Hyper-Gradient Methods for Bilevel Optimization with Manifold Lower-level Solution Set

Saeed Masiha[1], **Zebang Shen**[2], **Negar Kiyavash**[1], **and Niao He**[2]
[1]EPFL School of Management of Technology, Station 5, 1015 Lausanne, Switzerland
mohammadsaeed.masiha@epfl.ch, negar.kiyavash@epfl.ch
[2]ETH Department of Computer Science, Universitätstrasse 6, 8092 Zürich, Switzerland
zebang.shen@inf.ethz.ch, niao.he@inf.ethz.ch

Abstract

We study optimistic bilevel optimization when the lower-level problem has a non-isolated manifold of minimizers, a regime where standard hyper-gradient analyses break down because the hyper-objective can become nonsmooth due to minima selection. Under a local Polyak–Łojasiewicz condition, we show that if the optimistic minimizer is unique, then the hyper-objective is locally differentiable and admits an explicit hyper-gradient formula involving a Moore–Penrose pseudoinverse; with an additional non-degeneracy condition, the hyper-objective is locally smooth. Building on this structure, we propose HG-MS, a hyper-gradient method that combines Gibbs/Langevin sampling, Best-of-N minima selection, and efficient pseudoinverse approximation via Hessian–vector products. We establish near-stationarity and complexity guarantees that explicitly capture errors from sampling, selection, and linear solves, with rates depending on the intrinsic dimension of the lower-level solution manifold. Experiments on data hyper-cleaning and imbalanced-loss tuning show that explicit minima selection improves optimization stability and downstream performance over standard hyper-gradient and penalty-based baselines.

1 Introduction

Many problems in modern machine learning and AI can be cast as bilevel optimization, where an outer objective evaluates the outcome of an inner training procedure. Examples include hyperparameter optimization and meta-learning [Feurer and Hutter, 2019, Liu et al., 2021, Bertinetto et al., 2018, Finn et al., 2017], differentiable neural architecture search [Liu et al., 2019], example reweighting/data cleaning for robust learning [Ren et al., 2018], and bilevel formulations in reinforcement learning (e.g., actor–critic) [Hong et al., 2023], as well as bilevel formulations for LLM data reweighting, safe fine-tuning, and preference alignment [Pan et al., 2025, Shen et al., 2025b, Jian et al., 2025].

We focus on the optimistic[1] bilevel problem [Dempe et al., 2007, Ye et al., 1997, Ye and Ye, 1997]

$$\min_{\theta\in\Theta} F(\theta) \quad \text{with} \quad F(\theta) := \min_{x\in\mathcal{S}(\theta)} f(\theta, x) \text{ where } \mathcal{S}(\theta) := \arg\min_{x\in\mathbb{R}^d} g(\theta, x). \tag{1}$$

Here, to define the *hyper-objective* $F : \Theta \to \mathbb{R}$ on a compact convex set $\Theta \subseteq \mathbb{R}^m$, one minimizes the upper-level loss $f : \Theta \times \mathbb{R}^d \to \mathbb{R}$ over a lower-level solution set $\mathcal{S}(\theta) \subseteq \mathbb{R}^d$ of the lower-level loss $g : \Theta \times \mathbb{R}^d \to \mathbb{R}$. We refer to this inner minimization as *minima-selection*. We refer to the solutions to minima-selection as *optimistic minimizers*, denoted by

$$\mathtt{O}(\theta) := \arg\min_{x\in\mathcal{S}(\theta)} f(\theta, x). \tag{2}$$

[1]All of our conclusions can be similarly derived for the pessimistic formulation as well.

Preprint.

In particular, when $\texttt{O}(\theta)$ is a singleton, we denote the *unique optimistic minimizer* as $x^\star(\theta)$. In this work, we focus on the settings where the hyper-objective F can be proved to be differentiable. Since F is in general non-convex, we take the *first-order stationary point*[2] of the hyper-objective F as our solution concept. For works seeking more relaxed alternative solution concepts, see Appendix B.

In the existing literature, differentiability of the hyper-objective is most commonly established under the setting where the lower-level solution set $\mathcal{S}(\theta)$ is a *singleton* (in which case $\mathcal{S}(\theta) = \texttt{O}(\theta) = \{x^\star(\theta)\}$ and minima-selection is redundant) and that the lower-level loss $g(\theta, \cdot)$ satisfies *locally quadratic growth* around the unique minimizer $x^\star(\theta)$ [Pedregosa, 2016, Franceschi et al., 2017, Lorraine et al., 2020, Huang, 2024, Liu et al., 2022a, Kwon et al., 2023b] (see Appendix B.1 for details). In particular, with additional mild regularity assumptions on f and g, the hypergradient $\nabla F(\theta)$ admits an explicit expression. Recent work has sought to establish differentiability of the hyper-objective beyond the singleton-minimizer regime [Xiao et al., 2023, Shen et al., 2025a]. However, these extensions typically rely on assuming that the lower-level loss g is *convex* in the lower-level variable x, an assumption that rules out many practically important non-convex settings.

For instance, in bilevel learning problems such as hyperparameter tuning [Pedregosa, 2016, Franceschi et al., 2017, Lorraine et al., 2020], data cleaning [Ren et al., 2018], and recent LLM data reweighting and safe fine-tuning formulations [Pan et al., 2025, Shen et al., 2025b], the lower-level problem is often an empirical risk minimization (ERM) task over a training set using a neural-network-based model, while the upper-level objective evaluates a risk measure on a validation set. In this regime, distinct lower-level solutions may attain nearly identical training loss yet exhibit markedly different validation performance. This motivates defining $F(\theta)$ via an *optimistic* selection rule (as in eq. (1)) that, among all lower-level minimizers, chooses the one with the best validation performance. In such applications, the lower-level solution set $\mathcal{S}(\theta)$ is typically *neither a singleton nor convex*, and existing differentiability results therefore fail to capture these core use cases of bilevel optimization.

Manifold lower-level solution set and differentiability of the hyper-objective F. Without additional structure, F need not be continuous (let alone differentiable) in the worst case, [Dontchev and Rockafellar, 2009, Arbel and Mairal, 2022]. This motivates imposing further assumptions.

Under the common practice of neural network overparameterization, the lower-level loss g can admit many global minimizers that are not isolated points but instead form a *connected* set—often with *manifold* structure [Draxler et al., 2018, Garipov et al., 2018, Nguyen, 2019]. Guided by these empirical observations and prior work on the loss landscapes of overparameterized deep networks [Venturi et al., 2018, Liu et al., 2022b], we adopt the setting of [Masiha et al., 2025] and assume a *local PŁ condition* (specifically, PŁ° in [Gong et al., 2024]) for the lower-level loss g (with respect to the lower-level variable x). Under this condition, the lower-level solution set $\mathcal{S}(\theta)$ is well-structured and, in particular, forms an embedded submanifold of $\mathbb{R}^d$ (possibly of positive dimension). To the best of our knowledge, this local PŁ-type condition is one of the weakest assumptions in the literature that guarantees the hyper-objective F is *Lipschitz continuous*, a natural precursor to differentiability. However, we also highlight that this key regularity property is *insufficient* for our purpose.

In this work, we ask:

> In the setting where g satisfies the local PL condition, what are the sufficient conditions for the hyper-objective F to be differentiable or even smooth?

Our key observation is that, when there is `tie` in minima-selection, i.e., it has non-unique solutions, the set $\texttt{O}(\theta)$ is sensitive to changes in θ, constituting a major source of non-differentiability of F, even when the set-valued map $\theta \mapsto \mathcal{S}(\theta)$ varies smoothly with θ; see Example 3.1 for a concrete example. This suggests imposing the assumption that there is a *unique* optimistic minimizer $x^\star(\theta)$. Our first result shows that this intuition is indeed sufficient.

Theorem (Regularity; Informal)**.** *Suppose that g satisfies the local PŁ condition (see assumption 1).*

- *Assuming that the minima-selection step yields a unique solution $x^\star(\theta)$, i.e. $\texttt{O}(\theta)$ is a singleton, the hyper-objective F is continuously differentiable and the hyper-gradient admits the explicit form*

$$\nabla F(\theta) = \nabla_\theta f(\theta, x^\star(\theta)) - \nabla^2_{\theta x} g(\theta, x^\star(\theta)) \big[\nabla^2_{xx} g(\theta, x^\star(\theta))\big]^\dagger \nabla_x f(\theta, x^\star(\theta)). \tag{3}$$

[2] Stationary point in the gradient mapping sense, given the convex compact Θ constraint.

Here $[\cdot]^\dagger$ denotes the Moore–Penrose pseudoinverse.

- *Moreover, if $x^\star(\theta)$ is non-degenerate*[3]*, ∇F is Lipschitz continuous.*

We highlight that the unique optimistic minimizer assumption is *strictly weaker* than the singleton lower-level solution set assumption commonly encountered in the literature: When $\mathcal{S}(\theta)$ is singleton, the optimistic minimizer $x^\star(\theta)$ is clearly unique.

Moreover, as complementary results, we construct adversarial instances demonstrating that

- when the set $\mathtt{O}(\theta)$ is not a singleton, i.e., there is `tie` in minima-selection, F can be non-differentiable at *infinitely many* points within any prescribed *arbitrarily small* neighborhood;
- under the unique minimizer condition, if $x^\star$ is a degenerate point in the sense of Definition 3.4, ∇F can fail to be α-Hölder continuous with *any prescribed index* $\alpha > 0$.

Algorithmic consequence: select then differentiate. The differentiability result above, together with the corresponding hyper-gradient expression (3), suggests a practical recipe for bilevel optimization in regimes when the lower-level solution set exhibits manifold structure and minima-selection has a unique solution. Note that, in contrast to standard hyper-gradient methods for the singleton-minimizer case, our setting requires explicitly accounting for the minima-selection step.

To this end, we propose *Hyper-Gradient with Minima-Selection* (HG-MS); see Algorithm 1. The method proceeds in two stages:

- We tackle minima-selection with a *Best-of-N* (`BoN`) strategy: To mimic the enumeration over the non-convex set $\mathcal{S}(\theta)$, we generate N samples from a Gibbs measure $\mu_g^\lambda(\theta) \propto \exp(-g(\theta,\cdot)/\lambda)$ and output $\hat{x}_N^\lambda(\theta)$, the one achieving the minimum $f(\theta,\cdot)$ value, as an approximation for $x^\star(\theta)$. Note that when $\lambda \to 0$, $\mu_g^\lambda(\theta)$ exponentially concentrates around $\mathcal{S}(\theta)$, and we can show that

$$\forall \theta \in \Theta, \quad \mathbb{E}[\|x^\star(\theta) - \hat{x}_N^\lambda(\theta)\|^2] = \mathcal{O}(N^{-\frac{1}{\mathtt{k}}} + \lambda^{\frac{1}{2}}).$$

  Here $\mathtt{k} \geq 1$ denotes the *common intrinsic* dimension of the θ-dependent manifolds $\{\mathcal{S}(\theta)\}_{\theta\in\Theta}$.
- We approximate ∇F by directly using $\hat{x}_N^\lambda(\theta)$ as a proxy for $x^\star(\theta)$ in Equation (3).

For the convergence analysis, we view HG-MS as an inexact projected-gradient method applied to the hyper-objective F: the computed hyper-gradient can be written as $\nabla F(\theta_t) + e_t$. The main technical work is then to bound $\mathbb{E}[\|e_t\|^2]$ by decomposing it into contributions from (a) sampling from $\mu_g^\lambda(\theta)$, (b) `BoN` selection with finite N, and (c) approximating pseudo-inversion in eq. (3).

Empirically, we evaluate HG-MS on data hyper-cleaning and parametric loss tuning for imbalanced classification under fixed time budgets. Explicitly incorporating minima selection produces more stable upper-level trajectories and improves downstream performance: on data hyper-cleaning, it increases test accuracy from 81.2% to 87.0%, from 73.1% to 80.1%, and from 64.5% to 66.1% at corruption rates $\rho = 0.4, 0.6, 0.8$, respectively; on imbalanced loss tuning, it improves test balanced accuracy from 95.48% to 96.45%.

Contributions. We make three contributions:

- **Differentiability of the hyper-objective under manifold-structured lower-level minimizers.** We study the regime in which the lower-level function $g(\theta,\cdot)$ admits a non-singleton manifold of minimizers, as implied by the local PŁ condition considered in this paper. Under the unique optimistic minimizer assumption, we prove that the optimistic hyper-objective F is locally $\mathcal{C}^1$ and derive the corresponding hyper-gradient formula (Theorem 3.5). Under an additional non-degeneracy assumption on the selected optimistic minimizer (formalized in Assumption 3 using Definition 3.4), we further show that ∇F is Lipschitz continuous. Importantly, we construct hard instances showing that, in the worst case, these conditions are necessary for the local $\mathcal{C}^1$ regularity of F and for the Lipschitz continuity of ∇F, respectively.

[3]Here non-degenerate means that $f(\theta,\cdot)$ has strictly positive second-order curvature along every tangent direction of the lower-level manifold $\mathcal{S}(\theta)$ at $x^\star(\theta)$; see Definition 3.4.

- **A hyper-gradient method with BoN minima selection and complexity guarantees.** We propose HG-MS (Algorithm 1), which combines BoN approximate minima selection with approximate hyper-gradient computation. We prove that HG-MS finds an ϵ-stationary point of the hyper-objective F using $\mathcal{O}(\text{poly}(\epsilon^{-\mathtt{k}}))$ queries to the oracle for $\partial_x g(\theta, \cdot)$. Here, $\mathtt{k}$ denotes the common intrinsic dimension of $\mathcal{S}(\theta)$. Moreover, the $\epsilon^{-\mathtt{k}}$ dependence is necessary in the worst case due to the nonconvex nature of the problem [Nemirovskij and Yudin, 1983].
- **Empirical evidence.** On data hyper-cleaning and imbalanced-loss tuning, HG-MS yields more stable outer trajectories and improved validation/test performance compared to standard hyper-gradient and penalty/alternating baselines (Section 6).

Notations For any nonempty closed convex set $C \subset \mathbb{R}^m$, we write $\text{Proj}_C(u) := \arg\min_{v \in C} \|v - u\|$. For any embedded $\mathcal{C}^2$ submanifold $\mathcal{M} \subset \mathbb{R}^d$ and $x \in \mathcal{M}$, we write $\mathcal{T}_x\mathcal{M}$ and $\mathcal{N}_x\mathcal{M}$ for the tangent and normal spaces, and $P_{\mathcal{T}_x\mathcal{M}}$, $P_{\mathcal{N}_x\mathcal{M}}$ for the corresponding orthogonal projectors. When $\mathcal{M} = \mathcal{S}(\theta)$, we abbreviate $\mathcal{T}_x^\theta := \mathcal{T}_x\mathcal{S}(\theta)$ and $\mathcal{N}_x^\theta := \mathcal{N}_x\mathcal{S}(\theta)$.

2 Local PL Condition and the Manifold Solution Set

We now describe the PŁ$^\circ$ condition introduced by Gong et al. [2024], which is our central structural assumption. Besides the standard *local* Polyak–Łojasiewicz (PŁ) inequality around the local minimizers, the extra requirement in the PŁ$^\circ$ condition ensures that all local minima are connected (hence are all global minima). This supports efficient sampling-based approximations of the global optimal set.

Definition 2.1 (Local PŁ)**.** *Let $\mathcal{M}$ be the collection of all local minima of $g \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$. We say that g is* locally PŁ *if there exists $\mu > 0$ such that for any connected component $\mathcal{M}' \subseteq \mathcal{M}$, there exists an open neighborhood $\mathcal{N}(\mathcal{M}') \supset \mathcal{M}'$ satisfying*

$$g(x) - \min_{x' \in \mathcal{N}(\mathcal{M}')} g(x') \;\le\; (2\mu)^{-1} \|\nabla g(x)\|^2, \; \forall x \in \mathcal{N}(\mathcal{M}').$$

Definition 2.2 (PŁ$^\circ$ condition)**.** *A function $g : \mathbb{R}^d \to \mathbb{R}$ satisfies the PŁ$^\circ$ condition if: (1) $g \in \mathcal{C}^2$ is locally PŁ, and $g \in \mathcal{C}^4$ on every neighborhood $\mathcal{N}(\mathcal{M}')$; (2) Let $\mathcal{N}(\mathcal{M}) := \cup_{\mathcal{M}'}\mathcal{N}(\mathcal{M}')$. For any $x \in \mathbb{R}^d \setminus \mathcal{N}(\mathcal{M})$, if $\nabla g(x) = 0$, then $\nabla^2 g(x) \prec 0$; (3) The collection of all local minima $\mathcal{M}$ is contained in a compact set.*

We impose PŁ$^\circ$ uniformly over the parameter θ in the lower-level problem.

Assumption 1 (Lower-level PŁ$^\circ$)**.** *There exists a constant $\mu > 0$ such that for every $\theta \in \Theta$, the function $g(\theta, \cdot)$ satisfies the PŁ$^\circ$ condition, and its local PŁ neighborhoods in Definition 2.1 can be chosen with the same constant μ.*

Assumption 1 has both geometrical (manifold structure) and analytical (normal nondegeneracy) implications on the lower-level solution set $\mathcal{S}(\theta)$, summarized as follows.

Proposition 2.3 (Characterizations of $\mathcal{S}(\theta)$)**.** *Let Assumption 1 hold. Then, for every $\theta \in \Theta$,*

(i) $\mathcal{S}(\theta)$ is a connected compact $\mathcal{C}^2$ embedded submanifold of $\mathbb{R}^d$ without boundary [Gong et al., 2024, Prop. 3, Cor. 1]. In particular, all local minima of $g(\theta, \cdot)$ are global minima.

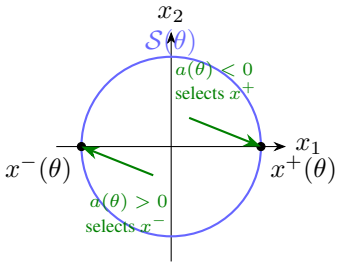
(ii) For any $x \in \mathcal{S}(\theta)$, the Hessian $H(\theta, x) := \nabla^2_{xx} g(\theta, x)$ satisfies $\ker H(\theta, x) = \mathcal{T}_x^\theta$ and $\langle v, H(\theta, x)v\rangle \ge c\|v\|^2$ for all $v \in \mathcal{N}_x^\theta$ for some constant $c > 0$ that is uniform over $\theta \in \Theta$ and $x \in \mathcal{S}(\theta)$ [Rebjock and Boumal, 2024, Cor. 2.17].

Moreover, under the additional regularity conditions required by [Masiha et al., 2025, Lem. 3.5] (which we include in Assumption 4), there exists an integer $\mathtt{k}$ such that

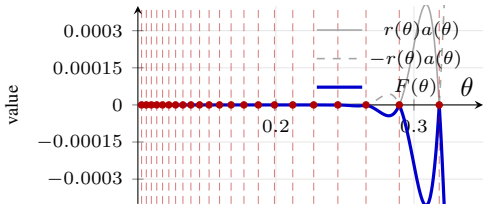
$$\dim(\mathcal{S}(\theta)) = \mathtt{k} \qquad \text{for all } \theta \in \Theta.$$

Remark 2.4 (Lipschitzness of hyper-objective)**.** *Under the lower-level local PŁ$^\circ$ condition and mild regularity on f and g, the hyper-objective function is Lipschitz; see [Masiha et al., 2025, Lem. 3.2]. This implies that F is differentiable almost everywhere. However, this is insufficient for our purpose.*

Curvature regularity of the solution manifolds. Our selection analysis uses volume comparison bounds for small geodesic balls on the compact manifolds $\mathcal{S}(\theta)$ (e.g., [Masiha et al., 2025, Lem. 4.1]),



(a) Two competing optimistic minimizers on $\mathcal{S}(\theta)$.



(b) Two branches and their envelope (kinks at ties $\theta_k = 10/(k\pi)$, a countably infinite set accumulating at 0, marked in red).

Figure 1: **Illustration of Example 3.1.** Left: the lower-level minimizer set $\mathcal{S}(\theta)$ is a smooth circle, but the optimistic solution switches between two competing endpoints depending on the sign of $a(\theta)$. Right: the optimistic hyper-objective $F(\theta) = -r(\theta)|a(\theta)|$ is the pointwise minimum of two smooth branches $\pm r(\theta)a(\theta)$. Kinks (nondifferentiability) occur at tie points $a(\theta) = 0$ (marked in red).

which requires curvature control. Assuming the following simple analytic condition on g and together with the normal spectral gap from Proposition 2.3 (iii), we show in Appendix F.5 a uniform bound on the second fundamental form of $\mathcal{S}(\theta)$.

Assumption 2 (regularity of g)**.** *There exist constants $\rho > 0$ and $L_{g,3} < \infty$ such that for all $\theta \in \Theta$, $g(\theta, \cdot)$ is $\mathcal{C}^3$ on the tube $\{x : \mathrm{dist}(x, \mathcal{S}(\theta)) \le \rho\}$ and $\sup_{\mathrm{dist}(x,\mathcal{S}(\theta))\le\rho} \left\|\nabla^3_{xxx} g(\theta, x)\right\|_{\mathrm{op}} \le L_{g,3}$.*

3 Differentiability of the Hyper-objective and the Hyper-gradient

This section analyzes the regularity of the optimistic hyper-objective F in four steps.

- Through a counterexample in Example 3.1, we show that when there is a `tie` in minima-selection, i.e. $\mathtt{0}(\theta)$ is non-singleton, F is not necessarily differentiable, even for f and g both in $\mathcal{C}^\infty$.
- We show in Theorem 3.2 that the uniqueness of optimistic minimizer is sufficient to obtain $F \in \mathcal{C}^1$, together with the explicit hyper-gradient formula (3).
- Through another counterexample in Example 3.3, we show that the unique optimistic minimizer condition alone does not guarantee α-Hölder continuity of ∇F for any $\alpha > 0$ (including Lipschitz continuity, which corresponds to $\alpha = 1$).
- We show that when enhancing the unique optimistic minimizer condition with the non-degeneracy condition from Assumption 3, the unique minimizer $x^\star(\theta)$ is a $\mathcal{C}^1$ function (see Theorem 3.5) and F is locally smooth (see Proposition 5.1).

`Tie` in minima-selection can create kinks in the hyper-objective. Even when $\mathcal{S}(\theta)$ is a $\mathcal{C}^\infty$ connected manifold and varies smoothly with θ, minimizing x over $\mathcal{S}(\theta)$ can still introduce kinks in F when the optimistic minimizer is not unique. The example below shows that, in any *arbitrarily small* neighborhood, even when f, g, and $\mathcal{S}(\theta)$ are all $\mathcal{C}^\infty$, F can have *infinitely* many non-differentiable points purely due to the `tie` phenomenon. W.l.o.g., we take the neighborhood to be centered at 0.

Example 3.1 (`Tie` in minima-selection creates kinks)**.** *Let $\theta \in \mathbb{R}$ and $x = [x_1, x_2] \in \mathbb{R}^2$. Define $a(\theta) := e^{-1/\theta^2} \sin(10/\theta)$ for $\theta \neq 0$ and $a(0) = 0$. Consider the losses*

$$f(\theta, x) := a(\theta)\, x_1 + x_2^2 \quad \text{and} \quad g(\theta, x) := \left(\|x\|^2 - (1 + \theta^2)\right)^2,$$

both of which are $\mathcal{C}^\infty(\mathbb{R} \times \mathbb{R}^2)$. The lower-level solution set is the circle (which is $\mathcal{C}^\infty$)

$$\mathcal{S}(\theta) = \{x \in \mathbb{R}^2 : \ \|x\| = \sqrt{1 + \theta^2}\}.$$

Write $r(\theta) = \sqrt{1 + \theta^2}$. On the circle, $x_2^2 = r(\theta)^2 - x_1^2$, hence $f(\theta, x) = a(\theta)x_1 + r(\theta)^2 - x_1^2$, a concave quadratic in x_1. Therefore its minimum over $x_1 \in [-r(\theta), r(\theta)]$ is attained at an endpoint:

for $a(\theta) \neq 0$ the constrained minimizer is unique and

$$x^\star(\theta) = [-\mathrm{sign}(a(\theta))r(\theta), 0].$$

At any θ with $a(\theta) = 0$, we have $\arg\min_{x \in \mathcal{S}(\theta)} f(\theta, x) = \{[-r(\theta), 0], [r(\theta), 0]\}$, i.e., the selected optimistic minimizer is not unique*. Consequently,*

$$F(\theta) = \min_{x \in \mathcal{S}(\theta)} f(\theta, x) = -r(\theta)\,|a(\theta)|.$$

Since $a(\theta) = 0$ at $\theta_k := 10/(k\pi)$ for each $k \in \mathbb{Z} \setminus \{0\}$ and these zeros are simple, F is not differentiable at every θ_k (a countably infinite set accumulating at 0).

This mechanism is geometric: although $\mathcal{S}(\theta)$ is a smooth manifold, the optimistic minimizer can switch between competing branches at θ's where $\mathtt{O}(\theta)$ is not unique, producing kinks in F.

Uniqueness in minima-selection implies differentiablity of hyper-objective. The above example suggests imposing the unique-solution condition on minima-selection, which we show is indeed sufficient for $F \in \mathcal{C}^1$.

Theorem 3.2 ($\mathcal{C}^1$ regularity of the hyper-objective under uniqueness)**.** *Let Assumption 1 hold and assume $f \in \mathcal{C}^1$ and $g \in \mathcal{C}^3$. Fix $\theta_0 \in \Theta$ and assume there exist an open set $\mathcal{U}$ containing θ_0 and a compact set $\mathcal{V} \subset \mathbb{R}^d$ such that, for all $\theta \in \mathcal{U}$,*

$$\mathcal{S}(\theta) \subseteq \mathcal{V}, \qquad \textit{and} \qquad \arg\min_{x \in \mathcal{S}(\theta)} f(\theta, x) \textit{ is a singleton.}$$

Then there exists an open neighborhood $\mathcal{U}_0 \subseteq \mathcal{U}$ of θ_0 such that F is continuously differentiable on $\mathcal{U}_0$. Moreover, for all $\theta \in \mathcal{U}_0$, the hyper-gradient is given by eq. (3)*.*

Proof sketch. We sketch the proof in three steps; Appendix C.2 contains the details.

Step 1: parameterize $\mathcal{S}(\theta)$ locally by solving the normal equations. Let $x_0 := x^\star(\theta_0)$ and abbreviate $\mathcal{T}_0 := \mathcal{T}_{x_0}^{\theta_0}$ and $\mathcal{N}_0 := \mathcal{N}_{x_0}^{\theta_0}$. Choose orthonormal bases $U_{\mathcal{T}_0}$ and $U_{\mathcal{N}_0}$ spanning $\mathcal{T}_0$ and $\mathcal{N}_0$. Write local coordinates $x = x_0 + U_{\mathcal{T}_0} u + U_{\mathcal{N}_0} v$ and consider

$$\Phi(\theta, u, v) := U_{\mathcal{N}_0}^\top \nabla_x g(\theta, x_0 + U_{\mathcal{T}_0} u + U_{\mathcal{N}_0} v).$$

By the normal nondegeneracy implied by Assumption 1 (see Proposition 2.3), $D_v \Phi(\theta_0, 0, 0) = U_{\mathcal{N}_0}^\top \nabla_{xx}^2 g(\theta_0, x_0) U_{\mathcal{N}_0}$ is invertible. Hence the implicit function theorem gives a C^2 map $v = \varphi(\theta, u)$ solving $\Phi(\theta, u, \varphi(\theta, u)) = 0$ near $(\theta_0, 0)$. Defining

$$\psi(\theta, u) := x_0 + U_{\mathcal{T}_0} u + U_{\mathcal{N}_0} \varphi(\theta, u),$$

we obtain a local C^2 chart of the nearby branch of $\mathcal{S}(\theta)$ around x_0.

Step 2: reduce to a fixed-domain minimization. Uniqueness and compactness imply that for θ close to θ_0, the selected minimizer stays inside this chart patch. Hence the optimistic problem reduces to a fixed-domain minimization

$$F(\theta) = \min_{u \in U} \tilde{f}(\theta, u), \qquad \tilde{f}(\theta, u) := f(\theta, \psi(\theta, u)),$$

over a compact set U independent of θ.

Step 3: apply a Danskin-type envelope argument and compute the hyper-gradient. We apply the Danskin-type envelope lemma in Appendix C.2, namely Lemma C.3. The point of this lemma is that for a value function of the form $F(\theta) = \min_{u \in U} \phi(\theta, u)$ over a fixed compact set, uniqueness of the minimizer allows one to differentiate F by differentiating only ϕ with respect to θ at the minimizing point; no derivative of the argmin map is needed. Applying this lemma pointwise, and then using continuity of the unique minimizer on the fixed compact domain, yields the local $\mathcal{C}^1$ regularity of F. Finally, differentiating the lower-level stationarity equation gives the pseudoinverse hyper-gradient formula (3). ■

Uniqueness in minima-selection is insufficient for Hölder continuity of ∇F. From an optimization perspective, however, $\mathcal{C}^1$ regularity is not enough. For a non-asymptotic analysis, we typically need ∇F to be α-Hölder continuous for some $\alpha > 0$. However, in the following example, we show that uniqueness alone is *insufficient* to guarantee this property.

Example 3.3 (Uniqueness does not imply local α-Hölder continuity of ∇F)**.** *Consider the following smooth construction. For $\theta \in \mathbb{R}$ and $x = (x_1, x_2) \in \mathbb{R}^2$, let*

$$a(\theta) := \begin{cases} \frac{1}{4} e^{-1/\theta^2} \sin(10/\theta), & \theta \neq 0, \\ 0, & \theta = 0, \end{cases} \qquad \eta(t) := \begin{cases} \operatorname{sign}(t) e^{-1/t^2}, & t \neq 0, \\ 0, & t = 0, \end{cases}$$

and define $\phi(t) := \int_0^t \eta(s)\,\mathrm{d}s$. Then $\eta, \phi \in \mathcal{C}^\infty(\mathbb{R})$, $\eta = \phi'$ is strictly increasing, and $\phi''(0) = \eta'(0) = 0$. Set

$$g(\theta, x) := \left(\|x\|^2 - (1+\theta^2)\right)^2, \qquad f(\theta, x) := \phi(x_2) + a(\theta) x_2 + \rho(-x_1),$$

where $\rho(s) = e^{-1/s}$ for $s > 0$ and $\rho(s) = 0$ for $s \leq 0$. Thus $\mathcal{S}(\theta) = \{x \in \mathbb{R}^2 : \|x\| = r(\theta)\}$ with $r(\theta) := \sqrt{1+\theta^2}$.

As in Example 3.1, the term $\rho(-x_1)$ forces the optimistic minimizer to lie on the right semicircle, so minimizing $f(\theta, \cdot)$ over $\mathcal{S}(\theta)$ reduces to the scalar problem $\min_{|t| \leq r(\theta)} \big(\phi(t) + a(\theta) t\big)$. Since η is strictly increasing, this problem has a unique critical point, determined by $\eta(t) + a(\theta) = 0$, namely

$$t^\star(\theta) = \begin{cases} -\operatorname{sign}(a(\theta)) \big(\log(1/|a(\theta)|)\big)^{-1/2}, & a(\theta) \neq 0, \\ 0, & a(\theta) = 0. \end{cases}$$

Moreover, $|t^\star(\theta)| \leq (\log 4)^{-1/2} < 1 \leq r(\theta)$, so this critical point is feasible and therefore is the unique minimizer. Hence the optimistic minimizer is unique for every θ and is given by $x^\star(\theta) = \big(\sqrt{r(\theta)^2 - (t^\star(\theta))^2},\, t^\star(\theta)\big)$.

By Theorem 3.2, the hyper-objective F is $\mathcal{C}^1$, and because the only θ-dependence in f is through the term $a(\theta)x_2$, the envelope formula gives $F'(\theta) = a'(\theta) t^\star(\theta)$. Let $\theta_k := 10/(k\pi)$, so $a(\theta_k) = 0$ and $a'(\theta_k) \neq 0$. Then $F'(\theta_k) = 0$. Since a' is continuous and nonzero at θ_k, we have $|a'(\theta)| \asymp 1$ for θ sufficiently close to θ_k. Also, because θ_k is a simple zero of a, one has $|a(\theta)| \asymp |\theta - \theta_k|$ near θ_k. Therefore

$$|F'(\theta) - F'(\theta_k)| = |F'(\theta)| = |a'(\theta)||t^\star(\theta)| \asymp \frac{1}{\sqrt{\log(1/|\theta - \theta_k|)}} \qquad \text{as } \theta \to \theta_k.$$

Consequently, for every $\alpha > 0$,

$$\frac{|F'(\theta) - F'(\theta_k)|}{|\theta - \theta_k|^\alpha} \to \infty \qquad \text{as } \theta \to \theta_k,$$

because $|\theta - \theta_k|^\alpha \sqrt{\log(1/|\theta - \theta_k|)} \to 0$. Therefore ∇F is not locally α-Hölder continuous near θ_k for any $\alpha > 0$, even though the optimistic minimizer is unique for every θ.

The key intuition bebind the above construction is that, at every zero of a, the tangential second derivative of the restricted objective is $\phi''(0) = 0$, so $x^\star(\theta)$ is degenerate in the following sense.

Definition 3.4 (Degenerate and non-degenerate optimistic minimizers)**.** *Fix $\theta \in \Theta$ and let $x \in \mathcal{O}(\theta)$. We say that x is a* non-degenerate optimistic minimizer *if*

$$\langle v, \nabla^2_{xx} f(\theta, x) v \rangle > 0 \qquad \forall v \in \mathcal{T}_x^\theta \setminus \{0\}.$$

Equivalently, $f(\theta, \cdot)$ has strictly positive second-order derivative along every tangent direction of the solution manifold $\mathcal{S}(\theta)$ at x. If this condition fails, we call x a degenerate optimistic minimizer*. When $\mathcal{O}(\theta)$ is a singleton, we also simply say that the point $x^\star(\theta)$ is non-degenerate or degenerate.*

Non-degeneracy of the unique optimistic minimizer implies local smoothness of F. To establish the stronger stability needed for the subsequent complexity analysis, we rule out this pathological case by assuming that the unique optimistic minimizer $x^\star(\theta)$ is non-degenerate.

To prove the local smoothness of F, the main missing ingredient is to establish $x^\star(\theta) \in \mathcal{C}^1$. Once this is available, we can write $F(\theta) = f(\theta, x^\star(\theta))$ locally and then control ∇F using the regularity of f, g together with the pseudoinverse hyper-gradient formula. The next theorem provides F is locally smooth via two applications of the implicit function theorem.

Theorem 3.5 (Smoothness of the hyper-objective under unique non-degenerate optimistic minimizer)**.** *Let f be $\mathcal{C}^2$ and let g be $\mathcal{C}^3$ and satisfy Assumption 1. Fix $\theta_0 \in \Theta$ and suppose that the selected optimistic minimizer $x^\star(\theta_0)$ is unique and non-degenerate in the sense of Definition 3.4. Then there exists a neighborhood $\mathcal{U}$ of θ_0 and a (unique) $\mathcal{C}^1$ selection $x^\star : \mathcal{U} \to \mathbb{R}^d$ with $x^\star(\theta) \in \mathcal{S}(\theta)$ such that $F(\theta) = f(\theta, x^\star(\theta))$ for all $\theta \in \mathcal{U}$. Consequently, F is locally smooth.*

We now sketch the two-IFT construction behind this stronger statement.

Proof sketch. Let $x_0 := x^\star(\theta_0)$ and abbreviate $\mathcal{T}_0 := \mathcal{T}_{x_0}^{\theta_0}$, $\mathcal{N}_0 := \mathcal{N}_{x_0}^{\theta_0}$. Let $\dim(\mathcal{T}_0) = \mathtt{k}$ and choose orthonormal bases $U_{\mathcal{T}_0} \in \mathbb{R}^{d\times \mathtt{k}}$ and $U_{\mathcal{N}_0} \in \mathbb{R}^{d\times(d-\mathtt{k})}$ spanning $\mathcal{T}_0$ and $\mathcal{N}_0$.

Step 1: parameterize $\mathcal{S}(\theta)$ locally by solving the normal equations. This is exactly the same normal-chart construction as in Step 1 of the proof sketch of Theorem 3.2. In the notation introduced there, the implicit function theorem yields a $\mathcal{C}^2$ map $v = \varphi(\theta, u)$ and hence the local chart

$$\psi(\theta, u) := x_0 + U_{\mathcal{T}_0} u + U_{\mathcal{N}_0} \varphi(\theta, u).$$

On a sufficiently small neighborhood, $u \mapsto \psi(\theta, u)$ parameterizes the local branch of $\mathcal{S}(\theta)$ near x_0.

Step 2: reduce the optimistic minimization to an unconstrained problem in intrinsic coordinates. Define the pullback objective $\tilde{f}(\theta, u) := f(\theta, \psi(\theta, u))$. A point $x = \psi(\theta, u)$ is a constrained critical point of $f(\theta, \cdot)$ on $\mathcal{S}(\theta)$ iff $\nabla_u \tilde{f}(\theta, u) = 0$. At $(\theta_0, 0)$, this stationarity holds, and Assumption 3 implies that the second-order variation of $\tilde{f}(\theta_0, \cdot)$ is positive definite at $u = 0$, hence $D_u(\nabla_u \tilde{f})(\theta_0, 0)$ is invertible. Applying the implicit function theorem again yields a $\mathcal{C}^1$ map $u^*(\theta)$, unique among nearby solutions, solving $\nabla_u \tilde{f}(\theta, u^*(\theta)) = 0$. Setting $x^\star(\theta) := \psi(\theta, u^*(\theta))$ gives the desired $\mathcal{C}^1$ selection and the local identity $F(\theta) = f(\theta, x^\star(\theta))$.

Step 3: obtain the hyper-gradient by the chain rule. Finally, since $F(\theta) = f(\theta, x^\star(\theta))$, the hyper-gradient follows from the chain rule. At a constrained minimizer, the tangential component of $\nabla_x f(\theta, x^\star(\theta))$ vanishes, so only the normal component of $D_\theta x^\star(\theta)$ is needed. Differentiating the lower-level stationarity condition $\nabla_x g(\theta, x^\star(\theta)) = 0$ and applying the pseudoinverse of $\nabla^2_{xx} g(\theta, x^\star(\theta))$ yields this normal sensitivity, and substitution gives (3). ■

Global assumption on minima-selection for complexity analysis. The above discussion is entirely local: we assume that the unique and non-degenerate optimistic minimizer condition holds only in a neighborhood, and the smoothness of F established above is therefore also local. For the subsequent complexity analysis, we strengthen this assumption and require it to hold for all $\theta \in \Theta$.

Assumption 3 (Unique and non-degenerate selected optimistic minimizer)**.** *For every $\theta \in \Theta$, the selected optimistic minimizer $x^\star(\theta)$ is unique and non-degenerate in the sense of Definition 3.4.*

4 Hyper-Gradient with Minima Selection (HG-MS)

Under Assumptions 1 and 3, we describe our algorithm HG-MS, for finding a stationary point of the hyper-objective F. We need to address the following challenges in our algorithm design:

- **Optimization over an implicit, non-convex set.** The feasible region $\mathcal{S}(\theta)$ of minima-selection is a non-convex Riemannian submanifold, so minima-selection itself is a non-convex problem. Beyond the intrinsic difficulty of non-convex global optimization, here $\mathcal{S}(\theta)$ is not given explicitly (as is typical in Riemannian optimization), but only implicitly as the minimizer set of $g(\theta, \cdot)$.
- **Stabilizing the pseudo-inverse computation off-manifold.** Because the Langevin-based proxy for $x^\star(\theta)$ will generally not lie exactly on $\mathcal{S}(\theta)$, the Hessian pseudo-inverse in eq. (3) must be computed in a manner that is robust to this mismatch.

We now describe in details the two components of HG-MS tailored to the non-singleton lower-level solution set regime: (i) *Best-of-N* (BoN) selection from a Gibbs measure concentrated near $\mathcal{S}(\theta)$, used to approximate the optimistic selection rule. and (ii) a *stabilized implicit step* that computes a pseudoinverse action via a ridge-regularized linear solve using Hessian–vector products.

BoN selection near $\mathcal{S}(\theta)$. Given the nonconvex nature of minima-selection, our strategy for any fixed $\theta \in \Theta$ is to approximate the range of values taken by $f(\theta, \cdot)$ over $\mathcal{S}(\theta)$. Since $\mathcal{S}(\theta)$ is defined only implicitly, however, this cannot be done directly. To circumvent this difficulty, we consider the Gibbs measure

$$\mu_g^\lambda(\theta)(\mathrm{d}x) \;\propto\; \exp\{-g(\theta, x)/\lambda\}\,\mathrm{d}x. \tag{4}$$

Under the manifold regularity induced by PŁ$^\circ$, as $\lambda \downarrow 0$, the measure μ_θ^λ concentrates in a thin tube around $\mathcal{S}(\theta)$, with exponential decay in both the distance to $\mathcal{S}(\theta)$ and the inverse temperature $1/\lambda$. This suggests treating samples from $\mu_g^\lambda(\theta)$ as a proxy for enumerating points on $\mathcal{S}(\theta)$.

Given candidate samples $X_1, \ldots, X_N$ at the outer iterate θ_t, we approximate the optimistic rule via the following *hard selection* step:

$$i^\star \in \arg\min_{1 \le i \le N} f(\theta, X_i), \qquad \hat{x}_N^\lambda(\theta) := X_{i^\star}.$$

For the analysis, it is convenient to interpret this BoN procedure as an effective lower-tail operation at level $\delta := 1/N$, meaning that the selected point lies approximately in the best δ fraction of lower-level candidates, as measured by $f(\theta, \cdot)$. The resulting selection accuracy, together with its dependence on the geometry of $\mathcal{S}(\theta)$, is developed in the theoretical part; see Appendix F.

To sample from $\mu_g^\lambda(\theta)$, we use the standard unadjusted Langevin algorithm (ULA) [Chewi et al., 2024], which converges in $\tilde{\mathcal{O}}(\lambda^{-1})$ steps thanks to the Poincaré inequality established in [Gong et al., 2024]. Together with the exponentially fast concentration of μ_θ^λ around $\mathcal{S}(\theta)$ as $\lambda \downarrow 0$, this yields a favorable trade-off between selection accuracy and computational efficiency.

Remark 4.1. *We note that the idea of approximating $\mathcal{S}(\theta)$ via the Gibbs measure $\mu_g^\lambda(\theta)$ has also appeared in [Masiha et al., 2025]. However, that work focuses only on approximating the value of the hyper-objective F, whereas here we must approximate the solution to minima-selection, which is a substantially more challenging task. As a result, the approximation analysis requires a completely different approach.*

Hyper-gradient evaluation at the selected point. Once a proxy $\hat{x}_N^\lambda(\theta)$ has been selected, we estimate the hyper-gradient by replacing it with $x^\star(\theta)$ the formula (3). For compactness, let θ_t be the t^{th} iterate and denote $\hat{x}_N^\lambda(\theta)$ by $\tilde{x}_t$. Letting $H_t := \nabla_{xx}^2 g(\theta_t, \tilde{x}_t)$, the only nontrivial operation is the action $H_t^\dagger \nabla_x f(\theta_t, \tilde{x}_t)$. We approximate this action by solving the linear system

$$\big(H_t + \gamma I\big)\,\tilde{v}_t = \nabla_x f(\theta_t, \tilde{x}_t)$$

using conjugate gradients (CG) with Hessian–vector products, stopping when the *residual* is below η_t, i.e.,

$$\big\|\nabla_x f(\theta_t, \tilde{x}_t) - (H_t + \gamma I)\,\tilde{v}_t\big\| \le \eta_t. \tag{5}$$

where $\gamma \ge 0$ is a ridge parameter. This regularization is motivated by two issues: (i) even on the manifold, $\nabla_{xx}^2 g(\theta, \cdot)$ is singular along tangential directions (hence the appearance of a pseudoinverse), and (ii) in our algorithm $\tilde{x}_t$ is only approximately on $\mathcal{S}(\theta_t)$, so H_t can be poorly conditioned in rare cases. To control rare ill-conditioned solves far from $\mathcal{S}(\theta_t)$, we optionally apply a standard safeguard by clipping the solve output:

$$v_t \;:=\; \mathrm{Proj}_{\mathbb{B}(0;R_v)}(\tilde{v}_t),$$

for a fixed radius $R_v > 0$ (this safeguard is only used in the theoretical guarantee and can be omitted in practice). We then compute

$$\widehat{h}_t \;:=\; \nabla_\theta f(\theta_t, \tilde{x}_t) \;-\; \nabla_{\theta x}^2 g(\theta_t, \tilde{x}_t)\, v_t, \tag{6}$$

and update θ by a projected step on the feasible region Θ.

Algorithm 1 summarizes the resulting outer loop. The standard ULA sampling primitive used in its first step is deferred to Appendix A; see Algorithm 2.

5 Convergence Analysis of HG-MS

Analytic regularity and growth assumptions. In addition to the geometric structure of $\mathcal{S}(\theta)$, our analysis requires standard smoothness and growth conditions to (i) control stability constants in the

Algorithm 1 HG-MS: Hyper-gradient with minima selection

Require: initial $\theta_0 \in \Theta$; stepsizes $\{\alpha_t\}$; temperature $\lambda > 0$; number of chains N; LMC steps K; LMC stepsize h; CG tolerances $\{\eta_t\}$; (optional) ridge $\gamma \geq 0$; (optional) clip radius $R_v > 0$.

Initialize chain states $X_{0,1}^{(0)}, \ldots, X_{0,N}^{(0)}$ (e.g., i.i.d. from a prior).
for $t = 0, 1, 2, \ldots$ **do**
  for $i = 1$ to N **in parallel do**
    $X_{t,i} \leftarrow \text{LMC-ULA}(\theta_t, g, X_{t,i}^{(0)}, \lambda, K, h)$
  end for
  $i_t^\star \leftarrow \arg\min_{1 \leq i \leq N} f(\theta_t, X_{t,i})$; $\tilde{x}_t \leftarrow X_{t,i_t^\star}$.
  Approximately solve for $\tilde{v}_t$:

$$\left(\nabla_{xx}^2 g(\theta_t, \tilde{x}_t) + \gamma I\right) \tilde{v}_t = \nabla_x f(\theta_t, \tilde{x}_t)$$

  by CG (HVPs only) up to residual η_t.
  (Safeguard; optional) $v_t \leftarrow \text{Proj}_{\mathbb{B}(0;R_v)}(\tilde{v}_t)$ (otherwise set $v_t = \tilde{v}_t$).
  $\widehat{h}_t \leftarrow \nabla_\theta f(\theta_t, \tilde{x}_t) - \nabla_{\theta x}^2 g(\theta_t, \tilde{x}_t)\, v_t$.
  (Update) $\theta_{t+1} \leftarrow \text{Proj}_\Theta\big(\theta_t - \alpha_t \widehat{h}_t\big)$.
end for

bilevel objective and (ii) ensure that the Gibbs measures used in sampling-based approximations are well-defined without imposing an *a priori* bounded domain in the lower-level variable. These assumptions will be invoked in Section 4 to define Gibbs/Langevin exploration and in Section 5 to control tube-width, mixing, and stability constants in the hyper-gradient error bounds. We collect these requirements below.

Assumption 4. *The following statements hold for all $\theta \in \Theta$.*

- ***Regularity of f.*** *f is $\mathcal{C}^1$ and $L_{f,1}$-Lipschitz, and $L_{f,2}$-smooth with respect to (θ, x). Moreover, for all $x \in \mathbb{R}^d$,*

$$|f(\theta, x)| \;\leq\; C_f \|x\|^{\mathsf{n}_f} + D_f,$$

  for constants $C_f, D_f \geq 0$ and an integer $\mathsf{n}_f \geq 1$.
- ***Regularity of g.*** *For every θ, the map $x \mapsto g(\theta, x)$ is $\mathcal{C}^2$ and $L_{g,2}$-smooth (i.e., $\nabla_x g(\theta, \cdot)$ is $L_{g,2}$-Lipschitz). Moreover, for all $x \in \mathbb{R}^d$,*

$$\|\partial_\theta g(\theta, x)\| \;\leq\; C_g \|x\|^{\mathsf{n}_g} + D_g,$$

  for constants $C_g, D_g \geq 0$ and an integer $\mathsf{n}_g \geq 1$.
- ***Continuity of the lower-level Hessian.*** *The Hessian $\partial_x^2 g(\theta, x)$ is continuous with respect to (θ, x).*
- ***Quadratic growth outside a compact set.*** *There exist constants $\mathsf{D} > 0$ and $\mu_{\mathsf{qg}} > 0$ such that for all $\|x\| \geq \mathsf{D}$,*

$$\frac{\mu_{\mathsf{qg}}}{2} \operatorname{dist}^2\big(x, \mathcal{S}(\theta)\big) \;\leq\; g(\theta, x) - \min_{z \in \mathbb{R}^d} g(\theta, z).$$

  W.l.o.g. we take D large enough so that $\mathcal{S}(\theta) \subseteq \mathbb{B}_d(0; \mathsf{D})$ for all $\theta \in \Theta$.

The smoothness and Lipschitz-type conditions above are standard in the bilevel optimization literature [Kwon et al., 2023b, Chen et al., 2024]. The main distinction in our setting is that we do not restrict the lower-level variable x to a bounded domain; instead, quadratic growth beyond a compact set provides the coercivity needed to control the tails of Gibbs measures [Hasenpflug et al., 2024]. Likewise, rather than assuming boundedness of f, we allow polynomial growth, which is compatible with many learning objectives while still yielding uniform bounds in the regimes we analyze.

Under Assumption 3, the local $\mathcal{C}^1$ selections from Theorem 3.5 patch together on the compact set Θ, so the local smoothness statement globalizes. The next proposition records the resulting Lipschitz continuity of ∇F on Θ (proof deferred to Appendix D).

Proposition 5.1 (Smoothness of F)**.** *Let Assumptions 1, 3 and 4 hold and assume f is $\mathcal{C}^2$ and g is $\mathcal{C}^3$. Then there exists $L_F < \infty$ such that for all $\theta, \theta' \in \Theta$,*

$$\|\nabla F(\theta) - \nabla F(\theta')\| \;\leq\; L_F \, \|\theta - \theta'\|.$$

In particular, F is L_F-smooth on Θ.

This section gives a standard outer-loop guarantee for projected gradient descent with inexact hyper-gradients, and then bounds the hyper-gradient error e_t for HG-MS in terms of selection, sampling, and linear-solve errors.

5.1 Outer-loop guarantee

Let Algorithm 1 produce iterates $\{\theta_t\}$ and hyper-gradient estimates $\widehat{h}_t$ in (6), and define the hyper-gradient error $e_t := \widehat{h}_t - \nabla F(\theta_t)$. We also define the projected gradient mapping

$$\mathcal{G}_\Theta(\theta, \nabla F(\theta); \alpha) := \alpha^{-1}\big(\theta - \mathrm{Proj}_\Theta(\theta - \alpha \nabla F(\theta))\big).$$

The following inexact projected-gradient bound is standard when F is uniformly smooth; we state it for completeness [e.g., Ghadimi et al., 2016, Nesterov, 2003].

Proposition 5.2 (Convergence to near-stationarity with inexact hyper-gradients)**.** *Let Assumptions 1, 3 and 4 hold. Suppose that F is L_F-smooth on Θ. Let $F_\star := \inf_{\theta\in\Theta} F(\theta) > -\infty$. Initialize $\theta_0 \in \Theta$ and run the projected update $\theta_{t+1} = \mathrm{Proj}_\Theta(\theta_t - \alpha\widehat{h}_t)$ with a constant stepsize $\alpha \le 1/L_F$. Then for any $T \ge 1$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\mathcal{G}_\Theta(\theta_t, \nabla F(\theta_t); \alpha)\|^2\big] \le \frac{8\,(F(\theta_0) - F_\star)}{\alpha T} + \frac{10}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|e_t\|^2\big]. \tag{7}$$

The proof is given in Appendix D. In the smooth regime, the outer-loop analysis reduces to controlling the hyper-gradient error. Thus, (7) makes explicit that all approximation effects from HG-MS enter only through the hyper-gradient error term e_t, which we bound next.

5.2 Bounding the hyper-gradient error

Proposition 5.2 reduces the outer-loop analysis to controlling the hyper-gradient error $e_t = \widehat{h}_t - \nabla F(\theta_t)$. In this subsection we bound its second moment $\mathbb{E}\|e_t\|^2$. This error arises because HG-MS (i) evaluates the implicit hyper-gradient formula at a selected candidate $\tilde{x}_t$ instead of the ideal optimistic point $x^\star(\theta_t)$, and (ii) replaces the pseudoinverse action by a ridge-regularized, inexact linear solve (CG/HVPs) with residual tolerance η_t. In particular, at each outer iteration the CG subroutine returns $\tilde{v}_t$ satisfying the residual condition (5), which we treat as part of the algorithmic specification.

We use a standard *tube event* around $\mathcal{S}(\theta_t)$ to ensure uniform conditioning of the ridge matrix and to control how often the selected point leaves this tube; see Appendix E (ridge invertibility) and Appendix F (tube tails for Gibbs/LMC candidates) for details.

Theorem 5.3 (Second-moment hyper-gradient error bound)**.** *Let the assumptions of Proposition 5.2 hold and recall $e_t = \widehat{h}_t - \nabla F(\theta_t)$. Fix $\gamma > 0$ and define $H_t := \nabla^2_{xx} g(\theta_t, \tilde{x}_t)$ and $\mathcal{E}_t := \{\mathrm{dist}(\tilde{x}_t, \mathcal{S}(\theta_t)) \le r(\gamma)\}$, where $r(\gamma)$ is the tube radius from Appendix E (Lemma E.1). Assume the CG output $\tilde{v}_t$ satisfies* (5) *with tolerance η_t, and that the algorithm uses the clipping safeguard $v_t = \mathrm{Proj}_{\mathbb{B}(0;R_v)}(\tilde{v}_t)$. Choose the clipping radius R_v as in Appendix E (Lemma E.5) so that $v_t = \tilde{v}_t$ on $\mathcal{E}_t$. Then for all t,*

$$\mathbb{E}\|e_t\|^2 \;\le\; 3C_x^2\Big(1+\tfrac{2}{\gamma}+\tfrac{2}{\gamma(c+\gamma)}\Big)^2\,\mathbb{E}\big[\|\tilde{x}_t - x^\star(\theta_t)\|^2 \mathbf{1}_{\mathcal{E}_t}\big] + \tfrac{12C_{\mathrm{lin}}^2}{\gamma^2}\,\eta_t^2 + 3C_{\mathrm{reg}}^2\,\gamma^2 + B_e^2\,\mathbb{P}(\mathcal{E}_t^c), \tag{8}$$

where $C_x, C_{\mathrm{lin}}, C_{\mathrm{reg}} < \infty$ are stability constants (depending on global smoothness/Lipschitz bounds and the normal spectral gap of $\nabla^2_{xx} g$ on Θ), $c > 0$ is the uniform normal spectral gap constant from Proposition 2.3, and $B_e < \infty$ is a uniform off-tube bound obtained from clipping. See Appendix E (Lemmas E.3 and E.5) for explicit definitions and proofs.

The proof is given in Appendix E.1.

The bound (8) reduces control of $\mathbb{E}\|e_t\|^2$ to bounding the squared selection error $\mathbb{E}\|\tilde{x}_t - x^\star(\theta_t)\|^2$, which we do next.

5.3 Bounding the selection error

Theorem 5.3 reduces bounding $\mathbb{E}\|e_t\|^2$ to bounding the *squared selection error* $\mathbb{E}\|\tilde{x}_t - x^\star(\theta_t)\|^2$ produced by the sampling-and-hard-selection step of Algorithm 1. The next theorem records a quantitative bound (details are in Appendix F).

Theorem 5.4 (Expected squared selection error bound under Rényi-2 candidate accuracy)**.** *Fix $\theta \in \Theta$ and let Assumptions 1, 3 and 4 hold. Assume in addition that the Gibbs measure μ_θ^λ satisfies a Poincaré inequality with constant at most C_{PI}. Let $X_1, \dots, X_M$ be* independent *candidates with marginal laws $\nu_1, \dots, \nu_M$, and define $\varepsilon_{\mathrm{R}}^2 := \max_{1\le i\le N} R_2(\nu_i \| \mu_\theta^\lambda)$. Let $\tilde{x} \in \arg\min_{1\le i\le M} f(\theta, X_i)$ be the hard-selection output. Then*

$$\mathbb{E}\|\tilde{x} - x^\star(\theta)\|^2 \le 2C_{\mathrm{tube},2}\, e^{\varepsilon_{\mathrm{R}}^2/2}\, \lambda \log(1+N) + 2\left(\frac{1}{c_{\mathrm{hg}}} + \frac{4\mathsf{D}^2}{\Delta_{r_0}}\right) \times \left(2L_{f,1}\, C_{\mathrm{tube}} \sqrt{\lambda \log(1+N)} + C_1\, N^{-1/k} + 4L_{f,1}\sqrt{N\, C_{\mathrm{PI}}}\left(e^{\varepsilon_{\mathrm{R}}^2} - 1\right)^{1/2}\right). \tag{9}$$

Here $L_{f,1}$ is from Assumption 4; the remaining constants are defined in Appendix F (Lemmas F.5, F.3, F.8, F.9).

The proof is given in Appendix F. By substituting the selection-error bound (9) into the hyper-gradient stability bound (8), we obtain the following bound on the hyper-gradient error in terms of the parameters of Algorithm 1. The proof is given in Appendix D.1.

Corollary 5.5 (Hyper-gradient error in terms of algorithmic parameters)**.** *Let Assumptions 1, 3 and 4 hold and assume f is $\mathcal{C}^2$ and g is $\mathcal{C}^3$. Assume in addition that each Gibbs measure $\mu_{\theta_t}^\lambda$ satisfies a Poincaré inequality with constant at most C_{PI}. Fix $\gamma \in (0, 1]$. At iteration t, let $X_{t,1}, \dots, X_{t,N}$ be independent candidates with marginal laws $\nu_{t,1}, \dots, \nu_{t,N}$, and define $\varepsilon_{\mathrm{R}}^2 := \max_{1\le i\le N} R_2(\nu_{t,i} \| \mu_{\theta_t}^\lambda)$. Then, uniformly over t,*

$$\mathbb{E}\|e_t\|^2 = O\Bigg(\frac{\sqrt{\lambda \log(1+N)} + N^{-1/k} + \sqrt{N\, C_{\mathrm{PI}}}\left(e^{\varepsilon_{\mathrm{R}}^2} - 1\right)^{1/2} + \eta_t^2}{\gamma^2} + \gamma^2 + N e^{\varepsilon_{\mathrm{R}}^2/2} \exp\Big(-\tfrac{c_{\mathrm{tube}}}{2}\tfrac{r(\gamma)^2}{\lambda}\Big)\Bigg).$$

Finally, we derive the total first-order oracle calls for the LMC sampler, together with the function evaluations used in hard selection, when the number of LMC iterations is chosen so that the Rényi error ε_{R} is sufficiently small. The proof is given in Appendix D.1.

Corollary 5.6 (Total oracle complexity (informal))**.** *Fix $\varepsilon \in (0, 1)$. Assume the setting of Proposition 5.2 and run* HG-MS *with $\alpha = 1/L_F$. With the parameter choices from Appendix D.1 ensuring $\sup_t \mathbb{E}\|e_t\|^2 = O(\varepsilon^2)$, Proposition 5.2 yields an ε-stationarity guarantee after $T = O\big(L_F(F(\theta_0) - F_\star)/\varepsilon^2\big)$ outer iterations. Moreover, under the explicit scaling $N = \lceil \varepsilon^{-4k} \rceil$, $\lambda = \varepsilon^8/\log(1+N)$, the Rényi tolerance $\varepsilon_{\mathrm{R}} = \varepsilon^{4+2k}$, and $K = \widetilde{O}(d\, C_{\mathrm{PI}}^2 \lambda^{-2} \varepsilon_{\mathrm{R}}^{-2})$, each outer iteration uses NK evaluations of $\nabla_x g$ (LMC) and N evaluations of f (hard selection). Hence the total number of oracle calls satisfies*

$$T(NK + N) = \widetilde{O}\Big(d\, C_{\mathrm{PI}}^2\, \varepsilon^{-26-8k}\Big),$$

dominated by the sampler term TNK.

6 Empirical Evaluations

Overview. We evaluate minima selection for hyper-gradients on two bilevel learning problems. In both, the lower level fits a model by minimizing a training objective $g(\theta, \cdot)$ on $\mathcal{D}_{\mathrm{tr}}$, while the upper level evaluates a validation objective $f(\theta, \cdot)$ on a *different* distribution $\mathcal{D}_{\mathrm{val}}$. This *train–validation mismatch* creates multiple lower-level solutions with similar training loss but noticeably different validation loss. A plain hyper-gradient method differentiates through whichever lower-level solution

the optimizer happens to reach, whereas HG-MS explicitly explores multiple lower-level candidates and selects the one with the smallest validation objective, better matching the optimistic bilevel model studied in Section 3. In Section 6.1, the mismatch is due to *label noise* (training labels are corrupted but validation labels are clean) and the upper level learns one weight per training example. In Section 6.2, the mismatch is due to *class imbalance* (the training set is highly imbalanced, while validation/test are class-balanced) and the upper level tunes class-wise loss parameters to improve balanced accuracy. We compare HG-MS to a plain hyper-gradient baseline (HG-BASELINE) and to penalty/alternating baselines (V-PBGD, G-PBGD, IAPTT-GM); citations are given when introducing each method below. Unless stated otherwise, we report train/test accuracies under a fixed wall-clock budget per algorithm and summarize performance with mean $\pm$ 95% confidence intervals across independent runs. Throughout, $\mathrm{CE}(\cdot,\cdot)$ denotes cross-entropy (negative log-likelihood).

6.1 Data hyper-cleaning (MNIST label noise)

Setup. We study *data hyper-cleaning* on MNIST, a standard bilevel benchmark in which the upper-level variable assigns a soft weight to each (potentially corrupted) training label. The lower-level (follower) trains a classifier by minimizing a weighted training loss on the (possibly noisy) training set, while the upper-level (leader) minimizes the validation loss on a small clean validation set. We use a 2-layer MLP (hidden size 300) and follow the split used in our code: we subsample the MNIST *training* split and randomly partition it into 5,000 lower-level training points, 100 validation points, and 10,000 test points (all drawn from the same underlying MNIST training set). We corrupt a fraction $\rho \in \{0.4, 0.6, 0.8\}$ of the training labels uniformly at random. This yields a challenging regime (small training set with substantial label noise) in which explicit hyper-cleaning is meaningful and differences between upper-level updates become visible under a fixed time budget.

Bilevel model. Let $\theta \in \mathbb{R}^{n_{\mathrm{tr}}}$ be training-example hyper-weights and let y denote the follower network parameters. Writing $w = \sigma(\theta)$ (sigmoid reparameterization), we consider the bilevel problem

$$\min_{\theta \in \mathbb{R}^{n_{\mathrm{tr}}}} \; \min_{y \in \arg\min_z g(\theta, z)} f(\theta, y),$$

which is the optimistic formulation: for each θ, among all minimizers of the weighted training loss $g(\theta, \cdot)$ we select the one with smallest validation loss $f(\theta, \cdot)$. The lower-level objective is a weighted training loss

$$g(\theta, y) \;=\; \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} w_i \cdot \mathrm{CE}\big(h_y(x_i^{\mathrm{tr}}),\; \widetilde{c}_i^{\mathrm{tr}}\big) \;+\; \frac{\lambda}{2}\|y\|_2^2,$$

and the upper-level objective is the validation loss on clean labels,

$$f(\theta, y) \;=\; \frac{1}{n_{\mathrm{val}}} \sum_{j=1}^{n_{\mathrm{val}}} \mathrm{CE}\big(h_y(x_j^{\mathrm{val}}),\; c_j^{\mathrm{val}}\big).$$

Methods. We compare the plain hyper-gradient baseline (HG-BASELINE; implicit differentiation at the attained inner solution) [Lorraine et al., 2020] against HG-MS, and include representative penalty-based baselines V-PBGD/G-PBGD [Kwon et al., 2023b, Shen and Chen, 2023] and a trajectory-truncation alternating baseline IAPTT-GM [Xiao et al., 2023, Shaban et al., 2019]. All methods use the same model class and AdamW in both levels (see code in the repository). In this label-noise regime, many inner solutions fit the corrupted training labels similarly well; by selecting the candidate that minimizes the clean validation loss, HG-MS yields hyper-gradients that are better aligned with the target objective.

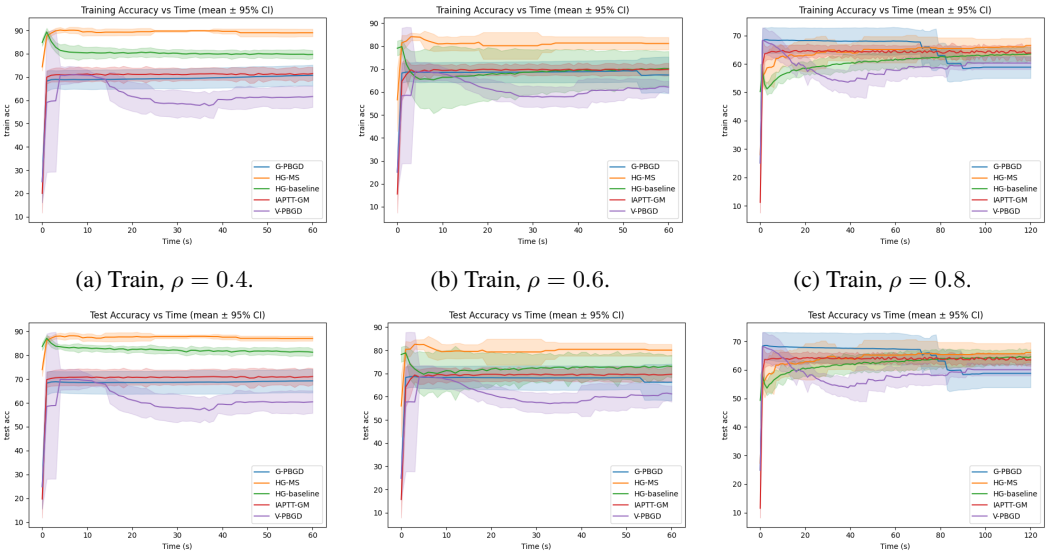
Time-budget comparison. Figure 2 reports mean and 95% confidence intervals under a fixed wall-clock budget *per algorithm* on a single NVIDIA H100 GPU (60s for $\rho \in \{0.4, 0.6\}$; 120s for $\rho = 0.8$), using 5 runs for each corruption rate. In the hardest setting $\rho = 0.8$, HG-MS improves the mean test accuracy from 64.5% (HG-BASELINE) to 66.1%, and outperforms the penalty/alternating baselines under the same time budget.

6.2 Parametric loss tuning for imbalanced data (MNIST)

Setup. We study bilevel *loss tuning* for class-imbalanced MNIST classification. From the MNIST training split, we construct an imbalanced lower-level training set by keeping each example of digit

Table 1: **Data hyper-cleaning (MNIST): train/test accuracies under different corruption rates.** Mean ± 95% CI accuracies (in %) at a fixed wall-clock budget per algorithm: 60s for $\rho \in \{0.4, 0.6\}$ (5 runs each), and 120s for $\rho = 0.8$ (5 runs).

	$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.8$	
Method	Train	Test	Train	Test	Train	Test
HG-baseline	79.69±1.78	81.22±1.45	70.19±7.56	73.09±4.86	63.48±2.48	64.54±2.62
HG-MS	**89.00±1.60**	**86.98±1.02**	**81.12±2.71**	**80.08±2.52**	**66.47±2.69**	**66.08±3.51**
V-PBGD	61.62±4.56	60.43±4.69	62.12±2.65	61.23±3.44	60.29±2.16	60.07±2.08
G-PBGD	70.73±4.48	69.21±4.93	67.40±7.94	66.26±7.86	58.84±3.87	58.77±4.93
IAPTT-GM	71.46±2.76	71.04±3.45	69.86±2.74	69.68±3.29	63.84±2.95	63.48±2.14



(a) Train, $\rho = 0.4$. (b) Train, $\rho = 0.6$. (c) Train, $\rho = 0.8$.

(d) Test, $\rho = 0.4$. (e) Test, $\rho = 0.6$. (f) Test, $\rho = 0.8$.

Figure 2: **Data hyper-cleaning (MNIST).** Train (top row) and test (bottom row) accuracies vs. time at corruption rates $\rho \in \{0.4, 0.6, 0.8\}$. Accuracies are plotted as mean ± 95% confidence intervals (fixed budget per algorithm: 60 s for $\rho \in \{0.4, 0.6\}$ with 5 runs, and 120 s for $\rho = 0.8$ with 5 runs).

$c \in \{0, \dots, 9\}$ with probability $p_c = b^c$ (we use $b = 0.3$), which yields a long-tailed label distribution in which larger digits are much rarer. We also hold out a small *class-balanced* validation set (1000 images total, roughly 100 per digit) for the upper-level objective, and we report performance on the standard MNIST test set. Because the training distribution is imbalanced while validation/test are class-balanced, we evaluate using *balanced accuracy* (macro-averaged recall). The upper-level variable parameterizes a class-wise transformation of the logits inside the cross-entropy loss,

$$\ell_{\delta,\gamma}(c;\, x) \;=\; \mathrm{CE}\Big(\, h(x) \odot \gamma + \delta,\; c \,\Big),$$

where $h(x) \in \mathbb{R}^{10}$ are the classifier logits, $\delta \in \mathbb{R}^{10}$ is a per-class bias, and $\gamma \in \mathbb{R}^{10}$ is a per-class scale (both bounded via $\tanh$ reparameterizations). The lower level minimizes the imbalanced training loss $g(\delta, \gamma, \cdot)$, while the upper level minimizes a class-balanced validation objective $f(\delta, \gamma, \cdot)$ (implemented via inverse-frequency class weights). In our implementation, the same logit shift/scale (δ, γ) is applied in both g and f, so θ affects the upper-level objective both directly and through the trained network parameters. Training uses data augmentation (denoted $\tilde{x}$), while validation and test examples are left unchanged.

Bilevel model. Let $\theta = (\delta, \gamma) \in \mathbb{R}^{10} \times \mathbb{R}^{10}$ be the loss parameters and let y denote the CNN parameters. Let n_{tr} and n_{val} denote the sizes of the (imbalanced) training set and the (class-balanced) validation set. We solve the optimistic bilevel problem

$$\min_{\theta} \; \min_{y \in \arg\min_z g(\theta, z)} f(\theta, y),$$

Table 2: **Imbalanced loss tuning (MNIST): final balanced accuracies.** Mean $\pm$ 95% CI balanced accuracy (%) at the last time point (120 s per algorithm).

Method	Train BalAcc	Test BalAcc
HG-baseline	99.46±0.20	95.48±0.77
HG-MS (ours)	**99.51±0.26**	**96.45±0.11**
V-PBGD	91.31±2.19	90.54±1.95
G-PBGD	75.88±2.17	77.40±1.41
IAPTT-GM	82.74±2.62	82.13±2.30

i.e., for each θ, among all minimizers of the training loss $g(\theta, \cdot)$ we select the one with the smallest validation loss $f(\theta, \cdot)$. The lower-level objective is

$$g(\theta, y) \;=\; \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \mathrm{CE}\big(h_y(\tilde{x}_i) \odot \gamma + \delta,\; c_i\big) \;+\; \frac{\lambda}{2}\|y\|_2^2,$$

and the upper-level objective is

$$f(\theta, y) \;=\; \frac{1}{n_{\mathrm{val}}} \sum_{j=1}^{n_{\mathrm{val}}} u_{c_j} \cdot \mathrm{CE}\big(h_y(x_j) \odot \gamma + \delta,\; c_j\big) \;+\; \frac{\lambda_\delta}{2}\|\delta\|_2^2 \;+\; \frac{\lambda_\gamma}{2}\|\gamma - \mathbf{1}\|_2^2,$$

where u_c is an inverse-frequency class weight computed on the validation set. Here $\tilde{x}_i$ denotes the train-time augmented view of x_i, while validation and test examples are uncorrupted.

Methods. We compare HG-BASELINE (plain hyper-gradient) [Lorraine et al., 2020] against HG-MS (our minima-selection variant), and include penalty-based baselines V-PBGD/G-PBGD [Kwon et al., 2023b, Shen and Chen, 2023] and a trajectory-truncation alternating baseline IAPTT-GM [Xiao et al., 2023, Shaban et al., 2019]. For HG-MS, we maintain a small ensemble of lower-level candidates trained on $g(\cdot)$ (one "base" run plus 4 LMC branches with slightly perturbed optimizer settings) and select the candidate that minimizes the validation objective $f(\cdot)$ at each outer step. Intuitively, when the training distribution is imbalanced but validation/test are balanced, different follower solutions can achieve similar training loss yet differ in balanced accuracy; selecting the best candidate for f yields a more informative hyper-gradient than differentiating through a single, potentially suboptimal solution.

Time-budget comparison. Figure 3 reports mean and 95% confidence intervals under a fixed wall-clock budget of 120 s *per algorithm* on a single NVIDIA H100 GPU. In this regime, HG-MS improves test balanced accuracy by about 0.97 percentage points over HG-BASELINE at the same wall-clock budget.

References

Pierre Ablin, Gabriel Peyré, and Thomas Moreau. Super-efficiency of automatic differentiation for functions defined as a minimum. In *International Conference on Machine Learning*, pages 32–41. PMLR, 2020.

Michael Arbel and Julien Mairal. Amortized implicit differentiation for stochastic bilevel optimization. *arXiv preprint arXiv:2111.14580*, 2021.

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

Adi Ben-Israel and Thomas N. E. Greville. *Generalized Inverses: Theory and Applications*. Springer, 2 edition, 2003.

Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018.

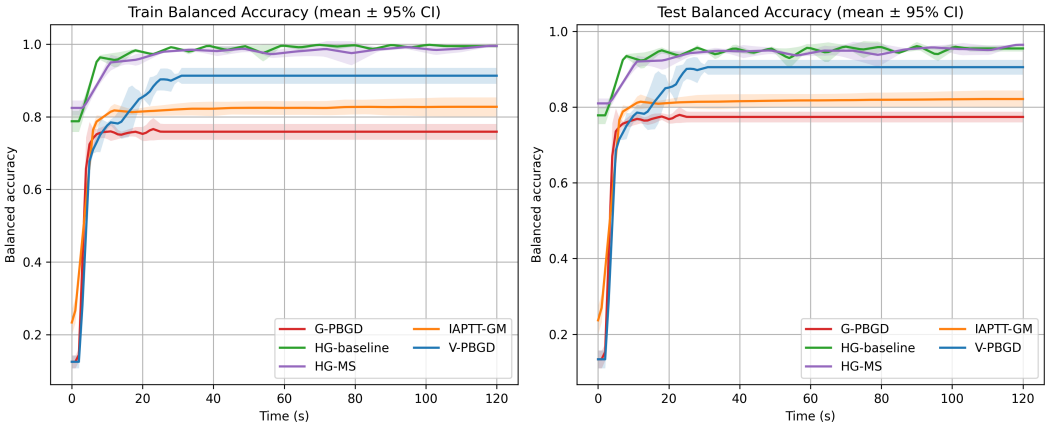


Figure 3: **Imbalanced loss tuning (MNIST).** Train/test balanced accuracy vs. time (mean and 95% CI; 120s per algorithm).

Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. *Advances in neural information processing systems*, 35:5230–5242, 2022.

Jérôme Bolte, Quoc-Tung Le, Edouard Pauwels, and Samuel Vaiter. Bilevel gradient methods and morse parametric qualification. *arXiv preprint arXiv:2502.09074*, 2025.

He Chen, Jiajin Li, and Anthony Man-Cho So. Set smoothness unlocks clarke hyper-stationarity in bilevel optimization. *arXiv preprint arXiv:2506.04587*, 2025.

Lesi Chen, Jing Xu, and Jingzhao Zhang. On finding small hyper-gradients in bilevel optimization: Hardness results and improved analysis. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 947–980. PMLR, 2024.

Sinho Chewi, Murat A Erdogdu, Mufan Li, Ruoqi Shen, and Matthew S Zhang. Analysis of langevin monte carlo from poincare to log-sobolev. *Foundations of Computational Mathematics*, pages 1–51, 2024.

Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Annals of Operations Research*, 153(1):235–256, 2007.

Chris Criscitiello, Quentin Rebjock, and Nicolas Boumal. If a smooth function is globally PŁ and coercive, then it has a unique minimizer, 2025. URL `www.racetothebottom.xyz/posts/PL-smooth-unique/`.

Stephan Dempe. *Foundations of bilevel programming*. Springer, 2002.

Stephan Dempe and Alain Zemkoho. Bilevel optimization. In *Springer optimization and its applications*, volume 161. Springer, 2020.

Stephan Dempe and Alain B Zemkoho. On the karush–kuhn–tucker reformulation of the bilevel optimization problem. *Nonlinear Analysis: Theory, Methods & Applications*, 75(3):1202–1218, 2012.

Stephan Dempe, Joydeep Dutta, and BS Mordukhovich. New necessary optimality conditions in optimistic bilevel programming. *Optimization*, 56(5-6):577–604, 2007.

Asen L Dontchev and R Tyrrell Rockafellar. *Implicit functions and solution mappings*, volume 543. Springer, 2009.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pages 1309–1318. PMLR, 2018.

Matthias Feurer and Frank Hutter. Hyperparameter optimization. *Automated machine learning: Methods, systems, challenges*, pages 3–33, 2019.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

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

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

Gene H Golub and Charles F Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4 edition, 2013.

Yun Gong, Niao He, and Zebang Shen. Poincare inequality for local log-polyak-lojasiewicz measures: Non-asymptotic analysis in low-temperature regime. *arXiv preprint arXiv:2501.00429*, 2024.

Mareike Hasenpflug, Daniel Rudolf, and Björn Sprungk. Wasserstein convergence rates of increasingly concentrating probability measures. *The Annals of Applied Probability*, 34(3):3320–3347, 2024.

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

Feihu Huang. Optimal hessian/jacobian-free nonconvex-pl bilevel optimization. *arXiv preprint arXiv:2407.17823*, 2024.

Chengtao Jian, Kai Yang, Tianhao Gao, Wuguang Ni, Keying Yang, Bowen Xiao, Jiajun Liu, and Ye Ouyang. Stable preference optimization: A bilevel approach to catastrophic preference shift. *arXiv preprint arXiv:2507.07723*, 2025. URL `https://arxiv.org/abs/2507.07723`.

Thomas Kleinert, Martine Labbé, Fränk Plein, and Martin Schmidt. There's no free lunch: on the hardness of choosing a correct big-m in bilevel optimization. *Operations research*, 68(6): 1716–1721, 2020.

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

Jeongyeol Kwon, Dohyun Kwon, Steve Wright, and Robert Nowak. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. *arXiv preprint arXiv:2309.01753*, 2023b.

John M Lee. *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media, 2006.

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.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59: 85–116, 2022b.

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

Risheng Liu, Jiaxin Gao, Jin Zhang, Deyu Meng, and Zhouchen Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10045–10067, 2021.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International conference on artificial intelligence and statistics*, pages 1540–1552. PMLR, 2020.

Zhi-Quan Luo, Jong-Shi Pang, and Daniel Ralph. *Mathematical programs with equilibrium constraints*. Cambridge University Press, 1996.

Saeed Masiha, Zebang Shen, Negar Kiyavash, and Niao He. Superquantile-gibbs relaxation for minima-selection in bi-level optimization. *arXiv preprint arXiv:2505.05991*, 2025.

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

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Quynh Nguyen. On connected sublevel sets in deep learning. In *International conference on machine learning*, pages 4790–4799. PMLR, 2019.

Rui Pan, Dylan Zhang, Hanning Zhang, Xingyuan Pan, Minrui Xu, Jipeng Zhang, Renjie Pi, Xiaoyu Wang, and Tong Zhang. ScaleBiO: Scalable bilevel optimization for LLM data reweighting. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31959–31982, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1543. URL https://aclanthology.org/2025.acl-long.1543/.

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

Quentin Rebjock and Nicolas Boumal. Fast convergence to non-isolated minima: four equivalent conditions for c 2 functions. *Mathematical Programming*, pages 1–49, 2024.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pages 4334–4343. PMLR, 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*, pages 1723–1732. PMLR, 2019.

H. Shen, Q. Xiao, and T. Chen. On penalty-based bilevel gradient descent method. *Mathematical Programming*, 2025a. doi: 10.1007/s10107-025-02194-4. URL https://doi.org/10.1007/s10107-025-02194-4.

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

Han Shen, Pin-Yu Chen, Payel Das, and Tianyi Chen. SEAL: safety-enhanced aligned LLM fine-tuning via bilevel data selection. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b. URL https://openreview.net/forum?id=VHguhvcoM5.

Luca Venturi, Afonso S Bandeira, and Joan Bruna. Spurious valleys in two-layer neural network optimization landscapes. *arXiv preprint arXiv:1802.06384*, 2018.

Luis N Vicente and Paul H Calamai. Bilevel and multilevel programming: A bibliography review. *Journal of Global optimization*, 5(3):291–306, 1994.

Quan Xiao, Songtao Lu, and Tianyi Chen. A generalized alternating method for bilevel optimization under the polyak-łojasiewicz condition. *arXiv preprint arXiv:2306.02422*, 2023.

Mengwei Xu and Jane J Ye. A smoothing augmented lagrangian method for solving simple bilevel programs. *Computational Optimization and Applications*, 59:353–377, 2014.

Jane J Ye and XY Ye. Necessary optimality conditions for optimization problems with variational inequality constraints. *Mathematics of Operations Research*, 22(4):977–997, 1997.

Jane J Ye and Daoli Zhu. New necessary optimality conditions for bilevel programs by combining the mpec and value function approaches. *SIAM Journal on Optimization*, 20(4):1885–1905, 2010.

Jane J Ye and DL Zhu. Optimality conditions for bilevel programming problems. *Optimization*, 33 (1):9–27, 1995.

JJ Ye, DL Zhu, and Qiji Jim Zhu. Exact penalization and necessary optimality conditions for generalized bilevel programming problems. *SIAM Journal on optimization*, 7(2):481–507, 1997.

Alain B Zemkoho and Shenglong Zhou. Theoretical and numerical comparison of the karush–kuhn–tucker and value function reformulations in bilevel optimization. *Computational Optimization and Applications*, 78(2):625–674, 2021.

Algorithm 2 LMC-ULA$(\theta, g, x^{(0)}, \lambda, K, h)$

Require: θ; lower objective $g(\theta, \cdot)$; initial state $x^{(0)}$; temperature λ; steps K; stepsize h.
1: $x \leftarrow x^{(0)}$.
2: **for** $s = 0$ to $K - 1$ **do**
3: Draw $\xi^{(s)} \sim \mathcal{N}(0, I_d)$.
4: $x \leftarrow x - h\nabla_x g(\theta, x) + \sqrt{2\lambda h}\, \xi^{(s)}$.
5: **end for**
6: **return** x.

A Standard ULA Sampling Primitive

For completeness, we record the standard unadjusted Langevin algorithm (ULA) used inside HG-MS; see, e.g., Chewi et al. [2024]. This is the sampling subroutine invoked in Step 3 of Algorithm 1.

B Related Work

Bilevel programming and set-valued lower-level solutions. Bilevel optimization has a long history in mathematical programming; see, e.g., Vicente and Calamai [1994], Dempe [2002], Colson et al. [2007], Luo et al. [1996], Dempe and Zemkoho [2020]. Classical approaches often reduce bilevel problems to single-level programs, either via value-function reformulations or by replacing the lower-level problem with its KKT/MPEC optimality system [Ye and Zhu, 1995, 2010, Dempe and Zemkoho, 2012, Zemkoho and Zhou, 2021]. Such reductions can lead to nonsmooth constraints or complementarity conditions, and practical implementations often rely on penalty, smoothing, or big-M relaxations [Xu and Ye, 2014, Kleinert et al., 2020]. In machine-learning applications, however, one typically seeks scalable gradient-based methods, which requires an explicit handle on how the lower-level solution depends on θ.

Hyper-gradient methods under singleton lower-level minimizers. Most gradient-based bilevel methods in machine learning operate in regimes where the lower-level solution map is single-valued, so that the hyper-objective is differentiable and hyper-gradients can be computed by implicit differentiation or truncated backpropagation [Pedregosa, 2016, Franceschi et al., 2017, Lorraine et al., 2020, Shaban et al., 2019]. This setting also underlies a growing literature on differentiable optimization layers and modular implicit differentiation frameworks [Ablin et al., 2020, Arbel and Mairal, 2021, Blondel et al., 2022]. While these works differ in how they approximate the implicit step (e.g., Hessian-free or fully first-order surrogates), they largely share the assumption that the lower-level minimizer is unique, so minima selection is absent.

Relaxing strong convexity via global regularity and penalties. Several recent analyses relax strong convexity by assuming global PŁ / error-bound conditions and deriving convergence guarantees for bilevel gradient methods, often through penalty perturbations or regularity of the perturbed objective $g + \sigma f$ [Chen et al., 2024, Kwon et al., 2023b, Huang, 2024, Liu et al., 2022a, Kwon et al., 2023a]. As discussed in Appendix B.1, when combined with standard boundedness/coercivity assumptions these global conditions typically rule out bounded non-singleton minimizer sets for smooth lower-level objectives [Criscitiello et al., 2025, Masiha et al., 2025], effectively reducing the optimistic formulation to the classical singleton-minimizer regime. Other lines of work obtain differentiability or algorithmic guarantees under convexity of the lower-level objective in x [Xiao et al., 2023, Shen et al., 2025a], which excludes many practically important nonconvex settings.

Minima selection beyond the singleton regime. When the lower-level solution set is non-singleton, the optimistic (and pessimistic) bilevel objectives are inherently selection problems over $\mathcal{S}(\theta)$ [Dempe et al., 2007, Ye et al., 1997, Ye and Ye, 1997], and the induced hyper-objective can be nonsmooth even for smooth f, g due to branch switching. Recent work has proposed to control non-uniqueness through explicit selection maps or qualification conditions for parameterized critical points [Arbel and Mairal, 2022, Bolte et al., 2025], and to target generalized stationarity notions for the resulting nonsmooth objectives [Chen et al., 2025]. Complementary to these directions, our prior work [Masiha et al., 2025] introduces a superquantile–Gibbs relaxation that turns minima selection into sampling

from a Gibbs measure and yields dimension-aware complexity guarantees for optimizing a nonsmooth Lipschitz hyper-objective. The present paper instead establishes *classical differentiability* of the optimistic hyper-objective in the manifold-minimizer regime under local PŁ$^\circ$ and a local identifiability assumption, yielding a correct pseudoinverse hyper-gradient and a practical hyper-gradient method with explicit minima selection.

B.1 Comparison with penalty-based differentiability analyses

We compare our standing assumptions to penalty-based differentiability analyses such as Kwon et al. [2023b] and Chen et al. [2024], which study the perturbed lower-level objective

$$h_\sigma(\theta, x) \;:=\; g(\theta, x) + \sigma f(\theta, x), \qquad \sigma > 0.$$

What is assumed in penalty-based differentiability works. A representative sufficient condition in this line of work is a σ-*uniform* error-bound / PŁ-type regularity of $h_\sigma(\theta, \cdot)$ in the lower-level variable x for all $\sigma \in [0, \sigma_0]$, with constants that do not degenerate as $\sigma \downarrow 0$. For example, Kwon et al. [2023b] assume a (small-error) proximal error-bound condition for $h_\sigma(\theta, \cdot)$ that holds uniformly for all $\sigma \in [0, \sigma_0]$, together with coercivity and smoothness assumptions. In the smooth unconstrained setting, such a proximal error bound is equivalent to a PŁ inequality [Chen et al., 2024, Prop. C.1]. In particular, since $\sigma = 0$ is included, these assumptions enforce a PŁ / error-bound condition for $g(\theta, \cdot)$ itself.

Implication: singleton lower-level minimizers. For C^2 objectives, a global PŁ condition implies that the minimizer set is either a singleton or unbounded [Masiha et al., 2025]. Therefore, under standard assumptions that rule out unbounded minimizers (e.g., coercivity, or compactness of $\mathcal{S}(\theta)$), the lower-level minimizer must be unique [Criscitiello et al., 2025]. In this regime, the optimistic bilevel objective reduces to the classical singleton-minimizer setting, where minima-selection is redundant. Several recent analyses in nonconvex bilevel optimization fall into this category (global PŁ / error bounds together with coercivity/compactness) and hence effectively assume a singleton lower-level minimizer; see, e.g., Huang [2024], Kwon et al. [2023b]. Other works enforce uniqueness more directly (without explicitly invoking global PŁ), e.g., Liu et al. [2022a].

What our assumptions imply about h_σ. Our standing assumptions do *not* require $h_\sigma(\theta, \cdot)$ to satisfy a σ-uniform PŁ/error-bound condition. Nevertheless, when PŁ$^\circ$ holds and the minima-selection problem $\min_{x \in \mathcal{S}(\theta)} f(\theta, x)$ has a unique nondegenerate minimizer $x^\star(\theta)$ (Assumption 3), the penalized objective $h_\sigma(\theta, \cdot) = g(\theta, \cdot) + \sigma f(\theta, \cdot)$ is locally well-behaved for each fixed (small) $\sigma > 0$, but with conditioning that deteriorates as $\sigma \downarrow 0$. The next lemma makes this precise.

Lemma B.1 (Local PŁ for h_σ in the manifold regime (non-uniform in σ))**.** *Fix $\theta \in \Theta$ and abbreviate $g(x) := g(\theta, x)$, $f(x) := f(\theta, x)$, and $h_\sigma(x) := g(x) + \sigma f(x)$. Assume Assumptions 1 and 3 hold at θ, and let $x^\star := x^\star(\theta) \in \mathcal{S}(\theta)$ be the unique optimistic minimizer. Define the* normal curvature *and* tangential curvature *constants*

$$c_\perp \;:=\; \min_{\substack{v \in \mathcal{N}^\theta_{x^\star} \\ \|v\|=1}} \big\langle v,\, \nabla^2_{xx} g(x^\star)\, v \big\rangle \;>\; 0, \qquad c_\parallel \;:=\; \min_{\substack{v \in \mathcal{T}^\theta_{x^\star} \\ \|v\|=1}} \big\langle v,\, \mathrm{Hess}_{\mathcal{S}(\theta)}\big(f|_{\mathcal{S}(\theta)}\big)(x^\star)[v] \big\rangle \;>\; 0,$$

where the second quantity is positive by Assumption 3. Then there exist $\sigma_0 > 0$ and $r > 0$ such that for every $\sigma \in (0, \sigma_0]$:

1. *h_σ has a unique critical point x_σ in $\mathbb{B}_d(x^\star; r)$, and this point is the unique minimizer of h_σ on $\mathbb{B}_d(x^\star; r)$.*
2. *h_σ satisfies a local PŁ inequality on $\mathbb{B}_d(x^\star; r)$ with a constant μ_σ that can be chosen as*

$$\mu_\sigma \;:=\; \tfrac{1}{16}\, \min\{c_\perp,\, \sigma c_\parallel\}. \tag{10}$$

That is, for all $x \in \mathbb{B}_d(x^\star; r)$,

$$h_\sigma(x) - h_\sigma(x_\sigma) \;\le\; (2\mu_\sigma)^{-1}\, \|\nabla h_\sigma(x)\|^2.$$

In particular, if $\dim(\mathcal{S}(\theta)) > 0$, then $\mu_\sigma = \Theta(\sigma)$ as $\sigma \downarrow 0$.

Proof. We sketch the standard local strong-convexity mechanism behind (10); the key point is that PŁ$^\circ$ supplies curvature in directions normal to $\mathcal{S}(\theta)$, while Assumption 3 supplies curvature along the manifold.

Step 1: Normal stationarity can be solved by the IFT. Let $k := \dim(\mathcal{S}(\theta))$ and set $\mathcal{T} := \mathcal{T}^{\theta}_{x^\star}$, $\mathcal{N} := \mathcal{N}^{\theta}_{x^\star}$. Choose orthonormal matrices $U_{\mathcal{T}} \in \mathbb{R}^{d\times k}$ and $U_{\mathcal{N}} \in \mathbb{R}^{d\times(d-k)}$ spanning $\mathcal{T}$ and $\mathcal{N}$. For (u, v) near $(0, 0)$, define $x(u, v) := x^\star + U_{\mathcal{T}}u + U_{\mathcal{N}}v$ and

$$\Phi_\sigma(u, v) \;:=\; U_{\mathcal{N}}^\top \nabla h_\sigma(x(u, v)) \in \mathbb{R}^{d-k}.$$

Since $x^\star \in \mathcal{S}(\theta)$, $\nabla g(x^\star) = 0$, hence $\Phi_0(0, 0) = 0$. Moreover,

$$D_v\Phi_0(0, 0) \;=\; U_{\mathcal{N}}^\top \nabla^2_{xx} g(x^\star)\, U_{\mathcal{N}}.$$

By PŁ$^\circ$(see Proposition 2.3), this matrix is SPD and its minimal eigenvalue is $c_\perp$. By continuity of $\nabla^2_{xx} g$ and $\nabla^2_{xx} f$, there exist $r > 0$ and $\sigma_0 > 0$ such that for all $\sigma \in [0, \sigma_0]$ and all (u, v) with $\|u\| + \|v\| \le r$,

$$\lambda_{\min}\big(D_v\Phi_\sigma(u, v)\big) \;\ge\; \tfrac{c_\perp}{2}. \tag{11}$$

Therefore, by the implicit-function theorem applied to $\Phi_\sigma(u, v) = 0$, there exists a unique C^1 map $v = \varphi_\sigma(u)$ (defined for $\|u\| \le r$) such that $\Phi_\sigma(u, \varphi_\sigma(u)) = 0$. Define the corresponding "normal-stationarity" manifold $\mathcal{M}_\sigma := \{x(u, \varphi_\sigma(u)) : \|u\| \le r\}$.

Step 2: Tangential curvature is $\Theta(\sigma)$. Define the reduced objective $\tilde{h}_\sigma(u) := h_\sigma(x(u, \varphi_\sigma(u)))$. By the chain rule and $\Phi_\sigma(u, \varphi_\sigma(u)) = 0$,

$$\nabla\tilde{h}_\sigma(u) \;=\; U_{\mathcal{T}}^\top \nabla h_\sigma(x(u, \varphi_\sigma(u))).$$

Thus $\nabla\tilde{h}_\sigma(u) = 0$ iff $\nabla h_\sigma(x(u, \varphi_\sigma(u))) = 0$, i.e., critical points of h_σ in the neighborhood correspond to critical points of $\tilde{h}_\sigma$.

At $\sigma = 0$, $\tilde{h}_0 \equiv \min g$ is locally constant because $\mathcal{M}_0 \subseteq \mathcal{S}(\theta)$. For $\sigma > 0$, $\tilde{h}_\sigma = \tilde{h}_0 + \sigma\,\tilde{f}_\sigma$ where $\tilde{f}_\sigma(u) := f(x(u, \varphi_\sigma(u)))$. Moreover, $\varphi_\sigma(u)$ is a C^1 perturbation of $\varphi_0(u)$ and $\varphi_0(\cdot)$ parametrizes $\mathcal{S}(\theta)$ near $x^\star$ (as in the first implicit-function step of Theorem 3.5 with θ fixed). Therefore, the Hessian $\nabla^2\tilde{f}_\sigma(0)$ converges to the intrinsic Hessian of the restriction $f|_{\mathcal{S}(\theta)}$ at $x^\star$ as $\sigma \downarrow 0$. By Assumption 3, this limiting intrinsic Hessian is positive definite with minimal eigenvalue $c_\| > 0$, hence for $\sigma \le \sigma_0$ (shrinking σ_0 if needed),

$$\lambda_{\min}\big(\nabla^2\tilde{h}_\sigma(0)\big) \;=\; \sigma\,\lambda_{\min}\big(\nabla^2\tilde{f}_\sigma(0)\big) \;\ge\; \tfrac{\sigma c_\|}{2}. \tag{12}$$

By continuity, (12) holds on a small ball $\|u\| \le r$ as well, so $\tilde{h}_\sigma$ is strongly convex there and has a unique minimizer u_σ. Set $x_\sigma := x(u_\sigma, \varphi_\sigma(u_\sigma)) \in \mathcal{M}_\sigma$. By construction, x_σ is the unique critical point of h_σ in $\mathbb{B}_d(x^\star; r)$, and it is the unique minimizer of h_σ on $\mathbb{B}_d(x^\star; r)$.

Step 3: From curvature to a local PŁ inequality. Fix $\sigma \in (0, \sigma_0]$ and a point $x = x(u, v)$ with $\|u\| + \|v\| \le r$. Write the Hessian in the orthonormal basis $U := [U_{\mathcal{T}}\ U_{\mathcal{N}}]$ as the block matrix

$$H(x) \;:=\; U^\top \nabla^2 h_\sigma(x)\, U \;=\; \begin{pmatrix} H_{uu}(x) & H_{uv}(x) \\ H_{vu}(x) & H_{vv}(x) \end{pmatrix}.$$

By definition, $H_{vv}(x) = D_v\Phi_\sigma(u, v)$, so (11) implies $H_{vv}(x) \succeq (c_\perp/2)I$, hence $\|H_{vv}(x)^{-1}\| \le 2/c_\perp$. Moreover, the Hessian of the reduced objective $\tilde{h}_\sigma$ is the Schur complement of H_{vv}:

$$\nabla^2\tilde{h}_\sigma(u) \;=\; H_{uu}(x) - H_{uv}(x)\, H_{vv}(x)^{-1} H_{vu}(x).$$

By (12) (and shrinking r if needed), we have $\nabla^2\tilde{h}_\sigma(u) \succeq (\sigma c_\|/2)I$ for $\|u\| \le r$. Therefore, using the standard block factorization

$$H(x) \;=\; \begin{pmatrix} I & H_{uv}(x)H_{vv}(x)^{-1} \\ 0 & I \end{pmatrix}^\top \begin{pmatrix} \nabla^2\tilde{h}_\sigma(u) & 0 \\ 0 & H_{vv}(x) \end{pmatrix} \begin{pmatrix} I & H_{uv}(x)H_{vv}(x)^{-1} \\ 0 & I \end{pmatrix},$$

we obtain, for all $z \in \mathbb{R}^d$,

$$z^\top H(x) z \;\ge\; \frac{1}{\|R^{-1}(x)\|^2} \cdot \min\left\{\tfrac{\sigma c_\|}{2},\, \tfrac{c_\perp}{2}\right\} \|z\|^2, \qquad R(x) := \begin{pmatrix} I & H_{uv}(x)H_{vv}(x)^{-1} \\ 0 & I \end{pmatrix}.$$

Finally, $H_{uv}(x) = U_{\mathcal{T}}^\top \nabla^2 h_\sigma(x) U_{\mathcal{N}}$ and $\nabla^2 h_\sigma = \nabla^2 g + \sigma\nabla^2 f$. Since $U_{\mathcal{T}}^\top \nabla^2 g(x^\star) U_{\mathcal{N}} = 0$ (tangent vectors lie in $\ker \nabla^2 g(x^\star)$) and the Hessians are continuous, we may shrink r (independently of σ)

so that $\|U_{\mathcal{T}}^\top \nabla^2 g(x) U_{\mathcal{N}}\| \le c_\perp/4$ on $\mathbb{B}_d(x^\star; r)$. Together with the uniform bound $\|\nabla^2 f(x)\| \le L_{f,2}$ from Assumption 4, this yields

$$\|H_{uv}(x) H_{vv}(x)^{-1}\| \;\le\; \frac{2}{c_\perp}\Big(\frac{c_\perp}{4} + \sigma L_{f,2}\Big) \;\le\; \frac{3}{4}$$

for all $\sigma \in (0, \sigma_0]$ (shrinking σ_0 if needed). Hence $\|R^{-1}(x)\| \le 1 + \|H_{uv}(x)H_{vv}(x)^{-1}\| \le 2$, and so

$$\nabla^2 h_\sigma(x) \;\succeq\; \tfrac{1}{8}\,\min\{c_\perp,\, \sigma c_\parallel\}\, I \qquad \text{for all } x \in \mathbb{B}_d(x^\star; r).$$

In particular, h_σ is m_σ-strongly convex on $\mathbb{B}_d(x^\star; r)$ with $m_\sigma := \frac{1}{8}\min\{c_\perp, \sigma c_\parallel\}$. For an m_σ-strongly convex differentiable function on a convex domain, one has the (local) PŁ inequality

$$h_\sigma(x) - h_\sigma(x_\sigma) \;\le\; \frac{1}{2m_\sigma}\,\|\nabla h_\sigma(x)\|^2,$$

which follows by applying strong convexity with $y = x_\sigma$ and maximizing the resulting upper bound over $\|x - x_\sigma\|$. This gives the claim with $\mu_\sigma := m_\sigma/2$, which is exactly (10). ■

Why our assumptions are weaker than σ-uniform local PŁ for $g + \sigma f$. In the compact manifold regime, the best possible local PŁ constant for $h_\sigma(\theta, \cdot)$ necessarily deteriorates as $\sigma \downarrow 0$ whenever $\mathcal{S}(\theta)$ is non-singleton and $f(\theta, \cdot)$ is non-constant along $\mathcal{S}(\theta)$. Indeed, one can upper bound any admissible PŁ constant by testing the inequality on points of the minimizer manifold.

Lemma B.2 (No σ-uniform local PŁ constant on neighborhoods intersecting a non-singleton $\mathcal{S}(\theta)$)**.** *Fix θ and let $x^\star \in \mathcal{S}(\theta)$ be the unique optimistic minimizer. Assume there exists $\bar{x} \in \mathcal{S}(\theta)$ such that $f(\theta, \bar{x}) > f(\theta, x^\star)$ (i.e., $f(\theta, \cdot)$ is not constant on $\mathcal{S}(\theta)$). Let $U \subset \mathbb{R}^d$ be any neighborhood containing both $x^\star$ and $\bar{x}$. If $h_\sigma(\theta, \cdot)$ satisfies a (local) PŁ inequality on U with constant $\mu_\sigma > 0$, i.e.,*

$$h_\sigma(\theta, x) - \min_{y \in U} h_\sigma(\theta, y) \;\le\; (2\mu_\sigma)^{-1}\,\|\nabla_x h_\sigma(\theta, x)\|^2 \qquad \forall x \in U,$$

then

$$\mu_\sigma \;\le\; \frac{\sigma}{2} \cdot \frac{\|\nabla_x f(\theta, \bar{x})\|^2}{f(\theta, \bar{x}) - f(\theta, x^\star)}. \tag{13}$$

In particular, $\mu_\sigma = O(\sigma)$ as $\sigma \downarrow 0$, so no constant bounded away from 0 can hold uniformly over $\sigma \in (0, \sigma_0]$ unless $\mathcal{S}(\theta)$ is (locally) a singleton or $f(\theta, \cdot)$ is constant on $\mathcal{S}(\theta)$.

Proof. Fix such a neighborhood U and point $\bar{x} \in \mathcal{S}(\theta) \cap U$. Since $x^\star \in U$, we have $\min_{y \in U} h_\sigma(\theta, y) \le h_\sigma(\theta, x^\star)$, hence

$$h_\sigma(\theta, \bar{x}) - h_\sigma(\theta, x^\star) \;\le\; h_\sigma(\theta, \bar{x}) - \min_{y \in U} h_\sigma(\theta, y).$$

Applying the PŁ inequality at $x = \bar{x}$ and using that $\bar{x} \in \mathcal{S}(\theta)$ (so $\nabla_x g(\theta, \bar{x}) = 0$ and $g(\theta, \bar{x}) = g(\theta, x^\star)$) gives

$$\sigma\big(f(\theta, \bar{x}) - f(\theta, x^\star)\big) \;\le\; (2\mu_\sigma)^{-1}\,\|\sigma \nabla_x f(\theta, \bar{x})\|^2,$$

which rearranges to (13). ■

C Proof of Section 3

This appendix provides the proofs for Section 3.

C.1 An explicit degenerate-selection example

This subsection provides the construction referenced in Section 3: the optimistic minimizer is unique for every θ, yet the selection map $\theta \mapsto x^\star(\theta)$ is not differentiable at infinitely many points because the minimum of $f(\theta, \cdot)$ over $\mathcal{S}(\theta)$ is degenerate at those parameters.

Example C.1 (Unique minimizer but non-differentiable selection)**.** *Let $\theta \in \mathbb{R}$ and $x = [x_1, x_2] \in \mathbb{R}^2$. Define the smooth function $\rho : \mathbb{R} \to \mathbb{R}$ by*

$$\rho(s) := \begin{cases} e^{-1/s}, & s > 0, \\ 0, & s \le 0. \end{cases}$$

Let $r(\theta) := \sqrt{1+\theta^2}$ and define the lower-level objective

$$g(\theta, x) \;:=\; \big(\|x\|^2 - r(\theta)^2\big)^2.$$

Then $g(\theta,\cdot) \geq 0$ and its minimizer set is the circle

$$\mathcal{S}(\theta) = \arg\min_x g(\theta, x) = \{x \in \mathbb{R}^2 :\; \|x\| = r(\theta)\},$$

which is a compact embedded manifold (without boundary) that depends on θ.

Next define $a(\theta) := e^{-1/\theta^2}\sin(10/\theta)$ for $\theta \neq 0$ and $a(0) = 0$. Define the upper-level objective

$$f(\theta, x) \;:=\; x_2^4 + a(\theta)\, x_2 + \rho(-x_1).$$

Since $\rho(-x_1) = 0$ when $x_1 \geq 0$ and $\rho(-x_1) > 0$ when $x_1 < 0$, any minimizer of $f(\theta,\cdot)$ on $\mathcal{S}(\theta)$ must satisfy $x_1 \geq 0$: if $x \in \mathcal{S}(\theta)$ has $x_1 < 0$, then reflecting it to $\bar{x} := (-x_1, x_2) \in \mathcal{S}(\theta)$ yields $f(\theta,\bar{x}) < f(\theta, x)$. Therefore the minimizer lies on the right semicircle $\{x \in \mathcal{S}(\theta) : x_1 \geq 0\}$, where $\rho(-x_1) \equiv 0$.

On this semicircle, minimizing f reduces to the one-dimensional problem

$$\min_{t \in [-r(\theta),\, r(\theta)]} \big(t^4 + a(\theta)\, t\big),$$

whose unique minimizer is

$$t^\star(\theta) = -\sqrt[3]{\frac{a(\theta)}{4}}.$$

Since $|a(\theta)| \leq 1$ and $r(\theta) \geq 1$, we have $|t^\star(\theta)| \leq 4^{-1/3} < r(\theta)$, so this minimizer lies in the interior. Thus the optimistic minimizer on $\mathcal{S}(\theta)$ is unique and given by

$$x^\star(\theta) = \big(\sqrt{r(\theta)^2 - (t^\star(\theta))^2},\; t^\star(\theta)\big).$$

The corresponding optimistic value is

$$F(\theta) = \min_{x \in \mathcal{S}(\theta)} f(\theta, x) = \min_{t \in [-r(\theta),\, r(\theta)]} \big(t^4 + a(\theta)\, t\big) = -\frac{3}{4^{4/3}}\, |a(\theta)|^{4/3},$$

because $4(t^\star(\theta))^3 + a(\theta) = 0$ implies

$$(t^\star(\theta))^4 + a(\theta) t^\star(\theta) = (t^\star(\theta))^4 - 4(t^\star(\theta))^4 = -3(t^\star(\theta))^4 = -\frac{3}{4^{4/3}}\, |a(\theta)|^{4/3}.$$

Hence F is $\mathcal{C}^1$ on $\mathbb{R}$ and

$$F'(\theta) = -\frac{1}{4^{1/3}}\, |a(\theta)|^{1/3} \operatorname{sign}(a(\theta))\, a'(\theta),$$

with $F'(\theta) = 0$ whenever $a(\theta) = 0$.

At any θ with $a(\theta) = 0$ (equivalently, $t^\star(\theta) = 0$), the restricted objective is $t \mapsto t^4$, whose second derivative at $t = 0$ is 0, so the optimistic minimizer is degenerate in the sense of Definition 3.4. Moreover, $a(\theta) = 0$ at $\theta_k := 10/(k\pi)$ for each $k \in \mathbb{Z} \setminus \{0\}$ and these zeros are simple (indeed $a'(\theta_k) = -10e^{-1/\theta_k^2}\cos(10/\theta_k)/\theta_k^2 \neq 0$). Therefore, $a(\theta) = a'(\theta_k)(\theta - \theta_k) + o(|\theta - \theta_k|)$ and

$$\left|\frac{t^\star(\theta) - t^\star(\theta_k)}{\theta - \theta_k}\right| \;=\; \frac{1}{4^{1/3}} \cdot \frac{|a(\theta)|^{1/3}}{|\theta - \theta_k|} \;\to\; \infty \qquad \text{as } \theta \to \theta_k,$$

so $\theta \mapsto x^\star(\theta)$ is not differentiable at every θ_k (a countably infinite set accumulating at 0). Furthermore,

$$\frac{|F'(\theta) - F'(\theta_k)|}{|\theta - \theta_k|} = \frac{|F'(\theta)|}{|\theta - \theta_k|} \;\sim\; \frac{|a'(\theta_k)|^{4/3}}{4^{1/3}}\, |\theta - \theta_k|^{-2/3} \;\to\; \infty \qquad \text{as } \theta \to \theta_k,$$

so ∇F is not locally Lipschitz at any θ_k.

C.2 $\mathcal{C}^1$ regularity of F under uniqueness (no non-degeneracy)

We prove Theorem 3.2. Example C.1 shows that even when the optimistic minimizer is unique, the selection map $\theta \mapsto x^\star(\theta)$ can fail to be differentiable at degenerate points. Nevertheless, uniqueness is still enough to ensure that the *value function* $F(\theta) = \min_{x \in \mathcal{S}(\theta)} f(\theta, x)$ is differentiable, and when uniqueness holds in a neighborhood it is in fact $\mathcal{C}^1$.

Remark C.2 (A convenient sufficient condition for $F \in \mathcal{C}^1(\Theta)$)**.** *Assume Θ is compact and $\arg\min_{x \in \mathcal{S}(\theta)} f(\theta, x)$ is a singleton for every $\theta \in \Theta$. If, in addition, the lower-level solution manifolds are uniformly bounded in the sense that there exists a compact set $\mathcal{V} \subset \mathbb{R}^d$ with $\mathcal{S}(\theta) \subseteq \mathcal{V}$ for all $\theta \in \Theta$ (e.g., $\mathcal{V} = \mathbb{B}_d(0; \mathsf{D})$ under Assumption 4), then F is $\mathcal{C}^1$ on Θ and* (3) *holds for all $\theta \in \Theta$. Indeed, Theorem 3.2 applies at each $\theta_0 \in \Theta$; compactness of Θ allows us to extract a finite subcover of the resulting local $\mathcal{C}^1$ neighborhoods.*

Lemma C.3 (A basic envelope/Danskin lemma)**.** *Let $U \subset \mathbb{R}^k$ be compact and let $\phi : \Theta \times U \to \mathbb{R}$ be continuous and $\mathcal{C}^1$ in θ, with $\nabla_\theta \phi$ continuous on a neighborhood of (θ_0, u_0). Define $F(\theta) := \min_{u \in U} \phi(\theta, u)$. If $\arg\min_{u \in U} \phi(\theta_0, u) = \{u_0\}$, then F is differentiable at θ_0 and*

$$\nabla F(\theta_0) = \nabla_\theta \phi(\theta_0, u_0).$$

Proof. We first note that the unique minimizer is continuous at θ_0. Let $\theta_n \to \theta_0$ and pick any minimizer $u_n \in \arg\min_{u \in U} \phi(\theta_n, u)$. By compactness of U, along a subsequence (not relabeled) we have $u_n \to \bar{u} \in U$. For any fixed $u \in U$, optimality of u_n gives $\phi(\theta_n, u_n) \le \phi(\theta_n, u)$. Taking $n \to \infty$ and using continuity of ϕ yields $\phi(\theta_0, \bar{u}) \le \phi(\theta_0, u)$ for all $u \in U$, so $\bar{u} \in \arg\min_{u \in U} \phi(\theta_0, u) = \{u_0\}$. Thus every convergent subsequence of u_n converges to u_0, hence $u_n \to u_0$.

Now fix any direction $d \in \mathbb{R}^m$ (where $\Theta \subseteq \mathbb{R}^m$) and let $t \downarrow 0$. Let $u_t \in \arg\min_{u \in U} \phi(\theta_0 + td, u)$. By the first part, $u_t \to u_0$ as $t \downarrow 0$. Using the inequalities

$$\phi(\theta_0 + td, u_t) - \phi(\theta_0, u_t) \;\le\; F(\theta_0 + td) - F(\theta_0) \;\le\; \phi(\theta_0 + td, u_0) - \phi(\theta_0, u_0),$$

divide by t and let $t \downarrow 0$. The right-hand side converges to $\langle \nabla_\theta \phi(\theta_0, u_0), d \rangle$ by differentiability of ϕ in θ. For the left-hand side, write the difference quotient as the integral of the directional derivative:

$$\frac{\phi(\theta_0 + td, u_t) - \phi(\theta_0, u_t)}{t} = \int_0^1 \big\langle \nabla_\theta \phi(\theta_0 + s\,td, u_t), d \big\rangle \, \mathrm{d}s.$$

Define $\psi_t(s) := \big\langle \nabla_\theta \phi(\theta_0 + s\,td, u_t), d \big\rangle$ for $s \in [0, 1]$. For each fixed $s \in [0, 1]$, we have $\theta_0 + s\,td \to \theta_0$ and $u_t \to u_0$, hence $(\theta_0 + s\,td, u_t) \to (\theta_0, u_0)$ and, by continuity of $\nabla_\theta \phi$, $\psi_t(s) \to \langle \nabla_\theta \phi(\theta_0, u_0), d \rangle$.

To justify exchanging the limit and the integral, we bound ψ_t uniformly in s for small t. Since $\nabla_\theta \phi$ is continuous at (θ_0, u_0), there exists $r > 0$ such that $\nabla_\theta \phi$ is bounded on the ball $\mathbb{B}((\theta_0, u_0); r)$; set

$$M := \sup \Big\{ \|\nabla_\theta \phi(\theta, u)\| : \; (\theta, u) \in \mathbb{B}((\theta_0, u_0); r) \Big\} < \infty.$$

Choose t small enough so that $\|u_t - u_0\| \le r/2$ and $t\|d\| \le r/2$. Then for all $s \in [0, 1]$, $\|\theta_0 + s\,td - \theta_0\| \le t\|d\| \le r/2$, so $(\theta_0 + s\,td, u_t) \in \mathbb{B}((\theta_0, u_0); r)$. Therefore, for all $s \in [0, 1]$ and small t,

$$|\psi_t(s)| \le \|\nabla_\theta \phi(\theta_0 + s\,td, u_t)\| \, \|d\| \le M\|d\|.$$

Since the dominating function $s \mapsto M\|d\|$ is integrable on $[0, 1]$, the dominated convergence theorem yields

$$\int_0^1 \psi_t(s)\, ds \;\to\; \int_0^1 \langle \nabla_\theta \phi(\theta_0, u_0), d \rangle \, ds = \langle \nabla_\theta \phi(\theta_0, u_0), d \rangle.$$

Therefore the directional derivative exists and equals $\langle \nabla_\theta \phi(\theta_0, u_0), d \rangle$ for every d. This map is linear in d, hence F is differentiable at θ_0 with gradient $\nabla_\theta \phi(\theta_0, u_0)$. ■

Proof of Theorem 3.2. Let $x^\star(\theta)$ denote the unique optimistic minimizer for $\theta \in \mathcal{U}$. We prove that F is $\mathcal{C}^1$ locally by (i) building a smooth chart for $\mathcal{S}(\theta)$ near x_0, (ii) reducing the bilevel problem to a fixed-domain minimization and applying an envelope/Danskin lemma (which yields both differentiability and continuity of ∇F), and (iii) computing the resulting gradient in the pseudoinverse form.

Step 1: A θ-dependent chart for $\mathcal{S}(\theta)$ near x_0. Let $\mathcal{T}_0 := \mathcal{T}_{x_0}^{\theta_0}$ and $\mathcal{N}_0 := \mathcal{N}_{x_0}^{\theta_0}$ and let $P_{\mathcal{T}_0}, P_{\mathcal{N}_0}$ be the corresponding orthogonal projectors. Let $k := \dim(\mathcal{T}_0)$ and choose orthonormal matrices $U_{\mathcal{T}_0} \in \mathbb{R}^{d\times k}$ and $U_{\mathcal{N}_0} \in \mathbb{R}^{d\times(d-k)}$ spanning $\mathcal{T}_0$ and $\mathcal{N}_0$. For (u, v) near $(0, 0)$, set $x(u, v) := x_0 + U_{\mathcal{T}_0}u + U_{\mathcal{N}_0}v$ and define

$$\Phi(\theta, u, v) := U_{\mathcal{N}_0}^\top \nabla_x g(\theta, x(u, v)) \in \mathbb{R}^{d-k}.$$

Then Φ is C^2 and $\Phi(\theta_0, 0, 0) = 0$. Moreover,

$$D_v\Phi(\theta_0, 0, 0) = U_{\mathcal{N}_0}^\top H(\theta_0, x_0)U_{\mathcal{N}_0}.$$

By Proposition 2.3(ii), $H(\theta_0, x_0)$ is positive definite on $\mathcal{N}_0$, hence $D_v\Phi(\theta_0, 0, 0)$ is invertible. Therefore the C^2 implicit function theorem yields neighborhoods $\mathcal{U}_\theta \subseteq \mathcal{U}$ of θ_0 and $\mathcal{U}_u$ of 0 and a unique C^2 map $h : \mathcal{U}_\theta \times \mathcal{U}_u \to \mathbb{R}^{d-k}$ such that $\Phi(\theta, u, h(\theta, u)) = 0$. Define the chart

$$\psi(\theta, u) := x_0 + U_{\mathcal{T}_0}u + U_{\mathcal{N}_0}h(\theta, u).$$

As in Step 1 of the proof of Theorem 3.5, shrinking neighborhoods if needed we may assume that for each $\theta \in \mathcal{U}_\theta$, the map $u \mapsto \psi(\theta, u)$ parametrizes $\mathcal{S}(\theta)$ on the local branch that stays near x_0. In particular, $\psi(\theta_0, 0) = x_0$.

Step 2: Reduce to a fixed-domain minimization and show $F \in \mathcal{C}^1$. Step 1 only gives a C^2 chart near x_0. Since $u \mapsto \psi(\theta, u)$ parametrizes only the part of $\mathcal{S}(\theta)$ near x_0, we first show that the minimizer stays in this chart patch for θ close to θ_0.

Step 2a: $x^\star(\theta)$ stays near x_0. Let $\theta_n \to \theta_0$ with $\theta_n \in \mathcal{U}_\theta$ and set $x_n := x^\star(\theta_n) \in \mathcal{S}(\theta_n) \subseteq \mathcal{V}$. By compactness of $\mathcal{V}$, along a subsequence (not relabeled) we have $x_n \to \bar{x} \in \mathcal{V}$. For any fixed $y \in \mathbb{R}^d$, since $x_n \in \arg\min_x g(\theta_n, x)$ we have $g(\theta_n, x_n) \le g(\theta_n, y)$. Taking $n \to \infty$ and using continuity of g yields $g(\theta_0, \bar{x}) \le g(\theta_0, y)$ for all y, hence $\bar{x} \in \mathcal{S}(\theta_0)$.

Moreover, $\psi(\theta_n, 0) \in \mathcal{S}(\theta_n)$ is a feasible competitor for the optimistic problem at θ_n, so

$$f(\theta_n, x_n) = F(\theta_n) \;\le\; f(\theta_n, \psi(\theta_n, 0)).$$

Sending $n \to \infty$ and using continuity of f and ψ gives $f(\theta_0, \bar{x}) \le f(\theta_0, x_0) = F(\theta_0)$. By uniqueness of $x_0 \in \arg\min_{x\in\mathcal{S}(\theta_0)} f(\theta_0, x)$ and $\bar{x} \in \mathcal{S}(\theta_0)$, we conclude $\bar{x} = x_0$. Thus every convergent subsequence of x_n converges to x_0, hence $x_n \to x_0$. After possibly shrinking $\mathcal{U}_\theta$, we may assume $x^\star(\theta)$ lies in the chart patch for all $\theta \in \mathcal{U}_\theta$.

Step 2b: fixed-domain reduction. Write $x^\star(\theta) = \psi(\theta, u^\star(\theta))$ for some $u^\star(\theta) \in \mathcal{U}_u$. Define $\tilde{f}(\theta, u) := f(\theta, \psi(\theta, u))$ and choose a compact neighborhood $U \subset \mathcal{U}_u$ of 0 such that $u^\star(\theta) \in U$ for all $\theta \in \mathcal{U}_\theta$. Then, for all $\theta \in \mathcal{U}_\theta$,

$$F(\theta) = \min_{x\in\mathcal{S}(\theta)} f(\theta, x) = \min_{u\in U} \tilde{f}(\theta, u).$$

Step 2c: differentiability via an envelope lemma. Since f is $\mathcal{C}^1$ and ψ is C^2, $\tilde{f}$ is continuous and $\mathcal{C}^1$ in θ with continuous $\nabla_\theta\tilde{f}$. We claim that for every $\theta \in \mathcal{U}_\theta$, the minimizer of $u \mapsto \tilde{f}(\theta, u)$ over U is unique and equals $u^\star(\theta)$. Indeed, by Step 2b,

$$F(\theta) = \min_{u\in U} \tilde{f}(\theta, u), \qquad \tilde{f}(\theta, u^\star(\theta)) = f(\theta, x^\star(\theta)) = F(\theta),$$

so $u^\star(\theta)$ is a minimizer. Conversely, if $\bar{u} \in U$ also minimizes $\tilde{f}(\theta, \cdot)$, then

$$f(\theta, \psi(\theta, \bar{u})) = \tilde{f}(\theta, \bar{u}) = F(\theta),$$

so $\psi(\theta, \bar{u}) \in \arg\min_{x\in\mathcal{S}(\theta)} f(\theta, x)$. By the uniqueness assumption in Theorem 3.2, this implies $\psi(\theta, \bar{u}) = x^\star(\theta) = \psi(\theta, u^\star(\theta))$. Since $u \mapsto \psi(\theta, u)$ is a chart parametrization on the patch, it is injective, hence $\bar{u} = u^\star(\theta)$. Applying Lemma C.3 at each $\theta \in \mathcal{U}_\theta$ yields that F is differentiable on $\mathcal{U}_\theta$ and

$$\nabla F(\theta) = \nabla_\theta\tilde{f}(\theta, u^\star(\theta)) \qquad \text{for all } \theta \in \mathcal{U}_\theta. \tag{14}$$

Importantly, (14) does *not* require differentiability of $\theta \mapsto u^\star(\theta)$.

Step 2d: continuity of ∇F (hence $F \in \mathcal{C}^1$). We show that $\theta \mapsto u^\star(\theta)$ is continuous on $\mathcal{U}_\theta$. Fix $\bar{\theta} \in \mathcal{U}_\theta$ and let $\theta_n \to \bar{\theta}$ with $\theta_n \in \mathcal{U}_\theta$. By compactness of U, along a subsequence (not

relabeled) we have $u^\star(\theta_n) \to \bar{u} \in U$. Optimality of $u^\star(\theta_n)$ gives $\tilde{f}(\theta_n, u^\star(\theta_n)) \le \tilde{f}(\theta_n, u)$ for all $u \in U$. Sending $n \to \infty$ and using continuity of $\tilde{f}$ yields $\tilde{f}(\bar{\theta}, \bar{u}) \le \tilde{f}(\bar{\theta}, u)$ for all $u \in U$. Thus $\bar{u} \in \arg\min_{u \in U} \tilde{f}(\bar{\theta}, u) = \{u^\star(\bar{\theta})\}$, hence $\bar{u} = u^\star(\bar{\theta})$. Therefore $u^\star(\theta_n) \to u^\star(\bar{\theta})$, proving continuity. Since $\nabla_\theta \tilde{f}$ is continuous, (14) implies that ∇F is continuous on $\mathcal{U}_\theta$. Setting $\mathcal{U}_0 := \mathcal{U}_\theta$, we conclude that F is $\mathcal{C}^1$ on $\mathcal{U}_0$.

Step 3: Compute $\nabla F(\theta)$ and conclude (3)**.** Fix any $\theta \in \mathcal{U}_0$ and write $u^\star := u^\star(\theta)$ and $x^\star := x^\star(\theta) = \psi(\theta, u^\star) \in \mathcal{S}(\theta)$. By (14) and the chain rule for $\tilde{f}(\theta, u) = f(\theta, \psi(\theta, u))$ (with u held fixed),

$$\nabla F(\theta) = \nabla_\theta f(\theta, x^\star) + \big(\nabla_x f(\theta, x^\star)\big)^\top \nabla_\theta \psi(\theta, u^\star).$$

Since $x^\star$ minimizes $f(\theta, \cdot)$ over the manifold $\mathcal{S}(\theta)$ (without boundary), its Riemannian gradient vanishes, i.e., $P_{\mathcal{T}^\theta_{x^\star}} \nabla_x f(\theta, x^\star) = 0$, hence $\nabla_x f(\theta, x^\star) \in \mathcal{N}^\theta_{x^\star}$.

On the other hand, for all $(\theta, u) \in \mathcal{U}_0 \times \mathcal{U}_u$ we have $\psi(\theta, u) \in \mathcal{S}(\theta)$, so $\nabla_x g(\theta, \psi(\theta, u)) = 0$. Differentiating this identity with respect to θ (with u held fixed) gives

$$H(\theta, \psi(\theta, u))\, \nabla_\theta \psi(\theta, u) + \nabla^2_{x\theta} g(\theta, \psi(\theta, u)) = 0.$$

Evaluating at $u = u^\star$ and multiplying by $H(\theta, x^\star)^\dagger$ yields

$$P_{\mathcal{N}^\theta_{x^\star}} \nabla_\theta \psi(\theta, u^\star) = -H(\theta, x^\star)^\dagger \nabla^2_{x\theta} g(\theta, x^\star),$$

where we used that $H(\theta, x^\star)^\dagger H(\theta, x^\star) = P_{\mathcal{N}^\theta_{x^\star}}$ by Proposition 2.3(ii). Since $\nabla_x f(\theta, x^\star) \in \mathcal{N}^\theta_{x^\star}$, we can insert the projector and obtain

$$\big(\nabla_x f(\theta, x^\star)\big)^\top \nabla_\theta \psi(\theta, u^\star) = -\big(\nabla_x f(\theta, x^\star)\big)^\top H(\theta, x^\star)^\dagger \nabla^2_{x\theta} g(\theta, x^\star).$$

Substituting into the expression for $\nabla F(\theta)$ and using $\nabla^2_{\theta x} g := (\nabla^2_{x\theta} g)^\top$ gives (3). ■

C.3 Proof of Theorem 3.5

We restate here the local pieces of Proposition 2.3 that are used in the proof. Under Assumption 1, we may equivalently write

$$\mathcal{S}(\theta) = \arg\min_{x \in \mathbb{R}^d} g(\theta, x) = \{x : \nabla_x g(\theta, x) = 0\}.$$

Assumption 5 (Geometry (from Proposition 2.3))**.** *For every $\theta \in \Theta$, the set $\mathcal{S}(\theta)$ is a non-empty, compact, embedded C^2 submanifold of $\mathbb{R}^d$ (without boundary). Moreover, there exists an integer k such that $\dim(\mathcal{S}(\theta)) = k$ for all $\theta \in \Theta$.*

Assumption 6 (Normal nondegeneracy (from Proposition 2.3))**.** *For every $\theta \in \Theta$ and every $x \in \mathcal{S}(\theta)$, $\nabla^2_{xx} g(\theta, x)$ is positive definite on the normal space $\mathcal{N}_x \mathcal{S}(\theta)$.*

Lemma C.4. *Fix θ and let $M, N \subset \mathbb{R}^d$ be embedded k-dimensional C^1 submanifolds (without boundary) and $p \in M \subseteq N$. Then there exists a neighborhood V of p in $\mathbb{R}^d$ such that $M \cap V = N \cap V$.*

Proof. Since N is an embedded submanifold, there exist a neighborhood V of p and a C^1 diffeomorphism $\varphi : V \to \varphi(V) \subset \mathbb{R}^d$ such that $\varphi(N \cap V) = \mathbb{R}^k \times \{0\}$ (after identifying $\mathbb{R}^d \simeq \mathbb{R}^k \times \mathbb{R}^{d-k}$). Then $\varphi(M \cap V)$ is a k-dimensional embedded submanifold of $\mathbb{R}^k \times \{0\} \simeq \mathbb{R}^k$, hence it is an open subset of $\mathbb{R}^k$. Shrinking V so that $\varphi(V) \cap (\mathbb{R}^k \times \{0\})$ lies inside that open subset yields $\varphi(N \cap V) \subseteq \varphi(M \cap V)$, hence $N \cap V \subseteq M \cap V$. The reverse inclusion is always true, so $M \cap V = N \cap V$. ■

Lemma C.5. $\ker(\nabla^2_{xx} g(\theta, x)) = \mathcal{T}_x \mathcal{S}(\theta)$.

Proof. Fix such θ and $x \in \mathcal{S}(\theta)$ near x_0. If $v \in \mathcal{T}_x \mathcal{S}(\theta)$, take a curve $x(t) \subset \mathcal{S}(\theta)$ with $\dot{x}(0) = v$; differentiating $\nabla_x g(\theta, x(t)) \equiv 0$ at $t = 0$ gives $\nabla^2_{xx} g(\theta, x) v = 0$, hence $\mathcal{T}_x \mathcal{S}(\theta) \subseteq \ker(\nabla^2_{xx} g(\theta, x))$. Conversely, if $v \in \ker(\nabla^2_{xx} g(\theta, x))$ and write $v = v_\mathcal{T} + v_\mathcal{N}$ according to $\mathbb{R}^d = \mathcal{T}_x \mathcal{S}(\theta) \oplus \mathcal{N}_x \mathcal{S}(\theta)$, then

$$0 = \langle v_\mathcal{N}, \nabla^2_{xx} g(\theta, x) v \rangle = \langle v_\mathcal{N}, \nabla^2_{xx} g(\theta, x) v_\mathcal{N} \rangle,$$

so Assumption 6 forces $v_\mathcal{N} = 0$ and $v \in \mathcal{T}_x \mathcal{S}(\theta)$. ■

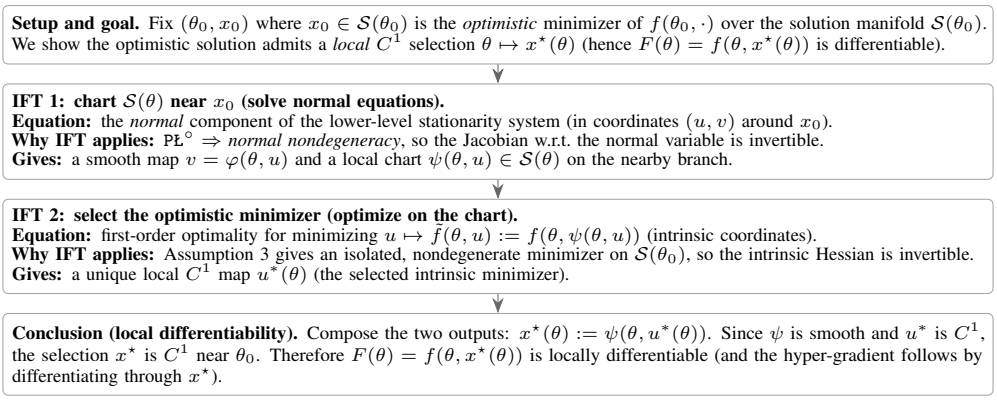


Figure 4: **Two implicit-function steps used in the proof of Theorem 3.5.** IFT 1 uses *normal nondegeneracy* (from PŁ$^\circ$) to build a local chart of the solution manifold $\mathcal{S}(\theta)$ near x_0. IFT 2 uses *local identifiability* of the optimistic minimizer on $\mathcal{S}(\theta_0)$ to obtain a unique C^1 intrinsic minimizer and hence a C^1 selection $x^\star(\theta)$.

Proof of Theorem 3.5. Fix θ_0 and let $x_0 := x^\star(\theta_0)$ be the unique minimizer of $x \mapsto f(\theta_0, x)$ over $\mathcal{S}(\theta_0)$. Let $\mathcal{T}_0 := \mathcal{T}_{x_0}\mathcal{S}(\theta_0)$, $\mathcal{N}_0 := \mathcal{N}_{x_0}\mathcal{S}(\theta_0)$, and let $P_{\mathcal{T}_0}, P_{\mathcal{N}_0}$ be the orthogonal projectors.

Figure 4 provides a roadmap of the two places where we invoke the implicit function theorem and why each invocation is valid.

Step 1: A θ-dependent local chart for $\mathcal{S}(\theta)$ near x_0. Let $k := \dim(\mathcal{T}_0)$ and choose orthonormal matrices $U_{\mathcal{T}_0} \in \mathbb{R}^{d\times k}$ and $U_{\mathcal{N}_0} \in \mathbb{R}^{d\times(d-k)}$ spanning $\mathcal{T}_0$ and $\mathcal{N}_0$. For (u, v) near $(0, 0)$ set $x(u, v) := x_0 + U_{\mathcal{T}_0}u + U_{\mathcal{N}_0}v$ and define

$$\Phi(\theta, u, v) := U_{\mathcal{N}_0}^\top \nabla_x g(\theta, x(u, v)) \in \mathbb{R}^{d-k}.$$

Since g is C^3 in (θ, x), Φ is C^2 in (θ, u, v) and $\Phi(\theta_0, 0, 0) = 0$ because $\nabla_x g(\theta_0, x_0) = 0$.

Moreover,

$$D_v\Phi(\theta_0, 0, 0) = U_{\mathcal{N}_0}^\top \nabla_{xx}^2 g(\theta_0, x_0) U_{\mathcal{N}_0}.$$

By Assumption 6, $\nabla_{xx}^2 g(\theta_0, x_0)$ is positive definite on $\mathcal{N}_0$, hence this matrix is SPD and invertible. We now apply the C^2 *implicit function theorem*: since Φ is C^2 and $D_v\Phi(\theta_0, 0, 0)$ is invertible, the implicit solution map inherits C^2 regularity. Therefore, there exist neighborhoods $\mathcal{U}_\theta$ of θ_0 and $\mathcal{U}_u$ of 0, and a unique C^2 map $h : \mathcal{U}_\theta \times \mathcal{U}_u \to \mathbb{R}^{d-k}$ such that $\Phi(\theta, u, h(\theta, u)) = 0$. Define

$$\psi(\theta, u) := x_0 + U_{\mathcal{T}_0}u + U_{\mathcal{N}_0}h(\theta, u).$$

Then ψ is C^2 in (θ, u). From the IFT construction above, for each $\theta \in \mathcal{U}_\theta$ the set

$$\widetilde{\mathcal{S}}(\theta) := \{x : \ U_{\mathcal{N}_0}^\top \nabla_x g(\theta, x) = 0\}$$

is an embedded C^2 submanifold of $\mathbb{R}^d$ of dimension k, and $\psi(\theta, \cdot)$ parametrizes $\widetilde{\mathcal{S}}(\theta)$ near x_0.

Moreover, since $\mathcal{S}(\theta) = \{x : \nabla_x g(\theta, x) = 0\}$, we have the inclusion $\mathcal{S}(\theta) \subseteq \widetilde{\mathcal{S}}(\theta)$.

By Assumption 5, for all $\theta \in \mathcal{U}_\theta$ the set $\mathcal{S}(\theta)$ is a k-dimensional embedded submanifold. Since $\widetilde{\mathcal{S}}(\theta)$ is also a k-dimensional embedded submanifold and $\mathcal{S}(\theta) \subseteq \widetilde{\mathcal{S}}(\theta)$, Lemma C.4 implies that $\mathcal{S}(\theta)$ and $\widetilde{\mathcal{S}}(\theta)$ coincide locally on this patch. Shrinking $\mathcal{U}_u$ and the ambient neighborhood if needed, we may therefore assume that, for all $\theta \in \mathcal{U}_\theta$, the map $u \mapsto \psi(\theta, u)$ provides a local chart of $\mathcal{S}(\theta)$ (on the local branch that stays near x_0).

Step 2: Reduce to an unconstrained problem and select $x^\star(\theta)$. Define the reduced objective $\tilde{f}(\theta, u) := f(\theta, \psi(\theta, u))$. Since f is C^2 and ψ is C^2, $\tilde{f}$ is C^2 in (θ, u), hence $\nabla_u \tilde{f}$ is C^1 in (θ, u).

Because x_0 minimizes $f(\theta_0, \cdot)$ over $\mathcal{S}(\theta_0)$ and $\psi(\theta_0, \cdot)$ parametrizes $\mathcal{S}(\theta_0)$ near x_0, $u_0 := 0$ is a minimizer of $u \mapsto \tilde{f}(\theta_0, u)$, hence $\nabla_u \tilde{f}(\theta_0, 0) = 0$.

We now show $D_u(\nabla_u \tilde{f})(\theta_0, 0) = \nabla^2_{uu} \tilde{f}(\theta_0, 0)$ is invertible. Let $A := D_u \psi(\theta_0, 0) : \mathbb{R}^k \to \mathbb{R}^d$. Since $\psi(\theta_0, \cdot)$ is a chart of the k-dim manifold $\mathcal{S}(\theta_0)$, A has full column rank and $\mathrm{range}(A) = \mathcal{T}_0$. For any $w \in \mathbb{R}^k$, set $v := Aw \in \mathcal{T}_0$. The second variation of the pullback along w coincides with the Riemannian Hessian quadratic form along v (standard chart calculus on manifolds), so Assumption 3 implies $\langle w, \nabla^2_{uu} \tilde{f}(\theta_0, 0) w \rangle > 0$ for all $w \neq 0$. Hence $\nabla^2_{uu} \tilde{f}(\theta_0, 0)$ is positive definite and invertible.

We now apply the C^1 *implicit function theorem* to the k equations $\nabla_u \tilde{f}(\theta, u) = 0$ at $(\theta_0, 0)$ (note $\nabla_u \tilde{f}$ is C^1 and its Jacobian in u is invertible at $(\theta_0, 0)$) to obtain a neighborhood $\mathcal{U}_\theta$ and a unique C^1 map $u^* : \mathcal{U}_\theta \to \mathcal{U}_u$ such that $\nabla_u \tilde{f}(\theta, u^*(\theta)) = 0$. Define

$$x^\star(\theta) := \psi(\theta, u^*(\theta)).$$

Then $x^\star(\theta) \in \mathcal{S}(\theta)$ and, by positive definiteness of $\nabla^2_{uu} \tilde{f}$ and continuity, $x^\star(\theta)$ is the unique local minimizer of $f(\theta, \cdot)$ over $\mathcal{S}(\theta)$ near x_0. In particular, locally

$$F(\theta) = \min_{x \in \mathcal{S}(\theta)} f(\theta, x) = f(\theta, x^\star(\theta)).$$

Step 3: Chain rule and normal-only dependence. Since $x^\star$ is C^1 and f is C^2, F is differentiable at θ_0 and

$$\nabla_\theta F(\theta_0) = \nabla_\theta f(\theta_0, x_0) + (\nabla_x f(\theta_0, x_0))^\top \nabla_\theta x^\star(\theta_0).$$

Because x_0 is a constrained minimizer on $\mathcal{S}(\theta_0)$, the Riemannian gradient vanishes, i.e. $P_{\mathcal{T}_0} \nabla_x f(\theta_0, x_0) = 0$, hence $\nabla_x f(\theta_0, x_0) \in \mathcal{N}_0$. Therefore only the normal component of $\nabla_\theta x^\star(\theta_0)$ contributes.

Step 4: Compute the normal component from $\nabla_x g(\theta, x^\star(\theta)) = 0$. Since $x^\star(\theta) \in \mathcal{S}(\theta)$, we have $\nabla_x g(\theta, x^\star(\theta)) = 0$ for all θ near θ_0. Differentiating at θ_0 gives

$$\nabla^2_{x\theta} g(\theta_0, x_0) + \nabla^2_{xx} g(\theta_0, x_0)\, \nabla_\theta x^\star(\theta_0) = 0. \tag{$\star$}$$

By Lemma C.5, $\ker(\nabla^2_{xx} g(\theta_0, x_0)) = \mathcal{T}_0$. Since $\nabla^2_{xx} g(\theta_0, x_0)$ is symmetric, $\mathrm{range}(\nabla^2_{xx} g(\theta_0, x_0)) = \mathcal{T}_0^\perp = \mathcal{N}_0$. Thus the Moore–Penrose pseudoinverse satisfies

$$(\nabla^2_{xx} g(\theta_0, x_0))^\dagger\, \nabla^2_{xx} g(\theta_0, x_0) = P_{\mathcal{N}_0}.$$

Apply $(\nabla^2_{xx} g(\theta_0, x_0))^\dagger$ to $(\star)$:

$$P_{\mathcal{N}_0} \nabla_\theta x^\star(\theta_0) = -(\nabla^2_{xx} g(\theta_0, x_0))^\dagger\, \nabla^2_{x\theta} g(\theta_0, x_0).$$

Since $\nabla_x f(\theta_0, x_0) \in \mathcal{N}_0$, inserting $P_{\mathcal{N}_0}$ into the chain rule yields

$$(\nabla_x f(\theta_0, x_0))^\top \nabla_\theta x^\star(\theta_0) = -(\nabla_x f(\theta_0, x_0))^\top (\nabla^2_{xx} g(\theta_0, x_0))^\dagger\, \nabla^2_{x\theta} g(\theta_0, x_0).$$

Therefore

$$\nabla_\theta F(\theta_0) = \nabla_\theta f(\theta_0, x_0) - (\nabla_x f(\theta_0, x_0))^\top (\nabla^2_{xx} g(\theta_0, x_0))^\dagger\, \nabla^2_{x\theta} g(\theta_0, x_0).$$

Equivalently, with $\nabla^2_{\theta x} g := (\nabla^2_{x\theta} g)^\top$,

$$\nabla F(\theta_0) = \nabla_\theta f(\theta_0, x_0) - \nabla^2_{\theta x} g(\theta_0, x_0)\, (\nabla^2_{xx} g(\theta_0, x_0))^\dagger\, \nabla_x f(\theta_0, x_0),$$

which is the claimed formula. ■

D Proofs for Section 5

This appendix proves Proposition 5.1 and Proposition 5.2, and records additional parameter-tuning details deferred from Section 5.

Proof of Proposition 5.1. For each $\theta_0 \in \Theta$, Assumption 3 and Theorem 3.5 yield an open neighborhood $\mathcal{U}_{\theta_0} \subseteq \Theta$ and a $\mathcal{C}^1$ map $x^\star_{\theta_0} : \mathcal{U}_{\theta_0} \to \mathbb{R}^d$ such that $x^\star_{\theta_0}(\theta) \in \mathcal{S}(\theta)$ and $F(\theta) = f(\theta, x^\star_{\theta_0}(\theta))$ on $\mathcal{U}_{\theta_0}$. Because the optimistic minimizer is unique for every $\theta \in \Theta$, these local branches agree on overlaps and therefore patch together into a global $\mathcal{C}^1$ selection $x^\star : \Theta \to \mathbb{R}^d$. Since Θ is compact, $x^\star$ is Lipschitz on Θ.

Let $\Gamma := \{(\theta, x^\star(\theta)) : \theta \in \Theta\}$. The graph Γ is compact. Since f is $\mathcal{C}^2$ and g is $\mathcal{C}^3$, the maps $(\theta, x) \mapsto \nabla_\theta f(\theta, x)$, $(\theta, x) \mapsto \nabla_x f(\theta, x)$, $(\theta, x) \mapsto \nabla^2_{\theta x} g(\theta, x)$, and $(\theta, x) \mapsto \nabla^2_{xx} g(\theta, x)$ are locally Lipschitz on a neighborhood of Γ, hence Lipschitz on Γ after enlarging the constants if needed. Moreover, Proposition 2.3 gives a uniform normal spectral gap on Γ, so the pseudoinverse map $(\theta, x) \mapsto [\nabla^2_{xx} g(\theta, x)]^\dagger$ is locally Lipschitz near Γ (constant-rank perturbation of symmetric matrices), and by compactness of Γ it is Lipschitz on Γ.

Applying the hyper-gradient formula from Theorem 3.5 on each neighborhood $\mathcal{U}_{\theta_0}$ and using uniqueness to identify the same global branch, we obtain for every $\theta \in \Theta$,

$$\nabla F(\theta) = \nabla_\theta f(\theta, x^\star(\theta)) - \nabla^2_{\theta x} g(\theta, x^\star(\theta)) \big[\nabla^2_{xx} g(\theta, x^\star(\theta))\big]^\dagger \nabla_x f(\theta, x^\star(\theta)).$$

Combining the Lipschitz bounds in this formula with the Lipschitz continuity of $x^\star$ on Θ yields the claim. ■

Proof of Proposition 5.2. Fix $T \geq 1$ and write $g_t := \widehat{h}_t = \nabla F(\theta_t) + e_t$. Define the (used) gradient mapping

$$\widetilde{\mathcal{G}}_t := \alpha^{-1}\big(\theta_t - \mathrm{Proj}_\Theta(\theta_t - \alpha g_t)\big) = \alpha^{-1}(\theta_t - \theta_{t+1}).$$

Step 1: a descent inequality in terms of $\widetilde{\mathcal{G}}_t$. Let $\Delta_t := \theta_{t+1} - \theta_t = -\alpha\, \widetilde{\mathcal{G}}_t$. By L_F-smoothness of F on Θ and convexity of Θ, we have

$$F(\theta_{t+1}) \leq F(\theta_t) + \langle \nabla F(\theta_t), \Delta_t \rangle + \frac{L_F}{2}\|\Delta_t\|^2.$$

Since $\theta_{t+1} = \mathrm{Proj}_\Theta(\theta_t - \alpha g_t)$ and $\theta_t \in \Theta$, the projection optimality condition yields

$$\langle g_t, \Delta_t \rangle \;\leq\; -\frac{1}{\alpha}\|\Delta_t\|^2.$$

Using $\nabla F(\theta_t) = g_t - e_t$ and the inequality $ab \leq \frac{1}{4}a^2 + b^2$ gives

$$\begin{aligned}
F(\theta_{t+1}) &\leq F(\theta_t) + \langle g_t, \Delta_t \rangle - \langle e_t, \Delta_t \rangle + \frac{L_F}{2}\|\Delta_t\|^2 \\
&\leq F(\theta_t) - \frac{1}{\alpha}\|\Delta_t\|^2 + \|e_t\|\,\|\Delta_t\| + \frac{L_F}{2}\|\Delta_t\|^2 \\
&\leq F(\theta_t) - \frac{1}{\alpha}\|\Delta_t\|^2 + \frac{1}{4\alpha}\|\Delta_t\|^2 + \alpha\|e_t\|^2 + \frac{L_F}{2}\|\Delta_t\|^2 \\
&= F(\theta_t) - \Big(\frac{3}{4\alpha} - \frac{L_F}{2}\Big)\|\Delta_t\|^2 + \alpha\|e_t\|^2.
\end{aligned}$$

Using $\alpha \leq 1/L_F$ gives $\frac{3}{4\alpha} - \frac{L_F}{2} \geq \frac{1}{4\alpha}$, hence

$$F(\theta_{t+1}) \leq F(\theta_t) - \frac{1}{4\alpha}\|\Delta_t\|^2 + \alpha\|e_t\|^2 = F(\theta_t) - \frac{\alpha}{4}\|\widetilde{\mathcal{G}}_t\|^2 + \alpha\|e_t\|^2.$$

Summing over $t = 0, \ldots, T-1$ and using $F(\theta_T) \geq F_\star$ yields

$$\sum_{t=0}^{T-1} \|\widetilde{\mathcal{G}}_t\|^2 \leq \frac{4\,(F(\theta_0) - F_\star)}{\alpha} + 4\sum_{t=0}^{T-1}\|e_t\|^2. \tag{15}$$

Step 2: relate $\widetilde{\mathcal{G}}_t$ to the true gradient mapping. By nonexpansiveness of projection,

$$\|\mathcal{G}_\Theta(\theta_t, \nabla F(\theta_t); \alpha) - \widetilde{\mathcal{G}}_t\| = \frac{1}{\alpha}\big\|\mathrm{Proj}_\Theta(\theta_t - \alpha g_t) - \mathrm{Proj}_\Theta(\theta_t - \alpha \nabla F(\theta_t))\big\| \leq \|e_t\|.$$

Therefore $\|\mathcal{G}_\Theta(\theta_t, \nabla F(\theta_t); \alpha)\|^2 \leq 2\|\widetilde{\mathcal{G}}_t\|^2 + 2\|e_t\|^2$, and combining with (15) gives

$$\sum_{t=0}^{T-1} \|\mathcal{G}_\Theta(\theta_t, \nabla F(\theta_t); \alpha)\|^2 \leq \frac{8\,(F(\theta_0) - F_\star)}{\alpha} + 10\sum_{t=0}^{T-1}\|e_t\|^2.$$

Dividing by T and taking expectations yields (7). ■

D.1 Parameter choices and oracle complexity

This appendix records a convenient way to pick algorithmic parameters so that the hyper-gradient error satisfies $\sup_t \mathbb{E}\|e_t\|^2 = O(\varepsilon^2)$, and hence Proposition 5.2 yields an ε-stationarity guarantee.

Proof of Corollary 5.5. By the Poincaré assumption and (38), each candidate law satisfies

$$\mathsf{W}_2(\nu_{t,i}, \mu_{\theta_t}^{\lambda}) \le 2\, C_{\mathrm{PI}}^{1/2} \left(e^{\varepsilon_{\mathrm{R}}^2} - 1\right)^{1/2}.$$

Combining (8) with (9) and the bound $\mathbb{E}\big[\|\tilde{x}_t - x^\star(\theta_t)\|^2 \mathbf{1}_{\mathcal{E}_t}\big] \le \mathbb{E}\|\tilde{x}_t - x^\star(\theta_t)\|^2$, we obtain uniformly over t

$$\begin{aligned}\mathbb{E}\|e_t\|^2 \;\le\; & 3C_x^2\Big(1 + \tfrac{2}{\gamma} + \tfrac{2}{\gamma(c+\gamma)}\Big)^2 \Bigg[2C_{\mathrm{tube},2}\, e^{\varepsilon_{\mathrm{R}}^2/2}\, \lambda \log(1+M) \\ & + \; 2\left(\frac{1}{c_{\mathrm{hg}}} + \frac{4\mathsf{D}^2}{\Delta_{r_0}}\right)\Big(2L_{f,1}\, C_{\mathrm{tube}} \sqrt{\lambda \log(1+M)} \;+\; C_1\, M^{-1/\mathtt{k}} \;+\; 4L_{f,1}\sqrt{M\, C_{\mathrm{PI}}}\left(e^{\varepsilon_{\mathrm{R}}^2} - 1\right)^{1/2}\Big)\Bigg] \\ & + \; \tfrac{12 C_{\mathrm{lin}}^2}{\gamma^2}\, \eta_t^2 \;+\; 3C_{\mathrm{reg}}^2\, \gamma^2 \;+\; B_e^2\, \mathbb{P}(\mathcal{E}_t^c).\end{aligned} \tag{16}$$

To control the off-tube probability, the tube tail from Appendix F (Lemma F.5) together with the Rényi-2 tail transfer (see the proof of Lemma F.5) yields

$$\mathbb{P}(\mathcal{E}_t^c) \le M\, e^{\varepsilon_{\mathrm{R}}^2/2}\, \sqrt{C_{\mathrm{tube}}} \exp\Big(-\tfrac{c_{\mathrm{tube}}}{2}\tfrac{r(\gamma)^2}{\lambda}\Big).$$

Since all remaining quantities in (16) are global problem-data constants independent of $M, \lambda, \varepsilon_{\mathrm{R}}, \gamma, \eta_t$, and t, absorbing them into big-O notation gives

$$\begin{aligned}\mathbb{E}\|e_t\|^2 = O\Bigg(& \Big(1 + \tfrac{1}{\gamma} + \tfrac{1}{\gamma(c+\gamma)}\Big)^2 \Big[e^{\varepsilon_{\mathrm{R}}^2/2}\lambda\log(1+M) + \sqrt{\lambda\log(1+M)} + M^{-1/\mathtt{k}} + \sqrt{M\, C_{\mathrm{PI}}}\left(e^{\varepsilon_{\mathrm{R}}^2} - 1\right)^{1/2}\Big] \\ & + \frac{\eta_t^2}{\gamma^2} + \gamma^2 \\ & + M e^{\varepsilon_{\mathrm{R}}^2/2} \exp\Big(-\tfrac{c_{\mathrm{tube}}}{2}\tfrac{r(\gamma)^2}{\lambda}\Big)\Bigg).\end{aligned}$$

Moreover, since $c > 0$ is fixed and $\gamma \in (0, 1]$,

$$\Big(1 + \tfrac{1}{\gamma} + \tfrac{1}{\gamma(c+\gamma)}\Big)^2 = O\Big(\tfrac{1}{\gamma^2}\Big).$$

Substituting this simplification into the previous display yields the claimed bound. ∎

Choosing parameters to achieve $\mathbb{E}\|e_t\|^2 = O(\varepsilon^2)$. Fix a target $\varepsilon \in (0, 1)$. To ensure $\sup_t \mathbb{E}\|e_t\|^2 = O(\varepsilon^2)$ it suffices to make each term on the right-hand side of (16) of order ε^2. One convenient choice is: (i) set $\gamma = \varepsilon$ so that $C_{\mathrm{reg}}\gamma = O(\varepsilon)$, (ii) solve the ridge system to residual $\eta_t = \gamma\varepsilon = \varepsilon^2$ so that $(C_{\mathrm{lin}}/\gamma)\eta_t = O(\varepsilon)$, and (iii) choose the sampling/selection parameters so that

$$e^{\varepsilon_{\mathrm{R}}^2/2}\lambda\log(1+M) \;+\; \sqrt{\lambda\log(1+M)} \;+\; M^{-1/\mathtt{k}} \;+\; \sqrt{M\, C_{\mathrm{PI}}}\left(e^{\varepsilon_{\mathrm{R}}^2} - 1\right)^{1/2} = O(\varepsilon^4).$$

A sufficient scaling for (iii) is to take

$$M = \left\lceil \varepsilon^{-4\mathtt{k}} \right\rceil, \qquad \lambda = \frac{\varepsilon^8}{\log(1+M)}, \qquad \varepsilon_{\mathrm{R}} = \varepsilon^{4+2\mathtt{k}}.$$

Then $\lambda\log(1+M) = \varepsilon^8$, $\sqrt{\lambda\log(1+M)} = \varepsilon^4$, and $M^{-1/\mathtt{k}} = O(\varepsilon^4)$. Moreover, for fixed C_{PI},

$$\sqrt{M\, C_{\mathrm{PI}}}\left(e^{\varepsilon_{\mathrm{R}}^2} - 1\right)^{1/2} = O\Big(\varepsilon^{-2\mathtt{k}} \cdot \varepsilon^{4+2\mathtt{k}}\Big) = O(\varepsilon^4),$$

since $e^u - 1 = O(u)$ as $u \downarrow 0$. Thus the displayed sampling/selection term is $O(\varepsilon^4)$, and the corresponding contribution to $\mathbb{E}\|e_t\|^2$ is $O(\varepsilon^2)$ after multiplication by the prefactor $O(1/\gamma^2) =$

$O(1/\varepsilon^2)$. Moreover, under this scaling the off-tube probability above is exponentially small in $r(\gamma)^2/\lambda$.

Substituting this control into Proposition 5.2 (with stepsize $1/L_F$) yields an ε-stationary guarantee after

$$T = O\left(\frac{L_F\,(F(\theta_0) - F_\star)}{\varepsilon^2}\right)$$

outer iterations.

Total oracle complexity (informal). Per outer iteration, HG-MS uses (a) MK evaluations of $\nabla_x g$ for the sampler (Algorithm 2), (b) M evaluations of $f(\theta_t, X_{t,i})$ for hard selection, and (c) $\#\mathrm{HVP}$ Hessian–vector products for CG to reach residual tolerance η_t. Thus the total first-order oracle cost is

$$O\big(T(MK + M + \#\mathrm{HVP})\big).$$

If the LMC/ULA sampler is run for K steps so that the order-2 Rényi sampling error satisfies $\varepsilon_{\mathrm{R}} = \varepsilon^{4+2\mathtt{k}}$, then Proposition F.11 gives $K = \widetilde{O}(d\,C_{\mathrm{PI}}^2\,\lambda^{-2}\varepsilon_{\mathrm{R}}^{-2})$. Under the above scaling of (M, λ) this gives $K = \widetilde{O}\big(d\,C_{\mathrm{PI}}^2\,\log^2(1+M)\,\varepsilon^{-24-4\mathtt{k}}\big)$ and hence

$$T\,M\,K = \widetilde{O}\big(d\,C_{\mathrm{PI}}^2\,M\log^2(1+M)\,\varepsilon^{-26-4\mathtt{k}}\big) = \widetilde{O}\big(d\,C_{\mathrm{PI}}^2\,\varepsilon^{-26-8\mathtt{k}}\big),$$

which dominates the additional TM selection-evaluation term and the CG term $T\,\#\mathrm{HVP}$.

E Derivation of the hyper-gradient stability bound

We prove a general deterministic stability bound for the error incurred when replacing the pseudoinverse action in Theorem 3.5 by a ridge-regularized, inexact linear solve. The main-text bound in Lemma E.3 follows by applying the lemma below on the tube event $\mathcal{E}_t$ and invoking Lemma E.1 to control the inverse norm by $2/\gamma$. The argument is standard and relies on basic residual-to-error relations for linear systems together with a spectral comparison between the normal-space inverse and its ridge regularization; see, e.g., Golub and Van Loan [2013, Ch. 5] or Ben-Israel and Greville [2003, Ch. 4].

Lemma E.1 (Tube radius ensuring ridge invertibility)**.** *For any fixed $\gamma > 0$, there exists $r(\gamma) > 0$ such that for all $\theta \in \Theta$ and all $x \in \mathbb{R}^d$ satisfying $\mathrm{dist}(x, \mathcal{S}(\theta)) \le r(\gamma)$, we have*

$$\nabla^2_{xx} g(\theta, x) + \gamma I \;\succeq\; \tfrac{\gamma}{2}\, I, \qquad \textit{and hence} \qquad \big\|\big(\nabla^2_{xx} g(\theta, x) + \gamma I\big)^{-1}\big\| \;\le\; \tfrac{2}{\gamma}.$$

Proof. Let $\mathcal{K} := \{(\theta, x) : \theta \in \Theta,\ x \in \mathcal{S}(\theta)\}$. By Assumption 4, $\mathcal{S}(\theta) \subseteq \mathbb{B}_d(0; \mathsf{D})$ for all θ, hence $\mathcal{K}$ is compact. Since $(\theta, x) \mapsto \nabla^2_{xx} g(\theta, x)$ is continuous (Assumption 4), it is uniformly continuous on a neighborhood of $\mathcal{K}$. Therefore, for any fixed $\gamma > 0$ there exists $r(\gamma) > 0$ such that $\lambda_{\min}(\nabla^2_{xx} g(\theta, x)) \ge -\gamma/2$ whenever $\theta \in \Theta$ and $\mathrm{dist}(x, \mathcal{S}(\theta)) \le r(\gamma)$. This implies $\nabla^2_{xx} g(\theta, x) + \gamma I \succeq (\gamma/2) I$, hence invertibility and the inverse-norm bound. ■

Lemma E.2 (General stability bound (technical))**.** *Let Assumptions 1, 3 and 4 hold, and let $x^\star : \Theta \to \mathbb{R}^d$ denote the global $\mathcal{C}^1$ selection obtained by patching the local branches from Theorem 3.5 under Assumption 3. Let $\mathcal{V} \subset \mathbb{R}^d$ be a compact set containing $\{x^\star(\theta) : \theta \in \Theta\}$. Define*

$$A(\theta, x) := \nabla^2_{\theta x} g(\theta, x), \qquad H(\theta, x) := \nabla^2_{xx} g(\theta, x), \qquad b(\theta, x) := \nabla_x f(\theta, x),$$

and the exact and ridge-regularized implicit maps

$$h(\theta, x) := \nabla_\theta f(\theta, x) - A(\theta, x)\, H(\theta, x)^\dagger b(\theta, x),$$
$$h_\gamma(\theta, x) := \nabla_\theta f(\theta, x) - A(\theta, x)\,\big(H(\theta, x) + \gamma I\big)^{-1} b(\theta, x).$$

Fix an iteration t with $\theta_t \in \Theta$ and a point $\tilde{x}_t \in \mathcal{V}$. Let $\tilde{v}_t$ satisfy the residual bound

$$\big\|\big(H(\theta_t, \tilde{x}_t) + \gamma I\big)\tilde{v}_t - b(\theta_t, \tilde{x}_t)\big\| \;\le\; \eta_t.$$

Define the estimator $\widehat{h}_t := \nabla_\theta f(\theta_t, \tilde{x}_t) - A(\theta_t, \tilde{x}_t)\tilde{v}_t$ and the error $e_t := \widehat{h}_t - \nabla F(\theta_t)$. Then

$$\|e_t\| \;\le\; C_x\Big(1 + \kappa_\gamma + \frac{\kappa_\gamma}{c + \gamma}\Big)\,\|\tilde{x}_t - x^\star(\theta_t)\| \;+\; C_{\mathrm{lin}} \cdot \kappa_\gamma\, \eta_t \;+\; C_{\mathrm{reg}} \cdot \gamma,$$

where

$$\kappa_\gamma \;:=\; \sup_{(\theta,x)\in\Theta\times\mathcal{V}} \big\|\big(H(\theta,x)+\gamma I\big)^{-1}\big\|$$

and one may take

$$C_{\mathrm{lin}} := \sup_{(\theta,x)\in\Theta\times\mathcal{V}} \|A(\theta,x)\|,$$

$$C_x := L_{\theta f,x} + L_{A,x}\,B_b + B_A\,L_{b,x} + B_A\,L_{H,x}\,B_b,$$

$$C_{\mathrm{reg}} := \sup_{\theta\in\Theta} \frac{\|A(\theta,x^\star(\theta))\|\,\|b(\theta,x^\star(\theta))\|}{c^2},$$

where $c>0$ is the normal spectral gap constant from Proposition 2.3,

$$B_A := \sup_{(\theta,x)\in\Theta\times\mathcal{V}} \|A(\theta,x)\|, \qquad B_b := \sup_{(\theta,x)\in\Theta\times\mathcal{V}} \|b(\theta,x)\|,$$

and $L_{\theta f,x}, L_{A,x}, L_{b,x}, L_{H,x} < \infty$ are uniform Lipschitz constants on $\mathcal{V}$ (uniform over $\theta\in\Theta$) for $x\mapsto\nabla_\theta f(\theta,x)$, $x\mapsto A(\theta,x)$, $x\mapsto b(\theta,x)$, and $x\mapsto H(\theta,x)$, respectively. Such constants exist and are finite because these maps are continuously differentiable and $\Theta\times\mathcal{V}$ is compact. We implicitly assume $\kappa_\gamma<\infty$, i.e., $H(\theta,x)+\gamma I$ is invertible for all $(\theta,x)\in\Theta\times\mathcal{V}$. If additionally $H(\theta,x)\succeq 0$ on $\Theta\times\mathcal{V}$, then $\kappa_\gamma\le 1/\gamma$.

Lemma E.3 follows by taking $\mathcal{V}=\mathbb{B}_d(0;\mathsf{D}+r(\gamma))$, restricting to the tube event $\operatorname{dist}(\tilde{x}_t,\mathcal{S}(\theta_t))\le r(\gamma)$, and applying Lemma E.1 to bound $\kappa_\gamma\le 2/\gamma$.

Proof of Lemma E.2. Write $x_t^\star := x^\star(\theta_t)$. By the global consequence of Theorem 3.5 under Assumption 3, $\nabla F(\theta_t) = h(\theta_t,x_t^\star)$. Let $v_{t,\gamma} := \big(H(\theta_t,\tilde{x}_t)+\gamma I\big)^{-1} b(\theta_t,\tilde{x}_t)$ and note that $h_\gamma(\theta_t,\tilde{x}_t) = \nabla_\theta f(\theta_t,\tilde{x}_t) - A(\theta_t,\tilde{x}_t)v_{t,\gamma}$. Add and subtract $h_\gamma(\theta_t,\tilde{x}_t)$ and $h_\gamma(\theta_t,x_t^\star)$:

$$e_t = \underbrace{\big(\widehat{h}_t - h_\gamma(\theta_t,\tilde{x}_t)\big)}_{(I)} + \underbrace{\big(h_\gamma(\theta_t,\tilde{x}_t) - h_\gamma(\theta_t,x_t^\star)\big)}_{(II)} + \underbrace{\big(h_\gamma(\theta_t,x_t^\star) - h(\theta_t,x_t^\star)\big)}_{(III)}.$$

For term (II), expand

$$\begin{aligned} h_\gamma(\theta_t,\tilde{x}_t) - h_\gamma(\theta_t,x_t^\star) \;=\; & \big(\nabla_\theta f(\theta_t,\tilde{x}_t) - \nabla_\theta f(\theta_t,x_t^\star)\big) \\ & - \Big(A(\theta_t,\tilde{x}_t)M(\theta_t,\tilde{x}_t)^{-1}b(\theta_t,\tilde{x}_t) \\ & \qquad - A(\theta_t,x_t^\star)M(\theta_t,x_t^\star)^{-1}b(\theta_t,x_t^\star)\Big), \end{aligned}$$

where $M(\theta,x) := H(\theta,x)+\gamma I$. The first difference is bounded by $L_{\theta f,x}\|\tilde{x}_t - x_t^\star\|$. For the second, add and subtract $A(\theta_t,x_t^\star)M(\theta_t,\tilde{x}_t)^{-1}b(\theta_t,\tilde{x}_t)$ to get

$$\begin{aligned} &\big\|A(\theta_t,\tilde{x}_t)M(\theta_t,\tilde{x}_t)^{-1}b(\theta_t,\tilde{x}_t) - A(\theta_t,x_t^\star)M(\theta_t,x_t^\star)^{-1}b(\theta_t,x_t^\star)\big\| \\ &\quad\le \|A(\theta_t,\tilde{x}_t) - A(\theta_t,x_t^\star)\|\,\|M(\theta_t,\tilde{x}_t)^{-1}\|\,\|b(\theta_t,\tilde{x}_t)\| \\ &\qquad + \|A(\theta_t,x_t^\star)\|\,\|M(\theta_t,\tilde{x}_t)^{-1}\|\,\|b(\theta_t,\tilde{x}_t) - b(\theta_t,x_t^\star)\| \\ &\qquad + \|A(\theta_t,x_t^\star)\|\,\big\|\big(M(\theta_t,\tilde{x}_t)^{-1} - M(\theta_t,x_t^\star)^{-1}\big)b(\theta_t,x_t^\star)\big\|. \end{aligned}$$

The first two terms are bounded by $L_{A,x}\,\kappa_\gamma\,B_b\,\|\tilde{x}_t - x_t^\star\|$ and $B_A\,\kappa_\gamma\,L_{b,x}\,\|\tilde{x}_t - x_t^\star\|$, respectively. For the last term, use the resolvent identity $M(\theta_t,\tilde{x}_t)^{-1} - M(\theta_t,x_t^\star)^{-1} = M(\theta_t,\tilde{x}_t)^{-1}\big(H(\theta_t,x_t^\star) - H(\theta_t,\tilde{x}_t)\big)M(\theta_t,x_t^\star)^{-1}$ to obtain

$$\big\|\big(M(\theta_t,\tilde{x}_t)^{-1} - M(\theta_t,x_t^\star)^{-1}\big)b(\theta_t,x_t^\star)\big\| \le \kappa_\gamma\,\|H(\theta_t,\tilde{x}_t) - H(\theta_t,x_t^\star)\|\,\big\|M(\theta_t,x_t^\star)^{-1}b(\theta_t,x_t^\star)\big\|.$$

Since $x_t^\star\in\mathcal{S}(\theta_t)$ and $b(\theta_t,x_t^\star)\in\mathcal{N}_{x_t^\star}^{\theta_t}$, Proposition 2.3 implies $\|M(\theta_t,x_t^\star)^{-1}b(\theta_t,x_t^\star)\| \le \|b(\theta_t,x_t^\star)\|/(c+\gamma) \le B_b/(c+\gamma)$. Moreover, $\|H(\theta_t,\tilde{x}_t) - H(\theta_t,x_t^\star)\| \le L_{H,x}\|\tilde{x}_t - x_t^\star\|$. Therefore, the last term is bounded by $B_A\,\kappa_\gamma\,L_{H,x}\,B_b\,\|\tilde{x}_t - x_t^\star\|/(c+\gamma)$. Combining these bounds yields

$$\|(II)\| \le C_x\Big(1 + \kappa_\gamma + \frac{\kappa_\gamma}{c+\gamma}\Big)\,\|\tilde{x}_t - x_t^\star\|.$$

For term (I), define the residual $r_t := \big(H(\theta_t, \tilde{x}_t) + \gamma I\big)\tilde{v}_t - b(\theta_t, \tilde{x}_t)$. Then $\tilde{v}_t - v_{t,\gamma} = \big(H(\theta_t, \tilde{x}_t) + \gamma I\big)^{-1} r_t$, hence

$$\|(I)\| = \|A(\theta_t, \tilde{x}_t)(\tilde{v}_t - v_{t,\gamma})\| \le \|A(\theta_t, \tilde{x}_t)\| \, \big\|\big(H(\theta_t, \tilde{x}_t) + \gamma I\big)^{-1}\big\| \, \|r_t\| \le C_{\mathrm{lin}} \, \kappa_\gamma \, \eta_t.$$

For term (III), note that $x_t^\star$ is a minimizer of $f(\theta_t, \cdot)$ over the manifold $\mathcal{S}(\theta_t)$, so the first-order condition implies $b(\theta_t, x_t^\star) \in \mathcal{N}_{x_t^\star}^{\theta_t}$. By Proposition 2.3, the restriction of $H(\theta_t, x_t^\star)$ to $\mathcal{N}_{x_t^\star}^{\theta_t}$ has eigenvalues in $[c, \infty)$. Let $\{u_i\}_{i=1}^m$ be an orthonormal eigenbasis of $\mathcal{N}_{x_t^\star}^{\theta_t}$ with $H(\theta_t, x_t^\star)u_i = \lambda_i u_i$ and $\lambda_i \ge c$. Since $b(\theta_t, x_t^\star) \in \mathcal{N}_{x_t^\star}^{\theta_t}$, we can write $b(\theta_t, x_t^\star) = \sum_{i=1}^m \beta_i u_i$. On $\mathcal{N}_{x_t^\star}^{\theta_t}$, the pseudoinverse coincides with the inverse, hence $H(\theta_t, x_t^\star)^\dagger b(\theta_t, x_t^\star) = \sum_{i=1}^m (\beta_i/\lambda_i)\, u_i$, while $\big(H(\theta_t, x_t^\star) + \gamma I\big)^{-1} b(\theta_t, x_t^\star) = \sum_{i=1}^m \big(\beta_i/(\lambda_i + \gamma)\big)\, u_i$. Therefore

$$\begin{aligned}\big\|\big(H(\theta_t, x_t^\star) + \gamma I\big)^{-1} b(\theta_t, x_t^\star) - H(\theta_t, x_t^\star)^\dagger b(\theta_t, x_t^\star)\big\| &= \Big\| \sum_{i=1}^m \beta_i \Big(\frac{1}{\lambda_i + \gamma} - \frac{1}{\lambda_i}\Big) u_i \Big\| \\ &\le \max_{i=1,\dots,m} \Big|\frac{1}{\lambda_i + \gamma} - \frac{1}{\lambda_i}\Big| \, \|b(\theta_t, x_t^\star)\|. \end{aligned} \tag{17}$$

Using $\big|\frac{1}{\lambda+\gamma} - \frac{1}{\lambda}\big| = \frac{\gamma}{\lambda(\lambda+\gamma)}$ and $\lambda_i \ge c$ gives

$$\big\|\big(H(\theta_t, x_t^\star) + \gamma I\big)^{-1} b(\theta_t, x_t^\star) - H(\theta_t, x_t^\star)^\dagger b(\theta_t, x_t^\star)\big\| \le \frac{\gamma}{c(c+\gamma)} \, \|b(\theta_t, x_t^\star)\| \le \frac{\gamma}{c^2} \, \|b(\theta_t, x_t^\star)\|.$$

Multiplying by $\|A(\theta_t, x_t^\star)\|$ and taking the supremum over $\theta \in \Theta$ yields $\|(III)\| \le C_{\mathrm{reg}}\gamma$. Combining the three bounds gives the claim. ■

Lemma E.3 (Stability of the inexact implicit term on the tube)**.** *Assume the setting of Lemma E.2 and fix $\gamma > 0$ with tube radius $r(\gamma)$ from Lemma E.1. Let $\mathcal{V} = \mathbb{B}_d(0; \mathsf{D} + r(\gamma))$. Fix an iteration t with $\theta_t \in \Theta$ and $\mathrm{dist}(\tilde{x}_t, \mathcal{S}(\theta_t)) \le r(\gamma)$, so that $\tilde{x}_t \in \mathcal{V}$. Assume clipping is inactive, i.e., $v_t = \tilde{v}_t$. Then the hyper-gradient error $e_t = \widehat{h}_t - \nabla F(\theta_t)$ obeys*

$$\|e_t\| \;\le\; A_\gamma \, \|\tilde{x}_t - x^\star(\theta_t)\| \;+\; \tfrac{2C_{\mathrm{lin}}}{\gamma} \, \eta_t \;+\; C_{\mathrm{reg}} \, \gamma,$$

where $C_{\mathrm{lin}}, C_{\mathrm{reg}}$ are as in Lemma E.2 and

$$A_\gamma := C_x \Big(1 + \tfrac{2}{\gamma} + \tfrac{2}{\gamma(c+\gamma)}\Big),$$

with $c > 0$ the normal spectral gap constant from Proposition 2.3 and C_x as in Lemma E.2.

Proof. Apply Lemma E.2 with $\mathcal{V} = \mathbb{B}_d(0; \mathsf{D} + r(\gamma))$. On the tube event, Lemma E.1 gives $\kappa_\gamma \le 2/\gamma$, and the claimed bound follows by substitution. ■

Lemma E.4 (Probability of leaving the tube)**.** *Let $X_{t,1}, \dots, X_{t,M}$ be the candidates at iteration t and let $\tilde{x}_t$ be any measurable selection of one of them (in particular, hard selection). Then for any $r > 0$,*

$$\mathbb{P}\big(\mathrm{dist}(\tilde{x}_t, \mathcal{S}(\theta_t)) > r\big) \;\le\; \frac{\mathbb{E}[\mathrm{dist}(\tilde{x}_t, \mathcal{S}(\theta_t))]}{r} \;\le\; \frac{\sum_{i=1}^M \mathbb{E}[\mathrm{dist}(X_{t,i}, \mathcal{S}(\theta_t))]}{r}.$$

Proof. The first inequality is Markov's inequality. For the second, use $\mathrm{dist}(\tilde{x}_t, \mathcal{S}(\theta_t)) \le \sum_{i=1}^M \mathrm{dist}(X_{t,i}, \mathcal{S}(\theta_t))$ and linearity of expectation. ■

Lemma E.5 (Bounding the off-tube term under clipping)**.** *Assume Algorithm 1 uses the safeguard $v_t = \mathrm{Proj}_{\mathbb{B}(0;R_v)}(\tilde{v}_t)$. Let $\mathcal{V} \subset \mathbb{R}^d$ be a compact set containing the candidates $\{X_{t,i}\}_{i=1}^M$ (and hence $\tilde{x}_t$) almost surely. Define*

$$\begin{aligned} B_{\theta f} &:= \sup_{(\theta,x)\in\Theta\times\mathcal{V}} \|\nabla_\theta f(\theta, x)\|, \\ B_A &:= \sup_{(\theta,x)\in\Theta\times\mathcal{V}} \|\nabla^2_{\theta x} g(\theta, x)\|, \\ B_b &:= \sup_{(\theta,x)\in\Theta\times\mathcal{V}} \|\nabla_x f(\theta, x)\|, \\ B_F &:= \sup_{\theta\in\Theta} \|\nabla F(\theta)\|. \end{aligned}$$

Then $\|e_t\| \le B_e := B_{\theta f} + B_A R_v + B_F$ for all t, and consequently

$$\mathbb{E}\big[\|e_t\|^2 \mathbf{1}_{\mathcal{E}_t^c}\big] \;\le\; B_e^2\, \mathbb{P}(\mathcal{E}_t^c).$$

Moreover, on $\mathcal{E}_t$ Lemma E.1 implies $\|\tilde{v}_t\| \le (2/\gamma)(B_b + \eta_t)$, so choosing $R_v \ge (2/\gamma)(B_b + \eta_t)$ ensures the safeguard is inactive on $\mathcal{E}_t$ (i.e., $v_t = \tilde{v}_t$).

Proof. Since $\|v_t\| \le R_v$ by construction and the derivative maps are continuous on the compact set $\Theta \times \mathcal{V}$, the suprema are finite and

$$\|\widehat{h}_t\| \le \|\nabla_\theta f(\theta_t, \tilde{x}_t)\| + \|\nabla^2_{\theta x} g(\theta_t, \tilde{x}_t)\|\, \|v_t\| \le B_{\theta f} + B_A R_v.$$

Since F is smooth on the compact set Θ, $B_F < \infty$, hence $\|e_t\| \le \|\widehat{h}_t\| + \|\nabla F(\theta_t)\| \le B_e$. The final claim follows from Lemma E.1 and the residual relation $(\nabla^2_{xx} g(\theta_t, \tilde{x}_t) + \gamma I)\tilde{v}_t = \nabla_x f(\theta_t, \tilde{x}_t) + r_t$ with $\|r_t\| \le \eta_t$. ■

E.1 Proof of Theorem 5.3

Proof. Fix an iteration t. Split

$$\mathbb{E}\|e_t\|^2 = \mathbb{E}\big[\|e_t\|^2 \mathbf{1}_{\mathcal{E}_t}\big] \;+\; \mathbb{E}\big[\|e_t\|^2 \mathbf{1}_{\mathcal{E}_t^c}\big].$$

On the tube. On $\mathcal{E}_t$, our choice of clipping radius ensures $v_t = \tilde{v}_t$ (Lemma E.5). Therefore Lemma E.3 gives

$$\|e_t\| \;\le\; A_\gamma\, \|\tilde{x}_t - x^\star(\theta_t)\| \;+\; \tfrac{2C_{\mathrm{lin}}}{\gamma}\, \eta_t \;+\; C_{\mathrm{reg}}\, \gamma.$$

Using $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ and that $\mathbf{1}_{\mathcal{E}_t} \le 1$ yields

$$\mathbb{E}\big[\|e_t\|^2 \mathbf{1}_{\mathcal{E}_t}\big] \;\le\; 3A_\gamma^2\, \mathbb{E}\big[\|\tilde{x}_t - x^\star(\theta_t)\|^2 \mathbf{1}_{\mathcal{E}_t}\big] \;+\; \tfrac{12C_{\mathrm{lin}}^2}{\gamma^2}\, \eta_t^2 \;+\; 3C_{\mathrm{reg}}^2\, \gamma^2.$$

Recalling $A_\gamma = C_x(1 + \frac{2}{\gamma} + \frac{2}{\gamma(c+\gamma)})$ from Lemma E.3 gives the first three terms in (8).

Off the tube. Lemma E.5 yields $\|e_t\| \le B_e$ for all t, and thus

$$\mathbb{E}\big[\|e_t\|^2 \mathbf{1}_{\mathcal{E}_t^c}\big] \le B_e^2\, \mathbb{P}(\mathcal{E}_t^c).$$

Combining the two bounds gives (8). ■

F Quality of the selected lower-level point

This appendix provides the main technical details behind the selection bound in Theorem 5.4 and the related discussion in Section 5. Fix $\theta \in \Theta$ and recall

$$\mathcal{S}(\theta) = \arg\min_{x \in \mathbb{R}^d} g(\theta, x), \qquad F(\theta) = \min_{x \in \mathcal{S}(\theta)} f(\theta, x).$$

Throughout, assume that the constrained problem admits a unique optimistic minimizer

$$x^\star(\theta) \in \arg\min_{x \in \mathcal{S}(\theta)} f(\theta, x),$$

as in Assumption 3. Given candidates $X_1, \dots, X_M \in \mathbb{R}^d$, our hard selection rule is

$$\tilde{x} \in \arg\min_{i \in [1:M]} f(\theta, X_i), \qquad \delta := \frac{1}{M}.$$

Since $\tilde{x}$ need not lie on $\mathcal{S}(\theta)$, we also define its projection

$$\bar{x} \in \arg\min_{x \in \mathcal{S}(\theta)} \|x - \tilde{x}\|, \qquad \text{so that} \qquad \|\tilde{x} - \bar{x}\| = \operatorname{dist}(\tilde{x}, \mathcal{S}(\theta)).$$

Empirical and Gibbs measures. Define the empirical measure of the candidates,

$$\widehat{\nu}_M := \frac{1}{M} \sum_{i=1}^{M} \delta_{X_i},$$

and recall the Gibbs measure used to sample near $\mathcal{S}(\theta)$,

$$\mu_\theta^\lambda(\mathrm{d}x) \propto \exp\{-g(\theta, x)/\lambda\}\, \mathrm{d}x.$$

Lower-tail superquantile (best-δ CVaR). We use the lower-tail superquantile functional $\mathrm{SQ}_\delta^{\mathrm{low}}$.

Definition F.1 (Lower δ-superquantile)**.** *Let Z be a real-valued random variable with (lower) quantile function $q_u(Z) := \inf\{t \in \mathbb{R} : \mathbb{P}(Z \le t) \ge u\}$ for $u \in (0,1)$. Define*

$$\mathrm{SQ}_\delta^{\mathrm{low}}(Z) := \frac{1}{\delta}\int_0^\delta q_u(Z)\,\mathrm{d}u, \qquad \delta \in (0,1).$$

If $X \sim \nu$ and $Z = f(\theta, X)$, we also write $\mathrm{SQ}_\delta^{\mathrm{low}}(f(\theta,X);\, X \sim \nu)$.

F.1 From hard selection to a value gap on $\mathcal{S}(\theta)$

Step 1: tail-min is controlled by an empirical lower superquantile. Hard selection is a deterministic lower-tail operation, and its output is always below the corresponding (lower) superquantile under the empirical law.

Lemma F.2 (Tail-selection $\Rightarrow$ superquantile control)**.** *Fix θ, candidates $X_1, \dots, X_M$, and $\delta \in (0,1)$. Let $\tilde{x} \in \arg\min_{i\in[1:M]} f(\theta, X_i)$. Then*

$$f(\theta, \tilde{x}) \;\le\; \mathrm{SQ}_\delta^{\mathrm{low}}\big(f(\theta,X);\, X \sim \widehat{\nu}_M\big).$$

Step 2: lifting $\tilde{x}$ back to $\mathcal{S}(\theta)$ via Lipschitzness. We will compare $x^\star(\theta)$ to $\bar{x} \in \mathcal{S}(\theta)$ (the projection of $\tilde{x}$) and separate (i) an off-manifold error $\|\tilde{x} - \bar{x}\|$ from (ii) an on-manifold suboptimality gap at $\bar{x}$.

Lemma F.3 (Suboptimality gap on $\mathcal{S}(\theta)$)**.** *Fix θ and let $x^\star(\theta)$ be the unique minimizer of $f(\theta,\cdot)$ over $\mathcal{S}(\theta)$. Assume $x \mapsto f(\theta,x)$ is $L_{f,1}$-Lipschitz (as in Assumption 4). Let $\tilde{x}$ be the output of hard selection and let $\bar{x}$ be its projection onto $\mathcal{S}(\theta)$. Then*

$$\begin{aligned} 0 \;\le\; f(\theta,\bar{x}) - f(\theta,x^\star(\theta)) \;&\le L_{f,1}\,\mathrm{dist}(\tilde{x},\mathcal{S}(\theta)) \\ &\quad + \Big[\mathrm{SQ}_\delta^{\mathrm{low}}\big(f(\theta,X);\, X\sim\widehat{\nu}_M\big) - f(\theta,x^\star(\theta))\Big]. \end{aligned} \tag{18}$$

Moreover, for any reference measure ν on $\mathbb{R}^d$,

$$\mathrm{SQ}_\delta^{\mathrm{low}}\big(f(\theta,X);\, X\sim\widehat{\nu}_M\big) - \mathrm{SQ}_\delta^{\mathrm{low}}\big(f(\theta,X);\, X\sim\nu\big) \;\le\; \frac{L_{f,1}}{\delta}\,\mathsf{W}_1(\widehat{\nu}_M,\nu), \tag{19}$$

and consequently, choosing $\nu = \mu_\theta^\lambda$ yields

$$\begin{aligned} f(\theta,\bar{x}) - f(\theta,x^\star(\theta)) \;&\le L_{f,1}\,\mathrm{dist}(\tilde{x},\mathcal{S}(\theta)) \\ &\quad + \frac{L_{f,1}}{\delta}\,\mathsf{W}_1(\widehat{\nu}_M,\mu_\theta^\lambda) + \Big[\mathrm{SQ}_\delta^{\mathrm{low}}\big(f(\theta,X);\, X\sim\mu_\theta^\lambda\big) - f(\theta,x^\star(\theta))\Big]. \end{aligned} \tag{20}$$

A hard-min bound under Gibbs sampling. *Assume additionally that $X_1,\dots,X_M \overset{\text{i.i.d.}}{\sim} \mu_\theta^\lambda$ and that $\tilde{x} \in \arg\min_i f(\theta,X_i)$. Then there exists a constant $C_1 > 0$ (depending only on the on-manifold geometry near $x^\star(\theta)$) such that*

$$\mathbb{E}\big[f(\theta,\bar{x}) - F(\theta)\big] \;\le\; 2L_{f,1}\,C_{\mathrm{tube}}\,\sqrt{\lambda\log(1+M)} \;+\; C_1\,M^{-1/\mathfrak{k}}, \tag{21}$$

where C_{tube} is the tube constant from Lemma F.5. More generally, if the candidates are independent with marginal laws ν_i satisfying $\mathsf{W}_2(\nu_i,\mu_\theta^\lambda) < \infty$ for all $i \in [1:M]$, then (21) *holds with an additional additive term*

$$2L_{f,1}\left(\sum_{i=1}^M \mathsf{W}_2(\nu_i,\mu_\theta^\lambda)^2\right)^{1/2}.$$

Proof. The deterministic bound (20) is valid for any δ, but when $\delta = 1/M$ it yields the loose factor $\delta^{-1} = M$ multiplying the empirical discrepancy term. We therefore analyze the hard-min selection directly. Write $\tilde{x} \in \arg\min_i f(\theta,X_i)$. By Lipschitzness and the definition of $\bar{x}$,

$$f(\theta,\bar{x}) \;\le\; f(\theta,\tilde{x}) + L_{f,1}\,\mathrm{dist}(\tilde{x},\mathcal{S}(\theta)).$$

Taking expectations gives

$$\mathbb{E}\big[f(\theta,\bar{x}) - F(\theta)\big] \;\le\; L_{f,1}\,\mathbb{E}\big[\mathrm{dist}(\tilde{x},\mathcal{S}(\theta))\big] + \mathbb{E}\Big[\min_{1\le i\le M} f(\theta,X_i)\Big] - F(\theta). \tag{22}$$

Hard-min bound under Gibbs sampling. Assume first that $X_1, \dots, X_M \overset{\text{i.i.d.}}{\sim} \mu_\theta^\lambda$. For each i, let $\Pi(X_i) \in \arg\min_{x \in \mathcal{S}(\theta)} \|x - X_i\|$ be a Euclidean projection onto $\mathcal{S}(\theta)$ (choose any measurable selection). By Lipschitzness,

$$f(\theta, X_i) \;\le\; f(\theta, \Pi(X_i)) + L_{f,1} \operatorname{dist}(X_i, \mathcal{S}(\theta)).$$

Taking the minimum over i and using $\min_i(a_i + b_i) \le \min_i a_i + \max_i b_i$ gives

$$\min_{1 \le i \le M} f(\theta, X_i) \;\le\; \min_{1 \le i \le M} f(\theta, \Pi(X_i)) + \; L_{f,1} \max_{1 \le i \le M} \operatorname{dist}(X_i, \mathcal{S}(\theta)).$$

Plugging this into (22) and using $\operatorname{dist}(\tilde{x}, \mathcal{S}(\theta)) \le \max_i \operatorname{dist}(X_i, \mathcal{S}(\theta))$ yields

$$\mathbb{E}\big[f(\theta, \bar{x}) - F(\theta)\big] \;\le\; 2L_{f,1}\, \mathbb{E}\Big[\max_{1 \le i \le M} \operatorname{dist}(X_i, \mathcal{S}(\theta))\Big] + \Big(\mathbb{E}\Big[\min_{1 \le i \le M} f(\theta, \Pi(X_i))\Big] - F(\theta)\Big). \quad (23)$$

The first term is controlled by the subgaussian tube concentration in Lemma F.5, giving the $\sqrt{\lambda \log(1 + M)}$ scaling.

For the on-manifold term, note that the projected points $\Pi(X_i)$ lie on $\mathcal{S}(\theta)$ and, under our standing regularity assumptions, the induced law on $\mathcal{S}(\theta)$ admits a continuous density bounded away from 0 in a neighborhood of $x^\star(\theta)$ (since μ_θ^λ concentrates in a tube and its pushforward converges to μ_θ^0; see [Masiha et al., 2025, Prop. 3.7]). Combined with geodesic volume scaling on $\mathcal{S}(\theta)$ [Masiha et al., 2025, Lem. 4.1], this yields a small-ball mass bound of the form $\mathbb{P}(d_{\mathcal{S}(\theta)}(\Pi(X_1), x^\star(\theta)) \le r) \ge c_0 r^{\mathtt{k}}$ for all sufficiently small $r > 0$. The nearest-neighbor argument (using this small-ball bound) then gives $\mathbb{E}[\min_i f(\theta, \Pi(X_i))] - F(\theta) \le C_1 M^{-1/\mathtt{k}}$ for a constant C_1 depending only on $(c_0, \mathtt{k}, \mathsf{D}, L_{f,1})$. Substituting these two bounds into (23) yields (21).

Adding an LMC sampling tolerance. If instead $X_1, \dots, X_M$ are independent with marginal laws ν_i satisfying $\mathsf{W}_2(\nu_i, \mu_\theta^\lambda) < \infty$ for all $i \in [1 : M]$, then for each i we may couple X_i with an independent $X_i^\lambda \sim \mu_\theta^\lambda$ so that $\mathbb{E}\|X_i - X_i^\lambda\|^2 \le \mathsf{W}_2(\nu_i, \mu_\theta^\lambda)^2$. Using

$$\operatorname{dist}(\tilde{x}, \mathcal{S}(\theta)) \le \max_{1 \le i \le M} \operatorname{dist}(X_i, \mathcal{S}(\theta)) \le \max_{1 \le i \le M} \operatorname{dist}(X_i^\lambda, \mathcal{S}(\theta)) + \max_{1 \le i \le M} \|X_i - X_i^\lambda\|$$

and

$$\min_{1 \le i \le M} f(\theta, X_i) \le \min_{1 \le i \le M} f(\theta, X_i^\lambda) + L_{f,1} \max_{1 \le i \le M} \|X_i - X_i^\lambda\|,$$

the exact-Gibbs argument above yields an additional term

$$2L_{f,1}\, \mathbb{E}\Big[\max_{1 \le i \le M} \|X_i - X_i^\lambda\|\Big].$$

Finally, $\max_i a_i \le (\sum_i a_i^2)^{1/2}$ gives

$$\mathbb{E}\Big[\max_{1 \le i \le M} \|X_i - X_i^\lambda\|\Big] \le \Big(\sum_{i=1}^M \mathbb{E}\|X_i - X_i^\lambda\|^2\Big)^{1/2} \le \left(\sum_{i=1}^M \mathsf{W}_2(\nu_i, \mu_\theta^\lambda)^2\right)^{1/2},$$

which proves the claimed additive term. ∎

F.2 Controlling the off-manifold term $\operatorname{dist}(\tilde{x}, \mathcal{S}(\theta))$

Step 3: a max-distance reduction. Because $\tilde{x}$ is one of the candidates, it satisfies the deterministic inequality

$$\operatorname{dist}(\tilde{x}, \mathcal{S}(\theta)) \;\le\; \max_{1 \le i \le M} \operatorname{dist}(X_i, \mathcal{S}(\theta)). \quad (24)$$

We now show that, under our standing assumptions, Gibbs samples concentrate around $\mathcal{S}(\theta)$ with a *subgaussian* tail in the normal distance $R := \operatorname{dist}(X, \mathcal{S}(\theta))$, which yields the sharper moment bound $\mathbb{E}[\max_i R_i] \lesssim \sqrt{\lambda \log M}$.

Step 4: a uniform quadratic growth bound for $g(\theta,\cdot)$. Recall from Assumption 4 that $g(\theta,\cdot)$ satisfies a quadratic growth bound *outside* a compact ball. The next lemma shows that, under local PŁ° and normal nondegeneracy on the minimizer manifold, this can be upgraded to a *global* quadratic growth inequality in terms of the distance to $\mathcal{S}(\theta)$, uniformly over $\theta \in \Theta$.

Lemma F.4 (Uniform quadratic growth of $g(\theta,\cdot)$ over Θ)**.** *Let Assumptions 1 and 4 hold. Then there exists a constant $\mu_{\mathrm{QG}} > 0$ such that for all $\theta \in \Theta$ and all $x \in \mathbb{R}^d$,*

$$g(\theta,x) - \min_{z\in\mathbb{R}^d} g(\theta,z) \;\geq\; \frac{\mu_{\mathrm{QG}}}{2}\,\mathrm{dist}^2\big(x,\mathcal{S}(\theta)\big). \tag{25}$$

Proof. Let D be the compactness radius from Assumption 4, so that $\mathcal{S}(\theta) \subseteq \mathbb{B}_d(0;\mathsf{D})$ for all $\theta \in \Theta$. Write $g_\star(\theta) := \min_z g(\theta,z)$.

Step 1: local quadratic growth in a uniform tube. Let $c > 0$ be the normal spectral gap from Proposition 2.3, i.e., $\langle v, \nabla^2_{xx} g(\theta,y) v\rangle \geq c\|v\|^2$ for all $\theta \in \Theta$, $y \in \mathcal{S}(\theta)$, and $v \in \mathcal{N}^\theta_y$. Since $(\theta,x) \mapsto \nabla^2_{xx} g(\theta,x)$ is continuous (Assumption 4) and the set $\mathcal{K} := \{(\theta,y) : \theta \in \Theta,\ y \in \mathcal{S}(\theta)\}$ is compact, there exists $r_{\mathrm{loc}} > 0$ such that for all $\theta \in \Theta$, all $y \in \mathcal{S}(\theta)$, and all x with $\|x-y\| \leq r_{\mathrm{loc}}$,

$$\langle v, \nabla^2_{xx} g(\theta,x) v\rangle \;\geq\; \tfrac{c}{2}\,\|v\|^2 \qquad \forall v \in \mathcal{N}^\theta_y. \tag{26}$$

Now fix any $\theta \in \Theta$ and $x \in \mathbb{R}^d$ with $R := \mathrm{dist}(x,\mathcal{S}(\theta)) \leq r_{\mathrm{loc}}$. Let $y \in \mathcal{S}(\theta)$ be a Euclidean projection of x onto $\mathcal{S}(\theta)$, so that $\|x-y\| = R$. By first-order optimality of the projection problem on the manifold, $u := x - y$ lies in the normal space $\mathcal{N}^\theta_y$. Using $\nabla_x g(\theta,y) = 0$ (since $y \in \mathcal{S}(\theta)$) and the integral form of Taylor's theorem,

$$g(\theta,x) - g(\theta,y) = \int_0^1 (1-t)\, u^\top \nabla^2_{xx} g(\theta, y+tu)\, u \,\mathrm{d}t.$$

Because $\|tu\| \leq R \leq r_{\mathrm{loc}}$, (26) applies along the segment $y + tu$, giving $u^\top \nabla^2_{xx} g(\theta, y+tu) u \geq (c/2)\|u\|^2$ for all $t \in [0,1]$. Therefore,

$$g(\theta,x) - g_\star(\theta) \geq g(\theta,x) - g(\theta,y) \geq \tfrac{c}{4}\,\|u\|^2 = \tfrac{c}{4}\,R^2. \tag{27}$$

Step 2: quadratic growth outside the compact ball. By Assumption 4, there exist constants $\mu_{\mathsf{qg}} > 0$ and $\mathsf{D} > 0$ such that for all $\|x\| \geq \mathsf{D}$,

$$g(\theta,x) - g_\star(\theta) \;\geq\; \tfrac{\mu_{\mathsf{qg}}}{2}\,\mathrm{dist}^2\big(x,\mathcal{S}(\theta)\big), \qquad \forall \theta \in \Theta. \tag{28}$$

Step 3: the middle region by compactness. Consider the set

$$\mathcal{A} := \Big\{(\theta,x) : \theta \in \Theta,\ \|x\| \leq \mathsf{D},\ \mathrm{dist}(x,\mathcal{S}(\theta)) \geq r_{\mathrm{loc}}\Big\}.$$

Since $\Theta \times \mathbb{B}_d(0;\mathsf{D})$ is compact and $(\theta,x) \mapsto \mathrm{dist}(x,\mathcal{S}(\theta))$ is lower semicontinuous (by compactness of $\mathcal{K}$ and existence of Euclidean projections), $\mathcal{A}$ is closed and hence compact. The map $(\theta,x) \mapsto g(\theta,x) - g_\star(\theta)$ is continuous on $\Theta \times \mathbb{B}_d(0;\mathsf{D})$, and is strictly positive on $\mathcal{A}$ (since $\mathrm{dist}(x,\mathcal{S}(\theta)) > 0$ implies $x \notin \mathcal{S}(\theta)$). Therefore, $m_{\min} := \min_{(\theta,x)\in\mathcal{A}} \big(g(\theta,x) - g_\star(\theta)\big) > 0$. Moreover, if $\|x\| \leq \mathsf{D}$ then $\mathrm{dist}(x,\mathcal{S}(\theta)) \leq 2\mathsf{D}$, so on $\mathcal{A}$ we have

$$g(\theta,x) - g_\star(\theta) \;\geq\; m_{\min} \;\geq\; \frac{m_{\min}}{(2\mathsf{D})^2}\,\mathrm{dist}^2(x,\mathcal{S}(\theta)).$$

Step 4: combine the three regions. Let

$$\mu_{\mathrm{QG}} := \min\Big\{\tfrac{c}{2},\ \mu_{\mathsf{qg}},\ \tfrac{2m_{\min}}{(2\mathsf{D})^2}\Big\}.$$

Then (27), (28), and the middle-region bound above imply (25) for all $x \in \mathbb{R}^d$ and all $\theta \in \Theta$. ∎

Step 5: subgaussian tube concentration for Gibbs samples. The uniform quadratic growth bound from Lemma F.4 implies that Gibbs samples have Gaussian-like concentration in the normal distance to $\mathcal{S}(\theta)$.

Lemma F.5 (Tube concentration for Gibbs samples)**.** *Assume the setting of Lemma F.4 and fix any $\theta \in \Theta$. Let $X \sim \mu_\theta^\lambda$ and define $R := \operatorname{dist}(X, \mathcal{S}(\theta))$. Then there exist constants $c_{\mathrm{tube}}, C_{\mathrm{tube}}, C_{\mathrm{tube},2} > 0$ (independent of θ and λ) such that*

$$\mathbb{P}(R \geq t) \;\leq\; C_{\mathrm{tube}} \exp\Big(- c_{\mathrm{tube}} \, \tfrac{t^2}{\lambda} \Big), \qquad \forall t \geq 0. \tag{29}$$

Consequently,

$$\mathbb{E}[R] \leq C_{\mathrm{tube}} \sqrt{\lambda}, \qquad \mathbb{E}[R^2] \leq C_{\mathrm{tube},2} \lambda.$$

Moreover, if $X_1, \dots, X_M \overset{\text{i.i.d.}}{\sim} \mu_\theta^\lambda$ and $\tilde{x}$ is any measurable selection of one of the candidates, then

$$\mathbb{E}\big[\operatorname{dist}(\tilde{x}, \mathcal{S}(\theta))\big] \;\leq\; \mathbb{E}\Big[\max_{1 \leq i \leq M} \operatorname{dist}(X_i, \mathcal{S}(\theta)) \Big] \;\leq\; C_{\mathrm{tube}} \sqrt{\lambda \, \log(1+M)}. \tag{30}$$

$$\mathbb{E}\big[\operatorname{dist}(\tilde{x}, \mathcal{S}(\theta))^2\big] \;\leq\; \mathbb{E}\Big[\max_{1 \leq i \leq M} \operatorname{dist}(X_i, \mathcal{S}(\theta))^2 \Big] \;\leq\; C_{\mathrm{tube},2} \, \lambda \, \log(1+M). \tag{31}$$

Approximate candidates (Rényi control). *If instead $X_1, \dots, X_M$ are candidates whose (not necessarily identical) marginal laws ν_i satisfy $R_2(\nu_i \| \mu_\theta^\lambda) \leq \varepsilon_{\mathrm{R}}^2$, then the same argument gives a subgaussian tail (with adjusted constants) and hence*

$$\mathbb{E}\big[\operatorname{dist}(\tilde{x}, \mathcal{S}(\theta))\big] \;\leq\; \mathbb{E}\Big[\max_{1 \leq i \leq M} \operatorname{dist}(X_i, \mathcal{S}(\theta)) \Big] \;\leq\; C'_{\mathrm{tube}} \sqrt{\lambda \, \log(1+M)}, \tag{32}$$

and

$$\mathbb{E}\big[\operatorname{dist}(\tilde{x}, \mathcal{S}(\theta))^2\big] \;\leq\; \mathbb{E}\Big[\max_{1 \leq i \leq M} \operatorname{dist}(X_i, \mathcal{S}(\theta))^2 \Big] \;\leq\; C'_{\mathrm{tube},2} \, \lambda \, \log(1+M), \tag{33}$$

where one may take $C'_{\mathrm{tube}} := C_{\mathrm{tube}} \exp(\varepsilon_{\mathrm{R}}^2/2)$ and $C'_{\mathrm{tube},2} := C_{\mathrm{tube},2} \exp(\varepsilon_{\mathrm{R}}^2/2)$ (up to universal numerical factors).

Proof. Fix θ and abbreviate $S := \mathcal{S}(\theta)$ and $g_\star := \min_z g(\theta, z)$. By Lemma F.4, $g(\theta, x) \geq g_\star + \frac{\mu_{\mathrm{QG}}}{2} \operatorname{dist}^2(x, S)$. Therefore, for any $t \geq 0$,

$$\mathbb{P}(R \geq t) = \frac{\int_{\operatorname{dist}(x,S) \geq t} e^{-g(\theta,x)/\lambda} \, \mathrm{d}x}{\int_{\mathbb{R}^d} e^{-g(\theta,x)/\lambda} \, \mathrm{d}x} \leq \frac{\int_{\operatorname{dist}(x,S) \geq t} \exp\big(- \frac{\mu_{\mathrm{QG}}}{2\lambda} \operatorname{dist}^2(x,S) \big) \, \mathrm{d}x}{\int_{\mathbb{R}^d} e^{-(g(\theta,x)-g_\star)/\lambda} \, \mathrm{d}x}.$$

To lower bound the denominator, fix any $y \in S$ and use smoothness of g on the ball $\mathbb{B}_d(0; \mathsf{D}+1)$ (Assumption 4) to obtain a constant $L_{\max} < \infty$ such that $g(\theta, y+u) - g_\star \leq \frac{L_{\max}}{2} \|u\|^2$ for all $\|u\| \leq 1$. Thus,

$$\int_{\mathbb{R}^d} e^{-(g(\theta,x)-g_\star)/\lambda} \, \mathrm{d}x \;\geq\; \int_{\|u\| \leq 1} \exp\Big(- \tfrac{L_{\max}}{2\lambda} \|u\|^2 \Big) \, \mathrm{d}u \;\geq\; c_0 \, \lambda^{d/2}$$

for a numerical constant $c_0 > 0$ depending only on $(d, L_{\max})$. For the numerator, we use a crude volume bound for distance shells around S. Since $S \subseteq \mathbb{B}_d(0; \mathsf{D})$, its r-neighborhood satisfies $\{x : \operatorname{dist}(x, S) \leq r\} \subseteq \mathbb{B}_d(0; \mathsf{D}+r)$ for all $r \geq 0$. Consequently, for any $\Delta r > 0$,

$$\begin{aligned} \operatorname{Vol}\Big(\{x : \operatorname{dist}(x,S) \in [r, r+\Delta r]\} \Big) &\leq \operatorname{Vol}\big(\mathbb{B}_d(0; \mathsf{D}+r+\Delta r) \big) - \operatorname{Vol}\big(\mathbb{B}_d(0; \mathsf{D}+r) \big) \\ &\leq C_d \, (\mathsf{D}+r)^{d-1} \, \Delta r, \end{aligned}$$

where $C_d > 0$ depends only on d. Using a layer-cake/coarea bound with $a := \mu_{\mathrm{QG}}/(2\lambda)$,

$$\int_{\operatorname{dist}(x,S) \geq t} e^{-a \operatorname{dist}^2(x,S)} \, \mathrm{d}x \;\leq\; C_d \int_{r=t}^{\infty} (\mathsf{D}+r)^{d-1} e^{-a r^2} \, \mathrm{d}r \;\leq\; C_0 \, \lambda^{d/2} \exp\Big(- \tfrac{\mu_{\mathrm{QG}}}{4\lambda} \, t^2 \Big),$$

where the last inequality is a standard Gaussian tail estimate (the polynomial prefactor is absorbed into the exponential). Combining the numerator/denominator bounds yields (29) with constants depending only on $(d, \mathsf{D}, \mu_{\mathrm{QG}}, L_{\max})$. Integrating (29) over $t \geq 0$ yields $\mathbb{E}[R] \leq C_{\mathrm{tube}} \sqrt{\lambda}$. Similarly,

$$\mathbb{E}[R^2] = 2 \int_0^\infty t \, \mathbb{P}(R \geq t) \, \mathrm{d}t \leq 2 C_{\mathrm{tube}} \int_0^\infty t \, \exp\Big(- c_{\mathrm{tube}} \, \tfrac{t^2}{\lambda} \Big) \, \mathrm{d}t \leq C_{\mathrm{tube},2} \lambda$$

after increasing the constant if needed.

Finally, for (30), use (24) and write

$$\mathbb{P}\Big(\max_{1\le i\le M} R_i \ge t\Big) \le \sum_{i=1}^{M} \mathbb{P}(R_i \ge t) \le M\, C_{\mathrm{tube}} \exp\Big(-c_{\mathrm{tube}}\, \tfrac{t^2}{\lambda}\Big).$$

To bound the expectation, use the tail integral representation

$$\mathbb{E}\Big[\max_{1\le i\le M} R_i\Big] = \int_0^\infty \mathbb{P}\Big(\max_{1\le i\le M} R_i \ge t\Big)\, \mathrm{d}t.$$

Let $t_0 := \sqrt{\frac{\lambda}{c_{\mathrm{tube}}} \log(1+M)}$. Splitting the integral and using the bound above yields

$$\mathbb{E}\Big[\max_{1\le i\le M} R_i\Big] \le t_0 + \int_{t_0}^\infty M\, C_{\mathrm{tube}} \exp\Big(-c_{\mathrm{tube}}\, \tfrac{t^2}{\lambda}\Big)\, \mathrm{d}t.$$

Using $\int_{t_0}^\infty e^{-at^2}\, \mathrm{d}t \le \frac{1}{2at_0}\, e^{-at_0^2}$ for $a > 0$ (with $a = c_{\mathrm{tube}}/\lambda$) gives

$$\begin{aligned}\int_{t_0}^\infty M\, C_{\mathrm{tube}} \exp\Big(-c_{\mathrm{tube}}\, \tfrac{t^2}{\lambda}\Big)\, \mathrm{d}t &\le M\, C_{\mathrm{tube}} \cdot \frac{\lambda}{2c_{\mathrm{tube}}\, t_0} \exp\Big(-c_{\mathrm{tube}}\, \tfrac{t_0^2}{\lambda}\Big) \\ &= \frac{M}{1+M} \cdot \frac{C_{\mathrm{tube}}}{2c_{\mathrm{tube}}} \cdot \frac{\lambda}{t_0} \\ &\le \frac{C_{\mathrm{tube}}}{2c_{\mathrm{tube}}} \frac{\sqrt{\lambda}}{\sqrt{\log(1+M)}}.\end{aligned}$$

Since $\log(1+M) \ge 1$, we conclude that $\mathbb{E}[\max_i R_i] \le C_{\mathrm{tube}} \sqrt{\lambda \log(1+M)}$ after adjusting C_{tube}. Similarly,

$$\mathbb{E}\Big[\max_{1\le i\le M} R_i^2\Big] = 2\int_0^\infty t\, \mathbb{P}\Big(\max_{1\le i\le M} R_i \ge t\Big)\, \mathrm{d}t.$$

Splitting at the same t_0 and using $\int_{t_0}^\infty t e^{-at^2}\, \mathrm{d}t = \frac{1}{2a} e^{-at_0^2}$ with $a = c_{\mathrm{tube}}/\lambda$, we get

$$\mathbb{E}\Big[\max_{1\le i\le M} R_i^2\Big] \le t_0^2 + \frac{MC_{\mathrm{tube}}\lambda}{c_{\mathrm{tube}}} \exp\Big(-c_{\mathrm{tube}}\, \tfrac{t_0^2}{\lambda}\Big) \le C_{\mathrm{tube},2}\lambda \log(1+M),$$

which yields (31) after increasing $C_{\mathrm{tube},2}$.

For (32), fix any candidate index i and let $A_t := \{R \ge t\}$. Under $R_2(\nu_i \| \mu_\theta^\lambda) \le \varepsilon_{\mathrm{R}}^2$, Cauchy–Schwarz yields

$$\nu_i(A_t) = \int \mathbf{1}_{A_t}\, \frac{\mathrm{d}\nu_i}{\mathrm{d}\mu_\theta^\lambda}\, \mathrm{d}\mu_\theta^\lambda \le e^{\varepsilon_{\mathrm{R}}^2/2}\, \mu_\theta^\lambda(A_t)^{1/2}.$$

Combining this with the Gibbs tail (29) gives $\mathbb{P}(R_i \ge t) \le e^{\varepsilon_{\mathrm{R}}^2/2} \sqrt{C_{\mathrm{tube}}} \exp\big(-\frac{c_{\mathrm{tube}}}{2}\frac{t^2}{\lambda}\big)$, and hence (by a union bound)

$$\mathbb{P}\Big(\max_{1\le i\le M} R_i \ge t\Big) \le M\, e^{\varepsilon_{\mathrm{R}}^2/2} \sqrt{C_{\mathrm{tube}}} \exp\Big(-\tfrac{c_{\mathrm{tube}}}{2}\tfrac{t^2}{\lambda}\Big).$$

Integrating this tail bound exactly as above (splitting at $t_0' := \sqrt{\frac{2\lambda}{c_{\mathrm{tube}}} \log(1+M)}$) yields (32). The same argument with the identity

$$\mathbb{E}\Big[\max_{1\le i\le M} R_i^2\Big] = 2\int_0^\infty t\, \mathbb{P}\Big(\max_{1\le i\le M} R_i \ge t\Big)\, \mathrm{d}t$$

and the same splitting point t_0' gives (33) (again with adjusted constants). ■

F.3 Interpreting the remaining terms in (20)

Step 5: empirical + mixing + discretization decomposition. To interpret $\mathsf{W}_1(\widehat{\nu}_M, \mu_\theta^\lambda)$, introduce the law $\nu_{\theta,T,h}^\lambda$ of a finite-step ULA/LMC chain and its stationary law $\nu_{\theta,h}^\lambda$, so that

$$\mathsf{W}_1(\widehat{\nu}_M, \mu_\theta^\lambda) \;\le\; \underbrace{\mathsf{W}_1(\widehat{\nu}_M, \nu_{\theta,T,h}^\lambda)}_{\text{empirical}} + \underbrace{\mathsf{W}_1(\nu_{\theta,T,h}^\lambda, \nu_{\theta,h}^\lambda)}_{\text{mixing}} + \underbrace{\mathsf{W}_1(\nu_{\theta,h}^\lambda, \mu_\theta^\lambda)}_{\text{discretization}}.$$

Step 6: Gibbs/superquantile bias on the manifold (typical rate). Define the Gibbs/superquantile bias

$$\varepsilon_{\mathrm{Gibbs}}(\theta;\lambda,\delta) := \mathrm{SQ}_{\delta}^{\mathrm{low}}\big(f(\theta,X);\, X\sim\mu_{\theta}^{\lambda}\big) - f(\theta,x^{\star}(\theta)).$$

Under the manifold regularity assumptions in [Masiha et al., 2025], this bias is controlled by an intrinsic tail term and a tube-width term.

Lemma F.6 (Superquantile–Gibbs bias (typical rate))**.** *Assume the setting of Appendix F and that $x \mapsto f(\theta,x)$ is $L_{f,1}$-Lipschitz. Then, under the manifold regularity assumptions in the Superquantile–Gibbs analysis of [Masiha et al., 2025], there exist constants $C_1, C_2 > 0$ such that for all sufficiently small $\lambda > 0$ and all $\delta \in (0,1)$,*

$$0 \;\le\; \varepsilon_{\mathrm{Gibbs}}(\theta;\lambda,\delta) \;\le\; C_1\,\delta^{1/\mathfrak{k}} + C_2\,\frac{\sqrt{\lambda}}{\delta}.$$

F.4 Converting the on-manifold gap to distance (link to local identifiability)

Step 7: Quadratic growth on $\mathcal{S}(\theta)$ from nondegeneracy. To convert an on-manifold value gap into a distance bound, we use a local growth inequality for the restriction of $f(\theta,\cdot)$ to the manifold $\mathcal{S}(\theta)$ around its optimistic minimizer $x^{\star}(\theta)$. The next lemma shows that, in our setting, this holds automatically with a *quadratic* exponent whenever the constrained minimizer is nondegenerate.

Lemma F.7 (Quadratic growth on $\mathcal{S}(\theta)$ implied by Assumption 3)**.** *Fix θ and suppose that $f(\theta,\cdot)$ is $\mathcal{C}^2$ in a neighborhood of $\mathcal{S}(\theta)$ and that $x^{\star}(\theta) \in \mathcal{S}(\theta)$ satisfies Assumption 3 (nondegenerate local minimizer on the manifold). Then there exist constants $c_{\mathrm{hg}} > 0$ and $r_0 > 0$ such that for all $x \in \mathcal{S}(\theta)$ with $d_{\mathcal{S}(\theta)}(x, x^{\star}(\theta)) \le r_0$,*

$$f(\theta,x) - f(\theta,x^{\star}(\theta)) \;\ge\; c_{\mathrm{hg}}\; d_{\mathcal{S}(\theta)}(x,x^{\star}(\theta))^2. \tag{34}$$

Proof. Let $\mathcal{M} := \mathcal{S}(\theta)$ be the embedded submanifold with the induced Riemannian metric, and define the restriction $\bar{f} : \mathcal{M} \to \mathbb{R}$ by $\bar{f}(x) := f(\theta,x)$. Since $f(\theta,\cdot)$ is $\mathcal{C}^2$ on a neighborhood of $\mathcal{M}$ and $\mathcal{M}$ is $\mathcal{C}^2$ embedded, $\bar{f}$ is $\mathcal{C}^2$ on $\mathcal{M}$.

By Assumption 3, $x^{\star} := x^{\star}(\theta)$ is a (local) minimizer of $\bar{f}$ on $\mathcal{M}$, so the Riemannian gradient vanishes: $\mathrm{grad}_{\mathcal{M}}\bar{f}(x^{\star}) = 0$. Moreover, the Riemannian Hessian of $\bar{f}$ at $x^{\star}$ is positive definite on $\mathcal{T}_{x^{\star}}\mathcal{M}$. Define its minimal eigenvalue

$$m \;:=\; \min_{\substack{v\in\mathcal{T}_{x^{\star}}\mathcal{M}\\ \|v\|=1}} \big\langle v,\, \mathrm{Hess}_{\mathcal{M}}\bar{f}(x^{\star})[v]\big\rangle \;>\; 0.$$

Since $\mathrm{Hess}_{\mathcal{M}}\bar{f}$ is continuous, there exists a neighborhood $\mathcal{U} \subset \mathcal{M}$ of $x^{\star}$ such that for all $y \in \mathcal{U}$ and all $w \in \mathcal{T}_y\mathcal{M}$,

$$\big\langle w,\, \mathrm{Hess}_{\mathcal{M}}\bar{f}(y)[w]\big\rangle \;\ge\; \tfrac{m}{2}\,\|w\|^2. \tag{35}$$

Let $r_0 > 0$ be such that the geodesic ball $\mathbb{B}_{\mathcal{M}}(x^{\star};r_0)$ is contained in $\mathcal{U}$.

Fix any $x \in \mathcal{M}$ with $d_{\mathcal{M}}(x,x^{\star}) \le r_0$. Since $\mathcal{M}$ is compact (and hence complete), the Hopf–Rinow theorem guarantees the existence of a minimizing geodesic $\gamma : [0,1] \to \mathcal{M}$ from $x^{\star}$ to x. See, e.g., [Lee, 2006, Hopf–Rinow Theorem]. Parameterize γ at constant speed so that $\|\dot\gamma(t)\| = d_{\mathcal{M}}(x,x^{\star})$ for all $t \in [0,1]$. Because γ is minimizing, $\gamma([0,1]) \subset \mathbb{B}_{\mathcal{M}}(x^{\star};r_0) \subset \mathcal{U}$.

Define $\phi(t) := \bar{f}(\gamma(t))$. Then $\phi'(t) = \langle \mathrm{grad}_{\mathcal{M}}\bar{f}(\gamma(t)), \dot\gamma(t)\rangle$ and in particular $\phi'(0) = 0$. Differentiating once more and using that γ is a geodesic yields the standard identity

$$\phi''(t) \;=\; \big\langle \dot\gamma(t),\, \mathrm{Hess}_{\mathcal{M}}\bar{f}(\gamma(t))[\dot\gamma(t)]\big\rangle.$$

Therefore, by (35) and constant speed,

$$\phi''(t) \;\ge\; \tfrac{m}{2}\,\|\dot\gamma(t)\|^2 \;=\; \tfrac{m}{2}\,d_{\mathcal{M}}(x,x^{\star})^2 \qquad \forall t \in [0,1].$$

Using $\phi'(0) = 0$ and the integral form of Taylor's theorem,

$$\phi(1) - \phi(0) = \int_0^1 (1-t)\,\phi''(t)\,\mathrm{d}t \ge \int_0^1 (1-t)\,\tfrac{m}{2}\,d_{\mathcal{M}}(x,x^{\star})^2\,\mathrm{d}t = \tfrac{m}{4}\,d_{\mathcal{M}}(x,x^{\star})^2.$$

Recalling $\phi(1) = \bar{f}(x)$ and $\phi(0) = \bar{f}(x^{\star})$ gives (34) with $c_{\mathrm{hg}} := m/4$. ∎

Step 7b: A uniform version over Θ. Under the global Assumption 3, the quadratic growth constants can be chosen uniformly over Θ.

Lemma F.8 (Uniform quadratic growth on $\mathcal{S}(\theta)$ over Θ)**.** *Assume the setting of Section 5, and let $x^\star : \Theta \to \mathbb{R}^d$ be the global $\mathcal{C}^1$ selection obtained by patching the local branches from Theorem 3.5 under Assumption 3. Then there exist constants $c_{\mathrm{hg}} > 0$ and $r_0 > 0$ such that for all $\theta \in \Theta$ and all $x \in \mathcal{S}(\theta)$ with $d_{\mathcal{S}(\theta)}(x, x^\star(\theta)) \le r_0$,*

$$f(\theta, x) - f(\theta, x^\star(\theta)) \;\ge\; c_{\mathrm{hg}}\, d_{\mathcal{S}(\theta)}(x, x^\star(\theta))^2.$$

Proof. For each $\theta \in \Theta$, let $\mathcal{M}_\theta := \mathcal{S}(\theta)$ and define the restriction $\bar{f}_\theta : \mathcal{M}_\theta \to \mathbb{R}$ by $\bar{f}_\theta(x) := f(\theta, x)$. By the global consequence of Theorem 3.5 under Assumption 3, for every $\theta \in \Theta$ the point $x^\star(\theta)$ is the (unique) minimizer of $\bar{f}_\theta$ on $\mathcal{M}_\theta$, and the Riemannian Hessian of $\bar{f}_\theta$ at $x^\star(\theta)$ is positive definite on $\mathcal{T}_{x^\star(\theta)}\mathcal{M}_\theta$. Define the minimal eigenvalue

$$m(\theta) \;:=\; \min_{\substack{v \in \mathcal{T}_{x^\star(\theta)}\mathcal{M}_\theta \\ \|v\|=1}} \big\langle v,\, \mathrm{Hess}_{\mathcal{M}_\theta} \bar{f}_\theta(x^\star(\theta))[v] \big\rangle \;>\; 0.$$

Since $x^\star(\cdot)$ is continuous and the Riemannian Hessian depends continuously on (θ, x), the map $\theta \mapsto m(\theta)$ is continuous on the compact set Θ. Therefore $m_{\min} := \min_{\theta \in \Theta} m(\theta) > 0$.

By continuity of the Riemannian Hessian and compactness of Θ, there exists $r_0 > 0$ such that for all $\theta \in \Theta$, all $y \in \mathcal{M}_\theta$ with $d_{\mathcal{M}_\theta}(y, x^\star(\theta)) \le r_0$, and all $w \in \mathcal{T}_y\mathcal{M}_\theta$,

$$\big\langle w,\, \mathrm{Hess}_{\mathcal{M}_\theta} \bar{f}_\theta(y)[w] \big\rangle \;\ge\; \tfrac{m_{\min}}{2}\, \|w\|^2.$$

Fix any $\theta \in \Theta$ and $x \in \mathcal{M}_\theta$ with $d_{\mathcal{M}_\theta}(x, x^\star(\theta)) \le r_0$. Since $\mathcal{M}_\theta$ is compact, the Hopf–Rinow theorem yields a minimizing geodesic $\gamma : [0, 1] \to \mathcal{M}_\theta$ from $x^\star(\theta)$ to x. Parameterize γ at constant speed so that $\|\dot\gamma(t)\| = d_{\mathcal{M}_\theta}(x, x^\star(\theta))$ for all $t \in [0, 1]$. As in the proof of Lemma F.7, define $\phi(t) := \bar{f}_\theta(\gamma(t))$. Then $\phi'(0) = 0$ and

$$\phi''(t) \;=\; \big\langle \dot\gamma(t),\, \mathrm{Hess}_{\mathcal{M}_\theta} \bar{f}_\theta(\gamma(t))[\dot\gamma(t)] \big\rangle \;\ge\; \tfrac{m_{\min}}{2}\, d_{\mathcal{M}_\theta}(x, x^\star(\theta))^2,$$

so the same integral argument yields

$$f(\theta, x) - f(\theta, x^\star(\theta)) \;=\; \phi(1) - \phi(0) \;\ge\; \tfrac{m_{\min}}{4}\, d_{\mathcal{M}_\theta}(x, x^\star(\theta))^2.$$

The claim follows with $c_{\mathrm{hg}} := m_{\min}/4$. ∎

Step 8: controlling $\mathbb{E}\|\bar{x} - x^\star(\theta)\|^2$ without assuming locality. The quadratic growth bound from Lemma F.8 only applies within a geodesic neighborhood of radius r_0 around $x^\star(\theta)$. To remove the explicit restriction $d_{\mathcal{S}(\theta)}(\bar{x}, x^\star(\theta)) \le r_0$, we use a value gap away from this neighborhood.

Lemma F.9 (Bounding $\mathbb{E}\|\bar{x} - x^\star(\theta)\|^2$ without a locality condition)**.** *Let (c_{hg}, r_0) be the constants from Lemma F.8 and define*

$$\mathcal{S}_{r_0}(\theta) := \{x \in \mathcal{S}(\theta) : d_{\mathcal{S}(\theta)}(x, x^\star(\theta)) \ge r_0\}, \qquad \Delta_{r_0} := \inf_{\theta \in \Theta} \min_{x \in \mathcal{S}_{r_0}(\theta)} \big(f(\theta, x) - F(\theta)\big),$$

and assume $\Delta_{r_0} > 0$. Then for any $\theta \in \Theta$,

$$\mathbb{E}\|\bar{x} - x^\star(\theta)\|^2 \;\le\; \left(\frac{1}{c_{\mathrm{hg}}} + \frac{4\mathsf{D}^2}{\Delta_{r_0}}\right) \mathbb{E}\big[f(\theta, \bar{x}) - F(\theta)\big].$$

Proof. Fix $\theta \in \Theta$ and define the suboptimality gap $\Delta(\bar{x}) := f(\theta, \bar{x}) - F(\theta) \ge 0$. Consider the event $\mathcal{A} := \{\Delta(\bar{x}) < \Delta_{r_0}\}$. By definition of Δ_{r_0}, on $\mathcal{A}$ we must have $d_{\mathcal{S}(\theta)}(\bar{x}, x^\star(\theta)) < r_0$, so Lemma F.8 gives

$$d_{\mathcal{S}(\theta)}(\bar{x}, x^\star(\theta)) \;\le\; \left(\frac{\Delta(\bar{x})}{c_{\mathrm{hg}}}\right)^{1/2}.$$

Since the ambient Euclidean distance is bounded by the geodesic distance, this implies $\|\bar{x} - x^\star(\theta)\|^2 \le \Delta(\bar{x})/c_{\mathrm{hg}}$ on $\mathcal{A}$. On $\mathcal{A}^c$, use $\|\bar{x} - x^\star(\theta)\|^2 \le 4\mathsf{D}^2$ (since $\mathcal{S}(\theta) \subseteq \mathbb{B}_d(0; \mathsf{D})$). Therefore,

$$\mathbb{E}\|\bar{x} - x^\star(\theta)\|^2 \;\le\; \frac{1}{c_{\mathrm{hg}}} \mathbb{E}\big[\Delta(\bar{x})\big] \;+\; 4\mathsf{D}^2\, \mathbb{P}(\mathcal{A}^c).$$

Finally, $\mathbb{P}(\mathcal{A}^c) = \mathbb{P}(\Delta(\bar{x}) \ge \Delta_{r_0}) \le \mathbb{E}[\Delta(\bar{x})]/\Delta_{r_0}$ by Markov, yielding the claim. ∎

Step 9: Final expected squared distance bound for the selected point. We now assemble the previous steps into a single bound on the squared (Euclidean) selection error $\mathbb{E}\|\tilde{x} - x^\star(\theta)\|^2$.

Corollary F.10 (Expected squared selection error)**.** *Fix $\theta \in \Theta$ and let $x^\star(\theta)$ denote the unique minimizer of $f(\theta, \cdot)$ over $\mathcal{S}(\theta)$. Assume in addition that μ_θ^λ satisfies a Poincaré inequality with constant at most C_{PI}. Let $X_1, \dots, X_M$ be independent candidates with marginal laws $\nu_1, \dots, \nu_M$, and define*

$$\varepsilon_{\mathrm{R}}^2 := \max_{1 \le i \le M} R_2(\nu_i \| \mu_\theta^\lambda).$$

Let $\tilde{x} \in \arg\min_{1 \le i \le M} f(\theta, X_i)$ be the hard-selection output. Then

$$\begin{aligned}\mathbb{E}\|\tilde{x} - x^\star(\theta)\|^2 \;&\le 2C_{\mathrm{tube},2}\, e^{\varepsilon_{\mathrm{R}}^2/2}\, \lambda \log(1+M) \\ &\quad +\; 2\left(\frac{1}{c_{\mathrm{hg}}} + \frac{4\mathsf{D}^2}{\Delta_{r_0}}\right)\left(2L_{f,1}\, C_{\mathrm{tube}} \sqrt{\lambda \log(1+M)} +\; C_1\, M^{-1/\mathtt{k}} \;+\; 4L_{f,1}\sqrt{M\, C_{\mathrm{PI}}}\left(e^{\varepsilon_{\mathrm{R}}^2} - 1\right)^{1/2}\right).\end{aligned} \tag{36}$$

Proof. By the Poincaré assumption and (38), each candidate law satisfies

$$\mathsf{W}_2(\nu_i, \mu_\theta^\lambda) \le 2\, C_{\mathrm{PI}}^{1/2}\left(e^{\varepsilon_{\mathrm{R}}^2} - 1\right)^{1/2}.$$

By the Euclidean triangle inequality and $(a+b)^2 \le 2a^2 + 2b^2$,

$$\|\tilde{x} - x^\star(\theta)\|^2 \le 2\|\tilde{x} - \bar{x}\|^2 + 2\|\bar{x} - x^\star(\theta)\|^2 = 2\,\mathrm{dist}(\tilde{x}, \mathcal{S}(\theta))^2 + 2\|\bar{x} - x^\star(\theta)\|^2.$$

Taking expectations and applying (33) from Lemma F.5 gives

$$\mathbb{E}\|\tilde{x} - x^\star(\theta)\|^2 \le 2C_{\mathrm{tube},2}\, e^{\varepsilon_{\mathrm{R}}^2/2}\, \lambda \log(1+M) + 2\,\mathbb{E}\|\bar{x} - x^\star(\theta)\|^2.$$

Then Lemma F.9 yields

$$\mathbb{E}\|\bar{x} - x^\star(\theta)\|^2 \le \left(\frac{1}{c_{\mathrm{hg}}} + \frac{4\mathsf{D}^2}{\Delta_{r_0}}\right)\mathbb{E}\big[f(\theta, \bar{x}) - F(\theta)\big].$$

Finally, (21) from Lemma F.3 bounds the remaining value gap by

$$\mathbb{E}\big[f(\theta, \bar{x}) - F(\theta)\big] \le 2L_{f,1}\, C_{\mathrm{tube}} \sqrt{\lambda \log(1+M)} + C_1\, M^{-1/\mathtt{k}} + 4L_{f,1}\sqrt{M\, C_{\mathrm{PI}}}\left(e^{\varepsilon_{\mathrm{R}}^2} - 1\right)^{1/2},$$

which gives (36). ∎

Step 10: Quantitative LMC mixing control via Rényi divergence. We restate a quantitative order-2 Rényi mixing bound for LMC from Masiha et al. [2025, Prop. 5.6] and record its implication for Wasserstein error under a Poincaré inequality.

Proposition F.11 (LMC convergence in order-2 Rényi divergence [Masiha et al., 2025, Prop. 5.6])**.** *Consider the unadjusted LMC iteration*

$$X^{k+1} = X^k - h\nabla G(X^k) + \sqrt{2h}\,\xi^k, \qquad \xi^k \sim \mathcal{N}(0, I_d),$$

targeting $\pi(\mathrm{d}x) \propto e^{-G(x)}\,\mathrm{d}x$. Assume ∇G is $L_{G,2}$-Lipschitz and that π satisfies a Poincaré inequality with constant C_{PI}. Let $\widehat{\mu}_n$ be the law of X^n. Then for any sufficiently small $\varepsilon \in (0,1)$ and an appropriate stepsize choice (as required by the LMC theory), one can ensure $R_2(\widehat{\mu}_n \| \pi) \le \varepsilon^2$ after

$$n \;=\; \Theta\Big(C_{\mathrm{PI}}^2 L_{G,2}^2\, d\, \varepsilon^{-2}\Big(R_3(\widehat{\mu}_0 \| \pi)^2 + \log^2(1/\varepsilon)\Big)\Big) \tag{37}$$

iterations, where $R_q(\cdot\|\cdot)$ denotes the order-q Rényi divergence. Moreover, using the inequality relating W_2 and R_2 under a Poincaré inequality, we obtain

$$\mathsf{W}_2(\widehat{\mu}_n, \pi) \;\le\; 2\, C_{\mathrm{PI}}^{1/2}\left(e^{R_2(\widehat{\mu}_n \| \pi)} - 1\right)^{1/2} \;\le\; 2\, C_{\mathrm{PI}}^{1/2}\, C^{1/2}\, \varepsilon, \tag{38}$$

where one may take the numerical constant $C := 2(e^{1/2} - 1)$.

Corollary F.12 (Specialization to Gibbs sampling (Rényi accuracy implies Wasserstein control))**.** *Apply Proposition F.11 with $G(x) = g(\theta, x)/\lambda$ and $\pi = \mu_\theta^\lambda$. Then $L_{G,2} = L_{g,2}/\lambda$ and the ULA update*

$$X^{k+1} = X^k - h\nabla_x g(\theta, X^k) + \sqrt{2\lambda h}\,\xi^k$$

corresponds to the LMC scheme with stepsize $h' = \lambda h$ for G. In particular, for any target Rényi tolerance $\varepsilon_{\mathrm{R}} \in (0, 1)$, running the chain for

$$n \;=\; \widetilde{O}\big(d\, C_{\mathrm{PI}}^2\, \lambda^{-2}\, \varepsilon_{\mathrm{R}}^{-2}\big) \tag{39}$$

iterations (where $\widetilde{O}(\cdot)$ hides $\log(1/\varepsilon_{\mathrm{R}})$ and the dependence on $R_3(\widehat{\mu}_0\|\mu_\theta^\lambda)$) ensures

$$R_2(\widehat{\mu}_n\|\mu_\theta^\lambda) \;\le\; \varepsilon_{\mathrm{R}}^2,$$

and therefore

$$\mathsf{W}_2(\widehat{\mu}_n, \mu_\theta^\lambda) \;\le\; 2\, C_{\mathrm{PI}}^{1/2}\, \big(e^{\varepsilon_{\mathrm{R}}^2} - 1\big)^{1/2}, \qquad \textit{and hence} \qquad \mathsf{W}_1(\widehat{\mu}_n, \mu_\theta^\lambda) \;\le\; 2\, C_{\mathrm{PI}}^{1/2}\, \big(e^{\varepsilon_{\mathrm{R}}^2} - 1\big)^{1/2}.$$

Plug-in form. With Proposition F.11 and Corollary F.12, one can run the sampler until the candidate laws satisfy $R_2(\nu_i\|\mu_\theta^\lambda) \le \varepsilon_{\mathrm{R}}^2$ and hence $\mathsf{W}_2(\nu_i, \mu_\theta^\lambda) \lesssim C_{\mathrm{PI}}^{1/2}(e^{\varepsilon_{\mathrm{R}}^2} - 1)^{1/2}$. This yields an explicit end-to-end bound on $\mathbb{E}\|\tilde{x} - x^\star(\theta)\|^2$ in terms of: (i) the number of candidates M, (ii) the Gibbs temperature λ, (iii) the achieved Rényi tolerance ε_{R} through both the tube term and the additive $\sqrt{M\, C_{\mathrm{PI}}}(e^{\varepsilon_{\mathrm{R}}^2} - 1)^{1/2}$ contribution, and (iv) the on-manifold approximation term $M^{-1/\mathtt{k}}$.

F.5 Bounded curvature from analytic regularity of g

This subsection proves the second fundamental form bound used in the volume comparison arguments of [Masiha et al., 2025, e.g., Lem. 4.1] from the analytic assumption on g stated in Assumption 2. Note that Assumption 2 is stronger than the baseline $\mathcal{C}^2/L_{g,2}$-smoothness in Assumption 4: it also controls how $\nabla_{xx}^2 g(\theta, \cdot)$ varies with x (via a uniform bound on $\nabla_{xxx}^3 g$ in a tube around $\mathcal{S}(\theta)$).

Proposition F.13 (Bounded second fundamental form under Assumption 2)**.** *Let Assumption 2 hold and fix any $\theta \in \Theta$. Then the second fundamental form II of $\mathcal{S}(\theta)$ satisfies*

$$\sup_{x \in \mathcal{S}(\theta)} \|\mathrm{II}_x\|_{\mathrm{op}} \;\le\; \frac{L_{g,3}}{c}.$$

In particular, since $(c, \rho, L_{g,3})$ in Assumption 2 are uniform over $\theta \in \Theta$, we obtain a uniform curvature bound over the family $\{\mathcal{S}(\theta)\}_{\theta \in \Theta}$.

Proof. Fix θ and abbreviate $S := \mathcal{S}(\theta)$. Let $H(x) := \nabla_{xx}^2 g(\theta, x)$. Fix $x \in S$ and unit tangent vectors $u, v \in \mathcal{T}_x^\theta$. Choose a $\mathcal{C}^2$ curve $\gamma : (-\varepsilon, \varepsilon) \to S$ with $\gamma(0) = x$ and $\dot\gamma(0) = v$, and choose a $\mathcal{C}^1$ tangent vector field $U(\cdot)$ along γ with $U(0) = u$ and $U(t) \in \mathcal{T}_{\gamma(t)}^\theta$.

Since $U(t) \in \ker H(\gamma(t)) = \mathcal{T}_{\gamma(t)}^\theta$ for all t, we have $H(\gamma(t))\, U(t) = 0$. Differentiating at $t = 0$ gives

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\Big(H(\gamma(t))\, U(t)\Big)\Big|_{t=0} = \big(D_v H(x)\big)\, u + H(x)\, \dot U(0),$$

where $D_v H(x)$ denotes the directional derivative of the Hessian in direction v. Let $P_{\mathcal{N}_x^\theta}$ denote the orthogonal projection onto the normal space $\mathcal{N}_x^\theta$. Since $H(x)$ is symmetric and $\ker H(x) = \mathcal{T}_x^\theta$, we have $\mathrm{range}(H(x)) = (\mathcal{T}_x^\theta)^\perp = \mathcal{N}_x^\theta$ and hence $P_{\mathcal{N}_x^\theta} H(x) = H(x) P_{\mathcal{N}_x^\theta}$. Applying $P_{\mathcal{N}_x^\theta}$ to the display above yields

$$H(x)\, P_{\mathcal{N}_x^\theta} \dot U(0) \;=\; -\, P_{\mathcal{N}_x^\theta}\big(D_v H(x)\big)\, u.$$

By the definition of the second fundamental form of an embedded submanifold in Euclidean space, $\mathrm{II}_x(v, u) = P_{\mathcal{N}_x^\theta} \dot U(0)$. Restricting $H(x)$ to $\mathcal{N}_x^\theta$ and using the uniform normal spectral gap gives

$$\|\mathrm{II}_x(v, u)\| \le \big\|\big(H(x)|_{\mathcal{N}_x^\theta}\big)^{-1}\big\|\, \big\|\big(D_v H(x)\big)\, u\big\| \le \frac{1}{c}\, \big\|\big(D_v H(x)\big)\, u\big\|.$$

Finally, by the bounded third derivative assumption, $\|(D_v H(x))\, u\| \le \|\nabla_{xxx}^3 g(\theta, x)\|_{\mathrm{op}} \|v\|\, \|u\| \le L_{g,3}$. Therefore $\|\mathrm{II}_x(v, u)\| \le L_{g,3}/c$ for all unit $u, v \in \mathcal{T}_x^\theta$, and taking the supremum over $x \in S$ yields the claim. ∎

G Experimental details

G.1 Common protocol

Metrics. For data hyper-cleaning we report *accuracy* (% correct). For imbalanced loss tuning we report *balanced accuracy* (macro recall), i.e., the average of per-class recall, reported as a percentage.

Confidence intervals. All tables/plots report mean $\pm$ 95% confidence intervals across independent seeds, using a two-sided t-interval.

Time budget. Time-budget plots measure wall-clock elapsed time of each algorithm and compare methods at fixed per-algorithm budgets.

G.2 Data hyper-cleaning (MNIST)

Data and corruption. We draw $15{,}100$ examples from the MNIST training split and randomly partition them into $n_{\mathrm{tr}} = 5{,}000$ (lower-level train), $n_{\mathrm{val}} = 100$ (clean validation), and $n_{\mathrm{test}} = 10{,}000$ (test). We corrupt a fraction $\rho \in \{0.4, 0.6, 0.8\}$ of the *training* labels uniformly at random; validation and test labels remain clean.

Model and optimization. All methods use the same 2-layer MLP (hidden size 300) and AdamW in both the lower and upper levels.

Budgets and runs. For the time-budget results in Table 1 and fig. 2, we use a fixed budget per algorithm on a single NVIDIA H100 GPU (60s for $\rho \in \{0.4, 0.6\}$; 120s for $\rho = 0.8$) and report results over 5 independent runs per ρ.

G.3 Parametric loss tuning for imbalanced data

Data and imbalance. We use MNIST and form an imbalanced lower-level training set by keeping a training example of class y with probability $p(y) = b^y$ (with $b = 0.3$). The upper level uses a small class-balanced validation set (100 examples in our main setting), creating a train–validation mismatch: the lower level sees an imbalanced class distribution while validation/test are class-balanced.

Model and optimization. The follower is a CNN classifier and the upper-level variables are class-wise logit shift/scale parameters $(\delta, \gamma) \in \mathbb{R}^{10} \times \mathbb{R}^{10}$ (implemented via bounded reparameterizations). All methods use AdamW updates for both levels.

Budgets. The time-budget results in Table 2 and fig. 3 use a fixed wall-clock budget per algorithm on a single NVIDIA H100 GPU (120s per algorithm) and 3 independent runs.