ONE PROPER SCORE FOR MANY DECISIONS: SHARP REGRET TRANSFER UNDER MISSPECIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Under model misspecification, minimizing different proper scoring rules can induce different downstream decisions. We characterize the smallest worst-case target-risk ratio achievable by minimizing one fixed proper score across populations and report classes, relative to the best predictor in each class. The target is the contextwise maximum of declared task regrets, augmented by a strictly proper baseline divergence. We show that the optimal risk ratio equals the smallest two-sided distortion between the target and a proper-score divergence. For finite-action tasks, decision boundaries force curvature charges that yield exact distortion m for an orthogonal family of m two-action tasks on K outcomes, with $m \leq K-1$. Independent shared components with common tie selections admit an exact covering formula that quantifies the benefit of shared geometry. Its optimal weights yield a baseline-augmented task-score sum that attains the optimum over all proper scores. We also derive exact binary and quadratic characterizations and a sharp local trade-off between smoothness and additive regret error at hard boundaries. Known-truth studies illustrate the distinction between worst-case guarantees and achieved projection risk.

1 INTRODUCTION

A probability forecast can be reused across decisions with different costs. When the true conditional distribution belongs to the report class, minimizing any strictly proper population score risk recovers it and supports every user's Bayes decision. Under misspecification, however, score choice determines which approximation is learned, and different proper scores can induce different actions (Elliott et al., 2016). We ask how well one proper score can protect several downstream decisions, relative to the best predictor available in the same class.

For outcomes in $[K]$, fix m tasks with declared Bayes decoders and deterministic tie selections. Let $R_j(p,q)$ be task j's regret when its decoder acts on report q and the truth is p. For positive weights λ_j, a strictly proper baseline divergence D_{F_0}, and $\varepsilon > 0$, define

$$T_\varepsilon(p,q) = \varepsilon D_{F_0}(p,q) + \max_{1\leq j\leq m} \lambda_j R_j(p,q). \tag{1}$$

The maximum protects the most adversely affected task in each context, rather than averaging over users. The baseline distinguishes probability reports even within regions where all task regrets vanish.

We choose the score for this task family before fixing a population P or report class $\mathcal{H}$. Write $\mathcal{R}_T(h) = \mathbb{E}_X T_\varepsilon(p_X, q_h(X))$, $\mathcal{R}_T^\star = \inf_{h\in\mathcal{H}} \mathcal{R}_T(h)$, and $h_S \in \arg\min_{h\in\mathcal{H}} \mathbb{E}_{X,Y} S(q_h(X), Y)$. The optimal uniform risk factor is

$$\inf_{S \text{ strictly proper}} \sup_{P,\mathcal{H}:\, \mathcal{R}_T^\star > 0} \frac{\mathcal{R}_T(h_S)}{\mathcal{R}_T^\star}. \tag{2}$$

Here P includes the context law and conditional truth. The supremum ranges over finite context spaces and report classes, with worst-case selection among score minimizers. The benchmark concerns fixed-score risk minimization, not arbitrary predictor-construction algorithms.

We show that this population risk problem has an exact pointwise characterization. Let $D_{F,f}$ be the conditional excess risk of a proper score with convex entropy F and selected subgradient f.

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.

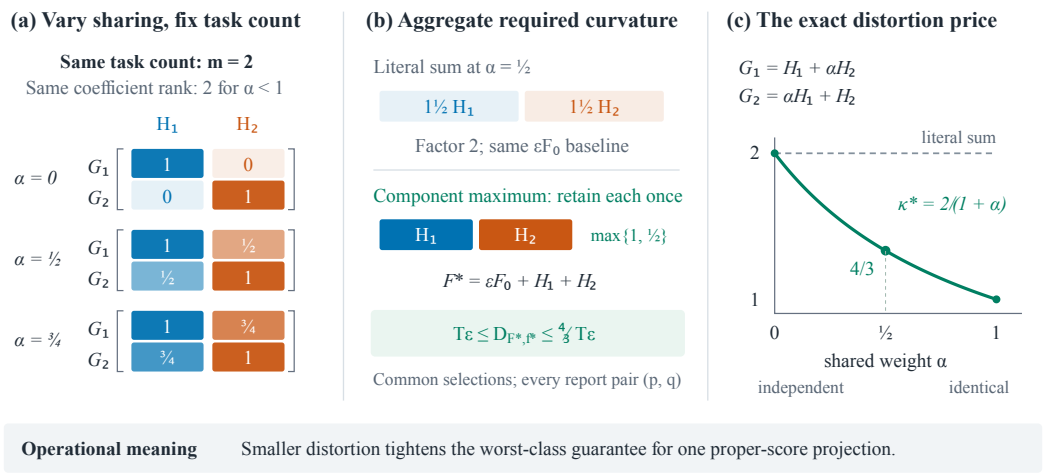


Figure 1: With independent components, common tie selections, and $0 \le \alpha \le 1$, the tasks $G_1 = H_1 + \alpha H_2$ and $G_2 = \alpha H_1 + H_2$ have optimal factor $2/(1+\alpha)$. Optimal weights give $(G_1 + G_2)/(1+\alpha) = H_1 + H_2$; the unweighted sum has factor two. Here $f^\star = \varepsilon \nabla F_0 + h_1 + h_2$; each baseline is counted once (Theorem 3.5).

Theorem 2.2 equates (2) with the least κ satisfying

$$T_\varepsilon(p,q) \le D_{F,f}(p,q) \le \kappa T_\varepsilon(p,q) \quad \text{for every } p, q \tag{3}$$

for some selected score. We call this factor the *proper-majorant distortion*, $\kappa^\star_\varepsilon$. The converse needs only two contexts and two candidate predictors, and remains sharp with a unique score minimizer. Thus the factor captures a possible mismatch between score-optimal and target-optimal projections even with unlimited data. Achieved ratios can be much smaller on a fixed population (Proposition 4.2).

The sharp factors depend on how the tasks share decision boundaries. For finite actions, regret grows linearly across a boundary, forcing a dominating entropy to carry a corresponding curvature charge. Charges on mutually singular facet supports must add, although the target takes a maximum over tasks. The universal factor m from a baseline-augmented task sum is sharp for an orthogonal family of m two-action tasks with $m \le K - 1$ (Theorem 3.3). For independent shared components with common tie selections, an exact covering problem yields optimal weights for a baseline-augmented sum of task scores. No other proper score achieves a smaller factor (Theorem 3.5). Figure 1 illustrates the gain from sharing. Binary tasks admit a pointwise least selected majorant, while quadratic tasks admit an exact ellipsoid characterization.

The boundary geometry also matters for optimization. Our hard task increments change score values across facets but leave ordinary within-cell gradients equal to the baseline's. We characterize a sharp local trade-off between entropy smoothness and additive regret error, and bound population approximation error in terms of report mass near the boundaries. Known-truth studies distinguish global guarantees, class-restricted estimates, and achieved projection risk.

2 Proper-majorant distortion

Let the outcome space be $[K]$, and let $\mathcal{P} \subseteq \Delta_K$ be compact and convex with nonempty relative interior. Task j has a finite action set A_j, bounded loss vectors $\ell_{ja} \in \mathbb{R}^K$, and a declared deterministic measurable Bayes decoder (for example, least-index tie breaking)

$$a_j(q) \in \arg\min_{a \in A_j} \langle q, \ell_{ja} \rangle.$$

Its negative Bayes envelope and selected subgradient are

$$G_j(p) = -\min_{a \in A_j} \langle p, \ell_{ja} \rangle, \qquad g_j(q) = -\ell_{j,a_j(q)}.$$

Thus G_j is convex piecewise linear and

$$R_j(p,q) = \langle p, \ell_{j,a_j(q)} \rangle - \min_a \langle p, \ell_{ja} \rangle = D_{G_j,g_j}(p,q).$$

A selected scoring pair consists of a finite convex entropy $F : \mathcal{P} \to \mathbb{R}$ and a selected $f(q) \in \partial F(q)$. With loss convention

$$S_{F,f}(q,y) = -F(q) - \langle f(q), e_y - q \rangle,$$

its conditional excess risk is

$$D_{F,f}(p,q) = F(p) - F(q) - \langle f(q), p - q \rangle.$$

Strict convexity of F gives strict propriety. Selections matter at kinks: the entropy alone does not specify either the score vector or a tied task action.

Throughout, F_0 is finite and strictly convex and is C^1 on a relative-open neighborhood of $\mathcal{P}$; $\varepsilon > 0$ and all task weights are positive. Subgradients and strict convexity are understood relative to $\mathrm{aff}(\mathcal{P})$.

Definition 2.1 (Proper-majorant distortion)**.** For T_ε in (1),

$$\kappa^\star_\varepsilon = \inf\{\kappa \geq 1 : \exists (F,f) \text{ satisfying (3)}\}.$$

For finite-valued nonnegative D, T that vanish on the diagonal, have a common zero set, and satisfy $T \not\equiv 0$, define their ratio range

$$a(D,T) = \inf_{T(p,q)>0} \frac{D(p,q)}{T(p,q)}, \qquad b(D,T) = \sup_{T(p,q)>0} \frac{D(p,q)}{T(p,q)}. \tag{4}$$

Let $\Gamma(D,T)$ be the supremum of

$$\frac{\mathbb{E}_X T(p_X, q_{h_D}(X))}{\min_{h \in \mathcal{H}} \mathbb{E}_X T(p_X, q_h(X))} \tag{5}$$

over finite context spaces, truth maps, context laws, and finite report classes with positive denominator, where h_D is chosen adversarially among exact minimizers of the D-risk. Finite contexts require no global measurability; set $\Gamma = 1$ when $T \equiv 0$.

Theorem 2.2 (Operational equivalence)**.** *If $0 < a(D,T) \leq b(D,T) < \infty$, then*

$$\Gamma(D,T) = \frac{b(D,T)}{a(D,T)}. \tag{6}$$

The same supremum is obtained if the score minimizer must be unique. Hence

$$\kappa^\star_\varepsilon = \inf_{(F,f) \text{ strictly proper}} \Gamma(D_{F,f}, T_\varepsilon), \tag{7}$$

with infinite factor when the lower ratio is zero or the upper ratio is infinite.

The upper bound follows by rescaling D by $1/a$ and comparing the two risk minimizers. For the converse, choose pairs with divergences D_-, D_+, target losses T_-, T_+, and ratios $r_- = D_-/T_-$, $r_+ = D_+/T_+$ near a, b. On two contexts, one hypothesis makes only the first error and the other only the second. Choosing context masses so that $\pi D_- = (1-\pi) D_+$ equalizes score risks, while the target-risk ratio is r_+/r_-. When $r_- < r_+$, a small mass perturbation makes the higher-target-risk hypothesis uniquely score-optimal. Taking limits gives the converse; Appendix C provides the full argument.

For finite-action tasks, a sum of task entropies gives a feasible score and a universal upper bound on this optimal risk factor.

Proposition 2.3 (Universal sum majorant)**.** *Let*

$$F_\Sigma = \varepsilon F_0 + \sum_{j=1}^m \lambda_j G_j, \qquad f_\Sigma = \varepsilon \nabla F_0 + \sum_{j=1}^m \lambda_j g_j.$$

Then $T_\varepsilon \leq D_{F_\Sigma, f_\Sigma} \leq m T_\varepsilon$, and hence $1 \leq \kappa^\star_\varepsilon \leq m$.

Indeed, writing $b_0 = \varepsilon D_{F_0}$ and $r_j = \lambda_j R_j$, the score divergence is $b_0 + \sum_j r_j$, whereas the target is $b_0 + \max_j r_j$. On the compact domain, finite infimal distortion is attained pointwise by a selected finite convex entropy, and the baseline makes it strictly convex (Theorem D.1, Appendix D). The concrete sums and joins below have Borel measurable selections: each combines the continuous baseline gradient with measurable task selections. We next determine when the factor m is unavoidable and how shared task geometry can reduce it.

3 SHARP AGGREGATION BOUNDS

We call piecewise-linear tasks and their kinks *hard*. The lower bounds use the following comparison between selected divergences.

Lemma 3.1 (Selected divergence comparison)**.** *For selected convex pairs (F, f), (G, g) and $\lambda \geq 0$,*

$$D_{F,f} \geq \lambda D_{G,g} \text{ pointwise} \tag{8}$$

if and only if $F - \lambda G$ is convex and $f(q) - \lambda g(q) \in \partial(F - \lambda G)(q)$ for every q.

Subtracting the divergences gives the subgradient inequality for $F - \lambda G$. At a kink, convexity alone is insufficient: the selected slopes must also be compatible.

Consider two full-dimensional cells of a finite-action envelope G meeting on a relative-open facet H. Their gradients satisfy

$$g^+ - g^- = \alpha_H n_H, \qquad \alpha_H > 0, \tag{9}$$

where n_H is the oriented unit normal in $\operatorname{aff}(\mathcal{P})$. Write $d = \dim \operatorname{aff}(\mathcal{P})$. The jump α_H, equivalently the singular Hessian mass on H, is its *curvature charge*.

Lemma 3.2 (Decision-facet charge)**.** *If $D_{F,f} \geq \lambda D_{G,g}$ globally for $\lambda > 0$, then*

$$D^2 F|_H \succeq \lambda \alpha_H n_H n_H^\top \mathscr{H}^{d-1}|_H$$

as positive-semidefinite matrix-valued measures. Consequently, an entropy that is C^1 near an interior point of H cannot dominate the hard decision regret with a positive global linear constant.

Lemma 3.1 gives $D^2 F \succeq \lambda D^2 G$, and the restriction of $D^2 G$ to H is the displayed rank-one measure. Equivalently, regret is first-order across the facet, whereas a C^1 score divergence is little-o of the crossing distance.

To see that the bound in Proposition 2.3 is sharp, specialize to $\mathcal{P} = \Delta_K$. Take $p_0 \in \operatorname{ri}(\Delta_K)$, let $m \leq K - 1$, and choose orthonormal simplex-tangent vectors $v_1, \ldots, v_m$. Set

$$G_j(p) = |\langle v_j, p - p_0 \rangle|.$$

Each envelope defines a two-action task.

Theorem 3.3 (Sharp transversal distortion)**.** *Under the standing baseline assumptions, these unit-weight tasks satisfy*

$$\kappa^\star_\varepsilon = m \qquad \text{for every } \varepsilon > 0.$$

The upper bound is Proposition 2.3. For the lower bound, domination implies $D^2 F - \varepsilon D^2 F_0 \succeq D^2 G_j$ for each j. The facet measures are mutually singular away from lower-dimensional intersections, so their charges add and $F - \varepsilon F_0 - \sum_j G_j$ is convex on an interior neighborhood of p_0. Along

$$q_t = p_0 - t \sum_j v_j, \qquad p_t = p_0 + t \sum_j v_j, \tag{10}$$

each task divergence is $2t$, while $D_{F_0}(p_t, q_t) = o(t)$. Every feasible ratio therefore has limit inferior at least m. Appendix F gives the measure restriction and selection argument.

Corollary 3.4 (Off-tie sharpness)**.** *For sufficiently small $t > 0$, both endpoints in* (10) *lie strictly inside decision cells. Every feasible proper score satisfies $\liminf_{t \downarrow 0} D_{F,f}(p_t, q_t)/T_\varepsilon(p_t, q_t) \geq m$.*

Partial sharing gives an exact intermediate factor. Let $p_0 \in \operatorname{ri}(\Delta_K)$, take linearly independent tangent vectors $v_1, \ldots, v_L$, and define

$$H_\ell(p) = |\langle v_\ell, p - p_0 \rangle|, \qquad G_j = \sum_{\ell=1}^{L} a_{j\ell} H_\ell, \quad A \in \mathbb{R}_+^{m \times L}.$$

Task weights are absorbed into A. Each component uses one deterministic sign at a tie, shared by all tasks. With component selections h_ℓ, set $g_j = \sum_\ell a_{j\ell} h_\ell$.

Theorem 3.5 (Exact distortion for independent shared components)**.** *Suppose $1 \le L \le K-1$, $m \ge 1$, and $c_\ell = \max_j a_{j\ell} > 0$ for every component. On the full simplex, under the standing baseline assumptions and common selections,*

$$\kappa^\star_\varepsilon = \max_{r\ge 0,\, Ar\le \mathbf{1}} c^\top r = \min_{u\ge 0,\, A^\top u\ge c} \mathbf{1}^\top u. \tag{11}$$

The entropy $\varepsilon F_0 + \sum_\ell c_\ell H_\ell$, with matching component selections, attains the optimum over all selected convex scores.

Each component must carry its largest task charge. Independent crossings realize arbitrary positive component-regret proportions, giving the matching lower bound (Appendix G). The covering problem also yields optimal task-score weights. For feasible u, set $F_u = \varepsilon F_0 + \sum_j u_j G_j$ and $f_u = \varepsilon \nabla F_0 + \sum_j u_j g_j$. Writing $b_0 = \varepsilon D_{F_0}(p,q)$ and $r_\ell = D_{H_\ell,h_\ell}(p,q)$, feasibility gives $A^\top u \ge c$ and $\mathbf{1}^\top u \ge 1$, since $c_\ell \le \sum_j a_{j\ell} u_j \le c_\ell \mathbf{1}^\top u$. Hence

$$T_\varepsilon = b_0 + \max_j (Ar)_j \le b_0 + u^\top A r = D_{F_u,f_u},$$
$$D_{F_u,f_u} \le b_0 + (\mathbf{1}^\top u) \max_j (Ar)_j \le (\mathbf{1}^\top u) T_\varepsilon.$$

Any covering minimizer therefore attains the all-score optimum using a weighted task sum, with the baseline included once. These training weights leave the declared target unchanged. For rows $(1,\alpha),(\alpha,1)$, $0\le\alpha\le 1$, weights $u_1 = u_2 = 1/(1+\alpha)$ give $(G_1+G_2)/(1+\alpha) = H_1 + H_2$ and factor $2/(1+\alpha)$. In general, equally optimal constructions need not select the same class projection. The covering upper bound extends to dependent convex components with common selections; the converse requires independence. Optimal distortion does not assert pointwise leastness at joint ties.

Under the assumptions of Theorem 3.5,

$$\kappa^\star_\varepsilon = 1 \quad\Longleftrightarrow\quad \exists j^\star :\ a_{j^\star \ell} = c_\ell \text{ for every } \ell.$$

Thus the factor is one precisely when a single task attains all componentwise maxima. In this case, the selected score generated by $\varepsilon F_0 + G_{j^\star}$ has divergence exactly T_ε, so score and target risks have the same minimizers in every report class. Duplicating an unchanged task leaves the optimal factor unchanged (Appendix G).

In one dimension, a pointwise least common majorant can be constructed from the selected charges on either side of each knot. Let C be the union of interior task knots and define

$$\ell_{j,c} = \lambda_j\{g_j(c) - G'_{j,-}(c)\}, \qquad r_{j,c} = \lambda_j\{G'_{j,+}(c) - g_j(c)\}. \tag{12}$$

Theorem 3.6 (Exact binary selected-charge join)**.** *Suppose $\mathcal{P} = [a,b]$, every task envelope is convex piecewise linear, and endpoint selections are inward: $g_j(a) = G'_{j,+}(a)$, $g_j(b) = G'_{j,-}(b)$. Set*

$$\ell^\vee_c = \max_j \ell_{j,c}, \qquad r^\vee_c = \max_j r_{j,c}, \qquad \alpha^\vee_c = \ell^\vee_c + r^\vee_c. \tag{13}$$

Modulo an affine term, let $H_\vee$ have curvature $\sum_{c\in C} \alpha^\vee_c \delta_c$. Select $h_\vee(c) = H'_{\vee,-}(c) + \ell^\vee_c$ at knots, the derivative elsewhere, and inward derivatives at the endpoints. Then

$$D_{H_\vee,h_\vee} \ge \lambda_j D_{G_j,g_j} \quad \text{for every } j, \tag{14}$$

and every other common selected majorant dominates it pointwise. Consequently,

$$\kappa^\star_\varepsilon = \sup_{p\ne q} \frac{\varepsilon D_{F_0}(p,q) + D_{H_\vee,h_\vee}(p,q)}{T_\varepsilon(p,q)}. \tag{15}$$

Domination forces at least $\ell^\vee_c$ and $r^\vee_c$ on the two selected sides of each knot; subtracting $H_\vee$ leaves a convex residual with a compatible selection (Appendix E.2). Two copies of $\alpha(q-c)_+$ require joined charge α with the same tie side, but 2α with opposite sides. Protection against all Bayes-optimal tied actions instead doubles the maximum full interior charge and adds one-sided endpoint charges (Theorem E.1, Appendix E.3).

For comparison, consider quadratic tasks on a compact convex affine chart $\mathcal{Z} \subset \mathbb{R}^d$ with nonempty interior. Let $F_0(z) = \frac{1}{2} z^\top A_0 z$, $A_0 \succ 0$, and $G_j(z) = \frac{1}{2} z^\top A_j z$, $A_j \succeq 0$. Define $Q(v) = \varepsilon v^\top A_0 v + \max_j \lambda_j v^\top A_j v$, $\mathcal{K} = \{v : Q(v) \le 1\}$, and $E_M = \{v : v^\top M v \le 1\}$. Theorem H.1 gives

$$\kappa^\star_\varepsilon = \inf\{\kappa \ge 1 : \exists M \succ 0,\ E_M \subseteq \mathcal{K} \subseteq \sqrt{\kappa} E_M\}.$$

The optimum over all finite convex entropies is attained by a quadratic entropy. Writing $r = \dim \operatorname{span}_j \operatorname{range}(A_j)$, we have $\kappa^\star_\varepsilon \le \min\{m, r\}$ for $r \ge 1$, and $\kappa^\star_\varepsilon = 1$ for $r = 0$. In the orthogonal case $d = m$, $A_0 = I_m$, $A_j = e_j e_j^\top$, and $\lambda_j = 1$,

$$\kappa^\star_\varepsilon = \frac{m(1+\varepsilon)}{1+m\varepsilon}.$$

These bounded continuous-action tasks provide a smooth counterpart to the finite-action construction. The full statement and proof are in Appendix H.

4 Misspecified learning and the hard–smooth interface

Fix a population P, conditional truth p_X, and a measurable report class $\{q_h(X) : h \in \mathcal{H}\}$. For a measurable divergence D, write

$$L_D(h) = \mathbb{E}_X D(p_X, q_h(X)), \qquad L_T(h) = \mathcal{R}_T(h) = \mathbb{E}_X T_\varepsilon(p_X, q_h(X)).$$

The two sides of the sandwich immediately transfer an approximation error in population score risk to target risk.

Corollary 4.1 (Approximate score minimization)**.** *Suppose a measurable selected score has divergence $D = D_{F,f}$ satisfying $T_\varepsilon \le D \le \kappa T_\varepsilon$. If $\widetilde{h}_F$ has population score risk within η of the infimum over $\mathcal{H}$, then*

$$\mathcal{R}_T(\widetilde{h}_F) \le \kappa \inf_{h\in\mathcal{H}} \mathcal{R}_T(h) + \eta. \tag{16}$$

Population score risk is a constant plus L_D, so the result follows by comparing the score minimizer with a target-risk comparator. Theorem 2.2 gives the sharp uniform converse when $\eta = 0$, after optimal scaling. With $\mathcal{R}^\star_T = \inf_{h\in\mathcal{H}} \mathcal{R}_T(h)$, the excess target risk is at most $(\kappa-1)\mathcal{R}^\star_T + \eta$. This multiplicative oracle bound does not imply exact $\mathcal{H}$-consistency (Awasthi et al., 2022); η measures population score suboptimality, not a finite-sample rate.

Fix a selected score satisfying (3) and a report class before observing an i.i.d. sample of size n. Assume the induced score class $\mathcal{S}_F = \{(x,y) \mapsto S_{F,f}(q_h(x), y) : h \in \mathcal{H}\}$ is measurable and takes values in a common interval of length B_F. If $\widehat{h}_F$ has empirical score-risk gap η_n, uniform convergence and Corollary 4.1 give, with probability at least $1-\delta$, $\delta \in (0,1)$,

$$\begin{aligned}\mathcal{R}_T(\widehat{h}_F) - \mathcal{R}^\star_T \le (\kappa-1)\mathcal{R}^\star_T + 4\mathfrak{R}_n(\mathcal{S}_F) \\ + 2B_F\sqrt{\frac{\log(2/\delta)}{2n}} + \eta_n.\end{aligned}$$

Here $\mathfrak{R}_n(\mathcal{S}_F)$ is the expected Rademacher complexity. The bound separates the score-projection penalty from estimation and optimization errors (Appendix I). A smaller distortion alone need not yield a smaller finite-sample bound, since the score class also changes.

A fixed population and class can admit a smaller factor. Define

$$a_{\mathcal{H},P} = \inf_{h: L_T(h)>0} \frac{L_D(h)}{L_T(h)}, \qquad b_{\mathcal{H},P} = \sup_{h: L_T(h)>0} \frac{L_D(h)}{L_T(h)}, \tag{17}$$

and let $\kappa_{\mathcal{H},P} = b_{\mathcal{H},P}/a_{\mathcal{H},P}$ be the *class-restricted risk distortion*.

Proposition 4.2 (Global, class-restricted, and achieved factors)**.** *Suppose $0 < a_{\mathcal{H},P} \le b_{\mathcal{H},P} < \infty$, the best target risk is positive, and h_D minimizes L_D over $\mathcal{H}$. Then*

$$\rho_{\mathcal{H},P}(D, T_\varepsilon) := \frac{L_T(h_D)}{\inf_{h\in\mathcal{H}} L_T(h)} \le \kappa_{\mathcal{H},P}(D, T_\varepsilon). \tag{18}$$

If also $0 < a(D,T_\varepsilon) \le b(D,T_\varepsilon) < \infty$, then $\kappa_{\mathcal{H},P}(D,T_\varepsilon) \le b(D,T_\varepsilon)/a(D,T_\varepsilon)$.

The first inequality follows by rescaling by $1/a_{\mathcal{H},P}$ and comparing minimizers. For the second, $L_D(h)/L_T(h)$ is a T_ε-weighted average of pointwise ratios, giving $a \le a_{\mathcal{H},P} \le b_{\mathcal{H},P} \le b$. If the class contains the truth, strict propriety makes both best risks zero; the ratio is undefined, and

we report agreement with the truth. Held-out risk ratios between fitted predictors are empirical comparisons, not certified values of the population projection ratio $\rho_{\mathcal{H},P}$.

These population comparisons do not provide an optimization algorithm. For hard majorants, the curvature required for a finite factor also limits the gradients available during fitting.

Proposition 4.3 (Hard-facet score geometry)**.** *Let B be convex and C^2, and let $J(q) = \max_a\{b_a + \langle s_a, q\rangle\}$. On a cell with unique active piece a, set $F = B + J$ and $f = \nabla B + s_a$. Then*

$$S_{F,f}(q,y) = S_{B,\nabla B}(q,y) - b_a - \langle s_a, e_y\rangle, \qquad \nabla_q S_{F,f}(q,y) = \nabla^2 B(q)(q - e_y). \tag{19}$$

For $J = \max\{\phi_-, \phi_+\}$, $\phi_\pm(q) = b_\pm + \langle s_\pm, q\rangle$, select s_- at a tie q_0 and orient the unit normal n toward the plus cell, with $s_+ - s_- = \alpha n$. Then $D_{J,s_-}(q_0 + tn, q_0) = \alpha t$ for $t > 0$ in the domain. Its log-sum-exp divergence is $o(t)$ at fixed temperature $\tau > 0$ and tends to $\alpha t/2$ at fixed t as temperature vanishes; tie curvature scales as $1/\tau$.

The task increment changes score values across a facet but contributes no within-cell gradient. This is a statement about ordinary pointwise derivatives. The population objective can also have a boundary-flux derivative, derived in Appendix J; it is not recovered by differentiating the finite empirical objective between sample crossings. Appendix E gives the log-cosh formula and its limits. Smoothing introduces a local task gradient, but cannot preserve the selected hard crossing regret without an additive deficit. For the left-selected hinge $G(t) = \alpha t_+$, $g(0) = 0$, Theorem E.2 shows that an L-smooth convex increment satisfying $D_J(t,0) \geq D_G(t,0) - \delta$ on $[0, R]$, with $R \geq \alpha/L$, must satisfy

$$\delta \geq \alpha^2/(2L).$$

Smoothness gives $D_J(t,0) \leq Lt^2/2$; maximizing the deficit proves the lower bound. The Huber hinge attains it. A common differentiable baseline cancels (Appendix E.5). Appendix E also treats finite-max smoothing away from facets, including duplicate affine pieces. Figure 2 shows the gradient obstruction and smoothing deficit.

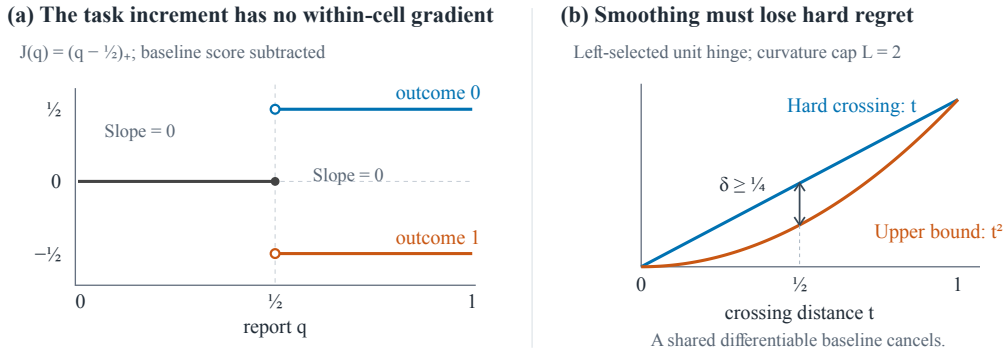


Figure 2: **The optimization cost of hard regret control.** (a) For $J(q) = (q - 1/2)_+$ with the left selection at the tie, the task score increment is $(1/2 - y)\mathbf{1}\{q > 1/2\}$ for $y \in \{0, 1\}$; its within-cell derivative is zero. The differentiable baseline is subtracted. (b) At a left-selected unit hinge, an $L = 2$ smooth increment has divergence at most t^2, below the hard crossing t by at least $1/4$ at $t = 1/2$. The statements concern pointwise score derivatives and divergence bounds.

For a smoothed score, let $\Delta_\tau(h)$ bound the absolute population hard–smooth risk discrepancy for predictor h. Corollary I.1 gives, for an η-optimal smoothed projection $\widehat{h}_\tau$ and any comparator h_T,

$$\mathcal{R}_T(\widehat{h}_\tau) \leq \kappa \mathcal{R}_T(h_T) + \eta + \Delta_\tau(\widehat{h}_\tau) + \Delta_\tau(h_T).$$

For finite log-sum-exp, the discrepancy depends on low-margin report mass. This bounds the hard–smooth approximation error, not the target-risk ordering of two fitted scores, which also depends on

their projection locations and task geometry. Appendices E and I give the constants and finite-sample extension.

5 From global certificates to fixed-distribution performance

A controlled binary study compares global factors, class-restricted factors (Proposition 4.2), and achieved projection ratios on 4,001 contexts with known truth, fixed tasks, and a common optimization protocol. Three nested heads calibrate a fixed base score: offset fits an intercept, Platt fits slope and intercept, and augmentation adds the nonlinear truth component. Its amplitude $\gamma = 0$ makes Platt realizable; $\gamma = 1$ requires augmentation. We fix unit task weights and $D_{F_0}(p, q) = (p - q)^2$.

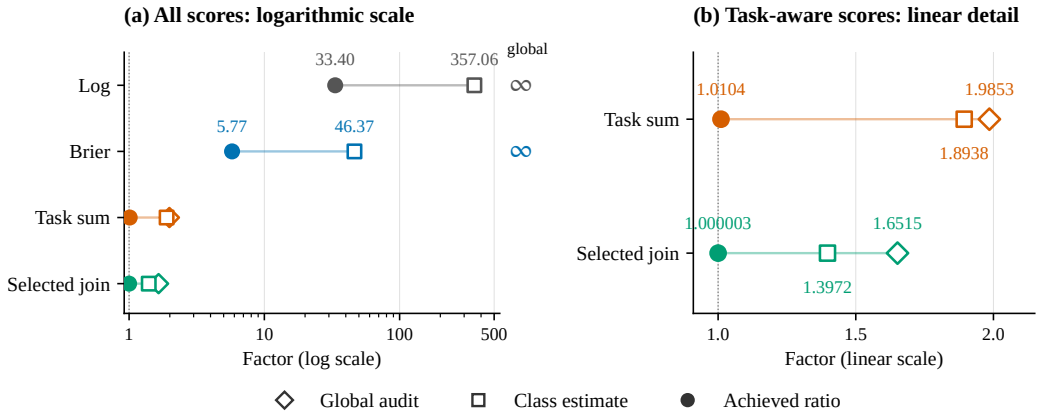


Figure 3: **Global factors, class factors, and achieved risk can be far apart.** Primary Platt population at $\gamma = 1$ and $\varepsilon = 0.05$. (a) All four scores on a logarithmic scale; analytic infinite global factors are displayed outside the numeric axis. (b) Linear detail separates the task-aware constructions near one. Finite global values are numerical audits; class factors are fitted estimates. Achieved ratios use the directly optimized target risk as denominator. Table 3 gives the values.

Global guarantees and achieved risk. The join nearly attains the directly optimized target reference, while its global audit and class-factor estimate remain above one (Figure 3). Log and Brier have infinite global hard-facet factors but finite estimated factors over the declared bounded report class. For log loss, the infinite-factor conclusion is already witnessed by interior crossings of a task boundary; boundary infinities are not needed. Their target risks are 1.879×10^{-3} and 3.25×10^{-4}, versus the direct reference's 5.6×10^{-5}: the small denominator amplifies the achieved ratios. Primary Platt extrema are interior; 43 extrema across the full parameter grid reach a declared bound. These numerical searches supply the finite-factor estimates (Appendix K).

The realizability and baseline-weight controls appear in Appendix K, Figures 4 and 5.

From the contextwise target to fixed users. The pointwise theory aggregates the worst task within each context, whereas a fixed-user objective aggregates each user's loss over contexts first. Write

$$B_{\mathrm{ctx}}(h) = \mathbb{E}_X \max_j \lambda_j R_j(p_X, q_h(X)), \qquad B_{\mathrm{fixed}}(h) = \max_j \mathbb{E}_X \lambda_j R_j(p_X, q_h(X)).$$

Then $B_{\mathrm{fixed}} \le B_{\mathrm{ctx}}$; adding the same proper baseline preserves this inequality. The score-projection guarantee uses the best contextwise target risk as its comparator. With observed labels, the fixed-user studies measure

$$C_{\mathrm{fixed}}(h) = \max_j \mathbb{E}_{X,Y} \lambda_j \left[\ell_{j,a_j(q_h(X))}\right]_Y,$$

where $[\cdot]_Y$ takes the outcome-Y component of the loss vector. User-specific Bayes-risk constants remain inside this maximum, so C_{fixed} need not share the worst user or optimizer of B_{fixed}.

Binary fixed-user study. The binary real-label study evaluates how score choice affects the costs of two declared users, with C_{fixed} measuring their worst mean cost. Average and worst-user costs can move differently; the paired intervals for the primary 1 mm worst-cost comparisons include zero. Appendix K gives the design, robustness panels, and paired intervals.

Four-class fixed-user study. The four-class study evaluates the *shared-component majorant* on held-out precipitation outcomes. The construction adds each shared convex or task-specific component once to the baseline; convex subtraction establishes it as a feasible majorant of each task. In the primary station-holdout comparison, its estimated cost difference relative to log is -0.000486, with a paired 95% interval $[-0.002937, 0.001977]$. Appendix L gives all 48 method contrasts, including the 12 shared-component-majorant comparisons. These fixed-population cost comparisons complement the uniform distortion analysis.

6 Related work

Surrogate regret bounds relate target excess risk to surrogate excess risk. Proper losses connect probability estimation and classification calibration (Reid & Williamson, 2010), with quantitative probability-error bounds studied by Bao & Takatsu (2025). Polyhedral embeddings provide consistent convex surrogates (Finocchiaro et al., 2024), and consistent polyhedral surrogates admit linear regret bounds (Frongillo & Waggoner, 2021). Awasthi et al. (2022) account explicitly for the hypothesis class through $\mathcal{H}$-consistency bounds. We choose a proper score to minimize the worst-case multiplicative target-risk ratio over populations and report classes, relative to the best target predictor in each class. The ratio can exceed one even at exact population score minima.

Cao et al. (2025) construct convex smooth Fenchel–Young losses with linear regret bounds and task-dependent prediction links. Their generator adds the target's negative Bayes risk to a base negentropy, sharing the additive structure of our task-augmented entropies. Their smoothness is in unconstrained score coordinates, whereas Lemma 3.2 concerns differentiability of the entropy in probability-report coordinates under the declared Bayes decoders. A nonsmooth entropy can generate a smooth surrogate, so these results are compatible.

Proper-score construction builds on convex-entropy representations and binary threshold mixtures (Gneiting & Raftery, 2007; Schervish, 1989). Score choice can favor different users under misspecification (Elliott et al., 2016). Decision-focused methods incorporate downstream costs into training (Donti et al., 2017; Elmachtoub & Grigas, 2022). Plaud et al. (2026) tailor strictly proper scores by matching local curvature to downstream estimation error. Our comparison is global and targets the contextwise maximum of declared decision regrets, augmented by a proper baseline.

Omniprediction seeks a single predictor with loss-by-loss guarantees (Gopalan et al., 2022); Gibbs & Tibshirani (2026) develop sample-efficient algorithms for proper losses. Our comparator instead minimizes the aggregate target T_ε within the report class. We restrict the learner to minimizing one fixed proper score in that class, so our lower bounds do not apply to arbitrary omniprediction algorithms.

Scoring rules have also been optimized for information-acquisition incentives (Li et al., 2022; Hartline et al., 2023), rather than ordered-pair regret distortion. Our facet-charge arguments bound the distortion of every dominating selected score. For independent shared components with common tie selections, optimally weighted task-score sums attain these bounds over all selected convex scores.

7 Conclusion and limitations

We characterize the optimal worst-class risk factor of proper-score minimization. Facet charges give matching bounds, and independent shared components admit optimal weighted task-score sums. The experiments distinguish uniform distortion guarantees from achieved projection risk and observed user costs on fixed populations. The hard-score experiments use low-dimensional calibration heads and derivative-free search. Scaling to high-dimensional training calls for smoothed objectives with the stated approximation slack or scalable estimators of the population facet flux. General facet arrangements, higher-dimensional protection against all tied actions, open-simplex attainment, and measurable-selection closure remain open.

AI USE STATEMENT

Generative AI assisted literature discovery, drafting, mathematical proofs, code, figures, and checks. The authors verified the outputs and take responsibility for the submission.

REPRODUCIBILITY STATEMENT

The appendix provides proofs and experimental protocols. Code and frozen outputs are available at `https://anonymous.4open.science/r/proper-majorants-1D60`. The repository's driver reaggregates reports and rebuilds the empirical figures. The repository README lists the figures and tables reproduced by the code and explains how to refit the models using public data.

REFERENCES

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. $\mathcal{H}$-consistency bounds for surrogate loss minimizers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1117–1174. PMLR, 2022. URL https://proceedings.mlr.press/v162/awasthi22c.html.

Han Bao and Asuka Takatsu. Proper losses regret at least 1/2-order. *Journal of Machine Learning Research*, 26(288):1–50, 2025. URL https://jmlr.org/papers/v26/24-1112.html.

Yuzhou Cao, Han Bao, Lei Feng, and Bo An. Establishing linear surrogate regret bounds for convex smooth losses via convolutional Fenchel–Young losses. In *Advances in Neural Information Processing Systems*, volume 38, pp. 33200–33237, 2025. doi: 10.52202/085713-1116. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2fbeed1dd7162f91804e7b9246e0c1a8-Abstract-Conference.html.

Priya L. Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5484–5494, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/3fc2c60b5782f641f76bcefc39fb2392-Abstract.html.

Graham Elliott, Dalia Ghanem, and Fabian Krüger. Forecasting conditional probabilities of binary outcomes under misspecification. *The Review of Economics and Statistics*, 98(4):742–755, 2016. doi: 10.1162/REST_a_00564. URL https://direct.mit.edu/rest/article/98/4/742/58337.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Jessie Finocchiaro, Rafael M. Frongillo, and Bo Waggoner. An embedding framework for the design and analysis of consistent polyhedral surrogates. *Journal of Machine Learning Research*, 25(63): 1–60, 2024. URL https://jmlr.org/papers/v25/22-0743.html.

Rafael M. Frongillo and Bo Waggoner. Surrogate regret bounds for polyhedral losses. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21569–21580, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b4572f47b7c69e27b8e46646d9579e67-Abstract.html.

Isaac Gibbs and Ryan J. Tibshirani. Sample-efficient omniprediction for proper losses. In *Proceedings of the 39th Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 2679–2719. PMLR, 2026. URL https://proceedings.mlr.press/v336/gibbs26a.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *13th Innovations in Theoretical Computer Science Conference*, volume 215 of *Leibniz International Proceedings in Informatics*, pp. 79:1–79:21. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022. doi: 10.4230/LIPIcs.ITCS.2022.79.

Jason D. Hartline, Liren Shan, Yingkai Li, and Yifan Wu. Optimal scoring rules for multi-dimensional effort. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2624–2650. PMLR, 2023. URL https://proceedings.mlr.press/v195/hartline23a.html.

Wonyeol Lee, Hangyeol Yu, and Hongseok Yang. Reparameterization gradient for non-differentiable models. In *Advances in Neural Information Processing Systems*, volume 31, pp. 5553–5563, 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/b096577e264d1ebd6b41041f392eec23-Abstract.html.

Yingkai Li, Jason D. Hartline, Liren Shan, and Yifan Wu. Optimization of scoring rules. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 988–989. ACM, 2022. doi: 10.1145/3490486.3538338. Full version: arXiv:2007.02905.

Matthew J. Menne, Imke Durre, Bryant Korzeniewski, Shelley McNeill, Kristy Thomas, Xungang Yin, Steven Anthony, Ron Ray, Russell S. Vose, Byron E. Gleason, and Tamara G. Houston. Global historical climatology network - daily (GHCN-Daily), version 3. NOAA National Climatic Data Center, 2012. URL https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily. doi: 10.7289/V5D21VHZ. 26-station subset; accessed 2026-08-23.

Roman Plaud, Alexandre Perez-Lebel, Antoine Saillenfest, Thomas Bonald, Marine Le Morvan, Gaël Varoquaux, and Matthieu Labeau. Tailoring strictly proper scoring rules for downstream tasks: An application to causal inference, 2026. URL https://arxiv.org/abs/2606.03332. Accepted to the 43rd International Conference on Machine Learning.

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/v11/reid10a.html.

Mark J. Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, 17(4):1856–1879, 1989. doi: 10.1214/aos/1176347398.

A GUIDE TO THE SUPPLEMENT

Appendices B–I contain the theory; Appendix J gives the population-gradient formula. The studies cover binary projections and costs (Appendix K) and four-class costs (Appendix L).

Notation. We use $[K] = \{1, \dots, K\}$, $\Delta_K = \{p \in \mathbb{R}_+^K : \sum_k p_k = 1\}$, and the yth coordinate vector e_y. On context space $\mathcal{X}$, $p_x = (P(Y = k \mid X = x))_{k=1}^K$ under population law P, with $q_h : \mathcal{X} \to \mathcal{P}$; $p, q \in \mathcal{P}$, $y \in [K]$, and $m \geq 1$. Binary outcomes are relabeled $0, 1$: a scalar p represents $(1-p, p)$, and $\mathcal{P} = [a, b] \subseteq [0, 1]$ uses scalar derivatives, with $a < b$. Thus $\frac{1}{2}\|(1-p, p) - (1-q, q)\|_2^2 = (p-q)^2$. For differentiable F, $D_F := D_{F,\nabla F}$; otherwise an omitted selection is the one declared locally. D^2F denotes the distributional Hessian, $\mathscr{H}^k$ Hausdorff measure, and δ_c the Dirac mass at c. We write $t_+ = \max\{t, 0\}$ and $\sigma(t) = (1 + e^{-t})^{-1}$. Table 1 separates the estimands and guarantee scopes.

Table 1: Quantifiers that distinguish the global theory, supporting constructions, and empirical studies.

Axis	Global theorem	Alternative	Interpretation
Interface	One probability report q	Correction head or multiple reports	Changes the report interface
Task aggregation	$B_{\mathrm{ctx}}(h)$	$B_{\mathrm{fixed}}(h)$; labeled evaluation: $C_{\mathrm{fixed}}(h)$	Context-adaptive, fixed-user regret, and observed fixed-user cost
Overlap	Independent shared components with common tie signs	Arbitrary CPWL arrangements	Exact covering factor versus open integrability problem
Ties	Declared decoder; selected score	Every Bayes-optimal tied action	Tie-robust least majorants: exact in 1D; general higher-dimensional case open
Learning	Worst misspecified report class	One fixed distribution/class	Global, class-restricted, and achieved factors

B SELECTED DIVERGENCE COMPARISON

B.1 PROOF OF LEMMA 3.1

Let $H = F - \lambda G$ and $h(q) = f(q) - \lambda g(q)$. Direct subtraction gives

$$D_{F,f}(p, q) - \lambda D_{G,g}(p, q) = H(p) - H(q) - \langle h(q), p - q\rangle.$$

Thus (8) is equivalent to

$$H(p) \geq H(q) + \langle h(q), p - q\rangle \qquad \text{for every } p, q. \tag{20}$$

If H is convex and $h(q) \in \partial H(q)$, this inequality is the subgradient inequality. Conversely, (20) supplies a global affine support at every q. Taking the supremum of these affine supports recovers H, so H is convex and $h(q) \in \partial H(q)$.

The compatibility clause cannot be dropped. Take $F(x) = G(x) = |x|$, $q = 0$, and $\lambda = 1$. Although $F - G = 0$ is convex, independently choosing $f(0) = 1$ and $g(0) = -1$ yields, for $p > 0$,

$$D_{F,f}(p, 0) = 0 \quad \text{and} \quad D_{G,g}(p, 0) = 2p.$$

C OPERATIONAL EQUIVALENCE

C.1 PROOF OF THEOREM 2.2

Write $a = a(D, T)$ and $b = b(D, T)$. Since multiplying D by a positive constant does not change its minimizers, normalize it as $\widetilde{D} = D/a$. Then

$$T \leq \widetilde{D} \leq \frac{b}{a} T.$$

For any admissible context law and report class, let h_D minimize the D-risk and let h_T minimize the T-risk. The sandwich gives

$$\begin{aligned} \mathbb{E}T(p_X, q_{h_D}(X)) &\leq \mathbb{E}\widetilde{D}(p_X, q_{h_D}(X)) \\ &\leq \mathbb{E}\widetilde{D}(p_X, q_{h_T}(X)) \\ &\leq \frac{b}{a}\mathbb{E}T(p_X, q_{h_T}(X)). \end{aligned}$$

Hence $\Gamma(D,T) \le b/a$.

If $a = b$, then $D = aT$ pointwise. Every score projection is target-optimal, and a singleton report class with positive target risk realizes $\Gamma = 1$ with a unique minimizer. It remains to consider $a < b$.

For the reverse inequality, fix $\xi > 0$ and choose pairs $z_- = (p_-, q_-)$ and $z_+ = (p_+, q_+)$ with positive target and ratios

$$r_- := \frac{D(z_-)}{T(z_-)} \le a + \xi, \qquad r_+ := \frac{D(z_+)}{T(z_+)} \ge b - \xi.$$

For sufficiently small ξ, $r_- < r_+$. Let the context space be $\{x_-, x_+\}$, with truths p_- and p_+. Define two report maps by

$$\begin{aligned} q_{h_-}(x_-) &= q_-, & q_{h_-}(x_+) &= p_+, \\ q_{h_+}(x_-) &= p_-, & q_{h_+}(x_+) &= q_+. \end{aligned}$$

Assign masses

$$w_- := \Pr(x_-) = \frac{D(z_+)}{D(z_-) + D(z_+)}, \qquad w_+ := \Pr(x_+) = \frac{D(z_-)}{D(z_-) + D(z_+)}.$$

Both hypotheses have score risk

$$c = w_- D(z_-) = w_+ D(z_+) > 0.$$

Their target risks are c/r_- and c/r_+. Adverse score tie breaking selects h_-, whereas h_+ is target-optimal, so the operational ratio is r_+/r_-. Letting $\xi \downarrow 0$ proves $\Gamma(D,T) \ge b/a$.

To make the score minimizer unique, replace w_- by $w_- - \zeta$ and w_+ by $w_+ + \zeta$ for sufficiently small $\zeta > 0$. Then h_- has strictly smaller score risk, and its target-risk ratio relative to h_+ converges to r_+/r_- as $\zeta \downarrow 0$. The definition of Γ is a supremum, so the same value follows under unique minimization.

If $a = 0$ or $b = \infty$, fix one pair with a positive finite ratio and choose the other with ratio tending to zero or infinity, respectively. The same two-context construction makes r_+/r_- arbitrarily large, also with a unique score minimizer after perturbing the masses. Thus $\Gamma = \infty$.

Finally, for fixed D, any rescaling that satisfies $T \le cD \le \kappa T$ must have $c \ge 1/a$ and $\kappa \ge cb \ge b/a$. For finite positive endpoints, $c = 1/a$ attains b/a; at either infinite-factor endpoint no finite sandwich is possible. Taking the infimum over strictly proper selected divergences gives (7). If $T_\varepsilon \equiv 0$, strict convexity of the baseline forces $\mathcal{P}$ to be a singleton, and both sides equal one by convention.

D ATTAINMENT

Theorem D.1 (Attainment). *Let $T : \mathcal{P}^2 \to [0,\infty)$ be bounded and vanish on the diagonal. Define*

$$\begin{aligned} \mathfrak{K}_T = \{\kappa \ge 1 : \exists F \textit{ finite convex},\ f(q) \in \partial_{\mathcal{P}} F(q), \\ T(p,q) \le D_{F,f}(p,q) \le \kappa T(p,q) \quad \forall p, q \in \mathcal{P}\}. \end{aligned}$$

If $\mathfrak{K}_T \neq \varnothing$, then $\inf \mathfrak{K}_T$ is attained by a finite convex entropy and a pointwise subgradient selection. If $T \ge \varepsilon D_{F_0}$, where F_0 is differentiable and strictly convex, the attaining entropy is strictly convex.

D.1 PROOF OF THEOREM D.1

Work in the linear space E parallel to the affine hull of $\mathcal{P}$ and identify subgradients with their orthogonal projections onto E. Let $\kappa^\star$ be the infimal distortion. Choose feasible pairs (F_n, f_n) and constants $\kappa_n \downarrow \kappa^\star$ such that

$$T(p,q) \le D_{F_n,f_n}(p,q) \le \kappa_n T(p,q) \qquad \text{for every } p, q \in \mathcal{P}. \tag{21}$$

After discarding finitely many terms, fix $\bar\kappa < \infty$ with $\kappa_n \le \bar\kappa$.

If $\dim E = 0$, then $\mathcal{P}$ is a singleton and the result is immediate. Hence assume $\dim E \ge 1$ below. All subgradients and convexity statements are relative to $\operatorname{aff}(\mathcal{P})$.

Choose $q_0 \in \mathcal{P}$ and apply the affine normalization

$$\widetilde{F}_n(p) = F_n(p) - F_n(q_0) - \langle f_n(q_0), p - q_0 \rangle,$$
$$\widetilde{f}_n(q) = f_n(q) - f_n(q_0). \tag{22}$$

Every selected divergence is unchanged. In particular,

$$\widetilde{F}_n(p) = D_{F_n, f_n}(p, q_0). \tag{23}$$

Let $C = \sup_{p,q \in \mathcal{P}} T(p, q)$ and $M = \bar{\kappa} C$. Equations (21)–(23) imply

$$0 \le \widetilde{F}_n(p) \le M \qquad \text{for every } p, n. \tag{24}$$

For arbitrary p, q, rearrange the divergence identity:

$$\langle \widetilde{f}_n(q), p - q \rangle = \widetilde{F}_n(p) - \widetilde{F}_n(q) - D_{\widetilde{F}_n, \widetilde{f}_n}(p, q).$$

The first difference lies in $[-M, M]$, while the divergence lies in $[0, M]$. Therefore

$$\left| \langle \widetilde{f}_n(q), p - q \rangle \right| \le 2M. \tag{25}$$

For $p_1, p_2 \in \mathcal{P}$, subtracting two such chord expressions gives

$$\left| \langle \widetilde{f}_n(q), p_1 - p_2 \rangle \right| \le 4M. \tag{26}$$

Because $\mathcal{P}$ has nonempty relative interior, its difference body $\mathcal{P} - \mathcal{P}$ contains r times the unit ball of E for some $r > 0$. Choosing the ball vector in the direction of $\widetilde{f}_n(q)$ in (26) gives

$$\|\widetilde{f}_n(q)\|_2 \le \frac{4M}{r} \qquad \text{for every } q, n. \tag{27}$$

The subgradient inequality at p and q now yields

$$|\widetilde{F}_n(p) - \widetilde{F}_n(q)| \le \frac{4M}{r} \|p - q\|_2.$$

Thus the normalized entropies are uniformly bounded and equi-Lipschitz.

Arzelà–Ascoli supplies a subsequence, not relabeled, converging uniformly on $\mathcal{P}$ to a finite function $F^\star$. A uniform limit of finite convex functions is convex. Fix a report q. By (27), the sequence $\widetilde{f}_n(q)$ has a convergent subsequence; write one of its cluster points as $f^\star(q)$. This subsequence may depend on q, but the uniform convergence of $\widetilde{F}_n$ is preserved along every subsequence. Writing $D_n = D_{\widetilde{F}_n, \widetilde{f}_n}$ and $D_\star = D_{F^\star, f^\star}$, along this report-specific subsequence

$$\sup_{p \in \mathcal{P}} |D_n(p, q) - D_\star(p, q)| \le 2 \|\widetilde{F}_n - F^\star\|_\infty + \operatorname{diam}(\mathcal{P}) \|\widetilde{f}_n(q) - f^\star(q)\|_2 \longrightarrow 0. \tag{28}$$

Hence, for every p, passage to the limit in (21) gives

$$T(p, q) \le F^\star(p) - F^\star(q) - \langle f^\star(q), p - q \rangle \le \kappa^\star T(p, q). \tag{29}$$

The lower inequality and $T \ge 0$ imply $f^\star(q) \in \partial F^\star(q)$. Choose one such cluster point for every q. Then $(F^\star, f^\star)$ is feasible and has distortion at most $\kappa^\star$; equality follows from the definition of the infimum.

Finally, suppose $T \ge \varepsilon D_{F_0}$. The lower inequality in (29) and Lemma 3.1 show that $F^\star - \varepsilon F_0$ is convex. Because F_0 is strictly convex and $\varepsilon > 0$, $F^\star$ is strictly convex.

The proof gives pointwise attainment. Extending it to a measurable-selection class requires a separate closure argument. The explicit finite-action constructions used elsewhere in the paper admit measurable selections.

Why the positive floor is structural for finite actions. The finitely many action-indifference hyperplanes partition $\mathcal{P}$ into finitely many relatively open faces. On one such face, every task has a constant optimal-action set. Hence for any two reports p, q in that face, the action selected at q is also optimal at p, so every selected task regret and their maximum $T_0(p, q)$ vanish. If $D_{F,f} \le \kappa T_0$, then $D_{F,f}(p, q) = 0$ within the face. Fixing q shows that F equals its selected affine support throughout each full-dimensional chamber. Thus F is affine there: the finite arrangement cannot separate the reports inside an infinite report region. The term εD_{F_0} prevents this zero-face degeneracy.

E FACET CHARGES

E.1 PROOF OF LEMMA 3.2

By Lemma 3.1, $W := F - \lambda G$ is convex. Convex functions have positive-semidefinite matrix-valued distributional Hessians, so

$$D^2F - \lambda D^2G = D^2W \succeq 0. \tag{30}$$

On the two cells adjacent to the relative-open facet H, the gradient of G equals g^- and g^+. The distributional derivative of this piecewise constant gradient field across H is

$$(g^+ - g^-)n_H^\top \mathscr{H}^{d-1}|_H.$$

Continuity of G along the facet makes the gradient jump normal to H, and convexity orients it as in (9). Therefore

$$D^2G|_H = \alpha_H n_H n_H^\top \mathscr{H}^{d-1}|_H. \tag{31}$$

Combining (30) and (31) proves the measure inequality.

For the smoothness conclusion, restrict the functions to almost any short line normal to H through an interior facet point. The directional derivative of G jumps upward by α_H. If F is C^1, its directional derivative has no jump. The directional derivative of $F - \lambda G$ would therefore jump by $-\lambda\alpha_H$, contradicting one-dimensional monotonicity of the derivative of a convex function.

E.2 PROOF OF THEOREM 3.6

Fix a knot c. For task j, the selected slope of the weighted envelope $\lambda_j G_j$ lies $\ell_{j,c}$ above its left derivative and $r_{j,c}$ below its right derivative. By construction, the atom of $H_\vee - \lambda_j G_j$ at c has mass

$$(\ell_c^\vee - \ell_{j,c}) + (r_c^\vee - r_{j,c}) \geq 0.$$

The selected difference $h_\vee(c) - \lambda_j g_j(c)$ lies $\ell_c^\vee - \ell_{j,c}$ above the left derivative of this difference and $r_c^\vee - r_{j,c}$ below its right derivative. It is therefore a compatible subgradient. Away from the knots, the distributional curvature difference is zero and the unique derivatives are compatible. Hence $H_\vee - \lambda_j G_j$ is convex with this compatible selection, and Lemma 3.1 proves (14).

Now let (H, h) be another common selected majorant. At c, write the atom of D^2H as $\beta_c\delta_c$ and write

$$h(c) = H'_-(c) + u_c, \qquad 0 \leq u_c \leq \beta_c.$$

Applying Lemma 3.1 to every weighted task forces

$$u_c \geq \ell_{j,c}, \qquad \beta_c - u_c \geq r_{j,c} \quad \text{for every } j.$$

Thus $u_c \geq \ell_c^\vee$ and $\beta_c - u_c \geq r_c^\vee$. The residual atom of $H - H_\vee$ is nonnegative, and all nonatomic curvature of H remains nonnegative. Moreover, the selected difference $h(c) - h_\vee(c)$ lies $u_c - \ell_c^\vee$ above the residual left derivative and $(\beta_c - u_c) - r_c^\vee$ below the residual right derivative. The difference is convex with a compatible selection. A final application of Lemma 3.1 gives $D_{H,h} \geq D_{H_\vee,h_\vee}$.

If (F, f) is feasible for T_ε, then $F - \varepsilon F_0$ is convex because its selected divergence is nonnegative. It is also a common selected majorant of the weighted task divergences. Therefore

$$D_{F,f} \geq \varepsilon D_{F_0} + D_{H_\vee,h_\vee}.$$

The right side is itself feasible. It is pointwise least, so its supremum ratio is the minimum distortion, proving (15).

E.3 SELECTION-ROBUST BINARY JOIN

Suppose $\mathcal{P} = [a, b]$ and every task envelope is convex continuous piecewise linear (CPWL). For every report q, let $\mathcal{S}_j(q)$ be the finite set of slopes induced by task j's Bayes-optimal actions and define

$$\overline{R}_j(p, q) = \max_{s \in \mathcal{S}_j(q)} D_{G_j,s}(p, q).$$

For each interior task knot c, and for the domain endpoints a, b, set

$$A_c = \max_j \lambda_j \{G'_{j,+}(c) - G'_{j,-}(c)\},$$
$$B_a = \max_{j, s \in \mathcal{S}_j(a)} \lambda_j \{G'_{j,+}(a) - s\}, \qquad B_b = \max_{j, s \in \mathcal{S}_j(b)} \lambda_j \{s - G'_{j,-}(b)\}.$$

Theorem E.1 (Exact all-ties binary join)**.** *Modulo an affine term, define $D^2 \overline{H}_\vee = \sum_{c \in C} 2A_c \delta_c$. Select the midpoint slope at each knot and the derivative elsewhere in (a, b). At the endpoints select $\overline{h}_\vee(a) = \overline{H}'_{\vee,+}(a) - B_a$ and $\overline{h}_\vee(b) = \overline{H}'_{\vee,-}(b) + B_b$. Its selected divergence dominates $\lambda_j D_{G_j,s}$ for every task, report, and $s \in \mathcal{S}_j(q)$, and every selected convex pair with this simultaneous domination property pointwise dominates it. Consequently, with $\overline{T}_\varepsilon = \varepsilon D_{F_0} + \max_j \lambda_j \overline{R}_j$, the exact robust binary distortion is*

$$\sup_{p \neq q} \frac{\varepsilon D_{F_0}(p, q) + D_{\overline{H}_\vee, \overline{h}_\vee}(p, q)}{\overline{T}_\varepsilon(p, q)}.$$

Proof. Fix an interior task knot c. Every deterministic action that is Bayes optimal at c supplies a slope in $[G'_{j,-}(c), G'_{j,+}(c)]$. The two endpoint slopes are supplied by the actions whose affine pieces are active immediately to the left and right of c, so both endpoints are attained. For any active slope s, its weighted left and right directional charges are

$$\ell_{j,c}(s) = \lambda_j \{s - G'_{j,-}(c)\}, \qquad r_{j,c}(s) = \lambda_j \{G'_{j,+}(c) - s\}.$$

They lie in $[0, \alpha_{j,c}]$, where $\alpha_{j,c} = \lambda_j \{G'_{j,+}(c) - G'_{j,-}(c)\}$. Because the endpoint slopes are active,

$$\max_{j,s} \ell_{j,c}(s) = \max_{j,s} r_{j,c}(s) = A_c.$$

The candidate atom $2A_c\delta_c$ and midpoint selection therefore leave, after subtracting any weighted task with any active slope, a nonnegative residual atom. The selected residual lies $A_c - \ell_{j,c}(s)$ above its left derivative and $A_c - r_{j,c}(s)$ below its right derivative. Both quantities are nonnegative.

At the left endpoint, subtracting a weighted task selection s from $\overline{h}_\vee(a) = \overline{H}'_{\vee,+}(a) - B_a$ leaves a selected residual that lies

$$B_a - \lambda_j \{G'_{j,+}(a) - s\} \geq 0$$

below its right derivative. It is therefore a valid relative subgradient on $[a, b]$. The analogous gap at b is $B_b - \lambda_j \{s - G'_{j,-}(b)\} \geq 0$ above the residual left derivative. Because the finite action sets induce a finite union of knots, these atomwise and endpoint inequalities define a globally convex residual. For every global deterministic Bayes selection, its selected residual lies in the residual subdifferential at every knot and endpoint and equals the unique derivative elsewhere. Lemma 3.1 therefore proves simultaneous domination.

Conversely, let (H, h) dominate every active deterministic selection. At c, write its curvature atom as $\beta_c \delta_c$ and its selection as $H'_-(c) + u_c$. Domination of a task attaining A_c with its right endpoint slope forces $u_c \geq A_c$; domination of its left endpoint slope forces $\beta_c - u_c \geq A_c$. The same maximizing task supplies both witnesses, so the argument does not depend on whether other tasks separately maximize the two directional requirements. Thus $\beta_c \geq 2A_c$.

At a, write $h(a) = H'_+(a) - v_a$ with $v_a \geq 0$. Domination of every active endpoint selection forces $v_a \geq B_a$. Likewise, writing $h(b) = H'_-(b) + v_b$ forces $v_b \geq B_b$. Subtracting the candidate therefore leaves nonnegative curvature and compatible selections at every interior knot and both endpoints. Another application of Lemma 3.1 proves pointwise leastness.

If (F, f) is feasible for $\overline{T}_\varepsilon$, then

$$D_{F - \varepsilon F_0,\, f - \varepsilon \nabla F_0} \geq \max_j \lambda_j \overline{R}_j,$$

so the residual pair simultaneously majorizes every active task selection. Pointwise leastness gives

$$D_{F,f} \geq \varepsilon D_{F_0} + D_{\overline{H}_\vee, \overline{h}_\vee}.$$

The pair on the right is itself feasible; hence its supremum ratio is a lower bound for every feasible distortion and is attained by that pair.

Finally, finite action sets make the active-set partition and $\mathcal{S}_j(q)$ finite and Borel. Thus $\overline{R}_j$ and the selected candidate are Borel measurable; compactness and bounded losses make the robust regrets bounded. Every term vanishes when $p = q$, whereas strict convexity of F_0 gives $\overline{T}_\varepsilon(p, q) > 0$ for $p \neq q$. The displayed ratio therefore excludes only the diagonal $0/0$ cases.

E.4 Proof of Proposition 4.3

On a cell where affine piece a is uniquely active, $J(q) = b_a + \langle s_a, q\rangle$. Substitution into the selected-score definition gives

$$\begin{aligned} S_{F,f}(q, y) &= -B(q) - J(q) - \langle \nabla B(q) + s_a, e_y - q\rangle \\ &= S_{B,\nabla B}(q, y) - b_a - \langle s_a, e_y\rangle. \end{aligned}$$

Differentiating on the affine hull yields $\nabla_q S_{F,f}(q, y) = \nabla^2 B(q)(q - e_y)$. In particular, the max-affine increment contributes no ordinary gradient inside the cell. At a facet tie, the scores selected from the two sides satisfy

$$S_{J,s_+}(q_0, y) - S_{J,s_-}(q_0, y) = -\alpha\langle n, e_y - q_0\rangle,$$

so the hard task information is instead carried by a score-vector jump.

For the smoothing calculation, put $\bar{\phi} = (\phi_- + \phi_+)/2$ and $z(q) = \langle n, q - q_0\rangle$. Since both affine pieces agree at q_0 and $s_+ - s_- = \alpha n$, their difference is $\phi_+(q) - \phi_-(q) = \alpha z(q)$. Therefore

$$J_\tau(q) = \bar{\phi}(q) + \tau \log\left(2\cosh\frac{\alpha z(q)}{2\tau}\right). \tag{32}$$

At the tie, $\nabla J_\tau(q_0) = (s_- + s_+)/2$. Hence, for $p_t = q_0 + tn$,

$$\begin{aligned} D_{J_\tau}(p_t, q_0) &= \tau \log\cosh\left(\frac{\alpha t}{2\tau}\right), \\ D_{J,s_-}(p_t, q_0) &= \phi_+(p_t) - \phi_-(q_0) - \langle s_-, p_t - q_0\rangle = \alpha t. \end{aligned}$$

Differentiating (32) twice at q_0 gives $\nabla^2 J_\tau(q_0) = \alpha^2 nn^\top/(4\tau)$. Finally, $\log\cosh u = u^2/2 + o(u^2)$ as $u \to 0$, whereas $\log\cosh u = u - \log 2 + o(1)$ as $u \to \infty$. These expansions give the two limits stated in Proposition 4.3; the latter also gives

$$\lim_{\tau\downarrow 0}\left(D_{J,s_-} - D_{J_\tau}\right)(p_t, q_0) = \frac{\alpha t}{2}.$$

Bregman divergences add under sums of generators, so adding the same differentiable baseline to the hard and smoothed increments leaves this additive gap unchanged.

E.5 Sharp one-facet smoothing frontier

Theorem E.2 (Sharp facet smoothness–slack frontier)**.** *Let $\alpha, L, R > 0$, $G(t) = \alpha t_+$, select the left slope at zero, and let J be convex and differentiable with L-Lipschitz derivative on an interval containing $[0, R]$. If*

$$D_J(t, 0) \geq D_G(t, 0) - \delta \quad \textit{for all } t \in [0, R], \tag{33}$$

then

$$\delta \geq \max_{0\leq t\leq R}\left\{\alpha t - \frac{L}{2}t^2\right\}. \tag{34}$$

If $R \geq \alpha/L$, this gives the sharp bound

$$\delta \geq \frac{\alpha^2}{2L}. \tag{35}$$

The bound is attained by the Moreau/Huber hinge: flat to the left of zero, quadratic with curvature L until α/L, and affine afterward.

Proof. Bregman divergences are invariant to affine terms, so normalize $J(0) = 0$ and $J'(0) = 0$. The L-Lipschitz derivative gives, for $t \geq 0$,

$$D_J(t, 0) = J(t) - J(0) - J'(0)t = \int_0^t \{J'(s) - J'(0)\}\, ds \leq \int_0^t Ls\, ds = \frac{L}{2}t^2.$$

Since the frozen left selection of $G(t) = \alpha t_+$ gives $D_G(t, 0) = \alpha t$, condition (33) implies

$$\delta \geq \alpha t - \frac{L}{2}t^2 \quad \text{for every } t \in [0, R].$$

The concave quadratic is maximized at R when $R < \alpha/L$ and at α/L otherwise, proving (34) and (35).

For sharpness, define, up to an affine term,

$$J_L(t) = \begin{cases} 0, & t \leq 0, \\ Lt^2/2, & 0 \leq t \leq \alpha/L, \\ \alpha t - \alpha^2/(2L), & t \geq \alpha/L. \end{cases}$$

Its derivative is 0, then Lt, then α. Hence J_L is convex, continuously differentiable, and has L-Lipschitz derivative. Direct substitution shows that $\sup_{t\in[0,R]}\{D_G(t, 0) - D_{J_L}(t, 0)\}$ equals the right side of (34). A common differentiable baseline adds the same divergence to both sides and cancels.

E.6 Margin-localized smoothing

Write $\phi_a(q) = b_a + \langle s_a, q\rangle$, $J = \max_{a\leq\nu} \phi_a$, and $J_\tau = \tau \log \sum_{a\leq\nu} e^{\phi_a/\tau}$, $\tau > 0$. Fix a measurable $a(q) \in \arg\max_a \phi_a(q)$ and set $j(q) = s_{a(q)}$. For a differentiable convex baseline B, write $D_{B+J} := D_{B+J,\nabla B+j}$. Let $R_{\mathcal{P}} = \sup_{p,q\in\mathcal{P}} \|p - q\|$, and $L_J = \max_{a,b} \|s_a - s_b\|$. Let $\Delta_J(q)$ be the gap between the unique active piece and the runner-up, with value zero at a tie.

Proposition E.3 (Margin-localized smoothing)**.** *For $\delta > 0$, set*

$$\eta_{\tau,\delta} = \frac{(\nu - 1)e^{-\delta/\tau}}{1 + (\nu - 1)e^{-\delta/\tau}}, \qquad \mathfrak{a}_{\tau,\delta} = \tau \log \nu + R_{\mathcal{P}} L_J \eta_{\tau,\delta}. \tag{36}$$

If $\Delta_J(q) \geq \delta$, then

$$\|\nabla J_\tau(q) - s_{a(q)}\| \leq L_J \eta_{\tau,\delta}, \qquad \|\nabla^2 J_\tau(q)\|_{\mathrm{op}} \leq \frac{L_J^2}{\tau}\eta_{\tau,\delta}, \tag{37}$$

and, uniformly in $p \in \mathcal{P}$,

$$\left|D_{J_\tau}(p, q) - D_{J,s_{a(q)}}(p, q)\right| \leq \mathfrak{a}_{\tau,\delta}. \tag{38}$$

The bound also holds with a common differentiable baseline. For any random pair $(U, V) \in \mathcal{P}^2$, set $\pi_\delta = \Pr\{\Delta_J(V) < \delta\}$. Then

$$\mathbb{E}|D_{B+J_\tau}(U, V) - D_{B+J}(U, V)| \leq (1 - \pi_\delta)\mathfrak{a}_{\tau,\delta} + \pi_\delta(\tau \log \nu + R_{\mathcal{P}} L_J). \tag{39}$$

If $T \leq D_{B+J} \leq \kappa T$, then on any declared pair set satisfying $\Delta_J(q) \geq \delta$ and $T(p, q) \geq r > \mathfrak{a}_{\tau,\delta}$,

$$\left(1 - \frac{\mathfrak{a}_{\tau,\delta}}{r}\right) T \leq D_{B+J_\tau} \leq \left(\kappa + \frac{\mathfrak{a}_{\tau,\delta}}{r}\right) T. \tag{40}$$

Proof. Let

$$w_a(q) = \frac{e^{\phi_a(q)/\tau}}{\sum_{b=1}^{\nu} e^{\phi_b(q)/\tau}}$$

be the softmax weights. If $\Delta_J(q) \geq \delta$ and $a(q)$ is the unique active piece, then

$$\sum_{b\neq a(q)} e^{(\phi_b(q) - \phi_{a(q)}(q))/\tau} \leq (\nu - 1)e^{-\delta/\tau}.$$

The total weight on inactive pieces is therefore at most $\eta_{\tau,\delta}$. Since $\nabla J_\tau(q) = \sum_a w_a(q) s_a$,

$$\|\nabla J_\tau(q) - s_{a(q)}\| \le L_J \sum_{b \neq a(q)} w_b(q) \le L_J \eta_{\tau,\delta}.$$

The log-sum-exp Hessian is the softmax covariance

$$\nabla^2 J_\tau(q) = \frac{1}{\tau} \sum_a w_a(q) \big(s_a - \bar{s}(q)\big) \big(s_a - \bar{s}(q)\big)^\top, \qquad \bar{s}(q) = \sum_a w_a(q) s_a.$$

This covariance is dominated by the corresponding second moment around $s_{a(q)}$. Hence

$$\|\nabla^2 J_\tau(q)\|_{\mathrm{op}} \le \frac{L_J^2}{\tau} \sum_{b \neq a(q)} w_b(q) \le \frac{L_J^2}{\tau} \eta_{\tau,\delta},$$

which proves (37).

Set $E_\tau = J_\tau - J$. The finite log-sum-exp inequality gives $0 \le E_\tau \le \tau \log \nu$. At a report with unique active piece,

$$\begin{aligned} D_{J_\tau}(p,q) - D_{J,s_{a(q)}}(p,q) = E_\tau(p) - E_\tau(q) \\ - \langle \nabla J_\tau(q) - s_{a(q)}, p - q \rangle. \end{aligned}$$

The first difference has absolute value at most $\tau \log \nu$; the second is at most $R_{\mathcal{P}} L_J \eta_{\tau,\delta}$. This proves (38). A common differentiable baseline adds the same Bregman divergence to the hard and smoothed scores and therefore cancels from their difference.

At an arbitrary report, including a tie, the softmax gradient is a convex combination of the affine slopes. Its distance from any selected active slope is at most L_J. The resulting global bound is $\tau \log \nu + R_{\mathcal{P}} L_J$. Splitting the expectation according to $\{\Delta_J(V) \ge \delta\}$ gives (39).

Finally, on the declared restricted pair set,

$$\begin{aligned} D_{B+J_\tau} &\ge D_{B+J} - \mathfrak{a}_{\tau,\delta} \ge T - \mathfrak{a}_{\tau,\delta} \ge \Big(1 - \frac{\mathfrak{a}_{\tau,\delta}}{r}\Big) T, \\ D_{B+J_\tau} &\le D_{B+J} + \mathfrak{a}_{\tau,\delta} \le \kappa T + \mathfrak{a}_{\tau,\delta} \le \Big(\kappa + \frac{\mathfrak{a}_{\tau,\delta}}{r}\Big) T. \end{aligned}$$

This is (40). When $\mathfrak{a}_{\tau,\delta} < r$, scaling the smoothed entropy by $(1 - \mathfrak{a}_{\tau,\delta}/r)^{-1}$ gives the stated restricted majorant and upper factor.

Duplicate affine pieces. If the distinct affine restrictions are ψ_u with multiplicities n_u, then the literal log-sum-exp is $\tau \log \sum_u n_u e^{\psi_u/\tau}$. Weighting each original exponential by $1/n_u$ instead yields $\tau \log \sum_u e^{\psi_u/\tau}$. Equal multiplicities add only a constant; unequal multiplicities generally change the divergence. Thus duplicate pieces must be accounted for when implementing the smoothing construction.

F TRANSVERSAL FINITE-ACTION TASKS

F.1 FINITE-ACTION REALIZATION

For $s \in \{-1, 1\}$, define

$$g_{j,s} = s v_j - s\langle v_j, p_0 \rangle \mathbf{1}, \qquad \ell_{j,s} = -g_{j,s}.$$

Because $\langle p, \mathbf{1} \rangle = 1$ on the simplex,

$$-\langle p, \ell_{j,s} \rangle = \langle p, g_{j,s} \rangle = s \langle v_j, p - p_0 \rangle.$$

Taking the maximum over the two actions gives $G_j(p) = |\langle v_j, p - p_0 \rangle|$. The losses are finite vectors and hence bounded on the finite outcome space. Its tie set is the hyperplane $H_j := \{p \in \mathcal{P} : \langle v_j, p - p_0 \rangle = 0\}$; away from H_j, the maximizing action is unique.

F.2 Proof of Theorem 3.3

Proposition 2.3 gives $\kappa^\star_\varepsilon \le m$. We prove the reverse inequality. Let (F, f) be any feasible selected scoring pair. Since

$$T_\varepsilon \ge \varepsilon D_{F_0} + D_{G_j}$$

for each j, Lemma 3.1 makes $F - \varepsilon F_0 - G_j$ convex. Here $\mathcal{P} = \Delta_K$ and $d = \dim \operatorname{aff}(\mathcal{P}) = K - 1$. Choose a relatively open Euclidean ball U around p_0 whose closure lies in $\operatorname{ri}(\mathcal{P})$, and define on U the matrix-valued measures

$$M = D^2 F - \varepsilon D^2 F_0, \qquad M_j = D^2 G_j. \tag{41}$$

Then

$$M - M_j \succeq 0 \qquad \text{for every } j. \tag{42}$$

The measure M_j equals $2 v_j v_j^\top \mathscr{H}^{d-1}|_{H_j \cap U}$. For

$$E_j = (H_j \cap U) \setminus \bigcup_{i \ne j} H_i, \qquad E_0 = U \setminus \bigcup_{j=1}^m E_j, \tag{43}$$

the sets $E_0, E_1, \dots, E_m$ form a measurable partition. Pairwise hyperplane intersections have codimension two, so M_j is carried by E_j.

We verify the sum inequality on an arbitrary Borel set $A \subset U$. On $A \cap E_j$, (42) gives

$$M(A \cap E_j) \succeq M_j(A \cap E_j).$$

On E_0, every M_j vanishes. Moreover M is positive semidefinite because $M = (M - M_1) + M_1$. Therefore

$$\begin{aligned} M(A) &= M(A \cap E_0) + \sum_{j=1}^m M(A \cap E_j) \\ &\succeq \sum_{j=1}^m M_j(A \cap E_j) = \sum_{j=1}^m M_j(A). \end{aligned}$$

Thus

$$D^2 \left(F - \varepsilon F_0 - \sum_{j=1}^m G_j \right) \succeq 0 \quad \text{on } U. \tag{44}$$

The distributional Hessian criterion implies that

$$H := F - \varepsilon F_0 - \sum_{j=1}^m G_j \tag{45}$$

is convex on U.

Choose $t > 0$ small enough that the segment between q_t and p_t in (10) lies in U. At q_t, every G_j is differentiable. Since F_0 is differentiable, the exact subgradient sum rule applied to (45) gives

$$f(q_t) - \varepsilon \nabla F_0(q_t) - \sum_j \nabla G_j(q_t) \in \partial H(q_t).$$

The corresponding Bregman divergence is nonnegative, so

$$D_{F,f}(p_t, q_t) \ge \varepsilon D_{F_0}(p_t, q_t) + \sum_{j=1}^m D_{G_j}(p_t, q_t). \tag{46}$$

For every j, $\langle v_j, q_t - p_0 \rangle = -t$ and $\langle v_j, p_t - p_0 \rangle = t$. The selected gradient at q_t points toward the negative branch, and direct substitution gives

$$D_{G_j}(p_t, q_t) = 2t. \tag{47}$$

Local C^1 regularity of F_0 yields

$$D_{F_0}(p_t, q_t) = o(t). \tag{48}$$

For example, write the divergence as the line integral of $\nabla F_0 - \nabla F_0(q_t)$ along the segment and use continuity of the gradient at p_0.

Combining (46)–(48),

$$\begin{aligned}\liminf_{t\downarrow 0} \frac{D_{F,f}(p_t, q_t)}{T_\varepsilon(p_t, q_t)} &\geq \lim_{t\downarrow 0} \frac{\varepsilon D_{F_0}(p_t, q_t) + 2mt}{\varepsilon D_{F_0}(p_t, q_t) + 2t} \\ &= m.\end{aligned}$$

Every feasible score therefore has distortion at least m, which matches the sum upper bound.

G Independent shared components

G.1 Proof of Theorem 3.5

For each task use the finite sign-action set $\{-1, 1\}^L$. Its negative loss vector for action s is

$$w_{j,s} = \sum_\ell a_{j\ell} s_\ell \{v_\ell - \langle v_\ell, p_0\rangle \mathbf{1}\}.$$

On the simplex its support envelope equals G_j. Thus these are bounded finite-action tasks, and common component signs specify valid Bayes decoders. The constant affine corrections have zero pairing with $p - q$.

For any report pair, put $b = \varepsilon D_{F_0}(p, q)$ and $r_\ell = D_{H_\ell, h_\ell}(p, q)$. Common selections give

$$D_{G_j, g_j}(p, q) = (Ar)_j, \qquad D_{F_c, f_c}(p, q) = b + c^\top r, \quad F_c = \varepsilon F_0 + \sum_\ell c_\ell H_\ell.$$

Here $f_c = \varepsilon \nabla F_0 + \sum_\ell c_\ell h_\ell$. Write $C = \sup_{r \geq 0, r \neq 0} c^\top r / \max_j (Ar)_j$. The denominator is strictly positive for every nonzero $r \geq 0$, since each $c_\ell > 0$. Coordinatewise domination gives $C \geq 1$, and

$$b + \max_j (Ar)_j \leq b + c^\top r \leq C\{b + \max_j (Ar)_j\}.$$

For $r = 0$ the same inequalities hold directly. This proves global feasibility and the upper bound, including ties and domain boundaries. The baseline makes F_c strictly convex.

For the converse, let (F, f) be any feasible scoring pair. Lemma 3.1 implies that $F - \varepsilon F_0 - G_j$ is convex for every j. On an interior ball U about p_0, the positive-semidefinite matrix-valued measures $M = D^2(F - \varepsilon F_0)$ and $\mu_\ell = D^2 H_\ell$ satisfy

$$M \succeq \sum_\ell a_{j\ell} \mu_\ell \quad \text{for every } j.$$

Distinct component hyperplanes intersect in sets of zero facet measure. The measures μ_ℓ are therefore mutually singular. Restrict to each hyperplane after removing all other component hyperplanes, and choose a row attaining c_ℓ there. On the remaining set use $M \succeq 0$. Adding these restrictions gives $M \succeq \sum_\ell c_\ell \mu_\ell$. The distributional Hessian criterion makes $F - F_c$ convex on U. At a report q off every component hyperplane, F_c is differentiable. The subgradient sum rule consequently gives

$$f(q) - \nabla F_c(q) \in \partial(F - F_c)(q), \qquad D_{F,f}(p, q) \geq D_{F_c, f_c}(p, q) \quad (p, q \in U).$$

This local comparison at off-tie reports supplies the lower bound below.

Given $r \in (0, \infty)^L$, linear independence provides a tangent vector w satisfying $\langle v_\ell, w\rangle = r_\ell/2$. For small $t > 0$, let $q_t = p_0 - tw$ and $p_t = p_0 + tw$. They lie in U, avoid all component ties, and obey $D_{H_\ell}(p_t, q_t) = t r_\ell$. Since the baseline is C^1, $D_{F_0}(p_t, q_t) = o(t)$. Therefore

$$\liminf_{t\downarrow 0} \frac{D_{F,f}(p_t, q_t)}{T_\varepsilon(p_t, q_t)} \geq \frac{c^\top r}{\max_j (Ar)_j}.$$

Positive vectors approximate all nonzero nonnegative vectors in this continuous homogeneous ratio. Taking the supremum proves the lower bound. Homogeneity yields the packing LP in (11); its feasible set is compact since $r_\ell \leq 1/c_\ell$. Standard LP duality yields the covering LP and completes the proof.

G.2 Consequences and examples

The factor lies in $[1, \min(m, L)]$: use $c^\top r \le \sum_j (Ar)_j$ for the first upper bound and $c_\ell r_\ell \le 1$ for the second. It equals one precisely when a row of A equals c. Otherwise a strictly positive r makes $c^\top r$ strictly greater than every $(Ar)_j$. Duplicating task rows adds redundant packing constraints, preserving the feasible set and optimal value.

For two rows $(1, \alpha), (\alpha, 1)$, summing the constraints gives $(1+\alpha)(r_1 + r_2) \le 2$. The point $r_1 = r_2 = 1/(1+\alpha)$ attains the bound, proving the factor $2/(1+\alpha)$. This requires at least three outcomes. The literal task sum has factor two throughout: its upper bound is Proposition 2.3, and balanced off-tie crossings have equal task regrets and approach that factor. For rows $(1, 1, 0), (1, 0, 1), (0, 1, 1)$, summing gives $2(r_1 + r_2 + r_3) \le 3$, attained at $r = (1/2, 1/2, 1/2)$. The exact factor is $3/2$; this example requires at least four outcomes.

For the attaining score, the lower ratio endpoint is one: choose distinct reports in one common chamber, where all task regrets vanish and only the positive baseline remains. Theorem 2.2 thus gives the same covering value as its optimally scaled worst-class projection factor. For a fixed population and class, achieved risk is governed by the separate factor hierarchy in Proposition 4.2.

Upper-bound scope. The construction also gives a bound for any finite convex component functions H_ℓ on the same domain, with selected subgradients h_ℓ, $G_j = \sum_\ell a_{j\ell} H_\ell$, and $g_j = \sum_\ell a_{j\ell} h_\ell$. Keep $A \ge 0$, $c_\ell = \max_j a_{j\ell} > 0$, and the standing baseline. Let $\kappa_{\mathrm{cov}}(A) = \max\{c^\top r : r \ge 0, Ar \le \mathbf{1}\}$. For $r_\ell = D_{H_\ell, h_\ell}(p, q) \ge 0$ and $b = \varepsilon D_{F_0}(p, q)$,

$$b + \max_j (Ar)_j \le b + c^\top r \le \kappa_{\mathrm{cov}}(A)\{b + \max_j (Ar)_j\}.$$

Thus the selected score with entropy $\varepsilon F_0 + \sum_\ell c_\ell H_\ell$ is strictly proper and has this upper factor, without independence or a common crossing. Exactness can fail: if $H_1 = H_2 = H$ and $A = I_2$, the covering value is two, whereas the identical-task target is the divergence of $\varepsilon F_0 + H$ and has optimal factor one. The general statement concerns selected divergences; its finite-action interpretation uses CPWL components and compatible task-action selections.

Each assumption has a specific role. Common tie selections permit a shared component construction; opposite binary choices require the directional charges of Theorem 3.6. Independence realizes arbitrary positive regret proportions at the common interior crossing. General CPWL arrangements require further integrability arguments, and selected leastness at joint ties remains separate from the distortion optimum. The four-class weather construction has four components in a three-dimensional tangent space and is analyzed separately as a feasible majorant.

H Quadratic tasks and ellipsoidal distortion

H.1 Ellipsoid characterization

On a compact convex affine chart $\mathcal{Z} \subset \mathbb{R}^d$ with nonempty interior, let

$$F_0(z) = \tfrac{1}{2} z^\top A_0 z, \quad A_0 \succ 0, \qquad G_j(z) = \tfrac{1}{2} z^\top A_j z, \quad A_j \succeq 0, \tag{49}$$

and define

$$Q(v) = \varepsilon v^\top A_0 v + \max_j \lambda_j v^\top A_j v, \qquad \mathcal{K} = \{v : Q(v) \le 1\}, \tag{50}$$

with $E_M = \{v : v^\top M v \le 1\}$.

Theorem H.1 (Quadratic ellipsoid characterization)**.** *For* (49)*,*

$$\kappa^\star_\varepsilon = \inf\{\kappa \ge 1 : \exists M \succ 0,\ E_M \subseteq \mathcal{K} \subseteq \sqrt{\kappa} E_M\}. \tag{51}$$

The optimizing quadratic entropy attains the optimum over all finite convex entropies. If $r = \dim \operatorname{span}_j \operatorname{range}(A_j) \ge 1$*, then*

$$\kappa^\star_\varepsilon \le \min\{m, r\}. \tag{52}$$

If $r = 0$*, all task matrices vanish and* $\kappa^\star_\varepsilon = 1$*. For* $d = m$*,* $A_0 = I_m$*,* $A_j = e_j e_j^\top$*, and unit weights,*

$$\kappa^\star_\varepsilon = \frac{m(1+\varepsilon)}{1 + m\varepsilon}. \tag{53}$$

Proof. For two chart points $z, z' \in \mathcal{Z}$, write $v = z - z'$. Their target divergence is $Q(v)/2$, where Q is defined in (50). Initially the attainable displacements lie in $\mathcal{Z} - \mathcal{Z}$, which contains a neighborhood of the origin because $\mathcal{Z}$ has nonempty interior. The quadratic inequalities below are homogeneous, so validity on that neighborhood extends them to every $v \in \mathbb{R}^d$. The quadratic entropy $F_M(z) = z^\top Mz/2$ has divergence $v^\top Mv/2$. Hence its global sandwich is exactly

$$Q(v) \le v^\top Mv \le \kappa Q(v) \qquad \text{for every } v. \tag{54}$$

If $E_M \subseteq \mathcal{K}$ and $v \neq 0$, scale v to the boundary of E_M to obtain the lower inequality. Conversely, the lower inequality sends every point of E_M into $\mathcal{K}$. The same homogeneity shows that the upper inequality is equivalent to $\mathcal{K} \subseteq \sqrt{\kappa} E_M$. Thus the best quadratic factor is the right side of (51). It is attained: the positive baseline bounds the eigenvalues of every feasible normalized M above and below once a finite comparison factor is fixed, so a minimizing sequence has a convergent positive-definite subsequence.

It remains to exclude an improvement by a nonquadratic entropy. Let (F, f) be any feasible pair with factor κ. A finite convex function is twice differentiable in the Alexandrov sense at almost every interior point. Choose such a point z_0, let $M = \nabla^2 F(z_0)$, and note that the selected subgradient is uniquely $f(z_0) = \nabla F(z_0)$. The single vector expansion is

$$F(z_0 + h) = F(z_0) + \langle \nabla F(z_0), h \rangle + \tfrac{1}{2} h^\top Mh + o(\|h\|_2^2).$$

Taking $h = tv$ therefore gives $D_F(z_0 + tv, z_0) = t^2 v^\top Mv/2 + o(t^2)$. The target equals $t^2 Q(v)/2$. Divide both sides of the global sandwich by $t^2/2$ and let $t \downarrow 0$. Equation (54) follows for this Hessian with the same κ; in particular $M \succeq \varepsilon A_0 \succ 0$. Therefore every feasible convex entropy has factor at least the best ellipsoidal factor. The optimizing quadratic entropy gives the reverse inequality globally.

For the task-sensitive dimension bound, put $u = A_0^{1/2} v$ and

$$B_j = A_0^{-1/2} (\lambda_j A_j) A_0^{-1/2}.$$

Let S be the span of their ranges. Its dimension is $r = \dim \operatorname{span}_j \operatorname{range}(A_j)$ because the whitening map is invertible. If $r = 0$, every A_j vanishes and the baseline itself has distortion one. Suppose henceforth that $r \ge 1$, and decompose $u = u_S + u_\perp$. Then

$$Q(v) = Q_S(u_S) + \varepsilon \|u_\perp\|_2^2, \qquad Q_S(w) = \varepsilon \|w\|_2^2 + \max_j w^\top B_j w.$$

The symmetric John theorem on S supplies $M_S \succ 0$ with

$$Q_S(w) \le w^\top M_S w \le r Q_S(w) \quad \text{for every } w \in S.$$

Extend it by $M_S \oplus \varepsilon I$ on $S \oplus S^\perp$. The lower quadratic sandwich is exact on the complement, and

$$u_S^\top M_S u_S + \varepsilon \|u_\perp\|^2 \le r Q_S(u_S) + \varepsilon \|u_\perp\|^2 \le r Q(v).$$

This proves the factor r; Proposition 2.3 supplies factor m.

For the orthogonal corollary, take $d = m$, $A_0 = I_m$, and $A_j = e_j e_j^\top$. The form $M = (1 + \varepsilon) I$ is feasible, and its ratio is

$$\frac{(1 + \varepsilon)\|v\|_2^2}{\varepsilon \|v\|_2^2 + \|v\|_\infty^2}.$$

Writing $s = \|v\|_2^2 / \|v\|_\infty^2 \in [1, m]$ gives the maximum $m(1 + \varepsilon)/(1 + m\varepsilon)$. Conversely, any local feasible Hessian satisfies $M_{jj} \ge 1 + \varepsilon$. Thus $\operatorname{tr} M \ge m(1 + \varepsilon)$, and averaging $u^\top Mu$ over $u \in \{-1, 1\}^m$ produces a sign direction with the matching ratio.

Finally, factor $A_j = C_j^\top C_j$. On the compact convex affine chart $\mathcal{Z}$,

$$\tfrac{1}{2} z^\top A_j z = \sup_{a \in \mathcal{A}_j} \{ \langle a, C_j z \rangle - \tfrac{1}{2} \|a\|_2^2 \}$$

for a sufficiently large compact action set $\mathcal{A}_j$. Each affine term is the negative expected loss of an action, with constants represented because probabilities sum to one. Compact action and chart ranges make the losses bounded.

I LEARNING CONSEQUENCES

I.1 PROOF OF COROLLARY 4.1

Conditional propriety gives

$$L_F(h) := \mathbb{E}S_{F,f}(q_h(X), Y) = C_F + \mathbb{E}_X D_{F,f}(p_X, q_h(X)), \qquad C_F = -\mathbb{E}_X F(p_X). \tag{55}$$

Since C_F is independent of h, population η-optimality and the sandwich imply $\mathcal{R}_T(\widetilde{h}_F) \le L_F(\widetilde{h}_F) - C_F \le L_F(h) - C_F + \eta \le \kappa\mathcal{R}_T(h) + \eta$ for every comparator h. Taking the infimum proves (16).

I.2 BOUNDARY-LAYER PROJECTION APPROXIMATION

Corollary I.1 (Boundary-layer approximation for smoothed scores)**.** *Let D_{hard} satisfy $T_\varepsilon \le D_{\mathrm{hard}} \le \kappa T_\varepsilon$, and suppose a proper smoothed divergence D_τ obeys*

$$|\mathbb{E}D_\tau(p_X, q_h(X)) - \mathbb{E}D_{\mathrm{hard}}(p_X, q_h(X))| \le \Delta_\tau(h) \quad (h \in \mathcal{H}).$$

If $\widehat{h}_\tau$ is η-optimal for population D_τ-risk, then for every comparator $h_T \in \mathcal{H}$,

$$\mathcal{R}_T(\widehat{h}_\tau) \le \kappa\mathcal{R}_T(h_T) + \eta + \Delta_\tau(\widehat{h}_\tau) + \Delta_\tau(h_T). \tag{56}$$

For the finite log-sum-exp construction, Proposition E.3 makes $\Delta_\tau(h)$ explicit: with ν affine pieces, domain diameter $R_{\mathcal{P}}$, and slope gap L_J,

$$\Delta_\tau(h) = (1 - \pi_\delta(h))\mathfrak{a}_{\tau,\delta} + \pi_\delta(h)(\tau \log \nu + R_{\mathcal{P}} L_J),$$

where the *occupancy* $\pi_\delta(h)$ is the report mass with affine-piece gap below δ, and $\mathfrak{a}_{\tau,\delta}$ is the off-layer smoothing error. This bound controls the hard–smooth objective discrepancy. Comparing fitted scores additionally requires their facet charges, signed action gaps, and projection locations.

Proof. Write $L_{\mathrm{hard}}(h) = \mathbb{E}D_{\mathrm{hard}}(p_X, q_h(X))$ and $L_\tau(h) = \mathbb{E}D_\tau(p_X, q_h(X))$. The discrepancy bound and population η-optimality give

$$\mathcal{R}_T(\widehat{h}_\tau) \le L_\tau(\widehat{h}_\tau) + \Delta_\tau(\widehat{h}_\tau) \le L_\tau(h_T) + \eta + \Delta_\tau(\widehat{h}_\tau).$$

Substitute $L_\tau(h_T) \le L_{\mathrm{hard}}(h_T) + \Delta_\tau(h_T) \le \kappa\mathcal{R}_T(h_T) + \Delta_\tau(h_T)$ to obtain (56). A uniform bound $\Delta_\tau(h) \le \overline{\Delta}_\tau$ gives total slack $2\overline{\Delta}_\tau$. The occupancy specialization follows from (39) by bounding absolute difference of expectations by expected absolute difference.

Standard finite-sample consequence. Fix a measurable selected score whose divergence satisfies $T_\varepsilon \le D_{F,f} \le \kappa T_\varepsilon$ and a report class before drawing the i.i.d. sample. Assume that the induced score class $\mathcal{S}_F$ is measurable and uniformly bounded in an interval of length B_F. If the score or report class is chosen using an independent preliminary sample, the result below holds conditionally on that sample; selection using the same sample is not covered without additional data-dependent complexity control. A smoothed score that does not itself satisfy this sandwich requires the preceding approximation bound, including its Δ_τ terms. Let $\mathfrak{R}_n(\mathcal{S}_F)$ be its expected Rademacher complexity and η_n the empirical optimization gap of $\widehat{h}_F$ for $\widehat{L}_F(h) = n^{-1}\sum_{i=1}^n S_{F,f}(q_h(X_i), Y_i)$. Standard uniform convergence gives, with probability at least $1 - \delta$, $\delta \in (0, 1)$,

$$\sup_h |L_F(h) - \widehat{L}_F(h)| \le 2\mathfrak{R}_n(\mathcal{S}_F) + B_F\sqrt{\frac{\log(2/\delta)}{2n}} =: U_n. \tag{57}$$

On this event, $L_F(\widehat{h}_F) \le \inf_h L_F(h) + \eta_n + 2U_n$. Applying Corollary 4.1 therefore yields

$$\mathcal{R}_T(\widehat{h}_F) \le \kappa \inf_{h\in\mathcal{H}} \mathcal{R}_T(h) + 4\mathfrak{R}_n(\mathcal{S}_F) + 2B_F\sqrt{\frac{\log(2/\delta)}{2n}} + \eta_n. \tag{58}$$

This is an i.i.d. learning bound for a fixed bounded score class, not a direct confidence guarantee for the weather experiment, whose observations have time-series dependence.

J POPULATION FACET FLUX

J.1 POPULATION FACET FLUX

Proposition J.1 (Population facet flux)**.** *Let $F = B + H$, where B is C^2 and strongly convex and $H = \max_a L_a$, $L_a(q) = b_a + \langle s_a, q\rangle$, is fixed finite max-affine. Under a common differentiable map $q_\beta(x)$, finite-sample gradients are baseline-only wherever all active cells remain fixed. For the population formula, assume the following on a parameter neighborhood: X has compact support in an open domain Ω and density ρ continuous near moving facets; the conditional truth $p(x) = p_x$ is continuous. The map $(\beta, x) \mapsto q_\beta(x)$ is C^2 on a neighborhood of that support. Active facet preimages form a finite piecewise C^1 partition; their closures satisfy a uniform positive lower bound on $\|\nabla_x g_{\beta,ab}\|$, and their junctions are transverse strata of codimension at least two, where $g_{\beta,ab} = L_a(q_\beta) - L_b(q_\beta)$. Let $M_{\beta,ab}$ be the preimage under q_β of the relative interior of the adjacent active facet $\{q : L_a(q) = L_b(q) = H(q)\}$, and count each oriented pair once. Fix a measurable $a_0(q) \in \arg\max_a L_a(q)$ and put $f_0(q) = \nabla B(q) + s_{a_0(q)}$, $\mathcal{L}_F(\beta) = \mathbb{E}S_{F,f_0}(q_\beta(X), Y)$. Then*

$$\nabla_\beta \mathcal{L}_F(\beta) = \mathbb{E}\big[J_\beta q_\beta(X)^\top \nabla^2 B(q_\beta(X))\{q_\beta(X) - p(X)\}\big] + \sum_{a\sim b} \int_{M_{\beta,ab}} \Delta c_{ab}(p(x)) \frac{\nabla_\beta g_{\beta,ab}(x)}{\|\nabla_x g_{\beta,ab}(x)\|} \rho(x)\, d\mathscr{H}^{d_x-1}(x), \tag{59}$$

where $\Delta c_{ab}(p) = -\langle s_a - s_b, p - q_\beta\rangle$ on the oriented facet, $d_x = \dim\Omega$, and $J_\beta q_\beta = \partial q_\beta/\partial\beta^\top$.

Proof. On a cell where a is uniquely active, substitution in the score representation gives

$$S_{F,f_0}(q, y) = S_{B,\nabla B}(q, y) - b_a - \langle s_a, e_y\rangle, \qquad \nabla_q S_{F,f_0}(q, y) = \nabla^2 B(q)(q - e_y). \tag{60}$$

If every fitted report in a finite sample has a positive active-cell margin, continuity leaves all active classes unchanged on a parameter neighborhood. The chain rule applied to (60) then yields only the common baseline gradient. This conclusion requires the same pointwise map $q_\beta(x)$; equality of the sets of attainable reports does not fix a parameter gradient.

For the population objective, work on a fixed local patch meeting only two adjacent active cells, with a smooth spatial cutoff χ. Orient them so that piece a is active when

$$g_{\beta,ab}(x) = L_a(q_\beta(x)) - L_b(q_\beta(x)) > 0.$$

The CPWL conditional score under active piece a is $c_a(p) = -b_a - \langle s_a, p\rangle$. Hence its population contribution can be written locally as

$$\mathcal{L}_{H,\chi}(\beta) = \int_\Omega \chi(x)\,[c_b(p(x)) + \mathbf{1}\{g_{\beta,ab}(x) > 0\}\Delta c_{ab}(p(x))]\,\rho(x)\,dx,$$

where $\Delta c_{ab} = c_a - c_b$. Differentiating the moving indicator and applying the coarea formula in the d_x-dimensional covariate space gives

$$\nabla_\beta \mathcal{L}_{H,\chi}(\beta) = \int_{M_{\beta,ab}} \chi(x)\Delta c_{ab}(p(x)) \frac{\nabla_\beta g_{\beta,ab}(x)}{\|\nabla_x g_{\beta,ab}(x)\|} \rho(x)\, d\mathscr{H}^{d_x-1}(x). \tag{61}$$

On $M_{\beta,ab}$, $(b_a - b_b) + \langle s_a - s_b, q_\beta(x)\rangle = 0$, so

$$\Delta c_{ab}(p(x)) = -\langle s_a - s_b, p(x) - q_\beta(x)\rangle.$$

The stated regularity makes each active surface locally a moving graph; its continuous density-weighted jump yields the displayed derivative. A partition of unity sums the local formulas. Transverse junctions of codimension at least two have swept volume of order $o(\|\Delta\beta\|)$ and contribute no first-order term; compact support removes covariate-domain boundary terms. Adding the baseline chain-rule term proves (59). This is the usual interior-plus-boundary decomposition for a nondifferentiable reparameterized expectation (Lee et al., 2018), specialized to the selected proper-score loss. A finite empirical objective has fixed indicators between sample crossings, so ordinary autodifferentiation does not recover the surface term.

A nonzero-flux binary example. Writing $\mathcal{L}_H$ for the population score risk of the selected increment, let

$$H_c(q) = c|q - 1/2|, \qquad q_\beta(x) = 1/2 + x - \beta, \qquad X \sim \mathrm{Unif}[-a, a],$$

Take $c > 0$, a constant truth $p_0 \neq 1/2$, $0 < a < 1/4$, and $|\beta| < a$. Then all reports remain in the probability interval and the moving facet lies inside the support. For every fixed $x \neq \beta$, the ordinary derivative of the CPWL score increment is zero. The facet preimage is the single point $x = \beta$, however, and (61) gives

$$\frac{d}{d\beta}\mathcal{L}_H(\beta) = \frac{c}{a}\left(p_0 - \frac{1}{2}\right) \neq 0. \tag{62}$$

Thus pointwise pathwise opacity and a nonzero exact population derivative are compatible; differentiation and integration do not commute across the moving score discontinuity.

K Binary projection and weather studies

This appendix moves from controlled population projections to held-out comparisons on observed precipitation labels. The first study isolates projection splitting and rejoining under known conditional truth. A deterministic fixed-class analysis then connects the global distortion certificate to the class-restricted and achieved factors. The real-label layer changes both the estimand and the evidence source: it evaluates fixed-user costs on real GHCN-Daily labels, for which the contextwise conditional truth is unobserved.

K.1 Known-truth projection mechanism

The controlled laboratory makes the source of projection differences observable. Let s range over a uniform 4,001-point grid on $[-2, 2]$ and set

$$h(s) = 0.8\sin(1.5s) + 0.3(s^2 - 4/3), \qquad p_\gamma(s) = \sigma\{-0.7 + 1.2s + \gamma h(s)\}.$$

The misspecification path is

$$\gamma \in \{0, 0.1, 0.25, 0.5, 0.75, 1, 1.5, 2\}.$$

The three nested report classes are

$$\begin{aligned} q_b(s) &= \sigma(1.2s + b), \\ q_{a,b}(s) &= \sigma(as + b), \\ q_{a,b,c}(s) &= \sigma(as + b + ch(s)). \end{aligned}$$

The first two classes are realizable at $\gamma = 0$; the augmented class is realizable for every γ.

Both declared tasks have three actions. Their loss matrices, with columns $y = 0, 1$ and rows ordered from no action to full action, are

$$L_1 = \begin{pmatrix} 0 & 1 \\ 5/21 & 2/7 \\ 19/21 & 0 \end{pmatrix}, \qquad L_2 = \begin{pmatrix} 0 & 1 \\ 5/26 & 11/26 \\ 7/13 & 0 \end{pmatrix}.$$

The corresponding convex decision entropies have thresholds $(0.25, 0.70)$ with charges $(20/21, 20/21)$ and $(0.25, 0.45)$ with charges $(10/13, 10/13)$. Thus the tasks share one facet and retain one distinct facet each. We use $\varepsilon = 0.05$ and compare log, Brier, the exact task sum, the selected binary join, the two task-specific scores, and direct minimization of T_ε.

Every method uses the same report class and bounds. An offset fit uses a parameter $b \in [-4, 3]$; Platt uses $(a, b) \in [0.05, 3] \times [-4, 3]$; augmentation adds $c \in [-1, 3]$. The offset search uses a 6,001-point grid followed by bounded scalar polishing. Platt and augmented fits use three differential-evolution starts (seeds 0, 1, 2), including independent searches for the direct target, followed by bounded Powell polishing. Differential evolution uses 140 iterations, population multiplier 18, and relative/absolute tolerances $10^{-7}/10^{-10}$; Powell uses at most 500 iterations and parameter/objective tolerances $10^{-8}/10^{-10}$. Known realizable parameters are included as controls. Additional direct-target polishing starts from every score solution; the lowest target risk across all candidates supplies the numerical reference. None of the 168 controlled fits ends on a parameter bound.

Table 2 reports the two central factor checks. The "global factor" is a dense whole-domain ratio evaluation augmented with crossings in both directions at every knot. It estimates the ratio supremum of an analytically feasible construction without certifying its exact value. The fitted-factor check compares each nonzero target-risk row with the directly optimized target reference in the same class.

Table 2: Controlled projection evaluation. The maximum ρ is over nonzero-risk misspecified cells. Exceedances use a tolerance of 0.002 above the evaluated global factor.

Proper construction	Evaluated global factor	Maximum fitted ρ	Exceedances
Task sum	1.98527	1.08093	0
Selected join	1.65154	1.00178	0

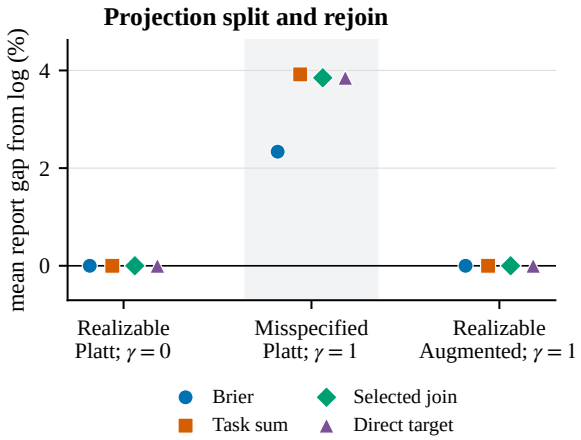


Figure 4: **Projection agreement and separation.** Proper-score projections separate under misspecification and rejoin when augmentation restores realizability. Figure 5 gives the baseline-weight paths.

The realizable controls rejoin to numerical precision: the maximum mean absolute report error in the preregistered realizable cells is 1.613×10^{-9}. At $\gamma = 2$, the largest Platt projection distance from the log projection is 0.03290, while augmented reports coincide again. Over all 24 γ–class cells, join-minus-sum target risk ranges from -2.149×10^{-3} to 1.04×10^{-17}, with the upper endpoint numerically zero. Across finite evaluated-pair diagnostics for all non-oracle scores, Spearman's rank association between ratio range and fitted ρ is 0.4303. The association is descriptive because the method–regime cells share truths and report classes.

K.2 Fixed-class certificate bridge and baseline-weight path

The fixed-class bridge places the global certificate, the integrated class-restricted factor, and the achieved projection ratio on the same population. This deterministic analysis reuses the controlled $\gamma = 1$ truth, tasks, 4,001 contexts, parameter bounds, and fitted score and direct-target projections, while adding 15 baseline weights and new ratio-endpoint optimizations. It was not part of the preregistered controlled panel. For each of the offset and Platt classes, it computes the lower and upper endpoints of

$$\frac{\mathbb{E}_X D(p_X, q_h(X))}{\mathbb{E}_X T_\varepsilon(p_X, q_h(X))}$$

over the declared bounded class using the same search schedule and tolerances as the controlled fits. Their estimated ratio approximates $\kappa_{\mathcal{H},P}$, and the achieved ratio compares the target risk at the score projection with the directly optimized target risk. The grid

$$\varepsilon \in \{0.001, 0.002, 0.005, 0.01, 0.02, 0.035, 0.05, 0.075, 0.1, 0.15, 0.25, 0.5, 1, 2, 5\}$$

provides 15 scale points.

Table 3: **Factor hierarchy** at $\gamma = 1$ and $\varepsilon = 0.05$. Class and achieved quantities use the misspecified Platt population. Finite factors are numerical estimates; infinite global factors follow analytically. Dashes denote class factors not evaluated.

Score	Global factor	Class estimate	Achieved $\rho_{\mathcal{H},P}$
Log	∞	357.06	33.40
Brier	∞	46.37	5.77
Task 1 score	∞	—	1.07785
Task 2 score	∞	—	1.01683
Task sum	1.9853	1.8938	1.0104
Selected join	1.6515	1.3972	1.000003

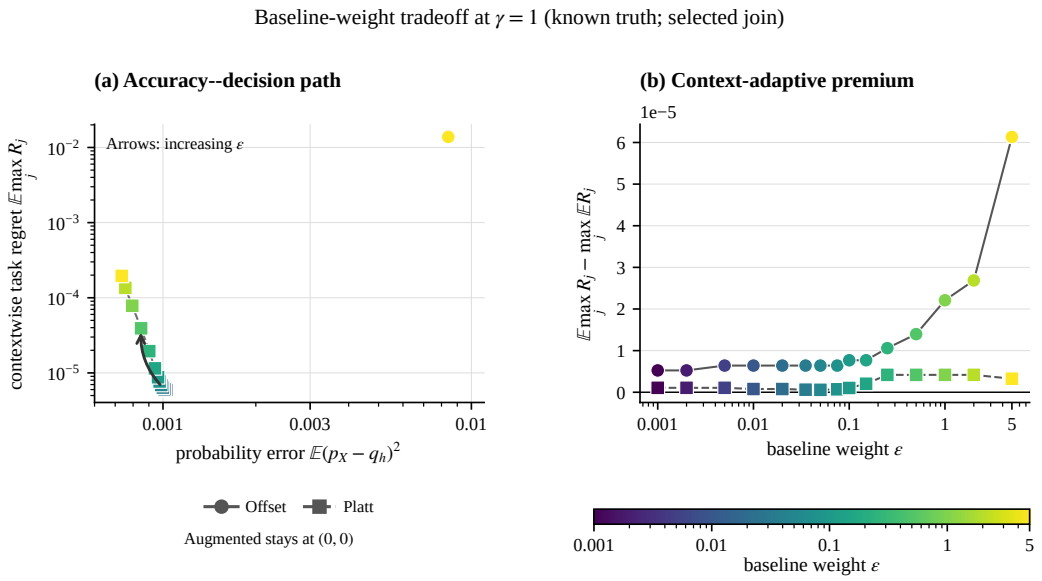


Figure 5: **Population paths for the selected join across baseline weights.** (a) Numerically fitted offset and Platt projections trace distinct accuracy–decision paths; the augmented realizable class remains at the origin. (b) The vertical gap is the premium of the context-adaptive target $\mathbb{E}\max_j R_j$ over the fixed-user target $\max_j \mathbb{E}R_j$. Colors encode the same 15 values of ε in both panels; the arrow marks increasing baseline weight along the Platt path.

The task-specific scores have target risks 6.064×10^{-5} (Task 1) and 5.721×10^{-5} (Task 2) in this panel. Their achieved ratios in Table 3 place single-user tailoring alongside the two multi-task constructions.

Figure 5 displays these population paths and the adaptivity premium. For the offset class, the adaptivity premium ranges from 5.24×10^{-6} to 6.13×10^{-5}; for Platt it ranges from 5.80×10^{-7} to 4.17×10^{-6}. It is zero in the augmented realizable control. Thus the fixed-user and context-adaptive objectives are close in this laboratory but are not identical, and their gap is model-class dependent.

The analysis comprises 120 factor evaluations and 480 endpoint-optimization diagnostics, with no optimizer failure. For constructions with finite evaluated global factors, the numerical estimates follow the hierarchy $\rho_{\mathcal{H},P} \le \kappa_{\mathcal{H},P} \le \kappa_{\text{global}}$, where $\kappa_{\text{global}} = b(D, T_\varepsilon)/a(D, T_\varepsilon)$ for the specified score. The sum and join are globally feasible by theorem; their displayed factors are numerical estimates from whole-domain searches and class-restricted optimization.

At $\varepsilon = 0.05$ in the Platt class, the directly optimized reference has target risk 5.6×10^{-5}, compared with 1.879×10^{-3} for log and 3.25×10^{-4} for Brier. The small reference risk explains how modest average report differences can yield large achieved ratios. Across the full grid, 43 endpoint extrema reach a declared parameter bound, mostly in the offset class and at extreme baseline weights; none does so in the primary Platt calculation at $\varepsilon = 0.05$.

K.3 BINARY REAL-LABEL COST EVALUATION

The real-label study evaluates a separate fixed-user estimand on observed next-day precipitation. It uses NOAA GHCN-Daily station files (Menne et al., 2012). The 26 prespecified station identifiers are

USW00003812	USW00012839†	USW00023174†	USW00024233†
USW00014739†	USW00094728	USW00013743	USW00013874
USW00012960	USW00003927	USW00014922	USW00094846
USW00014821	USW00014820	USW00093814	USW00093821
USW00013994	USW00013963	USW00023050†	USW00023183
USW00024127	USW00003017	USW00023169	USW00023234
USW00024229	USW00026451†		

where † marks the six geographic holdouts. All 26 station files were usable, meeting the prespecified minimum-station requirement.

For each station-day, features use only information available by that day: same-day precipitation, its one-day lag and three- and seven-day rolling summaries, daily maximum/minimum/average temperature, average wind speed, day-of-year sine and cosine, latitude, longitude, and elevation. Nonempty GHCN-Daily quality flags are treated as missing. A label is formed only when the next calendar day is present and has a valid precipitation observation. The primary label is next-day precipitation of at least 1 mm. The 5 mm and 10 mm labels are prespecified, separately refitted robustness panels.

Twenty nonholdout stations from 2000–2017 train a histogram-gradient-boosting base classifier. Its learning rate is 0.06, with 220 boosting iterations, 31 leaves, minimum leaf size 35, and ℓ_2 regularization 0.2. The base logit is frozen before score fitting. The 2018–2021 nonholdout observations fit one of three shared calibration heads:

$$\sigma(s+b), \qquad \sigma(as+b), \qquad \sigma\{as+b+c(s^2-\overline{s^2})/\operatorname{sd}(s^2)\}.$$

Both statistics use the calibration split; a standard deviation below 10^{-8} is set to one. The offset bound is $b \in [-4,4]$; Platt uses $(a,b) \in [0,4]\times[-5,5]$; the quadratic head adds $c \in [-2,2]$. The primary panel uses all three heads; robustness panels use the Platt head. The same seven score objectives as in the controlled study are compared, with the direct comparator minimizing the empirical maximum of the two users' mean costs plus the proper baseline. This comparator targets observed mean costs and does not observe the contextwise target. None of the 35 weather fits ends on a parameter bound.

The four analysis partitions contain 198,450 labeled station-days: 131,254 base-training observations, 29,220 score-fitting observations, 29,213 future-time test observations, and 8,763 station-holdout test observations. Both test partitions use 2022–2025 next-day outcome dates: future-time testing uses the twenty nonholdout stations, and geographic testing uses the six holdout stations. Partition years refer to the next-day outcome date. The full prepared dataset contains 256,124 rows; the remaining 57,674 comprise 9,464 with 1999 outcome dates, retained when constructing lagged features, and 48,210 from the six holdout stations with 2000–2021 outcome dates. These rows enter none of the four analysis partitions. We report the maximum and average of the two users' realized mean costs, together with Brier risk and negative log-likelihood (NLL). Uncertainty uses 500 paired station–month block-bootstrap replicates.

Table 4: Primary 1 mm Platt-head held-out results. Costs use real labels; worst and mean costs are the maximum and average of the two user costs.

Panel	Method	User 1	User 2	Worst	Mean	Brier	NLL
Future time	Log	0.171018	0.174087	0.174087	0.172553	0.147780	0.455752
Future time	Join	0.171659	0.174471	0.174471	0.173065	0.148198	0.457811
Station	Log	0.173457	0.176463	0.176463	0.174960	0.148783	0.455606
Station	Join	0.171555	0.176647	0.176647	0.174101	0.148371	0.455370

User 2 has the higher observed mean cost for every method and calibration head in both primary 1 mm test panels. Average and worst-user costs move differently in Table 4. Table 5 reports the paired worst-cost contrasts: neither primary interval excludes zero. The prespecified 5 mm station-holdout

interval excludes zero, whereas the other robustness intervals do not. These intervals describe the fixed panels and are not multiple-testing-adjusted claims.

Table 5: Join-minus-log worst-cost contrasts under the Platt head. The center is the mean of 500 paired bootstrap contrasts; the endpoints are their percentiles. These means need not equal the observed cost differences in Table 4. Negative values favor the join.

Event	Panel	Bootstrap mean	2.5%	97.5%
1 mm	Future time	0.000400	−0.000414	0.001175
1 mm	Station	0.000155	−0.001288	0.001589
5 mm	Future time	0.000161	−0.000247	0.000608
5 mm	Station	−0.000829	−0.001654	−0.000050
10 mm	Future time	−0.000257	−0.000660	0.000127
10 mm	Station	−0.000828	−0.001881	0.000204

Across every weather panel and capacity, join-minus-sum worst cost lies between -2.70×10^{-4} and zero. The improvement occurs in the offset head; Platt and quadratic heads usually select identical decisions for sum, join, and the direct declared-user comparator. These results separate the worst-case distortion certificate from strict gain on a fixed report distribution.

K.4 Interpretation

The controlled studies recover projection agreement under realizability and score-dependent projections under misspecification. The fixed-class analysis shows that estimated class factors and achieved ratios can be much smaller than the evaluated global factor. On observed precipitation labels, the primary worst-user cost intervals include zero, and one prespecified robustness interval favors the join. These cost comparisons concern the declared users and panels; the contextwise conditional truth is unobserved.

On a fixed population, boundary effects depend on both the truth and the fitted reports. A margin condition on p_X controls true probabilities near decision facets; the smoothing bound in Corollary I.1 instead uses the boundary occupancy of $q_h(X)$. Connecting the two under misspecification requires control of prediction error. The low-dimensional heads and separated binary thresholds offer a possible explanation for the close sum and join risks. Their relative performance also depends on where the fitted reports cross task boundaries: the join's least-majorant property concerns pointwise divergences, while the achieved target risks compare different fitted predictors.

L Four-class weather cost evaluation

This study evaluates fixed-user costs on observed next-day precipitation labels using a frozen four-class predictor and common calibration heads.

L.1 Data, partitions, and report classes

All 26 preregistered GHCN-Daily stations were usable. The weather features, station holdout, chronology, missing-value handling, and binary base model are the same as in Appendix K. Four-class next-day precipitation uses the bins dry (0 mm), light ($(0, 1)$ mm), moderate ($[1, 5)$ mm), and heavy (≥ 5 mm). Table 6 gives the sample counts.

Table 6: Four-class observations by partition.

Partition	Dry	Light	Moderate	Heavy
Train	93,252	8,677	12,770	16,555
Calibration	20,259	1,951	3,003	4,007
Future time	21,056	1,795	2,772	3,590
Held-out stations	6,222	594	953	994

The Laplace-smoothed training frequency is

$$p_0 = (0.710455743650, 0.066114065428, 0.097296926664, 0.126133264258).$$

One four-class histogram-gradient-boosting predictor is trained on 2000–2017 nonholdout stations. Its centered log probabilities are frozen before fitting one of three bounded calibration heads: a positive temperature; a temperature plus three class biases; or four positive diagonal scales plus three class biases. Explicitly, the heads are $\mathrm{softmax}(az)$, $\mathrm{softmax}(az + b)$, and $\mathrm{softmax}(a \odot z + b)$. Each logit scale lies in $[0.1, 3]$, each free bias in $[-3, 3]$, and $b_4 = 0$. The base model uses learning rate 0.06, 240 boosting iterations, 31 leaves, minimum leaf size 35, ℓ_2 regularization 0.2, seed 20260826, and no early stopping. Predicted probabilities are clipped below at 10^{-8} before taking and centering their logs.

L.2 Task families

For a severity vector w, action 0 incurs $w(Y)$ and action 1 incurs the constant $c = w^\top p_0$, with action 0 selected at a tie. The frozen components are

$$\begin{aligned} w_1 &= (0, 0, 1, 1), & c_1 &= 0.223430190922, \\ w_2 &= (0, 0, 0, 1), & c_2 &= 0.126133264258, \\ w_3 &= (0, 0.2, 0.6, 1), & c_3 &= 0.197734233342, \\ w_4 &= (0, 1, 1, 1), & c_4 &= 0.289544256350. \end{aligned}$$

The three-direction family uses three separate two-action tasks with severity vectors w_1, w_2, w_3, whose centered normals have rank three. For this family, the shared-component majorant equals the literal task sum. For component ℓ, write $H_\ell(q) = -\min\{w_\ell^\top q, c_\ell\}$. Each overlap task combines component 2 with one component $k \in \{1, 3, 4\}$: its action is a pair $(a_2, a_k) \in \{0, 1\}^2$, its loss is the sum of the two component losses, and its negative Bayes envelope is $H_2 + H_k$. The Bayes decoder acts componentwise with the same tie rule. The literal task sum therefore counts component 2 three times. The shared-component majorant counts each component once. Convex subtraction establishes this construction as a feasible majorant for the declared family.

L.3 Objectives and common optimization protocol

All methods use the same frozen base predictor, report class, parameter bounds, data split, calibration coreset, and optimization budget. The compared methods are log, Brier, the literal task sum, the shared-component majorant, and a nonproper direct comparator. Proper scores are fitted to observed labels. Task-augmented scores use $\varepsilon = 0.05$ and $F_0(q) = \frac{1}{2}\|q\|_2^2$. The direct comparator minimizes the empirical maximum mean task cost plus the Brier baseline, with baseline weight 0.05.

Optimization uses seeds 0 and 1, at most eight CPU threads, a deterministic 3,000-row calibration coreset, differential evolution followed by bounded Powell refinement, and a three-hour time limit. Every fitted solution is evaluated on the full calibration objective before selection. Direct-reference selection compares the complete set of fitted candidates on that objective. The maximum direct-minus-best-proper calibration gap is zero in the real-label evaluations. No fitted parameter reached a bound, and no fitting or evaluation failure was recorded. The coreset orders calibration rows by predicted class, then maximum predicted probability, with original row index breaking ties. It splits this order into 3,000 nearly equal blocks, selects the middle row of each block, and weights it by the block's fraction of the calibration sample.

L.4 Real-label cost evaluation

The real-label study evaluates the maximum mean cost across declared tasks on observed precipitation labels. The evaluation uses the same frozen base logits and calibration heads, with the overlap and three-direction families. Its primary contrast is the shared-component majorant minus log in the overlap family, using the temperature–bias head on held-out stations. All other family, head, method, and panel contrasts are secondary. Intervals use 1,000 paired station–month block-bootstrap replicates with seed 20260826 and common replicate indices.

The primary estimated cost difference is -0.000485712, with a paired 95% interval of $[-0.00293708, 0.00197659]$. Among all 12 shared-component-majorant contrasts against log, three

intervals lie below zero, nine include zero, and none lies above zero. Among all 36 proper-score contrasts, four intervals lie below zero, 28 include zero, and four lie above zero. Negative differences favor the candidate score. These counts are descriptive and use unadjusted intervals. Figures 6 and 7 retain all 48 method contrasts, including the 12 shared-component-majorant comparisons and the nonproper direct comparator.

Interpretation. The complete comparison shows how observed costs vary across task families, report classes, and evaluation panels. Task 1 attains the largest observed mean cost throughout the three-direction family; task 3 (components 2 and 4) does so throughout the overlap family, across all methods, heads, and test panels. Thus the worst-cost contrasts track a fixed user within each family. Table 7 displays the individual costs for the primary overlap head, including the differing directions of user-cost changes on held-out stations.

Table 7: Observed individual task costs in the overlap family with the temperature–bias head. Task 3 determines the worst-user cost in each row.

Panel	Method	Task 1	Task 2	Task 3
Future time	Log	0.243897	0.224896	0.296753
Future time	Component majorant	0.241989	0.223542	0.295008
Station	Log	0.236009	0.216293	0.292441
Station	Component majorant	0.237912	0.217525	0.291955

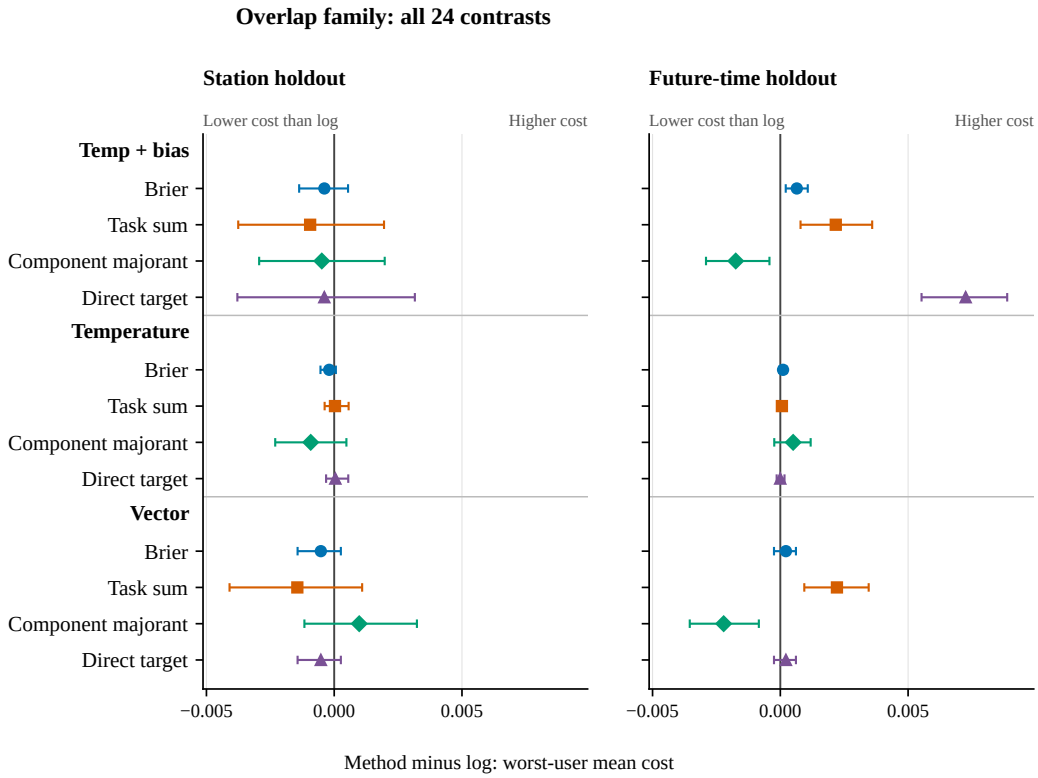


Figure 6: **Overlap-family real-label cost comparisons.** Each row is a paired worst-task cost difference from log, with station and future-time holdouts in separate columns; negative values favor the candidate. Every report class, holdout, and prespecified method is shown, including the nonproper direct comparator. The primary comparison is the shared-component majorant for the temperature–bias head on held-out stations; its interval includes zero. Intervals are descriptive and unadjusted.

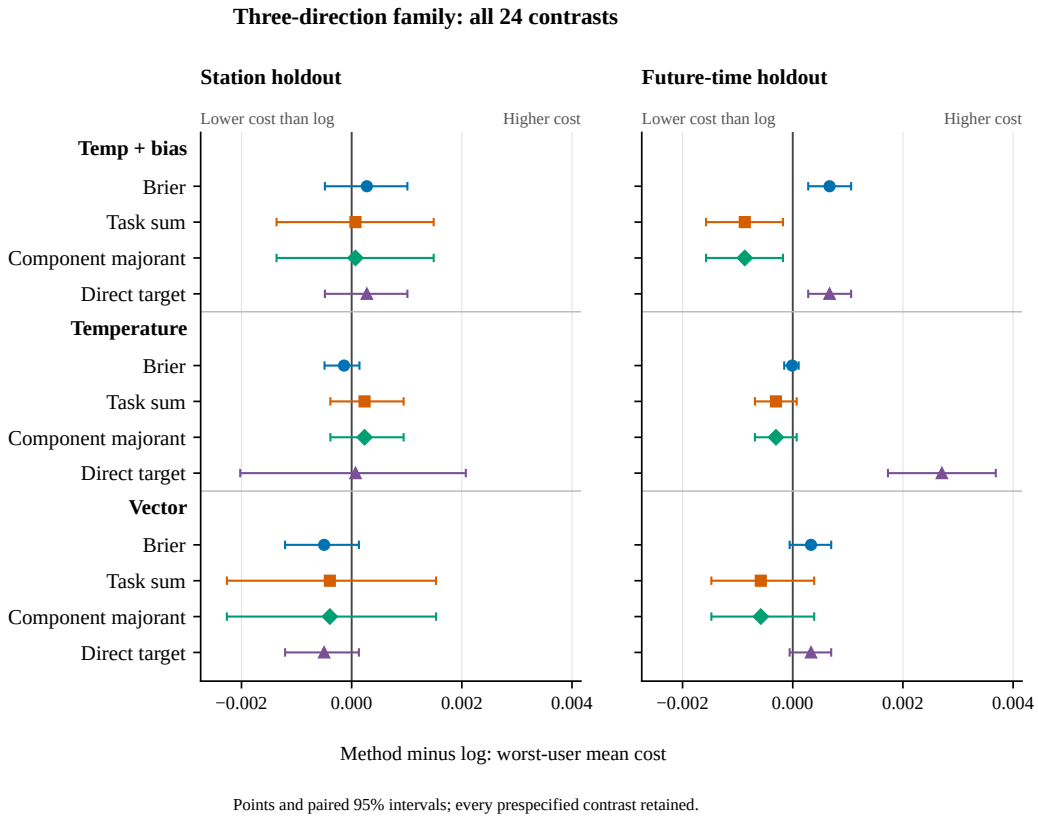


Figure 7: **Three-direction-family real-label cost comparisons.** With Figure 6, this completes the 48-contrast evaluation under the same convention. Across the 36 proper-score contrasts, four intervals lie below zero, 28 include zero, and four lie above zero. These counts are descriptive and use unadjusted intervals.