Robust Best-of-Both-Worlds Gap Estimators Based on Importance-Weighted Sampling

Anonymous Authors[1]

Abstract

We present a novel strategy for robust estimation of the gaps in multiarmed bandits that is based on importance-weighted sampling. The strategy is applicable in best-of-both-worlds setting, namely, it can be used in both stochastic and adversarial regime with no need for prior knowledge of the regime. It is based on a pair of estimators, one based on standard importance weighted sampling to upper bound the losses, and another based on importance weighted sampling with implicit exploration to lower bound the losses. We combine the strategy with the EXP3++ algorithm to achieve best-of-both-worlds regret guarantees in the stochastic and adversarial regimes, and in the stochastically constrained adversarial regime. We conjecture that the strategy can be applied more broadly to robust gap estimation in reinforcement learning, which will be studied in future work.

1. Introduction

Best-of-both-worlds algorithms are algorithms that perform well in stochastic, adversarial, and intermediate environments, with no need for prior knowledge about the nature of the environment. The idea and the term were introduced by Bubeck & Slivkins (2012), who studied multiarmed bandits, and have since spread to a broad range of other frameworks, including combinatorial bandits, linear bandits, bandits with graph feedback, bandits with delayed feedback, Markov Decision Processes (MDPs), and many more (Dann et al., 2023; Masoudian et al., 2024; Jin et al., 2023).

There exist two major approaches to deriving best-of-both-worlds algorithms. One is to start with an algorithm for stochastic environments and extend it to a best-of-both-algorithm by constantly monitoring whether the environment satisfies certain stochasticity tests, and if not, perform an irreversible switch into an adversarial operation mode. So far this approach failed to yield any practically applicable algorithms and to generalize beyond the multiarmed bandit setting (Bubeck & Slivkins, 2012; Auer & Chiang, 2016). The second approach is to start with an algorithm for adversarial bandits and to make adjustments (sometimes only in the analysis) to make it also work in stochastic environments. This category can be further subdivided into two. The first subcategory delivers stochastic regret guarantees through direct control of the gaps. This approach was introduced by Seldin & Slivkins (2014), who injected a bit extra exploration into the classical EXP3 algorithm with losses (Bubeck & Cesa-Bianchi, 2012) and obtained the first practically applicable best-of-both-worlds algorithm named EXP3++. The approach was further improved by Seldin & Lugosi (2017) and extended to additional settings, for example, bandits with graph feedback (Rouyer et al., 2022). An advantage of this approach is its intuitiveness and relative simplicity, making it relatively easy to generalize to new problems. A disadvantage is that the regret bounds are slightly suboptimal: the adversarial regret bound of Seldin & Lugosi (2017) is suboptimal by a $\ln K$ factor coming the analysis of EXP3 (where K is the number of arms) and the stochastic regret bound is suboptimal by a $\ln t$ factor coming from the control of the gaps (where t is the game round). The second subcategory is based on a self-bounding analysis introduced by Zimmert & Seldin (2021). This approach, known as Tsallis-INF, is currently the dominant one. It delivers minimax optimal regret guarantees in both the stochastic and adversarial environments (Zimmert & Seldin, 2021; Masoudian & Seldin, 2021; Ito, 2021), it also delivers minimax optimal regret guarantees in intermediate regimes, including stochastically constrained adversarial, and stochastic regime with adversarial corruptions (Zimmert & Seldin, 2021; Masoudian & Seldin, 2021), and it has been extended to a great variety of settings mentioned earlier (Dann et al., 2023; Jin et al., 2023; Masoudian et al., 2024). However, this approach is based solely on analysing properties of the distribution on arms played by the algorithm, and provides no gap estimates. In many practical cases knowledge the gaps could be interesting and valuable, but it is currently unknown whether this information can be

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the Workshop on Foundations of Reinforcement Learning and Control at the International Conference on Machine Learning (ICML). Do not distribute.

extracted from Tsallis-INF. A second disadvantage is that extension to new settings requires handcrafting of potential functions, which is not always intuitive.

Our work focuses on the first subcategory, namely, EXP3++ style approach. The best multiarmed bandits algorithm in this subcategory is the EXP3++ version introduced by Seldin & Lugosi (2017). It achieves $O\left(\sum_{a:\Delta(a)>0}\frac{(\ln t)^2}{\Delta(a)}\right)$ regret in the stochastic regime (where $\Delta(a)$ are the suboptimality gaps) and $O(\sqrt{Kt\ln K})$ regret in the adversarial regime. A disadvantage of the algorithm of Seldin & Lugosi is that its stochastic analysis is based on *plain* (or, in other words, *unweighted*) losses. Therefore, the stochastic regret guarantee applies only in the purely stochastic regime.

We introduce a novel modification of the algorithm, where both the stochastic and the adversarial analysis are based on importance-weighted losses. The modification preserves the same regret bounds in the stochastic and the adversarial regime as the regret bounds of Seldin & Lugosi, but provides an opportunity to achieve improved regret bounds in intermediate regimes, such as stochastically constrained adversarial. Moreover, it provides an explicit high-probability estimate of the gaps, which may be interesting in its own right, in particular if in the future the technique is extended to reinforcement learning, where using importance-weighted estimates is a common practice.

The primary challenge in high-probability gap estimation based on importance-weighted sampling are the high variance and range of importance-weighted samples. Our solution is based on using standard importance-weighted sampling to control loss deviations from above and importance-weighted sampling with implicit exploration (Neu, 2015) to control loss deviations from below. For the first the control is achieved by Bernstein's inequality for martingales, which only requires one-sided boundedness of the losses. For the second the control is achieved using the analysis of implicit exploration by Neu. We emphasize that using the combination of the two estimators is crucial, because due to high range each of the two estimators only allow deviation control in one direction.

In what follows, we start with outlining the problem setting in Section 2, present our gap estimation strategy in Section 3, combine it with the EXP3++ algorithm in Section 4, and finish with a discussion in Section 5. All proofs are deferred to the appendix.

2. Problem Setting

An environment generates a sequence of losses $\ell_1, \ell_2, \dots$, where $\ell_t \in [0,1]^K$. We consider three types of environments. In a *stochastic environment* each entry $\ell_t(a)$ is drawn from a distribution with a fixed expectation, $\mathbb{E}\left[\ell_t(a)\right] = \mu(a)$, independent of t. In an *oblivious adversarial environment* the vectors ℓ_t are generated arbitrarily before the game starts. Since the oblivious setting is the only adversarial setting we consider in the paper, we will simply refer to it as adversarial. In a *stochastically constrained adversarial environment* the vector entries are sampled independently from distributions that maintain the gaps, $\mathbb{E}\left[\ell_t(a) - \ell_t(a')\right] = \tilde{\Delta}_{a,a'}$, but the means are allowed to fluctuate over time. The stochastic environment is a special case of stochastically constrained adversarial environment, where the means do not fluctuate.

The game is played repeatedly, and at each step t the algorithm chooses an action $A_t \in \{1, \dots, K\}$ and observes only the loss of this action $\ell_t(A_t)$ at this time step.

The aim of the algorithm is to minimize the pseudo-regret, which is the difference between its cumulative loss and the cumulative loss of the best action in hindsight, defined as

$$R(t) = \sum_{s=1}^{t}\mathbb{E}[\ell_s(A_s)] - \min_a\left\{\mathbb{E}\left[\sum_{s=1}^{t}\ell_s(a)\right]\right\}.$$

In the oblivious adversarial setting the losses are considered deterministic and the second expectation can be dropped, making the pseudo-regret coincide with the expected regret

$$R(t) = \sum_{s=1}^{t}\mathbb{E}[\ell_s(A_s)] - \min_a\sum_{s=1}^{t}\ell_s(a).$$

In the stochastic regime action a is called optimal if $\mu(a) = \min_{a'}\{\mu(a')\}$. We use a^* to denote an optimal action (there may be more than one). We use $\Delta(a) = \mu(a) - \mu(a^*)$ to denote the suboptimality gap of action a. The definition of regret in the stochastic setting can then be rewritten as

$$R(t) = \sum_{a:\Delta(a)>0}\mathbb{E}[N_t(a)]\Delta(a), \tag{1}$$

where $N_t(a)$ denotes the number of times action a was played in the first t rounds of the game.

In the stochastically constrained adversarial regime we use $a^* \in \arg\min_a \tilde{\Delta}_{a,1}$ to denote an optimal action, and $\Delta(a) = \tilde{\Delta}_{a,a^*}$ the suboptimality gap of action a (Zimmert & Seldin, 2021). If the means do not fluctuate with time, this definition coincides with the definition of the gaps in the stochastic regime. In the stochastically constrained adversarial regime the regret can also be rewritten using equation (1).

3. Robust Gap Estimation

Our gap estimation strategy uses importance weighted losses, and importance weighted losses with implicit exploration. We denote the importance weighted loss of action

a at time t by:

$$\ell_t^{IW}(a) = \frac{\ell_t(a)\mathbb{1}(A_t = a)}{\tilde{p}_t(a)},$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. The importance weighted loss with implicit exploration of action a at time t is denoted by:

$$\ell_t^{IX}(a) = \frac{\ell_t(a)\mathbb{1}(A_t = a)}{\tilde{p}_t(a) + \gamma_t},$$

where γ_t is an implicit exploration parameter to be specified later.

$L_t^{IW}(a) = \sum_{s=1}^t \ell_s^{IW}(a)$ is the cumulative importance weighted loss of action a up to time t and $L_t^{IX}(a) = \sum_{s=1}^t \ell_s^{IX}(a)$ is the cumulative importance weighted loss with implicit exploration of action a up to time t. $L_t(a) = \sum_{s=1}^t \ell_t(a)$ is the true cumulative loss of action a up to time t.

In the following display we present our gap estimation algorithm, which we name Robust Importance Weighted Gap Estimation. The algorithm can be combined with any other algorithm (e.g., EXP3++) at the plug-in point marked in blue.

Algorithm 1 Robust Importance Weighted Gap Estimation

Remark: see text for definition of $\xi_t(a)$, and $\gamma_t(a)$
$\forall a : L_0^{IW}(a) = L_0^{IX}(a) = 0$
For t=1, 2,...
$\forall a : \tilde{\Delta}_t(a) = \left(L_{t-1}^{IX}(a) - \frac{\ln(4t)}{\gamma_{t-1}} - \min_a\left(L_{t-1}^{IW}(a) + \sqrt{2\nu_{t-1}(a)\ln(4t)} + \frac{\ln(4t)}{3}\right)\right)/(t-1)$
$\forall a : \hat{\Delta}_t(a) = \max\left(0, \tilde{\Delta}_t(a)\right)$
$\forall a : \epsilon_t(a) = \min\left\{\frac{1}{2K}, \frac{1}{2}\sqrt{\frac{\ln K}{tK}}, \xi_t(a)\right\}$
Let $p_t(a)$ be any distribution over $\{1, ..., K\}$ (plug in point for other algorithms).
$\forall a : \tilde{p}_t(a) = \epsilon_t(a) + (1 - \sum_{a'} \epsilon_t(a'))p_t(a)$ Draw action A_t according to $\tilde{p}_t(a)$ and play it
Observe and suffer the loss $\ell_t^{A_t}$
$\forall a : \ell_t^{IW}(a) = \frac{\ell_t(a)\mathbb{1}(A_t=a)}{\tilde{p}_t(a)}$
$\forall a : \ell_t^{IX}(a) = \frac{\ell_t(a)\mathbb{1}(A_t=a)}{\tilde{p}_t(a)+\gamma_t(a)}$
$\forall a : L_t^{IW}(a) = L_{t-1}^{IW}(a) + \ell_t^{IW}(a)$
$\forall a : L_t^{IX}(a) = L_{t-1}(a)^{IX} + \ell_t^{IX}(a)$
$\forall a : \nu_t(a) = \nu_{t-1}(a) + \epsilon_t(a)^{-1}$

The following proposition states the main property of the gap estimation algorithm, namely, that with an appropriate set of parameters it ensures that $\frac{1}{2}\Delta(a) \leq \hat{\Delta}_t(a) \leq \Delta$ with high probability for all sufficiently large t. Thus, $\hat{\Delta}_t(a)$ can be used as a reliable estimate of $\Delta(a)$ for any higher level purpose.

Proposition 3.1. *For $\gamma_t = \frac{\epsilon_t(a)\hat{\Delta}_t(a)}{\sqrt{1200}}$, and any a and t, the gap estimates $\hat{\Delta}_t(a)$ of Algorithm 1 in the stochastic regime satisfy:*

$$\mathbb{P}(\hat{\Delta}_t(a) \geq \Delta(a)) \leq \frac{1}{2t}. \quad (2)$$

Furthermore, for any choice of $\xi_t(a)$, such that $\xi_t(a) \geq \frac{1200\ln t}{t\hat{\Delta}_t(a)^2}$ and $t \geq t_{\min}(a)$, the gap estimates satisfy:

$$\mathbb{P}\left(\hat{\Delta}_t(a) \leq \frac{\Delta(a)}{2}\right) \leq \frac{1}{2t}, \quad (3)$$

where $t_{\min}(a) = \min_t \left\{t \geq \frac{4\cdot 1200(\ln t)^2 K}{\Delta(a)^4 \ln K}\right\}$ is the first time when $\frac{1200ln(t)}{t\Delta(a)^2} \leq \frac{1}{2}\sqrt{\frac{\ln K}{tK}}$.

A proof of this proposition is provided in Appendix B.

4. EXP3++ with Robust Importance Weighted Gap Estimation

In the following display we cite the EXP3++ algorithm of Seldin & Slivkins (2014).

Algorithm 2 EXP3++

Remark: see text for definition of η_t and $\xi_t(a)$
$\forall a : L_0^{IW}(a) = 0$
For $t = 1, 2, \ldots$
$\forall a : \epsilon_t(a) = \min\left\{\frac{1}{2K}, \frac{1}{2}\sqrt{\frac{\ln K}{tK}}, \xi_t(a)\right\}$
$\forall a : p_t(a) = e^{-\eta_t L_{t-1}^{IW}(a)} / \sum_{a'} e^{-\eta_t L_{t-1}^{IW}(a')}$
$\forall a : \tilde{p}_t(a) = \epsilon_t(a) + (1 - \sum_{a'} \epsilon_t(a'))p_t(a)$
Draw action A_t according to $\tilde{p}_t(a)$ and play it
Observe and suffer the loss $\ell_t(A_t)$
$\forall a : \ell_t^{IW}(a) = \frac{\ell_t(A_t)\mathbb{1}(A_t=a)}{\tilde{p}_t(a)}$
$\forall a : L_t^{IW}(a) = L_{t-1}^{IW}(a) + \ell_t^{IW}(a)$

We combine EXP3++ with our robust gap estimation by plugging the exploration parameters $\epsilon_t(a)$ from Algorithm 1 into EXP3++. The matching lines are highlighted in violet and the plug-in point in blue. Note that importance weigthed samples with implicit exploration are not used by EXP3++ and have no impact on its operation, they are only used within Algorithm 1.

We prove the following regret guarantee in the stochastic regime for EXP3++ with our robust gap estimation.

Theorem 4.1. *Let $\xi_t(a) = \frac{1200\ln t}{t\hat{\Delta}_t(a)^2}$, where $\hat{\Delta}_t(a)$ is the gap estimate from Algorithm 1. Then the expected regret of*

EXP3++ in the stochastic regime satisfies:

$$R(t) = O\left(\sum_{a:\Delta(a)>0} \frac{\ln^2 t}{\Delta(a)}\right) + \tilde{O}\left(\sum_{a:\Delta(a)>0} \frac{K}{\Delta(a)^3}\right), \tag{4}$$

where $\tilde{O}$ hides factors logarithmic in K.

We provide a proof of the theorem in Appendix C. We note that the regret bound matches the bound of Seldin & Lugosi (2017, Theorem 3), but we use importance-weighted gap estimates, opening potential for more applications.

The adversarial regret bound is taken directly from Seldin & Slivkins (2014), who provide a general adversarial analysis that holds for any choice of ξ_t.

Theorem 4.2 ((Seldin & Slivkins, 2014, Theorem 1))**.** *For $\eta_t = \frac{1}{2}\sqrt{\frac{\ln K}{tK}}$ and $\xi_t(a) \geq 0$ the regret of the EXP3++ algorithm in the adversarial regime for any t satisfies:*

$$R(t) \leq 4\sqrt{Kt\ln K}.$$

5. Discussion

We have provided a robust strategy for gap estimation based on importance weighted samples and implicit exploration. In combination with the EXP3++ algorithm it achieves regret of order $O\left(\sum_{a:\Delta(a)>0} \frac{(\ln t)^2}{\Delta(a)}\right)$ in the stochastic regime and regret of order $O(\sqrt{Kt\ln K})$ in the adversarial regime. While the regret bounds are the same as the bounds of Seldin & Lugosi (2017), the ability to use importance-weighted gap estimates opens the opportunity to achieve improved regret bounds in additional environments, such as stochastically constrained adversarial, to provide high-probability regret guarantees, and to expand to additional learning settings beyond multiarmed bandits. We emphasize that even though best-of-both-worlds algorithms like Tsallis-INF provide slightly tighter regret bounds, namely $O\left(\sum_{a:\Delta(a)>0} \frac{\ln t}{\Delta(a)}\right)$ in the stochastic regime and $O(\sqrt{Kt})$ in the adversarial regime, they provide neither gap estimates nor high-probability guarantees. The ability of our approach to provide high-probability gap estimates based on importance weighted samples might be valuable in its own right. We are looking forward to discuss these opportunities with workshop participants and explore them further in future work.

References

Auer, P. and Chiang, C.-K. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, 2016.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5, 2012.

Bubeck, S. and Slivkins, A. The best of both worlds: stochastic and adversarial bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, 2012.

Dann, C., Wei, C., and Zimmert, J. A blackbox approach to best of both worlds in bandits and beyond. In *Proceedings of the Conference on Learning Theory (COLT)*, 2023.

Ito, S. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Proceedings of the Conference on Learning Theory (COLT)*, 2021.

Jin, T., Liu, J., Rouyer, C., Chang, W., Wei, C.-Y., and Luo, H. No-regret online reinforcement learning with adversarial losses and transitions. In *Advances in Neural Information Processing Systems (NIPS)*, 2023.

Masoudian, S. and Seldin, Y. Improved analysis of the Tsallis-INF algorithm in stochastically constrained adversarial bandits and stochastic bandits with adversarial corruptions. In *Proceedings of the Conference on Learning Theory (COLT)*, 2021.

Masoudian, S., Zimmert, J., and Seldin, Y. A best-of-both-worlds algorithm for bandits with delayed feedback with robustness to excessive delays. Technical report, `https://arxiv.org/abs/2308.10675`, 2024.

Neu, G. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

Rouyer, C., van der Hoeven, D., Cesa-Bianchi, N., and Seldin, Y. A near-optimal best-of-both-worlds algorithm for online learning with feedback graphs. In *Advances in Neural Information Processing Systems (NIPS)*, 2022.

Seldin, Y. and Lugosi, G. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, 2017.

Seldin, Y. and Slivkins, A. One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

Zimmert, J. and Seldin, Y. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 2021.

A. Bernstein's Inequality for Martingales

We use the following concentration inequality of Seldin & Lugosi (2017) in our proofs. The important element for us that distinguishes it from the more broadly known Bernstein's inequality for martingales (**?**) is that it only requires one-sided boundedness of the martingale.

Theorem A.1 (Bernstein's inequality for martingales (Seldin & Lugosi, 2017))**.** *Let $X_1, ..., X_n$ be a martingale difference sequence with respect to filtration $F_1, ..., F_n$, where each X_j is bounded from above, and let $S_i = \sum_{j=1}^{i} X_j$ be the associated martingale. Let $v_n = \sum_{j=1}^{n} \mathbb{E}[(X_j)^2|F_{j-1}]$ and $c_n = \max_{1\leq j\leq n}\{X_j\}$. Then for any $\delta > 0$:*

$$\mathbb{P}\left\{\left(S_n \geq \sqrt{2\nu \ln\frac{1}{\delta}} + \frac{c\ln\frac{1}{\delta}}{3}\right) \wedge (v_n \leq \nu) \wedge (c_n \leq c)\right\} \leq \delta. \tag{5}$$

B. Proof of Proposition 1; Bounding the Probability of Failure

This section contains a proof of Proposition 1, that the gap estimates are reliable with high probability. It is comprised of two subsections, upper bounding the probability that the gap estimate is too large, and upper bounding the probability that it is too small once t passes a certain threshold. Before we begin with the proof, we introduce the following two inequalities:

$$\mathbb{P}\left(L_t^{IX}(a) - t\mu(a) \geq \frac{\ln(4(t+1))}{\gamma_t}\right) \leq \frac{1}{4(t+1)}, \tag{6}$$

$$\mathbb{P}\left(t\mu(a) \geq L_t^{IW}(a) + \sqrt{2\nu_t \ln(4(t+1))} + \frac{\ln(4(t+1))}{3}\right) \leq \frac{1}{4(t+1)}, \tag{7}$$

where line 6 follows from Neu (2015, Lemma 1), and 7 from Bernstein's inequality for Martingales, as stated in (Seldin & Lugosi, 2017, Theorem 9), a proof of line 7 is in Section D.

B.1. Upper Bound with High Probability

We want to show that

$$\mathbb{P}(\hat{\Delta}_t(a) \geq \Delta(a))$$

is small.

In the interest of legibility and without loss of generality, we prove this for $t+1$, though the proof would otherwise be the same. Firstly, we construct an upper bound on the probability that the gap estimate is larger than the true gap. Substituting in definitions, and then upper bounding using the inequalities on lines 6 and 7 leads to

$$\begin{aligned}
\mathbb{P}(\hat{\Delta}_{t+1}(a) \geq \Delta(a)) =& \mathbb{P}(t\hat{\Delta}_{t+1} \geq t\Delta(a)) \\
=& \mathbb{P}\left(L_t^{IX}(a) - \frac{\ln(4(t+1))}{\gamma_t} - \min_a\left(L_t^{IW}(a) + \sqrt{2\nu_t(a)\ln(4(t+1))} + \frac{\ln(4(t+1))}{3}\right) \geq t\mu(a) - t\mu(a^*)\right) \\
\leq& \mathbb{P}\left(L_t^{IX}(a) - \frac{\ln(4(t+1))}{\gamma_t} \geq t\mu(a)\right) + \mathbb{P}\left(\min_a\left(L_t^{IW}(a) + \sqrt{2\nu_t(a)\ln(4(t+1))} + \frac{\ln(4(t+1))}{3}\right) \leq t\mu(a^*)\right) \qquad (8) \\
\leq& \frac{1}{4(t+1)} + \mathbb{P}\left(\min_a\left(L_t^{IW}(a) + \sqrt{2\nu_t(a)\ln(4(t+1))} + \frac{\ln(4(t+1))}{3}\right) \leq t\mu(a^*)\right) \qquad (9) \\
\leq& \frac{1}{4(t+1)} + \mathbb{P}\left(L_t^{IW}(a^*) + \sqrt{2\nu_t(a^*)\ln(4(t+1))} + \frac{\ln(4(t+1))}{3} \leq t\mu(a^*)\right) \qquad (10) \\
\leq& \frac{1}{2(t+1)}, \qquad (11)
\end{aligned}$$

where line 9 follows by upper bounding the first term of line 8 using 6. Line 11 follows from 7.

B.2. Lower Bound with High Probability

It remains to show that the gap estimate is much smaller than the true gap with small probability. We want to show that

$$\mathbb{P}\left(\hat{\Delta}_t(a) \leq \frac{\Delta(a)}{2}\right)$$

is small.

Our approach involves substituting in the definitions, then splitting the probability into three terms handled separately, which, when combined, lead to an upper bound on the probability of interest.

We expand and then separate into three parts as shown in the following sections, where $c_t(x,y) = \sqrt{2x\ln(4(t+1))} + \frac{y\ln(4(t+1))}{3}$, and $a' = argmin_a(L_t^{IW}(a) + c_t(\nu_t(a),1))$.

B.2.1. Separate

Again, in the interest of legibility and without loss of generality, we prove this for $t+1$, though the proof would otherwise be the same. Substituting in the definitions of $\hat{\Delta}_{t+1}(a)$, $\Delta(a)$, and $c_t(x,y)$ we have:

$$\begin{aligned}\mathbb{P}\left(\hat{\Delta}_{t+1}(a) \leq \frac{\Delta(a)}{2}\right) &= \mathbb{P}\left(t\hat{\Delta}_{t+1}(a) \leq \frac{t\Delta(a)}{2}\right) \\ &= \mathbb{P}\left(L_t^{IX}(a) - \frac{\ln(4(t+1))}{\gamma_t} - \min_a\left(L_t^{IW}(a) + c_t(\nu_t(a),1)\right) \leq \frac{t\Delta(a)}{2}\right).\end{aligned}$$

Adding 0 terms, and rewriting the right side we have:

$$\begin{aligned}&= \mathbb{P}\left(L_t^{IX}(a) - \frac{\ln(4(t+1))}{\gamma_t} - \min_a\left(L_t^{IW}(a) + c_t(\nu_t(a),1)\right) + \sum_{s=1}^{t}\frac{\gamma_s}{\tilde{p}_s(a)} - \sum_{s=1}^{t}\frac{\gamma_s}{\tilde{p}_s(a)}\right. \\ &+ L_t^{IW}(a^*) - L_t^{IW}(a^*) + c_t\left(\nu_t(a^*), \frac{1}{\tilde{p}_t(a^*)}\right) - c_t\left(\nu_t(a^*), \frac{1}{\tilde{p}_t(a^*)}\right) + c_t\left(\nu_t(a), \frac{1}{\tilde{p}_t(a)}\right) - c_t\left(\nu_t(a), \frac{1}{\tilde{p}_t(a)}\right) \\ &\left.\leq t\mu(a) - t\mu(a^*) - \frac{t\Delta(a)}{2}\right).\end{aligned}$$

Rearranging and using the definition of a' leads to

$$= \mathbb{P}\left(L_t^{IX}(a) - t\mu(a) + c_t\left(\nu_t(a), \frac{1}{\tilde{p}_t(a)}\right) + \sum_{s=1}^{t}\frac{\gamma_s}{\tilde{p}_s(a)}\right. \tag{12}$$

$$+ t\mu(a^*) - L_t^{IW}(a^*) + c_t\left(\nu_t(a^*), \frac{1}{\tilde{p}_t(a^*)}\right) \tag{13}$$

$$\left.+ \frac{t\Delta(a)}{2} - \frac{\ln(4(t+1))}{\gamma_t} - c_t(\nu_t(a'),1) + L_t^{IW}(a^*) - L_t^{IW}(a') - c_t\left(\nu_t(a^*), \frac{1}{\tilde{p}_t(a^*)}\right) - c_t\left(\nu_t(a), \frac{1}{\tilde{p}_t(a)}\right) - \sum_{s=1}^{t}\frac{\gamma_s}{\tilde{p}_s(a)} \leq 0\right). \tag{14}$$

Let A denote the group of terms on line 12, B on line 13, and C on line 14.

$$\begin{aligned}&= \mathbb{P}(A \\ &+ B \\ &+ C \leq 0)\end{aligned} \tag{15}$$

Which can be upper bounded;

$$\leq \mathbb{P}(A \leq 0) + \mathbb{P}(B \leq 0) + \mathbb{P}(C \leq 0). \tag{16}$$

We next upper bound the first two probabilities in line 16 by $\frac{1}{4(t+1)}$ as shown in the following two sections.

B.2.2. BOUND A

First, we upper bound $\mathbb{P}(A \leq 0)$. We want to show:

$$\mathbb{P}\left(L_t^{IX}(a) - t\mu(a) + c_t\left(\nu_t(a), \frac{1}{\tilde{p}_t(a)}\right) + \sum_{s=1}^{t} \frac{\gamma_s}{\tilde{p}_s(a)} \leq 0\right) \leq \frac{1}{4(t+1)}.$$

Start by rearranging $\mathbb{P}(A \leq 0)$ to write:

$$\mathbb{P}\left(c_t\left(\nu_t(a), \frac{1}{\tilde{p}_t(a)}\right) + \sum_{s=1}^{t} \frac{\gamma_s}{\tilde{p}_s(a)} \leq t\mu(a) - L_t^{IX}(a)\right).$$

Then expand the right side:

$$t\mu(a) - L_t^{IX}(a) = t\mu(a) - \mathbb{E}[L_t^{IX}(a)] + \mathbb{E}[L_t^{IX}(a)] - L_t^{IX}(a). \tag{17}$$

The next step is to upper bound the first two terms on the right hand side of line 17.

$$\begin{aligned} t\mu(a) - \mathbb{E}[L_t^{IX}(a)] &= \sum_{s=1}^{t} \mu(a) \frac{\gamma_s}{\tilde{p}_s(a) + \gamma_s} \\ &\leq \sum_{s=1}^{t} \frac{\gamma_s}{\tilde{p}_s(a)}. \end{aligned} \tag{18}$$

The last two terms of line 17 can be lower bounded with probability at most δ by applying Bernstein's inequality for Martingales. Let S_t denote the last two terms of line 17, and let X_i be derived from S_t as follows:

$$\begin{aligned} S_t &= \mathbb{E}[L_t^{IX}(a)] - L_t^{IX}(a) \\ &= \sum_{i=1}^{t} \mathbb{E}[\ell_t^{IX}(a)] - \ell_t^{IX}(a) \\ &= \sum_{i=1}^{t} X_i. \end{aligned}$$

Each X_t is bounded from above:

$$\begin{aligned} X_t &= \mathbb{E}[\ell_t^{IX}(a)] - \ell_t^{IX}(a) \\ &\leq \frac{\tilde{p}_t(a)\ell_t(a)}{\tilde{p}_t(a) + \gamma_t} \\ &\leq 1, \end{aligned}$$

and has, by construction, expected value of 0 given the history up to and including time $t-1$:

$$\mathbb{E}[X_t|F_{t-1}] = 0.$$

Therefore, as $X_1, ..., X_t$ is a martingale difference sequence, $S_t = \sum_{i=1}^t X_i$ is the associated martingale. To apply Bernstein's inequality for Martingales we still need to bound the variance of S_t.

$$\begin{aligned}
\mathbb{E}[X_t^2|F_{t-1}] &= \mathbb{E}[(\mathbb{E}[\ell_t^{IX}(a)] - \ell_t^{IX}(a))^2] \\
&= \mathbb{E}[(\ell_t^{IX}(a) - \mathbb{E}[\ell_t^{IX}(a)])^2] \\
&= \mathbb{E}[(\ell_t^{IX}(a))^2] - \mathbb{E}[\ell_t^{IX}(a)]^2 \\
&= \mathbb{E}\left[\frac{\mathbb{1}(A_t = a)\ell_t(a)^2}{(\tilde{p}_t(a) + \gamma_t)^2}\right] - \frac{\tilde{p}_t(a)^2\mu(a)^2}{(\tilde{p}_t(a) + \gamma_t)^2} \\
&= \frac{\tilde{p}_t(a)\ell_t^2(a)}{(\tilde{p}_t(a) + \gamma_t)^2} - \frac{\tilde{p}_t(a)^2\ell_t^2(a)}{(\tilde{p}_t(a) + \gamma_t)^2} \\
&\leq \frac{\tilde{p}_t(a)}{(\tilde{p}_t(a) + \gamma_t)^2}(1 - \tilde{p}_t(a)) \\
&\leq \frac{1}{\tilde{p}_t(a) + \gamma_t} \\
&\leq \frac{1}{\tilde{p}_t(a)} \\
&\leq \epsilon_t^{-1}(a) \\
&\implies \\
v_t(a) = \sum_{j=1}^t \mathbb{E}[X_j^2|F_{j-1}] &\leq \sum_{j=1}^t \epsilon_j^{-1}(a) = \nu_t(a)
\end{aligned} \tag{19}$$

Applying Bernstein's Inequality for Martingales results in:

$$\mathbb{P}\bigg(\mathbb{E}[L_t^{IX}(a)] - L_t^{IX}(a) \geq c_t\Big(\nu_t(a), 1\Big)\bigg) \leq \frac{1}{4(t+1)}. \tag{20}$$

Putting the previous steps together to bound $\mathbb{P}(A \leq 0)$:

$$\begin{aligned}
\mathbb{P}(A \leq 0) =& \mathbb{P}\bigg(t\mu(a) - L_t^{IX}(a) \geq \sum_{s=1}^t \frac{\gamma_s}{\tilde{p}_s(a)} + c_t\Big(\nu_t(a), \frac{1}{\tilde{p}_t(a)}\Big)\bigg) \\
&= \mathbb{P}\bigg(\Big(t\mu(a) - \mathbb{E}[L_t^{IX}(a)]\Big) + \Big(\mathbb{E}[L_t^{IX}(a)] - L_t^{IX}(a)\Big) \geq \sum_{s=1}^t \frac{\gamma_s}{\tilde{p}_s(a)} + c_t\Big(\nu_t(a), \frac{1}{\tilde{p}_t(a)}\Big)\bigg) && (21) \\
&\leq \mathbb{P}\bigg(\mathbb{E}[L_t^{IX}(a)] - L_t^{IX}(a) \geq c_t\Big(\nu_t(a), \frac{1}{\tilde{p}_t(a)}\Big)\bigg) && (22) \\
&\leq \mathbb{P}\bigg(\mathbb{E}[L_t^{IX}(a)] - L_t^{IX}(a) \geq c_t\Big(\nu_t(a), 1\Big)\bigg) && (23) \\
&\leq \frac{1}{4(t+1)}, && (24)
\end{aligned}$$

where line 21 follows from expanding as in line 17 and line 22 follows from upper bounding $t\mu(a) - \mathbb{E}[L_t^{IX}(a)]$ as in 18. Line 23 follows by lower bounding $c_t\Big(\nu_t(a), \frac{1}{\tilde{p}_t(a)}\Big)$ with $c_t(\nu_t(a), 1)$. Finally, line 24 follows directly from 20.

B.2.3. Bound B

After bounding $\mathbb{P}(A \leq 0)$, we upper bound $\mathbb{P}(B \leq 0)$We want to show:

$$\mathbb{P}\left(t\mu(a^*) - L_t^{IW}(a^*) + c_t(\nu_t(a^*), \frac{1}{\tilde{p}_t(a^*)}) \leq 0\right) \leq \frac{1}{4(t+1)}.$$

The first step is to rewrite $\mathbb{P}(B \leq 0)$.

$$\mathbb{P}(B \leq 0) = \mathbb{P}\left(L_t^{IW}(a^*) - t\mu(a^*) \geq c_t\left(\nu_t(a^*), \frac{1}{\tilde{p}_t(a^*)}\right)\right)$$

Then, using the same technique as when bounding $\mathbb{P}(A \leq 0)$, let

$$X_t = \ell_t^{IW}(a^*) - \mu(a^*)$$

and

$$S_t = \sum_{i=1}^{t} X_i = L_t^{IW}(a^*) - t\mu(a^*).$$

In order to apply Bernstein's Inequality for Martingales we firstly show that $X_1, ..., X_t$ is a martingale difference sequence. Each term is bounded from above:

$$X_t \leq \ell_t^{IW}(a^*) \leq \frac{1}{\tilde{p}_t(a^*)}.$$

And $\mathbb{E}[X_t|F_{t-1}] = 0$:

$$\begin{aligned}\mathbb{E}[X_t|F_{t-1}] &= \mathbb{E}\left[\frac{\mathbb{1}(A_t = a^*)\ell_t(a^*)}{\tilde{p}_t(a^*)} - \mu(a^*)\right] \\ &= \frac{\tilde{p}_t(a^*)\mu(a^*)}{\tilde{p}_t(a^*)} - \mu(a^*) \\ &= 0.\end{aligned}$$

By construction, S_t is the associated martingale, and as before, in order to apply Bernstein's Inequality for Martingales, we now bound the variance of S_t. The first line follows directly from the definition of variance, and that $\mathbb{E}[\ell_t^{IW}(a^*)] = \mu(a^*)$.

$$\begin{aligned}\mathbb{E}[X_t^2|\mathcal{F}_{t-1}] &= \mathbb{E}[(\ell_t^{IW}(a^*))^2] - \mathbb{E}[\ell_t^{IW}(a^*)]^2 \\ &\leq \mathbb{E}\left[\left(\frac{\ell_t(a^*)\mathbb{1}\{A_t = a^*\}}{\tilde{p}_t(a^*)}\right)^2\right] = \mathbb{E}\left[\frac{\ell_t(a^*)^2\mathbb{1}\{A_t = a^*\}^2}{\tilde{p}_t(a^*)^2}\right] \\ &\leq \mathbb{E}\left[\frac{\ell_t(a^*)^2\mathbb{1}\{A_t = a^*\}}{\tilde{p}_t(a^*)^2}\right] = \frac{\mu(a^*)^2\tilde{p}_t(a^*)}{\tilde{p}_t(a^*)^2} \\ &\leq \frac{1}{\tilde{p}_t(a^*)} \\ &\leq \frac{1}{\epsilon_t(a^*)} \\ &\implies \\ v_t(a^*) = \sum_{j=1}^{t} \mathbb{E}[X_j^2|F_{j-1}] &\leq \sum_{j=1}^{t} \epsilon_j(a^*)^{-1} = \nu_t(a^*)\end{aligned} \tag{25}$$

Lastly, applying Bernstein's inequality for martingales results in:

$$\mathbb{P}(B \leq 0) = \mathbb{P}\left(L_t^{IW}(a^*) - t\mu(a^*) \geq c_t\left(\nu_t(a^*), \frac{1}{\tilde{p}_t(a^*)}\right)\right) \leq \frac{1}{4(t+1)}.$$

B.2.4. Bound C

To complete the bounding of line 16 it remains to upper bound $\mathbb{P}(C \leq 0)$.We want to show

$$\mathbb{P}\Bigg(\frac{t\Delta(a)}{2} - \frac{\ln(4(t+1))}{\gamma_t} - c_t(\nu_t(a'), 1) + L_t^{IW}(a^*) - L_t^{IW}(a') \\ - c_t(\nu_t(a^*), \frac{1}{\tilde{p}_t(a^*)}) - c_t(\nu_t(a), \frac{1}{\tilde{p}_t(a)}) - \sum_{s=1}^{t} \frac{\gamma_s}{\tilde{p}_s(a)} \leq 0\Bigg) \tag{26}$$

is small.

By adding $0 = c_t(\nu_t(a^*), 1) - c_t(\nu_t(a^*), 1)$ we rewrite C as:

$$C = \frac{t\Delta(a)}{2} - \frac{\ln(4(t+1))}{\gamma_t} + (L_t^{IW}(a^*) + c_t(\nu_t(a^*), 1)) - (L_t^{IW}(a') + c_t(\nu_t(a'), 1)) \\ - c_t(\nu_t(a^*), 1) - c_t\left(\nu_t(a^*), \frac{1}{\tilde{p}_t(a^*)}\right) - c_t\left(\nu_t(a), \frac{1}{\tilde{p}_t(a)}\right) - \sum_{s=1}^{t} \frac{\gamma_s}{\tilde{p}_s(a)}. \tag{27}$$

We define the following function:

$$F(t) = \frac{\ln(4(t+1))}{\gamma_t} + c_t(\nu_t(a^*), 1) + c_t\left(\nu_t(a^*), \frac{1}{\tilde{p}_t(a^*)}\right) + c_t\left(\nu_t(a), \frac{1}{\tilde{p}_t(a)}\right) + \sum_{s=1}^{t} \frac{\gamma_s}{\tilde{p}_s(a)}. \tag{28}$$

By definition of a' we have:

$$\left(L_t^{IW}(a^*) + c_t(\nu_t(a^*), 1)\right) \geq \left(L_t^{IW}(a') + c_t(\nu_t(a'), 1)\right).$$

Meaning that C, on line 27, can be lower bounded by:

$$C \geq \frac{t\Delta(a)}{2} - F(t) \\ \implies \\ \mathbb{P}(C \leq 0) \leq \mathbb{P}\left(F(t) \geq \frac{t\Delta(a)}{2}\right).$$

Substituting in the definition of c_t leads to:

$$F(t) = \frac{\ln(4(t+1))}{\gamma_t} + 2\sqrt{2\nu_t(a^*)\ln(4(t+1))} + \frac{\ln(4(t+1))}{3} + \frac{\ln(4(t+1))}{3\tilde{p}_t(a^*)} + \sum_{s=1}^{t} \frac{\gamma_s}{\tilde{p}_s(a)} + \sqrt{2\nu_t(a)\ln(4(t+1))} + \frac{\ln(4(t+1))}{3\tilde{p}_t(a)}. \tag{29}$$

Then, assuming that $\epsilon_t(a) = \xi_t(a)$, upper bound $\nu_t(a^*)$ and $\tilde{p}_t(a^*)^{-1}$, and substitute in the definition of $\nu_t(a)$. This will restrict the time interval to $t \geq t_{\min}$, which is addressed later. This leads to:

$$F(t) \leq \frac{\ln(4(t+1))}{3}(1 + 2\xi_t(a)^{-1}) + 3\sqrt{2\ln(4(t+1))\sum_{s=1}^{t} \epsilon_s(a)^{-1}} + \frac{\ln(4(t+1))}{\gamma_t} + \sum_{s=1}^{t} \gamma_s \epsilon_s(a)^{-1} \tag{30}$$

$$\tag{31}$$

Upper bounding the sum $\sum_{s=1}^{t} \epsilon_s(a)^{-1}$ as follows;

$$
\begin{aligned}
\sum_{s=1}^{t} \epsilon_s(a)^{-1} &\leq \sum_{s=1}^{t_{\min}} \epsilon_s(a)^{-1} + \sum_{s=t_{\min}}^{t} \epsilon_s(a)^{-1} \\
&\leq \sum_{s=1}^{t_{\min}} \frac{ct\Delta(a)^2}{\ln t} + \sum_{s=t_{\min}}^{t} \epsilon_s(a)^{-1} \\
&\leq \frac{ct^2\Delta(a)^2}{\ln t} + \sum_{s=t_{\min}}^{t} \xi_s(a)^{-1} \\
&\leq \frac{ct^2\Delta(a)^2}{\ln t} + \sum_{s=1}^{t} \xi_s(a)^{-1} \\
&= \frac{ct^2\Delta(a)^2}{\ln t} + \sum_{s=1}^{t} \frac{cs\hat{\Delta}_s(a)^2}{\ln s}. \qquad (32)
\end{aligned}
$$

Using line 32, and substituting in the definition of ξ we can upper bound $F(t)$ further;

$$
F(t) \leq \frac{\ln(4(t+1))}{3}(1 + \frac{2ct\hat{\Delta}_t(a)^2}{\ln t}) + 3\sqrt{2\ln(4(t+1))\left(\frac{ct^2\Delta(a)^2}{\ln t} + \sum_{s=1}^{t} \frac{cs\hat{\Delta}_s(a)^2}{\ln s}\right)} + \frac{\ln(4(t+1))}{\gamma_t} + \sum_{s=1}^{t} \gamma_s \epsilon_s(a)^{-1} \qquad (33)
$$

$$
\leq \frac{\ln(4(t+1))}{3}(1 + \frac{2ct\hat{\Delta}_t(a)^2}{\ln t}) + 3\sqrt{2c\ln(4(t+1))\left(\frac{t^2\Delta(a)^2}{\ln t} + \sum_{s=1}^{t} \frac{s\hat{\Delta}_s(a)^2}{\ln s}\right)} + \frac{\ln(4(t+1))}{\ln t} t\hat{\Delta}_t(a)\sqrt{c} + \sum_{s=1}^{t} \hat{\Delta}_s(a)\sqrt{c}. \qquad (34)
$$

Where line 33 follows from substituting in the definition of $\xi_t(a)$. The last two terms of 33 can be upper bounded by setting $\gamma_t = \epsilon_t(a)\hat{\Delta}_t(a)\sqrt{c}$ for all t, resulting in line 34. As we are bounding the probability of $\hat{\Delta}_t(a) \leq \frac{\Delta(a)}{2}$, we have $\hat{\Delta}_t(a) \leq \Delta(a)$, which, along with $\Delta(a)^2 \leq \Delta(a)$, allows for the following:

$$
\begin{aligned}
F(t) &\leq \frac{\ln(4(t+1))}{3}\left(1 + \frac{2ct\Delta(a)}{\ln t}\right) + 3\sqrt{2c\ln 4t\left(\frac{t^2\Delta(a)^2}{\ln t} + \sum_{s=1}^{t} \frac{s\Delta(a)^2}{\ln s}\right)} + \frac{\ln(4(t+1))}{\ln t} t\Delta(a)\sqrt{c} + t\Delta(a)\sqrt{c} \\
&= \frac{\ln(4(t+1))}{3}\left(1 + \frac{2ct\Delta(a)}{\ln t}\right) + 3\Delta(a)\sqrt{2c\left(\frac{t^2\ln(4(t+1))}{\ln t} + \sum_{s=1}^{t} \frac{s\ln 4t}{\ln s}\right)} + t\Delta(a)\sqrt{c}\left(\frac{\ln(4(t+1))}{\ln t} + 1\right) \qquad (35) \\
&\leq \frac{\ln(4(t+1))}{3}\left(1 + \frac{2ct\Delta(a)}{\ln t}\right) + 3\Delta(a)\sqrt{2c\left(\frac{t^2\ln(4(t+1))}{\ln t} + \sum_{s=1}^{t} \frac{t\ln(4(t+1))}{\ln t}\right)} + t\Delta(a)\sqrt{c}\left(\frac{\ln(4(t+1))}{\ln t} + 1\right) \qquad (36) \\
&= \frac{\ln(4(t+1))}{3} + \frac{2ct\Delta(a)\ln(4(t+1))}{3\ln t} + 6t\Delta(a)\sqrt{c\frac{\ln(4(t+1))}{\ln t}} + t\Delta(a)\sqrt{c}\left(\frac{\ln(4(t+1))}{\ln t} + 1\right) \\
&= \frac{\ln(4(t+1))}{3} + t\Delta(a)\left(\frac{2}{3}\frac{\ln(4(t+1))}{\ln t}c + \left(6\sqrt{\frac{\ln(4(t+1))}{\ln t}} + \frac{\ln(4(t+1))}{\ln t} + 1\right)\sqrt{c}\right), \qquad (37)
\end{aligned}
$$

where line 36 follows from 35 by $\frac{s}{\ln s} \leq \frac{t}{\ln t}$ for $s \geq t_{\min}$.

The next step is to strictly upper bound each term in line 37 by $\frac{t\Delta(a)}{4}$, in order to upper bound $F(t)$ by $\frac{t\Delta(a)}{2}$. Starting with the first term, the following holds for $t \geq t_{\min}$:

$$\frac{\ln(4(t+1))}{t} \leq \frac{3\Delta(a)}{4}.$$

To upper bound the second term for $t \geq t_{\min}$ note that $t_{\min} \geq \frac{e \cdot 4}{c} > 1 + \frac{4}{c} > \frac{4}{c}$, and substituting this value for t gives an upper bound on $\frac{\ln(4(t+1))}{\ln t}$. Using this, we do the following:

$$\begin{aligned} t\Delta(a)\left(\frac{2}{3}\frac{\ln\frac{16}{c}}{\ln\frac{4}{c}}c + \left(6\sqrt{\frac{\ln\frac{16}{c}}{\ln\frac{4}{c}}} + \frac{\ln\frac{16}{c}}{\ln\frac{4}{c}} + 1\right)\sqrt{c}\right) &< \frac{t\Delta(a)}{4} \\ \frac{2}{3}\frac{\ln\frac{16}{c}}{\ln\frac{4}{c}}c + \left(6\sqrt{\frac{\ln\frac{16}{c}}{\ln\frac{4}{c}}} + \frac{\ln\frac{16}{c}}{\ln\frac{4}{c}} + 1\right)\sqrt{c} &< \frac{1}{4} \qquad (38) \\ &\implies \\ \frac{2}{3}\frac{\ln\frac{16}{c}}{\ln\frac{4}{c}}c + \left(6\sqrt{\frac{\ln\frac{16}{c}}{\ln\frac{4}{c}}} + \frac{\ln\frac{16}{c}}{\ln\frac{4}{c}} + 1\right)\sqrt{c} - \frac{1}{4} &= 0. \qquad (39) \end{aligned}$$

Solving for the non-negative solution to line 39 gives:

$$c \geq \frac{1}{1200}. \tag{40}$$

Consequently, line 29 can be upper bounded by 0, for large enough t:

$$\begin{gathered} C \leq \frac{t\Delta(a)}{2} - F(t) \leq 0 \\ \implies \\ \mathbb{P}(C \leq 0) \leq \mathbb{P}\left(F(t) \geq \frac{t\Delta(a)}{2}\right) = 0. \end{gathered}$$

Putting all of the pieces together to bound the probability the gap estimate is too small,

$$\begin{aligned} \mathbb{P}\left(\hat{\Delta}_t(a) \leq \frac{\Delta(a)}{2}\right) &\leq \mathbb{P}(A \leq 0) + \mathbb{P}(B \leq 0) + \mathbb{P}(C \leq 0) \\ &\leq \frac{1}{4t} + \frac{1}{4t} + \mathbb{P}(C \leq 0) \\ &= 2\frac{1}{4t} \\ &= \frac{1}{2t} \text{ for } t \geq t_{min}(a). \end{aligned}$$

C. Proof of Theorem 1, Stochastic Regret Guarantee

We start by bounding $\mathbb{E}[N_T(a)]$. We split this into three parts, when the gap estimate is potentially too small during an initial period of the game, when it is either too large or too small at any time, and when the gap estimate is good but a sup-optimal action may be chosen regardless;

$$\mathbb{E}[N_a(t)] = \mathbb{E}[N_{1,a}(t)] + \mathbb{E}[N_{2,a}(t)] + \mathbb{E}[N_{3,a}(t)].$$

During the first $t \leq t_{min}(a)$ time steps, for any action, the gap estimate is not reliable, as it may be less than half the true gap. As such, during this period a sub optimal action may be played

$$\mathbb{E}[N_{1,a}(t)] \leq t_{\min}(a) = \tilde{O}\left(\frac{K}{\Delta(a)^4}\right)$$

times, where the $\tilde{O}$ notation hides the logarithmic factors.

The gap estimate may also fail after that time threshold, when $\hat{\Delta}(a) \geq \Delta(a)$ or $\hat{\Delta}(a) \leq \frac{\Delta(a)}{2}$, and a sub-optimal action may be played. The expected number of times this can happen for an action a is upper bounded by the following:

$$\begin{aligned}
\mathbb{P}(\hat{\Delta}(a) \geq \Delta(a)) + \mathbb{P}\bigg(\hat{\Delta}(a) \leq \frac{\Delta(a)}{2}\bigg) \leq 2\frac{1}{4t} + 2\frac{1}{4t} = 4\frac{1}{4t} = \frac{1}{t}& \\
\implies& \\
\mathbb{E}[N_{2,a}(t)] &\leq \sum_{s=1}^{t} \mathbb{P}(\hat{\Delta}_s(a) \geq \Delta(a)) + \mathbb{P}\bigg(\hat{\Delta}_s(a) \leq \frac{\Delta(a)}{2}\bigg) \\
&\leq \sum_{s=1}^{t} \frac{1}{s} \\
&= O(\ln t).
\end{aligned}$$

Even when the gap estimate is good, a sub-optimal action may still be played. This comes from $\tilde{p}_t(a)$, which is composed of two parts, handled separately as follows:

$$\mathbb{E}[N_{3,a}(t)] = \mathbb{E}\bigg[\sum_{s=t_{\min}(a)}^{t} \tilde{p}_s(a)\bigg] \leq \sum_{s=1}^{t} \mathbb{E}\bigg[\bigg(\epsilon_s(a) + (1 - \sum_{a'} \epsilon_s(a'))p_s(a)\bigg)\bigg]. \tag{41}$$

Starting with upper bounding the first term:

$$\begin{aligned}
\epsilon_s(a) &\leq \frac{\ln s}{cs\hat{\Delta}_s(a)^2} \\
&\implies \\
\sum_{s=1}^{t} \mathbb{E}\bigg[\epsilon_s(a)\bigg] &\leq \sum_{s=1}^{t} \mathbb{E}\bigg[\frac{\ln s}{cs\hat{\Delta}_s(a)^2}\bigg] \\
&\leq \sum_{s=1}^{t} \frac{4\ln s}{cs\Delta(a)^2} \\
&\leq \frac{4}{c\Delta(a)^2} \sum_{s=1}^{t} \frac{\ln t}{s} \\
&\leq \frac{4\ln^2 t}{c\Delta(a)^2} \\
&= O\bigg(\frac{\ln^2 t}{\Delta(a)^2}\bigg).
\end{aligned}$$

The first step to upper bound the second term of line 41 starts by upper bounding it by p:

$$(1 - \sum_{a'} \epsilon_s(a'))p_s(a) \leq p_s(a).$$

Upper bounding p is done nearly identically as in (Seldin & Lugosi, 2017, Proof of Theorem 3). The bound on the gap estimate being too large is of the same order, $\frac{1}{t}$, and as $\beta = \frac{1}{c} = 1200$ this satisfies the requirement that $\beta \geq 256$. The only difference is that in our analysis, we handle the $t_{\min}$ rounds of the game separately. Taking this we have:

$$\sum_{s=1}^{t} \mathbb{E}[p_s(a)] = O\bigg(\frac{(\ln t)^2}{\Delta(a)^2}\bigg).$$

Together we have:

$$\mathbb{E}[N_{3,a}] = O\bigg(\frac{(\ln t)^2}{\Delta(a)^2}\bigg).$$

Putting this together to get a bound on regret:

$$
\begin{aligned}
R(t) &= \sum_{a:\Delta(a)>0} \mathbb{E}[N_a(t)]\Delta(a) \\
&= \sum_{a:\Delta(a)>0} \left(O(\ln t) + \tilde{O}\left(\frac{K}{\Delta(a)^4}\right) + O\left(\frac{\ln^2 t}{\Delta(a)^2}\right)\right)\Delta(a) \\
&= O\left(\sum_{a:\Delta(a)>0} \ln t\Delta(a)\right) + \tilde{O}\left(\sum_{a:\Delta(a)>0} \frac{K}{\Delta(a)^3}\right) + O\left(\sum_{a:\Delta(a)>0} \frac{\ln^2 t}{\Delta(a)}\right).
\end{aligned}
$$

D. Proof of Second Inequality (line 7)

The proof of the inequality on line 7 is as follows. The left hand side of line 7 can be rewritten as:

$$
\mathbb{P}\left(t\mu(a) - L_t^{IW}(a) \geq \sqrt{2\nu_t \ln(4(t+1))} + \frac{\ln(4(t+1))}{3}\right).
$$

Let $S_t = t\mu(a) - L_t^{IW}(a)$, we show that S_t is a martingale, apply Bernstein's inequality for Martingales and arrive at line 7. Start by rewriting S_t:

$$
S_t = \sum_{s=1}^{t} \mu(a) - \ell_s^{IW}(a).
$$

Clearly $\mu(a) - \ell_s^{IW}(a) \leq 1$, meaning each term is upper bounded. We must also show that the expected value of each term with respect to the past is 0.

$$
\begin{aligned}
\mathbb{E}[\mu(a) - \ell_s^{IW}(a)|F_{s-1}] &= \mathbb{E}[\mu(a) - \ell_s^{IW}(a)] \\
&= \mathbb{E}[\mu(a)] - \mathbb{E}[\ell_s^{IW}(a)] \\
&= \mu(a) - \mathbb{E}\left[\frac{\ell_s(a)\mathbb{1}(A_s = a)}{\tilde{p}_s(a)}\right] \\
&= \mu(a) - \frac{\mu(a)\tilde{p}_s(a)}{\tilde{p}_s(a)} \\
&= 0
\end{aligned}
$$

As $\mu(a) - \ell_s^{IW}(a)$ is upper bounded and has expected value 0 for any s, it forms a martingale difference sequence, and by construction, S_t is the associated Martingale. In order to apply Bernstein's inequality for Martingales it remains to bound the variance of S_t .

$$
v_t = \sum_{s=1}^{t} \mathbb{E}[(\mu(a) - \ell_s^{IW}(a))^2|F_{s-1}]
$$

First note:

$$
\mathbb{E}[\mu(a)^2] + \mathbb{E}\left[\frac{-2\mu(a)^2\mathbb{1}(A_s = a)}{\tilde{p}_s(a)}\right] = \mu(a)^2 - 2\mu(a)^2\frac{\tilde{p}_s(a)}{\tilde{p}_s(a)} \leq 0 \tag{42}
$$

We start by bounding each term in v:

$$\begin{aligned}
\mathbb{E}[(\mu(a) - \ell_s^{IW}(a))^2] &= \mathbb{E}[\mu(a)^2] + \mathbb{E}[\ell_s^{IW}(a)^2] + \mathbb{E}[-2\mu(a)\ell_s^{IW}(a)] \\
&= \mathbb{E}[\mu(a)^2] + \mathbb{E}\left[\frac{\ell_s(a)^2 \mathbb{1}(A_s = a)^2}{\tilde{p}_s(a)^2}\right] + \mathbb{E}\left[\frac{-2\mu(a)^2 \mathbb{1}(A_s = a)}{\tilde{p}_s(a)}\right] \qquad (43) \\
&\leq \mathbb{E}\left[\frac{\ell_s(a)^2 \mathbb{1}(A_s = a)^2}{\tilde{p}_s(a)^2}\right] \qquad (44) \\
&\leq \mathbb{E}\left[\frac{\mathbb{1}(A_s = a)}{\tilde{p}_s(a)^2}\right] \\
&= \frac{\tilde{p}_s(a)}{\tilde{p}_s(a)^2} \\
&= \frac{1}{\tilde{p}_s(a)},
\end{aligned}$$

where line 44 follows from line 43 by upper bounding the first and last term with 0 as in line 42.

$$\begin{aligned}
v_t &= \sum_{s=1}^{t} \mathbb{E}[(\mu(a) - \ell_s^{IW}(a))^2 | F_{s-1}] \\
&\leq \sum_{s=1}^{t} \frac{1}{\tilde{p}_s(a)} \\
&\leq \sum_{s=1}^{t} \epsilon_s(a)^{-1} = \nu_t(a)
\end{aligned}$$

As this is an upper bound for v_t for all t, the probability of the second event in Bernstein's inequality for martingales is 1, and the same can be done for the probability of the third event using $c = 1$. We now apply Bernstein's inequality for martingales, with $\delta = \frac{1}{4(t+1)}$, directly resulting in line 7.