Supplemental Appendix for “Sequential Decision Problems with Missing Feedback”

Filippo Palomba

Princeton University

This version: July 25, 2025.

Abstract

This supplement contains all proofs, additional results, and other technical details. Section SA1 describes the setup and notation, states the assumptions we rely on, and introduces some auxiliary results. Section SA2 illustrates the main technical results. Section SA3 describes in detail the setting used for the simulation study. Section SA4 contains all the proofs.

Complete Contents

Table SA-1: Summary of Notation

Quantity	**Description**
Environment	
$\mathcal{A}$	set of actions
A	number of actions
R_a	reward for action $a \in \mathcal{A}$
C_a	censoring mechanism for action $a \in \mathcal{A}$
$\mathbf{X}_a$	observed covariates for action $a \in \mathcal{A}$
$\mathcal{R}$	support of rewards
$\mathcal{X}$	support of covariates
$\mathcal{Z}$	support of $(R_a, C_a, \mathbf{X}_a)$
$\mathcal{C}_j$	class of bandits, $j \in \{0, 1, 2\}$
ν_a	probability measure of $(R_a, C_a, \mathbf{X}_a)$ for action $a \in \mathcal{A}$
θ_a	unconditional mean reward for action $a \in \mathcal{A}$
$\overline{\theta}$	best mean reward
$a^\star$	action associated with the best mean reward
$\theta_a(\mathbf{X}_a)$	conditional (on $\mathbf{X}_a$) mean reward for action $a \in \mathcal{A}$
q_a	probability of censoring for action $a \in \mathcal{A}$
$A_{\mathtt{cen}}$	sum of inverse of censoring probabilities
$\underline{q}$	minimum probability of censoring across actions
$q_a(\mathbf{X}_a)$	conditional probability of censoring
σ_a	variance proxy of the reward for action $a \in \mathcal{A}$
$\bar{\sigma}$	largest variance proxy across all actions
$\overline{K}$	uniform upper bound for all θ_as
Algorithms	
T	number of rounds
Π	space of policy functions
π_t^{x}	policy function at round $t \in [T]$, $\mathsf{x} \in \{\mathsf{UCB}, \mathsf{DR}, \mathsf{ODR}\}$
A_t	action chosen at round $t \in [T]$
$\mathsf{Regret}_T^{\mathsf{x}}$	pseudo-regret of algorithm $\mathsf{x} \in \{\mathsf{UCB}, \mathsf{DR}, \mathsf{ODR}\}$
$\mathsf{Regret}_T^{\star}(\mathcal{C})$	minimax regret over class of bandits $\mathcal{C}$ and policies Π
δ	probability with which high-probability bounds do not hold
$P_a(t)$	times action $a \in \mathcal{A}$ has been played at the beginning of round $t \in [T]$
$N_a(t)$	times the reward of action $a \in \mathcal{A}$ has been observed at the beg. of round $t \in [T]$
$\widehat{R}_a^{\mathsf{x}}(t)$	mean reward estimator, $\mathsf{x} \in \{\mathsf{UCB}, \mathsf{DR}, \mathsf{ODR}\}$
$\widetilde{R}_a^{\mathsf{x}}(t, \delta)$	optimistic mean reward estimator, $\mathsf{x} \in \{\mathsf{UCB}, \mathsf{DR}, \mathsf{ODR}\}$
$b_a^{\mathsf{x}}(\delta)$	bonus term, $\mathsf{x} \in \{\mathsf{UCB}, \mathsf{DR}, \mathsf{ODR}\}$
$\underline{q}_\lambda$	minimum regularized probability of censoring across actions
$\hat{q}_a(\cdot)$	estimator of the conditional probability of censoring for action $a \in \mathcal{A}$
$\hat{\theta}_a(\cdot)$	estimator of the conditional expected reward for action $a \in \mathcal{A}$
$\mathsf{Err}_t(\mathsf{x})$	ℓ_2-error rate for estimator $\mathsf{x} \in \{(\hat{q}_a, \hat{\theta}_a), a \in \mathcal{A}\}$
λ	regularization parameter for UCB algorithm
K_{ODR}	regret constant

SA1 Introduction

This section introduces the notation used in this project (Section SA1.1), describes the setup and the algorithms analyzed (Sections SA1.2 and SA1.3), outlines the assumptions I rely on (Section SA1.4), and presents some auxiliary lemmas that will prove useful throughout (Section SA1.5).

SA1.1 Notation

Sets. In general, blackboard bold uppercase letters ($e.g., \mathbb{N}, \mathbb{R}$) are used to denote standard sets of numbers. All the other sets are denoted with uppercase calligraphic letters ($e.g., \mathcal{F}, \mathcal{G}$). The set of natural numbers is denoted with $\mathbb{N}$, the set of real numbers with $\mathbb{R}$, the set of non-negative real numbers with $\mathbb{R}_+$, and the set of positive real numbers with $\mathbb{R}_{++}$. I write $\mathbb{R}^d$ for $d \in \mathbb{N}$ to denote $\mathbb{R}^d = \times_{j=1}^d \mathbb{R}$, where $\times$ denotes the Cartesian product between sets. I denote an ordered set $\{1, \dots, n\}, n \in \mathbb{N}$ with $[n]$. The complement of a set $\mathcal{F}$ is denoted as $\overline{\mathcal{F}}$. I denote the sigma-algebra generated by a random variable X as $\sigma(X)$ and with $\mathscr{B}(\mathcal{S})$ the Borel σ-algebra on the topological space $\mathcal{S}$.

Linear algebra. Throughout the text, $\mathbf{0}_k$ and $\mathbf{1}_k$ denote the k-dimensional zero and one vectors, respectively. For a k by m matrix $\mathbf{A}$ I use $\mathbf{A}^\top$ to denote the transpose of $\mathbf{A}$ and, if $k = m$ and $\mathbf{A}$ is non-singular I use $\mathbf{A}^{-1}$ to denote the inverse of $\mathbf{A}$. For $x \in \mathbb{R}^k$, I write $\mathbf{x} \succeq \mathbf{0}_k$ to denote the component-wise inequality in $\mathbb{R}^k$. Let $\mathbf{x} \in \mathbb{R}^d$ I use $|\mathbf{x}| := \left(\sum_{i\in[d]} x_i^2\right)^{1/2}$ to the denote the Euclidean norm and $\|\mathbf{x}\|_\infty := \sup_{i\in[d]} |x_i|$ to denote the sup-norm.

Asymptotic statements. For two positive sequences $\{a_n\}_n, \{b_n\}_n$, I write $a_n = O(b_n)$ if $\exists\, M \in \mathbb{R}_{++} : a_n \leq M b_n$ for all large n, $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n b_n^{-1} = 0$, $a_n = \widetilde{O}(b_n)$ if $\exists\, k \in \mathbb{N}, C \in \mathbb{R}_{++} : a_n = O(b_n \ln^k(Cn))$ $a_n \lesssim b_n$ if there exists a constant $C \in \mathbb{R}_{++}$ such that $a_n \leq C b_n$ for all large n, and $a_n \sim b_n$ if $a_n/b_n \to 1$ as $n \to \infty$. For two sequences of random variables $\{A_n\}_n, \{B_n\}_n$, I write $A_n = o_{\mathbb{P}}(B_n)$ if $\forall\, \varepsilon \in \mathbb{R}_{++}, \lim_{n\to\infty} \mathbb{P}[|A_n B_n^{-1}| \geq \varepsilon] = 0$ and $A_n = O_{\mathbb{P}}(B_n)$ if $\forall\, \varepsilon \in \mathbb{R}_{++}, \exists\, M, n_0 \in \mathbb{R}_{++} : \mathbb{P}[|A_n B_n^{-1}| > M] < \varepsilon$, for $n > n_0$.

Statistical Distributions. I denote a (possibly multivariate) Gaussian random variable with $\mathsf{N}(\mathbf{a}, \mathbf{B})$, where $\mathbf{a}$ denotes the mean and $\mathbf{B}$ the variance-covariance, with $\mathsf{Be}(p)$ a Bernoulli distribution with $p \in (0, 1]$ denoting the success probability, and with $\mathsf{sG}(\sigma)$ a sub-Gaussian random variable. A random variable X is sub-Gaussian with variance proxy $\sigma > 0$ if $\forall\, \lambda \in \mathbb{R}, \mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2\sigma^2/2)$ and $\mathbb{E}[X] = 0$. With a slight abuse of terminology, I say that a random variable Y with non-zero mean is sub-Gaussian when $(Y - \mathbb{E}[Y]) \sim \mathsf{sG}(\sigma)$. If $\{X_t\}_{t=1}^\infty$ is an $\mathcal{F}$-adapted martingale difference sequence with respect to some filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t=1}^\infty$, then it is understood that $X_t \sim \mathsf{sG}(\sigma)$ requires

$\forall\, \lambda \in \mathbb{R}, \mathbb{E}[\exp(\lambda X_t) \mid \mathcal{F}_t] \leq \exp(\lambda^2\sigma^2/2)$ and $\mathbb{E}[X_t \mid \mathcal{F}_t] = 0$. All probability measures are assumed to belong to the set of all Borel probability measures on an appropriately defined topological space $\mathcal{S}$.

Table SA-1 summarizes the notation specific to this project. Notation that is only used in the proofs is omitted from the table and defined throughout.

SA1.2 Setup

I start by describing a generic instance of a stochastic MAB with possibly missing rewards and the decision-maker that interacts with such an environment.

Setting. A decision-maker faces a sequential decision problem over $T \in \mathbb{N}$ rounds in a stochastic environment. At the beginning of each round $t \in [T]$, using all the information available at that point, the decision-maker selects an action $A_t \in \mathcal{A} := \{1, \dots, A\}$. Each action $a \in \mathcal{A}$ is associated with a tuple of random variables: a reward $R_a \in \mathcal{R} \subseteq \mathbb{R}$, an indicator for not being missing $C_a \in \{0, 1\}$, and some covariates $\mathbf{X}_a \in \mathcal{X} \subseteq \mathbb{R}^k, k \in \mathbb{N}$. A stochastic MAB problem with missing rewards is defined as a collection of random variables $\{(R_{a,\ell}, C_{a,\ell}, \mathbf{X}_{a,\ell})\}_{a\in\mathcal{A},\ell\in[T]}$ with the first index running over the set of actions $\mathcal{A}$, the second index running over rounds, and satisfying the following three conditions for fixed actions $a, a' \in \mathcal{A} : a \neq a'$:

1. $\{(R_{a,t}, C_{a,t}, \mathbf{X}_{a,t})\}_{t\in[T]}$ are T independent draws from $(R_a, C_a, \mathbf{X}_a)$, which is distributed according to some (unknown) probability measure ν_a defined on the measurable space $(\mathcal{R} \times \{0, 1\} \times \mathcal{X}, \sigma(R_a, C_a, \mathbf{X}_a))$;
2. $(R_a, C_a, \mathbf{X}_a) \perp\!\!\!\perp (R_{a'}, C_{a'}, \mathbf{X}_{a'})$, so that the unknown joint distribution of $\{(R_a, C_a, \mathbf{X}_a)\}_{a\in\mathcal{A}}$ can be defined as $\nu = \prod_{a\in\mathcal{A}} \nu_a$;
3. the reward $R_{a,t}$ is observed by the decision-maker only if $C_{a,t} = 1$.

Each action $a \in \mathcal{A}$ has an associated mean reward

$$\theta_a \equiv \theta_a(\nu) = \mathbb{E}_\nu[R_a],$$

which I assume to be finite. I define the best (in hindsight) action, the associated best mean reward, and the sub-optimality gap as

$$a^\star := \arg\max_{a\in\mathcal{A}} \theta_a, \qquad \overline{\theta} := \max_{a\in\mathcal{A}} \theta_a, \qquad \Delta_a := \overline{\theta} - \theta_a,\ a \in \mathcal{A}.$$

It follows from the description above that all that is needed to characterize a MAB with possibly missing rewards is the collection of probability measures $\{\nu_a\}_{a\in\mathcal{A}}$. In this work, I focus on the following class of bandits

$$\mathcal{C} := \left\{(\nu_a)_{a\in\mathcal{A}} : \nu_a^{[R]} \in \mathcal{SG}(\sigma_a),\ \nu_a^{[C]} = \mathsf{Be}(q_a), q_a \in (0,1]\right\},$$

where $\nu_a^{[Y]}$ denotes the marginal of ν_a with respect to $Y \in \{C, R\}$ and $\mathcal{SG}(\sigma)$ denotes the space of sub-Gaussian probability distribution with variance proxy at most $\sigma > 0$, and $\mathsf{Be}(p)$ denotes the probability distribution of a Bernoulli random variable with parameter p. The class of bandits $\mathcal{C}$ is very general, as it only restricts the mean reward to be finite, the tails to be sub-Gaussian, and rules out the trivial case in which the reward of action $a \in \mathcal{A}$ is not observable (which would occur when $q_a = 0$). Finally, I define $\bar{\sigma} := \sqrt{\max_{a\in\mathcal{A}} \sigma_a^2}$. I conjecture that many of the results proved in this appendix can be extended to more general families of random variables, such as the class sub-Exponential or sub-Weibull random variables. For example, Lemma SA-1 can be shown to hold for sub-Exponential random variables using an identical strategy to the one used throughout.

Decision-maker. The interaction between the decision-maker and the environment produces the following collection of random variables

$$\{(A_t, \mathbf{Z}_t)\}_{t\in[T]}, \qquad \mathbf{Z}_j := (R_{A_j,j}, C_{A_j,j}, \mathbf{X}_{A_j,j}^\top)^\top.$$

Each decision-maker is characterized by a policy that maps the history up to round t, $\{(A_\ell, R_{A_\ell,\ell}, C_{A_\ell,\ell}, \mathbf{X}_{A_\ell,\ell}^\top)\}_{\ell\in[t-1]}$, to the space of probability distributions over actions $\Delta(\mathcal{A})$. Denote the space of policies as

$$\Pi := \left\{\pi : \pi = \{\pi_t\}_{t\in[T]}, \pi_t : (\mathcal{A}\times\mathcal{Z})^{t-1} \to \Delta(\mathcal{A})\right\}, \qquad \mathcal{Z} := \mathcal{R}\times\{0,1\}\times\mathcal{X}.$$

I use interchangeably the words "decision-maker", "algorithm", and "policy" when referring to $\pi \in \Pi$.

Protocol 1 below describes the interaction between the decision-maker and the MAB with censoring.

Regret. The *pseudo-regret* of a decision-maker following a policy π in a MAB with missing rewards $\nu \in \mathcal{C}$ is

$$\mathsf{Regret}_T(\pi;\nu) = \sum_{t=1}^{T}(\max_{a\in\mathcal{A}} \theta_a - \mathbb{E}_\nu[R_{A_t,t}]) = T\overline{\theta} - \sum_{t=1}^{T}\theta_{A_t},$$

Protocol 1 Multi-Armed Bandit with Missing Rewards

```
Consider a generic bandit $\nu \in \mathcal{C}$, where $\nu = (\nu_a)_{a \in \mathcal{A}}$
  for $\ell = 1, 2, \dots, T$ do
      Decision-maker chooses $A_\ell = a$ according to some policy $\pi_t$
      Nature samples $(C_{a,\ell}, R_{a,\ell}, \mathbf{X}_{a,\ell}) \sim \nu_a$
      if $C_{a,\ell} = 1$ then
          Decision-maker observes $R_{a,\ell}$
      else
          Decision-maker receives no feedback
      end if
  end for
```

which depends on ν via the average rewards, and it is a random quantity because the $\{A_t\}_{t \in [T]}$ are random. Note that the latter is true even if the policies considered are deterministic. The reason is that A_t depends on $\{\mathbf{Z}_\ell\}_{\ell=1}^{t-1}$ which are random. In what follows, I omit the dependence of the regret on ν and simply write $\mathsf{Regret}_T(\pi)$.

SA1.3 Algorithms

In what follows, I focus on the popular UCB algorithm (Auer et al., 2002) and modifications thereof as the algorithm used by the decision-maker to obtain a policy $\{\pi_t^{\mathsf{UCB}}\}_{t \in [T]}$. The UCB algorithm selects the optimal policy using optimistic estimates of previous rewards. Hereafter, I first describe the classic UCB algorithm and then showcase the novel doubly-robust version proposed in this project, first in its unfeasible (oracle) version and then in its feasible form.

SA1.3.1 Classic UCB Algorithm

Before formalizing the algorithm, it is necessary to introduce some notation:

- The number of times an arm $a \in \mathcal{A}$ has been pulled at the beginning of round $t \in [T]$
$$P_a(t) := \sum_{\ell=1}^{t-1} \mathbb{1}[A_\ell = a].$$
- The number of times the reward R_a has been observed at the beginning of round $t \in [T]$
$$N_a(t) := \sum_{\ell=1}^{t-1} \mathbb{1}[A_\ell = a] C_{a,\ell}.$$
- The (regularized) estimate for the mean reward of action $a \in \mathcal{A}$ at the beginning

of round $t \in [T]$ is

$$\widehat{R}_a^{\mathsf{UCB}}(t) = \frac{1}{N_a(t)+\lambda}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]C_{a,\ell}R_{a,\ell} = \frac{1}{P_a(t)+\lambda}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]\frac{C_{a,\ell}R_{a,\ell}}{\hat{q}_a(\ell)}, \quad (1)$$

where $\hat{q}_a(\ell) := \frac{N_a(\ell)+\lambda}{P_a(\ell)+\lambda}$ and $\lambda > 0$ is a regularization parameter that prevents the estimator from being ill-defined whenever, after initialization, it occurs that $N_a(t) = 0$ for some $a \in \mathcal{A}$ and $t \geq 1$.

- The optimistic mean reward estimate of action $a \in \mathcal{A}$ after $t \in [T]$ rounds is

  $$\widetilde{R}_a^{\mathsf{UCB}}(t,\delta) = \widehat{R}_a^{\mathsf{UCB}}(t) + b_{a,t}^{\mathsf{UCB}}(\delta),$$

  where the "bonus" term $b_{a,t}^{\mathsf{UCB}}(\delta)$ is chosen to make sure that the optimistic mean reward estimate $\widetilde{R}_a^{\mathsf{UCB}}(t,\delta)$ upper bounds the true mean reward θ_a with high probability. In this specific case, I define

  $$b_{a,t}^{\mathsf{UCB}}(\delta) := \frac{\bar{\sigma}}{\underline{q}}\sqrt{\frac{2\ln(2AT/\delta)}{P_a(t)+\lambda}} + \frac{\lambda\overline{K}}{N_a(t)+\lambda},$$

  where $\overline{K}$ is some constant larger than $\overline{\theta}$, $\underline{q} := \inf_{a,t}\hat{q}_a(t)$, and $\delta \in (0,1)$. Intuitively, the first term in $b_{a,t}^{\mathsf{UCB}}(\delta)$ governs the probability with which we want the optimistic estimate to overestimate the mean reward, whereas the second term takes into account the bias induced by the regularization term $\lambda > 0$. Lemma SA-5 formally justifies the particular choice of $b_{a,t}^{\mathsf{UCB}}(\delta)$ described above. Finally, note that the introduction of the regularization parameter $\lambda > 0$ makes $\underline{q}$ bounded away from 0.

The way the UCB algorithm works is straightforward: at round $t \in [T]$, it selects the arm a that has the highest optimistic mean reward estimate. Algorithm 1 below summarizes all the steps needed by the classic UCB algorithm.

Procedure 1 Update Estimators for UCB

```
for $a \in [A]$ do
    $N_a(t+1) \leftarrow N_a(t) + \mathbb{1}[A_t = a]C_{a,t}$
    $P_a(t+1) \leftarrow P_a(t) + \mathbb{1}[A_t = a]$
    $\widehat{R}_a(t+1) \leftarrow \frac{1}{N_a(t+1)+\lambda}\sum_{\ell=1}^{t}\mathbb{1}[A_\ell = a]C_{a,\ell}R_{a,\ell}$
    $b_{a,t+1}^{\mathsf{UCB}}(\delta) \leftarrow \frac{\bar{\sigma}}{\underline{q}_\lambda}\sqrt{\frac{2\ln(2AT/\delta)}{P_a(t+1)+\lambda}} + \frac{\lambda\overline{K}}{N_a(t+1)+\lambda}$
    $\widetilde{R}_a(t+1,\delta) \leftarrow \widehat{R}_a(t+1) + b_{a,t+1}^{\mathsf{UCB}}(\delta)$
end for
```

Algorithm 1 UCB algorithm

```
    Input: $\lambda > 0, \underline{q}_\lambda, \bar{\sigma}, T, \mathcal{A}, \delta, \overline{K}$
    Initialization: pull each arm once, get $\widehat{R}_a^{\mathsf{UCB}}(0)$, set $P_a(0) = 1, N_a(0) = C_{a,0}, \forall\, a \in \mathcal{A}$
    for $t = 1, 2, \dots, T$ do
        pull arm $a_t = \arg\max_{a \in \mathcal{A}} \widetilde{R}_a^{\mathsf{UCB}}(t, \delta)$ and set $\pi_t^{\mathsf{UCB}} = a_t$
        call Update Estimators for UCB (Procedure 1)
    end for
    Output: $\pi^{\mathsf{UCB}} = \{\pi_t^{\mathsf{UCB}}\}_{t \in [T]}$
```

SA1.3.2 Oracle Doubly-Robust UCB Algorithm

Let the true conditional mean reward and probability of rewards not being missing for arm $a \in \mathcal{A}$ as

$$\theta_a^\star(\mathbf{X}_a) := \mathbb{E}_\nu[R_a \mid \mathbf{X}_a], \qquad q_a^\star(\mathbf{X}_a) = \mathbb{E}_\nu[C_a \mid \mathbf{X}_a] \in [\underline{q}_a, 1]$$

almost surely. Throughout, I use interchangeably the terms "probability of rewards not being missing" and "probability of missingness".

The two doubly-robust versions of the classic UCB algorithm – one feasible, one unfeasible – described here differ from the standard one because they rely on alternative mean reward estimators and bonus terms.

The oracle doubly-robust UCB algorithm (ODR-UCB) uses the following mean reward estimator

$$\begin{aligned}\widehat{R}_a^{\mathsf{ODR}}(t) : &= \frac{1}{P_a(t)} \sum_{\ell=1}^{t-1} \mathbb{1}[A_\ell = a] \left(\frac{R_{a,\ell} C_{a,\ell}}{q_a(\mathbf{X}_{a,\ell})} - \frac{\theta_a(\mathbf{X}_{a,\ell})}{q_a(\mathbf{X}_{a,\ell})} \left(C_{a,\ell} - q_a(\mathbf{X}_{a,\ell})\right) \right) \\ &= \frac{1}{P_a(t)} \sum_{\ell=1}^{t-1} \mathbb{1}[A_\ell = a] \left(\frac{C_{a,\ell}(R_{a,\ell} - \theta_a(\mathbf{X}_{a,\ell}))}{q_a(\mathbf{X}_{a,\ell})} + \theta_a(\mathbf{X}_{a,\ell}) \right),\end{aligned} \tag{2}$$

for some known functions $\{\theta_a(\cdot), q_a(\cdot)\}_{a \in \mathcal{A}}$. The "Oracle" part comes from the requirement that $q_a(\cdot)$ and $\theta_a(\cdot)$ being known functions, whilst the "Double-Robust" follows from the fact that only one among $q_a(\cdot) = q_a^\star(\cdot)$ and $\theta_a(\cdot) = \theta_a^\star(\cdot)$ needs to be true to make $\widehat{R}_a^{\mathsf{ODR}}(t)$ a consistent estimator of θ_a (see Lemma SA-10). The optimistic mean reward estimator is defined accordingly as

$$\widetilde{R}_a^{\mathsf{ODR}}(t) = \widehat{R}_a^{\mathsf{ODR}}(t) + b_{a,t}^{\mathsf{ODR}}(\delta), \qquad b_{a,t}^{\mathsf{ODR}}(\delta) := K_{\mathsf{ODR}} \sqrt{\frac{2\ln(2AT/\delta)}{P_a(t)}},$$

where $K_{\mathsf{ODR}} := \frac{\bar{\sigma}}{\underline{q}} + \bar{\sigma}$ and $\underline{q} := \min_a \underline{q}_a$. Algorithm 2 details all the steps needed by ODR-UCB.

Procedure 2 Update Estimators for ODR-UCB

```
for $a \in [A]$ do
    $N_a(t+1) \leftarrow N_a(t) + \mathbb{1}[A_t = a]C_{a,\ell}$
    $P_a(t+1) \leftarrow P_a(t) + \mathbb{1}[A_t = a]$
    Update $q_a$ and $\theta_a$ if required
    $\widehat{R}_a^{\mathsf{ODR}}(t+1) := \frac{1}{P_a(t+1)} \sum_{\ell=1}^{t} \mathbb{1}[A_\ell = a] \left( \frac{C_{a,\ell}(R_{a,\ell} - \theta_a(\mathbf{X}_{a,\ell}))}{q_a(\mathbf{X}_{a,\ell})} + \theta_a(\mathbf{X}_{a,\ell}) \right)$
    $\widetilde{R}_a^{\mathsf{ODR}}(t+1,\delta) \leftarrow \widehat{R}_a^{\mathsf{ODR}}(t+1) + b_{a,t}^{\mathsf{ODR}}(\delta)$
end for
```

Algorithm 2 ODR-UCB algorithm

```
    Input: $\lambda > 0, T, \mathcal{A}, \{q_a(\cdot), \theta_a(\cdot)\}_{a\in\mathcal{A}}$
    Initialization: pull each arm once, get $\widehat{R}_a^{\mathsf{ODR}}(0)$ and set $P_a(0) = 1, N_a(0) = C_{a,0}, \forall\, a \in \mathcal{A}$

for $t = 1, 2, \ldots, T$ do
    pull arm $a_t = \arg\max_{a\in\mathcal{A}} \widetilde{R}_a^{\mathsf{ODR}}(t,\delta)$ and set $\pi_t^{\mathsf{ODR}} = a_t$    ▷ ties are broken randomly
    call *Update Estimators for* ODR-UCB (Procedure 2)
end for
    Output: $\pi^{\mathsf{ODR}} = \{\pi_t^{\mathsf{ODR}}\}_{t\in[T]}$
```

SA1.3.3 Feasible Doubly-Robust UCB Algorithm

The feasible doubly-robust UCB algorithm (DR-UCB) differs from ODR-UCB because it attempts to estimate the true conditional mean reward and probability of missingness for each arm. To grant good properties in terms of regret, such estimation needs to be conducted in appropriate ways, which is formalized in Assumptions SA5(c)-SA5(d).

Once $\hat{\theta}_a$ and $\hat{q}_a$ have been constructed, the following estimator for mean rewards can be obtained as follows:

$$\widehat{R}_a^{\mathsf{DR}}(t) := \frac{1}{P_a(t)} \sum_{\ell=1}^{t-1} \mathbb{1}[A_\ell = a] \left(\frac{C_{a,\ell}(R_{a,\ell} - \hat{\theta}_a(\mathbf{X}_{a,\ell}))}{\hat{q}_a(\mathbf{X}_{a,\ell})} + \hat{\theta}_a(\mathbf{X}_{a,\ell}) \right).$$

The optimistic mean reward estimator for DR-UCB is defined as

$$\begin{aligned}
\widetilde{R}_a^{\mathsf{DR}}(t,\delta) &= \widehat{R}_a^{\mathsf{DR}}(t) + b_{a,t}^{\mathsf{DR}}(\delta), & b_{a,t}^{\mathsf{DR}}(\delta) &= b_{a,t}^{\mathsf{ODR}}(\delta) + b_{a,t}^{[1]}(\delta) + b_{a,t}^{[2]}(\delta) + b_{a,t}^{[3]}(\delta), \\
b_{a,t}^{\mathsf{ODR}}(\delta) &= K_{\mathsf{ODR}} \sqrt{\frac{2\ln(2AT/\delta)}{P_a(t)}}, & b_{a,t}^{[1]}(\delta) &:= \frac{\bar{\sigma}}{\underline{q}^2} \sqrt{\frac{2\ln(2AT/\delta)}{P_a(t)}}\, \mathsf{Err}_t(\hat{q}_a), \\
b_{a,t}^{[2]}(\delta) &:= \frac{1}{\underline{q}} \sqrt{\frac{2\ln(2AT/\delta)}{P_a(t)}}\, \mathsf{Err}_t(\hat{\theta}_a), & b_{a,t}^{[3]}(\delta) &= \mathsf{Err}_t(\hat{\theta}_a)\mathsf{Err}_t(\hat{q}_a),
\end{aligned}$$

where

$$\mathsf{Err}_t(\hat{\theta}_a) := \sqrt{\frac{1}{P_a(t)} \sum_{\ell=1}^{t-1} \mathbb{1}[A_\ell = a](\hat{\theta}_a(\mathbf{X}_{a,\ell}) - \theta_a(\mathbf{X}_{a,\ell}))^2}\,,$$

and

$$\mathsf{Err}_t(\hat{q}_a) := \sqrt{\frac{1}{P_a(t)} \sum_{\ell=1}^{t-1} \mathbb{1}[A_\ell = a](\hat{q}_a(\mathbf{X}_{a,\ell}) - q_a(\mathbf{X}_{a,\ell}))^2}$$

are the sample ℓ_2 estimation errors the mean reward estimator incurs as it relies on the estimated counterparts of $\theta_a(\cdot)$ and $q_a(\cdot)$.

Finally, Algorithm 3 describes in greater detail how the DR-UCB algorithm works.

Procedure 3 Update Estimators for DR-UCB

```
for $a \in [A]$ do
    $N_a(t+1) \leftarrow N_a(t) + \mathbb{1}[A_t = a]C_{a,\ell}$
    $P_a(t+1) \leftarrow P_a(t) + \mathbb{1}[A_t = a]$
    Update $\hat{q}_a$ and $\hat{\theta}_a$ if required
    $\widehat{R}_a^{\mathsf{DR}}(t+1) := \frac{1}{P_a(t+1)} \sum_{\ell=1}^{t} \mathbb{1}[A_\ell = a] \left( \frac{C_{a,\ell}(R_{a,\ell} - \hat{\theta}_a(\mathbf{X}_{a,\ell}))}{\hat{q}_a(\mathbf{X}_{a,\ell})} + \hat{\theta}_a(\mathbf{X}_{a,\ell}) \right)$
    $\widetilde{R}_a^{\mathsf{DR}}(t+1, \delta) \leftarrow \widehat{R}_a^{\mathsf{DR}}(t+1) + b_{a,t}^{\mathsf{DR}}(\delta)$
end for
```

Algorithm 3 DR-UCB algorithm

```
Input: $\lambda > 0, T, \mathcal{A}, \delta, \{\hat{q}_a(\cdot), \hat{\theta}_a(\cdot)\}_{a \in \mathcal{A}}$
Initialization: pull each arm once, get $\widehat{R}_a^{\mathsf{DR}}(0)$ and set $P_a(0) = 1, N_a(0) = C_{a,0}, \forall\, a \in \mathcal{A}$
Nuisances: get estimates $\{\hat{q}_a(\mathbf{X}_{a,0}), \hat{\theta}_a(\mathbf{X}_{a,0})\}_{a \in \mathcal{A}}$ according to Assumption SA5(iii)
for $t = 1, 2, \ldots, T$ do
    pull arm $a_t = \arg\max_{a \in \mathcal{A}} \widetilde{R}_a^{\mathsf{DR}}(t, \delta)$ and set $\pi_t^{\mathsf{DR}} = a_t$
    call Update Estimators for DR-UCB (Procedure 3)
end for
Output: $\pi^{\mathsf{DR}} = \{\pi_t^{\mathsf{DR}}\}_{t \in [T]}$
```

SA1.4 Assumptions

Assumption SA1 is a strong assumption that implies that rewards are missing at random. Sub-gaussianity allows to control the tail behavior of rewards.

Assumption SA1 (Reward-Independent Missingness)**.** *For each action $a \in \mathcal{A}, C_a \perp\!\!\!\perp R_a$ and $R_a \sim \mathsf{sG}(\sigma_a)$.*

Assumption SA2 relaxes the previous assumption and instead assumes the process that causes missing data does not depend on rewards only conditional on a vector of observable variables $\mathbf{X}_a$. On top of that, Assumption SA2 has two other mild requirements: (i) the conditional expectation of each R_a is uniformly bounded over $\mathcal{X}$; (ii) the probability of observing rewards is non-zero ($q_a > 0$).

Assumption SA2 (Model- and Design-based Ignorability)**.** *For each $a \in \mathcal{A}$, either*

$$\mathbb{E}_\nu[R_a \mid \mathbf{X}_a, C_a] = \mathbb{E}_\nu[R_a \mid \mathbf{X}_a] =: \theta_a(\mathbf{X}_a) \qquad a.s. \tag{\textbf{MB}}$$

or

$$\mathbb{E}_\nu[C_a \mid \mathbf{X}_a, R_a] = \mathbb{E}_\nu[C_a \mid \mathbf{X}_a] =: q_a(\mathbf{X}_a) \qquad a.s. \tag{\textbf{DB}}$$

holds. Moreover, $R_a \mid \mathbf{X}_a, \sim \mathsf{sG}(\sigma_a)$ and $q_a(\mathbf{x}) \in [\underline{q}, 1]$, for some constants $0 \leq \overline{K}_\theta < \infty$ and $\underline{q} \in (0, 1]$.

Assumptions SA3 and SA4 are equivalent in all aspects and have been differentiated only for illustration and interpretation purposes. Assumption SA3 requires one between the true conditional probability of missingness, $q_a^\star(\mathbf{x})$, and the true conditional expectation of rewards, $\theta_a^\star(\mathbf{x})$, to be a known function. As mentioned, Assumption SA4 is equivalent, but requires such known functions to be the probability limit of an estimator.

Assumption SA3 (Oracle Double Robustness)**.** *For each $a \in \mathcal{A}, \mathbf{x} \in \mathcal{X}$, either*

$$q_a(\mathbf{x}) = q_a^\star(\mathbf{x}), \tag{a}$$

or

$$\theta_a(\mathbf{x}) = \theta_a^\star(\mathbf{x}), \tag{b}$$

holds for some known functions $q_a : \mathcal{X} \to [\underline{q}, 1]$ and $\theta_a : \mathcal{X} \to \mathcal{R}$.

Assumption SA4 (Double Robustness)**.** *For each $a \in \mathcal{A}, \mathbf{x} \in \mathcal{X}$, either*

$$q_a(\mathbf{x}) = q_a^\star(\mathbf{x}), \tag{a}$$

holds or

$$\theta_a(\mathbf{x}) = \theta_a^\star(\mathbf{x}), \tag{b}$$

where $q_a(\mathbf{x})$ and $\theta_a(\mathbf{x})$ are the probability limits as $T \to \infty$ of the estimators $\hat{q}_a(\cdot)$ and $\hat{\theta}_a(\cdot)$, respectively.

Assumption SA5 disciplines the estimation of the nuisance functions $\{(q_a(\cdot), \theta_a(\cdot), a \in \mathcal{A}\}$. Assumption SA5(b) requires the estimator to bound the estimated probability of missingness away from 0, a typical regularity condition in such problems. To avoid over-fitting biases, Assumption SA5(c) asks the nuisance functions to be estimated in an independent (conditional on $\mathbf{X}_a$) sample. Lastly, SA5(d) controls the estimation error of the nuisance estimators in two ways: (i) the estimation error of each nuisance need to be shrinking in $P_a(t)$; (ii) the product of the estimation errors must decay faster than $1/\sqrt{P_a(t)}$. These conditions make the sampling error dominate the estimation error induced by the fact that $\{(\theta_a(\cdot), q_a(\cdot)), a \in \mathcal{A}\}$ are estimated.

Assumption SA5 (Nuisance Estimation)**.** *For each $a \in \mathcal{A}$, the following are true:*

(a) (double robustness) either $\widetilde{q}_a(\mathbf{x}) = q_a(\mathbf{x})$ or $\widetilde{\theta}_a(\mathbf{x}) = \theta_a(\mathbf{x})$;

(b) (truncation) $\forall\, \mathbf{x} \in \mathcal{X}, \widehat{q}_a(\mathbf{x}) \in [\underline{q}, 1], \underline{q} \in (0, 1]$;

(c) (independence) $(\hat{q}_a(\mathbf{X}_a), \hat{\theta}_a(\mathbf{X}_a)) \perp\!\!\!\perp (R_a, C_a) \mid \mathbf{X}_a$;

(d) (ℓ_2-error rate) there exist rates $\alpha > 1/2, \alpha_q > 0$, and $\alpha_\theta > 0$ such that

$$\mathsf{Err}_t(\hat{q}_a) \lesssim \frac{1}{P_a(t)^{\alpha_q}}, \qquad \mathsf{Err}_t(\hat{\theta}_a) \lesssim \frac{1}{P_a(t)^{\alpha_\theta}}, \qquad \mathsf{Err}_t(\hat{q}_a)\mathsf{Err}_t(\hat{\theta}_a) \lesssim \frac{1}{P_a(t)^{\alpha}}$$

with probability $1 - \delta_\mathfrak{c}, \delta_\mathfrak{c} \in (0, 1)$.

SA1.5 Auxiliary Lemmas

The following lemma helps bound sums involving dependent random variables.

Lemma SA-1 (Freedman's Inequality)**.** *Let $\left\{\left(D_k, \mathcal{F}^k\right)\right\}_{k\geq 1}$ be a martingale difference sequence and let $\{\nu_k\}_{k=1}^n$ be random variables such that ν_k is $\mathcal{F}^{k-1}$-measurable. If $\forall\, \kappa \in \mathbb{R} \setminus \{0\}$*

$$\mathbb{E}\left[e^{\kappa D_k} \mid \mathcal{F}_{k-1}\right] \leq e^{\kappa^2 \nu_k^2/2} \quad a.s.,$$

then

$$\left|\sum_{k=1}^n D_k\right| \leq \sqrt{2\log(2/\delta)\sum_{k=1}^n \nu_k^2}$$

with probability at least $1 - \delta$. Furthermore, if $\sum_{k=1}^n \nu_k^2 \leq V$ a.s., then

$$\left|\sum_{k=1}^n D_k\right| \leq \sqrt{2V\log(2/\delta)}$$

with probability at least $1 - \delta$.

[Proof]

The next lemma shows that the product of a sub-Gaussian random variable and a Bernoulli random variable is still sub-Gaussian despite imposing no structure on the joint behavior of the two variates.

Lemma SA-2. *Let $X \sim \mathsf{sG}(\sigma), \sigma > 0, Y \sim \mathsf{Be}(p), p \in (0,1]$, and $Z := X \cdot Y$. Then $Z \sim \mathsf{sG}(\sigma)$.*

[Proof]

I had a much more slack result showing that $Z \sim \mathsf{sE}(2\sigma, 2\sqrt{2}\sigma)$, where sE denotes a sub-Exponential random variable. This tighter result was suggested by the user VHarisop in the following StackExchange post. Note that nothing is assumed on the joint of X and Y.

The next lemma shows that sub-Gaussianity of $Y \mid X$ is inherited by the fluctuations of $\mathbb{E}[Y \mid X]$ around $\mathbb{E}[Y]$.

Lemma SA-3. *Let $Y \mid X \sim \mathsf{sG}(\sigma), \sigma > 0$, and define $W := \mathbb{E}[Y \mid X] - \mathbb{E}[Y]$. Then, $W \sim \mathsf{sG}(\sigma)$.*

[Proof]

Finally, the next lemma shows some useful properties for the Kullback-Leibler divergence, D_{KL}, between two probability distributions.

Lemma SA-4. *Let P and Q be two probability distributions on $\mathcal{X} \times \mathcal{Y}$ that admit densities p and q, with respect to the Lebesgue measure. Then,*

$$D_{\mathsf{KL}}(P, Q) = D_{\mathsf{KL}}(P_X, Q_X) + \mathbb{E}_{X \sim P_X}[D_{\mathsf{KL}}(P_{Y|X}, Q_{Y|X})].$$

Furthermore, if $X \perp\!\!\!\perp Y$, then

$$D_{\mathsf{KL}}(P, Q) = D_{\mathsf{KL}}(P_X, Q_X) + D_{\mathsf{KL}}(P_Y, Q_Y).$$

[Proof]

SA2 Main Results

In this section, I analyze how the algorithms described in Section SA1.3 perform in the two variations of the MAB with missing rewards problem described in Section SA1.2. In Section SA2.1, I assume that rewards and censoring mechanisms are independent and show that the standard UCB algorithm still achieves nearly-optimal regret. In Section SA2.2, I relax the independence assumption and show that the UCB algorithm has linear regret, whereas ODR-UCB and DR-UCB both possess nearly-optimal regret rates. Specifically, I provide high-probability bounds that hold uniformly over the number of rounds T and a subclass of bandits contained in $\mathcal{C}$.

The probability measure underlying all these statements is a probability measure induced by the interaction between the policy $\pi \in \Pi$ and some bandit $\nu \in \mathcal{C}$. Formally, let $R_t = R_{A_t,t}, C_t = C_{A_t,t}$, and $\mathbf{X}_t = \mathbf{X}_{A_t,t}$. The probability measure considered throughout is the probability distribution associated to the tuple $(A_1, R_1, C_1, \mathbf{X}_1, \ldots, A_T, R_T, C_T, \mathbf{X}_T)$ on the measurable space $(\Omega_T, \mathcal{G}_T)$, where $\Omega_T := (\mathcal{A} \times \mathbb{R} \times \{0,1\} \times \mathcal{X})^{A \cdot T}$ and $\mathcal{G}_T := \mathscr{B}(\Omega_T)$. For more technical details on the construction of the appropriate measures and the underlying probability space, I refer the reader to Chapter 4.4 in Lattimore and Szepesvári (2020) and references therein.

SA2.1 Reward-independent Missingness

In what follows, I provide an upper bound for the regret of UCB using standard arguments. Namely, I consider a "good" event and show that it occurs with high probability in the setting considered throughout. In this spirit, define such an event as

$$\mathcal{G}(\delta_1, \delta_2) = \overline{\mathcal{F}^{\mathsf{UCB}}(\delta_1)} \cap \overline{\mathcal{F}^{\mathsf{MIS}}(\delta_2)},$$

for some $\delta_1, \delta_2 \in (0,1)$, where

$$\mathcal{F}^{\mathsf{UCB}}(\delta) := \left\{ \exists\, a \in \mathcal{A}, t \in [T], \left| \widehat{R}^{\mathsf{UCB}}_a(t) - \theta_a \right| \geq b^{\mathsf{UCB}}_{a,t}(\delta) \right\},$$

$$\mathcal{F}^{\mathsf{MIS}}(\delta) := \left\{ \exists\, a \in \mathcal{A}, t \in [T] : N_a(t) \leq (1-\delta) q_a P_a(t), P_a(t) \geq \underline{T}_a \right\},$$

where $\underline{T}_a := 1 + \frac{24 \ln(T)}{q_a}$. Under the good event: (i) for each action, the optimistic reward estimator always covers the true mean; (ii) the censoring mechanism is not too extreme in terms of percentage deviation from its mean.

The next lemma justifies the particular choice of bonus term for the UCB algorithm, that is

$$b_{a,t}^{\mathsf{UCB}}(\delta) := \frac{\bar{\sigma}}{\underline{q}_{\lambda}} \sqrt{\frac{2\ln(2AT/\delta)}{P_a(t)+\lambda}} + \frac{\lambda \overline{K}}{N_a(t)+\lambda}.$$

The following result shows that the absolute deviation between the mean reward estimator $\widehat{R}_a^{\mathsf{UCB}}(t)$ and the true mean reward θ_a is larger than $b_{a,t}^{\mathsf{UCB}}(\delta AT)$ with probability at most δ.

Lemma SA-5. *Let Assumption SA1 hold, $\delta \in (0,1)$, $a \in \mathcal{A}$, and $t \in [T]$. Then*

$$\left| \widehat{R}_a^{\mathsf{UCB}}(t) - \theta_a \right| \geq b_{a,t}^{\mathsf{UCB}}(\delta AT)$$

with probability at most δ.

[Proof]

Lemma SA-6, using Lemma SA-5 as a building block, quantifies the probability with which $\mathcal{F}^{\mathsf{UCB}}(\delta)$ realizes. Lemma SA-7 serves a similar purpose for $\mathcal{F}^{\mathsf{CEN}}(\delta)$. Lemma SA-8 shows an implication of the event $\overline{\mathcal{F}}^{\mathsf{UCB}}(\delta)$, which turns out to be useful when bounding the regret.

Lemma SA-6. *Let Assumption SA1 hold and $\delta \in (0,1)$. Then*

$$\mathbb{P}[\mathcal{F}^{\mathsf{UCB}}(\delta)] \leq \delta.$$

[Proof]

Lemma SA-7. *Let Assumption SA1 hold, $\delta \in (0,1)$, and $a \in \mathcal{A}$. Then*

$$\mathbb{P}[\mathcal{F}^{\mathsf{MIS}}(\delta)] \leq \frac{2}{\delta^2} T^{1-12\delta^2} A_{\mathtt{cen}},$$

where $A_{\mathtt{cen}} := \sum_{a \in \mathcal{A}} q_a^{-1}$.

[Proof]

Lemma SA-8. *Let Assumption SA1 hold and $\delta \in (0,1)$. With probability at least $1-\delta$ or, equivalently, under the event $\overline{\mathcal{F}^{\mathsf{UCB}}(\delta)}$, it holds that*

$$\forall\, a \in \mathcal{A}, t \in [T], \qquad \widetilde{R}_a^{\mathsf{UCB}}(t,\delta) > \theta_a.$$

[Proof]

Before stating the first theorem, I present Lemma SA-9. This is a technical lemma that plays a crucial role in bounding the reciprocal of the (random) number of times an arm has been pulled and its feedback observed. This quantity needs to be handled because of the second term in $b_{a,t}^{\mathsf{UCB}}(\delta)$ that addresses the presence of regularization bias.

Lemma SA-9. *Let Assumption SA1 hold and $\delta \in (0,1)$. With probability at least $1-\delta$ or, equivalently, under the event $\overline{\mathcal{F}^{\mathsf{MIS}}(\delta)}$, it holds that*

$$\sum_{t=1}^{T} \frac{1}{N_a(t)+\lambda} \leq \frac{A_{\mathsf{cen}}}{1-\delta} \ln\left(\frac{T}{A_{\mathsf{cen}}} + \frac{\lambda}{1-\delta}\right),$$

and

$$\sum_{t=1}^{T} \frac{1}{\sqrt{N_a(t)+\lambda}} \leq \frac{A_{\mathsf{cen}}}{\sqrt{1-\delta}} \sqrt{\frac{T}{A_{\mathsf{cen}}} + \frac{\lambda}{1-\delta}}.$$

[Proof]

I am now able to state and prove the first main result: the standard UCB algorithm has optimal (up to logarithmic factors) regret when the process that causes missing data does not depend on rewards.

Theorem SA-1. *Let Assumption SA1 hold, $\lambda = o(T^{1/2})$, $\delta_1 \in (0,1)$, $\delta_2 = \sqrt{\frac{1+\kappa}{12}}$, and $\kappa > 0$. Then, for any $T \in \mathbb{N}$ and bandit $\nu \in \mathcal{C}_1$*

$$\mathsf{Regret}_T(\pi^{\mathsf{UCB}}) \leq \frac{4\bar{\sigma}}{\underline{q}_{\lambda}} \sqrt{2AT\ln(2AT/\delta_1)} + o(\sqrt{T}) + A\overline{K},$$

with probability at least $1-\delta_1 - O(T^{-\kappa})$.

[Proof]

SA2.2 Reward-dependent Missingness

Now, I relax Assumption SA1 and instead assume that rewards and the censoring mechanisms are independent only conditional on a vector of covariates $\mathbf{X}_a$. Accordingly, in this section, the class of bandit considered is

$$\mathcal{C}_2 := \left\{\nu = (\nu_a)_{a\in\mathcal{A}} : \nu_a^{[C]} = \mathsf{Be}(q_a), q_a \in (0,1], \text{ and Assumption SA2 holds}\right\}.$$

In this more complex setting, the sample average of the observed rewards $\widehat{R}_a^{\mathsf{UCB}}(t)$, is not a consistent estimator of θ_a anymore. On the contrary, $\widehat{R}_2^{\mathsf{ODR}}(t)$ and $\widehat{R}_2^{\mathsf{DR}}(t)$ are consistent estimators for θ_a under Assumption SA3 and Assumptions SA4-SA5, respectively.

Next, I show via an example that the min-max regret of UCB grows linearly with the number of rounds T.

SA2.2.1 Classic UCB Algorithm

It is not hard to find instances of bandits in $\mathcal{C}_2$ that make the regret of the standard UCB algorithm grow linearly with T. For example, suppose that $\mathcal{A} = \{1,2\}, C_a \sim \mathsf{Be}(1/2), a \in \mathcal{A}$, and

$$\begin{cases} R_1 \sim \mathsf{Unif}([0,1/2]), & \text{if } C_1 = 1, \\ R_1 \sim \mathsf{Unif}([1/2,1]), & \text{if } C_1 = 0, \end{cases} \qquad R_2 \sim \mathsf{Unif}([0,3/4]).$$

As t grows large, the probability limit of $\widehat{R}_a^{\mathsf{UCB}}(t)$ is $\mathbb{E}_\nu[R_a \mid C_a = 1]$. Under Assumption SA3, the probability limit of $\widehat{R}_2^{\mathsf{ODR}}(t)$ is θ_a, whereas the same holds for $\widehat{R}_2^{\mathsf{DR}}(t)$ under Assumptions SA4-SA5 (see Lemma SA-10 and Lemma SA-11 for formal arguments). For the second arm, $\mathbb{E}_\nu[R_2 \mid C_2 = 1] = \theta_2$, thus $\widehat{R}_2(t)$ and $\widehat{R}_2^{\mathsf{ODR}}(t)$ share the same probability limit. However, for the first arm $\theta_1 = 1/2 > 1/4 = \mathbb{E}_\nu[R_1 \mid C_1 = 1]$, thus the two mean reward estimators converge to different values. In this example, the optimal arm is $a^\star = 1$ because $\theta_1 = 1/2 > 3/8 = \theta_2$. The ODR-UCB uses the right mean reward estimator and consistently chooses the first arm. On the contrary, the standard UCB algorithm will eventually end up stuck selecting the second arm because $\mathbb{E}[R_1 \mid C_1 = 1] = 1/4 < 3/8 = \mathbb{E}[R_2 \mid C_2 = 1]$.

The example above belongs to the class of bandits $\mathcal{C}_2$, for which the standard UCB algorithm consistently selects a suboptimal arm, leading to regret that grows linearly with T. More generally, the standard UCB algorithm has linear regret in all those bandits in which the censoring is negatively correlated with the rewards, that is smaller rewards are observed with higher probability. When such censoring is not properly addressed, the estimated ranking of actions might be a scrambled version of the true one.

Now, I proceed showing that ODR-UCB and DR-UCB achieve nearly optimal regret rates.

SA2.2.2 Oracle Doubly-Robust UCB Algorithm

The next lemma justifies the particular choice of bonus term for the ODR-UCB algorithm:

$$b_{a,t}^{\mathsf{ODR}}(\delta) := K_{\mathsf{ODR}} \sqrt{\frac{2\ln(2AT/\delta)}{P_a(t)}}.$$

This result will also be useful when illustrating the properties of the feasible version of this algorithm.

Lemma SA-10. *Let Assumptions SA2 and SA3 hold with $\underline{q} > 0$, $\delta \in (0,1)$, $a \in \mathcal{A}$, and $t \in [T]$. Then, with probability at most δ, it holds that*

$$\left|\widehat{R}_a^{\mathsf{ODR}}(t) - \theta_a\right| \geq b_{a,t}^{\mathsf{ODR}}(AT\delta).$$

[Proof]

The following theorem provides a high-probability regret bound for the ODR-UCB algorithm that holds uniformly over any horizon T and for any bandit in the class $\mathcal{C}_2$. Since this result is a special case of Theorem SA-3, I do not present a separate proof. Instead, I refer the reader to the proof of that more general result.

Theorem SA-2. *Let Assumptions SA2 and SA3 hold and $\delta \in (0,1)$. Then, for any horizon $T \in \mathbb{N}$ and bandit $\nu \in \mathcal{C}_2$*

$$\mathsf{Regret}_T(\pi^{\mathsf{ODR}}) \leq \frac{4\bar{\sigma}}{\underline{q}_\lambda}\sqrt{AT\ln(2AT/\delta)} + A\overline{K}$$

with probability $1-\delta$.

[Proof]

The above result claims that, under Assumption SA2, the ODR-UCB algorithm achieves a nearly optimal rate for the worst-case regret.

SA2.2.3 Feasible Doubly-Robust UCB Algorithm

The major drawback of the ODR-UCB algorithm is that it is unfeasible to use in practice. Indeed, Assumption SA3 is particularly stringent as it requires both the conditional probability of censoring $q_a(\cdot)$ and the conditional expected reward function $\theta_a(\cdot)$ to be known for each action $a \in \mathcal{A}$. This assumption can be relaxed by relying on appropriate estimators $\hat{q}_a(\mathbf{x})$ and $\hat{\theta}_a(\mathbf{x})$ (in the sense of Assumption SA5), whose probability limits are denoted as $q_a(\mathbf{x})$ and $\theta_a(\mathbf{x})$, respectively, for each $\mathbf{x} \in \mathcal{X}$. It then suffices to assume that at least one of them is correctly specified, i.e., either $q_a(\mathbf{x}) = q_a^\star(\mathbf{x})$ holds or $\theta_a(\mathbf{x}) = \theta_a^\star(\mathbf{x})$; see Assumption SA4 for a formalization of this concept.

Once appropriate $\hat{\theta}_a$ and $\hat{q}_a$ have been constructed for each $a \in \mathcal{A}$, the feasible doubly-robust mean reward estimator is

$$\widehat{R}_a^{\mathsf{DR}}(t) = \frac{1}{P_a(t)}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]\left(\frac{C_{a,\ell}(R_{a,\ell} - \hat{\theta}_a(\mathbf{X}_{a,\ell}))}{\hat{q}_a(\mathbf{X}_{a,\ell})} + \hat{\theta}_a(\mathbf{X}_{a,\ell})\right).$$

The next lemma shows that the bonus term for the DR-UCB algorithm has been chosen appropriately in the sense that it controls the probability with which $\widehat{R}_a^{\mathsf{DR}}(t)$ deviates from θ_a.

This result is of independent interest as it is the first one that provides high-probability bounds for a doubly-robust estimator under mild assumptions. The strategy of the proof is simple: it first decomposes $\widehat{R}_a^{\mathsf{DR}}(t)$ as the sum of $\widehat{R}_a^{\mathsf{ODR}}(t)$ and three residual terms $R_{a,j}(t), j \in \{1,2,3\}$. Then, it shows that each of the four terms in $b_{a,t}^{\mathsf{DR}}(\delta AT)$ serves a specific role in bounding in probability each of the terms mentioned above. In particular, $b_{a,t}^{\mathsf{ODR}}(\delta AT)$ controls $|\widehat{R}_a^{\mathsf{ODR}}(t) - \theta_a|$ (as already proven in Lemma SA-10), whilst each $b_{a,t}^{[j]}(\delta AT)$ controls $|R_{a,j}(t)$ for $j \in \{1,2,3\}$.

Lemma SA-11. *Let Assumptions SA2, SA4, and SA5 hold, $\delta \in (0,1), a \in \mathcal{A}$, and $t \in [T]$. Then,*

$$\left|\widehat{R}_a^{\mathsf{DR}}(t) - \theta_a\right| \geq b_{a,t}^{\mathsf{DR}}(\delta AT)$$

with probability at most δ.

[Proof]

Define the failure event

$$\mathcal{F}^{\mathsf{DR}}(\delta) := \left\{\exists\, a \in \mathcal{A}, t \in [T], \left|\widehat{R}_a^{\mathsf{DR}}(t) - \theta_a\right| \geq b_{a,t}^{\mathsf{DR}}(\delta)\right\}.$$

When $\mathcal{F}^{\mathsf{DR}}(\delta)$ occurs the optimistic doubly-robust reward estimator $\widetilde{R}_a^{\mathsf{DR}}(t) = \widehat{R}_a^{\mathsf{DR}}(t) + b_{a,t}^{\mathsf{DR}}(t)$ does not cover the true mean reward θ_a. The next lemma shows that such an event occurs with arbitrarily small probability.

Lemma SA-12. *Let Assumptions SA2, SA4, and SA5 hold and $\delta \in (0,1)$. Then,*

$$\mathbb{P}[\mathcal{F}^{\mathsf{DR}}(\delta)] \leq \delta.$$

[Proof]

Similarly to Lemma SA-8, the next lemma shows an implication of the event $\mathcal{F}^{\mathsf{DR}}(\delta)$, which turns out to be useful when bounding the regret.

Lemma SA-13. *Let Assumptions SA2, SA4, and SA5 hold and $\delta \in (0,1)$. With probability at least $1-\delta$ or, equivalently, under the event, $\overline{\mathcal{F}^{\mathsf{DR}}(\delta)}$ it holds that*

$$\forall\, a \in \mathcal{A}, t \in [T], \qquad \widetilde{R}_a^{\mathsf{DR}}(t,\delta) = \widehat{R}_a^{\mathsf{DR}}(t) + b_{a,t}^{\mathsf{DR}}(\delta) \geq \theta_a.$$

[Proof]

Finally, the next theorem shows that the regret of the DR-UCB algorithm is nearly optimal (up to logarithmic factors) and provides an upper bound that holds with high probability uniformly over the horizon T and the class of bandits $\mathcal{C}_2$.

Theorem SA-3. *Let Assumptions SA2, SA4, and SA5 hold with $\delta \in (0,1)$ and $\delta_{\mathfrak{c}} \in (0,1)$. Then, for any horizon $T \in \mathbb{N}$ and bandit $\nu \in \mathcal{C}_2$*

$$\mathsf{Regret}_T(\pi^{\mathsf{DR}}) \leq \frac{4\bar{\sigma}}{\underline{q}}\sqrt{AT\ln(2AT/\delta)} + \widetilde{o}(\sqrt{T}) + A\overline{K}$$

with probability $1-\delta-\delta_{\mathfrak{c}}$.

[Proof]

SA2.3 Lower Bound on Minimax Regret

In this section, I show that the minimax regret

$$\mathsf{Regret}^\star_T(\mathcal{C}_j) := \inf_{\pi\in\Pi}\sup_{\nu\in\mathcal{C}_j} \mathsf{Regret}_T(\pi;\nu), \quad j \in \{1,2\}$$

is lower bounded by a constant times $\sqrt{T}$.

The proof follows standard arguments using Le Cam's two-point method. Specifically, I analyze the regret of an arbitrary policy $\tilde{\pi} \in \Pi$ on two carefully chosen instances $\nu, \nu' \in \mathcal{C}$, and show that

$$\sup_{\tilde{\nu}\in\mathcal{C}} \mathsf{Regret}_T(\tilde{\pi};\tilde{\nu}) \geq \max\{\mathsf{Regret}_T(\tilde{\pi},\nu), \mathsf{Regret}_T(\tilde{\pi},\nu')\} \geq f(T)$$

for some function $f(\cdot)$. Since $\tilde{\pi}$ is arbitrary, this implies

$$\mathsf{Regret}^\star_T(\mathcal{C}) \geq f(T).$$

To lower bound minimax regret over $\mathcal{C}_1$ and $\mathcal{C}_2$, I construct the same lower bound for the Gaussian subclasses

$$\mathcal{C}_1^{\mathsf{gau}} := \{(\nu_a)_{a\in\mathcal{A}} : \nu_a^{[R]} = \mathcal{N}(\theta_a, 1),\ \nu_a^{[C]} = \mathrm{Be}(q_a),\ q_a \in (0,1]\} \subset \mathcal{C}_1$$

and

$$\mathcal{C}_2^{\mathsf{gau}} := \big\{(\nu_a)_{a\in\mathcal{A}} : \nu_a^{[R|X]} = \mathcal{N}(\theta_a(X), 1),\ \nu_a^{[C]} = \mathrm{Be}(q_a),\ q_a \in (0,1], \\ \text{and Assumption SA2 holds}\big\} \subset \mathcal{C}_2.$$

Because the supremum in the minimax regret is taken over a smaller class, the same lower bound extends to $\mathcal{C}_1$ and $\mathcal{C}_2$. The argument mirrors that of Theorem 15.2 in Lattimore and Szepesvári (2020), and the bounds are identical.

Theorem SA-4. *Let $T \in \mathbb{N}, T \geq A - 1$ and consider the classes of bandits $\mathcal{C}_1$ and $\mathcal{C}_2$. Then,*

$$\mathsf{Regret}_T^\star(\mathcal{C}_j) = \inf_{\pi \in \Pi} \sup_{\nu \in \mathcal{C}_j} \mathsf{Regret}_T(\pi; \nu) \geq \frac{\sqrt{T(A-1)}}{16\sqrt{e}}.$$

[Proof]

SA3 Simulations

This section provides more details about the simulation study presented in the main paper.

SA3.1 Setup

In this subsection, I suppress the dependence on a of each quantity and illustrate the data-generating process for a generic action. Let $X_j \overset{\text{iid}}{\sim} \mathsf{N}(0,1), j = 1,\dots,d$, and $u_j \overset{\text{iid}}{\sim} \mathsf{N}(0,\sigma_j^2), j \in \{C,R\}$. Define

$$C = \mathbb{1}\left(\sum_{\ell=1}^{d} X_\ell \beta_\ell + u_C > \tau(q)\right), \qquad R = \theta + \sum_{\ell=1}^{d} X_\ell \beta_\ell + u_R, \qquad \boldsymbol{\beta} := (\beta_1,\dots,\beta_d)^\top \in \mathbb{R}_{++},$$

where $\tau(q) : \mathbb{P}[\sum_{\ell=1}^{d} X_\ell\beta_\ell + u_C > \tau(q)] = q$. Note that

$$R \sim \mathsf{N}(\theta, \sigma_{\boldsymbol{\beta}}^2 + \sigma_R^2), \qquad C \sim \mathsf{Be}(p), \qquad C \perp\!\!\!\perp R \mid \mathbf{X},$$

where $\sigma_{\boldsymbol{\beta}}^2 := \|\boldsymbol{\beta}\|_2^2$. Define $W := \sum_{\ell=1}^{d} X_\ell \sim \mathsf{N}(0,\sigma_{\boldsymbol{\beta}}^2)$ and note that

$$\begin{aligned}
\mathbb{C}\text{ov}(R,C) &= \mathbb{E}[RC] - \mathbb{E}[R]\,\mathbb{E}[C] \\
&= \theta\,\mathbb{E}[C] + \mathbb{E}[WC] + \mathbb{E}[u_R C] - \theta\,\mathbb{E}[C] \\
&= \mathbb{E}[WC] = \mathbb{P}[C=1]\,\mathbb{E}[W \mid C=1] = q \cdot \mathbb{E}[W \mid C=1].
\end{aligned}$$

Define $V := W + u_C$, the quantity above can be rewritten as

$$\mathbb{E}[W \mid C=1] = \mathbb{E}[W \mid V > \tau(q)].$$

Note that

$$\begin{bmatrix} W \\ V \end{bmatrix} \sim \mathsf{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^2 & \sigma_{\boldsymbol{\beta}}^2 \\ \sigma_{\boldsymbol{\beta}}^2 & \sigma_{\boldsymbol{\beta}}^2 + \sigma_C^2 \end{bmatrix}\right), \qquad \rho = \frac{\sigma_{\boldsymbol{\beta}}^2}{\sigma_{\boldsymbol{\beta}}\sqrt{\sigma_{\boldsymbol{\beta}}^2 + \sigma_C^2}} = \frac{\sigma_{\boldsymbol{\beta}}}{\sqrt{\sigma_{\boldsymbol{\beta}}^2 + \sigma_C^2}},$$

thus, using the formulas for bivariate normal random variables,

$$W \mid V \sim \mathsf{N}\left(\rho^2 V, \rho^2 \sigma_C^2\right).$$

Hence,

$$\mathbb{E}[W \mid V > \tau(q)] = \mathbb{E}[\mathbb{E}[W \mid V] \mid V > \tau(q)] = \rho^2\,\mathbb{E}[V \mid V > \tau(q)].$$

Using formulas for the truncated expectation of a normal distribution, one gets

$$\mathbb{E}[V \mid V > \tau(q)] = \theta_V + \sigma_V \frac{\phi\left(\tilde{\tau}(q)\right)}{1-\Phi(\tilde{\tau}(q))} = \sqrt{\sigma_{\boldsymbol{\beta}}^2 + \sigma_C^2}\,\frac{\phi\left(\tilde{\tau}(q)\right)}{q}, \qquad \tilde{\tau}(q) := \frac{\tau(q)}{\sqrt{\sigma_{\boldsymbol{\beta}}^2 + \sigma_C^2}},$$

where the second equality also uses the fact that

$$1 - \Phi(\tilde{\tau}(q)) = 1 - \mathbb{P}[Z \leq \tau(q)/\sigma_V] = \mathbb{P}[V \geq \tau(q)] = q.$$

Therefore

$$\begin{aligned}\mathbb{C}\text{ov}(R, C) &= q \cdot \mathbb{E}[W \mid C = 1] \\ &= q \cdot \rho^2 \cdot \mathbb{E}[V \mid V > \tau(q)] \\ &= \rho^2 \sqrt{\sigma_{\boldsymbol{\beta}}^2 + \sigma_C^2}\,\phi\left(\tilde{\tau}(q)\right),\end{aligned}$$

and so

$$\text{Corr}(R, C) = \rho^2 \sqrt{\frac{\sigma_{\boldsymbol{\beta}}^2 + \sigma_C^2}{(\sigma_{\boldsymbol{\beta}}^2 + \sigma_R^2)q(1-q)}}\,\phi\left(\tilde{\tau}(q)\right) = \rho \frac{\sigma_{\boldsymbol{\beta}}\phi\left(\tilde{\tau}(q)\right)}{\sqrt{(\sigma_{\boldsymbol{\beta}}^2 + \sigma_R^2)q(1-q)}}. \tag{3}$$

The above equation makes it clear that once the unconditional probability of missingness q and the variance of the noise terms u_R and u_C have been specified, it is possible to search for $\boldsymbol{\beta} \in \mathbb{R}_{++}$ such that $\text{Corr}(R, C)$ matches a desired value.

SA3.2 Simulation Design

Throughout, I set $A = 2, T = 5000$, and $d = 1$. For each action and each simulation scenario, I parametrize the data-generating process with $\sigma_{a,R}^2 = 1, \sigma_{a,C}^2 = 2, \theta_1 = 0.5$, and $\theta_2 = 1$. The values of q_a and $\boldsymbol{\beta}$ are scenario-specific and are made explicit in Table SA-2.

Table SA-2: Parametrization of various simulation scenarios.

	Missingness	$\boldsymbol{\beta}$	(θ_1, θ_2)	$(\tilde{\theta}_1, \tilde{\theta}_2)$	(q_1, q_2)
1.	✗	$\mathbf{0}$	$(0.5, 1)$	$(0.5, 1)$	$(1, 1)$
2.	$C \perp\!\!\!\perp R$	$\mathbf{0}$	$(0.5, 1)$	$(0.5, 1)$	$(0.25, 0.9)$
3.	$C \perp\!\!\!\perp R \mid \mathbf{X}$	s.t. $\text{Corr}(C, R) = 0.2$	$(0.5, 1)$	$(1.16, 1.08)$	$(0.25, 0.9)$

More in detail:

1. Scenario 1 is no missing data, thus the data-generating process is that of a standard

multi-armed bandit, which is akin to specifying $\boldsymbol{\beta} = \mathbf{0}$ and $q_1 = q_2 = 1$;

2. Scenario 2 is reward-independent missingness, thus $\boldsymbol{\beta} = \mathbf{0}$ and $q_1, q_2 \in (0, 1)$;
3. Scenario 3 is reward-dependent missingness, thus $\boldsymbol{\beta}$ is selected so that $\mathrm{Corr}(C_a, R_a) = 0.2$ in Equation (3). In this case, the probability limit of the sample average of observed rewards, $\widehat{R}_a(T)$, is biased and different from θ_a.

Finally, the oracle versions of the UCB and DR-UCB algorithms are computed using knowledge of the underlying data-generating process.

More in detail, under scenarios 1 and 2, $\pi_\star^{\mathsf{UCB}}$ uses the following bonus term

$$\check{b}_{a,t}(\delta) = \mathfrak{q}_{1-\delta} \frac{\sigma_{a,R}}{\sqrt{N_a(t) + \lambda}},$$

whereas $\pi_\star^{\mathsf{UCB}}$ under scenario 3 uses the following bonus term

$$\dot{b}_{a,t}(\delta) = \mathfrak{q}_{1-\delta} \sqrt{\frac{\sigma_{a,R} + \|\boldsymbol{\beta}\|_2^2}{P_a(t)}},$$

where $\mathfrak{q}_{1-\delta}$ is the $(1-\delta)$th quantile of a standard normal distribution.

Finally, nuisance estimation is conducted using an auxiliary sample, $\hat{\theta}_a$ are estimated via least squares, and $\hat{q}_a$ using a probit model.

SA4 Proofs

SA4.1 Proof of Lemma SA-1

Proof. Fix $\delta \in (0,1)$. First of all, note that by assumption we get

$$\mathbb{E}\left[e^{\kappa D_k - \kappa^2 \nu_k^2/2} \mid \mathcal{F}^{k-1}\right] \leq 1 \quad \text{a.s..}$$

Let $\kappa \in \mathbb{R}\setminus\{0\}$, then by iteratively applying the law of iterated expectations

$$\mathbb{E}\left[e^{\sum_{k=1}^{n}(\kappa D_k - \kappa^2\nu_k^2/2)}\right] = \mathbb{E}\left[e^{\sum_{k=1}^{n-1}(\kappa D_k - \kappa^2\nu_k^2/2)}\,\mathbb{E}\left[e^{\kappa D_n - \kappa^2\nu_n^2/2} \mid \mathcal{F}_{n-1}\right]\right] \leq \cdots \leq 1.$$

Using Markov's inequality

$$\mathbb{P}\left[e^{\sum_{k=1}^{n}(\kappa D_k - \kappa^2\nu_k^2/2)} \geq 2\delta^{-1}\right] \leq \frac{\mathbb{E}\left[e^{\sum_{k=1}^{n}(\kappa D_k - \kappa^2\nu_k^2/2)}\right]}{2\delta^{-1}} \leq \delta/2.$$

Finally

$$\sum_{k=1}^{n}(\kappa D_k - \kappa^2\nu_k^2/2) \geq \log(2/\delta) \quad \text{w.p. } \delta/2 \qquad \Longrightarrow \qquad \sum_{k=1}^{n} D_k \leq \frac{\kappa}{2}\sum_{k=1}^{n}\nu_k^2 + \frac{1}{\kappa}\log(2/\delta) \quad \text{w.p. } 1-\delta/2.$$

The same logic can be applied to $-D_k$, by first noting that it is still a martingale difference sequence and then absorbing the minus sign into the κ. The two statements together give us

$$\left|\sum_{k=1}^{n} D_k\right| \leq \frac{\kappa}{2}\sum_{k=1}^{n}\nu_k^2 + \frac{1}{\kappa}\log(2/\delta).$$

Minimizing the upper bound over $\kappa \in \mathbb{R}\setminus\{0\}$, we get $\kappa^\star = \sqrt{2\log(2/\delta)(\sum_{k=1}^{n}\nu_k^2)^{-1}}$ and plugging it in the bound yields

$$\left|\sum_{k=1}^{n} D_k\right| \leq \sqrt{2\log(2/\delta)\sum_{k=1}^{n}\nu_k^2}\,,$$

which holds with probability at least $1-\delta$. The last statement of the lemma follows immediately. ■

SA4.2 Proof of Lemma SA-2

Proof. By Proposition 2.5.2, part (iv) in Vershynin (2018) we know that a random variable $Z \sim \mathsf{sG}(\sigma)$ if and only if

$$\mathbb{E}\left[\exp\left(\frac{Z^2}{C\sigma^2}\right)\right] \leq 2,$$

for some $C > 0$. Therefore, in our case

$$\mathbb{E}\left[\exp\left(\frac{Z^2}{C\sigma^2}\right)\right] = \mathbb{E}\left[\exp\left(\frac{X^2\cdot Y^2}{C\sigma^2}\right)\right] \overset{(i)}{\leq} \mathbb{E}\left[\exp\left(\frac{X^2}{C\sigma^2}\right)\right] \overset{(ii)}{\leq} 2,$$

where (i) follows from the fact that $Y^2 = Y \leq 1$ almost surely and $X \sim \mathsf{sG}(\sigma)$. Thus, we conclude that $Z \sim \mathsf{sG}(\sigma)$ which was to be shown. ■

SA4.3 Proof of Lemma SA-3

Proof. First, I prove an auxiliary fact that will turn out to be useful: conditional sub-Gaussianity implies unconditional sub-Gaussianity. As $Y \mid X \sim \mathsf{sG}(\sigma)$, by definition of sub-Gaussianity one gets

$$\forall\, \lambda \in \mathbb{R}, \qquad \mathbb{E}\left[e^{\lambda(Y-\mathbb{E}[Y|X])} \mid X\right] \leq e^{\frac{\lambda^2\sigma^2}{2}} \quad \text{a.s.}$$

It follows that $Y \sim \mathsf{sG}(\sigma)$. To see this, fix $\lambda \in \mathbb{R}$ and note that

$$\begin{aligned}
\mathbb{E}[e^{\lambda(Y-\mathbb{E}[Y])}] &= \mathbb{E}_Y[e^{\lambda(Y-\mathbb{E}_X[\mathbb{E}_Y[Y|X]])}] && \text{(LIE)}\\
&\leq \mathbb{E}_{X,Y}[e^{\lambda(Y-\mathbb{E}_Y[Y|X])}] && \text{(Jensen's inequality)}\\
&= \mathbb{E}_X[\mathbb{E}_{Y|X}[e^{\lambda(Y-\mathbb{E}_Y[Y|X])}] \mid X]\\
&\leq \mathbb{E}_X\left[e^{\frac{\lambda^2\sigma^2}{2}}\right] && (Y \mid X \sim \mathsf{sG}(\sigma))\\
&\leq e^{\frac{\lambda^2\sigma^2}{2}},
\end{aligned}$$

which was claimed.[1]

Now, consider $W := \mathbb{E}[Y \mid X] - \mathbb{E}[Y]$. Note that $\mathbb{E}[W] = 0$ and fix $\lambda \in \mathbb{R}$. Then,

$$\begin{aligned}
\mathbb{E}[e^{\lambda W}] &= \mathbb{E}[e^{\lambda(\mathbb{E}[Y|X]-\mathbb{E}[Y])}]\\
&= \mathbb{E}[e^{\lambda\,\mathbb{E}[Y|X]}]e^{-\lambda\,\mathbb{E}[Y]}\\
&\leq \mathbb{E}[\mathbb{E}[e^{\lambda Y} \mid X]]e^{-\lambda\,\mathbb{E}[Y]} && \text{(Jensen's inequality)}\\
&= \mathbb{E}[e^{\lambda Y}]e^{-\lambda\,\mathbb{E}[Y]}\\
&= \mathbb{E}[e^{\lambda(Y-\mathbb{E}[Y])}]\\
&\leq e^{\frac{\lambda^2\sigma^2}{2}}, && (Y \sim \mathsf{sG}(\sigma))
\end{aligned}$$

where the last line follows because of the fact proven above. Thus, it follows that $W \sim \mathsf{sG}(\sigma)$, which was to be shown. ■

SA4.4 Proof of Lemma SA-4

Proof. The first result follows from the fact that

$$\begin{aligned}
D_{\mathsf{KL}}(P, Q) &= \int_{\mathcal{X}}\int_{\mathcal{Y}} p(x,y)\,\ln\frac{p(x,y)}{q(x,y)}\,\mathrm{d}y\,\mathrm{d}x\\
&= \int_{\mathcal{X}}\int_{\mathcal{Y}} p(x)\,p(y|x)\,\ln\frac{p(x)\,p(y|x)}{q(x)\,q(y|x)}\,\mathrm{d}y\,\mathrm{d}x\\
&= \int_{\mathcal{X}} p(x)\ln\frac{p(x)}{q(x)}\,\mathrm{d}x + \int_{\mathcal{X}}\int_{\mathcal{Y}} p(x)\,p(y|x)\,\ln\frac{p(y|x)}{q(y|x)}\,\mathrm{d}y\,\mathrm{d}x
\end{aligned}$$

[1] Note that in principle one could make σ^2 a random variable that is $\sigma(X)$-measurable and integrable, and the result would also go through.

$$= D_{\mathsf{KL}}(P_X, Q_X) \; + \; \mathbb{E}_{X\sim P_X}\Big[D_{\mathsf{KL}}(P_{Y|X}, Q_{Y|X})\Big].$$

Furthermore, if $X \perp\!\!\!\perp Y$, then $P_{Y|X} = P_Y$ and $Q_{Y|X} = Q_Y$, thus

$$\mathbb{E}_{X\sim P_X}\Big[D_{\mathsf{KL}}(P_{Y|X}, Q_{Y|X})\Big] = D_{\mathsf{KL}}(P_Y, Q_Y),$$

which proves the second fact. ■

SA4.5 Proof of Lemma SA-5

Proof. Fix $a \in \mathcal{A}, t \in [T], \lambda > 0$, and $\delta \in (0,1)$. First of all, for some $\lambda > 0$ note that

$$\begin{aligned}
\widehat{R}_a^{\mathsf{UCB}}(t) - \theta_a &= \frac{1}{N_a(t)+\lambda}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]C_{a,\ell}R_{a,\ell} - \theta_a \\
&= \frac{1}{N_a(t)+\lambda}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]C_{a,\ell}(R_{a,\ell} - \theta_a) - \frac{\lambda\theta_a}{N_a(t)+\lambda}.
\end{aligned}$$

Define the auxiliary event

$$\mathcal{E}_{a,t}(\delta) := \left\{\left|\widehat{R}_a^{\mathsf{UCB}}(t) - \theta_a\right| \geq b_{a,t}^{\mathsf{UCB}}(\delta/AT)\right\}.$$

Via the triangular inequality, we have

$$\begin{aligned}
\mathcal{E}_{a,t}(\delta) &\subseteq \left\{\left|\frac{1}{N_a(t)+\lambda}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]C_{a,\ell}(R_{a,\ell} - \theta_a)\right| + \frac{\lambda\overline{K}}{N_a(t)+\lambda} \geq b_{a,t}^{\mathsf{UCB}}(\delta/AT)\right\} \\
&= \left\{\left|\frac{1}{N_a(t)+\lambda}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]C_{a,\ell}(R_{a,\ell} - \theta_a)\right| + \frac{\lambda\overline{K}}{N_a(t)+\lambda} \geq \frac{\bar{\sigma}}{\underline{q}_\lambda}\sqrt{\frac{2\ln(2/\delta)}{P_a(t)+\lambda}} + \frac{\lambda\overline{K}}{N_a(t)+\lambda}\right\} \\
&= \left\{\left|\frac{1}{N_a(t)+\lambda}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]C_{a,\ell}(R_{a,\ell} - \theta_a)\right| \geq \frac{\bar{\sigma}}{\underline{q}_\lambda}\sqrt{\frac{2\ln(2/\delta)}{P_a(t)+\lambda}}\right\} \\
&\subseteq \left\{\left|\frac{1}{P_a(t)+\lambda}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]C_{a,\ell}(R_{a,\ell} - \theta_a)\right| \geq \bar{\sigma}\sqrt{\frac{2\ln(2/\delta)}{P_a(t)+\lambda}}\right\} \\
&= \left\{\left|\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]C_{a,\ell}(R_{a,\ell} - \theta_a)\right| \geq \bar{\sigma}\sqrt{2\ln(2/\delta)(P_a(t)+\lambda)}\right\}.
\end{aligned}$$

Then, we get

$$\mathbb{P}\left[\mathcal{E}_{a,t}(\delta)\right] \leq \mathbb{P}\left[\left|\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]C_{a,\ell}(R_{a,\ell} - \theta_a)\right| \geq \bar{\sigma}\sqrt{2\ln(2/\delta)(P_a(t)+\lambda)}\right] \leq \delta, \tag{4}$$

where the last inequality follows from Freedman's inequality (Lemma SA-1). To justify the use of such inequality, I show that $\{W_{a,\ell}\}_{\ell=1}^{T}$, where $W_{a,\ell} := \mathbb{1}[A_\ell = a]C_{a,\ell}(R_{a,\ell} - \theta_a)$ is a martingale difference sequence for an appropriately defined filtration. Let such filtration be defined as $\{\mathcal{F}_\ell\}_{\ell=0}^{t}, \mathcal{F}_\ell = \sigma(\{(R_{A_j,j}, C_{A_j,j}), j = 1, \ldots, \ell\})$. It follows by construction that $\{W_{a,\ell}\}_{\ell=1}^{t-1}$ is $\{\mathcal{F}_\ell\}_{\ell=0}^{t-1}$-adapted and integrable. Note that $\mathbb{1}[A_\ell = a]$ is deterministic once we condition on $\mathcal{F}_{\ell-1}$ as the UCB algorithm picks A_ℓ with all the information available at the beginning of round ℓ (see Protocol 1 and Algorithm 1). Therefore, conditional on $\mathcal{F}_{\ell-1}$ either $W_{a,\ell} = 0$ a.s. or $W_{a,\ell} = C_{a,\ell}(R_{a,\ell} - \theta_a)$, hence, whenever $\{A_\ell \neq a\}$ realizes it follows immediately that $\mathbb{E}[W_{a,\ell} \mid \mathcal{F}_{\ell-1}] = 0$, whereas if $\{A_\ell = a\}$ occurs, then

$$\mathbb{E}[W_{a,\ell} \mid \mathcal{F}_{\ell-1}] = \mathbb{E}[C_{a,\ell}(R_{a,\ell} - \theta_a) \mid \mathcal{F}_{\ell-1}] = \mathbb{E}[C_{a,\ell}(R_{a,\ell} - \theta_a)] = \mathbb{E}[C_{a,\ell}]\,\mathbb{E}[(R_{a,\ell} - \theta_a)] = 0,$$

where the second equality follows because $(C_{a,\ell}, R_{a,\ell}) \overset{\text{iid}}{\sim} \nu_a$ and the third equality from Assumption SA1. Finally, we have that

$$\forall\, \kappa \in \mathbb{R}\setminus\{0\},\ \mathbb{E}\left[e^{\kappa W_{a,\ell}} \mid \mathcal{F}^{\ell-1}\right] \leq e^{\kappa^2 \nu_\ell^2/2} \quad \text{a.s.}$$

with

$$\nu_\ell^2 = \sigma_a^2 \mathbb{1}[A_\ell = a] \implies \sum_{\ell=1}^{t-1} \nu_\ell^2 \leq \bar{\sigma}^2 P_a(t) < \bar{\sigma}^2(P_a(t) + \lambda) \quad \text{a.s.,}$$

where the first inequality follows from the fact that: (i) $\mathbb{1}[A_\ell = a]$ is $\mathcal{F}_{\ell-1}$-measurable; (ii) $W_{a,\ell} \mid \mathcal{F}_{\ell-1} \sim \mathsf{sG}(\sigma_a)$ by Lemma SA-2 and the fact that $(R_{a,\ell}, C_{a,\ell}) \overset{\text{iid}}{\sim} \nu_a$; and (iii) for a random variable Z and sigma-algebra $\mathcal{F}$, if $Z \mid \mathcal{F} \sim \mathsf{sG}(\sigma)$, then $bZ \mid \mathcal{F} \sim \mathsf{sG}(|b|\sigma)$ for a random variable b that is $\mathcal{F}$-measurable. The result in (4) follows from Lemma SA-1. ■

SA4.6 Proof of Lemma SA-6

Proof. Fix some $\delta \in (0,1)$ to be chosen later and consider the failure event

$$\mathcal{F}^{\mathsf{UCB}}(\delta) = \left\{\exists\, a \in \mathcal{A}, t \in [T] : \left|\widehat{R}_a^{\mathsf{UCB}}(t) - \theta_a\right| \geq b_{a,t}^{\mathsf{UCB}}(\delta)\right\}.$$

Then

$$\begin{aligned}
\mathbb{P}[\mathcal{F}^{\mathsf{UCB}}(\delta)] &= \mathbb{P}\left[\bigcup_{a\in\mathcal{A}} \bigcup_{t\in[T]} \left\{\left|\widehat{R}_a^{\mathsf{UCB}}(t) - \theta_a\right| \geq b_{a,t}^{\mathsf{UCB}}(\delta)\right\}\right] \\
&\leq \sum_{a\in\mathcal{A}} \sum_{t\in[T]} \mathbb{P}\left[\left|\widehat{R}_a^{\mathsf{UCB}}(t) - \theta_a\right| \geq b_{a,t}^{\mathsf{UCB}}(\delta)\right] && \text{(union bound)} \\
&\leq \sum_{a\in\mathcal{A}} \sum_{t\in[T]} \frac{\delta}{AT} = \delta, && \text{(Lemma SA-5)}
\end{aligned}$$

which was to be shown. ■

SA4.7 Proof of Lemma SA-7

Proof. Fix $a \in \mathcal{A}, t, \kappa_a \in [T]$, and $\delta \in (0,1)$. Then

$$\mathbb{P}\left[\sum_{\ell=1}^{\kappa_a} \mathbb{1}[C_{a,\ell} = 1] \leq (1-\delta) q_a \kappa_a\right] \leq \exp\left\{-\frac{\delta^2 q_a \kappa_a}{2}\right\}, \tag{5}$$

where the inequality follows from the multiplicative version of a multiplicative Chernoff bound.

Pick $\delta \in (0,1)$, then the probability of the missingness event is

$$\begin{aligned}
\mathbb{P}[\mathcal{F}^{\mathsf{MIS}}(\delta)] &= \mathbb{P}\left[\bigcup_{a\in\mathcal{A}}\bigcup_{t\in[T]}\{N_a(t) \leq (1-\delta)q_a P_a(t), P_a(t) \geq \underline{T}_a\}\right] \\
&\leq \sum_{a\in\mathcal{A}}\sum_{t\in[T]} \mathbb{P}\left[\{N_a(t) \leq (1-\delta)q_a P_a(t), P_a(t) \geq \underline{T}_a\}\right] && \text{(union bound)} \\
&\leq \sum_{a\in\mathcal{A}}\sum_{t\in[T]} \mathbb{P}\left[\exists\, \kappa_a \in [T] : \left\{\sum_{\ell=1}^{\kappa_a} \mathbb{1}[C_{a,\ell} = 1] \leq (1-\delta)q_a\kappa_a, \kappa_a \geq \underline{T}_a\right\}\right] && \text{(Assumption SA1)} \\
&\leq \sum_{a\in\mathcal{A}}\sum_{t=1}^{T}\sum_{\kappa_a=\underline{T}_a}^{T} \mathbb{P}\left[\sum_{\ell=1}^{\kappa_a} \mathbb{1}[C_{a,\ell} = 1] \leq (1-\delta)q_a\kappa_a\right] && \text{(union bound)} \\
&\leq \sum_{a\in\mathcal{A}}\sum_{t=1}^{T}\sum_{\kappa_a=\underline{T}_a}^{T} \exp\left\{-\frac{\delta^2 q_a\kappa_a}{2}\right\} && \text{(by (5))} \\
&= T\cdot\sum_{a\in\mathcal{A}}\sum_{\kappa_a=\underline{T}_a}^{T} \exp\left\{-\frac{\delta^2 q_a\kappa_a}{2}\right\}. && (6)
\end{aligned}$$

Now, note that via an integral comparison, one gets

$$\begin{aligned}
\sum_{\kappa_a=\underline{T}_a}^{T} \exp\left\{-\frac{\delta^2 q_a\kappa_a}{2}\right\} &\leq \int_{\underline{T}_a-1}^{T} \exp\left\{-\frac{\delta^2 q_a u}{2}\right\} \mathrm{d}u \\
&= \left[-\frac{2}{\delta^2 q_a}\exp\left\{-\frac{\delta^2 q_a u}{2}\right\}\right]_{\underline{T}_a-1}^{T} \\
&= \left[-\frac{2}{\delta^2 q_a}\exp\left\{-\frac{\delta^2 q_a u}{2}\right\}\right]_{\underline{T}_a-1}^{\kappa_a} && \text{(summands are negative and } \kappa_a \leq T) \\
&\leq \frac{2}{\delta^2 q_a}\exp\left\{-\frac{\delta^2 q_a(\underline{T}_a-1)}{2}\right\} \\
&= \frac{2}{\delta^2 q_a}T^{-12\delta^2}. && \left(\underline{T}_a = 1 + \tfrac{24\ln(T)}{q_a}\right)
\end{aligned}$$

Therefore, using the above result in (6)

$$\mathbb{P}[\mathcal{F}^{\mathsf{MIS}}(\delta)] \leq T\cdot\sum_{a\in\mathcal{A}}\frac{2}{\delta^2 q_a}T^{-12\delta^2} = \frac{2}{\delta^2}T^{1-12\delta^2}A_{\mathsf{cen}},$$

where $A_{\mathsf{cen}} = \sum_{a\in\mathcal{A}} q_a^{-1}$, which was to be shown. ■

SA4.8 Proof of Lemma SA-8

Proof. Fix $\delta \in (0,1)$. Note that

$$\overline{\mathcal{F}^{\mathsf{UCB}}(\delta)} = \left\{\forall\, a \in \mathcal{A}, t \in [T], \left|\widehat{R}_a^{\mathsf{UCB}}(t) - \theta_a\right| < b_{a,t}^{\mathsf{UCB}}(\delta)\right\},$$

thus for all $a \in \mathcal{A}$ and $t \in [T]$ one has

$$\begin{aligned} \widetilde{R}_a^{\mathsf{UCB}}(t,\delta) &= \widehat{R}_a^{\mathsf{UCB}}(t) + b_{a,t}^{\mathsf{UCB}}(\delta) && \text{(definition)} \\ &= \theta_a + \widehat{R}_a^{\mathsf{UCB}}(t) - \theta_a + b_{a,t}^{\mathsf{UCB}}(\delta) \\ &> \theta_a, \end{aligned}$$

where the last inequality follows because under the event $\overline{\mathcal{F}^{\mathsf{UCB}}(\delta)}$ it occurs that

$$\forall\, a \in \mathcal{A}, t \in [T], \qquad -b_{a,t}^{\mathsf{UCB}}(\delta) < \widehat{R}_a^{\mathsf{UCB}}(t) - \theta_a < b_{a,t}^{\mathsf{UCB}}(\delta),$$

where the first inequality implies

$$\widehat{R}_a^{\mathsf{UCB}}(t) - \theta_a + b_{a,t}^{\mathsf{UCB}}(\delta) > 0.$$

■

SA4.9 Proof of Lemma SA-9

Proof. Fix $\delta \in (0,1)$, if $\overline{\mathcal{F}^{\mathsf{MIS}}(\delta)}$ holds then for all $a \in \mathcal{A}, t \in [T]$ we have that $N_a(t) > (1-\delta)q_a P_a(t)$. Hence for each $a \in \mathcal{A}$

$$\begin{aligned} (1-\delta)\sum_{t=1}^{T}\frac{1}{N_a(t)+\lambda} &\leq \sum_{t=1}^{T}\frac{1}{q_a P_a(t) + \frac{\lambda}{(1-\delta)}} && (\overline{\mathcal{F}^{\mathsf{MIS}}(\delta)} \text{ holds}) \\ &= \sum_{a\in\mathcal{A}}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]\frac{1}{q_a\cdot\ell + \frac{\lambda}{(1-\delta)}} && (T = \textstyle\sum_{a\in\mathcal{A}} P_a(T)) \\ &\overset{(i)}{\leq} \sum_{a\in\mathcal{A}}\int_0^{t-1}\mathbb{1}[A_\ell = a]\frac{1}{q_a\cdot u + \frac{\lambda}{(1-\delta)}}\,\mathrm{d}u \\ &\overset{(ii)}{\leq} \sum_{a\in\mathcal{A}}\frac{1}{q_a}\ln\left(q_a P_a(T) + \frac{\lambda}{(1-\delta)}\right), && (7) \end{aligned}$$

where (i) follows from an integral comparison and (ii) by standard computations.

Now the goal is to construct a generic upper bound on $\sum_{a\in\mathcal{A}}\frac{1}{q_a}\ln(q_a P_a(T)+\lambda)$ that holds for any process that causes missing data $\{q_a\}_{a\in\mathcal{A}}$. To do so, one can solve the following constrained optimization problem

$$\max_{\mathbf{x}\in\mathbb{R}^A}\sum_{a\in\mathcal{A}}\frac{1}{q_a}\ln(q_a x_a + \alpha), \qquad \text{s.to} \quad \mathbf{x}\succeq\mathbf{0}_A,\ \mathbf{1}_A^\top\mathbf{x} = T,$$

where $\alpha := \lambda/(1-\delta)$. The problem above is a standard convex problem (see the water-filling problem in Boyd and Vandenberghe, 2004, Example 5.2, p.245) and has the following KKT conditions

$$\begin{gathered} \mathbf{x}^\star \succeq \mathbf{0}_A, \quad \mathbf{1}_A^\top\mathbf{x}^\star = T, \quad \boldsymbol{\mu}^\star \succeq \mathbf{0}_A, \quad \theta_a^\star x_a^\star = 0, \quad a = 1,\ldots,A \\ -1/\left(\alpha + q_a x_a^\star\right) - \theta_a^\star + \nu^\star = 0, \quad a = 1,\ldots,A, \end{gathered}$$

where $\boldsymbol{\mu}^\star$ are the Lagrange multipliers for the inequality constraints and $\nu^\star$ is the multiplier of the equality constraint. The unique solution of this problem is given by

$$x_a^\star = \frac{1}{q_a}\left(\frac{1}{\nu^\star} - \frac{\lambda}{1-\delta}\right), \quad a = 1, \dots, A,$$

where $\nu^\star$ is such that $\mathbf{1}_A^\top \mathbf{x} = T$ and so $\nu^\star = \left(\frac{\lambda}{1-\delta} + \frac{T}{A_{\mathsf{cen}}}\right)^{-1}$ giving us

$$x_a^\star = \frac{T}{q_a A_{\mathsf{cen}}}, \quad a = 1, \dots, A,$$

which yields a maximum value of (7) equal to

$$A_{\mathsf{cen}} \ln\left(\frac{T}{A_{\mathsf{cen}}} + \frac{\lambda}{1-\delta}\right).$$

A similar logic can be used to show that

$$\sum_{t=1}^{T} \frac{1}{\sqrt{N_a(t) + \lambda}} \leq \frac{A_{\mathsf{cen}}}{\sqrt{1-\delta}} \sqrt{\frac{T}{A_{\mathsf{cen}}} + \frac{\lambda}{1-\delta}}.$$

■

SA4.10 Proof of Theorem SA-1

Proof. Define the good event $\mathcal{G}(\delta_1, \delta_2) = \overline{\mathcal{F}^{\mathsf{UCB}}(\delta_1)} \cap \overline{\mathcal{F}^{\mathsf{MIS}}(\delta_2)}$ for some $\delta_1, \delta_2 \in (0,1)$ to be chosen later and consider the regret of the UCB algorithm. Recall that under the UCB policy, the action at round t is chosen as $A_t := \arg\max_{a \in \mathcal{A}} \widetilde{R}_a^{\mathsf{UCB}}(t, \delta)$. Note that

$$\mathsf{Regret}_T(\pi^{\mathsf{UCB}}) = \sum_{t=1}^{T} \left(\overline{\theta} - \theta_{A_t}\right) = \sum_{t=1}^{T} \Delta_t,$$

with $\Delta_t := \overline{\theta} - \theta_{A_t}$ is the sub-optimality gap at time t. Furthermore, let $\Delta_t(\mathcal{E}) := \mathbb{E}_\nu[R_{a^\star,t} - R_{A_t,t} \mid \mathcal{E}]$ denote the sub-optimality gap conditional on the event $\mathcal{E}$.

First of all, via Lemma SA-6, Lemma SA-7, and a union bound we get

$$\mathbb{P}[\mathcal{G}(\delta_1, \delta_2)] \geq 1 - \delta_1 - \frac{2}{\delta_2^2} A_{\mathsf{cen}} T^{1-12\delta_2^2}. \tag{8}$$

Assume $\mathcal{G}(\delta_1, \delta_2)$ holds, then

$$\begin{aligned}
\Delta_t(\mathcal{G}(\delta_1, \delta_2)) &= \overline{\theta} - \theta_{A_t} \\
&\leq \widetilde{R}_{a^\star}(t, \delta_1) - \theta_{A_t} && \text{(Lemma SA-8)} \\
&\leq \widetilde{R}_{A_t}(t, \delta_1) - \theta_{A_t} && \text{(by UCB, } A_t := \arg\max_{a \in \mathcal{A}} \widetilde{R}_a(t, \delta_1)) \\
&= \widehat{R}_{A_t}(t) - \theta_{A_t} + b_{A_t,t}^{\mathsf{UCB}}(\delta_1) && \text{(definition of } \widetilde{R}_a(t, \delta_1)) \\
&\leq 2 b_{A_t,t}^{\mathsf{UCB}}(\delta_1). && (\overline{\mathcal{F}^{\mathsf{UCB}}(\delta_1)} \text{ holds})
\end{aligned}$$

Define $\widetilde{\lambda} := \frac{\lambda}{1-\delta_2}$. Then, using the result above and the fact that $\lambda > 0$

$$\begin{aligned}
\sum_{t=1}^{T} \Delta_t(\mathcal{G}(\delta_1, \delta_2)) &\le \frac{2\bar{\sigma}}{\underline{q}_\lambda} \sqrt{2\ln(2AT/\delta_1)} \sum_{t=1}^{T} \frac{1}{\sqrt{P_{A_t}(t)}} + 2\lambda \overline{K} \sum_{t=1}^{T} \frac{1}{N_{A_t}(t) + \lambda} \\
&\le \frac{2\bar{\sigma}}{\underline{q}_\lambda} \sqrt{2\ln(2AT/\delta_1)} \sum_{t=1}^{T} \frac{1}{\sqrt{P_{A_t}(t)}} + 2A_{\text{cen}} \overline{K} \widetilde{\lambda} \ln\left(\frac{T}{A_{\text{cen}}} + \widetilde{\lambda}\right) && \text{(Lemma SA-9)} \\
&= \frac{2\bar{\sigma}}{\underline{q}_\lambda} \sqrt{2\ln(2AT/\delta_1)} \sum_{a \in \mathcal{A}} \sum_{\ell=1}^{t-1} \mathbb{1}[A_\ell = a] \frac{1}{\sqrt{\ell}} + 2A_{\text{cen}} \overline{K} \widetilde{\lambda} \ln\left(\frac{T}{A_{\text{cen}}} + \widetilde{\lambda}\right) \\
&\le \frac{4\bar{\sigma}}{\underline{q}_\lambda} \sqrt{2\ln(2AT/\delta_1)} \sum_{a \in \mathcal{A}} \sqrt{P_a(T)} + 2A_{\text{cen}} \overline{K} \widetilde{\lambda} \ln\left(\frac{T}{A_{\text{cen}}} + \widetilde{\lambda}\right) && \left(\textstyle\sum_{j=1}^{k} \frac{1}{\sqrt{j}} \le 2\sqrt{k}\right) \\
&\le \frac{4\bar{\sigma}}{\underline{q}_\lambda} \sqrt{2\ln(2AT/\delta_1)} \sqrt{\sum_{a \in \mathcal{A}} 1 \cdot \sum_{a \in \mathcal{A}} P_a(T)} + 2A_{\text{cen}} \overline{K} \widetilde{\lambda} \ln\left(\frac{T}{A_{\text{cen}}} + \widetilde{\lambda}\right) \\
& && \text{(Cauchy-Schwarz)} \\
&\le \frac{4\bar{\sigma}}{\underline{q}_\lambda} \sqrt{2AT\ln(2AT/\delta_1)} + 2A_{\text{cen}} \overline{K} \widetilde{\lambda} \ln\left(\frac{T}{A_{\text{cen}}} + \widetilde{\lambda}\right).
\end{aligned}$$

Therefore, by (8)

$$\text{Regret}_T(\pi^{\text{UCB}}) \le \frac{4\bar{\sigma}}{\underline{q}_\lambda} \sqrt{2AT\ln(2AT/\delta_1)} + 2A_{\text{cen}} \overline{K} \widetilde{\lambda} \ln\left(\frac{T}{A_{\text{cen}}} + \widetilde{\lambda}\right)$$

with probability at least $1 - \delta_1 - \frac{2}{\delta_2^2} A_{\text{cen}} T^{1-12\delta_2^2}$. Note that for $\delta_2 = \sqrt{\frac{1+\kappa}{12}}$, one gets that $\frac{2}{\delta_2^2} A_{\text{cen}} T^{1-12\delta_2^2} = O(T^{-\kappa})$. Therefore, for $\kappa > 0$ and $\lambda = o(T^{1/2})$, it follows tht

$$\text{Regret}_T(\pi^{\text{UCB}}) \le \frac{4\bar{\sigma}}{\underline{q}_\lambda} \sqrt{2AT\ln(2AT/\delta_1)} + o(\sqrt{T}),$$

which was to be shown.

Finally, recall that each arm has been pulled once during the ’burn-in” period; thus, an additional factor of $\overline{K}$ needs to be taken into account. ■

SA4.11 Proof of Lemma SA-10

Proof. Recall that

$$\widehat{R}_a^{\text{ODR}}(t) = \frac{1}{P_a(t)} \sum_{\ell=1}^{t-1} \mathbb{1}[A_\ell = a] \left(\frac{R_{a,\ell} C_{a,\ell}}{q_a(\mathbf{X}_{a,\ell})} - \frac{\theta_a(\mathbf{X}_{a,\ell})}{q_a(\mathbf{X}_{a,\ell})} (C_{a,\ell} - q_a(\mathbf{X}_{a,\ell})) \right).$$

Define the auxiliary event

$$\mathcal{E}_{a,t}(\delta) := \left\{ \left| \widehat{R}_a^{\text{ODR}}(t) - \theta_a \right| \ge b_{a,t}^{\text{ODR}}(\widetilde{\delta}) \right\},$$

where

$$b_{a,t}^{\text{ODR}}(\widetilde{\delta}) = K_{\text{ODR}} \sqrt{\frac{2\ln(2/\delta)}{P_a(t)}}.$$

Note that

$$\widehat{R}_a^{\mathsf{ODR}}(t) = \frac{1}{P_a(t)} \sum_{\ell=1}^{t-1} (W_{a,\ell} + V_{a,\ell}),$$

with

$$W_{a,\ell} := \mathbb{1}[A_\ell = a] \frac{C_{a,\ell}}{q_a(\mathbf{X}_{a,\ell})} (R_{a,\ell} - \theta_a(\mathbf{X}_{a,\ell})), \qquad V_{a,\ell} = \mathbb{1}[A_\ell = a](\theta_a(\mathbf{X}_{a,\ell}) - \theta_a).$$

Using the fact that $\{|X| + |Y| \geq a + b\} \implies \{|X| \geq a\} \cup \{|Y| \geq b\}$ for any two random variables X, Y and $a, b \in \mathbb{R}$, one gets

$$\begin{aligned}
\mathcal{E}_{a,t}(\delta) &= \left\{ \left| \sum_{\ell=1}^{t-1} (W_{a,\ell} + V_{a,\ell}) \right| \geq \frac{\bar{\sigma}}{\underline{q}} \sqrt{2\ln(2/\delta) P_a(t)} + \bar{\sigma} \sqrt{2\ln(2/\delta) P_a(t)} \right\} \\
&= \left\{ \left| \sum_{\ell=1}^{t-1} W_{a,\ell} \right| \geq \frac{\bar{\sigma}}{\underline{q}} \sqrt{2\ln(2/\delta) P_a(t)} \right\} \bigcup \left\{ \left| \sum_{\ell=1}^{t-1} V_{a,\ell} \right| \geq \bar{\sigma} \sqrt{2\ln(2/\delta) P_a(t)} \right\}.
\end{aligned}$$

I now show that both $\{W_{a,\ell}\}_{\ell=1}^{t-1}$ and $\{V_{a,\ell}\}_{\ell=1}^{t-1}$ are martingale difference sequences with respect to appropriately defined filtrations. Define the collections of sigma-algebras $\{\mathcal{F}_\ell\}_{\ell=1}^{t}, \mathcal{F}_\ell = \sigma(\{\mathbf{X}_{a,\ell}, a \in \mathcal{A}\}) \otimes \sigma(\{\mathbf{Z}_j, j = 1, \dots, \ell - 1\})$ and $\{\mathcal{G}_\ell\}_{\ell=1}^{t}, \mathcal{G}_\ell = \sigma(\{\mathbf{Z}_j, j = 1, \dots, \ell - 1\})$, where $\mathbf{Z}_j = \{(R_{a,j}, C_{a,j}, \mathbf{X}_{a,j}), a \in \mathcal{A}\}$. It follows by construction that $\{V_{a,\ell}\}_{\ell=1}^{t-1}$ is $\{\mathcal{G}_\ell\}_{\ell=0}^{t-1}$-adapted and integrable and $\{W_{a,\ell}\}_{\ell=1}^{t-1}$ is $\{\mathcal{F}_\ell\}_{\ell=0}^{t-1}$-adapted and integrable. Note that $\mathbb{1}[A_\ell = a]$ is deterministic conditionally on either $\mathcal{F}_\ell$ or $\mathcal{G}_\ell$. Therefore, conditional on $\mathcal{F}_\ell$ or on $\mathcal{G}_\ell$, either $\{A_\ell \neq a\}$ realizes, and so $W_{a,\ell} = V_{a,\ell} = 0$ almost surely and $\mathbb{E}[V_{a,\ell} \mid \mathcal{G}_\ell] = \mathbb{E}[W_{a,\ell} \mid \mathcal{F}_\ell] = 0$ follow immediately. If instead $\{A_\ell = a\}$ realizes, then $\mathbb{E}[V_{a,\ell} \mid \mathcal{G}_\ell] = \mathbb{E}[V_{a,\ell}] = 0$ by the law of iterated expectations, whereas for $\mathbb{E}[W_{a,\ell|\mathcal{F}_\ell}]$ two cases need to be considered.

First, suppose Assumption SA3(a) holds, i.e., $q_a(\mathbf{x})$ is the conditional probability of missingness. Then

$$\begin{aligned}
\mathbb{E}[W_{a,\ell} \mid \mathcal{F}_\ell] &= \mathbb{E}[\mathbb{1}[A_\ell = a](C_{a,\ell} R_{a,\ell} - \theta_a(\mathbf{X}_{a,\ell})(C_{a,\ell} - q_a(\mathbf{X}_{a,\ell}) - \theta_a q_a(\mathbf{X}_{a,\ell}))/q_a(\mathbf{X}_{a,\ell}) \mid \mathcal{F}_\ell] \\
&= \mathbb{E}[(C_{a,\ell} R_{a,\ell} - \theta_a(\mathbf{X}_{a,\ell})(C_{a,\ell} - q_a(\mathbf{X}_{a,\ell}))/q_a(\mathbf{X}_{a,\ell}) \mid \mathcal{F}_\ell] - \theta_a && (\{A_\ell = a\} \text{ occurs}) \\
&= \mathbb{E}[(C_{a,\ell} R_{a,\ell} - \theta_a(\mathbf{X}_{a,\ell})(C_{a,\ell} - q_a(\mathbf{X}_{a,\ell}))/q_a(\mathbf{X}_{a,\ell})] - \theta_a && ((R_{a,\ell}, C_{a,\ell}, \mathbf{X}_{a,\ell}) \overset{\text{iid}}{\sim} \nu_a) \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{C_{a,\ell} R_{a,\ell}}{q_a(\mathbf{X}_{a,\ell})} \mid \mathbf{X}_{a,\ell} \right] \right] - \theta_a + \mathbb{E}\left[\mathbb{E}\left[\frac{C_{a,\ell} - q_a(\mathbf{X}_{a,\ell})}{q_a(\mathbf{X}_{a,\ell})} \mid \mathbf{X}_{a,\ell} \right] \right] \\
& && (\text{iterated expectations}) \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{C_{a,\ell} R_{a,\ell}}{q_a(\mathbf{X}_{a,\ell})} \mid \mathbf{X}_{a,\ell} \right] \right] - \theta_a && (\text{Assumption SA3(a)}) \\
&= \mathbb{E}[q_a(\mathbf{X}_{a,\ell})^{-1} \, \mathbb{E}[C_{a,\ell} \mid \mathbf{X}_{a,\ell}] \, \mathbb{E}[R_{a,\ell} \mid \mathbf{X}_{a,\ell}]] - \theta_a && (\text{Assumption SA2}) \\
&= \mathbb{E}[\mathbb{E}[R_{a,\ell} \mid \mathbf{X}_{a,\ell}]] - \theta_a = 0. && (\text{Assumption SA3(a)})
\end{aligned}$$

If instead Assumption SA3(b) holds, i.e., $\theta_a(\mathbf{x})$ is the conditional mean reward, then

$$\begin{aligned}
\mathbb{E}[W_{a,\ell} \mid \mathcal{F}_\ell] &= \mathbb{E}[\mathbb{1}[A_\ell = a](C_{a,\ell} R_{a,\ell} - \theta_a(\mathbf{X}_{a,\ell})(C_{a,\ell} - q_a(\mathbf{X}_{a,\ell}) - \theta_a q_a(\mathbf{X}_{a,\ell}))/q_a(\mathbf{X}_{a,\ell}) \mid \mathcal{F}_\ell] \\
&= \mathbb{E}[(C_{a,\ell} R_{a,\ell} - \theta_a(\mathbf{X}_{a,\ell})(C_{a,\ell} - q_a(\mathbf{X}_{a,\ell}))/q_a(\mathbf{X}_{a,\ell}) \mid \mathcal{F}_\ell] - \theta_a && (\{A_\ell = a\} \text{ occurs}) \\
&= \mathbb{E}[(C_{a,\ell} R_{a,\ell} - \theta_a(\mathbf{X}_{a,\ell})(C_{a,\ell} - q_a(\mathbf{X}_{a,\ell}))/q_a(\mathbf{X}_{a,\ell})] - \theta_a && ((R_{a,\ell}, C_{a,\ell}, \mathbf{X}_{a,\ell}) \overset{\text{iid}}{\sim} \nu_a) \\
&= \mathbb{E}[\theta_a(\mathbf{X}_{a,\ell})] - \theta_a + \mathbb{E}\left[q_a(\mathbf{X}_{a,\ell})^{-1} \, \mathbb{E}\left[C_{a,\ell}(R_{a,\ell} - \theta_a(\mathbf{X}_{a,\ell})) \mid \mathbf{X}_{a,\ell} \right] \right] \\
& && (\text{iterated expectations})
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[q_a(\mathbf{X}_{a,\ell})^{-1}\,\mathbb{E}\left[C_{a,\ell}(R_{a,\ell}-\theta_a(\mathbf{X}_{a,\ell})) \mid \mathbf{X}_{a,\ell}\right]\right] && \text{(Assumption SA3(b))}\\
&= \mathbb{E}[q_a(\mathbf{X}_{a,\ell})^{-1}\,\mathbb{E}[C_{a,\ell} \mid \mathbf{X}_{a,\ell}](\mathbb{E}[R_{a,\ell} \mid \mathbf{X}_{a,\ell}] - \theta_a(\mathbf{X}_{a,\ell}))] - \theta_a && \text{(Assumption SA2)}\\
&= 0. && \text{(Assumption SA3(b))}
\end{aligned}
$$

Moreover, it follows that

$$
\forall\,\kappa \in \mathbb{R},\ \mathbb{E}\left[e^{\kappa W_{a,\ell}} \mid \mathcal{F}_\ell\right] \leq e^{\kappa^2\nu_\ell^2/2} \quad \text{a.s.,} \quad \text{with } \nu_\ell^2 = \frac{\sigma_a^2}{\underline{q}^2}\mathbb{1}[A_\ell = a] \implies \sum_{\ell=1}^{t-1}\nu_\ell^2 \leq \frac{\bar{\sigma}^2}{\underline{q}^2}P_a(t) \quad \text{a.s.,}
$$

where the first inequality follows from the fact that: (i) $\mathbb{1}[A_\ell = a]$ is $\{\mathbf{Z}_j, j = 1, \ldots, \ell\}$-measurable; (ii) $\hat{q}_a(\mathbf{X}_{a,\ell}) - q_a(\mathbf{X}_{a,\ell})$ is $\sigma(\{\mathbf{X}_{a,\ell}, a \in \mathcal{A}\}) \otimes \sigma(\mathcal{D})$-measurable; (iii) by Assumption SA2 and Lemma SA-2, $C_{a,\ell}\epsilon_{a,\ell} \mid \mathcal{F}_\ell \sim \mathsf{sG}(\sigma_a)$; and (iv) for a random variable Z and sigma-algebra $\mathcal{F}$, if $Z \mid \mathcal{F} \sim \mathsf{sG}(\sigma)$, then $bZ \mid \mathcal{F} \sim \mathsf{sG}(|b|\sigma)$ for a random variable b that is $\mathcal{F}$-measurable.

With a similar argument and using Lemma SA-3, it also follows that

$$
\forall\,\kappa \in \mathbb{R},\ \mathbb{E}\left[e^{\kappa V_{a,\ell}} \mid \mathcal{G}_\ell\right] \leq e^{\kappa^2\xi_\ell^2/2} \quad \text{a.s.,} \quad \text{with } \xi_\ell^2 = \sigma_a^2\mathbb{1}[A_\ell = a] \implies \sum_{\ell=1}^{t-1}\xi_\ell^2 \leq \bar{\sigma}^2 P_a(t) \quad \text{a.s..}
$$

Put differently, all the requirements of Freedman's inequality (Lemma SA-1) are satisfied, thus for any fixed $a \in \mathcal{A}$ and round $t \in [T]$

$$
\mathbb{P}\left[\mathcal{E}_{a,t}(\delta)\right] \leq \mathbb{P}\left[\left|\sum_{\ell=1}^{t-1} W_{a,\ell}\right| \geq \frac{\bar{\sigma}}{\underline{q}}\sqrt{2\ln(2/\delta)P_a(t)}\right] + \mathbb{P}\left[\left|\sum_{\ell=1}^{t-1} V_{a,\ell}\right| \geq \bar{\sigma}\sqrt{2\ln(2/\delta)P_a(t)}\right] \leq 2\delta,
$$

where the inequality follows from a union bound. Reparametrizing δ yields the desired result. ■

SA4.12 Proof of Theorem SA-2

Proof. The proof of this result is a particular case of the proof of Theorem SA-3, because

$$
b_{a,t}^{\mathsf{DR}}(\delta) = b_{a,t}^{\mathsf{ODR}}(\delta) + b_{a,t}^{[1]}(\delta) + b_{a,t}^{[2]}(\delta).
$$

Thus, one can ignore the construction of the high-probability bounds on $b_{a,t}^{[1]}(\delta)$ and $b_{a,t}^{[2]}(\delta)$ and obtain a proof for this theorem. ■

SA4.13 Proof of Lemma SA-11

Proof. Before getting started with the actual proof, it is useful to think of the data-generating process as

$$
R_a = \theta_a^\star(\mathbf{X}_a) + \epsilon_a, \qquad C_a = q_a^\star(\mathbf{X}_a) + \xi_a, \quad \forall\, a \in \mathcal{A}.
$$

Note that the two equations above are *definitional* and do not impose conditions other than Assumption SA2 on the data-generating process. Furthermore, by Assumption SA2 it follows that $\epsilon_a \mid \mathbf{X}_a \sim \mathsf{sG}(\sigma_a)$

and $\xi_a \mid \mathbf{X}_a \in [-1,1]$ a.s., and so

$$\epsilon_a \mid \mathbf{X}_a \sim \mathsf{sG}(\sigma_a), \qquad \xi_a \mid \mathbf{X}_a \sim \mathsf{sG}(1), \quad \forall\, a \in \mathcal{A}.$$

Fix $a \in \mathcal{A}, t \in [T], \delta \in (0,1)$, and let $\widetilde{\delta} := AT\delta$. Consider the bonus term

$$b^{\mathsf{DR}}_{a,t}(\delta) = b^{\mathsf{ODR}}_{a,t}(\delta) + b^{[1]}_{a,t}(\delta) + b^{[2]}_{a,t}(\delta) + b^{[3]}_{a,t}(\delta),$$

where

$$b^{\mathsf{ODR}}_{a,t}(\delta) = K_{\mathsf{ODR}}\sqrt{\frac{2\ln(2AT/\delta)}{P_a(t)}}, \qquad b^{[1]}_{a,t}(\delta) := \frac{\bar{\sigma}}{\underline{q}^2}\sqrt{\frac{2\ln(2AT/\delta)}{P_a(t)}}\,\mathsf{Err}_t(\hat{q}_a),$$
$$b^{[2]}_{a,t}(\delta) := \frac{1}{\underline{q}}\sqrt{\frac{2\ln(2AT/\delta)}{P_a(t)}}\,\mathsf{Err}_t(\hat{\theta}_a), \qquad b^{[3]}_{a,t}(\delta) = \mathsf{Err}_t(\hat{\theta}_a)\mathsf{Err}_t(\hat{q}_a),$$

with

$$\mathsf{Err}_t(\hat{\theta}_a) := \sqrt{\frac{1}{P_a(t)}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a](\hat{\theta}_a(\mathbf{X}_{a,\ell}) - \theta_a(\mathbf{X}_{a,\ell}))^2}\,,\ \mathsf{Err}_t(\hat{q}_a) := \sqrt{\frac{1}{P_a(t)}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a](\hat{q}_a(\mathbf{X}_{a,\ell}) - q_a(\mathbf{X}_{a,\ell}))^2}\,.$$

Consider the following decomposition:

$$\begin{aligned}
\widehat{R}^{\mathsf{DR}}_a(t) - \theta_a &= \frac{1}{P_a(t)}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]\left(\frac{C_{a,\ell}(R_{a,\ell} - \hat{\theta}_a(\mathbf{X}_{a,\ell}))}{\hat{q}_a(\mathbf{X}_{a,\ell})} + \hat{\theta}_a(\mathbf{X}_{a,\ell})\right) - \theta_a \\
&= \frac{1}{P_a(t)}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]\left(\frac{C_{a,\ell}(R_{a,\ell} - \theta_a(\mathbf{X}_{a,\ell}))}{q_a(\mathbf{X}_{a,\ell})} + \theta_a(\mathbf{X}_{a,\ell})\right) - \theta_a && (:= R_{a,\mathsf{IF}}(t)) \\
&\quad + \frac{1}{P_a(t)}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]\frac{C_{a,\ell}(R_{a,\ell} - \theta_a(\mathbf{X}_{a,\ell}))}{\hat{q}_a(\mathbf{X}_{a,\ell})q_a(\mathbf{X}_{a,\ell})}\left(q_a(\mathbf{X}_{a,\ell}) - \hat{q}_a(\mathbf{X}_{a,\ell})\right) && (:= R_{a,1}(t)) \\
&\quad + \frac{1}{P_a(t)}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]\left(\hat{\theta}_a(\mathbf{X}_{a,\ell}) - \theta_a(\mathbf{X}_{a,\ell})\right)\left(\frac{q_a(\mathbf{X}_{a,\ell}) - C_{a,\ell}}{\hat{q}_a(\mathbf{X}_{a,\ell})}\right). && (:= R_{a,2}(t)) \\
&\quad + \frac{1}{P_a(t)}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]\left(\hat{\theta}_a(\mathbf{X}_{a,\ell}) - \theta_a(\mathbf{X}_{a,\ell})\right)\left(\frac{q_a(\mathbf{X}_{a,\ell}) - \hat{q}_a(\mathbf{X}_{a,\ell})}{\hat{q}_a(\mathbf{X}_{a,\ell})}\right). && (:= R_{a,3}(t))
\end{aligned}$$

First, Assumptions SA4 and SA5(b) ensure all quantities are well defined as q_a and $\hat{q}_a$ are bounded away from zero. Then, note that

$$\begin{aligned}
&\mathbb{P}\left[\left|\widehat{R}^{\mathsf{DR}}_a(t) - \theta_a\right| \geq b^{\mathsf{DR}}_{a,t}(\widetilde{\delta})\right] \\
&\overset{(1)}{\leq} \mathbb{P}\left[|R_{a,\mathsf{IF}}(t)| + |R_{a,1}(t)| + |R_{a,2}(t)| + |R_{a,3}(t)| \geq b^{\mathsf{ODR}}_{a,t}(\widetilde{\delta}) + b^{[1]}_{a,t}(\widetilde{\delta}) + b^{[2]}_{a,t}(\widetilde{\delta}) + b^{[3]}_{a,t}(\widetilde{\delta})\right] \\
&\overset{(2)}{\leq} \mathbb{P}\left[|R_{a,\mathsf{IF}}(t)| \geq b^{\mathsf{ODR}}_{a,t}(\widetilde{\delta})\right] + \mathbb{P}\left[|R_{a,1}(t)| \geq b^{[1]}_{a,t}(\widetilde{\delta})\right] + \mathbb{P}\left[|R_{a,2}(t)| \geq b^{[2]}_{a,t}(\widetilde{\delta})\right] + \mathbb{P}\left[|R_{a,3}(t)| \geq b^{[3]}_{a,t}(\widetilde{\delta})\right],
\end{aligned} \tag{9}$$

where (1) follows from the triangle inequality, (2) follows from the fact that $\{|X| + |Y| \geq a + b\} \implies \{|X| \geq a\} \cup \{|Y| \geq b\}$ for any two random variables X, Y and $a, b \in \mathbb{R}$. Therefore, the goal is to provide bounds for the three terms in (9). Towards this goal, recall that we defined $\epsilon_{a,\ell} := R_{a,\ell} - \theta_a(\mathbf{X}_{a,\ell})$ and

$\xi_{a,\ell} := C_{a,\ell} - q_a(\mathbf{X}_{a,\ell})$, $\xi_a \in [-1, 1]$ a.s., and, by the law of iterated expectations, $\mathbb{E}[\epsilon_{a,\ell}] = 0 = \mathbb{E}[\xi_{a,\ell}]$ for all $\ell \in [T]$ and $a \in \mathcal{A}$.

Bound on $\mathbb{P}[|R_{a,\mathsf{IF}}(t)| \geq b^{[\mathsf{ODR}]}_{a,t}(\widetilde{\delta})]$.

Note that Assumption SA4 is just Assumption SA3 where $q_a(\cdot)$ and $\theta_a(\cdot)$ are the probability limits of the nuisance estimators. Thus, it follows immediately from Lemma SA-10 that

$$\mathbb{P}\left[|R_{a,\mathsf{IF}}(t)| \geq b^{\mathsf{ODR}}_{a,t}(\widetilde{\delta})\right] \leq \delta. \tag{10}$$

Bound on $\mathbb{P}[|R_{a,1}(t)| \geq b^{[1]}_{a,t}(\widetilde{\delta})]$.

Note that

$$\begin{aligned}
\left\{|R_{a,1}(t)| \geq b^{[1]}_{a,t}(\widetilde{\delta})\right\} &= \left\{\left|\frac{1}{P_a(t)}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]\frac{C_{a,\ell}(R_{a,\ell} - \theta_a(\mathbf{X}_{a,\ell}))}{\hat{q}_a(\mathbf{X}_{a,\ell})q_a(\mathbf{X}_{a,\ell})}\left(q_a(\mathbf{X}_{a,\ell}) - \hat{q}_a(\mathbf{X}_{a,\ell})\right)\right| \geq \frac{\bar{\sigma}}{\underline{q}^2}\sqrt{\frac{2\ln(2/\delta)}{P_a(t)}}\mathsf{Err}_t(\hat{q}_a)\right\} \\
&\subseteq \left\{\frac{1}{\underline{q}^2}\left|\frac{1}{P_a(t)}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]C_{a,\ell}\epsilon_{a,\ell}\left(q_a(\mathbf{X}_{a,\ell}) - \hat{q}_a(\mathbf{X}_{a,\ell})\right)\right| \geq \frac{\bar{\sigma}}{\underline{q}^2}\sqrt{\frac{2\ln(2/\delta)}{P_a(t)}}\mathsf{Err}_t(\hat{q}_a)\right\} \\
&= \left\{\left|\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell = a]C_{a,\ell}\epsilon_{a,\ell}\left(q_a(\mathbf{X}_{a,\ell}) - \hat{q}_a(\mathbf{X}_{a,\ell})\right)\right| \geq \bar{\sigma}\sqrt{2\ln(2/\delta)P_a(t)}\mathsf{Err}_t(\hat{q}_a)\right\},
\end{aligned}$$

where the inclusion follows from Assumptions SA4 and SA5(b).

Denote with $\mathcal{D}$ the data (i.e., collection of random variables) used to estimate $\{\hat{q}_a, \hat{\theta}_a\}_{a\in\mathcal{A}}$ such that Assumption SA5(c) is satisfied and define $W_{a,\ell} := \mathbb{1}[A_\ell = a]C_{a,\ell}\epsilon_{a,\ell}(q_a(\mathbf{X}_{a,\ell}) - \hat{q}_a(\mathbf{X}_{a,\ell}))$ and the collection of sigma-algebras $\{\mathcal{F}_\ell\}_{\ell=1}^t, \mathcal{F}_\ell = \sigma(\{\mathbf{X}_{a,\ell}, a \in \mathcal{A}\}) \otimes \sigma(\{\mathbf{Z}_j, j = 1, \ldots, \ell - 1\}) \otimes \sigma(\mathcal{D})$, where $\mathbf{Z}_j = \{(R_{a,j}, C_{a,j}, \mathbf{X}_{a,j}), a \in \mathcal{A}\}$. It follows by construction that $\{W_{a,\ell}\}_{\ell=1}^{t-1}$ is $\{\mathcal{F}_\ell\}_{\ell=1}^{t-1}$-adapted and integrable. Again, if $\{A_\ell \neq a\}$ occurs, then $\mathbb{E}[W_{a,\ell} \mid \mathcal{F}_\ell] = 0$ a.s. follows immediately. If $\{A_\ell = a\}$ realizes, note that

$$\begin{aligned}
\mathbb{E}[W_{a,\ell} \mid \mathcal{F}_{\ell-1}] &= \mathbb{E}[C_{a,\ell}\epsilon_{a,\ell}(q_a(\mathbf{X}_{a,\ell}) - \hat{q}_a(\mathbf{X}_{a,\ell})) \mid \mathcal{F}_\ell] && (\{A_\ell = a\} \text{ occurs}) \\
&= \mathbb{E}[C_{a,\ell}\epsilon_{a,\ell} \mid \mathcal{F}_\ell](q_a(\mathbf{X}_{a,\ell}) - \hat{q}_a(\mathbf{X}_{a,\ell})) && (\text{Assumption SA5(d)}) \\
&= \mathbb{E}[C_{a,\ell}\epsilon_{a,\ell} \mid \mathbf{X}_{a,\ell}](q_a(\mathbf{X}_{a,\ell}) - \hat{q}_a(\mathbf{X}_{a,\ell})) && ((C_{a,\ell}, \epsilon_{a,\ell}) \mid \mathbf{X}_{a,\ell} \text{ are i.i.d.}) \\
&= \mathbb{E}[C_{a,\ell} \mid \mathbf{X}_{a,\ell}]\,\mathbb{E}[\epsilon_{a,\ell} \mid \mathbf{X}_{a,\ell}](q_a(\mathbf{X}_{a,\ell}) - \hat{q}_a(\mathbf{X}_{a,\ell})) && (\text{Assumption SA2}) \\
&= 0. && (\text{definition of } \epsilon_{a,\ell})
\end{aligned}$$

Moreover, for $\kappa \in \mathbb{R}$ we have

$$\forall\,\kappa \in \mathbb{R},\ \mathbb{E}\left[e^{\kappa W_{a,\ell}} \mid \mathcal{F}_\ell\right] \leq e^{\kappa^2\nu_\ell^2/2} \quad \text{a.s.},$$

with

$$\nu_\ell^2 = \bar{\sigma}^2\mathbb{1}[A_\ell = a](\hat{q}_a(\mathbf{X}_{a,\ell}) - q_a(\mathbf{X}_{a,\ell}))^2 \implies \sum_{\ell=1}^{t-1}\nu_\ell^2 \leq \bar{\sigma}^2 P_a(t)\mathsf{Err}_t(\hat{q}_a)^2 \quad \text{a.s.},$$

where the first inequality follows from the fact that: (i) $\mathbb{1}[A_\ell = a]$ is $\sigma(\{\mathbf{Z}_j, j = 1, \ldots, \ell - 1\})$-measurable; (ii) $\hat{q}_a(\mathbf{X}_{a,\ell}) - q_a(\mathbf{X}_{a,\ell})$ is $\sigma(\{\mathbf{X}_{a,\ell}, a \in \mathcal{A}\}) \otimes \sigma(\mathcal{D})$-measurable; (iii) by Assumption SA2 and Lemma SA-2, $C_{a,\ell}\epsilon_{a,\ell} \mid \mathcal{F}_\ell \sim \mathsf{sG}(\sigma_a)$; and (iv) for a random variable Z and sigma-algebra $\mathcal{F}$, if $Z \mid \mathcal{F} \sim \mathsf{sG}(\sigma)$, then $bZ \mid \mathcal{F} \sim \mathsf{sG}(|b|\sigma)$ for a random variable b that is $\mathcal{F}$-measurable.

All the conditions of Freedman's inequality (Lemma SA-1) are satisfied by $\{W_{a,\ell}\}_{\ell=1}^{t}$ and get

$$\mathbb{P}\left[\left|\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell=a]C_{a,\ell}\epsilon_{a,\ell}\left(q_a(\mathbf{X}_{a,\ell})-\hat{q}_a(\mathbf{X}_{a,\ell})\right)\right|\geq\bar{\sigma}\sqrt{2\ln(2/\delta)P_a(t)}\,\mathsf{Err}_t(\hat{q}_a)\right]\leq\delta,$$

which implies

$$\mathbb{P}[|R_{a,1}(t)|\geq b_{a,t}^{[1]}(\widetilde{\delta})]\leq\delta. \tag{11}$$

Bound on $\mathbb{P}[|R_{a,2}(t)|\geq b_{a,t}^{[2]}(\delta)]$.

Regarding $R_{a,2}(t)$, note that

$$\begin{aligned}|R_{a,2}(t)|&=\left|\frac{1}{P_a(t)}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell=a]\left(\hat{\theta}_a(\mathbf{X}_{a,\ell})-\theta_a(\mathbf{X}_{a,\ell})\right)\left(\frac{q_a(\mathbf{X}_{a,\ell})-C_{a,\ell}}{\hat{q}_a(\mathbf{X}_{a,\ell})}\right)\right|\\&\leq\frac{1}{\underline{q}}\left|\frac{1}{P_a(t)}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell=a]\left(\hat{\theta}_a(\mathbf{X}_{a,\ell})-\theta_a(\mathbf{X}_{a,\ell})\right)\xi_{a,\ell}\right|.\end{aligned}$$

Using a symmetric argument to the one used above to bound $R_{a,1}(t)$ in probability, one gets

$$\mathbb{P}[|R_{a,2}(t)|\geq b_{a,t}^{[2]}(\widetilde{\delta})]=\mathbb{P}\left[|R_{a,2}(t)|\geq\frac{1}{\underline{q}}\sqrt{2\ln(2/\delta)}\frac{\mathsf{Err}_t(\hat{\theta}_a)}{P_a(t)}\right]\leq\delta. \tag{12}$$

Bound on $\mathbb{P}[|R_{a,3}(t)|\geq b_{a,t}^{[3]}(\delta)]$.

By the Cauchy-Schwarz inequality and Assumption SA5(b)

$$\begin{aligned}|R_{a,3}(t)|&\leq\sqrt{\frac{1}{P_a(t)}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell=a](\hat{\theta}_a(\mathbf{X}_{a,\ell})-\theta_a(\mathbf{X}_{a,\ell}))^2}\cdot\sqrt{\frac{1}{P_a(t)}\sum_{\ell=1}^{t-1}\mathbb{1}[A_\ell=a](\hat{q}_a(\mathbf{X}_{a,\ell})-q_a(\mathbf{X}_{a,\ell}))^2}\\&=\frac{1}{\underline{q}}\mathsf{Err}_t(\hat{\theta}_a)\mathsf{Err}_t(\hat{q}_a),\end{aligned}$$

almost surely, which yields

$$\mathbb{P}[|R_{a,3}(t)|\geq b_{a,t}^{[3]}(\widetilde{\delta})]\leq\delta. \tag{13}$$

Final bound.

Finally, by using (10), (11), (12), and (13) in (9), it follows that

$$\mathbb{P}\left[\left|\widehat{R}_a^{\mathsf{DR}}(t)-\theta_a\right|\geq b_{a,t}^{\mathsf{DR}}(\widetilde{\delta})\right]\leq 3\delta.$$

Reparametrizing δ yields the desired result. ■

SA4.14 Proof of Lemma SA-12

Proof. Fix some $\delta \in (0,1)$ and consider the failure event

$$\mathcal{F}^{\mathsf{DR}}(\delta) = \left\{ \exists\, a \in \mathcal{A}, t \in [T] : \left| \widehat{R}_a^{\mathsf{DR}}(t) - \theta_a \right| \geq b_{a,t}^{\mathsf{DR}}(\delta) \right\}.$$

Then,

$$\begin{aligned}
\mathbb{P}[\mathcal{F}^{\mathsf{DR}}(\delta)] &= \mathbb{P}\left[\bigcup_{a \in \mathcal{A}} \bigcup_{t \in [T]} \left\{ \left| \widehat{R}_a^{\mathsf{DR}}(t) - \theta_a \right| \geq b_{a,t}^{\mathsf{DR}}(\delta) \right\} \right] \\
&\leq \sum_{a \in \mathcal{A}} \sum_{t \in [T]} \mathbb{P}\left[\left| \widehat{R}_a^{\mathsf{DR}}(t) - \theta_a \right| \geq b_{a,t}^{\mathsf{DR}}(\delta) \right] && \text{(union bound)} \\
&\leq \frac{\delta}{AT} \cdot AT = \delta. && \text{(Lemma SA-11)}
\end{aligned}$$

■

SA4.15 Proof of Lemma SA-13

Proof. Fix $\delta \in (0,1)$. Note that

$$\overline{\mathcal{F}^{\mathsf{DR}}(\delta)} = \left\{ \forall\, a \in \mathcal{A}, t \in [T], \left| \widehat{R}_a^{\mathsf{DR}}(t) - \theta_a \right| \leq b_{a,t}^{\mathsf{DR}}(\delta) \right\},$$

thus for all $a \in \mathcal{A}$ and $t \in [T]$ we have

$$\begin{aligned}
\widetilde{R}_{a,t}^{\mathsf{DR}}(t,\delta) &= \widehat{R}_a^{\mathsf{DR}}(t) + b_{a,t}^{\mathsf{DR}}(\delta) \\
&= \theta_a + \underbrace{\widehat{R}_a^{\mathsf{DR}}(t) - \theta_a + b_{a,t}^{\mathsf{DR}}(\delta)}_{\geq 0} \\
&\geq \theta_a,
\end{aligned}$$

where the last line follows because under $\overline{\mathcal{F}^{\mathsf{DR}}(\delta)}$ we have

$$\forall\, a \in \mathcal{A}, t \in [T], \qquad -b_{a,t}^{\mathsf{DR}}(\delta) \leq \widehat{R}_a^{\mathsf{DR}}(t) - \theta_a \leq b_{a,t}^{\mathsf{DR}}(\delta),$$

where the first inequality gives us

$$\widehat{R}_a^{\mathsf{DR}}(t) - \theta_a + b_{a,t}^{\mathsf{DR}}(\delta) \geq 0.$$

■

SA4.16 Proof of Theorem SA-3

Proof. Define the good event $\mathcal{G}(\delta) = \overline{\mathcal{F}^{\mathsf{DR}}(\delta)}$ for some $\delta \in (0,1)$ to be chosen later. Consider the regret of the DR-UCB algorithm that uses $\widehat{R}_a^{\mathsf{DR}}(t)$ as an estimator for mean rewards. Moreover, recall that under the UCB policy, the action at round t is chosen as $A_t := \arg\max_{a \in \mathcal{A}} \widetilde{R}_a^{\mathsf{DR}}(t, \delta)$. Note that

$$\mathsf{Regret}_T(\pi^{\mathsf{DR}}) = \sum_{t=1}^{T} \left(\overline{\theta} - \theta_{A_t}\right) = \sum_{t=1}^{T} \Delta_t,$$

with $\Delta_t := \overline{\theta} - \theta_{A_t}$ is the sub-optimality gap at time t. Furthermore, let $\Delta_t(\mathcal{E}) := \mathbb{E}_\nu[R_{a^\star,t} - R_{A_t,t} \mid \mathcal{E}]$ denote the sub-optimality gap conditional on the event $\mathcal{E}$.

Suppose $\mathcal{G}(\delta)$ holds. By Lemma SA-12, this occurs with probability at least $1 - \delta$. Then,

$$\begin{aligned}
\Delta_t(\mathcal{G}(\delta)) &= \overline{\theta} - \theta_{A_t} \\
&\leq \widetilde{R}_{a^\star}^{\mathsf{DR}}(t, \delta) - \theta_{A_t} && \text{(Lemma SA-13)} \\
&\leq \widetilde{R}_{A_t}^{\mathsf{DR}}(t, \delta) - \theta_{A_t} && \text{(by DR-UCB, } A_t := \arg\max_{a \in \mathcal{A}} \widetilde{R}_a^{\mathsf{DR}}(t, \delta)) \\
&= \widehat{R}_{A_t}^{\mathsf{DR}}(t) - \theta_{A_t} + b_{A_t,t}^{\mathsf{DR}}(\delta) && \text{(definition of } \widetilde{R}_{A_t}^{\mathsf{DR}}(t, \delta)) \\
&\leq 2b_{A_t,t}^{\mathsf{DR}}(\delta). && \text{(Lemma SA-12)}
\end{aligned}$$

Therefore, decomposing $b_{A_t,t}^{\mathsf{DR}}(\delta)$ and summing over rounds

$$\begin{aligned}
2\sum_{t=1}^{T} b_{A_t,t}^{\mathsf{ODR}}(\delta) &= \frac{2\overline{\sigma}}{\underline{q}} \sqrt{\ln(2AT/\delta)} \sum_{t=1}^{T} \frac{1}{\sqrt{P_{A_t}(t)}}, \\
&= \frac{2\overline{\sigma}}{\underline{q}} \sqrt{2\ln(2AT/\delta)} \sum_{a \in \mathcal{A}} \sum_{\ell=1}^{P_a(T)} \frac{1}{\sqrt{\ell}} \\
&\leq \frac{4\overline{\sigma}}{\underline{q}} \sqrt{2\ln(2AT/\delta)} \sum_{a \in \mathcal{A}} \sqrt{P_a(T)} && \left(\textstyle\sum_{j=1}^{k} \frac{1}{\sqrt{j}} \leq 2\sqrt{k}\right) \\
&\leq \frac{4\overline{\sigma}}{\underline{q}} \sqrt{2\ln(2AT/\delta)} \sqrt{\sum_{a \in \mathcal{A}} 1 \cdot \sum_{a \in \mathcal{A}} P_a(T)} && \text{(Cauchy-Schwarz)} \\
&\leq \frac{4\overline{\sigma}}{\underline{q}} \sqrt{2AT\ln(2AT/\delta)}.
\end{aligned}$$

Moreover, under Assumption SA5(d) with $\delta_{\mathfrak{c}} \in (0, 1)$, it follows that $\mathfrak{c}_q(P_a(t)) \lesssim P_a(t)^{-\alpha_q}$ for some $\alpha_q > 0$. Then,

$$\begin{aligned}
2\sum_{t=1}^{T} b_{A_t,t}^{[1]}(\delta) &= \frac{2\overline{\sigma}}{\underline{q}^2} \sqrt{2\ln(2AT/\delta)} \sum_{t=1}^{T} \sqrt{\frac{\mathsf{Err}_{P_{A_t}(t)}(\hat{q}_{A_t})}{P_{A_t}(t)}} \\
&\leq \frac{2\overline{\sigma}}{\underline{q}^2} \sqrt{2\ln(2AT/\delta)} \sum_{t=1}^{T} \frac{\mathfrak{c}_t^q}{\sqrt{P_{A_t}(t)}} \\
&\lesssim \frac{2\overline{\sigma}}{\underline{q}^2} \sqrt{2\ln(2AT/\delta)} \sum_{t=1}^{T} \frac{1}{P_{A_t}(t)^{1/2+\alpha_q}} \\
&= \tilde{o}\left(\sqrt{T}\right),
\end{aligned}$$

with probability $1 - \delta - \delta_{\mathfrak{c}}$ and where the last line follows from a comparison with $\sum_{t=1}^{T} P_{A_t}(t)^{-1/2}$. Similarly, using Assumption SA5(d) again, for some $\alpha_\theta > 0$ and $\alpha > 1/2$ we get

$$2\sum_{t=1}^{T} b_{A_t,t}^{[2]}(\delta) \leq 2\sum_{t=1}^{T} \mathsf{Err}_t(\hat{\theta}_a)\mathsf{Err}_t(\hat{\theta}_a) + \frac{2}{\underline{q}} \sqrt{2\ln(2AT/\delta)} \sum_{t=1}^{T} \frac{\mathsf{Err}_t(\hat{q}_a)}{\sqrt{P_a(t)}}$$

$$\lesssim T^{1-\alpha} + \frac{2}{\underline{q}}\sqrt{2\ln(2AT/\delta)}\sum_{t=1}^{T}\frac{1}{P_{A_t}(t)^{1/2+\alpha_\theta}} = \widetilde{o}\left(\sqrt{T}\right)$$

with probability $1-\delta-\delta_{\mathfrak{c}}$. Thus, one can conclude that

$$\mathsf{Regret}_T(\pi^{\mathsf{DR}}) = \sum_{t=1}^{T}\Delta_t(\mathcal{G}(\delta)) \leq \frac{4\bar{\sigma}}{\underline{q}}\sqrt{AT\ln(2AT/\delta)} + \widetilde{o}(\sqrt{T})$$

with probability $1-\delta-\delta_{\mathfrak{c}}$. Finally, recall that each arm has been pulled once during the "burn-in" period; thus, an additional factor of $A\overline{\theta}$ needs to be taken into account. ■

SA4.17 Proof of Theorem SA-4

Proof. Fix $T \in \mathbb{N}$ and choose a generic policy $\pi \in \Pi$ and two Gaussian-Bernoulli bandits ν and ν' in $\mathcal{C}_2^{\mathtt{gau}}$, defined as follows:

1. $\forall\, a \in \mathcal{A}, \nu = \{\nu_a\}_{a\in\mathcal{A}}$ is such that $\nu_a^{[R]} = \mathsf{N}(\theta_a, 1)$ with $\theta_a = \Delta\mathbb{1}[a=1]$ and $\nu_a^{[C]} = \mathsf{Be}(q_a), q_a \in [\underline{q}, 1]$;
2. $\forall\, a \in \mathcal{A}, \nu' = \{\nu'_a\}_{a\in\mathcal{A}}$ is such that $\nu_a^{[X]} = \nu_a^{'[X]}, \nu_a^{'[R|X]} = \mathsf{N}(\theta'_a, 1)$ with $\theta'_a = \Delta\mathbb{1}[a=1] + 2\Delta\mathbb{1}[a = i^\star]$, $\nu_a^{'[C|X]} = \mathsf{Be}(q_a)$ and where
$$i^\star := \arg\min_{a\in\mathcal{A}\setminus\{1\}} \mathbb{E}_\nu[P_a(T)].$$

The rationale for choosing these two bandits is that they are sufficiently hard to distinguish from each other, but they induce different strategies. To summarize, the two mean vectors are

$$(\Delta, 0, \ldots, 0) \quad \text{and} \quad (\Delta, 0, \ldots, 0, 2\Delta, 0, \ldots, 0).$$

This strategy exactly matches what an adversarial nature would play, making it the natural benchmark when focusing on minimax regret. Furthermore, by the definition of $i^\star$ and the fact that $P_a(T) \geq 0$ almost surely for each $a \in \mathcal{A}$

$$\sum_{a\in\mathcal{A}}\mathbb{E}_\nu[P_a(T)] = \mathbb{E}_\nu[P_1(T)] + \sum_{a\in\mathcal{A}\setminus\{1\}}\mathbb{E}_\nu[P_a(T)] \geq \sum_{a\in\mathcal{A}\setminus\{1\}}\mathbb{E}_\nu[P_a(T)] \geq (A-1)\,\mathbb{E}_\nu[P_{i^\star}(T)].$$

Then, because $\sum_{a\in\mathcal{A}}\mathbb{E}_\nu[P_a(T)] = T$ it follows that

$$\mathbb{E}_\nu[P_{i^\star}(T)] \leq \frac{T}{A-1}. \tag{14}$$

By the classical decomposition of regret see Lemma 4.2 in Lattimore and Szepesvári (2020), it follows that

$$\mathsf{Regret}_T(\nu) = \sum_{a\in\mathcal{A}}\Delta_a\,\mathbb{E}_\nu[P_a(T)] = \Delta\sum_{a\in\mathcal{A}\setminus\{1\}}\mathbb{E}_\nu[P_a(T)] = \Delta(T - \mathbb{E}_\nu[P_1(T)])$$

and

$$\mathsf{Regret}_T(\nu') = \sum_{a\in\mathcal{A}}\Delta_a\,\mathbb{E}_{\nu'}[P_a(T)] = \Delta\,\mathbb{E}_{\nu'}[P_1(T)] + 2\Delta\sum_{a\in\mathcal{A}\setminus\{1,i^\star\}}\mathbb{E}_{\nu'}[P_a(T)] \geq \Delta\,\mathbb{E}_{\nu'}[P_1(T)].$$

Then, define the event $\mathcal{A} := \{P_1(T) \leq T/2\}$ and note that

$$\mathsf{Regret}_T(\nu) = \Delta\,\mathbb{E}_\nu[T - P_1(T)] \geq \Delta\,\mathbb{E}_\nu[T - P_1(T) \mid \mathcal{A}]\mathbb{P}_\nu[\mathcal{A}] \geq \frac{\Delta T}{2}\mathbb{P}_\nu[P_1(T) \leq T/2]. \tag{15}$$

Similarly,

$$\mathsf{Regret}_T(\nu') = \Delta\,\mathbb{E}_{\nu'}[P_1(T)] \geq \Delta\,\mathbb{E}_{\nu'}[P_1(T) \mid \overline{\mathcal{A}}]\mathbb{P}_{\nu'}[\overline{\mathcal{A}}] \geq \frac{\Delta T}{2}\mathbb{P}_{\nu'}[P_1(T) > T/2]. \tag{16}$$

Thus, it follows that

$$\begin{aligned}
\mathsf{Regret}_T(\nu) + \mathsf{Regret}_T(\nu') &\geq \frac{\Delta T}{2}\left(\mathbb{P}_\nu[P_1(T) \leq T/2] + \mathbb{P}_{\nu'}[P_1(T) > T/2]\right) && \text{((15) and (16))}\\
&\geq \frac{\Delta T}{4}\exp\left\{-D_{\mathsf{KL}}(\mathbb{P}_\nu, \mathbb{P}_{\nu'})\right\} && \text{(Bretagnole-Huber inequality)}\\
&= \frac{\Delta T}{4}\exp\left\{-\sum_{a\in\mathcal{A}}\mathbb{E}_\nu[P_a(T)]D_{\mathsf{KL}}(\nu, \nu')\right\}, && \text{(Lemma 15.1, LS)}
\end{aligned}$$

where LS is short for Lattimore and Szepesvári (2020). Then, by Lemma SA-4 and the fact that ν and ν' differ only in action $i^\star$

$$\mathsf{Regret}_T(\nu) + \mathsf{Regret}_T(\nu') \geq \frac{\Delta T}{4}\exp\left\{-\mathbb{E}_\nu[P_{i^\star}(T)]D_{\mathsf{KL}}(\nu_{i^\star}, \nu'_{i^\star})\right\}$$

In both cases, $\nu, \nu' \in \mathcal{C}_2^{\mathsf{gau}}$, the Kullback-Leibler divergence between $\nu_{i^\star}$ and $\nu'_{i^\star}$ is $2\Delta^2/2$. Hence,

$$\begin{aligned}
\mathsf{Regret}_T(\nu) + \mathsf{Regret}_T(\nu') &\geq \frac{\Delta T}{4}\exp\left\{-\mathbb{E}_\nu[P_{i^\star}(T)]\frac{2\Delta^2}{2}\right\}\\
&\geq \frac{\Delta T}{4}\exp\left\{-\frac{2T\Delta^2}{2(K-1)}\right\} && \text{((14))}\\
&\geq \frac{T}{4}\sqrt{\frac{A-1}{4T}}\,e^{-1/2} && \left(\Delta = \sqrt{\tfrac{A-1}{4T}} \leq \tfrac{1}{2}\right)\\
&= \frac{\sqrt{T(A-1)}}{8\sqrt{e}}.
\end{aligned}$$

Finally, note that

$$\sup_{\tilde{\nu}\in\mathcal{C}_j}\mathsf{Regret}_T(\pi;\tilde{\nu}) \geq \max\{\mathsf{Regret}_T(\pi;\nu); \mathsf{Regret}_T(\pi;\nu')\} \geq \frac{1}{2}\left(\mathsf{Regret}_T(\nu) + \mathsf{Regret}_T(\nu')\right) \geq \frac{\sqrt{T(A-1)}}{16\sqrt{e}}.$$

Because $\pi \in \Pi$ was generically chosen, it follows that

$$\mathsf{Regret}_T^\star(\mathcal{C}_2^{\mathsf{gau}}) = \inf_{\pi\in\Pi}\sup_{\nu\in\mathcal{C}_2^{\mathsf{gau}}}\mathsf{Regret}_T(\pi;\nu) \geq \frac{\sqrt{T(A-1)}}{16\sqrt{e}},$$

which was to be shown. The proof for $\mathcal{C}_1^{\mathsf{gau}}$ is identical. ■

Supplement References

Auer, Peter, Nicolò Cesa-Bianchi, and Paul Fischer, "Finite-Time Analysis of the Multiarmed Bandit Problem," *Machine Learning*, May 2002, *47* (2), 235–256.

Boyd, Stephen P. and Lieven Vandenberghe, *Convex Optimization*, version 29 ed., Cambridge New York Melbourne New Delhi Singapore: Cambridge University Press, 2004.

Lattimore, Tor and Csaba Szepesvári, *Bandit Algorithms*, 1 ed., Cambridge University Press, July 2020.

Vershynin, Roman, *High-Dimensional Probability: An Introduction with Applications in Data Science*, 1 ed., Cambridge University Press, September 2018.