References

Agarwal, A. and Duchi, J. C. Distributed delayed stochastic optimization. In *Advances in Neural Information Processing Systems*, pp. 873–881, 2011.

Arya, S. and Yang, Y. Randomized allocation with nonparametric estimation for contextual multi-armed bandits with delayed rewards. *arXiv preprint arXiv:1902.00819*, 2019.

Audibert, J.-Y. and Bubeck, S. Minimax policies for bandits games. *COLT 2009*, 2009.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Carpentier, A. and Kim, A. K. Adaptive and minimax optimal estimation of the tail coefficient. *Statistica Sinica*, pp. 1133–1144, 2015.

Cesa-Bianchi, N., Gentile, C., and Mansour, Y. Nonstochastic bandits with composite anonymous feedback. In *Conference On Learning Theory*, pp. 750–773, 2018.

Chapelle, O. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1097–1105. ACM, 2014.

Chapelle, O. and Li, L. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pp. 2249–2257, 2011.

Cover, T. M. Universal portfolios. In *The Kelly Capital Growth Investment Criterion: Theory and Practice*, pp. 181–209. World Scientific, 2011.

De Haan, L. and Ferreira, A. *Extreme value theory: an introduction*. Springer Science & Business Media, 2007.

Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. Efficient optimal learning for contextual bandits. *arXiv preprint arXiv:1106.2369*, 2011.

Garcia, J., Ervin, F. R., and Koelling, R. A. Learning with prolonged delay of reinforcement. *Psychonomic Science*, 5(3):121–122, 1966.

Garg, S. and Akash, A. K. Stochastic bandits with delayed composite anonymous feedback. *arXiv preprint arXiv:1910.01161*, 2019.

Joulani, P., Gyorgy, A., and Szepesvári, C. Online learning under delayed feedback. In *International Conference on Machine Learning*, pp. 1453–1461, 2013.

Joulani, P., Gyorgy, A., and Szepesvári, C. Delay-tolerant online convex optimization: Unified analysis and adaptive-gradient algorithms. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1): 4–22, 1985.

Langford, J., Smola, A. J., and Zinkevich, M. Slow learners are fast. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, pp. 2331–2339, 2009.

Lattimore, T. and Szepesvári, C. Bandit algorithms. *preprint*, 2018.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. 2019. URL `http://downloads.tor-lattimore.com/book.pdf`.

Mandel, T., Liu, Y.-E., Brunskill, E., and Popović, Z. The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Mann, T. A., Gowal, S., Jiang, R., Hu, H., Lakshminarayanan, B., and Gyorgy, A. Learning from delayed outcomes with intermediate observations. *arXiv preprint arXiv:1807.09387*, 2018.

McMahan, B. and Streeter, M. Delay-tolerant algorithms for asynchronous distributed online learning. In *Advances in Neural Information Processing Systems*, pp. 2915–2923, 2014.

Pike-Burke, C., Agrawal, S., Szepesvari, C., and Grunewalder, S. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pp. 4102–4110, 2018.

Quanrud, K. and Khashabi, D. Online learning with adversarial delays. In *Advances in neural information processing systems*, pp. 1270–1278, 2015.

Sra, S., Yu, A. W., Li, M., and Smola, A. J. Adadelay: Delay adaptive distributed stochastic convex optimization. *arXiv preprint arXiv:1508.05003*, 2015.

Thune, T. S., Cesa-Bianchi, N., and Seldin, Y. Nonstochastic multiarmed bandits with unrestricted delays. *arXiv preprint arXiv:1906.00670*, 2019.

Vernade, C., Cappé, O., and Perchet, V. Stochastic bandit models for delayed conversions. *arXiv preprint arXiv:1706.09186*, 2017.

Vernade, C., Carpentier, A., Zappella, G., Ermis, B., and Brueckner, M. Contextual bandits under delayed feedback. *arXiv preprint arXiv:1807.02089*, 2018.

Weinberger, M. J. and Ordentlich, E. On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7):1959–1976, 2002.

Yoshikawa, Y. and Imai, Y. A nonparametric delayed feedback model for conversion rate prediction. *arXiv preprint arXiv:1802.00255*, 2018.

Zhou, Z., Xu, R., and Blanchet, J. Learning in generalized linear contextual bandits with stochastic delays. In *Advances in Neural Information Processing Systems 32*, pp. 5198–5209, 2019.

A. Proof of Theorems 1

The aim of this section is to bound the deviation of the estimator $\widehat{\mu}_i$ from the true mean μ_i. For this purpose, we first begin by defining the favorable event, that is, the event for which all confidence intervals hold for all arms at all time steps. We then prove with Hoeffding's inequality that this event occurs with high probability. And finally, we derive the desired results by bounding the bias incurred by $\widehat{\mu}_i$ and leveraging the properties of this favorable event.

Step 1: the favorable event. First let us define the following quantities (for $i \leq K$ and $u \leq T_i(T)$):

$$\bar{t}_i(u) = \inf\{t \geq 0 : \sum_{u \leq t} \mathbf{1}\{I_u = i\} = u\},$$

$$\overline{C}_{i,u} = C_{\bar{t}_i(u)},$$

$$\overline{D}_{i,u} = D_{\bar{t}_i(u)}.$$

Here $\bar{t}_i(u)$ is the time where we pulled arm i for the u-th time, $\overline{C}_{i,u}, \overline{D}_{i,u}$ are respectively the corresponding reward and delay. Note that

$$\sum_{u=1}^{t} X_{u,t-u} \mathbf{1}\{I_u = i\} = \sum_{u=1}^{T_i(t)} \overline{C}_{i,u} \mathbf{1}\{\bar{t}_i(u) + \overline{D}_{i,u} \leq t\}. \tag{7}$$

We define the event ξ as follows;

$$\xi \triangleq \left\{ \forall i \in \{1, ..., K\}, \forall t \in \{1, ..., T\}, \forall s \in \{1, ..., T_i(t)\} : \left| \sum_{u=1}^{s} \overline{C}_{i,u} \mathbf{1}\{\bar{t}_i(u) + \overline{D}_{i,u} \leq t\} - \sum_{u=1}^{s} \tau_i(t - \bar{t}_i(u))\mu_i \right| \leq \sqrt{2 \log \frac{2}{\delta} s} \right\}.$$

Note that for fixed $i \in \{1, ..., K\}, t \in \{1, ..., T\}, s \in \{1, ..., T_i(t)\}$, we have that

$$\left(\sum_{u \leq v} \left[\overline{C}_{i,u} \mathbf{1}\{\bar{t}_i(u) + \overline{D}_{i,u} \leq t\} - \tau_i(t - \bar{t}_i(u))\mu_i \right] \right)_{v \leq s}$$

is a martingale adapted to the filtration $\left(\sigma(\overline{C}_{i,v}, \overline{D}_{i,v}, \bar{t}_i(v))\right)_{v \leq s}$. And since the martingale increments $\left[\overline{C}_{i,u} \mathbf{1}\{\bar{t}_i(u) + \overline{D}_{i,u} \leq t\} - \tau_i(t - \bar{t}_i(u))\mu_i\right]$ belong to $[-1, 1]$ by assumption, it holds by Azuma-Hoeffding's inequality that with probability larger than $1 - \delta$

$$\left| \sum_{u \leq s} \left[\overline{C}_{i,u} \mathbf{1}\{\bar{t}_i(u) + \overline{D}_{i,u} \leq t\} - \tau_i(t - \bar{t}_i(u))\mu_i \right] \right| \leq \sqrt{2 \log \frac{2}{\delta} s}.$$

Since $T_i(t) \leq t$, it holds by a union bound that

$$\mathbb{P}(\xi) \geq 1 - KT^2\delta. \tag{8}$$

Step 2: Bound on the $|\widehat{\mu}_i(t) - \mu_i|$ on ξ. By Equation 7, it holds that

$$\xi \subset \left\{ \forall i \in \{1, ..., K\}, \forall t \in \{1, ..., T\}, \left| \widehat{\mu}_i(t) - \frac{1}{T_i(t)} \sum_{u=1}^{t} \tau_i(t-u)\mu_i \mathbf{1}\{I_u = i\} \right| \leq \sqrt{\frac{2 \log \frac{2}{\delta}}{T_i(t)}} \right\}. \tag{9}$$

Note that we have

$$\begin{aligned} \left| \frac{1}{T_i(t)} \sum_{u=1}^{t} \tau_i(t-u)\mu_i \mathbf{1}\{I_u = i\} - \mu_i \right| &\leq \frac{1}{T_i(t)} \sum_{u=1}^{t} |(1 - \tau_i(t-u))\mu_i \mathbf{1}\{I_u = i\}| \\ &\leq \frac{1}{T_i(t)} \sum_{u=1}^{t} \mathbf{1}\{I_u = i\}((t-u) \vee 1)^{-\alpha}, \end{aligned}$$

since $0 \leq \mu_i \leq 1$ and since by Assumption 1 it holds that $|\tau_i(m) - 1| \leq (m \vee 1)^{-\alpha}$. And since $\sum_{u=1}^{t} \mathbf{1}\{I_u = i\} = T_i(t)$, we have

$$\Big|\frac{1}{T_i(t)}\sum_{u=1}^{t}\tau_i(t-u)\mu_i\mathbf{1}\{I_u=i\} - T_i(t)\mu_i\Big| \leq \frac{1}{T_i(t)}\sum_{v=0}^{T_i(t)}(v\vee 1)^{-\alpha}$$
$$\leq \frac{1}{T_i(t)}\left[2 + \int_1^{T_i(t)} v^{-\alpha}dv\right], \tag{10}$$

whenever $\alpha \leq 1/2$:

$$\frac{1}{T_i(t)}\Big[2 + \int_1^{T_i(t)} v^{-\alpha}dv\Big] = \frac{1}{T_i(t)}\Big[2 + [\frac{-1}{1-\alpha}v^{1-\alpha}]_1^{T_i(t)}\Big]$$

$$= \frac{1}{T_i(t)}\Big[2 + \frac{1}{1-\alpha}[T_i(t)^{1-\alpha} - 1]\Big]$$
$$= \frac{1}{T_i(t)}\Big[-\frac{2\alpha}{1-\alpha} + \frac{1}{1-\alpha}T_i(t)^{1-\alpha}\Big]$$
$$\leq \frac{1}{T_i(t)}\Big[2T_i(t)^{1-\alpha}\Big]$$
$$\leq 2T_i(t)^{-\alpha} = 2T_i(t)^{-\alpha\wedge(1/2)}.$$

Now for $\alpha \geq 1/2$:

$$\frac{1}{T_i(t)}\Big[2 + \int_1^{T_i(t)} v^{-\alpha}dv\Big] \leq \frac{1}{T_i(t)}\Big[2 + \int_1^{T_i(t)} v^{-1/2}dv\Big]$$
$$= \frac{1}{T_i(t)}\Big[2 + 2[T_i(t)^{1/2} - 1]\Big] = 2T_i(t)^{1/2} = 2T_i(t)^{-\alpha\wedge(1/2)}.$$

Note that we have

$$|\widehat{\mu}_i(t) - \mu_i| = \Big|\widehat{\mu}_i(t) - \frac{1}{T_i(t)}\sum_{u=1}^{t}\tau_{t-u}\mu_i\mathbf{1}\{I_u=i\} + \frac{1}{T_i(t)}\sum_{u=1}^{t}\tau_{t-u}\mu_i\mathbf{1}\{I_u=i\} - \mu_i\Big|$$
$$\leq \Big|\widehat{\mu}_i(t) - \frac{1}{T_i(t)}\sum_{u=1}^{t}\tau_{t-u}\mu_i\mathbf{1}\{I_u=i\}\Big| + \Big|\frac{1}{T_i(t)}\sum_{u=1}^{t}\tau_{t-u}\mu_i\mathbf{1}\{I_u=i\} - \mu_i\Big|.$$

Now from the definition of the favorable event ξ in Equation (9) and from the bound of the bias in Equation (10), it holds on ξ that:

$$|\widehat{\mu}_i(t) - \mu_i| \leq \left(\frac{2\log\frac{2}{\delta}}{T_i(t)}\right)^{1/2} + 2T_i(t)^{-\alpha\wedge 1/2}. \tag{11}$$

B. Proof of Theorems 2 and 3

The proof of these theorems relies on the results obtained in Appendix A. Indeed, we will use the deviation results obtained on the event ξ to bound the number of pulls of sub-optimal arms in order to derive the upper bounds on the regret.

Step 1: upper bound on the number of pulls of sub-optimal arms. Let's assume that at some given time $t+1 > K$ the algorithm pulls a sub-optimal arm i (such that $\mu_i < \mu^*$). According to the algorithm's rules, we have: $UCB_i(t+1) \geq UCB_{k^*}(t+1)$. And so on ξ, we have because of Equation (11)

$$\mu^* \leq UCB_{k^*}(t+1) \leq UCB_i(t+1) \leq \mu_i + 2\left(\frac{2\log\frac{2}{\delta}}{T_i(t)}\right)^{1/2} + 4T_i(t)^{-\alpha\wedge 1/2}.$$

Rearranging the terms, we have on ξ:

$$\Delta_i = \mu^* - \mu_i \leq 2\left(\frac{2\log\frac{2}{\delta}}{T_i(t)}\right)^{1/2} + 4T_i(t)^{-\alpha\wedge 1/2},$$

which implies that on ξ

$$T_i(t) \leq \frac{16\log(2/\delta)}{\Delta_i^2} \vee \left(\frac{8}{\Delta_i}\right)^{-\frac{1}{\alpha}\vee 2} \vee 1,$$

and so on ξ, we have for any sub-optimal arm i

$$T_i(T) \leq \frac{16\log(2/\delta)}{\Delta_i^2} \vee \left(\frac{8}{\Delta_i}\right)^{\frac{1}{\alpha}\vee 2} \vee 1. \tag{12}$$

Step 2: Conclusion. Consider a sub-optimal arm i (such that $\mu_i < \mu^*$). Combining Equation(12) with Equation (8), and since $T_i(T) \leq T$ we have

$$\begin{aligned}\mathbb{E}[T_i(T)] &\leq \frac{16\log(2/\delta)}{\Delta_i^2} \vee \left(\frac{8}{\Delta_i}\right)^{\frac{1}{\alpha}\vee 2} \vee 1 + KT^3\delta \\ &\leq \frac{16\log(2KT^3)}{\Delta_i^2} \vee \left(\frac{8}{\Delta_i}\right)^{\frac{1}{\alpha}\vee 2} \vee 1 + 1,\end{aligned}$$

for $\delta \triangleq (KT^3)^{-1}$. Let us now consider some value $\Delta > 0$. We have by definition of the regret and with this $\delta \triangleq (KT^3)^{-1}$ as above:

$$\overline{R}_T \leq \sum_{i:\Delta_i>\Delta} \Delta_i\Big[\frac{16\log(2KT^3)}{\Delta_i^2} \vee \left(\frac{8}{\Delta_i}\right)^{\frac{1}{\alpha}\vee 2} \vee 1 + 1\Big] + \Delta \sum_{i:\Delta_i\leq\Delta} \mathbb{E}[T_i(T)]. \tag{13}$$

Taking $\Delta = 0$ and recalling that $K \leq T$, we obtain the result of Theorem 2, namely

$$\overline{R}_T \leq \sum_{i:\Delta_i>0} \Big[\frac{64\log(2T)}{\Delta_i} \vee \left(\frac{8}{\Delta_i}\right)^{\frac{1-\alpha}{\alpha}\vee 1}\Big] + 2K.$$

Now not that since the function $\Delta_i\Big[\frac{16\log(2KT^3)}{\Delta_i^2} \vee \left(\frac{8}{\Delta_i}\right)^{\frac{1}{\alpha}\vee 2} \vee 1 + 1\Big]$ increases when Δ_i decreases, we have for any $\Delta > 0$

$$\begin{aligned}\overline{R}_T &\leq K\Big[\frac{16\log(2KT^3)}{\Delta} \vee \left(\frac{8}{\Delta}\right)^{\frac{1-\alpha}{\alpha}\vee 1}\Big] + \Delta T + 2K \\ &\leq K\log(2T)\Big(\frac{64}{\Delta}\Big)^{\frac{1-\alpha}{\alpha}\vee 1} + \Delta T + 2K.\end{aligned}$$

And so for $\Delta = \left(\frac{K\log(2T)64^{\frac{1-\alpha}{\alpha}\vee 1}}{T}\right)^{\alpha\wedge 1/2}$, we have

$$\overline{R}_T \leq 2\times 64^{(1-\alpha)\vee 1/2}T^{1-\alpha\wedge 1/2}\Big(K\log(2T)\Big)^{\alpha\wedge 1/2} + 2K,$$

which concludes the proof of Theorem 3.

C. Proof of Theorem 4

In order to prove this result, we need to show that there exists a bandit problem in the family described in Section 2 such that the regret at T is $\Omega(T^{1-\alpha})$. To do so, we construct two problems in that family and show that for at least one of them, the regret is larger than the desired quantity.

First, for ease of notation, define

$$p = T^{-\alpha} \quad \text{and} \quad q = \frac{p}{4-2p}. \tag{14}$$

We construct two alternative problems with two arms, $K = 2$. In both cases, we fix arm 1 such that the distribution of the rewards ν_1 is $\mathcal{B}(1/2)$ and the distribution of the delays $\mathcal{D}_1$ is δ_0 (i.e. a Dirac mass in 0, no delays).

- Problem A: $\nu_2^{(A)}$ is $\mathcal{B}(1/2 - q)$ and $\mathcal{D}_2^{(A)}$ is δ_0.
- Problem B: $\nu_2^{(B)}$ is $\mathcal{B}(1/2 + q)$ and $\mathcal{D}_2^{(B)}$ is $(1-p)\delta_0 + p\delta_T$.

In Problem A, arm 1 is the best with gap $\Delta = q$ and there are no delays. In Problem B, arm 2 is the best, with gap $\Delta = q$ too, but delays are sending a proportion p of the rewards to $t \geq T$ so they cannot be used for learning. Thus, the conditional distribution of $X_{s,u}|I_s = 2$ is in fact $\mathcal{B}((1/2+q)(1-p))$. Note that

$$\left(\frac{1}{2} + q\right)(1-p) = \frac{1}{2} - \left(\frac{p}{2} - q + pq\right) = \frac{1}{2} - \left(\frac{p}{2} - \frac{p}{4-2p} + \frac{p^2}{4-2p}\right) = \frac{1}{2} - \frac{p}{4-2p} = \frac{1}{2} - q.$$

So the effective mean of arm 2 is $1/2 - q$, meaning that arm 2 has the same distribution in both problems. This implies in particular that

$$\mathbb{E}_A T_2(T) = \mathbb{E}_B T_2(T),$$

where $\mathbb{E}_a$ is the expectation in problem $a \in \{A, B\}$.

And so if we write $\overline{R}_T^{(a)}$ for the regret in scenario $a \in \{A, B\}$:

$$\max_{a \in \{1,2\}} \overline{R}_T^{(a)} \geq q \max(T - \mathbb{E}_B T_2(T), \mathbb{E}_A T_2(T)) \geq qT/2.$$

This concludes the proof as $q = p/(4-2p) \geq p/4 = T^{-\alpha}/4$.

D. Proof of Theorem 5

The proof of this theorem relies on the same tools as for Theorem 4 above, but uses a slightly different reasoning. Namely, we now fix $\alpha > 0$ and we restrict the family of algorithms to those that have a regret smaller than $T^{1-\alpha}/8$ for any stochastic bandit problem satisfying Assumption 1 for α. We denote this family $\mathcal{A}_\alpha$.

We want to prove that there exists a bandit problem satisfying Assumption 1 for some $\alpha' > \alpha$ such that any algorithm in $\mathcal{A}_\alpha$ has regret at least $T^{1-\alpha}/8 > T^{1-\alpha'}/8$. This proves that any algorithm minimax optimal for α is suboptimal for $\alpha' > \alpha$.

Similarly to the previous section, fix $p = T^{\alpha}$ and $q = p/(4-2p)$ as in Eq. (14), and consider the two problems,

- Problem A: $\nu_2^{(A)}$ is $\mathcal{B}(1/2 - q)$ and $\mathcal{D}_2^{(A)}$ is δ_0.
- Problem B: $\nu_2^{(B)}$ is $\mathcal{B}(1/2 + q)$ and $\mathcal{D}_2^{(B)}$ is $(1-p)\delta_0 + p\delta_T$.

Note that, Problem B satisfies Assumption 1 for α, while Problem A satisfies it for any $\alpha' > 0$ so in particular for $\alpha' > \alpha$. So for any algorithm in $\mathcal{A}_\alpha$, $q\mathbb{E}_B[T_1(t)] < 3T^{1-\alpha}/16$. But, as we proved above, because of the delays, the algorithm cannot distinguish both problems and we have $\mathbb{E}_A T_1(T) = \mathbb{E}_B T_1(T)$, i.e. the average number of pulls of arm 1 is the same in both problems. Thus,

$$\mathbb{E}_A T_1(T) = \mathbb{E}_B[T_1(t)] < q^{-1}T^{1-\alpha}/8.$$

Using that $q > p/4$ as before,

$$\max_{a\in\{A,B\}} \overline{R}_T^{(a)} \geq q(T - \mathbb{E}_A T_1(T)) > \frac{T^{-\alpha}}{4}T - T^{1-\alpha}/8 = T^{1-\alpha}/8.$$

E. Proof of Theorem 6

We start by stating the full version of the theorem that guarantees the performance of `Adapt-PatientBandits`.

Theorem 7. *Let $T > K \geq 1$ and $\alpha, \underline{\alpha}, c, \overline{\mu} > 0$, such that Assumption 2 holds. The regret of* `Adapt-PatientBandits` is bounded as

$$\overline{R}_T \leq 8^{17}\left(\frac{1}{c\overline{\mu}}\right)^4\left(\frac{2}{c}\right)^{4(\alpha\wedge 1/2)/\underline{\alpha}}\log(2T)^{13/2}T(K/T)^{\alpha\wedge(1/2)}.$$

Before proving Thereom 6, we first provide the following proposition that bounds the error on our estimator of α.

Proposition 1. *Let $\delta \in (0,1)$. There exists an event of probability larger than $1 - 2KT^2\delta$ such that for any $t \leq T$,*

$$\alpha\wedge 1/2 - \frac{\log\left(2^3\left(\left(\frac{2}{c}\right)^{(\alpha\wedge 1/2)/\underline{\alpha}} + B\right)\right)}{\log(\overline{T}_t)} \leq \widehat{\alpha}_t \leq \alpha\wedge 1/2 + \frac{\log\left(\frac{2^{7/2}B}{c\overline{\mu}}\right)}{\log\overline{T}_t},$$

where $B \triangleq B_\delta \triangleq \sqrt{2\log(2/\delta)}$.

Proof of Proposition 1. For a lighter notation we set $\overline{\mu}_t \triangleq \mu_{\overline{I}_t}$. Similarly to the analysis of the subset of ξ in Equation 9, we can prove that the event

$$\xi' \triangleq \left\{\forall t \leq T,\ |\overline{m}_{t,D_t} - \overline{\mu}_t\tau_{\overline{I}_t}(D_t)| \leq \sqrt{\frac{2\log(2/\delta)}{\overline{T}(t-D_t)}},\ |\overline{m}_{t,d_t} - \overline{\mu}_t\tau_{\overline{I}_t}(d_t)| \leq \sqrt{\frac{2\log(2/\delta)}{\overline{T}(t-d_t)}}\right\}$$

has probability larger than $1 - 2KT^2\delta$. Let us set

$$B \triangleq B_\delta \triangleq \sqrt{2\log(2/\delta)}.$$

Since $d_t \leq D_t$, by Assumption 2 we have that on ξ',

$$c\overline{\mu}_t d_t^{-\alpha} - \overline{\mu}_t D_t^{-\alpha} - \frac{2B}{\sqrt{\overline{T}(t-D_t)}} \leq \overline{m}_{t,D_t} - \overline{m}_{t,d_t} \leq \overline{\mu}_t d_t^{-\alpha} + \frac{2B}{\sqrt{\overline{T}(t-D_t)}}.$$

We now chose $d_t \triangleq \left\lfloor \left(\frac{c}{2}\right)^{1/\underline{\alpha}} D_t \right\rfloor$, for which we have that $c\overline{\mu}_t d_t^{-\alpha} - \overline{\mu}_t D_t^{-\alpha} \geq \frac{c}{2}\overline{\mu}_t d_t^{-\alpha}$ and therefore on ξ',

$$\frac{c}{2}\overline{\mu}_t d_t^{-\alpha} - \frac{2B}{\sqrt{\overline{T}(t-D_t)}} \leq \overline{m}_{t,D_t} - \overline{m}_{t,d_t} \leq \overline{\mu}_t d_t^{-\alpha} + \frac{2B}{\sqrt{\overline{T}(t-D_t)}}.$$

Here we have that $D_t = \lfloor \overline{T}_t/2 \rfloor$, and since $\overline{T}_t \geq 2$, we obtain

$$\overline{T}_t^{-\alpha} \leq D_t^{-\alpha} \leq 2^{2\alpha}\overline{T}_t^{-\alpha}.$$

Moreover, since we chose $d_t = \left\lfloor \left(\frac{c}{2}\right)^{1/\underline{\alpha}} D_t \right\rfloor$, we infer that

$$T_t^{-\alpha} \leq d_t^{-\alpha} \leq 2^{3\alpha}\left(\frac{2}{c}\right)^{\alpha/\underline{\alpha}}\overline{T}_t^{-\alpha}.$$

Therefore, since $\overline{T}(t-D_t) \leq \overline{T}_t \leq 2\overline{T}(t-D_t)$, we have that on ξ',

$$\frac{c}{2}\overline{\mu}\overline{T}_t^{-\alpha} - 2^{3/2}B\overline{T}_t^{-1/2} \leq \overline{m}_{t,D_t} - \overline{m}_{t,d_t} \leq \left(2^{3(\alpha\wedge 1/2)}\left(\frac{2}{c}\right)^{(\alpha\wedge 1/2)/\underline{\alpha}} + 2^{3/2}B\right)\overline{T}_t^{-\alpha\wedge 1/2} \triangleq C_\alpha\overline{T}_t^{-\alpha\wedge 1/2},$$

where we use the fact that the μ_k are in $[\overline{\mu}, 1]$.

First case — small α First, consider the case where $\frac{c}{4}\overline{\mu}\overline{T}_t^{-\alpha} \geq 2^{3/2}B\overline{T}_t^{-1/2}$. Then we have that on event ξ'

$$\frac{c}{4}\overline{\mu}\overline{T}_t^{-\alpha} \leq \overline{m}_{t,D_t} - \overline{m}_{t,d_t} \leq C_\alpha \overline{T}_t^{-\alpha\wedge 1/2},$$

which implies that on ξ',

$$-\log\left(\frac{c\overline{\mu}}{4}\right) + \alpha\log(\overline{T}_t) \geq -\log(\overline{m}_{t,D_t} - \overline{m}_{t,d_t}) \geq -\log(C_\alpha) + (\alpha\wedge 1/2)\log(\overline{T}_t).$$

Therefore, on ξ',

$$\alpha - \frac{\log(\frac{c\overline{\mu}}{4})}{\log(\overline{T}_t)} \geq -\frac{\log(\overline{m}_{t,D_t} - \overline{m}_{t,d_t})}{\log(\overline{T}_t)} \geq \alpha\wedge 1/2 - \frac{\log(C_\alpha)}{\log(\overline{T}_t)},$$

from which we finally get that on ξ',

$$\alpha\wedge 1/2 + \frac{\log(\frac{4}{c\overline{\mu}})}{\log(\overline{T}_t)} \geq \widehat{\alpha}_t \geq \alpha\wedge 1/2 - \frac{\log(C_\alpha)}{\log(\overline{T}_t)}. \tag{15}$$

Note that while the left hand side of the above inequality *only true under the assumption of the first case* that demands $\frac{c}{4}\overline{\mu}\overline{T}_t^{-\alpha} \geq 2^{3/2}B\overline{T}_t^{-1/2}$, the right hand side is also true when this assumption does not hold since we did not use it.

Second case — large α Now consider the case where $\frac{c}{4}\overline{\mu}\overline{T}_t^{-\alpha} \leq 2^{3/2}B\overline{T}_t^{-1/2}$. In this case it holds that

$$\overline{T}_t^{\alpha-1/2} \geq \frac{c\overline{\mu}}{2^{7/2}B} \triangleq b^{-1},$$

which means that

$$\alpha - 1/2 \geq -\frac{\log(b)}{\log\overline{T}_t},$$

and therefore,

$$\alpha\wedge 1/2 \geq 1/2 - \frac{\log(b)}{\log\overline{T}_t}.$$

Now by definition of $\widehat{\alpha}_t$, we have

$$\widehat{\alpha}_t \leq 1/2 \leq \alpha\wedge 1/2 + \frac{\log(b)}{\log\overline{T}_t}.$$

Taking only the right hand side of the first case in Equation 15, which does *not* use the assumption of the first case, unlike the left hand side (cf. the remark under Equation 15), we have that on ξ,

$$\widehat{\alpha}_t \geq \alpha\wedge 1/2 - \frac{\log(C_\alpha)}{\log(\overline{T}_t)}.$$

combining the two sides of the bound, we get that on ξ,

$$\alpha\wedge 1/2 + \frac{\log\left(\frac{2^{7/2}B}{c\overline{\mu}}\right)}{\log\overline{T}_t} \geq \widehat{\alpha}_t \geq \alpha\wedge 1/2 - \frac{\log(C_\alpha)}{\log(\overline{T}_t)}.$$

Notice that this inequality holds also in the first case (small α) since $4/(c\overline{\mu}) \leq \frac{2^{7/2}B}{c\overline{\mu}}$. The final result follows immediately since we have that $C_\alpha \leq 2^3\left(\left(\frac{2}{c}\right)^{(\alpha\wedge 1/2)/\underline{\alpha}} + B\right)$. □

Now leveraging these concentration bounds obtained on the error of the estimation of α as well as the theoretical results from Theorem 3 we prove the main result.

Proof of Theorem 6. Recall that, from Proposition 1 with probability larger than $1 - 2KT^2\delta$ we have that,

$$\alpha \wedge 1/2 + \frac{\log\left(\frac{2^{7/2}B}{c\overline{\mu}}\right)}{\log \overline{T}_t} \geq \widehat{\alpha}_t \geq \alpha \wedge 1/2 - \frac{\log\left(2^3\left(\left(\frac{2}{c}\right)^{(\alpha\wedge 1/2)/\underline{\alpha}} + B\right)\right)}{\log(\overline{T}_t)},$$

where $B \triangleq B_\delta = \sqrt{2\log(2/\delta)}$.
Let $L_\alpha \triangleq \log\left(\frac{2^{7/2}}{c\overline{\mu}}\right) + \log\left(2^3\left(\left(\frac{2}{c}\right)^{(\alpha\wedge 1/2)/\underline{\alpha}} + 1\right)\right)$. With probability larger than $1 - 2KT^2\delta$,

$$U_t \triangleq \frac{\log\left(\frac{2^{7/2}B}{c\overline{\mu}}\right) + \log\left(2^3\left(\left(\frac{2}{c}\right)^{(\alpha\wedge 1/2)/\underline{\alpha}} + B\right)\right)}{\log(\overline{T}_t)} \leq \frac{L_\alpha + \log(B)}{\log(\overline{T}_t)}$$

is a high probability lower deviation on the lower bound $\overline{\alpha}_t$ on α. Note that $\overline{T}_t$ is the number of pulls of the most pulled arm at time t and therefore $\overline{T}_t \geq t/K$. Furthermore, note that after the initialisation phase, we have guaranteed that $t \geq 2K$ from which we get

$$U_t \leq \frac{2L_\alpha + 2\log(B)}{\log(t)}.$$

Moreover, by Theorem 3, we have that conditionally on the event ξ from Proposition 1, the expected regret coming from the samples pulled *after* round t can be bounded as

$$128\sqrt{\log(2T)}T(K/T)^{\alpha\wedge(1/2)-U_t} + 2K.$$

The above bound implies that conditional on the event ξ' from Proposition 1, the expected regret due to the samples obtained after round $t = T^{1/2}$ can be bounded as

$$128\sqrt{\log(2T)}T(K/T)^{\alpha\wedge(1/2)-4(L_\alpha+\log B)/\log T} + 2K.$$

Now setting $\delta \triangleq (KT^3)^{-1}$ in the algorithm - where δ is used to define the event ξ' in Proposition 1 - we have on ξ' that,

$$\overline{R}_T \leq T^{1/2} + 128\sqrt{\log(2T)}T(K/T)^{\alpha\wedge(1/2)-4\left(L_\alpha+\sqrt{2}\log\left(2\log(2KT^3)\right)\right)/\log T} + 4K,$$

where the additional regret $T^{1/2}$ comes from the first $T^{1/2}$ samples to the bound computed above, and where the additional regret $2K$ comes from the case when the event ξ' from Proposition 1 does not hold - leading to a supplementary term bounded by $\mathbb{P}((\xi')^c)T \leq 2K$.

Notice that in the expression in the regret bound above,

$$T^{4\left(L_\alpha+\sqrt{2}\log\left(2\log(2KT^3)\right)\right)/\log T} \leq e^{4L_\alpha}(8\log(2T))^{4\sqrt{2}} \leq 8^{12}\left(\frac{1}{c\overline{\mu}}\right)^4\left(\frac{2}{c}\right)^{4(\alpha\wedge 1/2)/\underline{\alpha}}\log(2T)^6$$

and therefore we can simplify our guarantee to finally obtain

$$\overline{R}_T \leq T^{1/2} + 8^{16}\left(\frac{1}{c\overline{\mu}}\right)^4\left(\frac{2}{c}\right)^{4(\alpha\wedge 1/2)/\underline{\alpha}}\log(2T)^{13/2}T(K/T)^{\alpha\wedge(1/2)} + 4K.$$

□