POPULATION-PRESERVING LOSS CORRECTIONS FOR FINITE-SAMPLE DECISION SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Losses with identical population comparisons can select different decisions from finite samples. We study shared loss corrections learned across tasks and deployed by empirical minimization over fixed candidate banks. Conditionally centered controls preserve population preferences while changing joint empirical comparisons. We characterize harmful selection through multivariate Chernoff bounds, an entropy dual, and a convex-reference support certificate for zero regret. Regularized exponential construction in a reproducing kernel Hilbert space has a finite positive-cone representation, but convex comparison criteria admit an unbounded multiplicative gap to decision risk, even for bounded losses and one correction parameter. Direct episodic welfare fitting over a d-parameter affine family instead achieves excess risk of order $\sqrt{d\log(2M)/N}$ from N independent tasks with at most M candidates, without a margin assumption. A centered sensor construction gives a matching $\sqrt{d/N}$ lower bound for fixed bank size, even when each sampled task's law is revealed. Independent validation adds finite-library oracle and baseline-safety guarantees. Controlled sensor and contextual studies examine functional capacity, joint comparison information, and the difference between convex construction and direct decision-risk fitting.

1 INTRODUCTION

Decision-focused learning (DFL) trains predictors for the quality of their downstream decisions, using direct objectives, analytic surrogates, differentiable optimization, or learned losses (Donti et al., 2017; Wilder et al., 2019; Elmachtoub & Grigas, 2022; Shah et al., 2022; Mandi et al., 2024). These approaches connect prediction to the optimization problem that uses it. A further question concerns the learning procedure itself: among losses with the same population preferences, which one produces better decisions from a limited training sample?

Population alignment does not answer this question. An empirical learner selects a candidate through simultaneous comparisons with its competitors; their joint distribution determines the probability and cost of a selection error. Ranking-and-selection and sample-average optimization study related rare comparison events (Glynn & Juneja, 2004; Fu et al., 2004; Kleywegt et al., 2002). For loss design, the comparison distribution is itself adjustable. An observed nuisance with conditional mean zero can be added with a report-dependent coefficient, preserving every population loss difference while changing empirical rankings. This is a control-variate construction (Glasserman, 2004), with downstream selection risk as its objective.

We learn such coefficients as shared functions of observable metadata, reports, and context. Source tasks then determine a loss that is evaluated on a new task's own candidate bank and training sample, consistent with the task-distribution view of learning inductive bias (Baxter, 2000). Structural conditional centering preserves population comparisons separately on almost every task. The bank may vary across tasks but is fixed independently of their training observations. This setting covers selection among independently specified predictors and permits exact analysis of the finite-sample learning rule.

Our results separate three questions: what comparison information permits improvement, what an optimization criterion chooses, and how many source tasks are needed to learn a useful correction. Joint comparison laws yield a multivariate Chernoff bound, an entropy dual, and a support certificate

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.

using a convex combination of competitors. Centered RKHS corrections admit regularized convex construction and finite-space approximation. However, even with bounded losses, every convex comparison criterion can prefer a correction whose decision-risk ratio is arbitrarily large. We therefore analyze direct episodic welfare fitting over a continuous affine correction family. Counting the regions and tie faces of empirical comparisons gives an excess-risk bound of order $\sqrt{d\log(2M)/N}$ for d parameters, N source tasks, and bank size at most M, without a margin assumption. A centered sensor family supplies a matching $\sqrt{d/N}$ lower bound for fixed bank size. Independent validation can combine both construction approaches and certify improvement over direct loss.

Relation to prior theory. Chernoff bounds, entropy duality, convex separation, and the representer principle are classical (Dembo & Zeitouni, 1998; Rockafellar, 1970; Schölkopf et al., 2001). Here they characterize changes to joint learner comparisons under taskwise population constraints. DFL surrogates approximate downstream objectives (Elmachtoub & Grigas, 2022; Gupta & Huang, 2024; Schutte et al., 2024); CaVE aligns reports with action cones (Tang & Khalil, 2023). Our cones instead index predictors selected by empirical minimization. Direct episodic fitting uses the established theory of piecewise algorithm performance (Balcan et al., 2021) and empirical processes (Bartlett & Mendelson, 2002). Its upper bound, the lower bound within centered corrections, and the convex-criterion obstruction distinguish statistical learnability from the quality of convex construction.

2 Task model and loss transfer

A task τ consists of an observation law P_τ, a finite candidate bank $\mathcal{H}_\tau = \{h_{\tau,1}, \dots, h_{\tau,M_\tau}\}$ with $2 \le M_\tau \le M$, an observable descriptor a_τ, and a bounded realized welfare $w_\tau(h, O) \in [0,1]$. The bank is fixed before any training or evaluation observations. Write $W_{\tau,i} = \mathbb{E}_{P_\tau} w_\tau(h_{\tau,i}, O)$, $W_\tau^\star = \max_i W_{\tau,i}$, and $B_{\tau,i} = W_\tau^\star - W_{\tau,i}$. Tasks are i.i.d. from a meta-distribution Π. Observations include a context X, and a report $h_{\tau,i}(X)$ specifies a downstream action. We assume $w_\tau(h, O) = \widetilde{w}(a_\tau, h(X), O)$ for a known measurable reward map, so, for a report r, $S_0(a, r, O) = -\widetilde{w}(a, r, O)$ is an available direct decision loss (known positive affine equivalents induce the same selector). The descriptor contains observable problem metadata; the task law and its optimal candidate are unknown to the learner. All sample spaces are standard Borel; optimizing rules and empirical-process suprema are taken measurable (or interpreted in outer probability). Task-indexed maps and the probability kernel $\tau \mapsto P_\tau$ are measurable, and all population expectations used below exist.

In linear predict-then-optimize problems, reports are predicted cost vectors and downstream actions minimize the reported cost over a feasible set. Welfare normalizes negative realized cost. The downstream action geometry determines welfare gaps; the comparison cones below determine which predictor a finite sample selects. Appendix H formalizes this distinction and the links to DFL surrogates.

A shared loss is a measurable function $S(a, r, O)$. Given a target sample $D_m = (O_1, \dots, O_m) \sim P_\tau^m$, with X_j the context in O_j, use the smallest-index tie rule

$$\widehat{i}_{\tau,m,S} = \min \arg\min_i m^{-1} \sum_{j=1}^{m} S(a_\tau, h_{\tau,i}(X_j), O_j).$$

The target meta-risk is $\mathcal{R}_m(S) = \mathbb{E}_{\tau\sim\Pi} R_{\tau,m}(S)$ with $R_{\tau,m}(S) = \mathbb{E}_{P_\tau^m}[W_\tau^\star - W_{\tau,\widehat{i}_{\tau,m,S}}]$.

Construction, selection, and target tasks are independent draws from Π. The final loss is frozen before target training data are observed; every method receives the same target bank and target observations, and direct decision-loss empirical risk minimization (ERM) is included as a baseline and the safety comparator. Conditional on a task, all sample blocks are independent draws from its observation law. Banks may vary arbitrarily across tasks but cannot be fitted to these blocks. A shared loss remains applicable when source and target banks differ, because it scores reports through observable task inputs. Appendix C.3 discusses the corresponding information requirements.

Within a fixed task, suppress its index and descriptor and write M for its bank size. Let $\Delta_d = \{q \in \mathbb{R}^d : q \ge 0,\ \mathbf{1}^\top q = 1\}$, where $\mathbf{1}$ is the all-ones vector. Write $\Lambda = \{z \in \mathbb{R}^M : \mathbf{1}^\top z = 0\}$ and let P_Λ be orthogonal projection onto Λ, which removes the common component of a loss vector. Define

$$Z(S, O) = (S(h_1(X), O), \dots, S(h_M(X), O)), \quad D_r(S, O) = (Z_r - Z_i)_{i\ne r}.$$

For each candidate index r, let $C_r = \{z \in \mathbb{R}^M : z_r \le z_i \ \forall i \ne r\}$ be the cone of loss vectors for which candidate r minimizes the empirical loss. These are the minimizing normal cones of the candidate simplex Δ_M. Candidate r is harmful when $B_r > 0$; the unsafe set is $\mathcal{U} = \bigcup_{r:B_r>0} C_r$.

3 DISTRIBUTIONAL DECISION GEOMETRY

Fix a task and a loss. Write $R_m = R_{\tau,m}$ and $B_r = B_{\tau,r}$; $\mathbb{R}^d_+$ and $\mathbb{R}^d_-$ denote the nonnegative and nonpositive orthants. A harmful candidate is selected only if it beats *every* competitor on the same sample; the following result retains this joint event.

Theorem 1 (Distributional separation)**.** *Define, with $e^{-\infty} = 0$,*

$$J_r(S) = \sup_{\lambda \in \mathbb{R}^{M-1}_+} \{-\log \mathbb{E} \exp[-\lambda^\top D_r(S, O)]\}, \tag{1}$$

assuming the moment generating function is finite on this domain. For any $m \ge 1$ and nonnegative multipliers λ_r,

$$R_m(S) \le \min\left\{1, \sum_{r:B_r>0} B_r e^{-mJ_r(S)}\right\} \le \min\left\{1, \sum_{r:B_r>0} B_r (\mathbb{E} e^{-\lambda_r^\top D_r(S,O)})^m\right\}. \tag{2}$$

If O takes values $o_1, \dots, o_s$ with probabilities $p_j > 0$, then

$$J_r(S) = \inf_{\substack{q \in \Delta_s \\ \sum_j q_j D_r(S, o_j) \le 0}} D_{\mathrm{KL}}(q \| p), \tag{3}$$

where $D_{\mathrm{KL}}(q\|p) = \sum_j q_j \log(q_j/p_j)$ with $0 \log 0 = 0$, and an infeasible infimum is $+\infty$. If D_r has bounded support, let $K_r = \overline{\mathrm{conv}(\operatorname{ess\,ran} D_r)}$ be the closed convex hull of its essential range. Then

$$K_r \cap \mathbb{R}^{M-1}_- = \varnothing \quad \Longleftrightarrow \quad \exists q_r \in \Delta_{M-1},\ \delta_r > 0 : Z_r - \sum_{i \ne r} q_{r,i} Z_i \ge \delta_r \quad \text{a.s.} \tag{4}$$

If this holds for all harmful r, then $R_m(S) = 0$ for every m. On a finite alphabet, (4) *is also equivalent to $J_r(S) = +\infty$.*

Proof sketch and interpretation. Selection of r requires $\overline{D}_{r,m} = m^{-1} \sum_{j=1}^m D_r(S, O_j) \le 0$; a nonnegative combination of these inequalities gives an exponential Markov bound, and independence across observations yields its mth power. Gibbs' variational identity gives the entropy dual, and strict separation of the compact K_r from the negative orthant produces the convex reference in (4). Complete proofs are in Appendix B.

The entropy formula measures the smallest change of observation law that makes r an expected-loss minimizer. Finite-sample regret also depends on integer sample counts and the tie rule; the bound includes harmful minimizers excluded by that rule. Appendix C.1 accounts for these effects through empirical types. Averaging (2) over tasks gives a meta-risk bound. Writing $\lambda_r = \eta_r q_r$ with $\eta_r = \|\lambda_r\|_1$ and $q_r \in \Delta_{M-1}$ (arbitrary when $\eta_r = 0$), the scalar comparison is between candidate r and a convex combination of all competitors: the combination need not be deployed, and restricting to one reference can discard cancellations among competitors.

Corollary 1 (Gaussian specialization)**.** *Suppose $P_\Lambda Z$ is Gaussian with mean $g \in \Lambda$ and covariance Σ positive definite on Λ. With the inverse taken on Λ,*

$$J_r(S) = \frac{1}{2} \inf_{z \in C_r \cap \Lambda} (z - g)^\top \Sigma^{-1} (z - g). \tag{5}$$

For two candidates, let the harmful candidate have regret $B > 0$ and its harmful-minus-good comparison be $N(\mu_S, \sigma_S^2)$ with $\mu_S, \sigma_S > 0$. Let Φ be the standard normal distribution function. The Gaussian sample mean gives

$$R_m(S) = B\Phi\left(-\sqrt{m}\, \mu_S/\sigma_S\right), \tag{6}$$

Within this Gaussian family, maximizing the comparison signal-to-noise ratio μ_S/σ_S therefore minimizes decision risk.

Covariance geometry thus gives the Chernoff exponent exactly in the Gaussian model. Outside it, even exact joint means and covariances need not identify a zero-risk loss: Appendix G constructs two moment-matched worlds in which every deterministic moment-only designer incurs regret at least $4/25$ in one world, although a zero-risk loss exists in each. Support safety, conversely, requires knowledge of the true support; screening a finite sample cannot certify absent tails.

4 Functional construction of reusable losses

We optimize shared losses $S = S_{\mathrm{b}} + v$, where S_{b} is an integrable base loss and v is a correction. Let $\mathscr{H}$ be a real reproducing kernel Hilbert space (RKHS) on permitted inputs $\xi = (a, r, O)$, with measurable kernel satisfying $\sup_{\xi} k(\xi, \xi) \le \kappa^2$. For clarity take a common bank size M; variable sizes use the corresponding taskwise vector spaces with the same bound. For $v \in \mathscr{H}$, define

$$(\mathcal{A}v)(\tau, O) = P_{\Lambda}\big(v(a_\tau, h_{\tau,i}(X), O)\big)_{i=1}^{M}, \quad (\mathcal{M}v)(\tau) = \mathbb{E}_{P_\tau}(\mathcal{A}v)(\tau, O). \tag{7}$$

These maps take values in $L^2(\Pi(d\tau)P_\tau(dO); \mathbb{R}^M)$ and $L^2(\Pi; \mathbb{R}^M)$ with operator norms at most $\sqrt{M}\kappa$, so $\ker \mathcal{A} \subseteq \ker \mathcal{M}$ are closed, and the quotient $\ker \mathcal{M}/\ker \mathcal{A}$ identifies population-preserving changes that can affect empirical comparisons. Preservation is conditional on almost every task; the pooled equation $\mathbb{E}_{\tau,O}\mathcal{A}v = 0$ does not suffice.

Realizable correction spaces. We use a known closed subspace $\mathscr{V} \subseteq \ker \mathcal{M}$. For a bounded observed control vector $\zeta(O) \in \mathbb{R}^q$ with $\mathbb{E}[\zeta \mid X, \tau] = 0$ and bounded permitted features $\phi_p(a, r, X) \in \mathbb{R}^p$, fix a reference report $r_{\mathrm{ref}}(a, X)$ and optimize $\Theta \in \mathbb{R}^{q \times p}$:

$$v_\Theta(a, r, O) = \zeta(O)^\top \Theta \widetilde{\phi}_p(a, r, X), \qquad \widetilde{\phi}_p(a, r, X) = \phi_p(a, r, X) - \phi_p(a, r_{\mathrm{ref}}(a, X), X). \tag{8}$$

These corrections have zero mean for every candidate because their features depend only on metadata, reports, and contexts. Outcome-dependent features require a correspondingly stronger centering condition. Anchoring removes report-independent terms without changing comparisons. A product of a kernel on (a, r, X) and the control inner product realizes an RKHS of such corrections. Nested closed subspaces inherit one Hilbert norm, ensuring that enlarging capacity changes the available comparisons without changing the penalty on an existing function. Product-kernel and anchoring details are in Appendix D.

For a loss S, write $L_{\tau,S}(h) = \mathbb{E}_{P_\tau} S(a_\tau, h(X), O)$. Every $v \in \mathscr{V}$ satisfies, for almost every task and all bank members h, g,

$$L_{\tau,S_{\mathrm{b}}+v}(h) - L_{\tau,S_{\mathrm{b}}+v}(g) = L_{\tau,S_{\mathrm{b}}}(h) - L_{\tau,S_{\mathrm{b}}}(g). \tag{9}$$

Thus corrections preserve both population minimizers and their excess losses. For the direct base S_0, these excess losses equal the welfare gaps $B_{\tau,i}$. For other bases, any incorrect population ordering is also preserved. Proposition 11 quantifies the effect of approximate centering.

Entropic construction. For each of B_{con} construction tasks, collect n_b loss observations and an independent welfare block. Set $\overline{W}_{b,i}$ to candidate i's average welfare on the latter block, and $\widehat{B}_{br} = \max_i \overline{W}_{b,i} - \overline{W}_{b,r} \ge 0$. Write $\xi_{bij} = (a_b, h_{b,i}(X_{bj}), O_{bj})$. For multipliers $\lambda_{br} \in \mathbb{R}_+^{M-1}$ with a fixed budget $\|\lambda_{br}\|_1 \le T < \infty$, define

$$d_{brj}(v) = \sum_{i \ne r} \lambda_{br,i}[(S_{\mathrm{b}} + v)(\xi_{brj}) - (S_{\mathrm{b}} + v)(\xi_{bij})], \qquad \widehat{M}_{br}(v) = \frac{1}{n_b}\sum_j e^{-d_{brj}(v)}, \tag{10}$$

$$\widehat{F}(v; \lambda) = \frac{1}{B_{\mathrm{con}}}\sum_{b,r} \widehat{B}_{br}\widehat{M}_{br}(v)^m + \frac{\rho}{2}\|v\|_{\mathscr{H}}^2, \qquad \rho > 0. \tag{11}$$

The nonnegative weights emphasize candidates with larger estimated welfare gaps. Since both weights and moments are estimated, (11) serves as a construction objective; a population guarantee additionally requires control of source estimation error.

Theorem 2 (Positive-cone representation)**.** *For fixed multipliers,* (11) *has a unique minimizer in $\mathscr{V}$. Let $P_{\mathscr{V}}$ be orthogonal projection onto $\mathscr{V}$ and $k_\xi = k(\xi, \cdot)$. Define*

$$t_{brj} = P_{\mathscr{V}} \sum_{i \ne r} \lambda_{br,i}(k_{\xi_{brj}} - k_{\xi_{bij}}).$$

Its minimizer satisfies

$$v^\star = \sum_{b,r,j} \alpha_{brj} t_{brj}, \qquad \alpha_{brj} = \frac{m\widehat{B}_{br}}{\rho B_{\mathrm{con}} n_b} \widehat{M}_{br}(v^\star)^{m-1} e^{-d_{brj}(v^\star)} \geq 0. \tag{12}$$

The exponential objective is convex in v and the regularizer is strongly convex; differentiation in $\mathscr{V}$ gives (12): the correction combines estimated welfare-weighted comparison functionals projected into the admissible space, with larger weights on unfavorable observations—a finite representation even when the ambient RKHS is infinite dimensional (proof in Appendix B).

Construction bias. The data term substitutes empirical weights and moments into Theorem 1's bound. At horizons $m > 1$, powering an empirical moment introduces correction-dependent upward bias. For one task and candidate, abbreviate $n = n_b$ and $d_j = d_{brj}$. For $n \geq m$, the U-statistic averaging $\exp(-\sum_{j\in I} d_j(v))$ over all m-element subsets I of the n observations is unbiased for the population powered moment, retains convexity, and can be evaluated in $O(nm)$ operations (Appendix D.2). Correcting this estimation bias does not remove the more fundamental discrepancy with decision risk.

Theorem 3 (A convex-criterion obstruction)**.** *For every $0 < \eta < 1/2$, there is a bounded two-candidate task and a one-dimensional centered correction family S_t, $0 \leq t \leq 1$, based on its direct loss S_0, with the following properties. For every horizon m and every finite convex function $\Psi : \mathbb{R}^m \to \mathbb{R}$, the expectation $\mathbb{E}\Psi(D_{t,1}, \ldots, D_{t,m})$ is minimized at $t = 0$, where $D_{t,j}$ is the harmful-minus-good comparison on observation j. The minimizer is unique if Ψ is strictly convex or the penalty $\rho t^2/2$ is added with $\rho > 0$. Nevertheless,*

$$R_1(S_0) = \eta^{-1} \inf_{0\leq t\leq 1} R_1(S_t), \qquad \inf_{0\leq t\leq 1} R_1(S_t) > 0.$$

Even the Chernoff bound optimized over its multiplier uniquely prefers $t = 0$ at every horizon.

The construction has $D_t = D_0 + t\zeta$, with ζ independent and centered. Conditional Jensen favors $t = 0$, whereas the added noise corrects frequent small negative comparisons at the cost of rare errors. Appendix A gives the bounded four-atom construction and exact risks. Thus neither convexity nor population preservation ensures decision-risk calibration. Direct episodic learning in Section 5 controls this gap statistically.

Approximation and statistical scope. For fixed multipliers, nested finite-space optima converge in Hilbert norm when their union is dense (Corollary 3). A welfare-threshold margin condition converts loss approximation into risk approximation, and a uniform support margin lets sufficiently accurate approximations retain zero risk (Proposition 1; Corollary 4). A separate task-level bound controls source estimation of the entropic criterion, whose exponential envelope can be conservative (Proposition 2). Joint fitting of corrections and multipliers is generally nonconvex; the representation and approximation results condition on fixed multipliers. Appendix D also separates source estimation, approximation, optimization, and validation errors from the remaining construction-to-risk gap.

Relation to DFL losses. The Smart Predict-then-Optimize surrogate SPO+ (Elmachtoub & Grigas, 2022) is convex in the predicted cost report. Adding a conditionally centered correction that is affine in that report preserves both convexity and the population objective. Fisher consistency—optimal decisions at population minimizers over an unrestricted predictor class—does not ensure correct ordering within a fixed bank; a correction preserves any such preference reversal. Learned approximations to decision loss likewise require accuracy on the reports produced by the candidate class. For direct loss, population alignment is immediate, but preserving its discontinuities does not generally provide useful training gradients. Appendix H establishes these statements and the relation to action-cone methods.

5 Learning and validating by downstream risk

Independent episodes can optimize a correction family directly or select among previously constructed losses. An episode consists of a fresh task, m training observations, and $v \geq 1$ independent

evaluation observations. On task b, let $\widehat{i}_{b,S} = \widehat{i}_{\tau_b,m,S}$ be the candidate selected by loss S on the shared m training observations. Compute

$$\overline{W}_{b,i} = \frac{1}{v}\sum_{\ell=1}^{v} w_{\tau_b}(h_{\tau_b,i}, O'_{b\ell}), \qquad L_b(S) = \max_i \overline{W}_{b,i} - \overline{W}_{b,\widehat{i}_{b,S}}. \tag{13}$$

The empirical maximum welfare has an upward bias shared by every loss:

$$\mathbb{E}L_b(S) = \mathcal{R}_m(S) + \kappa_v, \qquad \kappa_v = \mathbb{E}_\tau[\mathbb{E}\max_i \overline{W}_{\tau,i} - W^\star_\tau] \geq 0. \tag{14}$$

The selected candidate is independent of the evaluation block conditional on its task and training block, so its evaluated welfare is unbiased, while Jensen's inequality gives the nonnegative common bias in (14). For any loss collection $\mathcal{L}$ evaluated on B episodes, the empirical maximum cancels from every difference of episode scores:

$$\arg\min_{S\in\mathcal{L}} \frac{1}{B}\sum_b L_b(S) = \arg\max_{S\in\mathcal{L}} \frac{1}{B}\sum_b \overline{W}_{b,\widehat{i}_{b,S}}, \tag{15}$$

so evaluating the distinct selected candidates suffices for finite-library comparison. A search over a continuous family can require more candidates. In both cases the optimal population welfare need not be known.

5.1 Direct learning over a correction family

Fix measurable functions $v_1, \ldots, v_d$ and a nonempty Borel parameter set $\mathcal{T} \subseteq \mathbb{R}^d$, independently of N fitting episodes. Consider $S_\vartheta = S_\mathrm{b} + \sum_{\ell=1}^d \vartheta_\ell v_\ell$, $\vartheta \in \mathcal{T}$. Choosing $v_\ell \in \ker \mathcal{M}$ preserves population comparisons; structural centering supplies such functions. Here d counts scalar correction parameters, including all control-feature products. Write $\widehat{L}_N(S) = N^{-1}\sum_{b=1}^N L_b(S)$.

Theorem 4 (Direct episodic learning)**.** *Let $d \geq 1$, $N \geq 1$, and $0 < \delta < 1$. For the fixed affine family above and independent episodes at horizon m, there is a universal constant $C > 0$ such that the following holds with*

$$\varepsilon_{N,d} = C\sqrt{\frac{d\log(2M) + \log(2/\delta)}{N}}.$$

With probability at least $1-\delta$, simultaneously for all $\vartheta \in \mathcal{T}$, $|\widehat{L}_N(S_\vartheta) - \mathcal{R}_m(S_\vartheta) - \kappa_v| \leq \varepsilon_{N,d}$. Consequently, for $\varepsilon_{\mathrm{opt}} \geq 0$, a measurable $\varepsilon_{\mathrm{opt}}$-approximate empirical minimizer $\widehat{\vartheta}$ satisfies

$$\mathcal{R}_m(S_{\widehat{\vartheta}}) \leq \inf_{\vartheta\in\mathcal{T}} \mathcal{R}_m(S_\vartheta) + 2\varepsilon_{N,d} + \varepsilon_{\mathrm{opt}}. \tag{16}$$

No margin, bounded correction norm, or comparison moment-generating function is required. Bounded welfare and finite loss evaluations suffice.

Why discontinuities remain learnable. Across N episodes, the selected indices are determined by at most $N\binom{M}{2}$ affine comparison hyperplanes in parameter space. Counting all their faces, including ties, bounds the number of score vectors by $[1 + NM(M-1)]^d$. An entropy refinement of this count, followed by symmetrization and bounded differences, gives the theorem. This applies the piecewise-performance viewpoint of data-driven algorithm design (Balcan et al., 2021) to shared loss corrections. The bound competes with the best correction in the family directly; its dependence on m enters through the target risk and acquisition cost, rather than an exponential envelope. Exact optimization is combinatorial in general. Histogram corrections admit independent threshold sweeps when each task uses only one cell coefficient. Appendix A gives proofs, capacity selection, and optimization details.

Theorem 5 (Task-information lower bound)**.** *For every $d, N \geq 1$, there is a family of bounded two-candidate task models with a d-parameter affine, exactly population-preserving correction class $\{S_t : t \in [-1/2, 1/2]^d\}$ such that every procedure returning a correction from this class after N independent source tasks satisfies, in at least one model,*

$$\mathbb{E}\Big[\mathcal{R}_1(S_{\widehat{t}}) - \inf_t \mathcal{R}_1(S_t)\Big] \geq \frac{1}{300}\min\{1, \sqrt{d/N}\}.$$

This holds even if the complete observation law of each sampled source task is revealed to the procedure.

The construction assigns unknown majority sensor orientations to d observable descriptor cells. Learning the best correction requires learning these majorities from independently drawn tasks; more observations within a task do not supply new draws of its orientation. Thus the square-root dependence on dimension and task count is minimax optimal for fixed bank size. The bound does not establish a sharp dependence on M or m.

5.2 Independent validation of a frozen library

Freeze $\mathcal{L} = \{S_0, \dots, S_{K-1}\}$, with direct loss first, and use B new validation episodes. The library can contain both entropic and directly fitted losses, with all construction data held fixed.

Theorem 6 (Actual-risk selection and safe improvement)**.** *Let B selection episodes be independent and let*

$$\widehat{L}_B(S) = B^{-1} \sum_{b=1}^{B} L_b(S), \qquad \epsilon_B = \sqrt{\frac{\log(2K/\delta)}{2B}}, \quad 0 < \delta < 1.$$

Conditional on all construction data and randomness, with probability at least $1 - \delta$, the following hold simultaneously. The empirical minimizer $\widehat{S} = \min \arg\min_{S \in \mathcal{L}} \widehat{L}_B(S)$, with ties resolved by library order, satisfies

$$\mathcal{R}_m(\widehat{S}) \le \min_{S \in \mathcal{L}} \mathcal{R}_m(S) + 2\epsilon_B, \tag{17}$$

$$\mathcal{R}_m(S) \le \min\{1, \widehat{L}_B(S) + \epsilon_B\}, \qquad S \in \mathcal{L}. \tag{18}$$

Define $S_{\mathrm{safe}} = \widehat{S}$ if $\widehat{L}_B(\widehat{S}) + 2\epsilon_B \le \widehat{L}_B(S_0)$, and $S_{\mathrm{safe}} = S_0$ otherwise. Then

$$\mathcal{R}_m(S_{\mathrm{safe}}) \le \mathcal{R}_m(S_0), \qquad \mathcal{R}_m(S_{\mathrm{safe}}) \le \min_{S \in \mathcal{L}} \mathcal{R}_m(S) + 4\epsilon_B. \tag{19}$$

The guarantee concerns expected risk on a fresh task, conditional on the constructed library. The gate can reject useful losses when the validation budget is small. A paired variance-adaptive gate and the extension to frozen predictor-training pipelines appear in Appendix B; the latter compares independently evaluated pipeline outputs without requiring a fixed candidate bank.

Construction and selection errors. The finite-library guarantee separates best-member risk from validation error. Direct fitting additionally controls excess risk within each affine family by (16). Exponential fitting supplies convex optimization and support certificates, while its decision risk can retain the gap in Theorem 3. Both types of fit can enter the same independently validated library.

6 Controlled experiments

We use controlled models to separate source estimation, functional capacity, and the construction-to-risk gap. Exact conditional risks and population oracles permit comparisons with the best correction in each function class. The sensor and contextual fits use nested histogram spaces with a common norm. Appendix E gives the protocols, solver diagnostics, and acquisition costs. Reported intervals are 95% confidence intervals over independent repetitions of construction and evaluation. Appendix I presents supplementary transfer studies.

6.1 Sensor cells: capacity, tail information, and the objective gap

Each task has a visible descriptor A uniform on $\mathcal{A}_{.2} = \{a \in [-1,1]^2 : |\sin(2\pi a_1) + .6a_2| \ge .2\}$ and an independent, equiprobable hidden sign $\theta \in \{-1, 1\}$. The orientation $e(a) = \mathrm{sign}(\sin(2\pi a_1) + .6a_2)$ is not supplied to the learner and is balanced: $\mathbb{P}(e(A) = 1) = 1/2$. Let independent V, U satisfy $V = .5, 3$ at probabilities $.8, .2$ and $U = -1, 1.5$ at $.2, .8$ (both have mean and variance one). Then $(Y_1, Y_2) = (\theta V, \theta U)$ when $e(A) = 1$, with the coordinates swapped otherwise. The learner observes A, Y_1, Y_2 but not θ or $e(A)$. The bank consists of the two constant actions $u \in \{-1, 1\}$, with welfare $w(u, O) = 1/2 + uY/6$, where $Y = (Y_1 + Y_2)/2$. With $Z = Y_1 - Y_2$, we learn $S_t(u, O) = -u[Y + t(A)Z]$, $|t(A)| \le 1/2$; at $t = 0$ this is a positive affine rescaling of negative welfare. Because $\mathbb{E}[Z \mid A, \theta] = 0$, every function preserves population preferences. With $s = e(A)t(A)$, the signed corrected observation $Q_s = (1/2 + s)V + (1/2 - s)U$ has mean one, variance

$1/2 + 2s^2$, and minimum support value $-1/4 + 3s/2$: variance minimization leaves a negative tail, $s > 1/6$ removes it, the direct loss ($t = 0$) has one-observation risk $4/75$, and the best constant correction has $2/75$ (Proposition 4).

The population entropic objective is even and strictly convex over constant corrections, and uniquely selects $t = 0$ for every positive quadratic regularizer (Proposition 7). In a mixed histogram cell, let π be the fraction of positive orientations. At multiplier four, the unregularized optimum crosses the safe threshold only when $\pi > 0.9821$ (or symmetrically $\pi < 0.0179$). At $\pi = .98$ its one-observation risk is fifty times the cell optimum (Appendix F.1).

Protocol. We use 100 repetitions, $N \in \{32, 128, 512, 2048\}$ source tasks, independent blocks of 64 pilot, 64 welfare, and 32 comparison observations, and nested histograms with $p \in \{1, 8, 32, 128\}$ cells. Cell coefficients lie in $[-1/2, 1/2]$. Entropic fits use multiplier four, horizon one, and $\rho = .01$. Baselines use the same blocks: direct loss, a fitted constant, cellwise variance minimization, sensor-endpoint selection, and a same-class pilot-score fit by threshold enumeration. The last method minimizes a welfare-weighted sign score against independent pilot signs, rather than an unbiased episodic welfare score. Population entropic and risk-optimal corrections are excluded from validation. Target risks use exact conditional enumeration and independent descriptor integration.

Results. The fitted constant attains risk 26.7×10^{-3}, approximately half the direct risk 53.3×10^{-3}, while variance minimization remains near the direct-loss level (Table 3). At $p = 128$ and $m = 1$, same-class empirical-risk fitting yields lower risk than entropic fitting: 2.43 versus 3.08 at $N = 512$, and 1.15 versus 3.37 at $N = 2048$, all in units of 10^{-3}. The difference persists at the population level: the entropic oracle attains 5.29×10^{-3}, compared with 1.08×10^{-3} for the risk-optimal oracle. As capacity increases from one to 128 cells, the descriptor mass of cells whose population entropic correction remains below the safe threshold decreases from 1.0 to 0.10 (Figure 1; Table 4). At $N = 512$, the entropic-only, non-entropic, and combined-library selectors attain risks 3.08, 2.42, and 2.43×10^{-3}, respectively (Table 5). The safety gate returns direct loss in every repetition: its threshold 0.0805 exceeds the largest possible population improvement, 0.0533 (Appendix I.1).

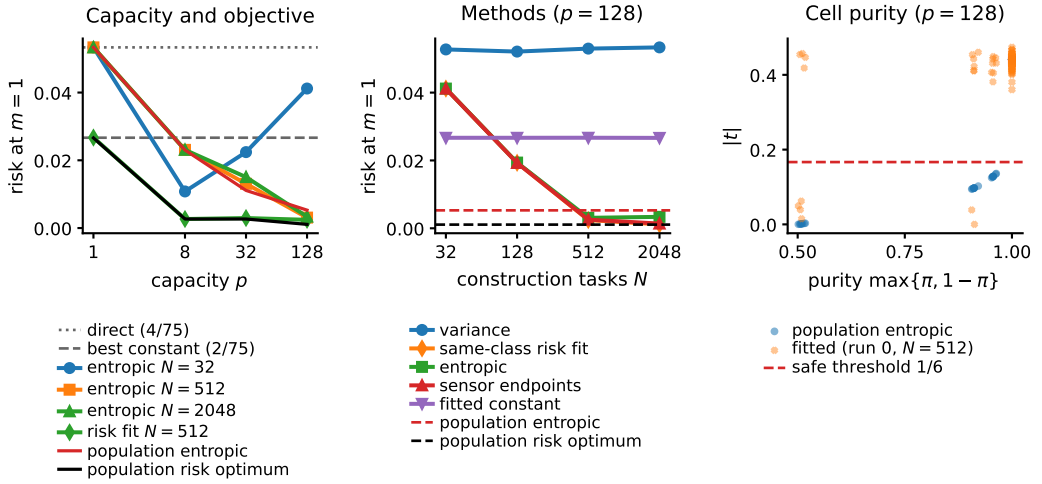

Figure 1: Sensor study at horizon $m = 1$. Left: risk versus capacity for sample-based entropic fits, the same-class empirical-risk baseline at $N = 512$, and the two population oracles; dotted lines are direct loss ($4/75$) and the best constant ($2/75$). Middle: equal-capacity methods at $p = 128$. Right: population entropic cell optima at $p = 128$ versus cell orientation purity, with fitted cells of the predesignated first repetition; the population optimum remains below the safe threshold except in nearly pure cells.

Direct episodic fitting. The preceding risk baseline uses estimated pilot signs. A separate study implements Theorem 4's welfare objective exactly: each source task supplies one training observa-

tion and 32 independent evaluation observations, with no pilot labels. Both episodic and entropic fits use these same blocks. At $p = 128$, entropic fitting is better at $N = 2048$ ($5.51 \pm .20$ versus $10.02 \pm .23$), but episodic fitting is better at $N = 8192$ ($1.56 \pm .04$ versus $2.95 \pm .08$), in risk units 10^{-3}. At $p = 1$, episodic fitting approaches $2/75$, whereas entropic fitting retains $4/75$. These 100-repetition results show both the estimation cost of direct fitting and its smaller population objective gap (Table 6). Rational comparison roots retain exact empirical ties; Appendix E.6 gives the protocol.

6.2 Joint comparisons versus tail information

The cancellation model has three candidates with welfare $(1, 1, 0)$ and losses $(a(t)Z, -a(t)Z, d + s(t)Y)$. Here Z, Y are independent Rademacher variables (equiprobable signs), $0 \le t \le 1$, $0 < \varepsilon < 1/2$, $d = 1/2$, $a(t) = 1 - (1 - \varepsilon)t$, $s(t) = \varepsilon + (1 - \varepsilon)t$. A single-reference criterion sees variance $a(t)^2 + s(t)^2$, uniquely minimized at $t = 1/2$ with $R_2 = 1/8$; the loss $t = 0$ is support-safe at every horizon. Indeed, the equal mixture of the first two losses is zero, while the third loss at $t = 0$ is at least $d - \varepsilon > 0$. This is the convex-reference certificate in (4), although neither good candidate alone dominates the third on every observation. Crossing single versus joint references with quadratic versus exponential objectives at budgets $T \in \{.5, 2, 4\}$ isolates the mechanism: both joint methods select $t = 0$, while both single-reference methods select $t = 1/2$. Thus full covariance already captures this cancellation; the sensor study separately tests tail information. Estimated-law variants recover these choices with 32 tasks and 16 observations per task (Tables 7 and 8).

Construction horizon and the powered-moment bias. We compare horizon-one, matched-horizon powered-moment, unbiased U-statistic, exact-law, and empirical-risk construction at $m \in \{1, 8, 32\}$, with fixed task count or acquisition budget (20 repetitions; Table 9). Removing powered-moment bias gives similar risks. Matching the horizon matters more: at $N = 512$, $p = 128$, and $m = 8$, risk falls from 19×10^{-6} to $1\text{–}3 \times 10^{-6}$, matching exact-law construction. Powered-moment fitting is 10–40 times faster than the U-statistic in this study. At $p = 1$, the construction objective and one-observation risk are anticorrelated (Spearman $\approx -.54$), as predicted by Proposition 7.

Unknown contextual nuisance and action-dependent shapes. We also study a four-action bank with task-specific mean costs and a shared nonlinear nuisance multiplied by an observed centered control. Corrections use context-dependent coefficients that are linear in the action coordinates, with an additional action interaction in the larger family (Appendix E). For the linear family, increasing capacity from $p = 1$ to $p = 32$ reduces risk by $(0.219 \pm 0.044) \times 10^{-3}$. At $p = 128$, adding the action interaction increases risk by $(0.050 \pm 0.013) \times 10^{-3}$, despite expanding the correction space beyond common outcome denoising. Outcome regression achieves lower risk than linear entropic fitting by $(0.058 \pm 0.010) \times 10^{-3}$. The entropic-only, non-entropic, and combined-library selectors attain 0.100, 0.072, and 0.075×10^{-3}, respectively (Table 10). Thus functional capacity helps in this model, but the additional entropic candidates do not improve library selection.

7 Conclusion

Population-preserving corrections expose a distinction between convex comparison control and direct decision-risk learning. Joint comparison laws provide support certificates and tractable entropic construction, but convex criteria can have an unbounded risk ratio. Direct episodic fitting learns the best correction in an affine family without a margin assumption; the task-information lower bound shows that its dependence on independent tasks is essential. Functional approximation and independent validation connect these results to reusable loss design. The analysis permits varying fixed banks and shared contextual corrections. Its limits are taskwise centering, representative source tasks, and the computational cost of direct fitting in general families. Extending the construction theory to adaptive candidate generation and finding efficient risk-consistent optimization beyond separable families remain open.

AI USE STATEMENT

Generative AI tools assisted with the theoretical and methodological development of this work and with manuscript revision. OpenAI Codex was used to refine the exposition, check notation and consistency, and verify the compiled LaTeX layout. The authors retain responsibility for the mathematical claims, experimental results, and final text.

REPRODUCIBILITY STATEMENT

Appendices A, B, and D provide proofs of the principal results and the approximation and statistical guarantees. Appendix I records the protocols and retained outputs of the supplementary exploratory studies, whose earlier implementations are not included in this source release.

REFERENCES

Maria-Florina Balcan, Dan DeBlasio, Travis Dick, Carl Kingsford, Tuomas Sandholm, and Ellen Vitercik. How much data is sufficient to learn high-performing algorithms? generalization guarantees for data-driven algorithm design. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021. doi: 10.1145/3406325.3451036.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002. URL `https://jmlr.org/papers/v3/bartlett02a.html`.

Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12: 149–198, 2000.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer, 2 edition, 1998.

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, 2017.

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

Michael C. Fu, Jian-Qiang Hu, Chun-Hung Chen, and Xiaoping Xiong. Optimal computing budget allocation under correlated sampling. In *Proceedings of the 2004 Winter Simulation Conference*, pp. 595–603, 2004.

Paul Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, 2004.

Peter W. Glynn and Sandeep Juneja. A large deviations perspective on ordinal optimization. In *Proceedings of the 2004 Winter Simulation Conference*, pp. 577–585, 2004.

Vishal Gupta and Michael Huang. Decision-focused learning with directional gradients. In *Advances in Neural Information Processing Systems*, volume 37, pp. 79194–79220, 2024. doi: 10.52202/079017-2514.

David Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik–Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995. doi: 10.1016/0097-3165(95)90052-7.

Anton J. Kleywegt, Alexander Shapiro, and Tito Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

Jayanta Mandi, James Kotary, Senne Berden, Maxime Mulamba, Victor Bucarey, Tias Guns, and Ferdinando Fioretto. Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research*, 80:1623–1701, 2024. doi: 10.1613/jair.1.15320.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *Computational Learning Theory*, pp. 416–426. Springer, 2001.

Noah Schutte, Krzysztof Postek, and Neil Yorke-Smith. Robust losses for decision-focused learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 4868–4875, 2024. doi: 10.24963/ijcai.2024/538.

Sanket Shah, Kai Wang, Bryan Wilder, Andrew Perrault, and Milind Tambe. Decision-focused learning without decision-making: Learning locally optimized decision losses. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1320–1332, 2022. doi: 10.52202/068431-0097.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Bo Tang and Elias B. Khalil. Cave: A cone-aligned approach for fast predict-then-optimize with binary linear programs. *arXiv preprint arXiv:2312.07718*, 2023.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1658–1665, 2019.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pp. 423–435. Springer, 1997. doi: 10.1007/978-1-4612-1880-7_29.

A DIRECT EPISODIC LEARNING AND ITS LIMITS

A.1 A QUANTITATIVE OBSTRUCTION FOR CONVEX COMPARISON CRITERIA

Proof of Theorem 3. Fix $0 < \eta < 1/2$. Let independent random variables D_0 and ζ have laws

$$\mathbb{P}(D_0 = -\eta/4) = 1 - \eta, \qquad \mathbb{P}(D_0 = 1/2) = \eta,$$
$$\mathbb{P}(\zeta = \eta/2) = 1 - \eta, \quad \mathbb{P}(\zeta = -(1-\eta)/2) = \eta.$$

Consider one task with constant context and two constant reports, indexed as good and bad. Define their observed welfares and losses by

$$w_{\mathrm{g}} = (1 + D_0)/2, \quad w_{\mathrm{b}} = (1 - D_0)/2,$$

$$S_t(\mathrm{g}, O) = -w_{\mathrm{g}}, \quad S_t(\mathrm{b}, O) = -w_{\mathrm{b}} + t\zeta, \quad 0 \le t \le 1.$$

Here $O = (D_0, \zeta)$. Both welfares lie in $[0, 1]$, S_0 is direct loss, and $\mathbb{E}\zeta = 0$. The correction is an observed centered control multiplied by the bad-report indicator. Its population mean is zero for each report, and the welfare gap is

$$b_\eta = \mathbb{E}D_0 = \eta(1 + \eta)/4 > 0.$$

The harmful-minus-good comparison is $D_t = D_0 + t\zeta$. The losses are uniformly bounded in absolute value by one, independently of η, t. For independent observations let $\mathbf{D}_t = (D_{t,1}, \ldots, D_{t,m})$. Independence and centering give $\mathbb{E}[\mathbf{D}_t \mid \mathbf{D}_0] = \mathbf{D}_0$. Conditional Jensen thus yields

$$\mathbb{E}\Psi(\mathbf{D}_t) \ge \mathbb{E}\Psi(\mathbf{D}_0)$$

for every finite convex $\Psi : \mathbb{R}^m \to \mathbb{R}$. For $t > 0$ the conditional distribution of $\mathbf{D}_t$ is nonconstant, so the inequality is strict when Ψ is strictly convex. Adding a penalty minimized at zero retains zero as a minimizer; a positive quadratic penalty makes it unique even when the comparison criterion is only convex.

At $t = 0$, the comparison is negative with probability $1 - \eta$. At $t = 1$, the four values, indexed by the two draws, are

$$\eta/4, \quad -(2 - \eta)/4, \quad (1 + \eta)/2, \quad \eta/2.$$

Only the second is negative; it has probability $\eta(1 - \eta)$. No endpoint comparison is zero. Consequently

$$R_1(S_0) = b_\eta(1 - \eta), \qquad R_1(S_1) = b_\eta\eta(1 - \eta).$$

In fact $\inf_{0 \le t \le 1} R_1(S_t) = R_1(S_1)$. The event $\{D_0 = -\eta/4, \zeta = -(1 - \eta)/2\}$ gives a strictly negative comparison for every admissible t, and has the latter probability. Every $t > 1/2$ attains this lower bound. Hence the ratio is exactly $1/\eta$.

The same family also defeats optimizing the Chernoff multiplier. Write $M_t(\lambda) = \mathbb{E}e^{-\lambda D_t}$. Independence gives $M_t(\lambda) = M_0(\lambda)\mathbb{E}e^{-\lambda t\zeta}$, and the last factor strictly exceeds one for $t, \lambda > 0$. For each t, $\mathbb{E}D_t = b_\eta > 0$ and D_t has both signs. Thus the maximizer of $-\log M_t(\lambda)$ over $\lambda \ge 0$ is finite and strictly positive: its derivative at zero is b_η, and the objective tends to $-\infty$ at infinity. Testing that maximizer in the strict moment comparison shows $J(S_t) < J(S_0)$ for every $t > 0$. The argument also holds with any positive finite multiplier budget. Thus even an exactly optimized exponential bound prefers zero. The ratio conclusion concerns one-observation risk; it does not assert that S_1 dominates at all horizons. □

The phenomenon is a consequence of convex order, the order defined by expectations of convex functions. Conditional averaging improves every convex comparison criterion, but the indicator of a harmful selection is not convex. It is compatible with the Gaussian comparison in Corollary 1, which imposes additional distributional structure. It is also compatible with the absolute regret in this family vanishing as $\eta \downarrow 0$; the impossibility is a uniform *multiplicative* guarantee.

A.2 COUNTING DECISIONS, INCLUDING TIES

For integers $H, d \geq 0$, define

$$A(H,d) = \sum_{j=0}^{\min(H,d)} 2^j \binom{H}{j}.$$

This quantity bounds the number of realized sign vectors in $\{-1,0,1\}^H$ of H affine functions on $\mathbb{R}^d$. One proof adds their zero hyperplanes successively. The nonempty sets having a fixed sign vector are relatively open convex faces. On adding one hyperplane, each face that is cut is replaced by two open pieces and its intersection, increasing the count by two. These intersections form faces of the restricted arrangement in dimension at most $d-1$. Faces contained in the new hyperplane do not split. Thus the maximal count satisfies

$$A_{\max}(H,d) \leq A_{\max}(H-1,d) + 2A_{\max}(H-1,d-1), \quad A_{\max}(0,d) = A_{\max}(H,0) = 1.$$

Induction and Pascal's identity give the displayed bound. Identically zero functions and coincident hyperplanes cannot increase the count. For $d \geq 1$, one may use the simpler bound

$$A(H,d) \leq \sum_{j=0}^{d} (2H)^j \leq (1+2H)^d,$$

where the last inequality follows by expanding the right side and using $\binom{d}{j} \geq 1$.

Fix N episodes. For the affine family in Theorem 4, each empirical candidate comparison is an affine function of $\vartheta \in \mathbb{R}^d$. There are at most $H = N\binom{M}{2}$ such functions, irrespective of how many observations form each average. On every sign face, all minimizing candidates and their smallest-index tie selections are fixed. Hence the vectors of episode scores take at most $A(H,d)$ distinct values. Restricting parameters to $\mathcal{T} \subseteq \mathbb{R}^d$ cannot increase this number. In particular, counting only full-dimensional cells would be insufficient: a parameter on an intersection can produce a distinct selection vector.

A.3 UNIFORM EPISODIC CONVERGENCE

Proof of Theorem 4. Condition on any independent data used to specify the family. Write E_b for the complete bth episode and $f_\vartheta(E_b) = L_b(S_\vartheta) \in [0,1]$. All suprema below are assumed measurable; otherwise the same argument uses outer expectations and outer probabilities. This convention also applies to the empirical optimizing rule. Finite-dimensional Borel parameterizations on standard Borel observation spaces allow universally measurable versions.

Let $\sigma_1, \ldots, \sigma_N$ be independent equiprobable signs, independent of the episodes, and set

$$G_N = \sup_{\vartheta \in \mathcal{T}} \left| \frac{1}{N} \sum_b f_\vartheta(E_b) - \mathbb{E} f_\vartheta(E) \right|.$$

Introducing independent copies of the episodes and using Jensen's inequality, then randomly interchanging each original-copy pair, gives the usual symmetrization bound (Bartlett & Mendelson, 2002)

$$\mathbb{E} G_N \leq 2\mathbb{E}_{E,\sigma} \sup_{\vartheta \in \mathcal{T}} \left| \frac{1}{N} \sum_b \sigma_b f_\vartheta(E_b) \right|.$$

Put $H = N\binom{M}{2}$. Conditional on the episodes, the supremum ranges over at most $2A(H,d)$ vectors after including negatives. Each vector x has $\|x\|_2 \leq \sqrt{N}$. Since $\mathbb{E}_\sigma e^{t\sum_b \sigma_b x_b/N} \leq e^{t^2/(2N)}$, the log-sum-exp bound followed by minimization over $t > 0$ gives

$$\mathbb{E}_\sigma \sup_\vartheta \left| N^{-1} \sum_b \sigma_b f_\vartheta(E_b) \right| \leq \sqrt{2\log(2A(H,d))/N}.$$

Changing one episode changes G_N by at most $1/N$. The bounded-difference inequality therefore adds at most $\sqrt{\log(1/\delta)/(2N)}$ with probability $1-\delta$. The arrangement bound gives the explicit,

slightly looser radius

$$e_{N,d} = 2\sqrt{\frac{2[d\log(1+NM(M-1))+\log 2]}{N}} + \sqrt{\frac{\log(1/\delta)}{2N}}.$$

We next remove the extra $\log N$ factor. Let $\mathcal{F} = \{f_\vartheta : \vartheta \in \mathcal{T}\}$ and let V be its pseudo-dimension, defined as the Vapnik–Chervonenkis dimension of the subgraph sets $\{(E,u) : u < f_\vartheta(E)\}$. Thus n episode-threshold pairs are shattered if all 2^n binary membership vectors are attained. Put $c = \binom{M}{2}$. Shattering requires $2^n \le A(nc,d)$, since each sign face fixes all n scores, and therefore their threshold comparisons. For $n \ge d$, the elementary binomial bound gives

$$A(nc,d) \le \sum_{j=0}^{d} \binom{2nc}{j} \le (2enc/d)^d.$$

The first inequality follows by pairing $2nc$ elements: choosing j distinct pairs and one of their elements counts $2^j\binom{nc}{j}$ subsets. The second is the usual bound on a truncated binomial sum. Writing $x = n/d$, shattering implies $2^x \le 2ecx$. This fails for $x \ge 4\log_2(4ec)$: the ratio $2^x/x$ is increasing there and exceeds $2ec$ at that endpoint. Consequently

$$V \le \left\lceil 4d\log_2\left(4e\tbinom{M}{2}\right)\right\rceil \le C_1 d\log(2M)$$

for a universal constant C_1. This includes all tie faces.

For completeness, the standard packing-to-complexity step is as follows. For a probability law Q on episodes, the symmetric-difference distance between two subgraphs under Q times uniform measure on $[0,1]$ equals $\|f-g\|_{L^1(Q)}$. The VC packing bound of Haussler (1995) therefore gives, for universal constants $C_2, C_3 > 0$ and $0 < \epsilon < 1$,

$$\log \mathcal{N}(\epsilon, \mathcal{F}, L^2(Q)) \le C_2(V+1)\log(C_3/\epsilon).$$

Here $\mathcal{N}$ is the minimum cardinality of an ϵ-cover; we used $\|f-g\|^2_{L^2(Q)} \le \|f-g\|_{L^1(Q)}$ to pass from L^1 radius ϵ^2 to L^2 radius ϵ. Apply this with the empirical episode law. Chaining the covers at radii 2^{-j} bounds the absolute Rademacher average by a universal multiple of

$$\frac{1}{\sqrt{N}} \sum_{j \ge 0} 2^{-j}\sqrt{(V+1)(j+1)} \le C_4\sqrt{\frac{V+1}{N}}.$$

Indeed, write each score vector as a coarse representative plus successive nearest-cover increments. An increment at level j has empirical L^2 norm at most $3\cdot 2^{-j}$; the exponential maximum bound used above controls all such increments by their norm times the square root of the log cover cardinality. The remainder is at most the final cover radius by Cauchy–Schwarz and tends to zero. The coarse representative contributes at most $N^{-1/2}$ to the absolute average. This is the usual entropy argument for Rademacher complexity (Bartlett & Mendelson, 2002). Symmetrization and the same bounded-difference inequality now prove the radius $\varepsilon_{N,d}$ in the theorem, absorbing $C_1, \dots, C_4$ and $V + 1 \le C_5 d\log(2M)$ into its universal constant C. Because training and evaluation are conditionally independent, (14) holds for every fixed parameter, and therefore

$$\sup_\vartheta |\widehat{L}_N(S_\vartheta) - \mathcal{R}_m(S_\vartheta) - \kappa_v| \le \varepsilon_{N,d}.$$

Apply this event at the empirical approximate minimizer and an arbitrary population comparator. The shared κ_v cancels, giving (16) after taking the infimum. For a prescribed sample there are finitely many attainable score vectors, so an empirical minimum exists whenever $\mathcal{T}$ is nonempty. The theorem allows optimization error and assumes a measurable choice of output. □

Capacity selection and approximation. Suppose P affine families $\mathcal{L}_1, \dots, \mathcal{L}_P$, with dimensions $d_1, \dots, d_P$, are specified independently of the episodes. Their bases and admissible parameter sets may differ. Applying the theorem to family p at confidence $\delta_p > 0$, with $\sum_p \delta_p \le \delta$, controls all families simultaneously. Let e_p be its radius, and let $\widehat{S}_p$ minimize its empirical episode score to accuracy $o_p \ge 0$. Select a family by minimizing $\widehat{L}_N(\widehat{S}_p) + e_p$. On the same event the result $\widehat{S}$ obeys

$$\mathcal{R}_m(\widehat{S}) \le \min_{1 \le p \le P} \left\{ \inf_{S \in \mathcal{L}_p} \mathcal{R}_m(S) + 2e_p + o_p \right\}.$$

Indeed, the penalized score upper-bounds risk plus the common bias for the selected family and is no greater than the penalized score of any competitor. A second uniform comparison gives the display. Relative to a larger class $\mathcal{L}_\infty$, this separates approximation error $\inf_{\mathcal{L}_p} \mathcal{R}_m - \inf_{\mathcal{L}_\infty} \mathcal{R}_m$ from statistical and optimization errors, without an exponential construction-to-risk term. Proposition 1 can control this approximation error for RKHS subspaces. A margin is needed for that approximation step, not for uniform learning within each affine family.

Optimization and data acquisition. The arrangement proof is a statistical argument; general arrangement enumeration has a number of faces polynomial in NM^2 only for fixed d. For a one-dimensional interval, sorting the comparison roots and evaluating the roots and intervening intervals suffices. For histogram corrections, when each task involves only its own cell coefficient, the problem separates into independent one-dimensional searches. Repeated roots are evaluated once and retain the exact prescribed tie rule. Only candidates selected by some admissible parameter need welfare evaluation; evaluating the full bank suffices in every case.

Many training/evaluation pairs from one task are dependent after averaging over its latent law. They cannot be substituted for independent tasks in the theorem. Additional within-task observations can reduce evaluation noise and computational work can reuse them, but the displayed bound counts independent episodes drawn with fresh tasks. The horizon m and evaluation size v may be arbitrary positive integers fixed in advance. The absence of m in the complexity factor does not equate risks across horizons or remove the cost of collecting m training observations.

A.4 A LOWER BOUND WITHIN CENTERED CORRECTION FAMILIES

Proof of Theorem 5. Let the visible descriptor A be uniform on $\{1, \dots, d\}$. Fix an unknown vector $\omega \in \{-1, 1\}^d$ and a number $0 < \alpha \le 1/2$. Given $A = j$, draw a task orientation $e \in \{-1, 1\}$ with

$$\mathbb{P}_\omega(e = 1 \mid A = j) = (1 + \alpha\omega_j)/2.$$

Independently draw a task sign $\theta \in \{-1, 1\}$ uniformly. Within each task draw independent V, U with

$$\mathbb{P}(V = 1/2) = 4/5, \quad \mathbb{P}(V = 3) = 1/5, \qquad \mathbb{P}(U = -1) = 1/5, \quad \mathbb{P}(U = 3/2) = 4/5.$$

Let $(Y_1, Y_2) = (\theta V, \theta U)$ for $e = 1$, and swap coordinates for $e = -1$. The bank has the two actions $u \in \{-1, 1\}$. Define

$$Y = (Y_1 + Y_2)/2, \quad Z = Y_1 - Y_2, \quad w(u, O) = 1/2 + uY/6,$$

$$S_t(A, u, O) = -u[Y + t_A Z]/6 - 1/2, \quad t \in [-1/2, 1/2]^d.$$

The zero correction is exactly negative welfare, the corrections are affine in t, and $\mathbb{E}[Z \mid A, e, \theta] = 0$. All losses and rewards are bounded uniformly over the family. Thus the construction satisfies the task model and exact taskwise population preservation. The target loss receives A and the sensor observations, but its coefficient t_A does not receive the latent orientation or sign.

Put $c_0 = 4/75$. Conditional on an orientation and averaging over θ, let $r(s)$ be one-observation regret at signed coefficient $s = et_A$. The signed corrected observation is $Q_s = (1/2+s)V + (1/2-s)U$, and the welfare gap is $1/3$. Its four values and probabilities are

	1	2	3	4
Q_s	$-1/4 + 3s/2$	$1 - s$	$1 + 4s$	$9/4 + 3s/2$
$\mathbb{P}$	$4/25$	$16/25$	$1/25$	$4/25$

with ties charged half the welfare gap after averaging the two task signs. In particular $r(s) \ge c_0$ when $s \le 0$, and $r(s) + r(-s) \ge c_0$ for all $s \in [-1/2, 1/2]$. Also $r(1/5) = 0$ and $r(-1/5) = c_0$. Therefore the risk-optimal coefficient in cell j is $t_j = \omega_j/5$, with conditional risk $c_0(1 - \alpha)/2$. Decode any learned coefficient as $\widehat{\omega}_j = 1$ if $\widehat{t}_j \ge 0$ and -1 otherwise. When $\widehat{\omega}_j \ne \omega_j$, the majority orientation has signed coefficient at most zero. The conditional risk is at least $c_0(1+\alpha)/2$, exceeding the optimum by at least $c_0\alpha$. Thus, also for randomized procedures,

$$\mathcal{R}_{1,\omega}(S_{\widehat{t}}) - \inf_t \mathcal{R}_{1,\omega}(S_t) \ge \frac{c_0\alpha}{d} \sum_{j=1}^{d} \mathbf{1}\{\widehat{\omega}_j \ne \omega_j\}.$$

Grant the procedure the complete law of each of its N sampled source tasks, equivalently (A, e, θ) together with the known observation mechanism. Conditional source observations add no information about ω once these quantities are given. If $\omega^{(j)}$ differs only in coordinate j, the KL divergence between the distributions of these N task draws is

$$D_{\mathrm{KL}}(P_\omega^N \| P_{\omega^{(j)}}^N) = \frac{N\alpha}{d} \log \frac{1+\alpha}{1-\alpha} \leq \frac{4N\alpha^2}{d}.$$

Choose $\alpha = \frac{1}{4}\min\{1, \sqrt{d/N}\}$. The divergence is at most $1/4$. Total variation is at most $\sqrt{D_{\mathrm{KL}}/2} \leq 1/2$ by Pinsker's inequality, so the sum of the two testing errors for coordinate j is at least $1/2$. Averaging over the uniform prior on $\{-1, 1\}^d$ gives error probability at least $1/4$ per coordinate. This is the hypercube testing argument underlying Assouad's method (Yu, 1997). Substituting into the preceding regret bound yields

$$\frac{c_0\alpha}{4} = \frac{1}{300}\min\{1, \sqrt{d/N}\}.$$

At least one ω attains the prior-average lower bound. Since the learner was granted each sampled task's law, the bound applies with any number of conditionally independent observations from those tasks. It does not apply to active selection of descriptors or to a larger loss family that is allowed to choose a coefficient using the target task's latent orientation. □

A.5 Construction and validation procedure

Functional loss construction and selection

1. Specify nested spaces, base losses, and a capacity and regularization grid. Draw independent construction and selection task collections.
2. Fit each correction on construction tasks: use the entropic criterion with independent reference and welfare blocks, or optimize actual episodic welfare at the target horizon.
3. Include direct loss and other specified baselines, then freeze the library.
4. On independent episodes, apply each loss at the target horizon and evaluate each distinct selected candidate's welfare; by (15), select the member with highest average evaluated welfare (evaluate the full bank only for absolute certificates).
5. Return the empirical-risk minimizer, optionally applying Theorem 6's baseline-safety gate.

Table 1 summarizes the assumptions used throughout the construction and learning results.

Table 1: Standing assumptions, where they are used, and what a violation entails.

Assumption	Needed for	If violated
Banks fixed independently of target training observations	Finite-bank comparison construction (Section 3)	Adaptive candidate generation needs a separate analysis
Observed controls have mean zero conditional on context and task	Population preservation (Section 4)	Correction can introduce population preference bias
Independent construction and selection task collections	Frozen-library oracle inequality (Theorem 6)	Selection guarantee does not apply directly
Affine family fixed before fitting episodes	Direct episodic learning (Theorem 4)	Adapting the features requires additional complexity control
Same source and target task distribution	Meta-risk transfer and selection	Only source-distribution performance is controlled
Fixed or independently piloted multipliers	Uniform construction bound (Proposition 2)	Jointly fitted multipliers need additional complexity control
Selection horizon equal to deployment horizon	Theorem 6 at the deployment horizon	A different risk functional is being validated
Population support knowledge	Support-safety certificate (Theorem 1)	Observed support alone cannot rule out unseen tails

B PROOFS OF THE PRINCIPAL RESULTS

B.1 DISTRIBUTIONAL SEPARATION

Proof of Theorem 1. Fix r, abbreviate the selected candidate index as $\widehat{i}$, and set $D_j = D_r(S, O_j)$. The event that r is selected is contained in $\{\sum_j D_j \le 0\}$. For any $\lambda \ge 0$, the latter event implies $\exp(-\lambda^\top \sum_j D_j) \ge 1$. Markov's inequality and independence across observations give

$$\mathbb{P}\{\widehat{i} = r\} \le \mathbb{P}\{\overline{D}_{r,m} \le 0\} \le \left(\mathbb{E}e^{-\lambda^\top D_r(S,O)}\right)^m.$$

Taking the infimum over λ gives e^{-mJ_r}, including when the infimum is zero and is not attained. Since $R_m(S) = \sum_r B_r \mathbb{P}\{\widehat{i} = r\}$ and $R_m(S) \le 1$, this proves (2). The choice $\lambda = 0$ ensures $J_r \ge 0$. The same upper bound holds for a minimizing mixture: its weight on r is at most the indicator that r belongs to the minimizing face.

For finite support, write $D_j = D_r(S, o_j)$ and $F(q, \lambda) = D_{\mathrm{KL}}(q\|p) + \lambda^\top \sum_j q_j D_j$. The finite Gibbs variational identity is

$$-\log \sum_j p_j e^{-\lambda^\top D_j} = \min_{q \in \Delta_s} F(q, \lambda).$$

For completeness, let $p_j^\lambda = p_j e^{-\lambda^\top D_j} / \sum_k p_k e^{-\lambda^\top D_k}$. Then $F(q, \lambda) = D_{\mathrm{KL}}(q\|p^\lambda) - \log \sum_j p_j e^{-\lambda^\top D_j}$, whose minimum is attained at p^λ.

To justify duality even without a strictly feasible entropy constraint, restrict λ to $[0, t]^{M-1}$ for $t > 0$, and write $x_+ = \max\{x, 0\}$. Both domains are compact and convex; F is continuous, convex in q, and affine in λ. The minimax theorem (Sion, 1958) gives

$$\sup_{\lambda \in [0,t]^{M-1}} \min_q F(q, \lambda) = \min_{q \in \Delta_s} \{D_{\mathrm{KL}}(q\|p) + t\, a(q)\}, \qquad a(q) = \sum_{i \ne r} \left(\sum_j q_j D_{j,i}\right)_+.$$

The left side increases to J_r as $t \to \infty$. If no feasible q exists, compactness gives $\min_q a(q) > 0$ and the right side tends to infinity. If a feasible q exists, let $q^\star$ minimize its relative entropy and q_t minimize the penalized problem. Since relative entropy is nonnegative, $t a(q_t) \le D_{\mathrm{KL}}(q^\star\|p)$, so each convergent subsequence of q_t has a feasible limit. Continuity of relative entropy on the simplex (all $p_j > 0$) implies that the limiting objective is at least $D_{\mathrm{KL}}(q^\star\|p)$. The opposite inequality follows by testing $q^\star$. This proves (3), including boundary and infeasible cases.

For (4), boundedness makes K_r a nonempty compact convex set. A compact convex set disjoint from a closed convex set admits strict separation. Thus disjointness from $\mathbb{R}_-^{M-1}$ gives a nonzero λ such that

$$\inf_{d \in K_r} \lambda^\top d > \sup_{y \in \mathbb{R}_-^{M-1}} \lambda^\top y.$$

The right side is finite only if $\lambda \ge 0$, in which case it is zero. Normalize by $\mathbf{1}^\top \lambda$ to obtain q_r and a positive minimum δ_r. Conversely, a continuous linear functional bounded below by $\delta_r > 0$ almost surely has the same lower bound on the essential range and its closed convex hull. For $y \le 0$ and $q_r \ge 0$, $q_r^\top y \le 0$, proving disjointness.

Averaging the strict comparison over any finite sample preserves positivity. It is incompatible with $Z_r \le Z_i$ for all $i \ne r$, so candidate r is never an empirical minimizer. Taking a finite union over harmful r proves zero risk. On a finite alphabet, $K_r = \mathrm{conv}\{D_j\}$, so its intersection with the negative orthant is exactly feasibility of the entropy constraint. All feasible distributions have finite relative entropy; hence infeasibility, support separation, and $J_r = \infty$ are equivalent. □

Bounded support is needed for the positive-margin conclusion. Even with bounded support, zero risk need not imply a uniform support margin: if a scalar harmful comparison is uniform on $(0, 1)$, every empirical mean is positive, but its closed support includes zero. In this example $J_r = \infty$; the equivalence of infinite exponent and strict support separation was asserted only for a finite alphabet.

B.2 GAUSSIAN SPECIALIZATION

Proof of Corollary 1. Let e_i denote the ith coordinate vector and $A_r : \Lambda \to \mathbb{R}^{M-1}$ have rows $e_r^\top - e_i^\top$, $i \neq r$. Then $D_r = A_r P_\Lambda Z$ is Gaussian with $\mu_r = A_r g$ and $V_r = A_r \Sigma A_r^\top$. The Gaussian moment generating function yields

$$J_r = \sup_{\lambda \geq 0} \left\{ \lambda^\top A_r g - \tfrac{1}{2} \lambda^\top A_r \Sigma A_r^\top \lambda \right\}.$$

Consider $\inf_{z \in \Lambda : A_r z \leq 0} \frac{1}{2}(z-g)^\top \Sigma^{-1}(z-g)$. Minimizing its Lagrangian over $z \in \Lambda$ gives $z = g - \Sigma A_r^\top \lambda$ and the displayed dual objective. The feasible set has a strictly feasible point in Λ (take $z_r = -(M-1)$ and $z_i = 1$ for $i \neq r$). Slater's condition and positive definiteness therefore give equality of primal and dual values (Boyd & Vandenberghe, 2004). Since $\{z \in \Lambda : A_r z \leq 0\} = C_r \cap \Lambda$, this proves the claim. □

B.3 OPERATORS, STRUCTURAL CENTERING, AND FUNCTIONAL CONSTRUCTION

For every ξ, the reproducing property gives $|v(\xi)| \leq \kappa \|v\|_{\mathscr{H}}$. Orthogonal projection is contractive, so $\|\mathcal{A}v\|_{L^2}^2 \leq M\kappa^2 \|v\|_{\mathscr{H}}^2$. Conditional Jensen gives $\|\mathcal{M}v\|_{L^2} \leq \|\mathcal{A}v\|_{L^2}$. Their kernels are consequently closed, and $\mathcal{A}v = 0$ implies $\mathcal{M}v = 0$. For fixed task and candidate, conditional centering gives

$$\mathbb{E}_{P_\tau} f_\ell(a_\tau, h_{\tau,i}(X), X)\zeta_\ell(O) = \mathbb{E}_{P_\tau}[f_\ell(a_\tau, h_{\tau,i}(X), X)\mathbb{E}_{P_\tau}[\zeta_\ell(O) \mid X]] = 0.$$

Bounded functions and controls ensure integrability. For the product kernel in (22), let ϕ_0 be a feature map of k_0 and use $(\zeta_\ell(O)\phi_0(a, r, X))_{\ell=1}^q$ in a direct sum of q feature spaces of k_0. Each RKHS function has the indicated sum representation with bounded f_ℓ. This proves the population-preserving claim. A closed subspace of such functions remains in $\ker \mathcal{M}$ by continuity.

Pooled centering is insufficient even with two equiprobable tasks: a correction whose projected coefficient is $(1, -1)$ on the first and $(-1, 1)$ on the second has pooled mean zero but reverses taskwise comparisons. No estimate of the pooled mean resolves this problem.

Proof of Theorem 2. For $v \in \mathscr{V}$, the reproducing property implies $d_{brj}(v) = d_{brj}(0) + \langle v, t_{brj} \rangle_{\mathscr{H}}$. Let $\mathscr{T}$ be the span of the finitely many t_{brj}. Writing $v = v_{\mathscr{T}} + v_\perp$, the data term depends only on $v_{\mathscr{T}}$ and the regularizer is $\rho(\|v_{\mathscr{T}}\|^2 + \|v_\perp\|^2)/2$. Thus any minimizer belongs to the finite-dimensional, closed space $\mathscr{T}$. The objective is continuous there, nonnegative, and coercive, so it attains a minimum.

Each $\widehat{M}_{br}$ is convex and positive. For integer $m \geq 1$, $x \mapsto x^m$ is convex and nondecreasing on $[0, \infty)$, so each $\widehat{M}_{br}^m$ is convex. The quadratic term makes the full objective ρ-strongly convex, hence its minimizer is unique. Its gradient within $\mathscr{V}$ is

$$\nabla \widehat{F}(v) = \rho v - \frac{1}{B_{\mathrm{con}}} \sum_{b,r} \frac{m\widehat{B}_{br}}{n_b} \widehat{M}_{br}(v)^{m-1} \sum_j e^{-d_{brj}(v)} t_{brj}.$$

At the unconstrained minimizer in the subspace this gradient vanishes, which gives (12) and nonnegative coefficients. This explicit sign structure supplements the usual representer principle. □

B.4 EPISODIC SELECTION

Proof of Theorem 6. Condition on the construction data and randomness, so $\mathcal{L}$ is fixed. For each S, conditional independence of training and evaluation on a task implies $\mathbb{E}[\overline{W}_{\tau,\widehat{\imath}_S} \mid \tau, D_m] = W_{\tau,\widehat{\imath}_S}$. Subtracting from the expected evaluation maximum proves (14); Jensen's inequality gives $\kappa_v \geq 0$. The episodes are i.i.d. and $0 \leq L_b(S) \leq 1$. Hoeffding's inequality and a union bound therefore give the event

$$\mathcal{E} = \left\{ \left| \widehat{L}_B(S) - \mathcal{R}_m(S) - \kappa_v \right| \leq \epsilon_B \text{ for every } S \in \mathcal{L} \right\}, \qquad \mathbb{P}(\mathcal{E} \mid \text{construction}) \geq 1 - \delta.$$

Let $S^\star$ minimize true risk over $\mathcal{L}$. On $\mathcal{E}$,

$$\mathcal{R}_m(\widehat{S}) \leq \widehat{L}_B(\widehat{S}) - \kappa_v + \epsilon_B \leq \widehat{L}_B(S^\star) - \kappa_v + \epsilon_B \leq \mathcal{R}_m(S^\star) + 2\epsilon_B.$$

Dropping $\kappa_v \geq 0$ and using risk at most one proves the simultaneous certificate (18).

When the gate accepts,

$$\mathcal{R}_m(\widehat{S}) - \mathcal{R}_m(S_0) \leq \widehat{L}_B(\widehat{S}) - \widehat{L}_B(S_0) + 2\epsilon_B \leq 0.$$

When it rejects, the returned loss is exactly S_0 and $\widehat{L}_B(S_0) < \widehat{L}_B(\widehat{S}) + 2\epsilon_B \leq \widehat{L}_B(S^\star) + 2\epsilon_B$. Hence

$$\mathcal{R}_m(S_0) \leq \widehat{L}_B(S_0) - \kappa_v + \epsilon_B < \widehat{L}_B(S^\star) - \kappa_v + 3\epsilon_B \leq \mathcal{R}_m(S^\star) + 4\epsilon_B.$$

Acceptance already has the sharper $2\epsilon_B$ oracle bound. This proves both assertions on the same event. Off that event, the excess over S_0 is at most one; hence $\mathbb{E}\mathcal{R}_m(S_{\mathrm{safe}}) \leq \mathcal{R}_m(S_0) + \delta$. □

The evaluation bias also satisfies $\kappa_v \leq \sqrt{\log M/(2v)}$. Indeed, each candidate's evaluation mean minus its population mean is conditionally sub-Gaussian given the task, with variance proxy $1/(4v)$. Bounding the expected maximum by $\log M/t + t/(8v)$ and minimizing over $t > 0$ gives this inequality. It affects the absolute certificate, not the oracle excess risk.

B.5 Selection of complete training pipelines

Corollary 2 (Frozen-pipeline selection)**.** *Write $W_\tau(h) = \mathbb{E}_{P_\tau} w_\tau(h, O)$ for a predictor's population welfare. Condition on construction and freeze K measurable procedures A_k mapping permitted task training data and internal randomness to feasible predictors. Let $W_\tau^\dagger \leq 1$ be a common measurable task benchmark, independent of training and dominating the population welfare of all possible outputs. Define $\mathcal{R}(\mathsf{A}_k) = \mathbb{E}[W_\tau^\dagger - W_\tau(\mathsf{A}_k(D_m))]$. Assume rewards lie in $[0,1]$ and evaluation data are independent of training and pipeline randomness conditional on the task. On B independent episodes, select the pipeline with largest mean evaluation welfare $\widehat{U}_k$. With ϵ_B as in Theorem 6, with probability at least $1-\delta$ its risk is at most $\min_k \mathcal{R}(\mathsf{A}_k) + 2\epsilon_B$. Accept it over an included baseline only if $\widehat{U}_{\widehat{k}} - \widehat{U}_0 \geq 2\epsilon_B$; this gate satisfies the two comparisons in* (19)*. A valid absolute certificate is $\mathcal{R}(\mathsf{A}_k) \leq \min\{1, 1 - \widehat{U}_k + \epsilon_B\}$.*

Proof. Each evaluation score is bounded and has expectation equal to its output's expected population welfare. Hoeffding and a union bound give $|\widehat{U}_k - \mathbb{E}W_\tau(\mathsf{A}_k(D_m))| \leq \epsilon_B$ simultaneously. The common benchmark cancels from risk differences. The selection and gate proofs of Theorem 6 therefore apply with scores $-\widehat{U}_k$. The absolute bound follows from $W_\tau^\dagger \leq 1$. □

The benchmark need not be known to the learner. When pipelines can output predictors outside a fixed bank, that bank's empirical maximum may fall below the pipeline benchmark. The extension therefore evaluates pipeline outputs directly; its guarantee concerns selection among frozen pipelines.

B.6 A paired variance-adaptive safety gate

For an included baseline 0, let $D_{bk} = L_b(S_k) - L_b(S_0) = \overline{W}_{b,\widehat{\imath}_0} - \overline{W}_{b,\widehat{\imath}_k} \in [-1,1]$. For pipelines use the difference of their evaluation welfares directly. Write $\mathcal{R}_k$ for member k's meta-risk and $\overline{D}_k = B^{-1}\sum_b D_{bk}$. In either case $\mathbb{E}D_{bk} = \mathcal{R}_k - \mathcal{R}_0$. For $B \geq 2$, define

$$\widehat{s}_k^2 = \frac{1}{B-1}\sum_b (D_{bk} - \overline{D}_k)^2, \qquad r_k = \sqrt{\frac{2\widehat{s}_k^2 \log(4K/\delta)}{B}} + \frac{14\log(4K/\delta)}{3(B-1)}.$$

Then $|\overline{D}_k - (\mathcal{R}_k - \mathcal{R}_0)| \leq r_k$ simultaneously with probability at least $1-\delta$. Indeed, apply the one-sided empirical-Bernstein inequality of Maurer & Pontil (2009, Theorem 4) to $(D_{bk}+1)/2$ and its complement, at failure probability $\delta/(2K)$ per tail. Rescaling multiplies the range term by two and gives the displayed variance term. A union bound proves the assertion. Consequently accepting any data-selected candidate only if $\overline{D}_k + r_k \leq 0$, and otherwise returning the baseline, guarantees no greater meta-risk on this event. This argument establishes baseline safety; an oracle inequality for the variance-adaptive rule requires a separate analysis. Confidence applies to one experiment setting. Simultaneous guarantees across horizons or gates require an additional confidence allocation.

C FINITE-SAMPLE REFINEMENTS AND APPROXIMATION BASELINES

C.1 A TYPE-BASED BRIDGE TO EXACT RISK

On the finite alphabet of Theorem 1, let $\mathcal{P}_m$ be the simplex points with coordinates integer multiples of $1/m$. Define

$$I_m(S) = \min_{\substack{q \in \mathcal{P}_m \\ \min \arg\min_i \sum_j q_j Z_i(S,o_j) \text{ is harmful}}} D_{\mathrm{KL}}(q\|p).$$

If this set is empty, $R_m(S) = 0$. Otherwise let $B_{\min}$ and $B_{\max}$ be the minimum and maximum positive candidate regrets. Then

$$B_{\min}(m+1)^{-s} e^{-mI_m(S)} \le R_m(S) \le B_{\max}(m+1)^s e^{-mI_m(S)}. \tag{20}$$

To see this, the probability of a type q lies between $(m+1)^{-s}e^{-mD_{\mathrm{KL}}(q\|p)}$ and $e^{-mD_{\mathrm{KL}}(q\|p)}$, and there are at most $(m+1)^s$ types. For the lower bound retain a minimizing harmful type; for the upper bound sum over all harmful types. This formulation handles ties and lattice effects without an unstated large-deviation regularity condition.

C.2 THE QUADRATIC CONSTRUCTION

If a scalar comparison D has a moment generating function in a neighborhood of zero, Taylor expansion gives

$$-\log \mathbb{E}e^{-\eta D} = \eta \mathbb{E}D - \frac{\eta^2}{2}\operatorname{Var}(D) + o(\eta^2).$$

For a population-preserving correction, the linear term is fixed. At fixed reference, maximizing this local exponent is therefore a variance minimization problem. More generally, let $\mathcal{T}$ be the bounded operator mapping corrections to weighted centered comparisons, let b_0 be the corresponding base-comparison function, and let $\mathcal{T}^*$ denote its Hilbert-space adjoint. The quadratic objective and normal equation are

$$\|\mathcal{T}v + b_0\|_{L^2}^2 + \rho\|v\|_{\mathscr{H}}^2, \qquad (\mathcal{T}^*\mathcal{T} + \rho I)v_\rho = -\mathcal{T}^* b_0,$$

where v_ρ is the minimizer. The positive regularizer makes this equation uniquely solvable. If $\rho = 0$ and the range of $\mathcal{T}$ is closed, the optimal attainable image is the orthogonal projection of $-b_0$ onto that range. The preimage need not be unique. These statements concern the quadratic approximation, not exact decision risk.

For a single harmful-reference comparison with mean $\mu > 0$, variance σ^2, and $|D - \mu| \le b$, write $\overline{D}_m = m^{-1}\sum_{j=1}^m D_j$ for its i.i.d. sample mean. Bernstein's inequality gives $\mathbb{P}(\overline{D}_m \le 0) \le \exp\{-m\mu^2/[2(\sigma^2 + b\mu/3)]\}$. The exponent favors small variance at fixed μ, b. The cancellation example in Appendix G shows that this preference can increase regret.

C.3 INFORMATION AVAILABLE TO THE LEARNER

A shared loss transfers across task-specific candidate banks: each target learner evaluates it on its own candidates and observations. A predictor selected on a source task need not belong to a future bank. This explains why the loss is a useful transfer object. The comparison with alternative procedures nevertheless depends on their information. Any unrestricted procedure given the same source data, target observations, and candidate bank can reproduce both loss construction and target selection. Thus the formulation supports comparisons with specified baselines, rather than dominance over all procedures with the same information.

D FUNCTIONAL APPROXIMATION AND STATISTICAL BRIDGES

D.1 APPROXIMATION OF FIXED-MULTIPLIER CONSTRUCTION

Product kernels and common norms. For bounded controls with $\mathbb{E}[\zeta_\ell \mid X, \tau] = 0$ and a bounded kernel k_0,

$$v(a, r, O) = \sum_{\ell=1}^{q} f_\ell(a, r, X)\zeta_\ell(O), \tag{21}$$

$$k_{\mathscr{V}}(\xi, \xi') = k_0((a, r, X), (a', r', X')) \sum_{\ell=1}^{q} \zeta_\ell(O)\zeta_\ell(O'). \tag{22}$$

The induced RKHS has this population-preserving representation. We may take $\mathscr{H} = \mathscr{V}$; no projection onto an unknown population nullspace is required. For numerical construction, choose an ordered landmark sequence, orthonormalize retained kernel sections in their common RKHS, and discard exact dependencies. Euclidean regularization in that basis is the inherited norm; independently rescaling each capacity would change the problem. Landmarks may be predetermined or frozen on independent feature-pilot tasks. Construction-adapted landmarks leave the deterministic theorem valid conditionally, but require additional analysis for source generalization.

Anchoring. For any fixed Θ and permissible reference report, subtracting its feature vector adds the same function of (a, X, O) to every candidate's loss. In a difference between reports r, r', these terms cancel exactly, leaving $\zeta^\top \Theta[\phi_p(a, r, X) - \phi_p(a, r', X)]$. This proves empirical and population comparison invariance, including ties. Regularized optimizers additionally depend on the chosen feature norm. For $u = (a, r, X)$, write $\bar{u} = (a, r_{\mathrm{ref}}(a, X), X)$. An explicit anchored kernel is

$$k_0^{\mathrm{anch}}(u, u') = k_0(u, u') - k_0(\bar{u}, u') - k_0(u, \bar{u}') + k_0(\bar{u}, \bar{u}').$$

With ϕ_0 a feature map of k_0, it is positive semidefinite because its feature map is $\phi_0(u) - \phi_0(\bar{u})$. Ordered sections of this kernel span nested subspaces with its fixed induced norm. When different parameter vectors represent the same correction, use the induced quotient norm and discard the dependent directions before counting effective dimension.

Let $\mathscr{V}_1 \subseteq \mathscr{V}_2 \subseteq \cdots \subseteq \mathscr{V}$ be closed subspaces with the inherited Hilbert norm.

Corollary 3 (Finite-space approximation)**.** *Fix the data, base, multipliers, and $\rho > 0$, and write $F(v) = \widehat{F}(v; \lambda)$. Let $v^\star$ minimize F in $\mathscr{V}$, $v_p^\star$ minimize it in $\mathscr{V}_p$, and P_p be orthogonal projection onto $\mathscr{V}_p$; I denotes the identity operator. Suppose $|d_{brj}(0)| \le A_0$ and $\|t_{brj}\| \le A_1$, and set $R_0 = \sqrt{2F(0)/\rho}$, $L_\rho = \rho + m^2 A_1^2 (B_{\mathrm{con}}^{-1} \sum_{b,r} \widehat{B}_{br}) \exp\{m(A_0 + A_1 R_0)\}$. Let $\varepsilon_{\mathrm{opt}} \ge 0$ and $\widehat{v}_p \in \mathscr{V}_p$ satisfy $F(\widehat{v}_p) \le F(v_p^\star) + \varepsilon_{\mathrm{opt}}$. Then*

$$0 \le F(v_p^\star) - F(v^\star) \le \tfrac{1}{2} L_\rho \|(I - P_p)v^\star\|^2, \tag{23}$$

$$\|\widehat{v}_p - v^\star\| \le \sqrt{\{L_\rho \|(I - P_p)v^\star\|^2 + 2\varepsilon_{\mathrm{opt}}\}/\rho}. \tag{24}$$

If the nested union is dense in $\mathscr{V}$, exact minimizers converge in norm to $v^\star$, and their objective values decrease to $F(v^\star)$.

Proof of Corollary 3. All gradients below are in $\mathscr{V}$. Nonnegativity of the data term and feasibility of zero imply $\|v^\star\|, \|v_p^\star\| \le R_0$. The projected competitor $P_p v^\star$ also lies in this ball. Expand a term as

$$\left[n^{-1} \sum_j e^{-d_j(0) - \langle v, t_j \rangle} \right]^m = n^{-m} \sum_{j_1, \ldots, j_m} e^{-\sum_{\ell=1}^m d_{j_\ell}(0) - \langle v, \sum_{\ell=1}^m t_{j_\ell} \rangle}.$$

Each exponential's Hessian is its value times the rank-one operator formed from $\sum_\ell t_{j_\ell}$. On the R_0 ball its value is at most $e^{m(A_0 + A_1 R_0)}$, and this operator has norm at most $m^2 A_1^2$. Thus $\|\nabla^2 F(v)\|_{\mathrm{op}} \le L_\rho$ on that ball. Smoothness along the segment from $v^\star$ to $P_p v^\star$, and $\nabla F(v^\star) = 0$, give

$$F(v_p^\star) - F(v^\star) \le F(P_p v^\star) - F(v^\star) \le (L_\rho/2)\|(I - P_p)v^\star\|^2.$$

Existence in a closed subspace follows either from the finite comparison span argument used in Theorem 2, or from weak lower semicontinuity and coercivity. Strong convexity gives $F(w) - F(v^\star) \geq (\rho/2)\|w - v^\star\|^2$ for every $w \in \mathscr{V}$. Apply this with $w = \widehat{v}_p$ and add its optimization error to the preceding upper bound. Density of the nested union implies $P_p v^\star \to v^\star$, proving convergence. Nested feasibility proves monotonicity of the exact optimal values. □

Optimization certificates. For unconstrained minimization within a finite subspace, completing the square in the strong-convexity inequality yields

$$F(v) - \min_{\mathscr{V}_p} F \leq \|\nabla_{\mathscr{V}_p} F(v)\|^2/(2\rho).$$

If the solver minimizes $\log F$, its gradient must first be rescaled: $\nabla F = F \nabla \log F$. In particular the log of this certificate is $2\log F + 2\log\|\nabla \log F\| - \log(2\rho)$. For a convex constrained set C, a valid alternative is the convex gap $\sup_{z\in C}\langle \nabla F(v), v - z\rangle$. For the box $C = [-b, b]^p$, writing $g_j = \partial F(v)/\partial v_j$, this equals $\sum_j (g_j v_j + b|g_j|)$ in its parameter coordinates. The full-gradient strong-convexity bound remains a valid upper bound on constrained suboptimality—for feasible v, $F(v) - \min_{z\in C} F(z) \leq F(v) - \min_z F(z) \leq \|\nabla F(v)\|^2/(2\rho)$—but it generally does not vanish at an active constrained optimum. We therefore use the box-constrained convex gap as the stationarity diagnostic.

D.2 Estimators of powered comparison moments

For one fixed comparison, write $d_j(v)$ for its value on observation j, n for the comparison-block size, $Y_j(v) = e^{-d_j(v)}$, $M(v) = \mathbb{E}Y_j(v)$, and $\widehat{M}(v) = n^{-1}\sum_j Y_j(v)$. All expectations condition on the task and independently fixed multipliers. For $m = 2$, $\mathbb{E}\widehat{M}(v)^2 = M(v)^2 + \mathrm{Var}(Y_j(v))/n$. For $n \geq m$ the alternative

$$U_{n,m}(v) = \binom{n}{m}^{-1} \sum_{I\subseteq\{1,\dots,n\}:|I|=m} \exp\left(-\sum_{j\in I} d_j(v)\right)$$

therefore removes a correction-dependent source of bias. Independence implies $\mathbb{E}U_{n,m}(v) = M(v)^m$; every summand is an exponential of an affine functional, so convexity and the comparison representer argument are preserved. Elementary symmetric polynomials compute the sum in $O(nm)$ operations: initialize $e_0 = 1$, $e_k = 0$ for $k > 0$, and for each Y_j update $e_k \leftarrow e_k + Y_j e_{k-1}$ in decreasing order of k. The result is $e_m/\binom{n}{m}$. Independent estimated welfare weights can still carry their common maximum bias. Neither unbiased moment estimation nor exact population optimization implies decision-risk optimality, as Theorem 3 shows.

D.3 Welfare-threshold margins

Proposition 1 (Risk continuity)**.** *Fix a horizon m, base loss, and finite-bank ERM rule. For a training sample $D_m = (O_1, \dots, O_m)$, let $\widehat{L}_i(S) = m^{-1}\sum_j S(a_\tau, h_{\tau,i}(X_j), O_j)$. For a regret threshold $t \in [0,1]$ and tolerance $\eta \geq 0$, put $G_{\tau,t} = \{i : B_{\tau,i} \leq t\}$ and*

$$H_{\tau,t}(S, D_m) = \min_{i\notin G_{\tau,t}} \widehat{L}_i(S) - \min_{i\in G_{\tau,t}} \widehat{L}_i(S), \qquad \mathfrak{A}_S(\eta) = \mathbb{E}_\tau \int_0^1 \mathbb{P}_{D_m|\tau}\{|H_{\tau,t}(S, D_m)| \leq \eta\}\, dt.$$

The good set is nonempty; set $H = +\infty$ if its complement is empty. For corrections $v, w \in \mathscr{V}$ satisfying $|v(\xi) - w(\xi)| \leq \kappa\|v - w\|$, one has

$$|\mathcal{R}_m(S_\mathrm{b} + w) - \mathcal{R}_m(S_\mathrm{b} + v)| \leq \mathfrak{A}_{S_\mathrm{b}+v}(2\kappa\|w - v\|).$$

Hence $\mathfrak{A}_{S_\mathrm{b}+v}(\eta) \leq C\eta^\alpha$ for small η, with $\alpha > 0$, gives a local Hölder risk bound with the same exponent.

Proof. Set $e = \kappa\|w - v\|$. Every empirical candidate loss changes by at most e, and so does each nonempty group's minimum. Thus H changes by at most $2e$ when both groups are nonempty. If the original $H > 2e$, both learners choose within the good set; if $H < -2e$, both choose outside it.

Threshold indicators can therefore differ only on $|H| \le 2e$, including all possible tie changes. The identity $\mathbb{E}B = \int_0^1 \mathbb{P}(B > t)\,dt$ for $B \in [0,1]$, followed by Tonelli's theorem and averaging over tasks, proves the bound. The all-good case contributes zero. □

This condition ignores switches among equally acceptable candidates. With the condition imposed at $v^\star$, Corollary 3 implies convergence to the risk of the full-space construction solution, whose risk can exceed the minimum over the correction class. An atom on a harmful tie can invalidate risk continuity despite uniform convergence of losses.

If $S^\star = S_{\mathrm{b}} + v^\star$ satisfies $\mathfrak{A}_{S^\star}(\eta) \le C\eta^\alpha$ for $0 < \eta \le \eta_0$, with $C, \alpha, \eta_0 > 0$, then

$$|\mathcal{R}_m(S_{\mathrm{b}} + \widehat{v}_p) - \mathcal{R}_m(S^\star)| \le C(2\kappa)^\alpha \left[\frac{L_\rho \|(I - P_p)v^\star\|^2 + 2\varepsilon_{\mathrm{opt}}}{\rho}\right]^{\alpha/2}, \tag{25}$$

whenever 2κ times the norm bound in (24) is at most η_0.

Corollary 4 (Inheritance of support safety)**.** *Suppose $S_{\mathrm{b}} + v^\star$ has margin at least $\delta > 0$ in the convex-reference support comparison of Theorem 1, for every harmful candidate on almost every task. If $2\kappa\|v - v^\star\| < \delta$, then $S_{\mathrm{b}} + v$ has zero risk at every positive integer sample size.*

Proof. A candidate-minus-convex-reference comparison changes by at most $2\kappa\|v - v^\star\|$ almost surely. All harmful comparisons retain a strict positive margin, so Theorem 1 applies at every horizon. □

The margin is a uniform population-support assumption, not a certificate inferred from a finite construction sample.

D.4 Task-level construction generalization

Condition on a feature space frozen independently of the construction tasks. Within each task use three conditionally independent blocks: a reference pilot determining all nonnegative multipliers, welfare evaluation determining $\widehat{B}_r$, and n comparison observations. Pilots and the feature space must not be adapted to the candidate correction v in the following bound. Let E denote a complete construction episode (task and its three blocks), let $\|\lambda_r\|_1 \le T$, and define

$$\Phi_E(v) = \sum_r \widehat{B}_r \left[n^{-1} \sum_j e^{-\lambda_r^\top D_r(S_{\mathrm{b}} + v, O_j)}\right]^m, \qquad \mathcal{C}_n(v) = \mathbb{E}_E \Phi_E(v).$$

Assume i.i.d. source tasks from the target task law, at most M candidates, rewards in $[0,1]$, and the same horizon m in this criterion and target risk.

Proposition 2 (Population criterion and finite-dimensional control)**.** *For every fixed v, $\mathcal{R}_m(S_{\mathrm{b}} + v) \le \mathcal{C}_n(v)$. Suppose $\mathscr{V}_p$ has Hilbert dimension d_p, base pairwise differences are bounded in absolute value by D_0, and evaluations obey $|v(\xi)| \le \kappa\|v\|$. Fix a norm radius $R > 0$, covering resolution $a > 0$, failure probability $\delta \in (0,1)$, and N construction episodes. Define*

$$H_R = Me^{mT(D_0 + 2\kappa R)}, \qquad L_R = 2mT\kappa H_R.$$

$$\varepsilon_{\mathrm{con}} = H_R \sqrt{\frac{d_p \log(1 + 2R/a) + \log(2/\delta)}{2N}} + 2L_R a.$$

With probability at least $1 - \delta$ over the N independent episodes,

$$\sup_{v \in \mathscr{V}_p : \|v\| \le R} \left| N^{-1} \sum_b \Phi_{E_b}(v) - \mathcal{C}_n(v) \right| \le \varepsilon_{\mathrm{con}}.$$

An $\varepsilon_{\mathrm{opt}}$-approximate minimizer $\widehat{v}$ of the regularized empirical criterion on this ball consequently satisfies

$$\mathcal{C}_n(\widehat{v}) + \tfrac{\rho}{2}\|\widehat{v}\|^2 \le \inf_{v \in \mathscr{V}_p : \|v\| \le R} [\mathcal{C}_n(v) + \tfrac{\rho}{2}\|v\|^2] + 2\varepsilon_{\mathrm{con}} + \varepsilon_{\mathrm{opt}}.$$

Proof. Condition on a task and its reference pilot. Theorem 1 bounds its risk by $\sum_r B_r M_r(v)^m$, where $M_r(v) = \mathbb{E}e^{-\lambda_r^\top D_r(S_\mathrm{b}+v,O)}$. Independent welfare observations give $\mathbb{E}[\widehat{B}_r \mid \tau] = B_r + b_\tau \geq B_r$, where $b_\tau = \mathbb{E}[\max_i \overline{W}_{\tau,i} \mid \tau] - W_\tau^\star$ is the nonnegative evaluation-maximum bias. Jensen's inequality for the nonnegative power x^m gives $\mathbb{E}[\widehat{M}_r(v)^m \mid \tau, \text{pilot}] \geq M_r(v)^m$. Conditional independence permits multiplication of these expectations. Summing over candidates and averaging over tasks and pilots proves the first claim.

On the radius-R ball each exponent has magnitude at most $T(D_0 + 2\kappa R)$. Since $0 \leq \widehat{B}_r \leq 1$, one has $0 \leq \Phi_E \leq H_R$. The norm of each exponent's derivative is at most $2T\kappa$; differentiation of the powered mean gives $\|\nabla\Phi_E\| \leq 2mT\kappa H_R = L_R$. An a-net of the ball has cardinality at most $(1 + 2R/a)^{d_p}$. Two-sided Hoeffding and a union bound yield the stated square-root term on this fixed net. Both empirical and population criteria are L_R-Lipschitz, so passing to a nearest net point adds at most $2L_R a$. Apply uniform deviation at the empirical approximate minimizer and a population comparator. The identical deterministic regularizer cancels; taking the infimum proves the final inequality. □

The independent units are tasks, not individual observations pooled across tasks. This bound explains a capacity–source-size tradeoff but can be vacuous because of its exponential envelope. The observed criterion alone is not a risk certificate. The statement does not apply to jointly fitted multipliers, construction-adapted features, or mismatched target horizons. Theorem 6 does not require any of these construction assumptions.

A deterministic norm ball. A radius can be derived from the objective itself. For a fit with objective no larger than the feasible zero correction, nonnegativity of the data term gives $\frac{\rho}{2}\|\widehat{v}\|^2 \leq F(0)$, and under the bounded-base and bounded-multiplier assumptions $F(0) \leq Me^{mTD_0}$, so every regularized fit lies in the deterministic ball $R_\mathrm{det} = \sqrt{2Me^{mTD_0}/\rho}$. This radius is deterministic, but the resulting generalization bound is too conservative to give an informative numerical certificate for our fits.

D.5 End-to-end risk decomposition

Combining approximation, source generalization, and validation separates the contributions to excess decision risk. Let $F_\mathrm{pop}(v) = \mathcal{C}_n(v) + \frac{\rho}{2}\|v\|^2$ be the population construction criterion of Proposition 2, with construction and deployment at the same horizon m and comparison-block size n, and let $v^\star$ minimize it over the admissible class $\mathscr{V}$. Define the population approximation error of the pth nested space

$$a_p = \inf_{v \in \mathscr{V}_p} F_\mathrm{pop}(v) - F_\mathrm{pop}(v^\star),$$

the construction-to-risk gap

$$\Delta_\mathrm{con} = \mathcal{R}_m(S_\mathrm{b} + v^\star) - \inf_{v \in \mathscr{V}} \mathcal{R}_m(S_\mathrm{b} + v),$$

and suppose the margin condition of Proposition 1 holds at $S^\star = S_\mathrm{b} + v^\star$, for constants $C, \alpha, \eta_0 > 0$: $\mathfrak{A}_{S^\star}(\eta) \leq C\eta^\alpha$ for $0 < \eta \leq \eta_0$.

Theorem 7 (End-to-end decomposition)**.** *Let the library $\{S_\mathrm{b} + \widehat{v}_p\}_p$ be frozen, where each $\widehat{v}_p$ is an $\varepsilon_{\mathrm{opt},p}$-approximate minimizer of the empirical criterion on $\mathscr{V}_p$, and let the uniform construction deviation of Proposition 2 at capacity p be at most $\varepsilon_{\mathrm{con},p}$ with confidence $1 - \delta_p$ on a deterministic ball containing both $\widehat{v}_p$ and a population minimizer in $\mathscr{V}_p$. Run the selection of Theorem 6 on B independent episodes at horizon m with failure probability δ_sel and library size K; set $\epsilon_B = \sqrt{\log(2K/\delta_\mathrm{sel})/(2B)}$. Then with probability at least $1 - \delta_\mathrm{sel} - \sum_p \delta_p$, simultaneously,*

$$F_\mathrm{pop}(\widehat{v}_p) - F_\mathrm{pop}(v^\star) \leq a_p + 2\varepsilon_{\mathrm{con},p} + \varepsilon_{\mathrm{opt},p}, \qquad \|\widehat{v}_p - v^\star\| \leq \sqrt{\frac{2a_p + 4\varepsilon_{\mathrm{con},p} + 2\varepsilon_{\mathrm{opt},p}}{\rho}},$$

for every p, and the selected loss $\widehat{S}$ satisfies

$$\mathcal{R}_m(\widehat{S}) - \inf_{v \in \mathscr{V}} \mathcal{R}_m(S_\mathrm{b} + v) \leq \Delta_\mathrm{con} + \min_p C(2\kappa)^\alpha \left(\frac{2a_p + 4\varepsilon_{\mathrm{con},p} + 2\varepsilon_{\mathrm{opt},p}}{\rho}\right)^{\alpha/2} + 2\epsilon_B, \quad (26)$$

provided each resulting $2\kappa\|\widehat{v}_p - v^\star\|$ is at most η_0.

Proof. The first display combines uniform deviation (applied at $\widehat{v}_p$ and at the population comparator in $\mathscr{V}_p$) with the optimization error; ρ-strong convexity converts the objective bound into the norm bound. For the risk bound, Theorem 6 gives $\mathcal{R}_m(\widehat{S}) \le \min_p \mathcal{R}_m(S_{\mathrm{b}} + \widehat{v}_p) + 2\epsilon_B$ on the selection event. For each p, Proposition 1 with the margin condition at $S^\star$ bounds $|\mathcal{R}_m(S_{\mathrm{b}} + \widehat{v}_p) - \mathcal{R}_m(S^\star)|$ by $C(2\kappa\|\widehat{v}_p - v^\star\|)^\alpha$, and the definition of Δ_{con} gives $\mathcal{R}_m(S^\star) = \inf_v \mathcal{R}_m(S_{\mathrm{b}} + v) + \Delta_{\mathrm{con}}$. A union bound over capacities combines the construction events with the selection event. □

The terms a_p, $\varepsilon_{\mathrm{con},p}$, $\varepsilon_{\mathrm{opt},p}$, and ϵ_B describe approximation, source estimation, optimization, and selection error. For a fixed admissible class and population criterion, Δ_{con} persists even when these errors vanish. Proposition 7 gives $\Delta_{\mathrm{con}} = 2/75$ for constant corrections in the balanced sensor model at horizon one. The experiments measure both the construction objective and decision risk to distinguish this gap from the other terms.

D.6 Characterizing common outcome corrections

Proposition 3. *Let $P_\Lambda = I - \mathbf{1}\mathbf{1}^\top/M$ and let $X_{\mathrm{act}} \in \mathbb{R}^{M\times d}$. A vector $v \in \mathbb{R}^M$ has the form $-X_{\mathrm{act}}b + c\mathbf{1}$ for some $b \in \mathbb{R}^d$, $c \in \mathbb{R}$ if and only if $P_\Lambda v \in \mathrm{range}(P_\Lambda X_{\mathrm{act}})$. It fails this condition if and only if there is a vector q with $\mathbf{1}^\top q = 0$, $X_{\mathrm{act}}^\top q = 0$, and $q^\top v \neq 0$.*

Proof. Projection removes the constant, proving necessity. Conversely, a solution of $P_\Lambda v = -P_\Lambda X_{\mathrm{act}} b$ leaves $v + X_{\mathrm{act}}b$ in $\ker P_\Lambda = \mathrm{span}\{\mathbf{1}\}$. Orthogonal decomposition within Λ proves the witness equivalence; the nonzero residual itself is a witness. For the four binary actions ordered as $(0,0),(0,1),(1,0),(1,1)$, $q = (1,-1,-1,1)$ annihilates constants and both action coordinates, while $q^\top(0,0,0,\zeta) = \zeta$. □

The residual norm on fresh observations quantifies departure from a common outcome correction. Its effect on decision risk is evaluated independently.

E Experimental protocols and additional results

This appendix specifies the sensor capacity, cancellation, horizon, and contextual nuisance studies reported in Section 6. Source files in `content-E/experiments/` comprise shared routines (`sensor_common.py`, `sensor_fits.py`), study scripts (`experiment_a.py`, `experiment_b.py`, `experiment_c.py`, `experiment_contextual.py`), and analytic checks (`verify_anchors.py`). Raw per-repetition records and aggregate tables are written to `content-E/results/`; `experiment_manifest.json` there lists every configuration and its status. We used Python with NumPy 2.0.2, SciPy 1.13.1, pandas 2.3.3, and Matplotlib 3.9.4, with one numerical-library thread. `content-E/README.md` gives the commands, seeds, and output descriptions. Sensor records include `experiment_a_primary.csv` (risks), `experiment_a_diagnostics.csv` (objectives, gaps, and norms), and `experiment_a_selection.csv` (selected members and gate decisions), with analogous records for the other studies. Supplementary transfer experiments have separate protocols in Appendix I.

E.1 A descriptor-dependent support separation

Proposition 4 (A functional advantage over constant corrections)**.** *Let A be uniform on $\mathcal{A}_{.2} = \{a \in [-1,1]^2 : |\sin(2\pi a_1) + .6a_2| \ge .2\}$, and put $e(a) = \mathrm{sign}(\sin(2\pi a_1) + .6a_2)$. Independently draw θ uniformly from $\{-1,1\}$. Let independent V, U have laws $\mathbb{P}(V = .5) = .8$, $\mathbb{P}(V = 3) = .2$, $\mathbb{P}(U = -1) = .2$, and $\mathbb{P}(U = 1.5) = .8$. Observe $(Y_1, Y_2) = (\theta V, \theta U)$ when $e(A) = 1$ and swap the sensors otherwise. Define $Y = (Y_1 + Y_2)/2$, $Z = Y_1 - Y_2$, $w(a_{\mathrm{act}}, O) = 1/2 + a_{\mathrm{act}}Y/6$ for $a_{\mathrm{act}} \in \{-1,1\}$, and $S_t = -a_{\mathrm{act}}(Y + t(A)Z)$. Every measurable $t : \mathcal{A}_{.2} \to [-.5, .5]$ preserves population comparisons. Conditional variance is uniquely minimized by $t = 0$ almost surely, whereas $\inf_{t\ \mathrm{constant}} \mathcal{R}_1(S_t) = 2/75$. Every t satisfying $e(a)t(a) > 1/6$ on the descriptor domain is safe at every horizon. Finite expansions of any universal continuous kernel on this compact domain include such functions.*

Proof. Both sensors have mean and variance one, so $\mathbb{E}[Z \mid A, \theta] = 0$. Conditional on the descriptor and task sign, the signed corrected observation has law

$$Q_s = (1/2 + s)V + (1/2 - s)U, \qquad s = e(A)t(A), \qquad \mathbb{E}Q_s = 1, \quad \mathrm{Var}(Q_s) = 1/2 + 2s^2.$$

This proves preservation and the unique almost-sure minimum of the integrated conditional variance. The action welfare gap is $1/3$. Symmetry of the descriptor domain and oddness of its defining function give equally likely orientations. For constant t, at least one of $t, -t$ is strictly below $1/6$. On that orientation, the event $(V, U) = (.5, -1)$, of probability $4/25$, gives $Q_s = -1/4 + 3s/2 < 0$. Thus risk is at least $(1/2)(4/25)(1/3) = 2/75$. For $t \in (1/6, 1/4)$ one orientation is strictly safe and the other's only negative support atom has mass $4/25$, attaining the bound without ties. Because both coefficients in Q_s are nonnegative when $|s| \leq .5$, its minimum support value is $-1/4 + 3s/2$. A positive minimum implies correct empirical selection at every sample size.

Finally $t^\circ(a) = .4e(a)$ is continuous on the compact domain, whose opposite-sign components are separated. Universality means that the kernel RKHS is uniformly dense in its continuous functions. Finite kernel sections are norm-dense in that RKHS, and bounded evaluation turns norm approximation into uniform approximation. Combining the two approximations gives a finite expansion within .05 of t°. It satisfies $e(a)t(a) > .35$ and $|t(a)| < .45$, hence is admissible and safe. □

The proposition establishes a functional advantage in the continuous-kernel family. The experiment below uses a finite histogram RKHS, whose fitted corrections may lack a positive support margin on parts of the domain. Proposition 8 in Appendix F analyses the population construction objective itself on this model.

E.2 Nested features and shared norms

Partition $[-1, 1]^2$ into rectangular grids of sizes $(1, 1), (4, 2), (8, 4), (16, 8)$, giving $p = 1, 8, 32, 128$ cells; a coarser auxiliary grid $(4, 4)$ with $p = 16$ is used only where noted. Assign each boundary point to the cell on its upper/right side, except at the outer boundary. Each partition refines the previous one. A scalar function with cell values c_j has norm $\|t\|^2 = p^{-1} \sum_j c_j^2$, the squared L^2 norm under normalized Lebesgue measure on the full square. This defines one finite RKHS at resolution 128 with kernel $128\mathbf{1}\{a, a' \text{ in the same finest cell}\}$. Coarser spaces inherit exactly this norm. Copying a parent value to its children preserves both the function and norm; orthonormal coordinates are $c_j/\sqrt{p}$. For vector functions sum these squared norms across coordinates. These spaces are predetermined: no feature-pilot observations are needed, and neither the sensor orientation nor the nuisance generator determines the partition. The experiments compare these finite capacities.

For the sensor loss, the action contrast $-a_{\mathrm{act}}$ has no report-independent component. We optimize the p cell values directly, with the box constraint $|c_j| \leq .5$. This enforces the global range of t without post-hoc clipping of an unconstrained fit. The box spaces are nested under parent copying, and the objective uses the inherited norm. Exact constrained objective minima are therefore nonincreasing in p for common data, multiplier, and ρ. The box-constrained convex gap in Appendix D provides an optimization certificate for these constrained fits. The contextual fits below are unconstrained and use the Hilbert-gradient certificate.

E.3 Sensor construction, baselines, and selection

Index the 100 independent repetitions by $\mathrm{rep} = 0, \ldots, 99$; their seeds are $530000 + \mathrm{rep}$. Rejection sampling generates uniform descriptors on $\mathcal{A}_{.2}$. Each repetition draws 2048 construction tasks and uses nested prefixes $N \in \{32, 128, 512, 2048\}$. Each task has 64 reference-pilot, 64 independent welfare, and 32 independent comparison observations. The learner sees descriptors and sensor observations, not θ, e, or true welfares. The sign estimate $\widehat{s}_b$ is the sign of the pilot mean of Y (negative on a tie); welfare weights are the empirical gaps of the two actions from the welfare block, with one weight zero. The welfare block also determines which action has the positive weight in the entropic objective. For every candidate the multiplier on its sole competitor is fixed at four. The primary entropic criterion is exactly (11) with construction horizon one and predeclared $\rho \in \{.001, .01\}$; the displayed setting is $\rho = .01$. The pilot is unnecessary for this fixed multiplier, but supplies independent sign estimates to the endpoint, constant, and empirical-risk baselines.

The box problem separates by cell. Forty-eight derivative-bisection steps solve each coefficient, including empty cells, whose regularizer selects zero; both endpoints are evaluated exactly. Bisection has no initialization dependence. The reported certificate is the box gap $\sum_j (g_j c_j + .5|g_j|)$, with $g_j = \partial \widehat{F}/\partial c_j$, in cell-value coordinates; across all 3,200 primary entropic fits (100 repetitions $\times$ 4 source sizes $\times$ 4 capacities $\times$ 2 regularizers) the maximum box gap is 1.04×10^{-10}.

The variance baseline minimizes within-task empirical variance over the same cell functions, using an analytic cellwise least-squares solution with the declared box constraint. Empty cells use zero. The sensor-endpoint baseline assigns each finest cell the sensor having fewer source sign errors relative to independent pilot signs; equal scores use the direct average $t = 0$. The fitted constant minimizes unweighted comparison sign errors over the ordered grid $(0, .2, -.2, .5, -.5)$, using pilot signs, and is therefore not an oracle. The same-class empirical-risk baseline uses $\widehat{B}_b = |\overline{Y}_b^{\text{welfare}}|/3$, where $\overline{Y}_b^{\text{welfare}}$ is the welfare-block mean, and minimizes, within each cell, the welfare-weighted empirical horizon-one sign error $\sum_{b \in \text{cell}} \widehat{B}_b \sum_j [\mathbf{1}\{\widehat{s}_b(Y_{bj} + tZ_{bj}) < 0\} + \frac{1}{2}\mathbf{1}\{\cdot = 0\}]$; since each observation's contribution flips only at $t = -Y_{bj}/Z_{bj}$ when $Z_{bj} \neq 0$, the score is piecewise constant and we evaluate it exactly on all breakpoints and interval midpoints, resolving ties toward smaller $|t|$ and then positive values. This baseline uses the same information and correction capacity as the entropic method, isolating the effect of the construction objective.

For independent copies $Q_s^{(1)}, \ldots, Q_s^{(m)}$ of Q_s, define $r_m(s) = \frac{1}{3}[\mathbb{P}(\sum_j Q_s^{(j)} < 0) + \frac{1}{2}\mathbb{P}(\sum_j Q_s^{(j)} = 0)]$; the half-weight averages over the two task signs under the fixed action tie rule. Two population oracles, computed from the known generator and excluded from the selection library, delimit the construction. The population entropic oracle minimizes per cell $\mu_c[\pi_c M_4(t) + (1-\pi_c)M_4(-t)]/3 + (\rho/2p)t^2$, where μ_c and π_c are the cell mass and orientation purity, numerically integrated from 2^{20} fixed-seed descriptors. The population risk-optimal oracle minimizes per cell the exact mixture risk $\pi_c r_m(t) + (1 - \pi_c) r_m(-t)$, using the piecewise-constant breakpoint structure of r_m (Section E.4); it is evaluated at each deployment horizon. The analytic safe oracle is $.4e(A)$.

The frozen library at each N has 19 members in declared order: direct loss, the fitted constant, sensor-endpoint selection, four variance fits, four empirical-risk fits, and eight entropic fits ($\rho \times p$). Every N has its own frozen library. Selection uses fresh episodes with $B = 2048$, target training horizon $m = 1$, and 128 independent evaluation observations; by (15) only the selected candidates' evaluated welfares enter the rule. Ties use library order, with direct first. We report the full, entropic-only, and non-entropic selectors, the predeclared representatives (entropic $\rho = .01$, $p = 128$; non-entropic sensor-endpoint), selection frequencies, and the Hoeffding gate of Theorem 6 at $\delta = .05$. Construction at horizon one is a generation heuristic for other horizons; Section 6 studies horizon matching directly.

E.4 Sensor risk enumeration and integration

Conditional on a cell coefficient and orientation, the signed corrected observation Q_s takes four values; multinomial types over m draws are enumerated exactly. Because twice the sensor values are the integers $(1, 1, 6, 6)$ and $(-2, 3, -2, 3)$ with probability numerators $(4, 16, 1, 4)/25$, the sign of each type's sum is the sign of an integer linear form in the stored coefficient (interpreted as an exact binary rational), and values within 10^{-12} of zero are rechecked with exact rational arithmetic. Deterministic action selection uses a strict-positive rule: action $+1$ only for a strictly positive corrected mean, and a zero sum is charged half to each action, matching the independent uniform task sign. Meta-risk integrates these conditional risks over 16,384 fresh descriptors per repetition (shared across methods), and we report 95% normal-approximation intervals (mean ± 1.96 standard errors) across the 100 repetitions, which include construction, selection, and integration variation. Anchor checks in `verify_anchors.py` validate the enumeration against brute-force Monte Carlo and reproduce the analytic risk values $4/75$ and $2/75$, the purity threshold 0.9821, and the safe region.

E.5 Additional sensor results

Table 2 gives the risk–capacity sweep behind the left panel of Figure 1; Table 4 gives the gap attribution; Table 5 gives the selector comparison and selection frequencies. The estimation term at $p = 128$ is negative (-2.21×10^{-3} at $N = 512$), meaning that the fitted corrections have lower mean

Table 2: Sensor study: target risk $\times 10^3$ at horizon $m = 1$ by capacity and construction size. Oracle rows use the known generator. Intervals are 95% over 100 repetitions.

Method	p	$N = 32$	$N = 128$	$N = 512$	$N = 2048$
Variance minimization	1	53.333 ± 0.000	53.333 ± 0.000	53.333 ± 0.000	53.333 ± 0.000
	8	52.639 ± 0.403	53.333 ± 0.000	53.333 ± 0.000	53.333 ± 0.000
	32	52.034 ± 0.301	52.994 ± 0.139	53.333 ± 0.000	53.333 ± 0.000
	128	52.731 ± 0.113	52.096 ± 0.151	52.973 ± 0.068	53.333 ± 0.000
Same-class empirical risk	1	26.679 ± 0.045	26.684 ± 0.045	26.711 ± 0.044	26.663 ± 0.045
	8	5.136 ± 0.580	2.894 ± 0.093	2.664 ± 0.018	2.662 ± 0.018
	32	21.924 ± 0.581	4.953 ± 0.262	2.951 ± 0.069	2.687 ± 0.019
	128	41.186 ± 0.175	19.311 ± 0.298	2.432 ± 0.111	1.152 ± 0.014
Entropic ($\rho = .01$)	1	53.333 ± 0.000	53.333 ± 0.000	53.333 ± 0.000	53.333 ± 0.000
	8	10.803 ± 1.018	19.204 ± 0.773	23.081 ± 0.040	23.081 ± 0.040
	32	22.419 ± 0.585	8.879 ± 0.373	13.069 ± 0.349	15.037 ± 0.391
	128	41.187 ± 0.175	19.369 ± 0.301	3.083 ± 0.114	3.371 ± 0.076
Population entropic (oracle)	1	53.333	53.333	53.333	53.333
	8	23.038	23.038	23.038	23.038
	32	11.109	11.109	11.109	11.109
	128	5.288	5.288	5.288	5.288
Population risk-optimal (oracle)	1	26.616	26.616	26.616	26.616
	8	2.670	2.670	2.670	2.670
	32	2.670	2.670	2.670	2.670
	128	1.083	1.083	1.083	1.083

Table 3: Sensor study at $p = 128$: the main-text block (top; horizon $m = 1$ across source sizes N) and the horizon block (bottom; $N = 512$ across horizons m). Intervals are 95% over 100 repetitions.

	$m = 1$			
Method	$N = 32$	$N = 128$	$N = 512$	$N = 2048$
Direct loss	53.333 ± 0.000	53.333 ± 0.000	53.333 ± 0.000	53.333 ± 0.000
Fitted constant	26.683 ± 0.045	26.683 ± 0.045	26.695 ± 0.044	26.652 ± 0.045
Variance minimization	52.731 ± 0.113	52.096 ± 0.151	52.973 ± 0.068	53.333 ± 0.000
Same-class empirical risk	41.186 ± 0.175	19.311 ± 0.298	2.432 ± 0.111	1.152 ± 0.014
Entropic	41.187 ± 0.175	19.369 ± 0.301	3.083 ± 0.114	3.371 ± 0.076
Sensor-endpoint selection	41.164 ± 0.176	19.264 ± 0.299	2.538 ± 0.110	1.389 ± 0.019
Population entropic (oracle)	5.288	5.288	5.288	5.288
Population risk-optimal (oracle)	1.083	1.083	1.083	1.083
Safe oracle $.4e(A)$	0	0	0	0

	$N = 512$		
Method	$m = 1$	$m = 5$	$m = 20$
Direct loss	53.333 ± 0.000	0.406 ± 0.000	0.000 ± 0.000
Fitted constant	26.695 ± 0.044	0.840 ± 0.001	0.000 ± 0.000
Variance minimization	52.973 ± 0.068	0.414 ± 0.007	0.000 ± 0.000
Same-class empirical risk	2.432 ± 0.111	0.065 ± 0.003	0.000 ± 0.000
Entropic	3.083 ± 0.114	0.042 ± 0.002	0.000 ± 0.000
Sensor-endpoint selection	2.538 ± 0.110	0.223 ± 0.007	0.000 ± 0.000
Population entropic (oracle)	5.288	0.019	0.000
Population risk-optimal (oracle)	1.083	0.019	0.000
Safe oracle $.4e(A)$	0	0	0

risk than the population entropic optimum in this setting. The sign of this estimation contribution can therefore differ from the sign of excess construction objective.

Table 4: Gap attribution for the entropic construction ($\rho = .01$) at $N = 512$, horizon $m = 1$, in risk units $\times 10^3$. The three rows below the rule partition the sampled risk: estimation (sampled minus population entropic), construction objective (population entropic minus population risk-optimal), and residual capacity (population risk-optimal, since the analytic safe oracle has zero risk).

	capacity p			
gap	1	8	32	128
sampled entropic	53.333	23.081	13.069	3.083
population entropic	53.333	23.038	11.109	5.288
population risk-optimal	26.616	2.670	2.670	1.083
estimation	$+0.000$	$+0.044$	$+1.961$	-2.205
construction objective	$+26.718$	$+20.367$	$+8.438$	$+4.205$
residual capacity	$+26.616$	$+2.670$	$+2.670$	$+1.083$

Table 5: Selected risk $\times 10^3$ at horizon $m = 1$ for the frozen-library selectors. Selection frequencies at $N = 512$: empirical-risk members $p = 8, 128, 32$ receive .42, .30, .01, sensor-endpoint .26, and one entropic member .01. The safety gate accepts in no repetition; see the power calculation in Appendix I.

Selector	$N = 32$	$N = 128$	$N = 512$	$N = 2048$
Full library	5.136 ± 0.580	2.881 ± 0.070	2.433 ± 0.086	1.217 ± 0.039
Entropic only	10.797 ± 1.022	8.879 ± 0.373	3.083 ± 0.114	3.371 ± 0.076
Non-entropic only	5.136 ± 0.580	2.881 ± 0.070	2.423 ± 0.087	1.217 ± 0.039
Safety-gated	53.333 ± 0.000	53.333 ± 0.000	53.333 ± 0.000	53.333 ± 0.000

E.6 Direct episodic welfare fitting

This additional study uses the same descriptor and sensor laws but has its own data and results. For repetition $r = 0, \ldots, 99$, seed $1200000+r$ generates 8192 source tasks with one training observation and 32 independent welfare observations each. Nested prefixes have $N \in \{128, 512, 2048, 8192\}$ tasks. The four fixed histograms have $p \in \{1, 8, 32, 128\}$ cells. Neither fitting routine receives latent signs or orientations. The episodic fit maximizes the sum of the actual evaluated welfares of the actions selected from the training block. In each cell, all rational comparison roots, box endpoints, zero, and one midpoint per open interval are evaluated. Actions choose -1 on a tie. Equal objective values prefer the evaluated coefficient with smallest absolute value and then the positive one; empty cells receive zero.

All training sensor sums and evaluation utility sums are integer after scaling by four (and control sums by two). Exact integer arithmetic compares a root or interval representative, represented as a rational fraction, with every training hyperplane. The objective is compared using integer utility sums, so neither the optimization score nor its tie rule uses a floating-point tolerance. This finite enumeration attains the empirical minimum in the cell family. It is an implementation of direct episodic fitting; the pilot-sign baseline in the preceding study has a different objective and is retained under its original protocol.

The entropic comparator uses the same observations, exact empirical welfare gaps from the evaluation block, multiplier four, horizon one, and $\rho = .01$ in the inherited histogram norm. The candidate with lower empirical evaluation welfare is the harmful candidate for this construction objective; no independent sign pilot is necessary. Bisection uses the existing convex sensor solver. Thus acquisition is $33N$ observations for either method and the difference is the construction objective.

Independent seed $1300000+r$ generates 65,536 target descriptors for each repetition. For each fitted cell value and orientation, all four sensor outcomes are enumerated, averaging over the independent equiprobable task sign and its deterministic action tie rule. Rational fitting coefficients retain their exact value in risk enumeration; entropic coefficients are interpreted as their stored binary rationals. The common descriptor sample integrates these conditional risks. The reported 95% intervals use Student's t quantile with 99 degrees of freedom over independent repetitions, including both fitting

Table 6: Direct episodic fitting: target risk $\times 10^3$ at $m = 1$. Intervals are 95% Student intervals over 100 independent repetitions. Both methods receive one training and 32 independent welfare observations per source task.

Capacity	Method	$N = 128$	$N = 512$	$N = 2048$	$N = 8192$
1	Episodic welfare	27.13 ± 0.34	26.67 ± 0.02	26.67 ± 0.02	26.67 ± 0.02
1	Entropic	53.33 ± 0.00	53.33 ± 0.00	53.33 ± 0.00	53.33 ± 0.00
8	Episodic welfare	13.95 ± 1.17	3.53 ± 0.22	2.89 ± 0.03	2.67 ± 0.01
8	Entropic	12.11 ± 1.13	15.68 ± 0.90	22.69 ± 0.26	23.04 ± 0.02
32	Episodic welfare	35.40 ± 0.67	13.08 ± 0.54	3.57 ± 0.12	2.80 ± 0.02
32	Entropic	30.97 ± 0.74	10.25 ± 0.47	11.43 ± 0.32	15.79 ± 0.42
128	Episodic welfare	47.28 ± 0.25	33.14 ± 0.35	10.02 ± 0.23	1.56 ± 0.04
128	Entropic	45.37 ± 0.29	27.90 ± 0.40	5.51 ± 0.20	2.95 ± 0.08

and integration variability. They are pointwise intervals, not simultaneous comparisons over all configurations.

The implementation, `experiment_episodic.py`, retains repetition-level risks, summary statistics, a table, and a protocol manifest in `content-E/results/`; the accompanying README lists the output files. At $p = 128$, entropic fitting has lower risk at $N = 128, 512, 2048$, while direct episodic fitting has lower risk at $N = 8192$. At $p = 8$, the empirical entropic risk rises toward its population level as N increases, whereas episodic fitting approaches the cellwise risk optimum. These finite-sample patterns are consistent with the distinct optimization targets; the asymptotic conclusion follows from the theory, not from these four source sizes.

E.7 Cancellation factorial protocol

The factorial study of Section 6.2 uses the cancellation model of Appendix G with $d = 1/2$, $\varepsilon \in \{.05, .1, .25, .4\}$, multiplier budgets $T \in \{.5, 2, 4\}$, and horizons $m \in \{1, 2, 5, 10, 20\}$. All four methods select t on a shared deterministic 501-point grid on $[0, 1]$. The exponential inner problems reduce to scalar convex minimization: equal multipliers $\lambda_1 = \lambda_2$ minimize the two-dimensional objective because $\cosh$ is even and increasing in the absolute value of its argument. The quadratic inner problems use closed-form simplex projection, cross-checked against `scipy.optimize` (agreement below 10^{-8}). Exact risks use the binomial enumeration of Appendix G; no Monte Carlo enters Stage 1. Stage 2 fits the same four methods from N source tasks with n comparison observations each (deterministic welfare is known; the single-reference methods use the first candidate by library order), seeds $770000 + \mathrm{rep}$, 100 repetitions. Table 8 gives the estimation stage; the selected corrections agree with the population optima at the smallest tested source size. The separation observed here therefore persists with population comparison information.

E.8 Horizon and moment-bias protocol

The horizon study of Section 6 uses 20 repetitions (seeds $860000 + \mathrm{rep}$), capacities $p \in \{1, 128\}$, deployment horizons $m \in \{1, 8, 32\}$, and comparison block sizes and task counts $(N, n) \in \{(512, 32), (128, 128), (32, 512)\}$ holding acquisition $Nn = 16{,}384$ fixed, plus $(128, 32)$ and $(128, 512)$ holding the task count fixed. All fitted variants share a deterministic 201-point grid on $[-.5, .5]$ so that search resolution is controlled and matched across criteria. The U-statistic $\binom{n}{m}^{-1} \sum_{|I|=m} \exp(-\sum_{j \in I} d_j)$ is evaluated by log-domain elementary-symmetric dynamic programming in $O(nm)$ per task and grid point; the powered V-statistic is $\widehat{M}^m$; horizon-one construction uses power one; the exact-law variant is the population entropic oracle at the target horizon. Criterion bias is probed at fixed grid points conditional on the majority orientation with pilot-correct signs, separating the powered moment bias from pilot contamination. Table 9 gives the risk comparison, with 95% normal-approximation intervals over repetitions; the bias probes and criterion–risk rank statistics are in `experiment_c_bias.csv` and `experiment_c_risks.csv`. At $p = 1$, the population criterion selects the zero constant (Proposition 7); finite-sample fits can differ. The construction objective is anticorrelated with one-observation risk across constant corrections (mean

Table 7: Factorial diagnostic, Stage 1 (exact population): selected loss parameter $t^\star$ and exact risks R_2, R_{10} at $T = 2$.

	$\varepsilon = 0.05$			$\varepsilon = 0.10$		
Method	t^*	R_2	R_{10}	t^*	R_2	R_{10}
Single-quadratic (SQ)	0.50	0.1250	0.0002403	0.50	0.1250	0.0002403
Full-quadratic (FQ)	0.00	0	0	0.00	0	0
Single-entropy (SE)	0.50	0.1250	0.0002403	0.50	0.1250	0.0002403
Full-entropy (FE)	0.00	0	0	0.00	0	0

	$\varepsilon = 0.25$			$\varepsilon = 0.40$		
Method	t^*	R_2	R_{10}	t^*	R_2	R_{10}
Single-quadratic (SQ)	0.50	0.1250	0.0002403	0.50	0.1250	0.003044
Full-quadratic (FQ)	0.00	0	0	0.00	0	0
Single-entropy (SE)	0.50	0.1250	0.0002403	0.50	0.1250	0.003044
Full-entropy (FE)	0.00	0	0	0.00	0	0

Table 8: Factorial diagnostic, estimation stage: mean exact risk at $m = 10$ ($\times 10^3$) with 95% intervals over 100 repetitions, $\varepsilon = .1$, $T = 2$.

		observations per task n		
Method	N	$n = 16$	$n = 64$	$n = 256$
Single-quadratic (SQ)	32	0.24 ± 0.00	0.24 ± 0.00	0.24 ± 0.00
	128	0.24 ± 0.00	0.24 ± 0.00	0.24 ± 0.00
	512	0.24 ± 0.00	0.24 ± 0.00	0.24 ± 0.00
Full-quadratic (FQ)	32	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	128	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	512	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Single-entropy (SE)	32	0.26 ± 0.06	0.24 ± 0.00	0.24 ± 0.00
	128	0.24 ± 0.00	0.24 ± 0.00	0.24 ± 0.00
	512	0.24 ± 0.00	0.24 ± 0.00	0.24 ± 0.00
Full-entropy (FE)	32	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	128	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	512	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

Spearman -0.54 at $m = 1$, with 72% of candidate pairs rank-reversed), while at $m \geq 8$ the correlation approaches one: the objective and risk rankings agree closely at these longer horizons.

E.9 Unknown contextual nuisance and action-dependent capacity

This controlled ablation uses the four actions $x \in \{0,1\}^2$, two context coordinates, and one centered control.

In repetition r (seed $630000 + r$), draw once $\omega_j \sim \mathrm{Unif}([-3,3]^2)$, $b_j \sim \mathrm{Unif}[0, 2\pi]$, and $B_j \sim \mathrm{Unif}([-1,1]^2)$ independently, for $j = 1, \ldots, 12$. Every task has a new hidden $\mu_\tau \sim \mathrm{Unif}([-.12, .12]^2)$. For independent $X \sim \mathrm{Unif}([-1,1]^2)$, Rademacher Z, and $\epsilon \sim \mathrm{Unif}([-.02, .02]^2)$, observe

$$C = \mu_\tau + .08X + B(X)Z + \epsilon, \qquad B(X) = 12^{-1} \sum_{j=1}^{12} B_j \cos(\omega_j^\top X + b_j).$$

The learner is not given the generator frequencies, phases, loadings, or task means. The known coordinate bound $|C_d| \leq 1.22$ guarantees that $w(x, C) = 1/2 - x^\top C/4.88 \in [0,1]$ for binary actions. Reported regret is normalized by 4.88, using the exact population costs $x^\top \mu_\tau$ of the constant-action bank. All candidate predictors are fixed reports $r_x = .12(1 - 2x)$, so minimizing predicted cost yields the candidate action x. The direct base is $x^\top C$; its positive scaling relative to negative welfare is fixed across the correction family.

Family A uses $v = -Z(x_1 f_1(X) + x_2 f_2(X))$, with $2p$ cell coefficients. Family B adds $-Z x_1 x_2 f_3(X)$, for $3p$ coefficients and an action interaction. Both are anchored at $(0,0)$ and preserve each candidate's mean. Family A is a shared outcome correction; family B's fresh-observation residual outside that space is $|f_3(X)|/2$. Writing c_{jd} for the value of f_d in cell j, both inherit $\|f\|^2 = p^{-1} \sum_{j,d} c_{jd}^2$.

Reported 95% intervals use Student-t quantiles over the 20 independent repetitions. Each repetition uses 128 construction tasks with $(64, 64, 32)$ reference, welfare, and comparison observations. The pilot chooses a reference action minimizing pilot mean cost. For each other candidate set the multiplier on that reference to four; for the reference itself use zero. Welfare weights come from the separate welfare block. Construction horizon is one, with fixed $\rho = .01$; no target-based regularization tuning is performed. Both families are fit entropically and with the quadratic exponential approximation $e^{-d} \approx 1 - d + d^2/2$ on identical comparisons. This quadratic criterion is convex but is not a tail probability or risk bound. L-BFGS uses at most 1200 iterations, gradient tolerance 10^{-10}, relative objective tolerance 10^{-14}, and 40 line-search steps. Each larger space starts at the embedded smaller solution. The gradient certificate in cell coordinates is $p\|\nabla_c F\|_2^2/(2\rho)$, because orthonormal coordinates are $c/\sqrt{p}$.

Outcome denoising directly regresses C on Z times the same cell indicators, with ridge coefficient .01 in the inherited norm, on pooled construction data. The one-/two-strength baselines scale its richest fitted map and optimize only those strengths using the entropic criterion. Squared loss and SPO+ use the same four reports and observations. The library has 26 members: direct, squared, SPO+, regression-denoised, two strength scalings, and the entropic and quadratic fits of both families at $p \in \{1, 8, 16, 32, 128\}$. All members are frozen before the 512 independent selection episodes with $(32, 128)$ training/evaluation observations. Evaluation uses 512 fresh target tasks with 32 training observations and exact constant-action population welfare. Target-bank ties follow a fixed deterministic rule.

The source-stage upper bound concerns horizon one. At target horizon 32, the losses are candidates whose actual risk is compared independently. The deterministic approximation and optimization certificates apply to these fits. The source-generalization bound in Appendix D is numerically too conservative to certify their risk. We report observed feature rank, norm, residual from constant-in-context functions, nonlinear-action residual, and solver certificates in `contextual_diagnostics.csv`. All 440 entropic/quadratic functional fits have the correctly scaled Hilbert-gradient certificate below 2.1×10^{-12}. Nested feasible spaces ensure that exact objective minima are nonincreasing in p on common data.

Table 9: Horizon study: target risk $\times 10^3$ for the construction variants at capacities $p = 1$ (top) and $p = 128$ (bottom), with 95% intervals over 20 repetitions. At $p = 128$ and $m = 8$, both target-horizon estimators reduce risk relative to horizon-one construction. All entries are rounded to three decimal places.

Capacity $p = 1$				
setting	variant	$m = 1$	$m = 8$	$m = 32$
$N = 512, n = 32$	horizon one	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	V-statistic	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	U-statistic	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	exact law	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	empirical risk	26.688 ± 0.089	0.110 ± 0.000	0.000 ± 0.000
$N = 128, n = 128$	horizon one	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	V-statistic	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	U-statistic	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	exact law	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	empirical risk	26.635 ± 0.088	0.109 ± 0.000	0.000 ± 0.000
$N = 32, n = 512$	horizon one	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	V-statistic	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	U-statistic	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	exact law	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	empirical risk	26.616 ± 0.086	0.109 ± 0.000	0.000 ± 0.000
$N = 128, n = 32$	horizon one	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	V-statistic	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	U-statistic	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	exact law	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	empirical risk	26.660 ± 0.089	0.109 ± 0.000	0.000 ± 0.000
$N = 128, n = 512$	horizon one	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	V-statistic	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	U-statistic	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	exact law	53.333 ± 0.000	0.005 ± 0.000	0.000 ± 0.000
	empirical risk	26.659 ± 0.089	0.109 ± 0.000	0.000 ± 0.000
Capacity $p = 128$				
setting	variant	$m = 1$	$m = 8$	$m = 32$
$N = 512, n = 32$	horizon one	3.114 ± 0.273	0.019 ± 0.003	0.000 ± 0.000
	V-statistic	3.114 ± 0.273	0.001 ± 0.000	0.000 ± 0.000
	U-statistic	3.114 ± 0.273	0.003 ± 0.000	0.000 ± 0.000
	exact law	5.288 ± 0.000	0.001 ± 0.000	0.000 ± 0.000
	empirical risk	2.459 ± 0.274	0.013 ± 0.003	0.000 ± 0.000
$N = 128, n = 128$	horizon one	19.533 ± 0.465	0.022 ± 0.004	0.000 ± 0.000
	V-statistic	19.533 ± 0.465	0.003 ± 0.000	0.000 ± 0.000
	U-statistic	19.533 ± 0.465	0.003 ± 0.000	0.000 ± 0.000
	exact law	5.288 ± 0.000	0.001 ± 0.000	0.000 ± 0.000
	empirical risk	19.447 ± 0.475	0.010 ± 0.002	0.000 ± 0.000
$N = 32, n = 512$	horizon one	40.385 ± 0.457	0.012 ± 0.002	0.000 ± 0.000
	V-statistic	40.385 ± 0.457	0.004 ± 0.000	0.000 ± 0.000
	U-statistic	40.385 ± 0.457	0.004 ± 0.000	0.000 ± 0.000
	exact law	5.288 ± 0.000	0.001 ± 0.000	0.000 ± 0.000
	empirical risk	40.375 ± 0.452	0.005 ± 0.000	0.000 ± 0.000
$N = 128, n = 32$	horizon one	20.064 ± 0.651	0.017 ± 0.003	0.000 ± 0.000
	V-statistic	20.064 ± 0.651	0.003 ± 0.000	0.000 ± 0.000
	U-statistic	20.064 ± 0.651	0.005 ± 0.000	0.000 ± 0.000
	exact law	5.288 ± 0.000	0.001 ± 0.000	0.000 ± 0.000
	empirical risk	19.977 ± 0.651	0.009 ± 0.002	0.000 ± 0.000
$N = 128, n = 512$	horizon one	19.567 ± 0.644	0.020 ± 0.003	0.000 ± 0.000
	V-statistic	19.567 ± 0.644	0.003 ± 0.000	0.000 ± 0.000
	U-statistic	19.567 ± 0.644	0.003 ± 0.000	0.000 ± 0.000
	exact law	5.288 ± 0.000	0.001 ± 0.000	0.000 ± 0.000
	empirical risk	19.515 ± 0.643	0.010 ± 0.003	0.000 ± 0.000

Table 10: Unknown-nuisance target regret $\times 10^3$, with 95% intervals over 20 repetitions, and selection frequencies out of 20. The library has 26 members; all methods use the same source episodes and 32 target observations.

Method	Target regret ($\times 10^3$)	Sel. freq.
Direct	0.372 ± 0.058	0
Squared	0.372 ± 0.058	0
SPO+	0.591 ± 0.081	0
Outcome regression ($p = 128$)	0.064 ± 0.006	5
One-strength scaling ($p = 128$)	0.065 ± 0.007	3
Two-strength scaling ($p = 128$)	0.065 ± 0.006	1
Entropic, action-linear ($p = 1$)	0.312 ± 0.052	0
Entropic, action-linear ($p = 8$)	0.133 ± 0.023	0
Entropic, action-linear ($p = 16$)	0.098 ± 0.012	2
Entropic, action-linear ($p = 32$)	0.094 ± 0.012	0
Entropic, action-linear ($p = 128$)	0.122 ± 0.012	0
Quadratic, action-linear ($p = 1$)	0.311 ± 0.053	0
Quadratic, action-linear ($p = 8$)	0.131 ± 0.024	0
Quadratic, action-linear ($p = 16$)	0.093 ± 0.012	0
Quadratic, action-linear ($p = 32$)	0.084 ± 0.011	4
Quadratic, action-linear ($p = 128$)	0.077 ± 0.008	4
Entropic, nonlinear-action ($p = 1$)	0.315 ± 0.053	0
Entropic, nonlinear-action ($p = 8$)	0.134 ± 0.023	0
Entropic, nonlinear-action ($p = 16$)	0.109 ± 0.012	0
Entropic, nonlinear-action ($p = 32$)	0.108 ± 0.013	0
Entropic, nonlinear-action ($p = 128$)	0.171 ± 0.020	0
Quadratic, nonlinear-action ($p = 1$)	0.313 ± 0.053	0
Quadratic, nonlinear-action ($p = 8$)	0.131 ± 0.026	0
Quadratic, nonlinear-action ($p = 16$)	0.100 ± 0.013	0
Quadratic, nonlinear-action ($p = 32$)	0.092 ± 0.013	1
Quadratic, nonlinear-action ($p = 128$)	0.098 ± 0.014	0
Selected (full library)	0.075 ± 0.011	–
Selected (entropic only)	0.100 ± 0.014	–
Selected (non-entropic only)	0.072 ± 0.008	–

Acquisition costs. At $N = 512$, the sensor study uses $512(64{+}64{+}32){+}2048(1{+}128) = 346112$ construction and selection observations. The contextual study uses $128(64 + 64 + 32) + 512(32 + 128) = 102400$. For T_{deploy} target deployments, divide these totals by T_{deploy} and add the respective target training horizons, one and 32, to obtain amortized observation costs. These counts exclude simulation-only evaluation of reported target risks.

E.10 Numerical verification

The primary studies comprise 3,200 sensor entropic fits, 1,600 variance fits, and 1,600 empirical-risk fits, plus the factorial and estimation grids, the horizon-study grid fits, and 440 contextual functional fits; the counts follow from the specified grids and are recorded in `experiment_manifest.json`. Solver certificates are floating-point diagnostics with the maxima reported above, not interval-arithmetic proofs. Exact enumeration is used for sensor conditional risks and cancellation risks, including their oracle comparisons. Integration over continuous descriptors is numerical; variation from fresh-descriptor sampling is included in the reported intervals. Recorded per-solve times range from milliseconds for the sensor bisection to at most a few seconds for the contextual L-BFGS and the U-statistic dynamic program on the execution machine. These times describe the implementations used here.

F Transfer mechanisms

In the models below, $R_m(S)$ abbreviates the meta-risk $\mathcal{R}_m(S)$, averaged over the specified task parameters as well as training samples.

Proposition 5 (Reusable nuisance correction)**.** *Let $\theta_\tau \in \{-d, d\}$ be equiprobable, independently across tasks, and let N, E be independent Rademacher variables, independent of θ_τ. The descriptor is constant and does not reveal θ_τ. Observe (Y, N), with $Y = \theta_\tau + \beta N + \sigma E$, where $0 \le \sigma < d$ and $\beta > d + \sigma$. Actions are $a \in \{-1, 1\}$ with welfare $w_\tau(a, O) = 1/2 + aY/(2C)$, $C = d + \beta + \sigma$. For $S_b(a, O) = -a(Y - bN)$: (i) all b have taskwise population losses $\mathbb{E}[S_b(a, O) \mid \theta_\tau] = -a\theta_\tau$; (ii) S_β has zero regret for every m; (iii) direct loss S_0 has one-observation meta-risk $d/(2C)$; and (iv) an action chosen from source data alone has the same risk. Moreover, with $\widehat{\beta} = N_s^{-1} \sum_{j=1}^{N_s} Y_j N_j$ from source observations that are independent conditional on their task parameters,*

$$\mathbb{P}\{|\widehat{\beta} - \beta| \ge (d - \sigma)/2\} \le 2 \exp[-N_s(d - \sigma)^2/(8C^2)].$$

On the complementary event $S_{\widehat{\beta}}$ has zero risk on every fresh task and every target sample size.

Proof. Since $\mathbb{E}N = 0$, population values are $-a\theta_\tau$. For $b = \beta$, $Y - \beta N = \theta_\tau + \sigma E$ has the sign of θ_τ, so every empirical average selects the optimal action. For $b = 0$, βN dominates $\theta_\tau + \sigma E$, giving an error probability $1/2$ at $m = 1$ and welfare gap d/C. A source-only action is independent of the new symmetric sign. Finally $\mathbb{E}[YN \mid \theta_\tau] = \beta$ and $|YN| \le C$. Conditional on all source-task parameters, the variables $Y_j N_j$ are independent with common mean β and range contained in $[-C, C]$. Hoeffding gives $\mathbb{P}(|\widehat{\beta} - \beta| \ge t) \le 2 \exp[-N_s t^2/(2C^2)]$. The bound does not depend on the parameters, so it also holds unconditionally. Take $t = (d - \sigma)/2$. On the complementary event, $|\widehat{\beta} - \beta| + \sigma < (d + \sigma)/2 < d$, so every corrected sample has the correct sign, simultaneously for both task signs and all sample sizes. □

Subtracting the fitted nuisance from the outcome produces the same empirical comparisons. Structural centering of N preserves population losses for every fitted coefficient; estimation accuracy determines whether the stronger zero-risk property holds.

For $d = .3$, $\sigma = .05$, $\beta = 1$, the sufficient condition for zero risk is $|\widehat{\beta} - 1| < .25$. The displayed concentration inequality instead controls the smaller event threshold .125; at $N_s = 1024$ its failure bound is $2 \exp[-1024(.25)^2/(8(1.35)^2)] \approx .0248$. The conservative bound can be sharpened because $YN - \beta = \theta_\tau N + \sigma EN$ has magnitude at most $d + \sigma$; the proposition uses the simpler common reward bound C.

Proposition 6 (Tail-sensitive two-sensor transfer)**.** *Let a task have independent equiprobable hidden sign $\theta \in \{-1, 1\}$. Let independent V, U have laws $\mathbb{P}(V = .5) = .8$, $\mathbb{P}(V = 3) = .2$, $\mathbb{P}(U = -1) =$*

.2, $\mathbb{P}(U = 1.5) = .8$. *Environment A observes* $(Y_1, Y_2) = (\theta V, \theta U)$*; environment B swaps the coordinates. Put* $Y = (Y_1 + Y_2)/2$, $Z = Y_1 - Y_2$, $w(a, O) = 1/2 + aY/6$, *and* $S_t(a, O) = -a(Y + tZ)$, $|t| \le 1/2$. *All losses preserve population preferences. In both environments,* $t = 0$ *uniquely minimizes variance and has one-observation risk* $4/75$. *Every* $t > 1/6$ *in A, or* $t < -1/6$ *in B, has zero risk at every horizon. For* $T = 4, \rho = .01$, *every global minimizer of the population horizon-one entropic objective with weight* $1/3$ *and regularizer* $\rho t^2/2$ *lies in this strict safe region, whether the multiplier is fixed at* T *or optimized over* $[0, T]$.

Proof. Both sensor laws have mean one and variance one. In A, the signed corrected observation is $Q_t = (1/2+t)V+(1/2-t)U$, with mean one and variance $1/2+2t^2$. Also $\mathbb{E}[Z \mid \theta] = 0$, and $|Y| \le 9/4$ ensures bounded rewards. The action welfare gap is $1/3$. At $t = 0$ only $(V, U) = (.5, -1)$ is negative, with probability $4/25$, giving risk $4/75$. The smallest support value is $-1/4+3t/2$, strictly positive for $t > 1/6$. For $t \le 1/6$, the harmful comparison $D = 2Q_t$ is nonpositive with probability at least $4/25$, so $\mathbb{E}e^{-\eta D} \ge 4/25$ for all $\eta \ge 0$. At $t = 1/2$, $D = 2V \ge 1$, giving objective at most $e^{-T}/3 + \rho/8 < 4/75$. A global minimizer therefore cannot be in the unsafe region. Continuity and compactness ensure existence in both the fixed- and optimized-multiplier formulations. Swapping sensors replaces t by $-t$ and proves all statements for B. □

Proposition 7 (Population inconsistency)**.** *In this sensor model, let* t *be constant and put* $M_4(t) = \mathbb{E}\exp(-8Q_t)$. *With exact welfare weights, multiplier four, and construction horizon one, the population objective is*

$$F_{4,1}(t) = \tfrac{1}{6}\{M_4(t) + M_4(-t)\} + \tfrac{\rho}{2}t^2, \qquad |t| \le \tfrac{1}{2}.$$

For every $\rho > 0$, *it is even and strictly convex, so its unique minimizer is* $t = 0$, *although* $\mathcal{R}_1(S_0) - \inf_{t \text{ constant}} \mathcal{R}_1(S_t) = 2/75$. *The gap persists with exact population information and global optimization.*

F.1 Population inconsistency of the fixed-multiplier construction

The same sensor model separates the exponential construction objective from decision risk at the population level. Define the fixed-multiplier comparison moment

$$M_\lambda(t) = \mathbb{E}\exp(-2\lambda Q_t), \qquad Q_t = \left(\tfrac{1}{2} + t\right)V + \left(\tfrac{1}{2} - t\right)U.$$

For $\lambda = 4$,

$$M_4(t) = \frac{4}{25}e^{2-12t} + \frac{16}{25}e^{-8+8t} + \frac{1}{25}e^{-8-32t} + \frac{4}{25}e^{-18-12t}. \tag{27}$$

Proposition 8 (Population inconsistency in sensor cells)**.** *Consider constant corrections* $t \in [-\frac{1}{2}, \frac{1}{2}]$, *exact welfare weights* $B_r = 1/3$, *multiplier* $\lambda = 4$, *and construction horizon one.*

(i) If both orientations are equally represented, the population construction objective is

$$F_{4,1}(t) = \frac{1}{6}\{M_4(t) + M_4(-t)\} + \frac{\rho}{2}t^2,$$

which is even and strictly convex; its unique minimizer is $t_{\text{ent}} = 0$ *for every* $\rho > 0$. *Hence*

$$R_1(S_{t_{\text{ent}}}) - \inf_{t \text{ constant}} R_1(S_t) = \frac{4}{75} - \frac{2}{75} = \frac{2}{75} > 0.$$

(ii) Within a descriptor cell in the model of Section 6.1, let $\pi = \mathbb{P}\{e(A) = 1 \mid A \text{ lies in the cell}\} \ge 1/2$. *Ignoring regularization, the cell objective is proportional to* $\pi M_4(t) + (1 - \pi)M_4(-t)$, *which is convex in* t. *Its minimizer exceeds the safe threshold* $1/6$ *only when*

$$\pi > \frac{-M_4'(-1/6)}{-M_4'(1/6) - M_4'(-1/6)} \approx 0.9821,$$

and a positive quadratic regularizer makes this requirement more stringent. Whenever the minimizer is strictly below the threshold, the cell's conditional one-observation risk is exactly the direct-loss risk $4/75$, *while the class-optimal constant correction achieves* $\frac{4}{75}\min\{\pi, 1 - \pi\}$; *at* $\pi = 0.98$ *the ratio is* 50.

Proof. (i) $t \mapsto Q_t$ is affine, so M_λ is a moment generating function composed with an affine map and is convex; it is nonnegative, so M_λ raised to any positive integer power is convex. The displayed objective is a symmetrized sum of such terms, hence even and convex, and $\frac{\rho}{2}t^2$ makes it strictly convex. An even strictly convex function has its unique minimizer at zero. For the risk values, at $t = 0$ the negative atom $(V, U) = (.5, -1)$ has probability $4/25$ and the welfare gap is $1/3$, giving $R_1(S_0) = 4/75$; the mixture risk of a constant t is $\frac{1}{2}(r_1(t) + r_1(-t))$ with r_1 the single-orientation risk, which equals $2/75$ for every $t \in (1/6, 1/4)$: one orientation is safe and the other's only negative atom has mass $4/25$.

(ii) Convexity is inherited from M_4. Both $M_4'(1/6) \approx -1.9135$ and $M_4'(-1/6) \approx -104.9169$ are negative, so the derivative of the cell objective at $t = 1/6$, $\pi M_4'(1/6) - (1-\pi)M_4'(-1/6)$, is negative—and the minimizer then lies strictly above $1/6$—exactly when π exceeds the displayed ratio. A quadratic regularizer adds a nonnegative slope at $t = 1/6$, so the safe region can only shrink. For the risk statement, $r_1(s) = 4/75$ for every $s \in (-1/6, 1/6)$: on this interval the only negative support atom of Q_s is $(V, U) = (.5, -1)$ with probability $4/25$. For $\pi \geq 1/2$, the derivative at zero is nonpositive, so the unique minimizer is nonnegative. A minimizer strictly below $1/6$ therefore lies in this interval and has conditional risk $\pi r_1(t) + (1-\pi)r_1(-t) = 4/75$. At $t = 1/6$, the zero atom requires the stated tie rule and this equality need not hold. The correction $t = .2$ is safe on the $e = 1$ orientation and incurs $4/75$ on the other, giving $\frac{4}{75}(1-\pi)$; the correction $t = -.2$ gives $\frac{4}{75}\pi$; the better sign gives $\frac{4}{75}\min\{\pi, 1-\pi\}$. That this is the exact class optimum follows from the piecewise-constant structure of r_1: its breakpoints on $[-\frac{1}{2}, \frac{1}{2}]$ are $-1/4$, $1/6$, and their reflections, and comparing the finitely many interval values gives $\min_t[\pi r_1(t) + (1-\pi)r_1(-t)] = \frac{4}{75}\min\{\pi, 1-\pi\}$. □

The positive risk gap arises at the exact population optimum. Exponential weighting magnifies the minority orientation's unfavorable comparison values: $M_4(-1/6) \approx 8.74$ is dominated by the rare negative atom. At $\pi = .98$, accepting errors on that orientation would instead reduce conditional decision risk by a factor of fifty. Together with Proposition 6, this shows how mixing orientations within a cell changes the behavior of the construction objective. Section 6 measures this gap alongside source-estimation and capacity effects.

G Limits of moment-based construction

G.1 Cancellation across optimal candidates

Let candidates $1, 2, 3$ have welfare $(1, 1, 0)$, and let Z, Y be independent Rademacher variables. For $0 < \varepsilon < 1/2$ and $d = 1/2$, define losses on three distinct constant reports by

$$Z(S_t, O) = (a(t)Z, -a(t)Z, d + s(t)Y), \quad a(t) = 1-(1-\varepsilon)t, \quad s(t) = \varepsilon+(1-\varepsilon)t, \quad 0 \leq t \leq 1.$$

They are mixtures of the losses at the two endpoints, and all have population coefficient $(0, 0, d)$. Every population minimizer is welfare-optimal.

Proposition 9 (Joint references and support separation)**.** *With a fixed valid Bernstein range bound, the single-reference exponent uniquely selects $t = 1/2$, where $R_2(S_{1/2}) = 1/8$. In contrast, $R_m(S_0) = 0$ for every m. A convex reference with equal weights on candidates $1, 2$ gives support margin $d - s(t)$. For every finite multiplier budget $T > 0$, the population full-comparison Chernoff construction uniquely prefers $t = 0$.*

Proof. Relative to either good candidate, the harmful comparison has mean d and variance

$$a(t)^2 + s(t)^2 = \frac{(1+\varepsilon)^2}{2} + 2(1-\varepsilon)^2(t - 1/2)^2.$$

Its Bernstein exponent is strictly decreasing in variance when the positive mean and the valid range bound are fixed. Thus it uniquely prefers $1/2$. At $t = 0$, the best good empirical loss is $-|\overline{Z}_m| \leq 0$ and the harmful empirical loss is at least $d - \varepsilon > 0$. At $t = 1/2$, set $q = (1+\varepsilon)/2 > d$. With $m = 2$, the harmful candidate is selected exactly when $\overline{Z}_2 = 0$ and $\overline{Y}_2 = -1$: if $|\overline{Z}_2| = 1$, the best good loss is $-q$, strictly below $d - q$; if $\overline{Y}_2 \geq 0$, the harmful loss is positive. The two required events have probabilities $1/2$ and $1/4$, independently. All relevant comparisons are strict, so the risk is $1/8$.

For multipliers (λ_1, λ_2), write $\eta = \lambda_1 + \lambda_2 \le T$. Independence gives

$$\mathbb{E}e^{-\lambda^\top D_3} = e^{-\eta d}\cosh(\eta s(t))\cosh((\lambda_2 - \lambda_1)a(t)).$$

For a fixed η, equal multipliers minimize the last factor to one. The remaining optimized negative log moment is $\max_{0\le\eta\le T}\{\eta d - \log\cosh(\eta s(t))\}$. Its derivative at $\eta = 0$ is $d > 0$, so any optimizer has $\eta > 0$. For any such η, the expression strictly decreases as $s(t) > 0$ increases. Testing an optimizer at a larger t therefore proves that its optimized value strictly decreases in t. The unique optimum is 0. Finally, the equal convex reference has sample loss zero, so its strict support margin is $d - s(t)$. □

With $\varepsilon = .1$, every $t \in [0, 4/9)$ has a strictly positive support margin and zero risk. At $t = 4/9$, the margin vanishes; risk remains zero under the prescribed smallest-index tie rule.

Exact risk is computed from independent binomial counts:

$$R_m(S_t) = 4^{-m}\sum_{k,l=0}^{m}\binom{m}{k}\binom{m}{l}\mathbf{1}\{md + s(t)(2l - m) < -a(t)|2k - m|\}.$$

The strict inequality follows because the harmful candidate has index 3. No Monte Carlo or asymptotic approximation is used in this diagnostic.

G.2 Two indistinguishable moment worlds

Let V equal $1/2$ with probability $4/5$ and 3 with probability $1/5$, and let U equal -1 with probability $1/5$ and $3/2$ with probability $4/5$. Both have mean one, second moment two, and variance one. There are two candidates, with welfare $(1, 0)$, and two dictionary losses $S^{(1)}, S^{(2)}$. The good candidate has loss zero under both; the bad candidate has losses (Y_1, Y_2). In World A these are independent (V, U), and in World B they are independent (U, V). Thus the full joint dictionary loss vector has the same mean and covariance in both worlds.

Proposition 10 (Moment-only design cannot recover a safe loss)**.** *For $S_t = (1-t)S^{(1)} + tS^{(2)}$ with $t \in [0, 1]$, every deterministic design rule using only these moments chooses a weight whose one-observation regret is at least $4/25$ in one of the two worlds. Nevertheless, World A has a zero-risk endpoint $t = 0$ and World B has a zero-risk endpoint $t = 1$.*

Proof. The harmful loss is $Y_t = (1 - t)Y_1 + tY_2$. In World A, the event $(Y_1, Y_2) = (1/2, -1)$ has probability $4/25$ and gives $Y_t = 1/2 - 3t/2 < 0$ when $t > 1/3$. In World B, the event $(Y_1, Y_2) = (-1, 1/2)$ has the same probability and gives $Y_t = -1 + 3t/2 < 0$ when $t < 2/3$. Every $t \in [0, 1]$ satisfies at least one strict inequality, and a moment-only deterministic rule has identical input in both worlds. On that event the uniquely welfare-optimal candidate is strictly beaten, giving regret one. At the stated endpoints, the harmful loss is $V > 0$ surely, giving zero risk at every sample size. □

The restriction concerns the information available during loss construction. The full comparison law or independent episodes distinguish the two worlds. For a randomized moment-only rule, the two world risks sum to at least $4/25$, so one world has expected risk at least $2/25$. This follows by averaging the same pointwise event bound.

H Population behavior and connections to DFL

H.1 Two geometries in DFL

For a compact feasible set $\mathcal{F}_\tau$, a fixed measurable tie rule defines $x^\star_\tau(r) \in \arg\min_{x\in\mathcal{F}_\tau} r^\top x$; for realized cost C, the decision loss is $\ell_{\tau,\mathrm{dec}}(r, C) = C^\top x^\star_\tau(r) - \min_{x\in\mathcal{F}_\tau} C^\top x$ and $\mathcal{D}_\tau(h) = \mathbb{E}\ell_{\tau,\mathrm{dec}}(h(X), C)$. If welfare normalizes negative cost by a known scale $q_\tau > 0$, then $B_{\tau,i} = [\mathcal{D}_\tau(h_{\tau,i}) - \min_{g\in\mathcal{H}_\tau}\mathcal{D}_\tau(g)]/q_\tau$: the realized optimal-cost term cancels. Downstream normal cones cover predicted cost space; the candidate-selection cones instead cover empirical loss vectors indexed by predictors, with the prescribed tie rule resolving overlaps. Our probability bounds concern this second geometry; downstream welfare identifies which predictor cones are harmful.

H.2 Population calibration preserved by functional corrections

Fix a task and suppress its index. Write $W(h) = \mathbb{E}w_\tau(h, O)$, $L_S(h) = \mathbb{E}S(h(X), O)$, $r(h) = W^\star - W(h)$, and

$$\psi_{S,\mathcal{H}}(\varepsilon) = \inf_{h\in\mathcal{H}: r(h)\geq\varepsilon} [L_S(h) - \inf_{g\in\mathcal{H}} L_S(g)], \qquad \inf \varnothing = +\infty.$$

All losses in the following statements are integrable with finite infima.

Proposition 11 (Calibration and approximate preservation)**.** *If $v \in \ker \mathcal{M}$, there is a task-dependent constant b_τ such that taskwise almost surely $L_{S_\mathrm{b}+v}(h) = L_{S_\mathrm{b}}(h) + b_\tau$ for every bank member, and $\psi_{S_\mathrm{b}+v,\mathcal{H}} = \psi_{S_\mathrm{b},\mathcal{H}}$. For a finite bank, this function is positive at every $\varepsilon > 0$ if and only if every population base-loss minimizer is welfare-optimal. More generally, let $\widehat{L}_S(h)$ be the training-sample average of $S(h(X), O)$ and let $\widehat{h}$ be the learner's output. For nonnegative centering, estimation, and optimization errors $\varepsilon_{\mathrm{null}}, \varepsilon_{\mathrm{est}}, \varepsilon_{\mathrm{opt}}$, suppose*

$$\inf_b \sup_h |\mathbb{E}v(h(X), O) - b| \leq \varepsilon_{\mathrm{null}}, \quad \sup_h |\widehat{L}_S(h) - L_S(h)| \leq \varepsilon_{\mathrm{est}}, \quad \widehat{L}_S(\widehat{h}) \leq \inf_h \widehat{L}_S(h) + \varepsilon_{\mathrm{opt}},$$

where $S = S_\mathrm{b} + v$. Then

$$L_{S_\mathrm{b}}(\widehat{h}) - \inf_h L_{S_\mathrm{b}}(h) \leq 2\varepsilon_{\mathrm{null}} + 2\varepsilon_{\mathrm{est}} + \varepsilon_{\mathrm{opt}}.$$

If $\psi_{S_\mathrm{b},\mathcal{H}}(\varepsilon)$ strictly exceeds this bound, then $r(\widehat{h}) < \varepsilon$.

Proof. The kernel condition makes the expected correction vector constant across candidates. Subtracting the minimum removes this constant. On a finite bank, each harmful candidate has a strictly positive loss gap exactly when no harmful candidate minimizes the loss; taking a minimum over any nonempty regret level set proves the equivalence. For approximate preservation, choose b with uniform error at most $\varepsilon_{\mathrm{null}} + \eta$. Empirical comparison gives $L_S(\widehat{h}) \leq \inf_h L_S(h) + 2\varepsilon_{\mathrm{est}} + \varepsilon_{\mathrm{opt}}$. Substituting the uniform correction bound at both endpoints adds $2(\varepsilon_{\mathrm{null}} + \eta)$; let $\eta \downarrow 0$. Approximate minimizers justify the same argument for infinite classes. The final conclusion follows directly from the definition of ψ. □

Sensitivity to centering error. For a product-structured correction, suppose $v(a, r, O) = f(a, r, X)^\top \zeta(O)$ with $\|\mathbb{E}[\zeta \mid X, \tau]\| \leq \delta$ and $\|f(a, r, X)\| \leq G$ on the permitted inputs. Then $|\mathbb{E}[v(a, h(X), O) \mid \tau]| \leq G\delta$ uniformly, so the proposition applies with $\varepsilon_{\mathrm{null}} \leq G\delta$: the population preference error is bounded by the product of the centering error and correction magnitude. The binary stress tests in Appendix I illustrate ordering reversal under a shifted control law. Cross-fitting can reduce dependence and estimation bias, but does not establish exact taskwise conditional centering. Bounding the correction norm limits amplification of a given centering error; a numerical guarantee additionally requires control of δ.

H.3 Restricted-class behavior of SPO+

For a report r, realized cost C, and the action map $x^\star(r) \in \arg\min_{x\in\mathcal{F}} r^\top x$, define the SPO+ loss

$$\ell_+(r, C) = \max_{x\in\mathcal{F}} (C - 2r)^\top x + 2r^\top x^\star(C) - C^\top x^\star(C).$$

Proposition 12 (A restricted-class preference reversal)**.** *Let $\mathcal{F} = [-1, 1]$, $C \sim \mathrm{Uniform}[.9, 1.1]$, and let independent X satisfy $\mathbb{P}(X = 0) = .9$, $\mathbb{P}(X = 1) = .1$. For $h_A \equiv .1$ and $h_B(0) = 1, h_B(1) = -.1$, expected decision regrets are $0, .2$, whereas expected SPO+ losses are $1.6, .24$. Thus mean-null corrections to SPO+ retain its incorrect ordering on this bank.*

Proof. Positive costs have optimal action -1; a negative prediction incurs regret $2C$. For positive c, scalar SPO+ equals $|c - 2r| + c - 2r = \max\{0, 2c - 4r\}$. Substitution gives the four stated expectations. The constant predictor 1 is an unrestricted population minimizer, but is not in the bank. For the auxiliary full-support variant $C_\sigma = 1 + \sigma G$, $G \sim N(0, 1)$ independent of X, decision regrets are $\mathbb{E}|C_\sigma| - 1$ and $\mathbb{E}|C_\sigma| - .8$. Their difference remains .2, and the strict surrogate reversal persists for sufficiently small positive σ by dominated convergence: for $\sigma \leq 1$ and the three fixed reports the losses are bounded by a constant times $1 + |G|$. □

The Fisher-consistency theorem of Elmachtoub & Grigas (2022, Theorem 4.2) concerns the unrestricted predictor class. Its assumptions include conditional central symmetry, a continuous conditional cost distribution on the full space, a unique conditional-mean optimizer, and suitable feasible-set conditions. The Gaussian variant illustrates the distinction under these stronger distributional conditions; its unbounded costs place it outside our bounded-reward selection theorem without a separate tail analysis.

H.4 Attained action cones versus statistical optimality

Proposition 13 (Relaxation cones)**.** *Let $P = \operatorname{conv} \mathcal{F} \subseteq Q$ and $x^\star \in \mathcal{F}$. Define $K_Q(x^\star) = \{r : r^\top(x - x^\star) \geq 0$ for every $x \in Q\}$. Then $K_Q(x^\star) \subseteq K_P(x^\star)$. Unique exposure of $x^\star$ over Q implies unique exposure over P. Closed-cone membership alone does not guarantee selection of $x^\star$ under every tie rule.*

Proof. The inequalities, and their strict versions for $x \neq x^\star$, restrict from Q to P. For the tie caveat take $r = 0$ and a feasible set containing two distinct points: every point minimizes the objective. □

The inclusion describes the geometry of attained action cones in relaxation-based alignment (Tang & Khalil, 2023). Guarantees for a training objective and its finite-sample decision risk require additional analysis. Under outcome noise, the realized cost optimizer and the conditional expected-cost optimizer can differ.

H.5 Approximation on the outputs that a learner uses

Proposition 14 (Class-relative surrogate approximation)**.** *Let $\mathcal{D}$ and ℓ_{dec} denote the task's expected and realized decision losses defined at the start of this appendix. Suppose $\alpha > 0$, $b \in \mathbb{R}$, $\varepsilon_{\mathrm{app}} \geq 0$, and $\sup_{h \in \mathcal{H}} |L_S(h) - \alpha \mathcal{D}(h) - b| \leq \varepsilon_{\mathrm{app}}$. Under the estimation and optimization conditions of Proposition 11,*

$$\mathcal{D}(\widehat{h}) - \inf_h \mathcal{D}(h) \leq (2\varepsilon_{\mathrm{app}} + 2\varepsilon_{\mathrm{est}} + \varepsilon_{\mathrm{opt}})/\alpha.$$

If the pointwise error $|S(h(X), O) - \alpha \ell_{\mathrm{dec}}(h(X), C) - b|$ is at most $\varepsilon_{\mathrm{loc}}$ on a region, at most G outside it, and every predictor exits that region with probability at most π, one may use $\varepsilon_{\mathrm{app}} = \varepsilon_{\mathrm{loc}} + G\pi$.

Proof. Empirical comparison bounds excess population surrogate loss by $2\varepsilon_{\mathrm{est}} + \varepsilon_{\mathrm{opt}}$. Apply the approximation bound at the selected predictor and an approximate decision-risk minimizer, then divide by α. For the local statement, split the expected absolute approximation error over the region and its complement. □

The approximation region must cover predictor outputs uniformly over the relevant class. These outputs can extend beyond perturbations around sampled labels, as emphasized in LODL's evaluation (Shah et al., 2022). A fitted quadratic supplies curvature to an approximation of decision regret, which is piecewise constant before smoothing. This curvature and the covariance of learner comparisons in our Gaussian result describe different quantities.

H.6 Continuity and gradient information

Proposition 15 (A continuity obstruction)**.** *Suppose two open report regions meet at r_0 and select fixed actions with different expected costs. No sample loss that is almost surely continuous at r_0 and has an integrable envelope in a neighborhood of r_0 can preserve population direct cost up to a report-independent constant throughout that neighborhood.*

Proof. Choose a sequence approaching r_0 within each region. Almost-sure continuity and dominated convergence force both expected-loss limits to equal the expected loss at r_0. Exact preservation instead gives two different constant limits, a contradiction. □

This does not forbid smooth interpolation at finitely many prescribed report values. Separately, an empirical correction depending on a report only through its chosen action is constant on every parameter region preserving the actions of all training observations. Thus corrected direct training retains the original gradient difficulty.

H.7 Conditional denoising and the population construction criterion

Proposition 16 (Conditional Jensen for comparison moments)**.** *Let $Z \in \mathbb{R}^M$ be an integrable loss vector, let $\mathscr{F}$ be a sub-σ-algebra, and put $Z^\circ = \mathbb{E}[Z \mid \mathscr{F}]$. For a vector z, write $D_r(z) = (z_r - z_i)_{i \neq r}$ and use $J_r(Z) = \sup_{\lambda \geq 0}\{-\log \mathbb{E}e^{-\lambda^\top D_r(Z)}\}$. For every candidate r and $\lambda \in \mathbb{R}_+^{M-1}$,*

$$\mathbb{E}e^{-\lambda^\top D_r(Z^\circ)} \leq \mathbb{E}e^{-\lambda^\top D_r(Z)}, \qquad J_r(Z^\circ) \geq J_r(Z).$$

Expectations of exponentials may be infinite. If all members of an admissible family have this same conditional mean and the family contains Z°, the latter simultaneously minimizes every comparison moment and the unregularized population cone–Chernoff risk bound.

Proof. Linearity gives $D_r(Z^\circ) = \mathbb{E}[D_r(Z) \mid \mathscr{F}]$. Conditional Jensen for the nonnegative convex exponential proves the moment inequality. Taking the infimum over multipliers and then the negative logarithm proves the exponent comparison; summing the nonnegative welfare-weighted powers proves the last statement. □

In the additive contextual model, condition on task, context, and residual outcome noise. The observed independent centered controls contribute a conditionally mean-zero perturbation to direct-loss coefficients. Oracle denoising removes it. The conclusion requires that the family contain this oracle; a source-fitted basis need not do so. Estimation and regularization can change the optimum. Nor does Jensen order exact finite-sample risk: selection indicators are not convex functions of comparison vectors.

Mean-preserving noise that helps the selector. Removing centered noise can strictly worsen finite-sample selection even while improving every exponential comparison moment. Let a harmful-minus-good direct-loss comparison satisfy

$$D_0 = \begin{cases} -0.1, & \text{with probability } 0.9, \\ 1, & \text{with probability } 0.1, \end{cases}$$

so $\mathbb{E}D_0 = 0.01 > 0$, and permit the mean-preserving family $D_c = D_0 + cZ$ with Z an independent Rademacher variable. At one observation, $\Pr(D_0 < 0) = 0.9$ but $\Pr(D_{0.2} < 0) = 0.45$: adding centered noise halves the selection error, and with bounded rewards $w_{\text{good}} = (1 + D_0)/2$, $w_{\text{bad}} = (1 - D_0)/2$ the one-observation regrets are 0.009 and 0.0045. Nevertheless, for every $\lambda \geq 0$,

$$\mathbb{E}e^{-\lambda D_c} = \mathbb{E}e^{-\lambda D_0}\cosh(\lambda c) \geq \mathbb{E}e^{-\lambda D_0},$$

so exponential construction and one-observation risk rank these corrections in opposite orders. As in the sensor model of Proposition 7, the disagreement holds with exact population information within the given correction family.

Table 11 summarizes the connections established in this appendix.

I Validation power and supplementary experiments

I.1 Safety-gate power

The gate of Theorem 6 certifies improvement only when $\widehat{L}_B(\widehat{S}) + 2\epsilon_B \leq \widehat{L}_B(S_0)$, so useful power requires the true improvement to exceed $2\epsilon_B = \sqrt{2\log(2K/\delta)/B}$. At the sensor library size $K = 19$, $\delta = .05$, and $B = 2048$, this threshold is 0.0805, exceeding the maximum achievable improvement over direct loss at horizon one ($4/75 \approx 0.0533$). Acceptance would therefore require an upward fluctuation in estimated improvement. The gate accepts in none of the 400 repetition–source-size combinations in the sensor study (Table 5). For the paired variance-adaptive gate of

Table 11: What the framework establishes for DFL methods. Full statements and proofs are in Appendix H and Corollary 2.

Approach	Justified connection	Not implied
Direct decision loss	Class-relative alignment survives mean-null correction	Informative gradients or finite-sample optimality
SPO+	Existing calibration, and restricted-class bias, are preserved	Universal class-relative correctness
CaVE / cone alignment	An attained tie-safe relaxation cone gives the desired action	Optimality of its training objective
LODL / learned losses	Approximation on relevant outputs controls decision excess	A Taylor expansion of discontinuous regret
Differentiable pipelines	Independent episodes compare actual downstream risk	Global optimization or unbiased surrogate gradients

Appendix B, the range term alone requires $B \gtrsim 14\log(4K/\delta)/(3\Delta)$ to certify an improvement of size Δ; for $\Delta = 10^{-3}$ and the library sizes used here this already exceeds 3×10^4 episodes, before accounting for the variance term. These calculations explain the limited power of the gates at the reported selection budgets.

The acquisition budget can be allocated between tasks and within-task evaluation. Write D_b for a paired episode difference averaged over v evaluation observations, with variance decomposition $\mathrm{Var}(D_b) = \sigma^2_{\mathrm{between}} + \sigma^2_{\mathrm{within}}/v$ (the first term includes fresh-task and target-training variability). The variance of the mean under acquisition cost $C = B(m+v)$ is $(m+v)(\sigma^2_{\mathrm{between}}+\sigma^2_{\mathrm{within}}/v)/C$, and, when both variance components are positive, its continuous relaxation is minimized at $v^\star = \sqrt{m\sigma^2_{\mathrm{within}}/\sigma^2_{\mathrm{between}}}$. A separate pilot can estimate the variance components and determine an integer allocation, which is then frozen before validation.

I.2 Exact diagnostics and binary transfer

The exact diagnostics of Appendix G use $d = .5$, $\varepsilon = .1$, and $m \in \{1, 2, 3, 5, 10, 20\}$. Integer-scaled comparisons in the binomial enumeration retain the prescribed tie rule. The two moment worlds are enumerated directly on their four-point joint supports; equality of moments and the risk lower bound are also proved in Appendix G. For the endpoint-selection diagnostic, each episode trains on one observation and evaluates the deterministic candidate welfares $(1, 0)$. The safe endpoint never incurs episode loss; the unsafe endpoint incurs loss one with probability .2 independently. World A puts the safe endpoint first, so it is always selected. World B puts it second, so the unsafe endpoint is selected only when all B episodes miss its negative observation, with probability $.8^B$. Multiplying by its true risk .2 gives the exact expected selected risk.

The transfer experiment uses $(d, \sigma, \beta) = (.3, .05, 1)$ and seed 590701 for 100 independent meta-design repetitions. Each repetition has 64 construction tasks with 16 observations each. A separate block of 128 observations per construction task estimates the better action. The moment correction is $\widehat{\beta} = \overline{YN}$. The quadratic correction is the pooled within-task least-squares slope of centered Y on centered N. If its denominator vanishes, the implementation returns coefficient zero. The scalar entropic correction minimizes

$$\frac{1}{64 \cdot 16} \sum_{b,j} \exp[-\widehat{s}_b(Y_{bj} - uN_{bj})], \qquad u \in [0, 2],$$

where $\widehat{s}_b$ is the estimated optimal sign from the separate block and u is the scalar correction coefficient. This diagnostic uses a unit exponential multiplier and a one-observation construction criterion, held fixed across all target horizons. The scalar coefficient is optimized directly. The frozen library consists of direct loss and these three corrections. Oracle correction and source-only deployment are analytic comparators. Outcome denoising with $\widehat{\beta}$ makes exactly the moment-correction decisions and therefore has identical risk.

For each $m \in \{1, 2, 5, 10, 20, 40\}$, we generate 8192 independent selection tasks and evaluate prefixes of sizes $B \in \{128, 512, 2048, 8192\}$, using 128 evaluation observations per task. These nested prefixes provide paired comparisons across B, not independent repetitions. Training and evaluation

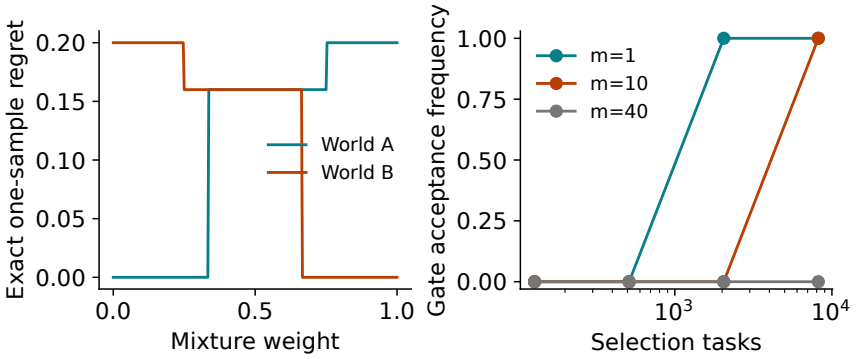

Figure 2: Auxiliary diagnostics. Left: moment-matched laws require different safe losses. Right: binary-transfer gate acceptance across 100 runs; smaller direct-risk gaps require more selection tasks.

Rademacher averages are sampled by their exact binomial laws. Target meta-risk is enumerated over both hidden task signs and the two binomial counts for every fitted coefficient. In floating-point calculations, scalar comparisons within 10^{-12} of zero are treated as ties in both training and enumeration. This floating-point enumeration is not an exact-arithmetic certificate for arbitrary fitted coefficients. The zero-risk result has a separate support certificate: the maximum values of $|b-1|$ over the 100 fitted moment, quadratic, and entropic coefficients are .03038, .00467, and .09449, respectively (rounded upward). Each is strictly below .25, so Proposition 5 certifies all three corrections at every sample size independently of numerical tie tolerances. Direct-loss risks are analytic binomial sums. The source cost, excluding reuse between diagnostic settings, is $64(16+128)+B(m+128)$ observations. The gate uses $\delta = .05$ and $K = 4$.

At $m = 1$, ordinary selection chooses a safe loss in all 100 runs, while the Hoeffding gate accepts in none at $B = 128, 512$ and all at $B = 2048, 8192$. At $m = 10$, direct risk is .03819 and acceptance requires $B = 8192$ in these runs. At $m = 40$, direct risk is .00662 and all gates fall back, including at $B = 8192$. Ordinary selection retains direct loss in two runs at $(m, B) = (40, 128)$, giving mean risk .0001324, and attains zero at larger B. The simultaneous library certificate covers the risks in all 100 repetitions of each setting. Confidence is specified separately for each setting. Figure 2 summarizes the moment-matching and gate diagnostics.

I.3 Contextual bank transfer

These supplementary experiments use banks of contextual predictors for combinatorial decisions. Their implementation is not included in `content-E/experiments/`. We evaluate four environments: shortest path without nuisance, shortest path with linear nuisance, knapsack with linear nuisance, and shortest path with nonlinear nuisance. Each environment has 20 independent meta-design repetitions. Within a repetition the nuisance loading is fixed across all construction, selection, and target tasks. Across repetitions it is redrawn; thus uncertainty intervals average over independent environments as well as task samples. Every invocation of the selection theorem is conditional on that environment, where source and target tasks have the same distribution.

Shortest path uses all 20 monotone paths on the directed right/down 4×4 grid: 24 edges and six edges per path. Knapsack has 12 items, task-specific independent weights uniform on $\{1, \dots, 5\}$, and capacity $\lfloor .4 \sum_e w_e \rfloor$. A vectorized dynamic program solves each knapsack exactly, including the empty set. All action optimizations use a fixed deterministic tie rule.

On every task, $X \sim \mathrm{Uniform}([-1,1]^5)$ and $\mu_\tau(X) = b_\tau + A_\tau X$. Path intercepts have independent entries uniform on $[-.5, .5]$; knapsack profit intercepts are uniform on $[.1, 1]$. Entries of A_τ are independent uniform on $[-.12, .12]$. The two observed controls Z are independent Rademacher variables, independent of X and the task. Realized costs or profits are

$$Y = \mu_\tau(X) + B_0 Z + \mathbf{1}_{\mathrm{nonlinear}} B_1 Z \tanh(X_1 + X_2) + E, \qquad E_e \sim \mathrm{Uniform}([-.05, .05]).$$

Entries of B_0, B_1 are independent uniform on $[-.75, .75]$ in nuisance environments, and zero in the no-nuisance environment. Consequently $|Y_e| \le 4.65$ in all environments. If $a(r)$ is the action selected by report r, define $c = Y$ for path and $c = -Y$ for knapsack, and use $w(r, O) = 1/2 - a(r)^\top c/(2L \cdot 4.65)$, with $L = 6$ or 12, respectively. This gives a fixed known $[0, 1]$ reward bound with no clipping.

Each task has 16 linear predictor maps, generated before all its samples:

$$h_i(X) = b_\tau + u_i + (A_\tau + V_i)X.$$

Four candidates are generated at each scale $s \in \{.08, .2, .45, .8\}$, with independent entries $u_{i,e} \sim \mathrm{Uniform}([-s, s])$ and $V_{i,ex} \sim \mathrm{Uniform}([-s/2, s/2])$. The simulator uses the latent map only to generate these banks and evaluate target welfare. Learners receive the candidate functions and problem descriptor; they do not receive μ_τ, the nuisance loading, or true candidate welfares. All methods receive identical fresh target banks.

I.4 Construction, baselines, and optimization

Each repetition uses 32 construction tasks, with 64 comparison observations and 128 independent welfare observations per task. The same construction observations fit a shared nuisance model by ridge regression of Y on $\phi = (Z_1, Z_2, Z_1 \tanh(X_1 + X_2), Z_2 \tanh(X_1 + X_2))$, with ridge 10^{-4}. Write its fitted loading as $(\widehat{B}_0, \widehat{B}_1)$. The constant conditional mean of each control is known to be zero; it is not estimated from source tasks.

The functional search uses

$$S_u(r, O) = a(r)^\top c - \mathrm{sgn}_{\mathrm{task}}\ a(r)^\top \{u_0 \widehat{B}_0 Z + u_1 \widehat{B}_1 Z \tanh(X_1 + X_2)\},$$

where $\mathrm{sgn}_{\mathrm{task}} = 1$ for costs and -1 for profits. We fit $u_0 \in \mathbb{R}$ with $u_1 = 0$ in linear environments, and both coefficients in the nonlinear environment. Conditional on construction, this is a bounded finite-rank RKHS of sample-loss corrections with coefficient regularizer $\rho\|u\|_2^2/2$. The family has one or two fitted coefficients and depends on the source nuisance estimate. Every fitted function still has zero taskwise mean because the controls are structurally centered. The base-loss coefficient is always one.

Direct loss and correction features are divided by the maximum of 10^{-8} and the standard deviation of direct losses over construction tasks, observations, and candidates. Predictive losses are normalized by their corresponding construction standard deviation (floored at 10^{-8}) when used in mixtures. Construction welfare gaps are divided by the maximum of their average and 10^{-8}, equivalently rescaling regularization when the average exceeds this floor. If all gaps vanish, the objective is only its regularizer: return the zero correction directly, recording objective zero (and $-\infty$ as a log-objective sentinel). The entropic library uses $\rho \in \{.01, .1\}$, $T \in \{.5, 2\}$, and single-best-reference or uniform-reference initialization, for eight losses. For each setting we perform eight alternating rounds: at most 120 BFGS iterations (gradient tolerance 10^{-8}) on the one- or two-dimensional correction coefficient, followed by 24 projected gradient steps on all multiplier vectors, with backtracking and projection onto the nonnegative ℓ_1 ball. After the final multiplier update we solve the coefficient block again (at most 200 iterations, tolerance 10^{-9}), retaining the previous coefficients if the objective worsens. A local grid checks all five coordinate offsets in $[-.2, .2]$ (25 points in two dimensions); any improving grid point initializes another coefficient solve. The two reference initializations provide distinct starting configurations. The target exponent is $m = 32$. Log-sum-exp is used for exponential averages. For the entropic objective, the solver evaluates $\log \widehat{F}$ and its exact gradient, including the regularizer, using log-sum-exp and normalized exponential weights. This strictly increasing transformation has exactly the same minimizers when $\widehat{F} > 0$ and requires no exponent cap; the all-zero-weight case is handled before taking logarithms. Backtracking also compares this log objective. The original correction subproblem is strongly convex; its logarithm need not be convex. The recorded objective and gradient are those of $\log \widehat{F}$ for entropic fits and of the quadratic objective otherwise. Let f denote the optimized objective (logarithmic for entropic fits), and let proj project onto the feasible multiplier set. We record final coefficient-gradient norms, projected multiplier residuals $\|\lambda - \mathrm{proj}(\lambda - \nabla_\lambda f)\|_2$, improvement from the final solve, and any remaining local-grid advantage. The residuals are blockwise diagnostics, not certified distances to a joint optimum. Finite-difference tests check coefficient and multiplier gradients at zero and nonzero coefficients, and extreme-coefficient tests check finite log-objective values and gradients.

The single-reference quadratic baseline fixes the estimated best reference. The full-reference quadratic baseline updates the reference multipliers. Write $\overline{d}_{br} = n_b^{-1} \sum_j d_{brj}$ and let $\widehat{\mathrm{Var}}(d_{br})$ be the corresponding sample variance with divisor n_b. Both use $(\rho, T) = (.01, 1)$ and the empirical approximation

$$\frac{1}{B_{\mathrm{con}}} \sum_{b,r} \widehat{B}_{br} \left\{ \tfrac{1}{2} \widehat{\mathrm{Var}}(d_{br}) - \overline{d}_{br} \right\} + \frac{\rho}{2} \|u\|_2^2.$$

Because empirical correction means are not exactly zero, the empirical mean term is retained. These moment baselines use a quadratic approximation to the log moment. The cancellation diagnostic in Appendix G separately analyzes a Bernstein criterion.

Other losses are direct decision loss, predictive squared error, SPO+, three direct/predictive mixtures with predictive weights $.25, .5, .75$, and shared outcome denoising with $u_0 = 1$ (and $u_1 = 1$ in the nonlinear environment). For cost prediction $\widehat{c}$ and true cost c, SPO+ is

$$\max_{a \in \mathcal{F}} (c - 2\widehat{c})^\top a + 2\widehat{c}^\top a^\star(c) - c^\top a^\star(c), \quad a^\star(c) \in \arg\min_{a \in \mathcal{F}} c^\top a.$$

Knapsack profits are negated before this calculation. The nonlinear library additionally includes linear-only denoising. Thus $K = 17$ in linear environments and $K = 18$ in the nonlinear environment. The fixed entropic baseline is the first declared setting $(\rho, T, \mathrm{init}) = (.01, .5, \mathrm{single})$, without risk selection. The full method selects over the complete library. The no-entropy ablation removes all eight entropic candidates before selecting; it uses the same selection episodes. A deterministic replay of construction and selection identifies its winner, and the raw target results for that baseline provide its paired evaluation. A mixture-only selector uses the direct/predictive sublibrary. All selection is performed after construction; no settings are added after inspecting selection outcomes.

I.5 Evaluation, uncertainty, and costs

There are 512 independent selection tasks, each with $m = 32$ training and $v = 128$ evaluation observations. The gate uses $\delta = .05$ and the actual library size. Each repetition evaluates 256 independent target tasks, with ten independent training blocks of size 32 per task. To evaluate candidate welfare, we integrate over 2048 independent target contexts, using the latent conditional mean to remove outcome-noise variance. This simulation-only integration is never available to a learner. The evaluation maximum has finite-integration bias common to all methods. Paired welfare differences cancel that maximum and remain unbiased for risk differences. Absolute meta-risks and certificate coverage are evaluated numerically. The binary experiment instead permits exact risk enumeration.

Reported 95% intervals use a Student-t interval across the 20 independent repetition means. We first average repeated training samples within each target task, then average tasks within a repetition. Paired differences are formed before averaging. This avoids treating repeated training samples as independent draws from the task distribution. Raw task-level data allow alternative clustered analyses. Timing includes loss evaluations and optimization on construction data, selection evaluation, and target evaluation separately; it is not a comparison of optimized implementations. Construction and selection times are recorded in the raw metadata. These are wall-clock measurements on the execution machine with one numerical-library thread; background load was not controlled. All 80 gates fall back and all 80 simultaneous contextual certificate diagnostics cover the integrated risk estimates. Neither this coverage nor small optimization residuals certify the empirical construction objective as a population bound.

Tables 12–14 report optimization diagnostics, paired risk comparisons, and safety-gate components for these contextual bank experiments. Across the 640 entropic fits, final coefficient residuals have mean $7.29 \cdot 10^{-9}$ and maximum $2.16 \cdot 10^{-7}$, but the maximum projected multiplier residual remains .0473 (rounded upward). No tested grid point improves the final objective. The largest improvements from the final coefficient solve are $3.64 \cdot 10^{-5}$ for the quadratic objective and $3.40 \cdot 10^{-4}$ for the entropic log objective (rounded upward). Adding entropic candidates changes selected risk by $(.27 \pm .56)10^{-6}$, $(1.27 \pm 1.47)10^{-6}$, and $(1.15 \pm 3.01)10^{-6}$ on linear paths, knapsack, and nonlinear paths, respectively. The intervals include zero in all three environments, so these runs do not establish an incremental benefit from the entropic candidates. Small coefficient residuals and zero grid advantage also leave open the convergence of the coupled coefficient–multiplier optimization.

Table 12: Final construction diagnostics over 80 fixed-reference, 80 full-reference quadratic, and 640 entropic fits. Coefficient and multiplier residuals use the identity objective for quadratic fits and the log objective for entropic fits, so their scales are not directly comparable. Multipliers are not optimized in the fixed-reference baseline. Grid max is the largest remaining objective advantage among the probed coefficient points.

Construction	Coef. mean	Coef. max	Mult. mean	Mult. max	Grid max
Fixed-reference quadratic	5.8×10^{-16}	3.7×10^{-15}	–	–	0
Full-reference quadratic	1.5×10^{-12}	5.4×10^{-11}	0.018	0.055	0
Entropic	7.3×10^{-9}	2.2×10^{-7}	0.02	0.047	0

Table 13: Selected risk minus comparator risk, $\times 10^6$, with paired 95% Student-t intervals across 20 repetitions. Negative favors selection. The extra-data direct comparator uses 720 target observations, corresponding to 128 deployments under equal amortized observation cost. Intervals are unadjusted comparisons, not simultaneous significance statements.

Environment	Direct	Denoising	No entropic losses	Direct, extra data
Path, none	-18.64 ± 1.04	-19.81 ± 1.05	$+0.00 \pm 0.00$	$+19.84 \pm 0.80$
Path, linear	-1144.31 ± 118.87	$+0.81 \pm 0.60$	$+0.27 \pm 0.56$	-90.95 ± 13.91
Knapsack	-328.96 ± 24.56	$+1.59 \pm 1.67$	$+1.27 \pm 1.47$	-26.36 ± 2.79
Path, nonlinear	-1322.38 ± 100.77	$+3.86 \pm 2.47$	$+1.15 \pm 3.01$	-107.46 ± 13.37

Table 14: Paired-gate components for the selected candidate and source-stage wall-clock times, averaged over 20 repetitions per environment. The range term dominates the variance term and the observed risk improvements.

Environment	Paired variance	Variance term	Range term	Construct (s)	Select (s)
Path, none	1.05×10^{-8}	1.71×10^{-5}	0.0659	10	0.508
Path, linear	6.2×10^{-6}	0.00041	0.0659	10.7	0.502
Knapsack	5.92×10^{-7}	0.000128	0.0659	10.8	1.79
Path, nonlinear	7.67×10^{-6}	0.000461	0.0664	23.3	0.494

Construction costs $32(64+128) = 6144$ observations; selection costs $512(32+128) = 81920$, totaling $N_{\text{meta}} = 88064$. For T_{deploy} deployments the effective observation cost is $32 + 88064/T_{\text{deploy}}$. We additionally run direct-loss ERM with $32 + \lceil 88064/T_{\text{deploy}} \rceil$ observations per target at $T_{\text{deploy}} \in \{128, 512, 2048\}$, namely $720, 204, 75$ observations. For each target, these three comparisons use prefixes of a separate stream, with ten training repetitions and the same welfare integration block. Source and target observations are counted equally in this accounting; application-specific acquisition costs may differ.

I.6 Assumption stress tests

We enumerate one-observation risk in the binary model for nuisance strengths $\beta \in \{0, .25, .5, 1, 2\}$, comparing uncorrected loss, the source correction $b = 1$, and the oracle correction $b = \beta$. A second test uses an independent target loading uniformly distributed over $\{.9, 1.1\}$ with $b = 1$; every corrected observation retains its task sign. A final test instead sets $\mathbb{P}(N = 1) = .95$ on targets, retains $\beta = 1$, and evaluates welfare using the new population mean $\theta_\tau + .9$. Both task signs now have optimal action $+1$. Direct loss chooses the wrong action with probability $.05$, giving risk $(.9/1.35)(.05) = 1/30$. The correction selects according to θ_τ, so it is wrong on the negative-sign tasks and has risk $(1/2)(.6/1.35) = 2/9$. This is deliberate centering misspecification and source-target shift; neither the population-preservation claim nor the episodic guarantee applies. These stress tests evaluate a fixed source correction without refitting.

I.7 Two-sensor estimation and exact target enumeration

The observations and rewards are defined in Proposition 6. Seeds are $810000 + 1000e + r$ for environment $e \in \{0, 1\}$ and repetition $r \in \{0, \dots, 99\}$. All draws of V, U are independent conditional on the task sign. On each of 64 source tasks, form comparison observations Y_{bj}, Z_{bj} for $j \le 32$ and an independent 128-observation welfare block. Let $\overline{Y}'_b$ be its mean, $\widehat{s}_b = \operatorname{sign}(\overline{Y}'_b)$ (negative on a tie), and $\widehat{B}_b = |\overline{Y}'_b|/3$. Let $\overline{Y}_b, \overline{Z}_b$ be the comparison-block means, and let $\operatorname{clip}_{[-1/2,1/2]}$ project a scalar onto that interval. The variance coefficient minimizes pooled within-task centered variance:

$$\widehat{t}_{\text{var}} = \operatorname{clip}_{[-1/2,1/2]} \left(- \frac{\sum_{b,j} (Y_{bj} - \overline{Y}_b)(Z_{bj} - \overline{Z}_b)}{\sum_{b,j} (Z_{bj} - \overline{Z}_b)^2} \right).$$

If the denominator vanishes, choose zero. The entropic coefficient minimizes the convex scalar objective

$$\frac{1}{64} \sum_b \widehat{B}_b \frac{1}{32} \sum_j e^{-8\widehat{s}_b (Y_{bj} + t Z_{bj})} + .01 t^2/2, \qquad |t| \le 1/2.$$

The factor eight is the fixed multiplier four times the harmful-minus-good loss difference factor two. Bounded scalar minimization uses tolerance 10^{-12} and explicitly checks both endpoints. The multiplier is fixed throughout this scalar optimization. Write $Y_{bj,k}$ for sensor k in comparison observation j on task b. The full-law baseline chooses the sensor with the lower pooled frequency of $\widehat{s}_b Y_{bj,k} < 0$, selecting sensor one on a tie. Sensor one and two correspond to $t = 1/2$ and $-1/2$. All procedures use the same source data; the support oracle is an analytic comparator, not a library candidate.

The library order is direct, variance, entropic, full-law, and score ties follow that order. For each horizon, a fresh set of 2048 selection tasks provides nested prefixes $B = 128, 512, 2048$, with 128 independent evaluation observations per task. Actions choose $+1$ only for strictly positive average corrected observations. Paired-gate confidence is .95 separately at each setting. Source cost for one setting is $64(32 + 128) + B(m + 128)$ observations. The fitted variance coefficients range from $-.03413$ to $.03975$, remaining near the population variance minimizer; all entropic coefficients lie in the appropriate strict safe region.

For target risk, enumerate all four-support multinomial counts. Their probability numerators are $(4, 16, 1, 4)$ with denominator 25; twice the sensor values are $(1, 1, 6, 6)$ and $(-2, 3, -2, 3)$. The stored fitted float is interpreted as its exact binary rational $t = n/d$. For summed twice-values v, u, the comparison sign in A is exactly the sign of $(d + 2n)v + (d - 2n)u$; in B negate n. Integer arithmetic determines signs and ties, and rational arithmetic sums masses. Ties are harmful for

exactly one of the two equiprobable task signs, giving risk $(p_{\text{negative}} + p_{\text{tie}}/2)/3$. Only the final reported risk is converted to floating point. This avoids a numerical tie tolerance in this sensor study.

At $m = 1$, mean paired variance is approximately .015 in both environments. For $B = 128, 512, 2048$, the mean variance-radius terms are approximately $.0365, .0184, .0092$, while the range terms are $.2120, .0527, .0132$. This explains why even zero-risk fitted losses are not certified over direct risk $4/75$ until $B = 2048$ in these runs. Longer horizons have smaller absolute risk differences. Raw outputs retain variance and both radius terms for every setting and repetition.

I.8 Fresh affine-predictor training

Seeds are $910000 + r$, $r = 0, \ldots, 19$. Use the shortest-path distribution above with linear nuisance: 24 edges, 20 feasible paths, five uniform context coordinates, task intercepts uniform on $[-.5, .5]$, mean slopes uniform on $[-.12, .12]$, shared nuisance loadings uniform on $[-.75, .75]$, two independent Rademacher controls, and edge noise uniform on $[-.05, .05]$. All source, selection, and target tasks have independent affine means. Each predictor is $h_\Theta(X) = \Theta\phi(X)$, $\phi(X) = (1, X^\top)^\top$, $\Theta \in \mathbb{R}^{24\times 6}$, $\|\Theta\|_F \le 6$. The latent mean lies in this ball since $\|\Theta_{\text{mean}}\|_F^2 \le 24(.5^2 + 5(.12)^2) < 36$. The evaluation benchmark is its conditional-mean optimal action. Learners do not receive this parameter or any latent conditional means.

Source construction uses 32 tasks with 64 observations each. Regress costs on the two controls with ridge 10^{-4} to obtain $\widehat{B}$. At source pilot reports $\widetilde{r} = Y - \widehat{B}Z$, compute a SPO+ subgradient $g = 2[x^\star(Y) - x^\star(2\widetilde{r} - Y)]$. Regress g on Z with the same ridge and negate the fitted loading to obtain G. The corrected loss is $(\ell_+(r, Y) + r^\top GZ)/6$. Corollary 2 applies to this source-fitted gradient-control heuristic. Conditional on the fit, $\mathbb{E}[Z \mid X, \tau] = 0$ preserves population SPO+ for every target predictor; the affine correction also preserves convexity in Θ.

A shared diagonal quadratic provides a learned-surrogate comparator. Repeat each source label four times, add independent $N(0, .3^2 I)$ report perturbations Δ, and regress realized decision regret of $Y + \Delta$ against the 24 squared coordinates Δ_e^2 using nonnegative least squares. Augment its design with $.01I$ and zero responses. Divide the weights by the maximum of their mean and 10^{-8}. Freeze this shared quadratic before selection and training. Uniform approximation on target predictor outputs is not assumed.

Let Φ have rows $\phi(X_j)^\top$ and let $\mathbf{Y}$ have rows $Y_j^\top$ for the target training block. Each method starts from the same ridge fit $\Theta_0 = [(\Phi^\top\Phi + .1I)^{-1}\Phi^\top\mathbf{Y}]^\top$, projected onto the radius-six ball. All objectives add $.005\|\Theta\|_F^2$: squared error uses $\|r - Y\|^2/48$; the learned quadratic uses its weighted version; plain and corrected SPO+ are divided by six. The smoothing method uses path probabilities proportional to $\exp(-r^\top x/.5)$ and minimizes the resulting expected realized path cost divided by six. This objective is generally nonconvex. All methods take 160 projected full-batch gradient or subgradient steps with size $2/\sqrt{1 + s/20}$ at iteration $s = 0, \ldots, 159$, returning the iterate with smallest observed empirical objective, including the final iterate. Optimization accuracy may differ. All five predictors deploy exact deterministic shortest-path optimization.

The 128 selection tasks each have 32 training and 128 independent evaluation observations. Select by mean evaluated welfare, using library order squared, SPO+, corrected SPO+, learned quadratic, smoothing. The 128 independent target tasks each have 32 training observations; evaluate on 2048 independent contexts using their conditional mean. Reward is $1/2 - Y^\top x/55.8$, as in the fixed-bank study. Source and selection cost is $32 \cdot 64 + 128(32 + 128) = 22528$ observations per repetition. Regret compares with the conditional-mean optimal action; the fixed-bank benchmark is the best bank member.

Analytic gradients pass central finite differences away from action ties; nonsmooth losses use valid subgradients. Writing g for the chosen objective subgradient in Θ and proj for projection onto its radius-six ball, mean update residuals $\|\Theta - \text{proj}(\Theta - g)\|_F$, averaged over all fitted selection and target predictors, are .00104 for squared loss, .09983 for SPO+, .09967 for corrected SPO+, .00157 for the learned quadratic, and .00570 for softmin. Their respective maxima are .00496, .24039, .28110, .00541, and .02380. For nonsmooth objectives these residuals use the chosen subgradient; they are not certified optimality gaps. Corrected SPO+ reduces risk by $(1.764 \pm .308)10^{-3}$ relative to plain SPO+ (mean and 95% interval half-width); squared loss has lower risk than both. Intervals use paired repetition means as independent replicates. Recorded times include training and evaluation.