilon)$, the total sample complexity reduces to $O(1/\epsilon^6)$.
However, it remains an open problem to develop a single-loop stochastic algorithm that achieves $O(1/\epsilon^6)$ complexity without requiring a large batch size or assuming mean-square smoothness (see next section for more discussion).

4.6 History and Notes

The optimization techniques presented in this chapter for stochastic compositional optimization are rooted in the pioneering work of Yuri Ermoliev (Ermoliev, 1976; Ermoliev and Wets, 1988). The monograph (Ermoliev, 1976), written in Ukrainian, laid the early foundations. Chapter 6 of the edited volume (Ermoliev and Wets, 1988) introduces an early form of the Stochastic Compositional Gradient Descent (SCGD) method, employing a sequence of moving average estimators $\mathbf{u}_t$ to track the inner function values at each iteration—referred to then simply as "averaging." The convergence analysis in these early works is largely limited to asymptotic results, if provided at all. Notably, these works considered a broader class of problems with functional constraints, which will be discussed further in Chapter 6.

The study of non-smooth compositional optimization, where a non-smooth convex function is composed with a smooth function, was first initiated in the works of Fletcher and Watson (1980); Fletcher (1982). Their proposed method, known as the *prox-linear method*, has since been extensively studied and developed in subsequent research (Lewis and Wright, 2009; Duchi and Ruan, 2018; Drusvyatskiy et al.,

2021; Duchi and Ruan, 2017; Drusvyatskiy and Paquette, 2019). We will consider non-smooth compositional optimization in next chapter.

The modern convergence analysis with non-asymptotic rates for stochastic compositional optimization was pioneered by Wang et al. (2017a). Their initial analysis established an $O(1/\epsilon^8)$ complexity for finding an ϵ-stationary solution to a smooth compositional problem, primarily due to suboptimal choices of learning rates. Subsequent works have aimed to improve this convergence rate (Ghadimi et al., 2020; Wang et al., 2017b; Chen et al., 2021a). The improved complexity of $O(1/\epsilon^5)$ for SCGD is derived by following the parameter settings introduced in Qi et al. (2021c). A further refined complexity of $O(1/\epsilon^4)$, under the assumption that the inner function is smooth, was achieved in Chen et al. (2021a). The use of a moving-average gradient estimator to attain the $O(1/\epsilon^4)$ complexity in stochastic compositional optimization is credited to (Ghadimi et al., 2020).

The modern variance-reduction technique for estimating the gradient of a smooth function originates from (Johnson and Zhang, 2013; Mahdavi and Jin, 2013; Zhang et al., 2013), and was inspired by earlier work (Schmidt et al., 2017) that established linear convergence for finite-sum problems with convex and smooth objectives. This technique is now widely known as SVRG. For the objective function $f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{w})$, the SVRG gradient estimator takes the form $\nabla f_i(\mathbf{w}_t) - \nabla f_i(\bar{\mathbf{w}}) + \nabla f(\bar{\mathbf{w}})$, where $\bar{\mathbf{w}}$ is a reference point whose full gradient $\nabla f(\bar{\mathbf{w}})$ is computed periodically.

For non-convex optimization, the variance reduction technique named SPIDER was initiated by Fang et al. (2018), which proposes a gradient estimator $\mathbf{v}_t = \mathbf{v}_{t-1} + \nabla f_i(\mathbf{w}_t) - \nabla f_i(\mathbf{w}_{t-1})$, with $\mathbf{v}$ being periodically re-initialized using either a full gradient or a large-batch gradient. This approach was earlier proposed under the name SARAH for convex optimization in (Nguyen et al., 2017). The technique later evolved into the STORM estimator (Cutkosky and Orabona, 2019), defined as $\mathbf{v}_t = (1-\beta)\mathbf{v}_{t-1} + \beta\nabla f(\mathbf{w}_t;\xi_t) + (1-\beta)\left[\nabla f(\mathbf{w}_t;\xi_t) - \nabla f(\mathbf{w}_{t-1};\xi_t)\right]$, which eliminates the need for periodic re-initialization.

Huo et al. (2018) applied the SVRG technique for finite-sum compositional optimization where both the inner and outer expectation is an average over a finite set. Hu et al. (2019) and Zhang and Xiao (2019) concurrently applied SARAH/SPIDER to compositional optimization with an expectation form and a finite-sum structure, and derived a complexity of $O(1/\epsilon^3)$ for the expectation form and $O(\sqrt{n}/\epsilon^2)$ for a finite-sum structure with n components. Qi et al. (2021a) applied the STORM estimator for SCO with a complexity of $O(1/\epsilon^3)$ and Chen et al. (2021b) applied the STORM estimator to only the inner function estimation for SCO with a complexity of $O(1/\epsilon^4)$.

The capped ℓ_1 norm for sparse regularization was introduced by Zhang (2013). The minimax concave penalty (MCP) was proposed by Zhang (2010), while the smoothly clipped absolute deviation (SCAD) regularizer was introduced by Fan and Li (2001). The proximal mappings for these non-convex regularizers were studied in (Gong et al., 2013). The non-convex piecewise affine regularization method for quantization was proposed by Ma and Xiao (2025). The theoretical analysis presented in Section 4.4 on non-convex optimization with non-convex regularizers fol-

lows the framework established by Xu et al. (2019a), whose results were applied by Deleu and Bengio (2021) to train sparse deep neural networks.

Stochastic weakly-convex–concave min–max optimization with a complexity of $O(1/\epsilon^6)$ was first studied by Rafique et al. (2018). When the problem is weakly-convex and strongly-concave, the complexity can be improved to $O(1/\epsilon^4)$ using double-loop algorithms (Rafique et al., 2018; Yan et al., 2020a). The analysis of SGDA for smooth non-convex min-max optimization was first established by Lin et al. (2020), who derived a complexity of $O(1/\epsilon^4)$ when using a large batch size on the order of $O(1/\epsilon^2)$ for problems that are strongly concave in the dual variable. Without employing a large batch size, the complexity degrades to $O(1/\epsilon^8)$, which also applies to problems lacking strong concavity. The analysis of the single-loop SMDA algorithm was provided by (Guo et al., 2021b), which also established the convergence guarantees for stochastic bilevel optimization using the first approach introduced in Section 4.5.3. A similar convergence result was achieved in Qiu et al. (2020), which employed moving-average gradient estimators for both the primal and dual variables. Chen et al. (2021a) obtained a complexity of $O(1/\epsilon^4)$ for smooth non-convex strongly-concave problems without relying on moving-average gradient estimators, under the stronger assumption that the Hessian/Jacobian matrix is Lipschitz continuous. An improved rate of $O(1/\epsilon^3)$ for smooth non-convex strongly-concave problems was established by (Huang et al., 2022) through the use of STORM estimators.

Bilevel optimization has a long and rich history (Bracken and McGill, 1973). The first complexity analysis of bilevel optimization was initiated by **?**, who employed the Neumann series to approximate the inverse of the Hessian. Their proposed double-loop stochastic algorithm achieves a sample complexity of $O(1/\epsilon^6)$ for solving the lower-level problem and $O(1/\epsilon^4)$ for the upper-level problem. Subsequent research has led to improved complexity bounds: $O(1/\epsilon^5)$ in (Hong et al., 2020), $O(1/\epsilon^4)$ in (Ji et al., 2020; Guo et al., 2021b; Chen et al., 2021a), and further down to $O(1/\epsilon^3)$ in (Yang et al., 2021; Khanduri et al., 2021; Guo et al., 2021a) under mean-square smoothness conditions. The analysis corresponding to Approach 1 in Section 4.5.3 can be found in (Qiu et al., 2022), while that of Approach 2 is provided in (Guo et al., 2021b).

Penalty-based metholds for bilevel optimization date back to (Ye et al., 1997), with recent developments appearing in (Liu et al., 2021, 2022; Shen and Chen, 2023). Lemma 4.26 is due to Kwon et al. (2023), which established a sample complexity of $O(1/\epsilon^7)$—comparable to Theorem 4.6—for a different double-loop algorithm. They also derived a complexity of $O(1/\epsilon^6)$ for an algorithm similar to update (4.77), except that the gradient estimators for both the lower- and upper-level functions are replaced with STORM estimators under stronger mean-square smoothness assumptions.

The complexity of $O(1/\epsilon^4)$ for stochastic compositional optimization is known to be optimal, as it matches the lower bound established for standard stochastic optimization (Arjevani et al., 2022). Moreover, under mean-square smoothness assumptions, a reduced complexity of $O(1/\epsilon^3)$ is also proven to be optimal (Arjevani et al., 2022).