Fast Swap-Agnostic Regression

Pahan Dewasurendra
Johns Hopkins University

Abstract

Swap-agnostic learning compares a forecaster with a different hypothesis on each level set of its predictions. A recent finite-class second-order approachability algorithm attains the sharp $\widetilde{O}(T^{1/3})$ time exponent. Extending it to infinite classes and optimizing its offset with an oracle were left open.

We give a black-box reduction to online square-loss regression. The algorithm maintains one regressor per prediction bucket, locates an adjacent sign crossing with logarithmically many oracle queries, and updates only the sampled bucket. A square-loss identity preserves the comparator's negative quadratic activity. This yields, for any base regret R,

$$\mathrm{SwapReg}_T = O\big(NR(T/N) + T/N^2\big)$$

up to loss and confidence factors. Logarithmic-regret regression therefore retains the $T^{1/3}$ exponent for infinite classes.

For half-Brier loss and bounded d-parameter linear predictors, constrained Online Newton Step gives $\widetilde{O}(T^{1/3}d^{2/3})$ contextual swap regret. A direct sum of recalibration instances gives a matching $\Omega(T^{1/3}d^{2/3})$ lower bound for a fixed class of pseudodimension d. Kernel ridge regression further gives a spectrum-adaptive log-determinant bound for bounded RKHS predictors. For every sequential-entropy exponent $p > 0$, fixed convex classes prove that both nonparametric rates are minimax optimal up to logarithmic factors. We also obtain fast canonical-link models and an oracle-efficient offline conversion with excess $\widetilde{O}((d/m)^{2/3})$.

1 Introduction

A probabilistic predictor is often reused after training. Different users may condition on its output and then choose different downstream hypotheses. Swap-agnostic learning formalizes this requirement. Instead of comparing with one $h \in \mathcal{H}$, it compares on every prediction level p with a separate $h_p \in \mathcal{H}$ (Gopalan et al., 2023). For a fixed loss ℓ, its online benchmark is

$$\sup_{\phi:\mathrm{range}(p_{1:T})\to\mathcal{H}} \sum_{t=1}^{T} \big[\ell(p_t, y_t) - \ell(\phi(p_t)(x_t), y_t)\big]. \quad (1)$$

This is contextual swap regret for squared loss (Garg et al., 2024; Luo et al., 2025a).

The benchmark looks like a separate agnostic-learning problem on every level set. Existing oracle reductions pay substantially for realizing the random level sets. Garg et al. (2024) obtain roughly $T^{3/4}$ contextual swap regret. Luo et al. (2025a) reduce pseudo contextual swap regret to online regression at rate $T/N^2 + Nd\log T$, but concentration for realized level sets leads to $\widetilde{O}(T^{3/5}d^{2/5})$. In a complementary line, Hu et al. (2026a) attain $\widetilde{O}(T^{1/3})$ squared swap-multicalibration. That statistic normalizes correlations by bucket mass, which does not control the comparator-dependent quadratic term in (1).

A new finite-class result identifies the missing structure (Okoroafor, 2026). For any fixed Lipschitz proper loss, a Bernstein-corrected approachability potential gives $\widetilde{O}(T^{1/3}(\log|\mathcal{H}|)^{2/3})$. Its discussion asks whether an infinite class can be handled by sequential complexity or a second-order weak learner, and whether the offset objective can be optimized by an oracle. We answer both questions through ordinary online square-loss regression.

Contributions. Our results are as follows.

- We give a black-box reduction from proper-loss hypothesis-swap regret to $N+1$ online square-loss learners. If each learner has concave regret R, the high-probability bound is

$$O\big(L\{(N+1)R(T/(N+1)) + T/N^2 + \log(1/\delta)\}\big).$$

If $R(n) = Cn^\alpha$, the time exponent is $(1+\alpha)/(3-\alpha)$. Thus logarithmic regret retains $T^{1/3}$, while a first-order $\sqrt{n}$ oracle gives $T^{3/5}$. Optimal online nonparametric regression bounds also yield explicit rates from sequential entropy. For every entropy exponent $p > 0$, we construct a fixed convex class with entropy $\Theta(\beta^{-p})$ and prove lower bounds matching both entropy regimes up to logarithmic factors. The crossing search uses $O(\log(N+1))$ oracle predictions per context and only one oracle update.

- For half-Brier loss and bounded d-parameter linear prediction, constrained Online Newton Step gives $\widetilde{O}(T^{1/3}d^{2/3})$. We prove a matching $\Omega(T^{1/3}d^{2/3})$ lower bound for a fixed class with pseudodimension exactly d. This closes the joint time-dimension rate up to logarithms. For bounded RKHS predictors, kernel ridge regression gives a polynomial-time bound in the log determinant of the realized kernel covariance.
- The same reduction applies whenever a proper loss's canonical slope is linear in the model parameter. This includes bounded-score logistic predictors under log loss. A self-normalized online-to-batch argument also gives an oracle-efficient $\widetilde{O}((d/m)^{2/3})$ offline learner.

The statistical offline exponent for infinite classes was already accessible by running the finite algorithm on a uniform cover (Okoroafor, 2026). Offline level-set boosting already used square-loss regression oracles for multicalibration (Globus-Harris et al., 2023), and Luo et al. (2025a) gave a slower oracle-efficient swap-agnostic rate. Our offline contribution is an oracle-efficient conversion that retains the sharp two-thirds exponent. The fast online infinite-class guarantee and the linear and nonparametric minimax rates are new.

2 Setup and Closest Guarantees

Outcomes are binary and contexts lie in an arbitrary set $\mathcal{X}$. A loss $\ell : [0,1] \times \{0,1\} \to \mathbb{R}$ is proper if $q = p$ minimizes $\mathbb{E}_{Y\sim\mathrm{Ber}(p)}\ell(q,Y)$. Its partial losses are L-Lipschitz if $|\ell(p,y) - \ell(q,y)| \le L|p-q|$ for both outcomes. Define

$$\Delta_\ell(q) = \ell(q,1) - \ell(q,0).$$

Properness makes Δ_ℓ nonincreasing. We use the uniform grid $\Gamma_N = \{0, 1/N, \dots, 1\}$ (Gneiting and Raftery, 2007; Reid and Williamson, 2010).

The adversary is nonanticipating. On round t, it chooses and reveals x_t from the past. After the forecaster announces a distribution on Γ_N, the adversary commits y_t without observing the fresh draw. The label is revealed after p_t is drawn. This is the adversary model used by the closest randomized calibration results.

Table 1: Contextual squared-loss swap regret for bounded d-dimensional linear classes. Logarithmic and confidence factors are suppressed.

Method	Realized regret	Efficient
Garg et al. (2024)	$dT^{3/4}$	yes
Luo et al. (2025a)	$T^{3/5}d^{2/5}$	yes
This work	$T^{1/3}d^{2/3}$	yes
This work, lower	$T^{1/3}d^{2/3}$	–

For a selector $\phi : \Gamma_N \to \mathcal{H}$, write

$$\mathrm{SwapReg}_T(\ell,\phi) = \sum_{t=1}^{T}[\ell(p_t,y_t) - \ell(\phi(p_t)(x_t),y_t)]$$

and let $\mathrm{SwapReg}_T(\ell,\mathcal{H}) = \sup_\phi \mathrm{SwapReg}_T(\ell,\phi)$.

Table 1 isolates the comparison most relevant to this paper. The new rate is smaller throughout the natural regime $d < T$, since the ratio of the two upper rates is $(d/T)^{4/15}$ before logarithms. Recent ℓ_2 swap-multicalibration rates are not entries in the table. If B_p is a residual correlation, M_p is a comparator's squared activity, and n_p is bucket mass, those results control $\sum_p B_p^2/n_p$. Contextual swap regret requires $\sum_p(2B_p - M_p)$. The activity M_p can be arbitrarily smaller than n_p. For example, on one bucket with $p = 1-\epsilon$, $y = 1$, and $h = 1$ on every round, $B_p = M_p = n_p\epsilon^2$. Thus $2B_p - M_p = n_p\epsilon^2$, while $B_p^2/n_p = n_p\epsilon^4$.

2.1 Where the prior exponent is lost

The distinction is visible before any new algorithm is introduced. The Blum–Mansour reduction (Blum and Mansour, 2007) used by Garg et al. (2024) assigns every bucket a regression learner and computes a stationary distribution over its proposed next predictions. The refinement of Luo et al. (2025a) combines Online Newton Step with unbiased grid rounding and proves the pseudo contextual-regret bound

$$\mathrm{PSReg}_T = O\left(\frac{T}{N^2} + Nd\log T\right). \tag{2}$$

This already has the desired optimization. The difficulty is that the comparator is selected after the random level sets are realized. Their uniform martingale comparison contributes, up to logarithms,

$$\sqrt{dNT}. \tag{3}$$

Balancing (2) and (3) gives $N \asymp (T/d)^{1/5}$ and regret $\widetilde{O}(T^{3/5}d^{2/5})$.

Our reduction never transfers a pseudo comparator guarantee to realized level sets. It updates the learner indexed by the sampled bucket, so its pathwise regret already holds on that realized subsequence. The only random quantity requiring concentration is the learner's own scalar offset. Its conditional variance is paid by the negative quadratic term in (5). This removes (3) rather than trying to sharpen a uniform bound on it.

2.2 Relation to multicalibration

Swap-agnostic learning was introduced through an equivalence with swap multicalibration and swap omniprediction (Gopalan et al., 2023). That equivalence is qualitative but can lose powers of the target accuracy. In the offline setting, Globus-Harris et al. (2023) characterize multicalibration through a squared-loss swap-like condition and iteratively regress each current level set. Their uniform-convergence analysis does not give the sharp two-thirds swap-agnostic rate proved here. Oracle-efficient online multicalibration first reduced the problem to online regression (Garg et al., 2024). The more recent algorithms of Luo et al. (2025a) and Hu et al. (2026a) attain a $T^{1/3}$ squared swap-multicalibration statistic for linear and general online-learnable classes. These results are not contextual squared-loss regret bounds for the activity-normalization reason above. Conversely, our method fixes one proper loss and does not provide simultaneous omniprediction over a loss family. The two guarantees answer different downstream questions. Farina and Perdomo (2026) give a general reduction from online learning to first-order multicalibration through expected variational inequalities, including RKHS test classes. Their objective controls expected residual correlation without the comparator-dependent negative activity needed for the fast log-determinant hypothesis-swap bound in Corollary 8. Ghuge et al. (2025) obtain direct ℓ_1 multicalibration for uniformly covered infinite classes, plus a slower offline-oracle-efficient rate for binary groups under transductive or separated contexts. This first-order objective has the same activity limitation. Sample-efficient statistical omniprediction over bounded left-continuous proper losses is known via direct ensembling (Gibbs and Tibshirani, 2026). Multiclass omniprediction is also possible for infinite comparator families (Hu et al., 2026d). Both benchmarks choose one comparator for each loss, rather than a prediction-indexed family that may choose a different hypothesis at every forecast value.

Two neighboring notions further clarify the boundary. Context-free swap regret replaces $\phi(p)(x)$ by a scalar map $\sigma(p)$. Luo et al. (2025b) obtain a simultaneous $\widetilde{O}(T^{1/3})$ guarantee for smooth proper losses in that setting, without contextual hypotheses. Decision swap regret permits context-dependent policies (Lu et al., 2025), but its finite-family calibration guarantee scales as $|A|\sqrt{T \log |\mathcal{H}|}$. Neither result supplies a second-order regression oracle for an infinite comparator class.

Large- and continuous-action swap regret is another neighboring notion (Dagan et al., 2024; Fishelson et al., 2025b; Anagnostides et al., 2026; Hu et al., 2026b; Fishelson et al., 2025a). It competes with a transformation of the learner's action. In forecasting applications this is a scalar map $\sigma(p)$, possibly against adaptive convex losses. The map sees p but not x, so it cannot choose a context-dependent hypothesis $h_p(x)$ and does not control (1). An offline decision-theoretic line constructs a testable calibration metric that is actionable for full swap regret after a rounded response (Bairaktari et al., 2026). Its response is likewise context-free.

3 The Offset-Regression Reduction

The proper-loss identity and smoothness of the negative Bayes risk imply (Okoroafor, 2026)

$$\ell(p,y) - \ell(q,y) \le 2L(y-p)g_{p,q} - Lg_{p,q}^2, \qquad g_{p,q} = \frac{\Delta_\ell(p) - \Delta_\ell(q)}{2L}. \tag{4}$$

Every $g_{p,q}$ lies in $[-1,1]$. For a bucket $p_i = i/N$, let

$$\mathcal{G}_i = \{x \mapsto g_{p_i,h(x)} : h \in \mathcal{H}\}.$$

Fix $\eta \in (0, 1/4]$. The elementary identity

$$\eta z g - 2\eta^2 g^2 = -\frac{1}{8}\left[(4\eta g - z)^2 - z^2\right] \tag{5}$$

turns the desired second-order score into square loss.

Definition 1 (Compatible regression oracle)**.** *An oracle for bucket i predicts $a_s \in [-1,1]$ and, after receiving $z_s \in [-1,1]$, incurs $(4\eta a_s - z_s)^2$. After n updates it satisfies*

$$\sum_{s=1}^{n}(4\eta a_s - z_s)^2 - \inf_{g\in\mathcal{G}_i}\sum_{s=1}^{n}(4\eta g(x_s) - z_s)^2 \le R_\eta(n). \tag{6}$$

Its predictions at the endpoint buckets have the corresponding signs, $a_s \ge 0$ at $p_0 = 0$ and $a_s \le 0$ at $p_N = 1$. Prediction queries are side-effect free, so the oracle state advances only when a target is supplied. All oracle instances below have this standard deterministic interface.

The sign condition holds whenever the oracle predicts in the pointwise convex hull of $\mathcal{G}_i$. It holds in particular for exponential aggregation and for constrained convex parametric classes. More generally, it can be enforced by clipping an endpoint prediction at zero. This cannot increase square loss, since $z = y$ is nonnegative at p_0 and $z = y - 1$ is nonpositive at p_N.

Algorithm. Maintain compatible oracles $A_0, \dots, A_N$. Given x_t, query A_0. If $a_{t,0} \le 0$, predict 0. Otherwise query A_N. If $a_{t,N} \ge 0$, predict 1. In the remaining case, maintain indices $l < r$ with $a_{t,l} \ge 0 \ge a_{t,r}$. Query the midpoint and replace l after a nonnegative answer or r after a negative answer. This binary search requires no monotonicity. It terminates with $r = l + 1$. Draw between p_l, p_r so that

$$\mathbb{E}_t[a_{t,I_t}] = 0. \tag{7}$$

After observing y_t, update only A_{I_t} with target $z_t = y_t - p_{I_t}$.

The method makes at most $2 + \lceil \log_2 N \rceil$ oracle predictions and one update per round. It does not construct a distribution over hypotheses or solve a Markov stationarity problem. The logarithmic dependence is query-optimal up to the endpoint checks. Even monotone $\{-1, +1\}$ response arrays can place their unique sign change at any of N adjacent pairs, while each query returns only one bit.

Lemma 2 (Offset control)**.** *For every $\eta \in (0, 1/4]$, with probability at least $1 - \delta$,*

$$\sum_{t=1}^{T} [\eta z_t a_{t,I_t} - 2\eta^2 a_{t,I_t}^2] \le \frac{2T}{N^2} + \log \frac{1}{\delta}. \tag{8}$$

Proof. At an interior crossing, direct calculation gives

$$\mathbb{E}_t[z_t a_{t,I_t}] \le N^{-1} \sqrt{\mathbb{E}_t[a_{t,I_t}^2]}.$$

The endpoint cases have nonpositive left side. The Bernstein inequality

$$e^{\eta z a - 2\eta^2 a^2} \le 1 + \eta z a - \frac{\eta^2 a^2}{4}$$

and Young's inequality make the exponential of the left side of (8), minus $2t/N^2$, a nonnegative supermartingale. Markov's inequality completes the proof. Appendix B gives all conditioning details. □

Theorem 3 (Regression reduction)**.** *Suppose Definition 1 holds pathwise in every bucket. Then, with probability at least $1 - \delta$,*

$$\mathrm{SwapReg}_T(\ell, \mathcal{H}) \le L \sum_{i=0}^{N} R_{1/4}(n_i) + \frac{16LT}{N^2} + 8L \log \frac{1}{\delta}, \tag{9}$$

where n_i is the number of updates to bucket i.

Proof. For every selector ϕ, apply (6) in bucket i to $g_{p_i,\phi(p_i)}$. Identity (5) gives

$$\sum_t [\eta z_t g_t - 2\eta^2 g_t^2] \le \sum_t [\eta z_t a_t - 2\eta^2 a_t^2] + \frac{1}{8} \sum_i R_\eta(n_i).$$

This is simultaneous over all selectors because the oracle regret is pathwise. Set $\eta = 1/4$, apply Lemma 2, multiply by $8L$, and use (4). □

Corollary 4 (Complexity interpolation)**.** *If R_η is nondecreasing and concave, the first term in (9) is at most $L(N+1)R_{1/4}(T/(N+1))$. If $R(n) = Cn^\alpha$ for $0 \le \alpha < 1$, optimizing N gives time exponent*

$$\frac{1+\alpha}{3-\alpha}.$$

For finite $\mathcal{H}$, exponential aggregation under square loss has $R(n) = O(\log |\mathcal{H}|)$ and recovers the finite-class rate of Okoroafor (2026). The reduction is useful when $\mathcal{H}$ is infinite and the oracle has regret much smaller than $\sqrt{n}$.

It also gives a direct sequential-complexity answer. Let $\mathcal{N}_2(\beta, \mathcal{G}_i, n)$ denote the worst-case sequential ℓ_2 covering number of the induced class in bucket i.

Corollary 5 (Nonparametric sequential-entropy rates)**.** *Suppose uniformly over buckets and horizons that $\log \mathcal{N}_2(\beta, \mathcal{G}_i, n) \le C\beta^{-p}$ for $\beta \in (0, 1]$ and some $p > 0$. There is a possibly inefficient forecaster such that, with probability at least $1 - \delta$,*

$$\mathrm{SwapReg}_T(\ell, \mathcal{H}) = \widetilde{O}(T^{q(p)} + \log(1/\delta)).$$

Here

$$q(p) = \begin{cases} (p+1)/(p+3), & 0 < p \le 2, \\ (2p-1)/(2p+1), & p \ge 2. \end{cases}$$

where the hidden constants depend on L, C, and p.

Indeed, optimal online nonparametric square-loss forecasters have regret $\widetilde{O}(n^{p/(p+2)})$ for $p \le 2$ and $\widetilde{O}(n^{1-1/p})$ for $p \ge 2$ (Rakhlin and Sridharan, 2014). A separate doubling schedule in each bucket makes these horizon-dependent strategies anytime without changing the exponents. Endpoint clipping makes them compatible. Substitution in Corollary 4 gives the result. The generic forecaster need not be tractable. The next results give efficient instances at the fast end of this spectrum.

Both entropy rates are unimprovable up to logarithmic factors even for half-Brier loss. The classes in the next theorem are fixed independently of the horizon.

Theorem 6 (Sequential-entropy minimax lower bound)**.** *For every $p > 0$, there exist a fixed domain, a*

fixed convex class $\mathcal{H}_p \subseteq [0,1]^{\mathcal{X}}$, and constants depending only on p such that the induced Brier offset classes satisfy

$$\log \mathcal{N}_2(\beta, \mathcal{G}_i, n) = \Theta_p(\beta^{-p})$$

whenever β is sufficiently small and $n \geq C_p \beta^{-p}$. The upper bound holds for every n. Moreover, for all sufficiently large T, every randomized forecaster admits a nonanticipating adversary satisfying

$$\mathbb{E}\,\mathrm{SwapReg}_T(\ell_B, \mathcal{H}_p) \geq \begin{cases} c_p T^{(p+1)/(p+3)}, \\ \dfrac{c_p T^{(2p-1)/(2p+1)}}{(\log T)^{2/(2p+1)}} \end{cases}$$

for $0 < p \leq 2$ and $p > 2$, respectively.

Proof sketch. Put $a_j = j^{-1/p}$ and $\mathcal{H}_p = \{(j,q,s) \mapsto q + s a_j w_j : w \in [-1,1]^{\mathbb{N}}\}$ on $\mathbb{N} \times [1/4, 3/4] \times [0, 1/8]$. Retaining $m = O_p(\beta^{-p})$ coordinates, zeroing the tail, and quantizing coordinate j at resolution proportional to β/a_j gives a static uniform cover of logarithmic size $O_p(\beta^{-p})$. Conversely, a constant-distance sign code on the first $m = \Theta_p(\beta^{-p})$ axes gives a sequential ℓ_2 packing of exponential size.

The scale-sensitive three-hypothesis lower bound derived from Hu et al. (2026c) by Okoroafor (2026) uses perturbation $s_n = K n^{-1/3}$ and forces $cn^{1/3}$ regret in an n-round block. Take $d \asymp T^{p/(p+3)}$ coordinates and $n = \lfloor T/d \rfloor$. In block j, feed context scale s_n/a_j. A sufficiently small constant in d ensures $s_n/a_j \leq 1/8$ for every active axis. A swap selector chooses all coordinate signs independently, so summing the block lower bounds gives $cdn^{1/3} = \Omega_p(T^{(p+1)/(p+3)})$. Appendix G gives the constants, depth-uniform packing, and conditioning details.

For $p > 2$, use a fixed multiscale gated class. At scale m, place $K_m = 2^m$ anchors a distance $\Theta(K_m^{-1})$ apart. Each anchor gates an independent polynomial-axis class with radius $b_m = 1/[16(m+1)^{1/p}]$, then take the convex hull. At scale β, its active gate supports and gatewise axis nets each have logarithmic size $O_p(\beta^{-p})$. At horizon T, choose $K = 2^{\lfloor \log_2(T \log T)/(2p+1) \rfloor}$ blocks and perturb each of their Bernoulli means by an independent sign. A selector maps the realized forecast to its nearest anchor. This choice balances its perturbation gain and the squared anchor spacing at order K^{-2}. Summing over $T - O(K)$ rounds gives $\Omega_p(T^{(2p-1)/(2p+1)}/(\log T)^{2/(2p+1)})$. □

4 Linear, Kernel, and Canonical-Link Models

For half-Brier loss $\ell_B(p,y) = (p-y)^2/2$ (Brier, 1950), a sharper normalization is available:

$$\ell_B(p,y) - \ell_B(q,y) = (y-p)(q-p) - \frac{1}{2}(q-p)^2. \quad (10)$$

Use $u_{p,q} = q - p$ in place of the generic g. At $\eta = 1/4$, the oracle loss is exactly ordinary outcome regression,

$$[(q-p) - (y-p)]^2 = (q-y)^2.$$

The proof of Theorem 3 now multiplies the offset score by 4 rather than $8L$.

Let $\mathcal{W} \subset \mathbb{R}^d$ be compact and convex, let $\|x\|_2 \leq 1$, and assume $h_w(x) = \langle w, x \rangle \in [0,1]$ for every $w \in \mathcal{W}$ and every context. Also assume $\sup_{w \in \mathcal{W}} \|w\|_2^2 \leq C_W d$ for a dimension-free constant C_W, and that quadratic projections onto $\mathcal{W}$ are efficient. Affine predictors are included by augmenting x. In bucket p, run constrained Online Newton Step on

$$f_t(w) = [4\eta(h_w(x_t) - p) - (y_t - p)]^2$$

with initial matrix $A_0 = I/d$. An exact quadratic expansion gives curvature independent of the Euclidean diameter of $\mathcal{W}$. The ONS potential then gives $R_\eta(n) = O(d \log(1+n))$ (Hazan et al., 2007). The precise log-determinant argument is given below.

Corollary 7 (Fast linear contextual swap regret)**.** *Under the preceding assumptions, there is a polynomial-time forecaster such that, with probability at least $1-\delta$,*

$$\mathrm{SwapReg}_T(\ell_B, \mathcal{H}) = O\left(T^{1/3}(d \log T)^{2/3} + \log \frac{1}{\delta}\right). \quad (11)$$

With dense linear algebra, each round uses $O(\log(N+1))$ inner products, one $O(d^2)$ Newton update, and one metric projection, while storing $N+1$ $d \times d$ Newton matrices.

The reduction also supports genuinely infinite-dimensional predictors. Let $\mathcal{F}$ be an RKHS with feature map ψ satisfying $\|\psi(x)\|_{\mathcal{F}} \leq \kappa$, and let

$$\mathcal{H}_B = \{x \mapsto \langle w, \psi(x) \rangle_{\mathcal{F}} : \|w\|_{\mathcal{F}} \leq B,\ h_w(\mathcal{X}) \subseteq [0,1]\}.$$

Write $C_T = \sum_{t=1}^T \psi(x_t) \otimes \psi(x_t)$. Determinants below are taken on the finite span of the realized features. For $\alpha > 0$, define

$$\Gamma_T(\alpha) = \log \det(I + C_T/\alpha).$$

Corollary 8 (Kernel and spectrum-adaptive rate)**.** *For every $\lambda > 0$ and fixed grid size N, there is an exact kernel forecaster such that, with probability at least $1-\delta$,*

$$\mathrm{SwapReg}_T(\ell_B, \mathcal{H}_B) \leq \frac{N+1}{2}[\lambda B^2 + \Gamma_T(\lambda(N+1))] + \frac{8T}{N^2} + 4 \log \frac{1}{\delta}. \quad (12)$$

Define $A_T(\lambda) = 1 + \lambda B^2 + \Gamma_T(\lambda)$. Given a deterministic bound $A \geq A_T(\lambda)$ with $A \leq T$, choosing

$N \asymp (T/A)^{1/3}$ *gives*

$$\mathrm{SwapReg}_T(\ell_B, \mathcal{H}_B) = O\left(T^{1/3}A^{2/3} + \log\frac{1}{\delta}\right).$$

If C_T *has rank* $r \geq 1$*, then*

$$\Gamma_T(\lambda(N+1)) \leq r\log\left(1 + \frac{\kappa^2 T}{\lambda r(N+1)}\right).$$

The algorithm runs clipped online kernel ridge regression independently in each bucket. It needs no finite feature representation. An exact Gram-matrix implementation uses polynomial time and space. The histories partition the past, so all $O(\log(N+1))$ queried buckets together require at most $O(t)$ kernel evaluations on round t, while the update of the sampled bucket takes at most $O(t^2)$ arithmetic operations. Classical online ridge and kernel second-order bounds also exhibit this log-determinant complexity (Vovk, 1997; Calandriello et al., 2017). For rank-d features with $\kappa \leq 1$, $B^2 = O(d)$, and $\lambda = 1$, the optimized bound recovers (11). Theorem 9 shows that its polynomial dependence on T and d is worst-case sharp, while (12) retains the finer realized spectrum.

The horizon-dependent choice of N is not essential. A doubling schedule restarts all regressors at horizons $1, 2, 4, \ldots$ and assigns confidence proportional to the inverse square of the epoch index. Since every exponent in Corollary 4 is below one, the epoch regrets form a geometric series of the same order. The result is an anytime forecaster with only lower-order logarithmic confidence terms.

The construction extends beyond Brier loss. Let $r_\ell(q) = -\Delta_\ell(q)$ be the canonical slope of a proper loss and restrict predictions to an interval I on which the partial losses are L-Lipschitz. For

$$\mathcal{H} = \{h_w : r_\ell(h_w(x)) = \langle w, x\rangle,\ w \in \mathcal{W}\},$$

the generic test is affine in w:

$$g_{p,h_w}(x) = \frac{\langle w, x\rangle - r_\ell(p)}{2L}.$$

ONS and Theorem 3 give the right side of (11) multiplied by L. For log loss on $I = [a, 1-a]$, $r_\ell(q) = \log(q/(1-q))$ and $L = 1/a$. Thus bounded-score logistic predictors are a direct instance. Appendix D states the parameter and inverse-link conditions precisely.

5 Matching Lower Bounds

The $d^{2/3}$ dependence in Corollary 7 is necessary. The proof tensorizes the fixed three-hypothesis lower bound of Okoroafor (2026), which itself follows from the sharp recalibration construction of Hu et al. (2026c).

Theorem 9 (Joint minimax lower bound)**.** *For every* d *there is a fixed domain and a fixed convex class* $\mathcal{H}_d$ *of half-Brier predictors with* $\mathrm{Pdim}(\mathcal{H}_d) = d$ *such that, for all* $T \geq c_0 d$*,*

$$\inf_{\text{forecasters}} \sup_{\text{nonanticipating adversaries}} \mathbb{E}\,\mathrm{SwapReg}_T(\ell_B, \mathcal{H}_d) \geq cT^{1/3}d^{2/3}. \tag{13}$$

Proof. Let the domain contain triples (j, q, s) with $j \in [d]$, $q \in [1/4, 3/4]$, and $s \in [0, 1/8]$. Define the fixed class

$$\mathcal{H}_d = \{h_w(j, q, s) = q + sw_j : w \in [-1, 1]^d\}.$$

It has pseudodimension exactly d. Split the horizon into d blocks of $n = \lfloor T/d \rfloor$ rounds. In block j, run an independent copy of the fixed three-hypothesis lower-bound process with hypotheses $q, q+s, q-s$.

For every realized prediction value p, a selector from $\mathcal{H}_d$ chooses a full vector $w(p)$. Its jth coordinate can select $0, +1$, or -1 independently of every other block. Hence global swap regret is at least the sum of the d within-block swap regrets. At the start of a block, condition on the public history. The global forecaster restricts to an arbitrary randomized one-dimensional forecaster, so invoke the one-dimensional hard process with fresh randomness. This is a nonanticipating continuation and gives conditional expected regret $cn^{1/3}$ per block. Summing yields $cd\lfloor T/d \rfloor^{1/3} = \Omega(T^{1/3}d^{2/3})$. On leftover rounds, set $q = 1/2$ and $s = 0$, then choose the label with nonnegative conditional expected excess. Thus the remainder does not reduce expected regret. □

The scale s used by the hard process may depend on the block length, but the class does not depend on T because s is part of the fixed domain. The full pseudodimension and conditioning arguments appear in Appendix F. This is also a bounded affine linear class satisfying the upper assumptions, with d free coefficients, $d+1$ ambient coordinates, unit-norm features, and squared radius $O(d)$.

6 Oracle-Efficient Offline Learning

Let $Z_1, \ldots, Z_m$ be i.i.d. from D. Run the online algorithm at $\eta = 1/10$ and store its pre-round prediction kernels $\pi_1, \ldots, \pi_m$. The historical predictor draws J uniformly from $[m]$ and, on input x, draws $p \sim \pi_J(x)$. A direct implementation stores the m historical states.

Let $\mathcal{N}_\infty(\rho, \mathcal{H})$ be the uniform prediction-space covering number. Let $\Phi_{\mathcal{H}}$ be the measurable selectors

from $[0,1]$ to $\mathcal{H}$. Define population swap excess, with $(X,Y) \sim D$ and $p \sim \widehat{p}(X)$, by

$$\mathrm{SwapAgn}_D(\ell, \mathcal{H}) = \sup_{\phi \in \Phi_{\mathcal{H}}} \mathbb{E}\big[\ell(p,Y) - \ell(\phi(p)(X),Y)\big].$$

Theorem 10 (Historical-predictor conversion)**.** *Suppose the oracle regret is nondecreasing and concave. There is a universal C such that, with probability at least $1-\delta$,*

$$\mathrm{SwapAgn}_D(\ell, \mathcal{H}) \le CL\Bigg[\frac{(N+1)R_{1/10}(m/(N+1))}{m} + \frac{1}{N^2} + \frac{(N+1)\log \mathcal{N}_\infty(\rho, \mathcal{H})}{m} + \frac{\log(1/\delta)}{m} + \rho\Bigg]. \tag{14}$$

The proof uses a finite cover only for population transfer. The learner still optimizes over the original infinite class. For a fixed selector test g, Freedman's inequality transfers realized bias and mass to their conditional population values. It gives

$$\mathrm{Bias}_D(g) \le 5\eta\, \mathrm{Mass}_D(g) + \frac{C\{\sum_i R_\eta(n_i) + m/N^2 + \lambda\}}{\eta m},$$

where $\lambda = (N+1)\log \mathcal{N}_\infty(\rho, \mathcal{H}) + \log(1/\delta)$. At $\eta = 1/10$, the first term cancels the negative mass in (4). Appendix E gives the proof.

For bounded linear predictors, $\log \mathcal{N}_\infty(\rho, \mathcal{H}) = O(d \log(B/\rho))$. Taking $\rho = 1/m$ and optimizing N gives the following.

Corollary 11 (Offline linear Brier rate)**.** *The historical predictor based on constrained Online Newton Step satisfies*

$$\mathrm{SwapAgn}_D(\ell_B, \mathcal{H}) = O\Bigg(\left(\frac{d\log(dm)}{m}\right)^{2/3} + \frac{\log(1/\delta)}{m}\Bigg)$$

with probability at least $1-\delta$.

The exponent and dimension dependence are also statistically sharp. The starting point is the batch multicalibration lower bound of Collina et al. (2026). Their Fano construction uses $k = \mathrm{polylog}(1/\epsilon)$ binary groups and requires $\widetilde{\Omega}(\epsilon^{-3})$ samples for mean ECE multicalibration. By the squared-loss equivalence of Gopalan et al. (2023), multicalibration error ϵ witnesses swap-agnostic squared-loss excess at least ϵ^2 against a bounded linear span of those groups.

Theorem 12 (Offline joint lower bound)**.** *For every $C\log^4 m \le d \le m/c_0$, there are a finite domain and a convex class of $[0,1]$-valued linear predictors with pseudodimension at most d, unit-norm features, and squared parameter radius $O(d)$ such that every possibly randomized learner outputting a $[0,1]$-valued predictor from m i.i.d. samples satisfies*

$$\sup_D \mathbb{E}\,\mathrm{SwapAgn}_D(\ell_B, \mathcal{H}_d) \ge \widetilde{\Omega}\left(\left(\frac{d}{m}\right)^{2/3}\right). \tag{15}$$

Proof sketch. The finite Fano family in Collina et al. (2026), together with the contrapositive direction of the equivalence in Gopalan et al. (2023), gives a prior for which any learner using n samples has expected half-Brier swap excess $\widetilde{\Omega}(n^{-2/3})$ against a $k = \mathrm{polylog}(n)$ dimensional bounded linear class. Take b independent copies on disjoint context blocks and choose the block uniformly. The comparator parameter is a product, so swap excess is the average of the block excesses. A sample contains at most $2m/b$ observations from any fixed block with probability at least $1/2$. Conditional on all other blocks, those observations are the only information about its independent Fano parameter. The expected excess is therefore $\widetilde{\Omega}((b/m)^{2/3})$. Fix a common parameter budget $\bar{k} = C\log^4 m$ that dominates every base instance used in the construction, pad unused coordinates, and take $b = \lfloor d/\bar{k} \rfloor$. Absorbing $\bar{k}$ into the polylogarithmic factor gives the result. Appendix H gives the product-prior and range details. □

7 Discussion and Limitations

The reduction shows that the relevant primitive is not ordinary residual correlation. It is offset regression, where disagreement with a comparator supplies the variance correction needed to realize random prediction levels. This explains why logarithmic square-loss regret changes the contextual swap rate while first-order multicalibration bounds do not. The countable-axis and multiscale-gate lower bounds further show that both complexity-interpolation exponents are intrinsic, rather than artifacts of bucketing the reduction. For $p > 2$, the swap exponent strictly exceeds the ordinary online-regression exponent $1 - 1/p$ (Rakhlin and Sridharan, 2014). The gain comes from letting each realized forecast level select a different gate.

There are three limitations. First, the adversary cannot condition the current label on the fresh grid draw. Second, the online algorithm stores $N+1$ regressor states, although it queries only logarithmically many and updates one, and the direct historical conversion stores m states. Third, the generic proper-loss theorem requires a square-loss oracle for the induced canonical-slope class. Efficient oracles need not exist

for arbitrary nonlinear hypothesis classes. Multiclass outcomes for the hypothesis-swap benchmark also remain open.

A Proper-Loss and Offset Details

For $p, q \in [0,1]$, let $L_p(q) = \mathbb{E}_{Y\sim\mathrm{Ber}(p)}\ell(q,Y)$. The canonical representation is

$$\ell(q,y) = L_q(q) + (y-q)\Delta_\ell(q).$$

Subtracting the representations at p and q gives

$$\ell(p,y) - \ell(q,y) = (y-p)(\Delta_\ell(p) - \Delta_\ell(q)) - D_{F_\ell}(p,q), \tag{16}$$

where $F_\ell(p) = -L_p(p)$ is convex. If the partial losses are L-Lipschitz, then Δ_ℓ is continuous and $-\Delta_\ell(q) \in \partial F_\ell(q)$. Hence F_ℓ is differentiable and $F_\ell' = -\Delta_\ell$ is $2L$-Lipschitz. The standard co-coercivity inequality in one dimension gives

$$D_{F_\ell}(p,q) \geq \frac{(\Delta_\ell(p) - \Delta_\ell(q))^2}{4L}.$$

Substitution in (16) proves (4).

For completeness, expanding the two square losses gives

$$\begin{aligned}&(4\eta a - z)^2 - (4\eta g - z)^2\\ &\quad = 8\{(\eta z g - 2\eta^2 g^2) - (\eta z a - 2\eta^2 a^2)\},\end{aligned}$$

which proves (5) and the oracle comparison used in Theorem 3.

B The Crossing Oracle and Supermartingale

Write $\rho = 1/N$. In the interior case, let $a \geq 0 \geq b$ be the two oracle predictions at adjacent grid values p and $p+\rho$. Put probability $\lambda = a/(a-b)$ on the right endpoint and $1-\lambda$ on the left. Then $(1-\lambda)a + \lambda b = 0$. For any label y committed before the draw,

$$\begin{aligned}&\mathbb{E}[(y - p_I)a_I]\\ &= (1-\lambda)(y-p)a + \lambda(y-p-\rho)b\\ &= \rho(1-\lambda)a \leq \rho\sqrt{(1-\lambda)a^2 + \lambda b^2}.\end{aligned}$$

At $p = 0$ with $a \leq 0$, the product ya is nonpositive. At $p = 1$ with $a \geq 0$, the product $(y-1)a$ is nonpositive.

We use the following scalar inequality from the Bernstein potential of Okoroafor (2026). For $\eta \leq 1/4$, $|a| \leq 1$, and $|z| \leq 1$,

$$\exp(\eta z a - 2\eta^2 a^2) \leq 1 + \eta z a - \frac{\eta^2 a^2}{4}. \tag{17}$$

Condition on the past, the current context, the announced grid mixture, and the nonanticipating label, but not the fresh draw. The preceding crossing bound and (17) imply

$$\begin{aligned}\mathbb{E}_t e^{\eta z_t a_t - 2\eta^2 a_t^2} &\leq 1 + \eta\rho\sqrt{\mathbb{E}_t a_t^2} - \frac{\eta^2}{4}\mathbb{E}_t a_t^2\\ &\leq e^{2\rho^2}.\end{aligned}$$

The last line uses $\eta\rho\sqrt{A} \leq \eta^2 A/8 + 2\rho^2$. Thus

$$\exp\left(\sum_{s=1}^{t}[\eta z_s a_s - 2\eta^2 a_s^2] - 2t\rho^2\right)$$

is a nonnegative supermartingale. Markov's inequality proves Lemma 2.

C Expanded Proof of the Online Bound

Fix a selector ϕ and abbreviate $g_t = g_{p_t,\phi(p_t)}(x_t)$. In bucket i, the oracle guarantee and the identity in Appendix A give

$$\sum_{t:p_t=p_i}[\eta z_t g_t - 2\eta^2 g_t^2] \leq \sum_{t:p_t=p_i}[\eta z_t a_t - 2\eta^2 a_t^2] + \frac{1}{8}R_\eta(n_i).$$

Sum over buckets and invoke Lemma 2. At $\eta = 1/4$,

$$2Lz_t g_t - Lg_t^2 = 8L[\eta z_t g_t - 2\eta^2 g_t^2].$$

Equation (4) proves (9). No union bound over selectors appears because (6) holds against every comparator on the realized bucket subsequence.

If R is concave, Jensen's inequality gives

$$\sum_{i=0}^{N} R(n_i) \leq (N+1)R\left(\frac{T}{N+1}\right).$$

For $R(n) = Cn^\alpha$, the two terms depending on N are of order $CT^\alpha N^{1-\alpha}$ and TN^{-2}. Equating them yields $N \asymp (T^{1-\alpha}/C)^{1/(3-\alpha)}$ and exponent $(1+\alpha)/(3-\alpha)$.

For Brier loss, put $u_t = \phi(p_t)(x_t) - p_t$ and use (10). The same calculation at $\eta = 1/4$ gives the sharper bound

$$\mathrm{SwapReg}_T(\ell_B, \mathcal{H}) \leq \frac{1}{2}\sum_i R_{1/4}(n_i) + \frac{8T}{N^2} + 4\log\frac{1}{\delta}.$$

D Linear, Kernel, and Canonical-Link Oracles

For Brier loss in bucket p_i, let

$$s_t(w) = 4\eta(\langle w, x_t\rangle - p_i), \qquad z_t = y_t - p_i,$$

and set $f_t(w) = [s_t(w) - z_t]^2$. If w_t is the current iterate and $g_t = \nabla f_t(w_t)$, exact expansion gives, for every $w \in \mathcal{W}$,

$$f_t(w_t) - f_t(w) = g_t^\top (w_t - w) - [s_t(w_t) - s_t(w)]^2.$$

Both $s_t(w_t)$ and z_t lie in $[-1, 1]$. Hence

$$f_t(w_t) - f_t(w) \le g_t^\top (w_t - w) - \frac{[g_t^\top (w_t - w)]^2}{16}.$$

Run projected Online Newton Step with curvature parameter $1/8$, initial matrix $A_0 = I/d$, and $A_t = A_{t-1} + g_t g_t^\top$. Its update is

$$w_{t+1} = \Pi_{\mathcal{W}}^{A_t}(w_t - 8A_t^{-1} g_t).$$

The metric projection inequality and the preceding curvature bound telescope to

$$\sum_t [f_t(w_t) - f_t(w)] \le \frac{1}{16}\|w_1 - w\|_{A_0}^2 + 4\sum_t g_t^\top A_t^{-1} g_t.$$

The determinant lemma bounds the final sum by $\log(\det A_{n_i} / \det A_0)$ (Hazan et al., 2007). Therefore

$$\begin{aligned} R_{\eta,i}(n_i) &\le C_W + C \log \frac{\det A_{n_i}}{\det A_0} \\ &\le C_W + C \log\det\left(I + 16d \sum_{t \in S_i} x_t x_t^\top\right). \end{aligned}$$

The first line uses $\|w - w_1\|_2^2/d \le 4C_W$, and the second uses $g_t g_t^\top \preceq 16 x_t x_t^\top$. The regret is therefore $O(d \log(1 + n_i))$. Matrix updates take $O(d^2)$ time, in addition to the assumed quadratic projection. Since $w_t \in \mathcal{W}$, the oracle prediction $a_{t,i} = \langle w_t, x_t\rangle - p_i$ has the required endpoint signs.

We next prove Corollary 8. Fix one bucket and index its selected subsequence by s. Put $\psi_s = \psi(x_s)$ and define

$$A_{s-1} = \lambda I + \sum_{u<s} \psi_u \otimes \psi_u, \qquad b_{s-1} = \sum_{u<s} y_u \psi_u.$$

Online ridge forms $m_s = \langle A_{s-1}^{-1} b_{s-1}, \psi_s\rangle$ and predicts $q_s = \Pi_{[0,1]}(m_s)$. Let $d_s = \langle \psi_s, A_{s-1}^{-1}\psi_s\rangle$. Completing the square after each rank-one update gives the exact sequential identity

$$\sum_s \frac{(y_s - m_s)^2}{1 + d_s} = \min_w \left\{ \lambda \|w\|_{\mathcal{F}}^2 + \sum_s (y_s - \langle w, \psi_s\rangle)^2 \right\}. \tag{18}$$

Indeed, before update s the regularized objective equals its minimum plus $\|w - A_{s-1}^{-1} b_{s-1}\|_{A_{s-1}}^2$. Minimizing after adding the new square increases its minimum by $(y_s - m_s)^2/(1 + d_s)$, which proves (18) by induction.

Write $e_s = y_s - q_s$. Projection toward $[0, 1]$ cannot increase loss because $y_s \in [0, 1]$, and $e_s^2 \le 1$. Therefore

$$e_s^2 = \frac{e_s^2}{1 + d_s} + \frac{d_s e_s^2}{1 + d_s} \le \frac{(y_s - m_s)^2}{1 + d_s} + \frac{d_s}{1 + d_s}.$$

The determinant lemma and $d/(1 + d) \le \log(1 + d)$ imply, for the bucket covariance C_i,

$$\sum_s \frac{d_s}{1 + d_s} \le \log\det(I + C_i/\lambda).$$

Comparing the minimum in (18) with any $h_w \in \mathcal{H}_B$ proves the pathwise bucket regret bound

$$R_i \le \lambda B^2 + \log\det(I + C_i/\lambda).$$

Clipping also ensures $q_s - p_i \ge 0$ at $p_i = 0$ and $q_s - p_i \le 0$ at $p_i = 1$, so the crossing oracle remains valid. Finally, $\sum_i C_i = C_T$ and concavity of $C \mapsto \log\det(I + C/\lambda)$ gives

$$\sum_{i=0}^{N} \log\det(I + C_i/\lambda) \le (N + 1) \cdot \log\det\left(I + \frac{C_T}{\lambda(N + 1)}\right).$$

Substitution into the sharper Brier bound proves (12). The rank statement follows from $\mathrm{tr}(C_T) \le \kappa^2 T$ and the arithmetic-geometric mean inequality applied to its nonzero eigenvalues.

For a general proper loss on $I = [a, b]$, use a grid of width $(b - a)/N$ and replace the endpoint buckets by a, b. If $r_\ell(h_w(x)) = \langle w, x\rangle$, then the oracle prediction

$$a_{t,i} = \frac{\langle w_t, x_t\rangle - r_\ell(p_i)}{2L}$$

lies in the convex range of the generic test class and has the endpoint signs. The same exact quadratic and ONS argument gives $O(d \log(1 + n))$ regret under the corresponding radius and efficient-projection assumptions. For the offline cover statement, also assume the inverse of r_ℓ is K-Lipschitz on $r_\ell(I)$. A Euclidean ρ/K parameter net then induces a uniform ρ prediction-space cover.

E Offline Transfer

Fix a selector from a finite uniform cover and write g_t for its test on training round t. Let

$$Z_t = (Y_t - p_t) g_t, \qquad Q_t = g_t^2,$$

and let $\mu_t = \mathbb{E}_t Z_t$, $\nu_t = \mathbb{E}_t Q_t$. The conditional expectation is over a fresh i.i.d. example and the online

algorithm's fresh prediction draw. If $V = \sum_t \nu_t$, fixed-time Freedman inequalities give, with probability at least $1 - 2e^{-\lambda}$,

$$\sum_t Q_t \le 2V + 2\lambda, \tag{19}$$

$$\sum_t \mu_t \le \sum_t Z_t + \sqrt{2V\lambda} + \frac{4\lambda}{3}. \tag{20}$$

The first bound uses $Q_t \in [0,1]$. For the second, note that $\mathbb{E}_t[(Z_t - \mu_t)^2] \le \nu_t$ and rescale the martingale difference by two (Cesa-Bianchi and Lugosi, 2006).

The online oracle and supermartingale bounds imply

$$\eta \sum_t Z_t - 2\eta^2 \sum_t Q_t \le B,$$
$$B = \frac{1}{8}\sum_i R_\eta(n_i) + \frac{2m}{N^2} + \log\frac{3}{\delta}.$$

Combining this display with (19) and (20), then using $\sqrt{2V\lambda} \le \eta V + \lambda/(2\eta)$, gives

$$\frac{1}{m}\sum_t \mu_t \le 5\eta\frac{V}{m} + \frac{B + C\lambda}{\eta m}. \tag{21}$$

The historical mixture identities are

$$\mathrm{Bias}_D(g, \widehat{p}) = \frac{1}{m}\sum_t \mu_t, \qquad \mathrm{Mass}_D(g, \widehat{p}) = \frac{V}{m}.$$

Let $K = \mathcal{N}_\infty(\rho, \mathcal{H})$. There are K^{N+1} selectors from the cover. Set $\lambda = (N+1)\log K + \log(6/\delta)$ and union bound (21). At $\eta = 1/10$, multiplying the bias by $2L$ cancels L times the population mass in (4). Finally, replace an arbitrary selector hypothesis in each bucket by its cover element. The L-Lipschitz loss changes by at most $L\rho$. Jensen's inequality for the oracle regrets proves Theorem 10.

For Brier loss, repeat the proof with $u = h(x) - p$ and (10). The cancellation at $\eta = 1/10$ is exact and the same rate follows with a smaller universal constant.

F Lower-Bound Details

We first recall the black-box consequence proved by Okoroafor (2026) from Hu et al. (2026c). There is a fixed domain $\mathcal{X}_\star = [1/4, 3/4] \times [0, 1/8]$ and the three fixed hypotheses

$$h_0(q,s) = q, \quad h_+(q,s) = q + s, \quad h_-(q,s) = q - s$$

such that every randomized forecaster admits a nonanticipating distribution with expected half-Brier swap regret at least $cn^{1/3}$ over n rounds. The statement applies after conditioning on any initial public history because the learner's continuation is still an arbitrary randomized forecaster.

For the direct sum, use the domain $\mathcal{X}_d = [d] \times \mathcal{X}_\star$ and

$$\mathcal{H}_d = \{(j,q,s) \mapsto q + sw_j : w \in [-1,1]^d\}.$$

This class is an affine translate of homogeneous linear functions with features se_j, so $\mathrm{Pdim}(\mathcal{H}_d) \le d$. For the reverse inequality, fix $s > 0$, take one point with $q = 1/2$ for each j, and use threshold $1/2$. Choosing the signs of $w_1, \dots, w_d$ realizes every dichotomy. Hence $\mathrm{Pdim}(\mathcal{H}_d) = d$.

Partition the first dn rounds into d contiguous blocks. In block j, present contexts (j, q_t, s_t) from an independent copy of the one-dimensional hard process. A selector $p \mapsto h_{w(p)}$ chooses every coordinate $w_j(p)$ independently. Restricting each coordinate to $\{0, +1, -1\}$ shows pointwise that

$$\mathrm{SwapReg}_{dn}(\ell_B, \mathcal{H}_d) \ge \sum_{j=1}^{d} \mathrm{SwapReg}_n^{(j)}$$
$$(\ell_B, \{h_0, h_+, h_-\}).$$

At the start of block j, condition on the public history. The learner's restriction to the block is a valid randomized forecaster for the one-dimensional lower bound. Invoke its hard process with fresh randomness. This continuation is nonanticipating, and taking conditional expectations and summing gives $cdn^{1/3}$. If T is not a multiple of d, ignore the remaining rounds in the block construction and take $n = \lfloor T/d \rfloor$. On each remaining round, present $(j, 1/2, 0)$. All comparators then predict $1/2$. Of the labels zero and one, at least one makes the conditional expected half-Brier excess nonnegative because their average is $\mathbb{E}_t[(p_t - 1/2)^2]/2$. Choosing that label before the fresh prediction draw preserves the lower bound. This proves Theorem 9.

G Sequential-Entropy Lower-Bound Details

G.1 Low-Entropy Regime

We first extract the scale in the one-dimensional lower bound. The recalibration theorem of Hu et al. (2026c), in the black-box form used by Okoroafor (2026), gives constants $c_0, \epsilon_0 > 0$ such that its hard process exists whenever $n \le c_0\epsilon^{-3}$ and $\epsilon \le \epsilon_0$. In the proof of the fixed three-hypothesis reduction, take

$$\epsilon = (c_0/n)^{1/3}, \qquad s_n = \epsilon/2.$$

If expected unscaled square-loss swap regret were at most $n\epsilon^2/64$, that proof would give both expected calibration error at most ϵ and expected external excess

at most ϵ^2, a contradiction. Thus, for all sufficiently large n, the hypotheses $q, q+s_n, q-s_n$ force expected half-Brier swap regret at least $n\epsilon^2/128 = cn^{1/3}$. In particular,

$$s_n = Kn^{-1/3} \tag{22}$$

for a universal $K > 0$.

Fix $p > 0$, set $a_j = j^{-1/p}$, and take

$$\begin{aligned} \mathcal{X}_p &= \mathbb{N} \times [1/4, 3/4] \times [0, 1/8], \\ \mathcal{H}_p &= \{h_w(j, q, s) = q + sa_j w_j : w \in [-1, 1]^{\mathbb{N}}\}. \end{aligned} \tag{23}$$

This is a fixed convex class with range in $[1/8, 7/8]$. We next prove its entropy claim. Constants from the half-Brier offset normalization can be absorbed into β, so it is enough to cover $\mathcal{H}_p$ itself. Given $\beta > 0$, retain the first $m = \lfloor (4\beta)^{-p} \rfloor$ coordinates and set the tail coordinates to zero. The tail changes a prediction by at most $\beta/2$. On coordinate $j \le m$, use a scalar net with coordinate accuracy $4\beta/a_j$. This also changes a prediction by at most $\beta/2$ and gives

$$\begin{aligned} \log |\mathcal{V}_\beta| &\le \sum_{j=1}^{m} \log\left(2 + \frac{a_j}{4\beta}\right) \\ &\le C_p m = O_p(\beta^{-p}). \end{aligned}$$

The second inequality follows by comparing the normalized sum with the integral of $\log(2 + x^{-1/p})$ on $(0, 1]$. This deterministic uniform cover is also a sequential ℓ_2 cover on every context tree.

For the reverse bound, take $m = \lfloor (c/\beta)^p \rfloor$ for a sufficiently small constant $c > 0$. Cycle the contexts $(j, 1/2, 1/8)$, $j \in [m]$, to fill any horizon $n \ge m$. A constant-distance binary code supplies $\exp(c_p m)$ vectors in $\{-1, +1\}^m$ with pairwise Hamming distance at least $m/4$. Each context occurs at least $\lfloor n/m \rfloor \ge n/(2m)$ times, so their empirical ℓ_2 prediction distance is at least a universal constant times a_m. Choosing c small enough makes this exceed 2β even after the half-Brier offset normalization. Hence every sequential β-cover on this tree has logarithmic size $\Omega_p(\beta^{-p})$. Translation and constant scaling in the Brier offset class preserve both entropy estimates.

It remains to prove the regret lower bound. Set

$$d = \lfloor \gamma_p T^{p/(p+3)} \rfloor, \qquad n = \lfloor T/d \rfloor,$$

where $\gamma_p > 0$ is a sufficiently small constant. For each $j \le d$, run one n-round hard process from (22). When that process presents (q_t, s_n), present the context $(j, q_t, s_n/a_j)$ to the global learner. Since $a_j \ge a_d$ and

$$\frac{s_n}{a_d} \le K(2d/T)^{1/3} d^{1/p} \le \frac{1}{8}$$

for small enough γ_p, every context belongs to $\mathcal{X}_p$. Choosing coordinate w_j equal to $0, +1$, or -1 exactly realizes the three hard hypotheses in block j.

Conditioning at the beginning of each block makes the learner's continuation an arbitrary randomized one-dimensional forecaster. Invoke fresh hard-process randomness and sum the conditional lower bounds. As in Appendix F, a swap selector chooses all coordinates independently and leftover rounds have nonnegative expected excess. Therefore

$$\mathbb{E}\,\mathrm{SwapReg}_T(\ell_B, \mathcal{H}_p) \ge cdn^{1/3} = \Omega_p\left(T^{(p+1)/(p+3)}\right).$$

For $0 < p \le 2$, this matches Corollary 5 and proves Theorem 6 in the low-entropy regime.

G.2 High-Entropy Regime

Fix $p > 2$. For every $m \ge 1$, define

$$\begin{aligned} K_m &= 2^m, & b_m &= \frac{1}{16(m+1)^{1/p}}, \\ q_{m,j} &= \frac{3}{8} + \frac{j-1}{4(K_m - 1)}, & & j \in [K_m]. \end{aligned}$$

Let the domain contain a null context $\varnothing$ and all triples (m, j, k) with $m, k \ge 1$ and $j \in [K_m]$. Define the baseline $q(\varnothing) = 1/2$ and $q(m, j, k) = q_{m,j}$. Define the atomic class

$$\mathcal{A}_p = \{q\} \cup \left\{h_{m,j,w} : m \ge 1,\ j \in [K_m],\ w \in [-1, 1]^{\mathbb{N}}\right\}, \tag{24}$$

where every atom is $1/2$ at the null context and

$$\begin{aligned} h_{m,j,w}(m', j', k) &= q_{m',j'} \\ &\quad + \mathbf{1}\{(m', j') = (m, j)\} b_m k^{-1/p} w_k. \end{aligned}$$

Take $\mathcal{H}_p = \mathrm{conv}(\mathcal{A}_p)$. This is a fixed convex class, and the range of every hypothesis lies in $[5/16, 11/16]$.

We establish its entropy upper bound directly. Every $h \in \mathcal{H}_p$ has a finite block representation

$$h = q + \sum_g \alpha_g v_g, \qquad \alpha_g \ge 0, \qquad \sum_g \alpha_g \le 1,$$

where $g = (m, j)$, the deviation v_g is supported on gate g, and $|v_g(m, j, k)| \le b_m k^{-1/p}$. Fix $\beta > 0$ and discard blocks with $\alpha_g b_m \le \beta/4$. The remaining scales satisfy $m \le C_p \beta^{-p}$. Moreover, their support can be encoded in $O_p(\beta^{-p})$ bits because a gate at scale m has a self-delimiting $O(m)$-bit name and

$$\sum_{g \ \mathit{active}} m \le \frac{4}{\beta} \max_{m \le C_p \beta^{-p}} b_m m \le C_p \beta^{-p}.$$

Here the first inequality uses $\sum_{g\ \mathit{active}} 1/b_m \le 4/\beta$, which follows from the active-block condition and $\sum_g \alpha_g \le 1$.

For each active gate set $r_g = \lceil 4\alpha_g b_m/\beta \rceil$. There are at most C/β active gates, and each $r_g \le C/\beta$, so encoding these integers costs $O(\beta^{-1}\log(1/\beta)) = O_p(\beta^{-p})$ bits. A scalar-axis net at block radius $r_g\beta/4$ and accuracy $\beta/4$ has logarithmic size at most $C_p r_g^p$, by the integral comparison used in the low-entropy proof. Finally,

$$\sum_{g\ mactive} r_g^p \le C_p \left\{ \beta^{-p} \sum_g \alpha_g^p b_m^p + \beta^{-1} \right\} \le C_p \beta^{-p}.$$

The last step uses $b_m \le b_1$ and $\sum_g \alpha_g^p \le (\sum_g \alpha_g)^p \le 1$. Since distinct blocks have disjoint support, the discarded blocks and blockwise nets give a static uniform β-cover. Thus $\log \mathcal{N}_2(\beta, \mathcal{H}_p, n) \le C_p \beta^{-p}$ at every depth. Translation and scaling give the same bound for each induced Brier offset class.

For the reverse bound, fix the gate $(m, j) = (1, 1)$. On that gate, $\mathcal{A}_p \subseteq \mathcal{H}_p$ contains all sequences $q_{1,1} + b_1 k^{-1/p} w_k$. Cycle the first $M = \lfloor (c_p b_1/\beta)^p \rfloor$ coordinates on a deterministic context tree. A constant-distance sign code on these coordinates has size $\exp(c_p M)$ and empirical ℓ_2 separation greater than 2β whenever $n \ge M$. Thus the sequential covering entropy is $\Omega_p(\beta^{-p})$ for $n \ge C_p \beta^{-p}$. This proves the entropy claim.

It remains to prove the regret lower bound. For a sufficiently large horizon, set

$$m = \left\lfloor \frac{\log_2(T \log T)}{2p+1} \right\rfloor, \quad K = 2^m,$$
$$n = \lfloor T/K \rfloor, \qquad \beta = b_m n^{-1/p}.$$

Use K blocks of n rounds. In block j, the contexts are (m, j, k) for $k = 1, \dots, n$. Independently sample signs $\epsilon_{j,k}$ and then binary labels with means

$$\mathbb{E}[Y_{j,k} \mid \epsilon_{j,k}] = q_{m,j} + \epsilon_{j,k}/8.$$

These means lie in $[1/4, 3/4]$. All signs and labels may be sampled before play. For each gate j, define the hindsight vector

$$w_k^j = \frac{\epsilon_{j,k}\beta}{b_m k^{-1/p}} \quad (k \le n), \qquad w_k^j = 0 \quad (k > n).$$

It is feasible because $k \le n$. Map every realized forecast a to $h_{m,J(a),w^{J(a)}}$, where $J(a)$ is the index of a nearest anchor $q_{m,j}$, with ties broken deterministically. This is one valid swap selector.

Consider a round in true block j and condition on the history and forecast a. If $J(a) = j$, write $d = a - q_{m,j}$. Averaging over the fresh sign and label, the selected comparator's expected half-Brier gain is exactly

$$\frac{1}{2}d^2 + \frac{1}{8}\beta - \frac{1}{2}\beta^2 \ge \frac{1}{16}\beta.$$

If $J(a) \ne j$, the selected gated hypothesis is inactive on the current context and predicts $q_{m,j}$. The anchor spacing is at least $1/(4K)$, so $|d| \ge 1/(8K)$ and its expected gain is

$$\frac{1}{2}d^2 \ge \frac{1}{128K^2}.$$

Consequently every block round contributes at least $c \min\{\beta, K^{-2}\}$ in conditional expectation.

The dyadic choice implies $K = \Theta_p((T \log T)^{1/(2p+1)})$, $m + 1 = \Theta_p(\log T)$, and

$$\beta^p = \Theta_p\left(\frac{K}{T \log T}\right) = \Theta_p(K^{-2p}).$$

Thus $\beta = \Theta_p(K^{-2})$. Also $Kn \ge T/2$ for large T. Summing the conditional bounds yields

$$\mathbb{E}\,\mathrm{SwapReg}_T(\ell_B, \mathcal{H}_p) \ge \frac{c_p T}{K^2} \ge \frac{c_p T^{(2p-1)/(2p+1)}}{(\log T)^{2/(2p+1)}}.$$

On the fewer than K leftover rounds, present the null context. All hypotheses then predict $1/2$. One of the two labels has nonnegative conditional expected excess against $1/2$, since the average over both labels is $\mathbb{E}_t[(p_t - 1/2)^2]/2$. Choosing that label before the fresh prediction draw preserves the lower bound. The randomized block construction also has a fixed oblivious realization attaining its expectation. This proves Theorem 6 for $p > 2$.

H Offline Product Lower Bound

We spell out the reduction used in Theorem 12. Theorem 14 of Collina et al. (2026) constructs, for every sufficiently small ϵ, a finite family of Bernoulli distributions and $k = O(\log^4(1/\epsilon))$ binary group functions such that any randomized learner achieving mean-ECE multicalibration error at most ϵ with probability $2/3$ needs

$$n = \Omega\left(\frac{1}{\epsilon^3 \log^3(1/\epsilon)}\right).$$

Their proof uses a uniform prior over a finite code family and Fano's inequality. Hence its contrapositive is a Bayes statement for that prior, not only a pointwise worst-case statement.

Write the binary groups as $c_1, \dots, c_k$, put $\varphi(x) = (1, c_1(x), \dots, c_k(x))/\sqrt{k+1}$, and define

$$\mathcal{W}_{\mathrm{base}} = \mathrm{conv}\{\sqrt{k+1}(pe_0 + ae_j) : j \in [k],\ 0 \le p \le 1,\ -p \le a \le 1-p\}.$$

Let $\mathcal{H}_{\mathrm{base}} = \{x \mapsto \langle w, \varphi(x)\rangle : w \in \mathcal{W}_{\mathrm{base}}\}$. This class is convex, $[0,1]$-valued, has pseudodimension at most

$k+1$, and satisfies $\|\varphi(x)\|_2 \le 1$ and $\sup_w \|w\|_2^2 \le 2(k+1)$. Ordinary mean-ECE multicalibration error is no larger than its swap counterpart. More directly, fix a binary group c witnessing ordinary error and, at prediction level p, put

$$a_p = \mathbb{E}[c(X)(Y-p) \mid p], \qquad h_p(x) = p + a_p c(x).$$

If $a_p > 0$, then $a_p \le 1-p$, and if $a_p < 0$, then $a_p \ge -p$. Thus $h_p \in [0,1]$. Expanding squared loss shows that its conditional improvement is at least a_p^2. Jensen's inequality over prediction levels makes the unscaled squared-loss improvement at least the square of the ordinary mean-ECE error. This is the contrapositive direction in the proof of Theorem 3.3 of Gopalan et al. (2023). Every h_p belongs to $\mathcal{H}_{\text{base}}$.

The calculation applies to randomized prediction kernels because a_p conditions on the realized output, and arbitrary output laws can be quantized with arbitrarily small error (Collina et al., 2026). Below their sample threshold, Fano and their decoding lemma force $\widetilde{\Omega}(n^{-1/3})$ mean-ECE error with constant probability, so the squared witness gives expected half-Brier swap excess at least

$$cn^{-2/3}/\operatorname{polylog}(n).$$

Now take b disjoint copies of the domain. Draw independent hard parameters $\theta_1, \dots, \theta_b$ from the Fano prior and define D_θ by first drawing J uniformly from $[b]$, then drawing from the Jth hard distribution. The comparator class is the Cartesian product of the b bounded linear classes. It is representable by at most $b(k+1)$ parameters, its block-sparse features have norm at most one, and its squared parameter radius is at most $2b(k+1)$. Its restriction to each block is independent of all other parameter coordinates. Consequently, for every learned predictor,

$$\operatorname{SwapAgn}_{D_\theta}(\ell_B, \mathcal{H}) = \frac{1}{b}\sum_{j=1}^{b} \operatorname{SwapAgn}^{(j)}_{D_{\theta_j}}(\ell_B, \mathcal{H}^{(j)}).$$

Let M_j be the number of training examples in block j. It is independent of θ_j and has mean m/b. Markov's inequality gives $\Pr(M_j \le 2m/b) \ge 1/2$. Conditional on M_j, on all samples from other blocks, and on all learner randomness used there, the learner restricted to block j is an arbitrary learner for the one-block Fano prior with at most $2m/b$ samples. The other blocks carry no information about θ_j. Applying the one-block Bayes lower bound and averaging over j gives

$$\mathbb{E}\operatorname{SwapAgn}_{D_\theta}(\ell_B, \mathcal{H}) \ge \frac{c}{\operatorname{polylog}(m)}\left(\frac{b}{m}\right)^{2/3}.$$

Set $\bar{k} = C_1 \log^4 m$ to dominate all relevant base dimensions, pad each block, and take $b = \lfloor d/\bar{k} \rfloor$. Then $b \ge 1$, the product uses at most d parameters, and its Bayes bound is $\widetilde{\Omega}((d/m)^{2/3})$. Some deterministic θ attains this average, proving (15).

References

Anagnostides, I., Farina, G., Fishelson, M., Luo, H., and Schneider, J. (2026). Swap regret minimization through response-based approachability. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 195–223. PMLR.

Bairaktari, K., Hu, L., Nguyen, H. L., and Ullman, J. (2026). Testable and actionable calibration for full swap regret. *arXiv preprint arXiv:2605.17749*.

Blum, A. and Mansour, Y. (2007). From external to internal regret. *Journal of Machine Learning Research*, 8:1307–1324.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Calandriello, D., Lazaric, A., and Valko, M. (2017). Second-order kernel online convex optimization with adaptive sketching. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 645–653. PMLR.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press.

Collina, N., Lu, J., Noarov, G., and Roth, A. (2026). The sample complexity of multicalibration. *arXiv preprint arXiv:2604.21923*.

Dagan, Y., Daskalakis, C., Fishelson, M., and Golowich, N. (2024). From external to swap regret 2.0: An efficient reduction for large action spaces. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 1216–1222. ACM.

Farina, G. and Perdomo, J. C. (2026). An efficient black-box reduction from online learning to multicalibration, and a new route to Φ-regret minimization. *arXiv preprint arXiv:2604.19592*.

Fishelson, M., Golowich, N., Mohri, M., and Schneider, J. (2025a). High-dimensional calibration from swap regret. In *Advances in Neural Information Processing Systems*, volume 38, pages 56149–56181.

Fishelson, M., Kleinberg, R., Okoroafor, P., Paes Leme, R., Schneider, J., and Teng, Y. (2025b). Full swap regret and discretized calibration. In *Proceedings of the 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pages 444–480. PMLR.

Garg, S., Jung, C., Reingold, O., and Roth, A. (2024). Oracle efficient online multicalibration and omniprediction. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2725–2792. SIAM.

Ghuge, R., Muthukumar, V., and Singla, S. (2025). Improved and oracle-efficient online ℓ_1-multicalibration. *arXiv preprint arXiv:2505.17365*.

Gibbs, I. and Tibshirani, R. J. (2026). Sample-efficient omniprediction for proper losses. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 2679–2719. PMLR.

Globus-Harris, I., Harrison, D., Kearns, M., Roth, A., and Sorrell, J. (2023). Multicalibration as boosting for regression. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 11459–11492. PMLR.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gopalan, P., Kim, M. P., and Reingold, O. (2023). Swap agnostic learning, or characterizing omniprediction via multicalibration. In *Advances in Neural Information Processing Systems*, volume 36, pages 39936–39956.

Hazan, E., Agarwal, A., and Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2–3):169–192.

Hu, L., Luo, H., Senapati, S., and Sharan, V. (2026a). Efficient swap multicalibration of elicitable properties. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 3314–3348. PMLR.

Hu, L., Schneider, J., and Wu, Y. (2026b). Near-optimal swap regret minimization for convex losses. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 3285–3313. PMLR.

Hu, L., Tian, K., and Yang, C. (2026c). Optimal recalibration of an online predictor. *arXiv preprint arXiv:2607.19689*.

Hu, L., Tian, K., and Yang, C. (2026d). Simultaneous blackwell approachability and applications to multiclass omniprediction. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 3593–3634. PMLR.

Lu, J., Roth, A., and Shi, M. (2025). Sample efficient omniprediction and downstream swap regret for non-linear losses. In *Proceedings of the Thirty-Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 3829–3878. PMLR.

Luo, H., Senapati, S., and Sharan, V. (2025a). Improved bounds for swap multicalibration and swap omniprediction. In *Advances in Neural Information Processing Systems*, volume 38, pages 23221–23263.

Luo, H., Senapati, S., and Sharan, V. (2025b). Simultaneous swap regret minimization via KL-calibration. In *Advances in Neural Information Processing Systems*, volume 38, pages 70267–70299.

Okoroafor, P. (2026). Fast rates for swap-agnostic learning of proper losses. *arXiv preprint arXiv:2607.28856*.

Rakhlin, A. and Sridharan, K. (2014). Online nonparametric regression. In *Proceedings of the 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 1232–1264. PMLR.

Reid, M. D. and Williamson, R. C. (2010). Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422.

Vovk, V. (1997). Competitive on-line linear regression. In *Advances in Neural Information Processing Systems*, volume 10, pages 364–370.