PERFECT PREFERENCE FIT, WRONG POLICY: LIMITS OF DPO UNDER AGGREGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a policy fit every observed preference perfectly and still allocate its probability mass incorrectly? We study population direct preference optimization (DPO) after response identities are replaced by semantic classes, using the class allocation from fine-policy optimization as the benchmark. Sparse comparison designs separate three guarantees: fitting preferences, repairing one environment, and preserving the target across environments. On paths, small within-class variation produces sharp regret $\Theta(\min\{n^2V^2, 1\})$ despite exact fitting and optimal choices of positive weights, temperature, and regular proper loss. This rate persists under arbitrary small offsets to the repeated class means. Closing the path restores preservation at the symmetric baseline, but nearby environments expose a gap: each admits exact recovery by balanced weights, while every shared balanced weighting has a worst-case lower bound even with a separately optimized temperature. Equal weights attain a matching upper bound. A variance-factored link representation extends the cycle rates to every sufficiently small-support common reward law, uniformly in graph size and atom masses. The benchmark measures loss of a KL-regularized objective; it can be positive with unchanged expected reward. Tree and dense-graph upper bounds give complementary positive cases, while sharp prediction on general sparse graphs remains open.

1 INTRODUCTION

Fitting every comparison need not recover the policy those comparisons were meant to train. Preference optimization turns local judgments into a global allocation of probability. If the identities of compared responses are discarded and only their semantic classes remain, the local probabilities change nonlinearly. We show how a comparison graph can accumulate that small change into a large policy error, how one additional comparison can repair it, and why that repair need not survive a change of environment. Figure 1 gives the construction before the formal statements.

Consider a preference-trained router over response intents such as clarification, a short explanation, and a worked example. Each intent has a pool of texts. For an annotation, independently draw a text from each of two pools using the reference probabilities conditional on the intent. A judge sees the texts, but the router stores only intent identifiers and the winner. Class-level DPO then adjusts intent masses. This is a possible data interface, not a claim about existing deployments. The known partition isolates the cost of discarding representative identity; responses within an intent may still differ in utility. Our benchmark is the intent allocation obtained by optimizing over all texts. A reference can encode a desired mix of response formats: suppressing equally valued formats can leave mean reward unchanged while raising the cost of departing from that mix. Our criterion measures this allocation loss, not necessarily poorer answers. A fixed renderer has a different optimal router in general (section 3.1).

Unrestricted population DPO (Rafailov et al., 2023) fits a Bradley–Terry (BT) score and applies a reference-relative exponential policy map. This established representation also underlies heterogeneous-preference distortion analysis (Gölz et al., 2025, App. E.3). Write $\mathsf{S}_\rho(a)_z = \rho_z e^{a_z} / \sum_u \rho_u e^{a_u}$, let v_w be the fitted class score, and let $C_\#$ aggregate a fine policy. The preservation question is

$$\underbrace{\mathsf{S}_\rho(v_w/\tau)}_{\text{aggregate, then optimize}} \quad = \quad \underbrace{C_\#\pi_\beta^{\mathrm{F}}}_{\text{optimize, then aggregate}} \quad . \tag{1}$$

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

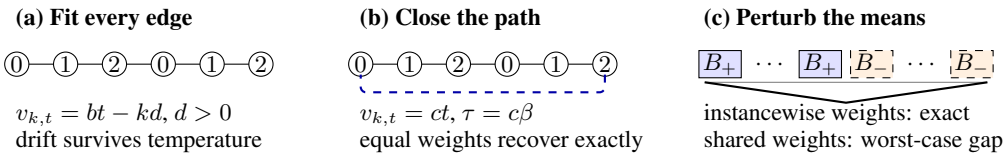

Figure 1: **Fit, repair, and stability are different guarantees.** Panels (a)–(b) repeat class means $0, 1, 2$; k indexes blocks and t their mean level. The dashed edge closes the full path. Panel (c) groups perturbed blocks $B_\pm = (0, a_\pm, 2a_\pm)$, $a_\pm = 1 \pm \xi/2$; the line illustrates equal-weight anchor drift. Swapping signs gives the paired lower-bound environment. Temperature calibrates (b), but suppressing drift in (a) also flattens its desired within-block differences. Theorems 1 to 3 give the stated guarantees in their neighborhoods: the perturbed-cycle lower bound covers every shared weighting of fixed balance, with a matching upper bound attained by equal weights.

We fix the target coefficient β and allow the training temperature τ to vary. Temperature rescales every score difference together. It can correct a common scale error; it cannot independently remove drift between classes that should receive the same reference-relative mass. Taking τ large also flattens the desired differences between classes. Our regret, $\beta\,\mathrm{KL}(p\|C_\#\pi_\beta^{\mathrm{F}})$, measures this tradeoff and equals the smallest fine-objective loss among lifts of p.

Fit, repair, and stability. A common translated within-class law cancels its log-MGF from the target, isolating distortion of the comparison link. Theorem 1 shows that this distortion accumulates on a path despite exact fitting and correct majority directions. Closing the path permits recovery; theorems 2 and 3 then separate environment-specific repair from shared-weight preservation. Each environment has its own fitted policy and optimal temperature. The sharp separation has constants and neighborhoods uniform in graph size. Corollary 4 removes exact mean repetition from the path rate; corollary 5 removes the cycle theorems' two-point restriction, without a lower bound on atom masses or on variance relative to support width.

Mixture misspecification and long-path distortion are established (Shirali et al., 2025; Gölz et al., 2025). Our contribution is quantitative preservation theory for these optimizer families. Class-dependent laws also move the target; the general residual and an identified three-class example expose that second effect. Tree and dense-graph bounds give positive cases. The cycle lower bounds retain their specific mean arrangements; general sparse sharp prediction and finite-data design selection remain open. All proofs are in the appendices.

2 Related Literature

From comparisons to policies. Preference-based reward learning (Christiano et al., 2017) and instruction tuning with human feedback (Ouyang et al., 2022) motivate the reward-fit/policy-optimize pipeline. DPO (Rafailov et al., 2023) reparameterizes the Bradley–Terry (BT) likelihood (Bradley & Terry, 1952) through a KL policy map. This representation is established, not our contribution. The ΨPO framework (Gheshlaghi Azar et al., 2024) and Nash learning (Munos et al., 2024) address limitations of replacing pairwise preferences with scalar rewards by changing the objective or solution concept. We instead fix the fine-policy target and test its preservation by ordinary population DPO and its equivalent BT-plus-KL implementation.

Heterogeneity and population distortion. Representative aggregation is a structured heterogeneous BT mixture (Lemma B.2). Shirali et al. (2025, Prop. 5.1, Eq. (14)) establish anonymous mean-ranking nonidentification and a variance expansion; Procaccia et al. (2025, Thm. 2.2) give a population impossibility and study clone robustness. Neither mixtures nor generic nonidentification are new here. Our auxiliary finite family has an identified target. Gölz et al. (2025, Thms. 6 and 9, App. F) prove utility-ratio distortion under a KL constraint, including unbounded distortion for sampling concentrated near a path. Thus path failure is also established. Our distinction is sharp additive regret under small common-law variation, correct coarse majority directions, all-temperature bounds, and the gap between shared and environment-specific weights. Sampling dependence and

sparse-design ranking errors also appear in Chen et al. (2026, Props. 28–29) and Tang et al. (2025, App. G).

Graph structure, projection, and statistical error. Topology-dependent minimax estimation (Shah et al., 2016) and BT score-error bounds (Li et al., 2022) quantify finite-sample uncertainty. Han et al. (2026, App. C, Thms. 3–4) connect comparison design to policy performance, assuming population target agreement in their upper bound. We test that agreement with exact probabilities. Potential/circulation decompositions are classical (Jiang et al., 2011, Sec. 5.1); BT Bernoulli-KL projection and optimum-preserving flows appear in Pukdee et al. (2026, Thm. 5.2, App. D.4). Xing (2026, Props. 2 and 7) gives local projection and comparison-KL expansions for a fixed graph. We use these tools, along with proper-composite derivatives (Reid & Williamson, 2010, Cor. 11); our sharp bounds require growing-graph uniformity and temperature optimization. The dense-coverage upper bound is a sufficient residual certificate, not a characterization of all safe sparse designs.

Alternative objectives and semantic targets. Distributionally robust DPO (Xu et al., 2025, Eq. (8), Sec. 5) optimizes worst-case training risk; our shared-weight theorem refits ordinary DPO in each environment. Parametric-policy misspecification (Gopalan et al., 2025, Prop. 3) differs from our unrestricted policy setting. Semantic entropy groups equivalent answers to measure uncertainty (Kuhn et al., 2023; Farquhar et al., 2024); semantic calibration (Nakkiran et al., 2025, Thm. 7) concerns sequence-data probabilities. Our benchmark is the class mass after fine optimization, whose log-MGF reward has an uncertainty-marginalization analogue (Hahami et al., 2026, Eq. (4)). Semantic per-pair temperatures (He et al., 2026) and hierarchical trajectory/action objectives (Gao et al., 2025) change the loss or annotations. They are outside our single-temperature model. Section E further separates these targets and guarantees.

3 Main Results

3.1 Experiment, target, and optimizer

Fix a finite response set $\mathcal{Y}$, a known surjection $C : \mathcal{Y} \to \mathcal{Z}$, a finite reward r, and a strictly positive reference π_{ref}. We suppress a single fixed prompt. For $n = |\mathcal{Z}| \geq 2$, let

$$\rho_z = \sum_{C(y)=z} \pi_{\mathrm{ref}}(y), \qquad \pi_{\mathrm{ref}}(y \mid z) = \pi_{\mathrm{ref}}(y)/\rho_z \quad (C(y) = z). \tag{2}$$

Here $\Delta_{\mathcal{Z}}$ is the probability simplex and $\Delta_{\mathcal{Z}}^{\circ}$ its strictly positive part; Δ_E° is defined in the same way. For a finite vector a, write $\mathrm{osc}(a) = \max_z a_z - \min_z a_z$. Let $G = (\mathcal{Z}, E)$ be a connected simple comparison graph with one orientation $e = (i, j)$ per edge and $(\mathsf{B}v)_e = v_j - v_i$. Independently sample $Y_i \sim \pi_{\mathrm{ref}}(\cdot \mid i)$ and $Y_j \sim \pi_{\mathrm{ref}}(\cdot \mid j)$ for each comparison. The fine BT scale is one:

$$\Pr(A = 1 \mid Y_i, Y_j) = \sigma(r(Y_j) - r(Y_i)), \qquad \sigma(t) = (1 + e^{-t})^{-1}. \tag{3}$$

Only the classes and label are retained. Their population probability is

$$Q_e = \mathbb{E}\sigma(r(Y_j) - r(Y_i)) \in (0, 1). \tag{4}$$

The representative conditionals are fixed, and Q is known exactly. Positive normalized pair weights $w \in \Delta_E^{\circ}$ apply equally to both outcomes of a pair. They change the objective weighting, leaving Q, r, π_{ref} fixed; there is no label-dependent selection.

For a fixed target coefficient $\beta > 0$, put

$$J_\beta(\pi) = \mathbb{E}_\pi r - \beta\,\mathrm{KL}(\pi \| \pi_{\mathrm{ref}}), \qquad \pi_\beta^{\mathrm{F}}(y) = \frac{\pi_{\mathrm{ref}}(y) e^{r(y)/\beta}}{\sum_{y'} \pi_{\mathrm{ref}}(y') e^{r(y')/\beta}}. \tag{5}$$

The desired class policy is $p_\beta^{\mathrm{F}} = C_{\#}\pi_\beta^{\mathrm{F}}$. Equivalently,

$$f_\beta(z) = \beta \log \mathbb{E}_{\pi_{\mathrm{ref}}(\cdot|z)} e^{r(Y)/\beta}, \qquad p_\beta^{\mathrm{F}} = \mathsf{S}_\rho(f_\beta/\beta). \tag{6}$$

The class regret and its exact fine-objective interpretation are

$$\mathrm{Reg}_\beta(p) = \beta\,\mathrm{KL}(p \| p_\beta^{\mathrm{F}}), \tag{7}$$

$$\mathrm{Reg}_\beta(p) = \min_{\pi : C_{\#}\pi = p} \{J_\beta(\pi_\beta^{\mathrm{F}}) - J_\beta(\pi)\}. \tag{8}$$

Thus lower bounds apply to every fine lift; upper bounds describe the best lift and do not assert recovery of within-class conditionals. Lemma B.3 proves these identities. This distinction matters for the router example. If its renderer keeps the reference conditionals, so $\pi_p(y) = p_{C(y)}\pi_{\mathrm{ref}}(y \mid C(y))$, its actual fine-objective loss is

$$J_\beta(\pi_\beta^{\mathrm{F}}) - J_\beta(\pi_p) = \mathrm{Reg}_\beta(p) + \beta \sum_z p_z \,\mathrm{KL}\big(\pi_{\mathrm{ref}}(\cdot \mid z) \| \pi_\beta^{\mathrm{F}}(\cdot \mid z)\big). \tag{9}$$

For this fixed renderer, with $m_z = \mathbb{E}_{\pi_{\mathrm{ref}}(\cdot|z)} r(Y)$,

$$J_\beta(\pi_p) = \sum_z p_z m_z - \beta\,\mathrm{KL}(p\|\rho), \qquad p_\beta^{\mathrm{router}} = \mathsf{S}_\rho(m/\beta). \tag{10}$$

Thus its optimal router generally differs from p_β^{F}; the two coincide under a common translated reward law. Class regret measures preservation of the fine-optimal allocation. Lemma B.3 also proves (10).

Unrestricted population logistic DPO at training temperature $\tau > 0$ is

$$\begin{aligned} v_w^{\mathrm{D}} &= \operatorname*{arg\,min}_{v:\,\sum_z v_z = 0} \sum_{e \in E} w_e \{\log(1 + e^{(\mathsf{B}v)_e}) - Q_e(\mathsf{B}v)_e\}, \\ p_{w,\tau}^{\mathrm{D}} &= \mathsf{S}_\rho(v_w^{\mathrm{D}}/\tau). \end{aligned} \tag{11}$$

The score is finite and unique, and independent of τ for fixed G, Q, w. This is the standard substitution $v_z = \tau \log(p_z/\rho_z)$, proved together with existence in lemma B.1. If fine representative identities were retained, the induced fine graph would be connected and BT-realizable, giving (5) at $\tau = \beta$. Hence (1) compares two exact population optimizations.

To state a stronger path result, let $\mathfrak{L}$ consist of differentiable binary losses with conditional derivative $\partial_s R_\ell(s;q) = (\sigma(s) - q)a_\ell(s)$ and $0 < a_\ell(s) < \infty$; see definition A.1 for all quantifiers. Let $\mathcal{A}$ contain the policies from all finite global minima, all $\ell \in \mathfrak{L}$, all $w \in \Delta_E^\circ$, and all $\tau > 0$. Define

$$\mathcal{R}_\beta^{\mathsf{env}} = \inf_{p \in \mathcal{A}} \mathrm{Reg}_\beta(p), \qquad \mathcal{R}_\beta^{\mathsf{temp}} = \inf_{\tau > 0} \mathrm{Reg}_\beta(p_{w^{\mathsf{unif}},\tau}^{\mathrm{D}}), \quad w_e^{\mathsf{unif}} = |E|^{-1}. \tag{12}$$

These are oracle approximation values; β stays fixed. Exact attainment requires finite positive temperature and strictly positive weights, although an infimum may approach their boundary. The larger envelope equals the logistic envelope after reweighting (lemma B.4); this equivalence is pointwise and does not preserve one shared weighting across environments.

3.2 Residual circulation and quantitative stability

Exact preservation requires $v = cf_\beta + \mathrm{constant}$, with $c = \tau/\beta > 0$. Define the target residual and its stationarity condition:

$$r_e(c) = \sigma\big(c(\mathsf{B}f_\beta)_e\big) - Q_e, \tag{13}$$

$$\mathsf{B}^\top \mathrm{diag}(w) r(c) = 0. \tag{14}$$

The weighted residuals must form a conserved flow. After discarding zero residuals and orienting the others by their signs, preservation under some positive weights is possible exactly when every directed edge lies on a directed cycle for some $c > 0$ (lemma B.4). This classical flow criterion covers arbitrary connected finite graphs and representative laws. A bridge cannot carry a nonzero circulation: on a tree, all target residuals must vanish. A cycle can preserve the target despite nonzero residuals, if their signs and weights permit conservation.

Two effects of discarding representative identity. Write $U_z = r(Y_z) - m_z$, $h_z = \beta \log \mathbb{E} e^{U_z/\beta}$, so $f_\beta = m + h$. Let $F_{ij}(t) = \mathbb{E}\sigma(t + U_j - U_i)$, with independent representatives, and put $\Delta m = m_j - m_i$, $\Delta h = h_j - h_i$. At the unscaled target the residual splits exactly as

$$r_{(i,j)}(1) = \underbrace{\sigma(\Delta m + \Delta h) - \sigma(\Delta m)}_{\text{aggregation-target shift}} + \underbrace{\sigma(\Delta m) - F_{ij}(\Delta m)}_{\text{comparison-link distortion}}. \tag{15}$$

A common translated law makes h_z constant and removes the first term; the path and cycle theorems isolate the second. With different class laws, both enter the divergence in proposition 1, and can reinforce or cancel. Small within-class variance controls the comparison term, but class-dependent log-MGFs need separate control (lemma B.9). The residual therefore links the restricted sharp theory to the general aggregation problem without identifying their scopes.

From residual divergence to policy regret. For a vertex i, consider its total comparison weight into any set containing at least $\lceil n/2 \rceil$ other classes. Set

$$\omega_G = \min_{i \in \mathcal{Z}} \min_{\substack{S \subseteq \mathcal{Z} \setminus \{i\} \\ |S| \geq \lceil n/2 \rceil}} \sum_{j \in S:\, \{i,j\} \in E} w_{ij}. \tag{16}$$

Here w_{ij} is the weight of the unordered pair, independent of its chosen orientation. The condition $\omega_G > 0$ holds on complete graphs and on a broader class of dense graphs; it need not hold on sparse graphs.

Proposition 1 (A residual bound under dense comparison coverage)**.** *In the general finite experiment, suppose $\omega_G > 0$. For $c > 0$, put*

$$b_c = \mathsf{B}^\top \operatorname{diag}(w) r(c), \quad M_c = c \operatorname{osc}(f_\beta), \quad \eta_c = \|b_c\|_\infty / \omega_G, \quad k_c = \sigma'(M_c + 1). \tag{17}$$

If $\eta_c \leq k_c/2$, the actual logistic optimizer satisfies

$$\operatorname{osc}(v_w^{\mathrm{D}} - c f_\beta) \leq \frac{2\eta_c}{k_c}, \qquad \inf_{\tau > 0} \operatorname{Reg}_\beta(p_{w,\tau}^{\mathrm{D}}) \leq \operatorname{Reg}_\beta(p_{w,c\beta}^{\mathrm{D}}) \leq \frac{\eta_c^2}{2c^2 \beta k_c^2}. \tag{18}$$

No common-law, small-support, or two-point assumption is needed. For common translated rewards of variance V, let $M = \operatorname{osc}(\mu)$ and $k_M = \sigma'(M + 1)$. If the minimum degree and weight ratio obey

$$d_{\min} + 1 \geq (1/2 + \gamma) n, \qquad \frac{\max_e w_e}{\min_e w_e} \leq \chi, \qquad 0 < \gamma \leq 1/2, \quad \chi \geq 1, \tag{19}$$

then

$$V \leq \frac{2\gamma k_M}{\chi} \quad \Longrightarrow \quad \operatorname{Reg}_\beta(p_{w,\beta}^{\mathrm{D}}) \leq \frac{\chi^2 V^2}{32 \gamma^2 \beta k_M^2}. \tag{20}$$

This holds for every centered finite-support common law, without a bound on its support width. For fixed γ, χ, M, β, the variance radius and bound are independent of n.

Section D.8 compares extreme score errors to at least half the classes. This is a sufficient upper bound using the true target; dense designs require order n^2 comparison types. Stars already show that density is unnecessary. The ratio η_c measures residual divergence against comparison mass opposing an extreme score error.

Why the aggregation target matters. Consider rewards $(-a, -\epsilon, \epsilon, a)$ with reference masses $(1/3, 1/6, 1/6, 1/3)$, merging the middle two responses. The three class means are $(-a, 0, a)$, but their effective rewards are

$$f_\beta = (-a, H_\beta, a), \qquad H_\beta = \beta \log \cosh(\epsilon/\beta) > 0. \tag{21}$$

The middle class gains fine-optimal mass because optimization can favor its better representative. Theorem 4 shows that the retained pairwise probabilities identify the target within this family, yet at $\beta = 1$ no choice in the optimizer envelope attains it. Its best regret is $\Theta(\epsilon^4)$ uniformly for $1 \leq a \leq 2$, $0 < \epsilon \leq 1$. Thus the noncancelling target effect can matter even when identification is not the obstacle.

3.3 Sharp amplification despite exact comparison fitting

Let ν be any finite-support centered law on $[-h, h]$, with variance V. Construct the fine experiment by

$$\mathcal{Y} = \mathcal{Z} \times \operatorname{supp} \nu, \quad C(z, u) = z, \quad \pi_{\mathrm{ref}}(z, u) = \rho_z \nu(u), \quad r(z, u) = \mu_z + u. \tag{22}$$

No lower bound on positive atom masses is imposed. For independent $U, U' \sim \nu$, put $D = U - U'$ and set

$$F_\nu(t) = \mathbb{E}\sigma(t + U - U'), \qquad g_\nu(t) = \operatorname{logit} F_\nu(t). \tag{23}$$

Symmetry of $U - U'$ makes g_ν odd, even if ν is asymmetric. The common exponential moment cancels from the target:

$$Q_{(i,j)} = F_\nu(\mu_j - \mu_i), \qquad p_\beta^{\mathrm{F}} = \mathsf{S}_\rho(\mu/\beta). \tag{24}$$

Moreover $F_\nu'(t) > 0$ and $F_\nu(0) = 1/2$, so every coarse majority agrees with the order of the class means. Here the common translated law isolates the comparison side of the problem: its log-MGF term is identical in every class and cancels, unlike (21). Thus the following negative results already hold when the target itself reduces to the mean-reward policy.

Table 1: **Parameters and scope.** n counts classes, not labels; $n = 3J$, and $J = 2L$ in the cycle lower bounds. The common law has support $[-h, h]$ and variance $V \le h^2$; in the two-point examples, $h = \epsilon$ and $V = \epsilon^2$. The mean-perturbation radius is ξ. All rates allow each environment its own temperature.

Result	Law and mean assumptions	Weights and guarantee
Path (Thm. 1; Cor. 4)	Any small-support common law; repeated means or arbitrary small offsets	All positive weights and allowed losses; sharp nV scale
Cycle upper (Thm. 2(i); Cor. 5(i))	Any small-support common law; every $\|z\|_\infty \le \xi$	Equal-weight logistic DPO; $nV\xi$ upper scale
Cycle lower (Thms. 2–3; Cor. 5)	Any small-support common law; specified two-region or paired means	Equal or shared fixed-balance weights; sharp $nV\xi$ scale
Positive cases (Prop. 1; Eq. (48))	General target residual; dense coverage or tree path sums	Sufficient upper bounds, not universal sharp rates

Theorem 1 (Sharp path amplification)**.** *Let $J \ge 2$, $n = 3J$, and index the classes by (k, t) with $0 \le k < J$ and $t \in \{0, 1, 2\}$. Set $\mu_{k,t} = t$, $\rho_z = 1/n$, and let G be the path in lexicographic order. There are universal $h_*, c_L, c_U > 0$ such that, for every common law* (22) *with $0 < h \le h_*$ and $V > 0$,*

$$c_L \min\{J^2V^2, 1\} \le \mathcal{R}_1^{\mathsf{env}} \le c_U \min\{J^2V^2, 1\}. \tag{25}$$

Every observed comparison is fitted exactly by every minimizer in the envelope. The constants are independent of graph size, support cardinality, atom masses, symmetry of ν, and V/h^2. Explicit values are given in equations (53) *and* (56)*.*

Since $n = 3J$, the order is also $\Theta(\min\{n^2V^2, 1\})$. Small support is essential to the stated uniformity; small variance alone is not the assumption. To see the mechanism, define

$$b = g_\nu(1), \qquad d_\nu = g_\nu(2) - 2g_\nu(1), \qquad \eta = d_\nu/b. \tag{26}$$

The variance-retaining expansion in lemma B.5 gives $d_\nu \asymp V > 0$. Every tree admits exact edge-logit fitting, so all positive weights and allowed proper losses give the same score, up to a constant:

$$v_{k,t} = -kd_\nu + tb. \tag{27}$$

For $s = b/\tau$, write $q_s(t) \propto e^{st}$ and $P_{a,J}(k) \propto e^{-ak}$. The KL factorization is exact:

$$\mathrm{Reg}_1(p^{\mathrm{D}}_{w,\tau}) = \mathrm{KL}(P_{\eta s,J} \| \,\mathrm{Unif}_J) + \mathrm{KL}(q_s \| q_1). \tag{28}$$

At $\tau = b$ ($s = 1$), both policies have type marginal q_1, so $\mathbb{E}_{p^{\mathrm{D}}}\mu = \mathbb{E}_{p_1^{\mathrm{F}}}\mu$. All regret is then the positive block KL: probability shifts between equally rewarded blocks increase the reference regularization cost. Thus "wrong policy" means failure to preserve the specified allocation, not a guaranteed reward decrease (equation (142)). For $s \ge 1/2$ the block term gives the lower bound; for $s \le 1/2$ the type term does. The scale nV measures accumulated score error.

This rate survives loss of exact repetition. Corollary 4 permits arbitrary $\mu_{k,t} = t + z_{k,t}$ with $\|z\|_\infty \le \zeta_{\mathrm{P}}$, where $\zeta_{\mathrm{P}} > 0$ is independent of n and V: $\mathcal{R}_\beta^{\mathsf{env}} \asymp_\beta \min\{n^2V^2, 1\}$ for every fixed $\beta > 0$. All target means may be distinct. The proof controls the perturbation of accumulated error by $O(nV\|z\|_\infty)$ and uses score geometry and separated groups in place of an exact KL factorization. On any tree of diameter D_G, proposition 2 instead bounds regret at $\tau = \beta$ by $D_G^2\mathcal{B}(h)^2V^2/(8\beta)$, with $\mathcal{B}(h) = (\cosh(2h) - 1)/(2h^2)$. A star on the same classes uses the same $n - 1$ comparison types and has an $O(V^2)$ upper bound independent of n. Bounded reference-mass ratios preserve the path order (lemma B.7).

3.4 Exact preservation and its local stability

Adding the endpoint pair closes the preceding path into a cycle. Proposition 3 proves that this is the unique single-pair addition permitting exact target agreement while every edge keeps positive weight. On this cycle, equal-weight DPO agrees with the target at $\tau = c_\nu\beta$, where

$$\sigma(c_\nu) + \sigma(2c_\nu) = F_\nu(1) + F_\nu(2). \tag{29}$$

This is a statement about one comparison type, not one observed label.

We now specialize to $U = \pm\epsilon$ with equal probabilities. On the cycle $i \to i+1$ modulo $n = 3J$, with $i = 3k + t$, use uniform class reference and means

$$\mu_i^0 = t, \qquad \mu_i = \mu_i^0 + z_i, \qquad \|z\|_\infty \le \xi \le 1/16, \quad 0 \le \epsilon \le \bar{\epsilon}. \tag{30}$$

The universal $\bar{\epsilon}$ is specified in equation (58). For positive ϵ this is the uniform reference on $2n$ fine responses; at zero it is the limiting degenerate common law. Put

$$F_\epsilon(t) = \tfrac{1}{4}\sigma(t-2\epsilon) + \tfrac{1}{2}\sigma(t) + \tfrac{1}{4}\sigma(t+2\epsilon),$$
$$Q_i = F_\epsilon(\mu_{i+1} - \mu_i), \qquad p^{\mathrm{F}}_{\mu,\beta} = \mathsf{S}_{\mathsf{unif}}(\mu/\beta), \tag{31}$$
$$\mathrm{Reg}_{\mu,\beta}(p) = \beta\,\mathrm{KL}(p\|p^{\mathrm{F}}_{\mu,\beta}), \qquad \mathcal{R}^{\mathsf{temp}}_{\mu,\beta} = \inf_{\tau>0} \mathrm{Reg}_{\mu,\beta}(p^{\mathrm{D}}_{w^{\mathsf{unif}},\tau}). \tag{32}$$

Define $\mathcal{R}^{\mathsf{env}}_{\mu,\beta}$ by using the same target regret in the envelope of (12). The baseline calibration is

$$\sigma(c_\epsilon) + \sigma(2c_\epsilon) = F_\epsilon(1) + F_\epsilon(2), \qquad c_0 = 1, \quad \tfrac{1}{2} \le c_\epsilon \le \tfrac{3}{2}. \tag{33}$$

Let $C_{\mathrm{sc}}, C_{\mathrm{up}} > 0$ be the universal constants specified in equation (60); their numerical values are not needed to interpret the rates.

Theorem 2 (Uniform upper bound and matching local lower bound)**.** *For the cycle experiment* (30)–(32)*:*

(i) For every $J \ge 2$, $\beta > 0$, and every allowed mean perturbation z, the actual equal-weight logistic score $v = v^{\mathrm{D}}_{w^{\mathsf{unif}}}$ satisfies

$$\mathrm{osc}(v/c_\epsilon - \mu) \le C_{\mathrm{sc}} n\epsilon^2\xi, \tag{34}$$
$$\mathrm{Reg}_{\mu,\beta}(p^{\mathrm{D}}_{w^{\mathsf{unif}},c_\epsilon\beta}) \le \frac{C_{\mathrm{up}}}{\beta} n^2\epsilon^4\xi^2. \tag{35}$$

In particular,

$$\mathcal{R}^{\mathsf{temp}}_{\mu,\beta} \le \min\left\{\frac{C_{\mathrm{up}}}{\beta} n^2\epsilon^4\xi^2, \frac{(2+2\xi)^2}{8\beta}\right\}. \tag{36}$$

No periodicity of the perturbed Q_i is assumed.

(ii) There exist universal constants $\epsilon_*, \xi_*, c_C, C_C > 0$, with $\epsilon_* \le \bar{\epsilon}$ and $\xi_* \le 1/16$, such that the following holds for every $0 < \epsilon \le \epsilon_*$, $0 < \xi \le \xi_*$, and every integer $L \ge 2$. Set $J = 2L$, $n = 6L$, $a_\pm = 1 \pm \xi/2$, and

$$\mu_{k,t} = \begin{cases} a_+ t, & 0 \le k < L, \\ a_- t, & L \le k < 2L. \end{cases} \tag{37}$$

Then $\|\mu - \mu^0\|_\infty \le \xi$, and at the fixed target coefficient one,

$$c_C \min\{n^2\epsilon^4\xi^2, 1\} \le \mathcal{R}^{\mathsf{temp}}_{\mu,1} \le C_C \min\{n^2\epsilon^4\xi^2, 1\}. \tag{38}$$

All four constants are independent of L, ϵ, and ξ within this neighborhood. The positive local radii are existence constants; part (ii) does not certify the whole numerical range in (30)*.*

Calibration removes the baseline mismatch. A mean perturbation changes its residual by order $\epsilon^2\xi$; the two-region construction makes these changes accumulate before cancelling around the cycle. Thus $n\epsilon^2\xi$ is score displacement after calibration. The upper bound permits every stated perturbation, while the lower bound uses the displayed construction and equal weights. The actual increments satisfy $v_{i+1} - v_i = \mathrm{logit}(Q_i + j)$ with a scalar root enforcing zero sum. Sections C.2 and C.3 retain the joint factor $\epsilon^2\xi$ through this equation and the all-temperature KL bound.

3.5 Shared weights versus instancewise recovery

Weights selected separately for each perturbation restore exact agreement: corollary 1 provides $1/(3n) \le w_i \le 3/n$ on a neighborhood $\xi \le \xi_{\mathrm{bal}}$ independent of n, ϵ, with the explicit radius given

in equation (58). Can one balanced weighting do as well throughout that neighborhood? The policy is refitted and the temperature may be chosen separately in every environment; only w is shared.

For fixed $\kappa \geq 3$, define

$$\mathcal{W}_\kappa(n) = \{w : \sum_i w_i = 1, \quad 1/(\kappa n) \leq w_i \leq \kappa/n\}, \tag{39}$$

and, writing $\mu(z) = \mu^0 + z$, set

$$\mathcal{R}_\kappa^{\text{fixed}}(n, \epsilon, \xi) = \inf_{w \in \mathcal{W}_\kappa(n)} \sup_{\|z\|_\infty \leq \xi} \inf_{\tau > 0} \text{Reg}_{\mu(z),1}(p^{\text{D}}_{w,\tau}(z)), \tag{40}$$

$$\mathcal{R}_\kappa^{\text{adapt}}(n, \epsilon, \xi) = \sup_{\|z\|_\infty \leq \xi} \inf_{w \in \mathcal{W}_\kappa(n),\, \tau > 0} \text{Reg}_{\mu(z),1}(p^{\text{D}}_{w,\tau}(z)). \tag{41}$$

Both use logistic DPO and target coefficient one.

Theorem 3 (A sharp gap between shared and instancewise weights)**.** *Fix $\kappa \geq 3$. There are constants $\epsilon_\kappa, c_\kappa, C_\kappa > 0$, with $\epsilon_\kappa \leq \bar{\epsilon}$, such that the following holds. Let $\xi_{\text{fixed}} > 0$ be the universal radius in equation* (58)*, and suppose*

$$n = 6L, \quad L \geq 2, \quad 0 \leq \epsilon \leq \epsilon_\kappa, \quad 0 \leq \xi \leq \xi_{\text{fixed}}.$$

Then

$$\mathcal{R}_\kappa^{\text{adapt}} = 0, \qquad c_\kappa \min\{n^2\epsilon^4\xi^2, 1\} \leq \mathcal{R}_\kappa^{\text{fixed}} \leq C_\kappa \min\{n^2\epsilon^4\xi^2, 1\}. \tag{42}$$

Two perturbations, fixed independently of w, suffice for the lower bound. With $k = 0, \ldots, 2L-1$, $t \in \{0, 1, 2\}$, and $s_k = 1$ for $k < L$ and $s_k = -1$ otherwise, put

$$\mu^{\pm}_{k,t} = (1 \pm s_k \xi/2)t. \tag{43}$$

For every $w \in \mathcal{W}_\kappa(n)$,

$$\max_{a \in \{+,-\}} \inf_{\tau > 0} \text{Reg}_{\mu^a,1}(p^{\text{D}}_{w,\tau}) \geq c_\kappa \min\{n^2\epsilon^4\xi^2, 1\}. \tag{44}$$

Constants and the heterogeneity neighborhood depend on κ, but not on L, ϵ, or ξ. The proof establishes existence of ϵ_κ; it does not certify a numerical value for that radius.

For clarity, put $\mathcal{E}(w) = \sup_{\|z\|_\infty \leq \xi} \inf_{\tau > 0} \text{Reg}_{\mu(z),1}(p^{\text{D}}_{w,\tau}(z))$ and $\Psi = \min\{n^2\epsilon^4\xi^2, 1\}$. The quantifiers are

$$\mathcal{E}(w) \geq c_\kappa \Psi \quad (w \in \mathcal{W}_\kappa(n)), \qquad \mathcal{E}(w^{\mathsf{unif}}) \leq C_\kappa \Psi. \tag{45}$$

The cycle root forces $n\epsilon^2\xi$ score displacement on equal-target anchors; section C.4 converts it to regret under separate temperatures.

At $z = 0$ and $\epsilon > 0$, fixed positive weights preserve the target exactly if and only if all short edges share one weight and all reset edges share another (proposition 4). Unequal short-edge weights can therefore leave positive regret even before perturbation. The appendix includes a one-off imbalance and its temperature endpoints; corollary 3 gives a repeated imbalance with a size-independent floor.

Fixed balance and broader common laws. The condition $nw_i \in [\kappa^{-1}, \kappa]$ keeps each comparison type within a fixed factor of uniform allocation. Six empirical distributions of $(nw_i)^{-1}$ lie in a common compact space, yielding size-uniform root derivatives. The restriction $\kappa \geq 3$ includes the recovery witness; a growing κ is not covered. Corollary 5 extends both cycle bounds and instancewise recovery to every sufficiently small-support common law, replacing ϵ^2 by V. The identity $F_\nu = \sigma + V\Phi$ keeps the factor V even for arbitrarily small atom masses. Two points simplify the canonical proofs; equal-target anchors and paired mean arrangements remain essential to the stated cycle lower-bound argument.

3.6 Growth and placement of aggregation error

Inside the stated neighborhoods, the sharp rates give

$$\mathcal{R}^{\mathsf{env}}_{1,n} \to 0 \iff nV_n \to 0 \quad \text{on the periodic paths.} \tag{46}$$

On cycles, set

$$\mathcal{E}_n^{\mathsf{unif}} = \sup_{\|z\|_\infty \le \xi_n} \mathcal{R}^{\mathsf{temp}}_{\mu^0+z,1}. \tag{47}$$

Both $\mathcal{E}_n^{\mathsf{unif}}$ and $\mathcal{R}_\kappa^{\mathrm{fixed}}$ vanish exactly when $n\epsilon_n^2\xi_n \to 0$; instancewise regret is zero. That condition is only necessary for arbitrary shared weights, whose baseline mismatch may persist (corollary 3). In contrast, $D_{G_n}V_n \to 0$ suffices on arbitrary trees, and $V_n \to 0$ suffices under the fixed coverage, balance, and mean-range conditions of proposition 1.

Placement also matters at a fixed graph and fixed local probability multiset. Alternating the same B_+, B_- blocks cancels their drift; grouping them accumulates it. At $\epsilon = \xi = n^{-1/3}$ the respective best-temperature regrets are $O(n^{-2})$ and $\Theta(1)$ (corollary 2 and figure 2). Graph topology and marginal comparison statistics alone therefore do not determine the regret.

4 Discussion

Exact fitting, instancewise repair, and preservation across environments are different guarantees. Regret measures a fixed KL-regularized objective: (28) includes a pure allocation failure with no expected mean-reward loss. The results also apply to equivalent BT-plus-KL procedures.

What extends to sparse graphs? Put $g_e = \operatorname{logit} Q_e$ and $\delta_e(c) = g_e/c - (\mathsf{B}f_\beta)_e$. For any tree, with path traversal signs $s_{ij,e}$,

$$\inf_{\tau>0} \mathrm{Reg}_\beta(p^{\mathrm{D}}_{w,\tau}) \le \inf_{c>0} \frac{1}{8\beta} \left(\max_{i,j} \left| \sum_{e\in P_{ij}} s_{ij,e}\delta_e(c) \right| \right)^2. \tag{48}$$

This bound permits arbitrary representative laws (lemma B.11). A short-diameter tree limits accumulation; a diameter bound discards cancellations visible in the path sums. Corollary 4 gives a size-uniform lower rate on an open neighborhood of the path means, with no exact anchors. It does not cover arbitrary sparse graphs. The cycle upper bound permits nonperiodic perturbations; its lower bounds still use repeated two-region blocks and, for shared weights, equal-target anchors. Fixed-size continuity (lemma B.12) gives no additional size-uniform cycle radius.

Graph, weights, link, and target coefficient. A closing edge permits residual circulation; reweighting changes its projection. If ν is known, correcting the link instead gives

$$d_e = F_\nu^{-1}(Q_e) = (\mathsf{B}\mu)_e, \qquad Q_e^{\mathrm{cal}} = \sigma(d_e). \tag{49}$$

DPO on these probabilities recovers the target at $\tau = \beta$ on any connected graph with positive weights (lemma B.10); this changes the comparison model. Likelihood correction under heterogeneity is already studied by Shirali et al. (2025, Sec. 5). The choice $\beta = 1$ normalizes presentation, not the data-generating reward. Proposition 5 proves the same orders for every fixed $\beta > 0$: calibrations and temperature cutoffs scale by β, and KL-conversion constants depend on β. No uniformity in extreme or size-dependent β is asserted.

Scope and design implications. The law extension preserves common-log-MGF cancellation; class-dependent laws also shift the target through (15). With common laws, a star gives $O(V^2)$ regret using the same number of pair types as the path's potentially constant regret. This is a concrete population design comparison, not a general optimal-design rule. Recovery weights, target-residual certificates, and known-law correction use oracle information. The explicit $\xi_{\mathrm{bal}} \approx 7.20 \times 10^{-5}$ is conservative; the lower-bound heterogeneity radii are existence constants. These guarantees establish structural limits and possibilities, not an empirical operating range or a finite-data collection algorithm.

AI USE STATEMENT

In this work, we used generative AI tools for spell and grammar check of English writing. We have not used generative AI tools for help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for identify relevant literature. We have reviewed all AI-assisted work. We checked all LLM-assisted survey results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All results concern finite, explicitly specified population experiments. Section A records the auxiliary statements and numerical constants, section B gives the supporting lemmas, and sections C and D give complete proofs. Section F proves the nonperiodic-path, fixed-target, and distribution-uniform cycle extensions. Local existence radii are distinguished from explicit sufficient radii. The single LaTeX source includes every appendix; the bibliography is provided separately as a BibTeX file.

REFERENCES

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika*, 39(3–4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324. URL https://academic.oup.com/biomet/article-abstract/39/3-4/324/326091.

Yurong Chen, Yu He, Michael I. Jordan, and Fan Yao. How sampling shapes LLM alignment: From one-shot optima to iterative dynamics, 2026. URL https://arxiv.org/html/2602.12180v1. arXiv preprint arXiv:2602.12180v1.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0.

Heyang Gao, Zexu Sun, Erxue Min, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Xu Chen. Solving the granularity mismatch: Hierarchical preference learning for long-horizon LLM agents, 2025. URL https://arxiv.org/html/2510.03253v1. arXiv preprint arXiv:2510.03253v1.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455. PMLR, 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Paul Gölz, Nika Haghtalab, and Kunhe Yang. Distortion of AI alignment: Does preference optimization optimize for preferences?, 2025. URL https://arxiv.org/html/2505.23749v1. arXiv preprint arXiv:2505.23749v1.

Aditya Gopalan, Sayak Ray Chowdhury, and Debangshu Banerjee. Why DPO is a misspecified estimator and how to fix it, 2025. URL https://arxiv.org/html/2510.20413v1. arXiv preprint arXiv:2510.20413v1.

Ely Hahami, Yoel Zimmermann, Ray Zhou, and Jack Benarroch Jedlicki. A unifying lens on reward uncertainty in RLHF, 2026. URL https://arxiv.org/html/2606.09073v1. arXiv preprint arXiv:2606.09073v1.

Jiangze Han, Vineet Goyal, and Will Ma. Which pairs to compare for LLM post-training?, 2026. URL https://arxiv.org/html/2606.19607v1. arXiv preprint arXiv:2606.19607v1.

Chaoyue He, Xin Zhou, Di Wang, Hong Xu, Wei Liu, and Chunyan Miao. SP^{2}DPO: An LLM-assisted semantic per-pair DPO generalization, 2026. URL https://arxiv.org/html/2601.22385v1. arXiv preprint arXiv:2601.22385v1.

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127:203–244, 2011. doi: 10.1007/s10107-010-0419-x. URL https://web.stanford.edu/~yyye/hodgeRank2011.pdf.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.09664.

Wanshan Li, Shamindra Shrotriya, and Alessandro Rinaldo. ℓ_∞-bounds of the MLE in the BTL model under general comparison graphs. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1178–1187. PMLR, 2022. URL https://proceedings.mlr.press/v180/li22g.html.

Remi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J. Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36743–36768. PMLR, 2024. URL https://proceedings.mlr.press/v235/munos24a.html.

Preetum Nakkiran, Arwen Bradley, Adam Goliński, Eugene Ndiaye, Michael Kirchhof, and Sinead Williamson. Trained on tokens, calibrated on concepts: The emergence of semantic calibration in LLMs, 2025. URL https://arxiv.org/html/2511.04869v1. arXiv preprint arXiv:2511.04869v1.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Ariel D. Procaccia, Benjamin Schiffer, and Shirley Zhang. Clone-robust AI alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49903–49926. PMLR, 2025. URL https://proceedings.mlr.press/v267/procaccia25a.html.

Rattana Pukdee, Maria-Florina Balcan, and Pradeep Ravikumar. What does preference learning recover from pairwise comparison data?, 2026. URL https://arxiv.org/html/2602.10286v2. arXiv preprint arXiv:2602.10286v2.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/html/2305.18290v3.

Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11(83):2387–2422, 2010. URL https://jmlr.org/papers/volume11/reid10a/reid10a.pdf.

Nihar B. Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17(58):1–47, 2016. URL https://jmlr.org/papers/v17/15-189.html.

Ali Shirali, Arash Nasr-Esfahany, Abdullah Alomar, Parsa Mirtaheri, Rediet Abebe, and Ariel Procaccia. Direct alignment with heterogeneous preferences, 2025. URL https://arxiv.org/html/2502.16320v1. arXiv preprint arXiv:2502.16320v1.

Shange Tang, Yuanhao Wang, and Chi Jin. Is Elo rating reliable? a study under model misspecification, 2025. URL https://arxiv.org/html/2502.10985v1. arXiv preprint arXiv:2502.10985v1.

Wenqian Xing. Attention limited reward learning, 2026. URL https://arxiv.org/html/2607.04590v1. arXiv preprint arXiv:2607.04590v1.

Zaiyan Xu, Sushil Vemuri, Kishan Panaganti, Dileep Kalathil, Rahul Jain, and Deepak Ramachandran. Distributionally robust direct preference optimization, 2025. URL https://arxiv.org/html/2502.01930v1. arXiv preprint arXiv:2502.01930v1.

A AUXILIARY RESULTS AND EXPLICIT CONSTANTS

A.1 LOSS FAMILY AND ORACLE ENVELOPE

Definition A.1 (Allowed proper losses)**.** Let $\mathfrak{L}$ be the family of differentiable binary score losses ℓ whose conditional risks $R_\ell(s;q) = q\ell(s,1) + (1-q)\ell(s,0)$ satisfy

$$\partial_s R_\ell(s;q) = (\sigma(s) - q)a_\ell(s), \qquad 0 < a_\ell(s) < \infty \quad (s \in \mathbb{R},\ q \in (0,1)). \tag{50}$$

The same loss is used on all pairs. Write

$$\begin{aligned} \mathcal{V}_{\ell,w} &= \operatorname*{arg\,min}_{v \in \mathbb{R}^n :\, \sum_z v_z = 0} \sum_e w_e R_\ell((\mathsf{B}v)_e; Q_e), \\ \mathcal{P}_{\ell,w,\tau} &= \{\mathsf{S}_\rho(v/\tau) : v \in \mathcal{V}_{\ell,w}\}. \end{aligned} \tag{51}$$

Only finite global minimizers are included; an empty argmin contributes no policies. Stationary points of a nonconvex proper risk are not included unless they are global minima.

This is the standard proper-composite derivative representation with a strictly positive derivative factor (Reid & Williamson, 2010, Cor. 11). Log loss belongs to $\mathfrak{L}$ with $a_\ell = 1$; the logistic-link Brier loss has $a_\ell = 2\sigma'$. Bare strict properness without the displayed regularity is not the assumption used here.

The attainable set used in the main text is

$$\mathcal{A} = \bigcup_{\ell \in \mathfrak{L},\ w \in \Delta_E^\circ,\ \tau > 0} \mathcal{P}_{\ell,w,\tau}. \tag{52}$$

All choices may depend on the population experiment and the fixed target. This is a feasibility envelope, not an estimator of those choices. In particular, $\mathcal{R}_\beta^{\mathsf{env}} \le \mathcal{R}_\beta^{\mathsf{temp}}$.

A.2 UNIFORM TREE BOUNDS AND CONSTANTS FOR THE MAIN THEOREMS

The constants used in the distribution-uniform statements are

$$\begin{aligned} \mathcal{B}(h) &= \frac{\cosh(2h) - 1}{2h^2}, \quad \mathcal{B}(0) = 1, \qquad \kappa_3 = \sigma'(3), \\ C_{\mathrm{mix}} &= \frac{1}{6\kappa_3} + \frac{1}{32\kappa_3^2}, \qquad K = 2\tanh(1/2) - \tanh 1 > 0, \\ h_* &= \min\left\{\frac{1}{4}, \sqrt{\frac{K}{6C_{\mathrm{mix}}}}\right\}. \end{aligned} \tag{53}$$

For orientation, $h_* \approx 0.03777$; its exact expression is used in all statements. The small-support hypothesis is stronger than small variance.

Proposition 2 (Uniform bound on any tree)**.** *In the common-law experiment* (22)*, let G be a tree of diameter D_G, with any positive class reference and any positive normalized pair weights. For every $h > 0$, $\beta > 0$, and $M = \mathrm{osc}(\mu)$,*

$$\mathrm{Reg}_\beta(p^{\mathrm{D}}_{w,\beta}) \le \frac{D_G^2 \mathcal{B}(h)^2 V^2}{8\beta}, \tag{54}$$

$$\inf_{\tau > 0} \mathrm{Reg}_\beta(p^{\mathrm{D}}_{w,\tau}) \le \frac{\min\{D_G^2 \mathcal{B}(h)^2 V^2, M^2\}}{8\beta}. \tag{55}$$

Every allowed proper loss has the same centered tree minimizer, so the bounds apply to all losses in $\mathfrak{L}$.

Valid constants in theorem 1 are

$$c_L = \frac{e^{-1}K^2}{1152}, \qquad c_U = \max\left\{\frac{9K^2}{8}, \log\frac{1 + e + e^2}{3} - 1\right\}. \tag{56}$$

The same hypotheses give the explicit drift estimates

$$\frac{1}{2} \le b \le \frac{3}{2}, \qquad \frac{K}{2}V \le d_\nu \le \frac{3K}{2}V, \qquad \frac{K}{3}V \le \eta \le 3KV. \tag{57}$$

For a nonuniform class reference with $\chi = \max_z \rho_z / \min_z \rho_z$, the path bound extends with constants c_L/χ and χc_U by lemma B.7. The representative law stays fixed. Uniformity in n requires a uniform bound on χ.

For theorems 2 and 3, the explicit upper-bound radius and the instancewise recovery radius are

$$\begin{aligned} \bar{\epsilon} &= \min\{h_*, 1/16, \sqrt{K/216}\}, \qquad \xi_{\mathrm{bal}} = K\kappa_3/102, \\ \xi_{\mathrm{fixed}} &= \min\{1/160, \xi_{\mathrm{bal}}\}. \end{aligned} \tag{58}$$

The local radii in the lower bounds can be smaller than these explicit upper-bound radii. The proofs establish that they are positive and independent of graph size.

The constants in the main cycle upper bound are

$$\alpha_0 = \sigma(-3), \qquad \kappa_* = (\alpha_0/2)(1 - \alpha_0/2), \tag{59}$$

$$C_{\mathrm{sc}} = 34/\kappa_*, \qquad C_{\mathrm{up}} = 289/(2\kappa_*^2). \tag{60}$$

The proof of theorem 2(i) in fact establishes the sharper bounds

$$\mathrm{osc}(v/c_\epsilon - \mu) \le \frac{17n\epsilon^2\xi}{c_\epsilon\kappa_*}, \tag{61}$$

$$\mathrm{Reg}_{\mu,\beta}(p^{\mathrm{D}}_{w^{\mathsf{unif}}, c_\epsilon\beta}) \le \frac{289n^2\epsilon^4\xi^2}{8\beta c_\epsilon^2\kappa_*^2}. \tag{62}$$

Using $c_\epsilon \ge 1/2$ gives the constants in the main statement. The deliberately loose derivative and residual bounds explain why these values are conservative.

A.3 An identified target outside the attainable set

Consider four fine responses with

$$r = (-a, -\epsilon, \epsilon, a), \quad \pi_{\mathrm{ref}} = (1/3, 1/6, 1/6, 1/3), \quad a > 0,\ \epsilon > 0. \tag{63}$$

The middle responses form class 2, and the outer responses form singleton classes 1 and 3. Thus $\rho = (1/3, 1/3, 1/3)$. Use all three edges, oriented as $(1,2), (2,3), (1,3)$. Define

$$\begin{aligned} u &= \cosh\epsilon, & q_s &= \frac{\sigma(a-\epsilon) + \sigma(a+\epsilon)}{2} = \frac{1}{2} + \frac{\sinh a}{2(\cosh a + u)}, \\ q_\ell &= \sigma(2a), & g_A &= \mathrm{logit}\, q_s = \log\frac{e^a + u}{e^{-a} + u}, \\ d_A &= a - g_A, & \zeta_A &= d_A/a. \end{aligned} \tag{64}$$

Here $Q = (q_s, q_s, q_\ell)$, $0 < g_A < a$, and $0 < d_A < \epsilon$. The long comparison has logit $2a > 2g_A$, so coarse comparisons are misspecified as a single BT model. Nevertheless, the full probabilities identify the parameters in this specified structural family:

$$a = \tfrac{1}{2}\,\mathrm{logit}\, q_\ell, \qquad u = \frac{\sinh a}{2q_s - 1} - \cosh a, \qquad \epsilon = \mathrm{arcosh}\, u. \tag{65}$$

This elementary inverse does not identify arbitrary within-class laws. For the target, put

$$H_\beta = \beta \log\cosh(\epsilon/\beta), \qquad p^{\mathrm{F}}_\beta = \frac{(e^{-a/\beta}, e^{H_\beta/\beta}, e^{a/\beta})}{Z_\beta}, \quad Z_\beta = 2\cosh(a/\beta) + \cosh(\epsilon/\beta). \tag{66}$$

Theorem 4 (Oracle attainability and a fourth-order floor)**.** *In* (63)–(66)*, the following hold.*

(i) The attainable set in (52) *is exactly*

$$\mathcal{A} = \left\{ \frac{(e^{-\alpha}, e^m, e^\alpha)}{2\cosh\alpha + e^m} : \alpha > 0,\ |m| < \zeta_A\alpha \right\}. \tag{67}$$

The same set is obtained by allowing only ordinary logistic DPO while retaining all positive pair weights and all $\tau > 0$.

(ii) The target is attained if and only if $H_\beta < d_A$. If $H_\beta = d_A$, then $\mathcal{R}_\beta^{\mathsf{env}} = 0$ but $p_\beta^{\mathrm{F}} \notin \mathcal{A}$. At equality, deleting edge $(1,2)$ while keeping the other two edges positive attains the target at $\tau = \beta$. If $H_\beta > d_A$, then $\mathcal{R}_\beta^{\mathsf{env}} > 0$. There is exactly one $\beta_c(a,\epsilon) > 1$ with $H_{\beta_c} = d_A$: exact positive-three-edge recovery holds for $\beta > \beta_c$ and positive oracle regret holds for $0 < \beta < \beta_c$.

(iii) For $H_\beta > d_A$, define

$$\lambda_A = \frac{g_A}{2a}, \qquad M_\beta = \max\{1, \lambda_A e^{(a+d_A)/\beta}, (1-\lambda_A)e^{(d_A-a)/\beta}\}.$$

Every p in the closure of $\mathcal{A}$ in the probability simplex obeys

$$\mathrm{Reg}_\beta(p) \geq \frac{\beta}{2M_\beta^2}\left(\frac{e^{H_\beta/\beta} - e^{d_A/\beta}}{Z_\beta}\right)^2. \tag{68}$$

(iv) Uniformly over every $1 \leq a \leq 2$ and $0 < \epsilon \leq 1$, at the fixed target coefficient $\beta = 1$,

$$c_A\epsilon^4 \leq \mathcal{R}_1^{\mathsf{env}} \leq \frac{\epsilon^4}{32}, \qquad c_A = \frac{1}{2e^8(e^2+\cosh 1)^2(2\cosh 2 + \cosh 1)^2}. \tag{69}$$

The lower bound remains valid if one pair may be omitted and the remaining support is connected. The stated upper bound follows from a policy obtained by equal-weight ordinary DPO at a finite positive training temperature.

Proof mechanism; complete proof in section D.2. At a minimizing score put $x = v_2 - v_1$ and $y = v_3 - v_2$. The positive residual-flow condition in lemma B.4 is equivalent to $x > g_A$, $y > g_A$, and $x + y < 2a$. Substituting $x = \tau(\alpha + m)$ and $y = \tau(\alpha - m)$ gives (67). The target has $(\alpha, m) = (a/\beta, H_\beta/\beta)$. The boundary and separation require closure arguments, not just exclusion from an open set. The cone implies $p_2 \leq p_1^{\lambda_A} p_3^{1-\lambda_A}$; a tangent bound for the weighted geometric mean followed by Pinsker gives (68). The proof treats the connected two-edge boundary separately and establishes part (iv) throughout the stated rectangle of parameters.

For an explicit existence witness when $H_\beta < d_A$, choose $g_A/(a - H_\beta) < c_0 < 1$ and $\tau = c_0\beta$. At the target score the edge residuals are

$$r_1 = \sigma(c_0(a + H_\beta)) - q_s > 0, \quad r_2 = \sigma(c_0(a - H_\beta)) - q_s > 0, \quad r_3 = \sigma(2c_0 a) - q_\ell < 0.$$

Weights proportional to $(r_1^{-1}, r_2^{-1}, -r_3^{-1})$ establish attainability. They are a proof witness using the true experiment.

The point of theorem 4 is that an identifiable target can lie outside the policies returned by every allowed objective. This is a restriction of the optimizer family, not an information-theoretic impossibility for all uses of Q. Raising the target β changes the question; it does not repair the original fixed target.

A.4 The unique single-pair addition

Proposition 3 (Endpoint comparison)**.** *Under the path construction and small-support conditions of theorem 1, now allow any positive class reference. Fix any $\beta > 0$. Among the unordered pairs absent from the path, the endpoint pair $\{(0,0),(J-1,2)\}$ is the unique pair whose addition permits $p_\beta^{\mathrm{F}} \in \mathcal{A}$ on the augmented graph, when all old and new edges must retain positive weight. This uniqueness allows every loss in $\mathfrak{L}$ and every positive global training temperature.*

On the cycle obtained from this endpoint addition, equal-weight ordinary DPO attains the target at $\tau = c_\nu\beta$, where c_ν is the unique solution of (29)*. It satisfies $b < c_\nu < g_\nu(2)/2$.*

Proof mechanism; complete proof in section D.3. Orient the closed path forward, including the final edge back to the first vertex. At $v_{k,t} = c_\nu t$, (29) makes the short-edge and reset-edge residuals equal, giving a constant circulation. For another added pair the graph is unicyclic. Nonzero stationary residual flows cannot occur on bridges. The uniqueness proof includes the breakpoints $c = b$ and $c = g_\nu(2)/2$, where one kind of edge residual vanishes. The result concerns one additional comparison *type*, not one label.

A.5 WHICH FIXED WEIGHTS PRESERVE THE UNPERTURBED CYCLE?

For later reference, the one-off imbalance is

$$w_0 = \frac{3}{2n}, \qquad w_1 = \frac{1}{2n}, \qquad w_i = \frac{1}{n} \quad (i \notin \{0,1\}), \tag{70}$$

and its three-periodic repetition is

$$w_{3k} = \frac{3}{2n}, \qquad w_{3k+1} = \frac{1}{2n}, \qquad w_{3k+2} = \frac{1}{n}. \tag{71}$$

Both are normalized and belong to $\mathcal{W}_3(n)$.

Proposition 4 (Baseline preservation at fixed weights)**.** *Consider the periodic cycle with $n = 3J$, $J \geq 2$, and $\mu^0_{k,t} = t$, under the common-law hypotheses of theorem 1. Allow any strictly positive class reference and fix $\beta > 0$. Write $b = g_\nu(1)$ and $g_2 = g_\nu(2)$, so $g_2 > 2b$. For a fixed positive normalized weighting w, the following are equivalent:*

(i) There is $\tau > 0$ such that $p^{\mathrm{D}}_{w,\tau} = p^{\mathrm{F}}_\beta$.

(ii) $\inf_{\tau>0} \mathrm{Reg}_\beta(p^{\mathrm{D}}_{w,\tau}) = 0$.

(iii) There are $w_s, w_r > 0$ such that

$$w_{3k} = w_{3k+1} = w_s, \qquad w_{3k+2} = w_r \quad (0 \leq k < J), \qquad 2Jw_s + Jw_r = 1. \tag{72}$$

If these conditions hold, the recovery temperature is $\tau = c\beta$, where $c \in (b, g_2/2)$ is the unique solution of

$$w_s\sigma(c) + w_r\sigma(2c) = w_sF_\nu(1) + w_rF_\nu(2). \tag{73}$$

Otherwise the infimum in (ii) is strictly positive for this fixed experiment and weighting. In particular, (70) *gives a positive baseline regret in the two-point experiment for every $0 < \epsilon \leq \bar{\epsilon}$, despite belonging to $\mathcal{W}_3(n)$. If the imbalance is instead repeated as in* (71)*, with uniform reference and a fixed common law ν, there is $\delta_{\nu,\beta} > 0$, independent of J, such that*

$$\inf_{\tau>0} \mathrm{Reg}_\beta(p^{\mathrm{D}}_{w,\tau}) = \delta_{\nu,\beta} \quad \textit{for every } n = 3J,\ J \geq 2. \tag{74}$$

This is a supporting specialization of the residual-circulation criterion, not a separate claim of a new general projection theorem. Outside the repeated construction in (74), the positive floor is pointwise in n, w, ν, ρ, β; no uniform rate is asserted. The full proof is in section D.6.

A.6 INSTANCEWISE BALANCED RECOVERY

Corollary 1 (Balanced oracle weights restore local attainability)**.** *In* (30)*, put $\xi_{\mathrm{bal}} = K\kappa_3/102$. For every $J \geq 2$, $\beta > 0$, $0 < \epsilon \leq \bar{\epsilon}$, and every z with $\|z\|_\infty \leq \xi \leq \xi_{\mathrm{bal}}$, there exists a normalized positive edge weighting such that*

$$\frac{1}{3n} \leq w_i \leq \frac{3}{n} \quad \textit{for all } i, \qquad p^{\mathrm{D}}_{w,c_\epsilon\beta} = p^{\mathrm{F}}_{\mu,\beta}. \tag{75}$$

At $\epsilon = 0$, equal weights already give exact agreement. The sufficient radius ξ_{bal} is independent of n and ϵ.

The existence witness is $w_i \propto r_i^{-1}$, where $r_i = \sigma(c_\epsilon(\mu_{i+1} - \mu_i)) - Q_i > 0$ is the target residual. In the stated neighborhood these residuals lie between one half and three halves of the baseline circulation. Their reciprocals therefore yield the bounds in (75). This uses population quantities. It supplies neither a learned weighting method nor a finite-sample guarantee. In particular, theorem 2 cannot be restated as a positive lower bound on $\mathcal{R}^{\mathsf{env}}_{\mu,1}$. The complete normalization and circulation bounds are in section D.4.

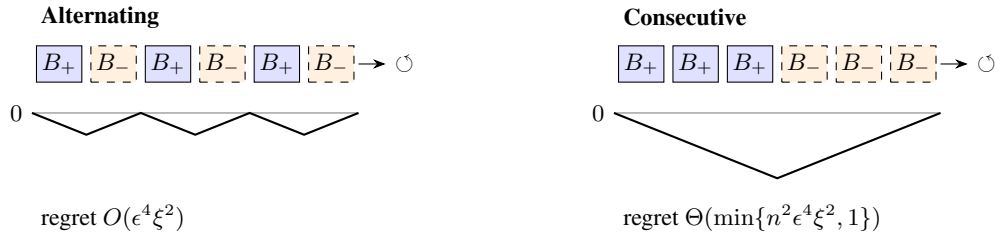

Figure 2: Placements from corollary 2 for positive ϵ, ξ ($L = 3$ shown). Each $B_\pm = (0, a_\pm, 2a_\pm)$ supplies two short edges and a reset. Lines show exact cumulative anchor-score drift in common block units: alternating blocks cancel locally, while consecutive blocks accumulate for L steps. The graph, weights, and class-mean and comparison-probability multisets agree.

A.7 Placement on an identical comparison graph

Corollary 2 (Same-graph placement separation)**.** *For $n = 6L$, $L \geq 2$, let $a_\pm = 1 \pm \xi/2$ and $B_\pm = (0, a_\pm, 2a_\pm)$. Place L blocks of each type on the equal-weight cycle, either alternately or in two consecutive regions. Use the same uniform reference and the same two-point law in both experiments. For sufficiently small positive ϵ, ξ, with constants and neighborhood independent of L,*

$$\mathcal{R}^{\mathsf{temp}}_{\text{alternating},1} \leq C_{\text{alt}}\epsilon^4\xi^2, \tag{76}$$

$$\mathcal{R}^{\mathsf{temp}}_{\text{regions},1} \asymp \min\{n^2\epsilon^4\xi^2, 1\}. \tag{77}$$

The graph, edge weights, multiset of class means, and multiset of oriented edge comparison probabilities agree in the two experiments. Taking $\epsilon = \xi = n^{-1/3}$ gives respective orders $O(n^{-2})$ and $\Theta(1)$ for all sufficiently large $n = 6L$.

This is a consequence of theorem 2. It isolates the placement of local comparison distortions as a quantity that graph spectra and marginal comparison statistics do not determine. The six-class reduction and the matching-multiset argument are proved in section D.5.

A.8 Consequences for sequences of finite experiments

For arbitrary tree sequences, the bound referenced below is

$$\inf_{\tau>0} \text{Reg}_1(p^{\text{D}}_{w,\tau}) \leq \frac{\mathcal{B}(h)^2}{8} D^2_{G_n} V_n^2. \tag{78}$$

Corollary 3 (Sharp criteria for vanishing regret)**.** *For the periodic paths of theorem 1, let $n = 3J \to \infty$ and allow any sequence of centered finite-support laws ν_n on $[-h_n, h_n]$, with $0 < h_n \leq h_*$ and $V_n = \text{Var}_{\nu_n}(U) > 0$. Then* (46) *holds. In particular, $\nu_n = \frac{1}{2}\delta_{-n^{-1/2}} + \frac{1}{2}\delta_{n^{-1/2}}$ gives $\mathcal{R}^{\mathsf{env}}_{1,n} = \Theta(1)$ for all sufficiently large n.*

For cycles $n = 6L \to \infty$, fix $\kappa \geq 3$ and take $0 \leq \epsilon_n \leq \epsilon_\kappa$, $0 \leq \xi_n \leq \xi_{\text{fixed}}$. Let $\Psi_n = \min\{n^2\epsilon_n^4\xi_n^2, 1\}$. Then

$$c_\kappa \Psi_n \leq \mathcal{R}^{\text{fixed}}_\kappa(n, \epsilon_n, \xi_n) \leq \mathcal{E}^{\mathsf{unif}}_n \leq C_\kappa \Psi_n, \qquad \mathcal{R}^{\text{adapt}}_\kappa(n, \epsilon_n, \xi_n) = 0. \tag{79}$$

Consequently the first two regret quantities converge to zero exactly when $n\epsilon_n^2\xi_n \to 0$. If any sequence $w_n \in \mathcal{W}_\kappa(n)$ has vanishing worst-case best-temperature regret, the same scalar condition is necessary. It is not sufficient for arbitrary w_n: at fixed $0 < \epsilon \leq \epsilon_\kappa$ and $\xi_n = n^{-2}$, the weights (71) *give an $\Omega(1)$ worst-case floor, while $\mathcal{E}^{\mathsf{unif}}_n = O(n^{-2})$. For arbitrary sequences of trees with common laws supported in a fixed $[-h, h]$,* (78) *holds and $D_{G_n} V_n \to 0$ is sufficient. For the periodic means and uniform references, equiprobable shocks $U = \pm n^{-1/4}$ give $\Theta(1)$ best attainable regret on paths and an $O(n^{-1})$ upper bound on stars for all sufficiently large $n = 3J$.*

These sequence criteria are consequences of the proved two-sided rates; they are not additional independent lower bounds.

B LEMMAS

The results in this appendix are supporting facts. Their classical mechanisms are not independent novelty claims.

Lemma B.1 (Population DPO and the centered BT optimizer)**.** *Under section 3.1, the DPO policy optimization is equivalent to* (11)*. The centered BT risk has a unique finite minimum. If fine representative identities are retained, its fine counterpart at $\tau = \beta$ is π^{F}_{β}.*

Proof. For a positive policy p, set $v_z = \tau \log(p_z/\rho_z)$. The DPO margin on edge $e = (i, j)$ is $(\mathsf{B}v)_e$, so its conditional expected negative log likelihood is

$$\begin{aligned} &- Q_e \log \sigma((\mathsf{B}v)_e) - (1 - Q_e) \log \sigma(-(\mathsf{B}v)_e) \\ &\quad = Q_e\{\log(1 + e^{(\mathsf{B}v)_e}) - (\mathsf{B}v)_e\} + (1 - Q_e) \log(1 + e^{(\mathsf{B}v)_e}) \\ &\quad = \log(1 + e^{(\mathsf{B}v)_e}) - Q_e(\mathsf{B}v)_e. \end{aligned}$$

Conversely, $p = \mathsf{S}_\rho(v/\tau)$ has $\tau \log(p_z/\rho_z) = v_z - \tau \log \sum_u \rho_u e^{v_u/\tau}$, which preserves all margins. Centering gives a bijection between positive policies and centered score vectors. For $q \in (0, 1)$,

$$\log(1 + e^s) - qs \geq \min\{q, 1 - q\}|s|.$$

Since all $w_e > 0$, the risk is bounded below by a positive constant times $\sum_e |(\mathsf{B}v)_e|$. On a connected graph the latter is a norm on the centered score space: its kernel consists only of constant vectors. Norm equivalence in this fixed finite-dimensional space gives coercivity. Continuity yields existence, and the Hessian $\mathsf{B}^\top \operatorname{diag}(w_e\sigma'((\mathsf{B}v)_e))\mathsf{B}$ is positive definite on centered scores, giving uniqueness. None of the uniform rate bounds below uses a constant from this norm-equivalence argument.

Every fine pair whose classes share an edge has positive mass, because the class weights and all reference conditionals are positive. Between different classes, a coarse path lifts to a fine path by selecting a representative in each intermediate class. Two responses in the same class connect through a response in a neighboring class; such a neighbor exists because $n \geq 2$ and the coarse graph is connected. Thus the induced fine graph is connected. The fine score r fits each fine BT probability exactly. Strict convexity after centering identifies it as the optimizer, and its reference-relative policy map at $\tau = \beta$ gives (5). □

Lemma B.2 (Reduction to a heterogeneous BT population)**.** *The retained probabilities* (4) *equal those of an anonymous heterogeneous BT population. In the common-law experiment, its mean-reward KL-optimal policy also equals p^{F}_{β}.*

Proof. Set $\Omega = \prod_z C^{-1}(z)$, let $\Lambda(\omega) = \prod_z \pi_{\mathrm{ref}}(\omega_z \mid z)$, and define $R_\omega(z) = r(\omega_z)$. For any oriented edge (i, j),

$$\begin{aligned} \mathbb{E}_\Lambda \sigma(R_\omega(j) - R_\omega(i)) = &\sum_{y_i \in C^{-1}(i)} \sum_{y_j \in C^{-1}(j)} \pi_{\mathrm{ref}}(y_i \mid i)\pi_{\mathrm{ref}}(y_j \mid j)\sigma(r(y_j) - r(y_i)) \\ &\cdot \prod_{z \notin \{i,j\}} \sum_{y \in C^{-1}(z)} \pi_{\mathrm{ref}}(y \mid z) = Q_{ij}. \end{aligned}$$

Every factor in the product is one. Under (22), $R_\omega(z) = \mu_z + U_z$ with centered coordinate shocks, so $\mathbb{E}_\Lambda R_\omega(z) = \mu_z$. Also $\mathbb{E}e^{(\mu_z+U)/\beta} = e^{\mu_z/\beta}\mathbb{E}e^{U/\beta}$; the common factor cancels on normalization. Hence both target policies are $\mathsf{S}_\rho(\mu/\beta)$. This is an exact reduction, not a new mixture model. □

Lemma B.3 (Class regret, optimal lifts, and oscillation)**.** *Equation* (8) *holds for every $p \in \Delta_{\mathcal{Z}}$. For any finite vectors $s, f \in \mathbb{R}^n$ and positive reference ρ,*

$$\beta \operatorname{KL}\big(\mathsf{S}_\rho(s/\beta) \| \mathsf{S}_\rho(f/\beta)\big) \leq \frac{\operatorname{osc}(s - f)^2}{8\beta}. \tag{80}$$

Proof. Write $Z_\beta = \sum_y \pi_{\mathrm{ref}}(y) e^{r(y)/\beta}$. Substituting the Gibbs formula, with $0 \log 0 = 0$, gives

$$\begin{aligned} \beta \operatorname{KL}(\pi \| \pi^{\mathrm{F}}_{\beta}) &= \beta \sum_y \pi(y) \left\{ \log \frac{\pi(y)}{\pi_{\mathrm{ref}}(y)} - \frac{r(y)}{\beta} + \log Z_\beta \right\} \\ &= \beta \operatorname{KL}(\pi \| \pi_{\mathrm{ref}}) - \mathbb{E}_\pi r + \beta \log Z_\beta = \beta \log Z_\beta - J_\beta(\pi). \end{aligned}$$

At $\pi = \pi_\beta^{\mathrm{F}}$ the divergence is zero, so $J_\beta(\pi_\beta^{\mathrm{F}}) = \beta \log Z_\beta$ and the difference of objectives is the displayed KL. For every z with $p_z > 0$, $\pi(y) = p_z \pi(y \mid z)$ and $\pi_\beta^{\mathrm{F}}(y) = p_\beta^{\mathrm{F}}(z)\pi_\beta^{\mathrm{F}}(y \mid z)$. Grouping the summands by class therefore gives

$$\begin{aligned}\mathrm{KL}(\pi \| \pi_\beta^{\mathrm{F}}) &= \sum_{z:p_z>0} \sum_{C(y)=z} p_z \pi(y \mid z) \left\{ \log \frac{p_z}{p_\beta^{\mathrm{F}}(z)} + \log \frac{\pi(y \mid z)}{\pi_\beta^{\mathrm{F}}(y \mid z)} \right\} \\ &= \mathrm{KL}(p \| p_\beta^{\mathrm{F}}) + \sum_{z:p_z>0} p_z \, \mathrm{KL}\big(\pi(\cdot \mid z) \| \pi_\beta^{\mathrm{F}}(\cdot \mid z)\big).\end{aligned}$$

The last sum is nonnegative and has minimum zero at the fine-optimal conditionals. This proves (8), including boundary class masses. Summing $\pi_{\mathrm{ref}}(y) e^{r(y)/\beta}$ within each class also gives $p_\beta^{\mathrm{F}}(z) \propto \rho_z \mathbb{E}_{\pi_{\mathrm{ref}}(\cdot|z)} e^{r(Y)/\beta}$, which proves (6).

For the reference-conditional renderer, $\pi_p(y) = p_z \pi_{\mathrm{ref}}(y \mid z)$ and $\pi_{\mathrm{ref}}(y) = \rho_z \pi_{\mathrm{ref}}(y \mid z)$ within class z. Consequently,

$$\begin{aligned}\mathbb{E}_{\pi_p} r &= \sum_z p_z \sum_{C(y)=z} \pi_{\mathrm{ref}}(y \mid z) r(y) = \sum_z p_z m_z, \\ \mathrm{KL}(\pi_p \| \pi_{\mathrm{ref}}) &= \sum_{z:p_z>0} \sum_{C(y)=z} p_z \pi_{\mathrm{ref}}(y \mid z) \log \frac{p_z \pi_{\mathrm{ref}}(y \mid z)}{\rho_z \pi_{\mathrm{ref}}(y \mid z)} = \mathrm{KL}(p \| \rho), \\ J_\beta(\pi_p) &= \beta \log \sum_z \rho_z e^{m_z/\beta} - \beta \, \mathrm{KL}\big(p \| \mathsf{S}_\rho(m/\beta)\big).\end{aligned}$$

This proves (10) and uniqueness of its maximizing class distribution. The equality $\mathsf{S}_\rho(m/\beta) = \mathsf{S}_\rho(f_\beta/\beta)$ holds exactly when $f_\beta - m$ is constant across classes. Under the common translated law, $m_z = \mu_z$ and $f_\beta(z) - m_z = \beta \log \mathbb{E} e^{U/\beta}$ is indeed constant. For the three-class diagnostic it is $(0, H_\beta, 0)$, so the two router targets differ when $\epsilon > 0$.

For (80), set $d = (f - s)/\beta$ and

$$A(t) = \log \sum_z \rho_z e^{s_z/\beta + t d_z}, \qquad p_t = \mathsf{S}_\rho((s + t(f - s))/\beta).$$

Differentiating this finite sum yields

$$A'(t) = \sum_z p_t(z) d_z, \qquad A''(t) = \sum_z p_t(z) d_z^2 - \left(\sum_z p_t(z) d_z \right)^2 = \mathrm{Var}_{p_t}(d).$$

Since $\log(p_0(z)/p_1(z)) = -d_z + A(1) - A(0)$, we obtain

$$\begin{aligned}\mathrm{KL}(p_0 \| p_1) &= A(1) - A(0) - A'(0) \\ &= \int_0^1 \{A'(t) - A'(0)\} \, dt = \int_0^1 (1 - t) A''(t) \, dt.\end{aligned}$$

For $m = (\min_z d_z + \max_z d_z)/2$, $\mathrm{Var}_{p_t}(d) = \inf_c \mathbb{E}_{p_t}(d - c)^2 \le \mathbb{E}_{p_t}(d - m)^2 \le \mathrm{osc}(d)^2/4$. Thus multiplying the integral by β gives $\beta \, \mathrm{KL}(p_0 \| p_1) \le \beta \, \mathrm{osc}(d)^2/8 = \mathrm{osc}(f - s)^2/(8\beta)$, as required. □

Lemma B.4 (Positive residual flow and the oracle loss envelope)**.** *Fix G, Q, ρ, a positive policy p, and $\tau > 0$, and put $v = \tau \log(p/\rho)$ and $r_e = \sigma((\mathsf{B}v)_e) - Q_e$. Discard the edges with $r_e = 0$. Orient the remaining edges forward if $r_e > 0$ and backward otherwise. There exists a strictly positive weighting for which p is the DPO minimizer if and only if every directed residual edge belongs to a directed cycle. Moreover, at each fixed τ,*

$$\bigcup_{\ell \in \mathfrak{L},\ w \in \Delta_E^\circ} \mathcal{P}_{\ell,w,\tau} = \bigcup_{w \in \Delta_E^\circ} \{p_{w,\tau}^{\mathrm{D}}\}. \tag{81}$$

If Q is BT-realizable on the connected graph, every allowed loss and positive weighting have the same unique centered minimizing score.

Proof. The DPO first-order equation is $\mathsf{B}^\top \operatorname{diag}(w) r = 0$. Convexity makes it sufficient as well as necessary. It specifies a circulation positive on every directed residual edge. To see that each such edge $u \to v$ lies on a directed cycle, let S be the set reachable from v through positive-flow edges. If $u \notin S$, no positive flow leaves S, while $u \to v$ brings positive flow into it, contradicting conservation summed over S. Thus a path from v back to u exists. Conversely, choose a directed cycle through each residual edge and sum their unit circulations. The result is positive on every residual edge. Dividing its edge magnitudes by $|r_e|$ gives the required positive weights. Zero-residual edges receive arbitrary positive weights before normalization. The all-zero residual case is included.

A finite global minimum of another loss satisfies $\mathsf{B}^\top \operatorname{diag}(w_e a_\ell((\mathsf{B}v)_e)) r = 0$. The positive factors define DPO weights after normalization. This proves one inclusion in (81); the other follows because log loss belongs to $\mathfrak{L}$. No sufficiency of stationarity for nonconvex proper losses is used. Finally, each conditional risk has its unique minimum at $s = \operatorname{logit} q$, by (50). When all these minima are jointly realizable they give the unique centered global score on a connected graph. □

The derivative representation is classical (Reid & Williamson, 2010), as are cycle-flow arguments. Related BT projection equations and zero-flow perturbations already appear in (Pukdee et al., 2026, App. D.4). Equation (81) is an equality after reweighting, not an equality across losses at fixed weights on a misspecified graph.

Lemma B.5 (Variance-retaining mixed-logit bounds)**.** *Under* (22) *and the constants in* (53)*, for every real* t*,*

$$|g_\nu(t) - t| \le \mathcal{B}(h) V. \tag{82}$$

For $h \le 1/4$ *and* $|t| \le 2$*,*

$$|g_\nu(t) - t + V \tanh(t/2)| \le C_{\mathrm{mix}} h^2 V. \tag{83}$$

Both hold when $V = 0$*, with zero right-hand sides.*

Proof. We have $\mathbb{E}D^2 = 2V$, $V \le h^2$, and $\mathbb{E}D^4 \le (2h)^2 \mathbb{E}D^2 = 8h^2 V$. For $X = e^D$ and $w(X) = (1 + e^t X)^{-1}$, an independent copy X' gives

$$2 \operatorname{Cov}(X, w(X)) = \mathbb{E}[(X - X')(w(X) - w(X'))] \le 0.$$

Since $F_\nu(t) = e^t \mathbb{E}[X w(X)]$ and $1 - F_\nu(t) = \mathbb{E}[w(X)]$,

$$e^{g_\nu(t) - t} = \frac{\mathbb{E}[X w(X)]}{\mathbb{E}[w(X)]} \le \mathbb{E}X.$$

Thus $g_\nu(t) - t \le \log \mathbb{E} e^D$. Symmetry of D and the same bound at $-t$ give $|g_\nu(t) - t| \le \log \mathbb{E} \cosh D$. For $|x| \le 2h$, $\cosh x - 1 \le x^2 \{\cosh(2h) - 1\}/(4h^2)$. Using $\log(1 + x) \le x$ proves (82).

For the derivative bounds used here and below, put $p = \sigma(t)$ and $z = 1 - 2p \in [-1, 1]$. Direct differentiation gives

$$\sigma'' = p(1-p)z, \quad \sigma''' = \tfrac{1}{2} p(1-p)(3z^2 - 1), \quad \sigma^{(4)} = p(1-p)z(3z^2 - 2).$$

Consequently $\sup |\sigma''| \le 1/4$, $\sup |\sigma'''| \le 1/4$, and $\sup |\sigma^{(4)}| \le 1/2$. Symmetry cancels odd Taylor terms, giving

$$F_\nu(t) = \sigma(t) + V \sigma''(t) + r(t), \qquad |r(t)| \le h^2 V/6, \qquad |F_\nu(t) - \sigma(t)| \le V/4.$$

For the stated t, h, both probabilities lie between $\sigma(-3)$ and $\sigma(3)$. On this interval $|\operatorname{logit}''| \le \kappa_3^{-2}$, and $\sigma'(t) \ge \kappa_3$. Expanding the logit yields

$$\left| g_\nu(t) - t - V \frac{\sigma''(t)}{\sigma'(t)} \right| \le \frac{h^2 V}{6\kappa_3} + \frac{V^2}{32 \kappa_3^2}.$$

Now use $\sigma''(t)/\sigma'(t) = -\tanh(t/2)$ and $V^2 \le h^2 V$. □

The variance correction to a mixed logistic probability is already explicit in heterogeneous-preference DPO analysis (Shirali et al., 2025, Eq. (14)). Here the factor V in the uniform remainder is needed to permit arbitrarily small atom masses in theorem 1; the Taylor principle itself is not new.

Lemma B.6 (Exact scalar flow on a cycle)**.** *For any $Q \in (0,1)^n$, equal-weight logistic DPO on the oriented cycle has the unique centered score specified by*

$$v_{i+1} - v_i = \operatorname{logit}(Q_i + j), \qquad \sum_i \operatorname{logit}(Q_i + j) = 0, \quad -\min_i Q_i < j < 1 - \max_i Q_i. \tag{84}$$

The scalar root exists and is unique.

Proof. The sum of logits is strictly increasing in j, tends to $-\infty$ at the left endpoint and to $+\infty$ at the right endpoint. Its root gives increments summing to zero, hence a score on the cycle. Every residual equals j, so incoming and outgoing weighted residuals cancel at every vertex. Strict convexity on centered scores proves the claim. □

Lemma B.7 (Common reference change)**.** *For positive ρ, $\chi = \max_i \rho_i / \min_i \rho_i$, uniform u, and any finite score vectors a, b,*

$$\chi^{-1} \operatorname{KL}(\mathsf{S}_u(a) \| \mathsf{S}_u(b)) \le \operatorname{KL}(\mathsf{S}_\rho(a) \| \mathsf{S}_\rho(b)) \le \chi \operatorname{KL}(\mathsf{S}_u(a) \| \mathsf{S}_u(b)). \tag{85}$$

Proof. At each score vector s, the ratio $\mathsf{S}_\rho(s)_i / \mathsf{S}_u(s)_i = \rho_i / \mathbb{E}_{\mathsf{S}_u(s)} \rho$ belongs to $[\chi^{-1}, \chi]$. The formula $\operatorname{Var}_p(f) = \inf_c \mathbb{E}_p(f-c)^2$ gives the same two-sided comparison of variances. Express each KL as the integral of variances along $s = a + t(b-a)$ with weight $1-t$, and apply the comparison inside the integral. □

Lemma B.8 (Geometric and three-type divergences)**.** *For $m \ge 2$ and $a \ge 0$, put*

$$P_{a,m}(k) = \frac{e^{-ak}}{\sum_{j=0}^{m-1} e^{-aj}}, \quad G_m(a) = \operatorname{KL}(P_{a,m} \| \operatorname{Unif}_m), \quad q_s(t) = \frac{e^{st}}{Z(s)}, \quad Z(s) = 1 + e^s + e^{2s}.$$

Then

$$\frac{e^{-1}}{32} \min\{m^2 a^2, 1\} \le G_m(a) \le \frac{(m-1)^2 a^2}{8}. \tag{86}$$

For $0 < s \le 1/2$, $\operatorname{KL}(q_s \| q_1) \ge e^{-2}/12$. For $1/2 \le a \le 2$ and $0 \le s \le a/2$, $\operatorname{KL}(q_s \| q_a) \ge e^{-4}/48$.

Proof. Let $A_m(a) = \log \sum_{k=0}^{m-1} e^{-ak}$. Direct substitution gives

$$G_m(a) = a A_m'(a) - A_m(a) + \log m, \qquad G_m'(a) = a \operatorname{Var}_{P_{a,m}}(k) \ge 0.$$

When $0 \le a \le 1/m$, the density ratio to the uniform law is at least e^{-1}. The variance-as-an-infimum formula used in lemma B.7 therefore gives

$$\operatorname{Var}_{P_{a,m}}(k) \ge e^{-1} \frac{m^2 - 1}{12} \ge \frac{e^{-1} m^2}{16}.$$

Integrate G_m' from zero to $\min\{a, 1/m\}$ and use monotonicity. This proves the lower bound, including $a > 1/m$. The upper bound is (80) with coefficient one and score range $(m-1)a$.

Let $\psi(s) = \log Z(s)$, so $\psi''(s) = \operatorname{Var}_{q_s}(t)$. For $0 \le u \le A$, the density ratio $q_u / \operatorname{Unif}_3$ is at least e^{-2A} and hence $\psi''(u) \ge (2/3) e^{-2A}$. For $0 \le s \le a$ the KL orientation gives

$$\operatorname{KL}(q_s \| q_a) = \psi(a) - \psi(s) - \psi'(s)(a-s) = \int_s^a (a-u) \psi''(u)\, du. \tag{87}$$

As a function of $s < a$, its derivative is $(s-a)\psi''(s) < 0$. Thus $0 < s \le 1/2$ gives a lower bound by its value at $s = 1/2$, $a = 1$, which is at least $(e^{-2}/3)(1 - 1/2)^2 = e^{-2}/12$. For $a \in [1/2, 2]$ and $s \le a/2$, (87) is at least $(e^{-4}/3)(a-s)^2 \ge e^{-4}/48$. □

These bounds are elementary exponential-family calculus. They are included to make the all-temperature constants explicit, not as contribution claims.

B.1 Structural clarifications: target shifts, link correction, and sparse graphs

The following supporting facts delimit the main theorems. The identities, correctly specified recovery argument, and continuity argument are not separate claims of a new learning algorithm or sharp general-graph theory.

Lemma B.9 (Comparison distortion and aggregation-target shift)**.** *In the general finite experiment, set $m_z = \mathbb{E}r(Y_z)$, $U_z = r(Y_z) - m_z$, $V_z = \mathbb{E}U_z^2$, and*

$$h_z = \beta \log \mathbb{E}e^{U_z/\beta}, \qquad F_{ij}(t) = \mathbb{E}\sigma(t + U_j - U_i),$$

where representatives from different classes are independent. Then $f_\beta = m + h$ and (15) *holds. In particular, for $e = (i, j)$,*

$$|r_e(1)| \le \frac{|h_j - h_i|}{4} + \frac{V_i + V_j}{8}. \tag{88}$$

Without a common law or additional tail control, vanishing $\max_z V_z$ does not force h to become constant, even when all class means agree. There are two-class finite experiments with $Q = 1/2$ and $\max_z V_z \to 0$ whose best-temperature regret diverges.

Proof. Since $r(Y_z) = m_z + U_z$,

$$f_\beta(z) = \beta \log \mathbb{E}e^{(m_z+U_z)/\beta} = m_z + \beta \log \mathbb{E}e^{U_z/\beta} = m_z + h_z,$$
$$Q_{(i,j)} = \mathbb{E}\sigma(m_j - m_i + U_j - U_i) = F_{ij}(m_j - m_i).$$

Writing $\Delta m = m_j - m_i$ and $\Delta h = h_j - h_i$, add and subtract $\sigma(\Delta m)$ in $r_e(1) = \sigma(\Delta m + \Delta h) - F_{ij}(\Delta m)$. This gives (15) exactly, with no sign assumption on either term. The first term satisfies

$$|\sigma(\Delta m + \Delta h) - \sigma(\Delta m)| = \left| \int_0^{\Delta h} \sigma'(\Delta m + t)\, dt \right| \le |\Delta h|/4.$$

For the second term, let $Z = U_j - U_i$. Independence and centering imply

$$\mathbb{E}Z = 0, \qquad \mathbb{E}Z^2 = \mathbb{E}U_j^2 + \mathbb{E}U_i^2 - 2\mathbb{E}U_j\mathbb{E}U_i = V_j + V_i.$$

Taylor's formula with integral remainder gives, for every real t, Z,

$$\sigma(t + Z) - \sigma(t) = Z\sigma'(t) + Z^2 \int_0^1 (1 - u)\sigma''(t + uZ)\, du.$$

Taking expectations, using $\sup |\sigma''| \le 1/4$, and computing $\int_0^1 (1 - u)\, du = 1/2$, we obtain

$$|F_{ij}(t) - \sigma(t)| \le \mathbb{E}Z^2 \int_0^1 (1 - u)\frac{1}{4}\, du = \frac{V_i + V_j}{8}.$$

The triangle inequality now proves (88). Summing the signed edge residuals with weights yields the divergence b_1 in proposition 1; replacing them by absolute bounds can lose cancellation. For a common centered law, every h_z is identical and the first term vanishes exactly.

For the final assertion fix $\beta > 0$, take two classes with $\rho = (1/2, 1/2)$ and $m_1 = m_2 = 0$, and let $U_1 = 0$. For $A \ge 1$, put $V_A = e^{-A/(2\beta)}$ and take the centered three-point law

$$\Pr(U_2 = A) = \Pr(U_2 = -A) = \frac{V_A}{2A^2}, \qquad \Pr(U_2 = 0) = 1 - \frac{V_A}{A^2}.$$

All three probabilities are strictly positive and sum to one. This is a finite experiment with four fine responses, and $\mathrm{Var}(U_2) = A^2\{V_A/(2A^2) + V_A/(2A^2)\} = V_A \to 0$. Symmetry and $\sigma(u) + \sigma(-u) = 1$ imply

$$Q_{(1,2)} = \left(1 - \frac{V_A}{A^2}\right)\frac{1}{2} + \frac{V_A}{2A^2}\{\sigma(A) + \sigma(-A)\} = \frac{1}{2}.$$

Hence the fitted BT score is constant and every temperature gives $p^{\mathrm{D}} = (1/2, 1/2)$. However, $h_1 = 0$ and

$$H_A := h_2 = \beta \log \left[1 + \frac{V_A}{A^2} \{\cosh(A/\beta) - 1\} \right],$$

$$H_A \geq \beta \log \left(\frac{V_A}{2A^2} e^{A/\beta} \right) = \frac{A}{2} - 2\beta \log A - \beta \log 2 \longrightarrow \infty.$$

The inequality retains just the positive atom's contribution to the exponential moment. For $p_\beta^{\mathrm{F}} = (1, e^{H_A/\beta})/(1 + e^{H_A/\beta})$, direct substitution gives

$$\begin{aligned} \mathrm{Reg}_\beta(p^{\mathrm{D}}) &= \frac{\beta}{2} \log \frac{1 + e^{H_A/\beta}}{2} + \frac{\beta}{2} \log \frac{1 + e^{H_A/\beta}}{2e^{H_A/\beta}} \\ &= \beta \log\{1 + e^{H_A/\beta}\} - \beta \log 2 - H_A/2 \\ &= \beta \log \cosh(H_A/(2\beta)) \longrightarrow \infty. \end{aligned}$$

This example does not contradict the dense-coverage bound: the target range grows, and the small-residual condition fails. It shows why the variance-only common-law conclusion cannot be read as a conclusion for arbitrary class-dependent laws. □

Lemma B.10 (Oracle recovery with a known comparison link)**.** *For a known centered finite common law ν, $F_\nu : \mathbb{R} \to (0, 1)$ is a strictly increasing bijection. In the common-law experiment,* (49) *recovers every observed mean difference. On any connected graph and for any positive pair weights, logistic DPO using Q^{cal} preserves p_β^{F} at $\tau = \beta$. The same recovery follows by fitting the correctly specified link F_ν directly to the original Q and applying $\mathsf{S}_\rho(v/\beta)$. For known class-dependent centered laws, inversion instead recovers $\mathsf{B}m$; adding $\mathsf{B}h$ before the policy map recovers the fine target.*

Proof. For each real t, finite support permits differentiation term by term:

$$F_\nu'(t) = \mathbb{E}\sigma'(t + U - U') > 0.$$

Moreover $\lim_{t \to -\infty} F_\nu(t) = 0$ and $\lim_{t \to \infty} F_\nu(t) = 1$, by the corresponding sigmoid limits in the finite sum. Continuity and strict monotonicity prove bijectivity. Since $Q_e = F_\nu((\mathsf{B}\mu)_e)$, inversion yields $d_e = (\mathsf{B}\mu)_e$. Integrating d along a spanning tree recovers μ up to one additive constant; every remaining edge is consistent because d is the gradient of the true μ.

For the corrected probabilities, $\mathrm{logit}\, Q_e^{\mathrm{cal}} = d_e = (\mathsf{B}\mu)_e$. The centered version of μ therefore minimizes every logistic conditional risk separately. All positive-weight global minimizers must attain every conditional minimum, so their differences equal $\mathsf{B}\mu$. Connectedness leaves only an additive constant. At temperature β their policy is $\mathsf{S}_\rho(\mu/\beta) = p_\beta^{\mathrm{F}}$.

Alternatively, define the conditional risk for the known link by

$$R_\nu(s; q) = -q \log F_\nu(s) - (1 - q) \log\{1 - F_\nu(s)\}.$$

For $q = F_\nu(d)$, subtraction at its realizable minimum gives

$$\begin{aligned} R_\nu(s; q) - R_\nu(d; q) &= q \log \frac{q}{F_\nu(s)} + (1 - q) \log \frac{1 - q}{1 - F_\nu(s)} \\ &= \mathrm{KL}\big(\mathrm{Ber}(q) \| \, \mathrm{Ber}(F_\nu(s))\big) \geq 0. \end{aligned}$$

Equality holds exactly when $F_\nu(s) = q$, hence exactly when $s = d$. The vector $v = \mu$ attains equality on every edge, so the positive weighted sum again has that unique centered global minimizer. This argument uses realizability, not a claim of global convexity for the general link. Except when the edge optima happen to coincide, its conditional optimum $F_\nu^{-1}(q)$ differs from $\mathrm{logit}\, q$, which is required by definition A.1. Thus this correction is outside the envelope used in the lower bounds.

For known centered laws ν_i, ν_j, the same derivative and limit arguments apply to $F_{ij}(t) = \mathbb{E}\sigma(t + U_j - U_i)$, even when this function is not symmetric about $1/2$. Hence

$$d_{ij} = F_{ij}^{-1}(Q_{(i,j)}) = m_j - m_i, \qquad \widetilde{d}_{ij} = d_{ij} + h_j - h_i = f_\beta(j) - f_\beta(i).$$

The shifts $h_z = \beta \log \mathbb{E}_{\nu_z} e^{U_z/\beta}$ are computable from the supplied laws. Logistic fitting to $\widetilde{Q}_{ij} = \sigma(\widetilde{d}_{ij})$ consequently has score f_β up to a constant and recovers p_β^{F} at $\tau = \beta$. Using F_{ij} without this last shift targets the mean policy instead. All assertions concern exact population probabilities and supplied laws; no claim is made that those laws are identifiable from anonymous Q. □

Lemma B.11 (Accumulated error on an arbitrary tree)**.** *Let the comparison graph be a tree, with arbitrary finite target f_β, probabilities $Q \in (0,1)^E$, and positive weights. Every allowed proper loss has the same centered fitted score v, characterized by $\mathsf{B}v = g$, where $g_e = \operatorname{logit} Q_e$. For $c > 0$ and $\delta_e(c) = g_e/c - (\mathsf{B}f_\beta)_e$,*

$$\operatorname{osc}(v/c - f_\beta) = \max_{i,j} \left| \sum_{e \in P_{ij}} s_{ij,e} \delta_e(c) \right|. \tag{89}$$

Here P_{ij} is the unique path from i to j, and $s_{ij,e}$ is 1 if its traversal agrees with the fixed edge orientation and -1 otherwise. This proves (48)*. Exact preservation at a finite temperature holds precisely when $g = c\mathsf{B}f_\beta$ for some $c > 0$.*

Proof. The conditional derivative in definition A.1 has the sign of $\sigma(s) - q$. It is strictly negative for $s < \operatorname{logit} q$ and strictly positive for $s > \operatorname{logit} q$. The unique conditional minimum is therefore $\operatorname{logit} q$. Choose a root o of the tree, set $v_o = 0$, and define

$$v_j = \sum_{e \in P_{oj}} s_{oj,e} g_e.$$

Because there is a unique path, this assignment is unambiguous and $v_j - v_i = g_{(i,j)}$ on each oriented edge. Thus all conditional minima are simultaneously attained. Positive weights force any global minimizer to attain every one of these minima; connectedness gives uniqueness up to constants. Centering removes the remaining constant.

Put $x = v/c - f_\beta$. On an oriented edge $e = (u, z)$,

$$x_z - x_u = (v_z - v_u)/c - \{f_\beta(z) - f_\beta(u)\} = \delta_e(c).$$

Summing along P_{ij} telescopes, with the sign reversing on each backward traversal:

$$x_j - x_i = \sum_{e \in P_{ij}} s_{ij,e} \delta_e(c).$$

The maximum absolute difference of a finite vector is its oscillation, proving (89). At temperature $\tau = c\beta$ the policy is $\mathsf{S}_\rho((v/c)/\beta)$. By lemma B.3,

$$\operatorname{Reg}_\beta(p^{\mathrm{D}}_{w,c\beta}) \le \frac{\operatorname{osc}(v/c - f_\beta)^2}{8\beta}.$$

Taking an infimum over $c > 0$ proves (48), since $c \mapsto c\beta$ covers every positive temperature.

Finally, two reference-relative softmax policies are equal precisely when their arguments differ by a constant. Thus exact preservation at $c\beta$ is equivalent to $v/c - f_\beta$ being constant, equivalent to $g/c - \mathsf{B}f_\beta = 0$. No periodicity, common representative law, or equal-target anchor is used. A diameter bound follows by replacing each path sum by $D_G \max_e |\delta_e(c)|$, thereby losing its cancellations. □

Lemma B.12 (Positive gaps persist at fixed graph size)**.** *Fix a finite connected graph, positive pair weights, a positive reference ρ, and $\beta > 0$. Write $v(Q)$ for its centered logistic optimizer and set*

$$R(Q, f) = \inf_{\tau > 0} \beta \operatorname{KL}\big(\mathsf{S}_\rho(v(Q)/\tau) \| \mathsf{S}_\rho(f/\beta)\big).$$

If $v(Q_0)$ is nonconstant and $R(Q_0, f_0) > 0$, then some $\delta, \eta > 0$ satisfy

$$\|Q - Q_0\|_\infty + \|f - f_0\|_\infty < \delta \quad \Longrightarrow \quad R(Q, f) \ge \eta.$$

The constants may depend on the fixed graph and weighting. No uniformity in graph size or over weights approaching zero is asserted.

Proof. On the centered score subspace the gradient equation is smooth, and its derivative with respect to the score is

$$\mathsf{B}^\top \operatorname{diag}\{w_e \sigma'((\mathsf{B}v)_e)\}\mathsf{B}.$$

For a nonzero centered vector a, its quadratic form equals $\sum_e w_e \sigma'((\mathsf{B}v)_e)((\mathsf{B}a)_e)^2 > 0$ by connectedness. It is therefore invertible on that subspace. The implicit function theorem, followed by uniqueness from lemma B.1, proves local continuity of $v(Q)$ at Q_0.

Suppose the conclusion fails. There are $Q_k \to Q_0$, $f_k \to f_0$ and $\tau_k > 0$ with

$$\beta\,\mathrm{KL}(p_k\|q_k) \longrightarrow 0, \qquad p_k = \mathsf{S}_\rho(v(Q_k)/\tau_k), \quad q_k = \mathsf{S}_\rho(f_k/\beta).$$

To obtain such temperatures, approximate each infimum within $1/k$. Since $q_k \to q_0 := \mathsf{S}_\rho(f_0/\beta)$, which is strictly positive, compactness of the probability simplex and continuity of KL imply $p_k \to q_0$. Indeed, every convergent subsequence has limit p with $\mathrm{KL}(p\|q_0) = 0$, hence $p = q_0$. After passing to a subsequence, τ_k has a limit in $[0, \infty]$.

If $\tau_k \to \tau_0 \in (0, \infty)$, continuity gives $q_0 = \mathsf{S}_\rho(v(Q_0)/\tau_0)$, contradicting $R(Q_0, f_0) > 0$. If $\tau_k \to \infty$, boundedness of the convergent scores gives $p_k \to \rho$. Then $q_0 = \rho$, and taking temperature to infinity at (Q_0, f_0) again gives $R(Q_0, f_0) = 0$, a contradiction.

It remains to rule out $\tau_k \to 0$. Put $A = \arg\max_i v_i(Q_0)$. Nonconstancy makes A a proper subset. For every $i \notin A$, fix $j \in A$ and let $d_i = v_j(Q_0) - v_i(Q_0) > 0$. Score continuity gives $v_j(Q_k) - v_i(Q_k) \geq d_i/2$ for all sufficiently large k. Hence

$$\frac{p_k(i)}{p_k(j)} = \frac{\rho_i}{\rho_j} \exp\left(\frac{v_i(Q_k) - v_j(Q_k)}{\tau_k}\right) \leq \frac{\rho_i}{\rho_j} e^{-d_i/(2\tau_k)} \longrightarrow 0.$$

Since $p_k(j) \leq 1$, this forces $p_k(i) \to 0$, contradicting $p_k \to q_0$ and $q_0(i) > 0$. All temperature limits are excluded. The contradiction proves the positive neighborhood claim. Small changes to repeated class means give continuous changes of both Q and f_β, so fixed-size examples can lose their exact anchors while retaining a positive gap. This argument supplies no rate for how their admissible perturbations depend on n. □

C Proofs of the main results

C.1 Proof of theorem 1

The target formula (24) follows from $\mathbb{E}e^{(\mu_z+U)/\beta} = e^{\mu_z/\beta}\mathbb{E}e^{U/\beta}$. It does not replace the comparison link F_ν by a BT link.

Uniform drift and the actual fitted score. Because $|D| \leq 2h \leq 1/2$, $\sigma(1-2h) \leq F_\nu(1) \leq \sigma(1+2h)$ and hence $1/2 \leq b \leq 3/2$. Write

$$g_\nu(t) = t - V\tanh(t/2) + r(t), \qquad |r(t)| \leq C_{\mathrm{mix}} h^2 V \quad (t = 1, 2).$$

Then $d_\nu = KV + r(2) - 2r(1)$ and $|r(2) - 2r(1)| \leq 3C_{\mathrm{mix}}h^2V \leq KV/2$. The bounds on d_ν and $\eta = d_\nu/b$ follow. For clarity, $K > 0$: with $z = \tanh(1/2)$, $K = 2z^3/(1+z^2)$. Also $K < 1$ since $2\tanh(1/2) < 1$. All these inequalities retain the variance factor V; no atom-mass lower bound or comparison between V and h^2 from below is used.

Within a block the fitted increments are b, b. Between consecutive blocks the fitted increment is $-g_\nu(2) = -(2b + d_\nu)$. Integrating from the first vertex proves (27): $v_{k,t} = -kd_\nu + tb$, up to a constant. Thus every comparison is fitted exactly, and the whole oracle envelope consists of the policies obtained by varying $\tau > 0$ for this one score.

All-temperature lower bound. Put $s = b/\tau > 0$. By uniformity of the reference, the score in (27) has normalization

$$\sum_{k=0}^{J-1}\sum_{t=0}^{2} e^{(-kd_\nu+tb)/\tau} = \left(\sum_{k=0}^{J-1} e^{-\eta sk}\right)\left(\sum_{t=0}^{2} e^{st}\right).$$

Consequently $p^{\mathrm{D}}_{w,\tau}(k,t) = P_{\eta s,J}(k)q_s(t)$ and $p^{\mathrm{F}}_1(k,t) = q_1(t)/J$. Substituting in KL gives

$$\begin{aligned}\mathrm{Reg}_1(p^{\mathrm{D}}_{w,\tau}) &= \sum_{k,t} P_{\eta s,J}(k)q_s(t)\left\{\log\frac{P_{\eta s,J}(k)}{1/J} + \log\frac{q_s(t)}{q_1(t)}\right\}\\ &= G_J(\eta s) + \mathrm{KL}(q_s\|q_1),\end{aligned}$$

since each factor sums to one. This proves (28) with the notation of lemma B.8. If $s \geq 1/2$, that lemma gives

$$G_J(\eta s) \geq \frac{e^{-1}}{32}\min\{J^2\eta^2 s^2, 1\} \geq \frac{e^{-1}}{128}\min\{J^2\eta^2, 1\}.$$

If $0 < s \le 1/2$, the other term is at least $e^{-2}/12$, which exceeds $e^{-1}/128$. Therefore the same lower bound holds for every $s > 0$. Using $\eta \ge KV/3$ and $K^2/9 < 1$ gives

$$\mathcal{R}_1^{\mathsf{env}} \ge \frac{e^{-1}K^2}{1152} \min\{J^2V^2, 1\}.$$

The two cases cover all positive temperatures, including any sequences tending to zero or infinity.

Upper bound. At $s = 1$ (the finite temperature $\tau = b$), the type divergence is zero, and lemma B.8 gives

$$G_J(\eta) \le \frac{(J-1)^2\eta^2}{8} \le \frac{9K^2J^2V^2}{8}.$$

As $s \downarrow 0$, the block divergence tends to zero and the type divergence tends to $\mathrm{KL}(q_0\|q_1) = \log((1+e+e^2)/3) - 1$. Taking the smaller of these bounds yields (25) with the stated c_U. The nonuniform-reference extension following the theorem uses lemma B.7 at each τ before taking infima; the common representative law and score stay fixed. □

C.2 Proof of theorem 2(i)

Write $c = c_\epsilon$ and $\Delta_i = \mu_{i+1} - \mu_i$. We first establish constants for the baseline calibration, and then bound the actual scalar root for the perturbed cycle.

Uniform calibration and residual derivatives. The logistic derivatives obey $\sup|\sigma''| \le 1/4$ and $\sup|\sigma'''| \le 1/4$. Taylor's formula with $\mathbb{E}D = 0$ and $\mathbb{E}D^2 = 2\epsilon^2$ gives, for every real t,

$$|F_\epsilon(t) - \sigma(t)| \le \epsilon^2/4, \qquad |F_\epsilon'(t) - \sigma'(t)| \le \epsilon^2/4. \tag{90}$$

The calibration function $H(c) = \sigma(c) + \sigma(2c)$ is strictly increasing. Using $\sigma(t - 2\epsilon) \le F_\epsilon(t) \le \sigma(t + 2\epsilon)$ at $t = 1, 2$ gives

$$H(1 - 2\epsilon) \le F_\epsilon(1) + F_\epsilon(2) \le H(1 + 2\epsilon).$$

Thus

$$1 - 2\epsilon \le c \le 1 + 2\epsilon, \qquad 7/8 \le c \le 9/8. \tag{91}$$

On the interval between c and one, $H'(u) = \sigma'(u) + 2\sigma'(2u) \ge 3\kappa_3$. By (90) and the mean value theorem,

$$|c - 1| \le \frac{\epsilon^2}{6\kappa_3} < 4\epsilon^2. \tag{92}$$

The numerical inequality uses $\kappa_3^{-1} = e^3 + 2 + e^{-3} < 24$; for example the exponential series gives $e < 11/4$ and hence $e^3 < 21$.

Define the residual left by this fixed calibration and its baseline value:

$$R_\epsilon(t) = \sigma(ct) - F_\epsilon(t), \qquad j_0 = R_\epsilon(1) = R_\epsilon(-2).$$

The equality follows from calibration and the symmetry of F_ϵ. For $|t| \le 3$, using $\sup|\sigma''| \le 1/4$ gives

$$\begin{aligned} |R_\epsilon'(t)| &\le |c-1|\sigma'(ct) + |\sigma'(ct) - \sigma'(t)| + |\sigma'(t) - F_\epsilon'(t)| \\ &\le \frac{1+|t|}{4}|c-1| + \frac{\epsilon^2}{4} \le \frac{17}{4}\epsilon^2. \end{aligned}$$

Let $\Delta_i^0 = \mu_{i+1}^0 - \mu_i^0 \in \{1, -2\}$. Because $|\Delta_i - \Delta_i^0| \le 2\xi$ and the intervening interval lies in $[-3, 3]$, the errors

$$e_i = R_\epsilon(\Delta_i) - j_0, \qquad |e_i| \le E_0 := \frac{17}{2}\epsilon^2\xi \tag{93}$$

are uniform in n and in the perturbation.

Admissible root bracket. Set $b_i^{\mathrm{prob}} = \sigma(c\Delta_i)$. We have $|\Delta_i| \le 2 + 2\xi \le 17/8$, so $|c\Delta_i| \le 153/64 < 3$. Consequently $b_i^{\mathrm{prob}} \in [\alpha_0, 1-\alpha_0]$ and

$$Q_i = b_i^{\mathrm{prob}} - j_0 - e_i.$$

Consider the two trial shifts $j_0 + \min_i e_i$ and $j_0 + \max_i e_i$. At either trial value, $Q_i + j$ differs from b_i^{prob} by at most $2E_0$. The numerical bounds in the statement imply

$$2E_0 = 17\epsilon^2\xi \le \frac{17}{4096} < \frac{\alpha_0}{2},$$

where $\alpha_0 = 1/(1+e^3) > 1/22$. Both trial values therefore lie in the admissible open interval of (84), and every probability there lies in $[\alpha_0/2, 1-\alpha_0/2]$. This verifies the domain before evaluating the root equation.

At the lower trial shift, $Q_i + j \le b_i^{\mathrm{prob}}$ for every i; at the upper trial shift the reverse holds. Moreover, $\sum_i \mathrm{logit}(b_i^{\mathrm{prob}}) = c\sum_i \Delta_i = 0$. The strictly increasing scalar equation in lemma B.6 thus gives

$$\min_i e_i \le j - j_0 \le \max_i e_i, \qquad |(Q_i + j) - b_i^{\mathrm{prob}}| \le 2E_0 \tag{94}$$

for the actual optimizer. On the displayed probability interval, $\mathrm{logit}'(p) = 1/(p(1-p)) \le 1/\kappa_*$, so

$$|(v_{i+1} - v_i) - c\Delta_i| \le \frac{2E_0}{\kappa_*} = \frac{17\epsilon^2\xi}{\kappa_*}.$$

Between any two vertices, integrate along a cycle path of at most n edges and divide by c. This proves (61). At $\tau = c\beta$ the policy is $\mathsf{S}_{\mathsf{unif}}((v/c)/\beta)$; lemma B.3 proves (62). Since $c = c_\epsilon \ge 1/2$,

$$\frac{17}{c\kappa_*} \le \frac{34}{\kappa_*} = C_{\mathrm{sc}}, \qquad \frac{289}{8c^2\kappa_*^2} \le \frac{289}{2\kappa_*^2} = C_{\mathrm{up}}.$$

This gives (34) and (35).

Finally, as $\tau \uparrow \infty$ the policy tends to the uniform reference. The mean range satisfies $\operatorname{osc}\mu \le 2 + 2\xi$, so lemma B.3 gives the second entry of (36). At $\epsilon = 0$ or $\xi = 0$, the same argument has $E_0 = 0$ and gives exact agreement at the stated finite calibrated temperature. Part (i) is proved. □

C.3 Proof of theorem 2(ii)

Throughout this proof the law is the specified equiprobable two-point law and the objective is equal-weight logistic DPO. In the two-region construction, the two short increments in a block of type a are a, a, and its reset increment is $-2a$. This remains true at both interfaces: the first mean in every block is zero. Define

$$\begin{aligned} D_\epsilon(a,j) &= 2\,\mathrm{logit}(F_\epsilon(a)+j) + \mathrm{logit}(1 - F_\epsilon(2a)+j), \\ D_\epsilon(a_+,j) + D_\epsilon(a_-,j) &= 0, \\ \lambda = -D_\epsilon(a_+,j) &= D_\epsilon(a_-,j), \qquad x_\pm = \mathrm{logit}(F_\epsilon(a_\pm)+j). \end{aligned} \tag{95}$$

There are L blocks of each type, so the full cycle root is exactly L times the equation displayed here. Its root is independent of L. Integrating the three increments in each block gives the actual score

$$v_{k,t} = \begin{cases} -k\lambda + tx_+, & 0 \le k < L, \\ -L\lambda + (k-L)\lambda + tx_-, & L \le k < 2L. \end{cases} \tag{96}$$

The last reset returns to $v_{0,0} = 0$ because the total drift is zero. The scalar equation and strict convexity ensure that this score, modulo its centering constant, is the actual global minimum.

Analytic root and both vanishing factors. Introduce $u = \epsilon^2$. The expression for $F_\epsilon(t)$ is analytic and even in ϵ near zero, jointly with t near the bounded values used here. Its power series thus defines a real analytic extension $\widetilde{F}(t,u)$ for u in an open neighborhood of zero, with

$$\widetilde{F}(t,u) = \sigma(t) + u\sigma''(t) + O(u^2).$$

Negative u is used only for this analytic extension. All experiments continue to have $u \geq 0$. Define $\widetilde{D}(a, u, j)$ by replacing F_ϵ with $\widetilde{F}$ in (95). The logits have arguments strictly between zero and one at $(a, u, j) = (1, 0, 0)$, so this function is analytic on a fixed neighborhood. Its first-order expansion is

$$\widetilde{D}(a, u, j) = -uK(a) + A(a)j + O(u^2 + |uj| + j^2), \tag{97}$$

uniformly for a near one, where

$$K(a) = 2\tanh(a/2) - \tanh a, \qquad A(a) = \frac{2}{\sigma'(a)} + \frac{1}{\sigma'(2a)} = 6 + 4\cosh a + 2\cosh(2a).$$

To verify the coefficients, $\widetilde{D}(a, 0, 0) = 2a - 2a = 0$, $\partial_u \widetilde{D}(a, 0, 0) = 2\sigma''(a)/\sigma'(a) - \sigma''(2a)/\sigma'(2a) = -K(a)$, and differentiating the three logits in j gives $A(a)$.

Set

$$\mathcal{G}(u, \xi, j) = \widetilde{D}(1 + \xi/2, u, j) + \widetilde{D}(1 - \xi/2, u, j).$$

At the origin, $\mathcal{G} = 0$ and $\partial_j \mathcal{G} = 2A(1) > 0$. The analytic implicit function theorem gives a unique analytic root $j = j(u, \xi)$ on a fixed neighborhood of $(0, 0)$. By shrinking it, all arguments of the logits remain in $(0, 1)$. The root then coincides with the unique global cycle root of lemma B.6 for physical parameters. Its neighborhood does not depend on L, since $\mathcal{G}$ does not.

The equation is even in ξ, so its unique local root is even in ξ. At $u = 0$, the value $j = 0$ solves it for every small ξ; hence $j(0, \xi) = 0$. Define the analytic drift $\lambda(u, \xi) = -\widetilde{D}(1 + \xi/2, u, j(u, \xi))$. It vanishes for $u = 0$ and for $\xi = 0$. Swapping ξ exchanges the two regions, and their drifts sum to zero, so $\lambda(u, -\xi) = -\lambda(u, \xi)$. Its convergent power series therefore factors as

$$\lambda(u, \xi) = u\xi\, \mathcal{H}(u, \xi^2) \tag{98}$$

for a real analytic $\mathcal{H}$ near the origin. This exact factorization is needed when ξ can be arbitrarily small relative to u.

Differentiate $\mathcal{G} = 0$ in u at zero, with ξ fixed. Using (97) and $j(0, \xi) = 0$ gives

$$j_u(0, \xi) = \frac{K(a_+) + K(a_-)}{A(a_+) + A(a_-)}, \qquad \lambda_u(0, \xi) = K(a_+) - A(a_+) j_u(0, \xi).$$

The first expression is even in ξ. Therefore

$$\mathcal{H}(0, 0) = \lambda_{u\xi}(0, 0) = \frac{1}{2}\left\{ K'(1) - K(1)\frac{A'(1)}{A(1)} \right\} = \frac{C_*}{2}.$$

Analyticity now gives the uniform factored remainder

$$\lambda = \frac{C_*}{2}\epsilon^2 \xi + O(\epsilon^4 \xi + \epsilon^2 \xi^3), \qquad C_* = K'(1) - K(1)\frac{A'(1)}{A(1)}. \tag{99}$$

The implicit constant here is on a fixed local neighborhood, independent of L. A bare $O(\epsilon^4)$ term would not give this conclusion.

Strict positivity without a numerical sign check. For $a > 0$ put $z = \tanh(a/2)$. Then

$$\frac{K(a)}{A(a)} = \frac{z^3(1 - z^2)^2}{2(1 + z^2)(3 + z^4)}, \qquad \frac{d}{dz}\log\frac{K(a)}{A(a)} = \frac{3}{z} - \frac{4z}{1 - z^2} - \frac{2z}{1 + z^2} - \frac{4z^3}{3 + z^4}.$$

At $a = 1$, $0 < z < 1/2$. The last display is strictly greater than $6 - 8/3 - 1 - 1/6 = 13/6$. Also $dz/da = (1 - z^2)/2 > 3/8$. Consequently

$$C_* = K(1)\left.\frac{d}{da}\log\frac{K(a)}{A(a)}\right|_{a=1} > \frac{13K(1)}{16} > 0. \tag{100}$$

Since $\mathcal{H}(0, 0) = C_*/2$ and $x_+(0, 0) = x_-(0, 0) = 1$, continuity permits positive constants $\epsilon_* \leq \bar{\epsilon}$ and $\xi_* \leq 1/16$ so that, throughout $0 < \epsilon \leq \epsilon_*$ and $0 < \xi \leq \xi_*$,

$$\frac{C_*}{4}\epsilon^2 \xi \leq \lambda \leq \frac{3C_*}{4}\epsilon^2 \xi, \qquad 0 < x_\pm \leq 2. \tag{101}$$

These are universal existence radii, independent of L. They need not equal the numerical upper-bound radii in part (i).

Full conditional KL decomposition. Fix any $\tau > 0$ and set $\gamma = \lambda/\tau$ and $s_\pm = x_\pm/\tau$. Under the target, the region masses are

$$(\theta_+, \theta_-) = \frac{(Z(a_+), Z(a_-))}{Z(a_+) + Z(a_-)},$$

the block index within each region is uniform on L points, and the conditional type laws are q_{a_+} and q_{a_-}. For the DPO score (96), let $S_L(\gamma) = \sum_{k=0}^{L-1} e^{-\gamma k}$. The unnormalized region masses are

$$Z(s_+)S_L(\gamma), \qquad Z(s_-)e^{-L\gamma}\sum_{r=0}^{L-1} e^{\gamma r} = e^{-\gamma}Z(s_-)S_L(\gamma).$$

Hence

$$(\widehat{\theta}_+, \widehat{\theta}_-) = \frac{(Z(s_+), e^{-\gamma}Z(s_-))}{Z(s_+) + e^{-\gamma}Z(s_-)}.$$

The block-index laws are $P_{\gamma,L}$ in the first region and its reflection in the second; the conditional type laws are q_{s_+} and q_{s_-}, independent of the block index within a region. Apply the KL chain rule and invariance of the uniform law under reflection to obtain the exact identity

$$\begin{aligned}\mathrm{Reg}_{\mu,1}(p^{\mathrm{D}}_{w^{\mathsf{unif}},\tau}) = {} & \mathrm{KL}(\widehat{\theta}\|\theta) + G_L(\gamma) \\ & + \widehat{\theta}_+ \,\mathrm{KL}(q_{s_+}\|q_{a_+}) + \widehat{\theta}_- \,\mathrm{KL}(q_{s_-}\|q_{a_-}).\end{aligned} \tag{102}$$

In particular, the type divergences retain their actual conditional weights, which are positive and sum to one. No uniform lower bound on an individual region mass is required.

Every positive temperature and explicit constants. For $0 < \tau \le 8$, lemma B.8 gives

$$G_L(\lambda/\tau) \ge \frac{e^{-1}}{32}\min\{L^2\lambda^2/\tau^2, 1\} \ge \frac{e^{-1}}{2048}\min\{L^2\lambda^2, 1\}.$$

For $\tau \ge 8$, (101) implies $0 < s_\pm \le 1/4 \le a_\pm/2$ and $a_\pm \in [1/2, 2]$. Both conditional type divergences are at least $e^{-4}/48$ by lemma B.8. Their weights sum to one, and $e^{-4}/48 > e^{-1}/2048$, so the same lower bound holds in this case. Since all other terms in (102) are nonnegative,

$$\inf_{\tau>0} \mathrm{Reg}_{\mu,1}(p^{\mathrm{D}}_{w^{\mathsf{unif}},\tau}) \ge \frac{e^{-1}}{2048}\min\{L^2\lambda^2, 1\}. \tag{103}$$

Combine this with (101) and $n = 6L$. One valid explicit lower constant is

$$c_C = \frac{e^{-1}}{2048}\min\left\{\frac{C_*^2}{576}, 1\right\}. \tag{104}$$

For the upper bound, the constructed perturbation has norm at most ξ. Part (i), $c_\epsilon \ge 1/2$, and $\xi \le 1/16$ imply

$$\mathcal{R}^{\mathsf{temp}}_{\mu,1} \le \min\left\{\frac{289}{2\kappa_*^2}n^2\epsilon^4\xi^2, \frac{289}{512}\right\}.$$

Thus $C_C = 289/(2\kappa_*^2)$ is a valid upper constant in (38). The constants and both local radii are independent of graph size. Part (ii) is proved. □

C.4 Proof of theorem 3

We keep the target coefficient equal to one. The proof first constructs a uniform sensitivity bound for the exact weighted cycle root. It then turns the difference between two score profiles into policy regret, allowing separate temperatures throughout.

Uniform scalar representation. Put $\eta = \epsilon^2$, $r_0 = 1/320$, and define

$$f(\eta, d) = \tfrac{1}{4}\sigma(d - 2\sqrt{\eta}) + \tfrac{1}{2}\sigma(d) + \tfrac{1}{4}\sigma(d + 2\sqrt{\eta}).$$

The function is smooth from the right at $\eta = 0$: the even Taylor expansion in $\sqrt{\eta}$ has only integer powers of η. In particular,

$$f(0, d) = \sigma(d), \qquad f_\eta(0, d) = \sigma''(d). \tag{105}$$

All derivatives used below are bounded on compact sets in (η, d). For example, the centered-difference integral identity and $|\sigma''| \le 1/4$ give

$$|f(\eta, d) - \sigma(d)| \le \eta/4. \tag{106}$$

For a weighting in (39), set $b_i = (nw_i)^{-1} \in [\kappa^{-1}, \kappa]$. Interpolate between the two environments by

$$\mu_{k,t}(r) = (1 + s_k r)t, \qquad |r| \le r_0, \qquad a_+ = 1 + r, \quad a_- = 1 - r.$$

The short edge differences in block k are both a_{s_k}, and its reset edge difference is $-2a_{s_k}$. The weighted stationarity equation has the exact form

$$v_{i+1} - v_i = \operatorname{logit}\{f(\eta, \mu_{i+1} - \mu_i) + jb_i\}, \qquad \sum_i (v_{i+1} - v_i) = 0. \tag{107}$$

Indeed, $w_i(\sigma(v_{i+1} - v_i) - Q_i)$ is constant around the cycle; j is n times that constant. Strict convexity identifies this root with the actual population optimizer.

It is important to make uniformity in n explicit. For each half $s \in \{+, -\}$ and position $t \in \{0, 1, 2\}$, let $M_{s,t}$ be the empirical distribution of b_{3k+t} over its L blocks. We enlarge these six empirical measures to arbitrary probability measures on $[\kappa^{-1}, \kappa]$. Define

$$\begin{aligned} h_s(j, \eta, a) = \sum_{t=0}^{1} \int \operatorname{logit}\{f(\eta, a) + jb\}\, dM_{s,t}(b) \\ + \int \operatorname{logit}\{f(\eta, -2a) + jb\}\, dM_{s,2}(b). \end{aligned} \tag{108}$$

The root solves

$$h_+(j, \eta, 1 + r) + h_-(j, \eta, 1 - r) = 0. \tag{109}$$

For empirical measures the unnormalized cycle equation is exactly L times (109); no approximation of the weights has occurred.

Here is a common compact domain for all six-measure choices. Put

$$\alpha = \sigma(-3), \quad k_0 = \alpha(1 - \alpha), \quad k_1 = (\alpha/2)(1 - \alpha/2), \quad j_0 = \frac{\alpha}{2\kappa},$$

and take

$$0 \le \eta \le \eta_0 := \min\{1/256,\, 8\alpha k_0/\kappa^2,\, \bar{\epsilon}^2\}.$$

For the edge arguments under consideration, all shifted sigmoid arguments lie in $[-3, 3]$. Thus $f \in [\alpha, 1 - \alpha]$, and $f + jb \in [\alpha/2, 1 - \alpha/2]$ for $|j| \le j_0$. At $j = 0$, each logit differs from its unperturbed edge difference by at most $\eta/(4k_0)$, by (106). There are six terms in (109), whereas its derivative in j is at least $24/\kappa$. Bracketing therefore gives a unique root satisfying

$$|j(\eta, r)| \le C_j \eta, \qquad C_j = \frac{\kappa}{16 k_0}. \tag{110}$$

The choice of η_0 ensures that the bracket is inside $[-j_0, j_0]$. For detail, let $\mathcal{H}(j, \eta, r) = h_+(j, \eta, 1 + r) + h_-(j, \eta, 1 - r)$. At $j = 0$ the six unperturbed edge differences sum to zero, so $|\mathcal{H}(0, \eta, r)| \le 6\eta/(4k_0)$. Each logit contributes derivative $b/[p(1 - p)] \ge 4/\kappa$, and hence $\mathcal{H}_j \ge 24/\kappa$. Integrating this derivative shows

$$\mathcal{H}(-C_j\eta, \eta, r) \le 0 \le \mathcal{H}(C_j\eta, \eta, r), \qquad C_j\eta_0 \le \frac{\kappa}{16k_0}\frac{8\alpha k_0}{\kappa^2} = j_0.$$

The intermediate value theorem and strict monotonicity prove the bracket. For $x, y \in \{\eta, r\}$, implicit differentiation gives

$$\begin{aligned} j_x &= -\frac{\mathcal{H}_x}{\mathcal{H}_j}, \\ j_{xy} &= -\frac{\mathcal{H}_{xy} + \mathcal{H}_{xj} j_y + \mathcal{H}_{yj} j_x + \mathcal{H}_{jj} j_x j_y}{\mathcal{H}_j}, \end{aligned}$$

with every derivative on the right evaluated at $(j(\eta, r), \eta, r)$. All probability denominators and $\mathcal{H}_j$ are bounded away from zero. These formulas, and their derivatives, consequently give continuous derivatives in (η, r), uniformly over the six measures. To justify this last uniformity, probability measures on the compact interval form a compact space in the weak topology; integrals of the displayed continuous bounded integrands and their derivatives are continuous in those measures. The unique root is also continuous, by strict monotonicity and the common derivative lower bound. Repeated implicit differentiation expresses its derivatives through these integrals and the same nonzero denominator. This argument contains no dimension-dependent norm equivalence.

A sign that no shared weighting can cancel. Let $\Psi(\eta, r) = h_+(j(\eta, r), \eta, 1 + r)$. For empirical measures, $L\Psi$ is the cumulative score increment across the first half. At $\eta = 0$, $j = 0$ and $\Psi = 0$ for every r. Set

$$\begin{aligned} d(a) &= 2\tanh(a/2) - \tanh a, \\ c(a) &= d'(a) = 4\{\sigma'(a) - \sigma'(2a)\}, \\ H_s(a) &= \frac{\int b\, dM_{s,0} + \int b\, dM_{s,1}}{\sigma'(a)} + \frac{\int b\, dM_{s,2}}{\sigma'(2a)}. \end{aligned}$$

Using (105) and $\sigma''(u)/\sigma'(u) = -\tanh(u/2)$ gives

$$(h_s)_\eta(0, 0, a) = -d(a), \qquad (h_s)_j(0, 0, a) = H_s(a).$$

Hence, with $H_s = H_s(a_s)$ and $d_s = d(a_s)$,

$$j_\eta(0, r) = \lambda(r) := \frac{d_+ + d_-}{H_+ + H_-}. \tag{111}$$

For clarity, differentiation of the root first gives, at general (η, r),

$$j_r = -\frac{(h_+)_a - (h_-)_a}{(h_+)_j + (h_-)_j}, \qquad \Psi_r = \frac{(h_+)_a (h_-)_j + (h_-)_a (h_+)_j}{(h_+)_j + (h_-)_j}.$$

At $\eta = 0$, both a derivatives vanish. Their total η derivatives along the root are $-c_s + \lambda H'_s(a_s)$. Therefore

$$\Psi_{\eta r}(0, r) = -\frac{c_+ H_- + c_- H_+}{H_+ + H_-} + (d_+ + d_-)\frac{H_+ H_-}{(H_+ + H_-)^2}(q_+ + q_-), \quad q_s = \frac{H'_s(a_s)}{H_s(a_s)}. \tag{112}$$

Since $1/\sigma'(a) = 2 + 2\cosh a$, q_s is a positive weighted average of $\tanh(a_s/2)$ and $2\tanh a_s$. In particular,

$$\tanh(a_s/2) \le q_s \le 2\tanh a_s.$$

Also $H_+ H_-/(H_+ + H_-)^2 \le 1/4$. Consequently

$$\Psi_{\eta r}(0, r) \le -\min_{|a-1| \le r_0} c(a) + \max_{|a-1| \le r_0} d(a) \max_{|a-1| \le r_0} 2\tanh a. \tag{113}$$

We verify the strict sign analytically. Let $t = \tanh(1/2)$ and $u = \tanh 1 = 2t/(1 + t^2)$. Then

$$\begin{aligned} c(1) - 2u\, d(1) &= u^2 - t^2 - 2u(2t - u) \\ &= \frac{t^2(3 - 10t^2 - t^4)}{(1 + t^2)^2} \ge \frac{7}{25}t^2 > \frac{1}{20}. \end{aligned}$$

Here $11/24 \le t < 1/2$: the lower bound follows from $\tanh x \ge x - x^3/3$ for $x \ge 0$, whose derivative inequality is $x^2 - \tanh^2 x \ge 0$. The displayed rational bound uses $t^2 < 1/4$, and $7(11/24)^2/25 = 847/14400 > 1/20$. On the relevant interval, $|c'| \le 3$, $|d'| \le 1$, and $|(2\tanh a)'| \le 2$. Moreover $d(1) \le 1$ and $2\tanh 1 \le 2$. Thus the gap in (113) is at least

$$\frac{1}{20} - 7r_0 - 2r_0^2 > \frac{1}{40}, \qquad r_0 = 1/320.$$

It follows that $\Psi_{\eta r}(0, r) \le -1/40$ uniformly in all six measures and $|r| \le r_0$.

Uniform continuity on the common compact parameter space now supplies $\eta_\kappa > 0$, no larger than η_0, such that $\Psi_{\eta r}(\eta, r) \le -1/80$ for $0 \le \eta \le \eta_\kappa$ and $|r| \le r_0$. Set $\gamma = 1/80$. Since $\Psi_r(0, r) = 0$, integration in η yields

$$\Psi_r(\eta, r) \le -\gamma\eta. \tag{114}$$

For an individual block, let $D_k(\eta, r)$ be the sum of its three increments in (107), using the actual root. At $\eta = 0$, $D_k = 0$ for every r. The same compact-domain argument, now also allowing its three individual b values in $[\kappa^{-1}, \kappa]^3$, bounds $\partial_\eta \partial_r D_k$ by some $M_\kappa \ge 1$. Therefore

$$|\partial_r D_k(\eta, r)| \le M_\kappa \eta. \tag{115}$$

Both (114) and (115) have neighborhoods independent of the number and arrangement of the weights.

Two score profiles separated on a positive fraction of classes. Let $0 < \xi \le 1/160$, and choose the additive gauge so that $x_k(r) := v_{3k}(r)$ satisfies $x_0(r) = 0$. Put

$$\Delta_k = x_k(\xi/2) - x_k(-\xi/2), \qquad \delta = \eta\xi.$$

From (114) and (115),

$$\Delta_0 = 0, \quad \Delta_L \le -\gamma L\delta, \quad |\Delta_{k+1} - \Delta_k| \le M_\kappa \delta. \tag{116}$$

Take $\vartheta = \gamma/(8M_\kappa)$, $m = \lfloor \vartheta L \rfloor$, and subsets of the $2L$ block anchors

$$A = \{0, \dots, m\}, \qquad B = \{L - m, \dots, L + m\}.$$

They are disjoint, each has at least a fraction $\rho_\kappa := \vartheta/2$ of the anchors, and

$$\min_{k \in A} \Delta_k - \max_{k \in B} \Delta_k \ge G := \frac{3\gamma L\delta}{4} = \frac{\gamma n \delta}{8}. \tag{117}$$

Indeed, the deviations from Δ_0 on A and from Δ_L on B are at most $M_\kappa m\delta \le \gamma L\delta/8$. The cardinality claim also holds when $m = 0$, since $m + 1 \ge \vartheta L$.

Small policy regret confines the temperature. Equations (106) and (110) imply the uniform edge bound

$$|(v_{i+1} - v_i) - (\mu_{i+1} - \mu_i)| \le E_\kappa \eta, \qquad E_\kappa = \frac{1}{4k_0} + \frac{\kappa C_j}{k_1}.$$

Reduce η_κ further so that $E_\kappa \eta_\kappa \le 1/8$. Each short score increment then lies in $[3/4, 5/4]$, and each block has target means $(0, a, 2a)$ with $a \in [15/16, 17/16]$.

Within a block, condition further on its first and last classes. Their fitted last-class probability is $\sigma((b_0^{\text{score}} + b_1^{\text{score}})/\tau)$, whereas the target probability is $\sigma(2a)$. The fitted mass of this two-class subset within its block is at least $1/2$, because both short increments are positive. For $\tau \ge 4$ or $\tau \le 1/4$, the difference of these two conditional probabilities is at least

$$d_T = \min\{\sigma(15/8) - \sigma(5/8),\, \sigma(6) - \sigma(17/8)\} > 0.$$

The KL chain rule and Bernoulli Pinsker inequality give conditional block KL at least d_T^2. Explicitly, if p_k, q_k are the fitted and target conditional laws in a block and $S = \{0, 2\}$, then

$$\mathrm{KL}(p_k \| q_k) \ge p_k(S)\, \mathrm{KL}(p_k(\cdot \mid S) \| q_k(\cdot \mid S)) \ge \tfrac{1}{2}\, 2d_T^2 = d_T^2.$$

Here $p_k(S) \ge 1/2$ follows from $e^{v_{k,0}/\tau} + e^{v_{k,2}/\tau} \ge e^{v_{k,1}/\tau}$. Applying the chain rule once more and summing over the fitted block masses, which add to one, proves

$$\mathrm{Reg}_{\mu,1}(p^{\mathrm{D}}_{w,\tau}) \ge d_T^2 \quad \text{if } \tau \notin (1/4, 4). \tag{118}$$

Converting the anchor separation into regret. Let $T = \{3k : 0 \le k < 2L\}$ be the set of anchors. All its target means are zero, so its target conditional distribution u_T is uniform in both environments. Its target mass is at least

$$q_0 = \{1 + e^{17/16} + e^{17/8}\}^{-1} > 0.$$

If a fitted policy gives T mass less than $q_0/2$, binary Pinsker already gives regret at least $q_0^2/2$. Otherwise the KL chain rule gives, writing R_a for the full regret in environment a,

$$\mathrm{KL}(p_a^{\mathrm{D}}(\cdot \mid T) \| u_T) \le 2R_a/q_0. \tag{119}$$

We use an elementary consequence of reverse KL which controls most uniformly weighted coordinates, rather than the variance of every log ratio. For $p > 0$ and uniform u on N points, put $t_k = p_k/u_k$ and $F(t) = t \log t - t + 1$. Since $\mathbb{E}_u t = 1$, $\mathrm{KL}(p\|u) = \mathbb{E}_u F(t)$. The derivative $F'(t) = \log t$ is negative below one and positive above one. Moreover, for $b \geq 0$,

$$F(e^b) - F(e^{-b}) = 2(b \cosh b - \sinh b) \geq 0,$$

because this expression vanishes at zero and has derivative $2b \sinh b \geq 0$. Thus the smaller boundary value outside $[e^{-b}, e^b]$ is $F(e^{-b})$. For $b > 0$,

$$\inf_{|\log t| > b} F(t) = F(e^{-b}) = \int_0^b s e^{-s}\, ds \geq \frac{1}{2e} \min\{b^2, 1\}.$$

The last inequality integrates over $[0, \min\{b, 1\}]$, where $e^{-s} \geq e^{-1}$. Multiplying this pointwise lower bound by the uniform mass of the exceptional coordinates and using $\mathbb{E}_u F(t) = \mathrm{KL}(p\|u)$ gives

$$u\{k : |\log(p_k/u_k)| > b\} \leq \frac{2e\ \mathrm{KL}(p\|u)}{\min\{b^2, 1\}}. \tag{120}$$

Allow arbitrary, separately chosen temperatures $\tau_+, \tau_- > 0$, and put $R = \max\{R_+, R_-\}$. If $R \geq c_0 := \min\{d_T^2, q_0^2/2\}$, a constant lower bound already holds. Otherwise (118) confines both temperatures to $(1/4, 4)$, and (119) holds for both. Choose $b = G/32$. At every anchor outside the exceptional sets in (120),

$$|x_k^a - C_a| \leq \tau_a b \leq 4b, \qquad C_a = \tau_a \log\left(\frac{1}{2L} \sum_j e^{x_j^a/\tau_a}\right).$$

Thus, if an anchor is good in both environments, $|\Delta_k - (C_+ - C_-)| \leq 8b$. Two such anchors have a difference in Δ at most $16b = G/2$. By (117), good anchors cannot exist in both A and B. The union of exceptional sets therefore has uniform anchor mass at least ρ_κ. On the other hand, (119) and (120) bound that union by $8eR/(q_0 \min\{b^2, 1\})$. It follows that

$$\begin{aligned} R &\geq \frac{q_0 \rho_\kappa}{8e} \min\{(G/32)^2, 1\} \\ &\geq \frac{q_0 \rho_\kappa}{8e} \left(\frac{\gamma}{256}\right)^2 \min\{n^2 \epsilon^4 \xi^2, 1\}. \end{aligned} \tag{121}$$

We may therefore take

$$c_\kappa = \min\left\{c_0, \frac{q_0 \rho_\kappa}{8e} \left(\frac{\gamma}{256}\right)^2\right\}, \qquad \epsilon_\kappa = \sqrt{\eta_\kappa}.$$

The local radius and M_κ, hence ρ_κ, are the positive constants established by the uniform scalar argument above. They depend only on κ and fixed logistic constants.

The two temperatures were arbitrary. Taking their infima independently proves (44), including possibly unattained infima by approximation. Both environments lie in the allowed perturbation set, so taking the infimum over w proves the lower bound in (42). Equal weights give the upper bound from theorem 2(i); for example one may take $C_\kappa = \max\{289/(2\kappa_*^2), 1\}$. For $\xi \leq \xi_{\mathrm{bal}}$, corollary 1 supplies an exact weighting in $\mathcal{W}_3(n) \subseteq \mathcal{W}_\kappa(n)$ for every perturbation, proving the zero adaptive value. At $\epsilon = 0$ ordinary BT gives exact recovery, and at $\xi = 0$ the baseline calibration with equal weights does so. These cases complete the statement. □

D Proofs of auxiliary results

D.1 Proof of proposition 2

On a tree every prescribed vector of edge logits has a potential, unique up to a constant: set its value at a root and integrate logits along the unique paths. Taking those logits to be $g_\nu((\mathsf{B}\mu)_e)$ fits every comparison. By lemma B.4, every positive weighting and every allowed proper loss has this same centered global minimum v. The error $e = v - \mu$ satisfies

$$|(\mathsf{B}e)_e| \leq \mathcal{B}(h) V$$

by lemma B.5. Between any two vertices, sum along their unique path, whose length is at most D_G. Thus $\operatorname{osc}(v-\mu) \le D_G \mathcal{B}(h) V$. Apply lemma B.3 at $\tau = \beta$ to obtain (54). As $\tau \uparrow \infty$, the policy converges to ρ; continuity of KL and lemma B.3 with score zero give the alternative bound $\operatorname{Reg}_\beta(\rho) \le M^2/(8\beta)$. Taking the infimum proves (55). This use of a limit does not claim that the infimum is attained at a finite temperature.

□

D.2 Proof of theorem 4

We first verify the parameter facts used in the statement. The identity for q_s in (64) follows by adding the two logistic fractions. It gives $q_s > 1/2$. Since $u > 1$, it also gives $q_s < \sigma(a)$, hence $0 < g_A < a$. Strict monotonicity of σ implies $q_s > \sigma(a-\epsilon)$, so $0 < d_A < \epsilon$; also $d_A < a$. Solving the expressions for q_ℓ and q_s gives (65). Summing the two middle exponential weights gives (66).

Part (i): the exact oracle envelope. Let v be a finite global minimizer of an allowed loss, and set $x = v_2 - v_1$, $y = v_3 - v_2$. Write its weighted derivative residuals as

$$f_1 = w_1 a_\ell(x)(\sigma(x) - q_s), \qquad f_2 = w_2 a_\ell(y)(\sigma(y) - q_s),$$
$$f_3 = w_3 a_\ell(x+y)(\sigma(x+y) - q_\ell).$$

The objective is constant along additive score shifts. Stationarity on the centered space therefore implies full stationarity, or $f_1 = f_2 = -f_3$. The common value cannot be zero: this would require $x = y = g_A$ and $x + y = 2a$, whereas $g_A < a$. It cannot be negative either: then $x, y < g_A$ but $x + y > 2a$. Consequently

$$x > g_A, \qquad y > g_A, \qquad x + y < 2a. \tag{122}$$

Conversely, every (x, y) satisfying (122) has positive short-edge residuals and a negative long-edge residual. Weights proportional to their inverse magnitudes make $f_1 = f_2 = -f_3$ for logistic loss. The unique centered logistic minimum has these increments. This proves both directions without asserting sufficiency of stationarity for a nonconvex proper loss.

Because ρ is uniform, every positive policy can be expressed uniquely using endpoint-centered log weights $(-\alpha, m, \alpha)$. At temperature τ, its score increments are $x = \tau(\alpha + m)$ and $y = \tau(\alpha - m)$. Condition (122) is equivalent to $\alpha > 0$, $|m| < \alpha$, and a nonempty interval

$$\frac{g_A}{\alpha - |m|} < \tau < \frac{a}{\alpha}.$$

That interval is nonempty exactly when $|m| < (1 - g_A/a)\alpha = \zeta_A \alpha$. Thus (67) holds, and logistic loss alone realizes the envelope.

Part (ii): attainment and its boundary. The target parameters are $\alpha = a/\beta$ and $m = H_\beta/\beta > 0$. Part (i) proves exact attainment if and only if $H_\beta < d_A$. The inverse-residual weights displayed after the theorem are valid: $c_0(a - H_\beta) > g_A$, $c_0(a + H_\beta) > g_A$, and $2c_0 a < 2a$, so all three required residual signs hold at $\tau = c_0\beta$.

If $H_\beta = d_A$, hold $\alpha = a/\beta$ fixed and choose $m_k = (1 - 1/(k+1))d_A/\beta$. The resulting policies belong to $\mathcal{A}$ and converge to p^{F}_β. Since the target is strictly positive, KL to the target is continuous on the closed simplex, with the convention $0 \log 0 = 0$. Hence $\mathcal{R}^{\mathsf{env}}_\beta = 0$, although part (i) excludes exact positive-three-edge attainment. If edge $(1, 2)$ is deleted, the score $(-a, d_A, a)$ at $\tau = \beta$ fits both remaining edges: $a - d_A = g_A$ and $a - (-a) = 2a$. The connected tree therefore has exactly the target policy as its minimum for every allowed loss and positive weighting. For $H_\beta > d_A$, part (iii) below gives a strictly positive lower bound, completing this case.

For the unique threshold, put $z = \epsilon/\beta$. Differentiation gives

$$\frac{dH_\beta}{d\beta} = \log\cosh z - z \tanh z < 0.$$

Indeed the right-hand side vanishes at $z = 0$ and its derivative in z is $-z \operatorname{sech}^2 z < 0$. The limits of H_β as $\beta \downarrow 0$ and $\beta \uparrow \infty$ are ϵ and zero, respectively. Because $0 < d_A < \epsilon$, continuity and strict

monotonicity give exactly one crossing. Finally,

$$u - e^{d_A} = \frac{u^2 - 1}{e^a + u} > 0 \tag{123}$$

implies $H_1 = \log u > d_A$, so $\beta_c > 1$.

Part (iii): separation on the closed simplex. Put $\lambda_A = g_A/(2a) \in (0, 1/2)$. From (67),

$$p_2 \le p_1^{\lambda_A} p_3^{1-\lambda_A}. \tag{124}$$

Both positive powers are continuous on $[0, 1]$, so the inequality also holds on the closure of $\mathcal{A}$. The weighted geometric mean $g(x, z) = x^{\lambda_A} z^{1-\lambda_A}$ is concave and homogeneous of degree one. Its tangent plane at the positive endpoint masses of p^{F}_β is $c_1 x + c_3 z$, where

$$c_1 = \lambda_A e^{(a+d_A)/\beta}, \qquad c_3 = (1 - \lambda_A) e^{(d_A - a)/\beta}.$$

Thus $T(p) = p_2 - c_1 p_1 - c_3 p_3 \le 0$ for every policy in that closure. At the target, homogeneity of g yields

$$T(p^{\mathrm{F}}_\beta) = \frac{e^{H_\beta/\beta} - e^{d_A/\beta}}{Z_\beta} > 0.$$

The coefficient vector of T has supremum norm M_β, so $M_\beta \|p - p^{\mathrm{F}}_\beta\|_1 \ge T(p^{\mathrm{F}}_\beta) - T(p) \ge T(p^{\mathrm{F}}_\beta)$. Pinsker's inequality with natural logarithms, $\mathrm{KL}(p\|q) \ge \|p - q\|_1^2/2$, proves (68).

Part (iv): uniform fourth-order bounds. For $1 \le a \le 2$ and $0 < \epsilon \le 1$, (123) gives

$$u - e^{d_A} \ge \frac{\epsilon^2}{e^2 + \cosh 1}, \qquad M_1 \le e^{2a} \le e^4, \qquad Z_1 \le 2\cosh 2 + \cosh 1.$$

Here $u^2 - 1 = \sinh^2 \epsilon \ge \epsilon^2$ was used in the first inequality. Substitution into part (iii) gives the displayed c_A. On the three possible connected two-edge supports, conditional properness gives, respectively,

$$(x, y) = (g_A, g_A), \quad (g_A, 2a - g_A), \quad (2a - g_A, g_A).$$

Each point lies in the closure of the open triangle (122). At any fixed positive temperature its policy is therefore in the closure of (67). This proves the same lower bound when one pair may be omitted.

For the upper bound, let s be the unique solution of

$$\sigma(s) + \sigma(2s) = q_s + q_\ell. \tag{125}$$

The left side is strictly increasing. Its values at $s = g_A$ and $s = a$ lie strictly below and above the right side, so $g_A < s < a$. At the score $(-s, 0, s)$, the two short-edge residuals coincide and the long-edge residual is their negative. Hence this is the actual equal-weight logistic minimum. At $\tau = s/a > 0$ its policy is proportional to $(e^{-a}, 1, e^a)$. The target changes only the middle reward by $\log\cosh\epsilon$, so lemma B.3 gives

$$\mathrm{Reg}_1(p^{\mathrm{D}}_{w^{\mathsf{unif}}, s/a}) \le \frac{(\log\cosh\epsilon)^2}{8} \le \frac{\epsilon^4}{32}.$$

The final inequality follows by integrating $\tanh t \le t$ for $t \ge 0$. All four parts of theorem 4 are proved. □

D.3 Proof of proposition 3

Write $b = g_\nu(1)$ and $g_2 = g_\nu(2)$; (57) gives $g_2 > 2b$. The function $H(c) = \sigma(c) + \sigma(2c)$ is continuous and strictly increasing, with limits zero and two. At the two relevant points,

$$H(b) = F_\nu(1) + \sigma(2b) < F_\nu(1) + F_\nu(2), \qquad H(g_2/2) = \sigma(g_2/2) + F_\nu(2) > F_\nu(1) + F_\nu(2).$$

This proves existence and uniqueness of c_ν, with $b < c_\nu < g_2/2$.

After adding the endpoint pair, orient the cycle forward along the old path and back to its first vertex. The nominal mean increments are $1, 1, -2$ in every block. At $v_{k,t} = c_\nu t$, a short edge has residual

$\sigma(c_\nu) - F_\nu(1)$, while a reset edge has residual $F_\nu(2) - \sigma(2c_\nu)$. These residuals are equal and strictly positive by the calibration equation and the preceding inequalities. Equal weights therefore give a constant circulation, hence the actual logistic minimum. At $\tau = c_\nu\beta$ its policy is $\mathsf{S}_\rho(\mu/\beta)$ for every positive reference.

We now prove uniqueness among all missing pairs. Suppose adding a chord permits the target at some positive temperature and some allowed loss with positive weights. The target score necessarily has the form $v = c\mu +$ constant, with $c = \tau/\beta > 0$; this follows by cancelling ρ in the reference-relative policy map. Its derivative residual flow is

$$f_e = w_e a_\ell((\mathsf{B}v)_e)\{\sigma((\mathsf{B}v)_e) - Q_e\}, \qquad \mathsf{B}^\top f = 0.$$

For any bridge, sum this conservation law over either component left by its deletion. Only the bridge crosses the cut, so its flow is zero. On the unique cycle of a path plus one chord, conservation then makes the cycle flow constant when oriented around the cycle. In particular, one zero-flow cycle edge forces zero flow on the whole graph.

For every old short edge the residual vanishes exactly at $c = b$. For every old reset edge it vanishes exactly at $c = g_2/2$; these two values are distinct. Consider all possibilities.

(a) If $c \notin \{b, g_2/2\}$, every old path edge has nonzero flow, since its weight and derivative factor are positive. Hence none can be a bridge, and all old edges must lie on the unique cycle. The chord must therefore join the two endpoints of the path.

(b) If $c = b$, all short edges have zero flow and all old reset edges have nonzero flow. The subpath closed by a missing pair has length at least two. Old reset edges are isolated from one another by short edges, so every such subpath contains a short edge. The unique cycle therefore contains a zero-flow edge. It follows that all flows vanish, contradicting the presence of at least one old reset edge because $J \geq 2$.

(c) If $c = g_2/2$, all reset edges have zero flow and all short edges have nonzero flow. If the unique cycle contains a reset edge, its flow is zero, again forcing all flows to vanish. If it contains no reset edge, it lies entirely within one three-vertex block. Another block exists because $J \geq 2$; its short edges are bridges with nonzero flow, also a contradiction.

These cases cover every $c > 0$ and every missing pair. They use only necessary stationarity for the proper losses and preserve strict positivity of every old and new weight. The endpoint pair is consequently the unique permitted addition. □

D.4 Proof of corollary 1

Suppose first that $0 < \epsilon \leq \bar{\epsilon}$. The baseline drift $d_\epsilon = g_\epsilon(2) - 2g_\epsilon(1)$ obeys $d_\epsilon \geq K\epsilon^2/2$ by (57) with $h = \epsilon$ and $V = \epsilon^2$. Calibration gives $g_\epsilon(1) < c_\epsilon < g_\epsilon(2)/2$ as in proposition 3. Therefore $j_0 = \sigma(c_\epsilon) - F_\epsilon(1) > 0$. For the function D_ϵ in (95),

$$D_\epsilon(1, 0) = -d_\epsilon, \qquad D_\epsilon(1, j_0) = 2c_\epsilon - 2c_\epsilon = 0.$$

For $0 \leq j \leq j_0$, the short-edge logits lie between $g_\epsilon(1)$ and c_ϵ; the reset logits lie between $-g_\epsilon(2)$ and $-2c_\epsilon$. The elementary mixture bounds $g_\epsilon(t) \in [t - 2\epsilon, t + 2\epsilon]$ and (91) put all these logits in $[-3, 3]$. Each logit derivative is therefore at most $1/\kappa_3$, and

$$d_\epsilon = \int_0^{j_0} \partial_j D_\epsilon(1, j)\, dj \leq \frac{3j_0}{\kappa_3}, \qquad j_0 \geq \frac{K\kappa_3}{6}\epsilon^2.$$

For any allowed perturbation, the residuals at the desired score $c_\epsilon\mu$ satisfy (93):

$$r_i = \sigma(c_\epsilon\Delta_i) - Q_i = j_0 + e_i, \qquad |e_i| \leq \frac{17}{2}\epsilon^2\xi.$$

If $\xi \leq \xi_{\mathrm{bal}} = K\kappa_3/102$, then $|e_i| \leq K\kappa_3\epsilon^2/12 \leq j_0/2$. Hence $j_0/2 \leq r_i \leq 3j_0/2$ for every edge. Define

$$w_i = \frac{r_i^{-1}}{\sum_{j=0}^{n-1} r_j^{-1}}.$$

Since $2/(3j_0) \le r_i^{-1} \le 2/j_0$, the denominator lies between $2n/(3j_0)$ and $2n/j_0$. Consequently $1/(3n) \le w_i \le 3/n$. The products $w_i r_i$ are a common positive constant, so their flow is conserved around the cycle. Logistic strict convexity makes $c_\epsilon \mu$, after centering, the unique global minimum. At $\tau = c_\epsilon \beta$ it maps to the target.

If $\epsilon = 0$, the comparison logits already equal the mean increments, and the ordinary BT-realizable argument gives exact agreement with equal weights, $c_0 = 1$, and $\tau = \beta$. No reciprocal of a zero residual is used in this case. The witness for positive ϵ is a population attainability argument, not a proposed learning rule. □

D.5 Proof of corollary 2

Both arrangements contain L copies of each block. Each block of type s contributes exactly the three oriented edge probabilities $F_\epsilon(a_s), F_\epsilon(a_s), F_\epsilon(-2a_s)$, because the next block starts at mean zero. Thus the edge-probability multisets, class-mean multisets, graph, and weights agree. The reset between the last and first blocks obeys the same formula. The common probability multiset also makes the scalar root j of lemma B.6 identical in both arrangements. The anchor-score increment of a block of type s is

$$D_s = 2\,\mathrm{logit}\{F_\epsilon(a_s) + j\} + \mathrm{logit}\{F_\epsilon(-2a_s) + j\}.$$

There are L blocks of each type, so the cycle equation is $L(D_+ + D_-) = 0$. The local sign calculation in section C.3 gives $D_+ < 0$ for positive ϵ, ξ in the stated neighborhood. Thus $D_- = -D_+ > 0$. Setting $\Lambda = -D_+$, the normalized anchor sums alternate between 0 and -1 in the alternating arrangement. In consecutive order they descend from 0 to $-L$ and then return to zero. These are precisely the block sums drawn in figure 2.

For alternating blocks, the data and uniform reference are invariant under a rotation by six classes. The unique centered BT minimizer has the same invariance: rotating a minimizer gives another centered minimizer. Its six distinct score entries minimize precisely the equal-weight six-cycle objective, since each of the six edge terms is repeated L times with weight $1/(6L)$. Both the fitted and target policies assign a six-class cell mass $1/L$. Hence, for every τ, their full KL equals the six-class KL:

$$\sum_{k=0}^{L-1} \sum_{j=0}^{5} \frac{p_j}{L} \log \frac{p_j/L}{q_j/L} = \sum_{j=0}^{5} p_j \log \frac{p_j}{q_j}.$$

Applying theorem 2(i) at size six and temperature c_ϵ gives (76); one may use $C_{\mathrm{alt}} = 36 \cdot 289/(2\kappa_*^2)$. The regional arrangement is exactly (37), so theorem 2(ii) gives (77). Substitution of $\epsilon = \xi = n^{-1/3}$ gives $\epsilon^4 \xi^2 = n^{-2}$, with both parameters eventually inside the fixed positive neighborhood. □

D.6 Proof of proposition 4

The common-law target is $p_\beta^{\mathrm{F}} = \mathsf{S}_\rho(\mu^0/\beta)$. If a score v maps to this target at temperature τ, then for every class i,

$$\frac{\rho_i e^{v_i/\tau}}{\sum_j \rho_j e^{v_j/\tau}} = \frac{\rho_i e^{\mu_i^0/\beta}}{\sum_j \rho_j e^{\mu_j^0/\beta}},$$

$$\frac{v_i}{\tau} - \frac{\mu_i^0}{\beta} = \log \sum_j \rho_j e^{v_j/\tau} - \log \sum_j \rho_j e^{\mu_j^0/\beta}.$$

The right side is independent of i. Thus $v = c\mu^0 + \text{constant}$ with $c = \tau/\beta > 0$. In the forward cycle orientation the target-score residuals are

$$r_s(c) = \sigma(c) - F_\nu(1), \qquad r_r(c) = \sigma(-2c) - F_\nu(-2) = F_\nu(2) - \sigma(2c). \tag{126}$$

The stationarity equation at vertex i is $w_{i-1} r_{i-1}(c) - w_i r_i(c) = 0$, with indices modulo n. Hence all the weighted residuals equal one scalar j.

Necessity of the two weight types. Suppose (i) holds. If $j = 0$, positivity of all weights implies $r_s(c) = r_r(c) = 0$, or $c = b$ and $2c = g_2$; this contradicts $g_2 > 2b$. If $j < 0$, both residuals in (126) are negative. Strict monotonicity of σ then gives $c < b$ and $c > g_2/2$, again impossible. Therefore $j > 0$ and

$$b < c < g_2/2, \qquad r_s(c) > 0, \qquad r_r(c) > 0.$$

Every short edge must have weight $j/r_s(c)$, and every reset edge weight $j/r_r(c)$. There are $2J$ short edges and J reset edges, so their normalization is $2Jw_s + Jw_r = 1$. This proves (iii).

Sufficiency and the unique recovery temperature. Suppose (iii) holds. Let

$$H_w(c) = w_s\sigma(c) + w_r\sigma(2c) - w_sF_\nu(1) - w_rF_\nu(2).$$

Then

$$\begin{aligned} H_w(b) &= w_r\{\sigma(2b) - F_\nu(2)\} < 0, \\ H_w(g_2/2) &= w_s\{\sigma(g_2/2) - F_\nu(1)\} > 0, \\ H_w'(c) &= w_s\sigma'(c) + 2w_r\sigma'(2c) > 0. \end{aligned}$$

The intermediate value theorem and strict monotonicity give one and only one root $c \in (b, g_2/2)$. At this root,

$$w_sr_s(c) - w_rr_r(c) = H_w(c) = 0.$$

Thus $c\mu^0$ has constant weighted residual flow around the cycle. After centering, it satisfies the first-order equation of the strictly convex logistic objective and is its unique global minimizer. Taking $\tau = c\beta$ maps this score to $\mathsf{S}_\rho(c\mu^0/(c\beta)) = p^{\mathrm{F}}_\beta$, proving (i) and the calibration formula. Implication (i)⇒(ii) follows immediately.

A zero infimum cannot hide at a temperature endpoint. Fix any positive weighting and let $v = v^{\mathrm{D}}_w$ be its finite score, which is independent of τ. This score is not constant: at a constant score all short residuals would equal $1/2 - F_\nu(1) < 0$ and all reset residuals $F_\nu(2) - 1/2 > 0$, preventing a common weighted flow. Let $M = \{i : v_i = \max_j v_j\}$, a nonempty proper subset. Subtracting $\max_j v_j$ before normalizing shows

$$\begin{aligned} \lim_{\tau\downarrow 0} p^{\mathrm{D}}_{w,\tau}(i) &= \lim_{\tau\downarrow 0} \frac{\rho_i e^{(v_i - \max_j v_j)/\tau}}{\sum_j \rho_j e^{(v_j - \max_u v_u)/\tau}} = \frac{\rho_i \mathbf{1}\{i \in M\}}{\sum_{j\in M}\rho_j}, \\ \lim_{\tau\uparrow\infty} p^{\mathrm{D}}_{w,\tau}(i) &= \rho_i. \end{aligned}$$

The first limit has zero mass outside M, while the target is strictly positive everywhere. The second is also different from the target, because μ^0 is nonconstant and β is finite. KL to the fixed positive target is continuous on the closed simplex, including zero coordinates, since $x \log x \to 0$ as $x \downarrow 0$. Reparameterize temperature by $t = \tau/(1+\tau) \in (0,1)$ and use the two limits to extend its regret continuously to $[0,1]$. Both endpoint values are strictly positive. If (i) fails, every interior value is strictly positive as well. The extended continuous function then attains a strictly positive minimum on the compact interval. This proves the contrapositive of (ii)⇒(i), and the stated positive floor.

The unequal-short-edge example. For (70),

$$\sum_i w_i = \frac{3}{2n} + \frac{1}{2n} + \frac{n-2}{n} = 1, \qquad \frac{1}{3n} \le w_i \le \frac{3}{n}.$$

The first two edges are short edges of the same block but have different weights, so (iii) fails. For the equiprobable two-point law with $0 < \epsilon \le \bar{\epsilon}$, the common-law hypotheses hold with $h = \epsilon$ and $V = \epsilon^2 > 0$; in particular $g_2 > 2b$. The preceding equivalence gives strictly positive best-temperature regret at $z = 0$. If this value is denoted $\delta_w > 0$ for a fixed experiment, then for every $\xi \ge 0$,

$$\sup_{\|z\|_\infty \le \xi} \inf_{\tau > 0} \mathrm{Reg}_{\mu(z),\beta}(p^{\mathrm{D}}_{w,\tau}(z)) \ge \delta_w,$$

because $z = 0$ is an allowed perturbation. At fixed positive ϵ and fixed n, $\min\{n^2\epsilon^4\xi^2, 1\} \to 0$ as $\xi \downarrow 0$. Thus no constant upper bound of that order can hold for this individual weighting throughout arbitrarily small perturbation radii.

Repeating the imbalance yields a size-independent floor. Now use (71) on every block and take $\rho_i = 1/n$. The weights are normalized because each block contributes $3/n$ and there are $J = n/3$ blocks; they also belong to $\mathcal{W}_3(n)$. Both the weights and Q are invariant under a rotation by three classes. Rotating the unique centered logistic minimizer therefore leaves it unchanged. Write this three-periodic score as $v_{3k+t} = s_t$, where $\sum_{t=0}^{2} s_t = 0$. Let $\alpha = (3/2, 1/2, 1)$ and $(Q_0, Q_1, Q_2) = (F_\nu(1), F_\nu(1), F_\nu(-2))$. The objective restricted to these scores is, with $s_3 = s_0$,

$$\sum_{k=0}^{J-1}\sum_{t=0}^{2}\frac{\alpha_t}{n}\{\log(1+e^{s_{t+1}-s_t}) - Q_t(s_{t+1}-s_t)\}$$
$$= \frac{1}{3}\sum_{t=0}^{2}\alpha_t\{\log(1+e^{s_{t+1}-s_t}) - Q_t(s_{t+1}-s_t)\}.$$

This strictly convex centered three-variable problem is independent of J, so its unique vector s is the same for every graph size. For $q_\tau(t) = e^{s_t/\tau}/\sum_u e^{s_u/\tau}$ and $q_\beta^{\mathrm{target}}(t) = e^{t/\beta}/\sum_u e^{u/\beta}$, the full fitted and target policies are respectively $q_\tau(t)/J$ and $q_\beta^{\mathrm{target}}(t)/J$. Hence

$$\mathrm{Reg}_\beta(p^{\mathrm{D}}_{w,\tau}) = \beta\sum_{k=0}^{J-1}\sum_{t=0}^{2}\frac{q_\tau(t)}{J}\log\frac{q_\tau(t)/J}{q_\beta^{\mathrm{target}}(t)/J}$$
$$= \beta\sum_{t=0}^{2} q_\tau(t)\log\frac{q_\tau(t)}{q_\beta^{\mathrm{target}}(t)}.$$

This identity holds at every temperature and is independent of J. At $J = 2$, condition (iii) fails because the two short weights differ, so the infimum of this same expression is strictly positive by the preceding proof. Call it $\delta_{\nu,\beta}$. Taking infima in the identity proves (74) for all $J \geq 2$. □

D.7 Proof of corollary 3

We first record the elementary limiting step. If $a_n \geq 0$, then

$$\min\{a_n^2, 1\} \longrightarrow 0 \quad\Longleftrightarrow\quad a_n \longrightarrow 0. \tag{127}$$

The reverse implication follows from continuity. For the forward implication, if a_n does not converge to zero, there are $\delta > 0$ and a subsequence on which $a_n \geq \delta$. The minimum is then at least $\min\{\delta^2, 1\} > 0$ on that subsequence, a contradiction.

For paths, $J = n/3$ and theorem 1 gives

$$\frac{c_L}{9}\min\{n^2V_n^2, 1\} \leq c_L\min\{n^2V_n^2/9, 1\} \leq \mathcal{R}^{\mathsf{env}}_{1,n} \leq c_U\min\{n^2V_n^2, 1\}.$$

The constants are independent of the changing law and of n. Apply (127) with $a_n = nV_n$. For the two-point example, $h_n = n^{-1/2} \leq h_*$ eventually and $V_n = 1/n$, so the displayed lower and upper bounds are respectively $c_L/9$ and c_U.

For cycles, the lower bound for the best shared weighting and the zero adaptive value are theorem 3. Equal weights belong to $\mathcal{W}_\kappa(n)$; taking an infimum over all shared weights therefore gives $\mathcal{R}^{\mathrm{fixed}}_\kappa \leq \mathcal{E}^{\mathsf{unif}}_n$. The equal-weight upper bound used in that theorem is uniform over the whole perturbation ball, so $\mathcal{E}^{\mathsf{unif}}_n \leq C_\kappa\Psi_n$. This proves (79). Applying (127) with $a_n = n\epsilon_n^2\xi_n$ proves both equivalences, including sequences with zero parameters. For an arbitrary $w_n \in \mathcal{W}_\kappa(n)$, its worst-case regret is bounded below by $c_\kappa\Psi_n$ by (45). Vanishing regret consequently forces $\Psi_n \to 0$, hence $n\epsilon_n^2\xi_n \to 0$. No per-weight upper bound is used or inferred.

For failure of the converse at arbitrary weights, fix $0 < \epsilon \leq \epsilon_\kappa$ and take $\xi_n = n^{-2}$, which lies in the required positive neighborhood for all sufficiently large $n = 6L$. Then $n\epsilon^2\xi_n = \epsilon^2/n \to 0$, and the equal-weight upper bound gives

$$\mathcal{E}^{\mathsf{unif}}_n \leq C_\kappa n^2\epsilon^4 n^{-4} = C_\kappa\epsilon^4 n^{-2}.$$

The periodically imbalanced weights belong to $\mathcal{W}_3(n) \subseteq \mathcal{W}_\kappa(n)$. Their perturbation ball contains the unperturbed experiment, whose best-temperature regret is $\delta_{\nu,1} > 0$ independently of n by (74),

for $\nu = \frac{1}{2}\delta_{-\epsilon} + \frac{1}{2}\delta_{\epsilon}$. Their worst-case regret is consequently bounded below by this same positive constant, proving the asserted separation.

Finally, applying proposition 2 with $\beta = 1$ and the common support bound h gives (78) for every n. The coefficient $\mathcal{B}(h)^2/8$ is fixed, so $D_{G_n} V_n \to 0$ makes this upper bound vanish. With shocks $\pm n^{-1/4}$, $V_n = n^{-1/2}$ and their support is contained in $[-h_*, h_*]$ eventually. On the periodic path, theorem 1 bounds regret above and below by fixed positive constants because $n^2 V_n^2 = n \to \infty$. On a star $D_{G_n} = 2$ and the tree upper bound, with support radius h_*, is $\mathcal{B}(h_*)^2/(2n)$. Both graphs have the same n classes and $n - 1$ edges. This proves the final comparison. □

D.8 Proof of proposition 1

We prove the statement for arbitrary $Q \in (0,1)^E$ and arbitrary finite f_β, without using the way in which representative sampling produces these quantities. Thus class-dependent representative laws are allowed. The graph and its positive weights are fixed throughout. Let

$$s = cf_\beta, \qquad v = v_w^{\mathrm{D}}, \qquad x = v - s, \qquad D = \max_i x_i - \min_i x_i, \qquad M = M_c = \operatorname{osc}(s).$$

Centering s is unnecessary because only differences and its oscillation will be used. Existence of the finite centered optimizer v follows from lemma B.1. Its full first-order equation and the target residual divergence are

$$\mathsf{B}^\top \operatorname{diag}(w)\{\sigma(\mathsf{B}v) - Q\} = 0,$$
$$b_c = \mathsf{B}^\top \operatorname{diag}(w)\{\sigma(\mathsf{B}s) - Q\}.$$

The sigmoid is applied coordinatewise. Subtracting these equations gives

$$-b_c(i) = \sum_{j:\,\{i,j\}\in E} w_{ij}\{\sigma(s_i - s_j + x_i - x_j) - \sigma(s_i - s_j)\} \quad (i \in \mathcal{Z}). \tag{128}$$

To check that this identity does not depend on the chosen edge orientation, an edge $e = (i, j)$ contributes minus its sigmoid difference at vertex i. Since $\sigma(-t) = 1 - \sigma(t)$, this is exactly the summand in (128); the contribution at j has the opposite sign. An edge with the reverse orientation gives the same expression.

An elementary logistic comparison. For $M \geq 0$, $a \in [-M, M]$, and $t \geq 0$, we claim

$$\sigma(a + t) - \sigma(a) \geq \sigma(M + t) - \sigma(M),$$
$$\sigma(a) - \sigma(a - t) \geq \sigma(M + t) - \sigma(M). \tag{129}$$

Indeed,

$$\sigma'(u) = \frac{1}{2 + 2\cosh u}$$

is even and nonincreasing in $|u|$. For $q \geq 0$, $|a + q| \leq M + q$ and $|a - q| \leq M + q$. Therefore

$$\sigma(a + t) - \sigma(a) = \int_0^t \sigma'(a + q)\, dq \geq \int_0^t \sigma'(M + q)\, dq = \sigma(M + t) - \sigma(M),$$
$$\sigma(a) - \sigma(a - t) = \int_0^t \sigma'(a - q)\, dq \geq \int_0^t \sigma'(M + q)\, dq = \sigma(M + t) - \sigma(M).$$

Coverage prevents an extreme error from being isolated. If $D = 0$, both asserted bounds hold. Suppose $D > 0$. Let i_+ and i_- be vertices attaining the maximum and minimum of x, and put

$$m_x = (\max_i x_i + \min_i x_i)/2, \qquad S_- = \{j : x_j \leq m_x\}, \qquad S_+ = \{j : x_j > m_x\}.$$

The two sets partition the n vertices, so at least one of them has cardinality at least $\lceil n/2 \rceil$. If $|S_-| \geq \lceil n/2 \rceil$, then $i_+ \notin S_-$ and $x_{i_+} - x_j \geq D/2$ for every $j \in S_-$. All summands in (128) at i_+

are nonnegative, including those with $j \notin S_-$. Monotonicity of the sigmoid and (129) imply

$$\begin{aligned}
\|b_c\|_\infty &\geq -b_c(i_+) \\
&= \sum_{j:\,\{i_+,j\}\in E} w_{i_+j}\{\sigma(s_{i_+} - s_j + x_{i_+} - x_j) - \sigma(s_{i_+} - s_j)\} \\
&\geq \sum_{j\in S_-:\,\{i_+,j\}\in E} w_{i_+j}\{\sigma(s_{i_+} - s_j + D/2) - \sigma(s_{i_+} - s_j)\} \\
&\geq \left(\sum_{j\in S_-:\,\{i_+,j\}\in E} w_{i_+j}\right)\{\sigma(M + D/2) - \sigma(M)\} \\
&\geq \omega_G\{\sigma(M + D/2) - \sigma(M)\}.
\end{aligned}$$

The last inequality is exactly the definition of ω_G. If $|S_+| \geq \lceil n/2\rceil$, use i_- instead. Now $i_- \notin S_+$ and $x_j - x_{i_-} > D/2$ for $j \in S_+$. All summands at i_- are nonpositive. Taking their negatives and using the second inequality in (129) gives

$$\begin{aligned}
\|b_c\|_\infty &\geq b_c(i_-) \\
&= \sum_{j:\,\{i_-,j\}\in E} w_{i_-j}\{\sigma(s_{i_-} - s_j) - \sigma(s_{i_-} - s_j + x_{i_-} - x_j)\} \\
&\geq \left(\sum_{j\in S_+:\,\{i_-,j\}\in E} w_{i_-j}\right)\{\sigma(M + D/2) - \sigma(M)\} \\
&\geq \omega_G\{\sigma(M + D/2) - \sigma(M)\}.
\end{aligned}$$

Consequently, in all cases,

$$\sigma(M + D/2) - \sigma(M) \leq \eta_c. \tag{130}$$

No bound on the unknown fitted score was assumed in this argument.

A nonlinear bound and its quadratic specialization. A slightly more general sufficient condition than the one in the proposition is $\eta_c < 1 - \sigma(M)$. Define

$$T_c = \operatorname{logit}\{\sigma(M) + \eta_c\} - M \geq 0.$$

The argument of the logit is then in $(0, 1)$. Inverting the monotone sigmoid in (130) yields

$$\operatorname{osc}(v - cf_\beta) \leq 2T_c, \qquad \operatorname{Reg}_\beta(p^{\mathrm{D}}_{w,c\beta}) \leq \frac{T_c^2}{2c^2\beta}. \tag{131}$$

For the second assertion, the fitted and target policies are $\mathsf{S}_\rho((v/c)/\beta)$ and $\mathsf{S}_\rho(f_\beta/\beta)$, respectively. Their score difference has oscillation D/c; lemma B.3 gives $\operatorname{Reg}_\beta \leq D^2/(8c^2\beta) \leq T_c^2/(2c^2\beta)$. This also proves exact agreement when $b_c = 0$.

Now assume $\eta_c \leq k_c/2$ with $k_c = \sigma'(M + 1)$. First,

$$1 - \sigma(M) > \sigma(M + 1) - \sigma(M) = \int_0^1 \sigma'(M + q)\, dq \geq k_c > \eta_c.$$

The nonlinear bound therefore applies and $0 \leq T_c < 1$. Because $\sigma(M + T_c) - \sigma(M) = \eta_c$,

$$\eta_c = \int_0^{T_c} \sigma'(M + q)\, dq \geq k_c T_c, \qquad T_c \leq \eta_c/k_c.$$

Substitution in (131) proves (18). Choosing $\tau = c\beta$ is allowed when taking an infimum over all positive temperatures. The proposition does not infer a lower bound for that infimum.

A concrete dense class and a variance-only condition. Suppose (19) holds and put $w_{\min} = \min_e w_e$, $w_{\max} = \max_e w_e$. For every i and $S \subseteq \mathcal{Z} \setminus \{i\}$ with $|S| \geq \lceil n/2\rceil$, at most $n - 1 - |S|$ neighbors of i lie outside S. Hence

$$\begin{aligned}
|N(i) \cap S| &\geq d_i - (n - 1 - |S|) \\
&\geq d_{\min} + 1 - n + \lceil n/2\rceil \\
&\geq (1/2 + \gamma)n - n + n/2 = \gamma n.
\end{aligned}$$

Multiplying by $w_{\min}$ and minimizing gives

$$\omega_G \geq \gamma n w_{\min} > 0. \tag{132}$$

Only this comparison-mass property is needed for the proof above.

For the common translated law, centering gives $m_z = \mu_z$ and the common log-MGF adds the same constant to each $f_\beta(z)$. Thus at $c = 1$, $\mathsf{B}f_\beta = \mathsf{B}\mu$ and $M_1 = M = \operatorname{osc}(\mu)$. Let U, U' be independent with this law and let $Z = U - U'$. Then

$$\mathbb{E}Z = 0, \qquad \mathbb{E}Z^2 = 2V.$$

For every real t, Z, Taylor's formula with its integral remainder is

$$\sigma(t+Z) = \sigma(t) + Z\sigma'(t) + Z^2 \int_0^1 (1-u)\sigma''(t+uZ)\,du.$$

The global bound $|\sigma''| \leq 1/4$ therefore gives

$$\begin{aligned} |F_\nu(t) - \sigma(t)| &= \left| \mathbb{E}\left[Z^2 \int_0^1 (1-u)\sigma''(t+uZ)\,du \right] \right| \\ &\leq \frac{1}{4}\,\mathbb{E}Z^2 \int_0^1 (1-u)\,du = V/4. \end{aligned} \tag{133}$$

This calculation uses finite variance but no bound on the support width, no minimum atom mass, and no symmetry of U. In the finite-response experiment the required moments and exponential moments exist. For each vertex,

$$|b_1(i)| \leq \sum_{e:\, i\in e} w_e |r_e(1)| \leq \frac{V}{4} \sum_{e:\, i\in e} w_e \leq \frac{(n-1)w_{\max}V}{4}.$$

Combining this with (132) and $w_{\max}/w_{\min} \leq \chi$ gives

$$\eta_1 \leq \frac{(n-1)w_{\max}V}{4\gamma n w_{\min}} \leq \frac{\chi V}{4\gamma}.$$

If $V \leq 2\gamma k_M/\chi$, this upper bound is at most $k_M/2$. Applying (18) at $c = 1$ yields

$$\operatorname{Reg}_\beta(p^{\mathrm{D}}_{w,\beta}) \leq \frac{1}{2\beta k_M^2}\left(\frac{\chi V}{4\gamma}\right)^2 = \frac{\chi^2 V^2}{32\gamma^2\beta k_M^2},$$

as asserted. For a complete graph, $d_{\min} = n - 1$ permits $\gamma = 1/2$ for every $n \geq 2$. Equal weights give $\chi = 1$, $\omega_G = \lceil n/2 \rceil / \binom{n}{2}$, and the stated $V^2/(8\beta k_M^2)$ bound follows.

Uniformity, retained heterogeneity, and graph sequences. The constants in (20) are uniform in n whenever γ is bounded below, χ and M are bounded above, and the target coefficient β is fixed. A sequence $V_n \to 0$ then eventually satisfies the positive variance threshold and has regret $O(V_n^2)$. For the repeated means of the path theorem, $M = 2$. With equiprobable shocks $U = \pm n^{-1/4}$, $V_n = n^{-1/2}$ and the dense-graph bound is $O(n^{-1})$. The path and star conclusions for the same fine experiment are proved in section D.7. Dense designs satisfying (19) have order n^2 comparison types by summing the degree condition; no equal-budget comparison to the sparse graphs is asserted.

To make the absence of a small-support requirement concrete, take any sequence $A_n \to \infty$ and $V_n \to 0$, and for all sufficiently large n put

$$\Pr(U = A_n) = \Pr(U = -A_n) = \frac{V_n}{2A_n^2}, \qquad \Pr(U = 0) = 1 - \frac{V_n}{A_n^2}.$$

This is a centered three-point law with variance exactly V_n and support width $2A_n \to \infty$. Each experiment still has finite rewards and exponential moments. The dense-graph guarantee is $O(V_n^2)$, although the common-support hypothesis of the path theorem does not apply to this sequence. Its possibly large common log-MGF still cancels from the target; cancellation need not hold for class-dependent laws. For those laws, the first part of the proposition applies directly to the true f_β and residual divergence instead. □

E FURTHER COMPARISON WITH RELATED LITERATURE

What is held fixed? The distinctions in Section 2 concern the statistical experiment, optimization criterion, and benchmark, rather than the use of the name DPO. Preference-based reward learning (Christiano et al., 2017) and human-feedback instruction tuning (Ouyang et al., 2022) provide the application background. The BT model (Bradley & Terry, 1952) and DPO reward/policy change of variables (Rafailov et al., 2023) supply our optimizer. These references do not justify assuming that marginalizing representatives preserves a BT link; that assumption is exactly what our examples test.

Changing the objective changes the question. Gheshlaghi Azar et al. (2024) formulate ΨPO directly from pairwise preferences, with RLHF, DPO, and identity preference optimization as special cases. Munos et al. (2024) learn a preference model and optimize a regularized two-player game. These are substantive alternatives to a scalar-reward reduction. A Nash policy or an identity-objective optimizer, however, need not equal the class marginal of our fixed fine Gibbs policy. Our lower bounds do not rule out such alternative goals, nor do they establish their failure. Likewise, the proper-loss quantifier in Theorem 1 retains the stipulated link and target; it is not a quantifier over every possible preference-learning objective.

More data and a better comparison design solve different problems. Shah et al. (2016) give topology-dependent minimax bounds for latent score estimation in parametric comparison models; Li et al. (2022) study the BTL maximum-likelihood estimator on general graphs. These are finite-sample guarantees under a specified statistical model. In our experiment every edge probability is already known. More observations of those same probabilities cannot remove the population discrepancy. Our dense-coverage proposition therefore complements statistical design bounds: it tests a deterministic population residual, under its own coverage and boundedness assumptions. It does not replace sample-complexity analysis or imply that dense comparison graphs are necessary. The tree certificate and star example provide sparse positive cases, while the cycle lower bounds retain their stated mean structure. The path extension allows arbitrary small mean offsets with a radius uniform in graph size.

Meaning-level probability is not a unique optimization target. Kuhn et al. (2023) introduce semantic entropy by aggregating linguistically equivalent generations; Farquhar et al. (2024) use meaning-level uncertainty to detect confabulations. This motivates studying class probabilities, but supplies neither our reward model nor our representative-sampling interface. We assume a known partition and do not analyze the errors of a semantic classifier. Semantic calibration (Nakkiran et al., 2025) and preservation of a fine-optimal class allocation are different requirements. In particular, a fixed reference-conditional renderer gives the mean-reward router target, whereas optimizing the fine response distribution first gives a log-MGF target. The targets coincide for common translated laws, not in general. Our residual decomposition records both the comparison distortion and this target shift. Thus semantic aggregation is the motivating operation; the strongest sharp rates isolate one controlled mechanism within it.

F UNIFORM EXTENSIONS AND INTERPRETATION OF REGRET

This appendix extends the canonical examples without changing the comparison experiment or optimizer. The first result permits arbitrary small changes to every path mean, including changes that remove all exactly equal means. The next result keeps the fine reward fixed and changes the target coefficient. The last result removes the two-point restriction from the cycle rates. All statements concern population probabilities.

F.1 PATH AMPLIFICATION WITHOUT EXACT REPETITION

Corollary 4 (Uniform amplification under arbitrary mean offsets)**.** *There are universal $h_{\mathrm{P}}, \zeta_{\mathrm{P}} > 0$ such that the following holds. In the path experiment of theorem 1, replace the means by $\mu_{k,t} = t + z_{k,t}$, with arbitrary $\|z\|_\infty \le \zeta_{\mathrm{P}}$. For every fixed $\beta > 0$ there are $c_\beta, C_\beta > 0$ for which*

$$c_\beta \min\{J^2V^2, 1\} \le \mathcal{R}_\beta^{\mathsf{env}} \le C_\beta \min\{J^2V^2, 1\}, \qquad J \ge 2, \tag{134}$$

for every finite centered common law on $[-h, h]$, $h \le h_{\mathrm{P}}$, with variance $V \ge 0$. Constants and radii are independent of J, atom masses, support cardinality, V/h^2, and the individual offsets. The envelope still allows every positive weighting, every loss in $\mathfrak{L}$, and every positive temperature. Every comparison is fitted exactly. No equality of target means across blocks is required.

Proof. We give explicit sufficient constants. Put

$$B_0 = \mathcal{B}(1/4), \quad k_4 = \sigma'(4), \quad C_g = \frac{1}{2k_4} + B_0, \quad C_e = \frac{6C_g}{K},$$

and take

$$h_{\mathrm{P}} = \min\{h_*, (8B_0)^{-1/2}\}, \qquad \zeta_{\mathrm{P}} = \min\{1/128, (32C_e)^{-1}\}. \tag{135}$$

These are conservative sufficient radii, not estimates of an operating range. Write $\zeta = \|z\|_\infty$ and suppose first $V > 0$. All norms and inner products below are under the uniform measure on $3J$ vertices. A tilde denotes subtraction of that mean.

The perturbation changes the accumulated error by order $JV\zeta$. Taylor's formula and $\mathbb{E}(U - U')^2 = 2V$ give $|F_\nu - \sigma| \le V/4$ and $|F'_\nu - \sigma'| \le V/4$. For $|a| \le 3$ and $h \le 1/4$, $F_\nu(a)$ lies between $\sigma(-4)$ and $\sigma(4)$. Since $x(1-x)$ is one-Lipschitz on $[0,1]$,

$$\begin{aligned} |g'_\nu(a) - 1| &= \frac{|F'_\nu(a) - F_\nu(a)(1 - F_\nu(a))|}{F_\nu(a)(1 - F_\nu(a))} \\ &\le \frac{|F'_\nu(a) - \sigma'(a)| + |\sigma(a)(1-\sigma(a)) - F_\nu(a)(1 - F_\nu(a))|}{k_4} \le \frac{V}{2k_4}. \end{aligned}$$

By lemma B.5, $|b - 1| \le B_0 V$, so $|g'_\nu(a) - b| \le C_g V$ on this interval. Exact tree fitting gives $v_{i+1} - v_i = g_\nu(\mu_{i+1} - \mu_i)$. Choose the additive constant so that $v_0 = b\mu_0$, and let $d = d_\nu$. The score can be written exactly as

$$v_{k,t} = b\mu_{k,t} - kd + e_{k,t}, \qquad e_{0,0} = 0.$$

If $\Delta_i^0 \in \{1, -2\}$ is the baseline increment, then

$$\begin{aligned} e_{i+1} - e_i &= g_\nu(\Delta_i^0 + z_{i+1} - z_i) - g_\nu(\Delta_i^0) - b(z_{i+1} - z_i) \\ &= \int_0^{z_{i+1} - z_i} \{g'_\nu(\Delta_i^0 + a) - b\}\, da. \end{aligned}$$

Every intervening argument lies in $[-3, 3]$. Thus $|e_{i+1} - e_i| \le 2C_g V\zeta$ and $\mathrm{osc}(e) \le 6JC_gV\zeta$. Define

$$D = Jd, \quad T_{k,t} = t - 1, \quad W_{k,t} = \frac{k - (J-1)/2}{J}, \quad Z = \widetilde{z}.$$

Using $d \ge KV/2$ from (57), we obtain

$$\widetilde{\mu} = T + Z, \quad \widetilde{v} = b\widetilde{\mu} - DW + \widetilde{e}, \quad \|Z\|_2 \le \zeta, \quad \|\widetilde{e}\|_2 \le C_e D\zeta, \quad \mathrm{osc}(e) \le 2C_eD\zeta. \tag{136}$$

The norm bound uses $\mathrm{Var}(e) \le \mathrm{osc}(e)^2/4$. This is the needed cancellation: bounding the raw score perturbation alone would lose the factor V.

Temperatures outside a fixed multiple of β have constant regret. Each short mean increment belongs to $[1-2\zeta, 1+2\zeta]$. Since $B_0 V \le 1/8$ and $\zeta \le 1/128$, its fitted increment lies in $[3/4, 5/4]$. The target mean difference from the first to the last vertex of a block lies in $[15/8, 17/8]$. Condition on these two vertices within a block. Their fitted last-vertex probability is $\sigma((v_{k,2} - v_{k,0})/\tau)$, and the target probability is $\sigma((\mu_{k,2} - \mu_{k,0})/\beta)$. The fitted mass of this pair within its block is at least $1/2$, since the middle score lies between the endpoint scores. Put

$$d_T(\beta) = \min\{\sigma(15/(8\beta)) - \sigma(5/(8\beta)),\, \sigma(6/\beta) - \sigma(17/(8\beta))\} > 0.$$

When $\tau \notin (\beta/4, 4\beta)$, the two conditional probabilities differ by at least $d_T(\beta)$. The KL chain rule, followed by binary Pinsker and then summation over fitted block masses, gives

$$\mathrm{Reg}_\beta(p^{\mathrm{D}}_{w,\tau}) \ge \beta d_T(\beta)^2. \tag{137}$$

This argument uses no repetition of means or block masses.

Bounded accumulated drift: a score component cannot be calibrated away. Now let $\tau = \beta s$ with $1/4 \le s \le 4$, and first suppose $D \le 1$. The baseline vectors obey

$$\langle T, W\rangle = 0, \qquad \|T\|_2^2 = 2/3, \qquad \|W\|_2^2 = (J^2-1)/(12J^2) \in [1/16, 1/12].$$

Consequently

$$\frac{|\langle W, \widetilde{\mu}\rangle|}{\|W\|_2 \|\widetilde{\mu}\|_2} \le \frac{\zeta}{\sqrt{2/3} - \zeta} < \frac{1}{2}, \qquad \mathrm{dist}(W, \mathrm{span}\{\widetilde{\mu}\}) \ge \frac{1}{8}.$$

Project $\widetilde{v}/s - \widetilde{\mu}$ orthogonally to $\widetilde{\mu}$. By (136),

$$\begin{aligned}\|\widetilde{v}/s - \widetilde{\mu}\|_2 &\ge \frac{D}{s}\left\{\mathrm{dist}(W, \mathrm{span}\{\widetilde{\mu}\}) - C_e\zeta\right\} \\ &\ge \frac{D}{16s} \ge \frac{D}{64}.\end{aligned}$$

Here $\operatorname{osc}\mu \le 3$ and $\operatorname{osc} v \le (3/2)3 + D + 2C_e D\zeta < 6$. For every logit vector on the line segment from $v/(\beta s)$ to μ/β, its oscillation is at most $24/\beta$. Its softmax probabilities are therefore at least $e^{-24/\beta}$ times uniform. For any vector a, minimizing over additive constants gives $\mathrm{Var}_p(a) \ge e^{-24/\beta}\,\mathrm{Var}_{\mathsf{unif}}(a)$. Apply the integral KL identity in lemma B.3 to obtain

$$\mathrm{Reg}_\beta(p^{\mathrm{D}}_{w,\tau}) \ge \frac{e^{-24/\beta}}{2\beta}\|\widetilde{v}/s - \widetilde{\mu}\|_2^2 \ge \frac{e^{-24/\beta}}{8192\beta} D^2. \tag{138}$$

Large accumulated drift: separated groups have the wrong mass ratio. Suppose $D \ge 1$. Let $m = \lceil J/4\rceil$, and choose the first m and last m vertices of type $t = 0$, calling these sets A and B. They are disjoint, and every index in B exceeds every index in A by at least $J - 2m + 1 \ge J/4$. From (136),

$$\min_A v - \max_B v \ge D/4 - 3\zeta - 2C_e D\zeta \ge D/8.$$

The sets have equal cardinality. Thus the fitted conditional probability of A within $S = A \cup B$ is at least $\sigma(D/(8\tau))$, hence at least $\sigma(1/(32\beta))$. The corresponding target probability is at most $\sigma(2\zeta/\beta) \le \sigma(1/(64\beta))$. Set

$$d_A(\beta) = \sigma(1/(32\beta)) - \sigma(1/(64\beta)) > 0, \qquad q_A(\beta) = e^{-3/\beta}/6.$$

Since $|S|/(3J) \ge 1/6$, its target mass is at least $q_A(\beta)$. If the fitted mass of S is below $q_A(\beta)/2$, binary Pinsker on S gives regret at least $\beta q_A(\beta)^2/2$. Otherwise conditional Pinsker within S gives $\beta\, p(S)\, 2d_A(\beta)^2 \ge \beta q_A(\beta) d_A(\beta)^2$. Thus

$$\mathrm{Reg}_\beta(p^{\mathrm{D}}_{w,\tau}) \ge \beta \min\{q_A(\beta)^2/2,\, q_A(\beta) d_A(\beta)^2\}. \tag{139}$$

Take the minimum of the positive constants in (137)–(139), and multiply it by $\min\{K^2/4, 1\}$. This gives c_β because $D \ge KJV/2$. For the upper bound, proposition 2 and $D_G < 3J$, $\operatorname{osc}\mu \le 3$ give

$$\inf_{\tau>0} \mathrm{Reg}_\beta(p^{\mathrm{D}}_{w,\tau}) \le \frac{1}{8\beta}\min\{9J^2 B_0^2 V^2, 9\},$$

so $C_\beta = 9\max\{B_0^2, 1\}/(8\beta)$ suffices. Every allowed loss and positive weighting has the same tree score; the bounds consequently hold for the full envelope. At $V = 0$ that score is μ up to a constant, and $\tau = \beta$ gives zero regret. □

F.2 CHANGING THE FIXED TARGET COEFFICIENT

Proposition 5 (Every fixed positive target coefficient)**.** *Keep the fine rewards and comparison probabilities fixed. The rates in theorem 1, theorem 2(ii), and theorem 3 hold with target coefficient $\beta > 0$ in place of one, with constants allowed to depend on that fixed β (and on κ in theorem 3). The score-side local radii can be kept independent of β. Exact cycle recovery uses $\tau = c_\epsilon\beta$; the path calibration uses $\tau = b\beta$. No uniform claim is made as $\beta \downarrow 0$ or $\beta \uparrow \infty$.*

Proof. The BT minimizing score depends on G, Q, w, not on the target coefficient or the output temperature. Changing β therefore leaves every score, root, and score-separation estimate unchanged. We give the necessary changes in converting these estimates to regret.

For the path, direct normalization gives

$$\mathrm{Reg}_\beta(p^{\mathrm{D}}_{w,\tau}) = \beta\{G_J(\eta s) + \mathrm{KL}(q_s\|q_{1/\beta})\}, \qquad s = b/\tau. \tag{140}$$

For $s \geq 1/(2\beta)$, the first term is at least $\beta e^{-1}\min\{J^2\eta^2/(4\beta^2), 1\}/32$. For $s \leq 1/(2\beta)$, condition q_s on $\{0, 2\}$. Its mass there is at least $1/2$, and its last-coordinate probability is at most $\sigma(1/\beta)$, versus $\sigma(2/\beta)$ for the target. Thus the second term is at least $\beta\{\sigma(2/\beta) - \sigma(1/\beta)\}^2 > 0$. The upper bounds follow at $s = 1/\beta$ and in the limit $s \downarrow 0$. This proves the path assertion; corollary 4 also proves it without exact repetition.

For theorem 2(ii), replace the target region probabilities in (102) by probabilities proportional to $Z(a_+/\beta), Z(a_-/\beta)$, replace the two target type laws by $q_{a_+/\beta}, q_{a_-/\beta}$, and multiply the entire identity by β. The block-index target remains uniform. For $\tau \leq 8\beta$,

$$\beta G_L(\lambda/\tau) \geq \frac{\beta e^{-1}}{32}\min\{L^2\lambda^2/(64\beta^2), 1\}.$$

For $\tau \geq 8\beta$, $s_\pm = x_\pm/\tau \leq 1/(4\beta)$ and $a_\pm/\beta \in [1/(2\beta), 2/\beta]$. Conditioning each type law on $\{0, 2\}$ gives a KL lower bound

$$d_C(\beta)^2, \quad d_C(\beta) = \min_{x\in[1/(2\beta),2/\beta]}\{\sigma(2x) - \sigma(x)\} > 0.$$

Continuity on this compact interval makes the minimum positive. The fitted conditional type weights sum to one. These two temperature ranges prove the lower bound; theorem 2(i) already supplies the upper bound for every β.

For theorem 3, the interval in (118) becomes $(\beta/4, 4\beta)$, and its lower bound becomes $\beta d_T(\beta)^2$, with d_T as in (137). The target anchor mass is at least

$$q_0(\beta) = \{1 + e^{17/(16\beta)} + e^{17/(8\beta)}\}^{-1}.$$

The alternative for insufficient fitted anchor mass is now $\beta q_0(\beta)^2/2$. Otherwise the conditional anchor KL is at most $2R_a/(\beta q_0(\beta))$. In the exceptional-set step choose $b = G/(32\beta)$, where G is the unchanged score separation in (117). A good anchor has $|x^a_k - C_a| \leq 4\beta b$, so two anchors good in both environments have a difference in Δ at most $16\beta b = G/2$. The same contradiction on A, B applies. The union of exceptional sets is at most $8eR/(\beta q_0(\beta)\min\{b^2, 1\})$, hence

$$R \geq \frac{\beta q_0(\beta)\rho_\kappa}{8e}\min\{(G/(32\beta))^2, 1\}. \tag{141}$$

Together with the two constant-regret alternatives and $G = \gamma n\epsilon^2\xi/8$, this is the asserted lower rate. All temperature choices remain independent between environments. The upper bound and the zero adaptive value follow at their existing calibrations multiplied by β. None of these changes shrinks the score-side neighborhoods. □

Expected reward and reference-relative allocation. For any effective reward f_β,

$$\mathrm{Reg}_\beta(p) = \mathbb{E}_{p^{\mathrm{F}}_\beta} f_\beta - \mathbb{E}_p f_\beta + \beta\{\mathrm{KL}(p\|\rho) - \mathrm{KL}(p^{\mathrm{F}}_\beta\|\rho)\}. \tag{142}$$

Indeed the Gibbs variational identity gives $\beta\log\sum_z \rho_z e^{f_\beta(z)/\beta} = \mathbb{E}_{p^{\mathrm{F}}_\beta} f_\beta - \beta\,\mathrm{KL}(p^{\mathrm{F}}_\beta\|\rho)$; subtract $\mathbb{E}_p f_\beta - \beta\,\mathrm{KL}(p\|\rho)$. For a common translated law, $f_\beta = \mu +$ constant, so f_β may be replaced by μ in the expectation difference. In the unperturbed path at $\tau = b\beta$, the two policies are $P_{\eta/\beta,J} \otimes q_{1/\beta}$ and $\mathrm{Unif}_J \otimes q_{1/\beta}$. Their expected class means coincide because both type marginals are $q_{1/\beta}$. Their excess reference KL is $G_J(\eta/\beta)$, so all regret here is regularization cost. A positive regret lower bound alone gives neither a positive expected-reward deficit nor a lower bound on a task-specific notion of answer quality.

F.3 CYCLE RATES FOR EVERY SMALL-SUPPORT COMMON LAW

Corollary 5 (Removing the two-point restriction)**.** *Replace the two-point law in the cycle experiment by any finite centered common law ν supported on $[-h, h]$, of variance V, and use its link F_ν and calibration c_ν from* (29)*. The following extensions hold.*

(i) If $h \le \bar{\epsilon}$ and $\xi \le 1/16$, the score and regret upper bounds in theorem 2(i) hold with ϵ^2 replaced by V, including their stated constants. If also $\xi \le \xi_{\mathrm{bal}}$, instancewise recovery holds with weights in $\mathcal{W}_3(n)$ at $\tau = c_\nu \beta$.

(ii) There exist universal positive h_C, ξ_C such that the same two-region means as in theorem 2(ii) satisfy, for every fixed $\beta > 0$ and $h \le h_C$, $0 < \xi \le \xi_C$,

$$\inf_{\tau>0} \mathrm{Reg}_{\mu,\beta}(p^{\mathrm{D}}_{w^{\mathsf{unif}},\tau}) \asymp_\beta \min\{n^2V^2\xi^2, 1\}, \qquad n = 6L,$$

where $L \ge 2$ and $V > 0$.

(iii) For each fixed $\kappa \ge 3$ there is $h_\kappa > 0$ such that theorem 3 extends to these laws with $h \le h_\kappa$, $\xi \le \xi_{\mathrm{fixed}}$, and every fixed $\beta > 0$:

$$\mathcal{R}^{\mathrm{adapt}}_{\kappa,\beta}(n, \nu, \xi) = 0, \qquad \mathcal{R}^{\mathrm{fixed}}_{\kappa,\beta}(n, \nu, \xi) \asymp_{\kappa,\beta} \min\{n^2V^2\xi^2, 1\}. \tag{143}$$

Here the definitions are (40)–(41) *with target coefficient β and common law ν. The same paired environments, fixed independently of the weights and law, prove the lower bound for every shared weighting; equal weights give the upper bound.*

All constants and local radii are independent of n, support cardinality, atom masses, symmetry of ν, and V/h^2. The lower-bound radii are existence radii. The comparison graph and lower-bound mean arrangements remain those of the original cycle constructions.

Proof. The crucial issue is uniformity when V is arbitrarily small relative to h^2. We supply a compact representation that retains V exactly, rather than replacing the remainder by an additive power of h.

An exact variance-factored family. For $V > 0$, put $D = U - U'$ and define a symmetric probability measure ω_ν on $[-2, 2]$ by

$$\int a(x)\, d\omega_\nu(x) = \frac{\mathbb{E}[D^2 a(D/h)]}{2V}.$$

It is a probability measure since $\mathbb{E}D^2 = 2V$; symmetry follows from exchangeability of U, U'. Let Ω be the compact space of all symmetric probability measures on $[-2, 2]$, with the weak topology. For $\omega \in \Omega$ define

$$\begin{aligned} \Phi_\omega(h, t) &= 2\int_{-2}^{2}\int_0^1 (1-a)\sigma''(t + ahx)\, da\, d\omega(x), \\ \mathsf{F}(u, h, \omega, t) &= \sigma(t) + u\Phi_\omega(h, t). \end{aligned} \tag{144}$$

Taylor's formula with integral remainder and $\mathbb{E}D = 0$ yield the exact identity

$$\begin{aligned} F_\nu(t) - \sigma(t) &= \mathbb{E}\left[D^2 \int_0^1 (1-a)\sigma''(t + aD)\, da\right] \\ &= V\Phi_{\omega_\nu}(h, t). \end{aligned}$$

Thus the physical link is $\mathsf{F}(V, h, \omega_\nu, t)$. The auxiliary parameter u can vary independently of h; intermediate values are used only to analyze the finite cycle equations. We have, for every h, ω,

$$\mathsf{F}(0, h, \omega, t) = \sigma(t), \qquad \partial_u \mathsf{F}(0, 0, \omega, t) = \sigma''(t), \qquad \mathsf{F}(u, h, \omega, -t) = 1 - \mathsf{F}(u, h, \omega, t).$$

The last identity uses oddness of σ'' and symmetry of ω.

For completeness, all derivative bounds needed below are uniform in ω. For any fixed integer $r \ge 0$, differentiating under the bounded integrals gives

$$\partial_t^r \Phi_\omega(h, t) = 2\int\int_0^1 (1-a)\sigma^{(r+2)}(t + ahx)\, da\, d\omega(x).$$

A second-order expansion in h, and $\int x\, d\omega(x) = 0$, imply

$$|\partial_t^r \Phi_\omega(h,t) - \sigma^{(r+2)}(t)| \le \frac{h^2}{3} \sup_{|y-t|\le 2|h|} |\sigma^{(r+4)}(y)|. \tag{145}$$

Indeed the factor is $2 \cdot \frac{1}{2} \cdot 4 \int_0^1 (1-a)a^2\, da = 1/3$. All finite mixed derivatives in h, t, u are bounded and jointly continuous in the weak parameter ω on compact domains. The same statements hold after the logits and their implicit roots are formed: choose u, h small enough that the logit arguments stay in one compact subinterval of $(0,1)$. At $u = 0$ they are ordinary logistic probabilities with bounded arguments, and their root derivatives are strictly positive. Uniform continuity and compactness give a common neighborhood. More explicitly, normalize each cycle equation by its number of blocks, and denote it by $\Gamma(u, j; p) = 0$, where p collects the remaining compact parameters. At $u = 0$, $j = 0$ is a root and $\Gamma_j(0, 0; p) \ge a > 0$ uniformly. On a smaller fixed admissible rectangle, bounded derivatives give $|\Gamma(u, 0; p)| \le C|u|$ and $\Gamma_j \ge a/2$. The trial values $j = \pm 4C|u|/a$ bracket a root; for sufficiently small $|u|$ they remain in that rectangle. Strict monotonicity makes the root unique and gives $|j| \le 4C|u|/a$. Implicit differentiation yields $j_u = -\Gamma_u/\Gamma_j$ and analogous formulas for the mean-perturbation and support parameters. Repeated differentiation expresses every derivative used below through bounded derivatives of Γ and powers of Γ_j^{-1}. They are therefore uniformly bounded and continuous, including in the weak measure parameters. This justifies differentiation at $u = 0$ without differentiating the law ν or dividing a bound by its smallest atom mass.

Upper bounds and instancewise recovery. The Taylor bounds $|F_\nu - \sigma|, |F_\nu' - \sigma'| \le V/4$ hold for every centered law. The support bound gives $1 - 2h \le c_\nu \le 1 + 2h$; for $h \le \bar\epsilon$ this interval is contained in $[7/8, 9/8]$. The calibration equation then gives $|c_\nu - 1| \le V/(6\kappa_3) < 4V$ exactly as in (92). The residual $R_\nu(t) = \sigma(c_\nu t) - F_\nu(t)$ satisfies

$$R_\nu(1) = R_\nu(-2) = j_0, \qquad |R_\nu'(t)| \le 17V/4 \quad (|t| \le 3).$$

Every mean perturbation therefore leaves edge errors $|e_i| \le 17V\xi/2$. Since $V \le h^2 \le \bar\epsilon^2$, the admissibility inequalities in (94) hold unchanged. They give the actual score oscillation bound with V in place of ϵ^2; the KL conversion and infinite-temperature cap do likewise.

Also $d_\nu \ge KV/2$ by lemma B.5. Integrating the same three-logit equation between its zero-flow value and j_0 gives $j_0 \ge K\kappa_3 V/6$. At $\xi \le \xi_{\mathrm{bal}}$ the errors are at most $j_0/2$. The weights proportional to $\{\sigma(c_\nu \Delta_i) - F_\nu(\Delta_i)\}^{-1}$ thus lie in $\mathcal{W}_3(n)$ and carry a constant circulation. Strict convexity proves exact recovery. This proves part (i).

Equal-weight lower bound: retaining both vanishing factors. Replace F_ϵ in (95) by $\mathsf{F}(u, h, \omega, \cdot)$, and write the resulting block function as $\mathsf{D}(a, u, h, \omega, j)$. Its cycle equation is

$$\mathsf{D}(1 + \xi/2, u, h, \omega, j) + \mathsf{D}(1 - \xi/2, u, h, \omega, j) = 0.$$

It is independent of L. At $u = h = 0$, the derivative in j is $2A(1) > 0$. The preceding compact-domain argument supplies a smooth root $j(u, h, \omega, \xi)$ uniformly in ω. Define $\lambda = -\mathsf{D}(1 + \xi/2, u, h, \omega, j)$. At $u = 0$, the root is zero and every block sum vanishes. At $\xi = 0$, each block sum is zero at the cycle root. Therefore

$$\lambda = u\xi H(u, h, \omega, \xi), \qquad H = \int_0^1 \int_0^1 \lambda_{u\xi}(au, h, \omega, b\xi)\, da\, db. \tag{146}$$

The function H is uniformly continuous. At $u = h = \xi = 0$, $\partial_u \mathsf{F} = \sigma''$, so the coefficient calculation in (97) gives, for every ω,

$$H(0, 0, \omega, 0) = \frac{1}{2}\{K'(1) - K(1)A'(1)/A(1)\} = C_*/2 > 0.$$

Uniform continuity now gives universal neighborhoods on which $C_*/4 \le H \le 3C_*/4$ and $0 < x_\pm \le 2$. Since the physical value is $u = V \le h^2$, restricting h to a universal positive radius keeps it in this neighborhood. Hence $C_* V\xi/4 \le \lambda \le 3C_* V\xi/4$, uniformly even when V/h^2 or ξ/V tends to zero. The exact conditional KL decomposition and the two temperature ranges in proposition 5 give part (ii).

Shared-weight lower bound: compactness now includes the law. Use the weighted-cycle equations of section C.4 with $f(\eta, d)$ replaced by $\mathsf{F}(u, h, \omega, d)$. The inverse weights still belong to $[\kappa^{-1}, \kappa]$; their six empirical measures and $\omega \in \Omega$ form one compact product space, independent of L. On it, $\mathsf{F} - \sigma = O(u)$ with all needed derivatives uniform. The root bracket, its positive derivative, and the $O(u)$ root bound follow by the same monotonicity estimates. The short-edge score error is consequently $O_\kappa(u)$.

Let Ψ be the half-cycle average block increment defined in that proof, now also depending on h, ω. At $u = 0$, ordinary BT gives $\Psi_r = 0$. At $u = h = 0$, $\mathsf{F}_u = \sigma''$, so the calculation (112)–(113) is identical for every ω and yields

$$\Psi_{ur}(0, 0, \omega, r) \le -1/40, \qquad |r| \le 1/320.$$

All six inverse-weight measures are included in this uniform inequality. Uniform continuity on their compact product with Ω supplies $u_\kappa, h'_\kappa > 0$ on which $\Psi_{ur} \le -1/80$. The mixed derivatives of each individual block increment are bounded there by a common M_κ. Integrating in u gives

$$\Psi_r \le -u/80, \qquad |\partial_r D_k| \le M_\kappa u.$$

Choose $h_\kappa \le \min\{\bar{\epsilon}, h'_\kappa, \sqrt{u_\kappa}\}$, shrinking it further so the uniform short-score error is at most $1/8$. At the physical value $u = V$, integrating between $r = -\xi/2$ and $r = \xi/2$ gives exactly (116) with $\delta = V\xi$. The same two sets of positive anchor fraction therefore have separation $G = \gamma n V \xi/8$. The all-temperature and anchor-KL arguments of proposition 5, especially (141), now prove the lower bound in part (iii). Part (i) gives both its upper bound and zero adaptive value. At $V = 0$, or at $\xi = 0$ with equal weights, exact recovery handles the zero-rate endpoints. This completes all parts. □