IMPROVED KKT COMPLEXITY FOR FIRST-ORDER BILEVEL OPTIMIZATION UNDER WEAK LOWER-LEVEL CONVEXITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We study deterministic first-order bilevel optimization under weak lower-level convexity, allowing nonconvex lower-level objectives and without assuming strong convexity, the Polyak–Łojasiewicz condition, or an error-bound property. We consider a δ-relaxed Moreau-gap constraint, with $\delta > 0$, for the lower-level stationarity condition and propose an inexact variable-smoothing penalty method (IVSP) for computing its approximate Karush–Kuhn–Tucker (KKT) points. For any fixed relaxation level δ, under a standard extended no-nonzero-abnormal-multiplier constraint qualification (ENNAMCQ), we prove finite stabilization of the adaptive penalty parameter and an overall $\widetilde{O}(\varepsilon^{-3})$ first-order complexity for computing an ε-KKT point. Notably, we give verifiable sufficient conditions for ENNAMCQ covering convex lower-level objectives without nonconstant affine segments (including the strictly convex case), and a class of nonconvex sample-reweighting models. The positive relaxation avoids the intrinsic constraint-qualification degeneracy of the exact Moreau-gap constraint while achieving an $\mathcal{O}(\sqrt{\delta})$ lower-level near-stationarity guarantee. Numerical experiments on synthetic and real-world bilevel learning problems illustrate the practical performance of IVSP.

1 INTRODUCTION

Bilevel optimization has found broad applications in hyperparameter optimization (Pedregosa, 2016; Franceschi et al., 2018), meta-learning (Franceschi et al., 2018; Rajeswaran et al., 2019), neural architecture search (Liu et al., 2019), and data reweighting or hypercleaning (Ren et al., 2018). We study bilevel problems of the form

$$\min_{x\in X,\, y\in Y} f(x,y) \qquad \text{s.t.} \qquad y \in \arg\min_{u\in Y} g(x,u).$$

In this paper, $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ are nonempty closed convex sets, f and g are smooth, and $g(x,\cdot)$ is uniformly weakly convex. Thus, the lower-level (LL) problem may be nonconvex and need not be strongly convex, satisfy the Polyak–Łojasiewicz (PL) condition, or admit a unique solution.

The modeling flexibility of bilevel optimization comes, however, with a substantial computational challenge: the upper-level (UL) objective depends implicitly on a solution of the lower-level problem. Classical hypergradient methods typically differentiate through the lower-level optimization process or use approximate implicit differentiation, which can require Hessian/Jacobian-vector products or linear-system solves (Ghadimi & Wang, 2018; Grazzi et al., 2020; Ji et al., 2021). These costs have motivated a growing line of Hessian-free and fully first-order methods (Liu et al., 2022; Kwon et al., 2023; Liu et al., 2024; Chen et al., 2025). Despite this progress, the strongest nonasymptotic guarantees remain concentrated in well-conditioned lower-level regimes, such as strong convexity, PL, or related error-bound conditions (Kwon et al., 2023; Huang, 2024; Chen et al., 2025).

Much less is known once such lower-level regularity is removed. For convex, possibly nonsmooth lower-level problems, Lu & Mei (2024) develop first-order penalty methods with approximate weak-KKT guarantees and a $\widetilde{O}(\varepsilon^{-4})$ complexity. For weakly convex/nonconvex lower levels, MEHA (Liu et al., 2024) and PNGBiO (Zhang et al., 2025) develop deterministic first-order methods based on

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Moreau envelope penalty reformulations. Their guarantees concern stationarity of a penalized reformulation together with lower-level feasibility rather than the full KKT system of the corresponding constrained problem. Requiring both $O(\varepsilon)$ stationarity and $O(\varepsilon)$ Moreau-gap feasibility gives an $O(\varepsilon^{-4})$ complexity. TSP (Lu, 2025) instead directly targets approximate KKT points of a Moreau envelope constrained reformulation using a stochastic primal–dual method.

A common route in the literature is to replace or relax the lower-level condition by nonlinear constraints, leading to a nonlinear-programming formulation of the bilevel problem. This viewpoint is particularly natural when the lower-level solution need not be unique, as may occur without strong convexity or PL-type regularity, so that a smooth reduced hyperobjective is generally unavailable. For such constrained formulations, the KKT conditions constitute the canonical first-order optimality system of nonlinear programming, encompassing stationarity, feasibility, and complementarity; see, e.g., Nocedal & Wright (2006). Existing bilevel methods nevertheless target several different solution notions, including stationarity of penalized reformulations, weak KKT-type conditions, and KKT systems of constrained reformulations; see Table 1. These distinctions become particularly important in the weakly convex lower-level regime, where both the attainable solution guarantee and its deterministic first-order complexity remain substantially less understood.

This motivates the following question:

Can a deterministic first-order method achieve a genuine KKT guarantee for bilevel optimization with a weakly convex lower level at an improved complexity?

1.1 Our approach and contributions

Our approach is to replace the lower-level condition by a δ-relaxed Moreau-gap constraint, with $\delta > 0$, yielding a smooth nonlinear-programming relaxation whose KKT system is our target solution concept. The positive relaxation is essential to this KKT viewpoint: because the Moreau gap is nonnegative, the exact constraint is intrinsically degenerate and fails standard constraint qualifications at every feasible point. A fixed $\delta > 0$ avoids this intrinsic degeneracy and permits a standard constraint-qualification-based KKT analysis, while retaining an $\mathcal{O}(\sqrt{\delta})$ lower-level near-stationarity guarantee; see Section 2.1 and Appendix A.4.

We solve the relaxed problem using a smoothed exact-penalty method with adaptive penalty updates and inexact Moreau evaluations. Our analysis is carried out under ENNAMCQ (Definition 3), which yields finite stabilization of the penalty parameter. We establish verifiable sufficient conditions for ENNAMCQ covering convex lower-level objectives without nonconstant affine segments (such as strictly and strongly convex cases) and a class of nonconvex sample-reweighting models; these conditions apply to our few-shot and data hyper-cleaning problems (Appendix B). For a suitable proximal parameter, the Moreau subproblems are strongly convex even when the lower-level objective is nonconvex. Our analysis requires these subproblems to be solved only to polynomially decreasing accuracy, so linearly convergent first-order inner solves add only a logarithmic overhead.

Our main contributions are as follows.

- **A KKT-oriented deterministic first-order method.** We develop the *inexact variable-smoothing penalty method* (IVSP), which uses only first-order information and projections, accommodates inexact Moreau evaluations, and admits either a fixed stepsize or an inexact backtracking scheme.
- **Improved KKT complexity under weak lower-level convexity.** For any fixed relaxation level $\delta > 0$, under ENNAMCQ, the adaptive penalty parameter stabilizes after finitely many updates (Lemma 5), and IVSP computes an ε-KKT point of the δ-relaxed Moreau-gap formulation with an overall $\widetilde{O}(\varepsilon^{-3})$ first-order complexity (Corollary 7). These guarantees do not require strong convexity, PL, or an error-bound property of the lower-level problem.
- **Empirical evaluation.** We compare IVSP with representative deterministic first-order bilevel methods on synthetic and real-world learning problems and observe competitive numerical performance.

1.2 Related works

Strong convexity, the Polyak–Łojasiewicz (PL) condition, and related error-bound properties provide quantitative control of lower-level optimization errors, which facilitates the approximation of lower-level responses and hypergradients. Under strong convexity, the lower-level solution is

Table 1: Comparison of representative deterministic first-order methods for bilevel optimization. The horizontal division separates methods relying on SC/PL lower-level regularity from those allowing convex or weakly convex lower-level problems without PL. SC/PL provides stronger control of lower-level errors and stability, while SC additionally guarantees uniqueness; such properties can be exploited to obtain sharper complexity guarantees. Complexity is reported using a norm-based ε-stationarity/KKT criterion whenever applicable.

Method	LL assumption	Approach	Guarantee	Complexity
BOME (Liu et al., 2022)	PL	barrier	KKT*	$O(\varepsilon^{-6})$–$O(\varepsilon^{-8})$
F^{2}SA (Kwon et al., 2023)	SC	penalty	hyperobj. stat.	$\widetilde{O}(\varepsilon^{-3})$
V-PBGD (Shen & Chen, 2023)	PL	penalty	pen. stat.	$\widetilde{O}(\varepsilon^{-2})$
SLM (Lu, 2023)	PL + reg.	primal-dual	KKT (fixed-δ relax.)	$O(\varepsilon^{-2})$
F^{2}BA (Chen et al., 2025)	SC	penalty	hyperobj. stat.	$\widetilde{O}(\varepsilon^{-2})$
Lu–Mei (Lu & Mei, 2024)	convex	penalty/minimax	w-KKT	$\widetilde{O}(\varepsilon^{-4})$
MEHA (Liu et al., 2024)	WC	ME penalty	pen. stat. + feas.	$O(\varepsilon^{-4})$
PNGBiO (Zhang et al., 2025)	WC	ME penalty	pen. stat. + feas.	$O(\varepsilon^{-4})$
IVSP (ours)	WC	ME-gap sm. pen.	KKT (fixed-δ relax.)	$\widetilde{O}(\varepsilon^{-3})$

SC/WC = strongly/weakly convex; PL = Polyak–Łojasiewicz; ME = Moreau envelope. $\widetilde{O}(\cdot)$ suppresses logarithmic factors. The target formulations, solution concepts, and rate conventions differ across methods; see Appendix E for details.

unique and classical approaches exploit implicit or iterative differentiation. More broadly, under SC/PL-type lower-level regularity, recent methods seek to avoid second-order computations through penalty, primal–dual, finite-difference, value-function, and other fully first-order constructions. Representative methods include BOME (Liu et al., 2022), F^{2}SA (Kwon et al., 2023), Prox-F^{2}BA (Kwon et al., 2024), SLM (Lu, 2023), HJFBiO (Huang, 2024), F^{2}BA (Chen et al., 2025), PBGD-Free (Jiang et al., 2025), and SGHA (Gu et al., 2026). In the strongly convex setting, fully first-order methods can attain the near-optimal $\widetilde{O}(\varepsilon^{-2})$ complexity for hyperobjective stationarity (Chen et al., 2025). Our interest is in regimes where strong convexity, PL, and error-bound conditions are unavailable.

Within this broader literature, relatively few methods directly target KKT or KKT-type solution concepts. BOME (Liu et al., 2022) introduces a KKT-type residual under the PL condition, although its criterion does not correspond to the full KKT system of a constrained bilevel reformulation. Lu & Mei (2024) develop first-order penalty methods with approximate weak-KKT guarantees for convex, possibly nonsmooth lower-level problems. SLM (Lu, 2023) establishes KKT guarantees for a positive value-function-gap formulation under the PL condition and additional regularity assumptions, while TSP (Lu, 2025) directly targets approximate KKT points of a Moreau envelope constrained formulation using a stochastic primal–dual method.

Turning specifically to weakly convex lower-level objectives, the Moreau envelope reformulation introduced by Gao et al. (2026) has led to several first-order methods. MEHA (Liu et al., 2024) develops a single-loop Hessian-free method, while PNGBiO (Zhang et al., 2025) extends this approach to generalized smoothness; both establish stationarity guarantees for penalized Moreau envelope reformulations together with lower-level feasibility. Our method instead combines a positive Moreau-gap constrained formulation with a smoothed exact penalty, allows inexact proximal computations, and establishes deterministic first-order complexity for computing approximate KKT points under weak lower-level convexity. Since the solution concepts in these works differ, Table 1 reports both the guarantee and its corresponding complexity.

2 Proposed method

2.1 Moreau-gap reformulation

Let $g(x,\cdot)$ be σ-weakly convex for every $x \in X$, where $\sigma \geq 0$. That is, $g(x,\cdot) + \frac{\sigma}{2}\|\cdot\|^2$ is convex for any $x \in X$. Fix $\rho > 0$ such that $\rho\sigma < 1$. Define

$$\psi_\rho(z;x,y) := g(x,z) + \frac{1}{2\rho}\|z-y\|^2, \qquad g_\rho(x,y) := \min_{z\in Y} \psi_\rho(z;x,y),$$

and let $z_\rho(x, y)$ denote the unique minimizer. Following the Moreau envelope reformulation of Gao et al. (2026), we consider

$$\min_{w \in C} f(w) \qquad \text{s.t.} \qquad R(w) := g(w) - g_\rho(w) - \delta \le 0, \tag{1}$$

where $w = (x, y)$, $C := X \times Y$, and $\delta > 0$. Since $\psi_\rho(\cdot; w)$ is $(\rho^{-1} - \sigma)$-strongly convex,

$$g(w) - g_\rho(w) \ge \frac{1 - \rho\sigma}{2\rho} \|y - z_\rho(w)\|^2,$$

and the optimality condition of the Moreau subproblem gives

$$g(w) - g_\rho(w) = 0 \iff y = z_\rho(w) \iff 0 \in \nabla_y g(w) + N_Y(y),$$

where N_Y denotes the normal cone to Y. Thus, under weak convexity, the exact Moreau-gap constraint characterizes lower-level first-order stationarity (Liu et al., 2024); when $g(x, \cdot)$ is convex, it is equivalent to lower-level global optimality and recovers the original bilevel constraint (Gao et al., 2026).

As in (Lu, 2025), we use a positive relaxation $\delta > 0$, which is important both for regularity and approximation. Since $g - g_\rho \ge 0$, every feasible point of the exact constraint $g - g_\rho \le 0$ is a minimizer of the gap over C; hence standard constraint qualifications such as MFCQ and NNAMCQ necessarily fail there (Lin et al., 2014; Ye et al., 2023; Gao et al., 2026; Bai et al., 2026). On the other hand, every feasible point of (1) satisfies

$$\|y - z_\rho(w)\| \le \sqrt{\frac{2\rho\delta}{1 - \rho\sigma}} \text{ and } \operatorname{dist}(0, \nabla_y g(x, z_\rho(w)) + N_Y(z_\rho(w))) \le \sqrt{\frac{2\delta}{\rho(1 - \rho\sigma)}}.$$

Hence, δ directly controls lower-level near stationarity at order $O(\sqrt{\delta})$. Further properties of the positive relaxation, including its behavior as $\delta \downarrow 0$, are given in Appendix A.4.

2.2 Overview of the smoothed exact-penalty method

Our approach to solving (1) begins with the exact-penalty model

$$\min_{w \in C} \; f(w) + \frac{1}{\gamma}[R(w)]_+,$$

where $[t]_+ := \max\{t, 0\}$ and $\gamma > 0$ is the inverse penalty parameter. Thus, decreasing γ strengthens the penalty. Under Assumption 1 and $\rho\sigma < 1$, the residual R is continuously differentiable with

$$\nabla R(w) = \begin{pmatrix} \nabla_x g(x, y) - \nabla_x g(x, z_\rho(w)) \\ \nabla_y g(x, y) - \rho^{-1}(y - z_\rho(w)) \end{pmatrix}; \tag{2}$$

see Appendix A.1. The exact-penalty model remains nonsmooth because of $[\cdot]_+$. We therefore replace it by a smooth approximation ϕ_μ with smoothing parameter $\mu > 0$ and consider

$$\min_{w \in C} M_{\gamma,\mu}(w) := \gamma f(w) + G_\mu(w), \qquad G_\mu(w) := \phi_\mu(R(w)). \tag{3}$$

We use a standard family of smooth approximations satisfying the conditions in Appendix A.2; see (Beck & Teboulle, 2012). In particular, ϕ_μ approximates $[\cdot]_+$ to accuracy $O(\mu)$, $0 \le \phi'_\mu \le 1$, and ϕ'_μ is $O(\mu^{-1})$-Lipschitz. A standard example is the softplus $\phi_\mu(t) = \mu \log(1 + \exp(t/\mu))$.

By the chain rule, $\nabla G_\mu(w) = \phi'_\mu(R(w)) \nabla R(w)$, and ∇G_μ is $O(\mu^{-1})$-Lipschitz; see Appendix A.2. Hence, an exact projected-gradient step for (3) takes the familiar form

$$w_L^k = P_C\Big(w^k - \frac{\mu}{L}\big[\gamma \nabla f(w^k) + \nabla G_\mu(w^k)\big]\Big),$$

where P_C denotes the Euclidean projection onto C. The difficulty is that computing this step requires the Moreau proximal point $z_\rho(w^k)$ due to (2), which is generally unavailable exactly. This motivates an inexact Moreau evaluation together with explicit control of the resulting value and gradient errors.

2.3 INEXACT MOREAU ENVELOPE EVALUATIONS

For a prescribed accuracy $\zeta \geq 0$, let $\widetilde{z} \in Y$ be an approximation of $z_\rho(w)$ satisfying

$$0 \leq \psi_\rho(\widetilde{z}; x, y) - g_\rho(x, y) \leq \zeta. \tag{4}$$

Define the corresponding inexact residual by $\widetilde{R}(w) \coloneqq g(x, y) - \psi_\rho(\widetilde{z}; x, y) - \delta$. Then

$$\widetilde{R}(w) \leq R(w) \leq \widetilde{R}(w) + \zeta. \tag{5}$$

Likewise, let $\widetilde{\nabla} R(w)$ be obtained from (2) by replacing $z_\rho(w)$ with $\widetilde{z}$, and set $\widetilde{\nabla} G_\mu(w) \coloneqq \phi'_\mu(\widetilde{R}(w))\widetilde{\nabla} R(w)$. Under Assumption 1,

$$\left\|\widetilde{\nabla} G_\mu(w) - \nabla G_\mu(w)\right\| = O\left(\sqrt{\zeta} + \frac{\zeta}{\mu}\right). \tag{6}$$

Explicit constants and detailed error estimates are given in Appendix A.3. The bounds (5)–(6) are the only properties of the inexact Moreau evaluations needed in the outer method.

Choice of inner solver. For a suitable proximal parameter, each Moreau subproblem is strongly convex. Standard first-order methods can therefore exploit strong convexity to obtain linear convergence. In particular, projected gradient (Beck, 2017) has this property under our assumptions and is the method used in our implementation. More generally, the complexity argument only requires an inner solver that computes a ζ-accurate solution in $O(\log(1/\zeta))$ iterations uniformly over the proximal subproblems. A computable residual-based stopping rule is given in Appendix C.

2.4 INEXACT VARIABLE-SMOOTHING PENALTY METHOD

We now combine the inexact Moreau evaluations with projected-gradient steps for (3). The outer variable-smoothing and adaptive-penalty mechanism of IVSP is adapted from sESQM (Xu et al., 2026), which was developed for single-level nonconvex optimization with nonlinear inequality constraints and applies one projected-gradient step to a smoothed exact penalty at each iteration while updating the smoothing and penalty parameters. In our bilevel setting, however, the Moreau-gap residual and its gradient depend on the generally unavailable proximal point $z_\rho(w)$. IVSP therefore incorporates the inexact Moreau evaluations of Section 2.3 and controls the resulting value and gradient errors in the outer method.

Let $\{\mu_k\}$ be a positive sequence with $\mu_k \downarrow 0$, let $\{\zeta_k\}$ be the prescribed proximal accuracies, and let $\{\widehat{\gamma}_t\}_{t\geq 0}$ be a positive inverse-penalty grid with $\widehat{\gamma}_t \downarrow 0$. We set $\gamma_k \coloneqq \widehat{\gamma}_{t_k}$, where $\{t_k\}$ is nondecreasing. Given a ζ_k-accurate proximal point $\widetilde{z}^k$ at w^k, form the corresponding $\widetilde{R}^k = \widetilde{R}(w^k)$ and $\widetilde{\nabla} G^k \coloneqq \phi'_{\mu_k}(\widetilde{R}^k)\widetilde{\nabla} R(w^k)$. For a parameter $L > 0$, define

$$w_L^k = P_C\left(w^k - \frac{\mu_k}{L}\left[\gamma_k \nabla f(w^k) + \widetilde{\nabla} G^k\right]\right). \tag{7}$$

The complete method is summarized in Algorithm 1.

Choice of L and inexact line search. We consider two strategies for selecting L. In Strategy I (fixed L), a sufficiently large constant L is used throughout; a sufficient choice L_{safe} is given in Lemma 4. This avoids line search but requires an a priori estimate of L_{safe}. Strategy II instead selects L adaptively by backtracking and does not require L_{safe} to be known in advance. Specifically, starting from $L_{k,0} \in [L_{\min}, L_{\max}]$, we increase L until an inexact Armijo condition is satisfied. For each trial w_L^k, compute a ζ_{k+1}-accurate proximal point $\widetilde{z}_{k,L}^+$ and the corresponding residual $\widetilde{R}_{k,L}^+$. We accept the trial if

$$\gamma_k f(w_L^k) + \phi_{\mu_k}\left(\widetilde{R}_{k,L}^+ + \zeta_{k+1}\right) \leq \gamma_k f(w^k) + \phi_{\mu_k}(\widetilde{R}^k) - \frac{c_1}{2\mu_k}\|w_L^k - w^k\|^2 + \tau_{k,L}, \tag{8}$$

where $\tau_{k,L} \geq 0$ is an explicit error allowance induced by the inexact Moreau evaluation, rather than an additional tuning parameter; its definition is given in Appendix D.1. If the condition fails, we increase L (e.g., $L \leftarrow 2L$) and recompute the trial point and its Moreau evaluation. In practice, $L_{k,0}$ may be initialized using a safeguarded Barzilai–Borwein spectral estimate (Barzilai & Borwein, 1988).

Algorithm 1 Inexact Variable-Smoothing Penalty Method (IVSP)

1: Choose $w^0 \in C$, $\{\mu_k\} \downarrow 0$, $\{\zeta_k\} \downarrow 0$, $\{\widehat{\gamma}_t\} \downarrow 0$, and set $t_0 = 0$. Compute a ζ_0-accurate proximal solution $\widetilde{z}^0$ at w^0 and form $\widetilde{R}^0$.
2: **for** $k = 0, 1, \ldots$ **do**
3: Set $\gamma_k = \widehat{\gamma}_{t_k}$ and form $\widetilde{\nabla} G^k$ from $\widetilde{z}^k$ and $\widetilde{R}^k$.
4: Choose L_k by Strategy I (fixed L) or Strategy II (backtracking via (8)), obtaining $w^{k+1} = w^k_{L_k}$ defined in (7) and a ζ_{k+1}-accurate proximal solution $\widetilde{z}^{k+1}$ at w^{k+1}.
5: Form $\widetilde{R}^{k+1}$ and set $t_{k+1} = t_k + 1$ if (9) holds; otherwise, set $t_{k+1} = t_k$.
6: **end for**

Adaptive penalty update. After obtaining w^{k+1} and $\widetilde{R}^{k+1}$, we strengthen the penalty whenever

$$\widetilde{R}^{k+1} > 2a_\phi \mu_k \qquad \text{and} \qquad \|w^{k+1} - w^k\| \le c_2 \mu_k \gamma_k, \tag{9}$$

where $a_\phi > 0$ is the smoothing-approximation constant defined in Appendix A.2. The first condition detects violation beyond the smoothing scale, which cannot be caused by the inexact Moreau evaluation since $\widetilde{R}^{k+1} \le R(w^{k+1})$; the second requires a small outer step. Thus, if a significant violation persists despite a small step, we move to the next value of the inverse-penalty grid $\{\widehat{\gamma}_t\}$.

3 Theoretical guarantees

3.1 Assumptions and stationarity measures

We impose the following standard assumptions.

Assumption 1. *The sets $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ are nonempty, compact, and convex. The functions f and g are continuously differentiable on neighborhoods of their respective domains, and ∇f and ∇g are L_f- and L_g-Lipschitz continuous, respectively, on $C = X \times Y$.*

By compactness, $f_{\inf} := \min_{w \in C} f(w)$ is well defined and finite. The Lipschitz continuity of ∇g implies that $g(x, \cdot)$ is weakly convex uniformly over $x \in X$; see, e.g., Bauschke & Combettes (2017, Lemma 2.64). We denote by $\sigma \ge 0$ a uniform weak-convexity modulus (and one may always take $\sigma = L_g$), and fix $\rho > 0$ such that $\rho\sigma < 1$. Since our target is the constrained relaxation (1), we measure convergence through its KKT residuals; cf. Lu (2025). This differs from penalty-stationarity criteria used, for example, in PBGD and MEHA (Shen et al., 2025; Shen & Chen, 2023; Liu et al., 2024), whose stationary systems depend explicitly on a prescribed penalty weight. By contrast, the KKT multiplier is determined jointly with the primal variables, while stationarity, feasibility, and complementarity are controlled separately.

Definition 2 (Approximate KKT point)**.** For $w \in C$ and $\lambda \ge 0$, define

$$\Phi^{\mathrm{s}}_\rho(w, \lambda) := \operatorname{dist}(0, \nabla f(w) + \lambda \nabla R(w) + N_C(w)),$$
$$\Phi^{\mathrm{f}}_\rho(w) := [R(w)]_+, \qquad \Phi^{\mathrm{c}}_\rho(w, \lambda) := |\lambda R(w)|.$$

We call w an $(\varepsilon_s, \varepsilon_f, \varepsilon_c)$-KKT point of (1) if there exists $\lambda \ge 0$ such that $\Phi^{\mathrm{s}}_\rho(w, \lambda) \le \varepsilon_s$, $\Phi^{\mathrm{f}}_\rho(w) \le \varepsilon_f$, and $\Phi^{\mathrm{c}}_\rho(w, \lambda) \le \varepsilon_c$. When the three tolerances equal ε, we simply call w an ε-KKT point. The case $\varepsilon = 0$ gives the standard KKT conditions.

To control the adaptive penalty parameter, we impose the *extended no-nonzero-abnormal-multiplier constraint qualification* (ENNAMCQ). Conditions of this type are standard in value-function reformulations of bilevel programs (Lin et al., 2014; Ye et al., 2023; Alcantara & Takeda, 2026), and have also been used for relaxed Moreau envelope reformulations; see, e.g., Gao et al. (2026). ENNAMCQ extends the classical NNAMCQ used in nonlinear programming (Ye, 2000) by imposing the same nonstationarity condition also at infeasible points. For the smooth single-constraint problem (1), it also coincides with the constraint qualification used in sESQM (Xu et al., 2026). This condition is the key ingredient in proving finite stabilization of the penalty parameter, which in turn is essential for the complexity analysis.

Definition 3 (ENNAMCQ)**.** We say that ENNAMCQ holds for (1) if, for every $w \in C$ with $R(w) \ge 0$, $0 \notin \nabla R(w) + N_C(w)$.

By compactness, ENNAMCQ yields a uniform separation from constraint stationarity over all active or infeasible points; see Lemma 21. We provide verifiable sufficient conditions in Appendix B under which ENNAMCQ holds automatically. For a positive Moreau-gap relaxation, Gao et al. (2026) established automatic ENNAMCQ under strict lower-level convexity in a more general setting with an x-dependent lower-level feasible set and additional *joint* weak-convexity assumptions. In our setting, strict convexity of $g(x,\cdot)$ over Y guarantees ENNAMCQ. We further extend this sufficient condition beyond strict convexity: for a convex lower-level problem, it suffices that every affine segment of $g(x,\cdot)$ on Y be constant. We further establish a sufficient condition for nonconvex sample-reweighting models, covering our few-shot and data hyper-cleaning settings.

3.2 Basic convergence properties

We first establish two properties underlying the complexity analysis: descent of the proposed step and finite stabilization of the adaptive penalty parameter. We define the Lyapunov function

$$Q_k \coloneqq \gamma_k\big(f(w^k) - f_{\inf}\big) + G_{\mu_k}(w^k) + c_\phi \mu_k,$$

where c_ϕ is the constant in the smoothing bound (23).

Lemma 4 (Well-definedness and descent)**.** *Under Assumption 1, there exists a constant $L_{\mathrm{safe}} > 0$, independent of k, such that the following hold.*

(i) Every trial parameter $L \geq L_{\mathrm{safe}}$ satisfies the inexact Armijo condition (8)*. Consequently, any fixed $L \geq L_{\mathrm{safe}}$ is admissible under Strategy I, while the backtracking procedure in Strategy II terminates finitely at every iteration. Moreover, under Strategy II, the accepted parameters L_k and the number of backtracking trials are uniformly bounded.*

(ii) Under either strategy, defining $L_k \coloneqq L$ for Strategy I, we have

$$Q_{k+1} \leq Q_k - \frac{c_1}{2\mu_k}\left\|d^k\right\|^2 + \tau_{k,L_k}, \quad \forall k \geq 0.$$

An explicit expression for L_{safe}, together with uniform bounds on L_k and the number of backtracking trials, is given in Appendix D.1.

Lemma 5 (Finite penalty stabilization)**.** *Suppose ENNAMCQ holds for* (1)*, the proximal accuracies satisfy $\zeta_k = o(\mu_k)$, and the inverse-penalty grid satisfies $\widehat{\gamma}_t \leq M_\gamma \widehat{\gamma}_{t+1}$ for some $M_\gamma > 1$ and every $t \geq 0$. Then there exists $\underline{\gamma} > 0$ such that $\gamma_k \geq \underline{\gamma}$ for all k. Consequently, only finitely many penalty updates occur, and γ_k is eventually constant.*

3.3 Complexity guarantees

We now establish the convergence rates of the proposed method under polynomial smoothing and proximal-accuracy schedules.

Theorem 6 (Outer-iteration complexity)**.** *Suppose Assumption 1 and ENNAMCQ hold, and suppose that the inverse-penalty grid satisfies $\widehat{\gamma}_t \leq M_\gamma \widehat{\gamma}_{t+1}$ for some $M_\gamma > 1$ and every $t \geq 0$. Let $\{\mu_k\}$ be positive and nonincreasing with $\mu_k = \Theta((1+k)^{-r})$ for some $r \in (0,1)$, and let $\{\zeta_k\}$ satisfy $\zeta_k \leq \bar{\zeta}\mu_k^q$ for some $\bar{\zeta} > 0$ and $q > 1/r$. Let $\{w^k\}$ be generated by Algorithm 1 and, for each $k \geq 0$, set $\widetilde{\lambda}_{k+1} \coloneqq \gamma_k^{-1}\phi'_{\mu_k}(\widetilde{R}^{k+1})$. Then, for every sufficiently large K, there exists $\widehat{k} \in [\lceil K/2 \rceil, K]$ such that*

$$\begin{aligned} &\Phi_\rho^{\mathrm{s}}(w^{\widehat{k}+1}, \widetilde{\lambda}_{\widehat{k}+1}) = O\Big(K^{-(1-r)/2}\Big), \qquad \Phi_\rho^{\mathrm{f}}(w^{\widehat{k}+1}) = O(K^{-r}), \\ &\Phi_\rho^{\mathrm{c}}(w^{\widehat{k}+1}, \widetilde{\lambda}_{\widehat{k}+1}) = O(K^{-r}). \end{aligned} \tag{10}$$

Consequently, within $K = O\Big(\max\Big\{\varepsilon_s^{-2/(1-r)}, \varepsilon_f^{-1/r}, \varepsilon_c^{-1/r}\Big\}\Big)$ outer iterations, we obtain an $(\varepsilon_s, \varepsilon_f, \varepsilon_c)$-KKT point of the δ-relaxed Moreau-gap formulation (1)*.*

The rates in (10) show that smaller r favors stationarity, whereas larger r favors feasibility and complementarity. Balancing $(1-r)/2 = r$ gives $r = 1/3$. Thus, for any $q > 3$, Algorithm 1 reaches an ε-KKT iterate within $O(\varepsilon^{-3})$ outer iterations. Accounting additionally for the logarithmic cost of the inexact Moreau evaluations yields the following overall first-order complexity.

Corollary 7 (Overall first-order complexity)**.** *Suppose the conditions of Theorem 6 hold and, in addition, choose $\zeta_k = \Theta(\mu_k^q)$. Suppose each required Moreau subproblem is solved by projected gradient to objective-gap accuracy ζ_k. Let B_{ls} denote a uniform bound on the number of trial Moreau evaluations per outer iteration, with $B_{\mathrm{ls}} = 1$ for Strategy I. Then the overall first-order/projection complexity of IVSP through outer iteration K is $O(B_{\mathrm{ls}}\, K \log K)$. In particular, for the balanced choice $r = 1/3$, IVSP computes an ε-KKT point of the δ-relaxed Moreau-gap formulation* (1) *with overall first-order/projection complexity $O\big(B_{\mathrm{ls}}\, \varepsilon^{-3} \log \frac{1}{\varepsilon}\big) = \widetilde{O}(\varepsilon^{-3})$.*

4 Experiments

In this section, we evaluate IVSP on a synthetic nonconvex problem and two real-world bilevel learning tasks: data hyper-cleaning and few-shot learning. We compare IVSP with representative deterministic first-order bilevel methods, including PNGBiO (Zhang et al., 2025), MEHA (Liu et al., 2024), SLM (Lu, 2023), F^{2}SA (Kwon et al., 2023), V-PBGD (Shen & Chen, 2023), and BOME (Liu et al., 2022). For IVSP, we use the blockwise smoothing schedule $\mu_k = \mu_0\big(1 + \lfloor k/B \rfloor\big)^{-r}$, which satisfies the smoothing condition of Theorem 6. Detailed problem formulations, implementation settings, and algorithmic hyperparameter values, including μ_0, B, and r are provided in Appendix F.

4.1 Synthetic numerical experiment

We first consider the smooth nonconvex toy problem used in (Liu et al., 2024):

$$\min_{\substack{x\in[-100,100]\\ y\in[-100,100]^n}} (x-a)^2 + \|y - ae - c\|^2 \quad \text{s.t.} \quad y \in \operatorname*{arg\,min}_{z\in[-100,100]^n} \sum_{i=1}^{n} \sin(x + z_i - c_i),$$

where e is the all-ones vector. The LL objective is smooth, nonconvex, and 1-weakly convex. We set $n = 1000$, $a = 2$, $c_i = 2$, $x^0 = -6$, and $y^0 = 0$. The optimal solution is $x^* = ((1-n)a + nC)/(1+n)$ and $y_i^* = C + c_i - x^*$, where $C \in \arg\min_{c\in\{-\pi/2+2k\pi:\, k\in\mathbb{Z}\}} |c - 2a|$.

Figure 1 reports the normalized errors in x and y and the KKT residual $\max\{\Phi_\rho^{\mathrm{s}}, \Phi_\rho^{\mathrm{f}}, \Phi_\rho^{\mathrm{c}}\}$ for IVSP. Since the competing methods target different reformulations and solution concepts, we report the IVSP KKT measure only for IVSP. IVSP achieves the fastest reduction and smallest final errors in both variables.

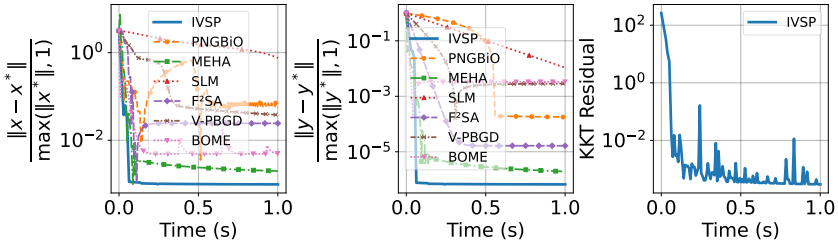
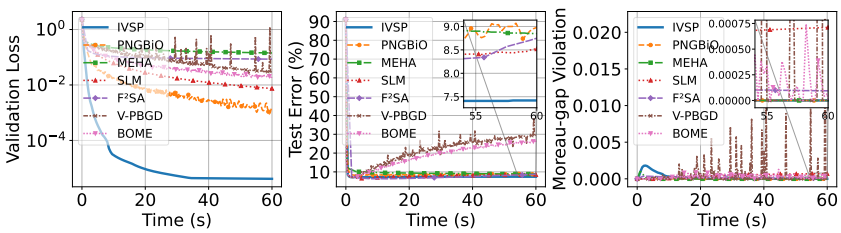


Figure 1: Convergence comparison on the nonconvex problem with $n = 1000$ measured by $\|x - x^*\|/\max\{\|x^*\|, 1\}$, $\|y - y^*\|/\max\{\|y^*\|, 1\}$, and the KKT residual $\max\{\Phi_\rho^{\mathrm{s}}, \Phi_\rho^{\mathrm{f}}, \Phi_\rho^{\mathrm{c}}\}$.

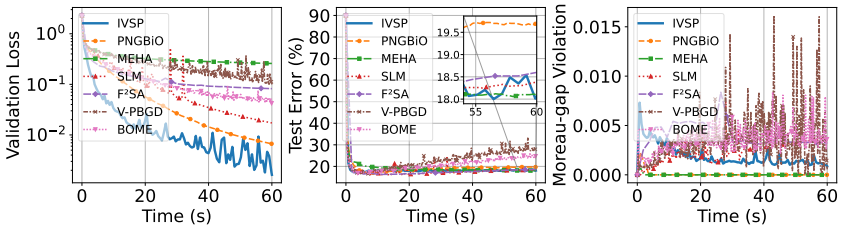
4.2 Real-world applications

Data hyper-cleaning. We consider data hyper-cleaning (Lu, 2023; Liu et al., 2024; Zhang et al., 2025) on MNIST (LeCun et al., 1998) and FashionMNIST (Xiao et al., 2017), where the UL variables assign weights to the training samples, half of which have corrupted labels, and the LL model is a two-layer fully connected network with sigmoid activation, yielding a smooth nonconvex LL problem. This problem falls within the sigmoid-reweighted classification model of Corollary 18, so ENNAMCQ holds for our choice $\delta = 10^{-3}$.

As shown in Figure 2, IVSP reduces the validation loss more rapidly than the compared methods and attains the lowest validation loss and lowest test error on both datasets. The estimated Moreau-gap constraint violation, which estimates $[R(w)]_+$, is driven essentially to zero on MNIST and remains



(a) Data hyper-cleaning: MNIST



(b) Data hyper-cleaning: FashionMNIST

Figure 2: Comparison results for data hyper-cleaning on MNIST and FashionMNIST. The third panels report the estimated Moreau-gap constraint violation.

small on FashionMNIST, although PNGBiO and MEHA reach zero violation there, consistent with their unrelaxed Moreau-gap feasibility target.

Few-shot learning. We consider 10-way 1-shot classification on Omniglot (Lake et al., 2015) following the bilevel meta-learning setting of Liu et al. (2024), with the goal of assessing whether the learned representation transfers effectively to unseen tasks. The UL variable parameterizes a shared Conv-4 feature extractor, while the LL variables are task-specific linear classifiers, yielding a convex but not strongly convex LL problem. Proposition 16 shows that ENNAMCQ holds for every $\delta > 0$. Hyperparameters are selected using only meta-validation query loss.

As shown in Figure 3, IVSP attains the lowest validation and test query losses as well as the lowest test classification error. The similar validation and test behavior suggests that the selected configuration generalizes to the held-out test classes in this experiment.

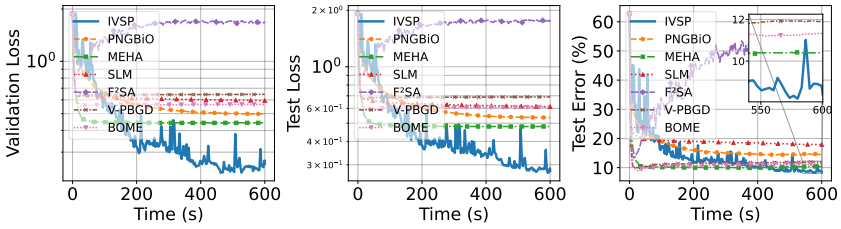


Figure 3: Comparison results for few-shot learning on Omniglot.

5 Conclusion

We developed IVSP, a deterministic first-order method for bilevel optimization with weakly convex lower-level objectives, using a positive Moreau-gap relaxation and a smoothed exact-penalty approach. Under ENNAMCQ, IVSP achieves finite penalty stabilization and an overall $\widetilde{O}(\varepsilon^{-3})$ first-order complexity for computing an ε-KKT point, without strong convexity, PL, or error-bound assumptions. Numerical experiments illustrate its competitive performance. Future work includes improving the complexity and further weakening the ENNAMCQ-type conditions required for finite penalty stabilization.

AI USE STATEMENT

ChatGPT (OpenAI), using the GPT-5.5 and GPT-5.6 Sol models over the course of the project, was used as an auxiliary tool in parts of the theoretical and experimental workflow, as well as in manuscript preparation. In the theoretical work, ChatGPT was used to assist with checking calculations, derivations, proof arguments, and error terms arising from inexact computations. For the numerical experiments, it was used to assist with code implementation and debugging, as well as with developing parameter-tuning strategies. ChatGPT was also used to help identify potentially relevant literature and to assist with restructuring, editing, and polishing parts of the manuscript for clarity and readability. All AI-assisted mathematical arguments, references, code, parameter choices, and reported results were reviewed and verified by the authors. The authors take full responsibility for the final content of the manuscript.

ETHICS STATEMENT

This work is primarily theoretical and does not involve human subjects, personally identifiable information, or sensitive data. The numerical experiments use standard publicly available datasets.

REPRODUCIBILITY STATEMENT

The assumptions, algorithmic details, and complete proofs of the theoretical results are provided in the main text and appendices. Detailed formulations of the experimental problems, implementation settings, and hyperparameter choices are provided in the appendix. Source code for the numerical experiments is included in the supplementary material.

REFERENCES

Jan Harold Alcantara and Akiko Takeda. Theoretical smoothing frameworks for nonsmooth simple bilevel problems. *Mathematics of Operations Research*, 2026.

Xiaoning Bai, Shangzhi Zeng, Jin Zhang, and Lezhi Zhang. Alternating gradient-type algorithm for bilevel optimization with inexact lower-level solutions via moreau envelope-based reformulation. *SIAM Journal on Optimization*, 36(1):350–380, 2026.

Jonathan Barzilai and Jonathan M. Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141–148, 1988.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2 edition, 2017.

Amir Beck. *First-Order Methods in Optimization*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

Amir Beck and Marc Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 22(2):557–580, 2012.

J. Frédéric Bonnans and Alexander Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2000.

Lesi Chen, Yaohua Ma, and Jingzhao Zhang. Near-optimal nonconvex-strongly-convex bilevel optimization with fully first-order oracles. *Journal of Machine Learning Research*, 26(109):1–56, 2025.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1568–1577. PMLR, 2018.

Lucy L. Gao, Jane J. Ye, Hai-An Yin, Shang-Zhi Zeng, and Jin Zhang. Moreau envelope-based difference of weakly convex reformulation and algorithm for bilevel programs. *Journal of the Operations Research Society of China*, 2026.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3748–3758. PMLR, 2020.

Zhihao Gu, Qilong Wu, and Junchi Yang. Sgha: A single-loop fully first-order algorithm for nonconvex-strongly-convex bilevel optimization, 2026.

Feihu Huang. Optimal Hessian/Jacobian-free nonconvex-PL bilevel optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19598–19621. PMLR, 2024.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4882–4892. PMLR, 2021.

Liuyuan Jiang, Quan Xiao, Lisha Chen, and Tianyi Chen. Beyond value functions: Single-loop bilevel optimization under flatness conditions. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Abderrahim Jourani. Constraint qualifications and lagrange multipliers in nondifferentiable programming problems. *Journal of Optimization Theory and Applications*, 81(3):533–548, 1994.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. A fully first-order method for stochastic bilevel optimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18083–18113. PMLR, 2023.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D. Nowak. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. In *International Conference on Learning Representations*, pp. 47473–47518, 2024.

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Gui-Hua Lin, Mengwei Xu, and Jane J. Ye. On solving simple bilevel programs with a nonconvex lower level program. *Mathematical Programming*, 144:277–305, 2014.

Bo Liu, Mao Ye, Stephen Wright, Peter Stone, and Qiang Liu. BOME! Bilevel optimization made easy: A simple first-order approach. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17248–17262, 2022.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. In *International Conference on Learning Representations*, 2019.

Risheng Liu, Zhu Liu, Wei Yao, Shangzhi Zeng, and Jin Zhang. Moreau envelope for nonconvex bi-level optimization: A single-loop and hessian-free solution strategy. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 31566–31596. PMLR, 2024.

Songtao Lu. SLM: A smoothed first-order lagrangian method for structured constrained nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 80414–80454, 2023.

Songtao Lu. TSP: A two-sided smoothed primal-dual method for nonconvex bilevel optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40665–40708. PMLR, 2025.

Zhaosong Lu and Sanyou Mei. First-order penalty methods for bilevel optimization. *SIAM Journal on Optimization*, 34(2):1937–1969, 2024.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2 edition, 2006.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 737–746. PMLR, 2016.

Aravind Rajeswaran, Chelsea Finn, Sham M. Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4334–4343. PMLR, 2018.

R. Tyrrell Rockafellar and Roger J. B. Wets. *Variational Analysis*. Springer, Berlin, 1998.

Thabo Samakhoana and Benjamin Grimmer. The optimal smoothings of sublinear functions and convex cones. *Set-Valued and Variational Analysis*, 34(2):24, 2026.

Han Shen and Tianyi Chen. On penalty-based bilevel gradient descent method. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 30992–31015. PMLR, 2023.

Han Shen, Quan Xiao, and Tianyi Chen. On penalty-based bilevel gradient descent method. *Mathematical Programming*, 214:539–589, 2025.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Jiefeng Xu, Ting Kei Pong, and Yongle Zhang. A smoothing extended sequential quadratic method for difference-of-convex optimization over a convex composite inequality constraint. *arXiv preprint arXiv:2606.13343*, 2026.

Jane J. Ye. Multiplier rules under mixed assumptions of differentiability and lipschitz continuity. *SIAM Journal on Control and Optimization*, 39(5):1441–1460, 2000.

Jane J. Ye, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. Difference of convex algorithms for bilevel programs with applications in hyperparameter selection. *Mathematical Programming*, 198(2):1583–1616, 2023.

Siqi Zhang, Xing Huang, and Feihu Huang. Generalized smooth bilevel optimization with nonconvex lower-level. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74997–75029. PMLR, 2025.

A Mathematical Preliminaries

A.1 Moreau envelope properties

We first collect some consequences of Assumption 1 that will be used throughout the analysis. Since ∇g is L_g-Lipschitz continuous on $C = X \times Y$, there exist finite constants $L_{yx}, L_{xy}, L_{yy} > 0$ such that

$$\|\nabla_y g(x_1, y) - \nabla_y g(x_2, y)\| \le L_{yx}\|x_1 - x_2\|, \tag{11}$$
$$\|\nabla_x g(x, y_1) - \nabla_x g(x, y_2)\| \le L_{xy}\|y_1 - y_2\|, \tag{12}$$
$$\|\nabla_y g(x, y_1) - \nabla_y g(x, y_2)\| \le L_{yy}\|y_1 - y_2\|. \tag{13}$$

In particular, one may take each of these constants no larger than L_g. Recall that $\sigma \ge 0$ denotes a uniform weak-convexity modulus of $g(x, \cdot)$ and that $\rho\sigma < 1$. For convenience, define

$$m_\rho := \rho^{-1} - \sigma > 0, \qquad \kappa_\rho := \frac{1}{1 - \rho\sigma}. \tag{14}$$

Proposition 8 (Moreau envelope regularity)**.** *Suppose that Assumption 1 holds, and let m_ρ, κ_ρ be given by* (14)*. The following statements hold.*

(i) For every $w = (x, y) \in C$, the function $\psi_\rho(\cdot; x, y)$ is m_ρ-strongly convex on Y. Consequently, $z_\rho(w)$ is well defined and unique.

(ii) The Moreau envelope g_ρ is continuously differentiable on C, with

$$\nabla_x g_\rho(x, y) = \nabla_x g(x, z_\rho(w)), \qquad \nabla_y g_\rho(x, y) = \rho^{-1}\big(y - z_\rho(w)\big). \tag{15}$$

(iii) For any $w_i = (x_i, y_i) \in C$, $i = 1, 2$,

$$\|z_\rho(w_1) - z_\rho(w_2)\| \le \kappa_\rho\left(\rho L_{yx}\|x_1 - x_2\| + \|y_1 - y_2\|\right). \tag{16}$$

In particular, z_ρ is Lipschitz continuous on C.

(iv) The residual R is continuously differentiable, with

$$\nabla R(w) = \begin{pmatrix} \nabla_x g(x, y) - \nabla_x g(x, z_\rho(w)) \\ \nabla_y g(x, y) - \rho^{-1}(y - z_\rho(w)) \end{pmatrix}. \tag{17}$$

Moreover, there exist constants $L_R, B_R > 0$ such that

$$\|\nabla R(w_1) - \nabla R(w_2)\| \le L_R\|w_1 - w_2\|, \qquad \text{and} \qquad \|\nabla R(w)\| \le B_R \tag{18}$$

for all $w, w_1, w_2 \in C$.

Proof. Fix $w = (x, y) \in C$. Since $g(x, \cdot)$ is σ-weakly convex, $z \mapsto g(x, z) + \frac{\sigma}{2}\|z\|^2$ is convex. Moreover,

$$\frac{1}{2\rho}\|z - y\|^2 - \frac{\sigma}{2}\|z\|^2 = \frac{m_\rho}{2}\|z\|^2 - \rho^{-1}\langle y, z\rangle + \frac{1}{2\rho}\|y\|^2.$$

It follows that $\psi_\rho(\cdot; x, y)$ is m_ρ-strongly convex on Y. Since Y is compact and $\psi_\rho(\cdot; x, y)$ is continuous, a minimizer exists, while strong convexity gives uniqueness; see (Beck, 2017, Theorem 5.25). This proves (i).

The differentiability and gradient formulas in (ii) follow from Danskin's theorem; see Bonnans & Shapiro (2000, Theorem 4.13 and Remark 4.14).

We next prove (iii). Let $w_i = (x_i, y_i)$ and $z_i = z_\rho(w_i)$ for $i = 1, 2$. The optimality conditions of the two proximal subproblems give

$$\big\langle \nabla_y g(x_1, z_1) + \rho^{-1}(z_1 - y_1), z_2 - z_1 \big\rangle \ge 0,$$
$$\big\langle \nabla_y g(x_2, z_2) + \rho^{-1}(z_2 - y_2), z_1 - z_2 \big\rangle \ge 0.$$

Adding these inequalities yields

$$\rho^{-1}\langle y_1 - y_2, z_1 - z_2\rangle \ge \langle \nabla_y g(x_1, z_1) - \nabla_y g(x_2, z_2), z_1 - z_2\rangle + \rho^{-1}\|z_1 - z_2\|^2. \tag{19}$$

On the other hand,

$$\begin{aligned}\langle \nabla_y g(x_1,z_1) - \nabla_y g(x_2,z_2), z_1 - z_2\rangle &= \langle \nabla_y g(x_1,z_1) - \nabla_y g(x_1,z_2), z_1 - z_2\rangle \\ &\quad + \langle \nabla_y g(x_1,z_2) - \nabla_y g(x_2,z_2), z_1 - z_2\rangle \\ &\geq -\sigma\|z_1 - z_2\|^2 - L_{yx}\|x_1 - x_2\|\|z_1 - z_2\|.\end{aligned}$$

For the first term above, we used the σ-weak convexity of $g(x_1,\cdot)$. For the second term, we used Cauchy–Schwarz along with (11).

Together with (19), we have

$$-\sigma\|z_1 - z_2\|^2 - L_{yx}\|x_1 - x_2\|\|z_1 - z_2\| + \rho^{-1}\|z_1 - z_2\|^2 \leq \rho^{-1}\langle y_1 - y_2, z_1 - z_2\rangle.$$

Hence,

$$\begin{aligned}(\rho^{-1} - \sigma)\|z_1 - z_2\|^2 &\leq L_{yx}\|x_1 - x_2\|\|z_1 - z_2\| + \rho^{-1}\langle y_1 - y_2, z_1 - z_2\rangle \\ &\leq L_{yx}\|x_1 - x_2\|\|z_1 - z_2\| + \rho^{-1}\|y_1 - y_2\|\|z_1 - z_2\|,\end{aligned}$$

where the last inequality again follows from Cauchy–Schwarz. If $z_1 = z_2$, the claim is immediate. Otherwise, dividing by $\|z_1 - z_2\|$ yields

$$(\rho^{-1} - \sigma)\|z_1 - z_2\| \leq L_{yx}\|x_1 - x_2\| + \rho^{-1}\|y_1 - y_2\|.$$

Therefore,

$$\begin{aligned}\|z_1 - z_2\| &\leq \frac{L_{yx}\|x_1 - x_2\| + \rho^{-1}\|y_1 - y_2\|}{\rho^{-1} - \sigma} \\ &= \frac{\rho L_{yx}\|x_1 - x_2\| + \|y_1 - y_2\|}{1 - \rho\sigma} \\ &= \kappa_\rho\left(\rho L_{yx}\|x_1 - x_2\| + \|y_1 - y_2\|\right),\end{aligned}$$

which proves (16).

Formula (17) follows directly from (15) and the definition of R. We next verify the Lipschitz continuity of ∇R. By (16) and Cauchy–Schwarz, there exists $K_z > 0$ such that

$$\|z_\rho(w_1) - z_\rho(w_2)\| \leq K_z\|w_1 - w_2\|, \qquad w_1, w_2 \in C.$$

Hence, the map $T(w) \coloneqq (x, z_\rho(w))$ is Lipschitz on C. Indeed,

$$\|T(w_1) - T(w_2)\|^2 = \|x_1 - x_2\|^2 + \|z_\rho(w_1) - z_\rho(w_2)\|^2 \leq (1 + K_z^2)\|w_1 - w_2\|^2.$$

Since ∇g is L_g-Lipschitz on C, it follows that $w \mapsto \nabla g(x, z_\rho(w))$ is Lipschitz on C. Therefore, each term in (17) is Lipschitz continuous in w, and hence there exists $L_R > 0$ such that

$$\|\nabla R(w_1) - \nabla R(w_2)\| \leq L_R\|w_1 - w_2\|, \qquad w_1, w_2 \in C.$$

Finally, C is compact and ∇R is continuous, so $B_R \coloneqq \max_{w\in C}\|\nabla R(w)\| < \infty$. This completes the proof of (iv). □

A.2 Smoothing estimates

We use a standard family of smooth approximations of the plus function (Beck & Teboulle, 2012). For every $\mu > 0$, let ϕ_μ be convex, continuously differentiable, and nondecreasing, and suppose that there exist constants $a_1, a_2, b_\phi, c_\phi \geq 0$ such that

$$0 \leq \phi_\mu'(t) \leq 1, \tag{20}$$

$$[t]_+ - a_1\mu \leq \phi_\mu(t) \leq [t]_+ + a_2\mu, \tag{21}$$

$$|\phi_\mu'(t) - \phi_\mu'(s)| \leq \frac{b_\phi}{\mu}|t - s|, \tag{22}$$

$$\phi_{\mu_1}(t) \leq \phi_{\mu_0}(t) + c_\phi(\mu_0 - \mu_1), \qquad 0 < \mu_1 \leq \mu_0. \tag{23}$$

$$a_\phi \coloneqq a_1 + a_2. \tag{24}$$

The next lemma makes precise the $O(1/\mu)$ smoothness property of G_μ used in Section 2.2.

Lemma 9. *Under Assumption 1 and the above conditions on ϕ_μ, the function $G_\mu = \phi_\mu \circ R$ is continuously differentiable and $\nabla G_\mu(w) = \phi'_\mu(R(w))\nabla R(w)$. Moreover, for every $\mu \in (0, \mu_0]$,*

$$\|\nabla G_\mu(w) - \nabla G_\mu(\bar{w})\| \le \frac{L_G}{\mu}\|w - \bar{w}\|, \qquad w, \bar{w} \in C,$$

where $L_G \coloneqq b_\phi B_R^2 + \mu_0 L_R$, with B_R and L_R given in (18)*. Furthermore, whenever $0 < \mu_1 \le \mu_0$,*

$$G_{\mu_1}(w) \le G_{\mu_0}(w) + c_\phi(\mu_0 - \mu_1), \qquad w \in C. \tag{25}$$

Proof. The gradient formula follows from the chain rule. The remaining estimates also follow as special cases of Xu et al. (2026, Lemma 2.3(i),(iii)); we provide a direct proof here for completeness. For $w, \bar{w} \in C$,

$$\|\nabla G_\mu(w) - \nabla G_\mu(\bar{w})\| \le |\phi'_\mu(R(w)) - \phi'_\mu(R(\bar{w}))|\|\nabla R(w)\| + |\phi'_\mu(R(\bar{w}))|\|\nabla R(w) - \nabla R(\bar{w})\|.$$

By Proposition 8, $\|\nabla R\| \le B_R$ and ∇R is L_R-Lipschitz. Hence $|R(w) - R(\bar{w})| \le B_R\|w - \bar{w}\|$. Using $0 \le \phi'_\mu \le 1$ by (20) and the (b_ϕ/μ)-Lipschitz continuity of ϕ'_μ by (22) gives

$$\|\nabla G_\mu(w) - \nabla G_\mu(\bar{w})\| \le \left(\frac{b_\phi B_R^2}{\mu} + L_R\right)\|w - \bar{w}\|.$$

Since $\mu \le \mu_0$, the right-hand side is bounded by $(L_G/\mu)\|w - \bar{w}\|$.

Finally, (23) directly gives

$$G_{\mu_1}(w) = \phi_{\mu_1}(R(w)) \le \phi_{\mu_0}(R(w)) + c_\phi(\mu_0 - \mu_1),$$

which proves (25). □

The following estimates correspond to bounds used in the proof of Xu et al. (2026, Lemma 3.2(ii)–(iii)); we give a direct proof under our notation for completeness.

Lemma 10 (Smoothing derivative estimates)**.** *Let a_ϕ be given by* (24) *and let $\theta > 1$. For every $\mu > 0$ and $t \in \mathbb{R}$, the following hold:*

(i) If $t > \theta a_\phi \mu$, then $\phi'_\mu(t) \ge \frac{\theta - 1}{\theta}$;

(ii) If $t \le \theta a_\phi \mu$, then $|\phi'_\mu(t)t| \le \theta a_\phi \mu$.

Proof. For (i), convexity of ϕ_μ gives

$$t\phi'_\mu(t) \ge \phi_\mu(t) - \phi_\mu(0). \tag{26}$$

For $t > 0$, the bound (21) yields $\phi_\mu(t) \ge t - a_1\mu$, and $\phi_\mu(0) \le a_2\mu$. Therefore,

$$\phi'_\mu(t) \ge 1 - \frac{(a_1 + a_2)\mu}{t} = 1 - \frac{a_\phi \mu}{t}.$$

If $t > \theta a_\phi \mu$, then $\frac{a_\phi \mu}{t} < \frac{1}{\theta}$, and hence the claim of part (i).

For (ii), first suppose $0 \le t \le \theta a_\phi \mu$. Since $0 \le \phi'_\mu(t) \le 1$, we have $|\phi'_\mu(t)t| \le t \le \theta a_\phi \mu$.

Now suppose $t < 0$. By (26), $(-t)\phi'_\mu(t) \le \phi_\mu(0) - \phi_\mu(t)$. Since $[t]_+ = 0$ for $t < 0$, (21) gives $\phi_\mu(0) \le a_2\mu$, and $\phi_\mu(t) \ge -a_1\mu$. Consequently, $(-t)\phi'_\mu(t) \le (a_1 + a_2)\mu = a_\phi \mu$. Since $t < 0$ and $\phi'_\mu(t) \ge 0$,

$$|\phi'_\mu(t)t| = (-t)\phi'_\mu(t) \le a_\phi \mu \le \theta a_\phi \mu.$$

This proves (ii). □

A.3 Error bounds for inexact Moreau evaluations

We next give the explicit error bounds underlying (6). Define

$$L_{a,\rho} \coloneqq \sqrt{L_{xy}^2 + \rho^{-2}}, \quad D_\rho(\zeta) \coloneqq L_{a,\rho}\sqrt{\frac{2\zeta}{m_\rho}}, \quad \text{and} \quad E_{\mu,\rho}(\zeta) \coloneqq D_\rho(\zeta) + \frac{b_\phi B_R}{\mu}\zeta, \tag{27}$$

with B_R given in (18).

Lemma 11 (Errors from an inexact Moreau evaluation)**.** *Let $w = (x, y) \in C$, $\zeta \geq 0$, and let $\widetilde{z} \in Y$ satisfy*

$$0 \leq \psi_\rho(\widetilde{z}; w) - g_\rho(w) \leq \zeta. \tag{28}$$

Define

$$\widetilde{g}_\rho(w) \coloneqq \psi_\rho(\widetilde{z}; w), \qquad \widetilde{R}(w) \coloneqq g(w) - \widetilde{g}_\rho(w) - \delta, \tag{29}$$

and

$$\widetilde{\nabla} R(w) \coloneqq \begin{pmatrix} \nabla_x g(x, y) - \nabla_x g(x, \widetilde{z}) \\ \nabla_y g(x, y) - \rho^{-1}(y - \widetilde{z}) \end{pmatrix}. \tag{30}$$

Then the following estimates hold:

(i) $\|\widetilde{z} - z_\rho(w)\| \leq \sqrt{\frac{2\zeta}{m_\rho}}$.

(ii) We have

$$\widetilde{R}(w) \leq R(w) \leq \widetilde{R}(w) + \zeta. \tag{31}$$

Consequently,

$$\phi_\mu(\widetilde{R}(w)) \leq G_\mu(w) \leq \phi_\mu(\widetilde{R}(w) + \zeta) \tag{32}$$

and

$$0 \leq \phi_\mu(\widetilde{R}(w) + \zeta) - \phi_\mu(\widetilde{R}(w)) \leq \zeta. \tag{33}$$

(iii) $\|\widetilde{\nabla} R(w) - \nabla R(w)\| \leq D_\rho(\zeta)$, *where $D_\rho(\zeta)$ is given in* (27)*.*

(iv) With $\widetilde{\nabla} G_\mu(w) = \phi'_\mu(\widetilde{R}(w))\widetilde{\nabla} R(w)$, *we have* $\|\widetilde{\nabla} G_\mu(w) - \nabla G_\mu(w)\| \leq E_{\mu,\rho}(\zeta)$, *where $E_{\mu,\rho}(\zeta)$ is defined in* (27)*.*

Proof. Since $\psi_\rho(\cdot; w)$ is m_ρ-strongly convex and $z_\rho(w)$ is its minimizer, we have from (Beck, 2017, Theorem 5.25) that

$$\psi_\rho(\widetilde{z}; w) \geq \psi_\rho(z_\rho(w); w) + \frac{m_\rho}{2}\|\widetilde{z} - z_\rho(w)\|^2. \tag{34}$$

Combining this with (28) and noting that $\psi_\rho(z_\rho(w); w) = g_\rho(w)$ give (i).

For part (ii), let $e \coloneqq \widetilde{g}_\rho(w) - g_\rho(w)$. Note that $e \in [0, \zeta]$ due to (28) and that $e = R(w) - \widetilde{R}(w)$ by noting (29). Hence, (31) holds. Since ϕ_μ is nondecreasing, this immediately yields (32). Moreover, $0 \leq \phi'_\mu \leq 1$, and therefore ϕ_μ is 1-Lipschitz. Hence (33) follows.

For (iii), subtract (17) from (30):

$$\widetilde{\nabla} R(w) - \nabla R(w) = \begin{pmatrix} \nabla_x g(x, z_\rho(w)) - \nabla_x g(x, \widetilde{z}) \\ \rho^{-1}(\widetilde{z} - z_\rho(w)) \end{pmatrix}.$$

Using (12) and part (i),

$$\|\widetilde{\nabla} R(w) - \nabla R(w)\| \leq \sqrt{L_{xy}^2 + \rho^{-2}}\,\|\widetilde{z} - z_\rho(w)\| \leq L_{a,\rho}\sqrt{\frac{2\zeta}{m_\rho}} = D_\rho(\zeta),$$

which proves (iii).

Finally, denote $\widetilde{R} = \widetilde{R}(w)$ and $R = R(w)$. Then

$$\begin{aligned} \|\widetilde{\nabla} G_\mu(w) - \nabla G_\mu(w)\| &= \|\phi'_\mu(\widetilde{R})\widetilde{\nabla} R(w) - \phi'_\mu(R)\nabla R(w)\| \\ &\leq |\phi'_\mu(\widetilde{R})|\|\widetilde{\nabla} R(w) - \nabla R(w)\| + |\phi'_\mu(\widetilde{R}) - \phi'_\mu(R)|\|\nabla R(w)\|. \end{aligned}$$

Together with $0 \le \phi'_\mu \le 1$, the (b_ϕ/μ)-Lipschitz continuity of ϕ'_μ, (31), and $\|\nabla R(w)\| \le B_R$, we obtain

$$\|\widetilde{\nabla} G_\mu(w) - \nabla G_\mu(w)\| \le D_\rho(\zeta) + \frac{b_\phi B_R}{\mu}\zeta = E_{\mu,\rho}(\zeta).$$

This proves (iv). □

A.4 Properties of the positive Moreau-gap relaxation

This subsection clarifies the role of the positive relaxation parameter $\delta > 0$ in (1) and the relation between the Moreau-gap constraint and lower-level stationarity.

Lower-level interpretation of the relaxation. Define

$$H_\rho(x,y) \coloneqq g(x,y) - g_\rho(x,y), \qquad \mathcal{F}_\delta \coloneqq \{(x,y) \in C : H_\rho(x,y) \le \delta\}. \tag{35}$$

Thus, $\mathcal{F}_\delta$ is the feasible set of (1), while $\mathcal{F}_0$ corresponds to the exact Moreau-gap constraint. We also define

$$\widetilde{S}(x) \coloneqq \{y \in Y : 0 \in \nabla_y g(x,y) + N_Y(y)\} \tag{36}$$

as the lower-level stationary-point mapping.

The Moreau envelope reformulation was introduced by Gao et al. (2026). Under convexity of the lower-level problem, the exact Moreau-gap constraint is equivalent to lower-level global optimality (Gao et al., 2026). For weakly convex lower-level objectives, the exact constraint instead characterizes lower-level first-order stationarity; see Liu et al. (2024). The proposition below recalls this equivalence and additionally gives quantitative bounds relating the Moreau gap to lower-level near-stationarity.

Proposition 12 (Stationarity properties of the Moreau gap)**.** *Suppose that $g(x,\cdot)$ is σ-weakly convex and $\rho\sigma < 1$. Let H_ρ and $\mathcal{F}_\delta$ be given by* (35)*. Then, for every $(x,y) \in C$,*

$$H_\rho(x,y) \ge \frac{1-\rho\sigma}{2\rho}\|y - z_\rho(x,y)\|^2. \tag{37}$$

Moreover,

$$\operatorname{dist}(0, \nabla_y g(x, z_\rho(x,y)) + N_Y(z_\rho(x,y))) \le \frac{1}{\rho}\|y - z_\rho(x,y)\| \le \sqrt{\frac{2H_\rho(x,y)}{\rho(1-\rho\sigma)}}. \tag{38}$$

In particular,

$$H_\rho(x,y) = 0 \iff y = z_\rho(x,y) \iff 0 \in \nabla_y g(x,y) + N_Y(y). \tag{39}$$

Hence,

$$\mathcal{F}_0 = \{(x,y) \in C : y \in \widetilde{S}(x)\}, \tag{40}$$

where $\widetilde{S}$ is given by (36)*.*

Proof. By strong convexity of $\psi_\rho(\cdot; x, y)$, as in (34), we have

$$\psi_\rho(y; x, y) - \psi_\rho(z_\rho(x,y); x, y) \ge \frac{m_\rho}{2}\|y - z_\rho(x,y)\|^2.$$

Since $\psi_\rho(y; x, y) = g(x,y)$ and $\psi_\rho(z_\rho(x,y); x, y) = g_\rho(x,y)$, this gives (37).

Since $z_\rho(x,y)$ minimizes $\psi_\rho(\cdot, x, y) = g(x,\cdot) + \frac{1}{2\rho}\|\cdot - y\|^2$, we have

$$\rho^{-1}(y - z_\rho(x,y)) \in \nabla_y g(x, z_\rho(x,y)) + N_Y(z_\rho(x,y)),$$

which proves the first inequality in (38); the second follows from (37).

Finally, (37) implies that $H_\rho(x,y) = 0$ only if $y = z_\rho(x,y)$, while the converse is immediate since $g_\rho(x,y) = g(x, z_\rho(x,y)) + \frac{1}{2\rho}\|z_\rho(x,y) - y\|^2$. By the optimality condition of the proximal subproblem, $y = z_\rho(x,y)$ holds if and only if $0 \in \nabla_y g(x,y) + N_Y(y)$. Thus, (39) holds. □

The preceding result also gives a direct interpretation of the feasibility residual used in our approximate KKT condition.

Corollary 13 (Interpretation of approximate feasibility)**.** *Let $w = (x, y) \in C$ satisfy $[R(w)]_+ \le \varepsilon_f$, where $R(w) = H_\rho(w) - \delta$. Then $H_\rho(w) \le \delta + \varepsilon_f$, and consequently*

$$\|y - z_\rho(x, y)\| \le \sqrt{\frac{2\rho(\delta + \varepsilon_f)}{1 - \rho\sigma}},$$

$$\operatorname{dist}(0, \nabla_y g(x, z_\rho(x, y)) + N_Y(z_\rho(x, y))) \le \sqrt{\frac{2(\delta + \varepsilon_f)}{\rho(1 - \rho\sigma)}}.$$

Proof. Since $R(w) \le [R(w)]_+$, we obtain $H_\rho(w) = R(w) + \delta \le \delta + \varepsilon_f$. The two estimates then follow from Proposition 12. □

Why a positive relaxation is needed for KKT analysis. The exact Moreau-gap constraint has an inherent constraint-qualification degeneracy. Recall that, for a feasible point w of a smooth inequality $R(w) \le 0$ over the closed convex set C, the *no nonzero abnormal multiplier constraint qualification* (NNAMCQ) requires that

$$0 \notin \nabla R(w) + N_C(w),$$

whenever $R(w) = 0$. For a single smooth inequality, this condition is equivalent to the Mangasarian–Fromovitz constraint qualification (MFCQ) relative to C (Jourani, 1994).

Proposition 14 (Degeneracy of the exact Moreau-gap constraint)**.** *Under Assumption 1, NNAMCQ (equivalently, MFCQ), fails at every feasible point of the exact Moreau-gap formulation $H_\rho \le 0$. On the other hand, for every $\delta > 0$, the relaxed constraint $H_\rho \le \delta$ admits a strictly feasible point.*

Proof. Since $H_\rho \ge 0$ on C, every $w \in \mathcal{F}_0$ is a global minimizer of H_ρ over C. The first-order necessary condition therefore gives $0 \in \nabla H_\rho(w) + N_C(w)$, so NNAMCQ, equivalently MFCQ, fails. For the second claim, fix any $x \in X$ and choose $y^\star \in \arg\min_{y \in Y} g(x, y)$. Then $0 \in \nabla_y g(x, y^\star) + N_Y(y^\star)$, and Proposition 12 gives $H_\rho(x, y^\star) = 0 < \delta$. □

Thus, the positive relaxation removes the unavoidable degeneracy of the exact formulation in the sense that strict feasibility becomes possible. Strict feasibility alone, however, does not guarantee the extended constraint qualification required in our convergence analysis. Sufficient conditions to guarantee ENNAMCQ are explored next in Appendix B.

Behavior as $\delta \downarrow 0$. The positive Moreau-gap relaxation is consistent with the exact constraint as the relaxation level vanishes. Indeed, since H_ρ is continuous and C is compact, the sets $\mathcal{F}_\delta = \{w \in C : H_\rho(w) \le \delta\}$ are nonempty and compact, and

$$\mathcal{F}_0 \subseteq \mathcal{F}_{\delta_1} \subseteq \mathcal{F}_{\delta_2} \qquad (0 \le \delta_1 \le \delta_2), \qquad \bigcap_{\delta > 0} \mathcal{F}_\delta = \mathcal{F}_0.$$

Consequently, $d_{\mathrm{H}}(\mathcal{F}_\delta, \mathcal{F}_0) \to 0$ as $\delta \downarrow 0$, where, for nonempty compact sets A and B,

$$d_{\mathrm{H}}(A, B) \coloneqq \max\left\{\sup_{a \in A} \operatorname{dist}(a, B), \sup_{b \in B} \operatorname{dist}(b, A)\right\}.$$

By (40), $\mathcal{F}_0$ corresponds to lower-level stationary points and, in general, need not be the feasible set of the original bilevel problem. When $\widetilde{S}(x) = S(x) \coloneqq \arg\min_{y \in Y} g(x, y)$ for every $x \in X$, in particular under lower-level convexity, the above Hausdorff convergence is to the original bilevel feasible set. This is consistent with the approximation result of Gao et al. (2026, Proposition 2), who show in the convex lower-level setting that sufficiently small positive relaxations admit local minimizers arbitrarily close to the bilevel solution set. Nevertheless, this geometric consistency does not imply uniform algorithmic complexity as $\delta \downarrow 0$; see Example 23 and the discussion following it. Accordingly, throughout our complexity analysis, $\delta > 0$ is fixed, and the resulting bounds are not claimed to be uniform as $\delta \downarrow 0$.

B SUFFICIENT CONDITIONS FOR ENNAMCQ

This section gives several sufficient conditions under which the ENNAMCQ assumption used in the main text holds. These conditions cover lower-level problems arising in many important bilevel optimization problems.

B.1 CONVEX LOWER-LEVEL PROBLEMS WITHOUT NONCONSTANT AFFINE SEGMENTS

For fixed $x \in X$, define the extended-valued lower-level objective

$$h_x(y) := g(x, y) + \iota_Y(y), \tag{41}$$

where ι_Y denotes the indicator function of Y.

We consider the following property:

$$\begin{aligned} &\text{if } y_0, y_1 \in Y \text{ and } t \mapsto g\big(x, (1-t)y_0 + ty_1\big) \text{ is affine on } [0,1], \\ &\text{then } g\big(x, (1-t)y_0 + ty_1\big) \text{ is constant on } [0,1]. \end{aligned} \tag{42}$$

In other words, affine segments are allowed, but only when the lower-level objective is constant along them. This condition is strictly weaker than strict convexity.

Proposition 15 (Automatic ENNAMCQ without nonconstant affine segments)**.** *Let $\delta > 0$, and suppose the standing smoothness assumptions hold. Fix $\bar{x} \in X$. Suppose that $g(\bar{x}, \cdot)$ is convex on Y and satisfies* (42)*. Then $0 \notin \nabla R(\bar{x}, y) + N_C(\bar{x}, y)$ for every $y \in Y$ such that $R(\bar{x}, y) \geq 0$. Consequently, if* (42) *holds for every $x \in X$, then ENNAMCQ holds on $X \times Y$.*

In particular, if $g(x, \cdot) + \iota_Y$ is strictly convex for all $x \in X$, then ENNAMCQ holds on $X \times Y$.

Proof. Fix $y \in Y$ with $R(\bar{x}, y) \geq 0$ and suppose, to the contrary, that

$$0 \in \nabla R(\bar{x}, y) + N_C(\bar{x}, y). \tag{43}$$

Let $z := z_\rho(\bar{x}, y)$, and $p := \rho^{-1}(y - z)$. Since $C = X \times Y$, taking the y-component of (43) and using $\nabla_y R(\bar{x}, y) = \nabla_y g(\bar{x}, y) - \rho^{-1}(y - z)$ from (17) gives

$$p \in \nabla_y g(\bar{x}, y) + N_Y(y) = \partial h_{\bar{x}}(y), \tag{44}$$

where $h_{\bar{x}}$ is defined as in (41). On the other hand, the optimality condition for the proximal subproblem defining z gives $0 \in \nabla_y g(\bar{x}, z) + \rho^{-1}(z - y) + N_Y(z)$, and therefore

$$p \in \nabla_y g(\bar{x}, z) + N_Y(z) = \partial h_{\bar{x}}(z). \tag{45}$$

Since $h_{\bar{x}}$ is convex, (44)–(45) yield

$$\begin{aligned} h_{\bar{x}}(z) &\geq h_{\bar{x}}(y) + \langle p, z - y \rangle, \\ h_{\bar{x}}(y) &\geq h_{\bar{x}}(z) + \langle p, y - z \rangle. \end{aligned}$$

Thus, the two inequalities must both hold with equality. Hence

$$h_{\bar{x}}(z) = h_{\bar{x}}(y) + \langle p, z - y \rangle. \tag{46}$$

For $t \in [0, 1]$, let $y_t = (1 - t)y + tz$. By convexity,

$$h_{\bar{x}}(y_t) \leq (1 - t)h_{\bar{x}}(y) + th_{\bar{x}}(z). \tag{47}$$

Using $p \in \partial h_{\bar{x}}(y)$ and (46),

$$h_{\bar{x}}(y_t) \geq h_{\bar{x}}(y) + t\langle p, z - y \rangle = (1 - t)h_{\bar{x}}(y) + th_{\bar{x}}(z).$$

Together with (47), this shows that $h_{\bar{x}}$, and hence $g(\bar{x}, \cdot)$, is affine on the segment $[y, z]$. By assumption (42), it must therefore be constant on this segment, so $h_{\bar{x}}(y) = h_{\bar{x}}(z)$. Combining this with (46) gives

$$0 = \langle p, y - z \rangle = \rho^{-1}\|y - z\|^2,$$

and so $y = z$. Therefore $H_\rho(\bar{x}, y) = 0$ and consequently $R(\bar{x}, y) = -\delta < 0$, contradicting $R(\bar{x}, y) \geq 0$. This completes the proof. □

We provide an application to a non-strictly-convex lower-level problem used in few-shot learning; see Appendix F.

Proposition 16 (ENNAMCQ for few-shot learning)**.** *Let $T, m, p \in \mathbb{N}$. For each task $i = 1, \dots, T$, let $S_i = \{(u_{i,r}, q_{i,r})\}_{r=1}^{n_i}$, with $q_{i,r} \in \{1, \dots, m\}$, be a finite support set. For fixed $x \in X$ and a finite dataset D, let $F_{x,D}(u) \in \mathbb{R}^p$ denote the feature vector produced for u by the shared feature extractor when its batch-dependent statistics are computed from D. For each i and r, set $F_{i,r}(x) := F_{x,S_i}(u_{i,r})$. Let $(W_i, b_i) \in \mathbb{R}^{m\times p} \times \mathbb{R}^m$ be the task-specific linear classifier, and set $y = ((W_i, b_i))_{i=1}^T \in Y$, where $Y \subseteq \prod_{i=1}^T (\mathbb{R}^{m\times p} \times \mathbb{R}^m)$ is convex. Define $\ell(s, q) := \log\left(\sum_{j=1}^m e^{s_j}\right) - s_q$, for $s = (s_1, \dots, s_m) \in \mathbb{R}^m$, and for all $y = ((W_i, b_i))_{i=1}^T \in Y$,*

$$g(x, y) := \frac{1}{T}\sum_{i=1}^{T} \frac{1}{n_i}\sum_{r=1}^{n_i} \ell\big(W_i F_{i,r}(x) + b_i, q_{i,r}\big).$$

Then, for every fixed $x \in X$, $g(x, \cdot)$ is convex on Y and satisfies (42)*. Consequently, ENNAMCQ holds for every $\delta > 0$.*

Proof. Fix x and $y_0, y_1 \in Y$. For $\nu = 0, 1$, write $y_\nu = \big((W_i^{[\nu]}, b_i^{[\nu]})\big)_{i=1}^T$, let $y_t = (1-t)y_0 + ty_1 = \big((W_i(t), b_i(t))\big)_{i=1}^T$, where $W_i(t) := (1-t)W_i^{[0]} + tW_i^{[1]}$, and $b_i(t) := (1-t)b_i^{[0]} + tb_i^{[1]}$, and write $\Delta W_i := W_i^{[1]} - W_i^{[0]}$, $\Delta b_i := b_i^{[1]} - b_i^{[0]}$. For $(u, q) \in S_i$, set $d_{i,r} := \Delta W_i F_{i,r}(x) + \Delta b_i$, and let $\varphi(t) := g(x, y_t)$.

For the softmax vector $p_j(s) := e^{s_j}/\sum_{k=1}^m e^{s_k}$, a direct calculation gives $\nabla^2\ell(s, q) = \mathrm{Diag}(p(s)) - p(s)p(s)^\top$ and, for any $d \in \mathbb{R}^m$,

$$d^\top \nabla^2 \ell(s, q) d = \sum_{j=1}^m p_j(s)\left(d_j - \sum_{k=1}^m p_k(s) d_k\right)^2 \geq 0.$$

Since $p_j(s) > 0$ for every j, equality holds if and only if $d = \alpha\mathbf{1}$ for some $\alpha \in \mathbb{R}$, where $\mathbf{1}$ is a vector of ones. Meanwhile, a direct calculation shows

$$\varphi''(t) = \frac{1}{T}\sum_{i=1}^{T} \frac{1}{n_i}\sum_{r=1}^{n_i} d_{i,r}^\top \nabla^2 \ell\big(W_i(t)F_{i,r}(x) + b_i(t), q_{i,r}\big) d_{i,r} \geq 0,$$

so $g(x, \cdot)$ is convex. If φ is affine on $[0, 1]$, then $\varphi''(t) = 0$ on $(0, 1)$. From the equivalence noted above, we have for each $(u, q) \in S_i$ that $d_{i,r} = \alpha_{i,r}\mathbf{1}$ for some $\alpha_{i,r} \in \mathbb{R}$. Thus,

$$W_i(t)F_{i,r}(x) + b_i(t) = W_i^{[0]}F_{i,r}(x) + b_i^{[0]} + t\alpha_{i,r}\mathbf{1}.$$

Moreover, for any $\beta \in \mathbb{R}$, $\ell(s + \beta\mathbf{1}, q) = \log(e^\beta \sum_j e^{s_j}) - s_q - \beta = \ell(s, q)$. Therefore, taking $\beta = t\alpha_{i,r}$,

$$\ell\big(W_i(t)F_{i,r}(x) + b_i(t), q_{i,r}\big) = \ell\big(W_i^{[0]}F_{i,r}(x) + b_i^{[0]}, q_{i,r}\big)$$

for every $t \in [0, 1]$, and hence

$$\varphi(t) = \frac{1}{T}\sum_{i=1}^{T} \frac{1}{n_i}\sum_{r=1}^{n_i} \ell\big(W_i^{[0]}F_{i,r}(x) + b_i^{[0]}, q_{i,r}\big) = \varphi(0).$$

Thus (42) holds, and Proposition 15 yields ENNAMCQ. □

B.2 Nonconvex lower level problem with sample reweighting

We next give a sufficient condition for ENNAMCQ for a specific class of nonconvex smooth loss functions that does not require convexity of the lower-level objective.

Proposition 17 (ENNAMCQ for sample-reweighting models)**.** *Suppose the standing assumptions hold. Let $X = [-M, M]^n$ and let Y be a compact set. Suppose*

$$g(x, y) = \frac{1}{n}\sum_{i=1}^{n} \omega(x_i)\ell_i(y), \qquad \omega \geq 0, \quad \omega' > 0 \text{ on } [-M, M], \tag{48}$$

where each ℓ_i is continuous on Y. Define $D_\ell := \max_{1\leq i\leq n}\left(\max_{y\in Y} \ell_i(y) - \min_{y\in Y} \ell_i(y)\right)$. If $\delta > \omega(-M)D_\ell$, then ENNAMCQ holds on $X \times Y$.

Proof. Suppose ENNAMCQ fails at (x, y) with $R(x, y) \geq 0$, and let $z := z_\rho(x, y)$ and $\Delta_i := \ell_i(y) - \ell_i(z)$. From (48), we have $\frac{\partial R}{\partial x_i}(x, y) = \frac{\omega'(x_i)}{n}\Delta_i$. Noting (17), the x_i-component of $0 \in \nabla R(x, y) + N_{X\times Y}(x, y)$ therefore gives

$$0 \in \frac{\omega'(x_i)}{n}\Delta_i + N_{[-M,M]}(x_i).$$

Since $\omega'(x_i) > 0$, the above together with the definition of normal cone gives

$$\Delta_i = \begin{cases} 0, & -M < x_i < M, \\ \geq 0, & x_i = -M, \\ \leq 0, & x_i = M. \end{cases}$$

Thus, using $\omega \geq 0$,

$$\begin{aligned} g(x, y) - g(x, z) &= \frac{1}{n}\sum_{i=1}^{n} \omega(x_i)\Delta_i \\ &= \frac{1}{n} \sum_{\{i:x_i=M \text{ or } -M\}} \omega(x_i)\Delta_i \\ &\leq \frac{\omega(-M)}{n} \sum_{\{i:x_i=-M\}} \Delta_i \\ &\leq \omega(-M)D_\ell. \end{aligned}$$

Consequently,

$$R(x, y) = g(x, y) - g(x, z) - \frac{1}{2\rho}\|z - y\|^2 - \delta \leq \omega(-M)D_\ell - \delta < 0,$$

where the last inequality holds from the choice of δ. Since $R(x, y) \geq 0$, we arrived at a contradiction. This completes the proof. □

A relevant specialization of Proposition 17 is obtained for sigmoid sample weights and multiclass cross-entropy loss. The data hyper-cleaning problems in our experiments are instances of this setting; see Appendix F.

Corollary 18 (ENNAMCQ for sigmoid-reweighted classification)**.** *Let $M, B_y > 0$, $n, m, d, d_h \in \mathbb{N}$, $X = [-M, M]^n$, and let $u_i \in \mathbb{R}^d$ and $q_i \in \{1, \dots, m\}$, $i = 1, \dots, n$. Let $\omega(t) := (1 + e^{-t})^{-1}$ denote the sigmoid sample-weight function. For $y = (A, a, V, b) \in Y \subseteq \mathbb{R}^{d_h\times d} \times \mathbb{R}^{d_h} \times \mathbb{R}^{m\times d_h} \times \mathbb{R}^m$, define $h_y(u) := V \operatorname{sigm}(Au+a)+b$, where* sigm *denotes the logistic sigmoid applied componentwise, and set $\ell_i(y) := \ell(h_y(u_i), q_i)$, where ℓ is defined in Proposition 16. Suppose Y is nonempty, compact and convex, and every component of V and b lies in $[-B_y, B_y]$, and let g be given by* (48)*. If $\delta > \big(2B_y(d_h + 1) + \log m\big)/(1 + e^M)$, then ENNAMCQ holds on $X \times Y$.*

Proof. Since $\omega(t) > 0$ and $\omega'(t) = \omega(t)(1 - \omega(t)) > 0$, Proposition 17 applies. Since $0 < \operatorname{sigm}(t) < 1$ for every $t \in \mathbb{R}$, for every $j = 1, \dots, m$, $|[h_y(u)]_j| \leq \sum_{\ell=1}^{d_h} |V_{j\ell}|\,|\operatorname{sigm}(Au + a)_\ell| + |b_j| \leq d_h B_y + B_y = B_y(d_h + 1)$. Hence, for every $q \in \{1, \dots, m\}$,

$$0 \leq \ell(h_y(u), q) \leq 2B_y(d_h + 1) + \log m,$$

and therefore $D_\ell \leq 2B_y(d_h + 1) + \log m$. Moreover, $\omega(-M) = (1 + e^M)^{-1}$, so

$$\omega(-M)D_\ell \leq \frac{2B_y(d_h + 1) + \log m}{1 + e^M} < \delta.$$

The conclusion follows from Proposition 17. □

Algorithm 2 Inexact Moreau evaluation $\mathrm{ME}(w, z^{\mathrm{init}}, \zeta)$

Require. A point $w = (x, y) \in C$, an initial point $z^{\mathrm{init}} \in Y$, and a tolerance $\zeta > 0$. Set $\eta = L_{\psi,\rho}^{-1}$, $z^0 = z^{\mathrm{init}}$, and $j = 0$.

Step 1. Compute $z^{j+1} = P_Y\big(z^j - \eta\left[\nabla_y g(x, z^j) + \rho^{-1}(z^j - y)\right]\big)$ and set $r^j := \frac{z^j - z^{j+1}}{\eta}$.

Step 2. If $\|r^j\|^2 \le 2m_\rho\zeta$, set $\widetilde{z} := z^{j+1}$ and go to **Step 3**. Otherwise, set $j \leftarrow j + 1$ and return to **Step 1**.

Step 3. Form $\widetilde{R}(w) = g(x, y) - g(x, \widetilde{z}) - \frac{1}{2\rho}\|\widetilde{z} - y\|^2 - \delta$. Return $\widetilde{z}$ and $\widetilde{R}(w)$.

C Complete algorithm

We first describe the inexact Moreau evaluation used by IVSP. Since the proximal objective-gap condition (4) involves the unknown value $g_\rho(x, y)$, we use projected gradient and terminate it using the computable residual criterion established below. The resulting routine is given in Algorithm 2.

Lemma 19 (Residual-based stopping rule for the Moreau evaluation)**.** *Let $L_{\psi,\rho} := L_{yy} + \rho^{-1}$, where L_{yy} is the Lipschitz constant in* (13)*. Given $w = (x, y) \in C$, define $h(\cdot) := \psi_\rho(\cdot; x, y)$. Let $u \in Y$, let $0 < \eta \le 2/L_{\psi,\rho}$, and define*

$$z^+ := P_Y(u - \eta\nabla h(u)) \qquad \textit{and} \qquad r_\eta(u) := \frac{u - z^+}{\eta}.$$

Then $0 \le h(z^+) - g_\rho(x, y) \le \frac{1}{2m_\rho}\|r_\eta(u)\|^2$, where m_ρ is given by (14) *Consequently,*

$$\|r_\eta(u)\|^2 \le 2m_\rho\zeta \tag{49}$$

guarantees that z^+ is ζ-accurate in the sense of (4)*.*

Proof. Let $z_\rho = z_\rho(w)$. By the optimality condition for the projection defining z^+ and noting that $z_\rho \in Y$,

$$\langle\nabla h(u), z^+ - z_\rho\rangle \le \langle r_\eta(u), z^+ - z_\rho\rangle. \tag{50}$$

By the $L_{\psi,\rho}$-smoothness of h, we have from (Beck, 2017, Lemma 5.7) that

$$h(z^+) \le h(u) + \langle\nabla h(u), z^+ - u\rangle + \frac{L_{\psi,\rho}}{2}\|z^+ - u\|^2.$$

On the other hand, the m_ρ-strong convexity of h (Proposition 8(i)) gives

$$h(z_\rho) \ge h(u) + \langle\nabla h(u), z_\rho - u\rangle + \frac{m_\rho}{2}\|z_\rho - u\|^2.$$

Subtracting the latter inequality from the former and using (50) yields

$$h(z^+) - h(z_\rho) \le \langle r_\eta(u), z^+ - z_\rho\rangle + \frac{L_{\psi,\rho}}{2}\|z^+ - u\|^2 - \frac{m_\rho}{2}\|z_\rho - u\|^2. \tag{51}$$

Since $z^+ = u - \eta r_\eta(u)$, the right-hand side of (51) equals

$$\langle r_\eta(u), u - z_\rho\rangle - \eta\left(1 - \frac{\eta L_{\psi,\rho}}{2}\right)\|r_\eta(u)\|^2 - \frac{m_\rho}{2}\|u - z_\rho\|^2. \tag{52}$$

Meanwhile, Young's inequality yields

$$\langle r_\eta(u), u - z_\rho\rangle \le \frac{1}{2m_\rho}\|r_\eta(u)\|^2 + \frac{m_\rho}{2}\|u - z_\rho\|^2. \tag{53}$$

Combining (53), (51), and (52), and using $\eta \le 2/L_{\psi,\rho}$, gives

$$0 \le h(z^+) - h(z_\rho) \le \frac{1}{2m_\rho}\|r_\eta(u)\|^2.$$

Here, the lower bound follows from the optimality of z_ρ, and $h(z_\rho) = g_\rho(x, y)$. □

Lemma 20 (Complexity of the residual stopping rule). *Let $\{z^j\}$ be generated by Algorithm 2 for a fixed $w = (x, y) \in C$, and let $q_\rho := 1 - \frac{1}{\kappa_{\psi,\rho}}$, where $\kappa_{\psi,\rho} := \frac{L_{\psi,\rho}}{m_\rho} \in [1, +\infty]$. If $\kappa_{\psi,\rho} > 1$, then there exists a constant $C_{\mathrm{ME}} > 0$, independent of ζ, such that*

$$\|r^j\|^2 \le C_{\mathrm{ME}} q_\rho^j, \qquad j \ge 0. \tag{54}$$

Consequently, the stopping condition (49) *is satisfied after $O\Big(\kappa_{\psi,\rho} \log \frac{1}{\zeta}\Big)$ iterations. Moreover, when w ranges over C and the initializations belong to Y, C_{ME} can be chosen uniformly.*

Proof. Let $z_\rho = z_\rho(w)$. Note that h is m_ρ-strongly convex with $L_{\psi,\rho}$-Lipschitz continuous gradient. Applying Beck (2017, Theorem 10.29(b)) to $h + \iota_Y$, with the constant stepsize $\eta = L_{\psi,\rho}^{-1}$, gives

$$\|z^j - z_\rho\|^2 \le q_\rho^j \|z^{\mathrm{init}} - z_\rho\|^2, \qquad j \ge 0. \tag{55}$$

Since $r^j = \eta^{-1}(z^j - z^{j+1})$ and $\eta = L_{\psi,\rho}^{-1}$, the triangle inequality and (55) yield

$$\begin{aligned} \|r^j\| &= L_{\psi,\rho}\|z^j - z^{j+1}\| \\ &\le L_{\psi,\rho}\big(\|z^j - z_\rho\| + \|z^{j+1} - z_\rho\|\big) \\ &\le L_{\psi,\rho}(1 + \sqrt{q_\rho}) q_\rho^{j/2} \|z^{\mathrm{init}} - z_\rho\| \\ &\le 2L_{\psi,\rho} q_\rho^{j/2} \|z^{\mathrm{init}} - z_\rho\|. \end{aligned}$$

Hence, $\|r^j\|^2 \le 4L_{\psi,\rho}^2 \|z^{\mathrm{init}} - z_\rho\|^2 q_\rho^j$, which proves (54) with

$$C_{\mathrm{ME}} = 4L_{\psi,\rho}^2 \|z^{\mathrm{init}} - z_\rho\|^2.$$

For $\kappa_{\psi,\rho} > 1$, $q_\rho^j = \Big(1 - \frac{1}{\kappa_{\psi,\rho}}\Big)^j \le \exp\Big(-\frac{j}{\kappa_{\psi,\rho}}\Big)$. Therefore, $\|r^j\|^2 \le 2m_\rho\zeta$ after $O(\kappa_{\psi,\rho}\log(1/\zeta))$ iterations.

Finally, let $D_Y := \mathrm{diam}(Y)$. Since $z^{\mathrm{init}}, z_\rho \in Y$, $\|z^{\mathrm{init}} - z_\rho\| \le D_Y$. Thus, $C_{\mathrm{ME}} \le 4L_{\psi,\rho}^2 D_Y^2$, which is independent of w, z^{init}, and ζ. Consequently, C_{ME} can be chosen uniformly over all Moreau evaluations. □

Combining Lemmas 19 and 20, Algorithm 2 terminates after $O(\kappa_{\psi,\rho}\log(1/\zeta))$ projected-gradient iterations and returns a point satisfying (4). The complete implementation of IVSP using Algorithm 2 is given in Algorithm 3.

D Proofs of the theoretical results

We recall that our Lyapunov function introduced in Section 3.2 is given by

$$Q_k = \gamma_k(f(w^k) - f_{\mathrm{inf}}) + G_{\mu_k}(w^k) + c_\phi \mu_k. \tag{59}$$

D.1 Lemma 4: Well-definedness of line search and descent inequality

The following is the complete statement of Lemma 4, including the constants used in the analysis.

Lemma 4 (complete statement). *Suppose Assumption 1 holds. Let $\hat\vartheta \in (0, 2)$ and $\underline\vartheta > 0$. For each iteration $k \ge 0$ and each trial parameter $L > 0$, define $d_{k,L} := w_L^k - w^k$, $E_k^0 := E_{\mu_k,\rho}(\zeta_k)$, and set $\vartheta_{k,L} := \max\Big\{\min\Big\{\frac{\mu_k E_k^0}{\|d_{k,L}\|}, \hat\vartheta L\Big\}, \underline\vartheta\Big\}$, where the ratio is interpreted as $+\infty$ when $d_{k,L} = 0$. Define $\tau_{k,L} := \frac{\mu_k}{2\vartheta_{k,L}}(E_k^0)^2 + \zeta_k + \zeta_{k+1}$. Furthermore, let*

$$L_{\mathrm{safe}} := \max\left\{\frac{\underline\vartheta}{\hat\vartheta}, \frac{c_1 + \widehat{\gamma}_0\mu_0 L_f + L_G}{2 - \hat\vartheta}\right\}.$$

Then, for every $k \ge 0$, the following hold.

Algorithm 3 Detailed implementation of Algorithm 1

Require. Constants $c_1, c_2, a_\phi, \underline{\vartheta} > 0$, $\hat{\vartheta} \in (0, 2)$, an initial point $w^0 \in C$, an initial proximal point $z^0_{\mathrm{init}} \in Y$, positive nonincreasing sequences $(\mu_k)_{k\geq 0}$ and $(\zeta_k)_{k\geq 0}$, and a strictly decreasing positive inverse-penalty grid $(\widehat{\gamma}_t)_{t\geq 0}$. Choose either: *Strategy I*, a fixed parameter $L > 0$; or *Strategy II*, constants $0 < L_{\min} \leq L_{\max} < \infty$ for backtracking. Set $t_0 = 0$ and $k = 0$.

Step 0. Compute the initial ζ_0-accurate Moreau evaluation $(\widetilde{z}^0_0, \widetilde{R}^0_0) = \mathrm{ME}(w^0, z^0_{\mathrm{init}}, \zeta_0)$. Form

$$\widetilde{\nabla} R^0_0 = \begin{pmatrix} \nabla_x g(x^0, y^0) - \nabla_x g(x^0, \widetilde{z}^0_0) \\ \nabla_y g(x^0, y^0) - \rho^{-1}(y^0 - \widetilde{z}^0_0) \end{pmatrix}.$$

Step 1. Set $\gamma_k = \widehat{\gamma}_{t_k}$, $E^0_k := E_{\mu_k,\rho}(\zeta_k)$, and $\widetilde{\nabla} G^0_k := \phi'_{\mu_k}(\widetilde{R}^0_k)\widetilde{\nabla} R^0_k$. Under Strategy I, set L to the prescribed fixed value. Under Strategy II, choose $L_{k,0} \in [L_{\min}, L_{\max}]$ and set $L = L_{k,0}$.

Step 2. Compute the trial point

$$w^k_L = P_C\left(w^k - \frac{\mu_k}{L}\left[\gamma_k \nabla f(w^k) + \widetilde{\nabla} G^0_k\right]\right). \tag{56}$$

Step 3. Using $\widetilde{z}^0_k$ as a warm start, compute a ζ_{k+1}-accurate Moreau evaluation at the trial point: $(\widetilde{z}^+_{k,L}, \widetilde{R}^+_{k,L}) = \mathrm{ME}(w^k_L, \widetilde{z}^0_k, \zeta_{k+1})$. Under Strategy II, define

$$\tau_{k,L} := \frac{\mu_k}{2\vartheta_{k,L}}(E^0_k)^2 + \zeta_k + \zeta_{k+1},$$

where $\vartheta_{k,L} := \max\left\{\min\left\{\frac{\mu_k E^0_k}{\|w^k_L - w^k\|}, \hat{\vartheta} L\right\}, \underline{\vartheta}\right\}$, where the ratio is interpreted as $+\infty$ when $w^k_L = w^k$.

Step 4. Under Strategy I, set $L_k := L$, $w^{k+1} := w^k_L$, and go to **Step 5**. Under Strategy II, if

$$\gamma_k f(w^k_L) + \phi_{\mu_k}(\widetilde{R}^+_{k,L} + \zeta_{k+1}) \leq \gamma_k f(w^k) + \phi_{\mu_k}(\widetilde{R}^0_k) - \frac{c_1}{2\mu_k}\|w^k_L - w^k\|^2 + \tau_{k,L} \tag{57}$$

does not hold, set $L \leftarrow 2L$ and return to **Step 2**. Otherwise, set $L_k := L$, $w^{k+1} := w^k_L$, and go to **Step 5**.

Step 5. Set $\widetilde{z}^+_k := \widetilde{z}^+_{k,L_k}$, and $\widetilde{R}^+_k := \widetilde{R}^+_{k,L_k}$. Then form

$$\widetilde{\nabla} R^+_k = \begin{pmatrix} \nabla_x g(x^{k+1}, y^{k+1}) - \nabla_x g(x^{k+1}, \widetilde{z}^+_k) \\ \nabla_y g(x^{k+1}, y^{k+1}) - \rho^{-1}(y^{k+1} - \widetilde{z}^+_k) \end{pmatrix}.$$

Set $d^k := w^{k+1} - w^k$, and $\omega_{k+1} := \frac{\|d^k\|}{\mu_k}$.

Step 6. If

$$\widetilde{R}^+_k > 2a_\phi \mu_k \qquad \text{and} \qquad \|d^k\| \leq c_2 \mu_k \gamma_k, \tag{58}$$

set $t_{k+1} = t_k + 1$; otherwise, set $t_{k+1} = t_k$.

Step 7. Retain the accepted Moreau evaluation as the current evaluation for the next outer iteration:

$$\widetilde{z}^0_{k+1} := \widetilde{z}^+_k, \qquad \widetilde{R}^0_{k+1} := \widetilde{R}^+_k, \qquad \widetilde{\nabla} R^0_{k+1} := \widetilde{\nabla} R^+_k.$$

Update $k \leftarrow k + 1$ and return to **Step 1**.

(i) Every trial parameter $L \geq L_{\mathrm{safe}}$ satisfies (57)*, i.e.,*

$$\gamma_k f(w^k_L) + \phi_{\mu_k}\big(\widetilde{R}^+_{k,L} + \zeta_{k+1}\big) \leq \gamma_k f(w^k) + \phi_{\mu_k}(\widetilde{R}^0_k) - \frac{c_1}{2\mu_k}\|w^k_L - w^k\|^2 + \tau_{k,L}. \tag{60}$$

Consequently, any fixed $L \geq L_{\mathrm{safe}}$ is admissible under Strategy I. Under Strategy II, the backtracking procedure terminates finitely, and the accepted parameter satisfies

$$L_k \leq \overline{L} := \max\{L_{\max}, 2L_{\mathrm{safe}}\}. \tag{61}$$

Furthermore, the number of trial evaluations at each iteration is bounded by

$$B_{\mathrm{ls}} := 1 + \max\left\{0, \left\lceil \log_2 \frac{L_{\mathrm{safe}}}{L_{\min}} \right\rceil\right\}. \tag{62}$$

(ii) Let $L_k = L$ under Strategy I, and let L_k denote the accepted parameter under Strategy II. Set $w^{k+1} := w^k_{L_k}$, $d^k := w^{k+1} - w^k$, and $\omega_{k+1} := \frac{\|d^k\|}{\mu_k}$. Moreover, let $\vartheta_k := \vartheta_{k,L_k}$ and $\tau_k := \tau_{k,L_k}$. Then, under either strategy,

$$Q_{k+1} \leq Q_k - \frac{c_1}{2}\mu_k \omega^2_{k+1} + \tau_k. \tag{63}$$

Proof of Lemma 4. Fix an iteration k and a trial parameter $L > 0$, and let $d \coloneqq d_{k,L} = w_L^k - w^k$. Using the optimality condition for the projected step (56) and noting that $w^k \in C$, we obtain

$$\gamma_k \langle \nabla f(w^k), d \rangle + \langle \widetilde{\nabla} G_k^0, d \rangle \le -\frac{L}{\mu_k} \|d\|^2.$$

Adding and subtracting $\nabla G_{\mu_k}(w^k)$ yields

$$\gamma_k \langle \nabla f(w^k), d \rangle + \langle \nabla G_{\mu_k}(w^k), d \rangle \le -\frac{L}{\mu_k} \|d\|^2 + \left\langle \nabla G_{\mu_k}(w^k) - \widetilde{\nabla} G_k^0, d \right\rangle. \tag{64}$$

Meanwhile, by the Cauchy–Schwarz inequality, Lemma 11(iv), and Young's inequality,

$$\begin{aligned} \left\langle \nabla G_{\mu_k}(w^k) - \widetilde{\nabla} G_k^0, d \right\rangle &\le \left\| \nabla G_{\mu_k}(w^k) - \widetilde{\nabla} G_k^0 \right\| \|d\| \\ &\le E_k^0 \|d\| \\ &\le \frac{\vartheta_{k,L}}{2\mu_k} \|d\|^2 + \frac{\mu_k}{2\vartheta_{k,L}} (E_k^0)^2. \end{aligned} \tag{65}$$

Since ∇f is L_f-Lipschitz and, by Lemma 9, ∇G_{μ_k} is (L_G/μ_k)-Lipschitz, we have (see (Beck, 2017, Lemma 5.7))

$$\begin{aligned} f(w^k + d) &\le f(w^k) + \langle \nabla f(w^k), d \rangle + \frac{L_f}{2} \|d\|^2, \\ G_{\mu_k}(w^k + d) &\le G_{\mu_k}(w^k) + \langle \nabla G_{\mu_k}(w^k), d \rangle + \frac{L_G}{2\mu_k} \|d\|^2. \end{aligned}$$

Multiplying the first inequality by γ_k, adding the second, and using (64)–(65) gives

$$\begin{aligned} &\gamma_k f(w^k + d) + G_{\mu_k}(w^k + d) \\ &\quad \le \gamma_k f(w^k) + G_{\mu_k}(w^k) - \frac{2L - \gamma_k \mu_k L_f - L_G - \vartheta_{k,L}}{2\mu_k} \|d\|^2 + \chi_{k,L}, \end{aligned} \tag{66}$$

where $\chi_{k,L} \coloneqq \frac{\mu_k}{2\vartheta_{k,L}} (E_k^0)^2$.

Now, suppose that $L \ge L_{\mathrm{safe}}$. By the definition of L_{safe}, we have $L \ge \frac{\vartheta}{\hat{\vartheta}}$, and therefore $\underline{\vartheta} \le \hat{\vartheta} L$. Since $\min\left\{ \frac{\mu_k E_k^0}{\|d_{k,L}\|}, \hat{\vartheta} L \right\}$ is also bounded above by $\hat{\vartheta} L$, it follows that $\vartheta_{k,L} \le \hat{\vartheta} L$. Since $\gamma_k \le \widehat{\gamma}_0$ and $\mu_k \le \mu_0$, we obtain

$$\begin{aligned} 2L - \gamma_k \mu_k L_f - L_G - \vartheta_{k,L} &\ge (2 - \hat{\vartheta}) L - \widehat{\gamma}_0 \mu_0 L_f - L_G \\ &\ge c_1, \end{aligned}$$

where the last inequality follows from the definition of L_{safe}. Hence, using (66), we have for all $L \ge L_{\mathrm{safe}}$ that

$$\gamma_k f(w_L^k) + G_{\mu_k}(w_L^k) \le \gamma_k f(w^k) + G_{\mu_k}(w^k) - \frac{c_1}{2\mu_k} \|w_L^k - w^k\|^2 + \chi_{k,L}. \tag{67}$$

We next relate the exact values in (67) to the quantities used in the inexact Armijo test. At the trial point w_L^k, we obtain the ζ_{k+1}-accurate minimizer of $\psi_\rho(\cdot, w_L^k)$. Applying (33) and (32) in Lemma 11, respectively, we obtain

$$\gamma_k f(w_L^k) + \phi_{\mu_k}\left(\widetilde{R}_{k,L}^+ + \zeta_{k+1}\right) \le \gamma_k f(w_L^k) + G_{\mu_k}(w_L^k) + \zeta_{k+1}. \tag{68}$$

On the other hand, at the current point w^k, the ζ_k-accurate solution gives

$$\gamma_k f(w^k) + G_{\mu_k}(w^k) \le \gamma_k f(w^k) + \phi_{\mu_k}(\widetilde{R}_k^0) + \zeta_k, \tag{69}$$

by using (32) and (33), respectively. Combining (67), (68), and (69), and noting that $\tau_{k,L} = \chi_{k,L} + \zeta_k + \zeta_{k+1}$, yields (60). Thus every $L \ge L_{\mathrm{safe}}$ satisfies the inexact Armijo condition.

Under Strategy I, the prescribed fixed parameter is therefore valid whenever $L \ge L_{\mathrm{safe}}$. Under Strategy II, L is doubled until the inexact Armijo condition holds, so the backtracking procedure

terminates finitely. Since $L_{k,0} \in [L_{\min}, L_{\max}]$, every accepted parameter satisfies (61). Moreover, the number of trial evaluations at any iteration is at most B_{ls} as given in (62). This completes the proof of (i).

We next prove (ii). Note from (32) that

$$\phi_{\mu_k}(\widetilde{R}_k^0) \le G_{\mu_k}(w^k) \qquad \text{and} \qquad G_{\mu_k}(w^{k+1}) \le \phi_{\mu_k}\big(\widetilde{R}_k^+ + \zeta_{k+1}\big).$$

Together with (60), we obtain[1]

$$\gamma_k f(w^{k+1}) + G_{\mu_k}(w^{k+1}) \le \gamma_k f(w^k) + G_{\mu_k}(w^k) - \frac{c_1}{2\mu_k}\|d^k\|^2 + \tau_k. \tag{70}$$

Meanwhile, since $\gamma_{k+1} \le \gamma_k$ and $f(w^{k+1}) - f_{\inf} \ge 0$,

$$\gamma_{k+1}\big(f(w^{k+1}) - f_{\inf}\big) \le \gamma_k\big(f(w^{k+1}) - f_{\inf}\big). \tag{71}$$

Moreover, inequality (25) in Lemma 9 gives

$$G_{\mu_{k+1}}(w^{k+1}) + c_\phi \mu_{k+1} \le G_{\mu_k}(w^{k+1}) + c_\phi \mu_k. \tag{72}$$

Combining (71) and (72) with (70), and using $\|d^k\| = \mu_k \omega_{k+1}$ and the definition of Q_k in (59), we obtain (63). □

D.2 Lemma 5: Finite penalty stabilization

To prove Lemma 5, we need the following two lemmas.

Lemma 21 (Uniform ENNAMCQ margin)**.** *Suppose that C is nonempty, compact, and convex, and that R is continuously differentiable on C. If ENNAMCQ holds for* (1) *and $\{w \in C : R(w) \ge 0\} \neq \varnothing$, then*

$$\eta_\delta := \min_{\substack{w \in C \\ R(w) \ge 0}} \operatorname{dist}(0, \nabla R(w) + N_C(w)) > 0. \tag{73}$$

Proof. Let $\eta := \inf_{\substack{w \in C \\ R(w) \ge 0}} \operatorname{dist}(0, \nabla R(w) + N_C(w))$. Since R is continuous, the set $\{w \in C : R(w) \ge 0\}$ is compact. Choose a sequence $\{w^j\}$ in this set and $n^j \in N_C(w^j)$ such that $\|\nabla R(w^j) + n^j\| \to \eta$. By compactness, we may assume without loss of generality that $w^j \to \bar{w}$, where $\bar{w} \in C$ and $R(\bar{w}) \ge 0$. It follows that

$$\eta \le \operatorname{dist}(0, \nabla R(\bar{w}) + N_C(\bar{w})) \tag{74}$$

Meanwhile, since ∇R is continuous and C is compact, $\{\nabla R(w^j)\}$ is bounded. Moreover, since $\{\nabla R(w^j) + n^j\}$ is bounded, $\{n^j\}$ is also bounded. Therefore, there exists a subsequence $\{j_r\}$ and some $\bar{n}$ such that $n^{j_r} \to \bar{n}$. Since C is closed and convex, the normal cone mapping N_C has closed graph by Rockafellar & Wets (1998, Proposition 6.6), and therefore $\bar{n} \in N_C(\bar{w})$. Consequently,

$$\operatorname{dist}(0, \nabla R(\bar{w}) + N_C(\bar{w})) \le \|\nabla R(\bar{w}) + \bar{n}\| = \lim_{r \to \infty} \|\nabla R(w^{j_r}) + n^{j_r}\| = \eta,$$

where the first equality holds by continuity of ∇R. Together with (74), we have

$$\eta = \operatorname{dist}(0, \nabla R(\bar{w}) + N_C(\bar{w})).$$

Thus the infimum is attained at $\bar{w}$ so that $\eta_\delta = \eta$. Finally, ENNAMCQ at $\bar{w}$ implies that $\operatorname{dist}(0, \nabla R(\bar{w}) + N_C(\bar{w})) > 0$, and so $\eta_\delta > 0$. □

Lemma 22 (Stationarity bound at a penalty update)**.** *Let $\overline{L} > 0$ be given by*

$$\overline{L} = \begin{cases} L, & \textit{under Strategy I}, \\ \max\{L_{\max}, 2L_{\mathrm{safe}}\}, & \textit{under Strategy II}, \end{cases} \tag{75}$$

[1] We note that under Strategy I, (70) follows directly from (67); in fact, in this case the stronger estimate with $\chi_k := \frac{\mu_k}{2\vartheta_k}(E_k^0)^2$ in place of τ_k holds.

with $L \geq L_{\text{safe}}$. Define $E_k^0 \coloneqq E_{\mu_k,\rho}(\zeta_k)$, $E_k^+ \coloneqq E_{\mu_k,\rho}(\zeta_{k+1})$, $D_k^+ \coloneqq D_\rho(\zeta_{k+1})$, and $A_0 \coloneqq M_f + c_2(\overline{L} + L_G)$, where $M_f \coloneqq \max_{w\in C} \|\nabla f(w)\|$. If a penalty update is triggered at iteration k, then

$$R(w^{k+1}) > 0, \qquad \text{and} \tag{76}$$

$$\operatorname{dist}\big(0, \nabla R(w^{k+1}) + N_C(w^{k+1})\big) \leq 2A_0\gamma_k + 2(E_k^0 + E_k^+) + D_k^+. \tag{77}$$

Proof. Since a penalty update is triggered, we have from the first condition in (58) that $\widetilde{R}_k^+ > 2a_\phi\mu_k > 0$. Thus, by (31), we have $R(w^{k+1}) \geq \widetilde{R}_k^+ > 0$. This proves (76).

From the optimality condition of (56) with $L = L_k$, there exists $n^{k+1} \in N_C(w^{k+1})$ such that

$$0 = \gamma_k \nabla f(w^k) + \widetilde{\nabla} G_k^0 + \frac{L_k}{\mu_k} d^k + n^{k+1}. \tag{78}$$

At the accepted point, let $\widetilde{s}_k^+ \coloneqq \phi'_{\mu_k}(\widetilde{R}_k^+)$, and $\widetilde{\nabla} G_k^+ \coloneqq \widetilde{s}_k^+ \widetilde{\nabla} R_k^+$. Then (78) gives

$$\widetilde{s}_k^+ \widetilde{\nabla} R_k^+ + n^{k+1} = -\gamma_k \nabla f(w^k) - \frac{L_k}{\mu_k} d^k + \widetilde{\nabla} G_k^+ - \widetilde{\nabla} G_k^0. \tag{79}$$

Meanwhile, by the triangle inequality,

$$\begin{aligned}\|\widetilde{\nabla} G_k^+ - \widetilde{\nabla} G_k^0\| \leq{}& \|\widetilde{\nabla} G_k^+ - \nabla G_{\mu_k}(w^{k+1})\| + \|\nabla G_{\mu_k}(w^{k+1}) - \nabla G_{\mu_k}(w^k)\| \\ &+ \|\nabla G_{\mu_k}(w^k) - \widetilde{\nabla} G_k^0\|.\end{aligned}$$

Using Lemma 11(iv) to bound the first and third terms, and using Lemma 9 to bound the second term, we obtain

$$\|\widetilde{\nabla} G_k^+ - \widetilde{\nabla} G_k^0\| \leq L_G \omega_{k+1} + E_k^0 + E_k^+.$$

Hence, together with (79) and noting that $L_k \leq \overline{L}$,

$$\|\widetilde{s}_k^+ \widetilde{\nabla} R_k^+ + n^{k+1}\| \leq M_f \gamma_k + (\overline{L} + L_G)\omega_{k+1} + E_k^0 + E_k^+.$$

Since $N_C(w^{k+1})$ is a cone and $\widetilde{s}_k^+ \geq 1/2$ by Lemma 10(i) with $\theta = 2$,

$$\operatorname{dist}\Big(0, \widetilde{\nabla} R_k^+ + N_C(w^{k+1})\Big) \leq 2M_f\gamma_k + 2(\overline{L} + L_G)\omega_{k+1} + 2(E_k^0 + E_k^+).$$

The second condition in (58) gives $\omega_{k+1} \leq c_2\gamma_k$, and therefore

$$\operatorname{dist}\Big(0, \widetilde{\nabla} R_k^+ + N_C(w^{k+1})\Big) \leq 2A_0\gamma_k + 2(E_k^0 + E_k^+). \tag{80}$$

Finally, Lemma 11(iii) gives $\|\widetilde{\nabla} R_k^+ - \nabla R(w^{k+1})\| \leq D_k^+$. Moreover, for every $n \in N_C(w^{k+1})$, the triangle inequality gives

$$\|\nabla R(w^{k+1}) + n\| \leq \|\nabla R(w^{k+1}) - \widetilde{\nabla} R_k^+\| + \|\widetilde{\nabla} R_k^+ + n\|.$$

Taking the infimum over $n \in N_C(w^{k+1})$ and using (80), we have

$$\begin{aligned}&\operatorname{dist}\big(0, \nabla R(w^{k+1}) + N_C(w^{k+1})\big) \\ &\quad\leq \|\nabla R(w^{k+1}) - \widetilde{\nabla} R_k^+\| + \operatorname{dist}\Big(0, \widetilde{\nabla} R_k^+ + N_C(w^{k+1})\Big) \\ &\quad\leq 2A_0\gamma_k + 2(E_k^0 + E_k^+) + D_k^+,\end{aligned}$$

which proves (77). □

We now give the complete statement of Lemma 5, including the explicit constant appearing in the lower bound on the inverse penalty parameter.

Lemma 5 (complete statement). *Suppose ENNAMCQ holds for* (1)*, $\zeta_k = o(\mu_k)$, and there exists $M_\gamma > 1$ such that*

$$\widehat{\gamma}_t \leq M_\gamma \widehat{\gamma}_{t+1} \qquad \forall t \geq 0. \tag{81}$$

Let $\overline{L}$ be given by (75)*, and define $A_0 \coloneqq M_f + c_2(\overline{L} + L_G)$.*

1. *If* $\{w \in C : R(w) \geq 0\} = \varnothing$, *then no penalty update occurs, and* $\gamma_k = \underline{\gamma} := \widehat{\gamma}_0$ *for all* $k \geq 0$.
2. *If* $\{w \in C : R(w) \geq 0\} \neq \varnothing$, *let* $\eta_\delta > 0$ *be defined by* (73). *Then there exists* $k_0 \geq 0$ *such that*

$$2[E_{\mu_k,\rho}(\zeta_k) + E_{\mu_k,\rho}(\zeta_{k+1})] + D_\rho(\zeta_{k+1}) \leq \frac{\eta_\delta}{2} \qquad \forall k \geq k_0.$$

Moreover,

$$\gamma_k \geq \underline{\gamma} := \min\left\{\widehat{\gamma}_{k_0}, \frac{\eta_\delta}{4M_\gamma A_0}\right\} > 0 \qquad \forall k \geq 0.$$

In either case, the penalty parameter is updated only finitely many times, and γ_k *is eventually constant.*

Proof of Lemma 5. If $\{w \in C : R(w) \geq 0\} = \varnothing$, then $R(w) < 0$ for every $w \in C$. Since $\widetilde{R}(w) \leq R(w)$, the first condition in (58) can never hold. Hence no penalty update is ever triggered, so $t_k = 0$ and $\gamma_k = \widehat{\gamma}_0$ for every k. Thus the conclusion holds with $\underline{\gamma} = \widehat{\gamma}_0$.

Suppose now that $\{w \in C : R(w) \geq 0\} \neq \varnothing$. By Lemma 21, $\eta_\delta > 0$ is well defined. Since $\zeta_k = o(\mu_k)$ and $\mu_k \leq \mu_0$, we have $\zeta_k/\mu_k \to 0$ and $\zeta_k \to 0$. Together with (27), we have

$$D_k^+ = L_{a,\rho}\sqrt{\frac{2\zeta_{k+1}}{m_\rho}} \to 0 \qquad \text{and} \qquad E_k^0 = L_{a,\rho}\sqrt{\frac{2\zeta_k}{m_\rho}} + b_\phi B_R \frac{\zeta_k}{\mu_k} \to 0.$$

Moreover, since $\mu_{k+1} \leq \mu_k$, $\zeta_{k+1}/\mu_k \leq \zeta_{k+1}/\mu_{k+1} \to 0$, and therefore

$$E_k^+ = L_{a,\rho}\sqrt{\frac{2\zeta_{k+1}}{m_\rho}} + b_\phi B_R \frac{\zeta_{k+1}}{\mu_k} \to 0.$$

Hence, there exists k_0 such that

$$2(E_k^0 + E_k^+) + D_k^+ \leq \frac{\eta_\delta}{2} \qquad \forall k \geq k_0. \tag{82}$$

Let $k \geq k_0$ be an iteration at which a penalty update is triggered. By Lemma 22, $R(w^{k+1}) > 0$, and hence the definition of η_δ gives

$$\eta_\delta \leq \operatorname{dist}\big(0, \nabla R(w^{k+1}) + N_C(w^{k+1})\big).$$

Combining this with (77) and (82), we obtain $\eta_\delta \leq 2A_0\gamma_k + \eta_\delta/2$. Thus, $\gamma_k \geq \frac{\eta_\delta}{4A_0}$. Since an update is triggered, $t_{k+1} = t_k + 1$. By (81), $\gamma_k = \widehat{\gamma}_{t_k} \leq M_\gamma \widehat{\gamma}_{t_k+1} = M_\gamma \gamma_{k+1}$. Hence

$$\gamma_{k+1} \geq \frac{\eta_\delta}{4M_\gamma A_0}. \tag{83}$$

Since $t_k \leq k$, we have for $k \leq k_0$ that $\gamma_k = \widehat{\gamma}_{t_k} \geq \widehat{\gamma}_{k_0} \geq \underline{\gamma}$. For $k \geq k_0$, γ_k changes only when a penalty update is triggered. If no update occurs at iteration k, then $\gamma_{k+1} = \gamma_k$. If an update is triggered, then (83) gives $\gamma_{k+1} \geq \eta_\delta/(4M_\gamma A_0) \geq \underline{\gamma}$. Therefore, $\gamma_k \geq \underline{\gamma}$ for all $k \geq 0$. This proves (ii).

Finally, since $\widehat{\gamma}_t \downarrow 0$, there are only finitely many indices t such that $\widehat{\gamma}_t \geq \underline{\gamma}$. Since every penalty update increases t_k by one, only finitely many penalty updates can occur. Hence t_k, and therefore γ_k, is eventually constant. □

The dependence of Lemma 5 on the relaxation level is worth noting. Although ENNAMCQ may hold for every fixed $\delta > 0$, the margin η_δ need not remain bounded away from zero as $\delta \downarrow 0$.

Example 23 (Deterioration of the ENNAMCQ margin)**.** Let X be any nonempty compact convex set, $Y = [-1, 1]$, and $g(x, y) = \frac{1}{2}y^2$. Then g is independent of x, and for every $\rho > 0$,

$$z_\rho(y) = \frac{y}{1+\rho}, \qquad H_\rho(y) = \frac{\rho}{2(1+\rho)}y^2.$$

Thus, for $0 < \delta < \rho/(2(1+\rho))$, the region $R(y) = H_\rho(y) - \delta \geq 0$ is characterized by $|y| \geq \sqrt{\frac{2(1+\rho)\delta}{\rho}}$. Moreover, $\nabla H_\rho(y) = \frac{\rho}{1+\rho} y$. Since the two points satisfying $|y| = \sqrt{\frac{2(1+\rho)\delta}{\rho}}$ lie in the interior of Y due to the choice of δ, their normal cones are $\{0\}$. Consequently, the minimum in Lemma 21 is attained at these points, and $\eta_\delta = \sqrt{\frac{2\rho}{1+\rho}}\sqrt{\delta}$. Hence $\eta_\delta = \Theta(\sqrt{\delta}) \to 0$ as $\delta \downarrow 0$.

Since the lower-level objective in Example 23 is strictly convex, ENNAMCQ holds for every fixed $\delta > 0$ by Proposition 15. Nevertheless, the corresponding η_δ may deteriorate as $\delta \downarrow 0$. Consequently, the lower bound γ in Lemma 5 may approach zero, and the stabilization index and constants entering the subsequent KKT estimates may also deteriorate. Accordingly, the complexity results below are stated for a fixed positive relaxation level and are not uniform in δ as $\delta \downarrow 0$.

D.3 Theorem 6 and Corollary 7: Complexity guarantees

We first establish the stationarity and feasibility estimates used in the complexity analysis.

Lemma 24 (Stationarity estimate)**.** *Let $E_k^0 := E_{\mu_k,\rho}(\zeta_k)$ and $E_k^+ := E_{\mu_k,\rho}(\zeta_{k+1})$, and define*

$$\widetilde{\lambda}_{k+1} := \gamma_k^{-1} \phi'_{\mu_k}(\widetilde{R}_k^+). \tag{84}$$

Then, for every $k \geq 0$,

$$\Phi_\rho^{\mathrm{s}}(w^{k+1}, \widetilde{\lambda}_{k+1}) \leq \gamma_k^{-1} \left(M_1 \omega_{k+1} + E_k^0 + E_k^+ \right), \tag{85}$$

where $\overline{L}$ is given by (75)*, $M_1 := \widehat{\gamma}_0 \mu_0 L_f + \overline{L} + L_G$, and Φ_ρ^{s} is given in Definition 2.*

Proof. From the proof of Lemma 5 in Appendix D.2, there exists $n^{k+1} \in N_C(w^{k+1})$ such that (78) holds. Then using the definition of $\widetilde{\lambda}_{k+1}$ in (84),

$$\begin{aligned}
&\gamma_k \nabla f(w^{k+1}) + \gamma_k \widetilde{\lambda}_{k+1} \nabla R(w^{k+1}) + n^{k+1} \\
&\quad = \gamma_k \big(\nabla f(w^{k+1}) - \nabla f(w^k) \big) - \frac{L_k}{\mu_k} d^k + \left[\phi'_{\mu_k}(\widetilde{R}_k^+) \nabla R(w^{k+1}) - \nabla G_{\mu_k}(w^{k+1}) \right] \\
&\qquad + \left[\nabla G_{\mu_k}(w^{k+1}) - \nabla G_{\mu_k}(w^k) \right] + \left[\nabla G_{\mu_k}(w^k) - \widetilde{\nabla} G_k^0 \right].
\end{aligned} \tag{86}$$

For the first bracketed term, since $\nabla G_{\mu_k}(w^{k+1}) = \phi'_{\mu_k}(R(w^{k+1})) \nabla R(w^{k+1})$, we have

$$\begin{aligned}
\left\| \phi'_{\mu_k}(\widetilde{R}_k^+) \nabla R(w^{k+1}) - \nabla G_{\mu_k}(w^{k+1}) \right\| &= \left| \phi'_{\mu_k}(\widetilde{R}_k^+) - \phi'_{\mu_k}(R(w^{k+1})) \right| \|\nabla R(w^{k+1})\| \\
&\leq \frac{b_\phi B_R}{\mu_k} \left| \widetilde{R}_k^+ - R(w^{k+1}) \right| \\
&\leq \frac{b_\phi B_R}{\mu_k} \zeta_{k+1} \leq E_k^+,
\end{aligned} \tag{87}$$

where we used (20), (18), and (31).

Taking norms in (86), applying the triangle inequality, and using (87),

$$\begin{aligned}
\left\| \gamma_k \nabla f(w^{k+1}) + \gamma_k \widetilde{\lambda}_{k+1} \nabla R(w^{k+1}) + n^{k+1} \right\| &\leq \gamma_k L_f \|d^k\| + \frac{L_k}{\mu_k} \|d^k\| + E_k^+ + \frac{L_G}{\mu_k} \|d^k\| + E_k^0 \\
&= \gamma_k L_f \|d^k\| + \frac{L_k + L_G}{\mu_k} \|d^k\| + E_k^0 + E_k^+.
\end{aligned}$$

where we also used L_f-Lipschitz continuity of ∇f, Lemma 9, and Lemma 11(iv). Since $\|d^k\| = \mu_k \omega_{k+1}$, $\gamma_k \leq \widehat{\gamma}_0$, $\mu_k \leq \mu_0$, and $L_k \leq \overline{L}$, we obtain

$$\left\| \gamma_k \nabla f(w^{k+1}) + \gamma_k \widetilde{\lambda}_{k+1} \nabla R(w^{k+1}) + n^{k+1} \right\| \leq M_1 \omega_{k+1} + E_k^0 + E_k^+.$$

Dividing by γ_k, noting that $\frac{n^{k+1}}{\gamma_k} \in N_C(w^{k+1})$ and by the definition of the distance function, we obtain (85). □

Lemma 25 (Feasibility and complementarity)**.** *Suppose that, at iteration k, no penalty update occurs but the second condition in* (58) *is satisfied. Then*

$$\Phi_\rho^{\mathrm{f}}(w^{k+1}) \le 2a_\phi\mu_k + \zeta_{k+1}, \tag{88}$$

$$\Phi_\rho^{\mathrm{c}}(w^{k+1}, \widetilde{\lambda}_{k+1}) \le \gamma_k^{-1}\left(2a_\phi\mu_k + \zeta_{k+1}\right), \tag{89}$$

where Φ_ρ^{f} and Φ_ρ^{c} are given in Definition 2

Proof. By the hypothesis, we have $\widetilde{R}_k^+ \le 2a_\phi\mu_k$. Thus, by (31), we have

$$R(w^{k+1}) \le \widetilde{R}_k^+ + \zeta_{k+1} \le 2a_\phi\mu_k + \zeta_{k+1}.$$

Consequently,

$$\Phi_\rho^{\mathrm{f}}(w^{k+1}) = [R(w^{k+1})]_+ \le [2a_\phi\mu_k + \zeta_{k+1}]_+ = 2a_\phi\mu_k + \zeta_{k+1},$$

proving (88).

For complementarity, write $R(w^{k+1}) = \widetilde{R}_k^+ + e_k^+$, where $0 \le e_k^+ \le \zeta_{k+1}$. Then

$$\begin{aligned}\left|\widetilde{\lambda}_{k+1}R(w^{k+1})\right| &\le \gamma_k^{-1}\left|\phi'_{\mu_k}(\widetilde{R}_k^+)\widetilde{R}_k^+\right| + \gamma_k^{-1}\phi'_{\mu_k}(\widetilde{R}_k^+)e_k^+ \\ &\le \gamma_k^{-1}\left|\phi'_{\mu_k}(\widetilde{R}_k^+)\widetilde{R}_k^+\right| + \gamma_k^{-1}\zeta_{k+1} \\ &\le \gamma_k^{-1}(2a_\phi\mu_k) + \gamma_k^{-1}\zeta_{k+1},\end{aligned}$$

where the second inequality holds by (20), and the last inequality holds by noting that $\widetilde{R}_k^+ \le 2a_\phi\mu_k$ and invoking Lemma 10(ii) with $\theta = 2$. By the definition of Φ_ρ^{c}, we get (89). □

Lemma 26 (KKT estimates at a non-update iteration)**.** *Suppose $\Xi_\infty := \sum_{k=0}^{\infty}\tau_k < \infty$, and let $\underline{\gamma} > 0$ be the lower bound from Lemma 5. Define*

$$t_* := \max\{t \ge 0 : \widehat{\gamma}_t \ge \underline{\gamma}\},$$

For $K \ge 0$, let $m := \lceil K/2 \rceil$ and define

$$S_K := \sum_{k=m}^{K}\mu_k, \qquad \widetilde{S}_K := S_K - t_*\mu_m. \tag{90}$$

$$\mathcal{E}_K := \max_{m\le j\le K}(E_j^0 + E_j^+), \tag{91}$$

$$\mathcal{Z}_K := \max_{m\le j\le K}\zeta_{j+1}. \tag{92}$$

If

$$\widetilde{S}_K > 0 \qquad \textit{and} \qquad \widetilde{S}_K^{-1} \le \frac{(c_2\underline{\gamma})^2}{M_2}, \tag{93}$$

where $M_2 := \frac{2(Q_0 + a_1\mu_0 + \Xi_\infty)}{c_1}$, then there exists an index $\widehat{k} \in [m, K]$ at which no penalty update occurs and such that

$$\Phi_\rho^{\mathrm{s}}(w^{\widehat{k}+1}, \widetilde{\lambda}_{\widehat{k}+1}) \le \underline{\gamma}^{-1}\left(M_1\sqrt{\frac{M_2}{\widetilde{S}_K}} + \mathcal{E}_K\right), \tag{94}$$

$$\Phi_\rho^{\mathrm{f}}(w^{\widehat{k}+1}) \le 2a_\phi\mu_m + \mathcal{Z}_K, \tag{95}$$

$$\Phi_\rho^{\mathrm{c}}(w^{\widehat{k}+1}, \widetilde{\lambda}_{\widehat{k}+1}) \le \underline{\gamma}^{-1}\left(2a_\phi\mu_m + \mathcal{Z}_K\right). \tag{96}$$

Proof. Since $\gamma_k \ge \underline{\gamma}$ for every k and $\widehat{\gamma}_t \downarrow 0$, the integer t_* is finite. Moreover, $t_0 = 0$ and every penalty update increases t_k by one. Since $\gamma_k = \widehat{\gamma}_{t_k} \ge \underline{\gamma}$, we have $t_k \le t_*$ for every k. Hence, the total number of penalty updates is at most t_*.

Let $\mathcal{I} \coloneqq \{k \geq 0 : t_{k+1} = t_k + 1\}$ be the set of penalty-update iterations, and define $\mathcal{N}_K \coloneqq \{m, \dots, K\} \setminus \mathcal{I}$. Thus, $\mathcal{N}_K$ consists of the non-update iterations in the second half of the first $K+1$ iterations.

Summing (63) from $k = m$ to K gives

$$\frac{c_1}{2} \sum_{k=m}^{K} \mu_k \omega_{k+1}^2 \leq Q_m - Q_{K+1} + \sum_{k=m}^{K} \tau_k. \tag{97}$$

On the other hand, summing (63) $k = 0$ to $m-1$ yields

$$Q_m \leq Q_0 + \sum_{k=0}^{m-1} \tau_k - \frac{c_1}{2} \sum_{k=0}^{m-1} \mu_k \omega_{k+1}^2 \leq Q_0 + \sum_{k=0}^{m-1} \tau_k. \tag{98}$$

Meanwhile, because $f(w) - f_{\inf} \geq 0$, $G_\mu(w) \geq -a_1 \mu$ by (21) and $\{\mu_k\}$ is nonincreasing, we have $Q_k \geq -a_1 \mu_0$ for all $k \geq 0$. Thus, combining (98) and (97) yields

$$\frac{c_1}{2} \sum_{k=m}^{K} \mu_k \omega_{k+1}^2 \leq Q_0 + a_1 \mu_0 + \sum_{k=0}^{K} \tau_k \leq Q_0 + a_1 \mu_0 + \Xi_\infty.$$

Hence, by the definition of M_2, we have

$$\sum_{k \in \mathcal{N}_K} \mu_k \omega_{k+1}^2 \leq M_2. \tag{99}$$

Since $\{\mu_k\}$ is nonincreasing and $|\mathcal{I}| \leq t_*$, we have from the definition of S_K and $\widetilde{S}_K$ in (90)

$$\begin{aligned} \sum_{k \in \mathcal{N}_K} \mu_k &= S_K - \sum_{k \in \mathcal{I} \cap \{m, \dots, K\}} \mu_k \\ &\geq S_K - |\mathcal{I} \cap \{m, \dots, K\}| \mu_m \\ &\geq S_K - t_* \mu_m = \widetilde{S}_K. \end{aligned} \tag{100}$$

Since $\widetilde{S}_K > 0$ by (93), the set $\mathcal{N}_K$ is nonempty. Combining (99) and (100), there exists $\widehat{k} \in \mathcal{N}_K$ such that

$$\omega_{\widehat{k}+1}^2 \leq \frac{\sum_{k \in \mathcal{N}_K} \mu_k \omega_{k+1}^2}{\sum_{k \in \mathcal{N}_K} \mu_k} \leq \frac{M_2}{\widetilde{S}_K}. \tag{101}$$

By (93) and $\gamma_{\widehat{k}} \geq \underline{\gamma}$,

$$\omega_{\widehat{k}+1} \leq \sqrt{\frac{M_2}{\widetilde{S}_K}} \leq c_2 \underline{\gamma} \leq c_2 \gamma_{\widehat{k}}.$$

Since $\omega_{\widehat{k}+1} = \dfrac{\left\| d^{\widehat{k}} \right\|}{\mu_{\widehat{k}}}$, it follows from the above inequality that the second condition in (58) holds. Meanwhile, $\widehat{k} \in \mathcal{N}_K$ so that no penalty update occurs at iteration $\widehat{k}$. Hence, Lemma 25 implies that

$$\Phi_\rho^{\mathrm{f}}(w^{\widehat{k}+1}) \leq 2 a_\phi \mu_{\widehat{k}} + \zeta_{\widehat{k}+1}.$$

Since $\widehat{k} \geq m$, the monotonicity of $\{\mu_k\}$ and the definition of $\mathcal{Z}_K$ in (92) imply $\mu_{\widehat{k}} \leq \mu_m$ and $\zeta_{\widehat{k}+1} \leq \mathcal{Z}_K$. Hence, (95) immediately follows. By a similar argument and noting that $\gamma_{\widehat{k}} \geq \underline{\gamma}$, we obtain (96).

Finally, Lemma 24, (101) and (91) yield

$$\begin{aligned} \Phi_\rho^{\mathrm{s}}(w^{\widehat{k}+1}, \widetilde{\lambda}_{\widehat{k}+1}) &\leq \gamma_{\widehat{k}}^{-1} \left(M_1 \omega_{\widehat{k}+1} + E_{\widehat{k}}^0 + E_{\widehat{k}}^+ \right) \\ &\leq \underline{\gamma}^{-1} \left(M_1 \sqrt{\frac{M_2}{\widetilde{S}_K}} + \mathcal{E}_K \right), \end{aligned}$$

which proves (94). □

To convert the preceding estimate into explicit polynomial rates, we record the following elementary consequences of the smoothing condition.

Lemma 27 (Polynomial smoothing estimates)**.** *Let $\{\mu_k\}$ be positive and nonincreasing and satisfy $\mu_k = \Theta((1+k)^{-r})$, with $r \in (0,1)$, and define S_K as in* (90) *with $m = \lceil K/2 \rceil$. Then*

$$S_K = \Theta(K^{1-r}), \qquad \mu_m = \Theta(K^{-r}). \tag{102}$$

Consequently, for $\widetilde{S}_K$ given in (90) *with any fixed $t_* \geq 0$, we have*

$$\widetilde{S}_K = \Theta(K^{1-r}). \tag{103}$$

In particular, $\widetilde{S}_K > 0$ for all sufficiently large K.

Proof. Since $K - m + 1 = \lfloor K/2 \rfloor + 1$, for $K \geq 2$,

$$\frac{K}{2} \leq K - m + 1 \leq K \tag{104}$$

By $\mu_k = \Theta((1+k)^{-r})$, there exist constants $c_\mu, C_\mu > 0$ such that, for all sufficiently large k,

$$c_\mu (1+k)^{-r} \leq \mu_k \leq C_\mu (1+k)^{-r}.$$

Hence, for all sufficiently large K and $k = m, \ldots, K$,

$$c_\mu (K+1)^{-r} \leq \mu_k \leq C_\mu (m+1)^{-r}.$$

Hence, summing from $k = m$ to $k = K$ and noting (104),

$$\frac{K}{2} c_\mu (K+1)^{-r} \leq S_K \leq K C_\mu (m+1)^{-r}.$$

Using $K + 1 \leq 2K$ and $m + 1 \geq K/2$ gives

$$\frac{c_\mu}{2^{r+1}} K^{1-r} \leq S_K \leq 2^r C_\mu K^{1-r}, \tag{105}$$

and therefore $S_K = \Theta(K^{1-r})$. Similarly, since $c_\mu (m+1)^{-r} \leq \mu_m \leq C_\mu (m+1)^{-r}$, and $K/2 \leq m + 1 \leq K$ for $K \geq 2$, we have

$$c_\mu K^{-r} \leq \mu_m \leq 2^r C_\mu K^{-r}. \tag{106}$$

Thus, $\mu_{\lceil K/2 \rceil} = \Theta(K^{-r})$. Hence, both the estimates in (102) hold.

Finally, using (105) and (106),

$$\frac{t_* \mu_m}{S_K} \leq \frac{t_* 2^r C_\mu K^{-r}}{(c_\mu / 2^{r+1}) K^{1-r}} = \frac{2^{2r+1} t_* C_\mu}{c_\mu K}, \tag{107}$$

Since

$$\frac{\widetilde{S}_K}{S_K} = 1 - \frac{t_* \mu_m}{S_K} \longrightarrow 1$$

by (107), we have $\widetilde{S}_K \sim S_K$. In particular, $\widetilde{S}_K = \Theta(S_K)$, and since $S_K = \Theta(K^{1-r})$, (103) holds. □

For instance, for any fixed integer $B \geq 1$, the blockwise schedule $\mu_k = \mu_0 (1 + \lfloor k/B \rfloor)^{-r}$ satisfies $\mu_k = \Theta((1+k)^{-r})$; the standard choice $\mu_k = \mu_0 (1+k)^{-r}$ corresponds to $B = 1$.

We now give the complete statement of Theorem 6, including the explicit estimates from which the rates in the main text follow.

Theorem 6 (complete statement). *Suppose Assumption 1 and ENNAMCQ hold, and suppose that the inverse-penalty grid satisfies* (81)*. Let $\{\mu_k\}$ be positive and nonincreasing with $\mu_k = \Theta((1+k)^{-r})$ for some $r \in (0,1)$, and let $\{\zeta_k\}$ satisfy $\zeta_k \leq \bar{\zeta} \mu_k^q$ for some $\bar{\zeta} > 0$ and $q > 1/r$.*

Let $\{w^k\}$ be generated by Algorithm 1 and, for each $k \geq 0$, set $\widetilde{\lambda}_{k+1} := \gamma_k^{-1}\phi'_{\mu_k}(\widetilde{R}^{k+1})$. Then, for every sufficiently large K, there exists $\widehat{k} \in [\lceil K/2 \rceil, K]$ such that

$$\Phi_\rho^{\mathrm{s}}(w^{\widehat{k}+1}, \widetilde{\lambda}_{\widehat{k}+1}) = O\Big(K^{-(1-r)/2}\Big), \tag{108}$$

$$\Phi_\rho^{\mathrm{f}}(w^{\widehat{k}+1}) = O(K^{-r}), \tag{109}$$

$$\Phi_\rho^{\mathrm{c}}(w^{\widehat{k}+1}, \widetilde{\lambda}_{\widehat{k}+1}) = O(K^{-r}). \tag{110}$$

Consequently, an $(\varepsilon_s, \varepsilon_f, \varepsilon_c)$-KKT point is obtained within

$$K = O\Big(\max\Big\{\varepsilon_s^{-2/(1-r)}, \varepsilon_f^{-1/r}, \varepsilon_c^{-1/r}\Big\}\Big) \tag{111}$$

outer iterations.

Proof. We verify the conditions required to invoke Lemma 26. Since $\zeta_k \leq \bar{\zeta}\mu_k^q$ with $q > 1/r > 1$, we have $\zeta_k/\mu_k \leq \bar{\zeta}\mu_k^{q-1} \to 0$. Hence $\zeta_k = o(\mu_k)$, and Lemma 5 yields $\gamma_k \geq \underline{\gamma} > 0$ for all $k \geq 0$ and finite stabilization of the penalty parameter.

Next, we verify the summability condition on $\{\tau_k\}$. From (27), we have $E_k^0 \leq c_D\mu_k^{q/2} + c_E\mu_k^{q-1}$, where $c_D := L_{a,\rho}\sqrt{2\bar{\zeta}/m_\rho}$ and $c_E := b_\phi B_R\bar{\zeta}$. Moreover, since $\mu_{k+1} \leq \mu_k$, $\zeta_{k+1} \leq \bar{\zeta}\mu_{k+1}^q \leq \bar{\zeta}\mu_k^q$, and therefore the same bound holds for E_k^+:

$$E_k^0 + E_k^+ \leq 2c_D\mu_k^{q/2} + 2c_E\mu_k^{q-1}. \tag{112}$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$, $\zeta_{k+1} \leq \bar{\zeta}\mu_k^q$, and $\vartheta_k \geq \underline{\vartheta}$ by definition, we obtain

$$\begin{aligned}\tau_k = \tau_{k,L_k} &= \frac{\mu_k}{2\vartheta_k}(E_k^0)^2 + \zeta_k + \zeta_{k+1} \\ &\leq \frac{c_D^2}{\underline{\vartheta}}\mu_k^{q+1} + \frac{c_E^2}{\underline{\vartheta}}\mu_k^{2q-1} + 2\bar{\zeta}\mu_k^q.\end{aligned}$$

Since $rq > 1$, we also have $r(q+1) > 1$ and $r(2q-1) = 2rq - r > 2 - r > 1$. Hence all three series on the above right-hand side are summable, and therefore $\Xi_\infty := \sum_{k=0}^\infty \tau_k < \infty$. In particular, the constant M_2 in Lemma 26 is finite.

Let $m = \lceil K/2 \rceil$. By Lemma 27, $S_K = \Theta(K^{1-r})$, $\mu_m = \Theta(K^{-r})$, and $\widetilde{S}_K = \Theta(K^{1-r})$. Thus $\widetilde{S}_K > 0$ for all sufficiently large K and $\widetilde{S}_K^{-1} = O(K^{-(1-r)}) \to 0$. Hence (93) holds for all sufficiently large K. At this point, for all sufficiently large K, the conditions of Lemma 26 are satisfied.

We derive the complexity of the right-hand sides of (94)–(96) in Lemma 26. For $m \leq j \leq K$, (112) and the monotonicity of μ_j give

$$E_j^0 + E_j^+ \leq 2c_D\mu_m^{q/2} + 2c_E\mu_m^{q-1}.$$

Using $\mu_m = \Theta(K^{-r})$ from (102), we obtain

$$\mathcal{E}_K = O\Big(K^{-rq/2} + K^{-r(q-1)}\Big). \tag{113}$$

Because $rq > 1$, we have $rq/2 > 1/2 > (1-r)/2$ and $r(q-1) = rq - r > 1 - r > (1-r)/2$. Therefore, together with (113),

$$\mathcal{E}_K = o\Big(K^{-(1-r)/2}\Big). \tag{114}$$

Similarly, for $m \leq j \leq K$, $\zeta_{j+1} \leq \bar{\zeta}\mu_{j+1}^q \leq \bar{\zeta}\mu_m^q$, and thus $\mathcal{Z}_K \leq \bar{\zeta}\mu_m^q = O(K^{-rq})$. Since $q > 1$,

$$\mathcal{Z}_K = o(K^{-r}). \tag{115}$$

Lemma 26 therefore yields, for every sufficiently large K, an index $\widehat{k} \in [m, K]$ satisfying (94)–(96). Since $\widetilde{S}_K^{-1/2} = O(K^{-(1-r)/2})$ from (103), (114) shows that the right-hand side of (94) is

$$O(K^{-(1-r)/2}) + o(K^{-(1-r)/2}) = O(K^{-(1-r)/2}).$$

Likewise, $\mu_m = \Theta(K^{-r})$ and (115) show that the right-hand sides of (95) and (96) are

$$O(K^{-r}) + o(K^{-r}) = O(K^{-r}).$$

This proves (108)–(110). Finally, solving the three rate bounds for K gives (111), which proves the claimed outer-iteration complexity. □

Remark 28 (Lower-level interpretation). The feasibility estimate (109) in Theorem 6, together with Corollary 13, shows that the returned iterate $w^{\widehat{k}+1} = (x^{\widehat{k}+1}, y^{\widehat{k}+1})$ satisfies $H_\rho(w^{\widehat{k}+1}) \le \delta + O(K^{-r})$, and consequently, its associated Moreau proximal point $z_\rho(w^{\widehat{k}+1})$ satisfies

$$\|y^{\widehat{k}+1} - z_\rho(w^{\widehat{k}+1})\| = O\Big(\sqrt{\delta + K^{-r}}\Big),$$

and

$$\operatorname{dist}\Big(0,\, \nabla_y g\Big(x^{\widehat{k}+1}, z_\rho(w^{\widehat{k}+1})\Big) + N_Y\Big(z_\rho(w^{\widehat{k}+1})\Big)\Big) = O\Big(\sqrt{\delta + K^{-r}}\Big).$$

Thus, the returned y-iterate is close to a point that is near-stationary for the lower-level problem. In particular, with $r = 1/3$ and $K = O(\varepsilon^{-3})$, both errors are $O(\sqrt{\delta + \varepsilon})$.

We next give the explicit oracle complexity obtained when the Moreau subproblems are solved by Algorithm 2. The argument is unchanged for any first-order inner method that computes a ζ-accurate Moreau evaluation in $O(\log(1/\zeta))$ iterations uniformly over the proximal subproblems; only the problem-dependent constant in the complexity bound changes.

Corollary 7 (Overall first-order complexity). *Suppose the conditions of Theorem 6 hold and, in addition, choose $\zeta_k = \Theta(\mu_k^q)$. Suppose Algorithm 2 is used for each required Moreau evaluation. Let B_{ls} denote a uniform bound on the number of trial Moreau evaluations per outer iteration, with $B_{\mathrm{ls}} = 1$ for Strategy I. Then the overall first-order/projection complexity of IVSP through outer iteration K is*

$$O(B_{\mathrm{ls}} \kappa_{\psi,\rho}\, K \log K)\,. \tag{116}$$

In particular, for the balanced choice $r = 1/3$, IVSP computes an ε-KKT point with overall first-order/projection complexity

$$O\left(B_{\mathrm{ls}} \kappa_{\psi,\rho}\, \varepsilon^{-3} \log \frac{1}{\varepsilon}\right) = \widetilde{O}(\varepsilon^{-3}). \tag{117}$$

Proof. By Lemma 20, each call to Algorithm 2 with tolerance ζ requires $O\Big(\kappa_{\psi,\rho} \log \frac{1}{\zeta}\Big)$ first-order/projection iterations for the Moreau subproblem. The implicit constant is uniform over the outer iterates and line-search trials.

At outer iteration k, every new trial-point Moreau evaluation is required to be ζ_{k+1}-accurate. Since $\zeta_k = \Theta(\mu_k^q)$ and $\mu_k = \Theta((1+k)^{-r})$, we have $\zeta_{k+1} = \Theta((k+2)^{-rq})$, and hence $\log(1/\zeta_{k+1}) = O(\log(k+2))$. Thus, each new trial-point proximal solve at outer iteration k requires $O(\kappa_{\psi,\rho} \log(k+2))$ first-order/projection iterations.

After the initial ζ_0-accurate Moreau evaluation, Strategy I requires one new proximal solve per outer iteration. Under Strategy II, every line-search trial requires one new proximal solve, and Lemma 4 bounds the number of trials per outer iteration by B_{ls}. The Moreau evaluation associated with an accepted trial is retained for the next outer iteration, so no additional current-point proximal solve is required. Therefore, the total first-order/projection work used by the Moreau subproblems through outer iteration K is

$$O\left(\kappa_{\psi,\rho} \log \frac{1}{\zeta_0}\right) + O\left(B_{\mathrm{ls}} \kappa_{\psi,\rho} \sum_{k=0}^{K-1} \log(k+2)\right) = O(B_{\mathrm{ls}} \kappa_{\psi,\rho}\, K \log K)\,.$$

The outer method itself requires only a constant number of first-order evaluations per accepted iteration and at most B_{ls} projections onto C per iteration. Its first-order/projection work is therefore $O(B_{\mathrm{ls}} K)$, which is dominated by (116). This proves the first claim.

For the balanced choice $r = 1/3$, Theorem 6 gives $K = O(\varepsilon^{-3})$. Substituting this into (116) gives (117). □

E Details on the complexity comparison

Table 1 compares representative deterministic first-order bilevel methods under a common norm-based ε convention whenever possible. Since the methods target different reformulations and solution concepts, the reported rates should be interpreted together with the "Guarantee" column rather than as complexities for an identical stopping criterion.

Solution criteria. For methods assuming a unique lower-level solution, *hyperobjective stationarity* refers to $\|\nabla\Phi(x)\| \le \varepsilon$, where $\Phi(x) = f(x, y^\star(x))$. *Penalty stationarity* refers to ε-stationarity of the penalized or smoothed surrogate used by the corresponding method. A KKT guarantee controls stationarity, constraint feasibility, and complementarity for the relevant constrained reformulation. BOME's KKT* criterion is instead based on the value-gap residual

$$\min_{\lambda \ge 0} \|\nabla f + \lambda \nabla q\|^2 + q, \qquad q = g - g^\star,$$

and is not the classical KKT residual. The weak-KKT notion of Lu–Mei (Lu & Mei, 2024) combines stationarity of their minimax reformulation with lower-level value-gap feasibility.

Rate conventions. When a method controls several residuals, the complexity displayed in Table 1 is the number of iterations required for all quantities indicated in the "Guarantee" column to be $O(\varepsilon)$. Rates originally stated using squared residuals are converted to this norm-based convention. For V-PBGD (Shen & Chen, 2023), the reported $\widetilde{O}(\varepsilon^{-2})$ rate treats its prescribed penalty coefficient as fixed. For SLM (Lu, 2023), we report the $O(\varepsilon^{-2})$ rate with its relaxation parameter treated as fixed, consistent with the fixed-relaxation setting of IVSP; the hidden constant depends polynomially on that relaxation level. For MEHA (Liu et al., 2024), stationarity of a fixed penalized problem admits the $\mathcal{O}(\nu\epsilon^{-2})$ rate reported in Lu (2025), whereas requiring both $\mathcal{O}(\epsilon)$ stationarity and $\mathcal{O}(\epsilon)$ unrelaxed Moreau-gap feasibility in its increasing-penalty analysis gives $\mathcal{O}(\epsilon^{-4})$. PNGBiO (Zhang et al., 2025) obtains an $\mathcal{O}(T^{-1/4})$ rate for both its stationarity measure and unrelaxed Moreau-gap violation, yielding the same $\mathcal{O}(\epsilon^{-4})$ joint criterion.

For IVSP, the relaxation level $\delta > 0$ is prescribed and fixed. The $\mathcal{O}(\epsilon^{-3})$ outer-iteration bound controls stationarity, feasibility, and complementarity for the KKT system of the relaxed constraint $g(x,y) - g_\rho(x,y) - \delta \le 0$. Thus, the IVSP rate and the above MEHA/PNGBiO rates do not correspond to an identical feasibility target: the latter drive the unrelaxed Moreau-gap violation to $\mathcal{O}(\epsilon)$, whereas IVSP drives $[g(x,y) - g_\rho(x,y) - \delta]_+$ to $\mathcal{O}(\epsilon)$ for fixed δ. Including the logarithmic work required by the inexact Moreau subproblems gives $\widetilde{\mathcal{O}}(\epsilon^{-3})$ first-order/projection complexity for this fixed-relaxation KKT target. Similar to SLM (Lu, 2023), the constants in the IVSP bound are not claimed to be uniform as $\delta \downarrow 0$; see Appendix A.4.

Per-iteration oracle costs. The complexity rates in Table 1 should not be interpreted as assuming identical per-iteration oracle costs. In particular, the iterations of MEHA (Liu et al., 2024) and PNGBiO (Zhang et al., 2025) are formed from first-order information without requiring lower-level objective-value evaluations. IVSP additionally requires such values to estimate the Moreau-gap residual. Under Strategy I, each outer iteration requires one new inexact Moreau solve and, after this solve, two evaluations of g to form $\widetilde{R}_{k+1}$. Thus, although its lower-level first-order/projection complexity is $O(\varepsilon^{-3}\log(1/\varepsilon))$, it also incurs $O(\varepsilon^{-3})$ lower-level value evaluations.

Under Strategy II, if B_k backtracking trials are performed at iteration k, each trial requires a new inexact Moreau solve, two lower-level value evaluations to form the trial residual, and an upper-level value evaluation for the Armijo test. Hence these trial-point costs scale by $B_k \le B_{\mathrm{ls}}$, whereas current-point first-order quantities are computed only once or retained from the preceding accepted iteration. Consequently, the corresponding value-oracle costs are $O(B_{\mathrm{ls}}\varepsilon^{-3})$, while Corollary 7 accounts separately for the $O(B_{\mathrm{ls}}\varepsilon^{-3}\log(1/\varepsilon))$ lower-level first-order/projection work.

F Experimental Details

This section provides the complete problem formulations and implementation details for the numerical experiments in Section 4.

We conducted all experiments on an Ubuntu 20.04 workstation equipped with an AMD EPYC 7413 24-Core Processor, 512 GB of system memory, and an NVIDIA RTX A5000 GPU with 24 GB of GPU memory. We used Python 3.10.12 and PyTorch 2.6.0 with CUDA 12.4.

All optimization runs are deterministic after fixing the problem instances, data splits, and initializations. In the real-world experiments, we use fixed training, validation, and test sets for all methods, and no stochastic mini-batching is used during optimization. The test set is used only for evaluation. All compared methods are run under the same wall-clock budget within each problem: 1 second for the synthetic experiment, 60 seconds for both MNIST and FashionMNIST hyper-cleaning, and 600 seconds for Omniglot few-shot learning. Wall-clock time includes all optimization computations, including inner-loop computations, but excludes data loading and offline metric evaluation. The main hyperparameters of all compared methods are tuned separately for each problem using method-specific candidate values. For the real-world tasks, we use the same validation-based tuning protocol and comparable tuning budgets for all methods within each problem. Test performance is not used for hyperparameter selection.

Implementation details. We make two practical implementation choices for Algorithm 3: a finite-precision relaxation of the inexact Armijo test and a safeguarded Barzilai–Borwein initialization of the backtracking parameter.

First, in finite-precision arithmetic, we use the following numerically relaxed version of the inexact Armijo condition (8):

$$\begin{aligned}&\gamma_k f(w_L^k) + \phi_{\mu_k}\big(\widetilde{R}_{k,L}^{+} + \zeta_{k+1}\big)\\ &\quad\le \gamma_k f(w^k) + \phi_{\mu_k}(\widetilde{R}^k) - \frac{c_1}{2\mu_k}\|w_L^k - w^k\|^2 + \tau_{k,L} + 10^{-12}.\end{aligned} \tag{118}$$

The additive tolerance 10^{-12} is used only to prevent numerical stalling due to floating-point round-off when the two sides of (8) are nearly identical, and is not part of the theoretical algorithm or convergence analysis.

Second, for the backtracking variant of IVSP, we initialize the line-search parameter using a safeguarded Barzilai–Borwein (BB) estimate (Barzilai & Borwein, 1988), similarly to the implementation of Xu et al. (2026). Specifically, define

$$\Delta w_k := w^k - w^{k-1}, \qquad \Delta v_k := \gamma_k\big(\nabla f(w^k) - \nabla f(w^{k-1})\big) + \widetilde{\nabla} G_k - \widetilde{\nabla} G_{k-1}, \qquad k \ge 1.$$

We form

$$L_k^{\mathrm{BB}} = \mu_k \frac{\|\Delta v_k\|^2}{|\langle \Delta w_k, \Delta v_k\rangle|}.$$

We set $L_{0,0} = L_{\mathrm{init}}$. For $k \ge 1$, if $|\langle \Delta w_k, \Delta v_k\rangle| > 10^{-24}$ and $L_k^{\mathrm{BB}} \in [L_{\min}, L_{\max}]$, we set $L_{k,0} = L_k^{\mathrm{BB}}$; otherwise, $L_{k,0} = \min\{L_{\max}, \max\{L_{\min}, L_{k-1,0}/2\}\}$. Starting from $L = L_{k,0}$, we repeatedly replace L by $2L$ until (118) is satisfied.

Evaluation of the Moreau-gap violation. For reporting the estimated Moreau-gap constraint violation, we compute an independent approximation $\widehat{z}$ of the proximal point $z_\rho(x, y)$ by applying projected gradient with monotone backtracking to the full lower-level Moreau subproblem, initialized at $z = y$. For each task, we use the same proximal parameter ρ as IVSP and apply it to all compared methods. The proximal solve is terminated when the norm of the projected-gradient mapping is at most 10^{-3}. Using the resulting $\widehat{z}$, we compute $\widehat{H}_\rho(x, y) = g(x, y) - g(x, \widehat{z}) - \frac{1}{2\rho}\|\widehat{z} - y\|^2$ and report $\widehat{\Phi}_\rho^{\mathrm{f}}(x, y) := [\widehat{H}_\rho(x, y) - \delta]_+$, which estimates the feasibility residual Φ_ρ^{f} in Definition 2. This independent proximal solve is used only for evaluation.

Common IVSP settings. For IVSP, following Xu et al. (2026), we use the centered-Huber smoothing of the plus function,

$$\phi_\mu(t) = \begin{cases} 0, & t \le -\mu/2, \\ \dfrac{(t + \mu/2)^2}{2\mu}, & |t| < \mu/2, \\ t, & t \ge \mu/2. \end{cases}$$

Table 2: Main hyperparameter settings for the synthetic nonconvex experiment.

Method	Hyperparameters
IVSP	$\mu_0 = 9 \times 10^{-3}$, $B = 1$, $\gamma_0 = 0.9$, $\bar{\zeta} = 3$, $\rho = 0.5$.
PNGBiO	$\alpha_k = 0.70(k+1)^{-1/2}$, $\beta_k = 0.20(k+1)^{-0.05}$, $\eta_k = 0.015$, $\gamma = 0.5$, $c_k = 0.004(k+1)^{0.21}$.
MEHA	$\alpha_k = 0.004$, $\beta_k = 0.02$, $\eta_k = 0.05$, $\gamma = 0.5$, $c_k = 0.1(k+1)^{0.28}$.
SLM	$\eta = 1.5 \times 10^{-3}$, $\alpha = 4.0 \times 10^{-3}$, $\gamma = 0.3$, $\tau = 10^{-8}$, $T_r = 10$.
F^{2}SA	$\alpha_k = 0.01$, $\gamma_k = 0.1$, $\lambda_0 = 0.0195$, $T = 1$.
V-PBGD	$\alpha = 0.003$, $\beta = 0.05$, $\gamma_k = 0.0025(k+1)^{0.6}$, $T = 1$.
BOME	$\xi = 0.05$, $\alpha = 0.05$, $\eta = 0.01$, $T = 1$.

This is the optimal inner $1/\mu$-smoothing of $[\cdot]_+$ in the sense of Samakhoana & Grimmer (2026), with uniform approximation error $\mu/8$. Accordingly, the constants in (24) and (22) are $a_\phi = 1/8$ and $b_\phi = 1$, respectively.

Across all experiments, we use $\mu_k = \mu_0(\lfloor k/B \rfloor + 1)^{-r}$, $\widehat{\gamma}_t = \gamma_0(t+1)^{-1/2}$, and $\zeta_k = \bar{\zeta}\mu_k^q$, with $r = 1/3$ and $q = 1/r + 0.01$. The common parameter settings are $L_{\min} = 10^{-11}$, $L_{\max} = 10^{11}$, $\delta = 10^{-3}$, $c_1 = 10^{-4}$, $c_2 = 1$, $\underline{\vartheta} = 10^{-24}$, $\hat{\vartheta} = 1.99$, and $B_R = 10^3$.

Synthetic nonconvex problem. For the synthetic experiment, we use the nonconvex bilevel problem described in Section 4.1, with $n = 1000$, $a = 2$, $c_i = 2$, $x^0 = -6$, and $y^0 = 0$.

For IVSP, we additionally use $\sigma = 1$, $L_{\text{init}} = 1$, $L_{xy} = \sqrt{1000}$, and $L_{yy} = 1$. The main method-specific hyperparameters are summarized in Table 2.

Data hyper-cleaning. For data hyper-cleaning (Ren et al., 2018; Liu et al., 2024; Zhang et al., 2025), let $\mathcal{D}_{\text{tr}} = \{(u_i, \widetilde{q}_i)\}_{i=1}^n$ denote the corrupted training set and let $\mathcal{D}_{\text{val}}$ denote the clean validation set. The UL variable $x \in [-100, 100]^n$ assigns the weight $\omega(x_i)$ to training example i, where $\omega(t) = 1/(1+e^{-t})$ is the sigmoid function.

The bilevel formulation is

$$\min_{\substack{x\in[-100,100]^{5000}\\ y\in[-100,100]^{25450}}} \quad \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{(u,q)\in\mathcal{D}_{\text{val}}} \ell(h_y(u), q)$$

$$\text{s.t.} \quad y \in \underset{\widetilde{y}\in[-100,100]^{25450}}{\arg\min} \frac{1}{|\mathcal{D}_{\text{tr}}|} \sum_{i=1}^n \omega(x_i)\ell(h_{\widetilde{y}}(u_i), \widetilde{q}_i).$$

The classifier h_y is a two-layer fully connected neural network with 784 input units, 32 sigmoid hidden units, and 10 output units. It therefore contains $784\times32+32+32\times10+10 = 25{,}450$ parameters. The sigmoid activation makes the LL objective smooth, while the neural-network parameterization makes it nonconvex. This problem is an instance of the sigmoid-reweighted classification model in Corollary 18. Indeed, with $M = B_y = 100$, $d_h = 32$, and $m = 10$,

$$\frac{2B_y(d_h+1) + \log m}{1 + e^M} < 2.5 \times 10^{-40} < 10^{-3} = \delta.$$

Hence, ENNAMCQ holds for both hyper-cleaning experiments.

For both MNIST (LeCun et al., 1998) and FashionMNIST (Xiao et al., 2017), we use 5,000 training, 5,000 validation, and 10,000 test examples. We randomly corrupt 50% of the training labels by replacing each selected label with an incorrect class label, while the validation and test sets remain clean. The same corrupted training set is used for all compared methods. The variables are initialized with $x^0 = 0$ and $y^0 = 0.01\,\varepsilon$, where ε has independent standard normal entries. We report validation loss, test error, and the Moreau-gap constraint violation as functions of wall-clock time.

Table 3: Main hyperparameter settings for data hyper-cleaning.

Method	MNIST	FashionMNIST
IVSP	$\mu_0 = 0.02, B = 20, \gamma_0 = 10, \bar{\zeta} = 9, \rho = 0.05$	$\mu_0 = 10, B = 2, \gamma_0 = 1, \bar{\zeta} = 5 \times 10^{-3}, \rho = 0.1$
PNGBiO	$\alpha_k = \beta_k = (k+1)^{-1/2}, \eta_k = 0.05, \gamma = 0.1, c_k = 0.2(k+1)^{0.25}$	$\alpha_k = \beta_k = (k+1)^{-1/2}, \eta_k = 0.05, \gamma = 0.4, c_k = 0.2(k+1)^{0.25}$
MEHA	$\alpha_k = 9, \beta_k = (k+1)^{-0.6}, \eta_k = 0.05, \gamma = 5, c_k = 0.2(k+1)^{0.005}$	$\alpha_k = 10, \beta_k = (k+1)^{-0.6}, \eta_k = 0.03, \gamma = 12, c_k = 0.1(k+1)^{0.005}$
SLM	$\eta = 2, \alpha = 2, \gamma = 0.2, \tau = 0.01, T_r = 1$	$\eta = 2, \alpha = 2, \gamma = 0.2, \tau = 0.01, T_r = 1$
F^{2}SA	$\alpha_k = 3(k+1)^{-1/3}, \gamma_k = 0.2, \lambda_0 = 0.1, T = 10$	$\alpha_k = 4(k+1)^{-1/3}, \gamma_k = 0.2, \lambda_0 = 0.1, T = 10$
V-PBGD	$\alpha = 2, \beta = 1, \gamma = 1, T = 1$	$\alpha = 1, \beta = 1, \gamma = 1, T = 1$
BOME	$\xi = 3, \alpha = 1, \eta = 0.5, T = 1$	$\xi = 2, \alpha = 1, \eta = 0.5, T = 1$

Table 4: Main hyperparameter settings for few-shot learning on Omniglot.

Method	Hyperparameters
IVSP	$\mu_0 = 1.5, B = 20, \gamma_0 = 0.035, \bar{\zeta} = 1, \rho = 0.1.$
PNGBiO	$\alpha_k = 0.35(k+1)^{-1/2}$, $\beta_k = 0.05(k+1)^{-1/2}$, $\eta_k = 10^{-3}$, $\gamma = 100$, $c_k = 0.067(k+1)^{0.08}$.
MEHA	$\alpha_k = 0.08(k+1)^{-1/2}$, $\beta_k = 0.05(k+1)^{-1/2}$, $\eta_k = 10^{-3}$, $\gamma = 100$, $c_k = 0.008(k+1)^{0.08}$.
SLM	$\eta = 55, \alpha = 10, \gamma = 0.01, \tau = 0.01, T_r = 10.$
F^{2}SA	$\alpha_k = 5(k+1)^{-1/3}, \gamma_k = 0.5, \xi = 0.15, \lambda_0 = 1, T = 5.$
V-PBGD	$\alpha = 1, \beta = 0.01, \gamma = 10, T = 10.$
BOME	$\xi = 3, \alpha = 1, \eta = 0.01, T = 10.$

We additionally use $L_{\text{init}} = 1$, $L_{xy} = 1$, $\sigma = 1$ for both hyper-cleaning tasks, with $L_{yy} = 1$ on MNIST and $L_{yy} = 20$ on FashionMNIST. The main method-specific hyperparameters are summarized in Table 3.

Few-shot learning. We consider episodic few-shot classification on Omniglot (Lake et al., 2015), following the bilevel meta-learning formulation used in Liu et al. (2024). The 1,623 character classes are randomly partitioned into 1,100 training, 100 validation, and 423 test classes using a fixed class split. The three class sets are mutually disjoint. We consider the 10-way 1-shot setting with five query examples per class. We generate 128 fixed training episodes, 100 validation episodes, and 100 test episodes.

For each training episode $i \in \{1, \dots, T\}$, where $T = 128$, let $\mathcal{S}_i$ and $\mathcal{Q}_i$ denote its support and query sets, respectively. Each support set contains one example from each of 10 classes, and each query set contains five examples from each class.

The UL variable x parameterizes a shared four-block convolutional feature extractor. Each block consists of a 3×3 convolution with 64 output channels, ReLU activation, 2×2 max pooling, and batch normalization. For Omniglot, the successive spatial resolutions are $28 \to 14 \to 7 \to 3 \to 1$, resulting in a 64-dimensional feature vector. Batch-normalization statistics are computed separately for each task from the image set being processed.

For an episode dataset $\mathcal{D} = \{(u_j, q_j)\}_{j=1}^{|\mathcal{D}|}$, let $F_{x,\mathcal{D}}(u_j) \in \mathbb{R}^{64}$ denote the feature vector of u_j produced by the shared feature extractor, with the batch-normalization statistics computed from $\mathcal{D}$. For each training episode i, the LL variable is a linear softmax classifier $y_i = (W_i, b_i)$, where $W_i \in \mathbb{R}^{10\times 64}$ and $b_i \in \mathbb{R}^{10}$. We define

$$\mathcal{L}_{\mathcal{D}}(x, y_i) = \frac{1}{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{D}|} \ell(W_i F_{x,\mathcal{D}}(u_j) + b_i, q_j)\,,$$

where ℓ denotes the 10-class cross-entropy loss. The corresponding bilevel problem is

$$\begin{aligned} &\min_{\substack{x\in[-5,5]^{111936}\\ y\in[-20,20]^{83200}}} \quad \frac{1}{T}\sum_{i=1}^{T}\mathcal{L}_{\mathcal{Q}_i}(x,y_i) \\ \text{s.t.} \quad &y \in \operatorname*{arg\,min}_{\widetilde{y}\in[-20,20]^{83200}} \quad \frac{1}{T}\sum_{i=1}^{T}\mathcal{L}_{\mathcal{S}_i}(x,\widetilde{y}_i). \end{aligned} \tag{119}$$

For fixed x, the LL objective in (119) is convex in y, since each task-specific model is a linear softmax classifier. It is not strongly convex: in particular, the softmax cross-entropy is invariant under a common shift of all class logits, which induces nontrivial flat directions in the LL objective. As established in Proposition 16, this structure nevertheless satisfies ENNAMCQ for every $\delta > 0$. Thus, this experiment provides a large-scale convex but non-strongly-convex LL instance for which the constraint qualification required by our analysis holds.

For each validation and test episode, we initialize a fresh linear classifier at zero and adapt it for 10 gradient steps on the support set with stepsize 0.1, while keeping the feature extractor fixed. We then evaluate query loss and classification error using the adapted classifier.

All training episodes are fixed before optimization and are shared across all compared methods. The optimization is therefore deterministic after fixing the dataset, episode construction, and initialization. The convolutional parameters are initialized according to the standard PyTorch convolutional initialization, the batch-normalization scale and bias are initialized to one and zero, respectively, and all task-specific linear classifiers are initialized at zero. We use data seed 42, initialization seed 43, and class-split seed 0.

For IVSP, we additionally use $\sigma = 0$, $L_{\text{init}} = 0.1$, $L_{xy} = 1$, and $L_{yy} = 15$. The main method-specific hyperparameters are summarized in Table 4.