Select-then-differentiate: Solving Bilevel Optimization with Manifold Lower-level Solution Sets

Saeed Masiha[1] **Zebang Shen**[2] **Negar Kiyavash**[1] **Niao He**[2]
[1]EPFL, Switzerland [2]ETH Zurich, Switzerland
{mohammadsaeed.masiha,negar.kiyavash}@epfl.ch
{zebang.shen,niao.he}@inf.ethz.ch

Abstract

We study optimistic bilevel optimization when the lower-level problem has a non-isolated manifold of minimizers. In this setting, the hyper-objective may be non-differentiable because the upper-level criterion must choose among multiple lower-level solutions. Under a Polyak–Łojasiewicz–type (PŁ°) condition, we show that differentiability does not require the lower-level solution set to be a singleton: uniqueness of the optimistic selection is sufficient. This yields an explicit pseudoinverse-based hyper-gradient formula extending the classical singleton-minimizer result. We further characterize the regularity of the hyper-objective: non-degeneracy of the selected minimizer along the solution manifold yields local smoothness, while failure of uniqueness can create many non-differentiable points and failure of non-degeneracy can destroy all positive Hölder regularity of the hyper-gradient. Motivated by this theory, we propose HG-MS, a select-then-differentiate method combining explicit optimistic selection with efficient pseudoinverse-based hyper-gradient computation. For the analyzed method, we prove convergence to stationarity with a conservative oracle-complexity bound whose accuracy exponent depends on the minimizer manifold's intrinsic dimension. Empirically, we test a practical variant of HG-MS for matched-budget LLM source reweighting. This variant preserves the select-then-differentiate principle and obtains the best GSM8K/MATH scores across the tested backbones, along with competitive or best MT-Bench instruction-following results.

1 Introduction

Many modern machine learning tasks involve an upper objective that evaluates the outcome of an inner training procedure. Such tasks can be formulated as bilevel optimization. Canonical examples include hyperparameter optimization and meta-learning [Feurer and Hutter, 2019, Liu et al., 2021a, Bertinetto et al., 2019, Finn et al., 2017], differentiable neural architecture search [Liu et al., 2019], bilevel formulations in reinforcement learning [Hong et al., 2023], and more recently, bilevel formulations for LLM data reweighting, safe fine-tuning, and preference alignment [Pan et al., 2025, Shen et al., 2025b, Jian et al., 2025]. The progress in aforementioned applications has been driven by the ability to compute *hyper-gradients*, which allows us to directly optimize the hyper-objective.

A standard assumption underlying much of the bilevel optimization literature is that the lower-level problem has a *unique* minimizer. In this singleton regime, the lower-level solution map is single-valued, and implicit differentiation yields a tractable hyper-gradient formula [Pedregosa, 2016, Franceschi et al., 2017, Lorraine et al., 2020, Huang, 2024, Liu et al., 2022a, Kwon et al., 2024]. However, this assumption is often violated in the nonconvex and overparameterized settings that dominate modern applications. Neural-network training objectives admit many global minimizers, which often are not isolated points but rather come from connected sets with manifold structure [Draxler et al., 2018, Garipov et al., 2018, Nguyen, 2019]. In such regimes, the lower-level solution

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

map is set-valued, and the hyper-objective critically depends on lower-level *minima selection*. Exact examples also arise in simple full-dataset autoencoder objectives: a tied one-filter linear convolutional model on whitened patches has a sphere of global minimizers, while a regularized linear autoencoder has a Stiefel family under suitable spectral conditions [Masci et al., 2011, Kunin et al., 2019]; see Examples B.4 and B.5 in Appendix B.6. For instance, in recent LLM data reweighting formulations [Pan et al., 2025, Shen et al., 2025b], the lower-level objective is typically the training loss, while the upper-level objective is evaluated on a validation set. Distinct lower-level minimizers can therefore attain nearly identical training loss but result in very different validation performance. This makes *optimistic bilevel optimization* a natural formulation: among all lower-level minimizers, select the one that is best for the upper level.

In this paper, we focus on the optimistic bilevel optimization problem

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

with compact convex $\Theta \subseteq \mathbb{R}^m$. The set $\mathcal{S}(\theta)$ is implicit and generally unavailable explicitly. Our method assumes no parameterization, enumeration, or projection oracle for it; the ULA candidate generator in Section 4 uses only evaluations of $\nabla_x g(\theta, \cdot)$. We refer to the inner optimization over $\mathcal{S}(\theta)$ as *minima selection* and define $\mathtt{O}(\theta) := \arg\min_{x\in\mathcal{S}(\theta)} f(\theta, x)$. In particular, when $\mathtt{O}(\theta)$ is a singleton, we denote the unique optimistic selection as $x^\star(\theta)$. In general, without additional structure, F need not be continuous, let alone differentiable [Dontchev and Rockafellar, 2009, Arbel and Mairal, 2022b].

Why optimistic selection requires a separate analysis. Classical implicit differentiation is simplest when a nonsingular lower Hessian determines a unique response [Pedregosa, 2016, Franceschi et al., 2017, Lorraine et al., 2020]. Nonunique-response analyses also study prescribed dynamics [Arbel and Mairal, 2022b], convex lower-level problems [Xiao et al., 2023, Shen et al., 2025a], or penalty-uniform regularity [Kwon et al., 2024, Chen et al., 2024]. Their selected objects and assumptions differ from our exact optimistic value on a nonconvex minimizer manifold; Appendix C gives the comparison.

Our starting point: manifold-structured lower-level minimizers. Motivated by the loss landscapes of overparameterized deep networks [Venturi et al., 2018, Liu et al., 2022b], we assume the lower-level loss g satisfies the $\mathtt{PŁ}^\circ$ condition of [Gong et al., 2024] with respect to the lower-level variable x. Under our lower-level $\mathtt{PŁ}^\circ$ assumption (Assumption 1), we obtain a structured non-singleton regime: the lower-level solution set $S(\theta)$ is an *embedded submanifold* of $\mathbb{R}^d$, and existing results further imply that the optimistic hyper-objective F is Lipschitz continuous [Masiha et al., 2025]. But Lipschitzness alone does not justify hyper-gradient methods [Chen et al., 2024]. This leads to the main technical question of the paper: When lower-level minimizers form a manifold, what additional conditions make the optimistic hyper-objective differentiable or smooth, and hence a good candidate for hyper-gradient optimization?

Main insight: uniqueness of the optimistic selection suffices. The main source of nonsmoothness is the instability of the minima selection. When $\mathtt{O}(\theta)$ contains multiple points, small perturbations of θ can change which minimizer is selected, making F non-differentiable even if the map $\theta \mapsto \mathcal{S}(\theta)$ varies smoothly. Thus, uniqueness of the optimistic selection is the key regularity condition. Under the same $\mathtt{PŁ}^\circ$ assumption (Assumption 1), we show that if $\mathtt{O}(\theta) = \{x^\star(\theta)\}$ throughout a θ-neighborhood, then F is continuously differentiable, even when $\mathcal{S}(\theta)$ is non-singleton and non-convex. The resulting hyper-gradient admits an explicit pseudoinverse-based formula that strictly generalizes the classical singleton-minimizer formula.

Theorem (Regularity; Informal)**.** Under Assumption 1:

- If the optimistic minimizer is unique throughout a θ-neighborhood, then F is continuously differentiable there, and

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

  where $\big[\nabla^2_{xx} g(\theta, x^\star(\theta))\big]^\dagger = \mathbf{N}_\theta \big[\mathbf{N}_\theta^\top \nabla^2_{xx} g(\theta, x^\star(\theta)) \mathbf{N}_\theta\big]^{-1} \mathbf{N}_\theta^\top$, and $\mathbf{N}_\theta$ is an orthonormal basis for the normal space of $\mathcal{S}(\theta)$ at $x^\star(\theta)$; see Remark 3.3.

- If, in addition, $x^\star(\theta)$ is non-degenerate (see Definition 3.4), then the selected branch is $\mathcal{C}^1$ and ∇F is locally Lipschitz.

Our uniqueness condition is *strictly weaker* than the standard singleton lower-level solution assumption. We further show that our conditions are nearly necessary: if $\mathtt{O}(\theta)$ is not a singleton, then F can be non-differentiable at *infinitely many* points in an *arbitrarily small* neighborhood; even under unique selection, if $x^\star$ is degenerate, then ∇F can fail to be α-Hölder continuous for *any* $\alpha > 0$. Together, these results give a sharp regularity picture for optimistic bilevel optimization in the manifold regime.

Algorithmic consequence: select, then differentiate. Our regularity theory suggests a simple, practical algorithmic principle for bilevel optimization in the regime when the lower-level solution set exhibits manifold structure and the minima selection yields a unique minimizer: one should first *select* the optimistic minimizer, and only then *differentiate*. To this end, we propose *Hyper-Gradient with Minima-Selection* (HG-MS); see Algorithm 1. The method proceeds iteratively in two stages. First, it approximates the minima selection using a *Best-of-N* (BoN) strategy: it generates N independent finite-step ULA candidates targeting the Gibbs measure $\mu_\theta^\lambda(\mathrm{d}x) \propto \exp(-g(\theta, x)/\lambda)\,\mathrm{d}x$, and outputs $\hat{x}_N^\lambda(\theta)$, the one achieving the smallest upper-level value $f(\theta, \cdot)$, as a proxy for $x^\star(\theta)$. Note that when $\lambda \to 0$, μ_θ^λ concentrates around $\mathcal{S}(\theta)$, allowing the procedure to explore the lower-level solution manifold. Second, the method approximates the hyper-gradient ∇F by plugging $\hat{x}_N^\lambda(\theta)$ in Equation (2), using ridge–CG, then takes a projected-gradient step.

For convergence analysis, we interpret the analyzed method as an inexact projected-gradient method applied to the hyper-objective F. The hyper-gradient error has three sources: (a) sampling error, including finite temperature and finite-step dynamics, (b) BoN selection error with finite N, and (c) approximation error of the pseudoinverse in (2). In particular, under the assumptions of Lemma F.20, the BoN selector satisfies $\mathbb{E}\big[\|x^\star(\theta) - \hat{x}_N^\lambda(\theta)\|^2\big] = \widetilde{\mathcal{O}}\big(\sqrt{\lambda} + \lambda + N^{-1/\mathtt{k}} + \sqrt{NC_{\mathrm{PI}}}\,\varepsilon_{\mathrm{R}}\big)$, where $\widetilde{\mathcal{O}}$ suppresses logarithmic factors in $1+N$. Here, $\mathtt{k} = \dim \mathcal{S}(\theta)$ denotes the common intrinsic dimension of the lower-level solution manifold, ε_{R} measures finite-step sampling error, and C_{PI} is the Gibbs Poincaré constant. There is an accuracy–efficiency tradeoff: smaller λ and larger N improve the temperature and coverage terms, yet increase the computational cost through slower Gibbs sampling and the need for more candidates. Balancing these errors yields our conservative oracle-complexity guarantee for finding an ϵ-stationary point of F; see Theorem 5.3. Intrinsic dimension controls the coverage exponent, while ambient dimension and sampling factors remain in the total cost.

Empirical evidence. We test the ULA/ridge–CG components on a synthetic problem with exact ground truth. Our main practical testbed is LLM source reweighting for mathematical reasoning, instruction following, and two-source GPT-2 data composition; two multi-seed MNIST benchmarks complement it. In the real-world LoRA benchmarks, one-step SGD branches, differentiation through the selected step, and Adam replace the analyzed components. Section 6 distinguishes these implementations and their scope.

Contributions. In summary, we make three contributions:

- **Differentiability of the hyper-objective under manifold-structured lower-level minimizers.** Under our structural assumptions, unique optimistic selection throughout a parameter neighborhood yields a $\mathcal{C}^1$ value and a pseudoinverse hyper-gradient; intrinsic non-degeneracy additionally gives a $\mathcal{C}^1$ selected branch and locally Lipschitz hyper-gradient.
- **A hyper-gradient method with BoN minima selection and oracle complexity guarantee.** HG-MS combines finite-step ULA candidates, BoN selection, and ridge–CG hyper-gradients. We prove stationarity with a conservative oracle bound whose selection exponent depends on intrinsic manifold dimension.
- **Empirical support for the select-then-differentiate principle.** Practical HG-MS achieves the best matched-budget GSM8K/MATH scores across the four tested math backbones, competitive or best MT-Bench scores, and lower two-source GPT-2 test loss than all baselines. Multi-seed MNIST experiments support gains in optimization and test performance. The synthetic study isolates selection and hyper-gradient mechanisms and compares analyzed HG-MS with four baselines.

Notations For any nonempty closed convex set $C \subset \mathbb{R}^m$, we write $\mathrm{Proj}_C(u) := \arg\min_{v \in C} \|v - u\|$. 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 $\mathcal{N}_x^\theta := \mathcal{N}_x\mathcal{S}(\theta)$. For probability measures ν and μ, we write $R_2(\nu\|\mu)$ for the order-2 Rényi divergence. Additional notations are collected in Section B.1.

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 local Polyak–Łojasiewicz (PŁ) inequalities, it restricts other critical points and local minimizers. With coercivity and $d \geq 2$, it gives connected global minimizers [Masiha et al., 2025]. This supports efficient sampling-based approximations of the global optimal set[1].

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

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

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 $\mathcal{V}$.

Global PŁ excludes suboptimal stationary points; PŁ$^\circ$ instead combines local PŁ near compact minimizers with the global requirements (2)–(3). The sphere quartic $\bar{g}(w) = \frac{1}{4}(\|w\|^2 - 1)^2$ is locally PŁ near its minimizer sphere, since $\|\nabla \bar{g}\|^2 = 4\|w\|^2 \bar{g}$, but not globally PŁ: the stationary origin is a suboptimal strict maximum. Conversely, $h(u, v) = u^2$ is globally PŁ but has an unbounded minimizer set, violating (3). We impose PŁ$^\circ$ uniformly over the parameter θ in the lower-level problem.

Assumption 1 (Lower-level PŁ$^\circ$)**.** Let $d \geq 2$ and let g be jointly $\mathcal{C}^3$ on $\mathcal{U} \times \mathbb{R}^d$, where $\mathcal{U}$ is an open neighborhood of Θ. For every $\theta \in \mathcal{U}$, $g(\theta, \cdot)$ is coercive and satisfies PŁ$^\circ$, with a common local PŁ constant $\mu > 0$ and all local minimizers contained in a common compact set $\mathcal{V}$.

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 $\mathcal{S}(\theta)$)**.** Under Assumption 1, the following hold 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 $\ker H(\theta, x) \;=\; \mathcal{T}^\theta_x$ and $\langle v, H(\theta, x) v \rangle \;\geq\; c\,\|v\|^2$ for all $v \in \mathcal{N}^\theta_x$ for some constant $c > 0$ that is uniform over $\theta \in \Theta$ and $x \in \mathcal{S}(\theta)$ [Rebjock and Boumal, 2025, Cor. 2.17]. Moreover, assuming $g \in \mathcal{C}^3$, there exists an integer $\mathtt{k}$ such that $\dim(\mathcal{S}(\theta)) = \mathtt{k}, \forall \theta \in \Theta$, [Masiha et al., 2025, Lem. 3.5]. Appendix B.3 supplies the parameter-dependent normal-graph argument and uniformity details.

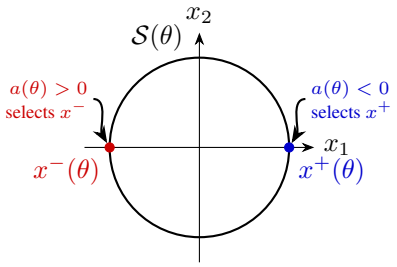
Such manifolds arise when parameters can vary continuously while the loss remains at its minimum. For example, exact data fitting may leave some degrees of freedom, while rotations of a linear autoencoder's latent coordinates can preserve its loss. Under suitable regularity conditions, these freedoms produce smooth families of minimizers. Appendix B.6 gives precise conditions and exact examples. These mechanisms explain the minimizer geometry, but do not by themselves establish the full PŁ$^\circ$ assumption.

Remark 2.4. Under Assumption 1 and the stated joint regularity of f and g, the hyper-objective function is Lipschitz continuous; see [Masiha et al., 2025, Lem. 3.2]. While this implies that F is differentiable almost everywhere, it remains insufficient for our purpose.

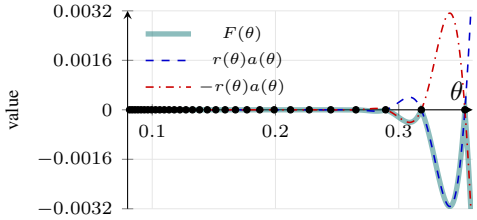
3 Local Differentiability of the Hyper-objective

Under Assumption 1, we give sufficient conditions for continuous differentiability and local smoothness of F, with counterexamples showing what can fail without them. Parameter assumptions and conclusions are local around θ_0; the lower-level assumptions still concern the full landscape. All formal proofs for this section are given in Appendix D. **Non-unique minima-selection can create kinks in the hyper-objective.** Even when $\mathcal{S}(\theta)$ is a $\mathcal{C}^\infty$ connected manifold and varies smoothly

[1] If $\mathcal{S}(\theta)$ has at least two connected components, then the mixing time of Langevin sampling of μ^λ_θ scales exponentially in λ^{-1}; see, e.g., Menz and Schlichting [2014].



(a) Two competing optimistic minimizers.



(b) Two smooth branches and the resulting hyper-objective (kinks at ties $\theta_k = 10/(k\pi)$, a countably infinite set accumulating at 0, marked in black).

Figure 1: **Illustration of Example 3.1.** Left: selection switches between two endpoints of $\mathcal{S}(\theta)$ according to the sign of $a(\theta)$. Right: $F(\theta) = -r(\theta)|a(\theta)|$ is obtained by taking the smaller value between the two branches $\pm r(\theta)a(\theta)$, creating kinks at ties $a(\theta) = 0$ (black dots).

with θ, minimizing $f(\theta, x)$ over $\mathcal{S}(\theta)$ can introduce kinks in F whenever the optimistic minimizer is not unique. The example below shows that, in an *arbitrarily small* neighborhood, F may have *infinitely* many non-differentiable points despite f, g, and $\mathcal{S}(\theta)$ being $\mathcal{C}^\infty$; the singularities arise solely from switching among non-unique minimizers.

Example 3.1. Let $\theta \in \mathbb{R}$ and $x = [x_1, x_2] \in \mathbb{R}^2$ and write $r(\theta) = \sqrt{1+\theta^2}$. Define $a(\theta) := e^{-1/\theta^2}\sin(10/\theta)$ for $\theta \neq 0$ and $a(0) = 0$. Consider the functions $f(\theta, x) := a(\theta)\, x_1 + x_2^2$ and $g(\theta, x) := \left(\|x\|^2 - (r(\theta))^2\right)^2$, both of which belong to $\mathcal{C}^\infty(\mathbb{R} \times \mathbb{R}^2)$. The lower-level solution set is the circle $\mathcal{S}(\theta) = \{x \in \mathbb{R}^2 : \|x\| = r(\theta)\}$. On $\mathcal{S}(\theta)$, $f(\theta, x) = a(\theta)x_1 + r(\theta)^2 - x_1^2$, which is a concave quadratic in x_1. Therefore, for $a(\theta) \neq 0$, the minimizer is unique and given by $x^\star(\theta) = [-\mathrm{sign}(a(\theta))r(\theta), 0]$; 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 $k \in \mathbb{Z} \setminus \{0\}$, F is not differentiable at every θ_k (a countably infinite set accumulating at 0); see Figure 1.

Uniqueness in minima-selection implies differentiability. The preceding example suggests that non-uniqueness of the optimistic minimizer is the main obstruction to differentiability. Our first positive result shows that, under the PŁ$^\circ$ geometry summarized in Proposition 2.3, uniqueness is sufficient to recover $\mathcal{C}^1$ regularity of F and an explicit hyper-gradient formula.

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 such that, for all $\theta \in \mathcal{U}$, $\arg\min_{x\in\mathcal{S}(\theta)} f(\theta, x)$ is a singleton. Then there exists an open neighborhood $\mathcal{U}_0 \subseteq \mathcal{U}$ of θ_0 such that $F \in \mathcal{C}^1(\mathcal{U}_0)$. Moreover, for all $\theta \in \mathcal{U}_0$, the hyper-gradient is given by eq. (2).

Remark 3.3. For compactness, denote $\mathbf{H} := \nabla^2_{xx} g(\theta, x^\star(\theta))$, and let $\mathbf{N}_\theta$ and $\mathbf{T}_\theta$ be orthogonal bases that span $\mathcal{N}^\theta_{x^\star(\theta)}$ and $\mathcal{T}^\theta_{x^\star(\theta)}$, respectively. Under the PŁ$^\circ$ condition, Proposition 2.3 implies $\mathbf{N}_\theta^\top H \mathbf{N}_\theta \succ 0$; Besides, we have $\mathbf{T}_\theta^\top \mathbf{H} \mathbf{T}_\theta = 0$. Consequently, $\mathbf{N}_\theta \left(\mathbf{N}_\theta^\top \mathbf{H} \mathbf{N}_\theta\right)^{-1} \mathbf{N}_\theta^\top$ coincides with the Moore–Penrose pseudoinverse of $\mathbf{H}$.

Uniqueness in minima-selection alone 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 Example D.4 (deferred to Appendix D.2), we show that uniqueness alone is *insufficient* to guarantee this property. This occurs when the upper-level objective restricted to the minimizer manifold has zero intrinsic second-order curvature at the selected point[2]. We therefore impose intrinsic non-degeneracy of the restricted objective $f(\theta, \cdot)|_{\mathcal{S}(\theta)}$ as a sufficient condition for stronger regularity.

Definition 3.4 (Degenerate and non-degenerate optimistic minimizers)**.** Consider the setting of Theorem 3.2. Let $f_\theta := f(\theta, \cdot)|_{\mathcal{S}(\theta)}$. We say that $x^*(\theta)$ is a *non-degenerate optimistic minimizer* for $\theta \in \mathcal{U}_0$ if the Riemannian Hessian[3] of the restricted objective is positive definite at $x^*(\theta)$, i.e.,

[2] One might instead try to require positivity of the ambient Hessian $\nabla^2_{xx} f$ on tangent vectors, but this is not an intrinsic condition (see Appendix D.2).

[3] See Equation (6) in Appendix B.4 for the convention used here.

$\left\langle v, \mathrm{Hess}_{\mathcal{S}(\theta)} \bar{f}_\theta(x^*(\theta))[v] \right\rangle > 0$ for all $v \in \mathcal{T}^\theta_{x^*(\theta)} \setminus \{0\}$. Equivalently, $f(\theta, \cdot)$ has strictly positive intrinsic second-order derivative along every tangent direction of the manifold $\mathcal{S}(\theta)$ at $x^*(\theta)$. If this condition fails, we call $x^*(\theta)$ a *degenerate optimistic minimizer*.

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 a $\mathcal{C}^1$ local branch for $x^\star(\theta)$. 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 proves this branch regularity by first deriving a parameter-dependent local chart for $\mathcal{S}(\theta)$ from Proposition 2.3, and then applying the implicit function theorem to the restricted problem on this chart.

Theorem 3.5 (Smoothness of the hyper-objective under unique non-degenerate optimistic minimizer)**.** Let $f \in \mathcal{C}^2$, $g \in \mathcal{C}^3$ and satisfy Assumption 1. Fix $\theta_0 \in \Theta$ and set $x_0 := x^\star(\theta_0)$. Suppose that x_0 is the unique optimistic minimizer and is 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 Lipschitz.

4 HG-MS: a Select-then-Differentiate Algorithm

The previous section gives a local regularity result. For algorithm design and convergence analysis on Θ, we impose uniqueness and non-degeneracy globally.

Assumption 2 (Unique and non-degenerate optimistic minimizer; **Global**)**.** For every $\theta \in \Theta$, the selected optimistic minimizer $x^\star(\theta)$ is unique and non-degenerate in the sense of Definition 3.4.

This yields global smoothness of the hyper-objective.

Proposition 4.1. Let $f \in \mathcal{C}^2$, $g \in \mathcal{C}^3$, and suppose Assumptions 1 and 2 hold. There exists $L_F < \infty$ such that $\|\nabla F(\theta) - \nabla F(\theta')\| \le L_F \|\theta - \theta'\|$, for all $\theta, \theta' \in \Theta$. In particular, F is L_F-smooth on Θ.

Thus, the natural upper-level update is projected gradient descent, $\theta_{t+1} := \mathrm{Proj}_\Theta\big(\theta_t - \alpha_t \nabla F(\theta_t)\big)$, with ∇F given by eq. (2). Implementing this update requires two primitives: Ⓐ an approximate optimistic selector $x^\star(\theta)$; Ⓑ an approximate pseudoinverse action $\left[\nabla^2_{xx} g(\theta, x^\star(\theta))\right]^\dagger \nabla_x f(\theta, x^\star(\theta))$. Both tasks are challenging because $\mathcal{S}(\theta)$ is an implicit non-convex manifold, and because the selected approximate lower-level solution may lie slightly off $\mathcal{S}(\theta)$, resulting in unstable Hessian pseudoinverse computations.

4.1 Algorithm Design

Algorithm 1 is the analyzed HG-MS method: finite-step ULA candidates, hard `BoN` selection, ridge–CG, and projected gradient. The practical method is given in Algorithm 3 (Appendix G.7).

Ⓐ `BoN` **selection near $\mathcal{S}(\theta)$.** ULA (unadjusted Langevin algorithm) is noisy gradient descent:

$$x^+ = x - h\nabla_x g(\theta, x) + \sqrt{2\lambda h}\,\xi, \qquad \xi \sim \mathcal{N}(0, I_d).$$

The gradient term lowers g, while Gaussian noise enables exploration. The finite-step law approximates the Gibbs target $\mu^\lambda_\theta(\mathrm{d}x) \propto e^{-g(\theta,x)/\lambda}\,\mathrm{d}x$, which concentrates near $\mathcal{S}(\theta)$ at low temperature (Lemma F.9). Thus ULA generates approximate lower-level solutions without computing $\mathcal{S}(\theta)$. From N independent ULA candidates, we select $i^\star \in \arg\min_i f(\theta, X_i)$ and set $\hat{x}^\lambda_N(\theta) = X_{i^\star}$.

In Appendix F.3, we derive a finite-time Rényi bound for the ULA candidate law relative to its Gibbs target (Corollary F.19) by specializing the ULA analysis of Chewi [2026]. The derivation uses Lipschitz gradients, a uniform Poincaré bound, and a λ-scaled Gaussian initialization. This bound transfers Gibbs concentration to ULA candidates, then controls selection error, hyper-gradient MSE, and outer convergence. Another sampler needs an analogous guarantee; arbitrary SGD branches do not inherit it.

Remark 4.2. Approximating $\mathcal{S}(\theta)$ via μ^λ_θ also appears in [Masiha et al., 2025], but their goal is to approximate only the **value** of F. Here we must approximate the selected **solution** itself in order to compute a hyper-gradient, which requires a different analysis.

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

Require: initial $\theta_0 \in \Theta$; stepsizes $\{\alpha_t\}$; temperature λ; number of chains N; ULA steps K and stepsize h; chain starts $\{X_{t,i}^{(0)}\}$; CG tolerances $\{\eta_t\}$; ridge $\gamma \geq 0$; clip radius $R_v > 0$

1: **for** $t = 0, 1, 2, \ldots$ **do**

2: Sample in parallel $X_{t,i} \leftarrow \mathrm{ULA}(\theta_t, g, X_{t,i}^{(0)}, \lambda, K, h)$

3: $i_t^\star \leftarrow \arg\min_{1 \leq i \leq N} f(\theta_t, X_{t,i})$; $\quad \hat{x}_N^\lambda(\theta_t) \leftarrow X_{t,i_t^\star}$

4: Solve $\big(\nabla_{xx}^2 g(\theta_t, \hat{x}_N^\lambda(\theta_t)) + \gamma I\big)\tilde{v}_t = \nabla_x f(\theta_t, \hat{x}_N^\lambda(\theta_t))$ by CG up to residual η_t, with the iteration cap and finite fallback in Remark F.22 (Appendix F.4)

5: Set $v_t \leftarrow \mathrm{Proj}_{\mathbb{B}_d(0;R_v)}(\tilde{v}_t)$ for the stated guarantees

6: $\widehat{h}_t \leftarrow \nabla_\theta f(\theta_t, \hat{x}_N^\lambda(\theta_t)) - \nabla_{\theta x}^2 g(\theta_t, \hat{x}_N^\lambda(\theta_t)) v_t$; $\quad \theta_{t+1} \leftarrow \mathrm{Proj}_\Theta\big(\theta_t - \alpha_t \widehat{h}_t\big)$

® Hyper-gradient evaluation at the selected point. After selecting $\hat{x}_N^\lambda(\theta_t)$, we estimate the hyper-gradient by substituting it for $x^\star(\theta_t)$ in eq. (2). Let $H_t := \nabla_{xx}^2 g(\theta_t, \hat{x}_N^\lambda(\theta_t))$ and $b_t := \nabla_x f(\theta_t, \hat{x}_N^\lambda(\theta_t))$. We approximate $H_t^\dagger b_t$ by solving the ridge-regularized system $(H_t + \gamma I)\tilde{v}_t = b_t$ with conjugate gradients, using only Hessian–vector products, until $\|b_t - (H_t + \gamma I)\tilde{v}_t\| \leq \eta_t$. The ridge parameter $\gamma \geq 0$ stabilizes the solve when $\hat{x}_N^\lambda(\theta_t)$ is close to, but not exactly in $\mathcal{S}(\theta_t)$: tangential zero eigenvalues of the exact Hessian can perturb near-zero eigenvalues off the manifold. For the complexity guarantee, we cap the number of CG iterations and stop early upon detecting nonpositive curvature; see Remark F.22 in Appendix F.4. We then clip the output: $v_t := \mathrm{Proj}_{\mathbb{B}_d(0;R_v)}(\tilde{v}_t)$. The hyper-gradient estimate is then

$$\widehat{h}_t := \nabla_\theta f(\theta_t, \hat{x}_N^\lambda(\theta_t)) - \nabla_{\theta x}^2 g(\theta_t, \hat{x}_N^\lambda(\theta_t)) v_t, \tag{3}$$

followed by a projected step on Θ. Algorithm 1 summarizes our approach.

5 Convergence Analysis of HG-MS

Additional regularity assumptions. In Appendix B.4, we list and explain the additional regularity assumptions needed for the proof of convergence.

Outer-loop guarantee. This section analyzes HG-MS as an inexact projected-gradient method for the hyper-objective F. Recall the hyper-gradient estimate $\widehat{h}_t$ in (3), and define the hyper-gradient error and the projected gradient mapping as $e_t := \widehat{h}_t - \nabla F(\theta_t)$ and $\mathcal{G}_\Theta(\theta, \nabla F(\theta); \alpha) := \alpha^{-1}\big(\theta - \mathrm{Proj}_\Theta(\theta - \alpha \nabla F(\theta))\big)$, respectively. Under Assumptions 1 and 2, Proposition 4.1 provides the required global smoothness of F on Θ. 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]. The objectives are deterministic; all expectations are over the randomness in ULA candidates, hard selection, hyper-gradient estimates, and outer iterates.

Proposition 5.1 (Convergence to near-stationarity with inexact hyper-gradients)**.** 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 $\alpha \leq 1/L_F$ (last line of Algorithm 1), then

$$\forall\, T \geq 1, \quad \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\mathcal{G}_\Theta(\theta_t, \nabla F(\theta_t); \alpha)\|^2\big] \leq \frac{8\,(F(\theta_0) - F_\star)}{\alpha T} + \frac{10}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|e_t\|^2\big]. \tag{4}$$

Compactness of Θ supplies uniform constants for the convergence analysis, but is not intrinsic to the local regularity results. Box constraints and projected updates model bounded hyperparameters [Liu et al., 2020, Kornowski et al., 2024]. Proposition 5.1 also holds on a closed convex, possibly unbounded domain if F is globally smooth and lower bounded. Extending Theorems 5.2 and 5.3 additionally requires uniform problem constants, or a separate bounded-iterate argument ensuring them; compactness cannot simply be dropped.

Hyper-gradient error bound. We now bound the error term $\mathbb{E}\|e_t\|^2$. Write $\tilde{x}_t := \hat{x}_N^\lambda(\theta_t)$ for the hard-selected lower-level point returned by Algorithm 1. The proof has two technical parts, which we defer to Appendix F.2 and Appendix F.3. Lemma F.1 controls the hyper-gradient error for a selected

point near $\mathcal{S}(\theta_t)$, with a clipped off-tube bound. Lemma F.20 controls the Best-of-N selection error. Combining the two ingredients results in the following parameter-level bound.

Theorem 5.2 (Hyper-gradient error bound)**.** Let Assumptions 1 to 4 hold. Suppose $\mathtt{k} \geq 1$, $0 < \lambda \leq \lambda_0$, $\lambda \log(1+N) \leq 1$, $0 < \gamma \leq 1$, and $\varepsilon_{\mathrm{R}} \leq 1$, where λ_0 is the low-temperature threshold from Lemma F.20. Choose the clipping radius R_v so that $R_v \geq (2/\gamma)(L_{f,1} + \eta_t)$. For all t, the error decomposes as

$$\begin{aligned} \mathbb{E}\|e_t\|^2 &\lesssim \gamma^{-2}(\mathcal{E}_{\mathrm{sel}} + \eta_t^2) + \gamma^2 + \mathcal{E}_{\mathrm{off},t}, \\ \mathcal{E}_{\mathrm{sel}} &= \mathcal{O}\Big(\sqrt{\lambda \log(1+N)} + N^{-1/\mathtt{k}} + \sqrt{N C_{\mathrm{PI}}}\,\varepsilon_{\mathrm{R}}\Big). \end{aligned} \tag{5}$$

Here $\mathcal{E}_{\mathrm{off},t}$ is the exponentially small off-tube contribution in (44); the smaller $\lambda \log(1+N)$ term is absorbed since $\lambda \log(1+N) \leq 1$. Here N is the number of parallel ULA candidates in Algorithm 1, and ε_{R} denotes a uniform square-root Rényi tolerance: $\max_i R_2(\nu_{t,i} \| \mu_{\theta_t}^{\lambda}) \leq \varepsilon_{\mathrm{R}}^2$, where $\nu_{t,i}$ is the conditional law of the i-th candidate. The hidden constant in the full bound depends only on problem-data constants, not on $(N, \lambda, \varepsilon_{\mathrm{R}}, \gamma, \eta_t, R_v)$ or t. The proof appears in Appendix F.4.

The three selection terms measure, respectively, displacement from the manifold at positive temperature, incomplete coverage by finitely many candidates, and finite-step ULA error relative to Gibbs. The CG residual η_t^2 and selection error are amplified by γ^{-2}; reducing γ lowers the ridge bias γ^2 but increases this sensitivity. The off-tube term covers selection outside the neighborhood where ridge stability holds. Full constants and the expanded expression remain in Appendix F.4. Finally, we translate this error bound to oracle complexity. We count evaluations of $\nabla_x g$ used by ULA, evaluations of f used for hard selection, and the Hessian–vector products used by CG; the number of ULA iterations is chosen so that the square-root Rényi tolerance ε_{R} is sufficiently small.

Theorem 5.3 (Total oracle complexity (informal))**.** Let Assumptions 1 to 4 hold and run HG-MS with $\alpha = 1/L_F$. Suppose $\mathtt{k} \geq 1$. Initialize every ULA chain from $\mathcal{N}(z, \tau\lambda I_d)$, with centers in a fixed compact set and $0 < \tau < 4/(3L_{g,2})$ (Lemma F.18). For all sufficiently small $\varepsilon \in (0,1)$, with the same parameter choices as in Appendix F.4 ensuring $\sup_t \mathbb{E}\|e_t\|^2 = \mathcal{O}(\varepsilon^2)$, Theorem 5.2 and proposition 5.1 yield an ε-stationarity guarantee after $T = \mathcal{O}\big(L_F(F(\theta_0) - F_\star)/\varepsilon^2\big)$ outer iterations. Moreover, under the explicit scaling $N = \lceil \varepsilon^{-4\mathtt{k}} \rceil$, $\lambda = \varepsilon^8 / \log(1+N)$, the square-root Rényi tolerance $\varepsilon_{\mathrm{R}} = \varepsilon^{4+2\mathtt{k}}$, $\gamma = \varepsilon$, $\eta_t = \varepsilon^2$, $R_v = \Theta(\varepsilon^{-1})$, and $K = \widetilde{\mathcal{O}}(d\, C_{\mathrm{PI}}^2 \lambda^{-4} \varepsilon_{\mathrm{R}}^{-2})$, with the joint stepsize/count choice in Proposition F.17 and corollary F.19, each outer iteration uses NK evaluations of $\nabla_x g$ (ULA), N evaluations of f (hard selection), and at most #HVP Hessian–vector products for capped CG (Remark F.22), plus a constant number of upper-gradient and mixed derivative–vector calls. Hence the total oracle complexity for achieving an ε-stationary point is $T(NK + N + \#\mathrm{HVP}) = \widetilde{\mathcal{O}}\Big(d\, C_{\mathrm{PI}}^2 \varepsilon^{-42-8\mathtt{k}}\Big)$. Here $C_{\mathrm{PI}} \geq 1$ is a common bound for all $\theta \in \Theta$ and sufficiently small λ (Proposition B.3); a temperature-dependent bound cannot be treated as fixed in this schedule. This is a conservative worst-case existence guarantee, not a runtime prediction.

Remark 5.4 (On the k-dependence)**.** The intrinsic dimension enters through the minima-selection subproblem, most visibly via the $N^{-1/\mathtt{k}}$ approximation term in Theorem 5.2. Here $\mathtt{k} = \dim \mathcal{S}(\theta) = \dim \ker \nabla_{xx}^2 g(\theta, x)$ at a minimizer. For the regularized autoencoder in Example B.5 (Appendix B.6), $\mathtt{k} = p\ell - \ell(\ell+1)/2$ while $d = 2pm$: $p = 3, \ell = 2$ gives $\mathtt{k} = 3, d = 6m$. Intrinsic dimension controls the coverage exponent; ambient dimension and Poincaré factors remain in sampling cost.

Strongly convex lower level. Then $\mathcal{S}(\theta)$ is a singleton and $\mathtt{k} = 0$, outside the theorem's $\mathtt{k} \geq 1$ regime. ULA/Best-of-N selection is unnecessary and the pseudoinverse becomes the Hessian inverse. With exact hyper-gradients (or $\sum_{t<T} \mathbb{E}\|e_t\|^2 = \mathcal{O}(1)$), Proposition 5.1 gives $\mathcal{O}(1/T)$ average squared stationarity, hence $T = \mathcal{O}(\varepsilon^{-2})$ for stationarity norm ε, consistent with Ji et al. [2021]. This is an outer-iteration comparison, not a separate end-to-end $\mathtt{k} = 0$ oracle bound.

6 Empirical Evaluations

Overview. We first study the ULA/ridge–CG/projected-gradient components of Algorithm 1 on a synthetic problem with exact ground truth. We then evaluate practical HG-MS (Algorithm 3 in Appendix G.7) on the real-world LoRA benchmarks: one differentiable SGD step per candidate, hard selection on a common held-out loss, differentiation through the selected step, and Adam on source logits. It uses $N = 4$ for reasoning (including Qwen2.5-32B), $N = 8$ for instruction following, and

Table 1: Synthetic objective gap $F(\theta) - F_\star$ after 60 seconds: mean $\pm$ bootstrap 95% interval half-width over 20 held-out seeds. Lower is better; best means are bold.

Method	$\mathtt{k} = 1$	$\mathtt{k} = 2$	$\mathtt{k} = 4$	$\mathtt{k} = 10$
HG-MS	$\mathbf{0.00113 \pm 0.00060}$	$\mathbf{0.00522 \pm 0.00279}$	$\mathbf{0.00440 \pm 0.00178}$	$\mathbf{0.0079 \pm 0.0018}$
HG-baseline	0.029 ± 0.010	0.032 ± 0.008	0.034 ± 0.008	0.042 ± 0.009
V-PBGD	0.101 ± 0.040	0.076 ± 0.029	0.067 ± 0.019	0.066 ± 0.017
G-PBGD	2.389 ± 0.863	2.386 ± 0.863	2.599 ± 0.861	2.385 ± 0.861
IAPTT-GM	2.160 ± 0.863	1.945 ± 0.863	1.729 ± 0.863	1.945 ± 0.863

$N = 1$ for HG-baseline. These components replace ULA, ridge–CG, and projected gradient; only parallel candidates, held-out-loss selection, and select-then-differentiate ordering carry over. The coverage, selected-point, MSE, stationarity, and oracle guarantees do not transfer. Proposition 5.1 requires controlled MSE relative to ∇F and its prescribed projected update; neither is established for the finite branches and Adam.

Appendix G.6 reports the full-model GPT-2 version of the practical algorithm with task-dependent inner-step counts; Appendices G.2 and G.3 report MNIST hyper-cleaning [Ren et al., 2018] and imbalanced-loss tuning. The lower-solver choices are summarized in Appendix G.7, Table 5.

6.1 Controlled ULA/ridge–CG mechanism study

For $\theta \in [-3, 3]$, the synthetic lower objective has a known sphere of minimizers of dimension $\mathtt{k} \in \{1, 2, 4, 10\}$ in $\mathbb{R}^{20}$. The optimistic solution and $F(\theta) = 2\sqrt{1.01 + \theta^2}$ are explicit. We compare this ULA/ridge–CG implementation, without clipping, with HG-baseline, V-PBGD, G-PBGD, and IAPTT-GM under paired, randomized-order, single-thread 60-second budgets: 20 held-out seeds give 400 runs. HG-MS has the lowest mean objective gap (Table 1), stationarity, and retained-point error at every tested dimension. Appendix G.5 gives all three metrics, uncertainty, protocols, and separate manifold-coverage, pseudoinverse/ridge–CG, and ULA-versus-Gibbs diagnostics. This is a mechanism study: the stationary origin is a saddle, violating the global critical-point clause of $\mathtt{PŁ}^\circ$.

6.2 Real-world LLM applications

We test the source-reweighting template on the LLM data-composition benchmarks of Pan et al. [2025]: mathematical reasoning and instruction following. These experiments use multi-source LoRA fine-tuning portfolios; Appendix G.7 gives the task definitions, source inventories, budgets, inference protocol, and judging details. Here we summarize the matched-budget outcomes.

Bilevel formulation. Both real-world benchmarks use the same source-reweighting formulation. Let p be the number of training sources, let $\theta \in \mathbb{R}^p$ be source logits, and let $\omega = \mathrm{softmax}(\theta) \in \Delta^{p-1}$ be the induced source weights. Let y denote the trainable LoRA adapter parameters for a fixed base model. With training sources $\mathcal{D}_1^{\mathrm{tr}}, \ldots, \mathcal{D}_p^{\mathrm{tr}}$ and a held-out validation set $\mathcal{D}^{\mathrm{val}}$, the reweighting stage uses $g(\theta, y) := \sum_{s=1}^p \omega_s |\mathcal{D}_s^{\mathrm{tr}}|^{-1} \sum_{\xi \in \mathcal{D}_s^{\mathrm{tr}}} \ell(y; \xi)$ and $f(y) := |\mathcal{D}^{\mathrm{val}}|^{-1} \sum_{\xi \in \mathcal{D}^{\mathrm{val}}} \ell_{\mathrm{val}}(y; \xi)$. Here ℓ is the standard token-level next-token cross-entropy loss used for supervised fine-tuning on source examples, and ℓ_{val} is the same token-level cross-entropy loss evaluated on held-out validation examples during source reweighting. After reweighting, the learned ω is converted into source counts for a final one-epoch supervised fine-tuning run, and the resulting checkpoint is evaluated on the task metrics reported below (GSM8K/MATH accuracy or MT-Bench score).

Mathematical reasoning benchmark. Following [Pan et al., 2025, Section 4.2.2], this benchmark uses a ten-source portfolio and evaluates the final fine-tuned checkpoint on GSM8K and MATH. All methods share the source portfolio, final one-epoch fine-tuning stage, and seed. Table 2 reports final test accuracies in percent after the shared fine-tuning: GSM8K covers grade-school word problems, while MATH covers competition-style mathematical reasoning. For all iterative methods, the matched reweighting stage is approximately 9h for `Qwen-2-7B`, 12h for `Llama-3-8B`, and 15h for `Gemma-2-9B`. Under these budgets, HG-MS improves over HG-BASELINE and is best in every reported GSM8K and MATH column. As supplementary diagnostics for Table 2, Appendix G.7 reports a `Qwen-2-7B` budget sweep in Table 21 and a fixed-budget branch-count sensitivity study in

Table 2: **Mathematical reasoning.** Scores for the ten-source math-reasoning portfolio.

	GSM8K			MATH		
Method	Llama-3-8B	Qwen-2-7B	Gemma-2-9B	Llama-3-8B	Qwen-2-7B	Gemma-2-9B
Uniform Weighting	60.06	60.03	78.08	30.24	30.26	37.29
RHO-LOSS	65.43	63.15	78.47	25.89	33.40	37.35
LESS	74.45	65.28	79.28	29.93	29.24	38.00
ScaleBiO	66.19	75.74	79.62	30.58	36.93	37.98
HG-baseline	73.77	68.46	80.14	30.49	34.27	38.23
HG-MS	**74.85**	**82.03**	**80.86**	**30.86**	**38.98**	**38.48**

Table 3: **Large-model math and instruction following.** Left: `Qwen2.5-32B` math results. Right: MT-Bench scores for the eighteen-source instruction portfolio.

(a) `Qwen2.5-32B` **math**

Method	GSM8K	MATH
Uniform Weighting	78.19	43.30
RHO-LOSS	OOM	OOM
LESS	OOM	OOM
ScaleBiO	88.4	51.29
HG-baseline	86.81	46.82
HG-MS	**92.58**	**53.32**

(b) Instruction following

Method	Llama-3-8B	Qwen-2-7B	Gemma-2-9B
Uniform Weighting	7.07	6.54	8.04
RHO-LOSS	7.09	6.37	8.29
LESS	7.06	6.80	8.19
ScaleBiO	**7.36**	6.58	8.11
HG-baseline	7.32	6.58	8.09
HG-MS	7.33	**6.92**	**8.30**

Table 22. The budget sweep changes only the iterative reweighting time (3h, 9h, 15h, and 21h), while the branch-count study keeps the nine-hour budget fixed and varies the number of candidate branches N. Across these diagnostics, HG-MS learns source distributions distinct from HG-BASELINE and SCALEBIO; the main $N = 4$, 9h configuration gives the strongest completed GSM8K/MATH results.

Large-model results. On Qwen2.5-32B, each completed method uses the same ten-source portfolio, 15h reweighting budget, and 20K-example final fine-tuning set; OOM denotes out-of-memory. HG-MS scores 92.58% on GSM8K and 53.32% on MATH, improving on ScaleBiO by 4.18 and 2.03 points (Table 3, left). Appendix G.7 gives the full evaluation protocol.

Instruction-following benchmark. Following Section 4.2.1 and Table 2 of Pan et al. [2025], this benchmark uses the released eighteen-source ScaleBiO portfolio, mixing general-chat data with multilingual Alpaca variants. Evaluation uses MT-Bench: 80 multi-turn questions scored by a GPT-4o judge on a 0–10 scale. Table 3 (right panel) reports the completed MT-Bench comparison. HG-MS is best on `Qwen-2-7B` (6.92) and `Gemma-2-9B` (8.30), while SCALEBIO is best on `Llama-3-8B` (7.36). The Llama gap is small (7.33 for HG-MS), thus, the table provides matched-budget evidence that explicit minima selection remains competitive in instruction-following source reweighting.

Conclusion and limitations. Optimistic selection on a lower-level minimizer manifold admits a pseudoinverse hyper-gradient under uniqueness; intrinsic non-degeneracy gives the stronger smoothness needed for optimization. Our analysis controls ULA/ridge–CG errors, while LLM experiments evaluate a distinct finite-branch method. Finite training runs, branch diversity, and downstream success do not establish global PŁ$^\circ$, compact connected global minimizers, unique optimistic selection, or uniform intrinsic non-degeneracy: these require a full-landscape characterization. The exact examples establish non-vacuous geometry, not these assumptions for LLMs. Theory for the optimizer-induced law $\nu_{\theta,B}$ remains future work; see Appendix A.

Acknowledgements and disclosures

Niao He is supported by the Swiss National Science Foundation (SNSF) through Project Funding No. 200021-207343 and an SNSF Starting Grant. The authors have no competing interests or relevant financial relationships to disclose.

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.

El Mehdi Achour, François Malgouyres, and Franck Mamalet. Existence, stability and scalability of orthogonal convolutional neural networks. *Journal of Machine Learning Research*, 23(347):1–56, 2022. URL `https://jmlr.org/papers/v23/22-0026.html`.

Michael Arbel and Julien Mairal. Amortized implicit differentiation for stochastic bilevel optimization. In *International Conference on Learning Representations*, 2022a. URL `https://openreview.net/forum?id=3PN4iyXBeF`.

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

Kuang Bai, Jane J Ye, and Shangzhi Zeng. Optimality conditions for bilevel programmes via moreau envelope reformulation. *Optimization*, 74(12):2685–2719, 2025.

Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=HyxnZh0ct7`.

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. Log-concave sampling, 2026. URL `https://chewisinho.github.io/main.pdf`. Manuscript, version 20 August 2026.

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.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL `https://arxiv.org/abs/2110.14168`.

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 `https://www.racetothebottom.xyz/posts/PL-smooth-unique/`.

Mathieu Dagréou, Pierre Ablin, Samuel Vaiter, and Thomas Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/aa84ec1ac3f5fdcf77bce2c22705ab77-Abstract-Conference.html.

John M. Danskin. *The Theory of Max-Min and Its Application to Weapons Allocation Problems*, volume 5 of *Econometrics and Operations Research*. Springer-Verlag, Berlin, Heidelberg, 1967.

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.

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

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.

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.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2103.03874.

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.

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

Feihu Huang. Optimal hessian/jacobian-free nonconvex-PL bilevel optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 19598–19621. PMLR, 2024. URL https://proceedings.mlr.press/v235/huang24a.html.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4882–4892. PMLR, 2021. URL https://proceedings.mlr.press/v139/ji21c.html.

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.

Xiaotian Jiang, Jiaxiang Li, Mingyi Hong, and Shuzhong Zhang. A correspondence-driven approach for bilevel decision-making with nonconvex lower-level problems. *arXiv preprint arXiv:2509.01148*, 2025.

Guy Kornowski, Swati Padmanabhan, Kai Wang, Jimmy Zhang, and Suvrit Sra. First-order methods for linearly constrained bilevel optimization. In *Advances in Neural Information Processing Systems*, volume 37, pages 141417–141460. Curran Associates, Inc., 2024. doi: 10.52202/079017-4491. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ffd86e56e6403d63dd6face033060e5a-Abstract-Conference.html.

Daniel Kunin, Jonathan Bloom, Aleksandrina Goeva, and Cotton Seed. Loss landscapes of regularized linear autoencoders. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3560–3569. PMLR, 2019. URL https://proceedings.mlr.press/v97/kunin19a.html.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D. Nowak. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CvYBvgEUK9.

Yann LeCun, Corinna Cortes, and Christopher J. C. Burges. The MNIST Database of Handwritten Digits. http://yann.lecun.com/exdb/mnist/, 1998.

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, Pan Mu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6305–6315. PMLR, 2020. URL https://proceedings.mlr.press/v119/liu20l.html.

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, 2021a.

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

Yuan Liu. The poincaré inequality and quadratic transportation-variance inequalities. *arXiv preprint arXiv:1902.04196*, 2020.

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.

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

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

Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In Timo Honkela, Włodzisław Duch, Mark Girolami, and Samuel Kaski, editors, *Artificial Neural Networks and Machine Learning – ICANN 2011*, volume 6791 of *Lecture Notes in Computer Science*, pages 52–59, Berlin, Heidelberg, 2011. Springer. doi: 10.1007/978-3-642-21735-7_7. URL https://doi.org/10.1007/978-3-642-21735-7_7.

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.

Georg Menz and André Schlichting. Poincaré and logarithmic sobolev inequalities by decomposition of the energy landscape. *The Annals of Probability*, 42(5):1809–1884, 2014. doi: 10.1214/14-AOP908.

Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15630–15649. PMLR, 2022. URL https://proceedings.mlr.press/v162/mindermann22a.html.

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.

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

Daisuke Oyama and Tomoyuki Takenawa. On the (non-)differentiability of the optimal value function when the optimal solution is unique. *Journal of Mathematical Economics*, 76:21–32, 2018. doi: 10.1016/j.jmateco.2018.02.004.

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.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with GPT-4. *arXiv preprint arXiv:2304.03277*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

Quentin Rebjock and Nicolas Boumal. Fast convergence to non-isolated minima: four equivalent conditions for C^2 functions. *Mathematical Programming*, 213(1–2):151–199, 2025. doi: 10.1007/s10107-024-02136-6.

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*, 214(1):539–589, 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`.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford Alpaca: An instruction-following Llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

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.

Paul Vicol, Jonathan Lorraine, David Duvenaud, and Roger Grosse. Implicit regularization in overparameterized bilevel optimization. In *ICML 2021 Beyond First Order Methods Workshop*, 2021.

Paul Vicol, Jonathan P Lorraine, Fabian Pedregosa, David Duvenaud, and Roger B Grosse. On implicit bias in overparameterized bilevel optimization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22234–22259. PMLR, 2022.

Jiayun Wang, Yubei Chen, Rudrasis Chakraborty, and Stella X. Yu. Orthogonal convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11505–11515, June 2020. URL `https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_Orthogonal_Convolutional_Neural_Networks_CVPR_2020_paper.html`.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 54104–54132. PMLR, 2024. URL `https://proceedings.mlr.press/v235/xia24c.html`.

Quan Xiao, Songtao Lu, and Tianyi Chen. A generalized alternating method for bilevel optimization under the polyak-łojasiewicz condition. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/c981fd12b1d5703f19bd8289da9fc996-Abstract-Conference.html.

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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Appendix

Table of Contents

A Limitations and Future Directions **18**

B Technical Preliminaries and Regularity Assumptions **18**
B.1 Appendix notation guide 18
B.2 ULA sampling and Poincaré preliminaries 20
B.3 Parameter-dependent minimizer geometry 20
B.4 Regularity assumptions for sampling and selection 21
B.5 Uniform geometry and low-temperature sampling constants 23
B.6 Mechanisms and exact examples of minimizer manifolds 25

C Related Work **26**
C.1 Comparison with penalty-based differentiability analyses 28

D Proof of Section 3 **29**
D.1 Proof of Theorem 3.2 29
D.2 Why uniqueness alone is not enough for local smoothness 33
D.3 Proof of Theorem 3.5 34

E Proofs for Section 4 **36**
E.1 Proof of Proposition 4.1 36

F Proofs and Technical Details for Section 5 **37**
F.1 Outer-loop guarantee 38
F.2 Hyper-gradient stability 39
F.3 Selection Error and Proof of Lemma F.20 43
F.4 Parameter choices and oracle complexity 59

G Experimental details **62**
G.1 Common protocol 62
G.2 Data hyper-cleaning (MNIST label noise) 63
G.3 Parametric loss tuning for imbalanced data (MNIST) 64
G.4 Completed outer updates in the MNIST fixed-budget runs 67
G.5 Controlled ULA/ridge-CG mechanism experiment 68
G.6 Controlled GPT-2 source reweighting 72
G.7 Detailed protocols for the real-world LLM experiments 74

A Limitations and Future Directions

The main theoretical contribution of Section 3 is an exact-selection regularity result for optimistic bilevel optimization with manifold lower-level solution sets. Under PŁ$^\circ$, the lower-level minimizers form a structured manifold and the Hessian of g has a normal spectral gap. We show that this non-singleton regime still admits a continuously differentiable optimistic hyper-objective and the pseudoinverse hyper-gradient formula when the optimistic selection is unique. Intrinsic non-degeneracy additionally yields a differentiable selected branch and locally Lipschitz hyper-gradient. These assumptions mark the clean smooth regime for exact optimistic selection: uniqueness prevents branch switching, non-degeneracy stabilizes the restricted minimizer on $\mathcal{S}(\theta)$, and Assumption 1 gives the geometry needed to differentiate through the moving solution set. A natural theoretical direction is to develop analogous exact-selection results under other geometric descriptions of the lower-level landscape, while preserving the same principle that regularity is governed by the stability of the selected lower-level solution.

Sections 4 and 5 turn this exact formula into a select-then-differentiate algorithm. The analyzed version of HG-MS uses ULA/Gibbs candidates because this gives a mathematically controlled candidate distribution near $\mathcal{S}(\theta)$. The resulting oracle-complexity bound is therefore best viewed as a worst-case existence guarantee for this analyzed instantiation, not as a practical runtime model. Its exponent is conservative because the proof separately uses low-temperature sampling, Rényi accuracy of finite-step ULA, finite-candidate selection, and the stabilized pseudoinverse solve. The dependence on the intrinsic dimension $\mathtt{k}$ in the selection term is unavoidable in the worst case, since selecting the best point on a $\mathtt{k}$-dimensional nonconvex minimizer manifold can be reduced to a $\mathtt{k}$-dimensional global optimization problem. The larger dimension-independent exponent comes from combining several uniform worst-case controls in a single end-to-end guarantee. Improving this part of the rate would require new techniques that preserve the present selection-error and hyper-gradient-stability mechanisms while coupling the sampling, selection, and linear-solve accuracies more tightly.

The applied experiments use practical finite-budget lower-level branches; the separate synthetic study in Appendix G.5 tests ULA and ridge–CG directly. In the real-world LLM LoRA source-reweighting tasks, each branch performs only one lower-level SGD update because the lower-level model is large and source reweighting must be run under a fixed wall-clock budget; in the MNIST experiments, the branches use longer AdamW/SGD lower-level runs. This practical version should not be understood as trying to sample the full Gibbs measure. Rather, the branch distribution is induced by initialization, minibatch randomness, optimizer state, perturbations, and the finite training budget. A promising next step is a candidate-generator-centered theory: if $\nu_{\theta,B}$ denotes the law of a B-step optimizer branch, one can study when $\nu_{\theta,B}$ places enough mass on lower-level candidates with strong upper-level validation performance. Then $\mathtt{BoN}$ selection amplifies this validation-good mass across branches. Such a theory would connect the exact optimistic hyper-gradient picture developed here with the finite-time optimizer dynamics used in large-scale bilevel applications.

The real-world LLM implementation replaces ULA/Gibbs sampling, ridge–CG, and projected gradient with one-step SGD branches, selected-step differentiation, and Adam, respectively. Its shared ingredients are parallel candidate generation, hard selection on a common held-out loss, and select-then-differentiate ordering. The Gibbs/Rényi coverage, selected-point error, hyper-gradient MSE, stationarity, and oracle-complexity guarantees do not transfer to this implementation. Proposition 5.1 is conditional on an MSE bound relative to ∇F and its specified projected-gradient update; neither premise is established for the finite branches and Adam.

Verifying global PŁ$^\circ$, compact connected global-minimizer geometry, unique optimistic selection, and uniform intrinsic non-degeneracy requires characterizing the full landscape across θ. Finite training runs, branch diversity, and downstream success do not verify these assumptions. The exact examples in Appendix B.6 establish that such geometry occurs, not that the LLM objectives satisfy it. The synthetic study explicitly discloses the global critical-point condition it violates.

B Technical Preliminaries and Regularity Assumptions

B.1 Appendix notation guide

This subsection supplements the notation paragraph in the introduction. It records mathematical notation used repeatedly in the appendix proofs.

Distance, balls, and spheres.
For a nonempty set $A \subseteq \mathbb{R}^d$, we write

$$\operatorname{dist}(x, A) := \inf_{y \in A} \|x - y\|.$$

We denote the closure and interior of A by $\overline{A}$ and $\operatorname{int}(A)$, respectively. For $a \in \mathbb{R}^d$ and $r \geq 0$, $\mathbb{B}_d(a; r) := \{x \in \mathbb{R}^d : \|x - a\| \leq r\}$ denotes the closed Euclidean ball. For $q \geq 1$, $\mathbb{S}^{q-1} := \{u \in \mathbb{R}^q : \|u\| = 1\}$ denotes the unit sphere in $\mathbb{R}^q$.

Manifold volume and geodesic distance.
If $\mathcal{M} \subset \mathbb{R}^d$ is an r-dimensional embedded submanifold, then $\mathrm{dVol}_{\mathcal{M}}(y)$ denotes the Riemannian volume element induced by the ambient Euclidean metric, and

$$\mathrm{Vol}_r(\mathcal{M}) := \int_{\mathcal{M}} \mathrm{dVol}_{\mathcal{M}}(y)$$

denotes its total r-dimensional volume. We write $d_{\mathcal{M}}(x, y)$ for the geodesic distance between $x, y \in \mathcal{M}$.

Riemannian second-order notation.
For a smooth scalar function h on $\mathcal{M}$, $\mathrm{Hess}_{\mathcal{M}} h(x)$ denotes the Riemannian Hessian of h at x. When $\mathcal{M} = \mathcal{S}(\theta)$, II_x^θ denotes the second fundamental form of $\mathcal{S}(\theta)$ at x.

Lower-level objective shorthand.
We write

$$g_\star(\theta) := \min_{z \in \mathbb{R}^d} g(\theta, z).$$

The common intrinsic dimension of $\mathcal{S}(\theta)$ is denoted by $\mathtt{k}$. In normal-coordinate arguments, we write $q := d - \mathtt{k}$ for the normal dimension.

Linear algebra.
For a matrix A, $\ker(A)$ denotes its null space and $A^\dagger$ denotes its Moore–Penrose pseudoinverse. If A is symmetric, $\lambda_{\min}(A)$ denotes its smallest eigenvalue. We write $\|A\|_{\mathrm{op}}$ for the operator norm.

Asymptotic notation.
We use $\mathcal{O}(\cdot)$ for upper bounds up to problem-dependent constants that are independent of the displayed scaling parameters. We use $\widetilde{\mathcal{O}}(\cdot)$ to additionally suppress logarithmic factors.

Probability measures.
For $x \in \mathbb{R}^d$, δ_x denotes the Dirac measure at x. Given candidates $X_1, \ldots, X_M$, their empirical measure is

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

The Gibbs measure is

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

The marginal laws of approximate ULA candidates are denoted by ν_i.

Lower-tail notation.
For a real random variable Z, its lower u-quantile is

$$q_u(Z) := \inf\{r \in \mathbb{R} : \mathbb{P}(Z \leq r) \geq u\}.$$

The lower-tail superquantile $\mathrm{SQ}_\delta^{\mathrm{low}}(Z)$ is defined in Definition F.6.

Selected candidates and tube events.
For candidates $X_1, \ldots, X_M$, the hard-selected point is

$$\tilde{x} \in \arg\min_{1 \leq i \leq M} f(\theta, X_i).$$

We write $\bar{x} \in \arg\min_{x \in \mathcal{S}(\theta)} \|x - \tilde{x}\|$ for a Euclidean projection of $\tilde{x}$ onto $\mathcal{S}(\theta)$. When each candidate is projected separately, $\Pi(X_i)$ denotes a measurable choice from $\arg\min_{x \in \mathcal{S}(\theta)} \|x - X_i\|$. At outer iteration t,

$$\tilde{x}_t := \hat{x}_N^\lambda(\theta_t), \qquad \mathsf{D}_t := \operatorname{dist}(\tilde{x}_t, \mathcal{S}(\theta_t)), \qquad \mathcal{E}_t := \{\mathsf{D}_t \leq r(\gamma)\}.$$

We write $\mathbf{1}_A$ for the indicator of an event or set A.

B.2 ULA sampling and Poincaré preliminaries

For completeness, we record the standard unadjusted Langevin algorithm (ULA) used as the candidate generator in the ULA/Gibbs instantiation of HG-MS; see, e.g., Chewi et al. [2024].

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

Require: θ; lower objective $g(\theta, \cdot)$; initial state $x^{(0)}$; temperature λ; steps K; stepsize h.

$x \leftarrow x^{(0)}$.
for $s = 0$ to $K - 1$ **do**
  Draw $\xi^{(s)} \sim \mathcal{N}(0, I_d)$.
  $x \leftarrow x - h\nabla_x g(\theta, x) + \sqrt{2\lambda h}\,\xi^{(s)}$.
return x.

Poincaré constant. A probability measure μ over $\mathbb{R}^d$ is said to satisfy the Poincaré inequality with constant C_{PI} if, for every test function $\varphi \in H^1(\mu)$,

$$\int \Big(\varphi - \int \varphi \,\mathrm{d}\mu\Big)^2 \mathrm{d}\mu \;\le\; C_{\mathrm{PI}} \int \|\nabla\varphi\|^2 \,\mathrm{d}\mu,$$

where $H^1(\mu)$ denotes the Sobolev space weighted by μ. For a fixed potential, Gong et al. [2024, Theorem 4] prove a low-temperature bound under geometric and tail assumptions. Parameter uniformity is a separate issue: Proposition B.3 verifies a common bound for $\mathtt{k} \ge 1$ and $0 < \lambda \le \lambda_{\mathrm{PI}}$ under the sampling assumptions stated below. We use $C_{\mathrm{PI}} = \max\{1, C^\star_{\mathrm{PI}}\}$, independent of both θ and the chosen temperature in that range.

B.3 Parameter-dependent minimizer geometry

Proof of Proposition 2.3 and parameter stability. Step 1: fixed-parameter geometry. Coercivity ensures existence of a global minimizer. The connectedness argument of Gong et al. [2024, Proposition 3 and Appendix C.1] uses their Assumptions 1, 2, and 3′: local PŁ, the critical-point condition, and coercivity with compact containment of local minima. These are precisely the fixed-parameter hypotheses here, with $d \ge 2$. Thus all local minima are global and form one connected compact set. The local PŁ–Morse–Bott equivalence [Rebjock and Boumal, 2025, Corollary 2.17], and the local $\mathcal{C}^4$ regularity in Definition 2.2, give the claimed boundaryless $\mathcal{C}^2$ manifold and tangent kernel. For a unit normal Hessian eigenvector with eigenvalue $a > 0$, Taylor expansion in the local PŁ inequality gives $a \le a^2/\mu$, hence $a \ge \mu$. Every other critical point has negative-definite Hessian: inside a local PŁ neighborhood a critical point is a local minimizer, and outside their union this is Definition 2.2(2).

Step 2: construct a graph over a fixed manifold. Fix $\bar\theta$, write $S_0 := \mathcal{S}(\bar\theta)$, and let $k_0 = \dim S_0 < d$. On a sufficiently small normal bundle of S_0, solve

$$P_{N_u S_0} \nabla_x g(\theta, u + n) = 0, \qquad n \in N_u S_0.$$

Its fiber derivative at $(\bar\theta, u, 0)$ is the normal Hessian, with inverse norm at most μ^{-1}. Finitely many local orthonormal normal frames cover compact S_0. Joint $\mathcal{C}^2$ regularity and the implicit-function theorem give a common small fiber radius and parameter neighborhood, with a unique solution $n_\theta(u)$ in that radius. Uniqueness makes the solutions agree on frame overlaps. The section and its derivative in u depend continuously on (θ, u); at $\bar\theta$ they vanish. Compactness therefore gives $\|n_\theta\|_{\mathcal{C}^1(S_0)} \to 0$. No global normal frame is required. The map $\Phi_\theta(u) := u + n_\theta(u)$ is an embedding by the tubular neighborhood theorem and, after shrinking the parameter neighborhood,

$$D\Phi_\theta(u)[T_u S_0] \oplus N_u S_0 = \mathbb{R}^d, \qquad z^\top \nabla^2_{xx} g(\theta, \Phi_\theta(u)) z \ge (\mu/2)\|z\|^2 \quad (z \in N_u S_0).$$

Indeed, the tangential projection of $D\Phi_\theta$ is within $1/2$ of the identity, and the Hessian inequality follows by uniform continuity.

Step 3: identify the graph and obtain uniform conclusions. Maximize $g(\theta, \Phi_\theta(u))$ over compact S_0. At a maximizing point its derivative vanishes on the graph tangent space; the normal equation makes the gradient orthogonal to $N_u S_0$. The direct sum above implies full stationarity. The positive normal direction excludes a negative-definite Hessian, so Step 1 makes this point a global minimizer. The

maximum of g on the graph is therefore its global minimum: every graph point belongs to $\mathcal{S}(\theta)$. Differentiating this stationarity identity along the graph places its k_0-dimensional tangent space in the Hessian kernel. Positivity on the complementary $d - k_0$ directions makes the kernel dimension at most k_0; equality follows. By Step 1 it is also the tangent space of $\mathcal{S}(\theta)$. The inclusion of the graph into $\mathcal{S}(\theta)$ thus has invertible differential and is locally open. Its image is nonempty and compact, hence closed. Connectedness gives $\mathcal{S}(\theta) = \Phi_\theta(S_0)$.

This proves Hausdorff continuity and local constancy of dimension; connectedness of Θ makes the dimension common. A finite parameter cover supplies a ball containing all solution sets, since each graph has uniformly small displacement from its compact reference. The solution graph is closed: pass to the limit in $g(\theta_j, x_j) \le g(\theta_j, z)$ for every fixed $z \in \mathbb{R}^d$. It is consequently compact. The proof uses no sampling hypothesis. ■

B.4 Regularity assumptions for sampling and selection

The assumptions in this subsection are used for the finite-time convergence and complexity analysis in Section 5; they are not needed for the local differentiability statements in Section 3. Their roles are as follows. Assumption 3 gives uniform third-derivative control of $g(\theta, \cdot)$ in a fixed tube around $\mathcal{S}(\theta)$. This is used in two places: first, to obtain the explicit tube radius

$$r(\gamma) = \min\left\{\rho, \frac{\gamma}{2\max\{L_{g,3}, 1\}}\right\}$$

for ridge invertibility in Lemma F.2; and second, through the second-fundamental-form bound in Proposition F.7, to justify the curvature/volume control used in the nearest-neighbor part of the selection analysis. Assumption 4 collects the global analytic controls needed because the ULA and Gibbs candidates are not compactly supported: smoothness and Lipschitz bounds make the hyper-gradient stability constants finite; polynomial growth of the derivative quantities $\partial_\theta g$ and $\nabla^2_{\theta x} g$ gives finite moments and controls the off-tube term in Lemma F.5; and the coercive quadratic-growth condition gives well-defined low-temperature Gibbs measures and Gaussian-type tube tails in Lemma F.9. The global Lipschitz condition on f already implies the linear growth of f needed for tail estimates, so no separate polynomial-growth assumption on f is imposed. Finally, the global intrinsic non-degeneracy in Assumption 2 yields the uniform restricted-curvature constant m_{R} in Proposition B.1. This constant is used only in the selection-error proof: Lemmas F.13, F.14 and F.16 convert a small validation-value gap on the manifold into a small distance to $x^\star(\theta)$.

Our sharp selection-rate analysis also uses volume comparison bounds for small geodesic balls on the compact manifolds $\mathcal{S}(\theta)$, which require the curvature control supplied by Assumption 3.

Assumption 3 (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

$$\sup_{\mathrm{dist}(x,\mathcal{S}(\theta))\le\rho} \left\|\nabla^3_{xxx} g(\theta, x)\right\|_{\mathrm{op}} \;\le\; L_{g,3}.$$

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

- **Regularity of f.** f is $\mathcal{C}^2$ with respect to (θ, x). Moreover, f is $L_{f,1}$-Lipschitz and $L_{f,2}$-smooth with respect to (θ, x). Since Θ is compact, this global Lipschitz condition implies the derived linear-growth bound $|f(\theta, x)| \le C(1 + \|x\|)$, uniformly over $\theta \in \Theta$.
- **Regularity of g.** g is $\mathcal{C}^3$ with respect to (θ, x). 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 $x \in \mathbb{R}^d$,

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

  for constants $C_g, D_g \ge 0$ and an integer $\mathsf{n}_g \ge 1$. We also assume polynomial growth of the mixed derivative used in the hyper-gradient:

  $$\|\nabla^2_{\theta x} g(\theta, x)\| \;\le\; C_{\theta x} \|x\|^{\mathsf{n}_{\theta x}} + D_{\theta x},$$

  for constants $C_{\theta x}, D_{\theta x} \ge 0$ and an integer $\mathsf{n}_{\theta x} \ge 0$. The Hessian $\partial^2_x g(\theta, x)$ is continuous with respect to (θ, x). Moreover, there exist constants $\mathsf{D} > 0$ and $\mu_{\mathsf{qg}} > 0$ such that for all $\|x\| \ge \mathsf{D}$,

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

W.l.o.g. we take D large enough so that $\mathcal{S}(\theta) \subseteq \mathbb{B}_d(0;\mathsf{D})$ for all $\theta \in \Theta$. For the uniform low-temperature Poincaré bound, also assume a uniform tail gradient bound: there exist $\mathsf{D}_{\mathrm{eb}}, \nu_{\mathrm{eb}} > 0$ such that

$$\|\nabla_x g(\theta, x)\| \geq \nu_{\mathrm{eb}} \operatorname{dist}(x, \mathcal{S}(\theta)) \qquad (\|x\| \geq \mathsf{D}_{\mathrm{eb}},\ \theta \in \Theta).$$

Increase D to be at least D_{eb}. This supplies the drift control used in Proposition B.3; quadratic growth alone does not imply this gradient bound. Conversely, global $L_{g,2}$-smoothness implies

$$g(\theta, x) - g_\star(\theta) \geq \frac{\|\nabla_x g(\theta, x)\|^2}{2L_{g,2}} \geq \frac{\nu_{\mathrm{eb}}^2}{2L_{g,2}} \operatorname{dist}(x, \mathcal{S}(\theta))^2 \quad (\|x\| \geq \mathsf{D}),$$

so this sufficient sampling condition also implies the tail quadratic growth above.

Proposition B.1 (Uniform restricted curvature from global non-degeneracy)**.** Let $f \in \mathcal{C}^2$, $g \in \mathcal{C}^3$, and suppose Assumptions 1 and 2 hold. For every $\theta \in \Theta$, let

$$\bar{f}_\theta : \mathcal{S}(\theta) \to \mathbb{R}, \qquad \bar{f}_\theta(x) := f(\theta, x),$$

denote the restriction of the upper-level objective to the lower-level solution manifold. There exists a constant $m_{\mathrm{R}} > 0$, independent of θ, such that

$$\left\langle v, \operatorname{Hess}_{\mathcal{S}(\theta)} \bar{f}_\theta(x^\star(\theta))[v] \right\rangle \ \geq\ m_{\mathrm{R}} \|v\|^2, \qquad \forall \theta \in \Theta,\ \forall v \in \mathcal{T}^\theta_{x^\star(\theta)}.$$

Proof. If $\mathtt{k} = 0$, the tangent spaces are trivial and the claim is vacuous; take any $m_{\mathrm{R}} > 0$. Assume $\mathtt{k} \geq 1$. For each $\bar{\theta} \in \Theta$, Theorem 3.5 and Assumption 2 give a neighborhood $U_{\bar{\theta}}$, a $\mathcal{C}^1$ selected branch $x^\star(\theta)$, and a $\mathcal{C}^2$ local chart $\psi_{\bar{\theta}}(\theta, u)$ for $\mathcal{S}(\theta)$ near $x^\star(\theta)$. Let $u^\star(\theta)$ be the local coordinate satisfying $x^\star(\theta) = \psi_{\bar{\theta}}(\theta, u^\star(\theta))$, and write

$$\tilde{f}_{\bar{\theta}}(\theta, u) := f(\theta, \psi_{\bar{\theta}}(\theta, u)), \qquad A_{\bar{\theta}}(\theta) := D_u \psi_{\bar{\theta}}(\theta, u^\star(\theta)).$$

For a tangent vector $v = A_{\bar{\theta}}(\theta) w$, the second variation of the pullback equals the Riemannian Hessian quadratic form, while $\|v\|^2 = w^\top A_{\bar{\theta}}(\theta)^\top A_{\bar{\theta}}(\theta) w$. Hence the local restricted curvature can be written as the smallest generalized eigenvalue

$$\mu_{\bar{\theta}}(\theta) := \min_{w \neq 0} \frac{w^\top \nabla^2_{uu} \tilde{f}_{\bar{\theta}}(\theta, u^\star(\theta)) w}{w^\top A_{\bar{\theta}}(\theta)^\top A_{\bar{\theta}}(\theta) w}.$$

This quantity is continuous on $U_{\bar{\theta}}$ and is positive at every point by the intrinsic non-degeneracy in Assumption 2. Shrinking $U_{\bar{\theta}}$ if necessary, there is a smaller neighborhood $V_{\bar{\theta}}$ with $\overline{V_{\bar{\theta}}} \subset U_{\bar{\theta}}$ on which $\inf_{\theta \in V_{\bar{\theta}}} \mu_{\bar{\theta}}(\theta) > 0$. Since Θ is compact, choose a finite subcover $V_{\bar{\theta}_1}, \ldots, V_{\bar{\theta}_J}$ and set

$$m_{\mathrm{R}} := \min_{1 \leq j \leq J} \inf_{\theta \in V_{\bar{\theta}_j}} \mu_{\bar{\theta}_j}(\theta) > 0.$$

This gives the stated uniform lower bound. ∎

Proposition B.1 is the uniform version of the local intrinsic non-degeneracy condition in Definition 3.4. The Hessian of the restricted objective includes the curvature of the constraint manifold, as shown by the identity below. With the convention for the second fundamental form used in Proposition F.7, for $v \in \mathcal{T}^\theta_{x^\star(\theta)}$,

$$\left\langle v, \operatorname{Hess}_{\mathcal{S}(\theta)} \bar{f}_\theta(x^\star(\theta))[v] \right\rangle = \langle v, \nabla^2_{xx} f(\theta, x^\star(\theta)) v \rangle + \langle \nabla_x f(\theta, x^\star(\theta)), \mathrm{II}^\theta_{x^\star(\theta)}(v, v) \rangle. \tag{6}$$

Here $\mathrm{II}^\theta_{x^\star(\theta)}$ denotes the second fundamental form of the embedded manifold $\mathcal{S}(\theta)$ at $x^\star(\theta)$. Thus the conclusion of Proposition B.1 holds, for instance, if the ambient tangent Hessian has a uniform positive margin and the second-fundamental-form correction is not too negative. Concretely, it is sufficient that for some $m_{\mathrm{amb}} > 0$ and $\kappa < m_{\mathrm{amb}}$,

$$\langle v, \nabla^2_{xx} f(\theta, x^\star(\theta)) v \rangle \geq m_{\mathrm{amb}} \|v\|^2, \qquad \langle \nabla_x f(\theta, x^\star(\theta)), \mathrm{II}^\theta_{x^\star(\theta)}(v, v) \rangle \geq -\kappa \|v\|^2$$

uniformly over $\theta \in \Theta$ and $v \in \mathcal{T}^\theta_{x^\star(\theta)}$. Using Proposition F.7, a further sufficient condition is

$$m_{\mathrm{amb}} > \sup_{\theta \in \Theta} \|\nabla_x f(\theta, x^\star(\theta))\| \frac{L_{g,3}}{c},$$

where c is the uniform normal spectral gap in Proposition 2.3.

We close this subsection by comparing these assumptions with common assumptions in related analyses. The smoothness and Lipschitz-type parts of Assumption 4 are standard in hyper-gradient analyses of bilevel optimization: they play the same role as the uniform derivative bounds used to control implicit-gradient stability and inexact lower-level solves [Kwon et al., 2024, Chen et al., 2024]. Many bilevel convergence results additionally avoid tail issues by imposing a bounded lower-level domain, bounded gradients, or bounded iterates. Our sampling-based method does not have compactly supported candidates, because the Gibbs and ULA laws live on $\mathbb{R}^d$. We therefore replace those bounded-domain requirements with derivative polynomial-growth bounds, the linear-growth consequence of Lipschitzness of f, and coercive quadratic growth away from $\mathcal{S}(\theta)$. This is the standard type of tail-control condition used in low-temperature Gibbs/Langevin analyses around minimizer sets [Hasenpflug et al., 2024, Chewi et al., 2024]. It is weaker than assuming all relevant quantities are bounded on $\mathbb{R}^d$, but strong enough to make the Gibbs normalizing constants finite, to prove tube concentration, and to control the off-tube contribution in the hyper-gradient error.

The manifold-specific assumptions have a different purpose. Assumption 3 is a quantitative version of local regularity near the minimizer manifold; it is analogous to the Hessian-Lipschitz or third-order smoothness assumptions often used when one needs stable local quadratic models, but here it is imposed only in a tube around $\mathcal{S}(\theta)$. Together with the normal spectral gap from PŁ$^\circ$, it provides the uniform curvature bounds required for the manifold-volume and nearest-neighbor estimates used in the selection analysis; related manifold Gibbs estimates appear in Masiha et al. [2025]. Finally, Proposition B.1 is the manifold analogue of a uniform second-order sufficient condition for the upper-level restricted problem $f(\theta,\cdot)|_{\mathcal{S}(\theta)}$. This conclusion is vacuous in the classical singleton-minimizer regime, but in the non-singleton manifold regime it is what converts validation-value suboptimality into distance to the unique optimistic selection. Thus the additional assumptions in this subsection are not substitutes for the local differentiability theory in Section 3; they are uniform, quantitative conditions needed for the finite-sample BoN selection rate and the resulting oracle-complexity bound.

B.5 Uniform geometry and low-temperature sampling constants

Lemma B.2 (Common tubes and finite-temperature projected mass)**.** Let Assumption 1 hold on its surrounding open parameter domain, and choose a compact convex $\mathcal{K}$ in that domain with $\Theta \subset \operatorname{int}\mathcal{K}$. Assume Assumptions 3 and 4 on Θ, and suppose $k := \mathtt{k} \geq 1$. No uniqueness or intrinsic non-degeneracy of upper-level selection is required here. There are common constants $r_{\mathrm{tube}}, J_{\min}, J_{\max}, V_{\min}, V_{\max}, r_{\mathrm{nn}}, v_-, D_{\mathrm{geo}} > 0$. Normal addition parametrizes the r_{tube}-tube, its volume Jacobian lies in $[J_{\min}, J_{\max}]$, and the manifold volumes lie in $[V_{\min}, V_{\max}]$. Intrinsic diameters are at most D_{geo}, and every intrinsic ball of radius $0 < r \leq r_{\mathrm{nn}}$ has volume at least $v_- r^k$.

Write $q = d - k$, $h_\theta = g(\theta,\cdot) - g_\star(\theta)$, and

$$A_{\lambda,\theta}(u) := \int_{\substack{n \in N_u\mathcal{S}(\theta) \\ \|n\| < r_{\mathrm{tube}}}} e^{-h_\theta(u+n)/\lambda} J_\theta(u,n)\,\mathrm{d}n, \qquad Z_\lambda(\theta) := \int_{\mathbb{R}^d} e^{-h_\theta(x)/\lambda}\,\mathrm{d}x.$$

For some common $\lambda_{\mathrm{cov}}, a_-, z_+ > 0$,

$$A_{\lambda,\theta}(u) \geq a_- \lambda^{q/2}, \qquad Z_\lambda(\theta) \leq z_+ \lambda^{q/2} \quad (0 < \lambda \leq \lambda_{\mathrm{cov}},\ u \in \mathcal{S}(\theta)).$$

Consequently a Gibbs sample X and any fixed measurable nearest-point map Π_θ satisfy

$$\mathbb{P}\big(d_{\mathcal{S}(\theta)}(\Pi_\theta X, p) \leq r\big) \geq c_0 r^k, \qquad c_0 := (a_-/z_+)v_- > 0, \quad p \in \mathcal{S}(\theta),\ 0 < r \leq r_{\mathrm{nn}}.$$

Reduce r_{nn} so that $c_0 r_{\mathrm{nn}}^k \leq 1/2$. Here $d_{\mathcal{S}(\theta)}$ is intrinsic distance.

Proof. Uniform geometric bounds. Proposition 2.3 gives a compact solution graph and finitely many normal-graph representations with tangent singular values in $[1/2, 3/2]$, after restricting each parameter neighborhood. The induced volume Jacobians lie in $[2^{-k}, (3/2)^k]$, giving common positive volume bounds. Mapping curves from each connected compact reference manifold bounds intrinsic diameters by $3/2$ times the largest reference diameter.

Lemma D.2 supplies jointly $\mathcal{C}^2$ coordinate maps on finitely many relatively compact neighborhoods of the solution graph. Their first derivatives have uniformly bounded singular values away from zero

and their second derivatives are bounded. In smaller coordinate neighborhoods, a fixed coordinate ball about any center is mapped into an intrinsic ball of a comparable radius: map the straight coordinate segments and use the derivative bound. Integrating the coordinate volume density gives $v_- r^k$ on a common radius range.

These charts also give a common tubular radius. Otherwise, there would be parameters θ_j and points at distance tending to zero with two distinct nearest minimizers. Compactness puts both minimizers in one such coordinate chart near a limiting minimizer. The coordinate Hessian of squared distance to the ambient point is $2D\psi^\top D\psi + 2\langle \psi - x, D^2\psi\rangle$; it is positive definite in a sufficiently small convex coordinate neighborhood, uniformly for nearby parameters and points. Both minimizers eventually lie there, contradicting strict convexity. The local normal map is invertible for small radii since $\|\mathrm{II}\| \le L_{g,3}/c$ by Proposition F.7. Its volume factor is $J_\theta(u,n) = \det(I - A_n)$, where the shape operator satisfies $\langle A_n v, w\rangle = \langle \mathrm{II}(u)[v,w], n\rangle$. Indeed the normal component of a normal-field derivative is orthogonal to its tangential component $-A_n v$; the latter changes the tangential volume by this determinant, while the fiber volume is unchanged. Thus, reducing the common radius, $J_{\min} \le J_\theta \le J_{\max}$. This reasoning uses local frames, not a global trivialization.

Fiber mass and normalization. Global smoothness gives $h_\theta(u+n) \le L_{g,2}\|n\|^2/2$. For $\sqrt{\lambda} \le r_{\mathrm{tube}}$, integration on $\|n\| \le \sqrt{\lambda}$ gives

$$A_{\lambda,\theta}(u) \ge J_{\min} e^{-L_{g,2}/2} \omega_q \lambda^{q/2} =: a_- \lambda^{q/2},$$

where ω_q is the Euclidean unit-ball volume. Lemma F.8, whose proof uses only normal curvature, compactness, and the tail assumption, gives $h_\theta(x) \ge \mu_{\mathrm{QG}} \operatorname{dist}(x, \mathcal{S}(\theta))^2/2$. Inside the tube, integration in normal coordinates therefore bounds the normalizer by $J_{\max} V_{\max} (2\pi/\mu_{\mathrm{QG}})^{q/2} \lambda^{q/2}$. Outside, split the exponential in half. One factor is at most $\exp[-\mu_{\mathrm{QG}} r_{\mathrm{tube}}^2/(4\lambda)]$; the other has a uniformly bounded integral for $\lambda \le 1$, since $\operatorname{dist}(x, \mathcal{S}(\theta)) \ge (\|x\| - \mathsf{D})_+$. The exponential absorbs $\lambda^{q/2}$, proving the normalizer bound.

For a Borel set $B \subset \mathcal{S}(\theta)$, the mass of $\{X \text{ in the tube}, \ \Pi_\theta X \in B\}$ is $\int_B A_{\lambda,\theta}(u)/Z_\lambda(\theta)\, \mathrm{dVol}(u)$. It is a subprobability measure, not the limiting manifold law or a conditional probability. Its density is at least a_-/z_+. The intrinsic-ball volume bound proves the last assertion, independently of how the nearest-point map is chosen outside the tube. ■

Proposition B.3 (Parameter-uniform low-temperature Poincaré bound)**.** Under the assumptions of Lemma B.2, there are $\lambda_{\mathrm{PI}} > 0$ and $C_{\mathrm{PI}}^\star < \infty$, independent of θ, such that

$$\mathrm{Var}_{\mu_\theta^\lambda}(\varphi) \le C_{\mathrm{PI}}^\star \int \|\nabla\varphi\|^2 \, \mathrm{d}\mu_\theta^\lambda \qquad (\theta \in \Theta,\ 0 < \lambda \le \lambda_{\mathrm{PI}}).$$

We use $C_{\mathrm{PI}} := \max\{1, C_{\mathrm{PI}}^\star\}$ below.

Proof. We give the parameter-uniform version of the tube and Lyapunov argument of Gong et al. [2024, Theorem 4]; their fixed-potential result alone does not establish this uniformity. Put $L = L_{g,2} > 0$, $D_\Delta = dL$, and $h_\theta = g(\theta,\cdot) - g_\star(\theta)$.

Step 1: common analytic bounds. Normal Taylor expansion, compactness of the solution graph, and the normal gap c give a common $r > 0$ with $\|\nabla h_\theta(x)\| \ge \nu \operatorname{dist}(x, \mathcal{S}(\theta))$ on the r-tube, where $\nu = c/2$. To see this, take a nearest minimizer u, put $n = x - u \in N_u\mathcal{S}(\theta)$, and integrate $n^\top \nabla^2 h_\theta(u + tn) n \ge (c/2)\|n\|^2$ along the segment. Also $|\Delta h_\theta| \le D_\Delta$. Choose $R \ge \max\{\mathsf{D}, 1\}$. Every nonminimum critical point lies in the compact region $\|x\| \le 2R$, outside the r-tube, by the tail and local gradient bounds. Their joint set is compact and their Hessians are negative definite by Proposition 2.3's proof. Thus an open neighborhood of that set has $\nabla^2 h_\theta \preceq -aI$ for one $a > 0$. The remaining compact region has gradient norm at least $b > 0$. Empty regions can be omitted, choosing arbitrary positive a, b.

Step 2: uniform Poincaré control in shrinking tubes. The finite normal-graph cover in Proposition 2.3 compares each manifold metric and volume to those of a fixed compact connected reference manifold. Comparing Rayleigh quotients gives a common positive lower bound for their first nonzero Laplace–Beltrami eigenvalues. On the radius-s normal disk bundle, use the product of normalized base volume and uniform fiber volume. Variance decomposition gives a Poincaré constant at most $\max\{C_{\mathrm{base}}, C_q s^2\}$: apply the Euclidean-ball inequality within each fiber and the base inequality

to the fiber mean. Normal parallel transport preserves fiber volume, so differentiation and Jensen's inequality control the gradient of that mean by horizontal gradient energy. This argument uses local frames and is valid for a nontrivial normal bundle. The tubular metric and density differ by bounded factors: its differential is $(v, w) \mapsto (I - A_n)v + w$, and its density is $\det(I - A_n)$. For small common s, Lemma B.2 therefore gives a common Poincaré constant for uniform volume in every ambient s-tube.

Set $C = 4D_\Delta/\nu^2$ and $s = \sqrt{C\lambda}$. In this tube $0 \le h_\theta \le Ls^2/2$. The Gibbs density relative to uniform volume consequently has ratio of maximum to minimum at most $e^{LC/2}$. Comparing variance and gradient integrals gives a common conditional Poincaré constant C_U, independent of θ, λ.

Step 3: control the complement. For $W = e^{h_\theta/(2\lambda)}$ and $\mathcal{L} = -\nabla h_\theta \cdot \nabla + \lambda\Delta$,

$$\frac{\mathcal{L}W}{\lambda W} = \frac{\Delta h_\theta}{2\lambda} - \frac{\|\nabla h_\theta\|^2}{4\lambda^2}.$$

Choose a common $\lambda_{\mathrm{PI}} \le 1$ so that $\sqrt{C\lambda_{\mathrm{PI}}} \le r$ and lies within the tube range of Step 2, and so that $\lambda_{\mathrm{PI}} \le \min\{b^2, \nu_{\mathrm{eb}}^2 R^2\}/(4D_\Delta)$. Between distances $\sqrt{C\lambda}$ and r, the local gradient bound makes this drift at most $-D_\Delta/(2\lambda)$. The same bound holds on the compact remainder by b, and on $\|x\| \ge 2R$ by $\|\nabla h_\theta(x)\| \ge \nu_{\mathrm{eb}}\|x\|/2$. Near nonminimum critical points the drift is at most $-da/(2\lambda)$. Writing U for the shrinking tube and $s_0 := \min\{D_\Delta, da\}/2 > 0$, we obtain

$$\frac{\mathcal{L}W}{\lambda W} \le -\sigma + B\mathbf{1}_U, \qquad \sigma = s_0/\lambda, \qquad B = (s_0 + D_\Delta/2)/\lambda.$$

For a smooth test function set $F = \varphi - \mathbb{E}[\varphi \mid U]$. Integration by parts gives $\int F^2[-\mathcal{L}W/(\lambda W)]\,\mathrm{d}\mu_\theta^\lambda \le \int \|\nabla\varphi\|^2\,\mathrm{d}\mu_\theta^\lambda$. This follows by expanding the nonnegative square $\|\nabla F - F\nabla W/W\|^2$; cutoffs are justified by the quadratic tail bound. The drift and conditional Poincaré inequality imply

$$\mathrm{Var}_{\mu_\theta^\lambda}(\varphi) \le [\lambda/s_0 + (1 + D_\Delta/(2s_0))C_U] \int \|\nabla\varphi\|^2\,\mathrm{d}\mu_\theta^\lambda.$$

Replace λ by λ_{PI} to define $C_{\mathrm{PI}}^\star$; density extends the inequality to H^1. All choices were uniform over Θ. No sampling bound outside Θ, nor global parameter-Lipschitz bound on $\nabla_x g$, was used. ■

B.6 Mechanisms and exact examples of minimizer manifolds

One general mechanism behind such solution manifolds is a zero-residual objective

$$g(\theta, x) = \tfrac{1}{2}\|r_\theta(x)\|^2, \qquad r_\theta : \mathbb{R}^d \to \mathbb{R}^q. \tag{7}$$

Assume that the residual map r_θ is smooth. If zero residual is attainable and $J_\theta(x) := \nabla_x r_\theta(x)$ has full row rank at every zero, the regular-level-set theorem gives $\mathcal{S}(\theta) = r_\theta^{-1}(0)$ the structure of a smooth manifold. At every such point,

$$\nabla_{xx}^2 g(\theta, x) = J_\theta(x)^\top J_\theta(x),$$

whose kernel is the tangent space and whose restriction to the normal space is positive definite. Continuous parameter symmetries likewise generate families of equivalent solutions, while tying, normalization, or regularization can fix scale and make those families compact. For instance, if $W \in \mathbb{R}^{M\times D}$ with $1 \le M < D$, then $\|WW^\top - I_M\|_F^2$ has the compact connected row-Stiefel manifold $\{W : WW^\top = I_M\}$ as its minimizer set; related orthogonality penalties are used in convolutional networks [Wang et al., 2020, Achour et al., 2022]. The strict inequality $M < D$ matters: when $M = D$, the orthogonal group has two connected components. These mechanisms explain how manifolds form, but they do not by themselves verify every global clause of PŁ$^\circ$.

Example B.4 (Tied linear convolutional autoencoder)**.** Let $z_1, \dots, z_M \in \mathbb{R}^D$, with $D = Cq^2 \ge 2$, be whitened $q \times q$ patches satisfying $M^{-1}\sum_{j=1}^M z_j z_j^\top = I_D$. A one-filter tied linear convolutional autoencoder with filter $w \in \mathbb{R}^D$ reconstructs z_j as $ww^\top z_j$. Its full-dataset loss is

$$g_{\mathrm{conv}}(w) = \frac{1}{2M}\sum_{j=1}^M \|ww^\top z_j - z_j\|^2 = \tfrac{1}{2}(\|w\|^2 - 1)^2 + \tfrac{D-1}{2}.$$

Consequently, $\arg\min g_{\mathrm{conv}} = \mathbb{S}^{D-1}$, a compact connected smooth manifold. The only other stationary point is the strict local maximum $w = 0$, so this example also satisfies the PŁ$^\circ$ geometry and specializes the tied convolutional architecture and reconstruction objective of Masci et al. [2011].

Example B.5 (Regularized linear autoencoder)**.** Let a centered data matrix $X \in \mathbb{R}^{m\times n}$ have singular value decomposition $X = U\Sigma V^\top$ with distinct positive singular values $\sigma_1 > \cdots > \sigma_m > 0$. Take latent width $p \le m$, encoder $W_1 \in \mathbb{R}^{p\times m}$, decoder $W_2 \in \mathbb{R}^{m\times p}$, and the unnormalized objective

$$g_{\mathrm{LAE}}(W_1, W_2) = \|X - W_2W_1X\|_F^2 + \lambda\big(\|W_1\|_F^2 + \|W_2\|_F^2\big).$$

For $0 < \ell < p$ and $\sigma_\ell^2 > \lambda > \sigma_{\ell+1}^2$, define $D_\ell = \mathrm{diag}_{i\le\ell}\sqrt{1-\lambda/\sigma_i^2}$ and let U_ℓ contain the leading ℓ left singular vectors. The complete global-minimizer set is

$$\left\{\big(OD_\ell U_\ell^\top,\ U_\ell D_\ell O^\top\big) : O \in \mathbb{R}^{p\times\ell},\ O^\top O = I_\ell\right\},$$

where the two entries are (W_1, W_2). It is parameterized by the compact connected Stiefel manifold $V_\ell(\mathbb{R}^p)$ and has intrinsic dimension $\mathtt{k} = p\ell - \ell(\ell+1)/2$, versus ambient parameter dimension $d = 2pm$. In particular, $p = 3, \ell = 2$ gives $\mathtt{k} = 3$ while $d = 6m$, illustrating why intrinsic dimension controls the BoN coverage exponent even though ambient dimension remains in sampling cost.

The threshold depends on loss normalization. If instead the objective is $\frac{1}{2n}\|X - W_2W_1X\|_F^2 + \frac{\lambda}{2}(\|W_1\|_F^2 + \|W_2\|_F^2)$, then the correct condition is $\sigma_\ell^2/n > \lambda > \sigma_{\ell+1}^2/n$, and D_ℓ contains $\sqrt{1-n\lambda/\sigma_i^2}$. The critical-point characterization follows from Kunin et al. [2019, Thm. 4.2]. Since $\lambda > 0$, the objective is coercive and every global minimum is critical. The strict threshold leaves exactly the admissible directions $1, \ldots, \ell$. For a critical family indexed by $I \subseteq \{1, \ldots, \ell\}$, the unnormalized objective value is $\sum_{i=1}^m \sigma_i^2 - \sum_{i\in I}(\sigma_i^2 - \lambda)^2/\sigma_i^2$. Each subtraction is positive and all $\ell < p$ directions fit within the latent width. Thus precisely $I = \{1, \ldots, \ell\}$ gives the global-minimizer family above. It establishes exact compact connected minimizer-manifold geometry; it does *not* by itself establish the complete global PŁ$^\circ$ assumption, because other critical manifolds can be saddles.

C Related Work

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 optimistic and pessimistic formulations already recognize that nonunique lower-level solutions create a genuine *selection* issue [Dempe et al., 2007, Ye et al., 1997, Ye and Ye, 1997]. Standard single-level reformulations, whether via value functions or KKT/MPEC systems [Ye and Zhu, 1995, 2010, Dempe and Zemkoho, 2012, Zemkoho and Zhou, 2021], provide an important optimization viewpoint. Our focus here is a different question: when does the induced hyper-objective admit a usable hyper-gradient?

Classical hyper-gradient methods. Most classical hyper-gradient methods study the regime in which the lower-level minimizer is a singleton, so the lower-level response can be treated as a single-valued and differentiable map of θ, typically because their assumptions (such as strong convexity) guarantee that the relevant lower-level Hessian is invertible. In this regime, the hyper-objective $F(\theta) = f(\theta, x(\theta))$ is differentiated exactly by implicit differentiation, while truncated backpropagation is used to approximate its hyper-gradient by differentiating through finitely many lower-level iterations [Pedregosa, 2016, Franceschi et al., 2017, Lorraine et al., 2020, Shaban et al., 2019]. Closely related methodological work studies how to differentiate through parameterized optimization problems more generally. Ablin et al. [2020] analyze gradients of functions defined as minima and compare analytic implicit gradients with automatic differentiation through iterative solvers. Arbel and Mairal [2022a] develop amortized inexact implicit differentiation for stochastic bilevel optimization with strongly convex inner problems. Blondel et al. [2022] develop a general framework for implicit differentiation of parameterized optimization problems: given an optimality system defining a single-valued solution map, they differentiate the solution by applying the implicit function theorem to that system. Their focus is modular implicit-differentiation machinery rather than minima selection from a non-singleton lower-level solution set. These works are closely related in technique, but they still largely assume a well-defined single-valued solution map, and therefore do not address the minima-selection issue that arises when the lower-level minimizer set is non-singleton.

Degenerate implicit differentiation and pseudoinverse formulas. A separate line studies bilevel differentiation when the lower-level Hessian is singular. Most relevant for us, Arbel and Mairal [2022b] study bilevel games in which the lower-level problem may have multiple critical points and a selection map chooses the particular critical point reached by a prescribed lower-level process (e.g., gradient flow). They show that, under a parametric Morse–Bott condition, this algorithmic critical-point selection map remains differentiable near local minima even when the lower Hessian is singular, and its derivative is characterized by a pseudoinverse-based implicit-differentiation formula. In a different overparameterized setting, Vicol et al. [2021, 2022] consider lower-level objectives with nonunique minimizers and study the minimum-norm lower-level solution selected by the optimization dynamics. Because this selected solution is defined through a singular linear system, its response Jacobian is characterized by the minimum-norm solution of that system, which is expressed using the Moore–Penrose pseudoinverse.

The pseudoinverse formulas in these works are technically related to ours, but the object being differentiated is different. Arbel and Mairal [2022b] differentiate a prescribed algorithmic critical-point selection map, and the overparameterized line [Vicol et al., 2021, 2022] differentiates a minimum-norm or algorithm-induced lower-level response. In contrast, we differentiate the exact optimistic value

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

where $\mathcal{S}(\theta)$ is the manifold of global lower-level minimizers and the selected point is determined by the upper-level objective itself. Thus our contribution is to show that, under PŁ° and local uniqueness of the optimistic minimizer, this hard optimistic value function is differentiable and its exact hyper-gradient admits a pseudoinverse formula.

Alternative formulations, selection-based models, and nonsmooth bilevel objectives. Several recent works address nonconvex or set-valued lower levels by modifying either the object being optimized or the solution concept. Liu et al. [2021b] propose a trajectory-based approximation framework for nonconvex followers, introducing an initialization auxiliary and a pessimistic trajectory truncation rule so that the upper-level method optimizes through finite lower-level dynamics rather than through a globally solved lower-level problem. Arbel and Mairal [2022b] introduce bilevel games with critical-point selection maps, where the lower-level object is a selected critical point reached by a prescribed optimization process, and they study how finite-budget unrolled methods relate to equilibria of these selection-based games. Under a Morse-type parametric qualification, Bolte et al. [2025] show that the graphs of lower-level critical points and local minima decompose into finitely many smooth branches and use this structure to define branch-wise hyper-objectives and gradient methods along individual branches. Jiang et al. [2025] argue that, for general nonconvex lower levels, the classical hyper-objective may be unsettled if global minimizers are unattainable, and instead propose a correspondence-driven hyper-objective tied to the output of a prescribed lower-level algorithm, together with Gaussian smoothing and projected stochastic-gradient guarantees. Chen et al. [2025] take a different nonsmooth route: rather than recovering classical differentiability, they introduce a set-smoothness condition that yields weak convexity/concavity of the induced hyper-objective and enables computation of approximate Clarke hyper-stationary points.

Another recent line develops Moreau-envelope single-level reformulations as an alternative to classical value-function and KKT-based reductions. Bai et al. [2025] derive directional necessary optimality conditions in this framework, Gao et al. [2023] propose a difference-of-weakly-convex reformulation with a sequentially convergent algorithm, and Lu [2025] develop a two-sided smoothed primal–dual method for general nonconvex bilevel problems with KKT-type convergence guarantees. Our focus is different from all of these directions: we retain the classical optimistic hyper-objective $F(\theta) = \min_{x \in \mathcal{S}(\theta)} f(\theta, x)$ itself and ask when, despite the lower-level solution set being manifold-valued, that object is differentiable and even locally smooth.

A recent sampling-based relaxation [Masiha et al., 2025] replaces hard minima selection by a superquantile–Gibbs surrogate and studies a zeroth-order method that optimizes the resulting nonsmooth Lipschitz objective through sampling-based function-value information rather than exact hyper-gradients. That approach is designed to circumvent the nondifferentiability of hard minima selection by changing the objective being optimized. The present paper takes a different route: we keep the hard optimistic objective, obtain $\mathcal{C}^1$ regularity and the exact pseudoinverse hyper-gradient from uniqueness of the optimistic minimizer under Assumption 1, and obtain local smoothness with the additional intrinsic non-degeneracy condition. We then analyze a ULA/ridge–CG/projected-gradient

method with an end-to-end stationarity guarantee. In this sense, the present paper can be viewed as a differentiable counterpart to that relaxation-based, zeroth-order line.

Relaxing strong convexity through global regularity or convexity assumptions. Recent works on bilevel optimization study nonconvex lower-level problems under *global* PŁ or closely related uniform error-bound assumptions [Chen et al., 2024, Kwon et al., 2024, Xiao et al., 2023, Huang, 2024, Shen et al., 2025a, Liu et al., 2022a]. One attraction of this line is that such global regularity helps guarantee a smooth and analytically tractable hyper-objective [Kwon et al., 2024, Chen et al., 2024]. From the viewpoint of the lower-level solution set, however, these assumptions are quite restrictive. For a $\mathcal{C}^2$ lower-level objective satisfying a global PŁ condition, the minimizer set $\mathcal{S}(\theta)$ can only be either a singleton or an unbounded subset of $\mathbb{R}^d$ [Criscitiello et al., 2025, Masiha et al., 2025]. Therefore, once one adds the boundedness or coercivity assumptions typically used in this literature, the lower-level problem collapses back to the singleton-minimizer regime, and bounded non-singleton minimizer manifolds are excluded.

This restriction is also visible in the specific assumptions used by these papers. Huang [2024] assume a nondegenerate lower Hessian at global minimizers, which effectively enforces local uniqueness; Liu et al. [2022a] explicitly assume that $\mathcal{S}(\theta)$ is a singleton; and Kwon et al. [2024] impose a proximal error bound for smooth g together with coercivity, which again yields a singleton solution [Chen et al., 2024, cf. Prop. C.1]. Other extensions obtain differentiability or algorithmic guarantees under convexity of the lower-level objective in x [Xiao et al., 2023, Shen et al., 2025a]; this allows non-strongly convex followers, but it still restricts $\mathcal{S}(\theta)$ to a convex geometry rather than the nonconvex compact manifold-valued setting that motivates our work. A related direction is Chen et al. [2025], who study approximate Clarke hyper-stationarity via a new set-smoothness property of the lower-level solution map. Their framework assumes, in particular, that for every θ, the lower-level solution set $\mathcal{S}(\theta)$ is nonempty, closed, and convex, together with a uniform error-bound condition (equivalently, under smoothness, a global PŁ condition). Under these assumptions, they show that the hyper-objective is weakly convex/concave and develop an algorithm for approximate Clarke hyper-stationarity. From the perspective of lower-level geometry, this is still quite different from our setting: their analysis is tailored to convex solution sets, whereas we study compact nonconvex manifold-valued minimizer sets under the local $\texttt{PŁ}^\circ$ condition.

By contrast, our analysis is based on the $\texttt{PŁ}^\circ$ condition of Gong et al. [2024], which is specifically designed to allow a connected compact set of global lower-level minimizers with $\mathcal{C}^2$ manifold structure. This regime remains rich enough to cover overparameterized deep-learning landscapes where connected minima have been observed [Draxler et al., 2018, Garipov et al., 2018, Nguyen, 2019], and it keeps minima selection genuinely relevant rather than collapsing the problem back to the singleton-minimizer case.

C.1 Comparison with penalty-based differentiability analyses

We compare our standing assumptions to penalty-based differentiability analyses such as Kwon et al. [2024] 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. [2024] 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 $\mathcal{C}^2$ objectives, a global PŁ condition implies that the minimizer set is either a singleton or unbounded [Criscitiello 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 [Masiha 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. [2024]. Other works enforce uniqueness more directly (without explicitly invoking global PŁ), e.g., Liu et al. [2022a].

Why our assumptions are weaker than σ-uniform local PŁ for $g + \sigma f$. The key obstruction is already visible by testing a putative PŁ inequality on points of the minimizer manifold. On $\mathcal{S}(\theta)$, the lower objective g is constant and $\nabla_x g = 0$. Thus the value gap of $h_\sigma = g + \sigma f$ between two manifold points is scaled by σ, while the gradient contribution from f is also scaled by σ. The following lemma shows that any PŁ constant on a neighborhood containing both the optimistic point and a higher-value manifold point must therefore be $\mathcal{O}(\sigma)$.

Lemma C.1 (No σ-uniform local PŁ constant on neighborhoods containing a higher-value manifold point)**.** 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.,

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

then

$$\mu_\sigma \;\le\; \frac{\sigma}{2} \cdot \frac{\|\nabla_x f(\theta, \bar{x})\|^2}{f(\theta, \bar{x}) - f(\theta, x^\star)}. \tag{8}$$

In particular, $\mu_\sigma = \mathcal{O}(\sigma)$ as $\sigma \downarrow 0$, so no constant bounded away from 0 can hold uniformly over $\sigma \in (0, \sigma_0]$ on any such neighborhood.

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

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

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 (8). ■

When $\mathtt{k} > 0$ and $x^\star$ is non-degenerate in the sense of Definition 3.4, such higher-value points occur arbitrarily close to $x^\star$: along any nonzero tangent direction, the restricted objective $f(\theta, \cdot)|_{\mathcal{S}(\theta)}$ increases quadratically for sufficiently small geodesic displacement. Thus the obstruction above is genuinely local in the non-singleton manifold regime.

D Proof of Section 3

This appendix provides the proofs for Section 3.

D.1 Proof of Theorem 3.2

Example D.4 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 $\mathcal{C}^1$.

The following is a standard envelope/Danskin lemma; see, e.g., Danskin [1967] and Oyama and Takenawa [2018, Prop. 2.1]. We provide the proof for completeness.

Lemma D.1 (A basic envelope/Danskin lemma)**.** Let $V \subset \mathbb{R}^m$ be an open neighborhood of θ_0, let $U \subset \mathbb{R}^\ell$ be compact, and let $\phi : V \times U \to \mathbb{R}$ be continuous. Suppose that $\phi(\cdot, u)$ is differentiable in θ for all $u \in U$, and that $(\theta, u) \mapsto \nabla_\theta \phi(\theta, u)$ is continuous on a neighborhood of (θ_0, u_0). Define $F(\theta) := \min_{u \in U} \phi(\theta, u)$ for $\theta \in V$. 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 every local choice of minimizers converges to the unique minimizer 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 $u_n \to u_0$.

We now prove Fréchet differentiability. Let $h \to 0$ in $\mathbb{R}^m$ with $\theta_0 + h \in V$, and let $u_h \in \arg\min_{u\in U} \phi(\theta_0 + h, u)$. By the first part, $u_h \to u_0$ as $h \to 0$. Set $a := \nabla_\theta \phi(\theta_0, u_0)$. Optimality of u_h and u_0 gives

$$\phi(\theta_0 + h, u_h) - \phi(\theta_0, u_h) \;\le\; F(\theta_0 + h) - F(\theta_0) \;\le\; \phi(\theta_0 + h, u_0) - \phi(\theta_0, u_0).$$

It remains to show that both the lower and upper bounds have first-order expansion $\langle a, h\rangle + o(\|h\|)$. The upper bound satisfies this immediately from differentiability of $\phi(\cdot, u_0)$ at θ_0:

$$\phi(\theta_0 + h, u_0) - \phi(\theta_0, u_0) - \langle a, h\rangle = o(\|h\|).$$

For the lower bound, take $\|h\|$ small enough so that $\{(\theta_0 + sh, u_h) : s \in [0, 1]\}$ lies in the neighborhood where $\nabla_\theta \phi$ is continuous, and write

$$\phi(\theta_0 + h, u_h) - \phi(\theta_0, u_h) - \langle a, h\rangle = \int_0^1 \big\langle \nabla_\theta \phi(\theta_0 + sh, u_h) - a,\; h\big\rangle \, \mathrm{d}s.$$

We claim that the integrand is uniformly $o(\|h\|)$. Indeed, since $u_h \to u_0$ and $\theta_0 + sh \to \theta_0$ uniformly in $s \in [0, 1]$, continuity of $\nabla_\theta \phi$ at (θ_0, u_0) gives

$$\sup_{s\in[0,1]} \big\|\nabla_\theta \phi(\theta_0 + sh, u_h) - a\big\| \to 0.$$

Consequently,

$$|\phi(\theta_0 + h, u_h) - \phi(\theta_0, u_h) - \langle a, h\rangle| \le \|h\| \sup_{s\in[0,1]} \big\|\nabla_\theta \phi(\theta_0 + sh, u_h) - a\big\| = o(\|h\|).$$

Combining the upper and lower bounds yields

$$F(\theta_0 + h) - F(\theta_0) - \langle a, h\rangle = o(\|h\|),$$

which proves differentiability of F at θ_0 with gradient $a = \nabla_\theta \phi(\theta_0, u_0)$. ∎

Lemma D.2 is the bridge from minimizing over the moving manifold $\mathcal{S}(\theta)$ to minimizing over fixed local coordinates: near x_0, points in $\mathcal{S}(\theta)$ can be written as $\psi(\theta, u)$ with u in a fixed neighborhood of $0 \in \mathbb{R}^{\mathtt{k}}$.

Lemma D.2 (Local parameter-dependent chart for the minimizer manifold)**.** Let $g \in \mathcal{C}^3$ and let Assumption 1 hold. Fix $\theta_0 \in \Theta$ and $x_0 \in \mathcal{S}(\theta_0)$. Then, after shrinking the neighborhood if needed, there exist neighborhoods $\mathcal{U}_\theta$ of θ_0, $\mathcal{U}_u$ of $0 \in \mathbb{R}^{\mathtt{k}}$, and $\mathcal{W}$ of x_0, together with a $\mathcal{C}^2$ map

$$\psi : \mathcal{U}_\theta \times \mathcal{U}_u \to \mathcal{W}, \qquad \psi(\theta_0, 0) = x_0,$$

such that $u \mapsto \psi(\theta, u)$ is an embedded chart and

$$\mathcal{S}(\theta) \cap \mathcal{W} = \{\psi(\theta, u) : u \in \mathcal{U}_u\}, \qquad \theta \in \mathcal{U}_\theta.$$

Proof. Let $\mathcal{T}_0 := \mathcal{T}_{x_0}^{\theta_0}$, $\mathcal{N}_0 := \mathcal{N}_{x_0}^{\theta_0}$, and choose orthonormal matrices $U_{\mathcal{T}_0} \in \mathbb{R}^{d\times \mathtt{k}}$ and $U_{\mathcal{N}_0} \in \mathbb{R}^{d\times(d-\mathtt{k})}$ spanning these spaces. For (u, v) near $(0, 0)$, write $x(u, v) := x_0 + U_{\mathcal{T}_0} u + U_{\mathcal{N}_0} v$ and define

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

By Proposition 2.3, $U_{\mathcal{N}_0}^\top \nabla_{xx}^2 g(\theta_0, x_0) U_{\mathcal{N}_0}$ is positive definite. The implicit function theorem therefore gives a $\mathcal{C}^2$ map $h(\theta, u)$, defined for (θ, u) near $(\theta_0, 0)$, such that $\Phi(\theta, u, h(\theta, u)) = 0$. Set

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

For each fixed θ, this parametrizes the $\mathtt{k}$-dimensional manifold

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

near x_0. This last statement is local in the following precise sense. The IFT gives open neighborhoods $\mathcal{U}_\theta^0$ of θ_0, $\mathcal{U}_u^0$ of $0 \in \mathbb{R}^{\mathtt{k}}$, and $\mathcal{U}_v^0$ of $0 \in \mathbb{R}^{d-\mathtt{k}}$ such that every solution of $U_{\mathcal{N}_0}^\top \nabla_x g(\theta, x_0 + U_{\mathcal{T}_0} u + U_{\mathcal{N}_0} v) = 0$ with $(\theta, u, v) \in \mathcal{U}_\theta^0 \times \mathcal{U}_u^0 \times \mathcal{U}_v^0$ is uniquely represented as $v = h(\theta, u)$. We now choose a connected open set $\mathcal{U}_u$ containing 0 whose closure is contained in $\mathcal{U}_u^0$, shrink the parameter neighborhood to $\mathcal{U}_\theta \subseteq \mathcal{U}_\theta^0$, and choose open coordinate balls $\mathcal{U}_u, \mathcal{U}_v$ with closures in the IFT domains so that $h(\theta, u) \in \mathcal{U}_v$ for all $(\theta, u) \in \mathcal{U}_\theta \times \mathcal{U}_u$. Set $\mathcal{W} := \{x_0 + U_{\mathcal{T}_0} u + U_{\mathcal{N}_0} v : u \in \mathcal{U}_u,\ v \in \mathcal{U}_v\}$. Then the projected-stationarity points in this fixed open cylinder are exactly the graph:

$$\widetilde{\mathcal{S}}(\theta) \cap \mathcal{W} = \{\psi(\theta, u) : u \in \mathcal{U}_u\}, \qquad \theta \in \mathcal{U}_\theta.$$

Finally, since every point of $\mathcal{S}(\theta)$ is a lower-level minimizer and hence satisfies the full first-order condition $\nabla_x g(\theta, x) = 0$, it satisfies the projected condition defining $\widetilde{\mathcal{S}}(\theta)$ and therefore $\mathcal{S}(\theta) \cap \mathcal{W} \subseteq \widetilde{\mathcal{S}}(\theta) \cap \mathcal{W}$.

It remains to show that this inclusion is equality on a smaller neighborhood. By the fixed-dimension and local Hausdorff-continuity conclusions proved in Appendix B.3, after shrinking $\mathcal{U}_\theta$ and $\mathcal{W}$, for every $\theta \in \mathcal{U}_\theta$,

$$\dim \mathcal{S}(\theta) = \mathtt{k} \qquad \text{and} \qquad \mathcal{S}(\theta) \cap \mathcal{W} \neq \emptyset.$$

For such a θ, define

$$Z_\theta := \{u \in \mathcal{U}_u : \psi(\theta, u) \in \mathcal{S}(\theta)\}.$$

The set Z_θ records which points of the graph for $\widetilde{\mathcal{S}}(\theta)$ are genuine lower-level minimizers. It is nonempty by the local Hausdorff-continuity step above: since $\mathcal{S}(\theta) \cap \mathcal{W} \neq \emptyset$ and $\mathcal{S}(\theta) \cap \mathcal{W} \subseteq \widetilde{\mathcal{S}}(\theta) \cap \mathcal{W}$, at least one point of $\mathcal{S}(\theta) \cap \mathcal{W}$ has the form $\psi(\theta, u)$. It is closed relative to $\mathcal{U}_u$ because

$$Z_\theta = \{u \in \mathcal{U}_u : \psi(\theta, u) \in \mathcal{S}(\theta)\} = \psi_\theta^{-1}(\mathcal{S}(\theta)) \cap \mathcal{U}_u, \qquad \psi_\theta(u) := \psi(\theta, u),$$

where ψ_θ is continuous and $\mathcal{S}(\theta)$ is compact, hence closed, by Proposition 2.3.

We next show that Z_θ is open. Take $u_1 \in Z_\theta$ and set $x_1 := \psi(\theta, u_1)$. Near x_1, both $\mathcal{S}(\theta)$ and $\widetilde{\mathcal{S}}(\theta)$ are embedded $\mathtt{k}$-dimensional submanifolds, and we already know that $\mathcal{S}(\theta) \subseteq \widetilde{\mathcal{S}}(\theta)$. Consider the inclusion map

$$i : \mathcal{S}(\theta) \hookrightarrow \widetilde{\mathcal{S}}(\theta), \qquad i(x) = x.$$

This map is smooth because both manifolds are embedded submanifolds of the same ambient space. Its derivative at x_1 is the identity map on tangent vectors:

$$Di(x_1) : \mathcal{T}_{x_1}\mathcal{S}(\theta) \to \mathcal{T}_{x_1}\widetilde{\mathcal{S}}(\theta), \qquad v \mapsto v.$$

Since the two tangent spaces both have dimension $\mathtt{k}$, this derivative is an isomorphism. The inverse function theorem for manifolds therefore gives a neighborhood $\mathcal{O}_S$ of x_1 in $\mathcal{S}(\theta)$ such that $i(\mathcal{O}_S)$ is an open neighborhood of x_1 in $\widetilde{\mathcal{S}}(\theta)$. But i is just inclusion, so $i(\mathcal{O}_S) = \mathcal{O}_S \subseteq \mathcal{S}(\theta)$. Thus some neighborhood of x_1 inside $\widetilde{\mathcal{S}}(\theta)$ is contained in $\mathcal{S}(\theta)$. Pulling this neighborhood back through the chart $u \mapsto \psi(\theta, u)$, we obtain a neighborhood of u_1 contained in Z_θ. Since this argument works for every $u_1 \in Z_\theta$, each point of Z_θ has a relative neighborhood in $\mathcal{U}_u$ contained in Z_θ. Thus Z_θ is open in $\mathcal{U}_u$.

Now Z_θ is nonempty and is both closed and open relative to $\mathcal{U}_u$. If $Z_\theta \neq \mathcal{U}_u$, then Z_θ and $\mathcal{U}_u \setminus Z_\theta$ would form a separation of $\mathcal{U}_u$: two nonempty disjoint relatively open sets whose union is $\mathcal{U}_u$. This contradicts the connectedness of $\mathcal{U}_u$. Hence $Z_\theta = \mathcal{U}_u$. Therefore every point of the local graph $\{\psi(\theta, u) : u \in \mathcal{U}_u\}$ belongs to $\mathcal{S}(\theta)$. Together with the earlier inclusion $\mathcal{S}(\theta) \cap \mathcal{W} \subseteq \widetilde{\mathcal{S}}(\theta) \cap \mathcal{W}$, this gives

$$\mathcal{S}(\theta) \cap \mathcal{W} = \{\psi(\theta, u) : u \in \mathcal{U}_u\},$$

after shrinking $\mathcal{W}$ within the chart neighborhood if needed. ■

Proof of Theorem 3.2. Let $x^\star(\theta)$ denote the unique optimistic minimizer for $\theta \in \mathcal{U}$. Set $x_0 := x^\star(\theta_0)$. 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. By Lemma D.2, after shrinking neighborhoods if needed there are an open neighborhood $\mathcal{U}_\theta \subseteq \mathcal{U}$, a bounded open ball $\mathcal{U}_u \subset \mathbb{R}^{\mathtt{k}}$ around 0, an open neighborhood $\mathcal{W}$ of x_0, and a $\mathcal{C}^2$ map

$$\psi : \mathcal{U}_\theta \times \mathcal{U}_u \to \mathcal{W}$$

such that $\psi(\theta_0, 0) = x_0$, each $u \mapsto \psi(\theta, u)$ is an embedded chart, and

$$\mathcal{S}(\theta) \cap \mathcal{W} = \{\psi(\theta, u) : u \in \mathcal{U}_u\}, \qquad \theta \in \mathcal{U}_\theta.$$

Step 2: Reduce to a fixed-domain minimization and show $F \in \mathcal{C}^1$. Step 1 only gives a $\mathcal{C}^2$ 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 $u^\star(\theta) \in \mathcal{U}_u$. The convergence $x^\star(\theta) \to x_0$ from Step 2a and the fact that $u \mapsto \psi(\theta, u)$ is a local chart imply $u^\star(\theta) \to 0$ as $\theta \to \theta_0$. Hence, after shrinking $\mathcal{U}_\theta$ once more, we may choose a compact neighborhood $U \subset \mathcal{U}_u$ of 0 such that $u^\star(\theta) \in U$ for all $\theta \in \mathcal{U}_\theta$. Define $\tilde{f}(\theta, u) := f(\theta, \psi(\theta, u))$. 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 $\mathcal{C}^2$, $\tilde{f}$ 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 D.1 at each $\theta \in \mathcal{U}_\theta$ yields that F is differentiable on $\mathcal{U}_\theta$ and

$$\nabla F(\theta) = \nabla_\theta \tilde{f}(\theta, u^\star(\theta)) \qquad \text{for all } \theta \in \mathcal{U}_\theta. \tag{9}$$

Importantly, (9) 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, (9) 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 (2). 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 (9) and the chain rule for $\tilde{f}(\theta, u) = f(\theta, \psi(\theta, u))$ (with u held fixed),

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

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

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

where we used $\nabla^2_{\theta x} g := (\nabla^2_{x\theta} g)^\top$ and the symmetry of $H(\theta, x^\star)^\dagger$. Substituting into the expression for $\nabla F(\theta)$ gives (2). ∎

Remark D.3 (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 (2) 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.

D.2 Why uniqueness alone is not enough for local smoothness

Under the PŁ$^\circ$ geometry in Proposition 2.3, the relevant local branch of the lower-level solution manifold is obtained in Lemma D.2; it should not be imposed as an additional assumption. The remaining obstruction is intrinsic degeneracy of the upper-level objective along that manifold. The next construction shows that degeneracy can destroy differentiability of the selected minimizer map and Lipschitz continuity of ∇F: even though the optimistic minimizer remains unique for every parameter value and $F \in \mathcal{C}^1$, the hyper-gradient need not be locally α-Hölder continuous for any $\alpha > 0$. We include a term that vanishes on the lower-level solution manifold but changes the ambient Hessian of f; this makes explicit why the non-degeneracy condition in Definition 3.4 must be stated for the restricted objective on $\mathcal{S}(\theta)$, rather than only for the ambient Hessian in tangent directions.

Example D.4 (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

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

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

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

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

The added b-term is identically zero on $\mathcal{S}(\theta)$, so it does not change the constrained minimization problem. 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, $|t^\star(\theta)| \le (\log 4)^{-1/2} < 1 \le r(\theta)$, 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 $\mathcal{C}^1$. Since the b-term vanishes on $\mathcal{S}(\theta)$, the scalar 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.

Finally, this example separates ambient and intrinsic second-order conditions. At any zero θ_k, we have $x^\star(\theta_k) = (r(\theta_k), 0)$, and the unit tangent direction is $v = (0, 1)$. The ambient tangent Hessian satisfies

$$\big\langle v, \nabla^2_{xx} f(\theta_k, x^\star(\theta_k))v \big\rangle = 2b > 0,$$

because the vanishing term $b(\|x\|^2 - r(\theta_k)^2)$ contributes $2bI$ to the ambient Hessian. However, this term is zero on $\mathcal{S}(\theta_k)$, so it contributes nothing to the restricted objective. The Riemannian Hessian of $f(\theta_k, \cdot)|_{\mathcal{S}(\theta_k)}$ in direction v is therefore 0, since the restricted scalar objective has second derivative $\phi''(0) = 0$. Thus a definition based only on the ambient tangent Hessian would incorrectly classify this example as non-degenerate, even though the intrinsic restricted problem is degenerate and ∇F has no positive Hölder modulus.

D.3 Proof of Theorem 3.5

Proof of Theorem 3.5. Let $x_0 := x^\star(\theta_0)$. We prove the result in four steps. First, we write the moving manifold $\mathcal{S}(\theta)$ in fixed local coordinates. Second, we use intrinsic non-degeneracy and the implicit function theorem to obtain a $\mathcal{C}^1$ local selected branch. Third, we use uniqueness at θ_0 to show that this local branch is the global optimistic branch for nearby parameters. Finally, we differentiate through lower-level stationarity to obtain the pseudoinverse formula and local Lipschitzness of ∇F.

Step 1: Reduce the moving-manifold problem to fixed local coordinates. By Lemma D.2, after shrinking neighborhoods if necessary, there are an open neighborhood $\mathcal{U}_\theta$ of θ_0, a bounded convex open ball $\mathcal{U}_u \subset \mathbb{R}^{\mathrm{k}}$ around 0 whose closure lies in a slightly larger coordinate domain, an open neighborhood $\mathcal{W}$ of x_0, and a $\mathcal{C}^2$ map

$$\psi : \mathcal{U}_\theta \times \mathcal{U}_u \to \mathcal{W}, \qquad \psi(\theta_0, 0) = x_0,$$

such that, for every $\theta \in \mathcal{U}_\theta$,

$$\mathcal{S}(\theta) \cap \mathcal{W} = \{\psi(\theta, u) : u \in \mathcal{U}_u\}.$$

We keep the same symbol ψ for its extension to this slightly larger coordinate domain and shrink $\mathcal{U}_u$, if needed, so that $u \mapsto \psi(\theta, u)$ is injective on a neighborhood of $\overline{\mathcal{U}_u}$ for all $\theta \in \mathcal{U}_\theta$. In these coordinates, the local restriction of the upper-level objective to $\mathcal{S}(\theta)$ is

$$\tilde{f}(\theta, u) := f(\theta, \psi(\theta, u)).$$

Since $f \in \mathcal{C}^2$ and $\psi \in \mathcal{C}^2$, we have $\tilde{f} \in \mathcal{C}^2$.

Step 2: Non-degeneracy gives a smooth local minimizer branch. At $\theta = \theta_0$, the point $u = 0$ corresponds to x_0. Since x_0 minimizes $f(\theta_0, \cdot)$ over $\mathcal{S}(\theta_0)$, $u = 0$ is a local minimizer of $\tilde{f}(\theta_0, \cdot)$, and hence

$$\nabla_u \tilde{f}(\theta_0, 0) = 0.$$

Let $A := D_u\psi(\theta_0, 0)$. Then A has full column rank and $\mathrm{range}(A) = \mathcal{T}_{x_0}^{\theta_0}$, because $u \mapsto \psi(\theta_0, u)$ is an embedded chart for $\mathcal{S}(\theta_0)$ around x_0. Because $u = 0$ is a critical point of the restricted objective, the second-derivative terms of the chart do not contribute to the Hessian quadratic form at $u = 0$. Therefore, for every $w \neq 0$, with $v := Aw$,

$$w^\top \nabla_{uu}^2 \tilde{f}(\theta_0, 0) w = \big\langle v, \mathrm{Hess}_{\mathcal{S}(\theta_0)} \bar{f}_{\theta_0}(x_0)[v] \big\rangle.$$

The non-degeneracy assumption makes the right-hand side positive for all $w \neq 0$, so $\nabla_{uu}^2 \tilde{f}(\theta_0, 0) \succ 0$. Applying the $\mathcal{C}^1$ implicit function theorem to

$$\nabla_u \tilde{f}(\theta, u) = 0$$

gives, after shrinking $\mathcal{U}_\theta$, a unique $\mathcal{C}^1$ map

$$u^\star : \mathcal{U}_\theta \to \mathcal{U}_u, \qquad u^\star(\theta_0) = 0,$$

such that $\nabla_u \tilde{f}(\theta, u^\star(\theta)) = 0$. By continuity of $\nabla_{uu}^2 \tilde{f}$, we can further shrink $\mathcal{U}_\theta$ and choose a closed ball $\mathcal{B}_u \subset \mathcal{U}_u$, with $u^\star(\theta) \in \mathrm{int}(\mathcal{B}_u)$, such that

$$\nabla_{uu}^2 \tilde{f}(\theta, u) \succ 0, \qquad \theta \in \mathcal{U}_\theta,\ u \in \mathcal{B}_u.$$

Since $\mathcal{B}_u$ is convex, $\tilde{f}(\theta, \cdot)$ is strictly convex on $\mathcal{B}_u$. Thus $u^\star(\theta)$ is the unique minimizer of $\tilde{f}(\theta, \cdot)$ on $\mathcal{B}_u$. Define

$$x_{\mathrm{loc}}(\theta) := \psi(\theta, u^\star(\theta)).$$

Then x_{loc} is $\mathcal{C}^1$, belongs to $\mathcal{S}(\theta)$, and uniquely minimizes $f(\theta, \cdot)$ over the local branch $\psi(\theta, \mathcal{B}_u)$.

Step 3: The local branch is the global optimistic branch. The previous step only controls the part of $\mathcal{S}(\theta)$ near x_0. We now rule out competitors on the rest of $\mathcal{S}(\theta)$. Choose a closed ball $\mathcal{B}_u^0$ around 0 such that $\mathcal{B}_u^0 \subset \mathrm{int}(\mathcal{B}_u)$, and then choose an open Euclidean ball $\mathcal{W}_0 := \{x : \|x - x_0\| < r\}$ such that $\overline{\mathcal{W}_0} \subset \mathcal{W}$ and

$$\mathcal{S}(\theta_0) \cap \overline{\mathcal{W}_0} \subset \psi(\theta_0, \mathrm{int}(\mathcal{B}_u^0)).$$

After shrinking $\mathcal{U}_\theta$, we may assume

$$\mathcal{S}(\theta) \cap \overline{\mathcal{W}_0} \subseteq \psi(\theta, \mathcal{B}_u), \qquad \theta \in \mathcal{U}_\theta.$$

Indeed, otherwise there would be $\theta_n \to \theta_0$ and points $z_n = \psi(\theta_n, u_n) \in \mathcal{S}(\theta_n) \cap \overline{\mathcal{W}_0}$ with $u_n \notin \mathcal{B}_u$. Passing to a subsequence, $u_n \to \bar{u} \in \overline{\mathcal{U}_u} \setminus \mathrm{int}(\mathcal{B}_u)$, and

$$z_n = \psi(\theta_n, u_n) \to \psi(\theta_0, \bar{u}) \in \mathcal{S}(\theta_0) \cap \overline{\mathcal{W}_0}.$$

This contradicts $\mathcal{S}(\theta_0) \cap \overline{\mathcal{W}_0} \subset \psi(\theta_0, \mathrm{int}(\mathcal{B}_u^0))$, the inclusion $\mathcal{B}_u^0 \subset \mathrm{int}(\mathcal{B}_u)$, and injectivity of the chart at θ_0.

Since x_0 is the unique minimizer of $f(\theta_0, \cdot)$ over the compact set $\mathcal{S}(\theta_0)$, there is a positive gap away from $\mathcal{W}_0$:

$$\Delta_0 := \min_{x \in \mathcal{S}(\theta_0) \setminus \mathcal{W}_0} \big(f(\theta_0, x) - f(\theta_0, x_0)\big) > 0,$$

with the convention that the minimum over an empty set is $+\infty$.

We claim that, after shrinking $\mathcal{U}_\theta$ once more, no optimistic minimizer at parameter $\theta \in \mathcal{U}_\theta$ lies outside $\mathcal{W}_0$. Suppose otherwise. Then there are $\theta_n \to \theta_0$ and $y_n \in \mathcal{S}(\theta_n) \setminus \mathcal{W}_0$ such that

$$f(\theta_n, y_n) \leq f(\theta_n, x_{\mathrm{loc}}(\theta_n)).$$

By Assumption 1, all lower-level minimizers lie in the fixed compact set $\mathcal{V}$. Passing to a subsequence, $y_n \to \bar{y} \in \mathcal{V}$. To see that $\bar{y} \in \mathcal{S}(\theta_0)$, fix any $z \in \mathcal{S}(\theta_0)$. Since y_n minimizes $g(\theta_n, \cdot)$, $g(\theta_n, y_n) \leq g(\theta_n, z)$. Taking $n \to \infty$ gives

$$g(\theta_0, \bar{y}) \leq g(\theta_0, z) = \min_x g(\theta_0, x),$$

so $\bar{y} \in \mathcal{S}(\theta_0)$. Also $\bar{y} \notin \mathcal{W}_0$, because $\mathcal{W}_0$ is open and $y_n \notin \mathcal{W}_0$. Taking limits in the displayed inequality and using $x_{\mathrm{loc}}(\theta_n) \to x_0$, we get

$$f(\theta_0, \bar{y}) \leq f(\theta_0, x_0),$$

which contradicts the positive gap Δ_0. Therefore every global optimistic minimizer for θ close to θ_0 lies in $\mathcal{W}_0$, hence in the charted branch $\psi(\theta, \mathcal{B}_u)$. Since $u^\star(\theta)$ is the unique minimizer on this branch, the global optimistic minimizer is unique and equals

$$x^\star(\theta) := x_{\mathrm{loc}}(\theta).$$

Thus $x^\star$ is $\mathcal{C}^1$ near θ_0, and

$$F(\theta) = f(\theta, x^\star(\theta)).$$

Step 4: Differentiate the selected branch. Because $x^\star(\theta) \in \mathcal{S}(\theta)$, lower-level stationarity gives

$$\nabla_x g(\theta, x^\star(\theta)) = 0.$$

Differentiating this identity with respect to θ yields

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

Let $H(\theta) := \nabla^2_{xx} g(\theta, x^\star(\theta))$. By Proposition 2.3,

$$\ker H(\theta) = \mathcal{T}^\theta_{x^\star(\theta)}, \qquad H(\theta)^\dagger H(\theta) = P_{\mathcal{N}^\theta_{x^\star(\theta)}}.$$

Multiplying the differentiated stationarity identity by $H(\theta)^\dagger$ gives the normal component of the branch sensitivity:

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

Since $x^\star(\theta)$ minimizes $f(\theta, \cdot)$ over $\mathcal{S}(\theta)$, the Riemannian first-order condition gives $\nabla_x f(\theta, x^\star(\theta)) \in \mathcal{N}^\theta_{x^\star(\theta)}$. Hence the tangent component of $\nabla_\theta x^\star(\theta)$ does not contribute to the chain rule:

$$\begin{aligned}\nabla F(\theta) &= \nabla_\theta f(\theta, x^\star(\theta)) + \big(\nabla_\theta x^\star(\theta)\big)^\top \nabla_x f(\theta, x^\star(\theta)) \\ &= \nabla_\theta f(\theta, x^\star(\theta)) - \nabla^2_{\theta x} g(\theta, x^\star(\theta))\, H(\theta)^\dagger \nabla_x f(\theta, x^\star(\theta)).\end{aligned}$$

This is the pseudoinverse hyper-gradient formula. Finally, $H(\theta)$ has constant rank along the branch and its nonzero eigenvalues are bounded away from zero by Proposition 2.3. Since $g \in \mathcal{C}^3$ and $x^\star \in \mathcal{C}^1$, the map $\theta \mapsto H(\theta)$ is $\mathcal{C}^1$, and the Moore–Penrose pseudoinverse is $\mathcal{C}^1$ on the constant-rank stratum. Therefore $\theta \mapsto H(\theta)^\dagger$ is $\mathcal{C}^1$ after shrinking the neighborhood if needed. The displayed formula, together with $f \in \mathcal{C}^2$ and $x^\star \in \mathcal{C}^1$, implies $\nabla F \in \mathcal{C}^1$. On a compact smaller neighborhood of θ_0, $D(\nabla F)$ is bounded, so ∇F is locally Lipschitz. ■

E Proofs for Section 4

This appendix contains the proof of the global smoothness result used in the algorithmic development of Section 4.

E.1 Proof of Proposition 4.1

Proof of Proposition 4.1. For each $\bar\theta \in \Theta$, Assumption 2 and Theorem 3.5 give an ambient open neighborhood $W_{\bar\theta} \subseteq \mathbb{R}^m$ on which the local selected branch exists and F is smooth. Choose a smaller open neighborhood $\mathcal{U}_{\bar\theta}$ of $\bar\theta$ such that $\overline{\mathcal{U}_{\bar\theta}} \cap \Theta \subset W_{\bar\theta}$, and set $V_{\bar\theta} := \mathcal{U}_{\bar\theta} \cap \Theta$. On this compact patch there is a finite constant $L_{\bar\theta}$ such that

$$\|\nabla F(\theta) - \nabla F(\theta')\| \le L_{\bar\theta} \|\theta - \theta'\|, \qquad \forall\, \theta, \theta' \in V_{\bar\theta}. \tag{10}$$

The local constant above can be obtained in the standard hyper-gradient way from the closed-form formula. On such a local patch, let $x^\star_{\bar\theta}$ denote the selected $\mathcal{C}^1$ branch and define $z(\theta) := (\theta, x^\star_{\bar\theta}(\theta))$. Write, for compactness,

$$a(z) := \nabla_\theta f(z), \qquad A(z) := \nabla^2_{\theta x} g(z), \qquad P(z) := \big[\nabla^2_{xx} g(z)\big]^\dagger, \qquad b(z) := \nabla_x f(z).$$

After shrinking the local patch if necessary, choose finite constants $\kappa_{\bar\theta}$, $B_{A,\bar\theta}, B_{P,\bar\theta}, B_{b,\bar\theta}$ and $L_{a,\bar\theta}, L_{A,\bar\theta}, L_{P,\bar\theta}, L_{b,\bar\theta}$ such that

$$\|z(\theta) - z(\theta')\| \le \kappa_{\bar\theta} \|\theta - \theta'\|,$$

and, on the local graph $\{z(\theta) : \theta \in V_{\bar\theta}\}$, the functions a, A, P, b have Lipschitz constants $L_{a,\bar\theta}, L_{A,\bar\theta}, L_{P,\bar\theta}, L_{b,\bar\theta}$, while A, P, b are bounded by $B_{A,\bar\theta}, B_{P,\bar\theta}, B_{b,\bar\theta}$. The pseudoinverse constants are finite because, along this local graph, the Hessian $\nabla^2_{xx} g$ has constant rank and its nonzero eigenvalues are bounded away from zero by the normal spectral gap in Proposition 2.3. We use the hyper-gradient formula

$$\nabla F(\theta) = a(z(\theta)) - A(z(\theta)) P(z(\theta)) b(z(\theta))$$

as follows. Fix $\theta, \theta' \in V_{\bar{\theta}}$ and abbreviate $z := z(\theta)$ and $z' := z(\theta')$. Then

$$\nabla F(\theta) - \nabla F(\theta') = a(z) - a(z') - \big(A(z)P(z)b(z) - A(z')P(z')b(z')\big),$$

and the product difference is decomposed by adding and subtracting $A(z')P(z)b(z)$ and $A(z')P(z')b(z)$:

$$\begin{aligned} A(z)P(z)b(z) - A(z')P(z')b(z') = &\big(A(z) - A(z')\big)P(z)b(z) \\ &+ A(z')\big(P(z) - P(z')\big)b(z) \\ &+ A(z')P(z')\big(b(z) - b(z')\big). \end{aligned}$$

Therefore, using the local Lipschitz and boundedness constants above,

$$\begin{aligned} \|\nabla F(\theta) - \nabla F(\theta')\| &\le \|a(z) - a(z')\| + \|A(z) - A(z')\|\,\|P(z)\|\,\|b(z)\| \\ &\quad + \|A(z')\|\,\|P(z) - P(z')\|\,\|b(z)\| \\ &\quad + \|A(z')\|\,\|P(z')\|\,\|b(z) - b(z')\| \\ &\le \Big(L_{a,\bar{\theta}} + L_{A,\bar{\theta}}B_{P,\bar{\theta}}B_{b,\bar{\theta}} + B_{A,\bar{\theta}}L_{P,\bar{\theta}}B_{b,\bar{\theta}} + B_{A,\bar{\theta}}B_{P,\bar{\theta}}L_{b,\bar{\theta}}\Big)\|z - z'\|. \end{aligned}$$

Combining this with $\|z - z'\| \le \kappa_{\bar{\theta}}\|\theta - \theta'\|$ yields the explicit local bound

$$L_{\bar{\theta}} \;:=\; \kappa_{\bar{\theta}}\Big(L_{a,\bar{\theta}} + L_{A,\bar{\theta}}B_{P,\bar{\theta}}B_{b,\bar{\theta}} + B_{A,\bar{\theta}}L_{P,\bar{\theta}}B_{b,\bar{\theta}} + B_{A,\bar{\theta}}B_{P,\bar{\theta}}L_{b,\bar{\theta}}\Big), \tag{11}$$

which is one admissible choice in (10).

The sets $\{V_{\bar{\theta}}\}_{\bar{\theta}\in\Theta}$ form an open cover of the compact metric space Θ (with the relative topology). Choose a finite subcover $V_1, \dots, V_J$, with corresponding local constants $L_1, \dots, L_J$, and define

$$L_F := \max_{1\le j\le J} L_j \quad \left(\text{equivalently, one may take } L_j = \sup_{\substack{\theta,\theta'\in V_j \\ \theta\neq\theta'}} \frac{\|\nabla F(\theta) - \nabla F(\theta')\|}{\|\theta - \theta'\|}\right). \tag{12}$$

This quantity is finite by the local smoothness bound (10).

It remains to show that the local constants globalize without any additional Lipschitz assumption on the selected branch. Let $\delta > 0$ be a Lebesgue number of the finite cover $\{V_j\}_{j=1}^{J}$ of Θ: every subset of Θ with diameter strictly smaller than δ is contained in some V_j. Fix $\theta, \theta' \in \Theta$. If $\theta = \theta'$, the claim is trivial. Otherwise, since Θ is convex, the line segment

$$\theta_s := \theta + s(\theta' - \theta), \qquad s \in [0, 1],$$

is contained in Θ. Choose an integer M such that $\|\theta - \theta'\|/M < \delta$, and set $\theta_\ell := \theta_{\ell/M}$ for $\ell = 0, \dots, M$. For each ℓ, the two-point set $\{\theta_\ell, \theta_{\ell+1}\}$ has diameter below δ, so it lies in some V_{j_ℓ}. Applying the local bound on that set gives

$$\|\nabla F(\theta_{\ell+1}) - \nabla F(\theta_\ell)\| \le L_{j_\ell}\|\theta_{\ell+1} - \theta_\ell\| \le L_F\|\theta_{\ell+1} - \theta_\ell\|.$$

Summing over $\ell = 0, \dots, M - 1$ yields

$$\|\nabla F(\theta') - \nabla F(\theta)\| \le L_F \sum_{\ell=0}^{M-1} \|\theta_{\ell+1} - \theta_\ell\| = L_F\|\theta' - \theta\|,$$

which proves the global smoothness claim. ■

F Proofs and Technical Details for Section 5

This appendix gives the complete proof pipeline behind the convergence analysis in Section 5. First, Section F.1 proves the inexact projected-gradient reduction in Proposition 5.1: once F is smooth, convergence to stationarity is controlled by the average hyper-gradient error $\mathbb{E}\|e_t\|^2$. Second, Section F.2 proves the hyper-gradient stability bound Lemma F.1. This step compares the exact pseudoinverse hyper-gradient at $x^\star(\theta_t)$ with the ridge-regularized, CG-approximated hyper-gradient

evaluated at the selected candidate $\tilde{x}_t$, separating the error into selected-point error, CG residual, ridge bias, and an off-tube moment term.

The remaining subsections bound these terms in algorithmic quantities. Section F.3 proves the selection estimate Lemma F.20: it combines curvature control for $\mathcal{S}(\theta)$, Gibbs/ULA tube concentration, a hard-selection value-gap argument, and the uniform restricted-curvature bound in Proposition B.1 to bound $\mathbb{E}\|\tilde{x}_t - x^\star(\theta_t)\|^2$. Finally, Section F.4 substitutes the stability and selection estimates into a single parameter-level bound for $\mathbb{E}\|e_t\|^2$, chooses $(N, \lambda, \varepsilon_{\mathrm{R}}, \gamma, \eta_t, R_v)$ so that this error is $\mathcal{O}(\varepsilon^2)$, and derives the conservative oracle-complexity bound reported in Section 5.

F.1 Outer-loop guarantee

Proof of Proposition 5.1. 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 $\nabla F(\theta_t) = g_t - e_t$, Cauchy–Schwarz, and the weighted Young inequality

$$uv \leq \frac{1}{4\alpha}u^2 + \alpha v^2, \qquad u, v \geq 0,$$

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

$$\sum_{t=0}^{T-1}\|\widetilde{\mathcal{G}}_t\|^2 \leq \frac{4\,(F(\theta_0) - F_\star)}{\alpha} + 4\sum_{t=0}^{T-1}\|e_t\|^2. \tag{13}$$

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 (13) 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 (4). ∎

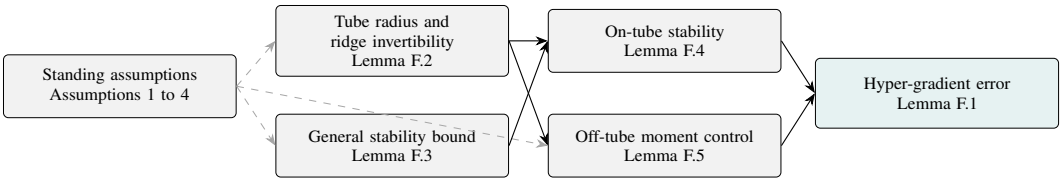


Figure 2: **Hyper-gradient error proof for Lemma F.1.** This figure summarizes the proof strategy presented in Appendix F.2: tube invertibility, deterministic stability, its on-tube specialization, and off-tube clipping control are combined to prove Lemma F.1.

F.2 Hyper-gradient stability and proof of Lemma F.1

We prove Lemma F.1 in the same order as the terms in (14). First, Lemma F.2 gives a quantitative tube radius $r(\gamma)$ on which the ridge matrix is uniformly invertible. Second, Lemma F.3 proves a deterministic stability estimate for a ridge-regularized, inexact linear solve. Third, Lemma F.4 specializes this estimate to the tube event $\mathcal{E}_t$. Fourth, Lemma F.5 controls the complement $\mathcal{E}_t^c$ using the clipping safeguard and the polynomial-growth condition on $\nabla^2_{\theta x} g$ in Assumption 4. The final proof then splits the second moment over $\mathcal{E}_t$ and $\mathcal{E}_t^c$.

Figure 2 summarizes the proof strategy and dependencies for this section.

Lemma F.1 (Second-moment hyper-gradient error bound)**.** Let Assumptions 1 to 4 hold. Fix $\gamma > 0$, define $\mathsf{D}_t := \mathrm{dist}(\tilde{x}_t, \mathcal{S}(\theta_t))$ and $\mathcal{E}_t := \{\mathsf{D}_t \le r(\gamma)\}$, where $r(\gamma)$ is the tube radius in Lemma F.2. Choose the clipping radius R_v in Algorithm 1 as in Lemma F.5. Then for all t,

$$\begin{aligned}\mathbb{E}\|e_t\|^2 \le 3C_x^2\Big(1 + \tfrac{2}{\gamma} + \tfrac{2}{\gamma(c+\gamma)}\Big)^2 \mathbb{E}\big[\|\tilde{x}_t - x^\star(\theta_t)\|^2 \mathbf{1}_{\mathcal{E}_t}\big] \\ + \ \tfrac{12C_{\mathrm{lin}}^2}{\gamma^2}\,\eta_t^2 \ + \ 3C_{\mathrm{reg}}^2\,\gamma^2 + \ C_{\mathrm{off}}(1+R_v^2)\,\mathbb{E}\big[(1+\mathsf{D}_t^{2\mathsf{n}_{\theta x}})\mathbf{1}_{\mathcal{E}_t^c}\big],\end{aligned} \tag{14}$$

where $C_x, C_{\mathrm{lin}}, C_{\mathrm{reg}} < \infty$ are stability constants depending on the global smoothness bounds and the normal spectral gap of $\nabla^2_{xx} g$ on Θ, $c > 0$ is the uniform normal spectral gap from Proposition 2.3, and $C_{\mathrm{off}} < \infty$ depends on $C_{\theta x}, D_{\theta x}, \mathsf{n}_{\theta x}$ from Assumption 4.

Lemma F.2 (Tube radius ensuring ridge invertibility)**.** Let Assumptions 1 and 3 hold. Let $\bar{L}_{g,3} := \max\{L_{g,3}, 1\}$ and define

$$r(\gamma) := \min\left\{\rho, \frac{\gamma}{2\bar{L}_{g,3}}\right\}.$$

Then for all $\theta \in \Theta$ and all $x \in \mathbb{R}^d$ satisfying $\mathrm{dist}(x, \mathcal{S}(\theta)) \le r(\gamma)$, we have

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

Proof. Let $y \in \mathcal{S}(\theta)$ be a nearest projection of x, so $\|x - y\| = \mathrm{dist}(x, \mathcal{S}(\theta)) \le r(\gamma)$. By Proposition 2.3, $\nabla^2_{xx} g(\theta, y) \succeq 0$. Since $r(\gamma) \le \rho$, the segment from y to x stays inside the ρ-tube around $\mathcal{S}(\theta)$. Assumption 3 therefore gives

$$\|\nabla^2_{xx} g(\theta, x) - \nabla^2_{xx} g(\theta, y)\| \le L_{g,3}\|x - y\| \le \bar{L}_{g,3} r(\gamma) \le \gamma/2.$$

Weyl's inequality yields $\lambda_{\min}(\nabla^2_{xx} g(\theta, x)) \ge -\gamma/2$. Thus $\nabla^2_{xx} g(\theta, x) + \gamma I \succeq (\gamma/2) I$, and the inverse-norm bound follows. ∎

Before stating the stability lemma, we clarify the notation for the selected branch and the candidate point. By Assumption 2, for every $\theta \in \Theta$ there is a unique optimistic minimizer, which we denote by $x^\star(\theta)$. For each $\bar{\theta} \in \Theta$, Theorem 3.5 gives a neighborhood of $\bar{\theta}$ on which this same selection coincides with a local $\mathcal{C}^1$ branch. On overlaps, the local branches agree by uniqueness. Thus $x^\star$ is a globally defined selection on Θ, and it is locally $\mathcal{C}^1$ around every point of Θ. In Lemma F.3, $\tilde{x}_t$ is a fixed candidate point at which the inexact implicit gradient is evaluated. The lemma is deterministic and does not use how $\tilde{x}_t$ was generated; in the ULA/Gibbs application below, $\tilde{x}_t = \hat{x}_N^\lambda(\theta_t)$, the `BoN`-selected point returned by the ULA candidate generator.

Lemma F.3 (General stability bound (technical))**.** Let Assumptions 1, 2 and 4 hold, and let $x^\star : \Theta \to \mathbb{R}^d$ denote the unique optimistic selection described above. Fix $\gamma > 0$, an iteration t, a parameter $\theta_t \in \Theta$, and a candidate point $\tilde{x}_t \in \mathbb{R}^d$. Let $\mathcal{V} \subset \mathbb{R}^d$ be any compact convex set containing $\{x^\star(\theta) : \theta \in \Theta\} \cup \{\tilde{x}_t\}$. Assume $\nabla^2_{xx} g(\theta_t, \tilde{x}_t) + \gamma I$ is invertible and satisfies

$$\left\| \left(\nabla^2_{xx} g(\theta_t, \tilde{x}_t) + \gamma I \right)^{-1} \right\| \le \kappa_\gamma.$$

Let $\tilde{v}_t$ satisfy the residual bound

$$\left\| \left(\nabla^2_{xx} g(\theta_t, \tilde{x}_t) + \gamma I \right) \tilde{v}_t - \nabla_x f(\theta_t, \tilde{x}_t) \right\| \;\le\; \eta_t.$$

Define

$$\widehat{h}_t := \nabla_\theta f(\theta_t, \tilde{x}_t) - \nabla^2_{\theta x} g(\theta_t, \tilde{x}_t) \tilde{v}_t, \qquad e_t := \widehat{h}_t - \nabla F(\theta_t).$$

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 one admissible choice of constants is

$$\begin{aligned} C_{\mathrm{lin}} &:= \sup_{(\theta, x) \in \Theta \times \mathcal{V}} \|\nabla^2_{\theta x} g(\theta, x)\|, \\ C_x &:= L_{\theta f, x} + L_{\theta x, x} \, B_{f,x} + B_{\theta x} \, L_{f_x, x} + B_{\theta x} \, L_{g_{xx}, x} \, B_{f,x}, \\ C_{\mathrm{reg}} &:= \sup_{\theta \in \Theta} \frac{\|\nabla^2_{\theta x} g(\theta, x^\star(\theta))\| \, \|\nabla_x f(\theta, x^\star(\theta))\|}{c^2}, \end{aligned}$$

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

$$B_{\theta x} := \sup_{(\theta, x) \in \Theta \times \mathcal{V}} \|\nabla^2_{\theta x} g(\theta, x)\|, \qquad B_{f,x} := \sup_{(\theta, x) \in \Theta \times \mathcal{V}} \|\nabla_x f(\theta, x)\|,$$

and $L_{\theta f, x}, L_{\theta x, x}, L_{f_x, x}, L_{g_{xx}, x} < \infty$ are uniform Lipschitz constants on $\mathcal{V}$ (uniform over $\theta \in \Theta$) for $x \mapsto \nabla_\theta f(\theta, x)$, $x \mapsto \nabla^2_{\theta x} g(\theta, x)$, $x \mapsto \nabla_x f(\theta, x)$, and $x \mapsto \nabla^2_{xx} g(\theta, x)$, respectively. Such constants exist and are finite because these maps are continuously differentiable and $\Theta \times \mathcal{V}$ is compact and convex.

Proof of Lemma F.3. Write $x^\star_t := x^\star(\theta_t)$. By Theorem 3.5 and Assumption 2,

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

Let

$$v_{t,\gamma} := \left(\nabla^2_{xx} g(\theta_t, \tilde{x}_t) + \gamma I \right)^{-1} \nabla_x f(\theta_t, \tilde{x}_t).$$

Add and subtract the ridge-regularized expressions at $\tilde{x}_t$ and $x^\star_t$, and write

$$e_t = (I) + (II) + (III),$$

where

$$\begin{aligned} (I) &:= \nabla^2_{\theta x} g(\theta_t, \tilde{x}_t) \, (v_{t,\gamma} - \tilde{v}_t), \\ (II) &:= \nabla_\theta f(\theta_t, \tilde{x}_t) - \nabla^2_{\theta x} g(\theta_t, \tilde{x}_t) v_{t,\gamma} - \nabla_\theta f(\theta_t, x^\star_t) \\ &\qquad + \nabla^2_{\theta x} g(\theta_t, x^\star_t) \left(\nabla^2_{xx} g(\theta_t, x^\star_t) + \gamma I \right)^{-1} \nabla_x f(\theta_t, x^\star_t), \end{aligned}$$

and

$$(III) := \nabla^2_{\theta x} g(\theta_t, x^\star_t) \Big(\big[\nabla^2_{xx} g(\theta_t, x^\star_t) \big]^\dagger - \left(\nabla^2_{xx} g(\theta_t, x^\star_t) + \gamma I \right)^{-1} \Big) \nabla_x f(\theta_t, x^\star_t).$$

For term (II), the $\nabla_\theta f$ difference is bounded by

$$\|\nabla_\theta f(\theta_t, \tilde{x}_t) - \nabla_\theta f(\theta_t, x^\star_t)\| \le L_{\theta f, x} \|\tilde{x}_t - x^\star_t\|.$$

For the product difference, add and subtract

$$\nabla^2_{\theta x} g(\theta_t, x^\star_t) \left(\nabla^2_{xx} g(\theta_t, \tilde{x}_t) + \gamma I \right)^{-1} \nabla_x f(\theta_t, \tilde{x}_t)$$

and

$$\nabla^2_{\theta x} g(\theta_t, x^\star_t) \left(\nabla^2_{xx} g(\theta_t, \tilde{x}_t) + \gamma I \right)^{-1} \nabla_x f(\theta_t, x^\star_t).$$

This gives

$$\begin{aligned}
&\Big\|\nabla^2_{\theta x} g(\theta_t,\tilde{x}_t)\big(\nabla^2_{xx} g(\theta_t,\tilde{x}_t)+\gamma I\big)^{-1}\nabla_x f(\theta_t,\tilde{x}_t)\\
&\qquad -\nabla^2_{\theta x} g(\theta_t,x_t^\star)\big(\nabla^2_{xx} g(\theta_t,x_t^\star)+\gamma I\big)^{-1}\nabla_x f(\theta_t,x_t^\star)\Big\|\\
&\le \|\nabla^2_{\theta x} g(\theta_t,\tilde{x}_t)-\nabla^2_{\theta x} g(\theta_t,x_t^\star)\|\,\big\|\big(\nabla^2_{xx} g(\theta_t,\tilde{x}_t)+\gamma I\big)^{-1}\big\|\,\|\nabla_x f(\theta_t,\tilde{x}_t)\|\\
&\quad + \|\nabla^2_{\theta x} g(\theta_t,x_t^\star)\|\,\big\|\big(\nabla^2_{xx} g(\theta_t,\tilde{x}_t)+\gamma I\big)^{-1}\big\|\,\|\nabla_x f(\theta_t,\tilde{x}_t)-\nabla_x f(\theta_t,x_t^\star)\|\\
&\quad + \|\nabla^2_{\theta x} g(\theta_t,x_t^\star)\|\,\Big\|\Big[\big(\nabla^2_{xx} g(\theta_t,\tilde{x}_t)+\gamma I\big)^{-1}\\
&\qquad\qquad -\big(\nabla^2_{xx} g(\theta_t,x_t^\star)+\gamma I\big)^{-1}\Big]\nabla_x f(\theta_t,x_t^\star)\Big\|.
\end{aligned}$$

The first two terms are bounded by

$$L_{\theta x,x}\kappa_\gamma B_{f,x}\|\tilde{x}_t-x_t^\star\| \quad\text{and}\quad B_{\theta x}\kappa_\gamma L_{f_x,x}\|\tilde{x}_t-x_t^\star\|.$$

For the last term, use the resolvent identity to obtain

$$\begin{aligned}
&\Big\|\Big[\big(\nabla^2_{xx} g(\theta_t,\tilde{x}_t)+\gamma I\big)^{-1}-\big(\nabla^2_{xx} g(\theta_t,x_t^\star)+\gamma I\big)^{-1}\Big]\nabla_x f(\theta_t,x_t^\star)\Big\|\\
&\qquad\le \kappa_\gamma\,\|\nabla^2_{xx} g(\theta_t,\tilde{x}_t)-\nabla^2_{xx} g(\theta_t,x_t^\star)\|\,\big\|\big(\nabla^2_{xx} g(\theta_t,x_t^\star)+\gamma I\big)^{-1}\nabla_x f(\theta_t,x_t^\star)\big\|.
\end{aligned}$$

Since $x_t^\star\in\mathcal{S}(\theta_t)$ and $\nabla_x f(\theta_t,x_t^\star)\in\mathcal{N}^{\theta_t}_{x_t^\star}$, Proposition 2.3 implies

$$\big\|\big(\nabla^2_{xx} g(\theta_t,x_t^\star)+\gamma I\big)^{-1}\nabla_x f(\theta_t,x_t^\star)\big\|\le\frac{\|\nabla_x f(\theta_t,x_t^\star)\|}{c+\gamma}\le\frac{B_{f,x}}{c+\gamma}.$$

Moreover,

$$\|\nabla^2_{xx} g(\theta_t,\tilde{x}_t)-\nabla^2_{xx} g(\theta_t,x_t^\star)\|\le L_{g_{xx},x}\|\tilde{x}_t-x_t^\star\|.$$

Therefore, the last product term is bounded by

$$B_{\theta x}\kappa_\gamma L_{g_{xx},x}B_{f,x}\,\frac{\|\tilde{x}_t-x_t^\star\|}{c+\gamma}.$$

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

$$r_t:=\big(\nabla^2_{xx} g(\theta_t,\tilde{x}_t)+\gamma I\big)\tilde{v}_t-\nabla_x f(\theta_t,\tilde{x}_t).$$

Then

$$\tilde{v}_t-v_{t,\gamma}=\big(\nabla^2_{xx} g(\theta_t,\tilde{x}_t)+\gamma I\big)^{-1}r_t,$$

and hence

$$\|(I)\|=\|\nabla^2_{\theta x} g(\theta_t,\tilde{x}_t)(\tilde{v}_t-v_{t,\gamma})\|\le\|\nabla^2_{\theta x} g(\theta_t,\tilde{x}_t)\|\,\big\|\big(\nabla^2_{xx} g(\theta_t,\tilde{x}_t)+\gamma I\big)^{-1}\big\|\,\|r_t\|\le C_{\mathrm{lin}}\,\kappa_\gamma\,\eta_t.$$

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 $\nabla_x f(\theta_t,x_t^\star)\in\mathcal{N}^{\theta_t}_{x_t^\star}$. By Proposition 2.3, the restriction of $\nabla^2_{xx} g(\theta_t,x_t^\star)$ to $\mathcal{N}^{\theta_t}_{x_t^\star}$ has eigenvalues in $[c,\infty)$. Let $\{u_i\}_{i=1}^{d-\mathtt{k}}$ be an orthonormal eigenbasis of $\mathcal{N}^{\theta_t}_{x_t^\star}$ with $\nabla^2_{xx} g(\theta_t,x_t^\star)u_i=\lambda_i u_i$ and $\lambda_i\ge c$. Since $\nabla_x f(\theta_t,x_t^\star)\in\mathcal{N}^{\theta_t}_{x_t^\star}$, we can write $\nabla_x f(\theta_t,x_t^\star)=\sum_{i=1}^{d-\mathtt{k}}\beta_i u_i$. On $\mathcal{N}^{\theta_t}_{x_t^\star}$, the pseudoinverse coincides with the inverse, hence

$$\big[\nabla^2_{xx} g(\theta_t,x_t^\star)\big]^\dagger\nabla_x f(\theta_t,x_t^\star)=\sum_{i=1}^{d-\mathtt{k}}(\beta_i/\lambda_i)u_i,$$

while

$$\big(\nabla^2_{xx} g(\theta_t,x_t^\star)+\gamma I\big)^{-1}\nabla_x f(\theta_t,x_t^\star)=\sum_{i=1}^{d-\mathtt{k}}\big(\beta_i/(\lambda_i+\gamma)\big)u_i.$$

Therefore

$$\big(\nabla^2_{xx} g(\theta_t, x_t^\star) + \gamma I\big)^{-1} \nabla_x f(\theta_t, x_t^\star) - \big[\nabla^2_{xx} g(\theta_t, x_t^\star)\big]^\dagger \nabla_x f(\theta_t, x_t^\star)$$
$$= \sum_{i=1}^{d-\mathtt{k}} \beta_i \Big(\frac{1}{\lambda_i + \gamma} - \frac{1}{\lambda_i}\Big) u_i.$$

Hence

$$\Big\| \big(\nabla^2_{xx} g(\theta_t, x_t^\star) + \gamma I\big)^{-1} \nabla_x f(\theta_t, x_t^\star) - \big[\nabla^2_{xx} g(\theta_t, x_t^\star)\big]^\dagger \nabla_x f(\theta_t, x_t^\star) \Big\|$$
$$\le \max_{i=1,\dots,d-\mathtt{k}} \Big| \frac{1}{\lambda_i + \gamma} - \frac{1}{\lambda_i} \Big| \, \|\nabla_x f(\theta_t, x_t^\star)\|.$$

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

$$\Big\| \big(\nabla^2_{xx} g(\theta_t, x_t^\star) + \gamma I\big)^{-1} \nabla_x f(\theta_t, x_t^\star) - \big[\nabla^2_{xx} g(\theta_t, x_t^\star)\big]^\dagger \nabla_x f(\theta_t, x_t^\star) \Big\|$$
$$\le \frac{\gamma}{c(c+\gamma)} \, \|\nabla_x f(\theta_t, x_t^\star)\| \le \frac{\gamma}{c^2} \, \|\nabla_x f(\theta_t, x_t^\star)\|.$$

Multiplying by $\|\nabla^2_{\theta x} g(\theta_t, x_t^\star)\|$ and taking the supremum over $\theta \in \Theta$ yields $\|(III)\| \le C_{\mathrm{reg}} \gamma$. Combining the three bounds gives the claim. ■

Lemma F.4 follows by taking a fixed compact set containing the ρ-tube around all $\mathcal{S}(\theta)$, restricting to $\mathrm{dist}(\tilde{x}_t, \mathcal{S}(\theta_t)) \le r(\gamma)$, and applying Lemma F.2 to bound $\kappa_\gamma \le 2/\gamma$.

Lemma F.4 (Stability of the inexact implicit term on the tube)**.** Assume the hypotheses of Lemma F.3 hold for the same $\theta_t, \tilde{x}_t, \tilde{v}_t$, and fix $\gamma > 0$ with tube radius $r(\gamma)$ from Lemma F.2. Let $\mathcal{V} = \mathbb{B}_d(0; \mathsf{D} + \rho)$. 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., the clipped vector used by the algorithm satisfies $v_t = \tilde{v}_t$. Then the algorithmic 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 F.3 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 F.3.

Proof. Because $r(\gamma) \le \rho$ and $\mathcal{S}(\theta_t) \subseteq \mathbb{B}_d(0; \mathsf{D})$, the tube condition implies $\tilde{x}_t \in \mathcal{V}$. Under this same condition, Lemma F.2 gives $\kappa_\gamma \le 2/\gamma$. Since clipping is inactive, the algorithmic $\widehat{h}_t$, which is formed with v_t, coincides with the quantity in Lemma F.3, which is formed with $\tilde{v}_t$. Applying Lemma F.3 with $\mathcal{V} = \mathbb{B}_d(0; \mathsf{D} + \rho)$ and substituting this inverse-norm bound gives the claim. ■

Lemma F.5 (Bounding the off-tube term under clipping)**.** Fix $\gamma > 0$ and let $r(\gamma)$ be the tube radius from Lemma F.2. Assume Algorithm 1 uses the safeguard $v_t = \mathrm{Proj}_{\mathbb{B}_d(0;R_v)}(\tilde{v}_t)$, and assume the CG output $\tilde{v}_t$ satisfies $\|b_t - (H_t + \gamma I)\tilde{v}_t\| \le \eta_t$. Let $\mathsf{D}_t := \mathrm{dist}(\tilde{x}_t, \mathcal{S}(\theta_t))$. Define the tube event $\mathcal{E}_t := \{\mathsf{D}_t \le r(\gamma)\}$. Under Assumption 4 and the smoothness of F on compact Θ, there exists a constant $C_{\mathrm{off}} < \infty$, independent of t, γ, η_t, R_v, such that

$$\|e_t\|^2 \;\le\; C_{\mathrm{off}}(1 + R_v^2)(1 + \mathsf{D}_t^{2\mathtt{n}_{\theta x}}).$$

Consequently,

$$\mathbb{E}\big[\|e_t\|^2 \mathbf{1}_{\mathcal{E}_t^c}\big] \;\le\; C_{\mathrm{off}}(1 + R_v^2) \, \mathbb{E}\big[(1 + \mathsf{D}_t^{2\mathtt{n}_{\theta x}}) \mathbf{1}_{\mathcal{E}_t^c}\big].$$

Moreover, on $\mathcal{E}_t$ Lemma F.2 implies $\|\tilde{v}_t\| \le (2/\gamma)(L_{f,1} + \eta_t)$, so choosing $R_v \ge (2/\gamma)(L_{f,1} + \eta_t)$ 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, Assumption 4 gives

$$\|\widehat{h}_t\| \le \|\nabla_\theta f(\theta_t, \tilde{x}_t)\| + \|\nabla^2_{\theta x} g(\theta_t, \tilde{x}_t)\| \, \|v_t\| \le L_{f,1} + \big(C_{\theta x} \|\tilde{x}_t\|^{\mathtt{n}_{\theta x}} + D_{\theta x}\big) R_v.$$

Since F is smooth on compact Θ, $B_F := \sup_{\theta \in \Theta} \|\nabla F(\theta)\| < \infty$. Using $(a+b+c)^2 \le 3(a^2 + b^2 + c^2)$, we obtain

$$\|e_t\|^2 \le C(1 + R_v^2)(1 + \|\tilde{x}_t\|^{2\mathtt{n}_{\theta x}})$$

for a finite constant C. Because $\mathcal{S}(\theta_t) \subseteq \mathbb{B}_d(0;\mathsf{D})$, we have $\|\tilde{x}_t\| \le \mathsf{D} + \mathsf{D}_t$, and hence $(1 + \|\tilde{x}_t\|^{2\mathsf{n}_{\theta x}}) \le C'(1 + \mathsf{D}_t^{2\mathsf{n}_{\theta x}})$. This proves the off-tube moment bound.

For the final claim, on $\mathcal{E}_t$, Lemma F.2 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 $\|r_t\| \le \eta_t$ give

$$\|\tilde{v}_t\| \le \frac{2}{\gamma}\big(\|\nabla_x f(\theta_t, \tilde{x}_t)\| + \eta_t\big) \le \frac{2}{\gamma}(L_{f,1} + \eta_t),$$

where the last inequality uses the global Lipschitzness of f. Under the stated choice of R_v, this gives $\|\tilde{v}_t\| \le R_v$, so the Euclidean projection onto $\mathbb{B}_d(0; R_v)$ leaves $\tilde{v}_t$ unchanged. ■

F.2.1 Proof of Lemma F.1

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 F.5). Therefore Lemma F.4 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 F.4 gives the first three terms in (14).

Off the tube. Lemma F.5 yields

$$\mathbb{E}\big[\|e_t\|^2 \mathbf{1}_{\mathcal{E}_t^c}\big] \le C_{\mathrm{off}}(1 + R_v^2)\,\mathbb{E}\big[(1 + \mathsf{D}_t^{2\mathsf{n}_{\theta x}})\mathbf{1}_{\mathcal{E}_t^c}\big].$$

Combining the two bounds gives (14). ■

F.3 Selection Error and Proof of Lemma F.20

We prove Lemma F.20 by reducing the selected-point error to two simpler quantities. Fix θ and let $\tilde{x}$ be the hard-selected candidate. Since $\tilde{x}$ is generally not on the minimizer manifold $\mathcal{S}(\theta)$, we let $\bar{x}$ be its Euclidean projection onto $\mathcal{S}(\theta)$. The proof then follows the decomposition

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

The first term is an *off-manifold tube error*: it is controlled by showing that Gibbs/ULA candidates remain close to $\mathcal{S}(\theta)$ with high probability. The second term is an *on-manifold selection error*: hard selection first gives a small value gap $f(\theta, \bar{x}) - F(\theta)$, and Proposition B.1 converts this value gap into squared distance on $\mathcal{S}(\theta)$. More concretely, the proof uses the following sequence:

1. establish geometric and sampling ingredients: curvature control, global quadratic growth of g, tube concentration, and Poincaré/Rényi transportation control for approximate ULA samples;
2. prove that hard selection gives a small on-manifold value gap for $\bar{x}$;
3. convert the on-manifold value gap into $\|\bar{x} - x^\star(\theta)\|^2$ using uniform quadratic growth of $f|_{\mathcal{S}(\theta)}$ and a positive away-from-minimizer gap;
4. combine the tube error and the on-manifold distance bound to obtain (43).

Figure 3 summarizes this dependency structure.

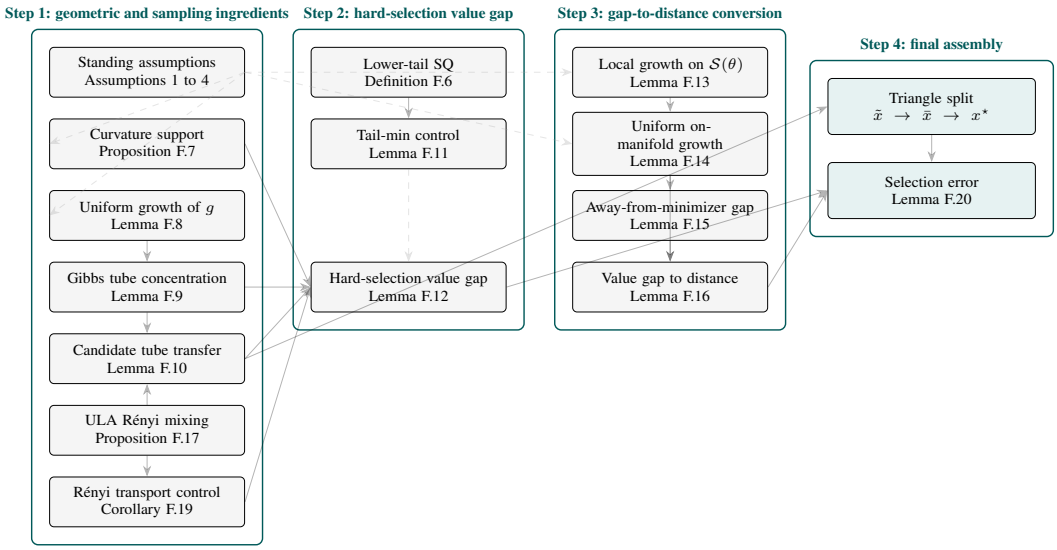


Figure 3: **Four-step selection-error proof for Lemma F.20.** Each lemma, proposition, corollary, or definition used in Appendix F.3 appears in its own box. The larger step boxes group the results according to the proof strategy stated at the beginning of Appendix F.3; dashed arrows mark supporting assumptions or intermediate deterministic superquantile steps.

Notation for the selection proof. We fix notation for the selection argument and record the lower-tail functional used in the deterministic superquantile bound inside Lemma F.12. 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 2. 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.6 (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.3.1 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]. The proof combines the manifold and normal-gap consequences of PŁ$^\circ$ from Proposition 2.3 with the analytic regularity condition in Assumption 3. Note that Assumption 3 is stronger than the $L_{g,2}$-smoothness in Assumption 4: it also controls how $\nabla^2_{xx} g(\theta, \cdot)$ varies with x (via a uniform bound on $\nabla^3_{xxx} g$ in a tube around $\mathcal{S}(\theta)$). We use this bound later in Lemma F.12, where the $M^{-1/\mathtt{k}}$ nearest-neighbor term follows from small-ball volume scaling on $\mathcal{S}(\theta)$. Thus Proposition F.7 is the step that verifies, from our assumptions on g, the curvature regularity needed to invoke the volume comparison result.

Proposition F.7 (Bounded second fundamental form from analytic regularity)**.** Let Assumptions 1 and 3 hold, and let $c > 0$ be the uniform normal spectral gap from Proposition 2.3. 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 $L_{g,3}$ and c 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 $\mathcal{S} := \mathcal{S}(\theta)$. Let $H(x) := \nabla^2_{xx} g(\theta, x)$. Fix $x \in \mathcal{S}$ and unit tangent vectors $u, v \in \mathcal{T}^\theta_x$. Choose a $\mathcal{C}^2$ curve $\gamma : (-\varepsilon, \varepsilon) \to \mathcal{S}$ 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 $U(t) \in \mathcal{T}^\theta_{\gamma(t)}$.

By Proposition 2.3, $\ker H(\gamma(t)) = \mathcal{T}^\theta_{\gamma(t)}$ for all t. Since $U(t)$ is tangent along γ, 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 $P_{\mathcal{N}^\theta_x}$ denote the orthogonal projection onto the normal space $\mathcal{N}^\theta_x$. Since $H(x)$ is symmetric and $\ker H(x) = \mathcal{T}^\theta_x$, we have $\mathrm{range}(H(x)) = (\mathcal{T}^\theta_x)^\perp = \mathcal{N}^\theta_x$ and hence $P_{\mathcal{N}^\theta_x} H(x) = H(x) P_{\mathcal{N}^\theta_x}$. Applying $P_{\mathcal{N}^\theta_x}$ to the display above yields

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

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

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

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

F.3.2 Controlling the off-manifold term $\mathrm{dist}(\tilde x, \mathcal{S}(\theta))$

Proof strategy. The off-manifold term is controlled in two steps. First, since the hard-selected point $\tilde x$ is one of the candidates, (15) reduces the problem to bounding the largest normal distance among the candidate points. Second, Lemma F.8 turns the lower-level objective gap $g(\theta, x) - g_\star(\theta)$ into a uniform quadratic lower bound in $\mathrm{dist}(x, \mathcal{S}(\theta))$. This converts the Gibbs density $\exp(-g(\theta, x)/\lambda)$ into a subgaussian tail for the distance to $\mathcal{S}(\theta)$, which Lemma F.9 integrates to control $\mathbb{E}[\max_i \mathrm{dist}(X_i, \mathcal{S}(\theta))]$ and $\mathbb{E}[\max_i \mathrm{dist}^2(X_i, \mathcal{S}(\theta))]$ for exact Gibbs candidates. Finally, Lemma F.10 transfers the same bounds to candidates with Rényi-controlled marginal laws ν_i, which is the case used for the ULA outputs.

Max-distance reduction. Because $\tilde x$ is one of the candidates, it satisfies the deterministic inequality

$$\mathrm{dist}(\tilde x, \mathcal{S}(\theta)) \;\le\; \max_{1 \le i \le M} \mathrm{dist}(X_i, \mathcal{S}(\theta)). \tag{15}$$

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

Uniform quadratic growth 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 PŁ$^\circ$, the stated tail quadratic growth, 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.8 (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 $x \in \mathbb{R}^d$,

$$g(\theta, x) - \min_{z \in \mathbb{R}^d} g(\theta, z) \;\geq\; \frac{\mu_{\mathrm{QG}}}{2} \operatorname{dist}^2\big(x, \mathcal{S}(\theta)\big). \tag{16}$$

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 $g_\star(\theta) := \min_z g(\theta, z)$.

Preliminaries on the minimizer family. Since each minimizer set is contained in $\mathbb{B}_d(0; \mathsf{D})$,

$$g_\star(\theta) = \min_{\|z\| \leq \mathsf{D}} g(\theta, z).$$

Thus $g_\star$ is continuous on the compact set Θ by Berge's maximum theorem. Consequently, the solution graph

$$\mathcal{K} := \{(\theta, y) : \theta \in \Theta,\ y \in \mathcal{S}(\theta)\} = \{(\theta, y) \in \Theta \times \mathbb{B}_d(0; \mathsf{D}) : g(\theta, y) = g_\star(\theta)\}$$

is compact. The standard local normal-chart argument for the zero set $\{\nabla_x g(\theta, x) = 0\}$, using the normal spectral gap in Proposition 2.3 around each $(\theta, y) \in \mathcal{K}$ and then patched by compactness, gives local Hausdorff continuity of $\theta \mapsto \mathcal{S}(\theta)$ on Θ. In particular, the map $(\theta, x) \mapsto \operatorname{dist}(x, \mathcal{S}(\theta))$ is continuous on $\Theta \times \mathbb{B}_d(0; \mathsf{D})$, and the normal spaces $\mathcal{N}_y^\theta$ vary continuously over $\mathcal{K}$.

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 $v \in \mathcal{N}_y^\theta$. Since $(\theta, x) \mapsto \nabla^2_{xx} g(\theta, x)$ is continuous (Assumption 4) and the normal bundle over $\mathcal{K}$ 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 $\|x - y\| \leq r_{\mathrm{loc}}$,

$$\langle v, \nabla^2_{xx} g(\theta, x) v \rangle \;\geq\; \tfrac{c}{2} \|v\|^2 \qquad \forall v \in \mathcal{N}_y^\theta. \tag{17}$$

Now fix any $\theta \in \Theta$ and $x \in \mathbb{R}^d$ with $R := \operatorname{dist}(x, \mathcal{S}(\theta)) \leq r_{\mathrm{loc}}$. 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 $\mathcal{N}_y^\theta$. 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 $\|tu\| \leq R \leq r_{\mathrm{loc}}$, (17) 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,

$$g(\theta, x) - g_\star(\theta) \geq g(\theta, x) - g(\theta, y) \geq \tfrac{c}{4} \|u\|^2 = \tfrac{c}{4} R^2. \tag{18}$$

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 $\|x\| \geq \mathsf{D}$,

$$g(\theta, x) - g_\star(\theta) \;\geq\; \tfrac{\mu_{\mathsf{qg}}}{2} \operatorname{dist}^2\big(x, \mathcal{S}(\theta)\big), \qquad \forall \theta \in \Theta. \tag{19}$$

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

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

By the continuity of the distance map just noted, $\mathcal{A}$ is closed in $\Theta \times \mathbb{B}_d(0; \mathsf{D})$, hence compact. If $\mathcal{A} = \emptyset$, the middle region is vacuous; in this case set $\mu_{\mathrm{mid}} := +\infty$. Otherwise, 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 $\operatorname{dist}(x, \mathcal{S}(\theta)) > 0$ implies $x \notin \mathcal{S}(\theta)$). In the nonempty case, therefore,

$$m_{\min} := \min_{(\theta, x) \in \mathcal{A}} \big(g(\theta, x) - g_\star(\theta)\big) > 0.$$

Moreover, if $\|x\| \le \mathsf{D}$ then $\operatorname{dist}(x, \mathcal{S}(\theta)) \le 2\mathsf{D}$, so on the nonempty set $\mathcal{A}$ we have

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

For this nonempty case, set $\mu_{\mathrm{mid}} := 2m_{\min}/(2\mathsf{D})^2$.

Step 4: combine the three regions. Let

$$\mu_{\mathrm{QG}} := \min\Big\{\tfrac{c}{2},\ \mu_{\mathtt{qg}},\ \mu_{\mathrm{mid}}\Big\}.$$

Then (18), (19), and the middle-region bound above imply (16) for all $x \in \mathbb{R}^d$ and all $\theta \in \Theta$. ■

Subgaussian tube concentration and Rényi transfer. The uniform quadratic growth bound from Lemma F.8 first gives Gaussian-like concentration of the normal distance to $\mathcal{S}(\theta)$ for exact Gibbs samples. We then transfer the same type of control to candidates whose marginal laws are close to the Gibbs law in order-2 Rényi divergence.

Lemma F.9 (Tube concentration for Gibbs samples)**.** Assume the setting of Lemma F.8, with $\mathtt{k} \ge 1$ and the common geometric bounds of Lemma B.2, 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 $\lambda_0, c_{\mathrm{tube}}, C_{\mathrm{tube}}, C_{\mathrm{tube},2} > 0$ (independent of θ and λ) with $\lambda_0 \le \min\{1, \lambda_{\mathrm{PI}}, \lambda_{\mathrm{cov}}\}$, such that, for every $0 < \lambda \le \lambda_0$,

$$\mathbb{P}(R \ge t) \;\le\; C_{\mathrm{tube}} \exp\Big(-c_{\mathrm{tube}}\, \frac{t^2}{\lambda}\Big), \qquad \forall t \ge 0. \tag{20}$$

Consequently,

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

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

$$\mathbb{E}\big[\operatorname{dist}(\tilde{x}, \mathcal{S}(\theta))\big] \;\le\; \mathbb{E}\Big[\max_{1 \le i \le M} \operatorname{dist}(X_i, \mathcal{S}(\theta))\Big] \;\le\; C_{\mathrm{tube}}\, \sqrt{\lambda\, \log(1 + M)}. \tag{21}$$

$$\mathbb{E}\big[\operatorname{dist}(\tilde{x}, \mathcal{S}(\theta))^2\big] \;\le\; \mathbb{E}\Big[\max_{1 \le i \le M} \operatorname{dist}(X_i, \mathcal{S}(\theta))^2\Big] \;\le\; C_{\mathrm{tube},2}\, \lambda\, \log(1 + M). \tag{22}$$

Proof of Lemma F.9. Fix θ and abbreviate $\mathcal{S} := \mathcal{S}(\theta)$ and $g_\star := \min_z g(\theta, z)$. By Lemma F.8, $g(\theta, x) \ge g_\star + \frac{\mu_{\mathrm{QG}}}{2} \operatorname{dist}^2(x, \mathcal{S})$. Therefore, for any $t \ge 0$,

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

Let $q := d - \mathtt{k}$ be the normal dimension of $\mathcal{S}$. Since $\mathcal{S}$ is compact and embedded in $\mathbb{R}^d$, we have $q \ge 1$. Lemma B.2 proves uniform tubular-coordinate estimates for the family $\{\mathcal{S}(\theta)\}_{\theta \in \Theta}$: there are constants $r_{\mathrm{tube}} > 0$, $0 < J_{\min} \le J_{\max} < \infty$, and $0 < V_{\min} \le V_{\max} < \infty$, independent of θ, such that the normal map

$$\Phi_\theta(y, z) = y + z, \qquad y \in \mathcal{S}, \quad z \in \mathcal{N}_y^\theta, \quad \|z\| < r_{\mathrm{tube}},$$

is a diffeomorphism onto the r_{tube}-tube around $\mathcal{S}$, its Jacobian lies in $[J_{\min}, J_{\max}]$, and the Riemannian volume of $\mathcal{S}$ lies in $[V_{\min}, V_{\max}]$.

We first lower bound the denominator. Let

$$Z_\lambda(\theta) := \int_{\mathbb{R}^d} e^{-(g(\theta,x) - g_\star)/\lambda}\, \mathrm{d}x.$$

The fiber lower bound and volume lower bound from Lemma B.2 give $c_{\mathrm{den}} = a_- V_{\min} > 0$ and

$$Z_\lambda(\theta) \ge c_{\mathrm{den}}\, \lambda^{q/2} \tag{23}$$

for all sufficiently small λ, with $c_{\mathrm{den}} > 0$ uniform in θ. The exponent is $q/2$, rather than $d/2$, because the Gibbs mass only concentrates in the $q = d - \mathtt{k}$ normal directions; integration along $\mathcal{S}$ contributes a bounded positive manifold-volume factor.

We now upper bound the numerator. Write $a := \mu_{\mathrm{QG}}/(2\lambda)$ and split the domain into the part inside the uniform tube and the part outside it.

Inside the tube. Every point in the tube has a unique representation $x = y + z$, where $y \in \mathcal{S}$ and $z \in \mathcal{N}_y^\theta$. Here $\mathrm{dVol}_{\mathcal{S}}(y)$ denotes the intrinsic $\mathtt{k}$-dimensional volume measure on $\mathcal{S}$ induced by the ambient Euclidean metric; for example, it is arclength when $\mathtt{k} = 1$ and surface area when $\mathtt{k} = 2$. Thus

$$\int_{\mathcal{S}} \mathrm{dVol}_{\mathcal{S}}(y) = \mathrm{Vol}_{\mathtt{k}}(\mathcal{S}).$$

In these coordinates, $\mathrm{dist}(y + z, \mathcal{S}) = \|z\|$ and the Euclidean volume element decomposes as

$$\mathrm{d}x = J_\theta(y, z)\, \mathrm{d}z\, \mathrm{dVol}_{\mathcal{S}}(y),$$

where $J_\theta(y, z) \le J_{\max}$. Therefore, after integrating out the y-variable using $\mathrm{Vol}_{\mathtt{k}}(\mathcal{S}) \le V_{\max}$, and identifying each normal fiber $\mathcal{N}_y^\theta$ with $\mathbb{R}^q$ through an arbitrary orthonormal basis, the tube contribution is bounded by a q-dimensional Gaussian tail:

$$\begin{aligned}
\int_{\substack{\mathrm{dist}(x,\mathcal{S}) \ge t \\ \mathrm{dist}(x,\mathcal{S}) < r_{\mathrm{tube}}}} e^{-a\,\mathrm{dist}^2(x,\mathcal{S})}\, \mathrm{d}x &\le J_{\max} V_{\max} \int_{\substack{z \in \mathbb{R}^q:\ \|z\| \ge t \\ \|z\| < r_{\mathrm{tube}}}} e^{-a\|z\|^2}\, \mathrm{d}z \\
&\le J_{\max} V_{\max}\, |\mathbb{S}^{q-1}| \int_t^\infty e^{-ar^2} r^{q-1}\, \mathrm{d}r \\
&\le C_1\, a^{-q/2} \exp\bigl(-at^2/2\bigr). \qquad (24)
\end{aligned}$$

The second inequality is just polar coordinates in the q-dimensional normal variable z: writing $z = r\omega$, with $\omega \in \mathbb{S}^{q-1}$, gives $\mathrm{d}z = r^{q-1}\mathrm{d}r\, \mathrm{d}\sigma(\omega)$, and integrating over ω gives the factor $|\mathbb{S}^{q-1}|$. We then upper bound the radial integral by replacing the upper limit r_{tube} with ∞. For the last inequality, substitute $u = \sqrt{a}\, r$ and use $\int_s^\infty u^{q-1} e^{-u^2}\, \mathrm{d}u \le C_q e^{-s^2/2}$ for all $s \ge 0$.

Outside the tube. Let $R := \max\{t, r_{\mathrm{tube}}\}$. We give a direct bound for this part. Since $\mathcal{S} \subseteq \mathbb{B}_d(0; \mathsf{D})$, every point with $\|x\| > \mathsf{D} + R$ satisfies

$$\mathrm{dist}(x, \mathcal{S}) \ge \|x\| - \mathsf{D}.$$

Hence

$$\begin{aligned}
\int_{\mathrm{dist}(x,\mathcal{S}) \ge R} e^{-a\,\mathrm{dist}^2(x,\mathcal{S})}\, \mathrm{d}x &\le \int_{\substack{\mathrm{dist}(x,\mathcal{S}) \ge R \\ \|x\| \le \mathsf{D}+R}} e^{-aR^2}\, \mathrm{d}x \;+\; \int_{\|x\| > \mathsf{D}+R} e^{-a(\|x\|-\mathsf{D})^2}\, \mathrm{d}x \\
&\le C_d(\mathsf{D} + R)^d e^{-aR^2} \;+\; C_d \int_R^\infty (\mathsf{D} + s)^{d-1} e^{-as^2}\, \mathrm{d}s,
\end{aligned}$$

where the last line uses polar coordinates with $s = \|x\| - \mathsf{D}$ for the radial tail. Because $R \ge r_{\mathrm{tube}} > 0$, the polynomial factors are dominated by the Gaussian decay. More explicitly, using the standard one-dimensional Gaussian-tail bound

$$\int_R^\infty (\mathsf{D} + s)^{d-1} e^{-as^2}\, \mathrm{d}s \le C_{d,\mathsf{D},r_{\mathrm{tube}}}\, a^{-1}(1 + R^{d-2}) e^{-aR^2}$$

for $R \ge r_{\mathrm{tube}}$, and increasing the constant to also cover the bounded-annulus term, there exists $a_0 < \infty$ such that for all $a \ge a_0$,

$$\int_{\mathrm{dist}(x,\mathcal{S}) \ge \max\{t, r_{\mathrm{tube}}\}} e^{-a\,\mathrm{dist}^2(x,\mathcal{S})}\, \mathrm{d}x \le C_5\, a^{-q/2} \exp\bigl(-at^2/2\bigr). \qquad (25)$$

Indeed, the last step follows because

$$\sup_{\substack{a \ge a_0 \\ R \ge r_{\mathrm{tube}}}} a^{q/2}(\mathsf{D} + R)^d e^{-aR^2/2} < \infty, \qquad \sup_{\substack{a \ge a_0 \\ R \ge r_{\mathrm{tube}}}} a^{q/2-1}(1 + R^{d-2}) e^{-aR^2/2} < \infty,$$

and $R \ge t$. Choose $\lambda_0 > 0$ small enough so that the denominator lower bound above holds and $a = \mu_{\mathrm{QG}}/(2\lambda) \ge a_0$ whenever $0 < \lambda \le \lambda_0$. Combining (24) and (25),

$$\int_{\mathrm{dist}(x,\mathcal{S}) \ge t} e^{-a\,\mathrm{dist}^2(x,\mathcal{S})}\, \mathrm{d}x \le C_{\mathrm{num}}\, a^{-q/2} \exp\bigl(-at^2/2\bigr) \le C'_{\mathrm{num}}\, \lambda^{q/2} \exp\Bigl(-\tfrac{\mu_{\mathrm{QG}}}{4\lambda} t^2\Bigr).$$

Dividing this numerator bound by (23) yields (20). Integrating (20) over $t \geq 0$ yields $\mathbb{E}[R] \leq C_{\mathrm{tube}}\sqrt{\lambda}$. Similarly,

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

after increasing the constant if needed.

Finally, set $R_i := \mathrm{dist}(X_i, \mathcal{S})$. Since $\tilde{x}$ is one of the candidates,

$$\mathrm{dist}(\tilde{x}, \mathcal{S}) \leq \max_{1\leq i\leq M} R_i.$$

By a union bound and (20),

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

To bound the expectation, use the tail integral representation

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

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

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

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

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

Since $M \geq 1$, $\log(1+M) \geq \log 2$, and we conclude that $\mathbb{E}[\max_i R_i] \leq C_{\mathrm{tube}}\sqrt{\lambda\log(1+M)}$ after adjusting C_{tube}. Similarly,

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

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

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

which yields (22) after increasing $C_{\mathrm{tube},2}$. ■

The next lemma is the form used for the candidates produced by the ULA/Gibbs instantiation of Algorithm 1. Those candidates are not assumed to be exact Gibbs samples: after finitely many ULA steps, the i-th candidate has marginal law ν_i. The lemma only requires each ν_i to be close to the ideal Gibbs law in order-2 Rényi divergence, and then transfers the Gibbs tube tail from Lemma F.9 by a Cauchy–Schwarz change-of-measure argument. This Rényi premise is not an additional modeling assumption for ULA candidates: Proposition F.17 below gives the required bound $R_2(\nu_i\|\mu_\theta^\lambda) \leq \varepsilon_{\mathrm{R}}^2$ once each ULA chain is run for the corresponding number of steps. Thus Lemma F.10 is the bridge from exact Gibbs concentration to the finite-time sampler used in the algorithm.

Lemma F.10 (Tube bounds for candidates with Rényi-controlled marginals)**.** Assume the setting of Lemma F.9. Let $X_1, \ldots, X_M$ be candidates whose marginal laws $\nu_1, \ldots, \nu_M$ are not necessarily identical and satisfy

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

Let $\tilde{x}$ be the hard-selection output

$$\tilde{x} \in \arg\min_{1\le i\le M} f(\theta, X_i),$$

with any measurable tie-breaking rule. Then

$$\mathbb{P}\Big(\max_{1\le i\le M} \operatorname{dist}(X_i, \mathcal{S}(\theta)) \ge t\Big) \le M\, e^{\varepsilon_{\mathrm{R}}^2/2}\sqrt{C_{\mathrm{tube}}}\exp\Big(-\tfrac{c_{\mathrm{tube}}}{2}\tfrac{t^2}{\lambda}\Big), \qquad t \ge 0. \tag{26}$$

Consequently,

$$\mathbb{E}\big[\operatorname{dist}(\tilde{x}, \mathcal{S}(\theta))\big] \;\le\; \mathbb{E}\Big[\max_{1\le i\le M} \operatorname{dist}(X_i, \mathcal{S}(\theta))\Big] \;\le\; C_{\mathrm{tube}}\, e^{\varepsilon_{\mathrm{R}}^2/2}\sqrt{\lambda\,\log(1+M)}, \tag{27}$$

and

$$\mathbb{E}\big[\operatorname{dist}(\tilde{x}, \mathcal{S}(\theta))^2\big] \;\le\; \mathbb{E}\Big[\max_{1\le i\le M} \operatorname{dist}(X_i, \mathcal{S}(\theta))^2\Big] \;\le\; C_{\mathrm{tube},2}\, e^{\varepsilon_{\mathrm{R}}^2/2}\lambda\,\log(1+M), \tag{28}$$

after increasing C_{tube} and $C_{\mathrm{tube},2}$ by universal numerical factors if necessary.

Proof of Lemma F.10. Fix θ and abbreviate $\mathcal{S} := \mathcal{S}(\theta)$ and $R_i := \operatorname{dist}(X_i, \mathcal{S})$. For any candidate index i, let $A_t := \{x : \operatorname{dist}(x, \mathcal{S}) \ge t\}$. 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 (20) gives $\mathbb{P}(R_i \ge t) \le e^{\varepsilon_{\mathrm{R}}^2/2}\sqrt{C_{\mathrm{tube}}}\exp\big(-\tfrac{c_{\mathrm{tube}}}{2}\tfrac{t^2}{\lambda}\big)$, and hence (by a union bound)

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

This proves (26). To integrate the tail, set

$$A_{\mathrm{R}} := e^{\varepsilon_{\mathrm{R}}^2/2}\sqrt{C_{\mathrm{tube}}}, \qquad \beta := \frac{c_{\mathrm{tube}}}{2\lambda}, \qquad t_0 := \sqrt{\beta^{-1}\log(1+M)}.$$

Then (26) gives

$$\mathbb{P}\Big(\max_{1\le i\le M} R_i \ge t\Big) \le \min\{1,\; MA_{\mathrm{R}}e^{-\beta t^2}\}.$$

Using the tail integral formula and splitting at t_0,

$$\begin{aligned}
\mathbb{E}\Big[\max_{1\le i\le M} R_i\Big] &\le t_0 + MA_{\mathrm{R}}\int_{t_0}^{\infty} e^{-\beta t^2}\,\mathrm{d}t \\
&\le t_0 + \frac{MA_{\mathrm{R}}}{2\beta t_0}e^{-\beta t_0^2} \\
&\le C(1+A_{\mathrm{R}})\sqrt{\lambda\log(1+M)} \le C' e^{\varepsilon_{\mathrm{R}}^2/2}\sqrt{\lambda\log(1+M)},
\end{aligned}$$

where the constants absorb C_{tube}, c_{tube}, and the fact that $\log(1+M) \ge \log 2$. After increasing C_{tube}, this proves (27).

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 gives

$$\begin{aligned}
\mathbb{E}\Big[\max_{1\le i\le M} R_i^2\Big] &\le t_0^2 + 2MA_{\mathrm{R}}\int_{t_0}^{\infty} te^{-\beta t^2}\,\mathrm{d}t \\
&= t_0^2 + \frac{MA_{\mathrm{R}}}{\beta}e^{-\beta t_0^2} \\
&\le C(1+A_{\mathrm{R}})\lambda\log(1+M) \le C' e^{\varepsilon_{\mathrm{R}}^2/2}\lambda\log(1+M).
\end{aligned}$$

After increasing $C_{\mathrm{tube},2}$, this proves (28). Finally, since the hard-selection output $\tilde{x} \in \arg\min_i f(\theta, X_i)$ is one of the candidates,

$$\operatorname{dist}(\tilde{x}, \mathcal{S}) \le \max_{1\le i\le M} R_i,$$

which gives the first inequalities in (27) and (28). ∎

F.3.3 From hard selection to a value gap on $\mathcal{S}(\theta)$

With the tube controls in place, we next bound the value gap of the projected selected point $\bar{x}$. The deterministic superquantile bound is useful for interpreting hard selection, but the final rate below uses a direct hard-min argument to avoid an unnecessary factor M.

Tail-min control 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.11 (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

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

Proof. Under the empirical law $\widehat{\nu}_M$, the random variable $f(\theta, X)$ takes values among $\{f(\theta, X_i)\}_{i=1}^M$. Let $m := \min_{1\le i\le M} f(\theta, X_i) = f(\theta, \tilde{x})$. For every $u \in (0,1)$, no value below m can have empirical probability at least u, hence $q_u(f(\theta, X)) \ge m$. Averaging this pointwise inequality over $u \in (0, \delta)$ gives the claim. ∎

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.12 (Suboptimality gap for Rényi-controlled candidates)**.** Fix θ and let $x^\star(\theta)$ be the unique minimizer of $f(\theta, \cdot)$ over $\mathcal{S}(\theta)$. Assume $\mathtt{k} \ge 1$ and that $x \mapsto f(\theta, x)$ is $L_{f,1}$-Lipschitz (as in Assumption 4). Given candidates $X_1, \dots, X_M$, let

$$\widehat{\nu}_M := \frac{1}{M}\sum_{i=1}^M \delta_{X_i}, \qquad \tilde{x} \in \arg\min_{1\le i\le M} f(\theta, X_i),$$

with any measurable tie-breaking rule, and let $\bar{x}$ be any measurable Euclidean projection of $\tilde{x}$ onto $\mathcal{S}(\theta)$. Then, for every $\delta \in (0,1)$,

$$\begin{aligned} 0 \;\le\; f(\theta, \bar{x}) - f(\theta, x^\star(\theta)) \;&\le L_{f,1}\,\mathrm{dist}(\tilde{x}, \mathcal{S}(\theta)) \\ &\quad + \Big[\mathrm{SQ}_\delta^{\mathrm{low}}\big(f(\theta, X);\, X \sim \widehat{\nu}_M\big) - f(\theta, x^\star(\theta))\Big]. \end{aligned} \tag{29}$$

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

$$\mathrm{SQ}_\delta^{\mathrm{low}}\big(f(\theta, X);\, X \sim \widehat{\nu}_M\big) - \mathrm{SQ}_\delta^{\mathrm{low}}\big(f(\theta, X);\, X \sim \nu\big) \;\le\; \frac{L_{f,1}}{\delta}\,\mathsf{W}_1(\widehat{\nu}_M, \nu), \tag{30}$$

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

$$\begin{aligned} f(\theta, \bar{x}) - f(\theta, x^\star(\theta)) \;&\le L_{f,1}\,\mathrm{dist}(\tilde{x}, \mathcal{S}(\theta)) \\ &\quad + \frac{L_{f,1}}{\delta}\,\mathsf{W}_1(\widehat{\nu}_M, \mu_\theta^\lambda) + \Big[\mathrm{SQ}_\delta^{\mathrm{low}}\big(f(\theta, X);\, X \sim \mu_\theta^\lambda\big) - f(\theta, x^\star(\theta))\Big]. \end{aligned} \tag{31}$$

Moreover, assume additionally that $0 < \lambda \le \lambda_0$, where λ_0 is chosen small enough for both the tube bound in Lemma F.9 and the projected-Gibbs small-ball bound used below, and that $X_1, \dots, X_M$ are independent candidates with marginal laws $\nu_1, \dots, \nu_M$, that μ_θ^λ satisfies a Poincaré inequality with constant at most C_{PI}, and that

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

Assume $\varepsilon_{\mathrm{R}} \le 1$. Then, for the hard-selection output $\tilde{x}$ and projection $\bar{x}$ defined above, there exists a constant $C_1 > 0$ (depending only on the on-manifold geometry near $x^\star(\theta)$; under the uniform assumptions above it can be chosen uniformly over θ) such that

$$\mathbb{E}\big[f(\theta, \bar{x}) - F(\theta)\big] \;\le\; 2L_{f,1}\,C_{\mathrm{tube}}\sqrt{\lambda \log(1+M)} \;+\; C_1\,M^{-1/\mathtt{k}} \;+\; 4\sqrt{e-1}\,L_{f,1}\sqrt{M\,C_{\mathrm{PI}}}\,\varepsilon_{\mathrm{R}}, \tag{32}$$

where C_{tube} is the tube constant from Lemma F.9.

Proof. We first prove the deterministic bounds. Since $\bar{x} \in \mathcal{S}(\theta)$ and $x^\star(\theta)$ minimizes $f(\theta, \cdot)$ over $\mathcal{S}(\theta)$, the left inequality in (29) holds. By Lipschitzness and the definition of $\bar{x}$,

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

Using Lemma F.11 to bound $f(\theta, \tilde{x})$ by the empirical lower superquantile gives (29).

To prove (30), use the variational representation

$$\mathrm{SQ}_\delta^{\mathrm{low}}(Z) = \sup_{\tau \in \mathbb{R}} \left\{ \tau - \frac{1}{\delta} \mathbb{E}[(\tau - Z)_+] \right\}.$$

For any coupling of $X \sim \widehat{\nu}_M$ and $Y \sim \nu$, the map $z \mapsto (\tau - f(\theta, z))_+$ is $L_{f,1}$-Lipschitz, and therefore, for every τ,

$$\tau - \frac{1}{\delta} \mathbb{E}[(\tau - f(\theta, X))_+] \le \tau - \frac{1}{\delta} \mathbb{E}[(\tau - f(\theta, Y))_+] + \frac{L_{f,1}}{\delta} \mathbb{E}\|X - Y\|.$$

Taking the supremum over τ and then the infimum over couplings gives (30). Combining (29) and (30) with $\nu = \mu_\theta^\lambda$ gives (31).

The deterministic bound (31) 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 $\bar{x}$,

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

Taking expectations gives

$$\mathbb{E}\big[f(\theta, \bar{x}) - F(\theta)\big] \;\le\; L_{f,1}\, \mathbb{E}\big[\operatorname{dist}(\tilde{x}, \mathcal{S}(\theta))\big] + \mathbb{E}\Big[\min_{1 \le i \le M} f(\theta, X_i)\Big] - F(\theta). \tag{33}$$

Step 1: hard-min bound under exact 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 (33) and using $\operatorname{dist}(\tilde{x}, \mathcal{S}(\theta)) \le \max_i \operatorname{dist}(X_i, \mathcal{S}(\theta))$ yields

$$\mathbb{E}\big[f(\theta, \bar{x}) - F(\theta)\big] \;\le\; 2L_{f,1}\, \mathbb{E}\Big[\max_{1 \le i \le M} \operatorname{dist}(X_i, \mathcal{S}(\theta))\Big] + \Big(\mathbb{E}\Big[\min_{1 \le i \le M} f(\theta, \Pi(X_i))\Big] - F(\theta)\Big). \tag{34}$$

The tube bound in Lemma F.9 gives

$$\mathbb{E}\Big[\max_{1 \le i \le M} \operatorname{dist}(X_i, \mathcal{S}(\theta))\Big] \le C_{\mathrm{tube}} \sqrt{\lambda \log(1 + M)}.$$

For the on-manifold term, apply Lemma B.2 directly to the finite-temperature projected Gibbs law. After restricting $\lambda_0 \le \lambda_{\mathrm{cov}}$, its common constants $c_0, r_{\mathrm{nn}} > 0$ give, uniformly over θ and $0 < \lambda \le \lambda_0$,

$$\mathbb{P}\big(d_{\mathcal{S}(\theta)}(\Pi(X_1), x^\star(\theta)) \le r\big) \ge c_0 r^{\mathtt{k}}, \qquad 0 < r \le r_{\mathrm{nn}}.$$

Let $D_M := \min_{1 \le i \le M} d_{\mathcal{S}(\theta)}(\Pi(X_i), x^\star(\theta))$. The small-ball bound gives

$$\mathbb{P}(D_M > r) \le (1 - c_0 r^{\mathtt{k}})^M \le \exp(-M c_0 r^{\mathtt{k}}), \qquad 0 < r \le r_{\mathrm{nn}},$$

Let $\operatorname{diam}_{\mathcal{S}}$ denote the finite geodesic diameter of $\mathcal{S}(\theta)$. It is uniformly bounded by D_{geo} from Lemma B.2. Integrating this tail bound gives

$$\begin{aligned}
\mathbb{E}[D_M] &= \int_0^{\operatorname{diam}_{\mathcal{S}}} \mathbb{P}(D_M > r)\, \mathrm{d}r \\
&\le \int_0^{r_{\mathrm{nn}}} \exp(-M c_0 r^{\mathtt{k}})\, \mathrm{d}r + \operatorname{diam}_{\mathcal{S}} \exp(-M c_0 r_{\mathrm{nn}}^{\mathtt{k}}) \\
&\le C_{\mathrm{nn}} M^{-1/\mathtt{k}},
\end{aligned}$$

where the last line uses $\mathtt{k} \geq 1$ and absorbs the exponentially small term into the same polynomial rate for all $M \geq 1$. Since $f(\theta, \cdot)$ is $L_{f,1}$-Lipschitz on $\mathcal{S}(\theta)$,

$$\mathbb{E}\Big[\min_{1\leq i\leq M} f(\theta, \Pi(X_i))\Big] - F(\theta) \leq L_{f,1}\mathbb{E}[D_M] \leq C_1 M^{-1/\mathtt{k}}.$$

Moreover, the display preceding (34) also gives

$$\begin{aligned} &L_{f,1}\,\mathbb{E}\Big[\max_{1\leq i\leq M} \operatorname{dist}(X_i, \mathcal{S}(\theta))\Big] + \mathbb{E}\Big[\min_{1\leq i\leq M} f(\theta, X_i)\Big] - F(\theta) \\ &\quad \leq 2L_{f,1}\,\mathbb{E}\Big[\max_{1\leq i\leq M} \operatorname{dist}(X_i, \mathcal{S}(\theta))\Big] + \mathbb{E}\Big[\min_{1\leq i\leq M} f(\theta, \Pi(X_i))\Big] - F(\theta). \end{aligned}$$

Substituting the tube and nearest-neighbor bounds into this display gives the exact-Gibbs reference bound

$$\begin{aligned} &L_{f,1}\,\mathbb{E}\Big[\max_{1\leq i\leq M} \operatorname{dist}(X_i, \mathcal{S}(\theta))\Big] \\ &\qquad + \mathbb{E}\Big[\min_{1\leq i\leq M} f(\theta, X_i)\Big] - F(\theta) \\ &\leq 2L_{f,1}\,C_{\mathrm{tube}}\,\sqrt{\lambda \log(1+M)} + \, C_1\, M^{-1/\mathtt{k}}. \end{aligned} \tag{35}$$

Step 2: transfer to Rényi-controlled candidates. We now return to independent candidates $X_1, \ldots, X_M$ with marginal laws ν_i satisfying $R_2(\nu_i \| \mu_\theta^\lambda) \leq \varepsilon_{\mathrm{R}}^2$. For each i, choose an optimal coupling of $X_i \sim \nu_i$ with a reference variable $X_i^\lambda \sim \mu_\theta^\lambda$ so that

$$\mathbb{E}\|X_i - X_i^\lambda\|^2 \leq \mathsf{W}_2(\nu_i, \mu_\theta^\lambda)^2.$$

Taking the product of these pairwise couplings gives a joint coupling whose first marginal is $\nu_1 \otimes \cdots \otimes \nu_M$, the law of the actual independent candidates, and whose second marginal is $(\mu_\theta^\lambda)^{\otimes M}$. Thus the reference variables $X_1^\lambda, \ldots, X_M^\lambda$ are i.i.d. from μ_θ^λ, and expectations of functions of $X_1, \ldots, X_M$ are unchanged under this coupling. We use the two deterministic comparisons

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

and

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

and combine them with (33), applied to the actual candidates, to obtain

$$\begin{aligned} \mathbb{E}\big[f(\theta, \bar{x}) - F(\theta)\big] \leq L_{f,1}\,\mathbb{E}\Big[\max_{1\leq i\leq M} \operatorname{dist}(X_i^\lambda, \mathcal{S}(\theta))\Big] + \mathbb{E}\Big[\min_{1\leq i\leq M} f(\theta, X_i^\lambda)\Big] - F(\theta) \\ + 2L_{f,1}\,\mathbb{E}\Big[\max_{1\leq i\leq M} \|X_i - X_i^\lambda\|\Big]. \end{aligned}$$

The first two terms on the right are controlled by the exact-Gibbs reference bound (35). It remains to bound the coupling error

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

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

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

and the Poincaré/Rényi transportation inequality [Liu, 2020] gives

$$\mathsf{W}_2(\nu_i, \mu_\theta^\lambda) \leq 2\,C_{\mathrm{PI}}^{1/2}\big(e^{R_2(\nu_i\|\mu_\theta^\lambda)} - 1\big)^{1/2} \leq 2\sqrt{(e-1)C_{\mathrm{PI}}}\,\varepsilon_{\mathrm{R}}.$$

The last inequality uses $\varepsilon_{\mathrm{R}} \leq 1$, so $e^{\varepsilon_{\mathrm{R}}^2} - 1 \leq (e-1)\varepsilon_{\mathrm{R}}^2$. Therefore

$$2L_{f,1}\,\mathbb{E}\Big[\max_{1\leq i\leq M} \|X_i - X_i^\lambda\|\Big] \leq 4\sqrt{e-1}\,L_{f,1}\sqrt{M\,C_{\mathrm{PI}}}\,\varepsilon_{\mathrm{R}}.$$

Combining this estimate with (35) proves (32). ■

F.3.4 Converting the on-manifold gap to distance (link to local identifiability)

Quadratic growth on $\mathcal{S}(\theta)$ from Riemannian non-degeneracy. 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 is the local deterministic step: positive Riemannian Hessian of the restricted objective implies quadratic growth along the manifold.

Lemma F.13 (Local quadratic growth from Riemannian non-degeneracy)**.** Fix θ and suppose that $f(\theta, \cdot)$ is $\mathcal{C}^2$ in a neighborhood of $\mathcal{S}(\theta)$ and that $x^\star(\theta) \in \mathcal{S}(\theta)$ is a local minimizer of $f(\theta, \cdot)|_{\mathcal{S}(\theta)}$. Assume moreover that the Riemannian Hessian of the restricted objective is positive definite at $x^\star(\theta)$, i.e., there exists $m_\theta > 0$ such that

$$\big\langle v, \mathrm{Hess}_{\mathcal{S}(\theta)} \bar{f}_\theta(x^\star(\theta))[v] \big\rangle \;\geq\; m_\theta \|v\|^2, \qquad \forall v \in \mathcal{T}^\theta_{x^\star(\theta)},$$

where $\bar{f}_\theta := f(\theta, \cdot)|_{\mathcal{S}(\theta)}$. Then there exist constants $c_{\mathrm{hg}} > 0$ and $r_0 > 0$ such that for all $x \in \mathcal{S}(\theta)$ with $d_{\mathcal{S}(\theta)}(x, x^\star(\theta)) \leq r_0$,

$$f(\theta, x) - f(\theta, x^\star(\theta)) \;\geq\; c_{\mathrm{hg}}\, d_{\mathcal{S}(\theta)}(x, x^\star(\theta))^2. \tag{36}$$

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 the local-minimizer assumption, $x^\star := x^\star(\theta)$ 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 assumed Riemannian non-degeneracy gives positive curvature of the restricted objective in every tangent direction. Define the minimal Riemannian Hessian 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 $w \in \mathcal{T}_y\mathcal{M}$,

$$\big\langle w,\, \mathrm{Hess}_{\mathcal{M}} \bar{f}(y)[w] \big\rangle \;\geq\; \tfrac{m}{2}\, \|w\|^2. \tag{37}$$

Let $r_0 > 0$ be such that the closed geodesic ball $\{y \in \mathcal{M} : d_{\mathcal{M}}(y, x^\star) \leq r_0\}$ is contained in $\mathcal{U}$.

Fix any $x \in \mathcal{M}$ with $d_{\mathcal{M}}(x, x^\star) \leq r_0$. 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, each subsegment is minimizing, and hence $d_{\mathcal{M}}(\gamma(t), x^\star) \leq d_{\mathcal{M}}(x, x^\star) \leq r_0$ for all $t \in [0, 1]$. Thus $\gamma([0, 1]) \subset \mathcal{U}$.

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 (37) and constant speed,

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

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

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

Recalling $\phi(1) = \bar{f}(x)$ and $\phi(0) = \bar{f}(x^\star)$ gives (36) with $c_{\mathrm{hg}} := m/4$. ∎

Uniform version over Θ. The restricted-curvature constant from Proposition B.1 makes the quadratic-growth constants uniform over Θ.

Lemma F.14 (Uniform quadratic growth on $\mathcal{S}(\theta)$ over Θ)**.** Assume the setting of Section 5 and let $m_{\mathrm{R}} > 0$ be the uniform restricted-curvature constant from Proposition B.1. Let $x^\star : \Theta \to \mathbb{R}^d$ be the global selected minimizer from Assumption 2. 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 $d_{\mathcal{S}(\theta)}(x, x^\star(\theta)) \leq r_0$,

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

Proof. If $\mathtt{k} = 0$, then each connected compact zero-dimensional manifold $\mathcal{S}(\theta)$ is a singleton. Hence the only point $x \in \mathcal{S}(\theta)$ is $x^\star(\theta)$, and the claim is trivial for any fixed positive c_{hg} and r_0. We therefore assume $\mathtt{k} \geq 1$.

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 Assumption 2, $x^\star(\theta)$ is the unique minimizer of $\bar{f}_\theta$ on $\mathcal{M}_\theta$. By Proposition B.1, for every $\theta \in \Theta$,

$$\big\langle v,\ \mathrm{Hess}_{\mathcal{M}_\theta} \bar{f}_\theta(x^\star(\theta))[v] \big\rangle \geq m_{\mathrm{R}} \|v\|^2, \qquad \forall v \in \mathcal{T}_{x^\star(\theta)} \mathcal{M}_\theta.$$

We next make the neighborhood on which this lower bound persists uniform. For each $\bar{\theta} \in \Theta$, continuity of the Riemannian Hessian in local charts gives a parameter neighborhood $V_{\bar{\theta}}$ and a radius $r_{\bar{\theta}} > 0$ such that, for every $\theta \in V_{\bar{\theta}}$, every $y \in \mathcal{M}_\theta$ with $d_{\mathcal{M}_\theta}(y, x^\star(\theta)) \leq r_{\bar{\theta}}$, and every $w \in \mathcal{T}_y \mathcal{M}_\theta$,

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

Choose a finite subcover $V_{\bar{\theta}_1}, \ldots, V_{\bar{\theta}_J}$ of Θ and set

$$r_0 := \min_{1 \leq j \leq J} r_{\bar{\theta}_j} > 0.$$

Then, for all $\theta \in \Theta$, all $y \in \mathcal{M}_\theta$ with $d_{\mathcal{M}_\theta}(y, x^\star(\theta)) \leq r_0$, and all $w \in \mathcal{T}_y \mathcal{M}_\theta$,

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

Fix any $\theta \in \Theta$ and $x \in \mathcal{M}_\theta$ with $d_{\mathcal{M}_\theta}(x, x^\star(\theta)) \leq r_0$. 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 $t \in [0, 1]$. Since γ is minimizing, $d_{\mathcal{M}_\theta}(\gamma(t), x^\star(\theta)) \leq r_0$ for all $t \in [0, 1]$, so the uniform Hessian lower bound above applies along the whole curve. As in the proof of Lemma F.13, define $\phi(t) := \bar{f}_\theta(\gamma(t))$. Then $\phi'(0) = 0$ and

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

so the same integral argument yields

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

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

Removing the locality restriction. The quadratic growth bound from Lemma F.14 only applies within a geodesic neighborhood of radius r_0 around $x^\star(\theta)$. To remove the explicit restriction $d_{\mathcal{S}(\theta)}(\bar{x}, x^\star(\theta)) \leq r_0$, we use a value gap away from this neighborhood.

Lemma F.15 (Uniform value gap away from the selected minimizer)**.** Let (c_{hg}, r_0) be the constants from Lemma F.14 and define

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

with the convention that empty sets are ignored in the infimum. If all $\mathcal{S}_{r_0}(\theta)$ are empty, choose any fixed positive value for Δ_{r_0}. Then $\Delta_{r_0} > 0$.

Proof. If all $\mathcal{S}_{r_0}(\theta)$ are empty, the claim is immediate by convention. Otherwise, consider

$$\mathcal{A}_{r_0} := \{(\theta, x) : \theta \in \Theta,\ x \in \mathcal{S}(\theta),\ d_{\mathcal{S}(\theta)}(x, x^\star(\theta)) \geq r_0\}.$$

Under the uniform regularity assumptions used in Lemma F.14, the compact manifolds $\mathcal{S}(\theta)$ vary continuously in local charts over compact Θ; hence their graph is compact and the geodesic distance to the continuous selected branch $x^\star(\theta)$ is continuous on this graph. Therefore $\mathcal{A}_{r_0}$ is compact. It is also disjoint from the selected graph $\{(\theta, x^\star(\theta)) : \theta \in \Theta\}$.

The gap map $(\theta, x) \mapsto f(\theta, x) - F(\theta)$ is continuous and strictly positive on $\mathcal{A}_{r_0}$: if it vanished at some $(\theta, x) \in \mathcal{A}_{r_0}$, then x would be another minimizer of $f(\theta, \cdot)$ over $\mathcal{S}(\theta)$, contradicting the uniqueness in Assumption 2. A positive continuous function on a compact set has a positive minimum, which is exactly Δ_{r_0}. ■

Lemma F.16 (Bounding $\mathbb{E}\|\bar{x} - x^\star(\theta)\|^2$ without a locality condition)**.** Let (c_{hg}, r_0) be the constants from Lemma F.14, and let $\Delta_{r_0} > 0$ be the uniform away-gap from Lemma F.15. 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.14 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. ∎

F.3.5 Sampler accuracy from Rényi divergence

The hard-selection value-gap bound above is stated directly for candidates whose laws are Rényi-close to the ideal Gibbs law. The next statements explain how finite-step ULA supplies this Rényi accuracy and recall the Poincaré/Rényi transportation control used in the proof of Lemma F.12.

We use the Poincaré branch of Chewi [2026, Theorem 6.2.9]. The following joint choice makes the initialization order and time-discretization requirements explicit.

Proposition F.17 (ULA convergence in order-2 Rényi divergence)**.** For $\pi(\mathrm{d}x) \propto e^{-G(x)}\mathrm{d}x$, suppose ∇G is $L_{G,2}$-Lipschitz and π has Poincaré constant C_{PI}. Enlarge both bounds to at least one. Let $R_4 := \mathcal{R}_4(\widehat{\mu}_0\|\pi) < \infty$, where $\mathcal{R}_q$ is order-q Rényi divergence. There are universal $C_T, c_h > 0$ and $\varepsilon_{\mathrm{src}} \in (0, 1/2]$ such that, for $0 < \varepsilon \le \varepsilon_{\mathrm{src}}$, set

$$T = C_T C_{\mathrm{PI}}[R_4 + \log(1/\varepsilon)], \qquad h_0 = c_h \min\left\{\frac{1}{4L_{G,2}^2 T}, \frac{1}{2L_{G,2}^{3/2}\sqrt{T}}, \frac{\varepsilon^2}{2L_{G,2}^2 dT}\right\}.$$

Choose $n = \lceil T/h_0 \rceil$ and $h = T/n$. The ULA recursion $X^{j+1} = X^j - h\nabla G(X^j) + \sqrt{2h}\xi^j$, with independent standard Gaussian ξ^j, satisfies $\mathcal{R}_2(\widehat{\mu}_n\|\pi) \le \varepsilon^2$, and

$$n \le C_{\mathrm{Ch}} C_{\mathrm{PI}}^2 L_{G,2}^2 d\varepsilon^{-2}[R_4^2 + \log^2(1/\varepsilon)] \tag{38}$$

for a universal C_{Ch}. In particular, the step size and iteration count must be chosen together; this is not a guarantee for arbitrary longer runs at a fixed step size. Moreover,

$$\mathsf{W}_2(\widehat{\mu}_n, \pi) \le 2\sqrt{C_{\mathrm{PI}}}\sqrt{e^{\mathcal{R}_2(\widehat{\mu}_n\|\pi)} - 1} \le 2\sqrt{(e-1)C_{\mathrm{PI}}}\,\varepsilon. \tag{39}$$

Proof. At target order two, the cited diffusion estimate uses the order-three initial divergence, bounded by R_4. Its discretization argument uses order four. With C_T large, the stated T supplies the required diffusion time. The first two restrictions on h_0 are those of the discretization estimate. Writing $\beta = L_{G,2}$, its remaining error is controlled by

$$\beta^3 h^2 T(d + R_4) + \beta^2 dhT \le \frac{c_h^2\varepsilon^4}{4}\frac{d + R_4}{\beta d^2 T} + \frac{c_h\varepsilon^2}{2} \le \frac{c_h^2\varepsilon^4}{4C_T}(1 + 1/\log 2) + \frac{c_h\varepsilon^2}{2}.$$

Small universal choices of $c_h, \varepsilon_{\mathrm{src}}$ give the accuracy claim. Since $T, \beta, d \ge 1$,

$$n \le 1 + c_h^{-1}\max\{4\beta^2T^2,\, 2\beta^{3/2}T^{3/2},\, 2\beta^2 dT^2\varepsilon^{-2}\} \le C\beta^2 dT^2\varepsilon^{-2},$$

which proves the count. The transportation consequence of Poincaré [Liu, 2020] gives the first inequality in (39); convexity of e^s on $[0, 1]$ gives $e^s - 1 \le (e-1)s$ for the second. ∎

Lemma F.18 (Order-4 Rényi warm start for λ-scaled Gaussian initialization)**.** Assume the setting of Lemma F.8, and let $g_\star(\theta) := \min_x g(\theta, x)$. Fix a compact set $\mathcal{Z} \subset \mathbb{R}^d$ and a variance parameter

$$0 < \tau < \frac{4}{3L_{g,2}}.$$

For $z \in \mathcal{Z}$, let

$$\widehat{\mu}_{0,z}^{\lambda} := \mathcal{N}(z, \tau\lambda I_d), \qquad 0 < \lambda \le 1.$$

Then there are constants $C_{\mathrm{init}}, C'_{\mathrm{init}} < \infty$, independent of θ, z, λ, such that

$$\mathcal{R}_4\big(\widehat{\mu}_{0,z}^{\lambda} \big\| \mu_\theta^{\lambda}\big) \le \frac{C_{\mathrm{init}}}{\lambda} \; + \; C'_{\mathrm{init}}\big(1 + \log(1/\lambda)\big). \tag{40}$$

Moreover, if $z \in \mathcal{S}(\theta)$, then the stronger bound

$$\mathcal{R}_4\big(\widehat{\mu}_{0,z}^{\lambda} \big\| \mu_\theta^{\lambda}\big) \le C'_{\mathrm{init}}\big(1 + \log(1/\lambda)\big) \tag{41}$$

holds after increasing C'_{init} if necessary.

Proof. We track the Gaussian-center dependence directly. Write $p_{z,\lambda}$ for the density of $\widehat{\mu}_{0,z}^{\lambda}$, and write

$$Z_\theta^{\lambda} := \int_{\mathbb{R}^d} \exp\{-g(\theta, x)/\lambda\}\, \mathrm{d}x$$

for the Gibbs normalizing constant. Since $\mu_\theta^{\lambda}(\mathrm{d}x) = (Z_\theta^{\lambda})^{-1} e^{-g(\theta,x)/\lambda}\mathrm{d}x$, the order-4 Rényi divergence is

$$\mathcal{R}_4\big(\widehat{\mu}_{0,z}^{\lambda} \big\| \mu_\theta^{\lambda}\big) = \frac{1}{3}\log\left[(Z_\theta^{\lambda})^3 \int_{\mathbb{R}^d} p_{z,\lambda}(x)^4 \exp\{3g(\theta, x)/\lambda\}\, \mathrm{d}x\right].$$

First, by Lemma F.8 and $\mathcal{S}(\theta) \subseteq \overline{\mathbb{B}}_d(0; \mathsf{D})$,

$$Z_\theta^{\lambda} \le e^{-g_\star(\theta)/\lambda} \int_{\mathbb{R}^d} \exp\Big\{-\frac{\mu_{\mathrm{QG}}}{2\lambda}\mathrm{dist}(x, \mathcal{S}(\theta))^2\Big\}\, \mathrm{d}x \le C_Z e^{-g_\star(\theta)/\lambda}$$

for all $0 < \lambda \le 1$, with C_Z uniform in θ. The last inequality follows because $\mathrm{dist}(x, \mathcal{S}(\theta)) \ge (\|x\| - \mathsf{D})_+$, and the resulting Gaussian tail is integrable uniformly for $\lambda \le 1$.

Next, set $u = x - z$. By $L_{g,2}$-smoothness in x,

$$g(\theta, z + u) \le g(\theta, z) + \langle \nabla_x g(\theta, z), u\rangle + \frac{L_{g,2}}{2}\|u\|^2.$$

Since

$$p_{z,\lambda}(z + u)^4 = (2\pi\tau\lambda)^{-2d} \exp\Big\{-\frac{2}{\tau\lambda}\|u\|^2\Big\},$$

and $a := 2/\tau - 3L_{g,2}/2 > 0$, Gaussian integration gives

$$\int_{\mathbb{R}^d} p_{z,\lambda}(x)^4 \exp\{3g(\theta, x)/\lambda\}\, \mathrm{d}x \le C\, \lambda^{-3d/2} \exp\left\{\frac{3g(\theta, z)}{\lambda} + \frac{9\|\nabla_x g(\theta, z)\|^2}{4a\lambda}\right\}.$$

Combining the two displays yields

$$\mathcal{R}_4\big(\widehat{\mu}_{0,z}^{\lambda} \big\| \mu_\theta^{\lambda}\big) \le \frac{g(\theta, z) - g_\star(\theta)}{\lambda} \; + \; \frac{3\|\nabla_x g(\theta, z)\|^2}{4a\lambda} \; + \; \frac{d}{2}\log(1/\lambda) + C.$$

The first two terms are uniformly bounded by $C_{\mathrm{init}}/\lambda$, because $\Theta \times \mathcal{Z}$ is compact and $g_\star$ is continuous on Θ. This proves (40). If $z \in \mathcal{S}(\theta)$, then $g(\theta, z) = g_\star(\theta)$ and $\nabla_x g(\theta, z) = 0$, so only the logarithmic term remains, proving (41). ∎

Corollary F.19 (Specialization to Gibbs sampling (Rényi accuracy implies Wasserstein control))**.** Assume the hypotheses of Proposition B.3. For $0 < \lambda \le \min\{1, \lambda_{\mathrm{PI}}\}$, use its common C_{PI} and apply Proposition F.17 with $G = g(\theta, \cdot)/\lambda$, $\pi = \mu_\theta^{\lambda}$. Taking $L_{g,2} \ge 1$, we have $L_{G,2} = L_{g,2}/\lambda \ge 1$. The ULA update

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

corresponds to the same ULA scheme with stepsize $h' = \lambda h$ for G. For $0 < \varepsilon_{\mathrm{R}} \le \varepsilon_{\mathrm{src}}$, the joint step-size and iteration-count choice of Proposition F.17 has sufficient count

$$n = \widetilde{\mathcal{O}}\big(d\, C_{\mathrm{PI}}^2 \lambda^{-2} \varepsilon_{\mathrm{R}}^{-2} \big(1 + \mathcal{R}_4(\widehat{\mu}_0 \| \mu_\theta^\lambda)^2\big)\big)$$

and ensures $R_2(\widehat{\mu}_n \| \mu_\theta^\lambda) \le \varepsilon_{\mathrm{R}}^2$. If, moreover, $\widehat{\mu}_0 = \widehat{\mu}_{0,z}^\lambda$ is the λ-scaled Gaussian initialization in Lemma F.18 with $z \in \mathcal{Z}$, then it suffices to run

$$n \;=\; \widetilde{\mathcal{O}}\big(d\, C_{\mathrm{PI}}^2\, \lambda^{-4}\, \varepsilon_{\mathrm{R}}^{-2}\big) \tag{42}$$

iterations. If the Gaussian center satisfies $z \in \mathcal{S}(\theta)$, then the sharper logarithmic bound (41) gives the improved sufficient count

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

In all cases, once

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

we have

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

F.3.6 Final assembly of the selection-error bound

We now assemble the tube bound, the hard-selection value gap, the on-manifold growth estimate, and the ULA accuracy bound into a single bound on the squared (Euclidean) selection error $\mathbb{E}\|\tilde{x} - x^\star(\theta)\|^2$.

Lemma F.20 (Expected squared selection error bound under Rényi-2 candidate accuracy)**.** Fix $\theta \in \Theta$ and let $x^\star(\theta)$ denote the unique minimizer of $f(\theta, \cdot)$ over $\mathcal{S}(\theta)$. Assume Assumptions 1 to 4 hold. Assume $\mathtt{k} \ge 1$, $0 < \lambda \le \lambda_0$, and that μ_θ^λ satisfies a Poincaré inequality with constant at most C_{PI}. Let $X_1, \dots, X_N$ be independent candidates with marginal laws $\nu_1, \dots, \nu_N$, and define

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

Assume $\varepsilon_{\mathrm{R}} \le 1$. In the ULA/Gibbs instantiation of Algorithm 1, these are the outputs of the N parallel ULA chains, i.e. N independent approximate Gibbs samples. Let $\tilde{x} \in \arg\min_{1 \le i \le N} f(\theta, X_i)$ be the hard-selection output, with a measurable tie-breaking rule, and let

$$\bar{x} \in \arg\min_{x \in \mathcal{S}(\theta)} \|x - \tilde{x}\|$$

be any measurable Euclidean projection of $\tilde{x}$ onto $\mathcal{S}(\theta)$. Here $L_{f,1}$ and D are from Assumption 4; C_{tube}, $C_{\mathrm{tube},2}$, and λ_0 are from Lemmas F.9 and F.10; C_1 is from Lemma F.12; c_{hg} and r_0 are from Lemma F.14; and $\Delta_{r_0} > 0$ is from Lemma F.15. Then

$$\begin{aligned}\mathbb{E}\|\tilde{x} - x^\star(\theta)\|^2 \;&\le 2e^{1/2} C_{\mathrm{tube},2}\, \lambda \log(1+N) \\ &\quad +\; 2\left(\frac{1}{c_{\mathrm{hg}}} + \frac{4\mathsf{D}^2}{\Delta_{r_0}}\right)\Bigg(2L_{f,1}\, C_{\mathrm{tube}} \sqrt{\lambda \log(1+N)} +\; C_1\, N^{-1/\mathtt{k}} \\ &\qquad\qquad +\; 4\sqrt{e-1}\, L_{f,1} \sqrt{N\, C_{\mathrm{PI}}}\, \varepsilon_{\mathrm{R}}\Bigg).\end{aligned} \tag{43}$$

Proof. 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 (28) from Lemma F.10, together with $e^{\varepsilon_{\mathrm{R}}^2/2} \le e^{1/2}$ since $\varepsilon_{\mathrm{R}} \le 1$, gives

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

Then Lemma F.16 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, the Rényi-controlled hard-selection bound (32) from Lemma F.12 gives

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

which gives (43). ∎

Plug-in form. With Proposition F.17, one can run the sampler until the candidate laws satisfy $R_2(\nu_i\|\mu_\theta^\lambda) \le \varepsilon_{\mathrm{R}}^2$. 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 N, (ii) the Gibbs temperature λ, (iii) the achieved Rényi tolerance ε_{R} through the additive $\sqrt{N\,C_{\mathrm{PI}}}\varepsilon_{\mathrm{R}}$ contribution, and (iv) the on-manifold approximation term $N^{-1/\mathtt{k}}$.

F.4 Parameter choices and oracle complexity

This appendix states the parameter-level hyper-gradient error bound obtained by combining (14) with the selection-error bound (43), and then records a convenient parameter choice ensuring $\sup_t \mathbb{E}\|e_t\|^2 = \mathcal{O}(\varepsilon^2)$. Together with Proposition 5.1, this yields the ε-stationarity guarantee and conservative polynomial oracle bound reported in Section 5.

Lemma F.21 (Hyper-gradient error in terms of algorithmic parameters)**.** Using the notation of Lemmas F.1 and F.20, suppose the assumptions therein hold. Assume additionally that $\varepsilon_{\mathrm{R}} \le 1$ and $0 < \gamma \le 1$. Then, uniformly over t,

$$\begin{aligned}\mathbb{E}\|e_t\|^2 = \mathcal{O}\Bigg(\frac{1}{\gamma^2}\Big[\lambda\log(1+N) \;+\; \sqrt{\lambda\log(1+N)} \;+\; N^{-1/\mathtt{k}} \;+\; \sqrt{N\,C_{\mathrm{PI}}}\,\varepsilon_{\mathrm{R}} \;+\; \eta_t^2\Big] \;+\; \gamma^2 \\ +\; (1+R_v^2)N\big(1 + r(\gamma)^{2\mathsf{n}_{\theta x}} + \lambda^{\mathsf{n}_{\theta x}}\big) \\ \times \exp\Big(-\tfrac{c_{\mathrm{tube}}}{4}\tfrac{r(\gamma)^2}{\lambda}\Big)\Bigg).\end{aligned} \tag{44}$$

Proof. We first condition on the current outer iterate θ_t. The constants in Lemma F.20 and Lemma F.10 are uniform over $\theta \in \Theta$, so the resulting conditional bounds can be averaged over θ_t without changing the constants. Combining (14) with (43) from Lemma F.20 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 the following bound, uniformly over t:

$$\begin{aligned}\mathbb{E}\|e_t\|^2 \le 3C_x^2 B_\gamma^2\,\mathcal{R}_{\mathrm{sel}}(N,\lambda,\varepsilon_{\mathrm{R}}) \;+\; \frac{12C_{\mathrm{lin}}^2}{\gamma^2}\,\eta_t^2 \;+\; 3C_{\mathrm{reg}}^2\gamma^2 \\ +\; C_{\mathrm{off}}(1+R_v^2)\,\mathbb{E}\big[(1+\mathsf{D}_t^{2\mathsf{n}_{\theta x}})\mathbf{1}_{\mathcal{E}_t^c}\big],\end{aligned} \tag{45}$$

where

$$B_\gamma := 1 + \frac{2}{\gamma} + \frac{2}{\gamma(c+\gamma)}$$

and

$$\begin{aligned}\mathcal{R}_{\mathrm{sel}}(N,\lambda,\varepsilon_{\mathrm{R}}) := 2e^{1/2}C_{\mathrm{tube},2}\,\lambda\log(1+N) \\ +\,2\left(\frac{1}{c_{\mathrm{hg}}} \;+\; \frac{4\mathsf{D}^2}{\Delta_{r_0}}\right)\Big(2L_{f,1}C_{\mathrm{tube}}\sqrt{\lambda\log(1+N)} \;+\; C_1 N^{-1/\mathtt{k}} \\ +\; 4\sqrt{e-1}\,L_{f,1}\sqrt{N\,C_{\mathrm{PI}}}\,\varepsilon_{\mathrm{R}}\Big).\end{aligned} \tag{46}$$

To control the off-tube moment, the candidate-marginal tube bound from Appendix F.3 (Lemma F.10) and $\varepsilon_{\mathrm{R}} \le 1$ yield, for $\mathsf{D}_t = \mathrm{dist}(\tilde{x}_t, \mathcal{S}(\theta_t))$,

$$\mathbb{P}(\mathsf{D}_t > s) \le N\,e^{1/2}\,\sqrt{C_{\mathrm{tube}}}\exp\Big(-\tfrac{c_{\mathrm{tube}}}{2}\tfrac{s^2}{\lambda}\Big), \qquad s \ge 0.$$

Indeed, $\mathsf{D}_t \le \max_i \mathrm{dist}(X_{t,i}, \mathcal{S}(\theta_t))$, so the maximum-tail bound in Lemma F.10 applies. Integrating the tail gives

$$\begin{aligned}\mathbb{E}\big[(1+\mathsf{D}_t^{2\mathsf{n}_{\theta x}})\mathbf{1}_{\mathcal{E}_t^c}\big] \le C_{\mathrm{tail}}\,N\big(1 + r(\gamma)^{2\mathsf{n}_{\theta x}} + \lambda^{\mathsf{n}_{\theta x}}\big) \\ \times \exp\Big(-\tfrac{c_{\mathrm{tube}}}{4}\tfrac{r(\gamma)^2}{\lambda}\Big),\end{aligned} \tag{47}$$

where C_{tail} depends only on $\mathsf{n}_{\theta x}$ and problem-data constants. One way to see (47) is to use the identity

$$\mathbb{E}\big[\mathsf{D}_t^{2p}\mathbf{1}_{\{\mathsf{D}_t > r\}}\big] = r^{2p}\mathbb{P}(\mathsf{D}_t > r) \;+\; \int_r^\infty 2p\,s^{2p-1}\mathbb{P}(\mathsf{D}_t > s)\,\mathrm{d}s,$$

with $p = \mathsf{n}_{\theta x}$ and $r = r(\gamma)$. In particular, if $\mathbb{P}(\mathsf{D}_t > s) \le K \exp(-as^2/\lambda)$, then

$$\mathbb{P}(\mathsf{D}_t > r) \;+\; \mathbb{E}\big[\mathsf{D}_t^{2p}\mathbf{1}_{\{\mathsf{D}_t > r\}}\big] \le C_p K\big(1 + r^{2p} + \lambda^p\big)\exp\Big(-\tfrac{a}{2}\tfrac{r^2}{\lambda}\Big),$$

which gives (47) with $a = c_{\mathrm{tube}}/2$ and $K = Ne^{1/2}\sqrt{C_{\mathrm{tube}}}$.

Substituting (47) into (45) gives

$$\begin{aligned}\mathbb{E}\|e_t\|^2 \le C\,B_\gamma^2\Big[\lambda\log(1+N) \;+\; \sqrt{\lambda\log(1+N)} \;+\; N^{-1/\mathtt{k}} \;+\; \sqrt{N\,C_{\mathrm{PI}}}\,\varepsilon_{\mathrm{R}}\Big] \\ + \frac{12C_{\mathrm{lin}}^2}{\gamma^2}\eta_t^2 \;+\; 3C_{\mathrm{reg}}^2\gamma^2 \\ + \;C(1+R_v^2)N\big(1 + r(\gamma)^{2\mathsf{n}_{\theta x}} + \lambda^{\mathsf{n}_{\theta x}}\big)\exp\Big(-\tfrac{c_{\mathrm{tube}}}{4}\tfrac{r(\gamma)^2}{\lambda}\Big),\end{aligned} \tag{48}$$

for a problem-data constant $C < \infty$ independent of $N, \lambda, \varepsilon_{\mathrm{R}}, \gamma, \eta_t, R_v$, and t. Since $c > 0$ is fixed and $0 < \gamma \le 1$,

$$B_\gamma^2 = \mathcal{O}\Big(\frac{1}{\gamma^2}\Big).$$

Applying this simplification to (48), and absorbing numerical and problem-data constants, gives (44). Since Theorem 5.2 is the same bound with the exponentially small off-tube term abbreviated as $\mathcal{E}_{\mathrm{off},t}$, and with $\lambda\log(1+N)$ absorbed by its square root when it is at most one, this proves the hyper-gradient error bound stated in Section 5. ■

Choosing parameters to achieve $\mathbb{E}\|e_t\|^2 = \mathcal{O}(\varepsilon^2)$. In the non-singleton regime $\mathtt{k} \ge 1$, fix a sufficiently small target $\varepsilon \in (0,1)$. As in Lemma F.20, ε_{R} denotes the square-root Rényi tolerance, so the sampler accuracy condition is $R_2(\nu_i\|\mu_{\theta_t}^{\lambda}) \le \varepsilon_{\mathrm{R}}^2$. We take ε small enough that the choices below satisfy $0 < \lambda \le \lambda_0 \le \min\{1, \lambda_{\mathrm{PI}}, \lambda_{\mathrm{cov}}\}$, $\varepsilon_{\mathrm{R}} \le \varepsilon_{\mathrm{src}}$, and $0 < \gamma \le 1$. The ULA stepsize and iteration count are chosen jointly as in Proposition F.17 and corollary F.19, rather than independently from their orders of magnitude. To ensure $\sup_t \mathbb{E}\|e_t\|^2 = \mathcal{O}(\varepsilon^2)$ it suffices to make each term on the right-hand side of (44) of order at most ε^2. One convenient choice is: (i) set $\gamma = \varepsilon$ so that $C_{\mathrm{reg}}\gamma = \mathcal{O}(\varepsilon)$, (ii) solve the ridge system to residual $\eta_t = \gamma\varepsilon = \varepsilon^2$, so that $\eta_t^2/\gamma^2 = \mathcal{O}(\varepsilon^2)$, (iii) set $R_v = (2/\gamma)(L_{f,1} + \eta_t)$, so $R_v = \mathcal{O}(\varepsilon^{-1})$, and (iv) choose the sampling/selection parameters so that the remaining bracketed terms satisfy

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

A sufficient scaling is

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

Then $\lambda\log(1+N) = \varepsilon^8$, $\sqrt{\lambda\log(1+N)} = \varepsilon^4$, and $N^{-1/\mathtt{k}} = \mathcal{O}(\varepsilon^4)$. Moreover, the common low-temperature bound C_{PI} from Proposition B.3 is independent of ε, so

$$\sqrt{N\,C_{\mathrm{PI}}}\,\varepsilon_{\mathrm{R}} = \mathcal{O}\Big(\varepsilon^{-2\mathtt{k}}\cdot\varepsilon^{4+2\mathtt{k}}\Big) = \mathcal{O}(\varepsilon^4).$$

Thus the displayed sampling/selection term is $\mathcal{O}(\varepsilon^4)$, and the corresponding contribution to $\mathbb{E}\|e_t\|^2$ is $\mathcal{O}(\varepsilon^2)$ after multiplication by the prefactor $\mathcal{O}(1/\gamma^2) = \mathcal{O}(1/\varepsilon^2)$. It remains to check the off-tube moment term. By Lemma F.2,

$$r(\gamma) = \min\left\{\rho, \frac{\gamma}{2\bar{L}_{g,3}}\right\},$$

so for all sufficiently small ε, $r(\gamma) \ge c_r\varepsilon$ for a problem-data constant $c_r > 0$. Consequently,

$$\frac{r(\gamma)^2}{\lambda} \;\ge\; c_r'\,\varepsilon^{-6}\log(1+N).$$

Since $R_v = \mathcal{O}(\varepsilon^{-1})$, the off-tube term

$$(1+R_v^2)N\big(1 + r(\gamma)^{2\mathsf{n}_{\theta x}} + \lambda^{\mathsf{n}_{\theta x}}\big)\exp\Big(-\tfrac{c_{\mathrm{tube}}}{4}\tfrac{r(\gamma)^2}{\lambda}\Big)$$

is bounded by a polynomial in ε^{-1} times $\exp(-c\,\varepsilon^{-6}\log(1+N))$ for a problem-data constant $c>0$. It is therefore super-polynomially small in ε^{-1} and, in particular, is $\mathcal{O}(\varepsilon^2)$.

Substituting this control into Proposition 5.1 (with stepsize $1/L_F$) yields an ε-stationary guarantee after

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

outer iterations.

Total oracle complexity. Per outer iteration, HG-MS uses (a) NK evaluations of $\nabla_x g$ for the sampler (Algorithm 2), (b) N evaluations of $f(\theta_t,X_{t,i})$ for hard selection, and (c) at most $\#\mathrm{HVP}$ lower-Hessian products for the capped CG solve in Remark F.22. Computing $\nabla_x f$, $\nabla_\theta f$, and one mixed Hessian–vector product adds only a constant number of oracle calls per outer iteration; this is absorbed by the displayed count since $N\ge 1$. First consider the sampler/function-evaluation cost,

$$\mathcal{O}\big(T(NK+N)\big).$$

Assume each ULA chain is initialized from the λ-scaled Gaussian warm start in Lemma F.18, with centers in a fixed compact set. If the ULA sampler is run for K steps so that the square-root Rényi tolerance satisfies $\varepsilon_{\mathrm{R}}=\varepsilon^{4+2\mathtt{k}}$, then Corollary F.19 gives $K=\widetilde{\mathcal{O}}(d\,C_{\mathrm{PI}}^2\,\lambda^{-4}\varepsilon_{\mathrm{R}}^{-2})$. Under the above scaling of $(N,\lambda,\varepsilon_{\mathrm{R}})$, this gives $K=\widetilde{\mathcal{O}}\big(d\,C_{\mathrm{PI}}^2\,\log^4(1+N)\,\varepsilon^{-40-4\mathtt{k}}\big)$ and hence

$$T\,N\,K=\widetilde{\mathcal{O}}\big(d\,C_{\mathrm{PI}}^2\,N\log^4(1+N)\,\varepsilon^{-42-4\mathtt{k}}\big)=\widetilde{\mathcal{O}}\big(d\,C_{\mathrm{PI}}^2\,\varepsilon^{-42-8\mathtt{k}}\big),$$

which dominates the additional TN hard-selection evaluations. If HVPs are counted as oracle calls, the total accounting becomes

$$\mathcal{O}\big(T(NK+N+\#\mathrm{HVP})\big)=\widetilde{\mathcal{O}}\big(d\,C_{\mathrm{PI}}^2\,\varepsilon^{-42-8\mathtt{k}}\big)+\mathcal{O}(T\,\#\mathrm{HVP}).$$

The following remark computes the CG/HVP count and shows that the displayed sampler term dominates it under the parameter choice above.

Remark F.22 (HVP count for the CG solve)**.** To obtain a bound for every run, not only on the tube event, cap the CG iterations. Let $\bar\kappa:=2(L_{g,2}+\gamma)/\gamma$, assume $\eta_t>0$, and use $J_t:=\lceil\sqrt{\bar\kappa}\log(2+2\sqrt{\bar\kappa}L_{f,1}/\eta_t)\rceil$. Initialize at zero, test the residual before each step, and stop if the search direction has nonpositive curvature. On termination return the last finite iterate, or zero if necessary, and clip as in the algorithm. On the tube event the matrix is positive definite and the estimate below proves that the cap suffices for residual η_t; otherwise the off-tube bound uses only the clipping radius, not a residual guarantee. Thus each run uses at most J_t lower-Hessian products, including unsuccessful solves. On the tube event, the ridge system solved by CG has coefficient matrix

$$A_t:=\nabla^2_{xx}g(\theta_t,\tilde x_t)+\gamma I.$$

Lemma F.2 gives $\lambda_{\min}(A_t)\ge\gamma/2$, while the $L_{g,2}$-smoothness of $g(\theta,\cdot)$ gives $\lambda_{\max}(A_t)\le L_{g,2}+\gamma$. Hence

$$\kappa(A_t)\le\frac{2(L_{g,2}+\gamma)}{\gamma}.$$

The standard CG estimate for an SPD system $Av=b$,

$$\|v_k-v_\star\|_A\le 2\left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^k\|v_0-v_\star\|_A,$$

see Nocedal and Wright [2006, Thm. 5.3], implies the residual bound

$$\|b-Av_k\|\le 2\sqrt{\kappa(A)}\left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^k\|b\|,$$

when initialized at $v_0=0$. Each CG iteration applies A_t once, which requires one Hessian-vector product with $\nabla^2_{xx}g(\theta_t,\tilde x_t)$ plus vector operations. Therefore, to reach the residual tolerance η_t in $\|b_t-(H_t+\gamma I)\tilde v_t\|\le\eta_t$, it suffices to take

$$\#\mathrm{HVP}=\mathcal{O}\left(\sqrt{\kappa(A_t)}\log\frac{\sqrt{\kappa(A_t)}\,\|\nabla_x f(\theta_t,\tilde x_t)\|}{\eta_t}\right)=\mathcal{O}\left(\gamma^{-1/2}\log\frac{\gamma^{-1/2}L_{f,1}}{\eta_t}\right),$$

Table 4: **Evaluation uncertainty for the real-world LLM tables.** The ranges summarize the 95% confidence-interval half-widths over the reported cells in the corresponding benchmark. GSM8K/MATH intervals use Wilson binomial intervals over saved binary correctness records; MT-Bench intervals use nonparametric bootstrap intervals over saved question-level judge scores.

Benchmark	CI construction	Evaluation units	95% half-width range
GSM8K, 7B/9B math table	Wilson binomial, $n = 1319$	accuracy points	2.07–2.64
MATH, 7B/9B math table	Wilson binomial, $n = 4989$	accuracy points	1.22–1.35
GSM8K, 32B panel	Wilson binomial, $n = 1319$	accuracy points	1.42–2.23
MATH, 32B panel	Wilson binomial, $n = 4989$	accuracy points	1.37–1.39
MT-Bench instruction table	Question-level bootstrap, $n = 80$	MT-Bench points	0.30–0.50

where the last step uses $\|\nabla_x f(\theta_t, \tilde{x}_t)\| \le L_{f,1}$. With $\gamma = \varepsilon$ and $\eta_t = \varepsilon^2$, this gives

$$\#\mathrm{HVP} = \widetilde{\mathcal{O}}(\varepsilon^{-1/2}), \qquad T\,\#\mathrm{HVP} = \widetilde{\mathcal{O}}(\varepsilon^{-5/2}),$$

which is negligible compared with the sampler-dominated cost $\widetilde{\mathcal{O}}(d\,C_{\mathrm{PI}}^2 \varepsilon^{-42-8\mathtt{k}})$. The final multiplication by $\nabla^2_{\theta x} g(\theta_t, \tilde{x}_t) v_t$ requires one additional mixed Hessian-vector product per outer iteration, which is also dominated by the sampler term.

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. For controlled GPT-2 source reweighting we report token-level validation/test loss and source-weight recovery error. For real-world LLM mathematical reasoning we report GSM8K and MATH accuracy (% correct), and for instruction following we report MT-Bench score on the standard 0–10 scale.

Confidence intervals. For the MNIST experiments, tables/plots report mean $\pm$ 95% confidence intervals across independent seeds, using a two-sided t-interval. The real-world LLM results in Table 2 are single-seed matched-budget comparisons. We complement them with uncertainty estimates computed from the saved evaluation files; see Table 4. For GSM8K and MATH, we treat each test example as correct or incorrect and compute a Wilson 95% binomial interval. For MT-Bench, we resample the saved question-level GPT-4o judge scores over the 80 standard questions. The table reports the range of 95% confidence-interval half-widths across the cells in each benchmark, measured in accuracy percentage points for GSM8K/MATH and MT-Bench score points for instruction following. These intervals quantify uncertainty from the finite evaluation set and do not estimate variance from rerunning source reweighting or final fine-tuning with different seeds.

Time budget. For time-budget plots and budget-sweep diagnostics, we measure wall-clock elapsed time for each algorithm. Fixed-budget tables compare methods after the same per-method wall-clock budget within each benchmark/backbone setting.

Code. Code and configuration files for the MNIST and LLM experiments are available in the following repository:
`https://github.com/smasiha/Experiment_HG_MS_paper`. The synthetic study has a separate repository, provided in Appendix G.5.

For the MNIST experiments, we compare HG-MS against HG-BASELINE [Lorraine et al., 2020], V-PBGD and G-PBGD [Kwon et al., 2024, Shen and Chen, 2023], and IAPTT-GM [Xiao et al., 2023, Shaban et al., 2019], and we report fixed-budget means $\pm$ 95% confidence intervals across seeds. Throughout, $\mathrm{CE}(\cdot,\cdot)$ denotes cross-entropy (negative log-likelihood).

Lower-level solvers and candidate generators. Table 5 summarizes the lower-level optimizer used by the iterative baselines and the candidate generator used by HG-MS in each experiment. For both MNIST benchmarks, the lower-level optimizer family is included in the same validation-loss pilot

Table 5: **Lower-level solvers and HG-MS candidate generators.** Summary of the optimizer used to generate HG-MS candidates and the corresponding lower-level optimizer used by iterative baselines. The iteration counts are lower-level optimizer updates per outer update or per candidate branch, up to the benchmark-specific time budget; detailed learning rates, batch sizes, and regularization values are given in the corresponding subsections.

Experiment	HG-MS candidate generator	Lower optimizer for iterative baselines	Lower-level updates
MNIST data hyper-cleaning	AdamW/SGD selected by validation-loss pilot sweep; reported HG-MS uses 8 optimizer branches	AdamW/SGD selected by validation-loss pilot sweep for HG-BASELINE, V-PBGD, G-PBGD, and IAPTT-GM; reported HG-BASELINE uses AdamW unroll	HG-MS/HG-BASELINE: 20; V-PBGD/G-PBGD/IAPTT-GM: same wall-clock budget
MNIST imbalanced loss tuning	AdamW/SGD selected by validation-loss pilot sweep; reported HG-MS uses one base model plus 4 perturbed optimizer branches	AdamW/SGD selected by validation-loss pilot sweep for HG-BASELINE, V-PBGD, G-PBGD, and IAPTT-GM; reported HG-BASELINE/HG-MS lower optimizer is AdamW	60 warm-start; HG-BASELINE: 30; HG-MS extra branches: 5 by default; V-PBGD: 20; IAPTT-GM: 5
Controlled GPT-2 source reweighting	SGD optimizer branches on the full GPT-2 parameter state	HG-BASELINE and SCALEBIO: SGD lower optimizer under the same validation-loss pilot-sweep protocol	Default 1; verified multilingual HG-MS: 8; configurations selected by validation loss
Real-world LLM math, 7B/8B/9B	$N = 4$ one-update SGD LoRA branches	HG-BASELINE and SCALEBIO: one SGD lower-model update	1 for HG-MS and iterative bilevel baselines
Real-world LLM instruction following	$N = 8$ one-update SGD LoRA branches	HG-BASELINE and SCALEBIO: one SGD lower-model update	1 for HG-MS and iterative bilevel baselines
Real-world `Qwen2.5-32B` math	$N = 4$ one-update SGD LoRA branches	HG-BASELINE and SCALEBIO: one SGD lower-model update	1 for HG-MS and iterative bilevel baselines

sweep for HG-MS and for the bilevel baselines: each HG-MS branch and each baseline lower-level solver is run with the better of AdamW and SGD under the selected configuration. For the controlled GPT-2 and real-world LLM experiments, the corresponding lower-level optimizer choices are stated in the table and detailed in the benchmark-specific subsections below.

G.2 Data hyper-cleaning (MNIST label noise)

This benchmark uses multi-step optimizer branches and a ridge–CG hyper-gradient, not the ULA candidate law of Algorithm 1 or the one-step LoRA implementation of Algorithm 3.

Setup. We study *data hyper-cleaning* on MNIST [LeCun et al., 1998], 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 sigmoid MLP with architecture $784 \to 300 \to 10$ 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, 500 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.

Table 6: **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

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

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

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

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

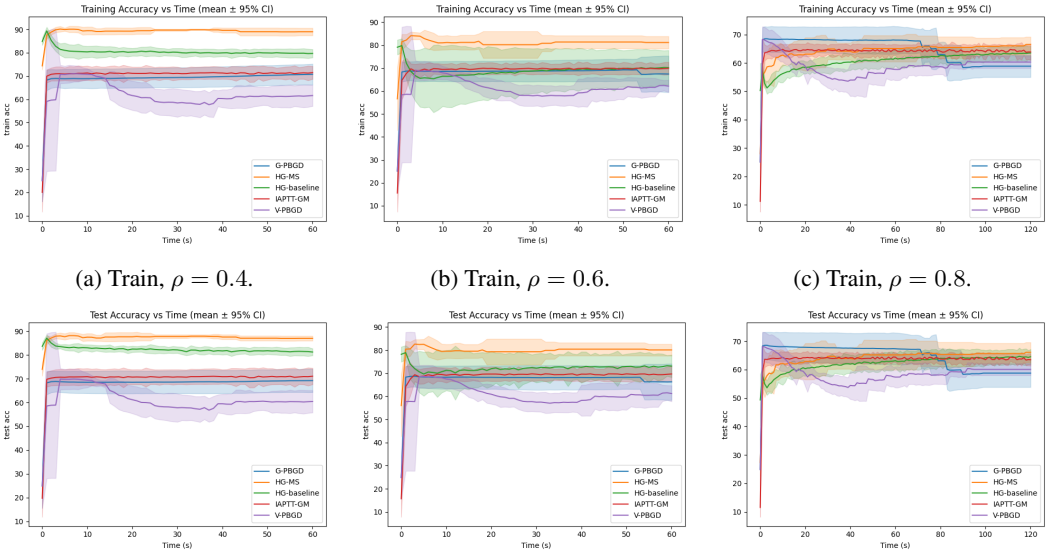
Methods and optimization. We use the following MNIST baselines: HG-BASELINE, V-PBGD, G-PBGD, and IAPTT-GM. All methods use the same 2-layer sigmoid MLP architecture. This label-noise regime can produce lower-level candidates with similar corrupted-training losses but different clean validation losses. HG-MS is designed to select the candidate with the smallest clean validation loss before computing the hyper-gradient.

Implementation details. The AdamW/SGD lower-optimizer choice is selected by the validation-loss pilot sweep described in the common protocol. For HG-BASELINE, the selected lower-level solver is a differentiable AdamW unroll with 20 inner steps per outer update, lower learning rate 5×10^{-3}, AdamW default betas, and weight decay 0. The upper optimizer is AdamW on the data-weight logits with learning rate 0.1 and weight decay 0. The lower regularization coefficient above is $\omega_{\mathrm{dhc}} = 10^{-4}$. For HG-MS, we use 8 candidate branches, and each branch uses the selected lower-optimizer family from the same AdamW/SGD sweep. Each branch takes 20 lower updates per outer step with branch learning rate 2×10^{-2}, minibatch size 256, and branch weight decay 0. The reported runs use no parameter initialization noise by default (`branch_init_noise_std` $= 0$) and no learning-rate diversity across branches. The upper learning rate starts at 0.1 and uses the default cosine schedule to 0.01; the explicit $\rho = 0.8$, 120s script uses an inverse schedule with decay steps 50. The conjugate-gradient solve in HG-MS uses ridge 10^{-2}, residual tolerance 10^{-3}, at most 30 iterations, and full-batch HVP/CG when `cg_batch_size` $= 0$.

Time-budget comparison. Figure 4 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.

G.3 Parametric loss tuning for imbalanced data (MNIST)

This benchmark uses multi-step AdamW branches before hyper-gradient computation; it does not implement either displayed algorithm verbatim.



(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 4: **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 $\pm$ 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).

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 $c \in \{0, \dots, 9\}$ 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. In the implementation, we optimize unconstrained raw variables $(\tilde{\delta}, \tilde{\gamma}) \in \mathbb{R}^{10} \times \mathbb{R}^{10}$ and use the transformed loss parameters

$$\delta = B_\delta \tanh(\tilde{\delta}), \qquad \gamma = \mathbf{1} + B_\gamma \tanh(\tilde{\gamma}),$$

with fixed bounds $B_\delta = 0.5$ and $B_\gamma = 0.15$. Thus every class scale is positive and the unshifted scale $\gamma = \mathbf{1}$ is attained at $\tilde{\gamma} = 0$. 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 $\theta = (\tilde{\delta}, \tilde{\gamma}) \in \mathbb{R}^{10} \times \mathbb{R}^{10}$ be the raw upper-level variable, let $\delta = \delta(\theta)$ and $\gamma = \gamma(\theta)$ denote the transformed loss parameters above, 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

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

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

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

and the upper-level objective is

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

where u_c is an inverse-frequency class weight computed on the validation set, and the regularizers are applied to the transformed loss parameters $\delta(\theta)$ and $\gamma(\theta)$. In the reported HG-MS/HG-BASELINE implementation, the regularization coefficients are $\omega_y = 5 \times 10^{-4}$ and $\omega_\delta = \omega_\gamma = 10^{-2}$. Here $\tilde{x}_i$ denotes the train-time augmented view of x_i, while validation and test examples are uncorrupted.

Methods. We use the common baselines listed in the overview above. For HG-MS, we maintain a small ensemble of lower-level candidates trained on $g(\cdot)$: one base run plus 4 optimizer branches with slightly perturbed optimizer settings. At each outer step, HG-MS selects the candidate that minimizes the validation objective $f(\cdot)$. In this finite-budget nonconvex training pipeline, different branches can produce candidate models with different validation losses and balanced accuracies, so explicit selection can change the point through which the hyper-gradient is computed.

Implementation details. The CNN has architecture $\mathrm{Conv2d}(1, 32, 5, \mathrm{pad} = 2) \to \mathrm{ReLU} \to \mathrm{MaxPool}$, $\mathrm{Conv2d}(32, 64, 5, \mathrm{pad} = 2) \to \mathrm{ReLU} \to \mathrm{MaxPool}$, flatten, $\mathrm{Linear}(64 \cdot 7 \cdot 7, 256) \to \mathrm{ReLU} \to \mathrm{Linear}(256, 10)$. The train and validation batch sizes are both 64, with full-batch validation used when `val_full_batch=True`. The AdamW/SGD lower-optimizer choice is selected by the validation-loss pilot sweep described in the common protocol for all HG-MS branches and for the lower-level solvers used by HG-BASELINE, V-PBGD, G-PBGD, and IAPTT-GM. For HG-BASELINE/HG-MS, the selected lower optimizer is AdamW with learning rate 10^{-3}, weight decay 5×10^{-4}, betas $(0.9, 0.999)$, cosine warm restart scheduler $T_0 = 20$, minimum learning rate 0.2 times the base learning rate, and gradient clipping at 1.0. The upper optimizer is AdamW by default with learning rate 10^{-2}, upper L_2 coefficient 10^{-4}, and upper-gradient clipping at 0.5. We use 60 lower warm-start steps and then 30 lower steps per outer update in the default `minsel_effective`/time-budget setting. HG-MS uses the base model plus 4 extra branches, for 5 candidates total. The branches use initialization noise 0.01; in the default `minsel_effective` setting, branch learning-rate and weight-decay multipliers are log-spaced from 0.5 to 2.0, while the showcase settings use the wider range $[0.25, 4.0]$. In time-budget mode, the default number of branch lower steps is 5 unless overridden by the environment.

For the penalty/alternating baselines, the reported configurations are selected by the same validation-loss pilot protocol. V-PBGD and G-PBGD use train/validation batch sizes $64/64$, lower regularization 5×10^{-4}, upper regularizers $\omega_\delta = \omega_\gamma = 10^{-2}$, $lrx = 2 \times 10^{-2}$, $lry = 5 \times 10^{-2}$, $\gamma_{\max} = 10$, and momentum 0.9; V-PBGD additionally uses `lr_inner` $= 5 \times 10^{-2}$ and `inner_itr` $= 20$. IAPTT-GM uses the default time-budget settings $K_{\mathrm{inner}} = 5$, `lr_inner` $= 0.03$, $lrx = lrz = 5 \times 10^{-4}$, lower regularization 5×10^{-4}, $\omega_\delta = \omega_\gamma = 10^{-4}$, the same CNN model, and the Adam-style inner option enabled.

Relation to the manifold theory. This benchmark tests the select-then-differentiate principle in a finite-budget nonconvex training pipeline, where the inner problems are not solved exactly and different optimizer branches can reach candidate models with different validation losses and balanced accuracies. This is precisely the practical regime in which explicit candidate selection can matter: HG-MS first selects the candidate with the best upper-level value and then computes the hyper-gradient through that selected candidate. The experiment is therefore complementary to the manifold theory in Section 3. Because the explicit L_2-regularization may make the exact global lower-level minimizers isolated ($\mathtt{k} = 0$), we do not use this benchmark as direct evidence for positive-dimensional minimizer manifolds. Rather, it shows that the same select-then-differentiate mechanism can be useful even in the singleton/isolated-minimizer limit when lower-level training is finite-budget and nonconvex.

Table 7: **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

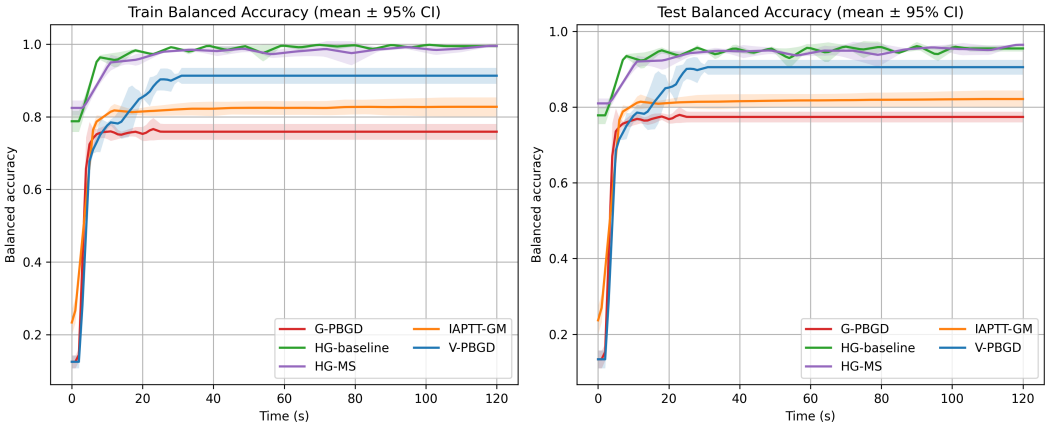


Figure 5: **Imbalanced loss tuning (MNIST).** Train/test balanced accuracy vs. time (mean and 95% CI; 120s per algorithm).

Time-budget comparison. Figure 5 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.

G.4 Completed outer updates in the MNIST fixed-budget runs

To make the fixed-budget comparisons more transparent, Table 8 reports the completed outer updates counted from the saved logs. For data hyper-cleaning, we report the mean over 5 runs with the min–max range in parentheses. For imbalanced loss tuning, we report the mean over 3 repetitions at 120s; a min–max range is included when it was recorded. These counts are not meant to equalize the cost of one outer step across methods: HG-MS completes fewer outer updates because each update is substantially heavier, as it trains multiple candidate branches, performs explicit minima selection, and then computes the selected hyper-gradient.

Table 8: **Completed outer updates in MNIST fixed-budget experiments.** Entries are mean completed outer updates, with min–max ranges in parentheses when available.

Data hyper-cleaning: mean over 5 runs			
Method	$\rho = 0.4$, 60s	$\rho = 0.6$, 60s	$\rho = 0.8$, 120s
HG-MS	312 (310 – −313)	311 (310 – −312)	626 (625 – −628)
HG-BASELINE	1,191 (1,185 – −1,199)	1,200 (1,185 – −1,205)	2,402 (2,380 – −2,416)
V-PBGD	18,747 (17,983 – −19,007)	18,680 (17,826 – −18,904)	37,016 (36,995 – −37,104)
G-PBGD	18,381 (18,337 – −18,452)	18,303 (18,271 – −18,324)	36,233 (36,196 – −37,284)
IAPTT-GM	5,263 (5,148 – −5,427)	5,287 (5,075 – −5,712)	10,745 (10,051 – −11,556)
Imbalanced loss tuning: mean over 3 repetitions at 120**s**			
Method	Completed outer updates		
HG-MS	155 (154 – −156)		
HG-BASELINE	620		
V-PBGD	11,164		
G-PBGD	10,998		
IAPTT-GM	3,568		

G.5 Controlled ULA/ridge-CG mechanism experiment

The GitHub source package[4] provides code and fixed-count mechanism diagnostics; the end-to-end comparison below uses the final 20-seed archive supplied in the supplementary material. This controlled construction isolates the mechanisms analyzed by the theory: manifold coverage, finite-step ULA versus its Gibbs target, and ridge-CG approximation of the pseudoinverse hyper-gradient, in ambient dimension 20. It is distinct from the controlled LLM source-reweighting experiment below.

Problem and exact ground truth. Let $\theta \in [-3, 3]$, $\mathtt{k} \in \{1, 2, 4, 10\}$, and $x = (y, z) \in \mathbb{R}^{\mathtt{k}+1} \times \mathbb{R}^{19-\mathtt{k}}$. Set

$$R_\theta = \sqrt{1+\theta^2}, \qquad a_\theta = (\cos\theta, \sin\theta, 0, \dots, 0) \in \mathbb{S}^{\mathtt{k}}, \qquad c_\theta = \frac{R_\theta}{\sqrt{1+R_\theta^2}},$$

and

$$u_\theta(x) = \frac{\langle y, a_\theta \rangle}{c_\theta \sqrt{1+\|x\|^2}}.$$

The ambient objectives are

$$g(\theta, x) = \sqrt{1 + \big(\|y\|^2 - R_\theta^2\big)^2} - 1 + \frac{1}{2}\|z\|^2,$$
$$f(\theta, x) = 2\sqrt{0.1^2 + \|x\|^2} + (1 + 0.3\theta)\big(u_\theta(x) + 1\big) + \frac{1}{2}\sin^2\big(2\pi(u_\theta(x) + 1)\big).$$

Both terms of g are nonnegative, and they vanish simultaneously exactly when $\|y\| = R_\theta$ and $z = 0$. Hence

$$\mathcal{S}(\theta) = \{(R_\theta v, 0) : v \in \mathbb{S}^{\mathtt{k}}\}.$$

On this sphere, $u_\theta(R_\theta v, 0) = \langle v, a_\theta \rangle$. Writing $s = \langle v, a_\theta \rangle \in [-1, 1]$, the nonconstant part of the restricted upper objective is

$$(1 + 0.3\theta)(s + 1) + \frac{1}{2}\sin^2\big(2\pi(s + 1)\big).$$

Since $1 + 0.3\theta \geq 0.1$, this expression is nonnegative and equals zero only at $s = -1$. Equality in $\langle v, a_\theta \rangle \geq -1$ then forces $v = -a_\theta$, proving the unique optimistic solution

$$x^\star(\theta) = (-R_\theta a_\theta, 0), \qquad F(\theta) = 2\sqrt{1.01 + \theta^2}, \qquad F'(\theta) = \frac{2\theta}{\sqrt{1.01 + \theta^2}}.$$

Therefore $\theta^\star = 0$ is the unique outer minimizer and $F^\star = 2\sqrt{1.01}$.

[4] `https://github.com/smasiha/HG-MS_Synthetic_Experiment_`

For completeness, at $v = -a_\theta$ the differential of $s(v) = \langle v, a_\theta \rangle$ vanishes on the tangent space of the unit sphere, and its Riemannian Hessian is the identity there. Rescaling from the unit sphere to the radius-R_θ sphere shows that every eigenvalue of the restricted/Riemannian Hessian of $f(\theta, \cdot)$ at $x^\star(\theta)$ equals

$$\frac{1 + 0.3\theta}{R_\theta^2} = \frac{1 + 0.3\theta}{1 + \theta^2} \geq \frac{0.1}{10} = 0.01.$$

Thus the optimistic solution is uniformly intrinsically nondegenerate over the stated interval.

Mechanism-level scope, not full assumption verification. This objective does *not* satisfy the manuscript's complete global PŁ$^\circ$ critical-point condition. Indeed, $(y, z) = (0, 0)$ is stationary and

$$\nabla^2_{xx} g(\theta, 0) = \operatorname{diag}\left(-\frac{2R_\theta^2}{\sqrt{1 + R_\theta^4}} I_{\mathtt{k}+1}, I_{19-\mathtt{k}} \right),$$

which is indefinite for every reported $\mathtt{k}$. The origin is therefore an off-manifold saddle, not a negative-definite strict maximum. The study is designed to test specific selection, sampling, and pseudoinverse/ridge-CG mechanisms; it is not an assumption-faithful end-to-end verification of PŁ$^\circ$.

ULA, selection, and ridge-CG protocol. At each outer iteration, independently refresh all chains: draw $v_i \sim \mathrm{Unif}(\mathbb{S}^{\mathtt{k}})$ and initialize

$$X_i^{(0)} \sim \mathcal{N}\big((2v_i, 0), \lambda I_{20}\big).$$

The end-to-end HG-MS configuration is $(N, K, h, \lambda) = (16, 100, 0.01, 10^{-3})$. After K ULA steps, hard-select $i^\star \in \arg\min_{1 \leq i \leq N} f(\theta, X_i^{(K)})$ and set $\widehat{x} = X_{i^\star}^{(K)}$. At the selected point, solve

$$\big(\nabla^2_{xx} g(\theta, \widehat{x}) + \gamma I\big) v = \nabla_x f(\theta, \widehat{x}), \qquad \widehat{h} = \nabla_\theta f(\theta, \widehat{x}) - \nabla^2_{\theta x} g(\theta, \widehat{x}) v,$$

with ridge $\gamma = 0.3$ and absolute CG residual tolerance 10^{-8}. CG starts from zero, has a cap of 64 iterations, and flags nonfinite or numerically nonpositive curvature ($p^\top A p \leq 10^{-30}$); there is no dense-solve fallback. The outer step is $\theta^+ = \Pi_{[-3,3]}(\theta - 0.01\widehat{h})$. No gradient clipping is used in this mechanism experiment; the 160 held-out HG-MS/HG-BASELINE records report no CG breakdown or tolerance failure.

Baselines and frozen pilot selection. The synthetic HG-BASELINE uses a single ULA chain with the same (K, h, λ), CG tolerance, and outer step, but its independently selected ridge is $\gamma = 0.8$. This is not the one-SGD-step baseline used for the LLM experiments, and the comparison is not a pure ablation of N. For each of the five methods, eight configurations were tested at $\mathtt{k} = 1$, for two seconds per pilot seed $11, 28, 47$. The configuration with the smallest mean logarithmic objective gap was frozen and reused at every held-out dimension.

For clarity, the other frozen implementations are as follows. Write $s_f = \min\{1, 1/\beta\}$ and $s_g = \min\{1, \beta\}$. V-PBGD uses the scaled value-penalty objective $s_f f(\theta, x) + s_g[g(\theta, x) - g(\theta, \widehat{\omega})]$, with $\beta = 30$, outer/lower step sizes $0.01/0.01$, and ten virtual lower steps of size 0.02; differentiation treats $\widehat{\omega}$ as fixed. G-PBGD uses $s_f f(\theta, x) + (s_g/2)\|\nabla_x g(\theta, x)\|^2$, with $\beta = 10$ and outer/lower step sizes $0.03/0.03$. IAPTT-GM takes 20 differentiable lower steps, selects the point with the *largest* upper loss among the updated points $x_1, \ldots, x_{20}$, excluding x_0 (pessimistic truncation), and differentiates through that point to update both θ and the lower initialization; its inner, outer, and initialization step sizes are all 0.03. All methods project the outer update onto $[-3, 3]$.

Final held-out comparison and uncertainty convention. The verified final comparison uses seeds $2000, \ldots, 2019$, five methods, and four values of $\mathtt{k}$, giving $20 \times 5 \times 4 = 400$ end-to-end runs. Each method receives a 60-second *wall-clock* budget with single-thread Apple Accelerate on the same machine. The implementation completes the update in progress when the budget expires; recorded elapsed times range from 60.0000 to 60.0163 seconds. Within each seed–dimension cell, all methods share $\theta_0 = (-1)^s U$, $U \sim \mathrm{Unif}[1, 2]$, generated with seed $s + 71003$, and the first lower candidate generated with seed $1009s + 17$. HG-MS can select a different candidate from its larger pool. Method execution order is randomized with seed $s + 99991\mathtt{k}$; the launcher fixes the BLAS thread counts to one.

Table 9 reports $F(\theta_T) - F^\star$, $|F'(\theta_T)|$, and $\|\widehat{x} - x^\star(\theta_T)\|$. Here $\widehat{x}$ is the lower point retained from the final update, evaluated against the exact optimistic solution at the final outer iterate, not a fresh sample at θ_T. For each method–dimension cell and metric, let $\bar{m}$ be the mean of the 20 seed endpoints. The reported interval is $[\bar{m} - H, \bar{m} + H]$, where

$$H = \mathrm{quantile}_{0.95}\big\{|\bar{m}^{*(b)} - \bar{m}| : b = 1, \ldots, 5000\big\},$$

and each $\bar{m}^{*(b)}$ resamples 20 seeds with replacement. Thus the table uses symmetric bootstrap 95% half-widths, not percentile endpoints. The aggregation uses random seed 92731, sorted dimension/method cells, and metric order gap, stationarity, and retained-point error.

Table 9: Final controlled synthetic comparison: 20 held-out seeds per method and dimension, 400 runs in total, with a 60-second single-thread wall-clock budget per run. Entries are means $\pm$ symmetric bootstrap 95% interval half-widths; all three metrics are minimized. Settings are selected separately for each method on pilot seeds and then frozen.

Method	$\mathtt{k} = 1$	$\mathtt{k} = 2$	$\mathtt{k} = 4$	$\mathtt{k} = 10$
(a) Upper-objective gap $F(\theta_T) - F^\star$				
HG-MS	$\mathbf{(1.13 \pm 0.60)10^{-3}}$	$\mathbf{(5.22 \pm 2.79)10^{-3}}$	$\mathbf{(4.40 \pm 1.78)10^{-3}}$	$\mathbf{(7.88 \pm 1.77)10^{-3}}$
HG-BASELINE	$.029 \pm .010$	$.032 \pm .008$	$.034 \pm .008$	$.042 \pm .009$
V-PBGD	$.101 \pm .040$	$.076 \pm .029$	$.067 \pm .019$	$.066 \pm .017$
G-PBGD	$2.389 \pm .863$	$2.386 \pm .863$	$2.599 \pm .861$	$2.385 \pm .861$
IAPTT-GM	$2.160 \pm .863$	$1.945 \pm .863$	$1.729 \pm .863$	$1.945 \pm .863$
(b) Stationarity $\|F'(\theta_T)\|$				
HG-MS	$\mathbf{.053 \pm .018}$	$\mathbf{.116 \pm .037}$	$\mathbf{.113 \pm .030}$	$\mathbf{.170 \pm .021}$
HG-BASELINE	$.304 \pm .059$	$.330 \pm .052$	$.351 \pm .042$	$.390 \pm .038$
V-PBGD	$.495 \pm .145$	$.459 \pm .111$	$.466 \pm .081$	$.467 \pm .073$
G-PBGD	$1.165 \pm .358$	$1.158 \pm .358$	$1.238 \pm .344$	$1.171 \pm .330$
IAPTT-GM	$.990 \pm .371$	$.908 \pm .371$	$.831 \pm .359$	$.920 \pm .359$
(c) Retained-point error $\|\widehat{x} - x^\star(\theta_T)\|$				
HG-MS	$\mathbf{.284 \pm .076}$	$\mathbf{.515 \pm .121}$	$\mathbf{.798 \pm .114}$	$\mathbf{1.049 \pm .058}$
HG-BASELINE	$1.264 \pm .268$	$1.431 \pm .226$	$1.329 \pm .181$	$1.469 \pm .098$
V-PBGD	$1.333 \pm .350$	$1.311 \pm .250$	$1.348 \pm .177$	$1.351 \pm .165$
G-PBGD	$2.271 \pm .480$	$2.260 \pm .475$	$2.384 \pm .439$	$2.365 \pm .374$
IAPTT-GM	$1.669 \pm .626$	$1.568 \pm .649$	$1.415 \pm .600$	$1.584 \pm .626$

HG-MS has the lowest mean in all 12 metric–dimension cells under this protocol. The means also reveal heterogeneous baseline outcomes: at $\mathtt{k} = 4$, IAPTT-GM reaches gap below 0.01 in 12 of 20 seeds, while the other eight terminate at the boundary $\theta = -3$; all consume the allotted budget. These are empirical outcomes for the frozen implementations, not a significance claim, a universal ranking, or a verification of the full PŁ$^\circ$ convergence theorem.

Fixed-count selection and linear-solver diagnostics. The exact-manifold BoN diagnostic samples candidates directly as $(R_\theta v_i, 0)$ at $\theta = 1$, uses 2,000 independent trials per N, and measures $\mathbb{E}\|\widehat{x}_N - x^\star\|^2$ at the final posted choices $N = 16, 256, 2048$. The ridge diagnostic compares the Moore–Penrose formula

$$\widehat{h}_{\mathrm{pinv}} = \nabla_\theta f - \nabla^2_{\theta x} g (\nabla^2_{xx} g)^\dagger \nabla_x f$$

with the exact derivative $F'(\theta)$, then separately measures ridge bias and the CG/direct-solve discrepancy.

The fixed-count records underlying the exact-manifold and linear-solver diagnostics are also present in the GitHub source package. We checked their manifest hashes and regenerated the displayed rounding from the machine-readable CSV files. For selection, let $E_N = \mathbb{E}\|\widehat{x}_N - x^\star\|^2$. The empirical values in Table 10 use 2,000 trials at every power of two from $N = 1$ through 2048; the reported log–log slopes are fitted over $N = 64, 128, \ldots, 2048$.

Table 10: Artifact-verified exact-manifold BoN selection diagnostic. The three displayed candidate counts are exactly the columns posted in the final OpenReview response; lower error is better. The fitted slope uses all six values $N = 64, 128, \ldots, 2048$.

k	E_{16}	E_{256}	E_{2048}	fitted slope
1	0.152	5.95×10^{-4}	9.24×10^{-6}	-2.016
2	0.590	0.0306	0.00367	-1.016
4	1.370	0.262	0.0907	-0.599
10	2.096	1.715	1.022	-0.158

The same checksummed records give

$$\max\left|\widehat{h}_{\mathrm{pinv}} - F'(\theta)\right| \le 6.7 \times 10^{-16},$$
$$\mathrm{slope}_{\gamma}\left|\widehat{h}_{\gamma} - \widehat{h}_{\mathrm{pinv}}\right| = 0.993,$$
$$\frac{1}{15}\sum_{j=1}^{15} \|v_{\mathrm{CG},j} - v_{\mathrm{dense},j}\| = 4.4 \times 10^{-12},$$

The pseudoinverse check uses exact optimistic points at $\mathtt{k} = 2$ and $\theta \in \{0.5, 1, 2\}$. The ridge-bias slope is fitted at $(\mathtt{k}, \theta) = (2, 1)$ over $\gamma \in \{0.3, 0.1, 0.03, 0.01, 0.003\}$. The last value is an arithmetic mean over 15 nonoptimistic on-manifold systems: rotate the optimistic direction by 0.3 radians at each of the three values of θ, and use each of the five ridge values, with absolute CG tolerance 10^{-8}. Thus the pseudoinverse identity is at machine precision, the ridge bias exhibits the predicted first-order dependence on γ, and a sufficiently tight CG tolerance recovers the dense ridge solve. The selection slopes are observed finite-range fits, not estimates of the theorem's worst-case exponent. They are consistent with the problem-specific $N^{-2/\mathtt{k}}$ asymptotic trend (with substantial finite-range effects at large k); the theorem's $N^{-1/\mathtt{k}}$ term remains a conservative envelope. The increasingly shallow slopes as k grows directly illustrate why intrinsic manifold dimension controls finite-candidate coverage.

Gibbs target and numerical direct-reference sampling. The Gibbs target factorizes in polar coordinates. Draw the angular component $v \sim \mathrm{Unif}(\mathbb{S}^{\mathtt{k}})$, draw $z \sim \mathcal{N}(0, \lambda I_{19-\mathtt{k}})$, and independently draw $r \ge 0$ with one-dimensional density

$$p_{\theta,\lambda}(r) \propto r^{\mathtt{k}} \exp\left\{-\frac{\sqrt{1 + (r^2 - R_\theta^2)^2} - 1}{\lambda}\right\}.$$

An ideal draw of r from this density makes (rv, z) an exact Gibbs draw. The released routine named `sample_exact_gibbs` implements an independent high-resolution numerical reference: it evaluates the radial density on $100{,}000$ equally spaced points over

$$\left[10^{-12},\ R_\theta + \max\left\{2, \frac{12\sqrt{\lambda}}{\max\{R_\theta, 0.25\}}\right\}\right],$$

forms the normalized CDF by the cumulative trapezoidal rule, and applies linear inverse-CDF interpolation. It generates the spherical direction by normalizing a standard Gaussian vector and samples z directly from $\mathcal{N}(0, \lambda I_{19-\mathtt{k}})$. Hence "exact Gibbs" means an independent direct reference rather than a Markov-chain output; the implementation still has finite radial truncation and grid/interpolation error. It is not adaptive quadrature and provides no certified numerical tolerance for these errors.

Final finite-step ULA versus Gibbs-reference sweeps. Both sweeps use $(\mathtt{k}, d, \theta, \lambda, h) = (2, 20, 1, 10^{-3}, 0.01)$, the Gaussian chain initialization above, and the same hard selection and selected-point ridge–CG estimator under both candidate laws. Here the ridge is $\gamma = \max\{0.3, 10\sqrt{\lambda}\} = 1/\sqrt{10}$, not the end-to-end value 0.3. CG starts from zero with absolute residual tolerance 10^{-8}, cap $4d = 80$, and no dense fallback or clipping. All recorded solves use at most three CG iterations and report no curvature breakdown. The K-sweep fixes $N = 32$ and uses $K \in \{10, 100, 500, 1000\}$; the N-sweep fixes $K = 100$ and uses $N \in \{1, 8, 32, 128\}$.

For each setting, 20 seeds $202603, \ldots, 202622$ each contribute 250 ULA selections and 5,000 independent numerical Gibbs-reference selections. These seeds are distinct from the end-to-end

seeds. The K-sweep shares trajectory checkpoints and a Gibbs reference within a seed. The N-sweep uses nested candidate prefixes within each law, with separate random streams for ULA and Gibbs. The two sweeps use different draws, so their $(K, N) = (100, 32)$ Monte Carlo estimates need not coincide.

Table 11 reports the mean of $(\widehat{h} - F'(\theta))^2$, averaging first within each seed and then across seeds. Unlike the symmetric intervals in the end-to-end table, the brackets here are 2.5th and 97.5th percentiles of 5,000 bootstrap resamples of the 20 seed means. A common resampled seed index is used for each ULA/Gibbs pair; the bootstrap seeds are 20260728 and 20260729 for the K- and N-sweeps, respectively.

Table 11: Final finite-step ULA/Gibbs-reference hyper-gradient MSE sweeps. Entries are means [bootstrap 95% percentile intervals], based on 20 seeds with 250 ULA and 5,000 numerical direct-reference selections per seed and setting. Both laws use the same ridge–CG estimator with $\gamma = 1/\sqrt{10}$.

Setting	ULA MSE [95% interval]	Gibbs-reference MSE [95% interval]
(a) K-sweep at $N = 32$		
$K = 10$	$0.710\ [0.683, 0.739]$	$0.681\ [0.676, 0.687]$
$K = 100$	$0.668\ [0.641, 0.693]$	$0.681\ [0.676, 0.687]$
$K = 500$	$0.657\ [0.635, 0.678]$	$0.681\ [0.676, 0.687]$
$K = 1000$	$0.663\ [0.640, 0.685]$	$0.681\ [0.676, 0.687]$
(b) N-sweep at $K = 100$		
$N = 1$	$2.413\ [2.298, 2.532]$	$2.283\ [2.266, 2.300]$
$N = 8$	$1.410\ [1.350, 1.471]$	$1.328\ [1.316, 1.342]$
$N = 32$	$0.687\ [0.663, 0.710]$	$0.687\ [0.680, 0.695]$
$N = 128$	$0.119\ [0.111, 0.127]$	$0.116\ [0.114, 0.117]$

Across the four K values, the absolute relative MSE differences from the Gibbs reference are $4.2\%, 2.0\%, 3.5\%, 2.8\%$, respectively. The N-sweep shows a decreasing empirical MSE for both laws as the candidate count grows. These finite-sample diagnostics support the sampling/selection mechanism but neither prove a Rényi-divergence bound nor imply monotone improvement with K, exact Gibbs sampling, or applicability of the full global $\mathtt{PŁ}^\circ$ convergence theorem to this construction.

G.6 Controlled GPT-2 source reweighting

Bilevel formulation. We use the two-source small-scale LLM tasks from ScaleBiO [Pan et al., 2025] to test whether the outer variable learns a nontrivial source mixture. Let $\theta \in \mathbb{R}^2$ be source logits, $\omega = \mathrm{softmax}(\theta) \in \Delta^1$ the induced source weights, and y the GPT-2 parameter vector. The lower-level and upper-level losses are

$$g(\theta, y) \;:=\; \sum_{s=1}^{2} \omega_s \, \frac{1}{n_s} \sum_{i=1}^{n_s} \ell(y; \xi_{s,i}^{\mathrm{tr}}), \qquad f(y) \;:=\; \frac{1}{n_{\mathrm{val}}} \sum_{j=1}^{n_{\mathrm{val}}} \ell(y; \xi_j^{\mathrm{val}}).$$

Here $\mathcal{D}_s^{\mathrm{tr}} = \{\xi_{s,i}^{\mathrm{tr}}\}_{i=1}^{n_s}$ are the two training sources, $\mathcal{D}^{\mathrm{val}} = \{\xi_j^{\mathrm{val}}\}_{j=1}^{n_{\mathrm{val}}}$ is the target validation set, and $\ell(y; \xi)$ is the token-level language-model negative log-likelihood. Thus the lower level trains on a source-weighted mixture, while the upper level evaluates held-out target loss.

Backbone and shared code path. The controlled experiments use GPT-2 (124M) [Radford et al., 2019] as the lower-level model, with maximum sequence length 128. Unlike the real-world LLM experiments below, this controlled GPT-2 code path does *not* use LoRA: it trains the full GPT-2 trainable parameter state loaded by `AutoModelForCausalLM`. Thus LoRA rank, scaling, and dropout are not applicable to this controlled experiment. This code follows the SGD-branch, hard-selection, and selected-branch differentiation structure of Algorithm 3, using the full GPT-2 parameter state and task-dependent inner step counts instead of one-step LoRA branches. Candidate diversity comes from minibatches, data order, seeds, and parameter perturbations. These branches are neither exact Gibbs samples nor finite-step ULA candidates. The Gibbs/Rényi, selected-point, hyper-gradient-MSE,

Table 12: **Controlled two-source LLM source reweighting adapted from ScaleBiO.** All methods use the same GPT-2 backbone and single-GPU code path. HG-BASELINE is the one-branch hyper-gradient variant, and HG-MS is the minima-selection variant. In denoising, $\omega = (\omega_{\text{clean}}, \omega_{\text{noisy}})$ and $\omega^\star = (1, 0)$; in multilingual training, $\omega = (\omega_{\text{ZH}}, \omega_{\text{EN}})$ and $\omega^\star = (0.6, 0.4)$. Lower is better.

Task	Method	Learned ω	$\|\omega - \omega^\star\|_1$	Val loss	Test loss
Denoising	Uniform	$(0.500,\ 0.500)$	1.000	2.571	2.610
	ScaleBiO	$(0.922,\ 0.078)$	0.156	**2.336**	2.443
	HG-baseline	$(0.930,\ 0.070)$	0.140	2.427	2.454
	HG-MS	$(0.962,\ 0.038)$	**0.077**	2.403	**2.437**
Multilingual	Uniform	$(0.500,\ 0.500)$	0.200	2.430	2.407
	ScaleBiO	$(0.623,\ 0.377)$	**0.047**	2.339	2.371
	HG-baseline	$(0.540,\ 0.460)$	0.119	2.327	2.307
	HG-MS	$(0.514,\ 0.486)$	0.172	**2.323**	**2.298**

stationarity, and oracle-complexity guarantees do not transfer. Shared validation batches reduce between-branch evaluation noise; all methods use the same single-GPU PyTorch infrastructure.

Task construction. For *data denoising*, we follow the two-source construction of Pan et al. [2025]: the training set contains 1000 clean Alpaca examples [Taori et al., 2023] and 9000 Alpaca examples with synthetically corrupted outputs, where each target sequence is replaced by ".". The validation and test sets each contain 1000 clean Alpaca examples. Thus the desired source weight is $\omega^\star = (1, 0)$. For *multilingual training*, the training set contains $40{,}000$ Alpaca-GPT4-ZH examples and $40{,}000$ Alpaca-GPT4-EN examples [Peng et al., 2023], while the validation and test sets each contain 600 ZH plus 400 EN examples, so the target source weight is $\omega^\star = (0.6, 0.4)$.

Methods and optimization details. We compare UNIFORM, SCALEBIO, HG-BASELINE, and HG-MS. UNIFORM fixes $\omega = (0.5, 0.5)$. Source logits are initialized at $\theta = 0$, hence the initial learned mixture is uniform, unless a fixed source mixture is explicitly passed to the script. For each tunable method, we selected the reported controlled-task configuration from a small method-specific pilot grid using only held-out validation loss. The grid varied learning-rate and update parameters appropriate to each method; test loss was not inspected for hyperparameter selection. In the reported HG-MS configurations, branch selection uses only the upper-level validation loss.

The default controlled script uses SGD as the lower optimizer, batch size 64, selection validation batch size 128, evaluation batch size 64, one inner step, two branches, branch noise 10^{-3}, model learning rate 10^{-5}, outer learning rate 10^{-2}, weight decay 10^{-2}, and source-logit gradient clipping at 10. The verified multilingual HG-MS configuration uses model learning rate 10^{-3}, outer learning rate 1.5, batch size 8, selection batch size 64, 8 selection batches, 8 inner steps, and 2 candidate branches. In both controlled tasks, the outer hyper-gradient is computed through the selected branch using shared validation batches.

Metrics. For both tasks we report token-level validation and test losses, together with the source-weight recovery error $\|\omega - \omega^\star\|_1$. In denoising, the latter remains a direct diagnostic because the clean-data optimum is unambiguous. In multilingual training, however, we treat held-out loss as the primary downstream metric and use $\|\omega - \omega^\star\|_1$ only as a calibration diagnostic, since finite-horizon nonconvex training can make the loss-optimal source mixture differ from the nominal validation composition.

Results. Table 12 reports the completed runs. In both tasks, the learned ω moves away from its initialization $(0.5, 0.5)$. On denoising, all reweighting methods improve over UNIFORM; HG-MS gives the best test loss (2.437) and best clean/noisy recovery, while SCALEBIO has the best validation loss. On multilingual training, HG-MS gives the best validation/test losses ($2.323/2.298$), whereas SCALEBIO is closest to the nominal source ratio. Since the bilevel objective optimizes held-out loss rather than source-ratio recovery, this is consistent with minima selection favoring branches with better validation behavior even when their mixture is farther from $\omega^\star$.

G.7 Detailed protocols for the real-world LLM experiments

Algorithm 3 specifies the real-world LLM procedure: $N = 4$ for mathematical reasoning, including `Qwen2.5-32B`; $N = 8$ for instruction following; and $N = 1$ for HG-BASELINE. One-step SGD, selected-step differentiation, and Adam replace ULA/Gibbs candidates, ridge–CG, and projected gradient, rather than approximating them. The scope qualifications in Appendix A apply to this implementation.

Algorithm 3 Practical HG-MS for real-world LLM source reweighting

Require: source logits θ_t; current LoRA state y_t; branch count N; inner SGD stepsize β; Adam state

1: **for** $i = 0, \dots, N-1$ in parallel **do**

2: Draw an independent minibatch and seed; set $y_{t,0}^{(0)} = y_t$ and, for $i > 0$, $y_{t,i}^{(0)} = y_t + \delta_{t,i}$ with an independent Gaussian perturbation $\delta_{t,i}$

3: Take one differentiable SGD step: $y_{t,i}^{+} = y_{t,i}^{(0)} - \beta \nabla_y g(\theta_t, y_{t,i}^{(0)}; \mathcal{B}_{t,i})$

4: Evaluate every $y_{t,i}^{+}$ on the same held-out objective and hard-select $i_t^\star \in \arg\min_i f(\theta_t, y_{t,i}^{+}; \mathcal{D}^{\mathrm{val}})$

5: Backpropagate the selected validation loss through branch $i_t^\star$'s SGD step to obtain the source-logit hyper-gradient $\widehat{h}_t$

6: Update source logits with Adam: $\theta_{t+1} = \mathrm{Adam}(\theta_t, \widehat{h}_t)$

Shared setup for the main real-world tables. The real-world LLM tables in the main text use single-seed, single-H100 LoRA pipelines in which the outer variable is the source-logit vector θ and the simplex weights are $\omega = \mathrm{softmax}(\theta)$. Unless stated otherwise, every reported run uses one H100 in bf16 precision[5], and each method/backbone pair is run independently on that single-GPU path. For every method, the learned source weights are converted into a final sampled supervised-fine-tuning set, and the target model is then fine-tuned for one fixed epoch. The final supervised fine-tuning optimizer is AdamW with learning rate 8×10^{-6}, no scheduler or warmup, default AdamW weight decay 0.01, and no explicit max-gradient-norm clipping in the checked code paths. Thus the comparison is intended to isolate the source-reweighting rule rather than the final fine-tuning budget. All main-table runs reported in Tables 2 and 3 use seed 2424. We interpret the real-world LLM results as controlled matched-budget comparisons under a fixed implementation and seed; Table 4 reports evaluation-set uncertainty, while the MNIST experiments provide multi-seed confidence intervals. The non-iterative baselines (UNIFORM, LESS, and RHO-LOSS) use the standard adapted pipeline without an outer reweighting horizon. For both tables, the iterative methods (SCALEBIO, HG-BASELINE, and HG-MS) use matched per-backbone wall-clock reweighting budgets of approximately 9h for `Qwen-2-7B`, 12h for `Llama-3-8B`, and 15h for `Gemma-2-9B`.

Choice of wall-clock budgets. The iterative LLM methods are finite-budget procedures: the reweighting horizon determines the learned source distribution, which is then used to sample the final fine-tuning set. We choose the wall-clock budgets to give the iterative methods substantial reweighting time under the single-H100 implementation, while keeping the comparison feasible and matched. The budget is fixed for a given task/backbone and then shared by SCALEBIO, HG-BASELINE, and HG-MS; because the backbones have different per-step costs, matching is by wall-clock time rather than by the number of completed outer updates.

The 9h budget used for the main `Qwen-2-7B` mathematical-reasoning comparison was fixed before evaluating the final checkpoints on the GSM8K and MATH test sets. The explicit `Qwen-2-7B` math sweep in Table 21 is included as an audit trail and sensitivity diagnostic, not as a test-set-based procedure for choosing the main stopping budget. Longer 15h and 21h runs are therefore reported to show budget sensitivity under the same pipeline, while the main table keeps the pre-specified 9h budget. The same fixed-budget principle is used for the other LLM task/backbone budgets: the chosen horizons are intended to give the iterative algorithms sufficient reweighting time under a matched wall-clock constraint, not to reward additional wall-clock after inspecting downstream test metrics.

[5] A 16-bit floating-point format is used to train/evaluate neural networks faster and with less GPU memory than float32. The bf16 has a much larger exponent range, so it is usually more numerically stable for LLM training. On H100 GPUs, bf16 is well supported and is the standard choice for large-model LoRA fine-tuning.

Table 13: **Instruction-following source portfolio.** Training sources used for the eighteen-source instruction-following benchmark, following Table 10 of Pan et al. [2025]. The first nine rows are the high-quality/general instruction or chat sources; the last nine rows are multilingual Alpaca variants.

Source	#Samples	Kind	License
AlpacaGPT4	52K	Instruction	Apache-2.0
ShareGPT4	6K	Conversation	Apache-2.0
SlimOrca	518K	Instruction	MIT
AlpacaChat	20K	Conversation	Apache-2.0
OpenOrcaGPT4	1M	Instruction	MIT
WildChat	1M	Conversation	AI2 ImpACT
LMSYS-Chat	1M	Conversation	LMSYS-Chat-1M
GPTeacher	89K	Instruction	MIT
Airoboros	59K	Conversation	CC-BY-4.0
Alpaca-es	52K	Instruction	CC-BY-4.0
Alpaca-de	50K	Instruction	Apache-2.0
Alpaca-ja	52K	Instruction	CC-BY-NC-SA-4.0
Alpaca-ko	50K	Instruction	CC-BY-NC-4.0
Alpaca-ru	30K	Instruction	CC-BY-4.0
Alpaca-it	52K	Instruction	CC-BY-NC-SA-4.0
Alpaca-fr	55K	Instruction	Apache-2.0
Alpaca-zh	49K	Instruction	CC-BY-4.0
Alpaca-pt	52K	Instruction	CC-BY-NC-4.0

Table 14: **Mathematical-reasoning source portfolio and validation sources.** Training sources above the midrule and validation/tuning sources below the midrule, following Table 11 of Pan et al. [2025].

Source	#Samples	Kind	License
DART-Math	591K	Math	MIT
SlimOrca	518K	Instruction	MIT
UltraInteract SFT	289K	Reasoning	MIT
MathInstruct	262K	Reasoning	MIT
Orca-Math word problems	200K	Math	MIT
WizardLM Evol-Instruct V2	196K	Instruction	MIT
Magicoder Evol-Instruct	110K	Coding	Apache-2.0
ShareGPT-Vicuna unfiltered	94K	Instruction	Apache-2.0
GPTeacher-General-Instruct	89K	Instruction	MIT
GPT4-LLM-Cleaned	55K	Instruction	Apache-2.0
GSM8K	7.5K	Math	MIT
Competition Math	12.5K	Math	MIT
Self-OSS-Instruct	50.7K	Coding	ODC-By
MMLU	116K	Science	MIT
ARC-Easy	5.2K	Instruction	CC-BY-SA-4.0
ARC-Challenge	2.6K	Instruction	CC-BY-SA-4.0

Relation to the ScaleBiO protocol. Although the table layouts follow the corresponding ScaleBiO tables, the numbers below are adapted single-H100 LoRA reproductions under our shared code path and final fine-tuning protocol, not direct copies of the original ScaleBiO results. Concretely, Table 2 follows the layout of Table 4 of Pan et al. [2025]; in Table 3, panel (a) follows the layout of their Table 5 and panel (b) follows the layout of their Table 2. The comparison with ScaleBiO should therefore be read as an adapted reproduction under our single-H100 LoRA implementation rather than as a direct transcription of the original ScaleBiO tables. For dataset transparency, Tables 13 and 14 mirror the dataset-summary tables from Appendix B.2 of Pan et al. [2025]; the reported sizes, dataset types, and licenses follow their Tables 10 and 11, while the experimental protocol text below states where our adapted reproduction differs.

Baselines and compared methods. We compare against both non-bilevel data-selection methods and bilevel source-reweighting methods. UNIFORM WEIGHTING is the source-agnostic control: it samples the final fine-tuning set uniformly across the available sources (10K examples for instruction following and 20K examples for mathematical reasoning) and then uses the same final supervised fine-tuning and evaluation pipeline as every other method. LESS [Xia et al., 2024] is an influence-aware data-selection baseline for targeted instruction tuning; in our adapted pipeline, it scores candidate examples from the same source portfolio using gradient/influence alignment with the target validation data and then selects the highest-scoring examples for final fine-tuning. RHO-LOSS [Mindermann et al., 2022] is a reference-model-based prioritized training baseline; it ranks examples by a reducible-loss criterion that subtracts a reference model's estimated irreducible loss from the target model loss, then uses the resulting scores to form the final training set. SCALEBIO [Pan et al., 2025] is the main bilevel competitor: it learns source logits through the scalable first-order bilevel reweighting update proposed in that work, converts the learned simplex weights into source counts, and samples the final training set from those counts. HG-BASELINE takes one differentiable SGD lower-model step and differentiates the held-out loss through that step, without multi-branch selection. HG-MS uses the same one-step construction with $N > 1$, hard-selects the lowest-held-out-loss branch, and differentiates through that branch. Neither method uses a ridge–CG hyper-gradient in the real-world LLM code path.

ScaleBiO also reports SOBA [Dagréou et al., 2022] as a bilevel baseline, but it runs out of memory in their real-world LLM tables. We do not include SOBA as a main-table row in our single-H100 reproduction for the same practical reason: its auxiliary-state and Hessian-system bookkeeping is not competitive with the memory budget needed for LoRA fine-tuning of the target LLMs. This omission keeps the main comparison focused on methods that can complete the shared single-H100 final-training pipeline.

HG-MS implementation details for the real-world LLM runs. The real-world runs use Algorithm 3, with source logits θ as the upper variable and LoRA adapter weights as the lower state. Each branch starts from the current adapter state, with an independent Gaussian perturbation for branch indices $i > 0$, and takes exactly one differentiable SGD step on g using an independent minibatch/seed. All branches are evaluated on the same held-out objective f. The lowest-loss branch is hard-selected, its loss is differentiated through that SGD step, and Adam updates the source logits. No branch learning-rate multipliers or optimizer diversity are used. ULA/Gibbs, ridge–CG, and projected gradient are replaced rather than approximated.

The branch counts used in the reported real-world LLM experiments are fixed as follows. For the mathematical-reasoning experiments using the ten-source portfolio in Table 14, HG-MS uses $N = 4$ branches for the reported `Llama-3-8B`, `Qwen-2-7B`, and `Gemma-2-9B` rows; the one-branch hyper-gradient baseline uses $N = 1$. The large-model `Qwen2.5-32B` configuration uses $N = 4$ branches. For the instruction-following experiments using the eighteen-source portfolio in Table 13, the reported HG-MS runs use $N = 8$ branches and HG-BASELINE uses $N = 1$.

All iterative real-world methods initialize source logits at $\theta = 0$, so the initial source distribution is uniform unless a fixed source mixture is explicitly loaded. For the source-logit update, HG-MS and HG-BASELINE use Adam, while SCALEBIO uses SGD. This is separate from the lower-model update: for the real-world LLM experiments, HG-MS, HG-BASELINE, and SCALEBIO all use one SGD lower-model update per branch or inner step. The method-specific outer learning rates are reported in Table 15.

The corresponding lower-model optimizer settings are summarized in Table 16. Each iterative row uses one SGD lower-model update per branch or inner step; the table separates this lower-model learning rate from the source-logit learning rates in Table 15. For SCALEBIO, `scalebio_alpha` is the coefficient multiplying its first-order source-reweighting surrogate, not an outer learning rate.

In the real-world code paths, the ScaleBiO source-logit update uses the softmax source weights and a scaled difference between weighted and unweighted source training losses: in the functional implementation this is $\alpha(\texttt{train_loss_w} - \texttt{train_loss_u})$, while in the 32B FSDP implementation it is $\alpha\big((\mathrm{softmax}(\theta) \odot \ell_w)^\top \mathbf{1} - (\mathrm{softmax}(\theta) \odot \ell_u)^\top \mathbf{1}\big)$, followed by the configured source-logit optimizer and optional gradient clipping. We did not find additional normalization beyond the softmax source weights and this optional clipping.

Table 15: **LLM reweighting implementation hyperparameters.** Outer learning rates for source-logit updates in the real-world LLM runs.

Experiment	Method	Outer LR
Math, `Qwen-2-7B`	HG-MS/HG-BASELINE	0.05
Math, `Qwen-2-7B`	SCALEBIO	1.0
Math, `Llama-3-8B/Gemma-2-9B`	HG-MS/HG-BASELINE	0.05
Math, `Llama-3-8B/Gemma-2-9B`	SCALEBIO	0.1
Instruction	HG-MS	0.05
Instruction	HG-BASELINE	10.0
Instruction	SCALEBIO	0.01
Large-model `Qwen2.5-32B`	HG-MS/HG-BASELINE	0.03
Large-model `Qwen2.5-32B`	SCALEBIO	0.05

Table 16: **Lower-model optimizer settings in the real-world LLM runs.** All rows use one SGD lower-model update per branch or inner step.

Experiment	Method	Lower-model LR	Additional settings
Math, `Qwen-2-7B`	HG-MS/HG-BASELINE	10^{-5}	HG-MS: $N = 4$, branch noise 10^{-3}
Math, `Qwen-2-7B`	SCALEBIO	10^{-4}	`scalebio_alpha` $= 10$
Math, `Llama-3-8B`	HG-MS/HG-BASELINE	10^{-4}	HG-MS: $N = 4$, branch noise 10^{-5}
Math, `Llama-3-8B`	SCALEBIO	10^{-4}	`scalebio_alpha` $= 10$
Math, `Gemma-2-9B`	HG-MS/HG-BASELINE	10^{-4}	HG-MS: $N = 4$, branch noise 10^{-5}
Math, `Gemma-2-9B`	SCALEBIO	10^{-4}	`scalebio_alpha` $= 10$
Instruction	HG-MS/HG-BASELINE	2×10^{-4}	HG-MS: $N = 8$, branch noise 5×10^{-4}
Instruction	SCALEBIO	10^{-5}	`scalebio_alpha` $= 100$
Large-model `Qwen2.5-32B`	HG-MS/HG-BASELINE	8×10^{-6}	HG-MS: $N = 4$, branch noise 5×10^{-4}
Large-model `Qwen2.5-32B`	SCALEBIO	5×10^{-5}	`scalebio_alpha` $= 100$

Final inference and judging. Across methods in a given real-world table, final checkpoints are evaluated with the same prompts, decoding code, and answer-extraction or judging procedure. The reported GSM8K/MATH evaluations and MT-Bench answer-generation stage use one generated response per prompt with greedy decoding (`do_sample=false`), with no sampling, beam search, or best-of-N inference. The `BoN` terminology in the theory and algorithmic description refers to the minima-selection mechanism inside HG-MS, not to the downstream LLM inference protocol. For instruction following, model answers are generated once and then scored by the same GPT-4o MT-Bench judge [Zheng et al., 2023] with temperature zero.

Main mathematical-reasoning protocol (Table 2). The math benchmark uses the same ten-source portfolio as Pan et al. [2025]; the full source and validation-source inventory is reported in Table 14. The ten training sources are DART-Math, SlimOrca, UltraInteract, MathInstruct, Orca-Math, WizardLM, Magicoder, ShareGPT-Vicuna, GPTeacher, and GPT4-LLM-Cleaned. Each backbone in the main table is equipped with LoRA adapters on the attention projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`) and is run through the same single-H100 adapted pipeline. The math pipeline uses LoRA rank 16, LoRA scaling 32, dropout 0.05, maximum sequence length 512, and greedy generation-based evaluation. The target evaluation benchmarks are GSM8K [Cobbe et al., 2021], a grade-school math word-problem benchmark, and MATH [Hendrycks et al., 2021], a competition-style mathematical problem-solving benchmark. During reweighting, we cache 500 training examples per source and use a GSM8K-only validation path; after reweighting, we sample a final 20K-example training set from the learned source mixture and fine-tune for one epoch with learning rate 8×10^{-6}, micro-batch size 1, and gradient accumulation 32. The completed main-table runs evaluate the final checkpoint on the full GSM8K and MATH test sets. For the iterative methods, the outer-stage wall-clock budget is matched within each backbone and is approximately 9h for `Qwen-2-7B`, 12h for `Llama-3-8B`, and 15h for `Gemma-2-9B`.

The `Qwen2.5-32B` results in Table 3 use the same ten-source portfolio and final GSM8K/MATH evaluation target. The completed iterative rows apply the corresponding source-reweighting method to the 32B model using the same LoRA code path, SDPA attention, gradient checkpointing, and matched large-model reweighting setup; the completed UNIFORM row uses equal source counts. For

the iterative 32B rows, the outer reweighting budget was fixed in advance at 15h per method before final GSM8K/MATH test evaluation. Each completed row then uses the resulting source mixture to form the full 20K-example final fine-tuning set and evaluates on the full GSM8K/MATH test sets. The completed HG-MS row attains 92.58% on GSM8K and 53.32% on MATH, the best completed scores in both columns. RHO-LOSS and LESS are marked as OOM in this large-model setting and remain in the table for comparability with the ScaleBiO large-model layout.

Main instruction-following protocol (Table 3). The instruction-following benchmark uses the released eighteen-source ScaleBiO portfolio, whose source inventory is reported in Table 13: `AlpacaGPT4`, `ShareGPT4`, `SlimOrca`, `AlpacaChat`, `OpenOrcaGPT4`, `WildChat`, `LMSYS-Chat`, `GPTeacher`, `Airoboros`, and the nine translated Alpaca variants (`es`, `de`, `ja`, `ko`, `ru`, `it`, `fr`, `zh`, `pt`). All completed instruction-following runs use LoRA adapters [Hu et al., 2022] on `q_proj`, `k_proj`, `v_proj`, and `o_proj`, maximum sequence length 256, LoRA rank 4, scaling 8, LoRA dropout 0.05, and one epoch of final supervised fine-tuning with learning rate 8×10^{-6}, micro-batch size 1, and gradient accumulation 64. Gradient checkpointing is enabled for the iterative jobs when needed by the single-H100 memory budget. During reweighting, we cache up to roughly 3000 examples per source from the released ScaleBiO assets, use a final 10K-example sampled training set, and cap MT-Bench prompts at 1024 input tokens with up to 512 generated tokens. The iterative methods use the matched per-backbone wall-clock reweighting budgets stated above before final fine-tuning: approximately 9h for `Qwen-2-7B`, 12h for `Llama-3-8B`, and 15h for `Gemma-2-9B`. The final evaluation is MT-Bench [Zheng et al., 2023] on the standard 80 FastChat questions, using a GPT-4o judge. In our implementation, MT-Bench generations are produced once per finished run and judged offline in a separate pass so that all methods share the same prompt set and judge model. The standard FastChat MT-Bench questions, generated answers, and GPT-4o scores used for the final reported evaluation are never used for source reweighting, checkpoint selection, hyperparameter tuning, or early stopping; those choices use only the released ScaleBiO MT-Bench-style validation and tuning splits described below.

For the instruction-following table, the time-budget comparison is therefore fixed at the reweighting stage: SCALEBIO, HG-BASELINE, and HG-MS receive the same single-H100 outer-loop budget within each backbone: approximately 9h for `Qwen-2-7B`, 12h for `Llama-3-8B`, and 15h for `Gemma-2-9B`. This is followed by the same one-epoch final fine-tuning and the same MT-Bench generation/judging pipeline. The non-iterative baselines do not use this outer-loop budget because they do not run bilevel source reweighting.

Following the ScaleBiO setup, the instruction portfolio intentionally mixes broadly useful instruction/chat data with multilingual Alpaca variants that are less directly aligned with MT-Bench. Each source is filtered by the maximum sequence length before entering the cached reweighting pool. The released ScaleBiO MT-Bench-style validation split is used as the outer validation set for source reweighting, and a separate MT-Bench-style tuning split is used to select the best reweighting checkpoint and tune method hyperparameters. This matches the role of the MT-Bench-style reference and tuning data described in Appendix B.2 of Pan et al. [2025], while our final reported score is computed only after one response is generated for each of the standard FastChat MT-Bench questions and judged by GPT-4o.

The method-specific data use is as follows. UNIFORM samples the final 10K training examples uniformly across the eighteen sources. LESS and RHO-LOSS score examples from the same portfolio using their adapted influence/reference-model criteria and then form the final training set. SCALEBIO, HG-BASELINE, and HG-MS learn source logits during the matched per-backbone reweighting stage, convert the selected source distribution into integer source counts, sample the final 10K training examples accordingly, and then run the identical one-epoch LoRA fine-tuning and MT-Bench evaluation pipeline.

Final source-mixture audit for mathematical reasoning. The model-specific tables in this paragraph report the final mathematical-reasoning source proportions stored in the run metadata for the main LLM tables; see Tables 17 to 20. These diagnostics are separate from the downstream test scores: they are included to make the learned or selected source mixtures auditable. For UNIFORM, the mixture is fixed by construction, so we report the common per-source mass rather than an arbitrary top-three ranking. For LESS and RHO-LOSS, the entries are empirical source proportions in the selected final SFT set, not learned source logits. For SCALEBIO, HG-BASELINE, and HG-MS, the

Table 17: **Final mathematical-reasoning source mixtures for `Llama-3-8B`.** Each non-uniform row reports the three largest final source proportions over the ten-source math portfolio.

Method	Final source mixture
Uniform Weighting	uniform over ten sources, each 0.100
RHO-LOSS	SlimOrca 0.352; GPT4-LLM-Cleaned 0.202; WizardLM 0.142
LESS	Orca-Math 0.119; MathInstruct 0.117; DART-Math 0.104
ScaleBiO	Magicoder 0.829; SlimOrca 0.096; GPTeacher 0.044
HG-baseline	MathInstruct 0.809; SlimOrca 0.091; GPT4-LLM-Cleaned 0.005
HG-MS	MathInstruct 0.902; DART-Math 0.036; Orca-Math 0.010

Table 18: **Final mathematical-reasoning source mixtures for `Qwen-2-7B`.** Each non-uniform row reports the three largest final source proportions over the ten-source math portfolio.

Method	Final source mixture
Uniform Weighting	uniform over ten sources, each 0.100
RHO-LOSS	MathInstruct 0.299; SlimOrca 0.215; GPT4-LLM-Cleaned 0.133
LESS	MathInstruct 0.140; DART-Math 0.128; Orca-Math 0.125
ScaleBiO	MathInstruct 0.996; SlimOrca 0.001; GPT4-LLM-Cleaned 0.001
HG-baseline	MathInstruct 0.996; GPT4-LLM-Cleaned 0.003; SlimOrca 0.000
HG-MS	DART-Math 0.940; MathInstruct 0.056; Orca-Math 0.004

entries are the largest components of the source distribution used to sample the final SFT set. We focus the source-mixture audit on mathematical reasoning, where the source shifts are most directly connected to the GSM8K/MATH improvements; the saved instruction-following source weights are included in the released experiment artifacts.

Run diagnostics. To make the fixed-budget mathematical-reasoning comparison auditable, each iterative run stores the elapsed reweighting wall-clock time, the number of completed outer steps, and the learned source weights used to sample the final fine-tuning set. Table 21 gives the budget-sensitivity reporting table for the `Qwen-2-7B` math setting, and Table 22 reports the corresponding fixed-budget branch-count sensitivity check. We fill the completed diagnostic rows from the available run metadata.

Table 21 should be read as a budget ledger rather than as an additional benchmark with a new protocol. Each row uses the same `Qwen-2-7B` math-reasoning pipeline described above, changes only the outer reweighting budget for the iterative method, then performs the same one-epoch final fine-tuning and greedy GSM8K/MATH evaluation. The "Outer steps" column records how many source-weight updates fit into the wall-clock budget, while the "Top-3 learned source weights" column reports the largest entries of the final source distribution used to sample the final training set.

Interpretation of the budget ledger. The learned distributions differ substantially across methods. SCALEBIO and HG-BASELINE usually become nearly single-source solutions: at 9h both put almost all mass on MathInstruct (0.996), and the longer HG-BASELINE runs remain similarly concentrated on MathInstruct. In contrast, HG-MS moves to a different source mixture. Its main 9h run places most mass on DART-Math (0.940) while keeping a secondary MathInstruct component (0.056); at 15h and 21h it remains DART-Math dominated but assigns a larger secondary mass to Orca-Math (0.156 and 0.142, respectively). Thus the gain of HG-MS is not just another collapse to the same source preferred by SCALEBIO/HG-BASELINE; minima selection changes which training sources are emphasized in the final fine-tuning set.

The main `Qwen-2-7B` math row uses the 9h HG-MS run because that stopping budget was fixed before the final GSM8K/MATH test evaluation. The 15h and 21h rows were run under the same final fine-tuning and evaluation protocol to check sensitivity to a longer reweighting horizon. In this run family, extending the HG-MS budget does not improve final accuracy even though it continues to update the source mixture; this is a post-hoc diagnostic observation, not the criterion used to choose the main-table stopping budget.

Table 19: **Final mathematical-reasoning source mixtures for `Gemma-2-9B`.** Each non-uniform row reports the three largest final source proportions over the ten-source math portfolio.

Method	Final source mixture
Uniform Weighting	uniform over ten sources, each 0.100
RHO-LOSS	SlimOrca 0.267; GPT4-LLM-Cleaned 0.178; WizardLM 0.144
LESS	Orca-Math 0.110; MathInstruct 0.109; DART-Math 0.103
ScaleBiO	UltraInteract 0.859; Magicoder 0.083; SlimOrca 0.033
HG-baseline	MathInstruct 0.975; SlimOrca 0.004; GPT4-LLM-Cleaned 0.004
HG-MS	MathInstruct 0.941; DART-Math 0.016; Orca-Math 0.010

Table 20: **Final source mixtures for the `Qwen2.5-32B` mathematical-reasoning rows.** Completed non-`OOM` rows use the ten-source math portfolio; RHO-LOSS and LESS do not have completed large-model final mixtures under the reported memory budget.

Method	Final source mixture
Uniform Weighting	uniform over ten sources, each 0.100
RHO-LOSS	`OOM`; no completed final source mixture
LESS	`OOM`; no completed final source mixture
ScaleBiO	GPT4-LLM-Cleaned 0.166; SlimOrca 0.120; GPTeacher 0.105
HG-baseline	Orca-Math 0.159; UltraInteract 0.155; DART-Math 0.153
HG-MS	DART-Math 0.165; Orca-Math 0.135; MathInteract 0.115

Fixed-budget branch-count sensitivity for HG-MS. To probe the role of explicit minima selection, we add a fixed-budget branch-count sensitivity study for the `Qwen-2-7B` math-reasoning setting. The study keeps the ten-source portfolio, seed, LoRA configuration, GSM8K-only validation path, nine-hour reweighting budget, final 20K-example fine-tuning set, and full GSM8K/MATH evaluation fixed, while changing the number N of candidate lower-level branches inside HG-MS. Because the comparison is wall-clock matched, changing N does not isolate branch count alone: larger N gives a richer minima-selection set per outer update but leaves fewer source-weight updates within the same budget. The $N = 1$ row corresponds to the one-branch HG-BASELINE, the $N = 4$ row is the completed main-table HG-MS configuration, and the $N = 2$ and $N = 8$ rows test whether the observed gain comes from a moderate amount of branch diversity or from simply increasing branch count. We therefore report outer steps and interpret this single-seed table as a finite-budget diagnostic rather than as an exhaustive tuning study over N.

Table 21: **Budget-sensitivity ledger for `Qwen-2-7B` math reasoning.**

Method	Budget	Outer steps	Top-3 learned source weights	GSM8K / MATH
ScaleBiO	3h	10527	Magicoder 0.848; GPTeacher 0.037; WizardLM 0.026	66.19 / 30.70
HG-baseline	3h	1950	GPT4-LLM-Cleaned 0.565; MathInstruct 0.370; SlimOrca 0.025	61.41 / 29.97
HG-MS	3h	1233	MathInstruct 0.993; GPT4-LLM-Cleaned 0.002; SlimOrca 0.002	68.61 / 33.75
ScaleBiO	9h	13804	MathInstruct 0.996; SlimOrca 0.001; GPT4-LLM-Cleaned 0.001	75.74 / 36.93
HG-baseline	9h	6129	MathInstruct 0.996; GPT4-LLM-Cleaned 0.003; SlimOrca 0.000	68.46 / 34.27
HG-MS	9h	3754	DART-Math 0.940; MathInstruct 0.056; Orca-Math 0.004	**82.03 / 38.98**
ScaleBiO	15h	23039	Magicoder 0.980; DART-Math 0.010; MathInstruct 0.003	70.66 / 29.65
HG-baseline	15h	10150	MathInstruct 1.000; GPT4-LLM-Cleaned 0.000; SlimOrca 0.000	70.05 / 34.14
HG-MS	15h	6283	DART-Math 0.828; Orca-Math 0.156; MathInstruct 0.016	81.96 / 37.38
ScaleBiO	21h	32208	Magicoder 0.978; DART-Math 0.010; MathInstruct 0.005	69.96 / 30.73
HG-baseline	21h	14220	MathInstruct 0.992; GPT4-LLM-Cleaned 0.002; SlimOrca 0.001	69.73 / 34.09
HG-MS	21h	8813	DART-Math 0.830; Orca-Math 0.142; MathInstruct 0.026	81.98 / 37.28

Table 22: **Fixed-budget branch-count sensitivity for `Qwen-2-7B` math reasoning.** All rows use the same nine-hour reweighting budget and one-epoch final fine-tuning protocol; N controls the number of candidate branches per outer update, so larger N trades richer branch selection against fewer source-weight updates.

Method	Branches N	Top-3 learned source weights	Outer steps	GSM8K / MATH
HG-baseline	1	MathInstruct 0.996; GPT4-LLM-Cleaned 0.003; SlimOrca 0.000	6129	68.46 / 34.27
HG-MS	2	Orca-Math 0.492; DART-Math 0.423; MathInstruct 0.085	5211	76.65 / 31.96
HG-MS	4	DART-Math 0.940; MathInstruct 0.056; Orca-Math 0.004	3754	**82.03 / 38.98**
HG-MS	8	DART-Math 0.760; Magicoder 0.190; Orca-Math 0.008	2516	66.19 / 29.37

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state the theoretical contribution and the empirical scope. The empirical claims are qualified in Section 6 and Appendix A, with Appendix G.7 documenting the single-seed status of the real-world LLM tables.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: Appendix A discusses the structural assumptions behind the theory, the conservative nature of the analyzed oracle-complexity bound, and the distinction between the ULA/Gibbs analysis and the finite-budget optimizer branches used in experiments. Appendix G.7 further qualifies the real-world LLM evidence as single-seed matched-budget comparisons and gives protocol and budget-sensitivity diagnostics.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes].

   Justification: The theoretical assumptions are stated in the theorem and assumption statements in Sections 3–5. Complete derivations and proofs are provided in the appendices, with cross-references from the main text.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: Section 6 and the experimental appendices describe the datasets, source portfolios, hyperparameters, budgets, evaluation procedures, and random seeds used for the reported experiments. The LLM protocol appendix also records branch counts, final fine-tuning sizes, decoding settings, and judging procedure, and Appendix G links to the GitHub code/configuration repository. Appendix G.5 specifies the exact synthetic objectives, frozen settings, pilot/held-out separation, and final 20-seed protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     I If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     II If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

III If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

IV We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: The paper identifies the public datasets and provides the experimental protocol. Appendix G links to a GitHub repository containing the experimental code and configuration files, while the appendices document the data splits, source portfolios, budgets, hyperparameters, and evaluation procedures needed to reproduce the reported results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The main experimental section gives the benchmark definitions and headline settings, while the appendices provide training budgets, source inventories, LoRA settings where applicable, candidate-generator/lower-optimizer choices, validation/evaluation procedures, seeds, and method-specific hyperparameters. Appendix G also links to a GitHub repository containing the experimental code and configuration files.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: The MNIST experiments report mean and 95% confidence intervals over independent runs. For the real-world LLM tables, Appendix G reports evaluation-set uncertainty estimates using Wilson binomial intervals for GSM8K/MATH and bootstrap intervals for MT-Bench, while explicitly noting that the LLM comparisons are single-seed and that these intervals do not estimate rerun variance. The controlled GPT-2 source-reweighting table is reported as a fixed-configuration diagnostic and is not used for statistical-significance claims. The synthetic comparison reports symmetric bootstrap 95% intervals over 20 held-out seeds; the separate ULA/Gibbs-reference sweeps report bootstrap percentile intervals over seed means, with both conventions defined in Appendix G.5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: The experimental section and appendices report the single-GPU setup, H100/bf16 LoRA path for the real-world LLM runs, H100 fixed-budget MNIST runs, wall-clock budgets, outer-step counts, and final fine-tuning/evaluation budgets for the main experiments. Appendix G.5 additionally specifies the synthetic single-thread Apple Accelerate implementation and 60-second budget for each of 400 held-out runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work uses standard public or released research datasets and benchmarks and does not introduce human-subject studies, crowdsourcing, or new high-risk datasets. The paper cites the relevant assets and reports dataset licenses where used in the LLM experiments.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No].

    Justification: The paper focuses on theory and controlled/source-reweighting experiments and does not include a dedicated broader-impact discussion. The main foreseeable impact is methodological: more efficient or targeted hyperparameter/data reweighting for ML and LLM fine-tuning.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release a new high-risk pretrained model, image generator, or scraped dataset. The LLM experiments fine-tune existing open/released models and datasets for evaluation of the proposed reweighting method.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: The paper cites the datasets, models, and baselines used in the experiments. Appendix G.7 includes source-portfolio tables with dataset kinds and licenses for the LLM source-reweighting experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: The new released asset is the experimental code/configuration package linked in Appendix G. The paper does not introduce a new dataset or pretrained model; the code asset is documented by the repository contents and the experimental-protocol appendices.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A].

Justification: The paper does not include crowdsourcing experiments or direct research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A].

Justification: The paper does not involve human-subject research, so IRB or equivalent approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A].

Justification: The proposed method does not use an LLM or agent as an important, original, or non-standard component of the core algorithm. LLMs are used as experimental lower-level models/evaluation targets in the source-reweighting experiments, and those experimental uses are described in Section 6 and Appendix G.7.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.