Your Hypergradient is Skewed: Antithetic Neumann Estimation for Bilevel Optimization

Jason Bohne[*]
Bloomberg
New York City, NY, USA
jbohne5@bloomberg.net

Paweł Polak[*†]
Stony Brook University, Stony Brook, NY, USA
Bloomberg, New York City, NY, USA
pawel.polak@stonybrook.edu

Gary Kazantsev
Bloomberg
New York City, NY, USA
gkazantsev@bloomberg.net

David Rosenberg
Bloomberg
Toronto, ON, Canada
drosenberg44@bloomberg.net

Abstract

Bilevel optimization is a core tool in practical machine learning workflows, including hyperparameter tuning, data reweighting, robust learning, reinforcement learning, and imbalanced classification. Hypergradient computation requires the inverse of the inner Hessian, typically estimated via a truncated Neumann series. While existing analyses focus on bias and variance, we show that the product structure of the Neumann series introduces a third source of error: skewness. We identify that, unlike ordinary gradients, whose asymmetry is suppressed by mini-batch averaging, the Neumann estimator amplifies skewness across factors, with the third-moment error growing, under weak contraction, with truncation depth, inner condition number, and per-factor noise asymmetry. This arises in many practical settings such as corrupted labels, outliers, heterogeneous or class-imbalanced datasets. To address this, we introduce the Antithetic Neumann Estimator (ANE) which pairs each Hessian sample with a mirrored counterpart around a reference Hessian and averages the resulting inverse-Hessian estimates. With an exact reference the antithetic product retains only even-order noise, eliminating all odd-order perturbation terms; with an approximate reference, the residual odd-order contribution is bounded linearly in the reference error. ANE requires only a modification of the Hessian–vector step and combines with existing stochastic or variance-reduced bilevel methods. We instantiate ANE-SGD and ANE-VR, the latter adding recursive variance reduction. Across robust regression, data cleaning, offline reinforcement learning, and imbalanced classification, ANE-based methods improve on stochastic and variance-reduced baselines on every task (20 seeds each), lowering loss by 4–25% relative to the best baseline; the largest gains, on robust regression, are a 25% lower test loss and a 54% higher test R^2. These results support skewness reduction as a practical, complementary mechanism for improving hypergradient quality.

[*]Jason Bohne and Paweł Polak contributed equally to this research.

[†]This research was conducted while participating in the Bloomberg Visiting Faculty Program during a sabbatical from Stony Brook University.

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

Table 1: **Neumann-based bilevel methods.** All methods inherit truncation bias from the Neumann inverse approximation. SUSTAIN applies recursive variance reduction across outer iterations. ANE modifies the within-estimate Neumann product by antithetic mirroring, which cancels odd perturbation-degree terms under an exact reference. These mechanisms act at different stages and can be combined; ANE may also affect variance because it removes the linear perturbation term.

Algorithm	Truncation bias	Recursive / across-iter. VR	Odd-degree cancellation
StocBiO [Ji et al., 2021]	✓	×	×
SUSTAIN [Khanduri et al., 2021]	✓	✓	×
ANE-SGD *(ours)*	✓	×	✓
ANE-VR *(ours)*	✓	✓	✓

1 Introduction

We study the bilevel optimization problem $\min_{\boldsymbol{\lambda}\in\mathcal{X}} F(\boldsymbol{\lambda}) \triangleq f(\boldsymbol{\lambda}, \boldsymbol{\beta}^*(\boldsymbol{\lambda}))$ with $\boldsymbol{\beta}^*(\boldsymbol{\lambda}) \in \arg\min_{\boldsymbol{\beta}} g(\boldsymbol{\lambda}, \boldsymbol{\beta})$, where $\boldsymbol{\lambda}$ denotes the outer (upper-level) variables, $\boldsymbol{\beta}^*(\boldsymbol{\lambda})$ is the optimal inner (lower-level) solution, f is the outer objective, and g is the inner objective. This formulation underlies hyperparameter optimization [Franceschi et al., 2017, Maclaurin et al., 2015, Lorraine et al., 2020], meta-learning [Finn et al., 2017, Rajeswaran et al., 2019], and data reweighting [Ren et al., 2018, Shu et al., 2019]. Gradient-based methods optimize F using the *hypergradient* $\nabla F(\boldsymbol{\lambda})$, which requires differentiating through the inner solution $\boldsymbol{\beta}^*(\boldsymbol{\lambda})$. The standard scalable approach approximates the resulting inverse Hessian via a truncated Neumann series [Ghadimi and Wang, 2018]. A substantial literature has since developed variance-reduced and oracle-efficient bilevel methods [Ji et al., 2021, Yang et al., 2021, Khanduri et al., 2021, Dagréou et al., 2022, Hong et al., 2023, Chen et al., 2024, Dagréou et al., 2024], but these works evaluate hypergradient estimators primarily through their bias and variance.

The Neumann estimator contains products of stochastic Hessian factors. In a scalar product model, asymmetric per-factor noise appears in odd perturbation-degree terms. These terms can make unnormalized third-moment proxies large when contraction is weak, as occurs along poorly conditioned directions, even though the standardized scalar skewness scales like $1/\sqrt{m}$ in the perturbative regime (Theorem 3.1). This third-order mechanism is present in Neumann-based bilevel methods such as StocBiO [Ji et al., 2021] and SUSTAIN [Khanduri et al., 2021] whenever the stochastic Hessian noise is asymmetric. Existing variance-reduced bilevel methods [Khanduri et al., 2021, Dagréou et al., 2022] reduce across-iteration variance but act at a different stage than the within-estimate Neumann product.

To address this, we introduce the *antithetic Neumann estimator* (ANE). ANE pairs each stochastic Hessian sample with a mirrored factor around a reference Hessian $\bar{H}$ and averages the two products: when $\bar{H}$ equals the true Hessian, all odd-order noise terms, including those carrying μ_3', cancel. We further embed ANE inside a recursive variance-reduction outer loop, yielding ANE-VR.

Throughout, $\gamma_1(Y) = \mathbb{E}[(Y - \mathbb{E}Y)^3]/\operatorname{Var}(Y)^{3/2}$ is signed standardized skewness for scalar Y, and $\mathcal{M}_3(\xi) = \mathbb{E}\|\xi - \mathbb{E}\xi\|^3$ is a scale-sensitive absolute third-moment proxy for vector error ξ (upper-bounding magnitudes of directional signed third moments, but itself *not* a signed skewness). The centered absolute moment can remain nonzero even for a symmetric error distribution.

Contributions.

1. We isolate a third-order mechanism in stochastic Neumann products: in a scalar model, the signed standardized skewness depends on the per-factor third central moment, while unnormalized third-moment contributions can be large under weak contraction (Theorem 3.1); a coordinate bound lifts scalar absolute third-moment effects to the vector inverse-Hessian estimate (Proposition 3.2). Under weak contraction, we further give lower bounds for the full truncated sum showing that, in this model, the vector proxy grows with truncation depth and conditioning (Lemma A.2).
2. We introduce ANE, an antithetic Neumann estimator that mirrors stochastic Hessian factors around a reference $\bar{H}$. With an exact reference and $\tau = 1$, ANE cancels all odd perturbation-degree terms in each product (Proposition 3.4); under uniform contraction, the deviation from

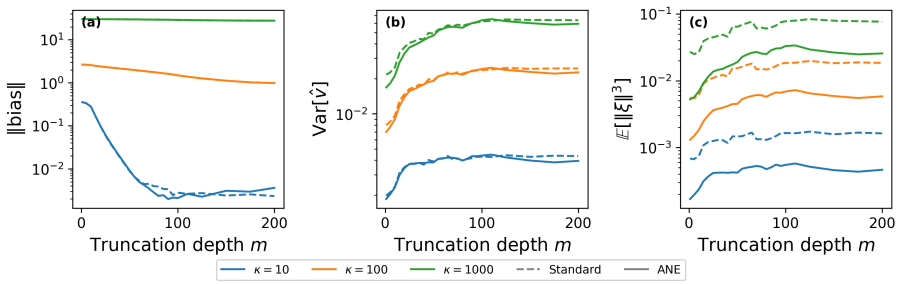


Figure 1: **Synthetic illustration of the ANE mechanism** (see the synthetic experiment in Section 5). Controlled quadratic setting $g(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^\top H \boldsymbol{\beta} - \boldsymbol{\lambda}^\top C \boldsymbol{\beta}$, prescribed $\kappa_g \in \{10, 100, 1000\}$, with stochastic Hessian–vector products (HVPs) $\hat{H}v = s_i \cdot Hv$, $\mathbb{E}[s_i] = 1$, positive third central moment ($d = 50$; dashed: Standard, solid: ANE). ANE and Standard share the same deterministic truncation target. (a) Bias: matched. (b) Variance: empirically comparable. (c) Absolute third-moment proxy $\mathbb{E}[\|\xi\|^3]$ (centered, i.e. $\mathcal{M}_3(\xi) = \mathbb{E}\|\xi - \mathbb{E}\xi\|^3$; a scale-sensitive vector diagnostic, not the signed skewness γ_1 of Theorem 3.1): lower for ANE (41–74% reduction across the swept κ_g, relative to a single standard estimate), consistent with odd perturbation-degree cancellation. The variance comparison is empirical for this setup, not a general consequence of the theory.

the exact-reference estimator is bounded linearly in the reference error (Proposition 3.6). The cancellation holds for noncommuting matrix-valued perturbations and does not require a shared eigenbasis. An iterate-level convergence bound (Theorem 4.1) shows the benefit: for Hessian error at level $\rho < 1$ at fixed depth, ANE shrinks the bound's inverse-Hessian third-moment term from $O(\rho^3)$ to $O(\rho^6)$ (Proposition A.4).

3. We instantiate ANE-SGD and ANE-VR as modifications of the inverse-Hessian–vector primitive. Recursive variance reduction and ANE act at different stages of the estimator, although ANE may also affect variance. Experiments on four bilevel tasks (robust regression, data cleaning, offline RL, imbalanced classification) show ANE-based variants improve the reported metrics, consistent with the proposed odd-order cancellation mechanism; matched synthetic controls isolate it, with no gain under symmetric noise, and runtime is measured at practical depths.

2 Background and Related Work

Assumptions. We adopt the standard assumptions for Neumann-based bilevel optimization [Ghadimi and Wang, 2018, Ji et al., 2021, Khanduri et al., 2021] (Assumptions 1–3), plus one additional assumption that goes beyond the existing literature (Assumption 4).

Assumption 1 (Smoothness)**.** *f and g are continuously differentiable; g is twice differentiable in $\boldsymbol{\beta}$. The maps ∇f, $\nabla_{\boldsymbol{\beta}} g$, $\nabla^2_{\boldsymbol{\lambda}\boldsymbol{\beta}} g$, and $\nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}} g$ are Lipschitz continuous with respective constants $\ell_{f,1}$, $\ell_{g,1}$, $\ell_{g,\boldsymbol{\lambda}\boldsymbol{\beta}}$, and $\ell_{g,2}$.*

Assumption 2 (Strong convexity)**.** *For every $\boldsymbol{\lambda} \in \mathcal{X}$, $g(\boldsymbol{\lambda}, \cdot)$ is μ_g-strongly convex. The inner condition number is $\kappa_g \triangleq \ell_{g,1}/\mu_g$.*

Assumption 3 (Bounded stochastic oracles)**.** *Stochastic gradient, Jacobian, and Hessian estimators are unbiased with uniformly bounded second moments: $\mathbb{E}[\|\widehat{H} - H\|^2] \le \sigma_H^2$ (and analogously for gradients and Jacobians with constants σ_g^2, σ_J^2).*

Assumption 4 (Higher-order regularity)**.** *(i)* Bounded third moment. *The stochastic Hessian estimator satisfies $\mathbb{E}[\|\widehat{H} - H\|^3] \le (\hat{\eta}\rho)^3$, i.e., the scaled error $E \triangleq (\widehat{H} - H)/\hat{\eta}$ has $(\mathbb{E}\|E\|^3)^{1/3} \le \rho$. The regime $\rho < 1$ is that of the standard estimator itself: with per-sample Hessians in $[0, \ell_{g,1} I]$ and the step size $\hat{\eta} > \ell_{g,1}$ of Section 2, $\|E\| \le \ell_{g,1}/\hat{\eta} < 1$. (ii)* Outer regularity. *For the iterate-level bound, assume directly that F is bounded below by $F_\star$, has L_F-Lipschitz gradient, and has L_2-Lipschitz Hessian on a domain containing the update segments. These properties are additional requirements: they do not follow from Assumptions 1–3 alone, and a Lipschitz $\nabla^2 F$ requires, for example, Lipschitz third derivatives of g.*

Assumptions 1–3 are standard [Ghadimi and Wang, 2018, Ji et al., 2021, Khanduri et al., 2021]. Assumption 4 collects the additional third-moment condition used with Theorem 4.1 (Proposition A.4;

Corollary 3.5 uses its almost-sure form) and the outer regularity conditions used in Proposition A.3. The analysis takes $\hat{\eta} > \ell_{g,1}$; the experiments use the practical choice $\hat{\eta} = \hat{L}_{\max}/2$ (Table 4), for which $\rho < 1$ is not guaranteed, with safeguards keeping the recursions bounded (Appendix C.7). Some algebraic results below impose additional local assumptions stated with each result: the scalar perturbative expansion (Theorem 3.1) requires a finite sixth moment; Corollary 3.5 assumes an almost-sure operator-norm bound on the Hessian perturbations; Proposition 3.6 assumes uniform contraction of both mirrored branches.

Under Assumptions 1–2, $\boldsymbol{\beta}^*(\boldsymbol{\lambda})$ is unique and the implicit hypergradient is $\nabla F(\boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}} f - \nabla^2_{\boldsymbol{\lambda}\boldsymbol{\beta}} g \left(\nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}} g\right)^{-1} \nabla_{\boldsymbol{\beta}} f$, evaluated at $(\boldsymbol{\lambda}, \boldsymbol{\beta}^*(\boldsymbol{\lambda}))$ [Ghadimi and Wang, 2018]. The bottleneck is the inverse-Hessian–vector product $(\nabla^2_{\boldsymbol{\beta}\boldsymbol{\beta}} g)^{-1} \nabla_{\boldsymbol{\beta}} f$, approximated via a truncated Neumann series. Setting $\hat{A}_j \triangleq I - \widehat{H}_j/\hat{\eta}$ for i.i.d. Hessian samples and $\hat{\eta} > \ell_{g,1}$:

$$\widehat{v}_{\mathrm{Neu}} = \frac{1}{\hat{\eta}} \sum_{k=0}^{m-1} \prod_{j=1}^{k} \widehat{A}_j \, b, \qquad \widetilde{\nabla} F(\boldsymbol{\lambda}) = \widehat{\nabla}_{\boldsymbol{\lambda}} f - \widehat{\nabla}^2_{\boldsymbol{\lambda}\boldsymbol{\beta}} g \, \widehat{v}_{\mathrm{Neu}}. \tag{1}$$

Truncation depth $m = O(\kappa_g \log(1/\epsilon))$ controls bias to accuracy ϵ [Ghadimi and Wang, 2018]. This stochastic hypergradient is shared by StocBiO [Ji et al., 2021] and SUSTAIN [Khanduri et al., 2021].

Bilevel optimization. Neumann-based implicit differentiation for bilevel optimization was introduced by Ghadimi and Wang [2018], with practical applications in hyperparameter optimization [Pedregosa, 2016, Maclaurin et al., 2015, Lorraine et al., 2020], meta-learning [Finn et al., 2017, Rajeswaran et al., 2019], and data reweighting [Ren et al., 2018, Shu et al., 2019]. Subsequent work developed single-loop, two-timescale, and variance-reduced algorithms with improved oracle complexity [Ji et al., 2021, Yang et al., 2021, Khanduri et al., 2021, Dagréou et al., 2022, Hong et al., 2023, Arbel and Mairal, 2022], including near-optimal lower bounds [Chen et al., 2024, Dagréou et al., 2024]. These analyses characterize bias and variance of the hypergradient estimator.

Variance reduction. Recursive variance reduction via SARAH [Nguyen et al., 2017], SPIDER [Fang et al., 2018], PAGE [Li et al., 2021], and momentum-based methods [Cutkosky and Orabona, 2019] has been widely adopted in bilevel optimization [Khanduri et al., 2021, Dagréou et al., 2022]. These methods operate on the hypergradient *after* the Neumann product has been formed, reducing the across-iteration variance.

Antithetic variates and adjacent work. Classical antithetic variates [Hammersley and Handscomb, 1964] pair scalar or vector samples to cancel the linear noise term. Alternative inverse approximations [Prakash et al., 2025] and heavy-tailed bilevel methods [Liu and Luo, 2025, Zhang et al., 2025] address different aspects of robustness. Online bilevel formulations [Tarzanagh et al., 2024, Bohne et al., 2024, 2026] also share the Neumann primitive.

3 Skewness in Neumann Hypergradients and Its Reduction

A mini-batch gradient is an average of i.i.d. contributions, so γ_1 decreases with batch size. The Neumann inverse approximation is structurally different: it is a *sum of products* of random Hessian factors. Those products do not admit term-wise averaging; their third-order behavior is governed by multiplicative interactions between perturbations. We work with the truncated-sum Neumann estimator $\hat{v}_{\mathrm{Neu}}$ of (1). We build the analysis bottom-up: a scalar product model isolating the multiplicative mechanism (Section 3.1), its coordinate-wise lift (Section 3.2), a decomposition of the full hypergradient error (Section 3.3), and finally the antithetic construction that cancels odd perturbation-degree terms (Section 3.4).

3.1 Scalar Product Model

As a stylized model, suppose the stochastic Hessian acts multiplicatively along an eigendirection v of H: for i.i.d. scalar multipliers s_j with $\mathbb{E}[s_j] = 1$, $\hat{H}_j v = s_j H v$. Then each Neumann factor shares the eigenbasis of H, and along eigendirection i with eigenvalue μ_i the factor reduces to a scalar variable $X_j^{(i)} = 1 - s_j \mu_i/\hat{\eta} = a_i + \delta_j^{(i)}$, with $a_i \triangleq 1 - \mu_i/\hat{\eta}$. This motivates the following scalar product model.

Theorem 3.1 (Scalar product third moments)**.** *Let $X_1, \dots, X_m$ be i.i.d. copies of $X = a + \delta$ with $0 < |a| < 1$, $\mathbb{E}\delta = 0$, $\mathbb{E}\delta^2 = \sigma^2$, $\mathbb{E}\delta^3 = \mu_3'$, and $\mathbb{E}|\delta|^6 < \infty$, and define $P_m \triangleq \prod_{j=1}^m X_j$. Then:*

(i) $\mathbb{E}[P_m] = a^m$ *and* $\mathrm{Var}(P_m) = (a^2 + \sigma^2)^m - a^{2m}$.

(ii) The third central moment is

$$\mathbb{E}[(P_m - a^m)^3] = (a^3 + 3a\sigma^2 + \mu_3')^m - 3(a^2 + \sigma^2)^m a^m + 2a^{3m}. \tag{2}$$

(iii) In the perturbative regime $m\sigma^2/a^2 = o(1)$ with $|a|$ bounded away from zero and $|\mu_3'| \le C\sigma^3$, $\mathrm{Var}(P_m) = m\sigma^2 a^{2m-2} + O(m^2\sigma^4|a|^{2m-4})$ and $\mathbb{E}[(P_m - a^m)^3] = m\mu_3' a^{3m-3} + O(m^2\sigma^4|a|^{3m-4})$, and therefore the signed standardized skewness satisfies $\gamma_1(P_m) = \mathrm{sgn}(a)^{m-1}\mu_3'/(\sigma^3\sqrt{m}) + O(\sqrt{m}\,\sigma)$. For $a > 0$ the sign factor is unity, recovering the positive-factor formula.

The scalar model shows that odd perturbation-degree terms are proportional to the per-factor third central moment. Their unnormalized contribution $m\mu_3' a^{3m-3}$ is most pronounced under weak contraction, when $|a|$ is close to one, whereas the signed standardized skewness scales as $1/\sqrt{m}$ in the perturbative regime; for fixed $|a| < 1$, the unnormalized contribution decays exponentially in m.

3.2 From Scalar to Vector

The Neumann estimator error is vector-valued. In the eigenbasis of H, each product term decomposes into scalar factors of the form analyzed above. The scalar formulas require the multiplicative eigendirection model; the odd-degree cancellation and reference-sensitivity results below apply to general matrix-valued noise and do not require a shared eigenbasis.

Proposition 3.2 (Coordinate contribution to vector absolute third moments)**.** *Under the multiplicative eigendirection model, let the depth-k product error along eigendirection i be $\xi_i^{(k)} \triangleq (b_i/\hat{\eta})(P_k^{(i)} - a_i^k)$, with $P_k^{(i)} \triangleq \prod_{j=1}^k (a_i + \delta_j^{(i)})$ and b_i the ith coordinate of b in the eigenbasis of H. Then*

$$\mathbb{E}[\|\xi^{(k)}\|^3] \ge \max_i \frac{|b_i|^3}{\hat{\eta}^3} \mathbb{E}[|P_k^{(i)} - a_i^k|^3],$$

so a large absolute third centered moment in any eigendirection lower-bounds the vector absolute third-moment proxy.

The proposition is a lower bound on $\mathcal{M}_3(\xi^{(k)})$: it shows that one-dimensional absolute third-moment effects survive in the multivariate proxy. It does not assert monotonic growth in k, nor does it bound signed vector skewness. Appendix A.6 lower-bounds the full truncated sum: under weak contraction, $\mathcal{M}_3$ grows at least as $\sigma^3 m^{9/2}$ for either sign of the factor-noise third moment, and as $\kappa_g^{3/2}$ along the smallest eigenvalue when $m = \Theta(\kappa_g)$; when that third moment μ_3' is nonnegative, the signed third moment grows at least as $\mu_3' m^4$.

3.3 Full Hypergradient Error: Where the m-Dependence Enters

Let $G \triangleq \nabla_{\boldsymbol{\lambda}} f - Jv^\star$ with $v^\star \triangleq H^{-1}\nabla_{\boldsymbol{\beta}} f$ and $J \triangleq \nabla^2_{\boldsymbol{\lambda}\boldsymbol{\beta}} g$, and define the stochastic hypergradient $\widehat{G} \triangleq \widehat{\nabla}_{\boldsymbol{\lambda}} f - \widehat{J}\hat{v}$, where $\hat{v}$ is any stochastic inverse-Hessian–vector estimator (standard Neumann or ANE).

Proposition 3.3 (Full-hypergradient error decomposition)**.** *The stochastic hypergradient error admits the exact decomposition*

$$\widehat{G} - G = (\widehat{\nabla}_{\boldsymbol{\lambda}} f - \nabla_{\boldsymbol{\lambda}} f) - (\widehat{J} - J)\hat{v} - J(\hat{v} - v^\star). \tag{3}$$

Consequently,

$$\|\widehat{G} - G\|^3 \le 9\|\widehat{\nabla}_{\boldsymbol{\lambda}} f - \nabla_{\boldsymbol{\lambda}} f\|^3 + 9\|(\widehat{J} - J)\hat{v}\|^3 + 9\|J\|^3\|\hat{v} - v^\star\|^3. \tag{4}$$

In particular, the explicit Neumann product and truncation-depth dependence enter through the inverse-Hessian–vector estimator $\hat{v}$; the outer-gradient and Jacobian oracles do not themselves contain Neumann products.

This motivates targeting the Neumann product: although other stochastic oracle terms may contribute to the full hypergradient error, the product expansion and its truncation-depth dependence arise inside $\hat{v}$. We use the Neumann-only absolute third-moment proxy

$$\mathcal{M}_3(\xi) \triangleq \mathbb{E}[\|\xi - \mathbb{E}\xi\|^3], \qquad \xi \triangleq \hat{v}_{\mathrm{Neu}} - v^\star, \tag{5}$$

as a scale-sensitive diagnostic. It upper-bounds the magnitude of every directional signed third moment since $|\mathbb{E}[(u^\top(\xi - \mathbb{E}\xi))^3]| \le \mathbb{E}[\|\xi - \mathbb{E}\xi\|^3]$ for every unit u, but $\mathcal{M}_3$ is itself *not* a signed skewness and can be nonzero for symmetric distributions.

3.4 The Antithetic Neumann Estimator

For each stochastic Hessian $\hat{H}_j$, ANE constructs a mirror around a reference Hessian $\bar{H}$. Let $\tau \in [0, 1]$ be a mirror strength and define

$$\hat{H}_j^+ \triangleq \hat{H}_j, \qquad \hat{H}_j^- \triangleq \bar{H} - \tau(\hat{H}_j - \bar{H}). \tag{6}$$

Setting $\hat{A}_j^\pm \triangleq I - \hat{H}_j^\pm/\hat{\eta}$, the antithetic Neumann estimator (ANE) is

$$\hat{v}_{\mathrm{ANE}} \triangleq \frac{1}{2\hat{\eta}} \sum_{k=0}^{m-1} \left[\prod_{j=1}^{k} \hat{A}_j^+ + \prod_{j=1}^{k} \hat{A}_j^- \right] b. \tag{7}$$

With an exact reference and $\tau = 1$, the average of forward and mirrored products cancels every term containing an odd number of perturbation factors (algebraic cancellation in the product expansion; this does not imply zero signed or absolute third moments overall). For $\tau < 1$ the linear term is not fully canceled, with residual $\propto 1 - \tau$ in the scalar expansion. The full ANE hypergradient replaces $\hat{v}_{\mathrm{Neu}}$ by $\hat{v}_{\mathrm{ANE}}$; in practice $\bar{H}$ can be maintained by periodic snapshots on a larger batch (four times the HVP batch in our experiments). The reference can be represented by its Hessian–vector-product operator, so a dense $p \times p$ matrix need not be stored.

Why does this cancel odd-degree terms? With two scalar factors, $\frac{1}{2}[(a + \delta_1)(a + \delta_2) + (a - \delta_1)(a - \delta_2)] = a^2 + \delta_1\delta_2$, so the linear terms cancel. The same mechanism holds at arbitrary depth.

Proposition 3.4 (Odd perturbation-degree cancellation)**.** *Assume $\bar{H} = H$ and $\tau = 1$. Define $A \triangleq I - H/\hat{\eta}$ and $E_j \triangleq (\hat{H}_j - H)/\hat{\eta}$, so that $\hat{A}_j^+ = A - E_j$ and $\hat{A}_j^- = A + E_j$. Then at each depth k,*

$$\frac{1}{2}\left(\prod_{j=1}^{k}(A - E_j) + \prod_{j=1}^{k}(A + E_j) \right) = \sum_{\substack{S \subseteq [k] \\ |S| \text{ even}}} \Pi_k(S), \tag{8}$$

where $\Pi_k(S)$ is the ordered product with factor E_j for $j \in S$ and factor A otherwise. All terms with an odd number of perturbation factors vanish.

Corollary 3.5 (Quadratic-in-ρ perturbation bound)**.** *Assume the setting of Proposition 3.4 and that $\|A\| \le a < 1$ and $\|E_j\| \le \rho$ almost surely for all j (the almost-sure form of Assumption 4(i)). Then for small ρ,*

$$\begin{aligned} \|\hat{v}_{\mathrm{ANE}} - \mathcal{N}_m(H; b)\| \le \frac{\|b\|\rho^2}{2\hat{\eta}} \sum_{k=0}^{m-1} k(k-1)a^{k-2} \\ + O\left(\frac{\|b\|\rho^4}{\hat{\eta}} \sum_{k=0}^{m-1} k^4 a^{k-4} \right), \end{aligned} \tag{9}$$

where $\mathcal{N}_m(H; b) = \hat{\eta}^{-1} \sum_{k=0}^{m-1} A^k b$. Consequently, if $a \le a_0 < 1$ is bounded away from one, then $\|\hat{v}_{\mathrm{ANE}} - \mathcal{N}_m(H; b)\| = O(\rho^2\|b\|/[\hat{\eta}(1 - a_0)^3])$ uniformly in m. In the weak-contraction finite-depth regime $a \approx 1$, the quadratic term scales at most as $O(m^3\rho^2\|b\|/\hat{\eta})$. In both cases the leading perturbation order in ρ is quadratic, whereas the standard truncated Neumann operator has a generically linear perturbation term.

Thus ANE removes the leading odd perturbation-degree terms when the reference is exact. This is the algebraic mechanism behind the smaller inverse-Hessian third-moment bound of Proposition A.4.

Proposition 3.6 (Sensitivity to approximate reference)**.** *The following is a local stability bound; it assumes both the exact-reference and approximate-reference mirrored factors are uniformly contractive (this is stronger than the moment assumptions on stochastic Hessian noise). Let $R = \bar{H} - H$ and $B = 2R/\hat{\eta}$. Assume $\tau = 1$ and $\|A + E_j\|, \|A - B + E_j\| \le c < 1$ for all j. Then*

$$\|\hat{v}_{\mathrm{ANE}}(\bar{H}) - \hat{v}_{\mathrm{ANE}}(H)\| \le \frac{\|B\|\|b\|}{2\hat{\eta}} \sum_{k=1}^{m-1} k c^{k-1} \le \frac{\|B\|\|b\|}{2\hat{\eta}(1-c)^2}. \tag{10}$$

An approximate reference shifts the mean of the mirrored branch ($\mathbb{E}[\hat{H}_j^-] = H + 2R$ when $\bar{H} = H + R$ and $\tau = 1$), so Proposition 3.6 is a sensitivity bound: under uniform contraction, the approximate-reference estimator differs from the exact-reference estimator by $O(\|R\|)$, controlling residual odd-degree terms and bias from a stale reference.

Remark 3.7 (Practical stabilization)**.** *When $2\bar{H} - \hat{H}_j$ becomes poorly conditioned, ANE can be stabilized by damped mirroring ($\tau < 1$), periodic refreshes of $\bar{H}$, or clipping in operator norm or Hessian–vector-product norm. These are stabilization choices that trade exact cancellation for robustness: the algebraic odd-degree cancellation in Proposition 3.4 applies exactly only when $\tau = 1$ and the reference is exact. With an exact reference, a degree-n perturbation term has coefficient $[1 + (-\tau)^n]/2$ relative to Standard. Thus odd terms retain $(1 - \tau^n)/2$: $\tau = 0.5$ leaves one quarter of the leading linear coefficient, while $\tau = 1$ removes it. At $\tau = 0$, ANE averages the Standard and reference branches rather than reducing to Standard.*

Relation to variance reduction. Although ANE is motivated by odd-degree cancellation, the same algebra also removes the linear perturbation term in the product expansion under an exact reference; ANE can therefore affect variance as well as third moments. Its complementarity with recursive variance reduction is algorithmic: ANE modifies the within-estimate Neumann product, whereas recursive variance reduction reduces across-iteration fluctuations of the hypergradient estimator.

4 Algorithms

ANE-SGD. The simplest use of ANE is as a drop-in replacement for the standard Neumann inverse approximation inside a bilevel method such as StocBiO [Ji et al., 2021]. We call the resulting method **ANE-SGD**. The inner solver, outer step size, and the surrounding single-loop structure are left unchanged; the only modification is the inverse-Hessian–vector estimate, where each stochastic Hessian factor is paired with its mirror around the current reference $\bar{H}_t$. Comparing a standard Neumann implementation against ANE-SGD isolates the effect of the antithetic estimator without the additional confound of recursive outer variance reduction.

ANE-VR. ANE targets the within-estimate asymmetry of the Neumann inverse approximation, while recursive variance reduction targets the across-iteration fluctuation of the full hypergradient. Since these affect different parts of the estimation pipeline, they compose naturally. Replacing the Neumann inverse approximation inside a PAGE-type recursive wrapper [Li et al., 2021] yields **ANE-VR** (Algorithm 3). Full algorithms appear in Appendix B; Algorithm 1 gives the compact PAGE variant. ANE-SGD is the simpler default, while ANE-VR is useful in an existing recursive-VR pipeline or when substantial across-iteration variance remains.

Computational cost. ANE evaluates forward and mirrored branches and also applies the reference operator. In a matrix-free implementation, a mirrored HVP uses $\hat{H}^- v = (1 + \tau)\bar{H}v - \tau\hat{H}v$ at the mirrored branch vector, which generally differs from the forward vector, so each Neumann factor costs three HVP-type operations (one stochastic HVP per branch and one reference HVP) against one for Standard. Refreshing the reference every $q_{\mathrm{ref}} = 10$ steps only resamples its batch (four times the HVP batch) and is negligible; applying it at every mirrored factor is included in the measured hypergradient cost (Appendix C.8). Storage can remain linear in parameter count; computational scaling also depends on architecture and reference-batch size.

Algorithm 1 ANE-VR (compact): ANE-SGD with recursive outer variance reduction

Input: Same as Algorithm 2, plus snapshot probability p.

Initialize: $\boldsymbol{\lambda}_0$, $\boldsymbol{\beta}_0$, $\bar{H}_0$, and $d_{-1} = 0$.

for $t = 0, 1, \ldots, T-1$ **do**

1. Steps 1–3 of Algorithm 2 $\to \widehat{G}_t^{\mathrm{ANE}}$.
2. **Variance reduction (PAGE).**
   (a) At $t = 0$ or with probability p: $d_t \leftarrow \widehat{G}_t^{\mathrm{ANE}}$ (snapshot from the current mini-batch).
   (b) Otherwise: $d_t \leftarrow d_{t-1} + \widehat{G}_t^{\mathrm{ANE}} - \widehat{G}^{\mathrm{ANE}}(\boldsymbol{\lambda}_{t-1}, \boldsymbol{\beta}_{t-1})$, where the second term is recomputed at the previous iterate on fresh samples with the current reference.
3. **Outer step.** $\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}_t - \alpha\, d_t$.

end for

Theorem 4.1 (Iterate-level guarantee)**.** *Let F be bounded below by $F_\star$, with L_F-Lipschitz gradient and L_2-Lipschitz Hessian, and let $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \alpha d_t$ with $0 < \alpha \le 1/(4L_F)$. Let $\mathcal{F}_t$ contain $\boldsymbol{\lambda}_t$ and all randomness preceding d_t, write $e_t = d_t - \mathbb{E}[d_t \mid \mathcal{F}_t]$ and $y_t = \boldsymbol{\lambda}_t - \alpha\mathbb{E}[d_t \mid \mathcal{F}_t]$, and assume that, almost surely for all t,*

$$\|\mathbb{E}[d_t \mid \mathcal{F}_t] - \nabla F(\boldsymbol{\lambda}_t)\|^2 \le \bar{r}^2, \qquad \mathbb{E}[\|e_t\|^2 \mid \mathcal{F}_t] \le \bar{\sigma}^2,$$
$$\mathbb{E}[\|e_t\|^3 \mid \mathcal{F}_t] \le \bar{M}_3, \qquad \mathbb{E}[e_t^\top \nabla^2 F(y_t) e_t \mid \mathcal{F}_t] \le \bar{Q}.$$

Then

$$\frac{1}{T}\sum_{t<T} \mathbb{E}\|\nabla F(\boldsymbol{\lambda}_t)\|^2 \le \frac{2(F(\boldsymbol{\lambda}_0) - F_\star)}{\alpha T} + \frac{5}{2}\bar{r}^2 + \alpha\bar{Q} + \frac{L_2\alpha^2}{3}\bar{M}_3. \tag{11}$$

The proof is in Appendix A.7. Since $\nabla^2 F \preceq L_F I$, taking $\bar{Q} = L_F\bar{\sigma}^2$ is always valid and gives the standard first-order variance term $L_F\alpha\bar{\sigma}^2$; the cubic term is the price of using the exact curvature, which pays off when the fluctuations e_t lie along low-curvature directions ($\bar{Q} < L_F\bar{\sigma}^2$). For $d_t = \widehat{G}$ with the Hessian error at level ρ (Assumption 4(i)), the decomposition of Proposition 3.3, applied to the centered error, splits $\bar{M}_3$ into an oracle part $\bar{M}_{\mathrm{orc}}$ from the outer-gradient and Jacobian noise, bounded alike for both estimators (Appendix A.8), and the inverse-Hessian part $9\|J\|^3\mathbb{E}\|\hat{v} - \mathcal{N}_m(H; b)\|^3$. At fixed depth, with an exact reference and $\tau = 1$ (Proposition A.4),

$$\bar{M}_3 \le \bar{M}_{\mathrm{orc}} + \begin{cases} O(\rho^3) & \text{standard Neumann,} \\ O(\rho^6) & \text{ANE.} \end{cases}$$

Before its constants are collected, ANE's bound keeps only the even-order terms of the standard bound, so it is never larger (Appendix A.8).

5 Experiments

Claims. All experiments compare StocBiO [Ji et al., 2021] and SUSTAIN [Khanduri et al., 2021] as stochastic bilevel baselines against our ANE methods. In particular, we compare SUSTAIN against ANE-VR, since both methods include variance reduction. We verify two claims: *(C1) The Neumann product exhibits nonzero odd-order perturbative contributions whenever the per-factor Hessian noise has nonzero third central moment μ_3'. Although standardized scalar skewness scales as $m^{-1/2}$ in the leading perturbative approximation, the scale-sensitive third-moment diagnostic $\mathcal{M}_3(\xi) = \mathbb{E}[\|\xi - \mathbb{E}\xi\|^3]$ can grow with truncation depth m and conditioning κ_g* (Theorem 3.1, Proposition 3.2). Lemma A.2 proves this growth for the multiplicative model under weak contraction. *(C2) Antithetic cancellation reduces the absolute third-moment proxy vs. a single standard Neumann estimate; this is associated with improved predictive metrics on the reported tasks.*

Synthetic. We consider a synthetic bilevel problem with inner quadratic $g(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^\top H\boldsymbol{\beta} - \boldsymbol{\lambda}^\top C\boldsymbol{\beta}$, prescribed κ_g, and stochastic HVPs $\hat{H}v = s_i \cdot Hv$ with $\mathbb{E}[s_i] = 1$ and positive third central moment $\mathbb{E}[(s_i - 1)^3] > 0$ (right-skewed). Sweeping $\kappa_g \in \{10, 100, 1000\}$ and $m \in \{1, \ldots, 200\}$

Table 2: **Summary of bilevel tasks.** The asymmetry source column identifies the origin of nonzero per-factor μ_3' in the Neumann product for each task. Appendix C.9 motivates this for each task.

Task	Problem	Dataset	Asymmetry source
Housing	Robust regr.	California Housing [Pace and Barry, 1997]	20% corrupted targets
Fashion	Data cleaning	Fashion-MNIST [Xiao et al., 2017]	10% of positives flipped
Acrobot	Offline RL	Acrobot-v1 [Brockman et al., 2016]	Bimodal episode lengths
Fraud	Imbalanced classif.	creditcard [Dal Pozzolo et al., 2015]	6% minority class

at $d = 50$, Figure 1 is consistent with the mechanism isolated by Theorem 3.1 and Proposition 3.4: under asymmetric scalar Hessian noise and weak contraction, the absolute third-moment proxy is lower for ANE than for a single standard Neumann estimate (41–74% reduction across the swept κ_g); the theorem explains the odd perturbation-degree terms while the plotted proxy is an empirical absolute third-moment diagnostic.

Matched controls. Matched synthetic controls ($\kappa_g = 100$, $m = 20$, eight seeds) isolate the asymmetry effect. With distributional mirroring, in which the mirrored branch draws its own multiplier, ANE matches Standard (2x), which averages two estimates, in variance and stochastic-HVP budget, yet has a much smaller coordinate skewness (0.42 ± 0.43 versus 1.30 ± 0.13); reference-only gives 1.87 ± 0.19. In a fixed-variance sweep the two are close for symmetric multipliers and separate as the multipliers become skewed (Appendix C.5).

Real tasks. We evaluate ANE-SGD/ANE-VR on four bilevel problems (Table 2); Table 3 reports final metrics and Figure 2 convergence trajectories (mean $\pm$ SD over 20 seeds). σ_i'' is per-sample loss curvature: when $\sigma_i'' \approx 0$ for some samples and $\gg 0$ for others, HVP factors $(I - \sigma_i'' J_i^\top J_i / \hat{\eta})$ have asymmetric noise and odd-degree contributions accumulate across Neumann factors (Theorem 3.1; per-task derivations and hyperparameters in Appendix C).

Task 1 (Housing). Pseudo-Huber regression on California Housing [Pace and Barry, 1997] with 20% corrupted targets; outer learns per-feature regularization on clean validation. Corrupted targets give $\sigma_i'' \approx 0$ for $\sim$20% of samples (worse conditioning, amplified odd-degree contributions, Theorem 3.1). ANE reaches $R^2 = 0.60$ vs. 0.39 ($+54\%$); loss $0.147 \to 0.110$ (-25%, Figure 2a).

Task 2 (Fashion-MNIST). MLP/softplus with one-directional label corruption (10% of positive labels flipped) [Xiao et al., 2017]; outer tunes per-feature regularization on clean validation. $\sigma_i'' \approx 0.25$ for corrupted vs. ≈ 0 for clean (asymmetric noise). ANE-SGD/ANE-VR reach 89.6%/89.5% vs. 87.5%/87.6% ($\sim$2 pp), test loss -11% (Figure 2b).

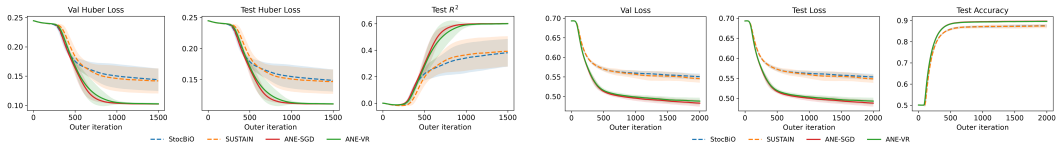
Task 3 (Acrobot). MLP value function on Acrobot-v1 [Brockman et al., 2016]; offline dataset 50% expert ($\sim$136 steps) / 50% random ($\sim$500 steps). Random episodes are $\sim$80% of transitions but $\sigma_i'' \approx 0$, leaving $\sim$21% effective curvature. ANE-SGD/ANE-VR reach $R^2 = 0.20$ vs. 0.16 (SUSTAIN), 0.14 (StocBiO); loss 0.311 vs. 0.325 (SUSTAIN; -4%, Figure 2c).

Task 4 (Fraud). MLP on credit-card transactions [Dal Pozzolo et al., 2015]; outer uses class-balanced validation. Boundary fraud ($\sigma_i'' \approx 0.25$) vs. high-margin legitimate ($\sigma_i'' \approx 0$): strong noise asymmetry. ANE-SGD/ANE-VR cut outer loss to $0.281/0.301$ vs. 0.339 (SUSTAIN; 11–17%), $+0.5$–0.6 pp AUPRC (vs. StocBiO) (Figure 2d).

Ablation and cost. An ablation over $\tau \in \{0.5, 0.7, 0.9, 1.0\}$ (Appendix C.4) shows that, with 20 seeds per value, all four mirror strengths give nearly identical predictive metrics (differences of at most 0.004 in loss, R^2, or AUPRC and 0.1 pp in accuracy), all well separated from the baselines. The hypergradient-step sweep in Appendix C.8, measured with each task's configuration, reports 2.7–3.3$\times$ ANE-SGD overhead across $m = 5, 10, 20, 50$.

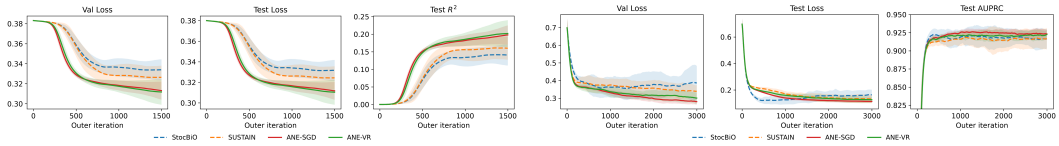
6 Conclusion and Limitations

The product structure of the Neumann series introduces a third source of error beyond bias and variance, skewness: the third-moment error grows with truncation depth and inner condition number under weak contraction (Lemma A.2), and with per-factor noise asymmetry (Theorem 3.1). ANE pairs each Hessian sample with a mirrored counterpart around a reference Hessian. With an exact refer-



(a) California Housing.

(b) Fashion-MNIST.



(c) Acrobot.

(d) Fraud detection.

Figure 2: **Convergence on bilevel tasks.** Each panel: validation (outer) loss, test loss, and the task-specific test metric (mean $\pm$ SD over 20 seeds) vs. outer iteration. Dashed: baselines (StocBiO, SUSTAIN); solid: ANE. On Housing, Fashion, and Acrobot, ANE-SGD and ANE-VR separate from baselines early and maintain a consistent gap; on Fraud, StocBiO has the lowest test loss mid-training and the AUPRC bands overlap. Appendix D enlarges each panel; curves are smoothed with an exponential moving average.

Table 3: **Final test metrics (mean $\pm$ SD over 20 seeds per task).** For Fraud, Loss is the class-balanced validation (outer) objective (Appendix C.3). Best per task in **bold**. In Acrobot, SUSTAIN beats StocBiO ($R^2 = 0.16$ vs. 0.14) but ANE-VR matches ANE-SGD, consistent with across-iteration VR adding little on top of ANE.

Task	Metric	StocBiO	SUSTAIN	ANE-SGD	ANE-VR
Housing	Loss $\downarrow$	$.147 \pm .018$	$.147 \pm .020$	$\mathbf{.110 \pm .001}$	$\mathbf{.111 \pm .001}$
	$R^2 \uparrow$	$.388 \pm .104$	$.393 \pm .116$	$\mathbf{.604 \pm .007}$	$\mathbf{.602 \pm .007}$
Fashion	Loss $\downarrow$	$.553 \pm .008$	$.548 \pm .011$	$\mathbf{.487 \pm .009}$	$\mathbf{.492 \pm .009}$
	Acc $\uparrow$	87.5 ± 0.8	87.6 ± 1.3	$\mathbf{89.6 \pm 0.3}$	$\mathbf{89.5 \pm 0.3}$
Acrobot	Loss $\downarrow$	$.332 \pm .011$	$.325 \pm .012$	$\mathbf{.311 \pm .010}$	$\mathbf{.310 \pm .013}$
	$R^2 \uparrow$	$.140 \pm .031$	$.159 \pm .034$	$\mathbf{.200 \pm .032}$	$\mathbf{.203 \pm .040}$
Fraud	Loss $\downarrow$	$.380 \pm .089$	$.339 \pm .035$	$\mathbf{.281 \pm .017}$	$\mathbf{.301 \pm .033}$
	AUPRC $\uparrow$	$.918 \pm .012$	$.916 \pm .017$	$\mathbf{.924 \pm .006}$	$\mathbf{.923 \pm .009}$

ence and $\tau = 1$ only even-order noise remains, so at fixed depth the inverse-Hessian third-moment term in the iterate-level bound (Theorem 4.1) is $O(\rho^6)$ in the Hessian-error level $\rho < 1$, versus $O(\rho^3)$ for Standard; with an approximate one, under uniform contraction, the odd-order residual is bounded linearly in the reference error. ANE modifies only the Hessian–vector step; ANE-SGD and ANE-VR lower loss by 4–25% relative to the best stochastic or variance-reduced baseline on every task (20 seeds). Skewness reduction is thus a promising, complementary mechanism for improving hypergradient quality. *Limitations:* odd-order cancellation is not a full hypergradient skewness theory; the approximate-reference bound assumes a contraction the real tasks lack (Appendix C.7), so the gains are not attributable to skewness correction alone. ANE costs about three times StocBiO's hypergradient step (Appendix C.8); matched-wall-clock evaluation is future work.

Acknowledgments

We sincerely thank Bloomberg's CTO Office for its support of the Visiting Faculty Program, through which Paweł Polak participated in this research.

References

M. Arbel and J. Mairal. Amortized implicit differentiation for stochastic bilevel optimization. In *International Conference on Learning Representations (ICLR)*, 2022.

J. Bohne, D. S. Rosenberg, G. Kazantsev, and P. Polak. Bregman geometry for stochastic online bilevel optimization. In *OPT 2024: Optimization for Machine Learning (NeurIPS Workshop)*, 2024.

J. Bohne, I. Petrulionyte, M. Arbel, J. Mairal, and P. Polak. Non-stationary functional bilevel optimization. *arXiv preprint arXiv:2601.15363*, 2026.

G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016.

L. Chen, J. Xu, and J. Zhang. On finding small hyper-gradients in bilevel optimization: Hardness results and improved analysis. In *Conference on Learning Theory (COLT)*, 2024.

A. Cutkosky and F. Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

M. Dagréou, P. Ablin, S. Vaiter, and T. Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

M. Dagréou, T. Moreau, S. Vaiter, and P. Ablin. A lower bound and a near-optimal algorithm for bilevel empirical risk minimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi. Calibrating probability with undersampling for unbalanced classification. In *IEEE Symposium Series on Computational Intelligence*, 2015.

C. Fang, C. J. Li, Z. Lin, and T. Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

L. Franceschi, M. Donini, P. Frasconi, and M. Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning (ICML)*, 2017.

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

J. M. Hammersley and D. C. Handscomb. *Monte Carlo Methods*. Methuen, 1964.

M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A two-timescale stochastic algorithm framework for bilevel optimization. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

K. Ji, J. Yang, and Y. Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning (ICML)*, 2021.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Z. Li, H. Bao, X. Zhang, and P. Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning (ICML)*, 2021.

Z. Liu and L. Luo. Stochastic bilevel optimization with heavy-tailed noise. *arXiv preprint arXiv:2509.14952*, 2025.

J. Lorraine, P. Vicol, and D. K. Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

D. Maclaurin, D. Duvenaud, and R. Adams. Gradient-based hyperparameter optimization through reversible learning. In *International Conference on Machine Learning (ICML)*, 2015.

L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning (ICML)*, 2017.

R. K. Pace and R. Barry. Sparse spatial autoregressions. *Statistics and Probability Letters*, 33(3): 291–297, 1997.

F. Pedregosa. Hyperparameter optimization with approximate gradient. In *International Conference on Machine Learning (ICML)*, 2016.

A. Prakash, N. He, D. Goktas, and A. Greenwald. Bi-level policy optimization with Nyström hypergradients. *arXiv preprint arXiv:2505.11714*, 2025.

A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

M. Ren, W. Zeng, B. Yang, and R. Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning (ICML)*, 2018.

J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, and D. Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

D. A. Tarzanagh, P. Nazari, B. Hou, L. Shen, and L. Balzano. Online bilevel optimization: Regret analysis of online alternating gradient methods. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

J. Yang, K. Ji, and Y. Liang. Provably faster algorithms for bilevel optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

X. Zhang, Y. Zhang, and H. Gao. Nonconvex decentralized stochastic bilevel optimization under heavy-tailed noises. *arXiv preprint arXiv:2509.15543*, 2025.

Appendix roadmap. Appendix A gives estimator proofs, the full-sum lower bounds, the conditional-moment iterate bound, and small-noise moment scaling. Appendix B records detailed algorithms. Appendix C provides hyperparameters, controls, runtime measurements, and diagnostic tables. Appendix D shows the 20-seed convergence trajectories.

A Skewness Analysis Proofs

A.1 Proof of Theorem 3.1

Proof. Let $X_j = a + \delta_j$ with δ_j i.i.d., $\mathbb{E}[\delta_j] = 0$, $\mathbb{E}[\delta_j^2] = \sigma^2$, and $\mathbb{E}[\delta_j^3] = \mu_3'$. Define

$$P_m \triangleq \prod_{j=1}^{m} X_j.$$

Part (i). By independence,

$$\mathbb{E}[P_m] = \prod_{j=1}^{m} \mathbb{E}[X_j] = a^m. \tag{12}$$

Also,

$$\mathbb{E}[P_m^2] = \prod_{j=1}^{m} \mathbb{E}[X_j^2] = (a^2 + \sigma^2)^m, \tag{13}$$

so

$$\mathrm{Var}(P_m) = (a^2 + \sigma^2)^m - a^{2m}. \tag{14}$$

Part (ii). Since

$$\mathbb{E}[X_j^3] = a^3 + 3a\sigma^2 + \mu_3',$$

we have

$$\mathbb{E}[P_m^3] = (a^3 + 3a\sigma^2 + \mu_3')^m. \tag{15}$$

Using the identity $\mathbb{E}[(Y - \mathbb{E}Y)^3] = \mathbb{E}[Y^3] - 3\mathbb{E}[Y^2]\mathbb{E}[Y] + 2(\mathbb{E}[Y])^3$ with $Y = P_m$ yields

$$\mathbb{E}[(P_m - a^m)^3] = \mathbb{E}[P_m^3] - 3\mathbb{E}[P_m^2]\mathbb{E}[P_m] + 2(\mathbb{E}[P_m])^3 \tag{16}$$

$$= (a^3 + 3a\sigma^2 + \mu_3')^m - 3(a^2 + \sigma^2)^m a^m + 2a^{3m}, \tag{17}$$

which is (2).

Part (iii). For the perturbative expansion we write $r \triangleq \sigma^2/a^2$ and $s \triangleq \mu_3'/a^3$. If the Neumann scale is chosen so that all scalar factors are positive, then $a \in (0, 1)$. More generally, the following moment expansions hold for $0 < |a| < 1$; the denominator in the standardized skewness uses $|a|$, which introduces the factor $\mathrm{sgn}(a)^{m-1}$. Then

$$\mathrm{Var}(P_m) = a^{2m}\big[(1 + r)^m - 1\big]. \tag{18}$$

Using $(1 + r)^m = 1 + mr + \binom{m}{2}r^2 + O(m^3 r^3)$ we get

$$\mathrm{Var}(P_m) = m\sigma^2 a^{2m-2} + O(m^2\sigma^4|a|^{2m-4}),$$

which is the variance expansion of Theorem 3.1(iii). Here $m\sigma^2/a^2 = o(1)$ implies $m^3 r^3 = o(m^2 r^2)$.

For the third central moment, (2) rewrites as $\mathbb{E}[(P_m - a^m)^3] = a^{3m}[(1+3r+s)^m - 3(1+r)^m + 2]$ and expanding gives

$$(1 + 3r + s)^m - 3(1 + r)^m + 2 = ms + \binom{m}{2}(6r^2 + 6rs + s^2) + O\big(m^3(r^3 + r^2|s| + rs^2 + |s|^3)\big).$$

Therefore

$$\mathbb{E}[(P_m - a^m)^3] = m\mu_3' a^{3m-3} + O(m^2\sigma^4|a|^{3m-4}) + O(m^2\sigma^2|\mu_3'||a|^{3m-5}) + O(m^2\mu_3'^2|a|^{3m-6}).$$

Under fixed a bounded away from zero and sufficiently small σ, with $|\mu_3'| \le C\sigma^3$, the last two error terms scale as $O(m^2\sigma^5|a|^{3m-5})$ and $O(m^2\sigma^6|a|^{3m-6})$ respectively and are dominated by $O(m^2\sigma^4|a|^{3m-4})$. Making the absolute value explicit when expanding the denominator,

$$\mathrm{Var}(P_m)^{3/2} = m^{3/2}\sigma^3|a|^{3m-3}\big(1 + O(m\sigma^2/a^2)\big),$$

so

$$\gamma_1(P_m) = \frac{m\mu_3' a^{3m-3}}{m^{3/2}\sigma^3|a|^{3m-3}} + O(\sqrt{m}\,\sigma) = \mathrm{sgn}(a)^{m-1}\frac{\mu_3'}{\sigma^3\sqrt{m}} + O(\sqrt{m}\,\sigma),$$

which is the standardized skewness expansion of Theorem 3.1(iii). □

A.2 Explicit next-order skewness expansion

Remark A.1 (Explicit next-order standardized-skewness expansion)**.** *With $r \triangleq \sigma^2/a^2$ and $s \triangleq \mu_3'/a^3$, the variance and third central moment of P_m admit the explicit binomial expansions*

$$\frac{\mathrm{Var}(P_m)}{a^{2m}} = mr + \binom{m}{2} r^2 + O(m^3 r^3), \tag{19}$$

$$\frac{\mathbb{E}[(P_m - a^m)^3]}{a^{3m}} = ms + \binom{m}{2}(6r^2 + 6rs + s^2) + O\big(m^3(r^3 + r^2|s| + rs^2 + |s|^3)\big). \tag{20}$$

Consequently, for $a > 0$, the standardized skewness expands as

$$\gamma_1(P_m) = \frac{\mu_3'}{\sigma^3\sqrt{m}} + O\left(\sqrt{m}\,\frac{\sigma}{|a|}\right) + O\left(\sqrt{m}\,\frac{|\mu_3'|}{a^2\sigma}\right). \tag{21}$$

In the bounded-shape perturbative regime where $|a|$ is bounded away from zero, $|\mu_3'| \le C\sigma^3$, and $m\sigma^2/a^2 = o(1)$, the last two error terms collapse to $O(\sqrt{m}\,\sigma)$ and (21) *reduces to the form stated in Theorem 3.1(iii).*

Proof. With $r = \sigma^2/a^2$ and $s = \mu_3'/a^3$, the exact identities from Theorem 3.1 become

$$\frac{\mathrm{Var}(P_m)}{a^{2m}} = (1+r)^m - 1, \tag{22}$$

$$\frac{\mathbb{E}[(P_m - a^m)^3]}{a^{3m}} = (1 + 3r + s)^m - 3(1+r)^m + 2. \tag{23}$$

Applying the binomial expansion to the variance term yields

$$\frac{\mathrm{Var}(P_m)}{a^{2m}} = mr + \binom{m}{2} r^2 + O(m^3 r^3), \tag{24}$$

which is (19). For the numerator,

$$(1 + 3r + s)^m = 1 + m(3r + s) + \binom{m}{2}(3r + s)^2 \tag{25}$$

$$+ O\big(m^3(|r| + |s|)^3\big), \tag{26}$$

and therefore

$$\frac{\mathbb{E}[(P_m - a^m)^3]}{a^{3m}} = \Big[1 + m(3r + s) + \binom{m}{2}(3r + s)^2\Big] \tag{27}$$

$$- 3\Big[1 + mr + \binom{m}{2} r^2\Big] + 2 \tag{28}$$

$$+ O\big(m^3(r^3 + r^2|s| + rs^2 + |s|^3)\big) \tag{29}$$

$$= ms + \binom{m}{2}(6r^2 + 6rs + s^2) \tag{30}$$

$$+ O\big(m^3(r^3 + r^2|s| + rs^2 + |s|^3)\big), \tag{31}$$

which is (20). Using

$$(mr + \binom{m}{2} r^2 + O(m^3 r^3))^{3/2} = (mr)^{3/2}\big(1 + O(mr)\big),$$

we obtain

$$\gamma_1(P_m) = \frac{ms + O\big(m^2(r^2 + r|s| + s^2)\big)}{(mr)^{3/2}\big(1 + O(mr)\big)} \tag{32}$$

$$= \frac{s}{r^{3/2}\sqrt{m}} + O\left(\frac{\sqrt{m}\,r^{1/2}}{1}\right) + O\left(\frac{\sqrt{m}\,|s|}{r^{1/2}}\right). \tag{33}$$

Returning to (σ, μ_3', a) with $a > 0$ gives

$$\gamma_1(P_m) = \frac{\mu_3'}{\sigma^3\sqrt{m}} + O\left(\sqrt{m}\,\frac{\sigma}{|a|}\right) + O\left(\sqrt{m}\,\frac{|\mu_3'|}{a^2\sigma}\right), \tag{34}$$

which is (21). If $|a|$ is bounded away from zero and $|\mu_3'| \le C\sigma^3$, the last two terms are both $O(\sqrt{m}\,\sigma)$. □

A.3 Proof of Proposition 3.4 and Corollary 3.5

Proof of Proposition 3.4. Fix a depth k. By distributivity,

$$P_k^{\pm} = \prod_{j=1}^{k}(A \pm E_j) = \sum_{S\subseteq[k]} (\pm 1)^{|S|}\, \Pi_k(S), \tag{35}$$

where $\Pi_k(S)$ is the ordered product with factor E_j when $j \in S$ and factor A otherwise. Averaging the two expansions gives

$$\frac{P_k^{+} + P_k^{-}}{2} = \sum_{S\subseteq[k]} \frac{1+(-1)^{|S|}}{2}\, \Pi_k(S) = \sum_{\substack{S\subseteq[k] \\ |S| \text{ even}}} \Pi_k(S), \tag{36}$$

which is exactly (8). □

Proof of Corollary 3.5. For each depth k, Proposition 3.4 shows that $\frac{1}{2}(P_k^{+} + P_k^{-}) - A^k$ contains only even-order perturbation terms. Every ordered product with exactly ℓ perturbation factors has norm at most $a^{k-\ell}\rho^{\ell}$, and there are $\binom{k}{\ell}$ ways to choose the locations of those perturbation factors. Therefore,

$$\left\| \frac{P_k^{+} + P_k^{-}}{2} - A^k \right\| \le \sum_{\substack{\ell=2 \\ \ell \text{ even}}}^{k} \binom{k}{\ell} a^{k-\ell}\rho^{\ell}. \tag{37}$$

Using the elementary identity

$$\sum_{\substack{\ell=0 \\ \ell \text{ even}}}^{k} \binom{k}{\ell} a^{k-\ell}\rho^{\ell} = \frac{(a+\rho)^k + (a-\rho)^k}{2}, \tag{38}$$

we obtain

$$\left\| \frac{P_k^{+} + P_k^{-}}{2} - A^k \right\| \le \frac{(a+\rho)^k + (a-\rho)^k - 2a^k}{2}. \tag{39}$$

Summing over $k = 0, \dots, m-1$ and multiplying by $\|b\|/\hat{\eta}$ yields (9).

For the asymptotic form, expand the right-hand side in ρ around 0:

$$(a+\rho)^k + (a-\rho)^k - 2a^k = k(k-1)a^{k-2}\rho^2 + O(\rho^4).$$

For fixed $a \le a_0 < 1$,

$$\sum_{k=0}^{\infty} k(k-1)a^{k-2} = \frac{2}{(1-a)^3} \le \frac{2}{(1-a_0)^3},$$

so the quadratic term is uniformly

$$O\!\left(\frac{\rho^2\|b\|}{\hat{\eta}(1-a_0)^3}\right).$$

If no gap from one is imposed, then using $a^{k-2} \le 1$ gives $\sum_{k=0}^{m-1} k(k-1) = O(m^3)$, so the finite-depth weak-contraction bound is $O(m^3\rho^2\|b\|/\hat{\eta})$. This proves the corollary. □

A.4 Proof of Proposition 3.6

Proof. Only the mirrored branch changes when H is replaced by $\bar{H}$. For a fixed depth k, define

$$M_j \triangleq A + E_j, \qquad \widetilde{M}_j \triangleq A - B + E_j. \tag{40}$$

Then the noncommutative telescoping identity gives

$$\prod_{j=1}^{k} \widetilde{M}_j - \prod_{j=1}^{k} M_j = \sum_{r=1}^{k} \left(\prod_{j=1}^{r-1} \widetilde{M}_j\right) (\widetilde{M}_r - M_r) \left(\prod_{j=r+1}^{k} M_j\right). \tag{41}$$

Since $\widetilde{M}_r - M_r = -B$ and each factor norm is at most c, we obtain

$$\left\| \prod_{j=1}^{k} \widetilde{M}_j - \prod_{j=1}^{k} M_j \right\| \leq k c^{k-1} \|B\|. \tag{42}$$

Summing over $k = 1, \ldots, m-1$ and including the prefactor $1/(2\hat{\eta})$ from the average of the two branches yields

$$\left\| \hat{v}_{\mathrm{ANE}}(\bar{H}) - \hat{v}_{\mathrm{ANE}}(H) \right\| \leq \frac{\|B\| \, \|b\|}{2\hat{\eta}} \sum_{k=1}^{m-1} k c^{k-1}. \tag{43}$$

Finally, because $c < 1$,

$$\sum_{k=1}^{m-1} k c^{k-1} \leq \sum_{k=1}^{\infty} k c^{k-1} = \frac{1}{(1-c)^2}, \tag{44}$$

which proves (10). □

A.5 Proof of Proposition 3.3

Proof. Write

$$G = \nabla_{\boldsymbol{\lambda}} f - J v^\star, \qquad \widehat{G} = \widehat{\nabla}_{\boldsymbol{\lambda}} f - \widehat{J}\, \hat{v}.$$

Add and subtract $J\hat{v}$:

$$\widehat{G} - G = \left(\widehat{\nabla}_{\boldsymbol{\lambda}} f - \nabla_{\boldsymbol{\lambda}} f\right) - \widehat{J}\, \hat{v} + J v^\star \tag{45}$$

$$= \left(\widehat{\nabla}_{\boldsymbol{\lambda}} f - \nabla_{\boldsymbol{\lambda}} f\right) - (\widehat{J} - J)\hat{v} - J(\hat{v} - v^\star), \tag{46}$$

which is (3).

For the cubic norm bound, define

$$x \triangleq \widehat{\nabla}_{\boldsymbol{\lambda}} f - \nabla_{\boldsymbol{\lambda}} f, \qquad y \triangleq -(\widehat{J} - J)\hat{v}, \qquad z \triangleq -J(\hat{v} - v^\star).$$

Then

$$\|\widehat{G} - G\|^3 = \|x + y + z\|^3 \leq (\|x\| + \|y\| + \|z\|)^3.$$

For nonnegative numbers a, b, c, Hölder's inequality implies

$$(a + b + c)^3 \leq 3^2 (a^3 + b^3 + c^3) = 9(a^3 + b^3 + c^3).$$

Applying this with $a = \|x\|$, $b = \|y\|$, and $c = \|z\|$ yields

$$\|\widehat{G} - G\|^3 \leq 9\|x\|^3 + 9\|y\|^3 + 9\|z\|^3, \tag{47}$$

that is,

$$\|\widehat{G} - G\|^3 \leq 9\left\|\widehat{\nabla}_{\boldsymbol{\lambda}} f - \nabla_{\boldsymbol{\lambda}} f\right\|^3 + 9\left\|(\widehat{J} - J)\hat{v}\right\|^3 + 9\|J\|^3 \left\|\hat{v} - v^\star\right\|^3, \tag{48}$$

which is (4). The final sentence of the proposition follows because the explicit truncation-depth parameter m appears only through the inverse-Hessian–vector estimator $\hat{v}$. □

A.6 Lower bounds for the full truncated sum

Claim (C1) in Section 5 states that the third-moment proxy $\mathcal{M}_3$ can grow with truncation depth m and conditioning κ_g; this subsection proves it for the full truncated Neumann sum in the multiplicative model under weak contraction. The growth is compatible with the $m^{-1/2}$ decay of standardized skewness in Theorem 3.1, which concerns a single product.

Lemma A.2 (Full-sum lower bounds)**.** *Let $0 < a < 1$ and let $\delta_1, \ldots, \delta_{m-1}$ be i.i.d. with $\mathbb{E}\delta_j = 0$, $\mathbb{E}\delta_j^2 = \sigma^2$, $\mathbb{E}|\delta_j|^3 < \infty$, and $\nu_3 \triangleq \mathbb{E}\delta_j^3$ (the μ_3' of Theorem 3.1). Let $P_k = \prod_{j=1}^{k}(a + \delta_j)$ and $S_m = \sum_{k=1}^{m-1}(P_k - a^k)$. For $m \geq 2$,*

$$\mathbb{E}|S_m|^3 \geq \left(\mathbb{E}S_m^2\right)^{3/2} \geq \sigma^3 a^{3m-6} \left[\frac{(m-1)m(2m-1)}{6}\right]^{3/2}, \tag{49}$$

and, if $\nu_3 \geq 0$,

$$\mathbb{E}|S_m|^3 \geq \mathbb{E}S_m^3 \geq \nu_3 \sum_{k,l,n=1}^{m-1} a^{k+l+n-3} \min(k,l,n) \geq \nu_3 a^{3m-6} \left[\frac{m(m-1)}{2}\right]^2. \tag{50}$$

If a coordinate of the Neumann error satisfies $\xi_i - \mathbb{E}\xi_i = (b_i/\hat{\eta})S_m$, then $\mathcal{M}_3(\xi) \geq (|b_i|/\hat{\eta})^3 \mathbb{E}|S_m|^3$. If also $a \geq 1/2$ and $m(1-a) \leq C$, the two bounds are $\Omega(\sigma^3 m^{9/2})$ and $\Omega(\nu_3 m^4)$, with constants depending only on C.

Proof. The first inequality in (49) is Lyapunov's. For $k \leq l$, $P_l = P_k \prod_{j=k+1}^{l}(a+\delta_j)$, where the product is independent of P_k and has mean a^{l-k}, so $\mathrm{Cov}(P_k, P_l) = a^{l-k}[(a^2+\sigma^2)^k - a^{2k}] \geq k\sigma^2 a^{k+l-2}$. Summing over k, l,

$$\mathbb{E}S_m^2 \geq \sigma^2 a^{2m-4} \sum_{k,l=1}^{m-1} \min(k,l) = \sigma^2 a^{2m-4} \frac{(m-1)m(2m-1)}{6}.$$

For (50), $P_k - a^k = \sum_{\emptyset \neq U \subseteq \{1,\ldots,k\}} a^{k-|U|} \prod_{j \in U} \delta_j$ has nonnegative coefficients. In $\mathbb{E}S_m^3$, monomials in which some index appears once vanish, and every other index contributes σ^2 or $\nu_3 \geq 0$, so all terms are nonnegative. Keeping, for each $j \leq \min(k,l,n)$, the product of the three terms linear in δ_j gives $\nu_3 a^{k+l+n-3} \min(k,l,n)$, and $\sum_{k,l,n} \min(k,l,n) = \sum_{j=1}^{m-1}(m-j)^3 = [m(m-1)/2]^2$. The vector bound follows from $\|\xi - \mathbb{E}\xi\| \geq |\xi_i - \mathbb{E}\xi_i|$, and the rates from $a^{3m-6} \geq e^{-6C}$, since $\log a \geq -2(1-a)$ for $a \geq 1/2$. □

Sign and conditioning. In the multiplicative model $\delta_j = -(\mu_i/\hat{\eta})(s_j - 1)$, so $\nu_3 = -(\mu_i/\hat{\eta})^3 \mathbb{E}[(s_j-1)^3]$: right-skewed multipliers, as in Figure 1, give $\nu_3 < 0$, and only (49) applies. Along the smallest eigenvalue, with b_i and $\hat{\eta} \propto \ell_{g,1}$ fixed, $1 - a = \Theta(1/\kappa_g)$ and $\sigma \propto \kappa_g^{-1}$; choosing $m = \Theta(\kappa_g)$ keeps $m(1-a)$ bounded, so the bound in (49) grows as $\kappa_g^{3/2}$ and, when $\nu_3 \geq 0$, the bound in (50) grows as κ_g.

A.7 An iterate-level bound with a third-order remainder

This subsection proves the curvature-weighted bound of Theorem 4.1 as Proposition A.3, which also records the first-order special case. The proposition carries the conditional bias, curvature-weighted variance, and centered third moment of the update direction d_t (the ANE-SGD estimate, or the recursive ANE-VR direction provided its carried error satisfies the bias bound) through to a stationarity bound on the outer iterates. The proof expands F to second order: the third moment enters through the cubic Taylor remainder, and keeping the exact curvature gives the curvature-weighted variance.

Proposition A.3 (Conditional-moment stationarity bound)**.** *Let $F : \mathbb{R}^p \to \mathbb{R}$ be bounded below by $F_\star$, with L_F-Lipschitz gradient and L_2-Lipschitz Hessian, and let $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \alpha d_t$ with $0 < \alpha \leq 1/(4L_F)$. Let $\mathcal{F}_t$ contain $\boldsymbol{\lambda}_t$ and all randomness preceding d_t; write $r_t = \mathbb{E}[d_t \mid \mathcal{F}_t] - \nabla F(\boldsymbol{\lambda}_t)$, $e_t = d_t - \mathbb{E}[d_t \mid \mathcal{F}_t]$, and $y_t = \boldsymbol{\lambda}_t - \alpha \mathbb{E}[d_t \mid \mathcal{F}_t]$; and suppose that, almost surely and uniformly in t, $\|r_t\|^2 \leq \bar{r}^2$, $\mathbb{E}[\|e_t\|^2 \mid \mathcal{F}_t] \leq \bar{\sigma}^2$, $\mathbb{E}[\|e_t\|^3 \mid \mathcal{F}_t] \leq \bar{M}_3$, and $\mathbb{E}[e_t^\top \nabla^2 F(y_t) e_t \mid \mathcal{F}_t] \leq \bar{Q}$. Then*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(\boldsymbol{\lambda}_t)\|^2 \leq \frac{2(F(\boldsymbol{\lambda}_0) - F_\star)}{\alpha T} + \frac{5}{2}\bar{r}^2 + \alpha\bar{Q} + \frac{L_2\alpha^2}{3}\bar{M}_3. \tag{51}$$

Since $\nabla^2 F \preceq L_F I$, one may take $\bar{Q} = L_F\bar{\sigma}^2$, which gives the first-order bound:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(\boldsymbol{\lambda}_t)\|^2 \leq \frac{2(F(\boldsymbol{\lambda}_0) - F_\star)}{\alpha T} + \frac{5}{2}\bar{r}^2 + L_F\alpha\bar{\sigma}^2 + \frac{L_2\alpha^2}{3}\bar{M}_3. \tag{52}$$

Proof. Condition on $\mathcal{F}_t$ and abbreviate $x = \boldsymbol{\lambda}_t$, $g = \nabla F(x)$, $h = g + r_t$, and $y = y_t = x - \alpha h$. By L_F-smoothness, $-\langle g, r_t\rangle \leq \|g\|^2/4 + \|r_t\|^2$, $\|g + r_t\|^2 \leq 2\|g\|^2 + 2\|r_t\|^2$, and $L_F\alpha \leq 1/4$,

$$F(y) \leq F(x) - \alpha\langle g, h\rangle + \tfrac{1}{2}L_F\alpha^2\|h\|^2 \leq F(x) - \tfrac{\alpha}{2}\|g\|^2 + \tfrac{5}{4}\alpha\|r_t\|^2.$$

The second-order Taylor bound about y gives

$$F(y-\alpha e_t) \le F(y) - \alpha\langle \nabla F(y), e_t\rangle + \frac{\alpha^2}{2} e_t^\top \nabla^2 F(y) e_t + \frac{L_2\alpha^3}{6}\|e_t\|^3.$$

Since y is $\mathcal{F}_t$-measurable, the linear term has zero conditional mean and the quadratic term has conditional mean at most $\frac{\alpha^2}{2}\bar{Q}$. Combining the two inequalities, taking expectations, summing over t, dividing by $\alpha T/2$, and using $\mathbb{E}F(\boldsymbol{\lambda}_T) \ge F_\star$ proves (51). □

A.8 Hessian-error level and centered third moments

This subsection bounds the inverse-Hessian component of the third-moment term $\bar{M}_3$ in Theorem 4.1. For independent factor perturbations at Hessian-error level $\rho \le 1$ (Assumption 4(i)), with an exact reference and $\tau = 1$, ANE's centered third moment is $O(\rho^6)$ at every depth, versus $O(\rho^3)$ for the standard Neumann estimator; the constants are uniform in m under a contraction margin.

Proposition A.4 (Third-moment scaling in the Hessian-error level)**.** *Fix m, A, b, and $\hat{\eta}$ with $\|A\| \le a < 1$, and assume an exact reference and $\tau = 1$. Let the scaled errors $E_j = (\hat{H}_j - H)/\hat{\eta}$ be independent with $\mathbb{E}E_j = 0$ and $(\mathbb{E}\|E_j\|^3)^{1/3} \le \rho \le 1$. Then*

$$\mathbb{E}\|\hat{v}_{\mathrm{Neu}} - \mathcal{N}_m(H;b)\|^3 \le \Big(\frac{K_1\|b\|}{\hat{\eta}}\Big)^3\rho^3, \qquad \mathbb{E}\|\hat{v}_{\mathrm{ANE}} - \mathcal{N}_m(H;b)\|^3 \le \Big(\frac{K_2\|b\|}{\hat{\eta}}\Big)^3\rho^6,$$

with $K_1 = \sum_{k=1}^{m-1} k(1+a)^{k-1}$ and $K_2 = \sum_{k=2}^{m-1}\binom{k}{2}(1+a)^{k-2}$; if moreover $a+\rho \le c < 1$, then K_1 and K_2 can be replaced by $(1-c)^{-2}$ and $(1-c)^{-3}$, uniformly in m. Both estimators have mean $\mathcal{N}_m(H;b)$, so these are their centered third moments.

Proof. For an ordered monomial with ℓ noise factors, submultiplicativity and independence give its L^3 norm at most $a^{k-\ell}\rho^\ell$. Minkowski's inequality applied to the ordered expansions gives

$$\big(\mathbb{E}\|\hat{v}_{\mathrm{Neu}} - \mathcal{N}_m\|^3\big)^{1/3} \le \frac{\|b\|}{\hat{\eta}}\sum_{k=0}^{m-1}\sum_{\ell=1}^{k}\binom{k}{\ell}a^{k-\ell}\rho^\ell,$$

$$\big(\mathbb{E}\|\hat{v}_{\mathrm{ANE}} - \mathcal{N}_m\|^3\big)^{1/3} \le \frac{\|b\|}{\hat{\eta}}\sum_{k=0}^{m-1}\sum_{\substack{\ell=2\\ \ell\ \mathrm{even}}}^{k}\binom{k}{\ell}a^{k-\ell}\rho^\ell.$$

Since $\rho \le 1$, $(a+\rho)^k - a^k \le k\rho(1+a)^{k-1}$ bounds the first sum by ρK_1. In the second, $\binom{k}{\ell} \le \binom{k}{2}\binom{k-2}{\ell-2}$ for $\ell \ge 2$ bounds each depth by $\rho^2\binom{k}{2}(a+\rho)^{k-2} \le \rho^2\binom{k}{2}(1+a)^{k-2}$, hence the sum by $\rho^2 K_2$; with $a+\rho \le c < 1$ the same steps give $\rho/(1-c)^2$ and $\rho^2/(1-c)^3$. The second sum keeps only the even-ℓ terms of the first, so it is no larger for every ρ. Including the $\ell = 0$ terms, the same expansion gives $(\mathbb{E}\|\hat{v}\|^3)^{1/3} \le \frac{\|b\|}{\hat{\eta}}\sum_{k=0}^{m-1}(a+\rho)^k$ for both estimators, so the Jacobian term of $\bar{M}_3$, which involves $\hat{v}$, is bounded alike. Every nonconstant ordered monomial contains each E_j at most once, so it has zero expectation by independence and centering, proving the mean identity. □

B Detailed Algorithms

Algorithm 2 ANE-SGD: antithetic Neumann hypergradients without recursive variance reduction

Input: outer steps T, inner steps K, truncation depth m, outer step size α, inner step size η, Neumann scale $\hat{\eta}$, mirror strength $\tau \in [0, 1]$, reference-refresh period q_{ref}.

Initialize: $\boldsymbol{\lambda}_0$, $\boldsymbol{\beta}_0$, reference Hessian $\bar{H}_0$.

for $t = 0, 1, \ldots, T-1$ **do**

1. **Inner solve.** Starting from $\boldsymbol{\beta}_t$, run K stochastic-gradient steps on $g(\boldsymbol{\lambda}_t, \cdot)$ with step size η to obtain $\boldsymbol{\beta}_{t+1}$.
2. **Reference refresh.** If $t \equiv 0 \pmod{q_{\mathrm{ref}}}$, update $\bar{H}_t$ using a large-batch or full-batch Hessian evaluation at $(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1})$; otherwise reuse the previous reference (or update it by an exponential moving average). The experiments use a reference batch four times the HVP batch, applied as an HVP operator.
3. **ANE hypergradient.** Draw stochastic Hessians $\widehat{H}_{t,1}, \ldots, \widehat{H}_{t,m-1}$. For each j, set
$$\widehat{H}^{+}_{t,j} = \widehat{H}_{t,j}, \qquad \widehat{H}^{-}_{t,j} = \bar{H}_t - \tau(\widehat{H}_{t,j} - \bar{H}_t).$$
Form
$$\hat{v}_{\mathrm{ANE}} = \frac{1}{2\hat{\eta}} \sum_{k=0}^{m-1} \left[\prod_{j=1}^{k} \left(I - \widehat{H}^{+}_{t,j}/\hat{\eta} \right) + \prod_{j=1}^{k} \left(I - \widehat{H}^{-}_{t,j}/\hat{\eta} \right) \right] b_t,$$
where $b_t = \widehat{\nabla}_{\boldsymbol{\beta}} f(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1})$. Then set
$$\widehat{G}_t \leftarrow \widehat{\nabla}_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}) - \widehat{J}_t \, \hat{v}_{\mathrm{ANE}}, \qquad \widehat{J}_t = \widehat{\nabla}^2_{\boldsymbol{\lambda}\boldsymbol{\beta}} g(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}).$$
4. **Outer step.** Update $\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}_t - \alpha \widehat{G}_t$ (or its projected/proximal variant).

end for

Algorithm 3 ANE-VR: antithetic Neumann hypergradients with recursive outer variance reduction

Input: outer steps T, inner steps K, truncation depth m, outer step size α, inner step size η, Neumann scale $\hat{\eta}$, mirror strength $\tau \in [0, 1]$, reference-refresh period q_{ref}, snapshot probability p.

Initialize: $\boldsymbol{\lambda}_0$, $\boldsymbol{\beta}_0$, reference Hessian $\bar{H}_0$, direction $d_{-1} = 0$.

for $t = 0, 1, \ldots, T-1$ **do**

1. **Inner solve.** Starting from $\boldsymbol{\beta}_t$, run K stochastic-gradient steps on $g(\boldsymbol{\lambda}_t, \cdot)$ to obtain $\boldsymbol{\beta}_{t+1}$.
2. **Reference refresh.** If $t \equiv 0 \pmod{q_{\mathrm{ref}}}$, refresh $\bar{H}_t$ using a large batch or a full batch at $(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1})$; otherwise reuse or exponentially average the previous reference.
3. **ANE hypergradient.** Form $\widehat{G}^{\mathrm{ANE}}_t$ exactly as in Algorithm 2, but with the mini-batch choice used by the outer variance-reduction wrapper.
4. **Recursive variance reduction.** At $t = 0$ and otherwise with probability p, set $d_t \leftarrow \widehat{G}^{\mathrm{ANE}}_t$. Otherwise, set $d_t \leftarrow d_{t-1} + \widehat{G}^{\mathrm{ANE}}_t - \widehat{G}^{\mathrm{ANE}}(\boldsymbol{\lambda}_{t-1}, \boldsymbol{\beta}_{t-1})$.
5. **Outer step.** Update $\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}_t - \alpha d_t$ (or its projected/proximal variant).

end for

C Extended Evaluation Details

This appendix records full experimental details for Section 5.

C.1 Estimator Diagnostics

Given L independent hypergradient estimates $\{\widehat{G}^{(\ell)}\}_{\ell=1}^{L}$ with mean $\bar{G}$, the primary estimator diagnostic is the empirical third-moment error:

$$\widehat{\mathcal{M}}_3(\widehat{G} - G) = \frac{1}{L}\sum_{\ell=1}^{L} \|\widehat{G}^{(\ell)} - \bar{G}\|^3. \tag{53}$$

This is a conservative third-moment proxy that upper-bounds directional signed third central moments: $|\mathbb{E}[(u^\top(\xi - \mathbb{E}\xi))^3]| \le \mathbb{E}[\|\xi - \mathbb{E}\xi\|^3]$ for any unit u. It measures centered fluctuation scale and need not vanish for symmetric errors. Signed third moments and standardized skewness require a fixed projection u: $Y^{(\ell)} = u^\top \widehat{G}^{(\ell)}$, $\hat{\mu}_3 = L^{-1}\sum_\ell (Y^{(\ell)} - \bar{Y})^3$, and $\hat{\gamma}_1 = \hat{\mu}_3/[L^{-1}\sum_\ell (Y^{(\ell)} - \bar{Y})^2]^{3/2}$. Bias requires a reference G, and vector variance is $L^{-1}\sum_\ell \|\widehat{G}^{(\ell)} - \bar{G}\|^2$. The raw error moment $L^{-1}\sum_\ell \|\widehat{G}^{(\ell)} - G\|^3$ is distinct from the centered proxy.

C.2 Synthetic Benchmark

The inner objective is $g(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^\top H \boldsymbol{\beta} - \boldsymbol{\lambda}^\top C \boldsymbol{\beta}$ with $H = \mathrm{diag}(\mu_g, \ldots, \ell_{g,1})$ having prescribed condition number $\kappa_g = \ell_{g,1}/\mu_g$, with $\ell_{g,1}$ fixed at $L_{\max} = 10$ and $\mu_g = L_{\max}/\kappa_g$. The coupling matrix C is random orthogonal. Stochastic HVPs use $s_i \cdot Hv$ where s_i are i.i.d. with $\mathbb{E}[s_i] = 1$ and positive third central moment. Dimension $d = 50$, 5 seeds, $L = 1000$ estimator samples per configuration. We sweep $\kappa_g \in \{10, 100, 1000\}$ and $m \in \{1, \ldots, 200\}$. In Figure 1 the mirrored branch draws its own multiplier (distributional mirroring; Section 5). The real tasks mirror the sampled mini-batch Hessian itself, as in (6).

C.3 Real Task Hyperparameters

Table 4: Full hyperparameter settings for each real task.

Parameter	Task 1 (Housing)	Task 2 (Fashion)	Task 3 (Acrobot)	Task 4 (Fraud)
K (inner steps)	10	50	10	10
η (inner lr)	0.01	0.01	0.01	0.01
α (outer lr)	0.01	0.001	0.01	0.01
Outer optimizer	Adam	Adam	Adam	Adam
m (truncation)	50	20	50	10
$\hat{\eta}$	$\hat{L}_{\max}/2$	$\hat{L}_{\max}/2$	$\hat{L}_{\max}/2$	$\hat{L}_{\max}/2$
Batch size (g, f)	8	32	8	8
Batch size (H)	4	8	4	4
T (outer iters)	1500	2000	1500	3000
Seeds	20	20	20	20
Gradient clip	1000	100	1000	1000

ANE uses $\tau = 1$ in all tasks, with a safeguard that damps τ for an individual factor only when the mirrored step would be unstable (the damped mirroring of Remark 3.7). ANE also caps the mirrored perturbation $\|\hat{H}v - \bar{H}v\|$ at $5\|v\|$, and both ANE and the standard Neumann estimator rescale a branch state whose norm exceeds $100\|b\|$ (the branch-norm guard). $\hat{L}_{\max}$ is a power-iteration estimate of the largest eigenvalue of the full-training inner Hessian near initialization. The reference is an HVP operator over a batch four times the HVP batch, drawn at the refresh point and refreshed every $q_{\mathrm{ref}} = 10$ steps; a dense Hessian is never formed. Adam is the outer optimizer in these experiments, so Proposition A.3 is a theoretical result for a separate constant-step gradient update.

Models and data. Housing uses a pseudo-Huber loss with $\delta = 0.5$ and an MLP with 32 hidden units on 8 PCA components of the inputs. We corrupt 20% of the training targets via $y \mapsto -y + 3\,\mathrm{sign}(y)$; validation and test targets are clean. Fashion separates classes 0–4 (positive) from 5–9 with a softplus loss, per-feature regularization, and an MLP with 16 hidden units on 50 PCA components. It has $10{,}000$ training samples, with 10% of their positive labels flipped to negative. Acrobot fits an MLP $9 \to 32 \to 1$ to return-to-go targets from 500 episodes, half expert and half random, using a pseudo-Huber loss with $\delta = 1$. Fraud trains an MLP $30 \to 32 \to 1$ on $4{,}500$ legitimate and 295 fraud transactions with per-parameter regularization. Its outer objective, a class-balanced validation loss, is also the reported loss. The data come from sklearn (California Housing), torchvision (Fashion-MNIST), Kaggle (credit-card fraud), and Gymnasium (Acrobot-v1). Housing and Acrobot split the samples (transitions) at random into $60/20/20\%$ train/validation/test, so Acrobot transitions from one episode can fall in several splits; Fashion holds out $2{,}500$ validation samples and uses the official test set; Fraud splits the fraud cases $60/20/20\%$ and adds $1{,}000$ legitimate validation transactions, with a test set of the same class balance. The Acrobot expert is an energy-pumping heuristic policy and the random policy samples actions uniformly.

C.4 Mirror Strength Ablation

Table 5 ablates $\tau \in \{0.5, 0.7, 0.9, 1.0\}$, where $\tau = 1$ is the a priori default. Smaller τ damps the mirror perturbation; at zero it averages Standard with a deterministic reference branch. All columns use 20 seeds (mean $\pm$ SD). The predictive metrics vary little with τ relative to the ANE–baseline gaps; on Fraud, $\tau = 1$ attains both the lowest loss and the highest AUPRC.

Table 5: **ANE-SGD sensitivity to mirror strength** τ (mean $\pm$ SD over 20 seeds for every column). SUSTAIN is shown for reference.

Task	Metric	SUSTAIN	τ=0.5	τ=0.7	τ=0.9	τ=1.0
Housing	Loss ↓	0.147 ± 0.020	0.110 ± 0.001	0.110 ± 0.001	0.110 ± 0.001	0.110 ± 0.001
	R^2 ↑	0.393 ± 0.116	0.607 ± 0.006	0.606 ± 0.007	0.605 ± 0.007	0.604 ± 0.007
Fashion	Loss ↓	0.548 ± 0.011	0.486 ± 0.008	0.487 ± 0.009	0.487 ± 0.009	0.487 ± 0.009
	Acc (%) ↑	87.6 ± 1.3	89.7 ± 0.3	89.6 ± 0.3	89.6 ± 0.3	89.6 ± 0.3
Acrobot	Loss ↓	0.325 ± 0.012	0.311 ± 0.011	0.310 ± 0.011	0.312 ± 0.010	0.311 ± 0.010
	R^2 ↑	0.159 ± 0.034	0.199 ± 0.034	0.202 ± 0.034	0.198 ± 0.032	0.200 ± 0.032
Fraud	Loss ↓	0.339 ± 0.035	0.284 ± 0.018	0.282 ± 0.021	0.282 ± 0.019	0.281 ± 0.017
	AUPRC ↑	0.916 ± 0.017	0.922 ± 0.007	0.923 ± 0.007	0.923 ± 0.006	0.924 ± 0.006

C.5 Controlled Estimator Comparisons

We distinguish estimator diagnostics from downstream predictive metrics. In the synthetic controls, skewness is summarized by the coordinate statistic $|\gamma_1|$, the mean over coordinates of the absolute standardized skewness of the estimate; it removes the scale of each coordinate's variance and differs from the centered absolute vector moment $\mathcal{M}_3(\xi) = \mathbb{E}\|\xi - \mathbb{E}\xi\|^3$, which depends on scale and need not vanish under symmetric noise. A reduction in either diagnostic alone does not identify the cause of every downstream improvement.

All controls use the synthetic problem with $\kappa_g = 100$, $m = 20$, multiplier standard deviation 0.3, $L = 2000$ estimates per run, and eight seeds (mean $\pm$ SD over seeds). The multipliers are $s = 1 + 0.3Z$, with Z a standardized gamma variable of the target skewness (a clipped Gaussian at skewness 0). Tables 6–8 share their Standard draws but draw ANE's mirrored multipliers at different points of the random stream, so their ANE entries at the shared setting (multiplier skewness 4, $\tau = 1$) agree only within seed variability. Standard (2x) averages two independent standard Neumann estimates and so has ANE's stochastic-HVP budget. Reference-only is the $\tau = 0$ construction, which uses the reference without mirroring.

Table 6: Matched controls with distributional mirroring (multiplier skewness 4).

Estimator	Coord. $\lvert\gamma_1\rvert$	Variance	$\mathcal{M}_3$
Standard	2.03 ± 0.28	7.4e−3	1.6e−3
Standard (2x)	1.30 ± 0.13	3.5e−3	4.1e−4
Reference-only ($\tau = 0$)	1.87 ± 0.19	1.8e−3	1.8e−4
ANE ($\tau = 1$)	0.42 ± 0.43	3.6e−3	5.2e−4

At matched stochastic-HVP budget and variance, ANE's coordinate skewness is far below that of Standard (2x), while reference-only keeps the skewness of Standard at a quarter of its variance (with a fixed exact reference, the $\tau = 0$ estimate is half of Standard plus a constant, so its standardized skewness equals Standard's by construction).

Table 7: Fixed-variance sweep with distributional mirroring: the multiplier variance is fixed and its standardized skewness is varied.

Multiplier skewness	Standard (2x) $\lvert\gamma_1\rvert$	ANE $\lvert\gamma_1\rvert$	Variance, 2x / ANE	$\mathcal{M}_3$, 2x / ANE
0 (symmetric)	0.21 ± 0.05	0.23 ± 0.04	3.6e−3 / 3.6e−3	3.3e−4 / 3.3e−4
2	0.52 ± 0.05	0.18 ± 0.09	3.6e−3 / 3.6e−3	3.3e−4 / 3.7e−4
4	1.30 ± 0.13	0.20 ± 0.10	3.5e−3 / 3.6e−3	4.1e−4 / 4.6e−4
6	1.98 ± 0.12	0.39 ± 0.18	3.7e−3 / 3.6e−3	5.3e−4 / 5.6e−4

In Table 7, the standard control's skewness increases with the skewness of the multipliers, while ANE's remains substantially lower at the three nonzero targets, at the same variance. For symmetric multipliers the two are close: this negative control shows no detectable advantage in the measured diagnostic.

C.6 Mirror Strength and Distributional Asymmetry

For an exact reference, a perturbation term of odd degree n has coefficient $(1-\tau^n)/2$ in the averaged construction, relative to the Standard coefficient. Thus $\tau = 0.5$ retains one quarter of the leading linear coefficient, whereas $\tau = 1$ cancels it exactly. Table 8 records the skewness sweep over τ with distributional mirroring (same setting as Appendix C.5, multiplier skewness 4), together with $\mathcal{M}_3$ and the variance at each mirror strength.

Table 8: Mirror-strength sweep with distributional mirroring (mean $\pm$ SD over eight seeds).

	Standard	$\tau = 0$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 1$
Coord. $\lvert\gamma_1\rvert$	2.03 ± 0.28	1.89 ± 0.33	1.70 ± 0.18	1.29 ± 0.47	0.55 ± 0.14	0.33 ± 0.39
$\mathcal{M}_3$	1.6e−3	1.9e−4	1.9e−4	2.5e−4	2.9e−4	4.8e−4
$\mathcal{M}_3$ below Standard	–	88%	88%	84%	81%	69%
Variance	7.4e−3	1.9e−3	1.9e−3	2.3e−3	2.7e−3	3.6e−3

The skewness diagnostic continues to improve beyond $\tau = 0.5$, even when the task metrics in Table 5 change only modestly, and $\mathcal{M}_3$ stays 69–88% below Standard at every mirror strength. Because $\mathcal{M}_3$ is scale-sensitive and also reflects variance, which rises toward $\tau = 1$ in this construction, its behavior need not coincide with the skewness curve. We use $\tau = 1$ as the a priori default for exact odd-degree cancellation.

C.7 Real-Task Reference and Estimator Diagnostics

This section reports, for the real tasks, the error of ANE's training reference, whether the Neumann recursion is contractive as Proposition 3.6 assumes, and, for completeness, the per-task estimator moments. All diagnostics use each task's training configuration (Table 4) at checkpoints t of a seed-0 ANE-SGD run of T outer steps, with that run's $\hat{\eta}$, mirror strength, safeguard, and current reference $\bar{H}$.

Reference error and contraction. ANE's reference (an HVP operator on a batch four times the HVP batch, refreshed every ten steps) has relative error 0.06–1.99 across tasks and checkpoints (Table 9). At the training $\hat{\eta} = \hat{L}_{\max}/2$, the depth-$m$ products are not contractive: their per-factor rates $(\| \prod_{j=1}^{m} \hat{A}_j^{\pm} b\| / \|b\|)^{1/m}$, averaged over eight draws, exceed one at most checkpoints (0.92–3.03 for the forward and 1.30–9.26 for the raw $\tau = 1$ mirrored branch), and on Acrobot the products overflow at several checkpoints. Proposition 3.6 therefore does not apply on these tasks, and both estimators stay bounded only through the branch-norm guard and, for ANE, the τ safeguard.

Table 9: Relative error $\|\bar{H} - H\| / \|H\|$ of ANE's training reference $\bar{H}$ with respect to the full training Hessian H (spectral norms by power iteration) at five checkpoints of the seed-0 ANE-SGD run.

	Checkpoint t/T				
Task	0	0.25	0.5	0.75	1
Housing	0.68	0.13	0.47	0.43	0.31
Fashion	0.06	1.09	1.99	0.96	1.42
Acrobot	0.37	0.08	0.34	0.50	0.45
Fraud	0.52	0.96	0.71	0.44	0.74

Estimator moments. Table 10 lists the per-task bias, variance, and skewness of Standard, Standard (2x), and ANE at each task's training depth; because the recursion is not contractive at the training $\hat{\eta}$, these magnitudes reflect the safeguards rather than antithetic cancellation.

Table 10: Estimator moments at each task's training depth m and mid-training checkpoint $t = T/2$ (mean $\pm$ SD over four repetitions of $L = 250$ estimates $\hat{g}$). Bias is $\|\mathbb{E}\hat{g} - g_{\mathrm{ref}}\|$ and variance $\mathbb{E}\|\hat{g} - \mathbb{E}\hat{g}\|^2$, with g_{ref} a deep Standard estimate (depth 200, batch 512, 32 draws) at the training $\hat{\eta}$; γ_1 is the standardized skewness of $u^\top \hat{g}$ along the fixed direction $u = g_{\mathrm{ref}}/\|g_{\mathrm{ref}}\|$; coordinate $|\gamma_1|$ is the mean over coordinates of the absolute standardized skewness. Standard (2x) averages two Standard estimates (ANE's stochastic-HVP budget).

	Error moments		Skewness	
Estimator	Bias	Variance	γ_1 along u	Coord. $\|\gamma_1\|$
Housing ($m = 50, t = 750$)				
Standard	$7.3\mathrm{e}{-3} \pm 1\mathrm{e}{-3}$	$0.014 \pm 5\mathrm{e}{-3}$	$+0.18 \pm 3.78$	2.09 ± 1.48
Standard (2x)	$6.6\mathrm{e}{-3} \pm 2\mathrm{e}{-3}$	$9.3\mathrm{e}{-3} \pm 6\mathrm{e}{-3}$	$+0.56 \pm 2.69$	2.62 ± 2.14
ANE	0.38 ± 0.08	17 ± 9	$+0.91 \pm 2.18$	1.15 ± 1.42
Fashion ($m = 20, t = 1000$)				
Standard	1.4 ± 0.4	70 ± 7	$+0.06 \pm 0.22$	0.22 ± 0.01
Standard (2x)	1.5 ± 0.1	33 ± 2	$+0.14 \pm 0.12$	0.17 ± 0.02
ANE	1.4 ± 0.3	203 ± 11	-0.13 ± 0.07	0.16 ± 0.01
Acrobot ($m = 50, t = 750$)				
Standard	0.90 ± 0.4	46 ± 2	-0.05 ± 0.05	0.43 ± 0.14
Standard (2x)	0.61 ± 0.2	22 ± 2	-0.07 ± 0.06	0.43 ± 0.10
ANE	0.85 ± 0.2	61 ± 8	-0.06 ± 0.24	0.52 ± 0.22
Fraud ($m = 10, t = 1500$)				
Standard	0.32 ± 0.01	1.2 ± 0.2	$+3.53 \pm 1.25$	5.96 ± 0.83
Standard (2x)	0.35 ± 0.02	1.1 ± 0.4	-3.50 ± 3.60	3.45 ± 0.41
ANE	$0.56 \pm 7\mathrm{e}{-3}$	5.2 ± 1	-2.57 ± 1.99	4.64 ± 0.81

C.8 Runtime Across Truncation Depths

Table 11 reports the time of one hypergradient computation at a shared mid-training checkpoint for $m \in \{5, 10, 20, 50\}$, with each task's configuration (Table 4, which also lists the task depths). Times exclude the shared K-step inner solve and are means over three seeds of 20 calls on one CPU thread (Apple M4 Pro). They include ANE's reference handling: the refresh is negligible, and applying the reference costs about one stochastic HVP per mirrored factor.

Table 11: Hypergradient-step wall-clock time (ms per call; ratio to StocBiO in parentheses).

Task	m	StocBiO	SUSTAIN	ANE-SGD	ANE-VR
Housing	5	1.02	1.69 (1.6×)	2.76 (2.7×)	3.19 (3.1×)
	10	1.86	2.96 (1.6×)	5.38 (2.9×)	7.08 (3.8×)
	20	3.47	5.20 (1.5×)	10.71 (3.1×)	13.75 (4.0×)
	50	8.79	13.23 (1.5×)	26.87 (3.1×)	33.22 (3.8×)
Fashion	5	0.93	1.41 (1.5×)	2.60 (2.8×)	3.35 (3.6×)
	10	1.69	2.63 (1.6×)	5.03 (3.0×)	6.15 (3.6×)
	20	3.23	5.00 (1.5×)	10.17 (3.1×)	11.29 (3.5×)
	50	7.79	11.96 (1.5×)	25.85 (3.3×)	29.49 (3.8×)
Acrobot	5	1.05	1.73 (1.7×)	2.88 (2.8×)	3.31 (3.2×)
	10	1.88	3.01 (1.6×)	5.69 (3.0×)	7.34 (3.9×)
	20	3.60	5.37 (1.5×)	11.05 (3.1×)	13.63 (3.8×)
	50	8.70	13.18 (1.5×)	27.28 (3.1×)	33.97 (3.9×)
Fraud	5	0.92	1.57 (1.7×)	2.67 (2.9×)	3.01 (3.3×)
	10	1.72	2.71 (1.6×)	5.14 (3.0×)	6.82 (4.0×)
	20	3.33	4.84 (1.5×)	10.34 (3.1×)	12.74 (3.8×)
	50	8.03	12.38 (1.5×)	25.17 (3.1×)	31.98 (4.0×)

Across tasks and depths, ANE-SGD costs 2.7–3.3 times StocBiO's hypergradient step (3.0–3.1 at the task depths), consistent with three HVP-type operations per factor against one; ANE-VR costs 3.1–4.0 times and SUSTAIN 1.5–1.7 times, reflecting the extra computation in their PAGE steps.

The mirrored Hessian acts as $\hat{H}^{-}v = (1+\tau)\bar{H}v - \tau\hat{H}v$. It can be applied through a reference-batch HVP without storing a dense $p \times p$ Hessian. Each factor remains an HVP computation with dependence on the model and reference-batch size, while ANE adds another branch and reference applications.

C.9 Per-Task Asymmetry Analysis

An illustrative source of curvature asymmetry is the generalized Gauss–Newton term $b^{-1}\sum_{i\in B}\sigma_i'' J_i^\top J_i$. For a nonlinear model, the full loss Hessian also contains $\ell'(z_i)\nabla^2 z_i$ (and regularization); the displayed term is therefore not an identity for every MLP Hessian. If directional Jacobian factors are treated as fixed, mini-batch factor noise has third moment proportional to

$$\mu_3' \propto -\frac{1}{b^2}\mathbb{E}[(\sigma_i'' - \bar{\sigma}'')^3]. \tag{54}$$

For a two-point curvature model with active fraction p and active curvature c, the exact identity is

$$\mathbb{E}[(\sigma_i'' - \bar{\sigma}'')^3] = pc^3(1-p)(1-2p). \tag{55}$$

This is nonzero away from $p \in \{0, 1/2, 1\}$. Its absolute value is not monotone as p moves toward zero; the standardized two-point skewness is $(1-2p)/\sqrt{p(1-p)}$. The following task descriptions motivate asymmetry qualitatively rather than certify scalar-model assumptions for the full Hessian.

Ill-conditioning from reduced effective samples (κ_g). For a small threshold $\varepsilon_{\mathrm{curv}} > 0$, let $n_{\mathrm{eff}} = |\{i : \sigma_i'' \geq \varepsilon_{\mathrm{curv}}\}|$. Curvature concentrated on a small subset can reduce directional coverage in the Gauss–Newton term and worsen conditioning. Pseudo-Huber curvature is strictly positive at finite residuals, so this counts substantial curvature rather than literal nonzero curvature.

In *Task 1 (Housing)*, pseudo-Huber regression gives $\sigma_i'' = (1 + (r_i/\delta)^2)^{-3/2}$ where $r_i = y_i - f(x_i;\boldsymbol{\beta})$ is the residual. Corrupted targets (20%) produce large residuals with $\sigma_i'' \approx 0$, while clean targets (80%) yield $\sigma_i'' > 0$. This reduces n_{eff} to $\sim$80%; the MLP Jacobian J_i varies with hidden-layer activations, so the reduced active subset provides non-uniform directional coverage, which can worsen conditioning. The resulting bimodal curvature distribution ($p = 0.8$ clean, 0.2 corrupted) gives $\mathbb{E}[(\sigma_i'' - \bar{\sigma}'')^3] \neq 0$ via (55), so $\mu_3' \neq 0$.

In *Task 3 (Acrobot)*, the same pseudo-Huber loss applies with $\sigma_i'' = (1 + (r_i/\delta)^2)^{-3/2}$. Expert episodes ($\sim$136 steps) and random episodes ($\sim$500 steps) contribute unequal transition counts: random transitions outnumber expert ones $\sim$4:1 and have large residuals with $\sigma_i'' \approx 0$, reducing n_{eff}

to $\sim 21\%$ of the training data. The bimodal curvature distribution ($p \approx 0.21$ expert, 0.79 random) gives $\mu_3' \neq 0$ via (54)–(55), motivating possible conditioning and asymmetry effects in this approximation.

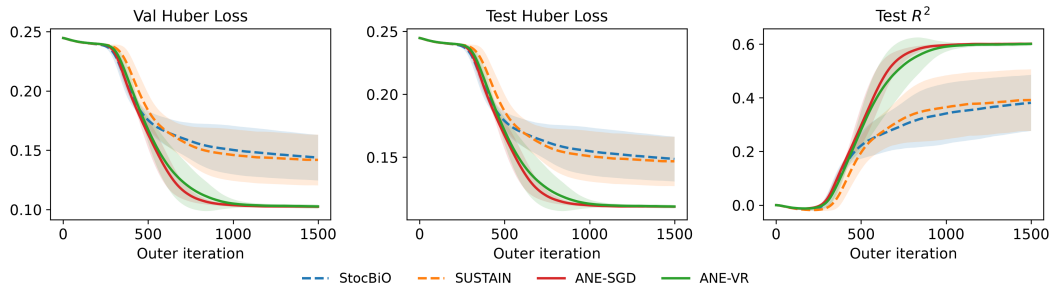
Noise asymmetry from bimodal curvature (μ_3'). In *Task 2 (Fashion)*, softplus classification gives $\sigma_i'' = \text{sigmoid}(z_i)(1 - \text{sigmoid}(z_i))$ where $z_i = y_i f(x_i; \boldsymbol{\beta})$ is the margin. Corrupted samples resist confident classification ($|z_i|$ small, $\sigma_i'' \approx 0.25$); clean samples are classified confidently ($|z_i|$ large, $\sigma_i'' \approx 0$). With $p \approx 0.05$ (10% of the positive half), equation (55) gives $\mathbb{E}[(\sigma_i'' - \bar{\sigma}'')^3] > 0$, so $\mu_3' \neq 0$.

In *Task 4 (Fraud)*, the same softplus loss applies. Fraud samples ($p = 0.06$) lie near the decision boundary ($\sigma_i'' \approx 0.25$); legitimate samples have large margins ($\sigma_i'' \approx 0$). The minority fraction can produce large standardized curvature skewness; it does not imply the largest raw third moment across tasks, since curvature scales and directional Jacobian factors differ.

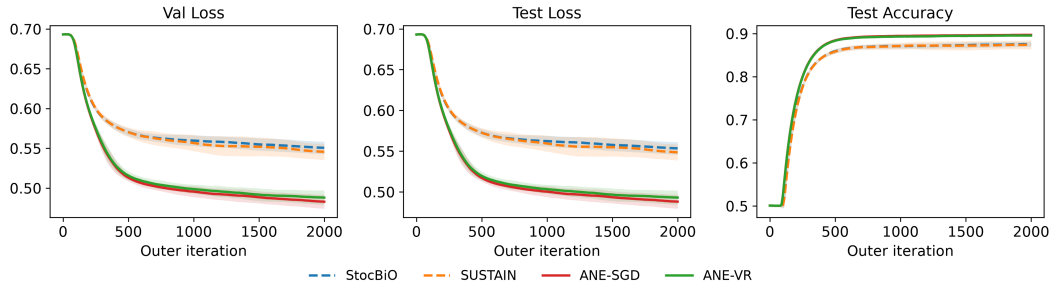
Interpretation. These mixture arguments explain why the per-sample curvature can be asymmetric, but they do not show that the scalar perturbative regime holds on a given task. At the mid-training and final checkpoints, the training Hessian is ill-conditioned on Housing ($\hat{\kappa}_g \approx 240$–$340$) and near-singular on Acrobot; on Fashion it is near-singular at mid-training and reaches $\hat{\kappa}_g \approx 180$ by the end. By contrast, Fraud is well-conditioned ($\hat{\kappa}_g \approx 40$) and has by far the largest per-factor curvature skewness.

D Enlarged Convergence Trajectories

These are enlarged versions of the panels in Figure 2 (mean $\pm$ SD over 20 seeds). Curves are smoothed with an exponential moving average (smoothing factor 0.1) before averaging.

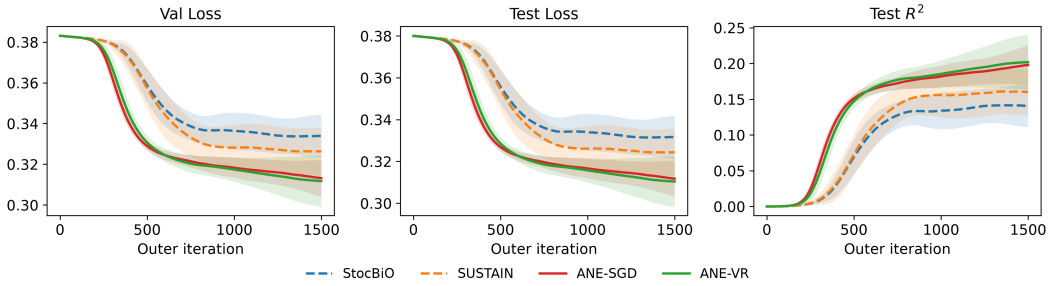


(a) California Housing. $n = 20$.

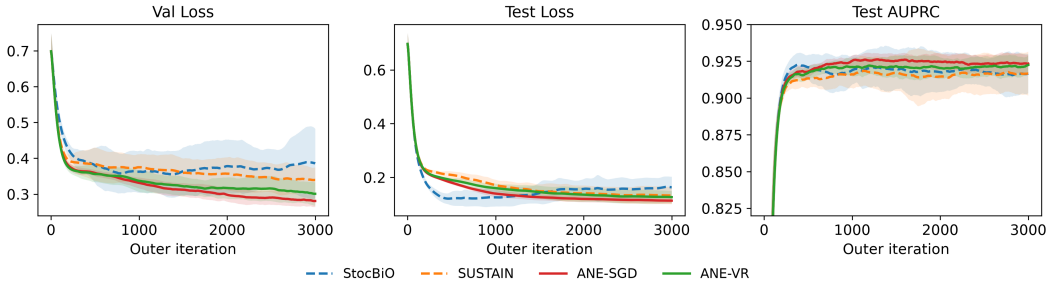


(b) Fashion-MNIST. $n = 20$.

Figure 3: **Convergence on bilevel tasks (enlarged).** Each panel: validation (outer) loss, test loss, and the task-specific test metric (mean $\pm$ SD over 20 seeds) vs. outer iteration. Dashed: baselines (StocBiO, SUSTAIN); solid: ANE. On Housing, Fashion, and Acrobot, ANE-SGD and ANE-VR separate from baselines early and maintain a consistent gap; on Fraud, StocBiO has the lowest test loss mid-training and the AUPRC bands overlap. These comparisons match outer iterations rather than wall-clock cost.



(c) Acrobot. $n = 20$.



(d) Fraud detection. $n = 20$; the validation panel shows the class-balanced outer objective.

Figure 3: **Convergence on bilevel tasks (continued).** Mean $\pm$ SD over 20 seeds, with the same conventions as the preceding panels.

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 two main claims: (C1) the centered third-moment proxy $\mathcal{M}_3$ of the Neumann estimator grows with truncation depth m under weak contraction (Lemma A.2), and ANE reduces it by 41–74% relative to a single standard estimate in the synthetic setting (Figure 1); (C2) ANE improves on stochastic and variance-reduced baselines on every task (20 seeds each), lowering loss by 4–25% relative to the best baseline; the largest gains, on robust regression (Housing), are a 25% lower test loss and a 54% higher test R^2 (Table 3, Figure 2).

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses the extra HVP cost, the contraction assumed by the approximate-reference analysis, and why the downstream gains are not attributed to skewness correction alone; the symmetric-noise control (Section 5) shows no gain when skewness is absent.

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 in Section 2. Complete proofs of all propositions and the main theorem are in Appendix A.

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: Appendix B gives the full algorithms. Full hyperparameters are in Appendix C (Table 4), together with the seed counts, the models and data for each task, and the synthetic setup (Appendix C.2). All datasets are publicly available (sklearn, torchvision, Kaggle, Gymnasium) and cited in Section 5.

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: [No]

   Justification: The code is not released. All datasets are publicly available and cited in Section 5. The algorithms, hyperparameters, seed counts, and task configurations needed to reimplement the experiments are given in Appendices B and C.

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: Section 5 describes all tasks, and Appendix C provides the complete hyperparameter table, optimizer details, and data preparation.

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: Table 3 and all convergence plots report mean $\pm$ standard deviation over 20 seeds per task, for every metric; the synthetic controls, ablations, and diagnostics state their own seed or repetition counts.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments run on CPU (no GPU required). One seed of one real task, all four methods, takes one to three minutes on a laptop CPU (Apple M4 Pro); the 20-seed comparison over the four tasks takes about 3 CPU-hours, and the synthetic controls and diagnostics a few CPU-hours more. Total compute for all experiments across all seeds is under 24 CPU-hours on a standard workstation.

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: This work is methodological (optimization algorithms) and does not involve human subjects, sensitive data, or dual-use concerns.

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 paper is foundational research on optimization algorithms with no direct societal impact beyond that typically associated with bilevel optimization in machine learning.

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 pre-trained models, generators, or scraped datasets.

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: All datasets are cited (sklearn California Housing, torchvision Fashion-MNIST, Kaggle creditcard, Acrobot-v1 from Gymnasium, the maintained successor of OpenAI Gym). All baseline algorithms are cited.

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: No new datasets, code, or pre-trained models are released.

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 human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects research.

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 a component of the core methodology.