
Fast Best-in-Class Regret for Contextual Bandits

Samuel Girard¹ Aurélien Bibaut² Jill-Jênn Vie¹ Arthur Gretton³ Nathan Kallus^{2,4} Houssam Zenati³

¹INRIA Soda

²Netflix

³Gatsby Computational Neuroscience Unit, University College London

⁴Cornell University and Cornell Tech

Abstract

We study the problem of stochastic contextual bandits in the agnostic setting, where the goal is to compete with the best policy in a given class without assuming realizability or imposing model restrictions on losses or rewards. In this work, we propose *Online Pessimistic Policy Learning* and establish the first fast rate for regret relative to the best-in-class policy. Our proposed algorithm updates the policy at every round by minimizing a pessimistic objective, defined as a clipped inverse-propensity estimate of the policy value plus a variance penalty. By leveraging entropy assumptions on the policy class and a Hölderian error-bound condition (a generalization of the margin condition), we achieve fast best-in-class regret rates, including polylogarithmic rates in the parametric case. Our analysis is driven by a novel sequential self-normalized maximal inequality for bounded martingale empirical processes, which yields uniform variance-adaptive confidence bounds and guarantees pessimism under adaptive data collection.

1 INTRODUCTION

Personalizing decisions to individuals is crucial across many user-facing domains, from improving health outcomes in medicine to increasing revenue in online advertising. In a new decision-making environment, one must learn how to personalize while making decisions that minimize loss, or maximize reward. This is the contextual bandit setting: contexts, representing individual information, arrive sequentially, and at each time the learner must choose an action, balancing good present decisions with learning for the future. In the stochastic variant, these instances are drawn iid from a fixed but unknown joint distribution over contexts and potential losses, one for each action.

A primary aim is sublinear regret, so that long-run time average loss converges to the average attainable if no learning were needed. To achieve sublinear regret, the literature on stochastic contextual bandits generally addresses two settings: value-based or policy-based. In the value-based setting, we assume the context-conditional expected loss function belongs to a given function class, whether linear [2, 6, 30, 40, 56], non-parametric [33, 48, 52], or a generic class with a regression oracle [22, 24, 26, 47, 58]. Regret in this setting corresponds to comparing against the personalization rule that minimizes the context-conditional expected loss.

Notably, regret rates in the value-based setting can be significantly sped up by accounting for the *margin condition* satisfied by the environment distribution [6, 7, 30, 33, 48, 52]: the less mass the context distribution places on hard contexts—those where the best and second-best actions have nearly equal loss—the faster the attainable regret.

In the policy-based setting, we are given a class of policies Π , make no realizability assumptions on the loss model, and instead consider competing with the best policy in the class [3, 9, 17, 23]. This approach is *agnostic* and *model-free*, imposing no assumptions on the context-loss joint distribution and competing with what is achievable inside the policy class. However, unlike value-based approaches, existing policy-based bandit algorithms do not enjoy *fast* best-in-class rates under low-noise / margin-like assumptions.

In this paper, we introduce *Online Pessimistic Policy Learning* (O2PL) and provide the first *fast* rates for best-in-class regret in stochastic contextual bandits in the model-agnostic policy-based setting. Our speedup depends on a Hölderian error bound (HEB), which generalizes the margin condition beyond the realizable setting. Our algorithm performs pessimistic policy learning at every round, minimizing over Π a clipped inverse-propensity estimate of policy value plus a variance-aware penalty. We do not impose *uniform* overlap—a floor $\pi(a \mid x) \geq \eta > 0$ that would bound the importance ratios outright; clipping instead controls large ratios adap-

tively. Our main technical contribution is a novel uniform, Bernstein-type maximal inequality for martingale empirical processes with bounded increments, localized by the *realized* rather than the predictable quadratic variation, which yields confidence radii that adapt to an empirical (clipped) importance-weight variance proxy. We expect this inequality to be of independent interest. The one quantity we leave *unclipped* is the importance-weight variance in the HEB, since it is precisely what governs the clipping bias.

Organization. Section 2 reviews related work, and Section 3 formalizes the agnostic best-in-class contextual bandit setting and assumptions. Section 4 presents O2PL, our clipped IPW algorithm with a variance-adaptive pessimistic penalty, and states its regret guarantee. Section 5 gives a master theorem for any algorithm satisfying abstract surrogate-risk, bias, and variance conditions; Section 6 then verifies these conditions for O2PL, using our new self-normalized maximal inequality. Section 7 reports experiments.

2 RELATED LITERATURE

We focus on the policy-based, best-in-class strand of contextual bandits. In adversarial settings, expert aggregation and exponential weights [5, 41, 64] led to Exp4-style algorithms and chaining-based nonparametric online learning [14, 15, 20, 43]. In stochastic settings, oracle-efficient methods progressed from epoch-greedy [39] to regret-optimal approaches based on cost-sensitive classification oracles [3, 17, 22, 23, 25].

Off-policy evaluation relies on inverse propensity weighting (IPW) [32, 55] and refinements such as doubly robust estimators [18, 19, 54, 65]. Policy learning optimizes such estimators over a policy class, often with normalization or variance control [4, 31, 34–38, 60, 68, 71]. For adaptively collected logs, risk guarantees were established by Bibaut et al. [10], Zhan et al. [70].

Our analysis is rooted in maximal inequalities from empirical-process theory [20, 61, 62]. For sequential dependence, Freedman-type martingale Bernstein inequalities [28] are central. We require a *uniform* self-normalized inequality over policy classes under adaptive sampling. Our development is closest in spirit to sequential complexity tools [50] and modern anytime self-normalized inference [66, 67], but adapted to clipped importance-weight losses and logit-scale entropy.

3 PROBLEM SETUP

We consider the stochastic contextual bandit setting: at each round $t = 1, \dots, T$, the learner observes a context $X_t \in \mathcal{X}$, selects a policy $\pi_t(\cdot | X_t) \in \Pi$, samples an action $A_t \sim \pi_t(\cdot | X_t)$, and incurs a loss $Y_t \in [-1, 0]$. We denote

$Z_t = (X_t, A_t, Y_t)$, and let $\mathcal{F}_{t-1} := \sigma(Z_1, \dots, Z_{t-1})$ be the sigma-field induced by $Z_1, \dots, Z_{t-1}$. We take the context space $\mathcal{X}$ to be a standard Borel space, the action set $\mathcal{A}$ to be finite or standard Borel, and policies $\pi(\cdot | x)$ to be measurable stochastic kernels on $\mathcal{A}$. Contexts are i.i.d. draws $X_t \sim P_X$ from a fixed distribution P_X on $\mathcal{X}$, and losses are generated by a fixed conditional kernel, so that the law of $Y_t | X_t = x, A_t = a$ is independent of t ; we write $\mu(x, a) := \mathbb{E}[Y | X = x, A = a]$ for its conditional mean. The policy π_t is $\mathcal{F}_{t-1}$ -measurable, i.e. determined only by data observed so far.

We are given a policy class Π and consider the cumulative pseudo-regret against the best-in-class policy:

$$\text{Regret}_T = \sum_{t=1}^T \left(R(\pi_t) - \inf_{\pi \in \Pi} R(\pi) \right), \quad (1)$$

where

$$R(\pi) = \mathbb{E}_{X \sim P_X} \mathbb{E}_{A \sim \pi(\cdot | X)} \mu(X, A), \quad (2)$$

and we write $\pi^* \in \arg \min_{\pi \in \Pi} R(\pi)$ for a best-in-class policy.

To obtain sublinear regret, we impose structure on the policy class Π together with a low-noise condition relating excess risk to the variance of importance weights; we do not require any realizability assumption on μ . The range $[-1, 0]$ is only a normalization: any bounded loss interval is admissible, with constants scaling in the range; we do not treat heavy-tailed or unbounded losses.

4 ONLINE PESSIMISTIC POLICY LEARNING

We now present our algorithm, *Online Pessimistic Policy Learning* (O2PL). The design principle is simple: at each round, optimize a lower confidence bound on a centered policy value, where uncertainty is measured by an empirical variance proxy. This yields conservative updates when coverage is poor, and aggressive updates when the data support a policy well.

Define the weight-truncated IPW counterfactual loss:

$$\ell_{t,\alpha}(\pi)(z) := \left\{ \min \left(\frac{\pi(a | x)}{\pi_t(a | x)}, \alpha \right) - 1 \right\} y. \quad (3)$$

We will later choose the truncation parameter α to grow with t to ensure truncation introduces only a vanishing bias.

Given this, define the centered clipped surrogate target

$$R_{t,\alpha}(\pi) := \frac{1}{t} \sum_{s=1}^t \mathbb{E}[\ell_{s,\alpha}(\pi)(Z_s) | \mathcal{F}_{s-1}], \quad (4)$$

$$R_t(\pi) := \frac{1}{t} \sum_{s=1}^t \mathbb{E} \left[\left(\frac{\pi}{\pi_s}(A_s | X_s) - 1 \right) Y_s | \mathcal{F}_{s-1} \right]. \quad (5)$$

The unclipped centered surrogate differs from the true risk only by an additive baseline:

$$R_t(\pi) = R(\pi) - \frac{1}{t} \sum_{s=1}^t R(\pi_s),$$

so that $R_t(\pi) - R_t(\pi^*) = R(\pi) - R(\pi^*)$.

We estimate this centered surrogate with the empirical clipped objective

$$\widehat{R}_{t,\alpha}(\pi) := \frac{1}{t} \sum_{s=1}^t \ell_{s,\alpha}(\pi)(Z_s), \quad (6)$$

and penalize using the clipped empirical variance proxy

$$\widehat{\sigma}_{t,\alpha}^2(\pi) := \frac{1}{t} \sum_{s=1}^t \left(\min \left(\alpha, \frac{\pi}{\pi_s}(A_s | X_s) \right) - 1 \right)^2 \quad (7)$$

For the theoretical analysis, we also define the unclipped predictable variance

$$\sigma_s^2(\pi) := \mathbb{E} \left[\left(\frac{\pi}{\pi_s}(A_s | X_s) - 1 \right)^2 \middle| \mathcal{F}_{s-1} \right], \quad (8)$$

$$\bar{\sigma}_t^2(\pi) := \frac{1}{t} \sum_{s=1}^t \sigma_s^2(\pi), \quad (9)$$

and abbreviate $\sigma_s^2 := \sigma_s^2(\pi^*)$, $\bar{\sigma}_t^2 := \bar{\sigma}_t^2(\pi^*)$. The unclipped variance $\bar{\sigma}_t^2$ is the natural quantity appearing in the Hölderian error bound (Condition 4.3) and in the clipping bias analysis, while the clipped proxy $\widehat{\sigma}_{t,\alpha}$ is what the algorithm can compute and what controls the empirical process.

O2PL relies at each time t on a pessimistic policy-update oracle PPUO_t : it returns the policy minimizing an upper confidence bound on the centered risk, namely the empirical estimate $\widehat{R}_t$ plus a variance-aware penalty Φ_p .

Definition 4.1 (Optimization oracle). The function $\text{PPUO}_t : (0, \infty) \times (0, 1) \times \mathcal{Z}^t \times \Pi^t \rightarrow \Pi$ is a pessimistic policy update oracle that maps a clipping level $\alpha > 0$, a confidence level $\delta \in (0, 1)$, and a sequence of observations $Z_1, \dots, Z_t \in \mathcal{Z}$ and policies $\pi_1, \dots, \pi_t \in \Pi$ to an element of

$$\arg \min_{\pi \in \Pi} \widehat{R}_t(\pi) + \Phi_p(\widehat{\sigma}_{t,\alpha}(\pi), \alpha, t, \delta), \quad (10)$$

where

$$\begin{aligned} \Phi_p(\sigma, \alpha, t, \delta) &:= \left(\frac{\alpha}{t} \right)^{2/(2+p)} + \frac{\sigma^{1-p/2}}{\sqrt{t}} \\ &\quad + \sigma \sqrt{\frac{\log(1/\delta)}{t}} + \frac{\alpha \log(1/\delta)}{t}. \end{aligned}$$

with the convention that $\sigma^{1-p/2} \equiv \sigma \cdot \text{polylog}(e/\sigma)$ when $p = 0$.

Algorithm 1 Online Pessimistic Policy Learning (O2PL)

Require: $T \geq 1$, $p \in [0, 2)$, $\beta > 0$, $\pi_0 \in \Pi$, $\delta \in (0, 1)$, $(\text{PPUO}_t)_{t=1}^T$.
Initialize $\pi_1 \leftarrow \pi_0$.
for $t = 1, \dots, T$ **do**
 Set α_t as in Theorem 4.4,
 $\delta_t \leftarrow \delta/(t(t+1))$,
 Receive context X_t , sample $A_t \sim \pi_t(\cdot | X_t)$, receive loss Y_t ,
 $\pi_{t+1} \leftarrow \text{PPUO}_t(\alpha_t, \delta_t, (Z_1, \dots, Z_t), (\pi_1, \dots, \pi_t))$,
end for

4.1 MAIN REGRET BOUND

We now present the main bound on the regret of O2PL to the best in class. Our result depends on the following two conditions which are discussed further in Section 4.2 and 4.3.

Condition 4.2 (Covering entropy of policy class on the logit scale). There exists $p \in [0, 2)$ such that for every $\epsilon \in (0, 1]$, $\log N(\epsilon, \log \Pi, \|\cdot\|_\infty) \lesssim \log(e/\epsilon)$ if $p = 0$, and $\lesssim \epsilon^{-p}$ if $p \in (0, 2)$.

Condition 4.3 (Hölderian policy error bound). There exists $\beta \in (0, 1]$ such that for every $\pi \in \Pi$,

$$\text{Var}_{X \sim P_X, A \sim \pi(\cdot | X)} \left(\frac{\pi^*(A | X)}{\pi(A | X)} \right) \lesssim (R(\pi) - R(\pi^*))^\beta. \quad (11)$$

Theorem 4.4. Suppose that Conditions 4.2 and 4.3 hold. Then the sequence of policies generated by Algorithm 1 satisfies with probability at least $1 - \delta$: if $\beta = 1$, for $\alpha_t := 1 + \log(et)$,

$$\text{Regret}_T = \begin{cases} \mathcal{O}(\log^2 T) & \text{for } p = 0, \\ \tilde{\mathcal{O}}\left(T^{\frac{p}{2+p}}\right) & \text{for } p \in (0, 2), \end{cases} \quad (12)$$

and, if $0 < \beta < 1$, for $\alpha_t := 2 \vee t^\gamma$ with $\gamma := \frac{2(1-\beta)}{2(2-\beta)+p}$,

$$\text{Regret}_T = \tilde{\mathcal{O}}\left(T^{\frac{2(1-\beta)+p}{2(2-\beta)+p}}\right), \quad (13)$$

where the $\tilde{\mathcal{O}}$ notation absorbs polylogarithmic factors in T and in $1/\delta$ (including the parametric $p = 0$ entropy logarithms).

The theorem does not require any overlap assumption and formalizes the central message of the paper: under entropy control and a Hölderian error-bound condition, policy-based agnostic contextual bandits admit fast best-in-class rates, with polylogarithmic regret in the parametric case.

On the overlap assumption. When the learner requires uniform overlap, $\pi_t(a \mid x) \geq \eta > 0$ for all t, a, x , the importance ratios are bounded by η^{-1} , so one may fix $\alpha_t \equiv \eta^{-1}$; clipping is then inactive and the surrogate coincides with the unbiased centered risk. Writing $\bar{\Delta}_t := \frac{1}{t} \sum_{s=1}^t (R(\pi_s) - R(\pi^*))$, the Hölderian bound and Jensen give $\bar{\sigma}_t^2 \lesssim \bar{\Delta}_t^\beta$, and the resulting fixed-point relation solves to $\text{Regret}_T = \tilde{O}(T^{\frac{2(1-\beta)+\beta p}{2(2-\beta)+\beta p}})$.

This is exactly the classical fast-rate exponent for full-information ERM under margin/error-bound conditions [44]: identifying $r = p/2$, $\theta = 1/\beta$ recovers $n^{-1/(2-\beta+\beta p/2)}$, reducing to $n^{-2/(2+p)}$ when $\beta = 1$ and to Stone’s rate $n^{-2k/(2k+d)}$ for smooth classes with $p = d/k$ [51, 59]. Thus, under overlap, partial monitoring carries no additional polynomial price. Without overlap one must let $\alpha_t \rightarrow \infty$ to control clipping bias, and this extra bias–variance tradeoff is exactly what yields the slower exponent in Theorem 4.4.

4.2 DISCUSSION OF THE POLICY CLASS ENTROPY CONDITION

Condition 4.2 quantifies the size of the policy class through a single exponent p :

$$\log N(\epsilon, \log \Pi, \|\cdot\|_\infty) \lesssim \begin{cases} \log(e/\epsilon), & p = 0, \\ \epsilon^{-p}, & p \in (0, 2). \end{cases}$$

Smaller p means lower complexity and therefore faster rates in Theorem 4.4. This is the same role played by metric-entropy exponents in classical empirical-process theory [20, 61, 62]. Typical examples are:

- Finite or finite-dimensional parametric policy classes (including compact softmax/logit models with uniformly Lipschitz logits in parameters): $\log N(\epsilon, \log \Pi, \|\cdot\|_\infty) = O(\log(1/\epsilon))$, so $p = 0$.
- Hölder logit classes $\log \Pi \subset C^s([0, 1]^d)$ with bounded C^s -norm: $\log N(\epsilon, \log \Pi, \|\cdot\|_\infty) \asymp \epsilon^{-d/s}$, so $p = d/s$.
- Sobolev/Matérn-type smooth classes on d -dimensional contexts (with enough smoothness for uniform control): the same heuristic scaling $p \approx d/s$ is standard.

The log-scale is significant. Our losses and confidence bounds depend on policy ratios $\pi/\pi_t = \exp(\log \pi - \log \pi_t)$, so uniform control of $\log \pi$ gives multiplicative control of importance ratios. This is exactly the right geometry for variance-sensitive policy learning.

It also requires the policy class to behave well near the boundary. If a class allows $\pi(a \mid x) = 0$, it implies $\log \pi(a \mid x) = -\infty$, and $\log \Pi$ may have infinite sup-norm entropy near that boundary. In practice, one works with interior policies (e.g., softmax with temperature/flooring,

ϵ -greedy mixtures, or explicit propensity floors), which keeps $\log \Pi$ well-behaved and simultaneously stabilizes importance-weight variance.

4.3 DISCUSSION OF THE HÖLDERIAN POLICY ERROR BOUND

Condition 4.3 introduced by Zenati et al. [68] is the structural assumption that drives fast rates in our analysis. It links a *variance/overlap* quantity to suboptimality:

$$\text{Var} \left(\frac{\pi^*(A \mid X)}{\pi(A \mid X)} \right) = \mathbb{E}_X [\chi^2(\pi^*(\cdot \mid X) \parallel \pi(\cdot \mid X))],$$

so the condition says that as a policy gets close in risk to π^* , its local mismatch with π^* (measured in χ^2 -divergence) must also shrink at a polynomial rate. This is exactly the quantity that controls the variance of importance-weighted evaluation, so the condition directly explains why pessimistic updates can accelerate.

This condition is a policy-learning analogue of Hölderian error bounds in optimization, where one controls a geometry term (distance to the solution set, gradient norm, KL-type divergence) by a power of objective suboptimality. In particular, for smooth objectives, the Polyak–Łojasiewicz (PL) inequality $\|\nabla f(\theta)\|^2 \gtrsim f(\theta) - f^*$ is one canonical instance of this geometry and yields linear-type convergence. In that literature, such bounds underlie restart-based acceleration and fast rates [8, 13, 21, 46, 49, 72, 73]; see also related online-learning uses [29].

Zenati et al. [68] discuss instances of losses and policy classes for which this condition holds. At a high level, Condition 4.3 should hold whenever two ingredients are present: (i) low excess risk forces near-agreement with the optimal policy on most contexts, and (ii) policy parameterization enforces enough overlap so that the χ^2 -mismatch is controlled by disagreement. The next lemma gives a concrete realizable-margin instantiation.

Lemma 4.5 (Realizability and Tsybakov margin imply HEB). *Suppose the action space $\mathcal{A} = \{1, \dots, K\}$ is finite. Define the conditional mean $\mu(x, a) := \mathbb{E}[Y \mid X = x, A = a]$, the optimal value $\mu^*(x) := \min_a \mu(x, a)$, and the optimal action selector $a^*(x) \in \arg \min_a \mu(x, a)$. Assume the following hold:*

- Realizability:** *There exists $\pi^* \in \Pi$ such that $R(\pi^*) = \mathbb{E}[\mu^*(X)]$.*
- Tsybakov margin condition [42]:** *Let the margin gap be $\Delta(x) := \min_{a \neq a^*(x)} (\mu(x, a) - \mu^*(x))$. There exist constants $C_m > 0$ and $\kappa > 0$ such that $\Pr(\Delta(X) \leq u) \leq C_m u^\kappa$ for all $u \geq 0$.*
- Optimal-action overlap:** *There exists a constant $c_0 > 0$ such that $\pi(a^*(X) \mid X) \geq c_0$ almost surely for all $\pi \in \Pi$.*

Then, there exists a constant $C > 0$ such that for all $\pi \in \Pi$,

$$\text{Var}_{X, A \sim \pi(\cdot|X)} \left(\frac{\pi^*(A|X)}{\pi(A|X)} \right) \leq C(R(\pi) - R(\pi^*))^{\frac{\kappa}{\kappa+1}}.$$

In particular, Condition 4.3 holds with $\beta = \kappa/(\kappa + 1)$.

Lemma 4.5, whose proof is given in Appendix A, formalizes the main intuition: margins convert risk suboptimality into small disagreement mass on near-tie contexts, and overlap converts this disagreement mass into low importance-weight variance. This is precisely the mechanism exploited by our sequential pessimistic updates.

The previous lemma is meant as a structural example for the variance–error relation itself; realizability is not necessary for Condition 4.3. Moreover, combining it simultaneously with the log-scale entropy condition typically requires working with a smoothed or floored version of the optimal policy. In practice, constructions such as ϵ -greedy or softmax policies with temperature/flooring are common ways to enforce the required optimal-action overlap.

Example 4.6 (No Realizability Instantiation). Take a single context, two actions with mean losses μ_0, μ_1 and gap $\Delta := \mu_0 - \mu_1 > 0$, and the floored class $\Pi = \{\pi_\theta : \pi_\theta(1) = \theta, \theta \in [c, 1 - c]\}$, $c \in (0, \frac{1}{2})$. The Bayes policy is deterministic and excluded, so realizability fails; the best-in-class policy is $\pi^* = \pi_{1-c}$. Then $R(\pi_\theta) - R(\pi^*) = (1 - c - \theta)\Delta$ while:

$$\text{Var}_{A \sim \pi_\theta}(\pi^*(A)/\pi_\theta(A)) = (1 - c - \theta)^2/(\theta(1 - \theta)),$$

which is quadratic in the excess risk. Hence Condition 4.3 holds with $\beta = 1$, the fastest regime, despite non-realizability.

5 A MASTER THEOREM FOR AGNOSTIC FAST RATES

This section isolates the reusable proof template behind the regret analysis. Rather than committing to a specific estimator, we state abstract conditions that separate three roles: (i) control of surrogate excess risk by pessimism, (ii) control of surrogate-to-true-risk bias, and (iii) a variance-to-excess-risk link. The key abstraction is the *critical radius*, which quantifies the resolution limit of the surrogate risk estimator over the policy class.

Consider a sequence of policies $(\pi_t)_{t \geq 1} \in \Pi^{\mathbb{N}}$. For every $t \geq 1$ and $\alpha > 0$, let $R_{t,\alpha} : \Pi \rightarrow \mathbb{R}$ be a surrogate risk functional and $\hat{R}_{t,\alpha}$ its empirical counterpart. Let $(\sigma_t)_{t \geq 1}, (\alpha_t)_{t \geq 1}$ be positive sequences, and write $\bar{\sigma}_t := (t^{-1} \sum_{s=1}^t \sigma_s^2)^{1/2}$. It is helpful to think of $R_{t,\alpha}$ as a clipped surrogate risk, α_t as a bias–complexity trade-off parameter, and σ_t^2 as a predictable variance proxy.

Critical radius. The accuracy of $\hat{R}_{t,\alpha}$ as a uniform estimate of $R_{t,\alpha}$ over Π depends on the complexity of the surrogate loss class. We capture this through a localized sequential complexity measure $\mathcal{H}_t(\cdot, \alpha) : (0, \infty) \rightarrow [0, \infty)$, which quantifies the metric entropy of the loss class localized to functions with variance at most r , relative to sample size t and envelope α . The *critical radius* is the smallest scale at which this complexity is dominated by a quadratic signal term:

$$r_{t,\alpha} := \inf\{r > 0 : \mathcal{H}_t(r, \alpha) \leq r^2\}. \quad (14)$$

Below $r_{t,\alpha}$, sampling noise overwhelms differences between surrogate risks; above it, uniform inference over the policy class becomes possible.

Condition 5.1 (Excess surrogate risk). There exist $b \in [0, 1]$ and a decreasing sequence $(\delta_t)_{t \geq 1}$ in $(0, 1)$ such that for every $t \geq 1$, with probability at least $1 - \delta_t$,

$$R_{t,\alpha_t}(\pi_{t+1}) - R_{t,\alpha_t}(\pi^*) \lesssim \frac{\bar{\sigma}_t^{1-b}}{\sqrt{t}} + r_{t,\alpha}^2, \quad (15)$$

where the critical radius satisfies $r_{t,\alpha} \asymp (\alpha_t/t)^{1/(2+2b)}$.

The parameter b plays the role of a complexity exponent: $b = 0$ corresponds to parametric classes and $b = p/2 \in (0, 1)$ to nonparametric classes with entropy exponent p . The condition says the pessimistic update controls surrogate excess risk up to a variance-adaptive term $\bar{\sigma}_t^{1-b}/\sqrt{t}$ that shrinks as the importance-weight variance decreases, and a critical-radius floor $r_{t,\alpha}^2$ set by class complexity and clipping level.

Condition 5.2 (Surrogate risk bias). For every $t \geq 1$,

$$(R(\pi_{t+1}) - R(\pi^*)) - (R_{t,\alpha_t}(\pi_{t+1}) - R_{t,\alpha_t}(\pi^*)) \lesssim \frac{1}{\alpha_t} \bar{\sigma}_t^2. \quad (16)$$

Condition 5.3 (Variance bound). There exists $\beta \in (0, 1]$ such that, for every $t \geq 1$, $\sigma_t^2 \lesssim (R(\pi_t) - R(\pi^*))^\beta$.

Remark 5.4 (Bias–complexity tradeoff). Conditions 5.1 and 5.2 encode a tension controlled by (α_t) . The critical-radius floor r_{t,α_t}^2 increases with α , since a larger envelope enlarges the localized complexity $\mathcal{H}_t(r, \alpha_t)$ and pushes the critical radius up; the surrogate bias $\bar{\sigma}_t^2/\alpha_t$ shrinks with α_t . Condition 5.3 breaks this deadlock: by linking $\bar{\sigma}_t^2$ to excess risk, it ensures the bias term decays along the trajectory, allowing α_t to grow slowly enough that the critical radius stays controlled. The proof of Theorem 5.5 chooses α_t to equalize the two competing rates. Notably, the variance-localization term $\bar{\sigma}_t^{1-b}/\sqrt{t}$ —the engine of the fast rate in full-information ERM—never binds here: because the running variance $\bar{\sigma}_t$ shrinks with the excess risk along the trajectory, it decays at least as fast as the balanced rate (equal only when $b = 0$ or $\beta = 1$). The rate is set entirely by the critical

radius against the clipping bias, so the Hölderian bound accelerates through the bias channel rather than through localization, in contrast to the overlap regime, where clipping is inactive and localization alone sets the rate.

Theorem 5.5 (Regret bound). *Suppose Conditions 5.1–5.3 hold. Then, with probability at least $1 - \delta$, if $\beta = 1$, for $\alpha_t := 1 + \log(et)$, $\delta_t = \delta/(t(t+1))$,*

$$\text{Regret}_T = \begin{cases} \mathcal{O}(\log^2 T) & \text{for } b = 0, \\ \tilde{\mathcal{O}}\left(T^{\frac{b}{1+b}}\right) & \text{for } b \in (0, 1], \end{cases} \quad (17)$$

and, if $0 < \beta < 1$, for $\alpha_t := 2 \vee t^\gamma$ with $\gamma := \frac{1-\beta}{2+b-\beta}$,

$$\text{Regret}_T = \tilde{\mathcal{O}}\left(T^{\frac{1-\beta+b}{2-\beta+b}}\right), \quad (18)$$

where $\tilde{\mathcal{O}}$ absorbs polylogarithmic factors in T and $1/\delta$.

Conditions 5.1–5.3 are intentionally modular: in Section 6 we verify them for O2PL using clipped importance-weighted losses, a self-normalized maximal inequality, and the Hölderian error-bound condition.

6 REGRET ANALYSIS

We now instantiate the master theorem for O2PL by verifying its conditions in sequence. First, we verify surrogate excess-risk control via pessimistic policy learning (Condition 5.1). We introduce a new maximal inequality to obtain uniform confidence control for the surrogate objective. Second, we control the surrogate bias (Condition 5.2). Eventually, the variance bound (Condition 5.3) is immediate from the Hölderian error bound (Condition 4.3).

6.1 EXCESS SURROGATE RISK CONTROL VIA PESSIMISTIC POLICY LEARNING

For every $t \geq 1$, let $\hat{R}_t$ be an $\mathcal{F}_t$ -measurable map $\Pi \rightarrow \mathbb{R}$, an estimator of R_{t,α_t} , with $(\alpha_t)_{t \geq 1}$ given. We suppose first that we can construct uniform-in- Π confidence intervals with widths depending on the estimated variance $\hat{\sigma}_{t,\alpha}(\pi)$.

Condition 6.1 (Uniform confidence intervals). There exists a known $\Phi : (0, \infty) \times \mathbb{N} \times (0, \infty) \times (0, 1) \rightarrow (0, \infty)$ and, for every $t \geq 1$, an $\mathcal{F}_t$ -measurable $\hat{\sigma}_{t,\alpha} : \Pi \rightarrow (0, \infty)$ such that with probability at least $1 - \delta_t$, for all $\pi \in \Pi$,

$$|R_{t,\alpha_t}(\pi) - \hat{R}_t(\pi)| \leq \Phi(\hat{\sigma}_{t,\alpha}(\pi), t, \alpha_t, \delta_t). \quad (19)$$

We define the pessimistic policy at round $t + 1$ as

$$\pi_{t+1} \in \arg \min_{\pi \in \Pi} \hat{R}_t(\pi) + \Phi(\hat{\sigma}_{t,\alpha}(\pi), t, \alpha_t, \delta_t). \quad (20)$$

Condition 6.2 (Empirical to predictable variance). With probability at least $1 - \delta_t$:

$$\Phi(\hat{\sigma}_{t,\alpha}(\pi^*), t, \alpha_t, \delta_t) \lesssim \bar{\sigma}_t^{1-b}/\sqrt{t} + r_{t,\alpha}^2$$

In the next subsection 6.2, we verify the Conditions 6.1, 6.2.

Proposition 6.3. *Suppose Conditions 6.1 and 6.2 hold. Then, with probability at least $1 - 2\delta_t$, Condition 5.1 holds.*

Proof. Fix $t \geq 1$ and work on the event where (19) holds uniformly over Π . By the pessimistic selection rule (20),

$$\begin{aligned} R_{t,\alpha_t}(\pi_{t+1}) &\leq \hat{R}_t(\pi_{t+1}) + \Phi(\hat{\sigma}_{t,\alpha}(\pi_{t+1}), t, \alpha_t, \delta_t) \\ &\leq \hat{R}_t(\pi^*) + \Phi(\hat{\sigma}_{t,\alpha}(\pi^*), t, \alpha_t, \delta_t) \\ &\leq R_{t,\alpha_t}(\pi^*) + 2\Phi(\hat{\sigma}_{t,\alpha}(\pi^*), t, \alpha_t, \delta_t), \end{aligned} \quad (21)$$

hence

$$R_{t,\alpha_t}(\pi_{t+1}) - R_{t,\alpha_t}(\pi^*) \leq 2\Phi(\hat{\sigma}_{t,\alpha}(\pi^*), t, \alpha_t, \delta_t).$$

The claim follows from Condition 6.2. $\square$

6.2 A MAXIMAL INEQUALITY FOR UNIFORM INFERENCE ON THE SURROGATE POLICY RISK

We seek a uniform control of the surrogate policy risk; to do so we verify Conditions 6.1, 6.2 by constructing the penalty Φ in three steps. First, we state a novel maximal inequality for martingale empirical processes localized by the *realized* quadratic variation, and specialize it to the clipped loss class; its deviation bound is the functional $\tilde{H}_t$ of (22). Second, entropy transfer (Proposition 6.5) reduces $\tilde{H}_t$ to the static sup-norm geometry of $\log \Pi$, yielding the critical radius $r_{\alpha,t}$. Third, peeling over variance shells self-normalizes the bound, producing the π -dependent penalty Φ (Theorem 6.6) exhibited in Conditions 6.1, 6.2.

Sequential bracketing entropy. For our maximal inequality, we use the bracketing entropy in the realized norm as a complexity measure. For any $t \geq 1$, index set Θ , and class of adapted processes $\{\xi_{1:t}(\theta) : \theta \in \Theta\}$ with each $\xi_s(\theta)$ being $\mathcal{F}_s$ -measurable, define the realized empirical norm

$$\hat{\rho}_t(\xi_{1:t}(\theta)) := \left(\frac{1}{t} \sum_{s=1}^t \xi_s(\theta)^2 \right)^{1/2}.$$

An ϵ -sequential bracketing is a family $\{(\Lambda_{1:t}^j, \Gamma_{1:t}^j)\}_{j=1}^N$ of $\mathcal{F}_s$ -measurable brackets such that every $\xi_{1:t}(\theta)$ is trapped $(\Lambda_s^j \leq \xi_s(\theta) \leq \Gamma_s^j$ for all s) by some bracket of realized width $\hat{\rho}_t(\Gamma^j - \Lambda^j) \leq \epsilon$; we write $N_{[]}(\epsilon, \Theta)$ for the smallest such N and

$$J_{[]}(\sigma, \Theta) := \int_0^\sigma \sqrt{\log N_{[]}(\epsilon, \Theta)} d\epsilon$$

for the associated entropy integral.

The maximal inequality. We state below our novel uniform deviation bound on which the confidence radii in our analysis are built.

Theorem 6.4 (Maximal inequality under realized-variation localization). *Let Θ index $\mathcal{F}_s$ -measurable processes with $|\xi_s(\theta)| \leq \alpha$ a.s., and set*

$$M_t(\theta) := \frac{1}{t} \sum_{s=1}^t (\xi_s(\theta) - \mathbb{E}[\xi_s(\theta) | \mathcal{F}_{s-1}]).$$

There is a universal $C > 0$ such that, for every $\sigma > 0$ and $\delta \in (0, 1]$, with probability at least $1 - \delta$,

$$\sup_{\theta \in \Theta} \mathbf{1}_{\{\hat{\rho}_t(\xi_{1:t}(\theta)) \leq \sigma\}} M_t(\theta) \leq C \tilde{H}_t(\sigma, \alpha, \delta),$$

where

$$\begin{aligned} \tilde{H}_t(\sigma, \alpha, \delta) &:= \frac{1}{\sqrt{t}} J_{[]}(\sigma, \Theta) + \frac{\alpha}{t} \log N_{[]}(\sigma, \Theta) \\ &\quad + \sigma \sqrt{\frac{\log(e/\delta)}{t}} + \alpha \frac{\log(e/\delta)}{t}. \end{aligned} \quad (22)$$

The proof is deferred to Appendix B. The key novelty is that our maximal inequality localizes the martingale empirical process by the *realized* quadratic variation, rather than by the usual predictable variation. This realized, hence observable, quadratic control is what makes it possible to chain over the policy class while retaining a data-dependent variance proxy under adaptive sampling. The bound has the same Bernstein-type structure as the scalar inequality in Proposition B.1: the first two terms of $\tilde{H}_t$ are the localized complexity contribution, while the last two are the standard Bernstein corrections, matching the classical scalar decomposition [1, 27].

Application to the clipped loss class. We apply Theorem 6.4 to the clipped importance-weighted loss $\ell_{s,\alpha}(\pi)$, an adapted process bounded with an envelope α and indexed by $\pi \in \Pi$, whose centered average is the deviation $\hat{R}_t(\pi) - R_{t,\alpha}(\pi)$ of the empirical clipped objective (6) from its predictable target. Since the inequality localizes by the realized norm, the class it must be applied to is the one restricted to a variance shell,

$$\mathcal{L}_{1:t}^\alpha(\sigma) := \{\ell_{1:t,\alpha}(\pi) : \hat{\rho}_t(\ell_{1:t,\alpha}(\pi)) \leq \sigma, \pi \in \Pi\},$$

so the entropy entering $\tilde{H}_t(\sigma, \alpha, \delta)$ in (22) is that of this localized class, $N_{[]}(\epsilon, \mathcal{L}_{1:t}^\alpha(\sigma))$. Controlling $\tilde{H}_t$ therefore reduces to bounding this localized entropy, which the next proposition transfers to the sup-norm geometry of $\log \Pi$; peeling over the shells σ then yields the π -dependent penalty Φ .

Entropy transfer to the log-policy class. The next proposition shows that the localized complexity of $\mathcal{L}_{1:t}^\alpha(\sigma)$ is governed by the sup-norm geometry of $\log \Pi$.

Proposition 6.5 (Entropy transfer on the log scale). *Assume $Y_s \in [-1, 0]$ a.s. and $\alpha \geq 1$. For every $t \geq 1$, $\sigma > 0$, $\epsilon > 0$, and a.e. realization,*

$$N_{[]}(\epsilon, \mathcal{L}_{1:t}^\alpha(\sigma)) \leq N\left(\frac{c\epsilon}{1+\sigma}, \log \Pi, \|\cdot\|_\infty\right)$$

for a universal constant $c > 0$.

The mechanism is simple: ratios satisfy $\frac{\pi}{\pi_t}(a | x) = \exp(\log \pi(a | x) - \log \pi_t(a | x))$, so sup-norm control of $\log \pi$ gives multiplicative control of the ratios; since $w \mapsto (\min(w, \alpha) - 1)y$ is monotone and Lipschitz for $y \in [-1, 0]$, brackets transfer directly (proof in Appendix C.1).

Localized complexity and critical radius. Combining Proposition 6.5 with Condition 4.2 gives, in the small-radius regime, $\log N_{[]}(\epsilon, \mathcal{L}_{1:t}^\alpha(\sigma)) \lesssim \log(e/\epsilon)$ for $p = 0$ and $\lesssim \epsilon^{-p}$ for $p \in (0, 2)$, up to constants depending on $1 + \sigma$. Substituting into (22),

$$\tilde{H}_t(\sigma, \alpha_t, \delta) \lesssim \mathcal{H}_t(\sigma, \alpha) + \sigma \sqrt{\frac{\log(e/\delta)}{t}} + \alpha_t \frac{\log(e/\delta)}{t}, \quad (23)$$

where

$$\mathcal{H}_t(\sigma, \alpha) := \begin{cases} \frac{\alpha_t}{t} \log(e/\sigma) + \frac{\sigma}{\sqrt{t}} \sqrt{\log(e/\sigma)}, & p = 0, \\ \frac{\alpha_t}{t} \sigma^{-p} + \frac{\sigma^{1-p/2}}{\sqrt{t}}, & p \in (0, 2). \end{cases}$$

Solving $\mathcal{H}_t(r, \alpha) = r^2$ from (14) gives $r_{\alpha,t} \asymp (\alpha_t/t)^{1/(2+p)}$, with the same scaling up to logarithmic factors when $p = 0$.

Appendix C.3 turns Equation (23) into a self-normalized maximal inequality by peeling over the shells of $\hat{\rho}_t(\ell_{1:t,\alpha}(\pi))$. Since $\hat{\sigma}_{t,\alpha_t}(\pi)$ upper-bounds $\hat{\rho}_t(\ell_{1:t,\alpha}(\pi))$ which results in the following theorem:

Theorem 6.6 (Self-normalized confidence bound). *Under Condition 4.2 with $p \in [0, 2)$, there is a universal $C > 0$ such that with probability at least $1 - \delta$, every $\pi \in \Pi$ satisfies*

$$\begin{aligned} |\hat{R}_t(\pi) - R_{t,\alpha_t}(\pi)| &\lesssim r_{\alpha,t}^2 + \frac{\hat{\sigma}_{t,\alpha_t}(\pi)^{1-p/2}}{\sqrt{t}} \\ &\quad + \hat{\sigma}_{t,\alpha_t}(\pi) \sqrt{\frac{\log(et/\delta)}{t}} + \alpha_t \frac{\log(et/\delta)}{t} \end{aligned}$$

with the second term interpreted up to a logarithmic factor when $p = 0$.

This bound is precisely the penalty Φ of Condition 6.1: $r_{\alpha,t}^2$ is the critical-radius floor, $\hat{\sigma}_{t,\alpha_t}(\pi)^{1-p/2}/\sqrt{t}$ is the localization gain from small empirical variance, and the rest are standard confidence corrections. Combining Theorem 6.6 with Proposition 6.3 and the concentration step transferring $\hat{\sigma}_{t,\alpha_t}(\pi^*)$ to $\bar{\sigma}_t$ yields:

Proposition 6.7 (Verification of Condition 6.2). *Under Conditions 4.2 and 4.3, with probability at least $1 - 2\delta_t$, $R_{t,\alpha_t}(\pi_{t+1}) - R_{t,\alpha_t}(\pi^*) \lesssim \bar{\sigma}_t^{1-p/2}/\sqrt{t} + r_{\alpha,t}^2$. In particular, Condition 5.1 then holds with $b = p/2$.*

Remark 6.8 (Why the variance estimator uses clipped weights). The pessimistic penalty uses the variance estimator $\hat{\sigma}_{t,\alpha_t}(\pi)$ of the clipped surrogate losses. This is natural, since the confidence analysis is carried out for the clipped class, and convenient: clipping ensures bounded increments for the squared loss process, allowing the empirical proxy at π^* to be compared to the predictable variance via Freedman’s inequality. An unclipped analogue would require an extra boundedness or tail assumption on π^*/π_t , essentially an optimal-action overlap $\pi_t(a^*(x) | x) > \eta > 0$.

6.3 SURROGATE BIAS CONTROL

The final step is to control the clipping approximation error. Standard analyses bound it by the clipping level alone, a deterministic $O(1/\alpha)$ that does not adapt to the difficulty of the problem. Our observation is that clipping bias is instead controlled by the *variance of the importance weights*: large ratios—the only ones clipping affects—are rare when $\sigma_s^2(\pi)$ is small, so the bias inherits the low-noise structure.

Lemma 6.9 (Clipping bias). *For any $\pi \in \Pi$ and $\alpha > 1$,*

$$R_t(\pi) \leq R_{t,\alpha}(\pi) \leq R_t(\pi) + \frac{1}{t(\alpha - 1)} \sum_{s=1}^t \sigma_s^2(\pi). \quad (24)$$

The proof (Appendix D.1) simply uses Chebyshev’s inequality and the astute observation that clipping can be decomposed as:

$$\min(\alpha, \frac{\pi}{\pi_t}) = \frac{\pi}{\pi_t} - \left(\frac{\pi}{\pi_t} - \alpha\right)_+$$

This is how the Hölderian bound propagates into the bias: near-optimal policies have small $\sigma_s^2(\pi^*)$, hence small clipping bias, permitting slower growth of α_t and a tighter critical radius. Consequently Condition 5.2 holds, since for any π and $\alpha > 1$,

$$\begin{aligned} (R_t(\pi) - R_t(\pi^*)) - (R_{t,\alpha}(\pi) - R_{t,\alpha}(\pi^*)) \\ \leq R_{t,\alpha}(\pi^*) - R_t(\pi^*) \leq \frac{1}{\alpha - 1} \bar{\sigma}_t^2, \end{aligned} \quad (25)$$

and $R_t(\pi) - R_t(\pi^*) = R(\pi) - R(\pi^*)$ gives the required control.

7 NUMERICAL INVESTIGATION

We report a compact numerical study organized around four questions aligned with the theory: (i) does pessimistic learning improve performance, and in particular does variance

penalization help under dependent data? (ii) how does our agnostic policy-based method compare to other agnostic policy-based methods? (iii) how does it compare to value-based methods under misspecification? (iv) does regret scaling track the H’olderian exponent? Additional experiments, implementation details, and comparisons are deferred to Appendix E.

7.1 VARIANCE PENALIZATION UNDER TEMPORAL DEPENDENCE

We first isolate the role of pessimism by comparing ERM to a variance-penalized variant under temporally dependent observations. Our self-normalized maximal inequality is designed for this setting: it provides confidence control in terms of an observable realized variance proxy, and therefore motivates a dependent-data analogue of the sample variance penalization (SVP) principle of Maurer and Pontil [45]. We adapt their classical SVP-versus-ERM experiment by replacing iid observations with coordinate-wise two-state Markov chains. For each coordinate k , we set $X_{t,k} = a_k + b_k S_{t,k}$ with $S_{t,k} \in \{-1, +1\}$ a two-state Markov chain of persistence p_k , so lag-one correlation $\chi_k = 2p_k - 1$. As in the iid setting, ERM ($\lambda = 0$) can over-select high-variance arms that look favorable by noise; dependence amplifies this because effective sample sizes shrink. We compare ERM to a variance-penalized score:

$$\hat{\mu}_k + \lambda \sqrt{\hat{v}_k / n_{\text{eff},k}} \quad \text{with } n_{\text{eff},k} = n(1 - \hat{\chi}_k)/(1 + \hat{\chi}_k)$$

. Figure 2 shows a clear gap in favor of variance penalization.

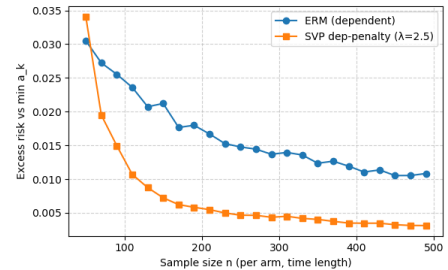


Figure 2: ERM vs. variance-penalized selection under dependent data.

7.2 COMPARISON TO OTHER AGNOSTIC METHODS

We also compare O2PL against other agnostic policy-learning baselines on the OpenML benchmark suite [12] where we do the standard multiclass-to-contextual-bandit conversion of Bietti et al. [11]. O2PL tends to outperform competing methods as environment complexity increases; Figure 1 reports pairwise normalized performance. Additional details are in Appendix E.

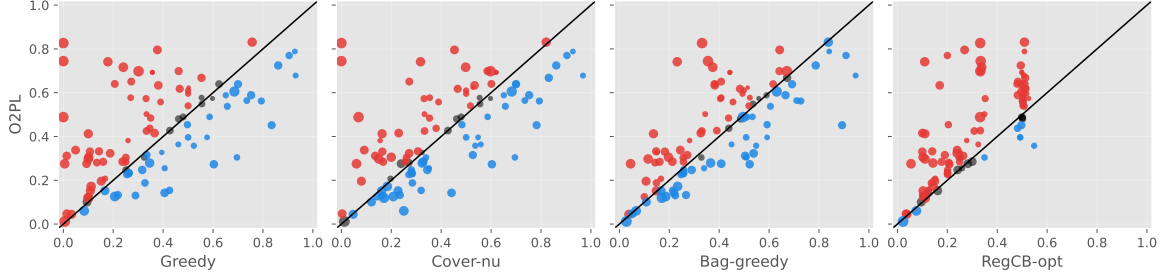


Figure 1: OpenML benchmark: pairwise normalized performance. Red points denote significant O2PL wins, blue points significant O2PL losses, and black points non-significant comparisons, according to a reward-based one-sided Z -test.

7.3 MISSPECIFICATION: THE WHEEL BANDIT

We compare our agnostic policy based method to value based methods in a setting where we mispecify the value function on purpose. We use the wheel bandit environment [53]: Contexts are uniform on the unit disk in $\mathbb{R}^2$ with five actions. Action 0 is a *safe* arm with constant reward μ_1 ; the other four depend on context. Inside the ball $\|x\| \leq \delta$ all four are suboptimal with mean $\mu_2 < \mu_1$; outside, exactly one becomes optimal per quadrant with reward $\mu_3 \gg \mu_1$. The optimal policy is inherently nonlinear—safe arm near the center, quadrant-dependent outside—so when δ is large the high-reward region is rare and linear methods like LinUCB/LinTS can be misled into the safe arm. Because O2PL learns policies directly from bandit feedback rather than fitting a globally linear reward model, it is more robust; Figure 3 shows it substantially outperforms LinUCB and LinTS.

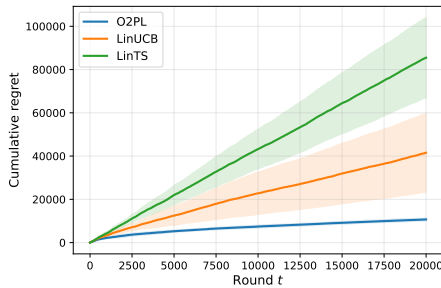


Figure 3: Cumulative regret on the wheel bandit with $\delta = 0.95$ and $(\mu_1, \mu_2, \mu_3) = (1.2, 1, 50)$.

7.4 REGRET SCALING WITH THE HÖLDERIAN EXPONENT

We examine how regret scales with the Hölderian exponent in a controlled synthetic environment.

Let $x_t = s_t z_t$ with $s_t \in \{-1, 1\}$ Rademacher and $z_t \sim \text{Beta}(\nu, 1)$, and choose mean rewards so $\Delta(x) \propto |x|$; then $\Pr(\Delta \leq u) \asymp u^\nu$, yielding $\beta = \nu/(\nu + 1)$. Varying ν tunes β . Figure 4 shows that as β increases, observed regret

growth slows, consistent with theory. Moreover, the empirical regret exponent variation matches that of the theoretical regret.

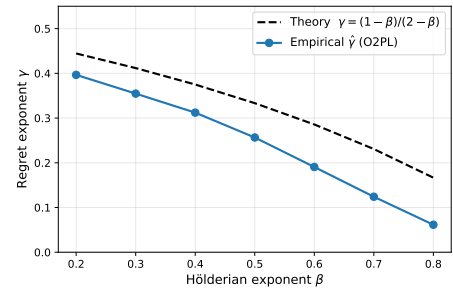


Figure 4: Empirical versus theoretical regret exponent as β varies. Larger β leads to slower regret growth, consistent with theory.

8 CONCLUSION

We studied agnostic policy-class contextual bandits and proved fast best-in-class regret for a fully online algorithm with every-round updates. The algorithmic ingredient is pessimistic policy improvement with variance-aware penalties; the main probabilistic ingredient is a new sequential self-normalized maximal inequality for martingale empirical processes, localized by realized quadratic variation. This inequality yields uniform variance-adaptive confidence bounds under adaptive sampling and is of separate interest beyond O2PL. Our regret analysis is organized through a master theorem: any algorithm satisfying surrogate-risk, bias, and variance conditions inherits fast rates, and O2PL verifies these conditions through clipping, pessimism, and the Hölderian error bound. The remaining open issue is computational: O2PL calls a variance-penalized policy-learning oracle at every round, and this objective is generally nonconvex. Understanding when practical approximations, such as majorization–minimization for variance-penalized IPW objectives [60], preserve the fast-rate guarantee remains an important direction.

References

- [1] *Sequential Analysis and Boundary Crossing Probabilities for Self-Normalized Statistics*, pages 243–265. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-540-85636-8. doi: 10.1007/978-3-540-85636-8_18. URL https://doi.org/10.1007/978-3-540-85636-8_18.
- [2] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.
- [3] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert E. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits, 2014. URL <https://arxiv.org/abs/1402.0555>.
- [4] Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.
- [5] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1): 48–77, 2002.
- [6] Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.
- [7] Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.
- [8] Stephen R. Becker, Emmanuel J. Candès, and Michael C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3):165–218, jul 2011.
- [9] Aurelien Bibaut, Antoine Chambaz, and Mark Laan. Generalized policy elimination: an efficient algorithm for nonparametric contextual bandits. In *Conference on Uncertainty in Artificial Intelligence*, pages 1099–1108. PMLR, 2020.
- [10] Aurelien Bibaut, Nathan Kallus, Maria Dimakopoulou, Antoine Chambaz, and Mark van der Laan. Risk minimization from adaptively collected data: Guarantees for supervised and policy learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 19261–19273. Curran Associates, Inc., 2021.
- [11] Alberto Bietti, Alekh Agarwal, and John Langford. A contextual bandit bake-off, 2021. URL <https://arxiv.org/abs/1802.04064>.
- [12] Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Pieter Gijsbers, Frank Hutter, Michel Lang, Rafael G. Mantovani, Jan N. van Rijn, and Joaquin Vanschoren. Openml benchmarking suites, 2021. URL <https://arxiv.org/abs/1708.03731>.
- [13] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4): 1205–1223, 2007.
- [14] Nicolò Cesa-Bianchi, Pierre Gaillard, Claudio Gentile, and Sébastien Gerchinovitz. Algorithmic chaining and the role of partial feedback in online nonparametric learning. In *Conference on Learning Theory*, pages 465–481. PMLR, 2017.
- [15] Niladri S. Chatterji, Aldo Pacchiano, and Peter L. Bartlett. Online learning with kernel losses. In *Proceedings of the 36th International Conference on Machine Learning*, pages 971–980. PMLR, 2019.
- [16] Mert Demirer, Vasilis Syrgkanis, Greg Lewis, and Victor Chernozhukov. Semi-parametric efficient policy learning with continuous actions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [17] Miroslav Dudík, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. *arXiv preprint arXiv:1106.2369*, 2011.
- [18] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning (ICML)*, 2011.
- [19] Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statist. Sci.*, 29(4):485–511, 11 2014. doi: 10.1214/14-STS500.
- [20] R. M. Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.
- [21] Alexandre d’Aspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. *Foundations and Trends® in Optimization*, 5(1-2):1–245, 2021. ISSN 2167-3888.
- [22] Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International conference on machine learning*, pages 3199–3210. PMLR, 2020.

- [23] Dylan J Foster and Akshay Krishnamurthy. Contextual bandits with surrogate losses: Margin bounds and efficient algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.
- [24] Dylan J Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems*, 34:18907–18919, 2021.
- [25] Dylan J. Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert E. Schapire. Practical contextual bandits with regression oracles. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1539–1548. PMLR, 2018.
- [26] Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.
- [27] David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975. ISSN 00911798, 2168894X. URL <http://www.jstor.org/stable/2959268>.
- [28] David A. Freedman. On tail probabilities for martingales. *Annals of Probability*, 3(1):100–118, 1975.
- [29] Pierre Gaillard and Olivier Wintenberger. Efficient online algorithms for fast-rate regret bounds under sparsity. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [30] Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1): 230–261, 2013.
- [31] Maxime Haddouche and Otmane Sakhi. Logarithmic smoothing for adaptive pac-bayesian off-policy learning, 2025. URL <https://arxiv.org/abs/2506.10664>.
- [32] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47 (260):663–685, 1952.
- [33] Yichun Hu, Nathan Kallus, and Xiaojie Mao. Smooth contextual bandits: Bridging the parametric and non-differentiable regret regimes. In *Conference on Learning Theory*, pages 2007–2010. PMLR, 2020.
- [34] Ying Jin, Zhimei Ren, Zhuoran Yang, and Zhao-ran Wang. Policy learning "without" overlap: Pessimism and generalized empirical bernstein's inequality, 2025. URL <https://arxiv.org/abs/2212.09900>.
- [35] Thorsten Joachims, Adith Swaminathan, and Maarten de Rijke. Deep learning with logged bandit feedback. In *International Conference on Learning Representations (ICLR)*, 2018.
- [36] Nathan Kallus. Balanced policy evaluation and learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS' 18, page 8909–8920, Red Hook, NY, USA, 2018. Curran Associates Inc.
- [37] Nathan Kallus, Yuta Saito, and Masatoshi Uehara. Optimal off-policy evaluation from multiple logging policies. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 5247–5256. PMLR, 2021.
- [38] Nathan Kallus, Xiaojie Mao, Kaiwen Wang, and Zhengyuan Zhou. Doubly robust distributionally robust off-policy evaluation and learning. In *International Conference on Machine Learning*, pages 10598–10632. PMLR, 2022.
- [39] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.
- [40] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.
- [41] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.
- [42] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6): 1808–1829, 1999. ISSN 00905364, 21688966. URL <http://www.jstor.org/stable/2673938>.
- [43] Matthieu Martin, Panayotis Mertikopoulos, Thibaud Rahier, and Houssam Zenati. Nested bandits. In *International Conference on Machine Learning*, pages 15093–15121. PMLR, 2022.
- [44] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006. doi: 10.1214/009053606000000786.
- [45] Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Conference on Learning Theory (COLT)*, 2009.
- [46] Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140: 125 – 161, 2012.

- [47] Aldo Pacchiano. Second order bounds for contextual bandits with function approximation. *arXiv preprint arXiv:2409.16197*, 2024.
- [48] Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *The Annals of Statistics*, 41(2), April 2013. ISSN 0090-5364. doi: 10.1214/13-aos1101. URL <http://dx.doi.org/10.1214/13-AOS1101>.
- [49] B.T Polyak. Convexity of quadratic transformations and its use in control and optimization. *J. Optim. Theory and App*, 99(3):553–583, 1998.
- [50] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability theory and related fields*, 161:111–153, 2015.
- [51] Alexander Rakhlin, Karthik Sridharan, and Alexandre B. Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017. doi: 10.3150/14-BEJ679.
- [52] Philippe Rigollet and Assaf Zeevi. Nonparametric bandits with covariates. *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, pages 54–66, 2010.
- [53] Carlos Riquelme, George Tucker, and Jasper Snoek. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling, 2018. URL <https://arxiv.org/abs/1802.09127>.
- [54] James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.
- [55] PAUL R. Rosaubem and DONALD B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983.
- [56] Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.
- [57] Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation, 2021. URL <https://arxiv.org/abs/2008.07146>.
- [58] David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3):1904–1931, 2022.
- [59] Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.
- [60] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning (ICML)*, 2015.
- [61] Sara A. van de Geer. *Applications of Empirical Process Theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2000.
- [62] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.
- [63] Ramon van Handel. On the minimal penalty for markov order estimation. *Probability Theory and Related Fields*, 150(3–4):709–738, April 2010. ISSN 1432-2064. doi: 10.1007/s00440-010-0290-y. URL <http://dx.doi.org/10.1007/s00440-010-0290-y>.
- [64] Vladimir G. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 371–386. Morgan Kaufmann, 1990.
- [65] Yu-Xiang Wang, Alekh Agarwal, and Miro Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning (ICML)*, 2017.
- [66] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting, 2022. URL <https://arxiv.org/abs/2010.09686>.
- [67] Ian Waudby-Smith, Lili Wu, Aaditya Ramdas, Nikos Karampatziakis, and Paul Mineiro. Anytime-valid off-policy inference for contextual bandits. *ACM / IMS J. Data Sci.*, 1(3), May 2024. doi: 10.1145/3643693. URL <https://doi.org/10.1145/3643693>.
- [68] Houssam Zenati, Eustache Diemert, Matthieu Martin, Julien Mairal, and Pierre Gaillard. Sequential counterfactual risk minimization. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.
- [69] Houssam Zenati, Alberto Bietti, Matthieu Martin, Eustache Diemert, Pierre Gaillard, and Julien Mairal. Counterfactual learning of stochastic policies with continuous actions. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.
- [70] Ruohan Zhan, Zhimei Ren, Susan Athey, and Zhengyuan Zhou. Policy learning with adaptively collected data, 2022. URL <https://arxiv.org/abs/2105.02344>.

- [71] Zhengyuan Zhou, Susan Athey, and Stefan Wager. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 0(0):null, 2022.
- [72] S. Łojasiewicz. *Une propriété topologique des sous ensembles analytiques réels*. Les équations aux dérivées partielles, 1963.
- [73] S. Łojasiewicz. Sur la géométrie semi et sous analytique. *Annales de l'institut Fourier*, 43(5):1575–1595, 1993. ISSN 2167-3888.

APPENDIX

This appendix is organized as follows:

- Appendix A: proof that realizability and Tsybakov margin imply the Hölderian error bound (Lemma 4.5).
- Appendix B: the core probability tools—Bernstein inequality with realized variation (Proposition B.1), sequential chaining, and the conditional expectation bound with realized-variation localization (Theorem B.4).
- Appendix C: complexity control and uniform confidence—entropy transfer on the log scale (Proposition 6.5), bounding the complexity measure via the critical radius, self-normalized confidence bound via peeling (Theorem 6.6), and verification of Condition 6.2.
- Appendix D: regret analysis—clipping bias control (Lemma 6.9) and proof of the master theorem (Theorem 5.5).
- Appendix E: implementation details, additional experiments, and supplementary results.

A SUFFICIENT CONDITION FOR THE HOLDERIAN ERROR BOUND

Lemma 4.5 provides one concrete sufficient condition for the Hölderian error bound, but Condition 4.3 itself is a single scalar inequality on the policy class that does not explicitly require realizability, a margin gap, or optimal-action overlap. It captures only their joint consequence: that importance-weight variance vanishes polynomially with excess risk. This makes it amenable to verification in settings beyond the realizable-margin framework of Lemma 4.5

Proof of Lemma 4.5. Applying the margin condition at $u = 0$ yields $\Pr(\Delta(X) = 0) = 0$, so Bayes-optimal actions are unique almost surely. Therefore $a^*(X)$ is well-defined almost surely.

Because π^* attains Bayes risk and each non-Bayes action has strictly larger conditional mean loss almost surely, necessarily

$$\pi^*(a^*(X) | X) = 1 \quad \text{a.s.}$$

Fix any $\pi \in \Pi$ and define

$$q_\pi(X) := 1 - \pi(a^*(X) | X) \in [0, 1].$$

Using $\pi^*(\cdot | X) = \delta_{a^*(X)}$, we have pointwise

$$\chi^2(\pi^*(\cdot | X) \| \pi(\cdot | X)) = \frac{1}{\pi(a^*(X) | X)} - 1 = \frac{q_\pi(X)}{\pi(a^*(X) | X)} \leq \frac{q_\pi(X)}{c_0},$$

hence, taking expectation over X ,

$$\text{Var}\left(\frac{\pi^*}{\pi}(A | X)\right) = \mathbb{E}[\chi^2(\pi^*(\cdot | X) \| \pi(\cdot | X))] \leq \frac{1}{c_0} \mathbb{E}[q_\pi(X)].$$

It remains to bound $\mathbb{E}[q_\pi(X)]$ by excess risk. Let $(X, A) \sim P_X \times \pi(\cdot | X)$. Then

$$R(\pi) - R(\pi^*) = \mathbb{E}\left[\sum_{a \in \mathcal{A}} \pi(a | X)(\mu(X, a) - \mu^*(X))\right] \geq \mathbb{E}[q_\pi(X)\Delta(X)] =: r.$$

Now fix any $u > 0$ and decompose:

$$\mathbb{E}[q_\pi(X)] \leq \mathbb{E}[\mathbf{1}\{\Delta(X) \leq u\}] + \mathbb{E}[q_\pi(X)\mathbf{1}\{\Delta(X) > u\}] \leq \Pr(\Delta(X) \leq u) + \frac{\mathbb{E}[q_\pi(X)\Delta(X)]}{u} \leq C_m u^\kappa + \frac{r}{u}.$$

The right-hand side is minimized at

$$u^* = \left(\frac{r}{\kappa C_m}\right)^{\frac{1}{\kappa+1}},$$

which gives

$$\mathbb{E}[q_\pi(X)] \leq \left(\kappa^{\frac{1}{\kappa+1}} + \kappa^{-\frac{\kappa}{\kappa+1}}\right) C_m^{\frac{1}{\kappa+1}} r^{\frac{\kappa}{\kappa+1}} =: C_{\kappa, C_m} r^{\frac{\kappa}{\kappa+1}}.$$

Since $r \leq R(\pi) - R(\pi^*)$, we conclude

$$\text{Var}\left(\frac{\pi^*}{\pi}(A | X)\right) \leq \frac{C_{\kappa, C_m}}{c_0} (R(\pi) - R(\pi^*))^{\frac{\kappa}{\kappa+1}},$$

which is the claim. $\square$

B SELF-NORMALIZED MAXIMAL INEQUALITY

The maximal inequality of Theorem B.4 controls the supremum of a martingale array localized by the *realized* quadratic variation $\hat{\rho}_n(X^\theta)^2 = n^{-1} \sum_{i=1}^n (X_i^\theta)^2$. This quantity is $\mathcal{F}_n$ -measurable but not predictable: it depends on the full sample path and cannot be expressed as a sum of conditional variances. The classical chaining argument of [63] localizes instead by the *predictable* variation $\sum_{i=1}^n \mathbb{E}[(X_i^\theta)^2 \mid \mathcal{F}_{i-1}]$, which is a predictable process and amenable to standard Freedman-type inequalities.

To work with the realized variation, we require a concentration inequality in which the quadratic control is itself non-predictable. Proposition B.1 below provides such a bound: it is a Bernstein-type concentration bound where the variance term is the observed sum of squares $S_j = \sum_{t=1}^j X_t^2$ rather than its conditional expectation. Once this ingredient is in place, the maximal inequality and chaining argument proceed along the lines of [63], with the predictable variation replaced throughout by the realized variation.

B.1 A FREEDMAN-TYPE INEQUALITY WITH REALIZED QUADRATIC CONTROL

Proposition B.1 (Bernstein inequality with the realizable variation). *Let $(\mathcal{F}_t)_{t \geq 0}$ be a filtration and let $(X_t)_{t \geq 1}$ be a sequence of real-valued random variables such that X_t is $\mathcal{F}_t$ -measurable and $|X_t| \leq B$ a.s. for all $t \geq 1$. Define*

$$M_j := \sum_{t=1}^j (X_t - \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]), \quad S_j := \sum_{t=1}^j X_t^2, \quad j \geq 1,$$

with $M_0 = S_0 = 0$. Then for all $u, v > 0$,

$$\Pr(\exists j \geq 1 : M_j \geq u, S_j \leq v) \leq \exp\left(-\frac{u^2}{4(v + Bu)}\right).$$

Proof. Write $D_t = X_t - \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]$ for the martingale difference.

Step 1: a scalar inequality. For every $x \geq -1/2$,

$$e^{x-x^2} \leq 1 + x. \quad (26)$$

Indeed, $g(x) := \log(1+x) - x + x^2$ satisfies $g'(x) = x(1+2x)/(1+x)$, which is nonpositive on $[-1/2, 0]$ and nonnegative on $[0, \infty)$. Since $g(0) = 0$, we have $g(x) \geq 0$ for all $x \geq -1/2$, and exponentiating gives (26).

Step 2: one-step bound. Fix $\lambda \in (0, 1/(2B)]$. Since $|\lambda X_t| \leq 1/2$, applying (26) with $x = \lambda X_t$ and taking conditional expectations yields

$$\mathbb{E}[e^{\lambda X_t - \lambda^2 X_t^2} \mid \mathcal{F}_{t-1}] \leq 1 + \lambda \mathbb{E}[X_t \mid \mathcal{F}_{t-1}].$$

Multiplying both sides by $e^{-\lambda \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]}$ and using the elementary inequality $e^{-y}(1+y) \leq 1$ for all $y \in \mathbb{R}$, we obtain

$$\mathbb{E}[e^{\lambda D_t - \lambda^2 X_t^2} \mid \mathcal{F}_{t-1}] \leq 1. \quad (27)$$

Step 3: supermartingale and stopping. Define $\mathcal{E}_j(\lambda) := \exp(\lambda M_j - \lambda^2 S_j)$. The multiplicative decomposition $\mathcal{E}_j = \mathcal{E}_{j-1} \cdot \exp(\lambda D_j - \lambda^2 X_j^2)$ together with (27) shows that $(\mathcal{E}_j(\lambda))_{j \geq 0}$ is a positive supermartingale with $\mathcal{E}_0 = 1$.

Let $\tau = \inf\{j \geq 1 : M_j \geq u, S_j \leq v\}$. On $\{\tau \leq m\}$ we have $\mathcal{E}_\tau(\lambda) \geq \exp(\lambda u - \lambda^2 v)$, so the optional stopping theorem applied to $\tau \wedge m$ gives $\Pr(\tau \leq m) \leq \exp(-\lambda u + \lambda^2 v)$ for every $m \geq 1$. Sending $m \rightarrow \infty$ by continuity from below,

$$\Pr(\exists j \geq 1 : M_j \geq u, S_j \leq v) \leq \exp(-\lambda u + \lambda^2 v). \quad (28)$$

Step 4: optimization over λ . Minimizing $f(\lambda) = -\lambda u + \lambda^2 v$ over $\lambda \in (0, 1/(2B)]$ gives two cases. If the unconstrained minimizer $\lambda_* = u/(2v)$ is admissible (i.e. $u \leq v/B$), then $f(\lambda_*) = -u^2/(4v)$. Otherwise, evaluating at the boundary $\lambda = 1/(2B)$ and using $v < Bu$ gives $f(1/(2B)) < -u/(4B)$. In both cases,

$$f(\lambda) \leq -\min\left\{\frac{u^2}{4v}, \frac{u}{4B}\right\} \leq -\frac{u^2}{4(v + Bu)},$$

where the last inequality follows from $u^2/(4v) \geq u^2/(4(v + Bu))$ and $u/(4B) = u^2/(4Bu) \geq u^2/(4(v + Bu))$. Substituting into (28) completes the proof. $\square$

B.2. MAXIMAL INEQUALITY FOR A FINITE FAMILY

Notation. For an event $A \in \mathcal{F}$ with $\Pr[A] > 0$ and an integrable random variable Z , we write

$$\mathbb{E}^A[Z] := \frac{\mathbb{E}[Z \mathbf{1}_A]}{\Pr[A]} = \mathbb{E}[Z | A].$$

We keep the event A explicit as a free parameter: the results below hold *simultaneously for every choice of A* , and this is what lets the high-probability form of Theorem B.4 be recovered by a direct substitution, with no separate concentration argument or union bound needed.

Lemma B.2 (Maximal inequality for a finite family). *Let $N \geq 1$, and let $X_1, \dots, X_N$ be real-valued random variables. Suppose there exist constants $v > 0$ and $B > 0$ such that*

$$\Pr(|X_i| \geq x) \leq \exp\left(-\frac{x^2}{4(v + Bx)}\right) \quad \text{for all } x > 0 \text{ and all } 1 \leq i \leq N.$$

Then for every event $A \in \mathcal{F}$ with $\Pr[A] > 0$,

$$\mathbb{E}^A \left[\max_{1 \leq i \leq N} |X_i| \right] \leq 4 \sqrt{v \log\left(1 + \frac{N}{\Pr[A]}\right)} + 16B \log\left(1 + \frac{N}{\Pr[A]}\right).$$

Corollary B.3 (Maximal inequality for a finite family of martingales). *Let $n \geq 1$, and let $(X_i^h)_{1 \leq i \leq n}$, for $h = 1, \dots, N$, be $\mathcal{F}_i$ -measurable arrays with $\|X_i^h\|_\infty \leq B$. Then for any event $A \in \mathcal{F}$ with $\Pr[A] > 0$ and any $R > 0$,*

$$\mathbb{E}^A \left[\max_{1 \leq h \leq N} \mathbf{1}_{\{\sum_{i=1}^n (X_i^h)^2 \leq R\}} \max_{1 \leq j \leq n} \sum_{i=1}^j (X_i^h - \mathbb{E}[X_i^h | \mathcal{F}_{i-1}]) \right] \leq 4 \sqrt{R \log\left(1 + \frac{N}{\Pr[A]}\right)} + 16B \log\left(1 + \frac{N}{\Pr[A]}\right).$$

B.3. CHAINING UNDER REALIZED-VARIATION LOCALIZATION

Theorem B.4 (Uniform deviation bound under realized-variation localization). *Fix $n \geq 1$. Let Θ be a class of $\mathcal{F}_i$ -measurable arrays $X^\theta = (X_i^\theta)_{1 \leq i \leq n}$ with $|X_i^\theta| \leq B$ a.s. for all i, θ . Define*

$$M_n^\theta := \frac{1}{n} \sum_{i=1}^n (X_i^\theta - \mathbb{E}[X_i^\theta | \mathcal{F}_{i-1}]), \quad \hat{\rho}_n(X^\theta) := \left(\frac{1}{n} \sum_{i=1}^n (X_i^\theta)^2 \right)^{1/2}.$$

For $\delta > 0$, let $N(\delta, \Theta, \hat{\rho}_n)$ denote the smallest number of predictable brackets $\{(\ell_p, u_p)\}_{p=1}^N$ such that every $\theta \in \Theta$ is captured by some bracket with $\ell_{p,i} \leq X_i^\theta \leq u_{p,i}$ and $\hat{\rho}_n(u_p - \ell_p) \leq \delta$. Define

$$H(R, B, \Theta, \hat{\rho}_n) := \frac{B}{n} \log N(\sqrt{R}, \Theta, \hat{\rho}_n) + \frac{1}{\sqrt{n}} \int_0^{\sqrt{R}} \sqrt{\log N(u, \Theta, \hat{\rho}_n)} du.$$

Then there exists a universal constant $C > 0$ such that, for every $R > 0$ and every $\delta \in (0, 1]$, with probability at least $1 - \delta$,

$$\sup_{\theta \in \Theta} \mathbf{1}_{\{\hat{\rho}_n(X^\theta) \leq \sqrt{R}\}} M_n^\theta \leq C \left(H(R, B, \Theta, \hat{\rho}_n) + \sqrt{\frac{R}{n} \log \frac{2}{\delta}} + \frac{B}{n} \log \frac{2}{\delta} \right).$$

Proof. Before proceeding to the proof, let us define the function

$$\Phi(x) := C \left(H(R, B, \Theta, \hat{\rho}_n) + \sqrt{\frac{R}{n} x} + \frac{B}{n} x \right), \quad x \geq 0.$$

We claim that in order to prove the Theorem, it actually suffices to prove the estimate

$$\mathbb{E}^A \left[\sup_{\theta \in \Theta} \mathbf{1}_{\{\hat{\rho}_n(X^\theta) \leq \sqrt{R}\}} M_n^\theta \right] \leq \Phi \left(\log \frac{2}{\Pr[A]} \right) \quad \text{for every event } A \text{ with } \Pr[A] > 0. \quad (*)$$

Indeed, if this is the case, then, writing $Z := \sup_{\theta \in \Theta} \mathbf{1}_{\{\hat{\rho}_n(X^\theta) \leq \sqrt{R}\}} M_n^\theta$, fixing $\delta \in (0, 1]$ and choosing $A := \{Z \geq \Phi(\log(2/\delta))\}$ (the case $\Pr[A] = 0$ being trivial), we may estimate

$$\Phi \left(\log \frac{2}{\delta} \right) \leq \mathbb{E}^A[Z] \leq \Phi \left(\log \frac{2}{\Pr[A]} \right),$$

where the first inequality holds because $Z \geq \Phi(\log(2/\delta))$ on A by construction, and the second is $(*)$ applied to this particular A . As Φ is strictly increasing (since $R, B > 0$), this forces $\Pr[A] \leq \delta$, i.e. $Z \leq \Phi(\log(2/\delta))$ with probability at least $1 - \delta$, from which the conclusion of the Theorem is immediate. We therefore concentrate, without loss of generality, on establishing $(*)$.

The proof of $(*)$ follows the chaining strategy of [63, Appendix A], with a control on the realized quadratic variation, which also implies the use of lower brackets in the chaining mechanism.

Setup. Fix an event $A \in \mathcal{F}$ with $\Pr[A] > 0$. Set $\eta_j = 2^{-j} \sqrt{R}$ and $N_j = N(\eta_j, \Theta, \hat{\rho}_n)$ for $j \geq 0$. For each j , choose a bracketing family $\mathcal{B}_j = \{(\ell_{j,p}, u_{j,p})\}_{p=1}^{N_j}$ at radius η_j . For each $\theta \in \Theta$ and level j , select an index $\pi(j, \theta)$ such that, writing $\Delta_{j,\theta,i} = u_{j,\pi(j,\theta),i} - \ell_{j,\pi(j,\theta),i}$ and $L_{j,\theta,i} = \ell_{j,\pi(j,\theta),i}$, we have

$$L_{j,\theta,i} \leq X_i^\theta \leq L_{j,\theta,i} + \Delta_{j,\theta,i}, \quad \hat{\rho}_n(\Delta_{j,\theta}) \leq \eta_j.$$

Decomposition. Fix $J \geq 1$ and positive truncation levels $(a_j)_{0 \leq j \leq J-1}$. Define the stopping rule $\tau_i^\theta = \min\{j \geq 0 : \Delta_{j,\theta,i} > a_j\} \wedge J$, which identifies the first level at which the bracket width exceeds the truncation threshold. With the convention $L_{-1,\theta,i} = L_{0,\theta,i}$, decompose

$$X_i^\theta = L_{0,\theta,i} + \sum_{j=0}^J b_{j,\theta,i} + \sum_{j=1}^J c_{j,\theta,i},$$

where

$$b_{j,\theta,i} = (X_i^\theta - L_{j,\theta,i} \vee L_{j-1,\theta,i}) \mathbf{1}_{\{\tau_i^\theta = j\}}$$

is the residual at the stopping level, and

$$c_{j,\theta,i} = (L_{j,\theta,i} \vee L_{j-1,\theta,i} - L_{j-1,\theta,i}) \mathbf{1}_{\{\tau_i^\theta = j\}} + (L_{j,\theta,i} - L_{j-1,\theta,i}) \mathbf{1}_{\{\tau_i^\theta > j\}}$$

is the chaining increment between consecutive bracket levels. The decomposition telescopes: on $\{\tau_i^\theta = m\}$, the increments $c_{k,\theta,i}$ for $k < m$ (where $\tau_i^\theta > k$) contribute $L_{k,\theta,i} - L_{k-1,\theta,i}$, and at $k = m$ they contribute $L_{m,\theta,i} \vee L_{m-1,\theta,i} - L_{m-1,\theta,i}$, so $\sum_{k=1}^m c_{k,\theta,i} = L_{m,\theta,i} \vee L_{m-1,\theta,i} - L_{0,\theta,i}$; adding $b_{m,\theta,i}$ gives $X_i^\theta - L_{0,\theta,i}$. Passing to centered averages, $M_n^\theta = A_n^\theta + \sum_{j=0}^J B_{j,n}^\theta + \sum_{j=1}^J C_{j,n}^\theta$, where each piece is the normalized martingale sum of the corresponding array.

Control of A_n^θ . On $\{\hat{\rho}_n(X^\theta) \leq \sqrt{R}\}$, the base bracket satisfies $n^{-1} \sum_i (L_{0,\theta,i})^2 \leq 4R$ (since $(L_{0,\theta,i})^2 \leq 2(X_i^\theta)^2 + 2\Delta_{0,\theta,i}^2$ and both empirical norms are at most $\sqrt{R}$). Since A_n^θ depends on θ only through $\pi(0, \theta)$, it takes at most N_0 values. Corollary B.3 with realized variance $4R$ and envelope B gives

$$\mathbb{E}^A \left[\sup_{\theta} \mathbf{1}_{\{\hat{\rho}_n(X^\theta) \leq \sqrt{R}\}} A_n^\theta \right] \leq C \left(\sqrt{\frac{R \ell_0}{n}} + \frac{B \ell_0}{n} \right), \quad \ell_0 = \log \frac{2N_0}{\Pr[A]}.$$

Control of the residuals $B_{j,n}^\theta$. Since $L_{j,\theta,i} \leq X_i^\theta$ and $L_{j-1,\theta,i} \leq X_i^\theta$, we have $L_{j,\theta,i} \vee L_{j-1,\theta,i} \leq X_i^\theta$, so $b_{j,\theta,i} \geq 0$, and $b_{j,\theta,i} \leq X_i^\theta - L_{j,\theta,i} \leq \Delta_{j,\theta,i}$. For $j < J$, since $\tau_i^\theta = j$ implies $\Delta_{j,\theta,i} > a_j$, we have $b_{j,\theta,i} \leq \Delta_{j,\theta,i}/a_j$, so $B_{j,n}^\theta \leq \eta_j^2/a_j$. For $j = J$, Cauchy–Schwarz gives $B_{J,n}^\theta \leq \eta_J$. Hence

$$\sum_{j=0}^J \mathbb{E}^A \left[\sup_{\theta} \mathbf{1}_{\{\hat{\rho}_n(X^\theta) \leq \sqrt{R}\}} B_{j,n}^\theta \right] \leq \sum_{j=0}^{J-1} \frac{\eta_j^2}{a_j} + \eta_J.$$

Control of the chain increments $C_{j,n}^\theta$. Fix $j \geq 1$. The array $c_{j,\theta}$ depends on θ only through $(\pi(0, \theta), \dots, \pi(j, \theta))$, so the supremum reduces to a maximum over $\tilde{N}_j = \prod_{p=0}^j N_p$ values. We establish two pointwise bounds. First, both brackets trap X_i^θ , so $|L_{j,\theta,i} - L_{j-1,\theta,i}| \leq \Delta_{j-1,\theta,i} \vee \Delta_{j,\theta,i}$, and the positive part satisfies $(L_{j,\theta,i} - L_{j-1,\theta,i})_+ \leq X_i^\theta - L_{j-1,\theta,i} \leq \Delta_{j-1,\theta,i}$. On $\{\tau_i^\theta = j\}$, we have $\tau_i^\theta > j-1$, so $\Delta_{j-1,\theta,i} \leq a_{j-1}$, giving $c_{j,\theta,i} = (L_{j,\theta,i} - L_{j-1,\theta,i})_+ \leq a_{j-1}$. On $\{\tau_i^\theta > j\}$, both $\Delta_{j-1,\theta,i} \leq a_{j-1}$ and $\Delta_{j,\theta,i} \leq a_j \leq a_{j-1}$ hold (assuming the a_j are nonincreasing, which our choice below will satisfy), so $|c_{j,\theta,i}| \leq a_{j-1}$. For the empirical second moment, on $\{\tau_i^\theta = j\}$ we have $c_{j,\theta,i} \leq \Delta_{j-1,\theta,i}$, and on $\{\tau_i^\theta > j\}$ we have $|c_{j,\theta,i}| \leq \Delta_{j-1,\theta,i} + \Delta_{j,\theta,i}$, so $n^{-1} \sum_i (c_{j,\theta,i})^2 \leq 2(\eta_{j-1}^2 + \eta_j^2) \leq 4\eta_{j-1}^2$. Corollary B.3 with $\ell_j = \log(2\tilde{N}_j / \Pr[A])$ gives

$$\mathbb{E}^A [\sup_\theta \mathbf{1}_{\{\hat{\rho}_n(X^\theta) \leq \sqrt{R}\}} C_{j,n}^\theta] \leq C \left(\frac{\eta_{j-1} \sqrt{\ell_j}}{\sqrt{n}} + \frac{a_{j-1} \ell_j}{n} \right).$$

Choice of truncation levels and conclusion. Set $a_j = \eta_j \sqrt{n} / \sqrt{\ell_{j+1}}$ for $0 \leq j \leq J-1$; this sequence is nonincreasing since η_j decreases geometrically and ℓ_{j+1} increases. Then $\eta_j^2 / a_j = \eta_j \sqrt{\ell_{j+1}} / \sqrt{n}$ and $a_{j-1} \ell_j / n = \eta_{j-1} \sqrt{\ell_j} / \sqrt{n}$. Combining all three contributions,

$$\mathbb{E}^A [\sup_\theta \mathbf{1}_{\{\hat{\rho}_n(X^\theta) \leq \sqrt{R}\}} M_n^\theta] \leq \eta_J + C \left(\sqrt{\frac{R \ell_0}{n}} + \frac{B \ell_0}{n} + \frac{1}{\sqrt{n}} \sum_{j=0}^J \eta_j \sqrt{\ell_j} \right).$$

Letting $J \rightarrow \infty$ removes the η_J term. Since $\ell_j = \log(2 / \Pr[A]) + \sum_{p=0}^j \log N_p$ and $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$, summing over j with $\sum_{j \geq 0} \eta_j = 2\sqrt{R}$ and $\sum_{j \geq p} \eta_j \leq 2\eta_p$ yields

$$\frac{1}{\sqrt{n}} \sum_{j \geq 0} \eta_j \sqrt{\ell_j} \leq C \left(\sqrt{\frac{R}{n} \log \frac{2}{\Pr[A]}} + \frac{1}{\sqrt{n}} \sum_{p \geq 0} \eta_p \sqrt{\log N_p} \right).$$

Finally, since $N(\cdot, \Theta, \hat{\rho}_n)$ is nonincreasing, on each interval $[\eta_{p+1}, \eta_p]$ we have $\sqrt{\log N(u)} \geq \sqrt{\log N_p}$, so $\sum_{p \geq 0} \eta_p \sqrt{\log N_p} \leq 2 \int_0^{\sqrt{R}} \sqrt{\log N(u, \Theta, \hat{\rho}_n)} du$. Substituting and using $\ell_0 \leq \log(2 / \Pr[A]) + \log N(\sqrt{R}, \Theta, \hat{\rho}_n)$ gives (*), and hence the Theorem. $\square$

C COMPLEXITY CONTROL AND UNIFORM CONFIDENCE

C.1 ENTROPY TRANSFER ON THE LOG SCALE (PROPOSITION 6.5)

Proposition C.1 (Entropy transfer on the log scale). *Assume $Y_s \in [-1, 0]$ almost surely and $\alpha \geq 1$. For every $t \geq 1$, $\sigma > 0$, and $\epsilon > 0$,*

$$N_{[]}(\epsilon, \mathcal{L}_{1:t}^\alpha(\sigma), \hat{\rho}_t) \leq N\left(\frac{c\epsilon}{1+\sigma}, \log \Pi, \|\cdot\|_\infty\right)$$

for a universal constant $c > 0$, where $\mathcal{L}_{1:t}^\alpha(\sigma) = \{\pi \in \Pi : \hat{\rho}_t(\ell_{1:t,\alpha}(\pi)) \leq \sigma\}$.

Proof. Fix a realization. Write $w_s^\pi = \pi(A_s | X_s)/\pi_s(A_s | X_s)$ and $\ell_s^\pi = (\min(w_s^\pi, \alpha) - 1)Y_s$. Let $\delta \in (0, 1]$ and let $\{g^j\}_{j=1}^N$ be a δ -cover of $\log \Pi$ in $\|\cdot\|_\infty$. For each $\pi \in \Pi$, choose j with $\|\log \pi - g^j\|_\infty \leq \delta$, and set $\tilde{w}_s^j = \exp(g^j(X_s, A_s))/\pi_s(A_s | X_s)$. Then $e^{-\delta}\tilde{w}_s^j \leq w_s^\pi \leq e^\delta\tilde{w}_s^j$ for every $s \leq t$.

The map $h_s : w \mapsto (\min(w, \alpha) - 1)Y_s$ is nonincreasing (since $Y_s \leq 0$) and $|Y_s|$ -Lipschitz. Define brackets $\lambda_s^j := h_s(e^\delta\tilde{w}_s^j)$, $\gamma_s^j := h_s(e^{-\delta}\tilde{w}_s^j)$. Monotonicity gives $\lambda_s^j \leq \ell_s^\pi \leq \gamma_s^j$.

We bound the bracket width directly in terms of clipped quantities, avoiding any appeal to the unclipped empirical norm. Write $b_s := e^{-\delta}\tilde{w}_s^j \leq w_s^\pi$, so the upper endpoint is $e^{2\delta}b_s$. Then

$$\gamma_s^j - \lambda_s^j = (\min(e^{2\delta}b_s, \alpha) - \min(b_s, \alpha)) |Y_s|.$$

Since $x \mapsto \min(x, \alpha)$ is concave with $\min(0, \alpha) = 0$, it is sublinear: $\min(e^{2\delta}b_s, \alpha) \leq e^{2\delta} \min(b_s, \alpha)$. Hence, using $b_s \leq w_s^\pi$,

$$\gamma_s^j - \lambda_s^j \leq (e^{2\delta} - 1) \min(b_s, \alpha) |Y_s| \leq (e^{2\delta} - 1) \min(w_s^\pi, \alpha) |Y_s| \leq C\delta \min(w_s^\pi, \alpha) |Y_s|,$$

for $\delta \in (0, 1]$ and some universal constant C . Now observe

$$\min(w_s^\pi, \alpha) |Y_s| = |(\min(w_s^\pi, \alpha) - 1)Y_s + Y_s| \leq |\ell_s^\pi| + |Y_s|.$$

Taking the empirical norm and applying the triangle inequality,

$$\hat{\rho}_t(\gamma^j - \lambda^j) \leq C\delta (\hat{\rho}_t(\ell^\pi) + \hat{\rho}_t(|Y|)) \leq C\delta(\sigma + 1),$$

where the last step uses $\pi \in \mathcal{L}_{1:t}^\alpha(\sigma)$ (which gives $\hat{\rho}_t(\ell^\pi) \leq \sigma$) and $|Y_s| \leq 1$ (which gives $\hat{\rho}_t(|Y|) \leq 1$). Choosing $\delta = c\epsilon/(1 + \sigma)$ for a small universal constant c completes the proof. $\square$

C.2 BOUNDING THE COMPLEXITY MEASURE VIA ENTROPY TRANSFER

We apply Theorem B.4 to the clipped importance-weighted loss class (3), where the envelope satisfies $|\ell_{\alpha,t}(\pi)(z)| \leq \alpha$.

Instantiation under the entropy assumption. In the small-radius regime ($\sigma \ll 1$), the rescaling factor $c\sigma/(1 + \sigma) \asymp \sigma$ in Proposition 6.5 is negligible. Substituting the entropy assumption $\log N(\epsilon, \log \Pi, \|\cdot\|_\infty) \lesssim \epsilon^{-p}$ into the complexity measure of Theorem B.4 gives the upper bound

$$\mathcal{H}_t(\sigma, \alpha) := \frac{\alpha}{t} \sigma^{-p} + \frac{\sigma^{1-p/2}}{\sqrt{t}}. \quad (29)$$

Solving the fixed-point equation $\mathcal{H}_t(r, \alpha) = r^2$ from (14) yields the critical radius $r_{\alpha,t} \asymp (\alpha_t/t)^{1/(2+p)}$.

The minimal point. The two terms in (29) balance at the scale

$$r_t^* \asymp \left(\frac{\alpha_t}{\sqrt{t}}\right)^{\frac{2}{2+p}}. \quad (30)$$

Below r_t^* , the entropy term $\alpha\sigma^{-p}/t$ dominates; above it, the Dudley-integral term $\sigma^{1-p/2}/\sqrt{t}$ dominates. For $\sigma \gg 1$, the rescaling saturates and $\mathcal{H}_t(\sigma) \lesssim \sigma/\sqrt{t}$.

Two-scale control. The following lemma collects the behaviour of $\mathcal{H}_t$ across the full range of localization radii.

Lemma C.2 (Two-scale control). *Under the entropy assumption 4.2 with $p \in (0, 2)$, suppose $r_t^* \leq c_0$ for a sufficiently small universal constant. Then:*

- (i) For $\sigma \in (0, c_0]$: $\mathcal{H}_t(\sigma) \lesssim \alpha_t \sigma^{-p}/t + \sigma^{1-p/2}/\sqrt{t}$.
- (ii) For $\sigma \in [r_{\alpha,t}, r_t^*]$: $\mathcal{H}_t(\sigma) \lesssim (r_{\alpha,t})^2$.
- (iii) For $\sigma \in [r_t^*, c_0]$: $\mathcal{H}_t(\sigma) \lesssim \sigma^{1-p/2}/\sqrt{t}$.
- (iv) For $\sigma \geq c_0$: $\mathcal{H}_t(\sigma) \lesssim \sigma/\sqrt{t}$.

Proof. Part (i) is immediate from (29). For (ii), on $[r_{\alpha,t}, r_t^*]$ the entropy term satisfies $\alpha_t \sigma^{-p}/t \leq \alpha_t (r_{\alpha,t})^{-p}/t = (r_{\alpha,t})^2$, and the Dudley term satisfies $(r_t^*)^{1-p/2}/\sqrt{t} \asymp \alpha_t (r_t^*)^{-p}/t \lesssim (r_{\alpha,t})^2$. For (iii), when $\sigma \geq r_t^*$ the entropy term is dominated by the Dudley term. Part (iv) follows from the saturation of the rescaling factor at $\sigma \gg 1$. $\square$

C.3 SELF-NORMALIZED BOUND VIA PEELING (THEOREM 6.6)

We combine Theorem B.4 with Lemma C.2 via a dyadic peeling argument over the shells of $\hat{\rho}_t(\ell_{\alpha,1:t}(\pi))$.

Theorem C.3 (Self-normalized bound via peeling). *Under the entropy condition (4.2) with $p \in (0, 2)$, there exists $C > 0$ such that with probability at least $1 - \eta$, every $\pi \in \Pi$ satisfies*

$$M_t^\pi \leq C \left((r_{\alpha,t})^2 + \frac{\hat{\rho}_t(\ell_{\alpha,1:t}(\pi))^{1-p/2}}{\sqrt{t}} + \hat{\rho}_t(\ell_{\alpha,1:t}(\pi)) \sqrt{\frac{L_t(\eta)}{t}} + \frac{\alpha_t}{t} L_t(\eta) \right),$$

where $L_t(\eta) := \log(\alpha_t/r_t^*) + \log(1/\eta)$.

Proof. Recall the localized deviation envelope from Theorem B.4: for any $\sigma > 0$ and $\eta \in (0, 1)$,

$$\varepsilon_t(\sigma, \eta) := C \left(\mathcal{H}_t(\sigma) + \sigma \sqrt{\frac{\log(2/\eta)}{t}} + \frac{\alpha_t}{t} \log \frac{2}{\eta} \right). \quad (31)$$

Shell 1 ($\hat{\rho}_t \leq r_{\alpha,t}$). Apply (31) at $r_{\alpha,t}$ with confidence $\eta/3$. The critical-radius property gives $\varepsilon_t(r_{\alpha,t}, \eta/3) \lesssim (r_{\alpha,t})^2 + (\alpha_t/t)L_t(\eta)$.

Shell 2 ($r_{\alpha,t} < \hat{\rho}_t \leq r_t^*$). Apply (31) at r_t^* with confidence $\eta/3$. Lemma C.2(ii) gives $\mathcal{H}_t(\sigma) \lesssim (r_{\alpha,t})^2$ throughout this range.

Shell 3 ($r_t^* < \hat{\rho}_t \leq \alpha_t$). Peel dyadically: $S_k = \{2^{k-1}r_t^* < \hat{\rho}_t \leq 2^k r_t^*\}$ for $k = 1, \dots, K$ with $K = \lceil \log_2(\alpha_t/r_t^*) \rceil$ and confidence $\eta_k = \eta/(3(k+1)^2)$ per shell. On S_k with $\sigma_k = 2^k r_t^*$:

- If $\sigma_k \leq c_0$: Lemma C.2(iii) gives $\mathcal{H}_t(\sigma_k) \lesssim \sigma_k^{1-p/2}/\sqrt{t}$, contributing $\hat{\rho}_t^{1-p/2}/\sqrt{t}$.
- If $\sigma_k > c_0$: Lemma C.2(iv) gives $\mathcal{H}_t(\sigma_k) \lesssim \sigma_k/\sqrt{t}$, absorbed by the deviation term.

The union bound over K shells costs $\log(\alpha_t/r_t^*)$, which is absorbed into $L_t(\eta)$. Collecting the three contributions completes the proof. $\square$

C.4 VERIFICATION OF CONDITION 6.2

The key step is to transfer from the empirical clipped norm $\hat{\rho}_t = \hat{\rho}_t(\ell_{\alpha,1:t}(\pi^*))$ to the predictable variance $\bar{\sigma}_t = \bar{\sigma}_t(\pi^*)$. Since π^* is a single policy, no uniformity is needed.

Proof. Set

$$X_s := (\min(w_s^{\pi^*}, \alpha_t) - 1)^2, \quad D_s := X_s - \mathbb{E}[X_s \mid \mathcal{F}_{s-1}],$$

so that $(D_s)_{s \leq t}$ is a martingale-difference sequence with

$$\hat{\sigma}_t^2 - \bar{\sigma}_{t,\text{clip}}^2 = \frac{1}{t} \sum_{s=1}^t D_s, \quad \bar{\sigma}_{t,\text{clip}}^2 := \frac{1}{t} \sum_{s=1}^t \mathbb{E}[(\min(w_s^{\pi^*}, \alpha_t) - 1)^2 \mid \mathcal{F}_{s-1}].$$

Note that this definition targets the clipped predictable variance directly, bypassing the factor Y_s^2 : in particular, $\hat{\sigma}_t^2 = t^{-1} \sum_s X_s$ is precisely the empirical clipped variance appearing in Condition 6.2.

By clipping, $\min(w_s^{\pi^*}, \alpha_t) \in [0, \alpha_t]$, hence $(\min(w_s^{\pi^*}, \alpha_t) - 1)^2 \leq \alpha_t^2$ (using $\alpha_t \geq 1$), so

$$0 \leq X_s \leq \alpha_t^2, \quad |D_s| \leq \alpha_t^2.$$

Moreover, clipping contracts toward 1, so $(\min(w, \alpha_t) - 1)^2 \leq (w - 1)^2$ pointwise; therefore

$$\bar{\sigma}_{t,\text{clip}}^2 \leq \frac{1}{t} \sum_{s=1}^t \mathbb{E}[(w_s^{\pi^*} - 1)^2 \mid \mathcal{F}_{s-1}] = \bar{\sigma}_t^2.$$

For the conditional second moment of D_s ,

$$\mathbb{E}[D_s^2 \mid \mathcal{F}_{s-1}] \leq \mathbb{E}[X_s^2 \mid \mathcal{F}_{s-1}] \leq \alpha_t^2 \mathbb{E}[X_s \mid \mathcal{F}_{s-1}],$$

since $X_s^2 \leq \alpha_t^2 X_s$. Setting $V_t := \sum_{s=1}^t \mathbb{E}[D_s^2 \mid \mathcal{F}_{s-1}]$,

$$V_t \leq \alpha_t^2 \sum_{s=1}^t \mathbb{E}[X_s \mid \mathcal{F}_{s-1}] = t\alpha_t^2 \bar{\sigma}_{t,\text{clip}}^2 \leq t\alpha_t^2 \bar{\sigma}_t^2.$$

Applying Freedman's inequality to $\sum_{s=1}^t D_s$ with increment bound α_t^2 yields, with probability at least $1 - \delta_t$,

$$\sum_{s=1}^t D_s \lesssim \sqrt{V_t \log(1/\delta_t)} + \alpha_t^2 \log(1/\delta_t).$$

Using the bound on V_t and dividing by t ,

$$\hat{\sigma}_t^2 \lesssim \bar{\sigma}_t^2 + \alpha_t \bar{\sigma}_t \sqrt{\frac{\log(1/\delta_t)}{t}} + \alpha_t^2 \frac{\log(1/\delta_t)}{t}.$$

Taking square roots,

$$\hat{\sigma}_t \leq c_1(\bar{\sigma}_t + \psi_t), \quad \psi_t := c_0 \alpha_t \sqrt{\frac{\log(1/\delta_t)}{t}}. \quad (32)$$

Condition 4.3 controls the predictable scale: because $\bar{\sigma}_t^2$ is bounded by a power of the excess risk and the latter is uniformly bounded, $\bar{\sigma}_t = O(1)$. Thus the predictable scale of the clipped process remains uniformly controlled. Clipping is essential: it yields the bounded-increment condition and the bound on V_t needed for Freedman's inequality.

From (32) to a penalty bound. We now work on the intersection of the concentration event (32) and the uniform confidence event, and bound $\Phi(\hat{\sigma}_t)$ directly, without case-splitting on the ordering of ψ_t and $r_{\alpha,t}$.

The penalty Φ takes the form $\Phi(\sigma) \asymp r_{\alpha,t}^2 + \sigma^{1-p/2}/\sqrt{t} + \sigma\sqrt{L_t}/t + \alpha_t L_t/t$, which (away from the floor $r_{\alpha,t}^2$) is a sum of terms of the form σ^q with $q \in \{1 - p/2, 1\}$, $q \in (0, 1]$ for $p \in [0, 2)$. On $[0, \infty)$, the map $x \mapsto x^q$ is subadditive for $q \in (0, 1]$: $(A + B)^q \leq A^q + B^q$. Consequently,

$$\Phi(c_1(\bar{\sigma}_t + \psi_t)) \leq C(\Phi(\bar{\sigma}_t) + \Phi(\psi_t))$$

for a universal constant C depending on c_1 and the exponents in Φ . Combined with (32) and the monotonicity of Φ on $[r_{\alpha,t}, \infty)$,

$$\Phi(\hat{\sigma}_t) \leq \Phi(c_1(\bar{\sigma}_t + \psi_t)) \leq C(\Phi(\bar{\sigma}_t) + \Phi(\psi_t)). \quad (33)$$

Absorption of $\Phi(\psi_t)$ into $r_{\alpha,t}^2$. With $\psi_t \asymp \alpha_t \sqrt{\log(1/\delta_t)/t}$,

$$\Phi(\psi_t) \asymp r_{\alpha,t}^2 + \frac{\psi_t^{1-p/2}}{\sqrt{t}} + \psi_t \sqrt{\frac{L_t}{t}} + \alpha_t \frac{L_t}{t}.$$

The last two terms are absorbed into $r_{\alpha,t}^2$ by the same reasoning used below for $\bar{\sigma}_t$. For the dominant term,

$$\frac{\psi_t^{1-p/2}}{\sqrt{t}} \asymp \alpha_t^{1-p/2} t^{(p-4)/4} (\log(1/\delta_t))^{(1-p/2)/2},$$

up to polylogarithmic factors. A direct comparison of exponents with $r_{\alpha,t}^2 \asymp (\alpha_t/t)^{4/(4+p)}$ shows that, for every $p \in [0, 2]$ and every polynomial schedule $\alpha_t \asymp t^\gamma$ with $\gamma \geq p/(2(2+p))$ in the range admitted by Theorem 5.5, one has $\Phi(\psi_t) \lesssim r_{\alpha,t}^2$.

Substituting into (33),

$$\Phi(\hat{\sigma}_t) \lesssim \Phi(\bar{\sigma}_t) + r_{\alpha,t}^2.$$

Expanding $\Phi(\bar{\sigma}_t)$, using $\bar{\sigma}_t = O(1)$, and absorbing the lower-order terms $\bar{\sigma}_t \sqrt{L_t/t} + \alpha_t L_t/t$ into $r_{\alpha,t}^2 + \bar{\sigma}_t^{1-p/2}/\sqrt{t}$ yields

$$\Phi(\hat{\sigma}_t) \lesssim r_{\alpha,t}^2 + \frac{\bar{\sigma}_t^{1-p/2}}{\sqrt{t}}.$$

This ends the proof. □

D REGRET ANALYSIS

In this section, we prove the master theorem, using the upper bound of the excess risk derived from conditions 5.1–5.3.

Proof of Theorem 5.5. Assume that, on the high-probability event from Condition 5.1, the one-step excess risk satisfies

$$R(\pi_{t+1}) - R(\pi^*) \lesssim \frac{\bar{\sigma}_t^{1-b}}{\sqrt{t}} + \left(\frac{\alpha_t}{t}\right)^{\frac{1}{1+b}} + \frac{\bar{\sigma}_t^2}{\alpha_t}, \quad (34)$$

where

$$\bar{\sigma}_t^2 := \frac{1}{t} \sum_{s=1}^t \sigma_s^2.$$

(As usual, the small-variance branch is absorbed into the critical-radius term $(\alpha_t/t)^{1/(1+b)}$.) Write

$$e_t := R(\pi_t) - R(\pi^*), \quad F_t := \sum_{s=1}^t e_s$$

By Condition 5.3, $\sigma_s^2 \lesssim e_s^\beta$. Since $u \mapsto u^\beta$ is concave on $[0, \infty)$ for $\beta \in (0, 1]$, Jensen's inequality gives

$$\bar{\sigma}_t^2 \lesssim \frac{1}{t} \sum_{s=1}^t e_s^\beta \leq \left(\frac{F_t}{t}\right)^\beta. \quad (35)$$

Let

$$E := \bigcap_{t \geq 1} E_t,$$

where E_t is the high-probability event from Condition 5.1 at round t . Since $\sum_{t \geq 1} \delta_t \leq \delta$, a union bound yields $\Pr(E) \geq 1 - \delta$. We work on E throughout.

We prove the claimed rates on E .

Case 1: $\beta \in (0, 1)$. Set

$$a := \frac{1}{2+b-\beta}, \quad c := \frac{1-\beta}{2+b-\beta}, \quad \alpha_t \asymp t^c.$$

We show that

$$F_t = \tilde{\mathcal{O}}(t^{1-a}),$$

which immediately implies

$$e_t = \tilde{\mathcal{O}}(t^{-a}) \quad \text{and} \quad \text{Regret}_T = \tilde{\mathcal{O}}(T^{1-a}).$$

To keep notation light, let L_t denote a generic nondecreasing polylogarithmic factor in t and $1/\delta$; its value may change from line to line.

Assume inductively that

$$F_t \lesssim t^{1-a} L_t.$$

Then by (35),

$$\bar{\sigma}_t^2 \lesssim t^{-a\beta} L_t^\beta, \quad \bar{\sigma}_t \lesssim t^{-a\beta/2} L_t^{\beta/2}.$$

Substituting into (34) gives

$$e_{t+1} \lesssim t^{-E_1} L_t^{\beta(1-b)/2} + t^{-E_2} + t^{-E_3} L_t^\beta,$$

where

$$E_1 := \frac{1}{2} + \frac{a\beta(1-b)}{2}, \quad E_2 := \frac{1-c}{1+b}, \quad E_3 := a\beta + c.$$

With the above choice of a, c ,

$$E_2 = \frac{1-c}{1+b} = \frac{1}{2+b-\beta} = a,$$

and

$$E_3 = a\beta + c = \frac{\beta}{2+b-\beta} + \frac{1-\beta}{2+b-\beta} = a.$$

Moreover,

$$E_1 - a = \frac{1}{2} + \frac{a\beta(1-b)}{2} - a = \frac{b(1-\beta)}{2(2+b-\beta)} \geq 0.$$

Hence all three terms are bounded by $\tilde{\mathcal{O}}(t^{-a})$, and therefore

$$e_{t+1} = \tilde{\mathcal{O}}(t^{-a}).$$

Summing from 1 to T yields

$$F_T = \sum_{t=1}^T e_t = \tilde{\mathcal{O}}\left(\sum_{t=1}^T t^{-a}\right) = \tilde{\mathcal{O}}(T^{1-a}),$$

since $a < 1$. This closes the induction and proves

$$e_t = \tilde{\mathcal{O}}\left(t^{-\frac{1}{2+b-\beta}}\right), \quad \text{Regret}_T = \tilde{\mathcal{O}}\left(T^{\frac{1+b-\beta}{2+b-\beta}}\right).$$

Case 2: $\beta = 1$ and $b \in (0, 1]$. Set

$$a := \frac{1}{1+b}, \quad \alpha_t := 1 + \log(et).$$

Again let L_t denote a generic nondecreasing polylogarithmic factor. Assume inductively that

$$F_t \lesssim t^{1-a} L_t.$$

Then by (35),

$$\bar{\sigma}_t^2 \lesssim t^{-a} L_t, \quad \bar{\sigma}_t \lesssim t^{-a/2} L_t^{1/2}.$$

Substituting into (34),

$$\begin{aligned} e_{t+1} &\lesssim \frac{\bar{\sigma}_t^{1-b}}{\sqrt{t}} + \left(\frac{\alpha_t}{t}\right)^{\frac{1}{1+b}} + \frac{\bar{\sigma}_t^2}{\alpha_t} \\ &\lesssim t^{-\frac{1}{2} - \frac{a(1-b)}{2}} L_t^{(1-b)/2} + t^{-\frac{1}{1+b}} \alpha_t^{1/(1+b)} + t^{-a} \frac{L_t}{\alpha_t}. \end{aligned}$$

Since $a = 1/(1+b)$,

$$\frac{1}{2} + \frac{a(1-b)}{2} = \frac{1}{1+b} = a,$$

so each term is $\tilde{\mathcal{O}}(t^{-a})$. Hence

$$e_{t+1} = \tilde{\mathcal{O}}(t^{-a}).$$

Summing yields

$$F_T = \tilde{\mathcal{O}}(T^{1-a}) = \tilde{\mathcal{O}}(T^{\frac{b}{1+b}}),$$

which proves

$$e_t = \tilde{\mathcal{O}}(t^{-1/(1+b)}), \quad \text{Regret}_T = \tilde{\mathcal{O}}(T^{b/(1+b)}).$$

Case 3: $\beta = 1$ and $b = 0$. Set $\alpha_t := 1 + \log(et)$. Then (34) becomes

$$e_{t+1} \lesssim \frac{\bar{\sigma}_t}{\sqrt{t}} + \frac{\alpha_t}{t} + \frac{\bar{\sigma}_t^2}{\alpha_t}.$$

By (35),

$$\bar{\sigma}_t^2 \lesssim \frac{F_t}{t}.$$

Hence

$$e_{t+1} \lesssim \frac{\sqrt{F_t}}{t} + \frac{\alpha_t}{t} + \frac{F_t}{t \alpha_t}. \tag{36}$$

We claim that $F_t = \mathcal{O}(\log^2 t)$. Indeed, if $F_t \leq A \log^2(et)$ for all $t \leq T$, then (36) gives

$$e_{t+1} \lesssim \frac{\log(et)}{t}.$$

Summing over $t \leq T$ yields

$$F_T = \sum_{t=1}^T e_t \lesssim \sum_{t=1}^T \frac{\log(et)}{t} = \mathcal{O}(\log^2 T),$$

so the bootstrap closes for A large enough. Therefore

$$\text{Regret}_T = \mathcal{O}(\log^2 T).$$

The one-step bound (36) then implies

$$e_t = \mathcal{O}\left(\frac{\log(et)}{t}\right),$$

and in particular

$$e_t = \mathcal{O}\left(\frac{\log^2 t}{t}\right),$$

which matches the stated theorem.

Case 4: $\beta \in (0, 1), b = 0$. Set $a := 1/(2 - \beta)$, $\gamma := (1 - \beta)/(2 - \beta)$, and $\alpha_t := 2 \vee t^\gamma$. Then (34) with $b = 0$ becomes

$$e_{t+1} \lesssim \frac{\bar{\sigma}_t}{\sqrt{t}} + \frac{\alpha_t}{t} + \frac{\bar{\sigma}_t^2}{\alpha_t}.$$

Assume inductively $F_t \leq C t^{1-a} L_t$. By (35), $\bar{\sigma}_t \lesssim t^{-a\beta/2} L_t^{\beta/2}$. The three terms decay as $t^{-E_1}, t^{-E_2}, t^{-E_3}$ with

$$E_1 = \frac{1}{2} + \frac{a\beta}{2}, \quad E_2 = 1 - \gamma = a, \quad E_3 = a\beta + \gamma = a.$$

Since $E_1 - a = (1 - a(2 - \beta))/2 = 0$, all three exponents equal a , so $e_{t+1} = \tilde{\mathcal{O}}(t^{-a})$. Summation gives

$$\text{Regret}_T = \tilde{\mathcal{O}}\left(T^{\frac{1-\beta}{2-\beta}}\right).$$

□

D.1 CLIPPING BIAS (LEMMA 6.9)

Proof of Lemma 6.9. For $s \leq t$, write

$$\rho_s^\pi := \frac{\pi(A_s | X_s)}{\pi_s(A_s | X_s)}.$$

Then

$$R_t(\pi) = \frac{1}{t} \sum_{s=1}^t \mathbb{E}[(\rho_s^\pi - 1)Y_s | \mathcal{F}_{s-1}], \quad R_{t,\alpha}(\pi) = \frac{1}{t} \sum_{s=1}^t \mathbb{E}[(\min(\rho_s^\pi, \alpha) - 1)Y_s | \mathcal{F}_{s-1}].$$

Since $\min(\rho_s^\pi, \alpha) \leq \rho_s^\pi$ and $Y_s \leq 0$,

$$R_t(\pi) \leq R_{t,\alpha}(\pi).$$

For the upper bound, using

$$\min(\rho_s^\pi, \alpha) = \rho_s^\pi - (\rho_s^\pi - \alpha)_+,$$

we get

$$R_{t,\alpha}(\pi) - R_t(\pi) = -\frac{1}{t} \sum_{s=1}^t \mathbb{E}[(\rho_s^\pi - \alpha)_+ Y_s | \mathcal{F}_{s-1}].$$

Because $Y_s \in [-1, 0]$,

$$-(\rho_s^\pi - \alpha)_+ Y_s \leq (\rho_s^\pi - \alpha)_+,$$

hence

$$R_{t,\alpha}(\pi) - R_t(\pi) \leq \frac{1}{t} \sum_{s=1}^t \mathbb{E}[(\rho_s^\pi - \alpha)_+ | \mathcal{F}_{s-1}].$$

Now, for any nonnegative Z ,

$$\mathbb{E}[Z | \mathcal{F}] = \int_0^\infty \Pr(Z \geq u | \mathcal{F}) du.$$

Applying this to $Z = (\rho_s^\pi - \alpha)_+$,

$$\mathbb{E}[(\rho_s^\pi - \alpha)_+ | \mathcal{F}_{s-1}] = \int_\alpha^\infty \Pr(\rho_s^\pi \geq v | \mathcal{F}_{s-1}) dv.$$

For $v \geq \alpha > 1$,

$$\{\rho_s^\pi \geq v\} \subseteq \{(\rho_s^\pi - 1)^2 \geq (v - 1)^2\},$$

so conditional Markov's inequality gives

$$\Pr(\rho_s^\pi \geq v | \mathcal{F}_{s-1}) \leq \frac{\mathbb{E}[(\rho_s^\pi - 1)^2 | \mathcal{F}_{s-1}]}{(v - 1)^2} = \frac{\sigma_s^2(\pi)}{(v - 1)^2}.$$

Therefore,

$$\mathbb{E}[(\rho_s^\pi - \alpha)_+ | \mathcal{F}_{s-1}] \leq \sigma_s^2(\pi) \int_\alpha^\infty \frac{dv}{(v - 1)^2} = \frac{\sigma_s^2(\pi)}{\alpha - 1}.$$

Averaging over $s = 1, \dots, t$ yields

$$R_{t,\alpha}(\pi) \leq R_t(\pi) + \frac{1}{t(\alpha - 1)} \sum_{s=1}^t \sigma_s^2(\pi).$$

Combining this with the lower bound proves (24). □

E ADDITIONAL NUMERICAL EXPERIMENTS AND DETAILS

In this section, we provide additional details about the experimental setup. All computations were performed on CPU; no GPUs, clusters, or cloud resources were used.

E.1 ERM VS. SVP

We evaluate SVP [45] under temporally dependent data using a Markov-chain design with $K = 500$ arms. For each coordinate k , generate $a_k \sim \text{Unif}[B, 1 - B]$ with $B = 0.25$, $b_k \sim \text{Unif}[0, B]$, and autocorrelation $\chi_k \sim \text{Unif}[0, \chi_{\max}]$ where $\chi_{\max} = 0.9$.

Data follows $X_{t,k} = a_k + b_k \cdot S_{t,k}$ where $(S_{t,k})$ is a Markov chain with $S_{0,k} \sim \text{Unif}\{\pm 1\}$ and persistence probability $p_k = \frac{1+\chi_k}{2}$, yielding $\text{Cov}(X_{t,k}, X_{t+1,k}) = \chi_k b_k^2$.

We compare ERM ($\arg \min_k \hat{\mu}_k$) against dependence-aware SVP:

$$\arg \min_k \left[\hat{\mu}_k + \lambda \sqrt{\frac{\hat{v}_k}{n_{\text{eff},k}}} \right], \quad n_{\text{eff},k} = n \cdot \frac{1 - \hat{\chi}_k}{1 + \hat{\chi}_k}$$

with $\lambda = 2.5$ and $\hat{\chi}_k$ the sample autocorrelation at lag 1. Performance is measured by excess risk $a_{\hat{k}} - \min_k a_k$ averaged over 1000 trials for $n \in \{50, 70, \dots, 500\}$.

E.2 SENSITIVITY TO OVERLAP IN OFFLINE CONTEXTUAL BANDITS

To study how offline learning degrades as overlap becomes weaker, we consider a synthetic contextual bandit environment with binary rewards and vary only the behavior policy. More precisely, we use a context-dependent logger of the form

$$\pi_b(a \mid x) = (1 - \varepsilon) \tilde{\pi}_b(a \mid x) + \frac{\varepsilon}{K},$$

where $\tilde{\pi}_b(\cdot \mid x)$ is a sharp contextual softmax policy and K is the number of actions. The parameter ε controls the amount of uniform exploration. Therefore the logging policy satisfies the uniform lower bound

$$\pi_b(a \mid x) \geq \frac{\varepsilon}{K} \quad \text{for all } x, a,$$

so decreasing ε directly weakens overlap.

We report results as a function of the guaranteed overlap floor ε/K . In addition to exact policy value, we also show simple overlap diagnostics computed from the full behavior policy, namely the mean minimum action probability, the mean maximum action probability, and the mean entropy of $\pi_b(\cdot \mid x)$. These diagnostics confirm that as the overlap floor decreases, the logger becomes increasingly concentrated on a small subset of actions.

Figure 5 shows that all methods deteriorate as overlap weakens, but O2PL remains the strongest method across the full sweep. The gap is clearest at moderate overlap levels, where clipping still stabilizes the objective without removing too much information from the logged sample. As the overlap floor becomes extremely small, the problem becomes difficult for all methods, but O2PL still degrades more gracefully than the standard IPW, direct-method, and doubly robust baselines.

E.3 COMPARISON TO AGNOSTIC METHODS

We evaluate O2PL on a benchmark of 100 OpenML multiclass datasets converted into contextual bandit problems, using 10 random seeds and a horizon capped at $T = 5000$ on each dataset. Rewards are shifted by +1 for all learners. For O2PL, we use the flat replay-Adam variant with learning rate 0.005, replay window and batch size 200, and warm start 50. Concretely, O2PL stores the most recent 200 interactions, and after the warm start it performs 5 Adam ascent steps every 5 interactions on the clipped-ratio pessimistic objective. The competing methods are off-the-shelf Vowpal Wabbit contextual-bandit learners using the feature map `a, ax`: Greedy, Cover-nu with ensemble size 3, Bag-greedy with ensemble size 3, and RegCB-opt in optimistic mode.

The datasets used in our experiments can be accessed at <https://www.openml.org/d/<id>>, where `<id>` is one of the following OpenML data IDs:

3, 6, 10, 11, 12, 14, 16, 18, 22, 23, 26, 28, 30, 31, 32, 36, 37, 39, 40, 41, 43, 44, 46, 48, 50, 53, 54, 59, 60, 61, 62, 150, 151, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 180, 181, 182, 183, 184, 187, 273, 275, 276, 277, 278, 279, 285, 293, 307, 310, 312, 313, 329, 333, 334, 335, 336, 337, 338, 339, 343, 346, 351, 354, 357, 375, 377, 444, 446, 448, 450, 457, 458, 459, 461, 462, 463, 464, 465, 467, 468, 469, 472, 475, 476, 477, 479, 480, 679, 682, 683.

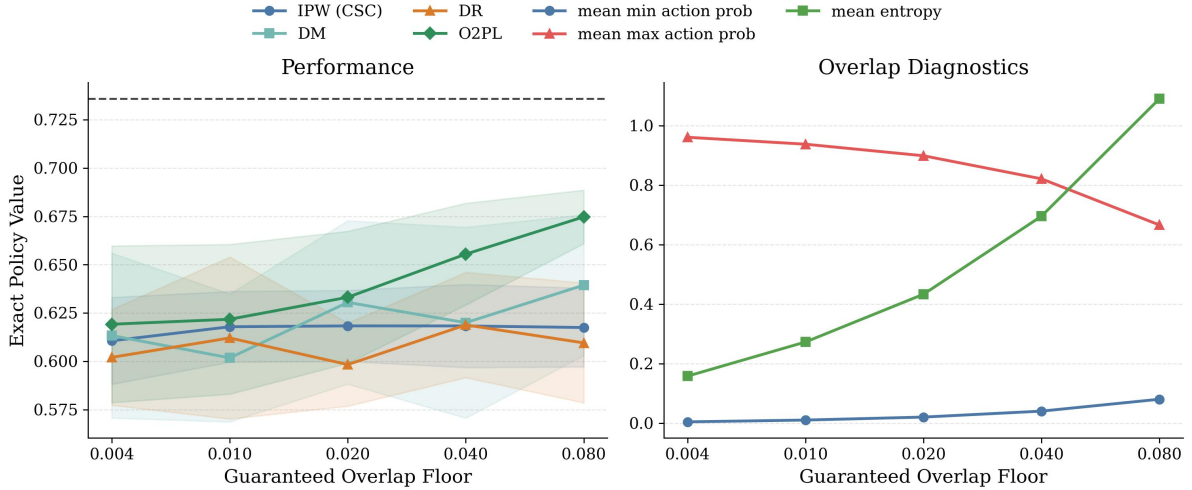


Figure 5: Sensitivity to overlap in offline contextual bandits. The left panel reports exact policy value as the guaranteed overlap floor ε/K decreases, where smaller values correspond to weaker overlap. The right panel reports overlap diagnostics for the same logging policies: the mean minimum action probability, the mean maximum action probability, and the mean entropy of the logger. As overlap weakens, the logger becomes more concentrated and all methods deteriorate, but O2PL remains the strongest method across the sweep.

Table 1: Dataset complexity for O2PL wins and losses against each baseline. Entries report the mean of dK with the number of datasets in parentheses. Significance uses a seed-level sign test with a minimum mean reward gap of 0.02.

Baseline	O2PL significant wins: mean dK (count)	O2PL significant losses: mean dK (count)
Greedy	421.16 (43)	152.24 (33)
Cover-nu	508.50 (30)	217.56 (34)
Bag-greedy	232.24 (25)	245.65 (20)
RegCB-opt	420.75 (57)	26.00 (4)

E.4 THE BENEFIT OF PESSIMISM

We also study O2PL for policy learning from logged bandit feedback, where the learner must construct a policy from a fixed batch of contextual bandit data. This setting highlights the benefit of pessimism under partial feedback.

We use the Open Bandit Pipeline (OBP) library [57], which provides both synthetic contextual bandit tasks (Figure 6) and multiclass-to-bandit reductions (Figure 7).

We compare O2PL with standard policy learning baselines in OBP: IPW with a cost-sensitive classification reduction (IPW-CSC), a direct method (DM), and a doubly robust method (DR). These baselines span importance-weighted, model-based, and doubly robust approaches to learning from logged bandit data.

All methods are trained on the same logged sample, and we report exact policy value. As shown in Figure 6, O2PL is especially strong under weak overlap and misspecification, where pessimistic policy learning is most useful. Dashed horizontal lines indicate the oracle policy value.

E.5 ADDITIONAL NUMERICAL RESULTS

We now study the relative performance of our online learning algorithm O2PL compared to the geometric batch approach in SCRM [68]. To illustrate the differences between our O2PL and SCRM, we set the following example. Gaussian policies $\pi_\theta = \mathcal{N}(\theta, \sigma^2)$ incur loss $l_t(a) = (a - y_t)^2 - 1$ with $y_t \sim \mathcal{N}(\theta^*, \sigma^2)$. Figure 8 reports loss over 15 rollouts for: (i) SCRM, trained adaptively with data from the last model π_{m-1} ; and (ii) O2PL, trained on data from all deployed models $\pi_0, \dots, \pi_{m-1}$. O2PL stabilizes the training and improves learning.

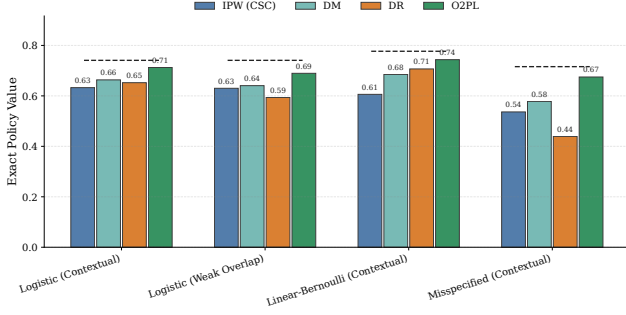


Figure 6: Exact policy value for policy learning on synthetic contextual bandit environments from logged data. The tasks vary the reward structure and the degree of overlap, including logistic, weak-overlap, linear Bernoulli, and misspecified settings.

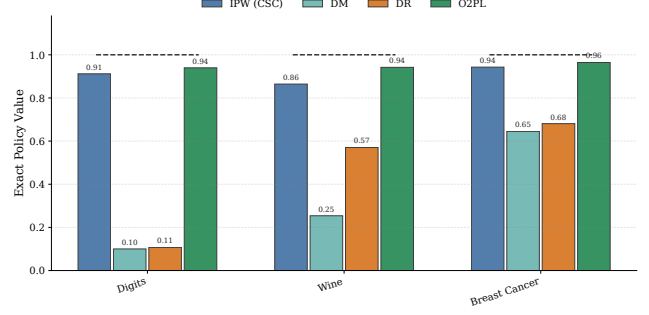


Figure 7: Exact policy value for policy learning on multiclass-to-bandit reductions from logged data. Each input is treated as a context, actions correspond to class labels, and rewards are binary. We compare the same baselines against O2PL. Dashed horizontal lines indicate the oracle policy value.

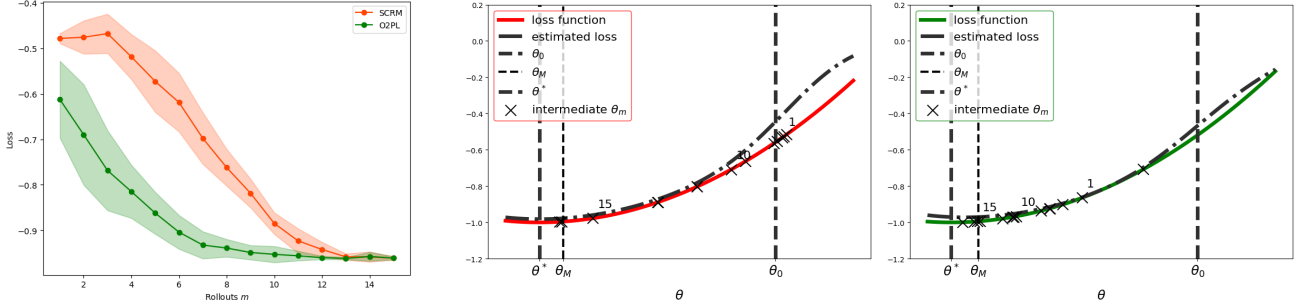


Figure 8: Loss evolution and optimization paths for (a) SCRM, (b) O2PL surrogate optimization, and (c) O2PL variance-penalized optimization.

To further compare our method to SCRM [68], we evaluate on both discrete and continuous action spaces and experiment in the batch setting.

Table 2: Loss values at $\lambda = \hat{\lambda}$ under **Uniform** and **Doubling trick**. Best (lowest) per environment within each block is in **bold**.

		Pricing	Advertising	Toy_env	Syn-5	Syn-20
Uniform	SCRM	-5.355 ± 0.196	-0.692 ± 0.026	-0.454 ± 0.002	0.534 ± 0.098	0.451 ± 0.059
	O2PL	-5.419 ± 0.269	-0.714 ± 0.003	-0.457 ± 0.002	0.230 ± 0.039	0.301 ± 0.036
Doubling	SCRM	-5.184 ± 0.406	-0.676 ± 0.048	-0.458 ± 0.003	0.570 ± 0.162	0.755 ± 0.109
	O2PL	-5.568 ± 0.036	-0.717 ± 0.002	-0.458 ± 0.003	0.226 ± 0.036	0.479 ± 0.193

Toy. Let $d = 10$. Contexts are iid $X \sim \mathcal{N}(0, I_d)$ and the action space is $\mathcal{A} = \mathbb{R}$. The environment has a fixed weight $w_{\text{env}} \in \mathbb{R}^d$; given (X, a) the stochastic reward is

$$R = (X^\top w_{\text{env}})a - a^2 + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 1) \text{ independent.}$$

Hence $\mathbb{E}[R | X, a] = (X^\top w_{\text{env}})a - a^2$ is concave in a with unique maximizer

$$a^*(X) = \frac{1}{2} X^\top w_{\text{env}}, \quad \mathbb{E}[R | X, a^*(X)] = \frac{1}{4} (X^\top w_{\text{env}})^2.$$

In the implementation, the environment returns *targets* $t(X) = a^*(X)$ and logs actions with a Gaussian behavior policy

$$A \sim \mathcal{N}(t(X), \sigma_0^2) \quad \text{with } \sigma_0 = \text{start_std} = 0.5,$$

so propensities are the corresponding Normal pdf values at the realized A . The learning loss is the negated reward, capped at 0 as in the code:

$$\ell(A, X) = \min\left\{-((X^\top w_{\text{env}})A - A^2 + \varepsilon), 0\right\}.$$

Pricing. As in [16], we use a personalized-pricing setup where a price p must be chosen from context x to maximize

$$r(x, p) = p(a(x) - b(x)p) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 1).$$

Think of $d(x, p) = a(x) + b(x)p + \varepsilon$ as an unknown, context-specific demand term. Contexts are $x \in [1, 2]^k$ with $k > 1$; only the first $\ell < k$ features affect demand (if $x = (z_1, \dots, z_\ell, \dots)$ we write $\bar{x} = (z_1, \dots, z_\ell)$). The logging policy is Gaussian, $p \sim \mathcal{N}(\bar{x}, 1)$, centered at $\bar{x}$. We use the quadratic specification $a(x) = 2x^2$ and $b(x) = 0.6x$ as in the original reference.

Advertising. As in [69], contexts $x \in \mathbb{R}^2$ are generated with `make_moons` (noise 0.05), yielding two user groups $g \in \{0, 1\}$. Each user has a latent responsiveness (target) $p_x \geq 0$ drawn from a group-specific Gaussian:

$$p_x \mid g = 0 \sim \mathcal{N}(\mu_0, \sigma_p^2), \quad p_x \mid g = 1 \sim \mathcal{N}(\mu_1, \sigma_p^2), \quad (\mu_0, \mu_1) = (3, 1), \quad \sigma_p = 0.5,$$

followed by an absolute value to enforce nonnegativity ($p_x := |p_x|$). Given a bid $a \in \mathbb{R}_+$, the one-step *reward* is

$$r(p_x, a) = \max\left(\begin{cases} \frac{a}{p_x}, & a < p_x, \\ 1 + \frac{1}{2}(p_x - a), & a \geq p_x, \end{cases}, -0.1\right),$$

and the learning loss used in experiments is the negated reward, $\ell(a, p_x) = -r(p_x, a)$. The logging policy is Gaussian,

$$a \sim \mathcal{N}(\mu_\ell, \sigma_\ell^2), \quad \mu_\ell = 2.0, \quad \sigma_\ell = 0.3,$$

with propensities computed from the exact Normal density.

Discrete actions. We construct a contextual bandit from a multi-class dataset $\{(x_i, l_i)\}_{i=1}^n$ by fitting a model $g_\theta(x) \in \mathbb{R}^K$ and defining action probabilities via softmax:

$$\pi_\theta(a \mid x) = \text{softmax}(g_\theta(x))_a, \quad a \in \{1, \dots, K\}.$$

For each context $x_i \in \mathbb{R}^d$, we sample $A_i \sim \pi_\theta(\cdot \mid x_i)$, observe binary loss $Y_i = \mathbf{1}\{A_i \neq l_i\}$, and record propensity $p_i = \pi_\theta(A_i \mid x_i)$, yielding bandit data $\mathcal{D}_{\text{bandit}} = \{(X_i, A_i, Y_i, p_i)\}_{i=1}^n$. We generate synthetic SYN data with `make_multiclass` from `scikit-learn`. The number of actions is 3 and we evaluate for context dimensions $d = 5$ ('Syn-5') and $d = 20$ ('Syn-20'). The logging policy π_0 is multinomial logistic regression (lbfgs, L_2 , $C=1$, `max_iter=10000`) trained on the first 10% of X_{train} . For a batch X , it outputs class probabilities $P \in [0, 1]^{n \times K}$, samples $A_i \sim \text{Cat}(P_i, \cdot)$, and returns propensities $w_i = P_{i, A_i}$. π_θ is linear-softmax: logits = XW with $W \in \mathbb{R}^{d \times K}$, probabilities via softmax, actions sampled by inverse-CDF; propensities $\pi_\theta(A \mid X)$ returned exactly. Greedy online evaluation uses $\arg \max_a$ logits and 0–1 loss on the test split.

Table 2 reports results for both continuous and discrete action settings. To strengthen our comparison with SCRM, we consider two sampling regimes: a fixed batch size (*uniform*) and exponentially increasing batch sizes (*doubling trick*). Across all environments, O2PL consistently matches or outperforms SCRM, highlighting the benefits of fully online updates.

Experiments were run on an average of 20 different seeds. For the discrete-action experiments, with the doubling trick we used $n_0 = 240$, $M = 5$, and an implementation floor parameter $\tau = 0.001$; with uniform splits we used $n_0 = 1500$, $M = 5$, and the same $\tau = 0.001$.

For the continuous-action environments, experiments were run on an average of 20 different seeds. With the doubling trick we used $n_0 = 150$, $M = 10$, and an implementation floor parameter $\tau = 0.001$; with uniform splits we used $n_0 = 1500$, $M = 10$, and the same $\tau = 0.001$.