GHOST HYPERGRADIENTS: VALUE-GAP ORACLES FOR OPTIMISTIC BILEVEL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimistic bilevel learning is fragile when the lower-level problem has many minimizers: differentiating an arbitrary selected response can follow a solver-dependent branch rather than the validation-preferred one. We study a value-function alternative based on a vanishing validation tilt. Let G be the lower-level training value and V_τ the value after adding τ times the validation loss to the lower objective. The scaled value gap $\Delta_\tau = (V_\tau - G)/\tau$ converges, under standard compactness and continuity conditions, to the optimistic objective that minimizes validation loss over the training-optimal set. Building on value-function penalties and equilibrium propagation, we make this perturbation identity algorithmic: by differencing conservative adjoint selections for V_τ and G, we obtain *ghost hypergradients*, exact conservative oracles for the ideal finite-tilt surrogate that avoid differentiating a set-valued argmin map. Read at one shared random outer point, their exact scaled difference is almost surely a surrogate gradient, and a certificate error bound lets solve residuals control the computed direction. We prove local surrogate-error bounds, sharpened to $O(\tau^{\theta/(1-\theta)})$ for $0 < \theta < 1$, projected-stationarity convergence, and inexact-solve bounds, exposing the bias–accuracy tradeoff in τ. On an exact flat-valley diagnostic, Ghost preserves the optimistic direction while selected-response gradients are misaligned. In our empirical study, Ghost achieves the best final test accuracy/runtime tradeoff among the compared bilevel oracles. These results support value gaps as practical first-order oracles for nonsmooth optimistic bilevel learning.

1 INTRODUCTION

Bilevel optimization is a natural language for validation-driven learning: the lower level trains a model, while the upper level tunes hyperparameters, data weights, regularizers, or algorithmic choices using validation performance. This viewpoint underlies gradient-based hyperparameter optimization, meta-learning, and data reweighting methods (Franceschi et al., 2017; 2018; Lorraine et al., 2020; Blondel et al., 2022). It is especially appealing for data hyper-cleaning, where a held-out validation split should decide which noisy training examples influence the learned model (Ren et al., 2018; Shaban et al., 2019; Wu et al., 2026).

The standard computational interface is a selected response: solve the lower problem, return one minimizer or one optimization trajectory, and differentiate validation loss through that choice. This is effective when the lower solution is stable and essentially single-valued. It is fragile in the regime created by nonsmooth penalties, constraints, sparsity, sample reweighting, and flat valleys, where the lower level can have many minimizers. In an *optimistic* bilevel problem, the upper level should not evaluate an arbitrary selected minimizer; it should evaluate the training minimizer with the best validation loss. A selected-response hypergradient can therefore optimize the solver's tie-breaking rule rather than the optimistic objective. Figure 1 shows this failure in the simplest flat-valley case.

Our premise is that, for this target, the obstruction comes from differentiating the wrong object. Rather than differentiate a set-valued argmin map, we work with lower-level values. For an outer state λ, let

$$G(\lambda) = \inf_{w \in \mathcal{W}(\lambda)} \mathcal{L}_{\mathrm{tr}}(w, \lambda),\ V_\tau(\lambda) = \inf_{w \in \mathcal{W}(\lambda)} \{\mathcal{L}_{\mathrm{tr}}(w, \lambda) + \tau \mathcal{L}_{\mathrm{val}}(w, \lambda)\},\ \Delta_\tau(\lambda) = \frac{V_\tau(\lambda) - G(\lambda)}{\tau}. \tag{1}$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

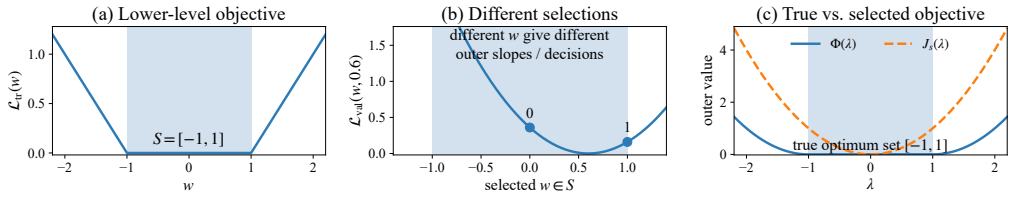


Figure 1: Why selected responses can be the wrong primitive for the optimistic target. The training loss has the flat minimizer set $S = [-1, 1]$, while validation loss $(w - \lambda)^2$ selects the closest point in S. The optimistic objective is $\Phi(\lambda) = \mathrm{dist}(\lambda, S)^2$, minimized by every $\lambda \in S$. A solver selection such as $s(\lambda) \equiv 0$ instead gives the surrogate λ^2, which, relative to the optimistic objective, wrongly makes 0 the unique optimum.

As $\tau \downarrow 0$, the tilted problem first enforces training optimality and then lets validation break ties. Under the compactness and continuity assumptions stated in Section 3,

$$\Delta_\tau(\lambda) \longrightarrow \Phi(\lambda) := \min_{w \in S(\lambda)} \mathcal{L}_{\mathrm{val}}(w, \lambda), \qquad S(\lambda) := \arg\min_{w \in \mathcal{W}(\lambda)} \mathcal{L}_{\mathrm{tr}}(w, \lambda). \tag{2}$$

The limit identity is a classical perturbation–envelope principle, related to Tikhonov, viscosity, and lexicographic selection (Willoughby, 1979; Moudafi & Maingé, 2006; Abernethy et al., 2024), and Δ_τ is the marginal function of a classical value-function penalty (Outrata, 1990; Ye & Zhu, 1995; Dempe et al., 2019). Our contribution is algorithmic: we turn this infinitesimal tie-break into a first-order oracle for optimistic bilevel learning.

The oracle is simple. At the same outer state, solve two lower-level value problems: the original training problem and the validation-tilted problem. Conservative-adjoint theory provides first-order fields for the values G and V_τ without requiring a differentiable or unique solution map (Bolte & Pauwels, 2021; Bolte et al., 2026). If $u_0(\lambda) \in D_G(\lambda)$ and $u_\tau(\lambda) \in D_{V_\tau}(\lambda)$ are such adjoint selections, then

$$d_\tau(\lambda) := \frac{u_\tau(\lambda) - u_0(\lambda)}{\tau} \tag{3}$$

integrates exactly to the finite-tilt surrogate Δ_τ along every path. We call d_τ a *ghost hypergradient*: it differentiates the validation preference expressed by two nearby training worlds, rather than differentiating one observed response trajectory. The construction is not a finite difference in λ; the outer state is fixed, and the quotient is taken across lower-level value functions that differ only by the validation tilt. Computing it from a base solve and a nudged solve follows the two-phase principle of equilibrium propagation (Scellier & Bengio, 2017; Zucchet & Sacramento, 2022).

This gives Ghost the same outer-loop interface as modern hypergradient methods while targeting a different object. In sample reweighting, the oracle is a loss-difference quotient between the base and tilted solutions, avoiding inverse Hessians and derivatives of a selected response. Smaller τ reduces modeling bias but amplifies solve and adjoint errors by $1/\tau$: a statistical–computational tradeoff, not an implementation detail.

Two further ingredients turn this route into a pointwise algorithm with a computable error. First, when G and V_τ are nonsmooth, two individually valid adjoints can come from different one-sided branches, so (3) need not be a gradient of Δ_τ where an algorithm uses it; Ghost reads both adjoints *compatibly*, at one shared, randomly perturbed outer point, where their exact difference is almost surely a gradient of Δ_τ (Section 4). Second, under a certificate error bound, the residuals of the two finite solves control the error of the computed direction, as the fixed-τ convergence analysis requires (Section 5).

Contributions. We make four contributions. First, we give a self-contained value-gap formulation of optimistic bilevel learning, showing how a vanishing validation tilt recovers the validation-best training minimizer. Second, we define ghost hypergradients as differences of compatible value adjoints read at a shared outer point, with a certificate-residual bound on the computed direction. Third, we prove local finite-tilt approximation (sharpened to $O(\tau^{\theta/(1-\theta)})$ for $0 < \theta < 1$), projected-stationarity convergence for the smooth fixed-τ surrogate, projected-PL value transfer to Φ, and

inexact-solve bounds that expose the bias–accuracy tradeoff in τ. Fourth, we validate the story on an exact flat-valley diagnostic and an SNLI noisy-label hyper-cleaning case study with a matched 100-epoch horizon, plus a clean-validation protocol and a pairing diagnostic.

2 Related work

Selected-response hypergradients. Most gradient-based bilevel algorithms differentiate the lower response returned by an inner solver: unrolled or truncated methods backpropagate through a training trajectory, while implicit methods differentiate local optimality conditions (Franceschi et al., 2017; 2018; Shaban et al., 2019; Lorraine et al., 2020; Blondel et al., 2022). Stochastic variants such as StocBiO, TTSA, SUSTAIN, VRBO, SABA/SOBA, and MA-SOBA reduce this cost via two-timescale updates, momentum, variance reduction, or single-loop designs (Ji et al., 2021; Hong et al., 2023; Khanduri et al., 2021; Yang et al., 2021; Dagréou et al., 2022; Chen et al., 2024b). These are natural hyper-cleaning baselines, but their cleanest guarantees presuppose a stable response—strong convexity, smoothness/PL structure, or a well-conditioned implicit system. They therefore optimize the solver-selected response, not necessarily the validation-preferred member of a nonunique solution set; that is the intended target when training dynamics are part of the model (Vicol et al., 2022; Arbel & Mairal, 2022; Jiang et al., 2026). For the optimistic target, Ghost addresses precisely this hidden failure mode.

Nonunique lower levels and selection. Recent theory explains the difficulty: without sufficient lower-level structure, finding small hypergradients can be intractable (Chen et al., 2024a). Existing remedies impose uniform convexity, set-smoothness, optimistic composite structure, or ADMM-style aggregation (Wu et al., 2026; Chen et al., 2025; Solla & Royset, 2026; Xiao et al., 2026). Closest to our target, HG-MS selects the optimistic minimizer explicitly and differentiates through a pseudoinverse, reaching stationary points of the optimistic objective under local PL structure (Masiha et al., 2026), a stronger target than ours; SQ-G relaxes minima selection into Langevin sampling and reaches Goldstein stationarity by zeroth-order steps (Masiha et al., 2025). Ghost keeps the optimistic problem intact and reveals the validation tie-break through the scalar finite-tilt value gap $\Delta_\tau = (V_\tau - G)/\tau$, rather than through an explicit selection step, a sampled minimizer, or a solver-selected argmin; it needs no manifold or PL structure, and its guarantees concern the fixed-τ surrogate.

Value-gap oracle viewpoint. Classical value-function reformulations enforce lower optimality by equality to the lower value, but usually add value constraints, penalties, barriers, or auxiliary variables (Outrata, 1990; Ye & Zhu, 1995; Dempe et al., 2019; Liu et al., 2021), and Δ_τ is the marginal function of such a penalty with weight $1/\tau$. Ghost instead uses value functions as an oracle: query the training value G and tilted value V_τ, subtract, scale, and differentiate; computing the derivative from a free and a nudged problem is the two-phase principle of equilibrium propagation (Scellier & Bengio, 2017), reviewed for bilevel optimization by Zucchet & Sacramento (2022). The ghost hypergradient is a difference of compatible conservative value adjoints; cancellation removes the common training direction and exposes the optimistic selector. Vanishing tilts relate to Tikhonov, viscosity, and lexicographic selection (Willoughby, 1979; Moudafi & Maingé, 2006; Abernethy et al., 2024), and conservative-field theory supplies value fields even for set-valued argmin maps (Bolte & Pauwels, 2021; Bolte et al., 2026). Thus we do not claim a universal complexity improvement in strongly convex regimes. The contribution is an oracle for optimistic selection under genuine degeneracy, with compatible readouts, a residual-to-direction error bound, and an explicit bias–accuracy tradeoff (Table 1, Appendix A).

3 Problem and Value-Gap Reformulation

We study bilevel learning when the lower problem may be nonsmooth and nonunique. For outer variables $\lambda \in \Lambda \subseteq \mathbb{R}^m$ and model parameters $w \in \mathbb{R}^d$, the lower level minimizes training loss over a feasible set $\mathcal{W}(\lambda)$. The upper level is *optimistic*: it evaluates validation loss at the validation-best training minimizer, not at an arbitrary solver-selected response.

Definition 1 (Optimistic bilevel objective)**.** Let $S(\lambda) := \arg\min_{u \in \mathcal{W}(\lambda)} \mathcal{L}_{\mathrm{tr}}(u, \lambda)$. The optimistic bilevel objective is

$$\min_{\lambda \in \Lambda} \Phi(\lambda), \qquad \Phi(\lambda) := \min_{w \in S(\lambda)} \mathcal{L}_{\mathrm{val}}(w, \lambda). \tag{4}$$

We do not assume that $S(\lambda)$ is singleton, differentiable, or continuously selectable. Specifically, our main empirical instance is validation-driven sample reweighting. Given noisy training examples $\{\zeta_i\}_{i=1}^n$ and a held-out validation set $\{\xi_j\}_{j=1}^{n_{\mathrm{val}}}$, a representative lower-level template is

$$\mathcal{L}_{\mathrm{tr}}(w,\lambda) = \frac{1}{n}\sum_{i=1}^{n} \sigma(\lambda_i)\,\ell(w;\zeta_i) + R(w) + I_{\mathcal{C}}(w), \qquad \mathcal{L}_{\mathrm{val}}(w,\lambda) = \frac{1}{n_{\mathrm{val}}}\sum_{j=1}^{n_{\mathrm{val}}} \ell_{\mathrm{val}}(w;\xi_j). \quad (5)$$

Here $\sigma(\lambda_i)$ is a learned example weight, R may be nonsmooth, and $I_{\mathcal{C}}$ encodes constraints. This covers standard hyper-cleaning/reweighting while allowing set-valued lower minimizers (Ren et al., 2018; Shaban et al., 2019; Wu et al., 2026).

First-order convention. Ghost differentiates value functions, not selected minimizers. For a locally Lipschitz $h : U \to \mathbb{R}$, a locally bounded set-valued map $D_h : U \rightrightarrows \mathbb{R}^m$ with closed graph and nonempty compact values is a conservative field (Bolte & Pauwels, 2021) if every absolutely continuous curve $\gamma : [0,1] \to U$ and measurable selection $v(t) \in D_h(\gamma(t))$ satisfy

$$h(\gamma(1)) - h(\gamma(0)) = \int_0^1 \langle v(t), \dot{\gamma}(t)\rangle\, dt. \quad (6)$$

If h is C^1, $\{\nabla h\}$ is a conservative field; conversely, every conservative field equals $\{\nabla h\}$ outside a Lebesgue-null set (Bolte & Pauwels, 2021, Thm. 1), but on that null set it may contain spurious elements.

3.1 From optimistic selection to a value gap

With the base training value, the validation-tilted value, and their finite-tilt gap in (1), the validation tilt is a variational tie-breaker: as $\tau \downarrow 0$, it preserves training optimality while selecting the validation-best point in a flat lower valley. Rearranging the tilted objective gives $\Delta_\tau(\lambda) = \inf_{w\in\mathcal{W}(\lambda)}\{\mathcal{L}_{\mathrm{val}}(w,\lambda) + (\mathcal{L}_{\mathrm{tr}}(w,\lambda) - G(\lambda))/\tau\}$, the marginal function of the value-function penalty with weight $1/\tau$.

Proposition 1 (Vanishing-tilt representation)**.** *Fix $\lambda \in \Lambda$. Assume $S(\lambda)$ is nonempty and compact; $w \mapsto \mathcal{L}_{\mathrm{val}}(w,\lambda)$ is continuous near $S(\lambda)$ and bounded below on $\mathcal{W}(\lambda)$; the tilted infimum defining $V_\tau(\lambda)$ is attained for all sufficiently small $\tau > 0$; and, for some $\varepsilon_0 > 0$, $\mathcal{N}_\lambda := \{u \in \mathcal{W}(\lambda) : \mathcal{L}_{\mathrm{tr}}(u,\lambda) \le G(\lambda) + \varepsilon_0\}$ is compact, with $u \mapsto \mathcal{L}_{\mathrm{tr}}(u,\lambda)$ lower semicontinuous on $\mathcal{N}_\lambda$. Then*

$$V_\tau(\lambda) = G(\lambda) + \tau\Phi(\lambda) + o(\tau), \qquad \Phi(\lambda) = \lim_{\tau\downarrow 0} \Delta_\tau(\lambda). \quad (7)$$

Appendix C.2 proves the proposition. Thus the optimistic objective can be recovered from scalar value gaps, without ever differentiating a selected response map. Proposition 1 is pointwise; uniform finite-tilt rates require the local error bound introduced in Section 5.

4 Ghost Hypergradient Oracle and Algorithm

We now describe the first-order oracle for the finite-tilt surrogate. At an outer evaluation point, the algorithm solves the base and the validation-tilted training problems, extracts one value-function adjoint from each, and differences the two adjoints.

4.1 Value-function adjoints

For $s \in \{0, \tau\}$, write

$$F_s(w,\lambda) := \mathcal{L}_{\mathrm{tr}}(w,\lambda) + s\mathcal{L}_{\mathrm{val}}(w,\lambda), \qquad v_s(\lambda) := \inf_{w\in\mathcal{W}(\lambda)} F_s(w,\lambda). \quad (8)$$

Thus $v_0 = G$ and $v_\tau = V_\tau$. An inner solver is asked for one optimality certificate for each value problem, $(w_s, \nu_s) \in \mathrm{Cert}_s(\lambda)$, $s \in \{0,\tau\}$, where ν_s denotes whatever dual, proximal, or stationarity information the solver exposes (Appendix B.1). In the abstract oracle, we only need the value-function interface:

$$u_s(\lambda) := \mathrm{Adj}(F_s, \lambda, w_s, \nu_s) \in D_{v_s}(\lambda), \qquad s \in \{0,\tau\}, \quad (9)$$

where D_{v_s} is a conservative field for the value function v_s. Conservative-adjoint theory provides such fields for broad definable parametric problems under local boundedness and qualification conditions, without differentiating the solution map (Bolte et al., 2026). In the smooth unconstrained case, (9) reduces to the familiar envelope output

$$u_s(\lambda) = \nabla_\lambda F_s(w_s, \lambda), \qquad w_s \in \arg\min_w F_s(w, \lambda), \tag{10}$$

at a global minimizer (a merely stationary point of a nonconvex lower problem does not certify the value); constrained and nonsmooth problems use the Lagrangian and conservative-field analogues (Appendix B.1).

Ghost hypergradient. Given paired adjoint outputs for the base and tilted values, define

$$d_\tau(\lambda) := \frac{u_\tau(\lambda) - u_0(\lambda)}{\tau}, \qquad u_\tau(\lambda) \in D_{V_\tau}(\lambda), \quad u_0(\lambda) \in D_G(\lambda). \tag{11}$$

Collecting all such differences gives the Cartesian field $D_{\Delta_\tau} := (D_{V_\tau} - D_G)/\tau$. With measurable paired selections, subtracting the conservative-chain-rule identities for V_τ and G and dividing by τ makes D_{Δ_τ} an exact conservative oracle for the finite-tilt surrogate $\Delta_\tau = (V_\tau - G)/\tau$ along paths. This holds for any measurable valid pair; the pointwise direction used by the algorithm must in addition be compatible (Section 4.2). Its connection to Φ is pathwise and finite-tilt (Section 5): $d_\tau(\lambda)$ is not a pointwise subgradient of Φ, nor does it by itself imply monotone descent on Φ. Algorithmically, (11) is the entire construction: two compatible value adjoints read at one outer point, and one difference quotient in the tilt strength.

4.2 Compatible readouts

Throughout, G and V_τ are locally Lipschitz on an open set $U \supseteq \Lambda$ with conservative fields D_G and D_{V_τ}. The pathwise identity does not make an arbitrary pair a gradient of Δ_τ at a point, since conservative fields may contain spurious elements where G or V_τ is not differentiable; in Example 1 (Appendix C.6) two valid readouts at a minimizer give $d_\tau = 2/\tau$ although $\nabla\Delta_\tau = 0$, and a projected step ascends. We therefore pair adjoints through a common approach to the point.

Definition 2 (Compatible pairs)**.** Let $R_\tau \subseteq U$ be the full-measure set where G and V_τ are both differentiable. For $\lambda \in U$ let $\mathcal{J}_\tau(\lambda)$ be the set of limits $(u_\tau, u_0) = \lim_k(\nabla V_\tau(\lambda_k), \nabla G(\lambda_k))$ along sequences $\lambda_k \in R_\tau$ with $\lambda_k \to \lambda$, and $D_\tau^{\mathrm{cmp}}(\lambda) = \{(u_\tau - u_0)/\tau : (u_\tau, u_0) \in \mathcal{J}_\tau(\lambda)\}$.

Proposition 2 (Compatible differences are generalized gradients)**.** *For every $\lambda \in U$, $\mathcal{J}_\tau(\lambda)$ is nonempty, compact, and contained in $\partial_C V_\tau(\lambda) \times \partial_C G(\lambda)$, and $D_\tau^{\mathrm{cmp}}(\lambda) \subseteq \partial_B\Delta_\tau(\lambda) \subseteq \partial_C\Delta_\tau(\lambda)$, where ∂_B and ∂_C are the Bouligand and Clarke subdifferentials. If Δ_τ is C^1 on a neighborhood of λ, then $D_\tau^{\mathrm{cmp}}(\lambda) = \{\nabla\Delta_\tau(\lambda)\}$.*

In that example $D_\tau^{\mathrm{cmp}} = \{0\}$ at the minimizer. A Clarke-valid direction is not by itself a descent direction, and the C^1 neighborhood cannot be weakened to differentiability at λ alone.

4.3 Implementation of the oracle

The hyper-cleaning experiment in Section 6 uses the sample-reweighting template in (5). In that experiment, the validation loss, regularizer, and feasible set have no explicit dependence on the sample-weight logits λ. Therefore the value-adjoint routine used for each world is the envelope partial derivative with respect to the logit λ_i, and the implemented ghost update for training example i is the loss-difference quotient:

$$[u_s(\lambda)]_i = \tfrac{1}{n}\sigma'(\lambda_i)\ell(w_s; \zeta_i), \quad s \in \{0, \tau\}, \qquad [d_\tau(\lambda)]_i = \tfrac{1}{n}\sigma'(\lambda_i)\frac{\ell(w_\tau; \zeta_i) - \ell(w_0; \zeta_i)}{\tau}. \tag{12}$$

No inverse-Hessian term or derivative of a selected response map is used: the dependence of the value on λ_i is the partial derivative of the lower objective at the returned lower-level state. Thus a positive component in (12) means that example i is fit less well in the validation-tilted world than in the base world, so a projected first-order step lowers its weight; a negative component tends to raise its weight. Algorithm 1 evaluates (12) at the shared point λ_t, where readouts at global minimizers are compatible (Lemma 1 in Appendix B.3) and approximate ones are controlled through certificate residuals (Theorem 4); the main SNLI runs of Section 6 used the unperturbed iterate.

ALGORITHM 1 GHOST

Input: initial $\lambda^0 \in \Lambda$, tilt $\tau > 0$, stepsizes $\{\eta_t\}$, radii $\delta_t \in (0, \bar{\delta}]$, a law μ with a density supported in the unit ball, inner solver SOLVE, adjoint routine ADJ

for $t = 0, 1, \ldots, T-1$ **do**
  draw $q_t \sim \mu$ independently of the past; $\bar{\lambda}_t \leftarrow \lambda^t + \delta_t q_t$ `// shared point`
  $(w_0, \nu_0) \leftarrow \text{SOLVE}(F_0, \bar{\lambda}_t)$
  $(w_\tau, \nu_\tau) \leftarrow \text{SOLVE}(F_\tau, \bar{\lambda}_t;\, w_0, \nu_0)$ `// warm start`
  $u_0 \leftarrow \text{ADJ}(F_0, \bar{\lambda}_t, w_0, \nu_0)$
  $u_\tau \leftarrow \text{ADJ}(F_\tau, \bar{\lambda}_t, w_\tau, \nu_\tau)$
  $d_t \leftarrow (u_\tau - u_0)/\tau$
  $\lambda^{t+1} \leftarrow \Pi_\Lambda(\lambda^t - \eta_t d_t)$ `// step from the unperturbed iterate`
return λ^T

4.4 SINGLE-WORLD GHOST UPDATE

Algorithm 1 uses one world and a Euclidean projected step. Both solves and readouts share $\bar{\lambda}_t$, whose conditional density avoids the fields' exceptional null sets almost surely, making exact readouts a compatible gradient pair; independent perturbations do not ensure this (Appendix D.8). The point is not projected: Assumption 2 requires values on $\Lambda + \bar{\delta}\,\mathbb{B}$. The step starts at λ^t, with perturbation error $L_\tau \delta_t$ (Theorem 4); floating-point draws are only approximately continuous. Each iteration costs two solves and two adjoint evaluations. Warm-starting the nearby tilted problem saves computation but does not ensure compatibility. Appendix B.4 gives a multi-start bundle as an ambiguity diagnostic.

5 MAIN THEORY

This section gives (i) a uniform finite-tilt approximation to the optimistic selector, (ii) an exact pathwise oracle for the computable surrogate, (iii) an explicit inexact-solve tradeoff, including the error of the directions computed by Algorithm 1, and (iv) projected-gradient and PL rates for the smooth fixed-τ surrogate.

Assumption 1 (Local finite-tilt and adjoint regime)**.** Let $K \subset \Lambda$ be compact. There are $\bar{\tau} > 0$, a compact lower-level tube $\mathcal{N}_K$, constants $L_K, \kappa_K, M_K > 0$, $m_K \in \mathbb{R}$, and $\theta \in (0, 1]$ such that, for every $\lambda \in K$ and $\tau \in (0, \bar{\tau}]$: (a) $S(\lambda)$ and $S_\tau(\lambda)$ are nonempty compact subsets of $\mathcal{W}(\lambda) \cap \mathcal{N}_K$, and $\Phi(\lambda) = \min_{w \in S(\lambda)} \mathcal{L}_{\text{val}}(w, \lambda)$ is attained; (b) the losses and feasible-set description are definable and locally Lipschitz on the tube, and the chosen lower-level certificates provide locally bounded conservative fields D_G and D_{V_τ} for G and V_τ near K; (c) $w \mapsto \mathcal{L}_{\text{val}}(w, \lambda)$ is L_K-Lipschitz on $\mathcal{N}_K$, bounded below by m_K, and $\Phi(\lambda) - m_K \le M_K$; and (d) the training valley obeys the feasible-point error bound

$$\operatorname{dist}(w, S(\lambda)) \le \kappa_K \big(\mathcal{L}_{\text{tr}}(w, \lambda) - G(\lambda)\big)_+^\theta \quad \forall w \in \mathcal{W}(\lambda) \cap \mathcal{N}_K. \tag{13}$$

The exponent θ is the sharpness of the flat lower valley: $\theta = 1$ for a sharp value error bound, and smaller θ for flatter degeneracy. The assumptions are sufficient conditions, not a description of every optimistic bilevel instance.

Theorem 1 (Uniform value-gap selection)**.** *Under Assumption 1, for every $\lambda \in K$ and $\tau \in (0, \bar{\tau}]$,*

$$0 \le \Phi(\lambda) - \Delta_\tau(\lambda) \le C_K \tau^\theta, \qquad C_K := L_K \kappa_K M_K^\theta. \tag{14}$$

Consequently, $\sup_{\lambda \in K} |\Delta_\tau(\lambda) - \Phi(\lambda)| \le C_K \tau^\theta \to 0$ *as* $\tau \downarrow 0$.

Thus the finite problem optimized by Ghost is a one-sided, uniformly consistent surrogate for the optimistic objective, and the validation tilt reveals the optimistic tie-break without selecting or differentiating a branch of $S(\lambda)$.

Corollary 1 (Sharpened finite-tilt bound)**.** *Under Assumption 1, for all $\lambda \in K$ and $\tau \in (0, \bar{\tau}]$, $0 \le \Phi(\lambda) - \Delta_\tau(\lambda) \le b_\tau$, where $b_\tau := \min\{C_K \tau^\theta, (\kappa_K L_K)^{1/(1-\theta)} \tau^{\theta/(1-\theta)}\}$ for $\theta < 1$ and $b_\tau := C_K \tau$ for $\theta = 1$. If $\theta = 1$ and $\tau < 1/(\kappa_K L_K)$, then $S_\tau(\lambda) \subseteq \arg\min_{w \in S(\lambda)} \mathcal{L}_{\text{val}}(w, \lambda)$ and $\Delta_\tau(\lambda) = \Phi(\lambda)$, so $b_\tau = 0$ may be taken. The exponent $\theta/(1-\theta)$ is attained by an instance satisfying Assumption 1.*

Appendix C.4 proves the corollary. Theorems 2 and 3 below hold verbatim with $C_K\tau^\theta$ replaced by b_τ, since their proofs use only the uniform endpoint bound.

Theorem 2 (Ghost is an exact conservative oracle for Δ_τ)**.** *Assume Assumption 1 and fix $\tau \in (0, \bar{\tau}]$. For any absolutely continuous $\gamma : [0,1] \to K$ and measurable paired selections $u_\tau(t) \in D_{V_\tau}(\gamma(t))$, $u_0(t) \in D_G(\gamma(t))$ for a.e. t, define $d_\tau(t) := (u_\tau(t) - u_0(t))/\tau$. Then $\Delta_\tau(\gamma(1)) - \Delta_\tau(\gamma(0)) = \int_0^1 \langle d_\tau(t), \dot{\gamma}(t)\rangle\, dt$.*

Moreover, relative to the limiting optimistic objective,

$$\left|\Phi(\gamma(1)) - \Phi(\gamma(0)) - \int_0^1 \langle d_\tau(t), \dot{\gamma}(t)\rangle\, dt\right| \le 2C_K\tau^\theta. \tag{15}$$

The oracle theorem does not ask whether $\lambda \mapsto S(\lambda)$ is single-valued, continuous, or differentiable; it only asks for value adjoints for G and V_τ, whose subtraction cancels the common training-value contribution. It is pathwise; the pointwise direction of the algorithm is covered by Proposition 2 and Theorem 4.

Theorem 3 (Inexact solves and the tilt tradeoff)**.** *Assume the conditions of Theorem 2. Let $d_\tau(\lambda) = (u_\tau(\lambda) - u_0(\lambda))/\tau$ be an exact paired ghost direction, and suppose the implemented adjoint selections are measurable along every absolutely continuous curve in K and obey $\|\tilde{u}_0(\lambda) - u_0(\lambda)\| \le \varepsilon_{0,K}$ and $\|\tilde{u}_\tau(\lambda) - u_\tau(\lambda)\| \le \varepsilon_{\tau,K}$ for all $\lambda \in K$. For $\tilde{d}_\tau(\lambda) := \frac{\tilde{u}_\tau(\lambda) - \tilde{u}_0(\lambda)}{\tau}$, we have $\|\tilde{d}_\tau(\lambda) - d_\tau(\lambda)\| \le \frac{\varepsilon_{0,K} + \varepsilon_{\tau,K}}{\tau}$, and, for every absolutely continuous $\gamma : [0,1] \to K$,*

$$\left|\Phi(\gamma(1)) - \Phi(\gamma(0)) - \int_0^1 \langle \tilde{d}_\tau(\gamma(t)), \dot{\gamma}(t)\rangle\, dt\right| \le 2C_K\tau^\theta + \frac{\varepsilon_{0,K} + \varepsilon_{\tau,K}}{\tau}\,\mathrm{Len}(\gamma). \tag{16}$$

Hence τ is not a nuisance parameter: large τ increases modeling bias, while very small τ magnifies solve and adjoint error. With a nearly uniform error envelope $\varepsilon_{0,K} + \varepsilon_{\tau,K} \le \bar{\varepsilon}_K$, balancing the two terms of this cruder bound gives $\tau \asymp \bar{\varepsilon}_K^{1/(\theta+1)}$ and pathwise error $O(\bar{\varepsilon}_K^{\theta/(\theta+1)})$, up to the path length and constants; with b_τ from Corollary 1 and $\theta < 1$, the balance is $\tau^\star = \Theta(\bar{\varepsilon}_K^{1-\theta})$ with error $O(\bar{\varepsilon}_K^\theta)$, and for $\theta = 1$ the largest $\tau < 1/(\kappa_K L_K)$ is best (Appendix C.8).

The direction computed by Algorithm 1 needs two further conditions.

Assumption 2 (Certified readouts at the shared point)**.** $\Lambda + \bar{\delta}\,\mathbb{B} \subseteq U$ ($\mathbb{B}$ the unit ball), and for $s \in \{0, \tau\}$ and $\lambda \in U$ there is a nonempty closed set $Z_s(\lambda)$ of exact certificates, for instance global lower solutions with qualified multipliers, whose readouts are valid value adjoints, $\mathrm{Adj}(F_s, \lambda, z) \in D_{v_s}(\lambda)$ for $z \in Z_s(\lambda)$. The readout $z \mapsto \mathrm{Adj}(F_s, \lambda, z)$ is B_s-Lipschitz, and the certificate $\tilde{z}_{s,t} = (w_s, \nu_s)$ returned at $\bar{\lambda}_t$ satisfies $\mathrm{dist}(\tilde{z}_{s,t}, Z_s(\bar{\lambda}_t)) \le \kappa_s R_{s,t}^{\alpha_s}$ for a computable residual $R_{s,t}$.

Assumption 3 (Smooth fixed-τ surrogate)**.** Δ_τ is differentiable on U with an L_τ-Lipschitz gradient.

Theorem 4 (Computed directions track the compatible gradient)**.** *Let Assumption 2 hold, and let $u_{s,t}$ and d_t be the readouts and direction computed by Algorithm 1 at iteration t. Almost surely, for every t: every exact readout at $\bar{\lambda}_t$ is the gradient, $\mathrm{Adj}(F_s, \bar{\lambda}_t, z) = \nabla v_s(\bar{\lambda}_t)$ for $z \in Z_s(\bar{\lambda}_t)$, and $\|u_{s,t} - \nabla v_s(\bar{\lambda}_t)\| \le \varepsilon_{s,t} := B_s\kappa_s R_{s,t}^{\alpha_s}$ for $s \in \{0, \tau\}$. If Assumption 3 also holds, then almost surely, for every t,*

$$\|d_t - \nabla\Delta_\tau(\lambda^t)\| \le \rho_t := L_\tau\delta_t\|q_t\| + \frac{\varepsilon_{0,t} + \varepsilon_{\tau,t}}{\tau}. \tag{17}$$

A small residual is not by itself a certificate: a stationary point of a nonconvex lower problem that is not globally optimal lies outside Z_s (strongly convex constants: Appendix C.9). Assumption 3 concerns Δ_τ, not G or V_τ separately (in Example 1 both values have a kink while Δ_τ is smooth).

Theorem 5 (Projected stationarity for smooth finite-tilt Ghost)**.** *Fix $\tau \in (0, \bar{\tau}]$ and let $\Lambda \subset \mathbb{R}^m$ be closed and convex. Suppose Δ_τ is differentiable near Λ, $\nabla\Delta_\tau$ is L_τ-Lipschitz on Λ, and $\Delta_\tau^{\inf} := \inf_{\lambda\in\Lambda}\Delta_\tau(\lambda) > -\infty$. With exact ghost directions $d_t = \nabla\Delta_\tau(\lambda^t)$, run $\lambda^{t+1} = \Pi_\Lambda(\lambda^t - \eta d_t)$ with $0 < \eta < 2/L_\tau$, and define $\mathcal{G}_{\eta,\tau}(\lambda) := \frac{1}{\eta}\big(\lambda - \Pi_\Lambda(\lambda - \eta\nabla\Delta_\tau(\lambda))\big)$. Then, for every $T \ge 1$,*

$$\min_{0\le t<T}\|\mathcal{G}_{\eta,\tau}(\lambda^t)\|^2 \le \frac{\Delta_\tau(\lambda^0) - \Delta_\tau^{\inf}}{\eta(1 - L_\tau\eta/2)T}. \tag{18}$$

If instead $\|d_t - \nabla\Delta_\tau(\lambda^t)\| \le \rho_t$ *and* $c = 1 - L_\tau\eta/2$, *then* $\min_{t<T}\|\mathcal{G}_{\eta,\tau}(\lambda^t)\|^2 \le 4(\Delta_\tau(\lambda^0) - \Delta_\tau^{\inf})/(\eta cT) + (4/(c^2T))\sum_{t<T}\rho_t^2$. *Under Assumptions 2–3, Algorithm 1 with* $\eta_t \equiv \eta$ *satisfies this almost surely with* ρ_t *from* (17), *and* $\delta_t \to 0$, $R_{s,t} \to 0$ *give* $\min_{t<T}\|\mathcal{G}_{\eta,\tau}(\lambda^t)\| \to 0$. *Under a projected PL inequality* $\frac{1}{2}\|\mathcal{G}_{\eta,\tau}\|^2 \ge \mu(\Delta_\tau - \inf_\Lambda \Delta_\tau)$ *with* $0 < 2\eta c\mu \le 1$ *and exact directions,* Δ_τ *contracts linearly and* $\Phi(\lambda^T) - \inf_\Lambda \Phi \le (1 - 2\eta c\mu)^T(\Delta_\tau(\lambda^0) - \inf_\Lambda \Delta_\tau) + b_\tau$ *for* $\Lambda \subseteq K$. *We state the full constants and nonsmooth stationarity qualification in Appendix C.10–C.11.*

Uniform closeness of values does not transfer stationarity, so we make no stationarity claim for Φ. Without Assumption 3, a zero projected compatible residual certifies only Clarke stationarity of Δ_τ; descent would need further structure, such as weak convexity and a proximal step.

6 Experiments

The experiments target the central oracle claim: Ghost should help when validation must choose among many valid training minimizers. An exact set-valued problem compares the optimistic value, finite-tilt gap, and biased selected-response gradient without approximation. We then test the accuracy–runtime tradeoff in SNLI hyper-cleaning under a shared-cache protocol, with both adjoints read at the unperturbed iterate; Section 6.2 adds a clean-validation protocol. Appendix D gives details, τ and inner-solve sweeps, and cancellation ablations.

6.1 Exact flat-valley nonuniqueness

We use the flat-valley lower objective

$$g(w; c, h) = \tfrac{1}{2}[c - h - w]_+^2 + \tfrac{1}{2}[w - c - h]_+^2, \qquad f_t(w) = \tfrac{1}{2}(w - t)^2, \tag{19}$$

with $c \in [-0.4, 1.4]$ and $h \in [0.15, 0.45]$. Every parameter has the continuum of minimizers $S(c,h) = [c-h, c+h]$. Hence $\Phi_t(c,h) = \min_{w\in S(c,h)} f_t(w)$ is realized by $\Pi_{[c-h,c+h]}(t)$, while a solver that selects the equally valid point $\Pi_{[c-h,c+h]}(a)$, $a < t$, can follow a different branch. We average 256 exact tasks with $t \sim \mathrm{Unif}[0.85, 1.15]$ and $a = t - \delta$, $\delta \sim \mathrm{Unif}[0.5, 0.9]$; directional cosines are measured only away from the zero-gradient plateau and projection kinks in Appendix D.3, where both values are C^1 and the pairing question of Section 4.2 does not arise.

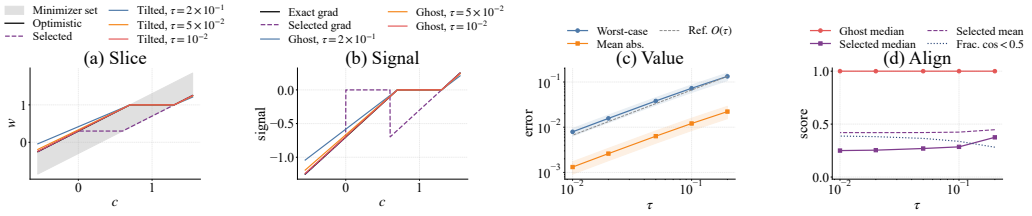


Figure 2: Exact flat-valley diagnostic with infinitely many lower-level minimizers. Left panels: a representative slice with validation target $t = 1$ and biased selector anchor $a = 0.3$, where a valid selected response follows the wrong branch. Right panels: averages over 256 randomized tasks. Δ_τ converges toward the optimistic value, Ghost keeps near-perfect directional alignment on the safe active region, and selected-response differentiation remains substantially degraded.

Figure 2 exposes the selection error. The selected response is a valid lower minimizer, but in the representative slice it remains pinned to the anchor-induced plateau after the optimistic selector has moved to the validation target. Over the randomized family, the value-gap surrogate has clean finite-tilt behavior: the average worst-case error drops from 1.33×10^{-1} at $\tau = 0.2$ to 7.92×10^{-3} at $\tau = 10^{-2}$, and the mean absolute error drops from 2.22×10^{-2} to 1.32×10^{-3}; the log–log slope 0.94 matches the $O(\tau)$ rate of Corollary 1 for this $\theta = 1/2$ valley, where Theorem 1 alone gives only $O(\tau^{1/2})$. Directionally, Ghost has average median cosine 1 throughout the sweep. The selected-response direction remains systematically misaligned, with median cosine between 0.25 and 0.38 and 28%–39% of the safe grid below cosine 0.5. This is the regime in which differentiating a stable-looking selected branch can be precisely wrong; the value gap instead differentiates the optimistic tie-break.

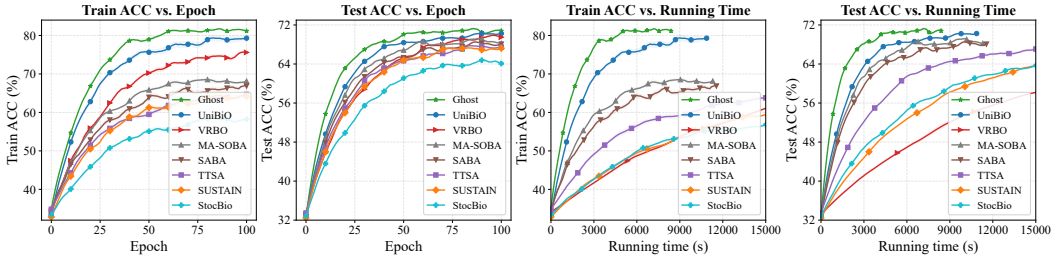


Figure 3: Mean of five SNLI runs (shared-cache protocol); time panels cropped at 15,000 s.

Method	Final Test (%, ↑)	Time (h, ↓)	Final Train (%, ↑)
Ghost	**70.80 ± 0.22**	**2.33 ± 0.03**	**81.15 ± 0.25**
UniBiO	70.25 ± 0.09	3.02 ± 0.09	79.30 ± 0.42
VRBO	69.50 ± 0.09	14.84 ± 0.24	75.60 ± 0.16
MA-SOBA	68.22 ± 0.18	3.14 ± 0.04	68.15 ± 0.32
SABA	68.07 ± 0.20	3.21 ± 0.06	66.85 ± 0.45
TTSA	67.44 ± 0.12	5.21 ± 0.17	64.68 ± 0.81
SUSTAIN	67.23 ± 0.21	9.31 ± 0.11	64.10 ± 0.20
StocBio	64.09 ± 0.28	5.96 ± 0.09	58.20 ± 0.17

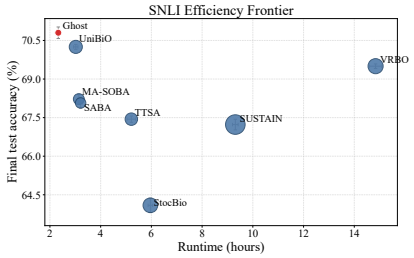


Figure 4: Full-budget SNLI results over five runs and accuracy–runtime frontier (shared-cache protocol); ± values are standard deviations. Accuracies and runtimes are final values after 100 epochs, not best-checkpoint values; training accuracy, on the noisy cache, is only a diagnostic.

6.2 Noisy-label hyper-cleaning on SNLI

We evaluate SNLI (Bowman et al., 2015) hyper-cleaning following Wu et al. (2026): each training label is independently corrupted with probability 0.2, and each example has a weight $\sigma(\lambda_i)$. Methods share one codebase, cached splits, batching, evaluation code, a 100-epoch horizon, and one A100 GPU. The training cache is noisy and the test cache clean; validation uses the shared comparison cache, not a regenerated strictly clean split. Baselines are StocBiO (Ji et al., 2021), TTSA (Hong et al., 2023), SABA (Dagréou et al., 2022), MA-SOBA (Chen et al., 2024b), SUSTAIN (Khanduri et al., 2021), VRBO (Yang et al., 2021), and UniBiO (Wu et al., 2026). Ghost is only approximate here, and the certificate assumptions of Section 5 are unverified. Both adjoints are read at the unperturbed iterate, with the tilted solve warm-started from the base: these runs test the value-gap direction, not Algorithm 1's shared-point readout. Final-run wall-clock times exclude search and preprocessing.

In Figures 3–4, Ghost reaches the highest mean final test accuracy ($70.80 \pm 0.22\%$) and the shortest runtime (2.33 ± 0.03 hours); UniBiO, the closest competitor, is 0.55 points lower and about $1.30\times$ slower. The frontier is consistent with the finite-compute tradeoff of Theorem 3, whose assumptions are unverified here. We read this as evidence for practical usefulness, not dataset-wide dominance.

Further evidence. At the kink of Appendix D.8, shared readouts avoid the ascent of arbitrary-branch and independent ones. Under a clean-validation protocol with HG-MS and SQ-G (Table 6), compatible Ghost has the highest mean accuracy, within seed spread (no significance claimed).

7 Conclusion and Limitations

Ghost recasts optimistic bilevel learning as validation-guided selection rather than solver-dependent differentiation. By comparing the lower training value with a validation-tilted value, the scaled gap Δ_τ exposes the optimistic tie-break while avoiding derivatives of a set-valued argmin map. Under a certificate error bound, compatible readouts connect this route to projected stationarity for the smooth fixed-τ surrogate. The flat valley illustrates optimistic alignment; shared-cache SNLI runs show a favorable accuracy–runtime frontier, with supporting clean-validation results. The method remains tilt- and solve-accuracy-sensitive. Guarantees do not transfer stationarity to Φ, certificate assumptions are unverified for deep networks, and optimistic selection can overfit a small validation set. Ghost thus offers a first-order interface for optimistic bilevel learning.

AI USE STATEMENT

The core of this paper was developed and written by the authors without generative AI: the value-gap formulation (Section 3), the Ghost oracle and Algorithm 1, Theorems 1–3 and 5 with their proofs, all experiments (Section 6, Appendix D), and the interpretation of results. The experiment code, including data processing and plotting, was written with the assistance of Claude (Anthropic). Generative AI was used mainly for material added in revision. From arguments developed by the authors, Claude Opus 5.5 (Anthropic) drafted Definition 2 and Proposition 2; Corollary 1, Assumption 2, and Theorem 4; their proofs (Appendices C.4, C.6, C.9) with Example 1; the scripts for the exponent and pairing checks (Appendices C.4 and D.8); the text of Section 6.2 and Appendix D.8; the additions to Section 2 and Table 1; and these statements. It also edited the text throughout for the ICLR format and page limit, helped check the proof steps and reference metadata, and prepared the anonymized code release. GPT-6 Astra (OpenAI) planned the revision, suggested the pairing check, drafted a compression edit, and reviewed drafts. Generative AI did not generate, clean, or relabel data, and was not used to interpret results. An author checked each AI-drafted proof line by line; the authors ran both scripts, inspected the anonymized code release, and revised the edited text. The authors take full responsibility for the paper, including text, claims, and code produced with the aid of generative AI.

ETHICS STATEMENT

The work is methodological. Validation-driven reweighting can inherit biases, label artifacts, or privacy issues from the validation set, and an optimistic target can overfit a small or unrepresentative validation split; learned example weights should be read as optimization diagnostics, not as judgments about individual examples. No new data were collected and no human-subject study was conducted; SNLI is used under its Creative Commons Attribution–ShareAlike 4.0 license, the GloVe vectors under the Public Domain Dedication and License, and the UCI handwritten digits data under Creative Commons Attribution 4.0.

REPRODUCIBILITY STATEMENT

Algorithm 1 and the readout (12) specify the method; Assumptions 1–3 state the conditions of every result, and Appendix C contains the proofs. Appendix D records the SNLI protocols, hyperparameters, and metrics and keeps the shared-cache and clean-validation protocols apart. The supplementary material contains anonymized code for the shared-cache SNLI runs (Ghost and all baselines in Figure 4, with data caching and result aggregation) and the standard-library scripts that print every number of the pairing check (Appendix D.8) and the exponent check (Appendix C.4). Code and per-seed logs of the clean-validation runs (Appendix D.8) are not included.

REFERENCES

Jacob A. Abernethy, Robert Schapire, and Umar Syed. Lexicographic optimization: Algorithms and stability. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2503–2511. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/abernethy24a.html.

E. Alpaydin and C. Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998.

Michael Arbel and Julien Mairal. Non-convex bilevel games with critical point selection maps. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 8013–8026. Curran Associates, Inc., 2022. doi: 10.52202/068431-0582. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/349a45f211fb1b3850da1ccd829e869e-Paper-Conference.pdf.

Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-Lopez, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Ad-*

vances in Neural Information Processing Systems, volume 35, pp. 5230–5242. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/228b9279ecf9bbafe582406850c57115-Paper-Conference.pdf.

Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188(1):19–51, 2021.

Jérôme Bolte, Edouard Pauwels, and Cheik Traoré. The adjoint state method for parametric definable optimization without smoothness or uniqueness. *arXiv preprint arXiv:2603.26503*, 2026.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642. Association for Computational Linguistics, 2015. URL https://aclanthology.org/D15-1075/.

He Chen, Jiajin Li, and Anthony Man-Cho So. Set smoothness unlocks Clarke hyper-stationarity in bilevel optimization. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=QBnfYm6Naa.

Lesi Chen, Jing Xu, and Jingzhao Zhang. On finding small hyper-gradients in bilevel optimization: Hardness results and improved analysis. In Shipra Agrawal and Aaron Roth (eds.), *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 947–980. PMLR, 30 Jun–03 Jul 2024a. URL https://proceedings.mlr.press/v247/chen24a.html.

Xuxing Chen, Tesi Xiao, and Krishnakumar Balasubramanian. Optimal algorithms for stochastic bilevel optimization under relaxed smoothness conditions. *Journal of Machine Learning Research*, 25(151):1–51, 2024b. URL http://jmlr.org/papers/v25/23-1323.html.

Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

Mathieu Dagréou, Pierre Ablin, Samuel Vaiter, and Thomas Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26698–26710. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/aa84ec1ac3f5fdcf77bce2c22705ab77-Paper-Conference.pdf.

Stephan Dempe, Boris S. Mordukhovich, and Alain B. Zemkoho. Two-level value function approach to non-smooth optimistic and pessimistic bilevel programs. *Optimization*, 68(2-3):433–455, 2019. doi: 10.1080/02331934.2018.1543294. URL https://doi.org/10.1080/02331934.2018.1543294.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1165–1173. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/franceschi17a.html.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1568–1577. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/franceschi18a.html.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023. doi: 10.1137/20M1387341. URL https://doi.org/10.1137/20M1387341.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference*

on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pp. 4882–4892. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ji21c.html.

Xiaotian Jiang, Jiaxiang Li, Jiawen Bi, Mingyi Hong, and Shuzhong Zhang. A correspondence-driven approach for bilevel decision-making with nonconvex lower-level problems, 2026. URL https://arxiv.org/abs/2509.01148.

Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 30271–30283. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/fe2b421b8b5f0e7c355ace66a9fe0206-Paper.pdf.

Risheng Liu, Xuan Liu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A value-function-based interior-point method for non-convex bi-level optimization. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6882–6892. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/liu21o.html.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1540–1552. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/lorraine20a.html.

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

Saeed Masiha, Zebang Shen, Negar Kiyavash, and Niao He. Select-then-differentiate: Solving bilevel optimization with manifold lower-level solution sets, 2026. URL https://arxiv.org/abs/2605.09209.

Abdellatif Moudafi and Paul-Emile Maingé. Towards viscosity approximations of hierarchical fixed-point problems. *Fixed Point Theory and Applications*, 2006(1):95453, 2006.

Jiří V Outrata. On the numerical solution of a class of Stackelberg problems. *Zeitschrift für Operations Research*, 34(4):255–277, 1990.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543. Association for Computational Linguistics, 2014. URL https://aclanthology.org/D14-1162/.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4334–4343. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ren18a.html.

Benjamin Scellier and Yoshua Bengio. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in Computational Neuroscience*, 11:24, 2017. ISSN 1662-5188. doi: 10.3389/fncom.2017.00024. URL https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2017.00024.

Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1723–1732. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/shaban19a.html.

Mattia Solla and Johannes O Royset. Optimistic bilevel optimization with composite lower-level problem. *arXiv preprint arXiv:2602.05417*, 2026.

Paul Vicol, Jonathan P Lorraine, Fabian Pedregosa, David Duvenaud, and Roger B Grosse. On implicit bias in overparameterized bilevel optimization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22234–22259. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/vicol22a.html.

Ralph A. Willoughby. Solutions of ill-posed problems (a. n. tikhonov and v. y. arsenin). *SIAM Review*, 21(2):266–267, 1979. doi: 10.1137/1021044. URL https://doi.org/10.1137/1021044.

Yuman Wu, Xiaochuan Gong, Jie Hao, and Mingrui Liu. Bilevel optimization with lower-level uniform convexity: Theory and algorithm. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=dJgb3ngAvT.

Yunhai Xiao, Anqi Liu, Peili Li, and Yanyun Ding. ADMM-based bilevel descent aggregation algorithm for sparse hyperparameter selection, 2026. URL https://arxiv.org/abs/2603.09546.

Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 13670–13682. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/71cc107d2e0408e60a3d3c44f47507bd-Paper.pdf.

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

Nicolas Zucchet and João Sacramento. Beyond backpropagation: Bilevel optimization through implicit differentiation and equilibrium propagation. *Neural Computation*, 34(12):2309–2346, 11 2022. ISSN 0899-7667. doi: 10.1162/neco_a_01547. URL https://doi.org/10.1162/neco_a_01547.

A Supplementary Details for Related Work

Table 1 compares the methods discussed in Section 2.

Table 1: Positioning of Ghost relative to bilevel hypergradient methods. ✓ means directly addressed; × means not targeted; △ means partial, conditional, or outside the main guarantee. Abbreviations: Cmp. = compared experimentally or theoretically here; NLL = nonunique lower level; Opt. sel. = optimistic/selection target; SRD = selected-response differentiation; V/C = value/conservative.

Method / line of work	Cmp.	Stoch.	NLL	Opt. sel.	No SRD	V/C oracle	Main restriction relative to Ghost
Unrolled / iterative differentiation (Franceschi et al., 2017; 2018; Shaban et al., 2019)	×	△	×	×	×	×	Differentiates a training trajectory; truncated variants trade memory for biased hypergradients.
Implicit differentiation (Lorraine et al., 2020; Blondel et al., 2022)	×	△	×	×	×	×	Requires differentiable optimality equations and a stable, effectively single-valued response.
Meta-reweighting (Ren et al., 2018)	×	✓	×	×	×	×	Uses validation data to learn weights, but usually through one-step or selected-trajectory meta-gradients.
StocBiO (Ji et al., 2021)	✓	✓	×	×	×	×	Designed for smooth nonconvex–strongly-convex bilevel problems with implicit selected-response hypergradients.
TTSA (Hong et al., 2023)	✓	✓	×	×	×	×	Two-timescale stochastic approximation under strongly-convex lower-level structure.
SUSTAIN (Khanduri et al., 2021)	✓	✓	×	×	×	×	Single-timescale double momentum, but still assumes a smooth selected-response regime.
VRBO (Yang et al., 2021)	✓	✓	×	×	×	×	Variance-reduced hypergradient estimator; relies on the standard implicit-gradient template.
SABA / SOBA (Dagréou et al., 2022)	✓	✓	×	×	×	×	Tracks lower solution, linear-system variable, and outer variable jointly; still selected-response based.
MA-SOBA (Chen et al., 2024b)	✓	✓	×	×	×	×	Hessian-inversion-free and weaker smoothness requirements, but not designed for nonunique optimistic selection.
UniBiO (Wu et al., 2026)	✓	✓	×	×	×	×	Weakens strong convexity to lower-level uniform convexity; still rules out true lower-level multiplicity.

B Supplementary Details for the Ghost Hypergradient Oracle

This section expands Section 4. The main paper presents the oracle through the value-function interface

$$u_s(\lambda) \in D_{v_s}(\lambda), \qquad d_\tau(\lambda) = \frac{u_\tau(\lambda) - u_0(\lambda)}{\tau}, \qquad s \in \{0, \tau\}.$$

Here we spell out what this interface means for constrained and nonsmooth lower-level solvers, how the adjoint output is formed, and how the optional bundle variant is implemented. The theoretical oracle statements in the main text remain single-world statements; the bundle variant below is included as a practical robustness and ambiguity diagnostic.

Table 1 (continued).

Method / line of work	Cmp.	Stoch.	NLL	Opt. sel.	No SRD	V/C oracle	Main restriction relative to Ghost
Hardness / PL tractability (Chen et al., 2024a)	×	△	△	△	△	×	Clarifies intractability and tractable PL regimes, but is not a practical optimistic value-gap oracle.
Superquantile-Gibbs selection (Masiha et al., 2025)	✓	△	✓	✓	✓	×	Pessimistic formulation (optimistic analogous); superquantile–Gibbs relaxation with Langevin sampling and zeroth-order steps to Goldstein stationarity; compared under the clean-validation protocol.
Select-then-differentiate (HG-MS) (Masiha et al., 2026)	✓	△	✓	✓	×	×	Explicit optimistic selection plus a pseudoinverse hypergradient under local PL on solution manifolds; reaches stationarity of the optimistic objective, a stronger target than ours.
Algorithm-defined targets (Vicol et al., 2022; Arbel & Mairal, 2022; Jiang et al., 2026)	×	△	✓	×	×	×	Target the solution returned by a training procedure or selection map, the right object when training dynamics are part of the model.
Value-function penalties (Outrata, 1990; Ye & Zhu, 1995; Dempe et al., 2019; Liu et al., 2021)	×	×	✓	✓	✓	△	Δ_τ is the marginal function of such a penalty; these methods solve penalized or constrained reformulations rather than read value adjoints.
Equilibrium propagation / two-phase (Scellier & Bengio, 2017; Zucchet & Sacramento, 2022)	×	△	×	×	✓	△	Same free/nudged two-solve structure that Ghost uses, developed for energy-based models with equilibria and reviewed for bilevel optimization.
Set-smoothness / Clarke stationarity (Chen et al., 2025)	×	×	✓	✓	✓	×	Targets nonsmooth hyper-stationarity via set smoothness and zeroth-order schemes rather than scalable adjoint differences.
Optimistic composite regularization (Solla & Royset, 2026)	×	×	✓	✓	△	×	Uses double regularization and gradient sampling for composite convex lower-level problems.
Ghost, this work	✓	△	✓	✓	✓	✓	Requires two value problems read at one shared outer point, certified value adjoints, and a finite-tilt bias–accuracy tradeoff in τ; guarantees concern the fixed-τ surrogate.

B.1 Two parametric value problems and solver certificates

For each tilt level $s \in \{0, \tau\}$, define

$$F_s(w, \lambda) := \mathcal{L}_{\mathrm{tr}}(w, \lambda) + s\mathcal{L}_{\mathrm{val}}(w, \lambda), \qquad v_s(\lambda) := \inf_{w \in \mathcal{W}(\lambda)} F_s(w, \lambda). \tag{20}$$

Thus $v_0 = G$ and $v_\tau = V_\tau$. When the feasible set is described explicitly, we write

$$\mathcal{W}(\lambda) = \{w \in \mathbb{R}^d : \ c(w, \lambda) \le 0, \ h(w, \lambda) = 0\}, \tag{21}$$

where c and h collect inequality and equality constraints. Nonsmooth regularizers may either be included in F_s through conservative-field calculus, or represented through an indicator function whose normal-cone term is part of the optimality certificate.

At a fixed outer state λ, the solver returns one certificate

$$\nu_s = (w_s, \alpha_s, \beta_s, q_s),$$

where w_s is a primal minimizer or approximate minimizer, α_s and β_s are multipliers for the explicit constraints in (21), and q_s denotes any additional stationarity data used by a proximal, projected, active-set, or nonsmooth solver. In the exact smooth constrained case, a certificate satisfies

$$\begin{aligned} &c(w_s, \lambda) \le 0, \qquad h(w_s, \lambda) = 0, \qquad \alpha_s \ge 0, \qquad \alpha_{s,j} c_j(w_s, \lambda) = 0, \\ &0 = \nabla_w F_s(w_s, \lambda) + J_w c(w_s, \lambda)^\top \alpha_s + J_w h(w_s, \lambda)^\top \beta_s. \end{aligned} \tag{22}$$

The nonsmooth version replaces $\nabla_w F_s$ by an admissible conservative-field or subdifferential element. This is the only point at which the details of the lower-level solver enter the oracle.

B.2 Conservative adjoint output set

Let $\Xi_s(w, \lambda) \subseteq \mathbb{R}^d \times \mathbb{R}^m$ be a joint conservative field for the map $(w, \lambda) \mapsto F_s(w, \lambda)$. In the smooth case,

$$\Xi_s(w, \lambda) = \{(\nabla_w F_s(w, \lambda), \nabla_\lambda F_s(w, \lambda))\}.$$

For locally Lipschitz definable objectives, Ξ_s can be chosen using conservative calculus and the adjoint value-function construction invoked in the main text (Bolte et al., 2026). To avoid treating arbitrary multiplier choices as value adjoints, we first record the feasibility part of an exact constrained certificate:

$$\mathcal{K}_s(\lambda) := \Big\{(w, \alpha, \beta) : c(w, \lambda) \le 0, \ h(w, \lambda) = 0, \ \alpha \ge 0, \ \alpha_j c_j(w, \lambda) = 0 \ \forall j\Big\}. \tag{23}$$

Given an admissible certificate $(w_s, \alpha_s, \beta_s, q_s)$ whose primal–dual part lies in $\mathcal{K}_s(\lambda)$, define the adjoint output set

$$\begin{aligned} \mathcal{B}_s(\lambda; w_s, \alpha_s, \beta_s) := \Big\{ &z \in \mathbb{R}^m : \ \exists (\xi_{w,s}, \xi_{\lambda,s}) \in \Xi_s(w_s, \lambda) \text{ such that} \\ &z = \xi_{\lambda,s} + J_\lambda c(w_s, \lambda)^\top \alpha_s + J_\lambda h(w_s, \lambda)^\top \beta_s, \\ &0 \in \xi_{w,s} + J_w c(w_s, \lambda)^\top \alpha_s + J_w h(w_s, \lambda)^\top \beta_s \Big\}, \end{aligned} \tag{24}$$

$$\mathcal{A}_s(\lambda; w_s, \alpha_s, \beta_s) := \begin{cases} \mathcal{B}_s(\lambda; w_s, \alpha_s, \beta_s), & (w_s, \alpha_s, \beta_s) \in \mathcal{K}_s(\lambda), \\ \emptyset, & \text{otherwise.} \end{cases}$$

The stationarity line in (24) is the lower-level KKT equation written with a conservative field element; the second line gates this candidate set by certificate feasibility, while (23) records primal feasibility, dual feasibility, and complementarity. Under the local boundedness and qualification assumptions used in the main theory, admissible exact certificates selected by the conservative-adjoint construction produce outputs

$$u_s(\lambda) \in \mathcal{A}_s(\lambda; w_s, \alpha_s, \beta_s) \subseteq D_{v_s}(\lambda), \tag{25}$$

where D_{v_s} is a conservative field for the value function v_s. This is an oracle-availability assumption about the returned certificate, not a claim that every algebraic multiplier satisfying stationarity is automatically a valid conservative value adjoint; in particular, a KKT point of a nonconvex lower problem that is not globally optimal does not certify the value (Assumption 2).

The unconstrained smooth envelope formula is recovered by taking $c = h = 0$ and $\Xi_s = \{(\nabla_w F_s, \nabla_\lambda F_s)\}$. If w_s is a global minimizer (so that $\nabla_w F_s(w_s, \lambda) = 0$), then

$$u_s(\lambda) = \nabla_\lambda F_s(w_s, \lambda). \tag{26}$$

The constrained smooth formula is similarly the partial λ-gradient of the Lagrangian evaluated at a KKT point,

$$u_s(\lambda) = \nabla_\lambda F_s(w_s, \lambda) + J_\lambda c(w_s, \lambda)^\top \alpha_s + J_\lambda h(w_s, \lambda)^\top \beta_s. \tag{27}$$

Paired selections. The ghost direction uses paired base and tilted adjoints:

$$d_\tau(\lambda) = \frac{u_\tau(\lambda) - u_0(\lambda)}{\tau}, \qquad u_\tau(\lambda) \in D_{V_\tau}(\lambda), \quad u_0(\lambda) \in D_G(\lambda). \tag{28}$$

A pair corresponds to one "world": one solver initialization, one base certificate, and the tilted certificate warm-started from that base solve. The pathwise conservative-oracle theorem only requires a measurable valid base adjoint and a measurable valid tilted adjoint. A pointwise direction additionally requires the pair to be compatible (Definition 2); Algorithm 1 obtains this by solving and reading out both problems at one shared random outer point. Warm-starting the tilted solve within the same world reduces cost and solver variance but does not by itself make the pair compatible.

B.3 Hyper-cleaning specialization

For the sample-reweighting model

$$\mathcal{L}_{\mathrm{tr}}(w, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \sigma(\lambda_i) \ell(w; \zeta_i) + R(w) + I_{\mathcal{C}}(w), \qquad \mathcal{L}_{\mathrm{val}}(w, \lambda) = \mathcal{L}_{\mathrm{val}}(w), \tag{29}$$

the explicit dependence on λ_i is only through the weight $\sigma(\lambda_i)$. Consequently, the explicit component of the value adjoint contains

$$[u_s(\lambda)]_i^{\exp} = \frac{1}{n} \sigma'(\lambda_i) \ell(w_s; \zeta_i), \qquad s \in \{0, \tau\}. \tag{30}$$

The explicit ghost component is

$$[d_\tau(\lambda)]_i^{\exp} = \frac{1}{n} \sigma'(\lambda_i) \frac{\ell(w_\tau; \zeta_i) - \ell(w_0; \zeta_i)}{\tau}. \tag{31}$$

Equation (31) is useful for interpreting sample weights. If example i has a positive explicit ghost component, then the validation-tilted world incurs larger training loss on that example than the base world; a descent step therefore lowers its weight. If the component is negative, the validation-tilted world fits the example more easily, so the update tends to raise its weight. The full adjoint may also contain constraint, regularizer, or explicit validation terms, but (31) is the dominant sample-wise diagnostic in the hyper-cleaning experiments.

If the validation loss depends explicitly on λ, the only change is the additional term

$$[u_s(\lambda)]_i^{\exp} = \frac{1}{n} \sigma'(\lambda_i) \ell(w_s; \zeta_i) + s\, \partial_{\lambda_i} \mathcal{L}_{\mathrm{val}}(w_s, \lambda),$$

with the same difference quotient used to form d_τ.

Lemma 1 (Global-minimizer readouts are gradients where the value is differentiable)**.** *Let $\mathcal{W}(\lambda) \equiv \mathcal{W}$ not depend on λ, let $F_s(w, \cdot)$ be differentiable for every $w \in \mathcal{W}$, and let $w^\star \in \arg\min_{w \in \mathcal{W}} F_s(w, \bar{\lambda})$. If v_s is differentiable at $\bar{\lambda}$, then $\nabla_\lambda F_s(w^\star, \bar{\lambda}) = \nabla v_s(\bar{\lambda})$.*

Proof. The function $\varphi(\lambda) = F_s(w^\star, \lambda) - v_s(\lambda)$ is nonnegative, vanishes at $\bar{\lambda}$, and is differentiable there, so $\bar{\lambda}$ is an unconstrained minimizer of φ and $\nabla\varphi(\bar{\lambda}) = 0$. □

For the sample-reweighting model the lemma applies with $\mathcal{W} = \mathcal{C}$, whatever the regularizer R, because λ enters only through the smooth weights $\sigma(\lambda_i)$. Algorithm 1 evaluates (30)–(31) at the shared point $\bar{\lambda}_t$, which lands almost surely where both values are differentiable; there, every pair of global-minimizer readouts is the gradient pair and hence compatible, even when the lower problems have many global minimizers. For approximate or merely stationary solutions, the readout error is what Theorem 4 controls.

Algorithm 2: Ghost oracle with optional bundle.

Input: outer iterate λ^t, shared perturbed point $\bar{\lambda}_t$ drawn as in Algorithm 1, tilt $\tau > 0$, number of worlds $K \geq 1$, inner solver SOLVE, adjoint routine ADJ, aggregation rule Agg, stepsize η_t.

1. For each world $k = 1, \ldots, K$:
   (a) Choose an initialization or solver perturbation ω_k.
   (b) $(w_0^{(k)}, \nu_0^{(k)}) \leftarrow \textsc{Solve}(F_0, \bar{\lambda}_t; \omega_k)$.
   (c) $u_0^{(k)} \leftarrow \textsc{Adj}(F_0, \bar{\lambda}_t, w_0^{(k)}, \nu_0^{(k)})$.
   (d) $(w_\tau^{(k)}, \nu_\tau^{(k)}) \leftarrow \textsc{Solve}(F_\tau, \bar{\lambda}_t; w_0^{(k)}, \nu_0^{(k)})$.
   (e) $u_\tau^{(k)} \leftarrow \textsc{Adj}(F_\tau, \bar{\lambda}_t, w_\tau^{(k)}, \nu_\tau^{(k)})$.
   (f) $d_\tau^{(k)} \leftarrow (u_\tau^{(k)} - u_0^{(k)})/\tau$.
2. Aggregate $\bar{d}_\tau \leftarrow \mathrm{Agg}\{d_\tau^{(k)}\}_{k=1}^K$.
3. Update $\lambda^{t+1} \leftarrow \Pi_\Lambda(\lambda^t - \eta_t \bar{d}_\tau)$.

B.4 SINGLE-WORLD AND BUNDLE IMPLEMENTATION

The main algorithm is the case $K = 1$ below. For highly nonunique lower levels, we optionally run K worlds from different initializations or small solver perturbations, then aggregate their ghost directions.

All worlds are solved and read out at the same shared point $\bar{\lambda}_t$; perturbing the model initialization ω_k is a different device from the outer perturbation and does not provide compatibility. The update in Algorithm 2 is the Euclidean projected step

$$\lambda^{t+1} = \Pi_\Lambda(\lambda^t - \eta_t \bar{d}_\tau). \tag{32}$$

For sample weights constrained to a box, (32) is usually sufficient. For simplex-constrained weights, an entropic mirror map gives a multiplicative-weights update; this extension is not covered by our analysis.

The bundle also provides an ambiguity diagnostic,

$$\mathfrak{a}_\tau(\lambda) := \frac{1}{K} \sum_{k=1}^{K} \big\| d_\tau^{(k)}(\lambda) - \bar{d}_\tau(\lambda) \big\|_2. \tag{33}$$

Small $\mathfrak{a}_\tau$ means that different lower-level worlds induce similar outer signals. Large $\mathfrak{a}_\tau$ means that the lower-level ambiguity has not been resolved by the chosen tilt and solver accuracy. In our implementation, the bundle is embarrassingly parallel across worlds; the tilted solve in each world is warm-started from that world's base certificate.

Status of the bundle variant. The pathwise conservative-oracle theorem applies to each single-world direction $d_\tau^{(k)}$ when the corresponding certificates are exact, and Theorem 4 applies to each world read at the shared point. An aggregated direction $\bar{d}_\tau$ should be interpreted as a robustness heuristic unless the aggregation rule is included in the field construction. A simple average can be covered by working with the convexified pointwise field, whereas a coordinatewise median is best viewed as an empirical robustification. For this reason, the main paper reports the single-world method as the theoretically justified oracle and treats bundle spread as a diagnostic for minima-selection ambiguity.

B.5 COST PROFILE AND PRACTICAL TOLERANCES

One outer iteration has cost

$$\mathrm{Cost}_{\mathrm{ghost}} \approx \mathrm{Cost}_{\mathrm{solve}}(F_0, \lambda) + \mathrm{Cost}_{\mathrm{solve}}(F_\tau, \lambda \mid w_0, \nu_0) + 2\,\mathrm{Cost}_{\mathrm{adj}}. \tag{34}$$

The second solve is warm-started from the base certificate because F_τ is a small perturbation of F_0. In solvers that already expose multipliers or stationarity residuals, the adjoint overhead is typically

small relative to the lower-level solve. The method therefore trades one additional warm-started value solve for avoiding backpropagation through a long inner trajectory.

The finite-tilt oracle has the usual bias–accuracy tension. Smaller τ makes Δ_τ closer to the limiting optimistic objective, but the difference quotient amplifies base and tilted adjoint errors by $1/\tau$. In practice we therefore use three safeguards:

1. solve the tilted problem from the base certificate rather than from a fresh random initialization (this saves computation; compatibility comes from the shared outer point);
2. keep primal, dual, and adjoint residuals at a scale commensurate with τ;
3. select τ from a small logarithmic grid using validation stability rather than driving τ to machine precision.

These choices are consistent with the inexact-oracle bound in the main theory: large τ creates modeling bias, while very small τ exposes solver noise.

C Supplementary Proofs for the Main Theory

This section first proves the pointwise value-gap representation from Section 3, then gives the detailed proofs for Section 5. The main-theory proofs separate the argument into four layers. First, a quantitative variational argument shows that the finite-tilt value gap approximates the optimistic objective under a local error bound. Second, conservative-chain-rule calculus shows that the difference of two value-function adjoints is an exact oracle for the finite-tilt surrogate. Third, a perturbation argument quantifies how inner-solve and adjoint errors are amplified by the difference quotient. Fourth, a standard projected-gradient argument gives an iteration-level stationarity guarantee for the finite-tilt surrogate in the smooth regime.

C.1 Notation and local regime

For a fixed outer parameter λ, recall

$$\begin{aligned} G(\lambda) &:= \inf_{w\in\mathcal{W}(\lambda)} \mathcal{L}_{\mathrm{tr}}(w,\lambda), & S(\lambda) &:= \arg\min_{w\in\mathcal{W}(\lambda)} \mathcal{L}_{\mathrm{tr}}(w,\lambda), \\ \Phi(\lambda) &:= \min_{w\in S(\lambda)} \mathcal{L}_{\mathrm{val}}(w,\lambda), & V_\tau(\lambda) &:= \inf_{w\in\mathcal{W}(\lambda)} \{\mathcal{L}_{\mathrm{tr}}(w,\lambda)+\tau\mathcal{L}_{\mathrm{val}}(w,\lambda)\}, \\ S_\tau(\lambda) &:= \arg\min_{w\in\mathcal{W}(\lambda)} \{\mathcal{L}_{\mathrm{tr}}(w,\lambda)+\tau\mathcal{L}_{\mathrm{val}}(w,\lambda)\}, & \Delta_\tau(\lambda) &:= \frac{V_\tau(\lambda)-G(\lambda)}{\tau}. \end{aligned} \tag{35}$$

C.2 Proof of the pointwise value-gap representation

We prove Proposition 1. The argument is pointwise in the outer variable and does not use the uniform error-bound assumptions needed later for rates.

Proof of Proposition 1. Fix λ. To simplify notation, write

$$\begin{aligned} f(w) &:= \mathcal{L}_{\mathrm{tr}}(w,\lambda), \qquad g(w) := \mathcal{L}_{\mathrm{val}}(w,\lambda), \qquad W := \mathcal{W}(\lambda), \\ G &:= G(\lambda), \qquad S := S(\lambda), \qquad \Phi := \Phi(\lambda). \end{aligned}$$

By compactness of S and continuity of g on a neighborhood of S, choose a validation-best training minimizer

$$w^\dagger \in S, \qquad g(w^\dagger) = \Phi.$$

For the upper bound, use $w^\dagger$ as a feasible point in the tilted problem:

$$V_\tau(\lambda) \le f(w^\dagger) + \tau g(w^\dagger) = G + \tau\Phi.$$

Therefore

$$\limsup_{\tau\downarrow 0} \frac{V_\tau(\lambda)-G}{\tau} \le \Phi. \tag{36}$$

For the lower bound, let w_τ be an exact minimizer of the tilted problem for sufficiently small $\tau > 0$. Since validation is bounded below on W, set

$$\underline{m} := \inf_{w \in W} g(w) > -\infty.$$

Optimality of w_τ and feasibility of $w^\dagger$ give

$$f(w_\tau) + \tau g(w_\tau) \le G + \tau\Phi.$$

Rearranging yields

$$0 \le f(w_\tau) - G \le \tau\big(\Phi - g(w_\tau)\big) \le \tau(\Phi - \underline{m}). \tag{37}$$

Hence, for all sufficiently small τ, the tilted minimizer lies in the compact near-level set

$$\mathcal{N}_\lambda = \{w \in W : f(w) \le G + \varepsilon_0\}.$$

Now take any sequence $\tau_k \downarrow 0$. After passing to a subsequence, compactness of $\mathcal{N}_\lambda$ gives $w_{\tau_k} \to \bar{w}$ with $\bar{w} \in \mathcal{N}_\lambda \subset W$. Lower semicontinuity of f on $\mathcal{N}_\lambda$ and (37) imply

$$f(\bar{w}) \le \liminf_{k\to\infty} f(w_{\tau_k}) \le G.$$

Because G is the infimum of f over W and $\bar{w} \in W$, we have $\bar{w} \in S$.

Since g is continuous on a neighborhood of S and $w_{\tau_k} \to \bar{w} \in S$, we also have $g(w_{\tau_k}) \to g(\bar{w})$. Moreover,

$$\frac{V_{\tau_k}(\lambda) - G}{\tau_k} = \frac{f(w_{\tau_k}) - G}{\tau_k} + g(w_{\tau_k}) \ge g(w_{\tau_k}).$$

Taking the limit inferior along the subsequence gives

$$\liminf_{k\to\infty} \frac{V_{\tau_k}(\lambda) - G}{\tau_k} \ge g(\bar{w}) \ge \min_{w\in S} g(w) = \Phi.$$

The original sequence was arbitrary, so

$$\liminf_{\tau\downarrow 0} \frac{V_\tau(\lambda) - G}{\tau} \ge \Phi. \tag{38}$$

Combining (36) and (38) proves

$$\frac{V_\tau(\lambda) - G(\lambda)}{\tau} \to \Phi(\lambda),$$

which is equivalent to

$$V_\tau(\lambda) = G(\lambda) + \tau\Phi(\lambda) + o(\tau).$$

□

The compactness used above is compactness of the near-level set itself, not merely containment of that set in a compact ambient set. Equivalently, one can encode feasibility by the extended-valued objective $f + I_W$ and require the corresponding near-level set to be compact; this is the formulation that rules out limits escaping the feasible set.

All statements below are local on a compact outer set $K \subset \Lambda$. We restate the main assumptions in a slightly expanded form to make the proof self-contained.

Assumption 4 (Supplemental local value-gap regime)**.** There exist $\bar{\tau} > 0$, a compact set $\mathcal{N}_K$ in the lower-level space, constants $L_K, \kappa_K, M_K > 0$, a constant $m_K \in \mathbb{R}$, and an exponent $\theta \in (0, 1]$ such that the following hold for all $\lambda \in K$ and all $\tau \in (0, \bar{\tau}]$.

1. The sets $S(\lambda)$ and $S_\tau(\lambda)$ are nonempty, compact subsets of $\mathcal{W}(\lambda) \cap \mathcal{N}_K$. The minimum defining $\Phi(\lambda)$ is attained.
2. The map $w \mapsto \mathcal{L}_{\mathrm{val}}(w, \lambda)$ is L_K-Lipschitz on $\mathcal{N}_K$.
3. The validation loss is bounded below on the tube by m_K, and the validation-best training minimizer is uniformly bounded above relative to this lower bound:

$$\mathcal{L}_{\mathrm{val}}(w, \lambda) \ge m_K \quad \forall w \in \mathcal{N}_K, \qquad \Phi(\lambda) - m_K \le M_K. \tag{39}$$

4. The lower-level training objective satisfies the local error bound on feasible points

$$\operatorname{dist}(w, S(\lambda)) \le \kappa_K \big(\mathcal{L}_{\mathrm{tr}}(w,\lambda) - G(\lambda)\big)_+^{\theta} \qquad \forall w \in \mathcal{W}(\lambda) \cap \mathcal{N}_K. \tag{40}$$

Equivalently, feasibility may be absorbed into the extended-valued training loss $\mathcal{L}_{\mathrm{tr}}(\cdot,\lambda) + I_{\mathcal{W}(\lambda)}$.

The compact-tube assumption is local. It is meant to rule out tilted minimizers escaping to distant parts of the lower-level landscape as $\tau \downarrow 0$. The error bound in (40) is imposed only on feasible lower-level points, which is all the proof uses. It is the quantitative step that turns the pointwise vanishing-tilt identity into a rate. When the lower-level valley is sharp, θ is close to one; flatter valleys correspond to smaller θ.

For the oracle theorem we also use the following conservative-field assumption, matching the main text.

Assumption 5 (Supplemental conservative-adjoint regime)**.** For every fixed $\tau \in (0, \bar{\tau}]$, there are locally bounded conservative fields D_G and D_{V_τ} for G and V_τ on a neighborhood of K. That is, for every absolutely continuous curve $\gamma : [0,1] \to K$ and every measurable pair of selections

$$u_0(t) \in D_G(\gamma(t)), \qquad u_\tau(t) \in D_{V_\tau}(\gamma(t)) \qquad \text{for a.e. } t,$$

one has the chain-rule identities

$$\begin{aligned} G(\gamma(1)) - G(\gamma(0)) &= \int_0^1 \langle u_0(t), \dot{\gamma}(t)\rangle \, dt, \\ V_\tau(\gamma(1)) - V_\tau(\gamma(0)) &= \int_0^1 \langle u_\tau(t), \dot{\gamma}(t)\rangle \, dt. \end{aligned} \tag{41}$$

In the solver-based interpretation, D_G and D_{V_τ} are produced by the conservative adjoint construction described in Section B. The theory below only needs the chain rules in (41); it does not depend on a particular lower-level algorithm.

C.3 Proof of the finite-tilt approximation theorem

The proof of Theorem 1 rests on three elementary facts. The first says that the finite-tilt value gap cannot exceed the optimistic objective, because the tilted problem may always test the validation-best exact training minimizer.

Lemma 2 (Upper sandwich)**.** *Under Assumption 4, for every $\lambda \in K$ and every $\tau \in (0, \bar{\tau}]$,*

$$\Delta_\tau(\lambda) \le \Phi(\lambda). \tag{42}$$

Proof. Fix $\lambda \in K$. Since $S(\lambda)$ is nonempty and contained in the compact tube, and since $\Phi(\lambda)$ is assumed attained on $S(\lambda)$, choose

$$w_\star \in \arg\min_{w \in S(\lambda)} \mathcal{L}_{\mathrm{val}}(w,\lambda).$$

Then $\mathcal{L}_{\mathrm{tr}}(w_\star,\lambda) = G(\lambda)$ and $\mathcal{L}_{\mathrm{val}}(w_\star,\lambda) = \Phi(\lambda)$. Using $w_\star$ as a feasible comparison point for the tilted problem gives

$$V_\tau(\lambda) \le \mathcal{L}_{\mathrm{tr}}(w_\star,\lambda) + \tau \mathcal{L}_{\mathrm{val}}(w_\star,\lambda) = G(\lambda) + \tau\Phi(\lambda).$$

Subtracting $G(\lambda)$ and dividing by $\tau > 0$ gives (42). □

The second fact says that a tilted minimizer cannot have much training error. If it did, the validation-best exact training minimizer would beat it in the tilted objective.

Lemma 3 (Training suboptimality of tilted minimizers)**.** *Under Assumption 4, let $w_\tau \in S_\tau(\lambda)$. Then*

$$0 \le \mathcal{L}_{\mathrm{tr}}(w_\tau,\lambda) - G(\lambda) \le \tau M_K. \tag{43}$$

Proof. The lower bound follows directly from the definition of $G(\lambda)$ as the lower-level infimum. For the upper bound, reuse the validation-best training minimizer $w_\star$ from the proof of Lemma 2. Optimality of w_τ for the tilted problem gives

$$\mathcal{L}_{\mathrm{tr}}(w_\tau, \lambda) + \tau \mathcal{L}_{\mathrm{val}}(w_\tau, \lambda) \le \mathcal{L}_{\mathrm{tr}}(w_\star, \lambda) + \tau \mathcal{L}_{\mathrm{val}}(w_\star, \lambda) = G(\lambda) + \tau \Phi(\lambda).$$

Rearranging,

$$\mathcal{L}_{\mathrm{tr}}(w_\tau, \lambda) - G(\lambda) \le \tau \big(\Phi(\lambda) - \mathcal{L}_{\mathrm{val}}(w_\tau, \lambda)\big).$$

Since $w_\tau \in \mathcal{N}_K$ and $\mathcal{L}_{\mathrm{val}}(w_\tau, \lambda) \ge m_K$, the right-hand side is bounded by $\tau(\Phi(\lambda) - m_K) \le \tau M_K$. □

The third fact turns training suboptimality into validation proximity to the optimistic value. It is the only place where the local error bound is used.

Lemma 4 (Validation proximity of tilted minimizers)**.** *Under Assumption 4, let $w_\tau \in S_\tau(\lambda)$. Then*

$$\mathcal{L}_{\mathrm{val}}(w_\tau, \lambda) \ge \Phi(\lambda) - L_K \kappa_K M_K^\theta \tau^\theta. \tag{44}$$

Proof. Because $w_\tau \in S_\tau(\lambda) \subset \mathcal{W}(\lambda) \cap \mathcal{N}_K$, Lemma 3 and the error bound (40) give

$$\operatorname{dist}(w_\tau, S(\lambda)) \le \kappa_K \big(\mathcal{L}_{\mathrm{tr}}(w_\tau, \lambda) - G(\lambda)\big)^\theta \le \kappa_K M_K^\theta \tau^\theta. \tag{45}$$

Because $S(\lambda)$ is compact, choose a nearest point

$$\bar{w}_\tau \in S(\lambda) \quad \text{such that} \quad \|w_\tau - \bar{w}_\tau\| = \operatorname{dist}(w_\tau, S(\lambda)).$$

Since $\bar{w}_\tau \in S(\lambda)$, the definition of $\Phi(\lambda)$ gives

$$\mathcal{L}_{\mathrm{val}}(\bar{w}_\tau, \lambda) \ge \Phi(\lambda).$$

Using the L_K-Lipschitz property of $w \mapsto \mathcal{L}_{\mathrm{val}}(w, \lambda)$ on $\mathcal{N}_K$,

$$\mathcal{L}_{\mathrm{val}}(w_\tau, \lambda) \ge \mathcal{L}_{\mathrm{val}}(\bar{w}_\tau, \lambda) - L_K \|w_\tau - \bar{w}_\tau\| \ge \Phi(\lambda) - L_K \operatorname{dist}(w_\tau, S(\lambda)).$$

Combining this inequality with (45) proves (44). □

We can now prove the main finite-tilt statement.

Proof of Theorem 1. Fix $\lambda \in K$, $\tau \in (0, \bar{\tau}]$, and choose $w_\tau \in S_\tau(\lambda)$. Lemma 2 gives

$$\Delta_\tau(\lambda) \le \Phi(\lambda),$$

so the error is nonnegative. For the lower estimate on Δ_τ, write

$$\begin{aligned} \Delta_\tau(\lambda) &= \frac{V_\tau(\lambda) - G(\lambda)}{\tau} \\ &= \frac{\mathcal{L}_{\mathrm{tr}}(w_\tau, \lambda) - G(\lambda)}{\tau} + \mathcal{L}_{\mathrm{val}}(w_\tau, \lambda). \end{aligned}$$

The first term is nonnegative because $G(\lambda)$ is the infimum of the training objective. Therefore

$$\Delta_\tau(\lambda) \ge \mathcal{L}_{\mathrm{val}}(w_\tau, \lambda).$$

By Lemma 4,

$$\Delta_\tau(\lambda) \ge \Phi(\lambda) - L_K \kappa_K M_K^\theta \tau^\theta.$$

Combining this lower bound with $\Delta_\tau(\lambda) \le \Phi(\lambda)$ yields

$$0 \le \Phi(\lambda) - \Delta_\tau(\lambda) \le C_K \tau^\theta, \qquad C_K := L_K \kappa_K M_K^\theta.$$

Since the constants are uniform over $\lambda \in K$, taking the supremum over K gives the uniform convergence claim. □

Remark 1 (Where the rate comes from). The power τ^θ is not a smoothness artifact. It is inherited from the error bound relating training-value suboptimality to distance from the lower-level solution set. If a tilted minimizer has training error $O(\tau)$, then (40) places it within $O(\tau^\theta)$ of $S(\lambda)$, and validation Lipschitzness converts that distance into an $O(\tau^\theta)$ value error. Comparing the tilted minimizer with its nearest point of $S(\lambda)$ instead of the validation-best one sharpens the exponent to $\theta/(1-\theta)$ (Corollary 1).

C.4 Proof of the sharpened finite-tilt bound

Proof of Corollary 1. Fix $\lambda \in K$, $\tau \in (0, \bar{\tau}]$, and $w_\tau \in S_\tau(\lambda)$. Let $\bar{w} \in S(\lambda)$ be a nearest point to w_τ, which exists because $S(\lambda)$ is compact, and set $g = \mathcal{L}_{\mathrm{tr}}(w_\tau, \lambda) - G(\lambda) \geq 0$ and $r = \|w_\tau - \bar{w}\| = \mathrm{dist}(w_\tau, S(\lambda))$; both points lie in $\mathcal{N}_K$. Since $\Delta_\tau(\lambda) = g/\tau + \mathcal{L}_{\mathrm{val}}(w_\tau, \lambda) \geq \mathcal{L}_{\mathrm{val}}(w_\tau, \lambda)$ and $\Phi(\lambda) \leq \mathcal{L}_{\mathrm{val}}(\bar{w}, \lambda)$,

$$0 \leq \Phi(\lambda) - \Delta_\tau(\lambda) \leq \mathcal{L}_{\mathrm{val}}(\bar{w}, \lambda) - \mathcal{L}_{\mathrm{val}}(w_\tau, \lambda) \leq L_K r.$$

Optimality of w_τ for the tilted problem, tested against $\bar{w}$, gives $g + \tau\mathcal{L}_{\mathrm{val}}(w_\tau, \lambda) \leq \tau\mathcal{L}_{\mathrm{val}}(\bar{w}, \lambda)$, hence $g \leq \tau L_K r$. With the error bound (40), $r \leq \kappa_K g^\theta \leq \kappa_K(\tau L_K r)^\theta$. If $\theta < 1$ and $r > 0$, dividing by r^θ gives $r \leq \kappa_K^{1/(1-\theta)}(\tau L_K)^{\theta/(1-\theta)}$, which also holds for $r = 0$, and so $\Phi - \Delta_\tau \leq (\kappa_K L_K)^{1/(1-\theta)}\tau^{\theta/(1-\theta)}$. Theorem 1 gives the other term of the minimum, and all constants are uniform over K. If $\theta = 1$ and $\kappa_K \tau L_K < 1$, then $r \leq \kappa_K \tau L_K r$ forces $r = 0$, so $w_\tau \in S(\lambda)$ and $g = 0$; hence $\Delta_\tau(\lambda) = \mathcal{L}_{\mathrm{val}}(w_\tau, \lambda) \geq \Phi(\lambda) \geq \Delta_\tau(\lambda)$, and every tilted minimizer is a validation-best training minimizer. □

The exponent is attained. Fix $\theta \in (0, 1)$, $p = 1/\theta$, and $L > 0$; let the outer variable be inert, $\mathcal{W} = \mathbb{R}$, $\mathcal{L}_{\mathrm{tr}}(w) = \mathrm{dist}(w, [-1, 1])^p$, and $\mathcal{L}_{\mathrm{val}}(w) = -Lw$. Then $S = [-1, 1]$, $G = 0$, $\Phi = -L$; the error bound holds with $\kappa = 1$, and on $\mathcal{N} = [-3, 3]$ one may take $m = -3L$ and $M = 2L$. The tilted objective is at least $-\tau L$ on $[-1, 1]$, larger for $w < -1$, and equals $r^p - \tau L(1 + r)$ at $w = 1 + r$ with $r \geq 0$, which is minimized at $r_\star = (\theta L\tau)^{\theta/(1-\theta)}$. Using $r_\star^p = r_\star \theta L \tau$, one gets $\Delta_\tau = -L - (1 - \theta)Lr_\star$ and

$$\Phi - \Delta_\tau = (1 - \theta)L(\theta L\tau)^{\theta/(1-\theta)},$$

a constant multiple $(1 - \theta)\theta^{\theta/(1-\theta)} \in (0, 1)$ of the bound $(\kappa L)^{1/(1-\theta)}\tau^{\theta/(1-\theta)}$. The exponent therefore cannot be improved under Assumption 1. A standard-library script confirms the closed form by brute-force minimization for $\theta \in \{1/4, 1/2, 2/3\}$ (relative error at most 1.2×10^{-4}, the grid resolution; log–log slopes $1/3$, 1, 2) and returns $\Delta_\tau = \Phi$ exactly for $\theta = 1$ and $\tau L < 1$.

C.5 Proof of the conservative-oracle theorem

The next proof is algebraic once the conservative fields for G and V_τ are available. We first record a selection-wise quotient lemma.

Lemma 5 (Difference quotients of conservative value fields)**.** *Assume Assumption 5. Fix $\tau \in (0, \bar{\tau}]$ and define $\Delta_\tau = (V_\tau - G)/\tau$. Let $\gamma : [0, 1] \to K$ be absolutely continuous, and let*

$$u_\tau(t) \in D_{V_\tau}(\gamma(t)), \qquad u_0(t) \in D_G(\gamma(t)) \qquad \textit{for a.e. } t$$

be measurable paired selections. Then

$$\Delta_\tau(\gamma(1)) - \Delta_\tau(\gamma(0)) = \int_0^1 \left\langle \frac{u_\tau(t) - u_0(t)}{\tau}, \dot{\gamma}(t) \right\rangle dt. \tag{46}$$

Proof. By the chain-rule identities in Assumption 5,

$$V_\tau(\gamma(1)) - V_\tau(\gamma(0)) = \int_0^1 \langle u_\tau(t), \dot{\gamma}(t)\rangle dt,$$

and

$$G(\gamma(1)) - G(\gamma(0)) = \int_0^1 \langle u_0(t), \dot{\gamma}(t)\rangle dt.$$

Subtract the second identity from the first and divide by τ:

$$\begin{aligned}\Delta_\tau(\gamma(1)) - \Delta_\tau(\gamma(0)) &= \frac{V_\tau(\gamma(1)) - G(\gamma(1))}{\tau} - \frac{V_\tau(\gamma(0)) - G(\gamma(0))}{\tau} \\ &= \int_0^1 \left\langle \frac{u_\tau(t) - u_0(t)}{\tau}, \dot{\gamma}(t) \right\rangle dt.\end{aligned}$$

This proves the claim. □

The following endpoint lemma converts the exact chain rule for Δ_τ into an approximate chain rule for the limiting optimistic objective Φ.

Lemma 6 (Endpoint transfer from Δ_τ to Φ)**.** *Suppose that, on K, $|\Phi(\lambda) - \Delta_\tau(\lambda)| \le b_\tau$ for some $b_\tau \ge 0$. If a measurable vector field $d(t)$ along an absolutely continuous curve $\gamma : [0,1] \to K$ satisfies*

$$\Delta_\tau(\gamma(1)) - \Delta_\tau(\gamma(0)) = \int_0^1 \langle d(t), \dot\gamma(t)\rangle dt,$$

then

$$\left|\Phi(\gamma(1)) - \Phi(\gamma(0)) - \int_0^1 \langle d(t), \dot\gamma(t)\rangle dt\right| \le 2b_\tau. \tag{47}$$

Proof. Add and subtract $\Delta_\tau(\gamma(1))$ and $\Delta_\tau(\gamma(0))$:

$$\begin{aligned}
&\Phi(\gamma(1)) - \Phi(\gamma(0)) - \int_0^1 \langle d(t), \dot\gamma(t)\rangle dt \\
&\quad = \big[\Phi(\gamma(1)) - \Delta_\tau(\gamma(1))\big] - \big[\Phi(\gamma(0)) - \Delta_\tau(\gamma(0))\big] \\
&\qquad + \left[\Delta_\tau(\gamma(1)) - \Delta_\tau(\gamma(0)) - \int_0^1 \langle d(t), \dot\gamma(t)\rangle dt\right].
\end{aligned}$$

The bracket on the last line is zero by assumption. Taking absolute values and using the uniform endpoint bound twice yields (47). □

Proof of Theorem 2. Let γ, $u_\tau(\cdot)$, and $u_0(\cdot)$ be as in the theorem, and set

$$d_\tau(t) := \frac{u_\tau(t) - u_0(t)}{\tau}.$$

The exact surrogate identity

$$\Delta_\tau(\gamma(1)) - \Delta_\tau(\gamma(0)) = \int_0^1 \langle d_\tau(t), \dot\gamma(t)\rangle dt$$

is precisely Lemma 5. This proves that paired ghost directions are exact conservative first-order oracles for the finite-tilt surrogate.

For the limiting objective Φ, Theorem 1 gives the uniform endpoint bound

$$|\Phi(\lambda) - \Delta_\tau(\lambda)| \le C_K \tau^\theta \qquad \forall \lambda \in K.$$

Apply Lemma 6 with $b_\tau = C_K\tau^\theta$ and $d = d_\tau$ to obtain

$$\left|\Phi(\gamma(1)) - \Phi(\gamma(0)) - \int_0^1 \langle d_\tau(t), \dot\gamma(t)\rangle dt\right| \le 2C_K\tau^\theta.$$

This is the desired asymptotic pathwise oracle statement. □

Remark 2 (Why the theorem uses paired selections). The result does not require a single-valued derivative of either value function. It only requires a measurable tilted adjoint selection and a measurable base adjoint selection along the same outer path. Mathematically, any measurable valid pair gives an element of the finite-tilt conservative field, and the identity holds along paths. It does not make an arbitrary pair a gradient at a single point; the algorithm therefore reads compatible pairs at one shared outer point (Proposition 2, Theorem 4). Computing both certificates in the same "world" with a warm start reduces cost and solver variance but is not the compatibility mechanism.

C.6 Compatible readouts

For a locally Lipschitz h, $\partial_B h(\lambda)$ denotes the limits $\lim_k \nabla h(\lambda_k)$ over differentiability points $\lambda_k \to \lambda$, and $\partial_C h(\lambda) = \operatorname{conv} \partial_B h(\lambda)$; more generally $\partial_C h(\lambda)$ is the convex hull of such limits along sequences avoiding any prescribed Lebesgue-null set (Clarke, 1990, Thm. 2.5.1).

Example 1 (Valid adjoints, invalid difference)**.** Let $\Lambda = [-1, 1]$, $\mathcal{L}_{\mathrm{tr}}(w, \lambda) = \mathrm{dist}(w, [-1, 1])^2 + |\lambda|$, and $\mathcal{L}_{\mathrm{val}}(w, \lambda) = \lambda^2/2$. Then $S(\lambda) = S_\tau(\lambda) = [-1, 1]$, $G = |\lambda|$, $V_\tau = |\lambda| + \tau\lambda^2/2$, and $\Delta_\tau = \Phi = \lambda^2/2$. The Clarke subdifferentials are conservative fields for G and V_τ and equal $[-1, 1]$ at $\lambda = 0$. The valid choices $u_0 = -1$, $u_\tau = 1$ give $d = 2/\tau$ although $\nabla\Delta_\tau(0) = 0$; for $0 < \eta < \tau/2$ the projected step moves the minimizer to $-2\eta/\tau$ and raises Δ_τ to $2\eta^2/\tau^2$. The Cartesian field $(D_{V_\tau} - D_G)/\tau$ is nevertheless a conservative field for Δ_τ: it is locally bounded with closed graph and compact values, and Lemma 5 gives its chain rule; its spurious element $2/\tau$ sits on the null set $\{0\}$, which is harmless for path integrals and harmful for an algorithm that evaluates the field there.

Proof of Proposition 2. Fix $\lambda \in U$ and a closed ball $\bar{B} \subset U$ around λ on which G and V_τ are Lipschitz with constants L_G and L_V. By Rademacher's theorem R_τ has full measure and is dense, so there are $\lambda_k \in R_\tau \cap \bar{B}$ with $\lambda_k \to \lambda$; the gradient pairs are bounded by (L_V, L_G), so a subsequence converges and $\mathcal{J}_\tau(\lambda) \neq \emptyset$. $\mathcal{J}_\tau(\lambda)$ is bounded by the same constants and closed by a diagonal argument: if $(a_j, b_j) \in \mathcal{J}_\tau(\lambda)$ converge to (a, b), pick $\mu_j \in R_\tau$ with $\|\mu_j - \lambda\| < 1/j$ and $\|(\nabla V_\tau(\mu_j), \nabla G(\mu_j)) - (a_j, b_j)\| < 1/j$. Each component of a compatible pair is a limit of gradients of the corresponding value at differentiability points converging to λ, hence lies in its Clarke subdifferential. At every $\mu \in R_\tau$, Δ_τ is differentiable with $\nabla\Delta_\tau(\mu) = (\nabla V_\tau(\mu) - \nabla G(\mu))/\tau$, so every element of $D_\tau^{\mathrm{cmp}}(\lambda)$ is a limit of gradients of Δ_τ and lies in $\partial_B\Delta_\tau(\lambda) \subseteq \partial_C\Delta_\tau(\lambda)$. If Δ_τ is C^1 near λ, continuity of $\nabla\Delta_\tau$ forces every such limit to equal $\nabla\Delta_\tau(\lambda)$. □

In Example 1, $R_\tau = \mathbb{R} \setminus \{0\}$, and the gradient pairs converge to $(1, 1)$ from the right and to $(-1, -1)$ from the left, so $\mathcal{J}_\tau(0) = \{(1, 1), (-1, -1)\}$ and $D_\tau^{\mathrm{cmp}}(0) = \{0\}$. Differentiability at λ alone does not suffice for the last statement: with $\mathcal{W} \equiv \{0\}$, $\mathcal{L}_{\mathrm{tr}} \equiv 0$, and $\mathcal{L}_{\mathrm{val}}(w, \lambda) = h(\lambda) = \lambda^2 \sin(1/\lambda)$, $h(0) = 0$, one has $\Delta_\tau = h$, $h'(0) = 0$, but $D_\tau^{\mathrm{cmp}}(0) = [-1, 1]$. A Clarke-valid direction is also not by itself a descent direction.

C.7 Proof of the inexact-oracle theorem

The inexact theorem uses only the triangle inequality. The important point is conceptual: differencing two approximate adjoints divides their errors by τ.

Lemma 7 (Pointwise perturbation of a ghost direction)**.** *Let $u_0, u_\tau, \tilde{u}_0, \tilde{u}_\tau \in \mathbb{R}^m$ satisfy*

$$\|\tilde{u}_0 - u_0\| \le \varepsilon_0, \qquad \|\tilde{u}_\tau - u_\tau\| \le \varepsilon_\tau.$$

Define

$$d_\tau := \frac{u_\tau - u_0}{\tau}, \qquad \tilde{d}_\tau := \frac{\tilde{u}_\tau - \tilde{u}_0}{\tau}.$$

Then

$$\|\tilde{d}_\tau - d_\tau\| \le \frac{\varepsilon_0 + \varepsilon_\tau}{\tau}. \tag{48}$$

Proof. Compute

$$\tilde{d}_\tau - d_\tau = \frac{(\tilde{u}_\tau - u_\tau) - (\tilde{u}_0 - u_0)}{\tau}.$$

The triangle inequality gives

$$\|\tilde{d}_\tau - d_\tau\| \le \frac{\|\tilde{u}_\tau - u_\tau\| + \|\tilde{u}_0 - u_0\|}{\tau} \le \frac{\varepsilon_\tau + \varepsilon_0}{\tau}.$$

□

Lemma 8 (Pathwise perturbation by an approximate ghost field)**.** *Let $\gamma : [0, 1] \to K$ be absolutely continuous. Suppose two measurable fields $d_\tau(t)$ and $\tilde{d}_\tau(t)$ along γ satisfy*

$$\|\tilde{d}_\tau(t) - d_\tau(t)\| \le \rho_\tau \qquad \textit{for a.e. } t \in [0, 1].$$

Then

$$\left| \int_0^1 \langle \tilde{d}_\tau(t) - d_\tau(t), \dot{\gamma}(t) \rangle dt \right| \le \rho_\tau \, \mathrm{Len}(\gamma), \qquad \mathrm{Len}(\gamma) := \int_0^1 \|\dot{\gamma}(t)\| dt. \tag{49}$$

Proof. By Cauchy–Schwarz and the assumed uniform bound,

$$\left|\int_0^1 \langle \tilde{d}_\tau(t) - d_\tau(t), \dot{\gamma}(t)\rangle dt\right| \le \int_0^1 \|\tilde{d}_\tau(t) - d_\tau(t)\|\,\|\dot{\gamma}(t)\| dt \le \rho_\tau \operatorname{Len}(\gamma).$$

□

Proof of Theorem 3. The pointwise error bound is Lemma 7 with

$$\varepsilon_0 = \varepsilon_{0,K}, \qquad \varepsilon_\tau = \varepsilon_{\tau,K}.$$

Thus, for every $\lambda \in K$,

$$\|\tilde{d}_\tau(\lambda) - d_\tau(\lambda)\| \le \frac{\varepsilon_{0,K} + \varepsilon_{\tau,K}}{\tau}.$$

Now let $\gamma : [0,1] \to K$ be absolutely continuous. Add and subtract the integral of the exact ghost field:

$$\begin{aligned}
&\left|\Phi(\gamma(1)) - \Phi(\gamma(0)) - \int_0^1 \langle \tilde{d}_\tau(\gamma(t)), \dot{\gamma}(t)\rangle dt\right| \\
&\quad\le \left|\Phi(\gamma(1)) - \Phi(\gamma(0)) - \int_0^1 \langle d_\tau(\gamma(t)), \dot{\gamma}(t)\rangle dt\right| \\
&\qquad + \left|\int_0^1 \langle d_\tau(\gamma(t)) - \tilde{d}_\tau(\gamma(t)), \dot{\gamma}(t)\rangle dt\right|.
\end{aligned}$$

The first term is bounded by $2C_K\tau^\theta$ from Theorem 2. The second term is bounded by Lemma 8 with

$$\rho_\tau = \frac{\varepsilon_{0,K} + \varepsilon_{\tau,K}}{\tau}.$$

Therefore

$$\left|\Phi(\gamma(1)) - \Phi(\gamma(0)) - \int_0^1 \langle \tilde{d}_\tau(\gamma(t)), \dot{\gamma}(t)\rangle dt\right| \le 2C_K\tau^\theta + \frac{\varepsilon_{0,K} + \varepsilon_{\tau,K}}{\tau} \operatorname{Len}(\gamma),$$

which proves the pathwise inexact-oracle bound. □

C.8 Balancing the tilt and numerical accuracy

The main text states the usual balancing rule when the combined adjoint error is essentially uniform in τ. We include the short calculation for completeness.

Proposition 3 (Elementary tilt balance)**.** *Assume the pathwise error is bounded, up to constants independent of τ, by*

$$E(\tau) = A\tau^\theta + \frac{B}{\tau}, \qquad A > 0, \quad B > 0, \quad \theta \in (0,1]. \tag{50}$$

Then any minimizer over $\tau > 0$ satisfies

$$\tau_\star = \left(\frac{B}{A\theta}\right)^{1/(\theta+1)}, \tag{51}$$

so $\tau_\star \asymp B^{1/(\theta+1)}$ when A and θ are treated as fixed. The corresponding error scales as

$$E(\tau_\star) = O\big(B^{\theta/(\theta+1)}\big). \tag{52}$$

Proof. Differentiate (50):

$$E'(\tau) = A\theta\tau^{\theta-1} - B\tau^{-2}.$$

Setting $E'(\tau) = 0$ gives

$$A\theta\tau^{\theta+1} = B,$$

which is (51). Substitution into (50) shows that both terms have order $B^{\theta/(\theta+1)}$, proving (52). □

In Theorem 3, the role of B is played by the combined adjoint-error envelope times the path length. If the inner solver becomes less accurate as τ decreases, then B is not constant and the same calculation should be repeated with the solver-specific error model.

With the sharpened bound of Corollary 1 and $\theta < 1$, the same calculation applies with the exponent $p = \theta/(1-\theta)$ and $A = 2(\kappa_K L_K)^{1/(1-\theta)}$ in place of θ and the cruder constant: $\tau^\star = (B/(Ap))^{1/(p+1)} = \Theta(B^{1-\theta})$, since $1/(p+1) = 1-\theta$, and both terms of $E(\tau^\star)$ are $O(B^\theta)$. For $\theta = 1$ the bias vanishes for $\tau < 1/(\kappa_K L_K)$, so the largest such τ is best. The rule $\tau \asymp B^{1/(1+\theta)}$ of Proposition 3 is the balance of the cruder bound and is superseded for small B.

C.9 Proof of the computed-direction theorem

Proof of Theorem 4. Let $\mathcal{F}_t$ be the σ-field generated by $\lambda^0, q_0, \dots, q_{t-1}$, and all solver randomness before iteration t; λ^t is $\mathcal{F}_t$-measurable, q_t is independent of $\mathcal{F}_t$, and δ_t is deterministic. *Step 1.* By Bolte & Pauwels (2021, Thm. 1), whose argument is local and applies verbatim on the open set U, there are Lebesgue-null Borel sets $N_0, N_\tau \subset U$ outside of which $D_G = \{\nabla G\}$ and $D_{V_\tau} = \{\nabla V_\tau\}$; both values are differentiable off $N = N_0 \cup N_\tau$. Conditionally on $\mathcal{F}_t$, $\mathbb{P}(\bar\lambda_t \in N \mid \mathcal{F}_t) = \mu(\delta_t^{-1}(N - \lambda^t)) = 0$ because μ has a density; taking expectations and a countable union over t, almost surely $\bar\lambda_t \notin N$ for every t. The point $\bar\lambda_t$ lies in U because $\lambda^t \in \Lambda$, $\|q_t\| \le 1$, $\delta_t \le \bar\delta$, and $\Lambda + \bar\delta\,\mathbb{B} \subseteq U$. *Step 2.* On this event, for $z \in Z_s(\bar\lambda_t)$, $\mathrm{Adj}(F_s, \bar\lambda_t, z) \in D_{v_s}(\bar\lambda_t) = \{\nabla v_s(\bar\lambda_t)\}$. With $z_s^\star$ a nearest point of the closed set $Z_s(\bar\lambda_t)$ to $\tilde z_{s,t}$, the Lipschitz readout and the certificate error bound give $\|u_{s,t} - \nabla v_s(\bar\lambda_t)\| \le B_s \operatorname{dist}(\tilde z_{s,t}, Z_s(\bar\lambda_t)) \le B_s \kappa_s R_{s,t}^{\alpha_s} = \varepsilon_{s,t}$. *Step 3.* Both values are differentiable at $\bar\lambda_t$, so $\nabla\Delta_\tau(\bar\lambda_t) = (\nabla V_\tau(\bar\lambda_t) - \nabla G(\bar\lambda_t))/\tau$ and $\|d_t - \nabla\Delta_\tau(\bar\lambda_t)\| \le (\varepsilon_{0,t} + \varepsilon_{\tau,t})/\tau$; under Assumption 3, $\|\nabla\Delta_\tau(\bar\lambda_t) - \nabla\Delta_\tau(\lambda^t)\| \le L_\tau \delta_t \|q_t\|$, and the triangle inequality gives (17). □

Step 2 needs certificates of the *global* value functions; a KKT point of a nonconvex lower problem that is not globally optimal lies outside Z_s, and no residual bound repairs that. For a μ_s-strongly convex smooth lower problem with $\nabla_\lambda F_s(\cdot, \bar\lambda)$ $L_{\lambda w,s}$-Lipschitz in w, taking $Z_s = \{w_s^\star\}$ and the readout $w \mapsto \nabla_\lambda F_s(w, \bar\lambda)$ gives $\kappa_s = 1/\mu_s$, $\alpha_s = 1$, $B_s = L_{\lambda w,s}$, hence $\varepsilon_s \le (L_{\lambda w,s}/\mu_s)\|\nabla_w F_s(w_s, \bar\lambda)\|$, and $\varepsilon_s \le L_{\lambda w,s}\sqrt{2q_s/\mu_s}$ for an objective gap q_s. Without Assumption 3, Steps 1–3 still show that d_t is within $(\varepsilon_{0,t} + \varepsilon_{\tau,t})/\tau$ of a gradient of Δ_τ at a random point within δ_t of λ^t; we do not analyze that nonsmooth regime.

C.10 Proof of the projected finite-tilt convergence theorem

We now prove Theorem 5. The proof is the standard nonconvex projected-gradient descent argument, included to make clear that the convergence is for the explicit finite-tilt surrogate Δ_τ.

Lemma 9 (Projected descent with an exact gradient)**.** *Let f be differentiable with L-Lipschitz gradient on a closed convex set Λ, and let*

$$x^+ = \Pi_\Lambda(x - \eta\nabla f(x)), \qquad \mathcal{G}_\eta(x) := \frac{1}{\eta}(x - x^+), \qquad 0 < \eta < 2/L.$$

Then

$$f(x^+) \le f(x) - \eta(1 - L\eta/2)\|\mathcal{G}_\eta(x)\|^2. \tag{53}$$

Proof. Write $s := x^+ - x = -\eta\mathcal{G}_\eta(x)$. The projection step is equivalently

$$x^+ = \arg\min_{z\in\Lambda}\left\{\langle\nabla f(x), z - x\rangle + \frac{1}{2\eta}\|z - x\|^2\right\}.$$

The first-order optimality condition, tested at $z = x$, gives

$$\langle\nabla f(x) + s/\eta, x - x^+\rangle \ge 0,$$

which is the same as

$$\langle\nabla f(x), s\rangle \le -\frac{1}{\eta}\|s\|^2.$$

By L-smoothness,

$$f(x^+) \le f(x) + \langle \nabla f(x), s\rangle + \frac{L}{2}\|s\|^2 \le f(x) - \left(\frac{1}{\eta} - \frac{L}{2}\right)\|s\|^2.$$

Substituting $\|s\| = \eta\|\mathcal{G}_\eta(x)\|$ proves (53). □

Lemma 10 (Projected descent with an inexact direction)**.** *Let f be differentiable with L-Lipschitz gradient on Λ. Let $d = \nabla f(x) + e$ with $\|e\| \le \rho$, and define*

$$x^+ = \Pi_\Lambda(x - \eta d), \qquad \widetilde{\mathcal{G}}_\eta(x; d) := \frac{1}{\eta}(x - x^+), \qquad c := 1 - L\eta/2 > 0.$$

Then

$$f(x^+) \le f(x) - \frac{\eta c}{2}\|\widetilde{\mathcal{G}}_\eta(x; d)\|^2 + \frac{\eta\rho^2}{2c}. \tag{54}$$

Moreover, if $\mathcal{G}_\eta(x)$ denotes the exact projected-gradient mapping, then

$$\|\mathcal{G}_\eta(x) - \widetilde{\mathcal{G}}_\eta(x; d)\| \le \rho. \tag{55}$$

Proof. Write $s := x^+ - x = -\eta\widetilde{\mathcal{G}}_\eta(x; d)$. The projection optimality condition for the direction d gives

$$\langle d, s\rangle \le -\frac{1}{\eta}\|s\|^2.$$

Since $d = \nabla f(x) + e$,

$$\langle \nabla f(x), s\rangle = \langle d, s\rangle - \langle e, s\rangle \le -\frac{1}{\eta}\|s\|^2 + \rho\|s\|.$$

The smoothness inequality therefore gives

$$f(x^+) \le f(x) - \left(\frac{1}{\eta} - \frac{L}{2}\right)\|s\|^2 + \rho\|s\|.$$

Using $\|s\| = \eta\|\widetilde{\mathcal{G}}_\eta(x; d)\|$ and $c = 1 - L\eta/2$ yields

$$f(x^+) \le f(x) - \eta c\|\widetilde{\mathcal{G}}_\eta(x; d)\|^2 + \eta\rho\|\widetilde{\mathcal{G}}_\eta(x; d)\|.$$

Apply Young's inequality $\rho a \le (c/2)a^2 + \rho^2/(2c)$ with $a = \|\widetilde{\mathcal{G}}_\eta(x; d)\|$ to obtain (54).

For (55), use nonexpansiveness of Euclidean projection:

$$\begin{aligned}\|\mathcal{G}_\eta(x) - \widetilde{\mathcal{G}}_\eta(x; d)\| &= \frac{1}{\eta}\,\|\Pi_\Lambda(x - \eta d) - \Pi_\Lambda(x - \eta\nabla f(x))\| \\ &\le \frac{1}{\eta}\|\eta(d - \nabla f(x))\| = \|e\| \le \rho.\end{aligned}$$

□

Proof of Theorem 5. Apply Lemma 9 with $f = \Delta_\tau$, $L = L_\tau$, and $x = \lambda^t$. With $c := 1 - L_\tau\eta/2 > 0$,

$$\Delta_\tau(\lambda^{t+1}) \le \Delta_\tau(\lambda^t) - \eta c\|\mathcal{G}_{\eta,\tau}(\lambda^t)\|^2.$$

Summing from $t = 0$ to $T - 1$ and using $\Delta_\tau(\lambda^T) \ge \Delta_\tau^{\inf}$ gives

$$\eta c\sum_{t=0}^{T-1}\|\mathcal{G}_{\eta,\tau}(\lambda^t)\|^2 \le \Delta_\tau(\lambda^0) - \Delta_\tau^{\inf}.$$

The minimum is bounded by the average, which proves (18).

For the inexact case, apply Lemma 10 at each iterate with $f = \Delta_\tau$, $d = d_t$, and $\|e_t\| \le \rho_t$:

$$\Delta_\tau(\lambda^{t+1}) \le \Delta_\tau(\lambda^t) - \frac{\eta c}{2}\|\widetilde{\mathcal{G}}_{\eta,t}\|^2 + \frac{\eta\rho_t^2}{2c}.$$

Summing and rearranging yields

$$\frac{\eta c}{2}\sum_{t=0}^{T-1}\|\widetilde{\mathcal{G}}_{\eta,t}\|^2 \le \Delta_\tau(\lambda^0) - \Delta_\tau^{\inf} + \frac{\eta}{2c}\sum_{t=0}^{T-1}\rho_t^2.$$

Dividing by $T\eta c/2$ bounds the average of $\|\widetilde{\mathcal{G}}_{\eta,t}\|^2$. Lemma 10 also gives $\|\mathcal{G}_{\eta,\tau}(\lambda^t)\| \le \|\widetilde{\mathcal{G}}_{\eta,t}\| + \rho_t$, so $\|\mathcal{G}_{\eta,\tau}(\lambda^t)\|^2 \le 2\|\widetilde{\mathcal{G}}_{\eta,t}\|^2 + 2\rho_t^2$, and since $2 \le 2/c^2$,

$$\min_{t<T}\|\mathcal{G}_{\eta,\tau}(\lambda^t)\|^2 \le \frac{1}{T}\sum_{t<T}\|\mathcal{G}_{\eta,\tau}(\lambda^t)\|^2 \le \frac{4(\Delta_\tau(\lambda^0)-\Delta_\tau^{\inf})}{\eta c T} + \frac{4}{c^2 T}\sum_{t<T}\rho_t^2,$$

the inexact statement of Theorem 5. For Algorithm 1, Theorem 4 supplies ρ_t almost surely; if $\delta_t \to 0$ and $R_{s,t} \to 0$, then $\rho_t \to 0$ because $\|q_t\| \le 1$, the Cesàro averages of ρ_t^2 vanish, and so does the bound. A zero of $\mathcal{G}_{\eta,\tau}$ means $-\nabla\Delta_\tau(\lambda) \in N_\Lambda(\lambda)$. In the exact case $\rho = 0$, the bound scales as $O(1/T)$ for the squared residual, or $O(T^{-1/2})$ for the residual itself, so $T = O(\varepsilon^{-2})$ gives an ε-stationary iterate. □

Corollary 2 (Linear convergence under projected PL geometry)**.** *Under the exact-oracle assumptions of Theorem 5, set* $c_\eta := 1 - L_\tau\eta/2$ *and* $\Delta_\tau^\star := \inf_{\lambda\in\Lambda}\Delta_\tau(\lambda)$*. If* $\frac{1}{2}\|\mathcal{G}_{\eta,\tau}(\lambda)\|^2 \ge \mu_\tau(\Delta_\tau(\lambda) - \Delta_\tau^\star)$ *on* Λ *and* $0 < 2\eta c_\eta\mu_\tau \le 1$*, then* $\Delta_\tau(\lambda^T) - \Delta_\tau^\star \le (1 - 2\eta c_\eta\mu_\tau)^T(\Delta_\tau(\lambda^0) - \Delta_\tau^\star)$*. For* $\eta = 1/L_\tau$*, the factor is* $1 - \mu_\tau/L_\tau$ *when* $\mu_\tau \le L_\tau$*.*

Proof of Corollary 2. Let

$$r_t := \Delta_\tau(\lambda^t) - \Delta_\tau^\star, \qquad c_\eta := 1 - L_\tau\eta/2.$$

The exact projected descent inequality established above gives

$$\Delta_\tau(\lambda^{t+1}) \le \Delta_\tau(\lambda^t) - \eta c_\eta\|\mathcal{G}_{\eta,\tau}(\lambda^t)\|^2.$$

Subtracting $\Delta_\tau^\star$ and applying the projected PL condition

$$\|\mathcal{G}_{\eta,\tau}(\lambda^t)\|^2 \ge 2\mu_\tau\big(\Delta_\tau(\lambda^t) - \Delta_\tau^\star\big) = 2\mu_\tau r_t$$

yields

$$r_{t+1} \le \big(1 - 2\eta c_\eta\mu_\tau\big)r_t.$$

The assumption $0 < 2\eta c_\eta\mu_\tau \le 1$ ensures that the contraction factor lies in $[0, 1)$. Iterating the recursion gives

$$r_T \le \big(1 - 2\eta c_\eta\mu_\tau\big)^T r_0,$$

which is the stated projected-PL rate. If $\eta = 1/L_\tau$, then $c_\eta = 1/2$, so the factor becomes $1 - \mu_\tau/L_\tau$; the additional condition $\mu_\tau \le L_\tau$ ensures that it is nonnegative. □

Value transfer to Φ. If $\Lambda \subseteq K$, Corollary 1 gives $\Delta_\tau \le \Phi \le \Delta_\tau + b_\tau$ on Λ, hence $\inf_\Lambda\Delta_\tau \le \inf_\Lambda\Phi$ and $\Phi(\lambda^T) - \inf_\Lambda\Phi \le \Delta_\tau(\lambda^T) - \Delta_\tau^\star + b_\tau \le (1 - 2\eta c_\eta\mu_\tau)^T r_0 + b_\tau$. This transfers values, not stationarity.

C.11 A SAFE STATIONARITY CERTIFICATE FOR THE FINITE-TILT SURROGATE

The main paper gives a projected-gradient theorem only in the smooth finite-tilt regime and avoids an unconditional monotone-descent theorem for arbitrary nonsmooth conservative-field selections or for the limiting objective Φ. The following statement is a weaker and safer nonsmooth certificate: for compatible directions, which are Clarke subgradients of the finite-tilt surrogate by Proposition 2, the usual projected-gradient residual certifies stationarity for Δ_τ.

Corollary 3 (Projected residual as a surrogate stationarity certificate)**.** *Assume that* $\Lambda \subset \mathbb{R}^m$ *is closed and convex and that, at the point* $\lambda \in \Lambda$*,*

$$d_\tau(\lambda) \in D_\tau^{\mathrm{cmp}}(\lambda) \subseteq \partial^{\mathrm{C}}\Delta_\tau(\lambda), \tag{56}$$

where the inclusion is Proposition 2. For the Cartesian field $(D_{V_\tau} - D_G)/\tau$ the inclusion can fail (Example 1), and a zero residual computed from an arbitrary valid pair certifies nothing. For $\eta > 0$, define

$$\begin{aligned} T_{\eta,\tau}(\lambda) &:= \arg\min_{z\in\Lambda} \left\{ \langle d_\tau(\lambda), z - \lambda\rangle + \frac{1}{2\eta}\|z-\lambda\|^2 \right\}, \\ \mathcal{G}_{\eta,\tau}(\lambda) &:= \frac{1}{\eta}\big(\lambda - T_{\eta,\tau}(\lambda)\big). \end{aligned} \tag{57}$$

If $\mathcal{G}_{\eta,\tau}(\lambda) = 0$, then

$$0 \in \partial^{\mathrm{C}}\Delta_\tau(\lambda) + N_\Lambda(\lambda), \tag{58}$$

so λ is Clarke-stationary for the finite-tilt surrogate over Λ.

Proof. The optimization problem defining $T_{\eta,\tau}(\lambda)$ is strongly convex, so the minimizer is unique. Its first-order optimality condition is

$$0 \in d_\tau(\lambda) + \frac{1}{\eta}\big(T_{\eta,\tau}(\lambda) - \lambda\big) + N_\Lambda(T_{\eta,\tau}(\lambda)).$$

If $\mathcal{G}_{\eta,\tau}(\lambda) = 0$, then $T_{\eta,\tau}(\lambda) = \lambda$, and the display reduces to

$$0 \in d_\tau(\lambda) + N_\Lambda(\lambda).$$

Using $d_\tau(\lambda) \in \partial^{\mathrm{C}}\Delta_\tau(\lambda)$ gives (58). □

Remark 3 (Why this is not a descent theorem). Corollary 3 only interprets the projected residual. It does not assert that every projected ghost step decreases Δ_τ or Φ. A monotone descent theorem would require additional assumptions, for example weak convexity and a genuine proximal step, or a smoothness model that upper-bounds the finite-tilt surrogate along the update. This distinction is important in a nonsmooth conservative-field setting.

C.12 Convergence-rate positioning against cited bilevel methods

This section records how the finite-tilt convergence theorem should be compared with the rates reported in the cited bilevel-optimization literature. The comparison is delicate because papers use different stationarity conventions. In this table, "squared-gradient accuracy" means a criterion of the form $\mathbb{E}\|\nabla h(x)\|^2 \le \epsilon$ or its projected analogue, while "norm accuracy" means $\|\nabla h(x)\| \le \epsilon$. Our main theorem is best read in squared-residual form:

$$R_T^{\mathrm{Ghost}} := \min_{0\le t<T} \|\mathcal{G}_{\eta,\tau}(\lambda^t)\|^2 \le \frac{\Delta_\tau(\lambda^0) - \Delta_\tau^{\inf}}{\eta(1 - L_\tau\eta/2)T}.$$

Thus exact full-batch Ghost gives $R_T^{\mathrm{Ghost}} = O(1/T)$ for the smooth finite-tilt surrogate, or $O(\epsilon^{-1})$ iterations for squared-residual accuracy $R_T^{\mathrm{Ghost}} \le \epsilon$. In norm-accuracy notation this is the usual $O(\epsilon^{-2})$ deterministic nonconvex first-order rate. Corollary 2 further gives a linear rate when the finite-tilt surrogate satisfies a projected PL inequality.

Table 2: Rate positioning relative to representative cited bilevel methods. Rates are stated in the conventions used by the corresponding papers, so the comparison should be interpreted structurally rather than as a single apples-to-apples oracle-count ranking.

Method / paper	Main structural regime	Reported stationarity target and rate	Implication for Ghost positioning
AID/ITD and stocBiO analyses (Ji et al., 2021); MRBO/VRBO (Yang et al., 2021)	Smooth nonconvex upper objective; strongly convex lower problem, hence a unique response map and a smooth hyper-objective.	stocBiO is reported at $\widetilde{O}(\epsilon^{-2})$ complexity for squared-gradient stationarity; MRBO and VRBO are reported at $\widetilde{O}(\epsilon^{-3/2})$ in the same stochastic strongly-convex-lower-level setting.	These methods have stronger stochastic sample-complexity rates in their unique-response regime. Ghost is not claimed to dominate them there; its point is to avoid differentiating a selected response when lower-level minimizers are nonunique.
SUSTAIN (Khanduri et al., 2021)	Stochastic unconstrained bilevel problems with strongly convex lower level and smooth possibly nonconvex upper objective.	The paper reports $O(\epsilon^{-3/2})$ iterations, each with $O(1)$ samples, to find a point whose squared outer-gradient norm is at most ϵ.	Faster stochastic complexity than deterministic projected Ghost in the unique-response regime; different assumptions and target objective.
SOBA/SABA (Dagréou et al., 2022)	Finite-sum stochastic bilevel problem with strongly convex lower problem and an auxiliary linear-system state.	SOBA has $\inf_{t\le T}\mathbb{E}\|\nabla h(x^t)\|^2 = O(\log(T)T^{-1/2})$; SABA has averaged $O(1/T)$ squared-gradient rate and linear convergence under a PL condition on the value function.	Our exact full-batch theorem matches the $O(1/T)$ squared-residual pattern and our PL corollary matches the linear-rate pattern, but under an optimistic finite-tilt value-gap formulation rather than a unique smooth response map.
Small hyper-gradients under lower-level PL (Chen et al., 2024a)	Nonconvex–nonconvex bilevel class with lower-level PL geometry; also establishes hardness for broad nonconvex–convex settings lacking strong enough structure.	Reported complexity bounds are $\widetilde{O}(\epsilon^{-2})$, $\widetilde{O}(\epsilon^{-4})$, and $\widetilde{O}(\epsilon^{-6})$ in deterministic, partially stochastic, and fully stochastic settings, respectively.	This supports the paper's cautious framing: without additional lower-level or value-gap structure, small-hypergradient guarantees can be impossible or expensive. Ghost assumes value-adjoint access and proves stationarity for Δ_τ, not a blanket guarantee for every nonsmooth optimistic objective.
Uniformly convex lower level / UniBiO (Wu et al., 2026)	Lower-level uniform convexity with exponent $p \ge 2$, interpolating between strong convexity and general convexity.	The reported stochastic oracle complexity is $\widetilde{O}(\epsilon^{-5p+6})$ for ϵ-stationarity, matching the optimal ϵ dependence for strongly convex lower levels when $p = 2$ under their convention.	UniBiO broadens unique-response-type theory through uniform convexity. Ghost instead targets validation-selected optimistic values and can represent lower-level nonuniqueness through the value gap.

Table 2 (continued).

Method / paper	Main structural regime	Reported stationarity target and rate	Implication for Ghost positioning
Set smoothness (Chen et al., 2025)	Nonsmooth hyper-objectives induced by multiple lower-level solutions under set-smoothness / lower-level error-bound or PL-type structure.	The zeroth-order theorem gives an expected squared approximate-Clarke stationarity measure of order $O(\sqrt{m}(\Delta + 1)/\sqrt{T})$ with neighborhood radius $\delta = O(T^{-1/4})$.	This is a stronger nonsmooth stationarity notion for a broad structural class, but it uses zeroth-order smoothing. Ghost is narrower in stationarity scope but gives a direct first-order value-gap oracle and the standard $O(1/T)$ squared-residual rate when Δ_τ is smooth.
Superquantile–Gibbs relaxation (Masiha et al., 2025)	Minima-selection with multiple lower minimizers under local lower-level PL geometry and intrinsic solution-manifold dimension k.	The current arXiv version reports $(O(\epsilon), \rho)$-Goldstein stationarity in expectation with $\widetilde{O}(m^{15+12k}\epsilon^{-2} \cdot (\epsilon\rho)^{-12k-14})$ lower-level gradient queries, stated for the pessimistic formulation with the optimistic case analogous.	This gives an explicit complexity for minima selection but depends strongly on intrinsic dimension and Langevin-style sampling. Ghost avoids sampling the minimizer manifold but proves stationarity only for the finite-tilt value gap.
Select-then-differentiate (HG-MS) (Masiha et al., 2026)	Local PL with manifold lower-level solution sets; uniqueness of the optimistic selection gives differentiability.	Convergence to stationary points of the optimistic objective, with complexity governed by the intrinsic dimension of the solution manifold.	A stronger stationarity target than ours, obtained under manifold and PL structure and with an explicit selection step.
Composite optimistic regularization (Solla & Royset, 2026)	Composite convex lower-level problem, not necessarily strongly convex, with double regularization and gradient sampling.	The paper establishes convergence of approximately stationary points of a regularized hyper-objective to stationary points of the actual optimistic problem; the stated abstract emphasizes convergence rather than a simple nonasymptotic rate.	Closest in spirit to optimistic regularization. Ghost uses a single scalar value gap $\Delta_\tau = (V_\tau - G)/\tau$ and conservative adjoints, giving a direct first-order projected-rate theorem for the smooth finite-tilt surrogate.
Ghost, this paper	Exact finite-tilt value gap with conservative adjoints; lower-level minimizers may be nonunique; smoothness is required only for the projected convergence theorem on Δ_τ.	$\min_{t<T} \|\mathcal{G}_{\eta,\tau}(\lambda^t)\|^2 = O(1/T)$ for exact projected Ghost, plus $O(T^{-1}\sum_t \rho_t^2)$ for the computed directions of Algorithm 1, with $\rho_t = L_\tau \delta_t \|q_t\| + (\varepsilon_{0,t} + \varepsilon_{\tau,t})/\tau$; linear value contraction under projected PL, transferred to Φ up to b_τ.	The honest claim is not universal rate dominance. The claim is comparable deterministic first-order behavior for the explicit finite-tilt surrogate, plus a more selection-robust objective interface for optimistic nonunique lower-level problems.

The conclusion from Table 2 is that the most defensible positioning is *comparable first-order convergence under more selection-robust modeling*, not a claim of beating the fastest stochastic strongly-convex-lower-level algorithms. Achieving a uniformly better rate would require additional algorithmic ingredients such as variance reduction, momentum, acceleration, or stochastic-gradient control for the ghost value-gap oracle. Those would be valuable extensions, but they are not consequences of the value-gap identity alone. The theorem in the main text is therefore deliberately phrased as a projected-gradient guarantee for the finite-tilt surrogate, plus a PL linear-rate corollary under an explicit additional outer-geometry assumption.

D Additional Experimental Diagnostics and Ablations

This appendix records diagnostics that support the experimental claims in Section 6. The main paper keeps the empirical section compact; here we provide additional toy checks, implementation details, SNLI sensitivity sweeps, and ablations used to interpret the value-gap oracle.

D.1 Which theorem is each experiment testing?

The experiments are read against the three main theoretical statements (Table 3).

Table 3: Theory-to-experiment map. The main paper includes only the compact diagnostics; the supplement keeps sensitivity and ablation details.

Theoretical claim	Diagnostic	Measured or checked empirical signature
Finite-tilt bias, Theorem 1	Sweep τ on exact toy and flat-valley problems and measure $\sup_\lambda \lvert\Phi - \Delta_\tau\rvert$.	The sweeps show decreasing value error as $\tau \downarrow 0$, with the nonunique flat-valley toy used in the main text.
Conservative oracle, Theorem 2	Compare the ghost direction with the exact optimistic direction and with selected-response directions.	The flat-valley diagnostic shows near-perfect Ghost alignment and persistent selected-response misalignment.
Inexact-solve tradeoff, Theorem 3	Vary τ and the lower-level solve budget on SNLI.	The sweeps show a single-peaked tilt curve and strong small-tilt sensitivity to insufficient inner solves.
Compatible readout, Proposition 2 and Theorem 4	Cartesian, independently perturbed, and shared readouts at a kink.	Only the shared readout avoids ascent (Appendix D.8).
Practical utility	SNLI noisy-label hyper-cleaning under a matched framework and fixed 100-epoch horizon.	The table, curves, and frontier report final accuracy, runtime, and trajectory stability.

D.2 Exact toy diagnostics

Table 4 gives an exact τ-sweep for a separate unique-minimizer finite-tilt sanity check. This diagnostic is kept only as a supplementary numerical check; the main-paper toy evidence for optimistic nonuniqueness is the flat-valley experiment in Section 6.1. The uniform approximation error decreases from 2.82×10^{-1} at $\tau = 0.2$ to 1.93×10^{-2} at $\tau = 10^{-2}$. Direction alignment improves at the same time: the median cosine between the ghost direction and the exact outer gradient rises from 0.9989 to 0.999996, and the minimum cosine on the differentiable region exceeds 0.99988 at $\tau = 10^{-2}$.

Figure 5 adds the controlled inexact-solve and bundle diagnostics. The inexact-solve panel replaces exact base and tilted minimizers by a small number of proximal-gradient steps. The mean direction error drops from 1.34×10^{-1} with one step to 1.15×10^{-4} with two steps and 3.48×10^{-6} by four steps, while the cosine is essentially one after two steps. This directly illustrates the perturbation term in Theorem 3. The bundle panel studies deliberately noisy one-step worlds: moderate bundling improves the median direction error, while very large bundles provide diminishing returns under a fixed inner budget.

Table 4: Exact toy diagnostics as a function of the tilt parameter τ.

τ	$\sup_\lambda \lvert\Phi - \Delta_\tau\rvert$	Mean cosine	Median cosine	Minimum cosine
2×10^{-1}	2.821×10^{-1}	0.99827	0.99889	0.98342
1×10^{-1}	1.634×10^{-1}	0.99945	0.99967	0.99277
5×10^{-2}	8.918×10^{-2}	0.99985	0.99991	0.99754
2×10^{-2}	3.785×10^{-2}	0.99997	0.99999	0.99954
1×10^{-2}	1.933×10^{-2}	0.99999	0.999996	0.99988

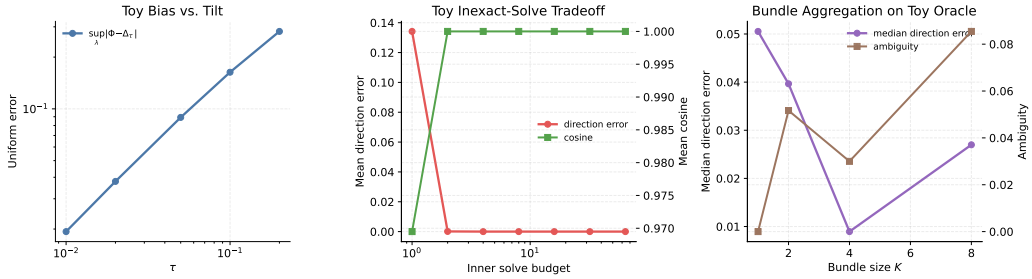


Figure 5: Additional exact and controlled toy diagnostics. Left: finite-tilt bias versus τ. Middle: inexact-solve direction error as the number of lower-level steps increases. Right: bundle aggregation under intentionally noisy one-step worlds.

D.3 A NONUNIQUE FLAT-VALLEY SANITY CHECK

The introduction uses the flat-valley example

$$\mathcal{L}_{\mathrm{tr}}(w) = \max(|w| - 1, 0), \qquad \mathcal{L}_{\mathrm{val}}(w, \lambda) = (w - \lambda)^2. \tag{59}$$

Here $S = [-1, 1]$ and

$$\Phi(\lambda) = \min_{w \in [-1,1]} (w - \lambda)^2 = \mathrm{dist}(\lambda, [-1, 1])^2. \tag{60}$$

A selected-response surrogate can give a different outer problem. For example, the constant selection $s(\lambda) \equiv 0$ gives $J_s(\lambda) = \lambda^2$, which incorrectly makes $\lambda = 0$ the unique optimum. By contrast, the value-gap formulation targets Φ. On any compact interval in λ, and for sufficiently small τ, the tilted minimizer is the closest point in $[-1, 1]$ to λ, so $\Delta_\tau(\lambda) = \Phi(\lambda)$ on that compact set. This toy is useful for motivation because it separates the proposed objective from arbitrary lower-level selections.

Randomized target-dependent flat-valley diagnostic. The main diagnostic in Figure 2 uses the two-parameter version

$$g(w; c, h) = \tfrac{1}{2}[c - h - w]_+^2 + \tfrac{1}{2}[w - c - h]_+^2, \qquad f_t(w) = \tfrac{1}{2}(w - t)^2,$$

so that $S(c, h) = [c - h, c + h]$ for every (c, h). The optimistic response is $\Pi_{[c-h,c+h]}(t)$; the selected-response baseline projects a biased anchor $a < t$ onto the same interval. We draw 256 tasks with $t \sim \mathrm{Unif}[0.85, 1.15]$ and $a = t - \delta$, $\delta \sim \mathrm{Unif}[0.5, 0.9]$. Cosines are reported only on the safe active region

$$\mathcal{A}_\eta(t) = \{(c, h) : \|\nabla_{c,h} \Phi_t(c, h)\|_2 > \eta,\ \mathrm{dist}(t, \{c - h, c + h\}) > \eta\},$$

which removes the zero-gradient validation plateau and projection kink boundaries. This is why the main text can interpret the cosine statistics as genuine directional agreement rather than numerical behavior at nondifferentiable or zero-gradient points.

D.4 SNLI IMPLEMENTATION AND EVALUATION DETAILS

This subsection records the implementation details behind the full-budget summary in Figure 4 and the trajectory plots in Figure 3. These runs use the shared-cache protocol; they read both adjoints at

the unperturbed iterate, with the tilted solve warm-started from the base solve, and do not use the shared outer perturbation of Algorithm 1. The experiments use the preprocessed SNLI benchmark pipeline inherited from the UniBiO comparison suite. In the reported runs, the dataset loader truncates each split to the first 10,000 examples, giving effective split sizes of 10,000 training examples, 10,000 validation examples, and 10,000 test examples. The label-corruption rate is 0.2. The outer variable is one sample-weight logit λ_i per training example, and the training weight is $\sigma(\lambda_i)$.

Validation and test protocol. The 100-epoch results in Figure 4 use the inherited cached files `snli_train_0.2.pkl`, `snli_dev_0.2.pkl`, and `snli_test_0.2.pkl`. In this shared comparison protocol, the training split is noisy, the validation split follows the noise-specific cached split from the comparison code, and the test split is the untouched clean test set duplicated under the noise-tagged filename. Thus the reported final test accuracies are measured on a clean test set, while the validation cache is the shared validation split rather than a newly generated strictly clean validation split. We state this distinction explicitly because the table evaluates the competing oracles under the same inherited pipeline.

Evaluation metrics. Final test accuracy is the accuracy of the final epoch on the clean test cache; it is not the best checkpoint selected by validation or test performance. Final training accuracy is computed against the labels in the loaded noisy training cache, i.e., the labels actually presented to the lower-level solver after corruption. It is therefore an optimization/fit diagnostic for the weighted noisy training objective, not a clean-label training score and not direct evidence that corrupted labels have been recovered. The best-to-final gap used below is the difference between the highest logged test accuracy along a run and the final-epoch test accuracy.

Model and optimizers. All methods use the same recurrent SNLI classifier: fixed 300-dimensional GLOVE Common Crawl word embeddings (Pennington et al., 2014), two recurrent encoding layers, encoder width 4096, classifier hidden dimension 1024, and three output classes. For Ghost, the base and tilted lower-level solves use plain SGD without momentum. The lower model is updated on the weighted training loss in the base world and on the weighted training loss plus the validation tilt in the tilted world. The baselines use the SGD-based implementations included in the UniBiO comparison repository and their corresponding hypergradient update rules.

Batches, inner steps, and Ghost hyperparameters. The reported Ghost runs use batch size 128 and inner batch size 128. VRBO uses inner batch size 256 in the comparison implementation. Ghost uses two lower-level SGD steps for both the base and tilted worlds per outer update. In the comparison code, StocBiO uses three inner steps, TTSA uses one, UniBiO uses one with periodic lower-level refresh (`update_interval` $= 2$), and SABA, MA-SOBA, SUSTAIN, and VRBO use their method-specific internal updates under the same 100-epoch outer horizon. Ghost uses outer learning rate 0.02, inner learning rate 0.1, tilt $\tau = 0.05$, bundle size $K = 1$, two burn-in epochs, one final retraining epoch, model perturbation scale 10^{-3} for extra worlds (this perturbs the model parameters of extra bundle worlds, is inactive at $K = 1$, and is not the outer perturbation of Algorithm 1), and lower-solver gradient clipping norm 1.0. The sample-weight logits are initialized as $\lambda_i = 1$, so the initial weight is $\sigma(1) \approx 0.731$, and logits are clipped after each outer update to $[-6, 6]$.

Tuning and runtime accounting. The baseline hyperparameters are inherited from the UniBiO comparison code and appendix; in that suite, upper- and lower-level learning rates were tuned over $[0.001, 0.1]$, and the selected values are used here. Ghost uses the manually tuned configuration above, selected from a small number of pilot runs rather than a fully logged exhaustive grid. The main-table tilt $\tau = 0.05$ should therefore be read as the pilot-selected configuration for the matched 100-epoch comparison, not as a claim that 0.05 is the globally optimal tilt. The separate sensitivity sweep below shows that smaller tilts can be preferable under its diagnostic fixed-budget setting. The table is therefore a matched-framework comparison with inherited tuned baseline settings and a manually tuned Ghost configuration. The reported wall-clock times are the final 100-epoch training runs only; they exclude hyperparameter-search time, failed pilot runs, and preprocessing outside the final run.

Diagnostic metric definitions. For the stability and cancellation plots, let d_k denote the outer update direction at logged epoch k, and write $\widehat{v} = v/(\|v\|_2 + 10^{-12})$ for numerical normalization. The *direction jitter* at epoch k is

$$J_k = 1 - \langle \widehat{d}_k, \widehat{d}_{k-1} \rangle,$$

so smaller values mean that successive update directions are more stable. The *reference direction* is an offline high-accuracy full Ghost direction $d_{\mathrm{ref},k}$ computed at the same checkpoint and diagnostic batch with contemporaneous base and tilted adjoints and a larger inner-solve budget; it is used only for measurement, never for training. *Reference alignment* is the cosine $\langle \widehat{d}_k, \widehat{d}_{\mathrm{ref},k} \rangle$. The *common-component cancellation* panel reports both

$$\cos(u_{\tau,k}, u_{0,k}) \quad \text{and} \quad \rho_k = \frac{\|u_{\tau,k} - u_{0,k}\|_2}{\|u_{\tau,k}\|_2 + \|u_{0,k}\|_2},$$

where a high cosine together with a small normalized residual ρ_k means that the two adjoints share a large training-value component that is removed by subtraction. Finally, because exact distances to conservative adjoint fields are unavailable in deep SNLI, the *inexact-oracle proxy* in Figure 7 uses stationarity residuals rather than exact adjoint errors:

$$\widehat{E}_\tau(m) = \frac{\widehat{\epsilon}_0(m) + \widehat{\epsilon}_\tau(m)}{\tau},$$

$$\widehat{\epsilon}_0(m) = \|\nabla_w \mathcal{L}_{\mathrm{tr}}(w_0^{(m)}, \lambda)\|_2, \qquad \widehat{\epsilon}_\tau(m) = \|\nabla_w (\mathcal{L}_{\mathrm{tr}} + \tau \mathcal{L}_{\mathrm{val}})(w_\tau^{(m)}, \lambda)\|_2,$$

evaluated on the diagnostic batch after m lower-level steps and averaged over logged states. This proxy has the same scaling as the inexact bound in Theorem 3, but it should be read as a solver-residual diagnostic rather than a direct estimate of the theorem's ϵ_0 and ϵ_τ.

D.5 Reproducibility, assets, and broader-impact notes

Open-source status. The supplementary material contains anonymized code for the shared-cache SNLI runs: Ghost and the baselines (adapted from the UniBiO comparison code), data caching, and result aggregation, together with the standard-library scripts of Appendices C.4 and D.8. As in those runs, the Ghost code reads both adjoints at the unperturbed iterate (Section 6.2). Code and per-seed logs of the clean-validation runs are not included.

Existing assets and licenses. The experiments use the Stanford Natural Language Inference corpus version 1.0 (Bowman et al., 2015), which is distributed under the Creative Commons Attribution–ShareAlike 4.0 International license at `https://nlp.stanford.edu/projects/snli/`. The shared pipeline uses GloVe `glove.840B.300d` word vectors (Pennington et al., 2014); the GloVe project page states that pretrained vectors are distributed under the Public Domain Dedication and License, while the GloVe software package is Apache-2.0 licensed. The digits control in Appendix D.8 uses the UCI Optical Recognition of Handwritten Digits data (Alpaydin & Kaynak, 1998), distributed under the Creative Commons Attribution 4.0 license. The UniBiO comparison implementation is a third-party MIT-licensed research codebase linked from Wu et al. (2026). We credit the original papers. The datasets and embeddings are not redistributed; the adapted UniBiO code in the supplementary material keeps its MIT license notice.

Compute accounting. The exact toy diagnostics are closed-form or vectorized calculations and do not require GPU resources. The reported SNLI main runs use one NVIDIA A100 80GB GPU, a 100-epoch outer horizon, and the per-method final-run wall-clock times in Figure 4. These wall-clock times exclude preprocessing, hyperparameter search, and failed pilot runs; those exploratory costs were not fully logged and should not be inferred from the final-run table.

Societal scope and safeguards. The work is methodological and targets more reliable validation-driven optimization, with possible positive effects on robustness and compute efficiency in data reweighting or hyperparameter selection. The main deployment risk is that any validation-driven reweighting method can inherit biases, privacy issues, or label artifacts from the validation set and may overfit to an unrepresentative validation protocol. The paper does not release a pretrained language model, generative model, scraped dataset, or high-risk dual-use artifact. Responsible use should audit validation data, monitor subgroup behavior where relevant, and treat learned example weights as optimization diagnostics rather than causal judgments about individual examples.

Human-subject status. No new crowdsourcing, data collection, or human-subject experiment was conducted for this work. The human-authored text examples used in the experiments come only through the existing public SNLI benchmark and its inherited comparison cache.

D.6 Additional SNLI summaries from the main runs

The cropped wall-clock trajectories in Figure 3 should be read only as an early-time comparison: Ghost reaches its high-accuracy regime within the displayed window, whereas several slower baselines continue beyond the cutoff and are summarized by their full-budget endpoints.

The main text reports the core SNLI table, performance curves, and efficiency frontier. Table 5 gives three derived diagnostics from the same five-run summary. The accuracy gap is measured relative to Ghost's final test accuracy. Final accuracy per hour is a coarse wall-clock efficiency score in percentage-points per hour. Runtime versus Ghost normalizes the wall-clock cost.

Table 5: Derived SNLI diagnostics from the five-run final-epoch summaries in Figure 4.

Method	Gap to Ghost (pts) ↓	Final acc./hour ↑	Runtime vs. Ghost ↓
Ghost	**0.00**	**30.39**	**1.00**×
UniBiO	0.55	23.26	1.30×
VRBO	1.30	4.68	6.37×
MA-SOBA	2.58	21.73	1.35×
SABA	2.73	21.21	1.38×
TTSA	3.36	12.94	2.24×
SUSTAIN	3.57	7.22	4.00×
StocBiO	6.71	10.75	2.56×

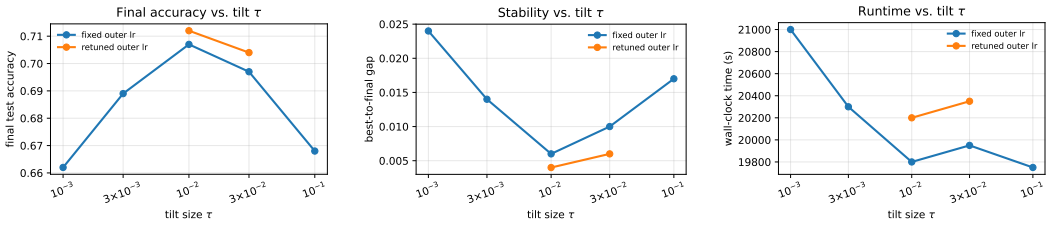
The derived table is consistent with the main efficiency plot: Ghost is the only method that simultaneously has the highest final test accuracy and the lowest runtime. UniBiO is close in final accuracy but slower. VRBO is competitive in final accuracy but much more expensive. MA-SOBA and SABA are efficient but finish several points below Ghost.

D.7 SNLI sensitivity and ablation studies

This subsection reports measured sensitivity diagnostics around the same SNLI hyper-cleaning pipeline. These studies help explain when the finite-tilt oracle is accurate and stable, and they are read as ablations of the Ghost design rather than as additional leaderboard comparisons.

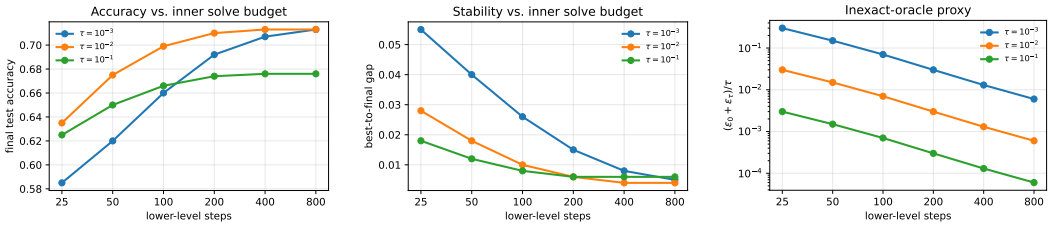
Tilt sensitivity on SNLI. Figure 6 sweeps the tilt parameter while keeping the inner budget fixed. With a fixed outer learning rate, final accuracy has a clear interior optimum: very small tilts are noisy because the adjoint difference is divided by τ, while very large tilts introduce visible finite-tilt bias. The measured peak is around $\tau = 10^{-2}$. This diagnostic does not contradict the main run at $\tau = 0.05$: the main run uses the pilot-selected configuration for the matched comparison, whereas this sweep is a separate fixed-budget diagnostic intended to expose the bias–accuracy tradeoff. Retuning the outer learning rate for the two strongest tilt values improves the same region further, pushing the 10^{-2} run slightly above the fixed-learning-rate curve and keeping 3×10^{-2} competitive. The stability panel tells the same story: the best-to-final gap is largest for the very small and very large tilts and is smallest near the accuracy optimum. Runtime is nearly flat within this diagnostic because the inner budget is fixed; the wall-clock panel should be compared across tilt values in the sweep, not against the final-run runtimes in Figure 4.

Inner-solve accuracy. Figure 7 varies the lower-level solve budget for three representative tilts. The smallest tilt, $\tau = 10^{-3}$, performs poorly with only 25–50 lower-level steps but improves sharply as the lower-level solve becomes more accurate. The intermediate tilt, $\tau = 10^{-2}$, is the most reliable across moderate and large inner budgets and reaches the best final accuracy once the solve budget is sufficient. The largest tilt, $\tau = 10^{-1}$, is comparatively robust to small budgets but saturates at a lower accuracy, consistent with finite-tilt bias. The measured inexact-oracle proxy defined above decreases monotonically with the inner budget and is largest for the smallest tilt, qualitatively matching the $(\epsilon_0 + \epsilon_\tau)/\tau$ scaling in the inexact-oracle bound.



(a) Final accuracy (b) Best-to-final gap (c) Wall-clock time

Figure 6: **Tilt sensitivity on SNLI.** We sweep τ under a fixed outer learning rate and then retune the outer learning rate for the two strongest tilt values. Final accuracy peaks near $\tau = 10^{-2}$, while both overly small and overly large tilts are less stable. The sweep is a diagnostic fixed-budget study and is not the same runtime setting as the main five-run comparison. This matches the finite-tilt tradeoff in Theorem 3: large τ increases surrogate bias, whereas tiny τ magnifies solve and adjoint noise through the factor $1/\tau$.



(a) Final accuracy (b) Best-to-final gap (c) Inexact-oracle proxy

Figure 7: **Inner-solve sensitivity on SNLI.** We vary the lower-level solve budget for $\tau \in \{10^{-3}, 10^{-2}, 10^{-1}\}$. Increasing the budget improves accuracy and reduces best-to-final degradation. The stationarity-residual proxy $(\widehat{\epsilon}_0 + \widehat{\epsilon}_\tau)/\tau$ falls with the number of lower-level steps and is largest for $\tau = 10^{-3}$, explaining why a very small tilt only becomes competitive after enough lower-level optimization.

Cancellation ablation. A natural concern is that Ghost is merely another implicit hypergradient computed on a tilted lower-level problem. Figure 8 isolates the subtraction mechanism by comparing the full direction

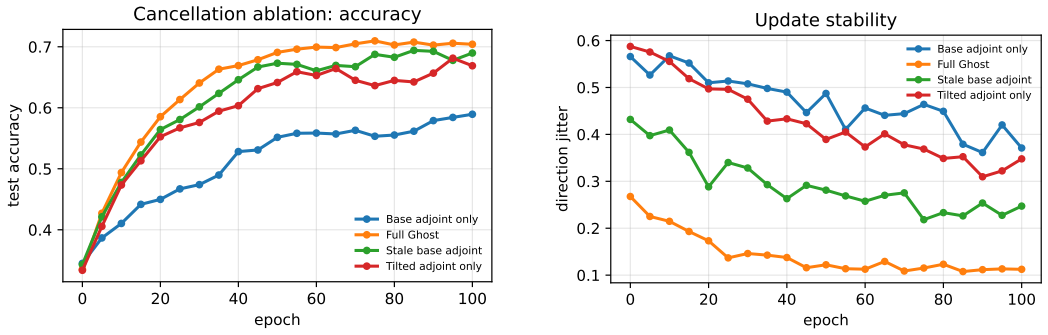
$$d_{\text{full}}(\lambda) = \frac{u_\tau(\lambda) - u_0(\lambda)}{\tau}$$

against a tilted-adjoint update using u_τ alone, a base-adjoint update using u_0 alone, and a stale-base update $(u_\tau - \tilde{u}_0)/\tau$. The full difference gives the strongest and most stable trajectory: it reaches the highest test accuracy, has the lowest direction jitter, and maintains the best alignment with the reference direction. The base-only update is substantially misaligned and remains at lower accuracy; the tilted-only and stale-base variants improve over base-only but lose stability and alignment relative to the contemporaneous difference. The common-component diagnostic explains why: u_τ and u_0 are highly correlated throughout training, while the normalized difference remains much smaller, indicating that the subtraction removes a large shared training-value component and exposes the validation-induced tie-breaking signal.

D.8 Clean-validation protocol, pairing check, and controls

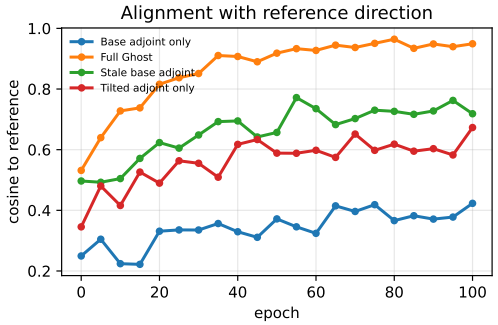
The results in this subsection are kept apart from the shared-cache runs above. Per-seed logs and configurations of the clean-validation runs and the controls are not included; the pairing check is an exact computation with a standard-library script in the supplementary material.

Clean-validation SNLI protocol. Table 6 uses a validation split with clean labels, 20% training-label corruption, 100 outer epochs, five seeds, and one A100 GPU per run; final test accuracy is the final-epoch accuracy on the clean test set, and time is final-run wall-clock excluding tuning and preprocessing. Compatible Ghost reads (12) on minibatches with the shared perturbation of Algorithm 1, normalizes by the batch size (which rescales the outer step), and updates the logits

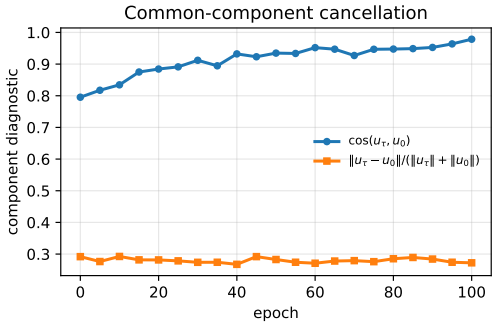


(a) Test accuracy

(b) Direction jitter



(c) Reference alignment



(d) Common-component cancellation

Figure 8: **Cancellation ablation on SNLI.** The full ghost direction outperforms the one-adjoint and stale-base variants in accuracy, update stability, and alignment with a reference direction. Panel (d) shows that u_τ and u_0 share a large common component; subtracting the contemporaneous base adjoint is therefore not a cosmetic two-adjoint trick, but the operation that isolates the validation-induced signal.

of the current batch; this stochastic block-coordinate variant lies outside the deterministic theory. The independent-pair row forms the same difference without the shared perturbation. Compatible Ghost has the highest mean final test accuracy, 0.36 and 0.52 percentage points above HG-MS and SQ-G; these gaps are of the order of the across-seed spreads, and we do not claim significance. It has the shortest runtime among the external baselines; equal epochs are not equal wall-clock or tuning budgets. The independent-pair ablation is 3% faster, 0.67 percentage points lower, and has about twice the spread, a single five-seed contrast that is consistent with pairing mattering beyond the constructed kink. SABA and SUSTAIN were not run under this protocol.

Table 6: Clean-validation SNLI hyper-cleaning with 20% training-label corruption, 100 outer epochs, five seeds. Mean $\pm$ standard deviation across seeds; time is final-run wall-clock, relative to compatible Ghost.

Method	Final test accuracy (%)	Time (h)	Relative time
Ghost, compatible (Algorithm 1)	$\mathbf{71.08 \pm 0.18}$	2.41 ± 0.04	$1.00\times$
Ghost, independent pair (ablation)	70.41 ± 0.34	2.33 ± 0.03	$0.97\times$
HG-MS	70.72 ± 0.24	5.86 ± 0.19	$2.43\times$
SQ-G	70.56 ± 0.28	8.94 ± 0.32	$3.71\times$
UniBiO	70.21 ± 0.11	3.05 ± 0.10	$1.27\times$
VRBO	69.47 ± 0.14	14.91 ± 0.27	$6.19\times$
MA-SOBA	68.18 ± 0.20	3.16 ± 0.05	$1.31\times$
TTSA	67.39 ± 0.16	5.26 ± 0.18	$2.18\times$
StocBiO	64.03 ± 0.30	6.01 ± 0.11	$2.49\times$

Pairing at a kink. On Example 1 with $\tau = 0.1$, $\eta = 0.01$, $\delta = 10^{-3}$, $q, q' \sim \mathrm{Unif}[-1, 1]$, 100 projected steps from $\lambda^0 = 0$, and 10^4 runs with a fixed seed, we compare three ways to read the two exact adjoints: the Cartesian rule picks each readout at the kink uniformly from the one-sided branches $\{-1, +1\}$; the independent rule reads u_0 at $\lambda + \delta q$ and u_τ at $\lambda + \delta q'$; the shared rule reads both at $\lambda + \delta q$. Of the four Cartesian vertex pairs at $\lambda = 0$, exactly two give $d = \pm 2/\tau$ and move the iterate to $|\lambda| = 0.2$, where $\Delta_\tau = 0.02$. Every run of the Cartesian and independent rules makes exactly one such jump (median final Δ_τ of 2.8×10^{-3} and 2.7×10^{-3}); no run of the shared rule does, its largest Δ_τ over all runs is 1.3×10^{-8}, and its first-step direction error is at most δ, as Theorem 4 predicts with $L_\tau = 1$. Randomization alone therefore does not repair the pairing; sharing the evaluation point does.

Further controls. Across 50 random strongly convex quadratic problems, the error between the Ghost direction and the implicit hypergradient has median log–log slope 0.998 in τ (median relative error 5.47×10^{-5} at $\tau = 10^{-4}$): consistency in the singleton regime, not a selection advantage. On affine solution manifolds $\{(x, z) : x = \lambda, z \in \mathbb{R}^k\}$, $k \in \{8, 16, 64\}$, Ghost has median cosine 1.000 with the exact optimistic pseudoinverse direction, a valid selected branch -0.877, -0.875, -0.875, and a Gibbs–Monte Carlo selection proxy with 512 samples, which is not the full SQ-G method, degrades from 0.582 to 0.115. Across 800 controlled strongly convex certificates, the residual-based bound of Assumption 2 held in every case (log correlation 0.997, largest error-to-bound ratio 0.535). On 20 held-out splits of the UCI Optical Recognition of Handwritten Digits data (Alpaydin & Kaynak, 1998) (8×8 images, digits 3 versus 8; 50 training and 100 clean-validation examples, 20% label flips, 65 parameters) with a median lower-level nullity of 15, a fixed budget of 100 outer steps gives Ghost $95.72 \pm 1.96\%$ test accuracy against $69.49 \pm 9.95\%$ for the exact optimistic direction (paired difference 26.23 points, two-sided paired t-test, 95% CI 21.55–30.91), while at closed-form convergence the two are essentially tied (95.72% versus 95.70%); the gain reflects faster progress under a finite budget, not a better target.