TO EXPLORE OR TO EXPLOIT: CONSERVATIVE OPTIMISM WITH PESSIMISTIC BASELINE FOR OFFLINE-TO-ONLINE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline-to-online learning begins with data collected before online interaction and then continues learning from newly acquired online feedback, raising a fundamental question: when should the learner trust the decision supported by offline data, and when is it worth paying the cost of exploration to improve upon it? The answer depends critically on the online horizon T: committing to a pessimistic lower-confidence-bound (LCB) decision avoids costly early exploration but incurs large regret for long horizon T if the decision is suboptimal, whereas optimistic upper-confidence-bound (UCB) exploration enables long-run improvement at the price of potentially large regret in the early stage. Yet choosing between these strategies in advance is generally impossible, since the relevant crossover depends on unknown instance gaps and the horizon may itself be unknown. We show that a simple cumulative-budget rule suffices to obtain direct baseline protection under the standard pseudo-regret criterion while preserving continued online learning. The resulting algorithm, Conservative Optimism with Pessimistic Baseline (COPB), maintains a cumulative performance certificate relative to the offline LCB decision: it follows the hybrid UCB recommendation when the certificate permits exploration, and otherwise plays the offline LCB baseline. For stochastic bandits, COPB simultaneously guarantees regret close to that of committing to the offline LCB decision, up to a prescribed tolerance, and hybrid UCB-type gap-dependent and gap-independent regret bounds with explicitly characterized conservative overhead. The bounds further capture how offline coverage reduces both online exploration cost and the cost of baseline protection. COPB is horizon-adaptive: its exploration-versus-exploitation behavior requires no horizon-dependent tuning and automatically shifts toward online improvement as evidence accumulates. We further extend the framework to episodic tabular reinforcement learning and validate it on synthetic and real-data experiments.

1 INTRODUCTION

Online learning is a central framework for sequential decision making under uncertainty, with applications in a wide range of domains such as online advertising, recommendation systems, financial decision making, and online model optimization. A canonical model for studying such problems is the multi-armed bandit model (Auer et al., 2002; Berry & Fristedt, 1985; Lattimore & Szepesvári, 2020; Chen et al., 2016; Liu et al., 2023; Thompson, 1933; Agrawal & Goyal, 2012; Russo et al., 2018; Bubeck & Cesa-Bianchi, 2012; Kleinberg & Slivkins, 2010; Kong & Li, 2023). In the classical online setting, the learner repeatedly interacts with an environment and improves its decisions from the feedback collected over time, with the goal of minimizing regret.

In many real-world sequential learning problems, however, online learning does not start from scratch: historical data are often already available before online interaction begins. For example, in sim-to-real robotics, a policy may first be trained using simulator data and then adapted after deployment using real-world interactions (Peng et al., 2018; Chebotar et al., 2019). Similarly, in large language model training, models are first optimized using existing datasets through pre-training and post-training procedures, and may later be further adapted using deployment-time user feedback (Brown et al., 2020; Ouyang et al., 2022; Han et al., 2025). These examples naturally motivate the **offline-to-online learning** setting, in which the learner starts with an offline dataset and subsequently continues to learn through online interaction. Recent studies have shown that such offline data can substantially reduce the burden of online exploration and improve learning efficiency (Cheung & Lyu, 2024; Zheng et al., 2023; Lee et al., 2022; Song et al., 2023; Li et al., 2023; Bu et al., 2022; Sentenac et al., 2025; Shivaswamy & Joachims, 2012; Zhou et al., 2026a;b; Li et al., 2026).

The presence of offline data gives the learner a meaningful alternative to immediate online exploration: when online interaction begins, it may already have a decision whose performance is well supported by historical evidence and can

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

therefore be exploited from the outset. This naturally raises a fundamental question: *when should the learner exploit the decision supported by offline data, and when is it worth paying the cost of exploration to improve beyond it?*

The answer to this question depends critically on the online horizon. In online bandit learning, a standard principle is *optimism in the face of uncertainty* (Auer et al., 2002; Bubeck & Cesa-Bianchi, 2012; Lattimore & Szepesvári, 2020), typically implemented by an upper confidence bound (UCB) rule that assigns uncertain actions optimistic values and thereby encourages exploration. By contrast, offline learning often adopts *pessimism in the face of uncertainty* (Liu et al., 2025; Swaminathan & Joachims, 2015; Li et al., 2022; Xiao et al., 2021), for example through a lower confidence bound (LCB) rule that favors actions whose performance can be reliably supported by the available data. These two principles lead to complementary behaviors in the offline-to-online setting. At a high level, committing to an offline LCB decision avoids spending early online rounds exploring poorly supported actions and can therefore be advantageous when only limited online interaction is available. However, if the offline decision is suboptimal, repeatedly exploiting it incurs regret that grows linearly with time. Conversely, UCB-style exploration may incur substantial transient regret at the beginning, but over a sufficiently long horizon it can discover better actions and achieve sublinear regret. Thus, neither pessimistic commitment nor optimistic exploration uniformly dominates across online horizons.

The discussion above suggests a natural static strategy: compare the regret guarantees of the LCB and UCB learners and switch between them at some horizon T_0. However, coarse gap-independent bounds are poorly suited to locating such a crossover: the sublinear-versus-linear dependence on T mainly reflects their long-run scaling and can obscure the finite-horizon exploration cost that motivates our question. More refined instance-dependent bounds could in principle define an oracle switching point, but such a threshold depends on unknown quantities such as the reward gaps and the quality and allocation of the offline data. Moreover, the eventual online horizon may itself be unknown. These limitations motivate a dynamic alternative: rather than fixing a switching time in advance, the learner should decide online whether further exploration is justified as information accumulates.

This dynamic perspective is closely related to (Sentenac et al., 2025), which recognizes the same horizon-dependent tension between pessimistic offline decisions and optimistic online learning. However, Sentenac et al. (2025) formalize the two sides using different comparison criteria: logging-policy regret, measured relative to the offline logging policy, for pessimistic protection and standard regret against optimality for optimistic learning. We ask whether both sides can instead be characterized directly under the standard online pseudo-regret criterion. (We defer a detailed comparison of results to Section 4.)

We show that a cumulative-budget mechanism can dynamically balance exploitation of the offline LCB decision with continued UCB exploration. Drawing inspiration from the cumulative-budget principle of Conservative UCB (Wu et al., 2016), we develop Conservative Optimism with Pessimistic Baseline (COPB) to achieve this balance in the offline-to-online setting. Specifically, COPB first selects a baseline action using the offline LCB rule. During online interaction, a hybrid UCB learner, initialized with the offline data and updated with online observations, proposes an optimistic action at each round. COPB then evaluates whether taking this action would keep its cumulative certified performance above a prescribed fraction of repeatedly playing the offline baseline. If so, the UCB action is accepted; otherwise, the algorithm plays the LCB baseline for that round. Thus, rather than committing to a fixed switching horizon, COPB makes the exploration decision online according to the accumulated performance budget, while maintaining direct protection relative to the offline baseline.

For stochastic K-armed bandits, COPB combines baseline protection with UCB-type learning under the standard pseudo-regret criterion. With high probability, its regret is at most $\alpha\mu_0 T$ above that of the offline LCB baseline, while achieving a simplified bound of $\widetilde{O}(\sqrt{KT} + K/(\alpha\mu_0))$. Here, $\mu_0 > 0$ is the mean reward of the baseline, $\alpha \in (0, 1)$ controls conservativeness, and $\widetilde{O}$ hides logarithmic factors. Table 1 summarizes sharper bounds capturing how offline data reduce both exploration and conservative costs; Section 4 provides the full analysis.

Our contributions are summarized as follows.

- We study the horizon-dependent tension between exploiting an pessimistic decision and continuing optimistic exploration under the standard pseudo-regret criterion, and show that a simple cumulative-budget mechanism is sufficient to make this choice dynamically, without requiring a pre-specified switching horizon or knowledge of the reward gaps.

- For stochastic multi-armed bandits, we establish regret guarantees showing that COPB combines the complementary strengths of LCB and UCB. The algorithm stays within an $\alpha\mu_0 T$ slack of committing to the offline LCB baseline, while retaining both gap-dependent and gap-independent UCB-type learning rates. Our analysis further provides a refined characterization of the additional cost of baseline protection, explicitly tracking its dependence on baseline quality, the user-controlled conservativeness parameter α, and the arm-wise allo-

Algorithm	Short Horizon R_1		Long Horizon R_T	
	Gap-dependent	Gap-independent	Gap-dependent	Gap-independent
UCB	$\min_i \left\{ \Delta_i + \max_{j \neq i} \frac{J}{\sqrt{N_j}} \right\}$	$\frac{J}{\sqrt{N_{\min}}}$	$\sum_{i:\Delta_i>0} \left(\frac{J^2}{\Delta_i} - N_i \Delta_i \right)_+$	$\sqrt{\frac{KT^2J^2}{T+N_{\min}}}^{\dagger}$
LCB	$\Delta_0 \leq \min_i \left\{ \Delta_i + \frac{J}{\sqrt{N_i}} \right\}$	$\Delta_0 \leq \frac{J}{\sqrt{N_{\min}}}$	$T\Delta_0$	$T\Delta_0$
COPB (ours)	$\Delta_0 + \alpha\mu_0$	$\Delta_0 + \alpha\mu_0$	$\sum_{i:\Delta_i>0} \left(\frac{J^2}{\Delta_i} - N_i \Delta_i \right)_+ + \frac{KJ^2}{\alpha\mu_0}$	$\sqrt{\frac{KT^2J^2}{T+N_{\min}}}^{\dagger} + \frac{KJ^2}{\alpha\mu_0}$

Table 1: Regret comparison of UCB, LCB, and COPB.[‡]

[†] The gap-independent UCB-type bounds in offline-to-online learning can be written in several equivalent forms (Cheung & Lyu, 2024; Sentenac et al., 2025; Zhou et al., 2026a;b). We display the form that most clearly shows the dependence on the online horizon T and is most convenient for comparison.

[‡] For readability, most bounds in this table are simplified and omit universal numerical constants and lower-order terms.

cation of offline samples. Together, these bounds quantify how offline data can reduce both the usual online exploration cost and the additional conservative overhead required to protect the baseline.

- We further extend the same conservative–optimistic principle to episodic tabular reinforcement learning, establish analogous regret guarantees, and validate the method through synthetic and real-data experiments.
- As a by-product, we extend Conservative UCB (Wu et al., 2016) to incorporate offline data. Our analysis further characterizes how the allocation of offline samples across arms affects the conservative cost.

Organization. We present the setup in Section 2, COPB in Section 3, and its regret analysis in Section 4, followed by experiments (Section 5). Section 6 concludes, while related work, the RL extension and proofs appear in Section 7.

2 Problem Setup

2.1 Notation

We use $(x)_+ = \max\{x, 0\}$. For a positive integer d, we write $[d] = \{1, \ldots, d\}$.

2.2 Offline-to-Online Bandits

We consider the stochastic K-armed bandits in the offline-to-online setting. Each arm $i \in [K]$ has an unknown mean reward $\mu_i \in [0, 1]$. Rewards from arm i are independent and 1-sub-Gaussian around their mean, that is, for any reward observation r from arm i and any $\lambda \in \mathbb{R}$, $\mathbb{E}[\exp(\lambda(r - \mu_i))] \leq \exp\left(\frac{\lambda^2}{2}\right)$. Let $i^\star \in \arg\max_{i \in [K]} \mu_i$ be an optimal arm and let $\mu^\star = \mu_{i^\star}$. The suboptimality gap of arm i is $\Delta_i = \mu^\star - \mu_i$.

Before online interaction starts, the learner receives an offline dataset $\mathcal{D}_0 = \left\{ \left(N_i, \{y_{i,s}\}_{s=1}^{N_i} \right) \right\}_{i=1}^{K}$, where N_i is the number of offline samples available for arm i, and $y_{i,s}$ is the s-th offline reward observation from arm i. We assume that the offline samples of arm i are drawn from the same reward distribution as the online rewards of arm i. Let $N_{\min} = \min_{i \in [K]} N_i$, $N = \sum_{i=1}^{K} N_i$. After observing $\mathcal{D}_0$, the learner enters the online phase. At each round $t = 1, \ldots, T$, the learner selects an arm $a_t \in [K]$ based on the offline data and the online observations collected so far. After pulling arm a_t, the learner observes a reward r_t with mean μ_{a_t}. We evaluate the learner by its online pseudo-regret $R_T = \sum_{t=1}^{T} (\mu^\star - \mu_{a_t})$, and by its expectation $\mathbb{E}[R_T]$ when an expected-regret guarantee is considered. The goal is to design an algorithm with small online pseudo-regret. In this paper, our main results are high-probability bounds on R_T. Only online rounds contribute to R_T: the offline samples are used to initialize or guide the learner, but are not counted as part of the online performance.

We next introduce useful hybrid statistics that combine offline and online samples. Let $T_i(t) = \sum_{\tau=1}^{t-1} \mathbf{1}\{a_\tau = i\}$ be the number of online pulls of arm i before round t, and define the total number of available samples for arm i before

round t as $\bar{N}_i(t) = N_i + T_i(t)$. The empirical mean based on both offline samples and online observations before round t is

$$\widehat{\mu}_i(t) = \frac{\sum_{s=1}^{N_i} y_{i,s} + \sum_{\tau=1}^{t-1} r_\tau \mathbf{1}\{a_\tau = i\}}{\bar{N}_i(t)}.$$

For high-probability guarantees, we use the shorthand $\ell_{\mathrm{mab}} = \log(KT/\delta)$ and $J_{\mathrm{mab}} = c\sqrt{\ell_{\mathrm{mab}}}$, where $c > 0$ is a sufficiently large universal constant. The analogous confidence factor for the RL setting will be denoted by J_{rl} in Section 7.2; when no confusion arises, we write J for either quantity.

3 Conservative Optimism with Pessimistic Baseline

The goal of COPB is to resolve the horizon-dependent trade-off between the offline LCB baseline and the hybrid UCB learner. As discussed in the introduction, LCB is preferable when the online horizon is short because it relies on a reliable offline decision, whereas UCB is necessary when the horizon is long because it continues to explore and learn. The difficulty is that the learner does not know in advance which regime is relevant, since the effective switching point depends on the unknown gaps and the quality of the offline data. COPB addresses this difficulty by treating the offline LCB arm as a conservative reference point and using a cumulative budget to decide, during online interaction, when optimistic UCB-style exploration is justified.

Algorithm 1 Conservative Optimism with Pessimistic Baselines

Require: K, ϵ_{off}, ϵ_{on}, α, offline data $\{N_i, \{y_{i,s}\}_{s=1}^{N_i}\}_{i=1}^{K}$

Compute offline estimates:
for $i = 1, \ldots, K$ **do**
  $\widehat{\mu}_i(0) \leftarrow \frac{1}{N_i} \sum_{s=1}^{N_i} y_{i,s}$
  $\mathrm{LCB}_i \leftarrow \widehat{\mu}_i(0) - \sqrt{\frac{\log(1/\epsilon_{\mathrm{off}})}{N_i}}$
end for
$i_0 \in \arg\max_i \mathrm{LCB}_i$ ▷ offline safe arm
Initialize $\mathcal{U}_0 \leftarrow \emptyset$, $\mathcal{B}_0 \leftarrow \emptyset$
Here $\mathcal{U}_t$ and $\mathcal{B}_t$ record accepted-UCB rounds and fallback rounds, respectively.
for $t = 1, 2, \ldots$ **do**
  Compute confidence intervals
  for $i = 1, \ldots, K$ **do**
    $rad_i(t) \leftarrow \sqrt{\frac{\log(1/\epsilon_{\mathrm{on}})}{N_i + T_i(t-1)}}$
    $U_i(t) \leftarrow \widehat{\mu}_i(t-1) + rad_i(t)$ ▷ UCB
    $L_i(t) \leftarrow \max\{0, \widehat{\mu}_i(t-1) - rad_i(t)\}$ ▷ LCB
  end for
  $A_t \leftarrow \arg\max_i U_i(t)$ ▷ optimistic arm
  Compute budget
  $\xi_t \leftarrow \sum_{s \in \mathcal{U}_{t-1}} L_{I_s}(t) + |\mathcal{B}_{t-1}|\, U_{i_0}(t) + L_{A_t}(t) - (1-\alpha) t\, U_{i_0}(t)$
  if $\xi_t \geq 0$ **then**
    $I_t \leftarrow A_t$ ▷ accept UCB recommendation
    $\mathcal{U}_t \leftarrow \mathcal{U}_{t-1} \cup \{t\}$, $\mathcal{B}_t \leftarrow \mathcal{B}_{t-1}$
  else
    $I_t \leftarrow i_0$ ▷ fallback to offline safe arm
    $\mathcal{U}_t \leftarrow \mathcal{U}_{t-1}$, $\mathcal{B}_t \leftarrow \mathcal{B}_{t-1} \cup \{t\}$
  end if
  Pull arm I_t, observe reward y_t, and update $T_{I_t}(t)$ and $\widehat{\mu}_{I_t}(t)$.
end for

The algorithm implements the above idea through a cumulative-budget viewpoint, which is natural in pricing and marketing scenarios where accumulated reliable rewards can be used to support later exploration. We first use the offline LCB rule at time $T = 0$ to select a baseline arm i_0. During the online phase, the algorithm maintains a budget ξ_t, which should be interpreted as the accumulated certified surplus of the algorithm relative to the LCB baseline. This budget is not a myopic comparison between the arm proposed by UCB at round t and the baseline arm. Rather, it

aggregates the certified rewards collected in all previous rounds, and therefore measures how much room the algorithm has earned for optimistic exploration.

At each round, the algorithm first computes the arm A_t recommended by the hybrid UCB rule. It then asks whether playing A_t would keep the cumulative budget nonnegative. If so, the algorithm follows UCB and plays A_t. This may increase the budget when A_t is well certified, or decrease the budget when A_t is uncertain or appears suboptimal; nevertheless, such a decrease is allowed as long as the accumulated surplus remains nonnegative. Only when the candidate UCB action would drive the budget below zero does the algorithm fall back to the offline LCB baseline arm i_0. In this sense, past certified gains finance future exploration, while the nonnegativity of the budget prevents the algorithm from falling behind the LCB benchmark. At a high level, this cumulative-budget design is inspired by the idea of online conservative bandits (Wu et al., 2016).

There is one further consideration. In some instances, the LCB baseline arm may already be very strong, and the arm proposed by UCB may be worse than the baseline, but not much worse. To allow such controlled exploration, we introduce a tolerance parameter α. This parameter formalizes the requirement that the UCB action should not perform too much worse than the LCB baseline. As a result, even when the algorithm plays the baseline arm, the budget increases by an α-fraction of the baseline reward. This leaves room for future exploration while still ensuring that, at any time, the cumulative performance of the algorithm remains comparable to that of the LCB baseline, up to the prescribed tolerance.

Of course, the learner does not know the true rewards of either the UCB arm or the baseline arm. We therefore implement the above budget condition using confidence bounds. Since the goal is to ensure that the algorithm does not fall behind the LCB baseline, we estimate the reward of the baseline arm conservatively using an upper confidence bound, and estimate the rewards of all other played arms using lower confidence bounds.

Let $\mathcal{U}_{t-1}$ and $\mathcal{B}_{t-1}$ denote the sets of rounds before t in which the gate accepted the UCB recommendation and triggered fallback to the LCB baseline, respectively. Then $|\mathcal{B}_{t-1}| = T_0(t-1)$, and the budget can be written as

$$\xi_t = \underbrace{\sum_{s \in \mathcal{U}_{t-1}} L_{I_s}(t) + |\mathcal{B}_{t-1}| U_{i_0}(t) + L_{A_t}(t)}_{\text{certified credit}} - \underbrace{(1-\alpha) t U_{i_0}(t)}_{\text{baseline requirement}}.$$

The certified credit consists of three parts: the lower-confidence rewards accumulated in previous rounds where the algorithm followed the UCB recommendation, the certified contribution from previous fallback rounds where the algorithm played the LCB baseline, and the lower-confidence reward that would be obtained if the current UCB candidate A_t were played. The last term is the tolerated baseline requirement: up to the tolerance parameter α, the algorithm must certify performance comparable to repeatedly playing the baseline arm i_0. Therefore, the condition $\xi_t \geq 0$ checks whether the algorithm has enough accumulated credit to afford playing the current optimistic action. Consequently, on the good confidence event, whenever the budget is nonnegative, the cumulative reward of our algorithm is certified to be no worse than that of the LCB baseline, up to the tolerance parameter.

In the next section, our regret analysis further shows that the algorithm also matches the performance of UCB up to confidence terms. Thus, the proposed method enjoys an anytime best-of-both-worlds guarantee: it retains the early-stage robustness of LCB while preserving the long-horizon advantage of UCB.

Remark 1 (Hybrid UCB versus offline LCB)**.** *The UCB and LCB components in COPB serve different purposes. The UCB component is a hybrid UCB rule, which uses the offline data as a warm start and keeps updating its estimates with online feedback. The LCB component, however, is used only to select a fixed offline baseline arm i_0 before online interaction starts. This fixed baseline provides a clean reference point for the cumulative budget constraint.*

A natural alternative is to use a hybrid LCB rule as the baseline. We do not adopt this choice because it brings little conceptual benefit for our purpose. Although hybrid LCB may become more accurate as online data accumulate, it can still remain conservative and suffer linear regret in the long run. Thus, using hybrid LCB would not fundamentally improve the pessimistic baseline, but would make the comparison and analysis less transparent. For this reason, we use the offline LCB arm as a stable conservative benchmark throughout the online phase.

Remark 2 (Unknown horizon)**.** *We use fixed-horizon confidence parameters for simplicity. Unknown horizons can be handled by a doubling schedule with summable failure probabilities, retaining the baseline and cumulative budget.*

4 Regret Analysis

We now present the regret guarantees for COPB. These results show that the proposed algorithm interpolates between pessimistic commitment to the offline LCB baseline and optimistic online exploration: it behaves similarly to the LCB

baseline in the short-horizon regime, while retaining the long-run learning and convergence properties of UCB. We focus on the main statements in this section and defer the proofs to the appendix. Specifically, Section 7.5 proves the LCB-type guarantee of Theorem 1, Section 7.6 proves the gap-dependent guarantee of Theorem 2, and Section 7.8 proves the gap-independent guarantee of Theorem 3.

Let i_0 be the arm selected by the offline LCB rule based on the offline data, and write $\mu_0 = \mu_{i_0}$ and $\Delta_0 = \mu^\star - \mu_0$. Specifically, i_0 is selected by the offline LCB rule implemented in Lines 1–6 of Algorithm 1. For clarity, we also provide the standalone pseudocode of the offline LCB rule in Appendix 7.3, Algorithm 5.

Theorem 1 (LCB baseline guarantee). *Let i_0 be the arm selected by the offline LCB rule based on the offline data, and write $\mu_0 = \mu_{i_0}$ and $\Delta_0 = \mu^\star - \mu_0$. Then, for every horizon $T \geq 1$ and an input $\alpha \in (0,1)$, with probability at least $1-\delta$, let $\epsilon_{off} = (\delta/(4K))^2$, $\epsilon_{on} = (\delta/(4K(T+1)))^2$, then the pseudo-regret of Algorithm 1 satisfies*

$$\boxed{R_T \leq (\Delta_0 + \alpha\mu_0)T.}$$

Theorem 2 (Gap-dependent regret of COPB). *For any input $\alpha \in (0,1)$, with probability at least $1-\delta$, let $\epsilon_{off} = (\delta/(4K))^2$, $\epsilon_{on} = (\delta/(4K(T+1)))^2$, then COPB satisfies*

$$\boxed{R_T \lesssim \sum_{i:\Delta_i>0} \left(\frac{J^2}{\Delta_i} - N_i\Delta_i\right)_+ + \frac{\Delta_0}{\alpha\mu_0}\left(B^\star + \sum_{i:\Delta_i>0} B_i\right) + \Delta_0.}$$

Here

$$J = c\sqrt{\log\frac{4K(T+1)}{\delta}}, \qquad \gamma = \Delta_0 + \alpha\mu_0, \qquad B^\star = \gamma\left(\frac{J}{\gamma} - \sqrt{N_{i^\star}}\right)_+^2,$$

and, for each suboptimal arm i with $\Delta_i > 0$,

$$B_i = \begin{cases} (\gamma - \Delta_i)\left(\dfrac{J}{\gamma-\Delta_i} - \sqrt{N_i}\right)_+^2 + \left(\dfrac{J^2}{\Delta_i} - N_i\Delta_i\right)_+, & 0 < \Delta_i < \frac{\gamma}{2}, \\ (\Delta_i - \gamma)\left(\dfrac{J^2}{\Delta_i^2} - N_i\right)_+ + 2J\left(\dfrac{J}{\Delta_i} - \sqrt{N_i}\right)_+ + \left(\dfrac{J^2}{\Delta_i} - N_i\Delta_i\right)_+, & \Delta_i \geq \frac{\gamma}{2}. \end{cases}$$

Theorem 3 (Gap-independent regret of COPB). *Let λ_{mab} satisfy $\sum_{i=1}^K (\lambda_{\mathrm{mab}} - N_i)_+ = T$, let $J = c\sqrt{\log\frac{4K(T+1)}{\delta}}$. Then, for any input $\alpha \in (0,1)$, with probability at least $1-\delta$, let $\epsilon_{off} = (\delta/(4K))^2$, $\epsilon_{on} = (\delta/(4K(T+1)))^2$, then COPB satisfies*

$$\boxed{R_T \lesssim J\sum_{i=1}^K \left(\sqrt{\lambda_{\mathrm{mab}}} - \sqrt{N_i}\right)_+ + \frac{\Delta_0}{\alpha\mu_0}(\Delta_0 + \alpha\mu_0)\sum_{i=1}^K \left(\frac{J}{\Delta_0+\alpha\mu_0} - \sqrt{N_i}\right)_+^2 + \Delta_0.}$$

Overall intuition. The first guarantee characterizes the protection provided by the LCB baseline. Theorem 1 shows that the regret of COPB is never much worse than the regret incurred by committing to the offline LCB arm, namely $\Delta_0 T$. The additional term $\alpha\mu_0 T$ reflects the exploration slack deliberately introduced by our algorithm. For fixed α and μ_0, this term is the price paid for preserving some online exploration capability rather than fully committing to the offline baseline. Thus, when the online horizon T is small, COPB behaves essentially like the LCB policy, which is consistent with the intended conservative behavior in the early stage.

Theorems 2 and 3 then characterize the long-horizon learning ability of COPB from gap-dependent and gap-independent perspectives, respectively. These results show that, despite the conservative LCB gate, COPB still enjoys UCB-type convergence behavior when T becomes large. In the gap-dependent bound, the leading term matches the standard hybrid UCB rate up to lower-order conservative terms. The two additional terms arise from the early conservative design of COPB. For fixed α and μ_0, these terms are of order close to, or smaller than, $\log(T)\Delta_0/\Delta_i$, and hence are dominated by the usual hybrid UCB term of order $\log(T)/\Delta_i$. Therefore, the conservative mechanism does not alter the leading regret order.

Role of α. The parameter $\alpha \in (0,1)$ provides a user-controlled trade-off between the conservative baseline guarantee and the UCB-type learning guarantees. The three bounds in Theorems 1–3 hold simultaneously, but they are informative in different regimes. If one cares more about the UCB-type gap-dependent or gap-independent guarantees, for example when the offline data are limited, the offline baseline is unreliable, or the horizon is large, then a larger α is preferable. Indeed, the additional conservative terms in Theorems 2 and 3 scale as $1/\alpha$, so increasing α reduces the price paid for maintaining the conservative constraint. On the other hand, if one cares more about short-horizon performance relative to the offline baseline, for example when the online budget is small or the offline data are reliable, then a smaller α is preferable. In this regime, Theorem 1 becomes tighter, since the slack relative to the offline baseline is of order $\alpha\mu_0 T$. Thus, α can be tuned depending on whether one wants COPB to behave more conservatively like the offline LCB baseline or more exploratorily like UCB.

Benefit of offline data. The regret bounds also make explicit how offline data reduce the cost of online exploration. We first look at the leading UCB-type terms. In the gap-dependent bound, the main term is $\sum_{i:\Delta_i>0}\left(\frac{J^2}{\Delta_i} - N_i\Delta_i\right)_+$. Compared with the pure online gap-dependent term $\sum_{i:\Delta_i>0} J^2/\Delta_i$, the offline samples of arm i reduce the regret contribution by $N_i\Delta_i$, until the term is clipped at zero. Thus, offline data are especially useful on suboptimal arms with large gaps: these are precisely the arms whose online exploration would incur large regret, and offline samples reduce the number of such costly online trials. The same phenomenon appears in the gap-independent bound. Its leading term also reflects a distribution-free saving effect through the quantities $\left(\sqrt{\lambda_{\mathrm{mab}}} - \sqrt{N_i}\right)_+$, summed over all arms. Here λ_{mab} is the water-filling level defined by $\sum_{i=1}^K(\lambda_{\mathrm{mab}} - N_i)_+ = T$. For arm i, the term $(\sqrt{\lambda_{\mathrm{mab}}} - \sqrt{N_i})_+$ shows that offline samples effectively skip the initial online pulls from count 1 up to count N_i. Since the regret of pulling an uncertain arm is controlled by a confidence radius of order $J/\sqrt{s}$ when the arm has been sampled s times, the cumulative saving from these skipped pulls is of order $\sum_{s=1}^{N_i}\frac{J}{\sqrt{s}} \asymp J\sqrt{N_i}$. Therefore, arms with sufficient offline coverage contribute little, or even nothing, to the remaining online exploration cost.

Beyond these leading UCB-type terms, offline data also affect the additional conservative terms introduced by COPB. In the gap-dependent bound, the quantities $B^\star$ and B_i contain positive-part terms involving $\sqrt{N_{i^\star}}$ and $\sqrt{N_i}$, and therefore decrease as the relevant arms receive more offline samples. In the gap-independent bound, the additional conservative term has the same structure through $\sum_{i=1}^K\left(\frac{J}{\Delta_0+\alpha\mu_0} - \sqrt{N_i}\right)_+^2$. Thus, the conservative overhead is not a fixed penalty. It becomes smaller when the offline data provide better coverage or when more offline samples are available. In this sense, offline data benefit both the optimistic UCB component and the conservative baseline-protection component of COPB.

Comparison with Sentenac et al. (2025). We next compare our guarantees with the offline-to-online method in Sentenac et al. (2025). Both works aim to combine pessimistic offline decisions with optimistic online exploration. The main distinction is how the comparison with the pessimistic LCB baseline is formalized. While Sentenac et al. (2025) also provides standard regret guarantees, its "no worse than LCB" comparison is established through a logging-regret criterion rather than through the standard pseudo-regret against the optimal arm. This distinction is important in the offline-to-online setting considered here. The horizon-dependent trade-off between LCB and UCB arises precisely because different algorithms can suffer very different regret during the online phase, especially when the online horizon T is small or large. Therefore, the most direct performance measure for this trade-off is the online pseudo-regret incurred by the deployed algorithm.

Under this criterion, a guarantee stated in terms of logging regret does not, by itself, directly control the standard online pseudo-regret. Define the logging-policy value as $\bar{\mu} := \frac{1}{N}\sum_{i=1}^K N_i\mu_i$ and the logging regret as $R_T^{\log} := \sum_{t=1}^T(\bar{\mu} - \mu_{a_t}) = R_T - T(\mu^\star - \bar{\mu})$, where $R_T = \sum_{t=1}^T(\mu^\star - \mu_{a_t})$ is the standard pseudo-regret. Thus, logging regret and pseudo-regret differ by the offset $T(\mu^\star - \bar{\mu})$. Since $\mu^\star - \bar{\mu}$ can be a constant, this offset can be of order T, and a small logging regret does not necessarily imply a small pseudo-regret. In contrast, our comparison with the LCB baseline is made directly under the standard pseudo-regret criterion. Moreover, our pseudo-regret guarantees can be translated into corresponding logging-regret guarantees, while the converse implication from its to ours does not hold; a detailed discussion is deferred to Appendix 7.9.

Another difference lies in how the conservative cost depends on the amount of offline data. From the Theorem 1 of Sentenac et al. (2025), the relevant parameter β of standard regret upper bound of OTO is of order $1/\sqrt{N}$, where N denotes the total amount of offline data. Consequently, the additional conservative term scaling as $1/\beta$ is of order $\sqrt{N}$. Thus, this part of the regret upper bound can increase as more offline data become available. Such a dependence is not desirable in offline-to-online learning: ideally, adding more offline data should help the online learner, or at least should not make the regret guarantee worse. This phenomenon is also observed in our experiments in Section 5,

where the performance of OTO may deteriorate as the amount of offline data increases. By contrast, our additional conservative terms do not change the leading UCB-type regret order and are tied to the quality of the offline baseline and the offline sample allocation. When the offline data provide more useful support for online learning, these terms become smaller.

Relation to Wu et al. (2016). Conservative UCB (ConUCB) (Wu et al., 2016) studies online exploration under a performance constraint relative to a prescribed default arm. We build on its cumulative-budget principle to address the offline-to-online trade-off between exploiting a pessimistic decision learned from historical data and exploring to improve upon it. COPB constructs the baseline through the offline LCB rule and incorporates offline observations into both optimistic action selection and performance certification. Our theoretical contribution further lies in characterizing how offline data affect the cost of maintaining this protection. Specifically, Theorems 2 and 3 explicitly track the allocation of offline samples across arms in both the hybrid UCB regret terms and the additional conservative terms. These bounds show how offline coverage can reduce both the cost of exploration and the overhead required to protect the offline baseline, a dependence absent from the purely online guarantees of ConUCB.

5 Experiments

We empirically evaluate COPB on multi-armed bandits tasks. The goal is to verify whether COPB can adaptively match the better side between pessimistic offline learning and optimistic online learning, without being substantially worse than either baseline. The main text focuses on the primary synthetic experiments, whose setup is described as follows. Due to space limitations, additional results are deferred to the appendix, including sensitivity analyses over the conservative hyperparameter α in Appendix 7.10.1 and the offline-data allocation parameter ρ in Appendix 7.10.2, as well as experiments on real-world datasets in Appendix 7.10.3. The code is available at `https://anonymous.4open.science/r/COPB-B598/`.

We compare COPB with three baselines: **LCB**, the purely pessimistic policy learned from offline data (Algorithm 5); **UCB**, a standard hybrid optimistic algorithm initialized with offline data (Algorithm 6); and **OTO**, the offline-to-online baseline from prior work (Algorithm 1 in Sentenac et al. (2025)). We first follow the experimental setting in Sentenac et al. (2025) and evaluate algorithms on a synthetic stochastic multi-armed bandit instance with $K = 20$ arms and offline data size $N = 2000$ or 3000. Rewards are Bernoulli. The offline dataset is generated by sampling each of the first 10 arms exactly 200 or 300 times, while arms 11 to 20 are not sampled in the offline phase. The mean rewards are set to $\mu_i = 0.5$ for $1 \le i \le 10$, $\mu_{12} = 0.75$, and $\mu_i = 0.25$ otherwise. As a result, the offline data provide accurate estimates only for a subset of non-optimal arms, while the best arm must be identified through online interaction. This creates a setting where the learner needs to balance reliance on offline evidence with sufficient exploration beyond the logging support. For the conservative hyperparameter α in our COPB method, we set $\alpha = 0.5$. For the OTO baseline, we set $\alpha = 0.2$ following the setting used in the original paper. All algorithms are evaluated under the same protocol, and results are averaged over 50 independent runs.

5.1 Main Results

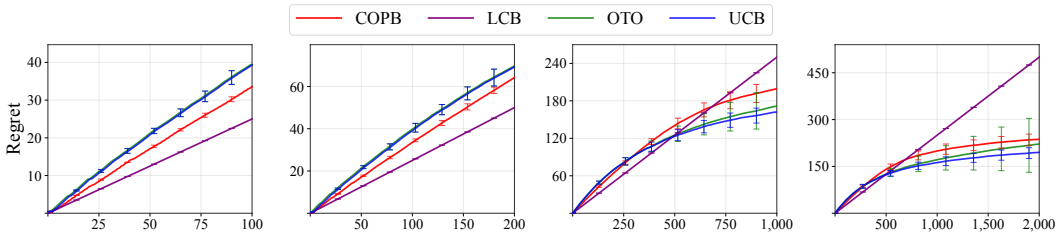

Figure 1: Main comparison with $N = 2000$.

Figures 1 and 2 report the cumulative regret over time under the same online horizon $T = 2000$, with offline data sizes $N = 2000$ and $N = 3000$, respectively. In each figure, the four panels are different zoomed-in views of the same cumulative-regret curve, rather than four independent experimental settings; they are shown to highlight the behavior at short, medium, and long horizons. Across both offline data sizes, LCB performs well in the early stage because it relies on the offline-supported pessimistic baseline, but its regret keeps increasing linearly due to insufficient online exploration. In contrast, UCB benefits from optimism and achieves good long-run performance, but it pays a larger price at short horizons because it explores aggressively from the beginning.

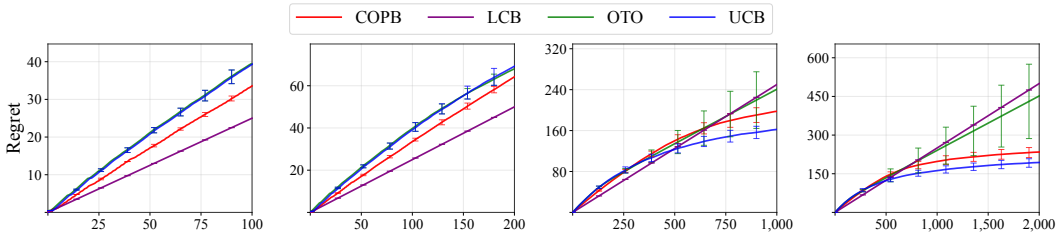

Figure 2: Main comparison with $N = 3000$.

COPB exhibits the desired horizon-adaptive behavior between these two baselines. At short horizons, it stays close to LCB by remaining conservative when online exploration is not yet sufficiently worthwhile. As more online samples are collected, it gradually allows optimistic exploration and becomes competitive with UCB at longer horizons. This behavior is consistent with our theoretical motivation: COPB is designed to preserve the short-horizon advantage of the pessimistic offline baseline while retaining the long-horizon benefit of optimistic online learning.

OTO is designed with a similar offline-to-online goal, but it does not consistently realize this balance in experiments. When $N = 2000$, OTO behaves too similarly to UCB in the early stage and therefore suffers from a relatively large short-horizon regret. This is consistent with the theoretical discussion in Section 4 that a guarantee in terms of logging regret does not directly imply an LCB-like guarantee under the standard pseudo-regret criterion. When $N = 3000$, OTO becomes closer to the offline-driven behavior, but it then fails to clearly exploit online exploration at longer horizons and exhibits larger variability. In other words, increasing the amount of offline data does not necessarily improve OTO. This observation is consistent with the theoretical analysis in their paper, where the UCB-type regret bound scales with $\sqrt{N}$ and therefore becomes larger as N increases. We further verify this phenomenon in the next Experiment, where we test different choices of N.

5.2 Effect of N

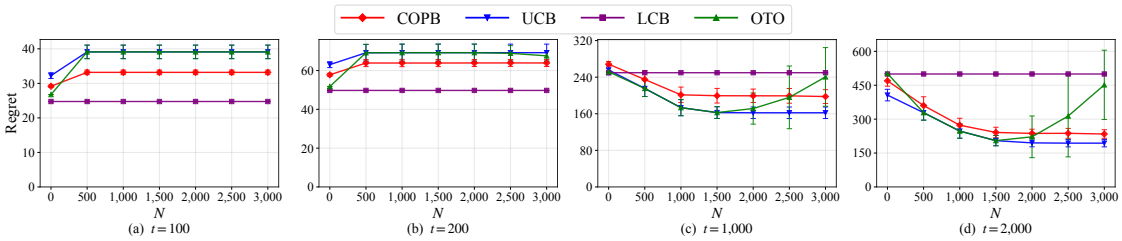

Figure 3: Sensitivity analysis with respect to the offline data size N.

We study the effect of the offline sample size N. As shown in Figure 3, COPB and OTO both interpolate between UCB and LCB. However, when N becomes large, OTO shows less stable behavior: it is closer to UCB at short horizons but shifts toward LCB at long horizons, with larger variance. COPB maintains a more stable conservative-optimistic balance, benefiting from improved offline estimates while still enabling safe online exploration.

6 Conclusion

We studied offline-to-online learning through the trade-off between pessimistic offline decisions and optimistic online exploration. We proposed COPB, an anytime algorithm that uses an offline LCB policy as a conservative baseline and explores only when UCB-style decisions can be justified. COPB achieves best-of-both-worlds regret guarantees: it remains competitive with the offline baseline for short horizons while recovering gap-dependent and worst-case UCB guarantees for long horizons. Our analysis also quantifies how offline data reduces exploration costs, and we extend the framework to reinforcement learning and validate it empirically.

A natural direction for future work is to replace the fixed conservativeness parameter α with an adaptive α_t that evolves with online evidence, although establishing guarantees for such schemes requires a finer time-dependent analysis.

AI USE STATEMENT

In this work, we have not used generative AI tools for tasks with required disclosure. Additionally, we used generative AI tools for English language polishing, including improving grammar, clarity, and readability, as well as assisting in generating code to visualize experimental results. We have reviewed all AI-assisted work. Specifically, the authors checked the revised text to ensure that it preserves the intended meaning and accurately conveys the technical content. AI-assisted visualization code was manually reviewed and executed, and the resulting figures were checked against the underlying experimental results for correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on the theoretical and algorithmic aspects of offline-to-online learning. We do not identify any specific ethical concerns arising from the methods and results presented in this paper.

REPRODUCIBILITY STATEMENT

We have taken several steps to support the reproducibility of our results. The complete implementation of the proposed methods and experimental pipeline is provided in the accompanying anonymous code repository, including the scripts and configurations used to reproduce the experiments reported in the paper. Experimental settings, datasets, baselines, and implementation details are described in the main text and appendix. For the theoretical results, all assumptions are stated explicitly and complete proofs of the main claims are provided in the appendix.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pp. 39.1–39.26. PMLR, 2012.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002. doi: 10.1023/A:1013689704352.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 263–272. PMLR, 2017.

Aharon Ben-Tal and Arkadi Nemirovski. Robust optimization—methodology and applications. *Mathematical Programming*, 92(3):453–480, 2002. doi: 10.1007/s101070100286.

Donald A. Berry and Bert Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, 1985.

Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011. doi: 10.1137/080734510.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Jinzhi Bu, David Simchi-Levi, and Yunzong Xu. Online pricing with offline data: Phase transition and inverse square law. *Management Science*, 68(12):8568–8588, 2022.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012. doi: 10.1561/2200000024.

Yevgen Chebotar, Ankur Handa, Viktor Makoviychuk, Miles Macklin, Jan Issac, Nathan D. Ratliff, and Dieter Fox. Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8973–8979. IEEE, 2019. doi: 10.1109/ICRA.2019.8793789. URL https://arxiv.org/abs/1810.05687.

Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50):1–33, 2016. URL https://jmlr.org/papers/v17/14-298.html.

Wang Chi Cheung and Lixing Lyu. Leveraging (biased) information: Multi-armed bandits with offline data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8286–8309. PMLR, 2024. URL https://proceedings.mlr.press/v235/cheung24a.html.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 208–214. PMLR, 2011.

Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pp. 191–198. ACM, 2016. doi: 10.1145/2959100.2959190. URL https://doi.org/10.1145/2959100.2959190.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2052–2062. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/fujimoto19a.html.

Eric Han, Jun Chen, Karthik Abinav Sankararaman, Xiaoliang Peng, Tengyu Xu, Eryk Helenowski, Kaiyan Peng, Mrinal Kumar, Sinong Wang, Han Fang, and Arya Talebzadeh. Reinforcement learning from user feedback. *arXiv preprint arXiv:2505.14946*, 2025. doi: 10.48550/arXiv.2505.14946. URL https://arxiv.org/abs/2505.14946.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010. URL https://www.jmlr.org/papers/v11/jaksch10a.html.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Àgata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind W. Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019. URL https://arxiv.org/abs/1907.00456.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1704–1713. PMLR, 2017.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143. PMLR, 2020.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5084–5096. PMLR, 2021. URL https://proceedings.mlr.press/v139/jin21e.html.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21810–21823, 2020.

Robert Kleinberg and Aleksandrs Slivkins. Sharp dichotomies for regret minimization in metric spaces. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 827–846. SIAM, 2010. doi: 10.1137/1.9781611973075.68. URL https://doi.org/10.1137/1.9781611973075.68.

Fang Kong and Shuai Li. Player-optimal stable regret for bandit learning in matching markets. In *Proceedings of the 2023 ACM-SIAM Symposium on Discrete Algorithms*, pp. 1512–1522. SIAM, 2023. doi: 10.1137/1.9781611977554.CH55. URL https://doi.org/10.1137/1.9781611977554.CH55.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In Aleksandra Faust, David Hsu, and Gerhard Neumann (eds.), *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 1702–1712. PMLR, 08–11 Nov 2022. URL https://proceedings.mlr.press/v164/lee22d.html.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020. URL https://arxiv.org/abs/2005.01643.

Bochao Li, Yao Fu, Wei Chen, and Fang Kong. Sample-mean anchored thompson sampling for offline-to-online learning with distribution shift, 2026. URL https://arxiv.org/abs/2605.10289.

Gen Li, Wenhao Zhan, Jason D. Lee, Yuejie Chi, and Yuxin Chen. Reward-agnostic fine-tuning: Provable statistical benefits of hybrid reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ade04fd4f26263f86b47ffb535c4cafb-Abstract-Conference.html.

Gene Li, Cong Ma, and Nathan Srebro. Pessimism for offline linear contextual bandits using ℓ_p confidence sets. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/8443219a991f068c34d9491ad68ffa94-Abstract-Conference.html.

Xutong Liu, Jinhang Zuo, Siwei Wang, John C. S. Lui, Mohammad Hajiesmaili, Adam Wierman, and Wei Chen. Contextual combinatorial bandits with probabilistically triggered arms. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22559–22593. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/liu23bf.html.

Xutong Liu, Xiangxiang Dai, Jinhang Zuo, Siwei Wang, Carlee Joe-Wong, John C. S. Lui, and Wei Chen. Offline learning for combinatorial multi-armed bandits. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 38251–38289. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/liu25i.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3803–3810. IEEE, 2018. doi: 10.1109/ICRA.2018.8460528. URL https://arxiv.org/abs/1710.06537.

Chengrui Qu, Laixi Shi, Kishan Panaganti, Pengcheng You, and Adam Wierman. Hybrid transfer reinforcement learning: Provable sample efficiency from shifted-dynamics data. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1054–1062. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/qu25a.html.

Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018. doi: 10.1561/2200000070.

Flore Sentenac, Ilbin Lee, and Csaba Szepesvári. Balancing optimism and pessimism in offline-to-online learning. *arXiv preprint arXiv:2502.08259*, 2025. URL https://arxiv.org/abs/2502.08259.

Pannagadatta Shivaswamy and Thorsten Joachims. Multi-armed bandit problems with history. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pp. 1046–1054. PMLR, 2012.

Yuda Song, Yifei Zhou, Ayush Sekhari, J. Andrew Bagnell, Akshay Krishnamurthy, and Wen Sun. Hybrid RL: Using both offline and online data can make RL efficient. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=yyBis80iUuU.

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16(52):1731–1755, 2015. URL http://jmlr.org/papers/v16/swaminathan15a.html.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3–4):285–294, 1933. doi: 10.1093/biomet/25.3-4.285.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1254–1262, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/wu16.html.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019. URL https://arxiv.org/abs/1911.11361.

Chenjun Xiao, Yifan Wu, Tor Lattimore, Bo Dai, Jincheng Mei, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. On the optimality of batch policy optimization algorithms. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11362–11371. PMLR, 2021. URL https://proceedings.mlr.press/v139/xiao21b.html.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27395–27407, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/e61eaa38aed621dd776d0e67cfeee366-Abstract.html.

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4065–4078, 2021.

Han Zheng, Xufang Luo, Pengfei Wei, Xuan Song, Dongsheng Li, and Jing Jiang. Adaptive policy learning for offline-to-online reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11372–11380, 2023. doi: 10.1609/aaai.v37i9.26345.

Kongchang Zhou, Zheshun Wu, Wei Chen, and Fang Kong. Regret analysis of hybrid linear bandits with biased offline data, 2026a. URL https://openreview.net/forum?id=NA4seLpkPd.

Kongchang Zhou, Tingyu Zhang, Wei Chen, and Fang Kong. Hybrid combinatorial multi-armed bandits with probabilistically triggered arms. *Transactions on Machine Learning Research*, 2026b. ISSN 2835-8856. URL https://openreview.net/forum?id=ZRzB7EVlui.

Yifei Zhou, Ayush Sekhari, Yuda Song, and Wen Sun. Offline data enhanced on-policy policy gradient with provable guarantees. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/9949e6906be6448230cdba9a4cb2d564-Abstract-Conference.html.