SHARPNESS-AWARE BILEVEL AND MINIMAX OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Bilevel and minimax optimization are foundational frameworks of nested optimization widely used in hyperparameter tuning, adversarial training, meta-learning, and reinforcement learning. Despite their versatility, these methods frequently suffer from poor generalization due to sharp minima and training instability, especially in overparameterized regimes. We introduce a unified first-order sharpness-aware optimization framework comprising Sharpness-Aware Bilevel Optimization (SABO) and Sharpness-Aware Minimax Optimization (SAMO). SABO improves generalization by jointly minimizing inner and outer objectives and their sharpness via a penalty-based reformulation that avoids costly implicit gradient computations. SAMO also yields a simplified algorithm tailored for sharpness-aware minimax settings. We establish theoretical convergence guarantees under the Polyak–Łojasiewicz condition and derive generalization bounds. Experiments on synthetic and real tasks show that SABO and SAMO consistently enhance generalization over standard baselines.

1 INTRODUCTION

Understanding why overparameterized models, whose parameters vastly outnumber training samples, generalize remains a core challenge in modern machine learning. Classical learning theory predicts overfitting in this regime (Hastie et al., 2009), yet such models often generalize well even when they perfectly fit the training data (Zhang et al., 2021). This tension (Belkin et al., 2019) has spurred extensive work relating loss-landscape geometry to generalization (Jiang et al., 2019; Keskar et al., 2016). Among these geometric properties, the *sharpness* of a minimum, i.e., how rapidly the loss changes around it, has emerged as a key predictor of generalization (Chaudhari et al., 2019; Yun et al., 2019): *flat* minima, stable under small perturbations, tend to generalize better, whereas sharp ones signal sensitivity to training-specific noise (Andriushchenko et al., 2024; Foret et al., 2020). Sharpness-Aware Minimization (SAM) (Foret et al., 2020) builds this insight into the objective, jointly minimizing training loss and sharpness on PAC–Bayesian grounds at first-order cost: sharpness is estimated with a single gradient-ascent step rather than Hessian- or sample-based surrogates (Chaudhari et al., 2019). SAM is effective in computer vision (Foret et al., 2020; Kwon et al., 2021; Yun et al., 2019; Zhong et al., 2022), language models (Zhong et al., 2022), and meta-learning (Abbas et al., 2022).

Yet most geometry-based generalization techniques, SAM included, are restricted to single-level optimization, while modern machine learning increasingly relies on *nested* problems with coupled, interdependent levels, as in hyperparameter optimization (Franceschi et al., 2018), neural architecture search (Liu et al., 2018), and reinforcement learning from human feedback (Shen et al., 2024). The two most prominent nested frameworks are *bilevel optimization* (BO), which minimizes an outer objective subject to the solution of an inner problem, and *minimax optimization* (MO), which minimizes a loss against an adversarial maximizer. Although overparameterization affects both levels, standard sharpness-aware methods do not directly apply to these structures, leaving a gap in obtaining flat, generalizable solutions for nested settings.

Contributions.

I. We introduce a unified, first-order framework comprising *Sharpness-Aware Bilevel Optimization* (SABO) and its minimax specialization, *Sharpness-Aware Minimax Optimization*

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.

(SAMO), controlling landscape geometry at every level. SABO rests on a novel *penalty-based* reformulation that bypasses implicit gradients and second-order terms entirely, and so stays efficient under fully *nonconvex* inner and outer objectives; SAMO exploits the adversarial structure of minimax games to yield a markedly simpler algorithm.

II. We establish theoretical guarantees under the Polyak–Łojasiewicz (PL) condition, proving an $\mathcal{O}(1/K)$ convergence rate (over K iterations) for SABO (Theorem 3.1) and SAMO (Theorem 4.3). We also derive generalization bounds by extending PAC–Bayesian theory to bilevel and minimax settings in Theorems D.1 and D.2 (Appendix D), showing that sharpness estimation of both the inner and outer objectives strengthens these bounds.

III. We validate both frameworks on synthetic and large-scale tasks: SABO and SAMO consistently find flatter minima and generalize better than standard baselines.

2 Sharpness-Aware Bilevel Optimization

Consider the bilevel optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^{d_1},\mathbf{y}\in\mathbb{R}^{d_2}} f(\mathbf{x},\mathbf{y}) \quad \text{subj. to} \quad \mathbf{y}\in\arg\min_{\mathbf{y}\in\mathbb{R}^{d_2}} g(\mathbf{x},\mathbf{y}). \tag{1}$$

Here, $f:\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}\to\mathbb{R}$ and $g:\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}\to\mathbb{R}$ denote the outer and inner objectives, respectively.

Standard BO frameworks, such as (1), are highly susceptible to converging toward sharp local minima, which ultimately results in poor generalization. We propose *sharpness-aware BO*, which seeks flatness at both levels by introducing random weight perturbation vectors $\boldsymbol{\epsilon}\in\mathbb{R}^{d_1}$ and $\boldsymbol{\nu}\in\mathbb{R}^{d_2}$ constrained within ℓ_2-balls of radii $\rho>0$ and $\eta>0$:

$$\begin{aligned}\min_{\mathbf{x}\in\mathbb{R}^{d_1},\,\mathbf{y}\in\mathbb{R}^{d_2}} \quad & \max_{\|\boldsymbol{\epsilon}\|\le\rho,\,\|\boldsymbol{\nu}\|\le\eta} f(\mathbf{x}+\boldsymbol{\epsilon},\,\mathbf{y}+\boldsymbol{\nu}) \\ \text{subj. to} \quad & \mathbf{y}\in\arg\min_{\mathbf{y}\in\mathbb{R}^{d_2}} \max_{\|\boldsymbol{\nu}\|\le\eta} g(\mathbf{x},\,\mathbf{y}+\boldsymbol{\nu}).\end{aligned} \tag{2}$$

The upper- and lower-level sharpness, respectively, are

$$\max_{\|\boldsymbol{\epsilon}\|\le\rho,\,\|\boldsymbol{\nu}\|\le\eta}\{f(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}+\boldsymbol{\nu})-f(\mathbf{x},\mathbf{y})\}, \qquad \max_{\|\boldsymbol{\nu}\|\le\eta}\{g(\mathbf{x},\mathbf{y}+\boldsymbol{\nu})-g(\mathbf{x},\mathbf{y})\}.$$

Thus, (2) minimizes both the objective and the sharpness at each level: the outer level reduces $f(\mathbf{x},\mathbf{y})$ and its perturbation sensitivity, while the inner level ensures that $\mathbf{y}$ minimizes $g(\mathbf{x},\mathbf{y})$ and its sharpness. This drives the solutions toward flat regions of both landscapes (see Figure 1). In the single-level case, (2) reduces to SAM (Foret et al., 2020):

$$\min_{\mathbf{x}\in\mathbb{R}^{d_1}} \max_{\|\boldsymbol{\epsilon}\|\le\rho} f(\mathbf{x}+\boldsymbol{\epsilon}), \tag{3}$$

which minimizes the worst-case objective under perturbations of radius ρ.

2.1 Penalty Reformulation for SABO

Single-level SAM (Foret et al., 2020) estimates sharpness with one first-order ascent step. The bilevel minimax problem (2), however, is nested and typically requires implicit differentiation with (approximate) second-order information, making direct SAM-style updates infeasible. Penalty reformulations (Domke, 2012; Ghadimi & Wang, 2018; Grazzi et al., 2020; Ji et al., 2021a; Pedregosa, 2016) reduce Hessian cost but do not address (2), and second-order methods that discourage sharp minima (Bartlett et al., 2023) do not scale. We instead reformulate (2) via a *value-function penalty* (Liu et al., 2022a; 2021a) and solve it with first-order steps: (i) an inner sharpness-aware problem, and (ii) outer Lagrangian updates penalizing inner optimality gaps.

2.1.1 Stage I: Sharpness-Aware Inner Optimization

We approximate the inner optimum $\mathbf{y}^*$ and the inner perturbation variable $\boldsymbol{\nu}^*$ using multiple SAM-style steps. Let

$$\boldsymbol{\nu}^*(\mathbf{x},\mathbf{y}) := \arg\max_{\|\boldsymbol{\nu}\|\le\eta} g(\mathbf{x},\mathbf{y}+\boldsymbol{\nu}). \tag{4}$$

Then, (2) is equivalent to

$$\min_{\mathbf{x}\in\mathbb{R}^{d_1},\mathbf{y}\in\mathbb{R}^{d_2}} \max_{\|\boldsymbol{\epsilon}\|\le\rho,\|\boldsymbol{\nu}\|\le\eta} f(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}+\boldsymbol{\nu}) \quad \text{subj. to} \quad \mathbf{y}\in\mathcal{S}(\mathbf{x}) := \arg\min_{\mathbf{y}\in\mathbb{R}^{d_2}} g(\mathbf{x},\mathbf{y}+\boldsymbol{\nu}^*(\mathbf{x},\mathbf{y})), \tag{5}$$

where $\mathcal{S}(\mathbf{x})$ is nonempty and closed.

The constraint in (5) requires that, for each $\mathbf{x}+\boldsymbol{\epsilon}$ and the corresponding inner perturbation $\boldsymbol{\nu}^*(\mathbf{x},\mathbf{y})$, the inner variable $\mathbf{y}$ be an optimizer of $g(\mathbf{x},\,\cdot+\boldsymbol{\nu}^*)$. When the arguments are clear from the context, we simply write $\boldsymbol{\nu}^* := \boldsymbol{\nu}^*(\mathbf{x},\mathbf{y})$. A convenient way to test whether a given $\mathbf{y}$ satisfies this requirement is to compare its objective value with the best achievable value. Define the *perturbed value function*

$$g^*(\mathbf{x}) := \min_{\mathbf{y}\in\mathbb{R}^{d_2}} g(\mathbf{x},\,\mathbf{y}+\boldsymbol{\nu}^*(\mathbf{x},\mathbf{y})), \tag{6}$$

and the corresponding inner optimality gap

$$g(\mathbf{x},\,\mathbf{y}+\boldsymbol{\nu}^*(\mathbf{x},\mathbf{y})) - g^*(\mathbf{x}). \tag{7}$$

The bilevel feasibility condition reduces to a single scalar constraint requiring the optimality gap to be zero. By the nonnegativity of the optimality gap, we have $\mathbf{y}\in\mathcal{S}(\mathbf{x})$ *if and only if* the optimality gap in (7) is nonpositive. Hence, (5) is equivalent to

$$\min_{\mathbf{x}\in\mathbb{R}^{d_1},\,\mathbf{y}\in\mathbb{R}^{d_2}} \max_{\|\boldsymbol{\epsilon}\|\le\rho,\,\|\boldsymbol{\nu}\|\le\eta} f(\mathbf{x}+\boldsymbol{\epsilon},\,\mathbf{y}+\boldsymbol{\nu}) \quad \text{subj. to} \quad g(\mathbf{x},\,\mathbf{y}+\boldsymbol{\nu}^*(\mathbf{x},\mathbf{y})) - g^*(\mathbf{x}) \;\le\; 0, \tag{8}$$

which compactly encodes that $\mathbf{y}$ is (perturbed) inner-optimal. The inequality form is useful in Appendix C.3: we relax it to $g(\mathbf{x},\,\mathbf{y}+\boldsymbol{\nu}^*(\mathbf{x},\mathbf{y})) - g^*(\mathbf{x}) \le \zeta_\gamma$ (for a small $\zeta_\gamma > 0$) to allow strict feasibility, and we directly penalize it in a first-order scheme.

To approximate g^*, and hence the optimality gap in (7), we need tractable approximations of $\boldsymbol{\nu}^*$ and a minimizer $\mathbf{y}^*\in\mathcal{S}(\mathbf{x})$. For radius $\eta>0$, a first-order expansion of g gives the SAM-style (Foret et al., 2020) inner perturbation (Lemma B.2, Appendix B.1)

$$\tilde{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y}) = \eta\frac{\nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y})}{\|\nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y})\|}. \tag{9}$$

Given the outer iterate $(\mathbf{x}_k,\mathbf{y}_k)$, we run T_k inner steps on $g(\mathbf{x}_k,\cdot)$ to approximate $\mathbf{y}^*$. Set $\mathbf{z}_1^{(k)} := \mathbf{y}_k$, apply (9), and update as follows:

$$\tilde{\boldsymbol{\nu}}_t^{(k)} := \tilde{\boldsymbol{\nu}}(\mathbf{x}_k,\mathbf{z}_t^{(k)}), \tag{10a}$$

$$\mathbf{z}_{t+1}^{(k)} = \mathbf{z}_t^{(k)} - \beta\,\nabla_{\mathbf{z}} g\Big(\mathbf{x}_k,\,\mathbf{z}_t^{(k)} + \tilde{\boldsymbol{\nu}}_t^{(k)}\Big), \quad t=1,\dots,T_k, \tag{10b}$$

$$\tilde{\boldsymbol{\nu}}_k := \tilde{\boldsymbol{\nu}}_{T_k}^{(k)}, \qquad \tilde{\mathbf{y}}_k := \mathbf{z}_{T_k+1}^{(k)}. \tag{10c}$$

When $T_k = 1$, this reduces to a single SAM-style step, i.e., the inner gradient is evaluated at a nearby point within a radius η, which encourages flatness in the inner landscape.

2.1.2 Stage II: Sharpness-Aware Lagrangian Minimization

We jointly update $(\mathbf{x},\mathbf{y})$ by minimizing a Lagrangian that adds an inner optimality-gap penalty to the perturbed outer objective. Define the perturbed Lagrangian

$$\mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu}) := f(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}+\boldsymbol{\nu}) + \gamma\left(g(\mathbf{x},\mathbf{y}+\boldsymbol{\nu}^*(\mathbf{x},\mathbf{y})) - g^*(\mathbf{x})\right), \tag{11}$$

where the penalty constant $\gamma>0$ balances the outer objective and the inner optimality gap defined in (7), and $\boldsymbol{\nu}^*(\mathbf{x},\mathbf{y})$ is defined in (4). SABO solves the minimax problem

$$\min_{\mathbf{x}\in\mathbb{R}^{d_1},\mathbf{y}\in\mathbb{R}^{d_2}} \max_{\|\boldsymbol{\epsilon}\|\le\rho,\|\boldsymbol{\nu}\|\le\eta} \mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu}). \tag{12}$$

The Lagrangian $\mathcal{L}_\gamma$ generalizes (8) to the Lagrangian setting: the first term in (11) controls the perturbed outer objective, while the penalty drives the inner suboptimality $g-g^*$ to zero, aligning $\mathbf{y}$ with the perturbed inner optimum. The main technical challenge is the potential non-smoothness of the gradient term $\nabla g^*(\mathbf{x})$. However, under the Polyak–Łojasiewicz (PL) condition (Polyak, 1964), this term becomes tractable.

Algorithm 1 Sharpness-Aware Bilevel Optimization (**SABO**)

Require: Init $(\mathbf{x}_1, \mathbf{y}_1)$; step sizes $\alpha, \beta > 0$; radii $\rho, \eta > 0$; penalty $\gamma > 0$; inner steps $\{T_k\}$; outer iterations K.

1: **for** $k = 1$ **to** K **do**
2: **Stage I:** From $\mathbf{y}_k$, run T_k inner steps of (10b) with perturbations from (9) to get $(\tilde{\mathbf{y}}_k, \tilde{\boldsymbol{\nu}}_k)$ as in (10c).
3: **Stage II:** Form $(\hat{\boldsymbol{\epsilon}}_k, \hat{\boldsymbol{\nu}}_k)$ via (14) and update $(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$ using (15).
4: **end for**
5: **return** $(\mathbf{x}_{K+1}, \mathbf{y}_{K+1})$.

Algorithm 2 Sharpness-Aware Minimax Optimization (**SAMO**)

Require: Initialization $(\mathbf{x}_1, \mathbf{y}_1) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$; step sizes $\alpha, \beta > 0$; perturbation radii $\rho, \eta > 0$; inner steps $\{T_k\}$; total iterations K

1: **for** $k = 1$ **to** K **do**
2: **Stage I:** From $\mathbf{y}_k$, run T_k inner steps of (19b) with perturbations from (19a) to get $(\boldsymbol{\epsilon}_k, \boldsymbol{\nu}_k, \mathbf{y}_{k+1})$ as in (19c).
3: **Stage II:** Update $\mathbf{x}_{k+1}$ using (20).
4: **end for**
5: **return** $\mathbf{x}_{K+1}$.

Assumption 2.1 (μ–PL Condition)**.** *There exists a constant $\mu > 0$ such that, for any given $\mathbf{x} \in \mathbb{R}^{d_1}$, the function $g(\mathbf{x}, \cdot)$ satisfies*

$$\|\nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y})\|^2 \;\geq\; \mu^{-1}\Big(g(\mathbf{x}, \mathbf{y}) - \min_{\mathbf{y}' \in \mathbb{R}^{d_2}} g(\mathbf{x}, \mathbf{y}')\Big), \qquad \forall \mathbf{y} \in \mathbb{R}^{d_2}. \tag{13}$$

Assumption 2.2. *Let $\mathbf{w} = [\mathbf{x}; \mathbf{y}] \in \mathbb{R}^{d_1+d_2}$ and $\mathbf{w}' = [\mathbf{x}'; \mathbf{y}'] \in \mathbb{R}^{d_1+d_2}$, where $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{d_1}$ and $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^{d_2}$. There exists a constant $\ell_{g,1} \in \mathbb{R}_{++}$ such that, for all $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^{d_1+d_2}$,*

$$\|\nabla g(\mathbf{w}) - \nabla g(\mathbf{w}')\| \;\leq\; \ell_{g,1} \|\mathbf{w} - \mathbf{w}'\|.$$

Assumption 2.3. *Let $\mathbf{w} = [\mathbf{x}; \mathbf{y}] \in \mathbb{R}^{d_1+d_2}$ and $\mathbf{w}' = [\mathbf{x}'; \mathbf{y}'] \in \mathbb{R}^{d_1+d_2}$, where $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{d_1}$ and $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^{d_2}$. For any $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^{d_1+d_2}$:*

I. *There exists $\ell_{f,0} \in \mathbb{R}_{++}$ such that $\big|f(\mathbf{w}) - f(\mathbf{w}')\big| \;\leq\; \ell_{f,0} \|\mathbf{w} - \mathbf{w}'\|$;*

II. *There exists $\ell_{f,1} \in \mathbb{R}_{++}$ such that $\|\nabla f(\mathbf{w}) - \nabla f(\mathbf{w}')\| \;\leq\; \ell_{f,1} \|\mathbf{w} - \mathbf{w}'\|$.*

Assumption 2.4. *The function $f(\mathbf{x}, \mathbf{y})$ is bounded below, i.e., $\inf_{(\mathbf{x},\mathbf{y})} f(\mathbf{x}, \mathbf{y}) > -\infty$.*

Assumptions 2.2, 2.3, and 2.4 are standard regularity conditions used in convergence analyses of BO (Ghadimi & Wang, 2018; Ji et al., 2020). The PL condition (13) is weaker than strong convexity and does not require convexity or uniqueness of the inner minimizer. It is widely used to provide optimization guarantees in modern overparameterized settings (Frei & Gu, 2021; Song et al., 2021; Yang et al., 2020). It ensures linear (gradient-dominated) convergence of the inner iterates and renders the value function g^* smooth and first-order tractable.

Next, we approximate the gradients of $\mathcal{L}_\gamma$ to solve (12). Following Lemma C.1 in Appendix C.1, at outer iteration k, we update $(\boldsymbol{\epsilon}, \boldsymbol{\nu})$ based on the approximation $\tilde{\mathbf{y}}_k$ from (10) by setting

$$\hat{\boldsymbol{\nu}}_k := \eta \frac{\nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k)}{\|\nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k)\|}, \quad \hat{\boldsymbol{\epsilon}}_k := \rho \frac{\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k)}{\|\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k)\|}. \tag{14}$$

The construction in (14) ensures that the outer perturbation $\hat{\boldsymbol{\epsilon}}_k$ incorporates both the sharpness-aware penalty and the discrepancy between the current inner iterate $\mathbf{y}_k$ and its proxy $\tilde{\mathbf{y}}_k$. Consequently, the update not only minimizes the perturbed outer objective but also steers the inner variable toward its approximate optimum.

We take $\tilde{\mathbf{y}}_k, \tilde{\boldsymbol{\nu}}_k$ from (10c) and $\hat{\boldsymbol{\epsilon}}_k, \hat{\boldsymbol{\nu}}_k$ from (14), define the corresponding gradient estimates of $\nabla \mathcal{L}_\gamma$, and apply gradient descent to perform the outer minimization in (12). Using (11), the resulting outer gradient estimates and the corresponding updates (with step size $\alpha > 0$) are

$$\begin{aligned}\widehat{\nabla}_{\mathbf{x}} \mathcal{L}_{\gamma,k} :=\; & \nabla_{\mathbf{x}} f(\mathbf{x}_k + \hat{\boldsymbol{\epsilon}}_k,\, \mathbf{y}_k + \hat{\boldsymbol{\nu}}_k) \\ & + \gamma\left(\nabla_{\mathbf{x}} g(\mathbf{x}_k,\, \mathbf{y}_k + \tilde{\boldsymbol{\nu}}_k) - \nabla_{\mathbf{x}} g(\mathbf{x}_k,\, \tilde{\mathbf{y}}_k + \tilde{\boldsymbol{\nu}}_k)\right), \end{aligned} \tag{15a}$$

$$\widehat{\nabla}_{\mathbf{y}} \mathcal{L}_{\gamma,k} := \nabla_{\mathbf{y}} f(\mathbf{x}_k + \hat{\boldsymbol{\epsilon}}_k,\, \mathbf{y}_k + \hat{\boldsymbol{\nu}}_k) + \gamma \nabla_{\mathbf{y}} g(\mathbf{x}_k,\, \mathbf{y}_k + \tilde{\boldsymbol{\nu}}_k), \tag{15b}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha\, \widehat{\nabla}_{\mathbf{x}} \mathcal{L}_{\gamma,k}, \qquad \mathbf{y}_{k+1} = \mathbf{y}_k - \alpha\, \widehat{\nabla}_{\mathbf{y}} \mathcal{L}_{\gamma,k}. \tag{15c}$$

2.1.3 ALGORITHM

Algorithm 1 summarizes the procedure. Each outer iteration: (i) computes inner perturbations and performs T_k SAM-style steps on $g(\mathbf{x}_k, \cdot)$ starting from $\mathbf{y}_k$ to obtain $\tilde{\mathbf{y}}_k$, (ii) estimates the outer perturbation $\boldsymbol{\epsilon}$ (and the inner perturbation $\boldsymbol{\nu}$) using ascent directions of the Lagrangian, and (iii) applies penalized gradient updates. This procedure remains first-order while steering both levels toward flatter solutions. Appendix C.3 studies the relationship between the bilevel problem (2) and its penalty reformulation (12).

3 THEORETICAL GUARANTEES FOR SABO

We now present the main convergence results of SABO and the iteration complexity for finding flat minima.

Theorem 3.1. *Suppose Assumptions 2.1, 2.3 and 2.4 hold. Let the penalty parameter, step sizes, and number of inner iterations be chosen to satisfy:*

$$\gamma \geq \ell_{f,0}\sqrt{\frac{\mu}{\varrho}},\ \forall \varrho > 0, \quad \alpha \leq \frac{1}{\ell_{f,1} + \gamma(\ell_{g,1} + \ell_{g,1}^2\mu)}, \quad \beta \leq \min\left\{\frac{1}{2\ell_{g,1}}, 2\mu\right\},$$
$$T_k \geq \log_{c_\beta} \min\{(288\mu^2\ell_{g,1}^2)^{-1}, (2\alpha k)^{-2}\},$$

with $c_\beta = 1 - (\beta/2\mu)$. Then, for the sequence $\{(\mathbf{x}_k, \mathbf{y}_k)\}$ generated by Algorithm 1 and any perturbation sizes $(\rho, \eta) > 0$, we have

$$\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla\mathcal{L}_\gamma(\mathbf{x}_k, \mathbf{y}_k)\right\|^2 \leq \frac{1}{\alpha K}\left(\Phi_0 + \frac{\Phi_1}{\alpha}\right) + \Phi_2\frac{\eta^2\mu^2}{\beta} + \Phi_3(\rho^2 + \eta^2),$$

where the constants are defined as $\Phi_0 = 13(\mathcal{L}_\gamma(\mathbf{x}_1, \mathbf{y}_1) - \inf_{(\mathbf{x},\mathbf{y})} f(\mathbf{x}, \mathbf{y}))$, $\Phi_1 = 180\,\ell_{g,1}^2\mu^2\ell_{f,0}^2$, $\Phi_2 = 180\,\gamma^2\ell_{g,1}^3$, $\Phi_3 = \max\{\Phi_4, \Phi_5\}$, where $\Phi_4 = 45\ell_{f,1}^2$ and $\Phi_5 = 45\ell_{f,1}^2 + 225\gamma^2\ell_{g,1}^2$.

The proof is in Appendix G.1. The bound has three parts: (i) an optimization error $\Phi_0/(\alpha K)$ from the initial suboptimality $\mathcal{L}_\gamma(\mathbf{x}_1, \mathbf{y}_1) - \inf_{(\mathbf{x},\mathbf{y})} f(\mathbf{x}, \mathbf{y})$; (ii) a bias term $\Phi_1/(\alpha^2 K)$ from inexact bilevel gradients; and (iii) sharpness terms $\Phi_2\eta^2\mu^2/\beta + \Phi_3(\rho^2 + \eta^2)$ governed by the radii (ρ, η). Four observations follow. First, γ enters the sharpness-dependent terms quadratically, so perturbation effects grow as the penalty becomes stronger. Second, the PL constant μ both constrains γ (via $\gamma \geq \ell_{f,0}\sqrt{\mu/\varrho}$) and scales Φ_2, so well-conditioned lower-level problems converge faster but require careful tuning of η. Third, the inner iterations T_k grow only logarithmically in k, keeping the nested scheme tractable. Finally, setting $\rho = \eta = 0$ removes the sharpness terms and recovers the PBGD rate (Shen & Chen, 2023), so SABO's extra terms are exactly the cost of sharpness awareness.

Corollary 3.2. *Under the conditions of Theorem 3.1, if $\rho = \eta = \mathcal{O}(K^{-1/2})$, then the sequence $\{(\mathbf{x}_k, \mathbf{y}_k)\}$ generated by Algorithm 1 satisfies $\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla\mathcal{L}_\gamma(\mathbf{x}_k, \mathbf{y}_k)\right\|^2 = \mathcal{O}(K^{-1})$.*

Corollary 3.2 gives an $\mathcal{O}(1/K)$ rate for SABO on the averaged squared Lagrangian-gradient norm. With perturbation radii $\rho = \eta = \mathcal{O}(K^{-1/2})$, the sharpness-dependent error terms vanish, yielding $\mathcal{O}(\varepsilon^{-2})$ gradient complexity, matching SAM (Liu et al., 2021b) and PBGD (Shen & Chen, 2023). This decay balances sharpness exploration against convergence. SABO's advantage thus lies in solution quality rather than complexity: its controlled perturbations steer iterates toward flatter regions, improving generalization (Appendix D.1) without sacrificing worst-case complexity.

Theorem 3.3. *Under the assumptions of Corollary 3.2, let $\{(\mathbf{x}_k, \mathbf{y}_k)\}$ be the sequence generated by Algorithm 1. If $\lim_{k\to\infty}(\mathbf{x}_k, \mathbf{y}_k) = (\mathbf{x}^*, \mathbf{y}^*)$, then $(\mathbf{x}^*, \mathbf{y}^*)$ is a stationary point of (12). Moreover, if $(\mathbf{x}^*, \mathbf{y}^*)$ is a local (respectively, global) solution of (12), then it is also a local (respectively, global) solution of (5).*

The proof is in Appendix G.2. Theorem 3.3 shows that any limit point of SABO is a stationary point of the penalized Lagrangian in (12), and that the local or global minima of (12) coincide with the solutions of the original bilevel problem (5). Together with Theorems C.8–C.9, this yields a complete characterization: SABO converges to stationary points that, under suitable conditions, solve the BO problem.

4 SHARPNESS-AWARE MINIMAX OPTIMIZATION

We next extend the sharpness-aware framework to the minimax setting, seeking solutions that are robust to adversarial perturbations and generalize well by accounting for the geometry of the loss landscape. The sharpness-aware minimax problem is

$$\min_{\mathbf{x}\in\mathbb{R}^{d_1}} \max_{\substack{\mathbf{y}\in\mathbb{R}^{d_2} \\ \|\boldsymbol{\epsilon}\|\le\rho,\,\|\boldsymbol{\nu}\|\le\eta}} f(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}+\boldsymbol{\nu}), \tag{16}$$

where $f:\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}\to\mathbb{R}$ denotes a continuously differentiable function.

Problem (16) is equivalent to minimizing a function $h(\cdot)$ as

$$\min_{\mathbf{x}\in\mathbb{R}^{d_1}} h(\mathbf{x}), \quad \text{with} \quad h(\mathbf{x}) := \max_{\substack{\mathbf{y}\in\mathbb{R}^{d_2} \\ \|\boldsymbol{\epsilon}\|\le\rho,\|\boldsymbol{\nu}\|\le\eta}} f(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}+\boldsymbol{\nu}). \tag{17}$$

We assume a PL condition on the ascent variable $\mathbf{y}$ to ensure that the inner maximization is tractable and that the envelope $h(\mathbf{x})$ remains smooth.

Assumption 4.1 ($\tilde{\mu}$–PL Condition)**.** *There exists a constant $\tilde{\mu}>0$ such that, for any given $\mathbf{x}\in\mathbb{R}^{d_1}$, the function $f(\mathbf{x},\cdot)$ satisfies $\|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})\|^2 \ge \tilde{\mu}^{-1}(\max_{\mathbf{y}'} f(\mathbf{x},\mathbf{y}') - f(\mathbf{x},\mathbf{y}))$ for all $\mathbf{y}\in\mathbb{R}^{d_2}$, where $\max_{\mathbf{y}'} f(\mathbf{x},\mathbf{y}')$ has a nonempty solution set and a finite optimal value.*

Assumption 4.2. *The function $f(\cdot)$ is lower bounded, i.e., $\inf_{\mathbf{x}\in\mathbb{R}^{d_1}} \max_{\mathbf{y}\in\mathbb{R}^{d_2}} f(\mathbf{x},\mathbf{y}) > -\infty$.*

Assumption 4.1 is standard in nonconvex–PL (NC-PL) minimax optimization (Huang, 2023a; Yang et al., 2022). Assumption 4.2 guarantees the feasibility of Problem (17).

It follows from Assumption 4.1 and Lemma C.2 in Appendix C.2 that

$$\nabla_{\mathbf{x}} h(\mathbf{x}) = \nabla_{\mathbf{x}} f(\mathbf{x}+\boldsymbol{\epsilon}^*,\,\mathbf{y}^*+\boldsymbol{\nu}^*), \qquad (\mathbf{y}^*,\boldsymbol{\epsilon}^*,\boldsymbol{\nu}^*) \in \underset{\mathbf{y}\in\mathbb{R}^{d_2},\|\boldsymbol{\epsilon}\|\le\rho,\|\boldsymbol{\nu}\|\le\eta}{\arg\max} f(\mathbf{x}+\boldsymbol{\epsilon},\,\mathbf{y}+\boldsymbol{\nu}).$$

Since solving the inner maximization over $\boldsymbol{\epsilon}$ and $\boldsymbol{\nu}$ exactly is computationally infeasible, we approximately solve it using Lemma C.1, and obtain the SAM-style perturbations as follows:

$$\boldsymbol{\epsilon}(\mathbf{x},\mathbf{y}) = \rho\,\frac{\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})}{\|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\|}, \qquad \boldsymbol{\nu}(\mathbf{x},\mathbf{y}) = \eta\,\frac{\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})}{\|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})\|}. \tag{18}$$

The algorithm then solves the inner maximization over $\mathbf{y}$ to approximate the gradient of the value function $h(\cdot)$. Specifically, at each iteration k, it employs gradient descent–ascent, performing an inner loop over $\mathbf{y}$ for each fixed $\mathbf{x}=\mathbf{x}_k$. Given the outer iterate $\mathbf{x}_k$, we run T_k inner steps on $f(\mathbf{x}_k,\cdot)$ to approximate $\mathbf{y}^*$: set $\mathbf{z}_1^{(k)} := \mathbf{y}_k$, apply (18), and update as follows:

$$\boldsymbol{\epsilon}_t^{(k)} := \boldsymbol{\epsilon}(\mathbf{x}_k,\mathbf{z}_t^{(k)}), \qquad \boldsymbol{\nu}_t^{(k)} := \boldsymbol{\nu}(\mathbf{x}_k,\mathbf{z}_t^{(k)}), \tag{19a}$$

$$\mathbf{z}_{t+1}^{(k)} = \mathbf{z}_t^{(k)} + \beta\,\nabla_{\mathbf{z}} f\big(\mathbf{x}_k+\boldsymbol{\epsilon}_t^{(k)},\mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)}\big), \qquad t=1,\dots,T_k, \tag{19b}$$

$$\boldsymbol{\epsilon}_k := \boldsymbol{\epsilon}_{T_k}^{(k)}, \qquad \boldsymbol{\nu}_k := \boldsymbol{\nu}_{T_k}^{(k)}, \qquad \mathbf{y}_{k+1} := \mathbf{z}_{T_k+1}^{(k)}. \tag{19c}$$

SAMO then applies gradient descent to the outer minimization in (17). Using Lemma C.2, the resulting outer gradient estimates and the corresponding updates (with step size $\alpha>0$) are

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha\,\nabla_{\mathbf{x}} f\big(\mathbf{x}_k+\boldsymbol{\epsilon}_k,\mathbf{y}_{k+1}+\boldsymbol{\nu}_k\big), \qquad k=1,\dots,K. \tag{20}$$

Algorithm 2 summarizes SAMO.

4.1 CONVERGENCE ANALYSIS FOR SAMO

We first define the convergence criterion. Because the function f is nonconvex in $\mathbf{x}$, obtaining its global optimum is generally intractable. Therefore, the goal is to design an algorithm that computes an ε-stationary point of the primal function $F(\mathbf{x}) := \max_{\mathbf{y}\in\mathbb{R}^{d_2}} f(\mathbf{x},\mathbf{y})$ (Huang, 2023a; Nouiehed et al., 2019). A point $\mathbf{x}$ is an ε-stationary point of a function F if $\|\nabla F(\mathbf{x})\|\le\varepsilon$, where $\varepsilon\in(0,1]$ and $\mathbf{x}$ is the output of an algorithm; if $\varepsilon=0$, then $\mathbf{x}$ is a stationary point.

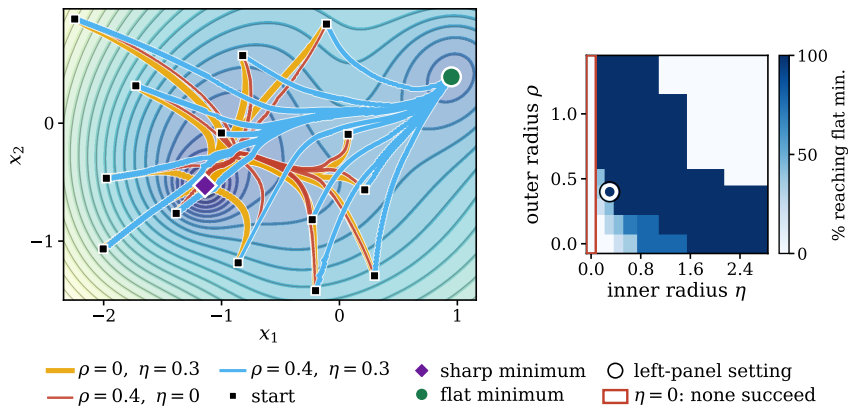

Figure 1: The two radii play different roles on instance (22); the sharp minimum is the purple diamond, the flat one the green circle. Left: paths from fourteen shared initializations under each radius alone and both together. Right: fraction of runs reaching the flat minimum over the (η, ρ) plane (left-panel setting circled); in the empty upper-right corner, runs reach neither minimum.

Theorem 4.3. *Suppose that Assumptions 2.3–II., 4.1 and 4.2 hold. For Algorithm 2, set*

$$\alpha \le \min\Big\{\tfrac{1}{2L_f}, \tfrac{\beta}{16L_f\tilde{\mu}}, \tfrac{\beta}{16\ell_{f,1}^2\tilde{\mu}^2}\Big\}, \quad \beta \le \min\Big\{\sqrt{\tfrac{2\tilde{\mu}}{\ell_{f,1}}}, \tfrac{1}{\ell_{f,1}}\Big\}, \quad T_k \ge \log_{d_\beta}(4\ell_{f,1}^2\tilde{\mu}^2\frac{\alpha}{\beta})^{-1},$$

where $d_\beta = 1 - \frac{\ell_{f,1}}{2\tilde{\mu}}\beta^2$ and L_f is defined in Lemma C.2 in Appendix C.2. Then, for any perturbation sizes $(\rho, \eta) > 0$ and any $K \ge 0$, we have

$$\frac{1}{K}\sum_{k=1}^{K}\|\nabla F(\mathbf{x}_k)\|^2 \le \frac{2}{\alpha K}\Big(2F(\mathbf{x}_1) - f(\mathbf{x}_1, \mathbf{y}_1) - \inf_{\mathbf{x}} F(\mathbf{x})\Big) + \Phi(\rho^2 + \eta^2),$$

where $\Phi := 2\ell_{f,1}^2 + \frac{1}{4\alpha^2} + \frac{1}{4\alpha\beta}$.

Corollary 4.4. *Under the conditions of Theorem 4.3, choosing $\rho = \eta = \mathcal{O}(K^{-1/2})$ yields $(1/K)\sum_{k=1}^{K}\|\nabla F(\mathbf{x}_k)\|^2 = \mathcal{O}(1/K)$. Similar to the state-of-the-art complexity for solving minimax problems (Nouiehed et al., 2019), the optimal gradient complexity of $\mathcal{O}(\varepsilon^{-2})$ is achieved for finding an ε-stationary solution.*

5 Numerical Experiments

We evaluate our algorithms on three tasks: controlled synthetic problems with closed-form sharp and flat solutions (Section 5.1), Data Hypercleaning for BO (Section 5.2), and GAN training for minimax optimization (Section 5.3). Each task compares the sharpness-aware methods (SABO, SAMO) with standard baselines (BO, MO).

5.1 Synthetic Experiments

We begin with a controlled instance whose sharp and flat solutions are known in closed form, to settle whether the two perturbation radii are separate conveniences, each improving its own level, or whether they require one another.

For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$, let the outer landscape be

$$V(\mathbf{x}) = a\|\mathbf{x}\|^2 - \sum_{i\in\{\mathrm{b,p,f}\}} d_i \exp\Big(-\tfrac{\|\mathbf{x}-\mathbf{c}_i\|^2}{2s_i^2}\Big) + b\exp\Big(-\tfrac{(\langle \mathbf{u}, \mathbf{x}\rangle - \tau)^2}{2w^2}\Big), \tag{21}$$

a curved ridge and three Gaussian wells: a wide shallow bowl with a narrow spike concentric inside it, together the sharp region, and a third, flat well (parameter values in Appendix L). Splitting the sharp region into spike and bowl separates two roles a single well would confound: the spike fixes

the curvature at the sharp minimum, 12.9 against 2.0 at the flat one. The *inner* radius reaches the outer variable only through the outer gradient: with $\mathbf{d} = \mathbf{y} - \kappa\mathbf{x}$, the penalty part of $\nabla_{\mathbf{x}}\mathcal{L}_\gamma$ carries $\frac{1}{2}\nabla_{\mathbf{x}} s\,(\|\mathbf{d}\|^2 + 2\eta\|\mathbf{d}\|)$, so where the inner curvature varies with $\mathbf{x}$ the perturbation drifts $\mathbf{x}$ towards better-conditioned regions. We let that curvature ramp along the axis joining the two minima, coordinatized by $t(\mathbf{x})$ (Appendix L):

$$g(\mathbf{x},\mathbf{y}) = \tfrac{1}{2}\, s(\mathbf{x})\, \|\mathbf{y} - \kappa\mathbf{x}\|^2, \qquad s(\mathbf{x}) = s_0 + A\left(1 + e^{\,t(\mathbf{x})/w_s}\right)^{-1}, \tag{22}$$

so that the inner problem is about five times stiffer at the sharp minimum than at the flat one. The outer objective is $f(\mathbf{x},\mathbf{y}) = V(\mathbf{x}) + \lambda\|\mathbf{y} - \kappa\mathbf{x} - \mathbf{q}\|^2$, whose only role is to hold $\mathbf{y}$ off the inner manifold so that $\mathbf{d} \neq \mathbf{0}$.

Three properties of (22), checked in Appendix L, rule out the explanations one would otherwise reach for: the inner solution set $\mathcal{S}(\mathbf{x})$ is a singleton, so η cannot select among inner solutions; the bilevel objective does not contain s; and the perturbed inner problem is still minimized on the inner manifold. The inner radius therefore acts through the gradient alone, by a drift that grows roughly sixfold from the sharp minimum to the flat one and so is weakest exactly where the iterates start.

Figure 1 reports fourteen initializations from which the unperturbed method reaches only the sharp minimum. The outer radius alone never succeeds: with $\eta = 0$ no run reaches the flat minimum, at any ρ up to 1.3. The inner radius on its own does succeed, but only at $\eta = 1.8$, four times the radius it needs when the two are combined. At $\rho = 0.4$ and $\eta = 0.3$ each radius alone returns 0% and the pair returns 100%, at the flat minimum's own curvature and within 0.02 of it. Started *at* the sharp minimum, either radius alone leaves it there and the pair carries it to the flat minimum, so the pair escapes rather than merely diverts. The division of labor: at $\rho = 0.4$ the outer perturbation erases the spike and cuts the barrier ahead of the iterate by two thirds, and that residue is what the inner drift covers; the far edge of the map shows the $\mathcal{O}(\rho^2 + \eta^2)$ floor of Theorem 3.1. Appendix K places this case among four regimes, in which only ρ, only η, both, or either is needed; the pair $(0.4, 0.9)$ succeeds in all four.

5.2 Sharpness-Aware Data Hypercleaning

Data Hypercleaning (DHC) learns per-sample weights that filter out corrupted training examples (Franceschi et al., 2017; Shaban et al., 2019). The lower level trains a classifier $\mathbf{y} \in \mathbb{R}^{d_2}$ on a noisy training set $\mathcal{D}_{tr} := \{(a_i, \bar{b}_i)\}$, in which each label $\bar{b}_i$ is independently flipped to a different class with probability $\tilde{p} \in (0,1)$ and each sample carries the weight $[\sigma(\mathbf{x})]_i$, where $\sigma(\mathbf{x}) = 1/(1+e^{-\mathbf{x}})$ acts element-wise; the upper level tunes $\mathbf{x} \in \mathbb{R}^{d_1}$ on a clean validation set $\mathcal{D}_{val}$. With perturbations $\boldsymbol{\epsilon}$ and $\boldsymbol{\nu}$ in Euclidean balls of radii ρ and η, the problem is

$$\min_{\mathbf{x},\mathbf{y}} \max_{\|\boldsymbol{\epsilon}\|\le\rho,\,\|\boldsymbol{\nu}\|\le\eta} \tfrac{1}{|\mathcal{D}_{val}|} \sum\nolimits_{\xi\in\mathcal{D}_{val}} \mathcal{L}(\mathbf{x},\mathbf{y};\boldsymbol{\epsilon},\boldsymbol{\nu};\xi)$$
$$\text{subj. to } \mathbf{y} \in \arg\min_{\mathbf{y}} \max_{\|\boldsymbol{\nu}\|\le\eta} \tfrac{1}{|\mathcal{D}_{tr}|} \sum\nolimits_{\zeta_i\in\mathcal{D}_{tr}} [\sigma(\mathbf{x})]_i\, \mathcal{L}(\mathbf{y};\boldsymbol{\nu};\zeta_i) + \tfrac{\lambda}{2}\|\mathbf{y}\|^2,$$

where the regularizer with $\lambda > 0$ makes the lower level strongly convex. We test noise rates of 20%, 40% and 50%, following standard protocols (Ji et al., 2021b; Shen & Chen, 2023), and report test accuracy and the F1 score of the learned cleaner, which measures how well it separates clean samples from corrupted ones. Appendix J gives the datasets, architectures and optimizer settings.

Figure 2 traces test accuracy on CIFAR-10; Appendix J adds FashionMNIST, the F1 of the cleaner and all final numbers (Table 1). First, the benefit of SABO grows with the noise rate. At 20% noise the two methods finish within 0.4 points of each other on every architecture; at 40% SABO leads by 2.5 to 5.5 points, and at 50% by 7.3 points on ResNet-18 and 3.9 on WRN-28-8. Second, the curves show when the gap forms. At 40% and 50% noise every run peaks early and then loses accuracy as the network begins to fit the corrupted labels; SABO's dip is shallower and its recovery stronger, so its advantage is built in the second half of training, when a sharp solution memorizes the noise. Third, the gain is largest on the smallest model and clearest in the cleaner itself: over both datasets SABO wins 16 of 18 F1 comparisons, by 19.7 to 32.8 points on ResNet-18. The two losses in each metric fall on the wider networks (e.g., WRN-16-8 at 50% noise trails BO by 0.8 accuracy points), which suggests that capacity partly substitutes for sharpness-awareness. A controlled replication in Appendix L isolates the mechanism: the inner radius η widens the gap between the mean weights given to clean and to corrupted samples from 0.35 to 0.53, so SABO improves the cleaner, not only the flatness of the classifier.

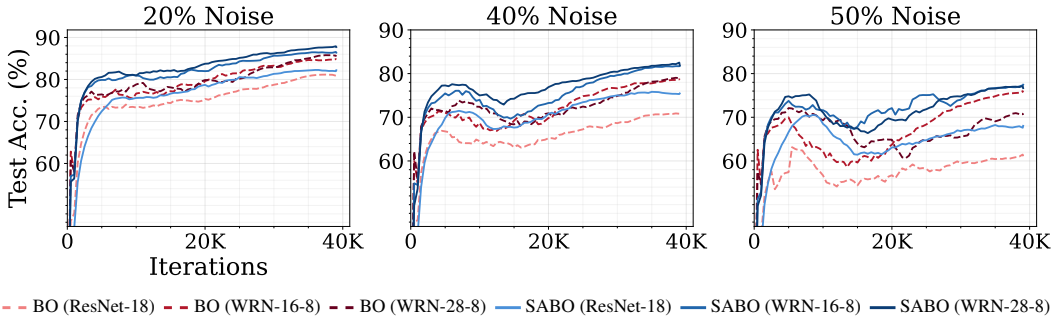

Figure 2: Test accuracy on CIFAR-10 during training under 20%, 40%, and 50% label noise; color marks the architecture, SABO is solid and BO dashed. At high noise every run peaks early and then declines as the corrupted labels are fitted; the SABO curves dip less and recover more.

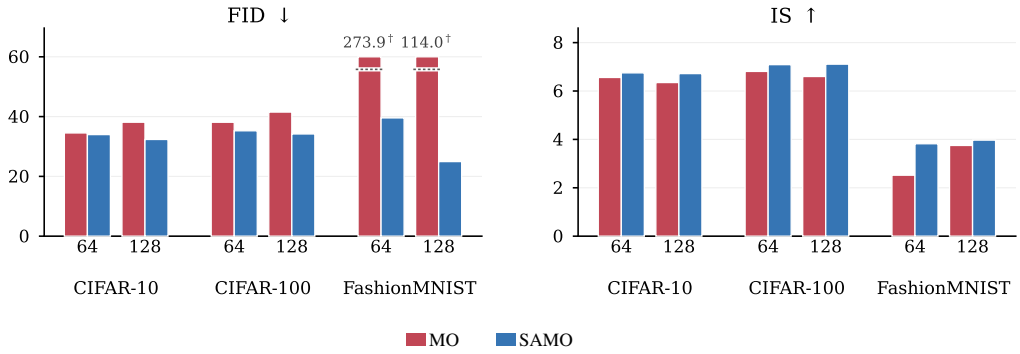

Figure 3: GAN results for DCGAN-64 and DCGAN-128 on three datasets: FID (left, lower is better) and IS (right, higher is better). † marks mode collapse of MO (FID > 100); those two bars run off the axis and carry their value. SAMO improves both metrics on all six configurations, with the narrowest margin on CIFAR-10 with DCGAN-64.

5.3 Sharpness-Aware GAN Training

GAN training is a two-player game between a generator $G_{\mathbf{x}} : z \mapsto u$ and a discriminator $D_{\mathbf{y}} : u \mapsto [0, 1]$ with parameters $\mathbf{x} \in \mathbb{R}^{d_1}$ and $\mathbf{y} \in \mathbb{R}^{d_2}$. Perturbations $\boldsymbol{\epsilon}$ and $\boldsymbol{\nu}$ move each player within a Euclidean ball of radius ρ_G or ρ_D. With $u \sim P_{data}$ and $z \sim P_z$, the losses are

$$\mathcal{L}_D := -\mathbb{E}_u\big[\log D_{\mathbf{y}+\boldsymbol{\nu}}(u)\big] - \mathbb{E}_z\big[\log\big(1 - D_{\mathbf{y}+\boldsymbol{\nu}}(G_{\mathbf{x}+\boldsymbol{\epsilon}}(z))\big)\big] \tag{23}$$

and $\mathcal{L}_G := -\mathbb{E}_z[\log D_{\mathbf{y}+\boldsymbol{\nu}}(G_{\mathbf{x}+\boldsymbol{\epsilon}}(z))]$, both functions of $(\mathbf{x}, \mathbf{y}; \boldsymbol{\epsilon}, \boldsymbol{\nu})$. The players solve

$$\min_{\mathbf{x}} \max_{\|\boldsymbol{\epsilon}\| \le \rho_G} \mathcal{L}_G(\mathbf{x}, \mathbf{y}; \boldsymbol{\epsilon}, \mathbf{0}) \qquad \text{and} \qquad \min_{\mathbf{y}} \max_{\|\boldsymbol{\nu}\| \le \rho_D} \mathcal{L}_D(\mathbf{x}, \mathbf{y}; \mathbf{0}, \boldsymbol{\nu}). \tag{24}$$

GAN instability—oscillation near saddle points, discriminator overfitting, mode collapse—is tied to each player's landscape sharpness, which (24) penalizes by requiring each network to hold its performance over a local neighborhood. The radii are asymmetric ($\rho_G \neq \rho_D$) since the players differ in scale; under $\mathcal{L}_G = -\mathcal{L}_D$ the framework reduces to Section 4's setting.

Figure 3 compares SAMO with minimax optimization (MO) in Fréchet Inception Distance (FID) and Inception Score (IS) at two DCGAN widths, with $\rho_G = 0.1$ and $\rho_D = 0.01$ (settings in Appendix J). The results fall into two regimes. On CIFAR-10 and CIFAR-100 both methods train stably and the gains are modest but grow with capacity: doubling the width worsens the FID of MO (from 34.5 to 38.1 on CIFAR-10 and from 38.1 to 41.5 on CIFAR-100) but improves that of SAMO (from 34.0 to 32.4 and from 35.3 to 34.2), so the largest gains, 5.8 and 7.3 FID, occur with DCGAN-128. This is consistent with a wider network giving both players more room to reach the sharp, unstable equilibria that the perturbations penalize. On FashionMNIST, MO does not merely lag but collapses, with FID 273.9 and 114.0 and an IS of 2.5 at DCGAN-64, while SAMO trains stably at both widths and reaches an FID of 25.0, the best in the figure. Sharpness-awareness therefore acts less as a refinement of sample quality than as a stabilizer for runs that would otherwise fail.

AI USE STATEMENT

We used generative AI tools to polish the writing, to write code for the experiments, to generate synthetic datasets, and to produce figures. All mathematical results and proofs were developed and validated by the authors; AI tools were not used to derive or prove any mathematical claim. All AI-generated code was reviewed, tested and run by the authors, and all AI-assisted text was checked by the authors. We take full responsibility for the content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops optimization methods and their analysis; it involves no human subjects and no personal or sensitive data. The experiments use public benchmark datasets (CIFAR-10, CIFAR-100 and FashionMNIST) and synthetic problems. For reproducibility, the paper and its appendices give all algorithmic details, hyperparameters and experimental settings needed to reproduce the algorithms and the reported results.

REFERENCES

Momin Abbas, Quan Xiao, Lisha Chen, Pin-Yu Chen, and Tianyi Chen. Sharp-MAML: Sharpness-aware model-agnostic meta learning. In *International conference on machine learning*, pp. 10–32. PMLR, 2022.

Lukas Adolphs, Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Local saddle point optimization: A curvature exploitation approach. In *Proceedings of AISTATS*, pp. 486–495. PMLR, 2019.

Alp Alacaoglu, Fei Huang, and Volkan Cevher. Stochastic variance reduction for weakly convex-concave min-max problems. In *NeurIPS*, pp. 2235–2246, 2022.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17(236):1–41, 2016.

Maksym Andriushchenko, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. Sharpness-aware minimization leads to low-rank features. *Advances in Neural Information Processing Systems*, 36, 2024.

Michael Arbel and Julien Mairal. Non-convex bilevel games with critical point selection maps. *Advances in Neural Information Processing Systems*, 35:8013–8026, 2022.

A. Auslender and M. Teboulle. Projected subgradient methods with non-Euclidean distances for non-differentiable convex minimization and variational inequalities. *Mathematical Programming*, 120:27–48, 2009.

Joel Bailey and Georgios Piliouras. Multiplicative weights update in zero-sum games. In *AAMAS*, pp. 1216–1224. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

David Balduzzi, Sam Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. In *Proceedings of ICML*, pp. 354–363. PMLR, 2018.

Fan Bao, Guoqiang Wu, Chongxuan Li, Jun Zhu, and Bo Zhang. Stability and generalization of bilevel programming in hyperparameter optimization. *Advances in neural information processing systems*, 34:4529–4541, 2021.

Behzad Barazandeh and Meisam Razaviyayn. Solving non-convex non-differentiable min-max games using proximal gradient method. In *ICASSP*, pp. 3162–3166. IEEE, 2020.

Peter L Bartlett, Philip M Long, and Olivier Bousquet. The dynamics of sharpness-aware minimization: Bouncing across ravines and drifting towards wide minima. *Journal of Machine Learning Research*, 24(316):1–36, 2023.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Radu I. Boţ and Andreas Böhm. Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. *SIAM Journal on Optimization*, 33(3):1884–1913, 2023.

Jerome Bracken and James T McGill. Mathematical programs with optimization problems in the constraints. *Operations Research*, 21(1):37–44, 1973.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2018.

Xiaowen Cai, Yuantao Tang, and Tianbao Yang. Accelerated algorithms for nonconvex-nonconcave minimax optimization. In *Proceedings of AISTATS*, 2024.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019 (12):124018, 2019.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *NeurIPS*, pp. 9236–9246, 2018.

Constantinos Daskalakis, Ioannis Panageas, Christos Papadimitriou, Georgios Piliouras, and Vasilis Syrgkanis. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *Proceedings of COLT*, pp. 787–803. PMLR, 2019.

Constantinos Daskalakis, Tristan Fiez, and Panayotis Mertikopoulos. Stay on the ridge: A framework for non-asymptotic convergence in min-max optimization. In *Proceedings of ICML*, pp. 7202–7228. PMLR, 2023.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1019–1028, 2017.

Justin Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pp. 318–326. PMLR, 2012.

Jiawei Du, Hanshu Yan, Jiashi Feng, Joey Tianyi Zhou, Liangli Zhen, Rick Siow Mong Goh, and Vincent YF Tan. Efficient sharpness-aware minimization for improved training of neural networks. *arXiv preprint arXiv:2110.03141*, 2021.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

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.

Spencer Frei and Quanquan Gu. Proxy convexity: A unified framework for the analysis of neural networks trained by gradient descent. *Advances in Neural Information Processing Systems*, 34: 7937–7949, 2021.

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pp. 2672–2680, 2014.

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

Petra Grnarova, Kevin Y. Levy, Aurelien Lucchi, Thomas Hofmann, and Andreas Krause. An online learning approach to generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1Yp-j1Cb.

Dong Han, Niao He, Chi Wang, Ming Liu, and Tianbao Yang. Lower complexity bounds and optimal algorithms for nonconvex finite-sum min-max optimization. In *Proceedings of AISTATS*, 2024.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York, 2nd edition, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, pp. 6626–6637, 2017.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020.

Fei Huang, Xu Wu, and He Huang. Efficient mirror descent ascent methods for nonsmooth minimax problems. In *NeurIPS*, pp. 10431–10443, 2021.

Fei Huang, Yiyuan Li, Tianbao Yang, and Mingyi Hong. Accelerated algorithms for stochastic finite-sum and multi-block nonconvex-concave min-max problems. In *Proceedings of ICML*, pp. 8990–9009. PMLR, 2022.

Feihu Huang. Enhanced adaptive gradient algorithms for nonconvex-PL minimax optimization. *arXiv preprint arXiv:2303.03984*, 2023a.

Feihu Huang. On momentum-based gradient methods for bilevel optimization with nonconvex lower-level. *arXiv preprint arXiv:2303.03944*, 2023b.

Feihu Huang. Optimal Hessian/Jacobian-free nonconvex-PL bilevel optimization. *Forty-first International Conference on Machine Learning*, 2024.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Provably faster algorithms for bilevel optimization and applications to meta-learning. *ArXiv*, abs/2010.07962, 2020.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, pp. 4882–4892. PMLR, 2021a.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Provably faster algorithms for bilevel optimization and applications to meta-learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021b.

Weisen Jiang, Hansi Yang, Yu Zhang, and James Kwok. An adaptive policy to employ sharpness-aware minimization. *arXiv preprint arXiv:2304.14647*, 2023.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International conference on machine learning*, pp. 4880–4889. PMLR, 2020a.

Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *Proceedings of ICML*, pp. 4880–4889. PMLR, 2020b.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pp. 795–811. Springer, 2016.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Minyoung Kim, Da Li, Shell X Hu, and Timothy Hospedales. Fisher SAM: Information geometry and sharpness aware minimisation. In *International Conference on Machine Learning*, pp. 11148–11161. PMLR, 2022.

Weijie Kong and R. D. C. Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *SIAM Journal on Optimization*, 31(4):2558–2585, 2021.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pp. 5905–5914. PMLR, 2021.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pp. 1302–1338, 2000.

Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Tianyi Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *Proceedings of AISTATS*, pp. 907–915. PMLR, 2019.

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, 2022a.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

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, 2021a.

Risheng Liu, Yaohua Liu, Shangzhi Zeng, and Jin Zhang. Towards gradient-based bilevel optimization with non-convex followers and beyond. *Advances in Neural Information Processing Systems*, 34:8662–8675, 2021b.

Shengrui Liu, Ming Liu, Chi Wang, and Tianbao Yang. First-order methods for min-max optimization under growing iterate bounds. In *Proceedings of NeurIPS*, pp. 1797–1808, 2021c.

Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12360–12370, 2022b.

Shiqian Lu, Vasileios Tsaknakis, Mingyi Hong, and Yang Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: Algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691, 2020.

Lu Luo, Yurong Xu, and Yinyu Ye. Stochastic variance reduction for finite-sum minimax problems. In *Proceedings of ICML*, pp. 6418–6428. PMLR, 2020.

Aleksander Madry, Aleksandr Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Panayotis Mertikopoulos, Baptiste Lecouat, Hadrien Zenati, Chastity-S Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg8jjC9KQ.

Peng Mi, Li Shen, Tianhe Ren, Yiyi Zhou, Xiaoshuai Sun, Rongrong Ji, and Dacheng Tao. Make sharpness-aware minimization stronger: A sparsified perturbation approach. *Advances in Neural Information Processing Systems*, 35:30950–30962, 2022.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Parvin Nazari, Bojian Hou, Davoud Ataee Tarzanagh, Li Shen, and George Michailidis. Stochastic regret guarantees for online zeroth-and first-order bilevel optimization. *Advances in Neural Information Processing Systems*, 39, 2025a.

Parvin Nazari, Ahmad Mousavi, Davoud Ataee Tarzanagh, and George Michailidis. A penalty-based method for communication-efficient decentralized bilevel programming. *Automatica*, 173: 112039, 2025b.

Mohammad Nouiehed, Mohammad Sanjabi, Tenghang Huang, Jinwoo D. Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *NeurIPS*, pp. 14905–14916, 2019.

David M. Ostrovskii, Aharon Lowy, and Meisam Razaviyayn. Efficient search of first-order Nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31 (4):2508–2538, 2021.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pp. 737–746. PMLR, 2016.

Boris T Polyak. Gradient methods for solving equations and inequalities. *USSR Computational Mathematics and Mathematical Physics*, 4(6):17–32, 1964.

Hamood Rafique, Mingyi Liu, Quan Lin, and Tianbao Yang. Weakly-convex-concave min-max optimization: Provable algorithms and applications in machine learning. *Optimization Methods and Software*, pp. 1–35, 2021.

J. Robinson. An iterative method of solving a game. *Annals of Mathematics*, pp. 296–301, 1951.

Mohammad Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jinwoo D. Lee. On the convergence and robustness of training GANs with regularized optimal transport. In *NeurIPS*, pp. 7091–7101, 2018.

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, Zhuoran Yang, and Tianyi Chen. Principled penalty-based methods for bilevel reinforcement learning and RLHF. *arXiv preprint arXiv:2402.06886*, 2024.

Amanjaneya Sinha, Hongseok Namkoong, and John C. Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hk6kPgZA-.

Ankur Sinha, Pekka Malo, and Kalyanmoy Deb. A review on bilevel optimization: from classical to evolutionary approaches and applications. *IEEE Transactions on Evolutionary Computation*, 22 (2):276–295, 2017.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Chaehwan Song, Ali Ramezani-Kebrya, Thomas Pethick, Armin Eftekhari, and Volkan Cevher. Subquadratic overparameterization for shallow neural networks. *Advances in Neural Information Processing Systems*, 34:11247–11259, 2021.

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.

Heinrich von Stackelberg. Theory of the market economy. *Oxford University Press*, 1952.

Davoud Ataee Tarzanagh, Parvin Nazari, Bojian Hou, Li Shen, and Laura Balzano. Online bilevel optimization: Regret analysis of online alternating gradient methods. In *International Conference on Artificial Intelligence and Statistics*, pp. 2854–2862. PMLR, 2024.

K. K. Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. In *NeurIPS*, pp. 12659–12670, 2019.

John von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928.

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

Quan Xiao, Songtao Lu, and Tianyi Chen. A generalized alternating method for bilevel learning under the Polyak-Łojasiewicz condition. *arXiv e-prints*, pp. arXiv–2306, 2023.

Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 33:1153–1165, 2020.

Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In *International conference on artificial intelligence and statistics*, pp. 5485–5517. PMLR, 2022.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Qihuang Zhong, Liang Ding, Li Shen, Peng Mi, Juhua Liu, Bo Du, and Dacheng Tao. Improving sharpness-aware minimization with fisher mask for better generalization on language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4064–4085, 2022.

CONTENTS

1 Introduction **1**

2 Sharpness-Aware Bilevel Optimization **2**

2.1 Penalty Reformulation for SABO 2

2.1.1 Stage I: Sharpness-Aware Inner Optimization 2

2.1.2 Stage II: Sharpness-Aware Lagrangian Minimization 3

2.1.3 Algorithm 5

3 Theoretical Guarantees for SABO **5**

4 Sharpness-Aware Minimax Optimization **6**

4.1 Convergence Analysis for SAMO 6

5 Numerical Experiments **7**

5.1 Synthetic Experiments 7

5.2 Sharpness-Aware Data Hypercleaning 8

5.3 Sharpness-Aware GAN Training 9

Appendix **18**

A Notation and Related Work **18**

A.1 Notation 18

A.2 Related Work 18

B Additional Preliminaries **19**

B.1 Auxiliary Lemmas 19

C Theoretical Discussions **20**

C.1 Missing Details from Section 2.1.2 20

C.2 Missing Details from Section 4 20

C.3 Linking the SABO Problem to Its Penalty Reformulation 20

D Generalization Discussions **23**

D.1 Generalization Analysis for SABO 23

D.2 Generalization Analysis for SAMO 24

E Proofs omitted from Section C.3 **25**

E.1 Proof of Lemma C.5 26

E.2 Proof of Lemma C.7 26

E.3 Proof of Theorem C.8 27

E.4 Proof of Theorem C.9 29

E.5 Proof of Corollary C.10 30

F Proofs omitted from Section 2.1.2 **31**

F.1 Proof of Lemma C.1 31

G Proofs omitted from Section 3 **32**

G.1 Proof of Theorem 3.1 35

G.2 Proof of Theorem 3.3 38

H Proofs omitted from Section D.1 **39**

H.1 Proof of Theorem D.1 39

I Proofs omitted from Section 4 **44**

I.1 Proof of Theorem 4.3 50

I.2 Proof of Theorem D.2 52

J Experimental Details **57**

J.1 Data Hypercleaning 57

J.2 GAN Training 58

K Which Radius Does the Work? Four Regimes **58**

L Additional Synthetic Experiments **60**

APPENDIX

In the supplement, Appendix A.1 presents the notation, while Appendix A.2 presents the related work. Appendix B.1 presents the auxiliary lemmas. Appendices C.1 and C.2 present the technical details of the baselines. Appendix C.3 establishes the connection between the SABO problem and its penalty reformulation. Appendices D.1 and D.2 provide the generalization analysis of SABO and SAMO, respectively. Appendix E provides the proof of the results in Section C.3. Appendices F–I provide the proofs of the theorems and lemmas in Sections 2 and 4. Appendix J provides the computational setup for the data hypercleaning and GAN experiments, together with the full hypercleaning results. Appendix K maps, on four synthetic instances, when the outer radius, the inner radius, or both are needed. Finally, Appendix L supplies additional experimental details and results that further validate the effectiveness of our method.

A NOTATION AND RELATED WORK

A.1 NOTATION

$\mathbb{R}^d$ is the d-dimensional real space; $\mathbb{R}^d_+$ and $\mathbb{R}^d_{++}$ denote its nonnegative and positive orthants. Bold roman/Greek are vectors (e.g., $\mathbf{x} \in \mathbb{R}^{d_1}$, $\boldsymbol{\epsilon} \in \mathbb{R}^{d_1}$). $\|\cdot\|$ is the Euclidean norm; $\langle \cdot, \cdot \rangle$ the standard inner product. For $\rho, \eta > 0$, let $\mathbb{B}_\rho^{(d_1)} = \{\boldsymbol{\epsilon} \in \mathbb{R}^{d_1} : \|\boldsymbol{\epsilon}\| \leq \rho\}$ and $\mathbb{B}_\eta^{(d_2)} = \{\boldsymbol{\nu} \in \mathbb{R}^{d_2} : \|\boldsymbol{\nu}\| \leq \eta\}$; we also write $\|\boldsymbol{\epsilon}\| \leq \rho$, $\|\boldsymbol{\nu}\| \leq \eta$. For a set $\mathcal{S}$ and point $\mathbf{z}$, define $d_{\mathcal{S}}(\mathbf{z}) = \min_{\mathbf{z}' \in \mathcal{S}} \|\mathbf{z} - \mathbf{z}'\|$. Let $\mathcal{N}(\mathbf{z}, r) := \{\mathbf{z}' \in \mathbb{R}^{\dim(\mathbf{z})} : \|\mathbf{z} - \mathbf{z}'\| \leq r\}$. Partial derivatives with respect to $\mathbf{x}$ and $\mathbf{y}$ are denoted by $\nabla_{\mathbf{x}} f$ and $\nabla_{\mathbf{y}} f$, respectively, and ∇f denotes the gradient. $\arg\min$ / $\arg\max$ return solution sets; $\mathcal{S}(\cdot)$ denotes a set-valued map and $\mathbf{y}^*$ any selection. Expectations are denoted by $\mathbb{E}[\cdot]$ (or $\mathbb{E}_{\xi \sim \mathcal{D}}[\cdot]$), and probabilities by $\mathbb{P}(\cdot)$. Let $\mathbf{I}_d$ denote a d-dimensional identity matrix. We use standard asymptotic notation: $g = \mathcal{O}(h)$ if $|g| \leq C|h|$ for some constant $C > 0$; $g = \Omega(h)$ if $|g| \geq c|h|$ for some constant $c > 0$; and $g = \Theta(h)$ if both conditions hold.

A.2 RELATED WORK

Sharpness-Aware Single-Level Minimization. Foret et al. (2020) introduce SAM, which improves generalization by jointly minimizing the loss function and sharpness. Bartlett et al. (2023) analyze SAM on convex quadratics, showing convergence to cycles along directions of largest curvature, while effectively reducing the minimum Hessian eigenvalue. SAM-trained networks have also been shown to learn lower-rank features (Andriushchenko et al., 2024).

Recent works have enhanced the scalability of sharpness-aware methods through adaptations to coordinate-wise ill-scaling (Kim et al., 2022; Kwon et al., 2021), gradient reuse and sparsification (Liu et al., 2022b; Mi et al., 2022), stochastic perturbations (Du et al., 2021), and adaptive switching to SGD (Jiang et al., 2023). However, most of these approaches—and work connecting loss geometry to generalization—focus solely on *single-level* optimization. In this paper, we study sharpness-aware *nested* optimization, where multiple levels are optimized jointly.

Bilevel Optimization. Since Stackelberg's formulation (Stackelberg, 1952) and Bracken–McGill's early model (Bracken & McGill, 1973), bilevel programming has been studied mainly via (i) reducing the nested structure to a single-level problem using the Karush–Kuhn–Tucker (KKT) conditions or penalty-based reformulations (Sinha et al., 2017), which remain algorithmically challenging, and (ii) gradient-based methods that estimate hypergradients using either iterative differentiation (Franceschi et al., 2017; Grazzi et al., 2020) or approximate implicit differentiation (Domke, 2012; Ghadimi & Wang, 2018; Grazzi et al., 2020; Ji et al., 2021a; Nazari et al., 2025b; Pedregosa, 2016), with asymptotic (Franceschi et al., 2018; Shaban et al., 2019) and finite-time (Ghadimi & Wang, 2018; Hong et al., 2020; Ji et al., 2020; 2021a) guarantees.

The majority of existing bilevel algorithms, particularly gradient-based frameworks, rely on the restrictive assumption of a convex inner objective. However, many applications are inherently nonconvex, e.g., neural architecture search (Liu et al., 2018), where the inner problem involves optimizing deep neural networks. To address this nonconvexity, several methods have emerged. These include auxiliary initialization schemes (Liu et al., 2021b), heuristic extensions of the implicit func-

tion theorem (Arbel & Mairal, 2022), and single-level or penalty-based reformulations with explicit convergence guarantees (Liu et al., 2022a; 2021a; Shen & Chen, 2023). Furthermore, alternating and momentum-based update schemes (Huang, 2023b; Tarzanagh et al., 2024; Xiao et al., 2023), alongside computationally efficient Hessian-free variants (Huang, 2024; Nazari et al., 2025a), have been actively explored. While existing stability and generalization analyses indicate that bilevel methods can theoretically outperform classic cross-validation paradigms (Bao et al., 2021), a critical gap remains. To our knowledge, *no prior work* has investigated the role of *loss-surface geometry and generalization performance of SAM-type BO*.

Minimax Optimization. Minimax optimization traces its origins back to von Neumann's minimax theorem (von Neumann, 1928) for bilinear matrix games of the form $\min_{\mathbf{x}\in\Delta^m}\max_{\mathbf{y}\in\Delta^n}\mathbf{x}^\top\mathbf{A}\mathbf{y}$, a foundational pillar of game theory. This work inspired algorithms with simultaneous minimax learning (Robinson, 1951) and Sion's convex-concave extension (Sion, 1958), followed by advances in continuous/discrete-time dynamics (Auslender & Teboulle, 2009; Liang & Stokes, 2019). While classical gradient-based frameworks provide rigorous complexity guarantees within convex–concave and strongly-convex–concave regimes (Auslender & Teboulle, 2009; Liang & Stokes, 2019), new machine learning applications have driven a shift toward more complex geometries. Consequently, recent literature has heavily focused on nonconvex–concave and fully nonconvex–nonconcave formulations (Grnarova et al., 2018; Kong & Monteiro, 2021; Lu et al., 2020; Ostrovskii et al., 2021; Rafique et al., 2021; Sanjabi et al., 2018; Sinha et al., 2018; Thekumparampil et al., 2019). Algorithmic innovations in this regime broadly fall into three categories: inexact or accelerated gradient descent-ascent (GDA) schemes (Grnarova et al., 2018; Jin et al., 2020b; Sanjabi et al., 2018; Sinha et al., 2018), structured nested-loop and single-loop architectures (Kong & Monteiro, 2021; Lu et al., 2020; Ostrovskii et al., 2021; Rafique et al., 2021; Thekumparampil et al., 2019), and extensions tailored to nonsmooth or composite settings (Barazandeh & Razaviyayn, 2020; Boţ & Böhm, 2023; Huang et al., 2021).

The advent of GANs (Goodfellow et al., 2014) and adversarial learning (Madry et al., 2018) has made nonconvex–nonconcave minimax optimization central in machine learning (ML), motivating research on alternative solution concepts, oscillation mitigation (Adolphs et al., 2019; Daskalakis & Panageas, 2018; Jin et al., 2020b), and accelerated training dynamics (Balduzzi et al., 2018; Heusel et al., 2017; Mertikopoulos et al., 2019). Some algorithms ensure stability or convergence under restrictive conditions (Adolphs et al., 2019; Balduzzi et al., 2018; Daskalakis & Panageas, 2018; Heusel et al., 2017; Mertikopoulos et al., 2019), and recent work extends theory to broader classes (Cai et al., 2024; Daskalakis et al., 2023; Liu et al., 2021c). Other research studies online/no-regret properties (Bailey & Piliouras, 2018; Daskalakis et al., 2019) and variance reduction in finite-sum settings (Alacaoglu et al., 2022; Han et al., 2024; Huang et al., 2022; Luo et al., 2020). Yet, despite this progress, the loss-surface geometry and generalization performance of MO remain largely unexplored.

B Additional Preliminaries

B.1 Auxiliary Lemmas

The following lemma relates the PL condition to the quadratic growth (QG) condition (Yang et al., 2020).

Lemma B.1. *(Yang et al., 2020, Lemma A.1) Suppose Assumption 2.1 holds. For all* $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$*, it holds that*

$$g(\mathbf{x},\mathbf{y}) - \min_{\mathbf{y}\in\mathbb{R}^{d_2}} g(\mathbf{x},\mathbf{y}) \geq \frac{1}{\mu} d^2_{\mathcal{S}(\mathbf{x})}(\mathbf{y}),$$

where $\mathcal{S}(\mathbf{x}) = \arg\min_{\mathbf{y}\in\mathbb{R}^{d_2}} g(\mathbf{x},\mathbf{y})$*, and* $d_{\mathcal{S}(\mathbf{z_2})}(\mathbf{z_1}) = \inf\{\|\mathbf{z_1}-\mathbf{z}\| \mid \mathbf{z}\in\mathcal{S}(\mathbf{z_2})\}$ *denotes the Euclidean distance from point* $\mathbf{z}_1$ *to the set* $\mathcal{S}(\mathbf{z}_2)$*.*

Lemma B.2. *There is an approximate solution,* $\tilde{\boldsymbol{\nu}}$*, for the inner maximization of Problem (4), given by*

$$\tilde{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y}) = \frac{\eta\nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y})}{\|\nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y})\|}.$$

Proof. The proof is similar to that of Foret et al. (2020). For a small value of η, using a first-order Taylor expansion around $\mathbf{y}$, the inner maximization in (4) becomes the following linearly constrained optimization problem

$$\begin{aligned}\boldsymbol{\nu}^*(\mathbf{x},\mathbf{y}) &= \arg\max_{\|\boldsymbol{\nu}\|\le\eta} g(\mathbf{x},\mathbf{y}+\boldsymbol{\nu}) \\ &\approx \arg\max_{\|\boldsymbol{\nu}\|\le\eta} g(\mathbf{x},\mathbf{y}) + \boldsymbol{\nu}^\top \nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y}) + \mathcal{O}(\eta^2) \\ &\approx \arg\max_{\|\boldsymbol{\nu}\|\le\eta} \boldsymbol{\nu}^\top \nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y}). \end{aligned} \tag{25}$$

To maximize $\boldsymbol{\nu}^\top \nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y})$, $\boldsymbol{\nu}$ should be in the same direction as $\nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y})$. By the Cauchy–Schwarz inequality, the dot product $\boldsymbol{\nu}^\top \nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y})$ is maximized when $\boldsymbol{\nu}$ is a scalar multiple of $\nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y})$. Let $\boldsymbol{\nu} = a\nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y})$ where a is a scalar. Then the constraint $\|\boldsymbol{\nu}\| \le \eta$ becomes

$$\|a\nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y})\| \le \eta, \quad \Rightarrow |a|\|\nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y})\| \le \eta, \quad \Rightarrow |a| \le \frac{\eta}{\|\nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y})\|}.$$

Since $|a|$ must be as large as possible but less than $\frac{\eta}{\|\nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y})\|}$, we take $a = \frac{\eta}{\|\nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y})\|}$. Hence, the solution of (25) is

$$\tilde{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y}) = \frac{\eta\nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y})}{\|\nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y})\|}.$$

□

C Theoretical Discussions

C.1 Missing Details from Section 2.1.2

Lemma C.1. *For any $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ and $\rho, \eta > 0$, there exist approximate solutions $\hat{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y})$ and $\hat{\boldsymbol{\epsilon}}(\mathbf{x},\mathbf{y})$ to the inner maximization in (12) such that*

$$\hat{\boldsymbol{\nu}}(\mathbf{x},\mathbf{y}) = \eta\frac{\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})}{\|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})\|}, \tag{26a}$$

$$\hat{\boldsymbol{\epsilon}}(\mathbf{x},\mathbf{y}) = \rho\frac{\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})}{\|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\|}, \tag{26b}$$

where $\mathbf{y}^* \in \mathcal{S}(\mathbf{x})$.

The proof of Lemma C.1 is in Appendix F.1. Intuitively, Lemma C.1 states that the maximizers of the inner problem can be well approximated by normalized gradient directions of the penalized objective.

C.2 Missing Details from Section 4

To evaluate the gradient of the value function $h(\cdot)$ in (17), we need the following lemma.

Lemma C.2. *(Nouiehed et al., 2019, Lemma A.5) Suppose Assumptions 2.3–II. and 4.1 hold. Let $F(\mathbf{x}) := \max_{\mathbf{y}\in\mathbb{R}^{d_2}} f(\mathbf{x},\mathbf{y})$, we have $\nabla_{\mathbf{x}} F(\mathbf{x}) = \nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}^*)$ for any $\mathbf{y}^* \in \arg\max_{\mathbf{y}\in\mathbb{R}^{d_2}} f(\mathbf{x},\mathbf{y})$. Moreover, F is L_f-smooth with $L_f := \ell_{f,1} + \ell_{f,1}^2/\tilde{\mu}$.*

C.3 Linking the SABO Problem to Its Penalty Reformulation

Before presenting the main results, we introduce the minimax game $\min_{\mathbf{w}} \max_{\mathbf{v}} \phi(\mathbf{w},\mathbf{v})$ and recall the following notions of minimax points (MMPs), approximate MMPs, and stationary points (Jin et al., 2020a).

Definition C.3. *For all $(\mathbf{w},\mathbf{v})$, a point $(\mathbf{w}^*,\mathbf{v}^*)$ is*

I. *a global MMP of ϕ, if $\phi(\mathbf{w}^*,\mathbf{v}) \le \phi(\mathbf{w}^*,\mathbf{v}^*) \le \max_{\bar{\mathbf{v}}} \phi(\mathbf{w},\bar{\mathbf{v}})$;*

II. *a local MMP of ϕ, if there exists $\delta_0 > 0$ and a function h satisfying $h(\delta) \to 0$ as $\delta \to 0$, such that, for any $\delta \in (0, \delta_0]$, and any $(\mathbf{w}, \mathbf{v})$ satisfying $\|\mathbf{w} - \mathbf{w}^*\| \le \delta$ and $\|\mathbf{v} - \mathbf{v}^*\| \le \delta$, we have*

$$\phi(\mathbf{w}^*, \mathbf{v}) \le \phi(\mathbf{w}^*, \mathbf{v}^*) \le \max_{\bar{\mathbf{v}}: \|\bar{\mathbf{v}} - \mathbf{v}^*\| \le h(\delta)} \phi(\mathbf{w}, \bar{\mathbf{v}});$$

III. *stationary point of ϕ, if for any local MMP $(\mathbf{w}^*, \mathbf{v}^*)$, we have*

$$\nabla_{\mathbf{w}^*} \phi(\mathbf{w}^*, \mathbf{v}^*) = \nabla_{\mathbf{v}^*} \phi(\mathbf{w}^*, \mathbf{v}^*) = 0.$$

Definition C.4. *For all $(\mathbf{w}, \mathbf{v})$ and any $\zeta > 0$, a point $(\mathbf{w}_\zeta, \mathbf{v}_\zeta)$ is*

I. *a ζ-global MMP of ϕ, if $\phi(\mathbf{w}_\zeta, \mathbf{v}) \le \phi(\mathbf{w}_\zeta, \mathbf{v}_\zeta) \le \max_{\bar{\mathbf{v}}} \phi(\mathbf{w}, \bar{\mathbf{v}}) + \zeta$;*

II. *a ζ-local MMP of ϕ, if there exists $\delta_0 > 0$ and a function h satisfying $h(\delta) \to 0$ as $\delta \to 0$, such that for any $\delta \in (0, \delta_0]$, and any $(\mathbf{w}, \mathbf{v})$ satisfying $\|\mathbf{w} - \mathbf{w}_\zeta\| \le \delta$ and $\|\mathbf{v} - \mathbf{v}_\zeta\| \le \delta$, we have*

$$\phi(\mathbf{w}_\zeta, \mathbf{v}) \le \phi(\mathbf{w}_\zeta, \mathbf{v}_\zeta) \le \max_{\bar{\mathbf{v}}: \|\bar{\mathbf{v}} - \mathbf{v}_\zeta\| \le h(\delta)} \phi(\mathbf{w}, \bar{\mathbf{v}}) + \zeta.$$

We use Definition C.3 which corresponds to a *Stackelberg equilibrium* (Jin et al., 2020a, Definition 9) instead of a *Nash equilibrium* (Jin et al., 2020a, Definition 1) because our formulation $\min_{\mathbf{w}} \max_{\mathbf{v}} \phi(\mathbf{w}, \mathbf{v})$ represents a *sequential game* where the adversary $\mathbf{v} = [\boldsymbol{\epsilon}; \boldsymbol{\nu}]$ observes model parameters $\mathbf{w} = [\mathbf{x}; \mathbf{y}]$ and then chooses optimal perturbations. This gives the max-player a second-mover advantage, captured by $\max_{\hat{\mathbf{v}}} \phi(\mathbf{w}, \hat{\mathbf{v}})$, whereas a Nash equilibrium assumes simultaneous play and does not reflect this asymmetry in adversarial optimization.

Recall the inner solution set $\mathcal{S}(\mathbf{x}) = \arg\min_{\mathbf{y} \in \mathbb{R}^{d_2}} g(\mathbf{x}, \mathbf{y} + \boldsymbol{\nu}^*(\mathbf{x}, \mathbf{y}))$, so that $\mathbf{y} \in \mathcal{S}(\mathbf{x})$ if and only if $d_{\mathcal{S}(\mathbf{x})}(\mathbf{y}) = 0$, where $d_{\mathcal{S}}(\mathbf{z}) = \min_{\mathbf{z}' \in \mathcal{S}} \|\mathbf{z} - \mathbf{z}'\|$. Problem (5) can then be equivalently reformulated as:

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^{d_1}, \mathbf{y} \in \mathbb{R}^{d_2}} \max_{\|\boldsymbol{\epsilon}\| \le \rho, \|\boldsymbol{\nu}\| \le \eta} \quad & f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu}) \\ \text{subj. to} \quad & d^2_{\mathcal{S}(\mathbf{x})}(\mathbf{y}) = 0. \end{aligned} \tag{27}$$

The squared distance function $d^2_{\mathcal{S}}$ is generally non-differentiable and nonconvex, making it computationally intractable for gradient-based optimization. This non-smoothness arises because the solution set $\mathcal{S}(\mathbf{x})$ may have complex geometry, and computing distances to such sets involves non-smooth minimization operations. To circumvent this difficulty, we provide a smooth and computationally tractable surrogate based on the inner optimality gap (7), denoted as:

$$p(\mathbf{x}, \mathbf{y}) := g(\mathbf{x}, \mathbf{y} + \boldsymbol{\nu}^*(\mathbf{x}, \mathbf{y})) - g^*(\mathbf{x}), \tag{28}$$

where $g^*(\mathbf{x}) = \min_{\mathbf{y}} g(\mathbf{x}, \mathbf{y} + \boldsymbol{\nu}^*(\mathbf{x}, \mathbf{y}))$.

The function $p(\mathbf{x}, \mathbf{y})$ measures how much the current inner objective value exceeds the best achievable value: when $p(\cdot) = 0$, the inner variable $\mathbf{y}$ is exactly optimal, and when $p(\cdot) > 0$, it quantifies the degree of suboptimality. The following lemma establishes the theoretical foundation for using $p(\cdot)$ as a surrogate for the distance-based constraint:

Lemma C.5. *Suppose Assumption 2.1 (the μ–PL condition) holds. Then, for all $(\mathbf{x}, \mathbf{y})$, we have $p(\mathbf{x}, \mathbf{y}) \ge 0$ and $\mathbf{y} \in \mathcal{S}(\mathbf{x})$ if and only if $p(\mathbf{x}, \mathbf{y}) = 0$. Moreover, there exists a PL constant $\mu > 0$ such that $d^2_{\mathcal{S}(\mathbf{x})}(\mathbf{y}) \le \mu\, p(\mathbf{x}, \mathbf{y})$.*

The proof draws on (Shen & Chen, 2023) and is provided in Appendix E.1. As shown by Karimi et al. (2016), the PL condition implies the quadratic growth (QG) condition. The lemma establishes that the optimality gap $p(\cdot)$ is a faithful surrogate for the distance constraint: the first part guarantees nonnegativity, the second provides exact equivalence between $p(\cdot) = 0$ and inner optimality, and the third ensures that controlling the optimality gap also controls proximity to the solution set. Consequently, we substitute the non-differentiable constraint $d^2_{\mathcal{S}}(\mathbf{y}) = 0$ in (27) with the tractable condition $p(\mathbf{x}, \mathbf{y}) = 0$, thereby enabling efficient optimization while preserving theoretical guarantees.

The constraint in (8) is not strictly feasible; it can only be satisfied with equality. Following Sow et al. (2022), we relax it by introducing a constant $\zeta_\gamma > 0$, yielding $p(\mathbf{x}, \mathbf{y}) - \zeta_\gamma \leq 0$, which admits strictly feasible points. With $\zeta_\gamma > 0$, we consider

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^{d_1}, \mathbf{y}\in\mathbb{R}^{d_2}} \max_{\|\boldsymbol{\epsilon}\|\leq\rho, \|\boldsymbol{\nu}\|\leq\eta} & f(\mathbf{x}+\boldsymbol{\epsilon}, \mathbf{y}+\boldsymbol{\nu}) \\ \text{subj. to} \quad & p(\mathbf{x}, \mathbf{y}) \leq \zeta_\gamma, \end{aligned} \tag{29}$$

Setting $\zeta_\gamma = 0$ recovers (5). When $\zeta_\gamma > 0$, (29) is a ζ_γ-approximation of (5) since, by Lemma C.5, $p(\cdot)$ upper-bounds $d^2_{S(\cdot)}$ and remains below ζ_γ.

Definition C.6. *We say that Problem (29) is a ζ-approximate formulation of the bilevel problem (5) if the following two conditions hold:*

I. *Under Assumption 2.1, every feasible solution $(\mathbf{x}, \mathbf{y})$ of (29) satisfies $d^2_{S(\mathbf{x})}(\mathbf{y}) \leq \mu\zeta$;*

II. *Under Assumptions 2.3–I., the difference in the optimal objective values of (5) and (29) is bounded as $|f^* - f^*_\zeta| \leq \ell_{f,0}\sqrt{\mu\zeta}$, where f^* and f^*_ζ denote, respectively, the optimal objective values of (5) and (29).*

Lemma C.7. *Suppose Assumptions 2.1 and 2.3–I. hold, then Problem (29) is an ζ-approximate problem of Problem (5).*

The proof is deferred to Appendix E.2. We analyze the relationship between the SABO in (5) and its corresponding Lagrangian in (12). We show that, under standard conditions on the inner function g, approximate global (or local) solutions to (5) can be obtained by globally (or locally) solving (12). To relate the solutions of (12) and (5), we require the solution $(\mathbf{x}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\epsilon}_\gamma, \boldsymbol{\nu}_\gamma)$ of (12) to be feasible for (29). By Assumption 2.3–I., the growth of $f(\mathbf{x}_\gamma + \boldsymbol{\epsilon}_\gamma, \cdot)$ is controlled. Hence, increasing γ in (12) penalizes $p(\cdot)$ and reduces $p(\mathbf{x}_\gamma, \mathbf{y}_\gamma)$.

Theorem C.8. *Suppose Assumptions 2.1 and 2.3–I. hold. For any $\varrho > 0$, let $\zeta = \ell_{f,0}\sqrt{\mu\varrho}/2$ and $\gamma^* = \ell_{f,0}\sqrt{\mu\varrho^{-1}}/2$.*

I. *Any global MMP of (5) is a ζ-global MMP of (12), provided $\gamma \geq \gamma^*$.*

II. *Conversely, if $(\mathbf{x}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\epsilon}_\gamma, \boldsymbol{\nu}_\gamma)$ is a ζ-global MMP of (12) with $\gamma \geq 3\gamma^*$, then $(\mathbf{x}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\epsilon}_\gamma, \boldsymbol{\nu}_\gamma)$ also solves (29) for some $\zeta_\gamma \leq \varrho$.*

The proof is deferred to Appendix E.3. Theorem C.8 shows that with a sufficiently large penalty $\gamma \approx \ell^2_{f,0}\mu/\zeta$, solving the penalized game (12) recovers (up to tolerance ζ) the solutions of SABO (5). When γ is large enough, the constraint via $p(\mathbf{x}, \mathbf{y})$ is strongly enforced, and (29) closely approximates (5). The threshold γ^* scales with the PL constant μ and $\ell^2_{f,0}$. Since (12) is generally nonconvex, we also establish a connection with local solutions.

Theorem C.9. *Suppose Assumptions 2.1 and 2.3–I. hold. If $(\mathbf{x}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\epsilon}_\gamma, \boldsymbol{\nu}_\gamma)$ is a local MMP solution of (12) with $\gamma \geq \ell_{f,0}\sqrt{\mu\varrho^{-1}}$ for some $\varrho > 0$ on $\mathcal{N}((\mathbf{x}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\epsilon}_\gamma, \boldsymbol{\nu}_\gamma), r)$, then it is also a local MMP solution of (29) for some $\zeta_\gamma \leq \varrho$.*

The proof of Theorem C.9 is in Appendix E.4. Theorems C.8 and C.9 show that solving (12) locally yields a global/local solution to an approximate version of (5). In particular, Theorem C.9 justifies the *continuation* schedule used in Algorithm 1: start with a moderate γ to locate a stable basin, then increase γ geometrically while reducing the outer step size in (15c) until the inner gap becomes small. The radii (ρ, η) determine how much flatness we enforce—larger values push the iterates toward broader valleys but also stiffen the landscape—so we tie their growth to the progress in Stage I (the inner steps T_k in (10)) to keep the proxy $\tilde{\mathbf{y}}_k$ aligned with $\mathbf{y}_k$.

Theorems C.8 and C.9 show that choosing $\gamma = \Omega(\varrho^{-0.5})$ guarantees $\zeta_\gamma \leq \varrho$. Next, we show that this bound cannot be improved further.

Corollary C.10. *(Lower bound on the penalty constant) Theorems C.8 and C.9 establish that obtaining $\zeta_\gamma = \mathcal{O}(\varrho)$ necessitates a lower bound of $\gamma = \Omega(\varrho^{-0.5})$.*

The proof of Corollary C.10 is in Appendix E.5. From Lemma C.5, the intractable distance constraint is replaced by the smooth gap $p(\mathbf{x}, \mathbf{y})$ in (28); the Lagrangian (11) then drives $p(\cdot) \to 0$,

thereby enforcing inner feasibility while minimizing the perturbed outer loss. In practice, one picks a target ζ_γ and monitors $p(\mathbf{x}_k, \mathbf{y}_k)$, the proxy mismatch $\|\mathbf{y}_k - \tilde{\mathbf{y}}_k\|$ for $\tilde{\mathbf{y}}_k$ in (10c), and whether the adversary in (14) reaches the ball boundary. If $p(\cdot) > \zeta_\gamma$ or $\|\mathbf{y}_k - \tilde{\mathbf{y}}_k\|$ stalls, one should increase γ and/or T_k, and reduce the step size. Remember, we optimize a *Stackelberg* game (Definition C.3): stationarity is insufficient if the inner maximization can still increase the objective; the global/local results certify robustness when γ is sufficiently large. Interpreting the constants: a larger PL constant μ (indicating an easier inner landscape) lowers the required γ for a given ζ_γ, while a larger $\ell_{f,0}$ (indicating a more sensitive outer problem) increases it.

D Generalization Discussions

We now analyze the generalization error of SABO and SAMO.

D.1 Generalization Analysis for SABO

To investigate the generalization properties of SABO, we establish theoretical bounds that explicitly incorporate the sharpness of the Lagrangian minimization problem (12) within a *stochastic* regime. These bounds are crucial for characterizing the gap between empirical minimizers and true population-level performance, particularly because sharpness-aware perturbations inherently encode the local geometry of the loss landscape. To formalize this analysis, we first define the relevant population-level objectives. Assume each data sample $(\xi, \zeta) \in \mathcal{Z}$ is generated from the underlying multivariate distributions $\mathcal{D}_f$ and $\mathcal{D}_g$, respectively. For a parameter space ϑ, let $(\mathbf{x}, \mathbf{y}) \in \vartheta$ denote the model parameters associated with an example (ξ, ζ). For parameters $(\mathbf{x}, \boldsymbol{\epsilon}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_1}$ and $(\mathbf{y}, \boldsymbol{\nu}) \in \mathbb{R}^{d_2} \times \mathbb{R}^{d_2}$, the standard and perturbed population risks are defined as:

$$\begin{aligned}\mathcal{L}_\gamma(\mathbf{x}, \mathbf{y}; \mathcal{D}_f, \mathcal{D}_g) &:= \mathbb{E}_{\xi\sim\mathcal{D}_f}[f(\mathbf{x}, \mathbf{y}; \xi)] \\ &\quad + \gamma\mathbb{E}_{\zeta\sim\mathcal{D}_g}[g(\mathbf{x}, \mathbf{y} + \boldsymbol{\nu}^*; \zeta) - g^*(\mathbf{x}; \zeta)], \qquad (30a)\\ \mathcal{L}_\gamma(\mathbf{x}, \mathbf{y}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{D}_f, \mathcal{D}_g) &:= \mathbb{E}_{\xi\sim\mathcal{D}_f}[f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu}; \xi)] \\ &\quad + \gamma\mathbb{E}_{\zeta\sim\mathcal{D}_g}[g(\mathbf{x}, \mathbf{y} + \boldsymbol{\nu}^*; \zeta) - g^*(\mathbf{x}; \zeta)], \qquad (30b)\end{aligned}$$

where $g^*(\cdot)$ is defined in (6).

Because the underlying distributions $\mathcal{D}_f$ and $\mathcal{D}_g$ are typically *unknown*, rendering the exact computation of the population risk intractable, standard practice relies on empirical approximations derived from finite training datasets. Let $\mathcal{S}_f = \{\xi_i\}_{i=1}^{n_1}$ and $\mathcal{S}_g = \{\zeta_i\}_{i=1}^{n_2}$ be datasets drawn i.i.d. from a sample space $\mathcal{Z}$ according to $\mathcal{D}_f$ and $\mathcal{D}_g$, respectively, yielding the following empirical perturbed losses:

$$\begin{aligned}\mathcal{L}_\gamma(\mathbf{x}, \mathbf{y}; \mathcal{S}_f, \mathcal{S}_g) &:= \frac{1}{n_1}\sum_{i=1}^{n_1} f(\mathbf{x}, \mathbf{y}; \xi_i) \\ &\quad + \frac{\gamma}{n_2}\sum_{i=1}^{n_2}\left(g(\mathbf{x}, \mathbf{y} + \boldsymbol{\nu}^*; \zeta_i) - g^*(\mathbf{x}; \zeta_i)\right), \qquad (31a)\\ \mathcal{L}_\gamma(\mathbf{x}, \mathbf{y}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{S}_f, \mathcal{S}_g) &:= \frac{1}{n_1}\sum_{i=1}^{n_1} f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu}; \xi_i) \\ &\quad + \frac{\gamma}{n_2}\sum_{i=1}^{n_2}\left(g(\mathbf{x}, \mathbf{y} + \boldsymbol{\nu}^*; \zeta_i) - g^*(\mathbf{x}; \zeta_i)\right). \qquad (31b)\end{aligned}$$

In the subsequent analysis, we adopt the following notations:

$$C(\mathbf{z}, \iota, n, d) := \frac{\pi^2}{6}\left(1 + \frac{\|\mathbf{z}\|^2}{\iota^2}\left(1 + \sqrt{\frac{\log(n)}{d}}\right)^2\right)^{\frac{d}{2}}(1 + d + n)^2. \qquad (32)$$

Theorem D.1. *For any $\rho > 0$ and $\eta > 0$, and for any distributions $\mathcal{D}_f$ and $\mathcal{D}_g$, assume $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \rho^2\mathbf{I}_{d_1}\right)$ and $\boldsymbol{\nu} \sim \mathcal{N}\left(\mathbf{0}, \eta^2\mathbf{I}_{d_2}\right)$. Suppose the flatness condition*

$$\mathcal{L}_\gamma(\mathbf{x}, \mathbf{y}; \mathcal{D}_f, \mathcal{D}_g) \le \mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{\nu}}[\mathcal{L}_\gamma(\mathbf{x}, \mathbf{y}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{D}_f, \mathcal{D}_g)] \qquad (33)$$

holds. Assume $|f| \leq M_f$ *and* $|g| \leq M_g$*. Then, for any* $\delta \in (0, 1]$*, with probability at least* $1 - \delta$ *over the draws* $\mathcal{S}_f \sim \mathcal{D}_f$ *and* $\mathcal{S}_g \sim \mathcal{D}_g$*,*

$$\mathcal{L}_\gamma(\mathbf{x}, \mathbf{y}; \mathcal{D}_f, \mathcal{D}_g) \leq \max_{\|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| \leq \eta} \mathcal{L}_\gamma(\mathbf{x}, \mathbf{y}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{S}_f, \mathcal{S}_g) + 4\frac{\acute{M}(\gamma)}{\sqrt{n}} + 2\sqrt{\frac{M(\gamma)}{n}} \left(1 + \log\big(C(\mathbf{x}, \rho, n_1, d_1)\, C(\mathbf{y}, \eta, n_2, d_2)\big) + \log\frac{1}{\delta}\right)^{\frac{1}{2}}, \tag{34}$$

where $n := \min(n_1, n_2)$, $M(\gamma) := \frac{M_f^2}{2} + 2\gamma^2 M_g^2$ and $\acute{M}(\gamma) := M_f + 2\gamma M_g$. The proof is provided in Appendix H.1. Theorem D.1 builds upon the general PAC-Bayes framework (Alquier et al., 2016) and accommodates generic bounded loss functions, rather than being restricted to the binary classification losses analyzed in single-level SAM (Foret et al., 2020). The bound in (34) reveals that the generalization error depends on three key factors: (i) the worst-case empirical loss within the perturbation radii (ρ, η), (ii) a stochastic term of order $\mathcal{O}(1/\sqrt{n})$ capturing sample uncertainty, and (iii) a complexity term that scales logarithmically with $C(\cdot)$.

The term $C(\cdot)$ shows how the geometry of the sharpness-aware perturbations influences generalization:

$$\left(1 + \frac{\|\mathbf{x}\|^2}{\rho^2}\right)^{d_1/2} \left(1 + \frac{\|\mathbf{y}\|^2}{\eta^2}\right)^{d_2/2} \mathrm{poly}(n_1, n_2, d_1, d_2).$$

As $(\rho, \eta) \to 0$ (corresponding to the standard bilevel setting without sharpness smoothing), $\log C(\cdot)$ diverges, rendering the bound vacuous. Conversely, finite perturbations $(\rho, \eta) > 0$ induce local averaging over flatter regions of the loss landscape, thereby tightening the bound and improving generalization guarantees. Moreover, since the logarithm decreases monotonically with respect to (ρ, η), increasing the perturbation strength (ρ, η) makes the bound tighter. This theoretical analysis formally substantiates the intuition that sharpness-aware perturbations actively steer the optimization trajectory toward flatter minima, thereby improving population-level performance.

The bound in (34) exhibits several asymptotic properties. As $n_1, n_2 \to \infty$, the statistical terms in the generalization bound decrease at a rate of $\mathcal{O}((\log n_1 + \log n_2)^{1/2}/n^{1/2})$, up to factors depending on the problem parameters. The bound also exhibits a sublinear dependence on the problem dimensions as $\mathcal{O}((d_1 + d_2)^{1/2})$ (arising from the Gaussian smoothing construction). In the small-perturbation regime $\rho, \eta \to 0$, where SABO reduces to the standard BO formulation, the penalty constant must scale as $\gamma = \mathcal{O}(\rho\eta)$ to preserve a non-vacuous bound. Combined with Corollary C.10, which establishes that attaining $\zeta_\gamma = \mathcal{O}(\varrho)$ requires $\gamma = \Omega(\varrho^{-1/2})$, these conditions jointly delineate the admissible scaling of γ as a function of the perturbation strength, with the sharpness-aware regime $\gamma \neq \rho\eta$ recovering the flat-minima behavior characterized in our convergence analysis. These results establish that SABO achieves *provably superior* generalization compared to standard bilevel methods, which lack such finite guarantees under this analysis framework. Finally, we note that setting $\mathbf{y} = 0$, $\boldsymbol{\nu} = 0$, and removing the inner objective reduces our framework to single-level optimization, and Theorem D.1 recovers the standard SAM bound (Foret et al., 2020).

D.2 Generalization Analysis for SAMO

We now analyze how solutions obtained by SAMO generalize from empirical samples to the population risk. Let $\mathcal{S} = \{\xi_i\}_{i=1}^n$ denote a sample set of size n, drawn i.i.d. from a space $\mathcal{Z}$ according to an *unknown* distribution $\mathcal{D}$. Let ϑ denote the parameter space, and let $\mathbf{x} \in \vartheta$ be the model parameters associated with an example $\xi \in \mathcal{Z}$. For a given $\mathbf{x}$, define $\mathbf{y}^* := \arg\max_{\mathbf{y} \in \mathbb{R}^{d_2}} \mathbb{E}_{\xi \sim \mathcal{D}} f(\mathbf{x}, \mathbf{y}; \xi)$ as the optimal ascent variable. Given the distribution $\mathcal{D}$, we define

$$F(\mathbf{x}; \mathcal{D}) := \mathbb{E}_{\xi \sim \mathcal{D}} f(\mathbf{x}, \mathbf{y}^*; \xi), \tag{35a}$$

$$F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{D}) := \mathbb{E}_{\xi \sim \mathcal{D}} f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y}^* + \boldsymbol{\nu}; \xi). \tag{35b}$$

For a dataset $\mathcal{S}$ with $|\mathcal{S}| = n$, define

$$F(\mathbf{x}; \mathcal{S}) := \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}, \mathbf{y}^*; \xi_i), \tag{36a}$$

$$F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{S}) := \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y}^* + \boldsymbol{\nu}; \xi_i). \tag{36b}$$

With these definitions in place, we are ready to establish the generalization performance.

Theorem D.2. *For any $(\rho, \eta) > 0$, and for any distribution $\mathcal{D}$, assume $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \rho^2 \mathbf{I}_{d_1}\right)$ and $\boldsymbol{\nu} \sim \mathcal{N}\left(\mathbf{0}, \eta^2 \mathbf{I}_{d_2}\right)$. Suppose the flatness condition*

$$F(\mathbf{x}; \mathcal{D}) \leq \mathbb{E}_{\boldsymbol{\epsilon}, \boldsymbol{\nu}}[F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{D})] \tag{37}$$

holds. Assume $|f| \leq M$. Then, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ over the draws $\mathcal{S} \sim \mathcal{D}$ of size n,

$$\begin{aligned} F(\mathbf{x}; \mathcal{D}) \leq & \max_{\|\boldsymbol{\epsilon}\| \leq \rho,\ \|\boldsymbol{\nu}\| \leq \eta} F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{S}) + \frac{4M}{\sqrt{n}} \\ & + \frac{\sqrt{2}M}{\sqrt{n}} \left(1 + \log\big(C(\mathbf{x}, \rho, n, d_1)\, C(\mathbf{y}^*, \eta, n, d_2)\big) + \log \frac{1}{\delta}\right)^{\frac{1}{2}}, \end{aligned} \tag{38}$$

where $C(\cdot)$ is defined in (32).

The proof is provided in Appendix I.2. Theorem D.2 bounds the population loss by the worst-case perturbed empirical loss plus two terms that decay with the sample size n. The key insight lies in the complexity term $C(\cdot)$ from (32), which depends inversely on the perturbation radii through the ratios $\|\mathbf{x}\|^2/\rho^2$ and $\|\mathbf{y}^*\|^2/\eta^2$. Small perturbations make these ratios large, inflating the bound and preventing meaningful generalization guarantees. Moderate perturbations enforce flatness by exploring neighborhoods around iterates, yielding solutions robust to input changes. However, excessively large perturbations distort the objective and hinder convergence. This reveals a fundamental trade-off: perturbations must be large enough to promote generalization yet small enough to maintain proximity to the original problem. Condition (37) formalizes this by requiring the solution to lie in a relatively flat region of the loss landscape.

E Proofs omitted from Section C.3

We first present the following lemma, whose proof follows the same arguments as those used in the proof of Lemma A.3 in (Yang et al., 2020).

Lemma E.1. *Suppose Assumption 2.1 holds. Then, for any fixed $\mathbf{x}$, the function $\phi_{\mathbf{x}}(\mathbf{y}) := \max_{\|\boldsymbol{\nu}\| \leq \eta} g(\mathbf{x}, \mathbf{y} + \boldsymbol{\nu})$ satisfies the PL condition with respect to $\mathbf{y}$ with the same constant μ.*

Proof. By Danskin's theorem (see Lemma A.5 of (Nouiehed et al., 2019)), it holds that,

$$\|\nabla_{\mathbf{y}} \phi_{\mathbf{x}}(\mathbf{y})\|^2 = \|\nabla_{\mathbf{y}} g(\mathbf{x}, \mathbf{y} + \boldsymbol{\nu}^*(\mathbf{x}, \mathbf{y}))\|^2,$$

where $\boldsymbol{\nu}^*(\mathbf{x}, \mathbf{y}) = \arg\max_{\|\boldsymbol{\nu}\| \leq \eta} g(\mathbf{x}, \mathbf{y} + \boldsymbol{\nu})$. By the PL condition for $g(\cdot, \boldsymbol{\nu})$ with constant μ in Assumption 2.1, we obtain

$$2\left[g(\mathbf{x}, \mathbf{y} + \boldsymbol{\nu}^*(\mathbf{x}, \mathbf{y})) - \min_{\hat{\mathbf{y}}} g(\mathbf{x}, \hat{\mathbf{y}} + \boldsymbol{\nu}^*(\mathbf{x}, \mathbf{y}))\right] \leq \mu \|\nabla_{\mathbf{y}} \phi_{\mathbf{x}}(\mathbf{y})\|^2. \tag{39}$$

Moreover, we have

$$g(\mathbf{x}, \hat{\mathbf{y}} + \boldsymbol{\nu}^*(\mathbf{x}, \mathbf{y})) \leq \max_{\|\boldsymbol{\nu}\| \leq \eta} g(\mathbf{x}, \hat{\mathbf{y}} + \boldsymbol{\nu}),$$

which implies

$$\min_{\hat{\mathbf{y}}} g(\mathbf{x}, \hat{\mathbf{y}} + \boldsymbol{\nu}^*(\mathbf{x}, \mathbf{y})) \leq \min_{\hat{\mathbf{y}}} \max_{\|\boldsymbol{\nu}\| \leq \eta} g(\mathbf{x}, \hat{\mathbf{y}} + \boldsymbol{\nu}) = \phi_{\mathbf{x}}^*, \tag{40}$$

where $\phi_{\mathbf{x}}^* := \min_{\mathbf{y}} \phi_{\mathbf{x}}(\mathbf{y})$. Using (39) and (40) together, we obtain

$$2\left(\phi_{\mathbf{x}}(\mathbf{y}) - \phi_{\mathbf{x}}^*\right) \leq \mu \|\nabla_{\mathbf{y}} \phi_{\mathbf{x}}(\mathbf{y})\|^2.$$

This concludes the proof. □

E.1 Proof of Lemma C.5

Proof. It is clear that

$$p(\mathbf{x},\mathbf{y}) = g(\mathbf{x},\mathbf{y}+\boldsymbol{\nu}^*) - g^*(\mathbf{x}) \geq 0,$$

where $\boldsymbol{\nu}^* = \arg\max_{\|\boldsymbol{\nu}\|\leq\eta} g(\mathbf{x},\mathbf{y}+\boldsymbol{\nu})$. Since the solution set of the inner Problem (5), $\mathcal{S}(\mathbf{x})$, is closed, it follows that

$$\mathbf{y}^* \in \mathcal{S}(\mathbf{x}), \quad \text{if and only if}, \quad d_{\mathcal{S}(\mathbf{x})}(\mathbf{y}) = 0.$$

Then by the definition of $\mathcal{S}(\mathbf{x})$, it holds for any $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$ that

$$\begin{aligned} &p(\mathbf{x},\mathbf{y}) = g(\mathbf{x},\mathbf{y}+\boldsymbol{\nu}^*) - g^*(\mathbf{x}) = 0. \\ &\text{if and only if}, \quad \mathbf{y} \in \mathcal{S}(\mathbf{x}), \end{aligned}$$

which implies

$$\begin{aligned} &p(\mathbf{x},\mathbf{y}) = g(\mathbf{x},\mathbf{y}+\boldsymbol{\nu}^*) - g^*(\mathbf{x}) = 0, \\ &\text{if and only if}, \quad d_{\mathcal{S}(\mathbf{x})}(\mathbf{y}) = 0. \end{aligned}$$

Therefore, it is sufficient to verify if $g(\mathbf{x},\mathbf{y}+\boldsymbol{\nu}^*) - g^*(\mathbf{x})$ is an upper bound of $d^2_{\mathcal{S}(\mathbf{x})}(\mathbf{y})$.
By Lemma E.1, which guarantees that $\max_{\|\boldsymbol{\nu}\|\leq\eta} g(\mathbf{x},\mathbf{y}+\boldsymbol{\nu})$ satisfies the PL condition, we apply Lemma B.1 to any $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$, to obtain

$$\begin{aligned} p(\mathbf{x},\mathbf{y}) &= g(\mathbf{x},\mathbf{y}+\boldsymbol{\nu}^*) - \min_{\mathbf{y}\in\mathbb{R}^{d_2}} g(\mathbf{x},\mathbf{y}+\boldsymbol{\nu}^*) \\ &= \max_{\|\boldsymbol{\nu}\|\leq\eta} g(\mathbf{x},\mathbf{y}+\boldsymbol{\nu}) - \min_{\mathbf{y}\in\mathbb{R}^{d_2}} \max_{\|\boldsymbol{\nu}\|\leq\eta} g(\mathbf{x},\mathbf{y}+\boldsymbol{\nu}) \\ &\geq \frac{1}{\mu} d^2_{\mathcal{S}(\mathbf{x})}(\mathbf{y}), \end{aligned}$$

where the second equality follows since $\boldsymbol{\nu}^*$ is a maximizer for every $\mathbf{y}$. □

E.2 Proof of Lemma C.7

Proof. By Lemma C.5, any feasible solution of (29) immediately satisfies $d^2_{\mathcal{S}(\mathbf{x})}(\mathbf{y}) \leq \mu\zeta$.
Next we prove $|f^* - f^*_\zeta| \leq \ell_{f,0}\sqrt{\mu\zeta}$. Let $(\mathbf{x}_\zeta,\mathbf{y}_\zeta,\boldsymbol{\epsilon}_\zeta,\boldsymbol{\nu}_\zeta)$ be a global MMP solution of (29) with $f^*_\zeta = f(\mathbf{x}_\zeta,\mathbf{y}_\zeta,\boldsymbol{\epsilon}_\zeta,\boldsymbol{\nu}_\zeta)$. Since $\mathcal{S}(\mathbf{x}_\zeta)$ is non-empty and closed, one can find $\bar{\mathbf{y}}_\zeta \in \mathcal{S}(\mathbf{x}_\zeta)$ such that $d_{\mathcal{S}(\mathbf{x}_\zeta)}(\mathbf{y}_\zeta) = \|\bar{\mathbf{y}}_\zeta - \mathbf{y}_\zeta\|$. Then it holds

$$\begin{aligned} f(\mathbf{x}_\zeta,\bar{\mathbf{y}}_\zeta,\boldsymbol{\epsilon},\boldsymbol{\nu}) - f(\mathbf{x}_\zeta,\mathbf{y}_\zeta,\boldsymbol{\epsilon},\boldsymbol{\nu}) &\leq \ell_{f,0}\|\bar{\mathbf{y}}_\zeta - \mathbf{y}_\zeta\| \\ &= \ell_{f,0} d_{\mathcal{S}(\mathbf{x}_\zeta)}(\mathbf{y}_\zeta) \\ &\leq \ell_{f,0}\sqrt{\mu\zeta}, \end{aligned} \tag{41}$$

where the last inequality holds because, from Lemma C.5, we have $d^2_{\mathcal{S}(\mathbf{x}_\zeta)}(\mathbf{y}_\zeta) \leq \mu\zeta$. Let $(\mathbf{x}^*,\mathbf{y}^*,\boldsymbol{\epsilon}^*,\boldsymbol{\nu}^*)$ be a global MMP solution of (5) so that

$$f^* = f(\mathbf{x}^*,\mathbf{y}^*,\boldsymbol{\epsilon}^*,\boldsymbol{\nu}^*).$$

Since $(\mathbf{x}_\zeta,\bar{\mathbf{y}}_\zeta,\boldsymbol{\epsilon}_\zeta,\boldsymbol{\nu}_\zeta)$ is feasible for (5), we have

$$f(\mathbf{x}^*,\mathbf{y}^*,\boldsymbol{\epsilon}^*,\boldsymbol{\nu}^*) \leq \max_{\|\boldsymbol{\epsilon}\|\leq\rho,\|\boldsymbol{\nu}\|\leq\eta} f(\mathbf{x}_\zeta,\bar{\mathbf{y}}_\zeta,\boldsymbol{\epsilon},\boldsymbol{\nu}). \tag{42}$$

The inequalities (42) and (41) imply

$$\begin{aligned} f^* - f^*_\zeta &= f^* - f(\mathbf{x}_\zeta,\mathbf{y}_\zeta,\boldsymbol{\epsilon}_\zeta,\boldsymbol{\nu}_\zeta) \\ &\leq \max_{\|\boldsymbol{\epsilon}\|\leq\rho,\|\boldsymbol{\nu}\|\leq\eta} \{f(\mathbf{x}_\zeta,\bar{\mathbf{y}}_\zeta,\boldsymbol{\epsilon},\boldsymbol{\nu})\} - f(\mathbf{x}_\zeta,\mathbf{y}_\zeta,\boldsymbol{\epsilon}_\zeta,\boldsymbol{\nu}_\zeta) \\ &\leq \max_{\|\boldsymbol{\epsilon}\|\leq\rho,\|\boldsymbol{\nu}\|\leq\eta} \{f(\mathbf{x}_\zeta,\bar{\mathbf{y}}_\zeta,\boldsymbol{\epsilon},\boldsymbol{\nu}) - f(\mathbf{x}_\zeta,\mathbf{y}_\zeta,\boldsymbol{\epsilon},\boldsymbol{\nu})\} \\ &\leq \ell_{f,0}\sqrt{\mu\zeta}, \end{aligned}$$

where the second inequality follows from $\max(a) - \max(b) \leq \max(a - b)$, and the fact that $(\mathbf{x}_\zeta, \mathbf{y}_\zeta, \boldsymbol{\epsilon}_\zeta, \boldsymbol{\nu}_\zeta)$ is a global minimax point. In particular, $(\boldsymbol{\epsilon}_\zeta, \boldsymbol{\nu}_\zeta)$ maximizes $f(\mathbf{x}_\zeta, \mathbf{y}_\zeta, \boldsymbol{\epsilon}, \boldsymbol{\nu})$ over the feasible perturbation set.
Since $(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\epsilon}^*, \boldsymbol{\nu}^*)$ is a global MMP solution of (5), the result follows from the fact that

$$f_\zeta^* = f(\mathbf{x}_\zeta, \mathbf{y}_\zeta, \boldsymbol{\epsilon}_\zeta, \boldsymbol{\nu}_\zeta) \leq \max_{\|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| \leq \eta} \{f(\mathbf{x}, \mathbf{y}, \boldsymbol{\epsilon}, \boldsymbol{\nu})\} \leq f(\mathbf{x}, \mathbf{y}, \boldsymbol{\epsilon}^*, \boldsymbol{\nu}^*) \leq f^*.$$

□

E.3 Proof of Theorem C.8

Proof. **To establish part I**: Based on Definition C.3, we prove that $(\mathbf{x}^\circ, \mathbf{y}^\circ, \boldsymbol{\epsilon}^\circ, \boldsymbol{\nu}^\circ)$ is a global minimax point of (5), i.e.,

$$f(\mathbf{x}^\circ + \boldsymbol{\epsilon}, \mathbf{y}^\circ + \boldsymbol{\nu}) \leq f(\mathbf{x}^\circ + \boldsymbol{\epsilon}^\circ, \mathbf{y}^\circ + \boldsymbol{\nu}^\circ) \leq \max_{\|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| \leq \eta} f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu}), \tag{43}$$

for all feasible $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, $\|\boldsymbol{\epsilon}\| \leq \rho$, and $\|\boldsymbol{\nu}\| \leq \eta$ if and only if

$$\begin{aligned} & f(\mathbf{x}^\circ + \boldsymbol{\epsilon}, \mathbf{y}^\circ + \boldsymbol{\nu}) + \gamma^* p(\mathbf{x}^\circ, \mathbf{y}^\circ) \\ \leq\ & f(\mathbf{x}^\circ + \boldsymbol{\epsilon}^\circ, \mathbf{y}^\circ + \boldsymbol{\nu}^\circ) + \gamma^* p(\mathbf{x}^\circ, \mathbf{y}^\circ) \\ \leq\ & \max_{\|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| \leq \eta} \{f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu}) + \gamma^* p(\mathbf{x}, \mathbf{y})\} + \zeta. \end{aligned} \tag{44}$$

Now, we prove that (43) implies (44). Recall that

$$\mathcal{S}(\mathbf{x}) = \arg\min_{\mathbf{y} \in \mathbb{R}^{d_2}} g(\mathbf{x}, \mathbf{y} + \boldsymbol{\nu}^*),$$

where $\boldsymbol{\nu}^* = \arg\max_{\|\boldsymbol{\nu}\| \leq \eta} g(\mathbf{x}, \mathbf{y} + \boldsymbol{\nu})$. Given any $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, since $\mathcal{S}(\mathbf{x})$ is closed and nonempty, let $\mathbf{y}_\mathbf{x} \in \arg\min_{\acute{\mathbf{y}} \in \mathcal{S}(\mathbf{x})} \|\mathbf{y} - \acute{\mathbf{y}}\|$. By Assumption 2.3–I., it holds that

$$\begin{aligned} & \max_{\|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| \leq \eta} f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y}_\mathbf{x} + \boldsymbol{\nu}) - \max_{\|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| \leq \eta} f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu}) \\ & \leq \ell_{f,0} \|\mathbf{y} - \mathbf{y}_\mathbf{x}\| \\ & = \ell_{f,0} d_{\mathcal{S}(\mathbf{x})}(\mathbf{y}), \end{aligned}$$

which implies that

$$\begin{aligned} & \max_{\|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| \leq \eta} f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu}) - \max_{\|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| \leq \eta} f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y}_\mathbf{x} + \boldsymbol{\nu}) \\ & \geq -\ell_{f,0} d_{\mathcal{S}(\mathbf{x})}(\mathbf{y}). \end{aligned}$$

Thus, it can be concluded that

$$\begin{aligned} & \max_{\|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| \leq \eta} f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu}) \\ & - \max_{\|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| \leq \eta} f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y}_\mathbf{x} + \boldsymbol{\nu}) + \gamma^* p(\mathbf{x}, \mathbf{y}) \\ & \geq -\ell_{f,0} d_{\mathcal{S}(\mathbf{x})}(\mathbf{y}) + \gamma^* p(\mathbf{x}, \mathbf{y}) \\ & \geq -\ell_{f,0} d_{\mathcal{S}(\mathbf{x})}(\mathbf{y}) + \frac{\gamma^*}{\mu} d^2_{\mathcal{S}(\mathbf{x})}(\mathbf{y}) \\ & \geq \min_{z \in \mathbb{R}_+} \left\{ -\ell_{f,0} z + \frac{\gamma^*}{\mu} z^2 \right\} =: -\zeta. \end{aligned} \tag{45}$$

Consequently, $\gamma^* = \ell_{f,0}^2 \mu / 4\zeta = \ell_{f,0} \sqrt{\mu \varrho^{-1}} / 2$.
Since $\mathbf{y}_\mathbf{x} \in \mathcal{S}(\mathbf{x})$, the pair $(\mathbf{x}, \mathbf{y}_\mathbf{x})$ is feasible for (5). Let $f^\circ := f(\mathbf{x}^\circ + \boldsymbol{\epsilon}^\circ, \mathbf{y}^\circ + \boldsymbol{\nu}^\circ)$ denote the optimal objective value of (5). For any feasible $(\mathbf{x}, \mathbf{y}, \boldsymbol{\epsilon}, \boldsymbol{\nu})$, with $\|\boldsymbol{\epsilon}\| \leq \rho$ and $\|\boldsymbol{\nu}\| \leq \eta$, we have

$$f(\mathbf{x}^\circ + \boldsymbol{\epsilon}, \mathbf{y}^\circ + \boldsymbol{\nu}) \leq f^\circ \leq \max_{\|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| \leq \eta} f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu}). \tag{46}$$

In particular, since $\mathbf{y}_\mathbf{x} \in \mathcal{S}(\mathbf{x})$, we obtain

$$f^\circ \leq \max_{\|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| \leq \eta} f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y}_\mathbf{x} + \boldsymbol{\nu}).$$

This, combined with (45), indicates that

$$\begin{aligned}-\zeta &\le \max_{\|\boldsymbol{\epsilon}\|\le\rho,\|\boldsymbol{\nu}\|\le\eta} f(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}+\boldsymbol{\nu})+\gamma^* p(\mathbf{x},\mathbf{y}) \\ &\quad - \max_{\|\boldsymbol{\epsilon}\|\le\rho,\|\boldsymbol{\nu}\|\le\eta} f(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}_{\mathbf{x}}+\boldsymbol{\nu}) \\ &\le \max_{\|\boldsymbol{\epsilon}\|\le\rho,\|\boldsymbol{\nu}\|\le\eta} f(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}+\boldsymbol{\nu})+\gamma^* p(\mathbf{x},\mathbf{y})-f^\circ \\ &\le \max_{\|\boldsymbol{\epsilon}\|\le\rho,\|\boldsymbol{\nu}\|\le\eta} \{f(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}+\boldsymbol{\nu})+\gamma^* p(\mathbf{x},\mathbf{y})\}-f^\circ, \end{aligned} \tag{47}$$

for all $(\mathbf{x},\mathbf{y})\in\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$.
From (7), recall that

$$p(\mathbf{x},\mathbf{y}) = g(\mathbf{x},\ \mathbf{y}+\boldsymbol{\nu}^*(\mathbf{x},\mathbf{y}))-g^*(\mathbf{x}),$$

where $g^*(\mathbf{x})=\min_{\mathbf{y}\in\mathbb{R}^{d_2}} g(\mathbf{x},\ \mathbf{y}+\boldsymbol{\nu}^*(\mathbf{x},\mathbf{y}))$.
Since $(\mathbf{x}^\circ,\mathbf{y}^\circ)$ is feasible, the constraint $p(\mathbf{x}^\circ,\mathbf{y}^\circ)=0$ holds. By (47), we have

$$\begin{aligned}&f(\mathbf{x}^\circ+\boldsymbol{\epsilon}^\circ,\mathbf{y}^\circ+\boldsymbol{\nu}^\circ)+\gamma^* p(\mathbf{x}^\circ,\mathbf{y}^\circ) \\ &\le \max_{\|\boldsymbol{\epsilon}\|\le\rho,\|\boldsymbol{\nu}\|\le\eta} \{f(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}+\boldsymbol{\nu})+\gamma^* p(\mathbf{x},\mathbf{y})\}+\zeta, \end{aligned} \tag{48}$$

for all $(\mathbf{x},\mathbf{y})\in\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$, which implies the right hand side of (44).
To prove the left-hand side of (44), note that $p(\mathbf{x}^\circ,\mathbf{y}^\circ)=0$. Therefore, using the left hand side of (46), we obtain

$$\begin{aligned}&f(\mathbf{x}^\circ+\boldsymbol{\epsilon},\mathbf{y}^\circ+\boldsymbol{\nu})+\gamma^* p(\mathbf{x}^\circ,\mathbf{y}^\circ) \\ &= f(\mathbf{x}^\circ+\boldsymbol{\epsilon},\mathbf{y}^\circ+\boldsymbol{\nu}) \\ &\le f(\mathbf{x}^\circ+\boldsymbol{\epsilon}^\circ,\mathbf{y}^\circ+\boldsymbol{\nu}^\circ) \\ &= f^\circ+\gamma^* p(\mathbf{x}^\circ,\mathbf{y}^\circ).\end{aligned}$$

Thus, we obtain that (43) implies (44). Together with the feasibility of the global solution of (5) for (12), this proves that the global solution of (5) attains the ζ_1-global-minimax for (12).
To establish part II: If $(\mathbf{x}_\gamma,\mathbf{y}_\gamma,\boldsymbol{\epsilon}_\gamma,\boldsymbol{\nu}_\gamma)$ is a ζ-global MMP of (12) with $\gamma\ge 3\gamma^*$ for some $\varrho>0$, then $(\mathbf{x}_\gamma,\mathbf{y}_\gamma,\boldsymbol{\epsilon}_\gamma,\boldsymbol{\nu}_\gamma)$ also solves

$$\begin{aligned}\min_{\mathbf{x}\in\mathbb{R}^{d_1},\mathbf{y}\in\mathbb{R}^{d_2}}\ &\max_{\|\boldsymbol{\epsilon}\|\le\rho,\|\boldsymbol{\nu}\|\le\eta} f(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}+\boldsymbol{\nu}) \\ \text{subj. to}\quad & p(\mathbf{x},\mathbf{y})\le\zeta_\gamma,\end{aligned} \tag{49}$$

with some $\zeta_\gamma\le\varrho$. Now, we want to prove that (44) implies (43).
Given that $(\mathbf{x}_\gamma,\mathbf{y}_\gamma,\boldsymbol{\epsilon}_\gamma,\boldsymbol{\nu}_\gamma)$ is a ζ-global-minimax solution, the following inequality holds for all $(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu})$ that are feasible for (12).

$$\begin{aligned}&f(\mathbf{x}_\gamma+\boldsymbol{\epsilon}_\gamma,\mathbf{y}_\gamma+\boldsymbol{\nu}_\gamma)+\gamma p(\mathbf{x}_\gamma,\mathbf{y}_\gamma) \\ &\le \max_{\|\boldsymbol{\epsilon}\|\le\rho,\|\boldsymbol{\nu}\|\le\eta} \{f(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}+\boldsymbol{\nu})+\gamma p(\mathbf{x},\mathbf{y})\}+\zeta \\ &\le \max_{\|\boldsymbol{\epsilon}\|\le\rho,\|\boldsymbol{\nu}\|\le\eta} \{f(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}+\boldsymbol{\nu})\}+\gamma p(\mathbf{x},\mathbf{y})+\zeta. \end{aligned} \tag{50}$$

By (50), we have

$$\begin{aligned}&f(\mathbf{x}_\gamma+\boldsymbol{\epsilon}_\gamma,\mathbf{y}_\gamma+\boldsymbol{\nu}_\gamma)-\max_{\|\boldsymbol{\epsilon}\|\le\rho,\|\boldsymbol{\nu}\|\le\eta}\{f(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}+\boldsymbol{\nu})\} \\ &\le \gamma\big(p(\mathbf{x},\mathbf{y})-p(\mathbf{x}_\gamma,\mathbf{y}_\gamma)\big)+\zeta.\end{aligned}$$

Now, we want to show that $(\mathbf{x}_\gamma,\boldsymbol{\epsilon}_\gamma,\mathbf{y}_\gamma,\boldsymbol{\nu}_\gamma)$ is feasible for (49). To this end, let $(\mathbf{x}^\circ,\mathbf{y}^\circ,\boldsymbol{\epsilon}^\circ,\boldsymbol{\nu}^\circ)$ be a global solution to Problem (5). Since $\mathbf{y}^\circ\in\mathcal{S}(\mathbf{x})$ by Lemma C.5, we have $p(\mathbf{x},\mathbf{y}^\circ)=0$ for every $\mathbf{x}$. Therefore, choosing $(\mathbf{x},\mathbf{y})=(\mathbf{x}^\circ,\mathbf{y}^\circ)$ in (50) yields

$$\begin{aligned}&f(\mathbf{x}_\gamma+\boldsymbol{\epsilon}_\gamma,\mathbf{y}_\gamma+\boldsymbol{\nu}_\gamma)+\gamma p(\mathbf{x}_\gamma,\mathbf{y}_\gamma) \\ &\le \max_{\|\boldsymbol{\epsilon}\|\le\rho,\|\boldsymbol{\nu}\|\le\eta} f(\mathbf{x}^\circ+\boldsymbol{\epsilon},\mathbf{y}^\circ+\boldsymbol{\nu})+\gamma p(\mathbf{x}^\circ,\mathbf{y}^\circ)+\zeta \\ &= f(\mathbf{x}^\circ+\boldsymbol{\epsilon}^\circ,\mathbf{y}^\circ+\boldsymbol{\nu}^\circ)+\gamma^* p(\mathbf{x}^\circ,\mathbf{y}^\circ)+\zeta \\ &\le \max_{\|\boldsymbol{\epsilon}\|\le\rho,\|\boldsymbol{\nu}\|\le\eta}\{f(\mathbf{x}+\boldsymbol{\epsilon},\mathbf{y}+\boldsymbol{\nu})+\gamma^* p(\mathbf{x},\mathbf{y})\}+2\zeta\quad\text{for any}\quad\mathbf{x},\mathbf{y} \\ &\le f(\mathbf{x}_\gamma+\boldsymbol{\epsilon}_\gamma,\mathbf{y}_\gamma+\boldsymbol{\nu}_\gamma)+\gamma^* p(\mathbf{x}_\gamma,\mathbf{y}_\gamma)+2\zeta,\end{aligned} \tag{51}$$

where the second inequality follows from (48) and last inequality follows since $(\mathbf{x}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\epsilon}_\gamma, \boldsymbol{\nu}_\gamma)$ is a global-minimax solution of Problem (12).
Then we have

$$(\gamma - \gamma^*)p(\mathbf{x}_\gamma, \mathbf{y}_\gamma) \leq 2\zeta.$$

Based on the above inequality and choice of $\gamma \geq 3\gamma^*$, we get

$$p(\mathbf{x}_\gamma, \mathbf{y}_\gamma) \leq \frac{2\zeta}{\gamma - \gamma^*} \leq \frac{\zeta}{\gamma^*} \leq \varrho.$$

Let $\zeta_\gamma := p(\mathbf{x}_\gamma, \mathbf{y}_\gamma)$. Then, we get

$$\zeta_\gamma \leq \varrho, \tag{52}$$

and $(\mathbf{x}_\gamma, \boldsymbol{\epsilon}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\nu}_\gamma)$ is feasible for (49).
By (51), we have

$$\begin{aligned} & f(\mathbf{x}_\gamma + \boldsymbol{\epsilon}_\gamma, \mathbf{y}_\gamma + \boldsymbol{\nu}_\gamma) - f(\mathbf{x}^\circ + \boldsymbol{\epsilon}^\circ, \mathbf{y}^\circ + \boldsymbol{\nu}^\circ) \\ \leq\ & \gamma \left(p(\mathbf{x}^\circ, \mathbf{y}^\circ) - p(\mathbf{x}_\gamma, \mathbf{y}_\gamma)\right) + \zeta \\ \leq\ & \zeta, \end{aligned}$$

where the last inequality holds since $p(\mathbf{x}^\circ, \mathbf{y}^\circ) = 0 \leq p(\mathbf{x}_\gamma, \mathbf{y}_\gamma)$.
Since, for any $(\mathbf{x}, \mathbf{y}, \boldsymbol{\epsilon}, \boldsymbol{\nu})$ feasible for Problem (5), we have

$$\begin{aligned} & f(\mathbf{x}_\gamma + \boldsymbol{\epsilon}_\gamma, \mathbf{y}_\gamma + \boldsymbol{\nu}_\gamma) - \max_{\|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| \leq \eta} \{f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu})\} \\ \leq\ & f(\mathbf{x}_\gamma + \boldsymbol{\epsilon}_\gamma, \mathbf{y}_\gamma + \boldsymbol{\nu}_\gamma) - f(\mathbf{x}^\circ + \boldsymbol{\epsilon}^\circ, \mathbf{y}^\circ + \boldsymbol{\nu}^\circ), \end{aligned}$$

we get

$$\begin{aligned} & f(\mathbf{x}_\gamma + \boldsymbol{\epsilon}_\gamma, \mathbf{y}_\gamma + \boldsymbol{\nu}_\gamma) - \max_{\|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| \leq \eta} \{f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu})\} \\ \leq\ & \zeta, \end{aligned}$$

which implies the right hand side of (43). Next, we want to prove the left-hand side of (43). By (52), we have $p(\mathbf{x}_\gamma, \mathbf{y}_\gamma) \leq \varrho$. Then, we have

$$f(\mathbf{x}_\gamma + \boldsymbol{\epsilon}, \mathbf{y}_\gamma + \boldsymbol{\nu}) + \gamma p(\mathbf{x}_\gamma, \mathbf{y}_\gamma) \leq f(\mathbf{x}_\gamma + \boldsymbol{\epsilon}_\gamma, \mathbf{y}_\gamma + \boldsymbol{\nu}_\gamma) + \gamma p(\mathbf{x}_\gamma, \mathbf{y}_\gamma),$$

which implies that

$$f(\mathbf{x}_\gamma + \boldsymbol{\epsilon}, \mathbf{y}_\gamma + \boldsymbol{\nu}) - f(\mathbf{x}_\gamma + \boldsymbol{\epsilon}_\gamma, \mathbf{y}_\gamma + \boldsymbol{\nu}_\gamma) \leq 0.$$

We obtain that (44) implies (43). Together with the feasibility of $(\mathbf{x}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\epsilon}_\gamma, \boldsymbol{\nu}_\gamma)$ for Problem (49), this shows that $(\mathbf{x}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\epsilon}_\gamma, \boldsymbol{\nu}_\gamma)$ is a global solution of Problem (49). □

E.4 Proof of Theorem C.9

Proof. From Definition C.3, since $(\mathbf{x}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\epsilon}_\gamma, \boldsymbol{\nu}_\gamma)$ is a local solution of (12), there exist $r_0 > 0$ and a function h satisfying $h(r) \to 0$ as $r \to 0$, such that, for any $r \in (0, r_0]$ and any $(\mathbf{x}, \mathbf{y}, \boldsymbol{\epsilon}, \boldsymbol{\nu})$ satisfying $\|(\mathbf{x}, \mathbf{y}) - (\mathbf{x}_\gamma, \mathbf{y}_\gamma)\| \leq r$ and $\|(\boldsymbol{\epsilon}, \boldsymbol{\nu}) - (\boldsymbol{\epsilon}_\gamma, \boldsymbol{\nu}_\gamma)\| \leq r$, the following condition holds:

$$\mathcal{L}_\gamma(\mathbf{x}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\epsilon}, \boldsymbol{\nu}) \leq \mathcal{L}_\gamma(\mathbf{x}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\epsilon}_\gamma, \boldsymbol{\nu}_\gamma) \leq \max_{(\boldsymbol{\nu}, \boldsymbol{\epsilon}) \in \mathcal{U}} \mathcal{L}_\gamma(\mathbf{x}, \mathbf{y}, \boldsymbol{\epsilon}, \boldsymbol{\nu}), \tag{53}$$

for any $(\mathbf{x}, \mathbf{y})$, where $\mathcal{U} := \{(\boldsymbol{\nu}, \boldsymbol{\epsilon}) : \|\boldsymbol{\nu} - \boldsymbol{\nu}_\gamma\| \leq h(r), \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\gamma\| \leq h(r), \|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| \leq \eta\}$ and

$$\mathcal{L}_\gamma(\mathbf{x}, \mathbf{y}, \boldsymbol{\epsilon}, \boldsymbol{\nu}) = f(\mathbf{x} + \boldsymbol{\epsilon},\ \mathbf{y} + \boldsymbol{\nu}) + \gamma\ (g(\mathbf{x},\ \mathbf{y} + \boldsymbol{\nu}^*) - g^*(\mathbf{x}))\,.$$

Since $(\mathbf{x}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\epsilon}_\gamma, \boldsymbol{\nu}_\gamma)$ is a local solution of (12), $\mathbf{y}_\gamma$ is a local solution of (12) with $(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}) = (\mathbf{x}_\gamma, \boldsymbol{\epsilon}_\gamma, \boldsymbol{\nu}_\gamma)$. By applying the first-order stationary condition at any local minimax point as defined in Definition C.3-III, it follows that

$$\nabla_{\mathbf{y}} f(\mathbf{x}_\gamma + \boldsymbol{\epsilon}_\gamma, \mathbf{y}_\gamma + \boldsymbol{\nu}_\gamma) + \gamma \nabla_{\mathbf{y}} g(\mathbf{x}_\gamma, \mathbf{y}_\gamma + \boldsymbol{\nu}^*) = 0,$$

which, together with Assumption 2.3–I., imply

$$\|\nabla_{\mathbf{y}} g(\mathbf{x}_\gamma, \mathbf{y}_\gamma + \boldsymbol{\nu}^*)\| \leq \frac{\ell_{f,0}}{\gamma}. \tag{54}$$

Since $g(\mathbf{x}_\gamma, \cdot + \boldsymbol{\nu}^*)$ satisfies Assumption 2.1, from (7), we obtain that:

$$\begin{aligned} p(\mathbf{x}_\gamma, \mathbf{y}_\gamma) &= g(\mathbf{x}_\gamma, \mathbf{y}_\gamma + \boldsymbol{\nu}^*) - \min_{\mathbf{y} \in \mathbb{R}^{d_2}} g(\mathbf{x}_\gamma, \mathbf{y} + \boldsymbol{\nu}^*) \\ &\leq \mu \|\nabla_{\mathbf{y}} g(\mathbf{x}_\gamma, \mathbf{y}_\gamma + \boldsymbol{\nu}^*)\|^2. \end{aligned} \tag{55}$$

Equation (55) together with (54) and $\gamma \geq \ell_{f,0}\sqrt{\mu\varrho^{-1}}$, implies

$$p(\mathbf{x}_\gamma, \mathbf{y}_\gamma) \leq \frac{\ell_{f,0}^2 \mu}{\gamma^2} \leq \varrho. \tag{56}$$

Let $\zeta_\gamma := p(\mathbf{x}_\gamma, \mathbf{y}_\gamma)$, then $\zeta_\gamma \leq \varrho$ and $(\mathbf{x}_\gamma, \boldsymbol{\epsilon}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\nu}_\gamma)$ is feasible for

$$\begin{aligned} &\min_{\mathbf{x} \in \mathbb{R}^{d_1}, \mathbf{y} \in \mathbb{R}^{d_2}} \max_{(\boldsymbol{\nu}, \boldsymbol{\epsilon}) \in \mathcal{U}} f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu}) \\ &\text{subj. to} \quad p(\mathbf{x}, \mathbf{y}) \leq \zeta_\gamma. \end{aligned} \tag{57}$$

Since $(\mathbf{x}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\epsilon}_\gamma, \boldsymbol{\nu}_\gamma)$ is a local solution of (12) on $\mathcal{N}((\mathbf{x}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\epsilon}_\gamma, \boldsymbol{\nu}_\gamma), r)$, it holds for any $(\mathbf{x}, \mathbf{y}, \boldsymbol{\epsilon}, \boldsymbol{\nu}) \in \mathcal{N}((\mathbf{x}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\epsilon}_\gamma, \boldsymbol{\nu}_\gamma), r)$ that

$$\begin{aligned} &f(\mathbf{x}_\gamma + \boldsymbol{\epsilon}_\gamma, \mathbf{y}_\gamma + \boldsymbol{\nu}_\gamma) + \gamma p(\mathbf{x}_\gamma, \mathbf{y}_\gamma) \\ &\leq \max_{(\boldsymbol{\nu}, \boldsymbol{\epsilon}) \in \mathcal{U}} \{f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu}) + \gamma p(\mathbf{x}, \mathbf{y})\} \\ &\leq \max_{(\boldsymbol{\nu}, \boldsymbol{\epsilon}) \in \mathcal{U}} \{f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu})\} + \gamma p(\mathbf{x}, \mathbf{y}). \end{aligned}$$

From the above inequality, for any $(\mathbf{x}, \mathbf{y}, \boldsymbol{\epsilon}, \boldsymbol{\nu}) \in \mathcal{N}((\mathbf{x}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\epsilon}_\gamma, \boldsymbol{\nu}_\gamma), r)$ that is feasible for Problem (57), we have $\max_{(\boldsymbol{\nu}, \boldsymbol{\epsilon}) \in \mathcal{U}} p(\mathbf{x}, \mathbf{y}) \leq \zeta_\gamma$. Consequently,

$$\begin{aligned} f(\mathbf{x}_\gamma + \boldsymbol{\epsilon}_\gamma, \mathbf{y}_\gamma + \boldsymbol{\nu}_\gamma) &\leq \max_{(\boldsymbol{\nu}, \boldsymbol{\epsilon}) \in \mathcal{U}} \{f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu})\} + \gamma \left(\max_{(\boldsymbol{\nu}, \boldsymbol{\epsilon}) \in \mathcal{U}} p(\mathbf{x}, \mathbf{y}) - p(\mathbf{x}_\gamma, \mathbf{y}_\gamma) \right) \\ &= \max_{(\boldsymbol{\nu}, \boldsymbol{\epsilon}) \in \mathcal{U}} \{f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu})\} + \gamma \left(\max_{(\boldsymbol{\nu}, \boldsymbol{\epsilon}) \in \mathcal{U}} p(\mathbf{x}, \mathbf{y}) - \zeta_\gamma \right) \\ &\leq \max_{(\boldsymbol{\nu}, \boldsymbol{\epsilon}) \in \mathcal{U}} \{f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu})\}. \end{aligned} \tag{58}$$

We next show the left side of (53) implies that

$$f(\mathbf{x}_\gamma + \boldsymbol{\epsilon}, \mathbf{y}_\gamma + \boldsymbol{\nu}) \leq f(\mathbf{x}_\gamma + \boldsymbol{\epsilon}_\gamma, \mathbf{y}_\gamma + \boldsymbol{\nu}_\gamma).$$

By (56), we have $p(\mathbf{x}_\gamma, \mathbf{y}_\gamma) \leq \varrho$. Then, we have

$$f(\mathbf{x}_\gamma + \boldsymbol{\epsilon}, \mathbf{y}_\gamma + \boldsymbol{\nu}) + \gamma p(\mathbf{x}_\gamma, \mathbf{y}_\gamma) \leq f(\mathbf{x}_\gamma + \boldsymbol{\epsilon}_\gamma, \mathbf{y}_\gamma + \boldsymbol{\nu}_\gamma) + \gamma p(\mathbf{x}_\gamma, \mathbf{y}_\gamma),$$

which implies that

$$f(\mathbf{x}_\gamma + \boldsymbol{\epsilon}, \mathbf{y}_\gamma + \boldsymbol{\nu}) - f(\mathbf{x}_\gamma + \boldsymbol{\epsilon}_\gamma, \mathbf{y}_\gamma + \boldsymbol{\nu}_\gamma) \leq \gamma \left(p(\mathbf{x}_\gamma, \mathbf{y}_\gamma) - p(\mathbf{x}_\gamma, \mathbf{y}_\gamma) \right) \leq 0. \tag{59}$$

By (58) and (59), together with the feasibility of $(\mathbf{x}_\gamma, \boldsymbol{\epsilon}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\nu}_\gamma)$ for (57), it follows that $(\mathbf{x}_\gamma, \boldsymbol{\epsilon}_\gamma, \mathbf{y}_\gamma, \boldsymbol{\nu}_\gamma)$ is a local solution of (57). □

E.5 Proof of Corollary C.10

Proof. Consider the following special case of (2):

$$\begin{aligned} &\min_{y \in \mathbb{R}} \max_{\|\epsilon\| \leq \rho,\ \|\nu\| \leq \eta} \{f(x + \epsilon,\ y + \nu) = y + \nu\} \\ &\quad \text{subj. to } y \in \arg\min_{y \in \mathbb{R}} \max_{\|\nu\| \leq \eta} \{g(x,\ y + \nu) = y^2 + \nu^2\}. \end{aligned} \tag{60}$$

Note that $\max_{\|\nu\| \leq \eta} \{y^2 + \nu^2\} = y^2 + \eta^2$, and hence

$$\arg\min_{y \in \mathbb{R}} \max_{\|\nu\| \leq \eta} (y^2 + \nu^2) = \arg\min_{y \in \mathbb{R}} (y^2 + \eta^2) = \{0\}.$$

Therefore, the feasible set of the constrained problem is $\{0\}$, and then $y = 0$ is the unique solution of (60).
For the inner maximization, we may choose $\nu^*(x, y) = \eta$. Thus, since $\eta \in \arg\max_{\|\nu\| \le \eta}\{y^2+\nu^2\}$, we get

$$g^*(x) = \min_y\{g(\mathbf{x},\, y + \nu^*(x, y))\} = \min_y\{y^2 + \eta^2\} = \eta^2,$$

which implies

$$p(x, y) = g(x,\, y + \nu^*(x, y)) - g^*(x) = (y^2 + \eta^2) - \eta^2 = y^2.$$

And Problem (12) is

$$\min_{y\in\mathbb{R}} \max_{\|\epsilon\|\le\rho, \|\nu\|\le\eta}\{\nu + y + \gamma y^2\} = \min_{y\in\mathbb{R}}\{\eta + y + \gamma y^2\},$$

with solution $-1/2\gamma$. Thus $-1/2\gamma$ is a solution of Problem (29) with $\zeta_\gamma = 1/(4\gamma^2)$. In this example, achieving $\zeta_\gamma = \mathcal{O}(\varrho)$, requires that γ grows at least as fast as $\varrho^{-0.5}$, i.e., $\gamma = \Omega(\varrho^{-0.5})$. Finally, the proof is concluded by noting that the conditions stated in Theorems C.8 and C.9 are satisfied in this example. □

F Proofs omitted from Section 2.1.2

F.1 Proof of Lemma C.1

Proof. **To establish (26a)**: As the exact solution to the inner maximization in (12) is hard to compute, we approximate it with the following first-order Taylor expansion

$$\begin{aligned}
\boldsymbol{\nu}^*(\mathbf{x}, \mathbf{y}) &= \arg\max_{\|\boldsymbol{\nu}\|\le\eta} \{\mathcal{L}_\gamma(\mathbf{x}, \mathbf{y}, \boldsymbol{\epsilon}, \boldsymbol{\nu})\} \\
&= \arg\max_{\|\boldsymbol{\nu}\|\le\eta} \{f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu}) + \gamma\, p(\mathbf{x}, \mathbf{y})\} \\
&= \arg\max_{\|\boldsymbol{\nu}\|\le\eta} \{f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu}) + \gamma\, (g(\mathbf{x},\, \mathbf{y} + \boldsymbol{\nu}^*) - g^*(\mathbf{x}))\} \\
&\approx \arg\max_{\|\boldsymbol{\nu}\|\le\eta} \left\{f(\mathbf{x}, \mathbf{y}) + \boldsymbol{\epsilon}^\top \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) + \boldsymbol{\nu}^\top \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})\right\} + \mathcal{O}(\eta^2) \\
&\approx \arg\max_{\|\boldsymbol{\nu}\|\le\eta} \left\{\boldsymbol{\nu}^\top \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})\right\}.
\end{aligned}$$

Let $B := \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$. To maximize $\boldsymbol{\nu}^\top B$, $\boldsymbol{\nu}$ should be in the same direction as B. By the Cauchy–Schwarz inequality, the dot product $\boldsymbol{\nu}^\top B$ is maximized when $\boldsymbol{\nu}$ is a scalar multiple of B. Let $\boldsymbol{\nu} = bB$ where b is a scalar. Then the constraint $\|\boldsymbol{\nu}\| \le \eta$ becomes:

$$\|bB\| \le \eta, \quad \Rightarrow |b|\|B\| \le \eta, \quad \Rightarrow |b| \le \frac{\eta}{\|B\|}.$$

Since $|b|$ must be as large as possible but less than $\frac{\eta}{\|B\|}$, we take $b = \frac{\eta}{\|B\|}$. Hence, we have

$$\hat{\boldsymbol{\nu}}(\mathbf{x}, \mathbf{y}) = \frac{\eta B}{\|B\|} = \frac{\eta \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})}{\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})\|}.$$

To establish (26b): As the exact solution to the inner maximization in (12) is hard to compute, we approximate it with the following first-order Taylor expansion

$$\begin{aligned}
\boldsymbol{\epsilon}^*(\mathbf{x}, \mathbf{y}) &= \arg\max_{\|\boldsymbol{\epsilon}\|\le\rho} \{\mathcal{L}_\gamma(\mathbf{x}, \mathbf{y}, \boldsymbol{\epsilon}, \boldsymbol{\nu})\} \\
&= \arg\max_{\|\boldsymbol{\epsilon}\|\le\rho} \{f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu}) + \gamma\, p(\mathbf{x}, \mathbf{y})\} \\
&= \arg\max_{\|\boldsymbol{\epsilon}\|\le\rho} \{f(\mathbf{x} + \boldsymbol{\epsilon}, \mathbf{y} + \boldsymbol{\nu}) + \gamma\, (g(\mathbf{x},\, \mathbf{y} + \boldsymbol{\nu}^*) - g^*(\mathbf{x}))\}.
\end{aligned}$$

Recall that $\mathbf{y}^* \in \mathcal{S}(\mathbf{x})$,

$$g^*(\mathbf{x}) = \min_{\mathbf{y}\in\mathbb{R}^{d_2}} g(\mathbf{x},\, \mathbf{y} + \boldsymbol{\nu}^*) = g(\mathbf{x},\, \mathbf{y}^* + \boldsymbol{\nu}^*),$$

and $\boldsymbol{\nu}^*(\mathbf{x},\mathbf{y}) = \arg\max_{\|\boldsymbol{\nu}\|\le\eta}\, g(\mathbf{x},\, \mathbf{y}+\boldsymbol{\nu})$. Then, we have

$$\begin{aligned}\boldsymbol{\epsilon}^*(\mathbf{x},\mathbf{y}) &\approx \arg\max_{\|\boldsymbol{\epsilon}\|\le\rho} \big\{ f(\mathbf{x},\mathbf{y}) + \boldsymbol{\epsilon}^\top \nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}) + \boldsymbol{\nu}^\top \nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}) \\ &\qquad\qquad + \gamma g(\mathbf{x},\mathbf{y}) + \gamma \boldsymbol{\nu}^{*\top} \nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y}) \\ &\qquad\qquad - \gamma g(\mathbf{x},\mathbf{y}^*) - \gamma \boldsymbol{\nu}^{*\top} \nabla_{\mathbf{y}} g(\mathbf{x},\mathbf{y}^*) \big\} + \mathcal{O}(\rho^2) \\ &\approx \arg\max_{\|\boldsymbol{\epsilon}\|\le\rho} \big\{ \boldsymbol{\epsilon}^\top \nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}) \big\}.\end{aligned}$$

To maximize $\boldsymbol{\epsilon}^\top \nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})$, $\boldsymbol{\epsilon}$ should be in the same direction as $\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})$. By the Cauchy–Schwarz inequality, the dot product $\boldsymbol{\epsilon}^\top \nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})$ is maximized when $\boldsymbol{\epsilon}$ is a scalar multiple of $\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})$. Let $\boldsymbol{\epsilon} = a \nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})$ where a is a scalar. Then the constraint $\|\boldsymbol{\epsilon}\| \le \rho$ becomes:

$$\|a \nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\| \le \rho, \quad \Rightarrow |a| \|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\| \le \rho, \quad \Rightarrow |a| \le \frac{\rho}{\|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\|}.$$

Since $|a|$ must be as large as possible but less than $\frac{\rho}{\|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\|}$, we take $a = \frac{\rho}{\|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\|}$. Hence, we have

$$\hat{\boldsymbol{\epsilon}}(\mathbf{x},\mathbf{y}) = \frac{\rho \nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})}{\|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\|}.$$

□

G Proofs omitted from Section 3

We now prove Theorems 3.1 and 3.3. We first establish several auxiliary lemmas, which will subsequently be used to prove Theorem 3.1.

Lemma G.1. *Suppose Assumption 2.2 holds and $\beta \le 1/\ell_{g,1}$. Then, the iterates $\{\mathbf{x}_k\}$ produced by Algorithm 1 satisfy:*

$$-\beta \langle \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)} + \tilde{\boldsymbol{\nu}}_t^{(k)}), \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)}) \rangle \le \frac{\eta^2 \ell_{g,1}}{2} - \frac{\beta}{2} \|\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2.$$

Proof. We first add and subtract terms to obtain:

$$\begin{aligned}&-\beta \langle \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)} + \tilde{\boldsymbol{\nu}}_t^{(k)}), \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)}) \rangle \\ &= -\beta \langle \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)} + \tilde{\boldsymbol{\nu}}_t^{(k)}) - \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)}), \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)}) \rangle \\ &\quad - \beta \langle \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)}), \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)}) \rangle. \end{aligned} \tag{61}$$

Recall that from (10a), we have

$$\tilde{\boldsymbol{\nu}}_t^{(k)} = \frac{\eta \, \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})}{\|\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|}.$$

For the first term on the right-hand side of (61), we have:

$$\begin{aligned}&\langle \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)} + \tilde{\boldsymbol{\nu}}_t^{(k)}) - \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)}), \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)}) \rangle \\ &= \frac{\|\nabla_{\mathbf{z}}\, g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|}{\eta} \langle \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)} + \tilde{\boldsymbol{\nu}}_t^{(k)}) - \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)}), \tilde{\boldsymbol{\nu}}_t^{(k)} \rangle \\ &\ge -\ell_{g,1} \frac{\|\nabla_{\mathbf{z}}\, g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|}{\eta} \|\tilde{\boldsymbol{\nu}}_t^{(k)}\|^2,\end{aligned}$$

where the inequality holds since from Assumption 2.2, we have

$$\langle \nabla L(w_1) - \nabla L(w_2), w_1 - w_2 \rangle \ge -\ell_{g,1} \|w_1 - w_2\|^2,$$

which implies

$$
\begin{aligned}
&-\beta\langle\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)} + \tilde{\boldsymbol{\nu}}_t^{(k)}) - \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)}), \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\rangle \\
&\leq \beta\ell_{g,1}\frac{\|\nabla_{\mathbf{z}}\, g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|}{\eta}\|\tilde{\boldsymbol{\nu}}_t^{(k)}\|^2.
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
(61) &\leq \beta\ell_{g,1}\frac{\|\nabla_{\mathbf{z}}\, g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|}{\eta}\|\tilde{\boldsymbol{\nu}}_t^{(k)}\|^2 - \beta\|\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 \\
&= \beta\ell_{g,1}\eta\|\nabla_{\mathbf{z}}\, g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\| - \beta\|\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 \\
&\leq \frac{\eta^2\ell_{g,1}}{2} + \frac{\beta^2\ell_{g,1}}{2}\|\nabla_{\mathbf{z}}\, g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 - \beta\|\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2,
\end{aligned}
$$

where the equality follows from $\|\tilde{\boldsymbol{\nu}}_t^{(k)}\| = \eta$, and the last inequality follows from Young's inequality: $2\langle a, b\rangle \leq v\|a\|^2 + \frac{1}{v}\|b\|^2$ for scalar $v > 0$.

Finally, from $\beta \leq 1/\ell_{g,1}$, we get the desired result. □

Lemma G.2. *Suppose Assumptions 2.1, 2.3–I. and 2.2 hold. Let $\beta \in (0, 1/\ell_{g,1}]$. For the sequence $\tilde{\mathbf{y}}_k$ generated by Algorithm 1, we have*

$$
\begin{aligned}
d^2_{\mathcal{S}(\mathbf{x}_k)}(\tilde{\mathbf{y}}_k) \leq \left(1 - \frac{\beta}{2\mu}\right)^{T_k}\left(2\mu^2\ell^2_{g,1}\eta^2 + \frac{4\mu^2}{\gamma^2\alpha^2}\|\mathbf{y}_k - \mathbf{y}_{k+1}\|^2 + \frac{4\mu^2\ell^2_{f,0}}{\gamma^2}\right) \\
+ \frac{2\mu^2\eta^2\ell_{g,1}}{\beta},
\end{aligned}
$$

where $\mathcal{S}(\mathbf{x}_k) = \arg\min_{\mathbf{y}\in\mathbb{R}^{d_2}} g(\mathbf{x}_k, \mathbf{y})$ and $d^2_{\mathcal{S}}(\mathbf{y}) := \min_{\mathbf{y}'\in\mathcal{S}}\|\mathbf{y} - \mathbf{y}'\|^2$.

Proof. From Assumption 2.2, we have

$$
\begin{aligned}
g(\mathbf{x}_k, \mathbf{z}_{t+1}^{(k)}) &\leq g(\mathbf{x}_k, \mathbf{z}_t^{(k)}) + \langle\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)}), \mathbf{z}_{t+1}^{(k)} - \mathbf{z}_t^{(k)}\rangle + \frac{\ell_{g,1}}{2}\|\mathbf{z}_{t+1}^{(k)} - \mathbf{z}_t^{(k)}\|^2 \\
&= g(\mathbf{x}_k, \mathbf{z}_t^{(k)}) - \beta\langle\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)}), \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)} + \tilde{\boldsymbol{\nu}}_t^{(k)})\rangle \\
&+ \frac{\ell_{g,1}\beta^2}{2}\|\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)} + \tilde{\boldsymbol{\nu}}_t^{(k)})\|^2,
\end{aligned}
$$

where the equality follows from (10b).
From $\|a\|^2 = 2\langle a, b\rangle - \|b\|^2 + \|a - b\|^2$, we have:

$$
\begin{aligned}
&\|\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)} + \tilde{\boldsymbol{\nu}}_t^{(k)})\|^2 \\
&= 2\langle\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)} + \tilde{\boldsymbol{\nu}}_t^{(k)}), \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\rangle - \|\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 \\
&+ \|\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)} + \tilde{\boldsymbol{\nu}}_t^{(k)}) - \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2.
\end{aligned}
$$

Thus, we get

$$
\begin{aligned}
g(\mathbf{x}_k, \mathbf{z}_{t+1}^{(k)}) \leq g(\mathbf{x}_k, \mathbf{z}_t^{(k)}) - \beta(1 - \ell_{g,1}\beta)\langle\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)}), \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)} + \tilde{\boldsymbol{\nu}}_t^{(k)})\rangle \\
- \frac{\ell_{g,1}\beta^2}{2}\|\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 \\
+ \frac{\ell_{g,1}\beta^2}{2}\|\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)} + \tilde{\boldsymbol{\nu}}_t^{(k)}) - \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2.
\end{aligned}
$$

From Lemma G.1, we have

$$
-\beta\langle\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)} + \tilde{\boldsymbol{\nu}}_t^{(k)}), \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\rangle \leq \frac{\eta^2\ell_{g,1}}{2} - \frac{\beta}{2}\|\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2.
$$

Thus, from the above inequalities, we obtain

$$\begin{aligned}
g(\mathbf{x}_k, \mathbf{z}_{t+1}^{(k)}) &\le g(\mathbf{x}_k, \mathbf{z}_t^{(k)}) + (1-\ell_{g,1}\beta)\frac{\eta^2\ell_{g,1}}{2} \\
&\quad - (1-\ell_{g,1}\beta)\frac{\beta}{2}\|\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 - \frac{\ell_{g,1}\beta^2}{2}\|\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 \\
&\quad + \frac{\ell_{g,1}\beta^2}{2}\|\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)} + \tilde{\boldsymbol{\nu}}_t^{(k)}) - \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 \\
&= g(\mathbf{x}_k, \mathbf{z}_t^{(k)}) + (1-\ell_{g,1}\beta)\frac{\eta^2\ell_{g,1}}{2} - \frac{\beta}{2}\|\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 \\
&\quad + \frac{\ell_{g,1}\beta^2}{2}\|\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)} + \tilde{\boldsymbol{\nu}}_t^{(k)}) - \nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 \\
&\le g(\mathbf{x}_k, \mathbf{z}_t^{(k)}) + \frac{\eta^2\ell_{g,1}}{2} - \frac{\beta}{2}\|\nabla_{\mathbf{z}} g(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 + \frac{\ell_{g,1}^3\beta^2}{2}\|\tilde{\boldsymbol{\nu}}_t^{(k)}\|^2,
\end{aligned}$$

where the last inequality follows from $1-\ell_{g,1}\beta \le 1$ and Assumption 2.2.
Let $\tilde{\mathbf{y}}_k^* \in \arg\min_{\mathbf{y}\in\mathbb{R}^{d_2}} g(\mathbf{x}_k, \mathbf{y})$. From Assumption 2.1 and $\|\tilde{\boldsymbol{\nu}}_t^{(k)}\| = \eta$, we also have

$$\begin{aligned}
g(\mathbf{x}_k, \mathbf{z}_{t+1}^{(k)}) - g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*) &\le g(\mathbf{x}_k, \mathbf{z}_t^{(k)}) - g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*) + \frac{\eta^2\ell_{g,1}}{2} \\
&\quad - \frac{\beta}{2\mu}\left(g(\mathbf{x}_k, \mathbf{z}_t^{(k)}) - g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*)\right) + \frac{\ell_{g,1}^3\beta^2}{2}\eta^2 \\
&\le \left(1-\frac{\beta}{2\mu}\right)\left(g(\mathbf{x}_k, \mathbf{z}_t^{(k)}) - g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*)\right) + \eta^2\ell_{g,1},
\end{aligned}$$

where the second inequality follows from $\beta \le 1/\ell_{g,1}$.
Iteratively applying the above inequality for $t = 1, \ldots, T_k$ yields

$$\begin{aligned}
&g(\mathbf{x}_k, \mathbf{z}_{T_k+1}^{(k)}) - g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*) \\
&\le \left(1-\frac{\beta}{2\mu}\right)^{T_k}\left(g(\mathbf{x}_k, \mathbf{z}_1^{(k)}) - g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*)\right) + \eta^2\ell_{g,1}\sum_{j=1}^{T_k}\left(1-\frac{\beta}{2\mu}\right)^j \\
&\le \left(1-\frac{\beta}{2\mu}\right)^{T_k}\left(g(\mathbf{x}_k, \mathbf{z}_1^{(k)}) - g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*)\right) + \frac{2\mu\eta^2\ell_{g,1}}{\beta},
\end{aligned} \tag{62}$$

where the last inequality follows from $\sum_{j=0}^{k-1}(1-\gamma)^j \le 1/\gamma$.

From Lemma B.1, the $1/\mu$-PL condition of $g(\mathbf{x}, \cdot)$ also implies the error bound of $g(\mathbf{x}, \cdot)$, which leads to

$$\frac{1}{\mu} d_{\mathcal{S}(\mathbf{x}_k)}^2(\tilde{\mathbf{y}}_k) \le g(\mathbf{x}_k, \mathbf{z}_{T_k+1}^{(k)}) - g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*), \tag{63}$$

where $\tilde{\mathbf{y}}_k = \mathbf{z}_{T_k+1}^{(k)}$ and $\mathcal{S}(\mathbf{x}_k) = \arg\min_{\mathbf{y}\in\mathbb{R}^{d_2}} g(\mathbf{x}_k, \mathbf{y})$. From (62) and (63), and using $\mathbf{z}_1^{(k)} = \mathbf{y}_k$, we get

$$d_{\mathcal{S}(\mathbf{x}_k)}^2(\tilde{\mathbf{y}}_k) \le \mu\left(1-\frac{\beta}{2\mu}\right)^{T_k}\underbrace{\left(g(\mathbf{x}_k, \mathbf{y}_k) - g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*)\right)}_{:=I} + \frac{2\mu^2\eta^2\ell_{g,1}}{\beta}. \tag{64}$$

For term I, by Assumption 2.1, we have

$$\begin{aligned}
I &\le \mu\|\nabla_{\mathbf{y}} g(\mathbf{x}_k, \mathbf{y}_k)\|^2 \\
&\le 2\mu\|\nabla_{\mathbf{y}} g(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{y}} g(\mathbf{x}_k, \mathbf{y}_k + \tilde{\boldsymbol{\nu}}_k)\|^2 + 2\mu\|\nabla_{\mathbf{y}} g(\mathbf{x}_k, \mathbf{y}_k + \tilde{\boldsymbol{\nu}}_k)\|^2 \\
&\le 2\mu\ell_{g,1}^2\|\tilde{\boldsymbol{\nu}}_k\|^2 + 2\mu\|\nabla_{\mathbf{y}} g(\mathbf{x}_k, \mathbf{y}_k + \tilde{\boldsymbol{\nu}}_k)\|^2 \\
&= 2\mu\ell_{g,1}^2\eta^2 + 2\mu\|\nabla_{\mathbf{y}} g(\mathbf{x}_k, \mathbf{y}_k + \tilde{\boldsymbol{\nu}}_k)\|^2,
\end{aligned} \tag{65}$$

where the second inequality follows from $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ and the third inequality follows from Assumption 2.2.

From (15c), we have

$$2\mu\|\nabla_{\mathbf{y}} g(\mathbf{x}_k, \mathbf{y}_k + \tilde{\boldsymbol{\nu}}_k)\|^2 = 2\mu \left\| \frac{\mathbf{y}_k - \mathbf{y}_{k+1} - \alpha\nabla_{\mathbf{y}} f(\mathbf{x}_k + \hat{\boldsymbol{\epsilon}}_k, \mathbf{y}_k + \hat{\boldsymbol{\nu}}_k)}{\alpha\gamma} \right\|^2 \leq \frac{4\mu}{\gamma^2\alpha^2}\|\mathbf{y}_k - \mathbf{y}_{k+1}\|^2 + \frac{4\mu\ell_{f,0}^2}{\gamma^2}, \tag{66}$$

where the inequality is by Assumption 2.3–I..
Substituting (66) into (65) yields

$$I \leq 2\mu\ell_{g,1}^2\eta^2 + \frac{4\mu}{\gamma^2\alpha^2}\|\mathbf{y}_k - \mathbf{y}_{k+1}\|^2 + \frac{4\mu\ell_{f,0}^2}{\gamma^2}. \tag{67}$$

Combining Equations (64) and (67), we obtain

$$d^2_{\mathcal{S}(\mathbf{x}_k)}(\tilde{\mathbf{y}}_k) \leq \left(1 - \frac{\beta}{2\mu}\right)^{T_k} \left(2\mu^2\ell_{g,1}^2\eta^2 + \frac{4\mu^2}{\gamma^2\alpha^2}\|\mathbf{y}_k - \mathbf{y}_{k+1}\|^2 + \frac{4\mu^2\ell_{f,0}^2}{\gamma^2}\right) + \frac{2\mu^2\eta^2\ell_{g,1}}{\beta}.$$

□

We next present the proof of Theorem 3.1.

G.1 Proof of Theorem 3.1

Proof. From (15), we have

$$\widehat{\nabla}\mathcal{L}_{\gamma,k} := \left(\widehat{\nabla}_{\mathbf{x}}\mathcal{L}_{\gamma,k}, \widehat{\nabla}_{\mathbf{y}}\mathcal{L}_{\gamma,k}\right),$$

where

$$\begin{aligned}\widehat{\nabla}_{\mathbf{x}}\mathcal{L}_{\gamma,k} &= \nabla_{\mathbf{x}} f(\mathbf{x}_k + \hat{\boldsymbol{\epsilon}}_k,\, \mathbf{y}_k + \hat{\boldsymbol{\nu}}_k) \\ &\quad + \gamma\left(\nabla_{\mathbf{x}} g(\mathbf{x}_k,\, \mathbf{y}_k + \tilde{\boldsymbol{\nu}}_k) - \nabla_{\mathbf{x}} g(\mathbf{x}_k,\, \tilde{\mathbf{y}}_k + \tilde{\boldsymbol{\nu}}_k)\right), \\ \widehat{\nabla}_{\mathbf{y}}\mathcal{L}_{\gamma,k} &= \nabla_{\mathbf{y}} f(\mathbf{x}_k + \hat{\boldsymbol{\epsilon}}_k,\, \mathbf{y}_k + \hat{\boldsymbol{\nu}}_k) + \gamma\nabla_{\mathbf{y}} g(\mathbf{x}_k,\, \mathbf{y}_k + \tilde{\boldsymbol{\nu}}_k),\end{aligned}$$

with $\bar{\nabla}_{\mathbf{x}} g(\mathbf{x}_k, \mathbf{y} + \boldsymbol{\nu}) = (\nabla_{\mathbf{x}} g(\mathbf{x}_k, \mathbf{y} + \boldsymbol{\nu}), \mathbf{0}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$. This, implies

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha\widehat{\nabla}\mathcal{L}_{\gamma,k},$$

where $\mathbf{w} := (\mathbf{x}, \mathbf{y})$. Let

$$\mathcal{L}_\gamma(\mathbf{w}_k) := f(\mathbf{w}_k) + \gamma\left(g(\mathbf{w}_k) - g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*)\right), \quad \text{where} \quad \tilde{\mathbf{y}}_k^* \in S(\mathbf{x}_k) := \arg\min_{\mathbf{y}\in\mathbb{R}^{d_2}} g(\mathbf{x}_k, \mathbf{y}).$$

By Assumption 2.2, the function g is $\ell_{g,1}$-smooth. By Lemma A.5 of (Nouiehed et al., 2019), under Assumption 2.1, $g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*)$ is $(\ell_{g,1} + \ell_{g,1}^2\mu)$-smooth. Thus, based on Assumption 2.3–I., $\mathcal{L}_\gamma(\mathbf{w}_k)$ is ℓ_γ-Lipschitz-smooth, where $\ell_\gamma = \ell_{f,1} + \gamma(\ell_{g,1} + \ell_{g,1}^2\mu)$. Then, we have

$$\nabla\mathcal{L}_\gamma(\mathbf{w}_k) = \nabla f(\mathbf{w}_k) + \gamma\left(\nabla g(\mathbf{w}_k) - \nabla g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*)\right).$$

Then by the Lipschitz-smoothness of $\mathcal{L}_\gamma$, it holds that

$$\begin{aligned}\mathcal{L}_\gamma(\mathbf{w}_{k+1}) &\leq \mathcal{L}_\gamma(\mathbf{w}_k) + \langle\nabla\mathcal{L}_\gamma(\mathbf{w}_k), \mathbf{w}_{k+1} - \mathbf{w}_k\rangle + \frac{\ell_\gamma}{2}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 \\ &\leq \mathcal{L}_\gamma(\mathbf{w}_k) + \underbrace{\langle\widehat{\nabla}\mathcal{L}_{\gamma,k}, \mathbf{w}_{k+1} - \mathbf{w}_k\rangle}_{:=A} \\ &\quad + \frac{1}{2\alpha}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 + \underbrace{\langle\nabla\mathcal{L}_\gamma(\mathbf{w}_k) - \widehat{\nabla}\mathcal{L}_{\gamma,k}, \mathbf{w}_{k+1} - \mathbf{w}_k\rangle}_{:=B},\end{aligned} \tag{68}$$

where the second inequality is by $\alpha \leq 1/\ell_\gamma$.

Now, let us derive upper bounds for the terms on the right-hand side of (68).
For the term A on the RHS of (68), we obtain:

$$A = \langle \widehat{\nabla}\mathcal{L}_{\gamma,k}, \mathbf{w}_{k+1} - \mathbf{w}_k \rangle = -\frac{1}{\alpha}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2. \tag{69}$$

For the last term on the RHS of (68), we get

$$\begin{aligned} B &= \langle \nabla\mathcal{L}_\gamma(\mathbf{w}_k) - \widehat{\nabla}\mathcal{L}_{\gamma,k}, \mathbf{w}_{k+1} - \mathbf{w}_k \rangle \\ &\leq \alpha \underbrace{\|\nabla\mathcal{L}_\gamma(\mathbf{w}_k) - \widehat{\nabla}\mathcal{L}_{\gamma,k}\|^2}_{:=I} + \frac{1}{4\alpha}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2, \end{aligned} \tag{70}$$

where the inequality follows from Young's inequality.
The first term in the above inequality can be bounded as

$$\begin{aligned} I = &\|\nabla f(\mathbf{x}_k, \mathbf{y}_k) + \gamma\left(\nabla g(\mathbf{x}_k, \mathbf{y}_k) - \nabla g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*)\right) \\ &- \left(\nabla f(\mathbf{x}_k + \hat{\boldsymbol{\epsilon}}_k, \mathbf{y}_k + \hat{\boldsymbol{\nu}}_k) + \gamma\left(\nabla g(\mathbf{x}_k, \mathbf{y}_k + \tilde{\boldsymbol{\nu}}_k) - \bar{\nabla}_{\mathbf{x}} g\left(\mathbf{x}_k, \tilde{\mathbf{y}}_k + \tilde{\boldsymbol{\nu}}_k\right)\right)\right)\|^2 \\ &\leq 3\|\nabla f(\mathbf{x}_k, \mathbf{y}_k) - \nabla f(\mathbf{x}_k + \hat{\boldsymbol{\epsilon}}_k, \mathbf{y}_k + \hat{\boldsymbol{\nu}}_k)\|^2 \\ &+ 3\gamma^2\|\nabla g(\mathbf{x}_k, \mathbf{y}_k) - \nabla g(\mathbf{x}_k, \mathbf{y}_k + \tilde{\boldsymbol{\nu}}_k)\|^2 \\ &+ 3\gamma^2\|\nabla g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*) - \bar{\nabla}_{\mathbf{x}} g\left(\mathbf{x}_k, \tilde{\mathbf{y}}_k + \tilde{\boldsymbol{\nu}}_k\right)\|^2. \end{aligned} \tag{71}$$

For each term on the RHS of (71), using Assumptions 2.3–II. and 2.2, we obtain:

$$\begin{aligned} \|\nabla f(\mathbf{x}_k, \mathbf{y}_k) - \nabla f(\mathbf{x}_k + \hat{\boldsymbol{\epsilon}}_k, \mathbf{y}_k + \hat{\boldsymbol{\nu}}_k)\|^2 &\leq \ell_{f,1}^2\|(\hat{\boldsymbol{\epsilon}}_k, \hat{\boldsymbol{\nu}}_k)\|^2 \\ &= \ell_{f,1}^2(\rho^2 + \eta^2), \end{aligned} \tag{72a}$$

where the equality holds because the perturbations $\|\hat{\boldsymbol{\epsilon}}_k\| = \rho$ and $\|\hat{\boldsymbol{\nu}}_k\| = \eta$.
Moreover

$$\begin{aligned} \|\nabla g(\mathbf{x}_k, \mathbf{y}_k) - \nabla g(\mathbf{x}_k, \mathbf{y}_k + \tilde{\boldsymbol{\nu}}_k)\|^2 &\leq \ell_{g,1}^2\|\tilde{\boldsymbol{\nu}}_k\|^2 \\ &\leq \ell_{g,1}^2\eta^2, \end{aligned} \tag{72b}$$

and

$$\begin{aligned} &\|\nabla g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*) - \bar{\nabla}_{\mathbf{x}} g\left(\mathbf{x}_k, \tilde{\mathbf{y}}_k + \tilde{\boldsymbol{\nu}}_k\right)\|^2 \\ &\leq 2\|\nabla g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*) - \nabla g(\mathbf{x}_k, \tilde{\mathbf{y}}_k)\|^2 \\ &+ 2\|\nabla g(\mathbf{x}_k, \tilde{\mathbf{y}}_k) - \bar{\nabla}_{\mathbf{x}} g\left(\mathbf{x}_k, \tilde{\mathbf{y}}_k + \tilde{\boldsymbol{\nu}}_k\right)\|^2 \\ &\leq 2\ell_{g,1}^2\|\tilde{\mathbf{y}}_k^* - \tilde{\mathbf{y}}_k\|^2 + 2\ell_{g,1}^2\|\tilde{\boldsymbol{\nu}}_k\|^2 \\ &\leq 2\ell_{g,1}^2\|\tilde{\mathbf{y}}_k^* - \tilde{\mathbf{y}}_k\|^2 + 2\ell_{g,1}^2\eta^2. \end{aligned} \tag{72c}$$

From (72a)–(72c), we obtain:

$$I \leq 3\ell_{f,1}^2\left(\rho^2 + \eta^2\right) + 9\gamma^2\ell_{g,1}^2\eta^2 + 6\gamma^2\ell_{g,1}^2\|\tilde{\mathbf{y}}_k^* - \tilde{\mathbf{y}}_k\|^2. \tag{73}$$

Substituting (73) in (70), we get

$$\begin{aligned} B \leq &3\alpha\ell_{f,1}^2\left(\rho^2 + \eta^2\right) + 9\alpha\gamma^2\ell_{g,1}^2\eta^2 \\ &+ 6\alpha\gamma^2\ell_{g,1}^2\|\tilde{\mathbf{y}}_k^* - \tilde{\mathbf{y}}_k\|^2 + \frac{1}{4\alpha}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2. \end{aligned} \tag{74}$$

Combining (74), (69) and (68) together, we have

$$\begin{aligned} \mathcal{L}_\gamma(\mathbf{w}_{k+1}) \leq &\mathcal{L}_\gamma(\mathbf{w}_k) + 3\alpha\ell_{f,1}^2\left(\rho^2 + \eta^2\right) + 9\alpha\gamma^2\ell_{g,1}^2\eta^2 \\ &+ 6\alpha\gamma^2\ell_{g,1}^2\|\tilde{\mathbf{y}}_k^* - \tilde{\mathbf{y}}_k\|^2 - \frac{1}{4\alpha}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2. \end{aligned}$$

From Lemma G.2, we have

$$\begin{aligned}\|\tilde{\mathbf{y}}_k^* - \tilde{\mathbf{y}}_k\|^2 &\leq \left(1 - \frac{\beta}{2\mu}\right)^{T_k} \left(2\mu^2 \ell_{g,1}^2 \eta^2 + \frac{4\mu^2}{\gamma^2\alpha^2}\|\mathbf{y}_k - \mathbf{y}_{k+1}\|^2 + \frac{4\mu^2 \ell_{f,0}^2}{\gamma^2}\right) \\ &\quad + \frac{2\mu^2\eta^2\ell_{g,1}}{\beta} \\ &\leq \eta^2 + \frac{\|\mathbf{w}_k - \mathbf{w}_{k+1}\|^2}{72\gamma^2\ell_{g,1}^2\alpha^2} + \frac{\mu^2\ell_{f,0}^2}{\alpha^2 k^2 \gamma^2} + \frac{2\mu^2\eta^2\ell_{g,1}}{\beta}, \end{aligned} \tag{75}$$

where the second inequality follows from

$$T_k \geq \max\{-\log_{c_\beta}(2\mu^2\ell_{g,1}^2), -\log_{c_\beta}(288\mu^2\ell_{g,1}^2), -2\log_{c_\beta}(2\alpha k)\}, \quad \text{with} \quad c_\beta = 1 - \frac{\beta}{2\mu}.$$

Since $c_\beta < 1$, we have $-\log_{c_\beta}(x) = \log_{c_\beta}(x^{-1})$ and $\log_{c_\beta}$ is decreasing; therefore

$$\max\{\log_{c_\beta}(a), \log_{c_\beta}(b)\} = \log_{c_\beta}(\min\{a, b\}).$$

Hence the bound is equivalent to

$$T_k \geq \log_{c_\beta} \min\{(288\mu^2\ell_{g,1}^2)^{-1}, (2\alpha k)^{-2}\}.$$

Then, we have

$$\begin{aligned}\mathcal{L}_\gamma(\mathbf{w}_{k+1}) &\leq \mathcal{L}_\gamma(\mathbf{w}_k) + 3\alpha\ell_{f,1}^2\left(\rho^2 + \eta^2\right) + 15\alpha\gamma^2\ell_{g,1}^2\eta^2 \\ &\quad - \frac{1}{6\alpha}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 + \frac{6\ell_{g,1}^2\mu^2\ell_{f,0}^2}{\alpha k^2} + \frac{12\alpha\gamma^2\mu^2\eta^2\ell_{g,1}^3}{\beta}.\end{aligned}$$

Summing over $k = 1, \ldots, K$ yields

$$\begin{aligned}\sum_{k=1}^K \left(\mathcal{L}_\gamma(\mathbf{w}_{k+1}) - \mathcal{L}_\gamma(\mathbf{w}_k)\right) &\leq 3\alpha\ell_{f,1}^2 \sum_{k=1}^K \left(\rho^2 + \eta^2\right) + 15\alpha\gamma^2\ell_{g,1}^2 \sum_{k=1}^K \eta^2 \\ - \frac{1}{6\alpha}\sum_{k=1}^K \|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 &+ \sum_{k=1}^K \frac{6\ell_{g,1}^2\mu^2\ell_{f,0}^2}{\alpha k^2} + \sum_{k=1}^K \frac{12\alpha\gamma^2\mu^2\eta^2\ell_{g,1}^3}{\beta}.\end{aligned} \tag{76}$$

We have

$$\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 = \alpha^2\|\widehat{\nabla}\mathcal{L}_{\gamma,k}\|^2. \tag{77}$$

Combining (77) and (76), and rearranging terms, we obtain:

$$\begin{aligned}\sum_{k=1}^K \|\widehat{\nabla}\mathcal{L}_{\gamma,k}\|^2 &\leq \frac{6}{\alpha}\sum_{k=1}^K \left(\mathcal{L}_\gamma(\mathbf{w}_k) - \mathcal{L}_\gamma(\mathbf{w}_{k+1})\right) \\ &\quad + 18\ell_{f,1}^2 \sum_{k=1}^K \left(\rho^2 + \eta^2\right) + 90\gamma^2\ell_{g,1}^2 \sum_{k=1}^K \eta^2 \\ &\quad + 36\sum_{k=1}^K \frac{\ell_{g,1}^2\mu^2\ell_{f,0}^2}{k^2\alpha^2} + 72\sum_{k=1}^K \frac{\gamma^2\mu^2\eta^2\ell_{g,1}^3}{\beta}.\end{aligned} \tag{78}$$

On the other hand, we have:

$$\begin{aligned}&\|\nabla\mathcal{L}_\gamma(\mathbf{w}_k) - \widehat{\nabla}\mathcal{L}_{\gamma,k}\|^2 \\ &= \|\nabla f(\mathbf{w}_k) + \gamma\left(\nabla g(\mathbf{w}_k) - \nabla g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*)\right) - \nabla f(\mathbf{x}_k + \hat{\boldsymbol{\epsilon}}_k, \mathbf{y}_k + \hat{\boldsymbol{\nu}}_k) \\ &- \gamma\left(\nabla g(\mathbf{x}_k, \mathbf{y}_k + \tilde{\boldsymbol{\nu}}_k) - \bar{\nabla} g(\mathbf{x}_k, \tilde{\mathbf{y}}_k + \tilde{\boldsymbol{\nu}}_k)\right)\|^2 \\ &\leq 3\|\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{x}_k + \hat{\boldsymbol{\epsilon}}_k, \mathbf{y}_k + \hat{\boldsymbol{\nu}}_k)\|^2 \\ &+ 3\gamma^2\|\nabla g(\mathbf{w}_k) - \nabla g(\mathbf{x}_k, \mathbf{y}_k + \tilde{\boldsymbol{\nu}}_k)\|^2 \\ &+ 3\gamma^2\|\nabla g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*) - \bar{\nabla} g(\mathbf{x}_k, \tilde{\mathbf{y}}_k + \tilde{\boldsymbol{\nu}}_k)\|^2,\end{aligned}$$

where the first inequality follows from $\|a+b+c\|^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$
Moreover, from $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we obtain

$$\begin{aligned}\|\nabla\mathcal{L}_\gamma(\mathbf{w}_k) - \widehat{\nabla}\mathcal{L}_{\gamma,k}\|^2 \leq\ & 3\|\nabla f(\mathbf{w}_k) - \nabla f(\mathbf{x}_k + \hat{\boldsymbol{\epsilon}}_k, \mathbf{y}_k + \hat{\boldsymbol{\nu}}_k)\|^2 \\ & + 3\gamma^2\|\nabla g(\mathbf{w}_k) - \nabla g(\mathbf{x}_k, \mathbf{y}_k + \tilde{\boldsymbol{\nu}}_k)\|^2 \\ & + 6\gamma^2\|\nabla g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^*) - \bar{\nabla} g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^* + \tilde{\boldsymbol{\nu}}_k)\|^2 \\ & + 6\gamma^2\|\bar{\nabla} g(\mathbf{x}_k, \tilde{\mathbf{y}}_k^* + \tilde{\boldsymbol{\nu}}_k) - \bar{\nabla} g(\mathbf{x}_k, \tilde{\mathbf{y}}_k + \tilde{\boldsymbol{\nu}}_k)\|^2,\end{aligned}$$

Then, from Assumptions 2.3–II. and 2.2, we have:

$$\begin{aligned}&\|\nabla\mathcal{L}_\gamma(\mathbf{w}_k) - \widehat{\nabla}\mathcal{L}_{\gamma,k}\|^2 \\ &\leq 3\ell_{f,1}^2\|(\hat{\boldsymbol{\epsilon}}_k, \hat{\boldsymbol{\nu}}_k)\|^2 + 3\gamma^2\ell_{g,1}^2\|\tilde{\boldsymbol{\nu}}_k\|^2 \\ &+ 6\gamma^2\ell_{g,1}^2\|\tilde{\boldsymbol{\nu}}_k\|^2 + 6\gamma^2\ell_{g,1}^2\|\tilde{\mathbf{y}}_k^* - \tilde{\mathbf{y}}_k\|^2 \\ &\leq 3\ell_{f,1}^2(\rho^2+\eta^2) + 9\gamma^2\ell_{g,1}^2\eta^2 + 6\gamma^2\ell_{g,1}^2\|\tilde{\mathbf{y}}_k^* - \tilde{\mathbf{y}}_k\|^2. \end{aligned} \tag{79}$$

Then, from $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and (79), we get

$$\begin{aligned}\|\nabla\mathcal{L}_\gamma(\mathbf{w}_k)\|^2 &\leq 2\|\nabla\mathcal{L}_\gamma(\mathbf{w}_k) - \widehat{\nabla}\mathcal{L}_{\gamma,k}\|^2 + 2\|\widehat{\nabla}\mathcal{L}_{\gamma,k}\|^2 \\ &\leq 6\ell_{f,1}^2(\rho^2+\eta^2) + 18\gamma^2\ell_{g,1}^2\eta^2 \\ &\quad + 12\gamma^2\ell_{g,1}^2\|\tilde{\mathbf{y}}_k^* - \tilde{\mathbf{y}}_k\|^2 + 2\|\widehat{\nabla}\mathcal{L}_{\gamma,k}\|^2. \end{aligned} \tag{80}$$

Combining (80), (75) and (77), yields:

$$\begin{aligned}\|\nabla\mathcal{L}_\gamma(\mathbf{w}_k)\|^2 \leq\ & 6\ell_{f,1}^2(\rho^2+\eta^2) + 30\gamma^2\ell_{g,1}^2\eta^2 \\ & + \frac{13}{6}\|\widehat{\nabla}\mathcal{L}_{\gamma,k}\|^2 + \frac{12\ell_{g,1}^2\mu^2\ell_{f,0}^2}{\alpha^2k^2} + \frac{24\gamma^2\ell_{g,1}^3\mu^2\eta^2}{\beta}. \end{aligned} \tag{81}$$

Finally, from (81) and (78), we have

$$\begin{aligned}\sum_{k=1}^K\|\nabla\mathcal{L}_\gamma(\mathbf{w}_k)\|^2 \leq\ & \frac{13}{\alpha}\sum_{k=1}^K(\mathcal{L}_\gamma(\mathbf{w}_k) - \mathcal{L}_\gamma(\mathbf{w}_{k+1})) \\ & + 45\ell_{f,1}^2\sum_{k=1}^K(\rho^2+\eta^2) + 225\gamma^2\ell_{g,1}^2\sum_{k=1}^K\eta^2 \\ & + 90\sum_{k=1}^K\frac{\ell_{g,1}^2\mu^2\ell_{f,0}^2}{k^2\alpha^2} + 180\sum_{k=1}^K\frac{\gamma^2\mu^2\eta^2\ell_{g,1}^3}{\beta}.\end{aligned}$$

We have

$$\sum_{k=1}^K\frac{1}{k^2} \leq 2 - \frac{1}{K} \leq 2.$$

For any $(\mathbf{x},\mathbf{y})$, $\mathcal{L}_\gamma(\mathbf{x},\mathbf{y}) \geq \inf_{(\mathbf{x},\mathbf{y})} f(\mathbf{x},\mathbf{y})$, then from the above inequalities, we obtain:

$$\begin{aligned}\sum_{k=1}^K\|\nabla\mathcal{L}_\gamma(\mathbf{w}_k)\|^2 \leq\ & \frac{13}{\alpha}\left(\mathcal{L}_\gamma(\mathbf{w}_1) - \inf_{(\mathbf{x},\mathbf{y})} f(\mathbf{x},\mathbf{y})\right) + \sum_{k=1}^K\left(\Phi_4\rho^2 + \Phi_5\eta^2\right) \\ & + 180\frac{\ell_{g,1}^2\mu^2\ell_{f,0}^2}{\alpha^2} + 180\sum_{k=1}^K\frac{\mu^2\eta^2\ell_{g,1}^3\gamma^2}{\beta}, \end{aligned} \tag{82}$$

where $\Phi_4 = 45\ell_{f,1}^2$ and $\Phi_5 = 45\ell_{f,1}^2 + 225\gamma^2\ell_{g,1}^2$. □

G.2 Proof of Theorem 3.3

Proof. Suppose $\lim_{k\to\infty}(\mathbf{x}_k,\mathbf{y}_k) = (\mathbf{x}^*,\mathbf{y}^*)$. Since $\nabla\mathcal{L}_\gamma(\mathbf{x}_k,\mathbf{y}_k)$ is continuous, it follows that

$$\lim_{k\to\infty}\nabla\mathcal{L}_\gamma(\mathbf{x}_k,\mathbf{y}_k) = \nabla\mathcal{L}_\gamma(\mathbf{x}^*,\mathbf{y}^*).$$

By Corollary 3.2, we have $\nabla\mathcal{L}_\gamma(\mathbf{x}^*,\mathbf{y}^*) = 0$, which implies that $(\mathbf{x}^*,\mathbf{y}^*)$ is a stationary point of (12). Note that setting $\zeta_\gamma = 0$ in (29) recovers (5). Furthermore, if $(\mathbf{x}^*,\mathbf{y}^*)$ is a local/global solution, then, by Theorems C.8 and C.9, the remaining result follows. □

H PROOFS OMITTED FROM SECTION D.1

H.1 PROOF OF THEOREM D.1

Proof. Let Q denote any posterior distribution that may depend on the training data $(\mathcal{S}_f, \mathcal{S}_g)$, and let P denote the prior distribution. Consider the joint distributions P and Q, defined as

$$P = P_{\mathbf{x}} \times P_{\mathbf{y}}, \quad Q = Q_{\mathbf{x}} \times Q_{\mathbf{y}}.$$

Here, $\mathbf{x}$ is distributed according to $P_{\mathbf{x}}$ and $\mathbf{y}$ according to $P_{\mathbf{y}}$. Since $Q = Q_{\mathbf{x}} \times Q_{\mathbf{y}}$ and $P = P_{\mathbf{x}} \times P_{\mathbf{y}}$, we have:

$$KL(Q||P) = KL(Q_{\mathbf{x}}||P_{\mathbf{x}}) + KL(Q_{\mathbf{y}}||P_{\mathbf{y}}), \tag{83}$$

where $KL(Q||P)$ is the Kullback–Leibler (KL) divergence,

$$KL(Q(h)||P(h)) = \int Q(h) \log \frac{Q(h)}{P(h)} = \mathbb{E}_{H\sim Q}\left[\log \frac{Q(H)}{P(H)}\right].$$

Using the general PAC-Bayesian generalization bound (Alquier et al., 2016), we have for any prior distribution P independent of the training data, the following holds with probability at least $1-\delta$ over the choice of the training sets $(\mathcal{S}_f, \mathcal{S}_g)$:

$$\begin{aligned}\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim Q}[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y};\mathcal{D}_f,\mathcal{D}_g)] \le \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim Q}[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y};\mathcal{S}_f,\mathcal{S}_g)] \\ + \frac{1}{\varpi}\left[KL(Q||P) + \log(\frac{1}{\delta}) + \Psi(\varpi, n_1, n_2)\right],\end{aligned} \tag{84}$$

for all $\varpi > 0$, where $\mathcal{L}_\gamma(\mathbf{x},\mathbf{y};\mathcal{D}_f,\mathcal{D}_g)$ and $\mathcal{L}_\gamma(\mathbf{x},\mathbf{y};\mathcal{S}_f,\mathcal{S}_g)$ are defined in (30a) and (31a), respectively,

$$\Psi(\varpi, n_1, n_2) = \log \mathbb{E}_P \mathbb{E}_{\mathcal{S}_f,\mathcal{S}_g}\left[\exp\{\varpi\left[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y};\mathcal{D}_f,\mathcal{D}_g) - \mathcal{L}_\gamma(\mathbf{x},\mathbf{y};\mathcal{S}_f,\mathcal{S}_g)\right]\}\right].$$

Recall that we assume $|f| \le M_f$, $|g| \le M_g$. By applying Hoeffding's inequality, we obtain:

$$\Psi(\varpi, n_1, n_2) \le \varpi^2\left(\frac{M_f^2}{2n_1} + \frac{2\gamma^2 M_g^2}{n_2}\right).$$

The task is to minimize the second term of RHS of (84), we thus choose

$$\varpi^2 = \frac{(KL(Q||P) + \log(\frac{1}{\delta}))}{\left(\frac{M_f^2}{2n_1} + \frac{2\gamma^2 M_g^2}{n_2}\right)}.$$

By this choice, we have

$$\Psi(\varpi, n_1, n_2) \le \varpi^2\left(\frac{M_f^2}{2n_1} + \frac{2\gamma^2 M_g^2}{n_2}\right) = KL(Q||P) + \log(\frac{1}{\delta}).$$

Using this bound, inequality (84) becomes:

$$\begin{aligned}\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim Q}[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y};\mathcal{D}_f,\mathcal{D}_g)] \le \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim Q}[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y};\mathcal{S}_f,\mathcal{S}_g)] \\ + \frac{2}{\varpi}\left[KL(Q||P) + \log(\frac{1}{\delta})\right].\end{aligned} \tag{85}$$

Then, the second term of RHS of (85) is equal to

$$\sqrt{KL(Q||P) + \log(\frac{1}{\delta})} \times 2\left(\frac{M_f^2}{2n_1} + \frac{2\gamma^2 M_g^2}{n_2}\right)^{1/2} \le \sqrt{KL(Q||P) + \log(\frac{1}{\delta})} \times \frac{2}{\sqrt{n}}\sqrt{M(\gamma)},$$

where $n := \min(n_1, n_2)$ and $M(\gamma) := \frac{M_f^2}{2} + 2\gamma^2 M_g^2$, which implies

$$\begin{aligned}\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim Q}[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y};\mathcal{D}_f,\mathcal{D}_g)] \le \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim Q}[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y};\mathcal{S}_f,\mathcal{S}_g)] \\ + \sqrt{KL(Q||P) + \log(\frac{1}{\delta})} \times \frac{2}{\sqrt{n}}\sqrt{M(\gamma)}.\end{aligned} \tag{86}$$

Moreover, if

$$P_{\mathbf{x}} = \mathcal{N}(0, \sigma_{\mathbf{x}}^2 I_{d_1}), \quad P_{\mathbf{y}} = \mathcal{N}(0, \sigma_{\mathbf{y}}^2 I_{d_2}),$$
$$Q_{\mathbf{x}} = \mathcal{N}(\mathbf{x}, \rho^2 I_{d_1}), \quad Q_{\mathbf{y}} = \mathcal{N}(\mathbf{y}, \eta^2 I_{d_2}),$$

then the KL divergence can be written as follows:

$$KL(Q_{\mathbf{x}}||P_{\mathbf{x}}) = \frac{1}{2}\left[\frac{d_1\rho^2 + \|\mathbf{x}\|^2}{\sigma_{\mathbf{x}}^2} - d_1 + d_1 \log(\frac{\sigma_{\mathbf{x}}^2}{\rho^2})\right], \tag{87a}$$

$$KL(Q_{\mathbf{y}}||P_{\mathbf{y}}) = \frac{1}{2}\left[\frac{d_2\eta^2 + \|\mathbf{y}\|^2}{\sigma_{\mathbf{y}}^2} - d_2 + d_2 \log(\frac{\sigma_{\mathbf{y}}^2}{\eta^2})\right]. \tag{87b}$$

For given posterior distribution $Q_{\mathbf{x}}$ with fixed ρ^2, to minimize the KL term $KL(Q_{\mathbf{x}}||P_{\mathbf{x}})$, the $\sigma_{\mathbf{x}}^2$ should be equal to $\rho^2 + \|\mathbf{x}\|^2/d_1$. In this case, the KL term is no less than $\frac{d_1}{2}\log(1+\frac{\|\mathbf{x}\|^2}{d_1\rho^2})$. Similarly, for the KL term $KL(Q_{\mathbf{y}}||P_{\mathbf{y}})$, we have the lower bound $\frac{d_2}{2}\log(1+\frac{\|\mathbf{y}\|^2}{d_2\eta^2})$. Thus, we have

$$KL(Q_{\mathbf{x}}||P_{\mathbf{x}}) + KL(Q_{\mathbf{y}}||P_{\mathbf{y}}) \geq \frac{1}{2}\left[d_1 \log(1 + \frac{\|\mathbf{x}\|^2}{d_1\rho^2}) + d_2 \log(1 + \frac{\|\mathbf{y}\|^2}{d_2\eta^2})\right].$$

Under the conditions

$$\|\mathbf{x}\|^2 > d_1\rho^2(\exp(n_1/d_1) - 1), \quad \text{and} \quad \|\mathbf{y}\|^2 > d_2\eta^2(\exp(n_2/d_2) - 1),$$

we get

$$KL(Q_{\mathbf{x}}||P_{\mathbf{x}}) + KL(Q_{\mathbf{y}}||P_{\mathbf{y}}) \geq \frac{n}{2}.$$

Thus, the second term on the right-hand side is

$$\sqrt{KL(Q||P) + \log(\frac{1}{\delta})} \times \frac{2}{\sqrt{n}}\sqrt{M(\gamma)} \geq \sqrt{\frac{n}{2}}\frac{2}{\sqrt{n}}\sqrt{M(\gamma)} = \sqrt{2}\sqrt{M(\gamma)}.$$

Hence, for any $\|\mathbf{x}\|^2 > d_1\rho^2(\exp(n_1/d_1) - 1)$ and $\|\mathbf{y}\|^2 > d_2\eta^2(\exp(n_2/d_2) - 1)$, the RHS of (34) is greater than the LHS of (34); thus, the inequality holds trivially.

In the remainder of the proof, we only consider the case:

$$\|\mathbf{x}\|^2 < d_1\rho^2(\exp(n_1/d_1) - 1), \quad \text{and} \quad \|\mathbf{y}\|^2 < d_2\eta^2(\exp(n_2/d_2) - 1).$$

Distribution P is Gaussian centered around 0 with variances $\sigma_{\mathbf{x}}^2 = \rho^2 + \|\mathbf{x}\|^2/d_1$ and $\sigma_{\mathbf{y}}^2 = \eta^2 + \|\mathbf{y}\|^2/d_2$, which are unknown at the time we set up the inequality, since $\mathbf{x}$ and $\mathbf{y}$ are unknown. Meanwhile we have to specify P in advance, since P is the prior distribution. To deal with this problem, we apply the union bound technique (Dziugaite & Roy, 2017; Foret et al., 2020). We set

$$c_{\mathbf{x}} = d_1\rho^2(1 + \exp(n_1/d_1)), \quad \text{and} \quad c_{\mathbf{y}} = d_2\eta^2(1 + \exp(n_2/d_2)),$$

then

$$P_{j_{\mathbf{x}}} = \mathcal{N}(0, \sigma_{j_{\mathbf{x}}}^2 I_{d_1}), \quad \sigma_{j_{\mathbf{x}}}^2 = c_{\mathbf{x}} \exp((1 - j_{\mathbf{x}})/d_1),$$
$$P_{j_{\mathbf{y}}} = \mathcal{N}(0, \sigma_{j_{\mathbf{y}}}^2 I_{d_2}), \quad \sigma_{j_{\mathbf{y}}}^2 = c_{\mathbf{y}} \exp((1 - j_{\mathbf{y}})/d_2),$$

and

$$\mathfrak{B} := \{(P_{j_{\mathbf{x}}}, P_{j_{\mathbf{y}}}) : j = 1, 2, \ldots\}.$$

Then the following inequality holds for a particular distribution P_j with probability $1 - \delta_{j_{\mathbf{x}}}\delta_{j_{\mathbf{y}}}$, where $\delta_{j_{\mathbf{x}}} = \frac{6\sqrt{\delta}}{\pi^2 j_{\mathbf{x}}^2}$ and $\delta_{j_{\mathbf{y}}} = \frac{6\sqrt{\delta}}{\pi^2 j_{\mathbf{y}}^2}$,

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim Q}[\mathcal{L}_\gamma(\mathbf{x}, \mathbf{y}; \mathcal{D}_f, \mathcal{D}_g)] \leq \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim Q}[\mathcal{L}_\gamma(\mathbf{x}, \mathbf{y}; \mathcal{S}_f, \mathcal{S}_g)] + \frac{1}{\varpi}\left[KL(Q||P_j) + \log(\frac{1}{\delta_{j_{\mathbf{x}}}\delta_{j_{\mathbf{y}}}}) + \Psi(\varpi, n_1, n_2)\right].$$

Since $\sum_{j=1}^{\infty} 1/j^2 = \pi^2/6$, by the union bound, the above inequality holds simultaneously for all $j_{\mathbf{x}}, j_{\mathbf{y}}$ with probability at least

$$1 - \sum_{j_{\mathbf{x}}=1}^{\infty} \sum_{j_{\mathbf{y}}=1}^{\infty} \frac{36\delta}{\pi^4 j_{\mathbf{x}}^2 j_{\mathbf{y}}^2} = 1 - \delta.$$

We pick

$$j_{\mathbf{x}}^* := \lfloor 1 - d_1 \log(\frac{d_1(\rho^2 + \|\mathbf{x}\|^2/d_1)}{c_{\mathbf{x}}}) \rfloor = \lfloor 1 - d_1 \log(\frac{(\rho^2 + \|\mathbf{x}\|^2/d_1)}{\rho^2(1+\exp(n_1/d_1))}) \rfloor,$$
$$j_{\mathbf{y}}^* := \lfloor 1 - d_2 \log(\frac{d_2(\eta^2 + \|\mathbf{y}\|^2/d_2)}{c_{\mathbf{y}}}) \rfloor = \lfloor 1 - d_2 \log(\frac{(\eta^2 + \|\mathbf{y}\|^2/d_2)}{\eta^2(1+\exp(n_2/d_2))}) \rfloor.$$

Therefore,

$$\begin{aligned}
&1 - j_{\mathbf{x}}^* = \lceil d_1 \log((\rho^2 + \|\mathbf{x}\|^2/d_1)/c_{\mathbf{x}}) \rceil \geq d_1 \log((\rho^2 + \|\mathbf{x}\|^2/d_1)/c_{\mathbf{x}}) \\
&\Rightarrow \log(\frac{(\rho^2 + \|\mathbf{x}\|^2/d_1)}{c_{\mathbf{x}}}) \leq \frac{1 - j_{\mathbf{x}}^*}{d_1} \leq \log(\frac{(\rho^2 + \|\mathbf{x}\|^2/d_1)}{c_{\mathbf{x}}}) + \frac{1}{d_1} \\
&\Rightarrow \rho^2 + \|\mathbf{x}\|^2/d_1 \leq c_{\mathbf{x}} \exp(\frac{1 - j_{\mathbf{x}}^*}{d_1}) \leq \exp(1/d_1)(\rho^2 + \|\mathbf{x}\|^2/d_1) \\
&\Rightarrow \rho^2 + \|\mathbf{x}\|^2/d_1 \leq \sigma_{j_{\mathbf{x}}^*}^2 \leq \exp(1/d_1)(\rho^2 + \|\mathbf{x}\|^2/d_1).
\end{aligned}$$

Thus, the KL term could be bounded as follow

$$\begin{aligned}
KL(Q_{\mathbf{x}}||P_{j_{\mathbf{x}}^*}) &= \frac{1}{2}\left[\frac{d_1\rho^2 + \|\mathbf{x}\|^2}{\sigma_{j_{\mathbf{x}}^*}^2} - d_1 + d_1 \log(\frac{\sigma_{j_{\mathbf{x}}^*}^2}{\rho^2})\right], \\
&\leq \frac{1}{2}\left[\frac{d_1\rho^2 + \|\mathbf{x}\|^2}{\rho^2 + \|\mathbf{x}\|^2/d_1} - d_1 + d_1 \log(\frac{\exp(1/d_1)(\rho^2 + \|\mathbf{x}\|^2/d_1)}{\rho^2})\right] \\
&= \frac{1}{2}\left[d_1 \log(\frac{\exp(1/d_1)(\rho^2 + \|\mathbf{x}\|^2/d_1)}{\rho^2})\right] \\
&= \frac{1}{2}\left[1 + d_1 \log(1 + \frac{\|\mathbf{x}\|^2}{d_1\rho^2})\right]. \qquad (88a)
\end{aligned}$$

By similar argument, we obtain:

$$KL(Q_{\mathbf{y}}||P_{j_{\mathbf{y}}^*}) \leq \frac{1}{2}\left[1 + d_2 \log(1 + \frac{\|\mathbf{y}\|^2}{d_2\eta^2})\right]. \qquad (88b)$$

From (83), (88a), and (88b), we obtain

$$\begin{aligned}
KL(Q||P_{j^*}) &\leq \frac{1}{2}\left[2 + d_1 \log(1 + \frac{\|\mathbf{x}\|^2}{d_1\rho^2}) + d_2 \log(1 + \frac{\|\mathbf{y}\|^2}{d_2\eta^2})\right] \\
&= \frac{1}{2}\left[2 + \log(1 + \frac{\|\mathbf{x}\|^2}{d_1\rho^2})^{d_1}(1 + \frac{\|\mathbf{y}\|^2}{d_2\eta^2})^{d_2}\right]. \qquad (89)
\end{aligned}$$

For the term $\log \frac{1}{\delta_{j_{\mathbf{x}}^*}\delta_{j_{\mathbf{y}}^*}}$, with recall that

$$c_{\mathbf{x}} = d_1\rho^2(1 + \exp(n_1/d_1)), \quad \text{and} \quad j_{\mathbf{x}}^* = \lfloor 1 - d_1 \log(\frac{(\rho^2 + \|\mathbf{x}\|^2/d_1)}{\rho^2(1+\exp(n_1/d_1))}) \rfloor,$$

we have:

$$\begin{aligned}
\log \frac{1}{\delta_{j^*_{\mathbf{x}}}\delta_{j^*_{\mathbf{y}}}} &= \log \frac{j^{*2}_{\mathbf{y}} j^{*2}_{\mathbf{x}} \pi^2}{6\delta} \\
&= \log \frac{1}{\delta} + 2\log\frac{\pi^2}{6} + 2\log(j^*_{\mathbf{x}}) + 2\log(j^*_{\mathbf{y}}) \\
&\le \log \frac{1}{\delta} + 2\log\frac{\pi^2}{6} + 2\log\left(1 + d_1 \log(\frac{\rho^2(1+\exp(n_1/d_1))}{(\rho^2 + \|\mathbf{x}\|^2/d_1)})\right) \\
&\qquad + 2\log\left(1 + d_2 \log(\frac{\eta^2(1+\exp(n_2/d_2))}{(\eta^2 + \|\mathbf{y}\|^2/d_2)})\right) \\
&\le \log \frac{1}{\delta} + 2\log\frac{\pi^2}{6} + 2\log\left(1 + d_1 \log(1+\exp(n_1/d_1))\right) \\
&\qquad + 2\log\left(1 + d_2 \log(1+\exp(n_2/d_2))\right),
\end{aligned}$$

which implies

$$\begin{aligned}
\log \frac{1}{\delta_{j^*_{\mathbf{x}}}\delta_{j^*_{\mathbf{y}}}} &\le \log \frac{1}{\delta} + 2\log\frac{\pi^2}{6} + 2\log\left(1 + d_1(1 + \frac{n_1}{d_1})\right) \\
&\qquad + 2\log\left(1 + d_2(1 + \frac{n_2}{d_2})\right) \\
&= \log \frac{1}{\delta} + \log(\frac{\pi^2}{6})^2 (1+d_1+n_1)^2(1+d_2+n_2)^2. \qquad (90)
\end{aligned}$$

Substituting (90) and (89) into (86), we obtain

$$\begin{aligned}
&\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim Q}[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y};\mathcal{D}_f,\mathcal{D}_g)] \\
&\le \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim Q}[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y};\mathcal{S}_f,\mathcal{S}_g)] + \frac{2}{\sqrt{n}}\sqrt{M(\gamma)}\sqrt{KL(Q||P_{j^*}) + \log(\frac{1}{\delta_{j^*_{\mathbf{x}}}\delta_{j^*_{\mathbf{y}}}})} \\
&\le \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim Q}[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y};\mathcal{S}_f,\mathcal{S}_g)] + \frac{2}{\sqrt{n}}\sqrt{M(\gamma)}\left(\log(1 + \frac{\|\mathbf{x}\|^2}{d_1\rho^2})^{\frac{d_1}{2}}(1 + \frac{\|\mathbf{y}\|^2}{d_2\eta^2})^{\frac{d_2}{2}} + A\right)^{1/2},
\end{aligned}$$

where $A := 1 + \log(\frac{1}{\delta}) + \log(\frac{\pi^2}{6})^2(1+d_1+n_1)^2(1+d_2+n_2)^2$. Fix

$$\rho'^2 = \frac{\rho^2}{d_1\left(1 + \sqrt{\frac{\log(n_1)}{d_1}}\right)^2}, \quad \text{and} \quad \eta'^2 = \frac{\eta^2}{d_2\left(1 + \sqrt{\frac{\log(n_2)}{d_2}}\right)^2}, \qquad (91)$$

and consider perturbations $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \rho'^2\mathbf{I}_{d_1})$ and $\boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \eta'^2\mathbf{I}_{d_2})$. Thus, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\rho'^2 I_{d_1}),\boldsymbol{\nu}\sim\mathcal{N}(0,\eta'^2 I_{d_2})}[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{D}_f,\mathcal{D}_g)] \\
&\le \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\rho'^2 I_{d_1}),\boldsymbol{\nu}\sim\mathcal{N}(0,\eta'^2 I_{d_2})}[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{S}_f,\mathcal{S}_g)] \\
&+ \frac{2}{\sqrt{n}}\sqrt{M(\gamma)}\left(\log(1 + \frac{\|\mathbf{x}\|^2}{d_1\rho'^2})^{\frac{d_1}{2}}(1 + \frac{\|\mathbf{y}\|^2}{d_2\eta'^2})^{\frac{d_2}{2}} + A\right)^{1/2}. \qquad (92)
\end{aligned}$$

Bounding the empirical expectation: The next step is to bound the expectation in the RHS of (92) by the worst-case loss. We make use of the following result from (Laurent & Massart, 2000):

$$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \rho'^2 I_{d_1}) \Rightarrow \|\boldsymbol{\epsilon}\|^2 \le \rho'^2 d_1\left(1 + \sqrt{\frac{\log(n_1)}{d_1}}\right)^2 = \rho^2, \qquad (93a)$$

with probability at least $1 - 1/\sqrt{n_1}$ and

$$\boldsymbol{\nu} \sim \mathcal{N}(0, \eta'^2 I_{d_2}) \Rightarrow \|\boldsymbol{\nu}\|^2 \le \eta'^2 d_2\left(1 + \sqrt{\frac{\log(n_2)}{d_2}}\right)^2 = \eta^2, \qquad (93b)$$

with probability at least $1 - 1/\sqrt{n_2}$.
The inequalities in (93) enable us to express the expectation in terms of the maximum value over a set of bounded perturbations. To this end, let

$$\boldsymbol{\zeta} := (\boldsymbol{\epsilon}, \boldsymbol{\nu}) \sim \mathcal{N}\big(0, \Sigma\big), \quad \Sigma := \operatorname{diag}\big(\rho^2 I_{d_1}, \eta^2 I_{d_2}\big).$$

Thus, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{N}\left(0,\Sigma\right)}[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{S}_f,\mathcal{S}_g)] \\
&\le \mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{N}\left(0,\Sigma\right)}\left[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{S}_f,\mathcal{S}_g) \mid \|\boldsymbol{\epsilon}\| \le \rho, \|\boldsymbol{\nu}\| \le \eta\right] \mathbb{P}(\|\boldsymbol{\epsilon}\| \le \rho, \|\boldsymbol{\nu}\| \le \eta) \\
&+ \mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{N}\left(0,\Sigma\right)}\left[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{S}_f,\mathcal{S}_g) \mid \|\boldsymbol{\epsilon}\| > \rho, \|\boldsymbol{\nu}\| \le \eta\right] \mathbb{P}(\|\boldsymbol{\epsilon}\| > \rho, \|\boldsymbol{\nu}\| \le \eta) \\
&+ \mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{N}\left(0,\Sigma\right)}\left[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{S}_f,\mathcal{S}_g) \mid \|\boldsymbol{\epsilon}\| \le \rho, \|\boldsymbol{\nu}\| > \eta\right] \mathbb{P}(\|\boldsymbol{\epsilon}\| \le \rho, \|\boldsymbol{\nu}\| > \eta) \\
&+ \mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{N}\left(0,\Sigma\right)}\left[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{S}_f,\mathcal{S}_g) \mid \|\boldsymbol{\epsilon}\| > \rho, \|\boldsymbol{\nu}\| > \eta\right] \mathbb{P}(\|\boldsymbol{\epsilon}\| > \rho, \|\boldsymbol{\nu}\| > \eta) \\
&\le (1 - \frac{1}{\sqrt{n_1}})(1 - \frac{1}{\sqrt{n_2}}) \max_{\|\boldsymbol{\epsilon}\|\le\rho, \|\boldsymbol{\nu}\|\le\eta} \mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{S}_f,\mathcal{S}_g) \\
&+ \big(1 - (1 - \frac{1}{\sqrt{n_1}})(1 - \frac{1}{\sqrt{n_2}})\big)\acute{M}(\gamma),
\end{aligned}$$

where the second inequality follows from $|\mathcal{L}_\gamma| \le \acute{M}(\gamma) := M_f + 2\gamma M_g$, which implies

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{N}\left(0,\Sigma\right)}[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{S}_f,\mathcal{S}_g)] \\
&\le \max_{\|\boldsymbol{\epsilon}\|\le\rho, \|\boldsymbol{\nu}\|\le\eta} \mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{S}_f,\mathcal{S}_g) \\
&+ \big(1 - (1 - \frac{1}{\sqrt{n_1}})(1 - \frac{1}{\sqrt{n_2}})\big)\big(\acute{M}(\gamma) - \max_{\|\boldsymbol{\epsilon}\|\le\rho, \|\boldsymbol{\nu}\|\le\eta} \mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{S}_f,\mathcal{S}_g)\big) \\
&\le \max_{\|\boldsymbol{\epsilon}\|\le\rho, \|\boldsymbol{\nu}\|\le\eta} \mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{S}_f,\mathcal{S}_g) + 2\acute{M}(\gamma)\big(1 - (1 - \frac{1}{\sqrt{n_1}})(1 - \frac{1}{\sqrt{n_2}})\big) \\
&\le \max_{\|\boldsymbol{\epsilon}\|\le\rho, \|\boldsymbol{\nu}\|\le\eta} \mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{S}_f,\mathcal{S}_g) + 2\acute{M}(\gamma)\frac{2}{\sqrt{n}},
\end{aligned} \tag{94}$$

where the last inequality holds since we have

$$1 - (1 - \frac{1}{\sqrt{n_1}})(1 - \frac{1}{\sqrt{n_2}}) = \frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_2}} - \frac{1}{\sqrt{n_1 n_2}} \le \frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_2}} \le \frac{2}{\sqrt{n}}.$$

From (94), we obtain

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\rho^2 I_{d_1}),\boldsymbol{\nu}\sim\mathcal{N}(0,\eta^2 I_{d_2})}[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{S}_f,\mathcal{S}_g)] \\
&\le \max_{\|\boldsymbol{\epsilon}\|\le\rho, \|\boldsymbol{\nu}\|\le\eta} \mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{S}_f,\mathcal{S}_g) + \frac{4}{\sqrt{n}}\acute{M}(\gamma).
\end{aligned} \tag{95}$$

Using assumption (33), we have

$$\mathcal{L}_\gamma(\mathbf{x},\mathbf{y};\mathcal{D}_f,\mathcal{D}_g) \le \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\rho^2 I_{d_1}),\boldsymbol{\nu}\sim\mathcal{N}(0,\eta^2 I_{d_2})}[\mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{D}_f,\mathcal{D}_g)]. \tag{96}$$

Substituting (95) and (96) into (92) yields

$$\begin{aligned}
&\mathcal{L}_\gamma(\mathbf{x},\mathbf{y};\mathcal{D}_f,\mathcal{D}_g) \\
&\le \max_{\|\boldsymbol{\epsilon}\|\le\rho, \|\boldsymbol{\nu}\|\le\eta} \mathcal{L}_\gamma(\mathbf{x},\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{S}_f,\mathcal{S}_g) + \frac{4}{\sqrt{n}}\acute{M}(\gamma) \\
&+ \frac{2}{\sqrt{n}}\sqrt{M(\gamma)}\left(\log(1 + \frac{\|\mathbf{x}\|^2}{d_1\rho'^2})^{\frac{d_1}{2}}(1 + \frac{\|\mathbf{y}\|^2}{d_2\eta'^2})^{\frac{d_2}{2}} + A\right)^{1/2}.
\end{aligned}$$

From (91), we have

$$\begin{aligned}
&\log(1 + \frac{\|\mathbf{x}\|^2}{d_1\rho'^2})^{\frac{d_1}{2}}(1 + \frac{\|\mathbf{y}\|^2}{d_2\eta'^2})^{\frac{d_2}{2}} \\
&= \log\left(1 + \frac{\|\mathbf{x}\|^2}{\rho^2}\left(1 + \sqrt{\frac{\log(n_1)}{d_1}}\right)^2\right)^{\frac{d_1}{2}}\left(1 + \frac{\|\mathbf{y}\|^2}{\eta^2}\left(1 + \sqrt{\frac{\log(n_2)}{d_2}}\right)^2\right)^{\frac{d_2}{2}},
\end{aligned}$$

which completes the proof. □

I PROOFS OMITTED FROM SECTION 4

This section provides proofs for Theorems 4.3 and D.2. We begin by presenting several auxiliary lemmas, which will later be used to establish Theorem 4.3.

Lemma I.1. *Suppose Assumption 2.3–II. holds and $\beta \le 1/\ell_{f,1}$. Then, the iterates $\{\mathbf{x}_k\}$ produced by Algorithm 2 satisfy:*

$$-\beta\langle\nabla_{\mathbf{z}} f(\mathbf{x}_k+\boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)}), \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\rangle \le \frac{\ell_{f,1}}{2}(\rho^2+\eta^2).$$

Proof. We first add and subtract terms to obtain:

$$-\beta\langle\nabla_{\mathbf{z}} f(\mathbf{x}_k+\boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)}), \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\rangle = I_1 + I_2, \tag{97}$$

where

$$I_1 := -\beta\langle\nabla_{\mathbf{z}} f(\mathbf{x}_k+\boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)}), \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\rangle,$$
$$I_2 := -\beta\langle\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)}), \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\rangle.$$

For the first term on the right-hand side of (97), we have:

$$\begin{aligned} I_1 &\le \frac{\beta}{2}\|\nabla_{\mathbf{z}} f(\mathbf{x}_k+\boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)})\|^2 \\ &\quad + \frac{\beta}{2}\|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 \\ &\le \frac{\beta}{2}\ell_{f,1}^2\|\boldsymbol{\epsilon}_t^{(k)}\|^2 + \frac{\beta}{2}\|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2, \end{aligned}$$

where the last inequality follows from Assumption 2.3–II..
Recall that, from Algorithm 2, we have

$$\boldsymbol{\nu}_t^{(k)} = \frac{\eta\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})}{\|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|}, \quad \text{and} \quad \boldsymbol{\epsilon}_t^{(k)} = \frac{\rho\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})}{\|\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|},$$

where $\|\boldsymbol{\nu}_t^{(k)}\| = \eta$ and $\|\boldsymbol{\epsilon}_t^{(k)}\| = \rho$.

Then, we have

$$I_1 \le \frac{\beta}{2}\ell_{f,1}^2\rho^2 + \frac{\beta}{2}\|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2. \tag{98}$$

For the second term on the right-hand side of (97), we have

$$\begin{aligned} &\langle\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)}), \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\rangle \\ &= \langle\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)}), \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\rangle + \|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 \\ &= \frac{\|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|}{\eta}\langle\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)}), \boldsymbol{\nu}_t^{(k)}\rangle \\ &+ \|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 \\ &\ge -\ell_{f,1}\frac{\|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|}{\eta}\|\boldsymbol{\nu}_t^{(k)}\|^2 + \|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2, \end{aligned}$$

where the inequality holds since, by Assumption 2.3–II., we have

$$\langle\nabla f(w_1) - \nabla f(w_2), w_1 - w_2\rangle \ge -\ell_{f,1}\|w_1 - w_2\|^2.$$

Then, we have

$$\begin{aligned} I_2 &\le \beta\ell_{f,1}\frac{\|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|}{\eta}\|\boldsymbol{\nu}_t^{(k)}\|^2 - \beta\|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 \\ &= \beta\ell_{f,1}\eta\|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\| - \beta\|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 \\ &\le \frac{\eta^2\ell_{f,1}}{2} + \frac{\beta^2\ell_{f,1}}{2}\|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 - \beta\|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2, \end{aligned} \tag{99}$$

where the inequality follows from Young's inequality: $2\langle a,b\rangle \le v\|a\|^2 + \frac{1}{v}\|b\|^2$ for scalar $v>0$. Substituting (98) and (99) into (97), we have

$$\begin{aligned}&-\beta\langle\nabla_{\mathbf{z}} f(\mathbf{x}_k+\boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)}), \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\rangle \\ &\le \frac{\ell_{f,1}^2}{2}\beta\rho^2 + \frac{\ell_{f,1}}{2}\eta^2 + \left(\frac{\beta^2\ell_{f,1}}{2} - \frac{\beta}{2}\right)\|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2.\end{aligned}$$

Finally, from $\beta \le 1/\ell_{f,1}$, we get the desired result. □

Lemma I.2. *Suppose Assumptions 2.3–II. and 4.1 hold. Let*

$$\alpha \le \min\left\{\frac{\beta}{16L_f\tilde{\mu}}, \frac{\beta}{16\ell_{f,1}^2\tilde{\mu}^2}, \frac{1}{16\tilde{\mu}\ell_{f,1}^2}\right\}, \quad \beta \le \frac{1}{\ell_{f,1}},$$

and $\Delta_k := f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_k, \mathbf{y}_k)$*, where* $\mathbf{y}_k^* \in \arg\max_{\mathbf{y}\in\mathbb{R}^{d_2}} f(\mathbf{x}_k, \mathbf{y})$*. Then, we have*

$$\begin{aligned}\sum_{k=1}^K(\Delta_{k+1}-\Delta_k) \le -\frac{\beta}{4\tilde{\mu}}\sum_{k=1}^K\Delta_k + \frac{\alpha}{4}\sum_{k=1}^K\|\nabla_{\mathbf{x}} f\big(\mathbf{x}_k+\boldsymbol{\epsilon}_k, \mathbf{y}_{k+1}+\boldsymbol{\nu}_k\big)\|^2 \\ + \frac{K}{8\alpha}(\rho^2+\eta^2),\end{aligned}$$

where L_f *is defined in Lemma C.2.*

Proof. From Assumption 2.3–II., we have

$$\begin{aligned}&-f(\mathbf{x}_{k+1}, \mathbf{z}_{t+1}^{(k)}) \\ &\le -f(\mathbf{x}_{k+1}, \mathbf{z}_t^{(k)}) - \langle\nabla_{\mathbf{z}} f(\mathbf{x}_{k+1}, \mathbf{z}_t^{(k)}), \mathbf{z}_{t+1}^{(k)} - \mathbf{z}_t^{(k)}\rangle + \frac{\ell_{f,1}}{2}\|\mathbf{z}_{t+1}^{(k)} - \mathbf{z}_t^{(k)}\|^2 \\ &= -f(\mathbf{x}_{k+1}, \mathbf{z}_t^{(k)}) - \beta\langle\nabla_{\mathbf{z}} f(\mathbf{x}_{k+1}, \mathbf{z}_t^{(k)}), \nabla_{\mathbf{z}} f(\mathbf{x}_k+\boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)})\rangle \\ &+ \frac{\ell_{f,1}}{2}\beta^2\|\nabla_{\mathbf{z}} f(\mathbf{x}_k+\boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)})\|^2, \qquad (100)\end{aligned}$$

where the equality follows from $\mathbf{z}_{t+1}^{(k)} = \mathbf{z}_t^{(k)} + \beta\,\nabla_{\mathbf{z}} f\big(\mathbf{x}_k+\boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)}\big)$.

For the second term on the right-hand side of (100), using

$$\langle a,b\rangle = \frac{1}{2}\left(\|a\|^2 + \|b\|^2 - \|a-b\|^2\right),$$

we have:

$$\begin{aligned}&-\beta\langle\nabla_{\mathbf{z}} f(\mathbf{x}_{k+1}, \mathbf{z}_t^{(k)}), \nabla_{\mathbf{z}} f(\mathbf{x}_k+\boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)})\rangle \\ &= -\frac{\beta}{2}\Big(\|\nabla_{\mathbf{z}} f(\mathbf{x}_{k+1}, \mathbf{z}_t^{(k)})\|^2 + \|\nabla_{\mathbf{z}} f(\mathbf{x}_k+\boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)})\|^2 \\ &-\|\nabla_{\mathbf{z}} f(\mathbf{x}_{k+1}, \mathbf{z}_t^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k+\boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)})\|^2\Big).\end{aligned}$$

Let $\mathbf{y}_k^* \in \arg\max_{\mathbf{y}\in\mathbb{R}^{d_2}} f(\mathbf{x}_k, \mathbf{y})$. Then, we have

$$\begin{aligned}&-\beta\langle\nabla_{\mathbf{z}} f(\mathbf{x}_{k+1}, \mathbf{z}_t^{(k)}), \nabla_{\mathbf{z}} f(\mathbf{x}_k+\boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)})\rangle \\ &\le -\frac{\beta}{2}\|\nabla_{\mathbf{z}} f(\mathbf{x}_{k+1}, \mathbf{z}_t^{(k)})\|^2 - \frac{\beta}{2}\|\nabla_{\mathbf{z}} f(\mathbf{x}_k+\boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)})\|^2 \\ &+ \frac{\beta}{2}\|\nabla_{\mathbf{z}} f(\mathbf{x}_{k+1}, \mathbf{z}_t^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k+\boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)})\|^2 \\ &\le -\frac{\beta}{2\tilde{\mu}}\Big(f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}^*) - f(\mathbf{x}_{k+1}, \mathbf{z}_t^{(k)})\Big) - \frac{\beta}{2}\|\nabla_{\mathbf{z}} f(\mathbf{x}_k+\boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)})\|^2 \\ &+ \frac{\beta}{2}\|\nabla_{\mathbf{z}} f(\mathbf{x}_{k+1}, \mathbf{z}_t^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k+\boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)}+\boldsymbol{\nu}_t^{(k)})\|^2, \qquad (101)\end{aligned}$$

where the second inequality follows from Assumption 4.1.
Combining (100) and (101), we have

$$
\begin{aligned}
&- f(\mathbf{x}_{k+1}, \mathbf{z}_{t+1}^{(k)}) \leq -f(\mathbf{x}_{k+1}, \mathbf{z}_t^{(k)}) + \frac{\ell_{f,1}}{2}\beta^2 \|\nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)} + \boldsymbol{\nu}_t^{(k)})\|^2 \\
&- \frac{\beta}{2\tilde{\mu}} \left(f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}^*) - f(\mathbf{x}_{k+1}, \mathbf{z}_t^{(k)}) \right) - \frac{\beta}{2} \|\nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)} + \boldsymbol{\nu}_t^{(k)})\|^2 \\
&+ \frac{\beta}{2} \|\nabla_{\mathbf{z}} f(\mathbf{x}_{k+1}, \mathbf{z}_t^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)} + \boldsymbol{\nu}_t^{(k)})\|^2. \qquad (102)
\end{aligned}
$$

Then, by rearranging (102), we obtain:

$$
\begin{aligned}
&f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}^*) - f(\mathbf{x}_{k+1}, \mathbf{z}_{t+1}^{(k)}) \\
&\leq \left(1 - \frac{\beta}{2\tilde{\mu}}\right) \left(f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}^*) - f(\mathbf{x}_{k+1}, \mathbf{z}_t^{(k)}) \right) \\
&+ \left(\frac{\ell_{f,1}}{2}\beta^2 - \frac{\beta}{2} \right) \|\nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)} + \boldsymbol{\nu}_t^{(k)})\|^2 \\
&+ \frac{\beta}{2} \|\nabla_{\mathbf{z}} f(\mathbf{x}_{k+1}, \mathbf{z}_t^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)} + \boldsymbol{\nu}_t^{(k)})\|^2.
\end{aligned}
$$

Since $\mathbf{y}_{k+1} = \mathbf{z}_{T_k+1}^{(k)}$, by applying T_k iterations, we obtain:

$$
\begin{aligned}
&f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}^*) - f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) \\
&\leq \left(1 - \frac{\beta}{2\tilde{\mu}}\right)^{T_k} \left(f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}^*) - f(\mathbf{x}_{k+1}, \mathbf{y}_k) \right) \\
&+ \left(\frac{\ell_{f,1}}{2}\beta^2 - \frac{\beta}{2} \right) \sum_{j=1}^{T_k} \left(1 - \frac{\beta}{2\tilde{\mu}}\right)^{T_k - j} \|\nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_j^{(k)}, \mathbf{z}_j^{(k)} + \boldsymbol{\nu}_j^{(k)})\|^2 \\
&+ \frac{\beta}{2} \sum_{j=1}^{T_k} \left(1 - \frac{\beta}{2\tilde{\mu}}\right)^{T_k - j} \|\nabla_{\mathbf{z}} f(\mathbf{x}_{k+1}, \mathbf{z}_j^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_j^{(k)}, \mathbf{z}_j^{(k)} + \boldsymbol{\nu}_j^{(k)})\|^2. \qquad (103)
\end{aligned}
$$

Next, we want to bound the second term on the right-hand side of (103). Using the $\ell_{f,1}$-smoothness of the function $-f(\cdot, \mathbf{y}_k)$ from Assumption 2.3–II., we have:

$$
-f(\mathbf{x}_{k+1}, \mathbf{y}_k) \leq -f(\mathbf{x}_k, \mathbf{y}_k) - \langle \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{\ell_{f,1}}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2.
$$

Then, we have

$$
\begin{aligned}
-f(\mathbf{x}_{k+1}, \mathbf{y}_k) + f(\mathbf{x}_k, \mathbf{y}_k) &\leq -\langle \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{\ell_{f,1}}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\
&= \langle \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k^*) - \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle \\
&\quad - \langle \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k^*), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{\ell_{f,1}}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2.
\end{aligned}
$$

Moreover, we have

$$
\begin{aligned}
&- f(\mathbf{x}_{k+1}, \mathbf{y}_k) + f(\mathbf{x}_k, \mathbf{y}_k) \\
&\leq \frac{1}{8\alpha} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + 2\alpha \|\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k^*)\|^2 \\
&- \langle \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k^*), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{\ell_{f,1}}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\
&\leq \frac{1}{8\alpha} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + 2\ell_{f,1}^2 \alpha \|\mathbf{y}_k - \mathbf{y}_k^*\|^2 \\
&+ f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}^*) + \frac{L_f}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + \frac{\ell_{f,1}}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\
&\leq 4\ell_{f,1}^2 \alpha \tilde{\mu} \left(f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_k, \mathbf{y}_k) \right) \\
&+ f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}^*) + \left(\frac{1}{8\alpha} + L_f \right) \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2,
\end{aligned}
$$

where the second inequality is due to the L_f-smoothness of function $f(\mathbf{x}, \mathbf{y}^*)$, and the last inequality holds by Lemma B.1 and $\ell_{f,1} \leq L_f$. Therefore we have

$$\begin{aligned} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}^*) - f(\mathbf{x}_{k+1}, \mathbf{y}_k) &= f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}^*) - f(\mathbf{x}_k, \mathbf{y}_k^*) + f(\mathbf{x}_k, \mathbf{y}_k^*) \\ &\quad - f(\mathbf{x}_k, \mathbf{y}_k) + f(\mathbf{x}_k, \mathbf{y}_k) - f(\mathbf{x}_{k+1}, \mathbf{y}_k) \\ &\leq \left(1 + 4\ell_{f,1}^2 \alpha\tilde{\mu}\right) \left(f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_k, \mathbf{y}_k)\right) \\ &\quad + \left(\frac{1}{8\alpha} + L_f\right) \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2. \end{aligned} \tag{104}$$

Substituting (104) in (103), we get

$$\begin{aligned} & f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}^*) - f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) \\ &\leq \left(1 - \frac{\beta}{2\tilde{\mu}}\right)^{T_k} \left(1 + 4\ell_{f,1}^2 \alpha\tilde{\mu}\right) \left(f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_k, \mathbf{y}_k)\right) \\ &\quad + \left(1 - \frac{\beta}{2\tilde{\mu}}\right)^{T_k} \left(\frac{1}{8\alpha} + L_f\right) \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &\quad + \left(\frac{\ell_{f,1}}{2}\beta^2 - \frac{\beta}{2}\right) \sum_{j=1}^{T_k} \left(1 - \frac{\beta}{2\tilde{\mu}}\right)^{T_k - j} \|\nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_j^{(k)}, \mathbf{z}_j^{(k)} + \boldsymbol{\nu}_j^{(k)})\|^2 \\ &\quad + \frac{\beta}{2} \sum_{j=1}^{T_k} \left(1 - \frac{\beta}{2\tilde{\mu}}\right)^{T_k - j} \|\nabla_{\mathbf{z}} f(\mathbf{x}_{k+1}, \mathbf{z}_j^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_j^{(k)}, \mathbf{z}_j^{(k)} + \boldsymbol{\nu}_j^{(k)})\|^2. \end{aligned} \tag{105}$$

Since $\alpha \leq \min\left\{\frac{\beta}{16\tilde{\mu}L_f}, \frac{\beta}{16\tilde{\mu}^2\ell_{f,1}^2}\right\}$, we have

$$\alpha \leq \frac{\beta}{16\tilde{\mu}^2\ell_{f,1}^2} \quad \Rightarrow \quad 4\ell_{f,1}^2\alpha\tilde{\mu} \leq \frac{\beta}{4\tilde{\mu}}.$$

Then, from $(1-a)^T(1+\frac{a}{2}) \leq 1 - \frac{a}{2} - \frac{a^2}{2} \leq 1 - \frac{a}{2}$ for any $a > 0$ and $T \geq 1$, we have

$$\left(1 - \frac{\beta}{2\tilde{\mu}}\right)^{T_k} \left(1 + 4\ell_{f,1}^2\alpha\tilde{\mu}\right) \leq \left(1 - \frac{\beta}{2\tilde{\mu}}\right)^{T_k} \left(1 + \frac{\beta}{4\tilde{\mu}}\right) \leq \left(1 - \frac{\beta}{4\tilde{\mu}}\right). \tag{106}$$

Thus from (106), $\left(1 - \frac{\beta}{2\tilde{\mu}}\right)^{T_k} \leq \left(1 - \frac{\beta}{2\tilde{\mu}}\right)$, and (105), we have

$$\begin{aligned} & f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}^*) - f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) \\ &\leq \left(1 - \frac{\beta}{4\tilde{\mu}}\right) \left(f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_k, \mathbf{y}_k)\right) \\ &\quad + \left(\frac{1}{8\alpha} + L_f - \frac{\beta}{16\alpha\tilde{\mu}} - \frac{L_f\beta}{2\tilde{\mu}}\right) \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &\quad + \left(\frac{\ell_{f,1}}{2}\beta^2 - \frac{\beta}{2}\right) \sum_{j=1}^{T_k} \left(1 - \frac{\beta}{2\tilde{\mu}}\right)^{T_k - j} \|\nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_j^{(k)}, \mathbf{z}_j^{(k)} + \boldsymbol{\nu}_j^{(k)})\|^2 \\ &\quad + \frac{\beta}{2} \sum_{j=1}^{T_k} \left(1 - \frac{\beta}{2\tilde{\mu}}\right)^{T_k - j} \|\nabla_{\mathbf{z}} f(\mathbf{x}_{k+1}, \mathbf{z}_j^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_j^{(k)}, \mathbf{z}_j^{(k)} + \boldsymbol{\nu}_j^{(k)})\|^2 \\ &\leq \left(1 - \frac{\beta}{4\tilde{\mu}}\right) \left(f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_k, \mathbf{y}_k)\right) + \frac{1}{8\alpha}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &\quad + \frac{\beta}{2} \sum_{j=1}^{T_k} \left(1 - \frac{\beta}{2\tilde{\mu}}\right)^{T_k - j} \|\nabla_{\mathbf{z}} f(\mathbf{x}_{k+1}, \mathbf{z}_j^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_j^{(k)}, \mathbf{z}_j^{(k)} + \boldsymbol{\nu}_j^{(k)})\|^2, \end{aligned} \tag{107}$$

where the second inequality follows from $\alpha \leq \min\left\{\frac{\beta}{16L_f\tilde{\mu}}, \frac{\beta}{16\ell_{f,1}^2\tilde{\mu}^2}\right\}$, and $\beta \leq \frac{1}{\ell_{f,1}}$, both of which imply that:

$$\alpha \leq \frac{\beta}{16L_f\tilde{\mu}} \Rightarrow L_f \leq \frac{\beta}{16\alpha\tilde{\mu}}.$$
$$\beta \leq \frac{1}{\ell_{f,1}} \Rightarrow \ell_{f,1} \leq \frac{1}{\beta}.$$

Rearranging (107) and summing over $k = 2, \ldots, K+1$ yields:

$$\sum_{k=2}^{K+1}(\Delta_k - \Delta_{k-1}) \leq -\frac{\beta}{4\tilde{\mu}}\sum_{k=2}^{K+1}\Delta_{k-1} + \frac{1}{8\alpha}\sum_{k=2}^{K+1}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2$$
$$+ \frac{\beta}{2}\sum_{k=2}^{K+1}\sum_{j=1}^{T_k}\left(1 - \frac{\beta}{2\tilde{\mu}}\right)^{T_k - j}\|\nabla_{\mathbf{z}} f(\mathbf{x}_{k+1}, \mathbf{z}_j^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_j^{(k)}, \mathbf{z}_j^{(k)} + \boldsymbol{\nu}_j^{(k)})\|^2.$$

From $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and Assumption 2.3–II., we have

$$\begin{aligned}
&\|\nabla_{\mathbf{z}} f(\mathbf{x}_{k+1}, \mathbf{z}_j^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_j^{(k)}, \mathbf{z}_j^{(k)} + \boldsymbol{\nu}_j^{(k)})\|^2 \\
&\leq 2\|\nabla_{\mathbf{z}} f(\mathbf{x}_{k+1}, \mathbf{z}_j^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_j^{(k)})\|^2 \\
&+ 2\|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_j^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_j^{(k)}, \mathbf{z}_j^{(k)} + \boldsymbol{\nu}_j^{(k)})\|^2 \\
&\leq 2\ell_{f,1}^2\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 + 2\ell_{f,1}^2(\|\boldsymbol{\epsilon}_j^{(k)}\|^2 + \|\boldsymbol{\nu}_j^{(k)}\|^2) \\
&= 2\ell_{f,1}^2\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 + 2\ell_{f,1}^2(\rho^2 + \eta^2),
\end{aligned}$$

where the last equality holds since $\|\boldsymbol{\epsilon}_j^{(k)}\|^2 = \rho^2$ and $\|\boldsymbol{\nu}_j^{(k)}\|^2 = \eta^2$.

Moreover, we have

$$\sum_{j=1}^{T_k}\left(1 - \frac{\beta}{2\tilde{\mu}}\right)^{T_k - j} = \sum_{i=0}^{T_k - 1}\left(1 - \frac{\beta}{2\tilde{\mu}}\right)^{i} \leq \frac{2\tilde{\mu}}{\beta},$$

and

$$\alpha \leq \frac{1}{16\tilde{\mu}\ell_{f,1}^2} \quad \Rightarrow \quad 2\tilde{\mu}\ell_{f,1}^2 \leq \frac{1}{8\alpha}.$$

The above inequalities yield

$$\sum_{k=1}^{K}(\Delta_{k+1} - \Delta_k) \leq -\frac{\beta}{4\tilde{\mu}}\sum_{k=1}^{K}\Delta_k + \frac{1}{4\alpha}\sum_{k=1}^{K}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + \frac{K}{8\alpha}(\rho^2 + \eta^2).$$

From Line 3 of Algorithm 2, where the update rule $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \alpha\,\nabla_{\mathbf{x}} f\big(\mathbf{x}_k + \boldsymbol{\epsilon}_k, \mathbf{y}_{k+1} + \boldsymbol{\nu}_k\big)$ is applied, we obtain the desired result.

□

Lemma I.3. *Suppose Assumptions 2.3–II. and 4.1 hold. Let* $\beta \leq \min\left\{\sqrt{\frac{2\tilde{\mu}}{\ell_{f,1}}}, \frac{1}{\ell_{f,1}}\right\}$. *For the sequence* $\mathbf{y}_k$ *generated by Algorithm 2, we have*

$$d^2_{\mathcal{S}(\mathbf{x}_k)}(\mathbf{y}_{k+1}) \leq \tilde{\mu}\left(1 - \frac{\ell_{f,1}\beta^2}{2\tilde{\mu}}\right)^{T_k}\Delta_k + \frac{2\tilde{\mu}^2}{\beta^2}(\rho^2 + \eta^2),$$

where Δ_k *is defined as in Lemma I.2. Moreover,*

$$\mathcal{S}(\mathbf{x}_k) := \arg\max_{\mathbf{y}\in\mathbb{R}^{d_2}} f(\mathbf{x}_k, \mathbf{y}), \quad \text{and} \quad d^2_{\mathcal{S}}(\mathbf{y}) = \min_{\mathbf{y}'\in\mathcal{S}}\|\mathbf{y} - \mathbf{y}'\|^2.$$

Proof. From Assumption 2.3–II. and $\mathbf{z}_{t+1}^{(k)} = \mathbf{z}_t^{(k)} + \beta \nabla_{\mathbf{z}} f\big(\mathbf{x}_k + \boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)} + \boldsymbol{\nu}_t^{(k)}\big)$, we have

$$\begin{aligned}
-f(\mathbf{x}_k, \mathbf{z}_{t+1}^{(k)}) &= -f(\mathbf{x}_k, \mathbf{z}_t^{(k)}) - \langle \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)}), \mathbf{z}_{t+1}^{(k)} - \mathbf{z}_t^{(k)} \rangle + \frac{\ell_{f,1}}{2} \|\mathbf{z}_{t+1}^{(k)} - \mathbf{z}_t^{(k)}\|^2 \\
&\leq -f(\mathbf{x}_k, \mathbf{z}_t^{(k)}) - \beta \langle \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)}), \nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)} + \boldsymbol{\nu}_t^{(k)}) \rangle \\
&\quad + \frac{\ell_{f,1}\beta^2}{2} \|\nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)} + \boldsymbol{\nu}_t^{(k)})\|^2.
\end{aligned}$$

From $\|a\|^2 = 2\langle a, b\rangle - \|b\|^2 + \|a - b\|^2$, we have

$$\begin{aligned}
&\|\nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)} + \boldsymbol{\nu}_t^{(k)})\|^2 \\
&= 2\langle \nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)} + \boldsymbol{\nu}_t^{(k)}), \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)}) \rangle - \|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 \\
&\quad + \|\nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)} + \boldsymbol{\nu}_t^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2.
\end{aligned}$$

Thus, we get

$$\begin{aligned}
&-f(\mathbf{x}_k, \mathbf{z}_{t+1}^{(k)}) \\
&\leq -f(\mathbf{x}_k, \mathbf{z}_t^{(k)}) - \beta(1 - \ell_{f,1}\beta) \langle \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)}), \nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)} + \boldsymbol{\nu}_t^{(k)}) \rangle \\
&\quad - \frac{\ell_{f,1}\beta^2}{2} \|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 \\
&\quad + \frac{\ell_{f,1}\beta^2}{2} \|\nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)} + \boldsymbol{\nu}_t^{(k)}) - \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2.
\end{aligned}$$

From Lemma I.1, we have

$$-\beta \langle \nabla_{\mathbf{z}} f(\mathbf{x}_k + \boldsymbol{\epsilon}_t^{(k)}, \mathbf{z}_t^{(k)} + \boldsymbol{\nu}_t^{(k)}), \nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)}) \rangle \leq \frac{\ell_{f,1}}{2}(\rho^2 + \eta^2).$$

Thus, from the above inequality and Assumption 2.3–II., we obtain:

$$\begin{aligned}
-f(\mathbf{x}_k, \mathbf{z}_{t+1}^{(k)}) &\leq -f(\mathbf{x}_k, \mathbf{z}_t^{(k)}) + (1 - \ell_{f,1}\beta)\frac{\ell_{f,1}}{2}\left(\rho^2 + \eta^2\right) \\
&\quad - \frac{\ell_{f,1}\beta^2}{2} \|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 + \frac{\ell_{f,1}^3\beta^2}{2} \|(\boldsymbol{\epsilon}_t^{(k)}, \boldsymbol{\nu}_t^{(k)})\|^2 \\
&\leq -f(\mathbf{x}_k, \mathbf{z}_t^{(k)}) + \frac{\ell_{f,1}}{2}(\rho^2 + \eta^2) \\
&\quad - \frac{\ell_{f,1}\beta^2}{2} \|\nabla_{\mathbf{z}} f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\|^2 + \frac{\ell_{f,1}^3\beta^2}{2}(\rho^2 + \eta^2),
\end{aligned}$$

where the second inequality follows from $1 - \ell_{f,1}\beta \leq 1$ and

$$\|(\boldsymbol{\epsilon}_t^{(k)}, \boldsymbol{\nu}_t^{(k)})\|^2 = \|\boldsymbol{\epsilon}_t^{(k)}\|^2 + \|\boldsymbol{\nu}_t^{(k)}\|^2 = (\rho^2 + \eta^2).$$

Let $\mathbf{y}_k^* \in \arg\max_{\mathbf{y} \in \mathbb{R}^{d_2}} f(\mathbf{x}_k, \mathbf{y})$. From Assumption 4.1, we also have

$$\begin{aligned}
f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_k, \mathbf{z}_{t+1}^{(k)}) &\leq f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_k, \mathbf{z}_t^{(k)}) + \frac{\ell_{f,1}}{2}(1 + \ell_{f,1}^2\beta^2)(\rho^2 + \eta^2) \\
&\quad - \frac{\ell_{f,1}\beta^2}{2\tilde{\mu}} \left(f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\right) \\
&\leq \left(1 - \frac{\ell_{f,1}\beta^2}{2\tilde{\mu}}\right) \left(f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_k, \mathbf{z}_t^{(k)})\right) \\
&\quad + \ell_{f,1}(\rho^2 + \eta^2),
\end{aligned}$$

where the second inequality follows from $\beta \leq 1/\ell_{f,1}$.
Iteratively applying the above inequality for $t = 1, \ldots, T_k$ yields

$$\begin{aligned}
f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_k, \mathbf{z}_{T_k+1}^{(k)}) &\leq \left(1 - \frac{\ell_{f,1}\beta^2}{2\tilde{\mu}}\right)^{T_k} \left(f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_k, \mathbf{z}_1^{(k)})\right) \\
&\quad + \ell_{f,1}(\rho^2 + \eta^2) \sum_{j=0}^{T_k - 1} \left(1 - \frac{\ell_{f,1}\beta^2}{2\tilde{\mu}}\right)^j. \qquad (108)
\end{aligned}$$

From Lemma B.1 and Assumption 4.1, we have:

$$\frac{1}{\tilde{\mu}} d^2_{\mathcal{S}(\mathbf{x}_k)}(\mathbf{y}_{k+1}) \leq f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_k, \mathbf{z}^{(k)}_{T_k+1}), \tag{109}$$

where $\mathbf{y}_{k+1} = \mathbf{z}^{(k)}_{T_k+1}$. From (108), (109) and $\mathbf{z}^{(k)}_1 = \mathbf{y}_k$, we get

$$\begin{aligned} d^2_{\mathcal{S}(\mathbf{x}_k)}(\mathbf{y}_{k+1}) \leq \tilde{\mu} \left(1 - \frac{\ell_{f,1}\beta^2}{2\tilde{\mu}}\right)^{T_k} (f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_k, \mathbf{y}_k)) \\ + \tilde{\mu}\ell_{f,1}(\rho^2 + \eta^2) \sum_{j=0}^{T_k-1} \left(1 - \frac{\ell_{f,1}\beta^2}{2\tilde{\mu}}\right)^j. \end{aligned}$$

We have

$$\sum_{j=0}^{T_k-1} \left(1 - \frac{\ell_{f,1}\beta^2}{2\tilde{\mu}}\right)^j \leq \frac{2\tilde{\mu}}{\ell_{f,1}\beta^2},$$

where we used $\sum_{j=0}^{k-1}(1-\gamma)^j \leq 1/\gamma$.

By combining the above inequalities, we obtain the desired result. □

Now, we are ready to proof Theorem 4.3.

I.1 Proof of Theorem 4.3

Proof. Recall that $\mathbf{y}_k^* \in \mathcal{S}(\mathbf{x}_k) = \arg\max_{\mathbf{y}\in\mathbb{R}^{d_2}} f(\mathbf{x}_k, \mathbf{y})$. Since $F(\mathbf{x}_k)$ is L_f-smooth with respect to $\mathbf{x}$ by Lemma C.2, we have

$$\begin{aligned} F(\mathbf{x}_{k+1}) &\leq F(\mathbf{x}_k) + \langle \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k^*), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L_f}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &= F(\mathbf{x}_k) - \alpha \langle \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k^*), \nabla_{\mathbf{x}} f\big(\mathbf{x}_k + \boldsymbol{\epsilon}_k, \mathbf{y}_{k+1} + \boldsymbol{\nu}_k\big)\rangle \\ &\quad + \frac{L_f}{2}\alpha^2 \|\nabla_{\mathbf{x}} f\big(\mathbf{x}_k + \boldsymbol{\epsilon}_k, \mathbf{y}_{k+1} + \boldsymbol{\nu}_k\big)\|^2, \end{aligned}$$

where the equality holds by Line 3 of Algorithm 2.
Then, by the binomial squares $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2$, we know that

$$\begin{aligned} &\langle \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k^*), \nabla_{\mathbf{x}} f\big(\mathbf{x}_k + \boldsymbol{\epsilon}_k, \mathbf{y}_{k+1} + \boldsymbol{\nu}_k\big)\rangle \\ &= \frac{1}{2}\|\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k^*)\|^2 + \frac{1}{2}\|\nabla_{\mathbf{x}} f\big(\mathbf{x}_k + \boldsymbol{\epsilon}_k, \mathbf{y}_{k+1} + \boldsymbol{\nu}_k\big)\|^2 \\ &\quad - \frac{1}{2}\|\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k^*) - \nabla_{\mathbf{x}} f\big(\mathbf{x}_k + \boldsymbol{\epsilon}_k, \mathbf{y}_{k+1} + \boldsymbol{\nu}_k\big)\|^2. \end{aligned}$$

It follows from the inequality above that:

$$\begin{aligned} F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) - \left(\frac{\alpha}{2} - \frac{L_f}{2}\alpha^2\right) \|\nabla_{\mathbf{x}} f\big(\mathbf{x}_k + \boldsymbol{\epsilon}_k, \mathbf{y}_{k+1} + \boldsymbol{\nu}_k\big)\|^2 \\ - \frac{\alpha}{2}\|\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k^*)\|^2 + \frac{\alpha}{2}\|\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k^*) - \nabla_{\mathbf{x}} f\big(\mathbf{x}_k + \boldsymbol{\epsilon}_k, \mathbf{y}_{k+1} + \boldsymbol{\nu}_k\big)\|^2. \end{aligned}$$

From $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we have

$$\begin{aligned} F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) - \left(\frac{\alpha}{2} - \frac{L_f}{2}\alpha^2\right) \|\nabla_{\mathbf{x}} f\big(\mathbf{x}_k + \boldsymbol{\epsilon}_k, \mathbf{y}_{k+1} + \boldsymbol{\nu}_k\big)\|^2 \\ - \frac{\alpha}{2}\|\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k^*)\|^2 + \alpha \underbrace{\|\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k^*) - \nabla_{\mathbf{x}} f\big(\mathbf{x}_k + \boldsymbol{\epsilon}_k, \mathbf{y}_k^* + \boldsymbol{\nu}_k\big)\|^2}_{:=I} \\ + \alpha \underbrace{\|\nabla_{\mathbf{x}} f\big(\mathbf{x}_k + \boldsymbol{\epsilon}_k, \mathbf{y}_k^* + \boldsymbol{\nu}_k\big) - \nabla_{\mathbf{x}} f\big(\mathbf{x}_k + \boldsymbol{\epsilon}_k, \mathbf{y}_{k+1} + \boldsymbol{\nu}_k\big)\|^2}_{:=J}. \end{aligned} \tag{110}$$

Before proceeding, we provide bounds for the terms I, and J, respectively.
From Assumption 2.3–II. and $\|\boldsymbol{\epsilon}_k\| = \rho$ and $\|\boldsymbol{\nu}_k\| = \eta$, we have:

$$I \leq \ell_{f,1}^2 \|(\boldsymbol{\epsilon}_k, \boldsymbol{\nu}_k)\|^2 = \ell_{f,1}^2 \|\boldsymbol{\epsilon}_k\|^2 + \ell_{f,1}^2 \|\boldsymbol{\nu}_k\|^2 = \ell_{f,1}^2(\rho^2 + \eta^2). \tag{111a}$$

From Lemma I.3, we have

$$d^2_{\mathcal{S}(\mathbf{x}_k)}(\mathbf{y}_{k+1}) \leq \tilde{\mu}\left(1 - \frac{\ell_{f,1}\beta^2}{2\tilde{\mu}}\right)^{T_k} \Delta_k + \frac{2\tilde{\mu}^2}{\beta^2}(\rho^2 + \eta^2),$$

where $\Delta_k = f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_k, \mathbf{y}_k)$.
Set

$$T_k \geq -\log_{d_\beta}(4\ell_{f,1}^2\tilde{\mu}^2\frac{\alpha}{\beta}), \quad \text{with} \quad d_\beta = 1 - \frac{\ell_{f,1}\beta^2}{2\tilde{\mu}}.$$

Then, we obtain:

$$d^2_{\mathcal{S}(\mathbf{x}_k)}(\mathbf{y}_{k+1}) \leq \frac{\beta}{4\ell_{f,1}^2\tilde{\mu}\alpha}\Delta_k + \frac{2\tilde{\mu}^2}{\beta^2}(\rho^2 + \eta^2). \tag{111b}$$

From Assumption 2.3–II. and (111b), we have

$$\begin{aligned}\alpha J \leq \alpha\ell_{f,1}^2\|\mathbf{y}_k^* - \mathbf{y}_{k+1}\|^2 &\leq \frac{\beta}{4\tilde{\mu}}\Delta_k + \frac{2\ell_{f,1}^2\tilde{\mu}^2\alpha}{\beta^2}(\rho^2 + \eta^2) \\ &\leq \frac{\beta}{4\tilde{\mu}}\Delta_k + \frac{1}{8\beta}(\rho^2 + \eta^2),\end{aligned} \tag{111c}$$

where the last inequality follows from $\alpha \leq \beta/16\ell_{f,1}^2\tilde{\mu}^2$.
Substituting (111a) and (111c) into (110), we get

$$\begin{aligned}F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) &- \left(\frac{\alpha}{2} - \frac{L_f}{2}\alpha^2\right)\|\nabla_{\mathbf{x}} f\big(\mathbf{x}_k + \boldsymbol{\epsilon}_k, \mathbf{y}_{k+1} + \boldsymbol{\nu}_k\big)\|^2 \\ &- \frac{\alpha}{2}\|\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k^*)\|^2 + (\ell_{f,1}^2\alpha + \frac{1}{8\beta})(\rho^2 + \eta^2) + \frac{\beta}{4\tilde{\mu}}\Delta_k.\end{aligned}$$

Rearranging terms and summing over $k = 1, \ldots, K$ yield:

$$\begin{aligned}\frac{\alpha}{2}\sum_{k=1}^K \|\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k^*)\|^2 \leq &\sum_{k=1}^K (F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})) \\ &- \left(\frac{\alpha}{2} - \frac{L_f}{2}\alpha^2\right)\sum_{k=1}^K \|\nabla_{\mathbf{x}} f\big(\mathbf{x}_k + \boldsymbol{\epsilon}_k, \mathbf{y}_{k+1} + \boldsymbol{\nu}_k\big)\|^2 \\ &+ (\ell_{f,1}^2\alpha + \frac{1}{8\beta})K(\rho^2 + \eta^2) + \frac{\beta}{4\tilde{\mu}}\sum_{k=1}^K \Delta_k.\end{aligned} \tag{112}$$

From Lemma I.2, we have

$$\begin{aligned}\sum_{k=1}^K (\Delta_{k+1} - \Delta_k) \leq -\frac{\beta}{4\tilde{\mu}}\sum_{k=1}^K \Delta_k + \frac{\alpha}{4}\sum_{k=1}^K \|\nabla_{\mathbf{x}} f\big(\mathbf{x}_k + \boldsymbol{\epsilon}_k, \mathbf{y}_{k+1} + \boldsymbol{\nu}_k\big)\|^2 \\ + \frac{K}{8\alpha}(\rho^2 + \eta^2),\end{aligned} \tag{113}$$

where $\Delta_k = f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_k, \mathbf{y}_k)$.
From (113) and (112), we have

$$\begin{aligned}&\frac{\alpha}{2}\sum_{k=1}^K \|\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k^*)\|^2 + \sum_{k=1}^K (\Delta_{k+1} - \Delta_k) \leq \sum_{k=1}^K (F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})) \\ &+ \left(-\frac{\alpha}{2} + \frac{L_f}{2}\alpha^2 + \frac{\alpha}{4}\right)\sum_{k=1}^K \|\nabla_{\mathbf{x}} f\big(\mathbf{x}_k + \boldsymbol{\epsilon}_k, \mathbf{y}_{k+1} + \boldsymbol{\nu}_k\big)\|^2 \\ &+ \left(\alpha\ell_{f,1}^2 + \frac{1}{8\alpha} + \frac{1}{8\beta}\right)K(\rho^2 + \eta^2).\end{aligned}$$

Moreover, from $\alpha \leq 1/2L_f$, we have

$$-\frac{\alpha}{2} + \frac{L_f}{2}\alpha^2 + \frac{\alpha}{4} \leq 0.$$

Let $\Phi := 2\ell_{f,1}^2 + \frac{1}{4\alpha^2} + \frac{1}{4\alpha\beta}$. Then, we have

$$\begin{aligned}
&\sum_{k=1}^{K} \|\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k^*)\|^2 \\
&\leq \frac{2}{\alpha}\left(F(\mathbf{x}_1) - F(\mathbf{x}_{K+1})\right) + \frac{2}{\alpha}\left(\Delta_1 - \Delta_{K+1}\right) + \Phi K(\rho^2 + \eta^2) \\
&\leq \frac{2}{\alpha}\left(F(\mathbf{x}_1) - \inf_{\mathbf{x}} F(\mathbf{x})\right) + \frac{2}{\alpha}\left(\Delta_1 - \Delta_{K+1}\right) + \Phi K(\rho^2 + \eta^2),
\end{aligned}$$

where the second inequality follows from Assumption 4.2.
Lemma C.2 implies that $\nabla_{\mathbf{x}} F(\mathbf{x}) = \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^*)$. Dividing both sides by K, we obtain

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K} \|\nabla F(\mathbf{x})\|^2 \\
&\leq \frac{2}{\alpha K}\left(F(\mathbf{x}_1) - \inf_{\mathbf{x}} F(\mathbf{x})\right) + \frac{2}{\alpha K}\left(\Delta_1 - \Delta_{K+1}\right) + \Phi(\rho^2 + \eta^2).
\end{aligned} \tag{114}$$

From $\Delta_k = f(\mathbf{x}_k, \mathbf{y}_k^*) - f(\mathbf{x}_k, \mathbf{y}_k)$, we have

$$\begin{aligned}
\Delta_1 - \Delta_{K+1} &= f(\mathbf{x}_1, \mathbf{y}_1^*) - f(\mathbf{x}_1, \mathbf{y}_1) - f(\mathbf{x}_{K+1}, \mathbf{y}_{K+1}^*) + f(\mathbf{x}_{K+1}, \mathbf{y}_{K+1}) \\
&\leq f(\mathbf{x}_1, \mathbf{y}_1^*) - f(\mathbf{x}_1, \mathbf{y}_1) - f(\mathbf{x}_{K+1}, \mathbf{y}_{K+1}^*) + f(\mathbf{x}_{K+1}, \mathbf{y}_{K+1}^*).
\end{aligned}$$

Moreover, we have

$$F(\mathbf{x}_1) - \inf_{\mathbf{x}} F(\mathbf{x}) = \max_{\mathbf{y}\in\mathbb{R}^{d_2}} f(\mathbf{x}_1, \mathbf{y}) - \inf_{\mathbf{x}} \max_{\mathbf{y}\in\mathbb{R}^{d_2}} f(\mathbf{x}, \mathbf{y}).$$

Substituting the above observations into (114), we obtain the statement of the theorem. □

I.2 Proof of Theorem D.2

Proof. For the finite-sample dataset $\mathcal{S}$, we have:

$$F(\mathbf{x}; \mathcal{S}) := \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}, \mathbf{y}^*; \xi_i).$$

Let Q denote any posterior distribution that may depend on the training data $\mathcal{S}$, and let P denote the prior distribution. Consider the joint distributions P and Q, defined as

$$P = P_{\mathbf{x}} \times P_{\mathbf{y}^*}, \quad Q = Q_{\mathbf{x}} \times Q_{\mathbf{y}^*}.$$

Here, $\mathbf{x}$ is distributed according to $P_{\mathbf{x}}$ and $\mathbf{y}^*$ according to $P_{\mathbf{y}^*}$. Since $Q = Q_{\mathbf{x}} \times Q_{\mathbf{y}^*}$ and $P = P_{\mathbf{x}} \times P_{\mathbf{y}^*}$, we have:

$$KL(Q||P) = KL(Q_{\mathbf{x}}||P_{\mathbf{x}}) + KL(Q_{\mathbf{y}^*}||P_{\mathbf{y}^*}), \tag{115}$$

where $KL(Q||P)$ is the Kullback–Leibler (KL) divergence. Using the general PAC-Bayesian generalization bound (Alquier et al., 2016), we have for any prior distribution P independent of the training data, the following holds with probability at least $1-\delta$ over the choice of the training set $\mathcal{S}$:

$$\begin{aligned}
\mathbb{E}_{(\mathbf{x},\mathbf{y}^*)\sim Q}[F(\mathbf{x}; \mathcal{D})] &\leq \mathbb{E}_{(\mathbf{x},\mathbf{y}^*)\sim Q}[F(\mathbf{x}; \mathcal{S})] \\
&\quad + \frac{1}{\varpi}\left[KL(Q||P) + \log(\frac{1}{\delta}) + \Psi(\varpi, n)\right],
\end{aligned} \tag{116}$$

for all $\varpi > 0$, where $F(\mathbf{x}; \mathcal{D})$ and $F(\mathbf{x}; \mathcal{S})$ are defined in (35a) and (36a), respectively,

$$\Psi(\varpi, n) = \log \mathbb{E}_P \mathbb{E}_{\mathcal{S}}\left[\exp\{\varpi\left[F(\mathbf{x}; \mathcal{D}) - F(\mathbf{x}; \mathcal{S})\right]\}\right].$$

By the assumption of the theorem, we have $|f| \leq M$. By applying Hoeffding's inequality, we obtain:

$$\Psi(\varpi, n) \leq \frac{\varpi^2 (2M)^2}{8n} = \frac{\varpi^2 M^2}{2n}.$$

The task is to minimize the second term of RHS of (117), we thus choose

$$\varpi = \frac{\sqrt{2n(KL(Q||P) + \log(\frac{1}{\delta}))}}{M}.$$

By this choice, we have

$$\Psi(\varpi, n) \leq \frac{\varpi^2 M^2}{2n} = KL(Q||P) + \log(\frac{1}{\delta}).$$

By invoking this bound, (116) becomes:

$$\mathbb{E}_{(\mathbf{x},\mathbf{y}^*)\sim Q}[F(\mathbf{x};\mathcal{D})] \leq \mathbb{E}_{(\mathbf{x},\mathbf{y}^*)\sim Q}[F(\mathbf{x};\mathcal{S})] + \frac{2}{\varpi}\left[KL(Q||P) + \log(\frac{1}{\delta})\right]. \tag{117}$$

Then, the second term of RHS of (117) is equal to

$$\sqrt{\frac{KL(Q||P) + \log(\frac{1}{\delta})}{n}} \times \sqrt{2}M,$$

which implies

$$\mathbb{E}_{(\mathbf{x},\mathbf{y}^*)\sim Q}[F(\mathbf{x};\mathcal{D})] \leq \mathbb{E}_{(\mathbf{x},\mathbf{y}^*)\sim Q}[F(\mathbf{x};\mathcal{S})] + \sqrt{\frac{KL(Q||P) + \log(\frac{1}{\delta})}{n}} \times \sqrt{2}M. \tag{118}$$

We choose

$$P_{\mathbf{x}} = \mathcal{N}(0, \sigma_{\mathbf{x}}^2 I_{d_1}), \quad P_{\mathbf{y}} = \mathcal{N}(0, \sigma_{\mathbf{y}}^2 I_{d_2}),$$
$$Q_{\mathbf{x}} = \mathcal{N}(\mathbf{x}, \rho^2 I_{d_1}), \quad Q_{\mathbf{y}} = \mathcal{N}(\mathbf{y}, \eta^2 I_{d_2}).$$

The KL divergence between Q and P, when they are Gaussian, is given by formula

$$KL(Q_{\mathbf{x}}||P_{\mathbf{x}}) = \frac{1}{2}\left[\frac{d_1\rho^2 + \|\mathbf{x}\|^2}{\sigma_{\mathbf{x}}^2} - d_1 + d_1 \log(\frac{\sigma_{\mathbf{x}}^2}{\rho^2})\right], \tag{119a}$$

$$KL(Q_{\mathbf{y}^*}||P_{\mathbf{y}^*}) = \frac{1}{2}\left[\frac{d_2\eta^2 + \|\mathbf{y}^*\|^2}{\sigma_{\mathbf{y}^*}^2} - d_2 + d_2 \log(\frac{\sigma_{\mathbf{y}^*}^2}{\eta^2})\right]. \tag{119b}$$

For given posterior distribution $Q_{\mathbf{x}}$ with fixed ρ^2, to minimize the KL term $KL(Q_{\mathbf{x}}||P_{\mathbf{x}})$, the $\sigma_{\mathbf{x}}^2$ should be equal to $\rho^2 + \|\mathbf{x}\|^2/d_1$. In this case, the KL term is no less than $\frac{d_1}{2}\log(1 + \frac{\|\mathbf{x}\|^2}{d_1\rho^2})$. Similarly, for the KL term $KL(Q_{\mathbf{y}^*}||P_{\mathbf{y}^*})$, we have the lower bound $\frac{d_2}{2}\log(1 + \frac{\|\mathbf{y}^*\|^2}{d_2\eta^2})$. Thus, we have

$$KL(Q_{\mathbf{x}}||P_{\mathbf{x}}) + KL(Q_{\mathbf{y}^*}||P_{\mathbf{y}^*}) \geq \frac{1}{2}\left[d_1 \log(1 + \frac{\|\mathbf{x}\|^2}{d_1\rho^2}) + d_2 \log(1 + \frac{\|\mathbf{y}^*\|^2}{d_2\eta^2})\right].$$

When we set

$$\|\mathbf{x}\|^2 > d_1\rho^2(\exp(n/d_1) - 1), \quad \text{and} \quad \|\mathbf{y}^*\|^2 > d_2\eta^2(\exp(n/d_2) - 1),$$

we get

$$KL(Q_{\mathbf{x}}||P_{\mathbf{x}}) + KL(Q_{\mathbf{y}^*}||P_{\mathbf{y}^*}) \geq n.$$

Thus, the second term on the right-hand side is

$$\sqrt{\frac{KL(Q_{\mathbf{x}}||P_{\mathbf{x}}) + KL(Q_{\mathbf{y}^*}||P_{\mathbf{y}^*}) + \log(\frac{1}{\delta})}{n}} \times \sqrt{2}M \geq \sqrt{\frac{2n}{n}}M > M.$$

Hence, for any $\|\mathbf{x}\|^2 > d_1\rho^2(\exp(n/d_1)-1)$ and $\|\mathbf{y}^*\|^2 > d_2\eta^2(\exp(n/d_2)-1)$, the RHS of (38) is greater than the LHS of (38); thus, the inequality holds trivially.

In what follows, we consider only the case:

$$\|\mathbf{x}\|^2 < d_1\rho^2(\exp(n/d_1)-1), \quad \text{and} \quad \|\mathbf{y}^*\|^2 < d_2\eta^2(\exp(n/d_2)-1).$$

Distribution P is Gaussian centered around 0 with variances $\sigma^2_{\mathbf{x}} = \rho^2 + \|\mathbf{x}\|^2/d_1$ and $\sigma^2_{\mathbf{y}^*} = \eta^2 + \|\mathbf{y}^*\|^2/d_2$, which are unknown at the time we set up the inequality, since $\mathbf{x}$ and $\mathbf{y}^*$ are unknown. Meanwhile we have to specify P in advance, since P is the prior distribution. To deal with this problem, we apply the union bound technique (Dziugaite & Roy, 2017; Foret et al., 2020). We set

$$c_{\mathbf{x}} = d_1\rho^2(1+\exp(n/d_1)), \quad \text{and} \quad c_{\mathbf{y}^*} = d_2\eta^2(1+\exp(n/d_2)),$$

then

$$P_{j_{\mathbf{x}}} = \mathcal{N}(0, \sigma^2_{j_{\mathbf{x}}} I_{d_1}), \quad \sigma^2_{j_{\mathbf{x}}} = c_{\mathbf{x}}\exp((1-j_{\mathbf{x}})/d_1),$$
$$P_{j_{\mathbf{y}^*}} = \mathcal{N}(0, \sigma^2_{j_{\mathbf{y}^*}} I_{d_2}), \quad \sigma^2_{j_{\mathbf{y}^*}} = c_{\mathbf{y}^*}\exp((1-j_{\mathbf{y}^*})/d_2),$$

and

$$\mathfrak{B} := \{(P_{j_{\mathbf{x}}}, P_{j_{\mathbf{y}^*}}) : j = 1, 2, \ldots\}.$$

Then the following inequality holds for a particular distribution P_j with probability $1 - \delta_{j_{\mathbf{x}}}\delta_{j_{\mathbf{y}^*}}$, where $\delta_{j_{\mathbf{x}}} = \frac{6\sqrt{\delta}}{\pi^2 j_{\mathbf{x}}^2}$ and $\delta_{j_{\mathbf{y}^*}} = \frac{6\sqrt{\delta}}{\pi^2 j_{\mathbf{y}^*}^2}$,

$$\mathbb{E}_{(\mathbf{x},\mathbf{y}^*)\sim Q}[F(\mathbf{x};\mathcal{D})] \le \mathbb{E}_{(\mathbf{x},\mathbf{y}^*)\sim Q}[F(\mathbf{x};\mathcal{S})] + \frac{1}{\varpi}\left[KL(Q||P_j) + \log(\frac{1}{\delta_{j_{\mathbf{x}}}\delta_{j_{\mathbf{y}^*}}}) + \Psi(\varpi, n)\right].$$

Since $\sum_{j=1}^{\infty}\frac{1}{j^2} = \frac{\pi^2}{6}$, the above inequality holds with every j, with probability $1 - \sum_{j_{\mathbf{x}}=1}^{\infty}\sum_{j_{\mathbf{y}}=1}^{\infty}\frac{36\delta}{\pi^4 j_{\mathbf{x}}^2 j_{\mathbf{y}^*}^2} = 1-\delta$. We set

$$j^*_{\mathbf{x}} := \lfloor 1 - d_1\log(\frac{d_1(\rho^2 + \|\mathbf{x}\|^2/d_1)}{c_{\mathbf{x}}})\rfloor = \lfloor 1 - d_1\log(\frac{(\rho^2 + \|\mathbf{x}\|^2/d_1)}{\rho^2(1+\exp(4n/d_1))})\rfloor,$$
$$j^*_{\mathbf{y}^*} := \lfloor 1 - d_2\log(\frac{d_2(\eta^2 + \|\mathbf{y}^*\|^2/d_2)}{c_{\mathbf{y}}})\rfloor = \lfloor 1 - d_2\log(\frac{(\eta^2 + \|\mathbf{y}^*\|^2/d_2)}{\eta^2(1+\exp(4n/d_2))})\rfloor.$$

Therefore,

$$1 - j^*_{\mathbf{x}} = \lceil d_1\log((\rho^2 + \|\mathbf{x}\|^2/d_1)/c_{\mathbf{x}})\rceil \ge d_1\log((\rho^2 + \|\mathbf{x}\|^2/d_1)/c_{\mathbf{x}})$$
$$\Rightarrow \log(\frac{(\rho^2 + \|\mathbf{x}\|^2/d_1)}{c_{\mathbf{x}}}) \le \frac{1-j^*_{\mathbf{x}}}{d_1} \le \log(\frac{(\rho^2 + \|\mathbf{x}\|^2/d_1)}{c_{\mathbf{x}}}) + \frac{1}{d_1}$$
$$\Rightarrow \rho^2 + \|\mathbf{x}\|^2/d_1 \le c_{\mathbf{x}}\exp(\frac{1-j^*_{\mathbf{x}}}{d_1}) \le \exp(1/d_1)(\rho^2 + \|\mathbf{x}\|^2/d_1)$$
$$\Rightarrow \rho^2 + \|\mathbf{x}\|^2/d_1 \le \sigma^2_{j^*_{\mathbf{x}}} \le \exp(1/d_1)(\rho^2 + \|\mathbf{x}\|^2/d_1).$$

Thus, the KL term could be bounded as follow

$$\begin{aligned} KL(Q_{\mathbf{x}}||P_{j^*_{\mathbf{x}}}) &= \frac{1}{2}\left[\frac{d_1\rho^2 + \|\mathbf{x}\|^2}{\sigma^2_{j^*_{\mathbf{x}}}} - d_1 + d_1\log(\frac{\sigma^2_{j^*_{\mathbf{x}}}}{\rho^2})\right], \\ &\le \frac{1}{2}\left[\frac{d_1\rho^2 + \|\mathbf{x}\|^2}{\rho^2 + \|\mathbf{x}\|^2/d_1} - d_1 + d_1\log(\frac{\exp(1/d_1)(\rho^2 + \|\mathbf{x}\|^2/d_1)}{\rho^2})\right] \\ &= \frac{1}{2}\left[d_1\log(\frac{\exp(1/d_1)(\rho^2 + \|\mathbf{x}\|^2/d_1)}{\rho^2})\right] \\ &= \frac{1}{2}\left[1 + d_1\log(1 + \frac{\|\mathbf{x}\|^2}{d_1\rho^2})\right]. \end{aligned} \tag{120a}$$

Applying the same argument, we obtain:

$$KL(Q_{\mathbf{y}^*}||P_{j^*_{\mathbf{y}^*}}) \le \frac{1}{2}\left[1 + d_2 \log(1 + \frac{\|\mathbf{y}^*\|^2}{d_2\eta^2})\right]. \tag{120b}$$

From (115), (120a), and (120b), we obtain

$$\begin{aligned} KL(Q||P_{j^*}) &\le \frac{1}{2}\left[2 + d_1 \log(1 + \frac{\|\mathbf{x}\|^2}{d_1\rho^2}) + d_2 \log(1 + \frac{\|\mathbf{y}^*\|^2}{d_2\eta^2})\right] \\ &= \frac{1}{2}\left[2 + \log(1 + \frac{\|\mathbf{x}\|^2}{d_1\rho^2})^{d_1}(1 + \frac{\|\mathbf{y}^*\|^2}{d_2\eta^2})^{d_2}\right]. \end{aligned} \tag{121}$$

For the term $\log \frac{1}{\delta_{j^*_{\mathbf{x}}}\delta_{j^*_{\mathbf{y}^*}}}$, with recall that

$$c_{\mathbf{x}} = d_1\rho^2(1 + \exp(n/d_1)), \quad \text{and} \quad j^*_{\mathbf{x}} = \lfloor 1 - d_1 \log(\frac{(\rho^2 + \|\mathbf{x}\|^2/d_1)}{\rho^2(1 + \exp(4n/d_1))})\rfloor,$$

we have:

$$\begin{aligned} \log \frac{1}{\delta_{j^*_{\mathbf{x}}}\delta_{j^*_{\mathbf{y}^*}}} &= \log \frac{j^{*2}_{\mathbf{y}^*} j^{*2}_{\mathbf{x}} \pi^2}{6\delta} \\ &= \log \frac{1}{\delta} + 2\log \frac{\pi^2}{6} + 2\log(j^*_{\mathbf{x}}) + 2\log(j^*_{\mathbf{y}^*}) \\ &\le \log \frac{1}{\delta} + 2\log \frac{\pi^2}{6} + 2\log\left(1 + d_1 \log(\frac{\rho^2(1 + \exp(n/d_1))}{(\rho^2 + \|\mathbf{x}\|^2/d_1)})\right) \\ &\qquad + 2\log\left(1 + d_2 \log(\frac{\eta^2(1 + \exp(n/d_2))}{(\eta^2 + \|\mathbf{y}^*\|^2/d_2)})\right) \\ &\le \log \frac{1}{\delta} + 2\log \frac{\pi^2}{6} + 2\log\left(1 + d_1 \log(1 + \exp(n/d_1))\right) \\ &\qquad + 2\log\left(1 + d_2 \log(1 + \exp(n/d_2))\right), \end{aligned}$$

which implies

$$\begin{aligned} \log \frac{1}{\delta_{j^*_{\mathbf{x}}}\delta_{j^*_{\mathbf{y}^*}}} &\le \log \frac{1}{\delta} + 2\log \frac{\pi^2}{6} + 2\log\left(1 + d_1(1 + \frac{n}{d_1})\right) \\ &\qquad + 2\log\left(1 + d_2(1 + \frac{n}{d_2})\right) \\ &= \log \frac{1}{\delta} + \log(\frac{\pi^2}{6})^2(1 + d_1 + n)^2(1 + d_2 + n)^2. \end{aligned} \tag{122}$$

Substituting (122) and (121) into (118), we obtain

$$\begin{aligned} &\mathbb{E}_{(\mathbf{x},\mathbf{y}^*)\sim Q}[F(\mathbf{x};\mathcal{D})] \\ &\le \mathbb{E}_{(\mathbf{x},\mathbf{y}^*)\sim Q}[F(\mathbf{x};\mathcal{S})] + \sqrt{2}M\sqrt{\frac{KL(Q||P_{j^*}) + \log(\frac{1}{\delta_{j^*_{\mathbf{x}}}\delta_{j^*_{\mathbf{y}^*}}})}{n}} \\ &\le \mathbb{E}_{(\mathbf{x},\mathbf{y}^*)\sim Q}[F(\mathbf{x};\mathcal{S})] + \frac{\sqrt{2}M}{\sqrt{n}}\left(\log(1 + \frac{\|\mathbf{x}\|^2}{d_1\rho^2})^{\frac{d_1}{2}}(1 + \frac{\|\mathbf{y}^*\|^2}{d_2\eta^2})^{\frac{d_2}{2}} + B\right)^{1/2}, \end{aligned}$$

where $B := 1 + \log(\frac{1}{\delta}) + \log(\frac{\pi^2}{6})^2(1 + d_1 + n)^2(1 + d_2 + n)^2$.

Fix $\rho'^2 = \frac{\rho^2}{d_1\left(1+\sqrt{\frac{\log(n)}{d_1}}\right)^2}$, and $\eta'^2 = \frac{\eta^2}{d_2\left(1+\sqrt{\frac{\log(n)}{d_2}}\right)^2}$, and consider perturbations $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \rho'^2\mathbf{I}_{d_1})$ and $\boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \eta'^2\mathbf{I}_{d_2})$. Thus, we have

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\rho'^2 I_{d_1}),\boldsymbol{\nu}\sim\mathcal{N}(0,\eta'^2 I_{d_2})}[F(\mathbf{x},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{D})] \\ &\le \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\rho'^2 I_{d_1}),\boldsymbol{\nu}\sim\mathcal{N}(0,\eta'^2 I_{d_2})}[F(\mathbf{x},\boldsymbol{\epsilon},\boldsymbol{\nu};\mathcal{S})] \\ &+ \frac{\sqrt{2}M}{\sqrt{n}}\left(\log(1 + \frac{\|\mathbf{x}\|^2}{d_1\rho^2})^{\frac{d_1}{2}}(1 + \frac{\|\mathbf{y}^*\|^2}{d_2\eta^2})^{\frac{d_2}{2}} + B\right)^{1/2}. \end{aligned} \tag{123}$$

Bounding the empirical expectation: The next step is to bound the expectation in the RHS of (123) by the worst-case loss. We make use of the following result from (Laurent & Massart, 2000):

$$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \rho'^2 I_{d_1}) \Rightarrow \|\boldsymbol{\epsilon}\|^2 \leq \rho'^2 d_1 \left(1 + \sqrt{\frac{\log(n)}{d_1}}\right)^2 = \rho^2, \tag{124a}$$

with probability at least $1 - 1/\sqrt{n}$ and

$$\boldsymbol{\nu} \sim \mathcal{N}(0, \eta'^2 I_{d_2}) \Rightarrow \|\boldsymbol{\nu}\|^2 \leq \eta'^2 d_2 \left(1 + \sqrt{\frac{\log(n)}{d_2}}\right)^2 = \eta^2, \tag{124b}$$

with probability at least $1 - 1/\sqrt{n}$.

The above inequalities allow us to bound the expectation in terms of the maximum value over a set of bounded perturbations. To this end, let

$$\boldsymbol{\zeta} := (\boldsymbol{\epsilon}, \boldsymbol{\nu}) \sim \mathcal{N}\big(0, \Sigma\big), \quad \Sigma := \operatorname{diag}\big(\rho^2 I_{d_1}, \eta^2 I_{d_2}\big).$$

Thus, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{N}\left(0,\Sigma\right)}[F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{S})] \\
&\leq \mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{N}\left(0,\Sigma\right)}[F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{S}) \mid \|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| \leq \eta]\ \mathbb{P}(\|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| \leq \eta) \\
&+ \mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{N}\left(0,\Sigma\right)}[F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{S}) \mid \|\boldsymbol{\epsilon}\| > \rho, \|\boldsymbol{\nu}\| \leq \eta]\ \mathbb{P}(\|\boldsymbol{\epsilon}\| > \rho, \|\boldsymbol{\nu}\| \leq \eta) \\
&+ \mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{N}\left(0,\Sigma\right)}[F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{S}) \mid \|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| > \eta]\ \mathbb{P}(\|\boldsymbol{\epsilon}\| \leq \rho, \|\boldsymbol{\nu}\| > \eta) \\
&+ \mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{N}\left(0,\Sigma\right)}[F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{S}) \mid \|\boldsymbol{\epsilon}\| > \rho, \|\boldsymbol{\nu}\| > \eta]\ \mathbb{P}(\|\boldsymbol{\epsilon}\| > \rho, \|\boldsymbol{\nu}\| > \eta) \\
&\leq (1 - \frac{1}{\sqrt{n}})^2 \max_{\|\boldsymbol{\epsilon}\|\leq\rho, \|\boldsymbol{\nu}\|\leq\eta} F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{S}) + \big(1 - (1 - \frac{1}{\sqrt{n}})^2\big) M,
\end{aligned}$$

which implies

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\zeta}\sim\mathcal{N}\left(0,\Sigma\right)}[F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{S})] \\
&\leq \max_{\|\boldsymbol{\epsilon}\|\leq\rho, \|\boldsymbol{\nu}\|\leq\eta} F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{S}) + \big(1 - (1 - \frac{1}{\sqrt{n}})^2\big)\big(M - \max_{\|\boldsymbol{\epsilon}\|\leq\rho, \|\boldsymbol{\nu}\|\leq\eta} F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{S})\big) \\
&\leq \max_{\|\boldsymbol{\epsilon}\|\leq\rho, \|\boldsymbol{\nu}\|\leq\eta} F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{S}) + 2M\big(1 - (1 - \frac{1}{\sqrt{n}})^2\big) \\
&\leq \max_{\|\boldsymbol{\epsilon}\|\leq\rho, \|\boldsymbol{\nu}\|\leq\eta} F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{S}) + \frac{4M}{\sqrt{n}},
\end{aligned} \tag{125}$$

where the second inequality follows from the assumption $|f| \leq M$, and the last inequality holds since we have

$$1 - (1 - \frac{1}{\sqrt{n}})^2 = \frac{2}{\sqrt{n}} - \frac{1}{n} \leq \frac{2}{\sqrt{n}}.$$

From (125), we obtain

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\rho^2 I_{d_1}), \boldsymbol{\nu}\sim\mathcal{N}(0,\eta^2 I_{d_2})}[F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{S})] \\
&\leq \max_{\|\boldsymbol{\epsilon}\|\leq\rho, \|\boldsymbol{\nu}\|\leq\eta} F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{S}) + \frac{4M}{\sqrt{n}}.
\end{aligned} \tag{126}$$

Using assumption (37), we have

$$F(\mathbf{x}; \mathcal{D}) \leq \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\rho^2 I_{d_1}), \boldsymbol{\nu}\sim\mathcal{N}(0,\eta^2 I_{d_2})}[F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{D})]. \tag{127}$$

Substituting (126) and (127) into (123) yields

$$\begin{aligned}
F(\mathbf{x}; \mathcal{D}) \leq &\max_{\|\boldsymbol{\epsilon}\|\leq\rho, \|\boldsymbol{\nu}\|\leq\eta} F(\mathbf{x}, \boldsymbol{\epsilon}, \boldsymbol{\nu}; \mathcal{S}) + \frac{4M}{\sqrt{n}} \\
&+ \frac{\sqrt{2}M}{\sqrt{n}} \left(\log(1 + \frac{\|\mathbf{x}\|^2}{d_1 \rho'^2})^{\frac{d_1}{2}} (1 + \frac{\|\mathbf{y}^*\|^2}{d_2 \eta'^2})^{\frac{d_2}{2}} + B\right)^{1/2}.
\end{aligned}$$

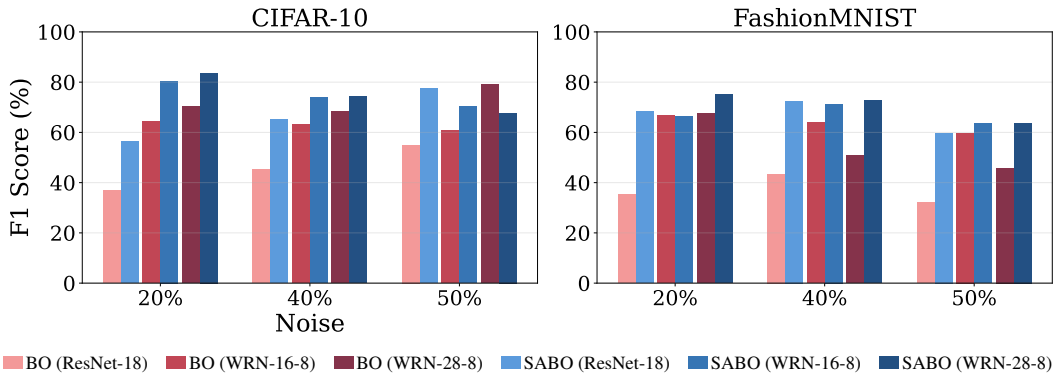

Figure 4: F1 score (data cleaner quality) at the best validation checkpoint across noise rates. SABO (blue) outperforms BO (red) across all settings, most notably on ResNet-18, where sharpness-aware optimization compensates for limited capacity.

From (124), we have

$$\log(1+\frac{\|\mathbf{x}\|^2}{d_1\rho'^2})^{\frac{d_1}{2}}(1+\frac{\|\mathbf{y}^*\|^2}{d_2\eta'^2})^{\frac{d_2}{2}}$$
$$=\log\left(1+\frac{\|\mathbf{x}\|^2}{\rho^2}\left(1+\sqrt{\frac{\log(n)}{d_1}}\right)^2\right)^{\frac{d_1}{2}}\left(1+\frac{\|\mathbf{y}^*\|^2}{\eta^2}\left(1+\sqrt{\frac{\log(n)}{d_2}}\right)^2\right)^{\frac{d_2}{2}},$$

which completes the proof. □

J Experimental Details

This appendix collects the settings for the two large-scale tasks of Section 5, together with the full hypercleaning results.

J.1 Data Hypercleaning

The dataset is randomly partitioned into three non-overlapping subsets: $\mathcal{D}_{tr}$ for training, $\mathcal{D}_{val}$ for validation, and $\mathcal{D}_{test}$ for testing. A predetermined fraction of the training samples in $\mathcal{D}_{tr}$ is then randomly corrupted; samples in the validation and test sets are not corrupted. The hyperparameter $\mathbf{x}$ is a trainable vector, each element of which corresponds to one training sample and so identifies noisy labels.

Setup. We evaluate on the CIFAR-10 (Krizhevsky & Hinton, 2009) and FashionMNIST (Xiao et al., 2017) datasets using three architectures of increasing capacity: ResNet-18 (He et al., 2016), WideResNet (WRN)-16-8 (Zagoruyko & Komodakis, 2016), and WRN-28-8 (Zagoruyko & Komodakis, 2016). The CIFAR-10 dataset contains 50k images with 10 classes. Each image is 32×32 with 3 channels for the RGB values. All experiments employ SGD with a momentum of 0.9, weight decay 5×10^{-4}, batch size 128, and 100 epochs. The inner learning rate follows a cosine schedule from 0.1 to 0, while the outer learning rate is fixed at 0.05—consistent with the bilevel optimization literature where the outer (hyperparameter) learning rate remains constant (Franceschi et al., 2018; Grazzi et al., 2020). The penalty coefficient γ is linearly annealed from 0 to 2.0 over the first 1000 iterations. Label smoothing (0.1) is applied only to the inner (training) loss, while the outer (validation) loss uses the standard cross-entropy loss. For SABO, the perturbation radii are set to $\rho = 0.1$ and $\eta = 0.1$.

Following Li et al. (2020), we report test accuracy and the F1 score at the best validation checkpoint. Because the F1 score directly measures the quality of the learned data cleaner, specifically, its ability to correctly separate clean samples from corrupted ones, it serves as our primary evaluation metric for bilevel optimization under label noise.

Table 1: Data Hypercleaning results at 20%, 40%, and 50% label noise. We report test accuracy and F1 score at the best validation checkpoint (100 epochs, cosine inner LR with $\eta_{\min} = 10^{-3}$, fixed outer LR). Best in each pair shown in **bold**.

		CIFAR-10				**FashionMNIST**			
		Test Acc (%)		F1 (%)		Test Acc (%)		F1 (%)	
Noise	**Model**	BO	SABO	BO	SABO	BO	SABO	BO	SABO
20%	ResNet-18	80.36	**80.50**	36.75	**56.40**	90.59	**91.96**	35.37	**68.18**
	WRN-16-8	**84.71**	84.68	64.37	**80.50**	89.78	**91.39**	**66.86**	66.26
	WRN-28-8	85.79	**86.14**	70.20	**83.64**	90.42	**91.46**	67.47	**74.98**
40%	ResNet-18	69.49	**74.97**	45.16	**65.22**	89.50	**90.70**	43.39	**72.49**
	WRN-16-8	78.70	**81.91**	63.18	**73.90**	89.27	**90.39**	63.85	**71.30**
	WRN-28-8	79.03	**81.49**	68.23	**74.23**	89.38	**90.94**	50.88	**72.77**
50%	ResNet-18	63.21	**70.54**	54.95	**77.36**	88.55	**90.46**	32.10	**59.73**
	WRN-16-8	**75.49**	74.66	60.68	**70.39**	88.75	**90.33**	59.74	**63.47**
	WRN-28-8	72.76	**76.70**	**78.94**	67.75	89.66	**90.44**	45.84	**63.69**

Figure 4 and Table 1 give the full hypercleaning results. Across all noise regimes SABO wins 16 of 18 accuracy comparisons and 16 of 18 F1 comparisons, by margins that widen at higher noise rates and on smaller architectures: at 50% noise it lifts F1 by 22.4 points on CIFAR-10 (ResNet-18) and 27.6 points on FashionMNIST (ResNet-18). Flatter optima thus act as a geometric compensation for limited model capacity.

J.2 GAN Training

We evaluate on CIFAR-10, CIFAR-100 (Krizhevsky & Hinton, 2009), and FashionMNIST using DCGAN-style architectures of varying capacity: DCGAN-64 (64 base features) and DCGAN-128 (128 base features), comparing SAMO against standard minimax optimization (MO). All runs use Adam with learning rate 10^{-4}, $\beta_1 = 0.0$, $\beta_2 = 0.999$, batch size 128, and 200 epochs with fixed learning rate. The choice of $\beta_1 = 0.0$ follows modern GAN practice (Brock et al., 2018; Miyato et al., 2018), which removes first-moment bias and provides a more challenging optimization landscape. For SAMO, we use asymmetric perturbation radii $\rho_G = 0.1$ and $\rho_D = 0.01$, reflecting the asymmetric roles of generator and discriminator.

To evaluate performance, we report the Fréchet Inception Distance (FID; lower is better) as a measure of distributional similarity, and the Inception Score (IS; higher is better) for sample quality and diversity.

K Which Radius Does the Work? Four Regimes

Section 5.1 exhibits one instance on which the two radii are complementary: each alone fails and the pair succeeds. Whether that is typical depends on where the sharpness sits, because the two radii act through different channels. The outer radius ρ acts on the outer landscape V, and it removes a sharp region only when that region is narrower than the ball. The inner radius η reaches $\mathbf{x}$ only through the penalty part of $\nabla_{\mathbf{x}}\mathcal{L}_\gamma$, which carries $\frac{1}{2}\nabla_{\mathbf{x}} s\,(\|\mathbf{d}\|^2 + 2\eta\|\mathbf{d}\|)$, so it moves the outer iterate only where the inner curvature $s(\mathbf{x})$ varies. Varying these two ingredients independently gives four regimes, shown in Figure 5.

Construction. All four instances have the form (21)–(22), with the same a, κ, λ, $\mathbf{q}$, well centers and ridge direction as Section 5.1 (Appendix L); the ramp is written $s(\mathbf{x}) = s_0 + A\left(1 + e^{(t(\mathbf{x})-t_0)/w_s}\right)^{-1}$, so that t_0 shifts it along the axis joining the minima. They differ only in the parameters of Table 2. Every run uses the updates (14)–(15) with $\alpha = 0.015$, $\gamma = 1.2$ and $K = 4000$: the outer perturbation enters f only, and the penalty gradient is evaluated at $\mathbf{x}_k$. The two ball maximizations are carried out exactly, over $\boldsymbol{\nu}$ in closed form and over $\boldsymbol{\epsilon}$ by search over 16 points of the

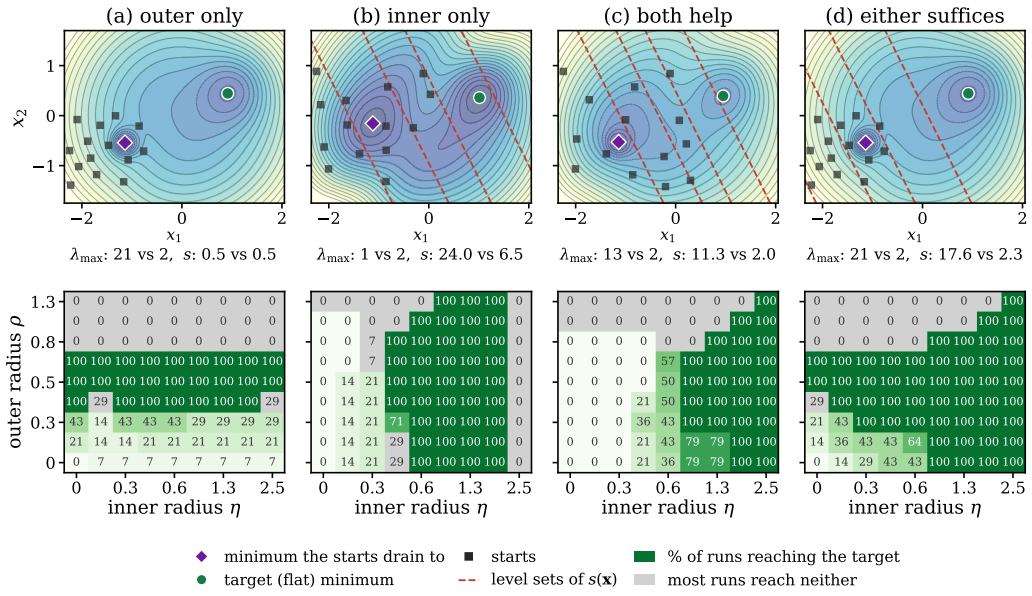

Figure 5: Where the sharpness sits decides which radius does the work. Top: the outer landscape V (fill) of each instance, with level sets of the inner curvature $s(\mathbf{x})$ (dashed; absent in (a), where s is constant), the fourteen starts, the minimum they drain to (diamond) and the target minimum (circle); the line under each panel gives $\lambda_{\max}$ of V and s at those two minima. Bottom: percentage of the fourteen runs that reach the target under the update rule (14)–(15), over the outer radius ρ (rows) and the inner radius η (columns); grey cells are those where most runs reach neither minimum. The bottom row of each map is η alone and the left column ρ alone.

circle, since the one-step rule stalls at large radii (Appendix L). Every instance has its own fourteen initializations, chosen by the same rule: spread over the region and each sent to the sharp minimum by the unperturbed method. The parameters of the three new instances were set by a coarse search so that each regime is clean; the 9×9 grids of Figure 5 were then run once.

(a) Outer sharpness only. The sharp minimum is a narrow well, $\lambda_{\max} = 20.9$ against 1.8, and the inner problem has the same curvature everywhere. The outer radius alone reaches the flat minimum in every run for $\rho \in [0.4, 0.65]$, once the ball is wide enough to remove the spike from the perturbed landscape; from $\rho = 0.8$ on, every run reaches neither minimum. The inner radius does nothing: η alone reaches the flat minimum in at most one run of fourteen at any radius up to 2.5, and each row of the map is nearly constant in η.

(b) Inner sharpness only. Here the outer landscape gives no reason to move. The well the starts drain to is not sharper than the target but slightly flatter ($\lambda_{\max} = 1.4$ against 2.3) and wider; what distinguishes the two minima is the inner curvature, 24.0 against 6.5. The outer radius alone never reaches the target, at any radius up to 1.3. The inner radius alone reaches it in every run for $\eta \in [0.6, 1.8]$, and a moderate outer radius lowers that threshold: at $\eta = 0.45$, 4 runs of 14 succeed alone and all 14 with $\rho \in [0.4, 0.8]$. This is the inner channel in isolation: η moves $\mathbf{x}$ down the gradient of s, towards the better-conditioned inner problem, even though the target is the sharper of the two outer minima.

(c) Both help. This is the instance of Section 5.1. The spike sits inside a wide bowl, so no ρ alone escapes: 0% at every radius. The ramp is flat at the spike, so the inner radius alone reaches the flat minimum in 11 runs of 14 at $\eta \in \{0.9, 1.3\}$ and in every run only from $\eta = 1.8$. The outer radius halves that requirement: with $\rho \in [0.3, 0.65]$, $\eta = 0.9$ succeeds in every run. It cannot replace the inner radius, however: at $\eta \leq 0.3$ no pair succeeds.

Table 2: The four instances. d_{b}, d_{p}, s_{p}: depth of the bowl and depth and width of the well at $\mathbf{c}_{\mathrm{s}}$; b: ridge height; A, w_s, t_0: height, width and position of the curvature ramp. The last four columns give the largest outer curvature $\lambda_{\max}$ of V and the inner curvature s at the minimum the starts drain to and at the target minimum. Instance (c) is that of Section 5.1.

								$\lambda_{\max}$		s	
Regime	d_{b}	d_{p}	s_{p}	b	A	w_s	t_0	start	target	start	target
(a) outer only	0	0.60	0.17	0.05	0	–	–	20.9	1.8	0.5	0.5
(b) inner only	0	0.40	0.75	0.30	30	0.80	0	1.4	2.3	24.0	6.5
(c) both help	0.45	0.35	0.17	0.16	12	0.55	0	12.9	2.0	11.3	2.0
(d) either suffices	0	0.60	0.17	0.05	35	0.80	-1.25	20.9	1.8	17.6	2.3

(d) Either suffices. Instance (d) has the outer landscape of (a) and a steep ramp placed across the spike, so both channels are open at once. The outer radius alone succeeds for $\rho \in [0.5, 0.65]$, as in (a), and the inner radius alone now succeeds for every $\eta \geq 0.9$, where in (a) it reached the target in at most one run. The only change between (a) and (d) is the inner curvature s, so this comparison isolates the inner channel: adding a gradient of s at the sharp minimum turns η from useless into sufficient. With both channels open the two radii also combine: $\rho = 0.4$, $\eta = 0.15$ and $\rho = \eta = 0.3$ succeed in every run, although each radius alone succeeds in at most 4 runs of 14 there. In all, 48 of the 81 cells succeed in every run, against 25 in (a), 39 in (b) and 26 in (c).

What the four regimes show. No single radius works across the four instances. The outer radius alone never reaches the target in (b) or (c), and the inner radius alone reaches it in at most one run of fourteen in (a). The pair $\rho = 0.4$, $\eta = 0.9$, in contrast, reaches the target in every run of all four instances. In practice one does not know whether sharpness sits in the outer landscape, in the conditioning of the inner problem, or in both, and the two radii read these separately; tuning both is therefore the robust default. The grey regions record the opposite failure: when a radius is too large, the iterates stop at neither minimum, the behavior the $\mathcal{O}(\rho^2 + \eta^2)$ floor of Theorem 3.1 anticipates.

L Additional Synthetic Experiments

This appendix stays with the instance of Section 5.1 and takes up three questions the main text leaves open: how far the benefit of the pair reaches beyond the fourteen initialisations shown there, what each radius does to the sharpness of each level, and why the two radii need each other. It closes with a controlled version of the hypercleaning task of Section 5.2.

The instance of Section 5.1 uses $a = 0.13$ with a curved ridge, $b = 0.16$, $\mathbf{u} = (0.8, -0.6)$, $\tau = -0.15$, $w = 0.45$. Two of its three wells are concentric at $\mathbf{c}_{\mathrm{b}} = \mathbf{c}_{\mathrm{p}} = (-1.15, -0.55)$: a wide shallow bowl, $d_{\mathrm{b}} = 0.45$, $s_{\mathrm{b}} = 0.85$, and a narrow spike inside it, $d_{\mathrm{p}} = 0.35$, $s_{\mathrm{p}} = 0.17$; the third is the flat well, $\mathbf{c}_{\mathrm{f}} = (1.05, 0.55)$, $d_{\mathrm{f}} = 0.35$, $s_{\mathrm{f}} = 0.48$. Its curvature ramp runs along $\mathbf{u}_0 = (\mathbf{c}_{\mathrm{f}} - \mathbf{c}_{\mathrm{p}})/\|\mathbf{c}_{\mathrm{f}} - \mathbf{c}_{\mathrm{p}}\|$, coordinatised by $t(\mathbf{x}) = \langle \mathbf{u}_0, \mathbf{x} - \frac{1}{2}(\mathbf{c}_{\mathrm{p}} + \mathbf{c}_{\mathrm{f}}) \rangle$. Its inner problem (22) uses $\kappa = 0.25$, $s_0 = 0.5$, $A = 12$, $w_s = 0.55$, and its outer objective $\lambda = 0.1$, $\mathbf{q} = (0, 1)$.

The three properties of (22) quoted in Section 5.1 are as follows. The inner solution set $\mathcal{S}(\mathbf{x}) = \{\kappa \mathbf{x}\}$ is a singleton, so η cannot act by selecting among inner solutions. The bilevel objective is $\Phi(\mathbf{x}) = f(\mathbf{x}, \kappa \mathbf{x}) = V(\mathbf{x}) + \lambda \|\mathbf{q}\|^2$, which does not contain s at all. And the perturbed inner problem is still minimised at $\mathbf{y} = \kappa \mathbf{x}$, so the perturbed outer objective is $V(\mathbf{x}) + \lambda(\|\mathbf{q}\| + \eta)^2$. The drift the inner radius exerts is therefore not uniform: its magnitude is $\gamma \|\mathbf{d}(\mathbf{x})\| \|\nabla_{\mathbf{x}} s(\mathbf{x})\|$ per unit η, and since $\|\mathbf{d}\| = 2\lambda \|\mathbf{q}\|/(2\lambda + \gamma s)$ grows as s falls, it grows by roughly a factor of six from the sharp minimum to the flat one.

All runs below use Algorithm 1 as written: the outer perturbation enters f only, as in (14), and the penalty is evaluated at the unperturbed point, as in (15). The maximisation over the outer ball is carried out by search over sixteen points of the circle, with $\alpha = 0.015$, $\gamma = 1.2$ and $K = 4000$.

The whole domain. Figure 6 starts the method from every point of a 62×48 grid covering the domain rather than from fourteen chosen points. Without perturbation the flat minimum drains about a third of the domain and the sharp region the rest. The outer radius on its own does not help: it

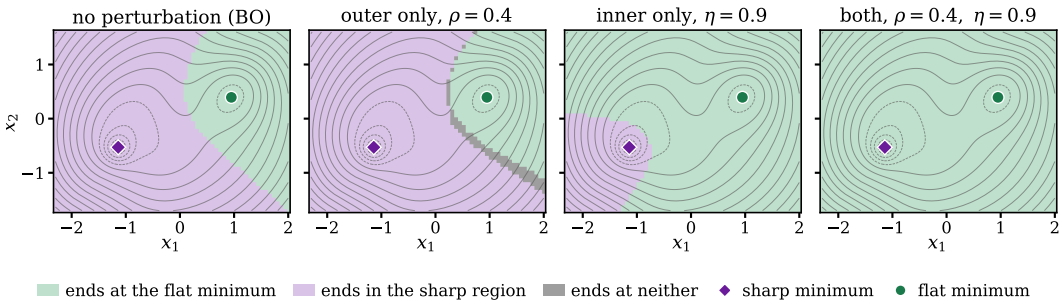

Figure 6: Where each start ends on instance (22), over a 62×48 grid of initialisations covering the domain, without perturbation, with each radius alone and with both. Contours are those of V.

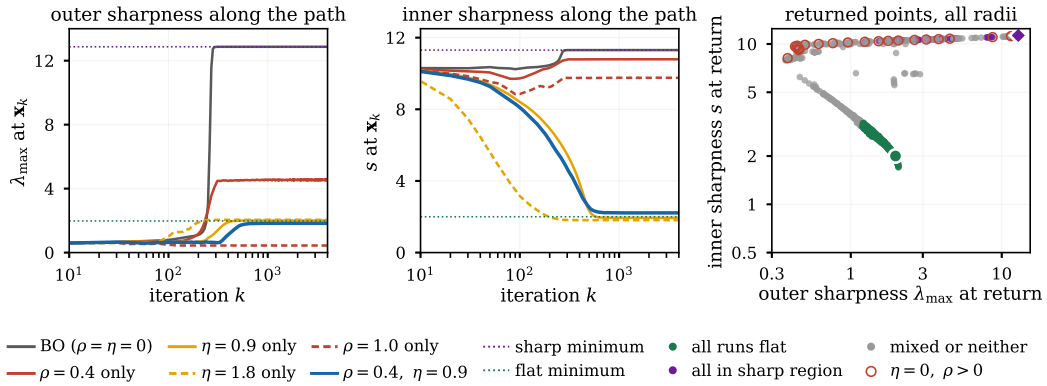

Figure 7: Sharpness at the two levels on instance (22), from the fourteen initialisations of Figure 1. Left and middle: the outer curvature $\lambda_{\max}(\nabla^2 V)$ and the inner curvature s at the iterate (medians over starts), with their values at the two minima dotted. Right: both at the returned point for every cell of the grid of Figure 8; circled points use the outer radius alone.

shrinks the flat basin slightly, since it lifts the flat well as well as the sharp one, and leaves a thin band of starts along the ridge at neither minimum. The inner radius on its own enlarges the flat basin to most of the domain but leaves a region around the sharp minimum, where the ramp of s is too flat for the drift to act. With both radii every start on the grid ends at the flat minimum, so the pair does not rescue a favourable set of initialisations but removes every other attractor from the domain.

Sharpness at each level. The instance has two sharpnesses that can be read off exactly at any point: the outer one, $\lambda_{\max}(\nabla^2 V(\mathbf{x}))$, and the inner one, the curvature $s(\mathbf{x})$ of $g(\mathbf{x}, \cdot)$. Figure 7 follows both along the paths. Without perturbation the iterates fall into the spike and the outer curvature jumps to that of the sharp minimum. The outer radius keeps them out of the spike but not out of the sharp region: at $\rho = 0.4$ they stop on its shoulder, and at $\rho = 1.0$ on the slope of the bowl, at an outer curvature below that of the flat minimum. The inner radius acts on the other axis, and s falls only while $\mathbf{x}$ moves down the ramp. The two quantities are genuinely separate, and one of them can mislead. With the outer radius alone the returned point can be flatter in the outer sense than the flat minimum while lying away from every minimum of V, and its inner curvature stays about five times that of the flat minimum at every radius tried. The right panel collects the returned points over the whole grid of radii of Figure 8: the runs that use the outer radius alone sit along the top of the plot, flat in the outer sense and sharp in the inner one, and the only point small on both axes is the flat minimum. A certificate of flatness at one level is not a certificate at the other, and a bilevel problem needs both radii because it has both sharpnesses.

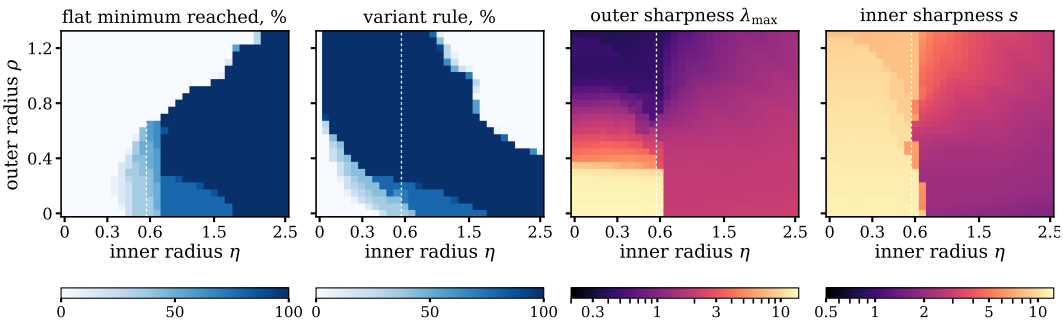

Figure 8: The (η, ρ) plane on instance (22), over the fourteen initialisations of Figure 1; the η axis is stretched below 0.6 (dotted line). From left: runs reaching the flat minimum under Algorithm 1; the same under the variant in which the outer perturbation also reaches the penalty; the outer and the inner curvature at the returned point under Algorithm 1 (medians over starts).

How the radii share the work. The first panel of Figure 8 maps the share of runs reaching the flat minimum over the (η, ρ) plane. The outer radius alone never succeeds, and the inner radius alone needs $\eta = 1.8$. With the outer radius between about 0.3 and 0.75 the inner radius needed falls to about 0.8, less than half, but beyond that window it rises again, above its value without any outer radius at $\rho = 1.3$. The perturbed outer landscape explains both halves. Along the segment joining the two minima, the ball lowers the barrier that separates them to about a third of its height while $\rho \leq 0.6$, which is what the inner drift then has to cover; from $\rho \approx 0.8$ on, the ball also erases the flat well, and the perturbed landscape rises all the way from the bowl to the flat minimum, so the drift has to work against the outer perturbation instead of with it. The outer radius is therefore useful to the inner one only in a window set by the widths of the two wells: wide enough to erase the spike, narrow enough to spare the flat well. A variant of the algorithm makes the pair stronger. If the perturbation also reaches the penalty, the inner-sharpness term is evaluated at $\mathbf{x} + \boldsymbol{\epsilon}$, where $\mathbf{y}$ is off the inner manifold by a distance of order $\kappa\rho$, and the drift $\gamma\eta\|\mathbf{d}\|\nabla_{\mathbf{x}} s$ grows accordingly. Under that variant (second panel) the inner radius needed drops much further, to 0.05 for large ρ. It is not the rule analysed in Section 3, but it suggests that the inner and outer perturbations could usefully share their evaluation point.

The price of the radii. Where the runs succeed they stop short of the flat minimum, and the distance is set almost entirely by the outer radius: fitted over the successful cells of the grid, the inner radius contributes almost nothing to it. This is the $\mathcal{O}(\rho^2 + \eta^2)$ floor of Theorem 3.1, and on this instance it is the outer term that dominates. Once the flat minimum is reached, the inner radius can therefore be taken generously at little cost, while the outer one should be kept at the smallest value that erases the sharp region.

One practical remark is worth recording. The single normalised ascent step of Lemma C.1 is a first-order approximation of the maximisation over the outer ball, and it degrades once ρ is comparable to the width of the basin being escaped. With the outer radius alone most runs do not leave the sharp minimum at all, since the gradient there carries no information about the basin around it, and the others stall at a point that is neither minimum, where the perturbed gradient vanishes although the true one does not. From $\rho = 0.4$ on, exact search lifts the runs out of the spike. With the inner radius included the two rules mostly agree, and at $\rho = 0.4$, $\eta = 0.9$ both send every run to the flat minimum. For the small radii of Sections 5.2 and 5.3 the one-step rule is adequate.

A controlled version of hypercleaning. We next reproduce the task of Section 5.2 in a regime where the outer problem is genuinely prone to overfitting. The inner variable $\mathbf{y} \in \mathbb{R}^{20}$ holds logistic-regression parameters and the outer variable $\mathbf{x} \in \mathbb{R}^{200}$ holds one weight per training sample, with

$$f(\mathbf{x}, \mathbf{y}) = \frac{1}{m} \sum_{j=1}^{m} \ell(\mathbf{x}, \mathbf{y}; \xi_j), \qquad g(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} [\sigma(\mathbf{x})]_i \, \ell(\mathbf{y}; \zeta_i) + \frac{\lambda}{2} \|\mathbf{y}\|^2, \tag{128}$$

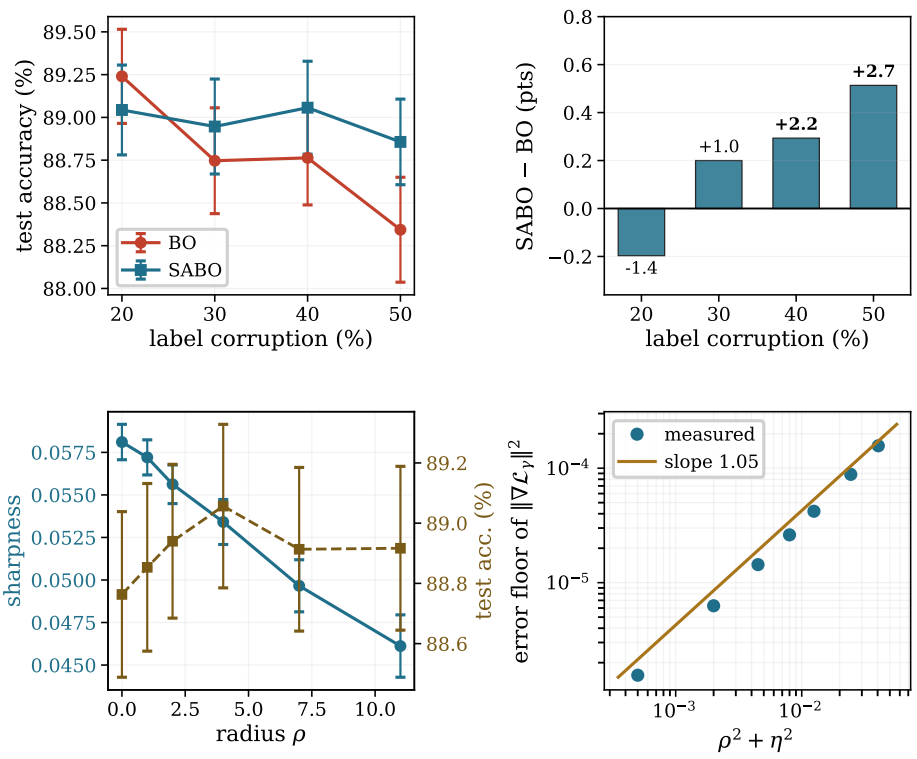

Figure 9: Synthetic experiments over 15 seeds; error bars are standard errors. Top left: test accuracy against label corruption for data hypercleaning (128) with a 30-sample validation set. Top right: the same data as a paired difference, annotated with the paired t-statistic; the SABO advantage grows with the corruption rate and is significant at 40% and 50%. Bottom left: sharpness of the outer objective at the returned solution (left axis, circles) decreases monotonically in ρ, while test accuracy (right axis, squares) peaks at an intermediate radius. Bottom right: the plateau of $(1/K)\sum_k \|\nabla\mathcal{L}_\gamma(\mathbf{x}_k, \mathbf{y}_k)\|^2$ against $\rho^2+\eta^2$; the fitted slope of 1.05 confirms the $\mathcal{O}(\rho^2+\eta^2)$ floor of Theorem 3.1.

where $n = 200$ training points of which a fraction is corrupted, only $m = 30$ clean validation points are available, and $\lambda = 5 \cdot 10^{-2}$ so that Assumption 2.1 again holds globally. We use $T_k = 5$, $\gamma = 5$, minibatches of 64, $K = 1500$, and radii $\rho = 4$ and $\eta = 0.8$ for SABO.

The outcome, reported in Figure 9, is that the benefit of sharpness-awareness is a function of how much there is to overfit. At 20% corruption SABO and BO are statistically indistinguishable, SABO in fact being 0.20 points worse with a paired t-statistic of -1.39; the advantage then grows monotonically with the noise level and becomes significant at 40% ($+0.29$ points, $t = +2.25$) and at 50% ($+0.51$ points, $t = +2.67$). Sharpness-awareness buys nothing when the outer problem is already well posed, and buys more the harder the problem becomes, which is precisely the pattern reported in Table 1. Independently of accuracy, the mechanism is the one the analysis assumes. The solutions SABO returns are measurably flatter at every corruption level: the worst-case increase of the outer loss over a ball of radius 0.5 around $\mathbf{y}$ falls from 0.0543 to 0.0507 at 20% corruption and from 0.0600 to 0.0553 at 50%, while the top Hessian eigenvalue of the outer loss falls in step, from 0.3557 to 0.3417 at 50%. This flattening is controlled by ρ exactly as the theory says it should be, decreasing monotonically over the whole range tested and by 21% between $\rho = 0$ and $\rho = 11$, while test accuracy peaks at an intermediate radius and then flattens — the trade-off anticipated in the discussion following Theorem D.2.

Which curvature, and where it comes from. Hypercleaning lets the curvature at the returned solution be measured exactly, since the Hessian of the logistic validation loss in $\mathbf{y}$ is available in closed

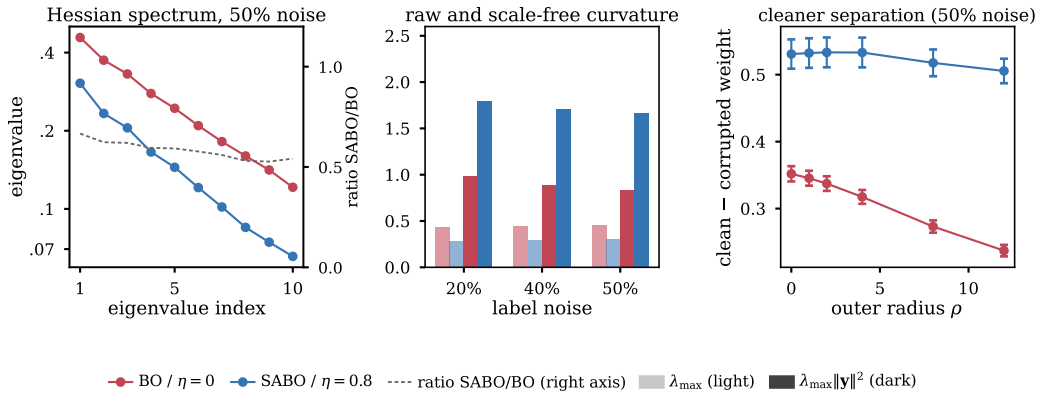

Figure 10: Curvature at the solutions of hypercleaning (128), averaged over 15 seeds. Left: the ten leading eigenvalues of the validation-loss Hessian in $\mathbf{y}$ at 50% corruption, for BO and for SABO ($\rho = 4$, $\eta = 0.8$), with their ratio dashed on the right axis. Middle: the largest eigenvalue (light) and its scale-free version $\lambda_{\max}\|\mathbf{y}\|^2$ (dark) at each corruption level. Right: the difference between the mean weight $\sigma(\mathbf{x})$ given to clean and to corrupted training samples as a function of ρ, without and with the inner radius; bars are standard errors.

form. We repeated the task over 15 seeds, three corruption levels, outer radii $\rho \in \{0, 1, 2, 4, 8, 12\}$ and inner radius $\eta \in \{0, 0.8\}$, and computed the full spectrum at every solution; Figure 10 summarises the result. Three findings stand out. First, the flattening is produced by the inner radius, not the outer one: $\eta = 0.8$ lowers the largest eigenvalue by about a third and the trace by about two fifths at every corruption level, whereas ρ, which perturbs the sample weights rather than the model, leaves the curvature in $\mathbf{y}$ essentially unchanged. Second, the reduction is uniform across the spectrum, every one of the ten leading eigenvalues shrinking by a similar factor of 0.53–0.67, and it is accompanied by a growth of $\|\mathbf{y}\|$ from about 1.4 to 2.4. For a logistic loss that is exactly the signature of a rescaling: larger margins saturate the loss and flatten it in every direction at once. Measured in a scale-free way, as $\lambda_{\max}\|\mathbf{y}\|^2$, the same solutions are *sharper*, not flatter, and across all runs neither measure predicts test accuracy from seed to seed. Curvature in parameter space is therefore a fragile summary of what sharpness-awareness does, in line with the observation that sharp minima can be reparameterised into flat ones (Dinh et al., 2017). Third, what does change in a way that bears on the task is the cleaner itself: the gap between the average weight given to clean and to corrupted samples grows from 0.35 to 0.53 when $\eta = 0.8$, and it is eroded again only by large outer radii. The inner radius, in other words, acts on hypercleaning through the quality of the learned weights more than through the flatness of the classifier.