Asymptotically Optimal Problem-Dependent Bandit Policies for Transfer Learning

Adrien PREVOST ADRIEN.PREVOST@INRIA.FR
Timothee MATHIEU TIMOTHEE.MATHIEU@INRIA.FR
Odalric-Ambrym MAILLARD ODALRIC.MAILLARD@INRIA.FR
Equipe Scool, Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189- CRIStAL, F-59000 Lille, France

Editors: Hung-yi Lee and Tongliang Liu

Abstract

We study the non-contextual multi-armed bandit problem in a transfer learning setting: before any pulls, the learner is given N'_k i.i.d. samples from each source distribution ν'_k, and the true target distributions ν_k lie within a known distance bound $d_k(\nu_k, \nu'_k) \leq L_k$. In this framework, we first derive a problem-dependent asymptotic lower bound on cumulative regret that extends the classical Lai–Robbins result to incorporate the transfer parameters (d_k, L_k, N'_k). We then propose KL-UCB-Transfer, a simple index policy that matches this new bound in the Gaussian case. Finally, we validate our approach via simulations, showing that KL-UCB-Transfer significantly outperforms the no-prior baseline when source and target distributions are sufficiently close.

Keywords: Multi-armed bandit; Transfer Learning; Problem-dependent optimality; Regret analysis

1. Introduction

In many sequential-decision tasks, one often has access to *historical data* collected under related but not identical conditions. Leveraging such data to speed up learning in a new environment is the goal of *transfer learning*, a paradigm that has been successfully applied in computer vision, reinforcement learning, and natural language processing. For example, pre-trained image classifiers built on large, generic datasets can be adapted to medical imaging; in robotics, policies learned in simulation help bootstrap real-world control; and in agriculture, past yield records from neighboring fields guide crop-management decisions in a new plot (Taylor and Stone, 2009).

In this work we study transfer in the simplest possible online setting: a *non-contextual multi-armed bandit* with K arms and i.i.d. rewards. Before interacting with the target bandit, the learner receives N'_k i.i.d. samples from each *source* arm k, whose true reward distribution we denote by ν'_k. The goal is to minimize cumulative regret on the *target* bandit, whose arm k has distribution ν_k. We assume the learner knows a family of (pseudo-)distance d_k and upper bounds $L_k > 0$ such that each ν_k lies within distance L_k of ν'_k:

$$\nu_k \in \left\{\gamma \in \mathcal{D} : d_k(\gamma, \nu'_k) \leq L_k\right\}, \quad k = 1, \ldots, K.$$

Intuitively, small L_k makes source samples informative; large N'_k reduces noise. This departs from prior transfer-bandit work assuming contextual, causal, or multi-task structure Rahul and Katewa (2024b); Wang et al. (2022). Instead, we study a minimal non-contextual

© 2025 A. PREVOST, T. MATHIEU & O.-A. MAILLARD.

setting with i.i.d. samples per arm and transfer via known distance bounds (d_k, L_k). This captures scenarios (e.g., warm-start from historical data or sim-to-real) while enabling sharp analysis. This *non-contextual transfer bandit* setting raises two key questions:

1. **Characterize the transfer-aware minimal regret.** Extend the classical Lai–Robbins lower bound on the regret Lai and Robbins (1985) so it explicitly depends on the transfer parameters (d_k, L_k, N'_k), yielding the minimal asymptotic cost any algorithm must pay.

2. **Design a matching policy.** Propose a simple index strategy that incorporates the N'_k prior samples and achieves the transfer-aware lower bound.

Our contributions

- We derive a *problem-dependent* lower bound on cumulative regret which holds for any algorithm with prior data $\{N'_k, \hat{\nu}'_k, d_k, L_k\}$ (see Section 4, Theorem 3).

- Focusing on Gaussian rewards with known variance, we introduce in Section 5.2 KL-UCB-Transfer, a minor modification of the KL-UCB index that adds a KL-penalty term for the source data. We prove that the regret of KL-UCB-Transfer *matches* our lower bound (see Section 6, Theorem 6). In the regret analysis, one must carefully handle potentially large or T-dependent N'_k: the index should leverage prior samples to avoid redundant exploration while preserving sufficient exploration when necessary.

- We validate via simulations that KL-UCB-Transfer can achieve substantial regret reductions when the priors on suboptimal arms are accurate, and we highlight the delicate short-term trade-offs when priors are placed on the optimal arm in Section 7.

By keeping our setting minimal—no contexts, fixed variances—we obtain a clean, intuitive theory and a useful algorithm for studying transfer learning.

2. Related Work

Classical non-contextual bandits. The stochastic multi-armed bandit was first studied by Thompson (1933) and rigorously analyzed by Robbins (1952). Lai and Robbins (1985) proved the fundamental problem-dependent asymptotic lower bound on regret, and a range of algorithms—such as KL-UCB Maillard et al. (2011); Cappé et al. (2013), IMED Honda and Takemura (2015), DMED Honda and Takemura (2010), and Thompson Sampling Kaufmann et al. (2012)—are known to match this bound.

Contextual transfer and causal approaches. Transfer in contextual bandits has been extensively studied: Deng et al. (2025) leverages causal transportability to handle covariate shift in latent contextual settings; Cai et al. (2024) derives minimax optimal rates under covariate shift; Bellot et al. (2023) proposes a transport-aware Thompson Sampling policy using graphical causal models; Zhang and Bareinboim (2017) analyzes a causal MAB with latent confounders and provides a problem-dependent lower bound under specific structural assumptions. However, these approaches rely on contextual or causal-structure assumptions and do not address the non-contextual offline–online transfer setting where one first observes i.i.d. source samples per arm, nor do they derive a problem-dependent lower bound, and consequently lack any algorithm matching such a bound in that regime.

Non-contextual transfer bandits. Several recent works consider transfer in the classical (non-contextual) bandit setting, but under different task structures Gheshlaghi azar et al. (2013); Saber et al. (2021). Wang et al. (2022) study a *multi-task* regime where multiple bandits run in parallel and Thompson Sampling borrows posterior mass across tasks for robustness, yielding minimax-style guarantees. Rahul and Katewa (2024b,a) analyze a *sequential-task* framework: they assume known "adjacent similarity" bounds to transfer raw reward samples via UCB, and in their follow-up work they learn the similarity radius online, each time achieving improved instance-dependent regret over no-transfer baselines.

By contrast, we consider a single target bandit and first receive N'_k i.i.d. samples from each source arm. We give the first *problem-dependent asymptotic lower bound* in this *offline, non-contextual* transfer setting (see Theorem 3) and design a KL-UCB–style index that *matches* it exactly in the Gaussian case (see Theorem 6).

While our setting may appear overly simplified, deriving a problem-dependent lower bound that explicitly accounts for arbitrary prior sample sizes N'_k is non-trivial and not covered by existing approaches, which often merge prior and online data without regard to lower-bound tightness. Also, despite the generic formulation, this setting has received little direct attention in the literature.

3. Problem formulation

In this section, we first recall the classical stochastic multi-armed bandit framework, then introduce our offline non-contextual transfer setting with prior samples from a related source.

Classical setting We consider a known family of distributions $\mathcal{D}$ and unknown arm distributions $\nu = (\nu_1, \ldots, \nu_K) \in \mathcal{D}^K$ with means μ_k. At each round t, a learner selects arm a_t and observes reward $X_t \sim \nu_{a_t}$. Assume w.l.o.g. that arm 1 has the highest mean. Let $\mathbf{1}\{\cdot\}$ denote the indicator. the cumulative regret is

$$R_T = \sum_{k=1}^{K} \mathbb{E}[N_k(T)]\,(\mu_1 - \mu_k), \quad \text{where} \quad N_k(T) = \sum_{t=1}^{T} \mathbf{1}\{a_t = k\}.$$

A strategy is *consistent* if for each suboptimal arm k, $\mathbb{E}[N_k(T)] = o(T^\alpha)$ for all $\alpha > 0$. For any consistent algorithm and suboptimal arm k, the following asymptotic lower bound holds Lattimore and Szepesvári (2020):

$$\liminf_{T\to\infty} \inf_{\substack{\tilde{\nu}_k \in \mathcal{D} \\ \mathbb{E}[\tilde{\nu}_k] > \mu_1}} \frac{\mathbb{E}[N_k(T)]\ \mathrm{KL}(\nu_k, \tilde{\nu}_k)}{\ln T} \ \geq\ 1, \tag{1}$$

results the following lower bound on the regret

$$R_T \geq \ln T \left(\sum_{k=2}^{K} (\mu_1 - \mu_k) \sup_{\substack{\tilde{\nu}_k \in \mathcal{D} \\ \mathbb{E}[\tilde{\nu}_k] > \mu_1}} \frac{1}{\mathrm{KL}(\nu_k, \tilde{\nu}_k)} \right) + o(\ln T)$$

Transfer learning setting Before interacting with the target ν, we observe N'_k i.i.d. samples from related source distributions $\nu' = (\nu'_1, \dots, \nu'_K) \in \mathcal{D}'^K$ satisfying $d_k(\nu_k, \nu'_k) \le L_k$. The regret on ν remains R_T as defined above. Our goal is to derive a transfer-aware lower bound analogous to (Eq(1)), explicitly involving (d_k, L_k, N'_k), and to design a KL-UCB–type algorithm that matches this bound.

4. Lower Bounds

In this section, we derive fundamental limits on cumulative regret in our offline transfer setting. Specifically, we extend the classical Lai–Robbins bound (Eq(1)) to incorporate the transfer parameters (d_k, L_k, N'_k), showing that no consistent algorithm can beat this rate (see Theorem 3), and we discuss several important special cases.

Remark 1 *In the rest of the paper, all $o(\cdot)$ and $O(\cdot)$ notation refer to the limit $T \to \infty$. Since offline sample sizes N'_k may depend arbitrarily on T (or even be $+\infty$), any $o(\cdot)$ or $O(\cdot)$ usage must explicitly account for possible dependence on N'_k.*

4.1. General Lower Bound in the Transfer Setting

Definition 2 *An algorithm is said to be* consistent *(w.r.t. $(\mathcal{D}, \mathcal{D}', d, L, N')$) if for every pair $(\nu, \nu') \in \mathcal{D}^K \times \mathcal{D}'^K$ and every suboptimal arm k,*

$$\forall\, \alpha > 0, \quad \mathbb{E}[N_k(T)] = o(T^\alpha) \quad (T \to \infty).$$

Theorem 3 *Let a consistent algorithm operate under $(\mathcal{D}, \mathcal{D}', d, L, N')$. For any suboptimal arm k,*

$$\liminf_{T\to\infty} \inf_{\substack{\tilde\nu_k \in \mathcal{D},\ \tilde\nu'_k \in \mathcal{D}' \\ \mathbb{E}[\tilde\nu_k] > \mu_1,\ d_k(\tilde\nu_k, \tilde\nu'_k) \le L_k}} \frac{\mathbb{E}[N_k(T)]\ \mathrm{KL}(\nu_k, \tilde\nu_k) + N'_k\ \mathrm{KL}(\nu'_k, \tilde\nu'_k)}{\ln T} \ge 1.$$

Equivalently, defining $\mathcal{K}_{\inf}(\nu'_k; d_k, \tilde\nu_k, L_k) := \inf\{\mathrm{KL}(\nu'_k, \tilde\nu'_k) \mid \ \tilde\nu'_k \in \mathcal{D}',\ d_k(\tilde\nu_k, \tilde\nu'_k) \le L_k\}$,

$$\liminf_{T\to\infty} \inf_{\substack{\tilde\nu_k \in \mathcal{D} \\ \mathbb{E}[\tilde\nu_k] > \mu_1}} \frac{\mathbb{E}[N_k(T)]\ \mathrm{KL}(\nu_k, \tilde\nu_k) + N'_k\, \mathcal{K}_{\inf}(\nu'_k; d_k, \tilde\nu_k, L_k)}{\ln T} \ge 1.$$

Offline sample size N'_k: We allow N'_k to be any non-negative integer (possibly depending on T) or even $+\infty$. If $N'_k = 0$ or $N'_k = o(\ln T)$, the offline term is negligible and the equation (1) is recovered; if $N'_k = \infty$, then either $\mathcal{K}_{\inf}(\nu'_k; d_k, \tilde\nu_k, L_k) = 0$ (reducing to equation (1)) or consistency imposes no constraint on $\mathbb{E}[N_k(T)]$, allowing $\mathbb{E}[N_k(T)] = o(\ln T)$.

Prior radius L_k: If $L_k \to \infty$, then for any $\tilde\nu_k$ the constraint $d_k(\tilde\nu_k, \tilde\nu'_k) \le L_k$ becomes vacuous, so $\mathcal{K}_{\inf}(\nu'_k; d_k, \tilde\nu_k, L_k) \to 0$ and the classical bound is recovered. If $L_k = 0$ and $d_k(\nu, \nu') = 0 \implies \nu = \nu'$, then $\nu'_k = \nu_k$ and

$$\mathbb{E}[N_k(T)]\ \mathrm{KL}(\nu_k, \tilde\nu_k) + N'_k\ \mathrm{KL}(\nu'_k, \tilde\nu_k) = (\mathbb{E}[N_k(T)] + N'_k)\ \mathrm{KL}(\nu_k, \tilde\nu_k),$$

yielding the classical bound (1) with effective sample size $\mathbb{E}[N_k(T)] + N'_k$.

General implications: Since $N'_k \mathcal{K}_{\inf}(\nu'_k; d_k, \tilde{\nu}_k, L_k) \geq 0$, Theorem 3 is tighter than the classical bound; prior data on the optimal arm do not affect this lower bound, as the optimal arm is pulled linearly and its mean is estimated arbitrarily well online.
The proof of Theorem 3 is deferred to Appendix A.

4.2. Gaussian case with mean-based distance

We now instantiate Theorem 3 when both target and source families are univariate Gaussians with known variances. We use a mean-based distance:

$$\mathcal{D} = \{\mathcal{N}(\mu, \sigma^2)\}, \quad \mathcal{D}' = \{\mathcal{N}(\mu, \sigma'^2)\}, \quad d_k(\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\tilde{\mu}, \sigma'^2)) = |\mu - \tilde{\mu}|.$$

Hence the constraint $|\tilde{\mu}_k - \tilde{\mu}'_k| \leq L_k$. We will also note $(x)_+ := \max\{x, 0\}$.

Corollary 4 (Gaussian lower bound) *Under the Gaussian setting with known variances and mean-based distance defined above, any consistent algorithm satisfies for each suboptimal arm k:*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_k(T)]}{\ln T} \frac{(\mu_1 - \mu_k)^2}{2\,\sigma^2} + \frac{N'_k}{\ln T} \frac{(\mu_1 - \mu'_k - L_k)_+^2}{2\,\sigma'^2} \geq 1. \tag{2}$$

Proof [Sketch of proof] Apply Theorem 3 with

$$\mathrm{KL}(\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\tilde{\mu}, \sigma^2)) = \frac{(\mu - \tilde{\mu})^2}{2\,\sigma^2},$$

and similarly for source-to-source KL. Constraining $\mathbb{E}[\tilde{\nu}_k] > \mu_1$ yields the first term in the bound, while enforcing $|\tilde{\mu}_k - \tilde{\mu}'_k| \leq L_k$ produces the $(\cdot)_+$ in the second term. ■

Equivalently, solving for the expected number of pulls yields:

$$\mathbb{E}[N_k(T)] \geq \frac{2\,\sigma^2}{(\mu_1 - \mu_k)^2} \Big(\ln T - N'_k \tfrac{(\mu_1 - \mu'_k - L_k)_+^2}{2\,\sigma'^2} \Big)_+ + o(\ln T). \tag{3}$$

5. KL-UCB Algorithm

In this section we first recall the classical KL-UCB algorithm, we then introduce our KL-UCB-Transfer variant, which augments the usual index with an extra penalty term to account for prior offline samples and will match Equation 3.

5.1. Classical KL-UCB

For the classical bandit problem, the KL-UCB algorithm introduced in Maillard et al. (2011) matches the asymptotic lower bound (Eq 1). The core idea of this algorithm is to select the arm with the highest potential mean. To do so, we define an index $U_a(t)$ for each arm a, and pull the arm with the highest index:

$$U_a(t) := \max\{\mathbb{E}[\nu] \mid N_a(t)\ \mathrm{KL}(\pi_{\mathcal{D}}(\hat{\nu}_a(t)), \nu) \leq \delta_t,\ \nu \in \mathcal{D}\}$$

where $\pi_{\mathcal{D}}(\hat{\nu}_k(t))$ is the projection of the empirical distribution $\hat{\nu}_k(t)$ onto the space of admissible distributions $\mathcal{D}$. This algorithm is optimal for bounded reward distributions (see Garivier and Cappé (2011); Cappé et al. (2013)). We typically choose δ_t to be of the order of $\ln t$, possibly with an additional lower-order term depending on $\mathcal{D}$.

5.2. KL-UCB-Transfer

In our transfer learning setting, assuming Gaussian distributions with known variance as in Section 4.2, we define the cost function with respect to the prior data as follows:

$$f_a^+(q) := N'_a \frac{(q - (\hat{\mu}'_a(t) + L_a))_+^2}{2\sigma'^2}$$

In this setting, the index becomes:

$$U_a(t) := \max\left\{ q \;\middle|\; N_a(t)\frac{(q - \hat{\mu}_a(t))_+^2}{2\sigma^2} + f_a^+(q) \le \delta_t \right\}, \tag{4}$$

with $\delta_t = \ln t + 3\ln(\max(1, \ln t))$, we compute $U_a(t)$ each turn and pull the arm with highest index; the positive part ensures a non-empty set. An explicit form of $U_a(t)$ is given in the supplement, enabling $O(1)$ computation. The idea behind the algorithm is to incorporate the cost of deviating from the prior information into the estimation of the highest potential mean. Thus, we recover the lower bound given in section 4.2.

Remark 5 *If there is no prior data,* KL-UCB-Transfer *reverts to classical* KL-UCB*.*

Design challenges and points of attention The key challenge in KL-UCB-Transfer is to balance prior and online data without merging them: unlike Rahul and Katewa (2024b), which fuses means from different but similar distributions, we keep empirical and prior means separate, using the positive part in $f_a^+(q)$ to ensure a non-empty confidence set while penalizing exceedances above the prior.

6. Regret bound

Working in the setting of section 4.2 with our KL-UCB-Transfer algorithm, we have control over the number of pulls for each suboptimal arm through the following theorem.

Theorem 6 (KL-UCB-Transfer Regret Bound) *Suppose for each arm k we have N'_k i.i.d. samples from $\nu'_k = \mathcal{N}(\mu'_k, \sigma'^2)$ and that the true target law is $\nu_k = \mathcal{N}(\mu_k, \sigma^2)$ with $|\mu_k - \mu'_k| \le L_k$. Then under* KL-UCB-Transfer*, every suboptimal arm a satisfies*

$$\mathbb{E}[N_a(T)] \;\le\; \frac{2\,\sigma^2}{(\mu_1 - \mu_a)^2}\left(\ln T - N'_a\,\frac{\big(\mu_1 - (\mu'_a + L_a)\big)_+^2}{2\,\sigma'^2}\right)_+ + O\big((\ln T)^{2/3}\big).$$

Hence, we recover optimality according to Eq (3). The exact expression of the $O\big((\ln T)^{2/3}\big)$ term can be derived in the proof of Theorem 6 in the Appendix B. The main challenge is handling N'_a, which may be arbitrarily large and depend on T.

7. Experiments

We evaluate KL-UCB-Transfer in three simulations on a 6-armed bandit with unit-variance Gaussian rewards and true means

$$\mu = (1.0,\ 0.9,\ 0.8,\ 0.7,\ 0.6,\ 0.5),$$

so arm 1 is optimal. When injecting prior data on arm k, we use $N'_k = 1000$ synthetic samples from $\mathcal{N}(\mu'_k, 1)$ to observe a noticeable improvement in the lower bound (Eq. (3)).

Exploration bonus in simulations Instead of $\delta_t = \log t + 3 \log\log t$, we set $\delta_t = (1+\varepsilon)\log t$ with $\varepsilon = \frac{1}{20}$, as in Cappé et al. (2013). This reduces over-exploration ($\varepsilon \log T \leq 3 \log\log T$ for $T < 10^{152}$) and inflates the $\log T$ regret term by only a factor $1+\varepsilon$. Adapting the theory to this choice of δ_t is straightforward, and simulations confirm its improved performance.

Simulation 1: Priors on all arms

In this study we apply the *same* prior shift (δ, L) to every arm. A prior on arm k yields asymptotic improvement whenever $\mu_1 > \mu'_k + L_k$. Concretely, we set

$$\mu'_k = \mu_k + \delta, \quad L_k = L \quad \text{for } k = 1, \ldots, 6,$$

and sweep through four increasingly accurate settings:

$$(\delta, L) \in \{(0.20, 0.40),\ (0.11, 0.20),\ (0.05, 0.10),\ (0.00, 0.05)\}.$$

We plot five curves: the no-prior baseline ($N'_k = 0$ for all arms) plus one for each (δ, L). In the first setting $(0.20, 0.40)$, for every suboptimal arm k we have $\mu_1 < \mu'_k + L_k$, so there is no long-term regret benefit. In the last three (tighter) settings, some arms satisfy $\mu_1 > \mu'_k + L_k$, hence those priors do improve long-term regret. Figure 1(a) shows that only when (δ, L) is small enough does KL-UCB-Transfer outperform the no-prior baseline.

Simulation 2: Priors only on the optimal arm

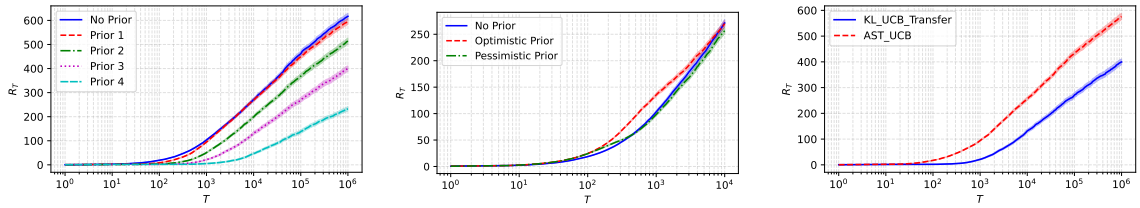
As shown in (Eq (4)), adding prior data for arm a reduces its index $U_a(t)$; hence a prior on a suboptimal arm is beneficial, while a prior on the optimal arm may not be. In our simulations, we focus on the effect of a prior placed only on the optimal arm. We restrict priors to the optimal arm ($k = 1$) and compare three configurations:

- No prior (baseline).
- Mildly optimistic (Prior 1): $\mu'_1 = \mu_1 + 0.001,\ L_1 = 0.004$.
- Pessimistic (Prior 2): $\mu'_1 = \mu_1 - 0.010,\ L_1 = 0.210$.

Define the hardship parameter $\eta := L_1 - (\mu'_1 - \mu_1)$ (see proof of Theorem 6). Although both priors are asymptotically harmless, a small η significantly slows early learning. Figure 1(b) illustrates that a very small η (mildly optimistic prior) leads to higher initial regret, whereas a larger η (pessimistic prior) grants an early advantage, even though both settings converge to the same asymptotic performance.

Simulation 3: Comparison with AST-UCB

We compare our algorithm KL-UCB-Transfer to AST-UCB from Rahul and Katewa (2024b). AST-UCB is designed for an *episodic* non-contextual bandit: the learner plays a sequence of episodes on the same arms, where at the start of each episode the arm means may shift by at most L from their values in the previous episode; AST-UCB uses past-episode data to compute the index in the new episode. In this simulation we use the third prior setting from Simulation 1, namely $(\delta, L) = (0.05, 0.10)$ on all arms, since it tightens the lower bound (Eq. (3)) and allows us to compare how efficiently our algorithm and AST-UCB leverage the prior data. We run AST-UCB with its matching episodic-shift parameter $L = 0.10$. Figure 1(c) demonstrates that KL-UCB-Transfer yields substantially lower regret than AST-UCB under this prior.



(a) Simulation 1 with Regret up to $T = 10^6$. Same priors on all arms.

(b) Simulation 2 with Regret up to $T = 10^4$. Priors on the optimal arm only.

(c) Simulation 3, with Regret up to $T = 10^6$ with prior $(\delta, L) = (0.05, 0.10)$.

Figure 1: Mean regret and error bars (100 runs per plot).

Takeaway message Overall, these simulations confirm that KL-UCB-Transfer can harness identical priors on every arm to reduce regret when the priors are accurate, reveal the nuanced short-term trade-offs when the prior is applied solely to the optimal arm, and consistently outperforms existing baselines in this transfer setting.

8. Conclusion

Asymptotic Lower Bound and KL-UCB-Transfer. In this work we have derived a sharp, problem-dependent asymptotic lower bound on cumulative regret that holds for any algorithm operating under arbitrary offline prior data. Focusing on the Gaussian bandit setting, we then introduced KL-UCB-Transfer, a simple modification of KL-UCB that incorporates priors via an extra KL penalty. We proved that KL-UCB-Transfer *matches* the general lower bound exactly when all arms are Gaussian with known variance and the prior samples N'_k are fixed, i.i.d. observations.

Simulation Results and Trade-offs. Our simulations confirm that, in regimes where the priors on suboptimal arms are sufficiently accurate, KL-UCB-Transfer enjoys strictly lower regret than the classical, no-prior KL-UCB. We also highlighted the short-term trade-offs that arise when one supplies priors on the optimal arm.

Future Directions. Looking forward, it would be interesting to extend our analysis beyond Gaussian rewards to exponential or bounded families, and to design transfer-aware algorithms that optimally balance prior bias against data-driven exploration in non-stationary or structured bandit environments. Moreover, in our current setting the prior sample sizes N'_k are non-random and the prior data are assumed i.i.d. A promising direction for future work is to study the case where prior data come from another online bandit process.

More broadly, our results offer a clean and tractable foundation to better understand transfer in simple settings, which can serve as a stepping stone toward principled transfer algorithms in richer and more realistic environments.

Acknowledgments

The authors gratefully acknowledge support from the French Ministry of Higher Education and Research, the Hauts-de-France Region, Inria and the MEL. They also thank the Scool research group for its excellent working environment. This work was supported by the PEPR FOUNDRY project (ANR-23-PEIA-0003).

References

Alexis Bellot, Alan Malek, and Silvia Chiappa. Transportability for bandits with data from different environments. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 44356–44381. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/8a8ce53beb3775522305e0a6033d4455-Paper-Conference.pdf.

Changxiao Cai, T. Tony Cai, and Hongzhe Li. Transfer learning for contextual multi-armed bandits. *The Annals of Statistics*, 52(1):207 – 232, 2024. doi: 10.1214/23-AOS2341. URL https://doi.org/10.1214/23-AOS2341.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, June 2013. doi: 10.1214/13-AOS1119. URL https://doi.org/10.1214/13-AOS1119.

Mingwei Deng, Ville Kyrki, and Dominik Baumann. Transfer learning in latent contextual bandits with covariate shift through causal transportability, 2025. URL https://arxiv.org/abs/2502.20153.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In Sham M. Kakade and Ulrike von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 359–376, Budapest, Hungary, 09–11 Jun 2011. PMLR. URL https://proceedings.mlr.press/v19/garivier11a.html.

Mohammad Gheshlaghi azar, Alessandro Lazaric, and Emma Brunskill. Sequential transfer in multi-armed bandit with finite set of models. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/062ddb6c727310e76b6200b7c71f63b5-Paper.pdf.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *COLT*, pages 67–79, 2010.

Junya Honda and Akimichi Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *Journal of Machine Learning Research*, 16(113):3721–3756, 2015. URL http://jmlr.org/papers/v16/honda15a.html.

Emilie Kaufmann, Nathaniel Korda, and Remi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory (ALT)*, 2012.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. ISSN 0196-8858. doi: https://doi.

org/10.1016/0196-8858(85)90002-8. URL https://www.sciencedirect.com/science/article/pii/0196885885900028.

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

Odalric-Ambrym Maillard. *Mathematics of Statistical Sequential Decision Making*. Habilitation à diriger des recherches, Université de Lille, Sciences et Technologies, February 2019. URL https://hal.science/tel-02162189.

Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In Sham M. Kakade and Ulrike von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 497–514, Budapest, Hungary, 09–11 Jun 2011. PMLR. URL https://proceedings.mlr.press/v19/maillard11a.html.

NR Rahul and Vaibhav Katewa. Exploiting adjacent similarity in multi-armed bandit tasks via transfer of reward samples. *arXiv preprint arXiv:2409.19975*, 2024a.

NR Rahul and Vaibhav Katewa. Transfer in sequential multi-armed bandits via reward samples. In *2024 European Control Conference (ECC)*, pages 2083–2089. IEEE, 2024b.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952. doi: 10.1090/S0002-9904-1952-09620-8. URL https://projecteuclid.org/euclid.bams/1183519660.

Hassan Saber, Léo Saci, Odalric-Ambrym Maillard, and Audrey Durand. Routine bandits: Minimizing regret on recurring problems. In Nuria Oliver, Fernando Pérez-Cruz, Stefan Kramer, Jesse Read, and Jose A. Lozano, editors, *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 3–18, Cham, 2021. Springer International Publishing. ISBN 978-3-030-86486-6.

Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7), 2009.

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. ISSN 00063444. URL http://www.jstor.org/stable/2332286.

Zhi Wang, Chicheng Zhang, and Kamalika Chaudhuri. Thompson sampling for robust transfer in multi-task bandits. In *Proceedings of the 39th International Conference on Machine Learning*, 2022. URL https://arxiv.org/abs/2206.08556. To appear in ICML 2022.

Junzhe Zhang and Elias Bareinboim. Transfer learning in multi-armed bandits: A causal approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1340–1346, 2017. doi: 10.24963/ijcai.2017/186. URL https://doi.org/10.24963/ijcai.2017/186.

Appendix A. Proof of Theorem 3

We now prove Theorem 3. For this, we begin with the following lemma.

Lemma 7 *Let (ν, ν') and $(\tilde{\nu}, \tilde{\nu}')$ be two sets of arm distributions in our setting. Denote by $\mathbb{P}_{(\nu,\nu')\pi,T}$ the canonical probability measure of the bandit model using strategy π and arm distributions (ν, ν') up to time T. Then,*

$$\mathrm{KL}(\mathbb{P}_{(\nu,\nu')\pi,T}, \mathbb{P}_{(\tilde{\nu},\tilde{\nu}')\pi,T}) = \sum_{k=1}^{K} \mathbb{E}_{(\nu,\nu')}\left[N_k(T)\right] \mathrm{KL}(\nu_k, \tilde{\nu}_k) + \sum_{k=1}^{K} N'_k \,\mathrm{KL}(\nu'_k, \tilde{\nu}'_k).$$

This lemma parallels Lemma 15.1 in Lattimore and Szepesvári (2020), except for the term $\sum_{k=1}^{K} N'_k \,\mathrm{KL}(\nu'_k, \tilde{\nu}'_k)$ accounting for initial offline data.

Proof [of Theorem 3]

Let k be a suboptimal arm for (ν, ν'). Consider a modification $(\tilde{\nu}, \tilde{\nu}') \in \mathcal{D} \times \mathcal{D}'$ such that arm k becomes optimal, while keeping the distributions $(\nu_i, \nu'_i)_{i \neq k}$ unchanged. Then, using the data-processing inequality (e.g., (Maillard, 2019, Lemma 3.3)) combined with the KL decomposition of Lemma 7, we get:

$$\mathbb{E}_{(\nu,\nu')}\left[N_k(T)\right] \mathrm{KL}(\nu_k, \tilde{\nu}_k) + N'_k \,\mathrm{KL}(\nu'_k, \tilde{\nu}'_k) \geq kl\left(\frac{\mathbb{E}_{(\nu,\nu')}[N_k(T)]}{T}, \frac{\mathbb{E}_{(\tilde{\nu},\tilde{\nu}')}[N_k(T)]}{T}\right).$$

Now using that for all $p, q \in (0, 1)$ we have

$$kl(p, q) = (1-p)\ln\left(\frac{1}{1-q}\right) + \underbrace{p\ln(1/q)}_{\geq 0} + \underbrace{p\ln(p) + (1-p)\ln(1-p)}_{\geq -\ln 2},$$

we have that $\mathbb{E}_{(\nu,\nu')}\left[N_k(T)\right] \mathrm{KL}(\nu_k, \tilde{\nu}_k) + N'_k \,\mathrm{KL}(\nu'_k, \tilde{\nu}'_k)$ is lower bounded by

$$\left(1 - \frac{\mathbb{E}_{(\nu,\nu')}[N_k(T)]}{T}\right) \ln\left(\frac{T}{T - \mathbb{E}_{(\tilde{\nu},\tilde{\nu}')}[N_k(T)]}\right) - \ln(2)$$

Using the consistency of the strategy π, we have for any $\alpha > 0$:

- $\mathbb{E}_{(\nu,\nu')}[N_k(T)] = o(T^{\alpha})$,
- $T - \mathbb{E}_{(\tilde{\nu},\tilde{\nu}')}[N_k(T)] = o(T^{\alpha})$.

Therefore, for any $\alpha > 0$,

$$\frac{\mathbb{E}_{(\nu,\nu')}[N_k(T)] \,\mathrm{KL}(\nu_k, \tilde{\nu}_k) + N'_k \,\mathrm{KL}(\nu'_k, \tilde{\nu}'_k)}{\ln T} \geq (1 - o(T^{-(1-\alpha)}))(1-\alpha) + o_{T\to\infty}(1).$$

Since this is true for all $\alpha > 0$, taking the infimum over all $(\tilde{\nu}, \tilde{\nu}') \in \mathcal{D} \times \mathcal{D}'$ such that arm k becomes optimal (with all other arms unchanged) yields the result in Theorem 3. ■

Proof of Lemma 7

Proof

We follow the same idea as in the proof of Lemma 15.1 from Lattimore and Szepesvári (2020). Let $p_{(\nu,\nu')\pi,T}$ denote the density of $\mathbb{P}_{(\nu,\nu')\pi,T}$, the probability of the space of T turn under the policy π with the arm distribution (ν,ν'). Let x' be the offline data. Then,

$$\begin{aligned} p_{(\nu,\nu')\pi,T}(x', a_1, x_1, ..., a_T, x_T) &= p_{\nu'}(x') \prod_{t=1}^{T} \pi_t(a_t \mid x', a_1, x_1, ..., a_{t-1}, x_{t-1}) p_{\nu_{a_t}}(x_t) \\ &= \prod_{k=1}^{K} \prod_{i=1}^{N'_k} p_{\nu'_k}(x'_{i,k}) \prod_{t=1}^{T} \pi_t(a_t \mid x', a_1, x_1, ..., a_{t-1}, x_{t-1}) p_{\nu_{a_t}}(x_t). \end{aligned}$$

Taking the likelihood ratio with $p_{(\tilde{\nu},\tilde{\nu}')\pi,T}$ gives:

$$\frac{p_{(\nu,\nu')\pi,T}(x', a_1, x_1, ..., a_T, x_T)}{p_{(\tilde{\nu},\tilde{\nu}')\pi,T}(x', a_1, x_1, ..., a_T, x_T)} = \prod_{k=1}^{K} \prod_{i=1}^{N'_k} \frac{p_{\nu'_k}(x'_{i,k})}{p_{\tilde{\nu}'_k}(x'_{i,k})} \prod_{t=1}^{T} \frac{p_{\nu_{a_t}}(x_t)}{p_{\tilde{\nu}_{a_t}}(x_t)}.$$

Taking the log of both sides:

$$\ln \frac{\mathbb{P}_{(\nu,\nu')\pi,T}}{\mathbb{P}_{(\tilde{\nu},\tilde{\nu}')\pi,T}}(x', a_1, x_1, ..., a_T, x_T) = \sum_{k=1}^{K} \sum_{i=1}^{N'_k} \ln \frac{p_{\nu'_k}(x'_{i,k})}{p_{\tilde{\nu}'_k}(x'_{i,k})} + \sum_{t=1}^{T} \ln \frac{p_{\nu_{a_t}}(x_t)}{p_{\tilde{\nu}_{a_t}}(x_t)}.$$

Taking expectation under (ν,ν') yields:

$$\begin{aligned} \mathbb{E}_{(\nu,\nu')}\left[\ln \frac{\mathbb{P}_{(\nu,\nu')\pi,T}}{\mathbb{P}_{(\tilde{\nu},\tilde{\nu}')\pi,T}}\right] &= \sum_{k=1}^{K} N'_k \,\mathrm{KL}(\nu'_k, \tilde{\nu}'_k) + \sum_{t=1}^{T} \mathbb{E}_{(\nu,\nu')}\left[\mathrm{KL}(\nu_{a_t}, \tilde{\nu}_{a_t})\right] \\ &= \sum_{k=1}^{K} N'_k \,\mathrm{KL}(\nu'_k, \tilde{\nu}'_k) + \sum_{t=1}^{T} \mathbb{E}_{(\nu,\nu')}\left[\sum_{k=1}^{K} \mathbf{1}(A_t = k)\,\mathrm{KL}(\nu_k, \tilde{\nu}_k)\right] \\ &= \sum_{k=1}^{K} N'_k \,\mathrm{KL}(\nu'_k, \tilde{\nu}'_k) + \sum_{k=1}^{K} \mathbb{E}_{(\nu,\nu')}\left[N_k(T)\right] \mathrm{KL}(\nu_k, \tilde{\nu}_k). \end{aligned}$$

■

Appendix B. Proof of Theorem 6

We will use the shorthand

$$f_a^+(q) = N'_a \frac{\left(q - (\hat{\mu}'_a + L_a)\right)_+^2}{2\,\sigma'^2}, \quad d(a,b) = \frac{(b-a)^2}{2\,\sigma^2}, \quad d^+(a,b) = \frac{(b-a)_+^2}{2\,\sigma^2}.$$

We use the following lemmas. Lemma 8 is proved in Section B.1; proof sketches for Lemmas 9, 10, and 11 appear in Sections B.2, B.3, and B.4, with full proofs in the Supplement.

Lemma 8 *For all $T > 1$, for all arm $a \neq 1$, we have the inequality*

$$\sum_{t=1}^{T} \mathbf{1}\left(A_t = a,\ \mu_1 \leq U_1(t)\right) \leq \sum_{s=1}^{T} \mathbf{1}\left(s\, d^+(\hat{\mu}_{a,s}, \mu_1) + f_a^+(\mu_1) \leq \delta_T\right).$$

Lemma 9 *For every $\delta_t \underset{T\to+\infty}{\longrightarrow} +\infty$ and for all arm $a \neq 1$, we have the inequality*

$$\mathbb{E}\left[(\delta_T - f_a^+(\mu_1))_+\right] \leq \left(\delta_T - N'_a \frac{(\mu_1 - (\mu'_a + L_a))_+^2}{2\sigma'^2}\right)_+ + O(\sqrt{\delta_T}).$$

Lemma 10 *There exists $C' > 0$ that depends only on $\sigma, \sigma', L_1, \mu_1, \mu'_1$, such that:*

$$\mathbb{E}\left[\exp\left(f_1^+(\mu_1)\right)\right] \leq C'.$$

Lemma 11 *For all $t \geq 1$, using the choice $\delta_t := \ln t + 3\ln(\max(1, \ln t))$,*

$$\mathbb{P}(\mu_1 > U_1(t)) = O\left(\frac{\mathbb{E}[e^{f_1^+(\mu_1)}]}{t \ln t}\right).$$

We bound the expected pulls of a suboptimal arm a by splitting into two cases:

1. Times when the upper confidence bound of the optimal arm underestimates its true mean ($\mu_1 > U_1(t)$). By choosing δ_t appropriately and using Lemma 10 and 11, this case contributes only $o(\ln T)$ to the regret.

2. Times when $U_1(t)$ is above μ_1 but arm a is still selected. Here one applies Lemma 8 to show that after a random threshold M, further pulls of a become unlikely, yielding the main $\ln T$ order term (up to lower-order corrections from Lemma 9).

Combining these two bounds gives the desired logarithmic bound on $\mathbb{E}[N_a(T)]$.
The changes in the proof, compared to that in Garivier and Cappé (2011); Cappé et al. (2013), include additional conditioning steps, Lemma 9, and Lemma 10, which arise from the inclusion of prior data.

Proof [of Theorem 6]

We begin by bounding the expected number of pulls of a suboptimal arm a:

$$\mathbb{E}\left[N_a(T)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left(A_t = a\right)\right] \leq \sum_{t=1}^{T} \mathbb{P}\left(\mu_1 > U_1(t)\right) + \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}\left(A_t = a,\ \mu_1 \leq U_1(t)\right)\right] \quad (5)$$

Now using Lemmas 10 and 11 we get:

$$\sum_{t=1}^{T} \mathbb{P}\left(\mu_1 > U_1(t)\right) = \sum_{t=1}^{T} O\left(\frac{1}{t \ln t}\right) = O(\ln \ln T) \quad (6)$$

For the second term in (5), using Lemma 8, define the random quantity:

$$M := \frac{(\delta_T - f_a^+(\mu_1))_+}{(1-\varepsilon)^2 d(\mu_a, \mu_1)}$$

Then:

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left(A_t=a,\ \mu_1\le U_1(t)\right)\right] \tag{7}$$

$$\begin{aligned}
&\overset{(a)}{=}\mathbb{E}\left[\sum_{s=1}^{T}\mathbb{P}\left(s\,d^+(\hat{\mu}_{a,s},\mu_1)+f_a^+(\mu_1)\le\delta_T\middle|f_a^+\right)\right]\\
&\le\mathbb{E}\left[M-1+\sum_{s=M}^{T}\mathbb{P}\left(s\,d^+(\hat{\mu}_{a,s},\mu_1)+f_a^+(\mu_1)\le\delta_T\middle|f_a^+\right)\right]\\
&\le\mathbb{E}\left[M-1+\sum_{s=M}^{T}\mathbb{P}\left(d^+(\hat{\mu}_{a,s},\mu_1)\le\frac{\delta_T-f_a^+(\mu_1)}{M}\middle|f_a^+\right)\right]\\
&\overset{(b)}{\le}\mathbb{E}\left[M-1+\sum_{s=M}^{T}\mathbb{P}\left(d^+(\hat{\mu}_{a,s},\mu_1)\le(1-\varepsilon)^2d(\mu_a,\mu_1)\right)\right]\\
&\le\frac{\mathbb{E}\left[(\delta_T-f_a^+(\mu_1))_+\right]}{(1-\varepsilon)^2d(\mu_a,\mu_1)}-1+\sum_{s=1}^{T}\mathbb{P}\left(d^+(\hat{\mu}_{a,s},\mu_1)\le(1-\varepsilon)^2d(\mu_a,\mu_1)\right),
\end{aligned} \tag{8}$$

where (a) follows by Lemma 8 and by conditioning and (b) by the definition of M. We now apply a standard Chernoff bound for Gaussian tails for the first inequality:

$$\begin{aligned}
\sum_{s=1}^{T}\mathbb{P}\left(d^+(\hat{\mu}_{a,s},\mu_1)\le(1-\varepsilon)^2d(\mu_a,\mu_1)\right)&=\sum_{s=1}^{T}\mathbb{P}\left(\hat{\mu}_{a,s}\ge\mu_a+\varepsilon(\mu_1-\mu_a)\right)\\
&\le\sum_{s=1}^{T}\exp\left(-\frac{\varepsilon^2(\mu_1-\mu_a)^2}{2\sigma^2}s\right)\\
&\le\frac{1}{1-\exp\left(-\frac{\varepsilon^2(\mu_1-\mu_a)^2}{2\sigma^2}\right)}\\
&=O(\varepsilon^{-2}).
\end{aligned} \tag{9}$$

Using Equations (8), (9) and Lemma 9, we finally obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left(A_t=a,\ \mu_1\le U_1(t)\right)\right]&\le\frac{\left(\delta_T-N_a'\frac{(\mu_1-(\mu_a'+L_a))_+^2}{2\sigma'^2}\right)_+}{(1-\varepsilon)^2d(\mu_a,\mu_1)}+O(\varepsilon^{-2})+O(\sqrt{\ln T})\\
&=\frac{2\sigma^2}{(\mu_1-\mu_k)^2}\left(\ln T-N_a'\frac{(\mu_1-(\mu_a'+L_a))_+^2}{2\sigma'^2}\right)_++O(\varepsilon\ln T)+O(\varepsilon^{-2})+O(\sqrt{\ln T}).
\end{aligned}$$

By taking $\varepsilon=(\ln T)^{-\frac{1}{3}}$ we get

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbf{1}\left(A_t=a,\ \mu_1\le U_1(t)\le\right)\right]\le\frac{2\sigma^2}{(\mu_1-\mu_k)^2}\left(\ln T-N_a'\frac{(\mu_1-(\mu_a'+L_a))_+^2}{2\sigma'^2}\right)_++O\left((\ln T)^{\frac{2}{3}}\right). \tag{10}$$

Combining Equations (5), (6), and (10), we conclude that,

$$\mathbb{E}\left[N_a(T)\right] \leq \frac{2\sigma^2}{(\mu_1-\mu_k)^2}\left(\ln T - N'_a\frac{(\mu_1-(\mu'_a+L_a))_+^2}{2\sigma'^2}\right)_+ + O\left((\ln T)^{\frac{2}{3}}\right).$$

■

B.1. Proof of Lemma 8

The proof mimes the one of Lemma 7 in Garivier and Cappé (2011)
Proof If $A_t = a$ and $\mu_1 \leq U_1(t)$, then $U_a(t) \geq U_1(t) \geq \mu_1$, and therefore:

$$N_a(t)d^+(\hat{\mu}_a(t),\mu_1) + f_a^+(\mu_1) \leq N_a(t)d^+(\hat{\mu}_a(t),U_a(t)) + f_a^+(U_a(t)) = \delta_t \leq \delta_T$$

Hence,

$$\begin{aligned}
\sum_{t=1}^{T}\mathbf{1}\left(A_t = a, \mu_1 \leq U_1(s)\right) &\leq \sum_{t=1}^{T}\mathbf{1}\left(A_t = a, N_a(t)d^+(\hat{\mu}_a(t),\mu_1) + f_a^+(\mu_1) \leq \delta_T\right)\\
&= \sum_{t=1}^{T}\sum_{s=1}^{t}\mathbf{1}\left(N_a(t) = s, A_t = a, sd^+(\hat{\mu}_{a,s},\mu_1) + f_a^+(\mu_1) \leq \delta_T\right)\\
&= \sum_{s=1}^{T}\mathbf{1}\left(sd^+(\hat{\mu}_{a,s},\mu_1) + f_a^+(\mu_1) \leq \delta_T\right)\sum_{t=s}^{T}\mathbf{1}\left(N_a(t) = s, A_t = a\right)\\
&\leq \sum_{s=1}^{T}\mathbf{1}\left(sd^+(\hat{\mu}_{a,s},\mu_1) + f_a^+(\mu_1) \leq \delta_T\right)
\end{aligned}$$

■

B.2. Proof Sketch of Lemma 9

Proof [Proof Sketch] Let

$$X := \delta_t - \frac{N'_a}{2\sigma'^2}\left(\mu_1 - (\hat{\mu}'_a + L_a)\right)_+^2, \quad \xi := \mu_1 - L_a - \mu'_a,\ \beta := \xi\sqrt{N'_a/\sigma'^2}.$$

Using that $\hat{\mu}'_a$ is gaussian we got $\mathbb{E}[X_+] = \int_0^{\delta_t}\Phi\left(\sqrt{2(\delta_t - t)} - \beta\right)dt =: I(\beta,\delta_t)$. Writing $\delta = \delta_t$, after some calculations one obtains the closed form

$$I(\beta,\delta) = \frac{2\delta - \beta^2 - 1}{2}\,\Phi(\sqrt{2\delta} - \beta) + \frac{\beta^2+1}{2}\,\Phi(-\beta) + \frac{\sqrt{2\delta}+\beta}{2}\,\phi(\sqrt{2\delta} - \beta) - \frac{\beta}{2}\,\phi(\beta).$$

Analyzing the three regimes $\beta \leq 0$, $0 \leq \beta \leq \sqrt{2\delta}$, and $\beta \geq \sqrt{2\delta}$ gives

$$I(\beta,\delta)\ \leq\ \left(\delta - \tfrac{\beta_+^2}{2}\right)_+\ +\ O(\sqrt{\delta}).$$

■

B.3. Proof Sketch of Lemma 10

Proof [Proof Sketch] Let $Z = \mu_1' - \hat{\mu}_1' \sim \mathcal{N}(0, \sigma'^2/N_1')$ and set

$$\eta := L_1 - (\mu_1' - \mu_1), \quad a := \frac{\eta}{\sigma'/\sqrt{N_1'}} > 0, \quad W := \frac{Z\sqrt{N_1'}}{\sigma'} \sim \mathcal{N}(0,1).$$

Then

$$f_1^+(\mu_1) = \tfrac{N_1'}{2\sigma'^2}\big(\mu_1 - (\hat{\mu}_1' + L_1)\big)_+^2 = \tfrac{1}{2}\,(W-a)_+^2,$$

so

$$\mathbb{E}\big[e^{f_1^+(\mu_1)}\big] = \int_{-\infty}^{a} \varphi(w)\,dw + \int_a^{\infty} e^{\frac{1}{2}(w-a)^2}\varphi(w)\,dw = \Phi(a) + \int_a^{\infty} \frac{1}{\sqrt{2\pi}} e^{-aw+\frac{a^2}{2}}\,dw = \Phi(a) + \frac{\varphi(a)}{a}.$$

Since $a \geq \frac{\eta}{\sigma'} > 0$, $\Phi(a) + \varphi(a)/a \leq 1 + \frac{\sigma'}{\eta\sqrt{2\pi}} =: C' < \infty$. This completes the proof. ■

B.4. Proof Sketch of Lemma 11

Proof [Proof Sketch] Denote for simplicity $f := f_1^+(\mu_1)$. Then

$$\mathbb{P}(\mu_1 > U_1(t)) = \mathbb{E}\big[\mathbb{P}(\mu_1 > U_1(t) \mid f)\big] \leq \mathbb{E}\Big[1_{f \geq \delta_t} + 1_{f<\delta_t}\,\mathbb{P}\big(N_1(t)d^+(\hat{\mu}_1, \mu_1) > \delta_t - f \mid f\big)\Big].$$

When $f < \delta_t$, use peeling arguments with $N_1(t) \in (n_{m-1}, n_m]$ with $n_m = \lceil \gamma^m \rceil$, $\gamma = \delta_t/(\delta_t - 1)$, $M := \lceil \ln t/\ln\gamma \rceil \leq \lceil \delta_t \ln t \rceil$. In block m, set

$$\varepsilon_m := \sqrt{\frac{2\sigma^2(\delta_t - f)}{n_m}}, \quad S_i := \sum_{j=1}^{i} X_j.$$

Using the martingale $W_{\lambda,i} := \exp(\lambda S_i - i\phi(\lambda))$, $\phi(\lambda) = \lambda\mu_1 + \frac{\lambda^2\sigma^2}{2}$, pick $\lambda = -\varepsilon_m/\sigma^2$. By Doob's maximal inequality,

$$\mathbb{P}(\exists i \in (n_{m-1}, n_m] : S_i \leq i(\mu_1 - \varepsilon_m)) \leq \exp\big(-(n_{m-1}+1)(\delta_t - f)/n_m\big) \leq \exp\big(-(\delta_t - f)/\gamma\big).$$

Summing $m = 1, \ldots, M$ gives

$$\mathbb{P}(\mu_1 > U_1(t) \mid f) \leq 1_{f \geq \delta_t} + Me^{-(\delta_t - f)/\gamma} \leq 1_{f\geq\delta_t} + \lceil \delta_t \ln t \rceil e^{-(\delta_t - f)/\gamma} = 1_{f\geq\delta_t} + O\Big(\frac{e^f}{t\ln t}\Big).$$

Since $1_{f\geq\delta_t} \leq e^{f-\delta_t} = O(e^f/(t\ln^3 t))$,

$$\mathbb{P}(\mu_1 > U_1(t)) = O\left(\frac{\mathbb{E}[e^{f_1^+(\mu_1)}]}{t\ln t}\right).$$

■