FROM EASY TO HARD: GEOMETRY-AWARE CONTINUATION FOR NONCONVEX BILEVEL OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Nonconvex bilevel optimization (BLO) is challenging due to its nested structure and potentially nonunique lower-level (LL) solutions. Value-function reformulations avoid explicit LL solution maps by recasting BLO as a single-level surrogate, but LL nonconvexity often makes the induced value function nonsmooth and difficult to optimize. Existing smoothed surrogates, such as Moreau-envelope-based reformulations, improve tractability by obtaining regularity through strong Euclidean regularization; however, when this regularization dominates the surrogate, the surrogate may become overly localized and mismatched with the anisotropic or non-Euclidean structure of constrained BLO. To overcome these issues, we propose GaCF, a single-loop geometry-aware continuation framework for nonconvex BLO. Instead of relying on a fixed Euclidean smoothing geometry, GaCF constructs a continuation path that starts from a well-conditioned auxiliary LL surrogate and progressively recovers the original nonconvex LL objective, thereby reducing early-stage dependence on rigid local surrogates. Along this path, GaCF uses the induced Bregman geometry to construct a Bregman-smoothed value-function surrogate and perform mirror-descent updates, so that the surrogate and update dynamics are aligned with a problem-adapted geometry. We establish a non-asymptotic convergence guarantee for GaCF under general nonconvex settings, and experiments on synthetic and real-world learning tasks demonstrate its convergence efficiency and predictive performance over existing BLO methods.

1 INTRODUCTION

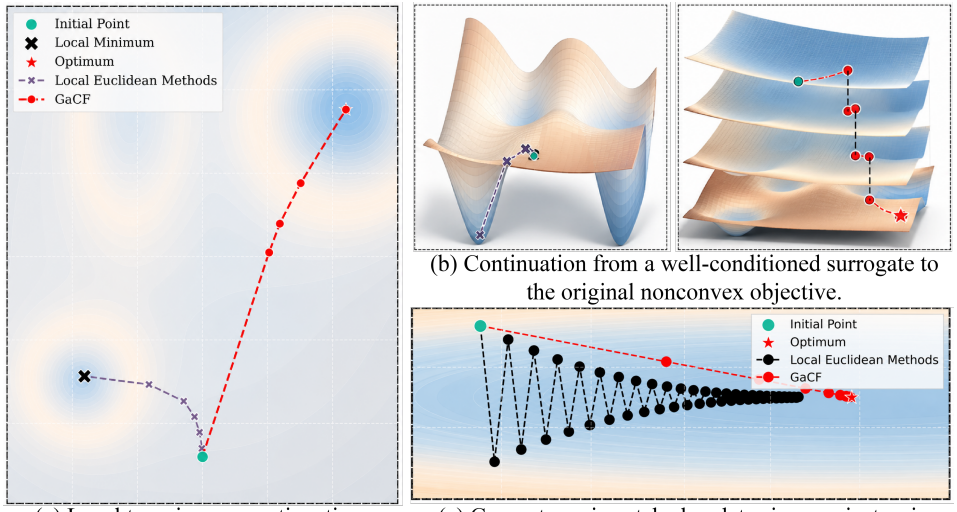
Bilevel optimization (BLO) has gained significant attention due to its powerful modeling capabilities in hierarchical learning across a wide range of real-world applications, including meta-learning (Hu et al., 2023; Killamsetty et al., 2022; Rajeswaran et al., 2019), hyperparameter optimization (Kim & Hospedales, 2025; Franceschi et al., 2018), reinforcement learning (Shen et al., 2025b; Thoma et al., 2024; Hong et al., 2023), continual learning (Wang et al., 2026; Hao et al., 2023), and fine-tuning large language models (Shen et al., 2025a; Pan et al., 2025; Lu et al., 2024). In this paper, we focus on the following class of constrained nonconvex BLO problems:

$$\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x})) \quad \text{s.t.} \quad \boldsymbol{y}^*(\boldsymbol{x}) \in \mathcal{S}(\boldsymbol{x}) := \arg\min_{\boldsymbol{y}\in\mathcal{Y}} g(\boldsymbol{x}, \boldsymbol{y}), \tag{1}$$

where the upper-level (UL) objective $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and the lower-level (LL) objective $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ are continuously differentiable and generally nonconvex, and $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{Y} \subseteq \mathbb{R}^m$ are nonempty compact convex sets. Due to the nested structure, evaluating the UL gradient (hypergradient) requires accounting for the implicit dependence of the LL solution $\boldsymbol{y}^*(\boldsymbol{x})$ on $\boldsymbol{x}$. Classical hypergradient methods, such as implicit differentiation (Chen et al., 2021; Bertrand et al., 2020) and iterative differentiation (Shaban et al., 2019; Franceschi et al., 2017), typically require second-order LL information or differentiation through long LL optimization trajectories, making them costly in large-scale or high-dimensional settings. Their theoretical guarantees also often rely on a well-defined and single-valued LL solution map, typically ensured by LL strong convexity, which limits their applicability to general nonconvex learning problems.

These limitations have motivated first-order penalty-based reformulations (Kwon et al., 2024; Shen & Chen, 2023; Liu et al., 2022), which replace solution-map differentiation with a single-level objective

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



(a) Local trapping vs. continuation-guided search.

(b) Continuation from a well-conditioned surrogate to the original nonconvex objective.

(c) Geometry-mismatched updates in an anisotropic landscape.

Figure 1: Geometric perspective on GaCF. (a) Direct optimization over the original nonconvex landscape may yield initialization-dependent trajectories and unfavorable stationary points, whereas GaCF follows a continuation path from a well-conditioned auxiliary surrogate to the original nonconvex LL objective. (b) The 3D view illustrates the progressive transition from the auxiliary surrogate to the original nonconvex objective. (c) Under anisotropic curvature, Euclidean-gradient updates may oscillate or progress slowly, while GaCF follows Bregman-metric directions adapted to the problem geometry, leading to more stable and efficient trajectories.

penalizing the LL optimality gap:

$$\min_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x},\boldsymbol{y}) + c(g(\boldsymbol{x},\boldsymbol{y}) - g^*(\boldsymbol{x})), \tag{2}$$

where $g^*(\boldsymbol{x}) := \min_{\boldsymbol{y}\in\mathcal{Y}} g(\boldsymbol{x},\boldsymbol{y})$ denotes the LL value function. This class of reformulations is attractive because it converts the bilevel problem into a single-level surrogate and avoids explicit hypergradient computation. Nevertheless, it does not remove the inherent difficulty: the nonuniqueness of LL solutions generally renders $g^*(\boldsymbol{x})$ nonsmooth and intractable to evaluate explicitly. A common remedy is to construct regularized or smoothed value-function surrogates, with Moreau-envelope-based reformulation being a representative example:

$$v_\gamma(\boldsymbol{x},\boldsymbol{y}) := \min_{\boldsymbol{\theta}\in\mathcal{Y}} g(\boldsymbol{x},\boldsymbol{\theta}) + \frac{1}{2\gamma}\|\boldsymbol{\theta}-\boldsymbol{y}\|^2, \tag{3}$$

where $\gamma > 0$ is the smoothing parameter.

Such surrogates improve tractability by regularizing the value-function structure and have enabled efficient single-loop first-order BLO algorithms (Lu, 2025; Liu et al., 2024a; Gao et al., 2023). However, this regularity often comes at the cost of an overly localized surrogate. In the Moreau envelope, for example, ensuring the desired regularity typically requires a small smoothing parameter γ, which makes the quadratic proximal term dominate the smoothed value-function surrogate. Consequently, the surrogate is strongly localized around the current reference point, restricting the search to a local neighborhood and making the trajectory sensitive to initialization, as illustrated in Figure 1(a). Beyond this locality issue, Euclidean quadratic regularization also imposes a rigid and isotropic search geometry, which can be misaligned with constrained nonconvex BLO, where the surrogate may exhibit direction-dependent curvature and the feasible sets may have non-Euclidean structures such as simplex or nonnegative constraints. Such mismatch can be amplified by Euclidean gradient or projection-based updates, leading to poorly aligned directions, oscillatory trajectories, or unstable progress under anisotropic landscapes, as illustrated in Figure 1(c). The key challenge is therefore *how to retain the tractability of single-level reformulations while avoiding the locality and metric rigidity induced by fixed Euclidean regularization, and adapt the search trajectory to the anisotropic and constraint-induced geometry of nonconvex BLO*.

1.1 MAIN CONTRIBUTIONS

To address this challenge, we propose a single-loop **G**eometry-**a**ware **C**ontinuation **F**ramework, termed GaCF, for constrained nonconvex BLO. Instead of imposing a fixed Euclidean regularization geometry, GaCF constructs the reformulation from a problem-adapted potential. This potential defines a continuation family of LL objectives that evolves from a well-conditioned surrogate to the original nonconvex LL objective, providing an easy-to-hard path that mitigates the rigidity of local surrogate search. Along this path, the potential-induced Bregman divergence smooths the value-function surrogate, while mirror descent updates the variables under the corresponding geometry. As a result, GaCF reduces excessive locality, stabilizes trajectories under anisotropic curvature or non-Euclidean constraints, and preserves single-loop first-order tractability. We further establish a non-asymptotic convergence analysis under general nonconvex settings. The main contributions are summarized as follows:

- **Geometry-aware continuation.** We propose GaCF, a single-loop framework for constrained nonconvex BLO. By constructing a continuation path from a well-conditioned auxiliary LL surrogate to the original nonconvex LL objective, GaCF mitigates the excessive locality and initialization sensitivity induced by rigid local surrogates.
- **Potential-induced smoothing and updates.** We develop a Bregman smoothing and mirror-descent scheme under a problem-adapted geometry. This couples the value-function surrogate and update dynamics in a consistent metric, enabling better alignment with anisotropic or non-Euclidean BLO structures without second-order hypergradient computation.
- **Convergence and validation.** We establish a non-asymptotic convergence guarantee for GaCF under general nonconvex settings, without assuming LL convexity or the Polyak–Łojasiewicz condition. Experiments on synthetic and real-world learning tasks, including data hypercleaning, few-shot learning, and parameter-efficient fine-tuning, demonstrate its improved stability and efficiency.

1.2 RELATED WORK

BLO algorithms and reformulation strategies. Classical gradient-based BLO methods typically rely on implicit differentiation (Khanduri et al., 2023; Bertrand et al., 2020) or iterative differentiation (Bolte et al., 2022; Shaban et al., 2019). Their analysis commonly assumes a locally unique LL solution map, often guaranteed by LL strong convexity, while hypergradient computation may require second-order information or differentiation through long LL trajectories. Within this setting, Huang et al. (Huang et al., 2022) developed deterministic and stochastic methods for nonconvex–strongly-convex BLO by incorporating Bregman proximal steps into the UL update. The reliance on a well-defined LL response and costly hypergradient evaluation limits their applicability to large-scale or nonconvex LL problems.

Value-function reformulations avoid explicit differentiation through the LL response. However, when the LL problem is nonconvex or admits multiple solutions, the resulting value function is generally nonsmooth and difficult to evaluate. Tikhonov-regularized methods (Liu et al., 2023a; Sow et al., 2022; Liu et al., 2021) introduce strongly convex regularization into the LL objective or response variable to stabilize the LL solution structure and smooth the induced value-function surrogate, but may incur asymptotic bias or numerical ill-conditioning. Moreau-envelope-based reformulations provide another route. Gao et al. (Gao et al., 2023) developed a double-loop inexact proximal difference algorithm, followed by more efficient single-loop first-order methods. In particular, Liu et al. (Liu et al., 2024a) proposed the Hessian-free MEHA method, and Lu (Lu, 2025) developed the two-sided smoothed primal-dual method (TSP) with improved convergence guarantees. Despite their efficiency, these Moreau-envelope-based methods rely on Euclidean quadratic smoothing, whose regularity often requires a small smoothing parameter and may result in an overly localized surrogate.

Trajectory geometry in nonconvex optimization. A related line of research studies how the geometry of optimization trajectories affects nonconvex optimization. Continuation, or homotopy, methods improve trajectory stability by evolving from a smoother or better-conditioned surrogate to the target problem, thereby reducing sensitivity to unfavorable local structures (Wu, 1996; Allgower & Georg, 2003; Mobahi & Fisher III, 2015; Dunlavy & O'Leary, 2005). Representative developments

include Gaussian homotopy (Starnes et al., 2023), continuous path learning (Lin et al., 2023), learning-assisted homotopy for vision problems with multiple spurious solutions (Hruby et al., 2022), and path-following strategies in convex optimization (Klingler & Netzer, 2025). In parallel, mirror descent and Bregman proximal schemes show that first-order dynamics can be substantially improved by replacing the ambient Euclidean metric with a potential-induced geometry (Nemirovskij & Yudin, 1983; Censor & Zenios, 1992; Beck & Teboulle, 2003; Bregman, 1967; Zhang & He, 2018; Ghadimi et al., 2016). These studies suggest that both the path followed by an algorithm and the metric used to follow it are critical in difficult nonconvex landscapes. However, existing trajectory-geometry methods do not directly address the geometry induced by value-function smoothing in nonconvex BLO. This observation motivates GaCF, which revisits the single-level reformulation from a geometry-aware continuation perspective and designs the surrogate and updates along this continuation path.

2 GEOMETRY-AWARE CONTINUATION REFORMULATION AND OPTIMIZATION

This section develops GaCF, a single-loop geometry-aware continuation framework for constrained nonconvex BLO. The key idea is to construct the single-level reformulation under a problem-adapted geometry, rather than imposing a fixed Euclidean regularization metric. Specifically, GaCF constructs a continuation path for the LL objective, builds a Bregman-smoothed value-function surrogate along this path, and performs first-order updates under the induced mirror geometry. In this way, the surrogate construction and optimization dynamics evolve consistently along the continuation process, enabling the trajectory to adapt to the constrained and anisotropic geometry of nonconvex BLO.

2.1 CONTINUATION-BASED PROBLEM REFORMULATION

Directly reformulating the original nonconvex LL objective may inherit a highly irregular value-function landscape and make the resulting surrogate strongly dependent on local regularization. To reduce the dependence on rigid local regularization, we introduce a potential function $\phi : \mathcal{Y} \to \mathbb{R}$ that is continuously differentiable and μ_ϕ-strongly convex on $\mathcal{Y}$, and define the continuation family

$$g_t(\boldsymbol{x}, \boldsymbol{y}) := tg(\boldsymbol{x}, \boldsymbol{y}) + (1-t)\phi(\boldsymbol{y}), \quad t \in [0, 1],$$

where t controls the transition from a well-conditioned surrogate to the target LL objective. When $t = 0$, the LL objective reduces to the strongly convex potential $\phi(\boldsymbol{y})$; when $t = 1$, it recovers the original objective $g(\boldsymbol{x}, \boldsymbol{y})$. Thus, increasing t gradually exposes the nonconvex LL structure, allowing the algorithm to follow a controlled path from a well-conditioned auxiliary surrogate to the original nonconvex objective. For each $t \in [0, 1]$, the corresponding continuation-based BLO problem is

$$\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}, \boldsymbol{y}_t^*(\boldsymbol{x})) \quad \text{s.t.} \quad \boldsymbol{y}_t^*(\boldsymbol{x}) \in \mathcal{S}_t(\boldsymbol{x}) := \arg\min_{\boldsymbol{y}\in\mathcal{Y}} g_t(\boldsymbol{x}, \boldsymbol{y}). \tag{4}$$

This continuation reformulation provides a smoother starting point, but it does not by itself remove the main difficulty of nonconvex BLO. As t increases, the LL landscape may become nonconvex, and the associated value function $g_t^*(\boldsymbol{x}) := \min_{\boldsymbol{y}\in\mathcal{Y}} g_t(\boldsymbol{x}, \boldsymbol{y})$ can still be nonsmooth when the LL problem admits nonunique solutions. Therefore, a tractable surrogate of the value function is needed. To this end, we define a Bregman-divergence-based smoothed surrogate:

$$v_\gamma^t(\boldsymbol{x}, \boldsymbol{y}) := \min_{\boldsymbol{\theta}\in\mathcal{Y}} \left\{ g_t(\boldsymbol{x}, \boldsymbol{\theta}) + \frac{1}{2\gamma} D_\phi(\boldsymbol{\theta}, \boldsymbol{y}) \right\}, \qquad \gamma > 0, \tag{5}$$

where D_ϕ is the Bregman divergence induced by ϕ, namely, $D_\phi(\boldsymbol{\theta}, \boldsymbol{y}) := \phi(\boldsymbol{\theta}) - \phi(\boldsymbol{y}) - \langle \nabla\phi(\boldsymbol{y}), \boldsymbol{\theta} - \boldsymbol{y} \rangle$. By construction, $g_t(\boldsymbol{x}, \boldsymbol{y}) \geq v_\gamma^t(\boldsymbol{x}, \boldsymbol{y})$. Unlike the Euclidean quadratic term in the Moreau envelope, the Bregman term is induced by the potential used to construct the continuation family. Hence, the smoothing geometry is not imposed independently of the reformulation path, but is tied to the geometry underlying the continuation reformulation. In GaCF, γ is scheduled together with t. At early stages, a relatively large γ avoids forcing the search to be overly localized, while the objective is still dominated by the well-conditioned potential. As t approaches 1, γ is reduced to maintain a strongly convex inner subproblem while the continuation recovers the original nonconvex objective.

Based on the smoothed surrogate v_γ^t, we formulate the following constrained problem:

$$\min_{\boldsymbol{x}\in\mathcal{X}, \boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}) \quad \text{s.t.} \quad g_t(\boldsymbol{x}, \boldsymbol{y}) - v_\gamma^t(\boldsymbol{x}, \boldsymbol{y}) \leq \varepsilon_{\text{rel}}, \tag{6}$$

where $\varepsilon_{\mathrm{rel}} \geq 0$ is a relaxation tolerance. For $t = 1$ and $\varepsilon_{\mathrm{rel}} = 0$, Proposition 1 in the appendix shows that, if $g(\boldsymbol{x}, \cdot)$ is ρ_g-weakly convex on $\mathcal{Y}$ and $\mu_\phi/(2\gamma) > \rho_g$, problem (6) is equivalent to

$$\min_{\boldsymbol{x}\in\mathcal{X},\, \boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}) \quad \text{s.t.} \quad \boldsymbol{y} \in \widetilde{\mathcal{S}}(\boldsymbol{x}), \tag{7}$$

where $\widetilde{\mathcal{S}}(\boldsymbol{x}) := \{\boldsymbol{y} \in \mathcal{Y} : 0 \in \nabla_{\boldsymbol{y}} g(\boldsymbol{x}, \boldsymbol{y}) + \mathcal{N}_{\mathcal{Y}}(\boldsymbol{y})\}$ denotes the set of LL stationary points. If $g(\boldsymbol{x}, \cdot)$ is convex, then $\widetilde{\mathcal{S}}(\boldsymbol{x}) = \mathcal{S}(\boldsymbol{x})$, and (7) coincides with the original BLO problem (1).

Write $\hat{V}_\gamma^t(\boldsymbol{x}, \boldsymbol{y}) := g_t(\boldsymbol{x}, \boldsymbol{y}) - v_\gamma^t(\boldsymbol{x}, \boldsymbol{y}) - \varepsilon_{\mathrm{rel}}$. To handle the inequality in (6), we use the augmented Lagrangian

$$\mathscr{L}_{\rho,\omega}^{\gamma,t}(\boldsymbol{x}, \boldsymbol{y}, \lambda) := f(\boldsymbol{x}, \boldsymbol{y}) + \frac{[\lambda + \rho \hat{V}_\gamma^t(\boldsymbol{x}, \boldsymbol{y})]_+^2 - \lambda^2}{2\rho} - \frac{\omega}{2}\lambda^2, \qquad \lambda \in [0, \Lambda], \tag{8}$$

where $[a]_+ := \max\{0, a\}$, $\rho > 0$ is the penalty parameter, $\omega > 0$ is the dual regularization parameter, and $\Lambda > 0$ is the projection bound.

2.2 Single-loop geometry-aware optimization algorithm

We develop a single-loop optimization procedure for solving (8). We first characterize the gradient of the Bregman-smoothed surrogate. Define

$$H_{\boldsymbol{x},\boldsymbol{y}}^{\gamma,t}(\boldsymbol{\theta}) := g_t(\boldsymbol{x}, \boldsymbol{\theta}) + \frac{1}{2\gamma} D_\phi(\boldsymbol{\theta}, \boldsymbol{y}).$$

Suppose $g(\boldsymbol{x}, \cdot)$ is ρ_g-weakly convex on $\mathcal{Y}$ for every $\boldsymbol{x} \in \mathcal{X}$. By Lemma 1, $H_{\boldsymbol{x},\boldsymbol{y}}^{\gamma,t}(\cdot)$ is strongly convex on $\mathcal{Y}$ for every $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$ whenever $(1 - t + 1/(2\gamma))\,\mu_\phi - t\rho_g > 0$. Hence the minimizer $\boldsymbol{\theta}_{\gamma,t}^*(\boldsymbol{x}, \boldsymbol{y})$ exists and is unique. By Danskin's theorem (Bertsekas, 1999; Danskin, 1967), the surrogate $v_\gamma^t(\boldsymbol{x}, \boldsymbol{y})$ is differentiable with respect to $(\boldsymbol{x}, \boldsymbol{y})$, and

$$\nabla v_\gamma^t(\boldsymbol{x}, \boldsymbol{y}) = \nabla_{(\boldsymbol{x},\boldsymbol{y})} H_{\boldsymbol{x},\boldsymbol{y}}^{\gamma,t}\big(\boldsymbol{\theta}_{\gamma,t}^*(\boldsymbol{x}, \boldsymbol{y})\big). \tag{9}$$

The expression in (9) indicates that evaluating $\nabla v_\gamma^t(\boldsymbol{x}, \boldsymbol{y})$ requires the minimizer $\boldsymbol{\theta}_{\gamma,t}^*(\boldsymbol{x}, \boldsymbol{y})$. Since computing this minimizer exactly at every iteration is generally expensive, we track it with an auxiliary sequence in the algorithm.

Inner tracking. At iteration k, given $(\boldsymbol{x}_k, \boldsymbol{y}_k, t_k, \gamma_k)$, the auxiliary variable $\boldsymbol{\theta}_{k+1}$ is updated by the mirror-descent step

$$\boldsymbol{\theta}_{k+1} = \arg\min_{\boldsymbol{\theta}\in\mathcal{Y}} \left\{ \left\langle \nabla_{\boldsymbol{\theta}} g_{t_k}(\boldsymbol{x}_k, \boldsymbol{\theta}_k) + \frac{1}{2\gamma_k} \nabla_{\boldsymbol{\theta}} D_\phi(\boldsymbol{\theta}_k, \boldsymbol{y}_k),\ \boldsymbol{\theta} - \boldsymbol{\theta}_k \right\rangle + \frac{1}{\eta} D_\phi(\boldsymbol{\theta}, \boldsymbol{\theta}_k) \right\}, \tag{10}$$

where $\eta > 0$ is the stepsize. With a suitable potential ϕ, this subproblem can often be solved in closed form. For example, negative entropy is natural for simplex constraints, while a weighted quadratic potential $\phi(\cdot) = \frac{1}{2}\|\cdot\|_{\mathbf{Q}}^2$ with $\mathbf{Q} \succ 0$ captures anisotropic coordinate scaling. When $\mathbf{Q} = \mathbf{I}$, the update reduces to the standard projected gradient form. To avoid the cost of exact inner minimization, we use $\boldsymbol{\theta}_{k+1}$ in (9) to approximate $\nabla v_{\gamma_k}^{t_k}(\boldsymbol{x}_k, \boldsymbol{y}_k)$ by $\nabla_{(\boldsymbol{x},\boldsymbol{y})} H_{\boldsymbol{x}_k,\boldsymbol{y}_k}^{\gamma_k,t_k}(\boldsymbol{\theta}_{k+1})$. The corresponding constraint approximation is

$$\tilde{V}_\gamma^t(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) := g_t(\boldsymbol{x}, \boldsymbol{y}) - H_{\boldsymbol{x},\boldsymbol{y}}^{\gamma,t}(\boldsymbol{\theta}) - \varepsilon_{\mathrm{rel}}.$$

Primal mirror-descent updates. We use the mirror potentials ψ on $\mathcal{X}$ and ϕ on $\mathcal{Y}$ for the primal updates. Replacing the constraint value and gradient in $\nabla_{\boldsymbol{x}} \mathscr{L}_{\rho,\omega}^{\gamma_k,t_k}$ by their approximations yields the direction

$$\boldsymbol{d}_x^k = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_k) + \big[\lambda_k + \rho \tilde{V}_{\gamma_k}^{t_k}(\boldsymbol{x}_k, \boldsymbol{y}_k, \boldsymbol{\theta}_{k+1})\big]_+ \nabla_{\boldsymbol{x}} \tilde{V}_{\gamma_k}^{t_k}(\boldsymbol{x}_k, \boldsymbol{y}_k, \boldsymbol{\theta}_{k+1}). \tag{11}$$

Then $\boldsymbol{x}$ is updated by

$$\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \langle \boldsymbol{d}_x^k, \boldsymbol{x} - \boldsymbol{x}_k \rangle + \frac{1}{\alpha} D_\psi(\boldsymbol{x}, \boldsymbol{x}_k) \right\}, \tag{12}$$

where $\alpha > 0$. After updating $\boldsymbol{x}$, the corresponding $\boldsymbol{y}$-direction is

$$\boldsymbol{d}_y^k = \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_{k+1}, \boldsymbol{y}_k) + \left[\lambda_k + \rho \tilde{V}_{\gamma_k}^{t_k}(\boldsymbol{x}_{k+1}, \boldsymbol{y}_k, \boldsymbol{\theta}_{k+1})\right]_+ \nabla_{\boldsymbol{y}} \tilde{V}_{\gamma_k}^{t_k}(\boldsymbol{x}_{k+1}, \boldsymbol{y}_k, \boldsymbol{\theta}_{k+1}). \tag{13}$$

The corresponding mirror step is

$$\boldsymbol{y}_{k+1} = \arg\min_{\boldsymbol{y} \in \mathcal{Y}} \left\{ \langle \boldsymbol{d}_y^k, \boldsymbol{y} - \boldsymbol{y}_k \rangle + \frac{1}{\beta} D_\phi(\boldsymbol{y}, \boldsymbol{y}_k) \right\}, \tag{14}$$

where $\beta > 0$. These Bregman proximal steps align the primal updates with the constraint geometry.

Dual and continuation updates. We initialize $\lambda_0 \in [0, \Lambda]$ and choose $0 < \rho\omega < 1$. A projected dual gradient ascent step of size ρ for (8), with the constraint replaced by its approximation, gives

$$\lambda_{k+1} = \Pi_{[0,\Lambda]} \left((1 - \rho\omega)\lambda_k + \rho \tilde{V}_{\gamma_k}^{t_k}(\boldsymbol{x}_{k+1}, \boldsymbol{y}_{k+1}, \boldsymbol{\theta}_{k+1}) \right). \tag{15}$$

Here $\Pi_{[0,\Lambda]}$ denotes the projection onto $[0, \Lambda]$. The projection limits multiplier amplification under inexact constraint evaluation, while the factor $1 - \rho\omega$ damps the contribution of past dual updates.

The continuation and smoothing parameters are then updated as

$$t_{k+1} = \min\{1, t_k + \Delta t\}, \qquad \gamma_{k+1} = \max\{\underline{\gamma}, \gamma_k - \Delta\gamma\}. \tag{16}$$

where $\Delta t > 0$, $\Delta\gamma > 0$, and $\underline{\gamma} > 0$ is a prescribed lower bound. The parameter updates drive (t_k, γ_k) to $(1, \underline{\gamma})$. Algorithm 1 summarizes the procedure.

Algorithm 1: Geometry-aware Continuation Framework (GaCF)

Input: Initial $(\boldsymbol{x}_0, \boldsymbol{y}_0, \boldsymbol{\theta}_0, \lambda_0)$ with $\lambda_0 \in [0, \Lambda]$; $t_0 = 0$, $\gamma_0 \geq \underline{\gamma} > 0$; positive parameters $\Delta t, \Delta\gamma, \alpha, \beta, \eta, \rho, \omega, \Lambda, \varepsilon_{\mathrm{rel}}$ with $0 < \rho\omega < 1$; iteration budget K

Output: $\boldsymbol{x}_K, \boldsymbol{y}_K$

for $k = 0, 1, \ldots, K-1$ **do**

- Update $\boldsymbol{\theta}_{k+1}$ by (10);
- Compute $\boldsymbol{d}_x^k$ by (11) and update $\boldsymbol{x}_{k+1}$ by (12);
- Compute $\boldsymbol{d}_y^k$ by (13) and update $\boldsymbol{y}_{k+1}$ by (14);
- Update λ_{k+1} by (15);
- Update $t_{k+1} \leftarrow \min\{1, t_k + \Delta t\}$;
- Update $\gamma_{k+1} \leftarrow \max\{\underline{\gamma}, \gamma_k - \Delta\gamma\}$;

end

return $\boldsymbol{x}_K, \boldsymbol{y}_K$

3 Convergence Analysis

In this section, we present a non-asymptotic convergence analysis of the proposed GaCF algorithm. We first state the regularity conditions on the objective and potential functions.

Assumption 1. *The functions f and g are continuously differentiable on an open neighborhood of $\mathcal{X} \times \mathcal{Y}$, with locally Lipschitz continuous gradients. For every $\boldsymbol{x} \in \mathcal{X}$, the function $g(\boldsymbol{x}, \cdot)$ is ρ_g-weakly convex on $\mathcal{Y}$.*

Assumption 2. *The mirror functions ψ and ϕ are strongly convex on $\mathcal{X}$ and $\mathcal{Y}$, with moduli $\mu_\psi, \mu_\phi > 0$, respectively. The gradients $\nabla\psi$, $\nabla\phi$, and ∇D_ϕ are locally Lipschitz continuous on open neighborhoods of $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Y} \times \mathcal{Y}$, respectively.*

Negative-entropy mirror functions are covered by the interior-localization result in the appendix.

Let $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Define the dual envelope and its unique maximizer by

$$\Phi_{\rho,\omega}(\boldsymbol{z}) := \max_{0 \leq \lambda \leq \Lambda} \mathscr{A}_{\rho,\omega}^{\underline{\gamma},1}(\boldsymbol{z}, \lambda), \quad \lambda_\omega^*(\boldsymbol{z}) = \Pi_{[0,\Lambda]} \left([\hat{V}_{\underline{\gamma}}^1(\boldsymbol{z})]_+ / \omega \right). \tag{17}$$

The residual

$$\mathcal{R}_\omega(\boldsymbol{z}, \lambda) := \mathrm{dist}^2 \left(-\nabla\Phi_{\rho,\omega}(\boldsymbol{z}), \mathcal{N}_{\mathcal{Z}}(\boldsymbol{z}) \right) + |\lambda - \lambda_\omega^*(\boldsymbol{z})|^2 \tag{18}$$

measures envelope stationarity and dual tracking error, where $\mathcal{N}_{\mathcal{Z}}$ denotes the normal cone.

Theorem 1 (Convergence Rate and Stationarity)**.** *Suppose Assumptions 1 and 2 hold. Let* $\varepsilon_{\mathrm{rel}}, \rho, \omega, \Lambda > 0$ *satisfy* $0 < \rho\omega < 1$. *Let* (t_k, γ_k) *be generated by* (16) *and satisfy* $\left(1 - t_k + \frac{1}{2\gamma_k}\right)\mu_\phi - t_k\rho_g > 0$ *for all* $k \geq 0$. *Then the iterates generated by Algorithm 1 with sufficiently small step sizes* $\eta, \alpha, \beta > 0$ *satisfy*

$$\min_{1 \leq k \leq K+1} \mathcal{R}_\omega(\boldsymbol{z}_k, \lambda_k) \leq \frac{C_\omega}{K+1}. \tag{19}$$

for $C_\omega > 0$ *independent of* K. *Moreover, every accumulation point of* $\{\boldsymbol{z}_k\}$ *is stationary for* $\Phi_{\rho,\omega}$.

The proof is deferred to the appendix. Theorem 1 gives an $\mathcal{O}(K^{-1})$ bound on the minimum of $\mathcal{R}_\omega$ over the first $K+1$ iterates, corresponding to an $\mathcal{O}(\varepsilon^{-2})$ complexity for attaining $\sqrt{\mathcal{R}_\omega} \leq \varepsilon$. Appendix Theorem 2 further establishes an $\mathcal{O}(K^{-1}) + \mathcal{O}(\omega^2)$ KKT residual bound, and Appendix Corollary 1 recovers the KKT conditions as $\omega \to 0$.

4 Experimental Results

We evaluate the convergence and practical performance of GaCF on synthetic problems and three real-world tasks: data hyper-cleaning, few-shot learning, and parameter-efficient fine-tuning. Parameter sensitivity and ablation studies are reported in the appendix.

4.1 Synthetic Numerical Experiments

LL convex case. We first consider a BLO problem with a convex LL subproblem:

$$\min_{x \in \mathbb{R}^n, \boldsymbol{y} \in \Delta} \frac{\nu}{2}||\boldsymbol{x} - \boldsymbol{a}||^2 + \frac{1}{2}||\boldsymbol{y} - \boldsymbol{b}||^2 \quad \text{s.t.} \quad \boldsymbol{y} \in \arg\min_{\boldsymbol{y} \in \Delta} \frac{1}{2}\boldsymbol{y}^\top \mathrm{diag}(\boldsymbol{q})\boldsymbol{y} + \boldsymbol{x}^\top \mathbf{A}\boldsymbol{y},$$

where Δ is the probability simplex, $\nu = 10^{-2}$, $n = 50$, and $m = 20000$. We sample $\boldsymbol{a}$ and $\mathbf{A}$ independently from zero-mean Gaussian distributions with standard deviations 0.5 and 0.1, respectively, and set $q_i = |z_i|$ for $z_i \sim \mathcal{N}(0, 1000^2)$. We choose $\boldsymbol{b}$ as an LL solution at $\boldsymbol{x} = \boldsymbol{a}$, use negative entropy for the simplex-constrained $\boldsymbol{y}$, and set $\psi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2$. Figure 2 compares GaCF with BDA (Liu et al., 2023b), BOME (Liu et al., 2022), F^{2}SA (Kwon et al., 2023), MEHA (Liu et al., 2024a), and TSP (Lu, 2025). The results show that GaCF reduces the relative solution error more rapidly and attains a lower final error. This behavior is consistent with a simplex-compatible Bregman metric, which avoids repeated Euclidean projections and yields updates better aligned with the feasible geometry.

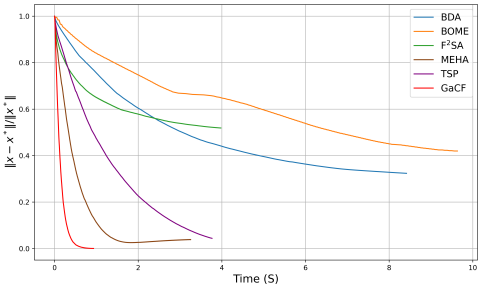


Figure 2: Convergence curves of optimization algorithms under the LL convex case.

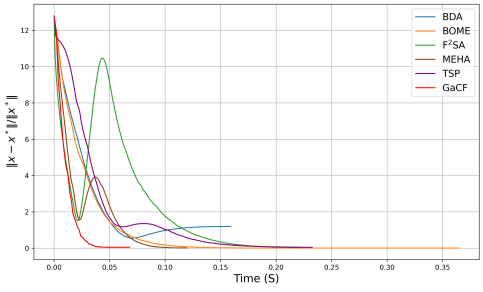


Figure 3: Convergence curves of optimization algorithms under the LL nonconvex case.

LL nonconvex case. We further evaluate GaCF on a BLO problem with a nonconvex LL problem:

$$\min_{x \in \mathbb{R},\ \boldsymbol{y} \in \mathbb{R}^{2m}} \frac{\tau}{2}x^2 + \frac{1}{2}\|\boldsymbol{y}_1\|^2 + \frac{1}{2}\|\boldsymbol{y}_2 - \mathbf{1}_m\|^2$$

$$\text{s.t.} \quad \boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2) \in \arg\min_{(\boldsymbol{y}_1, \boldsymbol{y}_2) \in \mathbb{R}^{2m}} \frac{\kappa}{2}\|\boldsymbol{y}_1 - x\mathbf{1}_m\|^2 + \frac{1}{4}(\boldsymbol{y}_2^\top \boldsymbol{y}_2 - 1)^2 + \boldsymbol{y}_1^\top \boldsymbol{y}_2,$$

Table 1: Comparison of different BLO algorithms on the data hyper-cleaning task. We report mean of 3 runs using different random seeds.

Methods	FashionMNIST			MNIST		
	Acc.(%)	F1 score	Time(S)	Acc.(%)	F1 score	Time(S)
BDA	82.81	87.21	40.13	90.94	91.01	42.86
BOME	83.01	88.93	13.03	92.04	91.05	12.61
F^2SA	84.56	87.77	11.21	91.93	90.98	11.32
MEHA	83.48	89.13	8.78	91.47	90.52	9.15
TSP	84.19	90.19	9.21	92.24	92.75	9.48
GaCF	**85.01**	**93.43**	**8.22**	**92.95**	**96.04**	**8.47**

where $\tau = 0.1$, $m = 20$, and $\mathbf{1}_m$ is the all-ones vector. To induce anisotropy, we set $\kappa = 10^4$, use $\phi(\boldsymbol{y}) = \frac{1}{2}\boldsymbol{y}^\top \operatorname{diag}(\boldsymbol{p})\boldsymbol{y}$ with $\boldsymbol{p} = [\kappa\mathbf{1}_m; \mathbf{1}_m]$, and set $\psi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2$. Figure 3 shows that GaCF converges faster and more smoothly than the baselines: continuation reduces early sensitivity to nonconvexity, while metric adaptation compensates for LL anisotropy.

4.2 Real-world applications

Data hyper-cleaning. Data hyper-cleaning (Shaban et al., 2019; Franceschi et al., 2017) learns sample weights to mitigate corrupted training labels. Let $\mathcal{D}_{\mathrm{tr}}$ and $\mathcal{D}_{\mathrm{val}}$ denote the corrupted training set and validation set. The problem is

$$\min_{\boldsymbol{x}\in\Delta} \sum_{\xi_i\in\mathcal{D}_{\mathrm{val}}} \ell(\boldsymbol{y}^*(\boldsymbol{x}); \xi_i), \quad \text{s.t.} \quad \boldsymbol{y}^*(\boldsymbol{x}) \in \arg\min_{\boldsymbol{y}} \sum_{\xi_i\in\mathcal{D}_{\mathrm{tr}}} x_i \ell(\boldsymbol{y}; \xi_i) + \frac{\sigma}{2}\|\boldsymbol{y}\|^2,$$

where ξ_i is the i-th sample, ℓ is cross-entropy, $\boldsymbol{x}$ contains sample weights, $\boldsymbol{y}$ contains model parameters, and $\sigma = 10^{-3}$. We use negative entropy for $\boldsymbol{x} \in \Delta$ and $\phi(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y}\|^2$. As shown in Table 1, GaCF achieves the best accuracy and F1 score on both MNIST (LeCun et al., 2002) and FashionMNIST (Xiao et al., 2017), while also requiring the least computational time. Since the sample-weight variable lies on the simplex, this task directly tests whether GaCF can exploit non-Euclidean feasible geometry. The results suggest that the entropy-induced geometry and continuation mechanism help stabilize sample reweighting under label corruption, leading to more reliable validation-guided learning.

Few-shot learning. Few-shot learning (Qiao et al., 2018; Vinyals et al., 2016) aims to adapt a model to new tasks using only a few labeled examples. In the N-way M-shot setting, each task contains M labeled samples from each of N classes. The objective is to learn a shared hyper-representation $\boldsymbol{x}$ while adapting a task-specific classifier $\boldsymbol{y}_j$ for each task j, leading to the BLO formulation

$$\min_{\boldsymbol{x}} \sum_j \ell\big(\boldsymbol{x}, \boldsymbol{y}_j^*(\boldsymbol{x}); \mathcal{D}_{\mathrm{val}}^j\big), \quad \text{s.t.} \quad \boldsymbol{y}^*(\boldsymbol{x}) \in \arg\min_{\boldsymbol{y}} \ell\big(\boldsymbol{x}, \boldsymbol{y}; \mathcal{D}_{\mathrm{tr}}^j\big) + \frac{1}{2}\|\boldsymbol{y}\|^2,$$

where $\mathcal{D}_{\mathrm{val}}^j$ and $\mathcal{D}_{\mathrm{tr}}^j$ denote the validation and training data for the j-th task, and ℓ is the cross-entropy loss. We use standard quadratic potentials for both ψ and ϕ, and evaluate all methods on Omniglot under 5-way 1-shot and 10-way 1-shot settings with at most 150 iterations. As shown in Figure 4 and Table 2, GaCF converges faster and achieves higher test accuracy than competing methods in both settings. This suggests that the continuation scheme is useful for few-shot learning, where optimization is sensitive to limited samples and task variability. By starting from a better-conditioned surrogate and gradually approaching the target nonconvex problem, GaCF stabilizes early updates and supports task adaptation, improving generalization to unseen classes.

Parameter-efficient fine-tuning (PEFT). PEFT (Houlsby et al., 2019; Hu et al., 2022) adapts pretrained language models to downstream tasks by updating only a small subset of task-specific parameters while keeping most pretrained weights fixed. Following DoRA (Liu et al., 2024b), we decompose each pretrained weight into a magnitude component $\boldsymbol{x}$ and a directional component $\boldsymbol{y}$.

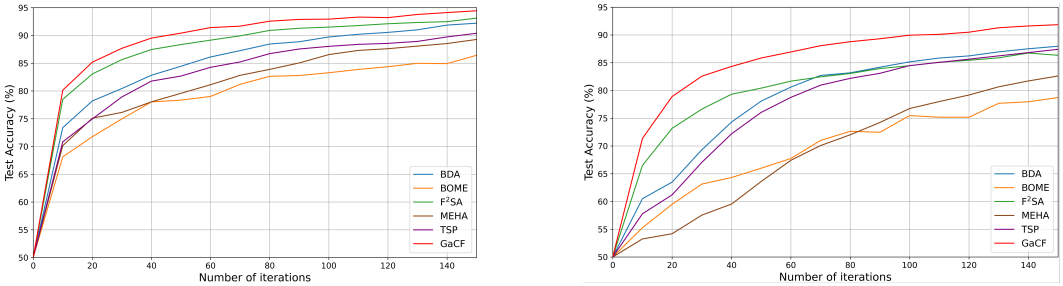


Figure 4: Comparison of various BLO algorithms on Omniglot few-shot learning (left: 5-way 1-shot; right: 10-way 1-shot) for convergence and generalization.

Table 2: Comparison of various BLO algorithms on the Omniglot dataset for few-shot learning. We report mean of 3 runs using different random seeds.

Methods	5-Way			10-Way		
	Val Acc. (%)	Test Acc. (%)	Time (s)	Val Acc. (%)	Test Acc. (%)	Time (s)
BDA	97.86	92.17	508.74	96.74	88.04	568.92
BOME	76.07	86.55	241.93	64.97	78.65	322.17
F^2SA	98.36	93.08	274.88	96.61	86.27	330.04
MEHA	96.71	89.35	**90.46**	93.41	82.76	94.52
TSP	99.51	90.31	111.02	96.08	87.36	122.61
GaCF	**99.72**	**94.47**	91.58	**98.79**	**91.78**	**93.11**

The directional component is parameterized by low-rank increments with rank fixed to 4, while the magnitude component is optimized separately. This leads to the following BLO formulation:

$$\min_{\boldsymbol{x}} \ \ell\big(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}); \mathcal{D}_{\text{val}}\big), \quad \text{s.t.} \quad \boldsymbol{y}^*(\boldsymbol{x}) \in \arg\min_{\boldsymbol{y}} \ \ell\big(\boldsymbol{x}, \boldsymbol{y}; \mathcal{D}_{\text{tr}}\big).$$

We conduct experiments on SST-2 using RoBERTa as the backbone and quadratic potentials. As shown in Figure 5, GaCF outperforms competing methods in test accuracy and F1 score. It reaches competitive performance within fewer iterations and maintains a stable trajectory, indicating improved stability and solution quality. This advantage is attributable to the continuation strategy, which improves early-stage conditioning and enables effective updates before gradually recovering the fine-tuning problem. Therefore, GaCF achieves more stable and effective parameter-efficient adaptation.

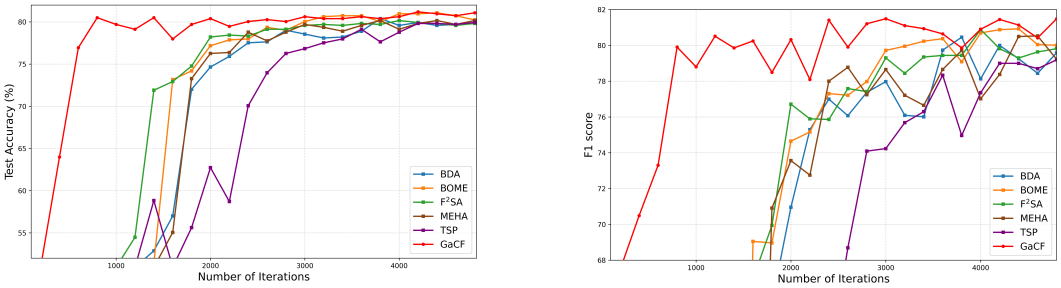


Figure 5: Performance of different BLO methods for PEFT on SST-2.

5 Conclusion

We proposed GaCF, a single-loop geometry-aware continuation framework for constrained nonconvex BLO with potentially nonunique LL solutions. GaCF follows a path from a well-conditioned surrogate to the original nonconvex LL objective and combines Bregman smoothing with mirror-descent updates under a shared problem-adapted geometry. We established a non-asymptotic convergence guarantee under general nonconvex settings. Experiments on synthetic and real-world tasks demonstrate improved convergence efficiency, stability, and predictive performance over existing BLO methods.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, LaTeX formatting, and the examination of mathematical derivations. All AI-assisted suggestions were independently reviewed and verified by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

To support reproducibility, complete proofs of all theoretical results are provided in the appendix. The appendix also specifies the problem settings, data-processing procedures, model architectures, hyperparameters, stopping criteria, and evaluation metrics used in the experiments. All reported results are averaged over three runs with different random seeds unless otherwise stated.

REFERENCES

Eugene L Allgower and Kurt Georg. *Introduction to numerical continuation methods*. SIAM, 2003.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Quentin Bertrand, Quentin Klopfenstein, Mathieu Blondel, Samuel Vaiter, Alexandre Gramfort, and Joseph Salmon. Implicit differentiation of lasso-type models for hyperparameter optimization. In *International Conference on Machine Learning*, pp. 810–821. PMLR, 2020.

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

Jerome Bolte, Pauwels Edouard, and Vaiter Samuel. Automatic differentiation of nonsmooth iterative algorithms. *Advances in Neural Information Processing Systems*, 35:26404–26417, 2022.

Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

Yair Censor and Stavros Andrea Zenios. Proximal minimization algorithm with d-functions. *Journal of Optimization Theory and Applications*, 73(3):451–464, 1992.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems*, 34: 25294–25307, 2021.

John M. Danskin. *The Theory of Max-Min and Its Application to Weapons Allocation Problems*. Springer, 1967.

Daniel M Dunlavy and Dianne P O'Leary. Homotopy optimization methods for global optimization. Technical report, Sandia National Laboratories, 2005.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International conference on machine learning*, pp. 1165–1173. PMLR, 2017.

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

Lucy L Gao, Jane J Ye, Haian Yin, Shangzhi Zeng, and Jin Zhang. Moreau envelope based difference-of-weakly-convex reformulation and algorithm for bilevel programs. *arXiv preprint arXiv:2306.16761*, 2023.

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305, 2016.

Jie Hao, Kaiyi Ji, and Mingrui Liu. Bilevel coreset selection in continual learning: A new formulation and algorithm. *Advances in Neural Information Processing Systems*, 36:51026–51049, 2023.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

Petr Hruby, Timothy Duff, Anton Leykin, and Tomas Pajdla. Learning to solve hard minimal problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5532–5542, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022.

Zixuan Hu, Li Shen, Zhenyi Wang, Baoyuan Wu, Chun Yuan, and Dacheng Tao. Learning to learn from apis: Black-box data-free meta-learning. In *International Conference on Machine Learning*, pp. 13610–13627. PMLR, 2023.

Feihu Huang, Junyi Li, Shangqian Gao, and Heng Huang. Enhanced bilevel optimization via bregman distance. *Advances in Neural Information Processing Systems*, 35:28928–28939, 2022.

Prashant Khanduri, Ioannis Tsaknakis, Yihua Zhang, Jia Liu, Sijia Liu, Jiawei Zhang, and Mingyi Hong. Linearly constrained bilevel optimization: A smoothed implicit gradient approach. In *International Conference on Machine Learning*, pp. 16291–16325. PMLR, 2023.

Krishnateja Killamsetty, Changbin Li, Chen Zhao, Feng Chen, and Rishabh Iyer. A nested bi-level optimization framework for robust few shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7176–7184, 2022.

Minyoung Kim and Timothy Hospedales. A stochastic approach to bi-level optimization for hyper-parameter optimization and meta learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17913–17920, 2025.

Andreas Klingler and Tim Netzer. Homotopy methods for convex optimization. *SIAM Journal on Optimization*, 35(3):1498–1523, 2025.

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

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

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

Xi Lin, Zhiyuan Yang, Xiaoyuan Zhang, and Qingfu Zhang. Continuation path learning for homotopy optimization. In *International Conference on Machine Learning*, pp. 21288–21311. PMLR, 2023.

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

Risheng Liu, Xuan Liu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A value-function-based interior-point method for non-convex bi-level optimization. In *International Conference on Machine Learning*, pp. 6882–6892. PMLR, 2021.

Risheng Liu, Xuan Liu, Shangzhi Zeng, Jin Zhang, and Yixuan Zhang. Value-function-based sequential minimization for bi-level optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15930–15948, 2023a.

Risheng Liu, Pan Mu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A general descent aggregation framework for gradient-based bi-level optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):38–57, 2023b.

Risheng Liu, Zhu Liu, Wei Yao, Shangzhi Zeng, and Jin Zhang. Moreau envelope for nonconvex bi-level optimization: A single-loop and hessian-free solution strategy. *International Conference on Machine Learning*, 2024a.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024b.

Han Lu, Yichen Xie, Xiaokang Yang, and Junchi Yan. Boundary matters: A bi-level active finetuning method. *Advances in Neural Information Processing Systems*, 37:35945–35972, 2024.

Songtao Lu. Tsp: A two-sided smoothed primal-dual method for nonconvex bilevel optimization. In *Forty-second International Conference on Machine Learning*, 2025.

Hossein Mobahi and John Fisher III. A theoretical analysis of optimization by gaussian continuation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Rui Pan, Dylan Zhang, Hanning Zhang, Xingyuan Pan, Minrui Xu, Jipeng Zhang, Renjie Pi, Xiaoyu Wang, and Tong Zhang. Scalebio: Scalable bilevel optimization for llm data reweighting. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31959–31982, 2025.

Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7229–7238, 2018.

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

Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1723–1732. PMLR, 2019.

Han Shen and Tianyi Chen. On penalty-based bilevel gradient descent method. In *International Conference on Machine Learning*, pp. 30992–31015. PMLR, 2023.

Han Shen, Pin-Yu Chen, Payel Das, and Tianyi Chen. Seal: Safety-enhanced aligned llm fine-tuning via bilevel data selection. *International Conference on Learning Representations*, 2025a.

Han Shen, Zhuoran Yang, and Tianyi Chen. Principled penalty-based methods for bilevel reinforcement learning and rlhf. *Journal of Machine Learning Research*, 26(114):1–49, 2025b.

Daouda Sow, Kaiyi Ji, Ziwei Guan, and Yingbin Liang. A primal-dual approach to bilevel optimization with multiple inner minima. *arXiv preprint arXiv:2203.01123*, 2022.

Andrew Starnes, Anton Dereventsov, and Clayton Webster. Gaussian smoothing gradient descent for minimizing functions (gsmoothgd). *arXiv preprint arXiv:2311.00521*, 2023.

Vinzenz Thoma, Barna Pásztor, Andreas Krause, Giorgia Ramponi, and Yifan Hu. Contextual bilevel reinforcement learning for incentive alignment. *Advances in Neural Information Processing Systems*, 37:127369–127435, 2024.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

Quanziang Wang, Renzhen Wang, Yichen Wu, Xixi Jia, Minghao Zhou, and Deyu Meng. Dual-cba: Improving online continual learning via dual continual bias adaptors from a bi-level optimization perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Zhijun Wu. The effective energy transformation scheme as a special continuation approach to global optimization with application to molecular conformation. *SIAM Journal on Optimization*, 6(3): 748–768, 1996.

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

Siqi Zhang and Niao He. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1806.04781*, 2018.

A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

Appendices A.2–A.3 establish reformulation equivalence and inner-problem regularity. Appendices A.4–A.9 develop the tracking, descent, and convergence analysis, while Appendices A.10–A.11 derive the KKT residual bounds and the vanishing-regularization result. Appendix A.12 provides the dataset, model, and parameter configurations for the real-world experiments, and Appendix A.13 reports the sensitivity of GaCF to the potential function, continuation increments, and penalty parameter. Appendix A.14 presents the ablation study.

A.1 NOTATION

Table 3 summarizes the notation. The optimization problems and Algorithm 1 are defined in the main text.

Table 3: Notation used in the convergence analysis.

Notation	Definition or shorthand	Description
$\boldsymbol{z}$, $\mathcal{Z}$	$\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$	Primal variable and feasible set
D_ϕ, D_ψ	$D_\phi(\boldsymbol{u}, \boldsymbol{v}) = \phi(\boldsymbol{u}) - \phi(\boldsymbol{v}) - \langle \nabla\phi(\boldsymbol{v}), \boldsymbol{u} - \boldsymbol{v} \rangle$ D_ψ is defined analogously	Bregman divergences
L_ψ, L_ϕ	Local Lipschitz constants of $\nabla\psi$, $\nabla\phi$	Uniform constants used in the analysis
$\mathcal{N}_{\mathcal{C}}(\boldsymbol{u})$	Normal cone to $\mathcal{C}$ at $\boldsymbol{u}$	Constraint geometry
g_t	$g_t(\boldsymbol{x}, \boldsymbol{y}) = tg(\boldsymbol{x}, \boldsymbol{y}) + (1-t)\phi(\boldsymbol{y})$	Homotopy objective
$H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}$, H_k	$H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}) = g_t(\boldsymbol{x}, \boldsymbol{\theta}) + D_\phi(\boldsymbol{\theta}, \boldsymbol{y})/(2\gamma)$ $H_k = H^{\gamma_k,t_k}_{\boldsymbol{x}_k,\boldsymbol{y}_k}$	Inner objective
$\boldsymbol{\theta}^*_{\gamma,t}$, $\boldsymbol{\theta}^*_k$	$\boldsymbol{\theta}^*_{\gamma,t}(\boldsymbol{x}, \boldsymbol{y}) = \arg\min_{\boldsymbol{\theta}\in\mathcal{Y}} H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta})$ $\boldsymbol{\theta}^*_k = \boldsymbol{\theta}^*_{\gamma_k,t_k}(\boldsymbol{x}_k, \boldsymbol{y}_k)$	Exact inner solutions
v^t_γ	$v^t_\gamma(\boldsymbol{x}, \boldsymbol{y}) = \min_{\boldsymbol{\theta}\in\mathcal{Y}} H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta})$	Smoothed value function
q^t_γ, $\hat{V}^t_\gamma$	$q^t_\gamma = g_t - v^t_\gamma$, $\hat{V}^t_\gamma = q^t_\gamma - \varepsilon_{\mathrm{rel}}$	Gap and relaxed constraint
$\tilde{V}^t_\gamma$	$\tilde{V}^t_\gamma(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) = g_t(\boldsymbol{x}, \boldsymbol{y}) - H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}) - \varepsilon_{\mathrm{rel}}$	Inexact constraint
$\mathcal{E}_k$	$D_\phi(\boldsymbol{\theta}^*_{k-1}, \boldsymbol{\theta}_k)$	Inner tracking error
$\mathscr{A}^{\gamma,t}_{\rho,\omega}$	Augmented Lagrangian in (8)	Primal–dual model
$\lambda^*_{\omega,\gamma,t}$, $\mu^*_{\omega,\gamma,t}$	Optimal dual multiplier and effective multiplier in (6-4)–(6-5)	Multiplier mappings
$\Phi^{\gamma,t}_{\rho,\omega}$	Envelope defined in (6-2)	Dual-maximized objective
$\mathcal{D}_k$	$\lvert\lambda_k - \lambda^*_{\omega,\gamma_k,t_k}(\boldsymbol{z}_k)\rvert^2$	Dual tracking error
$\mathcal{W}_k$	Lyapunov function defined in (9-2)	Descent analysis
$\mathcal{R}_\omega$, $\mathcal{R}_{\mathrm{KKT}}$	Residuals defined in (9-16) and (10-1)	Stationarity and KKT errors

A.2 EQUIVALENCE OF THE EXACT REFORMULATION

The following proposition establishes the equivalence between the Bregman-smoothed reformulation (6) and the relaxed bilevel problem (7).

Proposition 1 (Equivalence of the Exact Smooth Reformulation)**.** *Suppose Assumptions 1 and 2 hold. Let $t = 1$, $\varepsilon_{\mathrm{rel}} = 0$, and $\gamma > 0$ satisfy $\mu_\phi/(2\gamma) - \rho_g > 0$. Then, for every $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{Z}$,*

$$q^1_\gamma(\boldsymbol{x}, \boldsymbol{y}) \le 0 \quad \Longleftrightarrow \quad \boldsymbol{y} \in \widetilde{\mathcal{S}}(\boldsymbol{x}). \tag{2-1}$$

In this case, $\hat{V}^1_\gamma = q^1_\gamma$; hence, problem (6) *in the main text and the relaxed bilevel problem* (7) *have the same feasible set and objective.*

Proof. It suffices to prove (2-1). Let $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{Z}$ and recall the definition of v^1_γ in (5).

Necessity. Suppose $q_\gamma^1(\boldsymbol{x}, \boldsymbol{y}) \leq 0$. Since $\boldsymbol{y} \in \mathcal{Y}$, evaluating the inner objective at $\boldsymbol{\theta} = \boldsymbol{y}$ gives

$$g(\boldsymbol{x}, \boldsymbol{y}) \leq v_\gamma^1(\boldsymbol{x}, \boldsymbol{y}) \leq H_{\boldsymbol{x},\boldsymbol{y}}^{\gamma,1}(\boldsymbol{y}) = g(\boldsymbol{x}, \boldsymbol{y}).$$

Thus, $\boldsymbol{y}$ minimizes $H_{\boldsymbol{x},\boldsymbol{y}}^{\gamma,1}$ over $\mathcal{Y}$. Since $\nabla_{\boldsymbol{\theta}} D_\phi(\boldsymbol{y}, \boldsymbol{y}) = 0$, its first-order optimality condition yields

$$0 \in \nabla_{\boldsymbol{y}} g(\boldsymbol{x}, \boldsymbol{y}) + \mathcal{N}_{\mathcal{Y}}(\boldsymbol{y}).$$

Sufficiency. Conversely, suppose $\boldsymbol{y} \in \widetilde{\mathcal{S}}(\boldsymbol{x})$. By Assumptions 1 and 2, $H_{\boldsymbol{x},\boldsymbol{y}}^{\gamma,1}$ is strongly convex with modulus $\mu_\phi/(2\gamma) - \rho_g > 0$. Since $\nabla_{\boldsymbol{\theta}} D_\phi(\boldsymbol{y}, \boldsymbol{y}) = 0$, lower-level stationarity gives $0 \in \nabla_{\boldsymbol{\theta}} H_{\boldsymbol{x},\boldsymbol{y}}^{\gamma,1}(\boldsymbol{y}) + \mathcal{N}_{\mathcal{Y}}(\boldsymbol{y})$. Hence $\boldsymbol{y}$ is the unique minimizer of the inner problem, and

$$v_\gamma^1(\boldsymbol{x}, \boldsymbol{y}) = H_{\boldsymbol{x},\boldsymbol{y}}^{\gamma,1}(\boldsymbol{y}) = g(\boldsymbol{x}, \boldsymbol{y}).$$

Therefore, $q_\gamma^1(\boldsymbol{x}, \boldsymbol{y}) = 0$, proving the reverse implication in (2-1).

□

If $g(\boldsymbol{x}, \cdot)$ is convex, then $\widetilde{\mathcal{S}}(\boldsymbol{x}) = \mathcal{S}(\boldsymbol{x})$, so Proposition 1 also establishes equivalence with the lower-level optimality constraint.

A.3 Well-Posedness and Regularity of the Smoothed Inner Problem

This section establishes existence and uniqueness of the inner solution, local Lipschitz continuity of the solution mapping, and differentiability of the value function. These properties are used in the subsequent tracking and descent estimates.

Lemma 1 (Well-Posedness of the Smoothed Inner Problem)**.** *Let $t \in [0, 1]$ and $\gamma > 0$, and define*

$$\mu_H(t, \gamma) := \left(1 - t + \frac{1}{2\gamma}\right) \mu_\phi - t\rho_g > 0.$$

Then $H_{\boldsymbol{x},\boldsymbol{y}}^{\gamma,t}(\cdot)$ is $\mu_H(t, \gamma)$-strongly convex on $\mathcal{Y}$ and admits a unique minimizer.

Proof. By definition,

$$\nabla_{\boldsymbol{\theta}} H_{\boldsymbol{x},\boldsymbol{y}}^{\gamma,t}(\boldsymbol{\theta}) = t\nabla_{\boldsymbol{\theta}} g(\boldsymbol{x}, \boldsymbol{\theta}) + \left(1 - t + \frac{1}{2\gamma}\right) \nabla\phi(\boldsymbol{\theta}) - \frac{1}{2\gamma}\nabla\phi(\boldsymbol{y}). \tag{3-1}$$

By Assumptions 1 and 2, for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{Y}$,

$$\begin{aligned}\left\langle \nabla_{\boldsymbol{\theta}} H_{\boldsymbol{x},\boldsymbol{y}}^{\gamma,t}(\boldsymbol{\theta}_1) - \nabla_{\boldsymbol{\theta}} H_{\boldsymbol{x},\boldsymbol{y}}^{\gamma,t}(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\rangle &\geq -t\rho_g \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 + \left(1 - t + \frac{1}{2\gamma}\right) \mu_\phi \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 \\ &= \mu_H(t, \gamma) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2.\end{aligned}$$

Thus, $H_{\boldsymbol{x},\boldsymbol{y}}^{\gamma,t}$ is $\mu_H(t, \gamma)$-strongly convex in $\boldsymbol{\theta}$. Continuity on $\mathcal{Y}$ ensures existence of a minimizer, and strong convexity ensures uniqueness. □

Lemma 2 (Regularity of the Inner Solution Mapping)**.** *Given (t, γ) with $\mu_H(t, \gamma) > 0$, let $L_{g,x\theta}$ and L_ϕ be Lipschitz constants of $\nabla_{\boldsymbol{\theta}} g$ with respect to $\boldsymbol{x}$ and of $\nabla\phi$, respectively. Then*

$$\left\|\boldsymbol{\theta}_{\gamma,t}^*(\boldsymbol{x}_1, \boldsymbol{y}_1) - \boldsymbol{\theta}_{\gamma,t}^*(\boldsymbol{x}_2, \boldsymbol{y}_2)\right\| \leq \frac{tL_{g,x\theta}}{\mu_H(t, \gamma)} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + \frac{L_\phi}{2\gamma\mu_H(t, \gamma)} \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|.$$

Proof. For $i = 1, 2$, set $\boldsymbol{\theta}_i := \boldsymbol{\theta}^*_{\gamma,t}(\boldsymbol{x}_i, \boldsymbol{y}_i)$. The first-order optimality conditions at the two minimizers yield

$$\left\langle \nabla_{\boldsymbol{\theta}} H^{\gamma,t}_{\boldsymbol{x}_1,\boldsymbol{y}_1}(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2 - \boldsymbol{\theta}_1 \right\rangle \geq 0, \qquad \left\langle \nabla_{\boldsymbol{\theta}} H^{\gamma,t}_{\boldsymbol{x}_2,\boldsymbol{y}_2}(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\rangle \geq 0.$$

Adding these inequalities, then adding and subtracting $\nabla_{\boldsymbol{\theta}} H^{\gamma,t}_{\boldsymbol{x}_1,\boldsymbol{y}_1}(\boldsymbol{\theta}_2)$ and rearranging, yields

$$\left\langle \nabla_{\boldsymbol{\theta}} H^{\gamma,t}_{\boldsymbol{x}_1,\boldsymbol{y}_1}(\boldsymbol{\theta}_1) - \nabla_{\boldsymbol{\theta}} H^{\gamma,t}_{\boldsymbol{x}_1,\boldsymbol{y}_1}(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\rangle \leq \left\langle \nabla_{\boldsymbol{\theta}} H^{\gamma,t}_{\boldsymbol{x}_2,\boldsymbol{y}_2}(\boldsymbol{\theta}_2) - \nabla_{\boldsymbol{\theta}} H^{\gamma,t}_{\boldsymbol{x}_1,\boldsymbol{y}_1}(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\rangle.$$

By Lemma 1, the left-hand side is bounded below by $\mu_H(t,\gamma)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2$.

For fixed $\boldsymbol{\theta}_2$, the definition of $H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}$ and the Lipschitz bounds in Assumptions 1 and 2 give

$$\begin{aligned}
&\left\| \nabla_{\boldsymbol{\theta}} H^{\gamma,t}_{\boldsymbol{x}_2,\boldsymbol{y}_2}(\boldsymbol{\theta}_2) - \nabla_{\boldsymbol{\theta}} H^{\gamma,t}_{\boldsymbol{x}_1,\boldsymbol{y}_1}(\boldsymbol{\theta}_2) \right\| \\
&\quad \leq t\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{x}_2, \boldsymbol{\theta}_2) - \nabla_{\boldsymbol{\theta}} g(\boldsymbol{x}_1, \boldsymbol{\theta}_2)\| + \frac{1}{2\gamma}\|\nabla\phi(\boldsymbol{y}_2) - \nabla\phi(\boldsymbol{y}_1)\| \\
&\quad \leq tL_{g,x\theta}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + \frac{L_\phi}{2\gamma}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|.
\end{aligned}$$

The Cauchy-Schwarz inequality therefore yields

$$\begin{aligned}
\mu_H(t,\gamma)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 \leq t\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{x}_1, \boldsymbol{\theta}_2) - \nabla_{\boldsymbol{\theta}} g(\boldsymbol{x}_2, \boldsymbol{\theta}_2)\|\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \\
+ \frac{1}{2\gamma}\|\nabla\phi(\boldsymbol{y}_1) - \nabla\phi(\boldsymbol{y}_2)\|\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|.
\end{aligned}$$

Combining these bounds and dividing by $\mu_H(t,\gamma)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$ when $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ yields the estimate below; the case $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$ holds identically.

$$\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \leq \frac{tL_{g,x\theta}}{\mu_H(t,\gamma)}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + \frac{L_\phi}{2\gamma\mu_H(t,\gamma)}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|.$$

□

Lemma 3 (Differentiability of the Value Function)**.** *Under the conditions of Lemmas 1–2, v^t_γ is continuously differentiable on $\mathcal{Z}$, with partial gradients*

$$\nabla_{\boldsymbol{x}} v^t_\gamma(\boldsymbol{x},\boldsymbol{y}) = t\nabla_{\boldsymbol{x}} g\left(\boldsymbol{x}, \boldsymbol{\theta}^*_{\gamma,t}(\boldsymbol{x},\boldsymbol{y})\right), \qquad \nabla_{\boldsymbol{y}} v^t_\gamma(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2\gamma}\nabla_{\boldsymbol{y}} D_\phi\left(\boldsymbol{\theta}^*_{\gamma,t}(\boldsymbol{x},\boldsymbol{y}), \boldsymbol{y}\right). \tag{3-2}$$

Proof. Lemma 1 ensures uniqueness of the inner minimizer. Together with Lemma 2 and Assumptions 1–2, Danskin's theorem yields the partial gradients by differentiating $H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}$ with $\boldsymbol{\theta}$ fixed:

$$\nabla_{\boldsymbol{x}} H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}) = t\nabla_{\boldsymbol{x}} g(\boldsymbol{x},\boldsymbol{\theta}), \qquad \nabla_{\boldsymbol{y}} H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}) = \frac{1}{2\gamma}\nabla_{\boldsymbol{y}} D_\phi(\boldsymbol{\theta},\boldsymbol{y}).$$

Evaluating these expressions at $\boldsymbol{\theta} = \boldsymbol{\theta}^*_{\gamma,t}(\boldsymbol{x},\boldsymbol{y})$ gives the stated partial gradients. □

By Lemmas 2–3 and Assumptions 1–2, $\hat{V}^t_\gamma$ is continuously differentiable with a locally Lipschitz gradient for every fixed (t,γ) satisfying $\mu_H(t,\gamma) > 0$. Consequently, $\hat{V}^1_{\underline{\gamma}}$ and its gradient are bounded on $\mathcal{Z}$, and $\hat{V}^1_{\underline{\gamma}}$ is Lipschitz continuous there.

A.4 One-Step Inner Tracking

We first establish relative strong convexity and relative smoothness of the inner objective. These properties yield a one-step contraction estimate, from which we derive a recursion for the inner tracking error.

Lemma 4 (Relative Strong Convexity and Relative Smoothness)**.** *Let $L_{g,\theta}$ be a uniform Lipschitz constant of $\nabla_{\boldsymbol{\theta}} g(\boldsymbol{x}, \cdot)$ on $\mathcal{Y}$. Define*

$$\mu_{\mathrm{rel}}(t,\gamma) := 1 - t + \frac{1}{2\gamma} - \frac{t\rho_g}{\mu_\phi}, \qquad L_{\mathrm{rel}}(t,\gamma) := 1 - t + \frac{1}{2\gamma} + \frac{tL_{g,\theta}}{\mu_\phi}.$$

If $\mu_{\mathrm{rel}}(t,\gamma) > 0$, then, for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{Y}$,

$$\begin{aligned} \mu_{\mathrm{rel}}(t,\gamma) D_\phi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &\le H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}_1) - H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}_2) - \left\langle \nabla_{\boldsymbol{\theta}} H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\rangle \\ &\le L_{\mathrm{rel}}(t,\gamma) D_\phi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2). \end{aligned} \tag{4-1}$$

Proof. In the first-order remainder of $H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}$, the affine term and all terms independent of $\boldsymbol{\theta}$ cancel. Hence,

$$\begin{aligned} &H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}_1) - H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}_2) - \left\langle \nabla_{\boldsymbol{\theta}} H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\rangle \\ &\quad = t\Big[g(\boldsymbol{x}, \boldsymbol{\theta}_1) - g(\boldsymbol{x}, \boldsymbol{\theta}_2) - \langle \nabla_{\boldsymbol{\theta}} g(\boldsymbol{x}, \boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle \Big] + \left(1 - t + \frac{1}{2\gamma}\right) D_\phi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2). \end{aligned} \tag{4-2}$$

The weak convexity and gradient Lipschitz continuity in Assumption 1 give, respectively,

$$-\frac{t\rho_g}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 \le tg(\boldsymbol{x}, \boldsymbol{\theta}_1) - tg(\boldsymbol{x}, \boldsymbol{\theta}_2) - t\langle \nabla_{\boldsymbol{\theta}} g(\boldsymbol{x}, \boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle$$

and

$$tg(\boldsymbol{x}, \boldsymbol{\theta}_1) - tg(\boldsymbol{x}, \boldsymbol{\theta}_2) - t\langle \nabla_{\boldsymbol{\theta}} g(\boldsymbol{x}, \boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle \le \frac{tL_{g,\theta}}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2.$$

Strong convexity of ϕ in Assumption 2 further implies

$$-\frac{t\rho_g}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 \ge -\frac{t\rho_g}{\mu_\phi} D_\phi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \qquad \frac{tL_{g,\theta}}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 \le \frac{tL_{g,\theta}}{\mu_\phi} D_\phi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2).$$

Substituting these bounds into (4-2) yields

$$\begin{aligned} \mu_{\mathrm{rel}}(t,\gamma) D_\phi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &\le H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}_1) - H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}_2) - \left\langle \nabla_{\boldsymbol{\theta}} H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\rangle \\ &\le L_{\mathrm{rel}}(t,\gamma) D_\phi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2). \end{aligned}$$

□

Lemma 5 (One-Step Contraction of the Inner Update)**.** *Under the conditions of Lemma 4, let $\boldsymbol{\theta}_{k+1}$ be generated by* (10) *with step size η. Let*

$$\underline{\mu}_{\mathrm{rel}} := \inf_k \mu_{\mathrm{rel}}(t_k, \gamma_k) > 0, \qquad \overline{L}_{\mathrm{rel}} := \sup_k L_{\mathrm{rel}}(t_k, \gamma_k) < \infty.$$

If $0 < \eta \le 1/\overline{L}_{\mathrm{rel}}$ and $\tau := \eta\underline{\mu}_{\mathrm{rel}} \in (0,1)$, then

$$D_\phi(\boldsymbol{\theta}^*_k, \boldsymbol{\theta}_{k+1}) \le (1 - \tau) D_\phi(\boldsymbol{\theta}^*_k, \boldsymbol{\theta}_k). \tag{4-3}$$

Proof. Write $H_k := H_{\boldsymbol{x}_k,\boldsymbol{y}_k}^{\gamma_k,t_k}$. The first-order optimality condition for (10) gives, for every $\boldsymbol{\theta} \in \mathcal{Y}$,

$$\langle \eta \nabla H_k(\boldsymbol{\theta}_k) + \nabla\phi(\boldsymbol{\theta}_{k+1}) - \nabla\phi(\boldsymbol{\theta}_k), \boldsymbol{\theta} - \boldsymbol{\theta}_{k+1} \rangle \geq 0. \tag{4-4}$$

Set $\boldsymbol{\theta} = \boldsymbol{\theta}_k^*$ and invoke the three-point identity

$$\langle \nabla\phi(\boldsymbol{b}) - \nabla\phi(\boldsymbol{c}), \boldsymbol{a} - \boldsymbol{b} \rangle = D_\phi(\boldsymbol{a}, \boldsymbol{c}) - D_\phi(\boldsymbol{a}, \boldsymbol{b}) - D_\phi(\boldsymbol{b}, \boldsymbol{c}), \tag{4-5}$$

Substituting (4-5) into (4-4) with $\boldsymbol{a} = \boldsymbol{\theta}_k^*$, $\boldsymbol{b} = \boldsymbol{\theta}_{k+1}$, and $\boldsymbol{c} = \boldsymbol{\theta}_k$ gives

$$D_\phi(\boldsymbol{\theta}_k^*, \boldsymbol{\theta}_{k+1}) \leq D_\phi(\boldsymbol{\theta}_k^*, \boldsymbol{\theta}_k) - D_\phi(\boldsymbol{\theta}_{k+1}, \boldsymbol{\theta}_k) + \eta \left\langle \nabla H_k(\boldsymbol{\theta}_k), \boldsymbol{\theta}_k^* - \boldsymbol{\theta}_{k+1} \right\rangle. \tag{4-6}$$

The relative bounds (4-1) in Lemma 4 yield

$$\begin{aligned} H_k(\boldsymbol{\theta}_k^*) &\geq H_k(\boldsymbol{\theta}_k) + \langle \nabla H_k(\boldsymbol{\theta}_k), \boldsymbol{\theta}_k^* - \boldsymbol{\theta}_k \rangle + \underline{\mu}_{\mathrm{rel}} D_\phi(\boldsymbol{\theta}_k^*, \boldsymbol{\theta}_k), \\ H_k(\boldsymbol{\theta}_{k+1}) &\leq H_k(\boldsymbol{\theta}_k) + \langle \nabla H_k(\boldsymbol{\theta}_k), \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k \rangle + \overline{L}_{\mathrm{rel}} D_\phi(\boldsymbol{\theta}_{k+1}, \boldsymbol{\theta}_k). \end{aligned} \tag{4-7}$$

Since $\boldsymbol{\theta}_k^*$ globally minimizes H_k, $H_k(\boldsymbol{\theta}_k^*) \leq H_k(\boldsymbol{\theta}_{k+1})$. Combining this inequality with (4-7) and canceling $H_k(\boldsymbol{\theta}_k)$ yields

$$\langle \nabla H_k(\boldsymbol{\theta}_k), \boldsymbol{\theta}_k^* - \boldsymbol{\theta}_{k+1} \rangle \leq -\underline{\mu}_{\mathrm{rel}} D_\phi(\boldsymbol{\theta}_k^*, \boldsymbol{\theta}_k) + \overline{L}_{\mathrm{rel}} D_\phi(\boldsymbol{\theta}_{k+1}, \boldsymbol{\theta}_k). \tag{4-8}$$

Substituting (4-8) into (4-6) yields

$$D_\phi(\boldsymbol{\theta}_k^*, \boldsymbol{\theta}_{k+1}) \leq (1 - \eta\underline{\mu}_{\mathrm{rel}}) D_\phi(\boldsymbol{\theta}_k^*, \boldsymbol{\theta}_k) - (1 - \eta\overline{L}_{\mathrm{rel}}) D_\phi(\boldsymbol{\theta}_{k+1}, \boldsymbol{\theta}_k). \tag{4-9}$$

Since $\eta \leq 1/\overline{L}_{\mathrm{rel}}$, the last term in (4-9) is nonpositive. Omitting this term and using $\tau = \eta\underline{\mu}_{\mathrm{rel}}$ gives

$$D_\phi(\boldsymbol{\theta}_k^*, \boldsymbol{\theta}_{k+1}) \leq (1 - \tau) D_\phi(\boldsymbol{\theta}_k^*, \boldsymbol{\theta}_k),$$

□

Lemma 6 (Inner Tracking Error)**.** *Under the conditions of Lemma 5, there exist $C_x, C_y, C_t, C_\gamma > 0$ such that*

$$\begin{aligned} \mathcal{E}_{k+1} \leq \left(1 - \frac{\tau}{2}\right) \mathcal{E}_k + C_x \|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\|^2 + C_y \|\boldsymbol{y}_k - \boldsymbol{y}_{k-1}\|^2 \\ + C_t |t_k - t_{k-1}|^2 + C_\gamma |\gamma_k - \gamma_{k-1}|^2. \end{aligned} \tag{4-10}$$

Proof. The contraction estimate (4-3) in Lemma 5 gives

$$\mathcal{E}_{k+1} \leq (1 - \tau) D_\phi(\boldsymbol{\theta}_k^*, \boldsymbol{\theta}_k). \tag{4-11}$$

The three-point identity (4-5) gives

$$D_\phi(\boldsymbol{\theta}_k^*, \boldsymbol{\theta}_k) = D_\phi(\boldsymbol{\theta}_{k-1}^*, \boldsymbol{\theta}_k) + D_\phi(\boldsymbol{\theta}_k^*, \boldsymbol{\theta}_{k-1}^*) + \left\langle \nabla\phi(\boldsymbol{\theta}_{k-1}^*) - \nabla\phi(\boldsymbol{\theta}_k), \boldsymbol{\theta}_k^* - \boldsymbol{\theta}_{k-1}^* \right\rangle. \tag{4-12}$$

Lipschitz continuity of $\nabla\phi$ on $\mathcal{Y}$, as ensured by Assumption 2, implies

$$D_\phi(\boldsymbol{\theta}_k^*, \boldsymbol{\theta}_{k-1}^*) \leq \frac{L_\phi}{2} \|\boldsymbol{\theta}_k^* - \boldsymbol{\theta}_{k-1}^*\|^2.$$

By the Cauchy-Schwarz and Young inequalities and local Lipschitz continuity of $\nabla\phi$,

$$\begin{aligned}\left|\left\langle \nabla\phi(\boldsymbol{\theta}^*_{k-1}) - \nabla\phi(\boldsymbol{\theta}_k), \boldsymbol{\theta}^*_k - \boldsymbol{\theta}^*_{k-1}\right\rangle\right| &\le L_\phi\sqrt{\frac{2}{\mu_\phi}D_\phi(\boldsymbol{\theta}^*_{k-1},\boldsymbol{\theta}_k)}\|\boldsymbol{\theta}^*_k - \boldsymbol{\theta}^*_{k-1}\| \\ &\le \frac{\tau}{2}D_\phi(\boldsymbol{\theta}^*_{k-1},\boldsymbol{\theta}_k) + \frac{L_\phi^2}{\tau\mu_\phi}\|\boldsymbol{\theta}^*_k - \boldsymbol{\theta}^*_{k-1}\|^2.\end{aligned} \tag{4-13}$$

Combining (4-12)–(4-13) gives

$$D_\phi(\boldsymbol{\theta}^*_k,\boldsymbol{\theta}_k) \le \left(1+\frac{\tau}{2}\right)\mathcal{E}_k + C_\tau\|\boldsymbol{\theta}^*_k - \boldsymbol{\theta}^*_{k-1}\|^2. \tag{4-14}$$

where $C_\tau := L_\phi/2 + L_\phi^2/(\tau\mu_\phi)$. Also, $(1-\tau)(1+\tau/2) = 1-\tau/2-\tau^2/2 \le 1-\tau/2$.

The first-order optimality conditions for the exact inner solutions at iterations k and $k-1$ yield

$$\left\langle \nabla_{\boldsymbol{\theta}} H^{\gamma_k,t_k}_{\boldsymbol{x}_k,\boldsymbol{y}_k}(\boldsymbol{\theta}^*_k), \boldsymbol{\theta}^*_{k-1} - \boldsymbol{\theta}^*_k\right\rangle \ge 0, \qquad \left\langle \nabla_{\boldsymbol{\theta}} H^{\gamma_{k-1},t_{k-1}}_{\boldsymbol{x}_{k-1},\boldsymbol{y}_{k-1}}(\boldsymbol{\theta}^*_{k-1}), \boldsymbol{\theta}^*_k - \boldsymbol{\theta}^*_{k-1}\right\rangle \ge 0.$$

Adding these inequalities, then adding and subtracting $\nabla_{\boldsymbol{\theta}} H^{\gamma_k,t_k}_{\boldsymbol{x}_k,\boldsymbol{y}_k}(\boldsymbol{\theta}^*_{k-1})$ and rearranging, yields

$$\begin{aligned}&\left\langle \nabla_{\boldsymbol{\theta}} H^{\gamma_k,t_k}_{\boldsymbol{x}_k,\boldsymbol{y}_k}(\boldsymbol{\theta}^*_k) - \nabla_{\boldsymbol{\theta}} H^{\gamma_k,t_k}_{\boldsymbol{x}_k,\boldsymbol{y}_k}(\boldsymbol{\theta}^*_{k-1}), \boldsymbol{\theta}^*_k - \boldsymbol{\theta}^*_{k-1}\right\rangle \\ &\le \left\langle \nabla_{\boldsymbol{\theta}} H^{\gamma_{k-1},t_{k-1}}_{\boldsymbol{x}_{k-1},\boldsymbol{y}_{k-1}}(\boldsymbol{\theta}^*_{k-1}) - \nabla_{\boldsymbol{\theta}} H^{\gamma_k,t_k}_{\boldsymbol{x}_k,\boldsymbol{y}_k}(\boldsymbol{\theta}^*_{k-1}), \boldsymbol{\theta}^*_k - \boldsymbol{\theta}^*_{k-1}\right\rangle.\end{aligned}$$

Let $\underline{\mu}_H := \inf_k \mu_H(t_k,\gamma_k) > 0$. Strong monotonicity from Lemma 1 and the Cauchy–Schwarz inequality give

$$\underline{\mu}_H\|\boldsymbol{\theta}^*_k - \boldsymbol{\theta}^*_{k-1}\|^2 \le \left\|\nabla_{\boldsymbol{\theta}} H^{\gamma_k,t_k}_{\boldsymbol{x}_k,\boldsymbol{y}_k}(\boldsymbol{\theta}^*_{k-1}) - \nabla_{\boldsymbol{\theta}} H^{\gamma_{k-1},t_{k-1}}_{\boldsymbol{x}_{k-1},\boldsymbol{y}_{k-1}}(\boldsymbol{\theta}^*_{k-1})\right\| \|\boldsymbol{\theta}^*_k - \boldsymbol{\theta}^*_{k-1}\|.$$

Expanding $\nabla_{\boldsymbol{\theta}} H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}$ and applying Assumptions 1–2 with respect to $\boldsymbol{x}$, $\boldsymbol{y}$, t, and γ, we obtain constants $L_{\theta,x}, L_{\theta,y}, L_{\theta,t}, L_{\theta,\gamma} > 0$ such that

$$\begin{aligned}\|\boldsymbol{\theta}^*_k - \boldsymbol{\theta}^*_{k-1}\| \le L_{\theta,x}\|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\| + L_{\theta,y}\|\boldsymbol{y}_k - \boldsymbol{y}_{k-1}\| \\ + L_{\theta,t}|t_k - t_{k-1}| + L_{\theta,\gamma}|\gamma_k - \gamma_{k-1}|.\end{aligned} \tag{4-15}$$

Here one may take $L_{\theta,x} := (\sup_k t_k L_{g,x\theta})/\underline{\mu}_H$ and $L_{\theta,y} := L_\phi/(2\underline{\gamma}\,\underline{\mu}_H)$; $L_{\theta,t}$ and $L_{\theta,\gamma}$ are obtained by dividing the corresponding local Lipschitz constants in t and γ by $\underline{\mu}_H$.

Since $\gamma_k \ge \underline{\gamma} > 0$,

$$\left|\frac{1}{2\gamma_k} - \frac{1}{2\gamma_{k-1}}\right| \le \frac{|\gamma_k - \gamma_{k-1}|}{2\underline{\gamma}^2}.$$

Applying $(a_1+a_2+a_3+a_4)^2 \le 4\sum_{i=1}^4 a_i^2$ and setting

$$\begin{aligned}C_x &:= 4(1-\tau)C_\tau L_{\theta,x}^2, & C_y &:= 4(1-\tau)C_\tau L_{\theta,y}^2, \\ C_t &:= 4(1-\tau)C_\tau L_{\theta,t}^2, & C_\gamma &:= 4(1-\tau)C_\tau L_{\theta,\gamma}^2.\end{aligned} \tag{4-16}$$

Combining (4-11), (4-14), and (4-15), with the constants in (4-16), yields

$$\begin{aligned}\mathcal{E}_{k+1} \le \left(1-\frac{\tau}{2}\right)\mathcal{E}_k + C_x\|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\|^2 + C_y\|\boldsymbol{y}_k - \boldsymbol{y}_{k-1}\|^2 \\ + C_t|t_k - t_{k-1}|^2 + C_\gamma|\gamma_k - \gamma_{k-1}|^2.\end{aligned}$$

□

A.5 Function-Value and Gradient Errors of the Inner Approximation

We next bound the errors in the constraint value and its primal gradient in terms of the inner Bregman error.

Lemma 7 (Inner Approximation Error)**.** *Suppose Assumptions 1 and 2 hold and $\mu_H(t,\gamma) > 0$. Let $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*_{\gamma,t}(\boldsymbol{z})$. For fixed (t,γ), there exist constants $C_V, C_{\nabla V} > 0$, uniform over $\mathcal{Z} \times \mathcal{Y}$, such that*

$$\left|\tilde{V}^t_\gamma(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) - \hat{V}^t_\gamma(\boldsymbol{x},\boldsymbol{y})\right|^2 \le C_V D_\phi(\boldsymbol{\theta}^*,\boldsymbol{\theta}),$$

and

$$\left\|\nabla_{(\boldsymbol{x},\boldsymbol{y})}\tilde{V}^t_\gamma(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) - \nabla_{(\boldsymbol{x},\boldsymbol{y})}\hat{V}^t_\gamma(\boldsymbol{x},\boldsymbol{y})\right\|^2 \le C_{\nabla V} D_\phi(\boldsymbol{\theta}^*,\boldsymbol{\theta}).$$

Proof. Assumptions 1–2 imply that $\nabla_{\boldsymbol{\theta}} H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}$ is bounded on $\mathcal{Z} \times \mathcal{Y}$ by a constant $M_H < \infty$. By the definition of v^t_γ, the mean-value theorem, and the strong convexity of ϕ,

$$|\tilde{V}^t_\gamma - \hat{V}^t_\gamma| = \left|H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}) - H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}^*)\right| \le M_H\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \le M_H\sqrt{\frac{2}{\mu_\phi} D_\phi(\boldsymbol{\theta}^*,\boldsymbol{\theta})}.$$

Thus, with $C_V := 2M_H^2/\mu_\phi$,

$$\left|\tilde{V}^t_\gamma(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) - \hat{V}^t_\gamma(\boldsymbol{x},\boldsymbol{y})\right|^2 \le C_V D_\phi(\boldsymbol{\theta}^*,\boldsymbol{\theta}).$$

For the gradient error, (3-2) in Lemma 3, equivalently (9), gives

$$\nabla_{(\boldsymbol{x},\boldsymbol{y})}\tilde{V}^t_\gamma(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) - \nabla_{(\boldsymbol{x},\boldsymbol{y})}\hat{V}^t_\gamma(\boldsymbol{x},\boldsymbol{y}) = -\nabla_{(\boldsymbol{x},\boldsymbol{y})}H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}) + \nabla_{(\boldsymbol{x},\boldsymbol{y})}H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}^*). \tag{5-1}$$

By Assumptions 1–2, the partial gradient $\nabla_{(\boldsymbol{x},\boldsymbol{y})}H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}$ is Lipschitz continuous with respect to $\boldsymbol{\theta}$ on $\mathcal{Z} \times \mathcal{Y}$. Let $L_{H,z\theta}$ denote a corresponding constant. Then

$$\left\|\nabla_{(\boldsymbol{x},\boldsymbol{y})}H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}) - \nabla_{(\boldsymbol{x},\boldsymbol{y})}H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}^*)\right\| \le L_{H,z\theta}\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|. \tag{5-2}$$

Set $C_{\nabla V} := 2L^2_{H,z\theta}/\mu_\phi$. Combining (5-1)–(5-2) with strong convexity of ϕ yields

$$\left\|\nabla_{(\boldsymbol{x},\boldsymbol{y})}\tilde{V}^t_\gamma(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) - \nabla_{(\boldsymbol{x},\boldsymbol{y})}\hat{V}^t_\gamma(\boldsymbol{x},\boldsymbol{y})\right\|^2 \le C_{\nabla V} D_\phi(\boldsymbol{\theta}^*,\boldsymbol{\theta}).$$

□

A.6 Regularity of the Envelope

This section establishes the regularity of the envelope and its maximizing multiplier, including their dependence on the homotopy parameters.

For $t \in [0,1]$, $\gamma \in [\underline{\gamma}, \gamma_0]$, and $\rho, \omega, \Lambda > 0$, recall the augmented Lagrangian (8):

$$\mathscr{A}^{\gamma,t}_{\rho,\omega}(\boldsymbol{z},\lambda) := f(\boldsymbol{z}) + \frac{[\lambda + \rho\hat{V}^t_\gamma(\boldsymbol{z})]^2_+ - \lambda^2}{2\rho} - \frac{\omega}{2}\lambda^2, \qquad 0 \le \lambda \le \Lambda. \tag{6-1}$$

where $[a]_+ := \max\{0, a\}$. The optimal multiplier mapping and the envelope are

$$\lambda^*_{\omega,\gamma,t}(\boldsymbol{z}) := \arg\max_{0\le\lambda\le\Lambda} \mathscr{A}^{\gamma,t}_{\rho,\omega}(\boldsymbol{z},\lambda), \qquad \Phi^{\gamma,t}_{\rho,\omega}(\boldsymbol{z}) := \max_{0\le\lambda\le\Lambda} \mathscr{A}^{\gamma,t}_{\rho,\omega}(\boldsymbol{z},\lambda). \tag{6-2}$$

Along the iterates, write

$$\Phi_k := \Phi_{\rho,\omega}^{\gamma_k,t_k}, \qquad \lambda_{\omega,k}^* := \lambda_{\omega,\gamma_k,t_k}^*. \tag{6-3}$$

We use the shorthand $\Phi_{\rho,\omega} := \Phi_{\rho,\omega}^{\underline{\gamma},1}$ and $\lambda_\omega^* := \lambda_{\omega,\underline{\gamma},1}^*$.

Lemma 8 (Basic Properties of the Envelope)**.** *Suppose Assumptions 1 and 2 hold. Let $t \in [0,1]$, $\gamma \in [\underline{\gamma}, \gamma_0]$, and $\mu_H(t,\gamma) > 0$. For every $\boldsymbol{z} \in \mathcal{Z}$, the function $\mathscr{A}_{\rho,\omega}^{\gamma,t}(\boldsymbol{z},\cdot)$ is ω-strongly concave on $[0,\Lambda]$, with unique maximizer*

$$\lambda_{\omega,\gamma,t}^*(\boldsymbol{z}) = \Pi_{[0,\Lambda]}\left(\frac{[\hat{V}_\gamma^t(\boldsymbol{z})]_+}{\omega}\right). \tag{6-4}$$

Moreover, $\Phi_{\rho,\omega}^{\gamma,t}$ is continuously differentiable, with

$$\nabla\Phi_{\rho,\omega}^{\gamma,t}(\boldsymbol{z}) = \nabla f(\boldsymbol{z}) + \mu_{\omega,\gamma,t}^*(\boldsymbol{z})\nabla\hat{V}_\gamma^t(\boldsymbol{z}), \tag{6-5}$$

where $\mu_{\omega,\gamma,t}^*(\boldsymbol{z}) := [\lambda_{\omega,\gamma,t}^*(\boldsymbol{z}) + \rho\hat{V}_\gamma^t(\boldsymbol{z})]_+$. The envelope also satisfies

$$\Phi_{\rho,\omega}^{\gamma,t}(\boldsymbol{z}) \geq f(\boldsymbol{z}). \tag{6-6}$$

Proof. For any $(\boldsymbol{z},t,\gamma)$, continuous differentiability of $s \mapsto [s]_+^2$ implies that $\mathscr{A}_{\rho,\omega}^{\gamma,t}(\boldsymbol{z},\cdot)$ is continuously differentiable, with

$$\nabla_\lambda \mathscr{A}_{\rho,\omega}^{\gamma,t} = \begin{cases} \hat{V}_\gamma^t(\boldsymbol{z}) - \omega\lambda, & \lambda + \rho\hat{V}_\gamma^t(\boldsymbol{z}) \geq 0, \\ -(\rho^{-1}+\omega)\lambda, & \lambda + \rho\hat{V}_\gamma^t(\boldsymbol{z}) < 0. \end{cases}$$

The two expressions for the derivative coincide at $\lambda = -\rho\hat{V}_\gamma^t(\boldsymbol{z})$. Their slopes are $-\omega$ and $-(\rho^{-1}+\omega)$, respectively. Therefore, for all $\lambda_1, \lambda_2 \in [0,\Lambda]$,

$$\left(\nabla_\lambda \mathscr{A}_{\rho,\omega}^{\gamma,t}(\boldsymbol{z},\lambda_1) - \nabla_\lambda \mathscr{A}_{\rho,\omega}^{\gamma,t}(\boldsymbol{z},\lambda_2)\right)(\lambda_1 - \lambda_2) \leq -\omega|\lambda_1 - \lambda_2|^2.$$

Hence $\mathscr{A}_{\rho,\omega}^{\gamma,t}(\boldsymbol{z},\cdot)$ is ω-strongly concave and has a unique maximizer on the compact interval $[0,\Lambda]$.

If $\hat{V}_\gamma^t(\boldsymbol{z}) \leq 0$, then $\nabla_\lambda \mathscr{A}_{\rho,\omega}^{\gamma,t}(\boldsymbol{z},\lambda) \leq 0$ on $[0,\Lambda]$, and the unique maximizer is $\lambda = 0$. If $\hat{V}_\gamma^t(\boldsymbol{z}) > 0$, the derivative reduces to $\hat{V}_\gamma^t(\boldsymbol{z}) - \omega\lambda$, so the unique maximizer is $\Pi_{[0,\Lambda]}(\hat{V}_\gamma^t(\boldsymbol{z})/\omega)$. This proves (6-4).

Lemma 3 and differentiability of $s \mapsto [s]_+^2$ imply that $\mathscr{A}_{\rho,\omega}^{\gamma,t}$ is continuously differentiable in $\boldsymbol{z}$. Since the maximizer in (6-4) is unique, Danskin's theorem gives

$$\nabla\Phi_{\rho,\omega}^{\gamma,t}(\boldsymbol{z}) = \nabla_{\boldsymbol{z}}\mathscr{A}_{\rho,\omega}^{\gamma,t}(\boldsymbol{z},\lambda_{\omega,\gamma,t}^*(\boldsymbol{z})) = \nabla f(\boldsymbol{z}) + [\lambda_{\omega,\gamma,t}^*(\boldsymbol{z}) + \rho\hat{V}_\gamma^t(\boldsymbol{z})]_+\nabla\hat{V}_\gamma^t(\boldsymbol{z}),$$

which is (6-5). Evaluating the maximization problem in (6-2) at $\lambda = 0$ gives

$$\Phi_{\rho,\omega}^{\gamma,t}(\boldsymbol{z}) \geq f(\boldsymbol{z}) + \frac{\rho}{2}[\hat{V}_\gamma^t(\boldsymbol{z})]_+^2 \geq f(\boldsymbol{z}),$$

and proves (6-6). □

Lemma 9 (Uniform Regularity with Respect to the Homotopy Parameters)**.** *Suppose Assumptions 1 and 2 hold. Let (t,γ) range over a compact subset of $[0,1] \times [\underline{\gamma}, \gamma_0]$ on which $\mu_H(t,\gamma)$ is bounded away from zero. Then there exist $L_{V,z}, L_{V,p}, L_{\nabla V,z}, L_{\nabla V,p} > 0$ such that, for any $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{Z}$ and any $(t_1,\gamma_1), (t_2,\gamma_2)$ in the parameter set,*

$$|\hat{V}_{\gamma_1}^{t_1}(\boldsymbol{z}_1) - \hat{V}_{\gamma_2}^{t_2}(\boldsymbol{z}_2)| \le L_{V,z}\|\boldsymbol{z}_1 - \boldsymbol{z}_2\| + L_{V,p}\big(|t_1 - t_2| + |\gamma_1 - \gamma_2|\big), \tag{6-7}$$

$$\|\nabla\hat{V}_{\gamma_1}^{t_1}(\boldsymbol{z}_1) - \nabla\hat{V}_{\gamma_2}^{t_2}(\boldsymbol{z}_2)\| \le L_{\nabla V,z}\|\boldsymbol{z}_1 - \boldsymbol{z}_2\| + L_{\nabla V,p}\big(|t_1 - t_2| + |\gamma_1 - \gamma_2|\big). \tag{6-8}$$

Proof. Let $\boldsymbol{\theta}_i^* := \boldsymbol{\theta}_{\gamma_i,t_i}^*(\boldsymbol{z}_i)$ for $i = 1, 2$, and let $\underline{\mu}_H > 0$ be a uniform lower bound for $\mu_H(t, \gamma)$ on the parameter set.

By definition of $H_{\boldsymbol{x},\boldsymbol{y}}^{\gamma,t}$,

$$\nabla_{\boldsymbol{\theta}} H_{\boldsymbol{x},\boldsymbol{y}}^{\gamma,t}(\boldsymbol{\theta}) = t\nabla_{\boldsymbol{\theta}} g(\boldsymbol{x}, \boldsymbol{\theta}) + \left(1 - t + \frac{1}{2\gamma}\right)\nabla\phi(\boldsymbol{\theta}) - \frac{1}{2\gamma}\nabla\phi(\boldsymbol{y}),$$

Subtracting the two instances of (3-1) gives

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} H_{\boldsymbol{x}_1,\boldsymbol{y}_1}^{\gamma_1,t_1}(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} H_{\boldsymbol{x}_2,\boldsymbol{y}_2}^{\gamma_2,t_2}(\boldsymbol{\theta}) &= t_1\left[\nabla_{\boldsymbol{\theta}} g(\boldsymbol{x}_1, \boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} g(\boldsymbol{x}_2, \boldsymbol{\theta})\right] \\ &+ (t_1 - t_2)\left[\nabla_{\boldsymbol{\theta}} g(\boldsymbol{x}_2, \boldsymbol{\theta}) - \nabla\phi(\boldsymbol{\theta})\right] + \left(\frac{1}{2\gamma_1} - \frac{1}{2\gamma_2}\right)\left[\nabla\phi(\boldsymbol{\theta}) - \nabla\phi(\boldsymbol{y}_2)\right] \\ &- \frac{1}{2\gamma_1}\left[\nabla\phi(\boldsymbol{y}_1) - \nabla\phi(\boldsymbol{y}_2)\right].\end{aligned} \tag{6-9}$$

By Assumptions 1–2, there exist $L_{g,x\theta}, L_\phi, M_t, M_\gamma > 0$, uniform over the primal variables and inner solutions, such that

$$\begin{aligned}&\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{x}_1, \boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} g(\boldsymbol{x}_2, \boldsymbol{\theta})\| \le L_{g,x\theta}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|, \\ &\|\nabla\phi(\boldsymbol{y}_1) - \nabla\phi(\boldsymbol{y}_2)\| \le L_\phi\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|, \\ &\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{x}, \boldsymbol{\theta}) - \nabla\phi(\boldsymbol{\theta})\| \le M_t, \qquad \|\nabla\phi(\boldsymbol{\theta}) - \nabla\phi(\boldsymbol{y})\| \le M_\gamma.\end{aligned} \tag{6-10}$$

Moreover, $\gamma_1, \gamma_2 \ge \underline{\gamma} > 0$ implies

$$\frac{1}{2\gamma_1} \le \frac{1}{2\underline{\gamma}}, \qquad \left|\frac{1}{2\gamma_1} - \frac{1}{2\gamma_2}\right| \le \frac{|\gamma_1 - \gamma_2|}{2\underline{\gamma}^2}. \tag{6-11}$$

Substituting (6-10) and (6-11) into (6-9) yields

$$\begin{aligned}\|\nabla_{\boldsymbol{\theta}} H_{\boldsymbol{x}_1,\boldsymbol{y}_1}^{\gamma_1,t_1}(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} H_{\boldsymbol{x}_2,\boldsymbol{y}_2}^{\gamma_2,t_2}(\boldsymbol{\theta})\| &\le L_{g,x\theta}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + \frac{L_\phi}{2\underline{\gamma}}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\| \\ &+ M_t|t_1 - t_2| + \frac{M_\gamma}{2\underline{\gamma}^2}|\gamma_1 - \gamma_2|.\end{aligned} \tag{6-12}$$

Define

$$L_{H,z} := L_{g,x\theta} + \frac{L_\phi}{2\underline{\gamma}}, \qquad L_{H,p} := \max\left\{M_t, \frac{M_\gamma}{2\underline{\gamma}^2}\right\}. \tag{6-13}$$

Using $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|, \|\boldsymbol{y}_1 - \boldsymbol{y}_2\| \le \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|$, combining (6-12) and (6-13) gives

$$\|\nabla_{\boldsymbol{\theta}} H_{\boldsymbol{x}_1,\boldsymbol{y}_1}^{\gamma_1,t_1}(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} H_{\boldsymbol{x}_2,\boldsymbol{y}_2}^{\gamma_2,t_2}(\boldsymbol{\theta})\| \le L_{H,z}\|\boldsymbol{z}_1 - \boldsymbol{z}_2\| + L_{H,p}\big(|t_1 - t_2| + |\gamma_1 - \gamma_2|\big). \tag{6-14}$$

The first-order optimality conditions for the two inner minimizers give

$$\left\langle \nabla_{\boldsymbol{\theta}} H_{\boldsymbol{x}_1,\boldsymbol{y}_1}^{\gamma_1,t_1}(\boldsymbol{\theta}_1^*), \boldsymbol{\theta}_2^* - \boldsymbol{\theta}_1^* \right\rangle \ge 0, \left\langle \nabla_{\boldsymbol{\theta}} H_{\boldsymbol{x}_2,\boldsymbol{y}_2}^{\gamma_2,t_2}(\boldsymbol{\theta}_2^*), \boldsymbol{\theta}_1^* - \boldsymbol{\theta}_2^* \right\rangle \ge 0.$$

Adding these inequalities and adding and subtracting $\nabla_{\boldsymbol{\theta}} H^{\gamma_1,t_1}_{\boldsymbol{x}_1,\boldsymbol{y}_1}(\boldsymbol{\theta}^*_2)$, we obtain, by strong monotonicity and the Cauchy-Schwarz inequality,

$$\underline{\mu}_H \|\boldsymbol{\theta}^*_1 - \boldsymbol{\theta}^*_2\|^2 \leq \left\| \nabla_{\boldsymbol{\theta}} H^{\gamma_1,t_1}_{\boldsymbol{x}_1,\boldsymbol{y}_1}(\boldsymbol{\theta}^*_2) - \nabla_{\boldsymbol{\theta}} H^{\gamma_2,t_2}_{\boldsymbol{x}_2,\boldsymbol{y}_2}(\boldsymbol{\theta}^*_2) \right\| \cdot \|\boldsymbol{\theta}^*_1 - \boldsymbol{\theta}^*_2\|. \tag{6-15}$$

Combining (6-15) with (6-14) gives

$$\underline{\mu}_H \|\boldsymbol{\theta}^*_1 - \boldsymbol{\theta}^*_2\|^2 \leq \big[L_{H,z} \|\boldsymbol{z}_1 - \boldsymbol{z}_2\| + L_{H,p}\big(|t_1 - t_2| + |\gamma_1 - \gamma_2|\big) \big] \cdot \|\boldsymbol{\theta}^*_1 - \boldsymbol{\theta}^*_2\|.$$

Dividing by $\underline{\mu}_H \|\boldsymbol{\theta}^*_1 - \boldsymbol{\theta}^*_2\|$ when $\boldsymbol{\theta}^*_1 \neq \boldsymbol{\theta}^*_2$ yields the following bound, which also holds when the two minimizers coincide:

$$\|\boldsymbol{\theta}^*_1 - \boldsymbol{\theta}^*_2\| \leq \frac{L_{H,z}}{\underline{\mu}_H} \|\boldsymbol{z}_1 - \boldsymbol{z}_2\| + \frac{L_{H,p}}{\underline{\mu}_H} \big(|t_1 - t_2| + |\gamma_1 - \gamma_2|\big). \tag{6-16}$$

By optimality of the inner solutions,

$$v^{t_1}_{\gamma_1}(\boldsymbol{z}_1) - v^{t_2}_{\gamma_2}(\boldsymbol{z}_2) \leq H^{\gamma_1,t_1}_{\boldsymbol{x}_1,\boldsymbol{y}_1}(\boldsymbol{\theta}^*_2) - H^{\gamma_2,t_2}_{\boldsymbol{x}_2,\boldsymbol{y}_2}(\boldsymbol{\theta}^*_2).$$

Interchanging the two parameter points gives the reverse bound. Assumptions 1–2 imply that both g_t and $H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}$ are uniformly Lipschitz continuous in the primal variables and parameters. Hence there exist $L_{v,z}, L_{v,p} > 0$ such that

$$|v^{t_1}_{\gamma_1}(\boldsymbol{z}_1) - v^{t_2}_{\gamma_2}(\boldsymbol{z}_2)| \leq L_{v,z} \|\boldsymbol{z}_1 - \boldsymbol{z}_2\| + L_{v,p}\big(|t_1 - t_2| + |\gamma_1 - \gamma_2|\big).$$

Since $\hat{V}^t_\gamma = g_t - v^t_\gamma - \varepsilon_{\mathrm{rel}}$,

$$|\hat{V}^{t_1}_{\gamma_1}(\boldsymbol{z}_1) - \hat{V}^{t_2}_{\gamma_2}(\boldsymbol{z}_2)| \leq L_{V,z} \|\boldsymbol{z}_1 - \boldsymbol{z}_2\| + L_{V,p}\big(|t_1 - t_2| + |\gamma_1 - \gamma_2|\big).$$

The gradient formula in Lemma 3 gives

$$\nabla \hat{V}^t_\gamma(\boldsymbol{z}) = \nabla g_t(\boldsymbol{z}) - \nabla_{\boldsymbol{z}} H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{\theta}^*_{\gamma,t}(\boldsymbol{z})).$$

Subtracting the gradients at the two parameter points and inserting the corresponding intermediate terms yields

$$\begin{aligned} \|\nabla \hat{V}^{t_1}_{\gamma_1}(\boldsymbol{z}_1) - \nabla \hat{V}^{t_2}_{\gamma_2}(\boldsymbol{z}_2)\| &\leq \|\nabla g_{t_1}(\boldsymbol{z}_1) - \nabla g_{t_2}(\boldsymbol{z}_2)\| \\ &+ \|\nabla_{\boldsymbol{z}} H^{\gamma_1,t_1}_{\boldsymbol{x}_1,\boldsymbol{y}_1}(\boldsymbol{\theta}^*_1) - \nabla_{\boldsymbol{z}} H^{\gamma_1,t_1}_{\boldsymbol{x}_1,\boldsymbol{y}_1}(\boldsymbol{\theta}^*_2)\| + \|\nabla_{\boldsymbol{z}} H^{\gamma_1,t_1}_{\boldsymbol{x}_1,\boldsymbol{y}_1}(\boldsymbol{\theta}^*_2) - \nabla_{\boldsymbol{z}} H^{\gamma_2,t_2}_{\boldsymbol{x}_2,\boldsymbol{y}_2}(\boldsymbol{\theta}^*_2)\|. \end{aligned}$$

Since $g_t = tg + (1-t)\phi$, the first term is bounded by Assumptions 1–2. The same assumptions imply Lipschitz continuity of $\nabla_{\boldsymbol{z}} H^{\gamma,t}_{\boldsymbol{x},\boldsymbol{y}}$ in the remaining terms, since $\nabla_{\boldsymbol{y}} D_\phi$ is a component of the joint gradient of D_ϕ. Combining these bounds with (6-16), we obtain constants $L_{\nabla V,z}, L_{\nabla V,p} > 0$ such that

$$\|\nabla \hat{V}^{t_1}_{\gamma_1}(\boldsymbol{z}_1) - \nabla \hat{V}^{t_2}_{\gamma_2}(\boldsymbol{z}_2)\| \leq L_{\nabla V,z} \|\boldsymbol{z}_1 - \boldsymbol{z}_2\| + L_{\nabla V,p}\big(|t_1 - t_2| + |\gamma_1 - \gamma_2|\big).$$

The last two estimates are precisely (6-7) and (6-8). □

Lemma 10 (Uniform Regularity of the Optimal Multiplier Mapping)**.** *Under the conditions of Lemma 9, for any $\rho, \omega, \Lambda > 0$, the optimal multiplier mapping satisfies*

$$|\lambda^*_{\omega,\gamma_1,t_1}(\boldsymbol{z}_1) - \lambda^*_{\omega,\gamma_2,t_2}(\boldsymbol{z}_2)| \leq \frac{L_{V,z}}{\omega} \|\boldsymbol{z}_1 - \boldsymbol{z}_2\| + \frac{L_{V,p}}{\omega} \big(|t_1 - t_2| + |\gamma_1 - \gamma_2|\big). \tag{6-17}$$

Proof. By (6-4) and nonexpansiveness of the projection,

$$|\lambda^*_{\omega,\gamma_1,t_1}(\boldsymbol{z}_1) - \lambda^*_{\omega,\gamma_2,t_2}(\boldsymbol{z}_2)| \le \frac{1}{\omega}|\hat{V}^{t_1}_{\gamma_1}(\boldsymbol{z}_1) - \hat{V}^{t_2}_{\gamma_2}(\boldsymbol{z}_2)|, \tag{6-18}$$

Combining (6-18) with (6-7) proves (6-17). □

Lemma 11 (Uniform Regularity of the Envelope)**.** *Under the conditions of Lemmas 8–10, there exist $L_{\Phi,p}, L_{\nabla\Phi,p} > 0$ such that, for every $\boldsymbol{z} \in \mathcal{Z}$ and any $(t_1,\gamma_1),(t_2,\gamma_2)$ in the parameter set,*

$$\begin{aligned} &|\Phi^{\gamma_1,t_1}_{\rho,\omega}(\boldsymbol{z}) - \Phi^{\gamma_2,t_2}_{\rho,\omega}(\boldsymbol{z})| \le L_{\Phi,p}\big(|t_1-t_2| + |\gamma_1-\gamma_2|\big), \\ &\|\nabla\Phi^{\gamma_1,t_1}_{\rho,\omega}(\boldsymbol{z}) - \nabla\Phi^{\gamma_2,t_2}_{\rho,\omega}(\boldsymbol{z})\| \le L_{\nabla\Phi,p}\big(|t_1-t_2| + |\gamma_1-\gamma_2|\big). \end{aligned} \tag{6-19}$$

Proof. By (6-7), $\hat{V}^t_\gamma$ is uniformly bounded on $\mathcal{Z}$ and the parameter set. Since $s \mapsto [s]^2_+$ is Lipschitz continuous on bounded intervals, there exists $C_A > 0$ such that, for every $\lambda \in [0,\Lambda]$,

$$|\mathscr{A}^{\gamma_1,t_1}_{\rho,\omega}(\boldsymbol{z},\lambda) - \mathscr{A}^{\gamma_2,t_2}_{\rho,\omega}(\boldsymbol{z},\lambda)| \le C_A\big(|t_1-t_2| + |\gamma_1-\gamma_2|\big). \tag{6-20}$$

Let $\lambda^*_i := \lambda^*_{\omega,\gamma_i,t_i}(\boldsymbol{z})$. By optimality,

$$\Phi^{\gamma_1,t_1}_{\rho,\omega}(\boldsymbol{z}) - \Phi^{\gamma_2,t_2}_{\rho,\omega}(\boldsymbol{z}) \le \mathscr{A}^{\gamma_1,t_1}_{\rho,\omega}(\boldsymbol{z},\lambda^*_1) - \mathscr{A}^{\gamma_2,t_2}_{\rho,\omega}(\boldsymbol{z},\lambda^*_1) \le C_A\big(|t_1-t_2| + |\gamma_1-\gamma_2|\big). \tag{6-21}$$

Applying (6-20) to (6-21) and interchanging the two parameter pairs proves the function-value estimate in (6-19).

For the gradient estimate, let $\mu^*_i := [\lambda^*_i + \rho\hat{V}^{t_i}_{\gamma_i}(\boldsymbol{z})]_+$. Nonexpansiveness of projection onto the nonnegative half-line gives

$$|\mu^*_1 - \mu^*_2| \le |\lambda^*_1 - \lambda^*_2| + \rho|\hat{V}^{t_1}_{\gamma_1}(\boldsymbol{z}) - \hat{V}^{t_2}_{\gamma_2}(\boldsymbol{z})|. \tag{6-22}$$

By (6-5),

$$\|\nabla\Phi^{\gamma_1,t_1}_{\rho,\omega}(\boldsymbol{z}) - \nabla\Phi^{\gamma_2,t_2}_{\rho,\omega}(\boldsymbol{z})\| \le |\mu^*_1 - \mu^*_2|\,\|\nabla\hat{V}^{t_1}_{\gamma_1}(\boldsymbol{z})\| + |\mu^*_2|\,\|\nabla\hat{V}^{t_1}_{\gamma_1}(\boldsymbol{z}) - \nabla\hat{V}^{t_2}_{\gamma_2}(\boldsymbol{z})\|. \tag{6-23}$$

Both $|\mu^*_i|$ and $\|\nabla\hat{V}^{t_i}_{\gamma_i}(\boldsymbol{z})\|$ are uniformly bounded on $\mathcal{Z}$ and the parameter set. Using (6-22), (6-7), (6-8), and (6-17) in (6-23) gives

$$\|\nabla\Phi^{\gamma_1,t_1}_{\rho,\omega}(\boldsymbol{z}) - \nabla\Phi^{\gamma_2,t_2}_{\rho,\omega}(\boldsymbol{z})\| \le L_{\nabla\Phi,p}\big(|t_1-t_2| + |\gamma_1-\gamma_2|\big),$$

□

A.7 One-Step Dual Tracking

We next derive a recursion for the distance between the dual iterate and the regularized optimal multiplier.

Let $X_k := \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2$, $Y_k := \|\boldsymbol{y}_{k+1} - \boldsymbol{y}_k\|^2$, $\Delta_k := |t_{k+1} - t_k| + |\gamma_{k+1} - \gamma_k|$, and $\mathcal{D}_k := |\lambda_k - \lambda^*_{\omega,k}(\boldsymbol{z}_k)|^2$.

Lemma 12 (Dual Tracking Error)**.** *Suppose Assumptions 1 and 2 hold and $\inf_k \mu_H(t_k,\gamma_k) > 0$. For fixed $\omega > 0$ with $0 < \rho\omega < 1$, there exist $\chi_\lambda \in (0,1)$ and $C_{\lambda,z}, C_{\lambda,E}, C_{\lambda,p} > 0$ such that, for every $k \ge 0$,*

$$\mathcal{D}_{k+1} \le \chi_\lambda\mathcal{D}_k + C_{\lambda,z}\|\boldsymbol{z}_{k+1} - \boldsymbol{z}_k\|^2 + C_{\lambda,E}\mathcal{E}_{k+1} + C_{\lambda,p}\Delta^2_k. \tag{7-1}$$

Proof. The triangle inequality, strong convexity of ϕ, and (4-15) give

$$\begin{aligned}\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}^*_{\gamma_k,t_k}(\boldsymbol{z}_{k+1})\|^2 &\le 2\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}^*_k\|^2 + 2\|\boldsymbol{\theta}^*_k - \boldsymbol{\theta}^*_{\gamma_k,t_k}(\boldsymbol{z}_{k+1})\|^2 \\ &\le \frac{4}{\mu_\phi}\mathcal{E}_{k+1} + 4L^2_{\theta,x}X_k + 4L^2_{\theta,y}Y_k.\end{aligned}$$

Assumption 2 and Lemma 7 further yield

$$\begin{aligned}|\tilde{V}^{t_k}_{\gamma_k}(\boldsymbol{x}_{k+1},\boldsymbol{y}_{k+1},\boldsymbol{\theta}_{k+1}) - \hat{V}^{t_k}_{\gamma_k}(\boldsymbol{z}_{k+1})|^2 &\le C_V D_\phi(\boldsymbol{\theta}^*_{\gamma_k,t_k}(\boldsymbol{z}_{k+1}),\boldsymbol{\theta}_{k+1}) \\ &\le C_V\left(\frac{2L_\phi}{\mu_\phi}\mathcal{E}_{k+1} + 2L_\phi L^2_{\theta,x}X_k + 2L_\phi L^2_{\theta,y}Y_k\right) \le A_+(\mathcal{E}_{k+1}+X_k+Y_k),\end{aligned} \tag{7-2}$$

where one may take

$$A_+ := C_V \max\left\{\frac{2L_\phi}{\mu_\phi}, 2L_\phi L^2_{\theta,x}, 2L_\phi L^2_{\theta,y}\right\}.$$

Let $a := 1-\rho\omega \in (0,1)$. The explicit formula (6-4) gives the following fixed-point identity in each of the cases $\hat{V}^{t_k}_{\gamma_k}(\boldsymbol{z}) \le 0$, $0 < \hat{V}^{t_k}_{\gamma_k}(\boldsymbol{z}) < \omega\Lambda$, and $\hat{V}^{t_k}_{\gamma_k}(\boldsymbol{z}) \ge \omega\Lambda$:

$$\lambda^*_{\omega,k}(\boldsymbol{z}) = \Pi_{[0,\Lambda]}\left(a\lambda^*_{\omega,k}(\boldsymbol{z}) + \rho\hat{V}^{t_k}_{\gamma_k}(\boldsymbol{z})\right). \tag{7-3}$$

Applying nonexpansiveness to (15) and (7-3), and using (6-17) and (7-2), yields

$$\begin{aligned}|\lambda_{k+1} - \lambda^*_{\omega,k+1}(\boldsymbol{z}_{k+1})| &\le a|\lambda_k - \lambda^*_{\omega,k}(\boldsymbol{z}_{k+1})| + \rho\sqrt{A_+(\mathcal{E}_{k+1}+X_k+Y_k)} \\ &\quad + |\lambda^*_{\omega,k}(\boldsymbol{z}_{k+1}) - \lambda^*_{\omega,k+1}(\boldsymbol{z}_{k+1})| \\ &\le a|\lambda_k - \lambda^*_{\omega,k}(\boldsymbol{z}_k)| + \frac{aL_{V,z}}{\omega}\|\boldsymbol{z}_{k+1}-\boldsymbol{z}_k\| + \rho\sqrt{A_+(\mathcal{E}_{k+1}+X_k+Y_k)} + \frac{L_{V,p}}{\omega}\Delta_k.\end{aligned} \tag{7-4}$$

Set $\chi_\lambda := (1+a^2)/2 \in (a^2,1)$ and $\xi_\lambda := \chi_\lambda/a^2 - 1 = (1-a^2)/(2a^2) > 0$. For any $u, v \ge 0$, Young's inequality gives

$$(u+v)^2 \le (1+\xi_\lambda)u^2 + \left(1+\frac{1}{\xi_\lambda}\right)v^2. \tag{7-5}$$

In (7-4), take $u := a\sqrt{\mathcal{D}_k}$ and set

$$v := \frac{aL_{V,z}}{\omega}\|\boldsymbol{z}_{k+1}-\boldsymbol{z}_k\| + \rho\sqrt{A_+(\mathcal{E}_{k+1}+X_k+Y_k)} + \frac{L_{V,p}}{\omega}\Delta_k.$$

Both sides of (7-4) are nonnegative. Squaring, applying (7-5), and using $(1+\xi_\lambda)a^2 = \chi_\lambda$ yield

$$\begin{aligned}\mathcal{D}_{k+1} \le \chi_\lambda\mathcal{D}_k + \left(1+\frac{1}{\xi_\lambda}\right)\Bigg(&\frac{aL_{V,z}}{\omega}\|\boldsymbol{z}_{k+1}-\boldsymbol{z}_k\| \\ &+ \rho\sqrt{A_+(\mathcal{E}_{k+1}+X_k+Y_k)} + \frac{L_{V,p}}{\omega}\Delta_k\Bigg)^2.\end{aligned} \tag{7-6}$$

Applying $(r_1+r_2+r_3)^2 \le 3(r_1^2+r_2^2+r_3^2)$ to the bracketed sum in (7-6) and using $\|\boldsymbol{z}_{k+1}-\boldsymbol{z}_k\|^2 = X_k + Y_k$ gives

$$\mathcal{D}_{k+1} \le \chi_\lambda \mathcal{D}_k + 3\left(1+\frac{1}{\xi_\lambda}\right)\left[\left(\frac{a^2 L_{V,z}^2}{\omega^2} + \rho^2 A_+\right)(X_k + Y_k) + \rho^2 A_+ \mathcal{E}_{k+1} + \frac{L_{V,p}^2}{\omega^2}\Delta_k^2\right]. \tag{7-7}$$

Thus, one may choose

$$C_{\lambda,z} := 3\left(1+\frac{1}{\xi_\lambda}\right)\left(\frac{a^2 L_{V,z}^2}{\omega^2} + \rho^2 A_+\right), \qquad C_{\lambda,E} := 3\left(1+\frac{1}{\xi_\lambda}\right)\rho^2 A_+, \qquad C_{\lambda,p} := 3\left(1+\frac{1}{\xi_\lambda}\right)\frac{L_{V,p}^2}{\omega^2}. \tag{7-8}$$

With the constants defined in (7-8), (7-7) can be written as

$$\mathcal{D}_{k+1} \le \chi_\lambda \mathcal{D}_k + C_{\lambda,z}(X_k + Y_k) + C_{\lambda,E}\mathcal{E}_{k+1} + C_{\lambda,p}\Delta_k^2.$$

Since $X_k + Y_k = \|\boldsymbol{z}_{k+1} - \boldsymbol{z}_k\|^2$, this is (7-1). □

A.8 Primal Gradient Errors and Envelope Descent

The next two lemmas bound the errors in the primal directions and establish a descent inequality for the envelope in terms of the inner and dual tracking errors.

Write $\boldsymbol{z}_{k+\frac{1}{2}} := (\boldsymbol{x}_{k+1}, \boldsymbol{y}_k)$.

Lemma 13 (Approximation Error of the Primal Directions)**.** *Under the conditions of Lemma 12, there exist $G_{x,D}, G_{x,E}, G_{y,D}, G_{y,E}, G_{y,x} > 0$ such that*

$$\left\|\boldsymbol{d}_x^k - \nabla_{\boldsymbol{x}}\Phi_k(\boldsymbol{x}_k, \boldsymbol{y}_k)\right\|^2 \le G_{x,D}\mathcal{D}_k + G_{x,E}\mathcal{E}_{k+1}, \tag{8-1}$$

and

$$\left\|\boldsymbol{d}_y^k - \nabla_{\boldsymbol{y}}\Phi_k(\boldsymbol{x}_{k+1}, \boldsymbol{y}_k)\right\|^2 \le G_{y,D}\mathcal{D}_k + G_{y,E}\mathcal{E}_{k+1} + G_{y,x}X_k. \tag{8-2}$$

Proof. By (6-7)–(6-8) and continuity of $\tilde{V}_{\gamma_k}^{t_k}$, there exist constants $M_{\hat{V}}, M_{\nabla\hat{V}}, M_{\tilde{V}} > 0$, independent of k, such that

$$|\hat{V}_{\gamma_k}^{t_k}| \le M_{\hat{V}}, \qquad \|\nabla\hat{V}_{\gamma_k}^{t_k}\| \le M_{\nabla\hat{V}}, \qquad |\tilde{V}_{\gamma_k}^{t_k}| \le M_{\tilde{V}}. \tag{8-3}$$

At the $\boldsymbol{x}$-update point, $\mathcal{E}_{k+1} = D_\phi(\boldsymbol{\theta}_k^*, \boldsymbol{\theta}_{k+1})$ and Lemma 7 provide $A_x > 0$ such that

$$|\tilde{V}_{\gamma_k}^{t_k}(\boldsymbol{x}_k, \boldsymbol{y}_k, \boldsymbol{\theta}_{k+1}) - \hat{V}_{\gamma_k}^{t_k}(\boldsymbol{z}_k)|^2 + \left\|\nabla_{\boldsymbol{x}}\tilde{V}_{\gamma_k}^{t_k}(\boldsymbol{x}_k, \boldsymbol{y}_k, \boldsymbol{\theta}_{k+1}) - \nabla_{\boldsymbol{x}}\hat{V}_{\gamma_k}^{t_k}(\boldsymbol{z}_k)\right\|^2 \le A_x\mathcal{E}_{k+1}. \tag{8-4}$$

For the $\boldsymbol{y}$ update, let $\boldsymbol{\theta}_{k+\frac{1}{2}}^* := \boldsymbol{\theta}_{\gamma_k,t_k}^*(\boldsymbol{z}_{k+\frac{1}{2}})$. Strong convexity of ϕ, local Lipschitz continuity of $\nabla\phi$, and Lemma 2 give

$$D_\phi(\boldsymbol{\theta}_{k+\frac{1}{2}}^*, \boldsymbol{\theta}_{k+1}) \le \frac{L_\phi}{2}\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_{k+\frac{1}{2}}^*\|^2 \le \frac{2L_\phi}{\mu_\phi}\mathcal{E}_{k+1} + L_\phi L_{\theta,x}^2 X_k.$$

Lemma 7 then yields $A_y > 0$ such that

$$
\begin{aligned}
&|\tilde{V}_{\gamma_k}^{t_k}(\boldsymbol{x}_{k+1}, \boldsymbol{y}_k, \boldsymbol{\theta}_{k+1}) - \hat{V}_{\gamma_k}^{t_k}(\boldsymbol{z}_{k+\frac{1}{2}})|^2 \\
&\quad + \left\|\nabla_{\boldsymbol{y}} \tilde{V}_{\gamma_k}^{t_k}(\boldsymbol{x}_{k+1}, \boldsymbol{y}_k, \boldsymbol{\theta}_{k+1}) - \nabla_{\boldsymbol{y}} \hat{V}_{\gamma_k}^{t_k}(\boldsymbol{z}_{k+\frac{1}{2}})\right\|^2 \leq A_y(\mathcal{E}_{k+1} + X_k).
\end{aligned} \tag{8-5}
$$

Define

$$
\mu_{\omega,k}^*(\boldsymbol{z}) := [\lambda_{\omega,k}^*(\boldsymbol{z}) + \rho \hat{V}_{\gamma_k}^{t_k}(\boldsymbol{z})]_+, \qquad \overline{M}_\mu := \Lambda + \rho \max\{M_{\hat{V}}, M_{\tilde{V}}\}. \tag{8-6}
$$

Because $\lambda_k, \lambda_{\omega,k}^* \in [0, \Lambda]$, both the approximate and exact effective multipliers are bounded by $\overline{M}_\mu$. Nonexpansiveness of projection onto the nonnegative half-line gives

$$
\begin{aligned}
&|[\lambda_k + \rho \tilde{V}_{\gamma_k}^{t_k}(\boldsymbol{x}_k, \boldsymbol{y}_k, \boldsymbol{\theta}_{k+1})]_+ - \mu_{\omega,k}^*(\boldsymbol{z}_k)| \\
&\leq |\lambda_k - \lambda_{\omega,k}^*(\boldsymbol{z}_k)| + \rho |\tilde{V}_{\gamma_k}^{t_k}(\boldsymbol{x}_k, \boldsymbol{y}_k, \boldsymbol{\theta}_{k+1}) - \hat{V}_{\gamma_k}^{t_k}(\boldsymbol{z}_k)|.
\end{aligned} \tag{8-7}
$$

Subtracting (6-5) from the direction (11) gives

$$
\begin{aligned}
\boldsymbol{d}_x^k - \nabla_{\boldsymbol{x}} \Phi_k(\boldsymbol{z}_k) &= [\lambda_k + \rho \tilde{V}_{\gamma_k}^{t_k}(\boldsymbol{x}_k, \boldsymbol{y}_k, \boldsymbol{\theta}_{k+1})]_+ \left(\nabla_{\boldsymbol{x}} \tilde{V}_{\gamma_k}^{t_k} - \nabla_{\boldsymbol{x}} \hat{V}_{\gamma_k}^{t_k}\right) \\
&\quad + \left([\lambda_k + \rho \tilde{V}_{\gamma_k}^{t_k}(\boldsymbol{x}_k, \boldsymbol{y}_k, \boldsymbol{\theta}_{k+1})]_+ - \mu_{\omega,k}^*(\boldsymbol{z}_k)\right) \nabla_{\boldsymbol{x}} \hat{V}_{\gamma_k}^{t_k}(\boldsymbol{z}_k).
\end{aligned} \tag{8-8}
$$

The two gradients in brackets in (8-8) are evaluated at the same point $(\boldsymbol{x}_k, \boldsymbol{y}_k)$. Equations (8-4) and (8-7), together with $(a+b)^2 \leq 2a^2 + 2b^2$ therefore yield

$$
\left\|\boldsymbol{d}_x^k - \nabla_{\boldsymbol{x}} \Phi_k(\boldsymbol{z}_k)\right\|^2 \leq 4M_{\nabla \hat{V}}^2 \mathcal{D}_k + \left(2\overline{M}_\mu^2 A_x + 4\rho^2 M_{\nabla \hat{V}}^2 A_x\right) \mathcal{E}_{k+1}.
$$

Thus, one may take

$$
G_{x,D} := 4M_{\nabla \hat{V}}^2, \qquad G_{x,E} := 2\overline{M}_\mu^2 A_x + 4\rho^2 M_{\nabla \hat{V}}^2 A_x.
$$

For the $\boldsymbol{y}$ direction, use $\boldsymbol{z}_{k+\frac{1}{2}}$ as the comparison point. Equation (6-17) gives

$$
|\lambda_{\omega,k}^*(\boldsymbol{z}_{k+\frac{1}{2}}) - \lambda_{\omega,k}^*(\boldsymbol{z}_k)| \leq \frac{L_{V,z}}{\omega} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|.
$$

Together with (8-5), this gives

$$
\begin{aligned}
&\left|[\lambda_k + \rho \tilde{V}_{\gamma_k}^{t_k}(\boldsymbol{x}_{k+1}, \boldsymbol{y}_k, \boldsymbol{\theta}_{k+1})]_+ - \mu_{\omega,k}^*(\boldsymbol{z}_{k+\frac{1}{2}})\right| \\
&\leq \sqrt{\mathcal{D}_k} + \frac{L_{V,z}}{\omega} \sqrt{X_k} + \rho \sqrt{A_y(\mathcal{E}_{k+1} + X_k)}.
\end{aligned} \tag{8-9}
$$

Define

$$
\widetilde{\mu}_{y,k} := [\lambda_k + \rho \tilde{V}_{\gamma_k}^{t_k}(\boldsymbol{x}_{k+1}, \boldsymbol{y}_k, \boldsymbol{\theta}_{k+1})]_+, \qquad \mu_{k+\frac{1}{2}}^* := \mu_{\omega,k}^*(\boldsymbol{z}_{k+\frac{1}{2}}).
$$

The direction (13) and the envelope gradient (6-5) give

$$
\boldsymbol{d}_y^k - \nabla_{\boldsymbol{y}} \Phi_k(\boldsymbol{z}_{k+\frac{1}{2}}) = \widetilde{\mu}_{y,k} \left(\nabla_{\boldsymbol{y}} \tilde{V}_{\gamma_k}^{t_k} - \nabla_{\boldsymbol{y}} \hat{V}_{\gamma_k}^{t_k}\right) + (\widetilde{\mu}_{y,k} - \mu_{k+\frac{1}{2}}^*) \nabla_{\boldsymbol{y}} \hat{V}_{\gamma_k}^{t_k}(\boldsymbol{z}_{k+\frac{1}{2}}), \tag{8-10}
$$

where both constraint gradients are evaluated at $\boldsymbol{z}_{k+\frac{1}{2}}$. Equations (8-5) and (8-9) give

$$\|\nabla_{\boldsymbol{y}}\tilde{V}_{\gamma_k}^{t_k} - \nabla_{\boldsymbol{y}}\hat{V}_{\gamma_k}^{t_k}\|^2 \le A_y(\mathcal{E}_{k+1} + X_k),$$
$$|\widetilde{\mu}_{y,k} - \mu^*_{k+\frac{1}{2}}|^2 \le 3\mathcal{D}_k + \frac{3L_{V,z}^2}{\omega^2}X_k + 3\rho^2 A_y(\mathcal{E}_{k+1} + X_k).$$

Apply $(a+b)^2 \le 2a^2 + 2b^2$ to (8-10), together with $\widetilde{\mu}_{y,k} \le \overline{M}_\mu$ and $\|\nabla\hat{V}_{\gamma_k}^{t_k}\| \le M_{\nabla\hat{V}}$, and set

$$G_{y,D} := 6M_{\nabla\hat{V}}^2, \qquad G_{y,E} := 2\overline{M}_\mu^2 A_y + 6\rho^2 M_{\nabla\hat{V}}^2 A_y,$$
$$G_{y,x} := 2\overline{M}_\mu^2 A_y + 6M_{\nabla\hat{V}}^2\left(\frac{L_{V,z}^2}{\omega^2} + \rho^2 A_y\right).$$

Then

$$\left\|\boldsymbol{d}_y^k - \nabla_{\boldsymbol{y}}\Phi_k(\boldsymbol{z}_{k+\frac{1}{2}})\right\|^2 \le G_{y,D}\mathcal{D}_k + G_{y,E}\mathcal{E}_{k+1} + G_{y,x}X_k,$$

□

Lemma 14 (Inexact Descent of the Envelope)**.** *Under the conditions of Lemma 13, let $L_\Phi > 0$ be a uniform Lipschitz constant for $\nabla\Phi_k$ on $\mathcal{Z}$. Suppose the step sizes satisfy*

$$0 < \alpha < \frac{2\mu_\psi}{L_\Phi + 1 + G_{y,x}}, \qquad 0 < \beta < \frac{2\mu_\phi}{L_\Phi + 1}. \tag{8-11}$$

Define

$$a_x := \frac{\mu_\psi}{\alpha} - \frac{L_\Phi + 1 + G_{y,x}}{2} > 0, \qquad a_y := \frac{\mu_\phi}{\beta} - \frac{L_\Phi + 1}{2} > 0, \tag{8-12}$$

and

$$B_D := \frac{G_{x,D} + G_{y,D}}{2}, \qquad B_E := \frac{G_{x,E} + G_{y,E}}{2}. \tag{8-13}$$

Then the iterates satisfy

$$\begin{aligned}&\Phi_{k+1}(\boldsymbol{z}_{k+1}) - \Phi_k(\boldsymbol{z}_k)\\ &\le -a_x\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 - a_y\|\boldsymbol{y}_{k+1} - \boldsymbol{y}_k\|^2 + B_D\mathcal{D}_k + B_E\mathcal{E}_{k+1} + L_{\Phi,p}\Delta_k.\end{aligned} \tag{8-14}$$

Proof. For fixed (t_k, γ_k), (6-4), (6-7), and nonexpansiveness of the projection give

$$|\lambda^*_{\omega,k}(\boldsymbol{z}_1) - \lambda^*_{\omega,k}(\boldsymbol{z}_2)| \le \frac{L_{V,z}}{\omega}\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|.$$

Nonexpansiveness of projection onto the nonnegative half-line further gives

$$|\mu^*_{\omega,k}(\boldsymbol{z}_1) - \mu^*_{\omega,k}(\boldsymbol{z}_2)| \le \left(\frac{1}{\omega} + \rho\right)L_{V,z}\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|.$$

Let L_f be a Lipschitz constant of ∇f on $\mathcal{Z}$, as ensured by Assumption 1. By (6-5), Lemma 9, and the bounds (8-3) and (8-6),

$$\|\nabla\Phi_k(\boldsymbol{z}_1) - \nabla\Phi_k(\boldsymbol{z}_2)\| \le \left[L_f + M_{\nabla\hat{V}}\left(\frac{1}{\omega} + \rho\right)L_{V,z} + \overline{M}_\mu L_{\nabla V,z}\right]\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|.$$

Thus $\nabla\Phi_k$ is Lipschitz continuous with a constant L_Φ independent of k. We now apply this bound to the two primal updates.

Taking $\boldsymbol{x} = \boldsymbol{x}_k$ in the first-order optimality condition for (12) gives

$$\left\langle \boldsymbol{d}_x^k, \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \right\rangle \leq -\frac{\mu_\psi}{\alpha}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2.$$

Applying the descent lemma to $\Phi_k(\cdot, \boldsymbol{y}_k)$ and adding and subtracting $\boldsymbol{d}_x^k$ in the linear term yields

$$\begin{aligned}\Phi_k(\boldsymbol{x}_{k+1}, \boldsymbol{y}_k) - \Phi_k(\boldsymbol{x}_k, \boldsymbol{y}_k) \leq &-\left(\frac{\mu_\psi}{\alpha} - \frac{L_\Phi}{2}\right)\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 \\ &+ \left\langle \nabla_{\boldsymbol{x}}\Phi_k(\boldsymbol{z}_k) - \boldsymbol{d}_x^k, \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \right\rangle.\end{aligned}$$

Young's inequality and (8-1) bound the last term as

$$\Phi_k(\boldsymbol{z}_{k+\frac{1}{2}}) - \Phi_k(\boldsymbol{z}_k) \leq -\left(\frac{\mu_\psi}{\alpha} - \frac{L_\Phi + 1}{2}\right) X_k + \frac{G_{x,D}}{2}\mathcal{D}_k + \frac{G_{x,E}}{2}\mathcal{E}_{k+1}. \tag{8-15}$$

The first-order optimality condition for (14) gives, for every $\boldsymbol{y} \in \mathcal{Y}$,

$$\left\langle \boldsymbol{d}_y^k + \frac{1}{\beta}\big(\nabla\phi(\boldsymbol{y}_{k+1}) - \nabla\phi(\boldsymbol{y}_k)\big), \boldsymbol{y} - \boldsymbol{y}_{k+1} \right\rangle \geq 0.$$

Setting $\boldsymbol{y} = \boldsymbol{y}_k$ and using μ_ϕ-strong convexity of ϕ gives

$$\left\langle \boldsymbol{d}_y^k, \boldsymbol{y}_{k+1} - \boldsymbol{y}_k \right\rangle \leq -\frac{\mu_\phi}{\beta} Y_k.$$

Applying the descent lemma in the $\boldsymbol{y}$ component at $\boldsymbol{z}_{k+\frac{1}{2}}$ yields

$$\begin{aligned}\Phi_k(\boldsymbol{z}_{k+1}) - \Phi_k(\boldsymbol{z}_{k+\frac{1}{2}}) &\leq \langle \nabla_{\boldsymbol{y}}\Phi_k(\boldsymbol{z}_{k+\frac{1}{2}}), \boldsymbol{y}_{k+1} - \boldsymbol{y}_k\rangle + \frac{L_\Phi}{2} Y_k \\ &= \langle \boldsymbol{d}_y^k, \boldsymbol{y}_{k+1} - \boldsymbol{y}_k\rangle + \left\langle \nabla_{\boldsymbol{y}}\Phi_k(\boldsymbol{z}_{k+\frac{1}{2}}) - \boldsymbol{d}_y^k, \boldsymbol{y}_{k+1} - \boldsymbol{y}_k \right\rangle + \frac{L_\Phi}{2} Y_k.\end{aligned}$$

Applying Young's inequality to the error inner product and invoking (8-2) gives

$$\Phi_k(\boldsymbol{z}_{k+1}) - \Phi_k(\boldsymbol{z}_{k+\frac{1}{2}}) \leq -\left(\frac{\mu_\phi}{\beta} - \frac{L_\Phi + 1}{2}\right) Y_k + \frac{G_{y,D}}{2}\mathcal{D}_k + \frac{G_{y,E}}{2}\mathcal{E}_{k+1} + \frac{G_{y,x}}{2} X_k. \tag{8-16}$$

Adding (8-15) and (8-16) and using the definitions in (8-12)–(8-13) yields

$$\Phi_k(\boldsymbol{z}_{k+1}) - \Phi_k(\boldsymbol{z}_k) \leq -a_x X_k - a_y Y_k + B_D\mathcal{D}_k + B_E\mathcal{E}_{k+1}. \tag{8-17}$$

The parameter variation is bounded by (6-19):

$$\Phi_{k+1}(\boldsymbol{z}_{k+1}) - \Phi_k(\boldsymbol{z}_{k+1}) \leq L_{\Phi,p}\Delta_k. \tag{8-18}$$

Adding (8-17) and (8-18) proves (8-14). □

A.9 Lyapunov Descent and Convergence Rate

We combine the tracking recursions with the envelope descent estimate to derive a Lyapunov inequality and a stationarity rate for the problem associated with $(1, \underline{\gamma})$.

Throughout this section, the homotopy parameters follow (16), with fixed positive increments truncated at $(1, \underline{\gamma})$. Set $\boldsymbol{z}_{-1} := \boldsymbol{z}_0$, $t_{-1} := t_0$, and $\gamma_{-1} := \gamma_0$, giving $X_{-1} = Y_{-1} = \Delta_{-1} = 0$. The homotopy perturbations are summable:

$$\mathfrak{H} := \sum_{k=0}^{\infty} \left[\Delta_k + \Delta_k^2 + (1 - t_k)^2 + (\gamma_k - \underline{\gamma})^2\right] < \infty. \tag{9-1}$$

Proposition 2 (Lyapunov Descent and Summability of the Errors)**.** *Suppose Assumptions 1 and 2 hold. Let $\varepsilon_{\mathrm{rel}} > 0$ and $\rho, \omega, \Lambda > 0$ with $0 < \rho\omega < 1$, and assume that the inner strong-convexity modulus is uniformly positive along the homotopy path.*

Then, for sufficiently small positive step sizes η, α, and β, there exist $\Gamma_E, \Gamma_D, c_x, c_y, c, C_H > 0$ such that the Lyapunov function

$$\mathcal{W}_k := \Phi_k(\boldsymbol{z}_k) + \Gamma_E \mathcal{E}_k + \Gamma_D \mathcal{D}_k + c_x \|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\|^2 + c_y \|\boldsymbol{y}_k - \boldsymbol{y}_{k-1}\|^2 \tag{9-2}$$

satisfies

$$\mathcal{W}_{k+1} - \mathcal{W}_k \leq -c\big(\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 + \|\boldsymbol{y}_{k+1} - \boldsymbol{y}_k\|^2 + \mathcal{E}_k + \mathcal{D}_k\big) + C_H(\Delta_k + \Delta_{k-1}^2). \tag{9-3}$$

Moreover, $\sum_{k=0}^{\infty}(X_k + Y_k + \mathcal{E}_k + \mathcal{D}_k) < \infty$.

Proof. Let $\mathcal{P} := \overline{\{(t_k, \gamma_k) : k \geq -1\}}$ be the closure of the homotopy path. Continuity of μ_H gives $\inf_{(t,\gamma)\in\mathcal{P}} \mu_H(t, \gamma) > 0$. Assumptions 1–2 and compactness of the feasible and parameter sets ensure that the constants in Lemmas 6–14 can be chosen uniformly in k.

Step 1: Tracking recursions.

Let $C_{E,x}$ and $C_{E,y}$ denote the constants in (4-10) of Lemma 6 associated with changes in $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, and set $C_{E,p} := \max\{C_t, C_\gamma\}$. Lemma 6 gives

$$\mathcal{E}_{k+1} \leq \chi_E \mathcal{E}_k + C_{E,x} X_{k-1} + C_{E,y} Y_{k-1} + C_{E,p} \Delta_{k-1}^2, \qquad \chi_E := 1 - \frac{\tau}{2} \in (0, 1). \tag{9-4}$$

Lemma 12 gives

$$\mathcal{D}_{k+1} \leq \chi_\lambda \mathcal{D}_k + C_{\lambda,z}\,(X_k + Y_k) + C_{\lambda,E} \mathcal{E}_{k+1} + C_{\lambda,p} \Delta_k^2. \tag{9-5}$$

Step 2: Lyapunov weights and step sizes.

Define

$$\Gamma_D := \frac{B_D + 1}{1 - \chi_\lambda}, \tag{9-6}$$

$$P_E := B_E + \Gamma_D C_{\lambda,E}, \qquad \Gamma_E := \frac{P_E \chi_E + 1}{1 - \chi_E}, \tag{9-7}$$

and

$$c_x := (\Gamma_E + P_E) C_{E,x} + 1, \qquad c_y := (\Gamma_E + P_E) C_{E,y} + 1. \tag{9-8}$$

Choose Γ_D, Γ_E, c_x, and c_y in this order, and let the primal step sizes satisfy

$$\begin{aligned} 0 < \alpha \le \bar{\alpha} &:= \frac{\mu_\psi}{\frac{L_\Phi + 1 + G_{y,x}}{2} + \Gamma_D C_{\lambda,z} + c_x + 1}, \\ 0 < \beta \le \bar{\beta} &:= \frac{\mu_\phi}{\frac{L_\Phi + 1}{2} + \Gamma_D C_{\lambda,z} + c_y + 1}. \end{aligned} \tag{9-9}$$

Both upper bounds are positive, and (9-9) implies (8-11). By (8-12),

$$a_x \ge \Gamma_D C_{\lambda,z} + c_x + 1, \qquad a_y \ge \Gamma_D C_{\lambda,z} + c_y + 1. \tag{9-10}$$

By definition of $\mathcal{W}_k$,

$$\begin{aligned} \mathcal{W}_{k+1} - \mathcal{W}_k = {} & \Phi_{k+1}(\boldsymbol{z}_{k+1}) - \Phi_k(\boldsymbol{z}_k) + \Gamma_E(\mathcal{E}_{k+1} - \mathcal{E}_k) + \Gamma_D(\mathcal{D}_{k+1} - \mathcal{D}_k) \\ & + c_x(X_k - X_{k-1}) + c_y(Y_k - Y_{k-1}). \end{aligned} \tag{9-11}$$

Substituting (8-14) and (9-5) into (9-11), with $P_E = B_E + \Gamma_D C_{\lambda,E}$, gives

$$\begin{aligned} \mathcal{W}_{k+1} - \mathcal{W}_k \le {} & \big[B_D - \Gamma_D(1 - \chi_\lambda)\big]\mathcal{D}_k + (P_E + \Gamma_E)\mathcal{E}_{k+1} - \Gamma_E \mathcal{E}_k \\ & + \big[-a_x + \Gamma_D C_{\lambda,z} + c_x\big]X_k + \big[-a_y + \Gamma_D C_{\lambda,z} + c_y\big]Y_k \\ & - c_x X_{k-1} - c_y Y_{k-1} + L_{\Phi,p}\Delta_k + \Gamma_D C_{\lambda,p}\Delta_k^2. \end{aligned} \tag{9-12}$$

The coefficient $P_E + \Gamma_E$ of $\mathcal{E}_{k+1}$ is positive, so (9-4) yields

$$\begin{aligned} \mathcal{W}_{k+1} - \mathcal{W}_k \le {} & \big[B_D - \Gamma_D(1 - \chi_\lambda)\big]\mathcal{D}_k + \big[P_E\chi_E - \Gamma_E(1 - \chi_E)\big]\mathcal{E}_k \\ & + \big[(\Gamma_E + P_E)C_{E,x} - c_x\big]X_{k-1} + \big[(\Gamma_E + P_E)C_{E,y} - c_y\big]Y_{k-1} \\ & + \big[-a_x + \Gamma_D C_{\lambda,z} + c_x\big]X_k + \big[-a_y + \Gamma_D C_{\lambda,z} + c_y\big]Y_k \\ & + L_{\Phi,p}\Delta_k + \Gamma_D C_{\lambda,p}\Delta_k^2 + (\Gamma_E + P_E)C_{E,p}\Delta_{k-1}^2. \end{aligned} \tag{9-13}$$

The definitions in (9-6)–(9-8) give $B_D - \Gamma_D(1 - \chi_\lambda) = -1$, $P_E\chi_E - \Gamma_E(1 - \chi_E) = -1$, and $(\Gamma_E + P_E)C_{E,x} - c_x = (\Gamma_E + P_E)C_{E,y} - c_y = -1$.

By (9-10), the remaining two coefficients are at most -1. Since the parameter sequence lies in the compact set $[0,1] \times [\underline{\gamma}, \gamma_0]$, there exists $\overline{\Delta} > 0$ such that $\Delta_k^2 \le \overline{\Delta}\Delta_k$. Choose

$$C_H \ge L_{\Phi,p} + \Gamma_D C_{\lambda,p}\overline{\Delta} + (\Gamma_E + P_E)C_{E,p}.$$

Then

$$\mathcal{W}_{k+1} - \mathcal{W}_k \le -\big(\mathcal{D}_k + \mathcal{E}_k + X_{k-1} + Y_{k-1} + X_k + Y_k\big) + C_H(\Delta_k + \Delta_{k-1}^2). \tag{9-14}$$

Omitting the nonpositive terms $-X_{k-1}$ and $-Y_{k-1}$ from the right-hand side of (9-14) gives

$$\begin{aligned} \mathcal{W}_{k+1} - \mathcal{W}_k &\le -\big(X_k + Y_k + \mathcal{E}_k + \mathcal{D}_k\big) + C_H(\Delta_k + \Delta_{k-1}^2) \\ &= -\big(\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 + \|\boldsymbol{y}_{k+1} - \boldsymbol{y}_k\|^2 + \mathcal{E}_k + \mathcal{D}_k\big) + C_H(\Delta_k + \Delta_{k-1}^2). \end{aligned}$$

Thus, (9-3) holds with $c = 1$.

Step 3: Summability.

Continuity of f on $\mathcal{Z}$ gives $\inf_{\boldsymbol{z} \in \mathcal{Z}} f(\boldsymbol{z}) > -\infty$. Since (6-6) implies $\mathcal{W}_k \ge \Phi_k(\boldsymbol{z}_k) \ge f(\boldsymbol{z}_k)$, summing (9-3) from 0 to K and using (9-1) yields

$$\sum_{k=0}^{K}(X_k + Y_k + \mathcal{E}_k + \mathcal{D}_k) \le \frac{\mathcal{W}_0 - \inf_{\boldsymbol{z}\in\mathcal{Z}} f(\boldsymbol{z}) + C_H\mathfrak{H}}{c} < \infty. \tag{9-15}$$

This proves (9-3) and the asserted summability. □

To measure stationarity of the envelope and the dual tracking error, define

$$\begin{aligned}\mathcal{R}_\omega(\boldsymbol{z},\lambda) := \ & \operatorname{dist}^2\left(-\nabla_{\boldsymbol{x}}\Phi_{\rho,\omega}(\boldsymbol{z}), \mathcal{N}_{\mathcal{X}}(\boldsymbol{x})\right) + \operatorname{dist}^2\left(-\nabla_{\boldsymbol{y}}\Phi_{\rho,\omega}(\boldsymbol{z}), \mathcal{N}_{\mathcal{Y}}(\boldsymbol{y})\right) \\ & + |\lambda - \lambda^*_\omega(\boldsymbol{z})|^2.\end{aligned} \tag{9-16}$$

By (6-5), $\nabla\Phi_{\rho,\omega}(\boldsymbol{z})$ is the $\boldsymbol{z}$-gradient of the Lagrangian associated with (6) at $t = 1$ and $\gamma = \underline{\gamma}$, evaluated at the effective multiplier $\mu^*_\omega(\boldsymbol{z})$.

Theorem 1 (Convergence Rate and Stationarity)**.** *Suppose Assumptions 1 and 2 hold. Let $\varepsilon_{\mathrm{rel}}, \rho, \omega, \Lambda > 0$ satisfy $0 < \rho\omega < 1$. Let (t_k, γ_k) be generated by* (16) *and satisfy $\left(1 - t_k + \frac{1}{2\gamma_k}\right)\mu_\phi - t_k\rho_g > 0$ for all $k \ge 0$. Then the iterates generated by Algorithm 1 with sufficiently small step sizes $\eta, \alpha, \beta > 0$ satisfy*

$$\min_{1\le k\le K+1} \mathcal{R}_\omega(\boldsymbol{z}_k, \lambda_k) \le \frac{C_\omega}{K+1}. \tag{9-17}$$

for some $C_\omega > 0$ independent of K. Moreover,

$$\mathcal{R}_\omega(\boldsymbol{z}_k, \lambda_k) \longrightarrow 0. \tag{9-18}$$

Consequently, every accumulation point $\bar{\boldsymbol{z}}$ satisfies

$$0 \in \nabla\Phi_{\rho,\omega}(\bar{\boldsymbol{z}}) + \mathcal{N}_{\mathcal{Z}}(\bar{\boldsymbol{z}}).$$

Proof. By Proposition 2, the squared primal increments and the inner and dual tracking errors are summable. We bound the stationarity residual in terms of these quantities.

For the current parameters (t_k, γ_k), define

$$\begin{aligned}\mathcal{R}_{\omega,k}(\boldsymbol{z},\lambda) := \ & \operatorname{dist}^2\left(-\nabla_{\boldsymbol{x}}\Phi_k(\boldsymbol{z}), \mathcal{N}_{\mathcal{X}}(\boldsymbol{x})\right) \\ & + \operatorname{dist}^2\left(-\nabla_{\boldsymbol{y}}\Phi_k(\boldsymbol{z}), \mathcal{N}_{\mathcal{Y}}(\boldsymbol{y})\right) + |\lambda - \lambda^*_{\omega,k}(\boldsymbol{z})|^2.\end{aligned} \tag{9-19}$$

The optimality conditions of (12) and (14) are

$$\begin{aligned}0 &\in \boldsymbol{d}_x^k + \frac{1}{\alpha}\big(\nabla\psi(\boldsymbol{x}_{k+1}) - \nabla\psi(\boldsymbol{x}_k)\big) + \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}_{k+1}), \\ 0 &\in \boldsymbol{d}_y^k + \frac{1}{\beta}\big(\nabla\phi(\boldsymbol{y}_{k+1}) - \nabla\phi(\boldsymbol{y}_k)\big) + \mathcal{N}_{\mathcal{Y}}(\boldsymbol{y}_{k+1}).\end{aligned} \tag{9-20}$$

The first inclusion in (9-20) provides

$$\boldsymbol{n}_x^{k+1} := -\boldsymbol{d}_x^k - \frac{1}{\alpha}\big(\nabla\psi(\boldsymbol{x}_{k+1}) - \nabla\psi(\boldsymbol{x}_k)\big) \in \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}_{k+1}).$$

Since $\boldsymbol{n}_x^{k+1} \in \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}_{k+1})$, the definition of distance gives

$$\begin{aligned}\operatorname{dist}\left(-\nabla_{\boldsymbol{x}}\Phi_k(\boldsymbol{z}_{k+1}),\mathcal{N}_{\mathcal{X}}(\boldsymbol{x}_{k+1})\right) &\le \left\|-\nabla_{\boldsymbol{x}}\Phi_k(\boldsymbol{z}_{k+1})-\boldsymbol{n}_x^{k+1}\right\| \\ &\le \|\boldsymbol{d}_x^k-\nabla_{\boldsymbol{x}}\Phi_k(\boldsymbol{z}_k)\|+\|\nabla_{\boldsymbol{x}}\Phi_k(\boldsymbol{z}_{k+1})-\nabla_{\boldsymbol{x}}\Phi_k(\boldsymbol{z}_k)\| \\ &+\frac{1}{\alpha}\|\nabla\psi(\boldsymbol{x}_{k+1})-\nabla\psi(\boldsymbol{x}_k)\|.\end{aligned} \tag{9-21}$$

By (8-1), L_Φ-Lipschitz continuity of $\nabla\Phi_k$, and L_ψ-Lipschitz continuity of $\nabla\psi$ from Assumption 2, the three terms on the right are bounded by

$$\sqrt{G_{x,D}\mathcal{D}_k+G_{x,E}\mathcal{E}_{k+1}},\qquad L_\Phi\sqrt{X_k+Y_k},\qquad \frac{L_\psi}{\alpha}\sqrt{X_k}$$

respectively. Applying $(a+b+c)^2\le 3(a^2+b^2+c^2)$ gives

$$\begin{aligned}\operatorname{dist}^2\left(-\nabla_{\boldsymbol{x}}\Phi_k(\boldsymbol{z}_{k+1}),\mathcal{N}_{\mathcal{X}}(\boldsymbol{x}_{k+1})\right) &\le 3G_{x,D}\mathcal{D}_k+3G_{x,E}\mathcal{E}_{k+1} \\ &+\left(3L_\Phi^2+\frac{3L_\psi^2}{\alpha^2}\right)X_k+3L_\Phi^2Y_k.\end{aligned} \tag{9-22}$$

For the $\boldsymbol{y}$ component, the second inclusion in (9-20) provides

$$\boldsymbol{n}_y^{k+1}:=-\boldsymbol{d}_y^k-\frac{1}{\beta}\big(\nabla\phi(\boldsymbol{y}_{k+1})-\nabla\phi(\boldsymbol{y}_k)\big)\in\mathcal{N}_{\mathcal{Y}}(\boldsymbol{y}_{k+1}).$$

Since $\boldsymbol{n}_y^{k+1}\in\mathcal{N}_{\mathcal{Y}}(\boldsymbol{y}_{k+1})$, the same argument gives

$$\begin{aligned}\operatorname{dist}\left(-\nabla_{\boldsymbol{y}}\Phi_k(\boldsymbol{z}_{k+1}),\mathcal{N}_{\mathcal{Y}}(\boldsymbol{y}_{k+1})\right) &\le \|\boldsymbol{d}_y^k-\nabla_{\boldsymbol{y}}\Phi_k(\boldsymbol{z}_{k+\frac{1}{2}})\| \\ &+\|\nabla_{\boldsymbol{y}}\Phi_k(\boldsymbol{z}_{k+1})-\nabla_{\boldsymbol{y}}\Phi_k(\boldsymbol{z}_{k+\frac{1}{2}})\|+\frac{1}{\beta}\|\nabla\phi(\boldsymbol{y}_{k+1})-\nabla\phi(\boldsymbol{y}_k)\|.\end{aligned} \tag{9-23}$$

where

$$\|\nabla\Phi_k(\boldsymbol{z}_{k+1})-\nabla\Phi_k(\boldsymbol{z}_{k+\frac{1}{2}})\|\le L_\Phi\sqrt{Y_k},\qquad \|\nabla\phi(\boldsymbol{y}_{k+1})-\nabla\phi(\boldsymbol{y}_k)\|\le L_\phi\sqrt{Y_k}. \tag{9-24}$$

Substituting (8-2) and (9-24) into (9-23), followed by $(a+b)^2\le 2a^2+2b^2$, gives

$$\begin{aligned}\operatorname{dist}^2\left(-\nabla_{\boldsymbol{y}}\Phi_k(\boldsymbol{z}_{k+1}),\mathcal{N}_{\mathcal{Y}}(\boldsymbol{y}_{k+1})\right) &\le 2G_{y,D}\mathcal{D}_k+2G_{y,E}\mathcal{E}_{k+1}+2G_{y,x}X_k \\ &+2\left(L_\Phi+\frac{L_\phi}{\beta}\right)^2Y_k.\end{aligned} \tag{9-25}$$

Define

$$\begin{aligned}H_D&:=3G_{x,D}+2G_{y,D}, & H_E&:=3G_{x,E}+2G_{y,E},\\ H_x&:=3L_\Phi^2+\frac{3L_\psi^2}{\alpha^2}+2G_{y,x}, & H_y&:=3L_\Phi^2+2\left(L_\Phi+\frac{L_\phi}{\beta}\right)^2.\end{aligned} \tag{9-26}$$

By (6-17),

$$|\lambda_{k+1}-\lambda^*_{\omega,k}(\boldsymbol{z}_{k+1})|^2\le 2\mathcal{D}_{k+1}+\frac{2L_{V,p}^2}{\omega^2}\Delta_k^2. \tag{9-27}$$

Adding (9-27) to (9-22) and (9-25), and using (9-19) and (9-26), yields

$$\mathcal{R}_{\omega,k}(\boldsymbol{z}_{k+1},\lambda_{k+1}) \le H_x X_k + H_y Y_k + H_E \mathcal{E}_{k+1} + H_D \mathcal{D}_k + 2\mathcal{D}_{k+1} + \frac{2L_{V,p}^2}{\omega^2}\Delta_k^2, \tag{9-28}$$

The parameter Lipschitz bound in Lemma 11 provides $L_{\nabla\Phi,p} > 0$ such that

$$\|\nabla\Phi_k(\boldsymbol{z}) - \nabla\Phi_{\rho,\omega}(\boldsymbol{z})\| \le L_{\nabla\Phi,p}\big(|1-t_k| + |\gamma_k - \underline{\gamma}|\big). \tag{9-29}$$

The distance function to a nonempty closed set is 1-Lipschitz. Using this property, (9-29), (6-17), and $(a+b)^2 \le 2a^2 + 2b^2$, we obtain a constant $C_T > 0$ such that

$$\mathcal{R}_{\omega}(\boldsymbol{z},\lambda) \le 2\mathcal{R}_{\omega,k}(\boldsymbol{z},\lambda) + C_T\left[(1-t_k)^2 + (\gamma_k - \underline{\gamma})^2\right]. \tag{9-30}$$

Summing (9-4) from 0 to K gives

$$\sum_{k=0}^{K}\mathcal{E}_{k+1} \le \chi_E \sum_{k=0}^{K}\mathcal{E}_k + C_{E,x}\sum_{k=0}^{K}X_{k-1} + C_{E,y}\sum_{k=0}^{K}Y_{k-1} + C_{E,p}\sum_{k=0}^{K}\Delta_{k-1}^2. \tag{9-31}$$

Summing (9-5) gives

$$\sum_{k=0}^{K}\mathcal{D}_{k+1} \le \chi_\lambda \sum_{k=0}^{K}\mathcal{D}_k + C_{\lambda,z}\sum_{k=0}^{K}(X_k + Y_k) + C_{\lambda,E}\sum_{k=0}^{K}\mathcal{E}_{k+1} + C_{\lambda,p}\sum_{k=0}^{K}\Delta_k^2. \tag{9-32}$$

Let

$$S := \frac{\mathcal{W}_0 - \inf f + C_H\mathfrak{H}}{c} + X_{-1} + Y_{-1} + \mathfrak{H}.$$

By (9-15), the partial sums of X_k, Y_k, $\mathcal{E}_k$, and $\mathcal{D}_k$ are bounded by S. Equations (9-31)–(9-32) then provide constants $C_E^\Sigma, C_D^\Sigma > 0$, independent of K, such that

$$\sum_{k=0}^{K}\mathcal{E}_{k+1} \le C_E^\Sigma S, \qquad \sum_{k=0}^{K}\mathcal{D}_{k+1} \le C_D^\Sigma S. \tag{9-33}$$

Summing (9-28) and (9-30) from 0 to K and using (9-1), (9-15), and (9-33), we obtain a constant $C_\omega > 0$, independent of K, such that

$$\sum_{k=0}^{K}\mathcal{R}_{\omega}(\boldsymbol{z}_{k+1},\lambda_{k+1}) \le C_\omega. \tag{9-34}$$

Since the minimum does not exceed the arithmetic mean,

$$\min_{1\le j\le K+1}\mathcal{R}_{\omega}(\boldsymbol{z}_j,\lambda_j) \le \frac{1}{K+1}\sum_{k=0}^{K}\mathcal{R}_{\omega}(\boldsymbol{z}_{k+1},\lambda_{k+1}) \le \frac{C_\omega}{K+1},$$

This proves (9-17). Equation (9-34) also gives $\mathcal{R}_\omega(\boldsymbol{z}_k,\lambda_k) \to 0$, which is (9-18).

Since $\boldsymbol{z}_k \in \mathcal{Z}$, the sequence admits accumulation points. Let $\boldsymbol{z}_{k_j} \to \bar{\boldsymbol{z}}$ be an arbitrary convergent subsequence.

By (9-18),

$$\operatorname{dist}\left(-\nabla_{\boldsymbol{x}}\Phi_{\rho,\omega}(\boldsymbol{z}_{k_j}), \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}_{k_j})\right) \to 0,$$

The same conclusion holds for the $\boldsymbol{y}$ component. Hence there exist

$$\boldsymbol{n}_x^j \in \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}_{k_j}), \qquad \boldsymbol{n}_y^j \in \mathcal{N}_{\mathcal{Y}}(\boldsymbol{y}_{k_j}),$$

such that

$$\nabla_{\boldsymbol{x}} \Phi_{\rho,\omega}(\boldsymbol{z}_{k_j}) + \boldsymbol{n}_x^j \to 0, \qquad \nabla_{\boldsymbol{y}} \Phi_{\rho,\omega}(\boldsymbol{z}_{k_j}) + \boldsymbol{n}_y^j \to 0.$$

By continuity of $\nabla \Phi_{\rho,\omega}$ from Lemma 8, $\boldsymbol{n}_x^j \to -\nabla_{\boldsymbol{x}} \Phi_{\rho,\omega}(\bar{\boldsymbol{z}})$ and $\boldsymbol{n}_y^j \to -\nabla_{\boldsymbol{y}} \Phi_{\rho,\omega}(\bar{\boldsymbol{z}})$. Closedness of the normal-cone graphs therefore gives

$$-\nabla_{\boldsymbol{x}} \Phi_{\rho,\omega}(\bar{\boldsymbol{z}}) \in \mathcal{N}_{\mathcal{X}}(\bar{\boldsymbol{x}}), \qquad -\nabla_{\boldsymbol{y}} \Phi_{\rho,\omega}(\bar{\boldsymbol{z}}) \in \mathcal{N}_{\mathcal{Y}}(\bar{\boldsymbol{y}}).$$

Equivalently,

$$0 \in \nabla \Phi_{\rho,\omega}(\bar{\boldsymbol{z}}) + \mathcal{N}_{\mathcal{Z}}(\bar{\boldsymbol{z}}).$$

□

A.9.1 Negative-entropy kernel

We next consider the negative-entropy kernel. Let $\mathcal{Y}$ be the probability simplex

$$\Delta_m := \left\{ \boldsymbol{y} \in \mathbb{R}_+^m : \mathbf{1}^\top \boldsymbol{y} = 1 \right\},$$

and let $\phi(\boldsymbol{y}) := \sum_{i=1}^m y_i \log y_i$, with $0 \log 0 := 0$.

Proposition 3 (Convergence with the Negative-Entropy Kernel)**.** *In Proposition 2 and Theorem 1, let $\mathcal{Y} = \Delta_m$ and let ϕ be the negative-entropy kernel. Replace the regularity requirements on $\nabla\phi$ and ∇D_ϕ in Assumption 2 by the existence of compact sets $\mathcal{K}_y, \mathcal{K}_\theta \subset \mathrm{ri}(\Delta_m)$ such that*

$$\{\boldsymbol{y}_k : k \geq 0\} \subset \mathcal{K}_y, \qquad \{\boldsymbol{\theta}_k : k \geq 0\} \subset \mathcal{K}_\theta. \tag{9-35}$$

Then the descent estimate (9-3)*, the rate* (9-17)*, the limit* (9-18)*, and the accumulation-point conclusion remain valid. The same conclusion holds when ψ is the negative-entropy kernel on a probability simplex, provided that $\{\boldsymbol{x}_k\}$ is contained in a compact subset of* $\mathrm{ri}(\mathcal{X})$.

Proof. Compactness of $\mathcal{K}_y$ and $\mathcal{K}_\theta$ yields $\delta > 0$ such that $y_i \geq \delta$ and $\theta_i \geq \delta$ for every $\boldsymbol{y} \in \mathcal{K}_y$ and $\boldsymbol{\theta} \in \mathcal{K}_\theta$. On these sets, the negative-entropy kernel satisfies

$$\nabla^2 \phi(\boldsymbol{y}) = \mathrm{Diag}\left(\frac{1}{y_1}, \ldots, \frac{1}{y_m} \right), \qquad I \preceq \nabla^2 \phi(\boldsymbol{y}) \preceq \frac{1}{\delta} I, \tag{9-36}$$

and

$$\|\nabla^2 \phi(\boldsymbol{y}_1) - \nabla^2 \phi(\boldsymbol{y}_2)\| \leq \frac{1}{\delta^2} \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|. \tag{9-37}$$

Equations (9-36)–(9-37) imply strong convexity of ϕ and Lipschitz continuity of $\nabla\phi$ on these sets. Together with the definition of D_ϕ, they also give the required local Lipschitz continuity of ∇D_ϕ. In particular, one may take $\mu_\phi = 1$ and $L_\phi = 1/\delta$. The corresponding Bregman divergence is

$$D_\phi(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i=1}^m u_i \log \frac{u_i}{v_i}. \tag{9-38}$$

It remains to verify that the exact inner solution lies in the relative interior of the simplex. After removing terms independent of $\boldsymbol{\theta}$, the inner problem is equivalent to minimizing

$$tg(\boldsymbol{x}, \boldsymbol{\theta}) + \left(1 - t + \frac{1}{2\gamma}\right) \phi(\boldsymbol{\theta}) - \frac{1}{2\gamma} \langle \nabla\phi(\boldsymbol{y}), \boldsymbol{\theta} \rangle. \tag{9-39}$$

The coefficient $1 - t + 1/(2\gamma)$ in (9-39) is positive. At a point with a zero component, any feasible direction that increases that component has one-sided directional derivative $-\infty$ for the entropy term, whereas the remaining terms have finite directional derivatives. No boundary point can therefore minimize (9-39), and $\boldsymbol{\theta}^*_{\gamma,t}(\boldsymbol{x}, \boldsymbol{y}) \in \mathrm{ri}(\Delta_m)$. The same argument shows that an entropy mirror subproblem with an interior reference point and a finite direction also has an interior minimizer, so its first-order condition is well defined.

The inner objective is continuous on $\mathcal{X} \times \mathcal{K}_y \times \mathcal{P} \times \Delta_m$, and its minimizer is unique. Berge's maximum theorem therefore implies continuity of the solution mapping. Its image is a compact subset of $\mathrm{ri}(\Delta_m)$; enlarging $\mathcal{K}_\theta$ to the convex hull of this image and the original $\mathcal{K}_\theta$ preserves compactness and relative interiority. After reducing δ if necessary, (9-36)–(9-37) hold on the enlarged set and provide the uniform regularity required in Lemmas 1–14. The conclusions of Proposition 2 and Theorem 1 follow. When ψ is the negative-entropy kernel, the same compact-interior argument supplies the corresponding bounds for the $\boldsymbol{x}$-update. □

Remark 1. *The result extends to the negative-entropy choice of ψ on a probability simplex, provided that $\{\boldsymbol{x}_k\}$ remains in a compact subset of $\mathrm{ri}(\mathcal{X})$. The proof applies the preceding compact-interiority argument to the $\boldsymbol{x}$-update.*

A.10 KKT Residual Bounds for the Relaxed Problem

Given $\varepsilon_{\mathrm{rel}} > 0$, the Lagrangian associated with (6) at $t = 1$ and $\gamma = \underline{\gamma}$ is

$$\mathcal{L}(\boldsymbol{z}, \mu) := f(\boldsymbol{z}) + \mu \hat{V}^1_{\underline{\gamma}}(\boldsymbol{z}), \qquad \mu \geq 0.$$

Define its squared KKT residual by

$$\begin{aligned} \mathcal{R}_{\mathrm{KKT}}(\boldsymbol{z}, \mu) := \ & \mathrm{dist}^2\left(-\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{z}, \mu), \mathcal{N}_{\mathcal{X}}(\boldsymbol{x})\right) + \mathrm{dist}^2\left(-\nabla_{\boldsymbol{y}} \mathcal{L}(\boldsymbol{z}, \mu), \mathcal{N}_{\mathcal{Y}}(\boldsymbol{y})\right) \\ & + [\hat{V}^1_{\underline{\gamma}}(\boldsymbol{z})]^2_+ + |\mu \hat{V}^1_{\underline{\gamma}}(\boldsymbol{z})|^2. \end{aligned} \tag{10-1}$$

Lemma 15 (Comparison of Stationarity and KKT Residuals)**.** *For $\boldsymbol{z} \in \mathcal{Z}$ and $\lambda \in [0, \Lambda]$, define the effective multiplier*

$$\mu := [\lambda + \rho \hat{V}^1_{\underline{\gamma}}(\boldsymbol{z})]_+. \tag{10-2}$$

There exist $C_1, C_2 > 0$, depending only on the fixed problem parameters, such that, for every $\boldsymbol{z} \in \mathcal{Z}$ and $\lambda \in [0, \Lambda]$, the condition

$$\lambda^*_\omega(\boldsymbol{z}) < \Lambda, \tag{10-3}$$

implies

$$\mathcal{R}_{\mathrm{KKT}}(\boldsymbol{z}, \mu) \leq C_1 \mathcal{R}_\omega(\boldsymbol{z}, \lambda) + C_2 \omega^2. \tag{10-4}$$

Proof. Let $\lambda^* = \lambda^*_\omega(\boldsymbol{z})$ and $\mu^* := [\lambda^* + \rho \hat{V}^1_{\underline{\gamma}}(\boldsymbol{z})]_+$. By (6-7)–(6-8), there exist $M_{\hat{V}}, M_{\nabla \hat{V}} > 0$ such that $|\hat{V}^1_{\underline{\gamma}}| \leq M_{\hat{V}}$ and $\|\nabla \hat{V}^1_{\underline{\gamma}}\| \leq M_{\nabla \hat{V}}$ on $\mathcal{Z}$.

Nonexpansiveness of projection onto the nonnegative half-line gives

$$|\mu - \mu^*| \leq |\lambda - \lambda^*|. \tag{10-5}$$

By (6-5), $\nabla\Phi_{\rho,\omega}(\boldsymbol{z}) = \nabla f(\boldsymbol{z}) + \mu^*\nabla\hat{V}^1_{\underline{\gamma}}(\boldsymbol{z})$. Since distance to a closed set is 1-Lipschitz in its argument, the $\boldsymbol{x}$ component satisfies

$$\begin{aligned}&\operatorname{dist}\left(-\nabla_{\boldsymbol{x}} f(\boldsymbol{z}) - \mu\nabla_{\boldsymbol{x}}\hat{V}^1_{\underline{\gamma}}(\boldsymbol{z}), \mathcal{N}_{\mathcal{X}}(\boldsymbol{x})\right)\\ &\quad\le \operatorname{dist}\left(-\nabla_{\boldsymbol{x}}\Phi_{\rho,\omega}(\boldsymbol{z}), \mathcal{N}_{\mathcal{X}}(\boldsymbol{x})\right) + |\mu-\mu^*|\|\nabla_{\boldsymbol{x}}\hat{V}^1_{\underline{\gamma}}(\boldsymbol{z})\|.\end{aligned} \tag{10-6}$$

Applying the same argument to the $\boldsymbol{y}$ component, squaring and summing the two inequalities, and using $(a+b)^2 \le 2a^2 + 2b^2$ and (10-5) yields

$$\begin{aligned}&\operatorname{dist}^2(-\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{z},\mu), \mathcal{N}_{\mathcal{X}}(\boldsymbol{x})) + \operatorname{dist}^2(-\nabla_{\boldsymbol{y}}\mathcal{L}(\boldsymbol{z},\mu), \mathcal{N}_{\mathcal{Y}}(\boldsymbol{y}))\\ &\quad\le 2\mathcal{R}_\omega(\boldsymbol{z},\lambda) + 4M^2_{\nabla\hat{V}}|\lambda-\lambda^*|^2 \le (2+4M^2_{\nabla\hat{V}})\mathcal{R}_\omega(\boldsymbol{z},\lambda).\end{aligned} \tag{10-7}$$

We next estimate the feasibility and complementarity residuals. If $\hat{V}^1_{\underline{\gamma}}(\boldsymbol{z}) \le 0$, then (6-4) gives $\lambda^* = \mu^* = 0$, and

$$[\hat{V}^1_{\underline{\gamma}}(\boldsymbol{z})]_+ = 0, \qquad |\mu\hat{V}^1_{\underline{\gamma}}(\boldsymbol{z})|^2 \le M^2_{\hat{V}}|\lambda-\lambda^*|^2. \tag{10-8}$$

Here we used $|\hat{V}^1_{\underline{\gamma}}| \le M_{\hat{V}}$ and (10-5).

If $\hat{V}^1_{\underline{\gamma}}(\boldsymbol{z}) > 0$, condition (10-3) ensures that the upper projection bound is inactive. Equation (6-4) then gives

$$\hat{V}^1_{\underline{\gamma}}(\boldsymbol{z}) = \omega\lambda^*, \qquad 0 < \lambda^* < \Lambda. \tag{10-9}$$

Hence,

$$[\hat{V}^1_{\underline{\gamma}}(\boldsymbol{z})]^2_+ \le \omega^2\Lambda^2. \tag{10-10}$$

Moreover, $\mu^* = (1+\rho\omega)\lambda^*$. Combining (10-5), (10-9), and $\lambda^* \le \Lambda$ yields

$$|\mu\hat{V}^1_{\underline{\gamma}}(\boldsymbol{z})|^2 \le 2|\mu-\mu^*|^2\hat{V}^1_{\underline{\gamma}}(\boldsymbol{z})^2 + 2|\mu^*\hat{V}^1_{\underline{\gamma}}(\boldsymbol{z})|^2 \le 2\omega^2\Lambda^2|\lambda-\lambda^*|^2 + 2(1+\rho\omega)^2\omega^2\Lambda^4. \tag{10-11}$$

Combining (10-7)–(10-11) and choosing

$$C_1 := 2 + 4M^2_{\nabla\hat{V}} + M^2_{\hat{V}} + 2\omega^2\Lambda^2, \qquad C_2 := \Lambda^2 + 2(1+\rho\omega)^2\Lambda^4,$$

yields, in both cases,

$$\mathcal{R}_{\mathrm{KKT}}(\boldsymbol{z},\mu) \le C_1\mathcal{R}_\omega(\boldsymbol{z},\lambda) + C_2\omega^2.$$

□

Lemma 16 (Inactivity of the Upper Projection Bound)**.** *Suppose the conditions of Theorem 1 hold. Assume further that there exists $\sigma > 0$ such that every $\boldsymbol{z} \in \mathcal{Z}$ with $\hat{V}^1_{\underline{\gamma}}(\boldsymbol{z}) \ge \omega\Lambda$ admits a direction $\boldsymbol{d} \in \mathcal{T}_{\mathcal{Z}}(\boldsymbol{z})$ satisfying*

$$\langle\nabla\hat{V}^1_{\underline{\gamma}}(\boldsymbol{z}), \boldsymbol{d}\rangle \le -\sigma, \qquad \|\boldsymbol{d}\| \le 1, \tag{10-12}$$

where $\mathcal{T}_{\mathcal{Z}}(\boldsymbol{z})$ is the tangent cone to $\mathcal{Z}$ at $\boldsymbol{z}$. Let $M_f := \max_{\boldsymbol{z}\in\mathcal{Z}}\|\nabla f(\boldsymbol{z})\|$ and choose $\Lambda > M_f/\sigma$. If $\boldsymbol{z} \in \mathcal{Z}$ satisfies

$$\operatorname{dist}\left(-\nabla\Phi_{\rho,\omega}(\boldsymbol{z}),\mathcal{N}_{\mathcal{Z}}(\boldsymbol{z})\right) < \sigma\Lambda - M_f, \tag{10-13}$$

then $\lambda_\omega^*(\boldsymbol{z}) < \Lambda$.

Proof. Suppose, to the contrary, that $\lambda_\omega^*(\boldsymbol{z}) = \Lambda$. Equation (6-4) then implies $\hat{V}_{\underline{\gamma}}^1(\boldsymbol{z}) \geq \omega\Lambda > 0$, so (10-12) applies. Let

$$\mu^* := [\Lambda + \rho\hat{V}_{\underline{\gamma}}^1(\boldsymbol{z})]_+ = \Lambda + \rho\hat{V}_{\underline{\gamma}}^1(\boldsymbol{z}) \geq \Lambda.$$

Choose $\boldsymbol{n} \in \mathcal{N}_{\mathcal{Z}}(\boldsymbol{z})$ such that

$$\|\nabla\Phi_{\rho,\omega}(\boldsymbol{z}) + \boldsymbol{n}\| = \operatorname{dist}\left(-\nabla\Phi_{\rho,\omega}(\boldsymbol{z}),\mathcal{N}_{\mathcal{Z}}(\boldsymbol{z})\right).$$

By (6-5), $\nabla\Phi_{\rho,\omega} = \nabla f + \mu^*\nabla\hat{V}_{\underline{\gamma}}^1$. Since $\boldsymbol{d} \in \mathcal{T}_{\mathcal{Z}}(\boldsymbol{z})$ and $\boldsymbol{n} \in \mathcal{N}_{\mathcal{Z}}(\boldsymbol{z})$, $\langle \boldsymbol{n}, \boldsymbol{d}\rangle \leq 0$. Together with (10-12), this yields

$$\mu^*\sigma \leq \langle \nabla f(\boldsymbol{z}), \boldsymbol{d}\rangle - \langle \nabla\Phi_{\rho,\omega}(\boldsymbol{z}) + \boldsymbol{n}, \boldsymbol{d}\rangle \leq M_f + \operatorname{dist}\left(-\nabla\Phi_{\rho,\omega}(\boldsymbol{z}),\mathcal{N}_{\mathcal{Z}}(\boldsymbol{z})\right). \tag{10-14}$$

Since $\mu^* \geq \Lambda$, (10-14) implies

$$\sigma\Lambda - M_f \leq \operatorname{dist}\left(-\nabla\Phi_{\rho,\omega}(\boldsymbol{z}),\mathcal{N}_{\mathcal{Z}}(\boldsymbol{z})\right),$$

contradicting (10-13). □

Theorem 2 (KKT Residual Bound for the Relaxed Problem)**.** *Suppose the conditions of Theorem 1 hold, and let* $\mu_k := [\lambda_k + \rho\hat{V}_{\underline{\gamma}}^1(\boldsymbol{z}_k)]_+$. *Assume that there exists* $\sigma > 0$ *such that every* $\boldsymbol{z} \in \mathcal{Z}$ *satisfying* $\hat{V}_{\underline{\gamma}}^1(\boldsymbol{z}) \geq \omega\Lambda$ *admits* $\boldsymbol{d} \in \mathcal{T}_{\mathcal{Z}}(\boldsymbol{z})$ *with* $\|\boldsymbol{d}\| \leq 1$ *and* $\langle \nabla\hat{V}_{\underline{\gamma}}^1(\boldsymbol{z}), \boldsymbol{d}\rangle \leq -\sigma$. *Let* $M_f := \max_{\boldsymbol{z}\in\mathcal{Z}} \|\nabla f(\boldsymbol{z})\|$, *and suppose* $\Lambda > M_f/\sigma$.

Let $k_K \in \arg\min_{1\leq k\leq K+1} \mathcal{R}_\omega(\boldsymbol{z}_k, \lambda_k)$. *Then there exist* $C_1, C_2 > 0$ *and an integer* $K_0 \geq 0$, *all independent of* K, *such that, for every* $K \geq K_0$,

$$\mathcal{R}_{\mathrm{KKT}}(\boldsymbol{z}_{k_K}, \mu_{k_K}) \leq \frac{C_1C_\omega}{K+1} + C_2\omega^2. \tag{10-15}$$

In particular, the violation of the original relaxed constraint satisfies

$$[\hat{V}_{\underline{\gamma}}^1(\boldsymbol{z}_{k_K})]_+ \leq \sqrt{\frac{C_1C_\omega}{K+1} + C_2\omega^2}. \tag{10-16}$$

The full sequence also satisfies

$$\limsup_{k\to\infty} \mathcal{R}_{\mathrm{KKT}}(\boldsymbol{z}_k, \mu_k) \leq C_2\omega^2.$$

Proof. By (9-17) and the definition of k_K,

$$\mathcal{R}_\omega(\boldsymbol{z}_{k_K}, \lambda_{k_K}) \leq \frac{C_\omega}{K+1}.$$

Equation (9-16) and the product normal-cone identity $\mathcal{N}_{\mathcal{Z}}(\boldsymbol{z}) = \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}) \times \mathcal{N}_{\mathcal{Y}}(\boldsymbol{y})$ give

$$\operatorname{dist}\left(-\nabla\Phi_{\rho,\omega}(\boldsymbol{z}_{k_K}),\mathcal{N}_{\mathcal{Z}}(\boldsymbol{z}_{k_K})\right) \leq \sqrt{\frac{C_\omega}{K+1}}. \tag{10-17}$$

Since $\sigma\Lambda - M_f > 0$, choose K_0 so that, for every $K \geq K_0$,

$$\sqrt{\frac{C_\omega}{K+1}} < \sigma\Lambda - M_f. \tag{10-18}$$

By (10-17)–(10-18), Lemma 16 gives $\lambda^*_\omega(\boldsymbol{z}_{k_K}) < \Lambda$. Lemma 15 applied to $(\boldsymbol{z}_{k_K}, \lambda_{k_K})$ then gives (10-15), while (10-16) follows from the feasibility term in (10-1).

Moreover, (9-18) implies

$$\operatorname{dist}\left(-\nabla\Phi_{\rho,\omega}(\boldsymbol{z}_k), \mathcal{N}_{\mathcal{Z}}(\boldsymbol{z}_k)\right) \to 0.$$

Because $\sigma\Lambda - M_f > 0$, Lemma 16 ensures $\lambda^*_\omega(\boldsymbol{z}_k) < \Lambda$ for all sufficiently large k. Applying Lemma 15 to this tail of the sequence and using (9-18) gives

$$\limsup_{k\to\infty} \mathcal{R}_{\mathrm{KKT}}(\boldsymbol{z}_k, \mu_k) \leq C_2\omega^2.$$

□

A.11 Vanishing Regularization

We consider stagewise vanishing dual regularization with fixed $\varepsilon_{\mathrm{rel}} > 0$.

Let $\omega_s = \omega_0\chi^s$ with $0 < \chi < 1$. At stage s, run the algorithm with ω_s fixed, warm-started from the preceding stage, until an output $(\boldsymbol{z}^s, \lambda^s)$ satisfies

$$\mathcal{R}_{\omega_s}(\boldsymbol{z}^s, \lambda^s) \leq c_0\omega_s^2. \tag{11-1}$$

Corollary 1 (KKT Convergence as $\omega \to 0$)**.** *Suppose every stage satisfies the conditions of Theorem 1, and the descent condition of Lemma 16 holds uniformly with the same $\sigma > 0$. Choose $\Lambda > M_f/\sigma$ and define $\mu^s := [\lambda^s + \rho\hat{V}^1_{\underline{\gamma}}(\boldsymbol{z}^s)]_+$. Then*

$$\mathcal{R}_{\mathrm{KKT}}(\boldsymbol{z}^s, \mu^s) \longrightarrow 0. \tag{11-2}$$

Furthermore, for every accumulation point $\bar{\boldsymbol{z}}$, there exists $\bar{\mu} \geq 0$ such that

$$\begin{cases} 0 \in \nabla_{\boldsymbol{x}} f(\bar{\boldsymbol{z}}) + \bar{\mu}\nabla_{\boldsymbol{x}}\hat{V}^1_{\underline{\gamma}}(\bar{\boldsymbol{z}}) + \mathcal{N}_{\mathcal{X}}(\bar{\boldsymbol{x}}), \\ 0 \in \nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{z}}) + \bar{\mu}\nabla_{\boldsymbol{y}}\hat{V}^1_{\underline{\gamma}}(\bar{\boldsymbol{z}}) + \mathcal{N}_{\mathcal{Y}}(\bar{\boldsymbol{y}}), \\ \hat{V}^1_{\underline{\gamma}}(\bar{\boldsymbol{z}}) \leq 0, \qquad \bar{\mu}\hat{V}^1_{\underline{\gamma}}(\bar{\boldsymbol{z}}) = 0. \end{cases} \tag{11-3}$$

Proof. By (11-1),

$$\operatorname{dist}\left(-\nabla\Phi_{\rho,\omega_s}(\boldsymbol{z}^s), \mathcal{N}_{\mathcal{Z}}(\boldsymbol{z}^s)\right) \leq \sqrt{c_0}\,\omega_s \to 0.$$

Since $\sigma\Lambda - M_f > 0$, Lemma 16 ensures $\lambda^*_{\omega_s}(\boldsymbol{z}^s) < \Lambda$ for all sufficiently large s. The function $\hat{V}^1_{\underline{\gamma}}$ and its gradient are bounded on $\mathcal{Z}$, and $0 < \omega_s \leq \omega_0$; hence the constants in Lemma 15 can be chosen uniformly in s. Lemma 15 and (11-1) yield

$$\mathcal{R}_{\mathrm{KKT}}(\boldsymbol{z}^s, \mu^s) \leq (C_1c_0 + C_2)\omega_s^2 \to 0. \tag{11-4}$$

Let $\boldsymbol{z}^{s_j} \to \bar{\boldsymbol{z}}$ be any convergent subsequence.

Because $\lambda^{s_j} \in [0, \Lambda]$ and $\hat{V}^1_{\underline{\gamma}}(\boldsymbol{z}^{s_j})$ is bounded, the sequence $\mu^{s_j} = [\lambda^{s_j} + \rho\hat{V}^1_{\underline{\gamma}}(\boldsymbol{z}^{s_j})]_+$ is bounded. Passing to a further subsequence, without relabeling, gives

$$\mu^{s_j} \to \bar{\mu} \geq 0. \tag{11-5}$$

The feasibility and complementarity terms in (11-4) imply

$$[\hat{V}_{\underline{\gamma}}^1(\boldsymbol{z}^{s_j})]_+ \to 0, \qquad |\mu^{s_j}\hat{V}_{\underline{\gamma}}^1(\boldsymbol{z}^{s_j})| \to 0.$$

Continuity of $\hat{V}_{\underline{\gamma}}^1$ and (11-5) therefore give

$$\hat{V}_{\underline{\gamma}}^1(\bar{\boldsymbol{z}}) \leq 0, \qquad \bar{\mu}\hat{V}_{\underline{\gamma}}^1(\bar{\boldsymbol{z}}) = 0. \tag{11-6}$$

For the $\boldsymbol{x}$ component, the definition of distance provides $\boldsymbol{n}_x^j \in \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^{s_j})$ such that

$$\left\| \nabla_{\boldsymbol{x}} f(\boldsymbol{z}^{s_j}) + \mu^{s_j} \nabla_{\boldsymbol{x}} \hat{V}_{\underline{\gamma}}^1(\boldsymbol{z}^{s_j}) + \boldsymbol{n}_x^j \right\| \to 0. \tag{11-7}$$

Assumption 1, Lemma 3, (11-5), and (11-7) imply

$$\boldsymbol{n}_x^j \to -\nabla_{\boldsymbol{x}} f(\bar{\boldsymbol{z}}) - \bar{\mu}\nabla_{\boldsymbol{x}} \hat{V}_{\underline{\gamma}}^1(\bar{\boldsymbol{z}}).$$

The normal-cone mapping of a closed convex set has a closed graph. Since $\boldsymbol{x}^{s_j} \to \bar{\boldsymbol{x}}$,

$$-\nabla_{\boldsymbol{x}} f(\bar{\boldsymbol{z}}) - \bar{\mu}\nabla_{\boldsymbol{x}} \hat{V}_{\underline{\gamma}}^1(\bar{\boldsymbol{z}}) \in \mathcal{N}_{\mathcal{X}}(\bar{\boldsymbol{x}}). \tag{11-8}$$

The same argument for the $\boldsymbol{y}$ component gives

$$-\nabla_{\boldsymbol{y}} f(\bar{\boldsymbol{z}}) - \bar{\mu}\nabla_{\boldsymbol{y}} \hat{V}_{\underline{\gamma}}^1(\bar{\boldsymbol{z}}) \in \mathcal{N}_{\mathcal{Y}}(\bar{\boldsymbol{y}}). \tag{11-9}$$

Together, (11-6), (11-8), and (11-9) give the KKT conditions in (11-3). Since the convergent subsequence was arbitrary, every accumulation point admits a multiplier satisfying these conditions. □

A.12 Additional Experimental Details

This section provides additional implementation details for the real-world experiments. For each task, all methods use the same data splits, model architecture, and computational budget. Table 4 summarizes the task configurations, while Table 5 reports the principal parameter settings of GaCF.

Table 4: Dataset and model configurations for the real-world experiments.

Task	Dataset configuration	Model configuration
Data hyper-cleaning	MNIST and FashionMNIST; 5,000/5,000/10,000 training, validation, and test samples; 50% training-label corruption	Two-layer MLP with dimensions 784–250–10 and ReLU activation; one learnable weight for each training sample
Few-shot learning	Omniglot; 5-way and 10-way 1-shot classification; 15 query samples per class	Four convolutional blocks with 64 channels, followed by a linear classifier; meta-batch size 16
Parameter-efficient fine-tuning	SST-2 with the standard training, validation, and test partitions	RoBERTa-base with DoRA; rank 4, scaling factor 8, and maximum sequence length 128

Table 5: Principal parameter settings of GaCF.

Parameter	Data hyper-cleaning	Few-shot learning	PEFT
α	3×10^{-1}	10^{-3}	2×10^{-5}
β	3×10^{-1}	10^{-2}	3×10^{-5}
η	3×10^{-3}	10^{-3}	10^{-5}
Δt	2×10^{-2}	2×10^{-2}	8×10^{-4}
γ_0	1	1	1
$\underline{\gamma}$	3×10^{-2}	10^{-2}	10^{-2}
$\Delta\gamma$	2×10^{-2}	2×10^{-2}	10^{-3}
Constraint penalty	10^{-1}	10^{-2}	10^{-1}
Initial dual variable	0.2	0.1	0.1
Random seed	40	0	0

A.13 Sensitivity Analysis

We complement the experiments in Section 4 with a sensitivity study on the LL-convex synthetic problem, for which the reference solution is available. We examine the effects of the potential function, the continuation increments Δt and $\Delta\gamma$, and the penalty parameter ρ. In each experiment, only the indicated component is varied, while the initialization, stopping criterion, and all remaining parameters are held fixed.

The potential function determines the mirror geometry and is chosen to reflect the feasible set or variable scaling while retaining an efficient mirror update. We compare negative entropy, the Euclidean quadratic potential, and a weighted quadratic potential. The increments Δt and $\Delta\gamma$ control the variation between consecutive surrogates. In the corresponding experiments, we vary $\Delta t \in \{0.1, 0.05, 0.01\}$ and $\Delta\gamma \in \{0.1, 0.05, 0.01\}$ separately. The penalty parameter ρ controls the strength of constraint correction: a small value may slow this correction, whereas a large value can produce overly aggressive primal updates through the effective multiplier.

Table 6: Sensitivity analysis on the LL-convex synthetic problem. Only the component indicated in each block is varied.

Choice	Final f	$\|\boldsymbol{x} - \boldsymbol{x}^*\|/\|\boldsymbol{x}^*\|$	Iterations
Potential function			
Negative entropy	2.14×10^{-4}	2.48×10^{-4}	220
Euclidean quadratic	3.29×10^{-2}	1.46×10^{-2}	191
Weighted quadratic	4.44×10^{-1}	3.57×10^{-2}	90
Continuation increment Δt			
0.1	3.20×10^{-4}	3.50×10^{-4}	219
0.05	2.48×10^{-4}	2.04×10^{-4}	223
0.01	2.14×10^{-4}	2.48×10^{-4}	220
Continuation increment $\Delta\gamma$			
0.1	4.08×10^{-5}	2.93×10^{-4}	314
0.05	5.34×10^{-5}	3.11×10^{-4}	323
0.01	2.14×10^{-4}	2.48×10^{-4}	220
Penalty parameter ρ			
10^{-3}	2.64×10^{-4}	5.99×10^{-5}	245
10^{-2}	2.14×10^{-4}	2.48×10^{-4}	220
10^{-1}	5.73×10^{-4}	4.51×10^{-3}	171

Table 6 shows that negative entropy yields the lowest objective value and relative solution error among the tested potentials, consistent with the simplex geometry of the problem. The results remain stable

Table 7: Component ablation on the nonconvex synthetic problem with $\kappa = 10^4$. Results are averaged over 3 paired held-out seeds under an identical budget of 1000 iterations. Lower values are better.

Method	Upper-level objective	LL residual	Reference distance
Euclidean, no continuation	11.856	0.3770	0.6906
Continuation only	11.760	0.2038	0.6262
Geometry only	7.762	0.4706	0.5933
GaCF	**7.049**	**0.0602**	**0.3456**

across the tested continuation increments, although no monotone dependence on either increment is observed. For ρ, the intermediate choice provides a favorable balance between the final objective and convergence speed, whereas the largest value terminates earlier at the expense of a larger solution error.

A.14 Ablation Study

To examine the contributions of continuation and geometry adaptation, we conduct a component-wise ablation study on the nonconvex synthetic problem with $\kappa = 10^4$. We compare GaCF with three variants: a Euclidean baseline without continuation, a continuation-only variant using Euclidean geometry, and a geometry-only variant with the continuation mechanism disabled. All variants are evaluated using the same initialization for each of 3 held-out random seeds and an identical budget of 1000 iterations. As reported in Table 7, the two individual components provide different benefits. Continuation alone reduces the lower-level stationarity residual, while geometry adaptation improves the upper-level objective and the distance to the reference solution. Combining them yields the most consistent performance across all three metrics. GaCF achieves an upper-level objective of 7.049, a lower-level residual of 0.0602, and a reference distance of 0.3456. Compared with the continuation-only variant, it reduces the lower-level residual by 70.4% and the reference distance by 44.8%. These results suggest that continuation and geometry-aware updates play complementary roles in navigating the ill-conditioned nonconvex landscape.