Supplementary material

A Additional Motivating Examples

The bilevel optimization problem in (1) provides a versatile framework that covers a broad class of optimization problems. In addition to the motivating examples provided in the main body of the paper, here we also provide a generic example of stochastic convex constrained optimization that can be formulated as (1). We further present a more general form of the examples covered in the main body.

Generic Example: Stochastic convex optimization with many conic constraints. Consider the following convex optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \mathbb{E}[\tilde{f}(\mathbf{x},\theta)] \qquad \text{s.t.} \quad h(\mathbf{x},\xi) \in -\mathcal{K},\ \forall \xi \in \Omega,$$

where $\mathcal{K} \subseteq \mathbb{R}^d$ is a closed convex cone. This problem can be formulated as a special case of (1) by letting $\tilde{g}(\mathbf{x},\xi) = \frac{1}{2} d^2_{-\mathcal{K}}(h(\mathbf{x},\xi))$ where $d_{-\mathcal{K}}(\cdot) \triangleq \|\cdot - \mathcal{P}_{-\mathcal{K}}(\cdot)\|$ denotes the distance function and $\mathcal{P}_{-\mathcal{K}}(\cdot)$ denotes the projection map. Our proposed framework provides an efficient method for solving this class of problems when the projections onto $\mathcal{K}$ can be computed efficiently, while the projection onto the preimage $h^{-1}(-\mathcal{K},\xi)$ is not practical, e.g., when $\mathcal{K}$ is the positive semidefinite cone, computing a projection onto the preimage set requires solving a nonlinear SDP.

A.1 Lexicographic optimization

Example 1 (over-parameterized regression) can be generalized as a broader class of problem, which is known as lexicographic optimization [13] and uses the secondary loss to improve generalization. The problem can be formulated as the following stochastic simple bilevel optimization problem,

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\beta}) \quad \text{s.t.} \quad \boldsymbol{\beta} \in \arg\min_{\theta\in\mathcal{Z}} \ell_{\text{tr}}(\theta) = \mathbb{E}_{\mathcal{D}_{tr}}[\ell(y, \hat{y}_\theta(\mathbf{x}))] \tag{17}$$

In general, the lower-level problem could have multiple optimal solutions and be very sensitive to small perturbations. To tackle the issue, we use a secondary criterion $\mathcal{L}()$ to select some of the optimal solutions with our desired properties. For instance, we can find the optimal solutions with minimal ℓ_2-norm by letting $\mathcal{L}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|^2$, which is also known as *Lexicographic ℓ_2 Regularization*.

A.2 Lifelong learning

Example 2 (dictionary learning) is an instance of a popular framework known as lifelong learning, which can be formulated as follows,

$$\min_{\boldsymbol{\beta}} \frac{1}{n'} \sum_{i=1}^{n'} \ell\left(\langle \mathbf{x}'_i, \boldsymbol{\beta}\rangle, y'_i\right) \quad \text{s.t.} \quad \sum_{(\mathbf{x}_i,y_i)\in\mathcal{M}} \ell(\langle \mathbf{x}_i, \boldsymbol{\beta}\rangle, y_i) \le \sum_{(\mathbf{x}_i,y_i)\in\mathcal{M}} \ell(\langle \mathbf{x}_i, \boldsymbol{\beta}^{(t-1)}\rangle, y_i) \tag{18}$$

In this problem, the objective is the training loss on the current tasks $\mathcal{D}_t = \{(\mathbf{x}'_i, y'_i)\}_{i=1}^{n'}$. While the constraint enforces that the model parameterized by $\boldsymbol{\beta}$ performs no worse than the previous one on the episodic memory $\mathcal{M}$ (i.e., data samples from all the past tasks).

In the paper, we discuss a variant of the problem above, where we slightly change the constraint and ensure that the current model also minimizes the error on the past tasks. It can be formulated as the following finite-sum/stochastic simple bilevel optimization problem [12],

$$\min_{\boldsymbol{\beta}} \frac{1}{n'} \sum_{i=1}^{n'} \ell\left(\langle \mathbf{x}'_i, \boldsymbol{\beta}\rangle, y'_i\right) \quad \text{s.t.} \quad \boldsymbol{\beta} \in \operatorname*{argmin}_{\mathbf{z}} \sum_{(\mathbf{x}_i,y_i)\in\mathcal{M}} \ell\left(\langle \mathbf{x}_i, \mathbf{z}\rangle, y_i\right). \tag{19}$$

B Supporting lemmas

B.1 Proof of Lemma 4.1

Before we proceed to the proof for Lemma 4.1, we present the following technical lemma, which gives us an upper bound for a complex term appearing in the following analysis.

Lemma B.1. *Define $\rho_t = 1/(t+1)^\omega$ where $\omega \in (0,1]$ and $t \geq 1$. For all $t \geq 2$, let $\{s_t\}$ be a sequence of real numbers given by*

$$s_t = \sum_{\tau=2}^{t} \left(\rho_\tau \prod_{k=\tau}^{t} (1-\rho_k) \right)^2.$$

Then it holds that

$$s_t \leq \frac{1}{(t+1)^\omega}. \tag{20}$$

Proof. We prove the result by induction. For $t = 2$, we can verify that

$$S = \left(\frac{1}{3^\omega} \cdot \frac{3^\omega - 1}{3^\omega} \right)^2 \leq \frac{1}{3^{2\omega}} \leq \frac{1}{3^\omega}.$$

Now we suppose that the inequality in (20) holds when $t = T$ for some $T \geq 2$, i.e.,

$$s_T = \sum_{\tau=2}^{T} \left(\rho_\tau \prod_{k=\tau}^{T} (1-\rho_k) \right)^2 \leq \frac{1}{(t+1)^\omega}.$$

First note that the sequence $\{s_t\}$ satisfies the following recurrence relation:

$$\begin{aligned}
s_{T+1} = \sum_{\tau=2}^{T+1} \left(\rho_\tau \prod_{k=\tau}^{T+1} (1-\rho_k) \right)^2 &= (1-\rho_{T+1})^2 \sum_{\tau=2}^{T+1} \left(\rho_\tau \prod_{k=\tau}^{T} (1-\rho_k) \right)^2 \\
&= (1-\rho_{T+1})^2 \left[\sum_{\tau=2}^{T} \left(\rho_\tau \prod_{k=\tau}^{T} (1-\rho_k) \right)^2 + \rho_{T+1}^2 \right] \\
&= (1-\rho_{T+1})^2 (s_T + \rho_{T+1}^2).
\end{aligned}$$

Moreover, since $\alpha \in (0,1]$, we have $(T+2)^\omega - 1 \leq (t+1)^\omega$. Therefore, we obtain

$$\begin{aligned}
s_{T+1} &\leq \left(\frac{(T+2)^\omega - 1}{(T+2)^\omega} \right)^2 \left(\frac{1}{(t+1)^\omega} + \frac{1}{(T+2)^{2\omega}} \right) \\
&\leq \frac{((T+2)^\omega - 1)(t+1)^\omega}{(T+2)^{2\omega}} \left(\frac{1}{(t+1)^\omega} + \frac{1}{(T+1)^{2\omega}} \right) \\
&= \frac{(T+2)^\omega - 1}{(T+2)^{2\omega}} \frac{(T+1)^\omega + 1}{(T+1)^\omega} \\
&= \frac{(T+2)^\omega (t+1)^\omega + (T+2)^\omega - 1 - (t+1)^\omega}{(T+2)^{2\omega}(t+1)^\omega} \\
&\leq \frac{(T+2)^\omega (t+1)^\omega}{(T+2)^{2\omega}(t+1)^\omega} = \frac{1}{(T+2)^\omega}.
\end{aligned}$$

By induction, the inequality in (20) holds for all $t \geq 2$. □

Now we proceed to prove Lemma 4.1.

Proof of Lemma 4.1. We show the proof of part (i) here. The proof of part (ii) is very similar to part (i). The first step is to reformulate $\mathbf{e}_t = \widehat{\nabla} g_t - \nabla g(\mathbf{x}_t)$ as the sum of a martingale difference sequence. For $t \geq 1$, by unrolling the reucurrence we have

$$\begin{aligned}
\mathbf{e}_t &= (1-\beta_t)\mathbf{e}_{t-1} + \beta_t(\nabla \tilde{g}(\mathbf{x}_t, \xi_t) - \nabla g(\mathbf{x}_t)) \\
&\quad + (1-\beta_t)(\nabla \tilde{g}(\mathbf{x}_t, \xi_t) - \nabla \tilde{g}(\mathbf{x}_{t-1}, \xi_t) - (\nabla g(\mathbf{x}_t) - \nabla g(\mathbf{x}_{t-1})) \\
&= \prod_{k=2}^{t}(1-\beta_k)\mathbf{e}_1 + \sum_{\tau=2}^{t} \prod_{k=\tau}^{t} (1-\beta_k)(\nabla \tilde{g}(\mathbf{x}_\tau, \xi_\tau) - \nabla \tilde{g}(\mathbf{x}_{\tau-1}, \xi_\tau) - (\nabla g(\mathbf{x}_\tau) - \nabla g(\mathbf{x}_{\tau-1})) \\
&\quad + \sum_{\tau=2}^{t} \beta_\tau \prod_{k=\tau+1}^{t} (1-\beta_k)(\nabla \tilde{g}(\mathbf{x}_\tau, \xi_\tau) - \nabla g(\mathbf{x}_\tau)).
\end{aligned} \tag{21}$$

Thus, we can write $\mathbf{e}_t$ as the sum $\mathbf{e}_t = \sum_{\tau=1}^{t} \zeta_\tau$, where we define $\zeta_1 = \prod_{k=2}^{t}(1-\beta_k)\mathbf{e}_1$ and

$$\zeta_\tau = \prod_{k=\tau}^{t}(1-\beta_k)(\nabla\tilde{g}(\mathbf{x}_\tau,\xi_\tau) - \nabla\tilde{g}(\mathbf{x}_{\tau-1},\xi_\tau) - (\nabla g(\mathbf{x}_\tau) - \nabla g(\mathbf{x}_{\tau-1})) \tag{22}$$

$$+ \beta_\tau \prod_{k=\tau+1}^{t}(1-\beta_k)(\nabla\tilde{g}(\mathbf{x}_\tau,\xi_\tau) - \nabla g(\mathbf{x}_\tau)) \tag{23}$$

for $\tau > 1$. Recall that $\mathbf{e}_1 = \nabla\tilde{g}(\mathbf{x}_t,\zeta_1) - \nabla g(\mathbf{x}_1)$. We observe that $\mathbb{E}[\zeta_\tau|\mathcal{F}_{\tau-1}] = 0$ where $\mathcal{F}_{\tau-1}$ is the σ-field generated by $\{\mathbf{x}_1,\xi_1,\ldots,\mathbf{x}_{\tau-1},\xi_{\tau-1}\}$. Therefore, $\{\zeta_\tau\}_{\tau=1}^{t}$ is a martingale difference sequence.

Next, we derive upper bounds of $\|\zeta_\tau\|$. To begin with, we observe that for any $\tau = 1,2,\ldots,t$,

$$\prod_{k=\tau}^{t}(1-\beta_k) = \prod_{k=\tau}^{t}\left(1 - \frac{1}{(k+1)^\omega}\right) = \prod_{k=\tau}^{t}\frac{(k+1)^\omega - 1}{(k+1)^\omega} \leq \prod_{k=\tau}^{t}\frac{k^\alpha}{(k+1)^\omega} = \frac{\tau^\omega}{(t+1)^\omega}, \tag{24}$$

where we used the fact that $(k+1)^\omega - 1 \leq k^\omega$ in the last inequality. By using the above inequality, we can bound $\|\zeta_1\|$ as follows:

$$\|\zeta_1\| = \prod_{k=2}^{t}(1-\beta_k)\|\mathbf{e}_1\| \leq \frac{2^\omega}{(t+1)^\omega}\|\nabla\tilde{g}(\mathbf{x}_1,\xi_1) - \nabla g(\mathbf{x}_1)\| = \frac{2^\omega\sigma_1}{(t+1)^\omega}\frac{\|\nabla\tilde{g}(\mathbf{x}_1,\xi_1) - \nabla g(\mathbf{x}_1)\|}{\sigma_1}.$$

Define $c_1 = \frac{2^\omega\sigma_g}{(T+1)^\omega}$, then by Assumption 2.3(ii) we have $\mathbb{E}[\exp(\|\zeta_1\|^2/c_1^2)] \leq \exp(1)$. Moreover, for $\tau > 1$, by triangle inequality, $\|\zeta_\tau\|$ can be bounded by

$$\begin{aligned}
\|\zeta_\tau\| &\leq \prod_{k=\tau}^{t}(1-\beta_k)(\|\nabla\tilde{g}(\mathbf{x}_\tau,\xi_\tau) - \nabla\tilde{g}(\mathbf{x}_{\tau-1},\xi_\tau)\| + \|(\nabla g(\mathbf{x}_\tau) - \nabla g(\mathbf{x}_{\tau-1})\|) \\
&\quad + \beta_\tau \prod_{k=\tau+1}^{t}(1-\beta_k)\|\nabla\tilde{g}(\mathbf{x}_\tau,\xi_\tau) - \nabla g(\mathbf{x}_\tau)\| \\
&\leq 2L_g\|\mathbf{x}_\tau - \mathbf{x}_{\tau-1}\|\prod_{k=\tau}^{t}(1-\beta_k) + \|\nabla\tilde{g}(\mathbf{x}_\tau,\xi_\tau) - \nabla g(\mathbf{x}_\tau)\|\beta_\tau\prod_{k=\tau+1}^{t}(1-\beta_k) \\
&= 2L_g\gamma_\tau D\prod_{k=\tau}^{t}(1-\beta_k) + \|\nabla\tilde{g}(\mathbf{x}_\tau,\xi_\tau) - \nabla g(\mathbf{x}_\tau)\|\beta_\tau\prod_{k=\tau+1}^{t}(1-\beta_k) \\
&\leq 2L_gD\beta_\tau\prod_{k=\tau}^{t}(1-\beta_k) + \frac{3^\omega}{3^\omega-1}\|\nabla\tilde{g}(\mathbf{x}_\tau,\xi_\tau) - \nabla g(\mathbf{x}_\tau)\|\beta_\tau\prod_{k=\tau}^{t}(1-\beta_k) \\
&= \left(2L_gD + \frac{3^\omega}{3^\omega-1}\|\nabla\tilde{g}(\mathbf{x}_\tau,\xi_\tau) - \nabla g(\mathbf{x}_\tau)\|\right)\beta_\tau\prod_{k=\tau}^{t}(1-\beta_k) \\
&= \left(2L_gD + \frac{3^\omega\sigma_g}{3^\omega-1}\frac{\|\nabla\tilde{g}(\mathbf{x}_\tau,\xi_\tau) - \nabla g(\mathbf{x}_\tau)\|}{\sigma_g}\right)\beta_\tau\prod_{k=\tau}^{t}(1-\beta_k)
\end{aligned} \tag{25}$$

Define $c_\tau = (2L_gD + \frac{3^\omega\sigma_g}{3^\omega-1})\beta_\tau\prod_{k=\tau}^{t}(1-\beta_k)$. Note that if we have $\mathbb{E}[\exp(X_1^2/c_1^2)] \leq 1$ and $\mathbb{E}[\exp(X_2^2/c_2^2)] \leq 1$, then we have $\mathbb{E}[\exp((X_1+X_2)^2/(c_1+c_2)^2)] \leq 1$ [39]. Thus, we have $\mathbb{E}[\exp(\|\zeta_\tau\|^2/c_\tau^2)] \leq \exp(1)$ for all $1 \leq \tau \leq t$. Hence by proposition E.2, with probability $1-\delta'$

$$\|\mathbf{e}_t\| \leq c\cdot\sqrt{\sum_{\tau=1}^{t}c_\tau^2\log\frac{2d}{\delta'}} \tag{26}$$

where c is an absolute constant, d is the number of dimension, and $\sum_{\tau=1}^{T} c_\tau^2$ can be bounded by Lemma B.1 as follows,

$$\begin{aligned}\sum_{\tau=1}^{t} c_\tau^2 = c_1^2 + \sum_{\tau=2}^{t} c_\tau^2 &= \frac{2^{2\omega}\sigma_g^2}{(T+1)^{2\omega}} + (2L_gD + \frac{3^\omega}{3^\omega-1}\sigma_g)^2 \sum_{\tau=2}^{T}(\beta_\tau \prod_{k=\tau}^{T}(1-\beta_k))^2 \\ &\leq \frac{2^{2\omega}\sigma_g^2}{(T+1)^{2\omega}} + \frac{(2L_gD + \frac{3^\omega}{3^\omega-1}\sigma_g)^2}{(t+1)^\omega} \\ &\leq \frac{((\sqrt{2})^\omega\sigma_g)^2}{(t+1)^\omega} + \frac{(2L_gD + \frac{3^\omega}{3^\omega-1}\sigma_g)^2}{(t+1)^\omega} \\ &\leq \frac{2(2L_gD + \frac{3^\omega}{3^\omega-1}\sigma_g)^2}{(t+1)^\omega}\end{aligned} \tag{27}$$

where the last inequality follows from the fact that $(\sqrt{2})^\omega \leq 3^\omega/(3^\omega-1)$ for any $\omega \in (0,1]$. Combining (26) and (27), we have with probability at least $1-\delta^{'}$,

$$\|\nabla g(\mathbf{x}_t) - \widehat{\nabla} g_t\| \leq c\sqrt{2}(2L_gD + \frac{3^\omega}{3^\omega-1}\sigma_g)(t+1)^{-\omega/2}\sqrt{\log(2d/\delta^{'})} \stackrel{\text{def}}{=} K_{1,t} \tag{28}$$

Similarly with probability at least $1-\delta^{'}$,

$$|g(\mathbf{x}_t) - \hat{g}_t| \leq c\sqrt{2}(2L_lD + \frac{3^\omega}{3^\omega-1}\sigma_l)(t+1)^{-\omega/2}\sqrt{\log(2d/\delta^{'})} \stackrel{\text{def}}{=} K_{0,t} \tag{29}$$

and with probability at least $1-\delta^{'}$,

$$\|\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t\| \leq c\sqrt{2}(2L_fD + \frac{3^\omega}{3^\omega-1}\sigma_f)(t+1)^{-\omega/2}\sqrt{\log(2d/\delta^{'})} \stackrel{\text{def}}{=} K_{2,t} \tag{30}$$

where c is an absolute constant and d is the dimension of vectors. We can use union bound to obtain that these three inequalities hold for at least probability $1-3\delta^{'} = 1-\delta$. For simplicity, we define constant A_1^α and A_0^α such that,

$$A_1^\alpha(t+1)^{-\omega/2}\sqrt{\log(6d/\delta)} = K_{1,t} \quad \text{and} \quad A_0^\alpha(t+1)^{-\omega/2}\sqrt{\log(6d/\delta)} = K_{0,t} \tag{31}$$

and similarly $A_2^\alpha(t+1)^{-\omega/2}\sqrt{\log(6d/\delta)} = K_{2,t}$. □

B.2 Proof of Lemma 4.3

Proof. When $t = \lfloor t/q \rfloor q$, we have $\widehat{\nabla} g_t = \nabla g(\mathbf{x}_t)$, since we take the full batch. When $t \neq \lfloor t/q \rfloor q$, set $t_0 = \lfloor t/q \rfloor q$, and

$$\epsilon_{j,i} = \frac{1}{S}(\nabla g_{\mathcal{S}(i)}(\mathbf{x}_j) - \nabla g_{\mathcal{S}(i)}(\mathbf{x}_{j-1}) - \nabla g(\mathbf{x}_j) + \nabla g(\mathbf{x}_{j-1})) \tag{32}$$

where i is the index with $\mathcal{S}(i)$ denoting the i-th random component function selected at iteration t. Furthermore, from the update rule, we have $\|\mathbf{x}_j - \mathbf{x}_{j-1}\| = \gamma_j\|\mathbf{s}_{j-1} - \mathbf{x}_{j-1}\| \leq D\gamma$. And we have,

$$\begin{aligned}\|\epsilon_{j,i}\| &\leq \frac{1}{S}(\|\nabla g_i(\mathbf{x}_j) - \nabla g_i(\mathbf{x}_{j-1})\| + \|\nabla g(\mathbf{x}_j) - \nabla g(\mathbf{x}_{j-1})\|) \\ &\leq \frac{2L_g}{S}\|\mathbf{x}_j - \mathbf{x}_{j-1}\| \leq \frac{2L_gD\gamma}{S}\end{aligned} \tag{33}$$

for all $t_0 \leq j \leq t$ and $1 \leq i \leq S$. On the other hand, we have,

$$\begin{aligned}\|\widehat{\nabla} g_t - \nabla g(\mathbf{x}_t)\| &= \|\nabla g_{\mathcal{S}}(\mathbf{x}_k) - \nabla g_{\mathcal{S}}(\mathbf{x}_{k-1}) - \nabla g(\mathbf{x}_k) + \nabla g(\mathbf{x}_{k-1}) + (\nabla g_{k-1} - \nabla g(\mathbf{x}_{k-1}))\| \\ &= \|\sum_{j=t_0+1}^{T}(\nabla g_{\mathcal{S}}(\mathbf{x}_j) - \nabla g_{\mathcal{S}}(\mathbf{x}_{j-1}) - \nabla g(\mathbf{x}_j) + \nabla g(\mathbf{x}_{j-1})) \\ &+ (\nabla g_{\mathcal{S}_1}(\mathbf{x}_{t_0}) - \nabla g(\mathbf{x}_{t_0}))\| \\ &= \|\sum_{j=t_0+1}^{T}\sum_{i=1}^{S}\epsilon_{j,i} + \sum_{i=1}^{S}\epsilon_{t_0,i}\| = \|\sum_{j=t_0+1}^{T}\sum_{i=1}^{S}\epsilon_{j,i}\|\end{aligned} \tag{34}$$

Then by Proposition E.1, we have

$$\mathbb{P}(\|\widehat{\nabla} g_t - \nabla g(\mathbf{x}_t)\| \geq \lambda) \leq 4\exp(-\frac{\lambda^2}{4S(t-t_0)\frac{4L_g^2 D^2\gamma^2}{S^2}}) \leq 4\exp(-\frac{\lambda^2}{16L_g^2 D^2\gamma^2}) \tag{35}$$

where the last inequality follows from the fact $S = \sqrt{n}$ and $t - t_0 \leq q = \sqrt{n}$. By setting $\lambda = (4L_g D\gamma\sqrt{\log(4/\delta')})$ for some $\delta' \in (0,1)$, we have with probability at least $1-\delta'$,

$$\|\widehat{\nabla} g_t - \nabla g(\mathbf{x}_t)\| \leq 4L_g D\gamma\sqrt{\log(4/\delta')} \tag{36}$$

Similarly, with probability at least $1-\delta'$,

$$\|\hat{g}_t - g(\mathbf{x}_t)\| \leq 4L_l D\gamma\sqrt{\log(4/\delta')} \tag{37}$$

and with probability $1-\delta'$,

$$\|\widehat{\nabla} f_t - \nabla f(\mathbf{x}_t)\| \leq 4L_f D\gamma\sqrt{\log(4/\delta')} \tag{38}$$

Then by union bound and $\delta = 3\delta'$, we show these three equalities hold with probability $1-\delta$.

□

B.3 Proof of Lemma 4.2

Proof. Let $\mathbf{x}_g^*$ be any point in $\mathcal{X}_g^*$, i.e., any optimal solution of the lower-level problem. By definition, we have $g\left(\mathbf{x}_g^*\right) = g^*$. Since g is convex and $g^* \leq g\left(\mathbf{x}_0\right)$, we have

$$g\left(\mathbf{x}_0\right) - g\left(\mathbf{x}_t\right) \geq g^* - g\left(\mathbf{x}_t\right) = g\left(\mathbf{x}_g^*\right) - g\left(\mathbf{x}_t\right) \geq \left\langle \nabla g\left(\mathbf{x}_t\right), \mathbf{x}_g^* - \mathbf{x}_t \right\rangle \tag{39}$$

Add and subtract terms in (47), we have,

$$\langle \widehat{\nabla} g_t, \mathbf{x}_g^* - \mathbf{x}_t\rangle + \hat{g}_t - g(\mathbf{x}_0) \leq |\langle \widehat{\nabla} g_t - \nabla g(\mathbf{x}_t), \mathbf{x}_g^* - \mathbf{x}_t\rangle| + |\hat{g}_t - g(\mathbf{x}_t)| \tag{40}$$

Considering the random hyperplane we used in (9), we want to prove the following inequality holds with high probability,

$$\langle \widehat{\nabla} g_t, \mathbf{x}_g^* - \mathbf{x}_t\rangle + \hat{g}_t - g(\mathbf{x}_0) \leq K_t \tag{41}$$

Recall $K_t = K_{0,t} + DK_{1,t}$. And $K_{0,t}$ and $K_{1,t}$ were set as the high probability bounds of $\|\widehat{\nabla} g_t - \nabla g(\mathbf{x}_t)\|$ and $|\hat{g}_t - g(\mathbf{x}_t)|$ in Lemma 4.1 for Algorithm 1 or Lemma 4.3 for Algorithm 2. Then compare the two inequalities above and use Jensen's inequality, $|\langle \widehat{\nabla} g_t, \mathbf{x}_g^* - \mathbf{x}_t\rangle| + |\hat{g}_t - g(\mathbf{x}_0)| \leq K_t$ holds with high probability $1-\delta$ for all $t \geq 0$. Hence, Lemma 4.2 holds with probability $1-\delta$ for all $t \geq 0$. □

B.4 Improvement in one step

The following lemma characterizes the improvement of both the upper-level and lower-level objective values after one step of the algorithms.

Lemma B.2. *If Assumptions 2.1, 2.2, 2.3 are satisfied,*

(i) For all $t \geq 0$, assume that $\mathcal{X}_g^* \subset \mathcal{X}_t$. Then we have

$$\gamma_{t+1}\mathcal{G}(\mathbf{x}_t) \leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \gamma_{t+1} D\|\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t\| + \frac{L_f D^2 \gamma_{t+1}^2}{2} \tag{42}$$

As a corollary, if f is convex, we further have

$$f(\mathbf{x}_{t+1}) - f^* \leq (1-\gamma_{t+1})(f(\mathbf{x}_t) - f^*)) + \gamma_{t+1} D\|\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t\| + \frac{L_f D^2 \gamma_{t+1}^2}{2}. \tag{43}$$

(ii) We have

$$g(\mathbf{x}_{t+1}) - g(\mathbf{x}_0) \leq (1-\gamma_{t+1})(g(\mathbf{x}_t) - g(\mathbf{x}_0)) + D\gamma_{t+1}(\|\nabla g(\mathbf{x}_t) - \widehat{\nabla} g_t\| + K_{1,t}) + \gamma_{t+1}(\|g(x_t) - \hat{g}_t\| + K_{0,t}) + \frac{L_g D^2 \gamma_{t+1}^2}{2}. \quad (44)$$

Proof. (i) Based on the L_f-smoothness of the expected function f we show that $f(\mathbf{x}_{t+1})$ is bounded by

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L_f}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \quad (45)$$

Replace the terms $\mathbf{x}_{t+1} - \mathbf{x}_t$ by $\gamma_{t+1}(\mathbf{s}_t - \mathbf{x}_t)$ and add and subtract the term $\gamma_{t+1}\widehat{\nabla} f_t^T (\mathbf{s}_t - \mathbf{x}_t)$ to the right hand side to obtain,

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + \gamma_{t+1}(\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t)^\top (\mathbf{s}_t - \mathbf{x}_t) + \gamma_{t+1}\widehat{\nabla} f_t^\top (\mathbf{s}_t - \mathbf{x}_t) + \frac{L_f}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \quad (46)$$

By Lemma 4.2, $\mathcal{X}_g^* \subseteq \mathcal{X}_t$ with high probability $1-\delta$, for all $t = 1, \ldots, T$. Note that if we define $\mathbf{s}_t' = \arg\max_{\mathbf{s}\in\mathcal{X}_t}\{\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{s}\rangle\}$. Recall that FW gap is $\mathcal{G}(\hat{\mathbf{x}}) = \max_{s\in\mathcal{X}_g^*}\{\langle \nabla f(\hat{\mathbf{x}}), \hat{\mathbf{x}} - \mathbf{s}\rangle\}$. We can replace the inner product $\langle \widehat{\nabla} f_t, \mathbf{s}_t\rangle$ by its upper bound $\langle \widehat{\nabla} f_t, \mathbf{s}_t'\rangle$. Applying this substitution leads to

$$\begin{aligned}
f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \gamma_{t+1}(\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t)^\top (\mathbf{s}_t - \mathbf{x}_t) + \gamma_{t+1}\widehat{\nabla} f_t^\top (\mathbf{s}_t' - \mathbf{x}_t) + \frac{L_f}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&= f(\mathbf{x}_t) + \gamma_{t+1}(\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t)^\top (\mathbf{s}_t - \mathbf{x}_t) + \gamma_{t+1}(\widehat{\nabla} f_t - \nabla f(\mathbf{x}_t))^\top (\mathbf{s}_t' - \mathbf{x}_t) \\
&\quad - \gamma_{t+1}\nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{s}_t') + \frac{L_f}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&\leq f(\mathbf{x}_t) + \gamma_{t+1}(\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t)^\top (\mathbf{s}_t - \mathbf{s}_t') - \gamma_{t+1}\mathcal{G}(\mathbf{x}_t) + \frac{L_f}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&\leq f(\mathbf{x}_t) + \gamma_{t+1} D\|\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t\| - \gamma_{t+1}\mathcal{G}(\mathbf{x}_t) + \frac{L_f \gamma_{t+1}^2 D^2}{2}
\end{aligned} \quad (47)$$

Rearrange the terms for the inequality above, we can obtain,

$$\gamma_{t+1}\mathcal{G}(\mathbf{x}_t) \leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \gamma_{t+1} D\|\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t)\| + \frac{L_f \gamma_{t+1}^2 D^2}{2} \quad (48)$$

As a simple corollary, since $\mathcal{G}(\mathbf{x}_t) \geq f(\mathbf{x}_t) - f^*$ when f is convex, we have,

$$f(\mathbf{x}_{t+1}) - f^* \leq (1-\gamma_{t+1})(f(\mathbf{x}_t) - f^*)) + \gamma_{t+1} D\|\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t\| + \frac{L_f D^2 \gamma_{t+1}^2}{2} \quad (49)$$

(ii) Based on the L_g-smoothness of the expected function g we show that $g(\mathbf{x}_{t+1})$ is bounded by

$$g(\mathbf{x}_{t+1}) \leq g(\mathbf{x}_t) + \nabla g(\mathbf{x}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L_g}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \quad (50)$$

Replace the terms $\mathbf{x}_{t+1} - \mathbf{x}_t$ by $\gamma_{t+1}(\mathbf{s}_t - \mathbf{x}_t)$ and add and subtract the term $\gamma_{t+1}\widehat{\nabla} g_t^\top (\mathbf{s}_t - \mathbf{x}_t)$ to the right-hand side to obtain,

$$g(\mathbf{x}_{t+1}) \leq g(\mathbf{x}_t) + \gamma_{t+1}(\nabla g(\mathbf{x}_t) - \widehat{\nabla} g_t)^\top (\mathbf{s}_t - \mathbf{x}_t) + \gamma_{t+1}\widehat{\nabla} g_t^\top (\mathbf{s}_t - \mathbf{x}_t) + \frac{L_g}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \quad (51)$$

Now by definition of the set $\mathcal{X}_t$, using $\langle \widehat{\nabla} g_t, \mathbf{s}_t - \mathbf{x}_t\rangle \leq g(\mathbf{x}_0) - \hat{g}_t + K_{0,t} + DK_{1,t}$. In addition, we could use Cauchy–Schwarz inequality to upper bound the second term. Then add and subtract $\gamma_{t+1} g(\mathbf{x}_0)$ on the right hand side to obtain,

$$\begin{aligned}
g(\mathbf{x}_{t+1}) &\leq g(\mathbf{x}_t) + \gamma_{t+1}(g(\mathbf{x}_0) - g(\mathbf{x}_t)) + \gamma_{t+1} D\|\nabla g(\mathbf{x}_t) - \widehat{\nabla} g_t\| \\
&\quad + \gamma_{t+1}(g(\mathbf{x}_t) - \hat{g}_t) + \gamma_{t+1}(K_{0_t} + DK_{1_t}) + \frac{L_g}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2
\end{aligned} \quad (52)$$

Then subtract $g(\mathbf{x}_0)$ on both sides,

$$\begin{aligned}
g(\mathbf{x}_{t+1}) - g(\mathbf{x}_0) &\leq (1-\gamma_{t+1})(g(\mathbf{x}_t) - g(\mathbf{x}_0)) \\
&\quad + \gamma_{t+1}(D\|\nabla g(\mathbf{x}_t) - \widehat{\nabla} g_t\| + \|g(\mathbf{x}_t) - \hat{g}_t\| + K_{0,t} + DK_{1,t}) + \frac{L_g}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2
\end{aligned} \quad (53)$$

and the claim in the lemma follows. □

C Proof of Theorem for Algorithm 1

C.1 Proof of Theorem 4.4

Proof. For **lower-level**, by Lemma B.2, we have

$$\begin{aligned} g(\mathbf{x}_{t+1}) - g(\mathbf{x}_0) \leq (1-\gamma_{t+1})(g(\mathbf{x}_t) - g(\mathbf{x}_0)) + D\gamma_{t+1}(\|\nabla g(\mathbf{x}_t) - \widehat{\nabla g}_t\| + K_{1,t}) \\ + \gamma_{t+1}(\|g(x_t) - \hat{g}_t\| + K_{0,t}) + \frac{L_g D^2 \gamma_{t+1}^2}{2} \end{aligned} \tag{54}$$

By Lemma 4.1, we have $\|\nabla g(\mathbf{x}_t) - \widehat{\nabla g}_t\| \leq K_{1,t}$ and $\|g(x_t) - \hat{g}_t\| \leq K_{0,t}$ with probability $1-\delta^{'}$. Plug them in the inequality above to obtain,

$$\begin{aligned} g(\mathbf{x}_{t+1}) - g(\mathbf{x}_0) &\leq (1-\gamma_{t+1})(g(\mathbf{x}_t) - g(\mathbf{x}_0)) + 2\gamma_{t+1}(DK_{1,t} + K_{0,t}) + \frac{L_g D^2 \gamma_{t+1}^2}{2} \\ &\leq (1 - \frac{1}{t+1}) g(\mathbf{x}_t) - g(\mathbf{x}_0) \\ &+ \frac{2(DA_1^1\sqrt{\log(6d/\delta^{'})} + A_0^1\sqrt{\log(6/\delta^{'})})}{(t+1)^{3/2}} + \frac{L_g D^2}{2(t+1)^2} \end{aligned} \tag{55}$$

with probability $1-\delta^{'}$ for all t. Let $C_1 = 4(DA_1^1 + A_0^1)$ and $\delta = t\delta^{'}$. Then we can sum all the inequality up for all t to obtain,

$$\begin{aligned} g(\mathbf{x}_{t+1}) - g(\mathbf{x}_0) &\leq (1 - \frac{1}{t+1}) g(\mathbf{x}_t) - g(\mathbf{x}_0) + \frac{C_1/2\sqrt{\log(6d/\delta^{'})}}{(t+1)^{3/2}} + \frac{L_g D^2}{2(t+1)^2} \\ &= \prod_{i=1}^{t}(1 - \frac{1}{i+1})(g(\mathbf{x}_0) - g(\mathbf{x}_0)) + \sum_{k=1}^{t} \frac{C_1/2\sqrt{\log(6d/\delta^{'})}}{(k+1)^{3/2}} \prod_{i=k+1}^{t}(1 - \frac{1}{i+1}) \\ &+ \sum_{k=1}^{t} \frac{L_g D^2}{2(k+1)^2} \prod_{i=k+1}^{t}(1 - \frac{1}{i+1}) \\ &\leq 0 + \frac{C_1/2\sqrt{\log(6d/\delta^{'})}}{t+1} \sum_{k=1}^{t} \frac{1}{\sqrt{k+1}} + \frac{L_g D^2}{2(t+1)} \sum_{k=1}^{t} \frac{1}{k+1} \\ &\leq \frac{C_1\sqrt{\log(6d/\delta^{'})}}{\sqrt{t+1}} + \frac{L_g D^2}{2(t+1)}(1 + \log t) \\ &\leq \frac{C_1\sqrt{\log(6td/\delta)}}{\sqrt{t+1}} + \frac{L_g D^2 \log t}{t+1} \end{aligned} \tag{56}$$

with probability $1-\delta$.

For **upper-level**, by Lemma B.2, we have

$$f(\mathbf{x}_{t+1}) - f^* \leq (1-\gamma_{t+1})(f(\mathbf{x}_t) - f^*) + D\gamma_{t+1}(\|\nabla f(\mathbf{x}_t) - \widehat{\nabla f}_t\|) + \frac{L_g D^2 \gamma_{t+1}^2}{2} \tag{57}$$

By Lemma 4.1, we have $\|\nabla f(\mathbf{x}_t) - \widehat{\nabla f}_t\| \leq \frac{A_2^1\sqrt{log(6d/\delta^{'})}}{(t+1)^{1/2}}$ with probability $1-\delta^{'}$. Plug it in the inequality above to obtain,

$$f(\mathbf{x}_{t+1}) - f^* \leq (1 - \frac{1}{t+1})(f(\mathbf{x}_t) - f^*) + \frac{DA_2^1\sqrt{\log(6d/\delta^{'})}}{(t+1)^{3/2}} + \frac{L_g D^2}{2(t+1)^2} \tag{58}$$

with probability $1-\delta^{'}$ for all t. Then we can sum all the inequality up for all t to obtain,

$$\begin{aligned}
f(\mathbf{x}_{t+1}) - f^* &\leq (1-\frac{1}{t+1})(f(\mathbf{x}_t)-f^*) + \frac{DA_2^1\sqrt{\log(6d/\delta^{'})}}{(t+1)^{3/2}} + \frac{L_g D^2}{2(t+1)^2} \\
&= \prod_{i=1}^{t}(1-\frac{1}{i+1})(f(\mathbf{x}_0)-f^*) + \sum_{k=1}^{t}\frac{DA_2^1\sqrt{\log(6d/\delta^{'})}}{(k+1)^{3/2}}\prod_{i=k+1}^{t}(1-\frac{1}{i+1}) \\
&+ \sum_{k=1}^{T}\frac{L_g D^2}{2(k+1)^2}\prod_{i=k+1}^{T}(1-\frac{1}{i+1}) \\
&\leq \frac{f(\mathbf{x}_0)-f^*}{t+1} + \frac{DA_2^1\sqrt{\log(d/\delta^{'})}}{t+1}\sum_{k=1}^{T}\frac{1}{\sqrt{k+1}} + \frac{L_g D^2}{2(t+1)}\sum_{k=1}^{T}\frac{1}{k+1} \\
&\leq \frac{f(\mathbf{x}_0)-f^*}{t+1} + \frac{2DA_2^1\sqrt{\log(6d/\delta^{'})}}{\sqrt{t+1}} + \frac{L_g D^2}{2(t+1)}(1+\log t) \\
&\leq \frac{f(\mathbf{x}_0)-f^*}{t+1} + \frac{2DA_2^1\sqrt{\log(6td/\delta)}}{\sqrt{t+1}} + \frac{L_g D^2 \log t}{(t+1)}
\end{aligned} \tag{59}$$

with probability $1-\delta = 1-t\delta^{'}$. Let $C_2 = 2DA_2^1$. The theorem is obtained.

□

C.2 Proof of Theorem 4.5

Proof. For **lower-level**, by Lemma B.2, we have

$$\begin{aligned}
g(\mathbf{x}_{t+1}) - g(\mathbf{x}_0) \leq (1-\gamma_{t+1})(g(\mathbf{x}_t)-g(\mathbf{x}_0)) + D\gamma_{t+1}(\|\nabla g(\mathbf{x}_t) - \widehat{\nabla} g_t\| + K_{1,t}) \\
+ \gamma_{t+1}(\|g(x_t) - \hat{g}_t\| + K_{0,t}) + \frac{L_g D^2 \gamma_{t+1}^2}{2}
\end{aligned} \tag{60}$$

By Lemma 4.1, we have $\|\nabla g(\mathbf{x}_t) - \widehat{\nabla} g_t\| \leq K_{1,t}$ and $\|g(x_t) - \hat{g}_t\| \leq K_{0,t}$ with probability $1-\delta^{'}$. Plug them in the inequality above to obtain,

$$\begin{aligned}
g(\mathbf{x}_{t+1}) - g(\mathbf{x}_0) &\leq (1-\gamma_{t+1})(g(\mathbf{x}_t)-g(\mathbf{x}_0)) + 2\gamma_{T+1}(DK_{1,t}+K_{0,t}) + \frac{L_g D^2\gamma_{t+1}^2}{2} \\
&\leq (1-\frac{1}{(T+1)^{2/3}})g(\mathbf{x}_t) - g(\mathbf{x}_0) \\
&+ \frac{2D(A_1^{2/3}\sqrt{\log(6d/\delta^{'})} + A_0^{2/3}\sqrt{\log(6d/\delta^{'})})}{(t+1)^{1/3}(T+1)^{2/3}} + \frac{L_g D^2}{2(T+1)^{4/3}}
\end{aligned} \tag{61}$$

with probability $1-\delta^{'}$ for all t. Let $C_3 = 2(DA_1^{2/3} + A_0^{2/3})$.Then we can sum all the inequality up for all t to obtain,

$$\begin{aligned}
g(\mathbf{x}_{t+1}) - g(\mathbf{x}_0) &\leq (1-\frac{1}{(T+1)^{2/3}})(g(\mathbf{x}_t)-g(\mathbf{x}_0)) + \frac{C_3\sqrt{\log(6d/\delta^{'})}}{(t+1)^{1/3}(T+1)^{2/3}} + \frac{L_g D^2}{2(T+1)^{4/3}} \\
&\leq (1-\frac{1}{(T+1)^{2/3}})(g(\mathbf{x}_t)-g(\mathbf{x}_0)) + \frac{C_3\sqrt{\log(6Td/\delta)} + L_g D^2/2}{(t+1)^{1/3}(T+1)^{2/3}}
\end{aligned} \tag{62}$$

By induction, we have for all $t \geq 1$,

$$g(\mathbf{x}_{t+1}) - g(\mathbf{x}_0) \leq \frac{C_3\sqrt{\log(6Td/\delta)} + L_g D^2/2}{(T+1)^{1/3}} \tag{63}$$

with probability $1-\delta$, where $\delta = T\delta^{'}$.

For **upper-level**, by Lemma B.2, we have

$$\gamma_{t+1}\mathcal{G}(\mathbf{x}_t) \leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \gamma_{t+1}D\|\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t\| + \frac{L_f\gamma_{t+1}^2 D^2}{2} \tag{64}$$

By Lemma 4.1, we have $\|\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t\| \leq \frac{A_2^{2/3}\sqrt{log(6d/\delta')}}{(t+1)^{1/3}}$ with probability $1-\delta'$. Plug it and $\gamma_{t+1} = 1/(T+1)^{2/3}$ in inequality above to obtain,

$$\begin{aligned}\sum_{t=0}^{T-1}\gamma_{t+1}\mathcal{G}(\mathbf{x}_t) &\leq f(\mathbf{x}_0) - f(\mathbf{x}_T) + D\sum_{t=0}^{T-1}\gamma_{t+1}\|\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t\| + \frac{L_f D^2}{2}\sum_{t=0}^{T-1}\gamma_{t+1}^2 \\ &\leq f(\mathbf{x}_0) - f(\mathbf{x}_T) + D\sum_{t=0}^{T-1}\frac{A_2^{2/3}\sqrt{log(6d/\delta')}}{(t+1)^{1/3}(T+1)^{2/3}} + \frac{L_f D^2}{2}\sum_{t=0}^{T-1}\frac{1}{(T+1)^{4/3}} \\ &\leq f(\mathbf{x}_0) - f(\mathbf{x}_T) + \frac{3}{2}DA_2^{2/3}\sqrt{log(6d/\delta')} + \frac{L_f D^2}{2}\frac{1}{(T+1)^{1/3}}\end{aligned} \tag{65}$$

Let $t^* = \arg\min_{1\leq t\leq T}\mathcal{G}(\mathbf{x}_t)$, then

$$\begin{aligned}\mathcal{G}(\mathbf{x}_{t^*}) &\leq \frac{1}{\sum_{t=0}^{T-1}\gamma_{t+1}}\sum_{t=0}^{T-1}\gamma_{t+1}\mathcal{G}(\mathbf{x}_t) \\ &\leq \frac{1}{(T+1)^{1/3}}(f(\mathbf{x}_0) - f(\mathbf{x}_T) + \frac{3}{2}DA_2^{2/3}\sqrt{log(6Td/\delta)} + \frac{L_f D^2}{2}\frac{1}{(T+1)^{1/3}}) \\ &\leq \frac{1}{(T+1)^{1/3}}(f(\mathbf{x}_0) - \underline{f} + \frac{3}{2}DA_2^{2/3}\sqrt{log(6Td/\delta)} + \frac{L_f D^2}{2}\frac{1}{(T+1)^{1/3}})\end{aligned} \tag{66}$$

with probability $1-\delta$, where $\delta = T\delta'$. By letting $C_4 = \frac{3}{2}DA_2^{2/3}$, the theorem is obtained. □

D Proof of Theorem for Algorithm 2

D.1 Proof of Theorem 4.6

Proof. For **lower-level** By Lemma B.2, we have

$$\begin{aligned}g(\mathbf{x}_{t+1}) - g(\mathbf{x}_0) \leq (1-\gamma_{t+1})(g(\mathbf{x}_t) - g(\mathbf{x}_0)) + D\gamma_{t+1}(\|\nabla g(\mathbf{x}_t) - \widehat{\nabla} g_t\| + K_{1,t}) \\ + \gamma_{t+1}(\|g(x_t) - \hat{g}_t\| + K_{0,t}) + \frac{L_g D^2\gamma_{t+1}^2}{2}\end{aligned} \tag{67}$$

By Lemma 4.3, we have $\|\nabla g(\mathbf{x}_t) - \widehat{\nabla} g_t\| \leq 4L_g D\gamma\sqrt{\log(12/\delta')}$ and $\|g(x_t) - \hat{g}_t\| \leq 4L_l D\gamma\sqrt{\log(12/\delta')}$ with probability $1-\delta'$. Let $C_5 = 8D(DL_g + L_l)$ and $\delta = T\delta'$. Plug them in inequality above and let $\gamma_t = \gamma = \log T/T$ to obtain,

$$g(\mathbf{x}_{T+1}) - g(\mathbf{x}_0) \leq (1-\gamma)(g(\mathbf{x}_T) - g(\mathbf{x}_0)) + (C_5\sqrt{\log(12/\delta')} + L_g D^2/2)\gamma^2 \tag{68}$$

with probability $1-\delta/T$. Sum up the inequalities for all $1 \leq t \leq T$ to get,

$$\begin{aligned}g(\mathbf{x}_{T+1}) - g(\mathbf{x}_0) &= (1-\gamma)^T(g(\mathbf{x}_0) - g(\mathbf{x}_0)) + (C_5\sqrt{\log(12/\delta')} + L_g D^2/2)\gamma^2\sum_{k=1}^{T}(1-\gamma)^k \\ &\leq 0 + (C_5\sqrt{\log(12/\delta')} + L_g D^2/2)\gamma \leq \frac{(C_5\sqrt{\log(12T/\delta)} + L_g D^2/2)\log T}{T}\end{aligned} \tag{69}$$

with probability $1-\delta$.

For **upper-level**, by Lemma B.2, we have,

$$f(\mathbf{x}_T) - f^* \leq (1-\gamma_T)f(\mathbf{x}_{T-1}) - f^* + D\gamma_T\|\nabla f(\mathbf{x}_{T-1}) - \widehat{\nabla} f_{t-1}\| + \frac{L_f D^2\gamma_T^2}{2} \tag{70}$$

Now we proceed by replacing the terms $\|\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t\|$ by its upper bounds from Lemma 4.3, i.e. $\|\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t\| \leq 4L_f D\gamma\sqrt{\log(12/\delta')}$,

$$f(\mathbf{x}_T) - f^* \leq (1-\gamma)(f(\mathbf{x}_{T-1}) - f^*) + L_f D^2\gamma^2(4\sqrt{\log(12/\delta')} + 1/2) \tag{71}$$

with probability $(1-\delta^{'})$. And we can choose $\delta = 3T\delta^{'}$ Then by telescope, with $\gamma = \frac{\log T}{T}$, we can obtain,

$$\begin{aligned} f(\mathbf{x}_T) - f^* &\leq (1-\gamma)^T(f(\mathbf{x}_0) - f^*) + (4\sqrt{\log(12/\delta^{'})} + 1/2)L_f D^2\gamma^2 \sum_{i=1}^{T}(1-\gamma)^i \\ &\leq (1-\gamma)^T(f(\mathbf{x}_0) - f^*) + (4\sqrt{\log(12/\delta^{'})} + 1/2)L_f D^2\gamma \\ &\leq \exp(-\gamma T)(f(\mathbf{x}_0) - f^*) + (4\sqrt{\log(12/\delta^{'})} + 1/2)L_f D^2\gamma \\ &\leq (f(\mathbf{x}_0) - f^*)/T + (4\sqrt{\log(12T/\delta)} + 1/2)L_f D^2 \log T/T \end{aligned} \tag{72}$$

with probability $1-\delta$. Note that without loss of generality, we can assume $f(\mathbf{x}_0) - f^* \geq 0$. If it is less than 0, we can bound it by 0. By letting $C_6 = 5L_f D^2$, the theorem is obtained.

□

D.2 Proof of Theorem 4.7

Proof. For **lower-level**, by Lemma B.2, we have

$$\begin{aligned} g(\mathbf{x}_{t+1}) - g(\mathbf{x}_0) \leq (1-\gamma_{t+1})(g(\mathbf{x}_t) - g(\mathbf{x}_0)) + D\gamma_{t+1}(\|\nabla g(\mathbf{x}_t) - \widehat{\nabla} g_t\| + K_{1,t}) \\ + \gamma_{t+1}(\|g(x_t) - \hat{g}_t\| + K_{0,t}) + \frac{L_g D^2 \gamma_{t+1}^2}{2} \end{aligned} \tag{73}$$

By Lemma 4.3, we have $\|\nabla g(\mathbf{x}_t) - \widehat{\nabla} g_t\| \leq 4L_g D\gamma\sqrt{\log(12/\delta^{'})}$ and $\|g(x_t) - \hat{g}_t\| \leq 4L_l D\gamma\sqrt{\log(12/\delta^{'})}$ with probability $1-\delta^{'}$. Let $C_7 = 8D(DL_g + L_l)$ and $\delta = T\delta^{'}$. Plug them in inequality above and let $\gamma_t = 1/\sqrt{T}$ to obtain,

$$\begin{aligned} g(\mathbf{x}_{t+1}) - g(\mathbf{x}_0) &\leq (1 - \frac{1}{T^{1/2}})(g(\mathbf{x}_t) - g(\mathbf{x}_0)) + \frac{C_7\sqrt{\log(12/\delta^{'})}}{T} + \frac{L_g D^2}{2T} \\ &\leq (1 - \frac{1}{T^{1/2}})(g(\mathbf{x}_t) - g(\mathbf{x}_0)) + \frac{C_7\sqrt{\log(12/\delta^{'})} + L_g D^2/2}{T} \end{aligned} \tag{74}$$

with probability $1-\delta/T$. Sum up the inequalities for all $t \geq 1$ to get,

$$\begin{aligned} g(\mathbf{x}_{t+1}) - g(\mathbf{x}_0) &= (1 - \frac{1}{T^{1/2}})^t \mathbb{E}[g(\mathbf{x}_0) - g(\mathbf{x}_0)] + \frac{(C_7\sqrt{\log(12/\delta^{'})} + L_g D^2/2)}{T} \sum_{k=1}^{t}(1 - \frac{1}{T^{1/2}})^k \\ &\leq \frac{C_7\sqrt{\log(12T/\delta)} + L_g D^2/2}{T^{1/2}} \end{aligned} \tag{75}$$

with probability $1-\delta$.

For **upper-level**, by Lemma B.2, we have

$$\gamma_{t+1}\mathcal{G}(\mathbf{x}_t) \leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \gamma_{t+1}D\|\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t\| + \frac{L_f \gamma_{t+1}^2 D^2}{2} \tag{76}$$

By Lemma 4.3, we have $\|\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t\| \leq 4L_f D\gamma\sqrt{\log(12/\delta^{'})}$ with probability $1-\delta^{'}$. Plug it and $\gamma_{t+1} = 1/\sqrt{T}$ in inequality above to obtain,

$$\begin{aligned} \frac{1}{\sqrt{T}}\sum_{t=0}^{T-1}\mathcal{G}(\mathbf{x}_t) &\leq f(\mathbf{x}_0) - f(\mathbf{x}_T) + D\sum_{t=0}^{T-1}\gamma_{t+1}\|\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t\| + \frac{L_f D^2}{2}\sum_{t=0}^{T-1}\gamma_{t+1}^2 \\ &\leq f(\mathbf{x}_0) - f(\mathbf{x}_T) + L_f D^2(4\sqrt{\log(12\delta^{'})} + 1/2) \end{aligned} \tag{77}$$

Divide both sides by $\sqrt{T}$, we can get, Let $\mathbf{x}_o = \arg\min_{1\leq t\leq T}\mathcal{G}(\mathbf{x}_t)$, then

$$\mathcal{G}(\mathbf{x}_o) \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathcal{G}(\mathbf{x}_t) \leq \frac{f(\mathbf{x}_0) - \underline{f} + L_f D^2(4\sqrt{\log(12T/\delta)} + 1/2)}{T^{1/2}} \tag{78}$$

with probability $1-\delta$. By letting $C_8 = 5L_f D^2$, the theorem is obtained. □

E Azuma-Hoeffding-type inequalities

In this section, we present two useful vector versions of Azuma-Hoeffding-type concentration inequalities with uniform bound assumption or sub-gaussian assumption. They are crucial in our high probability analysis.

Proposition E.1. *(Pinelis and other 1994 [40], Theorem 3.5) Let $\zeta_1, \ldots, \zeta_t \in \mathbb{R}^d$ be a vector-valued martingale difference sequence w.r.t. a filtration $\{\mathcal{F}_t\}$, i.e. for each $\tau \in 1, \ldots, t$, we have $\mathbb{E}[\zeta_\tau | \mathcal{F}_{\tau-1}] = 0$. Suppose that $\|\zeta_\tau\| \leq c_\tau$ almost surely. Then $\forall t \geq 1$,*

$$P(\|\sum_{\tau=1}^{T} \zeta_\tau\| \geq \lambda) \leq 4\exp(-\frac{\lambda^2}{4\sum_{\tau=1}^{T} c_\tau^2}) \tag{79}$$

Proposition E.2. *(Jin et al. [41], Corollary 7) Let $\zeta_1, \ldots, \zeta_t \in \mathbb{R}^d$ be a vector-valued martingale difference sequence w.r.t. a filtration $\{\mathcal{F}_t\}$, i.e. for each $\tau \in 1, \ldots, t$, we have $\mathbb{E}[\zeta_\tau | \mathcal{F}_{\tau-1}] = 0$. Suppose that $\mathbb{E}[\exp(\|\zeta_\tau\|^2/c_\tau^2)] \leq \exp(1)$. Then there exists a absolute constant c such that, for any $\delta > 0$, with probability at least $1-\delta$,*

$$\|\sum_{\tau=1}^{T} \zeta_\tau\| \leq c \cdot \sqrt{\sum_{\tau=1}^{T} c_\tau^2 \log\frac{2d}{\delta}} \tag{80}$$

This proposition was also used in previous literature including [42] and [33]. It is common to use such martingale inequality to obtain some high-probability results recently.

F Experiment details

In this section, we include more details about the numerical experiments in Section 5. For completeness, we briefly introduce the update rules of aR-IP-SeG in [16] and DBGD in [13]. In the following, we use the notation $\Pi_{\mathcal{Z}}(\cdot)$ to denote the Euclidean projection onto the set $\mathcal{Z}$.
The aR-IP-SeG algorithm is given by,

$$\begin{aligned}
\mathbf{y}_{t+1} &= \Pi_{\mathcal{Z}}(\mathbf{x}_t - \gamma_t(\nabla \tilde{f}(\mathbf{x}_t, \theta_t)) + \rho_t \nabla \tilde{g}(\mathbf{x}_t, \xi_t)) \\
\mathbf{x}_{t+1} &= \Pi_{\mathcal{Z}}(\mathbf{x}_t - \gamma_t(\nabla \tilde{f}(\mathbf{y}_t, \theta_t^{'})) + \rho_t \nabla \tilde{g}(\mathbf{y}_t, \xi_t^{'})) \\
\Gamma_{t+1} &= \Gamma_t + (\gamma_t \rho_t)^r \\
\bar{\mathbf{y}}_{t+1} &= \frac{\Gamma_t \bar{\mathbf{y}}_t + (\gamma_t \rho_t)^r \mathbf{y}_{t+1}}{\Gamma_{t+1}}
\end{aligned} \tag{81}$$

where γ_t is the stepsize, ρ_t is the regularization parameter, and $\bar{\mathbf{y}}_T$ is the output of the algorithm. In this experiment, we choose $\gamma_t = \gamma_0/(t+1)^{3/4}$ and $\rho_t = \rho_0(t+1)^{1/4}$ for some constants γ_0 and ρ_0. The DBGD-sto is a stochastic version of DBGD, which simply replaces the gradients in DBGD with stochastic gradients. Although the stochastic version of DBGD does not have a theoretical guarantee, it has been used to solve stochastic simple bilevel optimization problems in [13], which worked pretty well empirically. Hence, we use it as a baseline for solving stochastic simple bilevel problems and compare it with our proposed algorithms. The DBGD algorithm is given by

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \gamma_k \left(\nabla f\left(\mathbf{x}_k\right) + \lambda_k \nabla g\left(\mathbf{x}_k\right)\right)$$

where γ_k is the stepsize and we set λ_k as

$$\lambda_k = \max\left\{\frac{\phi\left(\mathbf{x}_k\right) - \left\langle \nabla f\left(\mathbf{x}_k\right), \nabla g\left(\mathbf{x}_k\right)\right\rangle}{\left\|\nabla g\left(\mathbf{x}_k\right)\right\|^2}, 0\right\} \quad \text{and} \quad \phi(\mathbf{x}) = \min\left\{\alpha(g(\mathbf{x}) - \hat{g}), \beta\|\nabla g(\mathbf{x})\|^2\right\}$$

where α and β are hyperparameters and $\hat{g}$ is a lower bound of g^*. In this experiment, we choose $\hat{g} = 0$. We also note that [13] only considered unconstrained simple bilevel optimization, i.e. $\mathcal{Z} = \mathbb{R}^d$. We further project $\mathbf{x}_t$ onto $\mathcal{Z}$ for each iteration to ensure the constraints are satisfied.

F.1 Over-parameterized regression

Dataset generation. The original Wikipedia Math Essential dataset [26] composes of a data matrix of size 1068×731. We randomly select one of the columns as the outcome vector $\mathbf{b} \in \mathbb{R}^{1068}$ and the rest to be a new matrix $\mathbf{A} \in \mathbb{R}^{1068 \times 731}$. We set constraint parameter $\lambda = 10$ in this experiment.
Initialization. We run the algorithm, SPIDER-FW [36], with stepsize chosen as $\gamma_t = 0.1/(t+1)$ on the lower-level problem in (1). We terminate the process to get $\mathbf{x}_0$ as the initial point for both SBCGI 1 and SBCGF 2 after 10^5 stochastic oracle queries.
Implementation details. We query stochastic oracle 9×10^5 times with stepsize $\gamma_t = 0.01/(t+1)$ and $\gamma = 10^{-5}$ for SBCGI 1 and SBCGF 2 with $K_t = 10^{-4}/\sqrt{t+1}$, respectively. In each iteration, we need to solve the following subproblem induced by the methods,

$$\min_{\mathbf{s}} \langle \nabla f(\boldsymbol{\beta}_k), \mathbf{s} \rangle \quad \text{s.t.} \quad \|\mathbf{s}\|_1 \leq \lambda, \langle \nabla g(\boldsymbol{\beta}_k), \mathbf{s} - \boldsymbol{\beta}_k \rangle \leq g(\boldsymbol{\beta}_0) - g(\boldsymbol{\beta}_k). \tag{82}$$

Introduce $\mathbf{s}^+, \mathbf{s}^- \geq 0$ such that $\mathbf{s} = \mathbf{s}^+ - \mathbf{s}^-$. Then we can reformulate the problem above as follows,

$$\begin{aligned} &\min_{\mathbf{s}^+, \mathbf{s}^-} \left\langle \nabla f(\boldsymbol{\beta}_k), \mathbf{s}^+ - \mathbf{s}^- \right\rangle \\ &\text{s.t. } \mathbf{s}^+, \mathbf{s}^- \geq 0, \left\langle \mathbf{s}^+, \mathbf{1} \right\rangle + \left\langle \mathbf{s}^-, \mathbf{1} \right\rangle \leq \lambda, \left\langle \nabla g(\boldsymbol{\beta}_k), \mathbf{s}^+ - \mathbf{s}^- - \boldsymbol{\beta}_k \right\rangle \leq g(\boldsymbol{\beta}_0) - g(\boldsymbol{\beta}_k), \end{aligned} \tag{83}$$

where $1 \in \mathbb{R}^d$ is the all-one vector.
For aR-IP-SeG, we choose $\gamma_0 = 10^{-7}$ and $\rho_0 = 10^3$. For DBGD, we set $\alpha = \beta = 1$ and $\gamma_t = 10^{-6}$.

F.2 Dictionary learning

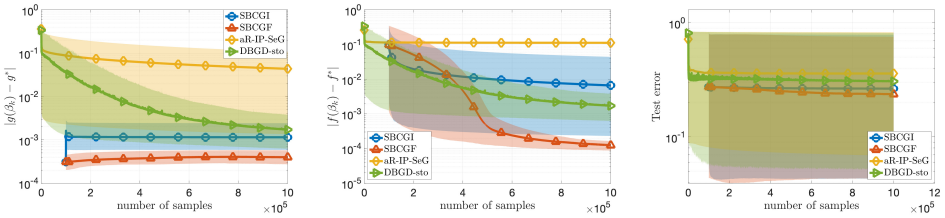
Dataset generation. We generate 500 sparse coefficient vectors $\{\mathbf{x}_i\}_{i=1}^{250}$ and $\{\mathbf{x}'_k\}_{k=1}^{250}$ with 5 random nonzero entries, whose absolute values are drawn uniformly from $[0.2, 1]$. The entries of the random noise vectors $\{\mathbf{n}_i\}_{i=1}^{250}$ and $\{\mathbf{n}'_k\}_{k=1}^{250}$ are drawn from i.i.d. Gaussian distribution with mean 0 and standard deviation 0.01.
Initialization. We use a similar initialization procedure as [12], which consists of two phases. In the first phase, we run the standard Frank-Wolfe algorithm on both the variables $\mathbf{D} \in \mathbb{R}^{25 \times 40}$ and $\mathbf{X} \in \mathbb{R}^{40 \times 250}$ for 10^4 iterations with the stepsize $\gamma_t = 1/\sqrt{t+1}$. Next, in the second phase, we fix the variable $\mathbf{X}$ and only update $\mathbf{D}$ using the Frank-Wolfe algorithm with exact line search for additional 10^4 iterations to obtain $\hat{\mathbf{D}}$ and $\hat{\mathbf{X}}$ as the initial point for the full bilevel problem.
Implementation Details. We choose $\delta = 3$ in both problems (5). To be fair, all four algorithms start from the same initial point. We slightly modify the initial point by letting $\tilde{\mathbf{D}} \in \mathbb{R}^{25 \times 50}$ be the concatenation of $\hat{\mathbf{D}} \in \mathbb{R}^{25 \times 40}$ and 10 columns of all zeros vectors. Furthermore, we initialize another variable $\tilde{\mathbf{X}}$ randomly by choosing its entries from a standard Gaussian distribution and then normalizing each column to have a ℓ_1-norm of δ. We choose the stepsize as $\gamma_t = 0.1/(t+1)^{2/3}$ and $\gamma = 10^{-3}$ for our SBCGI 1 and SBCGF 2 with $K_t = 0.01/(t+1)^{1/3}$, respectively. Empirically, we observe that taking one sample per iteration leads to a very unstable process in this problem. In this case, we choose a mini-batch of size 8 for SBCGI, aR-IP-SeG, and the stochastic version of DBGD. For each iteration, we will solve the following subproblem,

$$\min_{\tilde{\mathbf{D}}} \left\langle \nabla f_{\tilde{\mathbf{D}}}\left(\tilde{\mathbf{D}}_k, \tilde{\mathbf{X}}_k\right), \tilde{\mathbf{D}} \right\rangle \quad \text{s.t.} \quad \left\|\tilde{\mathbf{d}}_i\right\|_2 \leq 1, \left\langle \nabla g\left(\tilde{\mathbf{D}}_k\right), \tilde{\mathbf{D}} - \tilde{\mathbf{D}}_k \right\rangle \leq g\left(\tilde{\mathbf{D}}_0\right) - g\left(\tilde{\mathbf{D}}_k\right) \tag{84}$$

The above problem can be reformulated by using the KKT condition, which is equivalent to get a root of the following one-dimensional nonlinear equation involving $\lambda \geq 0$:

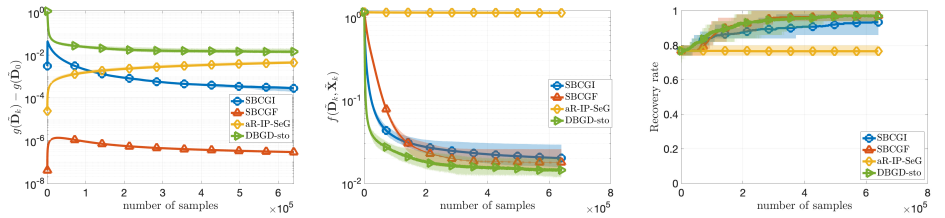
$$\tilde{\mathbf{D}} = \Pi_{\mathcal{Z}}\left(\nabla f_{\tilde{\mathbf{D}}}\left(\tilde{\mathbf{D}}_k, \tilde{\mathbf{X}}_k\right) + \lambda \nabla g\left(\tilde{\mathbf{D}}_k\right)\right), \quad \left\langle \nabla g\left(\tilde{\mathbf{D}}_k\right), \tilde{\mathbf{D}} - \tilde{\mathbf{D}}_k \right\rangle = g\left(\tilde{\mathbf{D}}_0\right) - g\left(\tilde{\mathbf{D}}_k\right) \tag{85}$$

where the projection onto $\mathcal{Z} = \left\{\tilde{\mathbf{D}} \in \mathbb{R}^{25 \times 40} : \left\|\tilde{\mathbf{d}}_i\right\|_2 \leq 1, i = 1, \ldots, 40\right\}$ is equivalent to project each column on the Euclidean ball. In practice, the reformulated problem can be solved efficiently by MATLAB's root-finding solver.
For aR-IP-SeG, we choose $\gamma_0 = 10^{-4}$ and $\rho_0 = 1$. For the stochastic version of DBGD, we set $\alpha = \beta = 100$ and $\gamma_t = 5 \times 10^{-3}$.



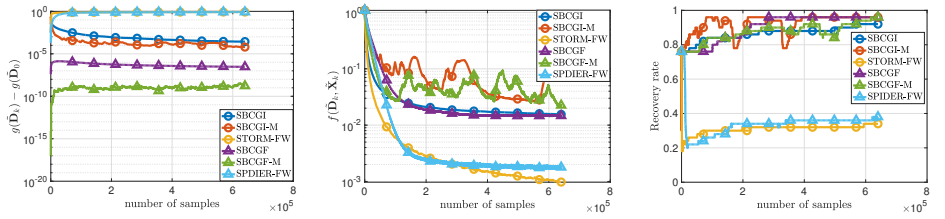
(a) Lower-level gap (b) Upper-level gap (c) Test error

Figure 3: Comparison of SBCGI, SBCGF, aR-IP-SeG, and DBGD-Sto for solving Problem (3) with 10 different random seeds



(a) Lower-level gap (b) Upper-level gap (c) Recovery rate

Figure 4: Comparison of SBCGI, SBCGF, aR-IP-SeG, and DBGD-Sto for solving Problem (5) with 10 different random seeds



(a) Lower-level gap (b) Upper-level gap (c) Recovery rate

Figure 5: Comparison of SBCGI, SBCGF, SBCGI-M, SBCGF-M, STORM-FW, and SPIDER-FW for solving Problem (5).

F.3 Experiments with different random seeds

We further repeat the experiment 10 times with different random seeds to see more realizations of the stochastic algorithms. The results are reported in Figure 3 and Figure 4. The solid lines denote the average statistics over 10 trials of the algorithms. While the shaded regions surrounding each line reflect the span of all the random instances involved. Figure 3 and Figure 4 present similar results as Figure 1 and Figure 2, which eliminates the possibility of choosing a particularly good instance.

F.4 Importance of the right cutting plane

In this section, we numerically illustrate the importance of choosing the right cutting plane on *Example 2 (dictionary learning)*. Specifically, we compare our proposed methods with the ones without a cutting plane and with an unregularized cutting plane (without additional term K_t).
If we replace the stochastic cutting plane (9) with the unregularized cutting plane (8) in SBCGI 1 and SBCGF 2, then the algorithm usually fail at some point in the process, depending on the datasets and parameters chosen, based on our experimental observations. More specifically, algorithms' failure means that the subproblem of dictionary learning (85) is infeasible. So we slightly modify it by adding a checkpoint before solving the subproblem. If the subproblem is infeasible at the current

iteration, then we choose the update direction $\mathbf{s}_t = \widehat{\nabla} g_t$. This adjustment prevents unnecessary interruptions during the process and enforce the algorithms to focus only on the lower-level problem when the subproblem is infeasible. We denote the modified algorithms SBCGI-M and SBCGF-M. Moreover, we also take SBCGI and SBCGF without cutting planes into consideration, denoted as STORM-FW and SPIDER-FW. In fact, in this case, the bilevel algorithms degenerate to single-level projection-free algorithms similar to algorithms in [33] and [36].

Figure 5 (a) indicates that SBCGI-M and SBCGF-M focus more on the lower-level problem due to the design of the algorithms and extremely unstable as we can see in Figure 5 (b)(c). While STORM-FW and SPIDER-FW only focus on the upper-level problem, which leads to terrible results on the lower-level gap and recovery rate.