FAST DYNAMIC REGRET WITH LOGARITHMIC STATE FOR ONLINE LEAST SQUARES

Anonymous authors
Paper under double-blind review

ABSTRACT

Online least squares in a non-stationary environment admits fast $T^{1/3}P_T^{2/3}$ dynamic regret under Euclidean comparator path length P_T. The continuous fixed-share implementation used for this rate keeps one Gaussian component for every possible start time, so its per-round cost grows with the horizon. We study which state is needed for the guarantee. The exact fixed-share posterior may require t positive Gaussian components even on a scalar stream; we instead retain only the state used by the analysis. Our learner applies sleeping Gaussian exponential weights to dyadic interval lifetimes. It also builds a causal, lossy orthonormal representation of the feature stream under a summable budget ε. The learner uses only the label bound, a reference point, and (for the lossy variant) ε; it does not use T, P_T, the feature scale, or the comparator radius. The analysis combines an interval regret bound valid for every real parameter with an exact identity: weighted birth variation is the minimum weighted segment-start cost over all couplings with prescribed Gaussian marginals. This yields $\widetilde{O}\Big(m_T + m_T^{2/3}T^{1/3}(P_T^{\mathrm{obs}})^{2/3}\Big) + O_D(\varepsilon + \varepsilon^2)$ with worst-case per-round time and space $O\big(d(1+m_t) + (1+m_t^2)\log(t{+}1)\big)$. Here the retained dimension $m_T \le d$, so the dynamic term scales as $m_T^{2/3}$, and $P_T^{\mathrm{obs}} \le P_T$ is the shortest path preserving every comparator prediction. We also give a scalar lower bound matching the T- and P_T-exponents, a stream of feature rank T whose gate retains one direction, an instance showing that the representation cross-term is unavoidable, and numerical checks of the reference implementation.

1 INTRODUCTION

Deployed learning systems often face changing data. Online convex optimization models these sequential prediction problems. Its standard performance measure is *dynamic regret*, the excess loss relative to a comparator sequence $u_1, \dots, u_T$. We measure comparator movement by the Euclidean path length $P_T = \sum_{t=2}^{T} \|u_t - u_{t-1}\|_2$ (Zinkevich, 2003). For square loss, curvature permits a $T^{1/3}P_T^{2/3}$ dependence, improving on the general convex-loss scale $\sqrt{TP_T}$ (Baby & Wang, 2021). A deployed learner must also keep its per-round computation bounded as the stream grows. This creates a computational problem for fast-rate methods: they mix over the time at which the environment may have changed, so the number of hypotheses usually grows with the horizon. We ask how much of this state is necessary. The literature first developed *tracking*. Fixed-share recycles a little prior mass into every expert each round. This gives shifting-comparator bounds over finitely many experts (Herbster & Warmuth, 1998) and, for linear prediction, over a continuum (Herbster & Warmuth, 2001). For square loss, the Vovk–Azoury–Warmuth forecaster gives the matching $O(d \log T)$ static guarantee (Vovk, 2001; Azoury & Warmuth, 2001). Zinkevich (2003) introduced the path-length formulation used today. A second line of work replaced switch counts by *interval* guarantees. Adaptive regret (Hazan & Seshadhri, 2009) and strongly adaptive regret over dyadic intervals (Daniely et al., 2015; Jun et al., 2017) use a logarithmic number of active learners. The sleeping-expert view explains the mass bookkeeping as learners enter and leave (Freund et al., 1997; Adamskiy et al., 2012; Koolen et al., 2012). These bounds are static inside each interval, so converting them into a path-length guarantee requires an additional argument and can lose the fast exponent (Zhang et al., 2018b). Expert pooling gives minimax dynamic regret for convex losses (Zhang et al., 2018a; Zhao et al., 2020). For curved losses, Baby & Wang (2021) reached the fast $T^{1/3}P_T^{2/3}$

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.

rate, paying polynomial dimension factors and partition machinery. Cost was studied next. Lu & Hazan (2022) traded exponent for state, attaining $T^{1/3+\epsilon}P_T^{2/3}/\epsilon$ with doubly logarithmically many experts, while Jacobsen & Cutkosky (2024) obtained efficient dynamic guarantees for online linear regression via discounting. Their guarantee uses a discounted variation instead of P_T, as does the discounted-to-dynamic reduction of Xie et al. (2026). Most recently,

Zhang et al. (2025) combined mixability with continuous fixed-share to obtain $\widetilde{O}(d + dT^{1/3}P_T^{2/3})$ for least squares under Euclidean path length. This is the statistical rate studied here. Their least-squares realization keeps one Gaussian component per start time, so per-round work scales as $O(d^2t)$ and continues to grow. Other approaches reduce dynamic regret to kernelized static regret (Jacobsen et al., 2025) or approximate an RKHS by a fixed subspace (Rokhlin & Karapetyants, 2026). These methods target variation measures or complexity notions different from Euclidean path length. Our comparison keeps the Euclidean-P_T benchmark and asks for bounded per-round state. Exact merging of the fixed-share posterior is impossible in general, which motivates our dyadic construction.

We propose a streaming Gaussian exponential-weights learner whose state grows only logarithmically with time. We compress the state needed for the guarantee while leaving the fixed-share posterior unapproximated. The analysis uses two ingredients. First, each Gaussian component has a *dyadic lifetime*, so at most $\lfloor \log_2 t \rfloor + 2$ components are active. Second, we couple smoothed comparator paths so that they switch rarely and their switch points have controlled second moments. An exact identity shows that the minimum cost of such starts is a weighted birth variation, a weighted version of maximal coupling (Pitman, 1976). This identity preserves the $d^{2/3}$ dimension factor in the conversion from interval regret to path regret.

We then reduce the matrix cost by causally freezing a lossy orthonormal representation of the feature stream under a summable budget ε, following the idea of online sketching (Luo et al., 2016). The resulting comparison error is bounded by a constant independent of T. The analysis also gives a sharper comparator quantity: $P_T^{\mathrm{obs}} \le P_T$ is the shortest path that preserves every comparator prediction. It can be zero even when P_T grows linearly on a full-rank stream. Table 1 compares the result with related guarantees. The contributions are:

- We characterize the minimum weighted segment-start cost of a coupling with prescribed marginals. The resulting transfer theorem converts whole-space interval regret to Euclidean-path dynamic regret without an extra dimension factor. For square loss it gives $\widetilde{O}(d + d^{2/3}T^{1/3}P_T^{2/3})$ with $\lfloor \log_2 t \rfloor + 2$ active Gaussian components and $O(d^2 \log(t{+}1))$ worst-case per-round time and space.
- We add a causally frozen lossy feature representation with a summable budget. The resulting guarantee controls the rate, the state, and the error: $\widetilde{O}\Big(m_T + m_T^{2/3}T^{1/3}(P_T^{\mathrm{obs}})^{2/3}\Big) + O_D(\varepsilon + \varepsilon^2)$ at per-round cost $O(d(1+m_t) + (1+m_t^2)\log(t{+}1))$, in which the retained dimension $m_T \le d$ and the additive representation cost is independent of T.
- We give a scalar lower bound matching the T- and P_T-exponents. We also give an instance showing that the representation cross-term cannot be reduced to squared energy, a true-rank-T stream whose gate retains one direction, a stream forcing $m_t = \min(t, d)$, and a proof that the fixed-share posterior admits no exact merging into fewer positive Gaussian components.

2 Related Work

2.1 Curvature and fast dynamic regret

Curvature is the source of the faster rates for square loss. The Vovk–Azoury–Warmuth forecaster attains $O(d \log T)$ static regret (Vovk, 2001; Azoury & Warmuth, 2001), and the same second-order structure is what lets dynamic regret improve on $\sqrt{T}$-type behaviour: mixability quantifies precisely how much a curved loss permits an aggregating forecaster to exploit (Vovk, 1998; van Erven et al., 2015). Zinkevich (2003) fixed the yardstick. For general convex losses the minimax dynamic regret $\Theta(\sqrt{T(1+P_T)})$ is reached by pooling learners of different rates (Zhang et al., 2018a; Zhao et al., 2020); curvature then improves the exponent, and Baby & Wang (2021) obtained $T^{1/3}P_T^{2/3}$ for exp-concave losses together with lower bounds delimiting it. Most sharply for our protocol, Zhang et al.

Table 1: Comparison of dynamic regret and per-round cost. The rows use different protocols and variation measures, so they are not directly comparable. Logarithmic factors are suppressed in $\widetilde{O}$ terms. Here $X_T := T^{1/3}P_T^{2/3}$, $X_T^{\mathrm{obs}} := T^{1/3}(P_T^{\mathrm{obs}})^{2/3}$, $X_T^{(1)} := T^{1/3}(P_T^{(1)})^{2/3}$, d is the ambient dimension, and m_t is the retained dimension. Here C_T is the ℓ_1 total variation used by Baby–Wang, $P_T^{(1)}$ is the ℓ_1 path used by Lu–Hazan, P_T^{β} and P_T^{γ} denote discounted variation measures, and $P_T^{\mathrm{obs}} \le P_T$ is the shortest prediction-preserving path.

Method	Loss	Variation	Dynamic regret	Per-round time/state
Zinkevich (2003)	convex	P_T	$O(\sqrt{T}(1+P_T))$	$O(d)$
Daniely et al. (2015); Jun et al. (2017)	convex	interval	static within each interval	$O(d\log T)$
Zhang et al. (2018a)	convex	P_T	$O(\sqrt{T(1+P_T)})$, minimax	$O(d\log T)$
Baby & Wang (2021)	exp-conc.	C_T	$\widetilde{O}(\mathrm{poly}(d)\,(1\vee T^{1/3}C_T^{2/3}))$	partitions / restarts
Lu & Hazan (2022)	convex	$P_T^{(1)}$	$\widetilde{O}(T^{\epsilon}X_T^{(1)}/\epsilon)$	$O(\log\log T/\epsilon)$ experts
Jacobsen & Cutkosky (2024)	square	P_T^{γ}	$O(d\log T\vee\sqrt{dTP_T^{\gamma}})$	$O(d^2\log T)$
Xie et al. (2026)	curved	P_T^{β}	$O(d\log T+\sqrt{dTP_T^{\beta}})$	$O(d^2)$ per base learner
Jacobsen et al. (2025)	square	temporal RKHS	$\lambda\|u\|_{\mathcal{H}}^2+d_{\mathrm{eff}}(\lambda)\log T$	kernel / effective dim.
Rokhlin & Karapetyants (2026)	square	$P_T^{\mathcal{H}}$	kernel-dependent	$O(m^2\log T+\log^2 T)$, m fixed
Zhang et al. (2025)	square	P_T	$\widetilde{O}(d+dX_T)$	t Gaussian components, $O(d^2t)$
This paper	square	$P_T^{\mathrm{obs}} \le P_T$	$\widetilde{O}(m_T+m_T^{2/3}X_T^{\mathrm{obs}})+O_D(\varepsilon+\varepsilon^2)$	$O(d(1+m_t)+(1+m_t^2)\log t)$

(2025) derived $\widetilde{O}(d+dT^{1/3}P_T^{2/3})$ for least squares under improper prediction, a relaxation whose necessity for fast rates is documented in related problems (Foster et al., 2018). These constructions use a hypothesis set over *when* the environment changed: a pool of learning rates, a restart or partition schedule, or a continuum of start times. State therefore grows with elapsed time, and the least-squares realization of the sharpest such bound holds one $d \times d$ Gaussian per start time. We keep their protocol, loss normalization, boundedness assumptions and variation measure unchanged, so our bound is directly comparable, and answer that question affirmatively, obtaining along the way a smaller dimension exponent in the dynamic term and a comparator quantity $P_T^{\mathrm{obs}} \le P_T$ that can vanish while P_T grows linearly. These results leave open whether the state growth in earlier constructions is necessary. We retain their protocol and assumptions and address this question directly.

2.2 Interval aggregation and path-length conversion

One way to control non-stationarity is to localize in *time* rather than measure comparator movement. Adaptive regret asks for small regret on every contiguous interval (Hazan & Seshadhri, 2009); strongly adaptive regret makes the dependence on interval length explicit and is achieved by covering the timeline with dyadic intervals, so that only $O(\log t)$ base learners are ever awake (Daniely et al., 2015; Jun et al., 2017). The sleeping-expert formalism explains why mass bookkeeping stays sound as learners enter and leave (Freund et al., 1997; Adamskiy et al., 2012; Koolen et al., 2012), and efficient tracking results show that large structured expert classes can be aggregated cheaply (György et al., 2012; Cesa-Bianchi & Lugosi, 2006). The tradeoff has also been studied head-on: Lu & Hazan (2022) run with doubly logarithmically many experts at the price of a T^{ϵ} factor. The recurring obstacle is that interval guarantees are *static inside each interval*. Turning them into a path-length bound requires committing to a partition and paying comparator movement at every boundary, and the standard route, charging P_T to pieces times diameter, either loses the fast exponent or inflates the dimension factor (Zhang et al., 2018b). We retain the dyadic interval scaffold but replace this conversion step. For any coupling with prescribed marginals, its segment-start cost is bounded below by a weighted birth variation, and a single coupling attains this bound for all nonnegative weights, extending maximal coupling in a weighted form (Pitman, 1976). This exact conversion preserves the $T^{1/3}P_T^{2/3}$ rate with only logarithmically many components. The proof requires the interval bound to hold for every parameter in $\mathbb{R}^d$ and charges switch points through a weighted density, since the coupled Gaussian comparator is unbounded.

2.3 Computational cost: comparison and representation

Cost can also be reduced by changing either the comparison class or the stored state. Discounting replaces the interval scaffold with an exponentially weighted covariance, giving efficient dynamic

guarantees for online linear regression (Jacobsen & Cutkosky, 2024); a discounted-to-dynamic reduction converts such bounds into dynamic ones for curved losses (Xie et al., 2026). More radically, the moving comparator can be represented as a static element of a temporal RKHS, so that dynamic regret becomes kernelized static regret (Jacobsen et al., 2025); earlier work instead equipped the comparator class with a dynamical model (Hall & Willett, 2013). Orthogonally one may compress the *state*: online sketching maintains a low-rank surrogate of the second-moment matrix (Luo et al., 2016; Ghashami et al., 2016), and in the RKHS setting a fixed finite-dimensional subspace supports discounted VAW bounds at $O(m^2 \log T + \log^2 T)$ per round, excluding basis construction (Rokhlin & Karapetyants, 2026); a recent line studies fast rates when only an approximate posterior is tracked (Ohn, 2026). Representation compression has two obstacles: projection error contains a comparator-residual cross term that cannot generally be charged to squared discarded energy, and a fixed-share posterior cannot be exactly merged into fewer positive Gaussian components. We therefore retain the original path-length yardstick and causally freeze a lossy orthonormal representation before each label; a summable threshold turns the resulting error into an additive penalty independent of T, d, and the reference point. The retained dimension adapts to the stream: one true-rank-T stream needs only one direction, while another forces $m_t = \min(t, d)$. Complexity is measured in the real-arithmetic model.

3 Problem Setup: Online Least Squares

3.1 Notation

We use $[n] = \{1, \dots, n\}$, $\|\cdot\|$ for the Euclidean norm on $\mathbb{R}^d$, and I_d for the identity. For symmetric matrices, $M \succ 0$, $M \succeq 0$, and $M \preceq M'$ denote positive definiteness, positive semidefiniteness, and the induced order. Let $\mathcal{N}(\cdot, \cdot)$ denote a Gaussian law, ϕ and Φ the standard normal density and CDF, and $\mathrm{TV}(p, q)$ total variation. We set $[z]_+ = \max\{z, 0\}$, $\mathrm{clip}_B(x) = \max\{-B, \min\{B, x\}\}$, $c_g = \sqrt{2/\pi}$, and $L_T = \log(eT)$; all logarithms are natural except $\log_2$. At round t, the feature and label are $a_t \in \mathbb{R}^d$ and $y_t \in \mathbb{R}$, and the learner emits $\hat{y}_t \in \mathbb{R}$. The prediction and parameter losses are $\ell(z, y) = \frac{1}{2}(z - y)^2$ and $f_t(u) = \frac{1}{2}(a_t^\top u - y_t)^2$, with $\hat{\ell}_t = \ell(\hat{y}_t, y_t)$; importantly, f_t is defined on all of $\mathbb{R}^d$. For an integer interval J, write $G_J = \sum_{t \in J} a_t a_t^\top$ and $S_T = \sum_{t \in [T]} \|a_t\|^2$. The dyadic family is $\mathcal{I} = \{I_{k,j}\} \cup \{I_\infty\}$, where $I_{k,j} = \{j2^k, \dots, (j+1)2^k - 1\}$ for $k \geq 0$, $j \geq 1$, and $I_\infty = \mathbb{N}_+$. Its active set is $\mathcal{H}_t = \{I \in \mathcal{I} : t \in I\}$; s_I, k_I, and $\pi_I > 0$ denote a finite interval's start, scale, and prior mass. Proof-only symbols, including σ, $\Gamma \succ 0$, and $X_{1:T}$, are introduced as needed and are not algorithm inputs.

3.2 Protocol, objective, and modelling choices

At round t, the environment reveals a_t, the learner outputs $\hat{y}_t$, and then observes y_t and updates its state. Thus $\hat{y}_t$ may depend on the past and on a_t, but not on y_t. For a comparator sequence $u_{1:T} \in \mathcal{W}^T$, we define the dynamic regret

$$R_T(u_{1:T}) \;=\; \sum_{t=1}^{T} \ell(\hat{y}_t, y_t) \;-\; \sum_{t=1}^{T} f_t(u_t), \qquad P_T \;=\; \sum_{t=2}^{T} \|u_t - u_{t-1}\|, \tag{1}$$

with $P_1 = 0$. Our bounds are pathwise inequalities $R_T(u_{1:T}) \leq \mathcal{B}(T, P_T, \dots)$ that hold simultaneously for every comparator sequence and every finite prefix. We report worst-case per-round time and space; Section 5.3 replaces P_T by a smaller valid comparator quantity.

Four modelling choices matter. Prediction is *improper*: $\hat{y}_t$ is an arbitrary scalar, while the benchmark remains in $\mathcal{W}$; this is the protocol used for fast least-squares rates (Zhang et al., 2025), and is necessary in related problems (Foster et al., 2018). Non-stationarity is measured by the Euclidean path length P_T in equation 1. The learner is horizon-free: T, P_T, A, and D are not inputs, and no doubling trick is used. Finally, all $O(\cdot)$ and $\widetilde{O}(\cdot)$ statements are one-sided upper bounds; R_T itself may be negative. We assume $B > 0$.

3.3 Standing assumptions

Assumption 1 (Adversarial bounded stream)**.** *The stream $(a_t, y_t)_{t \geq 1}$ is chosen by an arbitrary adversary that may adapt to the learner's past behaviour and to the prediction already submitted*

in the current round, subject to $|y_t| \le B$ *for all* t*, where* $B > 0$ *is known to the learner, and* $\max_{t \le T} \|a_t\| \le A$ *for some finite* A *that is* not *revealed to the learner.*

Assumption 2 (Comparator radius about a given reference point)**.** *The set* $\mathcal{W} \subseteq \mathbb{R}^d$ *is nonempty, compact and convex; a reference point* $w_0 \in \mathcal{W}$ *is supplied to the learner; and* $\sup_{u \in \mathcal{W}} \|u - w_0\| \le D < \infty$*. Comparators satisfy* $u_t \in \mathcal{W}$ *for every* t*, and* D *is not an input to the learner.*

Assumption 3 (Real-arithmetic cost model)**.** *Time and space are counted in the real-arithmetic model: scalar arithmetic, comparison,* exp, log *and square root each cost* $O(1)$*, and one real number occupies one storage cell, so a dense* $d \times d$ *matrix occupies* $O(d^2)$ *cells.*

Assumptions 1 and 2 match the setting used for the target rate and in Table 1 (Zhang et al., 2025); a known label bound is also standard and yields the $O(d \log T)$ static benchmark (Vovk, 2001; Azoury & Warmuth, 2001). The learner receives only B: $\eta = 1/(4B^2)$, while A, D, and S_T are evaluation quantities. The comparator assumption uses a radius about a supplied point, so the learner needs neither projection nor an optimization oracle. The guarantees hold for adaptive adversaries on realized sequences and require no stochastic-design, noise, or persistent-excitation assumptions. Smoothness and prediction-level exp-concavity follow from the square loss and the label bound. Complexity is measured in the real-arithmetic model; finite-precision stability is outside the scope of this analysis.

4 Algorithm: Dyadic Gaussian EW with Frozen Lossy Features

The learner has three parts. A *lifetime schedule* assigns each Gaussian component a dyadic interval and deletes it when the interval ends. Thus at most $\lfloor \log_2 t \rfloor + 2$ components are active at round t. A *conjugate core* updates each live component in closed form, using coordinates in the retained subspace. A *representation gate* decides, before the label is revealed, whether to keep the incoming feature direction. Rejected directions are discarded permanently and charged to an explicit error certificate. The only inputs are $B > 0$, the reference point w_0, and a budget $\varepsilon \ge 0$. When $\varepsilon = 0$, every direction with nonzero residual is admitted and the method recovers exact-rank tracking. Algorithm 1 states the procedure. Beyond the symbols of §3 we write $V_t \in \mathbb{R}^{d \times m_t}$ for the orthonormal basis of the retained subspace and m_t for its dimension, (β_I, K_I) for a component's mean and covariance *in those coordinates*, $o_t = a_t^\top w_0$ for the anchor offset, $\varsigma_t = 1 + B + |o_t| + \|a_t\|$ for the round's scale, and e_t^{f} for the discarded feature residual.

Deletion is scheduled rather than opportunistic: a finite component's lifetime is fixed at birth, so expiry on line 27 is part of the specification. The active set has size precisely $\lfloor \log_2 t \rfloor + 2$, and at most $1 + \lfloor \log_2 t \rfloor$ components are born in any round; hence births, updates, and deletions are all charged to the round in which they occur. Line 20 may subtract the running maximum before exponentiating, but the absolute values of $\log w_I$ are retained. New components start with prior π_I, and the proof compares all live masses with a total budget of one, so renormalizing survivors after deletion would change the algorithm. Clipping appears only on line 18; line 25 uses the unclipped parameter loss, preserving the closed-form update and exact evidence factor. The offset $o_t = a_t^\top w_0$ in line 17 is computed from the true feature and remains inside the predictive mean. Absorbing it into the label could violate the label bound needed for exp-concavity, so the components operate on an affine virtual stream.

When a direction is admitted, line 8 gives every live component unit prior variance in the new basis direction. This is correct for the frozen virtual stream because earlier virtual features lie in the old retained subspace; the discrepancy from the true stream is charged to W_T and Δ_T in Section 5.3. The gate is causal: τ_t depends only on quantities available before y_t is revealed. Since $1/(t(t+1))$ is summable, every finite prefix satisfies $W_T \le \varepsilon$ and $\Delta_T \le \varepsilon^2$. Although the algorithm stores reduced matrices, it represents an ambient-space method. With V_t and the component coordinates, define

$$\bar{m}_I = w_0 + V_t \beta_I, \qquad \bar{C}_I = I_d + V_t (K_I - I_{m_t}) V_t^\top.$$

Then Algorithm 1 produces exactly the predictions and masses of dyadic Gaussian exponential weights in $\mathbb{R}^d$ on the frozen virtual losses $\widetilde{f}_t(u) = \frac{1}{2}\big(o_t + \widetilde{a}_t^\top (u - w_0) - y_t\big)^2$. Since $m_t \le \min(d, t)$, the worst-case per-round time and space are $O\big(d(1 + m_t) + (1 + m_t^2)\log(t{+}1)\big)$, never worse than $O(d^2 \log(t{+}1))$ for the exact-rank instantiation. Sections A.1 and A.2 analyse the virtual stream, while Section 5.3 transfers the result back to the true one.

Algorithm 1 Dyadic Gaussian EW with causally frozen lossy features

Require: label bound $B > 0$; reference point $w_0 \in \mathbb{R}^d$; budget $\varepsilon \geq 0$
$\eta \leftarrow 1/(4B^2)$; $V_0 \leftarrow$ empty $d \times 0$ matrix; $m_0 \leftarrow 0$; $W_0 \leftarrow 0$; $\Delta_0 \leftarrow 0$
create permanent component I_∞ with $\beta \leftarrow 0$, $K \leftarrow I_{m_0}$, $\log w \leftarrow -\log 2$
for $t = 1, 2, \ldots$ **do**
  observe a_t; $o_t \leftarrow a_t^\top w_0$; $\varsigma_t \leftarrow 1 + B + |o_t| + \|a_t\|$; $\tau_t \leftarrow \varepsilon/\big(t(t+1)\varsigma_t\big)$
  $b \leftarrow V_{t-1}^\top a_t$; $q \leftarrow a_t - V_{t-1} b$; $\rho \leftarrow \|q\|$
  if $\rho > \tau_t$ **then** ▷ admit the direction: $\widetilde{a}_t = a_t$
    $V_t \leftarrow [\, V_{t-1},\, q/\rho\,]$; $m_t \leftarrow m_{t-1} + 1$; $b \leftarrow (b^\top, \rho)^\top$; $e_t^{\mathrm{f}} \leftarrow 0$
    for every live component: $\beta \leftarrow (\beta^\top, 0)^\top$, $K \leftarrow \mathrm{diag}(K, 1)$
  else ▷ freeze: discard for ever, $\widetilde{a}_t = V_t b$
    $V_t \leftarrow V_{t-1}$; $m_t \leftarrow m_{t-1}$; $e_t^{\mathrm{f}} \leftarrow q$
  end if
  $W_t \leftarrow W_{t-1} + \varsigma_t \|e_t^{\mathrm{f}}\|$; $\Delta_t \leftarrow \Delta_{t-1} + \|e_t^{\mathrm{f}}\|^2$ ▷ error certificates
  for all $k \geq 0$ such that 2^k divides t **do** ▷ at most $1 + \lfloor \log_2 t \rfloor$ births
    create $I = \{t, \ldots, t + 2^k - 1\}$ with $\beta_I \leftarrow 0$, $K_I \leftarrow I_{m_t}$, $\log w_I \leftarrow \log \pi_I$
  end for
  for all $I \in \mathcal{H}_t$ **do**
    $g_I \leftarrow K_I b$; $v_I \leftarrow b^\top g_I$; $c_I \leftarrow 1 + \eta v_I$; $\mu_I \leftarrow o_t + b^\top \beta_I$
    $z_I \leftarrow \mathrm{clip}_B(\mu_I / c_I)$
  end for
  $p_I \leftarrow$ softmax of $\{\log w_J\}_{J \in \mathcal{H}_t}$ ▷ absolute log-masses; shift only inside the softmax
  predict $\hat{y}_t \leftarrow \sum_{I \in \mathcal{H}_t} p_I z_I \in [-B, B]$
  observe y_t; $\hat{\ell}_t \leftarrow \frac{1}{2}(\hat{y}_t - y_t)^2$
  for all $I \in \mathcal{H}_t$ **do**
    $\log w_I \leftarrow \log w_I + \eta \hat{\ell}_t - \frac{1}{2} \log c_I - \eta(\mu_I - y_t)^2/(2c_I)$
    $\beta_I \leftarrow \beta_I + (\eta/c_I)\, g_I (y_t - \mu_I)$; $K_I \leftarrow K_I - (\eta/c_I)\, g_I g_I^\top$
  end for
  delete every $I \in \mathcal{H}_t$ whose last round is t ▷ scheduled expiry, not pruning
end for

5 Main results

The appendix is self-contained. It proves the auxiliary calculus and linear-algebra facts from scratch and does not invoke external dynamic-regret, strong-adaptivity, maximal-coupling, or PAC-Bayes results. We use four steps. First, we analyze one Gaussian update and obtain an interval bound valid for every $\theta \in \mathbb{R}^d$ (§A.1). Second, we replace a moving comparator by a randomized path and compute the replacement cost (§A.2). Third, we derive the fast rate with logarithmically many components (§5.1) and then refine it using the geometry of the observed features (§5.2). Finally, we discard feature directions under a budget and bound the resulting error (§5.3). Figure 1 summarizes this proof-to-bound pipeline and the two same-run refinements.

Throughout, $K_T = \lfloor \log_2 T \rfloor$, $H_T = 2(K_T + 1)$, $\Lambda_T = \log[2T(T+1)(K_T+1)(K_T+2)]$, and

$$\mathcal{C}_T = \frac{1}{\eta}\Big[\Lambda_T + \frac{d}{2}\log\Big(1 + \frac{\eta S_T}{d}\Big)\Big], \quad \mathcal{G}_T = H_T\Big[\mathcal{C}_T + \frac{D^2 + d + 1}{2\eta}\Big], \quad \mathcal{V}_T = S_T + 2B^2 T. \tag{2}$$

Definition 1 (Weighted birth variation)**.** For probability densities $q_1, \ldots, q_n$ on $\mathbb{R}^m$ and measurable $g : \mathbb{R}^m \to [0, \infty]$, $\mathsf{BV}_g(q_{1:n}) = \int g q_1 + \sum_{t=2}^n \int g\, [q_t - q_{t-1}]_+$.

5.1 Fast dynamic regret with logarithmically many components

Set $\Gamma = \sigma^2 I_d$ in the transfer theorem. Smoothing the square loss then costs exactly $\frac{\sigma^2}{2} S_T$, because its Hessian has rank one. The scale σ is used only in the analysis.

Theorem 1 (Fast dynamic regret and worst-case per-round cost)**.** *Under Assumptions 1–3, for every $T \geq 1$ and every $u_{1:T} \in \mathcal{W}^T$ the learner of §4 with $\varepsilon = 0$ satisfies, for every $0 < \sigma \leq 1$,*

$$R_T(u_{1:T}) \leq \mathcal{G}_T\Big(1 + \frac{P_T}{\sigma}\Big) + \frac{S_T \sigma^2}{2}, \tag{3}$$

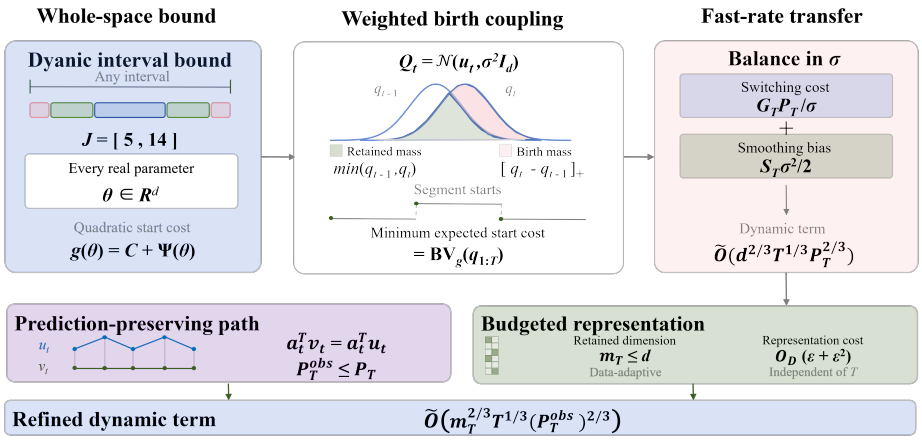

Figure 1: From interval regret to fast dynamic regret. The weighted birth-variation identity converts whole-space interval guarantees into a path-length bound; causal feature freezing then yields a data-adaptive retained dimension and a prediction-preserving comparator path.

and consequently, with $\mathcal{V}_T$ as in equation 2,

$$R_T(u_{1:T}) \le \min\Big\{2B^2T,\ \mathcal{G}_T + \tfrac{3}{2}\mathcal{G}_T^{2/3}\mathcal{V}_T^{1/3}P_T^{2/3}\Big\}. \tag{4}$$

The active set has size exactly $\lfloor \log_2 t \rfloor + 2$, and worst-case per-round time and space are $O(d^2 \log(t{+}1))$.

In equation 3, the smoothing scale controls two opposite costs. A smaller σ makes the coupled path switch more often, contributing $\mathcal{G}_T P_T/\sigma$. A larger σ increases the smoothing bias, contributing $S_T\sigma^2/2$. Balancing them at $\sigma \asymp (\mathcal{G}_T P_T/\mathcal{V}_T)^{1/3}$ gives the term $T^{1/3}P_T^{2/3}$. The linear term in $\mathcal{V}_T$ covers the case in which this choice would exceed one; then the trivial $2B^2T$ bound applies.

5.2 Local geometry: spectral and observation-equivalent refinements

The bound so far is stated in terms of d and P_T, but the algorithm's behaviour depends only on the geometry the stream actually exposes. Keeping Γ general and localizing to a subinterval turns Theorem 1 into an oracle. For $J = [s, e]$ write $G_J = \sum_{t\in J} a_t a_t^\top$, $\Xi_J = \log\det(I_d + \eta G_J)$, $r_J = \operatorname{rank} G_J$, $\mathcal{C}^{\mathrm{sp}}_{J,T} = \eta^{-1}(\Lambda_T + \Xi_J/2)$ and $h(n) = 2(\lfloor \log_2 n \rfloor + 1)$; let U_J be an orthonormal basis of $\operatorname{span}\{a_t : t \in J\}$, $\Pi_J = U_J U_J^\top$, $\xi_t = U_J^\top(u_t - w_0)$ and $\bar{G}_J = U_J^\top G_J U_J$.

Definition 2 (Observation-equivalent path length)**.** For the realized features and a comparator sequence $u_{1:T}$, let $\mathcal{V}_D(u; a) = \{v_{1:T} : \|v_t - w_0\| \le D,\ a_t^\top v_t = a_t^\top u_t\ (1 \le t \le T)\}$ and $P_T^{\mathrm{obs}} = \inf_{v\in\mathcal{V}_D(u;a)} \sum_{t=2}^T \|v_t - v_{t-1}\|$.

5.3 Lossy feature representation

For $\varepsilon > 0$, the learner in §4 tracks the *frozen virtual stream* $\widetilde{f}_t(u) = \frac{1}{2}(o_t + \widetilde{a}_t^\top(u - w_0) - y_t)^2$ instead of the true one. We first establish exact equivalence on the virtual stream and then bound the comparison error. Write $\widetilde{G}_T = \sum_{t\le T} \widetilde{a}_t\widetilde{a}_t^\top$, $\widetilde{S}_T = \operatorname{tr}\widetilde{G}_T$ and $m = \operatorname{rank}\widetilde{G}_T$.

Theorem 2 (A summable budget preserves the fast rate)**.** *With the threshold $\tau_t = \varepsilon/(t(t+1)\varsigma_t)$ of Algorithm 1, every finite prefix satisfies $W_T \le \varepsilon$ and $\Delta_{2,T} \le \varepsilon^2$, hence*

$$R_T(u) \le \min\Big\{2B^2T,\ \mathcal{U}_T(\widetilde{a}, u - w_0) + D\max\{1, D\}\varepsilon + \tfrac{D^2\varepsilon^2}{2}\Big\}, \tag{5}$$

and for fixed A, B, D and $m = m_T > 0$, $R_T(u) = \widetilde{O}(m + m^{2/3}T^{1/3}P_T^{2/3}) + D\max\{1, D\}\varepsilon + \frac{D^2\varepsilon^2}{2}$. When $m = 0$ the virtual regret is at most zero and only the explicit error terms remain. The algorithm uses B, w_0 and ε only; $\varepsilon = 0$ recovers Theorem 1.

The additive penalty is independent of T, d, and w_0. Combining Theorem 2 with Lemma A.5 gives the refined statement, in which the state is governed by the retained dimension and the comparator by what the stream can observe.

5.4 Optimality, necessity, and what cannot be compressed

Fix $A = B = D = 1$ and $w_0 = 0$, let $\mathcal{Z}_{T,d}(P)$ be the deterministic triples $(a_{1:T}, y_{1:T}, u_{1:T})$ obeying those bounds and $\sum_{t\geq 2} \|u_t - u_{t-1}\| \leq P$, and let $\mathfrak{R}_{T,d}(P) = \inf_{\mathcal{A}} \sup_{z\in\mathcal{Z}_{T,d}(P)} \mathbb{E}_{\mathcal{A}} R_T(z)$ over algorithms obeying the protocol of §3.

Theorem 3 (Scalar lower bound)**.** *For every integer $T \geq 1$ and every real $P \geq 0$, $\mathfrak{R}_{T,1}(P) \geq \frac{3}{4096} \min\{T, \log(T+1) + T^{1/3}P^{2/3}\}$.*

So for $d = 1$ and fixed boundedness parameters the pair of exponents in $T^{1/3}P^{2/3}$, and the additive logarithm, cannot be improved beyond logarithmic factors. Theorem 3 separates the two sources of this obstruction: a block construction gives the dynamic term $\min\{T, T^{1/3}P^{2/3}\}$, while a zero-path stochastic construction gives the $\log(T+1)$ term. Combining them yields a lower bound for the infimum over all protocol-compliant algorithms, including randomized ones. The logarithmic term already appears when $P = 0$, so it is a static statistical cost rather than a price for tracking. The truncation at T reflects the bounded per-round loss and keeps the statement valid for every path budget.

6 Experiments

6.1 Setup

We evaluate three synthetic streams, all satisfying Assumptions 1–2 with $A = B = D = 1$ and $w_0 = 0$: a switching stream with unit-norm features and piecewise-constant comparator shifts of magnitude $\approx \sqrt{2}D$; an observation-equivalent stream with $d = 3$, whose features alternate between two coordinates while the comparator moves in a third; and the compressible stream from Proposition A.2, with $T = d$, $a_1 = e_1$, $a_t = e_1 + \delta_t e_t$, and labels $\mathrm{clip}_B(a_t^\top u_t + \xi_t)$ for Gaussian ξ_t. We compare with the all-start continuous fixed-share method, a single-Gaussian VAW forecaster, and projected OGD. Our learner uses only B, w_0, and ε; baseline hyperparameters are grid-searched. We also report hindsight-tuned OGD and the all-start method as references, the latter with unbounded cost. Timings are wall-clock measurements on one CPU core under Assumption 3; each result averages three seeds, and code and raw outputs accompany the submission.

6.2 Cost and dynamic regret

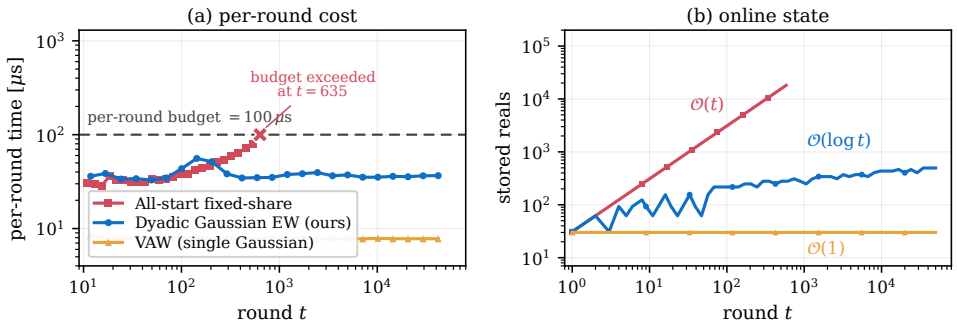

Figure 2: Anytime cost on the switching stream ($d = 5$, $T = 5 \times 10^4$). (a) Per-round wall-clock time in log-spaced round bins: the all-start baseline grows linearly and leaves a 100 μs budget at round 635, while ours stays flat at $\approx 36\,\mu$s. (b) Stored reals, showing the $\mathcal{O}(t)$, $\mathcal{O}(\log t)$ and $\mathcal{O}(1)$ regimes; the sawtooth in our curve is the dyadic birth-and-expiry cycle.

Table 2: Switching-stream cost and dynamic regret ($d = 5$, $T = 50,000$; five comparator shifts; label noise 0.05; mean $\pm$ one standard deviation over three seeds). The upper block contains tuning-free learners with horizon-independent per-round cost; the lower block gives reference methods. The all-start baseline exceeds the $100\,\mu$s budget at round 635.

Method	**State** (reals)	**Time** (μs/rd)	**Rounds** done	**Regret** R_T
Dyadic Gaussian EW (**ours**)	496	36.1	50,000	$\mathbf{909 \pm 16}$
OGD, theory step	5	6.3	50,000	935 ± 64
VAW (single Gaussian)	30	7.8	50,000	3975 ± 403
ref: OGD, oracle-tuned step	5	—	50,000	19 ± 1
ref: all-start fixed-share	19,695	58.9	635	9 ± 0.3

Figure 2 and Table 2 show the predicted state regimes: $\mathcal{O}(t)$ for the all-start baseline, $\mathcal{O}(\log t)$ for our learner, and constant state for the single-Gaussian forecaster. The all-start method exhausts the $100\,\mu$s budget after 635 rounds, whereas ours runs the full horizon at constant cost and has the lowest regret among tuning-free, budget-feasible learners, improving on VAW by $4.4\times$. Corollary A.2 predicts a $P_T^{2/3}$ dependence after the additive term is amortized. The left panel of Figure 3 gives slope 0.58 for our learner, close to the predicted $2/3$; the static VAW baseline is nearly flat (slope 0.10), and our learner is best at five of seven path lengths. A theory-step first-order method is ahead by less than 2% only at the two extreme points.

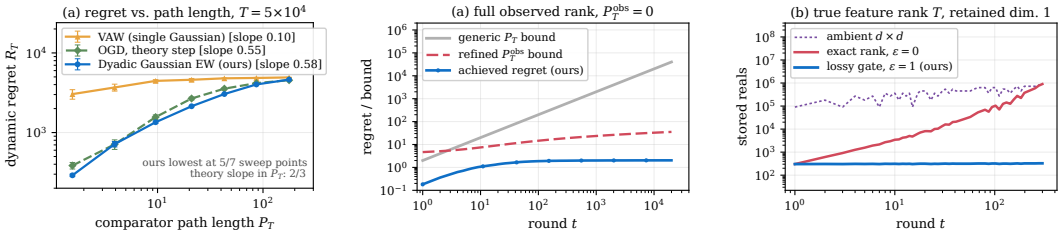

Figure 3: Main empirical checks. From left to right, the panels show dynamic regret against comparator path length, with fitted log-log slopes and one-standard-deviation error bars over three seeds; the observation-equivalent separation, where $P_T^{\mathrm{obs}} = 0$; and retained state on the lossy-representation instance.

6.3 Observation-equivalent path and lossy representation

Theorem A.5 replaces P_T by the shortest path preserving every comparator prediction. On the middle-panel stream, $P_T = 19999$ but $P_T^{\mathrm{obs}} = 0$; the refined guarantee is logarithmic and $1126\times$ tighter than the generic one, while the achieved regret is 2.03 (5.7% of the refined bound). Theorem 2 gives a causal certificate for the lossy representation: with $T = d = 300$, one retained direction stores 327 reals instead of 9.0×10^5 for exact-rank tracking, a $2761\times$ reduction (bottom panel and Appendix Table 3). The regret difference is 1.4×10^{-5}, and the measured certificate satisfies $W_T = 0.124 \le \varepsilon = 1$. Proposition A.2 shows that the retained dimension can instead grow as $m_t = \min(t, d)$.

7 Conclusion

We obtain the fast $T^{1/3}P_T^{2/3}$ dynamic-regret rate for non-stationary online least squares with logarithmically many active Gaussian components. Dyadic lifetimes and an exact coupling identity avoid one component per start time while preserving the improved $d^{2/3}$ dependence. Causal lossy features reduce state on compressible streams, provide an online error certificate, and enable the sharper observation-equivalent path. A scalar lower bound and an irreducible representation cross-term clarify the limits, while the transfer result extends beyond Gaussian square loss. The dimension gap, finite precision, data-independent state control, and learning-rate adaptation remain open.

AI USE STATEMENT

Generative AI tools were used solely to polish the language of this manuscript. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

The problem formulation and assumptions are stated explicitly, and the complete proofs are included in the appendices. The experiments use fully specified synthetic streams, with the constructions, parameter settings, baselines, evaluation protocols, and numerical checks documented in the appendix. Together with the algorithms and update rules in the paper, these materials provide the information needed to independently verify and reproduce the theoretical and empirical results.

REFERENCES

Dmitry Adamskiy, Wouter M. Koolen, Alexey Chernov, and Vladimir Vovk. A closer look at adaptive regret. In *International Conference on Algorithmic Learning Theory (ALT)*, 2012.

Katy S. Azoury and Manfred K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

Dheeraj Baby and Yu-Xiang Wang. Optimal dynamic regret in exp-concave online learning. In *Conference on Learning Theory (COLT)*, 2021.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Amit Daniely, Alon Gonen, and Shai Shalev-Shwartz. Strongly adaptive online learning. In *International Conference on Machine Learning (ICML)*, 2015.

Dylan J. Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *Conference on Learning Theory (COLT)*, 2018.

Yoav Freund, Robert E. Schapire, Yoram Singer, and Manfred K. Warmuth. Using and combining predictors that specialize. In *ACM Symposium on Theory of Computing (STOC)*, 1997.

Mina Ghashami, Edo Liberty, Jeff M. Phillips, and David P. Woodruff. Frequent directions: Simple and deterministic matrix sketching. *SIAM Journal on Computing*, 45(5):1762–1792, 2016.

András György, Tamás Linder, and Gábor Lugosi. Efficient tracking of large classes of experts. *IEEE Transactions on Information Theory*, 58(11), 2012.

Eric C. Hall and Rebecca M. Willett. Dynamical models and tracking regret in online convex programming. In *International Conference on Machine Learning (ICML)*, 2013.

Elad Hazan and Comandur Seshadhri. Efficient learning algorithms for changing environments. In *International Conference on Machine Learning (ICML)*, 2009.

Mark Herbster and Manfred K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2): 151–178, 1998.

Mark Herbster and Manfred K. Warmuth. Tracking the best linear predictor. *Journal of Machine Learning Research*, 1:281–309, 2001.

Andrew Jacobsen and Ashok Cutkosky. Online linear regression in dynamic environments via discounting. In *International Conference on Machine Learning (ICML)*, 2024.

Andrew Jacobsen, Alessandro Rudi, Francesco Orabona, and Nicolò Cesa-Bianchi. Dynamic regret reduces to kernelized static regret. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Kwang-Sung Jun, Francesco Orabona, Stephen Wright, and Rebecca Willett. Improved strongly adaptive online learning using coin betting. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Wouter M. Koolen, Dmitry Adamskiy, and Manfred K. Warmuth. Putting Bayes to sleep. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

Zhou Lu and Elad Hazan. On the computational efficiency of adaptive and dynamic regret minimization. *arXiv preprint arXiv:2207.00646*, 2022.

Haipeng Luo, Alekh Agarwal, Nicolò Cesa-Bianchi, and John Langford. Efficient second order online learning by sketching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Ilsang Ohn. Fast rates in Bayesian online learning with approximate posteriors. *arXiv:2608.25706*, 2026.

J. W. Pitman. On coupling of Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 35:315–322, 1976.

Dmitry B. Rokhlin and Georgiy A. Karapetyants. Dynamic regret for online regression in RKHS via discounted VAW and subspace approximation. *arXiv:2604.25021*, 2026.

Tim van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.

Vladimir Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153–173, 1998.

Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Yan-Feng Xie, Yu-Jie Zhang, Peng Zhao, and Zhi-Hua Zhou. Dynamic regret via discounted-to-dynamic reduction with applications to curved losses and Adam optimizer. In *International Conference on Machine Learning (ICML)*, 2026.

Lijun Zhang, Shiyin Lu, and Zhi-Hua Zhou. Adaptive online learning in dynamic environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018a.

Lijun Zhang, Tianbao Yang, Rong Jin, and Zhi-Hua Zhou. Dynamic regret of strongly adaptive methods. In *International Conference on Machine Learning (ICML)*, 2018b.

Yu-Jie Zhang, Peng Zhao, and Masashi Sugiyama. Non-stationary online learning for curved losses: Improved dynamic regret via mixability. In *International Conference on Machine Learning (ICML)*, 2025.

Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Dynamic regret of convex and smooth functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning (ICML)*, 2003.

A Deferred Results Used in the Main Proof

This section collects the auxiliary results cited in the main text. Their proofs appear in the following sections.

A.1 Interval and coupling results

Lemma A.1 (Exact Gaussian round). *Let $U \sim \mathcal{N}(\theta, C)$ with $C \succ 0$, let $a \in \mathbb{R}^d$ and $y \in [-B, B]$, and put $\mu = a^\top \theta$, $v = a^\top C a$, $c = 1 + \eta v$. Then*

$$Z := \mathbb{E} \exp\Big[-\frac{\eta}{2}(a^\top U - y)^2\Big] = c^{-1/2} \exp\Big[-\frac{\eta(\mu - y)^2}{2c}\Big]. \tag{6}$$

If $0 < \eta \le 1/B^2$, then $\ell(\mathrm{clip}_B(\mu/c), y) \le -\frac{1}{\eta} \log Z$. Moreover, reweighting $\mathcal{N}(\theta, C)$ by $e^{-\eta f(U)}$ produces the Gaussian with

$$C' = C - \frac{\eta}{c}(Ca)(Ca)^\top \succ 0, \qquad \theta' = \theta + \frac{\eta}{c} Ca\,(y - \mu). \tag{7}$$

Lemma A.2 (Dyadic bookkeeping). *The masses of §4 satisfy $\sum_{I \in \mathcal{I}} \pi_I \le 1$, and $|\mathcal{H}_t| = \lfloor \log_2 t \rfloor + 2$ for every $t \ge 1$. Every integer interval $[r, s] \subseteq [1, T]$ is the disjoint union of finite dyadic intervals using at most two per scale, hence at most H_T in total, and each satisfies $\log(1/\pi_I) \le \Lambda_T$.*

Lemma A.3 (Live mass never inflates). *For every stream and every t, $\sum_{I \in \mathcal{H}_t} w^+_{I,t} \le \sum_{I \in \mathcal{H}_t} w_{I,t}$. In the accounting that holds unborn intervals at their prior and freezes expired ones at their terminal mass, the total mass over $\mathcal{I}$ never exceeds one; consequently every individual mass is at most one throughout its life.*

Theorem A.1 (Whole-space interval regret). *For every finite dyadic $I \subseteq [1, T]$ and every $\theta \in \mathbb{R}^d$,*

$$\sum_{t \in I} \big[\hat{\ell}_t - f_t(\theta)\big] \le \frac{1}{\eta} \log \frac{1}{\pi_I} + \frac{\|\theta - w_0\|^2}{2\eta} + \frac{1}{2\eta} \log\det\Big(I_d + \eta \sum_{t \in I} a_t a_t^\top\Big). \tag{8}$$

Hence for every $[r, s] \subseteq [1, T]$ and every $\theta \in \mathbb{R}^d$, $\sum_{t=r}^{s} [\hat{\ell}_t - f_t(\theta)] \le H_T(\mathcal{C}_T + \|\theta - w_0\|^2/(2\eta))$, while the permanent component alone gives

$$\sum_{t=1}^{T} \big[\hat{\ell}_t - f_t(\theta)\big] \le \frac{\log 2}{\eta} + \frac{\|\theta - w_0\|^2}{2\eta} + \frac{d}{2\eta} \log\Big(1 + \frac{\eta S_T}{d}\Big). \tag{9}$$

Theorem A.2 (Weighted birth variation is the least segment-start cost). *Let $\mathfrak{C}(q_{1:n})$ be the set of joint laws on $(\mathbb{R}^m)^n$ with marginals $q_1, \dots, q_n$. Then*

$$\inf_{\mathbb{P} \in \mathfrak{C}(q_{1:n})} \mathbb{E}_{\mathbb{P}} \sum_{t \in \mathcal{J}(X)} g(X_t) = \mathsf{BV}_g(q_{1:n}), \tag{10}$$

and there is a single joint law attaining the infimum simultaneously for every nonnegative measurable g.

A.2 Moving-comparator results

Theorem A.3 (Transfer from interval regret to distributional variation). *Fix a stream of length n, let the algorithm's losses L_t be finite and the comparator losses F_t be nonnegative and measurable, and suppose that for some $C \ge 0$ and some measurable $\Psi \ge 0$, $\sum_{t=a}^{b} [L_t - F_t(x)] \le C + \Psi(x)$ for every subinterval and every $x \in \mathbb{R}^m$. Let $q_1, \dots, q_n$ be densities with $\int F_t q_t < \infty$ and $\mathsf{BV}_{C+\Psi}(q_{1:n}) < \infty$, and let $v_{1:n}$ be any path with $F_t(v_t)$ finite. Then*

$$\begin{aligned}\sum_{t=1}^{n} \big[L_t - F_t(v_t)\big] \le C\Big(1 + \sum_{t=2}^{n} \mathrm{TV}(q_t, q_{t-1})\Big) + \int \Psi q_1 + \sum_{t=2}^{n} \int \Psi\,(q_t - q_{t-1})_+ \\ + \sum_{t=1}^{n} \Big[\int F_t q_t - F_t(v_t)\Big].\end{aligned} \tag{11}$$

Lemma A.4 (Exact equal-covariance Gaussian birth moment). *Let $p = \mathcal{N}(v, \Gamma)$ and $q = \mathcal{N}(u, \Gamma)$ on $\mathbb{R}^r$ with $\Gamma \succ 0$. For $\delta = u - v \ne 0$ set $h = \sqrt{\delta^\top \Gamma^{-1} \delta}$, $\kappa = \|\delta\|^2/h^2$ and $\alpha = 2\Phi(h/2) - 1$, and*

set $h = \kappa = \alpha = 0$ when $\delta = 0$. Then $\mathrm{TV}(p, q) = \alpha$ *and*

$$\int \|x\|^2 \big[q(x) - p(x)\big]_+ dx = \alpha\Big(\frac{\|u\|^2 + \|v\|^2}{2} + \operatorname{tr}\Gamma\Big) + \frac{\|u\|^2 - \|v\|^2}{2} + h\kappa\,\phi(h/2). \tag{12}$$

Corollary A.1 (Exact switching and segment-start totals)**.** *Fix $J = [s, e]$ and marginals $Q_t = \mathcal{N}(\xi_t, \Gamma)$ with $\Gamma \succ 0$, and couple them as in Theorem A.2. With $\delta_t = \xi_t - \xi_{t-1}$, $h_t = \sqrt{\delta_t^\top \Gamma^{-1} \delta_t}$, $\alpha_t = 2\Phi(h_t/2) - 1$, $\kappa_t = \|\delta_t\|^2/h_t^2$ and $\mathsf{A}_J(\Gamma) = \sum_{t=s+1}^{e} \alpha_t$, the coupled path satisfies $\mathbb{E}N(X) = 1 + \mathsf{A}_J(\Gamma)$ and $\mathbb{E}\sum_{t\in\mathcal{J}(X)} \|X_t\|^2 = \mathsf{M}_J(\Gamma)$, where*

$$\mathsf{M}_J(\Gamma) = \frac{\|\xi_s\|^2 + \|\xi_e\|^2}{2} + \operatorname{tr}\Gamma + \sum_{t=s+1}^{e} \alpha_t\Big(\frac{\|\xi_t\|^2 + \|\xi_{t-1}\|^2}{2} + \operatorname{tr}\Gamma\Big) + \sum_{t=s+1}^{e} h_t\kappa_t\phi(h_t/2). \tag{13}$$

Moreover $2\Phi(h/2) - 1 \le \min\{1, c_g h/2\}$ and $h\phi(h/2) \le 2\phi(1)$ for all $h \ge 0$.

A.3 Local-geometry results

Corollary A.2 (Explicit rate)**.** *Fix A, B, D with $B > 0$ and set $L_T = \log(eT)$, $h_0 = 2(1+1/\log 2)$ and $C_* = \frac{h_0}{2\eta}[11 + D^2 + \log(1 + \eta A^2)]$. Then*

$$R_T(u_{1:T}) \le C_* d L_T^2 + \tfrac{3}{2} C_*^{2/3} (A^2 + 2B^2)^{1/3} d^{2/3} T^{1/3} P_T^{2/3} L_T^{4/3}, \tag{14}$$

so that $R_T = \widetilde{O}\big(d + d^{2/3}T^{1/3}P_T^{2/3}\big) \subseteq \widetilde{O}\big(d + dT^{1/3}P_T^{2/3}\big)$ at per-round cost $O(d^2 \log(t+1))$.

Corollary A.3 (Simultaneous static and piecewise-constant guarantees)**.** *For any $u \in \mathcal{W}$, $R_T(u, \dots, u) \le \frac{\log 2}{\eta} + \frac{D^2}{2\eta} + \frac{d}{2\eta}\log(1 + \eta S_T/d)$, which is $O(d \log(eT))$ for fixed A, B, D. If $u_{1:T}$ has M maximal constant stretches, then $R_T(u_{1:T}) \le M H_T(\mathcal{C}_T + D^2/(2\eta))$. Both hold for the same run as Theorem 1, so the minimum of the three bounds is valid without knowing M or P_T.*

Theorem A.4 (Local spectral oracle)**.** *Suppose $r_J > 0$ and let $\Gamma \succ 0$ be any $r_J \times r_J$ matrix. With $\mathsf{A}_J, \mathsf{M}_J$ as in Corollary A.1 and*

$$\mathsf{F}_{J,T}(\Gamma) = h(|J|)\Big[\mathcal{C}^{\mathrm{sp}}_{J,T}\big(1 + \mathsf{A}_J(\Gamma)\big) + \frac{\mathsf{M}_J(\Gamma)}{2\eta}\Big] + \tfrac{1}{2}\operatorname{tr}\big(\bar{G}_J\Gamma\big), \tag{15}$$

the learner satisfies $R_J(u) := \sum_{t\in J}[\hat{\ell}_t - f_t(u_t)] \le \min\{2B^2|J|, \inf_{\Gamma\succ 0} \mathsf{F}_{J,T}(\Gamma)\}$, simultaneously for all intervals and comparator sequences; if $r_J = 0$ then $R_J(u) = 0$. Neither Γ nor the infimum is computed by the algorithm.

Corollary A.4 (Observed rank and observed path length)**.** *Let $J = [1, T]$, $r = \operatorname{rank} G_J > 0$, $P_T^{\|} = \sum_{t=2}^{T} \|\Pi_J(u_t - u_{t-1})\|$ and $D_T^{\|} = \max_t \|\xi_t\|$. With $\mathcal{G}_T^{\mathrm{sp}} = H_T[\mathcal{C}^{\mathrm{sp}}_{J,T} + ((D_T^{\|})^2 + r + 1)/(2\eta)]$,*

$$R_T(u) \le \min\Big\{2B^2T,\ \mathcal{G}_T^{\mathrm{sp}} + \tfrac{3}{2}(\mathcal{G}_T^{\mathrm{sp}})^{2/3}\mathcal{V}_T^{1/3}(P_T^{\|})^{2/3}\Big\}. \tag{16}$$

for fixed A, B, D, where $r \le d$ and $P_T^{\|} \le P_T$.

Lemma A.5 (The minimum exists and dominates projection)**.** *$\mathcal{V}_D(u; a)$ is nonempty, the infimum in Definition 2 is attained, and $0 \le P_T^{\mathrm{obs}} \le P_T^{\|} \le P_T$. Moreover $P_T^{\mathrm{obs}} = 0$ if and only if some θ with $\|\theta - w_0\| \le D$ satisfies $a_t^\top \theta = a_t^\top u_t$ for all $t \le T$.*

Theorem A.5 (Fast rate in the observation-equivalent path length)**.** *Let $r = \operatorname{rank} G_{[1,T]} > 0$ and $\mathfrak{G}_T(D, r) = H_T[\eta^{-1}(\Lambda_T + \frac{1}{2}\log\det(I_d + \eta G_{[1,T]})) + (D^2 + r + 1)/(2\eta)]$. Then, without any change to the algorithm,*

$$R_T(u) \le \min\Big\{2B^2T,\ \mathfrak{G}_T(D, r) + \tfrac{3}{2}\mathfrak{G}_T(D, r)^{2/3}\mathcal{V}_T^{1/3}(P_T^{\mathrm{obs}})^{2/3}\Big\}. \tag{17}$$

and when $P_T^{\mathrm{obs}} = 0$ the sharper static bound $R_T(u) \le \frac{\log 2}{\eta} + \inf\{\|\theta - w_0\|^2/(2\eta)\} + \frac{1}{2\eta}\log\det(I_d + \eta G_{[1,T]})$ holds, the infimum being over all $\theta \in \mathbb{R}^d$ with $a_t^\top \theta = a_t^\top u_t$ for every t.

A.4 Lossy-representation results

Theorem A.6 (Exact equivalence on the virtual stream, and its cost)**.** *In the real-arithmetic model of Assumption 3, Algorithm 1 coincides, prediction by prediction and mass by mass, with dyadic*

Gaussian exponential weights run in $\mathbb{R}^d$ *on* $\widetilde{f}_{1:T}$*, the stored coordinates satisfying* $\bar{m}_I = w_0 + V_t\beta_I$ *and* $\bar{C}_I = I_d + V_t(K_I - I_{m_t})V_t^\top$*. Furthermore* $m_t = \operatorname{rank}(\sum_{s\le t}\widetilde{a}_s\widetilde{a}_s^\top) \le \min(d,t)$*, and the worst-case per-round time and space are* $O\big(d(1+m_t) + (1+m_t^2)\log(t{+}1)\big)$*.*

Lemma A.6 (Exact prediction-error transfer)**.** *For a candidate path* $x_{1:T} \subseteq \mathbb{R}^d$ *put* $p_t = a_t^\top u_t$*,* $\widetilde{p}_t = o_t + \widetilde{a}_t^\top x_t$*,* $h_t = \widetilde{p}_t - p_t$ *and* $E_T(x) = \sum_{t\le T} h_t^2$*, and let* $\mathcal{L}_T(u) = \sum_{t\le T} f_t(u_t)$*. Then*

$$\sum_{t=1}^{T}[\widetilde{f}_t(x_t) - f_t(u_t)] = \sum_{t=1}^{T}(p_t - y_t)h_t + \tfrac{1}{2}E_T(x) \;\le\; \sqrt{2\mathcal{L}_T(u)E_T(x)} + \tfrac{1}{2}E_T(x). \tag{18}$$

Theorem A.7 (Joint representation and comparator bound)**.** *For any label-independent admission rule and every comparator path obeying Assumption 2,*

$$R_T(u) \le \min\Big\{2B^2T,\ \inf_{x_{1:T}\in(\mathbb{R}^d)^T}\Big[\mathcal{U}_T(\widetilde{a},x) + \sqrt{2\mathcal{L}_T(u)E_T(x)} + \tfrac{1}{2}E_T(x)\Big]\Big\}. \tag{19}$$

Taking $x_t = u_t - w_0$ *and writing* $\Delta_{2,T} = \sum_t \|e_t^{\mathrm{f}}\|^2$ *and* $W_T = \sum_t \varsigma_t\|e_t^{\mathrm{f}}\|$ *gives* $R_T(u) \le \min\{2B^2T, \mathcal{U}_T(\widetilde{a}, u - w_0) + \mathcal{A}_T(D)\}$ *with* $\mathcal{A}_T(D) = \min\{D\sqrt{2\mathcal{L}_T(u)\Delta_{2,T}} + \frac{D^2}{2}\Delta_{2,T}, D\max\{1,D\}W_T + \frac{D^2}{2}\Delta_{2,T}\}$*.*

Corollary A.5 (Compression and observation equivalence together)**.** *Let* P_T^{obs} *be defined from the* true *features as in Definition 2, let* $m = m_T > 0$*, and put* $\widetilde{\mathfrak{G}}_T(D,m) = H_T[\eta^{-1}(\Lambda_T + \frac{1}{2}\log\det(I_d + \eta\widetilde{G}_T)) + (D^2+m+1)/(2\eta)]$ *and* $\widetilde{\mathfrak{V}}_T = \widetilde{S}_T + 2B^2T$*. Then the same run satisfies*

$$R_T(u) \le \min\Big\{2B^2T,\ \widetilde{\mathfrak{G}}_T(D,m) + \tfrac{3}{2}\widetilde{\mathfrak{G}}_T(D,m)^{2/3}\widetilde{\mathfrak{V}}_T^{1/3}(P_T^{\mathrm{obs}})^{2/3} + D\max\{1,D\}\varepsilon + \tfrac{D^2\varepsilon^2}{2}\Big\}. \tag{20}$$

so that for fixed A, B, D, ε*,* $R_T(u) = \widetilde{O}\big(m + m^{2/3}T^{1/3}(P_T^{\mathrm{obs}})^{2/3}\big) + O_D(\varepsilon + \varepsilon^2)$*.*

A.5 Lower-bound and compression results

Proposition A.1 (Multidimensional lower bound, and the remaining gap)**.** *Let* $1 \le r \le d$ *and* $x_r = (TP^2/r^2)^{1/3}$ *with* $2 \le x_r \le T/(2r^2)$*. Then there are admissible instances of feature rank exactly* r *on which every algorithm suffers* $\mathbb{E}R_T \ge \frac{3}{2048}r^{1/3}T^{1/3}P^{2/3}$*.*

Theorem A.8 (Determinant certificate for admitted directions)**.** *Let the admissions occur at* $t_1 < \cdots < t_m \le T$ *with residual norms* $\rho_{t_j} > 0$*. For every* $\lambda > 0$*,* $\sum_{j=1}^m \log(1 + \rho_{t_j}^2/\lambda) \le \log\det(I_d + G_T/\lambda)$*; if every admission threshold is at least* $\tau > 0$*, then* $m\log(1+\tau^2/\lambda) \le \log\det(I_d + G_T/\lambda)$*.*

Proposition A.2 (Separation, and its limits)**.** *Fix* $B > 0$*,* $w_0 = 0$*,* $0 < \varepsilon \le 1$ *and* $T \le d$*, and let* $a_1 = e_1$ *and* $a_t = e_1 + \delta_t e_t$ *with* $\delta_t = \varepsilon/[4(B+2)t(t+1)]$ *for* $2 \le t \le T$*. Then the true features have rank* T *while* $m_t = 1$ *for every* t*. Conversely, for* $a_t = e_t$ *with* $1 \le t \le \min(d,T)$ *every residual is admitted and* $m_t = \min(t,d)$*.*

Proposition A.3 (The cross term is unavoidable)**.** *Even for* $d = 1$*,* $A = B = D = 1$*,* $w_0 = 0$ *and a static comparator, the inequality in equation 18 is attained with equality, and no constant* C *can bound the comparison error by* $C\Delta_{2,T}$ *over all such instances.*

Proposition A.4 (No exact merging of the fixed-share posterior)**.** *For* $a_t = 1$*,* $y_t = 0$*, prior* $\mathcal{N}(0,1)$ *and continuous fixed-share with rate* $\mu \in (0,1)$*, the round-*t *predictive density is a mixture of* t *Gaussians with positive weights and pairwise distinct variances, and it cannot be represented exactly by fewer than* t *nondegenerate Gaussian components with positive weights.*

B Auxiliary Facts

This appendix is self-contained. The three lemmas of this section are the only external mathematics we use, and they are proved from elementary calculus and linear algebra; in particular no dynamic-regret, strong-adaptivity, maximal coupling or PAC-Bayes result is invoked anywhere.

Lemma A.7 (Clipping, scalar exp-concavity, and a logarithmic inequality)**.** *Let* $B > 0$ *and* $y \in [-B, B]$*. For every* $x \in \mathbb{R}$*,*

$$\ell(\mathrm{clip}_B(x), y) \le \ell(x, y). \tag{21}$$

For $0 < \eta \le 1/(4B^2)$, every finite collection $x_i \in [-B, B]$ and every $p_i \ge 0$ with $\sum_i p_i = 1$,

$$\exp\Big[-\eta\,\ell\Big(\sum_i p_i x_i, y\Big)\Big] \ge \sum_i p_i \exp\big[-\eta\,\ell(x_i, y)\big]. \tag{22}$$

Finally, $\log c \ge (c-1)/c$ for every $c \ge 1$.

Proof. We first prove equation 21. If $-B \le x \le B$ then clipping does not change x and the two sides are equal. If $x > B$ then $y \le B < x$, so $0 \le B - y \le x - y$ and therefore $(B-y)^2 \le (x-y)^2$. If $x < -B$ then $x < -B \le y$, so $0 \le y + B \le y - x$ and therefore $(-B-y)^2 \le (x-y)^2$. These three cases cover all real x, giving equation 21.

Next define $\varrho(x) = e^{-\eta(x-y)^2/2}$. Differentiating twice,

$$\varrho'(x) = -\eta(x-y)\varrho(x), \qquad \varrho''(x) = \eta\,\varrho(x)\big[\eta(x-y)^2 - 1\big].$$

When $x, y \in [-B, B]$ we have $|x-y| \le 2B$, hence $\eta(x-y)^2 \le 4\eta B^2 \le 1$ and therefore $\varrho''(x) \le 0$. Consequently ϱ' is nonincreasing, which we now convert into the chord inequality without invoking any general theory. For $a < z < b$, the mean value theorem applied on $[a, z]$ and on $[z, b]$ shows that $(\varrho(z) - \varrho(a))/(z-a)$ equals ϱ' at some point of (a, z) and $(\varrho(b) - \varrho(z))/(b-z)$ equals ϱ' at some point of (z, b); since ϱ' is nonincreasing and the first point precedes the second,

$$\frac{\varrho(z) - \varrho(a)}{z - a} \;\ge\; \frac{\varrho(b) - \varrho(z)}{b - z}.$$

Multiplying out by the positive quantities $z - a$ and $b - z$ and rearranging gives

$$\varrho(z) \;\ge\; \frac{b-z}{b-a}\varrho(a) + \frac{z-a}{b-a}\varrho(b), \tag{23}$$

which is equation 22 for two points, since $z = \frac{b-z}{b-a}a + \frac{z-a}{b-a}b$ and every convex combination of a and b arises this way. For n points we argue by induction. Let $q = \sum_{i<n} p_i$. If $q = 0$ then only the last point carries weight and equation 22 is an equality. If $q = 1$ then $p_n = 0$ and the inductive hypothesis applies directly. Otherwise $0 < q < 1$; write $\bar{x} = q^{-1}\sum_{i<n} p_i x_i$, which lies in $[-B, B]$ because it is a convex combination of points of that interval. The inductive hypothesis gives $\varrho(\bar{x}) \ge q^{-1}\sum_{i<n} p_i \varrho(x_i)$, and equation 23 applied to the two points $\bar{x}$ and x_n with weights q and $1-q$ gives $\varrho(q\bar{x} + (1-q)x_n) \ge q\varrho(\bar{x}) + (1-q)\varrho(x_n)$. Combining the two and noting $q\bar{x} + (1-q)x_n = \sum_i p_i x_i$ completes the induction.

Finally, for $c \ge 1$ the fundamental theorem of calculus and the bound $1/x \ge 1/c$ valid on $x \in [1, c]$ give $\log c = \int_1^c \frac{dx}{x} \ge \int_1^c \frac{dx}{c} = \frac{c-1}{c}$, with both sides zero at $c = 1$. □

Lemma A.8 (Gaussian moments and orthogonal invariance)**.** *Let $Z \sim \mathcal{N}(0, I_d)$, let $e \in \mathbb{R}^d$ be a unit vector, let $b \in \mathbb{R}^d$ and let $\sigma > 0$. Then $\mathbb{E}Z = 0$, $\mathbb{E}ZZ^\top = I_d$,*

$$\mathbb{E}|e^\top Z| = c_g = \sqrt{2/\pi} \le 1, \qquad \mathbb{E}\big[\|b + \sigma Z\|^2\,|e^\top Z|\big] = c_g\big(\|b\|^2 + (d+1)\sigma^2\big). \tag{24}$$

Proof. Write $J = \int_{\mathbb{R}} e^{-x^2/2}dx$. The integrand is nonnegative, so we may integrate over bounded regions and pass to the limit; changing to polar coordinates in the plane,

$$J^2 = \int_{\mathbb{R}^2} e^{-(x^2+y^2)/2}dx\,dy = \int_0^{2\pi}\int_0^\infty e^{-\varrho^2/2}\varrho\,d\varrho\,d\vartheta = 2\pi.$$

Since $J > 0$ we get $J = \sqrt{2\pi}$, so the standard normal density is $\varphi(z) = (2\pi)^{-1/2}e^{-z^2/2}$. The integrand $z\varphi(z)$ is odd, whence $\int z\varphi(z)dz = 0$. From $\varphi'(z) = -z\varphi(z)$ and integration by parts,

$$\int_{\mathbb{R}} z^2\varphi(z)dz = -\int_{\mathbb{R}} z\varphi'(z)dz = -\big[z\varphi(z)\big]_{-\infty}^{\infty} + \int_{\mathbb{R}}\varphi(z)dz = 1,$$

the boundary term vanishing because $|z|e^{-z^2/2} \to 0$. For the absolute moments,

$$\mathbb{E}|Z_1| = \frac{2}{\sqrt{2\pi}}\int_0^\infty ze^{-z^2/2}dz = \sqrt{2/\pi} = c_g, \qquad \mathbb{E}|Z_1|^3 = \frac{2}{\sqrt{2\pi}}\int_0^\infty z^3e^{-z^2/2}dz = 2c_g,$$

where the second integral uses the substitution $\varrho = z^2/2$, so that $z^3dz = 2\varrho\,d\varrho$, together with $\int_0^\infty \varrho e^{-\varrho}d\varrho = [-\varrho e^{-\varrho}]_0^\infty + \int_0^\infty e^{-\varrho}d\varrho = 1$.

The standard d-dimensional density is the product of one-dimensional densities, so the coordinates are independent; each has mean zero and second moment one, and distinct coordinates have product expectation zero, which gives $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ^\top = I_d$. For an orthogonal matrix O we have $\|Oz\| = \|z\|$ and $|\det O| = 1$, so substituting $x = Oz$ leaves the Gaussian density invariant; we may therefore rotate e to the first coordinate direction, and the first identity in equation 24 follows from the one-dimensional computation. Also $0 \le \mathbb{E}(|Z_1| - \mathbb{E}|Z_1|)^2 = 1 - c_g^2$, so $c_g \le 1$.

For the weighted second moment, work in an orthonormal frame whose first vector is e and expand $\|b + \sigma Z\|^2 = \|b\|^2 + 2\sigma b^\top Z + \sigma^2 \sum_{i=1}^d Z_i^2$. For each i we have $\mathbb{E}[Z_i|Z_1|] = 0$: for $i = 1$ the integrand is odd in Z_1, and for $i \neq 1$ independence and $\mathbb{E}Z_i = 0$ give the claim. Hence the cross term contributes nothing, while

$$\mathbb{E}\Big[\sum_{i=1}^d Z_i^2\,|Z_1|\Big] = \mathbb{E}|Z_1|^3 + \sum_{i=2}^d \mathbb{E}Z_i^2\,\mathbb{E}|Z_1| = 2c_g + (d-1)c_g = (d+1)c_g.$$

Substituting back gives the second identity in equation 24, and orthogonal transformations preserve $\|b\|$, so the result does not depend on the frame chosen. □

Lemma A.9 (Gaussian integral and a log-determinant bound)**.** *Let $M \succ 0$ and $b \in \mathbb{R}^d$. Then*

$$\int_{\mathbb{R}^d} (2\pi)^{-d/2} \exp\Big[-\tfrac{1}{2}(x-b)^\top M (x-b)\Big] dx = (\det M)^{-1/2}. \tag{25}$$

If $G \succeq 0$, then $\log\det(I_d + \eta G) \le d \log\big(1 + \eta \operatorname{tr} G/d\big)$.

Proof. We first record the orthogonal diagonalization of a real symmetric matrix, since we use it repeatedly. The map $x \mapsto x^\top M x$ is continuous on the unit sphere, which is compact, so it attains a maximum at some unit v. For any $h \perp v$ consider the curve $\gamma(\varrho) = (v + \varrho h)/\|v + \varrho h\|$ of unit vectors; differentiating $\gamma(\varrho)^\top M \gamma(\varrho)$ at $\varrho = 0$ and using that $\varrho = 0$ is a maximum yields $2h^\top M v = 0$. Hence Mv is orthogonal to every $h \perp v$, i.e. Mv is parallel to v and v is an eigenvector. Moreover $v^\top M h = (Mv)^\top h = 0$ for $h \perp v$, so the orthogonal complement of v is invariant under M; applying the same construction there and inducting on the dimension produces an orthonormal eigenbasis, so $M = O\,\mathrm{diag}(\lambda_1, \dots, \lambda_d) O^\top$ for some orthogonal O. If $M \succ 0$ then $\lambda_i = v_i^\top M v_i > 0$, and if $M \succeq 0$ then $\lambda_i \ge 0$.

In equation 25 substitute $z = O^\top(x - b)$, whose Jacobian has absolute value one. The integral becomes $\prod_{i=1}^d \big[(2\pi)^{-1/2} \int_{\mathbb{R}} e^{-\lambda_i z_i^2/2} dz_i\big]$, and the substitution $\varrho_i = \sqrt{\lambda_i} z_i$ together with the one-dimensional normalization from Lemma A.8 shows that the i-th bracket equals $\lambda_i^{-1/2}$. The product is $(\prod_i \lambda_i)^{-1/2} = (\det M)^{-1/2}$.

For the second claim let $\gamma_1, \dots, \gamma_d \ge 0$ be the eigenvalues of G, so that $\log\det(I_d + \eta G) = \sum_i \log(1 + \eta\gamma_i)$ and $\operatorname{tr} G = \sum_i \gamma_i$. The function $x \mapsto \log x$ has second derivative $-1/x^2 < 0$ on the positive axis, so by the finite weighted concavity argument of Lemma A.7 applied with all weights equal to $1/d$,

$$\frac{1}{d}\sum_{i=1}^d \log(1 + \eta\gamma_i) \le \log\Big(\frac{1}{d}\sum_{i=1}^d (1 + \eta\gamma_i)\Big) = \log\Big(1 + \frac{\eta \operatorname{tr} G}{d}\Big).$$

Multiplying by d finishes the proof. □

C Proofs for Section A.1

Proof of Lemma A.1. By Lemma A.9 the matrix $C \succ 0$ has a positive definite square root $C^{1/2}$, so we may write $U = \theta + C^{1/2} Z_0$ with $Z_0 \sim \mathcal{N}(0, I_d)$. Put $h = C^{1/2} a$, so that $a^\top U = \mu + h^\top Z_0$ and $\|h\|^2 = a^\top C a = v$.

If $v = 0$ then $h = 0$, so $a^\top U = \mu$ almost surely and equation 6 holds by direct substitution, both sides equalling $e^{-\eta(\mu - y)^2/2}$ since $c = 1$. If $v > 0$, rotating $h/\|h\|$ to the first coordinate direction and using the orthogonal invariance of Lemma A.8 shows that $a^\top U$ is distributed as $\mathcal{N}(\mu, v)$, whence

$$Z = \frac{1}{\sqrt{2\pi v}} \int_{\mathbb{R}} \exp\Big[-\frac{(x-\mu)^2}{2v} - \frac{\eta}{2}(x - y)^2\Big] dx.$$

Completing the square in the exponent,

$$\frac{(x-\mu)^2}{v}+\eta(x-y)^2=\frac{c}{v}\Big(x-\frac{\mu+\eta v y}{c}\Big)^2+\frac{\eta(\mu-y)^2}{c}, \tag{26}$$

which may be verified term by term: the coefficient of x^2 on the left is $1/v+\eta=c/v$; the coefficient of x is $-2\mu/v-2\eta y$, matching $-2(c/v)(\mu+\eta v y)/c$; and expanding the right-hand side leaves the constant $\mu^2/v+\eta y^2$, as on the left. The one-dimensional Gaussian integral now gives $Z=c^{-1/2}e^{-\eta(\mu-y)^2/(2c)}$, which is equation 6.

For the prediction inequality put $z_0=\mu/c$ and expand using equation 6:

$$-\frac{1}{\eta}\log Z-\frac{1}{2}(z_0-y)^2=\frac{\log c}{2\eta}+\frac{(\mu-y)^2}{2c}-\frac{1}{2}\Big(\frac{\mu}{c}-y\Big)^2=\frac{\log c}{2\eta}+\frac{(c-1)\mu^2}{2c^2}-\frac{(c-1)y^2}{2c}, \tag{27}$$

where the second equality follows by expanding both squares and collecting terms. Since $C\succ 0$ gives $v\geq 0$ and hence $c\geq 1$, the middle term of equation 27 is nonnegative. Using $y^2\leq B^2$ and $\log c\geq(c-1)/c$ from Lemma A.7,

$$-\frac{1}{\eta}\log Z-\ell(z_0,y)\;\geq\;\frac{c-1}{2c}\Big(\frac{1}{\eta}-B^2\Big)\;\geq\;0$$

whenever $\eta\leq 1/B^2$. Applying equation 21 to replace z_0 by $\mathrm{clip}_B(z_0)$ only decreases the loss, which yields the stated inequality. In particular $\eta=1/(4B^2)$ is admissible.

For the update, multiply the Gaussian density by the likelihood $e^{-\eta f(u)}$; the part of the exponent depending on u is

$$-\tfrac{1}{2}(u-\theta)^\top C^{-1}(u-\theta)-\tfrac{\eta}{2}(a^\top u-y)^2.$$

Expanding, the new precision matrix is $M=C^{-1}+\eta aa^\top$ and the new linear term is $C^{-1}\theta+\eta y a$. For $x\neq 0$ we have $x^\top Mx=x^\top C^{-1}x+\eta(a^\top x)^2>0$, so $M\succ 0$. Writing $g=Ca$ and using $C^{-1}g=a$, $aa^\top C=ag^\top$ and $a^\top g=v$, direct multiplication gives

$$\big(C^{-1}+\eta aa^\top\big)\Big(C-\frac{\eta}{c}gg^\top\Big)=I_d+\Big(\eta-\frac{\eta}{c}-\frac{\eta^2 v}{c}\Big)ag^\top=I_d,$$

because $\eta-(\eta+\eta^2 v)/c=\eta-\eta c/c=0$ by $c=1+\eta v$. Hence the matrix C' of equation 7 is M^{-1} and is therefore positive definite. For the mean,

$$C'\big(C^{-1}\theta+\eta y a\big)=\theta+\eta y g-\frac{\eta}{c}g(a^\top\theta)-\frac{\eta^2 y v}{c}g=\theta+\frac{\eta}{c}g(y-\mu),$$

using $g^\top C^{-1}\theta=\mu$, $g^\top a=v$ and $\eta y-\eta^2 yv/c=\eta y/c$. This is exactly equation 7, so each update performed by the algorithm is the closed-form Gaussian conditioning update. □

Proof of Lemma A.2. From $\frac{1}{n(n+1)}=\frac{1}{n}-\frac{1}{n+1}$ the partial sums telescope to $\sum_{n=1}^{N}\frac{1}{n(n+1)}=1-\frac{1}{N+1}$, so the full series equals one; substituting $n=k+1$ gives $\sum_{k\geq 0}\frac{1}{(k+1)(k+2)}=1$ as well. The valid starts and scales of finite dyadic intervals form a subset of all pairs (s,k) with $s\geq 1, k\geq 0$, and all terms are nonnegative, so

$$\sum_{I\neq I_\infty}\pi_I\leq\frac{1}{2}\sum_{s=1}^{\infty}\sum_{k=0}^{\infty}\frac{1}{s(s+1)(k+1)(k+2)}=\frac{1}{2}.$$

Adding the permanent mass $1/2$ gives $\sum_{I\in\mathcal{I}}\pi_I\leq 1$.

For the active count, fix t and a scale k. If $2^k\leq t$ then the unique interval of that scale containing t is $I_{k,\lfloor t/2^k\rfloor}$: indeed $j=\lfloor t/2^k\rfloor\geq 1$ and $j2^k\leq t<(j+1)2^k$. If $2^k>t$ then every admissible $j\geq 1$ has start $j2^k>t$, so no interval of that scale is active. The admissible scales are therefore $k=0,\ldots,\lfloor\log_2 t\rfloor$, and adding the permanent interval gives $|\mathcal{H}_t|=\lfloor\log_2 t\rfloor+2$.

For the covering claim, temporarily consider the complete dyadic grid starting at zero, so that every interval has a parent of twice the length obtained by dropping the least significant bit of its index. Two grid intervals are either disjoint or nested: if the shorter has length 2^k and the longer length 2^l with $k\leq l$, then both endpoints of the longer are integer multiples of 2^k, so the shorter cannot straddle either endpoint of the longer without being contained in it. Now fix $[r,s]$ and, for each integer $t\in[r,s]$, start from the singleton $[t,t]$ and repeatedly pass to the parent as long as the parent

is still contained in $[r, s]$. Lengths double at each iteration and cannot exceed $s-r+1$, so the process terminates, producing a maximal dyadic interval contained in $[r, s]$. These maximal intervals cover $[r, s]$, and if two of them intersect then by nestedness one contains the other, which by maximality of both forces them to coincide; distinct maximal intervals are therefore disjoint.

Fix a scale k and let I be a maximal interval of that scale. Its parent is not contained in $[r, s]$ but does intersect it, so the parent must cross the left endpoint r or the right endpoint s. At the fixed parent scale $k + 1$ the grid intervals are pairwise disjoint, so at most one parent crosses r and at most one crosses s. Each such parent contributes at most one maximal child: if both of its children were contained in $[r, s]$, their union, namely the parent itself, would also be contained in $[r, s]$, contradicting maximality. Hence at most two maximal intervals occur per scale.

Since $r \geq 1$, none of the selected intervals contains zero, so each has positive start and belongs to the family with $j \geq 1$ used by the algorithm. Their lengths are at most T, so their scales are at most K_T, and the total count is at most $2(K_T + 1) = H_T$. Finally, each selected interval has $s_I \leq T$ and $k_I \leq K_T$, so $1/\pi_I = 2s_I(s_I + 1)(k_I + 1)(k_I + 2) \leq 2T(T + 1)(K_T + 1)(K_T + 2)$, and taking logarithms gives $\log(1/\pi_I) \leq \Lambda_T$. □

Proof of Lemma A.3. Fix the active set at round t. By Lemma A.1, $\ell(z_{I,t}, y_t) \leq -\frac{1}{\eta} \log Z_{I,t}$; multiplying by $-\eta < 0$ reverses the inequality, and exponentiating gives

$$e^{-\eta \ell(z_{I,t}, y_t)} \;\geq\; Z_{I,t}. \tag{28}$$

Every $z_{I,t}$ lies in $[-B, B]$ and $\eta = 1/(4B^2)$, so the finite exp-concavity inequality equation 22 of Lemma A.7 together with equation 28 yields

$$e^{-\eta \hat{\ell}_t} = \exp\Big[-\eta\, \ell\Big(\sum_{I \in \mathcal{H}_t} p_{I,t} z_{I,t}, y_t\Big)\Big] \;\geq\; \sum_{I \in \mathcal{H}_t} p_{I,t} e^{-\eta \ell(z_{I,t}, y_t)} \;\geq\; \sum_{I \in \mathcal{H}_t} p_{I,t} Z_{I,t}.$$

Multiplying through by $e^{\eta \hat{\ell}_t} \sum_{J \in \mathcal{H}_t} w_{J,t} > 0$ and using $p_{I,t} = w_{I,t} / \sum_J w_{J,t}$ gives $\sum_I w_{I,t} e^{\eta \hat{\ell}_t} Z_{I,t} \leq \sum_I w_{I,t}$, which by the mass update of Algorithm 1 is precisely the asserted inequality.

For the global statement, assign to every interval in the countable family $\mathcal{I}$ a conceptual mass: an interval that has not yet started carries π_I; an active one carries the mass maintained by the algorithm; and an expired one is frozen at the value it had after its final update. The initial total is at most one by Lemma A.2. In any round only the finitely many active masses change, and their sum does not increase by the previous paragraph, while all other conceptual masses are unchanged; by induction on the round the total never exceeds one. All masses are nonnegative, so each individual mass is bounded by the total, hence by one. This ledger is used only in the analysis; the implementation stores neither future nor expired components. □

Proof of Theorem A.1. Fix a finite dyadic interval $I = [r, s]$ and write $\varphi_0(u) = (2\pi)^{-d/2} e^{-\|u - w_0\|^2/2}$ for its initial density. For $r \leq n \leq s$ define the accumulated evidence

$$E_n = \int \varphi_0(u) \exp\Big[-\eta \sum_{t=r}^{n} f_t(u)\Big] du, \qquad E_{r-1} = 1.$$

Each E_n is strictly positive and at most one, because the Gaussian density is strictly positive and every likelihood factor lies in $(0, 1]$.

We verify the component's posterior by induction. Before round r the density is φ_0. If before round n the density is $\varphi_0(u) e^{-\eta \sum_{t=r}^{n-1} f_t(u)} / E_{n-1}$, then multiplying by $e^{-\eta f_n(u)}$ and normalizing produces normalizing constant E_n / E_{n-1} and density $\varphi_0(u) e^{-\eta \sum_{t=r}^{n} f_t(u)} / E_n$. Lemma A.1 shows that the algorithm's matrix update realizes exactly this posterior, so the induction applies to the actual algorithm and moreover

$$Z_{I,n} = E_n / E_{n-1}, \qquad \prod_{t=r}^{s} Z_{I,t} = E_s. \tag{29}$$

Multiplying the mass updates along the interval therefore gives terminal mass $w_I^{\mathrm{end}} = \pi_I \exp\big(\eta \sum_{t=r}^{s} \hat{\ell}_t\big) E_s$, which is at most one by Lemma A.3. Taking logarithms and rearranging,

$$\sum_{t=r}^{s} \hat{\ell}_t \;\le\; \frac{1}{\eta} \log \frac{1}{\pi_I} - \frac{1}{\eta} \log E_s. \tag{30}$$

We now evaluate E_s exactly rather than bounding it. Set

$$M = I_d + \eta \sum_{t=r}^{s} a_t a_t^\top, \qquad b = w_0 + \eta \sum_{t=r}^{s} y_t a_t, \qquad c_0 = \|w_0\|^2 + \eta \sum_{t=r}^{s} y_t^2,$$

so that $M \succ 0$ and, writing $F(u) = \frac{1}{2}\|u - w_0\|^2 + \eta \sum_{t=r}^{s} f_t(u)$, expansion gives $F(u) = \frac{1}{2}(u^\top M u - 2b^\top u + c_0)$. With $u_* = M^{-1} b$ we obtain $F(u) = \frac{1}{2}(u - u_*)^\top M (u - u_*) + F(u_*)$ and $F(u_*) = \frac{1}{2}(c_0 - b^\top M^{-1} b)$; since the first term is nonnegative, $F(u_*) \le F(\theta)$ for every $\theta \in \mathbb{R}^d$. Substituting the completed square into the definition of E_s and applying equation 25 of Lemma A.9 gives $E_s = e^{-F(u_*)} (\det M)^{-1/2}$, whence

$$-\frac{1}{\eta} \log E_s = \frac{F(u_*)}{\eta} + \frac{\log \det M}{2\eta} \le \sum_{t=r}^{s} f_t(\theta) + \frac{\|\theta - w_0\|^2}{2\eta} + \frac{\log \det M}{2\eta}. \tag{31}$$

Inserting equation 31 into equation 30 and subtracting $\sum_{t \in I} f_t(\theta)$ from both sides proves equation 8.

Now fix an arbitrary interval $[r, s] \subseteq [1, T]$ and decompose it, by Lemma A.2, into disjoint finite dyadic intervals $I_1, \dots, I_m$ with $m \le H_T$. For each j, $\mathrm{tr}\big(\sum_{t \in I_j} a_t a_t^\top\big) = \sum_{t \in I_j} \|a_t\|^2 \le S_T$, so Lemma A.9 gives $\log\det(I_d + \eta \sum_{t \in I_j} a_t a_t^\top) \le d \log(1 + \eta S_T / d)$, while Lemma A.2 gives $\log(1/\pi_{I_j}) \le \Lambda_T$. Hence the right-hand side of equation 8 on each I_j is at most $\mathcal{C}_T + \|\theta - w_0\|^2/(2\eta)$. Summing over j, the left-hand sides add to the regret on $[r, s]$ because the I_j are disjoint and cover it, and the right-hand side is m copies of one nonnegative quantity, so replacing m by H_T preserves the inequality.

Finally, for the permanent component no dyadic decomposition is needed. Define $E_n^\infty = \int \varphi_0(u) \exp[-\eta \sum_{t=1}^{n} f_t(u)] du$ with $E_0^\infty = 1$. Exactly as above, induction on n using Lemma A.1 shows that the stored Gaussian has density $\varphi_0(u) e^{-\eta \sum_{t \le n} f_t(u)} / E_n^\infty$ and that $Z_{I_\infty, n} = E_n^\infty / E_{n-1}^\infty$; multiplying the mass updates and telescoping gives $w^+_{I_\infty, T} = \frac{1}{2} \exp(\eta \sum_{t \le T} \hat{\ell}_t) E_T^\infty \le 1$ by Lemma A.3. Taking logarithms, $\sum_{t \le T} \hat{\ell}_t \le \frac{\log 2}{\eta} - \frac{1}{\eta} \log E_T^\infty$. Completing the square with $M_T^\infty = I_d + \eta \sum_{t \le T} a_t a_t^\top$, $b_T^\infty = w_0 + \eta \sum_{t \le T} y_t a_t$ and $c_T^\infty = \|w_0\|^2 + \eta \sum_{t \le T} y_t^2$ exactly as before, and applying equation 25, gives $-\frac{1}{\eta} \log E_T^\infty \le \sum_{t \le T} f_t(\theta) + \frac{\|\theta - w_0\|^2}{2\eta} + \frac{\log \det M_T^\infty}{2\eta}$. Since $\mathrm{tr}(\sum_{t \le T} a_t a_t^\top) = S_T$, Lemma A.9 bounds the last term by $\frac{d}{2\eta} \log(1 + \eta S_T / d)$, which yields equation 9. No random event, failure probability or discretization of θ enters the argument, so all of these inequalities hold simultaneously for every interval and every $\theta \in \mathbb{R}^d$. □

D Proofs for Section A.2

Proof of Theorem A.2. Fix two adjacent marginals, written $p = q_{t-1}$ and $q = q_t$, and let (X, Y) be an arbitrary joint law with these marginals. Define two finite measures

$$\mathsf{D}(E) = \mathbb{P}(X = Y \in E), \qquad \mathsf{N}(E) = \mathbb{P}(Y \in E,\ X \ne Y).$$

For every measurable E, containment of events gives $\mathsf{D}(E) \le \mathbb{P}(X \in E) = \int_E p$ and $\mathsf{D}(E) \le \mathbb{P}(Y \in E) = \int_E q$, and by disjointness

$$\mathsf{N}(E) = \int_E q - \mathsf{D}(E). \tag{32}$$

Let $E_+ = \{x : q(x) > p(x)\}$ and first take g bounded, nonnegative and measurable. Since N is nonnegative, $\int g \, d\mathsf{N} \ge \int_{E_+} g \, d\mathsf{N}$. The set domination $\mathsf{D}(E) \le \int_E p$ implies $\int h \, d\mathsf{D} \le \int hp$ for every bounded nonnegative h: the inequality holds termwise for nonnegative simple functions, and a general bounded nonnegative h is the increasing limit of such functions. Combining this with

equation 32,

$$\mathbb{E}\big[g(Y)\mathbf{1}\{X \neq Y\}\big] \;\geq\; \int_{E_+} gq - \int_{E_+} g\, d\mathsf{D} \;\geq\; \int_{E_+} g(q-p) = \int g\,(q-p)_+. \tag{33}$$

All integrals here are finite because g is bounded and the measures involved have total mass at most one. For a general nonnegative measurable g, apply equation 33 to $g_K = \min\{g, K\}$ and let $K \uparrow \infty$; both sides are monotone limits of nonnegative quantities, so equation 33 persists and no undefined difference of infinities arises.

We now construct a transition attaining equality in equation 33. Put $\alpha = \int (q-p)_+$, and note $\alpha = \int (p-q)_+$ because $\int (q-p) = 0$. If $\alpha = 0$ then $p = q$ almost everywhere and we may take $Y = X$, which makes both sides of equation 33 zero. If $\alpha > 0$, define

$$a(x) = \begin{cases} \min\{1, q(x)/p(x)\}, & p(x) > 0, \\ 1, & p(x) = 0, \end{cases} \qquad r(y) = \frac{[q(y) - p(y)]_+}{\alpha},$$

and, given $X = x$, retain $Y = x$ with probability $a(x)$ and otherwise draw Y from the density r, independently. Because $p(x)a(x) = \min\{p(x), q(x)\}$, the retained branch contributes the density $\min\{p, q\}$; the probability of entering the resampling branch is $\int p(x)(1-a(x))dx = \int (p-q)_+ = \alpha$, so that branch contributes $\alpha r = (q-p)_+$. The two contributions sum to $\min\{p, q\} + (q-p)_+ = q$, so the marginal of Y is correct. The resampled value has a density with respect to Lebesgue measure, so for any fixed old value x the event $\{Y = x\}$ has probability zero; the two coordinates therefore differ exactly on the resampling branch, and

$$\mathbb{E}\big[g(Y)\mathbf{1}\{X \neq Y\}\big] = \alpha \int gr = \int g\,(q-p)_+. \tag{34}$$

Crucially the kernel was defined without reference to g, so equation 34 holds for every nonnegative measurable g simultaneously.

Finally, initialize from q_1 and apply the above kernel between each pair of consecutive marginals. Each kernel has total mass one, so the iterated integral $q_1(x_1)dx_1 K_2(x_1, dx_2) \cdots K_n(x_{n-1}, dx_n)$ defines a probability law on $(\mathbb{R}^m)^n$, and the marginal computation of the previous paragraph shows by induction on t that the t-th marginal is q_t. For any path,

$$\sum_{t \in \mathcal{J}(X)} g(X_t) = g(X_1) + \sum_{t=2}^{n} g(X_t)\mathbf{1}\{X_t \neq X_{t-1}\}.$$

Bounding each summand from below by equation 33 for an arbitrary coupling yields $\mathsf{BV}_g(q_{1:n})$, while the constructed law achieves equality in each summand by equation 34. Since the sum is finite and its terms nonnegative, expectations may be taken termwise, including when values are infinite. This proves equation 10 together with the simultaneity claim. □

Proof of Theorem A.3. Use Theorem A.2 to construct a path $X_{1:n}$ with the prescribed marginals attaining the weighted birth variation. Fix a realization and let its maximal constant stretches be $[\tau_j, \tau_{j+1} - 1]$ with stretch parameter $x_j = X_{\tau_j}$. The hypothesis holds for every subinterval and every parameter, so for each j,

$$\sum_{t=\tau_j}^{\tau_{j+1}-1} \big[L_t - F_t(X_t)\big] \leq C + \Psi(X_{\tau_j}).$$

The stretches are disjoint and cover $\{1, \ldots, n\}$, so summing over j gives

$$\sum_{t=1}^{n} \big[L_t - F_t(X_t)\big] \leq \sum_{t \in \mathcal{J}(X)} \big[C + \Psi(X_t)\big]. \tag{35}$$

The left side of equation 35 is integrable because each L_t is finite and each $\int F_t q_t$ is finite; the right side is integrable because under the coupling of Theorem A.2 its expectation equals $\mathsf{BV}_{C+\Psi}(q_{1:n})$, which is finite by assumption. Taking expectations in equation 35 and applying equation 10,

$$\sum_{t=1}^{n} L_t - \sum_{t=1}^{n} \int F_t q_t \;\leq\; \mathsf{BV}_{C+\Psi}(q_{1:n}).$$

Add $\sum_t \big(\int F_t q_t - F_t(v_t)\big)$ to both sides and expand the weighted birth variation according to Definition 1. The constant part contributes C from the initial term and, at each subsequent round, $C\int (q_t - q_{t-1})_+ = C\ \mathrm{TV}(q_t, q_{t-1})$, the last equality because $\int (q_t - q_{t-1}) = 0$ forces the positive and negative parts to have equal mass. The remaining terms are the stated weighted integrals of Ψ, which gives equation 11. No optimality, stationarity or Karush–Kuhn–Tucker property of $v_{1:n}$ was used, and the algorithm never samples X. □

Proof of Lemma A.4. If $u = v$ the two densities coincide and both assertions read $0 = 0$. Assume $u \neq v$ and set $\bar{m} = (u + v)/2$ and $e = \Gamma^{-1/2}\delta/h$, which is a unit vector because $h^2 = \delta^\top \Gamma^{-1}\delta$; the roots $\Gamma^{\pm 1/2}$ exist by the orthogonal diagonalization established in Lemma A.9. Subtracting the logarithms of the two densities and expanding the quadratic forms,

$$\log \frac{q(x)}{p(x)} = \delta^\top \Gamma^{-1}(x - \bar{m}). \tag{36}$$

Change coordinates by $x = \bar{m} + \Gamma^{1/2}(ze + Z_\perp)$ with $Z_\perp \in e^\perp$. Whitening maps the two means to $\pm he/2$ and both covariances to the identity, so by the orthogonal invariance and coordinate independence of Lemma A.8 we have $z \sim \mathcal{N}(h/2, 1)$ under q and $z \sim \mathcal{N}(-h/2, 1)$ under p, with $Z_\perp$ standard on $e^\perp$ and independent of z in both cases. By equation 36 the log-ratio equals hz, so the region $\{q > p\}$ is exactly $\{z > 0\}$.

Write $k = h/2$ and recall $\Phi(-k) = 1 - \Phi(k)$ since ϕ is even. Then

$$\int (q - p)_+ = \int_0^\infty \big[\phi(z - k) - \phi(z + k)\big] dz = \Phi(k) - \Phi(-k) = \alpha,$$

and since both densities integrate to one, $\mathrm{TV}(p, q) = \int (q - p)_+ = \alpha$.

For the quadratic weight we need three one-dimensional integrals:

$$\begin{aligned} I_0 &:= \int_0^\infty \big[\phi(z - k) - \phi(z + k)\big] dz = \alpha, \qquad I_1 := \int_0^\infty z\big[\phi(z - k) - \phi(z + k)\big] dz = k, \\ I_2 &:= \int_0^\infty z^2\big[\phi(z - k) - \phi(z + k)\big] dz = (1 + k^2)\alpha + 2k\phi(k). \end{aligned} \tag{37}$$

The first was just computed. From $\phi'(w) = -w\phi(w)$ we get $\int_{-k}^\infty w\phi(w)dw = \phi(k)$ and $\int_k^\infty w\phi(w)dw = \phi(k)$. Substituting $w = z - k$ and $w = z + k$ respectively,

$$\int_0^\infty z\phi(z - k)dz = \phi(k) + k\Phi(k), \qquad \int_0^\infty z\phi(z + k)dz = \phi(k) - k\Phi(-k),$$

and subtracting gives $k[\Phi(k) + \Phi(-k)] = k$, which is I_1. Integrating by parts,

$$\begin{aligned} \int_{-k}^\infty w^2\phi(w)dw &= \big[-w\phi(w)\big]_{-k}^\infty + \int_{-k}^\infty \phi(w)dw = \Phi(k) - k\phi(k), \\ \int_k^\infty w^2\phi(w)dw &= \big[-w\phi(w)\big]_k^\infty + \int_k^\infty \phi(w)dw = \Phi(-k) + k\phi(k), \end{aligned}$$

the boundary terms vanishing because $we^{-w^2/2} \to 0$. Expanding $(w + k)^2$ and $(w - k)^2$ then gives

$$\int_0^\infty z^2\phi(z - k)dz = (1 + k^2)\Phi(k) + k\phi(k), \qquad \int_0^\infty z^2\phi(z + k)dz = (1 + k^2)\Phi(-k) - k\phi(k),$$

and subtracting yields I_2, completing equation 37.

Next integrate over $Z_\perp$ at fixed z. Since $Z_\perp$ has mean zero and covariance $I_r - ee^\top$, the linear cross terms vanish and

$$\mathbb{E}_{Z_\perp}\big\|\bar{m} + \Gamma^{1/2}(ze + Z_\perp)\big\|^2 = \|\bar{m}\|^2 + 2z\,\bar{m}^\top\Gamma^{1/2}e + z^2 e^\top\Gamma e + \mathrm{tr}\,\Gamma - e^\top\Gamma e. \tag{38}$$

Because $\delta = h\Gamma^{1/2}e$ we have $e^\top\Gamma e = \|\delta\|^2/h^2 = \kappa$ and $\Gamma^{1/2}e = \delta/h$. Multiplying equation 38 by the two one-dimensional densities, subtracting and integrating over $z > 0$ with the help of equation 37,

$$\begin{aligned} \int \|x\|^2(q - p)_+ &= \big(\|\bar{m}\|^2 + \mathrm{tr}\,\Gamma - \kappa\big)\alpha + 2\bar{m}^\top\Gamma^{1/2}e\,k + \kappa\big[(1 + k^2)\alpha + 2k\phi(k)\big] \\ &= \alpha\Big(\|\bar{m}\|^2 + \mathrm{tr}\,\Gamma + \frac{\|\delta\|^2}{4}\Big) + \bar{m}^\top\delta + h\kappa\phi(h/2), \end{aligned}$$

where we used $\kappa k^2 = \|\delta\|^2/4$, $2k\bar{m}^\top\Gamma^{1/2}e = \bar{m}^\top\delta$ and $2k\phi(k) = h\phi(h/2)$. Substituting $\bar{m} = (u+v)/2$ and $\delta = u - v$ and using the identities $\|\bar{m}\|^2 + \|\delta\|^2/4 = (\|u\|^2 + \|v\|^2)/2$ and $\bar{m}^\top\delta = (\|u\|^2 - \|v\|^2)/2$ gives equation 12. Every integrand above is a Gaussian density times a quadratic polynomial, so all moments are finite and the interchanges and subtractions are legitimate. □

Proof of Corollary A.1. Under the coupling of Theorem A.2 the probability of a switch at round t equals $\mathrm{TV}(Q_t, Q_{t-1})$, which Lemma A.4 evaluates as α_t. Since $N(X) = 1 + \sum_{t=s+1}^{e} \mathbf{1}\{X_t \neq X_{t-1}\}$ identically, taking expectations gives $\mathbb{E}N(X) = 1 + \mathsf{A}_J(\Gamma)$.

For the segment-start second moment, the initial contribution is $\mathbb{E}\|X_s\|^2 = \|\xi_s\|^2 + \operatorname{tr}\Gamma$ by Lemma A.8. Each later contribution is given exactly by equation 34 with $g(x) = \|x\|^2$ combined with equation 12. Summing the signed endpoint terms telescopes,

$$\sum_{t=s+1}^{e} \frac{\|\xi_t\|^2 - \|\xi_{t-1}\|^2}{2} = \frac{\|\xi_e\|^2 - \|\xi_s\|^2}{2},$$

and adding the initial $\|\xi_s\|^2$ produces the average $(\|\xi_s\|^2 + \|\xi_e\|^2)/2$; the remaining terms are retained as written, which is exactly equation 13. When $s = e$ both sums are empty and equation 13 reduces to $\|\xi_s\|^2 + \operatorname{tr}\Gamma$, consistent with the initial moment.

For the two elementary bounds, $\phi(z) \le \phi(0) = 1/\sqrt{2\pi} = c_g/2$ gives $2\Phi(h/2) - 1 = 2\int_0^{h/2}\phi(z)dz \le h\phi(0) = c_g h/2$, and the same quantity is a total variation distance, hence at most one. Differentiating, $\frac{d}{dh}[h\phi(h/2)] = \phi(h/2)(1 - h^2/4)$, which is nonnegative on $[0,2]$ and nonpositive on $[2,\infty)$, so the maximum of $h\phi(h/2)$ is attained at $h = 2$ and equals $2\phi(1)$. □

We record two consequences of Lemma A.4 that are used in §5.1 and §5.2. The first is the crude, coordinate-free version of the birth cost; the point of its proof is that no factor $\sqrt{d}$ appears, because translating a Gaussian only involves a one-dimensional projection.

Lemma A.10 (Gaussian translation: switching and weighted birth cost)**.** *Let $u, v \in \mathcal{W}$, let q_u, q_v be the densities of $\mathcal{N}(u, \sigma^2 I_d)$ and $\mathcal{N}(v, \sigma^2 I_d)$ with $\sigma > 0$. Then*

$$\mathrm{TV}(q_u, q_v) \le \frac{c_g}{2\sigma}\|u - v\|, \qquad \int \|x - w_0\|^2 \big[q_u(x) - q_v(x)\big]_+ dx \le \frac{c_g}{\sigma}\big[D^2 + (d+1)\sigma^2\big]\|u - v\|. \tag{39}$$

Proof. If $u = v$ both left-hand sides vanish. Otherwise set $\delta = u - v$, $h = \|\delta\| > 0$, $e = \delta/h$ and $\bar{m}_\varrho = v + \varrho\delta$ for $\varrho \in [0,1]$. For fixed x, differentiating the Gaussian density along the segment of means gives $\frac{d}{d\varrho} q_{\bar{m}_\varrho}(x) = \frac{\delta^\top(x - \bar{m}_\varrho)}{\sigma^2} q_{\bar{m}_\varrho}(x)$, so by the fundamental theorem of calculus

$$q_u(x) - q_v(x) = \int_0^1 \frac{\delta^\top(x - \bar{m}_\varrho)}{\sigma^2} q_{\bar{m}_\varrho}(x)\, d\varrho, \qquad |q_u(x) - q_v(x)| \le \int_0^1 \frac{|\delta^\top(x - \bar{m}_\varrho)|}{\sigma^2} q_{\bar{m}_\varrho}(x)\, d\varrho, \tag{40}$$

the second inequality by the triangle inequality for integrals. The right-hand side is nonnegative, so the order of integration in x and ϱ may be exchanged. For fixed ϱ, substituting $x = \bar{m}_\varrho + \sigma z$ and using Lemma A.8,

$$\int \frac{|\delta^\top(x - \bar{m}_\varrho)|}{\sigma^2} q_{\bar{m}_\varrho}(x) dx = \frac{h}{\sigma}\,\mathbb{E}|e^\top Z| = \frac{h c_g}{\sigma}.$$

Integrating over the unit-length interval of ϱ and multiplying by $1/2$ gives the first bound in equation 39; observe that only the one-dimensional projection onto e occurred, which is why no dimensional factor arises.

For the weighted bound, use $[q_u - q_v]_+ \le |q_u - q_v|$, multiply the second relation of equation 40 by $\|x - w_0\|^2$ and integrate; interchanging the nonnegative integrals and substituting as before,

$$\int \|x - w_0\|^2 [q_u(x) - q_v(x)]_+ dx \le \frac{h}{\sigma}\int_0^1 \mathbb{E}\big[\|\bar{m}_\varrho - w_0 + \sigma Z\|^2\, |e^\top Z|\big] d\varrho.$$

Because $\bar{m}_\varrho - w_0 = (1-\varrho)(v - w_0) + \varrho(u - w_0)$, the triangle inequality and Assumption 2 give $\|\bar{m}_\varrho - w_0\| \le (1-\varrho)\|v - w_0\| + \varrho\|u - w_0\| \le D$. The weighted moment identity equation 24 therefore bounds the integrand by $c_g[D^2 + (d+1)\sigma^2]$, and integrating over $\varrho \in [0,1]$ gives the second bound in equation 39. All moments involved were computed to be finite in Lemma A.8. □

Lemma A.11 (Segment statistics of a Gaussian-smoothed comparator)**.** *Fix $u_{1:T} \in \mathcal{W}^T$ and $\sigma > 0$, and couple the marginals $Q_t = \mathcal{N}(u_t, \sigma^2 I_d)$ as in Theorem A.2. Then*

$$\mathbb{E}N(X) \le 1 + \frac{c_g P_T}{2\sigma}, \qquad \mathbb{E}\sum_{t\in\mathcal{J}(X)} \|X_t - w_0\|^2 \le D^2 + d\sigma^2 + \frac{c_g P_T}{\sigma}\big[D^2 + (d+1)\sigma^2\big], \quad (41)$$

and the smoothing bias is exactly

$$\sum_{t=1}^{T} \mathbb{E} f_t(X_t) - \sum_{t=1}^{T} f_t(u_t) = \frac{\sigma^2}{2} S_T. \quad (42)$$

Proof. Since $N(X) = 1 + \sum_{t=2}^{T} \mathbf{1}\{X_t \neq X_{t-1}\}$ identically, taking expectations and using that the switch probability equals the total variation distance, then the first bound of equation 39, gives the first claim in equation 41.

For the second, decompose

$$\sum_{t\in\mathcal{J}(X)} \|X_t - w_0\|^2 = \|X_1 - w_0\|^2 + \sum_{t=2}^{T} \|X_t - w_0\|^2 \mathbf{1}\{X_t \neq X_{t-1}\}.$$

As $X_1 \sim \mathcal{N}(u_1, \sigma^2 I_d)$, Lemma A.8 gives $\mathbb{E}\|X_1 - w_0\|^2 = \|u_1 - w_0\|^2 + d\sigma^2 \le D^2 + d\sigma^2$. For each $t \ge 2$, apply the exact birth identity equation 34 with $g(x) = \|x - w_0\|^2$ and then the second bound of equation 39; summing over t gives the second claim. This argument cannot be replaced by multiplying the expected number of segments by the second moment at a typical round, because the segment starts are selected by the coupling event and the two quantities are not independent.

For equation 42, note that although the X_t are dependent, each marginal is exactly $\mathcal{N}(u_t, \sigma^2 I_d)$, so for fixed t we may write $X_t = u_t + \sigma Z$ in distribution. Expanding the square,

$$f_t(u_t + \sigma Z) = \tfrac{1}{2}(a_t^\top u_t - y_t)^2 + \sigma(a_t^\top u_t - y_t) a_t^\top Z + \tfrac{\sigma^2}{2}(a_t^\top Z)^2.$$

By Lemma A.8 the linear term has zero expectation and $\mathbb{E}(a_t^\top Z)^2 = a_t^\top(\mathbb{E}ZZ^\top)a_t = \|a_t\|^2$, so $\mathbb{E}f_t(X_t) - f_t(u_t) = \frac{\sigma^2}{2}\|a_t\|^2$. Summing over $t = 1, \dots, T$ gives equation 42. This is an identity, not a generic smoothness bound: the rank-one Hessian of the square loss removes any additional dimensional factor. □

Corollary A.6 (Isotropic specialization of the exact birth total)**.** *In the setting of Corollary A.1 with $\Gamma = \sigma^2 I_r$ and $\|\xi_t\| \le D_J$ for all $t \in J$, writing $P = P_J^{\|}$,*

$$\mathsf{M}_J(\sigma^2 I_r) \le D_J^2 + r\sigma^2 + \frac{c_g P}{2\sigma}\big[D_J^2 + (r+1)\sigma^2\big]. \quad (43)$$

Proof. Here $h_t = \|\xi_t - \xi_{t-1}\|/\sigma$, $\kappa_t = \sigma^2$ and $\operatorname{tr}\Gamma = r\sigma^2$. In equation 13 the endpoint term is at most D_J^2, and the averaged endpoint norms multiplying α_t are at most D_J^2, so

$$\mathsf{M}_J \le D_J^2 + r\sigma^2 + (D_J^2 + r\sigma^2)\mathsf{A}_J + \sigma \sum_{t=s+1}^{e} \|\xi_t - \xi_{t-1}\|\, \phi(h_t/2),$$

using $h_t\kappa_t\phi(h_t/2) = \sigma\|\xi_t - \xi_{t-1}\|\phi(h_t/2)$. Applying $\mathsf{A}_J \le c_g P/(2\sigma)$ and $\phi(h_t/2) \le c_g/2$ from Corollary A.1 bounds the right-hand side by $D_J^2 + r\sigma^2 + \frac{c_g P}{2\sigma}(D_J^2 + r\sigma^2) + \frac{c_g \sigma P}{2}$, and combining the last two terms gives equation 43. The coefficient of the variation term is half of what the crude bound of Lemma A.10 would produce, and equation 13 retains the saturation and directional information that equation 43 discards. □

E Proofs for Section 5.1

Proof of Theorem 1. Fixing the data and constructing an auxiliary path. Fix the entire finite sequence $(a_t, y_t)_{t\le T}$, the predictions the algorithm makes on it, and an arbitrary $u_{1:T} \in \mathcal{W}^T$. Fix $\sigma > 0$ and construct $X_{1:T}$ with marginals $\mathcal{N}(u_t, \sigma^2 I_d)$ as in Theorem A.2. Nothing here requires the algorithm to know $u_{1:T}$ or to draw $X_{1:T}$: all randomness lives in the analysis of an already fixed sequence. For a realization, let the segment starts be $1 = \tau_1 < \cdots < \tau_M \le T$ with $\tau_{M+1} = T + 1$ and $M = N(X)$, and put $\theta_j = X_{\tau_j}$, so that $X_t = \theta_j$ for $\tau_j \le t \le \tau_{j+1} - 1$.

Applying the interval guarantee. The general-interval form of Theorem A.1 holds for all real parameters and all integer intervals simultaneously, so it may be applied even though the interval and the parameter were selected by the auxiliary coupling: $\sum_{t=\tau_j}^{\tau_{j+1}-1}[\hat{\ell}_t - f_t(\theta_j)] \le H_T(\mathcal{C}_T + \|\theta_j - w_0\|^2/(2\eta))$. Summing over j, and using that the segments are disjoint and cover $[1,T]$,

$$\sum_{t=1}^{T}\big[\hat{\ell}_t - f_t(X_t)\big] \le H_T\Big[\mathcal{C}_T N(X) + \frac{1}{2\eta}\sum_{t\in\mathcal{J}(X)}\|X_t - w_0\|^2\Big]. \tag{44}$$

The right-hand side has finite expectation by equation 41. The left-hand side does too: $\hat{\ell}_t \le 2B^2$, and each X_t has finite Gaussian second moment, so each $f_t(X_t)$ is integrable. We may therefore take expectations in equation 44.

Returning to the true comparator. Because the algorithm's losses do not depend on the auxiliary randomness once the data is fixed, we have the exact identity

$$R_T(u_{1:T}) = \mathbb{E}\Big[\sum_{t=1}^{T}\big(\hat{\ell}_t - f_t(X_t)\big)\Big] + \sum_{t=1}^{T}\big(\mathbb{E}f_t(X_t) - f_t(u_t)\big). \tag{45}$$

Substituting the expectation of equation 44 into equation 45 and then using the three statements of Lemma A.11 gives, for every $\sigma > 0$,

$$R_T(u_{1:T}) \le H_T\Big\{\mathcal{C}_T\Big(1 + \frac{c_g P_T}{2\sigma}\Big) + \frac{1}{2\eta}\Big[D^2 + d\sigma^2 + \frac{c_g P_T}{\sigma}\big(D^2 + (d+1)\sigma^2\big)\Big]\Big\} + \frac{\sigma^2}{2}S_T. \tag{46}$$

The argument does not use that the comparator is an offline optimum, any first-order condition for it, or independence between the comparator and the data; only the simultaneously valid interval guarantee and the explicitly constructed marginals were needed.

Simplifying for $0 < \sigma \le 1$. Since $c_g \le 1$ and $\sigma^2 \le 1$, the part of equation 46 not multiplying P_T/σ obeys $H_T[\mathcal{C}_T + (D^2 + d\sigma^2)/(2\eta)] \le H_T[\mathcal{C}_T + (D^2 + d + 1)/(2\eta)] = \mathcal{G}_T$, while the coefficient of P_T/σ obeys

$$H_T\Big[\frac{c_g\mathcal{C}_T}{2} + \frac{c_g\big(D^2 + (d+1)\sigma^2\big)}{2\eta}\Big] \le H_T\Big[\mathcal{C}_T + \frac{D^2 + d + 1}{2\eta}\Big] = \mathcal{G}_T.$$

Substituting both into equation 46 and retaining $S_T\sigma^2/2$ gives equation 3.

Handling the two boundary regimes. Since $\hat{y}_t, y_t \in [-B, B]$ we have $0 \le \hat{\ell}_t \le 2B^2$, and $f_t(u_t) \ge 0$, so

$$R_T(u_{1:T}) \le 2B^2 T. \tag{47}$$

If $P_T = 0$, then equation 3 gives $R_T \le \mathcal{G}_T + S_T\sigma^2/2$ for every $\sigma \in (0,1]$; letting $\sigma \downarrow 0$ the right-hand side converges to $\mathcal{G}_T$, so $R_T \le \mathcal{G}_T$. The limit is taken only in the analysis; the algorithm never holds a degenerate posterior and never divides by zero.

Now suppose $P_T > 0$. Since $B > 0$ and $T \ge 1$ we have $\mathcal{V}_T > 0$, and $\Lambda_T > 0$ gives $\mathcal{G}_T > 0$. Put $\rho = (\mathcal{G}_T P_T/\mathcal{V}_T)^{1/3}$ and $\mathsf{M}_T = \mathcal{G}_T^{2/3}\mathcal{V}_T^{1/3}P_T^{2/3}$. If $\rho \le 1$, take $\sigma = \rho$ in equation 3 and use $S_T \le \mathcal{V}_T$:

$$R_T \le \mathcal{G}_T + \frac{\mathcal{G}_T P_T}{\rho} + \frac{\mathcal{V}_T\rho^2}{2} = \mathcal{G}_T + \mathsf{M}_T + \tfrac{1}{2}\mathsf{M}_T = \mathcal{G}_T + \tfrac{3}{2}\mathsf{M}_T,$$

where the middle equality uses $\mathcal{G}_T P_T/(\mathcal{G}_T P_T/\mathcal{V}_T)^{1/3} = (\mathcal{G}_T P_T)^{2/3}\mathcal{V}_T^{1/3}$ and $\mathcal{V}_T(\mathcal{G}_T P_T/\mathcal{V}_T)^{2/3} = (\mathcal{G}_T P_T)^{2/3}\mathcal{V}_T^{1/3}$. If instead $\rho > 1$, then $\mathcal{G}_T P_T > \mathcal{V}_T$, hence $\mathsf{M}_T = (\mathcal{G}_T P_T)^{2/3}\mathcal{V}_T^{1/3} > \mathcal{V}_T \ge 2B^2 T$, and equation 47 gives $R_T \le 2B^2 T < \mathsf{M}_T \le \mathcal{G}_T + \frac{3}{2}\mathsf{M}_T$. Together with the case $P_T = 0$ this establishes the second entry of equation 4, and taking the minimum with equation 47 completes it. This also explains the role of the linear term inside $\mathcal{V}_T$: it serves only to cover the saturated regime in which the optimal analysis scale would exceed one.

Well-defined updates and per-round time. The covariance is initialized at $I_d \succ 0$, and Lemma A.1 shows the rank-one update preserves positive definiteness, so $v_{I,t} \ge 0$ and $c_{I,t} \ge 1$ for all components and rounds. Initial masses are strictly positive, the evidence factors $Z_{I,t}$ are strictly positive, and the mass multiplier is strictly positive, so all live masses stay positive; the active set always contains the permanent component and one singleton, so the normalizing denominator is strictly positive. Every prediction and update is therefore well defined. For a single component,

forming $C_{I,t}a_t$ takes $O(d^2)$ arithmetic operations, the inner products and the mean update take $O(d)$, the rank-one outer product and subtraction take $O(d^2)$, and the evidence and mass updates take $O(1)$; no matrix inverse is ever computed, the inverses in the proofs serving only to verify the rank-one formula. By Lemma A.2 there are $\lfloor \log_2 t \rfloor + 2$ active components, so a round costs $O\big(d^2(\lfloor \log_2 t \rfloor + 2)\big) = O(d^2 \log(t{+}1))$. When t is a power of two the number of births is largest, and even then only $\lfloor \log_2 t \rfloor + 1$ components are created, each initialized with $O(d^2)$ writes, which remains inside the same bound. The estimate is thus worst-case per round and does not rely on amortizing expensive initializations.

Memory and unbounded operation. Each active component stores a $d \times d$ covariance, a d-vector mean, one scalar log-mass and a constant number of interval markers, i.e. $O(d^2)$ cells; the cached $g_{I,t}$ and a few scalars add $O(d)$. Expired components are deleted immediately and no past losses, features or models are retained, while priors of future components are generated on demand from their start and scale. Hence the space is also $O(d^2 \log(t{+}1))$. The interval schedule, the priors and the learning rate contain no reference to the evaluation time T, so a single run satisfies the proven inequalities on every finite prefix and no doubling trick is needed. Finally, if the environment adapts to the algorithm's history it suffices to repeat the above deterministic inequalities on the realized sequence, since no distributional independence was assumed anywhere. □

Proof of Corollary A.2. First control the interval constants. For integers $T \geq 1$ we have $2^{T-1} \geq T$, by induction: equality at $T = 1$, and if it holds at T then $2^T \geq 2T \geq T + 1$. Hence $K_T + 1 = \lfloor \log_2 T \rfloor + 1 \leq T$ and $K_T + 2 \leq T + 1$, so

$$\Lambda_T \leq \log\big[2T^2(T+1)^2\big] \leq \log(8T^4) \leq 4\log(eT) = 4L_T,$$

the last inequality because $\log 8 = 3\log 2 < 3 < 4$. Also $H_T = 2(K_T{+}1) \leq 2\big(1{+}\log T/\log 2\big) \leq h_0 L_T$. From $S_T \leq A^2T$, $d \geq 1$ and $T \geq 1$ we get $1 + \eta S_T/d \leq 1 + \eta A^2 T \leq (1 + \eta A^2)T$, hence $\log(1{+}\eta S_T/d) \leq \log(1{+}\eta A^2){+}\log T \leq [1{+}\log(1{+}\eta A^2)]L_T$. Finally $D^2{+}d{+}1 \leq (D^2{+}2)d$, because the difference of the two sides is $(d-1)(D^2+1) \geq 0$. Substituting these into the definition of $\mathcal{G}_T$ in equation 2 and using $d \geq 1$, $L_T \geq 1$,

$$\mathcal{G}_T \leq \frac{h_0 L_T}{\eta}\Big[4dL_T + \frac{d}{2}\big(1 + \log(1 + \eta A^2)\big)L_T + \frac{(D^2+2)d}{2}L_T\Big] = C_* d L_T^2.$$

Moreover $\mathcal{V}_T = S_T + 2B^2T \leq (A^2 + 2B^2)T$. Inserting these two bounds into the second entry of equation 4 gives equation 14. Since $d^{2/3} \leq d$ and $L_T^{4/3} \leq L_T^2$, the right-hand side of equation 14 is in turn bounded by a constant depending only on A, B, D times $dL_T^2(1 + T^{1/3}P_T^{2/3})$, which is the stated rate; the complexity claim is immediate from Theorem 1. If A, B, D themselves vary with d, the dependence in C_* must be expanded rather than hidden in a "dimension-free" constant. □

Proof of Corollary A.3. For the static bound, take $\theta = u$ in equation 9 and use $\|u{-}w_0\|^2 \leq D^2$ from Assumption 2; the rate $O(d\log(eT))$ then follows from $S_T \leq A^2T$ and the logarithmic estimate established in the proof of Corollary A.2. The permanent component is exactly what preserves this static comparison without passing through an interval cover.

For the piecewise-constant bound, write the maximal constant stretches of $u_{1:T}$ as $[\tau_j, \tau_{j+1} - 1]$ with parameter $\theta_j \in \mathcal{W}$. The general-interval form of Theorem A.1 bounds the regret on each stretch by $H_T(\mathcal{C}_T + \|\theta_j - w_0\|^2/(2\eta)) \leq H_T(\mathcal{C}_T + D^2/(2\eta))$, and summing over the M disjoint stretches gives the claim. All three bounds are upper bounds on the same quantity for the same run of the same algorithm, so their minimum is also an upper bound. □

F Proofs for Section 5.2

Two lemmas prepare Theorem A.4. Throughout, the algorithm is the one of §4 with $\varepsilon = 0$; it knows nothing of the evaluation interval, the feature subspace, the comparator, or the analysis covariance, all of which are chosen in the analysis of an already realized sequence.

Lemma A.12 (Spectral form of the interval guarantee)**.** *Fix T and a nonempty integer interval $J = [s, e] \subseteq [1, T]$, and recall $\Xi_J = \log\det(I_d + \eta G_J)$ and $\mathcal{C}^{\mathrm{sp}}_{J,T} = \eta^{-1}(\Lambda_T + \Xi_J/2)$. For every*

nonempty integer subinterval $K \subseteq J$ and every $\theta \in \mathbb{R}^d$,

$$\sum_{t\in K}\big[\hat{\ell}_t - f_t(\theta)\big] \le h(|J|)\Big(\mathcal{C}^{\mathrm{sp}}_{J,T} + \frac{\|\theta - w_0\|^2}{2\eta}\Big). \tag{48}$$

Proof. We first check a matrix monotonicity property. If $I \subseteq J$ then $G_J - G_I = \sum_{t\in J\setminus I} a_t a_t^\top \succeq 0$. Writing $A_I = I_d + \eta G_I \succ 0$ and factoring,

$$I_d + \eta G_J = A_I^{1/2}\Big[I_d + \eta A_I^{-1/2}(G_J - G_I)A_I^{-1/2}\Big]A_I^{1/2}.$$

The second term inside the bracket is positive semidefinite, since its quadratic form at v is $\eta(A_I^{-1/2}v)^\top(G_J - G_I)(A_I^{-1/2}v) \ge 0$; by the orthogonal diagonalization of Lemma A.9 its eigenvalues are nonnegative, so all eigenvalues of the bracket are at least one and its determinant is at least one. Taking logarithms of determinants,

$$\log\det(I_d + \eta G_I) \le \Xi_J. \tag{49}$$

Using the maximal dyadic construction of Lemma A.2, partition K into disjoint finite dyadic intervals $I_1, \dots, I_m$. That construction contributes at most two intervals per scale, and every interval has length at most $|K| \le |J|$, so only the scales $0, \dots, \lfloor \log_2 |J| \rfloor$ occur and $m \le h(|J|)$. All these intervals lie in $[1, T]$, so their prior penalties are at most Λ_T. Applying equation 8 on each I_i and using equation 49,

$$\sum_{t\in I_i}\big[\hat{\ell}_t - f_t(\theta)\big] \le \frac{\Lambda_T}{\eta} + \frac{\|\theta - w_0\|^2}{2\eta} + \frac{\Xi_J}{2\eta}.$$

Summing over i, the left-hand sides add to the regret on K because the intervals are disjoint and cover it, and the right-hand side is a nonnegative quantity repeated $m \le h(|J|)$ times, which gives equation 48. □

Lemma A.13 (Local observable projection and exact smoothing)**.** *Let $r = r_J = \operatorname{rank} G_J > 0$ and let $U = U_J \in \mathbb{R}^{d\times r}$ have orthonormal columns spanning $\mathcal{S}_J = \operatorname{span}\{a_t : t \in J\}$. With $\xi_t = U^\top(u_t - w_0)$, $b_t = U^\top a_t$, $\bar{G}_J = U^\top G_J U$ and $\Pi_J = UU^\top$,*

$$\begin{aligned} a_t^\top u_t &= a_t^\top(w_0 + U\xi_t), \qquad D_J := \max_{t\in J}\|\xi_t\| \le D, \\ P_J^{\|} &:= \sum_{t=s+1}^{e}\|\xi_t - \xi_{t-1}\| \le \sum_{t=s+1}^{e}\|u_t - u_{t-1}\|, \end{aligned} \tag{50}$$

and, with $S_J = \operatorname{tr} G_J$, $\Xi_J = \log\det(I_r + \eta\bar{G}_J) \le r\log(1 + \eta S_J/r)$. Moreover, for every $\Gamma \succ 0$ and every family of random vectors X_t with mean ξ_t and covariance Γ,

$$\sum_{t\in J}\big\{\mathbb{E}f_t(w_0 + UX_t) - f_t(u_t)\big\} = \tfrac{1}{2}\operatorname{tr}\big(\bar{G}_J\Gamma\big). \tag{51}$$

Proof. We first identify the range of G_J. For any z, $z^\top G_J z = \sum_{t\in J}(a_t^\top z)^2$, which vanishes if and only if $a_t^\top z = 0$ for all $t \in J$. Since G_J is symmetric positive semidefinite, its orthogonal eigendecomposition shows $z^\top G_J z = 0$ if and only if $G_J z = 0$; hence the null space of G_J is $\mathcal{S}_J^\perp$ and its range is $\mathcal{S}_J$.

For any $v \in \mathbb{R}^d$ the residual $v - UU^\top v$ is orthogonal to the column space of U, because $U^\top(v - UU^\top v) = U^\top v - I_r U^\top v = 0$. Since $\|Uz\|^2 = z^\top U^\top U z = \|z\|^2$, the Pythagorean identity gives $\|v\|^2 = \|U^\top v\|^2 + \|v - UU^\top v\|^2$ and therefore $\|U^\top v\| \le \|v\|$. Taking $v = u_t - w_0$ and $v = u_t - u_{t-1}$, and summing in the latter case, gives the last two claims of equation 50.

Because $a_t \in \mathcal{S}_J$ we have $a_t = Ub_t$, so

$$a_t^\top(w_0 + U\xi_t) = a_t^\top w_0 + b_t^\top U^\top(u_t - w_0) = a_t^\top w_0 + a_t^\top(u_t - w_0) = a_t^\top u_t,$$

which is the first claim. Completing U to an orthonormal basis $[U, U_\perp]$ of $\mathbb{R}^d$, the matrix of G_J in that basis is $\operatorname{diag}(\bar{G}_J, 0)$, since every a_t has zero coordinates along $U_\perp$. Hence $\det(I_d + \eta G_J) = \det(I_r + \eta\bar{G}_J)$ and $\operatorname{tr}\bar{G}_J = S_J$, and applying the log-determinant bound of Lemma A.9 to the r nonnegative eigenvalues of $\bar{G}_J$ gives the stated estimate for Ξ_J.

Finally write $X_t = \xi_t + Z_t$ with $\mathbb{E}Z_t = 0$ and $\mathbb{E}Z_t Z_t^\top = \Gamma$, and put $\varpi_t = a_t^\top u_t - y_t$. Expanding the square and using the first claim of equation 50,

$$\mathbb{E}f_t(w_0 + UX_t) = \tfrac{1}{2}\mathbb{E}\big(\varpi_t + b_t^\top Z_t\big)^2 = \tfrac{1}{2}\varpi_t^2 + \varpi_t b_t^\top \mathbb{E}Z_t + \tfrac{1}{2}b_t^\top \mathbb{E}\big[Z_t Z_t^\top\big]b_t = f_t(u_t) + \tfrac{1}{2}b_t^\top \Gamma b_t.$$

Summing over $t \in J$ and expanding entrywise, $\sum_t b_t^\top \Gamma b_t = \sum_{i,j}\Gamma_{ij}\sum_t b_{t,i}b_{t,j} = \operatorname{tr}(\bar{G}_J\Gamma)$, which gives equation 51. □

Proof of Theorem A.4. Fix J and assume first that $r_J > 0$. Define $F_t(x) = f_t(w_0 + U_J x)$ for $x \in \mathbb{R}^{r_J}$ and $t \in J$. Lemma A.12 applies to every $\theta = w_0 + U_J x$, for which $\|\theta - w_0\|^2 = \|x\|^2$, so the hypothesis of Theorem A.3 holds on every subinterval of J with

$$C = h(|J|)\,\mathcal{C}^{\mathrm{sp}}_{J,T}, \qquad \Psi(x) = \frac{h(|J|)}{2\eta}\|x\|^2.$$

This guarantee is valid on all of $\mathbb{R}^{r_J}$. Therefore the full-support Gaussian $q_t = \mathcal{N}(\xi_t, \Gamma)$ is an admissible analysis marginal.

Fix any $\Gamma \succ 0$. All second moments are finite and each F_t is a nonnegative quadratic polynomial, so the integrability required by Theorem A.3 holds. In equation 11 the switching term is computed exactly as $\mathsf{A}_J(\Gamma)$ by Lemma A.4, the initial and newly born second-moment terms are computed exactly as $\mathsf{M}_J(\Gamma)$ by Corollary A.1, and the smoothing term is computed exactly as $\operatorname{tr}(\bar{G}_J\Gamma)/2$ by equation 51. Substituting these three evaluations yields $R_J(u) \le \mathsf{F}_{J,T}(\Gamma)$.

This holds for every positive definite Γ while the left-hand side does not depend on Γ, so $R_J(u)$ is a lower bound for the set $\{\mathsf{F}_{J,T}(\Gamma) : \Gamma \succ 0\}$ and hence does not exceed its infimum. No assumption that an optimal Γ exists is needed, and no singular covariance is ever substituted into a Gaussian density. Since $\hat{y}_t, y_t \in [-B, B]$ gives $\hat{\ell}_t \le 2B^2$ and $f_t(u_t) \ge 0$, we also have $R_J(u) \le 2B^2|J|$, and taking the smaller of the two bounds completes the estimate.

If $r_J = 0$, the null-space identification in the proof of Lemma A.13 forces $a_t = 0$ for every $t \in J$. Then each live component has $\mu = a_t^\top \bar{m} = 0$ and $v = a_t^\top \bar{C} a_t = 0$, so its prediction is zero and so is the aggregate; meanwhile $a_t^\top u_t = 0$. Both losses equal $y_t^2/2$ and therefore $R_J(u) = 0$.

Finally, the underlying interval guarantee is an algebraic inequality holding deterministically on each data sequence, with no probabilistic failure event attached to any interval or comparator, so the derivation is valid for all choices simultaneously on one realized sequence. □

Proof of Corollary A.4. Take $\Gamma = \sigma^2 I_r$ with $0 < \sigma \le 1$ in Theorem A.4 and abbreviate $P = P_T^{\|}$, $D' = D_T^{\|}$, $\mathcal{C} = \mathcal{C}^{\mathrm{sp}}_{J,T}$ and $\mathcal{G} = \mathcal{G}^{\mathrm{sp}}_T$. By Corollary A.1 and Corollary A.6,

$$R_T(u) \le H_T\Big[\mathcal{C}\Big(1 + \frac{c_g P}{2\sigma}\Big) + \frac{D'^2 + r\sigma^2}{2\eta} + \frac{c_g P}{4\eta\sigma}\big\{D'^2 + (r+1)\sigma^2\big\}\Big] + \frac{S_T\sigma^2}{2}.$$

Since $c_g/2 < 1$ and $\sigma^2 \le 1$, the bracketed part not involving P is at most $\mathcal{C} + (D'^2 + r + 1)/(2\eta)$, and the coefficient of P/σ is at most the same quantity. Hence

$$R_T(u) \le \mathcal{G}\Big(1 + \frac{P}{\sigma}\Big) + \frac{S_T\sigma^2}{2}, \qquad 0 < \sigma \le 1. \tag{52}$$

Both $\mathcal{G} > 0$ and $\mathcal{V}_T > 0$. If $P = 0$, substituting $\sigma = 1/n$ for each positive integer n gives $R_T(u) \le \mathcal{G} + S_T/(2n^2)$; were $R_T(u) > \mathcal{G}$, setting $\epsilon = R_T(u) - \mathcal{G} > 0$ and choosing n with $S_T/(2n^2) < \epsilon$ would be contradictory, so $R_T(u) \le \mathcal{G}$.

If $P > 0$, put $\rho = (\mathcal{G}P/\mathcal{V}_T)^{1/3} > 0$. When $\rho \le 1$ we may take $\sigma = \rho$ in equation 52 and compute $\mathcal{G}P/\rho = \mathcal{G}^{2/3}\mathcal{V}_T^{1/3}P^{2/3}$ together with $S_T\rho^2/2 = \frac{S_T}{2\mathcal{V}_T}\mathcal{G}^{2/3}\mathcal{V}_T^{1/3}P^{2/3} \le \frac{1}{2}\mathcal{G}^{2/3}\mathcal{V}_T^{1/3}P^{2/3}$, which gives the second entry of equation 16. When $\rho > 1$ we have $\mathcal{G}P > \mathcal{V}_T$ and hence $\mathcal{G}^{2/3}\mathcal{V}_T^{1/3}P^{2/3} = (\mathcal{G}P)^{2/3}\mathcal{V}_T^{1/3} > \mathcal{V}_T \ge 2B^2T$, so the already established trivial bound $R_T(u) \le 2B^2T$ implies the second entry as well. The three cases cover all $P \ge 0$, and combining with the trivial bound gives equation 16.

To exhibit the hidden factors, let $L = \log(eT) \ge 1$. From $H_T = 2(\lfloor \log_2 T \rfloor + 1)$ we get $H_T \le h_0 L$ with $h_0 = 2(1 + 1/\log 2)$. Also $K_T = \lfloor \log_2 T \rfloor \le T - 1$, verified directly at $T = 1$ and from $2^{T-1} \ge T$ for $T \ge 2$, so $\Lambda_T \le \log(8T^4) \le 5L$. Lemma A.13 gives $\Xi_J \le r\log(1 + \eta S_T/r) \le$

$r\log(1+\eta A^2 T) \le r[\log T + \log(1+\eta A^2)]$, and writing $c_A = \log(1+\eta A^2) \ge 0$ the bracket is at most $r(1+c_A)L$. Using $r \ge 1$, $L \ge 1$ and $D' \le D$, the bracket defining $\mathcal{G}_T^{\mathrm{sp}}$ is at most

$$\frac{5rL}{\eta} + \frac{r(1+c_A)L}{2\eta} + \frac{r(D^2+2)L}{2\eta} = \frac{13 + D^2 + c_A}{2\eta}\, rL,$$

so with $C_{\mathrm{sp}} = \frac{h_0}{2\eta}[13 + D^2 + \log(1+\eta A^2)]$ we obtain $\mathcal{G}_T^{\mathrm{sp}} \le C_{\mathrm{sp}} r L^2$, while $\mathcal{V}_T \le (A^2 + 2B^2)T$. Substituting into equation 16 gives

$$R_T(u) \le C_{\mathrm{sp}} r L^2 + \tfrac{3}{2} C_{\mathrm{sp}}^{2/3} (A^2 + 2B^2)^{1/3} r^{2/3} T^{1/3} (P_T^{\|})^{2/3} L^{4/3}, \tag{53}$$

which is the asserted rate with no constants hidden that depend on d, T, r or P_T. □

Proof of Lemma A.5. The path u itself satisfies the constraints, so $\mathcal{V}_D(u; a)$ is nonempty. For any feasible v the objective is a sum of norms, hence nonnegative, and $\|v_t - v_{t-1}\| \le 2D$ gives objective at most $2D(T-1)$; the infimum is therefore a finite real number.

Choose feasible paths $v^{(n)}$ whose objective is at most the infimum plus $1/n$, and regard each as a vector in $\mathbb{R}^{dT}$. All coordinates are uniformly bounded, so we may extract a subsequence convergent in the first coordinate, then a further subsequence convergent in the second, and so on; since there are only dT coordinates the process terminates and the final subsequence converges in every coordinate, hence in the Euclidean norm of $\mathbb{R}^{dT}$. Call the limit v^*. Norms and inner products are continuous, so each constraint $\|v_t^* - w_0\| \le D$ and $a_t^\top v_t^* = a_t^\top u_t$ survives the limit, and a finite sum of norms is continuous, so the objective of v^* equals the infimum: the minimum is attained.

For the comparison with projection, let Π_T be the orthogonal projection onto $\mathrm{span}\{a_t : t \le T\}$ and set $v_t = w_0 + \Pi_T(u_t - w_0)$. Orthogonal projection does not increase norms, as established via the Pythagorean identity in the proof of Lemma A.13, so $\|v_t - w_0\| \le D$. Since every a_t lies in the projected subspace, $a_t^\top \Pi_T = a_t^\top$, so this path preserves every round's prediction and is feasible; its path length is exactly $P_T^{\|}$. Substituting it into the minimization gives $P_T^{\mathrm{obs}} \le P_T^{\|}$, and applying non-expansiveness to each difference $u_t - u_{t-1}$ gives $P_T^{\|} \le P_T$.

If a parameter θ as in the statement exists, the constant path $v_t = \theta$ is feasible with objective zero, so the minimum is zero. Conversely, if the minimum is zero it is attained at some v^* whose objective is a sum of nonnegative terms, so each $\|v_t^* - v_{t-1}^*\|$ vanishes and all v_t^* coincide; taking $\theta = v_1^*$ gives a parameter with the required properties. □

Proof of Theorem A.5. Let v^* be the minimizing path provided by Lemma A.5. For every t we have $a_t^\top v_t^* = a_t^\top u_t$, so the two comparator sequences incur identical losses and $R_T(v^*) = R_T(u)$.

The proof of Corollary A.4 used no property of the comparator beyond the reference-point radius, so it applies to v^*. The projection of v^* onto the observed subspace has path length at most $\sum_{t=2}^{T} \|v_t^* - v_{t-1}^*\| = P_T^{\mathrm{obs}}$ and radius at most D. The coefficient in equation 16 is increasing in the squared radius, and the map $x \mapsto x + \frac{3}{2} x^{2/3} \mathcal{V}_T^{1/3} P^{2/3}$ is nondecreasing in $x \ge 0$ and in $P \ge 0$ separately, since each nonnegative power function is nondecreasing. We may therefore replace that coefficient by $\mathfrak{G}_T(D, r)$ and the projected path length by P_T^{obs}, which gives equation 17; the explicit constant computation equation 53 does not depend on which candidate path was used, so the asymptotic form follows as well.

When $P_T^{\mathrm{obs}} = 0$, Lemma A.5 supplies a static feasible parameter. For any real θ preserving all the predictions, the permanent-component guarantee equation 9 applied before the trace bound gives the asserted right-hand side without the infimum; the left-hand side does not involve θ, so we may take the infimum over all such θ. Existence of a minimum-norm parameter is not needed here. Finally, if $r = 0$ then every true feature vanishes, so both the algorithm and the comparator predict zero in every round and the regret is zero; no expression requiring division by r is evaluated in that case. □

G Proofs for Section 5.3

Because the learner runs on an affine virtual stream, the interval machinery must first be re-established with an offset. The offset is *not* absorbed into the label, and the following proof makes visible why.

Lemma A.14 (Affine interval guarantee)**.** *Consider scalar offsets $o_t \in \mathbb{R}$ revealed before the label, features $b_t \in \mathbb{R}^k$ and comparator losses $F_t(x) = \frac{1}{2}(o_t + b_t^\top x - y_t)^2$ with $|y_t| \le B$, and run the*

components of §4 initialized at $\mathcal{N}(0, I_k)$ with the priors π_I. Then for every finite dyadic I and every $x \in \mathbb{R}^k$,

$$\sum_{t\in I}\big[\hat{\ell}_t - F_t(x)\big] \le \frac{1}{\eta}\log\frac{1}{\pi_I} + \frac{\|x\|^2}{2\eta} + \frac{1}{2\eta}\log\det\Big(I_k + \eta\sum_{t\in I} b_t b_t^\top\Big), \tag{54}$$

and the general-interval and spectral forms of Lemma A.12 hold with $\sum_t b_t b_t^\top$ in place of G_J and with centre 0. Moreover, for X_t with mean x_t and covariance Γ one has $\mathbb{E}F_t(X_t) - F_t(x_t) = \frac{1}{2}b_t^\top \Gamma b_t$. If every b_t vanishes on I, then $\sum_{t\in I}[\hat{\ell}_t - F_t(x)] \le 0$, but not necessarily zero.

Proof. If $X \sim \mathcal{N}(\beta, K)$ then $o_t + b_t^\top X$ is a one-dimensional Gaussian with mean $\mu = o_t + b_t^\top\beta$ and variance $v = b_t^\top K b_t$, degenerating to the constant o_t when $b_t = 0$. The one-dimensional completion of the square equation 26 in the proof of Lemma A.1 used only the mean and variance of the random scalar, so it applies verbatim and yields the evidence $Z = q^{-1/2}\exp[-\eta(\mu - y_t)^2/(2q)]$ with $q = 1 + \eta v$.

For legality of the emitted scalar, repeating the computation equation 27 with μ and q gives

$$-\frac{1}{\eta}\log Z - \frac{1}{2}\Big(\frac{\mu}{q} - y_t\Big)^2 = \frac{\log q}{2\eta} + \frac{(q-1)\mu^2}{2q^2} - \frac{(q-1)y_t^2}{2q} \;\ge\; \frac{q-1}{2q}\Big(\frac{1}{\eta} - B^2\Big) \;\ge\; 0,$$

using $q \ge 1$, $\log q \ge (q-1)/q$, $y_t^2 \le B^2$ and $1/\eta = 4B^2 \ge B^2$; then clipping to $[-B, B]$ does not increase the loss against y_t by equation 21. Note that no bound on $|y_t - o_t|$ was used anywhere: had we instead shifted the label to $y_t - o_t$ and invoked the label bound, the argument would be invalid, since $|y_t - o_t|$ can exceed B.

For the posterior, the part of the exponent depending on x is $-\frac{1}{2}(x-\beta)^\top K^{-1}(x-\beta) - \frac{\eta}{2}(b_t^\top x + o_t - y_t)^2$, so the new precision is $M = K^{-1} + \eta b_t b_t^\top \succ 0$ and the new linear term is $K^{-1}\beta + \eta(y_t - o_t)b_t$. The direct multiplication performed in the proof of Lemma A.1 shows $M^{-1} = K - \frac{\eta}{q}(Kb_t)(Kb_t)^\top$. Writing $g = Kb_t$ and using $g^\top K^{-1}\beta = b_t^\top\beta$ and $g^\top b_t = v$,

$$M^{-1}\big[K^{-1}\beta + \eta(y_t - o_t)b_t\big] = \beta - \frac{\eta}{q}g\,(b_t^\top\beta) + \frac{\eta}{q}g\,(y_t - o_t) = \beta + \frac{\eta}{q}g\,(y_t - \mu),$$

which is the coordinate mean update, and the covariance update is M^{-1}; both are exactly the operations of Algorithm 1.

All component predictions lie in $[-B, B]$, so equation 22 together with the evidence lower bound gives, exactly as in the proof of Lemma A.3, that the live mass does not increase; the conceptual ledger over the countable family again starts at total mass at most one and changes only finitely many entries per round, so every terminal mass is at most one.

Fix a finite dyadic $I = [s, e]$ and let φ_0 be the standard Gaussian density on $\mathbb{R}^k$. With $\mathcal{E}_n = \int \varphi_0(x)\exp[-\eta\sum_{t=s}^n F_t(x)]dx$ and $\mathcal{E}_{s-1} = 1$, induction on n using the posterior computation shows the round-n evidence is $\mathcal{E}_n/\mathcal{E}_{n-1}$, so the evidences telescope and the terminal mass is $\pi_I\exp(\eta\sum_{t\in I}\hat{\ell}_t)\mathcal{E}_e \le 1$; taking logarithms, $\sum_{t\in I}\hat{\ell}_t \le \eta^{-1}\log(1/\pi_I) - \eta^{-1}\log\mathcal{E}_e$. Setting

$$M_I = I_k + \eta\sum_{t\in I} b_t b_t^\top, \quad h_I = \eta\sum_{t\in I}(y_t - o_t)b_t, \quad \varkappa_I = \eta\sum_{t\in I}(y_t - o_t)^2,$$

the function $J_I(x) = \frac{1}{2}\|x\|^2 + \eta\sum_{t\in I}F_t(x)$ equals $\frac{1}{2}(x^\top M_I x - 2h_I^\top x + \varkappa_I)$, so with $x_* = M_I^{-1}h_I$ we have $J_I(x) = J_I(x_*) + \frac{1}{2}(x - x_*)^\top M_I(x - x_*)$ and hence $J_I(x_*) \le J_I(x)$. Applying equation 25 gives $\mathcal{E}_e = e^{-J_I(x_*)}(\det M_I)^{-1/2}$ and therefore $-\eta^{-1}\log\mathcal{E}_e \le \sum_{t\in I}F_t(x) + \|x\|^2/(2\eta) + \log\det(M_I)/(2\eta)$, which yields equation 54. For a general interval, use the disjoint maximal dyadic cover of Lemma A.2 on each block together with the determinant monotonicity equation 49, exactly as in Lemma A.12.

For the smoothing identity, take X_t with mean x_t and covariance Γ and expand the square; the centred first moment kills the cross term, in whose coefficient the offset is the only place o_t appears, leaving $\mathbb{E}F_t(X_t) - F_t(x_t) = \frac{1}{2}b_t^\top\Gamma b_t$. Consequently Theorems A.2 and A.3 and Lemma A.4 apply unchanged, the objects involved being the same Gaussian densities, whole-space interval costs and integrable quadratics; substituting gives the same local spectral expression as Theorem A.4 with the present features and candidate means.

Finally, if every $b_t = 0$ on I then each component predicts $\mathrm{clip}_B(o_t)$ while each virtual comparator predicts o_t. Clipping can only decrease the loss, so the virtual regret is at most zero; it need not

equal zero, and the equality valid for genuinely zero features in the linear case cannot be copied to the affine one. □

Proof of Theorem A.6. The old basis has orthonormal columns, so $V^\top q = V^\top a_t - V^\top V b = b - b = 0$. When a direction is admitted, the new column q/ρ has unit norm and is orthogonal to the old columns, so the new basis is again orthonormal and the retained subspaces are nested. The current virtual feature always lies in the current retained subspace: on admission it is $a_t = Vb + q$, and on rejection it is Vb.

We show by induction that the retained subspace equals the span of all historical virtual features. Initially both are the zero subspace. If the current residual is rejected, the current virtual feature lies in the old subspace and neither space grows. If it is admitted, $\widetilde{a}_t = a_t$ has coordinate $\rho > \tau_t \geq 0$ along the appended direction, so it does not belong to the old subspace, and adjoining it to the span of the past virtual features adds exactly that direction. This gives the identity for m_t; at most one direction is appended per round and the number of columns cannot exceed the ambient dimension, so $m_t \leq \min(d, t)$.

Now compare the coordinate implementation with the ambient algorithm run on the virtual losses. The initial permanent component is $\mathcal{N}(w_0, I_d)$ in ambient terms, and with empty coordinates the claimed identities read $\bar{m}_I = w_0$ and $\bar{C}_I = I_d$, so they hold. Suppose they hold at the end of the previous round. If a column v is appended, then

$$[V, v]\begin{pmatrix}\beta_I \\ 0\end{pmatrix} = V\beta_I, \qquad [V, v]\Big[\mathrm{diag}(K_I, 1) - I_{m_t}\Big][V, v]^\top = V\big(K_I - I_{m_t-1}\big)V^\top,$$

so extending the coordinates changes neither the represented ambient mean nor the represented ambient covariance. The reason the fresh coordinate may carry unit prior variance is precisely that all historical *virtual* features lie in the old subspace, so the historical virtual likelihood contains no factor along the new direction; the identity just verified is the algebraic expression of that fact, and it does *not* require the historical true features to vanish there. If no column is appended the represented quantities are likewise unchanged, and a newborn component with zero mean and identity coordinate covariance represents $\mathcal{N}(w_0, I_d)$, matching the ambient initialization.

Write $\widetilde{a}_t = V_t b$. From the representation identities,

$$o_t + \widetilde{a}_t^\top(\bar{m}_I - w_0) = o_t + b^\top\beta_I, \qquad \widetilde{a}_t^\top \bar{C}_I \widetilde{a}_t = b^\top K_I b,$$
$$\bar{C}_I\widetilde{a}_t = V_t b + V_t(K_I - I_{m_t})b = V_t K_I b = V_t g_I,$$

so the predictive mean, predictive variance, denominator c_I and clipped prediction all coincide with those of the ambient algorithm. Equal pre-update masses give equal normalized weights and hence an equal aggregate prediction, and the same label then gives equal evidence factors and equal mass multipliers, so the post-update masses agree. Substituting the coordinate updates into the representation identities,

$$w_0 + V_t\beta_I^+ = \bar{m}_I + \frac{\eta}{c_I}\bar{C}_I\widetilde{a}_t(y_t - \mu_I), \qquad I_d + V_t(K_I^+ - I_{m_t})V_t^\top = \bar{C}_I - \frac{\eta}{c_I}\big(\bar{C}_I\widetilde{a}_t\big)\big(\bar{C}_I\widetilde{a}_t\big)^\top,$$

which are exactly the ambient updates. Positive definiteness of the ambient covariance was established in Lemma A.1, and since $K_I = V_t^\top \bar{C}_I V_t$ we get $z^\top K_I z = (V_t z)^\top \bar{C}_I (V_t z) > 0$ for $z \neq 0$. Deletion depends only on the interval endpoint, so both implementations remove the same components, completing the induction.

For the cost, write $m = m_t$. Reading a_t, forming the offset, the norm, the projection $V^\top a_t$ and the residual $a_t - Vb$ takes $O(d(1+m))$ real operations, each of the two matrix products costing $O(dm)$. At most one column is appended per round, at a cost of $O(d)$ in time and space. Extending one component's coordinate covariance to a larger square matrix costs $O(1+m^2)$, and the number of live components is $O(\log(t+1))$ by Lemma A.2, so even a round in which the dimension grows and every component is reallocated costs $O((1+m^2)\log(t+1))$. At most $1 + \lfloor \log_2 t \rfloor$ components are born, each initialized in $O(1+m^2)$; each component's matrix–vector product, quadratic form and rank-one update also cost $O(1+m^2)$; and normalization, scalar evidence updates and deletions cost $O(1)$ per component. Adding these gives the stated time bound, with every operation charged to the round in which it happens. The storage consists of w_0, the current feature and residual in $O(d)$ cells, the basis in $O(dm)$ cells, $O(1+m^2)$ cells per live component together with interval indices, and a constant number of certificate scalars, which gives the same expression. The historical virtual features used in the proof need never be stored. □

Proof of Lemma A.6. The virtual residual in round t is $p_t - y_t + h_t$. Using the identity $(\varrho+h)^2-\varrho^2 = 2\varrho h + h^2$ with $\varrho = p_t - y_t$, dividing by two and summing over t gives the equality in equation 18.

For the inequality, let $\varrho = (p_t - y_t)_{t\le T}$ and $h = (h_t)_{t\le T}$ be vectors in $\mathbb{R}^T$. If either has zero norm the inner product vanishes and the claim is immediate. Otherwise, for every real λ, $0 \le \|\varrho - \lambda h\|^2 = \|\varrho\|^2 - 2\lambda\langle\varrho, h\rangle + \lambda^2\|h\|^2$; substituting $\lambda = \langle\varrho, h\rangle/\|h\|^2$ yields $\langle\varrho, h\rangle^2 \le \|\varrho\|^2\|h\|^2$ and hence $\langle\varrho, h\rangle \le \|\varrho\|\|h\|$. By definition $\|\varrho\|^2 = 2\mathcal{L}_T(u)$ and $\|h\|^2 = E_T(x)$, which gives the stated bound. □

Proof of Theorem A.7. Fix any candidate path x. The algorithm's loss against the true labels does not depend on which comparator we evaluate against, so there is the algebraic identity

$$R_T(u) = \sum_{t=1}^{T}\big[\hat{\ell}_t - \widetilde{f}_t(x_t)\big] + \sum_{t=1}^{T}\big[\widetilde{f}_t(x_t) - f_t(u_t)\big].$$

By Lemma A.14, the virtual stream enjoys the whole-space interval guarantee and the exact smoothing identity, so the scale optimization carried out in the proof of Corollary A.4 applies verbatim on the virtual stream and bounds the first sum by $\mathcal{U}_T(\widetilde{a}, x)$. Lemma A.6 bounds the second sum. The resulting inequality holds for every finite real path x while the left-hand side does not depend on x, so we may pass to the infimum. Since the square loss is nonnegative and both the algorithm's prediction and the label lie in $[-B, B]$, we also have $R_T(u) \le 2B^2T$, and taking the smaller of the two gives equation 19.

Now specialize to $x_t = u_t - w_0$. Because $o_t = a_t^\top w_0$,

$$h_t = a_t^\top w_0 + \widetilde{a}_t^\top(u_t - w_0) - a_t^\top u_t = -\big(e_t^{\mathrm{f}}\big)^\top(u_t - w_0),$$

so the Cauchy–Schwarz inequality and Assumption 2 give $|h_t| \le D\|e_t^{\mathrm{f}}\|$ and hence $E_T(x) \le D^2\Delta_{2,T}$. The map $E \mapsto \sqrt{2\mathcal{L}_T(u)E}+E/2$ is nondecreasing on $[0,\infty)$, so the bound of Lemma A.6 is at most the first branch of $\mathcal{A}_T(D)$.

For the second branch, the true comparator residual obeys

$$|p_t - y_t| \le |o_t| + \|a_t\|\,\|u_t - w_0\| + B \le B + |o_t| + D\|a_t\|.$$

If $D \le 1$ this is at most ς_t, and if $D \ge 1$ it is at most $D\varsigma_t$; in both cases $|p_t - y_t| \le \max\{1, D\}\varsigma_t$. Each cross term in equation 18 is therefore at most $D\max\{1, D\}\varsigma_t\|e_t^{\mathrm{f}}\|$, and summing over t while bounding the quadratic term by $E_T(x) \le D^2\Delta_{2,T}$ gives the second branch. Both branches are valid, so their minimum is, and combining with the trivial loss bound completes the proof. □

Proof of Theorem 2. When a direction is admitted, $e_t^{\mathrm{f}} = 0$. When it is rejected, the residual norm is at most τ_t. In both cases $\varsigma_t\|e_t^{\mathrm{f}}\| \le \varepsilon/[t(t+1)]$. Since $1/[t(t+1)] = 1/t - 1/(t+1)$, the partial sums telescope to $1 - 1/(T+1) \le 1$, so summing gives $W_T \le \varepsilon$. Because $\varsigma_t \ge 1$ we also have $\sum_{t\le T}\|e_t^{\mathrm{f}}\| \le W_T \le \varepsilon$, and expanding the square of a sum of nonnegative numbers,

$$\Big(\sum_{t\le T}\|e_t^{\mathrm{f}}\|\Big)^2 = \Delta_{2,T} + 2\sum_{s<t}\|e_s^{\mathrm{f}}\|\,\|e_t^{\mathrm{f}}\| \;\ge\; \Delta_{2,T},$$

so $\Delta_{2,T} \le \varepsilon^2$. Substituting these two certificates into the second branch of $\mathcal{A}_T(D)$ in Theorem A.7 gives equation 5.

For the asymptotic form take $x_t = u_t - w_0$. The common orthogonal projection onto the final virtual subspace increases neither the radius nor any update, so $D_x \le D$ and $P_x \le P_T$. Each $\widetilde{a}_t$ is either a_t or an orthogonal projection of it, so $\|\widetilde{a}_t\| \le \|a_t\| \le A$ and therefore $\widetilde{S}_T \le A^2T$. For $m > 0$ the nonzero eigenvalues of the virtual Gram matrix sum to $\widetilde{S}_T$, so the log-determinant bound of Lemma A.9 gives $\log\det(I_d + \eta\widetilde{G}_T) \le m\log(1 + \eta A^2T/m)$. The termwise constant computation equation 53 therefore applies with this Gram matrix and radius and yields

$$\mathcal{U}_T(\widetilde{a}, u - w_0) \le C_{A,B,D}\,m\log^2(eT) + \tfrac{3}{2}C_{A,B,D}^{2/3}(A^2 + 2B^2)^{1/3}m^{2/3}T^{1/3}P_T^{2/3}\log^{4/3}(eT),$$

where one may take $C_{A,B,D} = \frac{2(1+1/\log 2)}{2\eta}\big[13 + D^2 + \log(1+\eta A^2)\big]$. Substituting into equation 5 gives the stated rate. If $m = 0$ then Lemma A.14 already shows the virtual regret is at most zero, so no determinant or log-mean bound is needed and no division by m occurs. The complexity claim is Theorem A.6. Finally, all thresholds are formed from the current feature, the reference point, the

label bound and the round index, each available before the label and none depending on an unknown path or on the horizon, so causality holds. □

Proof of Corollary A.5. By Lemma A.5 there is a path $v^* \in \mathcal{V}_D(u; a)$ attaining the minimum. It has exactly the same per-round true predictions as u, so $R_T(v^*) = R_T(u)$; its radius is at most D and its path length is P_T^{obs}. Theorem 2 holds for every real path obeying the radius condition, and the candidate path need not lie in the original convex set, because the comparator guarantee of Lemma A.14 is valid on the whole space.

Apply Theorem 2 to v^*, choosing in its proof the virtual comparator path $x_t = v_t^* - w_0$. The common orthogonal projection onto the final virtual subspace does not increase norms, so $D_x \le D$ and $P_x \le P_T^{\mathrm{obs}}$. The quadratic radius term and the nonnegative power of the path length in equation 5 are both nondecreasing in these quantities, so they may be replaced by D and P_T^{obs}; after the radius substitution the coefficient is exactly $\widetilde{\mathfrak{G}}_T(D, m)$. The representation error term depends only on the common radius D and on ε, not on which observation-equivalent path was selected, which gives equation 20.

If $m = 0$, every virtual comparator path has affine regret at most zero by Lemma A.14, so Theorem 2 directly gives the stated constant error bound. If $m > 0$, the determinant and logarithmic constants already expanded in the proof of Theorem 2 give the asymptotic form. The algorithm never receives v^* or P_T^{obs} as input; they serve only to select one of the simultaneously valid comparison guarantees. □

H Proofs for Section 5.4

Since $|y_t| \le 1$, clipping any algorithm's output to $[-1, 1]$ does not increase its loss, so when proving a lower bound we may assume the outputs are clipped: the original algorithm's regret is at least that of the clipped one, and a uniform bound for the latter transfers. All random variables appearing below are then integrable and no additional moment assumption on the predictions is required.

Lemma A.15 (A two-point mean test)**.** *Let $\epsilon \in (0, 1/2]$ and let $P_\pm$ be the distributions on $\{-1, +1\}$ with means $\pm\epsilon$. If $k\epsilon^2 \le 1/64$, then $\mathrm{TV}(P_+^{\otimes k}, P_-^{\otimes k}) < 1/4$. Consequently, if the unknown sign is equally likely to be positive or negative, every test based on those k samples and on independent extra randomness has average error at least $3/8$.*

Proof. For strictly positive distributions on a finite set define $\chi^2(p\|q) = \sum_x (p(x) - q(x))^2/q(x) = \sum_x p(x)^2/q(x) - 1$, the second form by expanding the square and using $\sum p = \sum q = 1$. Here $P_\pm(+1) = (1 \pm \epsilon)/2$ with the values at -1 interchanged, so

$$\chi^2(P_+\|P_-) = \frac{2\epsilon^2}{1-\epsilon} + \frac{2\epsilon^2}{1+\epsilon} = \frac{4\epsilon^2}{1-\epsilon^2} \le \frac{16}{3}\epsilon^2,$$

the last inequality using $\epsilon \le 1/2$. For product distributions the multiplicative distributivity of finite sums gives $1 + \chi^2(P_+^{\otimes k}\|P_-^{\otimes k}) = \big(\sum_x P_+(x)^2/P_-(x)\big)^k = [1 + \chi^2(P_+\|P_-)]^k$. Since $1 + x \le e^x$ for $x \ge 0$, because every term of the exponential series beyond the constant and linear ones is nonnegative,

$$1 + \chi^2(P_+^{\otimes k}\|P_-^{\otimes k}) \le \exp\big((16/3)k\epsilon^2\big) \le e^{1/12}.$$

Comparing power series termwise gives $e^x \le 1/(1-x)$ for $0 \le x < 1$, so $e^{1/12} \le 12/11$ and the product χ^2 is at most $1/11$. By the Cauchy–Schwarz inequality for finite sums,

$$\begin{aligned}\mathrm{TV}(p, q) &= \frac{1}{2}\sum_x \sqrt{q(x)}\,\frac{|p(x) - q(x)|}{\sqrt{q(x)}} \\ &\le \frac{1}{2}\Big(\sum_x q(x)\Big)^{1/2}\Big(\sum_x \frac{(p(x) - q(x))^2}{q(x)}\Big)^{1/2} = \frac{1}{2}\sqrt{\chi^2(p\|q)},\end{aligned}$$

so the total variation is at most $1/(2\sqrt{11}) < 1/4$. For $k = 0$ the joint sample space is a single point and the distance is zero, which also satisfies the bound.

For the testing statement, let $h(x) \in [0,1]$ be the probability of declaring the sign positive after observing x, the independent randomization having been averaged out. The average error is

$$\frac{1}{2}\sum_x \big[p(x)(1-h(x)) + q(x)h(x)\big] = \frac{1}{2}\Big[1 - \sum_x h(x)\big(p(x)-q(x)\big)\Big].$$

Since $h \le 1$ where $p > q$ and $h \ge 0$ where $p < q$, $\sum_x h(x)(p(x)-q(x)) \le \sum_{x:p(x)>q(x)}(p(x)-q(x)) = \mathrm{TV}(p,q)$, the last equality because $\sum(p-q) = 0$ makes the positive part and the absolute negative part equal. Hence the average error is at least $(1-\mathrm{TV})/2 \ge 3/8$. If further information independent of the unknown sign and of these samples is available, fix it, apply the bound just proved, and take expectations. □

Lemma A.16 (Regret on one unidentifiable block)**.** *Consider a scalar block of length $n \ge 1$ with $a_t = 1$, let the block mean θ be $\pm\epsilon$ with equal probability where $\epsilon = 1/(8\sqrt{n})$, and given θ let the labels be independent ± 1 with conditional mean θ, the block comparator being θ throughout. Then every online algorithm suffers expected regret at least $3/1024$ on the block, even given arbitrary information independent of the block's sign.*

Proof. Before the j-th round of the block the algorithm has seen $k = j-1 \le n-1$ labels generated from this mean, and $k\epsilon^2 \le n\epsilon^2 = 1/64$, so Lemma A.15 applies. Convert the prediction z into a test by declaring the sign positive if $z \ge 0$ and negative otherwise. If $\theta = +\epsilon$ and the test errs then $z < 0$, whence $(z-\theta)^2 \ge \epsilon^2$; if $\theta = -\epsilon$ and the test errs then $z \ge 0$, and again $(z-\theta)^2 \ge \epsilon^2$. Therefore $\mathbb{E}(z-\theta)^2 \ge \epsilon^2\,\mathbb{P}(\text{error}) \ge \frac{3}{8}\epsilon^2$.

Given the mean, the past labels and the current prediction, the current label has conditional mean θ, so expanding the square loss difference,

$$\mathbb{E}\Big[\tfrac{1}{2}(z-y)^2 - \tfrac{1}{2}(\theta-y)^2 \,\Big|\, \theta, \text{past}, z\Big] = \tfrac{1}{2}(z^2-\theta^2) - \theta(z-\theta) = \tfrac{1}{2}(z-\theta)^2.$$

Hence the expected regret in that round is at least $\frac{3}{16}\epsilon^2$, and summing over the n rounds gives $\frac{3}{16}n\epsilon^2 = 3/1024$. Information independent of the sign is handled by conditioning first and averaging afterwards, as in Lemma A.15. □

Proof of Theorem 3. Dynamic part. Put $x = (TP^2)^{1/3}$ and

$$M = \max\{1, \min\{T, \lfloor x \rfloor\}\}, \qquad n = \lfloor T/M \rfloor, \qquad \epsilon = 1/(8\sqrt{n}).$$

Since $1 \le M \le T$ we have $T/M \ge 1$, and for any $z \ge 1$ one has $\lfloor z \rfloor \ge z/2$: this is clear for $1 \le z < 2$ where the left side is one, and for $z \ge 2$ we have $\lfloor z \rfloor \ge z - 1 \ge z/2$. Hence $n \ge T/(2M) \ge 1$.

Form M blocks of length n, each using the independent sign and label mechanism of Lemma A.16, with $a_t = 1$ inside blocks and the block's $\pm\epsilon$ as comparator. On the remaining $T - Mn$ rounds set $a_t = 0$ and $y_t = 0$ and hold the comparator at the last block's value; those rounds contribute regret $\hat{y}_t^2/2 \ge 0$ and add nothing to the path length.

Every comparator has absolute value $\epsilon \le 1/8$, and features and labels obey the unit bounds. Consecutive blocks differ by at most 2ϵ, so on every realization $P_{\text{actual}} \le 2\epsilon(M-1) = (M-1)/(4\sqrt{n})$. If $M = 1$ this is zero. If $M > 1$ then $M \le x$ by construction and $1/\sqrt{n} \le \sqrt{2M/T}$, so

$$P_{\text{actual}} \le \frac{M}{4}\sqrt{\frac{2M}{T}} = \frac{\sqrt{2}}{4}\frac{M^{3/2}}{\sqrt{T}} \le \frac{\sqrt{2}}{4}\frac{x^{3/2}}{\sqrt{T}} = \frac{\sqrt{2}}{4}P \le P.$$

Thus every deterministic sequence in the support of the construction belongs to $\mathcal{Z}_{T,1}(P)$; the path budget is respected pathwise, not merely in expectation.

Apply Lemma A.16 to each block. The data of earlier blocks is independent of the current block's sign, so information gathered from the past does not violate the lemma's hypothesis, and the total expected regret is at least $3M/1024$. We claim $M \ge \frac{1}{2}\min\{T, x\}$: if $x < 2$ the right side is at most one while $M \ge 1$; if $2 \le x < T$ then $M = \lfloor x \rfloor \ge x/2$; and if $x \ge T$ then $\lfloor x \rfloor \ge T$ because T is an integer, so $M = T$. Therefore

$$\mathfrak{R}_{T,1}(P) \ge \frac{3}{2048}\min\{T, T^{1/3}P^{2/3}\}. \tag{55}$$

Passing from the randomized construction to a supremum over deterministic sequences needs no minimax theorem: for any fixed algorithm the supremum over deterministic sequences is at least

the average under any distribution over them, and our distribution yields the same bound for every algorithm, so the bound survives taking the infimum over algorithms.

Static logarithmic part. Take $a_t = 1$ for all t. Draw p uniformly on $[0, 1]$ and set $\theta = 2p - 1$; given p, let the labels be independent with $\mathbb{P}(y_t = +1 \mid p) = p$, and let the comparator be θ throughout. Each realization has zero path length and $|\theta| \le 1$, so the instance is admissible for every $P \ge 0$.

For nonnegative integers a, b put $I_{a,b} = \int_0^1 p^a (1-p)^b \, dp$. For $b = 0$ direct integration gives $I_{a,0} = 1/(a+1)$, and for $b \ge 1$ integration by parts with $dv = p^a \, dp$ and $u = (1-p)^b$ has vanishing boundary terms and yields $I_{a,b} = \frac{b}{a+1} I_{a+1,b-1}$. Applying this b times and substituting $I_{a+b,0} = 1/(a+b+1)$ gives

$$I_{a,b} = \frac{a!\, b!}{(a+b+1)!}. \tag{56}$$

After n labels of which k are positive, the posterior density of p is $p^k(1-p)^{n-k}/I_{k,n-k}$ by the definition of conditional density, so by equation 56,

$$\mathbb{E}[p \mid k] = \frac{I_{k+1,n-k}}{I_{k,n-k}} = \frac{k+1}{n+2}, \qquad \mathbb{E}[p^2 \mid k] = \frac{I_{k+2,n-k}}{I_{k,n-k}} = \frac{(k+1)(k+2)}{(n+2)(n+3)}.$$

Subtracting the square of the mean and placing over a common denominator, the numerator bracket is $(k+2)(n+2) - (k+1)(n+3) = n - k + 1$, so

$$\mathrm{Var}(p \mid k) = \frac{(k+1)(n-k+1)}{(n+2)^2(n+3)}. \tag{57}$$

The marginal probability of exactly k positive labels is $\binom{n}{k} I_{k,n-k} = \frac{n!}{k!(n-k)!} \cdot \frac{k!(n-k)!}{(n+1)!} = \frac{1}{n+1}$. To average equation 57, put $m = n + 1$ and $j = k + 1$ and compute

$$\sum_{k=0}^{n} (k+1)(n-k+1) = \sum_{j=1}^{m} j(m+1-j) = (m+1)\sum_{j=1}^{m} j - \sum_{j=1}^{m} j^2.$$

The identity $\sum_{j \le m} j = m(m+1)/2$ follows by adding the sum to its reversal, and summing $(j+1)^3 - j^3 = 3j^2 + 3j + 1$ from $j = 1$ to m gives $(m+1)^3 - 1$ on the left, whence $\sum_{j \le m} j^2 = m(m+1)(2m+1)/6$ after substituting the first-power sum. Hence the total equals $\frac{m(m+1)}{6}[3(m+1) - (2m+1)] = \frac{m(m+1)(m+2)}{6} = \frac{(n+1)(n+2)(n+3)}{6}$, and therefore

$$\mathbb{E}\,\mathrm{Var}(p \mid \text{first } n \text{ labels}) = \frac{1}{6(n+2)}, \qquad \mathbb{E}\,\mathrm{Var}(\theta \mid \text{first } n \text{ labels}) = \frac{2}{3(n+2)}.$$

Given the history and the algorithm's independent seed the prediction z is fixed, and writing $z - \theta = (z - \mathbb{E}[\theta \mid \text{history}]) + (\mathbb{E}[\theta \mid \text{history}] - \theta)$, whose cross term has zero conditional mean, gives $\mathbb{E}[(z-\theta)^2 \mid \text{history}, \text{seed}] \ge \mathrm{Var}(\theta \mid \text{history})$; the seed is independent of the parameter and of the label mechanism, so it does not alter the posterior. The square-loss conditional identity from the proof of Lemma A.16 applies again, so the expected regret in round $t = n + 1$ is at least $\frac{1}{3(n+2)} = \frac{1}{3(t+1)}$, and

$$\mathfrak{R}_{T,1}(P) \ge \frac{1}{3}\sum_{t=1}^{T} \frac{1}{t+1} \ge \frac{1}{6}\sum_{t=1}^{T} \frac{1}{t} \ge \frac{1}{6}\log(T+1), \tag{58}$$

the second inequality using $t + 1 \le 2t$ and the third comparing $1/x \le 1/t$ on $[t, t+1]$ and summing.

Combining. Write $a = \log(T+1)$, $b = T^{1/3}P^{2/3}$ and $c_0 = 3/2048$. Since $c_0 \le 1/6$, equation 55 and equation 58 give $\mathfrak{R}_{T,1}(P) \ge c_0 \max\{a, \min(T, b)\} \ge \frac{c_0}{2}[a + \min(T, b)]$. If $b \le T$ the bracket is $a + b \ge \min(T, a + b)$; if $b > T$ it is $a + T \ge T = \min(T, a + b)$. Hence $\mathfrak{R}_{T,1}(P) \ge \frac{3}{4096}\min\{T, \log(T+1) + T^{1/3}P^{2/3}\}$. □

Proof of Proposition A.1. Take $M = \lfloor x_r \rfloor \ge x_r/2 \ge 1$ and $n = \lfloor T/(rM) \rfloor$. From $M \le T/(2r^2)$ we get $T/(rM) \ge 2r$, so by the rounding bound used in the proof of Theorem 3, $n \ge T/(2rM) \ge r$. Put $\epsilon = 1/(8\sqrt{n})$.

Each block consists of r groups of n rounds. Before a block begins, choose signs $s_{j,i} \in \{-1, +1\}$ independently for each coordinate and fix the block comparator $u^{(j)} = \epsilon \sum_{i \le r} s_{j,i} e_i$. In group i

the feature is e_i and the labels are independent ± 1 with conditional mean $\epsilon s_{j,i}$; signs and labels of different blocks and coordinates are conditionally independent. The construction uses $rMn \le T$ rounds, and on the remainder the features and labels are zero with the comparator held at the last block's value.

The comparator norm is $\|u^{(j)}\| = \epsilon\sqrt{r} = \sqrt{r}/(8\sqrt{n}) \le 1/8$, using $n \ge r$. Consecutive blocks differ by at most $2\epsilon\sqrt{r}$, so on every realization, using $1/\sqrt{n} \le \sqrt{2rM/T}$ and $rx_r^{3/2} = \sqrt{T}P$,

$$P_{\text{actual}} \le 2\epsilon\sqrt{r}(M-1) \le \frac{\sqrt{r}\,M}{4\sqrt{n}} \le \frac{\sqrt{2}}{4}\frac{rM^{3/2}}{\sqrt{T}} \le \frac{\sqrt{2}}{4}\frac{rx_r^{3/2}}{\sqrt{T}} = \frac{\sqrt{2}}{4}P \le P.$$

The features are the first r standard basis vectors, each occurring at least once, so the feature rank is exactly r and every feature has unit norm. The information from other coordinates and other blocks is independent of the current coordinate's sign, so each of the rM groups satisfies the hypothesis of Lemma A.16; summing over groups, the expected regret is at least $3rM/1024$, and the remaining rounds contribute nonnegatively. Since $rM \ge rx_r/2$ and $rx_r = r^{1/3}T^{1/3}P^{2/3}$, we obtain $\mathbb{E}R_T \ge \frac{3}{2048}r^{1/3}T^{1/3}P^{2/3}$, and the supremum over deterministic sequences is at least the average of the construction. □

Proof of Theorem A.8. Write the admitted true features as $z_j = a_{t_j}$ and put $A_j = \lambda I_d + \sum_{i\le j} z_i z_i^\top$ with $A_0 = \lambda I_d$. Every admitted feature is retained in full, whereas a rejected virtual feature lies in the subspace retained at that moment; hence the retained subspace immediately before z_j is admitted is exactly $\operatorname{span}\{z_1, \dots, z_{j-1}\}$. Split $z_j = p_j + q_j$ with p_j in that subspace and q_j in its orthogonal complement, so $\|q_j\| = \rho_{t_j}$.

The matrix A_{j-1} leaves both subspaces invariant: each rank-one term $z_i z_i^\top$ vanishes on the complement, and the identity acts as the identity on each. Hence A_{j-1}^{-1} equals $\lambda^{-1}I$ on the complement with no cross terms, and by positive definiteness

$$z_j^\top A_{j-1}^{-1} z_j = p_j^\top A_{j-1}^{-1} p_j + \frac{\|q_j\|^2}{\lambda} \;\ge\; \frac{\rho_{t_j}^2}{\lambda}.$$

For the determinant recursion, set $v = A_{j-1}^{-1/2} z_j$, so that $A_j = A_{j-1}^{1/2}(I_d + vv^\top)A_{j-1}^{1/2}$. If $v = 0$ the middle factor has determinant one; if $v \ne 0$ then $I_d + vv^\top$ has eigenvalue $1 + \|v\|^2$ along $v/\|v\|$ and eigenvalue one on the orthogonal complement. Either way $\det A_j = \det A_{j-1}\big(1 + z_j^\top A_{j-1}^{-1} z_j\big)$. Taking logarithms, summing over $j = 1, \dots, m$ and using the quadratic-form bound above,

$$\sum_{j=1}^{m} \log\big(1 + \rho_{t_j}^2/\lambda\big) \le \log\det A_m - \log\det A_0 = \log\det\Big(I_d + \lambda^{-1}\sum_{j\le m} z_j z_j^\top\Big).$$

The admitted features are a subset of all the features, so $G_T - \sum_j z_j z_j^\top \succeq 0$ and the determinant monotonicity equation 49 gives the first claim. If every admission threshold is at least τ, then $\rho_{t_j} > \tau$ and monotonicity of the logarithm bounds the left side below by $m\log(1 + \tau^2/\lambda)$, which is the second claim. □

Proof of Proposition A.2. With $w_0 = 0$ the offsets vanish, so $\varsigma_1 = 1 + B + \|a_1\| = 2 + B$ and $\tau_1 = \varepsilon/[2(B+2)] < 1$, whereas the first residual is e_1 itself with norm one; the first direction is therefore admitted. Suppose only e_1 has been retained so far. Then the round-t residual is $\delta_t e_t$, and since $\delta_t < 1$ we have $\|a_t\| = \sqrt{1 + \delta_t^2} \le \sqrt{2}$, so $\varsigma_t = 1 + B + \|a_t\| \le 1 + B + \sqrt{2} < 4(B+2)$ and consequently

$$\delta_t = \frac{\varepsilon}{4(B+2)t(t+1)} < \frac{\varepsilon}{\varsigma_t\, t(t+1)} = \tau_t.$$

The residual is rejected and the retained subspace is unchanged; by induction $m_t = 1$ and $\widetilde{a}_t = e_1$ for every round. The basis is $d \times 1$ and each component stores a 1×1 matrix, so with $O(\log(t{+}1))$ live components the memory is $O(d + \log(t{+}1))$.

The true features are nonetheless linearly independent. Suppose $\sum_{t=1}^{T} \gamma_t a_t = 0$. Inspecting coordinate t for $t \ge 2$ gives $\gamma_t \delta_t = 0$, and $\delta_t > 0$ forces $\gamma_t = 0$; the first coordinate then gives $\gamma_1 = 0$. Hence the true feature rank is T, whereas exact-rank tracking would store a matrix of dimension T.

For the converse construction, take $a_t = e_t$ for $1 \le t \le \min(d, T)$. Each residual has norm one while the thresholds are smaller than one, so every direction is admitted and $m_t = \min(t, d)$.

This rules out any claim that the retained dimension is small for arbitrary data; it is not, however, a memory lower bound for other algorithms, since the growth of one particular method is not an information-theoretic obstruction. □

Proof of Proposition A.3. Fix $0 < \delta \le 1$, let the true feature be $a_t = \delta$ in every round, suppose every direction is discarded so that the virtual feature is zero, and take $u_t = 1$ and $y_t = 1$. All the boundedness conditions hold and the comparator path length is zero. The true comparator prediction is δ and the virtual one is zero, so

$$\mathcal{L}_T(u) = \frac{T}{2}(1-\delta)^2, \qquad \Delta_{2,T} = E_T(u) = T\delta^2,$$

and the actual loss difference is $\frac{T}{2}[1-(1-\delta)^2] = T\delta - \frac{T}{2}\delta^2$. Meanwhile the right-hand side of equation 18 equals

$$\sqrt{T(1-\delta)^2 \cdot T\delta^2} + \frac{T}{2}\delta^2 = T\delta(1-\delta) + \frac{T}{2}\delta^2 = T\delta - \frac{T}{2}\delta^2,$$

so the inequality is attained with equality. The ratio of the loss difference to $\Delta_{2,T}$ is $1/\delta - 1/2$, which is unbounded as $\delta \downarrow 0$, so no uniform constant C can exist.

The same instance also refutes an unconditional claim that a fixed nonzero tolerance leaves the static logarithmic guarantee intact. Running the fully discarding virtual algorithm, the offset is zero, every component predicts zero and so does the aggregate, so the regret against the true static comparator equals the loss difference above, which is of order $T\delta$. Note finally that this counterexample targets arbitrary discarding rules; the summable threshold of Algorithm 1 decreases with t and would eventually admit this direction. □

Proof of Proposition A.4. The stated representation, with positive weights. At $t = 1$ the density is $\mathcal{N}(0,1)$, which is the claimed form. Suppose the round-t density is $\sum_{s\le t} \omega_{s,t}\varphi(\cdot; 0, v_{s,t})$ with $v_{s,t} = [1+\eta(t-s)]^{-1}$, all $\omega_{s,t} > 0$. Multiplying by the round's likelihood $\exp(-\eta x^2/2)$ leaves all means at zero and increases each component's precision from $1+\eta(t-s)$ to $1+\eta(t+1-s)$; the associated evidence is a strictly positive Gaussian integral, so after normalization every existing weight remains strictly positive. Multiplying this posterior by $1-\mu > 0$ and injecting mass $\mu > 0$ of $\mathcal{N}(0,1)$ produces exactly the claimed form for $s = 1, \dots, t+1$. Since $\eta > 0$, the variances $v_{s,t}$ are pairwise distinct.

Any positive Gaussian representation needs at least t terms. Suppose a mixture with N pairwise distinct positive variances v_i and weights $p_i > 0$ coincides as a density with a mixture of K nondegenerate Gaussians of weights $q_j > 0$, means $m_j \in \mathbb{R}$ and variances $w_j > 0$. Equality of densities implies equality of moment generating functions at every real z. Completing the square gives

$$\int e^{zx}\varphi(x; m, v)\,dx = e^{mz+vz^2/2}\int \frac{1}{\sqrt{2\pi v}}\exp\Big[-\frac{(x-m-vz)^2}{2v}\Big]dx = e^{mz+vz^2/2},$$

and finite mixtures may be integrated termwise, so

$$\sum_{i=1}^{N} p_i e^{v_i z^2/2} = \sum_{j=1}^{K} q_j e^{m_j z + w_j z^2/2}. \tag{59}$$

Let $v_* = \max_i v_i$ and $w_* = \max_j w_j$. If $w_* > v_*$, divide equation 59 by $e^{v_* z^2/2}$ and let $z \to +\infty$: the left side tends to the finite positive total weight of the components with $v_i = v_*$, while on the right at least one term $q_j \exp[m_j z + (w_* - v_*)z^2/2]$ tends to infinity, the positive quadratic dominating the fixed linear term. If $w_* < v_*$, the same division makes every right-hand term tend to zero while the left side tends to a positive limit. Both are contradictions, so $w_* = v_*$.

Continue dividing by $e^{v_* z^2/2}$. Every term with $w_j < v_*$ tends to zero as $z \to \pm\infty$. If some term has $w_j = v_*$ and $m_j > 0$ it tends to infinity as $z \to +\infty$, and if some term has $w_j = v_*$ and $m_j < 0$ it tends to infinity as $z \to -\infty$; positivity of all weights precludes cancellation. Hence every maximal-variance term on the right has $m_j = 0$, and taking limits shows their weights sum exactly to the total left-hand weight at v_*.

Subtract from both sides the left-hand maximal-variance term together with all those right-hand maximal-variance zero-mean terms. As functions these are identical, so what remains is again an identity of the form equation 59 with all retained weights positive. If the left side still has terms

while the right side is empty, evaluating at $z = 0$ contradicts positivity of the left side. Otherwise repeat the maximal-variance argument. Each distinct left-hand variance thus consumes at least one right-hand component, and after N iterations $K \geq N$. Taking $N = t$ proves the claim. □

I Additional Experimental Results

Table 3: The two separation instances. Left: a stream of full observed rank on which the comparator path length grows linearly while the observation-equivalent path length is zero, so the refined guarantee of Theorem A.5 is $1126\times$ tighter than the generic one and the achieved regret is 5.7% of it. Right: a stream whose true feature rank is T but whose causal gate retains a single direction, giving $2761\times$ less state than exact-rank tracking at a regret difference of 1.4×10^{-5}, well inside the certified $D \max\{1, D\} W_T + \frac{1}{2} D^2 \Delta_{2,T} = 0.124$.

E2: observation equivalence		**E3: lossy representation**	
rounds T	20,000	rounds T = dimension d	300
observed rank r	2	true feature rank	300
path length P_T	19,999	retained dim. m_T (ours)	**1**
P_T^{obs}	**0**	retained dim. at ε=0	300
generic P_T bound	40,000	state, ours (reals)	**327**
refined P_T^{obs} bound	35.5	state, ε=0 (reals)	902,709
achieved regret	**2.03**	regret, ours	1.4049
bound ratio	$1126\times$	regret, ε=0	1.4049
		certificate $W_T \leq \varepsilon$	$0.124 \leq 1$