Riemannian Bilevel Optimization under the Polyak–Łojasiewicz Condition

Zhixuan Li[1], **Yuyang Zhang**[1], **Luke Jones**[1], **Muztaba Syed**[1], **William Chang**[1], **Andi Han**[2]

[1]University of California, Los Angeles [2]The University of Sydney

chang314@g.ucla.edu andi.han@sydney.edu.au

Abstract

This paper studies bilevel optimization on Riemannian manifolds where the upper-level objective is nonconvex and the lower-level problem satisfies a Riemannian Polyak–Łojasiewicz condition rather than geodesic strong convexity. The classical hypergradient formula then breaks down, since the lower-level Hessian may be singular or indefinite away from the minimizer. We address this with an intrinsic regularized tangent-space formulation based on spectral clipping, and develop a Riemannian bilevel algorithm that avoids Hessian-inverse solves and inner-loop differentiation. We establish an $O(1/T)$ rate on the averaged squared Riemannian gradient mapping and $O(\varepsilon^{-1})$ iteration complexity, and extend the method to the stochastic finite-sum setting. Experiments on Stiefel, Grassmannian, and Poincaré-ball problems with rank-deficient Hessians (BCI IV-2a, UCI Superconductivity) and Stiefel-constrained meta-learning on MiniImageNet show that the proposed method is the only one whose linear-system error stays at machine precision and whose Riemannian gradient norm decreases steadily, while Hessian-inversion and Neumann-series baselines fail to descend.

1 Introduction

Bilevel optimization studies hierarchical decision problems in which the upper-level objective depends on the solution of a lower-level problem. Since its early appearance in mathematical programming and model selection [10, 44, 93, 19, 64], bilevel optimization has become a central framework in modern machine learning, with influential applications in meta-learning, hyperparameter optimization, neural architecture search, reinforcement learning, and signal processing [28, 31, 74, 27, 63, 29, 50]. In this paper, we study the Riemannian bilevel problem

$$\min_{x\in\mathcal{M}_x} \Phi(x) := f(x, y^\star(x)) + \phi(x) \quad \text{such that} \quad y^\star(x) \in \arg\min_{y\in\mathcal{M}_y} g(x, y), \tag{1}$$

where $\mathcal{M}_x$ and $\mathcal{M}_y$ are Riemannian manifolds, f and g are smooth, and ϕ is an optional regularizer.

Our interest in the Riemannian setting is motivated by the growing role of non-Euclidean parameter spaces in optimization and learning. Many structured models are naturally posed on manifolds such as spheres, Stiefel manifolds, Grassmann manifolds, and the manifold of symmetric positive definite (SPD) matrices, with applications ranging from geometry-aware representation learning and SPD networks to optimal transport and manifold-valued statistical learning [2, 8, 6, 9, 60, 26, 3, 24, 66, 62, 4, 5, 45, 49, 57, 18, 72, 96]. In particular, manifold-constrained learning has already appeared in Riemannian meta-learning, SPD neural architecture search, generalized Bures–Wasserstein learning, and manifold optimal transport [98, 97, 42, 39, 82, 20, 21, 87, 34].

The rapid development of Riemannian min–max and variational methods [55, 59, 41, 40, 103, 51, 81, 11] further motivates a bilevel theory intrinsic to manifold geometry.

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

Our focus is the regime in which the upper-level objective is nonconvex while the lower-level problem satisfies a Riemannian Polyak–Łojasiewicz (PL) condition rather than geodesic strong convexity. This direction is natural from both optimization and modeling viewpoints. In the Euclidean space, the PL inequality is a well-established relaxation of strong convexity that still yields strong convergence guarantees [1]. The inequality also appears in recent structured learning models such as monotone, bi-Lipschitz, and PL networks [102]. However, the transition from strong convexity to PL geometry creates a substantial gap in current Riemannian bilevel theory. Lower-level problems that satisfy the Riemannian PL condition without geodesic strong convexity arise naturally: (a) least-squares and interpolation problems with rank-deficient designs, whose Hessian is a Gram matrix, whose PL constant is its smallest nonzero eigenvalue, and whose minimizer set is affine (our template (16)); (b) objectives invariant under a symmetry acting on the lower-level variable (rotations, permutations, or gauge freedom such as quotient lifts), whose minimizers form orbits, which rules out geodesic strong convexity while PL/Morse–Bott can hold transversally; (c) the PL models of [102]. Rank deficiency alone does not imply PL: unregularized cross-entropy heads on fixed features have a flat feature-Gram kernel (dimension at least 1900 of 2000 in our $\lambda = 0$ sweep) but, on a separable support set, no finite minimizer; we use them only as a stress test outside the theory (Section 4). We include additional related works in Appendix A.

Our contributions. In this work, **(i)** we study nonconvex bilevel optimization on Riemannian manifolds under a lower-level PL condition. We introduce an intrinsic regularized tangent-space system via spectral clipping of the Riemannian Hessian to deal with singular or indefinite Hessian along the iterates. The lower-level class is broader than the isolated nondegenerate minimizers of [54]: once the minimizer set has positive dimension, the target becomes a pseudoinverse hypergradient. Our contribution is a kernel-preserving clipped system that targets this pseudoinverse inside a single-loop Hessian/Jacobian-free tracking scheme, together with its Riemannian analysis. **(ii)** We propose a Riemannian bilevel algorithm (R-HJFBiO) that combines manifold gradient updates, transported auxiliary variables, and an operator-level Hessian approximation built from intrinsic finite-difference probes (or from a closed-form Hessian when one is available). The method avoids both Hessian-inverse linear-system solves and differentiation through the inner loop; we use "Hessian/Jacobian-free" in this specific sense, clarified in Remark 12. **(iii)** We derive iteration-complexity guarantees of the proposed method under deterministic and stochastic settings, with constants scaling polynomially in the curvature factor and the transport-distortion constant. **(iv)** We evaluate the clipped operator empirically on Stiefel, Grassmann, and Poincaré-ball geometries with deliberately rank-deficient lower-level Hessians, where Hessian-inversion baselines diverge or stall; Section 4 states which parts of the experiments lie outside the theorem's hypotheses.

2 Preliminary

Riemannian geometry preliminaries. A Riemannian manifold $\mathcal{M}$ is a smooth manifold equipped with a smoothly varying inner product $\langle \cdot, \cdot \rangle_z$ on each tangent space $T_z\mathcal{M}$, with induced norm $\|u\|_z := \sqrt{\langle u, u \rangle_z}$. The exponential map $\mathrm{Exp}_z(u) = c(1)$ takes the geodesic c with $c(0) = z$, $c'(0) = u$, and on a totally normal neighborhood it has a smooth inverse with $d(x, y) = \|\mathrm{Exp}_x^{-1}(y)\|_x$. Parallel transport $\Gamma_{z_1 \to z_2}$ along the unique minimizing geodesic is a linear isometry. For differentiable $f : \mathcal{M} \to \mathbb{R}$, the Riemannian gradient $\mathcal{G}f(z) \in T_z\mathcal{M}$ satisfies $\langle \mathcal{G}f(z), u \rangle_z = Df(z)[u]$, and the Riemannian Hessian $\mathrm{Hess}\, f(z)$ is the covariant derivative of the gradient. A function f is μ-geodesically strongly convex iff $\mathrm{Hess}\, f \succeq \mu\, \mathrm{id}$; we use the weaker Polyak–Łojasiewicz inequality (Assumption 2), which still yields linear convergence [1, 102].

Notation. Following [43], we write $\mathcal{G}\psi$ for the Riemannian gradient so that gradients and the cross-derivatives below share one alphabet; ∇ is reserved for Euclidean gradients in the experiments, and the gradient mapping $\mathcal{G}(x, w, \gamma)$ of Section 3 is always written with its arguments. For $\psi : \mathcal{M}_x \times \mathcal{M}_y \to \mathbb{R}$, write $\mathcal{G}_x\psi, \mathcal{G}_y\psi$ for the partial Riemannian gradients and $\mathrm{Hess}_x\psi, \mathrm{Hess}_y\psi$ for the corresponding Hessians. The cross-derivatives $\mathcal{G}^2_{xy}\psi(x, y) : T_y\mathcal{M}_y \to T_x\mathcal{M}_x$ and $\mathcal{G}^2_{yx}\psi(x, y) : T_x\mathcal{M}_x \to T_y\mathcal{M}_y$ are defined by $\mathcal{G}^2_{xy}\psi(x, y)[v] = D_y\mathcal{G}_x\psi(x, y)[v]$ and are mutually adjoint: $\mathcal{G}^2_{yx}\psi = \left(\mathcal{G}^2_{xy}\psi\right)^*$. For a linear $T : T_x\mathcal{M}_x \to T_y\mathcal{M}_y$, the adjoint satisfies $\langle T[u], v \rangle_y = \langle T^*[v], u \rangle_x$, with operator norm $\|T\|_{x \to y} := \sup_{\|u\|_x = 1} \|T[u]\|_y$.

Geometric setup. We work on neighborhoods $\mathcal{U}_x \subseteq \mathcal{M}_x$ and $\mathcal{U}_y \subseteq \mathcal{M}_y$ that contain all iterates, all lower-level solutions $y^\star(x)$, and all geodesic segments used in the analysis, and which are geodesically convex and totally normal (e.g., geodesic balls of radius below the convexity radius; see [8]). On such neighborhoods Exp^{-1}, the Riemannian distance, and parallel transport are well-defined and unique, and any geodesic triangle with vertices in $\mathcal{U}_y$ has all sides in $\mathcal{U}_y$ – a property used for (2) and (3) below. The geodesic convexity here is a property of the *domain* (geodesics stay inside) and is unrelated to convexity of g, which satisfies only the PL inequality of Assumption 2.

Curvature constants. Two standard constants handle curvature in the analysis [105, 8, 43]: with sectional curvatures of $\mathcal{U}_y$ bounded below by $\kappa_y \leq 0$ and $\mathrm{diam}(\mathcal{U}_y) \leq D_y$, set $\zeta_y := \sqrt{|\kappa_y|}\, D_y \coth\Big(\sqrt{|\kappa_y|}\, D_y\Big)$ when $\kappa_y < 0$ and $\zeta_y := 1$ when $\kappa_y \geq 0$; then

$$d(y_b, y_c)^2 \leq \zeta_y \big(d(y_a, y_b)^2 + d(y_a, y_c)^2\big) - 2\big\langle \mathrm{Exp}_{y_a}^{-1}(y_b), \mathrm{Exp}_{y_a}^{-1}(y_c)\big\rangle_{y_a} \tag{2}$$

for any geodesic triangle in $\mathcal{U}_y$. Define ζ_x analogously and $\zeta := \max\{\zeta_x, \zeta_y\}$; in Euclidean space $\zeta = 1$ and (2) reduces to the parallelogram identity. Second, there exists a transport-distortion constant $C_0 \geq 0$, depending only on κ_y and D_y, such that

$$\big\|\Gamma_{y_2 \to y_3} \circ \Gamma_{y_1 \to y_2} u - \Gamma_{y_1 \to y_3} u\big\|_{y_3} \leq C_0\, d(y_1, y_2)\, d(y_2, y_3)\, \|u\|_{y_1} \tag{3}$$

for all $y_1, y_2, y_3 \in \mathcal{U}_y$ and $u \in T_{y_1}\mathcal{M}_y$, measuring the failure of composed transports to match the direct transport [43, Lemma 2] ($C_0 = 0$ in Euclidean space).

Assumption 1 (Bounded sectional curvature and bounded diameter)**.** *The neighborhoods $\mathcal{U}_x$ and $\mathcal{U}_y$ have sectional curvatures bounded below by $\kappa_x, \kappa_y \in \mathbb{R}$ and finite diameters $D_x, D_y < \infty$. We further assume:*

(a) (Geodesic convexity.) *$\mathcal{U}_x, \mathcal{U}_y$ are geodesically convex and totally normal and contain all iterates (a small diameter alone does not imply convexity; geodesic balls of radius below the convexity radius are an example).*

(b) (Finite diameter.) *D_x, D_y are finite, so the constants $\zeta \geq 1$ and $C_0 \geq 0$ from (2)–(3) are finite.*

Clause (a) makes Exp^{-1}, parallel transport and the proximal step well defined; clause (b) keeps ζ and C_0 finite. Neither is automatic, even on Hadamard manifolds, since the iterates must remain in a bounded convex set. The Euclidean analysis of [54] corresponds to $\zeta = 1$ and $C_0 = 0$ (Remark 25).

We now state the core assumption on the lower-level objective $g(x, y)$.

Assumption 2 (Riemannian PL condition of the lower-level problem [41, Definition 2.3])**.** *There exists $\mu > 0$ such that, for every $x \in \mathcal{U}_x$,*

$$\tfrac{1}{2}\|\mathcal{G}_y g(x, y)\|_y^2 \geq \mu\big(g(x, y) - g(x, y^\star(x))\big) \qquad \text{for all } y \in \mathcal{U}_y, \tag{4}$$

where $y^\star(x) \in \arg\min_{y \in \mathcal{U}_y} g(x, y)$.

Although Assumption 2 is purely first-order, it has sharp second-order consequences on the lower-level minimizer set $\mathcal{M}^\star(x) := \arg\min_{y \in \mathcal{U}_y} g(x, y)$. By [89], whenever $g(x, \cdot)$ is C^2, the μ-PL inequality implies a Morse–Bott condition with the same constant μ around $\mathcal{M}^\star(x)$ (conversely, μ-Morse–Bott implies μ'-PL locally for every $\mu' < \mu$): the minimizer set is a C^1 embedded submanifold of $\mathcal{U}_y$, the kernel of $\mathrm{Hess}_y\, g(x, y^\star)$ coincides with $T_{y^\star}\mathcal{M}^\star(x)$ at every $y^\star \in \mathcal{M}^\star(x)$, and the Hessian is uniformly bounded between μ and L_g on the orthogonal complement $N_{y^\star}\mathcal{M}^\star(x)$. To turn this geometry into a usable hypergradient we will additionally require $\mathcal{G}_y f(x, y^\star) \in \mathrm{range}(\mathrm{Hess}_y\, g(x, y^\star))$ (Assumption 5), so that the Moore–Penrose pseudoinverse $\mathrm{Hess}_y\, g(x, y^\star)^\dagger\, \mathcal{G}_y f(x, y^\star)$ recovers $\mathcal{G}F(x)$ exactly even when the Hessian is singular along the tangent directions. The second-order structure used throughout the analysis is summarized in the following lemma.

Lemma 3 (Riemannian Morse–Bott condition at the lower-level minimizer set)**.** *For every $x \in \mathcal{U}_x$, the lower-level solution set $\mathcal{M}^\star(x) := \arg\min_{y \in \mathcal{U}_y} g(x, y)$ is a C^1 embedded submanifold of $\mathcal{U}_y$, and there exist constants $L_g \geq \mu > 0$ such that for every $y^\star \in \mathcal{M}^\star(x)$ the lower-level Hessian satisfies the μ-Morse–Bott property in the sense of [89]:*

(i) $\ker\big(\mathrm{Hess}_y\, g(x, y^\star)\big) = T_{y^\star}\mathcal{M}^\star(x)$;

(ii) $\mu\,\mathrm{id} \preceq \mathrm{Hess}_y\, g(x,y^\star)\big|_{N_{y^\star}\mathcal{M}^\star(x)} \preceq L_g\,\mathrm{id}$, *where* $N_{y^\star}\mathcal{M}^\star(x) := \big(T_{y^\star}\mathcal{M}^\star(x)\big)^\perp$.

Remark 4 (Morse–Bott is strictly weaker than local strong convexity)**.** *Lemma 3 does not impose local strong convexity at $y^\star(x)$ and does not require $\mathrm{Hess}_y\, g(x,y^\star)$ to be invertible: clause (ii) asks for a μ-lower bound only on the normal subspace $N_{y^\star}\mathcal{M}^\star(x)$, while (i) identifies the kernel with $T_{y^\star}\mathcal{M}^\star(x)$, so whenever $\dim \mathcal{M}^\star(x) > 0$ the Hessian is genuinely singular at every minimizer. By [89], for C^2 functions clauses (i)–(ii) are implied by Assumption 2 with the same constant μ (the converse holds with any $\mu' < \mu$); thus Lemma 3 is the second-order content of Assumption 2, not an additional hypothesis. By contrast, the compatibility condition below is an additional hypothesis on (f,g): it is automatic when $\dim \mathcal{M}^\star(x) = 0$ (in particular under geodesic strong convexity and in the setting of [54]), and a genuine restriction exactly when the minimizer set has positive dimension, where it plays the role that uniqueness plays under strong convexity: it makes $F(x) = f(x,y^\star(x))$ independent of the selection (Lemma 28). Minimal examples and the scope of our results without it are given in Appendix B.*

Assumption 5 (Bilevel compatibility)**.** *For every $x \in \mathcal{U}_x$ and every $y^\star \in \mathcal{M}^\star(x)$, $\mathcal{G}_y f(x,y^\star) \in \mathrm{range}\big(\mathrm{Hess}_y\, g(x,y^\star)\big)$; equivalently (by Lemma 3(i)), $\mathcal{G}_y f(x,y^\star) \perp T_{y^\star}\mathcal{M}^\star(x)$; equivalently, $f(x,\cdot)$ is constant on each connected component of $\mathcal{M}^\star(x)$. We further assume that $\mathcal{M}^\star(x)$ is connected for every $x \in \mathcal{U}_x$ (otherwise all statements below refer to the connected component containing the selected minimizer).*

Assumption 6 (First-order regularity and boundedness)**.** *There exist constants $C_{fy}, C_{gxy}, L_f, L_g > 0$ such that for all $(x,y) \in \mathcal{U}_x \times \mathcal{U}_y$, $\|\mathcal{G}_y f(x,y)\|_y \le C_{fy}$ and*

$\|\mathcal{G}^2_{xy} g(x,y)\|_{y\to x} \le C_{gxy}$, and the partial Riemannian gradients of f and g are Lipschitz continuous in the cross-tangent-space sense: for all $x, x_1, x_2 \in \mathcal{U}_x$ and $y, y_1, y_2 \in \mathcal{U}_y$, $\|\Gamma_{y_1\to y_2}\mathcal{G}_y f(x,y_1) - \mathcal{G}_y f(x,y_2)\|_{y_2} \le L_f\, d(y_1,y_2)$, and similarly for the three remaining partial-gradient/Lipschitz pairs (full statement deferred to Appendix B).

Assumption 7 (Second-order regularity)**.** *There exist constants $L_{gxy}, L_{gyy} > 0$ such that the cross-derivative $\mathcal{G}^2_{xy} g$ and the lower-level Hessian $\mathrm{Hess}_y g$ are Lipschitz continuous on $\mathcal{U}_x \times \mathcal{U}_y$ in the cross-tangent-space sense, with constants L_{gxy} and L_{gyy}. The full four-line statement (in x and y, for both operators) is deferred to Appendix B.*

Assumption 8 (Upper-level regularizer and lower-bounded objective)**.** *The upper-level regularizer $\phi : \mathcal{M}_x \to (-\infty, +\infty]$ is proper, lower semicontinuous, and geodesically convex on $\mathcal{U}_x$. The composite upper-level objective $\Phi(x) := F(x) + \phi(x)$ with $F(x) := f(x,y^\star(x))$ is bounded below on $\mathcal{U}_x$: $\Phi^\star := \inf_{x\in\mathcal{U}_x} \Phi(x) > -\infty$.*

Since ϕ is geodesically convex only on $\mathcal{U}_x$, the proximal subproblem (11) is posed over $u \in \mathrm{Exp}_x^{-1}(\mathcal{U}_x)$; as part of the standing requirement that all iterates remain in $\mathcal{U}_x$, we assume that its infimum is attained there (for $\phi \equiv 0$ the minimizer is $u = -\gamma w$). Remark 23 (Appendix C) compares each assumption with its counterpart in [43, 69, 54].

3 Methods and Convergence Guarantees

With the problem and assumptions fixed in Section 2, we now design and analyze our solution method. Since $\Phi(x) = f(x,y^\star(x)) + \phi(x)$ depends only on x, the natural strategy is Riemannian (proximal) gradient descent on Φ: $x_{t+1} = \mathrm{Exp}_{x_t}\big(-\eta\,\mathcal{G}F(x_t)\big)$, where $F(x) := f(x,y^\star(x))$ and $\mathcal{G}F$ is the *hypergradient*. The hypergradient is more delicate than an ordinary Riemannian gradient: since $y^\star(x)$ is implicit, the chain rule must be propagated through the optimality conditions of $g(x,\cdot)$, introducing couplings via cross-derivatives, the lower-level Hessian, and parallel transport. The next proposition — the Riemannian analogue of the classical implicit-differentiation identity [33] — makes this precise.

Proposition 9 (Hypergradient; pseudoinverse extension of [43])**.** *For every $x \in \mathcal{U}_x$, $\mathcal{G}F(x) = \mathcal{G}_x f(x,y^\star(x)) - \mathcal{G}^2_{xy} g(x,y^\star(x))\big[\mathrm{Hess}_y\, g(x,y^\star(x))^\dagger \mathcal{G}_y f(x,y^\star(x))\big]$.*

Under the geodesic strong convexity assumed by [43], $\mathrm{Hess}_y\, g(x,y^\star)$ is invertible and $\dagger$ may be read as the inverse. The same formula is asserted here under the strictly weaker Morse–Bott PL hypothesis of Lemma 3, where $\mathrm{Hess}_y\, g(x,y^\star)$ is singular whenever $\dim \mathcal{M}^\star(x) \ge 1$ and $\dagger$ must be read as the Moore–Penrose pseudoinverse on $T_{y^\star}\mathcal{M}_y$. The compatibility condition (Assumption 5),

$\mathcal{G}_y f(x, y^\star) \in \mathrm{range}(\mathrm{Hess}_y\, g(x, y^\star))$, ensures the right-hand side is well-defined and equals $\mathcal{G}F(x)$ exactly even in the singular case; the derivation, which holds for every smooth selection and at every point of $\mathcal{M}^\star(x)$, is given in Lemma 28(i) under a parametric regularity hypothesis on the solution sets stated there. Away from $\mathcal{M}^\star(x)$, $\mathrm{Hess}_y\, g$ may be indefinite, so the formula is not available along the trajectory, which motivates the regularized construction in the following. We address this with a regularized tangent-space construction. Note that $\big(\mathrm{Hess}_y\, g(x, y^\star)\big)^\dagger \mathcal{G}_y f(x, y^\star) \in T_{y^\star}\mathcal{M}_y$, and $\mathcal{G}^2_{xy} g(x, y^\star)$ then maps this $T_{y^\star}\mathcal{M}_y$-vector to $T_x\mathcal{M}_x$, yielding $\mathcal{G}F(x) \in T_x\mathcal{M}_x$.

Regularized tangent-space system. Since $\mathrm{Hess}_y\, g(x, y) : T_y\mathcal{M}_y \to T_y\mathcal{M}_y$ is self-adjoint, let $\{(\lambda_i, e_i)\}_{i=1}^{d_y}$ be an orthonormal eigenbasis. For a clip floor $0 < \mu_c \le \mu$, define the *kernel-preserving spectral clipping operator*

$$\mathcal{S}_{[\mu_c, L_g]}[\mathrm{Hess}_y\, g(x, y)][v] := \sum_{i\,:\,|\lambda_i| \ge \mu_c} \mathrm{clip}(\lambda_i; \mu_c, L_g)\, \langle e_i, v\rangle_y\, e_i, \qquad v \in T_y\mathcal{M}_y, \tag{5}$$

where $\mathrm{clip}(\lambda; \mu_c, L_g) := \min\{\max\{|\lambda|, \mu_c\}, L_g\}$ acts on magnitudes, so $\mathcal{S}_{[\mu_c, L_g]}[\cdot]$ is positive semidefinite with spectrum in $\{0\} \cup [\mu_c, L_g]$. Eigenvalues with $|\lambda_i| < \mu_c$ are mapped to zero rather than μ_c, preserving the Hessian's kernel structure along its small-eigenvalue directions. The floor μ_c is decoupled from the PL constant μ, which enters only through PL and quadratic growth (we use $\mu_c = 10^{-3}$, and $\mu_c = 10^{-2}$ on UCI Superconductivity).

Remark 10 (Why kernel-preserving)**.** *Hard clipping (raising magnitudes below μ_c to μ_c, then inverting) and* (5) *agree on the retained eigenvectors. Under Assumption 5 and $\mu_c \le \mu$, both return $\mathrm{Hess}_y\, g^\dagger \mathcal{G}_y f$ at a lower-level minimizer, and their difference vanishes as $y_t \to \mathcal{M}^\star(x)$. They differ away from $\mathcal{M}^\star(x)$, under noise in $\mathcal{G}_y f$, or without compatibility, where hard clipping amplifies near-kernel components of $\mathcal{G}_y f$ by up to $1/\mu_c$ while* (5) *discards them; moreover, only* (5) *reproduces $\mathrm{Hess}_y\, g(x, y^\star)$ exactly as an operator (Proposition 11).*

Define $v^\star(x, y) \in T_y\mathcal{M}_y$ as the minimum-norm solution of

$$v^\star(x, y) = \arg\min\big\{\|v\|_y^2 \;:\; \mathcal{S}_{[\mu_c, L_g]}[\mathrm{Hess}_y\, g(x, y)]\, v = \mathsf{P}_{\mathrm{rng}}(x, y)\, \mathcal{G}_y f(x, y)\big\}, \tag{6}$$

where $\mathsf{P}_{\mathrm{rng}}(x, y)$ is the orthogonal projector onto $\mathrm{range}\big(\mathcal{S}_{[\mu_c, L_g]}[\mathrm{Hess}_y\, g(x, y)]\big)$. Equivalently,

$$v^\star(x, y) = \big(\mathcal{S}_{[\mu_c, L_g]}[\mathrm{Hess}_y\, g(x, y)]\big)^\dagger\, \mathcal{G}_y f(x, y), \tag{7}$$

where $A^\dagger$ is the Moore–Penrose pseudoinverse on $T_y\mathcal{M}_y$. Since the spectrum of $\mathcal{S}_{[\mu_c, L_g]}[\mathrm{Hess}_y\, g(x, y)]$ on its range lies in $[\mu_c, L_g]$, the pseudoinverse has spectral norm at most $1/\mu_c$, giving $\|v^\star(x, y)\|_y \le \mu_c^{-1} \|\mathcal{G}_y f(x, y)\|_y$. Problem (6) has a unique solution even when $\mathrm{Hess}_y\, g(x, y)$ is singular or indefinite. Because clip jumps at $|\lambda| = \mu_c$, our analysis assumes that the lower-level spectra visited by the algorithm stay a margin away from μ_c (Assumption 27, Appendix G); near $\mathcal{M}^\star(x)$ this holds automatically with $\mu_c = \mu/2$ (Remark 26), and we check it along our deterministic runs (Table 5).

Proposition 11 (Consistency at the lower-level minimizer)**.** *Under Lemma 3, for every $y^\star \in \mathcal{M}^\star(x)$, $\mathcal{S}_{[\mu_c, L_g]}[\mathrm{Hess}_y\, g(x, y^\star)] = \mathrm{Hess}_y\, g(x, y^\star)$, and $v^\star(x, y^\star) = \mathrm{Hess}_y\, g(x, y^\star)^\dagger\, \mathcal{G}_y f(x, y^\star)$.*

Remark 24 (Appendix C) relates (6) and Proposition 11 to their Euclidean counterparts.

Using $v^\star(x, y)$ in place of the inverse-Hessian expression yields the intrinsic hypergradient estimator

$$\widehat{\mathcal{G}}_x F(x, y) = \mathcal{G}_x f(x, y) - \mathcal{G}^2_{xy} g(x, y)\big[v^\star(x, y)\big], \tag{8}$$

which is well-defined under PL geometry and, by Proposition 11, satisfies $\widehat{\mathcal{G}}_x F(x, y^\star(x)) = \mathcal{G}F(x)$ exactly. To avoid differentiating through the lower-level solver, we approximate the Jacobian-vector product and the Hessian operator using intrinsic finite-difference probes of $\mathcal{G}g$:

$$\widetilde{J}_R(x, y, v, \delta_\varepsilon) := \frac{\mathcal{G}_x g\big(x, \mathrm{Exp}_y(\delta_\varepsilon v)\big) - \mathcal{G}_x g\big(x, \mathrm{Exp}_y(-\delta_\varepsilon v)\big)}{2\delta_\varepsilon}, \qquad v \in T_y\mathcal{M}_y, \tag{9}$$

$$\widetilde{\mathcal{H}}_R(x, y, v, \delta_\varepsilon) := \frac{\Gamma_{\mathrm{Exp}_y(\delta_\varepsilon v) \to y} \mathcal{G}_y g\big(x, \mathrm{Exp}_y(\delta_\varepsilon v)\big) - \Gamma_{\mathrm{Exp}_y(-\delta_\varepsilon v) \to y} \mathcal{G}_y g\big(x, \mathrm{Exp}_y(-\delta_\varepsilon v)\big)}{2\delta_\varepsilon}. \tag{10}$$

The expression $\widetilde{J}_R$ is a tangent vector approximating $\mathcal{G}^2_{xy} g(x, y)[v]$; the expression $\widetilde{\mathcal{H}}_R$ is a tangent vector approximating $\mathrm{Hess}_y\, g(x, y)[v]$.

Algorithm 1 Riemannian Hessian/Jacobian-free Bilevel Optimization (R-HJFBiO)

Require: T; stepsizes $\lambda > 0$, $\gamma > 0$, $\tau > 0$; clip floor $0 < \mu_c \le \mu$ and ceiling L_g; finite-difference parameter $\delta_\varepsilon > 0$; radius $r_v \ge C_{fy}/\mu_c$; initial $x_1 \in \mathcal{M}_x$, $y_1 \in \mathcal{M}_y$, and $v_1 \in T_{y_1}\mathcal{M}_y$ with $\|v_1\|_{y_1} \le r_v$.

1: **for** $t = 1, 2, \dots, T$ **do**

2: **(Lower-level step)** $u_t \leftarrow \mathcal{G}_y g(x_t, y_t) \in T_{y_t}\mathcal{M}_y$

3: $y_{t+1} \leftarrow \mathrm{Exp}_{y_t}(-\lambda u_t)$ ▷ lines 4–9 use the current point (x_t, y_t, v_t)

4: **(Jacobian-free cross term)** $\widetilde{J}_t \leftarrow \widetilde{J}_R(x_t, y_t, v_t, \delta_\varepsilon)$ by (9)

5: $w_t \leftarrow \mathcal{G}_x f(x_t, y_t) - \widetilde{J}_t$ ▷ finite-difference version of (8)

6: **(Upper-level step)** $x_{t+1} \leftarrow P_\phi^\gamma(x_t, w_t)$ by (11)

7: **(Operator construction)** $\widehat{\mathcal{H}}_t := \widehat{\mathcal{H}}_R(x_t, y_t; \delta_\varepsilon) \in \mathrm{End}(T_{y_t}\mathcal{M}_y)$, self-adjoint, by (a) the symmetrized basis sweep of (10) ($2d_y$ probes of $\mathcal{G}_y g$) or (b) direct evaluation

8: **(Spectral clip)** $h_t \leftarrow \mathcal{S}_{[\mu_c, L_g]}[\widehat{\mathcal{H}}_t][v_t] - \mathcal{G}_y f(x_t, y_t)$ with $\mathcal{S}_{[\mu_c, L_g]}$ from (5)

9: **(Projected auxiliary update and transport)** $v_{t+1} \leftarrow \Gamma_{y_t \to y_{t+1}} \Pi_{r_v}^{y_t}\big(P_t(v_t - \tau h_t)\big) \in T_{y_{t+1}}\mathcal{M}_y$

P_t: orthogonal projector onto $\mathrm{range}(\mathcal{S}_{[\mu_c, L_g]}[\widehat{\mathcal{H}}_t])$ (from the eigendecomposition in line 8); $\Pi_{r_v}^y$: projection onto $\{u \in T_y\mathcal{M}_y : \|u\|_y \le r_v\}$

10: **end for**

11: **return** Choose uniformly at random from $\{x_t\}_{t=1}^T$

Operator-level Hessian approximation. The clip $\mathcal{S}_{[\mu_c, L_g]}$ acts via the eigendecomposition of its argument, so we need the Hessian as a self-adjoint operator. We denote this by $\widehat{\mathcal{H}}_R(x, y; \delta_\varepsilon)$, with two admissible realizations: (a) **Basis-sweep finite-difference:** for an orthonormal basis $\{e_i\}$ of $T_y\mathcal{M}_y$, set $\widehat{\mathcal{H}}_R^\circ e_i := \widetilde{\mathcal{H}}_R(x, y, e_i, \delta_\varepsilon)$, extend linearly ($2d_y$ probes of $\mathcal{G}_y g$ per outer iteration), and symmetrize, $\widehat{\mathcal{H}}_R := \frac{1}{2}\big(\widehat{\mathcal{H}}_R^\circ + (\widehat{\mathcal{H}}_R^\circ)^*\big)$ with the adjoint taken in $\langle\cdot,\cdot\rangle_y$. Columnwise finite differences are not self-adjoint in general, whereas (5) requires a self-adjoint argument; since $\mathrm{Hess}_y\, g$ is self-adjoint, symmetrization does not increase the Frobenius error; (b) **Direct evaluation:** when $\mathrm{Hess}_y\, g(x, y)$ is available in closed form (as in our meta-learning experiment), set $\widehat{\mathcal{H}}_R = \mathrm{Hess}_y\, g(x, y)$. The analysis only uses the error $\rho_F := \|\widehat{\mathcal{H}}_R - \mathrm{Hess}_y\, g\|_F$, with $\rho_F \le L_{gyy}\sqrt{d_y}\,\delta_\varepsilon$ for (a) and $\rho_F = 0$ for (b).

For $x \in \mathcal{M}_x$, $w \in T_x\mathcal{M}_x$, and $\gamma > 0$, define

$$P_\phi^\gamma(x, w) \;:=\; \mathrm{Exp}_x\Big(\operatorname*{arg\,min}_{u \in \mathrm{Exp}_x^{-1}(\mathcal{U}_x)} \big\{ \langle w, u\rangle_x + \tfrac{1}{2\gamma}\|u\|_x^2 + \phi(\mathrm{Exp}_x(u)) \big\} \Big). \tag{11}$$

Remark 12 (Scope of "Hessian/Jacobian-free")**.** *Following [54], the descriptor refers to two design choices: (i) the algorithm never solves a linear system in* $\mathrm{Hess}_y\, g$*, and (ii) it does not differentiate through the inner loop. The spectral clip step requires* $\mathrm{Hess}_y\, g$ *as an operator, realized either via* $2d_y$ *first-order probes of* $\mathcal{G}_y g$ *or via a closed-form Hessian when one is available; both routes use only first-order oracle calls plus an* $O(d_y^3)$ *eigendecomposition. A discussion of finite-difference robustness is in Remark 26.*

Intuition of Algorithm 1. A single-loop method (one y-update per x-update) with three coupled updates and a per-iteration operator construction. The lower-level step moves y_t along $-\mathcal{G}_y g(x_t, \cdot)$ as an inexact approximation to $y^\star(x_t)$. The auxiliary v_t tracks the regularized inverse-Hessian-vector term $v^\star(x_t, y_t)$ from (6) without solving a linear system; the spectral clip in line 8 makes the v-subproblem μ_c-strongly convex on the range of $\mathcal{S}_{[\mu_c, L_g]}[\widehat{\mathcal{H}}_t]$, and the projector in line 9 keeps the iterate on that range, so the step is well posed even when $\mathrm{Hess}_y\, g$ is singular or indefinite. The upper-level step updates x_t using the approximate hypergradient assembled from v_t and the finite-difference cross-derivative surrogate. All three updates are evaluated at the current point (x_t, y_t, v_t), which is the order analyzed in Appendix G; the new auxiliary vector is then transported to $T_{y_{t+1}}\mathcal{M}_y$. Whereas [43] proposes Hessian-inversion strategies, our method sidesteps the linear-system solve entirely and replaces the Hessian inverse with a kernel-preserving spectral-clip pseudoinverse, well-defined regardless of the spectrum of $\mathrm{Hess}_y\, g$.

3.1 Theoretical guarantees

Throughout, w_t is the approximate hypergradient of line 5 of Algorithm 1, and $\mathcal{G}(x_t, w_t, \gamma) := -\gamma^{-1} \operatorname{Exp}_{x_t}^{-1}(x_{t+1})$ is the Riemannian gradient mapping of the proximal step (11). We write $G(x) := g(x, y^\star(x))$ and $\Psi_1 := \Phi(x_1) + g(x_1, y_1) - G(x_1) + \|v_1 - v^\star(x_1, y_1)\|_{y_1}^2$ for the initial value of the Lyapunov function.

Theorem 13 (Informal version; full statement in Theorem 35)**.** *Assume the sequence $\{x_t, y_t, v_t\}_{t=1}^T$ is generated by Algorithm 1 under Assumptions 2–8, 1, and 27 and hypotheses (P1)–(P3) of Lemma 28. Choose $\tau \le 1/(6L_g)$ and γ, λ as in* (46)*, whose drift-dependent caps are $\gamma \le 3\mu_c\tau/(400\bar{L}^2)$ and $\lambda \le \min\{1/(2\zeta L_g),\, 3\mu_c\tau/(200\bar{L}^2)\}$ with $\bar{L}$ the drift constant* (31)*. Then*

$$\frac{1}{T}\sum_{t=1}^{T} \|\mathcal{G}(x_t, w_t, \gamma)\|_{x_t}^2 \le \frac{8(\Psi_1 - \Phi_\star)}{\gamma T} + O(\delta_\varepsilon^2) = O\!\left(\frac{1}{T}\right)$$

for $\delta_\varepsilon = O(T^{-1/2})$ as in Theorem 35. Hence Algorithm 1 attains an ε-stationary point in the sense of the Riemannian gradient mapping in $O(\varepsilon^{-1})$ iterations. If $\phi \equiv 0$, the proximal step reduces to $x_{t+1} = \operatorname{Exp}_{x_t}(-\gamma w_t)$ and $\mathcal{G}(x_t, w_t, \gamma) = w_t$, so the bound becomes an $O(\varepsilon^{-1})$ complexity bound for the smooth Riemannian bilevel problem.

Corollary 14 (True hypergradient)**.** *Under the conditions of Theorem 13 with $\phi \equiv 0$, $\frac{1}{T}\sum_{t=1}^{T} \|\mathcal{G}F(x_t)\|_{x_t}^2 = O(1/T)$ (proof in Appendix G.3).*

Below, $\bar{c} \ge 1$, $c_R, c_T \ge 0$ are the retraction and vector-transport constants of Assumption 22 (Appendix B), L_F is the Lipschitz constant of $\mathcal{G}F$ (Lemma 28(iv)), and $M_F := C_{fx} + C_{gxy}C_{fy}/\mu \ge \sup_x \|\mathcal{G}F(x)\|_x$, where $C_{fx} \ge \sup \|\mathcal{G}_x f\|$ is an additional bound required only here.

Corollary 15 (Extension to general retractions; full proof in Appendix G.4)**.** *Let $\phi \equiv 0$ and $\|\mathcal{G}_x f\| \le C_{fx}$ on $\mathcal{U}_x \times \mathcal{U}_y$. Replace, in Algorithm 1, the exponential maps of the lower-level and upper-level updates by a retraction* Retr*, and the auxiliary-vector transport by an* isometric *vector transport $\overline{\Gamma}$; the finite-difference probes* (9)–(10) *keep the exponential map and parallel transport (or the closed-form Hessian of route (b)). Define $\mathcal{G}_{\mathrm{Retr}}(x_t, w_t, \gamma) := -\gamma^{-1} \operatorname{Retr}_{x_t}^{-1}(x_{t+1}) = w_t$. Under Assumptions 2–8, 1, 27, and 22 and (P1)–(P3) of Lemma 28, with stepsizes chosen as in Theorem 35 but with L_F replaced by $\widetilde{L}_F := \bar{c}\, L_F + 2c_R\, M_F$, $\bar{L}^2$ replaced by $2\bar{c}\,\widetilde{\bar{L}}^2$ with $\widetilde{\bar{L}}^2 := \bar{L}^2 + c_T^2 C_{fy}^2/(4\mu_c^2)$, the two γ-caps of Lemma 32 divided by $\bar{c}$, and $\lambda \le 1/(2\bar{c}\,\zeta L_g + 4c_R L_g D_y)$, $\frac{1}{T}\sum_{t=1}^{T} \|\mathcal{G}_{\mathrm{Retr}}(x_t, w_t, \gamma)\|_{x_t}^2 = O(1/T)$, and the same $O(\varepsilon^{-1})$ iteration complexity follows. Setting $\bar{c} = 1$ and $c_R = c_T = 0$ recovers Theorem 13 with $\phi \equiv 0$ (up to a factor 2 in the drift caps). The proof is given in Appendix G.4.*

Complexity comparison. Table 2 (Appendix C) compares Algorithm 1 with the closest prior results. RHGD [43] assumes geodesic strong convexity, with oracle complexities polynomial in $\kappa_\ell = L/\mu$ (e.g. $G_g = \widetilde{O}(\kappa_\ell^5 \zeta \varepsilon^{-1})$ for RHGD-NS). Algorithm 1 matches the $O(\varepsilon^{-1})$ iteration complexity of the Euclidean method of [54], with constants scaled by ζ and C_0 through $\bar{L}$, under a Riemannian PL lower level. Each iteration costs $5 + 2d_y$ first-order gradient evaluations (basis sweep) or five plus one closed-form Hessian, plus an $O(d_y^3)$ eigendecomposition, so the first-order oracle complexity is $O(d_y \varepsilon^{-1})$ or $O(\varepsilon^{-1})$; the contribution is a geometry/generalization result rather than a complexity improvement over [54].

Relation to [54] and proof techniques. Relative to [54], the extension from an isolated nondegenerate minimizer to a Morse–Bott minimizer manifold changes the operator (the kernel-preserving clip), the target (Proposition 9 with †), and adds the compatibility condition (Assumption 5); the clipped tracking system is, to our knowledge, new even at $\zeta = 1$, $C_0 = 0$ (Remark 25). The geometric lift adds (a) the trigonometric distance bound (2), with the conservative inner stepsize $\lambda \le 1/(2\zeta L_g)$; (b) a conservative transport-distortion term $O(C_0)$ in the auxiliary-vector drift; and (c) transported Lipschitz conditions. Relative to [43], we avoid differentiating through the exponential map and inverting the Riemannian Hessian; the κ_ℓ^5 dependence of RHGD-NS is replaced by $1/\mu_c$ factors inside $\bar{L}$ and a factor $\mu_c\tau = O(\mu_c/L_g)$ in the stepsize caps (Appendix G).

3.2 Stochastic extension

We now consider the stochastic finite-sum setting $f(x,y) = \frac{1}{n}\sum_{i=1}^{n} f_i(x,y)$, $g(x,y) = \frac{1}{m}\sum_{j=1}^{m} g_j(x,y)$, with each f_i, g_j on $\mathcal{M}_x \times \mathcal{M}_y$. The stochastic counterpart (Algorithm 2, Appendix H) replaces the full gradients in the lower-level step, upper-level surrogate, and auxiliary update by mini-batch estimators on $\mathcal{B}_f \subseteq [n]$, $\mathcal{B}_g \subseteq [m]$: for a mini-batch $\mathcal{B}$, $\mathcal{G}_x f_{\mathcal{B}} := |\mathcal{B}|^{-1}\sum_{i\in\mathcal{B}} \mathcal{G}_x f_i$, and analogously for $\mathcal{G}_y f_{\mathcal{B}}$, $\mathcal{G}_x g_{\mathcal{B}}$, $\mathcal{G}_y g_{\mathcal{B}}$. We write $\widehat{w}_t$ for the resulting stochastic counterpart of w_t and $\mathcal{G}_\gamma(x_t, \widehat{w}_t) := \mathcal{G}(x_t, \widehat{w}_t, \gamma)$ for its gradient mapping; the variance constants $\sigma_{fx}, \sigma_{fy}, \sigma_{gx}, \sigma_{gy}, \sigma_J, \sigma_H$ are those of Assumption 16.

Coupled finite-difference sampling. A point specific to the Hessian/Jacobian-free setting – and absent from stochastic Riemannian methods that access second-order oracles directly [43] – is the variance of the central finite-difference estimators in lines 4–7 of Algorithm 2. Independent mini-batches at the $\pm\delta_\varepsilon$ perturbations give variance $\Theta(\sigma^2/(\delta_\varepsilon^2|\mathcal{B}|))$, breaking the schedule $\delta_\varepsilon = \Theta(\varepsilon^{1/2})$. We therefore use the standard common-random-numbers device: the same batch at both perturbations, so each component-wise difference is $O(\delta_\varepsilon)$, the δ_ε factors cancel, and the variance is $O(\sigma_J^2/|\mathcal{B}_t^J|)$ with σ_J^2 depending only on (L_g, r_v). The formal statement is deferred to Appendix H (Assumption 37, Lemma 38).

Assumption 16 (Stochastic manifold oracles)**.** *Under the stochastic setting, Assumptions 2–8 hold for each f_i, g_j with the same constants. The mini-batch gradients $\mathcal{G}_x f_{\mathcal{B}}(x,y), \mathcal{G}_y f_{\mathcal{B}}(x,y), \mathcal{G}_x g_{\mathcal{B}}(x,y), \mathcal{G}_y g_{\mathcal{B}}(x,y)$ are unbiased, with variance uniformly bounded by $\frac{\sigma_{fx}^2}{|\mathcal{B}|}, \frac{\sigma_{fy}^2}{|\mathcal{B}|}, \frac{\sigma_{gx}^2}{|\mathcal{B}|}, \frac{\sigma_{gy}^2}{|\mathcal{B}|}$ for some constants $\sigma_{fx}, \sigma_{fy}, \sigma_{gx}, \sigma_{gy} > 0$. The coupled finite-difference estimators $\widehat{J}_t, \widehat{H}_t$ in Algorithm 2 have variance bounded by $\sigma_J^2/|\mathcal{B}_t^J|$ and, per basis direction i, $\sigma_H^2/|\mathcal{B}_t^{H,i}|$ (hence $\sigma_H^2 d_y/B_H$ for the full operator in Frobenius norm), with σ_J^2, σ_H^2 depending only on (L_g, r_v) (full statement in Appendix H, Assumption 37).*

Theorem 17 (Stochastic convergence; constants traced in Appendix H.1)**.** *Suppose Assumptions 2–8, 1, 27, and 16 and (P1)–(P3) of Lemma 28 hold, and run Algorithm 2 with stepsizes γ, λ, τ from* (46)*, except that the caps $\mu_c\tau/(60C_{gxy}^2)$ and $\mu^2\lambda/(30A)$ are tightened to $\mu_c\tau/(72C_{gxy}^2)$ and $\mu^2\lambda/(36A)$ (Appendix H.1, Step 4), and a common mini-batch size B. Set $\Delta_1 := \Psi_1 - \Phi_\star$. Then there exist constants $C_1, C_2, C_3 > 0$, depending only on $(\mu, \mu_c, \rho, d_y, L_f, L_g, L_{gxy}, L_{gyy}, C_{fy}, C_{gxy}, r_v, \zeta, C_0)$ and the variance constants, such that*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\mathcal{G}_\gamma(x_t, \widehat{w}_t)\|_{x_t}^2 \;\le\; \frac{C_1\,\Delta_1}{T} \;+\; C_2\,\delta_\varepsilon^2 \;+\; \frac{C_3}{B}. \tag{12}$$

When $\phi \equiv 0$, $\mathcal{G}_\gamma(x_t, \widehat{w}_t) = \widehat{w}_t$, and (12) *bounds the averaged squared norm of the stochastic Riemannian hypergradient estimator itself.*

Corollary 18 (Complexity consequence)**.** *Under the assumptions of Theorem 17, choose $\delta_\varepsilon = \Theta(\varepsilon^{1/2}), B = \Theta(\varepsilon^{-1}), T = \Theta(\varepsilon^{-1})$. Let $\widetilde{t}$ be uniform on $\{1, \dots, T\}$ independent of the algorithm, and $\widetilde{x} := x_{\widetilde{t}}$. Then $\mathbb{E}\big\|\mathcal{G}_\gamma(\widetilde{x}, \widehat{w}_{\widetilde{t}})\big\|_{\widetilde{x}}^2 \le \varepsilon$. Hence SR-HJFBiO attains ε-stationarity in $O(\varepsilon^{-1})$ iterations and $O(\varepsilon^{-2})$ stochastic first-order oracle calls.*

The closest stochastic Riemannian result [43] assumes geodesic strong convexity and achieves the same $O(\varepsilon^{-2})$ order while calling separate stochastic Hessian/Jacobian oracles directly; the closest Euclidean baseline stocBiO [58] also requires strong convexity and obtains $\widetilde{O}(\kappa^{O(1)}\varepsilon^{-2})$. Algorithm 2 matches the standard $O(\varepsilon^{-2})$ stochastic order while extending the analysis to Riemannian PL and using only first-order $\mathcal{G}g$ evaluations.

4 Experiments

We compare R-HJFBiO with the four hypergradient strategies of [43] – direct Hessian inversion (HINV), conjugate gradient (CG), truncated Neumann series (NS), and automatic differentiation through the inner loop (AD) – on Stiefel, Grassmann, and Poincaré-ball geometries. The settings are: (i) bilevel ridge regression with deliberately rank-deficient lower-level Hessians (BCI IV-2a EEG and a synthetic Poincaré-ball generator); (ii) a UCI Superconductivity scaling experiment at

$d = 70$ that checks the same diagnostic picture in higher dimensions; and (iii) a 5-way 5-shot meta-learning task on MiniImageNet with a manifold-constrained convolutional backbone. Settings (i) and (ii) are in Appendix D; the main text focuses on (iii), which also sweeps the lower-level regularization.

Implemented variant and scope. "R-HJFBiO" in figures and tables denotes an exact-solve variant: line 9 is replaced by its fixed point $v = \mathcal{S}_{[\mu_c, L_g]}(\widehat{\mathcal{H}}_t)^\dagger \nabla_y f$, computed by spectral inversion. The runs also take $T_{\text{in}} > 1$ inner steps, clip the upper-level gradient (regression), and redraw tasks with a reset head (meta-learning). The theorems do not cover these changes (Appendix F): the experiments evaluate the clipped operator, not the tracking recursion. Two meta-learning cells are stress tests outside the theory: at $\lambda = 10^{-6}$, $\mu \le \lambda < \mu_c = 10^{-3}$, so Proposition 11 fails; at $\lambda = 0$, cross-entropy on separable support sets has no finite minimizer.

Task and backbone. 5-way 5-shot classification on MiniImageNet, using a seeded random 64:16:20 class split. Each task draws 5 classes with $N_s = 5$ support and $N_q = 15$ query images per class. The backbone is a 4-block CNN with 16 channels per layer producing a 400-dimensional feature embedding from 84×84 inputs; the first two convolutional kernels are constrained to Stiefel manifolds $W_1 \in \mathrm{St}(27, 16)$ and $W_2 \in \mathrm{St}(144, 16)$, with each column representing one orthonormalized output channel. The remaining two blocks are Euclidean. Given a backbone Θ and support features $F_s = \phi_\Theta(X_s)$, the task-specific linear head $w \in \mathbb{R}^{d\times K}$ minimizes $g(\Theta, w) = (N_s K)^{-1} \sum_{n=1}^{N_s K} \mathrm{CE}(F_s w, y_s) + (\lambda/2)\|w\|_F^2$, and the upper-level objective F is the cross-entropy of $w^\star(\Theta)$ on the query set. The full architecture, optimization details, and per-task closed-form Hessian – as well as the parallel runs that place W_1, W_2 on the Grassmann or Oblique manifolds (which reproduce every qualitative observation below) – are deferred to Appendix F.

Three regularization regimes. We sweep λ, the L2 weight on the inner classifier, through three values: $\lambda = 10^{-2}$ (*strongly convex*; the regime assumed in [43]), $\lambda = 10^{-6}$ (*tiny ridge*; nonsingular, but with $dK - N_s K(K-1) \ge 1900$ eigenvalues equal to λ), and $\lambda = 0$ (*singular Hessian*; rank of H at most $N_s K(K-1) = 100$ on a Hessian of dimension $dK = 2000$, so roughly 95% of the spectrum is exactly zero). All five methods use the same inner solver, outer-loop schedule, and divergence guard (any iterate with $\|\widehat{\mathcal{G}}F\|^2 > 10^8$ aborts cleanly so that subsequent iterates do not feed `Inf/NaN` downstream), so the baselines are the strategies of [43] inside an identical loop, which isolates the linear-system step. We report final-step metrics over a single meta-batch of 4 tasks (no repeated runs) and $T = 200$ outer iterations. Gradient norms use a *shared reference* $v^\star$ (Appendix D.3), evaluated at the inner-solver output (30 gradient steps from $w = 0$ on each task); it equals the hypergradient only at a lower-level minimizer under Assumption 5 and $\mu_c \le \mu$, so we report it as a clipped-surrogate diagnostic, not a stationarity certificate.

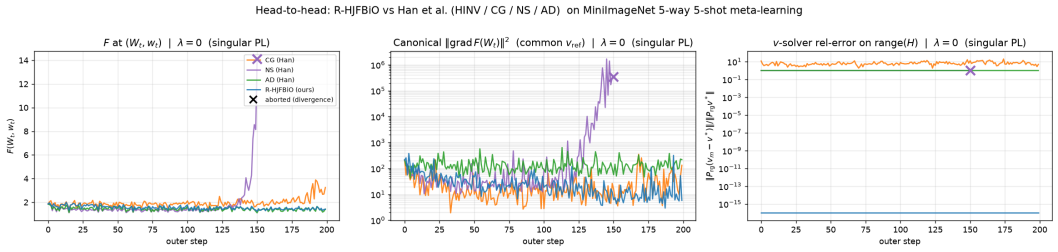


Figure 1: **MiniImageNet 5-way 5-shot meta-learning, Stiefel-constrained Conv-4 backbone, singular-Hessian regime ($\lambda = 0$).** Left: $F(W_t, w_t)$ on the meta-batch. Centre: shared-reference (clipped-surrogate) gradient norm $\|\mathcal{G}F(W_t)\|^2$; the reference $v^\star$ is built from the clipped pseudoinverse (7) (origin) and applied identically to every method's iterates (application). Right: v-solver relative error on the retained subspace $\mathcal{R}_{\mu_c}(H)$. Crosses ($\times$) mark runs aborted by the divergence guard ($\|\widehat{\mathcal{G}}F\|^2 > 10^8$); HINV diverges at step 0. R-HJFBiO (exact-solve variant) computes the reference vector itself, so its $\mathrm{err}_v = 0$ by construction. The strongly convex (λ=10^{-2}) and tiny-ridge (λ=10^{-6}) regimes are deferred to Appendix F, and show a similar pattern.

Findings. Three points stand out in Figure 1 and Table 4. *First*, a fixed tiny ridge does not stabilize direct inversion: HINV diverges at the first step for $\lambda \in \{10^{-6}, 0\}$, and in each of these regimes one of CG, NS is aborted mid-run and the other ends with a rising objective. *Second*, AD does not crash, but its shared-reference gradient norm sits at $\sim 10^2$, more than an order of magnitude above R-HJFBiO, and the Neumann proxy for its auxiliary vector (exact only for quadratic inner losses; Appendix D.3) stays above relative error 0.99 on $\mathcal{R}_{\mu_c}(H)$; its own-norm column hides this. *Third*, R-HJFBiO and AD are the only methods that complete all three regimes. R-HJFBiO's $\mathrm{err}_v = 0$ holds by construction, and AD has higher query accuracy in every regime (0.55–0.56 vs. 0.41–0.49): these runs probe the linear-system step, not meta-learning performance. Grassmann and Oblique backbones behave similarly (Appendix F).

The other six cells. The bilevel-regression setting (BCI IV-2a on Stiefel and Grassmann, plus the synthetic Poincaré-ball generator with $\zeta_x \leq 4.40$) and the UCI Superconductivity scaling at $d = 70$ on $\mathrm{St}(81, 70)$, $\mathrm{Gr}(81, 35)$, and B^8 appear in Appendix D. The baselines' errors on $\mathcal{R}_{\mu_c}(H)$ range from about 10^{-7} (HINV, UCI) to 10^{19} (CG, BCI Stiefel); HINV and CG stall or oscillate in most cells but are accurate on BCI Grassmann and Poincaré, NS and AD match R-HJFBiO on the objective and shared-reference norm on UCI, and no method separates in gradient norm on BCI Poincaré.

5 Conclusion

We studied bilevel optimization on Riemannian manifolds under a lower-level Polyak–Łojasiewicz condition. To address the failure of the classical inverse-Hessian hypergradient formula away from the lower-level minimizer, we introduced an intrinsic regularized tangent-space formulation based on kernel-preserving spectral clipping, and developed the Hessian/Jacobian-free method R-HJFBiO with an $O(1/T)$ rate, $O(\varepsilon^{-1})$ iteration complexity, and a stochastic extension matching the standard $O(\varepsilon^{-2})$ stochastic first-order complexity. Across nine head-to-head cells – bilevel regression on BCI IV-2a (Stiefel, Grassmann) and a Poincaré-ball generator, UCI Superconductivity scaling at $d = 70$, and a meta-learning stress test on MiniImageNet swept from strongly convex to a singular Hessian – an exact-solve variant of R-HJFBiO remains well defined where the inverse-Hessian baselines diverge or stall, while truncation-based baselines can match its objective and surrogate-gradient trajectories although their auxiliary vectors stay far from the clipped target.

Limitations. Our analysis uses the exponential map and parallel transport, while experiments use a QR-based retraction; Corollary 15 covers retractions only for $\phi \equiv 0$ with an isometric vector transport. The experimental variant and two meta-learning cells are outside the theory (Section 4). The analysis requires the iterates to stay in the bounded geodesically convex neighborhoods of Assumption 1, which is not automatic even on Hadamard manifolds, and the parametric and domain hypotheses (P1)–(P3) of Lemma 28; relaxing these is left to future work. The finite-difference probes (9)–(10) have a truncation/cancellation trade-off (Appendix F), which route (b) avoids. Finally, when $\dim \mathcal{M}^\star(x) > 0$ our guarantees require the compatibility condition (Assumption 5), and away from $\mathcal{M}^\star(x)$ they rely on the spectral margin of Assumption 27, which we verify empirically rather than prove; likewise, the clip floor μ_c is set heuristically rather than estimated from the PL constant μ, so the requirement $\mu_c \leq \mu$ is not certified in the experiments, and a principled data-driven choice of μ_c is left to future work.

Acknowledgments and Disclosure of Funding

The authors received no third-party funding or support related to this work during the 36 months preceding submission. Furthermore, the authors declare no competing interests.

References

[1] H. Abbaszadehpeivasti, E. de Klerk, and M. Zamani. Conditions for linear convergence of the gradient method for non-convex optimization. *Optimization Letters*, 17(5):1105–1125, 2023.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

[3] F. Alimisis, A. Orvieto, G. Bécigneul, and A. Lucchi. A continuous-time perspective for modeling acceleration in riemannian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1297–1307. PMLR, 2020.

[4] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM journal on matrix analysis and applications*, 29(1):328–347, 2007.

[5] R. Bhatia. Positive definite matrices. In *Positive Definite Matrices*. Princeton university press, 2009.

[6] S. Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

[7] H. Bonnel, L. Todjihoundé, and C. Udrişte. Semivectorial bilevel optimization on riemannian manifolds. *Journal of Optimization Theory and Applications*, 167(2):464–486, 2015.

[8] N. Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

[9] N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019.

[10] J. Bracken and J. T. McGill. Mathematical programs with optimization problems in the constraints. *Operations research*, 21(1):37–44, 1973.

[11] Y. Cai, M. I. Jordan, T. Lin, A. Oikonomou, and E.-V. Vlatakis-Gkaragkounis. Curvature-independent last-iterate convergence for games on riemannian manifolds. *arXiv preprint arXiv:2306.16617*, 2023.

[12] J. Cao, R. Jiang, N. Abolfazli, E. Yazdandoost Hamedani, and A. Mokhtari. Projection-free methods for stochastic simple bilevel optimization with convex lower-level problem. *Advances in Neural Information Processing Systems*, 36:6105–6131, 2023.

[13] L. Chen, J. Xu, and J. Zhang. Bilevel optimization without lower-level strong convexity from the hyper-objective perspective. 2023.

[14] L. Chen, J. Xu, and J. 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.

[15] T. Chen, Y. Sun, Q. Xiao, and W. Yin. A single-timescale method for stochastic bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2466–2488. PMLR, 2022.

[16] T. Chen, Y. Sun, and W. Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems*, 34:25294–25307, 2021.

[17] Z. Chen, B. Kailkhura, and Y. Zhou. An accelerated proximal algorithm for regularized nonconvex and nonsmooth bi-level optimization. *Machine Learning*, 112(5):1433–1463, 2023.

[18] A. Cherian and S. Sra. Riemannian dictionary learning and sparse coding for positive definite matrices. *IEEE transactions on neural networks and learning systems*, 28(12):2859–2871, 2016.

[19] B. Colson, P. Marcotte, and G. Savard. An overview of bilevel optimization. *Annals of operations research*, 153(1):235–256, 2007.

[20] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

[21] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[22] M. Dagréou, P. Ablin, S. Vaiter, and T. Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. *Advances in Neural Information Processing Systems*, 35:26698–26710, 2022.

[23] J. Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pages 318–326. PMLR, 2012.

[24] A. Douik and B. Hassibi. Manifold optimization over the set of doubly stochastic matrices: A second-order geometry. *IEEE Transactions on Signal Processing*, 67(22):5761–5774, 2019.

[25] S. Dutta, X. Cheng, and S. Sra. Riemannian bilevel optimization. *arXiv preprint arXiv:2405.15816*, 2024.

[26] O. P. Ferreira, M. S. Louzeiro, and L. Prudente. Gradient method for optimization on riemannian manifolds with lower bounded curvature. *SIAM Journal on Optimization*, 29(4):2517–2541, 2019.

[27] M. Feurer and F. Hutter. Hyperparameter optimization. In *Automated machine learning: Methods, systems, challenges*, pages 3–33. Springer, 2019.

[28] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[29] R. Flamary, A. Rakotomamonjy, and G. Gasso. Learning constrained task similarities in graphregularized multi-task learning. *Regularization, Optimization, Kernels, and Support Vector Machines*, 103:1, 2014.

[30] L. Franceschi, M. Donini, P. Frasconi, and M. Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International conference on machine learning*, pages 1165–1173. PMLR, 2017.

[31] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International conference on machine learning*, pages 1568–1577. PMLR, 2018.

[32] O. Ganea, G. Bécigneul, and T. Hofmann. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018.

[33] S. Ghadimi and M. Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

[34] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2066–2073. IEEE, 2012.

[35] S. Gould, B. Fernando, A. Cherian, P. Anderson, R. S. Cruz, and E. Guo. On differentiating parameterized argmin and argmax problems with application to bi-level optimization. *arXiv preprint arXiv:1607.05447*, 2016.

[36] R. Grazzi, L. Franceschi, M. Pontil, and S. Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pages 3748–3758. PMLR, 2020.

[37] K. Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, Nov. 2018.

[38] K. Hamidieh. Superconductivty Data. UCI Machine Learning Repository, 2018. DOI: https://doi.org/10.24432/C53P47.

[39] A. Han, B. Mishra, P. Jawanpuria, and J. Gao. Learning with symmetric positive definite matrices via generalized bures-wasserstein geometry. In *International Conference on Geometric Science of Information*, pages 405–415. Springer, 2023.

[40] A. Han, B. Mishra, P. Jawanpuria, and J. Gao. Nonconvex-nonconcave min-max optimization on riemannian manifolds. *Transactions on Machine Learning Research*, 2023.

[41] A. Han, B. Mishra, P. Jawanpuria, P. Kumar, and J. Gao. Riemannian hamiltonian methods for min-max optimization on manifolds. *SIAM Journal on Optimization*, 33(3):1797–1827, 2023.

[42] A. Han, B. Mishra, P. K. Jawanpuria, and J. Gao. On riemannian optimization over positive definite matrices with the bures-wasserstein geometry. *Advances in Neural Information Processing Systems*, 34:8940–8953, 2021.

[43] A. Han, B. Mishra, P. K. Jawanpuria, and A. Takeda. A framework for bilevel optimization on riemannian manifolds. *Advances in Neural Information Processing Systems*, 37:103829–103872, 2024.

[44] P. Hansen, B. Jaumard, and G. Savard. New branch-and-bound rules for linear bilevel programming. *SIAM Journal on scientific and Statistical Computing*, 13(5):1194–1217, 1992.

[45] M. Harandi, M. Salzmann, and R. Hartley. Dimensionality reduction on spd manifolds: The emergence of geometry-aware methods. *IEEE transactions on pattern analysis and machine intelligence*, 40(1):48–62, 2017.

[46] R. Hataya and M. Yamada. Nyström method for accurate and scalable implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, pages 4643–4654. PMLR, 2023.

[47] M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic, 2022.

[48] M. Hong, H.-T. Wai, Z. Wang, and Z. 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.

[49] I. Horev, F. Yger, and M. Sugiyama. Geometry-aware principal component analysis for symmetric positive definite matrices. In *Asian Conference on Machine Learning*, pages 1–16. PMLR, 2016.

[50] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

[51] Z. Hu, G. Wang, X. Wang, A. Wibisono, J. D. Abernethy, and M. Tao. Extragradient type methods for riemannian variational inequality problems. In *International Conference on Artificial Intelligence and Statistics*, pages 2080–2088. PMLR, 2024.

[52] F. Huang. Adaptive mirror descent bilevel optimization. *arXiv preprint arXiv:2311.04520*, 2023.

[53] F. Huang. On momentum-based gradient methods for bilevel optimization with nonconvex lower-level. *arXiv preprint arXiv:2303.03944*, 2023.

[54] F. Huang. Optimal hessian/jacobian-free nonconvex-pl bilevel optimization. *arXiv preprint arXiv:2407.17823*, 2024.

[55] F. Huang and S. Gao. Gradient descent ascent for minimax problems on riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8466–8476, 2023.

[56] F. Huang, J. Li, S. Gao, and H. Huang. Enhanced bilevel optimization via bregman distance. *Advances in Neural Information Processing Systems*, 35:28928–28939, 2022.

[57] Z. Huang and L. Van Gool. A riemannian network for spd matrix learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[58] K. Ji, J. Yang, and Y. Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pages 4882–4892. PMLR, 2021.

[59] M. Jordan, T. Lin, and E.-V. Vlatakis-Gkaragkounis. First-order algorithms for min-max optimization in geodesic metric spaces. *Advances in Neural Information Processing Systems*, 35:6557–6574, 2022.

[60] H. Kasai, H. Sato, and B. Mishra. Riemannian stochastic recursive gradient algorithm. In *International conference on machine learning*, pages 2516–2524. PMLR, 2018.

[61] P. Khanduri, S. Zeng, M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in neural information processing systems*, 34:30271–30283, 2021.

[62] M. Kochurov, R. Karimov, and S. Kozlukov. Geoopt: Riemannian optimization in pytorch. *arXiv preprint arXiv:2005.02819*, 2020.

[63] V. Konda and J. Tsitsiklis. Actor-critic algorithms. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

[64] G. Kunapuli, K. P. Bennett, J. Hu, and J.-S. Pang. Classification model selection via bilevel programming. *Optimization Methods & Software*, 23(4):475–489, 2008.

[65] J. Kwon, D. Kwon, S. Wright, and R. D. Nowak. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pages 18083–18113. PMLR, 2023.

[66] J. Li, L. Fuxin, and S. Todorovic. Efficient riemannian optimization on the stiefel manifold via the cayley transform. *arXiv preprint arXiv:2002.01113*, 2020.

[67] J. Li, B. Gu, and H. Huang. Improved bilevel model: Fast and optimal algorithm with theoretical guarantee. *arXiv preprint arXiv:2009.00690*, 2020.

[68] J. Li, B. Gu, and H. Huang. A fully single loop algorithm for bilevel optimization without hessian inverse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7426–7434, 2022.

[69] J. Li and S. Ma. Riemannian bilevel optimization. *Journal of Machine Learning Research*, 26(18):1–44, 2025.

[70] J. Liao and Z. Wan. Inexact restoration methods for semivectorial bilevel programming problem on riemannian manifolds. *Axioms*, 11(12):696, 2022.

[71] J. Liao and Z. Wan. On the karush-kuhn-tucker reformulation of the bilevel optimization problems on riemannian manifolds. *Filomat*, 36(11):3609–3624, 2022.

[72] Z. Lin. Riemannian geometry of symmetric positive definite matrices via cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1353–1370, 2019.

[73] B. Liu, M. Ye, S. Wright, P. Stone, and Q. Liu. Bome! bilevel optimization made easy: A simple first-order approach. *Advances in neural information processing systems*, 35:17248–17262, 2022.

[74] H. Liu, K. Simonyan, and Y. Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

[75] R. Liu, Y. Liu, W. Yao, S. Zeng, and J. Zhang. Averaged method of multipliers for bi-level optimization without lower-level strong convexity. In *International Conference on Machine Learning*, pages 21839–21866. PMLR, 2023.

[76] R. Liu, P. Mu, X. Yuan, S. Zeng, and J. Zhang. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *International conference on machine learning*, pages 6305–6315. PMLR, 2020.

[77] J. Lorraine, P. Vicol, and D. Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International conference on artificial intelligence and statistics*, pages 1540–1552. PMLR, 2020.

[78] S. Lu. Slm: A smoothed first-order lagrangian method for structured constrained nonconvex optimization. *Advances in Neural Information Processing Systems*, 36:80414–80454, 2023.

[79] Z. Lu and S. Mei. First-order penalty methods for bilevel optimization. *SIAM Journal on Optimization*, 34(2):1937–1969, 2024.

[80] D. Maclaurin, D. Duvenaud, and R. Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pages 2113–2122. PMLR, 2015.

[81] D. Martínez-Rubio, C. Roux, C. Criscitiello, and S. Pokutta. Accelerated methods for riemannian min-max optimization ensuring bounded geometric penalties. *arXiv preprint arXiv:2305.16186*, 2023.

[82] B. Mishra, N. T. V. Satyadev, H. Kasai, and P. Jawanpuria. Manifold optimization for non-linear optimal transport problems, 2021.

[83] G. M. Moore. *Bilevel programming algorithms for machine learning model selection*. Rensselaer Polytechnic Institute, 2010.

[84] M. Nickel and D. Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

[85] T. Okuno, A. Takeda, A. Kawana, and M. Watanabe. On lp-hyperparameter learning via bilevel nonsmooth optimization. *Journal of Machine Learning Research*, 22(245):1–47, 2021.

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

[87] G. Peyré and M. Cuturi. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019.

[88] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*, 2017.

[89] Q. Rebjock and N. Boumal. Fast convergence to non-isolated minima: four equivalent conditions for c 2 functions: Q. rebjock, n. boumal. *Mathematical Programming*, 213(1):151–199, 2025.

[90] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.

[91] A. Shaban, C.-A. Cheng, N. Hatch, and B. Boots. Truncated back-propagation for bilevel optimization, 2019.

[92] H. Shen and T. Chen. On penalty-based bilevel gradient descent method. In *International conference on machine learning*, pages 30992–31015. PMLR, 2023.

[93] C. Shi, J. Lu, and G. Zhang. An extended kuhn–tucker approach for linear bilevel programming. *Applied Mathematics and Computation*, 162(1):51–63, 2005.

[94] D. Sow, K. Ji, Z. Guan, and Y. Liang. A primal-dual approach to bilevel optimization with multiple inner minima. *arXiv preprint arXiv:2203.01123*, 2022.

[95] D. Sow, K. Ji, and Y. Liang. On the convergence theory for hessian-free bilevel algorithms. *Advances in Neural Information Processing Systems*, 35:4136–4149, 2022.

[96] S. Sra and R. Hosseini. Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization*, 25(1):713–739, 2015.

[97] R. S. Sukthanker, Z. Huang, S. Kumar, E. G. Endsjo, Y. Wu, and L. Van Gool. Neural architecture search of spd manifold networks. *arXiv preprint arXiv:2010.14535*, 2020.

[98] H. Tabealhojeh, P. Adibi, H. Karshenas, S. K. Roy, and M. Harandi. Rmaml: Riemannian meta-learning with orthogonality constraints. *Pattern Recognition*, 140:109563, 2023.

[99] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. R. Müller-Putz, et al. Review of the bci competition iv. *Frontiers in neuroscience*, 6:55, 2012.

[100] I. Tsaknakis, P. Khanduri, and M. Hong. An implicit gradient-type method for linearly constrained bilevel problems. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5438–5442. IEEE, 2022.

[101] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[102] R. Wang, K. Dvijotham, and I. R. Manchester. Monotone, bi-lipschitz, and polyak-lojasiewicz networks. *arXiv preprint arXiv:2402.01344*, 2024.

[103] X. Wang, D. Yuan, Y. Hong, Z. Hu, L. Wang, and G. Shi. Riemannian optimistic algorithms. *arXiv preprint arXiv:2308.16004*, 2023.

[104] Q. Xiao, H. Shen, W. Yin, and T. Chen. Alternating projected sgd for equality-constrained bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 987–1023. PMLR, 2023.

[105] H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In *Conference on learning theory*, pages 1617–1638. PMLR, 2016.

A Related Work

Classical and Euclidean bilevel optimization. Classical bilevel methods often relied on reformulations based on optimality conditions or constrained single-level surrogates [10, 44, 93, 83, 19]. In machine learning, the dominant modern line instead uses gradient-based bilevel methods. For nonconvex upper-level problems with strongly convex lower-level objectives, influential algorithmic templates include approximate implicit differentiation (AID), iterative differentiation (ITD), reversible or forward hypergradient propagation, and automatic-differentiation-based approaches [23, 86, 35, 80, 30, 31, 77, 36, 91, 33, 58]. Deterministic and stochastic improvements have been developed through two-timescale updates, alternating stochastic gradients, single-timescale methods, variance reduction, double-momentum ideas, and fully first-order schemes [47, 48, 16, 15, 22, 61, 73, 65]. Additional extensions address linearly constrained, equality-constrained, and beyond-singleton lower-level models [100, 104, 76, 85, 67].

A central issue in this Euclidean literature is hypergradient computation. Depending on the problem structure, the inverse-Hessian term can be approximated using conjugate-gradient solvers, Neumann-series expansions, truncated backpropagation, automatic differentiation, or low-rank approximations such as Nyström methods [33, 58, 36, 91, 46, 77]. These ideas have also played an important role in meta-learning and hyperparameter optimization, including MAML-type models, implicit-gradient meta-learning, differentiable closed-form solvers, and partial-parameter adaptation analyses [28, 31].

Beyond lower-level strong convexity. The strong convexity assumption is mathematically convenient because it guarantees uniqueness of the lower-level solution and invertibility of the lower-level Hessian at the solution, but it is restrictive in many modern models. This has led to a growing literature that weakens strong convexity to convexity, error-bound conditions, proximal error bounds, PL-type conditions, or broader non-singleton lower-level regimes [76, 13, 94, 79, 12]. For nonconvex–PL bilevel optimization in particular, recent work has established first-order methods, penalty methods, single-loop or momentum schemes, and Hessian/Jacobian-free algorithms [73, 92, 52, 53, 68, 95, 78, 17, 56, 54]. Hardness results and refined upper bounds in the absence of lower-level strong convexity were studied in [14], while related convex-lower-level or multiplier-based approaches were explored in [75]. Among the recent PL-based works, [54] is particularly close in spirit to ours: it develops an optimal Euclidean Hessian/Jacobian-free method for nonconvex–PL bilevel optimization with $O(1/T)$ convergence and $O(\varepsilon^{-1})$ iteration complexity.

Riemannian optimization and manifold-constrained learning. Our setting also relates to the broader literature on Riemannian optimization and manifold-constrained learning. Foundational algorithms and geometric tools appear in [2, 8, 105, 6, 9, 60, 26, 3], while manifold-constrained models for SPD and structured matrix learning have been developed in [4, 5, 45, 49, 57, 18, 72, 96, 24, 66]. These geometric ingredients are not merely contextual: they indicate why bilevel optimization on manifolds is practically relevant and why intrinsic analysis is needed once gradients, Hessians, and cross-derivatives live in different tangent spaces.

Bilevel optimization on manifolds. Compared with the Euclidean literature, bilevel optimization on Riemannian manifolds is still relatively undeveloped. Early work considered semivectorial formulations and KKT-type reformulations on manifolds [7, 71, 70]. These formulations are mathematically interesting, but they do not directly provide scalable gradient-based algorithms suitable for large-scale machine learning. More recently, [43] introduced an intrinsic framework for bilevel optimization on Riemannian manifolds and derived manifold hypergradient formulas together with estimation strategies based on Hessian inversion, conjugate gradient, truncated Neumann series, and automatic differentiation. [69] further developed gradient-based Riemannian bilevel algorithms and analyses. However, the existing gradient-based Riemannian bilevel theory still fundamentally assumes geodesic strong convexity of the lower-level problem.

Gap with current Riemannian bilevel theory. This is precisely where our work departs from the literature. Under geodesic strong convexity, the lower-level solution map is well behaved and the intrinsic hypergradient formula of [43, 69] is well defined through the inverse Riemannian Hessian. Under a Riemannian PL condition, by contrast, $\mathrm{Hess}_y\, g(x, y)$ need not be invertible along the iterates, and the standard inverse-Hessian hypergradient formula may fail away from the minimizer.

The analogous Euclidean difficulty already appears in nonconvex–PL bilevel optimization, where the lower-level Hessian can be singular or ill-conditioned away from the solution [17, 56, 54]; on manifolds the issue is harder because one must additionally control curvature, varying metrics, and operator comparisons across tangent spaces via parallel transport.

To address this difficulty, we introduce an intrinsic regularized tangent-space system based on spectral clipping of the Riemannian Hessian. This yields a well-defined auxiliary vector even when the lower-level Hessian is singular or indefinite along the iterates, and leads to a *Riemannian Hessian/Jacobian-free bilevel optimization* (R-HJFBiO) method based only on manifold first-order updates and intrinsic finite-difference approximations. In this way, our work connects the recent Euclidean literature on nonconvex–PL bilevel optimization [73, 17, 56, 14, 54] with the emerging Riemannian bilevel literature [43, 69], and extends manifold bilevel theory beyond geodesic strong convexity.

B Full statement of regularity assumptions

The main-text Assumptions 6 and 7 compress the Lipschitz-continuity conditions to a single representative line plus a pointer here. We restate them in full for use in the proofs.

Minimal example for Remark 4. $g(y_1, y_2) = y_2^2$ has minimizer set the y_1-axis, $\operatorname{Hess} g \equiv \operatorname{diag}(0, 2)$, and (global) PL constant $\mu = 2$, yet no neighborhood is strongly convex.

Examples for Assumption 5. We work in the Euclidean specialization for transparency. Take $\mathcal{M}_y = \mathbb{R}^2$ and $g(x, y) = \frac{1}{2}(y_2 - x)^2$. Then $\mathcal{M}^\star(x) = \{(t, x) : t \in \mathbb{R}\}$, $\operatorname{Hess}_y g = \operatorname{diag}(0, 1)$, the kernel is the tangent line of $\mathcal{M}^\star(x)$, $\mathcal{G}^2_{xy} g[v] = -v_2$, and PL holds with $\mu = 1$. *(i) Satisfied, with nontrivial f.* Let $f(x, y) = (y_2 - 1)^2 + (y_2 - x)\, y_1^2$. On $\mathcal{M}^\star(x)$, $\partial_{y_1} f = 2y_1(y_2 - x) = 0$, so Assumption 5 holds, even though both partial gradients depend on the point y_1 chosen on the line ($\mathcal{G}_x f = -y_1^2$ and $\partial_{y_2} f = 2(x - 1) + y_1^2$ there). Proposition 9 nevertheless returns $\mathcal{G}F(x) = -y_1^2 + \big(2(x - 1) + y_1^2\big) = 2(x - 1)$ for every selection, matching $F(x) = (x - 1)^2$: the selection-dependent pieces cancel exactly as the condition guarantees. *(ii) Violated.* Let $f(x, y) = y_1^2 + (y_2 - 1)^2$. The tangential component $\partial_{y_1} f = 2y_1$ is nonzero off $y_1 = 0$, and f equals $y_1^2 + (x - 1)^2$ on $\mathcal{M}^\star(x)$: any selection $s(x)$ gives $F(x) = s(x)^2 + (x - 1)^2$, so F and its gradient depend on the selection (the value function corresponds to $s \equiv 0$). *(iii) Reach.* If $\mathcal{M}^\star(x)$ is locally an orbit of a group acting by isometries under which $f(x, \cdot)$ is invariant, Assumption 5 holds automatically (invariance annihilates the tangential derivative); this covers gauge freedom such as the representative redundancy of the Grassmannian. In the least-squares template (16), with training and validation design matrices $\Phi_{\mathrm{tr}}(x)$ and $\Phi_{\mathrm{val}}(x)$, the condition is exactly $\ker \Phi_{\mathrm{tr}}(x) \subseteq \ker \Phi_{\mathrm{val}}(x)$: the upper level queries only directions that the lower level identifies.

Remark 19 (Role and scope of Assumption 5)**.** (i) Relation to [13, 14]. *Their hyper-objective is the value function $\varphi(x) = \min_{y \in \mathcal{M}^\star(x)} f(x, y)$. Under Assumption 5, $f(x, \cdot)$ is constant on the connected components of $\mathcal{M}^\star(x)$, so φ agrees locally with $f(x, y^\star(x))$ for every smooth selection within a component and inherits the gradient of Proposition 9: the assumption is the exact local condition under which the value-function and selection-based hyper-objectives coincide and are smooth.* (ii) Without it. *$F(x) = f(x, y^\star(x))$ is still well defined and differentiable for each fixed smooth selection (existence by [89] together with (P1) of Lemma 28) but selection-dependent: differentiating the stationarity condition determines $Dy^\star$ only up to $\ker \operatorname{Hess}_y g = T_{y^\star} \mathcal{M}^\star(x)$, and the tangential component of $\mathcal{G}_y f$ couples to this freedom, so no selection-free gradient exists; φ remains defined but is in general nonsmooth, consistent with the hardness results of [14].* (iii) Scope of Theorem 13. *Its proof uses Assumption 5 only through Lemma 28. Without it, Algorithm 1 is unchanged and w_t targets the field $\mathcal{G}_x f - \mathcal{G}^2_{xy} g\big[\operatorname{Hess}_y g^\dagger \, \mathsf{P}_{\mathrm{rng}} \mathcal{G}_y f\big]$ (the tangential part of $\mathcal{G}_y f$ never enters* (6)*); a certificate built from the gradient mapping of w_t then concerns this estimator field rather than a bona fide hyper-objective, and Lemma 28 is no longer available to identify it with $\mathcal{G}F$. This is the situation of our regression template whenever $\ker \Phi_{\mathrm{tr}} \not\subseteq \ker \Phi_{\mathrm{val}}$, which is one reason the diagnostic of Appendix D.3 measures the linear-system step on the retained subspace $\mathcal{R}_{\mu_c}(H)$.*

Assumption 20 (First-order regularity and boundedness, full statement)**.** *There exist constants $C_{fy}, C_{gxy}, L_f, L_g > 0$ such that for all $(x, y) \in \mathcal{U}_x \times \mathcal{U}_y$, $\|\mathcal{G}_y f(x, y)\|_y \le C_{fy}$ and*

$\|\mathcal{G}^2_{xy} g(x,y)\|_{y\to x} \le C_{gxy}$, *and for all* $x, x_1, x_2 \in \mathcal{U}_x$ *and* $y, y_1, y_2 \in \mathcal{U}_y$,

$$\begin{aligned}
\|\Gamma_{y_1\to y_2}\mathcal{G}_y f(x,y_1) - \mathcal{G}_y f(x,y_2)\|_{y_2} &\le L_f\, d(y_1,y_2),\\
\|\Gamma_{x_1\to x_2}\mathcal{G}_x f(x_1,y) - \mathcal{G}_x f(x_2,y)\|_{x_2} &\le L_f\, d(x_1,x_2),\\
\|\mathcal{G}_x f(x,y_1) - \mathcal{G}_x f(x,y_2)\|_{x} &\le L_f\, d(y_1,y_2),\\
\|\mathcal{G}_y f(x_1,y) - \mathcal{G}_y f(x_2,y)\|_{y} &\le L_f\, d(x_1,x_2),\\
\|\Gamma_{y_1\to y_2}\mathcal{G}_y g(x,y_1) - \mathcal{G}_y g(x,y_2)\|_{y_2} &\le L_g\, d(y_1,y_2),\\
\|\Gamma_{x_1\to x_2}\mathcal{G}_x g(x_1,y) - \mathcal{G}_x g(x_2,y)\|_{x_2} &\le L_g\, d(x_1,x_2),\\
\|\mathcal{G}_x g(x,y_1) - \mathcal{G}_x g(x,y_2)\|_{x} &\le L_g\, d(y_1,y_2),\\
\|\mathcal{G}_y g(x_1,y) - \mathcal{G}_y g(x_2,y)\|_{y} &\le L_g\, d(x_1,x_2).
\end{aligned}$$

The first four bounds control the partial Riemannian gradients of f, the last four those of g; the middle pair in each block (cross-variable Lipschitzness, e.g. $\mathcal{G}_x f$ in y and $\mathcal{G}_y f$ in x) is what is invoked when the proofs in Appendix G.3 bound differences of partial gradients across iterates whose x- and y-components both move. In the cross-variable lines both vectors already lie in the same tangent space ($T_x\mathcal{M}_x$ for $\mathcal{G}_x$, $T_y\mathcal{M}_y$ for $\mathcal{G}_y$), so no transport appears.

Assumption 21 (Second-order regularity, full statement)**.** *There exist constants $L_{gxy}, L_{gyy} > 0$ such that for all $x, x_1, x_2 \in \mathcal{U}_x$ and $y, y_1, y_2 \in \mathcal{U}_y$,*

$$\begin{aligned}
\|\Gamma_{x_1\to x_2}\circ \mathcal{G}^2_{xy} g(x_1,y) - \mathcal{G}^2_{xy} g(x_2,y)\| &\le L_{gxy}\, d(x_1,x_2),\\
\|\mathcal{G}^2_{xy} g(x,y_1) - \mathcal{G}^2_{xy} g(x,y_2)\circ \Gamma_{y_1\to y_2}\| &\le L_{gxy}\, d(y_1,y_2),\\
\|\operatorname{Hess}_y g(x_1,y) - \operatorname{Hess}_y g(x_2,y)\| &\le L_{gyy}\, d(x_1,x_2),\\
\|\Gamma_{y_1\to y_2}\circ \operatorname{Hess}_y g(x,y_1)\circ \Gamma_{y_2\to y_1} - \operatorname{Hess}_y g(x,y_2)\| &\le L_{gyy}\, d(y_1,y_2).
\end{aligned}$$

Assumption 22 (Retraction and vector-transport approximations)**.** *There exist constants $\bar{c} \ge 1$ and $c_R, c_T \ge 0$ such that, for all $x \in \mathcal{U}_x$, all $y, y_1, y_2 \in \mathcal{U}_y$, and all $u \in T_x\mathcal{M}_x$ (resp. $u \in T_{y_1}\mathcal{M}_y$),*

$$d^2\big(x, \operatorname{Retr}_x(u)\big) \;\le\; \bar{c}\,\|u\|_x^2, \tag{13}$$

$$\big\|\operatorname{Exp}_x^{-1}\big(\operatorname{Retr}_x(u)\big) - u\big\|_x \;\le\; c_R\,\|u\|_x^2, \tag{14}$$

$$\big\|\overline{\Gamma}_{y_1\to y_2} u - \Gamma_{y_1\to y_2} u\big\|_{y_2} \;\le\; c_T\,\|u\|_{y_1}\, d(y_1,y_2), \tag{15}$$

where Retr *is the retraction and $\overline{\Gamma}$ the vector transport used in the implementation in place of* Exp *and* Γ. *Bounds* (13) *and* (14) *are standard for first-order retractions [43, Assumption 5], [8, Sec. 3]; bound* (15) *is the standard isometric-up-to-leading-order condition on a vector transport [2, Sec. 8], [105]. On compact manifolds with smooth retractions and on Hadamard manifolds with the standard vector transport, c_R and c_T are finite and computable from sectional-curvature bounds. Setting $\bar{c} = 1$ and $c_R = c_T = 0$ recovers the exponential-map / parallel-transport case, and Assumption 22 is then automatic.*

C Comparison to Existing Works

Remark 23 (Comparison of assumptions)**.** *Table 1 relates each assumption to its counterpart in prior work. All but two have direct strong-convexity or Euclidean counterparts. The exceptions are the compatibility condition (Assumption 5), which replaces uniqueness of the lower-level solution once the minimizer set has positive dimension, and the spectral margin (Assumption 27), which is specific to the kernel-preserving clip.*

Remark 24 (Euclidean counterparts)**.** *System* (6) *generalizes the Euclidean auxiliary-vector construction of [54] (see also [68]) with two new ingredients: (a) it is posed intrinsically in the moving tangent space $T_y\mathcal{M}_y$, with the iterate transported between tangent spaces and the operator rebuilt at each outer step from geodesic probes; (b) the operator is the kernel-preserving clip followed by the pseudoinverse, rather than the clip-into-$[\mu, L_g]$-and-invert operator of [54]; the two agree on compatible right-hand sides at $\mathcal{M}^\star(x)$ and differ away from it, under noise, or without compatibility (Remark 10). Proposition 11 has a Euclidean counterpart only in degenerate form: at an isolated nondegenerate minimizer the clip is the identity and it reduces to $v^\star = \operatorname{Hess}_y g^{-1}\mathcal{G}_y f$. For $\dim \mathcal{M}^\star(x) > 0$, the clipped tracking system* (6) *and its consistency statement are, to our knowledge, new even in Euclidean space ($\zeta = 1$, $C_0 = 0$; cf. Remark 25).*

Table 1: Assumptions of this paper and their counterparts in prior work.

Assumption	Relation to prior work
1 (curvature, diameter)	Matches the curvature/diameter conditions of [43, 105].
2 (Riemannian PL)	Strictly weaker than the geodesic strong convexity of [43, 69]; Riemannian counterpart of the PL condition of [54].
5 (compatibility)	No strong-convexity counterpart; replaces uniqueness of the lower-level solution when $\dim \mathcal{M}^\star(x) > 0$, and is automatic when $\dim \mathcal{M}^\star(x) = 0$.
6–7 (regularity)	Transported analogues of the standard bilevel smoothness conditions of [33, 58, 54].
8 (regularizer)	Standard for proximal analysis.
27 (spectral margin)	Specific to the kernel-preserving clip; derived near $\mathcal{M}^\star(x)$ (Remark 26); no counterpart in [43, 54].

Table 2: Comparison with the closest Euclidean PL and Riemannian strongly-convex bilevel methods. Here $\kappa_\ell = L/\mu$ is the condition number and ζ denotes the curvature-dependent trigonometric-distance factor of (2).

Method	Geometry	Lower-level assumption	Structure	Complexity / oracle profile
RHGD-NS [43]	Riem.	Geodesic strong convexity	Double-loop	$G_f = O(\kappa_\ell^3 \varepsilon^{-1})$, $G_g = \widetilde{O}(\kappa_\ell^5 \zeta \varepsilon^{-1})$, $JV_g = O(\kappa_\ell^3 \varepsilon^{-1})$, $HV_g = \widetilde{O}(\kappa_\ell^4 \varepsilon^{-1} \log(\varepsilon^{-1}))$
RHGD-CG [43]	Riem.	Geodesic strong convexity	Double-loop	$G_f = O(\kappa_\ell^4 \varepsilon^{-1})$, $G_g = \widetilde{O}(\kappa_\ell^6 \zeta \varepsilon^{-1})$, $JV_g = O(\kappa_\ell^4 \varepsilon^{-1})$, $HV_g = \widetilde{O}(\kappa_\ell^{4.5} \varepsilon^{-1})$
RF^{2}SA [25]	Riem.	Geodesic strong convexity	Single-loop	Fully first-order (no JV_g/HV_g); deterministic $\widetilde{O}(\varepsilon^{-1.5})$; stochastic $\widetilde{O}(\varepsilon^{-2.5})$ (noise in $\mathcal{G}f$ only) and $\widetilde{O}(\varepsilon^{-3.5})$ (noise in both)
HJFBiO [54]	Euc.	PL / local-PL	Single-loop	$O(1/T)$ convergence, $O(\varepsilon^{-1})$ iteration complexity; seven first-order gradient evaluations per iteration
R-HJFBiO (Algorithm 1)	Riem.	Riemannian PL	Single-loop	$O(1/T)$ iteration convergence, $O(\varepsilon^{-1})$ iteration complexity with constants scaling polynomially in (ζ, C_0); per-iteration cost $5 + 2d_y$ first-order gradient evaluations plus an $O(d_y^3)$ eigendecomposition (basis-sweep route) or 5 gradient evaluations plus a closed-form Hessian and an $O(d_y^3)$ eigendecomposition; no JV_g or Hessian-inversion linear-system solves

D Additional bilevel-regression experiments

Datasets and licenses. The experiments in this appendix and in Section 4 use three publicly released benchmarks. **BCI Competition IV, Data set 2a** is the four-class motor-imagery EEG benchmark of the Institute for Knowledge Discovery, Graz University of Technology (Brunner, Leeb, Müller-Putz, Pfurtscheller, and Schlögl), distributed for free research use conditional on citing the BCI IV review article [99]. **UCI Superconductivity Data** [37, 38] is hosted by the UCI Machine Learning Repository as dataset #464 under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. **MiniImageNet** is the few-shot benchmark introduced by Vinyals et al. [101] from **ImageNet** (ILSVRC-2012) [90], whose standard split is that of Ravi and Larochelle [88] (our experiments instead use a seeded random 64:16:20 class split with the same class counts, Section 4); the underlying images are subject to the ImageNet Terms of Access, which permit non-commercial

research and educational use only. We used the assembled benchmark strictly within those terms and thank the dataset creators and curators in each case.

This appendix supplements Section 4 with the regression-style experiments referenced from the main text. We use a common ℓ_2-regression bilevel template to construct a singular-PL lower-level Hessian by deliberate design (Section D.1), instantiate it on three Riemannian upper-level manifolds (BCI IV-2a on $\mathrm{St}(22, 4)$ and $\mathrm{Gr}(22, 8)$, plus a synthetic Poincaré-ball regression), report results in Section D.4, and finally scale up to UCI Superconductivity to check the same diagnostic picture at $d = 70$, where the kernel of H has dimension at least 30 (Section E).

D.1 A common bilevel regression template

Each experiment instantiates a bilevel least-squares regression

$$\begin{aligned} \min_{x \in \mathcal{M}_x} \; F(x) \;=\; & \tfrac{1}{n_{\mathrm{val}}} \sum_{j=1}^{n_{\mathrm{val}}} \big(y_j - \langle\, \varphi(x; \mathbf{x}_j),\, \theta^\star(x) \,\rangle \big)^2, \\ & \text{s.t. } \theta^\star(x) \in \operatorname*{arg\,min}_{\theta \in \mathcal{M}_\theta} \tfrac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \big(y_i - \langle\, \varphi(x; \mathbf{x}_i),\, \theta \,\rangle \big)^2, \end{aligned} \tag{16}$$

where $\varphi(x; \cdot)$ is a smooth, manifold-aware feature map whose specific form changes between experiments and $\langle \cdot, \cdot \rangle$ is the inner product appropriate to the codomain of φ. Stacking training features into the design matrix $\Phi(x) \in \mathbb{R}^{n_{\mathrm{tr}} \times d}$, the Euclidean Hessian in θ is the Gram matrix $H(x) = \frac{2}{n_{\mathrm{tr}}} \Phi(x)^\top \Phi(x)$, which is independent of θ and has $\operatorname{rank} H(x) \le \min(n_{\mathrm{tr}}, d)$. By taking $n_{\mathrm{tr}} < d$ in every experiment, we force $\dim \ker H(x) \ge d - n_{\mathrm{tr}} > 0$ so $g(x, \cdot)$ is *not* strongly convex. The Polyak–Łojasiewicz inequality nonetheless holds with constant $\mu(x) = \lambda^+_{\min}(H(x))$, the smallest *nonzero* eigenvalue of $H(x)$, since the kernel component of $\theta - \theta^\star(x)$ contributes zero to both sides of PL. These statements concern a Euclidean coefficient vector θ; they apply to S2, S3 and the UCI cells, whose lower level is Euclidean, but not directly to S1 (see the caveat there).

S1 – BCI EEG on the Stiefel manifold. For each of the nine subjects $A0sT$ ($s = 1, \ldots, 9$) of BCI IV-2a, we extract the binary left-hand vs. right-hand motor-imagery contrast. Trials are bandpass-filtered to 8–30 Hz (4th-order Butterworth), epoched to a 2 s window starting 0.5 s after the cue, and reduced to their 22×22 sample covariance $C_i \in \mathbb{S}^{22}_{++}$ with a small SPD ridge and trace normalization. The upper-level variable is $W \in \mathrm{St}(22, 4)$ with the Euclidean (embedded) metric and a sign-fixed QR retraction; the lower-level variable is $M \in \mathbb{S}^4_{++}$ with the affine-invariant metric and the exponential-map retraction. The feature map is the W-projected covariance $\varphi(W; C_i) = W^\top C_i W \in \mathbb{S}^4$, paired with the Frobenius inner product. After identifying M with $\mathrm{vech}(M) \in \mathbb{R}^{10}$, we set $n_{\mathrm{tr}} = 6 < 10 = d$ (three trials per class), giving $\operatorname{rank} H(W) \le 6$ and at least four exact zero eigenvalues generically. *Caveat.* Here H is the Euclidean Hessian of the least-squares loss in the coordinates $\mathrm{vech}(M)$ (the Frobenius pairing requires a factor $\sqrt{2}$ on off-diagonal coordinates). It is not the affine-invariant Riemannian Hessian of $g(W, \cdot)$ on $\mathbb{S}^4_{++}$, which differs from it away from the minimizer and depends on M; the template's PL and rank statements therefore hold for the Euclidean coordinate system, and S1 is not covered by our verification of the theorem's assumptions.

S2 – BCI EEG on the Grassmannian. Identical preprocessing to S1; the upper-level variable is now $W \in \mathrm{Gr}(22, 8)$, represented by an orthonormal Stiefel representative modulo right-action of $\mathrm{O}(8)$. The horizontal projection is $P_W(Z) = Z - WW^\top Z$, with the QR retraction. The lower-level variable is $\theta \in \mathbb{R}^d$. Because W represents a subspace, φ must be $\mathrm{O}(8)$-invariant; we use the trace-normalised power-sum spectral features $\varphi_k(W; C) = \operatorname{tr}\big((B/\operatorname{tr} B)^{k+2}\big)$ with $B := W^\top C W$, $k = 0, \ldots, d-1$. We use $d = 8$ and $n_{\mathrm{tr}} = 6$.

S3 – Synthetic regression on the Poincaré ball. The BCI covariances have no canonical hyperbolic embedding, so we use a synthetic generator. The upper-level variable is $w \in B^n$ with the Poincaré-ball metric and the Möbius-addition exponential-map retraction; we project iterates back into a buffered ball of radius $1 - \varepsilon$ ($\varepsilon = 0.2$) for numerical stability, a standard fix in Poincaré-ball optimization [84, 32]. The lower level is Euclidean, $\theta \in \mathbb{R}^d$. The feature map uses bounded distance polynomials $\varphi_k(w; x) = \tanh\big(d_H(w, x)\big)^{k+1}$, $k = 0, \ldots, d-1$, where d_H is

the Poincaré-ball distance. These features are smooth except where w coincides with a data point ($\tanh d_H(w,x)$ is not differentiable at $w = x$); we do not enforce a separation margin, so S3 is an empirical cell and the smoothness assumptions are not claimed at such collisions. We use $n = 4$, $d = 8$, $n_{\mathrm{tr}} = 6$, $n_{\mathrm{val}} = 200$, hidden anchor at radius 0.3, and noise $\mathcal{N}(0, 0.05^2)$. With boundary buffer 0.2, the buffered ball has hyperbolic radius $2\,\mathrm{arctanh}(0.8) \approx 2.20$ and diameter $D_x = 4\,\mathrm{arctanh}(0.8) \approx 4.39$ (curvature -1), so the trigonometric-distance constant of [43] satisfies $\zeta_x \le D_x \coth(D_x) \approx 4.40$. S3 thus stresses the curvature-dependent terms in our complexity bound while holding the singular-PL trigger fixed.

D.2 Algorithms and shared hyperparameters

All five algorithms share the same outer Riemannian-gradient-descent loop: $x_{t+1} = \mathrm{Retr}_{x_t}\big(-\eta_x\, P_{x_t}(\widetilde{\nabla}_x F)\big)$. They differ only in how they assemble the auxiliary vector $v_m \approx H^{-1}\nabla_\theta f$ that enters the hypergradient $\widetilde{\nabla}_x F = \nabla_x f - (\nabla_x \nabla_\theta g)\, v_m$. **R-HJFBiO** (ours) builds $H(x_t)$ via d central finite-difference probes of $\nabla_\theta g$ with step $\delta_\varepsilon = 10^{-5}$, spectrally clips to $\mathcal{S}_{[\mu_c, L_g]}(H)$ with floor $\mu_c = 10^{-3}$ and ceiling L_g taken as the largest empirical eigenvalue, and computes the clipped pseudoinverse solution exactly from the eigendecomposition (the exact-solve variant of Section 4, not the one-step update of line 9 of Algorithm 1). **HINV** solves $(H + 10^{-8} I)v = \nabla_\theta f$ directly (the 10^{-8} is a numerical floor, not a meaningful regularizer). **CG** runs 20 conjugate-gradient steps on $Hv = \nabla_\theta f$. **NS** forms the truncated series $\alpha \sum_{k=0}^{4} (I - \alpha H)^k \nabla_\theta f$ with $\alpha = 5{\cdot}10^{-2}$. **AD** differentiates through 5 inner Euclidean GD steps with the same α, matching the AD baseline of [43].

Shared hyperparameters. Outer step $\eta_x = 5 \cdot 10^{-2}$; inner step $\eta_y = 5 \cdot 10^{-2}$; $T_{\mathrm{inner}} = 5$ Riemannian-GD steps on $\mathcal{M}_\theta$ between outer iterations (Riemannian on $\mathbb{S}^4_{++}$ for S1, Euclidean for S2/S3); $T_{\mathrm{outer}} = 100$. The Riemannian gradient is clipped to norm 1 before the retraction; the unclipped norm is what we plot. For S1 and S2 we run all 9 BCI subjects $\times$ 3 random splits (27 trajectories per algorithm); for S3 we run 5 data seeds $\times$ 3 initialization seeds (15 trajectories). Curves report the median across trajectories with shaded inter-quartile range. The choice $T_{\mathrm{inner}} = 5$ is the warm-started multi-step inner variant discussed in Appendix F (paragraph "Multi-step inner solver"), which carries over to the analysis only with the λ-cap divided by T_{in}; the gradient clipping and the exact auxiliary solve are not covered.

D.3 Comparison

A naive comparison of "gradient norms" across the five methods is misleading because each method computes a *different* gradient: each v_m defines its own $\widetilde{\nabla}_x F$, so the method's own norm $\|\widehat{\mathcal{G}} F\|^2$ is reported on five different scales. We therefore log two diagnostics that are applied identically to every method (both are defined relative to the clipped operator, see below).

Shared-reference gradient norm. At each iterate x_t we assemble a reference auxiliary vector $v^\star := \mathcal{S}_{[\mu_c, L_g]}(H)^\dagger \nabla_\theta f$ and report the norm of the resulting surrogate hypergradient (written $\|\mathcal{G}F\|_{\mathrm{sh}}$ or $\|\mathcal{G}F(x_t)\|$ in plots and tables). It equals the true hypergradient norm only at an exact lower-level minimizer under Assumption 5 and $\mu_c \le \mu$; at the inexact inner output, in the $\lambda = 10^{-6}$ cell ($\mu_c > \mu$), in the $\lambda = 0$ cell (no finite minimizer), for the S1 coordinate Hessian, and when $\ker \Phi_{\mathrm{tr}} \not\subseteq \ker \Phi_{\mathrm{val}}$ (Remark 19), it is a clipped-surrogate diagnostic and not a stationarity certificate. We separate its origin from its application: the reference is *constructed* from the same spectrally-clipped operator that R-HJFBiO uses, and *applied* uniformly to all five iterate sequences. Differences in this column therefore reflect each method's *trajectory* through $\mathcal{M}_x$, not differences in how the gradient is defined.

Why $v^\star$ is a neutral target. First, by Propositions 9 and 11, at the lower-level minimizer and for $\mu_c \le \mu$ the reference coincides with $\mathrm{Hess}_y\, g^\dagger \mathcal{G}_y f$, which under Assumption 5 yields the hypergradient: there it is the correct target independently of which method proposed the estimator. In the meta-learning experiment the reference is evaluated at the inner-solver output (30 gradient steps from $w = 0$ on each task; Appendix F); in the regression runs the inner residual is controlled by the PL linear contraction with $T_{\mathrm{in}} = 5$, and where $\ker \Phi_{\mathrm{tr}} \not\subseteq \ker \Phi_{\mathrm{val}}$ the reported quantity is the range component of Remark 19. Our bounds control the average over t, which dominates $\min_{t \le T}$; we additionally report $\min_{t \le T}$ of the squared norm in Table 4. Second, on the subspace where err_v

is measured, $\mathrm{span}\{u_i : |\lambda_i| \geq \mu_c\}$ in (17), the clip acts as the identity (the upper clamp equals the largest empirical eigenvalue), so there the reference is simply the solution of $Hv = \nabla_\theta f$ restricted to that subspace – the common target of HINV, CG, and NS. The single method-linked ingredient is the threshold μ_c delimiting the subspace, not the reference values on it. Because R-HJFBiO (exact-solve variant) computes exactly this vector, its $\mathrm{err}_v = 0$ holds by construction: err_v measures how far each baseline is from the clipped target, not the accuracy of the one-step tracking update of Algorithm 1.

Robustness to the reference. The shared reference $v^\star = \mathcal{S}_{[\mu_c, L_g]}(H)^\dagger \nabla_w f$ is built with the operator of R-HJFBiO, so we re-evaluate every reference-dependent panel against an independent reference: the minimum-norm least-squares solution $v_{\mathrm{LS}} = H^\dagger \nabla_w f$, computed in float64 with NumPy's default rank cutoff ($\sigma_{\max} \cdot dK \cdot \varepsilon_{\mathrm{mach}}$), untied to μ_c, at every iterate of every trajectory (Figure 2 and Table 3; no retraining). The two references can differ only on eigen-directions with $|\lambda| < \mu_c$. At $\lambda = 10^{-2}$ every eigenvalue exceeds μ_c and the references coincide: the gradient norms agree to within 0.14% and err_v to within $6 \cdot 10^{-5}$ (float32 precision). At $\lambda = 10^{-6}$ the least-squares cutoff keeps all 2000 directions, including the ridge-lifted flat directions with eigenvalues near 10^{-6}, on which v_{LS} is amplified by up to 10^6; under v_{LS} the final squared gradient norm is 10^9–10^{12} and $\mathrm{err}_v \approx 1$ for every method. At $\lambda = 0$ the numerical rank kept by least squares varies between 38 and 100 along the trajectories, and the spectrum shows no clean gap between the kernel and the signal band (eigenvalues below $9.9 \cdot 10^{-9}$ versus above $1.0 \cdot 10^{-8}$), so v_{LS} is dominated by directions at the level of numerical noise. Two conclusions follow. First, the Neumann proxy for AD's auxiliary vector is far from the solution under both references ($\mathrm{err}_v \geq 0.99$ in both low-λ regimes), so that finding is not an artifact of tying the reference to our operator; conversely, R-HJFBiO's own error under v_{LS} is also ≈ 1 at $\lambda = 10^{-6}$ (Table 3), so this check does not validate the clipped field as the true hypergradient. Second, in the low-λ regimes the unclipped pseudoinverse does not provide a usable reference, and the reference-dependent diagnostics are measured relative to μ_c, which determines the eigen-directions the hypergradient retains.

Table 3: Final-step diagnostics at $\lambda = 10^{-6}$ under the shared (clipped) reference and the independent minimum-norm least-squares reference (no μ_c; NumPy's default rank cutoff). err_v (ls) is measured on the subspace kept by the least-squares cutoff.

Method	$\|\mathcal{G}F\|^2_{\mathrm{sh}}$	$\|\mathcal{G}F\|^2_{\mathrm{ls}}$	err_v (shared)	err_v (ls)
HINV	—	—	—	—
CG	$4.8 \cdot 10^4$	$2.4 \cdot 10^{12}$	$1.9 \cdot 10^0$	$1.0 \cdot 10^0$
NS	$9.1 \cdot 10^4$	$2.8 \cdot 10^{11}$	$1.0 \cdot 10^0$	$1.0 \cdot 10^0$
AD	$2.0 \cdot 10^2$	$9.9 \cdot 10^9$	$1.0 \cdot 10^0$	$1.0 \cdot 10^0$
R-HJFBiO	$4.9 \cdot 10^0$	$2.4 \cdot 10^9$	0	$1.0 \cdot 10^0$

v-solver relative error on the retained subspace $\mathcal{R}_{\mu_c}(H)$. For each method's actual v_m, we report

$$\mathrm{err}_v \;=\; \frac{\|P_{\mathrm{rg}}(v_m - v^\star)\|}{\|P_{\mathrm{rg}}\, v^\star\|}, \qquad P_{\mathrm{rg}} \;=\; \text{projection onto } \mathcal{R}_{\mu_c}(H) := \mathrm{span}\{u_i : |\lambda_i| \geq \mu_c\}, \quad (17)$$

where $\{(\lambda_i, u_i)\}$ is the eigendecomposition of H. The retained subspace $\mathcal{R}_{\mu_c}(H)$ equals $\mathrm{range}(H)$ only when no nonzero eigenvalue lies below μ_c; it is strictly smaller for a positive ridge. Projecting onto $\mathcal{R}_{\mu_c}(H)$ before measuring the error is essential: in the singular regime the kernel of H is large, and any difference there is not meaningful (the spectrally-clipped pseudoinverse defines v to be zero on that subspace). For AD, which has no explicit v_m, we use the truncated Neumann series with AD's hyperparameters as a proxy: backprop through T_{ad} steps of GD with step α on a quadratic loss (fixed Hessian and cross-derivative) is equivalent to applying $\alpha \sum_{k=0}^{T_{\mathrm{ad}}-1} (I - \alpha H)^k$. For the cross-entropy loss of the meta-learning experiment the Hessian varies along the inner trajectory, so the AD entries of err_v there describe this proxy, not AD's actual unrolled derivative. The two diagnostics separate trajectory quality from the linear-system step, but both are defined relative to the clipped target.

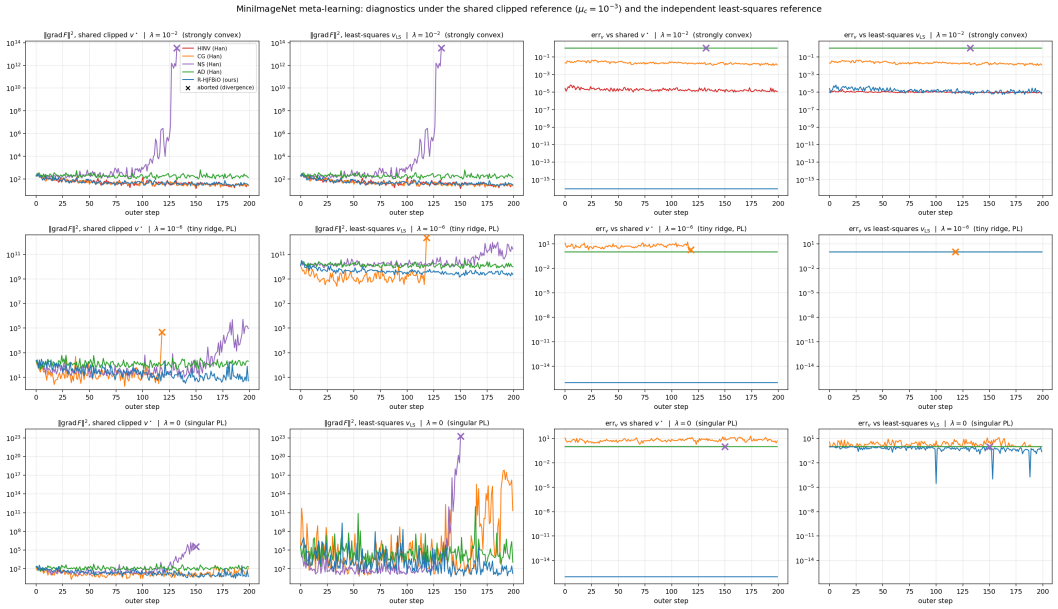


Figure 2: **MiniImageNet meta-learning under the shared and the independent reference.** Rows: $\lambda = 10^{-2}, 10^{-6}, 0$. Columns: squared gradient norm under the shared clipped reference $v^\star$ ($\mu_c = 10^{-3}$) and under the least-squares reference v_{LS}, then err_v against each. The F panel does not depend on the reference and is shown in Figures 1 and 9. Crosses ($\times$) mark divergence-guard aborts.

D.4 Results

Figures 3, 4, and 5 report panels (a) the upper-level objective $F(x_t)$ on validation, (c) the shared-reference Riemannian gradient norm, and (d) the v-solver relative error on $\mathcal{R}_{\mu_c}(H)$ from (17), on the BCI IV-2a Stiefel instance, the BCI IV-2a Grassmann instance, and the synthetic Poincaré-ball instance respectively. The three geometries behave differently and are discussed in turn; the Stiefel case is repeated at larger ambient dimension in Section E (Figure 6).

S1: Stiefel. On $\text{St}(22, 4)$ with $n_{\text{tr}} = 6 < 10 = d$, the rank-deficiency diagnosis is the same as for S2 below. Figure 3 shows R-HJFBiO at machine precision $\sim 10^{-17}$ on $\mathcal{R}_{\mu_c}(H)$ (by construction), while CG's v-solver error stays near 10^{19} (its un-regularized linear solve is overwhelmed by the kernel of H), NS rises from 10^1 to about 10^5, AD stays near 10^1, and HINV near 10^{-4}. All five shared-reference gradient norms decrease from about $2 \cdot 10^2$; R-HJFBiO's decreases furthest, to about 10^1 by steps 40–60, while HINV, CG and AD level off near 40 and NS near 70. All five methods reduce F to comparable values near 1.0 (AD slightly lowest) – the upper level is bounded – and R-HJFBiO attains the lowest shared-reference norm; this norm is a clipped-surrogate diagnostic of the Euclidean coordinate Hessian (see the S1 caveat), not a stationarity certificate. A similar v-solver picture holds at larger ambient dimension in Figure 6, where, however, NS and AD also descend.

S2: Grassmann. S2 is the cleanest diagnostic of the head-to-head protocol. The Grassmannian quotient removes the $\text{O}(p)$ gauge redundancy that S1 carries, so the residual differences between methods are entirely due to how each handles the rank-deficient H. In panel (d) of Figure 4, R-HJFBiO is at $\sim 10^{-17}$ (by construction), HINV and CG are close to the clipped target at about 10^{-6}–10^{-5}, and AD stays near 10^1. In panel (a), R-HJFBiO lowers F to about 0.85 while the four baselines stay near 1.0; in panel (c), R-HJFBiO, HINV and CG track each other at about 30–60, while NS and AD rise to about $3 \cdot 10^2$ before dropping near the end. On this cell the inverse-Hessian baselines therefore solve the linear system accurately, and the separation in panel (d) is with AD.

S3: Hyperbolic. On the Poincaré ball ($\zeta_x \leq 4.40$, Hadamard manifold), R-HJFBiO and AD descend F to about 0.63, while HINV and CG plateau near 0.85. In panel (c) no method separates:

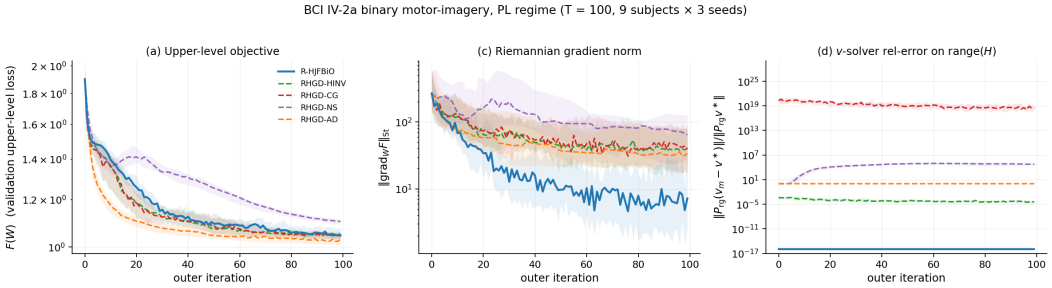


Figure 3: **S1 (Stiefel).** BCI IV-2a on $\mathrm{St}(22, 4)$ with $d = 10$, $n_{\mathrm{tr}} = 6$ (so $\dim \ker H \geq 4$). Median over 9 subjects $\times$ 3 random splits with shaded inter-quartile range. R-HJFBiO's v-solver error sits at machine precision on $\mathcal{R}_{\mu_c}(H)$ (by construction) and its shared-reference gradient norm decreases furthest (to about 10^1, against about 40–70 for the baselines); CG stays near 10^{19}, NS rises to about 10^5, AD stays near 10^1 and HINV near 10^{-4}.

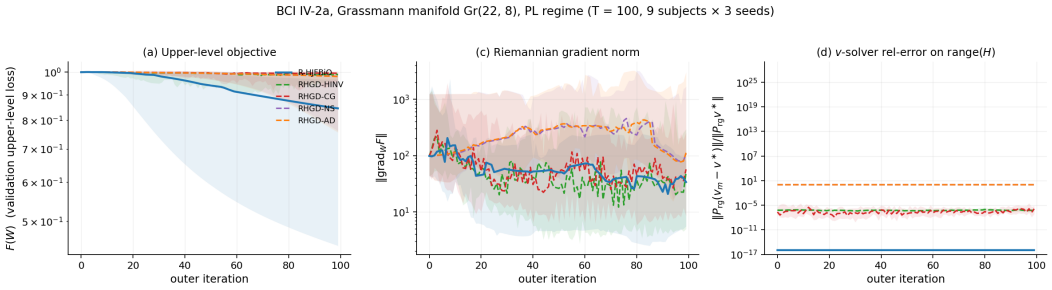


Figure 4: **S2 (Grassmann).** BCI IV-2a on $\mathrm{Gr}(22, 8)$, same protocol as S1 with $d = 8$, $n_{\mathrm{tr}} = 6$. v-solver error on $\mathcal{R}_{\mu_c}(H)$: R-HJFBiO at $\sim 10^{-17}$ (by construction), HINV and CG near 10^{-6}–10^{-5}, AD near 10^1. R-HJFBiO lowers F to about 0.85 while the baselines stay near 1.0.

R-HJFBiO's shared-reference norm dips from about $1.5 \cdot 10^3$ to about 35 near step 15, rises again (with a spike near step 65) and ends near $2 \cdot 10^2$, comparable to the baselines (about 2–$4 \cdot 10^2$). In panel (d), R-HJFBiO is at machine precision (by construction), AD stays near 10^1, CG falls from about 10^2 to about 10^{-5}, and HINV stays near 10^{-6}. AD thus matches R-HJFBiO on F while its auxiliary vector stays far from the clipped target; the inverse-Hessian baselines are close to the target in this cell but plateau higher in F. This cell shows no gradient-norm advantage for R-HJFBiO under negative curvature.

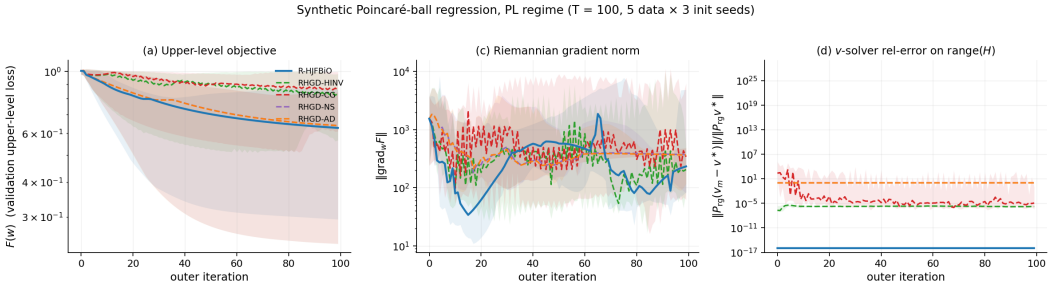


Figure 5: **S3 (Poincaré ball).** Synthetic regression on B^4, median over 5 data seeds $\times$ 3 init seeds. Under negative curvature ($\zeta_x \leq 4.40$), panel (a) shows R-HJFBiO and AD descending F to about 0.63 while HINV and CG plateau near 0.85; in panel (c) the shared-reference norms of all five methods end at comparable levels (about 2–$4 \cdot 10^2$). Panel (d) shows the linear-system step: R-HJFBiO at $\sim 10^{-17}$ (by construction), AD near 10^1, CG and HINV near 10^{-6}–10^{-5}.

Summary. Across the three geometries – positively curved Stiefel, quotient Grassmann, and Hadamard hyperbolic – the picture is mixed. CG is far from the clipped target only on S1 (about 10^{19}, driven by the kernel of H); on S2 and S3, HINV and CG reach relative errors of about 10^{-6}–10^{-5}, and HINV is near 10^{-4} on S1. AD stays near 10^{1} in all three cells and NS is far from the target on S1 – the diagnostic protocol of Section D.3 is what makes this bias visible. R-HJFBiO attains the lowest shared-reference gradient norm on S1, tracks HINV and CG on S2, and shows no gradient-norm advantage on S3; it lowers F furthest on S2, ties AD on S3, and is comparable to the baselines on S1. Its machine-precision v-solver diagnostic holds by construction. These runs use the exact-solve variant, so they do not test the tracking recursion analyzed in Section 3.

E Stochastic Minibatches in UCI Superconductivity

The BCI experiments above use small lower-level dimensions ($d \in \{8, 10\}$) chosen so the rank-deficiency is easy to visualise. We now scale to a regression problem – predicting the critical temperature of a superconducting compound from 81 chemical-formula features (UCI Superconductivity, 21 263 samples) – to check the diagnostic picture at $d = 70$, where the kernel of H has dimension at least 30 (at least 35 on the Grassmann cell). We run all three manifolds with the full training batch ($B = n_{\mathrm{tr}} = 40$) for the lower-level and auxiliary computations, so the only stochastic ingredient is the validation mini-batch used for the upper-level quantities, and the diagnostic protocol of Section D.3 measures only the linear-system step and the upper-level trajectory.

Dataset and bilevel formulation. We standardize each input feature on the training split, divide each row by $\sqrt{n_{\mathrm{in}}}$ to keep ambient norms order 1, and work with the same least-squares bilevel template (16). We use $n_{\mathrm{tr}} = 40$ and $n_{\mathrm{val}} = 20\,000$. The lower-level dimension is $d = 70$; with $n_{\mathrm{tr}} = 40 < d$, $\dim\ker H \geq 30$ by construction. On $\mathrm{St}(81, 70)$ and B^8, $\mathrm{rank}\, H \leq 40$, so at least 43% of the lower-level Hessian spectrum is exactly zero; on $\mathrm{Gr}(81, 35)$ the design matrix $XWW^\top R$ has rank at most 35, so $\dim\ker H \geq 35$ (at least 50%).

Manifold instances and feature maps. The three instances now cover larger ambient dimensions. *Stiefel* $W \in \mathrm{St}(81, 70)$ uses the linear feature map $\varphi(W; x) = W^\top x$. *Grassmann* $W \in \mathrm{Gr}(81, 35)$ represents a rank-35 subspace via a Stiefel rep; the feature map $\varphi(W; x) = R^\top WW^\top x$ uses a fixed Gaussian sketch $R \in \mathbb{R}^{81\times 70}$ to lift the projector $WW^\top$ to d features, and is $\mathrm{O}(35)$-invariant by construction. *Poincaré ball* $w \in B^8$ uses $\varphi_k(w; x) = (a_k^\top x)\, \exp\!\big(-d_H(w, p_k)^2/2\sigma^2\big)$, so the k-th feature is a linear projection along a fixed direction $a_k \in \mathbb{R}^{81}$, gated by a hyperbolic radial-basis weight measuring how close the upper-level point w is to the fixed anchor p_k in B^8. Compared to the polynomial-in-distance design used for the synthetic Poincaré experiment, the RBF gate lies in $(0, 1]$, so $|\varphi_k(w; x)| \leq |a_k^\top x|$ regardless of d, eliminating the exponential blow-up of d^k-style designs that otherwise renders H numerically singular even before the $n_{\mathrm{tr}} < d$ trigger.

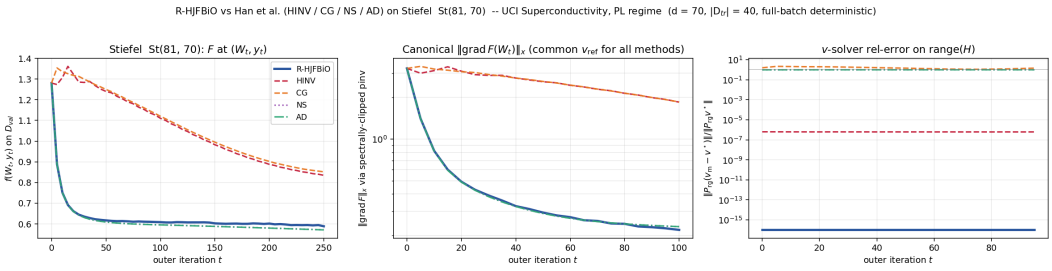


Figure 6: **UCI Superconductivity, $\mathrm{St}(81, 70)$, full training batch ($B = n_{\mathrm{tr}} = 40$).** Diagnostic protocol of Section D.3: $f(W_t, y_t)$ on D_{val} (left), the shared-reference (clipped-surrogate) gradient norm $\|\mathcal{G}F(W_t)\|_x$ (centre), and the v-solver relative error on $\mathcal{R}_{\mu_c}(H)$ (right). R-HJFBiO/NS/AD drop F to ≈ 0.6 within 50 iterations and are visually indistinguishable in the shared-reference gradient norm; HINV and CG descend extremely slowly and plateau near 0.85, an artifact of their inverse-Hessian solve being dominated by the kernel of H. R-HJFBiO sits at machine precision on $\mathcal{R}_{\mu_c}(H)$ by construction (right panel, $\sim 10^{-16}$); the four baselines stagnate at 10^{-7}–10^{0}.

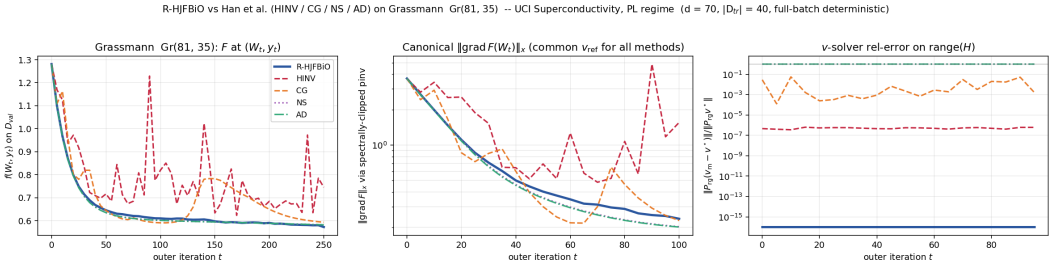


Figure 7: **UCI Superconductivity, $\mathrm{Gr}(81, 35)$, full training batch.** Same protocol as Figure 6. R-HJFBiO/NS/AD match each other on F and the shared-reference gradient norm; HINV's F-trajectory becomes visibly oscillatory (the kernel-amplified hypergradient produces large spurious steps that the QR retraction absorbs but does not damp), and CG oscillates similarly though with smaller amplitude and recovers in the second half of the run. Right panel: R-HJFBiO at machine precision by construction; the baselines span 10^{-7}–10^{0}.

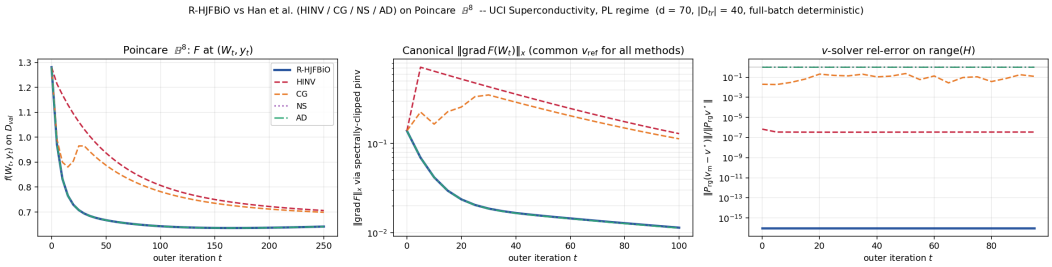


Figure 8: **UCI Superconductivity, B^8, full training batch.** Same protocol as Figure 6. R-HJFBiO/NS/AD are visually indistinguishable in panels (a) and (c) and reach $F \approx 0.65$; HINV and CG flatten at ≈ 0.7 with the inverse-Hessian solve again kernel-dominated. Right panel: R-HJFBiO at machine precision (by construction), baselines clustered at 10^{-7}–10^{0}.

Head-to-head deterministic baseline. Figures 6–8 repeat the apples-to-apples protocol of Section D.3 in full-training-batch mode ($B = n_{\mathrm{tr}}$). R-HJFBiO's v_m matches $v^\star$ on $\mathcal{R}_{\mu_c}(H)$ at every iterate by construction (rightmost panel of every figure, $\sim 10^{-16}$ vs. 10^{-7}–10^{0} for the four baselines), and its shared-reference gradient norm decreases steadily by 1–2 orders of magnitude on every manifold, as do those of NS and AD. Two new effects show up at this scale. *First*, on $\mathrm{St}(81, 70)$ (Figure 6), HINV and CG are indistinguishable in F: they descend extremely slowly, plateauing near 0.85 while R-HJFBiO/NS/AD reach 0.6 in fewer than 50 iterations. The story is the same as for BCI – the inverse-Hessian solve is dominated by the kernel of H – but the dynamic range is wider here ($d = 70$ vs. 10). *Second*, on $\mathrm{Gr}(81, 35)$ (Figure 7), HINV's F-trajectory becomes visibly *oscillatory*: the kernel-amplified hypergradient produces large spurious steps that the QR retraction absorbs but does not damp, producing the spike pattern in panel (a). NS and AD match R-HJFBiO on F and on the shared-reference norm in every panel, although the rightmost column shows their v_m is far from the clipped target – a distinction the F-only or own-norm view would miss. On the Poincaré ball B^8 (Figure 8), R-HJFBiO/NS/AD are visually indistinguishable in F and the shared-reference gradient norm but the same right-column gap appears, confirming that negative curvature does not rescue the inverse-Hessian baselines from the kernel-amplification problem.

Mini-batch sweep. The stochastic algorithm SR-HJFBiO of Appendix H comes with an $O(1/T + \delta_\varepsilon^2 + 1/B)$ upper bound (Theorem 17); an upper bound does not by itself predict a matching $\Theta(1/B)$ empirical decay. On the same UCI problem we ran SR-HJFBiO with $B \in \{1, 4, 16, 40\}$; these runs enter only the spectral-margin check of Table 5, which passes for $B = 40$ and fails for $B \leq 16$, and we draw no empirical conclusion from them about the dependence on B.

F Meta-learning experiment: full details

This appendix supplements Section 4 with the full architecture, optimization configuration, and additional discussion that were too long for the main text.

Numerical trade-off in the finite-difference probes. The central finite-difference probes in Eqs. (9)–(10) present a practical numerical trade-off. Under the Lipschitz-Hessian assumptions of our analysis the truncation error of the operator is $O(\delta_\varepsilon)$ (it enters our bounds squared, as $O(\delta_\varepsilon^2)$), and it competes with $O(\eta_{\mathrm{mach}}/\delta_\varepsilon)$ floating-point cancellation as $\delta_\varepsilon \to 0$; balancing these first-order terms gives $\delta_\varepsilon \sim \eta_{\mathrm{mach}}^{1/2}$. The familiar $\eta_{\mathrm{mach}}^{1/3}$ rule for central differences requires a second-order truncation bound, i.e. stronger (third-derivative) regularity. For the quadratic lower levels of our regression cells the central difference is exact up to rounding. Our experiments use $\delta_\varepsilon = 10^{-5}$ in double precision, comfortably above the cancellation regime, but very small stationarity tolerances under the schedule $\delta_\varepsilon = \Theta(\sqrt{\varepsilon})$ may eventually push the step toward it; the closed-form Hessian realization (Section 3, route (b)) sidesteps this entirely when $\mathrm{Hess}_y\, g$ is available, and is what we use in the meta-learning experiment.

Backbone architecture. The 4-block CNN is a standard Conv-4 design used widely in few-shot learning. Each block is $\mathrm{Conv}_{3\times3} \to \mathrm{BN} \to \mathrm{ReLU} \to \mathrm{MaxPool}_{2\times2}$, with 16 output channels per layer and an input resolution of 84×84. After four blocks the feature map is $5 \times 5 \times 16$, which we flatten to a 400-dimensional embedding. The first two convolutional kernels are reshaped to matrices of size $27 \times 16 = (3 \cdot 3 \cdot 3) \times 16$ and $144 \times 16 = (3 \cdot 3 \cdot 16) \times 16$ – so each column corresponds to one output channel viewed as a flattened receptive field – and constrained to lie on Stiefel manifolds, with the Euclidean (embedded) metric, whose Riemannian gradient is the tangent projection $P_W(Z) = Z - W\,\mathrm{sym}(W^\top Z)$ of the ambient gradient Z, and a sign-corrected QR retraction. We also report runs with W_1, W_2 on the Grassmannian $\mathrm{Gr}(n, 16)$ (using the horizontal projector $P_W(Z) = Z - WW^\top Z$ and the same QR retraction) and the Oblique manifold $\mathrm{OB}(n, 16) = (S^{n-1})^{16}$ (column-wise sphere normalization), keeping all other architectural and optimization choices identical. The Grassmann runs update orthonormal representatives with the horizontal projector; because batch normalization and ReLU act channel-wise, the network is not invariant under $W \mapsto WQ$, $Q \in \mathrm{O}(16)$, so these runs define an orthonormally constrained representative model rather than a well-defined objective on the Grassmannian.

Lower-level objective and closed-form Hessian. Given backbone parameters Θ and support features $F_s = \phi_\Theta(X_s) \in \mathbb{R}^{N_s K \times d}$ (with $K = 5$ classes, $N_s = 5$ support images per class, $d = 400$), the task-specific classifier $w \in \mathbb{R}^{d\times K}$ minimizes

$$g(\Theta, w) \;=\; \frac{1}{N_s K} \sum_{n=1}^{N_s K} \mathrm{CE}\big(F_s\, w,\; y_s\big) \;+\; \frac{\lambda}{2}\|w\|_F^2.$$

A direct computation gives

$$\nabla_w^2\, g(\Theta, w) \;=\; \frac{1}{N_s K} \sum_n \big(F_n F_n^\top\big) \otimes M_n \;+\; \lambda\, I_{dK}, \qquad M_n \;=\; \mathrm{diag}(p_n) - p_n p_n^\top,$$

where $p_n = \mathrm{softmax}(F_n^\top w) \in \mathbb{R}^K$. This closed-form Hessian is precisely the realization (b) of the operator-level Hessian approximation $\widehat{\mathcal{H}}_t$ described in Section 3: the operator $\widehat{\mathcal{H}}_t$ of line 7 of Algorithm 1 is constructed directly rather than via finite-difference probes, and the spectral clip of line 8 is then applied to it. The per-task cost is dominated by a single `eigh` on the $(dK) \times (dK) = 2000 \times 2000$ matrix and runs in $\sim$1 s on CPU. The matrix M_n is rank $K-1 = 4$ for every sample, so when $\lambda = 0$ the rank of $\nabla_w^2 g$ is at most $N_s K(K-1) = 100$, falling short of the ambient dimension $dK = 2000$ by a factor of 20 – roughly 95% of the spectrum is exactly zero. Rank deficiency does not imply PL: when $\lambda = 0$ and the $N_s K = 25$ support features are linearly independent (generic for $d = 400$), the support labels are separable, the infimum of $g(\Theta, \cdot)$ is not attained, and no uniform PL constant exists, so this regime lies outside Assumption 2. At $\lambda = 10^{-6}$ the problem is strongly convex, but its PL constant is at most $\lambda < \mu_c$. We emphasize that this materialization of the Hessian is consistent with the "Hessian/Jacobian-free" label in the sense clarified in Remark 12: no linear system is solved, and no implicit differentiation through the inner loop is performed.

Table 4: Final-step metrics (mean over the meta-batch) on the Stiefel backbone. Steps counts the completed outer steps out of $T = 200$; daggers (†) mark runs aborted by the divergence guard before T, for which every other column is taken at the last completed step. $\|\widehat{\mathcal{G}}F\|^2$ is the method's own hypergradient norm; $\|\mathcal{G}F\|^2_{\mathrm{sh}}$ is the shared-reference (clipped-surrogate) gradient norm, whose reference $v^\star$ is constructed from the clipped pseudoinverse (origin) and applied identically to every method's iterates (application; Appendix D.3); $\min_t$ is its minimum over $t \le T$; err_v is the relative v-solver error on $\mathcal{R}_{\mu_c}(H)$. R-HJFBiO and AD are the only methods that complete all three regimes; R-HJFBiO's own norm matches the shared reference by construction, and AD has the highest query accuracy (q-acc) in every regime. All values come from a single four-task meta-batch without repeated runs.

Regime	Method	Steps	$F(\Theta_t, w_t)$	q-acc	$\|\widehat{\mathcal{G}}F\|^2$	$\|\mathcal{G}F\|^2_{\mathrm{sh}}$	$\min_t \|\mathcal{G}F\|^2_{\mathrm{sh}}$	err_v
$\lambda = 10^{-2}$	HINV	200	1.29	0.51	$2.72 \cdot 10^{1}$	$2.72 \cdot 10^{1}$	$1.40 \cdot 10^{1}$	$1.13 \cdot 10^{-5}$
	CG	200	1.31	0.48	$2.48 \cdot 10^{1}$	$2.48 \cdot 10^{1}$	$1.85 \cdot 10^{1}$	$1.18 \cdot 10^{-2}$
	NS	133†	$1.44 \cdot 10^{4}$	0.25	$6.78 \cdot 10^{7}$	$3.44 \cdot 10^{13}$	$1.11 \cdot 10^{2}$	0.998
	AD	200	1.40	0.56	$1.49 \cdot 10^{0}$	$1.29 \cdot 10^{2}$	$7.98 \cdot 10^{1}$	0.997
	R-HJFBiO	**200**	1.27	0.49	$2.96 \cdot 10^{1}$	$2.96 \cdot 10^{1}$	$2.22 \cdot 10^{1}$	**0**
$\lambda = 10^{-6}$	HINV	0†	—	—	—	—	—	—
	CG	119†	5.82	0.29	$2.59 \cdot 10^{7}$	$4.81 \cdot 10^{4}$	$1.82 \cdot 10^{0}$	1.85
	NS	200	3.49	0.27	$1.18 \cdot 10^{3}$	$9.12 \cdot 10^{4}$	$9.64 \cdot 10^{0}$	0.999
	AD	200	1.43	0.55	$1.51 \cdot 10^{0}$	$2.02 \cdot 10^{2}$	$3.96 \cdot 10^{1}$	0.999
	R-HJFBiO	**200**	1.47	0.41	$4.93 \cdot 10^{0}$	$\mathbf{4.93 \cdot 10^{0}}$	$3.16 \cdot 10^{0}$	**0**
$\lambda = 0$	HINV	0†	—	—	—	—	—	—
	CG	200	3.25	0.29	$9.39 \cdot 10^{3}$	$1.32 \cdot 10^{2}$	$1.93 \cdot 10^{0}$	4.89
	NS	151†	14.11	0.35	$4.10 \cdot 10^{6}$	$3.52 \cdot 10^{5}$	$1.03 \cdot 10^{1}$	1.02
	AD	200	1.42	0.56	$1.42 \cdot 10^{0}$	$2.18 \cdot 10^{2}$	$4.16 \cdot 10^{1}$	0.999
	R-HJFBiO	**200**	1.41	0.43	$5.82 \cdot 10^{0}$	$\mathbf{5.82 \cdot 10^{0}}$	$2.94 \cdot 10^{0}$	**0**

Outer loop and divergence guard. The outer loop runs Riemannian SGD on the manifold parameters (Euclidean-metric tangent projection followed by sign-corrected QR retraction) and Euclidean SGD on the rest, for $T = 200$ outer iterations with meta-batch size 4. Every outer iteration draws a fresh meta-batch and runs 30 inner gradient steps from $w = 0$ on each task before assembling the hypergradient. We employ a divergence guard: any iteration whose own squared hypergradient norm $\|\widehat{\mathcal{G}}F\|^2$ exceeds 10^8 aborts cleanly so subsequent iterates do not feed `Inf/NaN` into the next `eigh`. Daggered entries in Table 4 indicate runs aborted by this guard; Figure 9 shows the corresponding trajectories for the strongly-convex and tiny-ridge regimes (the singular-Hessian regime is in Figure 1).

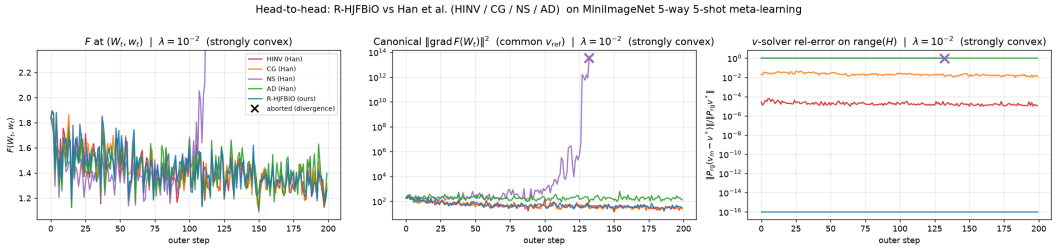
Multi-step inner solver. Algorithm 1 as analyzed in Section 3 performs a single inner step per outer iteration; the regression experiments of Appendix D use $T_{\mathrm{in}} = 5$ warm-started inner steps and this meta-learning experiment uses $T_{\mathrm{in}} = 30$ steps started from $w = 0$ on each freshly drawn task, in both cases with the hypergradient assembled *after* the inner steps. For a fixed objective with warm start, $y_t^{(0)} = y_t$, $y_t^{(k+1)} = \mathrm{Exp}_{y_t^{(k)}}\big(-\lambda\, \mathcal{G}_y g(x_t, y_t^{(k)})\big)$, and $y_{t+1} = y_t^{(T_{\mathrm{in}})}$, the proof of Lemma 32 with x_t fixed (so $\xi_t = 0$) gives, for $a_k := g(x_t, y_t^{(k)}) - G(x_t)$ and $d_k := d_y(y_t^{(k+1)}, y_t^{(k)})$,

$$a_{k+1} \le (1 - \lambda\mu)\, a_k - \frac{1}{4\lambda}\, d_k^2 \qquad \text{under } \lambda \le 1/(2\zeta L_g).$$

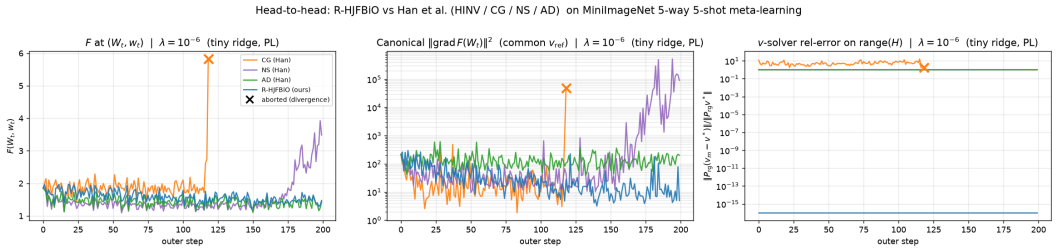
Iterating keeps geometric weights on the step terms:

$$a_{T_{\mathrm{in}}} \le (1-\lambda\mu)^{T_{\mathrm{in}}}\Delta_t - \frac{1}{4\lambda}\sum_k (1-\lambda\mu)^{T_{\mathrm{in}}-1-k} d_k^2 \le (1-\lambda\mu)^{T_{\mathrm{in}}}\Delta_t - \frac{(1-\lambda\mu)^{T_{\mathrm{in}}-1}}{4\lambda}\sum_k d_k^2.$$

The x-step half of that proof, applied at $y = y_{t+1}$, gives $\Delta_{t+1} \le (1 + 4L_g^2\gamma/\mu)\, a_{T_{\mathrm{in}}} + (\frac{1}{8\gamma} + L_G)\, d_x(x_{t+1}, x_t)^2$, and $(1 + 4L_g^2\gamma/\mu)(1-\lambda\mu)^{T_{\mathrm{in}}} \le (1 + 4L_g^2\gamma/\mu)(1-\lambda\mu) \le 1 - \lambda\mu/2$ under the



(a) Strongly-convex regime, $\lambda = 10^{-2}$.



(b) Tiny-ridge regime, $\lambda = 10^{-6}$.

Figure 9: **Trajectories for the two regimes deferred from Figure 1.** Same protocol, panels, and reference $v^\star$ as Figure 1. (a) At $\lambda = 10^{-2}$, four methods complete the trajectory and NS diverges in the second half of the run. (b) At $\lambda = 10^{-6}$, HINV aborts at step 0, CG diverges mid-run, and NS completes the run only with a sharply rising objective and gradient norm; R-HJFBiO and AD also complete the run – but AD's err_v remains ≈ 1 throughout. Crosses ($\times$) mark divergence-guard aborts.

caps of Lemma 32, so

$$\Delta_{t+1} \le \Big(1 - \frac{\lambda\mu}{2}\Big)\Delta_t + \Big(\frac{1}{8\gamma} + L_G\Big) d_x(x_{t+1}, x_t)^2 - \frac{(1-\lambda\mu)^{T_{\mathrm{in}}-1}}{4\lambda} \sum_{k=0}^{T_{\mathrm{in}}-1} d_k^2.$$

The weight $(1-\lambda\mu)^{T_{\mathrm{in}}-1}$ on the step terms is needed: with the full factor $(1-\lambda\mu/2)^{T_{\mathrm{in}}}$ on Δ_t and unweighted step terms, the right-hand side is negative for $g(y) = y^2/2$, $\lambda = 0.1$, $T_{\mathrm{in}} = 30$, $y_t = 1$ and no x-movement. If in addition $\lambda \le 1/(2\mu T_{\mathrm{in}})$, then $(1-\lambda\mu)^{T_{\mathrm{in}}-1} \ge 1/2$, and with $d_y(y_{t+1}, y_t)^2 \le T_{\mathrm{in}} \sum_k d_k^2$ the last term is at most $-\frac{1}{8\lambda T_{\mathrm{in}}} d_y(y_{t+1}, y_t)^2$. The drift term $\frac{50\bar{L}^2}{3\mu_c\tau} d_y(y_{t+1}, y_t)^2$ of Lemma 29 is then absorbed if $\lambda \le 3\mu_c\tau/(400\bar{L}^2 T_{\mathrm{in}})$, and the extra $L_G d_x^2 = L_G\gamma^2\|\mathcal{G}_t\|^2 \le \frac{\gamma}{32}\|\mathcal{G}_t\|^2$ (using $\gamma \le \lambda\mu/(16L_G)$ and $\lambda\mu \le 1/2$) fits in the slack of (53). With these caps the $O(\varepsilon^{-1})$ outer-iteration complexity persists for the evaluation order of Algorithm 1. The argument does *not* cover (a) the meta-learning procedure, which samples a fresh task and resets the head at every outer step, so the inner gap at the start of an outer step is not the Δ_t of the previous step; (b) evaluating the hypergradient after rather than before the inner steps; or (c) the exact auxiliary solve and gradient clipping used in all experiments. In these respects the experiments are empirical variants; likewise, an RHGD-style double loop [43] with our clipped-pseudoinverse v-step is an empirical variant, since its exact v-step is not the update of line 9. The head-to-head comparison of Section 4 is nevertheless controlled: all five methods share one outer loop, inner solver, and schedule and differ only in how v_m is assembled.

Spectral margin and rank along the runs. Assumption 27 is derived, not assumed, near $\mathcal{M}^\star(x)$ with $\mu_c = \mu/2$ (Remark 26); this value of μ_c only certifies that the margin holds automatically there. Our implementation instead sets μ_c without knowledge of μ (10^{-3} for the meta-learning, BCI, and Poincaré cells; 10^{-2} for UCI Superconductivity), and for this practical choice it is the empirical check below that confirms the required spectral separation on the realized trajectories. The clip already computes the full eigendecomposition of $\widehat{\mathcal{H}}_t$ at every outer iteration, so at no extra cost we log $m_t := \min_i \big||\lambda_i(\widehat{\mathcal{H}}_t)| - \mu_c\big|$ on every run and test $\min\{m_t, \mu_c\} > 2\rho_{\mathrm{op}}$ at every iteration: since m_t is measured on $\widehat{\mathcal{H}}_t$, Weyl's inequality certifies only a margin $\rho := \min\{m_t - \rho_{\mathrm{op}}, \mu_c\}$ for the

exact Hessian, and Assumption 27 needs $\rho > \rho_{\mathrm{op}}$. Here ρ_{op} is our measured estimate of the operator error (finite-difference error, or the floating-point floor of the eigensolver); it does not certify every upstream error in assembling the Hessian. For the regression cells with a Euclidean lower level (S2, S3, UCI) the Hessian does not depend on y, so this check also covers $y^\star(x_t)$. For S1 the logged operator is the Euclidean coordinate Hessian, not the affine-invariant Riemannian Hessian (Appendix D.1), so its row is a coordinate-operator diagnostic and does not certify Assumption 27. For the meta-learning cells the check is made at the inner-solver output y_t. At $\lambda = 10^{-2}$ every Hessian eigenvalue is at least λ everywhere, so $\min_t m_t = 9.0 \cdot 10^{-3}$, $\rho = \mu_c$, $c_\rho = 1.0$, the margin also holds at $y^\star(x_t)$, and $\mu_c \le \mu$ holds ($\mu \ge \lambda > \mu_c$); this is the only meta-learning cell that satisfies the theorem's lower-level hypotheses. The margin at $y^\star(x_t)$ is assumed at $\lambda = 10^{-6}$, while the $\lambda = 0$ stress test has no finite minimizer and is excluded from that statement. All three meta-learning cells are tabulated. Table 5 reports the worst run per cell. The test passes at every iteration of every tabulated run; the margin is smallest in the singular cells, where $\min_t m_t$ falls to 10^{-9}–10^{-7} and c_ρ is correspondingly large, so the constants of Theorem 35 are large there. In the mini-batch sweep of Appendix E, the test passes only for the full batch $B = 40$; for $B \le 16$ the mini-batch operator error exceeds the margin. The test certifies spectral separation only; it does not verify $\mu_c \le \mu$, which fails in the $\lambda = 10^{-6}$ cell ($\mu \le \lambda = 10^{-6} < \mu_c$), and the $\lambda = 0$ cell lies outside Assumption 2 altogether (no finite minimizer), so Table 5 does not establish all theorem assumptions.

Table 5: Pointwise spectral margin along the realized trajectories (worst run per cell). $\min_t m_t$ is the smallest distance of an eigenvalue magnitude of $\widehat{\mathcal{H}}_t$ from μ_c over all outer iterations; ρ_{op} is the measured operator error of line 7 (finite-difference error for the regression cells, floating-point floor of the eigensolver for the closed-form meta-learning Hessian); a run passes if $\min\{m_t, \mu_c\} > 2\rho_{\mathrm{op}}$ at every iteration; c_ρ is (18) with $\rho = \min\{\min_t m_t - \rho_{\mathrm{op}}, \mu_c\}$. The S1 row is a coordinate-operator diagnostic (see text). All three MiniImageNet rows ($\lambda = 10^{-2}$, 10^{-6} and 0) come from separate float64 CPU runs of R-HJFBiO (same code, seed and protocol; lines 7–8 in double precision). All other MiniImageNet results (Figures 1, 2 and 9, Tables 3 and 4, and the robustness comparison of Appendix D.3) come from the float32 GPU run of the head-to-head comparison.

Cell	μ_c	ρ_{op}	$\min_t m_t$	runs passing	c_ρ
MiniImageNet, $\lambda = 10^{-2}$	10^{-3}	$2.8 \cdot 10^{-10}$	$9.0 \cdot 10^{-3}$	$1/1$	1.0
MiniImageNet, $\lambda = 10^{-6}$	10^{-3}	$2.8 \cdot 10^{-10}$	$2.9 \cdot 10^{-7}$	$1/1$	$3.4 \cdot 10^{3}$
MiniImageNet, $\lambda = 0$	10^{-3}	$2.8 \cdot 10^{-10}$	$5.4 \cdot 10^{-7}$	$1/1$	$1.8 \cdot 10^{3}$
S1: BCI IV-2a, $\mathrm{St}(22, 4)$	10^{-3}	$1.8 \cdot 10^{-10}$	$1.6 \cdot 10^{-9}$	$27/27$	$8.1 \cdot 10^{5}$
S2: BCI IV-2a, $\mathrm{Gr}(22, 8)$	10^{-3}	$1.5 \cdot 10^{-13}$	$2.8 \cdot 10^{-7}$	$27/27$	$3.5 \cdot 10^{3}$
S3: Poincaré, B^4	10^{-3}	$1.5 \cdot 10^{-13}$	$5.6 \cdot 10^{-7}$	$15/15$	$1.8 \cdot 10^{3}$
UCI, $\mathrm{St}(81, 70)$	10^{-2}	$9.3 \cdot 10^{-14}$	$7.6 \cdot 10^{-4}$	$3/3$	13
UCI, $\mathrm{Gr}(81, 35)$	10^{-2}	$4.9 \cdot 10^{-14}$	$1.1 \cdot 10^{-3}$	$3/3$	9.0
UCI, B^8	10^{-2}	$7.2 \cdot 10^{-14}$	$7.0 \cdot 10^{-4}$	$3/3$	14

Hypergradient strategies and hyperparameters. All five methods compute a hypergradient of the form $\widehat{\nabla} F = \nabla_\Theta f - \nabla_\Theta \langle \nabla_w g, v_m \rangle$, differing only in v_m. HINV solves $(H + \rho I)\, v = \nabla_w f$ directly with $\rho = 10^{-8}$ as a numerical floor (too small to regularize the singular subspace, by design); CG runs $T_{\mathrm{cg}} = 20$ conjugate-gradient steps on the same system with no regularization; NS forms $\alpha \sum_{k=0}^{T_{\mathrm{ns}}-1} (I - \alpha H)^k \nabla_w f$ with $\alpha = 5 \cdot 10^{-2}$ and $T_{\mathrm{ns}} = 5$; AD differentiates through $T_{\mathrm{ad}} = 5$ inner Euclidean GD steps with the same α, keeping `create_graph=True`. R-HJFBiO uses the spectrally-clipped pseudoinverse $\big(\mathcal{S}_{[\mu_c, L_g]}(H)\big)^\dagger \nabla_w f$ from Eq. (5), with $\mu_c = 10^{-3}$ and L_g taken as the empirical maximum eigenvalue of H at the current iterate. We note that setting L_g to the empirical maximum eigenvalue makes the upper clamp in $\mathrm{clip}(\lambda; \mu_c, L_g)$ a no-op for every eigenvalue actually present in the spectrum, and H is positive semidefinite in all our experiments, so the magnitude clip leaves retained eigenvalues unchanged. The implementation therefore realizes the kernel-preserving operator from Eq. (5) in the degenerate form "zero out eigenvalues below μ_c, keep the rest unchanged, then invert," which equals the truncated Moore–Penrose pseudoinverse on the retained subspace $\mathcal{R}_{\mu_c}(H)$, and the computed v lies in this subspace. The upper clamp L_g in Eq. (5) is a worst-case device used in the smoothness step of the proof of Lemma 29; setting L_g adaptively to the empirical maximum eigenvalue does not affect this argument because the proof of

Lemma 29 only uses the eigenvalue floor μ_c on the support of $\mathcal{S}_{[\mu_c, L_g]}$ and a finite upper bound, both of which the empirical L_g satisfies. This is the operator analyzed in Section 3 and is what makes R-HJFBiO well-defined when H is genuinely rank-deficient; the experiments, however, apply its pseudoinverse exactly instead of the one-step update of line 9.

Generality across manifolds. Replacing the Stiefel constraint on W_1, W_2 with Grassmann or Oblique constraints (with the same retraction/projection conventions and the same hyperparameter table) gives the same qualitative pattern as Table 4: HINV aborts at step 0 on every manifold whenever $\lambda < 10^{-2}$, CG/NS destabilize after a varying number of steps, AD survives with a high shared-reference gradient floor and $\mathrm{err}_v \approx 1$, and R-HJFBiO completes the 200-step trajectory on every manifold $\times$ regime combination (with $\mathrm{err}_v = 0$ by construction). This suggests that the failure modes of HINV/CG/NS in the low-λ regimes stem from the linear-system step rather than from a particular retraction or manifold metric.

Practical implications. The meta-learning experiment pinpoints a failure mode that the regression experiments in Section D.4 establish but do not fully foreground: the "tiny ridge" practice common in hypergradient implementations – adding a small constant ρI before solving $Hv = \nabla_w f$ – is not wrong in principle: for a right-hand side in $\mathrm{range}(H)$ and a positive spectral gap, $(H + \rho I)^{-1} b \to H^{\dagger} b$ as $\rho \downarrow 0$. Its practical difficulty is conditioning: the ridge must be chosen relative to the smallest *nonzero* eigenvalue of H, which is unknown a priori, varies across iterates and tasks, and can be arbitrarily small, and any component of b outside $\mathrm{range}(H)$ (from inexact inner solutions or incompatibility) is amplified by $1/\rho$. R-HJFBiO sidesteps this with an *adaptive*, eigenvalue-aware kernel-preserving operator: it never inverts kernel components, regardless of how close to zero the spectrum decays. This is the mechanism that lets the convergence analysis carry through to the nonconvex-PL setting where the strong-convexity-based bounds of [43] do not apply.

G Proofs

Remark 25 (Relation to [54]: analysis versus algorithm)**.** Analysis. *The Euclidean analysis of [54] uses the law of cosines for points, $\|x - z\|^2 = \|x - y\|^2 + \|y - z\|^2 + 2\langle x - y, y - z\rangle$, and trivial parallel transport. On a curved manifold the former holds only up to the factor $\zeta \geq 1$ of* (2) *(the inner-product identity within a single tangent space is unchanged), and composed transports around a nondegenerate triangle disagree by $\mathcal{O}(C_0)$ via* (3)*. Our proofs therefore carry (conservatively) multiplicative ζ factors in the stepsize conditions (most visibly Lemma 32) and additive C_0-terms in the auxiliary-vector drift (Lemma 29). Both vanish when $\mathcal{M}_x, \mathcal{M}_y$ are Euclidean, and the remaining argument then follows that of [54] step by step, except for the steps involving the kernel-preserving clip (Lemma 29 and Assumption 27), which have no counterpart there.* Algorithm. *The two algorithms do not coincide in general. Their operators agree exactly on any Hessian whose spectrum lies in $[\mu, L_g]$ – in particular near an isolated nondegenerate minimizer, the setting of [54], so there the methods coincide once the iterates approach the solution and differ only transiently. They differ whenever eigenvalues of magnitude below the floor are present: [54] raises them to the floor and inverts, whereas we map them to zero and take the pseudoinverse. This difference is empty in the setting of [54]; in ours it vanishes at $\mathcal{M}^{\star}(x)$ for compatible $\mathcal{G}_y f$ but matters away from it, under noise, or without compatibility (Remark 10).*

Remark 26 (Spectral gap near $\mathcal{M}^{\star}(x)$ and finite-difference robustness)**.** *The kernel-preserving construction relies on a spectral gap separating the kernel and signal bands. Let $r_0 := \mu/(4L_{gyy})$, let $d(y, \mathcal{M}^{\star}(x)) \leq r_0$, and let $y^{\star} \in \mathcal{M}^{\star}(x)$ be a nearest point. By Lemma 3(i)–(ii), the spectrum of $\mathrm{Hess}_y\, g(x, y^{\star})$ lies in $\{0\} \cup [\mu, L_g]$; by the transported Hessian-Lipschitz bound of Assumption 7 and Weyl's inequality, every eigenvalue of $\mathrm{Hess}_y\, g(x, y)$ lies in $[-\mu/4, \mu/4] \cup [3\mu/4, L_g + \mu/4]$. With $\mu_c := \mu/2$, Assumption 27 below therefore holds on this neighborhood with $\rho = \mu/4$: there the margin is derived from Lemma 3 and Assumption 7, not assumed. By Weyl's inequality, an operator approximation with error $\rho_{\mathrm{op}} < \rho$ recovers the same kernel/signal partition as the exact Hessian. Away from this neighborhood the margin is a genuine assumption, which we verify on our runs (Appendix F); our implementation sets μ_c without knowledge of μ, and for this practical choice it is the empirical check that carries the load.*

Assumption 27 (Uniform spectral margin)**.** *There exists $\rho > \rho_{\mathrm{op}}$ such that, at every (x, y) the analysis touches, every eigenvalue λ of $\mathrm{Hess}_y\, g(x, y)$ satisfies $|\lambda| \notin (\mu_c - \rho, \mu_c + \rho)$. The points*

touched are the iterates (x_t, y_t) of Algorithms 1 and 2 (at which line 7 builds $\widehat{\mathcal{H}}_t$) and the lower-level solutions $(x_t, y^\star)$, $y^\star \in \mathcal{M}^\star(x_t)$ nearest to y_t; the condition is imposed only at these points, not along the geodesics joining them, so the rank of $\mathcal{S}_{[\mu_c, L_g]}$ may change between consecutive iterates. Here ρ_{op} is the operator-approximation error of line 7 of Algorithm 1: $\rho_{\mathrm{op}} \le L_{gyy}\sqrt{d_y}\,\delta_\varepsilon$ for the basis sweep, controlled through δ_ε, and $\rho_{\mathrm{op}} = 0$ for the closed-form realization.

Assumption 27 forbids only a fixed-width annulus around the threshold: eigenvalues that genuinely fall below μ_c away from $\mathcal{M}^\star(x)$ (e.g., the near-kernel directions of the $\lambda = 10^{-6}$ sweep) are simply truncated, as intended. Since a margin ρ implies every smaller margin, we assume without loss of generality that $\rho \le \mu_c$, and set

$$c_\rho \;:=\; \frac{\mu_c}{\rho - \rho_{\mathrm{op}}} \;\ge\; 1. \tag{18}$$

The margin enters the analysis only through the following two facts; they are needed because $\mathcal{S}_{[\mu_c, L_g]}$ is not Lipschitz in operator norm across the jump of clip at $|\lambda| = \mu_c$.

(M1) Support stability. By Weyl's inequality, if $\mathrm{Hess}_y\, g(x,y)$ satisfies the margin ρ and $\|\widehat{\mathcal{H}} - \mathrm{Hess}_y\, g(x,y)\|_{y\to y} \le \rho_{\mathrm{op}}$, then $\widehat{\mathcal{H}}$ satisfies it with $\rho' := \rho - \rho_{\mathrm{op}}$, and both operators retain the same number of eigenvalues.

(M2) Lipschitz bounds on the common support. For self-adjoint A, B on $T_y\mathcal{M}_y$ whose spectra both avoid $\{|\lambda| \in (\mu_c - \rho', \mu_c + \rho')\}$ with $0 < \rho' \le \mu_c$, let P_A, P_B be the orthogonal projectors onto $\mathrm{range}(\mathcal{S}_{[\mu_c, L_g]}[A])$, $\mathrm{range}(\mathcal{S}_{[\mu_c, L_g]}[B])$. Then

$$\big\|\mathcal{S}_{[\mu_c, L_g]}[A]^\dagger - \mathcal{S}_{[\mu_c, L_g]}[B]^\dagger\big\|_F \le \frac{\|A - B\|_F}{\rho' \mu_c}, \qquad \|P_A - P_B\|_F \le \frac{\|A - B\|_F}{2\rho'}. \tag{19}$$

Proof. On such spectra, $\mathcal{S}_{[\mu_c, L_g]}[\cdot]^\dagger = \varphi_1(\cdot)$ and $P_{(\cdot)} = \varphi_2(\cdot)$ for the even functions $\varphi_1, \varphi_2 : \mathbb{R} \to \mathbb{R}$ that vanish for $|\lambda| \le \mu_c - \rho'$, equal $1/\min\{|\lambda|, L_g\}$ and 1, respectively, for $|\lambda| \ge \mu_c + \rho'$, and are linear in $|\lambda|$ in between. Their Lipschitz constants are at most $\max\{(\mu_c + \rho')^{-2}, (2\rho'\mu_c)^{-1}\} \le (\rho'\mu_c)^{-1}$ (using $\rho' \le \mu_c$ and $L_g \ge \mu_c$) and $(2\rho')^{-1}$. For any L-Lipschitz φ and eigendecompositions $A = \sum_i \alpha_i p_i p_i^*$, $B = \sum_j \beta_j q_j q_j^*$, $\|\varphi(A) - \varphi(B)\|_F^2 = \sum_{i,j} |\varphi(\alpha_i) - \varphi(\beta_j)|^2 |\langle p_i, q_j\rangle|^2 \le L^2 \sum_{i,j} |\alpha_i - \beta_j|^2 |\langle p_i, q_j\rangle|^2 = L^2 \|A - B\|_F^2$. □

In particular, for $A = \mathrm{Hess}_y\, g(x,y)$ and $B = \widehat{\mathcal{H}}$ as in (M1), (19) holds with $(\rho'\mu_c)^{-1} = c_\rho/\mu_c^2$ and $(2\rho')^{-1} = c_\rho/(2\mu_c)$, and for two exact Hessians at points touched by the analysis it holds with $\rho' = \rho$, hence also with these constants.

G.1 Proof of Lemma 28

Lemma 28. *Under Assumptions 2–7, assume in addition that the solution graph $\mathcal{S}^\star := \{(x,y) : x \in \mathcal{U}_x,\ y \in \mathcal{M}^\star(x)\}$ is a C^1 embedded submanifold of $\mathcal{U}_x \times \mathcal{U}_y$ of dimension $\dim \mathcal{M}_x + \dim \mathcal{M}^\star(x)$ (a parametric Morse–Bott condition; the fixed-x result of [89] does not by itself provide it); call this (P1). Assume further (P2) that for every geodesic $c : [0,1] \to \mathcal{U}_x$ and every $y_0 \in \mathcal{M}^\star(c(0))$, the horizontal-lift equation of part (ii) has a solution on $[0,1]$ that remains in $\mathcal{U}_y$; and (P3) that for every $x \in \mathcal{U}_x$ the gradient flow of $g(x,\cdot)$ started in $\mathcal{U}_y$ remains in $\mathcal{U}_y$, so that Assumption 2 yields the quadratic growth $d(y, \mathcal{M}^\star(x))^2 \le \frac{2}{\mu}\big(g(x,y) - G(x)\big)$ for $y \in \mathcal{U}_y$, with $G(x) := g(x, y^\star(x))$. (P2) and (P3) are domain conditions; they can fail on incomplete or restricted domains even when (P1) holds. Then through every $(x_0, y_0) \in \mathcal{S}^\star$ passes a local C^1 selection $x \mapsto y^\star(x) \in \mathcal{M}^\star(x)$, and $F(x) := f(x, y^\star(x))$ is well defined and differentiable on $\mathcal{U}_x$; well-definedness is independent of the selection because $\mathcal{M}^\star(x)$ is connected and $\mathcal{G}_y f(x, y^\star) \perp T_{y^\star}\mathcal{M}^\star(x)$ by Assumption 5. Write H, B, C for $\mathrm{Hess}_y\, g$, $\mathcal{G}^2_{yx} g$, $\mathcal{G}^2_{xy} g$ evaluated at $(x, y^\star)$ and $N(x, y^\star) := -H^\dagger B$. Moreover, for all $x_1, x_2 \in \mathcal{U}_x$:*

(i) Hypergradient formula. *For every C^1 selection,*

$$Dy^\star(x) = N(x, y^\star(x)) + K(x), \qquad \mathrm{range}\, K(x) \subseteq \ker H = T_{y^\star(x)}\mathcal{M}^\star(x),$$

where K depends on the selection, and for every $y^\star \in \mathcal{M}^\star(x)$

$$\mathcal{G}F(x) = \mathcal{G}_x f(x, y^\star) - \mathcal{G}^2_{xy} g(x, y^\star)\Big[\big(\mathrm{Hess}_y\, g(x, y^\star)\big)^\dagger \mathcal{G}_y f(x, y^\star)\Big],$$

where † denotes the Moore–Penrose pseudoinverse.

(ii) Lipschitz lifts. *For every $y_1^\star \in \mathcal{M}^\star(x_1)$ there exists $y_2^\star \in \mathcal{M}^\star(x_2)$ with*

$$d\big(y_1^\star, y_2^\star\big) \le \kappa\, d(x_1, x_2), \qquad \kappa := \frac{C_{gxy}}{\mu}.$$

(iii) Lipschitz continuity of the normal derivative. *For $y_1^\star, y_2^\star$ as in (ii),*

$$\big\|\Gamma_{y_1^\star \to y_2^\star} \circ N(x_1, y_1^\star) \circ \Gamma_{x_2 \to x_1} - N(x_2, y_2^\star)\big\| \le L_y\, d(x_1, x_2),$$

where $L_y := \big(3C_{gxy}L_{gyy}/\mu^2 + L_{gxy}/\mu\big)(1+\kappa)$.

(iv) Lipschitz continuity of the Riemannian hypergradient. *For all $x_1, x_2 \in \mathcal{U}_x$,*

$$\|\Gamma_{x_1 \to x_2}\mathcal{G}F(x_1) - \mathcal{G}F(x_2)\|_{x_2} \le L_F\, d(x_1, x_2),$$

where $L_F = \left(L_f + L_f\kappa + C_{fy}\left(\frac{3C_{gxy}L_{gyy}}{\mu^2} + \frac{L_{gxy}}{\mu}\right)\right)(1+\kappa)$.

A selection $x \mapsto y^\star(x)$ need not itself be Lipschitz with constant κ (its tangential part K is arbitrary); (ii) replaces this by lifts along curves.

Proof. Fix $(x, y^\star) \in \mathcal{S}^\star$ and set

$$H := \mathrm{Hess}_y\, g(x, y^\star), \quad B := \mathcal{G}^2_{yx} g(x, y^\star), \quad C := \mathcal{G}^2_{xy} g(x, y^\star),$$
$$a := \mathcal{G}_x f(x, y^\star), \qquad b := \mathcal{G}_y f(x, y^\star).$$

By Lemma 3, H is symmetric with $\ker H = T_{y^\star}\mathcal{M}^\star(x)$ and $\mathrm{range}(H) = N_{y^\star}\mathcal{M}^\star(x)$, on which it is positive definite with spectrum in $[\mu, L_g]$, so $\|H^\dagger\| \le 1/\mu$, and Assumption 5 gives $b \in \mathrm{range}(H)$, hence $\|H^\dagger b\| \le C_{fy}/\mu$.

Tangent space of $\mathcal{S}^\star$ and local selections. Since $\mathcal{G}_y g \equiv 0$ on $\mathcal{S}^\star$, every tangent vector (v, w) of $\mathcal{S}^\star$ at $(x, y^\star)$ satisfies $Hw + Bv = 0$. The set of such (v, w) has dimension at most $\dim \mathcal{M}_x + \dim\ker H = \dim \mathcal{S}^\star$, with equality only if $Bv \in \mathrm{range}\, H$ for every v. Hence $T_{(x,y^\star)}\mathcal{S}^\star = \{(v, -H^\dagger Bv + k) : v \in T_x\mathcal{M}_x,\ k \in \ker H\}$, the projection $\mathcal{S}^\star \to \mathcal{U}_x$ is a submersion, and local C^1 sections (selections) through every point exist; for each fixed x, the structure of $\mathcal{M}^\star(x)$ is that of [89].

Part (i). For a C^1 selection, $(v, Dy^\star(x)[v])$ is tangent to $\mathcal{S}^\star$, so by the previous paragraph $Dy^\star(x) = -H^\dagger B + K$ with $\mathrm{range}\, K \subseteq \ker H$; the stationarity equation $HDy^\star + B = 0$ determines only the normal component. By the chain rule and $\mathcal{G}^2_{xy} g = (\mathcal{G}^2_{yx} g)^*$, for $v \in T_x\mathcal{M}_x$,

$$DF(x)[v] = \langle a, v\rangle - \langle b, H^\dagger Bv\rangle + \langle b, Kv\rangle = \langle a - C[H^\dagger b], v\rangle,$$

because $b \perp \ker H \supseteq \mathrm{range}\, K$. Thus $\mathcal{G}F(x) = a - C[H^\dagger b]$, which does not depend on K. Since a selection passes through every point of $\mathcal{M}^\star(x)$ and F does not depend on the selection, the formula holds at every $y^\star \in \mathcal{M}^\star(x)$.

Part (ii). Let $c : [0,1] \to \mathcal{U}_x$ be the minimizing geodesic from x_1 to x_2, with $\|c'(s)\| = d(x_1, x_2)$. Let $y(s)$ solve the horizontal-lift equation $y'(s) = N(c(s), y(s))[c'(s)]$, $y(0) = y_1^\star$. Since (c', Nc') is tangent to $\mathcal{S}^\star$, the solution stays in $\mathcal{S}^\star$, i.e., $y(s) \in \mathcal{M}^\star(c(s))$, and by (P2) it exists on $[0,1]$ and remains in $\mathcal{U}_y$. Set $y_2^\star := y(1)$. Then $\|N\| \le \|H^\dagger\|\|B\| \le C_{gxy}/\mu = \kappa$ gives

$$d\big(y_1^\star, y_2^\star\big) \le \int_0^1 \big\|y'(s)\big\|\, ds \le \kappa\, d(x_1, x_2).$$

A pseudoinverse perturbation bound. For self-adjoint operators A, A' on the same inner-product space, Wedin's identity $A'^\dagger - A^\dagger = -A'^\dagger(A'-A)A^\dagger + A'^\dagger A'^\dagger(A'-A)(I - AA^\dagger) + (I - A'^\dagger A')(A'-A)A^\dagger A^\dagger$ holds without any rank condition, so

$$\|A'^\dagger - A^\dagger\| \le 3\max\{\|A^\dagger\|, \|A'^\dagger\|\}^2\, \|A' - A\|. \tag{20}$$

We use (20) rather than the identity $(A'^\dagger - A^\dagger)w = A'^\dagger(A - A')A^\dagger w$ for $w \in \mathrm{range}\, A$, which does not hold in general when $\mathrm{range}\, A' \neq \mathrm{range}\, A$.

Part (iii). Write H_i, B_i at $(x_i, y_i^\star)$ with $y_i^\star$ as in (ii), and define the parallel-transported operators

$$\widetilde{H}_1 := \Gamma_{y_1^\star \to y_2^\star} \circ H_1 \circ \Gamma_{y_2^\star \to y_1^\star}, \qquad \widetilde{B}_1 := \Gamma_{y_1^\star \to y_2^\star} \circ B_1 \circ \Gamma_{x_2 \to x_1}.$$

Parallel transport is an isometry, so $\widetilde{H}_1^\dagger = \Gamma_{y_1^\star \to y_2^\star} \circ H_1^\dagger \circ \Gamma_{y_2^\star \to y_1^\star}$ and $\|\widetilde{H}_1^\dagger\| \le 1/\mu$, and

$$\Gamma_{y_1^\star \to y_2^\star} \circ N(x_1, y_1^\star) \circ \Gamma_{x_2 \to x_1} - N(x_2, y_2^\star) = -\big(\widetilde{H}_1^\dagger \widetilde{B}_1 - H_2^\dagger B_2\big) = -\widetilde{H}_1^\dagger(\widetilde{B}_1 - B_2) - (\widetilde{H}_1^\dagger - H_2^\dagger) B_2.$$

By (20) with $\|\widetilde{H}_1^\dagger\|, \|H_2^\dagger\| \le 1/\mu$, $\|(\widetilde{H}_1^\dagger - H_2^\dagger) B_2\| \le \frac{3}{\mu^2} \|\widetilde{H}_1 - H_2\| \, \|B_2\|$. By Assumption 7 and Part (ii),

$$\|\widetilde{H}_1 - H_2\| \le L_{gyy}(1+\kappa)\, d(x_1, x_2), \quad \|\widetilde{B}_1 - B_2\| \le L_{gxy}(1+\kappa)\, d(x_1, x_2), \quad \|B_2\| \le C_{gxy}.$$

Combining,

$$\Big\| \Gamma_{y_1^\star \to y_2^\star} \circ N(x_1, y_1^\star) \circ \Gamma_{x_2 \to x_1} - N(x_2, y_2^\star) \Big\| \le \left(\frac{3 C_{gxy} L_{gyy}}{\mu^2} + \frac{L_{gxy}}{\mu} \right) (1+\kappa)\, d(x_1, x_2).$$

Part (iv). Fix $y_1^\star \in \mathcal{M}^\star(x_1)$ and let $y_2^\star$ be as in (ii); by Part (i), $\mathcal{G}F(x_i)$ may be evaluated at $y_i^\star$. Set $v_i := H_i^\dagger b_i$, so $\mathcal{G}F(x_i) = a_i - C_i[v_i]$ and $\|v_i\| \le C_{fy}/\mu$, and define the transports

$$\widetilde{a}_1 := \Gamma_{x_1 \to x_2} a_1, \quad \widetilde{b}_1 := \Gamma_{y_1^\star \to y_2^\star} b_1, \quad \widetilde{C}_1 := \Gamma_{x_1 \to x_2} \circ C_1 \circ \Gamma_{y_2^\star \to y_1^\star}, \quad \widetilde{v}_1 := \Gamma_{y_1^\star \to y_2^\star} v_1.$$

By isometry, $\Gamma_{x_1 \to x_2}(C_1[v_1]) = \widetilde{C}_1[\widetilde{v}_1]$, so

$$\big\| \Gamma_{x_1 \to x_2} \mathcal{G}F(x_1) - \mathcal{G}F(x_2) \big\|_{x_2} \le \|\widetilde{a}_1 - a_2\| + \|\widetilde{C}_1[\widetilde{v}_1] - C_2[v_2]\|. \tag{21}$$

Lipschitz regularity of $\mathcal{G}_x f$ and Part (ii) give $\|\widetilde{a}_1 - a_2\| \le L_f(1+\kappa)\, d(x_1, x_2)$. Next, $\widetilde{v}_1 - v_2 = \widetilde{H}_1^\dagger(\widetilde{b}_1 - b_2) + (\widetilde{H}_1^\dagger - H_2^\dagger) b_2$, and by (20),

$$\|(\widetilde{H}_1^\dagger - H_2^\dagger) b_2\| \le \frac{3 C_{fy} L_{gyy}}{\mu^2} (1+\kappa)\, d(x_1, x_2), \quad \|\widetilde{H}_1^\dagger(\widetilde{b}_1 - b_2)\| \le \frac{L_f}{\mu} (1+\kappa)\, d(x_1, x_2).$$

Splitting $\widetilde{C}_1[\widetilde{v}_1] - C_2[v_2] = \widetilde{C}_1[\widetilde{v}_1 - v_2] + (\widetilde{C}_1 - C_2)[v_2]$ and using $\|\widetilde{C}_1\| \le C_{gxy}$, $\|v_2\| \le C_{fy}/\mu$, and $\|\widetilde{C}_1 - C_2\| \le L_{gxy}(1+\kappa)\, d(x_1, x_2)$,

$$\|\widetilde{C}_1[\widetilde{v}_1] - C_2[v_2]\| \le \left[L_f \kappa + C_{fy} \left(\frac{3 C_{gxy} L_{gyy}}{\mu^2} + \frac{L_{gxy}}{\mu} \right) \right] (1+\kappa)\, d(x_1, x_2),$$

where we used $\kappa = C_{gxy}/\mu$. Substituting into (21) gives the claim with the stated L_F. □

G.2 Supporting lemmas

Lemma 29. *Suppose the sequence $\{x_t, y_t, v_t\}_{t=1}^T$ is generated by Algorithm 1, and let*

$$v_t^\star := v^\star(x_t, y_t) = \Big(\mathcal{S}_{[\mu_c, L_g]}\big[\operatorname{Hess}_y g(x_t, y_t) \big] \Big)^\dagger \mathcal{G}_y f(x_t, y_t) \in T_{y_t} \mathcal{M}_y.$$

Under Assumptions 2–7, 1, and 27, if $0 < \tau \le \frac{1}{6 L_g}$ and $r_v \ge C_{fy}/\mu_c$, then

$$\begin{aligned} \|v_{t+1} - v_{t+1}^\star\|_{y_{t+1}}^2 \le \Big(1 - \frac{\mu_c \tau}{4}\Big) \|v_t - v_t^\star\|_{y_t}^2 - \frac{3}{4} \big\| v_{t+1} - P_t v_t \big\|^2 \\ + \frac{c_\rho^2 \, C_{fy}^2 \, \rho_F^2}{2 \mu_c^4} + \frac{50 \, \bar{L}^2}{3 \mu_c \tau} \Big(d(x_{t+1}, x_t)^2 + d(y_{t+1}, y_t)^2 \Big), \end{aligned} \tag{22}$$

where P_t is the projector of line 9 of Algorithm 1 and $\|v_{t+1} - P_t v_t\|$ is measured in $T_{y_t} \mathcal{M}_y$ before the transport, c_ρ is the margin constant (18) *of Assumption 27, and $\rho_F := \sup_t \|\widehat{\mathcal{H}}_t - \operatorname{Hess}_y g(x_t, y_t)\|_F$ satisfies $\rho_F \le L_{gyy} \sqrt{d_y}\, \delta_\varepsilon$ for the basis-sweep construction and $\rho_F = 0$ for the closed-form realization, in which case the ρ_F^2 term vanishes.*

Remark 30 (Role of the transport-distortion constant C_0)**.** *The Euclidean proof of [54, Lemma A.3] uses the identity $v_{t+1}^\star - v_t^\star$ and bounds it by Lipschitz continuity alone, because parallel transport is trivial in $\mathbb{R}^d$. On a manifold, the analogous quantity is $v_{t+1}^\star - \Gamma_{y_t \to y_{t+1}} v_t^\star$; the parallel transport $\Gamma_{y_t \to y_{t+1}}$ along the unique geodesic from y_t to y_{t+1} is itself well-defined only because $\mathcal{U}_y$ is totally normal under Assumption 1(a). If $v_t^\star$ and $v_{t+1}^\star$ were compared through a path of parallel transports through a third point, the composition would disagree with the direct transport by $O(C_0)$ per* (3)*. In the proof below both transports run along the single geodesic $y_t \to y_{t+1}$, whose two-edge loop has zero holonomy, so this term is a conservative over-estimate rather than a necessary correction; we nevertheless retain it in the constant $\bar{L}^2$, finite by Assumption 1(b), so that the recursion* (22) *formally mirrors the Euclidean one from [54], but with a strictly larger $\bar{L}$. On Hadamard manifolds the constant C_0 can be computed explicitly from sectional-curvature bounds (see [43, Section 3]); in Euclidean space $C_0 = 0$ and $\bar{L}$ reduces to the Euclidean value.*

Proof. Line 9 of Algorithm 1 computes the auxiliary update at the base point y_t; in Steps 1–3 we write v_{t+1} for the pre-transport vector $\Pi_{r_v}^{y_t}\big(P_t(v_t - \tau h_t)\big) \in T_{y_t}\mathcal{M}_y$, with $h_t = \mathcal{S}_{[\mu_c, L_g]}[\widehat{\mathcal{H}}_t][v_t] - \mathcal{G}_y f(x_t, y_t)$, $\widehat{\mathcal{H}}_t = \widehat{\mathcal{H}}_R(x_t, y_t; \delta_\varepsilon)$ (self-adjoint by construction), and P_t the orthogonal projector onto $\widehat{\mathcal{R}}_t := \mathrm{range}(\mathcal{S}_{[\mu_c, L_g]}[\widehat{\mathcal{H}}_t])$; the transport to y_{t+1} in line 9 is accounted for in (32) by isometry. All quantities are evaluated at (x_t, y_t, v_t), as in the algorithm. Write $\Lambda_{y_t} := \{v \in T_{y_t}\mathcal{M}_y : \|v\|_{y_t} \le r_v\}$, $\widehat{K}_t := \widehat{\mathcal{R}}_t \cap \Lambda_{y_t}$, and $e_t := \|v_t - v_t^\star\|_{y_t}^2$.

Step 1: the projected step is a projected-gradient step on the range. Define the tangent-space quadratic

$$\widehat{R}_t(v) := \frac{1}{2}\big\langle v, \mathcal{S}_{[\mu_c, L_g]}[\widehat{\mathcal{H}}_t][v]\big\rangle_{y_t} - \big\langle v, P_t \mathcal{G}_y f(x_t, y_t)\big\rangle_{y_t}, \qquad v \in T_{y_t}\mathcal{M}_y.$$

Since $\mathcal{S}_{[\mu_c, L_g]}[\widehat{\mathcal{H}}_t]$ vanishes on $\widehat{\mathcal{R}}_t^\perp$ and has spectrum in $[\mu_c, L_g]$ on $\widehat{\mathcal{R}}_t$, the restriction of $\widehat{R}_t$ to $\widehat{\mathcal{R}}_t$ is μ_c-strongly convex and L_g-smooth; this is the only form of strong convexity used below. The minimizer of $\widehat{R}_t$ over $\widehat{\mathcal{R}}_t$ is $\hat{v}_t^\star := \mathcal{S}_{[\mu_c, L_g]}[\widehat{\mathcal{H}}_t]^\dagger \mathcal{G}_y f(x_t, y_t)$. Let $u_t := P_t v_t$. Because $\mathcal{S}_{[\mu_c, L_g]}[\widehat{\mathcal{H}}_t] v_t = \mathcal{S}_{[\mu_c, L_g]}[\widehat{\mathcal{H}}_t] u_t$,

$$P_t(v_t - \tau h_t) = u_t - \tau\, \mathcal{G}\widehat{R}_t(u_t) \in \widehat{\mathcal{R}}_t, \qquad \mathcal{G}\widehat{R}_t(u) = \mathcal{S}_{[\mu_c, L_g]}[\widehat{\mathcal{H}}_t][u] - P_t \mathcal{G}_y f(x_t, y_t).$$

For $u \in \widehat{\mathcal{R}}_t$ and any z, $\|u - z\|^2 = \|u - P_t z\|^2 + \|(I - P_t)z\|^2$, and the radial projection onto the centred ball Λ_{y_t} maps $\widehat{\mathcal{R}}_t$ into itself; hence $\Pi_{r_v}^{y_t} \circ P_t$ is the projection onto the closed convex set $\widehat{K}_t$, and $v_{t+1} = \mathrm{proj}_{\widehat{K}_t}\big(u_t - \tau \mathcal{G}\widehat{R}_t(u_t)\big)$. For the analysis of Algorithm 2 we allow an inexact gradient $\mathcal{G}\widehat{R}_t(u_t) - \bar{\xi}_t$ with $\bar{\xi}_t \in \widehat{\mathcal{R}}_t$; for Algorithm 1, $\bar{\xi}_t = 0$. The optimality condition of the projection gives

$$\Big\langle \mathcal{G}\widehat{R}_t(u_t) - \bar{\xi}_t + \frac{1}{\tau}(v_{t+1} - u_t),\, v - v_{t+1} \Big\rangle_{y_t} \ge 0, \qquad \forall v \in \widehat{K}_t. \tag{23}$$

Step 2: contraction toward the projected target. Take the comparator $v_c := P_t v_t^\star$. It is feasible: $v_c \in \widehat{\mathcal{R}}_t$ and $\|v_c\| \le \|v_t^\star\| \le C_{fy}/\mu_c \le r_v$. Since $v_c - u_t$ and $v_{t+1} - u_t$ lie in $\widehat{\mathcal{R}}_t$, strong convexity and smoothness of $\widehat{R}_t$ on $\widehat{\mathcal{R}}_t$ give

$$\begin{aligned}
\widehat{R}_t(v_c) &\ge \widehat{R}_t(u_t) + \langle \mathcal{G}\widehat{R}_t(u_t), v_c - u_t\rangle + \tfrac{\mu_c}{2}\|v_c - u_t\|^2, \\
\widehat{R}_t(v_{t+1}) &\le \widehat{R}_t(u_t) + \langle \mathcal{G}\widehat{R}_t(u_t), v_{t+1} - u_t\rangle + \tfrac{L_g}{2}\|v_{t+1} - u_t\|^2.
\end{aligned}$$

Subtracting, and using (23) with $v = v_c$,

$$\widehat{R}_t(v_c) - \widehat{R}_t(v_{t+1}) \ge \frac{1}{\tau}\langle v_{t+1} - u_t,\, v_{t+1} - v_c\rangle + \langle \bar{\xi}_t,\, v_c - v_{t+1}\rangle + \frac{\mu_c}{2}\|v_c - u_t\|^2 - \frac{L_g}{2}\|v_{t+1} - u_t\|^2. \tag{24}$$

For the first inner product we use the polarization identity in $T_{y_t}\mathcal{M}_y$:

$$\langle v_{t+1} - u_t,\, v_{t+1} - v_c\rangle = \tfrac{1}{2}\|v_{t+1} - u_t\|^2 + \tfrac{1}{2}\|v_{t+1} - v_c\|^2 - \tfrac{1}{2}\|u_t - v_c\|^2. \tag{25}$$

For the error term, Young's inequality gives

$$\langle \bar{\xi}_t,\, v_c - v_{t+1}\rangle = \langle \bar{\xi}_t,\, v_c - u_t\rangle + \langle \bar{\xi}_t,\, u_t - v_{t+1}\rangle \geq -\frac{2}{\mu_c}\|\bar{\xi}_t\|^2 - \frac{\mu_c}{4}\|v_c - u_t\|^2 - \frac{\mu_c}{4}\|u_t - v_{t+1}\|^2. \tag{26}$$

For the left-hand side of (24): $v_{t+1} \in \widehat{\mathcal{R}}_t$, so $\widehat{R}_t(v_{t+1}) \geq \widehat{R}_t(\hat{v}_t^\star)$, and since $\widehat{R}_t$ is quadratic with vanishing gradient on $\widehat{\mathcal{R}}_t$ at $\hat{v}_t^\star$,

$$\widehat{R}_t(v_c) - \widehat{R}_t(v_{t+1}) \leq \widehat{R}_t(v_c) - \widehat{R}_t(\hat{v}_t^\star) \leq \frac{L_g}{2}\|v_c - \hat{v}_t^\star\|^2 \leq \frac{L_g}{2}\,\epsilon_t^2, \qquad \epsilon_t := \|v_t^\star - \hat{v}_t^\star\|,$$

where we used $\hat{v}_t^\star = P_t \hat{v}_t^\star$ and that P_t is nonexpansive. Plugging (25)–(26) into (24), multiplying by 2τ, and using $\|u_t - v_c\| = \|P_t(v_t - v_t^\star)\| \leq \|v_t - v_t^\star\|$,

$$\|v_{t+1} - v_c\|^2 \leq \Big(1 - \frac{\mu_c \tau}{2}\Big) e_t - \Big(1 - L_g\tau - \frac{\mu_c\tau}{2}\Big)\|v_{t+1} - u_t\|^2 + \tau L_g\,\epsilon_t^2 + \frac{4\tau}{\mu_c}\|\bar{\xi}_t\|^2. \tag{27}$$

Since $0 < \tau \leq \frac{1}{6L_g}$ and $\mu_c \leq L_g$, we have $1 - L_g\tau - \mu_c\tau/2 \geq \frac{3}{4}$ and $\tau L_g \leq \frac{1}{6}$.

Step 3: return to the exact target. The vector $v_{t+1} - v_c$ lies in $\widehat{\mathcal{R}}_t$, while $v_t^\star - v_c = (I - P_t)v_t^\star$ is orthogonal to it, so $\|v_{t+1} - v_t^\star\|^2 = \|v_{t+1} - v_c\|^2 + \|(I - P_t)v_t^\star\|^2$. Let Q_t be the orthogonal projector onto $\mathrm{range}(\mathcal{S}_{[\mu_c, L_g]}[\mathrm{Hess}_y\, g(x_t, y_t)])$, so that $v_t^\star = Q_t v_t^\star$ and $(I - P_t)v_t^\star = (Q_t - P_t)v_t^\star$. By Assumption 27 and (M1)–(M2) applied to $A = \mathrm{Hess}_y\, g(x_t, y_t)$, $B = \widehat{\mathcal{H}}_t$ (note $\rho_{\mathrm{op}} \leq \rho_F$),

$$\epsilon_t \leq \frac{c_\rho}{\mu_c^2}\,\rho_F\, C_{fy}, \qquad \|(I - P_t)v_t^\star\| \leq \frac{c_\rho}{2\mu_c}\,\rho_F \cdot \frac{C_{fy}}{\mu_c}.$$

Combining with (27) and $\frac{1}{6} + \frac{1}{4} = \frac{5}{12}$,

$$\|v_{t+1} - v_t^\star\|^2 \leq \Big(1 - \frac{\mu_c\tau}{2}\Big) e_t - \frac{3}{4}\|v_{t+1} - P_t v_t\|^2 + \frac{5 c_\rho^2 C_{fy}^2 \rho_F^2}{12\,\mu_c^4} + \frac{4\tau}{\mu_c}\|\bar{\xi}_t\|^2. \tag{28}$$

For the basis-sweep construction, ρ_F is controlled as follows. With the geodesics

$$\gamma_+^{(i)}(s) = \mathrm{Exp}_{y_t}(s\delta_\varepsilon e_i), \qquad \gamma_-^{(i)}(s) = \mathrm{Exp}_{y_t}(-s\delta_\varepsilon e_i), \qquad s \in [0, 1],$$

applied to each basis direction e_i, the Riemannian central-difference probe

$$\widetilde{\mathcal{H}}_R(x_t, y_t, e_i; \delta_\varepsilon) = \frac{\Gamma_{\gamma_+^{(i)}(1) \to y_t}\mathcal{G}_y g(x_t, \gamma_+^{(i)}(1)) - \Gamma_{\gamma_-^{(i)}(1) \to y_t}\mathcal{G}_y g(x_t, \gamma_-^{(i)}(1))}{2\delta_\varepsilon},$$

combined with the fundamental theorem of calculus along geodesics, gives $\widetilde{\mathcal{H}}_R(x_t, y_t, e_i; \delta_\varepsilon) = \bar{\mathcal{H}}_t^{(i)}\, e_i$, where $\bar{\mathcal{H}}_t^{(i)}$ averages $\mathrm{Hess}_y\, g(x_t, \cdot)$ along $\gamma_\pm^{(i)}$. The Lipschitz-Hessian property of Assumption 7 along each probe (geodesic length δ_ε) gives the column-wise bound

$$\|(\widehat{\mathcal{H}}_t^\circ - \mathrm{Hess}_y\, g(x_t, y_t))e_i\|_{y_t} \leq L_{gyy}\delta_\varepsilon$$

for the unsymmetrized operator, whose Frobenius error is therefore at most $L_{gyy}\delta_\varepsilon\sqrt{d_y}$; since $\mathrm{Hess}_y\, g(x_t, y_t)$ is self-adjoint, $\|\frac{1}{2}(E + E^*)\|_F \leq \|E\|_F$ for the error E, so the symmetrized $\widehat{\mathcal{H}}_t$ satisfies $\rho_F^2 \leq L_{gyy}^2\delta_\varepsilon^2 d_y$; for the closed-form realization, $\rho_F = 0$.

Step 4: drift of $v_t^\star$. Write $H_1 := \Gamma_{y_t \to y_{t+1}} \circ \mathrm{Hess}_y\, g(x_t, y_t) \circ \Gamma_{y_{t+1} \to y_t}$, $H_2 := \mathrm{Hess}_y\, g(x_{t+1}, y_{t+1})$, $b_1 := \Gamma_{y_t \to y_{t+1}}\mathcal{G}_y f(x_t, y_t)$, and $b_2 := \mathcal{G}_y f(x_{t+1}, y_{t+1})$. Parallel transport is an isometry, so $\Gamma_{y_t \to y_{t+1}} v_t^\star = \mathcal{S}_{[\mu_c, L_g]}[H_1]^\dagger b_1$, and H_1 has the same spectrum as $\mathrm{Hess}_y\, g(x_t, y_t)$, hence satisfies Assumption 27. By Assumptions 6–7, $\|b_1 - b_2\| \leq L_f\,(s_x + s_y)$ and $\|H_1 - H_2\|_F \leq \sqrt{d_y}\,\|H_1 - H_2\|_{y_{t+1} \to y_{t+1}} \leq \sqrt{d_y}\,L_{gyy}(s_x + s_y)$, where $s_x := d(x_{t+1}, x_t)$ and $s_y := d(y_{t+1}, y_t)$. Splitting $\mathcal{S}[H_1]^\dagger b_1 - \mathcal{S}[H_2]^\dagger b_2 = \mathcal{S}[H_1]^\dagger(b_1 - b_2) + (\mathcal{S}[H_1]^\dagger - \mathcal{S}[H_2]^\dagger)b_2$ and using $\|\mathcal{S}[\cdot]^\dagger\| \leq 1/\mu_c$, (M2), and $\|b_2\| \leq C_{fy}$,

$$\|v_{t+1}^\star - \Gamma_{y_t \to y_{t+1}} v_t^\star\|_{y_{t+1}} \leq \Big(\frac{L_f}{\mu_c} + \frac{c_\rho\sqrt{d_y}\,L_{gyy}C_{fy}}{\mu_c^2}\Big)(s_x + s_y).$$

Squaring with $(a+b)^2 \le 2a^2+2b^2$ and $(s_x+s_y)^2 \le 2(s_x^2+s_y^2)$, and retaining the transport-distortion correction governed by C_0 (cf. (3)), which only enlarges the bound, yields

$$\|v_{t+1}^\star - \Gamma_{y_t \to y_{t+1}} v_t^\star\|_{y_{t+1}}^2 \le \left(\frac{4L_f^2}{\mu_c^2} + \frac{4c_\rho^2\, d_y\, L_{gyy}^2 C_{fy}^2}{\mu_c^4} + C_0^2\, \frac{C_{fy}^2}{\mu_c^2}\, d(y_t, y_{t+1})^2 \right) \times \Big(d(x_{t+1}, x_t)^2 + d(y_{t+1}, y_t)^2 \Big). \tag{29}$$

The first two terms are the direct Lipschitz drift of $v^\star(\cdot,\cdot)$, which requires Assumption 27 because clip jumps at $|\lambda| = \mu_c$; the third is a conservative transport-distortion term: reverse parallel transport along the same geodesic $y_t \to y_{t+1}$ is the exact inverse, so it could be dropped, and retaining it only enlarges the bound (Remark 30). Bounding $d(y_t, y_{t+1}) \le D_y$ via Assumption 1(b) removes the residual factor and collects the three contributions into a single drift constant

$$L_v^2 \;:=\; \frac{4L_f^2}{\mu_c^2} + \frac{4c_\rho^2\, d_y\, L_{gyy}^2 C_{fy}^2}{\mu_c^4} + \frac{C_0^2 C_{fy}^2 D_y^2}{\mu_c^2}, \tag{30}$$

so that (29) reads $\|v_{t+1}^\star - \Gamma_{y_t \to y_{t+1}} v_t^\star\|_{y_{t+1}}^2 \le L_v^2 \left(d(x_{t+1}, x_t)^2 + d(y_{t+1}, y_t)^2 \right)$. This is the same constant later named L_v^2 in Lemma 31. In the Euclidean case $C_0 = 0$ and the third term of (30) vanishes.

For the recursion below, the $\alpha = \mu_c\tau/4$ Young's split in (32) multiplies L_v^2 by $1 + 4/(\mu_c\tau)$, and in combining with the contraction factor we will use the bookkeeping shorthand

$$\bar{L}^2 \;:=\; \frac{L_v^2}{4} \;=\; \frac{L_f^2}{\mu_c^2} + \frac{c_\rho^2\, d_y\, L_{gyy}^2 C_{fy}^2}{\mu_c^4} + \frac{C_0^2 C_{fy}^2 D_y^2}{4\mu_c^2}. \tag{31}$$

Since $\mu_c\tau \le \mu\tau \le L_g\tau \le \frac{1}{6}$, we have $1 + \frac{4}{\mu_c\tau} \le \frac{1/6+4}{\mu_c\tau} = \frac{25}{6\mu_c\tau}$, so the prefactor on $\bar{L}^2$ in (22) is $\frac{25}{6\mu_c\tau} \cdot 4 = \frac{50}{3\mu_c\tau}$. The factor 4 relating L_v^2 and $\bar{L}^2$ has no analytic content; it is purely a normalization that aligns the constants displayed in Lemmas 29 and 31 with the constants displayed in Theorem 35, where $\bar{L}^2$ enters the stepsize caps $\gamma \le 3\mu_c\tau/(400\bar{L}^2)$ and $\lambda \le 3\mu_c\tau/(200\bar{L}^2)$ in (46).

Step 5: combination. Decompose

$$\begin{aligned}
\|v_{t+1} - v_{t+1}^\star\|_{y_{t+1}}^2 &= \|v_{t+1} - \Gamma_{y_t \to y_{t+1}} v_t^\star + \Gamma_{y_t \to y_{t+1}} v_t^\star - v_{t+1}^\star\|_{y_{t+1}}^2 \\
&\le \Big(1 + \frac{\mu_c\tau}{4}\Big) \|v_{t+1} - \Gamma_{y_t \to y_{t+1}} v_t^\star\|_{y_{t+1}}^2 \\
&\quad + \Big(1 + \frac{4}{\mu_c\tau}\Big) \|\Gamma_{y_t \to y_{t+1}} v_t^\star - v_{t+1}^\star\|_{y_{t+1}}^2 \\
&= \Big(1 + \frac{\mu_c\tau}{4}\Big) \|\tilde{v}_{t+1} - v_t^\star\|_{y_t}^2 + \Big(1 + \frac{4}{\mu_c\tau}\Big) \|v_{t+1}^\star - \Gamma_{y_t \to y_{t+1}} v_t^\star\|_{y_{t+1}}^2,
\end{aligned} \tag{32}$$

where $\tilde{v}_{t+1} \in T_{y_t}\mathcal{M}_y$ is the pre-transport vector of line 9 (denoted v_{t+1} in Steps 1–3) and we used the isometry of parallel transport in the last step.

Combining (28) (with $\bar{\xi}_t = 0$), (29), and (32), and using

$$\Big(1 + \frac{\mu_c\tau}{4}\Big)\Big(1 - \frac{\mu_c\tau}{2}\Big) \le 1 - \frac{\mu_c\tau}{4}, \qquad -\Big(1 + \frac{\mu_c\tau}{4}\Big)\frac{3}{4} \le -\frac{3}{4},$$
$$\Big(1 + \frac{\mu_c\tau}{4}\Big)\frac{5}{12} \le \frac{25}{24} \cdot \frac{5}{12} \le \frac{1}{2}, \qquad 1 + \frac{4}{\mu_c\tau} \le \frac{25}{6\mu_c\tau},$$

and $L_v^2 = 4\bar{L}^2$, we arrive at (22). □

Lemma 31. *Assume the sequence $\{x_t, y_t, v_t\}_{t=1}^T$ is generated by Algorithm 1, and let $0 < \gamma \le \frac{1}{2L_F}$. Define*

$$\Phi(x) := F(x) + \phi(x), \qquad G(x) := g(x, y^\star(x)),$$

and the geodesic proximal update

$$x_{t+1} \;=\; P_{\phi(\cdot)}^\gamma(x_t, w_t) \;=\; \mathrm{Exp}_{x_t}\big(u_t^\star\big),$$

$$u_t^\star \;:=\; \operatorname*{arg\,min}_{u\in \mathrm{Exp}_{x_t}^{-1}(\mathcal{U}_x)} \Big\{ \langle w_t, u\rangle_{x_t} + \tfrac{1}{2\gamma}\|u\|_{x_t}^2 + \phi\big(\mathrm{Exp}_{x_t}(u)\big)\Big\}.$$

Define the Riemannian gradient mapping

$$\mathcal{G}(x_t, w_t, \gamma) \;:=\; -\frac{1}{\gamma}\,\mathrm{Exp}_{x_t}^{-1}(x_{t+1}) \;=\; -\frac{1}{\gamma}u_t^\star \;\in\; T_{x_t}\mathcal{M}_x.$$

Let $v_t^\star := v^\star(x_t, y_t) \in T_{y_t}\mathcal{M}_y$, let $y^\star(x_t)$ denote a point of $\mathcal{M}^\star(x_t)$ nearest to y_t (assumed to lie in $\mathcal{U}_y$), let $v_t^{\star,\infty} := v^\star(x_t, y^\star(x_t)) \in T_{y^\star(x_t)}\mathcal{M}_y$, and set

$$L_v^2 \text{ as in (30), equivalently } L_v^2 = 4\bar{L}^2, \qquad A \;:=\; L_f^2 + r_v^2 L_{gxy}^2 + 2C_{gxy}^2 L_v^2.$$

Then

$$\Phi(x_{t+1}) \;\le\; \Phi(x_t) - \frac{\gamma}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|_{x_t}^2 + \frac{12\gamma A}{\mu}\big(g(x_t, y_t) - G(x_t)\big) + 12\gamma C_{gxy}^2\|v_t - v_t^\star\|_{y_t}^2 + 2\gamma L_{gxy}^2\delta_\varepsilon^2 r_v^4.$$

Proof. Throughout we work in the tangent space $T_{x_t}\mathcal{M}_x$ and write $u_t^\star = \mathrm{Exp}_{x_t}^{-1}(x_{t+1})$, so $u_t^\star = -\gamma\mathcal{G}(x_t, w_t, \gamma)$, $\|u_t^\star\|_{x_t}^2 = \gamma^2\|\mathcal{G}(x_t, w_t, \gamma)\|_{x_t}^2$, and $d^2(x_t, x_{t+1}) = \|u_t^\star\|_{x_t}^2$. Set

$$\Psi(u) \;:=\; \langle w_t, u\rangle_{x_t} + \tfrac{1}{2\gamma}\|u\|_{x_t}^2 + \phi\big(\mathrm{Exp}_{x_t}(u)\big), \qquad u \in T_{x_t}\mathcal{M}_x.$$

The subproblem $u_t^\star = \arg\min_u \Psi(u)$ is well posed: by Assumption 1(a), $\mathcal{U}_x$ is totally normal so $\mathrm{Exp}_{x_t}^{-1}$ is single-valued on $\mathcal{U}_x$, and existence of a minimizer on $\mathrm{Exp}_{x_t}^{-1}(\mathcal{U}_x)$ is the attainment assumption stated after Assumption 8 (lower semicontinuity and coercivity alone do not give it on an open restricted domain).

Step 1: Descent inequality from the proximal step. The radial curve $t \in [0,1] \mapsto \mathrm{Exp}_{x_t}(t\,u_t^\star)$ is the unique minimizing geodesic from x_t to x_{t+1} (it lies inside $\mathcal{U}_x$ by Assumption 1(a)). Geodesic convexity of ϕ on $\mathcal{U}_x$ (Assumption 8) implies $t \mapsto \phi\big(\mathrm{Exp}_{x_t}(t\,u_t^\star)\big)$ is convex on $[0,1]$. Combined with the $\frac{1}{\gamma}$-strong convexity of $u \mapsto \langle w_t, u\rangle_{x_t} + \frac{1}{2\gamma}\|u\|_{x_t}^2$ on the Euclidean vector space $T_{x_t}\mathcal{M}_x$, the scalar function $\Theta(t) := \Psi(t\,u_t^\star)$ is $\frac{1}{\gamma}$-strongly convex on $[0,1]$. Since $u_t^\star$ minimizes Ψ, we have $\Theta(1) \le \Theta(0)$ and, by strong convexity,

$$\Theta(0) - \Theta(1) \;\ge\; \frac{1}{2\gamma}\big\|u_t^\star - 0\big\|_{x_t}^2 \;=\; \frac{\gamma}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|_{x_t}^2. \tag{33}$$

Expanding (33) using $\Theta(0) = \phi(x_t)$ and $\Theta(1) = \langle w_t, u_t^\star\rangle_{x_t} + \frac{1}{2\gamma}\|u_t^\star\|_{x_t}^2 + \phi(x_{t+1})$,

$$\phi(x_t) - \phi(x_{t+1}) - \langle w_t, u_t^\star\rangle_{x_t} - \frac{\gamma}{2}\|\mathcal{G}\|_{x_t}^2 \;\ge\; \frac{\gamma}{2}\|\mathcal{G}\|_{x_t}^2,$$

i.e.

$$\langle w_t, u_t^\star\rangle_{x_t} \;\le\; \phi(x_t) - \phi(x_{t+1}) - \gamma\|\mathcal{G}(x_t, w_t, \gamma)\|_{x_t}^2. \tag{34}$$

This is the upper bound on $\langle w_t, u_t^\star\rangle_{x_t}$ used below.

Step 2: Riemannian descent on F. By Lemma 28(iv), F is geodesically L_F-smooth, so

$$F(x_{t+1}) \;\le\; F(x_t) + \langle \mathcal{G}F(x_t), u_t^\star\rangle_{x_t} + \frac{L_F}{2}\|u_t^\star\|_{x_t}^2. \tag{35}$$

Add and subtract $\langle w_t, u_t^\star\rangle_{x_t}$:

$$F(x_{t+1}) \;\le\; F(x_t) + \langle w_t, u_t^\star\rangle_{x_t} + \langle \mathcal{G}F(x_t) - w_t, u_t^\star\rangle_{x_t} + \frac{L_F}{2}\|u_t^\star\|_{x_t}^2. \tag{36}$$

Step 3: Combining (34) **and** (36)**.** Substituting the upper bound (34) into (36), and using $u_t^\star = -\gamma\mathcal{G}$ together with Young's inequality $\langle a, b\rangle \le \|a\|^2 + \frac{1}{4}\|b\|^2$:

$$\langle \mathcal{G}F(x_t) - w_t, u_t^\star\rangle_{x_t} = -\gamma\langle \mathcal{G}F(x_t) - w_t, \mathcal{G}\rangle_{x_t} \le \gamma\|\mathcal{G}F(x_t) - w_t\|_{x_t}^2 + \frac{\gamma}{4}\|\mathcal{G}\|_{x_t}^2,$$

together with $\frac{L_F}{2}\|u_t^\star\|_{x_t}^2 = \frac{\gamma^2 L_F}{2}\|\mathcal{G}\|_{x_t}^2 \le \frac{\gamma}{4}\|\mathcal{G}\|_{x_t}^2$ since $\gamma \le \frac{1}{2L_F}$, we obtain

$$\Phi(x_{t+1}) \;\le\; \Phi(x_t) - \frac{\gamma}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|_{x_t}^2 + \gamma\|\mathcal{G}F(x_t) - w_t\|_{x_t}^2. \tag{37}$$

Step 4: Bounding $\|\mathcal{G}F(x_t) - w_t\|_{x_t}^2$. Recall

$$w_t = \mathcal{G}_x f(x_t, y_t) - \widetilde{J}_R(x_t, y_t, v_t, \delta_\varepsilon), \qquad \widehat{\mathcal{G}}_x F(x_t, y_t, v_t) := \mathcal{G}_x f(x_t, y_t) - \big(\mathcal{G}_{xy}^2 g(x_t, y_t)\big)[v_t],$$

and decompose

$$\|w_t - \mathcal{G}F(x_t)\|_{x_t}^2 \;\le\; 2\big\|\widehat{\mathcal{G}}_x F(x_t, y_t, v_t) - \mathcal{G}F(x_t)\big\|_{x_t}^2 + 2\big\|w_t - \widehat{\mathcal{G}}_x F(x_t, y_t, v_t)\big\|_{x_t}^2. \tag{38}$$

Exact-surrogate term. Write $y^\star := y^\star(x_t)$ for the nearest point above, so that $d(y_t, y^\star) = d(y_t, \mathcal{M}^\star(x_t))$, and let $\tilde{v}_t^{\star,\infty} := \Gamma_{y^\star \to y_t} v_t^{\star,\infty} \in T_{y_t}\mathcal{M}_y$ be the parallel transport of $v_t^{\star,\infty}$ from $T_{y^\star}\mathcal{M}_y$ to $T_{y_t}\mathcal{M}_y$. Since parallel transport is an isometry, $\|\tilde{v}_t^{\star,\infty}\|_{y_t} = \|v_t^{\star,\infty}\|_{y^\star} \le r_v$. By Lemma 28(i), which holds at every point of $\mathcal{M}^\star(x_t)$, and Proposition 11 (using $\mu_c \le \mu$), $\mathcal{G}F(x_t) = \mathcal{G}_x f(x_t, y^\star) - \big(\mathcal{G}_{xy}^2 g(x_t, y^\star)\big)[v_t^{\star,\infty}]$, hence

$$\begin{aligned}\widehat{\mathcal{G}}_x F(x_t, y_t, v_t) - \mathcal{G}F(x_t) = \big[\mathcal{G}_x f(x_t, y_t) - \mathcal{G}_x f(x_t, y^\star)\big] \\ - \Big[\big(\mathcal{G}_{xy}^2 g(x_t, y_t)\big)[v_t] - \big(\mathcal{G}_{xy}^2 g(x_t, y^\star)\big)[v_t^{\star,\infty}]\Big].\end{aligned} \tag{39}$$

Insert $\tilde{v}_t^{\star,\infty}$ to split the bracketed difference:

$$\begin{aligned}\big(\mathcal{G}_{xy}^2 g(x_t, y_t)\big)[v_t] - \big(\mathcal{G}_{xy}^2 g(x_t, y^\star)\big)[v_t^{\star,\infty}] = \underbrace{\big(\mathcal{G}_{xy}^2 g(x_t, y_t)\big)\big[v_t - \tilde{v}_t^{\star,\infty}\big]}_{=:T_a} \\ + \underbrace{\Big[\big(\mathcal{G}_{xy}^2 g(x_t, y_t)\big)[\tilde{v}_t^{\star,\infty}] - \big(\mathcal{G}_{xy}^2 g(x_t, y^\star)\big)[v_t^{\star,\infty}]\Big]}_{=:T_b}.\end{aligned}$$

By the operator-norm bound on $\mathcal{G}_{xy}^2 g$ (Assumption 6), $\|T_a\|_{x_t} \le C_{gxy}\,\|v_t - \tilde{v}_t^{\star,\infty}\|_{y_t}$. By the cross-tangent-space Lipschitz property of $\mathcal{G}_{xy}^2 g$ (Assumption 7, applied at fixed $x = x_t$ to the operator at y_t vs. at $y^\star$ pre-composed with the relevant transports), together with $\|v_t^{\star,\infty}\|_{y^\star} \le r_v$, $\|T_b\|_{x_t} \le L_{gxy}\, r_v\, d(y_t, y^\star)$. Combining, and using $\|T_1\|_{x_t} \le L_f\, d(y_t, y^\star)$ with T_1 the first bracket of (39), the inequality $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ gives

$$\|\widehat{\mathcal{G}}_x F(x_t, y_t, v_t) - \mathcal{G}F(x_t)\|_{x_t}^2 \;\le\; 3L_f^2\, d^2(y_t, y^\star) + 3L_{gxy}^2 r_v^2\, d^2(y_t, y^\star) + 3C_{gxy}^2\,\|v_t - \tilde{v}_t^{\star,\infty}\|_{y_t}^2. \tag{40}$$

Drift between $v_t^\star$ and $\tilde{v}_t^{\star,\infty}$. Using $\|v_t - \tilde{v}_t^{\star,\infty}\|_{y_t}^2 \le 2\|v_t - v_t^\star\|_{y_t}^2 + 2\|v_t^\star - \tilde{v}_t^{\star,\infty}\|_{y_t}^2$ and the Lipschitz drift of $v^\star(x_t, \cdot)$ on $\mathcal{U}_y$ (the same argument that yields (29) in the proof of Lemma 29, specialized to fixed $x = x_t$, with Assumption 27 applied at y_t and $y^\star$ and $d(y_t, y^\star) \le D_y$ used to absorb the residual quadratic factor),

$$\|v_t^\star - \tilde{v}_t^{\star,\infty}\|_{y_t}^2 \;\le\; \Big(\frac{4L_f^2}{\mu_c^2} + \frac{4c_\rho^2\, d_y\, L_{gyy}^2 C_{fy}^2}{\mu_c^4} + \frac{C_0^2 C_{fy}^2 D_y^2}{\mu_c^2}\Big) d^2(y_t, y^\star) \;=\; L_v^2\, d^2(y_t, y^\star). \tag{41}$$

Substituting (41) into (40) and collecting,

$$\begin{aligned}\|\widehat{\mathcal{G}}_x F(x_t, y_t, v_t) - \mathcal{G}F(x_t)\|_{x_t}^2 &\le 3\big(L_f^2 + L_{gxy}^2 r_v^2 + 2C_{gxy}^2 L_v^2\big) d^2(y_t, y^\star) + 6C_{gxy}^2\,\|v_t - v_t^\star\|_{y_t}^2 \\ &= 3A\, d^2(y_t, y^\star) + 6C_{gxy}^2\,\|v_t - v_t^\star\|_{y_t}^2.\end{aligned} \tag{42}$$

By the Riemannian PL inequality applied to $g(x_t, \cdot)$ and the standard quadratic-growth consequence $d^2(y_t, \mathcal{M}^\star(x_t)) \le \frac{2}{\mu}(g(x_t, y_t) - G(x_t))$ (valid on $\mathcal{U}_y$ by (P3) of Lemma 28), which controls the distance to the solution *set* and hence to the nearest point $y^\star$ (not to an arbitrary selected minimizer),

$$\|\widehat{\mathcal{G}}_x F(x_t, y_t, v_t) - \mathcal{G}F(x_t)\|_{x_t}^2 \;\le\; \frac{6A}{\mu}\big(g(x_t, y_t) - G(x_t)\big) + 6C_{gxy}^2\,\|v_t - v_t^\star\|_{y_t}^2. \tag{43}$$

Finite-difference term. By definition,

$$\widetilde{J}_R(x_t, y_t, v_t, \delta_\varepsilon) = \frac{\mathcal{G}_x g\big(x_t, \mathrm{Exp}_{y_t}(\delta_\varepsilon v_t)\big) - \mathcal{G}_x g\big(x_t, \mathrm{Exp}_{y_t}(-\delta_\varepsilon v_t)\big)}{2\delta_\varepsilon}.$$

Using the integral form of the Taylor remainder along the geodesics $k \mapsto \mathrm{Exp}_{y_t}(\pm k\delta_\varepsilon v_t)$ together with the L_{gxy}-Lipschitz continuity of $\mathcal{G}^2_{xy} g$ (Assumption 7),

$$\big\|w_t - \widehat{\mathcal{G}}_x F(x_t, y_t, v_t)\big\|^2_{x_t} = \big\|\widetilde{J}_R - \big(\mathcal{G}^2_{xy} g(x_t, y_t)\big)[v_t]\big\|^2_{x_t} \le L^2_{gxy}\delta^2_\varepsilon \|v_t\|^4_{y_t} \le L^2_{gxy}\delta^2_\varepsilon r_v^4. \quad (44)$$

Plugging (43) and (44) into (38),

$$\|w_t - \mathcal{G}F(x_t)\|^2_{x_t} \;\le\; \frac{12A}{\mu}\big(g(x_t, y_t) - G(x_t)\big) + 12C^2_{gxy}\|v_t - v_t^\star\|^2_{y_t} + 2L^2_{gxy}\delta^2_\varepsilon r_v^4. \quad (45)$$

Step 5: Conclude. Substituting (45) into (37) yields the claim. □

Lemma 32. *Suppose the sequence $\{x_t, y_t, v_t\}_{t=1}^T$ is generated by Algorithm 1, and let*

$$u_t := \mathcal{G}_y g(x_t, y_t) \in T_{y_t}\mathcal{M}_y, \qquad y_{t+1} = \mathrm{Exp}_{y_t}(-\lambda u_t).$$

Define

$$G(x) := g(x, y^\star(x)), \qquad y^\star(x) \in \arg\min_{y \in \mathcal{M}_y} g(x, y).$$

Under Assumptions 2–7 and 1, if

$$\gamma \le \min\Big\{\frac{\lambda\mu}{16L_G}, \frac{\lambda\mu^2}{16\zeta L_g^2}\Big\}, \qquad 0 < \lambda \le \frac{1}{2\zeta L_g},$$

and if G is L_G-smooth on $\mathcal{U}_x$, then

$$g(x_{t+1}, y_{t+1}) - G(x_{t+1}) \le \Big(1 - \frac{\lambda\mu}{2}\Big)\big(g(x_t, y_t) - G(x_t)\big) + \frac{1}{8\gamma} d_x(x_{t+1}, x_t)^2 \\ - \frac{1}{4\lambda} d_y(y_{t+1}, y_t)^2 + \lambda\big\|\mathcal{G}_y g(x_t, y_t) - u_t\big\|^2_{y_t}.$$

where

$$d_x(x_{t+1}, x_t) := \big\|\mathrm{Exp}^{-1}_{x_t}(x_{t+1})\big\|_{x_t}, \qquad d_y(y_{t+1}, y_t) := \big\|\mathrm{Exp}^{-1}_{y_t}(y_{t+1})\big\|_{y_t}.$$

Remark 33 (Role of ζ in the inner stepsize)**.** *Compared to the Euclidean counterpart [54, Lemma 4.3], which requires $\lambda \le 1/(2L_g)$, the Riemannian bound requires $\lambda \le 1/(2\zeta L_g)$ [105]. The extra ζ arises from applying* (2) *when expanding $g(x_{t+1}, y_{t+1})$ along the geodesic $y_t \to y_{t+1}$: the Euclidean parallelogram identity used in the Huang 2024 argument is replaced by the trigonometric comparison, which contributes the factor ζ. This factor is finite precisely because of the bounded-diameter clause Assumption 1(b), and matches the inner stepsize restriction $\eta_y = \Theta(\mu/L^2)$ in [43, Thm. 1] (their $L = \zeta L_g$ in our notation); in the Euclidean case ($\zeta = 1$) it recovers the Euclidean bound exactly. Under the transported-gradient Lipschitz condition of Assumption 6, the descent inequality along a geodesic in fact holds with coefficient $L_g/2$; the factor $\zeta \ge 1$ is therefore a conservative choice, not a derived necessity.*

Remark 34. *The L_G-smoothness of $G(x) = g(x, y^\star(x))$ follows from Assumptions 2–7 in the same manner as Lemma 28(iv); in particular, since $\mathcal{G}G(x) = \mathcal{G}_x g(x, y)$ for every $y \in \mathcal{M}^\star(x)$, the Lipschitz lifts give the constant $L_g(1+\kappa)$, and we fix the conservative choice $L_G := \max\{L_g(1 + 2\kappa), \zeta L_g\}$, which also satisfies the requirement $L_G \ge \zeta L_g$ used in the proof below. That G is smooth under PL alone (without strong convexity) uses the nondegeneracy of Lemma 3 and the Lipschitz lifts of Lemma 28(ii) (a selection itself need not be Lipschitz).*

Proof. Let

$$\xi_t := \mathrm{Exp}^{-1}_{x_t}(x_{t+1}) \in T_{x_t}\mathcal{M}_x, \qquad \eta_t := \mathrm{Exp}^{-1}_{y_t}(y_{t+1}) = -\lambda u_t \in T_{y_t}\mathcal{M}_y.$$

Both inverse exponentials are well defined and single-valued because $\mathcal{U}_x$ and $\mathcal{U}_y$ are totally normal by Assumption 1(a), which also ensures that the geodesic triangles with vertices in these neighborhoods used below have all sides contained in the neighborhoods, so that the trigonometric bound (2)

applies. Using the (Riemannian) L_g-smoothness of $g(x_{t+1}, \cdot)$ combined with (2) – whose curvature factor ζ is finite by Assumption 1(b) – we have

$$g(x_{t+1}, y_{t+1}) \le g(x_{t+1}, y_t) + \big\langle \mathcal{G}_y g(x_{t+1}, y_t),\, \eta_t \big\rangle_{y_t} + \frac{\zeta L_g}{2}\|\eta_t\|_{y_t}^2.$$

Since $\eta_t = -\lambda u_t$, it follows that

$$\begin{aligned}
\big\langle \mathcal{G}_y g(x_{t+1}, y_t),\, \eta_t \big\rangle_{y_t} &= -\lambda \big\langle \mathcal{G}_y g(x_{t+1}, y_t),\, u_t \big\rangle_{y_t} \\
&= -\frac{\lambda}{2}\Big(\|\mathcal{G}_y g(x_{t+1}, y_t)\|_{y_t}^2 + \|u_t\|_{y_t}^2 - \|\mathcal{G}_y g(x_{t+1}, y_t) - u_t\|_{y_t}^2 \Big) \\
&\le -\frac{\lambda}{2}\|\mathcal{G}_y g(x_{t+1}, y_t)\|_{y_t}^2 - \frac{1}{2\lambda}\|\eta_t\|_{y_t}^2 + \frac{\lambda}{2}\|\mathcal{G}_y g(x_{t+1}, y_t) - u_t\|_{y_t}^2.
\end{aligned}$$

By Lipschitz continuity of the partial Riemannian gradient in the x-variable,

$$\big\|\mathcal{G}_y g(x_{t+1}, y_t) - \mathcal{G}_y g(x_t, y_t)\big\|_{y_t} \le L_g\, d_x(x_{t+1}, x_t),$$

and hence

$$\|\mathcal{G}_y g(x_{t+1}, y_t) - u_t\|_{y_t}^2 \le 2L_g^2\, d_x(x_{t+1}, x_t)^2 + 2\|\mathcal{G}_y g(x_t, y_t) - u_t\|_{y_t}^2.$$

Therefore,

$$\begin{aligned}
\big\langle \mathcal{G}_y g(x_{t+1}, y_t),\, \eta_t \big\rangle_{y_t} \le -\frac{\lambda}{2}\|\mathcal{G}_y g(x_{t+1}, y_t)\|_{y_t}^2 - \frac{1}{2\lambda}\|\eta_t\|_{y_t}^2 \\
+ \lambda L_g^2\, d_x(x_{t+1}, x_t)^2 + \lambda\|\mathcal{G}_y g(x_t, y_t) - u_t\|_{y_t}^2.
\end{aligned}$$

Using the Riemannian PL inequality (Assumption 2) for $g(x_{t+1}, \cdot)$, which applies because $y_t \in \mathcal{U}_y$ by the convexity-radius clause Assumption 1(a),

$$\|\mathcal{G}_y g(x_{t+1}, y_t)\|_{y_t}^2 \ge 2\mu\big(g(x_{t+1}, y_t) - G(x_{t+1})\big),$$

so

$$\begin{aligned}
\big\langle \mathcal{G}_y g(x_{t+1}, y_t),\, \eta_t \big\rangle_{y_t} \le -\lambda\mu\big(g(x_{t+1}, y_t) - G(x_{t+1})\big) - \frac{1}{2\lambda}\|\eta_t\|_{y_t}^2 \\
+ \lambda L_g^2 d_x(x_{t+1}, x_t)^2 + \lambda\|\mathcal{G}_y g(x_t, y_t) - u_t\|_{y_t}^2.
\end{aligned}$$

Substituting this bound into the y-smoothness inequality gives

$$\begin{aligned}
g(x_{t+1}, y_{t+1}) \le g(x_{t+1}, y_t) - \lambda\mu\big(g(x_{t+1}, y_t) - G(x_{t+1})\big) - \frac{1}{2\lambda}\|\eta_t\|_{y_t}^2 \\
+ \lambda L_g^2 d_x(x_{t+1}, x_t)^2 + \lambda\|\mathcal{G}_y g(x_t, y_t) - u_t\|_{y_t}^2 + \frac{\zeta L_g}{2}\|\eta_t\|_{y_t}^2.
\end{aligned}$$

Rearranging,

$$\begin{aligned}
g(x_{t+1}, y_{t+1}) - G(x_{t+1}) \le (1 - \lambda\mu)\big(g(x_{t+1}, y_t) - G(x_{t+1})\big) - \frac{1}{2\lambda}\|\eta_t\|_{y_t}^2 \\
+ \lambda L_g^2 d_x(x_{t+1}, x_t)^2 + \lambda\|\mathcal{G}_y g(x_t, y_t) - u_t\|_{y_t}^2 + \frac{\zeta L_g}{2}\|\eta_t\|_{y_t}^2.
\end{aligned}$$

Next, by the Riemannian L_g-smoothness of $g(\cdot, y_t)$ combined with (2),

$$g(x_{t+1}, y_t) \le g(x_t, y_t) + \big\langle \mathcal{G}_x g(x_t, y_t),\, \xi_t \big\rangle_{x_t} + \frac{\zeta L_g}{2}\|\xi_t\|_{x_t}^2.$$

Thus,

$$\begin{aligned}
g(x_{t+1}, y_t) - g(x_t, y_t) &\le \big\langle \mathcal{G}_x g(x_t, y_t) - \mathcal{G}G(x_t),\, \xi_t \big\rangle_{x_t} + \big\langle \mathcal{G}G(x_t),\, \xi_t \big\rangle_{x_t} + \frac{\zeta L_g}{2}\|\xi_t\|_{x_t}^2 \\
&\le \frac{1}{8\gamma}\|\xi_t\|_{x_t}^2 + 2\gamma\|\mathcal{G}_x g(x_t, y_t) - \mathcal{G}G(x_t)\|_{x_t}^2 \\
&\quad + \big\langle \mathcal{G}G(x_t),\, \xi_t \big\rangle_{x_t} + \frac{\zeta L_g}{2}\|\xi_t\|_{x_t}^2.
\end{aligned}$$

Since $\mathcal{G}_y g$ vanishes on $\mathcal{M}^\star(x)$, the chain rule applied to any local selection gives $\mathcal{G}G(x) = \mathcal{G}_x g(x,y)$ for every $y \in \mathcal{M}^\star(x)$. Taking $y^\star(x_t)$ to be a point of $\mathcal{M}^\star(x_t)$ nearest to y_t and using Lipschitz continuity of $\mathcal{G}_x g$ in the y-variable,

$$\|\mathcal{G}_x g(x_t, y_t) - \mathcal{G}G(x_t)\|_{x_t} \le L_g \, d_y(y_t, y^\star(x_t)),$$

hence

$$g(x_{t+1}, y_t) - g(x_t, y_t) \le \frac{1}{8\gamma}\|\xi_t\|_{x_t}^2 + 2L_g^2\gamma \, d_y(y_t, y^\star(x_t))^2 + \big\langle \mathcal{G}G(x_t),\, \xi_t \big\rangle_{x_t} + \frac{\zeta L_g}{2}\|\xi_t\|_{x_t}^2.$$

Since G is L_G-smooth on $\mathcal{M}_x$ (again invoking (2), so the constant already absorbs a ζ factor),

$$G(x_{t+1}) \ge G(x_t) + \big\langle \mathcal{G}G(x_t),\, \xi_t \big\rangle_{x_t} - \frac{L_G}{2}\|\xi_t\|_{x_t}^2,$$

so

$$\big\langle \mathcal{G}G(x_t),\, \xi_t \big\rangle_{x_t} \le G(x_{t+1}) - G(x_t) + \frac{L_G}{2}\|\xi_t\|_{x_t}^2.$$

Therefore,

$$\begin{aligned} g(x_{t+1}, y_t) - g(x_t, y_t) \le{}& \frac{1}{8\gamma}\|\xi_t\|_{x_t}^2 + 2L_g^2\gamma \, d_y(y_t, y^\star(x_t))^2 \\ &+ G(x_{t+1}) - G(x_t) + \frac{L_G}{2}\|\xi_t\|_{x_t}^2 + \frac{\zeta L_g}{2}\|\xi_t\|_{x_t}^2. \end{aligned}$$

Using the quadratic-growth estimate implied by the Riemannian PL condition under (P3) of Lemma 28, which bounds the distance to $\mathcal{M}^\star(x_t)$ and hence to the nearest point $y^\star(x_t)$,

$$d_y(y_t, y^\star(x_t))^2 \le \frac{2}{\mu}\big(g(x_t, y_t) - G(x_t)\big),$$

and $\zeta L_g \le L_G$ (by the choice of L_G in Remark 34), we get

$$g(x_{t+1}, y_t) - g(x_t, y_t) \le \frac{4L_g^2\gamma}{\mu}\big(g(x_t, y_t) - G(x_t)\big) + G(x_{t+1}) - G(x_t) + \Big(\frac{1}{8\gamma} + L_G\Big)\|\xi_t\|_{x_t}^2.$$

Equivalently,

$$g(x_{t+1}, y_t) - G(x_{t+1}) \le \Big(1 + \frac{4L_g^2\gamma}{\mu}\Big)\big(g(x_t, y_t) - G(x_t)\big) + \Big(\frac{1}{8\gamma} + L_G\Big)\|\xi_t\|_{x_t}^2.$$

Substituting this into the previous bound yields

$$\begin{aligned} g(x_{t+1}, y_{t+1}) - G(x_{t+1}) \le{}& (1 - \lambda\mu)\Big(1 + \frac{4L_g^2\gamma}{\mu}\Big)\big(g(x_t, y_t) - G(x_t)\big) \\ &+ (1 - \lambda\mu)\Big(\frac{1}{8\gamma} + L_G\Big)\|\xi_t\|_{x_t}^2 - \frac{1}{2\lambda}\|\eta_t\|_{y_t}^2 \\ &+ \lambda L_g^2\|\xi_t\|_{x_t}^2 + \lambda\|\mathcal{G}_y g(x_t, y_t) - u_t\|_{y_t}^2 + \frac{\zeta L_g}{2}\|\eta_t\|_{y_t}^2. \end{aligned}$$

Hence

$$\begin{aligned} g(x_{t+1}, y_{t+1}) - G(x_{t+1}) \le{}& (1 - \lambda\mu)\Big(1 + \frac{4L_g^2\gamma}{\mu}\Big)\big(g(x_t, y_t) - G(x_t)\big) \\ &+ \Big(\frac{1}{8\gamma} + L_G - \frac{\lambda\mu}{8\gamma} - L_G\lambda\mu + L_g^2\lambda\Big)\|\xi_t\|_{x_t}^2 \\ &- \frac{1}{2}\Big(\frac{1}{\lambda} - \zeta L_g\Big)\|\eta_t\|_{y_t}^2 + \lambda\|\mathcal{G}_y g(x_t, y_t) - u_t\|_{y_t}^2. \end{aligned}$$

Finally, by the stepsize conditions

$$\gamma \le \min\Big\{\frac{\lambda\mu}{16L_G}, \frac{\lambda\mu^2}{16\zeta L_g^2}\Big\}, \qquad \lambda \le \frac{1}{2\zeta L_g},$$

we have

$$\frac{\lambda\mu}{2} \geq \frac{4L_g^2\gamma}{\mu} \quad (\text{from } \gamma \leq \tfrac{\lambda\mu^2}{16\zeta L_g^2} \leq \tfrac{\lambda\mu^2}{8L_g^2}), \qquad \frac{\lambda\mu}{8\gamma} \geq L_G + L_g^2\lambda, \qquad \frac{1}{2\lambda} \geq \zeta L_g.$$

Thus,

$$(1-\lambda\mu)\Big(1+\frac{4L_g^2\gamma}{\mu}\Big) \leq 1 - \frac{\lambda\mu}{2},$$

the coefficient of $\|\xi_t\|_{x_t}^2$ is at most $\frac{1}{8\gamma}$, and

$$-\frac{1}{2}\Big(\frac{1}{\lambda} - \zeta L_g\Big)\|\eta_t\|_{y_t}^2 \leq -\frac{1}{4\lambda}\|\eta_t\|_{y_t}^2.$$

Therefore,

$$g(x_{t+1},y_{t+1}) - G(x_{t+1}) \leq \Big(1-\frac{\lambda\mu}{2}\Big)\big(g(x_t,y_t) - G(x_t)\big) + \frac{1}{8\gamma}\|\xi_t\|_{x_t}^2 \\ - \frac{1}{4\lambda}\|\eta_t\|_{y_t}^2 + \lambda\|\mathcal{G}_y g(x_t,y_t) - u_t\|_{y_t}^2.$$

Recalling that

$$\|\xi_t\|_{x_t} = d_x(x_{t+1},x_t), \qquad \|\eta_t\|_{y_t} = d_y(y_{t+1},y_t),$$

completes the proof. □

G.3 Proof of Theorem 13

For the proof we restate Theorem 13 with all constants and stepsize ranges made explicit.

Theorem 35 (Full statement of Theorem 13)**.** *Assume the sequence $\{x_t, y_t, v_t\}_{t=1}^T$ is generated by Algorithm 1. Define*

$$G(x) := g(x, y^\star(x)), \qquad \Delta_t := g(x_t,y_t) - G(x_t),$$

and

$$v_t^\star := v^\star(x_t,y_t) = \Big(\mathcal{S}_{[\mu_c,L_g]}\big[\operatorname{Hess}_y g(x_t,y_t)\big]\Big)^\dagger \mathcal{G}_y f(x_t,y_t) \in T_{y_t}\mathcal{M}_y, \qquad e_t := \|v_t - v_t^\star\|_{y_t}^2.$$

Let

$$\Psi_t := \Phi(x_t) + \Delta_t + e_t, \qquad \bar{L}^2 := \frac{L_f^2}{\mu_c^2} + \frac{c_\rho^2\, d_y\, L_{gyy}^2 C_{fy}^2}{\mu_c^4} + \frac{C_0^2 C_{fy}^2 D_y^2}{4\mu_c^2}.$$

Suppose the assumptions of Lemmas 29, 31, and 32 hold (in particular Assumptions 1 and 27 and (P1)–(P3) of Lemma 28). If

$$0 < \gamma \leq \min\left\{\frac{1}{2L_F}, \frac{\lambda\mu}{16L_G}, \frac{\lambda\mu^2}{16\zeta L_g^2}, \frac{3\mu_c\tau}{400\bar{L}^2}, \frac{\mu_c\tau}{60\,C_{gxy}^2}, \frac{\mu^2\lambda}{30\,A}\right\}, \tag{46}$$

$$0 < \lambda \leq \min\left\{\frac{1}{2\zeta L_g}, \frac{3\mu_c\tau}{200\bar{L}^2}\right\}, \qquad 0 < \tau \leq \frac{1}{6L_g},$$

where $A := L_f^2 + r_v^2 L_{gxy}^2 + 2C_{gxy}^2 L_v^2$ and $L_v^2 = 4\bar{L}^2$ (see (30)*) are the constants from Lemma 31, then*

$$\frac{1}{T}\sum_{t=1}^T \|\mathcal{G}(x_t,w_t,\gamma)\|_{x_t}^2 \leq \frac{8\big(\Phi(x_1) + g(x_1,y_1) - G(x_1) + \|v_1 - v_1^\star\|_{y_1}^2 - \Phi_\star\big)}{T\gamma} \\ + 20L_{gxy}^2\delta_\varepsilon^2 r_v^4 + \frac{4c_\rho^2\, C_{fy}^2\, L_{gyy}^2\, d_y\, \delta_\varepsilon^2}{\gamma\,\mu_c^4}, \tag{47}$$

where $\Phi_\star := \inf_{x\in\mathcal{M}_x}\Phi(x)$. In particular, if δ_ε is small enough that $L_{gyy}\sqrt{d_y}\,\delta_\varepsilon < \rho$ and

$$\delta_\varepsilon^2 \leq \frac{1}{T\,\max\{L_{gxy}^2 r_v^4,\ c_\rho^2 C_{fy}^2 L_{gyy}^2 d_y/(\gamma\mu_c^4)\}},$$

then

$$\frac{1}{T}\sum_{t=1}^{T}\|\mathcal{G}(x_t,w_t,\gamma)\|_{x_t}^2 = O\Big(\frac{1}{T}\Big).$$

Hence the method achieves $O(\varepsilon^{-1})$ iteration complexity for obtaining an ε-stationary point in the sense of the Riemannian gradient mapping. If furthermore $\phi \equiv 0$, then the proximal step reduces to $x_{t+1} = \mathrm{Exp}_{x_t}(-\gamma w_t)$ and

$$\mathcal{G}(x_t,w_t,\gamma) = -\frac{1}{\gamma}\,\mathrm{Exp}_{x_t}^{-1}(x_{t+1}) = w_t,$$

so (47) *becomes an $O(\varepsilon^{-1})$ complexity bound for the smooth Riemannian bilevel problem.*

Remark 36 (Constants under alternative operator constructions)**.** *The displayed bound* (47) *uses $\rho_F^2 \le L_{gyy}^2\delta_\varepsilon^2 d_y$, the Frobenius-norm error of the basis-sweep finite-difference construction employed in Algorithm 1 (line 7). Two alternatives are admissible: (i) any other operator construction with Frobenius error ρ_F, for which $L_{gyy}^2 d_y \delta_\varepsilon^2$ is replaced by ρ_F^2 in* (47)*; (ii) the closed-form realization used in our meta-learning experiment (Appendix F), for which $\rho_F = \rho_{\mathrm{op}} = 0$ and the L_{gyy}^2-term vanishes entirely. The $O(\varepsilon^{-1})$ iteration complexity holds in all cases with the appropriate choice of δ_ε.*

Proof. Let

$$\mathcal{G}_t := \mathcal{G}(x_t,w_t,\gamma), \qquad d_x(x_{t+1},x_t)^2 = \big\|\mathrm{Exp}_{x_t}^{-1}(x_{t+1})\big\|_{x_t}^2 = \gamma^2\|\mathcal{G}_t\|_{x_t}^2.$$

Since Algorithm 1 uses the exact lower-level gradient,

$$u_t = \mathcal{G}_y g(x_t,y_t),$$

the error term in Lemma 32 vanishes:

$$\|\mathcal{G}_y g(x_t,y_t) - u_t\|_{y_t}^2 = 0.$$

From Lemma 31,

$$\Phi(x_{t+1}) \le \Phi(x_t) - \frac{\gamma}{2}\|\mathcal{G}_t\|_{x_t}^2 + \frac{12\gamma A}{\mu}\,\Delta_t + 12\gamma\,C_{gxy}^2\,e_t + 2\gamma\,L_{gxy}^2\,\delta_\varepsilon^2\,r_v^4, \tag{48}$$

where $A := L_f^2 + r_v^2 L_{gxy}^2 + 2C_{gxy}^2 L_v^2$ as in Lemma 31.

From Lemma 32,

$$\Delta_{t+1} \le \Big(1 - \frac{\lambda\mu}{2}\Big)\Delta_t + \frac{1}{8\gamma}d_x(x_{t+1},x_t)^2 - \frac{1}{4\lambda}d_y(y_{t+1},y_t)^2. \tag{49}$$

Using $d_x(x_{t+1},x_t)^2 = \gamma^2\|\mathcal{G}_t\|_{x_t}^2$,

$$\Delta_{t+1} \le \Big(1 - \frac{\lambda\mu}{2}\Big)\Delta_t + \frac{\gamma}{8}\|\mathcal{G}_t\|_{x_t}^2 - \frac{1}{4\lambda}d_y(y_{t+1},y_t)^2. \tag{50}$$

From Lemma 29, with $D_\delta := \frac{c_\rho^2 C_{fy}^2 \rho_F^2}{2\mu_c^4} \le \frac{c_\rho^2 C_{fy}^2 L_{gyy}^2 d_y\,\delta_\varepsilon^2}{2\mu_c^4}$,

$$e_{t+1} \le \Big(1 - \frac{\mu_c\tau}{4}\Big)e_t - \frac{3}{4}\|v_{t+1} - P_t v_t\|^2 + D_\delta + \frac{50\bar{L}^2}{3\mu_c\tau}\Big(d_x(x_{t+1},x_t)^2 + d_y(y_{t+1},y_t)^2\Big). \tag{51}$$

Again substituting $d_x(x_{t+1},x_t)^2 = \gamma^2\|\mathcal{G}_t\|_{x_t}^2$ gives

$$e_{t+1} \le \Big(1 - \frac{\mu_c\tau}{4}\Big)e_t - \frac{3}{4}\|v_{t+1} - P_t v_t\|^2 + D_\delta + \frac{50\bar{L}^2}{3\mu_c\tau}\gamma^2\|\mathcal{G}_t\|_{x_t}^2 + \frac{50\bar{L}^2}{3\mu_c\tau}d_y(y_{t+1},y_t)^2. \tag{52}$$

Adding (48), (50), and (52), we obtain

$$\begin{aligned}\Psi_{t+1} \le \Psi_t &- \Big(\frac{\gamma}{2} - \frac{\gamma}{8} - \frac{50\bar{L}^2}{3\mu_c\tau}\gamma^2\Big)\|\mathcal{G}_t\|_{x_t}^2 \\ &- \Big(\frac{\lambda\mu}{2} - \frac{12\gamma A}{\mu}\Big)\Delta_t - \Big(\frac{\mu_c\tau}{4} - 12\gamma C_{gxy}^2\Big)e_t \\ &- \Big(\frac{1}{4\lambda} - \frac{50\bar{L}^2}{3\mu_c\tau}\Big)d_y(y_{t+1}, y_t)^2 - \frac{3}{4}\|v_{t+1} - P_t v_t\|^2 \\ &+ 2\gamma L_{gxy}^2\delta_\varepsilon^2 r_v^4 + D_\delta. \end{aligned} \tag{53}$$

By the stated choices of γ, λ, τ in (46),

$$\frac{50\bar{L}^2}{3\mu_c\tau}\gamma^2 \le \frac{\gamma}{8}, \qquad \frac{12\gamma A}{\mu} \le \frac{2\lambda\mu}{5}, \qquad 12\,\gamma\, C_{gxy}^2 \le \frac{\mu_c\tau}{5}, \qquad \frac{50\bar{L}^2}{3\mu_c\tau} \le \frac{1}{4\lambda}.$$

Hence the coefficients of Δ_t and e_t in (53) are at most $-\lambda\mu/10$ and $-\mu_c\tau/20$, the remaining terms on the second and third lines are nonpositive, and therefore

$$\Psi_{t+1} \le \Psi_t - \frac{\gamma}{8}\|\mathcal{G}_t\|_{x_t}^2 - \frac{\lambda\mu}{10}\Delta_t - \frac{\mu_c\tau}{20}e_t + \frac{5}{2}\gamma L_{gxy}^2\delta_\varepsilon^2 r_v^4 + D_\delta, \tag{54}$$

where we only loosened the factor 2 to $\frac{5}{2}$ for convenience.

Summing (54) from $t = 1$ to T and using $\Psi_{T+1} \ge \Phi_\star$ (since $\Delta_{T+1} \ge 0$, $e_{T+1} \ge 0$, and $\Phi(x_{T+1}) \ge \Phi_\star$) yields

$$\sum_{t=1}^{T}\Big(\frac{\gamma}{8}\|\mathcal{G}_t\|_{x_t}^2 + \frac{\lambda\mu}{10}\Delta_t + \frac{\mu_c\tau}{20}e_t\Big) \le \Psi_1 - \Phi_\star + \frac{5}{2}T\gamma L_{gxy}^2\delta_\varepsilon^2 r_v^4 + TD_\delta, \tag{55}$$

with $\Psi_1 - \Phi_\star = \Phi(x_1) + g(x_1, y_1) - G(x_1) + \|v_1 - v_1^\star\|_{y_1}^2 - \Phi_\star$. Dropping the Δ_t and e_t terms and dividing by $T\gamma/8$ gives (47). The $O(T^{-1})$ consequence follows from the stated choice of δ_ε, under which $20L_{gxy}^2\delta_\varepsilon^2 r_v^4 \le 20/T$ and $8D_\delta/\gamma \le 4/T$. The final statement for $\phi \equiv 0$ is immediate from the definition of the Riemannian gradient mapping. □

Proof of Corollary 14. Let $\phi \equiv 0$, so $\mathcal{G}_t = w_t$. By (45) in the proof of Lemma 31, $\|\mathcal{G}F(x_t)\|^2 \le 2\|w_t\|^2 + 2\|w_t - \mathcal{G}F(x_t)\|^2 \le 2\|w_t\|^2 + \frac{24A}{\mu}\Delta_t + 24C_{gxy}^2 e_t + 4L_{gxy}^2\delta_\varepsilon^2 r_v^4$. By (55), each of $\frac{\gamma}{8}\cdot\frac{1}{T}\sum_t\|w_t\|^2$, $\frac{\lambda\mu}{10}\cdot\frac{1}{T}\sum_t\Delta_t$, and $\frac{\mu_c\tau}{20}\cdot\frac{1}{T}\sum_t e_t$ is at most $\frac{\Psi_1-\Phi_\star}{T} + \frac{5}{2}\gamma L_{gxy}^2\delta_\varepsilon^2 r_v^4 + D_\delta$, which is $O(1/T)$ under the choice of δ_ε in Theorem 35. Averaging the first display over t therefore gives $\frac{1}{T}\sum_{t=1}^{T}\|\mathcal{G}F(x_t)\|_{x_t}^2 = O(1/T)$. □

G.4 Proof of Corollary 15

We give the modifications needed to extend Theorem 13 to retractions and an isometric vector transport under Assumption 22, with $\phi \equiv 0$, $\|\mathcal{G}_x f\| \le C_{fx}$, and the finite-difference probes kept on the exponential map and parallel transport (or replaced by the closed-form Hessian). These restrictions are needed: geodesic convexity of ϕ does not imply convexity along retraction curves; a transport that is only first-order close to parallel transport changes squared norms at first order in the step and biases central-difference Hessian probes by a non-vanishing amount; and M_F requires a bound on $\mathcal{G}_x f$. Under them, the structure of the proof of Theorem 35 is preserved; only three inequalities inside Lemmas 29, 31, and 32 acquire additional terms, each of which is absorbed into a modified constant.

Modified algorithm. Replace, in Algorithm 1, $y_{t+1} = \mathrm{Exp}_{y_t}(-\lambda u_t)$ by $y_{t+1} = \mathrm{Retr}_{y_t}(-\lambda u_t)$, the transport $\Gamma_{y_t \to y_{t+1}}$ in line 9 by an isometric vector transport $\overline{\Gamma}_{y_t \to y_{t+1}}$, and (with $\phi \equiv 0$) the upper-level step by $x_{t+1} = \mathrm{Retr}_{x_t}(-\gamma w_t)$; the probes (9)–(10) are unchanged. The stationarity measure becomes

$$\mathcal{G}_{\mathrm{Retr}}(x_t, w_t, \gamma) \;:=\; -\gamma^{-1}\,\mathrm{Retr}_{x_t}^{-1}(x_{t+1}) \;=\; w_t.$$

Modification to Lemma 29 (auxiliary-vector recursion). Steps 1–3 take place in $T_{y_t}\mathcal{M}_y$ and are unchanged. In Step 5, for the pre-transport vector $\tilde{v}_{t+1}$, the isometry of $\overline{\Gamma}$ gives $\|v_{t+1} - \overline{\Gamma}_{y_t \to y_{t+1}} v_t^\star\|_{y_{t+1}} = \|\tilde{v}_{t+1} - v_t^\star\|_{y_t}$, exactly as before. In Step 4 the drift term becomes $\|v_{t+1}^\star - \overline{\Gamma}_{y_t \to y_{t+1}} v_t^\star\|_{y_{t+1}}^2$, and

$$\|v_{t+1}^\star - \overline{\Gamma}_{y_t \to y_{t+1}} v_t^\star\|_{y_{t+1}}^2 \;\le\; 2\,\|v_{t+1}^\star - \Gamma_{y_t \to y_{t+1}} v_t^\star\|_{y_{t+1}}^2 \;+\; 2\,\big\|(\overline{\Gamma} - \Gamma)_{y_t \to y_{t+1}} v_t^\star\big\|_{y_{t+1}}^2.$$

The first term is bounded as in the original proof. For the second, by (15) together with the uniform bound $\|v_t^\star\|_{y_t} \le C_{fy}/\mu_c$,

$$\big\|(\overline{\Gamma} - \Gamma)_{y_t \to y_{t+1}} v_t^\star\big\|_{y_{t+1}}^2 \;\le\; c_T^2\, \frac{C_{fy}^2}{\mu_c^2}\, d(y_t, y_{t+1})^2.$$

With $a := c_T^2 C_{fy}^2/\mu_c^2$, $s_x := d(x_{t+1}, x_t)$, $s_y := d(y_t, y_{t+1})$ and $L_v^2 = 4\bar{L}^2$, the drift is therefore at most $8\bar{L}^2(s_x^2 + s_y^2) + 2a\, s_y^2 \le 4 \cdot 2\widetilde{\bar{L}}^{\,2}(s_x^2 + s_y^2)$ with

$$\widetilde{\bar{L}}^{\,2} \;:=\; \bar{L}^2 + \frac{c_T^2 C_{fy}^2}{4\mu_c^2},$$

so the recursion of Lemma 29 holds with $\bar{L}^2$ replaced by $2\widetilde{\bar{L}}^{\,2}$, in terms of the geodesic distances s_x, s_y.

Modification to Lemma 31 (upper-level descent). With $\phi \equiv 0$, $\mathrm{Retr}_{x_t}^{-1}(x_{t+1}) = -\gamma w_t$ exactly, so Step 1 of the proof is replaced by the identity $\langle w_t, \mathrm{Retr}_{x_t}^{-1}(x_{t+1})\rangle_{x_t} = -\gamma\|w_t\|_{x_t}^2$. In the descent inequality

$$F(x_{t+1}) \;\le\; F(x_t) + \langle \mathcal{G}F(x_t), \mathrm{Exp}_{x_t}^{-1}(x_{t+1})\rangle_{x_t} + \tfrac{L_F}{2}\, d^2(x_t, x_{t+1}),$$

decompose

$$\langle \mathcal{G}F(x_t), \mathrm{Exp}_{x_t}^{-1}(x_{t+1})\rangle_{x_t} \;=\; -\gamma\langle \mathcal{G}F(x_t), w_t\rangle_{x_t} \;+\; \langle \mathcal{G}F(x_t), \mathrm{Exp}_{x_t}^{-1}(x_{t+1}) - \mathrm{Retr}_{x_t}^{-1}(x_{t+1})\rangle_{x_t}.$$

By Cauchy–Schwarz and (14), the second term is at most $M_F\, c_R\, \gamma^2\, \|w_t\|_{x_t}^2$, where

$$M_F := \sup_{x \in \mathcal{U}_x} \|\mathcal{G}F(x)\|_x \le C_{fx} + C_{gxy} C_{fy}/\mu$$

by Lemma 28(i). For the squared-distance term, (13) gives $d^2(x_t, x_{t+1}) \le \bar{c}\,\gamma^2\, \|w_t\|_{x_t}^2$. Following the Young's-inequality argument of the proof of Lemma 31 yields the same bound with $\mathcal{G} = \mathcal{G}_{\mathrm{Retr}} = w_t$, provided $\gamma \le 1/(2\widetilde{L}_F)$ with $\widetilde{L}_F := \bar{c}\, L_F + 2c_R\, M_F$.

Modification to Lemma 32 (lower-level descent). The proof expands $g(x_{t+1}, y_{t+1})$ along $y_t \to y_{t+1}$. Two changes occur. First, $d^2(y_t, y_{t+1}) \le \bar{c}\,\|\eta_t\|_{y_t}^2$ with $\eta_t = -\lambda u_t$ by (13), so the curvature term becomes $\frac{\bar{c}\zeta L_g}{2}\|\eta_t\|^2$. Second, the inner-product term acquires the correction

$$\big|\langle \mathcal{G}_y g(x_{t+1}, y_t), \mathrm{Exp}_{y_t}^{-1}(y_{t+1}) - \eta_t\rangle\big| \le c_R\, \|\mathcal{G}_y g(x_{t+1}, y_t)\|\, \|\eta_t\|_{y_t}^2 \le c_R L_g D_y \|\eta_t\|_{y_t}^2$$

by (14), using $\|\mathcal{G}_y g(x, y)\| \le L_g\, d(y, \mathcal{M}^\star(x)) \le L_g D_y$. Both are absorbed by the negative term $-\frac{1}{2\lambda}\|\eta_t\|^2$: the coefficient of $\|\eta_t\|^2$ is at most $-\frac{1}{4\lambda}$ provided $\lambda \le 1/(2\bar{c}\zeta L_g + 4c_R L_g D_y)$. The resulting inequality has the form of Lemma 32, with two differences that we keep explicit: its negative term is $-\frac{1}{4\lambda}\|\eta_t\|_{y_t}^2$ in the retraction-step length (not converted to d_y), and, because $d_x(x_{t+1}, x_t)^2 \le \bar{c}\,\gamma^2\|\mathcal{G}_{\mathrm{Retr}}\|^2$ instead of $= \gamma^2\|\mathcal{G}\|^2$, we apply the lemma with its parameter γ replaced by $\bar{c}\gamma$ (its two γ-caps divided by $\bar{c}$), so that its positive term satisfies $\frac{1}{8\bar{c}\gamma} d_x(x_{t+1}, x_t)^2 \le \frac{\gamma}{8}\|\mathcal{G}_{\mathrm{Retr}}\|^2$.

Conclusion. Substituting the three modified lemmas into the Lyapunov argument of Theorem 35, with geodesic distances and retraction-step lengths kept distinct ($d_x^2 \le \bar{c}\gamma^2\|\mathcal{G}_{\mathrm{Retr}}\|^2$ and $d_y^2 \le \bar{c}\|\eta_t\|^2$ in the auxiliary drift, against $-\frac{1}{4\lambda}\|\eta_t\|^2$ from the lower level), the drift coefficient $\frac{100\widetilde{\bar{L}}^{\,2}}{3\mu_c\tau}$ requires $\gamma \le 3\mu_c\tau/(800\bar{c}\,\widetilde{\bar{L}}^{\,2})$ and $\lambda \le 3\mu_c\tau/(400\bar{c}\,\widetilde{\bar{L}}^{\,2})$, i.e. the two drift caps of (46) with $\bar{L}^2 \to$

$2\bar{c}\,\widetilde{\bar{L}}^{\,2}$. Together with $L_F \to \widetilde{L}_F$, the two γ-caps of Lemma 32 divided by $\bar{c}$, and $\lambda \le 1/(2\bar{c}\,\zeta L_g + 4c_R L_g D_y)$, every cancellation of the proof holds, which gives

$$\frac{1}{T}\sum_{t=1}^{T} \|\mathcal{G}_{\mathrm{Retr}}(x_t, w_t, \gamma)\|_{x_t}^2 \;\le\; \frac{8\left(\Phi(x_1) + g(x_1, y_1) - G(x_1) + \|v_1 - v_1^\star\|_{y_1}^2 - \Phi^\star\right)}{T\gamma} + 20\, L_{gxy}^2\, \delta_\varepsilon^2\, r_v^4 + \frac{4c_\rho^2\, C_{fy}^2\, L_{gyy}^2\, d_y\, \delta_\varepsilon^2}{\gamma\, \mu_c^4}.$$

Choosing δ_ε as in Theorem 35 yields the $O(1/T)$ rate, and hence $O(\varepsilon^{-1})$ iteration complexity for ε-stationarity in the sense of $\mathcal{G}_{\mathrm{Retr}}$. Setting $\bar{c} = 1$, $c_R = c_T = 0$ recovers Theorem 35 with $\phi \equiv 0$, up to the factor 2 in the two drift caps, which comes from splitting off the transport error and is not needed when $c_T = 0$, completing the proof of Corollary 15. □

H Stochastic Algorithm and Analysis

This appendix supplements Section 3.2 by giving (i) the explicit pseudocode of SR-HJFBiO with coupled finite-difference sampling, (ii) the formal variance assumption on the coupled estimators, and (iii) the key one-line variance lemma that justifies the $O(1/B)$ term in Theorem 17.

Coupled-sampling design. The pseudocode below differs from a naive stochastic version of Algorithm 1 in exactly one place: the central finite differences in lines 4 and 7 reuse a *single* mini-batch ($\mathcal{B}_t^J$ or $\mathcal{B}_t^H$) at both the $+\delta_\varepsilon$ and $-\delta_\varepsilon$ perturbations, rather than drawing two independent batches. The reason was previewed in the body and is made precise in Lemma 38 below: with independent batches, the central difference is the difference of two uncorrelated noisy gradients, whose variance is $\Theta(1)$ in δ_ε, and the $1/(2\delta_\varepsilon)$ scaling then inflates the estimator variance to $\Theta(1/\delta_\varepsilon^2|\mathcal{B}|)$. With a coupled batch, the difference becomes a sample average of *component-wise* gradient differences, each of which is $2\delta_\varepsilon$ times a per-component Hessian–vector product up to an $O(\delta_\varepsilon^2)$ residual along the geodesic, and in any case bounded by $2\delta_\varepsilon L_g\|v\|$; the two δ_ε's cancel and the estimator variance becomes $O(1/|\mathcal{B}|)$ with constants determined by the gradient Lipschitz constant L_g of the components. This is the manifold analogue of common-random-numbers estimators in stochastic finite-difference optimization. The same construction underlies SPSA-style derivative-free methods in the Euclidean setting; here it is essential because it is what allows us to avoid stochastic Hessian/Jacobian oracles while retaining the $O(\varepsilon^{-2})$ stochastic complexity.

Algorithm 2 Stochastic Riemannian Hessian/Jacobian-free Bilevel Optimization (SR-HJFBiO)

Require: T; stepsizes $\lambda > 0$, $\gamma > 0$, $\tau > 0$; clip floor $0 < \mu_c \le \mu$ and ceiling L_g; finite-difference parameter $\delta_\varepsilon > 0$; radius $r_v \ge C_{fy}/\mu_c$; initial $x_1 \in \mathcal{M}_x$, $y_1 \in \mathcal{M}_y$, and $v_1 \in T_{y_1}\mathcal{M}_y$ with $\|v_1\|_{y_1} \le r_v$.

1: **for** $t = 1, 2, \dots, T$ **do**

2: Sample mini-batches $\mathcal{B}_t^g$, $\mathcal{B}_t^{f,x}$, $\mathcal{B}_t^{f,y}$, $\mathcal{B}_t^J$, and per-direction batches $\{\mathcal{B}_t^{H,i}\}_{i=1}^{d_y}$ ▷ coupled $\pm$ within $\mathcal{B}_t^J$ and within each $\mathcal{B}_t^{H,i}$

3: **(Stochastic lower-level step)**

$$u_t \leftarrow \mathcal{G}_y g_{\mathcal{B}_t^g}(x_t, y_t) \in T_{y_t}\mathcal{M}_y, \qquad y_{t+1} \leftarrow \mathrm{Exp}_{y_t}(-\lambda u_t)$$

4: **(Stochastic Jacobian-free cross term, coupled batch, at (x_t, y_t, v_t))**

$$\widehat{J}_t \leftarrow \frac{\mathcal{G}_x g_{\mathcal{B}_t^J}\big(x_t, \mathrm{Exp}_{y_t}(\delta_\varepsilon v_t)\big) - \mathcal{G}_x g_{\mathcal{B}_t^J}\big(x_t, \mathrm{Exp}_{y_t}(-\delta_\varepsilon v_t)\big)}{2\delta_\varepsilon}$$

5:

$$\widehat{w}_t \leftarrow \mathcal{G}_x f_{\mathcal{B}_t^{f,x}}(x_t, y_t) - \widehat{J}_t$$

6: **(Upper-level step)** $x_{t+1} \leftarrow P_\phi^\gamma(x_t, \widehat{w}_t)$

7: **(Operator construction, coupled-batch basis sweep)** For each orthonormal basis direction $e_i \in T_{y_t}\mathcal{M}_y$, $i = 1, \dots, d_y$, define the column

$$\begin{aligned}\widehat{\mathcal{H}}_t^\circ\, e_i \;\leftarrow\; \frac{1}{2\delta_\varepsilon}\Big[&\Gamma_{\mathrm{Exp}_{y_t}(\delta_\varepsilon e_i)\to y_t}\mathcal{G}_y g_{\mathcal{B}_t^{H,i}}\big(x_t, \mathrm{Exp}_{y_t}(\delta_\varepsilon e_i)\big) \\ &- \Gamma_{\mathrm{Exp}_{y_t}(-\delta_\varepsilon e_i)\to y_t}\mathcal{G}_y g_{\mathcal{B}_t^{H,i}}\big(x_t, \mathrm{Exp}_{y_t}(-\delta_\varepsilon e_i)\big)\Big],\end{aligned}$$

assemble $\widehat{\mathcal{H}}_t^\circ \in \mathrm{End}(T_{y_t}\mathcal{M}_y)$ by linearity, and symmetrize, $\widehat{\mathcal{H}}_t \leftarrow \frac{1}{2}\big(\widehat{\mathcal{H}}_t^\circ + (\widehat{\mathcal{H}}_t^\circ)^*\big)$. (When the Hessian is available in closed form, set $\widehat{\mathcal{H}}_t = \mathrm{Hess}_y\, g_{\mathcal{B}_t^H}(x_t, y_t)$ directly.)

8: **(Spectral clip and auxiliary update)**

$$\widehat{h}_t \leftarrow \mathcal{S}_{[\mu_c, L_g]}[\widehat{\mathcal{H}}_t][v_t] - \mathcal{G}_y f_{\mathcal{B}_t^{f,y}}(x_t, y_t)$$

9: $v_{t+1} \leftarrow \Gamma_{y_t \to y_{t+1}} \Pi_{r_v}^{y_t}\big(P_t(v_t - \tau \widehat{h}_t)\big)$, with P_t the orthogonal projector onto $\mathrm{range}(\mathcal{S}_{[\mu_c, L_g]}[\widehat{\mathcal{H}}_t])$

10: **end for**

11: **return** Choose uniformly at random from $\{x_t\}_{t=1}^T$

Assumption 37 (Coupled finite-difference variance, full statement)**.** *The mini-batch indices in $\mathcal{B}_t^J$ and in each per-direction batch $\mathcal{B}_t^{H,i}$ ($i = 1, \dots, d_y$) are drawn uniformly without replacement from $[m]$, independently of $\{(x_s, y_s, v_s)\}_{s \le t}$, and the same batch is used at the $+\delta_\varepsilon$ and $-\delta_\varepsilon$ perturbations within a given iteration t and direction i. There exist constants $\sigma_J^2, \sigma_H^2 > 0$, depending only on (L_g, r_v), such that for all $(x_t, y_t, v_t) \in \mathcal{U}_x \times \mathcal{U}_y \times \{v \in T_{y_t}\mathcal{M}_y : \|v\| \le r_v\}$,*

$$\begin{aligned}\mathbb{E}\big[\|\widehat{J}_t - \mathbb{E}\widehat{J}_t\|_{x_t}^2 \;\big|\; x_t, y_t, v_t\big] &\le \frac{\sigma_J^2}{|\mathcal{B}_t^J|}, \\ \mathbb{E}\big[\|(\widehat{\mathcal{H}}_t^\circ - \mathbb{E}\widehat{\mathcal{H}}_t^\circ)\, e_i\|_{y_t}^2 \;\big|\; x_t, y_t, v_t\big] &\le \frac{\sigma_H^2}{|\mathcal{B}_t^{H,i}|}, \quad i = 1, \dots, d_y.\end{aligned} \tag{56}$$

With a common per-direction batch size $|\mathcal{B}_t^{H,i}| = B_H$, summing across directions and using that symmetrization is a contraction in the Frobenius norm yields $\mathbb{E}[\|\widehat{\mathcal{H}}_t - \mathbb{E}\widehat{\mathcal{H}}_t\|_F^2] \le \sigma_H^2\, d_y / B_H$, which also bounds the squared operator norm. The d_y-fold mini-batch cost is precisely what yields the d_y factor in the basis-sweep first-order oracle complexity (cf. Remark 41). Moreover, $\widehat{J}_t$ and $\widehat{\mathcal{H}}_t$ are biased toward $\mathcal{G}_{xy}^2 g(x_t, y_t)[v_t]$ and $\mathrm{Hess}_y\, g(x_t, y_t)$ respectively, with bias at most $L_{gxy}\delta_\varepsilon r_v^2$ and, in Frobenius norm, $\sqrt{d_y}\, L_{gyy}\delta_\varepsilon$ (first order in δ_ε under the Lipschitz-Hessian assumptions; second order would require Lipschitz third derivatives).

Lemma 38 (Coupled finite-difference variance)**.** *Under Assumptions 16 and 37, with coupled mini-batches as in Algorithm 2 lines 4–7, the bound* (56) *holds with*

$$\sigma_J^2 = L_g^2 r_v^2, \qquad \sigma_H^2 = L_g^2.$$

With per-direction batch size B_H, this gives $\mathbb{E}[\|\widehat{\mathcal{H}}_t - \mathbb{E}\widehat{\mathcal{H}}_t\|_F^2] \le L_g^2 d_y / B_H$.

Proof. Fix t and condition on (x_t, y_t, v_t). Writing $y_\pm := \mathrm{Exp}_{y_t}(\pm\delta_\varepsilon v_t)$ and using the coupled batch $\mathcal{B} := \mathcal{B}_t^J$,

$$\widehat{J}_t = \frac{1}{2\delta_\varepsilon|\mathcal{B}|}\sum_{i\in\mathcal{B}}\big[\mathcal{G}_x g_i(x_t, y_+) - \mathcal{G}_x g_i(x_t, y_-)\big] = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} J_i,$$

with $J_i := \big(\mathcal{G}_x g_i(x_t, y_+) - \mathcal{G}_x g_i(x_t, y_-)\big)/(2\delta_\varepsilon)$. We bound the variances through *magnitudes* rather than the Lipschitz constants L_{gxy}, L_{gyy} of the derivatives, since for quadratic components $g_j(y) = a_j y^2/2$ the latter vanish while the variance does not. Since $d(y_+, y_-) = 2\delta_\varepsilon\|v_t\|$ and $\mathcal{G}_x g_i$ is L_g-Lipschitz in y (Assumption 20, inherited by each component under Assumption 16), $\|J_i\|_{x_t} \le L_g\|v_t\| \le L_g r_v$ for every i. The batch is drawn uniformly without replacement, so $\widehat{J}_t$ is a sample mean without replacement; its variance is the population variance of $\{J_i\}$ times $\frac{m-|\mathcal{B}|}{(m-1)|\mathcal{B}|} \le \frac{1}{|\mathcal{B}|}$, and the population variance is at most $\max_i \|J_i\|^2 \le L_g^2 r_v^2$. For the operator, fix a direction i and the coupled batch $\mathcal{B}_t^{H,i}$; the per-component summand $H_{i,j} := [\Gamma_+\mathcal{G}_y g_j(x_t, y_+^{(i)}) - \Gamma_-\mathcal{G}_y g_j(x_t, y_-^{(i)})]/(2\delta_\varepsilon)$ involves points $y_-^{(i)}, y_t, y_+^{(i)}$ on one geodesic, along which the composed parallel transports agree with the direct one, so the transported L_g-Lipschitz bound on $\mathcal{G}_y g_j$ gives $\|H_{i,j}\|_{y_t} \le L_g$. The same finite-population argument yields the per-direction bound $L_g^2/|\mathcal{B}_t^{H,i}|$, and the independence of $\{\mathcal{B}_t^{H,i}\}_{i=1}^{d_y}$ and the Frobenius contraction of symmetrization give the Frobenius bound. Finally, the conditional means are the deterministic probes of g, whose errors are bounded in (44) and in Step 3 of the proof of Lemma 29, which gives the bias statements of Assumption 37. □

Remark 39 (Why independent batches fail)**.** *If $\mathcal{B}_t^{J,+}$ and $\mathcal{B}_t^{J,-}$ are sampled independently (so the $\pm$ batches do not share indices), the per-component cancellation in the proof above no longer applies, because $\mathcal{G}_x g_i(x_t, y_+)$ and $\mathcal{G}_x g_j(x_t, y_-)$ for $i \ne j$ are evaluations of* different *component functions at* different *points and their difference does not collapse to a multiple of δ_ε. Instead, $\widehat{J}_t^{\mathrm{indep}} = [\mathcal{G}_x g_{\mathcal{B}^+}(x_t, y_+) - \mathcal{G}_x g_{\mathcal{B}^-}(x_t, y_-)]/(2\delta_\varepsilon)$, and bias–variance decomposition gives variance $\Theta(\sigma_{gx}^2/(\delta_\varepsilon^2|\mathcal{B}|))$. Under the schedule $\delta_\varepsilon = \Theta(\varepsilon^{1/2})$, $|\mathcal{B}| = \Theta(\varepsilon^{-1})$ this product is $\Theta(1)$ – bounded but not vanishing – and would replace the C_3/B term in Theorem 17 by an irreducible $\Theta(1)$ term, breaking the ε-stationarity argument of Corollary 18. We therefore impose coupling at the algorithm level, not as a post-hoc fix.*

Remark 40 (Comparison with stochastic Riemannian bilevel methods using direct second-order oracles)**.** *The stochastic Riemannian bilevel framework of [43] samples stochastic Hessian and Jacobian objects directly (its mini-batches B_3, B_4 are drawn for $\mathcal{G}_{xy}^2 g$ and $\mathrm{Hess}_y\, g$, not for finite differences of $\mathcal{G}g$). In that setting there is no δ_ε and the variance is $\Theta(1/|B_i|)$ by construction; the question of independent vs. coupled $\pm$ batches simply does not arise. The trade-off is access to second-order oracles. Our design avoids that requirement by paying the price of central differences, which is precisely why the coupled-sampling discipline is forced on us.*

Remark 41 (Per-iteration oracle count)**.** *With coupled batches, each iteration of Algorithm 2 requires $|\mathcal{B}_t^g| + |\mathcal{B}_t^{f,x}| + |\mathcal{B}_t^{f,y}| + 2|\mathcal{B}_t^J| + 2\sum_{i=1}^{d_y}|\mathcal{B}_t^{H,i}|$ first-order oracle calls; the factor 2 in each finite-difference term accounts for the $\pm$ pair sharing one batch but evaluating the gradient twice, and the basis-sweep operator construction in line 7 contributes one such pair per direction $i = 1, \dots, d_y$. With a common batch size B for all mini-batches this is $(5 + 2d_y)B$ calls per iteration, matching the per-iteration first-order cost of the deterministic algorithm. In the closed-form realization (with stochastic gradients but a directly evaluated Hessian) the cost reduces to $5B$ first-order calls plus the cost of evaluating one stochastic Hessian. This d_y-fold per-iteration cost in the basis-sweep operator construction is the source of the d_y factor in $\sigma_H^2 d_y/B_H$ in Lemma 38; equivalently, it is the price of obtaining a fresh stochastic Hessian–vector product for each of the d_y basis directions per iteration.*

H.1 Proof of Theorem 17 and Corollary 18

We extend the deterministic Lyapunov argument of Theorem 35 to the stochastic setting. The strategy is to derive stochastic counterparts of Lemmas 29, 31, and 32, each picking up explicit variance terms from the mini-batch oracles, and then to telescope the resulting expected Lyapunov inequality. The structure mirrors [54] for the Euclidean case, with the curvature- and transport-distortion-induced corrections of Section 3 absorbed into the deterministic constants $\bar{L}^2$, L_v^2, A as in Theorem 35.

Setup and notation. Let $\mathcal{F}_t$ denote the σ-algebra generated by the initial data (x_1, y_1, v_1) and all mini-batches drawn at iterations $s < t$. Then (x_t, y_t, v_t) is $\mathcal{F}_t$-measurable, and the mini-batches drawn at iteration t are independent of $\mathcal{F}_t$. Write $\mathbb{E}_t[\,\cdot\,] := \mathbb{E}[\,\cdot \mid \mathcal{F}_t]$. With a common batch size $|\mathcal{B}_t^\bullet| = B$, define the mean-zero stochastic noise terms

$$\xi_t^g := u_t - \mathcal{G}_y g(x_t, y_t), \qquad \xi_t^{f,x} := \mathcal{G}_x f_{\mathcal{B}_t^{f,x}}(x_t, y_t) - \mathcal{G}_x f(x_t, y_t),$$
$$\xi_t^{f,y} := \mathcal{G}_y f_{\mathcal{B}_t^{f,y}}(x_t, y_t) - \mathcal{G}_y f(x_t, y_t), \qquad \xi_t^J := \widehat{J}_t - \mathbb{E}[\widehat{J}_t \mid \mathcal{F}_t],$$
$$\xi_t^H := \widehat{\mathcal{H}}_t - \mathbb{E}[\widehat{\mathcal{H}}_t \mid \mathcal{F}_t].$$

By Assumption 16,

$$\mathbb{E}_t\big[\|\xi_t^g\|_{y_t}^2\big] \le \tfrac{\sigma_{gy}^2}{B}, \quad \mathbb{E}_t\big[\|\xi_t^{f,x}\|_{x_t}^2\big] \le \tfrac{\sigma_{fx}^2}{B}, \quad \mathbb{E}_t\big[\|\xi_t^{f,y}\|_{y_t}^2\big] \le \tfrac{\sigma_{fy}^2}{B},$$

and by Assumption 37,

$$\mathbb{E}_t\big[\|\xi_t^J\|_{x_t}^2\big] \le \tfrac{\sigma_J^2}{B}, \quad \mathbb{E}_t\big[\|\xi_t^H\|_{y_t \to y_t}^2\big] \le \mathbb{E}_t\big[\|\xi_t^H\|_F^2\big] \le \tfrac{\sigma_H^2 d_y}{B}.$$

The conditional means are biased toward the deterministic targets:

$$\mathbb{E}[\widehat{J}_t \mid \mathcal{F}_t] = \mathcal{G}_{xy}^2 g(x_t, y_t)[v_t] + b_t^J, \ \mathbb{E}[\widehat{\mathcal{H}}_t \mid \mathcal{F}_t] = \mathrm{Hess}_y\, g(x_t, y_t) + b_t^H, \tag{57}$$

with $\|b_t^J\|_{x_t} \le L_{gxy}\delta_\varepsilon r_v^2$ (by (44)) and $\|b_t^H\|_F \le \sqrt{d_y}\, L_{gyy}\delta_\varepsilon$ (by Step 3 of the proof of Lemma 29), since the conditional means are the deterministic probes of g; these bias bounds are first order in δ_ε, consistent with the deterministic analysis. Without loss of generality assume $\delta_\varepsilon \le 1$.

Step 1: Stochastic lower-level descent. Lemma 32 holds pathwise for any $u_t \in T_{y_t}\mathcal{M}_y$, with the residual term equal to $\lambda\|\mathcal{G}_y g(x_t, y_t) - u_t\|_{y_t}^2 = \lambda\|\xi_t^g\|_{y_t}^2$. Therefore

$$\Delta_{t+1} \le \Big(1 - \tfrac{\lambda\mu}{2}\Big)\Delta_t + \tfrac{1}{8\gamma} d_x(x_{t+1}, x_t)^2 - \tfrac{1}{4\lambda} d_y(y_{t+1}, y_t)^2 + \lambda\|\xi_t^g\|_{y_t}^2. \tag{58}$$

Substituting $d_x(x_{t+1}, x_t)^2 = \gamma^2\|\widehat{\mathcal{G}}_t\|_{x_t}^2$ with $\widehat{\mathcal{G}}_t := \mathcal{G}(x_t, \widehat{w}_t, \gamma)$ and taking $\mathbb{E}_t[\,\cdot\,]$,

$$\mathbb{E}_t[\Delta_{t+1}] \le \Big(1 - \tfrac{\lambda\mu}{2}\Big)\Delta_t + \tfrac{\gamma}{8}\mathbb{E}_t\big[\|\widehat{\mathcal{G}}_t\|_{x_t}^2\big] - \tfrac{1}{4\lambda}\mathbb{E}_t\big[d_y(y_{t+1}, y_t)^2\big] + \frac{\lambda\sigma_{gy}^2}{B}. \tag{59}$$

Step 2: Stochastic upper-level descent. Following the proof of Lemma 31 verbatim through equation (37) but with w_t replaced by $\widehat{w}_t$ and $\mathcal{G}_t$ by $\widehat{\mathcal{G}}_t$, the geodesic-smoothness argument gives the pathwise inequality

$$\Phi(x_{t+1}) \le \Phi(x_t) - \tfrac{\gamma}{2}\|\widehat{\mathcal{G}}_t\|_{x_t}^2 + \gamma\,\|\widehat{w}_t - \mathcal{G}F(x_t)\|_{x_t}^2. \tag{60}$$

We bound the residual by decomposing $\widehat{w}_t - \mathcal{G}F(x_t)$ into a mean-zero stochastic part and a deterministic biased part. Using (57) and the deterministic surrogate $\widehat{\mathcal{G}}_x F(x_t, y_t, v_t) := \mathcal{G}_x f(x_t, y_t) - \big(\mathcal{G}_{xy}^2 g(x_t, y_t)\big)[v_t]$,

$$\widehat{w}_t - \mathcal{G}F(x_t) = \big(\xi_t^{f,x} - \xi_t^J\big) + \big(\widehat{\mathcal{G}}_x F(x_t, y_t, v_t) - \mathcal{G}F(x_t)\big) - b_t^J. \tag{61}$$

Applying $(a+b+c)^2 \le 3(a^2+b^2+c^2)$,

$$\|\widehat{w}_t - \mathcal{G}F(x_t)\|_{x_t}^2 \le 3\|\xi_t^{f,x} - \xi_t^J\|_{x_t}^2 + 3\|\widehat{\mathcal{G}}_x F(x_t, y_t, v_t) - \mathcal{G}F(x_t)\|_{x_t}^2 + 3\|b_t^J\|_{x_t}^2. \tag{62}$$

The middle term is the deterministic exact-surrogate quantity bounded by equation (43) of the proof of Lemma 31:

$$\|\widehat{\mathcal{G}}_x F(x_t, y_t, v_t) - \mathcal{G}F(x_t)\|_{x_t}^2 \le \frac{6A}{\mu}\Delta_t + 6C_{gxy}^2 e_t, \tag{63}$$

with $A := L_f^2 + r_v^2 L_{gxy}^2 + 2C_{gxy}^2 L_v^2$ as in Lemma 31. Because Algorithm 2 evaluates $\widehat{w}_t$ at (x_t, y_t, v_t), the deterministic surrogate here is exactly the one bounded in (43), with $y^\star$ the point of $\mathcal{M}^\star(x_t)$ nearest to y_t, so (63) holds without further terms. The bias term satisfies $\|b_t^J\|^2 \le L_{gxy}^2 \delta_\varepsilon^2 r_v^4$ by (57). Taking $\mathbb{E}_t[\,\cdot\,]$ in (62) and using $\mathbb{E}_t\|\xi_t^{f,x} - \xi_t^J\|^2 \le 2(\sigma_{fx}^2 + \sigma_J^2)/B$,

$$\mathbb{E}_t\big[\|\widehat{w}_t - \mathcal{G}F(x_t)\|_{x_t}^2\big] \le \frac{18A}{\mu}\Delta_t + 18C_{gxy}^2 e_t + 3L_{gxy}^2\delta_\varepsilon^2 r_v^4 + \frac{6(\sigma_{fx}^2 + \sigma_J^2)}{B}. \tag{64}$$

Substituting into (60) and taking $\mathbb{E}_t[\,\cdot\,]$,

$$\mathbb{E}_t[\Phi(x_{t+1})] \le \Phi(x_t) - \tfrac{\gamma}{2}\mathbb{E}_t\big[\|\widehat{\mathcal{G}}_t\|_{x_t}^2\big] + \frac{18\gamma A}{\mu}\Delta_t + 18\gamma C_{gxy}^2 e_t + \gamma C_{w,1}\delta_\varepsilon^2 r_v^4 + \frac{\gamma C_{w,2}}{B}, \tag{65}$$

with $C_{w,1} := 3L_{gxy}^2$ and $C_{w,2} := 6(\sigma_{fx}^2 + \sigma_J^2)$.

Step 3: Stochastic auxiliary recursion. We follow the proof of Lemma 29 at (x_t, y_t, v_t), with h_t replaced by $\widehat{h}_t$ and the projected update of line 9 of Algorithm 2; only two steps change, and both ingredients of the deterministic proof (the range projector and the margin) carry over unchanged. (i) Since $\widehat{h}_t = \mathcal{S}_{[\mu_c, L_g]}[\widehat{\mathcal{H}}_t][v_t] - \mathcal{G}_y f(x_t, y_t) - \xi_t^{f,y}$, Step 1 holds with the inexact gradient $\mathcal{G}\widehat{R}_t(u_t) - \bar{\xi}_t$, $\bar{\xi}_t := P_t\xi_t^{f,y}$, so (28) carries the term $\frac{4\tau}{\mu_c}\|\bar{\xi}_t\|^2 \le \frac{4\tau}{\mu_c}\|\xi_t^{f,y}\|^2$. (ii) The operator error $\widehat{\mathcal{H}}_t - \mathrm{Hess}_y\, g(x_t, y_t) = \xi_t^H + b_t^H$ is now random. Take δ_ε small enough that $\|b_t^H\|_{y_t \to y_t} \le \sqrt{d_y} L_{gyy}\delta_\varepsilon \le \rho/4$, and let $\mathcal{A}_t := \{\|\xi_t^H\|_{y_t\to y_t} \le \rho/2 - \|b_t^H\|_{y_t\to y_t}\}$. On $\mathcal{A}_t$, Weyl's inequality shows that $\widehat{\mathcal{H}}_t$ satisfies the margin of Assumption 27 with $\rho/2$, so Step 3 of the deterministic proof applies with c_ρ replaced by $c'_\rho := 2\mu_c/\rho$ and ρ_F^2 by $\|\xi_t^H + b_t^H\|_F^2$. On the complement, the margin-free bounds $\|\mathcal{S}_{[\mu_c, L_g]}[\cdot]^\dagger\| \le 1/\mu_c$ and $\|(I - P_t)v^\star\| \le \|v^\star\| \le C_{fy}/\mu_c$ give $\epsilon_t^2/6 + \|(I - P_t)v^\star\|^2 \le \frac{5}{3}C_{fy}^2/\mu_c^2$, while Chebyshev's inequality and Assumption 37 give $\mathbb{P}(\mathcal{A}_t^c \mid \mathcal{F}_t) \le \mathbb{E}\|\xi_t^H\|_F^2/(\rho/4)^2 \le 16\sigma_H^2 d_y/(\rho^2 B)$. Using $\mathbb{E}\|\xi_t^H + b_t^H\|_F^2 \le 2\sigma_H^2 d_y/B + 2d_y L_{gyy}^2\delta_\varepsilon^2$, and the Young's split of (32) with $1 + \frac{4}{\mu_c\tau} \le \frac{25}{6\mu_c\tau}$ (the same substitution as in the deterministic proof), the stochastic analogue of (22) becomes

$$\begin{aligned}\mathbb{E}_t[e_{t+1}] \le{}& \Big(1 - \tfrac{\mu_c\tau}{4}\Big)e_t - \tfrac{3}{4}\mathbb{E}_t\big[\|v_{t+1} - P_t v_t\|^2\big] \\ &+ \frac{50\bar{L}^2}{3\mu_c\tau}\,\mathbb{E}_t\big[d_x(x_{t+1}, x_t)^2 + d_y(y_{t+1}, y_t)^2\big] + C_{e,1}\delta_\varepsilon^2 + \frac{C_{e,2}}{B},\end{aligned} \tag{66}$$

where

$$C_{e,1} := \frac{c_\rho'^2 C_{fy}^2 L_{gyy}^2 d_y}{\mu_c^4}, \qquad C_{e,2} := \frac{25\tau\sigma_{fy}^2}{6\mu_c} + \frac{c_\rho'^2 C_{fy}^2\sigma_H^2 d_y}{\mu_c^4} + \frac{32\, C_{fy}^2\sigma_H^2 d_y}{\mu_c^2\rho^2}.$$

The last term of $C_{e,2}$ is the one further $O(1/B)$ contribution of the event $\mathcal{A}_t^c$; it enlarges C_3 in Theorem 17 without changing the $O(\varepsilon^{-2})$ complexity.

Step 4: Combined Lyapunov recursion. Define $\Psi_t := \Phi(x_t) + \Delta_t + e_t$ as in Theorem 35. Adding (65), (59), and (66), and using $\mathbb{E}_t[d_x(x_{t+1}, x_t)^2] = \gamma^2\mathbb{E}_t[\|\widehat{\mathcal{G}}_t\|_{x_t}^2]$,

$$\begin{aligned}\mathbb{E}_t[\Psi_{t+1}] \le{}& \Psi_t - \Big(\tfrac{\gamma}{2} - \tfrac{\gamma}{8} - \tfrac{50\bar{L}^2}{3\mu_c\tau}\gamma^2\Big)\mathbb{E}_t\big[\|\widehat{\mathcal{G}}_t\|_{x_t}^2\big] \\ &- \Big(\tfrac{\lambda\mu}{2} - \tfrac{18\gamma A}{\mu}\Big)\Delta_t - \Big(\tfrac{\mu_c\tau}{4} - 18\gamma C_{gxy}^2\Big)e_t \\ &- \Big(\tfrac{1}{4\lambda} - \tfrac{50\bar{L}^2}{3\mu_c\tau}\Big)\mathbb{E}_t\big[d_y(y_{t+1}, y_t)^2\big] - \tfrac{3}{4}\mathbb{E}_t\big[\|v_{t+1} - P_t v_t\|^2\big] \\ &+ K_\delta\delta_\varepsilon^2 + \frac{K_B}{B},\end{aligned} \tag{67}$$

with

$$K_\delta := \gamma C_{w,1} r_v^4 + C_{e,1}, \qquad K_B := \gamma C_{w,2} + C_{e,2} + \lambda \sigma_{gy}^2.$$

The structure of (67) is identical to its deterministic counterpart (53), with the deterministic factors $12\gamma A/\mu$ and $12\gamma C_{gxy}^2$ replaced by $18\gamma A/\mu$ and $18\gamma C_{gxy}^2$ – a factor $3/2$ strengthening of the analytic requirement on the LHS. Because the deterministic cap (46) is already buffered (its $\gamma \le \mu_c\tau/(60C_{gxy}^2)$ gives $12\gamma C_{gxy}^2 \le \mu_c\tau/5 \le \mu_c\tau/4$, with buffer $5/4$), the resulting tightening of the *displayed* cap is by the smaller factor $(3/2)/(5/4) = 6/5$ – explicitly,

$$\gamma \le \tfrac{\mu_c\tau}{60\,C_{gxy}^2} \;\rightsquigarrow\; \gamma \le \tfrac{\mu_c\tau}{72\,C_{gxy}^2}, \qquad \gamma \le \tfrac{\mu^2\lambda}{30\,A} \;\rightsquigarrow\; \gamma \le \tfrac{\mu^2\lambda}{36\,A}$$

– guarantees $18\gamma A/\mu \le \lambda\mu/2$ and $18\gamma C_{gxy}^2 \le \mu_c\tau/4$, so the second and third lines of (67) are nonpositive. The remaining stepsize conditions ($\gamma \le 3\mu_c\tau/(400\bar{L}^2)$ for the $\|\widehat{\mathcal{G}}_t\|^2$ coefficient, $\lambda \le 3\mu_c\tau/(200\bar{L}^2)$ for the d_y^2 coefficient) are unchanged from (46). Consequently

$$\mathbb{E}_t[\Psi_{t+1}] \le \Psi_t - \tfrac{\gamma}{8}\mathbb{E}_t\big[\|\widehat{\mathcal{G}}_t\|_{x_t}^2\big] + K_\delta\,\delta_\varepsilon^2 + \frac{K_B}{B}. \tag{68}$$

Step 5: Telescoping. Taking total expectation in (68) and summing from $t = 1$ to T,

$$\tfrac{\gamma}{8}\sum_{t=1}^{T}\mathbb{E}\big[\|\widehat{\mathcal{G}}_t\|_{x_t}^2\big] \le \Psi_1 - \mathbb{E}[\Psi_{T+1}] + TK_\delta\delta_\varepsilon^2 + \frac{TK_B}{B}.$$

Since $\Delta_{T+1} \ge 0$, $e_{T+1} \ge 0$, and $\Phi(x_{T+1}) \ge \Phi_\star$, $\mathbb{E}[\Psi_{T+1}] \ge \Phi_\star$. Therefore

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\widehat{\mathcal{G}}_t\|_{x_t}^2\big] \le \frac{8(\Phi(x_1) + g(x_1,y_1) - G(x_1) + \|v_1 - v_1^\star\|_{y_1}^2 - \Phi_\star)}{T\gamma} + \frac{8K_\delta}{\gamma}\delta_\varepsilon^2 + \frac{8K_B}{\gamma B}. \tag{69}$$

Setting $C_1 := 8/\gamma$, $C_2 := 8K_\delta/\gamma$, and $C_3 := 8K_B/\gamma$ – each depending only on $(\mu, \mu_c, \rho, d_y, L_f, L_g, L_{gxy}, L_{gyy}, C_{fy}, C_{gxy}, r_v, \zeta, C_0)$ and the variance constants – yields (12) and completes the proof of Theorem 17. The case $\phi \equiv 0$, where $\mathcal{G}_\gamma(x_t, \widehat{w}_t) = \widehat{w}_t$, is immediate from the definition of the gradient mapping, exactly as in Theorem 35. □

Proof of Corollary 18. With the schedule $\delta_\varepsilon = \Theta(\varepsilon^{1/2})$, $B = \Theta(\varepsilon^{-1})$, $T = \Theta(\varepsilon^{-1})$, (69) gives

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\widehat{\mathcal{G}}_t\|_{x_t}^2\big] \le C_1\Delta_1 \cdot \Theta(\varepsilon) + C_2 \cdot \Theta(\varepsilon) + C_3 \cdot \Theta(\varepsilon) = O(\varepsilon).$$

Since batches are drawn without replacement, B cannot exceed the population sizes n, m; if $\Theta(\varepsilon^{-1})$ exceeds one of them, the corresponding oracle uses its full batch, whose variance is zero (the finite-population factor vanishes), so that contribution to C_3/B disappears and the bound still holds. Drawing $\widetilde{t}$ uniformly on $\{1, \dots, T\}$ independently of the algorithm and writing $\widetilde{x} := x_{\widetilde{t}}$, the law of total expectation gives

$$\mathbb{E}\big\|\mathcal{G}_\gamma(\widetilde{x}, \widehat{w}_{\widetilde{t}})\big\|_{\widetilde{x}}^2 = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\widehat{\mathcal{G}}_t\|_{x_t}^2\big] \le \varepsilon,$$

i.e. ε-stationarity in the sense of the Riemannian gradient mapping. The total stochastic first-order oracle cost is $T\cdot(5+2d_y)B = \Theta(d_y\varepsilon^{-2})$ in the basis-sweep realization and $T\cdot 5B = \Theta(\varepsilon^{-2})$ in the closed-form realization, by the per-iteration count of Remark 41. This establishes $O(\varepsilon^{-1})$ iterations and $O(\varepsilon^{-2})$ stochastic first-order oracle calls. □

Remark 42 (Comparison with Euclidean and Riemannian stochastic baselines)**.** *The $O(\varepsilon^{-2})$ stochastic first-order complexity established by Theorem 17 and Corollary 18 matches the order achieved by Euclidean stochastic bilevel methods such as stocBiO [58] and SUSTAIN [61] under the strictly stronger assumption of* strong convexity *of the lower level. The Riemannian stochastic framework of [43] also attains $O(\varepsilon^{-2})$ but accesses stochastic Hessian and Jacobian oracles directly; our design avoids those oracles entirely, at the price of the coupled finite-difference discipline of Lemma 38 and the δ_ε^2 term in* (12)*. Setting $\delta_\varepsilon = 0$ recovers the regime of stochastic gradients with exact second-order information, in which the bias term vanishes and* (12) *reduces to $C_1\Delta_1/T + C_3/B$ – the standard SGD-style complexity for the upper-level problem under the PL lower level.*

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 paper's contributions — an intrinsic regularized tangent-space formulation via spectral clipping, a single-loop Riemannian Hessian/Jacobian-free algorithm (R-HJFBiO), an $\mathcal{O}(1/T)$ convergence rate and $\mathcal{O}(\varepsilon^{-1})$ iteration complexity for ε-stationarity, a stochastic extension with an $O(1/T + \delta_\varepsilon^2 + 1/B)$ upper bound, and experiments on Stiefel/Grassmann/Poincaré geometries. Each claim is backed by a corresponding result: Theorem 13 (deterministic), Theorem 17 and Corollary 18 (stochastic), Section 4 (meta-learning experiments), and Appendix D (regression experiments). Section 4 states that the experiments run an exact-solve variant of the auxiliary step and that two meta-learning cells lie outside the theorem's hypotheses.

   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: Section 5 contains a dedicated "Limitations" paragraph. It discusses (i) the gap between the exponential-map / parallel-transport analysis and the QR retraction used in experiments, noting that Corollary 15 covers retractions only for $\phi \equiv 0$ with an isometric vector transport, and that the experiments' exact auxiliary solve, multiple inner steps, gradient clipping and task resets are not covered; (ii) the requirement that iterates stay in the bounded geodesically convex neighborhoods of Assumption 1, which is not automatic even on Hadamard manifolds, and hypotheses (P1)–(P3) of Lemma 28; and (iii) the standard truncation/cancellation trade-off of the central finite-difference probes in Eqs. (9)–(10), addressed in practice with $\delta_\varepsilon = 10^{-5}$ in double precision (discussed in Appendix F) and entirely sidestepped by the closed-form Hessian realization (route (b)); and (iv) the compatibility condition, the uncertified requirement $\mu_c \leq \mu$, the two meta-learning stress-test cells outside the theory, and the absence of the mini-batch sweep figures.

   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: All assumptions are stated explicitly (Assumptions 1, 2–8 in Section 2, including the compatibility condition of Assumption 5; the spectral margin of Assumption 27 in Appendix G; hypotheses (P1)–(P3) stated in Lemma 28; full statements in Assumptions 20–21 in Appendix B; Assumption 16 for the stochastic setting). Each theorem and lemma is numbered and cross-referenced. Complete proofs are provided in Appendix G (Lemma 28 in the first subsection; supporting Lemmas 29, 31, 32 in the next; Theorem 13 / Theorem 35 in the third). Theorem 13 is restated as Theorem 35 with all stepsize ranges and constants made explicit.

   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 4 specifies the task (5-way 5-shot MiniImageNet, 64:16:20 class split), backbone (Conv-4 with 16 channels, Stiefel-constrained first two kernels), and protocol ($T = 200$, meta-batch 4, divergence guard at 10^8). Appendix D (Sections D.1–D.3) describes the regression template, the BCI IV-2a preprocessing pipeline, the Grassmann/Poincaré feature maps, and the diagnostic protocol. Appendix E specifies the UCI Superconductivity setup. Appendix F gives the closed-form lower-level Hessian, hypergradient hyperparameters for all five methods ($\mu_c = 10^{-3}$, $\delta_\varepsilon = 10^{-5}$, $T_{\mathrm{cg}} = 20$, $T_{\mathrm{ns}} = T_{\mathrm{ad}} = 5$, $\alpha = 5 \cdot 10^{-2}$), and the divergence guard.

   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
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) 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).
  (d) 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 uses publicly available datasets (BCI IV-2a, UCI Superconductivity, MiniImageNet), all of which are referenced and described in detail sufficient for re-implementation. Supplementary code archive is submitted along with instructions.

   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: Hyperparameters are listed in Section 4 and Appendices D.2, E, and F, including outer/inner stepsizes ($\eta_x = \eta_y = 5 \cdot 10^{-2}$), inner iterations ($T_{\text{inner}} = 5$ for the regression experiments; $T_{\text{inner}} = 30$ for the meta-learning experiment), outer iterations ($T = 100$ for the BCI and Poincaré regression, $T = 250$ for the deterministic UCI Superconductivity [37, 38] head-to-head, and $T = 200$ for the MiniImageNet [101, 88] meta-learning experiment), seeds and meta-batch sizes (9 subjects $\times$ 3 seeds for BCI [99]; 5 data $\times$ 3 init for synthetic; meta-batch size 4 for MiniImageNet), and the mini-batch sweep $B \in \{1, 4, 16, 40\}$ used for the spectral-margin check on UCI Superconductivity. Optimizers, retractions (sign-corrected QR; affine-invariant exp on SPD; Möbius-addition exp on Poincaré ball [84, 32]), and gradient clipping are also specified.

   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: Curves report the median across multiple trajectories with a shaded interquartile range (S1, S2: 9 subjects $\times$ 3 seeds $= 27$ trajectories on BCI [99]; S3: 5 data $\times$ 3 init seeds $= 15$ trajectories on the synthetic Poincaré generator). Table 4 reports means over a single meta-batch of 4 tasks on MiniImageNet [101, 88] without repeated runs, so the meta-learning numbers carry no uncertainty estimate (stated in Section 4).

   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: [No]

Justification: The paper mentions that the per-task closed-form Hessian computation (a 2000×2000 `eigh`) runs in roughly 1 second on CPU (Appendix F), but does not provide comprehensive compute-resource documentation (hardware type, total wall-clock time, memory).

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 research is methodological and uses publicly available benchmark datasets with no human-subjects component beyond the BCI IV-2a EEG dataset, which is a longstanding public benchmark released for the express purpose of brain–computer-interface algorithm development. No proprietary, sensitive, or personally identifying data is collected or released.

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: [N/A]

Justification: This is foundational work on optimization algorithms for Riemannian bilevel problems. The contributions are theoretical (convergence analysis under PL geometry) and methodological (a Hessian/Jacobian-free single-loop algorithm). The paper does not introduce any model, data product, or system with a direct path to either positive or negative societal applications beyond the broad domain of optimization research itself.

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 any pretrained models, generative systems, or scraped datasets. The artifacts associated with the work are an algorithm and its convergence analysis, which carry no inherent misuse risk requiring safeguards.

    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 original sources of all datasets used: BCI Competition IV Data set 2a [99], UCI Superconductivity [37, 38], and MiniImageNet [101, 88], the latter built from ImageNet [90]. License names and terms of use are stated in Appendix D (BCI IV: research-use license conditional on citing [99]; UCI: CC BY 4.0; MiniImageNet/ImageNet: ImageNet Terms of Access for non-commercial research and educational use).

    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: [N/A]

    Justification: The paper does not release any new datasets or pretrained models. The contributions are an algorithm and a theoretical analysis; the experimental code is the only new asset; it is provided in the supplementary archive and addressed under question 5.

    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 involve crowdsourcing or any new collection of data from human participants. The BCI IV-2a dataset used in Appendix D is a previously collected and publicly distributed benchmark.

    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: No new human-subjects data was collected. The BCI IV-2a EEG benchmark is a publicly distributed dataset under an established release; no additional IRB approval was required for its use

    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: LLMs were not used as an important, original, or non-standard component of the core methodology. The contributions, the spectral-clipping construction, the Hessian/Jacobian-free algorithm, the convergence analysis, and the experimental protocol, were developed without LLMs as part of the research pipeline.

    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.