

Improved Offline Contextual Bandits with Second-Order Bounds: Betting and Freezing

J. Jon Ryu

Massachusetts Institute of Technology

JONGHA@MIT.EDU

Jeongyeol Kwon

University of Wisconsin-Madison

JEONGYEOL.KWON@WISC.EDU

Benjamin Koppe

Cornell University

BEK76@CORNELL.EDU

Kwang-Sung Jun

University of Arizona

KJUN@CS.ARIZONA.EDU

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We consider off-policy selection and learning in contextual bandits, where the learner aims to select or train a reward-maximizing policy using data collected by a fixed behavior policy. Our contribution is two-fold. First, we propose a novel off-policy selection method that leverages a new betting-based confidence bound applied to an inverse propensity weight sequence. Our theoretical analysis reveals that this method achieves a significantly improved, variance-adaptive guarantee over prior work. Second, we propose a novel and generic condition on the optimization objective for off-policy learning that strikes a different balance between bias and variance. One special case, which we call freezing, tends to induce low variance, which is preferred in small-data regimes. Our analysis shows that it matches the best existing guarantees. In our empirical study, our selection method outperforms existing methods, and freezing exhibits improved performance in small-sample regimes.

Keywords: offline contextual bandits; confidence bounds; martingale; second-order bounds

1. Introduction

The offline contextual bandit problem has emerged as a critical area of study in sequential decision-making, with significant implications for decision systems for various domains including recommendation (Li et al., 2010) and online advertising (Schwartz et al., 2017). In this problem, a behavior policy $\pi_{\text{ref}}(a|x)$ is deployed in the environment for a nontrivial period of time, where the policy defines a conditional distribution over the actions a (e.g., items to be recommended) given each context information x (e.g., user being served). Specifically, at each time step $t \in [n] \triangleq \{1, \dots, n\}$, an agent observes a context $x_t \sim \mathcal{D}$ from an unknown distribution \mathcal{D} , takes an action $a_t \sim \pi_{\text{ref}}(a|x_t)$, and then receives a reward $r_t = r(x_t, a_t) \in [0, 1]$ where r is an unknown (possibly stochastic) reward function. Given offline logs of interactions $D_n \triangleq \{(x_t, a_t, r_t)\}_{t=1}^n$ obtained via the behavior policy, we wish to find a policy π that maximizes the expected reward $\mu(\pi) \triangleq \mathbb{E}_{x \sim \mathcal{D}, a \sim \pi(a|x)}[r(x, a)]$, which we call the *value* of the policy. This setting is called *off-policy*, in contrast to its online counterpart, where a policy can be updated continually using feedback from the environment. While online interactions may allow for more effective policy optimization, in many real-world scenarios this is either infeasible due to system constraints or too costly due to operational risks. The offline problem naturally arises as a viable alternative in this context.

$$\underbrace{\sqrt{\mathbb{E}\left[\frac{(\tilde{r}_1^{\pi^*} - v(\pi^*))^2}{1 + \beta(\tilde{r}_1^{\pi^*} - v(\pi^*))}\right]}}_{\text{PUB (ours; see Eq. (3.8))}} \leq \mathbb{E}\left[\frac{(\tilde{r}_1^{\pi^*} - v(\pi^*))^2}{1 + \beta(\tilde{r}_1^{\pi^*} - v(\pi^*))}\right] \lesssim \underbrace{1 + \mathbb{E}\left[\frac{(\tilde{r}_1^{\pi^*})^2}{1 + \beta\tilde{r}_1^{\pi^*}}\right]}_{\text{Sakhi et al. (2024)}} \leq \underbrace{1 + \mathbb{E}\left[\frac{(\tilde{r}_1^{\pi^*})^2}{\pi^*r + \beta\tilde{r}_1^{\pi^*}}\right]}_{\text{Gabbianelli et al. (2024)}}.$$

(a) **Off-Policy Selection**

$$\underbrace{\mathbb{E}\left[\frac{(\tilde{r}_1^{\pi^*})^2}{c_u + \gamma\tilde{r}_1^{\pi^*}}\right]}_{\text{Solve (1.3) + Assumption 4.1 (ours; see Eq. (4.2)) \& Sakhi et al. (2024)}} \leq \underbrace{\mathbb{E}\left[\frac{(\tilde{r}_1^{\pi^*})^2}{\pi^*r + \gamma\tilde{r}_1^{\pi^*}}\right]}_{\text{Gabbianelli et al. (2024)}}.$$

(b) **Off-Policy Learning**

Figure 1: Comparison of different bounds on the offline regret (see Eq. (1.2)) for the off-policy (a) selection and (b) learning where $\beta \approx \sqrt{1/n}$ in (a), and $\gamma > 0$ in (b) is a hyperparameter. We hide a factor of $\sqrt{1/n}$ and other constants. The symbol above \lesssim holds for n sufficiently large. For selection, our method achieves an improved bound. For learning, we propose a broad family of methods that achieves the same order of bound as Sakhi et al. (2024).

The main challenge in such an offline setting is in the discrepancy between the behavior policy used to log the offline data and the set of candidate policies whose performance we wish to evaluate. That is, we cannot simply use the offline data D_n to estimate the expected reward of an arbitrary policy π , since π may choose actions that are different from a_t 's chosen by π_{ref} , in which case we have not observed the corresponding rewards. This is in stark contrast to the supervised learning setup where a classifier's generalization error can be estimated by simply computing the average error on a test dataset. To circumvent the problem, researchers have proposed unbiased estimators of the expected reward of a policy, such as the Importance Weighted (IW) estimator¹ (Horvitz and Thompson, 1952; Liu et al., 2020) and Doubly Robust (DR) estimator (Robins and Rotnitzky, 1995), with numerous extensions of them. The IW estimator is defined as

$$\hat{\mu}_n^{\text{IW}}(\pi) \triangleq \frac{1}{n} \sum_{t=1}^n \tilde{r}_t^{\pi}, \quad \text{where} \quad \tilde{r}_t^{\pi} = w_t^{\pi} r_t, \quad (1.1)$$

is called the *importance-weighted reward*, and we refer to $w_t^{\pi} \triangleq \frac{\pi(a_t|x_t)}{\pi_{\text{ref}}(a_t|x_t)}$ as the importance weight.

Depending on the goal, there are three representative types of off-policy (OP) problems. Below, we contrast each with its counterpart in a supervised learning setup.

- **Off-policy evaluation:** Given a policy π , we wish to estimate its value (i.e., expected reward). In supervised learning, this corresponds to estimating the generalization error of a classifier using test data or obtaining confidence bounds for it.
- **Off-policy selection:** Given Π , a finite set of candidate policies, we wish to find the best policy—that is, the one with the highest expected reward (i.e., value). In supervised learning, this corresponds to performing model selection using hold-out validation data.

1. Also known as Inverse Propensity Score (IPS) or Inverse Propensity Weighting (IPW) estimators.

- **Off-policy learning (optimization):** Given a policy class Π (typically with $|\Pi| = \infty$, as in the case of neural-network policies), we wish to find the best policy π that achieves the highest value. In supervised learning, this corresponds to learning a classifier using training data.

In selection and learning, we wish to establish a guarantee on the *offline regret* (or suboptimality gap) for the policy $\hat{\pi} \in \Pi$ selected by an algorithm, which is defined as, for $\pi^* \triangleq \arg \max_{\pi \in \Pi} \mu(\pi)$,

$$\text{Reg}_n(\hat{\pi}) \triangleq \mu(\pi^*) - \mu(\hat{\pi}). \quad (1.2)$$

We note that the key difference between selection and learning lies in the cardinality of the policy class Π . In the selection problem, since Π is finite, we can exhaustively evaluate the performance of each policy. In the learning (optimization) problem, however, Π is typically a continuously parameterized class (e.g., neural networks), and thus solving it requires *computational efficiency* in the optimization process. In the literature, such requirement on the learning objective is called *oracle efficiency* (Langford and Zhang, 2007; Wang et al., 2024), meaning that the learning objective is optimizable efficiently assuming access to an optimization oracle. More concretely, we prefer objectives that are convex, or at least amenable to stochastic gradient-based optimization.

Contributions. In this paper, we make two main contributions. First, we propose a novel OP selection method called PUB (**P**essimism via semi-**U**nbounded-coin-**B**etting). PUB is an algorithm for computing a lower confidence bound (LCB) of any nonnegative random variable and is based on a variation of betting-based confidence bound (Waudby-Smith and Ramdas, 2024; Orabona and Jun, 2024; Ryu and Bhatt, 2024). By applying our new LCB to the importance-weighted rewards $\{\tilde{r}_t^\pi\}_{t=1}^n$ (defined in Eq. (1.1)) for each policy under consideration, we can establish a guarantee on the performance measure called offline regret (defined in Eq. (1.2)), which is strictly tighter than existing works to our knowledge. We highlight two features in our guarantee. First, our regret bound scales with the *standard deviation* of \tilde{r}^π , significantly improving the prior art scaling with the raw second moment. Second, more crucially, we achieve the improved guarantee *without any hyperparameter tuning*. This is crucial in practice as tuning a parameter in the existing estimators is infeasible in general due to the lack of knowledge on the second-order statistics of \tilde{r}^π . We summarize the comparison in Figure 1(a) and provide details in Section 3. Our LCB can be also applied to OP evaluation to construct both lower- and upper- confidence bounds, provably converging to the value of a target policy; we defer this discussion to Appendix B.6.

Second, we propose a broad family of optimization objectives for OP learning in the form of

$$\hat{\pi}_n \triangleq \arg \max_{\pi \in \Pi} \sum_{t=1}^n \phi(\beta \tilde{r}_t^\pi), \quad (1.3)$$

where β is a hyperparameter, and $\phi: \mathbb{R}_+ \rightarrow \mathbb{R}$ is a *score function*. Under a mild condition on ϕ , we show that such an optimal policy $\hat{\pi}_n$ guarantees a regret bound that depends on a second moment, matching the rate of the state-of-the-art method known as *logarithmic smoothing* (Sakhi et al., 2024). One extreme instance of our generic family is called *freezing*, in which the score function $\phi(x)$ is zero for x sufficiently large. This greatly reduces variance at the cost of introducing bias, and we empirically show that freezing achieves the best performance especially in the small-data regime. Our analysis not only matches the same smoothed second-order bound as the state-of-the-art (Sakhi et al., 2024), but also reveals that, depending on the problem instance, more aggressive methods such as freezing may be preferable. We summarize our achieved bounds along with those of existing work in Figure 1(b), and explain the details in Section 4.

Finally, we conduct an empirical evaluation of the proposed selection and learning methods, following the suite of experiments in (Wang et al., 2024). We demonstrate that PUB outperforms all existing methods, and that the new learning methods either outperform or match the performance of baseline methods. We conclude the paper by outlining promising directions for future research. Due to space constraints, we defer the discussion of related work to Appendix A.

Notation. For a random variable X , we denote its expectation and variance by $\mathbb{E}[X]$ and $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, respectively. We use $a_{1:n}$ to denote a sequence of numbers a_1, \dots, a_n . For real numbers $a, b \in \mathbb{R}$, we use shorthand notations $a \wedge b \triangleq \min\{a, b\}$ and $a \vee b \triangleq \max\{a, b\}$.

2. Problem Setting

We are given a log of interactions $D_n = \{(x_t, a_t, r_t)\}_{t=1}^n$ from a contextual bandit, collected using a behavior (or reference) policy $\pi_{\text{ref}}(a|x)$. That is, for each $t \geq 1$, $(x_t, a_t, r_t) \sim p(x)\pi_{\text{ref}}(a|x)p(r|x, a)$. Based on the bandit-logged data D_n , our goal is to evaluate the value of a target policy $\pi(a|x)$:

$$\mu(\pi) \triangleq \mathbb{E}_{(x,a,r) \sim p(x)\pi(a|x)p(r|x,a)}[r] .$$

With a slight abuse of notation, we will occasionally write $r = r(x, a)$, where $r(x, a) \in [0, 1]$ denotes a (possibly stochastic) reward function. One simple yet popular unbiased estimator for $\mu(\pi)$ is the importance weighted (IW) estimator (Horvitz and Thompson, 1952) defined in Eq. (1.1). Hereafter, we denote the variance of the importance-weighted reward by

$$\tilde{\mathbb{V}}(\pi) \triangleq \mathbb{V}[\tilde{r}_1^\pi].$$

While the IW estimator is unbiased, i.e., $\mathbb{E}[\hat{\mu}_n^{\text{IW}}(\pi)] = \mu(\pi)$, the variance $\tilde{\mathbb{V}}(\pi)$ can be undesirably large for policies that frequently choose different actions not explored by π_{ref} . The effect is exacerbated when the IW estimator is used as the objective function for selection and/or learning tasks. That is, we may end up choosing a poor policy because, if \tilde{r}_1^π exhibits disproportionately high variance, then the value can be largely overestimated with nontrivial probability.

This led to development of the *pessimism* principle (Swaminathan and Joachims, 2015), which aims to find the policy that maximizes a lower confidence bound on the value. This has the benefit of penalizing policies with large variance, effectively serving as a form of regularization to promote stability. Theoretically, pessimism is known to enjoy a property called *single-policy concentrability*, which means that the primary factor determining the convergence of the offline regret (see Eq. (1.2)) scales with a quantity that depends on the variability of the optimal policy π^* , rather than the worst-case variability over all policies $\pi \in \Pi$, a condition commonly referred to as *all-policy concentrability*. Intuitively, single-policy concentrability ensures that the convergence to the optimal policy is not affected by ill-behaved policies, as long as the optimal policy remains well-behaved.

3. Off-Policy Selection

In the selection problem, we wish to choose a policy that maximizes the expected reward from a set of finite policies. Developing a new lower confidence bound technique, we will follow the standard *pessimism* under uncertainty: construct the LCB on the expected reward for each policy, and choose the policy that maximizes the LCB. In what follows, we first introduce a betting-based (time-uniform) confidence bound for mean-parameter estimation when the random variables are $[0, \infty)$ -valued. We then show how the pessimism strategy with our confidence bound performs in the selection task (Theorem 3.3) and discuss its superiority against existing methods.

3.1. New Betting-Based Lower Confidence Bound for $[0, \infty)$ -Valued Random Variables

To construct a confidence bound, we draw ideas from gambling and the martingale theory, which have been widely used in the recent literature (Orabona and Jun, 2024; Waudby-Smith and Ramdas, 2020, 2024; Waudby-Smith et al., 2022; Ryu and Bhatt, 2024; Ryu and Wornell, 2024). The most general form of gambling is stock market investment (Cover and Thomas, 2006), but we focus on betting in a two-stock market, following the convention of Ryu and Bhatt (2024).

3.1.1. A GENERIC BETTING-BASED CONSTRUCTION

Suppose that there are two stocks, say stock 1 and stock 2. On each day $t \in \mathbb{N}$, a gambler must make her betting $\mathbf{b}_t = (b_t, 1 - b_t)$, for $b_t \in [0, 1]$, over the two stocks at the beginning. At the end of the day, the *price relative vector* $\mathbf{x}_t = (x_{t1}, x_{t2}) \in \mathbb{R}_+^2$ is revealed, where $x_{ti} > 0$ captures the multiplicative change in the price of stock i . Note that the betting \mathbf{b}_t must be causal, that is, \mathbf{b}_t can be only a function of the past observations $\mathbf{x}_{1:t-1}$. If we denote the gambler's wealth at day t by W_t , then the multiplicative gain of the wealth can be written as

$$\frac{W_t}{W_{t-1}} = \mathbf{b}_t^\top \mathbf{x}_t = b_t x_{t1} + (1 - b_t) x_{t2}.$$

If we assume that $(\mathbf{x}_t)_{t=1}^\infty$ is stochastic and satisfies $\mathbb{E}[\mathbf{x}_t | \mathbf{x}_{1:t-1}] \leq [1, 1]^\top$ (coordinate-wise), then it is easy to check that the wealth process $(W_t)_{t=1}^\infty$ is *super-martingale*, i.e., $\mathbb{E}[W_t | \mathbf{x}_{1:t-1}] \leq W_{t-1}$, regardless of the choice of \mathbf{b}_t .

Now, suppose that we have a random process $(Y_t)_{t=1}^\infty$ such that $\mathbb{E}[Y_t | Y^{t-1}] = \mathbb{E}[Y_1] \triangleq \mu$ for any $t \geq 1$, and we wish to construct a confidence set for the unknown mean parameter μ . We can then construct a confidence sequence based on betting as follows. First, we construct a *hypothetical* stock market $\mathbf{x}_t(\nu)$ as a function of candidate mean parameter ν , such that any wealth process from the market becomes super-martingale when $\nu = \mu$. If we denote the resulting wealth by $W_n(\mathbf{x}_{1:n}(\nu))$, then, by applying Ville's inequality (Ville, 1939) to a super-martingale $(W_t(\mathbf{x}_{1:t}(\mu)))_{t=1}^\infty$, we have

$$1 - \delta \leq \mathbb{P} \left(\sup_{n \geq 1} W_n(\mathbf{x}_{1:n}(\mu)) < \frac{1}{\delta} \right).$$

Given that this good event happens, we can now construct a confidence set at level $(1 - \delta)$ at each time step t , by collecting all candidate parameters that result in wealth that *does not exceed* the threshold $1/\delta$, as they cannot be μ . This outlines the general recipe for constructing confidence sequences based on betting. To derive a confidence sequence based on this meta-algorithm, one needs to specify: (1) how to construct the hypothetical stock market, and (2) which betting strategy to employ.

The construction for bounded random processes have been extensively studied in the recent literature (Waudby-Smith and Ramdas, 2020, 2024; Orabona and Jun, 2024; Ryu and Bhatt, 2024). Specifically, for a $[0, 1]$ -valued random process $(Y_t)_{t=1}^\infty$ with mean parameter $\mu \in (0, 1)$, one can set the stock market $\mathbf{x}_t(\nu) \triangleq [\frac{Y_t}{\nu}, \frac{1-Y_t}{1-\nu}]^\top$ for each $\nu \in (0, 1)$. In particular, Orabona and Jun (2024) showed that a variant of Cover's universal portfolio leads to confidence bounds that are of empirical-Bernstein type (i.e., confidence width adapts to the empirical variance), and provably never worse than the Bernoulli-KL-based confidence bound. The latter property does not hold for the empirical Bernstein bound (Maurer and Pontil, 2009) in the small-sample regime.

What if we are interested in a nonnegative random process $(Y_t)_{t=1}^\infty$ that may be unbounded (i.e., $Y_t \in [0, \infty)$)? The unbounded nature of Y_t breaks the nonnegativity of the market sequence $\mathbf{x}_t(\nu)$

in the construction above, thereby violating Ville’s inequality and preventing us from obtaining confidence bounds. As a solution, [Waudby-Smith et al. \(2022\)](#) considered a stock market which, when rephrased in the two-stock market language of [Ryu and Bhatt \(2024\)](#), takes the form

$$\mathbf{x}_t(\nu) \triangleq \left[\frac{Y_t}{\nu}, 1 \right]^\top,$$

so that the resulting wealth process remains nonnegative. Here, the first stock depends on the underlying process Y_t , while the second stock can be interpreted as *cash*. We call this a *one-sided betting* formulation. For a betting strategy $(\mathbf{x}_{1:t-1}(\nu) \mapsto \mathbf{b}_t)_{t=1}^\infty$, the cumulative wealth is then

$$W_n(\mathbf{x}_{1:n}(\nu)) \triangleq W_n(Y_{1:n}; \nu) \triangleq \prod_{t=1}^n \left(1 - b_t + b_t \frac{Y_t}{\nu} \right),$$

assuming that we start from a unit initial wealth $W_0 = 1$. Thus, for $\delta \in (0, 1)$, defining

$$C_{\text{bet}}^{(\delta)}(Y_{1:n}) \triangleq \left\{ \nu \in (0, 1) : W_n(Y_{1:n}; \nu) \leq \frac{1}{\delta} \right\} \quad \text{and} \quad \hat{\mu}_{\text{bet}}^{(\delta)}(Y_{1:n}) \triangleq \inf C_{\text{bet}}^{(\delta)}(Y_{1:n}),$$

we have that $\hat{\mu}_{\text{bet}}^{(\delta)}(Y_{1:n})$ is a $(1 - \delta)$ -lower confidence bound (LCB) for $\mathbb{E}[Y_1]$ by Ville’s inequality:

Proposition 3.1 ([Waudby-Smith et al., 2022, Proposition 1](#)) *Let $(Y_t)_{t=1}^\infty$ a non-negative real-valued random process $(Y_t)_{t=1}^\infty$ such that $\mathbb{E}[Y_t | Y^{t-1}] \triangleq \mathbb{E}[Y_1] = \mu > 0$ for any $t \geq 1$. For any causal betting strategy, $C_{\text{bet}}^{(\delta)}(Y_{1:n})$ is a (time-uniform) lower confidence set at level $1 - \delta$, that is,*

$$\mathbb{P}(\forall t \geq 1, \mu \geq \hat{\mu}_{\text{bet}}^{(\delta)}(Y_{1:t})) \geq 1 - \delta.$$

The proof of the result above, deferred to Appendix, is based on a standard martingale argument. We note that $(\hat{\mu}_{\text{bet}}^{(\delta)}(Y_{1:t}))_{t=1}^\infty$ satisfies a strong, *time-uniform* guarantee, but in its application to OP problems below, we will only invoke the LCB only for the last time step. We also note in passing that obtaining an *upper* confidence bound for $[0, \infty)$ -valued random variables is nontrivial and is beyond the scope of our work.

While any choice of betting strategy yields a valid, time-uniform LCB, we need a *good* betting strategy to obtain a *tight* (i.e., sample-efficient) LCB. Specifically, since the LCB is a random variable as well, one may wish to establish its sample efficiency of the LCB by bounding the gap between the LCB and the true mean. To make such a bound meaningful, it is important that the bound consists of *deterministic* quantities (e.g., variance) because random quantities (e.g., the empirical variance) may behave poorly—even with a large number of samples—rendering the bound unreliable. This consideration is particularly important for $[0, \infty)$ -valued random variables, for which the empirical variance may fail to converge.² To the best of our knowledge, however, existing LCBs for this setting either do not provide any sample efficiency guarantees ([Waudby-Smith et al., 2022](#)), or provide regret guarantees that scale worse than $\sqrt{\text{Var}(\tilde{r}_1^\pi)}$ (or a comparable quantity) ([Gabbianelli et al., 2024](#); [Sakhi et al., 2024](#)), which we consider to be a desirable dependence. Alternatively, some methods rely on additional assumptions, such as an upper bound on the variance or convergence of the sample variance ([Wang and Ramdas, 2023](#)). This observation motivates the novel LCB we introduce below.

2. For $[0, 1]$ -valued random variables, one can easily show that the empirical variance will converge to the true variance.

3.1.2. LCB INDUCED BY UNIVERSAL PORTFOLIO

We propose to use [Cover \(1991\)](#)’s universal portfolio as the betting strategy, in the same spirit as [Orabona and Jun \(2024\)](#), who first studied its application to *bounded* processes. A constant betting strategy $\mathbf{b}_t = (b, 1 - b)$, also referred to as a *constantly rebalanced portfolio* (CRP), yields

$$W_n^{\text{CRP}(b)}(Y_{1:n}; \nu) \triangleq \prod_{t=1}^n \left(1 - b + b \frac{Y_t}{\nu}\right) \quad (3.1)$$

as the cumulative wealth, for some $b \in [0, 1]$. [Cover \(1991\)](#) proposed a strategy called the *w-weighted universal portfolio*, or *universal portfolio* (UP) in short, to track the wealth achieved by the best CRP in hindsight asymptotically up to the first-order exponent. Cover’s UP is defined as the *mixture* of CRP wealths with respect to a mixture distribution $w(b)$ over $b \in [0, 1]$, that is,

$$W_n^{\text{UP}}(Y_{1:n}; \nu) \triangleq \int_0^1 W_n^{\text{CRP}(b)}(Y_{1:n}; \nu) w(b) db. \quad (3.2)$$

Intuitively, Cover’s UP can be understood as a buy-and-hold strategy of the set of constant betting strategies, where a unit wealth is distributed according to the weight $w(b)$. [Cover and Ordentlich \(1996\)](#) showed that, with the particular choice of weight distribution $w(b) = \frac{1}{\sqrt{\pi b(1-b)}}$, which is the density of the Beta($\frac{1}{2}, \frac{1}{2}$) distribution and the default in our work, the UP’s wealth is minimax optimal with respect to the class of CRPs. For our specific stock market, the guarantee of [Cover and Ordentlich \(1996, Theorem 2\)](#) simplifies as follows: for any sequence $y_{1:n} \in \mathbb{R}_+^n$,

$$W_n^{\text{UP}}(y_{1:n}; \nu) \geq W_n^{\text{pCRP}^*}(y_{1:n}; \nu) \triangleq \frac{1}{\sqrt{\pi(n+1)}} \sup_{b \in (0,1)} W_n^{\text{CRP}(b)}(y_{1:n}; \nu). \quad (3.3)$$

In words, this shows that the Cover’s UP can achieve the performance of the best CRP up to a polynomial factor $\sqrt{\pi(n+1)}$. We refer to the right-hand side as the *penalized best CRP wealth*.

In [Figure 2](#), we visualize the logarithmic wealth functions $\nu \mapsto \ln W_n^{\text{UP}}(Y_{1:n}; \nu)$ of different CRPs and that of Cover’s UP, at different time steps. We first note that the wealth function of each CRP is log-convex and monotonically decreasing, and thus so is that of Cover’s UP. This ensures that there exists a unique root for the equation $W_n^{\text{UP}}(Y_{1:n}; \nu) = \delta^{-1}$, and thus

$$\hat{\mu}_{\text{UP}}^{(\delta)}(Y_{1:n}) \triangleq \min \left\{ \nu > 0 : W_n^{\text{UP}}(Y_{1:n}; \nu) \leq \frac{1}{\delta} \right\} \quad (3.4)$$

is well-defined and a valid $(1 - \delta)$ -LCB. We refer to the resulting bound as the *UP-LCB*. We also remark that the curve of Cover’s UP, $\nu \mapsto W_n^{\text{UP}}(Y_{1:n}; \nu)$, closely approximates the frontier $\nu \mapsto \sup_{b \in [0,1]} W_n^{\text{CRP}(b)}(Y_{1:n}; \nu)$. This follows from the fact that Cover’s UP asymptotically tracks the wealth of the best CRP for any stock market, as implied by [Eq. \(3.3\)](#).

One can also use the penalized best CRP wealth in [Eq. \(3.3\)](#) to construct an LCB, which is slightly looser than UP-LCB, yet simpler to compute; see below for computational details. Specifically,

$$\hat{\mu}_{\text{pCRP}^*}^{(\delta)}(Y_{1:n}) \triangleq \min \left\{ \nu > 0 : W_n^{\text{pCRP}^*}(Y_{1:n}; \nu) \leq \frac{1}{\delta} \right\} \quad (3.5)$$

is a valid $(1 - \delta)$ -LCB, which we call the *penalized-best-CRP-LCB* or *pCRP*-LCB* in short.

3.1.3. FINITE-SAMPLE GUARANTEES

Our main technical contribution in this section is the following statement, which establishes the rate of convergence of UP-LCB and pCRP*-LCB to the true mean, automatically adapting to the underlying variance. For n sufficiently large, we further show that the convergence is proportional to a *smoothed variance*, defined as follows. This guarantee will be handy later for comparing the bound with (Sakhi et al., 2024). For a nonnegative random variable Y , we define a b -smoothed variance

$$\mathbb{W}_b[Y] \triangleq \mathbb{E} \left[\frac{(Y - \mathbb{E}[Y])^2}{1 + b \frac{Y - \mathbb{E}[Y]}{\mathbb{E}[Y]}} \right] = \mathbb{E}[Y] \mathbb{E} \left[\frac{(Y - \mathbb{E}[Y])^2}{bY + (1 - b) \mathbb{E}[Y]} \right]$$

for $b \in [0, 1]$. We note that $\mathbb{W}_b[Y]$ interpolates the two extreme quantities $\mathbb{W}_0[Y] = \mathbb{V}[Y]$, the variance, and $\mathbb{W}_1[Y] = \mathbb{E}[Y] \mathbb{E}[\frac{(Y - \mathbb{E}[Y])^2}{Y}]$. Under a mild assumption, i.e., $\mathbb{E}[Y] < \infty$ and $\mathbb{E}[\frac{1}{Y}] < \infty$, $b \mapsto \mathbb{W}_b[Y]$ is strictly convex, unless Y is constant with probability 1. Under such condition,

$$\mathbb{W}_b[Y] < \mathbb{W}_0[Y] \wedge \mathbb{W}_1[Y] = \mathbb{V}[Y] \wedge \mathbb{E}[Y] \mathbb{E} \left[\frac{(Y - \mathbb{E}[Y])^2}{Y} \right].$$

In what follows, we assume that $(Y_t)_{t=1}^\infty$ is an independent and identically distributed (i.i.d.), nonnegative random process, with

$$\mu \triangleq \mathbb{E}[Y_1] \quad \text{and} \quad \sigma^2 \triangleq \mathbb{V}[Y_1].$$

Theorem 3.2 (Convergence rate of UP-LCB and pCRP*-LCB) *Let $n \geq 1$ and define $F_n^{(\delta)} \triangleq \ln \frac{\sqrt{\pi(n+1)}}{\delta^2}$. Then, with probability $\geq 1 - 2\delta$,*

$$0 \leq \mu - \hat{\mu}_{\text{UP}}^{(\delta)}(Y_{1:n}) \leq \mu - \hat{\mu}_{\text{pCRP}^*}^{(\delta)}(Y_{1:n}) \leq \sqrt{\frac{48\sigma^2}{n} F_n^{(\delta)}} \vee \frac{12\mu}{n} F_n^{(\delta)}.$$

Moreover, if $n \geq 108(1 \vee 36 \frac{\mu^2}{\sigma^2}) F_n^{(\delta)}$, for $b_n^{(\delta)} \triangleq \sqrt{\frac{\mu^2}{2\sigma^2} \frac{F_n^{(\delta)}}{n}}$, we further have

$$\mu - \hat{\mu}_{\text{pCRP}^*}^{(\delta)}(Y_{1:n}) \leq \inf_{b \in (0, \frac{3}{4}]} \left\{ \frac{b}{\mu} \mathbb{W}_b[Y_1] + \frac{\mu}{b} \frac{F_n^{(\delta)}}{n} \right\} \leq 2 \sqrt{\frac{\mathbb{W}_{b_n^{(\delta)}}[Y_1]}{n} F_n^{(\delta)}}. \quad (3.6)$$

This shows that both UP-LCB and pCRP*-LCB converge to the true mean from below at the rate of $O(\sqrt{\frac{\sigma^2}{n} \ln \frac{n^{1/4}}{\delta}} \vee \frac{\mu}{n} \ln \frac{n^{1/4}}{\delta})$. It is analogous to Bernstein's inequality, but importantly, it holds uniformly over time and applies to any $[0, \infty)$ -valued random variables. To our knowledge, this is the first LCB with a finite-sample guarantee for $[0, \infty)$ -valued random variables, which is much stronger than merely statistically valid bounds.

Empirical-Bernstein-type relaxation. We can also derive a loose outer bound of the UP-LCB and pCRP*-LCB, which is simply a function of the empirical mean and variance in a similar spirit to (Orabona and Jun, 2024, Theorem 6); see Theorem B.14 for a formal statement. This relaxation can be understood as a statistically valid empirical Bernstein inequality for $[0, \infty)$ -valued random variables. While this bound is statistically valid, it does not characterize the rate of convergence or even the asymptotic consistency of the LCB as an estimator of the mean.

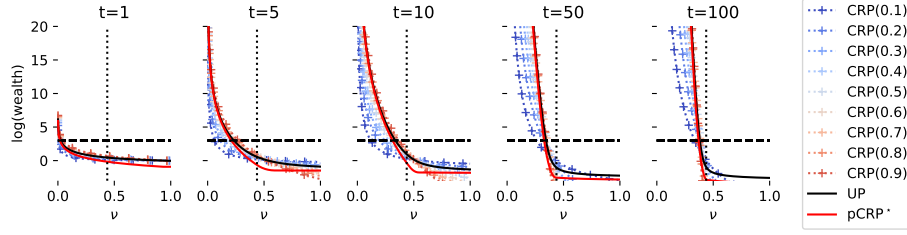


Figure 2: Example of the evolution of cumulative wealths achieved by different CRPs in Eq. (3.1), Cover’s UP in Eq. (3.2), and the penalized best CRP wealth in Eq. (3.3). The underlying process is a sequence of independent and identically distributed (i.i.d.) Gamma random variables with shape and scale parameters of 6 and $1/8$, respectively, and thus mean $3/4$.

Implementation. We can compute the UP-LCB up to numerical precision using a dynamic programming approach combined with binary search over ν , similar to the method described in (Ryu and Bhatt, 2024). Its time complexity for processing a length- n trajectory is, however, $O(n^2)$. In practice, one can implement pCRP^* -LCB (Eq. (3.5)), or techniques in Orabona and Jun (2024) or Ryu and Bhatt (2024) to compute reasonably accurate proxy in linear complexity $O(n)$. In the experiment below, we used the pCRP^* -LCB. We defer the detailed discussion about implementation to Appendix B.1, including its fast, approximate version with $O(n)$ complexity.

On the gambling technique. At first glance, using the gambling technique may seem excessive when strong time-uniformity is not needed, especially given their looser bounds. However, in this paper, Cover’s UP plays a central role due to its principled adaptation to the process statistics *without prior knowledge*, enabled by mixture wealth. Here, time-uniformity is a *byproduct* of the gambling framework, not the main goal. The $\log n$ regret term reflects the cost of this adaptation. If additional information—e.g., bounds on variance or sub-Gaussianity—were known, UP might be unnecessary and the $\log n$ term potentially avoidable. Whether this dependence can be reduced remains open.

3.2. Off-Policy Selection with Pessimism via semi-Unbounded-coin-Betting (PUB)

We propose a selection strategy, termed *Pessimism by semi-Unbounded-coin-Betting* (PUB), as follows: apply the UP-LCB $\hat{\mu}_{\text{UP}}^{(\delta)}(\cdot)$ from Eq. (3.4) or the pCRP^* -LCB $\hat{\mu}_{\text{pCRP}^*}^{(\delta)}(\cdot)$ from Eq. (3.5) to the underlying processes $\{\tilde{r}_{1:n}^\pi\}_{\pi \in \Pi}$, and select the policy with the highest lower confidence bound:

$$\hat{\pi}_{\text{UP}}^{(\delta)} \triangleq \arg \max_{\pi \in \Pi} \hat{\mu}_{\text{UP}}^{(\delta)}(\tilde{r}_{1:n}^\pi) \quad \text{and} \quad \hat{\pi}_{\text{pCRP}^*}^{(\delta)} \triangleq \arg \max_{\pi \in \Pi} \hat{\mu}_{\text{pCRP}^*}^{(\delta)}(\tilde{r}_{1:n}^\pi). \quad (3.7)$$

Hereafter, we will omit the superscript (δ) . The following guarantee is immediate from Theorem 3.2.

Theorem 3.3 (Selection) *Let $\delta' = |\Pi|/\delta$. With probability $\geq 1 - 2\delta$, for any $\pi^* \in \Pi$ and $\hat{\pi} \in \{\hat{\pi}_{\text{UP}}, \hat{\pi}_{\text{pCRP}^*}\}$, we have*

$$0 \leq \mu(\pi^*) - \mu(\hat{\pi}) \leq \sqrt{\frac{48\tilde{V}(\pi^*)}{n} F_n^{(\delta')}} \vee \frac{12\mu(\pi^*)}{n} F_n^{(\delta')}.$$

Moreover, if $n \geq 108(1 \vee 36 \frac{\mu(\pi^*)^2}{\tilde{V}(\pi^*)}) F_n^{(\delta')}$, for $b_n^{(\delta)} \triangleq \sqrt{\frac{\mu^2 F_n^{(\delta)}}{2\sigma^2 n}}$, we further have

$$\mu(\pi^*) - \mu(\hat{\pi}) \leq \inf_{b \in (0, \frac{3}{4}]} \left\{ \frac{b}{\mu(\pi^*)} \mathbb{W}_b[\tilde{r}_1^{\pi^*}] + \frac{\mu(\pi^*)}{b} \frac{F_n^{(\delta')}}{n} \right\} \leq 2\sqrt{\frac{F_n^{(\delta')}}{n} \mathbb{W}_{b_n^{(\delta')}}[\tilde{r}_1^{\pi^*}]}. \quad (3.8)$$

We compare the guarantee with that of the Logarithmic Smoothing (LS) estimator of [Sakhi et al. \(2024\)](#); see the definition in Appendix D.3. In Proposition 6 therein, it is proved that the estimator achieves the regret bound $\beta \mathbb{E} \left[\frac{(\tilde{r}_1^{\pi^*})^2}{1 + \beta \tilde{r}_1^{\pi^*}} \right] + \frac{2}{\beta n} \ln \frac{2|\Pi|}{\delta}$. First, we note that, our first term inside the infimum can be viewed as a *centered* version of their first term. Also, similar to that the first term of their regret is always bounded by $\beta \mu(\pi^*)$, our first term is also bounded by $\frac{b}{\mu(\pi^*)} \mathbb{W}_0[\tilde{r}_1^{\pi^*}] \wedge \mathbb{W}_1[\tilde{r}_1^{\pi^*}]$. In the second part of the statement, we further show in Eq. (3.8) that, for n sufficiently large, the regret scales as $\tilde{O}(\frac{1}{\sqrt{n}})$, where the leading factor is $\sqrt{\mathbb{W}_{b_n^{(\delta')}}[\tilde{r}_1^{\pi^*}]}$, which can be rewritten as

$$\sqrt{\mathbb{W}_{b_n^{(\delta')}}[\tilde{r}_1^{\pi^*}]} = \sqrt{\mathbb{E} \left[\frac{(\tilde{r}_1^{\pi^*} - \mu(\pi^*))^2}{1 + O(\sqrt{1/n})(\tilde{r}_1^{\pi^*} - \mu(\pi^*))} \right]} \quad \text{vs.} \quad \underbrace{1 + \mathbb{E} \left[\frac{(\tilde{r}_1^{\pi^*})^2}{1 + O(\sqrt{1/n})\tilde{r}_1^{\pi^*}} \right]}_{\text{Sakhi et al. (2024, Proposition 6)}}.$$

As noted in Figure 1, we note that $\sqrt{\mathbb{W}_{b_n^{(\delta')}}[\tilde{r}_1^{\pi^*}]} \approx \tilde{V}(\pi^*)$ is strictly smaller than $1 + \mathbb{E}[(\tilde{r}_1^{\pi^*})^2]$ for n sufficiently large. Arguably the most appealing property of our selection method is that we achieve these bounds in a *parameter-free* sense, unlike the existing methods that require tuning β . This is implemented by the infimum in the bound unlike the LS estimator, which shows that our estimator automatically adapts to the “optimal” hyperparameter, as a consequence of applying the wealth of Cover’s UP or penalized-best-CRP. Note that the price of adaptivity is only a logarithmic factor.

4. Off-Policy Learning

As a natural extension of the betting-based method for OP learning, we can formulate an optimization problem via Lagrange multipliers as follows:

$$\max_{\pi \in \Pi} \max_{\alpha \geq 0} \min_{\nu} \left\{ \nu + \alpha \left(\max_{b \in [0,1]} \sum_{t=1}^n \ln \left(1 + b \frac{\tilde{r}_t^{\pi} - \nu}{\nu} \right) - \ln \frac{\sqrt{\pi(n+1)}}{\delta/|\Pi|} \right) \right\}.$$

However, this optimization is not straightforward to implement in practice. Thus motivated, in this section, we consider a broad class of *pessimistic* objective functions that take the following simple form, involving a *score function* $\phi: \mathbb{R}_+ \rightarrow \mathbb{R}$, where $\beta \geq 0$ is a hyperparameter:

$$\hat{\pi}_n \triangleq \arg \max_{\pi \in \Pi} \sum_{t=1}^n \phi(\beta \tilde{r}_t^{\pi}). \quad (4.1)$$

This form of objective admits a practical optimization with stochastic-gradient-based algorithms. To guarantee statistical efficiency, we further restrict our attention to the following assumption:

Assumption 4.1 *For some $c_1, c_2 \in (0, 1]$, a score function $\phi: \mathbb{R}_+ \rightarrow \mathbb{R}$ satisfies*

$$-\ln \left(1 - x + \frac{x^2}{c_1 + c_2 x} \right) \leq \phi(x) \leq \ln(1 + x).$$

Below, we provide a few concrete examples from this family.

Proposition 4.2 (Examples of score functions) *The following satisfy Assumption 4.1:*

- *Logarithmic smoothing* ([Sakhi et al., 2024](#)): $\phi^{LS}(x) = \ln(1 + x)$ with $c_1 = c_2 = 1$.
- *Clipping*: $\phi^{clipping}(x) = \ln(1 + (x \wedge 1))$ with $c_1 = c_2 = \frac{1}{2}$.
- *Freezing*: $\phi^{freezing}(x) = \ln(1 + x \cdot \mathbb{1}\{x \leq 1\})$ with $c_1 = c_2 = \frac{1}{2}$.

The clipping score function simply truncates the score function at $\ln 2$. Note that the clipping is applied directly to \tilde{r}^π , rather than applied to w_t^π . Penalizing large values of \tilde{r}^π , clipping may help reduce variance and improve sample efficiency by implementing a more aggressive pessimism.

The freezing score function implements an even more aggressive pessimism by zeroing out \tilde{r}^π higher than 1. The potential benefit of freezing is to effectively remove samples (x_t, a_t, r_t) for policies π whose \tilde{r}_t^π is too large. This may have an even higher degree of variance reduction effect.

While Sakhi et al. (2024) also proposed a family of choices for implementing pessimism, their condition only guarantees *correctness* of the pessimism (Sakhi et al., 2024, Corollary 4), but without *sample efficiency*. Their sample efficiency result is only proved for the logarithmic smoothing. In contrast, we establish a sample-efficiency guarantee for a broad class of score functions.

Main Result. In the following, we show a smoothed second-order bound for a fairly large class of score functions satisfying Assumption 4.1. Notably, our guarantee only depends on the optimal policy π^* . The proof is deferred to Appendix C.

Theorem 4.3 (Learning) *Let $\hat{\pi}_n$ denote the estimator defined in Eq. (4.1) with a score function ϕ satisfying Assumption 4.1, and let π^* be the optimal policy. Then, with probability $\geq 1 - \delta$, we have*

$$\mu(\pi^*) - \mu(\hat{\pi}_n) \leq \beta \mathbb{E} \left[\frac{(\tilde{r}^{\pi^*})^2}{c_1 + c_2 \tilde{r}^{\pi^*}} \right] + \frac{2}{\beta n} \ln \frac{|\Pi|}{\delta} - F_\beta(\phi), \quad (4.2)$$

where we define the functional $F_\beta(\phi) \triangleq \frac{1}{\beta} \ln(\mathbb{E}[e^{\phi(\beta \tilde{r}_1(\hat{\pi}_n)) - \mathbb{E}[\beta \tilde{r}_1(\hat{\pi}_n)]}])$ as the negative influence induced by ϕ , and it satisfies $F_\beta(\phi) \geq 0$.

As a special case, our theorem recovers the guarantee of logarithmic smoothing (Sakhi et al., 2024, Proposition 6). The main term $\mathbb{E} \left[\frac{(\tilde{r}^{\pi^*})^2}{c_1 + c_2 \tilde{r}^{\pi^*}} \right]$ is a smoothed second-moment term that specializes to that of Sakhi et al. (2024) when $c_1 = c_2 = 1$. Thus, our bound inherits all the benefits such as being strictly better than IX (Gabbianelli et al., 2024) and being bounded from above with probability 1.

Additionally, different choices of ϕ induce a nontrivial tradeoff in the resulting bound, e.g.,

$$F_\beta(\phi^{\text{freezing}}) \leq F_\beta(\phi^{\text{clipping}}) \leq F_\beta(\phi^{\text{LS}}).$$

Thus, by using a score function with $c_1, c_2 < 1$ (e.g., freezing or clipping), we induce a larger negative influence, which may lead to improved performance in practice—especially in the small-sample regime—as we demonstrate in our experiments below. We believe this tradeoff arises from a delicate balance between bias and variance. A more precise characterization is left for future work.

Finally, we note that β is a hyperparameter that can be tuned using a holdout set, in conjunction with our proposed selection method from the previous section.

5. Experiments

We demonstrate the efficacy of our ideas, betting and freezing, under a synthetic, controlled experiment setup. We first demonstrate the qualitative advantage of the UP-LCB against the existing

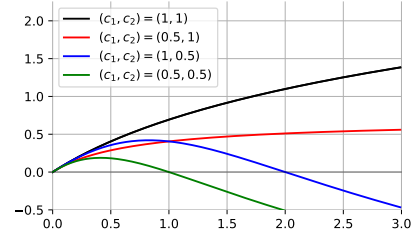


Figure 3: Examples of the score function ϕ in Assumption 4.1.

baselines under heavy-tailed distributions. We then closely follow the setting of Wang et al. (2024), to which we refer for detailed descriptions. We first present the learning experiment, followed by the selection experiment, in which we use the best policies from the learning phase across baselines.

5.1. Synthetic Evaluation of UP-LCB under Heavy-Tail Distribution

Here we empirically show that the betting-based LCB is not only statistically valid, but also converges to the target parameter in a stable manner even for heavy-tailed data. We construct a synthetic contextual bandit setting to demonstrate such robustness as follows. For a countably infinite context space $\mathcal{X} = \{1, 2, \dots\}$ and a binary action space $\mathcal{A} = \{1, 2\}$, consider: (1) context distribution: $p(x = i) = \frac{6}{\pi^2} \frac{1}{i^2}$ for $i \in \mathbb{N}$; (2) behavior policy: $\pi_{\text{ref}}(a|x = i) = \text{Bern}(a|\frac{1}{i^\beta})$; (3) reward distribution: $p(r = 1|x = i, a = 1) = 1, p(r|x = i, a = 0) = \text{Bern}(r|1 - \frac{1}{i})$. We can show that the fourth raw moment $\mathbb{E}[(\tilde{r}_t^\pi)^4]$ does not exist; see Proposition D.1 in Appendix for a formal statement.

We simulated the environment using $\beta = 3$, where a higher β leads to a heavier tail behavior of \tilde{r}_t^π , and thus a more erratic behavior for the existing baselines. We generated the trajectory of interactions of length $n = 10^4$ for $N = 100$ random trials, and visualize in Figure 4 the sample mean trajectory, the LCB based on the empirical Bernstein (EB) of Maurer and Pontil (2009), and a betting-based LCB proposed by Waudby-Smith et al. (2022) (see Appendix B.2 for its definition), Logarithmic Smoothing (LS) of Sakhi et al. (2024), and the UP-LCB, all averaged over the random trials. The shaded areas indicate empirical 10% and 90% quantiles. Here, following (Wang et al., 2024), for the EB-LCB, we used $\hat{\mu}_t(\pi) - \sqrt{2\hat{V}_t(\pi) \ln \frac{2}{\delta}}$, where $\hat{\mu}_t(\pi)$ and $\hat{V}_t(\pi)$ are the empirical mean and variance of the importance weighted rewards $(\tilde{r}_i^\pi)_{i=1}^t$.

As shown in Figure 4, the UP-LCB provides a stable lower estimate of the target mean despite the heavy tail. The estimates from the EB-LCB present erratic behaviors whenever we encounter a sample from the heavy tail. The LS-LCB is also sub-optimal as expected, being unable to adapt to the underlying variance. The

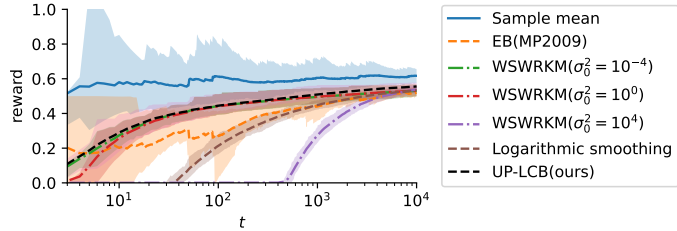


Figure 4: Comparison of UP-LCB with baselines.

WSWRKM-LCB has a hyperparameter σ_0^2 , which we set to $\{10^{-4}, 1, 10^4\}$. WSWRKM-LCBs with $\sigma_0^2 \in \{10^{-4}, 1\}$ both show a fairly close or almost equal performance to UP-LCB on average. However, these two WSWRKM-LCBs exhibit a larger variability of $\sim 7 \times 10^{-2}$ measured by the difference between the 10% and 90% sample quantiles, compared to that of $\sim 4 \times 10^{-2}$ of WSWRKM-LCB. On the other hand, for WSWRKM-LCB with $\sigma_0^2 = 10^4$, the variability was similar to UP-LCB. This shows that the hyperparameter σ_0^2 trades off variability with the tightness of the confidence bound. We also recall that the WSWRKM-LCB does not have a sample-efficiency guarantee. We defer a more extensive discussion on the WSWRKM-LCB to Appendix B.2. In Appendix D.1, we include some realizations of the experiments to further demonstrate the actual behavior for erratic trajectories.

5.2. Off-Policy Learning

Datasets. The contextual bandit data were simulated using three multi-class classification datasets from OpenML (Vanschoren et al., 2013). Each dataset has 10^6 data points. Other statistics of the

datasets are summarized in Table 3 in Appendix. For each learning method, we viewed each dataset as a multi-class regression problem, where each class corresponds to an action. We then treated a classifier, which maps a feature to a probability vector, as a deterministic policy that chooses the action of the maximum probability. Among various configurations considered in (Wang et al., 2024, Section 4), we specifically considered the real-valued cost function and a single logging policy $\pi_{\text{good}, \varepsilon=0.1}$ therein. This logging policy is defined as a random mixture of a deterministic policy (induced from a separately trained classifier) and a uniform-random policy, where $\varepsilon = 0.1$ defines the probability of using the uniform-random policy. In Appendix D.4, we report additional results with two additional policies $\pi_{\text{good}, \varepsilon=0.01}$ and $\pi_{\text{bad}, \varepsilon=0.1}$ as done in (Wang et al., 2024), where the latter combines a *badly* trained deterministic classifier with the uniform-random policy. We tested different fractions $\{0.01, 0.1, 1\}$ of datasets for training.

Baselines. We consider seven different methods in the learning experiment. The default baseline is the minimizer of the IW estimators without any regularizer (denoted as IW). We then consider a naive method (Naive, which naively maximizes $\sum_{t=1}^n \pi(a_t|x_t)r_t$ ignoring $\pi_{\text{ref}}(a_t|x_t)$), pseudo-loss (PL) (Wang et al., 2024), clipped IW (ClippedIW), Implicit Exploration (IX) (Gabbianelli et al., 2024), Logarithmic Smoothing (LS) (Sakhi et al., 2024), and lastly LS with freezing (LS+freezing), which we propose in Section 4. We include explicit definitions of the estimators in Appendix D.3. For optimization (i.e., learning), we used the linear regression approach. After training, we computed the relative improvement of each estimator against the IW baseline: (relative improvement of π) $\triangleq \frac{\hat{\mu}(\pi_{\text{IW}}) - \hat{\mu}(\pi)}{\hat{\mu}(\pi_{\text{IW}})}$. Here, we used the IW estimator to estimate the value $\hat{\mu}(\pi)$ of each policy π . Similar to Wang et al. (2024), the hyperparameter β in each estimator (see Appendix D.3) were tuned based on our PUB method with $\delta = 0.1$, using a 50/50 split of the data. For each learning method, we swept the hyperparameter β over eight values $\{0, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1\}$, and thus there are $48 = 6 \times 8$ instances of methods in total. For each dataset, we ran each experiment with 50 different seeds.

Results. The results are summarized in Figure 5. As predicted by our analysis in Section 4, LS+freezing (light blue) consistently outperforms LS (blue) in the small-sample regime and remains competitive more broadly. We attribute this to reduced variance from aggressive freezing, which effectively filters out outliers, an advantage that is especially pronounced when data is limited.

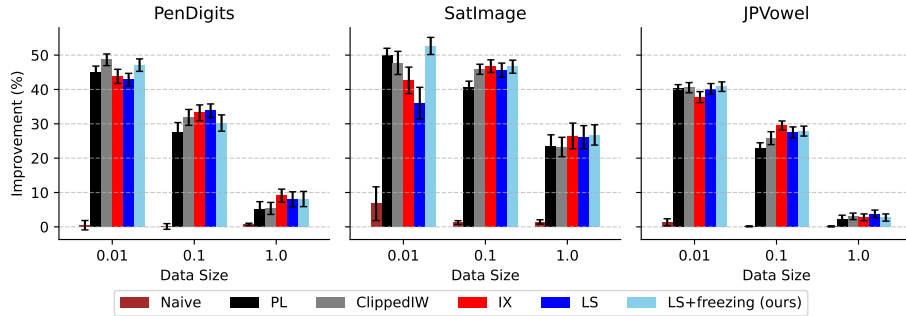


Figure 5: Results from the OP learning experiment, showing relative improvement of each method against the no-pessimism baseline. We highlight the nearly consistent improvement of LS+freezing over LS.

5.3. Off-Policy Selection

Setup. We reused the synthetic bandit data from the OP learning experiment. For selection, we considered all the 7×8 policies from the learning experiment as a policy class. For each dataset and each run, we selected the best method using PUB, LS, WSWRKM (pessimism with WSWRKM-LCB using $\sigma_0^2 = 1$), and the original empirical Bernstein (EB) of [Maurer and Pontil \(2009, Theorem 4\)](#), all with $\delta = 0.1$. Here, we aim to simulate the scenario where the practitioner tries out various policies with various hyperparameters on the training data and then chooses the best policy using the validation set.

Evaluation. We again computed the relative performance improvement of each selected policy against that of the IW baseline. To avoid misleading conclusion due to randomness, we performed the paired t -test for all pairs of the selection policies over the 50 random trials. We report the best-performing selection method and indicate statistically indistinguishable selection methods (with a p -value > 0.05) in boldface.

Results. The results are summarized in Table 1. Remarkably, for all cases, the proposed method PUB performs the best, or is statistically indistinguishable from the best. This demonstrates that PUB is not only statistically valid, but also perform empirically well, corroborating the practical benefit of variance-adaptive guarantee. We also note that, despite the competitive performance of WSWRKM-LCB in Section 5.1, this experiment reveals a failure mode of WSWRKM, particularly in the relatively large-sample regime. Additional OP selection results in Appendix D.4 highlight even more severe failure cases.

Table 1: Summary of OP selection experiment. Size is the fraction of data used for training. The best is **boldfaced** for each column, and those which do not pass paired t-test with the best (i.e., those that are not statistically distinguishable from the best) at significance level 0.05 are underlined. PUB is either the best or indistinguishable from the best.

Dataset	PenDigits			SatImage			JPVowel		
Size	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1
EB	<u>24.67</u>	35.83	11.66	<u>34.00</u>	<u>35.41</u>	24.03	<u>25.96</u>	29.75	<u>1.91</u>
LS	<u>22.82</u>	<u>35.22</u>	<u>9.99</u>	31.80	<u>30.29</u>	24.03	<u>24.39</u>	<u>27.64</u>	<u>1.65</u>
WSWRKM	36.55	<u>29.45</u>	2.19	40.75	46.24	15.68	<u>20.39</u>	9.73	<u>1.76</u>
PUB (ours)	<u>34.80</u>	<u>31.69</u>	<u>7.81</u>	<u>37.10</u>	<u>35.41</u>	24.03	33.77	<u>20.82</u>	2.08

6. Conclusion

In this paper, we have established new state-of-the-art bounds for off-policy (OP) problems. Our work opens up several interesting research directions. First, since we have developed a selection method that adapts to the variance of the IW estimator, a natural next step is to characterize the optimal rate for offline regret. Second, it is worth investigating whether similar or stronger bounds can be achieved for a doubly robust estimator. Third, it remains an open question whether one can design an objective that is amenable to optimization while matching the statistical rate of our offline selection method. Last but not least, our LCB for nonnegative random variables is, to our knowledge, the strongest in the literature—particularly in terms of convergence rate—since it requires only the existence of variance, not higher-order moments. Exploring its potential applications or extensions to other learning-theoretic problems could lead to improved guarantees and deeper insights.

Acknowledgments

KS was supported in part by the National Science Foundation under grant CCF-2327013 and Meta Platforms, Inc.

References

- Imad Aouali, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. Exponential smoothing for off-policy learning. In *Proc. Int. Conf. Mach. Learn.*, pages 984–1017, 2023.
- Thomas M Cover. Universal portfolios. *Math. Financ.*, 1(1):1–29, 1991.
- Thomas M Cover and Erik Ordentlich. Universal portfolios with side information. *IEEE Trans. Inf. Theory*, 42(2):348–363, 1996.
- Thomas M Cover and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons, 2006.
- Germano Gabbianelli, Gergely Neu, and Matteo Papini. Importance-weighted offline learning done right. In *Proc. Int. Conf. Algorithmic Learn. Theory*, pages 614–634. PMLR, 2024.
- Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.*, 47(260):663–685, 1952.
- Ying Jin, Zhimei Ren, Zhuoran Yang, and Zhaoran Wang. Policy learning “without” overlap: Pessimism and generalized empirical Bernstein’s inequality. *arXiv preprint arXiv:2212.09900*, 2022.
- Nikos Karampatziakis, John Langford, and Paul Mineiro. Empirical likelihood for contextual bandits. *Adv. Neural Inf. Proc. Syst.*, 33:9597–9607, 2020.
- Nikos Karampatziakis, Paul Mineiro, and Aaditya Ramdas. Off-policy confidence sequences. In *Proc. Int. Conf. Mach. Learn.*, pages 5301–5310, 2021.
- Ilja Kuzborskij and Csaba Szepesvári. Efron-Stein PAC-Bayesian Inequalities. *arXiv:1909.01931*, 2019.
- Ilja Kuzborskij, Claire Vernade, Andras Gyorgy, and Csaba Szepesvari. Confident off-policy evaluation and selection through self-normalized importance weighting. In *Proc. Int. Conf. Artif. Int. Statist.*, 2021.
- John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Adv. Neural Inf. Proc. Syst.*, 2007.
- Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *Proc. Int. Conf. World Wide Web*, pages 661–670, 2010.
- Lydia T Liu, Horia Mania, and Michael I Jordan. Competing Bandits in Matching Markets. In *Proc. Int. Conf. Artif. Int. Statist.*, volume 108, 2020.

- Ben London and Ted Sandler. Bayesian counterfactual risk minimization. In *Proc. Int. Conf. Mach. Learn.*, pages 4125–4133, 2019.
- Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Conf. Learn. Theory*, 2009.
- Francesco Orabona and Kwang-Sung Jun. Tight concentrations and confidence sequences from the regret of universal portfolio. *IEEE Trans. Inf. Theory*, 70(1):436–455, 2024.
- James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Statist. Assoc.*, 90(429):122–129, 1995.
- J. Jon Ryu and Alankrita Bhatt. On confidence sequences for bounded random processes via universal gambling strategies. *IEEE Trans. Inf. Theory*, 70(10):7143–7161, 2024.
- Jongha Jon Ryu and Gregory W. Wornell. Gambling-based confidence sequences for bounded random vectors. In *Proc. Int. Conf. Mach. Learn.*, volume 235, pages 42856–42869. PMLR, 21–27 Jul 2024.
- Otmame Sakhi, Pierre Alquier, and Nicolas Chopin. Pac-bayesian offline contextual bandits with guarantees. In *Proc. Int. Conf. Mach. Learn.*, pages 29777–29799, 2023.
- Otmame Sakhi, Imad Aouali, Pierre Alquier, and Nicolas Chopin. Logarithmic smoothing for pessimistic off-policy evaluation, selection and learning. In *Adv. Neural Inf. Proc. Syst.*, 2024.
- Eric M Schwartz, Eric T Bradlow, and Peter S Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.
- Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *J. Mach. Learn. Res.*, 16(52):1731–1755, 2015.
- Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013.
- Jean Ville. Etude critique de la notion de collectif. *Bull. Amer. Math. Soc*, 45(11):824, 1939.
- Hongjian Wang and Aaditya Ramdas. Catoni-style confidence sequences for heavy-tailed mean estimation. *Stoch. Process. Their Appl.*, 163:168–202, 2023.
- Lequn Wang, Akshay Krishnamurthy, and Aleksandrs Slivkins. Oracle-efficient pessimism: Offline policy optimization in contextual bandits. In *Proc. Int. Conf. Artif. Int. Statist.*, 2024.
- Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. In *Adv. Neural Inf. Proc. Syst.*, volume 33, pages 20204–20214. Curran Associates, Inc., 2020.
- Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *J. R. Stat. Soc. B*, 86(1):1–27, 2024.
- Ian Waudby-Smith, Lili Wu, Aaditya Ramdas, Nikos Karampatziakis, and Paul Mineiro. Anytime-valid off-policy inference for contextual bandits. *arXiv preprint arXiv:2210.10768*, 2022.
- Houssam Zenati, Eustache Diemert, Matthieu Martin, Julien Mairal, and Pierre Gaillard. Sequential counterfactual risk minimization. In *Proc. Int. Conf. Mach. Learn.*, pages 40681–40706, 2023.

Appendix

A Related Work	17
B Deferred Discussions and Proofs for Off-Policy Selection	18
B.1 Implementation of Proposed LCBs	18
B.1.1 Computing UP-LCB with Dynamic Programming	19
B.1.2 Computing pCRP* Wealth	20
B.1.3 Lower-Bound Universal Portfolio: A Fast Alternative	20
B.2 Comparison to the Betting Strategy of Waudby-Smith et al. (2022)	23
B.3 Proof of Theorem 3.2 (Convergence Rate Analysis for UP-LCB and pCRP*-LCB)	24
B.3.1 Proof of Theorem B.5	25
B.3.2 Proof of Theorem B.6	28
B.3.3 Technical Lemmas	30
B.4 Empirical-Bernstein-Type Relaxation of UP-LCB	32
B.5 Proof for Theorem 3.3 (Regret Analysis for PUB)	33
B.6 Off-Policy Evaluation with Betting	33
C Deferred Proofs for Off-Policy Learning	34
C.1 Proof of Proposition 4.2 (Examples of Score Functions)	34
C.2 Proof of Theorem 4.3 (Regret Analysis for Learning Algorithm)	34
D On Experiments and Additional Results	36
D.1 On the Heavy-Tail Setup in Section 5.1	36
D.2 On OP Learning and Selection Datasets	36
D.3 OP Learning Baselines	38
D.4 Additional Experiments for OP Learning and Selection	38

Appendix A. Related Work

Since the work of [Swaminathan and Joachims \(2015\)](#), there have been numerous studies on off-policy contextual bandits and reinforcement learning. An exhaustive literature review with a detailed comparison would warrant a separate survey paper. Here, we focus on categorizing representative works from a theoretical perspective based on the level of guarantees they provide.

No Finite-Time Correctness Guarantee. Methods in this category do not provide a provable guarantee of correctness for the proposed confidence bound on the performance of a given policy under evaluation. Notable examples include the empirical likelihood approach ([Karampatziakis et al., 2020](#)) and the self-normalized estimator ([Swaminathan and Joachims, 2015](#)). These lack explicit finite-time correctness guarantees, let alone sample efficiency guarantees. Moreover, coverage violation of [Karampatziakis et al. \(2020\)](#) was empirically observed in [Kuzborskij et al. \(2021, Figure 3\)](#).

Finite-Time Correctness Without Sample Efficiency Guarantee. Several approaches only come with a finite-sample correctness guarantee of the proposed confidence bound, but without a convergence rate guarantee, and consequently with no offline regret guarantee. This includes the seminal work of [London and Sandler \(2019\)](#) leveraging PAC-Bayesian bounds, exponential weighting ([Aouali](#)

et al., 2023), empirical Bernstein style bound (Sakhi et al., 2023), Efron-Stein semi-empirical bound for the self-normalized importance weight (Kuzborskij and Szepesvári, 2019; Kuzborskij et al., 2021), and betting-based bounds (Karampatziakis et al., 2021; Waudby-Smith et al., 2022).

Sample Efficiency Guarantee Under Bounded Probability Ratios. Including many works mentioned above, several works have assumed a finite upper bound on the weights $w_{1:n}^\pi$. Many works mentioned above make this assumption. Of those that provide sample efficiency guarantees, the following works either assume bounded weight or their guarantees become vacuous when the weight is unbounded: Jin et al. (2022, Corollary 4.3), Wang et al. (2024), and Zenati et al. (2023).

Sample Efficiency Guarantee Without Bounded Probability Ratios. Only recently have methods with sample efficiency guarantees that remain valid without the bounded probability ratio assumption been proposed. These methods allow the behavior policy to assign arbitrarily small probabilities to certain actions. While such bounds can still be vacuous in the worst case, they may remain meaningful even when $\pi_{\text{ref}}(a | x)$ approaches zero, depending on the distribution of the context x and the reward function. Early studies in this direction established sample efficiency guarantees that depend on empirical quantities, such as those in Jin et al. (2022, Theorem 4.1). However, these guarantees are challenging to interpret and compare with other bounds, as they depend on the specific randomness in the bound’s construction. In a seminal work, Gabbianelli et al. (2024) provided the first deterministic sample efficiency bound, which was later improved by Sakhi et al. (2024). Our work falls into this category, achieving the strongest sample efficiency guarantees for selection and evaluation while matching the bound of Sakhi et al. (2024) for learning.

Appendix B. Deferred Discussions and Proofs for Off-Policy Selection

B.1. Implementation of Proposed LCBs

In this section, we discuss the implementation of the proposed LCBs, i.e., UP-LCB and pCRP*-LCB, in detail and their complexity; see Table 2 for a summary. We make a distinction of the *online complexity* (when constructing LCB at each time step) and *offline complexity* (when constructing LCB only for the last time step). We also provide a computationally efficient version, LBUP-LCB, based on a similar trick of Ryu and Bhatt (2024).

Table 2: Comparison of complexity of different LCBs. We present the time complexity to compute a LCB for each time step in “Online Complexity” for a length- n trajectory, and that to compute a LCB only for the last step with n samples in “Offline Complexity”. Here, M denotes the maximum time complexity for a root finding procedure. For example, if we use a bisection algorithm with target precision ε , $M = O(\ln \frac{1}{\varepsilon})$.

Algorithm	Online complexity	Offline complexity	Rate guarantee
UP-LCB	$\Theta(Mn^2)$	$\Theta(n^2 + Mn)$	Yes
pCRP*-LCB	$\Theta(M^2n^2)$	$\Theta(M^2n)$	Yes
LBUP-LCB	$\Theta(Mn)$	$\Theta(n + M)$	No

B.1.1. COMPUTING UP-LCB WITH DYNAMIC PROGRAMMING

As alluded to earlier, we can compute the exact UP wealth using dynamic programming. A similar statement was proved by [Ryu and Bhatt \(2024\)](#) in the context of confidence sequences for bounded stochastic processes, and the original dynamic programming argument for UP can be found in ([Cover and Ordentlich, 1996](#)).

Recall that the wealth of UP is defined as a mixture wealth of CRPs:

$$\mathbb{W}_t^{\text{UP}}(y_{1:t}; \nu) \triangleq \int_0^1 \mathbb{W}_t^{\text{CRP}(b)}(y_{1:t}; \nu) w(b) db.$$

The following proposition holds for any weight distribution $w(b)$.

Proposition B.1 *The wealth of UP can be computed as*

$$\mathbb{W}_t^{\text{UP}}(y_{1:t}; \nu) = \sum_{k=0}^t \frac{1}{\nu^k} \psi_t^w(k) y^{(t)}(k), \quad (\text{B.1})$$

where we define

$$\psi_t^w(k) \triangleq \int_0^1 b^k (1-b)^{t-k} dw(b), \quad (\text{B.2})$$

$$y^{(t)}(k) \triangleq \sum_{x_{1:t} \in \{0,1\}^t \text{ s.t. } k(x_{1:t})=k} \prod_{i=1}^t y_i^{x_i}, \quad (\text{B.3})$$

and $k(x_{1:t}) \triangleq \sum_{i=1}^t x_i$. Furthermore, for each $t \geq 1$, we have

$$y^{(t)}(k) = \begin{cases} y^{(t-1)}(0) & \text{if } k = 0, \\ y_t y^{(t-1)}(k-1) + y^{(t-1)}(k) & \text{if } 1 \leq k \leq t-1 \\ y_t y^{(t-1)}(t-1) & \text{if } k = t. \end{cases} \quad (\text{B.4})$$

Proof We first note that we can write the cumulative wealth of any constant bettor b as

$$\mathbb{W}_t^{\text{CRP}(b)}(y_{1:t}; \nu) = \sum_{x_{1:t} \in \{0,1\}^t} \prod_{i=1}^t \left(\frac{y_i b}{\nu} \right)^{x_i} (1-b)^{1-x_i}, \quad (\text{B.5})$$

where the equality follows by the distributive law. To see Eq. (B.1), we first note that continuing from Eq. (B.5), we have

$$\begin{aligned} \mathbb{W}_t^{\text{CRP}(b)}(y_{1:t}; \nu) &= \sum_{k=0}^t \nu^{-k} b^k (1-b)^{t-k} \sum_{x_{1:t} \in \{0,1\}^t \text{ s.t. } k(x_{1:t})=k} \prod_{i=1}^t y_i^{x_i} \\ &= \sum_{k=0}^t \nu^{-k} b^k (1-b)^{t-k} y^{(t)}(k), \end{aligned} \quad (\text{B.6})$$

and thus integrating over b with respect to $w(b)$ leads to (B.1). The recursive update in Eq. (B.4) is straightforward. \blacksquare

This proposition shows that, the recursive update takes $O(t)$ at time step t , and thus the online complexity is $O(Mn^2)$. Even for the offline setting where we only need to compute the LCB with the entire samples once, we need to run the recursive update in Eq. (B.4) for each $t = 1, \dots, n$ and evaluating the wealth defined in Eq. (B.1) takes $O(t)$, which leads to the complexity $O(n^2 + Mn)$.

B.1.2. COMPUTING PCR^{*} WEALTH

Recall that the PCR^{*}-LCB in Eq. (3.5) is defined by the (unique) root ν of the equation

$$W^{\text{PCR}^*}(Y_{1:n}; \nu) = \frac{1}{\sqrt{\pi(n+1)}} \sup_{b \in (0,1)} W^{\text{CRP}(b)}(Y_{1:n}; \nu) = \frac{1}{\delta}.$$

Here, for each ν , the maximizer b can be found by finding the root of the derivative $\frac{d}{db} W^{\text{CRP}(b)}(Y_{1:n}; \nu) = 0$. Hence, we can numerically find the root by the bisection algorithm over both $\nu > 0$ and $b \in (0, 1)$. Note that the CRP wealth evaluation takes $O(t)$ at time step t , and thus computing the LCB takes $O(M^2t)$. Therefore, the online and offline complexities are $O(M^2n^2)$ and $O(M^2n)$, respectively.

B.1.3. LOWER-BOUND UNIVERSAL PORTFOLIO: A FAST ALTERNATIVE

Adapting the development of [Ryu and Bhatt \(2024\)](#) for $[0, 1]$ -valued random processes, here we present a fast alternative approach that tightly approximates the UP wealth. The idea is to directly compute a mixture of very tight lower bounds on the CRP wealths. The mixture of lower bounds can be computed efficiently by numerical integration, by viewing the lower bound as an (unnormalized) exponential family distribution. While there is no guarantee on the approximation error, the resulting bound is empirically a very good proxy to the UP-LCB, even better than the PCR^{*}-UCB, when sample size is sufficiently large; see Figure 6.

Tight Lower-Bound on CRP Wealth. We start with the following lemma from ([Ryu and Bhatt, 2024](#)). We note that ([Sakhi et al., 2024](#)) also proved a similar statement (see Lemma 10 therein), but the domain is restricted to \mathbb{R}_+ and thus not sufficient for our purpose.

Lemma B.2 ([Ryu and Bhatt, 2024, Lemma 25](#)) *For an integer $\ell \geq 1$, if we define*

$$f_\ell(t) \triangleq \begin{cases} \frac{\ln(1+t) - \sum_{k=1}^{\ell-1} \frac{(-t)^k}{k}}{\frac{(-t)^\ell}{\ell}} & \text{if } t > -1 \text{ and } t \neq 0, \\ -1 & \text{if } t = 0, \end{cases}$$

then $t \mapsto f_\ell(t)$ is continuous and strictly increasing over $(-1, \infty)$.

We can then prove the following lower bound. As noted in ([Ryu and Bhatt, 2024](#)), the positive integer $r \geq 1$ in the statement can be understood as the approximation order. Empirical results show that a higher order r results in a tighter lower bound, but we do not have a formal proof.

Lemma B.3 *For any $r \in \mathbb{N}$, $b \in [0, 1]$, and $z \geq 0$, we have*

$$\ln(1 - b + bz) \geq \sum_{k=1}^{2r-1} \frac{b^k}{k} \{(1-z)^{2r} - (1-z)^k\} + (1-z)^{2r} \ln(1-b).$$

Proof Note that the right hand side diverges to $-\infty$ and thus the inequality becomes vacuously true for $b = 1$. We now assume that $b < 1$, which ensures $b(z - 1) \geq -b > -1$. Hence, from Lemma B.2, we have $f_{2r}(b(z - 1)) \geq f_{2r}(-b)$, which is equivalent to

$$\frac{\ln(1 + b(z - 1)) - \sum_{k=1}^{2r-1} \frac{(-b(z-1))^k}{k}}{\frac{(-b(z-1))^{2r}}{2r}} \geq \frac{\ln(1 - b) - \sum_{k=1}^{2r-1} \frac{(-b)^k}{k}}{\frac{(-b)^{2r}}{2r}}.$$

Rearranging the terms concludes the proof. \blacksquare

The lower bound in the statement can be understood as the logarithm of an unnormalized exponential family distribution over z , i.e.,

$$\ln(1 - b + bz) \geq \ln \psi_r(z|b), \quad (\text{B.7})$$

where $\psi_r(z|b)$ is an unnormalized exponential family distribution defined as

$$\psi_r(z|b) \triangleq \exp(\boldsymbol{\theta}_r(b)^\top \mathbf{T}_r(z)). \quad (\text{B.8})$$

Here, $\boldsymbol{\theta}_r(b)$ is the natural parameter defined as

$$\boldsymbol{\theta}_r(b) \triangleq \begin{bmatrix} b \\ b^2/2 \\ \vdots \\ b^{2r-1}/(2r-1) \\ \ln(1-b) \end{bmatrix},$$

and $\mathbf{T}_r(z)$ is the sufficient statistics defined as

$$\mathbf{T}_r(z) \triangleq \begin{bmatrix} (1-z)^{2r} - (1-z) \\ (1-z)^{2r} - (1-z)^2 \\ \vdots \\ (1-z)^{2r} - (1-z)^{2r-1} \\ (1-z)^{2r} \end{bmatrix} = \sum_{j=0}^{2r} \binom{2r}{j} (-1)^j z^j \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} \sum_{j=0}^1 \binom{1}{j} (-1)^j z^j \\ \sum_{j=0}^2 \binom{2}{j} (-1)^j z^j \\ \vdots \\ \sum_{j=0}^{2r-1} \binom{2r-1}{j} (-1)^j z^j \\ 0 \end{bmatrix}.$$

From this definition, it is easy to check that

$$\prod_{t=1}^n \psi_r(z_t|b) = \exp\left(\boldsymbol{\theta}_r(b)^\top \sum_{t=1}^n \mathbf{T}_r(z_t)\right),$$

and $\sum_{t=1}^n \mathbf{T}_r(z_t)$ is a function of $(s_j(z_{1:n}))_{j=0}^{2r}$, where we denote the (unnormalized) empirical j -th moment for $j \in \mathbb{N}$ by

$$s_j(z_{1:n}) \triangleq \sum_{t=1}^n z_t^j.$$

This implies that the lower bound can be readily computed from the empirical moments, unlike the CRP wealth or UP wealth that requires storing the entire history $z_{1:n}$.

Mixture of Lower-Bounds on CRP Wealths. For computational tractability, we now consider a mixture weight in the form of the *conjugate prior* of $\psi_r(z|b)$, defined as

$$w_r(b; \alpha) \triangleq \frac{\exp(\theta_r(b)^\top \alpha)}{Z_r(\alpha)}. \quad (\text{B.9})$$

Here, $\alpha \in \mathbb{R}^{2r}$ is a hyperparameter of the conjugate prior, and

$$Z_r(\alpha) \triangleq \int_0^1 \exp(\theta_r(b)^\top \alpha) db$$

is the *partition function*. By the following theorem, computing the mixture of the CRP wealths with respect to this conjugate prior only requires to compute the normalization constant efficiently:

Theorem B.4 *Let $r \geq 1$. For any $y_{1:t} \in \mathbb{R}_{\geq 0}^t$ and $\nu > 0$, we have*

$$\int W_n^{\text{CRP}(b)}(y_{1:n}; \nu) w_r(b; \alpha) db \geq \frac{Z_r(\sum_{t=1}^n \mathbf{T}_r(\frac{y_t}{\nu}) + \alpha)}{Z_r(\alpha)}. \quad (\text{B.10})$$

In the special case of $r = 1$, we can compute $Z_1(\alpha)$ in an analytical form if $\alpha_1 \geq 0$:

$$Z_1(\alpha) = e^{\alpha_1} \alpha_1^{-\alpha_2-1} \gamma(\alpha_2 + 1, \alpha_1). \quad (\text{B.11})$$

Here, $\gamma(s, x) \triangleq \int_0^x t^{s-1} e^{-t} dt$ for $s > 0$ denotes the lower incomplete gamma function. For $r > 1$, we need a numerical integration library to compute the partition function.

We note that the conjugate prior is not same as the beta prior of Cover’s UP in general. In particular, however, if we set $\alpha = \mathbf{0}$, then the prior $w_r(b; \alpha)$ boils down the uniform distribution over $[0, 1]$, and the resulting mixture wealth lower bound can be viewed as a lower bound to Cover’s UP with the uniform prior (i.e., Beta(1, 1) prior). Following [Ryu and Bhatt \(2024\)](#), we refer to the resulting wealth lower bound the *lower-bound UP wealth of approximation order r* , or LBUP(r) in short. We refer to the resulting LCB as the LBUP(r)-LCB.

Implementation and Complexity. We can numerically compute the LBUP(r)-LCB using the bisection method. Since we only need to keep track of the $2r$ empirical moments $(s_j(y_{1:t}))_{j=1}^{2r}$, the storage complexity is $O(r)$ and per-step time complexity for function evaluation is $O(r)$ at any time step. Consequently, for computing the LBUP(r)-LCB, the online complexity is $O(Mnr)$ and the offline complexity is $O(n + Mr)$.

Simulation. We simulated the UP-LCB, pCRP*-LCB, and LBUP(r)-LCB for $r \in \{1, 2, 3\}$ for the same synthetic setting used in Figure 2. We generated $n = 10^4$ i.i.d. Gamma random variables with shape and scale parameters of 6 and $1/8$, respectively, and thus of mean $3/4$. The results are summarized in Figures 6, 7, and 8. In particular, we remark that LBUP(r)-LCBs (especially with $r \geq 2$) very closely approximate the UP-LCB (Figure 7) better than pCRP* in a large sample regime, exhibiting better scalability over n (Figure 8). We note, however, that we do not have a formal guarantee for the closeness of LBUP-LCB to UP-LCB, and LBUP-LCBs require some burn-in samples ($\sim 10^2$ samples in this example) to become sufficiently close to UP-LCB. For an off-policy inference setting with large-scale data, practitioners may consider using LBUP-LCB if the sample trajectory is sufficiently long, and otherwise may prefer pCRP* for guaranteed performance with moderate complexity.

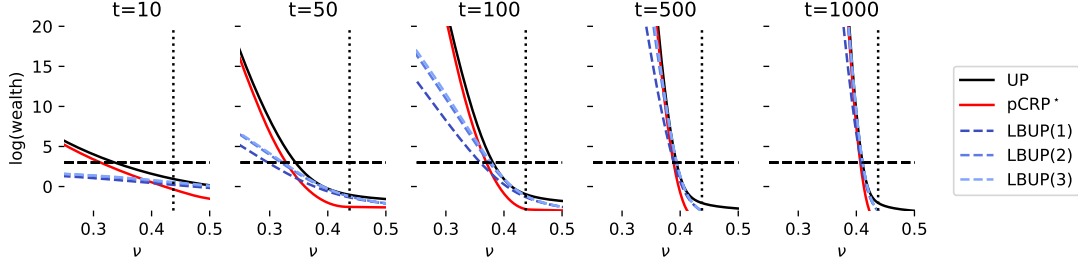


Figure 6: Example of the evolution of cumulative wealths achieved by Cover’s UP in Eq. (3.2), and the penalized best CRP wealth in Eq. (3.3), and the lower-bound universal portfolio in Appendix B.1.3. The setting is exactly same as Figure 2, except that we use larger time steps and depict with a different range for ν .

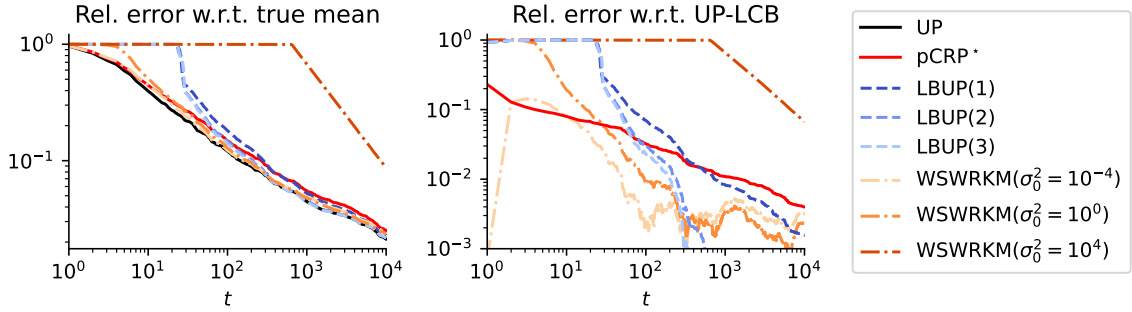


Figure 7: Convergence UP-LCB, pCRP*-LCB, and LBUP(r)-LCB for $r \in \{1, 2, 3\}$. The left panel and right panel present the relative convergence of each LCB with respect to true mean and UP-LCB, respectively. WSWRKM refers to the method in (Waudby-Smith et al., 2022), whose definition and discussion of the result can be found in Appendix B.2.

B.2. Comparison to the Betting Strategy of Waudby-Smith et al. (2022)

As alluded to earlier, Waudby-Smith et al. (2022) proposed the general construction of a time-uniform lower confidence bound for non-negative random variables, based on the one-sided betting in disguise. Beyond the meta strategy, they also suggested to use a certain betting strategy in their Eq. (12), Eq. (14), and Eq. (15). Concretely, for a hyperparameter $c \in [1/4, 3/4]$, they proposed a betting scheme defined as³

$$b_t(\nu) \triangleq \min \left\{ \nu \sqrt{\frac{2 \log \frac{1}{\delta}}{\hat{\sigma}_{t-1}^2 t \log(t+1)}}, c \right\}$$

when $\nu > 0$ is a candidate mean parameter, where

$$\hat{\sigma}_t^2 \triangleq \frac{1}{t+1} \left(\sigma_0^2 + \sum_{i=1}^t (\tilde{r}_i^\pi - \bar{r}_t^\pi)^2 \right) \quad \text{and} \quad \bar{r}_t^\pi \triangleq \min \left\{ \frac{1}{t} \sum_{i=1}^t \tilde{r}_i^\pi, 1 \right\}.$$

3. The original proposal considered a doubly robust estimator, and we simplify it by setting $k_t = 0$ therein.

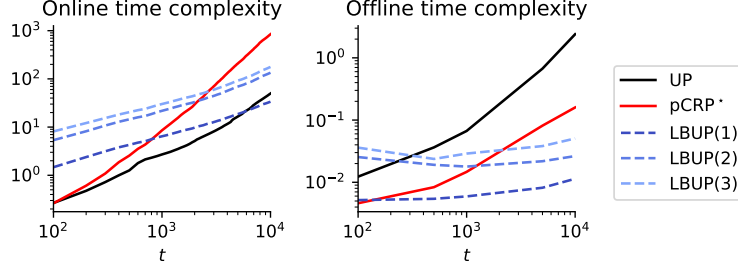


Figure 8: Online and offline time complexity for computing UP-LCB, pCRP*-LCB, and LBUP(r)-LCB for $r \in \{1, 2, 3\}$.

In the definition above, $\hat{\sigma}_t^2$ acts as a *regularized* empirical variance, where σ_0^2 , which is a hyperparameter acting as a *prior* on the variance, critically influences the performance in the small-to-moderate sample regime. When σ_0^2 is too large (compared to the true variance), the amount of betting $b_t(\nu)$ will be very small and thus will not be able to sufficiently increase the log wealth, resulting in very loose confidence bounds. When σ_0^2 is too small, the betting $b_t(\nu)$ is sensitive to the variability in the variance estimate and becomes unstable. Consequently, the confidence bounds tend to have a larger variability as we have observed in the toy experiment in Section 5.

Simulation. In the simulation setup in Appendix B.1.3, we demonstrate the performance of this method; see WSWRKM in Figure 7. In this experiment, we set $c = 1/2$ as suggested, and varied $\sigma_0^2 \in \{10^{-4}, 1, 10^4\}$ to demonstrate the effect of σ_0^2 .

We first examine the role of σ_0^2 . As alluded to above, an extremely small σ_0^2 in this case (i.e., $\sigma_0^2 = 10^{-4}$) starts off closely following the UP-LCB, but then dominated by a moderate $\sigma_0^2 = 1$ when $t \gtrsim 10^3$. If we set σ_0^2 extremely large (i.e., $\sigma_0^2 = 10^4$), it takes a significant amount of observations to result in a nonvacuous LCB. We note that this is a rather benign setting, since with Gamma random variables exhibit a light tail. The behavior of WSWRKM under a heavy-tail setting was discussed in Section 5 in the main text.

Overall, the WSWRKM-LCB performs reasonably well, but it is outperformed by pCRP* in the small-sample regime, and LBUP(r)-LCB with $r \geq 2$ in the large-sample regime. We also remark that Waudby-Smith et al. (2022) only proved its statistical validity, without establishing a finite-sample convergence rate of the LCB to the true mean. Compared to the hyperparameter-free nature of UP-LCB, the presence of additional hyperparameters, c and σ_0^2 , in WSWRKM-LCB is also undesirable in practice.

B.3. Proof of Theorem 3.2 (Convergence Rate Analysis for UP-LCB and pCRP*-LCB)

We restate Theorem 3.2 in two separate statements, and prove them separately. Technical lemmas are deferred to Appendix B.3.3.

Theorem B.5 (First part of Theorem 3.2) *Let $n \geq 1$ and define $F_n^{(\delta)} \triangleq \ln \frac{\sqrt{\pi(n+1)}}{\delta^2}$. Then, with probability $\geq 1 - 2\delta$,*

$$0 \leq \mu - \hat{\mu}_{\text{UP}}^{(\delta)}(Y_{1:n}) \leq \mu - \hat{\mu}_{\text{pCRP}^*}^{(\delta)}(Y_{1:n}) \leq \sqrt{\frac{48\sigma^2}{n} F_n^{(\delta)}} \vee \frac{12\mu}{n} F_n^{(\delta)}.$$

Recall the definition of the smoothed variance

$$\mathbb{W}_b[Y] \triangleq \mathbb{E} \left[\frac{(Y - \mathbb{E}[Y])^2}{1 + \frac{b}{\mathbb{E}[Y]}(Y - \mathbb{E}[Y])} \right].$$

Theorem B.6 (A Full Version of Second Part of Theorem 3.2) *Pick any $\varepsilon \in (0, \frac{1}{2}]$. Suppose that $(Y_t)_{t=1}^\infty$ is an independent identically distributed (i.i.d.), nonnegative random process, with $\mu \triangleq \mathbb{E}[Y_1]$ and $\sigma^2 \triangleq \mathbb{V}[Y_1]$. Let $b_n^{(\delta)} \triangleq \sqrt{\frac{\mu^2}{2\sigma^2} \frac{F_n^{(\delta)}}{n}}$. With probability $\geq 1 - 2\delta$, for any*

$$n \geq \left(12 \left(1 + \frac{4}{\varepsilon} \right) \vee 48 \left(1 + \frac{4}{\varepsilon} \right)^2 \frac{\mu^2}{\sigma^2} \right) F_n^{(\delta)},$$

we have

$$\begin{aligned} 0 \leq \mu - \hat{\mu}_{\text{UP}}(Y_{1:n}) &\leq \mu - \hat{\mu}_{\text{pCRP}^*}(Y_{1:n}) \\ &\leq \inf_{b \in (0, 1-\varepsilon]} \left\{ \frac{b}{\mu} \mathbb{W}_b[Y_1] + \frac{\mu}{b} \frac{F_n^{(\delta)}}{n} \right\} \end{aligned} \quad (\text{B.12})$$

$$\leq 2 \sqrt{\frac{F_n^{(\delta)}}{n} \mathbb{W}_{b_n^{(\delta)}}[Y_1]}. \quad (\text{B.13})$$

B.3.1. PROOF OF THEOREM B.5

Theorem B.5 is an immediate consequence of Lemma B.7 and B.8 below.

Lemma B.7 *With probability $\geq 1 - \delta$, $\mu \geq \hat{\mu}_{\text{UP}}^{(\delta)}(Y_{1:n}) \geq \hat{\mu}_{\text{pCRP}^*}^{(\delta)}(Y_{1:n})$ for any $n \geq 1$.*

Proof Since $(W_t^{\text{UP}}(Y_{1:t}; \mu))_{t=1}^n$ is a nonnegative martingale, by Ville's inequality, we have

$$\mathbb{P} \left(\sup_{t \geq 1} W_t^{\text{UP}}(Y_{1:t}; \mu) \geq \frac{1}{\delta} \right) \leq \delta,$$

which concludes the proof for the first inequality. The second inequality is trivial by 3.3 ■

Lemma B.8 *Let*

$$G_n^{(\delta)} \triangleq \sqrt{\frac{12\sigma^2}{n} F_n^{(\delta)}} \vee \frac{6\mu}{n} F_n^{(\delta)}.$$

With probability $\geq 1 - \delta$,

$$\mu \leq \hat{\mu}_{\text{pCRP}^*}^{(\delta)}(Y_{1:n}) + 2G_n^{(\delta)} \leq \hat{\mu}_{\text{UP}}^{(\delta)}(Y_{1:n}) + 2G_n^{(\delta)}$$

for any $n \geq 1$.

Proof It suffices to prove the inequality for $\hat{\mu}_{\text{pCRP}^*}^{(\delta)}(Y_{1:n})$ due to Eq. (3.3). We first note that, if $n < 12F_n^{(\delta)}(1 \vee \frac{4\sigma^2}{\mu^2})$, we deterministically have $G_n^{(\delta)} > \frac{\mu}{2}$, which implies that

$$\mu - \hat{\mu}_{\text{pCRP}^*}^{(\delta)}(Y_{1:n}) \leq \mu < 2G_n^{(\delta)},$$

which proves the claim.

Hence, hereafter, we thus assume $n \geq 12F_n^{(\delta)}(1 \vee \frac{4\sigma^2}{\mu^2})$ and show a slightly stronger bound

$$\mu - \hat{\mu}_{\text{pCRP}^*}^{(\delta)}(Y_{1:n}) \leq G_n^{(\delta)}. \quad (\text{B.14})$$

In this regime, if we define

$$\nu_o \triangleq \mu - G_n^{(\delta)},$$

we have $\nu_o > 0$, since $G_n^{(\delta)} \leq \frac{\mu}{2} < \mu$.

Recall that $\nu \mapsto W_n^{\text{pCRP}^*}(Y_{1:n}; \nu)$ is monotonically decreasing and $\hat{\mu}_{\text{pCRP}^*}^{(\delta)}(Y_{1:n})$ is the unique root of $W_n^{\text{pCRP}^*}(Y_{1:n}; \nu) = \frac{1}{\delta}$. Therefore, to prove the desired claim in Eq. (B.14), it suffices to show that

$$W_n^{\text{pCRP}^*}(Y_{1:n}; \nu_o) > \frac{1}{\delta}, \quad (\text{B.15})$$

since it implies that $\nu_o < \hat{\mu}_{\text{pCRP}^*}^{(\delta)}(Y_{1:n})$. By the definition of $W_n^{\text{pCRP}^*}$, it suffices to show that there exists $b^* \in (0, 1)$ such that

$$\frac{1}{n} \ln W_n^{\text{CRP}(b^*)}(Y_{1:n}; \nu_o) > \frac{1}{n} \ln \frac{\sqrt{\pi(n+1)}}{\delta}. \quad (\text{B.16})$$

We will construct such b^* below.

Define

$$A \triangleq \frac{\mu - \nu_o}{\nu_o} \quad \text{and} \quad B \triangleq \frac{\sigma^2 + (\mu - \nu_o)^2}{\nu_o^2},$$

and set

$$b^* \triangleq \frac{A}{2(A+B)} \leq \frac{1}{2}.$$
⁴

By applying Lemma B.9 with $\nu_o = \mu - G_n^{(\delta)}$ and b^* chosen above, we have: with probability $\geq 1 - \delta$, for any $n \geq 1$, for any $\pi^* \in \Pi$,

$$\begin{aligned} \frac{1}{n} \ln W_n^{\text{CRP}(b^*)}(Y_{1:n}; \nu_o) &\geq b^* \frac{\mu - \nu_o}{\nu_o} - \frac{(b^*)^2}{1 - b^*} \frac{\sigma^2 + (\mu - \nu_o)^2}{\nu_o^2} - \frac{1}{n} \ln \frac{1}{\delta} \\ &= b^* A - \frac{(b^*)^2}{1 - b^*} B - \frac{1}{n} \ln \frac{1}{\delta} \end{aligned}$$

4. The optimal choice of b is $1 - \sqrt{\frac{B}{A+B}}$, but the rate does not change in the current analysis.

$$= \frac{A^2}{2(A+2B)} - \frac{1}{n} \ln \frac{1}{\delta}.$$

The last equality follows from the choice of b^* .

To show Eq. (B.16), it remains to show that

$$\frac{A^2}{A+2B} \geq \frac{2F_n^{(\delta)}}{n}.$$

We prove by contradiction: if $\frac{A^2}{A+2B} < \frac{2F_n^{(\delta)}}{n}$, or equivalently

$$\frac{(\mu - \nu_o)^2}{2(\sigma^2 + (\mu - \nu_o)^2) + (\mu - \nu_o)\nu_o} < \frac{2F_n^{(\delta)}}{n}, \quad (\text{B.17})$$

then $G_n^{(\delta)} = \mu - \nu_o < G_n^{(\delta)}$. We consider the following two cases separately.

Case 1. $\sigma^2 + (\mu - \nu_o)^2 \geq (\mu - \nu_o)\nu_o$.

In this case, from Eq. (B.17), we have

$$\frac{2F_n^{(\delta)}}{n} > \frac{(\mu - \nu_o)^2}{3(\sigma^2 + (\mu - \nu_o)^2)},$$

which implies that

$$(G_n^{(\delta)})^2 = (\mu - \nu_o)^2 < \frac{\frac{6F_n^{(\delta)}}{n}}{1 - \frac{6F_n^{(\delta)}}{n}} \sigma^2 \leq \frac{12\sigma^2 F_n^{(\delta)}}{n},$$

which is a contradiction. Here, the last inequality follows from the assumption $n \geq 12F_n$.

Case 2. $\sigma^2 + (\mu - \nu_o)^2 < (\mu - \nu_o)\nu_o$.

In this case, from Eq. (B.17), we have

$$\frac{2F_n^{(\delta)}}{n} > \frac{(\mu - \nu_o)^2}{3(\mu - \nu_o)\nu_o} = \frac{\mu - \nu_o}{3\nu_o},$$

which implies that

$$G_n^{(\delta)} = \mu - \nu_o < \frac{6\nu_o F_n^{(\delta)}}{n} < \frac{6\mu F_n^{(\delta)}}{n},$$

which is a contradiction. Here, the last inequality follows since $\nu_o = \mu - G_n^{(\delta)} < \mu$. This conclude the proof. \blacksquare

B.3.2. PROOF OF THEOREM B.6

Let $\hat{v} \triangleq \hat{v}_{\text{UP}}^{(\delta)}(Y_{1:n})$. Note that $W_n^{\text{UP}}(Y_{1:n}; \hat{v}) = \frac{1}{\delta}$ by the definition of UP-LCB. Let $Z_1 \triangleq \frac{Y_1}{\hat{v}}$. Since $W_n^{\text{UP}}(Y_{1:n}; \nu) \geq W_n^{\text{pCRP}^*}(Y_{1:n}; \nu)$ from the regret guarantee of Cover's UP in Eq. (3.3), by the definition of $W_n^{\text{pCRP}^*}(Y_{1:n}; \nu)$, we have, for any $b \in (0, 1)$,

$$\begin{aligned} \frac{1}{n} \ln \frac{\sqrt{\pi(n+1)}}{\delta} &\geq \frac{1}{n} \ln W_n^{\text{CRP}(b)}(Y_{1:n}; \hat{v}) \\ &= \frac{1}{n} \sum_{t=1}^n \ln \left(1 - b + b \frac{Y_t}{\hat{v}} \right) \\ &\geq b(\mathbb{E}[Z_1] - 1) - \mathbb{E} \left[\frac{b^2(Z_1 - 1)^2}{1 + b(Z_1 - 1)} \right] - \frac{1}{n} \ln \frac{1}{\delta}, \end{aligned}$$

where the last inequality holds with probability $\geq 1 - \delta$ by Lemma B.10. We define $\Delta \triangleq \mu - \hat{v}$, and we assume that $\Delta \geq 0$, which happens with probability $\geq 1 - \delta$. Note that $\frac{\Delta}{\hat{v}} = \frac{\mu}{\hat{v}} - 1 = \mathbb{E}[Z_1] - 1 \geq 0$. Rearranging the inequality, we then have

$$\Delta \leq \hat{v} \left(b \mathbb{E} \left[\frac{(Z_1 - 1)^2}{1 + b(Z_1 - 1)} \right] + \frac{1}{b} \frac{F_n^{(\delta)}}{n} \right). \quad (\text{B.18})$$

We bound the first term as follows:

$$\begin{aligned} \mathbb{E} \left[\frac{(Z_1 - 1)^2}{1 + b(Z_1 - 1)} \right] &\leq 2 \mathbb{E} \left[\frac{(Z_1 - \mathbb{E}[Z_1])^2 + (\mathbb{E}[Z_1] - 1)^2}{1 + b(Z_1 - 1)} \right] \\ &\leq 2 \mathbb{E} \left[\frac{(Z_1 - \mathbb{E}[Z_1])^2}{1 + b(Z_1 - \mathbb{E}[Z_1])} \right] + \frac{2\Delta^2}{\hat{v}^2} \mathbb{E} \left[\frac{1}{1 + b(Z_1 - 1)} \right] \quad (\because \mathbb{E}[Z_1] \geq 1) \\ &\stackrel{(a)}{\leq} 2 \mathbb{E} \left[\frac{(Z_1 - \mathbb{E}[Z_1])^2}{1 + b(Z_1 - \mathbb{E}[Z_1])} \right] + \frac{\Delta}{2\hat{v}}. \end{aligned}$$

Here, we show that (a) is true given $n \geq (\frac{12}{c} \vee \frac{48\sigma^2}{c^2\mu^2})F_n^{(\delta)}$ for $c = \frac{\varepsilon}{\varepsilon+4}$ and $b \in (0, 1 - \varepsilon]$. To see this, Theorem 3.2 along with the requirement on n ensures

$$0 \leq \Delta = \mu - \hat{v} \leq c\mu,$$

which is equivalent to

$$\frac{4}{\varepsilon + 4} \mu = (1 - c)\mu \leq \hat{v} \leq \mu \quad (\text{B.19})$$

or

$$0 \leq \Delta \leq \frac{c}{1 - c} \hat{v} = \frac{\varepsilon}{4} \hat{v}.$$

This leads to, using $Z_1 \geq 0$ and $b \in (0, 1 - \varepsilon]$,

$$\frac{2\Delta^2}{\hat{v}^2} \mathbb{E} \left[\frac{1}{1 + b(Z_1 - 1)} \right] \leq \frac{2\Delta^2}{\hat{v}^2} \frac{1}{1 - b} \leq \frac{2\Delta}{\hat{v}^2} \frac{\Delta}{\varepsilon} \leq \frac{2\Delta}{\hat{v}^2} \frac{\hat{v}}{4} = \frac{\Delta}{2\hat{v}},$$

concluding the proof of (a) above.

We now apply the upper bound of $\mathbb{E}\left[\frac{(Z_1-1)^2}{1+b(Z_1-1)}\right]$ above to Eq. (B.18) and solve it for Δ to obtain

$$\Delta \leq (2-b)\Delta \leq \frac{4b}{\hat{v}} \mathbb{E}\left[\frac{(Y_1-\mu)^2}{1+\frac{b}{\hat{v}}(Y_1-\mu)}\right] + \frac{\hat{v}}{b} \frac{F_n^{(\delta)}}{n} \triangleq h\left(\frac{b}{\hat{v}}\right), \quad (\text{B.20})$$

where $h(q) \triangleq 4q \mathbb{E}\left[\frac{(Y_1-\mu)^2}{1+q(Y_1-\mu)}\right] + \frac{1}{q} \frac{F_n^{(\delta)}}{n}$. Taking infimum over $b \in (0, 1-\varepsilon]$,

$$\Delta \leq \inf_{b \in [0, 1-\varepsilon]} h\left(\frac{b}{\hat{v}}\right) = \inf_{q \in [0, \frac{1-\varepsilon}{\hat{v}})} h(q) \leq \inf_{q \in [0, \frac{1-\varepsilon}{\mu})} h(q) = \inf_{b \in [0, 1-\varepsilon]} h\left(\frac{b}{\mu}\right).$$

The second inequality holds since we assume $\mu \geq \hat{v}$. This concludes the proof for the first inequality in Eq. (B.12).

To prove the second inequality in Eq. (B.13), we rewrite the inequality in Eq. (B.12) as

$$\Delta \leq \inf_{b \in (0, 1-\varepsilon]} \{f(b) + g(b)\} \leq \inf_{b \in (0, \frac{1}{4}]} \{f(b) + g(b)\}, \quad (\text{B.21})$$

where $f(b) \triangleq \frac{b}{\mu} \mathbb{E}\left[\frac{(Y_1-\mu)^2}{1+\frac{b}{\mu}(Y_1-\mu)}\right]$ and $g(b) \triangleq \frac{\mu}{b} \frac{F_n^{(\delta)}}{n}$. Note that $f(b)$ is monotonically increasing and $g(b)$ is monotonically decreasing over $b \in [0, 1]$. We now show that $f(\frac{1}{4}) \geq g(\frac{1}{4})$, which implies that $f(b_o) = g(b_o)$ for some $0 < b_o \leq \frac{1}{4}$. To show this, note that

$$\begin{aligned} f\left(\frac{1}{4}\right) &= \mathbb{E}\left[\frac{(Y_1-\mu)^2}{4\mu + (Y_1-\mu)}\right] \\ &= \mathbb{E}\left[\frac{(Y_1-\mu)^2}{3\mu + Y_1}\right] \\ &\geq \mathbb{E}\left[\frac{(Y_1-\mu)^2}{2Y_1} \mathbf{1}\{Y_1 \geq 3\mu\}\right] \\ &\geq \mathbb{E}\left[\frac{\frac{Y_1^2}{2} - \mu^2}{2Y_1} \mathbf{1}\{Y_1 \geq 3\mu\}\right] && (\because (a-b)^2 \geq \frac{1}{2}a^2 - b^2) \\ &\geq \mathbb{E}\left[\frac{\frac{Y_1^2}{2} - \frac{Y_1^2}{9}}{2Y_1} \mathbf{1}\{Y_1 \geq 3\mu\}\right] \\ &\geq \mathbb{E}\left[\frac{7}{36} Y_1 \mathbf{1}\{Y_1 \geq 3\mu\}\right] \\ &\geq \frac{7}{12} \mu \\ &\geq 4\mu \frac{F_n^{(\delta)}}{n} = g\left(\frac{1}{4}\right). \end{aligned}$$

Here, the last inequality follows from the assumption that $n \geq 7F_n^{(\delta)}$. Hence, if we plug in the root b_o to Eq. (B.21), then we have

$$\Delta \leq f(b_o) + g(b_o) = 2\sqrt{\frac{F_n^{(\delta)}}{n} \mathbb{E}\left[\frac{(Y_1-\mu)^2}{1+\frac{b_o}{\mu}(Y_1-\mu)}\right]}.$$

To further upper bound this term, it suffices to find a deterministic lower bound on b_o , since, by Lemma B.13 stated below, if $0 \leq b_\ell \leq b_o$,

$$\mathbb{E} \left[\frac{(Y_1 - \mu)^2}{1 + \frac{b_o}{\mu}(Y_1 - \mu)} \right] \leq 2 \mathbb{E} \left[\frac{(Y_1 - \mu)^2}{1 + \frac{2b_\ell}{\mu}(Y_1 - \mu)} \right].$$

To find such a lower bound b_ℓ , we note that, if we define $\eta(b) \triangleq \frac{2\sigma^2}{\mu}b \geq f(b)$ and $b \mapsto \eta(b)$ is monotonically increasing, and thus the root b'_o of the equation $\eta(b) = g(b)$ must be smaller than b_o . Hence, solving $\eta(b) = \frac{2\sigma^2}{\mu}b = \frac{\mu}{b} \frac{F_n^{(\delta)}}{n} = g(b)$ yields the root

$$b'_o = b_n^{(\delta)} \triangleq \sqrt{\frac{\mu^2}{2\sigma^2} \frac{F_n^{(\delta)}}{n}}.$$

Note that we require $n > \frac{\mu^2}{\sigma^2} F_n^{(\delta)}$ to ensure that the root $b_n^{(\delta)}$ lies in $(0, \frac{1}{2})$, which is assumed in the statement. Finally, we have $\Delta \leq f(b_o) + g(b_o) \leq f(b_n^{(\delta)}) + g(b_n^{(\delta)})$, which concludes the proof. ■

B.3.3. TECHNICAL LEMMAS

Here, we state and prove technical lemmas used in the proofs above. Note that we obtain time-uniform guarantees below immediately by applying Ville's inequality in place of Markov's inequality.

Lemma B.9 *Let Y_1, \dots, Y_t be i.i.d. nonnegative random variables. For any “betting” $b \in [0, 1]$ and a “reference” mean $\nu > 0$, we have*

$$\mathbb{P} \left(\frac{1}{n} \sum_{t=1}^n \ln \left(1 - b + b \frac{Y_t}{\nu} \right) \geq b \frac{\mu - \nu}{\nu} - \frac{b^2}{1-b} \frac{\mathbb{V}[Y_1] + (\mu - \nu)^2}{\nu^2} - \frac{1}{n} \ln \frac{1}{\delta} \right) \geq 1 - \delta.$$

Proof Applying Lemma B.12 to Lemma B.10 concludes the proof. ■

Lemma B.10 *Let Y_1, \dots, Y_t be i.i.d. nonnegative random variables. For any “betting” $b \in [0, 1]$ and a “reference” mean $\nu > 0$, we have*

$$\mathbb{P} \left(\frac{1}{n} \sum_{t=1}^n \ln \left(1 - b + b \frac{Y_t}{\nu} \right) \geq b \frac{\mu - \nu}{\nu} - \mathbb{E} \left[\frac{b^2 \frac{(Y_1 - \nu)^2}{\nu^2}}{1 + b \frac{Y_1 - \nu}{\nu}} \right] - \frac{1}{n} \ln \frac{1}{\delta} \right) \geq 1 - \delta.$$

Proof Use Lemma B.11 with $Z_t \leftarrow b \frac{Y_t - \nu}{\nu}$. ■

Lemma B.11 *Let Z_1, \dots, Z_t be i.i.d. random variables supported over $(-1, \infty)$. Then, we have*

$$\mathbb{P} \left(-\frac{1}{n} \sum_{t=1}^n \ln (1 + Z_t) + \mathbb{E}[Z_1] \leq \mathbb{E} \left[\frac{Z_1^2}{1 + Z_1} \right] + \frac{1}{n} \ln \frac{1}{\delta} \right) \geq 1 - \delta.$$

Proof Note that the following is a nonnegative random variable.

$$M_n = \prod_{t=1}^n \frac{1}{\mathbb{E}\left[\frac{1}{1+Z_t}\right]}.$$

Thus, by Markov's inequality $\mathbb{P}(\ln \frac{M_n}{\mathbb{E}[M_n]} \geq \ln \frac{1}{\delta}) \leq \delta$, or equivalently, w.p. at least $1 - \delta$, we have

$$\begin{aligned} -\frac{1}{n} \sum_{t=1}^n \ln(1 + Z_t) - \frac{1}{n} \ln \frac{1}{\delta} &< \ln \mathbb{E}\left[\frac{1}{1 + Z_1}\right] \\ &\leq \mathbb{E}\left[\frac{1}{1 + Z_1}\right] - 1 \quad (\because \ln(x) \leq x - 1) \\ &= \mathbb{E}\left[\frac{-Z_1}{1 + Z_1}\right] \\ &= \mathbb{E}\left[\frac{Z_1^2}{1 + Z_1}\right] - \mathbb{E}[Z_1]. \end{aligned}$$

The last equality holds since $-\frac{t}{1+t} = \frac{t^2}{1+t} - t$. ■

Lemma B.12 *We have*

$$\mathbb{E}\left[\frac{b^2 \frac{(Y_1 - \nu)^2}{\nu^2}}{1 + b \frac{Y_1 - \nu}{\nu}}\right] \leq \frac{b^2}{1 - b} \frac{\mathbb{V}[Y_1] + (\mu - \nu)^2}{\nu^2}.$$

Proof Consider

$$\frac{(Y_1 - \nu)^2}{1 + b \frac{Y_1 - \nu}{\nu}} = \frac{\nu(Y_1 - \nu)^2}{bY_1 + (1 - b)\nu} \leq \frac{(Y_1 - \nu)^2}{(1 - b)}.$$

Taking the expectation, we have $\mathbb{E}[(Y_1 - \nu)^2] = \mathbb{E}[(Y_1 - \mu + \mu - \nu)^2] = \mathbb{V}[Y_1] + (\mu - \nu)^2$, which concludes the proof. ■

Lemma B.13 *For any $y \geq 0$, $0 \leq b' \leq b \leq \frac{1}{2}$, we have*

$$\frac{1}{1 + b \frac{y - \mu}{\mu}} \leq \frac{2}{1 + 2b' \frac{y - \mu}{\mu}}.$$

Proof Note that the denominators in both sides are positive. Hence, the inequality is equivalent to

$$1 + 2b' \frac{y - \mu}{\mu} \leq 2 + 2b \frac{y - \mu}{\mu} \Leftrightarrow (b - b') \left(1 - \frac{y}{\mu}\right) \leq 1.$$

The last inequality readily follows from the assumptions $0 \leq b' \leq b \leq \frac{1}{2}$ and $y \geq 0$. ■

B.4. Empirical-Bernstein-Type Relaxation of UP-LCB

As alluded to earlier in Section 3.1.3, here we provide an empirical-Bernstein-type relaxation of pCRP*-LCB.

Theorem B.14 (Empirical-Bernstein-type relaxation of pCRP*-LCB) *Let $\hat{\mu}_n \triangleq \frac{1}{n} \sum_{t=1}^n Y_t$ and $\hat{V}_n \triangleq \frac{1}{n} \sum_{t=1}^n (Y_t - \hat{\mu}_n)^2$ denote the empirical mean and variance, respectively. Let $H_n^{(\delta)} \triangleq \ln \frac{\sqrt{\pi(n+1)}}{\delta}$ and let $\hat{\mu}_{\text{EB}}^{(\delta)}(Y_{1:n}) \triangleq \hat{\mu}_n - \Delta_n^{(\delta)}$, where*

$$\Delta_n^{(\delta)} \triangleq \frac{1}{1 - \frac{2}{n} H_n^{(\delta)}} \left(\frac{\hat{\mu}_n}{n} H_n^{(\delta)} + \sqrt{\frac{\hat{\mu}_n^2}{n^2} (H_n^{(\delta)})^2 + \frac{4\hat{V}_n}{n} H_n^{(\delta)} \left(1 - \frac{2}{n} H_n^{(\delta)}\right)} \right).$$

Under the same setting of Proposition 3.1, with probability at least $1 - \delta$, for all $n \geq 1$ such that $H_n^{(\delta)} < \frac{1}{2}$, we have $\mu \geq \hat{\mu}_{\text{EB}}^{(\delta)}(Y_{1:n})$.

Proof By Ville's inequality, with probability $1 - \delta$, we have, for any $n \geq 1$,

$$\ln \frac{1}{\delta} \geq \ln W_t^{\text{UP}}(Y_{1:n}; \nu) \stackrel{(a)}{\geq} \ln W^{\text{pCRP}^*}(Y_{1:n}; \nu) = \sup_{b \in [0,1]} \ln W^{\text{CRP}(b)}(Y_{1:n}; \nu) - \ln \sqrt{\pi(n+1)},$$

which is equivalent to

$$\frac{1}{n} \sup_{b \in [0,1]} \sum_{t=1}^n \ln \left(1 - b + b \frac{Y_t}{\nu} \right) \leq \frac{1}{n} H_n^{(\delta)}.$$

Here, (a) follows from Eq. (3.3).

Now, we apply Lemma B.3 for $n = 1$ and obtain

$$\begin{aligned} \ln(1 - b + bZ) &\geq b((1 - Z)^2 - (1 - Z)) + (1 - Z)^2 \ln(1 - b) \\ &= b(Z^2 - Z) + (Z^2 - 2Z + 1) \ln(1 - b), \end{aligned}$$

which holds for any $b \in [0, 1)$ and $Z > 0$. Applying this inequality to each summand, we have

$$\begin{aligned} \frac{H_n^{(\delta)}}{n} &\geq \frac{1}{n} \sup_{b \in [0,1]} \sum_{t=1}^n \ln \left(1 - b + b \frac{Y_t}{\nu} \right) \\ &\geq \frac{1}{\nu^2} \sup_{b \in [0,1]} \left\{ ((\hat{V}_n + \hat{\mu}_n^2) - \hat{\mu}_n \nu) b + ((\hat{V}_n + \hat{\mu}_n^2) - 2\hat{\mu}_n \nu + \nu^2) \ln(1 - b) \right\} \\ &= \frac{1}{\nu^2} \sup_{b \in [0,1]} \left\{ Bb + (B - A) \ln(1 - b) \right\} \\ &\stackrel{(b)}{\geq} \frac{1}{\nu^2} \sup_{b \in [0,1]} \left\{ Bb + (B - A) \frac{-b}{1 - b} \right\} \\ &\stackrel{(c)}{\geq} \frac{1}{\nu^2} \frac{A^2}{2(2B - A)}, \end{aligned}$$

where $A \triangleq (\hat{\mu}_n - \nu)\nu$ and $B \triangleq (\hat{V}_n + \hat{\mu}_n^2) - \hat{\mu}_n\nu$. Note that (b) follows from the elementary inequality $\ln(1 - b) \geq \frac{-b}{1-b}$ for $b < 1$, and (c) follows by setting $b = \frac{A}{2B}$ to derive a lower bound. We now wish to solve the equation

$$\frac{H_n^{(\delta)}}{n}\nu^2 = \frac{A^2}{2(2B - A)}$$

with respect to ν , which becomes equivalent to

$$\left(1 - 2\frac{H_n^{(\delta)}}{n}\right)x^2 - 2\frac{H_n^{(\delta)}}{n}\hat{\mu}_n x - 4\frac{H_n^{(\delta)}}{n}\hat{V}_n = 0,$$

if we let $x \triangleq \hat{\mu}_n - \nu$. It is easy to check that $x = \hat{\mu}_n - \hat{\mu}_{\text{EB}}^{(\delta)}(Y_{1:n})$ is the solution to this quadratic equation and thus a valid lower bound for μ . \blacksquare

B.5. Proof for Theorem 3.3 (Regret Analysis for PUB)

We provide a proof for $\hat{\pi} = \hat{\pi}_{\text{UP}}$, and the other case follows immediately by the same logic. It suffices to show the second inequality. Letting $2G_n^{(\delta)}[Y_1]$ denote the upper bound in Theorem 3.2, we apply Theorem 3.2 to the process $\tilde{r}_{1:n}^\pi$ for each $\pi \in \Pi$ and take a union bound with $\delta \leftarrow \delta' = \frac{\delta}{|\Pi|}$. Under the good event with probability $\geq 1 - 2\delta$, we have

$$\mu(\pi^*) - \mu(\hat{\pi}_{\text{UP}}) \stackrel{(a)}{\leq} \mu(\pi^*) - \hat{\mu}_{\text{UP}}(\tilde{r}_{1:n}^{\hat{\pi}_{\text{UP}}}) \stackrel{(b)}{\leq} \mu(\pi^*) - \hat{\mu}_{\text{UP}}(\tilde{r}_{1:n}^{\pi^*}) \stackrel{(c)}{\leq} 2G_n^{(\delta')}[\tilde{r}_1^{\pi^*}].$$

Here, (a) follows since $\mu(\tilde{r}_{1:n}^{\hat{\pi}_{\text{UP}}}) \geq \hat{\mu}_{\text{UP}}(\tilde{r}_{1:n}^{\hat{\pi}_{\text{UP}}})$, (b) from the definition of the selection method in Eq. (3.7), and (c) from the upper bound of Theorem 3.2.

The second part of the statement follows from the second part of Theorem 3.2 in place of the first part. \blacksquare

B.6. Off-Policy Evaluation with Betting

We can immediately apply the UP-LCB and pCRP*-LCB for off-policy evaluation as well. Similar to Waudby-Smith et al. (2022), we can construct the upper confidence bound (UCB) of the value of a policy using our LCB machinery, since

$$\check{r}_t^\pi \triangleq w_t^\pi(1 - r_t) = w_t^\pi - \tilde{r}_t^\pi$$

is also a nonnegative random process. Using $\mathbb{E}[\check{r}_t^\pi] = 1 - \mu(\pi)$, we can construct the LCB from $\check{r}_{1:n}^\pi$, from which we can construct the UCB of $\mu(\pi)$. More precisely, we have:

Proposition B.15 *Pick any policy π . With probability $\geq 1 - 2\delta$,*

$$\hat{\mu}_{\text{pCRP}^*}^{(\delta)}(\check{r}_{1:n}^\pi) \leq \hat{\mu}_{\text{UP}}^{(\delta)}(\check{r}_{1:n}^\pi) \leq \mu(\pi) \leq 1 - \hat{\mu}_{\text{UP}}^{(\delta)}(\check{r}_{1:n}^\pi) \leq 1 - \hat{\mu}_{\text{pCRP}^*}^{(\delta)}(\check{r}_{1:n}^\pi).$$

Unlike Sakhi et al. (2024), our guarantee provides a direct control over the width of the confidence bounds. The following guarantee is immediate from Theorem 3.2:

Theorem B.16 (Evaluation) *Pick any policy π . Let $\check{V}(\pi) \triangleq \mathbb{V}[\check{r}_1^\pi]$. With probability $\geq 1 - 4\delta$,*

$$\begin{aligned}
 - \left(\sqrt{\frac{48\check{V}(\pi)}{n}} F_n^{(\delta)} \vee \frac{12(1 - \mu(\pi))}{n} F_n^{(\delta)} \right) &\leq \mu(\pi) - (1 - \hat{\mu}_{\text{pCRP}^*}^{(\delta)}(\check{r}_{1:n}^\pi)) \\
 &\leq \mu(\pi) - (1 - \hat{\mu}_{\text{UP}}^{(\delta)}(\check{r}_{1:n}^\pi)) \\
 &\leq 0 \\
 &\leq \mu(\pi) - \hat{\mu}_{\text{UP}}^{(\delta)}(\check{r}_{1:n}^\pi) \\
 &\leq \mu(\pi) - \hat{\mu}_{\text{pCRP}^*}^{(\delta)}(\check{r}_{1:n}^\pi) \leq \sqrt{\frac{48\check{V}(\pi)}{n}} F_n^{(\delta)} \vee \frac{12\mu(\pi)}{n} F_n^{(\delta)}.
 \end{aligned}$$

Appendix C. Deferred Proofs for Off-Policy Learning

C.1. Proof of Proposition 4.2 (Examples of Score Functions)

Proof Logarithmic smoothing is trivial by definition. For freezing, the upper bound side is obvious. For the lower bound, we need to find c_1 and c_2 such that

$$f(x) \triangleq \frac{\exp(-\phi(x)) - 1 + x}{x^2} \leq \frac{1}{c_1 + c_2 x}$$

For this, if $x \leq 1$ then we have $f(x) = \frac{1}{1+x}$. If $x > 1$, then we have $f(x) = 1/x$. Thus, using $\mathbb{1}\{x > 1\} \leq \frac{x}{1+x}$,

$$\begin{aligned}
 f(x) &\leq \mathbb{1}\{x \leq 1\} \frac{1}{1+x} + \mathbb{1}\{x > 1\} \frac{1}{x} \\
 &\leq \mathbb{1}\{x \leq 1\} \frac{1}{1+x} + \frac{2}{1+x} \\
 &\leq \frac{2}{1+x}.
 \end{aligned}$$

Thus, we have $c_1 = c_2 = \frac{1}{2}$. For clipping, similar to freezing, if $x \leq 1$ then we have $f(x) = \frac{1}{1+x}$. If $x > 1$, then we have $f(x) = \frac{-\frac{1}{2}+x}{x^2} \leq \frac{1}{x}$. We can then proceed the identical derivation to Freezing to obtain $c_1 = c_2 = \frac{1}{2}$. \blacksquare

C.2. Proof of Theorem 4.3 (Regret Analysis for Learning Algorithm)

To derive the desired regret bound for our general estimator $\hat{\pi}_n \triangleq \arg \max_{\pi \in \Pi} \sum_{t=1}^n \phi(\beta \tilde{r}_t^\pi)$, we consider the following two martingales:

$$\begin{aligned}
 \text{(Upper deviation): } U_n^\pi &\triangleq \prod_{t=1}^n \frac{e^{\phi(\beta \tilde{r}_t^\pi)}}{\mathbb{E}[e^{\phi(\beta \tilde{r}_t^\pi)}]}, \\
 \text{(Lower deviation): } L_n^\pi &\triangleq \prod_{t=1}^n \frac{e^{-\phi(\beta \tilde{r}_t^\pi)}}{\mathbb{E}[e^{-\phi(\beta \tilde{r}_t^\pi)}]}.
 \end{aligned}$$

Throughout the proof we omit the subscript t from \tilde{r}_t^π inside the expectation, and use \tilde{r}^π for simplicity.

By applying Ville's inequality (Ville, 1939) and taking the union bound over $\pi \in \Pi$, we have: with probability at least $1 - 2\delta$, $U_n^\pi \leq \frac{1}{\delta}$ and $L_n^\pi \leq \frac{1}{\delta}$ for all $\pi \in \Pi$. Given this good event, we have

$$\begin{aligned} -\ln \left(\mathbb{E}[e^{-\phi(\beta \tilde{r}^{\pi^*})}] \right) - \frac{1}{n} \ln \frac{|\Pi|}{\delta} &\stackrel{(a)}{\leq} \frac{1}{n} \sum_{t=1}^n \phi(\beta \tilde{r}_t^{\pi^*}) \\ &\stackrel{(b)}{\leq} \frac{1}{n} \sum_{t=1}^n \phi(\beta \tilde{r}_t^{\hat{\pi}_n}) \\ &\stackrel{(c)}{\leq} \ln \left(\mathbb{E}[e^{\phi(\beta \tilde{r}^{\hat{\pi}_n})}] \right) + \frac{1}{n} \ln \frac{|\Pi|}{\delta}, \end{aligned} \quad (\text{C.1})$$

where (a) follows from $L_n^{\pi^*} \leq \frac{1}{\delta}$, (b) follows by the definition of $\hat{\pi}_n$, and (c) follows from $U_n^{\hat{\pi}_n} \leq \frac{1}{\delta}$.

We now further upper- and lower-bound this inequality. Note that

$$\ln(\mathbb{E}[e^{\phi(\beta \tilde{r}^{\hat{\pi}_n})}]) = \beta \mathbb{E}[\tilde{r}^{\hat{\pi}_n}] + \ln(\mathbb{E}[e^{\phi(\beta \tilde{r}^{\hat{\pi}_n}) - \mathbb{E} \beta \tilde{r}^{\hat{\pi}_n}}])$$

Thus,

$$\frac{1}{n} \sum_t \frac{1}{\beta} \phi(\tilde{r}_t^{\hat{\pi}_n}) - \mathbb{E}[\tilde{r}^{\hat{\pi}_n}] \leq \underbrace{\frac{1}{\beta} \ln \frac{1}{\mathbb{E}[e^{\phi(\beta \tilde{r}^{\hat{\pi}_n}) - \mathbb{E} \beta \tilde{r}^{\hat{\pi}_n}}]}}_{=-F_\beta(\phi)} + \frac{1}{n\beta} \ln(1/\delta) \quad (\text{C.2})$$

Note that $F_\beta(\phi) \geq 0$ by $\phi(x) \leq \ln(1+x)$.

For the lower bound, we have Note that

$$\begin{aligned} \ln(\mathbb{E}[e^{-\phi(\beta \tilde{r}^{\pi^*})}]) &\leq \mathbb{E}[e^{-\phi(\beta \tilde{r}^{\pi^*})}] - 1 \\ &\leq -\mathbb{E}[\beta \tilde{r}^{\pi^*}] + \mathbb{E}\left[\frac{\beta^2 (\tilde{r}^{\pi^*})^2}{c_1 + c_2 \beta \tilde{r}^{\pi^*}}\right]. \end{aligned}$$

Therefore,

$$\frac{1}{n} \sum_t -\frac{1}{\beta} \phi(\beta \tilde{r}_t^{\pi^*}) + \mathbb{E}[\tilde{r}^{\pi^*}] \leq \beta \mathbb{E}\left[\frac{(\tilde{r}^{\pi^*})^2}{c_1 + c_2 \beta \tilde{r}^{\pi^*}}\right] + \frac{1}{n\beta} \ln(1/\delta). \quad (\text{C.3})$$

By combining Eq. (C.2) and Eq. (C.3) through Eq. (C.1), we have

$$\begin{aligned} v(\pi^*) - v(\hat{\pi}_n) &= \mathbb{E}[\tilde{r}^{\pi^*}] - \mathbb{E}[\tilde{r}^{\hat{\pi}_n}] \\ &\leq \beta \mathbb{E}\left[\frac{(\tilde{r}^{\pi^*})^2}{c_1 + c_2 \beta \tilde{r}^{\pi^*}}\right] + F_\beta(\phi) + \frac{2}{\beta n} \ln \frac{|\Pi|}{\delta}, \end{aligned}$$

which proves the desired claim. ■

Appendix D. On Experiments and Additional Results

D.1. On the Heavy-Tail Setup in Section 5.1

We first provide a formal statement and its proof on the nonexistence of the fourth moment in the setup of Section 5.1.

Proposition D.1 *Consider the discrete context space $\mathcal{X} = \mathbb{N} = \{1, 2, \dots\}$ and a discrete action space $\mathcal{A} = \{1, \dots, K\}$, where the context probability $p(x)$ is assigned such that $p(x = i) \propto \frac{1}{i^2}$. Suppose that $p(r = 1 | x = i, a = 1) \geq \tau > 0$, and a behavior policy is defined as $\pi_{\text{ref}}(a = 1 | x) \triangleq \frac{1}{x^\beta}$ for $\beta \geq 1$. If $\pi(a = 1 | x) \geq c$ for any $x \in \mathcal{X}$ for some $c > 0$, then the fourth moment $\mathbb{E}[(\tilde{r}_t^\pi)^4]$ does not exist.*

Proof Consider

$$\begin{aligned}
 \mathbb{E}[(\tilde{r}_t^\pi)^4] &= \mathbb{E}_{p(x)\pi_{\text{ref}}(a|x)p(r|a,x)} \left[\left(\frac{\pi(A|X)}{\pi_{\text{ref}}(A|X)} \right)^4 R^4 \right] \\
 &= \mathbb{E}_{p(x)} \left[\sum_{a \in \{0,1\}} \frac{\pi(a | X)^4}{\pi_{\text{ref}}(a | X)^3} \mathbb{E}_{p(r|a,X)}[R^4] \right] \\
 &\geq \tau \mathbb{E}_{p(x)} \left[\frac{\pi(a = 1 | X)^4}{\pi_{\text{ref}}(a = 1 | X)^3} \right] \\
 &\gtrsim \sum_{i=1}^{\infty} \frac{1}{i^2} \frac{\pi(a = 1 | x = i)^4}{\pi_{\text{ref}}(a = 1 | x = i)^3} \\
 &\gtrsim \sum_{i=1}^{\infty} \frac{1}{i^2} \frac{1}{(1/i^\beta)^3} \\
 &\gtrsim \sum_{i=1}^{\infty} i^{3\beta-2} = \infty.
 \end{aligned}$$

This concludes the proof. ■

In Figure 9, we present some realizations of the trajectories and LCBs that correspond to the summarization in Figure 4.

D.2. On OP Learning and Selection Datasets

Table 3 summarizes the dimensions of each dataset.

Table 3: Summary statistics of the datasets.

Dataset	PenDigits	SatImage	JPVowel
# features	16	36	14
# classes	10	6	9

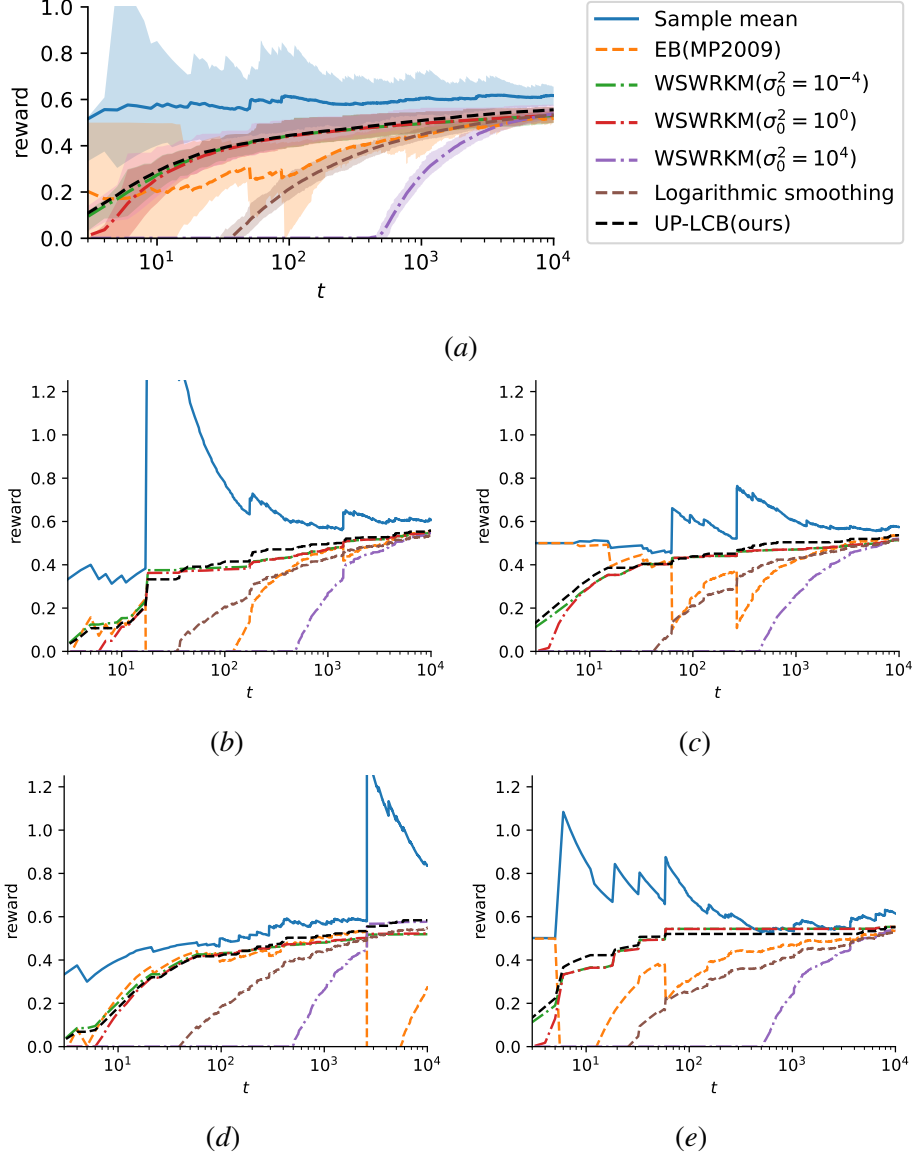


Figure 9: Comparison of the UP-based LCB with baseline LCBs. The average behavior over $N = 100$ random trials is presented in (a), and (b)-(d) show some realizations of the random runs. These instances clearly demonstrate the failure cases of empirical-Bernstein-type bounds, which rely on the concentration of the empirical variance.

D.3. OP Learning Baselines

The estimators we tested in the OP learning experiment are defined as follows:

$$\begin{aligned}
 \hat{\pi}_{\text{PL}} &\triangleq \arg \max_{\pi \in \Pi} \sum_{t=1}^n \left(\tilde{r}_t^\pi - \beta \sum_{a \in \mathcal{A}} \frac{\pi(a|x_t)}{\pi_{\text{ref}}(a|x_t)} \right), \\
 \hat{\pi}_{\text{ClippedIW}} &\triangleq \arg \max_{\pi \in \Pi} \sum_{t=1}^n \frac{\pi(a_t|x_t)}{\pi_{\text{ref}}(a_t|x_t) \vee \beta} r_t, \\
 \hat{\pi}_{\text{IX}} &\triangleq \arg \max_{\pi \in \Pi} \sum_{t=1}^n \frac{\pi(a_t|x_t)}{\pi_{\text{ref}}(a_t|x_t) + \beta} r_t, \\
 \hat{\pi}_{\text{LS}} &\triangleq \arg \max_{\pi \in \Pi} \sum_{t=1}^n \ln(1 + \beta \tilde{r}_t^\pi), \\
 \hat{\pi}_{\text{LS+freezing}} &\triangleq \arg \max_{\pi \in \Pi} \sum_{t=1}^n \ln(1 + \beta \tilde{r}_t^\pi) \mathbb{1}\{x \leq \beta \tilde{r}_t^\pi\}.
 \end{aligned}$$

In each case, $\beta > 0$ is an hyperparameter.

D.4. Additional Experiments for OP Learning and Selection

In this section, we report the OP learning and selection results with two more different policies $\pi_{\text{good}, \varepsilon=0.01}$ and $\pi_{\text{bad}, \varepsilon=0.1}$ for completeness. Note that the experimental results in the main text were with the policy $\pi_{\text{good}, \varepsilon=0.1}$.

Figure 10 and Table 4 summarize the OP learning and selection results, respectively. For $\pi_{\text{good}, \varepsilon=0.01}$, the behavior of the estimators aligns with that under $\pi_{\text{good}, \varepsilon=0.1}$ in Figure 5 and Table 1. Specifically, LS+freezing consistently improves upon LS in learning, and PUB achieves the best or near-best performance in selection.

In contrast, for $\pi_{\text{bad}, \varepsilon=0.01}$, OP learning results show mixed behavior, likely due to the poor quality of the behavior policy. In OP selection, all methods perform poorly in the small-sample regime, failing to improve upon the IW baseline. We particularly note that, while EB, LS, and PUB show comparable performance, WSWRKM performs substantially worse in this erratic setting.

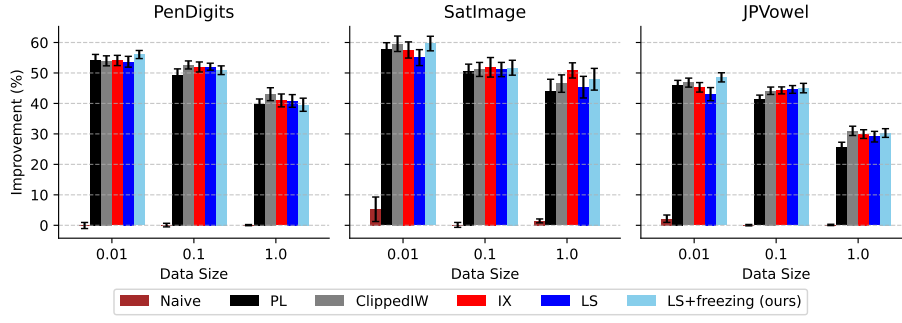
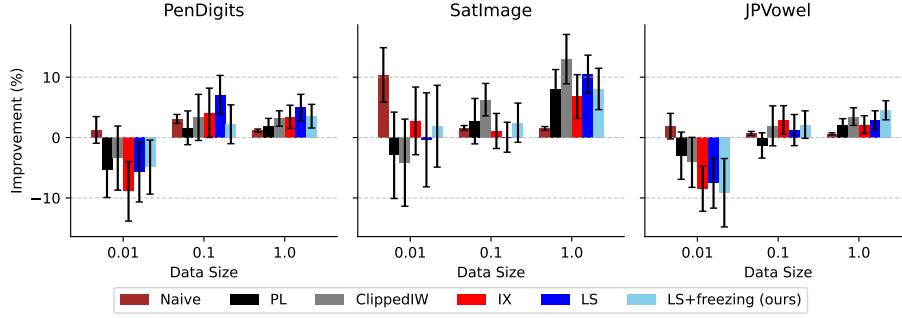
(a) Good policy + uniform policy with probability $\varepsilon = 0.01$ of choosing the good policy.(b) Bad policy + uniform policy with probability $\varepsilon = 0.1$ of choosing the bad policy.

Figure 10: Additional OP learning results. Compare to Figure 5 in the main text.

Table 4: Additional OP selection results. Compare to Table 1 in the main text.

(a) Good policy + uniform policy with probability $\varepsilon = 0.01$ of choosing the good policy.

Dataset	PenDigits			SatImage			JPVowel		
Size	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1
EB	<u>50.90</u>	<u>47.32</u>	<u>40.24</u>	52.34	<u>33.73</u>	<u>33.98</u>	<u>41.33</u>	48.41	<u>32.09</u>
LS	25.09	46.72	<u>39.30</u>	29.40	<u>23.88</u>	<u>35.94</u>	18.38	<u>46.88</u>	<u>32.15</u>
WSWRKM	<u>35.22</u>	<u>50.00</u>	18.18	44.01	35.09	39.00	45.76	<u>37.62</u>	<u>31.17</u>
PUB (ours)	51.02	51.35	41.33	<u>50.60</u>	<u>32.05</u>	<u>37.10</u>	<u>35.26</u>	<u>46.88</u>	32.89

(b) Bad policy + uniform policy with probability $\varepsilon = 0.1$ of choosing the bad policy.

Dataset	PenDigits			SatImage			JPVowel		
Size	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1
EB	-8.86	3.27	1.22	-8.36	0.93	2.68	-7.89	9.23	-0.79
LS	-8.86	3.27	1.22	-8.36	0.93	2.68	-7.89	<u>7.65</u>	-0.79
WSWRKM	-15.79	-0.56	-1.50	-46.89	-4.59	1.41	-37.90	-8.19	-39.76
PUB (ours)	-8.86	3.27	1.22	-8.36	0.93	2.68	<u>-8.66</u>	<u>7.65</u>	-0.79