CATEGORICAL REGRESSION IN REBEL: RATE PRESERVATION, BOUNDARY BIAS, AND NO GENERAL ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

REBEL reduces policy improvement to squared-loss regression of reward gaps onto policy log-ratio differences and asks whether cross-entropy does better. We answer for the tied HL-Gauss loss, which encodes the scalar as a cross-entropy-trained histogram. Inside the support the loss is two-sidedly calibrated to the squared loss with constants free of support width and bin count, so REBEL's bound carries over, with its square-root rate sharp under the stated assumption class. At the boundary the encoding winsorizes noisy targets, a bias absent from squared regression, bounded by how far and how often targets cross the boundary. Preregistered bandit experiments confirm this picture: under symmetric contamination HL trails MSE in all 20 seeds, by 3.4% of clean regret on average. One-sided, feature-correlated spikes give a replicated benefit of 0.7% of clean regret near the interpolation threshold, below the preregistered smallest effect of interest and, in the registered runs, not separated from curvature-matched Huber. Widening the support attenuates both effects; at LLM scale HL and Huber are equivalent within the registered margin of ± 0.035 clean-reward units. The replacement keeps REBEL's rate but buys no general robustness: what matters is where a loss clamps, not whether it is categorical; curvature-matched Huber remains the conservative default.

1 INTRODUCTION

REBEL (Gao et al., 2024) reduces policy improvement to iterated least squares: it regresses relative rewards $a = r(x, y) - r(x, y')$ onto policy log-ratio differences, and the exact minimizer is a mirror-descent update (Kakade, 2001). Its authors pose as their first open question whether a loss other than the square loss, "e.g. log loss or cross-entropy", could yield better practical performance or tighter theoretical bounds. On the value side of deep RL, HL-Gauss cross-entropy, a histogram loss with Gaussian-smoothed targets, is reported to improve learning broadly (Farebrother et al., 2024; Imani & White, 2018). The tempting transfer is that the policy-side regression should inherit that robustness for free.

One idea organizes the answer. Inside the support, the tied HL-Gauss loss is locally quadratic in the scalar it fits, so nothing changes: the guarantee, its rate, and its lower bounds are those of squared regression. At the boundary, the finite support winsorizes the target, so the two losses differ only when noisy targets reach it, and the sign of the difference depends on the structure of the noise (Figure 1).

Contributions.

- *Rate preservation.* A two-sided calibration with constants free of the support width and bin count (Theorem 1) carries REBEL's bound over with transfer coefficient one (Theorem 2). A matching construction shows the square-root rate is sharp in this bound format (Theorem 3), and a uniform reward gap restores first-order rates (Theorem 4).
- *Boundary bias.* With noisy targets the best scalar fit is displaced inward from the target mean even under symmetric noise (Proposition 2); Theorem 6 bounds the displacement by how far, and how often, targets cross into the compression zone of the support.
- *A preregistered phase diagram.* On a linear bandit and an on-policy MLP, the sign of HL$-$MSE depends on the contamination structure and its size on how often targets reach the boundary.

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.

Support widening and a two-hot encoding show that the categorical form is not the operative variable. The first registration's prediction failed and is reported, the second replicated on fresh seeds, and every result is in Table 1.

Related work. REBEL (Gao et al., 2024) belongs to the NPG/mirror-descent lineage (Kakade, 2001; Schulman et al., 2015; 2017); DRO-REBEL (Sahu & Wells, 2025) robustifies it distributionally while keeping the squared loss, and DPO (Rafailov et al., 2023) fits a logistic link on the same log-ratio object. HL-Gauss (Farebrother et al., 2024; Imani & White, 2018) and regression-as-classification more broadly (Stewart et al., 2023; Hui & Belkin, 2021) report conditional gains on the value side, where the categorical head is untied. The categorical projection of distributional RL (Rowland et al., 2018; Lyle et al., 2019) appears here as a policy-side regression target whose boundary, not its expressiveness, is the operative axis. The small-loss guarantees that the log loss yields elsewhere (Foster & Krishnamurthy, 2021; Wang et al., 2023; 2024) are structurally unavailable in REBEL's bound format (Corollary C.1). Huber's estimator (Huber, 1964) is the residual-space robust arm; two-hot targets (Schrittwieser et al., 2020; Hafner et al., 2025) and reward clipping (Mnih et al., 2015) are the target-space relatives of the winsorization studied here. Extended discussion: Appendix I. Evidence status per claim: Table 4.

2 Setup and the HL-REBEL objective

The REBEL reduction Let $x \sim \rho$ be a context, $\mathcal{Y}$ a finite or countable response space, and $r : \mathcal{X} \times \mathcal{Y} \to [-R_{\max}, R_{\max}]$ a deterministic reward, with objective $J(\pi) := \mathbb{E}_{x\sim\rho,\, y\sim\pi(\cdot|x)}[r(x, y)]$. At iteration t, REBEL (Gao et al., 2024) draws triples $(x, y, y') \sim \mathcal{D}_t := \rho \otimes \pi_t \otimes \mu_t$, with π_t the current policy and μ_t a base distribution, and regresses relative rewards onto policy log-ratio differences. With $f_{\theta,t}(x, y) := \frac{1}{\eta} \log \frac{\pi_\theta(y|x)}{\pi_t(y|x)}$, $g_{\theta,t}(x, y, y') := f_{\theta,t}(x, y) - f_{\theta,t}(x, y')$, and $a_t(x, y, y') := r(x, y) - r(x, y')$, squared-loss REBEL approximately minimizes $\mathbb{E}_{\mathcal{D}_t}(g_{\theta,t} - a_t)^2$; exact minimization recovers the mirror-descent update $\pi_{t+1} \propto \pi_t e^{\eta r}$, and $|a_t| \le 2R_{\max}$. The identity requires support preservation; softmax and autoregressive softmax classes qualify, with $\mathcal{S}_x = \mathcal{Y}$.

Assumption 1 (Common support and support preservation)**.** For ρ-almost every x there is a nonempty $\mathcal{S}_x \subseteq \mathcal{Y}$ such that every policy in the class has support exactly $\mathcal{S}_x$; for every t, $\operatorname{supp} \mu_t(\cdot \mid x) \subseteq \mathcal{S}_x$; and every comparator satisfies $\operatorname{supp} \pi^\star(\cdot \mid x) \subseteq \mathcal{S}_x$.

The normalized HL-Gauss map Following Farebrother et al. (2024) and Imani & White (2018), fix a symmetric support $I = [-W, W]$, an integer $m \ge 2$, and a scale $\sigma > 0$; partition I into m bins of width $\varsigma := 2W/m$ and set $\beta := \sigma/\varsigma$. For $s \in \mathbb{R}$, let $P_i(s)$ be the Gaussian mass of bin i at location s and $Z(s) := \sum_i P_i(s)$. The encoding is the bin-index law of $\mathcal{N}(s, \sigma^2)$ truncated to I,

$$\Gamma_\sigma(s) := (p_1(s), \dots, p_m(s)), \qquad p_i(s) := P_i(s)/Z(s).$$

The normalization makes $\mathrm{CE}(q, p) - \mathrm{H}(q) = \mathrm{KL}(q\|p)$ an identity on Δ^m; the *unnormalized* masses define a different, displaced objective (F.2). Fix $c_\delta > 0$ and put $\delta := c_\delta\sigma$ and $I_\delta := [-W + \delta, W - \delta]$, the interior; the part of the support outside I_δ is the *compression zone*. The configuration used throughout is $(\beta, c_\delta) = (0.75, 3)$. The support is designed so that targets and predictions live in I_δ: choosing a design bound A_{des} before training requires $W \ge \max\{2R_{\max}, 2A_{\mathrm{des}}\} + c_\delta\sigma$, and oversizing at fixed (σ, β) is free: a larger support costs bins, never constants.

The tied training objective HL-REBEL keeps the reduction and replaces only the loss. At iteration t it approximately minimizes $\widehat{L}_t(\theta) := \sum_{(x,y,y')\in\widehat{\mathcal{D}}_t} \mathrm{CE}(\Gamma_\sigma(a_t), \Gamma_\sigma(g_{\theta,t}))$. The target is an optimization constant, and the prediction differentiates through the tied scalar $g_{\theta,t}$ alone; there is no independent categorical head (Figure 1a; Algorithm 1, Appendix H). Tying is essential: an untied m-way head decouples the loss from the policy, and Proposition 1 requires the tie; the tested decoded-mean variant pays a uniform price but also changes the potential class and parameter count (Appendix H.5.6). For the analysis we impose, as in the squared-loss case:

Assumption 2 (Population categorical risk)**.** For every t there is $\varepsilon_t^{\mathrm{CE}} \ge 0$ with $\mathbb{E}_{\mathcal{D}_t} \mathrm{KL}(\Gamma_\sigma(a_t)\|\Gamma_\sigma(g_t)) \le \varepsilon_t^{\mathrm{CE}}$.

Assumption 3 (Bounded centered update)**.** For every t, the centered update $\widetilde{A}_t := f_t - \mathbb{E}_{y\sim\pi_t(\cdot|x)} f_t$ satisfies $|\widetilde{A}_t| \le A_t \le A_{\mathrm{des}}$.

Assumption 3 is REBEL's own mirror-descent hypothesis, reused to size the support so that $g_t, a_t \in I_\delta$.

Proposition 1 (Exact fit recovers mirror descent). *Under Assumption 1, if the categorical fit is exact, $\mathrm{KL}(\Gamma_\sigma(a_t) \,\|\, \Gamma_\sigma(g_t)) = 0$ for $\mathcal{D}_t$-almost every pair, then for ρ-almost every x and every $y \in \mathcal{S}_x$, $\pi_{t+1}(y \mid x) = \pi_t(y \mid x)\, e^{\eta r(x,y)} / \sum_{z \in \mathcal{S}_x} \pi_t(z \mid x)\, e^{\eta r(x,z)}$.*

Proof: A.4.

Notation.

$W,\ m,\ \varsigma = 2W/m$	support half-width, bin count, bin width	$a_t,\ g_t$	reward gap, fitted log-ratio difference
$\sigma,\ \beta = \sigma/\varsigma$	smoothing scale, its ratio to the bin width (0.75)	$\alpha_{\mathrm{HL}},\ \kappa_{\mathrm{HL}}$	calibration constants (Theorem 1)
$c_\delta,\ \delta = c_\delta\sigma,\ I_\delta$	margin multiplier (3), margin, interior	$\varepsilon_t^{\mathrm{CE}},\ A_t,\ A_{\mathrm{des}}$	risk budget, update bound, design value
$\Gamma_\sigma(s)$	bin law of $\mathcal{N}(s, \sigma^2)$ truncated to I	$C_t(\pi^\star),\ D_0(\pi^\star)$	coverage, initial KL of the comparator

Full guide: Appendix H.2.

3 Rates are preserved: calibration, reduction, and lower bounds

Complete proofs: Appendices A–D.

3.1 Two-sided calibration

For a categorical target $q \in \Delta^m$ write $\ell_q(g) := \mathrm{CE}(q, \Gamma_\sigma(g))$, so that $\mathrm{KL}(\Gamma_\sigma(u) \| \Gamma_\sigma(v)) = \ell_{\Gamma_\sigma(u)}(v) - \ell_{\Gamma_\sigma(u)}(u)$. Let $v(c_\delta)$ be the explicit lower bound of Lemma A.2 on the truncated Gaussian's variance relative to σ^2 at margin c_δ; $v(3) = 0.970630$. The calibration constants are $\alpha_{\mathrm{HL}} := (v(c_\delta) - 1/(4\beta^2))/\sigma^2$ and $\kappa_{\mathrm{HL}} := 2/\alpha_{\mathrm{HL}} = 2\sigma^2/(v(c_\delta) - 1/(4\beta^2))$, defined whenever $v(c_\delta) > 1/(4\beta^2)$; $\kappa_{\mathrm{HL}} = \Theta(\sigma^2)$, and at $(\beta, c_\delta) = (0.75, 3)$ it equals $3.8009\sigma^2$.

Theorem 1 (Two-sided HL-Gauss calibration). *Let $m \geq 2$ and use the matched normalized map Γ_σ.*

1. *For all $u, v \in \mathbb{R}$, $\mathrm{KL}(\Gamma_\sigma(u) \| \Gamma_\sigma(v)) \leq \dfrac{(u-v)^2}{2\sigma^2}$.*
2. *If $v(c_\delta) > 1/(4\beta^2)$ and $u, v \in I_\delta$, then $(u-v)^2 \leq \kappa_{\mathrm{HL}}\ \mathrm{KL}(\Gamma_\sigma(u) \| \Gamma_\sigma(v))$.*
3. *For every $q \in \Delta^m$, $g \mapsto \ell_q(g)$ is α_{HL}-strongly convex on I_δ, and its derivative is globally $1/\sigma^2$-Lipschitz.*

Proof idea. Conditioned on a bin, the truncated Gaussian is a one-parameter exponential family in s/σ^2, so the derivative and curvature of ℓ_q are exact conditional-mean and conditional-variance identities (Lemma A.1). Popoviciu's inequality (Popoviciu, 1935) bounds the within-bin variance by $\varsigma^2/4$, the Brascamp–Lieb inequality (Brascamp & Lieb, 1976) bounds every variance by σ^2, and on I_δ the full truncated variance is at least $v(c_\delta)\sigma^2$ (Lemma A.2). Together these give $\alpha_{\mathrm{HL}} \leq \ell_q'' \leq 1/\sigma^2$, and Taylor's theorem does the rest (Appendix A).

Remarks. W and m enter only through $\beta = \sigma/\varsigma$, so refining the bins or enlarging the support at fixed $(\sigma, \beta, c_\delta)$ leaves κ_{HL} unchanged (confirmed numerically over 84 configurations, Appendix H.6). The compact range is necessary, since no finite constant works on all of $\mathbb{R}$ (Proposition A.5), and the statement is a two-sided, explicit-constant member of the classification-calibration lineage (Zhang, 2004; Bartlett et al., 2006; Steinwart, 2007).

Corollary 1 (CE-to-pairwise-L_2 bridge). *Under Assumptions 1–3 and the support design, $\mathbb{E}_{\mathcal{D}_t}(g_t - a_t)^2 \leq \kappa_{\mathrm{HL}}\, \varepsilon_t^{\mathrm{CE}}$ for every t.*

No decoded-mean, clipping, projection, or discretization term appears; on the experimental platform the realized aggregate ratio is $0.543\,\kappa_{\mathrm{HL}}$ (Appendix H.6).

3.2 Performance bound

For a comparator $\pi^\star$, let $D_0(\pi^\star) := \mathbb{E}_{x \sim \rho}\, \mathrm{KL}(\pi^\star(\cdot \mid x) \| \pi_0(\cdot \mid x))$ and $C_t(\pi^\star) := \operatorname{ess\,sup}_{x,y}\ \pi^\star(y \mid x)/\mu_t(y \mid x)$.

Theorem 2 (Agnostic HL-REBEL performance bound)**.** *Fix $\eta > 0$ and a comparator $\pi^\star$ with $C_t(\pi^\star) < \infty$ for every t and $D_0(\pi^\star) < \infty$. Under Assumptions 1–3 and the support design, the uniform mixture $\bar{\pi}_T := \frac{1}{T}\sum_{t=0}^{T-1} \pi_t$ satisfies*

$$J(\pi^\star) - J(\bar{\pi}_T) \le \frac{D_0(\pi^\star)}{\eta T} + \frac{\eta}{2T}\sum_{t=0}^{T-1} A_t^2 + \frac{1}{T}\sum_{t=0}^{T-1}\sqrt{C_t(\pi^\star)\,\kappa_{\mathrm{HL}}\,\varepsilon_t^{\mathrm{CE}}},$$

and some iterate $\pi_{\widehat{t}}$ has value at least $J(\bar{\pi}_T)$ and obeys the same bound.

With uniform constants ($A_t \le A$, $C_t(\pi^\star) \le C$, $\varepsilon_t^{\mathrm{CE}} \le \varepsilon_{\mathrm{CE}}$, the same constants valid over the candidate step sizes), $\eta = \sqrt{2D_0(\pi^\star)/(A^2T)}$ gives $J(\pi^\star) - J(\bar{\pi}_T) \le A\sqrt{2D_0(\pi^\star)/T} + \sqrt{C\,\kappa_{\mathrm{HL}}\,\varepsilon_{\mathrm{CE}}}$ (Corollary 2, B.4).

The bound has the squared-loss REBEL shape, with $\kappa_{\mathrm{HL}}\varepsilon_{\mathrm{CE}}$ in place of the pairwise mean squared error; on the same predictor the two regression terms differ by at most the factor $\sqrt{\kappa_{\mathrm{HL}}/(2\sigma^2)} = 1.38$ (Appendix H.3). Only one step of the proof depends on the loss, the passage from categorical risk to pairwise squared error via Theorem 1; every other step is loss-agnostic (Table 3). Two constants improve on REBEL's own proof and carry over to the squared loss verbatim. Applying the decomposition of Lemma B.1 once, jointly, gives transfer coefficient 1 where REBEL's term-by-term application lands at $\sqrt{10\,C\varepsilon}$; Hoeffding's lemma (Lemma B.2) frees the bound of the side condition $\eta A_t \le 1$ and of initialization beyond $D_0(\pi^\star)$. As in squared-loss REBEL, $\varepsilon_t^{\mathrm{CE}}$ is a supplied population quantity (Assumption 2); the fitting step itself is compared under matched sample and optimizer budgets (Section 5).

3.3 Sharpness of the square-root term

Theorem 3 (informal; full statement and proof in Appendix C). *For every coverage level $C \ge 2$, admissible HL-Gauss parameters, and every sufficiently small $\varepsilon > 0$, there is a one-context C-action instance with deterministic reward, a full-support tabular softmax class, a uniform base distribution, and a deterministic comparator $\pi^\star$ on which every iterate satisfies all hypotheses of Theorem 2 with $C_t(\pi^\star) = C$, categorical risk at most ε, and $A_t \le \sigma\sqrt{C\varepsilon}$, yet has regret $J(\pi^\star) - J(\pi_t) \ge \frac{\sigma}{4}\sqrt{C\varepsilon}$.*

Theorems 2 and 3 together settle the rate question *under the stated assumption class*: the construction satisfies every hypothesis of Theorem 2 with the stated budgets, so no analysis relying only on them can improve the exponent. This is sharpness within the stated class, not an information-theoretic lower bound over all algorithms. Sharpness here is exponent-level: κ_{HL} is explicit and (W, m)-free, and no lower bound is claimed for the constant itself. Coverage does not rescue the rate: Proposition C.2 exhibits a two-action instance with *perfect* coverage ($C_t(\pi^\star) = 1$) and the same square root.

First-order (small-loss) bounds are unavailable in the stated bound format. Relative rewards are specified only up to a common offset that cancels from every pairwise target, update, and regret, so the instance of Theorem 3 can be rewritten in cost form with optimal cost zero. There a small-loss bound $O(\sqrt{L^\star\varepsilon} + \varepsilon)$ (Foster & Krishnamurthy, 2021; Wang et al., 2023; 2024) would force regret $O(\varepsilon)$ against $\Omega(\sqrt{\varepsilon})$ (Corollary C.1). The argument uses only local quadraticity, so it covers every smooth real-gap loss and, in the same supplied-budget format, the preference-label (DPO-style) reduction (Remark C.3); label-noise adaptivity is left open.

3.4 First-order rates under a uniform gap

Faster budget dependence requires reward geometry.

Theorem 4 (uniform reward gap; proof in Appendix D.2). *For a deterministic comparator $\pi^\star = \delta_{y^\star(x)}$ with reward gap in $[\Delta, B_r]$ over every other action, $C_t^\star := \operatorname{ess\,sup}_x 1/\mu_t(y^\star(x) \mid x) < \infty$, $A_t \le A$, $\eta A \le 1$, and $\eta \le 1/(8B_r)$,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\big[J(\pi^\star) - J(\pi_t)\big] \le \frac{2D_0(\pi^\star)}{\eta T} + \frac{4\eta}{T}\sum_{t=0}^{T-1}\kappa_{\mathrm{HL}}\varepsilon_t^{\mathrm{CE}} + \frac{2}{\Delta T}\sum_{t=0}^{T-1} C_t^\star\,\kappa_{\mathrm{HL}}\varepsilon_t^{\mathrm{CE}}.$$

The $1/\Delta$ shape is necessary (Proposition D.3), a Tsybakov-style margin (Tsybakov, 2004) interpolates between the agnostic and gapped regimes with exponent $(\nu+1)/(\nu+2)$ (Theorem D.2), and a greedy readout attains the same $1/\Delta$ shape (Theorem D.4).

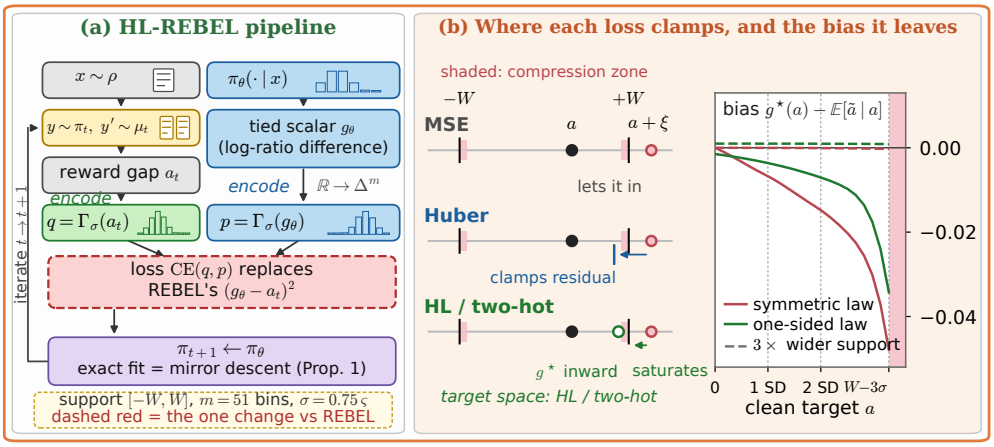


Figure 1: **HL-REBEL at a glance.** (a) Only the loss changes: gap a_t and tied scalar g_θ pass through the same map Γ_σ; an exact fit is mirror descent (Proposition 1). (b) Boundary saturation is target-space winsorization: MSE lets a spike in, Huber clamps the residual, HL and two-hot saturate at $\pm W$ and pull the best scalar fit $g^\star$ inward (shaded: compression zone); right, the mixture bias of Proposition 2 for the two tested noise laws (Table 15), which on a $3\times$ wider support (dashed) collapses $\approx 200\times$ for the symmetric law and leaves a $\approx +0.001$ residual of opposite sign for the one-sided law. The tied encoding and the residual-space and target-space loss shapes are in Figure 4.

4 Why noisy targets are different: boundary saturation and the mixture bias

Section 3 treats the *in-support case*: deterministic targets inside I_δ, where target and prediction pass through the *same* map Γ_σ and the bridge is exact (Figure 1). The experiments occupy the complementary *boundary case*, entered when realized targets reach the compression zone; there the realized updates exceed the design bound of Assumption 3 (Section 6.3). Two things change once the scalar target is random. First, the Bayes categorical target is no longer $\Gamma_\sigma(\mathbb{E}\tilde{a})$ but the mixture $\mathbb{E}[\Gamma_\sigma(\tilde{a})]$. Second, near the boundary the mixture is displaced inward, so the best scalar fit is biased even under symmetric noise. Proposition 2 isolates this bias, Theorem 6 bounds it, and Section 5 measures it.

Proposition 2 (Mixture decomposition)**.** *Let $\tilde{a}$ be a random scalar target conditional on a fixed pair, with $\mathbb{E}|\tilde{a}| < \infty$ and the displayed KL expectations finite, and let $\bar{q} := \mathbb{E}[\Gamma_\sigma(\tilde{a})]$. For every categorical prediction p with positive entries, $\mathbb{E}_{\tilde{a}}\,\mathrm{KL}(\Gamma_\sigma(\tilde{a}) \| p) = \mathbb{E}_{\tilde{a}}\,\mathrm{KL}(\Gamma_\sigma(\tilde{a}) \| \bar{q}) + \mathrm{KL}(\bar{q} \| p)$, so the Bayes categorical target is the mixture $\bar{q}$. With $g^\star \in \arg\min_{g \in I_\delta} \mathrm{KL}(\bar{q} \| \Gamma_\sigma(g))$, $\bar{a} := \mathbb{E}\tilde{a}$, and $\mathrm{excess}_{\mathrm{CE}}(g)$ the excess population CE risk, every $g \in I_\delta$ satisfies*

$$(g - \bar{a})^2 \le 2\kappa_{\mathrm{HL}}\,\mathrm{excess}_{\mathrm{CE}}(g) + 2(g^\star - \bar{a})^2.$$

The second term, $(g^\star - \bar{a})^2$, is the boundary bias: a genuine mixture bias that does not vanish under matched Gaussian smoothing. For *unnormalized* masses it would, since $\mathbb{E}_\xi P_i^\sigma(a+\xi) = P_i^{\sqrt{\sigma^2+\tau^2}}(a)$ exactly; but the normalized target carries the random denominator $Z_\sigma(a+\xi)$, and the truncated projection saturates at the support boundary, renormalizing excluded mass inward (F). The upshot is *target-space winsorization*: symmetric, mean-zero perturbations are rectified by boundary saturation into an inward displacement of the best-fitting scalar. This is a property of *where the clamping happens*, not of the perturbations themselves; Table 15 shows the inward displacement for both laws at the frozen support, and we claim no general sign theorem. The saturation, not the normalization that realizes it, is load-bearing: every finite-support encoding tested here compresses target differences at the boundary, and the denominator-free two-hot variant is indistinguishable in both decisive settings (Section 5.5).

The Huber contrast follows (Figure 1b). Huber clamps *influence in residual space*: targets are untouched, and the location estimand is still the mean under symmetric noise. HL clamps *in target space*, before any fitting, so comparable-curvature losses can respond oppositely to the same contamination; this is the three-loss contrast the experiments measure.

In the known-Gaussian corner, with predictions smoothed at the matched per-context scale, the log loss carries per-context inverse-variance weights intrinsically (Theorem 5, Appendix F.4). This is

a structural statement, not a faster rate, and its in-family hypothesis excludes from scope exactly the renormalization term of Proposition 2.

Theorem 6 (informal; full statement and proof in Appendix F.5). *Write $\bar{m}(s)$ and $V_I(s)$ for the mean and variance of $\mathcal{N}(s, \sigma^2)$ truncated to the support; $\bar{m}$ is a soft clip of s. For $\bar{a} \in I_\delta$, up to a discretization ripple of order ς, the bias $|g^\star - \bar{a}|$ is sandwiched between constant multiples of the Jensen gap $|\bar{m}(\bar{a}) - \mathbb{E}\,\bar{m}(\tilde{a})|$ of this clip, which is at most the expected excursion of the noisy target beyond I_δ plus $(1 - v(c_\delta))\,\mathbb{E}|\tilde{a} - \bar{a}|$.*

Outside that corner the last term has relative size 0.03 at $c_\delta = 3$, so the bias is controlled by how far, and how often, targets cross the boundary: the compression-zone occupancy of Section 5.5 (Table 15). The ripple term is bounded but not shown to be second order (Section 6.3).

5 Experiments: when does the replacement help?

The experiments ask whether the categorical loss buys robustness once targets are noisy, in the boundary regime of Section 4; the answer is a phase diagram (Figure 2; Table 1).

5.1 Platform, protocol, and evidence status

Platform. We use a log-linear contextual bandit (C=20 contexts, K=100 actions, d=32 Gaussian features; Appendix H.2). Each round draws n fresh pairs from the fixed law $\rho \otimes U \otimes U$, independent of the current policy. A linear potential increment is fitted by 60 full-batch Adam steps on identical data at matched loss scale. The three losses are MSE, Huber with threshold $W/2$, its curvature matched to MSE, and the tied HL-Gauss cross-entropy (m=51, β=0.75), written HL; the support is set once from a 32768-pair clean warmup by a rule fixed in advance. This is a fixed-data mechanism platform, not an implementation of Algorithm 1; the on-policy checks are in Appendices H.5.1 and H.5.7. Every training cell exceeds Assumption 3's design bound and leaves I_δ (Table 28), so only Proposition 2's mixture identity and Theorem 6, whose targets may leave I_δ, reach the experiments. The primary outcome is the value of the uniform round mixture of the post-update iterates, which obeys Theorem 2's bound (Appendix B.4), paired within seed; a seed indexes one 15-round data draw. The clean MSE regret is 0.145 (E1 series) and 0.157 (E2/E3 series); the preregistered smallest effect size of interest (SESOI) is 10% of it, and every contrast is reported as a percentage of that regret (Table 1).

Evidence status. Table 1 marks each result as confirmatory (a registration or a registered extension), replication, or descriptive (protocol and interpretive branches fixed before execution). Registration v1 predicted an HL advantage under symmetric contamination and was refuted. Registration v2, filed after that result and after a six-configuration exploration disclosed in Table 6, is the confirmatory experiment; its contrast is about half the exploratory one, the expected winner's-curse direction. Registration history: Appendix H.4.

5.2 Symmetric contamination: a consistent deficit

Design. Setting E1 adds zero-mean mixture noise, $\xi \sim 0.95\,\mathcal{N}(0, \tau^2) + 0.05\,\mathcal{N}(0, (10\tau)^2)$, to each reward endpoint before differencing, with 512 pairs per round, so each fit averages over many pairs: an averaging regime (20 seeds). **Result.** HL is worse than MSE in every seed, by 3.4% of baseline regret, a third of the SESOI (Table 1). Huber matches MSE. On clean rewards HL also trails MSE, by a clean deficit of 0.3% with a different, interior origin (Section 5.5). **Reading.** Differencing two endpoints re-symmetrizes symmetric noise, so MSE is unaffected in the averaging regime. HL pays the boundary bias of Section 4: boundary saturation realizes the $(g^\star - \bar{a})^2$ term of Proposition 2. Huber clamps residuals rather than targets and pays nothing, so bounded influence is not the cause; bounded targets are.

5.3 Structured contamination: a small, replicated, regime-bound benefit

Design. Setting E2 marks each reward endpoint with probability 0.05. A marked endpoint whose feature is aligned with a fixed direction h ($\langle \varphi(c,k), h \rangle > 0$, $\cos(h, \theta^\star) = -0.7$) receives a positive one-sided spike $+|z| \cdot 10\tau$; all others receive base noise $z \cdot \tau$. Training uses 32 pairs per round at $d = 32$, the nominal n=d interpolation threshold, so the fit can interpolate the data: the interpolation regime. The model was registered as a synthetic twin of reward-model length bias (Singhal et al.,

Setting	Contamination	Regime	Pairs / seeds	Status	HL−MSE [95% CI]	HL−Huber	Verdict
E1	symmetric mixture, zero mean	averaging	512 / 20	conf. v1	-3.4% $[-3.7, -3.1]$	-3.4%	deficit
Clean, E1 series	none	averaging	512 / 20	conf. v1	-0.3%	$-0.4\%^{\ddagger}$	clean deficit
E2	one-sided spikes, feature-correlated	interpolation	32 / 40	conf. v2	$+0.7\%^{\dagger}$ $[-0.02, +1.37]$	$+0.5\%$ n.s.	benefit
E2, replication	same, fresh seeds	interpolation	32 / 80	replication	$+0.7\%^{\dagger}$ $[+0.3, +1.2]$	$+0.4\%$ n.s.	benefit
E2, 1200 seeds (factorial)	same	interpolation	32 / 1200	descr.	$+0.8\%$	$+0.7\%$	benefit
E2, doubled rate	rate 0.10, same amplitude	interpolation	32 / 40	descr.	$+1.9\%$	$+1.9\%$	benefit
E2, Adam(600)	same, $10\times$ budget	interpolation	32 / 40	descr.	$+2.5\%$	$+3.5\%$	benefit
E2, L-BFGS	same, high-precision solve	interpolation	32 / 40	descr.	$+0.1\%$ n.s.	$+0.9\%$ n.s.	unresolved
E2, n=64	same	exit interpolation	64 / 40	descr.	-0.1% n.s. $[-0.4, +0.2]$	-1.1%	null; Huber ahead
E3	symmetrized spikes	interpolation	32 / 40	descr.	-1.2% $[-2.3, -0.1]$	-1.3%	deficit; $\approx$0 at 400 seeds
Clean, E2 series	none	interpolation	32 / 40	conf. v2	-0.1%	-0.2%	clean deficit
On-policy MLP, k=1	symmetric mixture, $\mu_t = \pi_t$	averaging	512 / 400	conf. ext.	$-1.7\%^{\dagger}$ $[-1.9, -1.4]$	$-2.6\%^{\dagger}$	deficit
On-policy MLP, k=3	same, support $\times 3$	averaging	512 / 400	conf. ext.	-0.2% n.s.	-0.1%	attenuated
LLM TL;DR, contaminated	injected length bias	–	500 / 12	conf. ext.	$+0.014$ $[-0.009, +0.037]$ incon.	-0.001 equiv.	HL $\equiv$ Huber
LLM TL;DR, clean	none	–	500 / 8	conf. ext.	$+0.023$ $[-0.008, +0.054]$ incon.	–	inconclusive

Table 1: **The phase diagram at a glance.** Paired per-seed contrasts against MSE and against Huber, as a percentage of the clean MSE regret of the same series (0.1452 for the E1 series and the MLP platform, 0.1566 for the E2/E3 series; the SESOI is 10% of it, 0.01452 and 0.01566); brackets: 95% confidence intervals in the same units; n.s.: not significant at 0.05; †one-sided test, as registered; ‡difference of the recorded HL−MSE and Huber−MSE means. Status: conf. = confirmatory (a registration or a registered extension); descr. = descriptive, protocol fixed before execution. LLM rows: clean-reward-model score units, 90% intervals, equivalence margin ± 0.035. Raw values and p-values: Tables 9, 19, 25, and 24.

2024). **Result.** HL beats MSE by 0.7% of clean regret in the registered run and again in the replication on 80 fresh seeds (Table 1). Both runs meet the registered sign criterion exactly at its floor, so per-seed variability, not the mean, is the fragile margin (Appendix H.5.1). The benefit is 7% of the SESOI: it establishes a sign and, with Sections 5.4–5.5, its boundary mechanism, not a practically material improvement. *Huber.* In the registered runs HL is not separated from curvature-matched Huber (Table 1); the separation appears only in descriptive rows, at 1200 seeds, at the doubled contamination rate, and under the longer Adam budgets. *Interpolation regime only.* At 64 pairs per round the benefit is absent and Huber is ahead (Table 18).

5.4 How the benefit varies: structure, amplitude, regime, budget

Table 2 varies the E2 benefit along four axes on the same fixed data. The sign of the contamination effects lies in the conjunction of one-sidedness and feature correlation, not in either alone; the benefit belongs to the finite-step Adam path, whereas the E1 deficit is a property of the fitted objective.

5.5 Where the loss clamps: boundary winsorization versus bin-curvature ripple

Anatomy. Two target-space variants join the three losses in E2 and E3: **two-hot**, a mean-preserving two-point target with no normalization denominator, and **hard winsorization** of the MSE target at $\pm(W - 3\sigma)$. Two-hot is indistinguishable from HL in both; hard winsorization captures most of the E2 benefit but pays a larger E3 deficit (Table 17). Both effects are carried by target-space boundedness; the clip geometry matters. **Support widening** attenuates every contamination effect to the level of the clean deficit, in both regimes, whether all scales or only the boundary move (Figure 2c; Appendix H.5.3). The clean deficit itself stays: every contamination effect here is narrow-support winsorization.

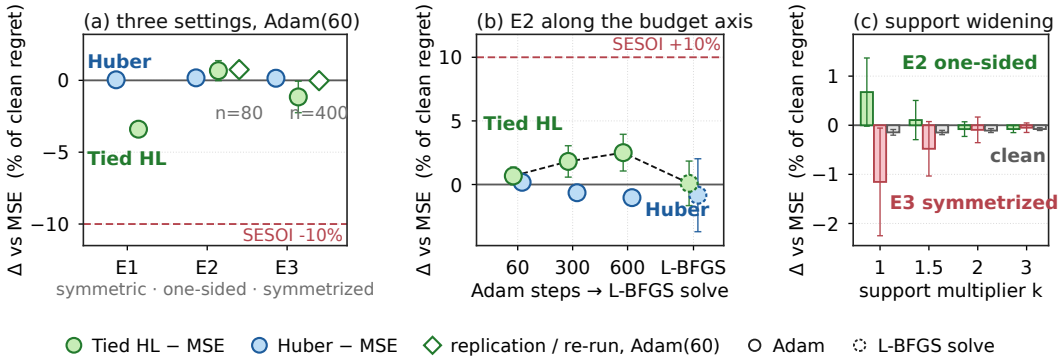


Figure 2: **The three-loss phase diagram** (means; whiskers 95% CIs; vertical axes in percent of the series' clean MSE regret, $1\% = 1.45 \times 10^{-3}$ for E1 and 1.57×10^{-3} for E2/E3; dashed red: the preregistered SESOI, $\pm 10\%$). (a) HL−MSE and Huber−MSE at the frozen Adam(60) budget in the three settings of Table 1; diamonds: the 80-seed replication (E2) and the 400-seed re-run (E3). (b) E2 as the optimization budget varies from Adam(60) to per-round L-BFGS solves (dashed edges). (c) Proportional support widening attenuates both contamination effects to the level of the clean deficit by $k = 2$.

Axis	Intervention	Reading
Structure	spikes symmetrized at matched nominal rate (E3); moment-matched 2×2 factorial of one-sidedness × feature correlation	E3 removes the benefit; both factorial main effects and their interaction are significant; the one-sided model is separated by the feature-conditional mean $\mathbb{E}[\xi \mid \varphi]$, not $\mathbb{E}[\xi]$
Amplitude	spike amplitude doubled; contamination rate doubled at fixed amplitude	the symmetric, feature-independent cell turns positive at the doubled amplitude while its rate ladder is flat; the one-sided effect triples at the doubled rate
Regime	pairs per round $n \in \{16, \ldots, 64\}$ at d=32; ℓ_2 penalties	positive for every $n \leq 36$, small at 48, absent at 64; the tested ℓ_2 penalties do not reproduce it
Budget	Adam 60 → 300/600 steps; per-round float64 L-BFGS solves	the E1 deficit is stable under high-precision solves and matches the population mixture bias in sign and order; the clean deficit decays with precision; the benefit grows with Adam steps and is unresolved at the high-precision solves

Table 2: **Moderators of the E2 benefit** (descriptive controls on the fixed data of Section 5.1). Records: Appendix H.5.2 and Figure 9 (structure); Figure 8 (amplitude); Table 18 (regime); Tables 15 and 19 (budget).

Smoothing. Clean targets essentially never reach the compression zone (0.099% beyond $W - 3\sigma$), and doubling the smoothing scale removes the deficit while leaving the contaminated benefit unchanged (Table 14): the clean deficit is bin-curvature ripple, not boundary winsorization. **Occupancy.** Compression-zone occupancy, the fraction of realized targets beyond $W - 3\sigma$, organizes the whole record: across all 19 settings the effect's magnitude tracks occupancy while structure and amplitude set the sign (Figure 3), and we claim no deterministic sign rule.

5.6 Beyond the linear platform

MLP pilots reproduce the symmetric-model deficit at every width but not the linear n=d threshold (Table 22). **On-policy MLP test of the prediction.** A preregistered extension closes Algorithm 1's loop ($\mu_t = \pi_t$) with a persistent-trunk MLP under the symmetric law (512 pairs, 400 seeds). Its three registered predictions hold under Holm (Table 1): the deficit of Corollary F.2, its attenuation at k=3, and HL below residual-space Huber, which shows no deficit against MSE. Appending outer bins at fixed Huber threshold reproduces the attenuation (Appendix H.5.7). **At LLM scale** we run REBEL with LoRA (r=1024, 150 iterations) on the TL;DR summarization stack of Huang et al. (2024): a Pythia-1B policy and training reward model, a held-out Pythia-2.8B evaluator, and 500 prompts (Biderman et al., 2023; Stiennon et al., 2020). Contamination is injected length bias (12 contaminated and 8 clean seeds). The three losses' pooled means differ by under 0.02: HL and Huber are equivalent within the prespecified ± 0.035, while HL−MSE and the clean-reward contrast are inconclusive at the achieved precision (Table 1).

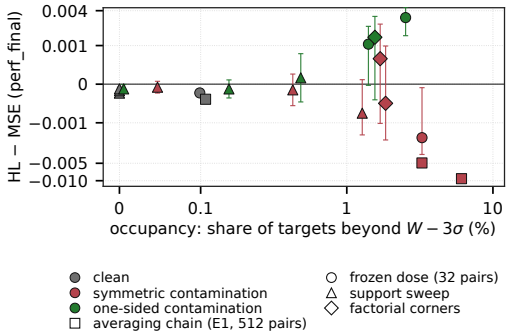


Figure 3: **Occupancy versus effect across the full record** (n=19 cells). x: share of realized targets beyond the cell's interior bound $W - 3\sigma$ (symlog); y: paired HL−MSE with 95% CI (symlog). |effect| tracks occupancy (Spearman 0.833; +0.943 within one-sided cells, −0.786 within symmetric cells): an association within one platform, not a prediction rule.

6 Discussion

6.1 Practitioner's takeaway

The axis is *where the loss clamps*, not "categorical versus squared." Three rules follow. (1) *Widen the support first.* On adequately wide support the replacement is inert: the three losses differ by no more than the clean deficit. (2) *When the reward scale is unknown or drifts, use curvature-matched Huber.* Widening presupposes that W covers unseen contamination; otherwise narrow support is the operative regime. There Huber needs no support calibration and is never significantly below MSE in the tested conditions. It trails HL by at most 0.7–3.5% of clean regret where HL wins, leads it by 2.6–3.4% where HL pays (Table 1), and is equivalent to HL at LLM scale only within the tested margin. (3) *Monitor compression-zone occupancy.* On real data the calibration recipe leaves the zone empty, and a factor-two miscalibration exposes only 2.1–2.6% of targets, so activation needs calibration error or drift rather than everyday tails (Appendix H.5.6).

6.2 Mechanism and open problem

Both contamination effects are boundary phenomena: widening attenuates them and a two-hot encoding reproduces them, so they belong to the bounded encoding, not to the categorical form; the clean deficit is bin-curvature ripple. The optimizer axis separates the two contamination effects: the averaging-regime deficit is stable under high-precision solves, the benefit optimizer-path dependent. The benefit remains to be explained; an influence-cap conjecture did not survive a closed-form test (Appendix H.5.5). The bias of Theorem 6 fixes the statistical target; the policy effect adds the finite-sample fit, and the two compete: at equal occupancy, doubling the rate deepens the deficit and doubling the amplitude reverses it (Appendix H.5.7).

6.3 Limitations

Natural reward-model noise, preregistered policy training, and a calibrated held-out evaluator have not been combined in one experiment. An earlier LLM-scale study whose evaluators failed pre-specified calibration checks carries no confirmatory evidence (Appendix H.5.6). Confirmatory evidence is synthetic and on one environment instance, with contamination families built to separate mechanisms, not to mirror natural reward noise. The phase diagram is linear and from fixed behavior data (Appendix H.5.1); the on-policy MLP test is nonlinear and closed-loop but on one noise family (Appendix H.5.7). Only Proposition 2 and Theorem 6 reach the training cells (Section 5.1); Corollary F.2 is instead tested as a prediction, and holds (Appendix H.5.7). Theorem 6's ripple term and the monotonicity of V_I in $|s|$ remain the most direct open problems.

REPRODUCIBILITY STATEMENT

Definitions. One training round and its readout can be reconstructed from the text: setup and assumptions (Section 2), complete proofs (Appendices A–G), the synthetic platform (Appendix H.2), the LLM configuration (Appendix H.5.6), and the statistical conventions of every contrast (Appendices H.4 and H.5).

Reproduction. The supplementary repository carries the frozen specifications, per-seed and per-prompt raw arrays, and the scripts that regenerate every table and figure, with run instructions at three levels: regenerating the tables from the released records, re-running the core synthetic settings on a few seeds, and full re-execution.

Traceability. Every run executed under a specification fixed in advance and bound by SHA-256 across specification, environment lock, run manifest, and approval; confirmatory runs were zero-edit executions of protocols frozen before execution (registry: Appendix H.4). Every reported statistic was recomputed from raw records by an implementation independent of the executing one; the digests, executor identities, audit inventory, registration timeline, and a seed-0 deterministic checkpoint are in the process record released with the repository (summary: Appendix H.8).

AI USE STATEMENT

We used generative AI for the following tasks that require disclosure. *Formulating mathematical claims and developing the theoretical framework*: AI assistance contributed to statements and proof routes; the authors verified every statement and proof (Appendices A–F). *Proposing hypotheses* and *designing experiments*: the hypotheses, decision rules, and pre-worded outcome branches of the post-closure extensions (Appendix H.4) were drafted by AI agents and frozen by the authors before execution; registrations v1 and v2 were authored by the authors. *Implementing methods*: the experiment harness and the analysis, figure, and audit scripts, run as zero-edit executions of hash-pinned scripts. *Assisting in the writing of proofs*: proof drafting, with full proofs in Appendices A–F. *Interpreting results*: recomputation of every reported statistic from raw records by an AI auditor independent of the executing agent, and mechanical re-derivation of the pre-worded interpretive branches. Generative AI was not used for translation; generating, cleaning, or reformatting datasets and qualitative or thematic analysis are not applicable to this work. We also used generative AI to draft and edit the manuscript from a researcher-curated dossier of fixed records and to format references. We have reviewed all AI-assisted work: the authors checked every proof, every reported statistic was recomputed independently, and generative AI was not used to fabricate, select, or filter results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Herm Jan Brascamp and Elliott H. Lieb. On extensions of the Brunn–Minkowski and Prékopa–Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22:366–389, 1976.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust DPO: Aligning language models with noisy feedback. In *International Conference on Machine Learning (ICML)*, 2024.

Jesse Farebrother, Jordi Orbay, Quan Vuong, Adrien Ali Taïga, Yevgen Chebotar, Ted Xiao, Alex Irpan, Sergey Levine, Pablo Samuel Castro, Aleksandra Faust, Aviral Kumar, and Rishabh Agarwal. Stop regressing: Training value functions via classification for scalable deep RL. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Dylan J. Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Zhaolin Gao, Jonathan D. Chang, Wenhao Zhan, Owen Oertell, Gokul Swamy, Kianté Brantley, Thorsten Joachims, J. Andrew Bagnell, Jason D. Lee, and Wen Sun. REBEL: Reinforcement learning via regressing relative rewards. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2404.16767.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949–986, 2022.

Shengyi Huang, Michael Noukhovitch, Arian Hosseini, Kashif Rasul, Weixun Wang, and Lewis Tunstall. The N+ implementation details of RLHF with PPO: A case study on TL;DR summarization. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2403.17031.

Peter J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1): 73–101, 1964.

Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. In *International Conference on Learning Representations (ICLR)*, 2021.

Ehsan Imani and Martha White. Improving regression performance with distributional losses. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Sham Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2001.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Clare Lyle, Marc G. Bellemare, and Pablo Samuel Castro. A comparative analysis of expected and distributional reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2019.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

Arkadi Nemirovski and David Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

Tiberiu Popoviciu. Sur les équations algébriques ayant toutes leurs racines réelles. *Mathematica (Cluj)*, 9:129–145, 1935.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Mark Rowland, Marc G. Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Sharan Sahu and Martin T. Wells. Online distributionally robust LLM alignment via regression to relative reward, 2025. arXiv:2509.19104.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588: 604–609, 2020.

John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. arXiv:1707.06347.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in RLHF. In *Conference on Language Modeling (COLM)*, 2024.

Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

Lawrence Stewart, Francis Bach, Quentin Berthet, and Jean-Philippe Vert. Regression as classification: Influence of task formulation on neural network features. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Kaiwen Wang, Kevin Zhou, Runzhe Wu, Nathan Kallus, and Wen Sun. The benefits of being distributional: Small-loss bounds for reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Kaiwen Wang, Owen Oertell, Alekh Agarwal, Nathan Kallus, and Wen Sun. More benefits of being distributional: Second-order bounds for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2024.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

Jin Peng Zhou, Kaiwen Wang, Jonathan Chang, Zhaolin Gao, Nathan Kallus, Kilian Q. Weinberger, Kianté Brantley, and Wen Sun. $Q^\sharp$: Provably optimal distributional RL for LLM post-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2502.20548.

APPENDIX ROADMAP

Question	Where	Key statements, tables, figures
Why is the tied HL-Gauss loss locally quadratic, with which constants?	Appendix A	Lemmas A.1–A.2, Corollary A.3, Proposition A.5
How does the categorical risk enter REBEL's bound?	Appendices B, G	Lemmas B.1–B.2; Table 3
Why can the square-root exponent not be improved in this format?	Appendix C	Theorem 3, Corollary C.1, Remark C.3
When are first-order rates available?	Appendix D	Theorem 4, Theorem D.2, Propositions D.3–D.4
What is the local Fisher and GGN geometry of the scalar loss?	Appendix E	Propositions E.1–E.2
What happens to noisy targets at the boundary?	Appendix F	Proposition 2, Theorems 5 and 6, Corollaries F.1–F.2
How was the synthetic platform built, calibrated, and run?	Appendix H.2	Algorithm 2; Table 5; the seed-0 checkpoint is in the released process record
How does the calibration constant behave at other corners, and why not a Pinsker route?	Appendix H.3	–
What was preregistered, when, and with what outcome?	Appendix H.4	Table 6; the event-by-event timeline is in the released process record
What is the evidence status of each claim?	Appendix H.1	Table 4
What do the controls and ablations show, by question?	Appendix H.5	Tables 9, 15, 17, 24; Figure 3
Do the theorems hold numerically on the pipeline?	Appendix H.6	Tables 28–29; Figures 5–6
Does the boundary-bias prediction hold under $\mu_t = \pi_t$ with a nonlinear potential?	Appendix H.5.7	Tables 25–27
Where are the per-seed, dose, and mechanism figures?	Appendix H.7	Figures 7–9
How is every number traced to a pinned artifact?	Appendix H.8	the released process record (digests, audit inventory, seed index sets)

Dependencies among the theory statements: Lemma A.1 $\to$ Lemma A.2 $\to$ Corollary A.3 $\to$ Theorem 1 $\to$ Corollary 1 $\to$ Theorem 2; Theorem 3 and Corollary C.1 are constructions satisfying Theorem 2's hypotheses; Theorem 4 and Theorem D.2 add Lemma D.1 to the chain; Proposition 2 $\to$ Theorem 6 $\to$ Corollary F.2, with Theorem 5 the in-family special case.

A HL-GAUSS GEOMETRY AND CALIBRATION PROOFS

Throughout, $Y_s \sim \mathcal{N}(s, \sigma^2)$ denotes the untruncated Gaussian at location s, $\widetilde{Y}_s$ denotes Y_s conditioned on $Y_s \in I$, $-W = e_0 < e_1 < \cdots < e_m = W$ are the bin edges, and $B(y) \in \{1, \ldots, m\}$ is the index of the bin containing y, so that $P_i(s) = \Pr(Y_s \in [e_{i-1}, e_i])$ and $p_i(s) = \Pr(B(\widetilde{Y}_s) = i)$. For each bin, define

$$m_i(s) := \mathbb{E}[Y_s \mid Y_s \in [e_{i-1}, e_i]], \qquad V_i(s) := \mathrm{Var}(Y_s \mid Y_s \in [e_{i-1}, e_i]),$$

and define

$$\bar{m}(s) := \mathbb{E}[\widetilde{Y}_s], \qquad V_I(s) := \mathrm{Var}(\widetilde{Y}_s).$$

The truncated family has density proportional to

$$\exp\left(\frac{sy}{\sigma^2}\right) \exp\left(-\frac{y^2}{2\sigma^2}\right) \mathbf{1}\{y \in I\},$$

so it is a one-parameter exponential family with natural parameter s/σ^2. The same statement holds after conditioning on a fixed bin.

A.1 SCORE AND CURVATURE

Lemma A.1 (Score and curvature identities)**.** *For every i and finite s,*

$$\partial_s \log p_i(s) = \frac{m_i(s) - \bar{m}(s)}{\sigma^2}.$$

Consequently, for every $q \in \Delta^m$,

$$\ell'_q(g) = \frac{\bar{m}(g) - \sum_i q_i m_i(g)}{\sigma^2},$$

$$\ell''_q(g) = \frac{V_I(g) - \sum_i q_i V_i(g)}{\sigma^4}.$$

For the in-family target $q = \Gamma_\sigma(s)$, the scalar Fisher information is

$$\mathcal{I}_{\mathrm{HL}}(s) = \sum_i p_i(s)\big(\partial_s \log p_i(s)\big)^2 = \frac{V_I(s) - \sum_i p_i(s) V_i(s)}{\sigma^4}.$$

Proof. Differentiation under the integral gives

$$\partial_s P_i(s) = P_i(s)\frac{m_i(s) - s}{\sigma^2}, \qquad \partial_s Z(s) = Z(s)\frac{\bar{m}(s) - s}{\sigma^2}.$$

Subtracting logarithmic derivatives proves the score identity. Exponential-family calculus gives

$$\frac{d}{ds} m_i(s) = \frac{V_i(s)}{\sigma^2}, \qquad \frac{d}{ds}\bar{m}(s) = \frac{V_I(s)}{\sigma^2}.$$

Differentiating ℓ'_q gives the curvature identity. For $q = p(s)$, the score has mean zero, and the law of total variance gives

$$V_I(s) = \sum_i p_i(s) V_i(s) + \mathrm{Var}_{i\sim p(s)}(m_i(s)),$$

which proves the Fisher identity. □

A.2 Variance and smoothness bounds

The margin function of Section 2 is, for $c_\delta > 0$,

$$v(c_\delta) := 1 - 2\big(c_\delta \varphi(c_\delta) + \Phi(-c_\delta)\big) - \Big(\frac{2\varphi(c_\delta)}{1 - 2\Phi(-c_\delta)}\Big)^2,$$

so that $v(3) = 0.970630$; it is strictly increasing on $(0, \infty)$ (Appendix F.5).

Lemma A.2 (Variance bounds)**.** *For every bin and every finite s,*

$$0 \le V_i(s) \le \min\left\{\frac{\varsigma^2}{4}, \sigma^2\right\}.$$

For the full truncated interval,

$$0 \le V_I(s) \le \sigma^2 \qquad \textit{for every } s \in \mathbb{R}.$$

If $s \in I_\delta$, then

$$V_I(s) \ge \sigma^2 v(c_\delta).$$

Proof. Popoviciu's inequality gives $V_i(s) \le \varsigma^2/4$. The Gaussian density restricted to a convex interval remains $1/\sigma^2$-strongly log-concave, so the one-dimensional Brascamp–Lieb variance inequality (Brascamp & Lieb, 1976) gives $V_i(s) \le \sigma^2$ and $V_I(s) \le \sigma^2$.

For the lower bound, write $\widehat{Y} = \widetilde{Y}_s - s$. Since s is at least $c_\delta\sigma$ from each boundary,

$$\mathbb{E}[\widehat{Y}^2] \ge \sigma^2\left[1 - 2\big(c_\delta\varphi(c_\delta) + \Phi(-c_\delta)\big)\right].$$

Moreover,

$$|\mathbb{E}\widehat{Y}| \le \frac{2\sigma\varphi(c_\delta)}{1 - 2\Phi(-c_\delta)}.$$

Subtracting the squared mean yields $V_I(s) \ge \sigma^2 v(c_\delta)$. □

Corollary A.3 (Strong convexity and global smoothness)**.** *For every $q \in \Delta^m$ and $g \in I_\delta$,*

$$\ell_q''(g) \geq \frac{\sigma^2 v(c_\delta) - \varsigma^2/4}{\sigma^4} = \alpha_{\mathrm{HL}}.$$

For every $g \in \mathbb{R}$,

$$|\ell_q''(g)| \leq \frac{1}{\sigma^2}.$$

Proof. The interior lower bound follows from Lemmas A.1 and A.2. Globally, both $V_I(g)$ and $\sum_i q_i V_i(g)$ lie in $[0, \sigma^2]$, so their difference has absolute value at most σ^2. □

A.3 Injectivity

Let $z_i = (e_{i-1} + e_i)/2$ and define

$$M_\sigma(s) := \sum_{i=1}^{m} z_i p_i(s).$$

Lemma A.4 (Global injectivity)**.** *For every finite s,*

$$M_\sigma'(s) = \frac{1}{\sigma^2} \operatorname{Cov}(z_{B(\widetilde{Y}_s)}, \widetilde{Y}_s) > 0.$$

Consequently, $s \mapsto \Gamma_\sigma(s)$ is injective on $\mathbb{R}$.

Proof. Lemma A.1 gives

$$M_\sigma'(s) = \frac{1}{\sigma^2} \sum_i z_i p_i(s)(m_i(s) - \bar{m}(s)) = \frac{1}{\sigma^2} \operatorname{Cov}(z_B, \widetilde{Y}_s).$$

For independent copies Y, Y' of $\widetilde{Y}_s$,

$$2 \operatorname{Cov}(z_B(Y), Y) = \mathbb{E}[(z_B(Y) - z_B(Y'))(Y - Y')].$$

The integrand is nonnegative because the bin-center map is nondecreasing, and it is strictly positive on an event of positive probability because $m \geq 2$ and the truncated Gaussian has positive density throughout I. Thus M_σ is strictly increasing. Equality of two categorical vectors implies equality of their decoded center means, hence equality of the scalar parameters. □

A.4 Proofs of Proposition 1 and Theorem 1

Proof of Proposition 1. Zero KL implies equality of the two positive categorical vectors, and Lemma A.4 gives $g_t = a_t$ for $\mathcal{D}_t$-almost every sampled pair. Because $\mathcal{Y}$ is countable and π_t is positive on $\mathcal{S}_x$, every pair (y, y') with $y \in \mathcal{S}_x$ and $y' \in \operatorname{supp} \mu_t(\cdot \mid x)$ is an atom of the conditional product support. Fixing one $y_0' \in \operatorname{supp} \mu_t(\cdot \mid x)$ gives

$$f_t(x, y) - r(x, y) = f_t(x, y_0') - r(x, y_0') =: c_t(x)$$

for all $y \in \mathcal{S}_x$. Therefore

$$\pi_{t+1}(y \mid x) = \pi_t(y \mid x) e^{\eta(r(x,y) + c_t(x))}.$$

Assumption 1 ensures that this identity covers the full support of π_{t+1}. Normalizing over $\mathcal{S}_x$ yields the stated mirror-descent update. □

Proof of Theorem 1. Fix u and set $q = \Gamma_\sigma(u)$. Then

$$\mathrm{KL}(q \| \Gamma_\sigma(v)) = \ell_q(v) - \ell_q(u), \qquad \ell_q'(u) = 0.$$

Taylor's formula along the segment from u to v gives

$$\ell_q(v) - \ell_q(u) = (v - u)^2 \int_0^1 (1 - s) \ell_q''(u + s(v - u))\, ds.$$

The global upper curvature bound $\ell_q'' \le 1/\sigma^2$ yields

$$\mathrm{KL}(\Gamma_\sigma(u)\|\Gamma_\sigma(v)) \le \frac{(u-v)^2}{2\sigma^2}.$$

If $u, v \in I_\delta$, the segment lies in I_δ and Corollary A.3 gives

$$\mathrm{KL}(\Gamma_\sigma(u)\|\Gamma_\sigma(v)) \ge \frac{\alpha_{\mathrm{HL}}}{2}(u-v)^2.$$

This is equivalent to the lower calibration inequality. The strong-convexity and global gradient-Lipschitz statements follow directly from Corollary A.3. □

Proof of Corollary 1. Assumption 3 and the support design imply $|g_t| \le 2A_{\mathrm{des}} \le W - \delta$, while $|a_t| \le 2R_{\max} \le W - \delta$. Apply the lower calibration inequality pointwise and take expectation under $\mathcal{D}_t$. □

A.5 No global CE-to-MSE calibration

Proposition A.5 (Boundedness is necessary)**.** *There is no finite κ such that*

$$(u-v)^2 \le \kappa\,\mathrm{KL}(\Gamma_\sigma(u)\|\Gamma_\sigma(v))$$

for all $u, v \in \mathbb{R}$. More strongly, for every $\varepsilon > 0$ and $L > 0$, there is a one-context two-action policy-realizable pair distribution with

$$\mathbb{E}[\mathrm{KL}] \le \varepsilon, \qquad \mathbb{E}[(g-a)^2] \ge L.$$

Proof. Lemma A.1 implies

$$|\ell_q'(g)| \le \frac{2W}{\sigma^2},$$

because both means in the derivative formula lie in $[-W, W]$. Hence

$$\mathrm{KL}(\Gamma_\sigma(u)\|\Gamma_\sigma(v)) \le \frac{2W}{\sigma^2}|u-v|,$$

which rules out a global quadratic lower bound.

For an explicit policy construction, take one context and two actions y_0, y_1. Let $\mu = \delta_{y_0}$, let $\pi_t(y_1) = w \in (0,1)$, and set $r \equiv 0$. Choose an unnormalized update potential $h(y_0) = 0$, $h(y_1) = \Delta$, define $\pi_{t+1} \propto \pi_t e^{\eta h}$, and let $f = h - \eta^{-1}\log \mathbb{E}_{\pi_t} e^{\eta h}$. Then $g = \Delta$ on the event $(y, y') = (y_1, y_0)$ and $g = 0$ otherwise, while $a = 0$ always. Therefore

$$\mathbb{E}[\mathrm{KL}] \le w\frac{2W}{\sigma^2}\Delta, \qquad \mathbb{E}[(g-a)^2] = w\Delta^2.$$

Set $\Delta = \varepsilon\sigma^2/(2Ww)$. The KL risk is at most ε, whereas the MSE equals $\varepsilon^2\sigma^4/(4W^2w)$ and exceeds any prescribed L for sufficiently small w. □

B Proof of the main performance theorem

Define the pointwise scalar residual

$$e_t(x,y) := f_t(x,y) - r(x,y),$$

and its conditional means

$$m_{\pi,t}(x) := \mathbb{E}_{y\sim\pi_t(\cdot|x)} e_t(x,y), \qquad m_{\mu,t}(x) := \mathbb{E}_{y\sim\mu_t(\cdot|x)} e_t(x,y).$$

B.1 EXACT PAIRWISE VARIANCE DECOMPOSITION

The identity below is the expansion carried out in the proof of REBEL's Lemma 1 (Gao et al., 2024, Appendix D); we state it separately because the transfer uses it jointly rather than term by term.

Lemma B.1 (Exact decomposition)**.** *For every t,*

$$\mathbb{E}_{\mathcal{D}_t}(g_t - a_t)^2 = \mathbb{E}_{x\sim\rho}\left[\mathrm{Var}_{\pi_t}(e_t) + \mathrm{Var}_{\mu_t}(e_t) + (m_{\pi,t} - m_{\mu,t})^2\right].$$

Consequently,

$$\mathbb{E}_{\rho,\mu_t}(e_t - m_{\pi,t})^2 \le \mathbb{E}_{\mathcal{D}_t}(g_t - a_t)^2,$$

and also

$$\mathbb{E}_{\rho,\pi_t}(e_t - m_{\pi,t})^2 \le \mathbb{E}_{\mathcal{D}_t}(g_t - a_t)^2.$$

Proof. Condition on x and let $Y \sim \pi_t(\cdot \mid x)$ and $Y' \sim \mu_t(\cdot \mid x)$ be independent. Since

$$g_t - a_t = e_t(Y) - e_t(Y'),$$

expanding the square gives

$$\mathbb{E}(e_t(Y) - e_t(Y'))^2 = \mathrm{Var}_{\pi_t}(e_t) + \mathrm{Var}_{\mu_t}(e_t) + (m_{\pi,t} - m_{\mu,t})^2.$$

Moreover,

$$\mathbb{E}_{\mu_t}(e_t - m_{\pi,t})^2 = \mathrm{Var}_{\mu_t}(e_t) + (m_{\mu,t} - m_{\pi,t})^2,$$

and

$$\mathbb{E}_{\pi_t}(e_t - m_{\pi,t})^2 = \mathrm{Var}_{\pi_t}(e_t).$$

Each is a subsum of the exact decomposition. □

B.2 MIRROR DESCENT VIA HOEFFDING'S LEMMA

Lemma B.2 (Support-valid KL telescope)**.** *Under Assumptions 1 and 3, for every comparator with finite $D_0(\pi^\star)$,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{\rho,\pi^\star}\widetilde{A}_t \le \frac{D_0(\pi^\star)}{\eta T} + \frac{\eta}{2T}\sum_{t=0}^{T-1}A_t^2.$$

Proof. On the common support,

$$\pi_{t+1}(y \mid x) = \pi_t(y \mid x)e^{\eta f_t(x,y)}.$$

Since f_t and $\widetilde{A}_t$ differ by an x-dependent constant,

$$\pi_{t+1}(y \mid x) = \frac{\pi_t(y \mid x)e^{\eta\widetilde{A}_t(x,y)}}{Z_t(x)}, \qquad Z_t(x) := \mathbb{E}_{y\sim\pi_t(\cdot|x)}e^{\eta\widetilde{A}_t(x,y)}.$$

For fixed x,

$$\mathrm{KL}(\pi^\star\|\pi_{t+1}) - \mathrm{KL}(\pi^\star\|\pi_t) = -\eta\mathbb{E}_{\pi^\star}\widetilde{A}_t + \log Z_t(x).$$

By construction, $\mathbb{E}_{\pi_t}\widetilde{A}_t = 0$. Since $\widetilde{A}_t \in [-A_t, A_t]$, Hoeffding's lemma gives

$$\log Z_t(x) \le \frac{\eta^2 A_t^2}{2}.$$

Average over x, sum over t, telescope the KL terms, and use nonnegativity of the terminal KL. □

B.3 Proof of Theorem 2

Proof of Theorem 2. Define the true reward advantage

$$A_t^r(x,y) := r(x,y) - \mathbb{E}_{z\sim\pi_t(\cdot|x)} r(x,z).$$

Since $f_t = r + e_t$,

$$\widetilde{A}_t = A_t^r + e_t - m_{\pi,t}, \qquad A_t^r - \widetilde{A}_t = m_{\pi,t} - e_t.$$

Corollary 1 and Lemma B.1 give

$$\mathbb{E}_{\rho,\mu_t}(m_{\pi,t} - e_t)^2 \le \kappa_{\mathrm{HL}} \varepsilon_t^{\mathrm{CE}}.$$

By change of measure and Cauchy–Schwarz,

$$\begin{aligned}\mathbb{E}_{\rho,\pi^\star}(A_t^r - \widetilde{A}_t) &\le \left(\mathbb{E}_{\rho,\mu_t}\left[\frac{\pi^\star}{\mu_t}(m_{\pi,t} - e_t)^2\right]\right)^{1/2} \\ &\le \sqrt{C_t(\pi^\star)\kappa_{\mathrm{HL}}\varepsilon_t^{\mathrm{CE}}}.\end{aligned}$$

The contextual performance-difference identity is

$$J(\pi^\star) - J(\pi_t) = \mathbb{E}_{\rho,\pi^\star} A_t^r.$$

Add and subtract $\widetilde{A}_t$, apply the preceding transfer bound and Lemma B.2, and average over t. Finally,

$$J(\bar{\pi}_T) = \frac{1}{T}\sum_{t=0}^{T-1} J(\pi_t)$$

by linearity of J in the policy. This proves Theorem 2. □

Proof of Corollary 2. Substitute the uniform bounds into Theorem 2. The optimized expression follows by minimizing

$$\frac{D_0(\pi^\star)}{\eta T} + \frac{\eta A^2}{2}$$

over $\eta > 0$, under the stated uniformity condition. □

B.4 Uniform constants

Corollary 2 (Uniform constants)**.** *If $A_t \le A$, $C_t(\pi^\star) \le C$, and $\varepsilon_t^{\mathrm{CE}} \le \varepsilon_{\mathrm{CE}}$ for every t, and the same constants remain valid over the candidate step sizes, then $\eta = \sqrt{2D_0(\pi^\star)/(A^2T)}$ gives $J(\pi^\star) - J(\bar{\pi}_T) \le A\sqrt{2D_0(\pi^\star)/T} + \sqrt{C\,\kappa_{\mathrm{HL}}\,\varepsilon_{\mathrm{CE}}}$.*

The choice balances the first two terms of Theorem 2; the third is η-free.

Remark (post-update mixture). The experiments report the uniform mixture of the post-update iterates, $\bar{\pi}_T^+ := \frac{1}{T}\sum_{t=1}^{T}\pi_t$, whereas Theorem 2 averages $\pi_0, \dots, \pi_{T-1}$. The two are related exactly by $J(\bar{\pi}_T^+) = J(\bar{\pi}_T) + \big(J(\pi_T) - J(\pi_0)\big)/T$, and for any comparator with $J(\pi^\star) \ge J(\pi_0)$, applying Theorem 2 over $T+1$ rounds and dropping the nonnegative $t=0$ term gives $J(\pi^\star) - J(\bar{\pi}_T^+) \le \frac{T+1}{T}\big[J(\pi^\star) - J(\bar{\pi}_{T+1})\big]$: the bound of Theorem 2 at horizon $T+1$, up to the factor $(T+1)/T$. Both comparators of the record, the mirror-descent oracle and the per-context maximizer, dominate the uniform π_0.

C Lower bounds

Theorem 3 (Sharpness of the regression-error exponent)**.** *Fix an integer $C \ge 2$, admissible HL-Gauss parameters, and $\varepsilon > 0$. There is a one-context, C-action instance with deterministic reward, a full-support tabular softmax policy class, a uniform base distribution, and a deterministic comparator $\pi^\star$ such that, for every t, $C_t(\pi^\star) = C$, $\mathbb{E}_{\mathcal{D}_t}\mathrm{KL}(\Gamma_\sigma(a_t)\|\Gamma_\sigma(g_t)) \le \varepsilon$, and $A_t \le \sigma\sqrt{C\varepsilon}$, while every iterate satisfies $J(\pi^\star) - J(\pi_t) \ge \frac{\sigma}{4}\sqrt{C\varepsilon}$. For any fixed $R_{\max} > 0$, the construction is valid for all sufficiently small ε.*

C.1 Proof of Theorem 3

Proof of Theorem 3. Let $\mathcal{X} = \{x_0\}$ and let $\mathcal{Y}$ contain $C-1$ baseline actions and one distinguished action y_1. Put $w = 1/C$. Let all baseline actions have reward c_0 and let

$$r(y_1) = c_0 + \frac{\Delta}{2}, \qquad \Delta := \sigma\sqrt{C\varepsilon}.$$

Take $\mu_t = \pi_0$ to be uniform and $\pi^\star = \delta_{y_1}$. Use a full-support tabular softmax class. For any fixed $R_{\max}$, choose ε sufficiently small that the rewards can be placed in $[-R_{\max}, R_{\max}]$; a common offset c_0 does not change pairwise targets or regret.

At every round define the unnormalized update potential

$$h_t(y) := r(y) - \Delta \mathbf{1}\{y = y_1\},$$

then define

$$Z_t := \mathbb{E}_{z\sim\pi_t} e^{\eta h_t(z)}, \qquad \pi_{t+1}(y) := \frac{\pi_t(y) e^{\eta h_t(y)}}{Z_t},$$

and

$$f_t(y) := h_t(y) - \frac{1}{\eta}\log Z_t.$$

Thus f_t is exactly $\eta^{-1}\log(\pi_{t+1}/\pi_t)$, while

$$g_t(y, y') = h_t(y) - h_t(y').$$

The effective update value of y_1 is $c_0 - \Delta/2$, below the baseline value c_0, so $\pi_t(y_1)$ is nonincreasing and $\pi_t(y_1) \le w$. The scalar error is

$$g_t - a_t = -\Delta\big(\mathbf{1}\{y = y_1\} - \mathbf{1}\{y' = y_1\}\big).$$

It is nonzero only when exactly one endpoint is y_1, an event of probability at most

$$\pi_t(y_1) + \mu_t(y_1) \le 2w.$$

By the global upper half of Theorem 1, every such pair contributes at most $\Delta^2/(2\sigma^2)$ KL. Hence

$$\mathbb{E}_{\mathcal{D}_t} \mathrm{KL}(\Gamma_\sigma(a_t) \| \Gamma_\sigma(g_t)) \le 2w\frac{\Delta^2}{2\sigma^2} = \varepsilon.$$

The potential h_t takes two values separated by $\Delta/2$, so its centered version — and therefore $\widetilde{A}_t$ — has absolute value at most $\Delta/2 \le \Delta$. The common-support condition is automatic for the softmax class, and the support can be designed with $A_{\mathrm{des}} = \Delta$.

Finally,

$$J(\pi^\star) - J(\pi_t) = \frac{\Delta}{2}(1 - \pi_t(y_1)) \ge \frac{\Delta}{2}(1 - w) \ge \frac{\Delta}{4}.$$

Substitution of $\Delta = \sigma\sqrt{C\varepsilon}$ proves the lower bound. The same construction supplies a pairwise squared-error budget of order $\sigma^2\varepsilon$, so the square-root obstruction is not specific to cross-entropy. □

Corollary C.1 (No first-order dependence in this assumption class)**.** *There are no exponent $p > \frac{1}{2}$, function $\psi(T) \to 0$, and finite nonnegative coefficient functions h_1, h_2 of $(A, C, D_0(\pi^\star), \sigma, \beta, c_\delta, R_{\max}, |\mathcal{Y}|)$ such that*

$$J(\pi^\star) - J(\bar{\pi}_T) \le h_1\,\psi(T) + h_2 \cdot \big(\max_t \varepsilon_t^{\mathrm{CE}}\big)^p$$

for every instance and iterate sequence satisfying the hypotheses of Theorem 2.

Proof. Pin $R_{\max}$ at any prescribed value by the free offset c_0, fix C (hence $|\mathcal{Y}| = C$ and $D_0(\pi^\star) = \log C$), and fix an admissible update bound $A_0 > 0$ independent of ε, sizing the support with $A_{\mathrm{des}} = A_0$. For every sufficiently small ε the instance above satisfies every hypothesis of Theorem 2 with budgets $\varepsilon_t^{\mathrm{CE}} \le \varepsilon$ and centered updates $|\widetilde{A}_t| \le \sigma\sqrt{C\varepsilon} \le A_0$, so it is admissible with the fixed tuple $(A_0, C, \log C, \sigma, \beta, c_\delta, R_{\max}, C)$, and h_1, h_2 are fixed numbers along the family. Choose $T = T(\varepsilon)$ so large that $h_1\psi(T) \le \frac{\sigma}{8}\sqrt{C\varepsilon}$; since every iterate — hence the mixture — has regret at least $\frac{\sigma}{4}\sqrt{C\varepsilon}$, the proposed bound forces $h_2\,\varepsilon^p \ge \frac{\sigma}{8}\sqrt{C}\,\varepsilon^{1/2}$ for all small ε, contradicting $p > \frac{1}{2}$. Fixing A_0 rather than shrinking the update bound with ε is what allows h_2 to depend on A: once A no longer varies along the family, no coefficient in A can re-create the square root. □

C.2 Perfect coverage does not remove the square root

Proposition C.2 (Perfect comparator coverage)**.** *There is a two-action instance with*

$$\mu_t = \pi^\star, \qquad C_t(\pi^\star) = 1,$$

such that the CE risk is at most ε while every iterate has regret at least $\sigma\sqrt{2\varepsilon}/4$.

Proof. Let $\mathcal{Y} = \{y_0, y_1\}$, set

$$r(y_0) = 0, \qquad r(y_1) = \frac{\Delta}{2},$$

and take $\pi^\star = \mu_t = \delta_{y_1}$ while π_0 is uniform. Define

$$h_t(y) = r(y) - \Delta \mathbf{1}\{y = y_1\},$$

and construct π_{t+1} and the exact log-ratio f_t from h_t as in Appendix C.1. Under $\mathcal{D}_t$, $y' = y_1$ always. The scalar pair error has magnitude Δ exactly when $y = y_0$, so

$$\mathbb{E}_{\mathcal{D}_t} \mathrm{KL} \le (1 - \pi_t(y_1))\frac{\Delta^2}{2\sigma^2} \le \frac{\Delta^2}{2\sigma^2}.$$

The update makes y_1 no more likely than its initial probability $1/2$, and therefore

$$J(\pi^\star) - J(\pi_t) = \frac{\Delta}{2}(1 - \pi_t(y_1)) \ge \frac{\Delta}{4}.$$

Set $\Delta = \sigma\sqrt{2\varepsilon}$. □

C.3 Scope of the exclusion across pairwise losses

Remark C.3 (Which pairwise losses the exclusion covers). Corollary C.1 is stated for the tied HL-Gauss risk, but the construction behind it is loss-agnostic in the following sense.

(a) Every smooth gap-fitting loss is covered. Call $\ell : \mathbb{R}^2 \to [0, \infty)$ a *gap-fitting loss* if $\ell(a, a) = 0$ for every a and ℓ is locally upper-quadratic at its minimum: $\ell(a + u, a) \le c_\ell\, u^2$ for all $|u| \le u_0$, for some $c_\ell < \infty$ and $u_0 > 0$. The squared loss, Huber at any threshold, any Bregman divergence of a C^2 convex potential (on a compact neighborhood, where its curvature is bounded), and—by Theorem 1—the tied HL-Gauss loss in its excess form $\ell(g, a) = \mathrm{KL}(\Gamma_\sigma(a) \| \Gamma_\sigma(g))$ (with $c_\ell = 1/(2\sigma^2)$ and $u_0 = \infty$; the raw cross-entropy has floor $H(\Gamma_\sigma(a))$ and is not itself gap-fitting) all qualify. In the instance of Theorem 3 the realized pairwise error lies in $\{0, \pm\Delta\}$ and is nonzero on an event of probability at most $2/C$, so for $\Delta \le u_0$ the population ℓ-risk is at most $(2/C)\, c_\ell \Delta^2$. Setting $\Delta = \sqrt{C\varepsilon/(2c_\ell)}$ gives budget $\varepsilon_t^\ell \le \varepsilon$ at every t, while every iterate has regret at least $\Delta/4 = \frac{1}{4}\sqrt{C/(2c_\ell)}\,\sqrt{\varepsilon}$ and the centered update obeys $A_t \le \Delta$ exactly as before. Repeating the proof of Corollary C.1 with $\varepsilon_t^{\mathrm{CE}}$ replaced by ε_t^ℓ excludes every exponent $p > \frac{1}{2}$ under the correspondingly restated assumption class, with (c_ℓ, u_0) adjoined to the coefficient functions' arguments in place of the HL-specific $(\sigma, \beta, c_\delta)$. Three quantities must be kept apart here: the comparator's task cost $L^\star_{\mathrm{task}}$ (the optimal cost in the offset-free cost form of Section 3.3), the raw surrogate risk $R^\star_\ell$, and the excess surrogate risk $\mathcal{E}_\ell(g) = R_\ell(g) - R^\star_\ell$. Assumption 2 is stated in the excess (KL) form, whose floor is zero at the exact fit; the raw cross-entropy floor is the target entropy $H(\Gamma_\sigma(a))$ and is never zero. The small-loss form $O(\sqrt{L^\star_{\mathrm{task}}\varepsilon} + \varepsilon)$ of Section 3.3 is excluded because the offset invariance of targets, updates, and regret pins $L^\star_{\mathrm{task}} = 0$ on the shifted instance while the regret stays $\Omega(\sqrt{\varepsilon})$; realizability of the exact fit (the tabular softmax class realizes every potential) makes the excess budget attainable but is not what drives the exclusion. The obstruction is the target geometry of the reduction—pairwise gaps specified only up to a common offset—not any property of cross-entropy. As in Theorem 3, this is sharpness under the stated assumption class, not an information-theoretic lower bound over all algorithms.

(b) Bernoulli preference labels: the same format-level exclusion, a different open question. The preference-label reduction (Rafailov et al., 2023) replaces the real-valued gap by a sampled label $z \in \{0, 1\}$ with $\Pr(z = 1 \mid x, y, y') = \sigma_{\mathrm{lgt}}(\eta a_t)$ (σ_{lgt} the logistic sigmoid) and fits g by the logistic log loss. On the instance family of Theorem 3 the raw floor $R^\star_{\mathrm{lgt}}$ is the conditional label entropy $\mathbb{E}\, H(\sigma_{\mathrm{lgt}}(\eta a_t)) \to \log 2$ as $\varepsilon \to 0$, whereas the same event accounting bounds the *excess* logistic

risk of the iterates by $(2/C)\,(\eta\Delta)^2/8$—the Bernoulli KL at logit gap $\eta\Delta$—while the regret is $\Omega(\Delta)$; and the label law, the update, and the regret are all invariant to a common reward offset, so $L^\star_{\mathrm{task}} = 0$ on the shifted instance exactly as in (a). Hence, in the supplied-budget format of Corollary C.1 with the excess logistic risk as the budget, neither first-order dependence nor the small-loss form $O(\sqrt{L^\star_{\mathrm{task}}\varepsilon} + \varepsilon)$ is available for the preference-label reduction either: the exclusion is a property of offset-invariant pairwise reductions whose supplied budget is a locally quadratic excess risk, not of the real-gap format as such; offset invariance alone does not force the square root—an absolute-error budget $\varepsilon^{(1)}_t = \mathbb{E}_{\mathcal{D}_t}|g_t - a_t|$ is offset-invariant, lies outside this class, and transfers linearly, $\mathbb{E}_{\rho,\pi^\star}[m_{\pi,t} - e_t] \le C_t(\pi^\star)\,\varepsilon^{(1)}_t$, by the same coverage step. (Reading the raw floor $R^\star_{\mathrm{lgt}} \approx \log 2$ as the $L^\star$ of a small-loss bound would make such a bound read $\Theta(\sqrt{\varepsilon})$ and appear consistent with the construction; but that floor is neither the task cost of Foster & Krishnamurthy (2021) nor an excess risk, so the appearance of consistency is an artifact of mixing the three quantities.) What the construction does not address is adaptivity to the label-noise level itself: $R^\star_{\mathrm{lgt}}$ is a variance-type quantity that vanishes precisely in high-margin regimes, and a guarantee parametrized by it—the preference-label analogue of the variance adaptivity of Theorem 5—is, to our knowledge, open. None of this is an algorithm-level impossibility for DPO-style methods; as throughout this appendix, the statements concern bounds that see the fitting step only through a supplied population risk budget.

D Additional fast-rate results

D.1 A second-order exponential-weights inequality

Lemma D.1. *Suppose $\eta\widetilde{A}_t(x,y) \le 1$ pointwise. Then*

$$\sum_{t=0}^{T-1} \mathbb{E}_{\rho,\pi^\star}\widetilde{A}_t \le \frac{D_0(\pi^\star)}{\eta} + \eta \sum_{t=0}^{T-1} \mathbb{E}_{\rho,\pi_t}\widetilde{A}_t^2.$$

Proof. For $u \le 1$, $e^u \le 1 + u + u^2$. Hence, using $\mathbb{E}_{\pi_t}\widetilde{A}_t = 0$,

$$\log \mathbb{E}_{\pi_t} e^{\eta\widetilde{A}_t} \le \log\left(1 + \eta^2 \mathbb{E}_{\pi_t}\widetilde{A}_t^2\right) \le \eta^2 \mathbb{E}_{\pi_t}\widetilde{A}_t^2.$$

Substitute this bound into the support-valid KL identity in Lemma B.2 and telescope. □

D.2 Statement and proof of Theorem 4

Theorem 4 is summarized in Section 3.4; its full statement follows.

Theorem 4 (Uniform reward gap)**.** *Let $\pi^\star = \delta_{y^\star(x)}$ be deterministic, with $y^\star(x)$ the unique maximizer and $\Delta \le r(x,y^\star(x)) - r(x,y) \le B_r$ for every $y \ne y^\star(x)$; let $C^\star_t := \operatorname{ess\,sup}_x 1/\mu_t(y^\star(x) \mid x) < \infty$, $A_t \le A$, $\eta A \le 1$, and $\eta \le 1/(8B_r)$. Then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\big[J(\pi^\star) - J(\pi_t)\big] \le \frac{2D_0(\pi^\star)}{\eta T} + \frac{4\eta}{T}\sum_{t=0}^{T-1}\kappa_{\mathrm{HL}}\varepsilon^{\mathrm{CE}}_t + \frac{2}{\Delta T}\sum_{t=0}^{T-1} C^\star_t\,\kappa_{\mathrm{HL}}\varepsilon^{\mathrm{CE}}_t.$$

Proof of Theorem 4. Let

$$R_t(x) := r(x,y^\star(x)) - \mathbb{E}_{y\sim\pi_t} r(x,y), \qquad R_t := \mathbb{E}_x R_t(x).$$

The uniform gap gives

$$R_t(x) \ge \Delta\big(1 - \pi_t(y^\star(x) \mid x)\big).$$

Define

$$q_t(x) := \mathbb{E}_{y\sim\pi_t}\big(e_t(x,y) - e_t(x,y^\star(x))\big)^2$$

and

$$B_t(x) := m_{\pi,t}(x) - e_t(x,y^\star(x)).$$

The integrand defining $B_t(x)$ vanishes at the optimal action, so Cauchy–Schwarz gives

$$B_t(x) \le \sqrt{q_t(x)\big(1 - \pi_t(y^\star(x) \mid x)\big)} \le \sqrt{\frac{R_t(x)q_t(x)}{\Delta}}.$$

A second Cauchy–Schwarz inequality gives

$$B_t := \mathbb{E}_x B_t(x) \le \sqrt{\frac{R_t Q_t}{\Delta}}, \qquad Q_t := \mathbb{E}_x q_t(x).$$

The pairwise MSE contains the event $y' = y^\star(x)$, so Corollary 1 yields

$$\kappa_{\mathrm{HL}} \varepsilon_t^{\mathrm{CE}} \ge \mathbb{E}_x \left[\mu_t(y^\star(x) \mid x) q_t(x) \right] \ge \frac{Q_t}{C_t^\star}.$$

Therefore

$$B_t \le \sqrt{\frac{C_t^\star \kappa_{\mathrm{HL}} \varepsilon_t^{\mathrm{CE}} R_t}{\Delta}} \le \frac{R_t}{4} + \frac{C_t^\star \kappa_{\mathrm{HL}} \varepsilon_t^{\mathrm{CE}}}{\Delta}.$$

Next,

$$\mathbb{E}_{\rho,\pi_t} \widetilde{A}_t^2 \le 2\mathbb{E}_{\rho,\pi_t} (A_t^r)^2 + 2\mathbb{E}_{\rho,\pi_t} (e_t - m_{\pi,t})^2.$$

Let $L_t(x,y) = r(x, y^\star(x)) - r(x,y) \in [0, B_r]$. Then

$$\mathbb{E}_{\pi_t} (A_t^r)^2 = \mathrm{Var}_{\pi_t}(L_t) \le \mathbb{E}_{\pi_t} L_t^2 \le B_r R_t(x).$$

Lemma B.1 and Corollary 1 give

$$\mathbb{E}_{\rho,\pi_t} (e_t - m_{\pi,t})^2 \le \kappa_{\mathrm{HL}} \varepsilon_t^{\mathrm{CE}}.$$

Since

$$R_t = \mathbb{E}_{\rho,\pi^\star} \widetilde{A}_t + B_t,$$

Lemma D.1 and the preceding bounds imply

$$\sum_t R_t \le \frac{D_0}{\eta} + 2\eta B_r \sum_t R_t + 2\eta \sum_t \kappa_{\mathrm{HL}} \varepsilon_t^{\mathrm{CE}} + \frac{1}{4} \sum_t R_t + \frac{1}{\Delta} \sum_t C_t^\star \kappa_{\mathrm{HL}} \varepsilon_t^{\mathrm{CE}}.$$

Because $\eta \le 1/(8B_r)$, the coefficients of $\sum_t R_t$ on the right sum to at most $1/2$. Move them to the left, multiply by two, and divide by T. □

D.3 Tsybakov-style context margin

To avoid conflict with $\beta = \sigma/\varsigma$, denote the context-margin exponent by $\nu > 0$.

Theorem D.2 (Tsybakov-style refinement). *Let $\pi^\star = \delta_{y^\star(x)}$ and suppose the assumptions of Theorem 2 hold. Assume:*

1. *$y^\star(x)$ is the unique maximizer and*
$$\Delta(x) := r(x, y^\star(x)) - \sup_{y \ne y^\star(x)} r(x,y) > 0;$$
2. *all reward gaps are at most B_r;*
3. *for constants $M > 0$, $\nu > 0$, and $u_0 > 0$,*
$$\rho\{x : \Delta(x) \le u\} \le M u^\nu \qquad (0 < u \le u_0);$$
4. *$C_t^\star \le C^\star$ and*
$$\kappa_{\mathrm{HL}} \varepsilon_t^{\mathrm{CE}} \le \bar{\varepsilon}_2$$
for every t;
5. *$A_t \le A$, $\eta A \le 1$, and $\eta \le 1/(8B_r)$.*

If

$$\tau_\star := \left(\frac{C^\star \bar{\varepsilon}_2}{M} \right)^{1/(\nu+2)} \le u_0,$$

then

$$\boxed{\frac{1}{T} \sum_{t=0}^{T-1} \left[J(\pi^\star) - J(\pi_t) \right] \le \frac{2D_0(\pi^\star)}{\eta T} + 4\eta \bar{\varepsilon}_2 + 4M^{1/(\nu+2)} (C^\star \bar{\varepsilon}_2)^{(\nu+1)/(\nu+2)}.}$$

Proof. Fix $\tau \in (0, u_0]$ and split contexts into

$$H_\tau := \{x : \Delta(x) > \tau\}, \qquad L_\tau := \{x : \Delta(x) \le \tau\}.$$

On H_τ, the uniform-gap calculation with Δ replaced by τ gives

$$B_{t,H_\tau} \le \frac{R_t}{4} + \frac{C^\star \bar{\varepsilon}_2}{\tau}.$$

On L_τ, Cauchy–Schwarz and the margin condition give

$$B_{t,L_\tau} \le \sqrt{\rho(L_\tau) Q_t} \le \sqrt{M \tau^\nu C^\star \bar{\varepsilon}_2}.$$

Thus

$$B_t \le \frac{R_t}{4} + \frac{C^\star \bar{\varepsilon}_2}{\tau} + \sqrt{M C^\star \bar{\varepsilon}_2}\, \tau^{\nu/2}.$$

The last two terms are equal at $\tau = \tau_\star$, and each equals

$$M^{1/(\nu+2)} (C^\star \bar{\varepsilon}_2)^{(\nu+1)/(\nu+2)}.$$

Combine this estimate with the second-order mirror-descent calculation from Appendix D.2, absorb one half of $\sum_t R_t$, and divide by T. □

D.4 Necessity of the $1/\Delta$ coefficient

Proposition D.3 (Margin dependence is necessary)**.** *Fix $\Delta > 0$ and an integer $C \ge 2$. There is a one-context C-action instance with uniform reward gap Δ, uniform base distribution, full-support tabular softmax policies, and a deterministic comparator such that*

$$C_t^\star = C_t(\pi^\star) = C,$$

and the CE risk is at most

$$\bar{\varepsilon}_{\mathrm{CE}} := \frac{4\Delta^2}{C\sigma^2},$$

while every iterate has regret at least $\Delta/2$. Consequently:

***(a)** a bound of the form*

$$J(\pi^\star) - J(\pi_t) \le h(\Delta) C_t^\star \bar{\varepsilon}_{\mathrm{CE}}$$

valid on this family requires

$$h(\Delta) \ge \frac{\sigma^2}{8\Delta};$$

***(b)** along the coupled diagonal $\Delta = \frac{\sigma}{2}\sqrt{C\bar{\varepsilon}_{\mathrm{CE}}}$ the regret is at least $\frac{\sigma}{4}\sqrt{C\bar{\varepsilon}_{\mathrm{CE}}}$: a margin hypothesis improves the rate only through a fixed margin scale, and uniformly over margins coupled to the error scale the square root persists;*

***(c)** with $C_{\infty,t} = C$ on this family, the greedy readout of Theorem D.4 (D.5 below) selects a suboptimal action and incurs regret exactly Δ, so any bound of the form $h(\Delta)\, C_{\infty,t}\, \bar{\varepsilon}_{\mathrm{CE}}$ on the greedy regret requires $h(\Delta) \ge \sigma^2/(4\Delta)$ — the $1/\Delta$ shape of Theorem D.4, which is therefore exponent-tight in Δ (a match in the exponent, not in the constant).*

Proof. Use the construction of Appendix C.1 with reward $r(y_1) = c_0 + \Delta$ and baseline reward c_0. Define

$$h_t(y) = r(y) - 2\Delta \mathbf{1}\{y = y_1\}$$

and normalize it to obtain the exact policy log-ratio. The effective update value of y_1 is $c_0 - \Delta$, so its probability decreases from $1/C$. The scalar error has magnitude 2Δ when exactly one endpoint is y_1, an event of probability at most $2/C$. Hence

$$\mathbb{E}\,\mathrm{KL} \le \frac{2}{C} \frac{(2\Delta)^2}{2\sigma^2} = \frac{4\Delta^2}{C\sigma^2}.$$

Every iterate has regret

$$\Delta(1 - \pi_t(y_1)) \ge \Delta(1 - 1/C) \ge \frac{\Delta}{2}.$$

Substitution into the proposed first-order form gives the lower bound on $h(\Delta)$, which is (a). For (b): the stated budget is $\bar{\varepsilon}_{\mathrm{CE}} = 4\Delta^2/(C\sigma^2)$, so $\Delta = \frac{\sigma}{2}\sqrt{C\bar{\varepsilon}_{\mathrm{CE}}}$ and the per-iterate regret bound $\Delta/2$ reads $\frac{\sigma}{4}\sqrt{C\bar{\varepsilon}_{\mathrm{CE}}}$. For (c): the update potential has $h_t(y_1) = c_0 - \Delta$ strictly below the baseline value c_0, and f_t orders actions as h_t does (they differ by an x-constant), so the greedy readout $\arg\max_y f_t$ selects some $y \neq y_1$ and incurs regret exactly Δ; substituting the budget into $h(\Delta)\, C\, \bar{\varepsilon}_{\mathrm{CE}} \geq \Delta$ gives $h(\Delta) \geq \sigma^2/(4\Delta)$. These statements and Theorem 4 are complementary, not contradictory: the first-order bound holds for each fixed Δ, and the family shows its Δ-dependence cannot be improved. □

D.5 Greedy readout for finite action spaces

Theorem D.4 (Margin plus greedy readout)**.** *Let $|\mathcal{Y}| < \infty$ and fix t. Suppose the CE-to-MSE assumptions hold, every suboptimal action has reward gap at least $\Delta > 0$, and*

$$C_{\infty,t} := \operatorname*{ess\,sup}_{x,y} \frac{1}{\mu_t(y \mid x)} < \infty.$$

Define

$$\widehat{y}_t(x) := \arg\max_y f_t(x, y),$$

with arbitrary tie-breaking. Then

$$\mathbb{E}_{x\sim\rho}\left[\max_y r(x,y) - r(x, \widehat{y}_t(x))\right] \leq \frac{2C_{\infty,t}\kappa_{\mathrm{HL}}\varepsilon_t^{\mathrm{CE}}}{\Delta},$$

and the greedy choice is suboptimal with probability at most $2C_{\infty,t}\kappa_{\mathrm{HL}}\varepsilon_t^{\mathrm{CE}}/\Delta^2$.

Proof. Fix x, write $\widetilde{e}_t = e_t - m_{\mu,t}$, and let $d(x) := \max_y r(x,y) - r(x, \widehat{y}_t(x))$. If $d(x) = 0$ there is nothing to prove. Otherwise choose an optimal action $y^\star$; then $d(x) \geq \Delta$ and $f_t(x, \widehat{y}) \geq f_t(x, y^\star)$, so

$$\widetilde{e}_t(x, \widehat{y}) - \widetilde{e}_t(x, y^\star) = \big(f_t(x, \widehat{y}) - f_t(x, y^\star)\big) + \big(r(x, y^\star) - r(x, \widehat{y})\big) \geq d(x).$$

Hence

$$d(x) \leq \frac{d(x)^2}{\Delta} \leq \frac{\big(\widetilde{e}_t(x, \widehat{y}) - \widetilde{e}_t(x, y^\star)\big)^2}{\Delta} \leq \frac{2}{\Delta}\left(\widetilde{e}_t(x, \widehat{y})^2 + \widetilde{e}_t(x, y^\star)^2\right).$$

The two actions are distinct and each carries μ_t-mass at least $1/C_{\infty,t}$, so

$$\widetilde{e}_t(x, \widehat{y})^2 + \widetilde{e}_t(x, y^\star)^2 \leq C_{\infty,t}\, \mathbb{E}_{z\sim\mu_t(\cdot|x)} \widetilde{e}_t(x, z)^2 = C_{\infty,t}\, \mathrm{Var}_{\mu_t}(e_t)(x).$$

Average over x and use Lemma B.1 with Corollary 1: $\mathbb{E}_x\, \mathrm{Var}_{\mu_t}(e_t) \leq \mathbb{E}_{\mathcal{D}_t}(g_t - a_t)^2 \leq \kappa_{\mathrm{HL}}\varepsilon_t^{\mathrm{CE}}$. For the probability statement bound the indicator $\mathbf{1}\{d(x) > 0\} \leq (\widetilde{e}_t(x, \widehat{y}) - \widetilde{e}_t(x, y^\star))^2/\Delta^2$ instead of $d(x)$ and repeat the same steps. □

Multiplying the error probability by the reward range $\bar{R} := \operatorname{ess\,sup}_x(\max_y r(x,y) - \min_y r(x,y)) \leq 2R_{\max}$ would give the weaker form $2\bar{R}\, C_{\infty,t}\kappa_{\mathrm{HL}}\varepsilon_t^{\mathrm{CE}}/\Delta^2$; since $\bar{R} \geq \Delta$ whenever a suboptimal action exists, the direct bound above is never worse, and no Δ^{-2} dependence of the regret is forced. The greedy policy is not the KL-conservative REBEL iterate, and exact maximization can be intractable in sequence spaces. The result is mainly relevant to small finite action spaces. The $1/\Delta$ coefficient shape is exponent-tight: Proposition D.3(c).

E Local Fisher and GGN geometry

E.1 Local scalar expansion

For

$$\ell(g, a) := \mathrm{CE}(\Gamma_\sigma(a), \Gamma_\sigma(g)),$$

Lemma A.1 gives

$$\partial_g \ell(a, a) = 0, \qquad \partial_{gg}\ell(a, a) = \mathcal{I}_{\mathrm{HL}}(a).$$

Proposition E.1 (Target-centered expansion)**.** *Let*

$$B_3 := \frac{(2W)^3 + \varsigma^3}{\sigma^6}.$$

For every finite a, g,

$$\ell(g, a) = \mathrm{H}(\Gamma_\sigma(a)) + \frac{1}{2}\mathcal{I}_{\mathrm{HL}}(a)(g - a)^2 + \mathcal{R}_3(g, a),$$

with

$$|\mathcal{R}_3(g, a)| \le \frac{B_3}{6}|g - a|^3.$$

Proof. Differentiating the conditional-variance identities shows that the third derivative is a difference of third central moments divided by σ^6. A random variable supported on an interval of length d has absolute third central moment at most d^3. Taylor's theorem gives the result. □

On I_δ,

$$\alpha_{\mathrm{HL}} \le \mathcal{I}_{\mathrm{HL}}(a) \le \frac{1}{\sigma^2}.$$

E.2 Benign scalar landscape

Proposition E.2 (Unique minimizer and bounded tails)**.** *Fix $a \in \mathbb{R}$ and $q = \Gamma_\sigma(a)$. Then*

$$\operatorname{sign} \ell_q'(g) = \operatorname{sign}(g - a) \qquad (g \ne a).$$

Thus $g = a$ is the unique critical point and global minimizer. Moreover,

$$|\ell_q'(g)| \le \frac{2W}{\sigma^2}$$

globally. Hence the scalar loss is uniformly quadratic on I_δ and has at most linear growth in its far tails.

Proof. For $g > a$, the Gaussian likelihood ratio between locations g and a is increasing in the latent scalar. Integrating over ordered bins implies monotone-likelihood-ratio, hence first-order stochastic, dominance of $\Gamma_\sigma(g)$ over $\Gamma_\sigma(a)$ in the bin index. Since $i \mapsto m_i(g)$ is strictly increasing,

$$\bar{m}(g) > \sum_i \Gamma_\sigma(a)_i m_i(g),$$

and Lemma A.1 gives $\ell_q'(g) > 0$. The case $g < a$ is symmetric. The gradient bound follows because the two means in Lemma A.1 lie in $[-W, W]$. □

Quantitative tail bound. With $\mathsf{M}(q) := \sum_i z_i q_i$ the decoded mean of q, if $|a| \le W - c_\delta\sigma$ and $W \ge \sigma$, then for every $g \ge W$,

$$\ell_q'(g) \ge \frac{(W - a) - 0.84\,\sigma - \varsigma/2 - |\mathsf{M}(q) - a|}{\sigma^2},$$

and symmetrically for $g \le -W$: the restoring gradient does not vanish outside the support, so an optimizer that overshoots the support is pushed back at a rate the design controls.

Proof. For $g \ge W$: $\bar{m}(g) \ge \bar{m}(W)$, since $\bar{m}' = V_I/\sigma^2 \ge 0$ by the same exponential-family calculus as Lemma A.1. The exact truncated-mean formula at the edge gives $\bar{m}(W) \ge W - c_0\sigma$ with $c_0 = \varphi(0)/\big(\frac{1}{2} - \Phi(-2W/\sigma)\big) \le 0.84$ for $W \ge \sigma$. Each conditional mean lies below its bin's upper edge, so $\sum_i q_i m_i(g) \le \sum_i q_i e_i = \mathsf{M}(q) + \varsigma/2 \le a + |\mathsf{M}(q) - a| + \varsigma/2$. Insert both into the gradient identity of Lemma A.1. □

E.3 THREE DISTINCT LOCAL QUADRATIC MODELS

Let $\bar{\theta}$ be a parameter value and define the pairwise score feature

$$d_{\bar{\theta},t}(x, y, y') := \nabla_\theta \log \pi_{\bar{\theta}}(y \mid x) - \nabla_\theta \log \pi_{\bar{\theta}}(y' \mid x).$$

The scalar output linearization is

$$g_{\bar{\theta}+\delta,t} \approx \bar{g} + \frac{1}{\eta} d_{\bar{\theta},t}^\top \delta.$$

Target-centered surrogate. Combining this linearization with Proposition E.1 gives

$$\frac{1}{2}\mathbb{E}_{\mathcal{D}_t}\left[\mathcal{I}_{\mathrm{HL}}(a_t)\left(\bar{g} + \frac{1}{\eta} d^\top \delta - a_t\right)^2\right].$$

Its minimum-Euclidean-norm solution is

$$\delta_{\mathrm{target}} = \eta\left(\mathbb{E}[\mathcal{I}_{\mathrm{HL}}(a_t) d d^\top]\right)^\dagger \mathbb{E}[\mathcal{I}_{\mathrm{HL}}(a_t) d (a_t - \bar{g})].$$

This is a reward-gap-dependent pairwise metric. Unless the weights are effectively constant, it is not the Fisher matrix of a single policy distribution.

Exact prediction-centered quadratic model. At the current policy parameter θ_t, $g_{\theta_t,t} = 0$. The exact scalar Taylor model around the current prediction uses

$$s(0, a) := \partial_g \ell(0, a),$$

and

$$H(0, a) := \partial_{gg} \ell(0, a) = \frac{V_I(0) - \sum_i \Gamma_\sigma(a)_i V_i(0)}{\sigma^4}.$$

Thus the exact current-prediction curvature is in general target dependent; it coincides with the model Fisher at $a = 0$ but need not elsewhere (for a symmetric two-bin encoding, for instance, $V_1(0) = V_2(0)$ and $H(0, a)$ is constant in a).

Fisher scoring. Replacing $H(0, a)$ by the model Fisher $\mathcal{I}_{\mathrm{HL}}(0)$ gives a constant-weight pairwise least-squares model with pseudo-target

$$a_{\mathrm{pseudo}}(a) := -\frac{s(0, a)}{\mathcal{I}_{\mathrm{HL}}(0)}.$$

Smoothness of the matched family gives

$$a_{\mathrm{pseudo}}(a) = a + O(a^2)$$

as $a \to 0$. In the same regime,

$$\mathcal{I}_{\mathrm{HL}}(a) = \mathcal{I}_{\mathrm{HL}}(0) + O(a), \qquad H(0, a) = \mathcal{I}_{\mathrm{HL}}(0) + O(a).$$

Therefore all three models agree to leading order for small reward gaps. At finite gaps, only Fisher scoring uses a constant model-Fisher weight; the target-centered and exact prediction-centered models retain sample-dependent curvatures. None of these local approximations is used in the performance theorems.

F STOCHASTIC TARGETS AND STRUCTURAL RESULTS

F.1 MIXTURE BAYES TARGET AND EXPLICIT BIAS

Let $\tilde{a}$ be a random scalar target conditional on a fixed pair and define

$$\bar{q} := \mathbb{E}[\Gamma_\sigma(\tilde{a})].$$

Proposition 2 (Mixture decomposition; restated)**.** *Suppose $\mathbb{E}|\tilde{a}| < \infty$ and the displayed KL expectations are finite. For every categorical prediction p with positive entries,*

$$\mathbb{E}_{\tilde{a}}\,\mathrm{KL}(\Gamma_\sigma(\tilde{a}) \| p) = \mathbb{E}_{\tilde{a}}\,\mathrm{KL}(\Gamma_\sigma(\tilde{a}) \| \bar{q}) + \mathrm{KL}(\bar{q} \| p).$$

Thus the Bayes categorical target is $\bar{q}$, not generally $\Gamma_\sigma(\mathbb{E}\tilde{a})$.

Let

$$g^\star \in \arg\min_{g \in I_\delta} \mathrm{KL}(\bar{q} \| \Gamma_\sigma(g)), \qquad \bar{a} := \mathbb{E}\tilde{a},$$

and define

$$\mathrm{excess}_{\mathrm{CE}}(g) := \mathbb{E}_{\tilde{a}}\, \mathrm{KL}(\Gamma_\sigma(\tilde{a}) \| \Gamma_\sigma(g)) - \inf_{u \in I_\delta} \mathbb{E}_{\tilde{a}}\, \mathrm{KL}(\Gamma_\sigma(\tilde{a}) \| \Gamma_\sigma(u)).$$

Then, for every $g \in I_\delta$,

$$(g - g^\star)^2 \le \kappa_{\mathrm{HL}} \left[\mathrm{KL}(\bar{q} \| \Gamma_\sigma(g)) - \mathrm{KL}(\bar{q} \| \Gamma_\sigma(g^\star)) \right],$$

and consequently

$$\boxed{(g - \bar{a})^2 \le 2\kappa_{\mathrm{HL}}\, \mathrm{excess}_{\mathrm{CE}}(g) + 2(g^\star - \bar{a})^2.}$$

Proof. Cross-entropy is linear in its first argument, so the expected risk differs from $\mathrm{KL}(\bar{q} \| p)$ by a constant independent of p. Strong convexity of $g \mapsto \mathrm{CE}(\bar{q}, \Gamma_\sigma(g))$ on I_δ gives the first inequality. The second follows from

$$(g - \bar{a})^2 \le 2(g - g^\star)^2 + 2(g^\star - \bar{a})^2.$$

□

The last term is a genuine mixture or approximation bias. If $\xi \sim \mathcal{N}(0, \tau^2)$ and P_i^c denotes the unnormalized mass of bin i at Gaussian scale c, Gaussian convolution gives

$$\mathbb{E}_\xi P_i^{\sigma_0}(a + \xi) = P_i^{\sqrt{\sigma_0^2 + \tau^2}}(a),$$

but normalized targets contain the random denominator $Z_{\sigma_0}(a + \xi)$. In general,

$$\mathbb{E}\left[\frac{P_i}{Z}\right] \neq \frac{\mathbb{E}P_i}{\mathbb{E}Z}.$$

Therefore matched Gaussian smoothing does not by itself prove $g^\star = \bar{a}$ on a finite normalized support. A variance-weighted mean-reward theorem for noisy normalized targets would require a separate quantitative bound on this bias; Appendix F.5 supplies a two-sided one (Theorem 6). Appendix F.4 records what survives when that step is granted by hypothesis: Theorem 5, whose in-family risk condition at the matched scale excludes from its scope exactly the renormalization gap displayed above.

F.2 Normalized and unnormalized bin masses are different objectives

Let $\widetilde{p}_i(g) := P_i(g)$. For a normalized target q,

$$\mathrm{CE}(q, \widetilde{p}(g)) = \mathrm{CE}(q, \Gamma_\sigma(g)) - \log Z(g).$$

For $q = \Gamma_\sigma(a)$, the normalized objective has derivative zero at $g = a$, whereas

$$\frac{d}{dg}\big[-\log Z(g)\big]\Big|_{g=a} = -\frac{\bar{m}(a) - a}{\sigma^2}.$$

Indeed, $\partial_g \log Z(g) = (\bar{m}(g) - g)/\sigma^2$. For $a > 0$,

$$\mathbb{E}\big[(Y_a - a)\mathbf{1}\{Y_a \in [-W, W]\}\big] = \sigma^2\big(\varphi_\sigma(W + a) - \varphi_\sigma(W - a)\big) < 0,$$

so $\bar{m}(a) < a$; the case $a < 0$ is symmetric. Thus $\bar{m}(a) \neq a$ for every $a \neq 0$, and the unnormalized objective generally displaces the scalar minimizer. The calibration and exact-MD results apply to the normalized construction.

F.3 Only linear pairwise remappings preserve all potentials

Proposition 3 (Only linear pairwise remappings preserve all potentials). *Let $B \in (0, \infty]$ and let ψ be continuous on $[-B, B]$. Suppose that for every triple (r_1, r_2, r_3) whose pairwise differences all lie in $[-B, B]$ there exist potentials (f_1, f_2, f_3) such that*

$$f_i - f_j = \psi(r_i - r_j)$$

for all pairs. Then

$$\psi(u) = cu \qquad \text{on } [-B, B]$$

for a constant c.

Proof. Fix u, v with $u, v, u + v \in [-B, B]$. The triple $(u + v,\, v,\, 0)$ has pairwise differences $\pm u, \pm v, \pm(u + v)$, all in $[-B, B]$, so telescoping gives

$$\psi(u + v) = f_1 - f_3 = (f_1 - f_2) + (f_2 - f_3) = \psi(u) + \psi(v).$$

This restricted additive equation gives $\psi(0) = 0$ and $\psi(-u) = -\psi(u)$, and, by induction on k, $\psi(ku/n) = (k/n)\,\psi(u)$ for integers $0 \le k \le n$ and $|u| \le B$; continuity extends the multiplier to all of $[0, 1]$, so $\psi(u) = (u/B)\,\psi(B)$ on $[0, B]$ and, by oddness, on $[-B, B]$ (for $B = \infty$ apply this on every finite $[-B', B']$; the constants agree). The domain must be symmetric: on $[0, 1]$ the only admissible triples are constant, and $\psi(u) = u^2$ satisfies the hypothesis vacuously. □

Thus hard clipping or a generic nonlinear pointwise transformation of pairwise targets can violate cycle consistency: a policy-induced pairwise prediction must be a difference of a scalar potential.

F.4 Statement and proof of Theorem 5 (conditional variance-weighted bound)

Theorem 5 is summarized in Section 4 and stated in full here; throughout, $\sigma_0 = \beta_0 \varsigma$ is the base encoding scale, $s(x)^2 := \sigma_0^2 + \tau^2(x)$ the matched per-context prediction scale, and $s_{\max}^2 := \sigma_0^2 + \tau_{\max}^2$.

Theorem 5 (Variance-weighted bound in the known-Gaussian corner). *Fix $\eta > 0$ and a comparator $\pi^\star$ with $C := \max_t C_t(\pi^\star) < \infty$ and $D_0(\pi^\star) < \infty$. Suppose Assumptions 1 and 3 hold with $A_t \le A$, that $v(c_\delta) > 1/(4\beta_0^2)$, that $W \ge \max\{2R_{\max}, 2A\} + c_\delta s_{\max}$, and that for every t the* in-family *risk condition $\mathbb{E}_{\mathcal{D}_t}\big[\mathrm{KL}\big(\Gamma_{s(x)}(a_t) \,\big\|\, \Gamma_{s(x)}(g_t)\big)\big] \le \varepsilon_{\mathrm{CE}}$ holds, where Γ_c denotes the construction of Section 2 at scale c (same bins, same support). Then, with $c_\kappa := 2/(v(c_\delta) - 1/(4\beta_0^2))$ and $\bar{S}^2 := \mathbb{E}_{x\sim\rho}[\sigma_0^2 + \tau^2(x)]$,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\big[J(\pi^\star) - J(\pi_t)\big] \le \frac{D_0(\pi^\star)}{\eta T} + \frac{\eta}{2T}\sum_{t=0}^{T-1} A_t^2 + \sqrt{c_\kappa\, C\, \varepsilon_{\mathrm{CE}}\, \bar{S}^2}.$$

Proof of Theorem 5. Write $s(x)^2 = \sigma_0^2 + \tau^2(x) \in [\sigma_0^2, s_{\max}^2]$ and $c_\kappa = 2/(v(c_\delta) - 1/(4\beta_0^2))$. Four steps; only the first, second, and fourth deviate from Appendix B, and each deviation is a pointwise-in-x manipulation.

Step 1 (per-context calibration). For each x, apply Theorem 1 at scale $s(x)$: the effective ratio is $\beta(x) = s(x)/\varsigma \ge \beta_0$, so $v(c_\delta) - 1/(4\beta(x)^2) \ge v(c_\delta) - 1/(4\beta_0^2) > 0$, and the per-context calibration constant is at most $c_\kappa s(x)^2$. The enlarged margin $c_\delta s_{\max} \ge c_\delta s(x)$ together with $|a_t| \le 2R_{\max}$ and $|g_t| \le 2A$ (Assumption 3, as in Section 2) places every target and prediction in the interior required at scale $s(x)$. Pointwise,

$$(g_t - a_t)^2 \le c_\kappa\, s(x)^2\, \mathrm{KL}\big(\Gamma_{s(x)}(a_t) \,\big\|\, \Gamma_{s(x)}(g_t)\big),$$
$$\text{hence} \qquad \mathbb{E}_{\mathcal{D}_t}\big[s(x)^{-2}(g_t - a_t)^2\big] \le c_\kappa\, \varepsilon_{\mathrm{CE}}.$$

Step 2 (weighted exact decomposition). Lemma B.1 is an identity conditionally on x. Multiplying it by $s(x)^{-2} \ge 0$ and taking $\mathbb{E}_{x\sim\rho}$ gives

$$\mathbb{E}_x\Big[s(x)^{-2}\,\mathbb{E}_{y'\sim\mu_t}\big(e_t - m_{\pi,t}\big)^2\Big] \le c_\kappa\, \varepsilon_{\mathrm{CE}}.$$

Step 3 (mirror-descent term). Lemma B.2 is loss-free and unchanged; its hypotheses are among the theorem's.

Step 4 (weighted transfer). By Cauchy–Schwarz over x, then Jensen and the per-context change of measure ($\mathbb{E}_{\pi^\star(\cdot|x)}h \le C\,\mathbb{E}_{\mu_t(\cdot|x)}h$ for $h \ge 0$, which commutes with any x-weight),

$$\begin{aligned}\mathbb{E}_{\rho,\pi^\star}\big(A_t^r - \widetilde{A}_t\big) &\le \mathbb{E}_x\Big[s(x)\cdot s(x)^{-1}\,\mathbb{E}_{\pi^\star}\big|e_t - m_{\pi,t}\big|\Big]\\ &\le \sqrt{\mathbb{E}_\rho[s^2]}\,\sqrt{\mathbb{E}_x\Big[s^{-2}\,C\,\mathbb{E}_{\mu_t}\big(e_t - m_{\pi,t}\big)^2\Big]}\\ &\le \sqrt{c_\kappa\,C\,\varepsilon_{\mathrm{CE}}\,\bar{S}^2}.\end{aligned}$$

Assembling as in Appendix B.3 — the performance-difference identity plus Lemma B.2, averaged over t — proves Theorem 5. The dependency audit of Appendix G carries over row by row: the only loss-specific step is still G-1, applied per context at the matched scale. □

Remark 1. *Hypothesis.* The empirical objective encodes targets $\Gamma_{\sigma_0}(\tilde{a})$; cross-entropy is linear in the target, and at the level of unnormalized masses $\mathbb{E}_\xi P_i^{\sigma_0}(a_t+\xi) = P_i^{s(x)}(a_t)$ exactly (Appendix F.1), so up to boundary renormalization the induced population target is $\Gamma_{s(x)}(a_t)$, the object the condition is stated against. The renormalization gap is the mixture bias of Proposition 2, bounded in Appendix F.5; the hypothesis excludes it from scope. *Degeneration.* As $\tau \to 0$: $s(x) \to \sigma_0$, $\bar{S}^2 \to \sigma_0^2$, $c_\kappa \bar{S}^2 \to \kappa_{\mathrm{HL}}$, and the bound reduces to the uniform-constant form of Theorem 2.

F.5 Statement and proof of Theorem 6 (two-sided bound on the mixture bias)

Theorem 6 is summarized in Section 4 and stated in full here. Notation is that of Appendix A: $\bar{m}(s) = \mathbb{E}[\widetilde{Y}_s]$ and $V_I(s) = \mathrm{Var}(\widetilde{Y}_s)$ are the mean and variance of the Gaussian at location s truncated to $I = [-W, W]$, $m_i(s)$ and $V_i(s)$ their conditional counterparts on bin i, and $v(\cdot)$ is the margin function of Section 2, which is strictly increasing on $(0,\infty)$ (its first part has derivative $2t^2\varphi(t)$, and the subtracted ratio $\varphi(t)/(1-2\Phi(-t))$ is decreasing). Write $d(s) := W - |s|$ for the signed distance from s to the nearer support edge, and $v_+(t) := \max\{v(t), 0\}$ for $t > 0$, $v_+(t) := 0$ for $t \le 0$. For a reference point $u \in \mathbb{R}$ define the *frozen decoder* and the *ripple*

$$D_u(a) := \sum_{i=1}^m \Gamma_\sigma(a)_i\, m_i(u), \qquad \rho_u(a) := D_u(a) - \bar{m}(a) = \sum_{i=1}^m \Gamma_\sigma(a)_i\big[m_i(u) - m_i(a)\big],$$

where the second form uses $\bar{m}(a) = \sum_i \Gamma_\sigma(a)_i\, m_i(a)$ (law of total expectation); thus $\rho_u(u) = 0$ and $D_u(u) = \bar{m}(u)$. Throughout, $\ell := \ell_{\bar{q}}$, i.e. $\ell(g) = \mathrm{CE}(\bar{q}, \Gamma_\sigma(g))$.

Theorem 6 (Two-sided bound on the mixture bias)**.** *In the setting of Proposition 2, suppose $\bar{a} = \mathbb{E}\tilde{a} \in I_\delta$ and let $g^\star \in \arg\min_{g\in I_\delta}\ell(g)$. Then:*

(i) (Gradient sandwich.) *$|g^\star - \bar{a}| \le \frac{\kappa_{\mathrm{HL}}}{2}\,|\ell'(\bar{a})|$, and if $g^\star$ lies in the interior of I_δ then also $|g^\star - \bar{a}| \ge \sigma^2|\ell'(\bar{a})|$, where*

$$\sigma^2\ell'(\bar{a}) = D_{\bar{a}}(\bar{a}) - \mathbb{E}_{\tilde{a}}\big[D_{\bar{a}}(\tilde{a})\big] = \big[\bar{m}(\bar{a}) - \mathbb{E}_{\tilde{a}}\bar{m}(\tilde{a})\big] - \mathbb{E}_{\tilde{a}}\big[\rho_{\bar{a}}(\tilde{a})\big].$$

(ii) (Ripple.) *For all $u, a \in \mathbb{R}$, $|\rho_u(a)| \le \min\{\varsigma,\ |a-u|/(4\beta^2)\}$.*

(iii) (Clip term.)

$$\begin{aligned}\big|\bar{m}(\bar{a}) - \mathbb{E}_{\tilde{a}}\bar{m}(\tilde{a})\big| &\le \mathbb{E}_{\tilde{a}}\Big|\int_{\bar{a}}^{\tilde{a}}\big(1 - v_+(d(s)/\sigma)\big)\,ds\Big|\\ &\le \big(1 - v(c_\delta)\big)\,\mathbb{E}|\tilde{a} - \bar{a}| + \mathbb{E}\big[(|\tilde{a}| - (W - c_\delta\sigma))_+\big].\end{aligned}$$

(iv) (Soft-clip pullback.) *Let $\tilde{g} := \bar{m}^{-1}\big(\mathbb{E}_{\tilde{a}}\bar{m}(\tilde{a})\big)$, well defined because $\bar{m}$ is a strictly increasing bijection of $\mathbb{R}$ onto $(-W, W)$. If $\tilde{g} \in I_\delta$, then*

$$|g^\star - \tilde{g}| \le \frac{\kappa_{\mathrm{HL}}}{2\sigma^2}\,\big|\mathbb{E}_{\tilde{a}}\rho_{\tilde{g}}(\tilde{a})\big| \le \frac{\kappa_{\mathrm{HL}}}{2\sigma^2}\,\min\{\varsigma,\ \mathbb{E}|\tilde{a} - \tilde{g}|/(4\beta^2)\}.$$

Consequently,

$$\boxed{|g^\star - \bar{a}| \le \frac{\kappa_{\mathrm{HL}}}{2\sigma^2}\Big\{\mathbb{E}_{\tilde{a}}\Big|\int_{\bar{a}}^{\tilde{a}} \big(1 - v_+(d(s)/\sigma)\big)\, ds\Big| + \big|\mathbb{E}_{\tilde{a}}\rho_{\bar{a}}(\tilde{a})\big|\Big\}, \qquad \frac{\kappa_{\mathrm{HL}}}{2\sigma^2} = \frac{1}{v(c_\delta) - 1/(4\beta^2)}.}$$

Proof of Theorem 6. **Step 1 (sandwich).** By Corollary A.3 applied with $q = \bar{q}$, $\ell'' \ge \alpha_{\mathrm{HL}}$ on I_δ and $|\ell''| \le 1/\sigma^2$ on $\mathbb{R}$. Fix $g \in I_\delta$. Minimality of $g^\star$ over the interval I_δ gives $\ell'(g^\star)(g - g^\star) \ge 0$, and integrating $\ell'' \ge \alpha_{\mathrm{HL}}$ along the segment between $g^\star$ and g, which lies in I_δ, gives $(\ell'(g) - \ell'(g^\star))(g - g^\star) \ge \alpha_{\mathrm{HL}}(g - g^\star)^2$. Adding the two inequalities,

$$\ell'(g)\,(g - g^\star) \ge \alpha_{\mathrm{HL}}\,(g - g^\star)^2, \qquad \text{hence} \qquad |g - g^\star| \le \frac{|\ell'(g)|}{\alpha_{\mathrm{HL}}} = \frac{\kappa_{\mathrm{HL}}}{2}\,|\ell'(g)|,$$

valid for every $g \in I_\delta$; $g = \bar{a}$ is the upper bound in (i). If $g^\star$ is interior, $\ell'(g^\star) = 0$, and the global curvature bound gives $|\ell'(\bar{a})| = |\ell'(\bar{a}) - \ell'(g^\star)| \le |\bar{a} - g^\star|/\sigma^2$, the lower bound.

Step 2 (the gradient is a Jensen gap). Lemma A.1 with $q = \bar{q}$ gives, for every $u \in \mathbb{R}$, $\sigma^2\ell'(u) = \bar{m}(u) - \sum_i \bar{q}_i\, m_i(u)$. Since $\bar{q}_i = \mathbb{E}_{\tilde{a}}[\Gamma_\sigma(\tilde{a})_i]$ and the sum is finite, $\sum_i \bar{q}_i\, m_i(u) = \mathbb{E}_{\tilde{a}}[D_u(\tilde{a})]$, while $D_u(u) = \bar{m}(u)$; the second form in (i) is $D_u = \bar{m} + \rho_u$ at $u = \bar{a}$.

Step 3 (ripple). By the exponential-family calculus of Lemma A.1, $m_i'(s) = V_i(s)/\sigma^2$, and Lemma A.2 gives $0 \le V_i \le \varsigma^2/4$; hence $0 \le m_i' \le \varsigma^2/(4\sigma^2) = 1/(4\beta^2)$ and $|m_i(u) - m_i(a)| \le |a - u|/(4\beta^2)$. Also $m_i(\cdot) \in [e_{i-1}, e_i]$, so $|m_i(u) - m_i(a)| \le \varsigma$. Averaging against the weights $\Gamma_\sigma(a)_i$ proves (ii).

Step 4 (clip term). Put $c(s) := s - \bar{m}(s)$, so that $c'(s) = 1 - V_I(s)/\sigma^2 \in [0, 1]$ by Lemma A.1 and the upper half of Lemma A.2. The lower half of Lemma A.2 uses only that s lies at distance at least $c_\delta\sigma$ from each edge; applied with the margin $d(s)/\sigma$ in place of c_δ it gives $V_I(s) \ge \sigma^2 v(d(s)/\sigma)$ for $|s| < W$, and $V_I \ge 0$ always, so $c'(s) \le 1 - v_+(d(s)/\sigma)$ for every $s \in \mathbb{R}$. Because $\mathbb{E}\tilde{a} = \bar{a}$, $\bar{m}(\bar{a}) - \mathbb{E}_{\tilde{a}}\bar{m}(\tilde{a}) = \mathbb{E}_{\tilde{a}}[c(\tilde{a}) - c(\bar{a})]$, and $|c(\tilde{a}) - c(\bar{a})| = |\int_{\bar{a}}^{\tilde{a}} c'| \le |\int_{\bar{a}}^{\tilde{a}} (1 - v_+(d(s)/\sigma))\, ds|$, the first inequality of (iii). For the second, the segment from $\bar{a} \in I_\delta$ to $\tilde{a}$ meets I_δ in a set of length at most $|\tilde{a} - \bar{a}|$, on which $d(s) \ge c_\delta\sigma$ and hence, v being increasing, $v_+(d(s)/\sigma) \ge v(c_\delta)$; it meets the complement of I_δ in a set of length $(|\tilde{a}| - (W - c_\delta\sigma))_+$, on which the integrand is at most 1.

Step 5 (pullback). Step 2 at the reference point $u = \tilde{g}$ (Lemma A.1 does not use $\mathbb{E}\tilde{a} = u$) gives $\sigma^2\ell'(\tilde{g}) = \bar{m}(\tilde{g}) - \mathbb{E}_{\tilde{a}}[\bar{m}(\tilde{a}) + \rho_{\tilde{g}}(\tilde{a})] = -\mathbb{E}_{\tilde{a}}\rho_{\tilde{g}}(\tilde{a})$, since $\bar{m}(\tilde{g}) = \mathbb{E}_{\tilde{a}}\bar{m}(\tilde{a})$. Apply the sandwich of Step 1 with $g = \tilde{g} \in I_\delta$, then (ii). The boxed inequality combines (i)–(iii). □

Corollary F.1 (Sign of the clip term under one-sided target laws)**.** *Suppose that $V_I(s)$ is nonincreasing in $|s|$ on $\mathbb{R}$ (property (M); equivalently, $\bar{m}$ is concave on $[0, \infty)$ and, being odd, convex on $(-\infty, 0]$). If $\tilde{a} \ge 0$ almost surely and $\bar{a} \in I_\delta$, then $\mathbb{E}_{\tilde{a}}\bar{m}(\tilde{a}) \le \bar{m}(\bar{a})$, so the clip contribution to $\sigma^2\ell'(\bar{a})$ is nonnegative, $\tilde{g} \in [0, \bar{a}]$, and*

$$g^\star \le \bar{a} + \frac{\kappa_{\mathrm{HL}}}{2\sigma^2}\min\{\varsigma,\ \mathbb{E}|\tilde{a} - \tilde{g}|/(4\beta^2)\}:$$

the soft clip displaces the Bayes scalar toward the origin for every target law supported on one side of the origin, symmetric or not, up to the ripple. The mirror statement holds for $\tilde{a} \le 0$.

Proof. Under (M), $\bar{m}' = V_I/\sigma^2$ is nonincreasing on $[0, \infty)$, so $\bar{m}$ is concave there and Jensen's inequality gives $\mathbb{E}_{\tilde{a}}\bar{m}(\tilde{a}) \le \bar{m}(\mathbb{E}\tilde{a}) = \bar{m}(\bar{a})$ for $\tilde{a} \ge 0$. Since $\bar{m}$ is increasing and $\bar{m}(0) = 0 \le \bar{m}(\tilde{a})$, this reads $0 \le \tilde{g} \le \bar{a}$, so $\tilde{g} \in I_\delta$, and Theorem 6(iv) gives the displayed bound. □

Corollary F.2 (Noisy-target transfer)**.** *Let Assumptions 1 and 3 and the support design hold, and suppose that at every round the realized target is $\tilde{a}_t = a_t + \xi_t$ with $\mathbb{E}[\xi_t \mid x, y, y'] = 0$, so that $\bar{a}_t = a_t \in I_\delta$ for every pair. Let $g_t^\star$ be the pairwise Bayes scalar of Proposition 2 and $B_t^2 := \mathbb{E}_{\mathcal{D}_t}[(g_t^\star - a_t)^2]$, which Theorem 6 bounds pair by pair. If the fitted scalar satisfies $g_t \in I_\delta$ and $\mathbb{E}_{\mathcal{D}_t}[\mathrm{excess}_{\mathrm{CE}}(g_t)] \le \varepsilon_t^{\mathrm{mix}}$, the excess population cross-entropy risk over the mixture target, then*

$$\mathbb{E}_{\mathcal{D}_t}(g_t - a_t)^2 \le 2\kappa_{\mathrm{HL}}\,\varepsilon_t^{\mathrm{mix}} + 2B_t^2,$$

and the conclusion of Theorem 2 holds with $\kappa_{\mathrm{HL}}\varepsilon_t^{\mathrm{CE}}$ replaced by $2\kappa_{\mathrm{HL}}\varepsilon_t^{\mathrm{mix}} + 2B_t^2$. If instead $\mathbb{E}[\xi_t \mid x, y, y'] =: \Delta_t \ne 0$, the same statements hold with a_t replaced by $\bar{a}_t = a_t + \Delta_t$, and against the clean gap $\mathbb{E}_{\mathcal{D}_t}(g_t - a_t)^2 \le 2\,\mathbb{E}_{\mathcal{D}_t}(g_t - \bar{a}_t)^2 + 2\,\mathbb{E}_{\mathcal{D}_t}\Delta_t^2$, where the Δ_t term is common to every loss.

Table 3: Loss dependence of each step in the proof of Theorem 2.

Step	Mathematical content	Depends on the training loss?
G-1	CE risk implies pairwise MSE through Theorem 1	Yes: this is the HL-Gauss-specific step.
G-2	Exact conditional variance decomposition in Lemma B.1	No.
G-3	$\pi_{t+1} \propto \pi_t e^{\eta \widetilde{A}_t}$ on the common support	No; follows from the log-ratio definition and support preservation.
G-4	KL telescope and Hoeffding bound in Lemma B.2	No; uses only $\lvert\widetilde{A}_t\rvert \le A_t$.
G-5	Contextual performance-difference identity	No.
G-6	Change of measure from μ_t to $\pi^\star$	No; uses only $C_t(\pi^\star)$.

Proof. Proposition 2 holds for every pair; take $\mathbb{E}_{\mathcal{D}_t}$. The proof of Theorem 2 uses the training loss only through the pairwise squared error of Corollary 1 (step G-1 of Table 3); substituting the displayed bound there and repeating Appendix B verbatim proves the second claim. For biased noise, $(g_t - a_t)^2 \le 2(g_t - \bar{a}_t)^2 + 2\Delta_t^2$ pointwise. □

Remark 2. *Reading.* Up to the ripple, the mixture bias is the Jensen gap of the soft clip $\bar{m}$ under the target law. The map $\bar{m}$ is 1-Lipschitz with slope $V_I/\sigma^2 \ge v_+(d(s)/\sigma)$: it is the identity to within $1 - v(t)$ at distance $t\sigma$ from an edge (3×10^{-2} at $t = 3$, 10^{-3} at $t = 4$) and saturates at $\pm W$. The clip term is therefore controlled by the expected excursion of targets beyond I_δ, the first-moment version of the occupancy statistic of Figure 3; under proportional widening $(W, \varsigma, \sigma) \to k(W, \varsigma, \sigma)$ the excursion of a fixed noise law decays like the Gaussian tail of its widest component, which is the $\approx 200\times$ collapse of Table 15. The ripple is first order in ς and does not cancel for skewed laws; it is the $\approx +0.001$ residual of the one-sided law at $k{=}3$. *Tightness and scope.* The lower bound in (i) shows the sandwich is tight up to the factor $\kappa_{\mathrm{HL}}/(2\sigma^2) = 1.90$ at the frozen (β, c_δ); the realized values are in Table 16. Nothing here bounds the ripple to second order for symmetric laws, and property (M) is verified numerically (Appendix H.5), not proved—nor is it a consequence of log-concavity alone: for the Laplace base $e^{-|y|}$ the tilted variance $2(1 + \theta^2)/(1 - \theta^2)^2$ increases in $|\theta|$, so a proof must use the Gaussian curvature of the truncated family; the no-general-sign-theorem statement of Section 4 stands, the one-sided law at $k{=}3$ being a case where the ripple outweighs the clip and reverses the sign.

G Proof dependency audit

The proof of Theorem 2 uses the ingredients listed in Table 3.

No least-squares normal equation or projection property enters the main performance proof. Squared-loss REBEL obtains the pairwise MSE directly from its regression assumption; HL-REBEL obtains the same quantity through Theorem 1. The exact decomposition controls the comparator second moment with coefficient one, avoiding constants introduced by repeated applications of $(a + b)^2 \le 2a^2 + 2b^2$.

H Experimental platform, protocol registry, and full per-arm records

This appendix collects the material summarized in the main text: the evidence status of every claim class (H.1); the synthetic platform as executed—environment, streams, contamination models, potential fit, losses, optimizer, support calibration, evaluation, and the loop (H.2); the full remarks on the calibration constant (H.3); the registration record and the disclosed exploration (H.4); the secondary records and controls, organized by the scientific question each answers (H.5); the engineering-validity table and diagnostics (H.6); the per-seed, dose, and mechanism figures (H.7); and the traceability summary (H.8). The registration timeline, the seed index sets, the deterministic seed-0 checkpoint, and the pinned digests are in the process record released with the repository.

H.1 EVIDENCE-STATUS REGISTRY

Every claim in this paper reads against Table 4, which records for each claim class its evidence type and supported scope.

H.2 THE SYNTHETIC PLATFORM: DEFINITIONS AS EXECUTED

This subsection is the reference definition of the synthetic platform. It is written from the frozen harness that produced every synthetic number in this paper (`scripts/s1_run_synthetic.py` at its manifest-pinned digest; the E1 chain ran the same harness at its earlier pinned digest, H.2.2) and is organized so that a reader can construct one training round—design matrix, targets, fitted increment, updated policy—and the reported readout from the text alone, without consulting the process records of H.8. The sampling law, the three losses, the optimizer, and the evaluation are shared by every cell; H.2.2 states what each contamination model changes, and Table 5 what persists across rounds.

Summary (the registered description). The platform is a log-linear contextual bandit with $C = 20$ contexts, $K = 100$ actions, and $d = 32$ features, with an exact tabular mirror-descent oracle as the evaluation benchmark, step size $\eta = 1.0$, and full-batch Adam (Kingma & Ba, 2015) with identical optimizer settings across three losses: MSE (artifact name `mse`); a curvature-matched Huber loss, `2×huber` with $\delta_H = W/2$ (twice the standard Huber loss, so that its quadratic-region curvature equals that of the squared loss—the Hessian-scale gate of Appendix H.6 confirms the match); and HL, the tied HL-Gauss cross-entropy (`hl_tied`), with $m = 51$ bins and $\beta = \sigma/\varsigma = 0.75$. Support and scale parameters are frozen from a 32768-pair clean warmup under a precommitted rule: half-width $W = 0.91543$ (recorded as V in the artifacts), bin width $\varsigma = 0.035899$, $\sigma = 0.026924$, Huber $\delta_H = W/2 = 0.45771$, contamination base scale $\tau = 0.3 \cdot \mathrm{std} = 0.076166$, and the frozen loss-scale constant $S := 2/\operatorname{median} \ell''$, the median being taken over the clean-warmup bulk $|a| \le W - 3\sigma$ of the fp64 central-difference curvature of the unscaled tied cross-entropy at $g = a$; the precommitted rule forbids the analytic shortcut $S = 2\sigma^2$, and the realized value is $S = 0.0016645 = 2.296\sigma^2$. Step size $\eta = 1.0$; the primary contrast $d_s = \mathrm{perf}(\mathrm{HL}) - \mathrm{perf}(\mathrm{MSE})$ is paired per seed through exogenous common random number streams: contexts, actions, and noise draws come from fixed streams that do not depend on the current policy, so every loss sees the identical design matrices and targets in every round (common streams here mean common training data, not merely common random numbers). Performance is `perf_final`, the negative regret of the uniform round mixture of the post-update iterates—the average over the 15 rounds of the policy value, which obeys Theorem 2's bound through the post-update remark of Appendix B.4 (higher is better); the registration's primary-outcome field reads, verbatim, `perf_final` $= -(J(\pi^\star) - J(\bar{\pi}_T))$, averaged over the 20 contexts, $T = 15$—the uniform-mixture readout has been the registered endpoint since the first registration; per-round values additionally yield the final-iterate and best-iterate readouts of the optimizer-precision control.

Algorithm 1 (HL-REBEL)**.** *Require:* reward bound $R_{\max}$; policy class Π; base distributions $\{\mu_t\}_{t<T}$; step size $\eta > 0$; design bound A_{des}; ratio $\beta = \sigma/\varsigma$ and margin c_δ with $v(c_\delta) > 1/(4\beta^2)$; support half-width $W \ge \max\{2R_{\max}, 2A_{\mathrm{des}}\} + c_\delta\sigma$.

1. Initialize $\pi_0 \in \Pi$.
2. For $t = 0, \ldots, T-1$: collect $\widehat{\mathcal{D}}_t = \{(x, y, y')\}$ with $x \sim \rho$, $y \sim \pi_t(\cdot \mid x)$, $y' \sim \mu_t(\cdot \mid x)$; for each triple set $a \leftarrow r(x, y) - r(x, y')$ and encode $q \leftarrow \Gamma_\sigma(a)$; compute $\theta_{t+1} \approx \arg\min_\theta \widehat{L}_t(\theta)$ and set $\pi_{t+1} = \pi_{\theta_{t+1}}$.
3. Return the uniform mixture $\bar{\pi}_T = T^{-1} \sum_{t<T} \pi_t$, or any iterate with value at least $J(\bar{\pi}_T)$ (Theorem 2 guarantees one exists).

Notation guide $\beta = \sigma/\varsigma$ is the bin-smoothing ratio, distinct from the Tsybakov margin exponent ν (Theorem D.2). $\delta = c_\delta\sigma$ is the absolute support margin with preregistered multiplier $c_\delta = 3$, distinct from the Huber threshold δ_H. a_t is the realized relative-reward target ($|a_t| \le 2R_{\max}$; noise enters additively as $a_t + \xi$, Section 4); it is distinct from Assumption 3's chain $|\widetilde{A}_t| \le A_t \le A_{\mathrm{des}}$—the centered update, its per-round bound, and the design bound. Table 28 reports realized values of $\widetilde{A}_t$.

Random streams. Every random object is drawn from a stateless keyed stream: the key is a string `namespace|part|...|part`, its seed is the first 8 bytes of SHA-256(key) read as a big-endian

Table 4: **Evidence status of every claim class**, with the question each leaves open. The E2 benefit carries the two qualifications of Section 5.3 (interpolation regime only; not separated from curvature-matched Huber).

Claim	Status	Supported scope	Left open
Two-sided calibration; REBEL transfer; noisy-target transfer; lower bounds; offset exclusion (Thms 1–3, Cor. C.1, Cor. F.2)	Theorem	under the stated assumption class; noisy targets may leave I_δ (Cor. F.2)	constants outside the assumption class; Cor. F.2 takes the excess risk over the mixture target as a supplied budget, as Assumption 2 does
Mixture decomposition; two-sided mixture-bias bound; conditional variance-weighted corner (Prop. 2, Thms 5–6)	Theorem	$\lvert g^\star - \bar{a}\rvert$ sandwiched within the factor $\kappa_{\mathrm{HL}}/(2\sigma^2)$ by the soft-clip Jensen gap, clip term bounded by the expected excursion beyond I_δ (F.5); in-family corner excludes the term from scope	second-order control of the ripple term; property (M) verified numerically, not proved
E1 symmetric deficit	Preregistered confirmatory (REFUTED verdict)	tested synthetic averaging regime	boundary intervention run at both doses: the deficit collapses $60\times$ (rate 0.05) and $75\times$ (rate 0.10) at $k{=}3$ (H.5.3)
E2 benefit	Preregistered + fresh-seed replication	within-generator; interpolation regime; not separated from Huber (fixed data, $n{\le}80$)	sign reproduced on eight fresh environment instances ($8/8$) and under policy-to-sampling feedback (160 seeds; H.5.1); high-precision limit unresolved
E3 control	Descriptive, protocol frozen first	mechanism evidence, not standalone attribution	magnitude does not extend at $n{=}400$, and the deficit is instance-specific: $7/8$ fresh environment instances give a non-negative contrast (sign nominal)
Five-branch ablation suite (W, β, N, R, F)	Descriptive; protocol and branches fixed before execution (post-closure)	narrows the mechanism; no confirmatory generalization	$n{=}40$ precision per branch
Boundary attribution of both contamination effects	Descriptive; interpretive branches frozen before execution	moving the boundary alone removes the 32-pair effects and collapses the E1 deficit 60–$75\times$; the clean ripple is interior	clip location and hardness not separated (anatomy arms move both); the finite-step Adam path behind the benefit
Factorial statistics; optimizer-precision control; occupancy assembly; RM bridge; pure-boundary ablation; bound instantiation; persistent-trunk MLP; untied-head cells; powered factorial	Post-closure descriptive controls and instruments (audit-side authored)	robustness, mechanism-separation, design-space, and diagnostic evidence	within one platform; descriptive
MLP check	Descriptive boundary check	partial transfer: tax robust, benefit width-dependent	width dependence unexplained; persistent representation changes the sign structure
LLM check	Descriptive pilot ($n = 3$ seeds)	no LLM benefit claim; pilot clean-tax direction not replicated by the powered round	$n{=}3$ seeds
Powered LLM extension (H-P1–P3)	Extension-registered confirmatory (amended prereg.; Holm+TOST)	HL–Huber statistically equivalent within ± 0.035; HL$-$MSE and clean tax inconclusive at the achieved precision	H-P1 and H-P3 at higher precision; natural (uninjected) reward-model noise
On-policy MLP test (H1–H3)	Extension-registered confirmatory (Holm; maintenance r1 before the rerun)	symmetric law, averaging regime, $\mu_t = \pi_t$, persistent-trunk MLP: the deficit, its $k{=}3$ attenuation, and HL below residual-space Huber (H.5.7)	one environment instance and one noise family; the fixed-data benefit is unresolved under feedback at 32 pairs; block-B readout on MLP arms confounded by off-distribution evaluation
Pure-widening follow-up (H2′)	Extension-registered confirmatory (one hypothesis)	the $k{=}3$ attenuation reproduced with bin width, σ, and Huber threshold fixed ($D' = -0.00216$, $p = 1.3 \times 10^{-24}$; H.5.7)	one environment instance and one noise family; the top two rungs of the ladder are not separated
Influence-cap mechanism	Closed-form companion, descriptive	M-B: not supported—the prespecified leverage prediction is not met (CI $[-0.0035, +0.0067]$); Adam-path mechanism open	the Adam-path mechanism inside the row space of the design: the null-space share of ω_t does not carry the benefit at $n{=}32$ and carries about half of the point estimate at $n < d$ (H.5)

integer, and the generator is NumPy's Philox. No key contains the loss, the cell, or the current policy, so the same key is replayed identically by every arm; the stream identities are pinned by the digests of H.8.

H.2.1 Environment and true reward (drawn once)

Features $\varphi(c,k) \in \mathbb{R}^d$ for $c \in \{1,\dots,C\}$ and $k \in \{1,\dots,K\}$, $(C,K,d) = (20,100,32)$, have i.i.d. $\mathcal{N}(0,1/d)$ entries (a standard normal array divided by $\sqrt{d}$), drawn once from the key `HLREBEL_SYNTH_ENV_v1|20260802`; from the same stream, $\theta^\star \sim \mathcal{N}(0,I_d)$ is normalized to $\|\theta^\star\|_2 = 1$. The true reward is the inner product with no further transformation,

$$r^\star(c,k) = \langle \varphi(c,k), \theta^\star \rangle$$

(realized range $[-0.591, 0.669]$ and standard deviation 0.183 over the 2000 context–action pairs; so $R_{\max}$ is of order 0.7 and clean gaps have standard deviation 0.254). The policy class is log-linear, $\pi_\theta(k \mid c) \propto \exp\langle \theta, \varphi(c,k)\rangle$ with $\theta \in \mathbb{R}^d$, initialized at the uniform policy ($\theta_0 = 0$); the context law ρ is uniform. The contamination direction comes from the key `HLREBEL_SYNTH_HACKDIR_v1|0`: $h_{\mathrm{raw}} \sim \mathcal{N}(0,I_d)$, $h_\perp$ the unit vector along $h_{\mathrm{raw}} - \langle h_{\mathrm{raw}}, \theta^\star\rangle\theta^\star$, and

$$h = \frac{\cos\alpha\,\theta^\star + \sqrt{1-\cos^2\alpha}\,h_\perp}{\left\|\cos\alpha\,\theta^\star + \sqrt{1-\cos^2\alpha}\,h_\perp\right\|}, \qquad \cos\alpha = -0.7$$

for the registered model (-0.5 and -0.9 in the two sensitivity rows, which share $h_\perp$; the realized cosines equal their targets to 10^{-12}).

What is fixed and what a seed changes. The environment $(\varphi, \theta^\star, r^\star)$, the direction h, the clean warmup, and every constant derived from it ($W, \varsigma, \sigma, \delta_H, \tau, S$; H.2.6) are generated once and are identical in every seed, cell, arm, and chain of this record (the E1 chain's metrics file carries the same constants to every printed digit). A seed indexes only the training data of H.2.2: one 15-round realization of contexts, action pairs, and endpoint noise. Every confidence interval and p-value in the synthetic record is therefore a statement about training-data randomness within one fixed environment instance—the "within-generator" qualifier used throughout—and no variation across environment instances enters them. A descriptive resampling of the instance itself—eight fresh draws of $\varphi, \theta^\star, h$ and their derived constants, training streams held fixed—is reported in H.5: the E2 sign reproduces in $8/8$ instances and the registered instance ranks fourth of nine.

H.2.2 Training pairs and the contamination models

Round $t \in \{1,\dots,T\}$, $T = 15$, of seed s uses n pairs ($n = 32$ in the registered cells; $n = 64$ in the sensitivity row, whose first 32 pairs are the $n = 32$ pairs; $n \in \{16,\dots,64\}$ in the regime sweep, prefixes likewise). Pair $i \in \{0,\dots,n-1\}$ draws, from the key `HLREBEL_SYNTH_DATA_v2|s|t|i` in this order, a context $c_i \sim U\{1,\dots,C\}$ and two actions $k_i, k_i' \sim U\{1,\dots,K\}$, independently of each other and of everything else: both endpoints come from the uniform law, so the sampling law is $Q = \rho \otimes U \otimes U$ in every round and under every policy, and $k_i = k_i'$ is allowed (probability $1/K$; realized 1.1% of pairs), in which case the design row is zero. Indices are drawn as 0-based integers (`integers(0,C)`, then `integers(0,K)` twice, in this order); the sets $\{1,\dots,C\}$ and $\{1,\dots,K\}$ above are the same objects relabeled from 1. At $n = 32$ the design matrix is rank-deficient on 182 of the 600 registered rounds (40 seeds $\times$ 15 rounds; H.5). The design row and the clean gap are

$$\Delta\varphi_i = \varphi(c_i,k_i) - \varphi(c_i,k_i'), \qquad a_i = r^\star(c_i,k_i) - r^\star(c_i,k_i'),$$

and $X_t \in \mathbb{R}^{n\times d}$ is the matrix with rows $\Delta\varphi_i^\top$ (stored in single precision). Write $a_{i,0} := k_i$ and $a_{i,1} := k_i'$ for the two endpoint actions. For each endpoint position $e \in \{0,1\}$ the key `HLREBEL_SYNTH_NOISE_v2|s|t|i|e` yields $u_{i,e} \sim U(0,1)$ and then $z_{i,e} \sim \mathcal{N}(0,1)$; the key carries the position, not the action index, so the two endpoints of a pair always draw independent noise, including the $k_i = k_i'$ pairs, whose zero design row then carries a nonzero target. The training target is $\tilde{a}_i = a_i + \xi_{i,0} - \xi_{i,1}$, with endpoint noise

$$\text{clean cells:} \quad \xi_{i,e} \equiv 0 \quad \text{(targets are the exact gaps; no noise stream is read);}$$

$$\text{E2 model (registered):} \quad \xi_{i,e} = \begin{cases} |z_{i,e}| \cdot 10\tau, & u_{i,e} < \varrho \text{ and } \langle \varphi(c_i,a_{i,e}), h\rangle > 0, \\ z_{i,e}\,\tau, & \text{otherwise;} \end{cases}$$

E3 symmetrized model: $\xi_{i,e} = \begin{cases} z_{i,e} \cdot 10\tau, & u_{i,e} < 0.05, \\ z_{i,e}\,\tau, & \text{otherwise}, \end{cases}$

with $\varrho = 0.05$ (0.10 in the doubled-rate row, which the frozen protocols label "10× dose" after the value 0.10: a doubling of the mark rate at unchanged spike amplitude 10τ), the symmetrized model reading its own stream (namespace `HLREBEL_SYNTH_NOISE_N1_v1`). Every endpoint of a contaminated cell thus carries base noise of scale τ; a marked endpoint of the E2 model that lies on the positive side of h carries instead a positive spike of scale 10τ; the symmetrized model keeps the mark, drops the alignment gate, and makes the spike signed. Implied rates for the E2 model: endpoint mark ϱ, endpoint spike $\varrho \Pr(\langle \varphi, h \rangle > 0) \approx 2.4\%$ (realized 2.46%), pair-level spike (either endpoint) $\approx 4.9\%$; for the symmetrized model the pair-level spike rate is $1 - 0.95^2 = 9.75\%$. The E1 chain (registration v1) used the same environment and constants with 20 seeds, $n = 512$, streams keyed per round (`HLREBEL_SYNTH_DATA_v1|s|t` for the block of contexts and actions, `HLREBEL_SYNTH_NOISE_v1|cell|s|t` for the block of (u, z)), and the symmetrized law above with $\varrho \in \{0.05, 0.10\}$: the E1 and E3 noise laws are the same mixture, $\xi \sim 0.95\,\mathcal{N}(0, \tau^2) + 0.05\,\mathcal{N}(0, (10\tau)^2)$ per endpoint. Within a seed the data $(X_t, \tilde{a}_t)_{t \le T}$ are identical for the three losses and for every arm that shares the cell (the loss is in no key): "paired" contrasts are paired on identical training data, not merely on common random numbers.

Sampling protocol of the synthetic platform versus Algorithm 1 Algorithm 1 and the theory of Section 3 sample $(x, y, y') \sim \mathcal{D}_t = \rho \otimes \pi_t \otimes \mu_t$. The synthetic platform does not: both endpoints of every pair are drawn from a fixed behavior law $Q_t \equiv Q = \rho \otimes U \otimes U$ (ρ uniform over the 20 contexts, U uniform over the 100 actions), read from exogenous streams that do not depend on the current policy, with noise drawn per endpoint; consequently, in the contaminated cells, the $\approx 1.1\%$ of pairs whose endpoints coincide carry a nonzero target on a zero design row, are non-interpolable, and add a θ-independent constant to every loss (on clean cells such pairs carry target zero). The platform is therefore a *fixed-data mechanism platform*: it isolates the loss and optimizer axis on identical data, at the price of removing the feedback by which a loss changes the policy and the policy changes the next round's data; it is not an implementation of the on-policy protocol of Algorithm 1, and the LLM stack of Appendix H.5.6 differs from it (both endpoints there are sampled from the current policy, $\mu_t = \pi_t$). Theorem 2 conditions on the population risk under $\mathcal{D}_t$ and is indifferent to how the fit was obtained, but a risk measured under Q is a different object; since $\mathrm{d}\mathcal{D}_t / \mathrm{d}Q = K \pi_t(y \mid x) \le K = 100$ on this platform, $\mathbb{E}_{\mathcal{D}_t} \mathrm{KL} \le \Lambda_t \, \mathbb{E}_Q \mathrm{KL}$ with $\Lambda_t := \operatorname{ess\,sup}_{x,y} K \pi_t(y \mid x)$, so the hypothesis of Theorem 2 is implied by the Q-risk up to this factor *when both are evaluated against the same clean target*; this change of measure keeps the loss's target fixed, and converting the noisy-target risk of a contaminated cell into the clean-gap risk that Theorem 2 needs requires the separate bias control of Section 4, not a factor. Training-pair averages such as $\hat{\varepsilon}^{\,\mathrm{CE}}$ in Table 28 are empirical estimates of the Q-risk, not of the $\mathcal{D}_t$-risk. An on-policy sensitivity run closes this loop (H.5, EXT-ONPOLICY): with both endpoints drawn from the current policy on 160 seeds, the E2 sign survives ($+0.0025$, one-sided $p = 0.012$), at the larger variance that non-identical data across losses imply.

H.2.3 The potential fit and the policy update

Each round fits an *increment* $\omega_t \in \mathbb{R}^d$ of the policy parameter, not the parameter itself. With the linear potential $u_\omega(c, k) := \langle \varphi(c, k), \omega \rangle$, the prediction for pair i is

$$g_i(\omega) = u_\omega(c_i, k_i) - u_\omega(c_i, k_i') = \langle \Delta\varphi_i, \omega \rangle = (X_t \omega)_i;$$

ω is initialized at 0, optimized against $\tilde{a}_t$ for the fixed budget of H.2.5, and the policy is updated by $\theta_t = \theta_{t-1} + \omega_t$, $\pi_t = \pi_{\theta_t}$. Because the class is log-linear, this is exactly the multiplicative update of Algorithm 1 with $\eta = 1$:

$$\pi_t(k \mid c) = \frac{\pi_{t-1}(k \mid c)\, e^{u_{\omega_t}(c,k)}}{\sum_j \pi_{t-1}(j \mid c)\, e^{u_{\omega_t}(c,j)}},$$

and therefore

$$\frac{1}{\eta} \log \frac{\pi_t(k \mid c)}{\pi_{t-1}(k \mid c)} - \frac{1}{\eta} \log \frac{\pi_t(k' \mid c)}{\pi_{t-1}(k' \mid c)} = \langle \varphi(c, k) - \varphi(c, k'), \omega_t \rangle,$$

so the fitted scalar $g_i(\omega)$ *is* the policy log-ratio difference $g_{\theta,t}$ of Section 2: REBEL's $f_{\theta,t}$ equals u_{ω_t} up to the context-wise constant $-\eta^{-1} \log Z_t(c)$, which the pairwise difference removes. There is no separate policy parameterization and no projection step. Table 5 lists what persists across rounds and what is reset.

Table 5: **Persistence and reset rules of the synthetic loop** (as executed; identical for the three losses).

Object	Rule
Environment $\varphi, \theta^\star, r^\star, h$; clean warmup; $W, \varsigma, \sigma, \delta_H, \tau, S$	generated once; shared by every seed, cell, arm, and chain
Training pairs $(X_t, \tilde{a}_t)$	fresh pairs every round from the fixed law Q (no round reuses data; no dependence on π_{t-1}); identical across the three losses within a seed
Fitted increment ω_t	initialized at 0 every round; 60 full-batch Adam steps; never carried to the next round
Adam state (moments, step counter)	a new optimizer instance every round; identical hyperparameters for all losses
Policy parameter θ_t	$\theta_0 = 0$ (uniform policy); $\theta_t = \theta_{t-1} + \omega_t$ accumulates over the 15 rounds
Readout	$J(\pi_t)$ after every update; `perf_final` averages $t = 1, \ldots, 15$ (H.2.7)

H.2.4 Losses and their numerical implementation

With bin edges $e_j := -W + j\varsigma$ ($j = 0, \ldots, m$), the map of Section 2 is

$$\Gamma_\sigma(s)_j = \frac{\Phi\big((e_j - s)/\sigma\big) - \Phi\big((e_{j-1} - s)/\sigma\big)}{\Phi\big((W - s)/\sigma\big) - \Phi\big((-W - s)/\sigma\big)}, \qquad j = 1, \ldots, m,$$

applied unchanged to any real s: a target or prediction beyond $\pm W$ is neither clipped nor projected; only its truncated mass is renormalized. The per-pair losses are

$$\ell_{\mathrm{MSE}}(g, \tilde{a}) = (g - \tilde{a})^2,$$

$$\ell_{\mathrm{Huber}}(g, \tilde{a}) = 2H_{\delta_H}(g - \tilde{a}), \qquad H_\delta(e) = \begin{cases} \frac{1}{2}e^2, & |e| \le \delta, \\ \delta\big(|e| - \frac{\delta}{2}\big), & |e| > \delta, \end{cases}$$

$$\ell_{\mathrm{HL}}(g, \tilde{a}) = -S \sum_{j=1}^{m} \Gamma_\sigma(\tilde{a})_j \log \Gamma_\sigma(g)_j,$$

with $\delta_H = W/2$ and the frozen scale S of H.2.6 (the factor 2 makes the Huber quadratic region coincide with the squared loss). The target side of ℓ_{HL} is a constant (detached), the prediction side is differentiated through the scalar g alone, and both sides call the same projection routine. The round objective is the *mean* over the n pairs, $F_t(\omega) = \frac{1}{n}\sum_{i=1}^{n} \ell\big((X_t\omega)_i, \tilde{a}_i\big)$; independent seeds are vectorized by summing their per-seed means, which couples neither gradients nor Adam states. Numerics: X_t, $\tilde{a}_t$, ω, and the losses are single precision; $\log \Gamma_\sigma$ is evaluated entirely in the log domain—$\log \Phi$ via `log_ndtr` (the survival form on bins lying above the location), bin masses via `logdiffexp`, renormalization via `logsumexp`—with no probability floor and no clamping, and every projected row is asserted to sum to one within the frozen tolerance; the curvature calibration of S and every evaluation quantity are double precision.

The local geometry of the scalar loss $g \mapsto \mathrm{CE}(\Gamma_\sigma(a), \Gamma_\sigma(g))$—its target-centered expansion, the uniqueness of the minimizer with bounded tail gradients, and the three local quadratic models—is collected in Appendix E.

H.2.5 Optimizer

The inner solver is Adam (Kingma & Ba, 2015) on the full batch of n pairs: 60 steps per round, learning rate 0.05, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 0, no AMSGrad, no gradient clipping, no learning-rate schedule, a new optimizer instance every round, and the same values for the three losses, written into the run manifest before any result existed (the specification forbids per-loss tuning). The optimizer-precision control (H.5) changes this rule and nothing else: 300 or 600 Adam steps with the same constants, or one double-precision L-BFGS solve per round (strong-Wolfe line search, at most 200 iterations, history 50, gradient tolerance 10^{-12}, change tolerance 10^{-16}, ω initialized at 0, θ accumulated in double precision). The E1 chain used the same Adam constants; its

harness averaged the loss over seeds and pairs jointly, a constant factor $1/20$ relative to the per-seed mean that Adam's per-coordinate normalization absorbs up to the ϵ term. The MLP cells re-initialize the network every round and train it by the same rule; the persistent-trunk variant of H.5.6 is the one deliberate exception. What "identical optimizer settings" certifies is stated in H.6 and H.5: matched loss scale and effective step at initialization (the fairness gates) and, at the frozen budget, matched final gradient norms and objective gaps across the three losses; it is not a statement that the three finite-step trajectories coincide.

H.2.6 Support calibration: the complete construction

Warmup. From the key `HLREBEL_SYNTH_WARMUP_v1|20260801`, $N_{\mathrm{w}} = 32768$ triples $(c, k, k') \sim Q$—drawn as three vectorized calls in the order c, then k, then k' (`integers`, 0-based, 32768 values each)—give clean gaps $a^{\mathrm{w}} = r^\star(c,k) - r^\star(c,k')$ (no noise; the round-0 clean sample of the specification). *Upper confidence bound.* UCB is the percentile-bootstrap 95% one-sided upper confidence bound of the 0.999-quantile of $|a^{\mathrm{w}}|$: $B = 10000$ resamples of the 32768 values with replacement (`PCG64`, seed 20260802), the type-7 (linear-interpolation) 0.999-quantile of each resample, and the type-7 0.95-quantile of the B replicate statistics. Realized: point estimate $\widehat{Q}_{99.9} = 0.82584$, UCB $= 0.83465$. *Geometry.* With $m = 51$ and $\beta = \sigma/\varsigma = 0.75$ fixed in advance, the half-width is the solution of $W - 3\sigma = \mathrm{UCB}$ under $\varsigma = 2W/m$ and $\sigma = \beta\varsigma$:

$$W = \frac{\mathrm{UCB}}{1 - 6\beta/m} = \frac{0.83465}{0.91176} = 0.91543, \qquad \varsigma = \frac{2W}{m} = 0.035899, \qquad \sigma = \beta\varsigma = 0.026924,$$

so that $W - 3\sigma = 0.83465 = \mathrm{UCB}$: the interior bound of I_δ at $c_\delta = 3$ is placed at the clean 99.9% upper confidence quantile, and the identity is asserted to $10^{-12}W$ at load time. *Huber and contamination scales.* $\delta_H = W/2 = 0.45771$; $\tau = 0.3 \cdot \mathrm{sd}(a^{\mathrm{w}}) = 0.3 \times 0.25389 = 0.076166$ (population standard deviation of the warmup gaps). *Loss scale.* On the warmup bulk $\{|a^{\mathrm{w}}| \le W - 3\sigma\}$, the curvature $\ell''(a)$ of the unscaled tied cross-entropy at $g = a$ is computed by a double-precision central difference with step $10^{-3}\sigma$, and $S = 2/\operatorname{median} \ell'' = 2/1201.55 = 0.0016645 = 2.296\sigma^2$; the specification disallows the analytic shortcut $S = 2\sigma^2$. *Sharing.* The constants are calibrated once and used unchanged in every cell of both chains; the ablations that change W, m, or β re-derive S by the same rule, and the LLM stack applies the same rule with $m = 101$ to its own warmup (H.5.6). *Design versus realized support.* The theory's support condition is pointwise—every target and prediction inside I_δ (Assumption 3)—whereas the calibration is a clean-sample quantile rule: it uses no information about the contamination law, enforces nothing, and no clipping or projection is applied at any point of training. This is why contaminated targets leave I_δ at the rates below, and why Table 28 reports the design chain of Assumption 3 as exceeded in every cell.

Target occupancy of the compression zone Fraction of realized training targets (per cell, recomputed from the frozen scenario builders) beyond the interior bound $W - 3\sigma$ and beyond the support edge W: clean 0.099% / 0.031%; feature-correlated contaminated (E2 model) 1.41% / 1.09%—but 24.73% of the spiked pairs themselves ($n = 934$); symmetric model 3.27% / 2.70%; E1-chain cells (512 pairs, own support): clean 0.108% / 0.042%, symmetric 3.27% / 2.67%, doubled rate 6.09% / 5.09%. Clean targets essentially never reach the zone, so boundary clamping is an unlikely origin of the clean-arm deficit; the smoothing ablation is consistent with bin-curvature ripple (item 9; Section 5.5). Prediction-side occupancy ships with the ablation-suite records; the occupancy-versus-effect assembly across all 19 recorded cells is Figure 3.

H.2.7 Evaluation, comparators, and statistical units

For any policy, $J(\pi) = \frac{1}{C}\sum_{c=1}^{C}\sum_{k=1}^{K}\pi(k \mid c)\, r^\star(c,k)$ is computed exactly in double precision on the full $C \times K$ table: the evaluation benchmark is tabular and draws no samples. The exact mirror-descent oracle from the uniform policy with $\eta = 1$ is $\pi_t^{\mathrm{MD}}(k \mid c) \propto \exp\{t\, r^\star(c,k)\}$; it never supplies targets and serves only two readouts. *Comparator and primary outcome.* The registered outcome uses $\pi^\star := \pi_T^{\mathrm{MD}}$ with $T = 15$:

$$\texttt{perf_final} = -\Big(J(\pi^\star) - \frac{1}{T}\sum_{t=1}^{T} J(\pi_t)\Big),$$

where π_t is the policy after the t-th update (π_0 is not included): the negative regret of the post-update uniform round mixture $\bar{\pi}_T^{+}$, which differs from the mixture of Theorem 2 by $(J(\pi_T) - J(\pi_0))/T$

and obeys the same bound at horizon $T+1$ (Appendix B.4). Final-iterate ($J(\pi_T)$) and best-iterate ($\max_t J(\pi_t)$) readouts are formed from the same per-round values. *Recovery diagnostic.* $\mathrm{KL}(\pi_1^{\mathrm{MD}} \| \pi_1)$ per context; its median over seeds and contexts on the clean cell is the first-round recovery diagnostic of Proposition 1 (H.6). *Theorem-2 instantiation.* The bound diagnostic of Table 28 instantiates the theorem's comparator as the per-context reward-maximizing action, as the theorem permits any comparator; it is a different object from $\pi^\star$ above and is used nowhere else. *Contrasts and units.* $d_s = \texttt{perf_final}(\mathrm{HL}, s) - \texttt{perf_final}(\mathrm{MSE}, s)$ on identical data; the pairing unit is the seed; paired t over seeds, one-sided as registered for the E2 direction and two-sided elsewhere; 95% t-intervals; the SESOI of a chain is 10% of $|\mathrm{mean}_s\, \texttt{perf_final}(\mathrm{MSE}, \mathrm{clean}, s)|$.

H.2.8 The loop as executed, and a deterministic checkpoint

Algorithm 2 (The synthetic loop as executed)**.** *Input:* environment $(\varphi, r^\star, h)$ and constants $(W, \varsigma, \sigma, \delta_H, \tau, S)$ of H.2.1 and H.2.6; seed s; cell (noise law and n); loss ℓ.

1. $\theta_0 \leftarrow 0$.
2. For $t = 1, \ldots, 15$: read $(c_i, k_i, k_i')_{i<n}$ and (u_e, z_e) from the keyed streams and form $X_t, \tilde{a}_t$ (H.2.2); $\omega \leftarrow 0$; new Adam(0.05; $0.9, 0.999$; 10^{-8}); repeat 60 times: one Adam step on $\nabla_\omega \frac{1}{n}\sum_i \ell\big((X_t\omega)_i, \tilde{a}_i\big)$; $\theta_t \leftarrow \theta_{t-1} + \omega$; record $J(\pi_{\theta_t})$ and $\mathrm{KL}(\pi_t^{\mathrm{MD}} \| \pi_{\theta_t})$.
3. Return $\texttt{perf_final} = \frac{1}{15}\sum_{t=1}^{15} J(\pi_{\theta_t}) - J(\pi_{15}^{\mathrm{MD}})$.

A deterministic checkpoint for seed 0—per-round objective values, gradient norms, increments, and policy masses for both cells and all three losses, reproduced to 8×10^{-9} by a clean-room NumPy implementation written from this appendix alone—is in the process record released with the repository (EXT-CHECKPOINT, H.8).

H.3 Remarks on the calibration constant

Provenance of the headline constant. The frozen pipeline constant $3.8009405\sigma^2$ derives from the six-decimal value $v(3) = 0.970630$; the exact expression gives $3.8009396\sigma^2$, a relative difference of 2.3×10^{-7}. The numerical verification in Appendix H.6 uses the exact, and therefore stricter, value.

First, the admissible corner. At other admissible choices, $\kappa_{\mathrm{HL}}/\sigma^2 = 7.11$ at $(\beta, c_\delta) = (0.75, 2)$, 4.20 at $(1, 2)$, and 2.78 at $(1, 3)$, against 3.80 at the $(0.75, 3)$ configuration used throughout: the interior margin and the smoothing ratio move the constant by small factors, and neither W nor m enters. *Second, no free lunch from the smoothing scale.* The theorem is two-sided, so on I_δ the realized categorical risk $R_t^{\mathrm{CE}} := \mathbb{E}_{\mathcal{D}_t}\, \mathrm{KL}(\Gamma_\sigma(a_t) \| \Gamma_\sigma(g_t))$ and the pairwise squared error are pinned to each other within the factor $\kappa_{\mathrm{HL}}/(2\sigma^2) = (v(c_\delta) - 1/(4\beta^2))^{-1} = 1.90$ at $(0.75, 3)$; a budget $\varepsilon_t^{\mathrm{CE}} \geq R_t^{\mathrm{CE}}$ (Assumption 2) inherits only the upper direction $\mathbb{E}(g_t - a_t)^2 \leq \kappa_{\mathrm{HL}}\varepsilon_t^{\mathrm{CE}}$, so the sandwich is a statement about realized risk, not about an arbitrary budget. Since the pairwise squared error is σ-free, shrinking σ shrinks κ_{HL} but inflates the realized $\varepsilon_t^{\mathrm{CE}}$ in proportion: the product $\kappa_{\mathrm{HL}}\varepsilon_t^{\mathrm{CE}}$ that enters Theorem 2 cannot be tuned away, which is why Appendix H.6 reports *tightness*—the realized fraction of the bound—rather than the bound alone. *Third, the route matters.* The alternative mean-difference argument through total variation and Pinsker's inequality yields the constant $R_I^2/(2\lambda^2)$, with $R_I = 2W$ the support width and $\lambda = \inf_{s \in I_\delta} M_\sigma'(s)$ the worst-case slope of the decoded mean (A.3): it scales with the squared support width—looser than κ_{HL} by $\Theta(m^2/\beta^2)$—and its only closed-form lower bound on λ degrades like $e^{-R_I^2/\sigma^2}$. Nothing in this paper uses it; the strong-convexity route is what makes the constant support-independent.

H.4 Registration timeline and protocol registry

Section 5.1 summarizes the registration record: registration v1 (refuted), the six-configuration exploration (Table 6), the amended registration v2 and its confirmatory run, the fresh-seed replication, and the registered extensions of Table 4. The event-by-event timeline, with dates, status, and the itemized wording of every entry, is in the process record released with the repository.

All protocols, manifests, and approvals are mutually bound by SHA-256; the pinned digests, executor identities, execution records, and per-round audit inventory are in the process record released with

Table 6: The six-configuration exploration that preceded registration v2 (same support, loss-scale constant, and optimizer as the frozen v1 run; exploratory, not confirmatory). "hack" denotes the exploratory over-scoring model that v2 later formalized as feature-correlated one-sided spikes; row 1 is the registered v1 run itself; row 6 is an independent replication of row 5 on fresh streams. Only rows 5–6 give HL a positive effect; v2 promoted their joint condition (interpolation regime plus feature-correlated one-sided over-scoring) to the preregistered hypothesis, and the confirmatory mean contrast $+0.0010573$ is smaller than row 6's $+0.00213$ ($t \approx 1.97$ versus 3.6).

#	Configuration (regime × noise)	HL−MSE	Huber−MSE	HL−Huber
1	n=512 × frozen symmetric mixture (= v1, 20 seeds)	-0.00494 (0/20, t= − 23)	$+0.00006$ (n.s.)	-0.00499 (0/20)
2	n=512 × hack, random direction	-0.00234 (0/20)	$+0.00052$ (16/20, t=2.3)	-0.00286 (0/20)
3	n=512 × hack, adversarial direction ($\cos = -0.7$)	-0.00112 (0/20)	$+0.00302$ (20/20, t=13)	-0.00414 (0/20)
4	n=32 × frozen symmetric mixture	-0.00054 (n.s.)	$+0.00013$ (n.s.)	-0.00067 (n.s.)
5	n=32 × hack, adversarial (20 seeds)	$+0.00126$ (15/20, t=2.2)	$+0.00063$ (n.s.)	$+0.00064$ (n.s.)
6	n=32 × hack, adversarial (replication, fresh streams, 40 seeds)	$+0.00213$ (29/40, t=3.6)	$+0.00039$ (n.s.)	$+0.00174$ (29/40, t=2.9)

the repository, summarized in H.8. The E2 contamination model is registered under namespace `HLREBEL_NOISE_v2` (spec v1.4, §10; the v1.5 revision changes no scientific definition, H.8).

H.5 Secondary records and controls, by scientific question

The records are grouped by the question they answer (Table 7 maps every protocol code to its question and paragraph). Each control is recorded as a question, an intervention, a readout, and a verdict; frozen branch wordings are quoted verbatim in the verdict. Process facts (digests, executors, wall times, audit counts), superseded readings, and execution narratives are in the process record released with the repository (summary: H.8).

Table 7: **Protocol codes used in this appendix**, by the question each record answers.

Code	Question	Where
E1, E2, E3; SUPP-E2REP	main effects and the fresh-seed replication	H.5.1
EXT-ENVRESAMPLE	does the effect depend on the environment instance?	H.5.1
EXT-ONPOLICY	does the effect survive policy-to-sampling feedback?	H.5.1
EXT-FACTPOWER; EXT-FACTORIAL, EXT-FACTSTATS	which contamination factors carry the sign?	H.5.2
SUPP-ANATOMY; EXT-SWEEPS (support, smoothing); EXT-PUREWIDE; EXT-E1BOUNDARY, -D10; EXT-OCCEFF	boundary or interior geometry?	H.5.3
EXT-SWEEPS (regime, ridge); EXT-OPTCONTROL	sample size, ridge, and optimization budget	H.5.4
EXT-MINNORM; EXT-NULLSPACE	is the influence-cap conjecture supported?	H.5.5
EXT-NEURAL, EXT-NEURALFULL, EXT-UNTIED; EXT-LLM-LENGTH, EXT-LLMPOWER; EXT-RMBRIDGE	does the sign structure transfer beyond the linear platform?	H.5.6
EXT-ONPOLICYMLP	does the boundary-bias prediction hold under $\mu_t = \pi_t$ with a nonlinear potential?	H.5.7
EXT-ONPOLICYMLP-PW	is the k=3 attenuation a property of the support extent alone?	H.5.7
SUPP-CALSWEEP; R6-BOUNDDIAG, EXT-BOUNDDIAG2; EXT-CHECKPOINT	engineering instruments (no scientific hypothesis)	H.6, H.2.8

H.5.1 MAIN EFFECTS AND THEIR REPLICATION

Robustness checks on the primary contrasts Table 8 collects the secondary readouts of the two registered contrasts and of the replication; the narrative points follow. For E1, across a $4\times$ threshold range Huber stays within ± 0.0014 of MSE, never approaching the HL tax, so the conservative-default recommendation is not δ_H-fragile. For E2, along the alignment geometry the effect is non-monotone and weakens only at strong alignment; the alignment -0.7 was fixed in the v2 registration and the two neighbouring cells are themselves preregistered descriptive sensitivity rows, so neither was used to choose it. For the replication, parametric, nonparametric, and trimmed estimates agree on fresh randomness, and Huber$-$MSE is of the same order as the HL$-$Huber readout.

Table 8: **Robustness checks on the primary contrasts** (values as recorded; seeds+ in parentheses). Boundary case: the replication's HL$-$Huber contrast is one-sided $p = 0.049979$, inside the 0.05 threshold by 2×10^{-5}, a single seed's worth of noise, while the protocol's frozen readout for this contrast is two-sided ($p = 0.100$); the two-sided readout applies as frozen, and the qualification of Section 5.3 stands for the registered and replication runs, while the powered factorial (H.5.2) and the longer Adam budgets (H.5.4) detect a limited positive difference in the same setting and the on-policy test (H.5.7) a larger negative one under symmetric contamination.

Contrast	Readouts
E1, HL$-$MSE	20% trimmed mean -0.004916; Wilcoxon $p = 1.9 \times 10^{-6}$; preregistered mirrored difference-in-differences B1 $= -0.004546$ $(0/20)$
E1, Huber$-$MSE, $\delta_H \in \{W/4, W/2, W\}$	$+0.0013150$ $(15/20)$, $+0.0000564$ $(9/20)$, -0.0001771 $(7/20)$
E2, realized contamination monitors	mark rate 5.11%, spike rate 2.46%, pair-mark rate 9.95%, alignment rate 0.4835, induced gap mean $+1.3 \times 10^{-4}$
E2, HL$-$MSE	Wilcoxon $p = 0.0817$; 20% trimmed mean $+0.000669$; B1 interaction contrast $+0.0012837$ ($p = 0.0218$, CI $[2.0 \times 10^{-4},\ 2.37 \times 10^{-3}]$)
E2, alignment sensitivity, $\cos(h, \theta^\star)$	-0.5: $+0.001049$ ($p = 0.033$); -0.7 (registered): $+0.0010573$; -0.9: $+0.000485$ ($p = 0.324$)
Replication, HL$-$MSE	20% trimmed mean $+0.000974$ (83% of the untrimmed mean; in the original E2 cell trimming halves the estimate); Wilcoxon two-sided $p = 0.0053$ (one-sided 0.0027); Cohen's $d = 0.34$
Replication, Huber$-$MSE	$+0.00049$ $(46/80$, two-sided $p = 0.072)$

E2 dose and sensitivity (deferred from Section 5.3) At the doubled contamination rate ($\varrho = 0.10$, amplitude unchanged) the effect strengthens (HL$-$MSE $= +0.0030384$, $33/40$, $t = 3.95$; HL$-$Huber $= +0.0029946$, $t = 4.00$; Figure 8). At 64 pairs per round the benefit vanishes, and the Huber comparison reverses significantly (Table 9; Huber$-$MSE $= +0.001562$, $p = 1.9 \times 10^{-5}$); alignment-geometry sensitivity rows below.

The full phase table Table 9 is the complete phase diagram behind Figure 2(a–c), with confidence intervals for every cell; Figure 7 shows the per-seed contrasts behind it.

Does the effect depend on the environment instance? (EXT-ENVRESAMPLE)

Question Every seed of the record shares one environment instance ($\varphi, \theta^\star, h$ and the constants derived from its clean warmup; H.2.1), so the confidence intervals of the phase diagram do not cover instance-to-instance variation.

Intervention The platform's full recipe (instance $\to \varphi \to \theta^\star \to r^\star \to h \to$ warmup $\to W, \varsigma, \sigma, \delta_H, \tau, S$) is applied to eight fresh keys with the training and noise streams held fixed (seeds 0–39), so that the instance is the only varied factor and the runs are paired on identical stream indices; instance 0 is the frozen instance. Each instance runs the E2, symmetrized, and clean cells under the three losses.

Readout Table 10. The E2 HL$-$MSE contrast is positive in $8/8$ fresh instances (mean over instances $+0.00097$, sd 0.00041; one-sided $p < 0.05$ in 4 of eight at the per-instance power of the registered run), and the registered instance ($+0.0010573$) ranks fourth of nine. HL$-$Huber is positive in $7/8$ (mean $+0.00051$), Huber$-$MSE in $8/8$ (mean $+0.00046$), and the clean deficit reproduces in $8/8$ (mean -0.000228, sd 0.000041). The symmetrized-model contrast does not reproduce: $7/8$ fresh instances give a non-negative HL$-$MSE (mean $+0.00020$, sd 0.00061), and the registered instance's -0.0018072 is the most negative of the nine, consistent with the readout dependence of

Table 9: **Conditional phase diagram of the HL-for-MSE swap.** Paired per-seed contrasts in `perf_final`; positive favors HL. Shares of the series' SESOI (HL−MSE, rows in order): -34%, -12% ($\approx$0 at n=400: -0.00003, 95% CI $[-0.0007, +0.0006]$), $+7\%$, $+7\%$, -1%. At the doubled rate ($\varrho = 0.10$) the E2 row strengthens to HL−MSE $= +0.0030384$ ($33/40$, $t = 3.95$).

Contamination structure	Regime	HL − MSE	HL − Huber	n
Zero-mean symmetric mixture (E1)	Averaging (512 pairs/round)	-0.0049361 ($0/20$, $t = -23.41$, $p = 1.78 \times 10^{-15}$; CI $[-0.005377, -0.004495]$)	-0.0049926 ($0/20$, $t = -24.46$, $p = 6.7 \times 10^{-16}$; CI $[-0.005420, -0.004565]$)	20
Zero-mean symmetric (E3 control)	Interpolation (32 pairs/round)	-0.0018072 ($14/40$, $p = 0.0396$; CI $[-0.003524, -0.000090]$)	-0.0020377 ($13/40$, $p = 0.0409$)	40
Feature-correlated one-sided spikes (E2)	Interpolation (32 pairs/round)	$+0.0010573$ ($26/40$, one-sided $p = 0.0282$; CI $[-0.000030, +0.002145]$)	$+0.000785$ ($20/40$, two-sided $p = 0.117$, n.s.)	40
Same model, fresh-seed replication (§5.3)	Interpolation (32 pairs/round)	$+0.0011726$ ($48/80$, one-sided $p = 0.0015$; CI $[+0.000412, +0.001933]$)	$+0.000682$ ($44/80$, $p = 0.100$, n.s.)	80
Same model, sensitivity (E2, $n = 64$)	Exit interpolation (64 pairs/round)	-0.000177 ($p = 0.44$; CI $[-0.000632, +0.000278]$)	-0.001739 ($p = 6.1 \times 10^{-8}$, reversal)	40
Clean (all three experiments)	—	-0.0003904 to -0.0002264 (significant, $p = 1.3 \times 10^{-12}$ and 1.1×10^{-5} for the E1 and E2 chains, $0/20$ and $9/40$ positive; $\lvert d \rvert \ll$ SESOI, $= 0.01452$ for E1's chain and 0.01566 for the E2/E3 chain)	same direction	20/40/40

Table 10: **Environment-instance resampling** (descriptive; 40 paired seeds per cell; training and noise streams identical across instances, so instances are paired on stream indices). Instance 0 is the registered instance; its nine arms reproduce the frozen records. W is the instance's own calibrated support half-width (τ ranges over 0.0733–0.0762).

Instance	W	E2 HL−MSE (pos)	E2 HL−Huber	E3 HL−MSE (pos)	clean HL−MSE
registered (0)	0.9154	+0.001057 (26/40)	+0.000785	−0.001807 (14/40)	−0.000226
fresh 1	0.9205	+0.001306 (27/40)	+0.001171	+0.000019 (21/40)	−0.000251
fresh 2	0.9377	+0.001498 (24/40)	+0.001266	−0.000970 (18/40)	−0.000248
fresh 3	0.9020	+0.000851 (24/40)	+0.000036	+0.000451 (17/40)	−0.000162
fresh 4	0.9371	+0.000968 (26/40)	+0.000475	+0.000321 (22/40)	−0.000171
fresh 5	0.9125	+0.000534 (22/40)	−0.000413	+0.000411 (21/40)	−0.000272
fresh 6	0.8732	+0.000866 (26/40)	+0.000303	+0.000109 (16/40)	−0.000219
fresh 7	0.8806	+0.000329 (15/40)	+0.000184	+0.000047 (18/40)	−0.000236
fresh 8	0.9275	+0.001372 (27/40)	+0.001055	+0.001224 (24/40)	−0.000262
mean over fresh (sd)		+0.000966 (0.000410)	+0.000510 (0.000601)	+0.000201 (0.000610)	−0.000228 (0.000041)

the E3 deficit (optimizer-precision control) and with the near-zero symmetric-independent corner of the powered factorial (corner D, +0.00003). Instance constants vary little ($W \in [0.873, 0.938]$, $\tau \in [0.0733, 0.0762]$, pair-level spike rate 4.8–5.2%). The sd of the eight instance means (0.00041) lies below the standard error of one 40-seed mean ($0.0034/\sqrt{40} = 0.00054$): no instance component of variance is resolvable at this design, since every instance mean carries its own 40-seed sampling error on the shared stream indices.

Verdict Per the pre-frozen branch wording: *Re-drawing the environment instance—φ, $\theta^\star$, h and the constants derived from its clean warmup—eight times with the training streams held fixed reproduces the sign of the E2 HL−MSE contrast in $8/8$ instances (mean over instances +0.00097; the registered instance ranks fourth of nine): the sign is not specific to the frozen environment instance.* The magnitude stays an order of magnitude below the SESOI in every instance, and the branch counts the signs of instance means, not per-instance significance. Exploratory-descriptive; protocol and branch wordings frozen before execution.

Does the effect survive policy-to-sampling feedback? (EXT-ONPOLICY)

Question The platform draws both endpoints of every pair from the fixed law Q, which removes the feedback by which a loss changes the policy and the policy changes the next round's data (H.2.2); whether the E2 sign depends on that removal is what this run measures.

Intervention Everything is held fixed except the sampling law: contexts come from the frozen stream, and the endpoints are drawn by inverse-CDF sampling from the current policy $\pi_{t-1}(\cdot \mid x)$—both endpoints (`pi_pi`; Algorithm 1 with $\mu_t = \pi_t$, as in the LLM stack) or the first endpoint only, with the second uniform (`pi_u`)—from a fresh uniform-number stream (key `HLREBEL_SYNTH_ONPOLICY_v1` | s | t | i) shared by the three losses; noise law, optimizer, and readout are unchanged. Because the policies diverge from round 2, the three losses no longer see identical data: contrasts are paired on the seed, not on the data, and their variance is correspondingly larger, so the run uses 160 seeds (0–39 and 100–219).

Readout Table 11. Under `pi_pi` the E2 HL−MSE contrast is +0.0025 (95/160, one-sided $p = 0.012$; 95% interval [+0.0003, +0.0047]; per-seed sd 0.0139 against 0.0034 on fixed data), HL−Huber +0.0033 (95/160; two-sided $p = 0.009$, a descriptive side readout without multiplicity control), Huber−MSE −0.0008; under `pi_u` it is +0.0019 (86/160, one-sided $p = 0.024$), with HL−Huber +0.0012 ($p = 0.24$). On the registered seeds 0–39 alone, listed separately as the protocol requires, the `pi_pi` E2 contrast is +0.0001 (19/40, $p = 0.97$) and the `pi_u` contrast +0.0018 (22/40, $p = 0.34$): the 160-seed sign is carried by the 120 fresh seeds, and the registered subset sits at the platform's noise floor under feedback. On clean rewards the three losses are within noise of one another under feedback (+0.00001 and −0.00014; the fixed-data ripple of −0.00023 lies below the resolution of these cells). The feedback changes the data in the direction the mechanism predicts: the policies concentrate on high-reward actions, which lie on the negative side of h ($\cos(h, \theta^\star) = -0.7$), so fewer sampled endpoints are eligible for a spike and the realized pair-level spike rate halves (2.5% against 4.86%), while 6.8% of pairs draw the same action twice.

Verdict Per the pre-frozen branch wording: *Closing the loop—both endpoints of every pair drawn from the current policy, as in Algorithm 1 with $\mu_t = \pi_t$, so that each loss shapes its own data—the E2 HL−MSE contrast keeps its sign at 160 seeds (+0.0025, one-sided $p = 0.012$, 95/160): the fixed-data reading survives policy-to-sampling feedback on this platform.* The run is a sign check at the platform's own precision, not a second confirmatory cell, and it does not change the fixed-data

Table 11: **Policy-to-sampling feedback** (descriptive; 160 seeds, paired on the seed; contexts and noise from the frozen streams, endpoints by inverse-CDF sampling from the current policy). p_1 one-sided (HL−MSE > 0); sd is the per-seed standard deviation of the HL−MSE contrast; "same" is the share of pairs drawing the same action twice; spike is the realized pair-level spike rate. The fixed-data rows are the frozen record on seeds 0–39 ; the subset rows list the registered seeds 0–39 alone under feedback (HL−MSE only, as the protocol lists them; p_2 two-sided).

Sampling law	cell	HL−MSE (pos; p_1)	HL−Huber	Huber−MSE	sd	same	spike
fixed data (registered; seeds 0–39)	E2	$+0.001057$ (26/40; $p_1 = 0.028$)	$+0.000785$	$+0.000272$	0.0034	1.1%	4.86%
fixed data (registered; seeds 0–39)	clean	-0.000226 (9/40)	-0.000315	$+0.000089$	0.0003	1.1%	0
pi_pi: $y, y' \sim \pi_{t-1}$	E2	$+0.002503$ (95/160; $p_1 = 0.012$)	$+0.003316$	-0.000813	0.0139	6.8%	2.52%
pi_pi: $y, y' \sim \pi_{t-1}$	clean	$+0.000010$ (78/160; $p_1 = 0.483$)	-0.000170	$+0.000180$	0.0030	6.2%	0.00%
pi_u: $y \sim \pi_{t-1}, y' \sim U$	E2	$+0.001931$ (86/160; $p_1 = 0.024$)	$+0.001166$	$+0.000765$	0.0122	1.1%	3.65%
pi_u: $y \sim \pi_{t-1}, y' \sim U$	clean	-0.000136 (70/160; $p_1 = 0.852$)	-0.000024	-0.000112	0.0016	1.0%	0.00%
pi_pi, seeds 0–39 subset	E2	$+0.000095$ (19/40; $p_2 = 0.97$)	–	–	0.0148	–	–
pi_u, seeds 0–39 subset	E2	$+0.001843$ (22/40; $p_2 = 0.34$)	–	–	0.0121	–	–

Table 12: **Ablation suite III: moment-matched factorial** (corner A = E2; monitors: realized spike rate, signed spike mean (symmetry), cov(gap noise, feature projection) (independence).

corner	HL−MSE	p_2	pos	rate	signed mean	cov
A	$+0.0010573$	0.056	26/40	2.461%	$+0.641$	$+0.00507$
B	-0.0004958	0.503	18/40	2.461%	$+0.001$	-0.00044
C	$+0.0013904$	0.125	26/40	2.456%	$+0.614$	$+0.00033$
D	$+0.0006600$	0.433	26/40	2.456%	-0.014	$+0.00020$

platform's role as the mechanism instrument; the magnitude under feedback is not separated from the fixed-data value. The HL−Huber side readout is the second nominal one-sided separation in this record, after the replication's $p = 0.049979$ (boundary case above); both are descriptive, and neither alters the frozen two-sided qualification, which is a fixed-data statement. Exploratory-descriptive; protocol and branch wordings frozen before execution.

H.5.2 WHICH CONTAMINATION FACTORS MOVE THE EFFECT

The powered factorial (1200 seeds per corner) and its dose ladders locate the sign of the benefit in the conjunction of one-sidedness and feature correlation at the tested amplitude. Table 12 is the earlier n=40 factorial's monitor table (the powered run's corner A reuses it) and Table 13 its formal statistics: the superseded reading, whose narrative is in the process record.

Which contamination factors carry the sign? (EXT-FACTPOWER: powered factorial and dose ladders)

Question Whether the E2 benefit needs both one-sidedness and feature correlation, or whether contamination dose (rate, amplitude) alone reproduces it; the n=40 factorial left this unresolved (Table 13).

Intervention Corner A is the frozen E2 cell extended to n=1200 seeds per corner; corners B–D are its moment-matched symmetric and feature-independent counterparts (monitors: Table 12); the symmetrized-control rung is anchored to the frozen E3 realization. Two dose ladders vary the contamination rate over symmetric feature-independent noise (1.19–5.0% realized) and the spike amplitude (2×); the symmetrized control is re-run at n=400 under its own rule.

Readout HL−MSE means: A $+0.0012417$ ($p = 2.3 \times 10^{-29}$, 742/1200 positive), B $+0.0001380$ ($p = 0.29$; 95% CI $[-0.00012, +0.00040]$), C $+0.0000310$ ($p = 0.81$; CI $[-0.00022, +0.00028]$), D $+0.0000304$ ($p = 0.82$; CI $[-0.00024, +0.00030]$): as a descriptive comparison of these intervals with A's point estimate, each of B–D lies below a third of A's effect; all three structural effects are established (one-sidedness $+0.0005521$, $p = 4.3 \times 10^{-5}$; correlation $+0.0006592$, $p = 5.6 \times 10^{-10}$; interaction $+0.0011032$, $p = 2.3 \times 10^{-11}$; $\alpha = 0.001$ with margin 0.00025). At the same precision corner A also separates HL from curvature-matched Huber: HL−Huber $+0.0010796$ (two-sided $p = 7.8 \times 10^{-20}$, 701/1200 positive, 95% CI $[+0.00085, +0.00131]$; descriptive, uncorrected), a contrast the registered and replication cells leave unresolved (Table 9); corners B–D give -0.00021, -0.00016, -0.00033 ($p = 0.13, 0.26, 0.022$). The rate ladder is flat: rung means -0.00027, -0.00029, -0.00032, -0.00003 with SE ≤ 0.00033, every rung's 95% interval inside $[-0.001, +0.0007]$; paired first-versus-last difference $+0.00024$, 95% CI $[-0.00047, +0.00095]$,

Table 13: **Formal factorial statistics** (HL−MSE, $n = 40$ CRN-paired seeds; percentile bootstrap $B = 20{,}000$). Effect conventions: one-sidedness $= ((A-B)+(C-D))/2$; correlation $= ((A-C)+(B-D))/2$; interaction $= (A-B)-(C-D)$.

Quantity	Mean	95% CI (t)	95% CI (boot)	p (two-sided)
Corner A (corr, one-sided)	$+0.0010573$	$[-0.0000303, +0.0021448]$	$[+0.0000642, +0.0021472]$	0.056
Corner B (corr, symmetric)	-0.0004958	$[-0.0019799, +0.0009883]$	$[-0.0019488, +0.0008728]$	0.503
Corner C (indep, one-sided)	$+0.0013904$	$[-0.0004047, +0.0031854]$	$[-0.0003408, +0.0031133]$	0.125
Corner D (indep, symmetric)	$+0.0006600$	$[-0.0010255, +0.0023454]$	$[-0.0010330, +0.0021997]$	0.433
One-sidedness main effect	$+0.0011417$	$[-0.0002463, +0.0025297]$	$[-0.0000904, +0.0025634]$	0.104
Correlation main effect	-0.0007444	$[-0.0018138, +0.0003250]$	$[-0.0017923, +0.0002644]$	0.167
Interaction (diff-of-diffs)	$+0.0008227$	$[-0.0010097, +0.0026551]$	$[-0.0008727, +0.0025902]$	0.369

Table 14: **Ablation suite I: support scaling and smoothing** (Section 5.5; descriptive, $n = 40$ paired seeds; HL−MSE mean (positive seeds); S recalibrated per configuration by the frozen rule).

cell	k=1	k=1.5	k=2	k=3
E2 model	$+0.0010573$ (26/40)	$+0.0001652$ (21/40)	-0.0001233 (14/40)	-0.0001221 (12/40)
symmetric model	-0.0018072 (14/40)	-0.0007508 (18/40)	-0.0001470 (18/40)	-0.0000783 (18/40)
clean	-0.0002264 (9/40)	-0.0002319 (5/40)	-0.0001707 (7/40)	-0.0001216 (6/40)
Smoothing ($\sigma'=\beta'\varsigma$; columns $\beta' \in \{0.75, 1.5, 3\}$):				
clean	-0.0002264 (9/40)	$+0.0000046$ (20/40)	$+0.0000242$ (28/40)	
contaminated	$+0.0010573$ (26/40)	$+0.0011188$ (26/40)	$+0.0011372$ (26/40)	

$p = 0.50$; per-seed slope over the tested span $+0.00013$, CI $[-0.00037, +0.00062]$: flat within about ± 0.001, the resolution of these arms. The amplitude ladder is dose-monotone in outlier scale: at $2\times$ spike scale the symmetric-independent cell flips positive ($+0.006251$, $p = 7 \times 10^{-22}$). The n=400 re-run of the symmetrized control reads -0.0000333 (95% interval $[-0.00067, +0.00060]$) against the frozen -0.0018072, consistent across seed sets with the fresh-seed n1 cell of the untied extension (-0.00003, seeds 100..179).

Verdict Frozen wordings: "the interaction is established at powered precision (n=1200, SE $\sim 10^{-4}$): neither factor is sufficient alone—the conjunction reading returns, now with power."; "the rate axis is flat at powered precision across the tested ladder"; "the powered re-run of the symmetrized control at its own dose confirms the sign with a shifted magnitude (reported)". Read at the tested amplitude and precision: single factors did not produce an effect of A's size and the interaction supports a joint contribution; the other corners are bounded, not zero; scale moves the symmetric family at higher amplitude, structure at the frozen dose. For the symmetrized control the conservative reading, which we adopt, is that the sign is retained only nominally and the magnitude is indistinguishable from zero: the registered n=40 cell stands, and its magnitude does not extend. The conjunction requirement is consistent with the influence-cap conjecture (alignment supplies the leverage correlation, one-sidedness the net direction), which this piece leaves untested and the closed-form companion (H.5.5) subsequently does not support. Exploratory-descriptive; protocol, thresholds, and readings frozen before execution.

H.5.3 Boundary or interior geometry

How large is the excluded mixture term, and does Theorem 6 bound it? (companion computation and BIAS-SANDWICH)

Question Proposition 2's term $(g^\star - \bar{a})^2$ is excluded from the deterministic theory: how large is it for the tested noise laws at the frozen support, and does the sandwich of Theorem 6 bound it?

Intervention For the tested laws the population tied-CE minimizer $g^\star(a)$ is evaluated numerically and the static bias $g^\star(a) - \mathbb{E}[\tilde{a} \mid a]$ reported at $a \in \{0, \mathrm{SD}, 2\mathrm{SD}, W - 3\sigma\}$ for the frozen (k=1) and the tripled (k=3) support (Table 15). The E1 pair law is the exact four-component Gaussian mixture; the one-sided law is a stylized pair-level surrogate (spike rate 2.46%, discretized half-normal amplitudes; endpoint-level one-sidedness partially symmetrizes at pair level in the real generator, so this row is an upper-bound-flavored illustration). The sandwich instrument (`bias_sandwich_20260906`) reuses the same noise laws, the 96-node Gauss–Hermite quadrature, and the $g^\star$ solver, and evaluates at $a \in \{\mathrm{SD}, W - 3\sigma\}$ every quantity of Theorem 6 (Table 16): the gradient $\sigma^2 \ell'_{\bar{q}}(\bar{a})$ (the lower bound), its $\kappa_{\mathrm{HL}}/(2\sigma^2) = 1.90$ multiple (the upper bound), the clip and ripple parts of the gradient, the path-integral bound of Theorem 6(iii), and the expected excursion $\mathbb{E}[(|\tilde{a}| - (W - 3\sigma))_+]$;

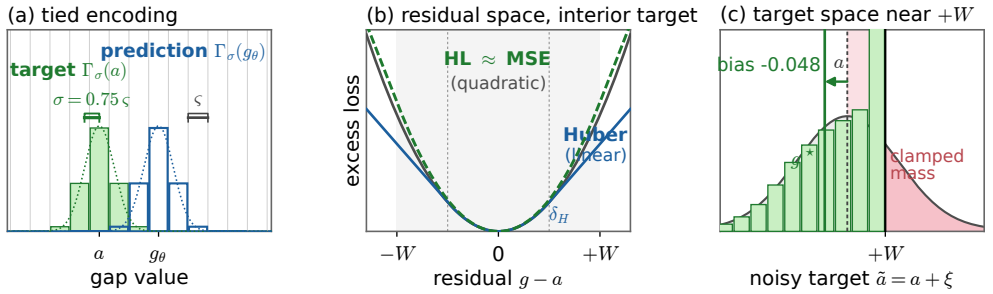


Figure 4: **The tied encoding and where each loss clamps** (the former panels of Figure 1). (a) The encoding at the support used throughout ($m = 51$, $\sigma = 0.75\varsigma$). (b) In residual space, at an interior target, the tied loss is quadratic like MSE and only Huber bends, at its threshold. (c) In target space near $+W$, symmetric noise at $a = W - 3\sigma$ is clamped at $+W$ and renormalized inward, so the best scalar fit $g^\star$ sits below a by the Table 15 value.

Table 15: **Static mixture-term bias** $g^\star(a) - \mathbb{E}[\tilde{a} \mid a]$ by law and support (columns: $a = 0$, 1SD, 2SD, $W - 3\sigma$).

law, support	a=0	1SD	2SD	$W-3\sigma$
symmetric mixture (E1 law), k=1	+0.00000	−0.00674	−0.01479	−0.04788
symmetric mixture (E1 law), k=3	−0.00000	−0.00003	−0.00008	−0.00027
one-sided spike (stylized), k=1	−0.00148	−0.00362	−0.00705	−0.03425
one-sided spike (stylized), k=3	+0.00099	+0.00098	+0.00096	+0.00087

every expectation is exact under the quadrature law, to which Theorem 6 applies verbatim. Static population quantities, not propagated through training: direction and scale anchors only.

Readout Symmetric law: -0.0067 at one target-SD, already the order of the measured E1-chain deficit; -0.048 at the interior bound, collapsing to -0.00027 at k=3. One-sided law: -0.034 at the interior bound, the benefit's winsorization mechanism; its boundary term too vanishes at k=3, and what remains is a nearly constant residual of $\approx$+0.001, essentially independent of a and opposite in sign, an interior discretization term of the skewed law at the tripled bin width. The sandwich holds in every cell; at the interior bound the clip part carries 97% of the gradient for the symmetric law at the frozen support and 91% at k=3, and the one-sided law's k=3 residual is ripple alone (clip $\approx$0). Read as a bound rather than a decomposition: at the interior bound the upper side of the sandwich is 1.73× the realized bias (its two sides differ by the factor 1.90), the clip term carries 97% of the exact gradient, the path bound is 1.30× and the excursion bound 1.64× the exact clip term, and once the $O(\varsigma)$ ripple bound is added the fully analytic bound is 3.6× (path form) to 4.3× (excursion form) the realized bias, so the slack sits almost entirely in the ripple term (Section 6.3). Property (M) of Corollary F.1 was checked on the same functions: $V_I(s)$ is nonincreasing in $|s|$ for W/σ on a 40-point logarithmic grid in $[0.25, 200]$ (the frozen ratio is 34), and v is increasing on $[0.05, 8]$. The same computation on a finer 22-point grid of a underlies Figure 1b; the four tabulated values are reproduced to 10^{-9}.

Verdict The excluded term is of the E1 deficit's order at the frozen support and is the one-sided benefit's mechanism; Theorem 6 bounds it within a factor set by the ripple term. The bound tracks the bias, not the policy effect: the two dose cells of H.5.7 with equal occupancy have opposite signs. The quadrature law differs from the continuous law by at most 3×10^{-4} in any column; audit: process record (H.8). Descriptive; no training quantity is touched.

Is the categorical form or the bounded encoding the operative variable? (SUPP-ANATOMY)

Question Whether the two contamination effects belong to the categorical cross-entropy or to any bounded target encoding, and whether the normalization denominator of Γ_σ is a necessary cause.

Intervention Two synthetic-only target-space variants join the three registered losses on bit-paired streams of the E2 cell (E2's feature-correlated model) and the E3 cell (E3's symmetrized model), 40 seeds each: **two-hot**, the mean-preserving two-point target on adjacent bin centers (decode-identity on the convex hull of the bin centers, $|a| \le W - \varsigma/2$; no normalization denominator; prediction side

Table 16: **The mixture-bias sandwich of Theorem 6 on the laws of Table 15** ($\bar{a} = \mathbb{E}[\tilde{a} \mid a]$; "path" is the integral bound of Theorem 6(iii); "excursion" is $\mathbb{E}[(|\tilde{a}| - (W - 3\sigma))_+]$; frozen 96-node quadrature law; instrument and audit in H.8).

law, support	a	$g^\star - \bar{a}$	$\sigma^2 \ell'_{\bar{q}}(\bar{a})$	$\frac{\kappa_{\mathrm{HL}}}{2} \lvert\ell'_{\bar{q}}(\bar{a})\rvert$	clip	ripple	path	excursion
E1 law, $k{=}1$	1SD	-0.00674	$+0.00624$	0.01186	$+0.00607$	$+0.00017$	0.01152	0.01255
E1 law, $k{=}1$	$W{-}3\sigma$	-0.04788	$+0.04365$	0.08295	$+0.04246$	$+0.00119$	0.05499	0.06961
E1 law, $k{=}3$	1SD	-0.00003	$+0.00002$	0.00004	$+0.00002$	$+0.00000$	0.00004	0.00006
E1 law, $k{=}3$	$W{-}3\sigma$	-0.00027	$+0.00024$	0.00046	$+0.00022$	$+0.00002$	0.00032	0.00045
one-sided, $k{=}1$	1SD	-0.00362	$+0.00334$	0.00634	$+0.00403$	-0.00069	0.00446	0.00481
one-sided, $k{=}1$	$W{-}3\sigma$	-0.03425	$+0.03049$	0.05795	$+0.03044$	$+0.00005$	0.04227	0.05710
one-sided, $k{=}3$	1SD	$+0.00098$	-0.00087	0.00165	$+0.00001$	-0.00087	0.00001	0.00002
one-sided, $k{=}3$	$W{-}3\sigma$	$+0.00087$	-0.00077	0.00146	$+0.00009$	-0.00086	0.00014	0.00020

Table 17: **Loss-geometry anatomy.** Four clamping geometries—none (MSE), residual-space (Huber), boundary-level target compression (HL, two-hot), interior hard winsorization—on the two decisive cells, on bit-paired data streams. Paired per-seed contrasts, $n = 40$, two-sided p throughout; the HL$-$MSE benefit-cell entry is the preregistered E2 contrast, whose registered one-sided value is $p = 0.028$. Two-hot is statistically indistinguishable from HL in both cells, with confidence intervals tight enough to exclude all but a small fraction of the corresponding HL$-$MSE effect; interior hard winsorization is significantly worse than HL in the E3 cell and captures 83% of the benefit (n.s.).

Contrast	E3 cell (symmetrized)	E2 cell (feature-correlated)
HL $-$ MSE	-0.0018072 ($14/40$, $p = 0.040$)	$+0.0010573$ ($26/40$, $p = 0.056$)
Huber $-$ MSE	$+0.0002305$ ($19/40$, $p = 0.60$)	$+0.0002721$ ($22/40$, $p = 0.42$)
HL $-$ Huber	-0.0020377 ($13/40$, $p = 0.041$)	$+0.0007852$ ($20/40$, $p = 0.117$)
TwoHot $-$ MSE	-0.0018023 ($14/40$, $p = 0.040$)	$+0.0011111$ ($26/40$, $p = 0.046$)
Winsorized $-$ MSE	-0.0027008 ($13/40$, $p = 0.008$)	$+0.0008791$ ($19/40$, $p = 0.17$)
TwoHot $-$ HL	$+0.0000048$ ($20/40$, $p = 0.79$; CI $[-3.1 \times 10^{-5}, +4.1 \times 10^{-5}]$)	$+0.0000538$ ($20/40$, $p = 0.14$; CI $[-1.8 \times 10^{-5}, +1.3 \times 10^{-4}]$)
Winsorized $-$ HL	-0.0008937 ($12/40$, $p = 0.007$)	-0.0001782 ($19/40$, $p = 0.37$)

unchanged at $\Gamma_\sigma(g)$; scale constant calibrated by the same frozen $S = 2/\operatorname{median} \ell''$ rule), and **hard winsorization** (`mse_clip`), squared loss against the target clamped at the interior bound $\pm(W - 3\sigma)$. Hard winsorization and HL differ in two respects at once, the clip location ($W - 3\sigma$ against HL's soft saturation at $\pm W$) and its hardness, and no arm separates them. 14 descriptive contrasts (7 per cell) from a protocol frozen before execution; no multiplicity correction, because the section registers no confirmatory claim and the two contrasts the reading turns on are the two *null* ones, which multiplicity control would only make harder to reject. The E1 cell (512 pairs per round) is not part of this anatomy: its boundary reading rests on the population computation above, on its stability under the high-precision solves (Table 19), and on the pure-boundary widening of the 512-pair chain below.

Readout Table 17. Two-hot is statistically indistinguishable from HL in both cells (TwoHot$-$HL $+0.0000048$, CI $[-3.1 \times 10^{-5}, +4.1 \times 10^{-5}]$ in E3; $+0.0000538$, CI $[-1.8 \times 10^{-5}, +1.3 \times 10^{-4}]$ in E2), with intervals tight enough to exclude all but a small fraction of the corresponding HL$-$MSE effect; hard winsorization captures 83% of the E2 benefit (n.s.) and is significantly worse than HL in the E3 cell (-0.0008937, $12/40$, $p = 0.007$), a gap attributed to the clip geometry, not to either factor alone.

Verdict The preregistered expectation was that two-hot and winsorization would pay *less* than HL, isolating the normalization denominator as the dominant term; the data selected the frozen alternative branch (Section 5.5). Removing the normalization denominator (two-hot) removes neither effect, so the denominator is not a necessary cause; what the two encodings share, finite support and compression of extreme targets at the boundary, carries them. The comparison changes the target encoding only: it narrows the mechanism to that shared property and does not by itself establish the optimization dynamics behind the interpolation-regime benefit (H.5.4). Descriptive; interpretive branches for both outcomes frozen before execution.

Boundary or interior: proportional widening and smoothing at fixed support (EXT-SWEEPS, support and smoothing)

Question Whether the clean deficit and the contamination effects are boundary or interior properties. Proportional widening scales (W, ς, σ) together at fixed bin count m and fixed $\beta = \sigma/\varsigma$, so its attenuation moves the boundary and the interior discretization at once; smoothing at fixed support moves only the interior. Clean targets essentially never reach the compression zone (0.099% beyond $W - 3\sigma$), so a clean-arm deficit caused by boundary clamping should be insensitive to the interior smoothing, whereas a bin-curvature ripple should shrink as the encoding is smoothed.

Intervention (i) Proportional widening by $k \in \{1.5, 2, 3\}$; the loss-scale constant S is re-derived at each k by the frozen calibration rule and lands on the exact k^2 scaling to $\leq 1.1 \times 10^{-7}$ relative deviation, and κ_{HL} depends on β alone (the 84-configuration sweep, spread $\leq 8.1\%$ across m and W at fixed β; Appendix H.6), so widening at fixed β holds curvature matching and the bin count fixed but scales the bin width with W. (ii) Smoothing at fixed support: with W, m, every bin edge, and the truncation at $\pm W$ held fixed, the target smoothing is widened from $\sigma = 0.75\varsigma$ to $\sigma' = \beta'\varsigma$, $\beta' \in \{1.5, 3\}$ (HL arm only; MSE and Huber do not depend on β), and $S(\beta')$ is recalibrated by the frozen rule ($0.0016645 \to 0.0060141 \to 0.023412$). A side effect, disclosed before execution, is that the interior bound $W - 3\sigma'$ moves inward, so the share of clean targets beyond it rises from 0.099% to 0.26% and 1.9%, the direction opposite to the ripple prediction, which keeps the two readings separable.

Readout Table 14. Widening attenuates every contamination effect to the level of the clean deficit by k=2 (E2 $+0.0010573 \to -0.0001233$; symmetric $-0.0018072 \to -0.0001470$) while the clean deficit itself stays ($-0.0002264 \to -0.0001707 \to -0.0001216$). Smoothing: the clean HL$-$MSE contrast goes from -0.000226 ($9/40$) to $+0.000005$ ($20/40$) and $+0.000024$ ($28/40$), while the E2-model benefit is unchanged ($+0.00106 \to +0.00112 \to +0.00114$; $26/40$ throughout); the clean residual of $+0.00002$ at $\beta' = 3$ is within noise ($p = 0.23$).

Verdict Per the pre-frozen branch wording: *the clean-arm deficit shrinks under smoother target smearing at frozen calibration: consistent with bin-curvature ripple, not boundary clamping; the clean-arm attribution in the paper is rewritten accordingly.* The contamination benefit does not depend on the smoothing scale, so the boundary that carries it is the truncation at $\pm W$, not the 3σ interior margin used to define occupancy. Proportional widening does not by itself isolate the boundary; the pure-boundary variants below do. The bin count at fixed W was not varied. Descriptive; protocol and branch wordings frozen before execution.

Does the boundary alone carry the effects? (EXT-PUREWIDE on the 32-pair cells; EXT-E1BOUNDARY and EXT-E1BOUNDARY-D10 on the 512-pair chain)

Question Proportional widening scales W, ς, and σ together, so its attenuation cannot say whether the contamination effects depend on where the boundary sits or on the interior discretization; the boundary attribution of the E1 deficit rested on the population computation and on its stability under the high-precision solves.

Intervention The boundary alone is moved: σ, the bin width, and every interior bin edge are held fixed, and $j \in \{13, 26, 51\}$ bins of the frozen width are appended per side ($k_j \in \{1.51, 2.02, 3.00\}$; j=51 matches the proportional-widening k=3 anchor); the baseline arms are reused unchanged. Cells: the two 32-pair interpolation-regime cells (E2 model; E3 symmetrized model), and the HL arm of the 512-pair averaging-regime chain (20 seeds, symmetric mixture contamination at the frozen and at the doubled rate) with environment, constants, streams, and training loop taken from the pinned v1.3 harness.

Readout 32-pair cells, HL$-$MSE at $j = 0/13/26/51$: contaminated $+0.00106 \to +0.00012 \to -0.00017 \to -0.00017$ with occupancy $0.0141 \to 0.0044 \to 0.0012 \to 0$; symmetric $-0.00181 \to -0.00071 \to -0.00012 \to -0.00007$; clean $-0.00023 \to -0.00030 \to -0.00030 \to -0.00030$, a one-time shift of -0.00007 (SE 0.000033), the previously truncated $\leq 0.27\%$ tail mass entering the normalizer, constant across j and inside the frozen invariance criterion. 512-pair chain: E1 contaminated $-0.0049361 \to -0.0018105 \to -0.0005556 \to -0.0000823$ ($0/20$, $0/20$, $0/20$, $5/20$ positive) as the share of targets beyond the extended inner bound falls $3.27\% \to 1.06\% \to 0.29\% \to 0.007\%$; E1 clean $-0.0003904 \to -0.0003110 \to -0.0003107 \to -0.0003107$, a one-time shift of $+0.00008$ (SE 0.000007), constant across j and inside the invariance criterion; at the doubled rate (protocol label "10× dose") $-0.0092274 \to -0.0033902 \to -0.0010853 \to -0.0001232$, occupancy $6.09\% \to 0.036\%$, $4/20$ positive at j=51. The E1 deficit collapses 60× at k=3, tracking occupancy, and 75× at the doubled rate.

Verdict Per the pre-frozen branch wordings. 32-pair cells: *moving the boundary alone—interior bins, bin width, and smoothing scale held bitwise fixed—attenuates both contamination effects below a quarter of their frozen magnitude while leaving the clean-arm ripple in place: the contamination phase structure is a boundary property (winsorization), the clean-arm deficit an interior property (bin curvature); the two mechanisms separate experimentally.* 512-pair chain: *Moving the boundary alone on the 512-pair averaging-regime chain—interior bins, bin width, and smoothing scale held bitwise fixed—attenuates the symmetric-contamination deficit below a quarter of its frozen magnitude*

Table 18: **Ablation suite II: regime controls** (HL−MSE by pairs per round; ridge−MSE at $n \in \{30, 32, 36\}$).

	n=16	24	30	32	36	48	64
HL−MSE	+0.0011254	+0.0008369	+0.0008083	+0.0010573	+0.0010775	+0.0002556	−0.0001773

	n=30	n=32	n=36
$\lambda = 0.001$	−0.0035995	−0.0036171	−0.0035313
$\lambda = 0.01$	−0.0266673	−0.0273765	−0.0274979

($-0.0049 \to -0.0001$) while leaving the clean-arm ripple in place: the E1 deficit is a boundary property (winsorization), as the population mixture-bias computation predicts. Doubled rate: *At the $10\times$ dose the same boundary-only widening attenuates the E1 deficit below a quarter of its frozen magnitude ($-0.0092 \to -0.0001$): the boundary attribution of the averaging-regime deficit holds at both tested doses.* The intervention excludes the reading that the proportional-widening attenuation needed the interior discretization to change; like every widening, it moves the boundary of the categorical encoding rather than replacing it, and it does not by itself account for the sample-size and budget dependence of the benefit under finite-step Adam (H.5.4). Exploratory-descriptive; protocol and branch wordings frozen before execution.

Does one statistic organize the record? (EXT-OCCEFF, occupancy versus effect; post-closure control)

Question Whether the fraction of realized targets beyond the interior bound of the support a cell actually used organizes the sign and size of the contamination effects across the record.

Intervention Every recorded cell with a paired HL−MSE contrast and a measurable target distribution is assembled against its occupancy: the four frozen-dose v2-chain cells, the nine support-sweep cells, the three moment-matched factorial corners, and the three E1-chain cells (n=19; Figure 3); robustness (v1.1) by leave-one-block-out over the four source blocks.

Readout Effect magnitude tracks occupancy (Spearman 0.833, $p = 9.5 \times 10^{-6}$); within families the signed effect is ordered by occupancy in opposite directions (one-sided $+0.943$, $p = 0.005$; symmetric -0.786, $p = 0.021$); clean cells sit at $\approx$0 occupancy with the ripple-level deficit only. Leave-one-block-out keeps the magnitude association in $[0.76, 0.85]$; excluding the factorial corners sharpens the symmetric-family trend to -0.94 ($p = 0.005$). One recorded exception: the symmetric-independent factorial corner is positive at $+0.00066$. The powered-factorial dose-ladder cells postdate this frozen assembly and are not among the 19; at $2\times$ spike amplitude the symmetric-independent cell turns positive ($+0.0063$, Section 5.4), extending the corner-D exception.

Verdict Occupancy is the dose variable; the two families' signed trends run opposite, and amplitude moves the symmetric family across zero, so no deterministic sign rule is claimed. Cells share generators, seeds, and streams within blocks: a within-platform descriptive association, not an externally validated predictor. Descriptive; post-closure control.

H.5.4 Dependence on optimization budget and sample size

Sample size and ridge (EXT-SWEEPS, regime and ridge controls)

Question Whether the E2 benefit is tied to the near-interpolation regime $n \approx d$, and whether an explicit ℓ_2 penalty on the MSE fit reproduces it.

Intervention Pairs per round varied at fixed $d = 32$ on the E2 model (the n=32 column is the registered cell; the other columns use prefixes of the frozen 64-pair streams); an explicit ℓ_2 penalty added to the MSE fit at $n \in \{30, 32, 36\}$ with $\lambda \in \{10^{-3}, 10^{-2}\}$.

Readout Table 18: the benefit is positive at every tested $n \leq 36$, small at 48, absent at 64; ridge is harmful relative to MSE at both penalties (-0.0036 and -0.027 at n=32, 3–25$\times$ the benefit).

Verdict The reading of Section 5.4: a near-interpolation-regime property on this platform that the tested ℓ_2 penalties do not reproduce. Weaker penalties were not run on the Adam platform; in the gradient-descent idealization $\lambda \leq 10^{-3}$ is indistinguishable from the clip (H.5.5). Descriptive; the optimizer-precision control below varies the budget on the same data.

Objective or path? (EXT-OPTCONTROL, optimizer-precision and readout control; post-closure control)

Question Whether a contrast survives when each round's fit is carried closer to its optimum, which separates a property of the fitted objective from a property of the finite-step path.

Table 19: **Optimizer-precision control** (paired means, mixture readout; $n = 40$ seeds v1.4 cells, $n = 20$ E1 cells). HL−MSE under the frozen Adam(60), $10\times$ Adam, and per-round float64 L-BFGS; Huber−MSE at the frozen dose and at the L-BFGS64 solves. The $5\times$-Adam column and full CIs are in the result artifact; parenthetical entries are two-sided p for the Adam(600) and L-BFGS columns (the E1-clean L-BFGS64 entry is machine-ε scale). On the contaminated cells the L-BFGS64 intervals are 2.5–$5\times$ wider than the Adam(60) ones and contain the frozen-dose effects (text).

	HL−MSE			Huber−MSE	
Cell	Adam(60)	Adam(600)	L-BFGS64	Adam(60)	L-BFGS64
clean	−0.0002264	−0.0000715 (.014)	−0.0000457 (.134)	+0.0000885	−0.0000098
E2	+0.0010573	+0.0039217 (.001)	+0.0001435 (.916)	+0.0002721	−0.0013184
E2 10×	+0.0030384	+0.0091457 (<.001)	+0.0035456 (.374)	+0.0000437	+0.0014904
E3	−0.0018072	+0.0075989 (.001)	−0.0026668 (.501)	+0.0002305	−0.0000000
E1 clean	−0.0003904	−0.0000007 (.193)	+0.0000000 (.001)	+0.0001910	−0.0000000
E1 contaminated	−0.0049361	−0.0049158 (<.001)	−0.0049163 (<.001)	+0.0000564	−0.0000608

Intervention Both chains' frozen streams are replayed with optimization precision as the only axis: the frozen Adam(60), $5\times$ and $10\times$ Adam (300/600 steps, same frozen hyperparameters), and per-round float64 L-BFGS solves (strong Wolfe, 200 iterations, gradient tolerance 10^{-12}; H.2.5), the *high-precision arm*; for HL and Huber it is high precision relative to the best value any solver reached, not a certificate of global optimality (only MSE has a closed-form floor). Readouts: `perf_final` (the uniform-round mixture), the final iterate, and the best iterate. Three layers are read separately, fit quality, uniqueness of the fitted parameter, and policy value; the full three-layer narrative is in the process record.

Readout Table 19. Fit quality: at the frozen dose the three losses' final per-round gradient norms match within 4% (medians 3.68–3.81×10^{-3} on E2) and their float64 objective gaps within 8% (medians 1.17–1.26×10^{-3}), so identical steps deliver matched suboptimality; the high-precision arm reaches median final gradient $\sim 4 \times 10^{-9}$ (3.1×10^{-9} on the E1 chain) and median objective gap $\leq 2.5 \times 10^{-16}$ (95th percentile $\leq 7 \times 10^{-16}$) on the contaminated cells against the best-known reference, with the MSE floor cross-checked against closed-form least squares; a few rounds stay unsolved (the L-BFGS objective above the least-squares floor by more than 10^{-10} on $4/600$ E2 and $6/600$ E3 rounds for MSE, worst gaps 2.1×10^{-3} and 1.7×10^{-2}; final gradient above 10^{-6} on at most $1/600$ rounds for MSE and $9/600$ for HL). Uniqueness: at n=32 the design matrix is rank-deficient on $182/600$ rounds, so equal objective values do not certify equal policies; the well-posed exact fit is the minimum-norm companion of H.5.5 (clip−raw at $t{=}\infty$ $= +0.0046$). Policy value: the E1 deficit is unchanged under $10\times$ Adam and the L-BFGS solves ($-0.0049361 \to -0.0049158 \to -0.0049163$); the clean deficit decays with precision; the E2 and E3 HL−MSE contrasts increase under $5\times/10\times$ Adam and are not resolvable at the L-BFGS solves, whose 95% intervals (E2 $[-0.0026, +0.0029]$, E2 doubled rate $[-0.0044, +0.0115]$, E3 $[-0.0106, +0.0053]$) have standard errors 2.5–$5\times$ the frozen-dose ones and each contain the frozen-dose effect; absolute performance degrades for every loss at the high-precision solves in the interpolation regime (E2 mean `perf_final` $-0.175 \to -0.284$; E3 $-0.164 \to -0.307$) while the E1 chain moves by less than 0.001 ($-0.1452 \to -0.1445$). On E2 the HL−Huber contrast follows the budget axis: $+0.0038546$ at Adam(300) ($30/40$, two-sided $p = 0.0020$) and $+0.0055388$ at Adam(600) ($31/40$, $p = 0.00019$), against $+0.0007852$ ($p = 0.117$) at the frozen Adam(60) and an unresolved $+0.0014619$ ($p = 0.54$) at the L-BFGS solves. Final- and best-iterate readouts agree in sign with the mixture readout on E1 (both settings), E2, the doubled-rate row, and clean at every optimizer level; the symmetrized control's deficit is mixture- and budget-specific (final-iterate $+0.0025$ at the frozen dose; positive under 5–$10\times$ Adam at every readout), the one readout-dependent cell, reported in full.

Verdict The averaging-regime deficit is a property of the fitted objective, stable under high-precision solves and matched in sign and order by the population computation of Table 15; the interpolation-regime benefit is a property of the finite-step Adam path, and its high-precision limit is not resolved by this control at the achieved precision: the L-BFGS64 column shows a loss of resolution, not a demonstrated collapse to zero, and the informative optimizer axis is the $5\times/10\times$-Adam column. Compute: on the linear platform the tied categorical loss dominates per-arm cost (≈ 21 s versus ≈ 0.4 s per 40-seed arm); at LLM scale the three losses are wall-clock indistinguishable (ledger in the process record). Descriptive; protocol frozen before execution with pre-declared reading rules (OC-1/OC-2, both realized as declared).

Table 20: Closed-form spectral-filter companion (descriptive): clip−raw in `perf_final` on the one-sided model under budget-matched $t^\star$ gradient descent, by pairs per round. One-sided p (clip > raw); positive seeds out of 40. The regime signature of the Adam record (positive across n=16..36, small at 48, absent at 64) does not reproduce.

n	$t^\star$	clip−raw	one-sided p	pos
16	11	-0.000732	0.984	16/40
24	28	-0.000668	0.931	17/40
30	83	$+0.000030$	0.479	15/40
32	156	$+0.000545$	0.191	19/40
36	69	$+0.000022$	0.483	20/40
48	21	-0.000212	0.732	19/40
64	11	-0.000357	0.958	17/40

H.5.5 A CONCRETE MECHANISM MODEL: THE INFLUENCE-CAP CONJECTURE

Is the influence-cap conjecture supported? (EXT-MINNORM, closed-form spectral-filter companion)

Question The leading conjecture of Section 6.2 is a specific mechanism model: partial optimization fits high-leverage directions first, leverage-aligned spikes are therefore fitted preferentially, and bounded encoding acts as a per-sample influence cap.

Intervention The model is tested in its own idealization, where every quantity is closed-form: per round, t-step gradient descent from zero on the pairwise least squares is the spectral-filter solution $\omega_t = VG_tS^+U^\top y$, $g_t(s) = 1-(1-s^2/s^2_{\max})^t$, with per-sample partial-fit leverage $\ell_i = \sum_j g_t(s_j)U^2_{ij}$ and signed value-influence $I_i = w^\top VG_tS^+U^\top e_i$ (w the exact value gradient of the multiplicative update); hard winsorization at $\pm(W-3\sigma)$ represents bounded encoding; the budget $t^\star$ is matched per n to the frozen Adam(60) MSE median objective gap (0.0012609; realized $t^\star$ in Table 20, with $\mathrm{gap}(t^\star) \le \mathrm{target} < \mathrm{gap}(t^\star-1)$), on the frozen v2-series streams, seeds 0–39, with the t=∞ arm at the exact least-squares fit. Five signatures of the Adam record were prespecified as the model's predictions: preferential fitting of spiked pairs, a one-sided-specific influence cap, the regime shape in n, non-equivalence with ridge, and erasure of the contrast at the exact fit.

Readout Per the frozen decision rules, the conjecture's core clause fails its prespecified test (one-sided $p < 0.01$ with $\ge 30/40$ seeds positive): spiked pairs are not *detectably* preferentially fitted, $\Delta\ell(\mathrm{E2\ model}) = +0.001599$ (one-sided $p = 0.265$, 27/40; 95% CI $[-0.0035, +0.0067]$, a null at the registered threshold, not an equivalence statement, since no minimum effect was prespecified), and the one-sided-minus-symmetrized difference is $+0.0014$ ($p = 0.309$). The cap splits: spikes are systematically harmful (mean $\Sigma = -0.0184$, one-sided $p = 6.6\times10^{-5}$, the one clause that holds), but the winsorization cut of $|\Sigma|$ is not specific to the one-sided model ($+0.0017$ there, $p = 0.17$, versus $+0.0064$ on the symmetrized model, $p = 0.023$) and does not transfer to the endpoint (clip−raw = $+0.000545$, 19/40, one-sided $p = 0.19$). The remaining signatures do not reproduce either: the regime shape is absent (Table 20; significantly *negative* at n=16), small ridge penalties are indistinguishable from the clip ($\lambda \le 10^{-3}$: $p \ge 0.115$; $\lambda = 3\times10^{-3}$: $p = 0.032$) while $\lambda \ge 10^{-2}$ is significantly worse ($p < 10^{-4}$), and the exact fit does not erase the contrast (clip−raw at t=∞ is $+0.004609$, 8.5× the $t^\star$ value, nominally n.s. at $p = 0.051$ but far outside the frozen $\frac{1}{4}$ contraction bound) though it degrades both finite-budget arms absolutely ($+0.112$/$+0.108$, $p < 10^{-14}$).

Verdict Branch M-B, as frozen (historical wording; the reading adopted in Section 6.2 reads "not detectably preferentially fitted"): *"A closed-form spectral-filter companion (Appendix H.5) does not support the influence-cap reading: spiked pairs are not preferentially fitted under the one-sided model, so the overparameterization residue remains unexplained."* The gradient-descent idealization reproduces none of the five signatures, in particular not the sample-size dependence that is the Adam record's most distinctive feature; the two null clauses are nulls at the registered threshold, not demonstrations of zero. Scope: the companion tests the gradient-descent idealization, not Adam's coordinate-wise adaptive update. For every loss the round gradient is $\nabla_\omega F_t(\omega) = \frac{1}{n}X_t^\top s(\omega)$ with $s_i = \partial_g \ell\big((X_t\omega)_i, \tilde{a}_i\big)$, so plain gradient descent from $\omega = 0$ keeps every iterate in the row space of X_t, and the spectral-filter solution lies there by construction; Adam's step $-\alpha\,\hat{m} \oslash (\sqrt{\hat{v}} + \epsilon)$ rescales coordinates, so under Adam ω_t can acquire a component in $\ker X_t$ that leaves every training prediction unchanged ($X_t\omega = X_t(\omega + v)$ for $v \in \ker X_t$) and moves the policy only along directions the round did not observe; it can exist on the $182/600$ rank-deficient rounds at n=32 and on every round at $n < d$, and it is absent at full rank. The companion therefore could not have shown a null-space effect even in principle; the counterfactual below tests it on the Adam record.

Table 21: **Null-space counterfactual** (descriptive; 40 paired seeds; three losses per cell, HL shown for the null share). "Rank-deficient" counts rounds with $\ker X_t \neq \{0\}$ over 40 seeds $\times$ 15 rounds; "null share" is the mean of $\|\omega_t^{\text{null}}\|^2/\|\omega_t\|^2$ on those rounds. The last three columns are the HL−MSE contrast in `perf_final` for the recorded (full), row-space-only, and null-space-only trajectories, and the paired full−row difference (sd).

Cell	rank-def.	null share	full	row-only	null-only	full−row
E2 model, n=16	600/600	17.4%	+0.001125 (24/40)	+0.000609 (23/40)	+0.000584	+0.000517 (0.001861)
E2 model, n=24	600/600	8.7%	+0.000837 (25/40)	+0.000475 (24/40)	+0.000191	+0.000362 (0.001974)
E2 model, n=30	600/600	2.5%	+0.000808 (23/40)	+0.000565 (23/40)	+0.000219	+0.000244 (0.000935)
E2 model, n=32 (registered)	182/600	1.5%	+0.001057 (26/40)	+0.001127 (26/40)	−0.000075	−0.000069 (0.000547)
symmetrized, n=32	182/600	1.5%	−0.001807 (14/40)	−0.001771 (14/40)	+0.000003	−0.000036 (0.000711)
clean, n=32	182/600	1.3%	−0.000226 (9/40)	−0.000222 (7/40)	−0.000017	−0.000005 (0.000064)

Exploratory-descriptive; protocol (v2), decision rules, and all four branch wordings frozen before execution.

Is the benefit carried by the unconstrained directions? (EXT-NULLSPACE, null-space counterfactual)

Question Whether the E2 benefit is carried by the directions the design leaves unconstrained, the $\ker X_t$ component that Adam can produce and gradient descent cannot.

Intervention The frozen loop is transcribed with every round's increment ω_t recorded; each ω_t is split by the SVD of X_t into a row-space part and a null-space part (rank by the default threshold, never by condition number; on full-rank rounds the null part is exactly zero). Because each round's fit starts from $\omega = 0$ and does not depend on θ_{t-1}, the counterfactual trajectories need no retraining: $\theta_t^{\text{row}} = \sum_{s\le t} P_{\text{row}}(X_s)\omega_s$ and $\theta_t^{\text{null}} = \sum_{s\le t} P_{\text{null}}(X_s)\omega_s$ are read off the record, the training fit of every round being identical by construction. Readouts: `perf_final` of the full, row-only, and null-only trajectories, and the one-step value change from dropping the null part at the recorded θ_{t-1}. Cells: the E2 model at $n \in \{16, 24, 30, 32\}$, the symmetrized model and clean at n=32, three losses each.

Readout Table 21. At n=32 the null part exists on the 182/600 rank-deficient rounds only and carries 1.4–1.5% of $\|\omega_t\|^2$ there, alike for the three losses; removing it moves the E2 HL−MSE contrast from +0.0010573 to +0.0011267 (26/40 both; paired difference −0.00007, sd 0.00055), the symmetrized contrast from −0.0018072 to −0.0017708, and the clean contrast from −0.0002264 to −0.0002218; the null-only trajectories stay at the uniform-policy value (−0.377) for every loss. Below d the null part is large (17.4% of $\|\omega_t\|^2$ at n=16, on every round) and carries about half of the point estimate (n=16: +0.0011254 → +0.0006087, null-only +0.00058; n=24: +0.0008369 → +0.0004754; n=30: +0.0008083 → +0.0005646), with paired differences that are positive but not resolved at 40 seeds ($p \ge 0.09$).

Verdict Per the pre-frozen branch wording: *Removing the null-space component of every round's Adam increment—the part that leaves the training fit unchanged—moves the E2 HL−MSE contrast by less than a quarter of its value (+0.0011 → +0.0011): the benefit lives in the row space of the design, not in the directions the training data do not constrain.* The Adam-path residue of Section 6.2 is therefore a row-space property at the registered cell: what Adam does differently from gradient descent inside the span of the observed pairs, not outside it. Below d the unconstrained directions contribute, at a precision that does not separate their share from zero. Exploratory-descriptive; protocol and branch wordings frozen before execution.

H.5.6 Transfer beyond the linear platform

Does the sign structure transfer to an MLP and to an LLM pilot? (EXT-NEURAL, EXT-LLM-LENGTH)

Question Does the sign structure of the linear platform survive a nonlinear potential class, and an LLM-scale stack? Two post-closure extensions ask this (registry: H.4).

Intervention MLP potentials with $h \in \{8, 32, 128\}$ on the frozen streams (all parameters trained within each round from a shared per-(seed, width) initialization restored at every round, so the trunk is not frozen but the representation is not carried across rounds). LLM: the 3×2×3 matrix (seeds 99–101 × contaminated/clean × three losses) on the terminated phase's frozen stack, with injected, spec-registered length bias and the pinned clean reward model as ground truth: a transfer probe by design, not an LLM-scale gain claim.

Table 22: **MLP direction check** (Appendix H.5.6; descriptive, $n = 40$ paired seeds per cell, 32 pairs per round unless noted). p_1 one-sided (HL−MSE > 0); p_2 two-sided.

Cell	HL − MSE	pos	p_1	HL − Huber	p_2
E2 model, $h = 8$	-0.0000817	$21/40$	0.573	$+0.0004106$	0.489
E2 model, $h = 32$	$+0.0008081$	$22/40$	0.028	$+0.0006618$	0.101
E2 model, $h = 128$	-0.0004084	$14/40$	0.812	-0.0003081	0.584
E2 model, $h = 8$, 64 pairs	$+0.0012938$	$23/40$	0.013	-0.0017528	0.0013
E3 model, $h = 8$	-0.0028838	$11/40$	0.996	-0.0031394	0.0022
E3 model, $h = 32$	-0.0013793	$17/40$	0.935	-0.0013540	0.158
E3 model, $h = 128$	-0.0020817	$16/40$	0.980	-0.0018824	0.086

Table 23: **LLM-side direction check under injected length bias** (Appendix H.5.6; descriptive, $n = 500$ paired prompts per cell, clean evaluation-RM score units, two-sided p throughout). "share" is the per-prompt positive share; "len" is mean generated tokens (HL/MSE/Huber). Seed-level: contaminated HL−MSE positive at $2/3$ seeds, clean at $0/3$. Eighteen contrasts are shown; no entry survives a Bonferroni correction at $0.05/18 \approx 0.0028$; all rows are descriptive.

Cell	HL − MSE	share	p_2	HL − Huber (p_2)	Huber − MSE (p_2)	len
cont., s99	$+0.0877194$	0.478	0.121	-0.0695187 (0.195)	$+0.1572380$ (0.004)	33.5/33.6/34.0
cont., s100	$+0.0158531$	0.468	0.795	-0.0264603 (0.667)	$+0.0423134$ (0.496)	33.1/33.7/33.7
cont., s101	-0.0280665	0.420	0.605	-0.0456680 (0.433)	$+0.0176015$ (0.754)	33.2/33.9/33.0
clean, s99	-0.0099848	0.408	0.868	-0.0933358 (0.121)	$+0.0833510$ (0.192)	32.8/33.1/33.6
clean, s100	-0.0341640	0.418	0.538	-0.0818673 (0.133)	$+0.0477034$ (0.372)	32.6/33.0/33.0
clean, s101	-0.0603399	0.410	0.317	-0.0295423 (0.645)	-0.0307976 (0.628)	33.0/32.9/33.4

Readout Tables 22–23. MLP: the deficit replicates at every width (h=8 HL−Huber $p = 0.0022$); the benefit appears only at $h = 32$ ($+0.0008$, one-sided $p = 0.028$), the interpolation-regime shape of Section 5.3. LLM: contaminated per-prompt HL−MSE $+0.0877/+0.0159/-0.0281$ (all n.s.), clean cells negative at all three seeds (the pilot's deficit direction; the powered round below does not replicate it), Huber−MSE $+0.157$ ($p = 0.004$) the only nominally significant entry; 18 contrasts, none survives correction (n=3).

Verdict Per the pre-frozen branch wordings. MLP: *the linear-platform sign structure does not transfer (fully) to the tested nonlinear class; the paper's empirical claims are correspondingly restricted to the linear platform, and the nonlinear cells are reported in full as boundary evidence.* LLM: *the synthetic-platform direction does not (cleanly) transfer at pilot scale; reported in full as boundary evidence, consistent with the paper's restriction of claims to the tested platform.* Neither pilot is confirmatory.

Persistent representation and the untied head (EXT-NEURALFULL, EXT-UNTIED)

Question Does the MLP sign structure depend on restarting the representation every round, and does an untied categorical head change the picture?

Intervention (i) A persistent-trunk variant of the MLP pilot: trunk initialized once and trained through all 15 rounds, heads zero-initialized per round. (ii) The untied categorical head in its realizable coupled form, a per-endpoint distribution decoded to a bounded mean that *is* the policy potential, trained as a fourth arm on both platforms (linear $\Theta \in \mathbb{R}^{d \times m}$ on seeds 100–179; trainable-trunk $h \to m$ logits on the MLP grid).

Readout Persistent trunk: the symmetric-cell deficit no longer holds at every width ($+0.0021$ at h=32), the benefit no longer sits at the interpolation width (-0.0016 at h=32), and every cell is small ($|\cdot| \leq 0.0025$, 16–23/40 seeds). Untied head: untied−tied $= -0.039$ on every linear cell ($0/80$ seeds, all three settings) and -0.020 to -0.069 on the MLP grid, growing with width, 20–60× any measured HL−MSE effect (the head has $d \times m = 1632$ parameters against 32 pairs per round; the interpolation-regime qualifier travels). Two descriptive side-reads: the fresh-seed symmetric cell's HL−MSE is -0.00003 (the symmetrized family's fragility again), while the clean-arm ripple replicates at -0.00024.

Verdict Per the pre-frozen branch wordings: *the MLP sign structure under persistent representation departs from the per-round-restart record; reported in full as boundary evidence, the paper's claims already*

restricted to the tested platforms; and, on both platforms, *the bounded softmax-mixture potential class pays its own price; reported in full.* Traceability: H.8.

The powered LLM-scale extension Configuration. All LLM-scale runs use the TL;DR summarization stack of Huang et al. (2024) on Pythia (Biderman et al., 2023): the policy and the reference policy are the public Pythia-1B-deduped SFT checkpoint (`vwxyzjn/EleutherAI_pythia-1b-deduped__sft__tldr`, revision `20d6d190`; no adapter on the reference), the training reward model is the matching Pythia-1B reward checkpoint (`..._reward__tldr`, revision `33b95d01`), and the *evaluation* reward model is the Pythia-2.8B-deduped reward checkpoint (revision `1dffce01`), which never enters training (a design constraint of the frozen specification; no 2.8B policy was ever run). Training: LoRA r=1024, α=2048, learning rate 3×10^{-6} with linear decay, 512 prompts per iteration, 2 samples per prompt, 4 inner epochs, temperature 0.1, at most 53 new tokens, η=1.0, bfloat16 training with float32 losses, 150 iterations per run; the extension reuses the frozen training path unchanged. Prompts: the TL;DR filtered preprocessing of the summarize-from-feedback data (Stiennon et al., 2020), train split for training and the 500-prompt `anchor_eval` slice of the validation split, disjoint from training, for every reward-model endpoint reported here; the 2000 human-preference pairs of the reward-model bridge are a sample of the `comparisons` validation data and serve only as the reward-model calibration standard (evaluation-RM agreement with humans 0.63275, below the frozen gate), not as training data. **Initial-policy baseline.** The SFT policy's mean evaluation-RM score on the 500 anchor prompts is -1.182872 (one sample per prompt, per-prompt sd 1.9109, SE 0.0855; KL to itself 0). After 150 iterations the pooled means over the 20 seed-cells are -0.5768 (MSE), -0.5591 (HL), -0.5641 (Huber); contaminated cells -0.5713/-0.5573/-0.5563, clean cells -0.5849/-0.5618/-0.5758. Each loss therefore moves the policy by $\approx +0.62$ evaluation-RM points, while the three differ by at most 0.0177 (2.9% of the movement; the ± 0.035 equivalence margin is 5.7% of it): the equivalence statements below are not equivalence of untrained policies. The baseline and the trained arms differ in two disclosed generation conventions (one sample per prompt and a uniform-driven token loop against G=8 samples per prompt and the frozen HF generation configuration, same temperature and length cap; mean generated length 31.86 versus 33.24), so the movement is a magnitude reference, not a paired statistic.

Does the replacement help or hurt at LLM scale under injected contamination? (EXT-LLMPOWER; H-P1–H-P3 registered)

Question The pilot's confirmatory gap: whether, on the frozen TL;DR stack with injected length-bias contamination, HL differs from MSE and from curvature-matched Huber in clean-reward-model score at a preregistered equivalence margin.

Intervention A powered, separately registered extension on the same frozen stack (amended-preregistration semantics: hypotheses, endpoints, SESOI, and all five outcome wordings per hypothesis frozen before any new seed existed; the main chain's terminated channel is unaffected). Design: contaminated seeds 102–113 (n=12) and clean seeds 102–109 (n=8), three losses, 150 LoRA iterations per arm, 60 arms; the endpoint is the mean of 4000 clean-RM scores per arm (500 frozen prompts $\times$ G=8 generations under derived per-call seeds; no judge anywhere). Three reward objects are kept apart: the 1B training reward r_{train}, its contaminated version $\widetilde{r}_{\text{train}} = r_{\text{train}} + \xi$ used by the contaminated arms, and the 2.8B evaluation reward r_{eval} that never enters training; clean arms use no injected perturbation (the EOS and KL processing below still applies), and every evaluation endpoint is r_{eval}, the fixed evaluation target of this controlled proxy task, not a human-preference ground truth (the same model scored 0.63275 agreement with human labels, below the frozen gate). Contamination rule (frozen specification, namespace `HLREBEL_NOISE_v2`): each endpoint draws (u, z) from a stateless stream keyed by seed, iteration, prompt, and sample; it is *marked* if $u < 0.05$ and *aligned* if its generated length exceeds the batch median; a marked and aligned endpoint receives $\xi = |z| \cdot 10\tau$, every other endpoint of a contaminated arm receives $\xi = z \cdot \tau$ (τ the base scale of the frozen support artifact); non-terminated generations receive the overlong penalty -1.0, and every endpoint reward subtracts 0.05 times the policy-to-reference log-ratio. The reward that enters the regression is therefore $r_t^{\text{used}}(x, y) = \big[r_{\text{RM}}(x, y) + \xi\big]\mathbf{1}\{\text{terminated}\} - 1.0 \cdot \mathbf{1}\{\text{not terminated}\} - 0.05\big[\log \pi_t(y \mid x) - \log \pi_{\text{ref}}(y \mid x)\big]$, with π_t the rollout policy: the log-probabilities are computed once per iteration under no-grad, the pairwise target $a = r_t^{\text{used}}(x, y) - r_t^{\text{used}}(x, y')$ is cached with the rollout and held fixed over the four inner epochs, and the same rule applies to all three losses. Both endpoints of a pair are sampled from the current policy ($\mu_t = \pi_t$); generation and the REBEL log-ratio use the same frozen temperature 0.1 (logits divided by the temperature in both), and sequence log-probabilities are sums over generated tokens up to and including EOS; LoRA rank 1024, $\alpha = 2048$,

Table 24: **Powered LLM-scale extension**: seed-level paired differences in mean clean-RM score. Holm-adjusted two-sided p; TOST at the preregistered ± 0.035; outcomes per the frozen five-way classification.

Hypothesis (pair)	mean	sd	90% CI	Holm / TOST p	outcome
H-P1: cont. HL−MSE (n=12)	$+0.0140$	0.0433	$[-0.0085, +0.0365]$	0.60 / 0.061	inconclusive
H-P2: cont. HL−Huber (n=12)	-0.0010	0.0438	$[-0.0237, +0.0217]$	0.94 / **0.011**	**equivalent**
H-P3: clean HL−MSE (n=8)	$+0.0232$	0.0464	$[-0.0079, +0.0542]$	0.60 / 0.25	inconclusive

dropout 0, on the `query_key_value` projections. The reported endpoint is the raw r_{eval} score of each generated response, with no overlong or KL processing; seed-level intervals condition on the fixed 500-prompt evaluation set and do not separately quantify prompt-sampling uncertainty. This is a fixed-target regression within each iteration whose target carries a policy-dependent regularizer, so the LLM runs instantiate the reduction of Section 2 with a KL-regularized reward rather than a fixed deterministic one; seed-level pairing is a block design, not common random numbers (what a run seed fixes, identically for the three losses, is listed in the process record, together with the pilot's variance decomposition and the Stage-0 engineering probe that set the tier). Hypotheses: Holm $\alpha = 0.05$, two-sided, with TOST at the preregistered $\text{SESOI} = 0.035$ clean-reward units (dual-anchored at half the pilot's Huber−MSE benefit and at the pilot's clean-deficit magnitude); outcomes classified by the frozen five-way rule (Table 24).

Readout Table 24. H-P2, contaminated HL−Huber: -0.0010, 90% CI $[-0.0237, +0.0217]$, TOST $p = 0.011$: equivalent within ± 0.035. H-P1, contaminated HL−MSE: $+0.0140$, CI $[-0.0085, +0.0365]$: neither rejects zero nor establishes equivalence (the CI upper end misses the equivalence margin by 0.0015; the frozen rules forbid post hoc widening). H-P3, clean HL−MSE: $+0.0232$, CI $[-0.0079, +0.0542]$: inconclusive, and the pilot's clean-deficit direction does not replicate (the point estimate is *positive*; leave-one-out sign-robust, maximum shift 0.0134). Precision, disclosed: realized seed-level SD 0.043–0.046 implies, after removing the per-seed evaluation-sampling SE of 0.0207 at G=8 (protocol §6), training-side SD 0.038–0.042, the frozen power table's pessimistic row (power ≈0.69 at SESOI). For scale, the sixty arm-level absolute means span $[-0.65, -0.50]$ clean-reward units (per-loss means: MSE -0.577, HL -0.559, Huber -0.564; cross-arm sd 0.033); mean generated length 33.4 (contaminated) / 33.0 (clean): the mean difference is small (no length-equivalence test is reported).

Verdict Per the frozen outcome wordings, the powered LLM-scale cell *bounds the contaminated HL−Huber difference within* ± 0.035 *clean-reward units* (TOST at $\alpha = 0.05$, $n = 12$ seeds); at the tested dose the replacement neither helps nor hurts, the record's one confirmatory LLM-scale statement. H-P1 and H-P3 are inconclusive at the achieved precision, that precision's preregistered outcome, not a selection artifact; the clean-deficit direction's platform scope narrows to the synthetic and MLP platforms. The instrument-calibration failure that closed the main chain's channel is unchanged (two commercial LLM judges failed order-stability and an evaluation reward model failed human agreement before any confirmatory data existed), and the extension's endpoint avoids it by construction: contamination is injected, so the pinned uncontaminated evaluation reward model is the fixed evaluation target, with no judge in the loop. Compute and traceability: process record (H.8).

Support constants and boundary occupancy of the LLM runs. The frozen support artifact of the LLM stack (digest `B0746489...7684`; 32768 iteration-0 SFT pairs, one per prompt) gives $W = 5.797$ ($m = 101$ bins of width $\varsigma = 0.1148$, $\sigma = 0.0861$, $\beta = 0.75$), loss scale $S = 0.01702$, base noise scale $\tau = 0.309$ ($0.3\times$ the clean-gap sd 1.030), and Huber threshold $\delta_H = 2.898$, from the bootstrap UCB 5.539 of the 99.9% quantile; the real-data bridge below re-derives a support ($W = 8.00$) on a different sample and is not the training support. Every arm logged its support-validity window over the 150 iterations: targets beyond $W - 3\sigma$ average $0.09 - -0.15\%$ on the clean arms and $0.54 - -0.56\%$ on the contaminated arms (beyond W: $0.06 - -0.11\%$ and $0.45 - -0.46\%$), predictions beyond $W - 3\sigma$ at most 0.02%, all under the 2% gate. By the occupancy reading of Section 5.5, an occupancy of half a percent, against 3% in the synthetic symmetric cell, predicts contrasts near the ripple level, which is what the extension finds.

A reward-model bridge: the diagnostic on real data (EXT-RMBRIDGE; post-closure control)

Question Whether real relative-reward gaps from a reward model have the tail and length structure the contamination models isolate, and whether the frozen support-calibration recipe leaves the compression zone empty on such data.

Intervention The 2000-pair human-preference validation sample scored by the pinned 2.8B evaluation reward model (the instrument-calibration data of the terminated LLM phase) supplies real relative-reward gaps $a = s(y) - s(y')$. The frozen support-calibration recipe (bootstrap $Q_{99.9}$ UCB $\rightarrow$ derive-support, `llm` profile) is applied to this sample at scale 1, 0.75, and 0.5 (a factor-two miscalibration), and in held-out form (v1.1: 20 random half/half splits, calibrate on one half, evaluate on the other, then swap: 40 readings). No policy update is performed anywhere in this bridge.

Readout Distribution: sd 1.667, skewness -0.09, excess kurtosis $+0.42$, $|z| > 3$ at 0.70% ($2.6\times$ the normal rate). Structure: $\mathrm{corr}(a, \Delta\mathrm{len}) = +0.355$ ($p = 2.5 \times 10^{-60}$); the longer response scores higher in 62.3% of pairs overall (length-tied pairs, $21/2000$, excluded) but in 79% of the top-decile $|a|$ pairs, the length-aligned one-sided tail concentration E2's contamination model isolates. Support: $W = 8.00$, and the compression zone is empty (interior and edge occupancy 0 at scale 1; still 0 at 0.75); the factor-two miscalibration exposes 2.55% / 2.05%, inside the occupancy range where the synthetic phase map's effects operate (Figure 3); held-out calibration keeps the deployed zone effectively empty (mean interior exit 0.08%, worst reading 0.4%) while the halved support averages 4.6% (worst 8.8%). Error geometry from the human labels: the reward model misranks 36.7% of pairs overall but only 16.1% of the top-decile $|a|$ pairs; length alignment is higher among correctly ranked pairs (66.0%) than among misranks (56.4%); of the pairs a halved support would expose, 11.8% are misranks ($P(\text{exposed} \mid \text{misrank}) = 0.8\%$ vs 3.6% given correct).

Verdict On this instrument the ordering of extreme gaps agrees with the human labels more often than average; whether their magnitudes are calibrated, and whether clipping them would help or hurt policy value, is not tested here: ecological plausibility and a deployable diagnostic, not policy validation. Natural reward-model error, preregistered policy training, and a calibrated held-out evaluator have not yet been combined in a single experiment. Descriptive; single instrument and dataset; scores ship with the S2R audit records.

H.5.7 On-policy MLP test of the boundary-bias prediction

Does the boundary-bias prediction hold under $\mu_t = \pi_t$ with a nonlinear potential? (EXT-ONPOLICYMLP; H1–H3 confirmatory under Holm, blocks B–H descriptive; protocol frozen before execution)

Question Whether the boundary-bias prediction of Section 4 holds where the record's two remaining scope restrictions bind: a closed loop ($\mu_t = \pi_t$) and a nonlinear potential class.

Intervention Two controlled increments over the frozen loop, each transcribed from an audited executor: both endpoints of every pair are drawn from the current policy ($\mu_t = \pi_t$; inverse-CDF sampling from the on-policy stream, as in the on-policy sensitivity run), and the potential is $f(x) = W_2 \tanh(W_1 x + b_1)$ with a trunk that persists across rounds and a head re-zeroed each round (as in the persistent-trunk MLP). Everything else is the frozen platform: environment, support and loss-scale constants, the three losses through the same channel, Adam(60, 0.05), $T = 15$, and the uniform-mixture regret against π_{15}^{MD}; noise laws are keyed by pair slot, not by action, so a stream is the same object under every policy. Confirmatory cells: the symmetric mixture law of E1/E3 in the averaging regime (512 pairs per round) with 400 seeds, at k=1 (C1a) and at k=3 (C1b: (W, ς, σ) scaled with the loss scale recalibrated by the frozen rule; the Huber threshold $\delta_H = W/2$ scales with W, as in the support sweep of H.5.3, so the k=3 Huber arm is nearly quadratic on the realized residual range and its contrasts are those of a re-thresholded Huber; the protocol text's remark that only the HL arm depends on the support misstates the text, not the executed code, which transcribes the sweep; ledger entry of 2026-09-08). Registered hypotheses, paired one-sided t tests under Holm: H1 the deficit of Corollary F.2, H2 its attenuation at k=3, H3 HL below residual-space Huber. Descriptive blocks vary the law (rate 0.10; amplitude 20τ; a variance-matched t_3 law), the width ($h \in \{8, 128\}$), the base distribution ($\mu_t = U$), a reward-scale-heterogeneous environment (per-context scale $\exp U(-\ln 3, \ln 3)$ with the support re-derived by the frozen rule: $W' = 1.802$, $\tau' = 0.0926$), and the budget (300 Adam steps; $T = 60$ against π_{60}^{MD}); block B evaluates the fitted potential each round on the 32768 exogenous warmup triples, binned by $|a|$, and averages the signed error $(g - a)\,\mathrm{sign}(a)$ (negative = inward). Gates: process record.

Readout Table 25: all three registered predictions hold (H1 -0.00243, H2 -0.00220, H3 -0.00373; Holm $p \le 2.5 \times 10^{-28}$). In the confirmatory cell the three losses sit at -0.15056 (HL), -0.14683 (Huber), and -0.14813 (MSE); Huber is ahead of MSE by $+0.00130$ (95% CI $[+0.00090, +0.00171]$; descriptive, not a registered hypothesis) under the symmetric mixture, so the ordering is target-space clamping $<$ none $<$ residual-space clamping; H3 compares the two clamping geometries and says nothing by itself about Huber against MSE. At k=3 the three losses are not separated at this precision (HL$-$MSE -0.00023, $199/400$, one-sided $p = 0.12$, 95% CI $[-0.00062, +0.00016]$, within $\pm 0.4\%$ of clean regret; no equivalence margin was registered) and the compression-zone occupancy collapses from 0.0307 to 0.0002. The linear companion cell gives -0.00475 ($0/160$),

the fixed-data E1 magnitude, so feedback alone does not change the deficit and the nonlinear potential halves it. Descriptive blocks (Table 26): the contrast is not monotone in the excursion depth, since doubling the contamination rate deepens the deficit (-0.00440) but doubling the spike amplitude to 20τ, at nearly the same occupancy (0.0591 against 0.0590) and twice the excursion depth (0.057 against 0.025), reverses it ($+0.00484$, $112/160$; Huber$-$MSE $+0.01021$), because spikes far beyond W inflate the variance of the unclipped squared fit more than clipping biases HL, the amplitude reversal of Section 5.4 in the averaging regime; the t_3 law gives a small unresolved deficit (-0.00054); the deficit is present at every width, largest at h=8 (-0.00431, near the linear value) and about half that at $h \geq 32$; the $\mu_t = U$ instantiation agrees in sign; under one decade of reward-scale heterogeneity the three losses match on clean and symmetric rewards (occupancy 0.0091: the support, calibrated to the pooled scale, keeps most contexts interior, and no scale-invariance advantage of the categorical loss appears); on the budget axis the deficit persists at $T = 60$ (-0.00051, one-sided $p = 0.002$) but is null at 300 Adam steps ($+0.00014$; per-seed sd doubles and every loss degrades), the interpolating fit at which the record's benefit lives; the fixed-data E2 benefit itself is not resolved under feedback at 32 pairs ($+0.00052$, $79/160$). Realized target-space bias (block B, Table 27): the linear cell shows the signature of Theorem 6, HL displaced inward by -0.02082 in the outermost interior bin against -0.00346 for MSE and -0.00357 for Huber, stable across rounds and of the order of the population value -0.048 at the interior bound (Table 15); splitting that error into the Bayes term $g^\star(a) - a$ of Theorem 6 and the fit residual $g - g^\star$ attributes -0.01890 of the -0.02082 to the statistical target and -0.00192 to the fit, the latter no larger than the whole MSE and Huber errors, which are fit alone; only in the outermost bin (25 pairs, 0.08% of the evaluation set) does the linear fit fall short of the Bayes displacement (-0.03068 against -0.06665). For the MLP cells this readout is not a bias estimate: the fit is trained on policy-concentrated pairs and evaluated on uniform ones, and the off-distribution squared error dominates (mean e^2 in the same bin 0.172, 0.155, 0.121 for MSE, Huber, HL against 0.0048 for the linear HL fit); an on-policy evaluation set is the appropriate descriptive follow-up.

Verdict The three registered predictions hold under Holm: the deficit of Corollary F.2, its attenuation when the support is widened, and HL below curvature-matched Huber, which clamps in residual space. One maintenance revision (r1) of the output writer, after a NaN from a structurally empty bin, was authorized by the researcher and re-pinned before the rerun; the crashed cell is bitwise-equal on rerun, so it changed no training or statistic (execution record: H.8). Confirmatory (H1–H3); blocks B–H descriptive.

Is the attenuation a property of the support extent alone? (EXT-ONPOLICYMLP-PW; one registered hypothesis, frozen before execution)

Question The k=3 cell rescales the bin width and the Huber threshold together with W; whether the attenuation of H2 is a property of the support extent, of the re-thresholded Huber arm, or of the coarser bins of the scaled cell.

Intervention Outer bins of the frozen width ς are appended to the HL support up to $k \in \{1.51, 2.02, 3.00\}$ ($j \in \{13, 26, 51\}$ bins per side; k=3.00 is the C1b extent) with σ, the interior edges, and $\delta_H = W/2$ bitwise unchanged and the loss scale recalibrated by the frozen rule on the extended interior (as in H.5.3); the HL arm is retrained on the C1a streams and seeds, and the MSE and Huber arms are the C1a records themselves, since neither loss reads the support extent. Registered test: the one-sided paired t on $D'_s = d_s(k{=}1) - d_s(k{=}3)$ at $\alpha = 0.05$. Gates: process record.

Readout HL$-$MSE moves from -0.00243 to -0.00028 ($D' = -0.00216$, 95% CI $[-0.00255, -0.00177]$, $t = -10.9$, one-sided $p = 1.3 \times 10^{-24}$; Wilcoxon 3.4×10^{-22}), not monotonically along k (-0.00108, -0.00007, -0.00028); at k=3.00 HL$-$MSE is -0.00028 ($184/400$, one-sided $p = 0.079$, 95% CI $[-0.00066, +0.00011]$), and the top two rungs are not separated. Along the ladder the compression-zone occupancy falls $0.0307 \to 0.0104 \to 0.0029 \to 0.0001$ and the excursion depth $0.0125 \to 0.0036 \to 0.0009 \to 0.00003$ (each under its own extended interior bound), and HL stays below the C1a Huber arm at every rung (-0.00158 at k=3.00, $p = 5.3 \times 10^{-13}$), which itself sits $+0.00130$ above MSE on this cell (Table 26, PW rows).

Verdict Pure boundary widening on the on-policy MLP platform, outer bins appended up to the k=3 support extent at unchanged bin width, σ, and Huber threshold, with the MSE and Huber arms the C1a records themselves, attenuates the deficit: the attenuation of H2 is a property of the support extent, not of the re-thresholded Huber arm or of the coarser bins of the scaled cell. The cap overrun of this run is disclosed in the process record (H.8). Confirmatory (H2′); the ladder descriptive.

H.6 Engineering validity and the calibration sweep

The pipeline's numerical claims are realized, not assumed: first-round mirror-descent KL medians are 1.1–1.4×10^{-5} (averaging regime) and 1.1–1.2×10^{-3} (interpolation regime), far below the

Table 25: **On-policy MLP test: the three registered hypotheses** (C1a: symmetric law, 512 pairs, $\mu_t = \pi_t$, persistent-trunk MLP h=32, 400 seeds; C1b: the same at k=3). H1 is the deficit of Corollary F.2, H2 its attenuation under support widening (Theorem 6(iii)), H3 HL below residual-space clamping. Paired one-sided t tests, Holm over the three; per-seed sd 0.00407, 0.00426, 0.00362; Wilcoxon and the 20% trimmed mean (-0.00239 for H1) are robustness readouts.

Hypothesis	mean	95% CI	t	raw p	Holm p	Wilcoxon p	decision
H1: HL$-$MSE (C1a)	-0.00243	$[-0.00283, -0.00203]$	-11.9	1.3×10^{-28}	2.5×10^{-28}	5.4×10^{-26}	reject
H2: $d_{k=1} - d_{k=3}$	-0.00220	$[-0.00262, -0.00178]$	-10.3	1.3×10^{-22}	1.3×10^{-22}	2.9×10^{-20}	reject
H3: HL$-$Huber (C1a)	-0.00373	$[-0.00409, -0.00338]$	-20.6	3.8×10^{-65}	1.1×10^{-64}	1.8×10^{-49}	reject

Table 26: **On-policy MLP test: all cells.** Setting lists law, pairs per round n, support scale k, potential class, base distribution μ_t, and Adam steps/T; contrasts are paired per-seed means of perf_final (positive = HL better); $p_<$ is the one-sided p for HL$-$MSE < 0; occ. and exc. are the share of realized targets beyond $W-3\sigma$ and the mean excursion depth $\mathbb{E}[(|\tilde{a}| - (W-3\sigma))_+]$ under the cell's own support. PW rows (pure-widening follow-up): HL retrained with outer bins appended to the stated k at fixed bin width, σ, and δ_H; MSE and Huber are the C1a arms, so Huber$-$MSE repeats C1a's value; occ. and exc. under the extended interior bound $kW-3\sigma$.

Cell	Setting	seeds	HL$-$MSE	pos.	$p_<$	HL$-$Huber	Huber$-$MSE	occ.	exc.
C1a	symmetric; n=512, k=1, MLP$_{32}$, μ_t=π_t, 60/15	400	-0.00243	103/400	1.3e-28	-0.00373	$+0.00130$	0.0307	0.01245
C1b	symmetric; n=512, k=3, MLP$_{32}$, μ_t=π_t, 60/15	400	-0.00023	199/400	0.12	-0.00011	-0.00012	0.0002	0.00006
C2	clean; n=512, k=1, MLP$_{32}$, μ_t=π_t, 60/15	160	$+0.00004$	86/160	0.87	$+0.00007$	-0.00003	0.0006	0.00005
C3	E2 one-sided; n=32, k=1, MLP$_{32}$, μ_t=π_t, 60/15	160	$+0.00052$	79/160	0.70	-0.00061	$+0.00114$	0.0074	0.00238
C4	symmetric; n=512, k=1, linear, μ_t=π_t, 60/15	160	-0.00475	0/160	4.0e-90	-0.00495	$+0.00020$	0.0307	0.01246
C5	symmetric; n=32, k=1, MLP$_{32}$, μ_t=π_t, 60/15	160	$+0.00017$	79/160	0.55	-0.00093	$+0.00110$	0.0311	0.01277
D1	symmetric, rate 0.10; n=512, k=1, MLP$_{32}$, μ_t=π_t, 60/15	160	-0.00440	45/160	6.5e-14	-0.00758	$+0.00318$	0.0590	0.02502
D2	symmetric, 20τ; n=512, k=1, MLP$_{32}$, μ_t=π_t, 60/15	160	$+0.00484$	112/160	1.00	-0.00537	$+0.01021$	0.0591	0.05692
D3	t_3; n=512, k=1, MLP$_{32}$, μ_t=π_t, 60/15	160	-0.00054	73/160	0.083	-0.00125	$+0.00071$	0.0204	0.00581
E128	symmetric; n=512, k=1, MLP$_{128}$, μ_t=π_t, 60/15	160	-0.00246	50/160	1.4e-11	-0.00307	$+0.00061$	0.0306	0.01242
E8	symmetric; n=512, k=1, MLP$_{8}$, μ_t=π_t, 60/15	160	-0.00431	18/160	1.5e-32	-0.00409	-0.00022	0.0306	0.01241
F1	symmetric; n=512, k=1, MLP$_{32}$, μ_t=U, 60/15	160	-0.00247	52/160	4.8e-10	-0.00392	$+0.00145$	0.0445	0.01542
F3	E2 one-sided; n=32, k=1, MLP$_{32}$, μ_t=U, 60/15	160	$+0.00071$	80/160	0.78	$+0.00082$	-0.00011	0.0189	0.00417
G0	clean; scale-het.; n=512, k=1, MLP$_{32}$, μ_t=π_t, 60/15	160	-0.00018	76/160	0.056	-0.00024	$+0.00005$	0.0002	0.00004
G1	symmetric; scale-het.; n=512, k=1, MLP$_{32}$, μ_t=π_t, 60/15	160	$+0.00008$	86/160	0.57	-0.00146	$+0.00154$	0.0091	0.00376
H300a	symmetric; n=512, k=1, MLP$_{32}$, μ_t=π_t, 300/15	160	$+0.00014$	84/160	0.59	$+0.00031$	-0.00017	0.0307	0.01248
H300c	E2 one-sided; n=32, k=1, MLP$_{32}$, μ_t=π_t, 300/15	160	$+0.00013$	85/160	0.56	-0.00018	$+0.00031$	0.0076	0.00255
H60a	symmetric; n=512, k=1, MLP$_{32}$, μ_t=π_t, 60/60	160	-0.00051	63/160	0.002	-0.00218	$+0.00168$	0.0283	0.01137
H60c	E2 one-sided; n=32, k=1, MLP$_{32}$, μ_t=π_t, 60/60	160	-0.00025	83/160	0.35	-0.00045	$+0.00019$	0.0024	0.00078
PW13	symmetric; n=512, appended bins to k=1.51, MLP$_{32}$, μ_t=π_t, 60/15	400	-0.00108	161/400	1.2e-8	-0.00238	$+0.00130$	0.0104	0.00363
PW26	symmetric; n=512, appended bins to k=2.02, MLP$_{32}$, μ_t=π_t, 60/15	400	-0.00007	201/400	0.37	-0.00137	$+0.00130$	0.0029	0.00085
PW51	symmetric; n=512, appended bins to k=3.00, MLP$_{32}$, μ_t=π_t, 60/15	400	-0.00028	184/400	0.079	-0.00158	$+0.00130$	0.0001	0.00003

Table 27: **Realized target-space bias in the linear on-policy cell C4** (symmetric law, 512 pairs, 160 seeds): per bin of $|a|$ on the exogenous warmup triples, the round-averaged inward error $\mathbb{E}[(g-a)\,\mathrm{sign}(a)]$ / mean squared error; bin shares of the evaluation set in the header. The last two rows split the HL inward error into the population Bayes term $g^\star(a)-a$ of Theorem 6 (E1 law at the frozen support, averaged over the warmup pairs of the bin; zero for MSE and Huber, whose population minimizer under mean-zero noise is the clean gap) and the remainder $g-g^\star$; offline companion computation (H.8).

Loss	$\lvert a\rvert \le$ SD 68.44%	(SD, 2SD] 26.89%	(2SD, $W-3\sigma$] 4.60%	$> W-3\sigma$ 0.08%
MSE	-0.00073 / 0.00532	-0.00209 / 0.00575	-0.00346 / 0.00646	-0.00469 / 0.00778
Huber	-0.00074 / 0.00226	-0.00215 / 0.00244	-0.00357 / 0.00273	-0.00456 / 0.00327
HL	-0.00411 / 0.00368	-0.01240 / 0.00409	-0.02082 / 0.00480	-0.03068 / 0.00612
HL: Bayes term $g^\star(a)-a$	-0.00304	-0.00975	-0.01890	-0.06665
HL: fit residual $g-g^\star$	-0.00107	-0.00265	-0.00192	$+0.03597$

preregistered 0.05-nat recovery gate (Proposition 1); the two-sided bridge holds on the live pipeline with 0 violations at a realized inner aggregate ratio $\sum(u-v)^2/\sum \mathrm{KL} = 0.543\,\kappa_{\mathrm{HL}}$ (the calibration sweep's tightness, by contrast, is the pointwise maximum ratio over κ_{HL}); the frozen loss-scale constant $S = 2.296\sigma^2$ sits strictly inside the predicted interval $[2\sigma^2, \kappa_{\mathrm{HL}}]$ (Corollary A.3); the preregistered fairness gates pass (gradient-scale ratio HL/MSE $= 1.096$, Hessian $= 1.000$, Huber/MSE $= 0.937$, all within $[0.8, 1.2]$); runs are CPU-bitwise deterministic, and the v1.5 budget re-anchor (zero change to any scientific definition) was cascade re-emitted bit-identical in $840/840$ cells. Numerical hygiene across all monitored cells: NaN/Inf $= 0$, normalization error 0.0, silent clamping $= 0$ (bin probabilities are evaluated in the log domain—$\log\Phi$ via `log_ndtr`, bin masses via `logdiffexp`, renormalization via `logsumexp`, no probability floor, and a static scan that rejects naive CDF subtraction—and every projected row is asserted to sum to one within the frozen tolerance of the specification; the support-derivation identity $W - 3\sigma = \mathrm{UCB}$ is asserted to $10^{-12}\,W$); the dual-environment test suite is green in both a local CPU and a server GPU environment ($173/173$). The fairness gates are global by design: they certify optimizer-level comparability—loss scale and effective step—over the full stream, not equality of boundary-zone curvature, which the comparison cannot and should not equate: Huber's curvature vanishes beyond δ_H by construction (that is the treatment under test), and the two-hot arm of Section 5.5 supplies the paired control at identical boundary geometry.

The post-closure calibration sweep extends the bridge check across 84 configurations—$W \in \{0.5, 0.915, 2.0\}$, $m \in \{21, 51, 101, 201\}$, $\beta \in \{0.55, 0.6, 0.75, 1.0, 1.5, 2.0\}$, plus a documented out-of-condition $\beta = 0.5$ (where $v(3) < 1/(4\beta^2)$ and κ_{HL} is undefined; the upper inequality still holds there). Result: zero violations of either inequality—the lower across the 72 configurations for which κ_{HL} is defined—with tightness rising monotonically from 0.171–0.186 at $\beta = 0.55$ to 0.931–0.939 at $\beta = 2.0$ and a relative spread across (m, W) of at most 8.1% at fixed β: Theorem 1's claimed m- and W-invariance realized numerically. Table 29 summarizes the gates, diagnostics, and audit verdicts; Figure 5 visualizes the bridge, recovery, and calibration diagnostics, and Figure 6 the sweep.

Theorem-2 terms on the live platform: an assumption-mismatch diagnostic A deterministic replay of the fidelity triplet (nine arms, $40/40$ bitwise per arm) extracts, per round, the realized centered update $\widetilde{A}_t$, the empirical per-round pair-set CE excess $\hat{\varepsilon}_t^{\mathrm{CE}}$ (an empirical instantiation, so named), and the support-violation rate, and evaluates the three right-hand terms of Theorem 2 at the frozen (η, T) with the empirical coverage $C_{\mathrm{emp}} = 133.0$ (comparator: per-context argmax reward; base law: empirical frequency of the frozen action streams, maximum deviation from uniform 6.1×10^{-3}; $D_0 = \ln 100$). Table 28 is a mismatch diagnostic, not a certificate. The design chain of Assumption 3 is exceeded in every cell: the realized $\max|\widetilde{A}_t|$ is $\approx 2.9 \times A_{\mathrm{des}}$ on the clean cells and 6.6–10.1$\times$ under contamination. Since A_{des} enters Theorem 2 only through the support design, what this violation touches is the interior control of updates and predictions. A second, independent mismatch is the exit of realized *targets* from I_δ (occ., $\Pr(|a_t + \xi| > W - \delta)$ over the training pairs, $\xi \equiv 0$ on clean cells; prediction-side exits are recorded with the ablation-suite records): 0.1% of pairs on clean rewards against 1.4–3.3% under contamination; the A_t term itself is instantiated at the

Table 28: **Assumption mismatch and the plug-in expression of Theorem 2 on the live platform** (per cell: maximum realized $|\widetilde{A}_t|$ over A_{des}; mean support-violation rate; mean realized uniform-mixture regret against the per-context reward-maximizing action ($J = 0.4646$ on this instance, against $J = 0.3808$ for the mirror-descent oracle of the primary outcome); then two instantiations of the third term—the empirical CE excess against the realized (noisy) targets with the empirical coverage $C_{\mathrm{emp}} = 133$, and the CE excess against the clean gaps that Theorem 2's a_t denotes with the theoretical coverage $C = 100$; plug-in = mean of the expression; ratio = mean over seeds of the per-seed expression/realized quotient). No cell satisfies the design chain $|\widetilde{A}_t| \le A_{\mathrm{des}}$; the LLM instantiation is deferred to the powered extension's artifacts.

				noisy targets, C_{emp}			clean gaps, $C = 100$		
cell	$\max \lvert\widetilde{A}_t\rvert / A_{\mathrm{des}}$	occ.	realized	$\hat{\varepsilon}^{\,\mathrm{CE}}$	plug-in	ratio	$\hat{\varepsilon}^{\,\mathrm{CE}}$	plug-in	ratio
cont. MSE	7.32	0.0141	0.175	1.89	1.94	11.2	23.0	3.52	20.3
cont. Huber	7.52	0.0141	0.175	1.81	1.96	11.4	23.0	3.56	20.6
cont. HL	6.62	0.0141	0.174	1.55	1.77	10.3	18.4	3.22	18.7
symm. MSE	9.79	0.0327	0.164	3.25	2.65	16.3	37.4	4.70	28.9
symm. Huber	10.14	0.0327	0.164	3.01	2.68	16.5	37.5	4.76	29.4
symm. HL	7.44	0.0327	0.166	2.30	2.19	13.3	27.1	3.98	24.2
clean MSE	2.94	0.0010	0.157	0.03	0.77	4.9	0.034	0.76	4.8
clean Huber	2.94	0.0010	0.157	0.03	0.77	4.9	0.033	0.76	4.8
clean HL	2.97	0.0010	0.157	0.04	0.78	4.9	0.037	0.76	4.9

realized per-round maxima, which Theorem 2 permits, though substituting these maxima does not restore the violated support conditions. Two further caveats: $\hat{\varepsilon}^{\,\mathrm{CE}}$ is an empirical pair-set average and, in the contaminated cells, is measured against the realized (noisy) targets rather than the clean gaps that Theorem 2's a_t denotes, so those rows instantiate the fitting term rather than the theorem's $\varepsilon_t^{\mathrm{CE}}$ (on clean rows the targets coincide, but the sampling law is still Q); and $C_{\mathrm{emp}} = 133.0$ is the empirical-frequency coverage, whereas the exogenous generator is uniform over 100 actions (maximum deviation 6.1×10^{-3}), so the theoretical coverage of the deterministic comparator is 100. With that instantiation the plug-in expression exceeds the realized regret by $\approx 4.9\times$ (clean) and 10–16$\times$ (contaminated); that ordering is a numerical observation about the expression, not a verification of the theorem's applicability. Both caveats are resolved by a recomputation on the same replay (EXT-BOUNDDIAG2, H.8; nine arms bitwise, and the noisy-target $\hat{\varepsilon}$ matched to this record to 10^{-12}): with $\hat{\varepsilon}^{\,\mathrm{CE}}$ measured against the clean gaps that Theorem 2's a_t denotes and the theoretical coverage $C = 100$, the contaminated-cell excess is 11.5–12.7$\times$ larger (18.4–37.5 against 1.55–3.25), because the fitted predictor tracks the spikes it was trained on, and the plug-in exceeds the realized regret by 19–29$\times$ (clean cells: 4.8$\times$, unchanged); Table 28 carries both instantiations, and the mismatch reading above only strengthens under the theorem's own quantities. The LLM instantiation is deferred to the powered extension's artifacts; $R_{\max}$ and the reward scale are not known a priori on that platform—the measured form of the takeaway's narrow-support premise. Independently audited (165 checks, including per-seed bound-term recomputation and cross-artifact realized-regret consistency).

H.7 Per-seed, dose, and mechanism figures

Figures 7, 8, and 9 collect the per-seed contrasts, the dose response, and the mechanism discriminant; the registration-timeline figure is in the process record released with the repository.

H.8 Traceability summary

Every protocol, manifest, and approval is bound by SHA-256 into a chain across specification, environment lock, run manifest, and approval, with an append-only ledger of amendments. Two anchors fix the data, the feature streams of all three losses and the re-emitted E1′ chain (digest `DE4CC04A...F797`) and the targets; the fresh-seed replication executor had to reproduce both bitwise on the registered seeds before its own run was admissible (Section 5.3), and did, and the confirmatory protocols were frozen with their outcome wordings before any new seed existed. Seventeen role-separated internal audit rounds recomputed every reported statistic from raw per-seed or per-prompt records with an independent implementation, with 0 mismatches in every round, re-derived every frozen interpretive branch mechanically from the protocol text, and re-hashed manifests,

Table 29: **Engineering validity.** Preregistered fairness gates, bridge and recovery diagnostics, and determinism; cascade verification and the audit verdicts are process records (H.8).

Check	Realized value	Criterion
Fairness gates (grad / Hessian; vs MSE)	HL 1.096 / 1.000; Huber 0.937 / 1.000	all in $[0.8, 1.2]$
Two-sided bridge	0 violations; inner aggregate ratio $0.543\,\kappa_{\mathrm{HL}}$	0 violations required
First-round MD-KL median	1.1–1.4×10^{-5} (E1, averaging); 1.1–1.2×10^{-3} (E2, interpolation)	≤ 0.05 nats
Numerical hygiene	NaN/Inf $= 0$; normalization error 0.0; silent clamping $= 0$	zeros required
Determinism	CPU-bitwise (pinned threading); test suite $173/173$ in both environments	bitwise / green
Loss-scale constant S	$2.296\sigma^2$ (analytic shortcut $2\sigma^2$ disallowed)	$[2\sigma^2, \kappa_{\mathrm{HL}}]$ $= [2.0, 3.8009]\sigma^2$ (Cor. A.3)
Calibration sweep (post-closure)	84 configs (72 in-condition for the lower bound): 0 violations; tightness spread $\leq 8.1\%$ per β	0 violations required

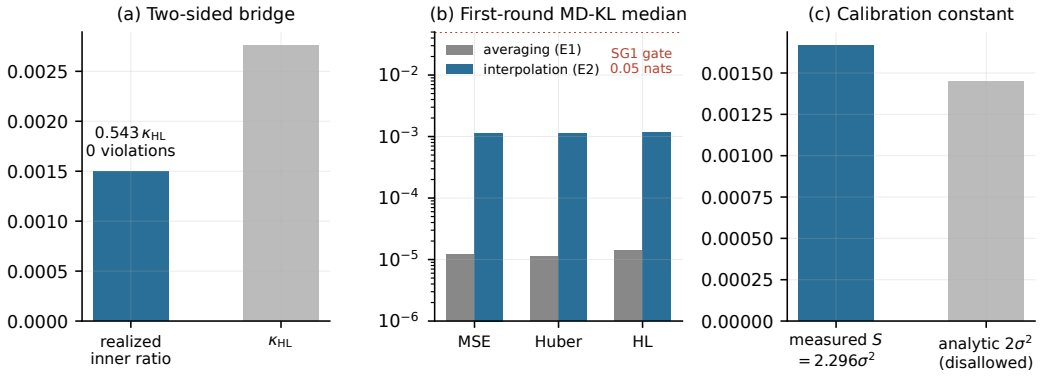


Figure 5: Numerical confirmation of the theory on the live pipeline: (a) the two-sided bridge holds with zero violations and realized inner aggregate ratio $0.543\,\kappa_{\mathrm{HL}}$; (b) first-round mirror-descent KL medians across the three losses—1.1–1.4×10^{-5} in the averaging regime (E1) and 1.1–1.2×10^{-3} in the interpolation regime (E2)—both far below the preregistered 0.05-nat recovery gate (Proposition 1 diagnostics); (c) the frozen loss-scale constant $S = 2/\operatorname{median} \ell''$ is measured at $2.296\sigma^2$, inside the interval $[2\sigma^2, \kappa_{\mathrm{HL}}] = [2.0\sigma^2, 3.8009\sigma^2]$ that the two-sided curvature bound of Corollary A.3 predicts, and 15% away from the disallowed analytic shortcut $2\sigma^2$.

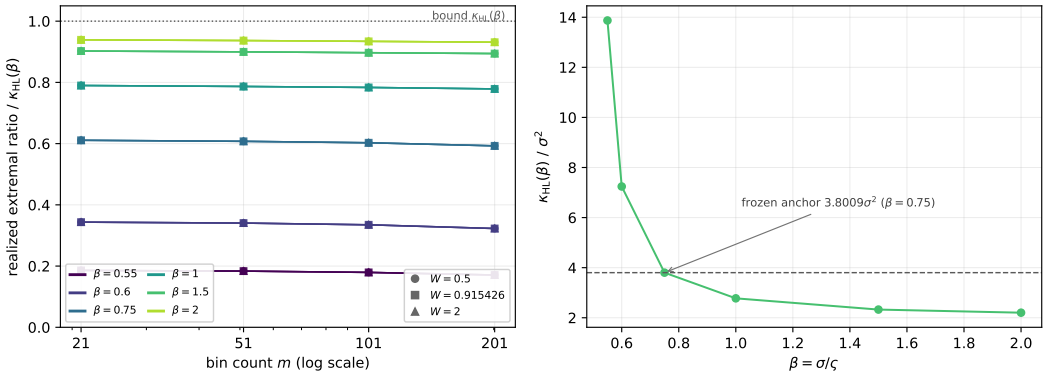


Figure 6: The calibration constant's invariance, swept. Left: realized extremal ratio over $\kappa_{\mathrm{HL}}(\beta)$ (tightness) versus bin count m (log scale) for $\beta \in \{0.55, \ldots, 2.0\}$ and $W \in \{0.5, 0.915, 2.0\}$—flat in m and W at fixed β (spread $\leq 8.1\%$), always below 1, approaching the bound as β grows. Right: $\kappa_{\mathrm{HL}}(\beta)/\sigma^2$ with the frozen pipeline anchor $3.8009\sigma^2$ at $\beta = 0.75$. Zero violations: the upper inequality across all 84 configurations, the lower across the 72 in-condition ones ($\beta \geq 0.55$).

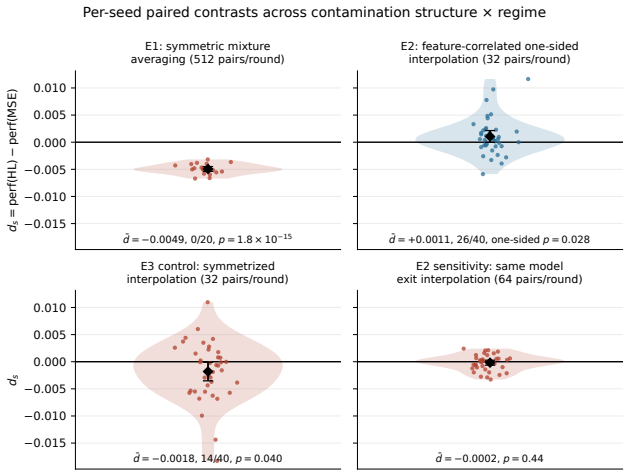
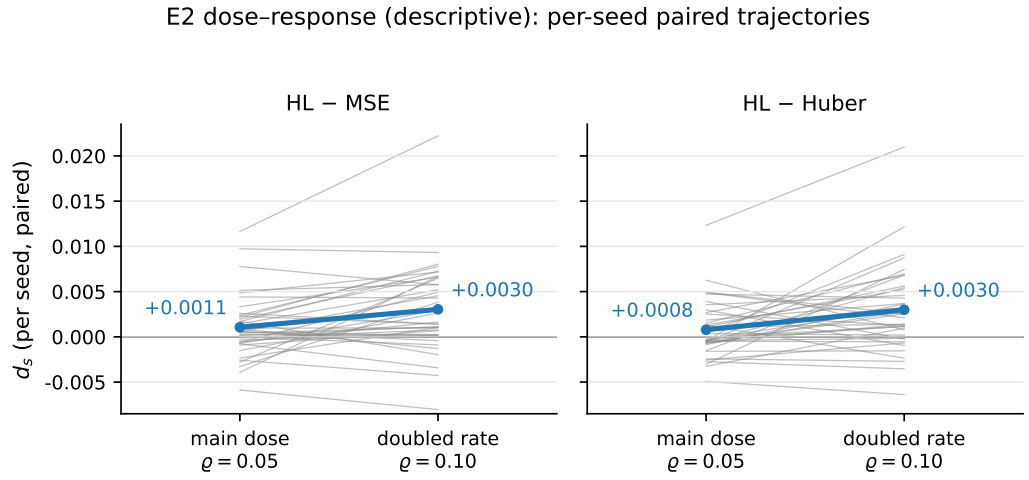


Figure 7: Per-seed paired contrasts $d_s = \mathrm{perf}(\mathrm{HL}) - \mathrm{perf}(\mathrm{MSE})$ across the phase cells; dots seeds, diamonds means, whiskers 95% CIs (values: Table 9). The sign is conditional, not universal.



te row doubles the endpoint mark rate ϱ (0.05 → 0.10) at unchanged spike amplitude 10τ; the frozen protocols label

Figure 8: Per-seed paired trajectories of d_s from the E2 main dose ($\varrho = 0.05$) to the descriptive doubled-rate row ($\varrho = 0.10$, spike amplitude unchanged; thin lines; bold line = mean). The marginal main-dose effect ($+0.0010573$, $26/40$) strengthens to $+0.0030384$ ($33/40$, $t = 3.95$) at the doubled rate; HL−Huber strengthens likewise to $+0.0029946$ ($t = 4.00$). Descriptive, not preregistered-confirmatory.

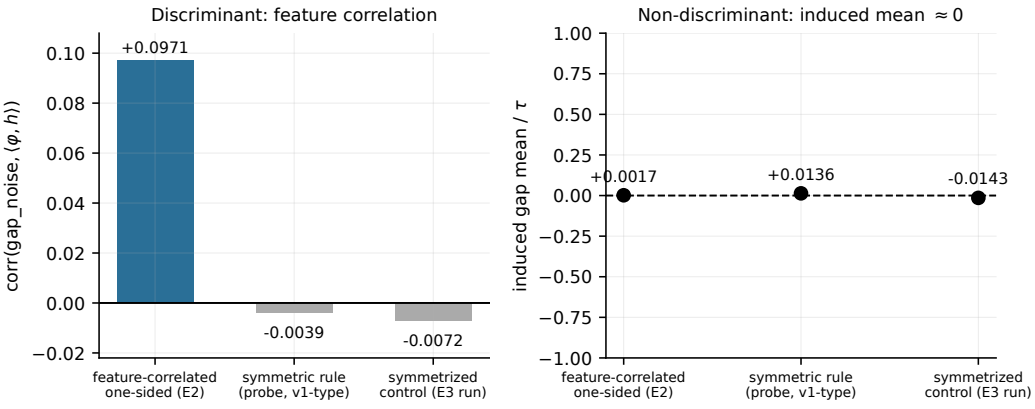


Figure 9: The discriminating statistic across contamination models is feature correlation, not the mean (induced gap mean $+1.3 \times 10^{-4} \approx 0$ in all three): $\mathrm{corr}(\mathrm{gap_noise}, \langle \varphi, h \rangle)$ is $+0.0971$ under the feature-correlated one-sided model versus -0.0039 (v1-type symmetric, probe recomputation on the same frozen streams) and -0.0072 (symmetrized control run), while the induced mean of the pairwise gap is ≈ 0 in all three (E2: $+1.3 \times 10^{-4}$). Before moment matching, feature correlation is the statistic separating the E2 model from both symmetric controls; the diagnostic is descriptive and does not by itself establish attribution. The n=40 moment-matched factorial left the structural moderators unresolved (Table 13); the powered factorial (n=1200, Section 5.4) establishes the conjunction at the frozen dose, and its rate ladder is flat.

pins, and sidecars against disk; the two companion computations (the mixture-bias quantification and the Theorem-6 sandwich) and the two engineering instruments (the bound diagnostic and the checkpoint replay) were audited on the same pattern. The full process record, released with the repository as `PROCESS_RECORD.tex`, lists the protocol, executor, and result digests of every extension, self-test and check counts, execution dates and wall times against their caps (the one cap overrun is disclosed in H.5.7), the maintenance amendment r1 of the on-policy MLP test, the per-round audit inventory, and the seed index set of every experiment (seeds index training streams, so overlapping seed sets are not independent replications).

I EXTENDED RELATED WORK

REBEL (Gao et al., 2024) reduces policy improvement to iterated squared regression in the NPG/mirror-descent lineage (Kakade, 2001; Nemirovski & Yudin, 1983; Schulman et al., 2015; 2017); replacing that loss is its authors' first open question. Farebrother et al. (2024) report broad value-side gains from HL-Gauss cross-entropy, building on Imani & White (2018); $Q^\sharp$ (Zhou et al., 2025) pursues a distributional critic for KL-regularized RL—our categorical object is instead the pairwise regression target of a critic-free reduction. Outside RL, cross-entropy versus square loss (Hui & Belkin, 2021) and regression-as-classification (Stewart et al., 2023) report conditional, task-dependent advantages that our phase diagram parallels on the policy-improvement side. Foster & Krishnamurthy (2021) obtain small-loss guarantees through the log loss, extended by Wang et al. (2023; 2024); these shapes are structurally unavailable here (Corollary C.1), so a categorical policy-side loss cannot rest on small-loss adaptivity—our one weighted statement (Theorem 5) is variance-weighted, conditional, in-family. DRO-REBEL (Sahu & Wells, 2025) robustifies REBEL distributionally while retaining the squared loss, orthogonal and composable; curvature-matched Huber is our robust-regression representative. DPO (Rafailov et al., 2023) and its iterative descendants optimize a logistic link on the same log-ratio object; ours is a real-valued reward gap mapped to an m-bin categorical law; Remark C.3 (Appendix C) shows the small-loss exclusion covers every locally upper-quadratic gap-fitting loss and the Bernoulli logistic excess risk in the supplied-budget format, and leaves open adaptivity to the label-noise level. Length bias—reward models over-scoring longer responses—is the canonical structured over-scoring in RLHF (Singhal et al., 2024); the frozen specification registers our feature-correlated model as its synthetic twin (Section 5.3), a correspondence the reward-model bridge's error geometry later qualifies (Appendix H.5).

Calibration, categorical projections, and robust alignment Theorem 1 sits in the classification-calibration lineage—surrogate-risk transfer through explicit comparison functions (Zhang, 2004; Bartlett et al., 2006; Steinwart, 2007)—instantiated here two-sided, with explicit constants, for the matched binned family derived from the latent truncated-Gaussian exponential family on a compact support. Analyses of categorical distributional RL (Rowland et al., 2018; Lyle et al., 2019) study the value-side categorical projection and its fixed points; ours is the policy-side pairwise regression target, where the projection's boundary, not its expressiveness, is the operative axis. On the robustness side, preference-noise-robust alignment (Chowdhury et al., 2024) modifies the preference-label objective against label flips; our contamination acts on real-valued reward gaps upstream of any label—the regime Remark C.3 leaves open for preference-label reductions.

Interpolation and double descent The registered E2 cell has $n{=}d$, near a nominal interpolation threshold; double-descent and minimum-norm analyses (Belkin et al., 2019; Hastie et al., 2022; Bartlett et al., 2020) motivate the comparison but do not establish a variance peak for this rank-deficient, finite-step platform, where any effective shrinkage of extreme targets can still act as implicit regularization. The suite's regime and ridge branches sharpen the question the 64-pair row raises: the benefit spans the near-interpolation region ($n \leq 36$ at $d = 32$) and fades on exit, and explicit ℓ_2 at the two tested penalties ($\lambda \in \{10^{-3}, 10^{-2}\}$, both harmful) does not reproduce it; weaker penalties were not run on the Adam platform, while in the gradient-descent idealization $\lambda \leq 10^{-3}$ is indistinguishable from the clip (EXT-MINNORM). Whether bounded encoding acts as a per-sample influence cap unavailable to explicit regularization was tested in closed form (EXT-MINNORM, Appendix H.5) and not supported at the level of the gradient-descent idealization; the finite-budget Adam-path mechanism remains the open question (Section 5.4).

Bounded targets in deep RL practice Two-hot value targets are standard in modern model-based agents (Schrittwieser et al., 2020; Hafner et al., 2025); DreamerV3 applies them to symlog-transformed returns; an endpoint transform $\psi(r_i) - \psi(r_j)$ remains a potential difference, whereas Proposition 3 concerns remapping the pairwise gap $\psi(r_i - r_j)$ itself, which admits potentials only if linear. Reward clipping and normalization (Mnih et al., 2015) are the everyday target-space analogue of winsorization; Huber's location estimator (Huber, 1964) is the ancestor of the residual-space, curvature-matched arm.