

807 Appendix / supplemental material

808 A Omitted Proofs in Section 3

809 A.1 Proof of Lemma 3.1

810 First, we show that if the point $\hat{\mathbf{x}}$ is an (ϵ_f, ϵ_g) -stationary point as defined in Definition 3.1, then the
 811 two conditions in Lemma 3.1 are satisfied. For any $\delta > 0$, let $r = \min\{2\delta\sqrt{\epsilon_g}/L_g, 2\delta\sqrt{\epsilon_f}/(\lambda L_g +$
 812 $L_f)\}$. For any \mathbf{x} satisfying $\|\mathbf{x} - \hat{\mathbf{x}}\| \leq r$, Using the fact that g is L_g -smooth and $\|\nabla g(\hat{\mathbf{x}})\|^2 \leq \epsilon_g$, it
 813 holds that

$$\begin{aligned} g(\mathbf{x}) &\geq g(\hat{\mathbf{x}}) + \langle \nabla g(\hat{\mathbf{x}}), \mathbf{x} - \hat{\mathbf{x}} \rangle - \frac{L_g}{2} \|\hat{\mathbf{x}} - \mathbf{x}\|^2 \\ &\geq g(\hat{\mathbf{x}}) - \sqrt{\epsilon_g} \|\mathbf{x} - \hat{\mathbf{x}}\| - \delta \sqrt{\epsilon_g} \|\hat{\mathbf{x}} - \mathbf{x}\| = g(\hat{\mathbf{x}}) - (1 + \delta) \sqrt{\epsilon_g} \|\mathbf{x} - \hat{\mathbf{x}}\|, \end{aligned}$$

814 where we used $\|\mathbf{x} - \hat{\mathbf{x}}\| \leq r \leq 2\delta\sqrt{\epsilon_g}/L_g$ in the second inequality. Thus, the first condition in
 815 Lemma 3.1 is satisfied. Moreover, Consider any \mathbf{x} that satisfies $\|\mathbf{x} - \hat{\mathbf{x}}\| \leq r$ and $g(\mathbf{x}) \leq g(\hat{\mathbf{x}})$.
 816 Since f is L_f -smooth, it holds that $f(\mathbf{x}) \geq f(\hat{\mathbf{x}}) + \langle \nabla f(\hat{\mathbf{x}}), \mathbf{x} - \hat{\mathbf{x}} \rangle - \frac{L_f}{2} \|\hat{\mathbf{x}} - \mathbf{x}\|^2$. By using
 817 $\|\nabla f(\hat{\mathbf{x}}) + \lambda \nabla g(\hat{\mathbf{x}})\| \leq \sqrt{\epsilon_f}$, we further have

$$\begin{aligned} f(\mathbf{x}) &\geq f(\hat{\mathbf{x}}) + \langle \nabla f(\hat{\mathbf{x}}) + \lambda \nabla g(\hat{\mathbf{x}}), \mathbf{x} - \hat{\mathbf{x}} \rangle - \lambda \langle \nabla g(\hat{\mathbf{x}}), \mathbf{x} - \hat{\mathbf{x}} \rangle - \frac{L_f}{2} \|\hat{\mathbf{x}} - \mathbf{x}\|^2 \\ &\geq f(\hat{\mathbf{x}}) - \sqrt{\epsilon_f} \|\mathbf{x} - \hat{\mathbf{x}}\| - \lambda \langle \nabla g(\hat{\mathbf{x}}), \mathbf{x} - \hat{\mathbf{x}} \rangle - \frac{L_f}{2} \|\hat{\mathbf{x}} - \mathbf{x}\|^2. \end{aligned}$$

818 Using the smoothness of g , we also have $g(\mathbf{x}) \geq g(\hat{\mathbf{x}}) + \langle \nabla g(\hat{\mathbf{x}}), \mathbf{x} - \hat{\mathbf{x}} \rangle - \frac{L_g}{2} \|\mathbf{x} - \hat{\mathbf{x}}\|^2$. Hence, we
 819 get $-\langle \nabla g(\hat{\mathbf{x}}), \mathbf{x} - \hat{\mathbf{x}} \rangle \geq -\frac{L_g}{2} \|\mathbf{x} - \hat{\mathbf{x}}\|^2$. Thus, this leads to

$$f(\mathbf{x}) \geq f(\hat{\mathbf{x}}) - \sqrt{\epsilon_f} \|\mathbf{x} - \hat{\mathbf{x}}\| - \lambda \frac{L_g}{2} \|\hat{\mathbf{x}} - \mathbf{x}\|^2 - \frac{L_f}{2} \|\hat{\mathbf{x}} - \mathbf{x}\|^2.$$

820 Since $\|\hat{\mathbf{x}} - \mathbf{x}\| \leq r \leq 2\delta\sqrt{\epsilon_f}/(\lambda L_g + L_f)$, we obtain $f(\mathbf{x}) \geq f(\hat{\mathbf{x}}) - (1 + \delta) \sqrt{\epsilon_f} \|\mathbf{x} - \hat{\mathbf{x}}\|$. This
 821 shows that the second condition in Lemma 3.1 is also satisfied.

822 For the other direction, assume that $\hat{\mathbf{x}}$ satisfies both conditions in Lemma 3.1. Consider any direction
 823 $\mathbf{d} \in \mathbb{R}^n$. Then Condition (i) implies that, for all t small enough, we have $g(\hat{\mathbf{x}} - t\mathbf{d}) \geq g(\hat{\mathbf{x}}) - (1 +$
 824 $\delta)\epsilon_g t \|\mathbf{d}\|$, which can be rewritten as $\frac{g(\hat{\mathbf{x}}) - g(\hat{\mathbf{x}} - t\mathbf{d})}{t} \leq (1 + \delta)\epsilon_g \|\mathbf{d}\|$. By taking the limit $t \rightarrow 0$, we
 825 obtain $\langle \nabla g(\hat{\mathbf{x}}), \mathbf{d} \rangle \leq (1 + \delta)\epsilon_g \|\mathbf{d}\|$. By taking $\mathbf{d} = \nabla g(\hat{\mathbf{x}})$, this implies that $\|\nabla g(\hat{\mathbf{x}})\| \leq (1 + \delta)\epsilon_g$.
 826 Since this holds for any $\delta > 0$, taking the limit $\delta \rightarrow 0$ yields $\|\nabla g(\hat{\mathbf{x}})\| \leq \epsilon_g$. Moreover, let
 827 $\mathbf{d} \in \mathbb{R}^n$ be any direction that satisfies $\langle \nabla g(\hat{\mathbf{x}}), \mathbf{d} \rangle > 0$. Then for all t small enough, it holds that
 828 $g(\hat{\mathbf{x}} - t\mathbf{d}) \leq g(\hat{\mathbf{x}})$. Thus, using Condition (ii), we have

$$f(\hat{\mathbf{x}} - t\mathbf{d}) \geq f(\hat{\mathbf{x}}) - (1 + \delta)\epsilon_f t \|\mathbf{d}\| \quad \Rightarrow \quad \frac{f(\hat{\mathbf{x}}) - f(\hat{\mathbf{x}} - t\mathbf{d})}{t} \leq (1 + \delta)\epsilon_f \|\mathbf{d}\|.$$

829 Similarly, by taking the limits $t \rightarrow 0$ and $\delta \rightarrow 0$, we obtain $\langle \nabla f(\hat{\mathbf{x}}), \mathbf{d} \rangle \leq \epsilon_f \|\mathbf{d}\|$. Since this holds
 830 for any \mathbf{d} that satisfies $\langle \nabla g(\hat{\mathbf{x}}), \mathbf{d} \rangle > 0$, continuity ensures that it also holds for any \mathbf{d} such that
 831 $\langle \nabla g(\hat{\mathbf{x}}), \mathbf{d} \rangle \geq 0$. If $\langle \nabla f(\hat{\mathbf{x}}), \nabla g(\hat{\mathbf{x}}) \rangle \geq 0$, then by setting $\mathbf{d} = \nabla f(\hat{\mathbf{x}})$, we obtain that $\|\nabla f(\hat{\mathbf{x}})\| \leq \epsilon_f$.
 832 Otherwise, if $\langle \nabla f(\hat{\mathbf{x}}), \nabla g(\hat{\mathbf{x}}) \rangle < 0$, let $\lambda = -\frac{\langle \nabla f(\hat{\mathbf{x}}), \nabla g(\hat{\mathbf{x}}) \rangle}{\|\nabla g(\hat{\mathbf{x}})\|^2} > 0$ and set $\mathbf{d} = \nabla f(\hat{\mathbf{x}}) + \lambda \nabla g(\hat{\mathbf{x}})$.
 833 Note that this choice of λ ensures that $\langle \nabla g(\hat{\mathbf{x}}), \mathbf{d} \rangle = 0$, and hence $\langle \nabla f(\hat{\mathbf{x}}), \mathbf{d} \rangle = \|\mathbf{d}\|^2 \leq \epsilon_f \|\mathbf{d}\|$,
 834 which implies that $\|\mathbf{d}\| \leq \epsilon_f$. This completes the proof.

835 A.2 Proof of Theorem 3.2

836 Suppose $\hat{\mathbf{x}}$ is an (ϵ_f, ϵ_g) -stationary point of Problem (I), the second inequality in Definition 3.2
 837 is satisfied with $\epsilon_d = \epsilon_g$. Now, we start to prove the first inequality in Definition 3.2 by setting
 838 $\mathbf{w} = \lambda(\hat{\mathbf{x}} - \mathbf{x}^*)$, where \mathbf{x}^* denotes the stationary point closest to $\hat{\mathbf{x}}$.

$$\begin{aligned}
\|\nabla f(\hat{\mathbf{x}}) + \nabla^2 g(\hat{\mathbf{x}})\mathbf{w}\| &\leq \|\nabla f(\hat{\mathbf{x}}) + \nabla^2 g(\mathbf{x}^*)\mathbf{w}\| + \|\nabla^2 g(\hat{\mathbf{x}}) - \nabla^2 g(\mathbf{x}^*)\|\|\mathbf{w}\| \\
&\leq \|\nabla f(\hat{\mathbf{x}}) + \lambda \nabla g(\hat{\mathbf{x}})\| + \lambda \|\nabla g(\hat{\mathbf{x}}) - \nabla^2 g(\mathbf{x}^*)(\hat{\mathbf{x}} - \mathbf{x}^*)\| + \lambda L_H \|\hat{\mathbf{x}} - \mathbf{x}^*\|^2 \\
&\leq \epsilon_f + \lambda \|\nabla g(\hat{\mathbf{x}}) - \nabla g(\mathbf{x}^*) - \nabla^2 g(\mathbf{x}^*)(\hat{\mathbf{x}} - \mathbf{x}^*)\| + \lambda L_H \|\hat{\mathbf{x}} - \mathbf{x}^*\|^2 \\
&\leq \epsilon_f + 2\lambda L_H \|\hat{\mathbf{x}} - \mathbf{x}^*\|^2 \leq \epsilon_f + \lambda \|\nabla g(\hat{\mathbf{x}})\| \cdot 2L_H c^2 \epsilon_g \\
&= \mathcal{O}(\epsilon_f + \lambda \|\nabla g(\hat{\mathbf{x}})\| \epsilon_g)
\end{aligned}$$

where the second and fourth inequalities follow from the Lipschitz continuity of $\nabla^2 g(\mathbf{x})$, the third follows from the second condition in Definition 3.1 and the last follows from Assumption 3.1. Hence, the first condition in Definition 3.2 holds with $\epsilon_p = \mathcal{O}(\epsilon_f + \lambda \|\nabla g(\hat{\mathbf{x}})\| \epsilon_g)$.

B Omitted Proofs in Section 5

B.1 Proof of Lemma 5.1

From Assumption 2.1, f has an L_f -Lipschitz continuous gradient, hence,

$$\begin{aligned}
f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq -\eta \nabla f(\mathbf{x}_k)^\top \mathbf{d}_k + \frac{L_f}{2} \eta^2 \|\mathbf{d}_k\|^2 \\
&= -\eta (\nabla f(\mathbf{x}_k) - \mathbf{d}_k)^\top \mathbf{d}_k - \eta \left(1 - \frac{L_f}{2} \eta\right) \|\mathbf{d}_k\|^2 \\
&= \eta \lambda_k \nabla g(\mathbf{x}_k)^\top \mathbf{d}_k - \eta \left(1 - \frac{L_f}{2} \eta\right) \|\mathbf{d}_k\|^2
\end{aligned}$$

where in the last equality we used $\nabla f(\mathbf{x}_k) = \mathbf{d}_k - \lambda_k \nabla g(\mathbf{x}_k)$. Since \mathbf{d}_k is the optimal solution of subproblem (12) with the corresponding optimal dual multiplier λ_k , the complementarity slackness implies that $\lambda_k (\nabla g(\mathbf{x}_k)^\top \mathbf{d}_k - \beta \|\nabla g(\mathbf{x}_k)\|^2) = 0$. Hence, we further obtain

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\eta \left(1 - \frac{L_f}{2} \eta\right) \|\mathbf{d}_k\|^2 + \eta \lambda_k \beta \|\nabla g(\mathbf{x}_k)\|^2.$$

By dividing both sides by η and rearranging the inequality, we obtain (17).

Moreover, from Assumption 2.1, g has an L_g -Lipschitz continuous gradient, which implies that

$$g(\mathbf{x}_{k+1}) - g(\mathbf{x}_k) \leq -\eta \nabla g(\mathbf{x}_k)^\top \mathbf{d}_k + \frac{L_g}{2} \eta^2 \|\mathbf{d}_k\|^2 \leq -\eta \beta \|\nabla g(\mathbf{x}_k)\|^2 + \frac{L_g}{2} \eta^2 \|\mathbf{d}_k\|^2,$$

where we used $\nabla g(\mathbf{x}_k)^\top \mathbf{d}_k \geq \beta \|\nabla g(\mathbf{x}_k)\|^2$ from (12) in the last inequality. Dividing both sides by η and rearranging the inequality yields (18).

B.2 Proof of Lemma 5.2

By Assumption 2.1, the gradient of f is bounded by G_f . Thus, we have

$$\lambda_k \leq \beta + \frac{|\langle \nabla f(\mathbf{x}_k), \nabla g(\mathbf{x}_k) \rangle|}{\|\nabla g(\mathbf{x}_k)\|^2} \leq \beta + \frac{G_f}{\|\nabla g(\mathbf{x}_k)\|}.$$

This completes the proof.

B.3 Proof of Lemma 5.3

By combining Lemma 5.2 with (17), we have $(1 - \frac{\eta L_f}{2}) \|\mathbf{d}_k\|^2 \leq \frac{\Delta f_k}{\eta} + \beta^2 \|\nabla g(\mathbf{x}_k)\|^2 + \beta G_f \|\nabla g(\mathbf{x}_k)\|$. Substituting the upper bound on $\|\nabla g(\mathbf{x}_k)\|$ in (18) and combining terms, we arrive at $(1 - \frac{\eta(L_f + \beta L_g)}{2}) \|\mathbf{d}_k\|^2 \leq \frac{\Delta f_k + \beta \Delta g_k}{\eta} + \sqrt{\beta} G_f \sqrt{\frac{\Delta g_k}{\eta}} + \frac{L_g}{2} \eta \|\mathbf{d}_k\|^2$. Since $\eta \leq \frac{1}{L_f + L_g} \leq \frac{1}{L_f + \beta L_g}$, the left side of this inequality can be lower bounded by $\frac{1}{2} \|\mathbf{d}_k\|^2$. By multiplying both sides by 2 the claim follows.

861 B.4 Proof of Lemma 5.4

862 Since $x \leq A + B\sqrt{x}$, we have $(\sqrt{x} - \frac{B}{2})^2 \leq A + \frac{B^2}{4}$, which further implies $\sqrt{x} - \frac{B}{2} \leq \sqrt{A + \frac{B^2}{4}}$. By
 863 adding $\frac{B}{2}$ to both sides, taking the square, and using Young's inequality we obtain $x \leq (\sqrt{A + \frac{B^2}{4}} +$
 864 $\frac{B}{2})^2 = A + \frac{B^2}{2} + B\sqrt{A + \frac{B^2}{4}} \leq A + \frac{B^2}{2} + \frac{B^2}{4} + (A + \frac{B^2}{4}) = 2A + B^2$. This completes the proof.

865 B.5 Proof of Theorem 5.5

866 Multiplying (18) by $\frac{1}{L_g\eta}$ and adding it to (21), implies

$$\frac{\|\mathbf{d}_k\|^2}{2} + \frac{\beta\|\nabla g(\mathbf{x}_k)\|^2}{L_g\eta} \leq \frac{4(\Delta f_k + \beta\Delta g_k)}{\eta} + \frac{3\Delta g_k}{L_g\eta^2} + 2\beta G_f^2 L_g\eta.$$

867 Define the potential function as $\mathcal{G}_k \triangleq \frac{1}{2}\|\mathbf{d}_k\|^2 + \frac{\beta}{L_g\eta}\|\nabla g(\mathbf{x}_k)\|^2$. Averaging the above inequality
 868 over $k = 0$ to $K - 1$ and noting that $\sum_{k=0}^{K-1} \Delta f_k = f(\mathbf{x}_0) - f(\mathbf{x}_K) \leq f(\mathbf{x}_0) - \inf f = \Delta_f$ and
 869 $\sum_{k=0}^{K-1} \Delta g_k = g(\mathbf{x}_0) - g(\mathbf{x}_K) \leq g(\mathbf{x}_0) - g^* = \Delta_g$, we obtain

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathcal{G}_k \leq \frac{4(\Delta_f + \beta\Delta_g)}{\eta K} + \frac{3\Delta_g}{L_g\eta^2 K} + 2\beta G_f^2 L_g\eta.$$

870 Since $\eta = \frac{1}{LK^{1/(3+p)}}$ and $\beta = \frac{1}{K^{p/(3+p)}}$, by letting $k^* = \operatorname{argmin}_{0 \leq k \leq K-1} \mathcal{G}_k$, we get

$$\mathcal{G}_{k^*} \leq \frac{4L\Delta_f}{K^{(2+p)/(3+p)}} + \frac{4L\Delta_g}{K^{(2+2p)/(3+p)}} + \frac{3L^2\Delta_g}{L_g K^{(1+p)/(3+p)}} + \frac{2G_f^2}{K^{(1+p)/(3+p)}}.$$

871 Finally, since $\mathcal{G}_{k^*} = \frac{1}{2}\|\mathbf{d}_{k^*}\|^2 + \frac{\beta}{L_g\eta}\|\nabla g(\mathbf{x}_{k^*})\|^2$, it follows that $\|\mathbf{d}_{k^*}\|^2 \leq 2\mathcal{G}_{k^*}$ and $\|\nabla g(\mathbf{x}_{k^*})\|^2 \leq$
 872 $\frac{L_g\eta}{\beta}\mathcal{G}_{k^*} = \frac{L_g\mathcal{G}_{k^*}}{LK^{(1-p)/(3+p)}}$. By the definition $\mathbf{d}_k = \nabla f(\mathbf{x}_k) + \lambda_k \nabla g(\mathbf{x}_k)$ and the fact that $\lambda_k \geq 0$,
 873 the proof is complete.

874 C Other Choices of $\phi(\mathbf{x})$ and their connection to methods considered in the 875 literature.

876 In this section, we briefly discuss the connection between other methods studied in the literature and
 877 the general algorithmic framework described in (11)-(12).

878 **Lower-level linearization based methods.** If we set $\phi(\mathbf{x}) = \alpha(g(\mathbf{x}) - g^*)$ in the update (12),
 879 where $\alpha = 1/\eta$, the resulting method closely aligns with the lower-level linearization-based approach
 880 introduced in [2]. This method was originally developed to solve simple bilevel optimization problems
 881 with a *convex* lower-level objective. The key idea of this type of method is to construct a halfspace to
 882 approximate the lower-level solution set \mathcal{X}_g^* . Specifically, the approximated set is constructed using a
 883 linear approximation of the lower-level objective as follows,

$$\mathcal{X}_k = \{\mathbf{x} \in \mathbb{R}^n : g(\mathbf{x}_k) + \nabla g(\mathbf{x}_k)^\top (\mathbf{x} - \mathbf{x}_k) \leq g^*\}$$

884 If g is convex, then the constructed set \mathcal{X}_k contains \mathcal{X}_g^* for all k . The update of the projection variant
 885 of the algorithm in [2] is as follows,

$$\mathbf{x}_{k+1} = \Pi_{\mathcal{X}_k}(\mathbf{x}_k - \eta \nabla f(\mathbf{x}_k))$$

886 which would be equivalent to

$$\mathbf{x}_{k+1} = \operatorname{argmin}_{\mathbf{x}} \|\mathbf{x} - (\mathbf{x}_k - \eta \nabla f(\mathbf{x}_k))\|^2 \quad \text{s.t.} \quad g(\mathbf{x}_k) + \nabla g(\mathbf{x}_k)^\top (\mathbf{x} - \mathbf{x}_k) \leq g^*$$

887 Through change of variables and defining $\mathbf{d} = (\mathbf{x}_k - \mathbf{x})/\eta$, we can equivalently reformulate the
 888 above subproblem as

$$\mathbf{d}_k = \operatorname{argmin}_{\mathbf{d}} \|\mathbf{d} - \nabla f(\mathbf{x}_k)\|^2 \quad \text{s.t.} \quad \nabla g(\mathbf{x}_k)^\top \mathbf{d} \geq (g(\mathbf{x}_k) - g^*)/\eta.$$

This is a special instance of [12] with $\phi(\mathbf{x}) = (g(\mathbf{x}) - g^*)/\eta$. This choice of $\phi(\mathbf{x})$ is suitable for convex problems, as the solution set \mathcal{X}_g^* is convex and can be contained within \mathcal{X}_k . However, when the lower-level loss is nonconvex, \mathcal{X}_g^* is also nonconvex, meaning the inclusion $\mathcal{X}_g^* \subseteq \mathcal{X}_k$ is not guaranteed. To address this, $\phi(\mathbf{x})$ must be adapted, and using the gradient norm offers a natural extension to the nonconvex case.

Orthogonal projection methods. BiLevel Optimization with Orthogonal Projection (BLOOP) [5] was recently proposed for stochastic simple bilevel optimization. Its key idea is projecting the upper-level gradient to be orthogonal to the lower-level gradient. In the deterministic version, the descent direction \mathbf{d}_k for the update $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \mathbf{d}_k$ is chosen as

$$\mathbf{d}_k = \beta \nabla g(\mathbf{x}_k) + \left[\nabla f(\mathbf{x}_k) - \frac{\nabla f(\mathbf{x}_k)^\top \nabla g(\mathbf{x}_k)}{\|\nabla g(\mathbf{x}_k)\|^2} \nabla g(\mathbf{x}_k) \right]$$

The second part of \mathbf{d}_k is the projection of the upper-level gradient onto the orthogonal space of the lower-level gradient. If we rearrange the terms in \mathbf{d}_k , \mathbf{d}_k is equivalent to

$$\mathbf{d}_k = \underset{\mathbf{d}}{\operatorname{argmin}} \|\mathbf{d} - \nabla f(\mathbf{x}_k)\|^2 \quad \text{s.t.} \quad \nabla g(\mathbf{x}_k)^\top \mathbf{d} = \beta \|\nabla g(\mathbf{x}_k)\|^2.$$

This is a special case of [12] with $\phi(\mathbf{x}) = \beta \|\nabla g(\mathbf{x})\|^2$, but with an equality constraint instead of an inequality. Solving the equality-constrained subproblem with the chosen $\phi(\mathbf{x})$ ensures convergence of the lower-level objective but not the upper-level one [5]. In contrast, we show that solving the inequality-constrained problem also guarantees convergence for the upper level.

D Connections with Algorithms for General Bilevel Problems

In this section, we discuss why most algorithms designed for general bilevel problems are not directly applicable to our simple bilevel setting and highlight the connections between the two classes of algorithms. In the general form of bilevel problems, the upper-level function f may also depend on an additional variable $\mathbf{y} \in \mathbb{R}^m$ that in turn influences the lower-level problem:

$$\min_{\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m} f(\mathbf{x}, \mathbf{y}) \quad \text{s.t.} \quad \mathbf{x} \in \arg \min_{\mathbf{z} \in \mathbb{R}^n} g(\mathbf{z}, \mathbf{y})$$

However, in our considered simple bilevel setting, there is no additional upper-level variable. As a result, the upper-level updates present in algorithms for general bilevel problems become invalid. When these updates are removed, some algorithms—such as those in [43, 44, 35]—reduce to standard gradient descent on g , i.e., $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \nabla g(\mathbf{x}_k)$. Many other methods [33, 34, 31, 45, 32] reduce to the update,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k (\nabla f(\mathbf{x}_k) + \lambda_k \nabla g(\mathbf{x}_k)),$$

which we refer to as the penalty method for nonconvex simple bilevel problems. We include this method as a baseline in our experiments in Section 6. The key challenge for the penalty method lies in selecting an appropriate penalty parameter λ_k . The choices of λ_k used in general bilevel problems are not suitable for the simple bilevel setting, as they are based on different stationarity metrics. Therefore, determining the appropriate value of λ_k for this method requires a tailored analysis specific to the simple bilevel setting. Note that DBGD algorithm essentially provides a dynamic scheme for selecting λ_k , as described in [14].

D.1 Connections with Stationarity Metrics for General Bilevel Problems

Besides the algorithms themselves, the stationarity metrics for general bilevel problems are also not directly applicable to the simple bilevel setting. For instance, [46, 31, 32] adopt the norm of the hyper-gradient as a measure of stationarity. Recall that the hyper-objective [47] is defined as follows:

$$\min_{\mathbf{y} \in \mathbb{R}^m} \varphi(\mathbf{y}), \quad \text{where } \varphi(\mathbf{y}) = \min_{\mathbf{x} \in X^*(\mathbf{y})} f(\mathbf{x}, \mathbf{y}),$$

where $X^*(\mathbf{y}) \triangleq \arg \min_{\mathbf{z}} g(\mathbf{z}, \mathbf{y})$. However, in the simple bilevel setting without upper-level variables \mathbf{y} , the norm of the hyper-gradient constant and thus fails to serve as a valid metric. Furthermore, most existing approaches rely on strong convexity or the Polyak–Łojasiewicz (PL) condition for the lower-level problem—assumptions that are violated in our case, where the hyper-gradient may not even be well-defined.

Other works, such as [35], consider alternative stationarity metrics. When rewritten in the context of our simple bilevel setting, their condition becomes: there exists $\mathbf{w} \in \mathbb{R}^n$ such that

$$\|\nabla^2 g(\hat{\mathbf{x}})(\nabla f(\hat{\mathbf{x}}) + \nabla^2 g(\hat{\mathbf{x}})\mathbf{w})\|^2 \leq \epsilon_f, \quad \|\nabla g(\hat{\mathbf{x}})\|^2 \leq \epsilon_g.$$

Intuitively, the first condition ensures that the component of $\nabla f(\hat{\mathbf{x}}) + \nabla^2 g(\hat{\mathbf{x}})\mathbf{w}$ projected onto the kernel of $\nabla^2 g(\hat{\mathbf{x}})$ is small, i.e.,

$$\text{Proj}_{\text{Ker}(\nabla^2 g(\hat{\mathbf{x}}))}(\nabla f(\hat{\mathbf{x}}) + \nabla^2 g(\hat{\mathbf{x}})\mathbf{w}) \approx 0.$$

This stationary metric is generally weaker than the metric defined in Definition 3.2

D.2 Additional Related Works on General Bilevel Problems

To go beyond strongly convex lower-level objectives, additional assumptions on the lower-level problem are necessary to ensure meaningful guarantees, particularly in light of the negative results for general bilevel optimization with merely convex lower-level objectives [32]. A common strategy is to assume that the nonconvex lower-level objective satisfies the Polyak–Łojasiewicz (PL) condition. Specifically, a penalty-based gradient method was introduced in [34] for both unconstrained and constrained nonconvex-PL bilevel optimization. Later, [35] proposed GALET, a Hessian-vector-product-based method with non-asymptotic convergence guarantees to the modified KKT points of a gradient-based reformulation. In [31], nonconvex bilevel optimization under the proximal error-bound (EB) condition was studied, which is analogous to the PL condition. More recently, in [36], a Hessian/Jacobian-free method was developed that achieves optimal convergence complexity for nonconvex-PL bilevel problems. Besides imposing the PL condition on the lower-level problem, these works also rely on different additional assumptions. For example, [33] additionally assumes that both the upper- and lower-level function values, as well as the norms of their gradients, are bounded, and the lower-level optimal solution is unique. The work in [35] requires both PL and convexity assumptions on the lower-level problem to guarantee convergence. The studies in [31] and [32] impose the condition that a weighted sum of the upper- and lower-level objectives satisfies the PL condition. Finally, in [36] it is assumed that $\nabla^2 g(\mathbf{x})$ is non-singular at the minimizer of g .

D.3 On the Role of the PL Condition

The PL condition plays a central role in the analyses of the aforementioned works in general bilevel optimization. For example, [32] heavily relies on the fact that the PL condition induces a "strongly convex subspace" around any minimizer of the lower-level objective. This structural property enables the adaptation of proof techniques similar to those in [45], which developed an algorithm for general bilevel problems with a strongly convex lower-level objective. Essentially, in general bilevel settings, the PL condition ensures the continuity of the hyper-objective $\varphi(\mathbf{y})$, thereby guaranteeing the existence of the hyper-gradient. This facilitates rapid convergence to a neighborhood of $X^*(\mathbf{y})$. However, in our considered simple bilevel setting, the hyper-objective and its gradient are not well-defined, and we instead rely on alternative stationarity metrics. Consequently, the PL condition is less applicable and offers limited benefit compared to its role in general bilevel problems.

E Experiments Details

In this section, we include more details of the numerical experiments in Section 6. All simulations are implemented using MATLAB R2022a on a PC running macOS Sonoma with an Apple M1 Pro chip and 16GB Memory.

Toy Example. Recall that for Problem (24), the optimal solution set of the lower-level problem is given by $\mathcal{X}_g^* = \{\mathbf{x} \in \mathbb{R}^2 : x_2 = \sin(10x_1)\}$. The optimal solution of the bilevel problem is $\mathbf{x}^* = (-\frac{\pi}{20}, -1)$. We apply DBGD using $\phi(\mathbf{x}) = \|\nabla g(\mathbf{x}_k)\|^2$, i.e., with $\beta = 1$, and also employ the Penalty methods introduced in Section D with $\lambda \in \{1, 10, 100, 1000\}$. Both methods are initialized at the point $\mathbf{x}_0 = (-3, -1)$, using a base stepsize of $\eta = 10^{-2}$ and a total of $K = 10^3$ iterations. Since the penalty methods become unstable for large values of λ , we further scale the stepsize by a factor of $1/(1 + \lambda)$ in each independent run.

Matrix Factorization. For Problem (25), we set $n = r = 10$ to generate \mathbf{U}_* and construct $\mathbf{M} = \mathbf{U}_* \mathbf{U}_*^\top + \epsilon \mathbf{I}_n$, where $\epsilon \sim \mathcal{N}(0, 0.01)$ and $\mathbf{I}_n \in \mathbb{R}^{n \times n}$ denotes the identity matrix. We

977 apply DBGD with $\beta \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ and compare it against the penalty
 978 methods described in Section D using $\lambda \in \{1, 2, 5, 10, 20, 50, 100, 200, 500, 1000\}$. Both methods
 979 use a stepsize of $\eta = 10^{-5}$ and are run for $K = 10^6$ iterations. Since the penalty methods become
 980 unstable for large values of λ , we further scale the stepsize by a factor of $1/(1 + \lambda)$ in each independent
 981 run. The hyperparameter α in both f_1 and f_2 is set to 1.

982 E.1 Additional Experiment

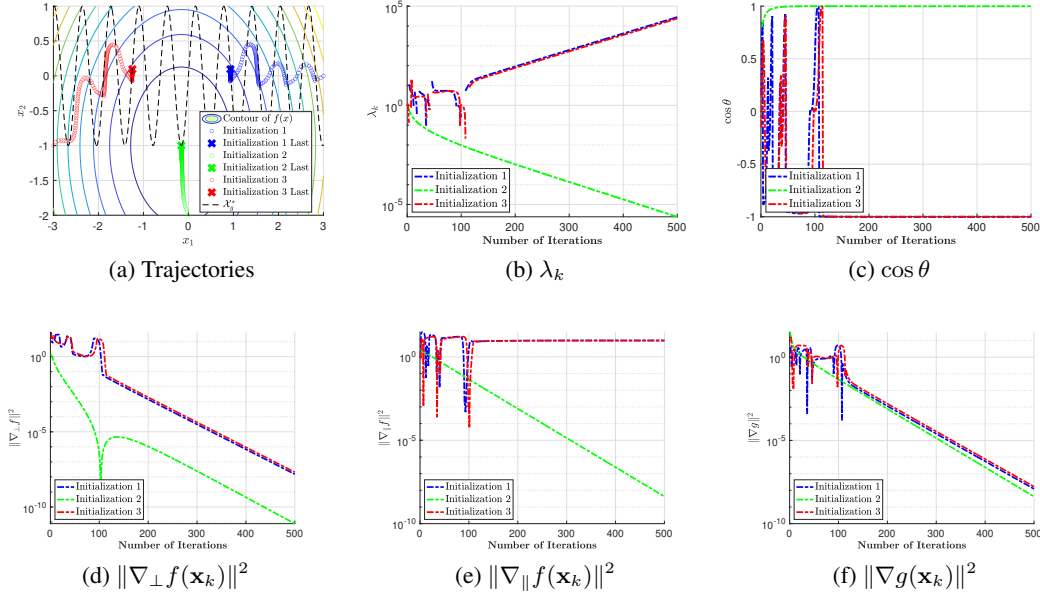


Figure 3: Solving Problem (24) with different Initializations

983 In this additional experiment, we analyze the exact stationary points to which DBGD converges and
 984 examine the effect of different λ_k values at these points, as discussed in Section 3.

985 We consider the problem in Equation (24) from Section 6 and run DBGD with $\phi(\mathbf{x}) = \|\nabla g(\mathbf{x})\|^2$
 986 on the specified instance. As shown in Figure 3, the algorithm converges to three distinct stationary
 987 points, depending on the initialization. This behavior corresponds to the two scenarios discussed in
 988 Section 3, further supporting our theoretical insights.

- 989 • **Case I:** For Initialization 2 (green), DBGD converges to a point where both $\|\nabla f(\mathbf{x}_k)\|$ and
 990 $\|\nabla g(\mathbf{x}_k)\|$ are small. As shown in Figure 3(d), (e), and (f), all three metrics decrease. As illustrated
 991 in Figure 3(c), the cosine of the angle between $\nabla f(\mathbf{x}_k)$ and $\nabla g(\mathbf{x}_k)$ remains positive and eventually
 992 approaches 1. Figure 3(b) shows that λ_k decreases to 0, aligning with the closed-form expression
 993 (16).
- 994 • **Case II:** For Initializations 1 and 3 (blue and red), DBGD converges to stationary points where
 995 $\|\nabla g(\mathbf{x}_k)\|$ is small, as shown in Figure 3(f). Additionally, $\nabla f(\mathbf{x}_k)$ has minimal energy in directions
 996 orthogonal to $\nabla g(\mathbf{x}_k)$, as seen in Figure 3(d). The remaining energy of $\nabla f(\mathbf{x}_k)$ is entirely in the
 997 opposite direction of $\nabla g(\mathbf{x}_k)$, since $\|\nabla_{\parallel} f(\mathbf{x}_k)\|$ does not converge (Figure 3(e)), and the angle
 998 between $\nabla f(\mathbf{x}_k)$ and $\nabla g(\mathbf{x}_k)$ is close to 180° , as shown in Figure 3(c). In this case, λ_k cannot
 999 be bounded by an absolute constant, as depicted in Figure 3(b), which is also consistent with our
 1000 theoretical results.