
Single-Loop Penalty Methods for Bilevel Optimization with Non-Smooth Constrained Lower Levels

Samuel Rey
King Juan Carlos University
samuel.rey.escudero@urjc.es

Antonio G. Marques
King Juan Carlos University
antonio.garcia.marques@urjc.es

Tianyi Chen
Cornell University
tianyi.chen@cornell.edu

Abstract

We study bilevel optimization (BLO) problems whose lower level (LL) is strongly convex but *non-smooth*, comprising a smooth term together with up to two convex proximable non-smooth terms. Convex constraints are included within this two-term template by taking one of the non-smooth terms to be an indicator function, whose proximal operator is the corresponding projection. Such LLs are pervasive in machine learning, statistics, and signal processing—sparse regression, robust estimation with sparse outliers, fused-Lasso and total-variation models, low-rank estimation through nuclear-norm penalties, and constrained structured estimation—yet they preclude the direct use of hypergradient methods, which differentiate through the LL solution map or its optimality conditions. We develop a single-loop, hypergradient-free algorithm based on a value-function penalty. The method only requires (i) an inexact LL solve and (ii) proximal access to the non-smooth terms; it never differentiates the non-smooth solution map. Our analysis has three pillars: a regularity result showing that the LL value function is differentiable with a Lipschitz gradient even when the LL is non-smooth, with a gradient computable from the LL minimizer; a single-loop inexact proximal-gradient scheme that attains an $O(1/K)$ stationarity rate under a controlled inner-accuracy schedule; and a clean dichotomy between two inner-solver regimes—forward-backward splitting when the composite proximal operator is tractable, and Davis–Yin three-operator splitting when two non-smooth terms must be handled separately. We instantiate the framework on two representative problems: robust regression with sparse outlier rejection (one non-smooth term) and fused-Lasso/total-variation regression (two non-smooth terms), the latter additionally requiring an inexact *outer* proximal step that we analyze. We outline experiments validating the predicted rates and the practical behavior of both regimes.

1 Introduction

Bilevel optimization (BLO) provides a principled language for hierarchical problems in which an upper-level (UL) objective is optimized subject to the optimality of a lower-level (LL) problem [1, 2]. In machine learning it underlies hyperparameter optimization, meta-learning, adversarial training, and data-driven model selection [3–5]. A recurring feature of many of these problems, especially in statistics and signal processing, is that the lower level is *non-smooth*: it contains sparsity-inducing, low-rank-inducing, or more general structure-inducing regularizers (ℓ_1 , group/fused ℓ_1 , total variation,

nuclear norm) and/or hard convex constraints. Examples include sparse and robust regression, estimators with sparse outlier variables, fused-Lasso and total-variation reconstruction, low-rank matrix estimation, dictionary learning, and constrained structured estimation.

Two methodological families dominate the BLO literature. *Hypergradient* methods approximate the gradient of the UL objective by differentiating through the LL optimality conditions, via implicit differentiation or unrolling [3]. These are powerful when the LL is smooth and well-conditioned, but they become delicate when the LL is non-smooth: the solution map may be non-differentiable, active sets change, and projections appear, so one must resort to non-smooth implicit differentiation [6–9] or to *smoothing* the non-smooth terms, which trades exactness for conditioning. *Penalty* (value-function) methods instead relax the bilevel constraint by penalizing the LL optimality gap $g(\mathbf{x}, \mathbf{y}) - v(\mathbf{x})$, where v is the LL value function [10, 11]. They yield single-loop schemes and are attractive in the non-smooth regime, but their theory has mostly been developed for smooth LLs.

This paper. We develop a single-loop penalty method for bilevel problems whose lower level is strongly convex and of the form

$$g(\mathbf{x}, \mathbf{y}) = g_s(\mathbf{x}, \mathbf{y}) + g_{n,1}(\mathbf{y}) + g_{n,2}(\mathbf{y}), \quad (1)$$

where g_s is smooth and $g_{n,1}, g_{n,2}$ are convex and proximal. This notation covers the single non-smooth case by setting one term to zero, and it covers convex constraints by taking, for example, $g_{n,2} = \iota_{\mathcal{O}}$, in which case $\text{prox}_{g_{n,2}} = \text{proj}_{\mathcal{O}}$. Thus, the Davis–Yin regime considered here separates at most two non-smooth terms; it does not assume access to three independent proximal/projection operators. The method optimizes the value-function penalty by inexact proximal-gradient descent, approximating the value-function gradient through an inexact LL minimizer. Crucially, it *never differentiates the non-smooth solution map*: it only needs an inexact LL solve and the proximal operators of the non-smooth terms. In this sense, the algorithm is single-loop at the penalty level: each outer iteration performs one proximal-gradient update, supported by a controlled number of inner splitting steps prescribed by the accuracy schedule.

1.1 Contributions

- **Value-function regularity for non-smooth LLs.** We show that, under strong convexity of the LL and Lipschitz smoothness of the UL-dependence, the LL solution map is single-valued and Lipschitz and the value function v is differentiable with a Lipschitz gradient, given in closed form by $\nabla v(\mathbf{x}) = \nabla_{\mathbf{x}} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$ (Theorem 1), *even when the lower level contains one or two non-smooth proximal terms*. We state the result in a form tailored to value-function penalty algorithms with exact proximal non-smooth LL terms.
- **A single-loop, hypergradient-free penalty algorithm.** We propose a proximal-gradient scheme on the value-function penalty that uses only an inexact LL solve and proximal access to the non-smooth terms (Algorithm 1), and we prove an $O(1/K)$ stationarity rate under a controlled inner-accuracy schedule (Theorem 2). We additionally handle the case in which the *outer* proximal step is itself intractable in closed form, via an inexact-prox extension (Proposition 1).
- **Two inner-solver regimes.** We characterize precisely when forward–backward splitting suffices and when Davis–Yin three-operator splitting is needed, and translate the inner-accuracy schedule into inner-iteration counts for both (Corollary 1, Section 6). In the Davis–Yin regime, the inner guarantee (Lemma 3) covers one finite-valued (locally Lipschitz) non-smooth term, e.g. any norm-type penalty, together with a second term that may be a constraint indicator; it does not cover two arbitrary extended-valued terms (Remark 3). We instantiate the theory on (i) robust regression with sparse outlier rejection (one non-smooth term, forward–backward), and (ii) fused-Lasso/total-variation regression (two non-smooth terms, Davis–Yin), and outline experiments validating the predicted behavior.

1.2 Related work

Penalty / value-function BLO. Value-function penalties reformulate the bilevel problem as a single-level program by penalizing the optimality gap. [10] analyze a value-function penalty under a (proximal) Polyak–Łojasiewicz / strong-convexity condition and establish a closed form for the value-function gradient, but for *smooth* LLs. [11] study first-order penalty methods for bilevel problems whose lower level is possibly non-smooth convex, but their subproblems take a structured

minimax form rather than a single-loop proximal scheme. A growing body of recent work develops constrained and single-loop penalty methods, typically under smoothness of the coupling or of the constraints. Our analysis is closest in spirit to [10], which we extend to LLs handled by proximal splitting with up to two non-smooth convex terms, one of which may be a constraint indicator.

Non-smooth hypergradients. The alternative route differentiates the non-smooth solution map. [6] differentiate fixed points of proximal (coordinate) descent for sparse, non-smooth learning; [7] provide a rigorous foundation via conservative Jacobians; [8] approximate the LL solution map by a differentiable iterative scheme; and [9] analyze total-variation LLs through Bouligand-subdifferential systems. [12] learn variational-model parameters via bilevel optimization. These methods either differentiate (a smoothed or carefully reparameterized) solution map or smooth the non-smoothness; our penalty route sidesteps both.

Non-smooth LLs in data science. Lower-level objectives with ℓ_1 sparsity, group/fused sparsity, total variation, nuclear-norm low-rank regularization, sparse outliers, or convex constraints arise throughout sparse regression, robust statistics, imaging, matrix recovery, and structured estimation. We do not survey applications exhaustively; rather, we use two of them, robust regression with sparse outliers [13] and fused-Lasso/total-variation regression [14–16], as concrete instances of the abstract model (1), and we collect further instances in Section 7.3.

1.3 Notation

We use boldface lowercase for vectors, boldface uppercase for matrices, and calligraphic letters for sets. For an integer N , $[N] := \{1, \dots, N\}$. $\|\cdot\|_2$ is the Euclidean (Frobenius) norm, $\|\cdot\|_1$ the entrywise ℓ_1 norm, and $\|\cdot\|_*$ the nuclear norm. For a closed convex set $\mathcal{C}$, $\iota_{\mathcal{C}}$ is its indicator, $\text{proj}_{\mathcal{C}}$ the Euclidean projection, and $N_{\mathcal{C}}$ the normal cone. For a proper closed convex function ϕ and $\beta > 0$, $\text{prox}_{\beta\phi}(\mathbf{z}) := \arg \min_{\mathbf{u}} \{\frac{1}{2}\|\mathbf{u} - \mathbf{z}\|_2^2 + \beta\phi(\mathbf{u})\}$. $d_{\mathcal{C}}(\mathbf{z}) := \min_{\mathbf{u} \in \mathcal{C}} \|\mathbf{z} - \mathbf{u}\|_2$.

2 Problem formulation and preliminaries

2.1 Bilevel problem and value-function penalty

We follow the standard BLO convention: $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^n$ are the UL variables and $\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^m$ the LL variables. We consider

$$\begin{aligned} \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) \\ \text{s.t. } \mathbf{y}^*(\mathbf{x}) = \arg \min_{\mathbf{y} \in \mathcal{Y}} g(\mathbf{x}, \mathbf{y}), \quad g(\mathbf{x}, \mathbf{y}) := g_s(\mathbf{x}, \mathbf{y}) + g_{n,1}(\mathbf{y}) + g_{n,2}(\mathbf{y}), \end{aligned} \quad (2)$$

where f is smooth (possibly non-convex) with Lipschitz gradient, g_s is smooth with Lipschitz gradient, $g_{n,1}, g_{n,2}$ are proper closed convex and proximal, and $g(\mathbf{x}, \cdot)$ is strongly convex (so $\mathbf{y}^*(\mathbf{x})$ is unique). We let $\mathcal{S}(\mathbf{x})$ denote the (singleton) LL minimizer set. If a hard convex constraint $\mathbf{y} \in \mathcal{O}$ is part of the lower level, it is included as one of the two non-smooth terms, e.g., $g_{n,2} = \iota_{\mathcal{O}}$; its proximal operator is then $\text{proj}_{\mathcal{O}}$.

Remark 1 (Generality of the splitting). *The analysis covers any LL of the form (smooth) + (one or two convex proximal terms). The single-term case is recovered by setting $g_{n,2} \equiv 0$. A constrained problem with one regularizer, $g_s + g_{n,1} + \iota_{\mathcal{O}}$, is recovered by setting $g_{n,2} = \iota_{\mathcal{O}}$. A problem with two regularizers and an additional independent constraint would contain three non-smooth terms and is not covered by standard Davis–Yin splitting unless two of them can be grouped into a proximal composite.*

Value-function penalty. Let $v(\mathbf{x}) := \min_{\mathbf{y} \in \mathcal{Y}} g(\mathbf{x}, \mathbf{y})$ be the LL value function. The optimality gap $g(\mathbf{x}, \mathbf{y}) - v(\mathbf{x}) \geq 0$ vanishes iff $\mathbf{y} = \mathbf{y}^*(\mathbf{x})$. For $\gamma > 0$ we minimize

$$\min_{\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}} F_{\gamma}(\mathbf{x}, \mathbf{y}) := f(\mathbf{x}, \mathbf{y}) + \gamma(g(\mathbf{x}, \mathbf{y}) - v(\mathbf{x})). \quad (3)$$

Local minimizers of (3) for large enough γ are approximate bilevel solutions, made precise next.

Definition 1 (ϵ -approximate bilevel problem [10]). For $\epsilon > 0$, an optimization problem is an ϵ -approximate bilevel problem associated with (2) if, for some $\rho > 0$: (i) every feasible $(\mathbf{x}, \mathbf{y})$ satisfies $d_{S(\mathbf{x})}^2(\mathbf{y}) \leq \rho\epsilon$ with $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$; and (ii) its optimal value f_ϵ^* satisfies $|f^* - f_\epsilon^*| \leq L\sqrt{\rho\epsilon}$, where f^* is the optimal value of (2).

2.2 Standing assumptions

AS1. For any $\mathbf{x} \in \mathcal{X}$, $f(\mathbf{x}, \cdot)$ is L_c -Lipschitz continuous on the bounded region containing the LL solution path $\{\mathbf{y}^*(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$, the outer y -iterates, and the LL estimates produced by the inner solver.

AS2. For any $\mathbf{x} \in \mathcal{X}$, $g(\mathbf{x}, \cdot)$ is σ_g -strongly convex on $\mathcal{Y}$.

AS3. f and g_s are L_f - and L_g -Lipschitz smooth in both arguments.

AS4. $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{Y} \subseteq \mathbb{R}^m$ are nonempty, closed, and convex sets.

AS5. g is differentiable in $\mathbf{x}$, and $\nabla_{\mathbf{x}}g(\mathbf{x}, \mathbf{y})$ is L_x -Lipschitz in both arguments.

In both instances of Section 7 the smooth part g_s is quadratic, so ∇g_s is globally Lipschitz and Assumption 3 holds globally, without any artificial restriction of $\mathcal{Y}$. For the quadratic validation losses of Section 7, Assumption 1 likewise follows from boundedness of the relevant iterates and solution path.

2.3 How strong convexity is met, and the role of $\mathbf{x}$

Assumption 2 is satisfied in two ways. It can hold *natively*, when the convex LL already contains a strongly convex smooth term (e.g. a ridge / elastic-net term or a data-fidelity term with full-rank design). When it does not, it can be *induced* by adding a proximal (Tikhonov-type) term

$$\frac{\eta}{2} \|\mathbf{y} - \mathbf{p}\|_2^2, \quad \eta > 0, \quad (4)$$

to a reference point $\mathbf{p}$. This is classical proximal regularization. What matters for our framework is that, when $\mathbf{p}$ is treated as an *upper-level* variable (so $\mathbf{x}$ contains $\mathbf{p}$), the term (4) places the UL dependence in the *smooth* part g_s , which keeps the non-smooth terms independent of $\mathbf{x}$ (the regime of Theorem 2). This preserves the Lipschitzness of $\nabla_{\mathbf{x}}g$ (Assumption 5) while rendering the lower level function strongly convex. In particular, Theorem 1 below then gives $\nabla v(\mathbf{x}) = \eta(\mathbf{p} - \mathbf{y}^*(\mathbf{x}))$ as an immediate consequence, requiring only the LL minimizer. We introduce (4) purely as one potential way to induce strong convexity; many others exist and richer structural consequences of the particular quadratic coupling in (4) are neither developed nor needed here.

3 Value-function regularity

We first record that strong convexity yields quadratic growth, then prove that the value function of a *non-smooth* lower level is differentiable with a Lipschitz gradient.

Lemma 1 (Quadratic growth). For any $\mathbf{x} \in \mathcal{X}$, if $g(\mathbf{x}, \cdot)$ is σ_g -strongly convex, then with $\mu := 2/\sigma_g$,

$$g(\mathbf{x}, \mathbf{y}) - v(\mathbf{x}) \geq \frac{1}{\mu} d_{\mathbf{y}^*(\mathbf{x})}^2(\mathbf{y}), \quad \forall \mathbf{y} \in \mathcal{Y}. \quad (5)$$

Proof. Strong convexity gives $g(\mathbf{x}, \mathbf{y}) - g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) \geq \frac{\sigma_g}{2} \|\mathbf{y} - \mathbf{y}^*(\mathbf{x})\|_2^2$ for all $\mathbf{y} \in \mathcal{Y}$ [17, Th. 5.25]; substitute $\mu = 2/\sigma_g$ and $v(\mathbf{x}) = g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$. $\square$

Theorem 1 (Solution-map Lipschitzness and value-function smoothness). Suppose $g(\mathbf{x}, \mathbf{y})$ decomposes according to (2), where (i) g_s is differentiable in both variables and $\nabla_{\mathbf{x}}g_s(\mathbf{x}, \mathbf{y})$ is L_x -Lipschitz in both arguments (Assumption 5); (ii) $g_{n,1}, g_{n,2}$ are proper, closed, convex, proximable, possibly non-smooth, and independent of $\mathbf{x}$; and (iii) $g(\mathbf{x}, \cdot)$ is σ_g -strongly convex on the convex set $\mathcal{Y}$ for every $\mathbf{x} \in \mathcal{X}$ (Assumptions 2 and 4). Let $\mathbf{y}^*(\mathbf{x})$ be the unique LL minimizer. Then: (a) $\mathbf{y}^*$ is L_S -Lipschitz on $\mathcal{X}$, with $L_S = (L_x + \sqrt{L_x^2 + 2\sigma_g L_x})/\sigma_g$; (b) v is differentiable on $\mathcal{X}$ with $\nabla v(\mathbf{x}) = \nabla_{\mathbf{x}}g_s(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) = \nabla_{\mathbf{x}}g(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$, and ∇v is L_v -Lipschitz with $L_v = L_x(1 + L_S)$.

Proof. The proof is given in Appendix A. $\square$

Algorithm 1: Single-loop value-function penalty method for non-smooth constrained BLO.

Input: $(\mathbf{x}^1, \mathbf{y}^1) \in \mathcal{Z}$; initial inner state state^1 (warm start); stepsizes $\alpha, \beta > 0$; penalty $\gamma > 0$; inner counts $\{T_k\}$; iterations K .

```

1 for  $k = 1, \dots, K$  do
  // (1) inexact LL solve, warm-started (FB or Davis-Yin; Section 6)
2    $(\text{state}^{k+1}, \hat{\mathbf{y}}^{k+1}) \leftarrow \text{InnerSolve}(\mathbf{x}^k, \text{state}^k, T_k)$ ; // FB: state is the primal iterate and
    $\hat{\mathbf{y}}^{k+1}$  its last value; DY: state is the governing variable  $\mathbf{s}$  and  $\hat{\mathbf{y}}^{k+1}$  the
   shadow output (18)
  // (2) UL proximal-gradient step
3   Form  $\widehat{\nabla} \bar{F}_\gamma(\mathbf{z}^k \mid \hat{\mathbf{y}}^{k+1})$  as in (6);
4    $\mathbf{z}^{k+1} \leftarrow \text{prox}_{\alpha r}(\mathbf{z}^k - \alpha \widehat{\nabla} \bar{F}_\gamma(\mathbf{z}^k \mid \hat{\mathbf{y}}^{k+1}))$ ; //  $r = \gamma(g_{n,1} + g_{n,2}) + \iota_{\mathcal{Z}}$ 
5 end
Output:  $(\mathbf{x}^{K+1}, \mathbf{y}^{K+1})$  and LL estimate  $\hat{\mathbf{y}}^{K+1}$ .

```

Remark 2. *Theorem 1 requires no differentiability of the non-smooth terms in $\mathbf{y}$ and no differentiation of the solution map; the regularity of v comes solely from strong convexity (uniqueness and Lipschitzness of $\mathbf{y}^*$) and smoothness of the $\mathbf{x}$ -dependence. The single-non-smooth-term case ($g_{n,2} \equiv 0$) is included as the forward-backward regime, while the two-term case motivates the Davis-Yin regime analyzed below.*

4 Single-loop penalty algorithm

Composite structure and inexact gradient. Write $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, and split $F_\gamma = \bar{F}_\gamma + r$ with smooth part $\bar{F}_\gamma(\mathbf{z}) := f(\mathbf{x}, \mathbf{y}) + \gamma(g_s(\mathbf{x}, \mathbf{y}) - v(\mathbf{x}))$ and non-smooth part $r(\mathbf{z}) := \gamma(g_{n,1}(\mathbf{y}) + g_{n,2}(\mathbf{y})) + \iota_{\mathcal{Z}}(\mathbf{z})$, where $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. If a convex constraint involving $\mathbf{y}$ is present, it is one of the two terms $g_{n,i}$, not an additional third proximal operator. By Theorem 1, $\nabla v(\mathbf{x}) = \nabla_{\mathbf{x}} g_s(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$ because the non-smooth LL terms are independent of $\mathbf{x}$. Replacing $\mathbf{y}^*(\mathbf{x}^k)$ by an inexact LL solution $\hat{\mathbf{y}}^{k+1}$ gives the following blockwise inexact smooth gradient:

$$\widehat{\nabla} \bar{F}_\gamma(\mathbf{z}^k \mid \hat{\mathbf{y}}^{k+1}) := \nabla f(\mathbf{z}^k) + \gamma \begin{bmatrix} \nabla_{\mathbf{x}} g_s(\mathbf{x}^k, \mathbf{y}^k) - \nabla_{\mathbf{x}} g_s(\mathbf{x}^k, \hat{\mathbf{y}}^{k+1}) \\ \nabla_{\mathbf{y}} g_s(\mathbf{x}^k, \mathbf{y}^k) \end{bmatrix}. \quad (6)$$

The UL update is a proximal-gradient step $\mathbf{z}^{k+1} = \text{prox}_{\alpha r}(\mathbf{z}^k - \alpha \widehat{\nabla} \bar{F}_\gamma(\mathbf{z}^k \mid \hat{\mathbf{y}}^{k+1}))$. The inner solver only needs to drive the LL gap below a tolerance ϵ_k (no high-accuracy solve), which makes the method single-loop.

Lemma 2 (Inexact proximal-gradient descent). *Let $\Phi = \psi + r$ with ψ having L -Lipschitz gradient and r proper closed convex. Fix $\mathbf{z} \in \text{dom } r$, $\mathbf{e} \in \mathbb{R}^d$, $\alpha \in (0, 1/L]$, and set $\bar{\mathbf{z}}^+ := \text{prox}_{\alpha r}(\mathbf{z} - \alpha \nabla \psi(\mathbf{z}))$, $\mathbf{z}^+ := \text{prox}_{\alpha r}(\mathbf{z} - \alpha(\nabla \psi(\mathbf{z}) + \mathbf{e}))$, $G_\alpha(\mathbf{z}) := \frac{1}{\alpha}(\mathbf{z} - \bar{\mathbf{z}}^+)$. Then for universal $c_1, c_2 > 0$,*

$$\Phi(\mathbf{z}^+) \leq \Phi(\mathbf{z}) - c_1 \alpha \|G_\alpha(\mathbf{z})\|_2^2 + c_2 \alpha \|\mathbf{e}\|_2^2. \quad (7)$$

Proof. The proof (with $c_1 = 3/16$, $c_2 = 19/8$) is in Appendix C. $\square$

5 Convergence analysis

We first analyze the case in which the outer proximal map $\text{prox}_{\alpha r}$ is available exactly (e.g. a single proximable term, or two terms with separable joint prox; this covers Example 7.1), then relax it (Proposition 1, used by Example 7.2).

Theorem 2 (Stationarity rate). *Consider (3) with $g_{n,1}, g_{n,2}$ independent of $\mathbf{x}$ and an exactly evaluable $\text{prox}_{\alpha r}$. Suppose Assumptions 1–5 hold, that v is differentiable with L_v -Lipschitz gradient satisfying $\nabla v(\mathbf{x}) = \nabla_{\mathbf{x}} g_s(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$ (Theorem 1), and that $F_\gamma^{\text{inf}} := \inf_{\mathbf{z} \in \mathcal{Z}} F_\gamma(\mathbf{z}) > -\infty$. Run Algorithm 1 with $\alpha = 1/L$, $L := L_f + \gamma(L_g + L_v)$, and an inner solver returning $\hat{\mathbf{y}}^{k+1} \in \mathcal{Y}$ with*

$$g_s(\mathbf{x}^k, \hat{\mathbf{y}}^{k+1}) + g_{n,1}(\hat{\mathbf{y}}^{k+1}) + g_{n,2}(\hat{\mathbf{y}}^{k+1}) - v(\mathbf{x}^k) \leq \epsilon_k, \quad \epsilon_k \leq c/(\alpha \gamma k)^2, \quad (8)$$

for a constant $c > 0$. Let $\mu := 2/\sigma_g$ and choose $\gamma \geq L_c \sqrt{\mu/\delta}$ for some $\delta > 0$. Denote the exact proximal-gradient mapping $\mathbf{G}_\gamma(\mathbf{z}^k) := \alpha^{-1}(\mathbf{z}^k - \bar{\mathbf{z}}^{k+1})$ with $\bar{\mathbf{z}}^{k+1} := \text{prox}_{\alpha r}(\mathbf{z}^k - \alpha \nabla \bar{F}_\gamma(\mathbf{z}^k))$. Then, with c_1, c_2 from Lemma 2:

$$\frac{1}{K} \sum_{k=1}^K \|\mathbf{G}_\gamma(\mathbf{z}^k)\|_2^2 \leq \frac{F_\gamma(\mathbf{z}^1) - F_\gamma^{\text{inf}}}{c_1 \alpha K} + \frac{c_2 c \pi^2 L_x^2 \mu}{6 c_1 \alpha^2 K}. \quad (9)$$

Moreover, if $\mathbf{z}^k \rightarrow \mathbf{z}^*$ then $\mathbf{0} \in \nabla \bar{F}_\gamma(\mathbf{z}^*) + \partial r(\mathbf{z}^*)$, i.e. $\mathbf{z}^*$ is stationary for F_γ ; and, independently of the algorithm, any local minimizer of (3) lying in a neighborhood on which $f(\mathbf{x}, \cdot)$ is L_c -Lipschitz continuous is a local solution of an ϵ -approximate problem of (2) with $\epsilon \leq \delta$ [10, Thm. 2]. The corresponding global statement holds under the stronger requirements of [10, Prop. 3], namely $f(\mathbf{x}, \cdot)$ Lipschitz continuous on all of $\mathcal{Y}$ (e.g., $\mathcal{X} \times \mathcal{Y}$ compact), rather than only on the bounded region of Assumption 1. When an indicator $g_{n,2} = \iota_{\mathcal{O}}$ is present, the correspondence is applied with $\mathcal{O}$ absorbed into the feasible set, so that the penalty $g - v$ is evaluated as the finite-valued $g_s + g_{n,1} - v$ on it, matching the continuity hypothesis of [10].

Proof. The proof is given in Appendix B. $\square$

Corollary 1 (Inner-iteration complexity). *Under Theorem 2, the schedule $\epsilon_k = O(1/(\alpha\gamma k)^2)$ is realized as follows. If the inner solver contracts linearly, $\epsilon_k \leq C_{\text{FB}} \rho^{T_k}$ with $\rho \in (0, 1)$, then $T_k = \Omega(\log(\alpha\gamma k))$ suffices. If it returns the ergodic shadow average of a Davis–Yin solve, with sublinear gap $\epsilon_k \leq C_{\text{DY}}/T_k$ (Lemma 3), then $T_k = \Omega((\alpha\gamma k)^2)$ suffices. The outer rate (9) holds with the same constants in both cases. Here C_{DY} is treated as uniform in k : by Lemma 3 it depends on k only through $D_0^k = \|\mathbf{s}_0^k - \mathbf{s}^*(\mathbf{x}^k)\|_2$ and $G_*^k = \|\nabla_{\mathbf{y}} g_s(\mathbf{x}^k, \mathbf{y}^*(\mathbf{x}^k))\|_2$, and both are bounded — hence $\sup_k C_{\text{DY}}^k < \infty$ — whenever the outer iterates, the warm-started governing variables, and the LL solution path remain in a bounded region, as posited in Assumption 1 (see also Lemma 5).*

Regarding the ergodic shadow average estimate of the Davis–Yin scheme, we note that in practical implementations one may instead return a structure-preserving shadow iterate. Interestingly, as long as the returned point dominates the ergodic average in LL objective, the same bound applies (see Section 6 for additional details on this).

Inexact outer prox. When $\text{prox}_{\alpha r}$ has no closed form (two non-separable proximable terms, as in Example 7.2), we evaluate it by an inner splitting subroutine. Given the inexact-gradient point

$$\mathbf{q}^k := \mathbf{z}^k - \alpha \widehat{\nabla} \bar{F}_\gamma(\mathbf{z}^k \mid \hat{\mathbf{y}}^{k+1}),$$

we call $\mathbf{z}^{k+1}$ a δ_k -approximate outer-prox point if

$$\frac{1}{2} \|\mathbf{z}^{k+1} - \mathbf{q}^k\|_2^2 + \alpha r(\mathbf{z}^{k+1}) \leq \inf_{\mathbf{u}} \left\{ \frac{1}{2} \|\mathbf{u} - \mathbf{q}^k\|_2^2 + \alpha r(\mathbf{u}) \right\} + \delta_k. \quad (10)$$

Incorporating this prox error into Lemma 2 adds a summable perturbation controlled by $\sqrt{\delta_k}$.

Proposition 1 (Rate under inexact outer prox). *Under the hypotheses of Theorem 2, suppose the outer prox is computed according to (10). Assume additionally that the exact and inexact outer-prox points, and the corresponding prox arguments $\mathbf{q}^k$, remain in a bounded set on which $\nabla \bar{F}_\gamma$ is bounded. If $D_\delta := \sum_{k=1}^\infty \sqrt{\delta_k} < \infty$; for example, $\delta_k = O(k^{-2-\varsigma})$ for any $\varsigma > 0$, then there exists a constant $C_{\text{op}} > 0$, independent of K , such that*

$$\frac{1}{K} \sum_{k=1}^K \|\mathbf{G}_\gamma(\mathbf{z}^k)\|_2^2 \leq \frac{F_\gamma(\mathbf{z}^1) - F_\gamma^{\text{inf}}}{c_1 \alpha K} + \frac{c_2 c \pi^2 L_x^2 \mu}{6 c_1 \alpha^2 K} + \frac{C_{\text{op}} D_\delta}{K}. \quad (11)$$

In particular, the $O(1/K)$ stationarity rate is preserved.

The proof is given in Appendix D.

6 Inner solvers

Theorem 2 reduces inner-solver design to the gap bound (8). The lower level contains at most two non-smooth proximable terms. If one of them is a constraint indicator, say $g_{n,2} = \iota_{\mathcal{O}}$, then

its proximal operator is the projection $\text{proj}_{\mathcal{O}}$. Throughout the splitting analysis we take $\mathcal{Y} = \mathbb{R}^m$ without loss of generality: a lower-level constraint set strictly smaller than $\mathbb{R}^m$ is encoded through one of the two non-smooth terms (in the DY regime, through the indicator slot $g_{n,2}$, per the convention below), so that the unconstrained optimality condition $\mathbf{0} \in \nabla_{\mathbf{y}} g_s(\mathbf{x}, \mathbf{y}^*) + \partial g_{n,1}(\mathbf{y}^*) + \partial g_{n,2}(\mathbf{y}^*)$ used in Appendix E is exactly the constrained one, and the requirement $\hat{\mathbf{y}}^{k+1} \in \mathcal{Y}$ of Theorem 2 is automatic. Both examples of Section 7 have $\mathcal{Y} = \mathbb{R}^m$ natively. The relevant question is whether the composite proximal operator of the complete non-smooth part is tractable, i.e. whether

$$\text{prox}_{\beta(g_{n,1}+g_{n,2})} = (\text{closed form / separable}). \quad (12)$$

When (12) holds, forward-backward (FB) splitting applies; otherwise Davis-Yin (DY) three-operator splitting is used to separate the smooth term and the two non-smooth terms. Standard DY therefore covers $g_s + g_{n,1} + \iota_{\mathcal{O}}$, or $g_s + g_{n,1} + g_{n,2}$, but not $g_s + g_{n,1} + g_{n,2} + \iota_{\mathcal{O}}$ unless two non-smooth terms can be grouped into a proximable composite.

When does (12) hold? It holds when the two proximal operators act *separably* on disjoint blocks (so they do not interfere)—e.g. a single ℓ_1 term, or an ℓ_1 term on one block together with an entrywise box/nonnegativity constraint represented as $g_{n,2} = \iota_{\mathcal{O}}$, where the joint prox is soft-thresholding composed with clipping. It *fails* when two non-smooth terms act on the *same* variable with no separable joint prox—e.g. $\lambda_1 \|\beta\|_1 + \lambda_2 \|\mathbf{D}\beta\|_1$ (Example 7.2)—or for coupling constraints (simplex, row-sum) combined with soft-thresholding.

Solver A: projected forward-backward. When (12) holds, the LL is a smooth-plus-(composite-proximable) problem and FB applies:

$$\hat{\mathbf{y}}_{t+1}^k = \text{prox}_{\beta(g_{n,1}+g_{n,2})}(\hat{\mathbf{y}}_t^k - \beta \nabla_{\mathbf{y}} g_s(\mathbf{x}^k, \hat{\mathbf{y}}_t^k)), \quad \beta \in (0, 1/L_g]. \quad (13)$$

When the strong convexity of Assumption 2 resides in the *smooth* part— $g_s(\mathbf{x}, \cdot)$ σ_g -strongly convex, as in both examples of Section 7 and whenever it is induced by the coupling (4)—the FB operator with stepsize $\beta = 1/L_g$ is a contraction in distance with factor $\rho = 1 - \sigma_g/L_g \in [0, 1)$ [17]. More generally, strong convexity of the full sum $g(\mathbf{x}, \cdot)$, and in fact any proximal error-bound / proximal-PL condition, yields linear decay of the LL value gap, $g(\mathbf{x}^k, \hat{\mathbf{y}}_t^k) - v(\mathbf{x}^k) \leq C_{\text{FB}} \rho^t$ with $\rho \in (0, 1)$ [18, 19], which is the property Corollary 1 consumes. Either way, $T_k = \Omega(\log(\alpha\gamma k))$ suffices (Corollary 1).

Solver B: Davis-Yin three-operator splitting. When (12) fails, the three operators $g_s, g_{n,1}, g_{n,2}$ are handled separately by DY [20]. DY does *not* iterate a proximal output directly: it maintains a *governing variable* $\mathbf{s}_t^k$ and produces a *shadow* sequence $\tilde{\mathbf{y}}_t^k = \text{prox}_{\beta g_{n,2}}(\mathbf{s}_t^k)$ that converges to the LL minimizer. Across outer iterations DY is warm-started on the governing variable, $\mathbf{s}_0^k \leftarrow \mathbf{s}_{T_{k-1}}^{k-1}$ (not on the LL estimate $\hat{\mathbf{y}}^k$), while the gradient surrogate (6) uses the shadow output (18). With this initialization, for $t = 0, \dots, T_k - 1$:

$$\tilde{\mathbf{y}}_t^k = \text{prox}_{\beta g_{n,2}}(\mathbf{s}_t^k), \quad (14)$$

$$\check{\mathbf{y}}_t^k = 2\tilde{\mathbf{y}}_t^k - \mathbf{s}_t^k - \beta \nabla_{\mathbf{y}} g_s(\mathbf{x}^k, \tilde{\mathbf{y}}_t^k), \quad (15)$$

$$\mathbf{u}_t^k = \text{prox}_{\beta g_{n,1}}(\check{\mathbf{y}}_t^k), \quad (16)$$

$$\mathbf{s}_{t+1}^k = \mathbf{s}_t^k + (\mathbf{u}_t^k - \tilde{\mathbf{y}}_t^k), \quad (17)$$

with $\beta \in (0, 2/L_g)$ (an over-relaxation λ_t may be inserted in (17); we take $\lambda_t = 1$; see Remark 4 for the stepsize ranges). The carried variable is the governing variable $\mathbf{s}_t^k$ in (17), *not* the proximal output $\mathbf{u}_t^k$. Returning $\mathbf{u}_{T_k-1}^k$ as the next input, as in a plain forward-backward sweep, does *not* solve $g_s + g_{n,1} + g_{n,2}$, since the two proximal steps need not agree at the resulting fixed point. The point returned to the outer loop is built from the shadow $\{\tilde{\mathbf{y}}_t^k\}$. The output carrying the theoretical guarantee is the ergodic average

$$\bar{\mathbf{y}}_{T_k}^k := \frac{1}{T_k} \sum_{t=0}^{T_k-1} \tilde{\mathbf{y}}_t^k, \quad \hat{\mathbf{y}}^{k+1} \in \arg \min_{\mathbf{y} \in \{\tilde{\mathbf{y}}_0^k, \dots, \tilde{\mathbf{y}}_{T_k-1}^k, \bar{\mathbf{y}}_{T_k}^k\}} g(\mathbf{x}^k, \mathbf{y}), \quad (18)$$

where the returned point $\hat{\mathbf{y}}^{k+1}$ is the best, in LL objective, among the shadow iterates and their ergodic average. Since $\bar{\mathbf{y}}_{T_k}^k$ is one of the candidates, $\hat{\mathbf{y}}^{k+1}$ always inherits the ergodic guarantee

of Lemma 3; and when a shadow iterate is at least as good (the typical case near convergence) the returned point is an actual iterate, hence structure-preserving. Selecting among these candidates is cheap because the LL objective $g(\mathbf{x}^k, \cdot)$ is evaluated directly. Crucially, the averaging touches only the *output* sequence: every inner iteration still applies $\text{prox}_{\beta g_{n,1}}$ and $\text{prox}_{\beta g_{n,2}}$ separately and never the joint proximal operator of $g_{n,1} + g_{n,2}$, so Example 7.2 remains a genuine two-prox instance. If $g_{n,2} = \iota_{\mathcal{O}}$, then (14) is the projection $\tilde{\mathbf{y}}_t^k = \text{proj}_{\mathcal{O}}(\mathbf{s}_t^k)$; this is still the same two-non-smooth-term DY template. Recall that when a hard constraint is present in the DY regime, we assign the indicator to the term producing the returned shadow, i.e., we take $g_{n,2} = \iota_{\mathcal{O}}$; otherwise the roles of $g_{n,1}$ and $g_{n,2}$ should be swapped.

Lemma 3 (Davis–Yin inner guarantee). *Fix $\mathbf{x}^k$ and abbreviate $\phi(\mathbf{y}) := g(\mathbf{x}^k, \mathbf{y}) = g_s(\mathbf{x}^k, \mathbf{y}) + g_{n,1}(\mathbf{y}) + g_{n,2}(\mathbf{y})$. Assume $g_s(\mathbf{x}^k, \cdot)$ is convex with L_g -Lipschitz gradient, $g_{n,1}, g_{n,2}$ are proper, closed, and convex, $g_{n,1}$ is L_{g_1} -Lipschitz continuous on an open convex set containing the inner iterates $\{\tilde{\mathbf{y}}_t^k, \mathbf{u}_t^k\}_{t < T_k}$ and $\mathbf{y}^* := \mathbf{y}^*(\mathbf{x}^k)$, and ϕ attains its minimum at $\mathbf{y}^*$ (Assumption 2); $g_{n,2}$ may be an indicator $\iota_{\mathcal{O}}$. Run (14)–(17) with $\beta \in (0, 1/L_g]$ from an arbitrary $\mathbf{s}_0^k$, and let $\mathbf{s}^*$ be a fixed point of the Davis–Yin operator. Then the ergodic shadow average $\bar{\mathbf{y}}_{T_k}^k$ of (18) satisfies*

$$g(\mathbf{x}^k, \bar{\mathbf{y}}_{T_k}^k) - v(\mathbf{x}^k) \leq \frac{C_{\text{DY}}}{T_k}, \quad C_{\text{DY}} := \frac{(D_0 + \beta(L_{g_1} + G_*))^2}{2\beta} + 2(L_{g_1} + G_*)D_0, \quad (19)$$

with $D_0 := \|\mathbf{s}_0^k - \mathbf{s}^*\|_2$ and $G_* := \|\nabla_{\mathbf{y}} g_s(\mathbf{x}^k, \mathbf{y}^*)\|_2$, and hence so does the returned point $\hat{\mathbf{y}}^{k+1}$ of (18), since $\bar{\mathbf{y}}_{T_k}^k$ is among its candidates. If $g_{n,2} = \iota_{\mathcal{O}}$, all shadow iterates and their average are feasible and (19) holds unchanged (see Remark 3).

Proof. The proof is given in Appendix E; it must be self-contained because the published ergodic bounds for Davis–Yin evaluate the two non-smooth terms along two *different* sequences, so they cannot be invoked verbatim for the single-point gap in (19). $\square$

Lemma 3 is the interface between the Davis–Yin inner solver and the outer analysis: the returned point (18) enters Theorem 2 only through the gap condition (8), and the lemma certifies exactly that gap, at the ergodic rate $O(1/T_k)$. Its implications are twofold. First, via Corollary 1 the rate fixes the inner budget of the DY regime, $T_k = \Omega((\alpha\gamma k)^2)$; this sublinear budget is the price of handling two genuinely non-smooth terms by separate proximal steps alone. Second, although the guarantee is carried by an ergodic average, the returned point is the best candidate in LL objective, so the bound is retained while the output is a structure-preserving shadow iterate whenever one attains the minimum.

Remark 3 (Role of the two non-smooth terms; linear regime). (i) *Hard constraints.* The proof of Lemma 3 (Appendix E) uses Lipschitz continuity of $g_{n,1}$ only, and nothing beyond convexity of $g_{n,2}$; this is why a constraint indicator $\iota_{\mathcal{O}}$ must be assigned to $g_{n,2}$, the shadow-producing term, as prescribed above: every shadow iterate $\tilde{\mathbf{y}}_t^k = \text{proj}_{\mathcal{O}}(\mathbf{s}_t^k)$ is then feasible, so is the ergodic average by convexity of $\mathcal{O}$, and the finite-valued gap in (19) is exactly the quantity required in (8). Under the opposite assignment the shadow iterates are in general infeasible, so $g(\mathbf{x}^k, \bar{\mathbf{y}}_{T_k}^k) = +\infty$ and the inner bound would not be certified; the assignment convention is thus necessary, not cosmetic. We stress that the Lipschitz continuity of $g_{n,1}$ is not required for the convergence of the DY iteration itself, which holds for any proper closed convex pair $g_{n,1}, g_{n,2}$ [20]; it is consumed only by the single-point ergodic rate, in the step of Appendix E that merges the two output sequences (passing the value of $g_{n,1}$ from the averaged $\mathbf{u}$ -sequence to the averaged shadow). Moreover, since any finite-valued convex function is Lipschitz continuous on bounded sets, and Fejér monotonicity confines the inner iterates to a ball, the hypothesis holds automatically for norm-type penalties; its only real content is to exclude extended-valued $g_{n,1}$, i.e. to fix which of the two slots an indicator may occupy. (ii) *No best-iterate claim; linear regime.* We claim no non-ergodic or best-iterate rate; these require additional regularity. If one of $g_{n,1}, g_{n,2}$ has a Lipschitz gradient—so that a single genuinely non-smooth term remains—the DY operator contracts, in the regime $(\mu_A + \mu_B + \mu_C)(1/L_A + 1/L_B) > 0$ of [20, 21] (strong convexity of g_s alone is not sufficient), giving $\|\tilde{\mathbf{y}}_t^k - \mathbf{y}^*(\mathbf{x}^k)\|_2 = O(\rho^t)$ with $\rho \in (0, 1)$; local Lipschitz continuity of $g(\mathbf{x}^k, \cdot)$ around $\mathbf{y}^*(\mathbf{x}^k)$ then turns this into the linear gap bound $g(\mathbf{x}^k, \tilde{\mathbf{y}}_{T_k-1}^k) - v(\mathbf{x}^k) = O(\rho^{T_k})$. Both terms are genuinely non-smooth in Example 7.2 ($\ell_1 + \text{TV}$), so there the sublinear ergodic bound is the operative one.

Remark 4 (Stepsize range). With unit relaxation the DY iteration (14)–(17) is Fejér monotone and convergent for any $\beta \in (0, 2/L_g)$ [20]. For $\beta \in (0, 4/L_g)$, convergence holds with relaxation

parameters $\lambda_t \in (0, 2 - \beta L_g/2]$ satisfying $\sum_t \lambda_t(2 - \beta L_g/2 - \lambda_t) = +\infty$ [22, Thm. 3.3]; in particular, a constant relaxation must satisfy $0 < \lambda < 2 - \beta L_g/2$. The restriction $\beta \in (0, 1/L_g]$ in Lemma 3 is what our self-contained ergodic argument requires: the Young step absorbing the gradient cross term in Appendix E consumes exactly the cocoercivity margin. Since β is an algorithmic choice and only the order $O(1/T_k)$ enters Corollary 1, nothing is lost.

An accelerated primal-dual alternative for problems with a non-smooth linear-operator term, with an optimal duality-gap rate, is OS3X [23]; we keep Davis–Yin as the canonical two-prox mechanism and do not pursue OS3X here. Table 1 summarizes the two inner-solver regimes, and Remark 5 clarifies how the same forward–backward / Davis–Yin dichotomy governs the outer proximal step.

Table 1: Inner-solver regimes and the resulting inner-loop accuracy schedule. The Davis–Yin rate is the ergodic $O(1/T_k)$ guarantee of Lemma 3.

Non-smooth structure of the LL	Inner solver	Inner-loop rate
Single ℓ_1 (or ℓ_1 + entrywise box)	Projected FB	Linear ($T_k = \Omega(\log k)$)
Two prox on the same variable (e.g. $\ell_1 + \ \mathbf{D} \cdot\ _1$)	Davis–Yin	Sublinear ($T_k = \Omega(k^2)$)
ℓ_1 + coupling constraint (simplex, row-sum)	Davis–Yin	Sublinear ($T_k = \Omega(k^2)$, constrained gap)

Remark 5 (Outer vs. inner prox). *The same dichotomy governs the outer proximal step. In the FB regime the outer prox_{α_T} is closed-form (Theorem 2 applies directly). In the DY regime the outer prox is itself a two-prox problem, computed by a finite DY subroutine (14)–(17) run until its returned point satisfies the prox-error inequality (10) to accuracy δ_k . As in the inner solve, the output carrying that guarantee is the ergodic average (18), and the resulting error is handled by Proposition 1.*

7 Representative instances

7.1 Example I: robust regression with sparse outlier rejection

Model. We observe $\mathbf{b} = \mathbf{A}\mathbf{w}_{\text{true}} + \mathbf{o}_{\text{true}} + \boldsymbol{\varepsilon} \in \mathbb{R}^p$, with design matrix $\mathbf{A} \in \mathbb{R}^{p \times d}$, dense noise $\boldsymbol{\varepsilon}$, and a *sparse* gross-error vector $\mathbf{o}_{\text{true}}$ flagging corrupted observations [13]. The lower level jointly estimates the regression vector $\mathbf{w}$ and the outliers $\mathbf{o}$, the latter penalized by an ℓ_1 term.

Bilevel formulation. Let $\mathbf{p} \in \mathbb{R}^d$ be an upper-level reference for $\mathbf{w}$, as introduced in Section 2.3, allowing the upper-level problem to influence the selection of lower-level optimal weights while also inducing strong convexity. With LL variable $\mathbf{y} = (\mathbf{w}, \mathbf{o})$ and UL variable $\mathbf{x} = \mathbf{p}$,

$$(\mathbf{w}^*(\mathbf{p}), \mathbf{o}^*(\mathbf{p})) = \arg \min_{\mathbf{w}, \mathbf{o}} \underbrace{\frac{1}{2} \|\mathbf{b} - \mathbf{A}\mathbf{w} - \mathbf{o}\|_2^2 + \frac{\eta}{2} \|\mathbf{w} - \mathbf{p}\|_2^2}_{g_s(\mathbf{p}, \mathbf{w}, \mathbf{o})} + \underbrace{\nu \|\mathbf{o}\|_1}_{g_{n,1}(\mathbf{o})}, \quad (20)$$

and the upper level minimizes a held-out (or downstream) loss

$$\min_{\mathbf{p}} \frac{1}{2} \|\mathbf{b}_{\text{val}} - \mathbf{A}_{\text{val}} \mathbf{w}^*(\mathbf{p})\|_2^2. \quad (21)$$

Mapping to the theory. Here $g_{n,2} \equiv 0$, so (20) is a single-non-smooth-term lower level. The Hessian of g_s in $(\mathbf{w}, \mathbf{o})$ is $\begin{bmatrix} \mathbf{A}^\top \mathbf{A} + \eta \mathbf{I} & \mathbf{A}^\top \\ \mathbf{A} & \mathbf{I} \end{bmatrix}$, which is positive definite (the associated quadratic form equals $\|\mathbf{A}\mathbf{w} + \mathbf{o}\|_2^2 + \eta \|\mathbf{w}\|_2^2$, zero only at $(\mathbf{w}, \mathbf{o}) = \mathbf{0}$); hence g_s is jointly strongly convex with $\sigma_g > 0$, and the data term alone already provides strong convexity in $\mathbf{o}$. The UL dependence sits in g_s , so $\nabla_{\mathbf{p}} g = \eta(\mathbf{p} - \mathbf{w})$ is η -Lipschitz (Assumption 5, $L_x = \eta$), and Theorem 1 gives

$$\nabla v(\mathbf{p}) = \eta(\mathbf{p} - \mathbf{w}^*(\mathbf{p})). \quad (22)$$

The non-smooth term $\nu \|\mathbf{o}\|_1$ has separable prox (soft-thresholding on $\mathbf{o}$, identity on $\mathbf{w}$), so this is the **forward–backward / Case A** regime, with linear inner convergence; the *outer* prox is likewise closed-form, so Theorem 2 applies directly.

Why this example. It exhibits a genuinely non-smooth lower level, strong convexity that is partly native (data term, in $\mathbf{o}$) and partly induced (coupling, in $\mathbf{w}$), a closed-form prox, and a single-loop FB instantiation with the full $O(1/K)$ guarantee. Grouped outliers ($\|\mathbf{o}\|_{2,1}$) or sign/box constraints on $\mathbf{o}$ fit verbatim (the latter pushing toward the DY regime).

7.2 Example II: fused-Lasso / total-variation regression

Model. Many estimation problems require simultaneously *sparse* and *piecewise-constant* solutions, leading to a lower level with two non-smooth terms: an ℓ_1 penalty and a one-dimensional fused-Lasso / total-variation penalty on consecutive differences. Specifically, we consider the penalty $\|\mathbf{D}\beta\|_1$, where $\mathbf{D} \in \mathbb{R}^{(d-1) \times d}$ is the first-difference matrix defined by $[\mathbf{D}\beta]_i = \beta_{i+1} - \beta_i$ for $i = 1, \dots, d-1$ [14, 15].

Bilevel formulation. With LL variable $\mathbf{y} = \beta \in \mathbb{R}^d$, UL reference $\mathbf{x} = \mathbf{p} \in \mathbb{R}^d$,

$$\beta^*(\mathbf{p}) = \arg \min_{\beta} \underbrace{\frac{1}{2}\|\mathbf{b} - \mathbf{A}\beta\|_2^2 + \frac{\eta}{2}\|\beta - \mathbf{p}\|_2^2}_{g_s(\mathbf{p}, \beta)} + \underbrace{\lambda_1\|\beta\|_1}_{g_{n,1}(\beta)} + \underbrace{\lambda_2\|\mathbf{D}\beta\|_1}_{g_{n,2}(\beta)}, \quad (23)$$

and the upper level minimizes a held-out (or downstream) loss

$$\min_{\mathbf{p}} \frac{1}{2}\|\mathbf{b}_{\text{val}} - \mathbf{A}_{\text{val}}\beta^*(\mathbf{p})\|_2^2. \quad (24)$$

Mapping to the theory. The coupling gives strong convexity with modulus $\geq \eta$ regardless of $\mathbf{A}$, so $\sigma_g \geq \eta$, $L_x = \eta$, and $\nabla v(\mathbf{p}) = \eta(\mathbf{p} - \beta^*(\mathbf{p}))$. The two non-smooth terms act on the *same* variable: $\text{prox}_{\beta_{g_{n,1}}}$ is soft-thresholding and $\text{prox}_{\beta_{g_{n,2}}}$ is one-dimensional total-variation / fused-Lasso denoising (an exact, fast operator [16]), but the prox of their *sum* has no closed form. Hence this is a setup where the Davis–Yin-based solver (referred to as solver B in Section 6) must be used. Specifically, the inner solve runs (14)–(17) and returns the output (18). The guarantee is carried by the ergodic shadow average; the returned point need not itself be sparse or exactly piecewise constant, but it satisfies the LL gap bound $g(\mathbf{p}^k, \hat{\mathbf{y}}^{k+1}) - v(\mathbf{p}^k) = O(1/T_k)$ required by the theorem, and in practice it coincides with a structure-preserving shadow iterate whenever one attains the candidate minimum. Because both $\|\beta\|_1$ and $\|\mathbf{D}\beta\|_1$ are genuinely non-smooth, the linear branch of Lemma 3 does not apply, so we rely on the sublinear guarantee and $T_k = \Omega((\alpha\gamma k)^2)$ suffices (Corollary 1). By Remark 5, the outer prox is also a two-prox problem, computed by a finite DY subroutine that returns its ergodic point to accuracy δ_k and is analyzed by Proposition 1.

Why this example. It is a clean instance where forward–backward is provably insufficient and Davis–Yin is the natural mechanism—both for the inner solve and for the outer prox—so it exercises the full machinery, including Proposition 1. It is precisely the regime where differentiating the (doubly non-smooth) solution map, or smoothing both penalties, is most awkward.

7.3 Further instances

The abstract model (2) covers many other bilevel problems with non-smooth LLs; Table 2 lists representative ones, grouped by the source of non-smoothness and the inner-solver regime. These examples are not analyzed in full detail here; they delimit the scope of the framework and clarify which problems fall in the forward–backward or Davis–Yin regimes. The relevant algorithmic distinction is whether the non-smooth part is *single-prox* (FB) or *two-prox* (DY). The listed instances are covered by Theorem 2 under the parameterizations indicated in the last column, namely when the UL variables enter the smooth LL part or the smooth coupling and the non-smooth terms remain independent of $\mathbf{x}$.

The instances in Table 2 are covered in the parameterizations where the upper-level variables enter the smooth part of the LL objective, or the smooth strong-convexity coupling, while the non-smooth terms $g_{n,1}, g_{n,2}$ remain fixed functions of the LL variable. This distinction is important. Variants that learn weights or operators inside a non-smooth term, such as $\lambda\|\mathbf{y}\|_1, \lambda_1\|\mathbf{y}\|_1 + \lambda_2\|\mathbf{D}\mathbf{y}\|_1$ with (λ_1, λ_2) upper-level variables, constraints of the form $\mathbf{y} \in \mathcal{O}(\mathbf{x})$, or SVM model selection in which C scales the hinge loss or the kernel parameters enter it [25], fall outside Theorem 2. In such cases the value function may no longer be differentiable in the simple form $\nabla v(\mathbf{x}) = \nabla_{\mathbf{x}}g_s(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$, and additional smoothing, sensitivity analysis, or block-coordinate arguments are needed.

8 Experiments

We outline the planned evaluation; full protocols and tables are deferred to Appendix ???. The goals are to (i) validate the predicted rates, (ii) contrast the FB and DY regimes, and (iii) compare against hypergradient and smoothing baselines.

Table 2: Instances of the non-smooth constrained BLO template (2).

Instance	LL non-smooth term(s)	Solver	UL variable
Robust regression (Ex. I)	$\nu \ \mathbf{o}\ _1$	FB	reference $\mathbf{p}$
Fused-Lasso / TV (Ex. II)	$\lambda_1 \ \boldsymbol{\beta}\ _1 + \lambda_2 \ \mathbf{D}\boldsymbol{\beta}\ _1$	DY	reference $\mathbf{p}$
Low-rank + sparse decomposition	$\tau \ \cdot\ _* + \rho \ \cdot\ _1$ (sep.)	FB	downstream task
Sparse-and-low-rank (same var.)	$\tau \ \cdot\ _* + \rho \ \cdot\ _1$	DY	downstream task
Dictionary learning [24]	ℓ_1 (elastic net)	FB	dictionary
Constrained sparse estimation	$\ell_1 + \iota_{\mathcal{O}}$ (coupling $\mathcal{O}$)	DY	operator/design in smooth term; fixed $\mathcal{O}$

Experiment 1: convergence diagnostics. On controlled synthetic instances of (20) and (23) we will track the proximal-gradient mapping $\|\mathbf{G}_\gamma(\mathbf{z}^k)\|_2^2$, the LL gap $g(\mathbf{x}^k, \hat{\mathbf{y}}^{k+1}) - v(\mathbf{x}^k)$, and $F_\gamma(\mathbf{z}^k)$, verifying the $O(1/K)$ outer decay and the effect of the penalty γ , the coupling η , and the inner schedule T_k (fixed vs. increasing), as well as exact vs. inexact LL solves. For Example II we additionally vary the outer-prox accuracy δ_k (Proposition 1).

Experiment 2: robust regression with sparse outliers (FB). We will generate contaminated linear models with varying outlier fraction, noise level, and sample size, and compare the proposed penalty method against a sequential robust-regression baseline, an unrolled proximal-gradient baseline, and (where available) an implicit-differentiation baseline. Metrics: validation MSE, outlier-support recovery, convergence curves, and runtime.

Experiment 3: fused-Lasso / TV regression (DY). On piecewise-constant / piecewise-sparse signals we will compare the proposed Davis–Yin instantiation against a *smoothed-TV* surrogate (the standard route enabling differentiation), an unrolled baseline, and a sequential baseline. Metrics: reconstruction error, change-point recovery, the stationarity measure, and sensitivity to T_k and δ_k . This experiment isolates the benefit of handling the two non-smooth terms exactly.

9 Concluding remarks

We presented a single-loop, hypergradient-free penalty method for bilevel problems with strongly convex but non-smooth LLs comprising up to two convex proximable terms, one of which may encode a convex constraint through an indicator function. The analysis rests on a value-function regularity result valid for one or two non-smooth proximable terms, an inexact proximal-gradient scheme with an $O(1/K)$ stationarity rate (including the inexact-outer-prox case), and a clean forward–backward/Davis–Yin dichotomy, instantiated on robust regression with sparse outliers and on fused-Lasso/TV regression.

Limitations. The method targets stationary points of the penalized problem; strong convexity of the lower level is required (natively or induced); the Davis–Yin regime incurs a sublinear inner rate and an inexact outer prox; and LLs whose non-smooth terms depend on the UL variable fall outside the jointly convex analysis (last paragraph in Section 7.3). Ways and implications of relaxing the strong convexity assumption are briefly discussed in Appendix F.

Future work. Promising directions include identifying structural conditions under which Davis–Yin inner solves admit faster rates, developing stochastic and mini-batch variants, extending the framework to lower-level constraints that couple LL and UL variables more tightly, and designing specialized schemes for the application classes listed in Table 2.

References

- [1] Jonathan F. Bard. *Practical Bilevel Optimization: Algorithms and Applications*, volume 30 of *Nonconvex Optimization and Its Applications*. Springer, New York, 2013. doi: 10.1007/978-1-4757-2836-1. URL <https://doi.org/10.1007/978-1-4757-2836-1>. Original copyright 1998; eBook published 2013.

- [2] Ankur Sinha, Pekka Malo, and Kalyanmoy Deb. A review on bilevel optimization: From classical to evolutionary approaches and applications. *IEEE Transactions on Evolutionary Computation*, 22(2): 276–295, 2018. doi: 10.1109/TEVC.2017.2712906. URL <https://doi.org/10.1109/TEVC.2017.2712906>.
- [3] 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*, pages 1568–1577. PMLR, July 2018. URL <https://proceedings.mlr.press/v80/franceschi18a.html>.
- [4] Yihua Zhang, Prashant Khanduri, Ioannis Tsaknakis, Yuguang Yao, Mingyi Hong, and Sijia Liu. An introduction to bilevel optimization: Foundations and applications in signal processing and machine learning. *IEEE Signal Processing Magazine*, 41(1):38–59, 2024. doi: 10.1109/MSP.2024.3358284. URL <https://doi.org/10.1109/MSP.2024.3358284>.
- [5] Risheng Liu, Jiaxin Gao, Jin Zhang, Deyu Meng, and Zhouchen Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10045–10067, 2022. doi: 10.1109/TPAMI.2021.3132674. URL <https://doi.org/10.1109/TPAMI.2021.3132674>.
- [6] Quentin Bertrand, Quentin Klopfenstein, Mathurin Massias, Mathieu Blondel, Samuel Vaiter, Alexandre Gramfort, and Joseph Salmon. Implicit differentiation for fast hyperparameter selection in non-smooth convex learning. *Journal of Machine Learning Research*, 23(149):1–43, 2022. URL <https://jmlr.org/papers/v23/21-0486.html>.
- [7] Jérôme Bolte, Tam Le, Edouard Pauwels, and Antonio Silvetti-Falls. Nonsmooth implicit differentiation for machine-learning and optimization. In *Advances in Neural Information Processing Systems*, volume 34, pages 13537–13549. Curran Associates, Inc., 2021. URL <https://proceedings.neurips.cc/paper/2021/hash/70afbf2259b4449d8ae1429e054df1b1-Abstract.html>.
- [8] Peter Ochs, René Ranftl, Thomas Brox, and Thomas Pock. Bilevel optimization with nonsmooth lower level problems. In *Scale Space and Variational Methods in Computer Vision*, volume 9087 of *Lecture Notes in Computer Science*, pages 654–665. Springer, 2015. doi: 10.1007/978-3-319-18461-6_52. URL https://doi.org/10.1007/978-3-319-18461-6_52.
- [9] Juan Carlos De los Reyes and David Villacís. Bilevel optimization methods in imaging. In Ke Chen, Carola-Bibiane Schönlieb, Xue-Cheng Tai, and Laurent Younes, editors, *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging*, pages 909–941. Springer, Cham, 2023. doi: 10.1007/978-3-030-98661-2_66. URL https://doi.org/10.1007/978-3-030-98661-2_66.
- [10] 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*, pages 30992–31015. PMLR, July 2023. URL <https://proceedings.mlr.press/v202/shen23c.html>.
- [11] Zhaosong Lu and Sanyou Mei. First-order penalty methods for bilevel optimization. *SIAM Journal on Optimization*, 34(2):1937–1969, 2024. doi: 10.1137/23M1566753. URL <https://doi.org/10.1137/23M1566753>.
- [12] Karl Kunisch and Thomas Pock. A bilevel optimization approach for parameter learning in variational models. *SIAM Journal on Imaging Sciences*, 6(2):938–983, 2013. doi: 10.1137/120882706. URL <https://doi.org/10.1137/120882706>.
- [13] Gonzalo Mateos and Georgios B. Giannakis. Robust nonparametric regression via sparsity control with application to load curve data cleansing. *IEEE Transactions on Signal Processing*, 60(4):1571–1584, 2012. doi: 10.1109/TSP.2011.2181837. URL <https://doi.org/10.1109/TSP.2011.2181837>.
- [14] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005. doi: 10.1111/j.1467-9868.2005.00490.x. URL <https://doi.org/10.1111/j.1467-9868.2005.00490.x>.
- [15] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1–4):259–268, 1992. doi: 10.1016/0167-2789(92)90242-F. URL [https://doi.org/10.1016/0167-2789\(92\)90242-F](https://doi.org/10.1016/0167-2789(92)90242-F).
- [16] Laurent Condat. A direct algorithm for 1-d total variation denoising. *IEEE Signal Processing Letters*, 20(11):1054–1057, 2013. doi: 10.1109/LSP.2013.2278339. URL <https://doi.org/10.1109/LSP.2013.2278339>.

- [17] Amir Beck. *First-Order Methods in Optimization*, volume 25 of *MOS-SIAM Series on Optimization*. SIAM, 2017. doi: 10.1137/1.9781611974997. URL <https://doi.org/10.1137/1.9781611974997>.
- [18] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, volume 9851 of *Lecture Notes in Computer Science*, pages 795–811. Springer, Cham, 2016. doi: 10.1007/978-3-319-46128-1_50. URL https://doi.org/10.1007/978-3-319-46128-1_50.
- [19] Dmitriy Drusvyatskiy and Adrian S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018. doi: 10.1287/moor.2017.0889. URL <https://doi.org/10.1287/moor.2017.0889>.
- [20] Damek Davis and Wotao Yin. A three-operator splitting scheme and its optimization applications. *Set-Valued and Variational Analysis*, 25(4):829–858, 2017. doi: 10.1007/s11228-017-0421-z. URL <https://doi.org/10.1007/s11228-017-0421-z>.
- [21] Fabian Pedregosa and Gauthier Gidel. Adaptive three operator splitting. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4085–4094. PMLR, July 2018. URL <https://proceedings.mlr.press/v80/pedregosa18a.html>.
- [22] Francisco J. Aragón-Artacho and David Torregrosa-Belén. A direct proof of convergence of Davis–Yin splitting algorithm allowing larger stepsizes. *Set-Valued and Variational Analysis*, 30(3):1011–1029, 2022. doi: 10.1007/s11228-022-00631-6. URL <https://doi.org/10.1007/s11228-022-00631-6>.
- [23] Seyoon Ko and Joong-Ho Won. Optimal minimization of the sum of three convex functions with a linear operator. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1185–1194. PMLR, April 2019. URL <https://proceedings.mlr.press/v89/ko19a.html>.
- [24] Julien Mairal, Francis Bach, and Jean Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804, 2012. doi: 10.1109/TPAMI.2011.156. URL <https://doi.org/10.1109/TPAMI.2011.156>.
- [25] Gautam Kunapuli, Kristin P. Bennett, Jing Hu, and Jong-Shi Pang. Bilevel model selection for support vector machines. In *Data Mining and Mathematical Programming*, volume 45 of *CRM Proceedings and Lecture Notes*, pages 129–158. American Mathematical Society, Providence, RI, 2008. doi: 10.1090/crm/045/07. URL <https://doi.org/10.1090/crm/045/07>.
- [26] Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32, 2019.
- [27] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 2 edition, 1999.
- [28] J. Frédéric Bonnans and Alexander Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research. Springer, New York, 2000. doi: 10.1007/978-1-4612-1394-9. URL <https://doi.org/10.1007/978-1-4612-1394-9>.
- [29] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: A general approach. *Annals of Operations Research*, 46(1):157–178, 1993. doi: 10.1007/BF02096261. URL <https://doi.org/10.1007/BF02096261>.

A Proof of Theorem 1

Steps 1–3 establish the Lipschitz continuity of the solution mapping; Step 4 establishes differentiability and Lipschitz smoothness of the value function.

Step 1: setup. Throughout the proof we work with the convex LL objective

$$g(\mathbf{x}, \cdot) = g_s(\mathbf{x}, \cdot) + g_{n,1}(\cdot) + g_{n,2}(\cdot),$$

where $g_{n,1}$ and $g_{n,2}$ are proper, closed, convex, and independent of $\mathbf{x}$. If the LL contains a convex constraint, it is represented by setting one of these two terms equal to an indicator function. By Assumption 2, for every fixed $\mathbf{x} \in \mathcal{X}$ the minimizer is unique; denote it by $\mathbf{y}^*(\mathbf{x})$.

Step 2: quadratic growth of the LL objective. Strong convexity in the LL variable gives, for every $\mathbf{y} \in \mathcal{Y}$,

$$g(\mathbf{x}, \mathbf{y}) - g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) \geq \frac{\sigma_g}{2} \|\mathbf{y} - \mathbf{y}^*(\mathbf{x})\|_2^2. \quad (25)$$

This is the same quadratic-growth estimate as Lemma 1, specialized to the unique-minimizer case.

Step 3: Lipschitz continuity of the solution mapping. Fix $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and let $\mathbf{y}_i^* := \mathbf{y}^*(\mathbf{x}_i)$ for $i = 1, 2$. Applying (25) with $\mathbf{x} = \mathbf{x}_2$ and $\mathbf{y} = \mathbf{y}_1^*$ yields

$$g(\mathbf{x}_2, \mathbf{y}_1^*) - g(\mathbf{x}_2, \mathbf{y}_2^*) \geq \frac{\sigma_g}{2} \|\mathbf{y}_1^* - \mathbf{y}_2^*\|_2^2. \quad (26)$$

Since $\mathbf{y}_1^*$ minimizes $g(\mathbf{x}_1, \cdot)$ over $\mathcal{Y}$ and $\mathbf{y}_2^* \in \mathcal{Y}$,

$$g(\mathbf{x}_1, \mathbf{y}_1^*) - g(\mathbf{x}_1, \mathbf{y}_2^*) \leq 0.$$

Adding this inequality to (26) gives

$$[g(\mathbf{x}_2, \mathbf{y}_1^*) - g(\mathbf{x}_1, \mathbf{y}_1^*)] - [g(\mathbf{x}_2, \mathbf{y}_2^*) - g(\mathbf{x}_1, \mathbf{y}_2^*)] \geq \frac{\sigma_g}{2} \|\mathbf{y}_1^* - \mathbf{y}_2^*\|_2^2. \quad (27)$$

Because $g_{n,1}$ and $g_{n,2}$ are independent of $\mathbf{x}$, only g_s contributes to the two bracketed differences. By the mean-value theorem, there exist $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in [\mathbf{x}_1, \mathbf{x}_2] \subseteq \mathcal{X}$ (the segment lies in $\mathcal{X}$ by convexity, Assumption 4) such that

$$g(\mathbf{x}_2, \mathbf{y}_i^*) - g(\mathbf{x}_1, \mathbf{y}_i^*) = \langle \nabla_{\mathbf{x}} g_s(\boldsymbol{\xi}_i, \mathbf{y}_i^*), \mathbf{x}_2 - \mathbf{x}_1 \rangle, \quad i = 1, 2.$$

Substitution into (27) gives

$$\langle \nabla_{\mathbf{x}} g_s(\boldsymbol{\xi}_1, \mathbf{y}_1^*) - \nabla_{\mathbf{x}} g_s(\boldsymbol{\xi}_2, \mathbf{y}_2^*), \mathbf{x}_2 - \mathbf{x}_1 \rangle \geq \frac{\sigma_g}{2} \|\mathbf{y}_1^* - \mathbf{y}_2^*\|_2^2. \quad (28)$$

By Cauchy–Schwarz and then Assumption 5,

$$\begin{aligned} \langle \nabla_{\mathbf{x}} g_s(\boldsymbol{\xi}_1, \mathbf{y}_1^*) - \nabla_{\mathbf{x}} g_s(\boldsymbol{\xi}_2, \mathbf{y}_2^*), \mathbf{x}_2 - \mathbf{x}_1 \rangle &\leq \|\nabla_{\mathbf{x}} g_s(\boldsymbol{\xi}_1, \mathbf{y}_1^*) - \nabla_{\mathbf{x}} g_s(\boldsymbol{\xi}_2, \mathbf{y}_2^*)\|_2 \|\mathbf{x}_2 - \mathbf{x}_1\|_2 \\ &\leq L_x (\|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|_2 + \|\mathbf{y}_1^* - \mathbf{y}_2^*\|_2) \|\mathbf{x}_2 - \mathbf{x}_1\|_2 \\ &\leq L_x (\|\mathbf{x}_1 - \mathbf{x}_2\|_2 + \|\mathbf{y}_1^* - \mathbf{y}_2^*\|_2) \|\mathbf{x}_2 - \mathbf{x}_1\|_2, \end{aligned} \quad (29)$$

where the last inequality uses $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2 \in [\mathbf{x}_1, \mathbf{x}_2]$. Combining (28) and (29), and setting $a := \|\mathbf{y}_1^* - \mathbf{y}_2^*\|_2$ and $b := \|\mathbf{x}_1 - \mathbf{x}_2\|_2$, yields

$$\frac{\sigma_g}{2} a^2 \leq L_x b^2 + L_x a b.$$

Solving this quadratic inequality in a and discarding the negative root gives

$$\|\mathbf{y}_1^* - \mathbf{y}_2^*\|_2 \leq L_S \|\mathbf{x}_1 - \mathbf{x}_2\|_2, \quad L_S := \frac{L_x + \sqrt{L_x^2 + 2\sigma_g L_x}}{\sigma_g}. \quad (30)$$

This proves part (a).

Step 4: smoothness of the value function. For any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, using Assumption 5 and (30),

$$\begin{aligned} \|\nabla_{\mathbf{x}} g_s(\mathbf{x}_1, \mathbf{y}_1^*) - \nabla_{\mathbf{x}} g_s(\mathbf{x}_2, \mathbf{y}_2^*)\|_2 &\leq L_x (\|\mathbf{x}_1 - \mathbf{x}_2\|_2 + \|\mathbf{y}_1^* - \mathbf{y}_2^*\|_2) \\ &\leq L_x (1 + L_S) \|\mathbf{x}_1 - \mathbf{x}_2\|_2. \end{aligned} \quad (31)$$

Similar to [26, Lemma A.5], by Danskin-type sensitivity for the unique-minimizer case [27, 28], the directional derivative of v in any direction $\mathbf{d}$ satisfies

$$v'(\mathbf{x}; \mathbf{d}) = \langle \nabla_{\mathbf{x}} g_s(\mathbf{x}, \mathbf{y}^*(\mathbf{x})), \mathbf{d} \rangle.$$

Therefore v is differentiable and

$$\nabla v(\mathbf{x}) = \nabla_{\mathbf{x}} g_s(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) = \nabla_{\mathbf{x}} g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})),$$

where the last equality uses the $\mathbf{x}$ -independence of the non-smooth terms. Inequality (31) proves that ∇v is L_v -Lipschitz with $L_v = L_x(1 + L_S)$. $\square$

B Proof of Theorem 2

Write $\mathbf{z} = (\mathbf{x}, \mathbf{y})$. By the assumed smoothness of v , $\nabla \bar{F}_\gamma$ is L -Lipschitz with $L = L_f + \gamma(L_g + L_v)$.

Step 1: bounding the gradient approximation error. The only inexact component of $\widehat{\nabla} \bar{F}_\gamma$ is the surrogate

$$\nabla v(\mathbf{x}^k) = \nabla_{\mathbf{x}} g_s(\mathbf{x}^k, \mathbf{y}^*(\mathbf{x}^k)) \approx \nabla_{\mathbf{x}} g_s(\mathbf{x}^k, \hat{\mathbf{y}}^{k+1}).$$

The $\mathbf{y}$ -block of $\widehat{\nabla} \bar{F}_\gamma$ coincides with the $\mathbf{y}$ -block of $\nabla \bar{F}_\gamma$. Therefore, by Assumption 5 and Lemma 1,

$$\begin{aligned} \|\nabla \bar{F}_\gamma(\mathbf{z}^k) - \widehat{\nabla} \bar{F}_\gamma(\mathbf{z}^k | \hat{\mathbf{y}}^{k+1})\|_2 &\leq \gamma L_x \|\hat{\mathbf{y}}^{k+1} - \mathbf{y}^*(\mathbf{x}^k)\|_2 \\ &\leq \gamma L_x \sqrt{\mu \epsilon_k}. \end{aligned} \quad (32)$$

Step 2: applying the inexact proximal-gradient descent lemma. Identify $\psi = \bar{F}_\gamma$ and

$$\mathbf{e}^k := \widehat{\nabla} \bar{F}_\gamma(\mathbf{z}^k | \hat{\mathbf{y}}^{k+1}) - \nabla \bar{F}_\gamma(\mathbf{z}^k).$$

Lemma 2, with $\alpha = 1/L$, gives

$$F_\gamma(\mathbf{z}^{k+1}) \leq F_\gamma(\mathbf{z}^k) - c_1 \alpha \|\mathbf{G}_\gamma(\mathbf{z}^k)\|_2^2 + c_2 \alpha \|\mathbf{e}^k\|_2^2. \quad (33)$$

Using (32),

$$F_\gamma(\mathbf{z}^{k+1}) \leq F_\gamma(\mathbf{z}^k) - c_1 \alpha \|\mathbf{G}_\gamma(\mathbf{z}^k)\|_2^2 + c_2 \alpha \gamma^2 L_x^2 \mu \epsilon_k. \quad (34)$$

Step 3: telescoping and bounding the error sum. Summing (34) for $k = 1, \dots, K$ and using $F_\gamma(\mathbf{z}^{K+1}) \geq F_\gamma^{\inf}$,

$$c_1 \alpha \sum_{k=1}^K \|\mathbf{G}_\gamma(\mathbf{z}^k)\|_2^2 \leq F_\gamma(\mathbf{z}^1) - F_\gamma^{\inf} + c_2 \alpha \gamma^2 L_x^2 \mu \sum_{k=1}^K \epsilon_k. \quad (35)$$

By the schedule $\epsilon_k \leq c/(\alpha \gamma k)^2$,

$$c_2 \alpha \gamma^2 L_x^2 \mu \sum_{k=1}^K \epsilon_k \leq \frac{c_2 c L_x^2 \mu}{\alpha} \sum_{k=1}^K \frac{1}{k^2} \leq \frac{c_2 c \pi^2 L_x^2 \mu}{6 \alpha}.$$

Substituting into (35) and dividing by $c_1 \alpha K$ proves (9).

Step 4: stationarity. The rate bound implies $\liminf_{k \rightarrow \infty} \|\mathbf{G}_\gamma(\mathbf{z}^k)\|_2 = 0$, so there exists a subsequence $\{k_j\}$ such that $\mathbf{G}_\gamma(\mathbf{z}^{k_j}) \rightarrow \mathbf{0}$. The exact proximal-gradient mapping is continuous on $\mathcal{Z}$: $\nabla \bar{F}_\gamma$ is continuous by assumption and $\text{prox}_{\alpha r}$ is non-expansive since r is proper, closed, and convex. Hence, if $\mathbf{z}^k \rightarrow \mathbf{z}^*$, then

$$\mathbf{G}_\gamma(\mathbf{z}^*) = \lim_{j \rightarrow \infty} \mathbf{G}_\gamma(\mathbf{z}^{k_j}) = \mathbf{0}.$$

Equivalently,

$$\mathbf{z}^* = \text{prox}_{\alpha r}(\mathbf{z}^* - \alpha \nabla \bar{F}_\gamma(\mathbf{z}^*)).$$

The first-order optimality condition of the proximal map yields

$$\mathbf{0} \in \nabla \bar{F}_\gamma(\mathbf{z}^*) + \partial r(\mathbf{z}^*),$$

which is the asserted stationarity condition. The approximate-bilevel statement follows from [10, Thm. 2, Prop. 3], as qualified in the theorem statement (local version under Assumption 1; global version under Lipschitz continuity of $f(\mathbf{x}, \cdot)$ on all of $\mathcal{Y}$); it depends on the penalty formulation and quadratic growth, not on the algorithmic errors. $\square$

C Proof of Lemma 2

We establish the inexact proximal-gradient descent inequality (7). Let $\mathbf{g} := \nabla\psi(\mathbf{z})$ denote the exact gradient and $\tilde{\mathbf{g}} := \mathbf{g} + \mathbf{e}$ the inexact gradient, so that

$$\bar{\mathbf{z}}^+ = \text{prox}_{\alpha r}(\mathbf{z} - \alpha \mathbf{g}), \quad \mathbf{z}^+ = \text{prox}_{\alpha r}(\mathbf{z} - \alpha \tilde{\mathbf{g}}).$$

Step 1: descent on the smooth part. By L -smoothness of ψ and $\alpha L \leq 1$,

$$\psi(\mathbf{z}^+) \leq \psi(\mathbf{z}) + \langle \mathbf{g}, \mathbf{z}^+ - \mathbf{z} \rangle + \frac{1}{2\alpha} \|\mathbf{z}^+ - \mathbf{z}\|_2^2. \quad (36)$$

Step 2: proximal optimality. Since $\mathbf{z}^+$ minimizes the 1-strongly convex function $\frac{1}{2}\|\mathbf{w} - (\mathbf{z} - \alpha \tilde{\mathbf{g}})\|_2^2 + \alpha r(\mathbf{w})$, Fermat's rule gives $\mathbf{0} \in \mathbf{z}^+ - (\mathbf{z} - \alpha \tilde{\mathbf{g}}) + \alpha \partial r(\mathbf{z}^+)$, i.e. $\frac{1}{\alpha}(\mathbf{z} - \alpha \tilde{\mathbf{g}} - \mathbf{z}^+) \in \partial r(\mathbf{z}^+)$. The subgradient inequality at $\mathbf{z}^+$ with this particular subgradient reads $r(\mathbf{u}) \geq r(\mathbf{z}^+) + \frac{1}{\alpha} \langle \mathbf{z} - \alpha \tilde{\mathbf{g}} - \mathbf{z}^+, \mathbf{u} - \mathbf{z}^+ \rangle$ for every $\mathbf{u} \in \text{dom } r$. Rearranging yields (37)

$$r(\mathbf{z}^+) - r(\mathbf{u}) \leq \frac{1}{\alpha} \langle \mathbf{z} - \alpha \tilde{\mathbf{g}} - \mathbf{z}^+, \mathbf{z}^+ - \mathbf{u} \rangle. \quad (37)$$

Setting $\mathbf{u} = \mathbf{z}$ in (37),

$$r(\mathbf{z}^+) - r(\mathbf{z}) \leq -\langle \tilde{\mathbf{g}}, \mathbf{z}^+ - \mathbf{z} \rangle - \frac{1}{\alpha} \|\mathbf{z}^+ - \mathbf{z}\|_2^2. \quad (38)$$

Step 3: combine the two bounds. Adding (36) and (38),

$$\begin{aligned} \Phi(\mathbf{z}^+) - \Phi(\mathbf{z}) &\leq \langle \mathbf{g} - \tilde{\mathbf{g}}, \mathbf{z}^+ - \mathbf{z} \rangle - \frac{1}{2\alpha} \|\mathbf{z}^+ - \mathbf{z}\|_2^2 \\ &= -\langle \mathbf{e}, \mathbf{z}^+ - \mathbf{z} \rangle - \frac{1}{2\alpha} \|\mathbf{z}^+ - \mathbf{z}\|_2^2. \end{aligned} \quad (39)$$

Step 4: relate the inexact and exact proximal steps. By non-expansiveness of $\text{prox}_{\alpha r}$ applied to the inputs $\mathbf{z} - \alpha \mathbf{g}$ and $\mathbf{z} - \alpha \tilde{\mathbf{g}}$,

$$\|\mathbf{z}^+ - \bar{\mathbf{z}}^+\|_2 \leq \alpha \|\mathbf{e}\|_2. \quad (40)$$

Since $\alpha G_\alpha(\mathbf{z}) = \mathbf{z} - \bar{\mathbf{z}}^+$, the triangle inequality gives

$$\|\mathbf{z}^+ - \mathbf{z}\|_2^2 \geq \frac{1}{2} \|\bar{\mathbf{z}}^+ - \mathbf{z}\|_2^2 - \|\mathbf{z}^+ - \bar{\mathbf{z}}^+\|_2^2 \geq \frac{\alpha^2}{2} \|G_\alpha(\mathbf{z})\|_2^2 - \alpha^2 \|\mathbf{e}\|_2^2. \quad (41)$$

Step 5: absorb the cross term. Young's inequality with parameter $1/(8\alpha)$ gives

$$|\langle \mathbf{e}, \mathbf{z}^+ - \mathbf{z} \rangle| \leq \frac{1}{8\alpha} \|\mathbf{z}^+ - \mathbf{z}\|_2^2 + 2\alpha \|\mathbf{e}\|_2^2.$$

Combining this bound with (39) and (41),

$$\begin{aligned} \Phi(\mathbf{z}^+) - \Phi(\mathbf{z}) &\leq -\frac{3}{8\alpha} \|\mathbf{z}^+ - \mathbf{z}\|_2^2 + 2\alpha \|\mathbf{e}\|_2^2 \\ &\leq -\frac{3}{8\alpha} \left(\frac{\alpha^2}{2} \|G_\alpha(\mathbf{z})\|_2^2 - \alpha^2 \|\mathbf{e}\|_2^2 \right) + 2\alpha \|\mathbf{e}\|_2^2 \\ &= -\frac{3\alpha}{16} \|G_\alpha(\mathbf{z})\|_2^2 + \frac{19\alpha}{8} \|\mathbf{e}\|_2^2. \end{aligned}$$

Thus (7) holds with $c_1 = 3/16$ and $c_2 = 19/8$. $\square$

D Proof of Proposition 1

Let

$$\tilde{\mathbf{z}}^{k+1} := \text{prox}_{\alpha r}(\mathbf{q}^k), \quad \mathbf{q}^k := \mathbf{z}^k - \alpha \widehat{\nabla} \bar{F}_\gamma(\mathbf{z}^k \mid \hat{\mathbf{y}}^{k+1}),$$

be the exact outer-prox point associated with the inexact gradient, and let $\mathbf{z}^{k+1}$ satisfy the δ_k -approximate prox condition (10). Note that (10) forces $\frac{1}{2}\|\mathbf{z}^{k+1} - \mathbf{q}^k\|_2^2 + \alpha r(\mathbf{z}^{k+1}) < \infty$, hence $r(\mathbf{z}^{k+1}) < \infty$ and $\mathbf{z}^{k+1} \in \text{dom } r$; the exact prox point $\tilde{\mathbf{z}}^{k+1}$ lies in $\text{dom } r$ as well, so all quantities below are finite and the indicator part of r is respected by both points. Define the proximal subproblem

$$Q_k(\mathbf{u}) := \frac{1}{2}\|\mathbf{u} - \mathbf{q}^k\|_2^2 + \alpha r(\mathbf{u}).$$

Since Q_k is 1-strongly convex and $\tilde{\mathbf{z}}^{k+1}$ is its exact minimizer,

$$Q_k(\mathbf{z}^{k+1}) \geq Q_k(\tilde{\mathbf{z}}^{k+1}) + \frac{1}{2}\|\mathbf{z}^{k+1} - \tilde{\mathbf{z}}^{k+1}\|_2^2.$$

Combining this with $Q_k(\mathbf{z}^{k+1}) \leq Q_k(\tilde{\mathbf{z}}^{k+1}) + \delta_k$ gives

$$\|\mathbf{z}^{k+1} - \tilde{\mathbf{z}}^{k+1}\|_2 \leq \sqrt{2\delta_k}. \quad (42)$$

Apply Lemma 2 to the exact inexact-gradient step $\tilde{\mathbf{z}}^{k+1} = \text{prox}_{\alpha r}(\mathbf{q}^k)$. With

$$\mathbf{e}^k := \widehat{\nabla} \bar{F}_\gamma(\mathbf{z}^k \mid \hat{\mathbf{y}}^{k+1}) - \nabla \bar{F}_\gamma(\mathbf{z}^k),$$

we obtain

$$F_\gamma(\tilde{\mathbf{z}}^{k+1}) \leq F_\gamma(\mathbf{z}^k) - c_1 \alpha \|\mathbf{G}_\gamma(\mathbf{z}^k)\|_2^2 + c_2 \alpha \|\mathbf{e}^k\|_2^2. \quad (43)$$

It remains to pass from $\tilde{\mathbf{z}}^{k+1}$ to the computed point $\mathbf{z}^{k+1}$. The boundedness assumption implies that there are constants $M_\gamma, R_\gamma < \infty$ such that, for all relevant k ,

$$\|\nabla \bar{F}_\gamma(\mathbf{u})\|_2 \leq M_\gamma \quad \text{on the segment between } \tilde{\mathbf{z}}^{k+1} \text{ and } \mathbf{z}^{k+1}, \quad \|\tilde{\mathbf{z}}^{k+1} + \mathbf{z}^{k+1} - 2\mathbf{q}^k\|_2 \leq R_\gamma.$$

The approximate-optimality inequality $Q_k(\mathbf{z}^{k+1}) \leq Q_k(\tilde{\mathbf{z}}^{k+1}) + \delta_k$ yields

$$\begin{aligned} \alpha(r(\mathbf{z}^{k+1}) - r(\tilde{\mathbf{z}}^{k+1})) &\leq \delta_k + \frac{1}{2}(\|\tilde{\mathbf{z}}^{k+1} - \mathbf{q}^k\|_2^2 - \|\mathbf{z}^{k+1} - \mathbf{q}^k\|_2^2) \\ &= \delta_k + \frac{1}{2}\langle \tilde{\mathbf{z}}^{k+1} + \mathbf{z}^{k+1} - 2\mathbf{q}^k, \tilde{\mathbf{z}}^{k+1} - \mathbf{z}^{k+1} \rangle \\ &\leq \delta_k + \frac{R_\gamma}{2}\|\mathbf{z}^{k+1} - \tilde{\mathbf{z}}^{k+1}\|_2. \end{aligned} \quad (44)$$

Moreover, by the mean-value theorem and the boundedness of $\nabla \bar{F}_\gamma$,

$$\bar{F}_\gamma(\mathbf{z}^{k+1}) - \bar{F}_\gamma(\tilde{\mathbf{z}}^{k+1}) \leq M_\gamma \|\mathbf{z}^{k+1} - \tilde{\mathbf{z}}^{k+1}\|_2. \quad (45)$$

Combining (44) and (45) gives

$$\begin{aligned} F_\gamma(\mathbf{z}^{k+1}) - F_\gamma(\tilde{\mathbf{z}}^{k+1}) &= \bar{F}_\gamma(\mathbf{z}^{k+1}) - \bar{F}_\gamma(\tilde{\mathbf{z}}^{k+1}) + r(\mathbf{z}^{k+1}) - r(\tilde{\mathbf{z}}^{k+1}) \\ &\leq \frac{\delta_k}{\alpha} + \left(M_\gamma + \frac{R_\gamma}{2\alpha}\right) \|\mathbf{z}^{k+1} - \tilde{\mathbf{z}}^{k+1}\|_2. \end{aligned}$$

Using (42) in the last bound yields

$$F_\gamma(\mathbf{z}^{k+1}) \leq F_\gamma(\tilde{\mathbf{z}}^{k+1}) + \frac{\delta_k}{\alpha} + \sqrt{2} \left(M_\gamma + \frac{R_\gamma}{2\alpha}\right) \sqrt{\delta_k}.$$

Absorbing the finite number of terms for which $\delta_k > 1$ into the constant gives a constant $C_{\text{op}} > 0$ such that

$$F_\gamma(\mathbf{z}^{k+1}) \leq F_\gamma(\tilde{\mathbf{z}}^{k+1}) + C_{\text{op}} \sqrt{\delta_k}. \quad (46)$$

Using (43), (46), and the gradient-error bound (32), we get

$$F_\gamma(\mathbf{z}^{k+1}) \leq F_\gamma(\mathbf{z}^k) - c_1 \alpha \|\mathbf{G}_\gamma(\mathbf{z}^k)\|_2^2 + c_2 \alpha \gamma^2 L_x^2 \mu \epsilon_k + C_{\text{op}} \sqrt{\delta_k}. \quad (47)$$

Summing (47) for $k = 1, \dots, K$, using $F_\gamma(\mathbf{z}^{K+1}) \geq F_\gamma^{\text{inf}}$, the same bound on $\sum_k \epsilon_k$ as in the proof of Theorem 2, and $\sum_{k=1}^K \sqrt{\delta_k} \leq D_\delta$, yields

$$\frac{1}{K} \sum_{k=1}^K \|\mathbf{G}_\gamma(\mathbf{z}^k)\|_2^2 \leq \frac{F_\gamma(\mathbf{z}^1) - F_\gamma^{\text{inf}}}{c_1 \alpha K} + \frac{c_2 c \pi^2 L_x^2 \mu}{6 c_1 \alpha^2 K} + \frac{C_{\text{op}} D_\delta}{c_1 \alpha K}.$$

Renaming $C_{\text{op}}/(c_1 \alpha)$ as C_{op} gives (11). Hence the $O(1/K)$ rate is preserved. $\square$

E Proof of Lemma 3

Fix k and drop it from the notation: write $\mathbf{s}_t, \tilde{\mathbf{y}}_t, \mathbf{u}_t$ for the inner iterates, $T = T_k$, $g_s(\cdot) = g_s(\mathbf{x}^k, \cdot)$, $g_i = g_{n,i}$, $\phi^* = v(\mathbf{x}^k) = \phi(\mathbf{y}^*)$, and $\mathbf{w}_t := \mathbf{u}_t - \tilde{\mathbf{y}}_t$, so that $\mathbf{s}_{t+1} = \mathbf{s}_t + \mathbf{w}_t$, $\bar{\mathbf{y}}_T = \frac{1}{T} \sum_{t=0}^{T-1} \tilde{\mathbf{y}}_t$ and $\bar{\mathbf{u}}_T := \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{u}_t$.

Step 0: fixed points and Fejér monotonicity. Optimality of $\mathbf{y}^*$ provides $\boldsymbol{\xi}_1 \in \partial g_1(\mathbf{y}^*)$ and $\boldsymbol{\xi}_2 \in \partial g_2(\mathbf{y}^*)$ with $\nabla g_s(\mathbf{y}^*) + \boldsymbol{\xi}_1 + \boldsymbol{\xi}_2 = \mathbf{0}$ (by the Lipschitz hypothesis of Lemma 3, g_1 is finite and continuous on a neighborhood of $\mathbf{y}^*$, so the convex subdifferential sum rule applies at $\mathbf{y}^*$). Then $\mathbf{s}^* := \mathbf{y}^* + \beta \boldsymbol{\xi}_2$ is a fixed point of the Davis–Yin operator [20, Lemma 2.2]: indeed $\text{prox}_{\beta g_2}(\mathbf{s}^*) = \mathbf{y}^*$ and $\text{prox}_{\beta g_1}(2\mathbf{y}^* - \mathbf{s}^* - \beta \nabla g_s(\mathbf{y}^*)) = \text{prox}_{\beta g_1}(\mathbf{y}^* + \beta \boldsymbol{\xi}_1) = \mathbf{y}^*$. Moreover, since the L_{g_1} -Lipschitz continuity of g_1 on an open set containing $\mathbf{y}^*$ bounds its subgradients there, $\|\boldsymbol{\xi}_1\|_2 \leq L_{g_1}$ and

$$\|\mathbf{s}^* - \mathbf{y}^*\|_2 = \beta \|\boldsymbol{\xi}_2\|_2 = \beta \|\boldsymbol{\xi}_1 + \nabla g_s(\mathbf{y}^*)\|_2 \leq \beta(L_{g_1} + G_*). \quad (48)$$

Because g_s is convex with L_g -Lipschitz gradient, ∇g_s is $(1/L_g)$ -cocoercive (Baillon–Haddad), so for $\beta \in (0, 2/L_g)$ and unit relaxation the iteration is Fejér monotone with respect to the fixed-point set [20, Theorem 2.1]; in particular

$$\|\mathbf{s}_t - \mathbf{s}^*\|_2 \leq D_0 \quad \forall t, \quad \left\| \sum_{t=0}^{T-1} \mathbf{w}_t \right\|_2 = \|\mathbf{s}_T - \mathbf{s}_0\|_2 \leq 2D_0. \quad (49)$$

Step 1: one inequality per operator. The proximal characterization ($\mathbf{p} = \text{prox}_{\beta\varphi}(\mathbf{z}) \iff \langle \mathbf{z} - \mathbf{p}, \mathbf{w} - \mathbf{p} \rangle \leq \beta(\varphi(\mathbf{w}) - \varphi(\mathbf{p})) \forall \mathbf{w}$), applied at $\mathbf{w} = \mathbf{y}^*$ to (14) and (16), yields

$$\beta(g_2(\tilde{\mathbf{y}}_t) - g_2(\mathbf{y}^*)) \leq \langle \mathbf{s}_t - \tilde{\mathbf{y}}_t, \tilde{\mathbf{y}}_t - \mathbf{y}^* \rangle, \quad (50)$$

$$\beta(g_1(\mathbf{u}_t) - g_1(\mathbf{y}^*)) \leq \langle 2\tilde{\mathbf{y}}_t - \mathbf{s}_t - \beta \nabla g_s(\tilde{\mathbf{y}}_t) - \mathbf{u}_t, \mathbf{u}_t - \mathbf{y}^* \rangle. \quad (51)$$

For the smooth term, convexity and L_g -smoothness give the strengthened convexity bound [17, Theorem 5.8]

$$\beta(g_s(\tilde{\mathbf{y}}_t) - g_s(\mathbf{y}^*)) \leq \beta \langle \nabla g_s(\tilde{\mathbf{y}}_t), \tilde{\mathbf{y}}_t - \mathbf{y}^* \rangle - \frac{\beta}{2L_g} \|\boldsymbol{\delta}_t\|_2^2, \quad \boldsymbol{\delta}_t := \nabla g_s(\tilde{\mathbf{y}}_t) - \nabla g_s(\mathbf{y}^*). \quad (52)$$

Step 2: per-iteration energy inequality. Add (50)–(52). The gradient terms combine into $-\beta \langle \nabla g_s(\tilde{\mathbf{y}}_t), \mathbf{w}_t \rangle$, while, using $2\tilde{\mathbf{y}}_t - \mathbf{s}_t - \mathbf{u}_t = -(\mathbf{s}_t - \tilde{\mathbf{y}}_t) - \mathbf{w}_t$ and $\mathbf{s}_{t+1} = \mathbf{s}_t + \mathbf{w}_t$,

$$\langle \mathbf{s}_t - \tilde{\mathbf{y}}_t, \tilde{\mathbf{y}}_t - \mathbf{y}^* \rangle + \langle 2\tilde{\mathbf{y}}_t - \mathbf{s}_t - \mathbf{u}_t, \mathbf{u}_t - \mathbf{y}^* \rangle = \langle \mathbf{w}_t, \mathbf{y}^* - \mathbf{s}_t \rangle - \|\mathbf{w}_t\|_2^2 = \frac{1}{2} \|\mathbf{s}_t - \mathbf{y}^*\|_2^2 - \frac{1}{2} \|\mathbf{s}_{t+1} - \mathbf{y}^*\|_2^2 - \frac{1}{2} \|\mathbf{w}_t\|_2^2.$$

Hence, writing the mixed-point gap $\Delta_t := g_s(\tilde{\mathbf{y}}_t) + g_2(\tilde{\mathbf{y}}_t) + g_1(\mathbf{u}_t) - \phi^*$ and splitting $\nabla g_s(\tilde{\mathbf{y}}_t) = \nabla g_s(\mathbf{y}^*) + \boldsymbol{\delta}_t$ with Young’s inequality $-\beta \langle \boldsymbol{\delta}_t, \mathbf{w}_t \rangle \leq \frac{\beta}{2L_g} \|\boldsymbol{\delta}_t\|_2^2 + \frac{\beta L_g}{2} \|\mathbf{w}_t\|_2^2$,

$$\beta \Delta_t \leq \frac{1}{2} \|\mathbf{s}_t - \mathbf{y}^*\|_2^2 - \frac{1}{2} \|\mathbf{s}_{t+1} - \mathbf{y}^*\|_2^2 - \frac{1}{2} (1 - \beta L_g) \|\mathbf{w}_t\|_2^2 - \beta \langle \nabla g_s(\mathbf{y}^*), \mathbf{w}_t \rangle, \quad (53)$$

where the third term on the right-hand side is nonpositive because $\beta \leq 1/L_g$.

Step 3: telescoping and averaging. Sum (53) over $t = 0, \dots, T-1$, drop $-\frac{1}{2} \|\mathbf{s}_T - \mathbf{y}^*\|_2^2 \leq 0$, and bound the last term by Cauchy–Schwarz and (49): $-\beta \langle \nabla g_s(\mathbf{y}^*), \sum_t \mathbf{w}_t \rangle \leq \beta G_* \|\mathbf{s}_T - \mathbf{s}_0\|_2 \leq 2\beta G_* D_0$. Dividing by βT and applying Jensen’s inequality to the convex functions g_s, g_1, g_2 ,

$$g_s(\bar{\mathbf{y}}_T) + g_2(\bar{\mathbf{y}}_T) + g_1(\bar{\mathbf{u}}_T) - \phi^* \leq \frac{1}{T} \sum_{t=0}^{T-1} \Delta_t \leq \frac{1}{T} \left(\frac{\|\mathbf{s}_0 - \mathbf{y}^*\|_2^2}{2\beta} + 2G_* D_0 \right). \quad (54)$$

Step 4: from mixed points to the shadow average. Since $\bar{\mathbf{u}}_T - \bar{\mathbf{y}}_T = \frac{1}{T} \sum_t \mathbf{w}_t = (\mathbf{s}_T - \mathbf{s}_0)/T$, inequality (49) gives $\|\bar{\mathbf{y}}_T - \bar{\mathbf{u}}_T\|_2 \leq 2D_0/T$. Both averages are convex combinations of inner iterates, hence lie in the convex set on which g_1 is L_{g_1} -Lipschitz, so $g_1(\bar{\mathbf{y}}_T) \leq g_1(\bar{\mathbf{u}}_T) + 2L_{g_1} D_0/T$. Adding this to (54) and bounding $\|\mathbf{s}_0 - \mathbf{y}^*\|_2 \leq D_0 + \beta(L_{g_1} + G_*)$ via (48) yields (19). The returned point $\hat{\mathbf{y}}^{k+1}$ minimizes $g(\mathbf{x}^k, \cdot)$ over a candidate set containing $\bar{\mathbf{y}}_T$, so it satisfies the same bound.

Finally, if $g_{n,2} = \iota_{\mathcal{O}}$ then $\tilde{\mathbf{y}}_t = \text{proj}_{\mathcal{O}}(\mathbf{s}_t) \in \mathcal{O}$ for every t and $\bar{\mathbf{y}}_T \in \mathcal{O}$ by convexity of $\mathcal{O}$, so $g_2(\tilde{\mathbf{y}}_t) = g_2(\bar{\mathbf{y}}_T) = g_2(\mathbf{y}^*) = 0$; every step above holds verbatim (g_2 enters only through (50), evaluated at feasible points), and no Lipschitz continuity of $g_{n,2}$ is used anywhere. $\square$

F Relaxing strong convexity: quadratic growth, PL, and the role of uniqueness

Assumption 2 (strong convexity of the lower level) is invoked throughout the paper, but a careful reading of the proofs shows that it is used *only* as a convenient sufficient condition for two logically independent properties of the lower-level (LL) solution:

- (P1) a *uniform quadratic-growth* (equivalently, proximal Polyak–Łojasiewicz / error-bound) condition, and
- (P2) *single-valuedness* of the LL solution map.

This appendix disentangles the two. We show that each result of the paper requires only a specific subset of (P1)–(P2), and we restate the key theorems under the weakest condition that suffices. The practical conclusion is that strong convexity can be replaced by “uniform quadratic growth + unique minimizer” *without altering any proof beyond the substitution* $\sigma_g/2 \mapsto 1/\mu$, and that the genuine watershed—the property that cannot be dropped—is uniqueness (P2), which is needed precisely by the results that rely on differentiability of the value function.

Throughout, $v(\mathbf{x}) = \min_{\mathbf{y} \in \mathcal{Y}} g(\mathbf{x}, \mathbf{y})$ is the LL value function, $\mathcal{S}(\mathbf{x}) = \arg \min_{\mathbf{y} \in \mathcal{Y}} g(\mathbf{x}, \mathbf{y})$ is the (possibly set-valued) LL solution map, and $d_{\mathcal{S}(\mathbf{x})}(\mathbf{y}) := \min_{\mathbf{u} \in \mathcal{S}(\mathbf{x})} \|\mathbf{y} - \mathbf{u}\|_2$.

F.1 The two properties bundled by strong convexity

AS 6 (Uniform quadratic growth, set form). *There exists $\mu > 0$ such that, for all $\mathbf{x} \in \mathcal{X}$ and all $\mathbf{y} \in \mathcal{Y}$,*

$$g(\mathbf{x}, \mathbf{y}) - v(\mathbf{x}) \geq \frac{1}{\mu} d_{\mathcal{S}(\mathbf{x})}^2(\mathbf{y}). \quad (55)$$

AS 7 (Single-valued solution map). *For every $\mathbf{x} \in \mathcal{X}$ the LL minimizer is unique, i.e. $\mathcal{S}(\mathbf{x}) = \{\mathbf{y}^*(\mathbf{x})\}$.*

AS 8 (Uniform quadratic growth at the minimizer). *Assumption 7 holds and there exists $\mu > 0$ such that, for all $\mathbf{x} \in \mathcal{X}$ and all $\mathbf{y} \in \mathcal{Y}$,*

$$g(\mathbf{x}, \mathbf{y}) - v(\mathbf{x}) \geq \frac{1}{\mu} \|\mathbf{y} - \mathbf{y}^*(\mathbf{x})\|_2^2. \quad (56)$$

Assumption 8 is exactly Assumption 6 + Assumption 7: under uniqueness $d_{\mathcal{S}(\mathbf{x})}(\mathbf{y}) = \|\mathbf{y} - \mathbf{y}^*(\mathbf{x})\|_2$, so (55) and (56) coincide. We keep the two forms separate because, as we detail below, several results need only the *set* form (55), which tolerates a continuum of minimizers, whereas the value-function regularity results need the *point* form (56).

Relation to PL. For a convex composite LL of the form $g(\mathbf{x}, \cdot) = g_s(\mathbf{x}, \cdot) + g_{n,1} + g_{n,2}$ with g_s smooth, the natural gradient-domination condition is the *proximal-PL* inequality of [18]: writing the LL proximal-gradient (forward–backward) residual at step β as

$$\mathcal{G}_\beta(\mathbf{x}, \mathbf{y}) := \frac{1}{\beta} \left(\mathbf{y} - \text{prox}_{\beta(g_{n,1} + g_{n,2})}(\mathbf{y} - \beta \nabla_{\mathbf{y}} g_s(\mathbf{x}, \mathbf{y})) \right),$$

the LL satisfies proximal-PL with modulus $\kappa > 0$ if $\frac{1}{2} \|\mathcal{G}_\beta(\mathbf{x}, \mathbf{y})\|_2^2 \geq \kappa (g(\mathbf{x}, \mathbf{y}) - v(\mathbf{x}))$ for all $\mathbf{x}, \mathbf{y}$. For convex objectives, proximal-PL, the quadratic-growth condition (55), and the proximal error bound are mutually equivalent up to constants [18, 19]. We therefore treat “PL” and “QG” as the same tier and work with QG, which is the form the proofs actually consume. Crucially, *none* of these conditions implies uniqueness: e.g. overparameterized least squares is PL/QG but has a continuum of minimizers. This is why “PL/QG” and “PL/QG + unique minimizer” are genuinely different hypotheses.

What strong convexity gives. If $g(\mathbf{x}, \cdot)$ is σ_g -strongly convex then Assumption 8 holds with $\mu = 2/\sigma_g$ (this is Lemma 1, restated for the unique-minimizer case), and Assumption 7 holds automatically. Thus strong convexity is the conjunction (P1) $\wedge$ (P2) with $1/\mu = \sigma_g/2$, and it is strictly stronger than either: it imposes a curvature lower bound at *every* point, whereas Assumption 8 is a one-sided bound anchored at the solution.

F.2 What each result actually needs

Table 3 summarizes, for each formal statement of the paper, the weakest condition among (55)–(56)–Assumption 7 under which its proof goes through unchanged (up to the constant substitution $\sigma_g/2 \mapsto 1/\mu$).

Table 3: Lower-level regularity actually consumed by each result. “QG (set)” = Assumption 6; “QG (pt)+U” = Assumption 8 (equivalently Assumption 6+Assumption 7). “PL” is interchangeable with “QG” throughout.

Result	Suffices	What it is used for
Lemma 1 (quadratic growth)	QG (set)	is the QG property; trivial once assumed
Value/feasibility correspondence: Def. 1 and the approximate-bilevel claim of Thm. 2 (via [10, Prop. 3])	QG (set)	bounds optimal-value gap and $d_{S(\mathbf{x})}(\mathbf{y})$; set-based, no uniqueness
Inner-loop gap decay used in Cor. 1	QG (set) / PL	drives $g(\mathbf{x}^k, \hat{\mathbf{y}}^{k+1}) - v(\mathbf{x}^k) \leq \epsilon_k$; linear FB rate needs an LL error bound, not full SC (see Rem. 7)
Lemma 2 (inexact prox-grad descent)	— (none)	generic smooth+prox lemma; independent of the LL
Value-function regularity, Thm. 1	QG (pt) + U	Lipschitz solution map and $v \in C^{1,1}$ with $\nabla v(\mathbf{x}) = \nabla_{\mathbf{x}} g_s(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$
Gradient-surrogate error step $\ \hat{\mathbf{y}}^{k+1} - \mathbf{y}^*(\mathbf{x}^k)\ \leq \sqrt{\mu \epsilon_k}$ inside Thm. 2	QG (pt) + U	needs distance to the minimizer feeding the surrogate
Stationarity rate, Thm. 2	QG (pt) + U	inherits $C^{1,1}$ of v from Thm. 1
Inexact-outer-prox rate, Prop. 1	QG (pt) + U	inherits the hypotheses of Thm. 2

The table makes the dividing line explicit. Every result that lives at the level of *values and distances to the solution set* needs only **(P1)** and is indifferent to uniqueness. Every result that forms or differentiates the value-function gradient ∇v needs **(P2)** as well, because that is what makes ∇v exist as a single vector rather than a Clarke subdifferential (see Section F.4).

F.3 Restated results under the relaxed hypotheses

We now state the relaxed versions. Proofs are identical to those in Appendices A–B after the substitutions indicated; we point to the exact lines that change.

Lemma 4 (Quadratic growth, relaxed). *Under Assumption 6, inequality (55) holds by definition; under Assumption 8 it holds with $d_{S(\mathbf{x})}(\mathbf{y}) = \|\mathbf{y} - \mathbf{y}^*(\mathbf{x})\|_2$. In particular every use of Lemma 1 in the paper remains valid with $\frac{1}{\mu}$ in place of $\frac{\sigma_g}{2}$.*

Theorem 3 (Solution-map Lipschitzness and value-function smoothness, relaxed). *Replace Assumption 2 by Assumption 8, and keep Assumptions 3, 4, 5. Then the conclusions of Theorem 1 hold verbatim: (a) $\mathbf{y}^*$ is L_S -Lipschitz on $\mathcal{X}$ with*

$$L_S = \frac{\mu}{2} \left(L_x + \sqrt{L_x^2 + \frac{4L_x}{\mu}} \right);$$

(b) v is differentiable with $\nabla v(\mathbf{x}) = \nabla_{\mathbf{x}} g_s(\mathbf{x}, \mathbf{y}^(\mathbf{x}))$ and ∇v is L_v -Lipschitz with $L_v = L_x(1 + L_S)$.*

Proof (changes only). The proof of Theorem 1 in Appendix A uses strong convexity at a *single point*: equation (25), namely $g(\mathbf{x}, \mathbf{y}) - g(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) \geq \frac{\sigma_g}{2} \|\mathbf{y} - \mathbf{y}^*(\mathbf{x})\|_2^2$. This is precisely Assumption 8 with $\frac{1}{\mu}$ in place of $\frac{\sigma_g}{2}$. Substituting it, Step 3 yields $\frac{1}{\mu} a^2 \leq L_x b^2 + L_x a b$ with $a = \|\mathbf{y}_1^* - \mathbf{y}_2^*\|_2$, $b = \|\mathbf{x}_1 - \mathbf{x}_2\|_2$; solving the quadratic in a gives the stated L_S (which reduces to the original constant when $\mu = 2/\sigma_g$). Step 4 is unchanged: uniqueness (Assumption 7, part of Assumption 8) is what licenses the Danskin-type formula $\nabla v(\mathbf{x}) = \nabla_{\mathbf{x}} g_s(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$, and $L_v = L_x(1 + L_S)$ follows from (31). $\square$

Theorem 4 (Stationarity rate, relaxed). *Replace Assumption 2 by Assumption 8. Then Theorem 2 holds verbatim, with μ the quadratic-growth modulus of (56): under $\alpha = 1/L$, $L = L_f + \gamma(L_g + L_v)$, the inner schedule $\epsilon_k \leq c/(\alpha\gamma k)^2$, and $\gamma \geq L_c\sqrt{\mu/\delta}$,*

$$\frac{1}{K} \sum_{k=1}^K \|\mathbf{G}_\gamma(\mathbf{z}^k)\|_2^2 \leq \frac{F_\gamma(\mathbf{z}^1) - F_\gamma^{\inf}}{c_1 \alpha K} + \frac{c_2 c \pi^2 L_x^2 \mu}{6 c_1 \alpha^2 K}.$$

The approximate-bilevel conclusion holds under the weaker Assumption 6 alone.

Proof (changes only). The convergence proof in Appendix B touches the LL through exactly two channels, both already expressed via μ . First, the gradient-error bound (32), $\|\nabla \bar{F}_\gamma(\mathbf{z}^k) - \widehat{\nabla} \bar{F}_\gamma(\mathbf{z}^k | \hat{\mathbf{y}}^{k+1})\|_2 \leq \gamma L_x \|\hat{\mathbf{y}}^{k+1} - \mathbf{y}^*(\mathbf{x}^k)\|_2 \leq \gamma L_x \sqrt{\mu \epsilon_k}$, where the first inequality uses $\nabla v(\mathbf{x}) = \nabla_{\mathbf{x}} g_s(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$ (Theorem 3, hence Assumption 8) and the second uses the point-form quadratic growth (56). Second, L_v enters L through Theorem 3. Every other step (Lemma 2, telescoping, stationarity) is independent of the LL regularity. The final approximate-bilevel statement comes from [10, Prop. 3], which depends only on the penalty formulation and the value gap, hence holds under Assumption 6 alone. $\square$

Remark 6 (Inexact outer prox). *Proposition 1 uses the LL only through Theorem 2 and the same error bound (32); it therefore holds verbatim under Assumption 8.*

Remark 7 (Inner-iteration complexity). *The interface required by Corollary 1 is only that the inner solver returns a point with LL value gap $\leq \epsilon_k$, which is a set-based requirement (Assumption 6). The constants are finer. The linear forward-backward rate $\rho = 1 - \sigma_g \beta$ quoted for Solver A uses strong convexity, but it survives the relaxation: a uniform quadratic-growth / error-bound condition on the LL already implies Q -linear convergence of proximal-gradient on the value gap [18, 19, 29], so $T_k = \Omega(\log(\alpha\gamma k))$ still suffices under Assumption 8. The Davis–Yin ergodic $O(1/T_k)$ rate of Solver B does not use strong convexity of the LL (only its convex, Lipschitz-continuous structure), so it too is unaffected, provided the regularity assumptions of Lemma 3 remain in force.*

F.4 Why uniqueness is the watershed

It is worth isolating why **(P2)** can be dropped for the value-level results but not for the gradient-level ones, since this delimits exactly how far the relaxation can go.

The value function $v(\mathbf{x}) = \min_{\mathbf{y} \in \mathcal{Y}} g(\mathbf{x}, \mathbf{y})$ is Lipschitz whenever $\nabla_{\mathbf{x}} g_s$ is bounded on the relevant region, *regardless of uniqueness*, as an infimum of uniformly Lipschitz functions. By Danskin’s theorem its directional derivative is

$$v'(\mathbf{x}; \mathbf{d}) = \min_{\mathbf{y}^* \in \mathcal{S}(\mathbf{x})} \langle \nabla_{\mathbf{x}} g_s(\mathbf{x}, \mathbf{y}^*), \mathbf{d} \rangle,$$

and its Clarke subdifferential is $\partial v(\mathbf{x}) = \text{conv}\{\nabla_{\mathbf{x}} g_s(\mathbf{x}, \mathbf{y}^*) : \mathbf{y}^* \in \mathcal{S}(\mathbf{x})\}$. When $\mathcal{S}(\mathbf{x})$ is a singleton, $v'(\mathbf{x}; \cdot)$ is linear and v is differentiable with $\nabla v(\mathbf{x}) = \nabla_{\mathbf{x}} g_s(\mathbf{x}, \mathbf{y}^*(\mathbf{x}))$; when $\mathcal{S}(\mathbf{x})$ is a continuum, $v'(\mathbf{x}; \cdot)$ is a minimum of linear functions, hence concave and generally *not* linear, so v fails to be differentiable.

This has two consequences for our algorithm. (i) The smooth descent machinery of Lemma 2 requires $\bar{F}_\gamma = f + \gamma(g_s - v)$ to have an L -Lipschitz gradient, i.e. $v \in C^{1,1}$; without uniqueness $\bar{F}_\gamma$ is merely Lipschitz and the $O(1/K)$ gradient-mapping rate is not available. (ii) The hypergradient-free surrogate $\nabla v(\mathbf{x}) \approx \nabla_{\mathbf{x}} g_s(\mathbf{x}, \hat{\mathbf{y}}^{k+1})$ presupposes that $\nabla_{\mathbf{x}} g_s$ is essentially constant over $\mathcal{S}(\mathbf{x})$; with multiple minimizers the inner iterate $\hat{\mathbf{y}}^{k+1}$ may satisfy the gap bound while landing near an arbitrary vertex of $\partial v(\mathbf{x})$, so the error $\|\nabla v(\mathbf{x}^k) - \nabla_{\mathbf{x}} g_s(\mathbf{x}^k, \hat{\mathbf{y}}^{k+1})\|_2$ controlled in (32) is not even well defined. The penalty-to-bilevel correspondence of [10, Prop. 3], by contrast, is stated through the value gap and the distance $d_{\mathcal{S}(\mathbf{x})}(\mathbf{y})$ to the solution set, both of which remain meaningful for set-valued $\mathcal{S}(\mathbf{x})$; that is exactly why it transfers under Assumption 6 alone.

Remark 8 (Dropping uniqueness entirely). *Removing **(P2)** is not a matter of weaker constants but of a different analysis: one would replace the smooth proximal-gradient framework by a nonsmooth one (Clarke subgradients of v , or a Moreau/Tikhonov smoothing of the value function), obtaining a weaker stationarity notion and, in general, a slower rate. We do not pursue this here; within the present single-loop smooth-penalty framework, uniqueness of the LL minimizer is essential.*

Remark 9 (Net effect). *Replacing Assumption 2 by Assumption 8 changes no proof beyond the substitution $\sigma_g/2 \mapsto 1/\mu$ and leaves every rate identical, while broadening the scope to lower levels that are not strongly convex but have a unique, quadratically-sharp minimizer (e.g. restricted/essentially strongly convex objectives, or objectives strongly convex only near the solution). In both instances of Section 7 the coupling term $\frac{\eta}{2} \|\cdot\|_2^2$ supplies full strong convexity with $\sigma_g \geq \eta$, so the relaxation is immaterial there and is offered as a generalization of the statements rather than a change to the running examples.*

G Improved γ -dependence via a strengthened inner-accuracy schedule

This appendix records an improvement in the dependence on the penalty parameter γ of the stationarity rates of Theorem 2 and Proposition 1, obtained by tightening the inner-accuracy schedule by a factor α . We emphasize at the outset that *at fixed γ nothing changes*: the $O(1/K)$ rates of the main text are correct as stated. What improves is the γ -dependence of the constants, from $\Theta(\gamma^2)$ to $\Theta(\gamma)$, which matters because the approximation guarantee of Theorem 2 takes $\gamma \geq L_c \sqrt{\mu/\bar{\epsilon}}$, i.e. large.

G.1 The observation: the error term of (9) scales as γ^2

Recall $\alpha = 1/L$ with $L = L_f + \gamma(L_g + L_v)$, where $L_v = L_x(1 + L_S)$ and L_S (Theorem 1) do *not* depend on γ . Hence, setting $\gamma_0 := L_f/(L_g + L_v)$, for all $\gamma \geq \gamma_0$,

$$(L_g + L_v) \gamma \leq \frac{1}{\alpha} \leq 2(L_g + L_v) \gamma, \quad \text{so} \quad \alpha \gamma = \Theta(1), \quad \frac{1}{\alpha} = \Theta(\gamma). \quad (57)$$

The two terms of (9) therefore scale differently in γ : the first is $\frac{F_\gamma(\mathbf{z}^1) - F_\gamma^{\text{inf}}}{c_1 \alpha K} = \Theta(\gamma \Delta_1/K)$ with $\Delta_1 := F_\gamma(\mathbf{z}^1) - F_\gamma^{\text{inf}}$ (see Remark 10 for the initialization needed to make $\Delta_1 = O(1)$), whereas the second is $\frac{c_2 c \pi^2 L_x^2 \mu}{6 c_1 \alpha^2 K} = \Theta(\gamma^2/K)$. The origin of the mismatch is visible in the proof of Theorem 2 (Appendix B): the per-iteration error contribution is $c_2 \alpha \|\mathbf{e}^k\|_2^2 \leq c_2 \alpha \gamma^2 L_x^2 \mu \epsilon_k$ by (32), and the schedule $\epsilon_k \leq c/(\alpha \gamma k)^2$ of (8) cancels the factor $\alpha \gamma^2$ only up to a residual $1/\alpha = \Theta(\gamma)$:

$$c_2 \alpha \gamma^2 L_x^2 \mu \cdot \frac{c}{\alpha^2 \gamma^2 k^2} = \frac{c_2 c L_x^2 \mu}{\alpha k^2} = \Theta\left(\frac{\gamma}{k^2}\right). \quad (58)$$

Consequently, with $\gamma = L_c \sqrt{\mu/\bar{\epsilon}}$, driving the average squared gradient mapping below a target ε requires $K = O(\mu L_c^2/(\bar{\epsilon} \varepsilon))$ outer iterations, rather than the $K = O(L_c \sqrt{\mu/\bar{\epsilon}}/\varepsilon)$ one would hope for from the first term alone.

G.2 Strengthened schedule and improved rate

The fix is to tighten the inner tolerance by one factor of α : replace (8) by

$$g_s(\mathbf{x}^k, \hat{\mathbf{y}}^{k+1}) + g_{n,1}(\hat{\mathbf{y}}^{k+1}) + g_{n,2}(\hat{\mathbf{y}}^{k+1}) - v(\mathbf{x}^k) \leq \epsilon_k, \quad \epsilon_k \leq \frac{c}{\alpha \gamma^2 k^2}. \quad (59)$$

The gap bound (59) *implies* the distance bound

$$\|\hat{\mathbf{y}}^{k+1} - \mathbf{y}^*(\mathbf{x}^k)\|_2 \leq \sqrt{\mu \epsilon_k} \leq \sqrt{c \mu / \alpha} / (\gamma k), \quad (60)$$

by the quadratic growth of Lemma 1. We stress that this implication is *directional*: gap control certifies distance control, but the converse holds only in the smooth (forward-backward) regime. Indeed, a distance certificate implies a gap bound through $g(\mathbf{x}, \mathbf{y}) - v(\mathbf{x}) \leq \frac{L_g}{2} \|\mathbf{y} - \mathbf{y}^*(\mathbf{x})\|_2^2$, which is valid only where $g(\mathbf{x}, \cdot)$ is L_g -smooth; near the kinks of the non-smooth terms $g_{n,1}, g_{n,2}$ the gap grows *linearly* in the distance (the subgradient of the non-smooth part is bounded away from zero), so no quadratic gap-from-distance bound holds. Thus, in the Davis–Yin regime, (60) is a consequence of (59), not an equivalent certificate; in the FB/smooth regime a distance certificate with constant $L_g/2$ conversely certifies the gap, and the two forms are interchangeable up to the condition number L_g/μ . The analysis below is driven entirely by the gap form (59), so this asymmetry does not affect any of the rates; it only concerns which quantity may legitimately be used as the inner solver’s stopping certificate. Since $\alpha < 1$, (59) is stronger than (8) by exactly the factor α .

Theorem 5 (Stationarity rate, strengthened schedule). *Under the hypotheses of Theorem 2, with (8) replaced by (59),*

$$\frac{1}{K} \sum_{k=1}^K \|\mathbf{G}_\gamma(\mathbf{z}^k)\|_2^2 \leq \frac{F_\gamma(\mathbf{z}^1) - F_\gamma^{\text{inf}}}{c_1 \alpha K} + \frac{c_2 c \pi^2 L_x^2 \mu}{6 c_1 \alpha K}. \quad (61)$$

All remaining conclusions of Theorem 2 (stationarity of limit points; the ϵ -approximation guarantee with $\epsilon \leq \bar{\epsilon}$ for $\gamma \geq L_c \sqrt{\mu/\bar{\epsilon}}$) hold unchanged. In particular, by (57) both terms of (61) are $\Theta(\gamma/K)$ whenever $F_\gamma(\mathbf{z}^1) - F_\gamma^{\text{inf}} = O(1)$.

Proof (changes only). Exactly two displays of Appendix B change. The per-iteration error bound becomes $c_2 \alpha \|\mathbf{e}^k\|_2^2 \leq c_2 \alpha \gamma^2 L_x^2 \mu \epsilon_k \leq c_2 c L_x^2 \mu k^{-2}$, now free of α and γ ; summing, $\sum_{k=1}^K c_2 \alpha \|\mathbf{e}^k\|_2^2 \leq c_2 c L_x^2 \mu \pi^2/6 =: \bar{c}$ (previously this sum carried an extra factor $1/\alpha$). Dividing the telescoped descent inequality by $c_1 \alpha K$ yields (61). Every other step is untouched. $\square$

Corollary 2 (Inner-iteration complexity, strengthened schedule). *The schedule (59) is realized as follows. If the inner solver contracts linearly, $\epsilon_k \leq C_{\text{FB}} \rho^{T_k}$, then $T_k = \Omega(\log(\alpha \gamma^2 k^2)) = \Omega(\log(\gamma k))$ suffices — versus $\Omega(\log k)$ under (8): an additive $\log \gamma$, i.e. asymptotically free. If it returns the Davis–Yin output (18) with ergodic gap $\epsilon_k \leq C_{\text{DY}}/T_k$ (Lemma 3), then $T_k = \Omega(\alpha \gamma^2 k^2) = \Omega(\gamma k^2)$ suffices — versus $\Omega(k^2)$: a multiplicative factor γ per outer iteration (but see Section G.5 for the total-work comparison, which nevertheless favors the strengthened schedule).*

G.3 A uniform bounded region, and the inexact outer prox

Proposition 1 inherits the improvement of its middle term verbatim. Its *third* term, however, hides its own γ^2 : tracing the proof in Appendix D, the constant multiplying $\sqrt{\delta_k}$ is $C_{\text{op}} = \sqrt{2}(M_\gamma + R_\gamma/(2\alpha))$, where $M_\gamma := \sup \|\nabla \bar{F}_\gamma\|_2$ and R_γ bounds the prox displacements over the set in which the iterates live. The next lemma discharges the boundedness hypothesis of Proposition 1 and, crucially, provides a radius for the iterates that does *not* grow with γ ; this is what keeps $M_\gamma = O(\gamma)$ (the γ coming only from the penalty prefactor, not from a growing radius) and $R_\gamma = O(1)$.

Lemma 5 (Uniform bounded region; iterate radius nonincreasing in γ). *Suppose f is bounded below on $\mathcal{Z}$ with $f^{\text{inf}} := \inf_{\mathcal{Z}} f > -\infty$, and that the upper-level component of the objective is coercive along the trajectory in the following uniform sense: there is a radius $R_x < \infty$, independent of $\gamma \geq \gamma_0$, such that $f(\mathbf{z}) \leq F^0 + \bar{c}$ implies $\|\mathbf{x}\|_2 \leq R_x$, and the LL minimizer is bounded on that set, $\|\mathbf{y}^*(\mathbf{x})\|_2 \leq R_y$ whenever $\|\mathbf{x}\|_2 \leq R_x$. Assume moreover that the algorithm is warm-started so that $F_\gamma(\mathbf{z}^1) \leq F^0$ with F^0 independent of γ (Remark 10). Then, under the strengthened schedule (59), every iterate $\mathbf{z}^k = (\mathbf{x}^k, \mathbf{y}^k)$ satisfies*

$$\|\mathbf{y}^k - \mathbf{y}^*(\mathbf{x}^k)\|_2 \leq \sqrt{\frac{\mu \Delta_0}{\gamma}}, \quad \Delta_0 := F^0 - f^{\text{inf}} + \bar{c}, \quad (62)$$

with $\bar{c} = c_2 c L_x^2 \mu \pi^2/6$ the (γ -free) sum of the gradient-error terms from Theorem 5. The iterates $\mathbf{z}^k$ lie in a ball whose radius $R_x + R_y + \sqrt{\mu \Delta_0/\gamma}$ is nonincreasing in γ ; the prox arguments $\mathbf{q}^k$ and the exact outer-prox points lie in a bounded convex set $\mathcal{B}$ whose radius, while larger, is independent of γ over $\gamma \geq \gamma_0$. For an inexact outer prox, the same conclusions hold under the additional bounded-prox-points hypothesis of Proposition 1, with Δ_0 enlarged by the (γ -free) surcharge $C_{\text{op}} D_\delta = O(1)$ of the tolerance rule in Proposition 2.

Proof. Summing the per-step descent inequality and using $\sum_k c_2 \alpha \|\mathbf{e}^k\|_2^2 \leq \bar{c}$ (proof of Theorem 5) gives $F_\gamma(\mathbf{z}^k) \leq F_\gamma(\mathbf{z}^1) + \bar{c} \leq F^0 + \bar{c}$ for all k ; this bound holds pointwise along the trajectory and presupposes no *a priori* radius. This descent-sum argument concerns the *exact* outer prox, as does the lemma. With an inexact outer prox the descent inequality carries the additional terms $C_{\text{op}} \sqrt{\delta_k}$ of Proposition 1; boundedness of the prox points is then part of the *hypotheses* of that proposition, not a conclusion of this lemma, and under the tolerance rule of Proposition 2 the total surcharge $C_{\text{op}} D_\delta = O(1)$ is uniform in γ , so the bounds below hold with Δ_0 enlarged by that (γ -free) constant.

Bounded LL gap and $\mathbf{y}$ -iterates. By definition of the value function, $g(\mathbf{x}, \mathbf{y}) - v(\mathbf{x}) \geq 0$. Moreover, on $\text{dom } r$ the penalized objective is $F_\gamma(\mathbf{z}) = f(\mathbf{z}) + \gamma(g(\mathbf{x}, \mathbf{y}) - v(\mathbf{x}))$, where $g = g_s + g_{n,1} + g_{n,2}$.

Since $f \geq f^{\inf}$ and $F_\gamma(\mathbf{z}^k) \leq F^0 + \bar{c}$,

$$\gamma(g(\mathbf{x}^k, \mathbf{y}^k) - v(\mathbf{x}^k)) = F_\gamma(\mathbf{z}^k) - f(\mathbf{z}^k) \leq F^0 + \bar{c} - f^{\inf} = \Delta_0,$$

so the LL gap obeys $g(\mathbf{x}^k, \mathbf{y}^k) - v(\mathbf{x}^k) \leq \Delta_0/\gamma$. Quadratic growth (Lemma 1) converts this into (62).

Bounded x-iterates. From $f(\mathbf{z}^k) \leq F_\gamma(\mathbf{z}^k) \leq F^0 + \bar{c}$ and the uniform coercivity hypothesis, $\|\mathbf{x}^k\|_2 \leq R_x$ with R_x independent of γ . The triangle inequality with $\|\mathbf{y}^*(\mathbf{x}^k)\|_2 \leq R_y$ and (62) then bounds $\|\mathbf{y}^k\|_2 \leq R_y + \sqrt{\mu\Delta_0/\gamma}$, so every $\mathbf{z}^k$ lies in a ball of radius $R_x + R_y + \sqrt{\mu\Delta_0/\gamma}$, nonincreasing in γ .

Bounded prox points. On this ball the inexact gradient is uniformly bounded along the trajectory: by (32) and the schedule (59),

$$\|\widehat{\nabla} \bar{F}_\gamma(\mathbf{z}^k)\|_2 \leq M_\gamma + \|\mathbf{e}^k\|_2 \leq M_\gamma + \gamma L_x \sqrt{\mu\epsilon_k} \leq M_\gamma + L_x \sqrt{c\mu/\alpha} \frac{1}{k},$$

with $M_\gamma = O(\gamma)$; since $1/\sqrt{\alpha} = O(\sqrt{\gamma})$, the second term is $O(\sqrt{\gamma})$ and is dominated by the first, so $\|\widehat{\nabla} \bar{F}_\gamma(\mathbf{z}^k)\|_2 = O(\gamma)$ uniformly in k . Consequently $\alpha\|\widehat{\nabla} \bar{F}_\gamma(\mathbf{z}^k)\|_2 = O(1)$ uniformly in γ (as $\alpha = O(1/\gamma)$), so the prox arguments $\mathbf{q}^k = \mathbf{z}^k - \alpha\widehat{\nabla} \bar{F}_\gamma(\mathbf{z}^k)$ are uniformly bounded over $\gamma \geq \gamma_0$. Because r is proper, closed and convex, $\text{prox}_{\alpha r}$ is nonexpansive; moreover, its value at any fixed $\mathbf{z}_0 \in \text{dom } r$ is bounded uniformly over $\gamma \geq \gamma_0$, since the $\mathbf{y}$ -part of αr is $(\alpha\gamma)(g_{n,1} + g_{n,2})$ with $\alpha\gamma = \Theta(1)$ by (57), the indicator part does not scale with γ , and $g_{n,1}, g_{n,2}$, being proper closed convex, admit affine minorants, so the quadratic term of the prox subproblem dominates uniformly over $\alpha\gamma$ in a compact interval. Nonexpansiveness around this uniform anchor transfers the boundedness of $\{\mathbf{q}^k\}$ to the exact outer-prox points $\tilde{\mathbf{z}}^{k+1} = \text{prox}_{\alpha r}(\mathbf{q}^k)$, and the inexact ones remain uniformly bounded whenever $\sup_k \delta_k < \infty$. Taking $\mathcal{B}$ to be the closed convex hull of all these points (intersected with the convex $\mathcal{Z}$) gives a bounded convex set, of radius independent of γ , containing every iterate, prox argument, and exact/inexact prox point. $\square$

Proposition 2 (Rate under inexact outer prox, strengthened schedules). *Under the hypotheses of Proposition 1 and Lemma 5, with the LL schedule (59) and outer-prox accuracies $\{\delta_k\}$, write $D_\delta := \sum_{k \geq 1} \sqrt{\delta_k} < \infty$. Then, with $C_{\text{op}} = \sqrt{2}(M_\gamma + R_\gamma/(2\alpha)) = O(\gamma)$,*

$$\frac{1}{K} \sum_{k=1}^K \|\mathbf{G}_\gamma(\mathbf{z}^k)\|_2^2 \leq \frac{F_\gamma(\mathbf{z}^1) - F_\gamma^{\inf}}{c_1 \alpha K} + \frac{c_2 c \pi^2 L_x^2 \mu}{6 c_1 \alpha K} + \frac{C_{\text{op}} D_\delta}{c_1 \alpha K}, \quad (63)$$

and the third term is $O(\gamma^2 D_\delta/K)$. Choosing $\delta_k = O(\gamma^{-2} k^{-2-\varsigma})$ for any $\varsigma > 0$ gives $D_\delta = O(1/\gamma)$ and hence a third term of order $O(\gamma/K)$; with $F_\gamma(\mathbf{z}^1) - F_\gamma^{\inf} = O(1)$ all three terms are then $O(\gamma/K)$.

Proof (changes only). The middle term changes as in Theorem 5. For the third, Lemma 5 provides a γ -uniform bounded convex set on which $M_\gamma = O(\gamma)$ (the γ coming only from $\nabla \bar{F}_\gamma = \nabla f + \gamma \nabla(g_s - v)$, the radius not growing) and $R_\gamma = O(1)$, so $C_{\text{op}} = O(\gamma)$. The proof of Proposition 1 then gives $F_\gamma(\mathbf{z}^{k+1}) \leq F_\gamma(\tilde{\mathbf{z}}^{k+1}) + C_{\text{op}} \sqrt{\delta_k}$; summing and dividing by $c_1 \alpha K = \Theta(K/\gamma)$ yields the third term of (63), of order $O(\gamma) \cdot O(\gamma) \cdot D_\delta/K = O(\gamma^2 D_\delta/K)$. The stated tolerance rule makes $D_\delta = O(1/\gamma)$. $\square$

Since the outer-prox subroutine has ergodic gap $O(1/T)$ on the 1-strongly convex subproblem Q_k , certifying δ_k costs $T \propto 1/\delta_k$, i.e. a factor γ^2 more per outer iteration than under the original schedule; Section G.5 shows the trade is still favorable in total work.

G.4 Initialization

Remark 10 (Warm start; why it is needed). *The improvement of Theorem 5 is only as good as the first term of (61). Note that $\Delta_1 = F_\gamma(\mathbf{z}^1) - F_\gamma^{\inf}$ contains $\gamma(g(\mathbf{x}^1, \mathbf{y}^1) - v(\mathbf{x}^1))$: if the initial LL gap is $\Theta(1)$, then $\Delta_1 = \Theta(\gamma)$ and the first term is again $\Theta(\gamma^2/K)$, nullifying the gain*

regardless of the schedule. A single warm-start inner solve at $k = 0$, run to gap $\epsilon_0 \leq c_0/\gamma$, ensures $F_\gamma(\mathbf{z}^1) \leq f(\mathbf{z}^1) + c_0 =: F^0$; since $F_\gamma \geq f \geq f^{\inf}$ pointwise, $F_\gamma^{\inf} \geq f^{\inf}$ uniformly in γ , hence $\Delta_1 \leq f(\mathbf{z}^1) - f^{\inf} + c_0 = O(1)$. The one-off cost is $O(\log \gamma)$ FB iterations in the linear regime, or $O(\gamma)$ Davis–Yin iterations, and is negligible. This same warm start supplies the γ -independent constant F^0 used in Lemma 5. (The caveat applies equally to any $O(\gamma/K)$ claim derived from a value-function penalty.)

G.5 Total inner work: the Davis–Yin trade-off

In the FB regime the strengthened schedule is free (Corollary 2). In the DY regime it costs a factor γ per outer iteration, but it *reduces* the total inner work. To drive the average squared gradient mapping below ε (with $\Delta_1 = O(1)$): under the original schedule, $K_A = \Theta(\gamma^2/\varepsilon)$ outer iterations with $T_k = \Theta(k^2)$ give total inner work $\sum_{k \leq K_A} T_k = \Theta(K_A^3) = \Theta(\gamma^6 \varepsilon^{-3})$; under the strengthened schedule, $K_B = \Theta(\gamma/\varepsilon)$ with $T_k = \Theta(\gamma k^2)$ give $\Theta(\gamma K_B^3) = \Theta(\gamma^4 \varepsilon^{-3})$. The cubic reduction in outer iterations dominates the linear per-iteration surcharge, for a net factor of γ^2 . The same comparison for the outer-prox subroutine of Proposition 2 yields $\Theta(\gamma^{5+\varsigma} \varepsilon^{-(3+\varsigma)})$ versus $\Theta(\gamma^{6+2\varsigma} \varepsilon^{-(3+\varsigma)})$, again in favor of the strengthened schedule. We therefore recommend adopting (59) in *both* regimes.

G.6 Summary and scope

We close by comparing the two inner-accuracy schedules side by side. Recall the ground rule stated at the beginning of the appendix: *at fixed γ nothing changes* — both schedules yield the $O(1/K)$ stationarity rate. What distinguishes them is how the constant in front of $1/K$ grows with γ , and this matters because the approximation guarantee forces $\gamma = L_c \sqrt{\mu/\bar{\varepsilon}}$, i.e. large when the target accuracy $\bar{\varepsilon}$ of the approximate bilevel problem is small. Throughout this comparison $\alpha = 1/L = \Theta(1/\gamma)$ by (57), and the warm start of Remark 10 is assumed, so that $F_\gamma(\mathbf{z}^1) - F_\gamma^{\inf} = O(1)$; without it, the initialization term alone is $\Theta(\gamma^2/K)$ and the comparison below becomes void, *regardless* of the schedule.

The two schedules and the bounds they give. The *original* schedule of the main text, (8), requires the inner solver to return a point with LL gap $\epsilon_k \leq c/(\alpha \gamma k)^2 = \Theta(1/k^2)$. Under it, the two terms of the stationarity bound (9) scale differently: the initialization term is $\Theta(\gamma/K)$, but each per-iteration gradient-error contribution retains a residual factor $1/\alpha = \Theta(\gamma)$ (see (58)), so the error term — and hence the overall bound — is $O(\gamma^2/K)$. The *strengthened* schedule (59) tightens the tolerance by exactly one factor of α , requiring $\epsilon_k \leq c/(\alpha \gamma^2 k^2) = \Theta(1/(\gamma k^2))$. This removes the residual: the per-iteration error becomes $\Theta(1/k^2)$, free of γ , both terms of (61) are $\Theta(\gamma/K)$, and the overall bound improves to $O(\gamma/K)$. The inexact-outer-prox bound (63) inherits the same improvement provided the prox tolerances are tightened analogously, to $\delta_k = O(\gamma^{-2} k^{-2-\varsigma})$, which makes its third term $O(\gamma/K)$ as well.

What this means in outer complexity. Fix the two targets: $\bar{\varepsilon}$ for the approximation level of the penalized problem (which sets $\gamma = L_c \sqrt{\mu/\bar{\varepsilon}}$) and ε for the average squared gradient mapping. Equating each bound to ε and solving for K , the original schedule needs $K = O(\gamma^2/\varepsilon) = O(\mu L_c^2/(\bar{\varepsilon} \varepsilon))$ outer iterations, whereas the strengthened schedule needs $K = O(\gamma/\varepsilon) = O(L_c \sqrt{\mu/\bar{\varepsilon}}/\varepsilon)$: the dependence on the penalty accuracy improves from $1/\bar{\varepsilon}$ to $1/\sqrt{\bar{\varepsilon}}$. In the natural regime where the two accuracies are coupled, $\bar{\varepsilon} = \varepsilon$, the outer complexity drops from $O(\varepsilon^{-2})$ to $O(\varepsilon^{-3/2})$.

What the tighter tolerance costs. The price is paid inside the inner loop, and it differs across regimes (Corollary 2). In the linearly convergent FB regime the extra accuracy costs only an *additive* $\log \gamma$ per inner solve ($T_k = \Omega(\log(\gamma k))$ versus $\Omega(\log k)$), i.e. it is asymptotically free. In the sublinear DY regime it costs a *multiplicative* factor γ per outer iteration ($T_k = \Theta(\gamma k^2)$ versus $\Theta(k^2)$); nevertheless, the cubic reduction in the number of outer iterations dominates this linear surcharge, and the *total* inner work to reach ε falls from $\Theta(\gamma^6 \varepsilon^{-3})$ to $\Theta(\gamma^4 \varepsilon^{-3})$, a net saving of γ^2 (Section G.5, where the analogous comparison for the outer-prox subroutine is also favorable). We therefore recommend the strengthened schedule in both regimes. Table 4 collects the comparison.

Table 4: The two inner-accuracy schedules compared (warm start of Remark 10 assumed; $\alpha = \Theta(1/\gamma)$; $K(\varepsilon)$ is the outer-iteration count to reach average squared stationarity ε , with $\gamma = L_c \sqrt{\mu/\bar{\varepsilon}}$).

	Original (8)	Strengthened (59)
Inner tolerance ϵ_k	$c/(\alpha\gamma k)^2 = \Theta(1/k^2)$	$c/(\alpha\gamma^2 k^2) = \Theta(1/(\gamma k^2))$
Stationarity bound	$O(\gamma^2/K)$ (9)	$O(\gamma/K)$ (61)
$K(\varepsilon)$	$O(\mu L_c^2/(\bar{\varepsilon}\varepsilon))$	$O(L_c \sqrt{\mu/\bar{\varepsilon}}/\varepsilon)$
coupled ($\bar{\varepsilon} = \varepsilon$)	$O(\varepsilon^{-2})$	$O(\varepsilon^{-3/2})$
T_k , FB regime	$\Omega(\log k)$	$\Omega(\log(\gamma k))$
T_k , DY regime	$\Theta(k^2)$	$\Theta(\gamma k^2)$
Total inner work, DY	$\Theta(\gamma^6 \varepsilon^{-3})$	$\Theta(\gamma^4 \varepsilon^{-3})$

Scope. Finally, these modifications are orthogonal to the relaxation of strong convexity in Appendix F: the proof of Theorem 5 touches the lower level through the same two channels as that of Theorem 2 (the gradient-error bound and the constant L_v), so Theorem 4 admits the identical improvement, with μ reinterpreted as the quadratic-growth modulus of (56).