How Neural Reward Models Learn Features for Policy Optimization: A Single-Index Analysis

Rei Higuchi[1,2]* **Ryotaro Kawata**[1,2] **Akifumi Wachi**[3]
Shokichi Takakura[3] **Kohei Miyaguchi**[3] **Taiji Suzuki**[1,2]
[1]The University of Tokyo [2]RIKEN AIP [3]LY Corporation

Abstract

Reward modeling is not only a prediction problem: in KL-regularized policy optimization, the learned reward is exponentiated to define the deployed policy, so downstream value depends on errors in reward-tilted regions. We study this feedback in a Gaussian single-index model with $r^*(x) = \sigma^*(\langle\theta^*, x\rangle)$ and $x \sim \mathcal{N}(0, I_d)$. We analyze a two-stage neural reward model that first learns the hidden direction θ^* from reward-weighted samples and then fits the readout layer by weighted ridge regression. Exponential reward weighting changes the Hermite signal available to the first layer; for any feature-learning temperature β_1 above a dimension-free $O(1)$ threshold, a constant fraction of neurons recover the hidden direction, with weak-recovery complexity governed by the generative exponent. After feature recovery, we derive tilted-policy value-gap bounds for an idealized label-weighted fit with weights e^{y/β_2} and a more practical surrogate-weighted fit with weights $e^{r_{a_0}(x)/\beta_2}$. Keeping the β_2-dependence explicit yields an admissible set of deployment temperatures, balancing the gain from lowering β_2 against the learning cost amplified by exponential weighting; in the surrogate-weighted case, proxy-dependent factors shrink this admissible set.

1 Introduction

Reward modeling (RM) is a key component of aligning large language models [Christiano et al., 2017, Stiennon et al., 2020, Ouyang et al., 2022], but it differs from ordinary supervised prediction. A learned reward is not merely evaluated on static data; it is optimized to produce a new policy, thereby dictating the distribution on which future rewards are collected and evaluated. In KL-regularized reward maximization, this optimized policy takes the form of an exponential tilt of a reference policy Peng et al. [2019], Korbak et al. [2022], Go et al. [2023], Peters and Schaal [2007]:

$$\pi_\beta(dx) \propto e^{r(x)/\beta}\pi_{\mathrm{ref}}(dx),$$

where r is the reward function and the temperature $\beta > 0$ controls how strongly the policy concentrates on high-reward regions. Here x denotes the evaluated input to the reward model; in language-modeling applications, this input is a prompt-response pair. Consequently, reward modeling faces a policy-induced distribution shift.

Existing theoretical research on offline alignment provides insights into this distribution shift, but typically assumes linear reward models over fixed, known features [Zhu et al., 2023, Ji et al., 2025, Scheid et al., 2024, Foster et al., 2025]. While these settings clarify the interaction between coverage and policy optimization, they typically do not address nonlinear reward landscapes or the feature learning problem inherent to training neural networks. To analyze how neural networks discover

*Correspondence to: higuchi-rei714@g.ecc.u-tokyo.ac.jp.

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

relevant features and approximate nonlinear rewards under policy-induced distribution shifts, we formulate the problem within a Gaussian single-index model (SIM) , where

$$x \sim \pi_{\mathrm{ref}} = \mathcal{N}(0, I_d), \qquad r^*(x) = \sigma^*(\langle \theta^*, x \rangle), \qquad y = r^*(x) + \zeta.$$

Here θ^* is the hidden target direction, σ^* is the target link function, and ζ is the observation noise. We focus on upper-bounded polynomial links, which provide a tractable nonlinear reward class while effectively ruling out extreme scenarios of unboundedly large rewards.

In Gaussian SIMs, recovering the hidden direction is governed by Hermite-based exponents. The Information Exponent (IE) captures the first nonzero Hermite signal available to correlation-based methods, while the Generative Exponent (GE) captures the smallest such exponent obtainable after nonlinear transformations of the target link [Bietti et al., 2022, Lee et al., 2024, Tsiolis et al., 2025]. Recent work shows that suitable nonlinear transformations can reduce the effective learning bottleneck from IE to GE [Lee et al., 2024, Tsiolis et al., 2025, Nishikawa et al., 2025].

Our analysis proceeds in two stages: first-layer feature recovery using a feature-learning temperature β_1, and second-layer nonlinear reward fitting to deploy a policy at a potentially different temperature β_2. The feature-learning temperature β_1 controls the stability of the early feature-learning signal, whereas β_2 governs the downstream tilted policy; the two temperatures need not coincide. For downstream alignment, the relevant metric is not merely prediction risk under the reference distribution, but the value of the policy induced by the learned reward. We study policy-level gaps of the form $\mathbb{E}_{x\sim\pi_{\beta^*}}[r^*(x)] - \mathbb{E}_{x\sim\widehat{\pi}_{\beta_2}}[r^*(x)]$, where π_{β^*} is a target finite-temperature tilted policy and $\widehat{\pi}_{\beta_2}$ is the learned tilted policy induced by the fitted reward. Our main contributions are summarized as follows:

- **Feature Learning under Reward-Induced Distribution Shift.** We characterize first-layer feature learning under exponential reward weighting in a Gaussian SIM. The weighting acts as a nonlinear label transformation and changes the Hermite signal available for recovering the hidden direction, reducing the effective bottleneck from the IE to the GE. Under upper-bounded polynomial assumptions on the target and student links, bounded noise, and standard Guassian SIM assumptions, a constant fraction of neurons recover the hidden direction for any fixed feature-learning temperature β_1 above a dimension-free $O(1)$ threshold. This replaces the dimension-growing polylogarithmic-in-d temperature conditions required in prior clipped-exponential analyses [Nishikawa et al., 2025].
- **From Weighted Prediction to Tilted-Policy Regret.** After feature recovery, we evaluate fitted rewards by the true reward obtained under the deployed tilted policy rather than by prediction risk under the reference distribution. The resulting regret bounds separate temperature mismatch, projected-truncation, and learning terms, and explicitly link the policy-value rates to the local shape of the nonlinear reward maximum.
- **Label- and Surrogate-Weighted Regret Bounds.** This type of bound was derived for two fitting schemes: an idealized label-weighted estimator with weights e^{y/β_2} and a practical surrogate-weighted estimator with weights $e^{r_{a_0}(x)/\beta_2}$ defined by a frozen proxy reward. For both schemes, we prove downstream projected-truncation regret bounds. The surrogate-weighted bound isolates additional proxy-dependent factors that amplify approximation and finite-sample errors when the weighting measure is defined by an imperfect proxy. This identifies a statistical mechanism behind proxy mismatch and reward overoptimization.

2 Related Work

Reward Modeling and Distribution Shift. Reward modeling is a standard component of modern alignment pipelines (RLHF), where a reward signal learned from human feedback is subsequently optimized to produce a policy [Christiano et al., 2017, Ziegler et al., 2019, Stiennon et al., 2020, Ouyang et al., 2022]. This downstream optimization changes the learning problem: models are trained on reference data but evaluated on regions heavily emphasized by the optimized policy. For KL-regularized reward maximization, this policy-induced distribution shift is mathematically formalized as an exponential tilting of the reference policy [Peters and Schaal, 2007, Go et al., 2023, Peng et al., 2019, Korbak et al., 2022]. Consequently, policy-level evaluation translates into an exponentially weighted prediction problem under the reference distribution, drawing natural connections to classical importance weighting under covariate shift [Sugiyama et al., 2007, Gretton et al., 2008].

This distribution shift is also connected to reward overoptimization in offline alignment, where optimizing an imperfect reward signal can exploit proxy errors or move beyond data-supported regions [Gao et al., 2023, Huang et al., 2025].

Alignment Theory under Fixed Representations. The statistical impact of policy-induced distribution shift has been extensively studied in the context of offline alignment and preference-based RL. Recent theoretical works address this mismatch through coverage or optimal-design assumptions, providing suboptimality, regret, sample-complexity, or query-complexity guarantees in fixed-feature linear reward models, contextual bandits, and offline RLHF settings [Zhu et al., 2023, Ji et al., 2025, Scheid et al., 2024, Foster et al., 2025]. While these frameworks clarify the statistical cost of distribution shift under fixed representations, they circumvent the challenge of feature learning and fitting nonlinear link functions. Our work is complementary to this literature: by moving to a nonlinear feature-learning setting, we explicitly analyze the bottleneck that is inherently absent in fixed linear reward classes.

Single-Index Feature Learning and Nonlinear Transformations. A broader line of work studies representation learning beyond fixed kernels or linearization, including one-step feature learning, low-dimensional structure, mean-field training, and spiked-covariance perspectives [Bai and Lee, 2020, Ba et al., 2022, 2023, Mahankali et al., 2023, Wang et al., 2024]. Within this landscape, Gaussian single-index models provide a particularly tractable setting for studying nonlinear feature learning [Dudeja and Hsu, 2018]: targets take the form $f^*(x) = \sigma^*(\langle\theta^*, x\rangle)$ with $x \sim \mathcal{N}(0, I_d)$, and recovery of the hidden direction θ^* is governed by the Hermite coefficients of the link function. The Information Exponent (IE) captures the first nonzero Hermite signal available to correlational methods [Bietti et al., 2022]. Recent work shows that suitable training dynamics or nonlinear label transformations can reduce the effective exponent from IE toward the Generative Exponent (GE) [Lee et al., 2024, Tsiolis et al., 2025, Nishikawa et al., 2025]. Our analysis builds on this IE/GE framework, treating the exponential weighting inherent to reward modeling as a reward-induced nonlinear transformation whose effect is controlled by the policy temperature β.

Downstream Evaluation vs. Prediction Error. In two-layer model analyses, the final step often fits the readout layer after fixing the first-layer features, with guarantees stated in terms of prediction error or population risk [Rahimi and Recht, 2007, Ba et al., 2022, Lee et al., 2024]. While natural for standard regression, this metric does not directly match the objectives of reward modeling, where the learned reward is exponentiated to deploy a policy. To address this discrepancy, we depart from standard prediction bounds and instead evaluate the fitted network via the true reward obtained under the induced policy. By translating weighted prediction guarantees into tilted-policy regret, we use an evaluation criterion that aligns with the practical endpoints of modern alignment pipelines.

3 Problem Setup

Let $d \in \mathbb{N}$, let $\mathbb{S}^{d-1}$ be the unit sphere in $\mathbb{R}^d$, and write $\pi_{\mathrm{ref}} := \mathcal{N}(0, I_d)$. Asymptotic notation is with respect to the ambient dimension and sample sizes; $\lesssim, \gtrsim$ hide constants independent of scaling parameters, and $\widetilde{O}(\cdot), \widetilde{\Theta}(\cdot)$ hide logarithmic factors in the problem parameters. For a probability measure π on $\mathbb{R}^d$, all $L^p(\pi)$ norms and inner products are taken with respect to the explicitly displayed distribution; $\|f\|^p_{L^p(\pi)} := \mathbb{E}_{x\sim\pi}[f(x)^p]$.

We consider a single-index reward model with target direction $\theta^* \in \mathbb{S}^{d-1}$. Fix a noise radius $\tau > 0$. A sample is generated by

$$y := r^*(x) + \zeta, \qquad x \sim \pi_{\mathrm{ref}} = \mathcal{N}(0, I_d), \qquad \zeta \sim \mathrm{Unif}[-\tau, \tau], \tag{1}$$

where $r^*(x) = \sigma^*(\langle\theta^*, x\rangle)$ and ζ is independent of x. Here σ^* is a nonconstant upper-bounded polynomial of degree $q^* \in \mathbb{N}$, and let $B_* := \sup_{u\in\mathbb{R}} \sigma^*(u) < \infty$. Here, the upper boundedness of the target link function allows for the analysis of a more realistic setting by excluding situations where extreme outputs yield infinitely high rewards.

For a measurable reward estimate $\widehat{r} : \mathbb{R}^d \to \mathbb{R}$ and a temperature $\beta > 0$, the ideal population target is the prediction error under the reward-tilted distribution associated with the true reward. For the corresponding tilted distribution π_β induced by r^*, whose density ratio satisfies $d\pi_\beta/d\pi_{\mathrm{ref}} \propto$

$e^{r^*(x)/\beta}$, this target is proportional to the following reference-distribution objective:

$$\mathbb{E}_{x\sim\pi_{\mathrm{ref}},\,\zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[e^{r^*(x)/\beta}\left(y-\widehat{r}(x)\right)^2\right].$$

Since this ratio depends on the unknown reward r^*, we analyze the observable label-weighted analogue

$$\mathcal{L}_{\mathrm{lbl},\beta}[\widehat{r}] := \mathbb{E}_{x\sim\pi_{\mathrm{ref}},\,\zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[e^{y/\beta}\left(y-\widehat{r}(x)\right)^2\right].$$

This loss isolates the exponential weighting effect that later appears in the first-layer oracle and the label-weighted ridge objective.

As the student model, we consider a two-layer network with N neurons:

$$r_a(x;W,b) := \tfrac{1}{N}\textstyle\sum_{j=1}^{N} a_j\sigma(\langle w_j,x\rangle + b_j). \tag{2}$$

where $a=(a_1,\dots,a_N)\in\mathbb{R}^N$, $W=(w_1,\dots,w_N)\in(\mathbb{S}^{d-1})^N$, and $b=(b_1,\dots,b_N)\in\mathbb{R}^N$. We also assume that σ is an upper-bounded polynomial of degree $q\in\mathbb{N}$ with $q\geq q^*$, and let $B:=\sup_{u\in\mathbb{R}}\sigma(u)<\infty$. For the feature-recovery phase, we use the standard random initialization from Gaussian single-index feature-learning analyses Lee et al. [2024], Tsiolis et al. [2025]: the first-layer directions are initialized independently and uniformly on the sphere, $w_j^{(0)}\sim\mathrm{Unif}(\mathbb{S}^{d-1})$, and the initial readout coefficients have independent symmetric signs, $(a_0)_j\sim\mathrm{Unif}\{\pm s_{\mathrm{init}}\}$ with $s_{\mathrm{init}}>0$. The first-layer recovery phase does not use random biases. After feature recovery, we sample the biases used for the second-layer approximation as

$$b_j\sim\mathrm{Unif}[-C_b,C_b],\qquad C_b=\mathrm{polylog}(d),\quad j=1,\dots,N. \tag{3}$$

4 Reward-Weighted Feature Recovery

The training protocol separates feature recovery from coefficient fitting. In this section, feature recovery refers to the phase in which the first-layer directions acquire alignment with the hidden target direction θ^*. Weak recovery means escaping the random-initialization scale and reaching $\langle w_j,\theta^*\rangle\geq c_{\mathrm{wk}}$ for a threshold $c_{\mathrm{wk}}\gtrsim 1/\mathrm{polylog}(d)$; strong recovery refers to the subsequent regime $\langle w_j,\theta^*\rangle\geq 1-\varepsilon$ for an accuracy parameter ε.

During feature recovery, we initialize the second-layer coefficients at $a=a_0\in\mathbb{R}^N$ and keep them fixed, while the directions w_j are trained by online spherical SGD [Arous et al., 2021, Damian et al., 2023, Lee et al., 2024, Arnaboldi et al., 2024] using the bias-free first-layer features. Following standard single-index feature-learning analyses, we study the leading first-layer correlation update induced by the label-weighted objective, which isolates the teacher-side Hermite signal governing early alignment [Lee et al., 2024, Tsiolis et al., 2025]; Appendix A derives this oracle from the weighted squared loss in the small-readout regime. After feature recovery, we sample the biases in (3), freeze $(w_j,b_j)_{j=1}^N$, and fit only $a\in\mathbb{R}^N$ with a weighted ridge objective. We summarize the feature-recovery guarantee here and defer the detailed recovery statements and proofs to Appendix A.

Definition 4.1 (Information and generative exponents)**.** Let $G\sim\mathcal{N}(0,1)$. For $f\in L^2(\mathcal{N}(0,1))$, define $\mathsf{H}_i[f]:=\mathbb{E}_{G\sim\mathcal{N}(0,1)}\left[f(G)\mathrm{He}_i(G)\right]$, and $i\geq 0$, where He_i is the probabilists' Hermite polynomial. The information exponent and generative exponent are

$$\mathrm{IE}(f):=\inf\{i\geq 1:\mathsf{H}_i[f]\neq 0\},\qquad \mathrm{GE}(f):=\inf_{\Phi\in L^2(P_{f(G)})}\mathrm{IE}(\Phi\circ f),$$

where $P_{f(G)}$ is the law of $f(G)$, with the convention that the infimum of the empty set is ∞.

For Gaussian single-index models, feature recovery is governed by the Hermite expansion of the one-dimensional signal along θ^* [Lee et al., 2024, Tsiolis et al., 2025]. Near random initialization, the lowest nonzero Hermite degree of the teacher-side signal, together with the corresponding student-side coefficient, determines the leading drift of $\langle\theta^*,w_j\rangle$ and hence the weak-recovery sample complexity. Let $U_i(\beta_1)$ denote the i-th Hermite coefficient of the teacher-side reward-weighted signal; see Appendix A for its formal definition. To state the teacher-side exponent-reduction bound, let

$$p_{\mathrm{gen}}:=\mathrm{GE}(\sigma^*),\qquad I_*:=\min\left\{I\geq 1:\mathrm{IE}\big((\sigma^*)^I\big)=p_{\mathrm{gen}}\right\}.$$

The following lemma characterizes how this teacher-side signal scales with the temperature β_1.

Lemma 4.2 (Dimension-free exponent reduction by exponential weighting)**.** *For every $\beta_1 > 0$, the lower-degree signals vanish: $U_i(\beta_1) = 0$ for $1 \le i < p_{\mathrm{gen}}$. Moreover, as $\beta_1 \to \infty$,*

$$U_{p_{\mathrm{gen}}}(\beta_1) = \frac{\mathsf{H}_{p_{\mathrm{gen}}}[(\sigma^*)^{I_*}]}{(I_* - 1)!\,\beta_1^{I_*-1}} + O(\beta_1^{-I_*}).$$

Consequently, there exist constants $\underline{\beta}_1 > 0$ and $c_{\mathrm{sig}} > 0$, which are independent of the dimension d, such that for every $\beta_1 \ge \underline{\beta}_1$, the p_{gen}-degree signal satisfies

$$|U_{p_{\mathrm{gen}}}(\beta_1)| \ge c_{\mathrm{sig}} \beta_1^{-(I_*-1)}.$$

The proof is given in Appendix A. With this signal lower bound established, we can now analyze the online SGD process. Lemma 4.2 provides the magnitude of the teacher-side signal. By applying this signal bound to standard Gaussian SIM feature-learning results for random spherical initialization, a constant fraction of neurons are recovered.

Theorem 4.3 (Feature recovery by reward-weighted SGD)**.** *Assume the student-side nondegeneracy condition*

$$V_{p_{\mathrm{gen}}-1} := \mathbb{E}_{h\sim\mathcal{N}(0,1)}[\sigma'(h)\mathrm{He}_{p_{\mathrm{gen}}-1}(h)] \neq 0.$$

Under the random initialization specified in Section 3, for any fixed $\beta_1 \ge \underline{\beta}_1$, the hidden direction θ^* is recovered for a constant fraction of neurons with high probability over the initialization and online samples. The weak-recovery time T_{wk} and the strong-recovery time T_{str} are

$$T_{\mathrm{wk}} = \widetilde{O}\left(\beta_1^{2(I_*-1)} d^{(p_{\mathrm{gen}}-1)\vee 1}\right), \qquad T_{\mathrm{str}} = \widetilde{O}\left(\beta_1^{2(I_*-1)} d\varepsilon^{-2}\right).$$

The precise high-probability statement is given in Appendix A.

Numerical illustration of dimension-free weak recovery. We simulate the online spherical SGD update in Theorem A.1 for a degree-four target with IE $= 3$ and GE $= 1$. Figure 1 shows the weak-recovery rate across six dimensions (d from 32 to 1024) and eight temperatures (β_1 from 1 to 12). For each pair, we run 40 independent trials with initialization conditioned on the single-neuron assumptions of the theorem. The sample budget and learning rate use common constants across all dimensions and temperatures. At every tested dimension, $\beta_1 = 6$ was the smallest tested temperature at which all 40 trials achieved weak recovery. This is consistent with Theorem 4.3, which gives a sufficient temperature for weak recovery that does not increase with dimension. Experimental details are given in Appendix H.

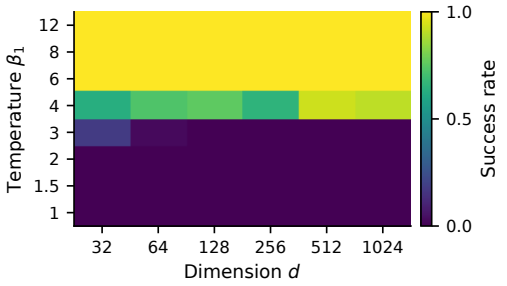


Figure 1: Color shows the fraction of 40 runs with final alignment $|\langle w^{(T)}, \theta^*\rangle| \ge 0.5$.

Note that the threshold $\underline{\beta}_1 = O(1)$ is independent of the dimension d, matching the sample complexity of Lee et al. [2024], Tsiolis et al. [2025], Nishikawa et al. [2025] while improving upon the dimension-growing $\mathrm{polylog}(d)$ bounds required by prior clipping-based analyses Nishikawa et al. [2025]. This is not merely a technical refinement, but addresses a fundamental limitation: an artificially high $\mathrm{polylog}(d)$ temperature would make the exponential weighting nearly uniform, nullifying the benefit of exponent reduction. Lemma 4.2 makes this explicit. The target learning signal scales as $\beta_1^{-(I_*-1)}$, meaning a $\mathrm{polylog}(d)$ temperature would diminish the signal strength to roughly $1/\mathrm{polylog}(d)$. On the other hand, the existence of an $O(1)$ lower bound itself is practically natural. Before the target feature is reliably represented, setting β_1 arbitrarily small would cause the update to undesirably over-concentrate on misleading samples.

The role of the lower bound $\beta_1 \ge \underline{\beta}_1$ is to prevent premature temperature decay during the initial feature-learning phase, rather than to restrict the final deployment policy. In the early stages of training, a premature decay of the temperature would force the model to overfit to a few misleading high-reward samples before the underlying representation is reliably formed. Theorem 4.3 establishes that a dimension-free $O(1)$ temperature is sufficient to prevent this representation bottleneck. Once the hidden direction is successfully recovered, the temperature can be safely lowered to prioritize reward maximization, a regime we analyze in the subsequent section.

5 Tilted-Policy Value Gap After Feature Recovery

5.1 Second-layer weighted ridge setup

Second-layer ridge fitting starts after feature recovery has produced enough recovered features. We condition on this recovery event and treat the first-layer directions as fixed: at least $\tilde{\Theta}(N)$ of the fixed directions satisfy $\langle w_j, \theta^* \rangle \geq 1 - \varepsilon$.

The only remaining optimization variable is the second-layer coefficient vector $a \in \mathbb{R}^N$, which we fit using T_2 fresh ridge-fitting samples $\{(x_i, \zeta_i, y_i)\}_{i=1}^{T_2}$ generated independently according to (1). The goal of this section is to bound the tilted-policy value gap induced by the fitted reward after feature recovery.

In the reward-modeling pipeline, a learned reward $\widehat{r}$ is not used directly; it is exponentiated and combined with the reference distribution to form a tilted policy. Small prediction error under $x \sim \pi_{\mathrm{ref}}$ therefore does not, by itself, control the value of the induced policy, since the exponential weighting can amplify mild prediction errors in regions of large reward. We therefore evaluate the fitted reward through tilted-policy value rather than reference-distribution prediction error [Zhu et al., 2023, Gao et al., 2023, Huang et al., 2025].

In this section, we analyze the downstream policy value using a deployment temperature β_2. Crucially, β_2 can be chosen independently of, and typically strictly smaller than, the feature-recovery temperature β_1. This two-stage setup mirrors practical annealing schedules for KL penalties: a larger β_1 encourages exploration and stable representation learning, while a smaller β_2 allows for aggressive reward maximization once the features are frozen.

We compare against a finite-temperature target policy π_{β^*} rather than an argmax of r^* because in open-ended generation the desired behavior is rarely a single reward mode [Holtzman et al., 2020, Suzgun et al., 2023, Jinnai et al., 2024]; β_2 denotes the temperature at which the fitted reward is deployed, which need not equal β^*. Assuming the policy optimization step is carried out perfectly, we use the exact analytical solution of the KL-regularized reward maximization problem Peters et al. [2010], Korbak et al. [2022], Go et al. [2023] as our deployed policy. For each $\beta_2 > 0$, write

$$\pi_{\beta_2}(dx) := \frac{e^{r^*(x)/\beta_2}}{Z_{\beta_2}} \pi_{\mathrm{ref}}(dx), \qquad Z_{\beta_2} := \mathbb{E}_{x \sim \pi_{\mathrm{ref}}} \left[e^{r^*(x)/\beta_2} \right].$$

For polynomial reward models, the learned tilted policy may be difficult to control globally without an additional restriction on where the policy is evaluated. After feature recovery, the first-layer directions are fixed, and both r^* and the fitted reward models considered below depend on x only through the directions $S := \mathrm{span}\{\theta^*, w_1, \ldots, w_N\}$. Let P_S denote the orthogonal projection onto S. We therefore truncate only the coordinates relevant to the reward model. Fix a projected truncation radius R with $R \gtrsim \sqrt{\dim S}$, where the implicit constant is taken sufficiently large, and write

$$B_R := \{x \in \mathbb{R}^d : \ \|P_S x\|_2 \leq R\}.$$

To isolate the effect of the weighting rule on policy value, we compare label and surrogate weighting on the same set B_R. Including θ^* in S ensures that the true index and each learned index are bounded in absolute value by R on this set. Surrogate weighting is motivated by the practical setting in which a learned reward model supplies the regression weights. In this setting, the weight of a sample can be computed from the frozen proxy r_{a_0}, but membership in B_R generally cannot be determined because θ^* is unknown. Appendix I therefore analyzes surrogate-weighted regression and policy deployment with a truncation set defined by the learned directions alone, and shows how the resulting projection error enters the value-gap bound.

The corresponding truncated target policy is

$$\pi_{\beta_2,R}(dx) := \frac{\mathbf{1}_{B_R}(x) e^{r^*(x)/\beta_2}}{Z_{\beta_2,R}} \pi_{\mathrm{ref}}(dx), \qquad Z_{\beta_2,R} := \mathbb{E}_{x \sim \pi_{\mathrm{ref}}} \left[\mathbf{1}_{B_R}(x) e^{r^*(x)/\beta_2} \right].$$

For a learned reward $\widehat{r}_R$, let $\widehat{\pi}_{\beta_2,R}$ and $\widehat{Z}_{\beta_2,R}$ denote the policy and normalizing constant obtained from the same construction after replacing r^* by $\widehat{r}_R$. The natural performance measure for the fitted reward is the value gap between the target policy at temperature β^* and this learned truncated tilted policy at temperature β_2,

$$\mathcal{R}_R := \mathbb{E}_{x \sim \pi_{\beta^*}}[r^*(x)] - \mathbb{E}_{x \sim \widehat{\pi}_{\beta_2,R}}[r^*(x)].$$

Adding and subtracting $\mathbb{E}_{x\sim\pi_{\beta_2}}[r^*(x)]$ and $\mathbb{E}_{x\sim\pi_{\beta_2,R}}[r^*(x)]$ decomposes $\mathcal{R}_R$ into three conceptually distinct contributions:

$$\mathcal{R}_R = T_{\mathrm{temp}} + T_{\mathrm{cut}}(R) + T_{\mathrm{learn}}(R),$$

where

$$T_{\mathrm{temp}} := \mathbb{E}_{x\sim\pi_{\beta^*}}[r^*(x)] - \mathbb{E}_{x\sim\pi_{\beta_2}}[r^*(x)], \quad T_{\mathrm{cut}}(R) := \mathbb{E}_{x\sim\pi_{\beta_2}}[r^*(x)] - \mathbb{E}_{x\sim\pi_{\beta_2,R}}[r^*(x)],$$

and

$$T_{\mathrm{learn}}(R) := \mathbb{E}_{x\sim\pi_{\beta_2,R}}[r^*(x)] - \mathbb{E}_{x\sim\widehat{\pi}_{\beta_2,R}}[r^*(x)].$$

The first two terms depend only on r^*, β^*, β_2, and R, and describe properties of the target tilted policy that are independent of the learning procedure; we control them in Section 5.2. Only $T_{\mathrm{learn}}(R)$ depends on the second-layer fitting rule, and bounding it through a weighted reward prediction error is the subject of Section 5.3.

5.2 Temperature mismatch and truncation

Define $m_{\beta_2} := \mathbb{E}_{x\sim\pi_{\beta_2}}[r^*(x)]$ and $p_{\beta_2,R} := \pi_{\beta_2}(B_R)$. The identities used in the next two propositions are

$$T_{\mathrm{temp}} = \int_{1/\beta_2}^{1/\beta^*} \mathrm{Var}_{x\sim\pi_{\beta'}}[r^*(x)]\, d\left(\frac{1}{\beta'}\right), \quad T_{\mathrm{cut}}(R) = \frac{1}{p_{\beta_2,R}}\mathbb{E}_{x\sim\pi_{\beta_2}}\left[(r^*(x) - m_{\beta_2})\,\mathbf{1}_{B_R^c}(x)\right].$$

The first two terms in the value-gap decomposition describe the target tilted policy alone, and both can be controlled at low temperature once the local behavior of σ^* near its maximizers is understood. Let $u_1, \ldots, u_m$ be the global maximizers of σ^*, and let p_i be the local order at u_i, meaning $\sigma^*(u_i + t) = B_* - c_i t^{2p_i} + o(t^{2p_i})$ for some $c_i > 0$. Define $p_{\max} := \max_{1\le i\le m} p_i$ and $I_{\max} := \{i : p_i = p_{\max}\}$. For a constant $\kappa > 0$ determined by these local orders and defined in Appendix B, the low-temperature variance expansion is

$$\mathrm{Var}_{x\sim\pi_{\beta_2}}[r^*(x)] = \frac{{\beta_2}^2}{2p_{\max}} + O({\beta_2}^{2+\kappa}),$$

from which the following common bound follows by standard low-temperature and truncation estimates; the proof is given in Appendix C.

Proposition 5.1 (Common low-temperature non-learning terms)**.** *For every sufficiently small $\overline{\beta}_2 > 0$, there exist constants $C_{\mathrm{temp}} > 0$ and $C_{R,\overline{\beta}_2} < \infty$ such that*

$$|T_{\mathrm{temp}}| \le \left(\frac{1}{2p_{\max}} + C_{\mathrm{temp}}{\overline{\beta}_2}^{\kappa}\right)|\beta^* - \beta_2|, \qquad |T_{\mathrm{cut}}(R)| \le C_{R,\overline{\beta}_2}\beta_2,$$

for all $0 < \beta_2, \beta^* \le \overline{\beta}_2$.

5.3 Label/surrogate-weighted ridge fitting

For $a \in \mathbb{R}^N$, write $r_a(x) := \frac{1}{N}\sum_{j=1}^N a_j\sigma(\langle w_j, x\rangle + b_j)$. Given a positive weighting rule $W(x_i, y_i)$, define the weighted ridge estimator

$$\hat{a}_R^W \in \arg\min_{a\in\mathbb{R}^N}\left\{\frac{1}{T_2}\sum_{i=1}^{T_2}\mathbf{1}_{B_R}(x_i)W(x_i,y_i)\,(y_i - r_a(x_i))^2 + \lambda\|a\|_2^2\right\}, \qquad \widehat{r}_R^W := r_{\hat{a}_R^W}.$$

We study two choices of W, denoted by $\mathrm{w} \in \{\mathrm{lbl}, \mathrm{surr}\}$.

- **Label-weighted fitting.** The label-weighted scheme uses the observed reward label in the weight, $W_{\mathrm{lbl}}(x_i, y_i) := e^{y_i/\beta_2}$. This is the cleanest theoretical benchmark for reward-induced exponential weighting. We write $\widehat{r}_R^{\mathrm{lbl}} := \widehat{r}_R^{W_{\mathrm{lbl}}}$ and $\nu_R^{\mathrm{lbl}} := \pi_{\beta_2,R}$. The learned reward induces

  $$\widehat{\pi}_{\beta_2,R}^{\mathrm{lbl}}(dx) := \left(\widehat{Z}_{\beta_2,R}^{\mathrm{lbl}}\right)^{-1}\mathbf{1}_{B_R}(x)e^{\widehat{r}_R^{\mathrm{lbl}}(x)/\beta_2}\pi_{\mathrm{ref}}(dx),$$

  where $\widehat{Z}_{\beta_2,R}^{\mathrm{lbl}}$ is the normalizing constant. This is the learned policy used in $\mathcal{R}_R$ and $T_{\mathrm{learn}}(R)$.

- **Surrogate-weighted fitting.** The surrogate-weighted scheme is closer to plug-in reward-modeling pipelines, where sample weights are computed from a learned surrogate reward rather than from the observed reward labels. Treating the surrogate and the fitted weights as jointly changing makes the objective highly nonconvex and theoretically intractable. A natural way to avoid this is to update the model in blocks: fix the surrogate reward for T_2/K samples, update the target weights, and iterate this process K times.

  For analytical clarity, we state the single-block version, corresponding to $K = 1$: the initial readout vector a_0 from the feature-recovery phase is frozen and treated as fixed throughout the T_2 ridge-fitting samples. This choice keeps the notation and proof structure focused on one weighted ridge problem. The same argument can be applied block by block in the K-block sequential setting. With $r_{a_0}(x) := \frac{1}{N}\sum_{j=1}^{N}(a_0)_j\sigma(\langle w_j, x\rangle + b_j)$, the surrogate-weighted rule is $W_{\mathrm{surr}}(x_i, y_i) := e^{r_{a_0}(x_i)/\beta_2}$. We write $\widehat{r}_R^{\mathrm{surr}} := \widehat{r}_R^{W_{\mathrm{surr}}}$ and define

$$\nu_R^{\mathrm{surr}}(dx) := \left(Z_{\beta_2,R}^{a_0}\right)^{-1} \mathbf{1}_{B_R}(x) e^{r_{a_0}(x)/\beta_2}\pi_{\mathrm{ref}}(dx),$$

  where $Z_{\beta_2,R}^{a_0}$ is the corresponding normalizing constant. Let $\widehat{\pi}_{\beta_2,R}^{\mathrm{surr}}$ be the policy obtained from the same truncated-tilt construction with r_{a_0} replaced by $\widehat{r}_R^{\mathrm{surr}}$, and let $\widehat{Z}_{\beta_2,R}^{\mathrm{surr}}$ denote its normalizing constant. The corresponding surrogate-weighted value gap is

$$\mathcal{R}_R^{\mathrm{surr}} := \mathbb{E}_{x\sim\pi_{\beta^*}}[r^*(x)] - \mathbb{E}_{x\sim\widehat{\pi}_{\beta_2,R}^{\mathrm{surr}}}[r^*(x)].$$

We next relate the weighted prediction error of each fitted reward to the corresponding learning term in the policy value decomposition. For each $\mathrm{w} \in \{\mathrm{lbl}, \mathrm{surr}\}$, define the interpolation path

$$\pi_{\beta_2,t,R}^{\mathrm{w}}(dx) := \left(Z_{\beta_2,t,R}^{\mathrm{w}}\right)^{-1} \mathbf{1}_{B_R}(x) e^{(r^*(x)+t(\widehat{r}_R^{\mathrm{w}}(x)-r^*(x)))/\beta_2}\pi_{\mathrm{ref}}(dx), \qquad t\in[0,1],$$

where $Z_{\beta_2,t,R}^{\mathrm{w}}$ is the normalizing constant, and set

$$\mathcal{D}_{\mathrm{w},R} := \int_0^1 \left\| \frac{d\pi_{\beta_2,t,R}^{\mathrm{w}}}{d\nu_R^{\mathrm{w}}} \right\|_{L^2(\nu_R^{\mathrm{w}})} dt.$$

Lemma 5.2 (Density-ratio bridge for the learning term)**.** *Let* $M_R := \sup_{x\in B_R}|r^*(x)|$. *For each* $\mathrm{w} \in \{\mathrm{lbl}, \mathrm{surr}\}$, *define* $T_{\mathrm{learn}}^{\mathrm{w}}(R) := \mathbb{E}_{x\sim\pi_{\beta_2,R}}[r^*(x)] - \mathbb{E}_{x\sim\widehat{\pi}_{\beta_2,R}^{\mathrm{w}}}[r^*(x)]$. *Then*

$$|T_{\mathrm{learn}}^{\mathrm{w}}(R)| \le 2{\beta_2}^{-1} M_R \mathcal{D}_{\mathrm{w},R} \inf_{c\in\mathbb{R}} \|\widehat{r}_R^{\mathrm{w}} - r^* - c\|_{L^2(\nu_R^{\mathrm{w}})}.$$

The quantity $\mathcal{D}_{\mathrm{w},R}$ measures how well the weighted regression distribution ν_R^{w} covers the interpolation path from the target tilted policy to the learned policy. It plays a similar role to coverage/concentrability quantities in offline alignment analyses [Zhu et al., 2023, Huang et al., 2025].

The infimum over c reflects the fact that tilted policies are invariant to global reward shifts. This shift-invariant form is useful for label weighting: at the population level, the label-weighted squared loss targets $r^* + m_{\zeta,\beta_2}$ rather than r^* (see Appendix D.1), where

$$m_{\zeta,\beta_2} := \mathbb{E}_{\zeta\sim\nu_{\zeta,\beta_2}}[\zeta], \qquad \nu_{\zeta,\beta_2}(d\zeta) := \left(\mathbb{E}_{\zeta'\sim\mathrm{Unif}[-\tau,\tau]}[e^{\zeta'/\beta_2}]\right)^{-1} e^{\zeta/\beta_2}\mathrm{Unif}[-\tau,\tau](d\zeta).$$

This additive shift is absorbed by the infimum over c in Lemma 5.2.

5.4 Value-gap bounds under projected truncation

In this subsection, assume the first-layer recovery condition from the second-layer weighted ridge setup and fix $0 < \beta_2, \beta^* \le \overline{\beta}_2$, where $\overline{\beta}_2 > 0$ is sufficiently small. Define the temperature and truncation contribution

$$\Gamma_{R,\overline{\beta}_2}(\beta_2,\beta^*) := \left(\frac{1}{2p_{\max}} + C_{\mathrm{temp}}{\overline{\beta}_2}^{\kappa}\right)|\beta^*-\beta_2| + C_{R,\overline{\beta}_2}\beta_2.$$

Throughout this subsection, $\lesssim$ hides constants depending only on the fixed problem parameters, the truncation radius R, the frozen first-layer directions, and the upper temperature cutoff $\overline{\beta}_2$ for simplicity.

Theorem 5.3 (Label-weighted projected-truncation value gap)**.** *Define* $\alpha := 1/(2p_{\max})$*. Choose the regularization parameter* λ *at the label-weighted ridge scale specified in* (6)*. Then, with probability at least* $1 - 4\delta_0$*,*

$$|\mathcal{R}_R| \lesssim \Gamma_{R,\overline{\beta}_2}(\beta_2, \beta^*) + M_R \mathcal{D}_{\mathrm{lbl},R} \left[{\beta_2}^{-(\alpha+3)/2} (N^{-1} + \sqrt{\varepsilon}) + {\beta_2}^{-1-(\alpha+1)/4} (T_2\delta_0)^{-1/4} \right].$$

Theorem 5.4 (Surrogate-weighted projected-truncation value gap)**.** *Define*

$$M_{0,R} := \sup_{x \in B_R} r_{a_0}(x), \qquad \mathcal{C}_{0,R}(\beta_2) := \left(Z^{a_0}_{\beta_2,R} \right)^{-1} e^{M_{0,R}/\beta_2}.$$

Choose the regularization parameter λ *at the surrogate-weighted ridge scale specified in* (5)*. Then, with probability at least* $1 - 4\delta_0$*,*

$$|\mathcal{R}^{\mathrm{surr}}_R| \lesssim \Gamma_{R,\overline{\beta}_2}(\beta_2, \beta^*) + M_R \mathcal{D}_{\mathrm{surr},R} \left[\mathcal{C}_{0,R}(\beta_2)^{1/2} {\beta_2}^{-1} (N^{-1} + \sqrt{\varepsilon}) + \mathcal{C}_{0,R}(\beta_2)^{1/4} {\beta_2}^{-1} (T_2\delta_0)^{-1/4} \right].$$

Theorem 5.3 and Theorem 5.4 follow by combining Lemma 5.2 with the common non-learning bound in Proposition 5.1, and then applying the corresponding weighted ridge learning bound.

Combining Theorem 5.4 with Lemma E.5 in Appendix E.1 gives the explicit polynomial upper bound $\mathcal{C}_{0,R}(\beta_2) \lesssim \beta_2^{-d_S}$, where $d_S := \dim S$. More refined local information about the maxima of r_{a_0} can yield a smaller exponent $\alpha_0 \le d_S$; see Remark E.6.

Corollary 5.5 (Polynomial surrogate-weighted value-gap bound)**.** *In the setting of Theorem 5.4, let* $d_S := \dim S$*. Then, for all sufficiently small* $\beta_2 > 0$*, with probability at least* $1 - 4\delta_0$*,*

$$|\mathcal{R}^{\mathrm{surr}}_R| \lesssim \Gamma_{R,\overline{\beta}_2}(\beta_2, \beta^*) + M_R \mathcal{D}_{\mathrm{surr},R} \left[{\beta_2}^{-1-d_S/2} (N^{-1} + \sqrt{\varepsilon}) + {\beta_2}^{-1-d_S/4} (T_2\delta_0)^{-1/4} \right].$$

The three bounds decompose the policy value gap into interpretable components. The common term $\Gamma_{R,\overline{\beta}_2}(\beta_2, \beta^*)$ captures the temperature mismatch and projected-truncation costs, which are independent of the learning rule. The remaining terms are learning costs, scaled by $\mathcal{D}_{\mathrm{lbl},R}$ or $\mathcal{D}_{\mathrm{surr},R}$.

Inside the brackets, the approximation error is proportional to $N^{-1} + \sqrt{\varepsilon}$, reflecting how well r^* can be approximated using the recovered features, while the finite-sample error decays as $(T_2\delta_0)^{-1/4}$.

The difference between label and surrogate weighting is the extra factor $\mathcal{C}_{0,R}(\beta_2)$ in the surrogate learning cost. As defined in Theorem 5.4, $\mathcal{C}_{0,R}(\beta_2)$ upper bounds the density of the surrogate-induced measure ν^{surr}_R relative to π_{ref} on B_R. When this factor is large, the surrogate weights may concentrate mass on a narrow high-surrogate-reward region. This makes the ridge objective emphasize errors on that region and can also reduce the effective sample efficiency. Consequently, both the approximation error and the finite-sample error are amplified in the surrogate-weighted bound.

Choosing the deployment temperature. The value-gap bounds suggest a resource-aware choice of the deployment temperature. Smaller β_2 better matches a low-temperature target policy, but also amplifies the learning terms through exponential weighting and the coverage quantity $\mathcal{D}_{\mathrm{w},R}$. Since $\mathcal{D}_{\mathrm{w},R}$ depends on the fitted reward, we choose β_2 using a deterministic coverage envelope $\overline{\mathcal{D}}_{\mathrm{w},R}$. For a tolerance $\eta > 0$, let $\mathcal{S}^{\mathrm{adm}}_w(\eta) \subset (0, \overline{\beta}_2]$ be the set of temperatures for which the learning-term upper bound is at most η. Choosing $\beta^{\mathrm{w}}_{2,\eta} \in \arg\min_{\beta \in \mathcal{S}^{\mathrm{adm}}_w(\eta)} \Gamma_{R,\overline{\beta}_2}(\beta, \beta^*)$ gives

$$|\mathcal{R}^{\mathrm{w}}_R| \lesssim \inf_{\beta \in \mathcal{S}^{\mathrm{adm}}_w(\eta)} \Gamma_{R,\overline{\beta}_2}(\beta, \beta^*) + \eta.$$

Thus the target temperature is used when it is admissible under the coverage envelope; otherwise one chooses the admissible temperature with the best temperature-mismatch and truncation tradeoff. Appendix G gives explicit forms of $\mathcal{S}^{\mathrm{adm}}_w(\eta)$ and its resource dependence.

Numerical comparison of label and surrogate weighting. We isolate second-layer fitting by using the true direction and a degree-four Hermite basis that represents r^* exactly, with matched temperatures $\beta_2 = \beta^*$. We compare label weights e^{y/β_2} with weights $e^{r_{a_0}(x)/\beta_2}$ from the fixed proxy $r_{a_0}(x) = r^*(x) - 2\langle\theta^*, x\rangle$. Figure 2 compares the sample sizes needed to attain a median absolute learning gap $|T_{\mathrm{learn}}(R)| \leq 0.05$ over 60 trials. At $\beta_2 = 0.5$, the persistent sample threshold for a median gap at most 0.05 is 16 under label weighting and 2048 under surrogate weighting. The additional sample cost in the tested low-temperature conditions is consistent with the proxy-dependent learning cost in Theorem 5.4. Experimental details are given in Appendix H.

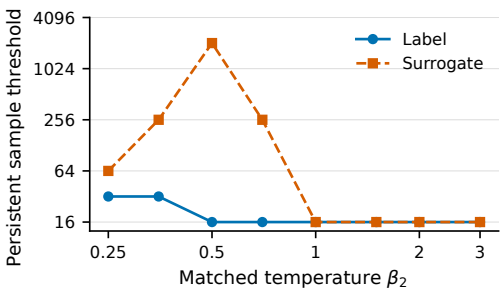


Figure 2: Each point gives the smallest tested T_2 for which the median absolute learning gap is at most 0.05 at T_2 and every larger tested sample size.

6 Conclusion and Discussion

This paper studied reward modeling in a controlled Gaussian single-index setting, with emphasis on how reward prediction interacts with downstream tilted-policy value. The main point is that the relevant statistical target is not only prediction accuracy under $x \sim \pi_{\mathrm{ref}}$, but the tilted-policy value of the learned truncated tilted policy. We analyzed this interaction through first-layer feature recovery under an exponentially weighted reward oracle and second-layer weighted ridge fitting.

For feature recovery, we examined the implications of the IE/GE theory for the reward-induced exponential oracle. The exponential weighting changes the Hermite signal seen by the first layer, and the analysis makes explicit how this signal depends on the policy temperature β_1. In our setting, the lower bound on β_1 is independent of the dimension d. Crucially, while matching the sample complexity of [Lee et al., 2024, Tsiolis et al., 2025, Nishikawa et al., 2025], this improves the polylogarithmic-in-d scale required by clipping-based analyses [Nishikawa et al., 2025] to a fixed $O(1)$ bound. This bound is consistent with the interpretation that, before first-layer feature recovery, too small a temperature may over-emphasize high-reward samples before the relevant feature is reliably represented.

For second-layer ridge fitting, we translated weighted reward prediction guarantees into tilted-policy value bounds. After first-layer feature recovery, the frozen network reduces the remaining problem to a weighted ridge fit over the second-layer coefficients. We analyzed both a label-weighted objective, which serves as a clean theoretical benchmark for reward-induced exponential weighting, and a surrogate-weighted objective, which is closer to the plug-in reward-modeling pipeline. In both cases, the final bounds decompose the tilted-policy value gap into a temperature mismatch term, a projected-truncation term, and a learning term controlled by weighted reward prediction error. The dependence on β_2 is kept explicit throughout the ridge-fitting analysis.

While the Gaussian single-index model is a theoretical abstraction, it provides a foundational characterization of the interaction between reward modeling and policy value. Future work could build upon this framework by relaxing several technical assumptions, such as extending the analysis to richer input distributions or multi-index reward structures beyond the current Gaussian setting. Another promising direction is to remove the projected truncation to further generalize the theoretical results.

Acknowledgments and Disclosure of Funding

RH and RK were partially supported by JSPS KAKENHI (25H01107) and JST BOOST (JPMJBS2418). TS was partially supported by JSPS KAKENHI (24K02905) and JST CREST (JPMJCR2015). This research is supported by the National Research Foundation, Singapore and the Ministry of Digital Development and Information under the AI Visiting Professorship Programme (award number AIVP-2024-004). Any opinions, findings and conclusions or recommendations ex-

pressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and the Ministry of Digital Development and Information.

References

Emmanuel Abbe, Enric Boix Adserà, and Theodor Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. In Gergely Neu and Lorenzo Rosasco, editors, *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*, Proceedings of Machine Learning Research, pages 2552–2623. PMLR, 2023. URL https://proceedings.mlr.press/v195/abbe23a.html.

Luca Arnaboldi, Yatin Dandi, Florent Krzakala, Luca Pesce, and Ludovic Stephan. Repetita iuvant: Data repetition allows SGD to learn high-dimensional multi-index functions. *CoRR*, abs/2405.15459, 2024. doi: 10.48550/ARXIV.2405.15459. URL https://doi.org/10.48550/arXiv.2405.15459.

Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *J. Mach. Learn. Res.*, 22:106:1–106:51, 2021. URL https://jmlr.org/papers/v22/20-1288.html.

Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/f7e7fabd73b3df96c54a320862afcb78-Abstract-Conference.html.

Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. Learning in the presence of low-dimensional structure: A spiked random matrix perspective. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/38a1671ab0747b6ffe4d1c6ef117a3a9-Abstract-Conference.html.

Yu Bai and Jason D. Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rkllGyBFPH.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/3fb6c52aeb11e09053c16eabee74dd7b-Abstract-Conference.html.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D. Lee. Smoothing the landscape boosts the signal for SGD: optimal sample complexity for learning single index models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December*

10 - 16, 2023, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/02763667a5761ff92bb15d8751bcd223-Abstract-Conference.html.

Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, Proceedings of Machine Learning Research, pages 1887–1930. PMLR, 2018. URL http://proceedings.mlr.press/v75/dudeja18a.html.

Dylan J. Foster, Zakaria Mhammedi, and Dhruv Rohatgi. Is a good foundation necessary for efficient reinforcement learning? the computational role of the base model in exploration. In Nika Haghtalab and Ankur Moitra, editors, *The Thirty Eighth Annual Conference on Learning Theory, 30-4 July 2025, Lyon, France*, Proceedings of Machine Learning Research, pages 2026–2142. PMLR, 2025. URL https://proceedings.mlr.press/v291/foster25a.html.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. Aligning language models with preferences through f-divergence minimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 11546–11583. PMLR, 2023. URL https://proceedings.mlr.press/v202/go23a.html.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. In Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors, *Dataset Shift in Machine Learning*. The MIT Press, 12 2008. ISBN 9780262255103. doi: 10.7551/mitpress/7921.003.0013. URL https://doi.org/10.7551/mitpress/7921.003.0013.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D. Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J. Foster. Correcting the mythos of kl-regularization: Direct alignment without overoptimization via chi-squared preference optimization. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=hXm0Wu2U9K.

Kaixuan Ji, Jiafan He, and Quanquan Gu. Reinforcement learning from human feedback with active queries. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=EScatQaRxz.

Yuu Jinnai, Ukyo Honda, Tetsuro Morimura, and Peinan Zhang. Generating diverse and high-quality texts by minimum bayes risk decoding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 8494–8525. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.503. URL https://doi.org/10.18653/v1/2024.findings-acl.503.

Tomasz Korbak, Hady Elsahar, Germán Kruszewski, and Marc Dymetman. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/67496dfa96afddab795530cc7c69b57a-Abstract-Conference.html.

Jason D. Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with SGD near the information-theoretic limit. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/6bd5fca2074dcd9ede9de50f71f7ec28-Abstract-Conference.html.

Arvind V. Mahankali, Haochen Zhang, Kefan Dong, Margalit Glasgow, and Tengyu Ma. Beyond NTK with vanilla gradient descent: A mean-field analysis of neural networks with polynomial width, samples, and time. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/b3748cdac932d91f0a51a37db90dec50-Abstract-Conference.html.

Naoki Nishikawa, Yujin Song, Kazusato Oko, Denny Wu, and Taiji Suzuki. Nonlinear transformers can perform inference-time feature learning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/nishikawa25a.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *CoRR*, abs/1910.00177, 2019. URL http://arxiv.org/abs/1910.00177.

Jan Peters and Stefan Schaal. Reinforcement learning by reward-weighted regression for operational space control. In Zoubin Ghahramani, editor, *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, ACM International Conference Proceeding Series, pages 745–750. ACM, 2007. doi: 10.1145/1273496.1273590. URL https://doi.org/10.1145/1273496.1273590.

Jan Peters, Katharina Mülling, and Yasemin Altun. Relative entropy policy search. In Maria Fox and David Poole, editors, *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*, pages 1607–1612. AAAI Press, 2010. doi: 10.1609/AAAI.V24I1.7727. URL https://doi.org/10.1609/aaai.v24i1.7727.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pages 1177–1184. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

Antoine Scheid, Etienne Boursier, Alain Durmus, Michael I. Jordan, Pierre Ménard, Eric Moulines, and Michal Valko. Optimal design for reward modeling in RLHF. *CoRR*, abs/2410.17055, 2024. doi: 10.48550/ARXIV.2410.17055. URL https://doi.org/10.48550/arXiv.2410.17055.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human

feedback. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.*, 8:985–1005, 2007. doi: 10.5555/1314498.1390324. URL https://dl.acm.org/doi/10.5555/1314498.1390324.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. Follow the wisdom of the crowd: Effective text generation via minimum bayes risk decoding. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Findings of ACL, pages 4265–4293. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.262. URL https://doi.org/10.18653/v1/2023.findings-acl.262.

Konstantinos Christopher Tsiolis, Alireza Mousavi Hosseini, and Murat A. Erdogdu. From information to generative exponent: Learning rate induces phase transitions in SGD. *CoRR*, abs/2510.21020, 2025. doi: 10.48550/ARXIV.2510.21020. URL https://doi.org/10.48550/arXiv.2510.21020.

Zhichao Wang, Denny Wu, and Zhou Fan. Nonlinear spiked covariance matrices and signal propagation in deep neural networks. In Shipra Agrawal and Aaron Roth, editors, *The Thirty Seventh Annual Conference on Learning Theory, June 30 - July 3, 2023, Edmonton, Canada*, Proceedings of Machine Learning Research, pages 4891–4957. PMLR, 2024. URL https://proceedings.mlr.press/v247/wang24b.html.

Banghua Zhu, Michael I. Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 43037–43067. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhu23f.html.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593, 2019. URL http://arxiv.org/abs/1909.08593.

A Reward-weighted feature recovery

This appendix gives the detailed feature-recovery statements summarized by Theorem 4.3. The analysis leverages the online spherical SGD framework of Tsiolis et al. [2025] to treat the exponentially weighted reward oracle induced by our objective.

We use the information exponent, generative exponent, and Hermite-coefficient notation from Definition 4.1. For the target activation, write

$$p_{\mathrm{gen}} := \mathrm{GE}(\sigma^*).$$

Since σ^* is a nonconstant polynomial, the polynomial-link characterization of Lee et al. [2024, Proposition 6 and Lemma 8] implies that this level is attained by a power transformation; see also Tsiolis et al. [2025, Lemma 2.3]. We therefore write

$$I_* := \min\left\{I \geq 1 : \mathrm{IE}\big((\sigma^*)^I\big) = p_{\mathrm{gen}}\right\}.$$

During feature recovery, the second-layer coefficients are kept fixed at a_0, and each direction is updated by online SGD with spherical projection. The update below is the leading correlation part of the label-weighted squared loss gradient near initialization Abbe et al. [2023], Lee et al. [2024], Tsiolis et al. [2025]. Indeed, for

$$\mathcal{L}_{\beta_1}[W, a] := \mathbb{E}_{x\sim\pi_{\mathrm{ref}},\, \zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[e^{y/\beta_1}(y - r_a(x; W))^2\right],$$

the spherical first-layer gradient-descent direction for neuron j, up to the common positive scale $2/N$, is

$$a_j \mathbb{E}_{x,\zeta}\left[e^{y/\beta_1}(y - r_a(x;W))\sigma'(\langle w_j, x\rangle)P_{w_j}^{\perp}x\right].$$

Decomposing $e^{y/\beta_1}(y - r_a(x;W)) = ye^{y/\beta_1} - r_a(x;W)e^{y/\beta_1}$ separates the leading term

$$a_j \mathbb{E}_{x,\zeta}\left[ye^{y/\beta_1}\sigma'(\langle w_j, x\rangle)P_{w_j}^{\perp}x\right]$$

from the correction involving the current network output. In the small-readout regime used for initial feature learning, this leading correlation term is the teacher-side signal analyzed by the single-index oracle. For the coefficient-level statement below, the relevant population oracle is the exponentially weighted label transform

$$G_{\beta_1}(u) := \mathbb{E}_{\zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[(u+\zeta)e^{(u+\zeta)/\beta_1}\right].$$

For each $i \geq 1$, define the teacher-side signal coefficient

$$U_i(\beta_1) := \mathbb{E}_{g\sim\mathcal{N}(0,1)}\left[G_{\beta_1}(\sigma^*(g))\mathrm{He}_i(g)\right].$$

For a fixed second-layer coefficient a, define the student-side coefficient

$$V_{i-1} := \mathbb{E}_{h\sim\mathcal{N}(0,1)}\left[\sigma'(h)\mathrm{He}_{i-1}(h)\right],$$

and set

$$\mu_i(\beta_1, a) := i\, a\, U_i(\beta_1)V_{i-1}.$$

Proof of Lemma 4.2. The label transform G_{β_1} can be written as

$$G_{\beta_1}(u) = e^{u/\beta_1}\left(u\,\mathbb{E}_{\zeta\sim\mathrm{Unif}[-\tau,\tau]}[e^{\zeta/\beta_1}] + \mathbb{E}_{\zeta\sim\mathrm{Unif}[-\tau,\tau]}[\zeta e^{\zeta/\beta_1}]\right).$$

Since σ^* is an upper-bounded polynomial and ζ is bounded, $G_{\beta_1} \in L^2(P_{\sigma^*(g)})$. Therefore $G_{\beta_1}\circ\sigma^*$ is an admissible label transform in the definition of $\mathrm{GE}(\sigma^*)$, and the definition gives $U_i(\beta_1) = 0$ for $1 \leq i < p_{\mathrm{gen}}$.

It remains to identify the first nonzero coefficient. Let $p = p_{\mathrm{gen}}$, $t = \beta_1^{-1}$, and $S = \sigma^*(g) + \zeta$. By Taylor's formula with remainder,

$$Se^{tS} = \sum_{m=0}^{I_*-1}\frac{t^m S^{m+1}}{m!} + R_{I_*}(t,S),$$

where, if $\sigma^*(g) \leq B_*$,

$$|R_{I_*}(t,S)| \leq \frac{t^{I_*}}{I_*!}|S|^{I_*+1}e^{t(B_*+\tau)_+}.$$

The right-hand side is integrable after multiplication by $|\mathrm{He}_p(g)|$, because S is a polynomial of the Gaussian variable plus bounded noise. Hence the remainder contributes $O(\beta_1^{-I_*})$ to $U_p(\beta_1)$.

For $m < I_* - 1$, the term $\mathbb{E}_{\zeta\sim\mathrm{Unif}[-\tau,\tau]}[S^{m+1}]$ is a polynomial in $\sigma^*(g)$ involving only powers $(\sigma^*)^k$ with $k < I_*$. By the definition of I_* and the fact that p is the generative exponent, each such power has zero p-th Hermite coefficient. For $m = I_* - 1$,

$$\mathbb{E}_{\zeta\sim\mathrm{Unif}[-\tau,\tau]}[S^{I_*}] = (\sigma^*)^{I_*} + \text{lower powers of } \sigma^*,$$

and the lower powers again have zero p-th Hermite coefficient. Thus

$$U_p(\beta_1) = \frac{\mathsf{H}_p[(\sigma^*)^{I_*}]}{(I_*-1)!\,\beta_1^{I_*-1}} + O(\beta_1^{-I_*}).$$

The leading coefficient is nonzero by the definition of I_*. The displayed lower bound follows for all sufficiently large β_1. □

Theorem A.1 (Weak recovery for the exponentially weighted first-layer oracle)**.** *For any failure parameter $\delta \in (0,1)$, there exists a constant $C_{\mathrm{wk}} > 0$ such that the following holds. Consider the feature-recovery update with step size $\eta_1 > 0$:*

$$w_j^{(t+1)} = \frac{w_j^{(t)} + \eta_1 (a_0)_j y_t e^{y_t/\beta_1} \sigma'(\langle w_j^{(t)}, x_t \rangle) P^{\perp}_{w_j^{(t)}} x_t}{\left\| w_j^{(t)} + \eta_1 (a_0)_j y_t e^{y_t/\beta_1} \sigma'(\langle w_j^{(t)}, x_t \rangle) P^{\perp}_{w_j^{(t)}} x_t \right\|_2},$$

where (x_t, ζ_t) are i.i.d. samples from $\pi_{\mathrm{ref}} \otimes \mathrm{Unif}[-\tau, \tau]$, $y_t = r^*(x_t) + \zeta_t$, and $P^{\perp}_w := I_d - ww^\top$. Fix a neuron j and suppose that

$$\mu_{p_{\mathrm{gen}}}(\beta_1, (a_0)_j) > 0.$$

If

$$\eta_1 \le C_{\mathrm{wk}} \delta \, \mu_{p_{\mathrm{gen}}}(\beta_1, (a_0)_j) d^{-(p_{\mathrm{gen}}/2 \vee 1)},$$

then, conditional on an initialization satisfying $\langle w_j^{(0)}, \theta^* \rangle \asymp d^{-1/2}$, the iterate reaches weak recovery,

$$\langle w_j^{(T_{\mathrm{wk}})}, \theta^* \rangle \ge c_{\mathrm{wk}}, \qquad c_{\mathrm{wk}} \gtrsim 1/\mathrm{polylog}\, d,$$

with probability at least $1 - \delta$ after

$$T_{\mathrm{wk}} = \widetilde{O}\left(\eta_1^{-1} \mu_{p_{\mathrm{gen}}}(\beta_1, (a_0)_j)^{-1} d^{((p_{\mathrm{gen}}-2)/2) \vee 0} \right)$$

iterations. With the largest stable step size, this becomes

$$T_{\mathrm{wk}} = \widetilde{O}\left(\mu_{p_{\mathrm{gen}}}(\beta_1, (a_0)_j)^{-2} d^{(p_{\mathrm{gen}}-1) \vee 1} \right).$$

Corollary A.2 (Large-fixed-temperature weak recovery dependence on β_1)**.** *Under the assumptions of Theorem A.1, suppose additionally that $|(a_0)_j V_{p_{\mathrm{gen}}-1}| \ge c_{\mathrm{stu}} > 0$ and the sign of $(a_0)_j V_{p_{\mathrm{gen}}-1}$ is compatible with $U_{p_{\mathrm{gen}}}(\beta_1)$. Then for every $\beta_1 \ge \underline{\beta}_1$, the largest-stable-step weak recovery time satisfies*

$$T_{\mathrm{wk}} = \widetilde{O}\left({\beta_1}^{2(I_*-1)} d^{(p_{\mathrm{gen}}-1) \vee 1} \right).$$

Theorem A.3 (Strong recovery after weak recovery)**.** *For any failure parameter $\delta \in (0,1)$, there exists a constant $C_{\mathrm{str}} > 0$ such that the following holds. Work under the same feature-recovery update and coefficient condition as in Theorem A.1. Suppose that a neuron has already reached weak recovery,*

$$\langle w_j^{(0)}, \theta^* \rangle \ge c_{\mathrm{wk}}, \qquad c_{\mathrm{wk}} \gtrsim 1/\mathrm{polylog}\, d.$$

For any $\varepsilon \in (0,1)$, if

$$\eta_1 \le C_{\mathrm{str}} \delta \, d^{-1} \varepsilon \mu_{p_{\mathrm{gen}}}(\beta_1, (a_0)_j) c_{\mathrm{wk}}^{p_{\mathrm{gen}}-1},$$

then the iterate reaches strong recovery,

$$\langle w_j^{(T_{\mathrm{str}})}, \theta^* \rangle \ge 1 - \varepsilon,$$

with probability at least $1 - \delta$ after

$$T_{\mathrm{str}} = \widetilde{O}\left(\eta_1^{-1} \varepsilon^{-1} \mu_{p_{\mathrm{gen}}}(\beta_1, (a_0)_j)^{-1} \right)$$

additional iterations. With the largest stable step size, this gives

$$T_{\mathrm{str}} = \widetilde{O}\left(d \, \varepsilon^{-2} c_{\mathrm{wk}}^{-(p_{\mathrm{gen}}-1)} \mu_{p_{\mathrm{gen}}}(\beta_1, (a_0)_j)^{-2} \right).$$

Corollary A.4 (Large-fixed-temperature strong recovery dependence on β_1)**.** *Under the assumptions of Theorem A.3, suppose additionally that $|(a_0)_j V_{p_{\mathrm{gen}}-1}| \ge c_{\mathrm{stu}} > 0$ and the sign of $(a_0)_j V_{p_{\mathrm{gen}}-1}$ is compatible with $U_{p_{\mathrm{gen}}}(\beta_1)$. Then for every $\beta_1 \ge \underline{\beta}_1$, the largest-stable-step strong recovery time satisfies*

$$T_{\mathrm{str}} = \widetilde{O}\left(d \, \varepsilon^{-2} c_{\mathrm{wk}}^{-(p_{\mathrm{gen}}-1)} \beta_1^{2(I_*-1)} \right).$$

Proof of Theorem 4.3. Let $p = p_{\mathrm{gen}}$. By Lemma 4.2, for every $\beta_1 \geq \underline{\beta}_1$,

$$|U_p(\beta_1)| \gtrsim \beta_1^{-(I_*-1)}.$$

The student-side nondegeneracy assumption $V_{p-1} \neq 0$ ensures that the corresponding student-side Hermite coefficient is nonzero. Under the random initialization specified in Section 3, the standard Gaussian SIM feature-learning guarantee for random spherical initialization therefore applies to the reward-weighted teacher signal. It gives weak recovery for a constant fraction of neurons with high probability, with iteration complexity

$$\widetilde{O}\Big(|U_p(\beta_1)|^{-2} d^{(p-1)\vee 1}\Big) = \widetilde{O}\Big(\beta_1^{2(I_*-1)} d^{(p_{\mathrm{gen}}-1)\vee 1}\Big).$$

Starting from weak recovery, the same feature-learning guarantee gives strong recovery after an additional

$$\widetilde{O}\big(|U_p(\beta_1)|^{-2} d\varepsilon^{-2}\big) = \widetilde{O}\Big(\beta_1^{2(I_*-1)} d\varepsilon^{-2}\Big)$$

iterations. □

B Low-temperature facts for the target tilted policy

This appendix records the one-dimensional low-temperature facts used for the temperature mismatch and projected-truncation bounds in Sections 5.2 and 5.4. Let

$$\phi(u) := (2\pi)^{-1/2} e^{-u^2/2}$$

be the standard Gaussian density on $\mathbb{R}$. For $\beta_2 > 0$, define

$$Z_{\beta_2} := \int_{\mathbb{R}} e^{\sigma^*(u)/\beta_2} \phi(u)\, du, \qquad \mu_{\beta_2}(du) := \frac{e^{\sigma^*(u)/\beta_2}\phi(u)}{Z_{\beta_2}}\, du.$$

Proposition B.1 (Maximizers of an upper-bounded polynomial)**.** *Let $\sigma^* : \mathbb{R} \to \mathbb{R}$ be a nonconstant polynomial that is bounded above, and use the constant $B_* = \sup_{u\in\mathbb{R}} \sigma^*(u)$ from the setup. Then σ^* attains its maximum, and the maximizer set*

$$\mathcal{M} := \{u \in \mathbb{R} : \sigma^*(u) = B_*\}$$

is finite. Writing $\mathcal{M} = \{u_1, \dots, u_m\}$, for each $i \in \{1, \dots, m\}$ there exist $p_i \in \mathbb{N}$ and $c_i > 0$ such that

$$\sigma^*(u_i + t) = B_* - c_i t^{2p_i} + o(t^{2p_i}) \qquad (t \to 0).$$

Proof. Since σ^* is a nonconstant polynomial and is bounded above on $\mathbb{R}$, its degree is even and its leading coefficient is negative. Hence

$$\sigma^*(u) \to -\infty \qquad (|u| \to \infty).$$

Therefore σ^* attains its maximum on a sufficiently large compact interval. If $\mathcal{M}$ were infinite, then the nonzero polynomial $\sigma^*(u) - B_*$ would have infinitely many roots, which is impossible. Thus $\mathcal{M}$ is finite.

Fix $u_i \in \mathcal{M}$. The polynomial $\sigma^*(u_i + t) - B_*$ is not identically zero. Let k_i be the order of its first nonzero Taylor coefficient at $t = 0$. Since u_i is a local maximum, this first nonzero order must be even and its coefficient must be negative. Thus $k_i = 2p_i$ for some $p_i \in \mathbb{N}$, and the coefficient can be written as $-c_i$ with $c_i > 0$. □

Define

$$p_{\max} := \max_{1\leq i\leq m} p_i, \qquad I_{\max} := \{i : \ p_i = p_{\max}\},$$

and

$$\kappa := \min\left\{\frac{1}{2p_{\max}}, \ \min_{i\notin I_{\max}} \left(\frac{1}{2p_i} - \frac{1}{2p_{\max}}\right)\right\},$$

where the second minimum is interpreted as $+\infty$ when $I_{\max} = \{1, \dots, m\}$. Also define, for $i \in I_{\max}$,

$$A_i := \phi(u_i) \int_{\mathbb{R}} e^{-c_i s^{2p_i}}\, ds.$$

Lemma B.2 (Laplace asymptotics and zero-temperature weak limit)**.** *As $\beta_2 \downarrow 0$,*

$$Z_{\beta_2} = e^{B_*/\beta_2} {\beta_2}^{1/(2p_{\max})} \left(\sum_{i \in I_{\max}} A_i + o(1) \right).$$

Moreover,

$$\mu_{\beta_2} \Rightarrow \mu_0 := \sum_{i \in I_{\max}} w_i \delta_{u_i}, \qquad w_i := \frac{A_i}{\sum_{j \in I_{\max}} A_j}.$$

Equivalently, for every bounded continuous function $f : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}_{U \sim \mu_{\beta_2}}[f(U)] \to \sum_{i \in I_{\max}} w_i f(u_i).$$

Proof. For a bounded continuous function $f : \mathbb{R} \to \mathbb{R}$, define

$$N_{\beta_2}[f] := \int_{\mathbb{R}} f(u) e^{\sigma^*(u)/\beta_2} \phi(u)\, du.$$

Thus $Z_{\beta_2} = N_{\beta_2}[1]$.

Choose $\rho > 0$ small enough that the intervals

$$(u_i - \rho, u_i + \rho), \qquad i = 1, \dots, m,$$

are disjoint. We also choose ρ small enough that, for each i, the local expansion in Proposition B.1 implies

$$\sigma^*(u_i + t) \le B_* - \frac{c_i}{2} t^{2p_i} \qquad (|t| < \rho).$$

First consider the contribution near a fixed maximizer u_i. Let

$$\alpha_i := \frac{1}{2p_i}, \qquad g_f(u) := f(u)\phi(u).$$

After the change of variables $u = u_i + {\beta_2}^{\alpha_i} s$, we obtain

$$\begin{aligned} &e^{-B_*/\beta_2} {\beta_2}^{-\alpha_i} \int_{u_i - \rho}^{u_i + \rho} f(u) e^{\sigma^*(u)/\beta_2} \phi(u)\, du \\ &= \int_{|s| < \rho {\beta_2}^{-\alpha_i}} \exp\left(\frac{\sigma^*(u_i + {\beta_2}^{\alpha_i} s) - B_*}{\beta_2} \right) g_f(u_i + {\beta_2}^{\alpha_i} s)\, ds. \end{aligned}$$

For each fixed $s \in \mathbb{R}$, the integrand converges to

$$e^{-c_i s^{2p_i}} g_f(u_i).$$

Moreover, for all sufficiently small β_2 the integrand is dominated by

$$\|g_f\|_{L^\infty(\mathbb{R})} e^{-(c_i/2)s^{2p_i}},$$

which is integrable over $\mathbb{R}$. By dominated convergence,

$$\int_{u_i - \rho}^{u_i + \rho} f(u) e^{\sigma^*(u)/\beta_2} \phi(u)\, du = e^{B_*/\beta_2} {\beta_2}^{1/(2p_i)} \left(f(u_i)\phi(u_i) \int_{\mathbb{R}} e^{-c_i s^{2p_i}}\, ds + o(1) \right).$$

Next consider the complement of these neighborhoods. Since $\sigma^*(u) \to -\infty$ as $|u| \to \infty$ and all global maximizers have been removed, there exists $\eta_\rho > 0$ such that

$$\sigma^*(u) \le B_* - \eta_\rho$$

for all

$$u \notin \bigcup_{i=1}^{m} (u_i - \rho, u_i + \rho).$$

Therefore

$$\left| \int_{\mathbb{R} \setminus \cup_{i=1}^{m} (u_i - \rho, u_i + \rho)} f(u) e^{\sigma^*(u)/\beta_2} \phi(u)\, du \right| \leq \|f\|_{L^\infty(\mathbb{R})} e^{(B_* - \eta_\rho)/\beta_2},$$

which is

$$o\left(e^{B_*/\beta_2} {\beta_2}^{1/(2p_{\max})} \right).$$

If $p_i < p_{\max}$, then

$${\beta_2}^{1/(2p_i)} = o\left({\beta_2}^{1/(2p_{\max})} \right).$$

Combining the local contributions and the complement estimate gives

$$N_{\beta_2}[f] = e^{B_*/\beta_2} {\beta_2}^{1/(2p_{\max})} \left(\sum_{i \in I_{\max}} f(u_i) A_i + o(1) \right).$$

Taking $f \equiv 1$ gives the stated asymptotic for Z_{β_2}. Since $A_i > 0$ for each $i \in I_{\max}$, division by the asymptotic for Z_{β_2} yields

$$\mathbb{E}_{U \sim \mu_{\beta_2}}[f(U)] = \frac{N_{\beta_2}[f]}{Z_{\beta_2}} \to \frac{\sum_{i \in I_{\max}} f(u_i) A_i}{\sum_{j \in I_{\max}} A_j} = \sum_{i \in I_{\max}} w_i f(u_i).$$

This is exactly weak convergence of μ_{β_2} to μ_0. □

Lemma B.3 (Log-partition derivatives and low-temperature variance)**.** *Define, for $\lambda > 0$,*

$$\mathcal{Z}(\lambda) := \int_{\mathbb{R}} e^{\lambda \sigma^*(u)} \phi(u)\, du, \qquad \psi(\lambda) := \log \mathcal{Z}(\lambda).$$

Then

$$\psi'(\lambda) = \mathbb{E}_{U \sim \mu_{1/\lambda}}[\sigma^*(U)], \qquad \psi''(\lambda) = \mathrm{Var}_{U \sim \mu_{1/\lambda}}[\sigma^*(U)].$$

Moreover, as $\lambda \to \infty$,

$$\psi''(\lambda) = \frac{1}{2p_{\max}} \lambda^{-2} + O(\lambda^{-2-\kappa}),$$

where κ is defined above from the local orders of the dominant maximizers of σ^*.

Proof. Differentiation under the integral is justified by the upper bound $\sigma^*(u) \leq B_*$ and the polynomial growth of σ^*. This gives

$$\psi'(\lambda) = \frac{\int_{\mathbb{R}} \sigma^*(u) e^{\lambda \sigma^*(u)} \phi(u)\, du}{\int_{\mathbb{R}} e^{\lambda \sigma^*(u)} \phi(u)\, du} = \mathbb{E}_{U \sim \mu_{1/\lambda}}[\sigma^*(U)]$$

and the usual exponential-family identity

$$\psi''(\lambda) = \mathrm{Var}_{U \sim \mu_{1/\lambda}}[\sigma^*(U)].$$

It remains to identify the leading low-temperature behavior of this variance. Write

$$D(u) := B_* - \sigma^*(u) \geq 0.$$

Since adding the constant B_* does not change variance,

$$\mathrm{Var}_{U \sim \mu_{1/\lambda}}[\sigma^*(U)] = \mathrm{Var}_{U \sim \mu_{1/\lambda}}[D(U)].$$

The same localization argument used in Lemma B.2, now applied to the weighted integrals with factors $D(u)$ and $D(u)^2$, gives

$$\mathbb{E}_{U \sim \mu_{1/\lambda}}[D(U)] = \frac{1}{2p_{\max}} \lambda^{-1} + O(\lambda^{-1-\kappa}),$$

and

$$\mathbb{E}_{U \sim \mu_{1/\lambda}}[D(U)^2] = \left(\frac{1}{2p_{\max}} + \frac{1}{4p_{\max}^2} \right) \lambda^{-2} + O(\lambda^{-2-\kappa}).$$

Indeed, near each dominant maximizer u_i with $i \in I_{\max}$, the change of variables

$$u = u_i + \lambda^{-1/(2p_{\max})} s$$

turns $D(u)$ into $\lambda^{-1} c_i s^{2p_{\max}} + o(\lambda^{-1})$. Under the limiting local density proportional to $e^{-c_i s^{2p_{\max}}}$, the random variable $c_i s^{2p_{\max}}$ has first two moments

$$\frac{1}{2p_{\max}}, \qquad \frac{1}{2p_{\max}} \left(\frac{1}{2p_{\max}} + 1 \right),$$

respectively. The non-dominant maximizers and the complement of the maximizer neighborhoods contribute only to the stated remainder order. Hence

$$\begin{aligned} \mathrm{Var}_{U \sim \mu_{1/\lambda}}[D(U)] &= \mathbb{E}_{U \sim \mu_{1/\lambda}}[D(U)^2] - \left(\mathbb{E}_{U \sim \mu_{1/\lambda}}[D(U)] \right)^2 \\ &= \frac{1}{2p_{\max}} \lambda^{-2} + O(\lambda^{-2-\kappa}). \end{aligned}$$

This proves the asserted expansion for $\psi''(\lambda)$. □

Lemma B.4 (Low-temperature variance asymptotics)**.** *As $\beta_2 \to 0$,*

$$\mathrm{Var}_{x \sim \pi_{\beta_2}}[r^*(x)] = \frac{{\beta_2}^2}{2p_{\max}} + O({\beta_2}^{2+\kappa}).$$

Proof of Lemma B.4. Set $\lambda = 1/\beta_2$. Since $r^*(x) = \sigma^*(\langle \theta^*, x \rangle)$ and $U = \langle \theta^*, x \rangle$ has marginal distribution μ_{β_2} under $x \sim \pi_{\beta_2}$,

$$\mathrm{Var}_{x \sim \pi_{\beta_2}}[r^*(x)] = \mathrm{Var}_{U \sim \mu_{\beta_2}}[\sigma^*(U)] = \psi''(1/\beta_2).$$

Lemma B.3 therefore gives

$$\mathrm{Var}_{x \sim \pi_{\beta_2}}[r^*(x)] = \frac{{\beta_2}^2}{2p_{\max}} + O({\beta_2}^{2+\kappa}),$$

as claimed. □

C Temperature-mismatch and projected-truncation bounds

Building on the low-temperature variance expansion in Appendix B, this appendix proves the two common non-learning controls used in the value-gap theorems: the temperature-mismatch bound and the projected-truncation bound.

Proof of the temperature-mismatch part of Proposition 5.1. By the identity for T_{temp} in Section 5.2,

$$T_{\mathrm{temp}} = \int_{1/\beta_2}^{1/\beta^*} \mathrm{Var}_{x \sim \pi_{\beta'}}[r^*(x)] \, d\left(\frac{1}{\beta'} \right).$$

By Lemma B.4, after choosing $\overline{\beta}_2 > 0$ sufficiently small, there exists $C_{\mathrm{temp}} > 0$ such that

$$\mathrm{Var}_{x \sim \pi_{\beta'}}[r^*(x)] \le \left(\frac{1}{2p_{\max}} + C_{\mathrm{temp}} {\overline{\beta}_2}^{\kappa} \right) (\beta')^2$$

for all $0 < \beta' \le \overline{\beta}_2$. Hence, for all $0 < \beta_2, \beta^* \le \overline{\beta}_2$,

$$\begin{aligned} |T_{\mathrm{temp}}| &\le \left(\frac{1}{2p_{\max}} + C_{\mathrm{temp}} {\overline{\beta}_2}^{\kappa} \right) \left| \int_{1/\beta_2}^{1/\beta^*} s^{-2} \, ds \right| \\ &= \left(\frac{1}{2p_{\max}} + C_{\mathrm{temp}} {\overline{\beta}_2}^{\kappa} \right) |\beta^* - \beta_2|. \end{aligned}$$

□

Lemma C.1 (Uniform projected-truncation bound for the tilted tail probability)**.** *For $R > 0$ and $\beta_2 > 0$, define*

$$\alpha_{\beta_2,R} := \pi_{\beta_2}(B_R^c).$$

If

$$R > \max_{i \in I_{\max}} |u_i|,$$

then

$$\overline{\alpha}_R := \sup_{\beta_2 > 0} \alpha_{\beta_2,R} < 1.$$

Proof. Let

$$S_0 := S \cap (\theta^*)^\perp.$$

Write the projected coordinates as

$$P_S x = U\theta^* + Z, \qquad U := \langle \theta^*, x \rangle, \qquad Z \in S_0.$$

Under $x \sim \pi_{\beta_2}$, the coordinate U has the tilted one-dimensional marginal μ_{β_2}, while $Z \sim \gamma_{S_0}$ is independent of U. Since

$$B_R^c = \{x : \|P_S x\|_2 > R\},$$

we have

$$B_R^c = \{U^2 + \|Z\|_2^2 > R^2\}.$$

Define

$$g_R(u) := \mathbb{P}_{Z \sim \gamma_{S_0}} \left[u^2 + \|Z\|_2^2 > R^2\right].$$

Then g_R is bounded and continuous at each $u_i \in I_{\max}$, and

$$\alpha_{\beta_2,R} = \mathbb{E}_{U \sim \mu_{\beta_2}}[g_R(U)].$$

By Lemma B.2,

$$\alpha_{\beta_2,R} \to \alpha_{0,R} := \sum_{i \in I_{\max}} w_i g_R(u_i).$$

If $R > \max_{i \in I_{\max}} |u_i|$, then $g_R(u_i) < 1$ for every $i \in I_{\max}$, and therefore $\alpha_{0,R} < 1$.

For every fixed $\beta_2 > 0$, we have $\alpha_{\beta_2,R} < 1$ because π_{β_2} has a strictly positive density with respect to π_{ref} and $\pi_{\mathrm{ref}}(B_R) > 0$. To turn this pointwise strict inequality into a uniform one, we control the two endpoints $\beta_2 \downarrow 0$ and $\beta_2 \to \infty$, and then use compactness on the remaining middle range.

The map

$$\beta_2 \mapsto \alpha_{\beta_2,R}$$

is continuous on $(0, \infty)$ because both

$$\mathbb{E}_{x \sim \pi_{\mathrm{ref}}}\left[\mathbf{1}_{B_R^c}(x) e^{r^*(x)/\beta_2}\right] \quad \text{and} \quad \mathbb{E}_{x \sim \pi_{\mathrm{ref}}}\left[e^{r^*(x)/\beta_2}\right]$$

are continuous in $\beta_2 > 0$ by dominated convergence, using the upper bound $r^*(x) \le B_*$. The denominator is strictly positive, so the ratio defining $\alpha_{\beta_2,R}$ is continuous on $(0, \infty)$.

As $\beta_2 \to \infty$, dominated convergence gives

$$\alpha_{\beta_2,R} \to \pi_{\mathrm{ref}}(B_R^c) < 1.$$

Together with $\alpha_{\beta_2,R} \to \alpha_{0,R} < 1$ as $\beta_2 \downarrow 0$, this implies that there exist $0 < \beta_- < \beta_+ < \infty$ and $\eta > 0$ such that

$$\alpha_{\beta_2,R} \le 1 - \eta$$

whenever $\beta_2 \in (0, \beta_-] \cup [\beta_+, \infty)$.

On the compact interval $[\beta_-, \beta_+]$, continuity implies that $\alpha_{\beta_2,R}$ attains its maximum. Since $\alpha_{\beta_2,R} < 1$ for every fixed $\beta_2 > 0$, this maximum is also strictly smaller than 1. Combining the three regions gives

$$\sup_{\beta_2 > 0} \alpha_{\beta_2,R} < 1.$$

□

Proof of the projected-truncation part of Proposition 5.1. Let

$$\alpha_{\beta_2,R} := \pi_{\beta_2}(B_R^c).$$

By the identity for $T_{\mathrm{cut}}(R)$ in Section 5.2,

$$T_{\mathrm{cut}}(R) = \frac{1}{1-\alpha_{\beta_2,R}} \mathbb{E}_{x\sim\pi_{\beta_2}}\left[(r^*(x) - m_{\beta_2})\mathbf{1}_{B_R^c}(x)\right].$$

By Cauchy–Schwarz,

$$|T_{\mathrm{cut}}(R)| \le \frac{\sqrt{\mathrm{Var}_{x\sim\pi_{\beta_2}}[r^*(x)]\,\alpha_{\beta_2,R}}}{1-\alpha_{\beta_2,R}}.$$

Fix a sufficiently small $\overline{\beta}_2 > 0$. Since R is taken sufficiently large, $R > \max_{i\in I_{\max}} |u_i|$. Then Lemma C.1 gives $\overline{\alpha}_R < 1$ such that $\alpha_{\beta_2,R} \le \overline{\alpha}_R$ for all $\beta_2 > 0$. Lemma B.4 implies, after reducing the allowed size of $\overline{\beta}_2$ if necessary, that

$$\mathrm{Var}_{x\sim\pi_{\beta_2}}[r^*(x)] \le C_{\mathrm{var},\overline{\beta}_2}{\beta_2}^2$$

for all $0 < \beta_2 \le \overline{\beta}_2$. Therefore

$$|T_{\mathrm{cut}}(R)| \le \frac{\sqrt{\overline{\alpha}_R C_{\mathrm{var},\overline{\beta}_2}}}{1-\overline{\alpha}_R}\,\beta_2.$$

Absorbing the prefactor into $C_{R,\overline{\beta}_2}$ proves the claim. □

D From weighted prediction to tilted-policy value

This appendix explains why a weighted prediction bound is sufficient for controlling the tilted-policy value gap. The key point is that, along the interpolation path from the true reward to the fitted reward, differentiating the tilted expectation produces a covariance between the true reward and the reward error. The density-ratio factor in Lemma 5.2 is the cost of controlling this covariance by an L^2 prediction error under the weighted regression measure. We first record the label-weighted population shift, then prove a general three-factor bridge inequality, and finally apply its simplified L^2 form to prove Lemma 5.2.

D.1 Label-weighted population shift

For each fixed $x \in B_R$,

$$\begin{aligned}&\mathbb{E}_{\zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[e^{(r^*(x)+\zeta)/\beta_2}\left(r^*(x)+\zeta-f(x)\right)^2\right]\\ &\quad= e^{r^*(x)/\beta_2} Z_\zeta(\beta_2)\left\{\left(f(x)-r^*(x)-m_{\zeta,\beta_2}\right)^2 + \mathrm{Var}_{\zeta\sim\nu_{\zeta,\beta_2}}[\zeta]\right\},\end{aligned}$$

where

$$Z_\zeta(\beta_2) := \mathbb{E}_{\zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[e^{\zeta/\beta_2}\right], \qquad \nu_{\zeta,\beta_2}(d\zeta) := \frac{e^{\zeta/\beta_2}}{Z_\zeta(\beta_2)}\mathrm{Unif}[-\tau,\tau](d\zeta),$$

and

$$m_{\zeta,\beta_2} := \mathbb{E}_{\zeta\sim\nu_{\zeta,\beta_2}}[\zeta].$$

Thus the label-weighted population risk targets $r^* + m_{\zeta,\beta_2}$, an additive shift of the true reward.

Lemma D.1 (Three-factor bridge for the learning term)**.** *Let ν_R be a probability measure supported on B_R, and let $(\pi_t)_{t\in[0,1]}$ be probability measures supported on B_R such that $\pi_t \ll \nu_R$ for every $t \in [0,1]$. Let $\widehat{r}_R : B_R \to \mathbb{R}$ be any learned reward and define*

$$Y_t(x) := r^*(x) - \mathbb{E}_{z\sim\pi_t}[r^*(z)].$$

For Hölder exponents $p_{\mathrm{Holder}} \in (2,\infty]$ and $q_{\mathrm{Holder}} \in [2,\infty)$ satisfying

$$\frac{1}{p_{\mathrm{Holder}}} + \frac{1}{q_{\mathrm{Holder}}} = \frac{1}{2},$$

set

$$\mathcal{A}_{p_{\mathrm{Holder}},q_{\mathrm{Holder}},R}[\nu_R,\pi_\cdot] := \int_0^1 \|Y_t\|_{L^{q_{\mathrm{Holder}}}(\pi_t)} \left\| \left(\frac{d\pi_t}{d\nu_R}\right)^{1-1/q_{\mathrm{Holder}}} \right\|_{L^{p_{\mathrm{Holder}}}(\nu_R)} dt.$$

Then

$$\left| \int_0^1 \mathbb{E}_{x\sim\pi_t}\left[Y_t(x)(\widehat{r}_R(x)-r^*(x))\right] dt \right| \le \mathcal{A}_{p_{\mathrm{Holder}},q_{\mathrm{Holder}},R}[\nu_R,\pi_\cdot] \inf_{c\in\mathbb{R}} \|\widehat{r}_R - r^* - c\|_{L^2(\nu_R)}.$$

The following simplified form follows from the same argument. If

$$M_R := \sup_{x\in B_R} |r^*(x)|$$

and

$$\mathcal{D}_R[\nu_R,\pi_\cdot] := \int_0^1 \left\| \frac{d\pi_t}{d\nu_R} \right\|_{L^2(\nu_R)} dt,$$

then

$$\left| \int_0^1 \mathbb{E}_{x\sim\pi_t}\left[Y_t(x)(\widehat{r}_R(x)-r^*(x))\right] dt \right| \le 2M_R\mathcal{D}_R[\nu_R,\pi_\cdot] \inf_{c\in\mathbb{R}} \|\widehat{r}_R - r^* - c\|_{L^2(\nu_R)}. \tag{4}$$

Proof. Fix $c\in\mathbb{R}$ and write

$$h_c(x) := \widehat{r}_R(x) - r^*(x) - c, \qquad \rho_t(x) := \frac{d\pi_t}{d\nu_R}(x).$$

Since

$$\mathbb{E}_{x\sim\pi_t}[Y_t(x)] = 0,$$

the left-hand side is unchanged if $\widehat{r}_R - r^*$ is replaced by h_c. For each $t\in[0,1]$,

$$\begin{aligned}
|\mathbb{E}_{x\sim\pi_t}[Y_t(x)h_c(x)]| &= |\mathbb{E}_{x\sim\nu_R}\left[Y_t(x)h_c(x)\rho_t(x)\right]| \\
&= \left| \mathbb{E}_{x\sim\nu_R}\left[\left(Y_t(x)\rho_t(x)^{1/q_{\mathrm{Holder}}}\right) h_c(x)\rho_t(x)^{1-1/q_{\mathrm{Holder}}} \right] \right| \\
&\le \|Y_t\|_{L^{q_{\mathrm{Holder}}}(\pi_t)} \|h_c\|_{L^2(\nu_R)} \left\| \rho_t^{1-1/q_{\mathrm{Holder}}} \right\|_{L^{p_{\mathrm{Holder}}}(\nu_R)}.
\end{aligned}$$

Here the last line is the three-factor Hölder inequality under $x\sim\nu_R$, using $1/q_{\mathrm{Holder}} + 1/2 + 1/p_{\mathrm{Holder}} = 1$. Integrating over t and then taking the infimum over $c\in\mathbb{R}$ proves the first claim. It remains to prove (4). Use

$$\begin{aligned}
|\mathbb{E}_{x\sim\pi_t}[Y_t(x)h_c(x)]| &\le \sup_{x\in B_R} |Y_t(x)| \mathbb{E}_{x\sim\nu_R}[|h_c(x)|\rho_t(x)] \\
&\le 2M_R \|h_c\|_{L^2(\nu_R)} \|\rho_t\|_{L^2(\nu_R)},
\end{aligned}$$

where

$$\sup_{x\in B_R} |Y_t(x)| \le 2M_R,$$

because π_t is supported on B_R. □

We now apply the simplified L^2 form of the preceding lemma to the interpolation paths defined in Subsection 5.3. This gives the bridge inequality stated in Lemma 5.2.

Proof of Lemma 5.2. Fix $\mathrm{w}\in\{\mathrm{lbl},\mathrm{surr}\}$, and write

$$\pi_t := \pi^{\mathrm{w}}_{\beta_2,t,R}, \qquad \nu_R := \nu^{\mathrm{w}}_R, \qquad \widehat{r}_R := \widehat{r}^{\mathrm{w}}_R.$$

By differentiating the interpolation path,

$$T^{\mathrm{w}}_{\mathrm{learn}}(R) = -\frac{1}{\beta_2}\int_0^1 \mathbb{E}_{x\sim\pi_t}\left[\left(r^*(x) - \mathbb{E}_{z\sim\pi_t}[r^*(z)]\right)\left(\widehat{r}_R(x) - r^*(x)\right)\right] dt.$$

Thus

$$|T^{\mathrm{w}}_{\mathrm{learn}}(R)| \le \frac{1}{\beta_2}\left|\int_0^1 \mathbb{E}_{x\sim\pi_t}\left[Y_t(x)(\widehat{r}_R(x)-r^*(x))\right]dt\right|,$$

where

$$Y_t(x) := r^*(x) - \mathbb{E}_{z\sim\pi_t}[r^*(z)].$$

Applying the simplified bound (4) from Lemma D.1, with

$$\mathcal{D}_R[\nu_R,\pi_\cdot] = \mathcal{D}_{\mathrm{w},R},$$

gives

$$|T^{\mathrm{w}}_{\mathrm{learn}}(R)| \le \frac{2M_R\mathcal{D}_{\mathrm{w},R}}{\beta_2}\inf_{c\in\mathbb{R}}\|\widehat{r}_R - r^* - c\|_{L^2(\nu_R)}.$$

This is the desired bound. □

E Weighted ridge estimates after feature recovery

This appendix proves the finite-sample weighted ridge estimates used by the label-weighted and surrogate-weighted learning bounds. For brevity, write $\hat{r}_R := \widehat{r}^{\mathrm{lbl}}_R$ and $\hat{r}^{(0)}_R := \widehat{r}^{\mathrm{surr}}_R$ in this appendix. For the vector notation used in this appendix, define

$$\psi_N(x) := \frac{1}{N}\left(\sigma(\langle w_j, x\rangle + b_j)\right)_{j=1}^N$$

so that $r_a(x) = \langle a, \psi_N(x)\rangle$. We first collect the weighted moment notation used in the exact bounds. For label weighting, define

$$Z_\zeta := \mathbb{E}_{\zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[e^{\zeta/\beta_2}\right], \qquad M_{*,R} := \sup_{x\in B_R} r^*(x),$$

and

$$m_{\zeta,\beta_2} := \mathbb{E}_{\zeta\sim\nu_{\zeta,\beta_2}}[\zeta], \qquad \nu_{\zeta,\beta_2}(d\zeta) := \frac{e^{\zeta/\beta_2}}{Z_\zeta}\mathrm{Unif}[-\tau,\tau](d\zeta).$$

Set

$$r^*_{\mathrm{lbl},\beta_2}(x) := r^*(x) + m_{\zeta,\beta_2}.$$

The unnormalized label-weighted moments are

$$M_{2,\mathrm{wt},R} := \mathbb{E}_{x\sim\pi_{\mathrm{ref}},\,\zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[\mathbf{1}_{B_R}(x)e^{2(r^*(x)+\zeta)/\beta_2}(\zeta-m_{\zeta,\beta_2})^2\|\psi_N(x)\|_2^2\right],$$

$$M_{4,\mathrm{wt},R} := \mathbb{E}_{x\sim\pi_{\mathrm{ref}},\,\zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[\mathbf{1}_{B_R}(x)e^{2(r^*(x)+\zeta)/\beta_2}\|\psi_N(x)\|_2^4\right],$$

and

$$G_{4,\mathrm{wt},R} := \mathbb{E}_{x\sim\pi_{\mathrm{ref}},\,\zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[\mathbf{1}_{B_R}(x)e^{2(r^*(x)+\zeta)/\beta_2}|r^*_{\mathrm{lbl},\beta_2}(x)|^4\right].$$

For surrogate weighting, define

$$M^{(0)}_{2,\mathrm{wt},R} := \mathbb{E}_{x\sim\pi_{\mathrm{ref}},\,\zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[\mathbf{1}_{B_R}(x)e^{2r_{a_0}(x)/\beta_2}\zeta^2\|\psi_N(x)\|_2^2\right],$$

$$M^{(0)}_{4,\mathrm{wt},R} := \mathbb{E}_{x\sim\pi_{\mathrm{ref}},\,\zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[\mathbf{1}_{B_R}(x)e^{2r_{a_0}(x)/\beta_2}\|\psi_N(x)\|_2^4\right],$$

and

$$G^{(0)}_{4,\mathrm{wt},R} := \mathbb{E}_{x\sim\pi_{\mathrm{ref}},\,\zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[\mathbf{1}_{B_R}(x)e^{2r_{a_0}(x)/\beta_2}|r^*(x)|^4\right].$$

Lemma E.1 (Second-layer comparators from recovered first-layer features)**.** *Work under the first-layer recovery condition from the second-layer weighted ridge setup. Let $P_{T_2} = T_2^{-1}\sum_{i=1}^{T_2}\delta_{x_i}$ be the empirical distribution of the ridge-fitting inputs. Then, for each of the two targets*

$$r_{\mathrm{target}} \in \{r^*,\ r^*_{\mathrm{lbl},\beta_2}\}, \qquad r^*_{\mathrm{lbl},\beta_2}(x) := r^*(x) + m_{\zeta,\beta_2},$$

there exists a comparator $a^\sharp \in \mathbb{R}^N$ such that

$$\|r_{a^\sharp} - r_{\mathrm{target}}\|^2_{L^2(P_{T_2})} \lesssim N^{-2} + \varepsilon, \qquad \|a^\sharp\|_2^2 \lesssim N.$$

The implicit constants may depend on the fixed reward and activation classes and on τ, but not on $N, T_2, \lambda, \delta_0$.

Proof. Under the first-layer recovery condition, the unit vectors satisfy

$$\|w_j - \theta^*\|_2 = \sqrt{2(1 - \langle w_j, \theta^* \rangle)} \le \sqrt{2\varepsilon}.$$

Both r^* and $r^*_{\mathrm{lbl},\beta_2} = r^* + m_{\zeta,\beta_2}$ are finite-degree ridge-polynomial targets. Apply Lee et al. [2024, Appendix B.6, Lemma 22] to each target with distance tolerance $\sqrt{2\varepsilon}$. □

For the abstract weighted ridge estimate, fix a measurable weight $W_R(x,\zeta) \ge 0$ supported on B_R, a regression target $r_{\mathrm{target}} : B_R \to \mathbb{R}$, and a residual $\xi_R(x,\zeta)$ such that the observed label can be written as

$$y = r_{\mathrm{target}}(x) + \xi_R(x,\zeta).$$

We use the ridge-fitting sample $(x_i, \zeta_i, y_i)_{i=1}^{T_2}$ defined in the Section 5.1. Assume

$$0 < Z_W := \mathbb{E}_{x\sim\pi_{\mathrm{ref}},\, \zeta\sim\mathrm{Unif}[-\tau,\tau]}[W_R(x,\zeta)] < \infty.$$

Define the probability measure ν_W on B_R by

$$\nu_W(dx) := \frac{\mathbb{E}_{\zeta\sim\mathrm{Unif}[-\tau,\tau]}[W_R(x,\zeta)]}{Z_W}\pi_{\mathrm{ref}}(dx).$$

Define the unnormalized weighted moments

$$M_{2,W,R} := \mathbb{E}_{x\sim\pi_{\mathrm{ref}},\, \zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[W_R(x,\zeta)^2\xi_R(x,\zeta)^2\|\psi_N(x)\|_2^2\right],$$

$$M_{4,W,R} := \mathbb{E}_{x\sim\pi_{\mathrm{ref}},\, \zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[W_R(x,\zeta)^2\|\psi_N(x)\|_2^4\right],$$

and

$$G_{4,W,R} := \mathbb{E}_{x\sim\pi_{\mathrm{ref}},\, \zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[W_R(x,\zeta)^2|r_{\mathrm{target}}(x)|^4\right].$$

For a measurable $h : B_R \to \mathbb{R}$, write

$$\|h\|^2_{W,T_2} := \frac{1}{T_2}\sum_{i=1}^{T_2} W_R(x_i,\zeta_i)h(x_i)^2, \qquad \|h\|^2_{W,\mathrm{pop}} := \mathbb{E}_{x\sim\pi_{\mathrm{ref}},\, \zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[W_R(x,\zeta)h(x)^2\right].$$

Lemma E.2 (Abstract truncated weighted ridge prediction bound)**.** *Assume the following weighted residual orthogonality:*

$$\mathbb{E}_{x\sim\pi_{\mathrm{ref}},\, \zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[W_R(x,\zeta)\xi_R(x,\zeta)\psi_N(x)\right] = 0.$$

Let

$$\hat{a}_W \in \arg\min_{a\in\mathbb{R}^N}\left\{\frac{1}{T_2}\sum_{i=1}^{T_2} W_R(x_i,\zeta_i)(y_i - r_a(x_i))^2 + \lambda\|a\|_2^2\right\}.$$

For any comparator $a^\sharp \in \mathbb{R}^N$, set

$$\mathcal{E}_{W,T_2}(a^\sharp) := \|r_{a^\sharp} - r_{\mathrm{target}}\|^2_{W,T_2}.$$

There exists a universal constant $C > 0$ such that, with probability at least $1 - 4\delta_0$ over the ridge-fitting sample, if

$$\lambda \ge C\sqrt{\frac{M_{4,W,R}}{T_2\delta_0}},$$

then

$$\|r_{\hat{a}_W} - r_{\mathrm{target}}\|^2_{L^2(\nu_W)} \lesssim \frac{1}{Z_W}\left(\mathcal{E}_{W,T_2}(a^\sharp) + \frac{M_{2,W,R} + M_{4,W,R} + G_{4,W,R}}{T_2\lambda\delta_0} + \sqrt{\frac{G_{4,W,R}}{T_2\delta_0}} + \lambda\|a^\sharp\|_2^2\right).$$

Proof. The argument is the weighted analogue of the second-layer ridge proof of Lee et al. [2024, Appendix B.6, Lemma 20]; the same unweighted step is also used as a black-box prediction-error argument in Tsiolis et al. [2025, Proposition B.10]. In this proof, write $\mathbb{E}_{\mathrm{sample}}$ for expectation over

$$(x_i,\zeta_i)_{i=1}^{T_2} \sim (\pi_{\mathrm{ref}} \otimes \mathrm{Unif}[-\tau,\tau])^{T_2}.$$

First, ridge optimality compares $\hat{a}_W$ with $a^\sharp$. Since $y = r_{\text{target}} + \xi_R$, expanding the two weighted empirical objectives gives

$$\|r_{\hat{a}_W} - r_{\text{target}}\|_{W,T_2}^2 + \lambda\|\hat{a}_W\|_2^2 \le \mathcal{E}_{W,T_2}(a^\sharp) + 2\langle a^\sharp - \hat{a}_W, V\rangle + \lambda\|a^\sharp\|_2^2,$$

where

$$V := \frac{1}{T_2}\sum_{i=1}^{T_2} W_R(x_i,\zeta_i)\xi_R(x_i,\zeta_i)\psi_N(x_i).$$

The weighted residual orthogonality implies

$$\mathbb{E}_{\text{sample}}[V] = 0, \qquad \mathbb{E}_{\text{sample}}[\|V\|_2^2] \le \frac{M_{2,W,R}}{T_2}.$$

Thus Markov's inequality and Young's inequality imply, with probability at least $1-\delta_0$,

$$\|r_{\hat{a}_W} - r_{\text{target}}\|_{W,T_2}^2 + \frac{\lambda}{2}\|\hat{a}_W\|_2^2 \le \mathcal{E}_{W,T_2}(a^\sharp) + C\frac{M_{2,W,R}}{T_2\lambda\delta_0} + \frac{3\lambda}{2}\|a^\sharp\|_2^2.$$

Second, compare the weighted empirical seminorm with the weighted population seminorm. Define

$$\Sigma_T := \frac{1}{T_2}\sum_{i=1}^{T_2} W_R(x_i,\zeta_i)\psi_N(x_i)\psi_N(x_i)^\top, \qquad \Sigma := \mathbb{E}_{x\sim\pi_{\text{ref}},\,\zeta\sim\text{Unif}[-\tau,\tau]}\left[W_R(x,\zeta)\psi_N(x)\psi_N(x)^\top\right],$$

$$u_T := \frac{1}{T_2}\sum_{i=1}^{T_2} W_R(x_i,\zeta_i)r_{\text{target}}(x_i)\psi_N(x_i), \qquad u := \mathbb{E}_{x\sim\pi_{\text{ref}},\,\zeta\sim\text{Unif}[-\tau,\tau]}\left[W_R(x,\zeta)r_{\text{target}}(x)\psi_N(x)\right],$$

and

$$s_T := \frac{1}{T_2}\sum_{i=1}^{T_2} W_R(x_i,\zeta_i)r_{\text{target}}(x_i)^2, \qquad s := \mathbb{E}_{x\sim\pi_{\text{ref}},\,\zeta\sim\text{Unif}[-\tau,\tau]}\left[W_R(x,\zeta)r_{\text{target}}(x)^2\right].$$

Then

$$\|r_{\hat{a}_W} - r_{\text{target}}\|_{W,T_2}^2 = \|r_{\hat{a}_W} - r_{\text{target}}\|_{W,\text{pop}}^2 + \hat{a}_W^\top(\Sigma_T - \Sigma)\hat{a}_W - 2\hat{a}_W^\top(u_T - u) + (s_T - s).$$

The three deviations are controlled by the moments defined above:

$$\mathbb{E}_{\text{sample}}[\|\Sigma_T - \Sigma\|_F^2] \le \frac{M_{4,W,R}}{T_2},$$

$$\mathbb{E}_{\text{sample}}[\|u_T - u\|_2^2] \le \frac{M_{4,W,R} + G_{4,W,R}}{2T_2},$$

where $2|r_{\text{target}}(x)|^2\|\psi_N(x)\|_2^2 \le |r_{\text{target}}(x)|^4 + \|\psi_N(x)\|_2^4$, and

$$\mathbb{E}_{\text{sample}}[(s_T - s)^2] \le \frac{G_{4,W,R}}{T_2}.$$

Applying Markov's inequality to these three bounds and using Young's inequality for the cross term gives, on an event of probability at least $1-3\delta_0$,

$$\begin{aligned}\|r_{\hat{a}_W} - r_{\text{target}}\|_{W,\text{pop}}^2 \le{}& \|r_{\hat{a}_W} - r_{\text{target}}\|_{W,T_2}^2 + \sqrt{\frac{M_{4,W,R}}{T_2\delta_0}}\|\hat{a}_W\|_2^2 \\ &+ \frac{\lambda}{4}\|\hat{a}_W\|_2^2 + C\frac{M_{4,W,R} + G_{4,W,R}}{T_2\lambda\delta_0} + \sqrt{\frac{G_{4,W,R}}{T_2\delta_0}}.\end{aligned}$$

Combining this with the ridge-optimality bound and taking a union bound yields, with probability at least $1-4\delta_0$,

$$\begin{aligned}&\|r_{\hat{a}_W} - r_{\text{target}}\|_{W,\text{pop}}^2 + \left(\frac{\lambda}{4} - \sqrt{\frac{M_{4,W,R}}{T_2\delta_0}}\right)\|\hat{a}_W\|_2^2 \\ &\quad\lesssim \mathcal{E}_{W,T_2}(a^\sharp) + \frac{M_{2,W,R} + M_{4,W,R} + G_{4,W,R}}{T_2\lambda\delta_0} + \sqrt{\frac{G_{4,W,R}}{T_2\delta_0}} + \lambda\|a^\sharp\|_2^2.\end{aligned}$$

The lower bound on λ makes the coefficient of $\|\hat{a}_W\|_2^2$ nonnegative, so this term can be dropped. Finally,

$$\mathbb{E}_{x\sim\pi_{\text{ref}},\,\zeta\sim\text{Unif}[-\tau,\tau]}\left[W_R(x,\zeta)(r_{\hat{a}_W}(x) - r_{\text{target}}(x))^2\right] = Z_W\|r_{\hat{a}_W} - r_{\text{target}}\|_{L^2(\nu_W)}^2$$

converts the weighted population seminorm into an $L^2(\nu_W)$ error. □

Lemma E.3 (Label-weighted truncated ridge estimate)**.** *Under the first-layer recovery condition of Lemma E.1, if*

$$\lambda \geq C\sqrt{\frac{M_{4,\mathrm{wt},R}}{T_2\delta_0}},$$

then the label-weighted ridge estimator $\hat{r}_R$ satisfies, with probability at least $1-4\delta_0$ over the ridge-fitting sample,

$$\inf_{c\in\mathbb{R}} \|\hat{r}_R - r^* - c\|^2_{L^2(\pi_{\beta_2,R})} \lesssim \frac{1}{Z_{\beta_2,R}Z_\zeta}\Bigg(e^{(M_{*,R}+\tau)/\beta_2}(N^{-2}+\varepsilon) + \frac{M_{2,\mathrm{wt},R}+M_{4,\mathrm{wt},R}+G_{4,\mathrm{wt},R}}{T_2\lambda\delta_0} + \sqrt{\frac{G_{4,\mathrm{wt},R}}{T_2\delta_0}} + \lambda N\Bigg).$$

Proof. Apply Lemma E.2 with

$$W_R(x,\zeta) = \mathbf{1}_{B_R}(x)e^{(r^*(x)+\zeta)/\beta_2}, \qquad r_{\mathrm{target}}(x) = r^*_{\mathrm{lbl},\beta_2}(x), \qquad \xi_R(x,\zeta) = \zeta - m_{\zeta,\beta_2}.$$

Then $Z_W = Z_{\beta_2,R}Z_\zeta$ and $\nu_W = \pi_{\beta_2,R}$. The required weighted residual orthogonality follows from

$$\mathbb{E}_{\zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[e^{\zeta/\beta_2}(\zeta - m_{\zeta,\beta_2})\right] = 0.$$

Moreover, on $B_R\times[-\tau,\tau]$,

$$W_R(x,\zeta) \leq e^{(M_{*,R}+\tau)/\beta_2},$$

so the comparator bound in Lemma E.1 gives

$$\mathcal{E}_{W,T_2}(a^\sharp) \lesssim e^{(M_{*,R}+\tau)/\beta_2}(N^{-2}+\varepsilon).$$

The same lemma gives $\|a^\sharp\|_2^2 \lesssim N$, which turns the comparator regularization term in Lemma E.2 into λN. Since $r^*_{\mathrm{lbl},\beta_2} = r^* + m_{\zeta,\beta_2}$, taking the infimum over constants converts the $L^2(\pi_{\beta_2,R})$ error for $r^*_{\mathrm{lbl},\beta_2}$ into the claimed shift-invariant error for r^*. □

Lemma E.4 (Surrogate-weighted truncated ridge estimate)**.** *Under the first-layer recovery condition of Lemma E.1, if*

$$\lambda \geq C\sqrt{\frac{M^{(0)}_{4,\mathrm{wt},R}}{T_2\delta_0}},$$

then, with probability at least $1-4\delta_0$ over the ridge-fitting sample,

$$\inf_{c\in\mathbb{R}} \|\hat{r}^{(0)}_R - r^* - c\|^2_{L^2(\pi^{(0)}_{\beta_2,R})} \leq \mathcal{E}^{(0)}_R,$$

where

$$\mathcal{E}^{(0)}_R := \frac{1}{Z^{(0)}_{\beta_2,R}}\left(e^{M_{0,R}/\beta_2}(N^{-2}+\varepsilon) + \frac{M^{(0)}_{2,\mathrm{wt},R}+M^{(0)}_{4,\mathrm{wt},R}+G^{(0)}_{4,\mathrm{wt},R}}{T_2\lambda\delta_0} + \sqrt{\frac{G^{(0)}_{4,\mathrm{wt},R}}{T_2\delta_0}} + \lambda N\right).$$

Proof. Apply Lemma E.2 with

$$W_R(x,\zeta) = \mathbf{1}_{B_R}(x)e^{r_{a_0}(x)/\beta_2}, \qquad r_{\mathrm{target}}(x) = r^*(x), \qquad \xi_R(x,\zeta) = \zeta.$$

Then $Z_W = Z^{(0)}_{\beta_2,R}$ and $\nu_W = \pi^{(0)}_{\beta_2,R}$. The required weighted residual orthogonality follows from

$$\mathbb{E}_{\zeta\sim\mathrm{Unif}[-\tau,\tau]}[\zeta] = 0,$$

because W_R is independent of ζ. On B_R,

$$W_R(x,\zeta) \leq e^{M_{0,R}/\beta_2},$$

so the comparator bound gives

$$\mathcal{E}_{W,T_2}(a^\sharp) \lesssim e^{M_{0,R}/\beta_2}(N^{-2}+\varepsilon).$$

The same lemma gives $\|a^\sharp\|_2^2 \lesssim N$, which turns the comparator regularization term in Lemma E.2 into λN. The claim follows from Lemma E.2. □

Derivation of the surrogate-weighted learning bound. Let

$$\mathcal{C}_{0,R}(\beta_2) = \frac{e^{M_{0,R}/\beta_2}}{Z^{(0)}_{\beta_2,R}}.$$

On the projected truncation set B_R, the functions $\|\psi_N(x)\|_2^2$, $\|\psi_N(x)\|_2^4$, and $|r^*(x)|^4$ are bounded: ψ_N depends only on the coordinates in S, and B_R bounds $\|P_S x\|_2$. Since $r_{a_0}(x) \le M_{0,R}$ on B_R,

$$\frac{M^{(0)}_{2,\mathrm{wt},R}}{Z^{(0)}_{\beta_2,R}} \lesssim e^{M_{0,R}/\beta_2}, \qquad \frac{M^{(0)}_{4,\mathrm{wt},R}+G^{(0)}_{4,\mathrm{wt},R}}{Z^{(0)}_{\beta_2,R}} \lesssim e^{M_{0,R}/\beta_2}.$$

Also,

$$M^{(0)}_{4,\mathrm{wt},R} \lesssim e^{M_{0,R}/\beta_2} Z^{(0)}_{\beta_2,R} = \frac{e^{2M_{0,R}/\beta_2}}{\mathcal{C}_{0,R}(\beta_2)}.$$

We choose the regularization parameter at the scale

$$\lambda = C_{\lambda,\mathrm{surr},R,N} e^{M_{0,R}/\beta_2} \left(\mathcal{C}_{0,R}(\beta_2) T_2 \delta_0\right)^{-1/2}. \tag{5}$$

where $C_{\lambda,\mathrm{surr},R,N}$ is sufficiently large and independent of β_2, T_2, δ_0. The preceding bound on $M^{(0)}_{4,\mathrm{wt},R}$ shows that this choice satisfies the lower bound required by Lemma E.4.

Start from Lemma E.4. The approximation term is

$$\frac{e^{M_{0,R}/\beta_2}}{Z^{(0)}_{\beta_2,R}}(N^{-2}+\varepsilon) = \mathcal{C}_{0,R}(\beta_2)(N^{-2}+\varepsilon).$$

The weighted moment term satisfies

$$\frac{M^{(0)}_{2,\mathrm{wt},R}+M^{(0)}_{4,\mathrm{wt},R}+G^{(0)}_{4,\mathrm{wt},R}}{Z^{(0)}_{\beta_2,R} T_2 \lambda \delta_0} \lesssim \frac{\mathcal{C}_{0,R}(\beta_2)^{1/2}}{\sqrt{T_2\delta_0}}.$$

For the square-root term, the bound on $G^{(0)}_{4,\mathrm{wt},R}$ gives

$$\frac{1}{Z^{(0)}_{\beta_2,R}}\sqrt{\frac{G^{(0)}_{4,\mathrm{wt},R}}{T_2\delta_0}} \lesssim \frac{\mathcal{C}_{0,R}(\beta_2)^{1/2}}{\sqrt{T_2\delta_0}}.$$

The regularization term is controlled by

$$\frac{\lambda N}{Z^{(0)}_{\beta_2,R}} \lesssim \frac{\mathcal{C}_{0,R}(\beta_2)^{1/2}}{\sqrt{T_2\delta_0}},$$

where the implicit constant may depend on N. Combining these estimates gives the displayed reward prediction error bound.

The learning-term bound follows from Lemma 5.2:

$$|T^{(0)}_{\mathrm{learn}}(R)| \le \frac{2M_R \mathcal{D}_{\mathrm{surr},R}}{\beta_2} \inf_{c\in\mathbb{R}} \|\hat{r}^{(0)}_R - r^* - c\|_{L^2(\pi^{(0)}_{\beta_2,R})}.$$

Taking square roots and using $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$ gives the surrogate-weighted learning contribution in Theorem 5.4. Corollary 5.5 follows by combining this bound with Lemma E.5 below. □

E.1 Polynomial control of $\mathcal{C}_{0,R}(\beta_2)$

Let $d_S := \dim S$, and let γ_S denote the standard Gaussian measure on S. Write

$$K_R := \{z \in S : \|z\|_2 \le R\}.$$

Since the frozen surrogate r_{a_0} depends only on $P_S x$, define $f_0 : K_R \to \mathbb{R}$ by

$$f_0(z) := r_{a_0}(x) \qquad \text{for any } x \text{ such that } P_S x = z.$$

Then

$$Z^{(0)}_{\beta_2,R} = \int_{K_R} e^{f_0(z)/\beta_2}\, \gamma_S(dz), \qquad M_{0,R} = \sup_{z\in K_R} f_0(z),$$

and

$$\mathcal{C}_{0,R}(\beta_2) = \left(\int_{K_R} e^{(f_0(z)-M_{0,R})/\beta_2}\, \gamma_S(dz)\right)^{-1}.$$

Lemma E.5 (Surrogate near-maximizer mass on the projected ball)**.** *There exists a finite constant $\bar{C}_{0,R}$, independent of β_2, such that for all sufficiently small $\beta_2 > 0$,*

$$\mathcal{C}_{0,R}(\beta_2) \le \bar{C}_{0,R}{\beta_2}^{-d_S}.$$

Proof. Since r_{a_0} is a finite linear combination of polynomial ridge features, f_0 is a polynomial function on S. Hence f_0 is Lipschitz on the compact set K_R. Let $L_{0,R}$ be a Lipschitz constant and set $L_+ := \max\{L_{0,R}, 1\}$.

Let $z_\star \in K_R$ be a maximizer of f_0. For

$$A_{\beta_2} := K_R \cap B\left(z_\star, \frac{\beta_2}{L_+}\right),$$

the Lipschitz property gives

$$f_0(z) \ge M_{0,R} - \beta_2 \qquad \text{for all } z \in A_{\beta_2}.$$

Therefore

$$Z^{(0)}_{\beta_2,R} \ge e^{M_{0,R}/\beta_2 - 1}\gamma_S(A_{\beta_2}).$$

The Gaussian density on S is bounded below by a positive constant on K_R. Moreover, since K_R is a Euclidean ball in S, there exists $c_{R,S} > 0$ such that

$$\mathrm{Vol}_{d_S}\left(K_R \cap B(z_\star, \rho)\right) \ge c_{R,S}\rho^{d_S}$$

for all sufficiently small $\rho > 0$, uniformly over $z_\star \in K_R$. Applying this with $\rho = \beta_2/L_+$ yields

$$\gamma_S(A_{\beta_2}) \ge c'_{0,R}{\beta_2}^{d_S}$$

for all sufficiently small $\beta_2 > 0$. Hence

$$Z^{(0)}_{\beta_2,R} \ge c''_{0,R}e^{M_{0,R}/\beta_2}{\beta_2}^{d_S}.$$

Rearranging proves

$$\mathcal{C}_{0,R}(\beta_2) \le \bar{C}_{0,R}{\beta_2}^{-d_S}.$$

□

Remark E.6 (Sharper exponents from local Taylor expansions). The exponent d_S in Lemma E.5 is a worst-case exponent obtained from Lipschitz continuity alone. If the local Taylor expansion of f_0 around its maximizers is analyzed more precisely, the exponent can often be improved. For example, a nondegenerate interior maximizer gives

$$\mathcal{C}_{0,R}(\beta_2) \lesssim {\beta_2}^{-d_S/2}.$$

More generally, the exponent is determined by the leading nonzero negative terms in the local Taylor expansion near the maximizers, analogously to the low-temperature analysis for the true reward in Appendix B. These sharper estimates are not needed for the main value-gap bound.

F Low-temperature simplification for label weighting

This appendix derives the low-temperature moment ratios used to simplify the label-weighted reward prediction error. Write

$$\alpha := \frac{1}{2p_{\max}}, \qquad L_* := B_* + \tau.$$

Here $s > 0$ is a generic temperature parameter; below we apply these estimates with $s = \beta_2$ and $s = \beta_2/2$. Define

$$\pi_{s,R}(dx) := \frac{\mathbf{1}_{B_R}(x)e^{r^*(x)/s}}{Z_{s,R}}\pi_{\mathrm{ref}}(dx), \qquad Z_{s,R} := \mathbb{E}_{x\sim\pi_{\mathrm{ref}}}\left[\mathbf{1}_{B_R}(x)e^{r^*(x)/s}\right].$$

Also define the exponentially weighted uniform-noise normalization at temperature s by

$$Z_\zeta(s) := \mathbb{E}_{\zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[e^{\zeta/s}\right],$$

and the corresponding probability measure by

$$\nu_{\zeta,s}(d\zeta) := \frac{e^{\zeta/s}}{Z_\zeta(s)}\mathrm{Unif}[-\tau,\tau](d\zeta), \qquad m_{\zeta,s} := \mathbb{E}_{\zeta\sim\nu_{\zeta,s}}[\zeta].$$

Thus, specializing the generic temperature s to the deployment temperature β_2, we write $Z_\zeta = Z_\zeta(\beta_2)$ and $m_{\zeta,\beta_2} = m_{\zeta,s}|_{s=\beta_2}$.

Lemma F.1 (Exact formulas for the exponentially weighted uniform noise)**.** *For every $s > 0$,*

$$Z_\zeta(s) = \frac{s}{\tau}\sinh\left(\frac{\tau}{s}\right), \qquad m_{\zeta,s} = \tau\coth\left(\frac{\tau}{s}\right) - s.$$

Moreover,

$$\mathrm{Var}_{\zeta\sim\nu_{\zeta,s}}[\zeta] = s^2 - \tau^2\,\mathrm{csch}^2\left(\frac{\tau}{s}\right).$$

At the two temperatures β_2 and $\beta_2/2$,

$$\frac{Z_\zeta(\beta_2/2)}{Z_\zeta(\beta_2)} = \cosh\left(\frac{\tau}{\beta_2}\right),$$

and

$$V^{(2)}_{\zeta,\beta_2} := \mathbb{E}_{\zeta\sim\nu_{\zeta,\beta_2/2}}\left[(\zeta - m_{\zeta,\beta_2})^2\right] = \frac{{\beta_2}^2}{2} - \beta_2\tau\,\mathrm{csch}\left(\frac{2\tau}{\beta_2}\right).$$

Consequently, as $\beta_2 \downarrow 0$,

$$Z_\zeta(\beta_2) = \frac{\beta_2}{2\tau}e^{\tau/\beta_2}\left(1 + o(1)\right), \qquad m_{\zeta,\beta_2} = \tau - \beta_2 + O(e^{-2\tau/\beta_2}),$$

$$\frac{Z_\zeta(\beta_2/2)}{Z_\zeta(\beta_2)} = \frac{1}{2}e^{\tau/\beta_2}\left(1 + o(1)\right), \qquad V^{(2)}_{\zeta,\beta_2} = \frac{{\beta_2}^2}{2} + O\left(\beta_2\tau e^{-2\tau/\beta_2}\right).$$

Proof. The identity for $Z_\zeta(s)$ follows directly from

$$\mathbb{E}_{\zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[e^{\zeta/s}\right] = \frac{1}{2\tau}\int_{-\tau}^{\tau} e^{z/s}\,dz.$$

Let

$$\Lambda_\zeta(\eta) := \log\mathbb{E}_{\zeta\sim\mathrm{Unif}[-\tau,\tau]}\left[e^{\eta\zeta}\right].$$

Then $\nu_{\zeta,s}$ is the exponentially weighted uniform law with $\eta = 1/s$. Differentiating this log-normalization gives

$$\Lambda'_\zeta(\eta) = \mathbb{E}_{\zeta\sim\nu_{\zeta,1/\eta}}[\zeta], \qquad \Lambda''_\zeta(\eta) = \mathrm{Var}_{\zeta\sim\nu_{\zeta,1/\eta}}[\zeta].$$

Differentiating $\Lambda_\zeta(\eta) = \log(\sinh(\tau\eta)/(\tau\eta))$ and setting $\eta = 1/s$ gives the formulas for $m_{\zeta,s}$ and $\mathrm{Var}_{\zeta\sim\nu_{\zeta,s}}[\zeta]$. The ratio $Z_\zeta(\beta_2/2)/Z_\zeta(\beta_2) = \cosh(\tau/\beta_2)$ follows from $\sinh(2a) = 2\sinh(a)\cosh(a)$. Finally,

$$V^{(2)}_{\zeta,\beta_2} = \mathrm{Var}_{\zeta\sim\nu_{\zeta,\beta_2/2}}[\zeta] + \left(m_{\zeta,\beta_2/2} - m_{\zeta,\beta_2}\right)^2,$$

where

$$m_{\zeta,\beta_2/2} - m_{\zeta,\beta_2} = \frac{\beta_2}{2} - \tau \operatorname{csch}\left(\frac{2\tau}{\beta_2}\right).$$

Substituting this identity and $\mathrm{Var}_{\zeta\sim\nu_{\zeta,\beta_2/2}}[\zeta] = {\beta_2}^2/4 - \tau^2 \operatorname{csch}^2(2\tau/\beta_2)$ gives the displayed formula for $V^{(2)}_{\zeta,\beta_2}$. The asymptotics follow from $\sinh a \sim e^a/2$, $\cosh a \sim e^a/2$, $\coth a = 1 + O(e^{-2a})$, and $\operatorname{csch} a = O(e^{-a})$ as $a \to \infty$. □

The next lemma gives the truncated Laplace estimate needed for feature moments. Let

$$S_0 := S \cap (\theta^*)^\perp.$$

Let γ_{S_0} denote the standard Gaussian measure on S_0. For bounded continuous $F : B_R \to \mathbb{R}$ that depend only on $P_S x$, define

$$\mathcal{A}_R[F] := \sum_{i\in I_{\max}} \phi(u_i) \left(\int_{\mathbb{R}} e^{-c_i t^{2p_{\max}}}\, dt\right) \times \int_{S_0} \mathbf{1}_{\{u_i^2+\|z\|_2^2\le R^2\}} F(u_i\theta^* + z)\, d\gamma_{S_0}(z).$$

Set

$$A_R := \mathcal{A}_R[1].$$

Since R is taken sufficiently large, $R > \max_{i\in I_{\max}} |u_i|$, so $A_R > 0$. Define

$$\mu_{0,R}[F] := \frac{\mathcal{A}_R[F]}{A_R}.$$

Lemma F.2 (Truncated feature Laplace asymptotics)**.** *For every bounded continuous function $F : B_R \to \mathbb{R}$ that depends only on the projected coordinate $P_S x$,*

$$\mathbb{E}_{x\sim\pi_{\mathrm{ref}}}\left[\mathbf{1}_{B_R}(x) e^{r^*(x)/s} F(x)\right] = e^{B_*/s} s^\alpha \left(\mathcal{A}_R[F] + o(1)\right) \qquad (s \downarrow 0).$$

In particular,

$$Z_{s,R} = e^{B_*/s} s^\alpha (A_R + o(1)), \qquad \mathbb{E}_{x\sim\pi_{s,R}}[F(x)] \to \mu_{0,R}[F].$$

Moreover, if $F_s : B_R \to \mathbb{R}$ are bounded continuous functions that depend only on $P_S x$ and $\|F_s - F_0\|_{L^\infty(B_R)} \to 0$, then

$$\mathbb{E}_{x\sim\pi_{s,R}}[F_s(x)] \to \mu_{0,R}[F_0].$$

Proof. Let

$$S_0 := S \cap (\theta^*)^\perp.$$

Write the projected coordinates as

$$P_S x = u\theta^* + z, \qquad u := \langle \theta^*, x\rangle, \qquad z \in S_0.$$

Under $x \sim \pi_{\mathrm{ref}}$, the coordinates $u \sim \mathcal{N}(0,1)$ and $z \sim \gamma_{S_0}$ are independent, where γ_{S_0} denotes the standard Gaussian measure on S_0. Since

$$B_R = \{x : \|P_S x\|_2 \le R\},$$

the truncation event is

$$u^2 + \|z\|_2^2 \le R^2.$$

Define

$$I_s[F] := \mathbb{E}_{x\sim\pi_{\mathrm{ref}}}\left[\mathbf{1}_{B_R}(x) e^{r^*(x)/s} F(x)\right].$$

Using $r^*(x) = \sigma^*(\langle\theta^*, x\rangle) = \sigma^*(u)$, we get

$$I_s[F] = \int_{\mathbb{R}} \int_{S_0} \mathbf{1}_{\{u^2+\|z\|_2^2\le R^2\}} e^{\sigma^*(u)/s} F(u\theta^* + z)\phi(u)\, d\gamma_{S_0}(z)\, du.$$

The indicator implies $|u| \le R$. Choose disjoint neighborhoods U_j of the global maximizers of σ^* that intersect $[-R, R]$. For each $i \in I_{\max}$, choose U_i small enough that $U_i \subset (-R, R)$, which is possible because $R > |u_i|$.

Let $I_s^{\text{off}}[F]$ be the part of $I_s[F]$ with $u \notin \cup_j U_j$. By continuity of σ^* on the compact set $[-R, R] \setminus \cup_j U_j$, there is $\eta > 0$ such that $\sigma^*(u) \le B_* - \eta$ on this set. Hence

$$|I_s^{\text{off}}[F]| \le \|F\|_{L^\infty(B_R)} e^{(B_* - \eta)/s},$$

and therefore

$$e^{-B_*/s} s^{-\alpha} |I_s^{\text{off}}[F]| \to 0.$$

Now fix $i \in I_{\max}$, and let $I_{s,i}^{\text{loc}}[F]$ be the part of $I_s[F]$ with $u \in U_i$. Set $u = u_i + s^\alpha t$. Then

$$\frac{\sigma^*(u_i + s^\alpha t) - B_*}{s} \to -c_i t^{2p_{\max}}$$

for every fixed t. After this change of variables,

$$\begin{aligned} e^{-B_*/s} s^{-\alpha} I_{s,i}^{\text{loc}}[F] = \int_{T_{s,i}} \int_{S_0} \mathbf{1}_{\{(u_i + s^\alpha t)^2 + \|z\|_2^2 \le R^2\}} e^{(\sigma^*(u_i + s^\alpha t) - B_*)/s} \\ \times F((u_i + s^\alpha t)\theta^* + z) \phi(u_i + s^\alpha t)\, d\gamma_{S_0}(z)\, dt, \end{aligned}$$

where $T_{s,i} := \{t : u_i + s^\alpha t \in U_i\}$.

After shrinking U_i if necessary, the Taylor expansion at u_i implies

$$\sigma^*(u_i + h) \le B_* - \frac{c_i}{2} h^{2p_{\max}}$$

for $u_i + h \in U_i$. Thus the absolute value of the last integrand is bounded by

$$\|F\|_{L^\infty(B_R)} e^{-(c_i/2) t^{2p_{\max}}},$$

which is integrable with respect to $dt\, d\gamma_{S_0}(z)$.

For γ_{S_0}-almost every z, the point z is not on the boundary

$$\|z\|_2^2 = R^2 - u_i^2.$$

Hence

$$\mathbf{1}_{\{(u_i + s^\alpha t)^2 + \|z\|_2^2 \le R^2\}} \to \mathbf{1}_{\{u_i^2 + \|z\|_2^2 \le R^2\}}.$$

The continuity of F gives

$$F((u_i + s^\alpha t)\theta^* + z) \to F(u_i \theta^* + z).$$

Dominated convergence gives

$$e^{-B_*/s} s^{-\alpha} I_{s,i}^{\text{loc}}[F] \to \mathcal{A}_{R,i}[F],$$

where

$$\begin{aligned} \mathcal{A}_{R,i}[F] :=& \phi(u_i) \left(\int_{\mathbb{R}} e^{-c_i t^{2p_{\max}}}\, dt \right) \\ & \times \int_{S_0} \mathbf{1}_{\{u_i^2 + \|z\|_2^2 \le R^2\}} F(u_i \theta^* + z)\, d\gamma_{S_0}(z). \end{aligned}$$

If u_j is a global maximizer with $j \notin I_{\max}$, then $p_j < p_{\max}$. The same local argument with the scale $s^{1/(2p_j)}$ gives

$$I_{s,j}^{\text{loc}}[F] = O\left(e^{B_*/s} s^{1/(2p_j)}\right) = o\left(e^{B_*/s} s^\alpha\right).$$

Combining the off-neighborhood estimate, the dominant local limits, and the non-dominant local bounds yields

$$I_s[F] = e^{B_*/s} s^\alpha \left(\sum_{i \in I_{\max}} \mathcal{A}_{R,i}[F] + o(1) \right) = e^{B_*/s} s^\alpha \left(\mathcal{A}_R[F] + o(1) \right).$$

Taking $F \equiv 1$ gives the formula for $Z_{s,R}$, and division gives

$$\mathbb{E}_{x\sim\pi_{s,R}}[F(x)] \to \mu_{0,R}[F].$$

The uniformly convergent version follows by writing

$$\left|\mathbb{E}_{x\sim\pi_{s,R}}[F_s(x)] - \mathbb{E}_{x\sim\pi_{s,R}}[F_0(x)]\right| \le \|F_s - F_0\|_{L^\infty(B_R)}$$

and applying the fixed-F result to F_0. □

Lemma F.3 (Low-temperature normalized label-weighted moments)**.** *Recall the limiting functional $\mu_{0,R}$ from Lemma F.2. Define*

$$\Psi_{2,R,N} := \mu_{0,R}[\|\psi_N\|_2^2], \qquad \Psi_{4,R,N} := \mu_{0,R}[\|\psi_N\|_2^4].$$

With $M_{2,\mathrm{wt},R}$, $M_{4,\mathrm{wt},R}$, and $G_{4,\mathrm{wt},R}$ defined in Appendix E, as $\beta_2 \downarrow 0$,

$$Z_{\beta_2,R} = e^{B_*/\beta_2}{\beta_2}^{\alpha}(A_R + o(1)),$$

$$\frac{Z_{\beta_2/2,R}}{Z_{\beta_2,R}} = 2^{-\alpha}e^{B_*/\beta_2}(1 + o(1)).$$

Since $M_{*,R} = B_*$,

$$\frac{e^{(M_{*,R}+\tau)/\beta_2}}{Z_{\beta_2,R}Z_\zeta(\beta_2)} = \frac{2\tau}{A_R}{\beta_2}^{-(\alpha+1)}(1 + o(1)).$$

Moreover,

$$\frac{M_{2,\mathrm{wt},R}}{Z_{\beta_2,R}Z_\zeta(\beta_2)} = 2^{-\alpha-2}\Psi_{2,R,N}{\beta_2}^2 e^{L_*/\beta_2}(1 + o(1)),$$

$$\frac{M_{4,\mathrm{wt},R}}{Z_{\beta_2,R}Z_\zeta(\beta_2)} = 2^{-\alpha-1}\Psi_{4,R,N}e^{L_*/\beta_2}(1 + o(1)),$$

and

$$\frac{G_{4,\mathrm{wt},R}}{Z_{\beta_2,R}Z_\zeta(\beta_2)} = O\left(e^{L_*/\beta_2}\right).$$

If $B_* + \tau \neq 0$, the last display can be sharpened to

$$\frac{G_{4,\mathrm{wt},R}}{Z_{\beta_2,R}Z_\zeta(\beta_2)} = 2^{-\alpha-1}|B_* + \tau|^4 e^{L_*/\beta_2}(1 + o(1)).$$

Proof. First reduce the weighted moments to normalized expectations. The random variables $x \sim \pi_{\mathrm{ref}}$ and $\zeta \sim \mathrm{Unif}[-\tau, \tau]$ are independent, and

$$e^{2(r^*(x)+\zeta)/\beta_2} = e^{r^*(x)/(\beta_2/2)}e^{\zeta/(\beta_2/2)}.$$

Thus each unnormalized weighted moment splits into an x-integral normalized by $Z_{\beta_2/2,R}$ and a noise integral normalized by $Z_\zeta(\beta_2/2)$. After division by $Z_{\beta_2,R}Z_\zeta(\beta_2)$,

$$\frac{M_{2,\mathrm{wt},R}}{Z_{\beta_2,R}Z_\zeta(\beta_2)} = \frac{Z_{\beta_2/2,R}}{Z_{\beta_2,R}}\frac{Z_\zeta(\beta_2/2)}{Z_\zeta(\beta_2)}V^{(2)}_{\zeta,\beta_2}\mathbb{E}_{x\sim\pi_{\beta_2/2,R}}[\|\psi_N(x)\|_2^2],$$

$$\frac{M_{4,\mathrm{wt},R}}{Z_{\beta_2,R}Z_\zeta(\beta_2)} = \frac{Z_{\beta_2/2,R}}{Z_{\beta_2,R}}\frac{Z_\zeta(\beta_2/2)}{Z_\zeta(\beta_2)}\mathbb{E}_{x\sim\pi_{\beta_2/2,R}}[\|\psi_N(x)\|_2^4],$$

and

$$\frac{G_{4,\mathrm{wt},R}}{Z_{\beta_2,R}Z_\zeta(\beta_2)} = \frac{Z_{\beta_2/2,R}}{Z_{\beta_2,R}}\frac{Z_\zeta(\beta_2/2)}{Z_\zeta(\beta_2)}\mathbb{E}_{x\sim\pi_{\beta_2/2,R}}[|r^*(x) + m_{\zeta,\beta_2}|^4].$$

Here the noise factor in the first display is

$$\mathbb{E}_{\zeta\sim\nu_{\zeta,\beta_2/2}}[(\zeta - m_{\zeta,\beta_2})^2] = V^{(2)}_{\zeta,\beta_2}.$$

We now take the low-temperature limit. The asymptotic for $Z_{\beta_2,R}$ is Lemma F.2 with $F \equiv 1$. Applying the same lemma at temperature $\beta_2/2$ gives

$$Z_{\beta_2/2,R} = e^{2B_*/\beta_2}(\beta_2/2)^{\alpha}(A_R + o(1)),$$

which implies the displayed ratio $Z_{\beta_2/2,R}/Z_{\beta_2,R}$. Since R is taken sufficiently large, $R > |u_i|$ for all $i \in I_{\max}$. Hence the truncated ball contains a global maximizer of r^*, and $M_{*,R} = B_*$. Combining the asymptotic for $Z_{\beta_2,R}$ with Lemma F.1 gives the normalization ratio involving $e^{(M_{*,R}+\tau)/\beta_2}$.

For the weighted moments, use the normalized identities proved above. Lemma F.2 gives

$$\mathbb{E}_{x\sim\pi_{\beta_2/2,R}}[\|\psi_N(x)\|_2^2] \to \Psi_{2,R,N}, \qquad \mathbb{E}_{x\sim\pi_{\beta_2/2,R}}[\|\psi_N(x)\|_2^4] \to \Psi_{4,R,N}.$$

The noise formulas give

$$\frac{Z_\zeta(\beta_2/2)}{Z_\zeta(\beta_2)} = \frac{1}{2}e^{\tau/\beta_2}(1+o(1)), \qquad V^{(2)}_{\zeta,\beta_2} = \frac{{\beta_2}^2}{2}(1+o(1)).$$

This proves the $M_{2,\mathrm{wt},R}$ and $M_{4,\mathrm{wt},R}$ formulas. Finally,

$$m_{\zeta,\beta_2} \to \tau,$$

so the functions

$$F_{\beta_2}(x) := |r^*(x) + m_{\zeta,\beta_2}|^4$$

converge uniformly on B_R to $F_0(x) = |r^*(x)+\tau|^4$. The uniform version of Lemma F.2 gives

$$\mathbb{E}_{x\sim\pi_{\beta_2/2,R}}[|r^*(x) + m_{\zeta,\beta_2}|^4] \to \mu_{0,R}[|r^*+\tau|^4].$$

Under $\mu_{0,R}$, the coordinate $\langle\theta^*, x\rangle$ belongs to $\{u_i : i \in I_{\max}\}$ almost surely. Since $r^*(x) = \sigma^*(\langle\theta^*, x\rangle)$, this implies $r^*(x) = B_*$ $\mu_{0,R}$-almost surely. Therefore the last limit is $|B_*+\tau|^4$. This gives the sharpened asymptotic when $B_* + \tau \neq 0$. Without the nondegeneracy condition, the same convergence only gives that this expectation is bounded, which yields the displayed upper bound. □

Derivation of the label-weighted learning bound. Start from Lemma E.3:

$$\inf_{c\in\mathbb{R}} \|\hat{r}_R - r^* - c\|^2_{L^2(\pi_{\beta_2,R})} \lesssim \frac{1}{Z_{\beta_2,R}Z_\zeta(\beta_2)}\Bigg(e^{(M_{*,R}+\tau)/\beta_2}(N^{-2}+\varepsilon) + \frac{M_{2,\mathrm{wt},R} + M_{4,\mathrm{wt},R} + G_{4,\mathrm{wt},R}}{T_2\lambda\delta_0} + \sqrt{\frac{G_{4,\mathrm{wt},R}}{T_2\delta_0}} + \lambda N\Bigg).$$

The approximation term is controlled by

$$\frac{e^{(M_{*,R}+\tau)/\beta_2}}{Z_{\beta_2,R}Z_\zeta(\beta_2)} \lesssim {\beta_2}^{-(\alpha+1)}.$$

The moment terms are controlled by Lemma F.3:

$$\frac{M_{2,\mathrm{wt},R}}{Z_{\beta_2,R}Z_\zeta(\beta_2)} \lesssim {\beta_2}^2 e^{L_*/\beta_2}, \qquad \frac{M_{4,\mathrm{wt},R} + G_{4,\mathrm{wt},R}}{Z_{\beta_2,R}Z_\zeta(\beta_2)} \lesssim e^{L_*/\beta_2}.$$

For the square-root term, the same lemma implies

$$\frac{G_{4,\mathrm{wt},R}}{Z_{\beta_2,R}Z_\zeta(\beta_2)} \lesssim e^{L_*/\beta_2}, \qquad Z_{\beta_2,R}Z_\zeta(\beta_2) \lesssim e^{L_*/\beta_2}{\beta_2}^{\alpha+1}.$$

Multiplying these two bounds gives

$$G_{4,\mathrm{wt},R} \lesssim e^{2L_*/\beta_2}{\beta_2}^{\alpha+1}.$$

The lower bound

$$Z_{\beta_2,R}Z_\zeta(\beta_2) \gtrsim e^{L_*/\beta_2}{\beta_2}^{\alpha+1}$$

then gives

$$\frac{1}{Z_{\beta_2,R}Z_\zeta(\beta_2)}\sqrt{\frac{G_{4,\mathrm{wt},R}}{T_2\delta_0}} \lesssim \frac{{\beta_2}^{-(\alpha+1)/2}}{\sqrt{T_2\delta_0}}.$$

Finally,

$$\frac{\lambda N}{Z_{\beta_2,R} Z_\zeta(\beta_2)} \lesssim \lambda N {\beta_2}^{-(\alpha+1)} e^{-L_*/\beta_2}.$$

Combining these estimates gives, for every λ satisfying the lower bound in Lemma E.3,

$$\begin{aligned}\inf_{c\in\mathbb{R}} \|\hat{r}_R - r^* - c\|^2_{L^2(\pi_{\beta_2,R})} \lesssim\ & C_{\mathrm{app},R} {\beta_2}^{-(\alpha+1)}(N^{-2}+\varepsilon) \\ & + C_{\mathrm{mom},R,N} \frac{e^{L_*/\beta_2}}{T_2\lambda\delta_0}(1+{\beta_2}^2) \\ & + C_{\sqrt{G},R} \frac{{\beta_2}^{-(\alpha+1)/2}}{\sqrt{T_2\delta_0}} + C_{\lambda,R}\lambda N {\beta_2}^{-(\alpha+1)} e^{-L_*/\beta_2}.\end{aligned}$$

The lower bound on λ required by Lemma E.3 is $\lambda \geq C\sqrt{M_{4,\mathrm{wt},R}/(T_2\delta_0)}$. By Lemma F.3,

$$M_{4,\mathrm{wt},R} = \frac{M_{4,\mathrm{wt},R}}{Z_{\beta_2,R} Z_\zeta(\beta_2)} Z_{\beta_2,R} Z_\zeta(\beta_2) \lesssim {\beta_2}^{\alpha+1} e^{2L_*/\beta_2},$$

and therefore it is sufficient that

$$\lambda \geq C_{\lambda,\mathrm{low},R,N} \frac{{\beta_2}^{(\alpha+1)/2} e^{L_*/\beta_2}}{\sqrt{T_2\delta_0}}.$$

We choose the regularization parameter at the scale

$$\lambda = C_{\lambda,\mathrm{choice},R,N} {\beta_2}^{(\alpha+1)/2} e^{L_*/\beta_2} \left(T_2\delta_0\right)^{-1/2}. \tag{6}$$

where $C_{\lambda,\mathrm{choice},R,N}$ is sufficiently large and independent of β_2, T_2, δ_0. This choice satisfies the lower bound above. Substituting this value of λ into the intermediate prediction-error bound gives

$$\inf_{c\in\mathbb{R}} \|\hat{r}_R - r^* - c\|^2_{L^2(\pi_{\beta_2,R})} \lesssim C_{\mathrm{app},R} {\beta_2}^{-(\alpha+1)}(N^{-2}+\varepsilon) + C_{\mathrm{stat},R,N} \frac{{\beta_2}^{-(\alpha+1)/2}}{\sqrt{T_2\delta_0}}.$$

Finally, Lemma 5.2 gives

$$|T_{\mathrm{learn}}(R)| \leq \frac{2M_R \mathcal{D}_{\mathrm{lbl},R}}{\beta_2} \left(\inf_{c\in\mathbb{R}} \|\hat{r}_R - r^* - c\|^2_{L^2(\pi_{\beta_2,R})} \right)^{1/2}.$$

Applying $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ to the prediction-error bound yields the stated learning bound. □

G Admissible deployment temperatures

This appendix makes explicit the admissible temperature sets referenced in Section 5.4. Write

$$\mathcal{R}_R^{\mathrm{lbl}} := \mathcal{R}_R, \qquad \mathcal{R}_R^{\mathrm{surr}} := \mathcal{R}_R^{(0)}, \qquad \rho_N := N^{-1} + \sqrt{\varepsilon}.$$

For label weighting, define

$$\mathcal{L}_{\mathrm{lbl}}(\beta) := \beta^{-(\alpha+3)/2}\rho_N + \beta^{-(\alpha+5)/4}(T_2\delta_0)^{-1/4}.$$

For surrogate weighting, assume that for some $\alpha_0 \geq 0$,

$$\mathcal{C}_{0,R}(\beta) \lesssim \beta^{-\alpha_0}$$

for all sufficiently small $\beta > 0$. Lemma E.5 gives the valid worst-case choice $\alpha_0 = d_S$, while Remark E.6 explains when smaller exponents are available. Under this polynomial control, define

$$\mathcal{L}_{\mathrm{surr}}(\beta) := \beta^{-1-\alpha_0/2}\rho_N + \beta^{-1-\alpha_0/4}(T_2\delta_0)^{-1/4}.$$

We assume a deterministic coverage envelope $\overline{\mathcal{D}}_{w,R} : (0, \overline{\beta}_2] \to [1, \infty]$ such that

$$\mathcal{D}_{w,R}(\beta) \leq \overline{\mathcal{D}}_{w,R}(\beta)$$

for every considered deployment temperature β. As in standard coverage or concentrability assumptions, this envelope is treated as a problem-dependent quantity rather than estimated from the same ridge-fitting samples.

For a tolerance $\eta > 0$, define the admissible temperature set

$$\mathcal{S}_w^{\mathrm{adm}}(\eta; N, T_2, \varepsilon, \delta_0) := \left\{\beta \in (0, \overline{\beta}_2] : M_R \overline{\mathcal{D}}_{w,R}(\beta)\mathcal{L}_w(\beta) \le \eta\right\}.$$

Here $w \in \{\mathrm{lbl}, \mathrm{surr}\}$, with $\mathcal{L}_w = \mathcal{L}_{\mathrm{lbl}}$ or $\mathcal{L}_{\mathrm{surr}}$, respectively.

Corollary G.1 (Admissible deployment temperature)**.** *Fix $w \in \{\mathrm{lbl}, \mathrm{surr}\}$ and $\eta > 0$. On the event of Theorem 5.3 for $w = \mathrm{lbl}$, or Corollary 5.5 for $w = \mathrm{surr}$, every $\beta_2 \in \mathcal{S}_w^{\mathrm{adm}}(\eta; N, T_2, \varepsilon, \delta_0)$ satisfies*

$$|\mathcal{R}_R^w| \lesssim \Gamma_{R,\overline{\beta}_2}(\beta_2, \beta^*) + \eta.$$

Consequently, if

$$\beta_{2,\eta}^w \in \arg\min_{\beta \in \mathcal{S}_w^{\mathrm{adm}}(\eta; N, T_2, \varepsilon, \delta_0)} \Gamma_{R,\overline{\beta}_2}(\beta, \beta^*),$$

then

$$|\mathcal{R}_R^w| \lesssim \inf_{\beta \in \mathcal{S}_w^{\mathrm{adm}}(\eta; N, T_2, \varepsilon, \delta_0)} \Gamma_{R,\overline{\beta}_2}(\beta, \beta^*) + \eta.$$

Proof. By definition of the admissible temperature set,

$$M_R \overline{\mathcal{D}}_{w,R}(\beta_2)\mathcal{L}_w(\beta_2) \le \eta.$$

For $w = \mathrm{lbl}$, this upper bounds the learning term in Theorem 5.3; for $w = \mathrm{surr}$, it upper bounds the learning term in Corollary 5.5. Substituting this into the corresponding bound gives the first claim. The second claim follows by minimizing the remaining $\Gamma_{R,\overline{\beta}_2}$ term over the admissible temperature set. □

Remark G.2 (Resource dependence of the admissible temperature set). Fix the coverage envelope $\overline{\mathcal{D}}_{w,R}$, the tolerance η, the failure parameter δ_0, and, in the surrogate case, the exponent α_0. If

$$N'^{-1} + \sqrt{\varepsilon'} \le N^{-1} + \sqrt{\varepsilon}, \qquad T_2' \ge T_2,$$

then

$$\mathcal{S}_w^{\mathrm{adm}}(\eta; N, T_2, \varepsilon, \delta_0) \subseteq \mathcal{S}_w^{\mathrm{adm}}(\eta; N', T_2', \varepsilon', \delta_0).$$

Thus, under a fixed coverage envelope, improving feature recovery or increasing the number of second-stage samples enlarges the admissible temperature region.

In particular, for any fixed candidate temperature $\beta_0 \in (0, \overline{\beta}_2]$ with $\overline{\mathcal{D}}_{w,R}(\beta_0) < \infty$, the temperature β_0 belongs to $\mathcal{S}_w^{\mathrm{adm}}(\eta; N, T_2, \varepsilon, \delta_0)$ for all sufficiently large N and T_2, and sufficiently small ε.

H Numerical Experiments: Setup and Full Results

We give the experimental details for Figures 1 and 2. Both experiments use $x \sim \mathcal{N}(0, I_d)$ and a unit vector θ^*. The true reward and observed label are

$$z = \langle \theta^*, x \rangle, \qquad r^*(x) = \sigma^*(z) = 3 + 0.35\,\mathrm{He}_3(z) - \mathrm{He}_4(z), \qquad y = \sigma^*(z) + \zeta,$$

where He_k is the probabilists' Hermite polynomial and $\zeta \sim \mathrm{Unif}[-0.1, 0.1]$ is independent of x. This reward is bounded above and satisfies $\mathrm{IE} = 3$, $\mathrm{GE} = 1$, and $I_* = 2$. Figure 1 examines first-layer weak recovery. Figure 2 compares the two second-layer weighting schemes given exact recovery of the true direction.

H.1 Weak recovery by online spherical SGD

We use the online update of Theorem A.1. The activation and its derivative are

$$\sigma(h) = \tfrac{1}{2}h + \tfrac{1}{6}h^3 - \tfrac{1}{80}h^4, \qquad \sigma'(h) = \tfrac{1}{2} + \tfrac{1}{2}h^2 - \tfrac{1}{20}h^3,$$

giving $V_0 = 1$. We test all combinations of

$$d \in \{32, 64, 128, 256, 512, 1024\}, \qquad \beta_1 \in \{1, 1.5, 2, 3, 4, 6, 8, 12\}.$$

Each update uses one fresh sample. The number of updates and the learning rate are

$$T = 64\beta_1^2 d, \qquad \eta_1 = \frac{0.01}{\beta_1 d}.$$

The coefficients 64 and 0.01 are fixed across conditions. The number of updates follows the $\beta_1^2 d$ scaling for this example with $I_* = 2$ and $p_{\mathrm{gen}} = 1$.

By rotational symmetry, take $\theta^* = e_1$. We fix $a = -1$, so that $aU_1(\beta_1)V_0 > 0$ at every tested temperature, and draw $w^{(0)}$ uniformly from the sphere conditional on

$$\frac{0.5}{\sqrt{d}} \le \langle w^{(0)}, \theta^* \rangle \le \frac{2}{\sqrt{d}}.$$

We run 40 independent trials per condition. For a given dimension and trial number, the initial alignment is shared across temperatures. The update samples are independent across temperatures and trials. The color in Figure 1 gives the fraction of trials under this initialization for which the final iterate satisfies $|\langle w^{(T)}, \theta^* \rangle| \ge 0.5$.

H.2 Second-layer weighted fitting

We use the true coordinate $z = \langle \theta^*, x \rangle \sim \mathcal{N}(0, 1)$ and the five-dimensional feature vector

$$\phi(z) = \big(\mathrm{He}_k(z)/\sqrt{k!}\big)_{k=0}^{4}.$$

The span of these features represents σ^* and its constant shifts exactly, allowing us to compare label and surrogate weighting in the absence of feature approximation error. We fix the proxy reward at $r_{a_0}(z) = \sigma^*(z) - 2z$ in all conditions and test the temperature and sample-size grids

$$\mathcal{B}_2 = \{0.25, 0.35, 0.5, 0.7, 1, 1.5, 2, 3\}, \qquad \mathcal{T}_2 = \{16, 32, 64, 128, 256, 512, 1024, 2048, 4096\}.$$

Fitting and repetitions. In each of 60 independent trials, we draw 4096 standard Gaussian coordinates z_i and add independent noise to the corresponding rewards to obtain labels y_i. For sample size T_2, we fit on the first T_2 observations. Within a trial, we use the same observations for all temperatures and both weighting schemes.

For $\mathrm{w} \in \{\mathrm{lbl}, \mathrm{surr}\}$, the weights are

$$W_i^{\mathrm{lbl}} = e^{y_i/\beta_2}, \qquad W_i^{\mathrm{surr}} = e^{r_{a_0}(z_i)/\beta_2}, \qquad \omega_i^{\mathrm{w}} = \frac{W_i^{\mathrm{w}}}{\sum_{j=1}^{T_2} W_j^{\mathrm{w}}}.$$

The fitted reward is $\widehat{r}^{\mathrm{w}}(z) = \phi(z)^\top \widehat{c}^{\mathrm{w}}$, where

$$\widehat{c}^{\mathrm{w}} = \arg\min_{c \in \mathbb{R}^5} \left\{ \sum_{i=1}^{T_2} \omega_i^{\mathrm{w}} (\phi(z_i)^\top c - y_i)^2 + 10^{-6}\|c\|_2^2 \right\}.$$

The penalty applies to all five coefficients, including the constant term. We fit on samples from the untruncated Gaussian distribution and apply truncation only at policy evaluation.

Policy evaluation. The true value of the policy induced by a reward f is

$$V_{\beta,R}(f) = \frac{\int_{-R}^{R} \sigma^*(z) e^{f(z)/\beta} \varphi(z)\, dz}{\int_{-R}^{R} e^{f(z)/\beta} \varphi(z)\, dz},$$

where φ is the standard Gaussian density. We chose $R = 2.3$ so that the evaluation interval $[-R, R]$ contains the maximizer of σ^* at $z \approx 1.83$.

We match the temperatures, $\beta^* = \beta_2$, and evaluate the policies induced by the true and fitted rewards on the same interval. We measure

$$G^{\mathrm{w}}(\beta_2, T_2) = \left| V_{\beta_2,R}(\sigma^*) - V_{\beta_2,R}(\widehat{r}^{\mathrm{w}}) \right| = |T^{\mathrm{w}}_{\mathrm{learn}}(R)|.$$

This comparison isolates the contribution of second-layer learning to the value gap.

Persistent sample threshold. Let $m^{\mathrm{w}}(\beta, T)$ denote the median of $G^{\mathrm{w}}(\beta, T)$ over the 60 trials. Figure 2 shows the smallest sample size satisfying

$$\widehat{T}^{\mathrm{w}}(\beta) = \min\{T \in \mathcal{T}_2 : m^{\mathrm{w}}(\beta, T') \le 0.05 \text{ for every } T' \in \mathcal{T}_2 \text{ with } T' \ge T\}.$$

Finite-sample learning curves can be nonmonotone, so the median must remain at or below 0.05 at every larger tested sample size. For example, under surrogate weighting at $\beta_2 = 0.35$ and 0.7, the median first falls below 0.05 at $T_2 = 32$, but the persistent threshold is 256 in both cases.

Table 1: Smallest tested sample size at which the median absolute learning gap stays at or below 0.05, using 60 trials per condition.

$\beta_2 = \beta^*$	0.25	0.35	0.5	0.7	1	1.5	2	3
Label	32	32	16	16	16	16	16	16
Surrogate	64	256	2048	256	16	16	16	16

The differences in Table 1 are consistent with the predicted dependence of sample requirements on the proxy reward. This experiment compares the two weighting schemes. It does not test monotonicity in temperature or estimate the exponent in the theoretical upper bound.

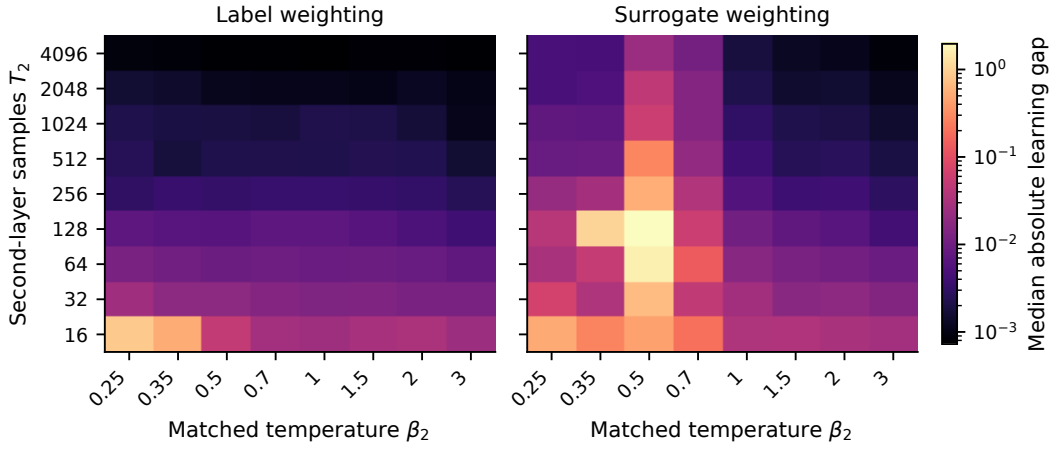


Figure 3: Median absolute learning gaps for all combinations of eight temperatures and nine sample sizes. Both panels use the same logarithmic color scale.

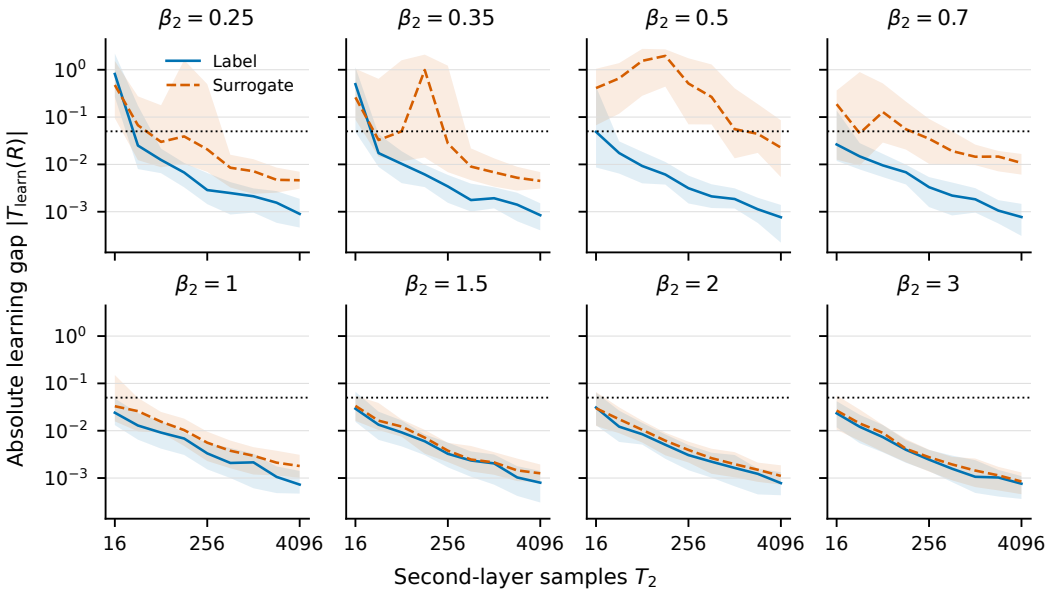


Figure 4: Learning curves at each tested temperature. Blue solid and orange dashed lines show the median absolute learning gap for label and surrogate weighting, respectively. Bands span the 25th to 75th percentiles across 60 trials. The dotted line marks the tolerance 0.05.

I Surrogate-weighted fitting on the learned subspace

We use the surrogate-weighted regression and policy deployment rules of Sections 5.1 and 5.3. The T_2 fresh pairs (x_i, ζ_i) are drawn i.i.d. from $\pi_{\mathrm{ref}} \otimes P_\zeta$, independently of the first-stage history, with centered bounded noise. In both regression and deployment, we replace the truncation set B_R by

$$\widehat{B}_R = \{x : \|Px\| \le R\}, \qquad P = P_{\widehat{S}}, \qquad \widehat{S} = \mathrm{span}\{w_1, \dots, w_N\}.$$

Write $\hat{r}$ and $\hat{\pi}$ for the resulting reward estimate and policy. To bound the regression weight by one, subtract the supremum of the proxy reward on $\widehat{B}_R$ and set

$$\omega(x) = \mathbf{1}_{\widehat{B}_R}(x) \exp\left(\frac{r_{a_0}(x) - \sup_{\widehat{B}_R} r_{a_0}}{\beta_2}\right), \qquad \mathcal{C} = \left(\mathbb{E}_{\mathrm{ref}}[\omega]\right)^{-1}, \qquad d\nu = \mathcal{C}\omega \, d\pi_{\mathrm{ref}}.$$

Draw the biases i.i.d. as in (3), independently of the pre-bias history. Assume that with probability at least $1 - \delta_1$, at least ρN directions satisfy $\langle w_j, \theta^* \rangle \ge 1 - \varepsilon$, where $0 < \rho \le 1$ and $\rho N \ge 4 \log(2N/\delta_b)$. Denote the conditional mean of the true reward given the projection by $g(z) = \mathbb{E}[r^*(x) \mid Px = z]$.

Proposition I.1 (Prediction error of the surrogate-weighted ridge estimator)**.** *For the weight ω, take $\lambda = 8 \sup_{|v| \le R + C_b} |\sigma(v)|^2 / (N\sqrt{\mathcal{C} T_2 \delta_0})$. Then, with probability at least $1 - \delta_1 - \delta_b - \delta_x - 4\delta_0$,*

$$\|\hat{r} - g \circ P\|_{L^2(\nu)} \le B_{\mathrm{pred}} = \widetilde{O}\left(\mathcal{C}^{1/2}\left[\frac{1}{\rho N} + \sqrt{\varepsilon}\right] + \frac{\mathcal{C}^{1/4}}{\sqrt{\rho}(T_2\delta_0)^{1/4}}\right). \tag{7}$$

Under the proxy mass condition of Lemma E.5 on the projected ball, its proof gives $\mathcal{C} \lesssim \beta_2^{-\dim \widehat{S}}$ for small β_2.

Theorem I.2 (Value gap of the surrogate policy)**.** *Take λ as in Proposition I.1. Under the temperature condition of Proposition 5.1 and the radius condition of Lemma C.1, with probability at least $1 - \delta_1 - \delta_b - \delta_x - 4\delta_0$,*

$$\begin{aligned} |V(\pi_{\beta^*}) - V(\hat{\pi})| &\le C_{\mathrm{temp}} |\beta^* - \beta_2| + C_{\mathrm{cut}} \beta_2 \\ &\quad + \frac{K}{\beta_2} \sqrt{B_{\mathrm{pred}}^2 + D_*^2 \eta^2 (1 + R^{q^*-1})^2}, \qquad \eta = \|(I - P)\theta^*\| \le \sqrt{2\varepsilon}. \end{aligned} \tag{8}$$

Here $V(\pi) = \mathbb{E}_\pi[r^*(x)]$, D_* depends only on σ^*, and K is the finite coverage factor obtained by applying Lemma D.1 on $\widehat{B}_R$.

Proof of Proposition I.1. Condition on the frozen directions, biases, and a_0 when taking expectations over the second-stage samples. Under the reference Gaussian, $\langle (I - P)\theta^*, x \rangle$ is independent of Px and has distribution $\mathcal{N}(0, \eta^2)$. Consequently,

$$g(z) = \mathbb{E}_G[\sigma^*(\langle P\theta^*, z \rangle + \eta G)], \qquad G \sim \mathcal{N}(0, 1).$$

Since ω depends only on Px, changing the reference law to ν preserves the conditional law given Px. The noise is independent and centered, so $g(Px) = \mathbb{E}_\nu[y \mid Px]$.

On recovery, put $\gamma = \|P\theta^*\| > 0$ and $e = P\theta^*/\gamma$. Then $g(z) = h(\langle e, z \rangle)$, where $h(u) = \mathbb{E}_G[\sigma^*(\gamma u + \eta G)]$ is a polynomial of degree at most $q^* \le q$ with coefficients uniformly bounded over $\gamma^2 + \eta^2 = 1$. To approximate h with the learned features, first write

$$h(u) = \frac{1}{2C_b} \int_{-C_b}^{C_b} v_h(b) \sigma(u + b) \, db.$$

Since the leading coefficient of σ is nonzero, matching the coefficients of u in descending degree determines the required moments of v_h of orders 0 through q. The moment matrix of $1, b, \dots, b^q$ under the uniform law is positive definite. Thus a polynomial v_h with these moments can be chosen linearly in the coefficients of h, and $V_b = \sup_h \|v_h\|_\infty < \infty$ over this family.

Let M be the number of recovered directions and partition the bias interval into $m = \lfloor M/(2\log(2M/\delta_b)) \rfloor$ equal bins. Conditional on the directions, the probability that a given bin contains none of their M biases is $(1 - 1/m)^M \le e^{-M/m}$. The probability that some bin is empty

is therefore at most $me^{-M/m} \le \delta_b$. When all bins are occupied, select in each bin I the least index j of a recovered direction with bias in I, set $a^\sharp_j = N(2C_b)^{-1}\int_I v_h(b)\,db$, and set all other coefficients to zero. Since $\rho N \le M \le N$,

$$\frac{1}{m} \le \frac{4\log(2N/\delta_b)}{\rho N}, \qquad \|a^\sharp\|^2 \le \frac{N^2 V_b^2}{m} \le \frac{4V_b^2 N\log(2N/\delta_b)}{\rho}.$$

For each selected direction, $\|w_j - e\|^2 = 2 - 2\langle w_j, \theta^*\rangle/\gamma \le 2\varepsilon$. By a Gaussian union bound, with conditional probability at least $1-\delta_x$, the projections of all second-stage samples onto e, the selected w_j, and $(w_j - e)/\|w_j - e\|$ whenever $w_j \ne e$, are at most $H = \sqrt{2\log(2(2N+1)T_2/\delta_x)}$ in absolute value. When the integral on each bin is approximated by its selected feature, the activation arguments differ by at most $2C_b/m$ from the bias displacement and $H\sqrt{2\varepsilon}$ from the direction displacement. The mean value theorem and summation over the bins give

$$\max_i |r_{a^\sharp}(x_i) - g(Px_i)| \le E, \qquad E = V_b \sup_{|u|\le H + C_b} |\sigma'(u)| \left(\frac{8C_b\log(2N/\delta_b)}{\rho N} + H\sqrt{2\varepsilon}\right).$$

We now apply the argument of Lemma E.2 with target $g(Px)$ and residual $\xi = y - g(Px)$. Let $Q(a)$ and $Q_{T_2}(a)$ denote the population and empirical unnormalized weighted squared errors relative to $g(Px)$. Write $\omega_i = \omega(x_i)$, $\psi_i = \psi_N(x_i)$, and $g_i = g(Px_i)$. Expanding the empirical error gives

$$Q_{T_2}(a) = \frac{1}{T_2}\sum_i \omega_i (a^\top\psi_i - g_i)^2 = a^\top\widehat{\Sigma}a - 2a^\top\widehat{b} + \widehat{c},$$

where

$$\widehat{\Sigma} = \frac{1}{T_2}\sum_i \omega_i\psi_i\psi_i^\top, \qquad \widehat{b} = \frac{1}{T_2}\sum_i \omega_i g_i\psi_i, \qquad \widehat{c} = \frac{1}{T_2}\sum_i \omega_i g_i^2.$$

Replacing these three sample averages by their expectations gives $Q(a)$. Fitting the observed labels $y_i = g_i + \xi_i$ also introduces the residual score $u = T_2^{-1}\sum_i \omega_i\xi_i\psi_i$.

Define the moments

$$V_2 = \mathbb{E}\left[\omega^2\xi^2\|\psi_N\|^2\right], \quad V_4 = \mathbb{E}\left[\omega^2\|\psi_N\|^4\right], \quad J_4 = \mathbb{E}\left[\omega^2 g(Px)^4\right].$$

Each summand of the score has mean zero because $\mathbb{E}[\xi \mid Px] = 0$ and ω, ψ_N depend only on Px. Sample independence then gives $\mathbb{E}\left[\|u\|^2\right] = V_2/T_2$. Second moments likewise control the deviations of $\widehat{\Sigma}, \widehat{b}$ and $\widehat{c}$ from their expectations. For $\widehat{b}$, Cauchy–Schwarz gives $\mathbb{E}\left[\omega^2 g(Px)^2\|\psi_N\|^2\right] \le \sqrt{J_4 V_4}$. Applying Markov's inequality to each squared deviation and taking a union bound shows that, with conditional probability at least $1 - 4\delta_0$,

$$\|u\|^2 \le \frac{V_2}{T_2\delta_0}, \qquad \left\|\widehat{\Sigma} - \mathbb{E}\left[\widehat{\Sigma}\right]\right\|_F^2 \le \frac{V_4}{T_2\delta_0},$$
$$\left\|\widehat{b} - \mathbb{E}\left[\widehat{b}\right]\right\|^2 \le \frac{\sqrt{J_4 V_4}}{T_2\delta_0}, \qquad \left|\widehat{c} - \mathbb{E}[\widehat{c}]\right|^2 \le \frac{J_4}{T_2\delta_0}.$$

Since $\hat{a}$ minimizes the ridge objective for the observed labels, its objective value is at most that of any a. Substituting $y_i = g_i + \xi_i$ and expanding the squares cancels the ξ_i^2 terms and gives

$$Q_{T_2}(\hat{a}) + \lambda\|\hat{a}\|^2 \le Q_{T_2}(a) + \lambda\|a\|^2 + 2\langle u, \hat{a} - a\rangle.$$

Cauchy–Schwarz and Young's inequality bound the last term by

$$\begin{aligned} 2\langle u, \hat{a} - a\rangle &\le 2\|u\|\|\hat{a}\| + 2\|u\|\|a\| \\ &\le \tfrac{\lambda}{2}(\|\hat{a}\|^2 + \|a\|^2) + \tfrac{4}{\lambda}\|u\|^2 \\ &\le \tfrac{\lambda}{2}(\|\hat{a}\|^2 + \|a\|^2) + \frac{4V_2}{T_2\delta_0\lambda}. \end{aligned}$$

In passing from $Q_{T_2}(\hat{a})$ to $Q(\hat{a})$, the covariance deviation contributes at most $\lambda\|\hat{a}\|^2/8$ when $\lambda \ge 8\sqrt{V_4/(T_2\delta_0)}$. By Young's inequality, the deviation of $\widehat{b}$ contributes at most $\lambda\|\hat{a}\|^2/8 +$

$8\sqrt{J_4 V_4}/(T_2\delta_0\lambda)$. Absorbing the terms in $\|\hat{a}\|^2$ into the penalty on the left and including the deviation of $\widehat{c}$, we obtain

$$Q(\hat{a}) \le Q_{T_2}(a) + \tfrac{3}{2}\lambda\|a\|^2 + \frac{4V_2 + 8\sqrt{J_4 V_4}}{T_2\delta_0\lambda} + \sqrt{\frac{J_4}{T_2\delta_0}}.$$

Put $F = \sup_{|v|\le R+C_b} |\sigma(v)|$. On $\widehat{B}_R$, $\|\psi_N\|^2 \le F^2/N$. Let

$$J_1 = \sup_{\substack{|v|\le R\\ 0\le t\le 1}} \left|\mathbb{E}[\sigma^*(v+tG)]\right|, \qquad J_2 = \sup_{\substack{|v|\le R\\ 0\le t\le 1}} \mathbb{E}\left[\sigma^*(v+tG)^2\right] + \mathbb{E}\left[\zeta^2\right].$$

Then $|g| \le J_1$ and $\mathbb{E}\left[\xi^2 \mid Px\right] \le J_2$ on $\widehat{B}_R$. Since $\omega^2 \le \omega$ and $\mathbb{E}_{\mathrm{ref}}[\omega] = 1/\mathcal{C}$,

$$V_2 \le \frac{F^2 J_2}{N\mathcal{C}}, \qquad V_4 \le \frac{F^4}{N^2\mathcal{C}}, \qquad J_4 \le \frac{J_1^4}{\mathcal{C}}.$$

The stated λ therefore satisfies the required lower bound. Insert $a = a^\sharp$ and this λ into the ridge inequality and use $Q_{T_2}(a^\sharp) \le E^2$. Since $\|\hat{r} - g\circ P\|^2_{L^2(\nu)} = \mathcal{C}Q(\hat{a})$, multiplying by $\mathcal{C}$ gives the squared error bound

$$B^2_{\mathrm{pred}} = \mathcal{C}E^2 + \frac{\sqrt{\mathcal{C}}}{\sqrt{T_2\delta_0}}\left[\frac{48F^2V_b^2\log(2N/\delta_b)}{\rho} + \frac{J_2}{2} + 2J_1^2\right]. \tag{9}$$

A union bound gives probability at least $1 - \delta_1 - \delta_b - \delta_x - 4\delta_0$ that recovery, bias occupancy, the projection bounds for the second-stage samples, and the ridge deviation bounds all hold. For fixed R and C_b, taking square roots in (9) gives (7). □

Proof of Theorem I.2. On the joint event used in the proof of Proposition I.1, the prediction bound holds and at least one direction is recovered. For a fixed fit, $\hat{r}(x)$ is a function of Px. Since $g(Px) = \mathbb{E}_\nu[r^*(x) \mid Px]$,

$$\mathbb{E}_\nu[(\hat{r}(x) - g(Px))(g(Px) - r^*(x)) \mid Px] = 0.$$

Indeed, conditional on Px, the first factor is constant and the second has mean zero. Squaring $\hat{r} - r^* = (\hat{r} - g\circ P) + (g\circ P - r^*)$ and taking expectations thus gives

$$\|\hat{r} - r^*\|^2_{L^2(\nu)} = \|\hat{r} - g\circ P\|^2_{L^2(\nu)} + \mathbb{E}_\nu[\mathrm{Var}(r^*(x) \mid Px)].$$

Put $v = \langle P\theta^*, Px\rangle$. Given Px, the true reward has the law of $\sigma^*(v + \eta G)$. The Gaussian Poincaré inequality yields

$$\mathrm{Var}(r^*(x) \mid Px) \le \eta^2 \mathbb{E}_G\left[|\sigma^{*\prime}(v+\eta G)|^2\right].$$

On $\widehat{B}_R$, $|v| \le \|P\theta^*\|\,\|Px\| \le R$. Since $\sigma^{*\prime}$ has degree $q^* - 1$, $\eta \le 1$, and all Gaussian moments are finite,

$$\mathbb{E}_G\left[|\sigma^{*\prime}(v+\eta G)|^2\right] \le D_*^2(1 + R^{q^*-1})^2 \qquad (|v| \le R).$$

Multiplying by η^2 gives the squared projection error in (8). Every recovered w_j belongs to $\widehat{S}$, so $\eta \le \|w_j - \theta^*\| \le \sqrt{2\varepsilon}$.

On $\widehat{B}_R$, interpolate between the true and fitted rewards by $f_t = (1-t)r^* + t\hat{r}$ for $0 \le t \le 1$. The corresponding policy and its normalizer are

$$d\pi_t = Z_t^{-1}\mathbf{1}_{\widehat{B}_R} e^{f_t/\beta_2}\, d\pi_{\mathrm{ref}}, \qquad Z_t = \int_{\widehat{B}_R} e^{f_t(x)/\beta_2}\, d\pi_{\mathrm{ref}}(x).$$

Here π_0 is the truncated true policy and $\pi_1 = \hat{\pi}$ is the fitted policy. Their value difference is the integral of $\frac{d}{dt}V(\pi_t)$ from $t = 0$ to $t = 1$.

Take $0 < r < R$ and let $A = \{x : \|x\| \le r\}$. Since $\|Px\| \le \|x\|$, $A \subseteq \widehat{B}_R$. The continuous functions r^* and $\hat{r}$ have a common lower bound m_0 on the compact ball A. For every t,

$$Z_t \ge \int_A e^{f_t(x)/\beta_2}\, d\pi_{\mathrm{ref}}(x) \ge \pi_{\mathrm{ref}}(A)e^{m_0/\beta_2} > 0.$$

Both $\hat{r}$ and r_{a_0} are bounded on the compact projected ball, and r^* is bounded above globally. Thus $f_t - r_{a_0}$ is bounded above uniformly in t. Together with the lower bound on Z_t, this gives a uniform bound on the density ratio for the realized fit: on $\widehat{B}_R$,

$$\frac{d\pi_t}{d\nu}(x) = \frac{\int_{\widehat{B}_R} e^{r_{a_0}(v)/\beta_2}\, d\pi_{\mathrm{ref}}(v)}{Z_t} \exp\!\left(\frac{f_t(x) - r_{a_0}(x)}{\beta_2}\right).$$

Differentiating the exponential weight in $V(\pi_t)$ introduces the factor $(\hat{r} - r^*)/\beta_2$. The weight is uniformly bounded above and the remaining factors grow at most polynomially in x. They are integrable under the reference Gaussian, so we can differentiate the numerator and denominator under the integral sign. The quotient rule gives

$$\begin{aligned}\frac{d}{dt}V(\pi_t) &= \frac{1}{\beta_2}\Big(\mathbb{E}_{\pi_t}[r^*(\hat{r} - r^*)] - \mathbb{E}_{\pi_t}[r^*]\ \mathbb{E}_{\pi_t}[\hat{r} - r^*]\Big) \\ &= \frac{1}{\beta_2}\operatorname{Cov}_{\pi_t}(r^*, \hat{r} - r^*).\end{aligned}$$

Apply Lemma D.1 to the integral of this covariance. Write $K = \mathcal{A}_{p,q',R}$ for its factor, with $q' > 2$ and $1/p + 1/q' = 1/2$. The density-ratio bound and the finite polynomial moments of the Gaussian give $K < \infty$, and the absolute learning term satisfies

$$|V(\pi_1) - V(\pi_0)| \le \frac{K}{\beta_2}\|\hat{r} - r^*\|_{L^2(\nu)}.$$

Finally, $\widehat{S} \subseteq S$ implies $\widehat{B}_R^c \subseteq B_R^c$. Let $m = \dim S$ and $U = \langle \theta^*, x\rangle$. Under π_{β_2}, $\|P_S x\|^2$ has the law of U^2 plus an independent χ^2_{m-1} variable. The space S is spanned by $N+1$ directions in $\mathbb{R}^d$, so $m \le k = \min(d, N+1)$. For independent standard Gaussian variables $G_1, \dots, G_{k-1}$ independent of U,

$$U^2 + \sum_{j=1}^{m-1} G_j^2 \le U^2 + \sum_{j=1}^{k-1} G_j^2.$$

The probability that the left-hand side exceeds R^2 is therefore at most the corresponding probability for the right-hand side. For $R > \max_{i \in I_{\max}} |u_i|$, Lemma C.1 at the deterministic dimension k gives, for some $\bar{\alpha}_{R,k} < 1$,

$$\pi_{\beta_2}(\widehat{B}_R^c) \le \pi_{\beta_2}(B_R^c) \le \bar{\alpha}_{R,k}.$$

Cauchy–Schwarz and Lemma B.4 bound the absolute cut error by

$$\frac{\sqrt{\operatorname{Var}_{\pi_{\beta_2}}(r^*)}\sqrt{\bar{\alpha}_{R,k}}}{1 - \bar{\alpha}_{R,k}} \le C_{\mathrm{cut}}\beta_2,$$

where C_{cut} depends on R and k. Adding the temperature term from Proposition 5.1 gives (8). □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's scope as a Gaussian single-index analysis of reward-weighted feature recovery and tilted-policy value gaps. The stated contributions are developed in the main technical sections, and the claims are restricted to the paper's upper-bounded polynomial, bounded-noise, and projected-truncation setting.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discusses limitations in the conclusion and discussion, including the Gaussian single-index abstraction and the use of projected truncation in the policy-value analysis.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are stated in the problem setup and in the theorem statements, and the main theoretical results are accompanied by proofs in the appendix or supplemental sections, with the relevant lemmas and intermediate results cross-referenced.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendix H specifies the synthetic data distribution, training procedures, parameter grids, number of trials, and evaluation criteria for both experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The simulation code and per-trial results are not publicly available. Appendix H describes how the synthetic data are generated and how both experiments are run.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix H specifies the initialization, learning rate, and sample budget for spherical SGD, as well as the weighting rules, ridge penalty, and policy evaluation procedure for the second-layer experiment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Figure 1 reports the fraction of 40 independent trials that achieve weak recovery at each setting. Figure 4 shows the median absolute learning gap across 60 trials, with bands spanning the 25th to 75th percentiles to indicate trial-to-trial variability.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [No]

   Justification: We omit runtime and memory measurements because the experiments evaluate how temperature and sample size affect learning, rather than computational efficiency.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and believe that the research conforms to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The paper is a theoretical study of an idealized two-layer neural reward model in a Gaussian single-index setting, and it does not release data, models, or deployed systems. We therefore do not identify direct positive or negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release data or models with a high risk of misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The paper does not use existing code, datasets, or model assets as part of the research contribution.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.