Proceedings of Machine Learning Research vol 195:1–68, 2023

A Blackbox Approach to Best of Both Worlds in Bandits and Beyond

Christoph Dann CDANN@CDANN.NET
Google Research

Chen-Yu Wei CHENYUW@MIT.EDU
MIT Institute for Data, Systems, and Society

Julian Zimmert ZIMMERT@GOOGLE.COM
Google Research

Abstract

Best-of-both-worlds algorithms for online learning which achieve near-optimal regret in both the adversarial and the stochastic regimes have received growing attention recently. Existing techniques often require careful adaptation to every new problem setup, including specialised potentials and careful tuning of algorithm parameters. Yet, in domains such as linear bandits, it is still unknown if there exists an algorithm that can simultaneously obtain $O(\log(T))$ regret in the stochastic regime and $\tilde{O}(\sqrt{T})$ regret in the adversarial regime. In this work, we resolve this question positively and present a general reduction from best of both worlds to a wide family of follow-the-regularized-leader (FTRL) and online-mirror-descent (OMD) algorithms. We showcase the capability of this reduction by transforming existing algorithms that are only known to achieve worst-case guarantees into new algorithms with best-of-both-worlds guarantees in contextual bandits, graph bandits and tabular Markov decision processes.

1. Introduction

Multi-armed bandits and its various extensions have a long history (Lai et al., 1985; Auer et al., 2002a;b). Traditionally, the stochastic regime in which all losses/rewards are i.i.d. and the adversarial regime in which an adversary chooses the losses in an arbitrary fashion have been studied in isolation. However, it is often unclear in practice whether an environment is best modelled by the stochastic regime, a slightly contaminated regime, or a fully adversarial regime. This is why the question of automatically adapting to the hardness of the environment, also called achieving *best-of-both-worlds* guarantees, has received growing attention (Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014; Auer and Chiang, 2016; Seldin and Lugosi, 2017; Wei and Luo, 2018; Zimmert and Seldin, 2019; Ito, 2021b; Ito et al., 2022a; Masoudian and Seldin, 2021; Gaillard et al., 2014; Mourtada and Gaïffas, 2019; Ito, 2021a; Lee et al., 2021; Rouyer et al., 2021; Amir et al., 2022; Huang et al., 2022; Tsuchiya et al., 2022; Erez and Koren, 2021; Rouyer et al., 2022; Ito et al., 2022b; Kong et al., 2022; Jin and Luo, 2020; Jin et al., 2021; Chen et al., 2022; Saha and Gaillard, 2022; Masoudian et al., 2022; Honda et al., 2023). One of the most successful approaches has been to derive carefully tuned FTRL algorithms, which are canonically suited for the adversarial regime, and then show that these algorithms are also close to optimal in the stochastic regime as well. This is achieved via proving a crucial self-bounding property, which then translates into stochastic guarantees. This type of algorithms have been proposed for multi-armed bandits (Wei and Luo, 2018; Zimmert and Seldin, 2019; Ito, 2021b; Ito et al., 2022a), combinatorial semi-bandits (Zimmert et al., 2019), bandits with graph feedback (Ito et al., 2022b), tabular MDPs (Jin and Luo, 2020; Jin et al., 2021) and others. Algorithms with self-bounding properties also automatically adapt to

Setting	Algorithm	Adversarial	Stochastic	$o(C)$	Rd. 1	Rd. 2
Linear bandit	Lee et al. (2021)	$\sqrt{dT\log(\lvert\mathcal{X}\rvert T)\log T}$	$\frac{d\log(\lvert\mathcal{X}\rvert T)\log(T)}{\Delta}$			
	Theorem 3	$d\sqrt{\nu L_\star \log T}$	$\frac{d^2\nu\log T}{\Delta}$	✓		
	Corollary 7	$d\sqrt{T\log T}$	$\frac{d^2\log T}{\Delta}$	✓	✓	
	Corollary 12	$\sqrt{dT\log\lvert\mathcal{X}\rvert}$	$\frac{d\log\lvert\mathcal{X}\rvert\log T}{\Delta}$	✓	✓	✓
Contextual bandit	Corollary 13	$\sqrt{KT\log\lvert\mathcal{X}\rvert}$	$\frac{K\log\lvert\mathcal{X}\rvert\log T}{\Delta}$	✓	✓	✓
Graph bandit Strongly observable	Ito et al. (2022b)	$\sqrt{\alpha T\log T\log(KT)}$	$\frac{\alpha\log^2(KT)\log T}{\Delta}$	✓		
	Rouyer et al. (2022)	$\sqrt{\widetilde{\alpha}T\log K}$	$\frac{\widetilde{\alpha}\log(KT)\log T}{\Delta}$			
	Corollary 15	$\sqrt{\min\{\widetilde{\alpha},\alpha\log K\}T\log K}$	$\frac{\min\{\widetilde{\alpha},\alpha\log K\}\log K\log T}{\Delta}$	✓	✓	✓
Graph bandit Weakly observable	Ito et al. (2022b)	$(\delta\log(KT)\log T)^{\frac{1}{3}}T^{\frac{2}{3}}$	$\frac{\delta\log(KT)\log T}{\Delta^2}$	✓		
	Corollary 19	$(\delta\log K)^{\frac{1}{3}}T^{\frac{2}{3}}$	$\frac{\delta\log K\log T}{\Delta^2}$	✓	✓	✓
Tabular MDP	Jin et al. (2021)	$\sqrt{H^2S^2AT\log^2 T}$	$\frac{H^6S^2A\log^2 T}{\Delta'}$	✓		
	Theorem 26	$\sqrt{HS^2AL_\star\log^4 T}$	$\frac{H^2S^2A\log^3 T}{\Delta}$	✓	✓	✓

Table 1: Overview of regret bounds. The $o(C)$ column specifies whether the algorithm achieves the optimal dependence on the amount of corruption ($\sqrt{C}$ or $C^{\frac{2}{3}}$, depending on the setting). "Rd. 1" and "Rd. 2" indicate whether the result leverages the first and the second reductions described in Section 4, respectively. $L_\star$ is the cumulative loss of the best action or policy. ν in Theorem 3 is the self-concordance parameter; it holds that $\nu \leq O(d)$ (see Appendix B for details). Δ' in Jin et al. (2021) is the gap in the $Q^\star$ function, which is different from the policy gap Δ we use; it holds that $\Delta \leq \Delta'$. $\alpha, \widetilde{\alpha}, \delta$ are complexity measures of feedback graphs defined in Section 5.

intermediate regimes of stochastic losses with adversarial corruptions (Lykouris et al., 2018; Gupta et al., 2019; Zimmert and Seldin, 2019; Ito, 2021a), which highlights the strong robustness of this algorithm design. However, every problem variation required careful design of potential functions and learning rate schedules. For linear bandits, it has been still unknown whether self-bounding is possible and therefore the state-of-the-art best-of-both-worlds algorithm (Lee et al., 2021) neither obtains optimal $\log T$ stochastic rate, nor canonically adapts to corruptions.

In this work, the make the following contributions: 1) We propose a general reduction from best of both worlds to typical FTRL/OMD algorithms, sidestepping the need for customized potentials and learning rates. 2) We derive the first best-of-both-worlds algorithm for linear bandits that obtains $\log T$ regret in the stochastic regime, optimal adversarial worst-case regret and adapts canonically to corruptions. 3) We derive the first best-of-both-worlds algorithm for linear bandits and tabular MDPs with first-order guarantees in the adversarial regime. 4) We obtain the first best-of-both-worlds algorithms for bandits with graph feedback and bandits with expert advice with optimal $\log T$ stochastic regret.

2. Related Work

Our reduction procedure is related to a class of model selection algorithms that uses a meta bandit algorithm to learn over a set of base bandit algorithms, and the goal is to achieve a comparable performance as if the best base algorithm is run alone. For the adversarial regime, Agarwal et al. (2017) introduced the Corral algorithm to learn over adversarial bandits algorithm that satisfies the stability condition. This framework is further improved by Foster et al. (2020); Luo et al. (2022). For the stochastic regime, Arora et al. (2021); Abbasi-Yadkori et al. (2020); Cutkosky et al. (2021) introduced another set of techniques to achieve similar guarantees, but without explicitly relying on the stability condition. While most of these results focus on obtaining a $\sqrt{T}$ regret, Arora et al. (2021); Cutkosky et al. (2021); Wei et al. (2022); Pacchiano et al. (2022) made some progress in obtaining polylog(T) regret in the stochastic regime. Among them, Wei et al. (2022); Pacchiano et al. (2022) are most related to us since they also pursue a best-of-both-worlds guarantee across adversarial and stochastic regimes. However, their regret bounds, when applied to our problems, all suffer from highly sub-optimal dependence on $\log(T)$ and the amount of corruptions.

3. Preliminaries

We consider sequential decision making problems where the learning interacts with the environment in T rounds. In each round t, the environment generates a loss vector $(\ell_{t,u})_{u\in\mathcal{X}}$ and the learner generates a distribution over actions p_t and chooses an action $A_t \sim p_t$. The learner then suffers loss ℓ_{t,A_t} and receives some information about $(\ell_{t,u})_{u\in\mathcal{X}}$ depending on the concrete setting. In all settings but MDPs, we assume $\ell_{t,u} \in [-1, 1]$.

In the adversarial regime, $(\ell_{t,u})_{u\in\mathcal{X}}$ is generated arbitrarily subject to the structure of the concrete setting (e.g., in linear bandits, $\mathbb{E}[\ell_{t,u}] = \langle u, \ell_t\rangle$ for arbitrary $\ell_t \in \mathbb{R}^d$). In the stochastic regime, we further assume that there exists an action $x^\star \in \mathcal{X}$ and a gap $\Delta > 0$ such that $\mathbb{E}[\ell_{t,x} - \ell_{t,x^\star}] \geq \Delta$ for all $x \neq x^\star$. We also consider the corrupted stochastic regime, where the assumption is relaxed to $\mathbb{E}[\ell_{t,x} - \ell_{t,x^\star}] \geq \Delta - C_t$ for some $C_t \geq 0$. We define $C = \sum_{t=1}^T C_t$. The goal of the learner is to minimize the pseudo-regret, defined as $\text{Reg} = \max_{u\in\mathcal{X}} \mathbb{E}\left[\sum_{t=1}^T (\ell_{t,A_t} - \ell_{t,u})\right]$.

4. Main Techniques

4.1. The standard global self-bounding condition and a new linear bandit algorithm

To obtain a best-of-both-world regret bound, previous works show the following property for their algorithm:

Definition 1 (α-global-self-bounding condition, or α-GSB)

$$\forall u \in \mathcal{X}: \ \mathbb{E}\left[\sum_{t=1}^T (\ell_{t,A_t} - \ell_{t,u})\right] \leq \min\left\{c_0^{1-\alpha}T^\alpha, \ \ (c_1 \log T)^{1-\alpha}\mathbb{E}\left[\sum_{t=1}^T (1 - p_{t,u})\right]^\alpha\right\} + c_2 \log T$$

where c_0, c_1, c_2 are problem-dependent constants and $p_{t,u}$ is the probability of the learner choosing u in round t.

With this condition, one can use the standard *self-bounding* technique to obtain a best-of-both-world regret bounds:

Proposition 2 *If an algorithm satisfies α-GSB, then its pseudo-regret is bounded by $O\left(c_0^{1-\alpha}T^\alpha + c_2 \log T\right)$ in the adversarial regime, and by $O\big(c_1 \log(T)\Delta^{-\frac{\alpha}{1-\alpha}} + (c_1 \log T)^{1-\alpha}\left(C\Delta^{-1}\right)^\alpha + c_2 \log(T)\big)$ in the corrupted stochastic regime.*

For completeness, we provide a proof of Proposition 2 in Appendix D. Previous works have found algorithms with GSB in various settings, such as multi-armed bandits, semi-bandits, graph-bandits, and MDPs. Here, we provide one such algorithm (Algorithm 3) for linear bandits based on the framework of SCRiBLe (Abernethy et al., 2008). The guarantee of Algorithm 3 is stated in the next theorem under the more general "learning with predictions" setting introduced in Rakhlin and Sridharan (2013), where in every round t, the learner receives a loss predictor m_t before making decisions.

Theorem 3 *In the "learning with predictions" setting, Algorithm 3 achieves a second-order regret bound of* $O\left(d\sqrt{\nu \log(T) \sum_{t=1}^{T}(\ell_{t,A_t} - m_{t,A_t})^2} + d\nu \log T\right)$ *in the adversarial regime, where* d *is the feature dimension,* $\nu \leq O(d)$ *is the self-concordance parameter of the regularizer, and* $m_{t,x} = \langle x, m_t \rangle$ *is the loss predictor. This also implies a first-order regret bound of* $O\left(d\sqrt{\nu \log(T) \sum_{t=1}^{T} \ell_{t,u}}\right)$ *if* $\ell_{t,x} \geq 0$ *and* $m_t = \mathbf{0}$*; it simultaneously achieves* $\mathrm{Reg} = O\left(\frac{d^2\nu \log T}{\Delta} + \sqrt{\frac{d^2\nu \log T}{\Delta} C}\right)$ *in the corrupted stochastic regime.*

See Appendix B for the algorithm and the proof. We call this algorithm Variance-Reduced SCRiBLe (VR-SCRiBLe) since it is based on the original SCRiBLe updates, but with some refinement in the construction of the loss estimator to reduce its variance. A good property of a SCRiBLe-based algorithm is that it simultaneously achieves data-dependent bounds (i.e., first- and second-order bounds), similar to the case of multi-armed bandit using FTRL/OMD with log-barrier regularizer (Wei and Luo, 2018; Ito, 2021b). Like Rakhlin and Sridharan (2013); Bubeck et al. (2019); Ito (2021b); Ito et al. (2022a), we can also use another procedure to learn m_t and obtain path-length or variance bounds. The details are omitted.

We notice, however, that the the bound in Theorem 3 is sub-optimal in d since the best-known regret for linear bandits is either $d\sqrt{T \log T}$ or $\sqrt{dT \log |\mathcal{X}|}$, depending on whether the number of actions is larger than T^d. These bounds also hint the possibility of getting better dependence on d in the stochastic regime if we can also establish GSB for these algorithms. Therefore, we ask: can we achieve best-of-both-world bounds for linear bandits with the optimal d dependence?

An attempt is to try to show GSB for existing algorithms with optimal d dependence, including the EXP2 algorithm (Bubeck et al., 2012) and the logdet-FTRL algorithm (Zimmert and Lattimore, 2022)[1]. Based on our attempt, it is unclear how to adapt the analysis of logdet-FTRL to show GSB. For EXP2, using the learning-rate tuning technique by Ito et al. (2022b), one can make it satisfy a similar guarantee as GSB, albeit with an additional $\log(T)$ factor, resulting in a bound $\frac{d \log(|\mathcal{X}|T) \log(T)}{\Delta}$ in the stochastic regime. This gives a sub-optimal rate of $O(\log^2 T)$. Therefore, we further ask: can we achieve $O(\log T)$ regret in the stochastic regime with an optimal d dependence?

Motivated by these questions, we resort to approaches that do not rely on GSB, which appears for the first time in the best-of-both-world literature. Our approach is in fact a general *reduction* — it not only successfully achieves the desired bounds for linear bandits, but also provides a principled way to convert an algorithm with a worst-case guarantee to one with a best-of-both-world guarantee for general settings. In the next two subsections, we introduce our reduction techniques.

1. Another potential algorithm is the truncated continuous exponential weight algorithm by Ito et al. (2020), which obtains optimal regret dependence on d, is computationally efficient, and can achieve first/second-order bounds. However, the regret bound shown by Ito et al. (2020) has sub-optimal dependence on $\log T$, making it impossible to obtain a $O(\log T)$ regret bound in the stochastic regime using the analysis framework of Ito et al. (2020).

Algorithm 1 BOBW via LSB algorithms

Input: LSB algorithm $\mathcal{A}$
$T_1 \leftarrow 0,\ T_0 \leftarrow -c_2 \log(T)$.
$\widehat{x}_1 \sim \text{unif}(\mathcal{X})$.
$t \leftarrow 1$.
for $k = 1, 2 \ldots,$ **do**
  Initialize $\mathcal{A}$ with candidate $\widehat{x}_k$.
  Set counters $N_k(x) = 0$ for all $x \in \mathcal{X}$.
  for $t = T_k + 1, T_k + 2, \ldots$ **do**
    Play action A_t as suggested by $\mathcal{A}$, and advance $\mathcal{A}$ by one step.
    $N_k(A_t) \leftarrow N_k(A_t) + 1$
    if $t - T_k \geq 2(T_k - T_{k-1})$ *and* $\exists x \in \mathcal{X} \setminus \{\widehat{x}_k\}$ *such that* $N_k(x) \geq \frac{t-T_k}{2}$ **then**
      $\widehat{x}_{k+1} \leftarrow x$.
      $T_{k+1} \leftarrow t$.
      break.
    end
  end
end

4.2. First reduction: Best of Both Worlds → Local Self-Bounding

Our reduction approach relies on an algorithm to satisfy a *weaker* condition than GSB, defined in the following:

Definition 4 (α-local-self-bounding condition, or α-LSB) *We say an algorithm satisfies the α-local-self-bounding condition if it takes a candidate action $\widehat{x} \in \mathcal{X}$ as input and satisfies the following pseudo-regret guarantee for any stopping time $t' \in [1, T]$:*

$\forall u \in \mathcal{X}$:

$$\mathbb{E}\left[\sum_{t=1}^{t'} (\ell_{t,A_t} - \ell_{t,u})\right] \leq \min\left\{ c_0^{1-\alpha} \mathbb{E}[t']^\alpha,\ \ (c_1 \log T)^{1-\alpha} \mathbb{E}\left[\sum_{t=1}^{t'} (1 - \mathbb{I}\{\boldsymbol{u = \widehat{x}}\} p_{t,u})\right]^\alpha \right\} + c_2 \log T$$

where c_0, c_1, c_2 are problem dependent constants.

The difference between LSB in Definition 4 and GSB in Definition 1 is that LSB only requires the self-bounding term $\sum_t (1 - p_{t,u})$ to appear when u is a particular action $\widehat{x}$ given as an input to the algorithm; for all other actions, the worst-case bound suffices. A minor additional requirement is that the pseudo-regret needs to hold for any stopping time (because our reduction may stop this algorithm during the learning procedure), but this is relatively easy to satisfy — for all algorithms in this paper, without additional efforts, their regret bounds naturally hold for any stopping time.

For linear bandits, we find that an adaptation of the logdet-FTRL algorithm (Zimmert and Lattimore, 2022) satisfies $\frac{1}{2}$-LSB, as stated in the following lemma. The proof is provided in Appendix C.

Lemma 5 *For d-dimensional linear bandits, by transforming the action set into $\left\{ \binom{x}{0} \,\middle|\, x \in \mathcal{X} \setminus \{\widehat{x}\} \right\} \cup \left\{ \binom{\mathbf{0}}{1} \right\} \subset \mathbb{R}^{d+1}$, Algorithm 4 satisfies $\frac{1}{2}$-LSB with $c_0 = O((d+1)^2 \log T)$ and $c_1 = c_2 = O((d+1)^2)$.*

With the LSB condition defined, we develop a reduction procedure (Algorithm 1) which turns any algorithm with LSB into a best-of-both-world algorithm that has a similar guarantee as in Proposition 2. The guarantee is formally stated in the next theorem.

Theorem 6 *If $\mathcal{A}$ satisfies α-LSB, then the regret of Algorithm 1 with $\mathcal{A}$ as the base algorithm is upper bounded by $O\left(c_0^{1-\alpha}T^\alpha + c_2\log^2 T\right)$ in the adversarial regime and by $O\big(c_1\log(T)\Delta^{-\frac{\alpha}{1-\alpha}} + (c_1\log T)^{1-\alpha}\left(C\Delta^{-1}\right)^\alpha + c_2\log(T)\log(C\Delta^{-1})\big)$ in the corrupted stochastic regime.*

See Appendix E.1 for a proof of Theorem 6. In particular, Theorem 6 together with Lemma 5 directly lead to a better best-of-both-world bound for linear bandits.

Corollary 7 *Combining Algorithm 1 and Algorithm 4 results in a linear bandit algorithm with $O\big(d\sqrt{T\log T}\big)$ regret in the adversarial regime and $O\Big(\frac{d^2\log T}{\Delta}+d\sqrt{\frac{C\log T}{\Delta}}\Big)$ regret in the corrupted stochastic regime simultaneously.*

Ideas of the first reduction Algorithm 1 proceeds in epochs. In epoch k, there is an action $\widehat{x}_k \in \mathcal{X}$ being selected as the candidate ($\widehat{x}_1$ is randomly drawn from $\mathcal{X}$). The procedure simply executes the base algorithm $\mathcal{A}$ with input $\widehat{x} = \widehat{x}_k$, and monitors the number of draws to each action. If in epoch k there exists some $x \neq \widehat{x}_k$ being drawn more than half of the time, and the length of epoch k already exceeds two times the length of epoch $k-1$, then a new epoch is started with $\widehat{x}_{k+1}$ set to x.

Here, we give a proof sketch for Theorem 6 with $\alpha = \frac{1}{2}$. It is straightforward in the adversarial regime: Let $\tau_k = T_{k+1} - T_k$ be the length of epoch k. In each epoch k, the regret is upper bounded by $O(\sqrt{c_0\tau_k})$ due to the LSB property of the base algorithm. Thus, the over all regret is upper bounded by $O(\sum_k\sqrt{c_0\tau_k}) = O(\sqrt{c_0\sum_k\tau_k}) = O(\sqrt{c_0T})$ because $\tau_k \geq 2\tau_{k-1}$ (except for the last epoch).

Next, we analyze the regret in the stochastic regime assuming $c_2 = 0$ and $C = 0$ (see the formal proof for the general case). Let $m = \max\{k:\ \widehat{x}_k \neq x^\star\}$ be the index of the last epoch in which $\widehat{x}_k$ is not $x^\star$. Below, we show that $\widehat{x}_m \neq x^\star$ is drawn for at least $\Omega(T_{m+1})$ times. If $\tau_m > 2\tau_{m-1}$, then by the termination condition of epoch m, we know that just one step before epoch m terminates, $\widehat{x}_m$ is drawn for at least half of the times in epoch m, which is at least $\frac{\tau_m}{2} - 1 = \Omega(\tau_m) = \Omega(T_{m+1})$ times. If $\tau_m \leq 2\tau_{m-1}$, then by the termination condition of epoch $m-1$, we know that $\widehat{x}_m$ must have been drawn for $\frac{\tau_{m-1}}{2} = \Omega(T_{m+1})$ times in epoch $m-1$.

Notice that the regret up to the end of epoch m is upper bounded by $O(\sqrt{c_1T_{m+1}\log T})$, but also lower bounded by $\Omega(T_{m+1}\Delta)$ by the argument above. This implies that $T_{m+1} = O(\frac{c_1\log T}{\Delta^2})$ and thus the regret up to the end of epoch m is upper bounded by $O(\frac{c_1\log T}{\Delta})$. Finally, if epoch m is not the last epoch, then in the last epoch $m+1$, it must be that $\widehat{x}_{m+1} = x^\star$. By the LSB property, in the last epoch, the regret is upper bounded by $O(\sqrt{c_1\sum_{t=T_{m+1}+1}^{T}(1-p_{t,x^\star})\log T}$ and lower bounded by $\sum_{t=T_{m+1}+1}^{T}(1-p_{t,x^\star})\Delta$. This implies that $\sum_{t=T_{m+1}+1}^{T}(1-p_{t,x^\star}) = O(\frac{c_1\log T}{\Delta^2})$ and thus the regret is upper bounded by $O(\frac{c_1\log T}{\Delta})$.

Theorem 6 is not sufficient to be considered a black-box reduction approach, since algorithms with LSB are not common. Therefore, our next step is to present a more general reduction that makes use of recent advances of Corral algorithms.

4.3. Second reduction: Local Self-Bounding $\rightarrow$ Importance-Weighting Stable

In this subsection, we show that one can achieve LSB using the idea of *model selection*. Specifically, we will run a variant of the Corral algorithm (Agarwal et al., 2017) over two instances: one is $\widehat{x}$, and the other is a *importance-weighting-stable* algorithm (see Definition 8) over the action set $\mathcal{X}\backslash\{\widehat{x}\}$. Here, we focus on the case of $\alpha = \frac{1}{2}$, which is the case for most standard settings where the worst-case regret is $\sqrt{T}$; examples for $\alpha = \frac{2}{3}$ in the graph bandit setting is discussed in Section 5.

First, we introduce the notion of importance-weighting stablity.

Definition 8 ($\frac{1}{2}$-iw-stable) *An algorithm is* $\frac{1}{2}$-iw-stable *(importance-weighting stable), if given an adaptive sequence of weights $q_1, q_2, \cdots \in (0, 1]$ and the assertion that the feedback in round t is observed with probability q_t, it obtains the following pseudo regret guarantee for any stopping time t':*

$$\mathbb{E}\left[\sum_{t=1}^{t'}(\ell_{t,A_t} - \ell_{t,u})\right] \leq \mathbb{E}\left[\sqrt{c_1 \sum_{t=1}^{t'} \frac{1}{q_t}} + \frac{c_2}{\min_{t\leq t'} q_t}\right].$$

Definition 8 requires that even if the algorithm only receives the desired feedback with probability q_t in round t, it still achieves a meaningful regret bound that smoothly degrades with $\sum_t \frac{1}{q_t}$. In previous works on Corral and its variants (Agarwal et al., 2017; Foster et al., 2020), a similar notion of $\frac{1}{2}$-stability is defined as having a regret of $\sqrt{c_1 \left(\max_{\tau\leq t'} \frac{1}{q_\tau}\right) t'}$, which is a weaker assumption than our $\frac{1}{2}$-iw-stability. Our stronger notion of stability is crucial to get a best-of-both-world bound, but it is still very natural and holds for a wide range of algorithms.

Below, we provide two examples of $\frac{1}{2}$-iw-stable algorithms. Their analysis is mostly standard FTRL analysis (considering extra importance weighting) and can be found in Appendix G.

Lemma 9 *For linear bandits, EXP2 with adaptive learning rate (Algorithm 9) is $\frac{1}{2}$-iw-stable with constants $c_1 = c_2 = O(d \log |\mathcal{X}|)$.*

Lemma 10 *For contextual bandits, EXP4 with adaptive learning rate (Algorithm 10) is $\frac{1}{2}$-iw-stable with constants $c_1 = O(K \log |\mathcal{X}|), c_2 = 0$.*

Our reduction procedure is Algorithm 2, which turns an $\frac{1}{2}$-iw-stable algorithm into one with $\frac{1}{2}$-LSB. The guarantee is formalized in the next theorem, whose proof is in Appendix F.1.

Theorem 11 *If $\mathcal{B}$ is a $\frac{1}{2}$-iw-stable algorithm with constants (c_1, c_2), then Algorithm 2 with $\mathcal{B}$ as the base algorithm satisfies $\frac{1}{2}$-LSB with constants (c_0', c_1', c_2') where $c_0' = c_1' = O(c_1)$ and $c_2' = O(c_2)$.*

Cascading Theorem 11 and Theorem 6, and combining it with Lemma 9 and Lemma 10, respectively, we get the following corollaries:

Corollary 12 *Combining Algorithm 1, Algorithm 2 and Algorithm 9 results in a linear bandit algorithm with $O\big(\sqrt{dT \log |\mathcal{X}|}\big)$ regret in the adversarial regime and $O\Big(\frac{d \log |\mathcal{X}| \log T}{\Delta} + \sqrt{\frac{d \log |\mathcal{X}| \log T}{\Delta} C}\Big)$ regret in the corrupted stochastic regime simultaneously.*

Algorithm 2 LSB via Corral (for $\alpha = \frac{1}{2}$)

Input: candidate action $\widehat{x}$, $\frac{1}{2}$-iw-stable algorithm $\mathcal{B}$ over $\mathcal{X}\backslash\{\widehat{x}\}$ with constants c_1, c_2.
Define: $\psi_t(q) = -\frac{2}{\eta_t}\sum_{i=1}^{2}\sqrt{q_i} + \frac{1}{\beta}\sum_{i=1}^{2}\ln\frac{1}{q_i}$.
$B_0 = 0$.
for $t = 1, 2, \dots$ **do**
Let

$$\bar{q}_t = \operatorname*{argmin}_{q\in\Delta_2}\left\{\left\langle q, \sum_{\tau=1}^{t-1} z_\tau - \begin{bmatrix} 0 \\ B_{t-1}\end{bmatrix}\right\rangle + \psi_t(q)\right\}, \quad q_t = \left(1 - \frac{1}{2t^2}\right)\bar{q}_t + \frac{1}{4t^2}\mathbf{1},$$
$$\text{with} \quad \eta_t = \frac{1}{\sqrt{t} + 8\sqrt{c_1}}, \qquad \beta = \frac{1}{8c_2}.$$

Sample $i_t \sim q_t$.
if $i_t = 1$ **then** draw $A_t = \widehat{x}$ and observe ℓ_{t,A_t} ;
else draw A_t according to base algorithm $\mathcal{B}$ and observe ℓ_{t,A_t} ;
Define $z_{t,i} = \frac{(\ell_{t,A_t}+1)\mathbb{I}\{i_t=i\}}{q_{t,i}} - 1$ and

$$B_t = \sqrt{c_1\sum_{\tau=1}^{t}\frac{1}{q_{\tau,2}} + \frac{c_2}{\min_{\tau\le t} q_{\tau,2}}}.$$

end

Corollary 13 *Combining Algorithm 1, Algorithm 2 and Algorithm 10 results in a contextual bandit algorithm with $O\big(\sqrt{KT\log|\mathcal{X}|}\big)$ regret in the adversarial regime and $O\Big(\frac{K\log|\mathcal{X}|\log T}{\Delta} + \sqrt{\frac{K\log|\mathcal{X}|\log T}{\Delta}C}\Big)$ regret in the corrupted stochastic regime simultaneously, where K is the number of arms.*

Ideas of the second reduction Algorithm 2 is related to the Corral algorithm (Agarwal et al., 2017; Foster et al., 2020; Luo et al., 2022). We use a special version which is essentially an FTRL algorithm (with a hybrid $\frac{1}{2}$-Tsallis entropy and log-barrier regularizer) over two base algorithms: the candidate $\widehat{x}$, and an algorithm $\mathcal{B}$ operating on the reduced action set $\mathcal{X}\backslash\{\widehat{x}\}$. For base algorithm $\mathcal{B}$, the Corral algorithm adds a bonus term that upper bounds the regret of $\mathcal{B}$ under importance weighting (i.e., the quantity B_t defined in Algorithm 6). In the regret analysis, the bonus creates a negative term as well as a bonus overhead in the learner's regret. The negative term can be used to cancel the positive regret of $\mathcal{B}$, and the analysis reduces to bounding the bonus overhead. Showing that the bonus overhead is upper bounded by the order of $\sqrt{c_1\log(T)\sum_{t=1}^{T}(1-p_{t,\widehat{x}})}$ is the key to establish the LSB property.

Combining Algorithm 2 and Algorithm 1, we have the following general reduction: *as long as we have an $\frac{1}{2}$-iw-stable algorithm over $\mathcal{X}\backslash\{\widehat{x}\}$ for any $\widehat{x}\in\mathcal{X}$, we have an algorithm with the best-of-both-world guarantee when the optimal action is unique.* Notice that $\frac{1}{2}$-iw-stable algorithms are quite common – usually it can be just a FTRL/OMD algorithm with adaptive learning rate.

The overall reduction is reminiscent of the G-COBE procedure by Wei et al. (2022), which also runs model selection over $\widehat{x}$ and a base algorithm for $\mathcal{X}\backslash\{\widehat{x}\}$ (similar to Algorithm 2), and dynamically restarts the procedure with increasing epoch lengths (similar to Algorithm 1). However, besides being more complicated, G-COBE only guarantees a bound of $\frac{\text{polylog}(T)}{\Delta}+C$ in the corrupted stochastic regime (omitting dependencies on c_1, c_2), which is sub-optimal in both C and T.[2]

5. Case Study: Graph Bandits

In this section, we show the power of our reduction by improving the state of the art of best-of-both-worlds algorithms for bandits with graph feedback. In bandits with graph feedback, the learner is given a directed feedback graph $G = (V, E)$, where the nodes $V = [K]$ correspond to the K-arms of the bandit. Instead of observing the loss of the played action A_t, the player observes the loss $\ell_{t,j}$ iff $(A_t, j) \in E$. Learnable graphs are divided into *strongly observable graphs* and *weakly observable graphs*. In the first case, every arm $i \in [K]$ must either receive its own feedback, i.e. $(i, i) \in E$, or *all* other arms do, i.e. $\forall j \in [K] \setminus \{i\} : (j, i) \in E$. In the weakly observable case, every arm $i \in [K]$ must be observable by at least one arm, i.e. $\exists j \in [K] : (j, i) \in E$. The following graph properties characterize the difficulty of the two regimes. An independence set is any subset $V' \subset V$ such that no edge exists between any two distinct nodes in V', i.e. $\forall i, j \in V', i \neq j$ we have $(i, j) \notin E$ and $(j, i) \notin E$. The independence number α is the size of the largest independence set. A related number is the weakly independence number $\widetilde{\alpha}$, which is the independence number of the subgraph (V, E') obtained by removing all one-sided edges, i.e. $(i, j) \in E'$ iff $(i, j) \in E$ and $(j, i) \in E$. For undirected graphs, the two notions coincide, but in general $\widetilde{\alpha}$ can be larger than α by up to a factor of K. Finally, a weakly dominating set is a subset $D \subset V$ such that every node in V is observable by some node in D, i.e. $\forall j \in V\, \exists i \in D : (i, j) \in E$. The weakly dominating number δ is the size of the smallest weakly dominating set.

Alon et al. (2015) provides a almost tight characterization of the adversarial regime, providing upper bounds of $O(\sqrt{\alpha T \log T \log K})$ and $O(\sqrt{\widetilde{\alpha} T \log K})$ for the strongly observable case and $O((\delta \log K)^{\frac{1}{3}} T^{\frac{2}{3}})$ for the weakly observable case, as well as matching lower bounds up to all log factors. Zimmert and Lattimore (2019) have shown that the $\log(T)$ dependency can be avoided and that hence α is the right notion of difficulty for strongly observable graphs even in the limit $T \to \infty$ (though they pay for this with a larger $\log K$ dependency).

State-of-the-art best-of-both-worlds guarantees for bandits with graph feedback are derivations of EXP3 and obtain either $O(\frac{\widetilde{\alpha} \log^2(T)}{\Delta})$ or $O(\frac{\alpha \log(T) \log^2(TK)}{\Delta})$ regret for strongly observable graphs and $O(\frac{\delta \log^2(T)}{\Delta^2})$ for weakly observable graphs (Rouyer et al., 2022; Ito et al., 2022b)[3]. Our blackbox reduction leads directly to algorithms with optimal $\log(T)$ regret in the stochastic regime.

5.1. Strongly observable graphs

We begin by providing an examples of $\frac{1}{2}$-iw-stable algorithm for bandits with strongly observable graph feedback. The algorithm and the proof are in Appendix G.3.

2. However, Wei et al. (2022) handles a more general type of corruption.
3. Rouyer et al. (2022) obtains better bounds when the gaps are not uniform, while Ito et al. (2022b) can handle graphs that are unions of strongly observable and weakly observable sub-graphs. We do not extend our analysis to the latter case for the sake of simplicity.

Lemma 14 *For bandits with strongly observable graph feedback, $(1 - 1/\log K)$-Tsallis-INF with adaptive learning rate (Algorithm 11) is $\frac{1}{2}$-iw-stable with constants $c_1 = O(\min\{\widetilde{\alpha}, \alpha \log K\} \log K), c_2 = 0$.*

Before we apply the reduction, note that Algorithm 2 requires that the player observes the loss ℓ_{t,A_t} when playing arm A_t, which is not necessarily the case for all strongly observable graphs. However, this can be overcome via defining an observable surrogate loss that is used in Algorithm 2 instead. We explain how this works in detail in Section H and assume from now that this technical issue does not arise.

Cascading the two reduction steps, we immediately obtain the following.

Corollary 15 *Combining Algorithm 1, Algorithm 2 and Algorithm 11 results in a graph bandit algorithm with $O\big(\sqrt{\min\{\widetilde{\alpha}, \alpha \log K\} T \log K}\big)$ regret in the adversarial regime and $O\Big(\frac{\min\{\widetilde{\alpha},\alpha \log K\} \log T \log K}{\Delta} + \sqrt{\frac{\min\{\widetilde{\alpha},\alpha \log K\} \log T \log K}{\Delta} C}\Big)$ regret in the corrupted stochastic regime simultaneously.*

5.2. Weakly observable graphs

Weakly observable graphs motivate our general definition of LSB stable beyond $\alpha = \frac{1}{2}$. We first need to define the equivalent of $\frac{1}{2}$-iw-stable for this regime.

Definition 16 ($\frac{2}{3}$-iw-stable) *An algorithm is* $\frac{2}{3}$-iw-stable *(importance-weighting stable), if given an adaptive sequence of weights $q_1, q_2, \dots, q_\tau$ and the feedback in any round being observed with probability q_t, it obtains the following pseudo regret guarantee for any stopping time t':*

$$\mathbb{E}\left[\sum_{t=1}^{t'} (\ell_{t,A_t} - \ell_{t,u})\right] \le \mathbb{E}\left[c_1^{\frac{1}{3}} \left(\sum_{t=1}^{t'} \frac{1}{\sqrt{q_t}}\right)^{\frac{2}{3}} + \max_{t \le t'} \frac{c_2}{q_t}\right].$$

An example of such an algorithm is given by the following lemma (see Appendix G.4 for the algorithm and the proof).

Lemma 17 *For bandits with weakly observable graph feedback, EXP3 with adaptive learning rate (Algorithm 12) is $\frac{2}{3}$-iw-stable with constants $c_1 = c_2 = O(\delta \log K)$.*

Similar to the $\frac{1}{2}$-case, this allows for a general reduction to LSB algorithms.

Theorem 18 *If $\mathcal{B}$ is a $\frac{2}{3}$-iw-stable algorithm with constants (c_1, c_2), then Algorithm 5 with $\mathcal{B}$ as the base algorithm satisfies $\frac{2}{3}$-LSB with constants (c_0', c_1', c_2') where $c_0' = c_1' = O(c_1)$ and $c_2' = O(c_1^{\frac{1}{3}} + c_2)$.*

See Appendix F.2 for the proof. Algorithm 5 works almost the same as Algorithm 2, but we need to adapt the bonus terms to the definition of $\frac{2}{3}$-iw-stable. The major additional difference is that we do not necessarily observe the loss of the action played and hence need to play exploratory actions with probability γ_t in every round to estimate the performance difference between $\widehat{x}$ and $\mathcal{B}$. A second point to notice is that the base algorithm $\mathcal{B}$ uses the action set $\mathcal{X} \setminus \{\widehat{x}\}$, but is still allowed to use $\widehat{x}$ in its internal exploration policy. This is necessary because the sub-graph with one arm removed might have a larger dominating number, or might even not be learnable at all. By allowing $\widehat{x}$ in the internal exploration, we guarantee the the regret of the base algorithm is not larger than over the full action set. Finally, cascading the lemma and the reduction, we obtain

Corollary 19 *Combining Algorithm 1, Algorithm 5 and Algorithm 12 results in a graph bandit algorithm with* $O\big((\delta\log K)^{\frac{1}{3}}T^{\frac{2}{3}}\big)$ *regret in the adversarial regime and* $O\Big(\frac{\delta\log K\log T}{\Delta^2}+\Big(\frac{C^2\delta\log K\log T}{\Delta^2}\Big)^{\frac{1}{3}}\Big)$ *regret in the corrupted stochastic regime simultaneously.*

6. More Adaptations

To demonstrate the versatility of our reduction framework, we provide two more adaptations. The first demonstrates that our reduction can be easily adapted to obtain a *data-dependent* bound (i.e., first- or second-order bound) in the adversarial regime, provided that the base algorithm achieves a corresponding data-dependent bound. The second tries to eliminate the undesired requirement by the corral algorithm (Algorithm 2) that the base algorithm needs to operate in $\mathcal{X}\backslash\{\widehat{x}\}$ instead of the more natural $\mathcal{X}$. We show that under a stronger stability assumption, we can indeed just use a base algorithm that operates in $\mathcal{X}$. This would be helpful for settings where excluding one single action/policy in the algorithm is not straightforward (e.g., MDP). Finally, we combine the two techniques to obtain a first-order best-of-both-world bound for tabular MDPs.

6.1. Reduction with first- and second-order bounds

A first-order regret bound refers to a regret bound of order $O\big(\sqrt{c_1\text{polylog}(T)L_\star}+c_2\text{polylog}(T)\big)$, where $L_\star=\min_{x\in\mathcal{X}}\mathbb{E}\big[\sum_{t=1}^T\ell_{t,x}\big]$ is the best action's cumulative loss. This is meaningful under the assumption that $\ell_{t,x}\geq 0$ for all t,x. A second-order regret bound refers to a bound of order $O\Big(\sqrt{c_1\text{polylog}(T)\mathbb{E}\big[\sum_{t=1}^T(\ell_{t,A_t}-m_{t,A_t})^2\big]}+c_2\text{polylog}(T)\Big)$, where $m_{t,x}$ is a *loss predictor* for action x that is available to the learner at the beginning of round t. We refer the reader to Wei and Luo (2018); Ito (2021b) for more discussions on data-dependent bounds.

We first define counterparts of the LSB condition and iw-stability condition with data-dependent guarantees:

Definition 20 ($\frac{1}{2}$-dd-LSB) *We say an algorithm satisfies $\frac{1}{2}$-dd-LSB (data-dependent LSB) if it takes a candidate action $\widehat{x}\in\mathcal{X}$ as input and satisfies the following pseudo-regret guarantee for any stopping time $t'\in[1,T]$:*

$$\forall u\in\mathcal{X}:$$

$$\mathbb{E}\left[\sum_{t=1}^{t'}(\ell_{t,A_t}-\ell_{t,u})\right]\leq\sqrt{(c_1\log T)\mathbb{E}\left[\sum_{t=1}^{t'}\left(\sum_x p_{t,x}\xi_{t,x}-\mathbb{I}\{\boldsymbol{u}=\widehat{\boldsymbol{x}}\}p_{t,u}^2\xi_{t,u}\right)\right]}+c_2\log T$$

where c_1,c_2 are some problem dependent constants, $\xi_{t,x}=(\ell_{t,x}-m_{t,x})^2$ in the second-order bound case, and $\xi_{t,x}=\ell_{t,x}$ in the first-order bound case.

Definition 21 ($\frac{1}{2}$-dd-iw-stable) *An algorithm is* $\frac{1}{2}$-dd-iw-stable *(data-dependent-iw-stable) if given an adaptive sequence of weights $q_1,q_2,\dots\in(0,1]$ and the assertion that the feedback in round t is observed with probability q_t, it obtains the following pseudo regret guarantee for any stopping time t':*

$$\mathbb{E}\left[\sum_{t=1}^{t'}(\ell_{t,A_t}-\ell_{t,u})\right]\leq\sqrt{c_1\mathbb{E}\left[\sum_{t=1}^{t'}\frac{\textit{upd}_t\cdot\xi_{t,A_t}}{q_t^2}\right]}+\mathbb{E}\left[\frac{c_2}{\min_{t\leq t'}q_t}\right],$$

where $\mathsf{upd}_t = 1$ *if feedback is observed in round* t *and* $\mathsf{upd}_t = 0$ *otherwise, and* $\xi_{t,x}$ *is defined in the same way as in Definition 20.*

We can turn a dd-iw-stable algorithm into one with dd-LSB (see Appendix F.3 for the proof):

Theorem 22 *If* $\mathcal{B}$ *is* $\frac{1}{2}$*-dd-iw-stable with constants* (c_1, c_2)*, then Algorithm 6 with* $\mathcal{B}$ *as the base algorithm satisfies* $\frac{1}{2}$*-dd-LSB with constants* (c_1', c_2') *where* $c_1' = O(c_1)$ *and* $c_2' = O(\sqrt{c_1} + c_2)$*.*

Then we turn an algorithm with dd-LSB into one with data-dependent best-of-both-world guarantee (see Appendix E.2 for the proof):

Theorem 23 *If an algorithm* $\mathcal{A}$ *satisfies* $\frac{1}{2}$*-dd-LSE, then the regret of Algorithm 1 with* $\mathcal{A}$ *as the base algorithm is upper bounded by* $O\Big(\sqrt{c_1 \mathbb{E}\left[\sum_{t=1}^T \xi_{t,A_t}\right] \log^2 T} + c_2 \log^2 T\Big)$ *in the adversarial regime and* $O\Big(\frac{c_1 \log T}{\Delta} + \sqrt{\frac{c_1 \log T}{\Delta} C} + c_2 \log(T) \log(C\Delta^{-1})\Big)$ *in the corrupted stochastic regime.*

6.2. Achieving LSB without excluding $\widehat{x}$

Our reduction in Section 4.3 requires that the base algorithm $\mathcal{B}$ to operate in the action set of $\mathcal{X} \backslash \{\widehat{x}\}$. This is sometimes not easy to implement for structural problems where actions share common components (e.g., MDPs or combinatorial bandits). To eliminate this requirement so that we can simply use a base algorithm $\mathcal{B}$ that operates in the original action space $\mathcal{X}$, we propose the following stronger notion of iw-stability that we called *strongly-iw-stable*:

Definition 24 ($\frac{1}{2}$-strongly-iw-stable) *An algorithm is* $\frac{1}{2}$-strongly-iw-stable *if the following holds: given an adaptive sequence of weights* $q_1, q_2, \cdots \in (0,1]^{\mathcal{X}}$ *and the assertion that the feedback in round* t *is observed with probability* $q_t(x)$ *if the algorithm chooses* $A_t = x$*, it obtains the following pseudo regret guarantee for any stopping time* t'*:*

$$\mathbb{E}\left[\sum_{t=1}^{t'} (\ell_{t,A_t} - \ell_{t,u})\right] \le \mathbb{E}\left[\sqrt{c_1 \sum_{t=1}^{t'} \frac{1}{q_t(A_t)}} + \frac{c_2}{\min_{t \le t'} \min_x q_t(x)}\right].$$

Compared with iw-stability defined in Definition 8, strong-iw-stability requires the bound to have an additional flexibility: if choosing action x results in a probability $q_t(x)$ of observing the feedback, the regret bound needs to adapts to $\sum_t \frac{1}{q_t(A_t)}$. For the class of FTRL/OMD algorithms, strong-iw-stability holds if the *stability term* is bounded by a constant no matter what action is chosen. Examples include log-barrier-OMD/FTRL for multi-armed bandits, SCRiBLe (Abernethy et al., 2008) or a truncated version of continuous exponential weights (Ito et al., 2020) for linear bandits. In fact, Upper-Confidence-Bound (UCB) algorithms also satisfy strong-iw-stability, though it is mainly designed for the pure stochastic setting; however, we will need this property when we design algorithms for adversarial MDPs, where the transition estimation part is done through UCB approaches. This will be demonstrated in Section 6.3. With strong-iw-stability, the second reduction only requires a base algorithm over $\mathcal{X}$:

Theorem 25 *If* $\mathcal{B}$ *is a* $\frac{1}{2}$*-strongly-iw-stable algorithm with constants* (c_1, c_2)*, then Algorithm 7 with* $\mathcal{B}$ *as the base algorithm satisfies* $\frac{1}{2}$*-LSB with constants* (c_0', c_1', c_2') *where* $c_0' = c_1' = O(c_1)$ *and* $c_2' = O(\sqrt{c_1} + c_2)$*.*

The proof is provided in Appendix F.4.

6.3. Application to MDPs

Finally, we combine the techniques developed in Section 6.1 and Section 6.2 to obtain a best-of-both-world guarantee for tabular MDPs with a first-order regret bound in the adversarial regime. We use Algorithm 13 as the base algorithm, which is adapted from the UCB-log-barrier Policy Search algorithm by Lee et al. (2020) to satisfy both the data-dependent property (Definition 21) and the strongly iw-stable property (Definition 24). The corral algorithm we use is Algorithm 8, which takes a base algorithm with the dd-strongly-iw-stable property and turns it into a data-dependent best-of-both-world algorithm. The details and notations are all provided in Appendix I. The guarantee of the algorithm is formally stated in the next theorem.

Theorem 26 *Combining Algorithm 1, Algorithm 8, and Algorithm 13 results in an MDP algorithm with* $O\left(\sqrt{S^2AHL_\star \log^2(T)\iota^2} + S^5A^2\log^2(T)\iota^2\right)$ *regret in the adversarial regime, and* $O\left(\frac{H^2S^2A\iota^2\log T}{\Delta} + \sqrt{\frac{H^2S^2A\iota^2\log T}{\Delta}C} + S^5A^2\iota^2\log(T)\log(C\Delta^{-1})\right)$ *in the corrupted stochastic regime, where* S *is the number of states,* A *is the number of actions,* H *is the horizon,* $L_\star$ *is the cumulative loss of the best policy, and* $\iota = \log(SAT)$.

7. Conclusion

We provided a general reduction from the best-of-both-worlds problem to a wide range of FTRL/OMD algorithms, which improves the state of the art in several problem settings. We showed the versatility of our approach by extending it to preserving data-dependent bounds.

Another potential application of our framework is partial monitoring, where one might improve the $\log^2(T)$ rates of to $\log(T)$ for both the $T^{\frac{1}{2}}$ and the $T^{\frac{2}{3}}$ regime using our respective reductions.

A weakness of our approach is the uniqueness requirement of the best action. While this assumption is typical in the best-of-both-worlds literature, it is not merely an artifact of the analysis for us due to the doubling procedure in the first reduction. Additionally, our reduction can only obtain worst-case Δ dependent bounds in the stochastic regime, which can be significantly weaker than more refined notions of complexity.

References

Yasin Abbasi-Yadkori, Aldo Pacchiano, and My Phan. Regret balancing for bandit and rl model selection. *arXiv preprint arXiv:2006.05491*, 2020.

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *21st Annual Conference on Learning Theory, COLT 2008*, 2008.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.

Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. From bandits to experts: A tale of domination and independence. *Advances in Neural Information Processing Systems*, 26, 2013.

Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, pages 23–35. PMLR, 2015.

Idan Amir, Guy Azov, Tomer Koren, and Roi Livni. Better best of both worlds bounds for bandits with switching costs. In *Advances in Neural Information Processing Systems*, 2022.

Raman Arora, Teodor Vanislavov Marinov, and Mehryar Mohri. Corralling stochastic bandit algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 2116–2124. PMLR, 2021.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 116–120. PMLR, 2016.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1. JMLR Workshop and Conference Proceedings, 2012.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Sham M Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, pages 41–1. JMLR Workshop and Conference Proceedings, 2012.

Sébastien Bubeck, Yuanzhi Li, Haipeng Luo, and Chen-Yu Wei. Improved path-length regret bounds for bandits. In *Conference On Learning Theory*, pages 508–528. PMLR, 2019.

Cheng Chen, Canzhe Zhao, and Shuai Li. Simultaneously learning stochastic and adversarial bandits under the position-based model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6202–6210, 2022.

Ashok Cutkosky, Christoph Dann, Abhimanyu Das, Claudio Gentile, Aldo Pacchiano, and Manish Purohit. Dynamic balancing for model selection in bandits and rl. In *International Conference on Machine Learning*, pages 2276–2285. PMLR, 2021.

Liad Erez and Tomer Koren. Towards best-of-all-worlds online learning with feedback graphs. *Advances in Neural Information Processing Systems*, 34:28511–28521, 2021.

Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33:11478–11489, 2020.

Pierre Gaillard, Gilles Stoltz, and Tim Van Erven. A second-order bound with excess losses. In *Conference on Learning Theory*, pages 176–196. PMLR, 2014.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 1562–1578. PMLR, 2019.

Junya Honda, Shinji Ito, and Taira Tsuchiya. Follow-the-perturbed-leader achieves best-of-both-worlds for bandit problems. In *International Conference on Algorithmic Learning Theory*. PMLR, 2023.

Jiatai Huang, Yan Dai, and Longbo Huang. Adaptive best-of-both-worlds algorithm for heavy-tailed multi-armed bandits. In *International Conference on Machine Learning*, pages 9173–9200. PMLR, 2022.

Shinji Ito. On optimal robustness to adversarial corruption in online decision problems. *Advances in Neural Information Processing Systems*, 34:7409–7420, 2021a.

Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Conference on Learning Theory*, pages 2552–2583. PMLR, 2021b.

Shinji Ito, Shuichi Hirahara, Tasuku Soma, and Yuichi Yoshida. Tight first-and second-order regret bounds for adversarial linear bandits. *Advances in Neural Information Processing Systems*, 33: 2028–2038, 2020.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Adversarially robust multi-armed bandit algorithm with variance-dependent regret bounds. In *Conference on Learning Theory*, pages 1421–1422. PMLR, 2022a.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, 2022b.

Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic mdps with known transition. *Advances in neural information processing systems*, 33:16557–16566, 2020.

Tiancheng Jin, Longbo Huang, and Haipeng Luo. The best of both worlds: stochastic and adversarial episodic mdps with unknown transition. *Advances in Neural Information Processing Systems*, 34:20491–20502, 2021.

Fang Kong, Yichi Zhou, and Shuai Li. Simultaneously learning stochastic and adversarial bandits with general graph feedback. In *International Conference on Machine Learning*, pages 11473–11482. PMLR, 2022.

Tze Leung Lai, Herbert Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. *Advances in neural information processing systems*, 33:15522–15533, 2020.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang, and Xiaojin Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *International Conference on Machine Learning*, pages 6142–6151. PMLR, 2021.

Haipeng Luo. Homework 3 solution, introduction to online optimization/learning. http://haipeng-luo.net/courses/CSCI659/2022_fall/homework/HW3_solutions.pdf, November 2022.

Haipeng Luo, Mengxiao Zhang, Peng Zhao, and Zhi-Hua Zhou. Corralling a larger band of bandits: A case study on switching regret for linear bandits. In *Conference on Learning Theory*, pages 3635–3684. PMLR, 2022.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122, 2018.

Saeed Masoudian and Yevgeny Seldin. Improved analysis of the tsallis-inf algorithm in stochastically constrained adversarial bandits and stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 3330–3350. PMLR, 2021.

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, 2022.

Jaouad Mourtada and Stéphane Gaïffas. On the optimality of the hedge algorithm in the stochastic regime. *Journal of Machine Learning Research*, 20:1–28, 2019.

Aldo Pacchiano, Christoph Dann, and Claudio Gentile. Best of both worlds model selection. In *Advances in Neural Information Processing Systems*, 2022.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. *Advances in Neural Information Processing Systems*, 26, 2013.

Chloé Rouyer, Yevgeny Seldin, and Nicolo Cesa-Bianchi. An algorithm for stochastic and adversarial bandits with switching costs. In *International Conference on Machine Learning*, pages 9127–9135. PMLR, 2021.

Chloé Rouyer, Dirk van der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. A near-optimal best-of-both-worlds algorithm for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, 2022.

Aadirupa Saha and Pierre Gaillard. Versatile dueling bandits: Best-of-both world analyses for learning from relative preferences. In *International Conference on Machine Learning*, pages 19011–19026. PMLR, 2022.

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 1743–1759. PMLR, 2017.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, pages 1287–1295. PMLR, 2014.

Taira Tsuchiya, Shinji Ito, and Junya Honda. Best-of-both-worlds algorithms for partial monitoring. *arXiv preprint arXiv:2207.14550*, 2022.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291. PMLR, 2018.

Chen-Yu Wei, Christoph Dann, and Julian Zimmert. A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pages 1043–1096. PMLR, 2022.

Julian Zimmert and Tor Lattimore. Connections between mirror descent, thompson sampling and the information ratio. *Advances in Neural Information Processing Systems*, 32, 2019.

Julian Zimmert and Tor Lattimore. Return of the bias: Almost minimax optimal high probability bounds for adversarial linear bandits. In *Conference on Learning Theory*, pages 3285–3312. PMLR, 2022.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 467–475. PMLR, 2019.

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, pages 7683–7692. PMLR, 2019.

Appendices

Appendix A. FTRL Analysis

Lemma 27 *The optimistic FTRL algorithm over a convex set Ω:*

$$p_t = \operatorname*{argmin}_{p\in\Omega}\left\{\left\langle p, \sum_{\tau=1}^{t-1}\ell_\tau\right\rangle + m_t + \psi_t(p)\right\}$$

guarantees the following for any t':

$$\sum_{t=1}^{t'}\langle p_t - u, \ell_t\rangle \le \psi_0(u) - \min_{p\in\Omega}\psi_0(p)$$
$$+\sum_{t=1}^{t'}(\psi_t(u)-\psi_{t-1}(u)-\psi_t(p_t)+\psi_{t-1}(p_t)) + \sum_{t=1}^{t'}\underbrace{\max_{p\in\Omega}\left(\langle p_t - p, \ell_t - m_t\rangle - D_{\psi_t}(p,p_t)\right)}_{\textit{stability}}.$$

Proof Let $L_t \triangleq \sum_{\tau=1}^{t}\ell_\tau$. Define $F_t(p) = \langle p, L_{t-1}+m_t\rangle + \psi_t(p)$ and $G_t(p) = \langle p, L_t\rangle + \psi_t(p)$. Therefore, p_t is the minimizer of F_t over Ω. Let p'_{t+1} be minimizer of G_t over Ω. Then by the first-order optimality condition, we have

$$F_t(p_t) - G_t(p'_{t+1}) \le F_t(p'_{t+1}) - G_t(p'_{t+1}) - D_{\psi_t}(p'_{t+1},p_t) = -\langle p'_{t+1}, \ell_t - m_t\rangle - D_{\psi_t}(p'_{t+1},p_t). \tag{1}$$

By definition, we also have

$$G_t(p'_{t+1}) - F_{t+1}(p_{t+1}) \le G_t(p_{t+1}) - F_{t+1}(p_{t+1}) = \psi_t(p_{t+1}) - \psi_{t+1}(p_{t+1}) - \langle p_{t+1}, m_{t+1}\rangle. \tag{2}$$

Thus,

$$\sum_{t=1}^{t'}\langle p_t,\ell_t\rangle$$
$$\le \sum_{t=1}^{t'}\left(\langle p_t,\ell_t\rangle - \langle p'_{t+1},\ell_t-m_t\rangle - D_{\psi_t}(p'_{t+1},p_t) + G_t(p'_{t+1}) - F_t(p_t)\right) \qquad \text{(by (1))}$$
$$= \sum_{t=1}^{t'}\left(\langle p_t,\ell_t\rangle - \langle p'_{t+1},\ell_t-m_t\rangle - D_{\psi_t}(p'_{t+1},p_t) + G_{t-1}(p'_t) - F_t(p_t)\right) + G_{t'}(p'_{t'+1}) - G_0(p'_1)$$
$$= \sum_{t=1}^{t'}\left(\langle p_t,\ell_t\rangle - \langle p'_{t+1},\ell_t-m_t\rangle - D_{\psi_t}(p'_{t+1},p_t) - \psi_t(p_t) + \psi_{t-1}(p_t) - \langle p_t, m_t\rangle\right) + G_{t'}(p'_{t'+1}) - G_0(p'_1)$$
$$\le \sum_{t=1}^{t'}\left(\langle p_t - p'_{t+1},\ell_t-m_t\rangle - D_{\psi_t}(p'_{t+1},p_t) - \psi_t(p_t) + \psi_{t-1}(p_t)\right) + G_{t'}(u) - \min_{p\in\Omega}\psi_0(p)$$
$$\text{(by (2), using that } p'_{t'+1} \text{ is the minimizer of } G_{t'})$$
$$= \sum_{t=1}^{t'}\left(\max_{p\in\Omega}\{\langle p_t - p,\ell_t-m_t\rangle - D_{\psi_t}(p,p_t)\} - \psi_t(p_t) + \psi_{t-1}(p_t)\right)$$

$$+ \sum_{t=1}^{t'} \langle u, \ell_t \rangle + \psi_{t'}(u) - \min_{p \in \Omega} \psi_0(p)$$

$$= \sum_{t=1}^{t'} \left(\max_{p \in \Omega} \{ \langle p_t - p, \ell_t - m_t \rangle - D_{\psi_t}(p, p_t) \} - \psi_t(p_t) + \psi_{t-1}(p_t) \right)$$

$$+ \sum_{t=1}^{t'} \langle u, \ell_t \rangle + \psi_{t'}(u) - \min_{p \in \Omega} \psi_0(p)$$

$$= \sum_{t=1}^{t'} \left(\max_{p \in \Omega} \{ \langle p_t - p, \ell_t - m_t \rangle - D_{\psi_t}(p, p_t) \} + (\psi_t(u) - \psi_{t-1}(u) - \psi_t(p_t) + \psi_{t-1}(p_t)) \right)$$

$$+ \sum_{t=1}^{t'} \langle u, \ell_t \rangle + \psi_0(u) - \min_{p \in \Omega} \psi_0(p)$$

Re-arranging finishes the proof. ■

Lemma 28 *Consider the optimistic FTRL algorithm with bonus $b_t \geq \mathbf{0}$:*

$$p_t = \operatorname*{argmin}_{p \in \Omega} \left\{ \left\langle p, \sum_{\tau=1}^{t-1} (\ell_\tau - b_\tau) \right\rangle + m_t + \psi_t(p) \right\}$$

with $\psi_t(p) = \frac{1}{\eta_t} \psi^{\mathsf{Ts}}(p) + \frac{1}{\beta} \psi^{\mathsf{Lo}}(p)$ where $\psi^{\mathsf{Ts}}(p) = \frac{-1}{1-\alpha} \sum_i p_i^\alpha$ for some $\alpha \in (0,1)$, $\psi^{\mathsf{Lo}}(p) = \sum_i \ln \frac{1}{p_i}$, and η_t is non-increasing. We have

$$\sum_{t=1}^{t'} \langle p_t - u, \ell_t \rangle$$

$$\leq \frac{1}{1-\alpha} \frac{K^{1-\alpha}}{\eta_0} + \frac{1}{1-\alpha} \sum_{t=1}^{t'} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \left(\sum_{i=1}^K p_{t,i}^\alpha - 1 \right) + \frac{K \ln \frac{1}{\delta}}{\beta} + \frac{1}{\alpha} \sum_{t=1}^{t'} \eta_t \sum_{i=1}^K p_{t,i}^{2-\alpha} (\ell_{t,i}^{\mathsf{Ts}} - x_t)^2$$

$$+ 2\beta \sum_{t=1}^{t'} \sum_{i=1}^K p_{t,i}^2 (\ell_{t,i}^{\mathsf{Lo}} - w_t)^2 + \left(1 + \frac{1}{\alpha} \right) \sum_{t=1}^{t'} \langle p_t, b_t \rangle - \sum_{t=1}^{t'} \langle u, b_t \rangle + \delta \sum_{t=1}^{t'} \left\langle -u + \frac{1}{K} \mathbf{1}, \ell_t - b_t \right\rangle.$$

for any $\delta \in (0,1)$ and any $\ell_t^{\mathsf{Ts}}, b_t^{\mathsf{Ts}} \in \mathbb{R}^K$, $\ell_t^{\mathsf{Lo}}, b_t^{\mathsf{Lo}} \in \mathbb{R}^K$, $x_t, w_t \in \mathbb{R}$ such that

$$\ell_t^{\mathsf{Ts}} + \ell_t^{\mathsf{Lo}} = \ell_t - m_t, \tag{3}$$

$$b_t^{\mathsf{Ts}} + b_t^{\mathsf{Lo}} = b_t \tag{4}$$

$$\eta_t p_{t,i}^{1-\alpha} (\ell_{t,i}^{\mathsf{Ts}} - x_t) \geq -\frac{1}{4}, \tag{5}$$

$$\beta p_{t,i} (\ell_{t,i}^{\mathsf{Lo}} - w_t) \geq -\frac{1}{4}, \tag{6}$$

$$0 \leq \eta_t p_{t,i}^{1-\alpha} b_{t,i}^{\mathsf{Ts}} \leq \frac{1}{4}, \tag{7}$$

$$0 \leq \beta p_{t,i} b_{t,i}^{\mathsf{Lo}} \leq \frac{1}{4} \tag{8}$$

for all t, i.

Proof Let $u' = (1-\delta)u + \frac{\delta}{K}\mathbf{1}$. By Lemma 27, we have

$$\begin{aligned}
\sum_{t=1}^{t'}\langle p_t - u', \ell_t - b_t\rangle &\le \frac{1}{1-\alpha}\frac{1}{\eta_0}\sum_{i=1}^{K} p_{1,i}^{\alpha} + \frac{1}{1-\alpha}\sum_{t=1}^{t'}\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right)\sum_{i=1}^{K}\left(p_{t,i}^{\alpha} - u_i'^{\alpha}\right) + \frac{1}{\beta}\sum_{i=1}^{K}\ln\frac{p_{1,i}}{u_i'} \\
&\quad + \sum_{t=1}^{t'}\max_p\left(\langle p_t - p, \ell_t - b_t - m_t\rangle - \frac{1}{\eta_t}D_{\psi^{\mathsf{Ts}}}(p,p_t) - \frac{1}{\beta}D_{\psi^{\mathsf{Lo}}}(p,p_t)\right) \\
&\le \frac{1}{1-\alpha}\frac{K^{1-\alpha}}{\eta_0} + \frac{1}{1-\alpha}\sum_{t=1}^{t'}\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right)\left(\sum_{i=1}^{K}p_{t,i}^{\alpha} - 1\right) + \frac{K\ln\frac{1}{\delta}}{\beta} \\
&\quad + \underbrace{\sum_{t=1}^{t'}\max_p\left(\langle p_t - p, \ell^{\mathsf{Ts}} - x_t\mathbf{1}\rangle - \frac{1}{2\eta_t}D_{\psi^{\mathsf{Ts}}}(p,p_t)\right)}_{\textbf{stability-1}} \\
&\quad + \underbrace{\sum_{t=1}^{t'}\max_p\left(\langle p_t - p, -b_t^{\mathsf{Ts}}\mathbf{1}\rangle - \frac{1}{2\eta_t}D_{\psi^{\mathsf{Ts}}}(p,p_t)\right)}_{\textbf{stability-2}} \\
&\quad + \underbrace{\sum_{t=1}^{t'}\max_p\left(\langle p_t - p, \ell^{\mathsf{Lo}} - w_t\mathbf{1}\rangle - \frac{1}{2\beta}D_{\psi^{\mathsf{Lo}}}(p,p_t)\right)}_{\textbf{stability-3}} \\
&\quad + \underbrace{\sum_{t=1}^{t'}\max_p\left(\langle p_t - p, -b_t^{\mathsf{Lo}}\mathbf{1}\rangle - \frac{1}{2\beta}D_{\psi^{\mathsf{Lo}}}(p,p_t)\right)}_{\textbf{stability-4}}
\end{aligned}$$

where in the last inequality we use $\langle p_t - p, \mathbf{1}\rangle = 0$.

By Problem 1 in Luo (2022), we can bound **stability-1** by $\frac{\eta_t}{\alpha}\sum_{i=1}^{K} p_{t,i}^{2-\alpha}(\ell_{t,i}^{\mathsf{Ts}} - x_t)^2$ under the condition (5). Similarly, **stability-2** can be upper bounded by $\frac{\eta_t}{\alpha}\sum_{i=1}^{K} p_{t,i}^{2-\alpha} b_{t,i}^{\mathsf{Ts}2} \le \frac{1}{\alpha}\sum_{i=1}^{K} p_{t,i} b_{t,i}^{\mathsf{Ts}}$ under the condition (7). Using Lemma 30, we can bound **stability-3** by $2\beta\sum_{i=1}^{K} p_{t,i}^2(\ell_{t,i}^{\mathsf{Lo}} - w_t)^2$ under the condition (6). Similarly, we can bound **stability-4** by $2\beta\sum_{i=1}^{K} p_{t,i}^2 b_{t,i}^{\mathsf{Lo}2} \le \sum_{i=1}^{K} p_{t,i} b_{t,i}^{\mathsf{Lo}}$ under (8). Collecting all terms and using the definition of u' finishes the proof. ∎

Lemma 29 *Consider the optimistic FTRL algorithm with bonus $b_t \ge \mathbf{0}$:*

$$p_t = \operatorname*{argmin}_{p\in\Omega}\left\{\left\langle p, \sum_{\tau=1}^{t-1}(\ell_\tau - b_\tau)\right\rangle + m_t + \psi_t(p)\right\}$$

with $\psi_t(p) = \frac{1}{\eta_t}\sum_i \ln \frac{1}{p_i}$*, and* η_t *is non-increasing. We have*

$$\sum_{t=1}^{t'} \langle p_t - u, \ell_t \rangle \leq \frac{K \ln \frac{1}{\delta}}{\eta_{t'}} + 2\sum_{t=1}^{t'} \eta_t \sum_{i=1}^{K} p_{t,i}^2 (\ell_{t,i} - m_{t,i} - x_t)^2$$
$$+ 2\sum_{t=1}^{t'} \langle p_t, b_t \rangle - \sum_{t=1}^{t'} \langle u, b_t \rangle + \delta \sum_{t=1}^{t'} \left\langle -u + \frac{1}{K}\mathbf{1}, \ell_t - b_t \right\rangle .$$

for any $\delta \in (0,1)$ *and any* $x_t \in \mathbb{R}$ *if the following hold:* $\eta_t p_{t,i}(\ell_{t,i} - m_{t,i} - x_t) \geq -\frac{1}{4}$ *and* $\eta_t p_{t,i} b_{t,i} \leq \frac{1}{4}$ *for all* t, i*.*

Proof Let $u' = (1-\delta)u + \frac{\delta}{K}\mathbf{1}$. By Lemma 27 and letting $\eta_0 = \infty$, we have

$$\sum_{t=1}^{t'} \langle p_t - u', \ell_t - b_t \rangle \leq \sum_{t=1}^{t'} \sum_{i=1}^{K} \left(\ln \frac{p_{t,i}}{u'_i} \right) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) + \sum_{t=1}^{t'} \max_p \left(\langle p_t - p, \ell_t - b_t - m_t \rangle - D_{\psi_t}(p, p_t) \right)$$
$$\leq \frac{K \ln \frac{K}{\delta}}{\eta_{t'}} + \sum_{t=1}^{t'} \underbrace{\max_p \left(\langle p_t - p, \ell_t - m_t - x_t \mathbf{1} \rangle - \frac{1}{2} D_{\psi_t}(p, p_t) \right)}_{\textbf{stability-1}}$$
$$+ \sum_{t=1}^{t'} \underbrace{\max_p \left(\langle p_t - p, -b_t \rangle - \frac{1}{2} D_{\psi_t}(p, p_t) \right)}_{\textbf{stability-2}}$$

where in the last inequality we use $\langle p_t - p, \mathbf{1} \rangle = 0$.

Using Lemma 30, we can bound **stability-1** by $2\eta_t \sum_{i=1}^{K} p_{t,i}^2 (\ell_{t,i} - m_{t,i} - x_t)^2$, and bound **stability-2** by $2\eta_t \sum_{i=1}^{K} p_{t,i}^2 b_{t,i}^2 \leq \sum_{i=1}^{K} p_{t,i} b_{t,i}$ under the specified conditions. Collecting all terms and using the definition of u' finishes the proof. ■

Lemma 30 (Stability under log barrier) *Let* $\psi(p) = \frac{1}{\beta}\sum_i \ln \frac{1}{p_i}$*, and let* $\ell_t \in \mathbb{R}^K$ *be such that* $\beta p_i \ell_{t,i} \geq -\frac{1}{2}$*. Then*

$$\max_{p \in \Delta([K])} \{ \langle p_t - p, \ell_t \rangle - D_\psi(p, p_t) \} \leq \sum_i \beta p_{t,i}^2 \ell_{t,i}^2 .$$

Proof

$$\max_{p \in \Delta([K])} \{ \langle p_t - p, \ell_t \rangle - D_\psi(p, p_t) \} \leq \max_{q \in \mathbb{R}_+^K} \{ \langle p_t - q, \ell_t \rangle - D_\psi(q, p_t) \}$$

Define $f(q) = \langle p_t - q, \ell_t \rangle - D_\psi(q, p_t)$. Let $q^\star$ be the solution in the last expression. Next, we verify that under the specified conditions, we have $\nabla f(q^\star) = 0$. It suffices to show that there exists $q \in \mathbb{R}_+^K$ such that $\nabla f(q) = 0$ since if such q exists, then it must the maximizer of f and thus $q^\star = q$.

$$[\nabla f(q)]_i = -\ell_{t,i} - [\nabla \psi(q)]_i + [\nabla \psi(p_t)]_i = -\ell_{t,i} + \frac{1}{\beta q_i} - \frac{1}{\beta p_{t,i}}$$

By the condition, we have $-\frac{1}{\beta p_{t,i}} - \ell_{t,i} < 0$ for all i. and so $\nabla f(q) = \mathbf{0}$ has solution in $\mathbb{R}_+$, which is $q_i = \left(\frac{1}{p_{t,i}} + \eta_{t,i}\ell_{t,i}\right)^{-1}$.

Therefore, $\nabla f(q^\star) = -\ell_t - \nabla\psi_t(q^\star) + \nabla\psi_t(p_t) = 0$, and we have

$$\max_{q\in\mathbb{R}_+^K}\left\{\langle p_t - q, \ell_t\rangle - D_\psi(q, p_t)\right\} = \langle p_t - q^\star, \nabla\psi(p_t) - \nabla\psi(q^\star)\rangle - D_\psi(q^\star, p_t) = D_\psi(p_t, q^\star).$$

It remains to bound $D_\psi(p_t, q^\star)$, which by definition can be written as

$$D_\psi(p_t, q^\star) = \sum_i \frac{1}{\beta} h\left(\frac{p_{t,i}}{q_i^\star}\right)$$

where $h(x) = x - 1 - \ln(x)$. By the relation between $q_i^\star$ and $p_{t,i}$ we just derived, it holds that $\frac{p_{t,i}}{q_i^\star} = 1 + \beta p_{t,i}\ell_{t,i}$. By the fact that $\ln(1+x) \geq x - x^2$ for all $x \geq -\frac{1}{2}$, we have

$$h\left(\frac{p_{t,i}}{q_i^\star}\right) = \beta p_{t,i}\ell_{t,i} - \ln(1 + \beta p_{t,i}\ell_{t,i}) \leq \beta^2 p_{t,i}^2 \ell_{t,i}^2$$

which gives the desired bound.

■

Lemma 31 (Stability under negentropy) *If ψ is the negentropy $\psi(p) = \sum_i p_i \log p_i$ and $\ell_{t,i} > 0$, then for any $\eta > 0$ the stability is bounded by*

$$\max_{p\in\mathbb{R}_{\geq 0}^K} \langle p_t - p, \ell_t\rangle - \frac{1}{\eta} D_\psi(p, p_t) \leq \frac{\eta}{2}\sum_i p_{t,i}\ell_{t,i}^2\,.$$

If $\ell_t > -\frac{1}{\eta}$, then the stability is bounded by

$$\max_{p\in\mathbb{R}_{\geq 0}^K} \langle p_t - p, \ell_t\rangle - \frac{1}{\eta} D_\psi(p, p_t) \leq \eta\sum_i p_{t,i}\ell_{t,i}^2\,.$$

instead.

Proof Let $f_i(p_i) = (p_{t,i} - p_i)\ell_{ti} - \frac{1}{\eta}\left(p_i(\log p_i - 1) - p_i \log p_{t,i} + p_{t,i}\right)$. Then we maximize $\sum_i f_i(p_i)$, which is upper bounded by maximizing the expression in each coordinate for $p_i \geq 0$. We have

$$f_i'(p) = -\ell_{t,i} - \frac{1}{\eta}(\log p - \log p_{t,i}),$$

and hence the maximum is obtained for $p^\star = p_{t,i}\exp(-\eta\ell_{t,i})$. Plugging this in leads to

$$\begin{aligned} f_i(p^\star) &= p_{t,i}\ell_{t,i}(1 - \exp(-\eta\ell_{t,i})) - \frac{1}{\eta}\left(-p_{t,i}\exp(-\eta\ell_{t,i})\eta\ell_{t,i} + p_{t,i}(1 - \exp(-\eta\ell_{t,i}))\right) \\ &= p_{t,i}\ell_{t,i} - \frac{p_{t,i}}{\eta}(1 - \exp(-\eta\ell_{t,i})) \leq p_{t,i}\ell_{t,i} - \frac{p_{t,i}}{\eta}\left(\eta\ell_{t,i} - \frac{1}{2}\eta^2\ell_{t,i}^2\right) = \frac{\eta}{2}p_{t,i}\ell_{t,i}^2\,, \end{aligned}$$

for non-negative ℓ_t and

$$f_i(p^*) \leq \eta p_{t,i}\ell_{t,i}^2$$

for $\ell_{t,i} \geq -1/\eta$ respectively, where we used the bound $\exp(-x) \leq 1 - x + \frac{x^2}{2}$ which holds for any $x \geq 0$ and $\exp(-x) \leq 1 - x + x^2$, which holds for $x \geq -1$ respectively. Summing over all coordinates finishes the proof. ■

Lemma 32 (Stability of Tsallis-INF) *For the potential $\psi(p) = -\sum_i \frac{p_i^\alpha}{\alpha(1-\alpha)}$, any $p_t \in \Delta([K])$, any non-negative loss $\ell_t \geq 0$ and any positive learning rate $\eta > 0$, we have*

$$\max_p \langle p - p_t, \ell_t\rangle - \frac{1}{\eta}D_\psi(p, p_t) \leq \frac{\eta}{2}\sum_i p_i^{2-\alpha}\ell_{t,i}^2\,.$$

Proof We upper bound this term by maximizing over $p \geq 0$ instead of $p \in \Delta([K])$. Since ℓ_t is positive, the optimal $p^\star$ satisfies $p_i^\star \leq p_i$ in all components. We have

$$D_\psi(p^\star, p_t) = \sum_{i=1}^K \int_{p_{t,i}}^{p_i^\star} \frac{p_{t,i}^{\alpha-1} - p^{\alpha-1}}{1-\alpha}\,dp = \sum_{i=1}^K \int_{p_{t,i}}^{p_i^\star}\int_{p_{t,i}}^{\widetilde{p}} p^{\alpha-2}\,dp\,d\widetilde{p} \geq \sum_{i=1}^K \int_{p_{t,i}}^{p_i^\star}\int_{p_{t,i}}^{\widetilde{p}} p_{t,i}^{\alpha-2}\,dp\,d\widetilde{p}$$
$$= \sum_{i=1}^K \frac{1}{2}(p_i^\star - p_{t,i})^2 p_{t,i}^{\alpha-2}\,.$$

By AM-GM inequality, we further have

$$\max_p \langle p - p_t, \ell_t\rangle - \frac{1}{\eta}D_\psi(p, p_t) \leq \max_p \sum_i \left((p_i - p_{t,i})\ell_{t,i} - \frac{(p_i - p_{t,i})^2 p_{t,i}^{\alpha-2}}{2\eta}\right) \leq \frac{\eta}{2}\sum_i p_{t,i}^{2-\alpha}\ell_{t,i}^2\,.$$

■

Appendix B. Analysis for Variance-Reduced SCRiBLe (Algorithm 3 / Theorem 3)

Algorithm 3 VR-SCRiBLe

Define: Let $\psi(\cdot)$ be a ν-self-concordant barrier of $\text{conv}(\mathcal{X}) \subset \mathbb{R}^d$.

for $t = 1, 2, \ldots$ **do**

Receive $m_t \in \mathbb{R}^d$. Compute

$$w_t = \operatorname*{argmin}_{w \in \text{conv}(\mathcal{X})} \left\{ \left\langle w, \sum_{\tau=1}^{t-1} \widehat{\ell}_\tau + m_t \right\rangle + \frac{1}{\eta_t} \psi(w) \right\} \tag{9}$$

where

$$\eta_t = \min\left\{ \frac{1}{16d}, \sqrt{\frac{\nu \log T}{\sum_{\tau=1}^{t-1} \|\widehat{\ell}_\tau - m_\tau\|^2_{H_\tau^{-1}}}} \right\}.$$

where $H_t = \nabla^2 \psi(w_t)$.

Sample s_t uniformly from $\mathbb{S}_d$ (the unit sphere in d-dimension).

Define

$$w_t^+ = w_t + H_t^{-\frac{1}{2}} s_t, \qquad w_t^- = w_t - H_t^{-\frac{1}{2}} s_t.$$

Find distributions q_t^+ and q_t^- over actions such that

$$w_t^+ = \sum_{x \in \mathcal{X}} q_{t,x}^+ x, \qquad w_t^- = \sum_{x \in \mathcal{X}} q_{t,x}^- x$$

Sample $A_t \sim q_t \triangleq \frac{q_t^+ + q_t^-}{2}$, receive $\ell_{t,A_t} \in [-1, 1]$ with $\mathbb{E}[\ell_{t,x}] = \langle x, \ell_t \rangle$, and define

$$\widehat{\ell}_t = d(\ell_{t,A_t} - m_{t,A_t}) \left(\frac{q_{t,A_t}^+ - q_{t,A_t}^-}{q_{t,A_t}^+ + q_{t,A_t}^-} \right) H_t^{\frac{1}{2}} s_t + m_t$$

where $m_{t,x} := \langle x, m_t \rangle$.

end

Lemma 33 *In Algorithm 3, we have* $\mathbb{E}\left[\widehat{\ell}_t\right] = \ell_t$ *and*

$$\mathbb{E}\left[\left\| \widehat{\ell}_t - m_t \right\|^2_{\nabla^{-2}\psi(w_t)} \right] \leq d^2 \mathbb{E}\left[\sum_{x \in \mathcal{X}} \min\{p_{t,x}, 1 - p_{t,x}\} (\ell_{t,x} - m_{t,x})^2 \right].$$

where $p_{t,x}$ *is the probability of choosing action* x *in round* t.

Proof

$$
\begin{aligned}
\mathbb{E}\left[\widehat{\ell}_t\right] &= \mathbb{E}\left[d(\ell_{t,A_t}-m_{t,A_t})\left(\frac{q^+_{t,A_t}-q^-_{t,A_t}}{q^+_{t,A_t}+q^-_{t,A_t}}\right)H_t^{\frac{1}{2}}s_t+m_t\right]\\
&= \mathbb{E}\left[d\mathbb{E}\left[(\ell_{t,A_t}-m_{t,A_t})\left(\frac{q^+_{t,A_t}-q^-_{t,A_t}}{q^+_{t,A_t}+q^-_{t,A_t}}\right)\middle|s_t\right]H_t^{\frac{1}{2}}s_t+m_t\right]\\
&\hspace{14em}\text{(note that } q_t^+, q_t^- \text{ depend on } s_t\text{)}\\
&= \mathbb{E}\left[d\mathbb{E}\left[\sum_x q_{t,x}(\ell_{t,x}-m_{t,x})\left(\frac{q^+_{t,x}-q^-_{t,x}}{q^+_{t,x}+q^-_{t,x}}\right)\middle|s_t\right]H_t^{\frac{1}{2}}s_t+m_t\right]\\
&= \mathbb{E}\left[d\mathbb{E}\left[\sum_x \langle x,\ell_t-m_t\rangle\left(\frac{q^+_{t,x}-q^-_{t,x}}{2}\right)\middle|s_t\right]H_t^{\frac{1}{2}}s_t+m_t\right]\\
&= \mathbb{E}\left[d\mathbb{E}\left[\frac{\langle w_t^+-w_t^-,\ell_t-m_t\rangle}{2}\middle|s_t\right]H_t^{\frac{1}{2}}s_t+m_t\right]\\
&= \mathbb{E}\left[d\langle H^{-\frac{1}{2}}s_t,\ell_t-m_t\rangle H_t^{\frac{1}{2}}s_t+m_t\right]\\
&= \mathbb{E}\left[dH_t^{\frac{1}{2}}s_ts_t^\top H_t^{-\frac{1}{2}}(\ell_t-m_t)+m_t\right]\\
&= \ell_t. \hspace{16em}\text{(because } \mathbb{E}[s_ts_t^\top]=\tfrac{1}{d}I\text{)}
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\ell}_t-m_t\right\|^2_{\nabla^{-2}\psi(w_t)}\right] &= \mathbb{E}\left[d^2(\ell_{t,A_t}-m_{t,A_t})^2\left(\frac{q^+_{t,A_t}-q^-_{t,A_t}}{q^+_{t,A_t}+q^-_{t,A_t}}\right)^2\left\|H_t^{\frac{1}{2}}s_t\right\|^2_{H_t^{-1}}\right]\\
&= \mathbb{E}\left[d^2(\ell_{t,A_t}-m_{t,A_t})^2\left(\frac{q^+_{t,A_t}-q^-_{t,A_t}}{q^+_{t,A_t}+q^-_{t,A_t}}\right)^2\right]\\
&= \mathbb{E}\left[\mathbb{E}\left[d^2(\ell_{t,A_t}-m_{t,A_t})^2\left(\frac{q^+_{t,A_t}-q^-_{t,A_t}}{q^+_{t,A_t}+q^-_{t,A_t}}\right)^2\middle|s_t\right]\right]\\
&\leq \mathbb{E}\left[\mathbb{E}\left[\sum_x q_{t,x}d^2(\ell_{t,x}-m_{t,x})^2\left|\frac{q^+_{t,x}-q^-_{t,x}}{q^+_{t,x}+q^-_{t,x}}\right|\middle|s_t\right]\right]
\end{aligned}
$$

For any x, we have $q_{t,x}\left|\frac{q^+_{t,x}-q^-_{t,x}}{q^+_{t,x}+q^-_{t,x}}\right| \leq q_{t,x}$ and

$$
q_{t,x}\left|\frac{q^+_{t,x}-q^-_{t,x}}{q^+_{t,x}+q^-_{t,x}}\right| = \frac{|q^+_{t,x}-q^-_{t,x}|}{2} \leq 1-\frac{q^+_{t,x}+q^-_{t,x}}{2} = 1-q_{t,x}.
$$

Therefore, we continue to bound $\mathbb{E}\left[\left\|\widehat{\ell}_t - m_t\right\|^2_{\nabla^{-2}\psi(w_t)}\right]$ by

$$\mathbb{E}\left[\mathbb{E}\left[\sum_{x\in\mathcal{X}} \min\{q_{t,x}, 1-q_{t,x}\} d^2(\ell_{t,x} - m_{t,x})^2 \,\middle|\, s_t\right]\right]$$
$$\leq \mathbb{E}\left[\sum_{x\in\mathcal{X}} \min\left\{p_{t,x}, 1-p_{t,x}\right\} d^2(\ell_{t,x} - m_{t,x})^2\right].$$
$$(\mathbb{E}[\min(\cdot,\cdot)] \leq \min(\mathbb{E}[\cdot], \mathbb{E}[\cdot]) \text{ and } p_{t,x} = \mathbb{E}[\mathbb{E}[q_{t,x} \mid s_t]])$$

■

Lemma 34 *If* $\eta_t \leq \frac{1}{16d}$*, then* $\max_w \left(\langle w_t - w, \widehat{\ell}_t - m_t\rangle - \frac{1}{\eta_t} D_\psi(w, w_t)\right) \leq 8\eta_t \|\widehat{\ell}_t - m_t\|^2_{\nabla^{-2}\psi(w_t)}$.

Proof We first show that $\eta_t \|\widehat{\ell}_t - m_t\|_{\nabla^{-2}\psi(w_t)} \leq \frac{1}{16}$. By the definition of $\widehat{\ell}_t$, we have

$$\eta_t \left\|\widehat{\ell}_t - m_t\right\|_{\nabla^{-2}\psi(w_t)} \leq \frac{1}{16d} \cdot d\|H_t^{\frac{1}{2}} s_t\|_{H_t^{-1}} \leq \frac{1}{16}.$$

Define

$$F(w) = \langle w_t - w, \widehat{\ell}_t - m_t\rangle - \frac{1}{\eta_t} D_\psi(w, w_t).$$

Define $\lambda = \|\widehat{\ell}_t - m_t\|_{\nabla^{-2}\psi(w_t)}$. Let w' be the maximizer of F. it suffices to show $\|w' - w_t\|_{\nabla^2\psi(w_t)} \leq 8\eta_t\lambda$ because this leads to

$$F(w') \leq \|w' - w_t\|_{\nabla^2\psi(w_t)} \|\widehat{\ell}_t - m_t\|_{\nabla^{-2}\psi(w_t)} \leq 8\eta_t\lambda^2.$$

To show $\|w' - w_t\|_{\nabla^2\psi(w_t)} \leq 8\eta_t\lambda$, it suffices to show that for all u such that $\|u - w_t\|_{\nabla^2\psi(w_t)} = 8\eta_t\lambda$, $F(u) \leq 0$. To see why this is the case, notice that $F(w_t) = 0$, and $F(w') \geq 0$ because w' is the maximizer of F. Therefore, if $\|w' - w_t\|_{\nabla^2\psi(w_t)} > 8\eta_t\lambda$, then there exists u in the line segment between w_t and w' with $\|u - w_t\|_{\nabla^2\psi(w_t)} = 8\eta_t\lambda$ such that $F(u) \leq 0 \leq \min\{F(w_t), F(w')\}$, contradicting that F is strictly concave.

Below, consider any u with $\|u - w_t\|_{\nabla^2\psi(w_t)} = 8\eta_t\lambda$. By Taylor expansion, there exists u' in the line segment between u and w_t such that

$$F(u) \leq \|u - w_t\|_{\nabla^2\psi(w_t)} \|\widehat{\ell}_t - m_t\|_{\nabla^{-2}\psi(w_t)} - \frac{1}{2\eta_t}\|u - w_t\|^2_{\nabla^2\psi(u')}.$$

Because ψ is a self-concordant barrier and that $\|u' - w_t\|_{\nabla^2\psi(w_t)} \leq \|u - w_t\|_{\nabla^2\psi(w_t)} = 8\eta_t\lambda \leq \frac{1}{2}$, we have $\nabla^2\psi(u') \succeq \frac{1}{4}\nabla^2\psi(w_t)$. Continuing from the previous inequality,

$$F(u) \leq \|u - w_t\|_{\nabla^2\psi(w_t)} \|\widehat{\ell}_t - m_t\|_{\nabla^{-2}\psi(w_t)} - \frac{1}{8\eta_t}\|u - w_t\|^2_{\nabla^2\psi(w_t)} = 8\eta_t\lambda \cdot \lambda - \frac{(8\eta_t\lambda)^2}{8\eta_t} = 0.$$

This concludes the proof. ∎

Proof [Proof of Theorem 3] By the standard analysis of optimistic-FTRL and Lemma 34, we have for any u,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}(\ell_{t,A_t}-\ell_{t,u})\right] \\
&\le O\left(\mathbb{E}\left[\frac{\nu\log T}{\eta_T}+\sum_{t=1}^{T}\eta_t\left\|\widehat{\ell}_t-m_t\right\|^2_{\nabla^{-2}\psi(w_t)}\right]\right) \\
&= O\left(\sqrt{\nu\log T\,\mathbb{E}\left[\sum_{t=1}^{T}\left\|\widehat{\ell}_t-m_t\right\|^2_{\nabla^{-2}\psi(w_t)}\right]}+d\nu\log(T)\right). \qquad \text{(by the tuning of } \eta_t\text{)}
\end{aligned} \tag{10}$$

In the adversarial regime, using Lemma 33, we continue to bound (10) by

$$O\left(d\sqrt{\nu\log T\,\mathbb{E}\left[\sum_{t=1}^{T}(\ell_{t,A_t}-m_{t,A_t})^2\right]}+d\nu\log(T)\right)$$

When losses are non-negative and $m_t=\mathbf{0}$, we can further upper bound it by

$$O\left(d\sqrt{\nu\log T\,\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,A_t}\right]}+d\nu\log(T)\right).$$

Then solving the inequality for $\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,A_t}\right]$, we can further get the first-order regret bound of

$$\mathbb{E}\left[\sum_{t=1}^{T}(\ell_{t,A_t}-\ell_{t,u})\right]\le O\left(d\sqrt{\nu\log T\,\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,u}\right]}+d\nu\log(T)\right).$$

In the corrupted stochastic setting, using Lemma 33, we continue to bound (10) by

$$\begin{aligned}
&O\left(d\sqrt{\nu\log(T)\mathbb{E}\left[\sum_{t=1}^{T}\left((1-p_{t,u})(\ell_{t,u}-m_{t,u})^2+\sum_{x\neq u}p_{t,x}(\ell_{t,x}-m_{t,x})^2\right)\right]}\right) \\
&\le O\left(d\sqrt{\nu\log(T)\mathbb{E}\left[\sum_{t=1}^{T}(1-p_{t,u})\right]}\right).
\end{aligned}$$

Then by the self-bounding technique stated in Proposition 2, we can further bound the regret in the corrupted stochastic setting by

$$O\left(\frac{d^2\nu\log(T)}{\Delta}+\sqrt{\frac{d^2\nu\log(T)}{\Delta}C}\right).$$

∎

Appendix C. Analysis for LSB Log-Determinant FTRL (Algorithm 4 / Lemma 5)

Algorithm 4 LSB-logdet

Input: $\mathcal{X}(=\{\binom{x}{0}\}), \widehat{x} = \binom{\mathbf{0}}{1}$.
Define: $H(\alpha) := \mathbb{E}_{x\sim\alpha}[xx^\top]$, $\mu_\alpha := \mathbb{E}_{x\sim\alpha}[x]$.
Let $\pi_{\mathcal{X}}$ be John's exploration over $\mathcal{X}$, i.e., $\pi_{\mathcal{X}}$ is such that $\max_{x\in\mathcal{X}} \|x\|^2_{H(\pi_{\mathcal{X}})^{-1}} \le d$.
for $t = 1, 2, \dots$ **do**
Let

$$\eta_t = \min\left\{\frac{1}{4d}, \sqrt{\frac{\log(T)}{\sum_{\tau=1}^{t-1}(1-p_{\tau,\widehat{x}})}}\right\},$$

$$p_t := \operatorname*{argmin}_{\alpha\in\Delta(\mathcal{X}\cup\{\widehat{x}\})} \left\langle \mu_\alpha, \sum_{\tau=1}^{t-1}\widehat{\ell}_\tau \right\rangle - \frac{1}{\eta_t}\log\det\left(H(\alpha) - \mu_\alpha\mu_\alpha^\top\right),$$

$$\widetilde{p}_t := (1-\eta_t d)p_t + \eta_t d((1-p_{t,\widehat{x}})\pi_{\mathcal{X}} + p_{t,\widehat{x}}\pi_{\widehat{x}})$$

where $\pi_{\widehat{x}}$ denotes the sampling distribution that picks $\widehat{x}$ with probability 1.
Sample an action $A_t \sim \widetilde{p}_t$.
Construct loss estimator:

$$\widehat{\ell}_t(a) = a^\top\left(H(\widetilde{p}_t) - \mu_{\widetilde{p}_t}\mu_{\widetilde{p}_t}^\top\right)^{-1}(a_t - \mu_{\widetilde{p}_t})\ell_t(a_t).$$

end

We begin by using the following simplifying notation. For a distribution $\alpha \in \Delta(\mathcal{X}\cup\{\widehat{x}\})$, we define $\alpha_{\mathcal{X}}$ as the restricted distribution over $\mathcal{X}$ such that $\alpha_{\mathcal{X}} \propto \alpha$ over $\mathcal{X}$, i.e. $\alpha_{\mathcal{X},x} = \frac{\alpha_x}{1-\alpha_{\widehat{x}}}$ for any $x \in \mathcal{X}$. We further define

$$\begin{aligned}
&H(\alpha) = \mathbb{E}_{x\sim\alpha}[xx^\top], \quad \mu_\alpha = \mathbb{E}_{x\sim\alpha}[x],\\
&H_\alpha^{\mathbb{V}} = H(\alpha) - \mu_\alpha\mu_\alpha^\top,\\
&G(\alpha) = H(\alpha_{\mathcal{X}}), \quad m_\alpha = \mu_{\alpha_{\mathcal{X}}},\\
&G_\alpha^{\mathbb{V}} = G(\alpha) - m_\alpha m_\alpha^\top.
\end{aligned}$$

Lemma 35 *Assume $\alpha \in \Delta(\mathcal{X}\cup\{\widehat{x}\})$ is such that $G_\alpha^{\mathbb{V}}$ is of rank $d-1$ (i.e. full rank over $\mathcal{X}$). Then we have the following properties:*

$$H_\alpha^{\mathbb{V}} = (1-\alpha_{\widehat{x}})\left(G_\alpha^{\mathbb{V}} + \alpha_{\widehat{x}}(m_\alpha - \widehat{x})(m_\alpha-\widehat{x})^\top\right),$$

$$[H_\alpha^{\mathbb{V}}]^{-1} = \frac{1}{1-\alpha_{\widehat{x}}}\left([G_\alpha^{\mathbb{V}}]^+ + [G_\alpha^{\mathbb{V}}]^+ m_\alpha\widehat{x}^\top + \widehat{x}m_\alpha^\top[G_\alpha^{\mathbb{V}}]^+ + \left(\|m_\alpha\|^2_{[G_\alpha^{\mathbb{V}}]^+} + \frac{1}{\alpha_{\widehat{x}}}\right)\widehat{x}\widehat{x}^\top\right),$$

where $[G_\alpha^{\mathbb{V}}]^+$ denotes the pseudo-inverse.

Proof The first identity is a simple algebraic identity

$$\begin{aligned}H_\alpha^{\mathbb{V}} &= H(\alpha) - \mu_\alpha\mu_\alpha^\top = (1-\alpha_{\widehat{x}})G(\alpha) + \alpha_{\widehat{x}}\widehat{x}\widehat{x}^\top - ((1-\alpha_{\widehat{x}})m_\alpha + \alpha_{\widehat{x}}\widehat{x})((1-\alpha_{\widehat{x}})m_\alpha + \alpha_{\widehat{x}}\widehat{x})^\top \\ &= (1-\alpha_{\widehat{x}})\left(G_\alpha^{\mathbb{V}} + \alpha_{\widehat{x}}(m_\alpha - \widehat{x})(m_\alpha - \widehat{x})^\top\right),\end{aligned}$$

which holds for any α. For the second identity note that by the definition of $\widehat{x} = \binom{\mathbf{0}}{1}$ and $\forall x \in \mathcal{X}$: $\langle x, \widehat{x}\rangle = 0$, we have $G_\alpha^{\mathbb{V}}\widehat{x} = [G_\alpha^{\mathbb{V}}]^+\widehat{x} = 0$. Furthermore $G_\alpha^{\mathbb{V}}[G_\alpha^{\mathbb{V}}]^+ = I - \widehat{x}\widehat{x}^{-1}$ due to the rank $d-1$ assumption. Multiplying the two matrices yields

$$\begin{aligned}&\left(G_\alpha^{\mathbb{V}} + \alpha_{\widehat{x}}(m_\alpha - \widehat{x})(m_\alpha - \widehat{x})^\top\right)\cdot\left([G_\alpha^{\mathbb{V}}]^+ + [G_\alpha^{\mathbb{V}}]^+ m_\alpha\widehat{x}^\top + \widehat{x}m_\alpha^\top[G_\alpha^{\mathbb{V}}]^+ + \left(\|m_\alpha\|^2_{[G_\alpha^{\mathbb{V}}]^+} + \frac{1}{\alpha_{\widehat{x}}}\right)\widehat{x}\widehat{x}^\top\right) \\ &= I - \widehat{x}\widehat{x}^\top + m_\alpha\widehat{x}^\top + \alpha_{\widehat{x}}(m_\alpha - \widehat{x})\left([G_\alpha^{\mathbb{V}}]^+ m_\alpha + \|m_\alpha\|^2_{[G_\alpha^{\mathbb{V}}]^+}\,\widehat{x} - [G_\alpha^{\mathbb{V}}]^+ m_\alpha - \left(\|m_\alpha\|^2_{[G_\alpha^{\mathbb{V}}]^+} + \frac{1}{\alpha_{\widehat{x}}}\right)\widehat{x}\right)^\top \\ &= I - \widehat{x}\widehat{x}^\top + m_\alpha\widehat{x}^\top - (m_\alpha - \widehat{x})\widehat{x}^\top = I\end{aligned}$$

■

which implies for any $x, y \in \operatorname{span}(\mathcal{X})$

$$\langle x, [H_\alpha^{\mathbb{V}}]^{-1}y\rangle = \frac{\langle x, [G_\alpha^{\mathbb{V}}]^+ y\rangle}{1-\alpha_{\widehat{x}}}, \tag{11}$$

$$\langle x, [H_\alpha^{\mathbb{V}}]^{-1}(\widehat{x} - m_\alpha)\rangle = 0, \tag{12}$$

$$\|\widehat{x} - m_\alpha\|^2_{[H_\alpha^{\mathbb{V}}]^{-1}} = \frac{1}{(1-\alpha_{\widehat{x}})\alpha_{\widehat{x}}}. \tag{13}$$

Lemma 36 *Let $\pi_1, \pi_2 \in \Delta(\mathcal{X} \cup \{\widehat{x}\})$, then for any $\lambda \in [0,1]$:*

$$H^{\mathbb{V}}_{\lambda\pi_1+(1-\lambda)\pi_2} \succeq \lambda H^{\mathbb{V}}_{\pi_1}.$$

Proof Simple algebra shows

$$H^{\mathbb{V}}_{\lambda\pi_1+(1-\lambda)\pi_2} = \lambda H^{\mathbb{V}}_{\pi_1} + (1-\lambda)H^{\mathbb{V}}_{\pi_2} + \lambda(1-\lambda)(\mu_{\pi_1} - \mu_{\pi_2})(\mu_{\pi_1} - \mu_{\pi_2})^\top.$$

■

Lemma 37 *Let $\pi \in \Delta(\mathcal{X})$ be arbitrary, and $\tilde{\pi} = (1-\eta_t d)\pi + \eta_t d\pi_{\mathcal{X}}$ for $(\eta_t d) \in (0,1)$, then it holds for any $x \in \mathcal{X}$:*

$$\|m_\pi - m_{\tilde{\pi}}\|_{[G^{\mathbb{V}}_{\tilde{\pi}}]^+} \le \sqrt{\frac{\eta_t d}{1-\eta_t d}}$$

$$\|x - m_{\tilde{\pi}}\|_{[G^{\mathbb{V}}_{\tilde{\pi}}]^+} \le \frac{2}{\sqrt{\eta_t}}$$

Proof We have

$$G^{\mathbb{V}}_{\tilde{\pi}} = \eta_t d G^{\mathbb{V}}_{\pi_{\mathcal{X}}} + (1-\eta_t d)G^{\mathbb{V}}_{\pi} + \eta_t d(1-\eta_t d)(m_\pi - m_{\pi_{\mathcal{X}}})(m_\pi - m_{\pi_{\mathcal{X}}})^\top ,$$

hence

$$\begin{aligned}\|m_\pi - m_{\tilde{\pi}}\|^2_{[G^{\mathbb{V}}_{\tilde{\pi}}]^+} &= (\eta_t d)^2 \|m_\pi - m_{\pi_{\mathcal{X}}}\|^2_{\left[\eta_t d G^{\mathbb{V}}_{\pi_{\mathcal{X}}} + (1-\eta_t d)G^{\mathbb{V}}_{\pi} + \eta_t d(1-\eta_t d)(m_\pi - m_{\pi_{\mathcal{X}}})(m_\pi - m_{\pi_{\mathcal{X}}})^\top\right]^{-1}} \\ &\leq (\eta_t d)^2 \|m_\pi - m_{\pi_{\mathcal{X}}}\|^2_{\left[\eta_t d(1-\eta_t d)(m_\pi - m_{\pi_{\mathcal{X}}})(m_\pi - m_{\pi_{\mathcal{X}}})^\top\right]^+} = \frac{\eta_t d}{1-\eta_t d} .\end{aligned}$$

For the second inequality, we have

$$\begin{aligned}\|x - m_{\tilde{\pi}}\|_{[G^{\mathbb{V}}_{\tilde{\pi}}]^+} &\leq \|x - m_{\pi_{\mathcal{X}}}\|_{[G^{\mathbb{V}}_{\tilde{\pi}}]^+} + \|m_{\tilde{\pi}} - m_{\pi_{\mathcal{X}}}\|_{[G^{\mathbb{V}}_{\tilde{\pi}}]^+} \\ &\leq \frac{1}{\sqrt{\eta_t d}}\left(\|x - m_{\pi_{\mathcal{X}}}\|_{[G^{\mathbb{V}}_{\pi_{\mathcal{X}}}]^+} + \|m_{\tilde{\pi}} - m_{\pi_{\mathcal{X}}}\|_{[G^{\mathbb{V}}_{\pi_{\mathcal{X}}}]^+}\right) \leq \frac{2}{\sqrt{\eta_t}} ,\end{aligned}$$

where the last inequality uses that John's exploration satisfies $\|x - m_{\pi_{\mathcal{X}}}\|^2_{[G^{\mathbb{V}}_{\pi_{\mathcal{X}}}]^+} \leq d$ for all $x \in \mathcal{X}$.
■

Lemma 38 *The Bregman divergence between two distributions α, β over $\mathcal{X} \cup \{\widehat{x}\}$ with respect to the function $F(\alpha) = -\log\det\left(H^{\mathbb{V}}_{\alpha}\right)$ is bounded by*

$$D(\alpha,\beta) \geq D_{\log}(\alpha_{\widehat{x}}, \beta_{\widehat{x}}) + D_{\log}(1-\alpha_{\widehat{x}}, 1-\beta_{\widehat{x}}) + \frac{1-\alpha_{\widehat{x}}}{1-\beta_{\widehat{x}}}\|m_\alpha - m_\beta\|^2_{[G^{\mathbb{V}}_{\beta}]^+}$$

where $D_{\log}$ is the Bregman divergence of $-\log(x)$.

Proof We begin by simplifying $F(\alpha)$. Note that

$$\begin{aligned}H^{\mathbb{V}}_{\alpha} &= (1-\alpha_{\widehat{x}})\left(G^{\mathbb{V}}_{\alpha} + \alpha_{\widehat{x}}\widehat{x}^\top\right)^{\frac{1}{2}} M \left(G^{\mathbb{V}}_{\alpha} + \alpha_{\widehat{x}}\widehat{x}^\top\right)^{\frac{1}{2}} , \\ M &= I + (\sqrt{\alpha_{\widehat{x}}}[G^{\mathbb{V}}_{\alpha}]^+ m_\alpha - \widehat{x})(\sqrt{\alpha_{\widehat{x}}}[G^{\mathbb{V}}_{\alpha}]^+ m_\alpha - \widehat{x}))^\top - \widehat{x}\widehat{x}^\top .\end{aligned}$$

M is a matrix with $d-2$ eigenvalues of size 1, since it is the identity with two rank-1 updates. The product of the remaining eigenvalues is given by considering the determinant of the 2×2 sub-matrix with coordinates $\frac{[G^{\mathbb{V}}_{\alpha}]^+ m_\alpha}{\|[G^{\mathbb{V}}_{\alpha}]^+ m_\alpha\|}$ and $\widehat{x}$. We have that

$$\begin{aligned}\det(M) &= \widehat{x}^\top M \widehat{x} \cdot \frac{m_\alpha^\top [G^{\mathbb{V}}_{\alpha}]^+}{\|[G^{\mathbb{V}}_{\alpha}]^+ m_\alpha\|} M \frac{[G^{\mathbb{V}}_{\alpha}]^+ m_\alpha}{\|[G^{\mathbb{V}}_{\alpha}]^+ m_\alpha\|} - \left(\widehat{x}^\top M \frac{[G^{\mathbb{V}}_{\alpha}]^+ m_\alpha}{\|[G^{\mathbb{V}}_{\alpha}]^+ m_\alpha\|}\right)^2 \\ &= 1 \cdot \left(1 + \alpha_{\widehat{x}}\left\|[G^{\mathbb{V}}_{\alpha}]^+ m_\alpha\right\|^2\right) - \alpha_{\widehat{x}}\left\|[G^{\mathbb{V}}_{\alpha}]^+ m_\alpha\right\|^2 = 1.\end{aligned}$$

Hence we have

$$F(\alpha) = -\log(1-\alpha_{\widehat{x}}) - \log(\alpha_{\widehat{x}}) - \log\det_{d-1}((1-\alpha_{\widehat{x}})G^{\mathbb{V}}_{\alpha})$$

Where $\det_{d-1}$ is the determinant over the first $d-1$ eigenvalues of the submatrix of the first $(d-1)$ coordinates. The derivative term is given by

$$
\begin{aligned}
\langle \alpha - \beta, \nabla F(\beta)\rangle &= \sum_{x\in\mathcal{X}\cup\{\widehat{x}\}} (\alpha_x - \beta_x)\left\|x-\mu_\beta\right\|^2_{[H^{\mathbb{V}}_\beta]^{-1}} \\
&= \sum_{x\in\mathcal{X}\cup\{\widehat{x}\}} \alpha_x \left\|x-\mu_\beta\right\|^2_{[H^{\mathbb{V}}_\beta]^{-1}} - d \\
&= \operatorname{Tr}\left(\sum_{x\in\mathcal{X}\cup\{\widehat{x}\}} \alpha_x (x-\mu_\beta)(x-\mu_\beta)^\top [H^{\mathbb{V}}_\beta]^{-1}\right) - d \\
&= \operatorname{Tr}\left(\left(H^{\mathbb{V}}_\alpha + (\mu_\alpha-\mu_\beta)(\mu_\alpha-\mu_\beta)^\top\right)[H^{\mathbb{V}}_\beta]^{-1}\right) - d \\
&= \operatorname{Tr}\left(H^{\mathbb{V}}_\alpha [H^{\mathbb{V}}_\beta]^{-1}\right) + \left\|\mu_\alpha-\mu_\beta\right\|^2_{[H^{\mathbb{V}}_\beta]^{-1}} - d.
\end{aligned}
$$

The first term is

$$
\begin{aligned}
&\operatorname{Tr}\left(H^{\mathbb{V}}_\alpha [H^{\mathbb{V}}_\beta]^{-1}\right) \\
&= \operatorname{Tr}\left((1-\alpha_{\widehat{x}})G^{\mathbb{V}}_\alpha [H^{\mathbb{V}}_\beta]^{-1}\right) + \alpha_{\widehat{x}}(1-\alpha_{\widehat{x}})\left\|\widehat{x}-m_\alpha\right\|^2_{[H^{\mathbb{V}}_\beta]^{-1}} && \text{(by Lemma 35)} \\
&= \operatorname{Tr}_{d-1}\left(\frac{1-\alpha_{\widehat{x}}}{1-\beta_{\widehat{x}}} G^{\mathbb{V}}_\alpha [G^{\mathbb{V}}_\beta]^{+}\right) + \alpha_{\widehat{x}}(1-\alpha_{\widehat{x}})\left(\left\|\widehat{x}-m_\beta\right\|^2_{[H^{\mathbb{V}}_\beta]^{-1}} + \left\|m_\alpha-m_\beta\right\|^2_{[H^{\mathbb{V}}_\beta]^{-1}}\right) \\
& && \text{(by Lemma 35)} \\
&= \operatorname{Tr}_{d-1}\left(\frac{1-\alpha_{\widehat{x}}}{1-\beta_{\widehat{x}}} G^{\mathbb{V}}_\alpha [G^{\mathbb{V}}_\beta]^{+}\right) + \frac{\alpha_{\widehat{x}}(1-\alpha_{\widehat{x}})}{\beta_{\widehat{x}}(1-\beta_{\widehat{x}})} + \frac{\alpha_{\widehat{x}}(1-\alpha_{\widehat{x}})\left\|m_\alpha-m_\beta\right\|^2_{[G^{\mathbb{V}}_\beta]^{+}}}{1-\beta_{\widehat{x}}}. && \text{(by (13))}
\end{aligned}
$$

The norm term is

$$
\begin{aligned}
&\left\|\mu_\alpha-\mu_\beta\right\|^2_{[H^{\mathbb{V}}_\beta]^{-1}} \\
&= \left\|\mu_\alpha - (\alpha_{\widehat{x}}\widehat{x} + (1-\alpha_{\widehat{x}})m_\beta)\right\|^2_{[H^{\mathbb{V}}_\beta]^{-1}} + \left\|(\alpha_{\widehat{x}}\widehat{x} + (1-\alpha_{\widehat{x}})m_\beta) - \mu_\beta\right\|^2_{[H^{\mathbb{V}}_\beta]^{-1}} \\
&= (1-\alpha_{\widehat{x}})^2\left\|m_\alpha-m_\beta\right\|^2_{[H^{\mathbb{V}}_\beta]^{-1}} + (\alpha_{\widehat{x}}-\beta_{\widehat{x}})^2\left\|\widehat{x}-m_\beta\right\|^2_{[H^{\mathbb{V}}_\beta]^{-1}} \\
&= \frac{(1-\alpha_{\widehat{x}})^2\left\|m_\alpha-m_\beta\right\|^2_{[G^{\mathbb{V}}_\beta]^{+}}}{1-\beta_{\widehat{x}}} + \frac{(\alpha_{\widehat{x}}-\beta_{\widehat{x}})^2}{\beta_{\widehat{x}}(1-\beta_{\widehat{x}})}.
\end{aligned}
$$

Combining both terms

$$
\begin{aligned}
&\langle \alpha - \beta, \nabla F(\beta)\rangle \\
&= \operatorname{Tr}_{d-1}\left(\frac{1-\alpha_{\widehat{x}}}{1-\beta_{\widehat{x}}} G^{\mathbb{V}}_\alpha [G^{\mathbb{V}}_\beta]^{+}\right) + \frac{\alpha_{\widehat{x}}(1-\alpha_{\widehat{x}}) + (\alpha_{\widehat{x}}-\beta_{\widehat{x}})^2}{\beta_{\widehat{x}}(1-\beta_{\widehat{x}})} + \frac{1-\alpha_{\widehat{x}}}{1-\beta_{\widehat{x}}}\left\|m_\alpha-m_\beta\right\|^2_{[G^{\mathbb{V}}_\beta]^{+}} - d \\
&= \operatorname{Tr}_{d-1}\left(\frac{1-\alpha_{\widehat{x}}}{1-\beta_{\widehat{x}}} G^{\mathbb{V}}_\alpha [G^{\mathbb{V}}_\beta]^{+}\right) + \frac{\alpha_{\widehat{x}}}{\beta_{\widehat{x}}} + \frac{1-\alpha_{\widehat{x}}}{1-\beta_{\widehat{x}}} - 1 + \frac{1-\alpha_{\widehat{x}}}{1-\beta_{\widehat{x}}}\left\|m_\alpha-m_\beta\right\|^2_{[G^{\mathbb{V}}_\beta]^{+}} - d\,.
\end{aligned}
$$

Combining the everything

$$
\begin{aligned}
&D(\alpha, \beta) \\
&= D_{\log}(\alpha_{\widehat{x}}, \beta_{\widehat{x}}) + D_{\log}(1-\alpha_{\widehat{x}}, 1-\beta_{\widehat{x}}) + \frac{1-\alpha_{\widehat{x}}}{1-\beta_{\widehat{x}}} \|m_\alpha - m_\beta\|^2_{[G^{\mathbb{V}}_\beta]^+} + D_{d-1}\left((1-\alpha_{\widehat{x}})G^{\mathbb{V}}_\alpha, (1-\beta_{\widehat{x}})G^{\mathbb{V}}_\beta\right) \\
&\geq D_{\log}(\alpha_{\widehat{x}}, \beta_{\widehat{x}}) + D_{\log}(1-\alpha_{\widehat{x}}, 1-\beta_{\widehat{x}}) + \frac{1-\alpha_{\widehat{x}}}{1-\beta_{\widehat{x}}} \|m_\alpha - m_\beta\|^2_{[G^{\mathbb{V}}_\beta]^+} \;,
\end{aligned}
$$

where the last inequality follows from the positiveness of Bregman divergences. ■

Lemma 39 *For any $b \in (0,1)$, any $x \in \mathbb{R}$ and $\eta \leq \frac{1}{2|x|}$, it holds that*

$$
\sup_{\alpha \in [0,1]} |a-b|x - \frac{1}{\eta} D_{\log}(a,b) \leq \eta b^2 x^2 \,.
$$

Proof The statement is equivalent to

$$
\sup_{a \in [0,1]} f(a) = \sup_{a \in [0,1]} (b-a)x - \frac{1}{\eta} D_{\log}(a,b) \leq \eta b^2 x^2 \,,
$$

since x can take positive or negative sign. The function is concave, so setting the derivative to 0 is the optimal solution if that value lies in $(0, \infty)$.

$$
f'(a) = -x + \frac{1}{\eta}\left(\frac{1}{a} - \frac{1}{b}\right) \qquad a^\star = \frac{b}{1+\eta b x} \,.
$$

Plugging this in, leads to

$$
\begin{aligned}
f(a^\star) &= \frac{\eta b^2 x^2}{1+\eta b x} - \frac{1}{\eta}\left(\log(1+\eta b x) + \frac{1}{1+\eta b x} - 1\right) \\
&\leq \frac{\eta b^2 x^2}{1+\eta b x} - \frac{1}{\eta}\left(\eta b x - \eta^2 b^2 x^2 - \frac{\eta b x}{1+\eta b x}\right) = \eta b^2 x^2 \,,
\end{aligned}
$$

where the last line uses $\log(1+x) \geq x - x^2$ for any $x \geq -\frac{1}{2}$. ■

Lemma 40 *The stability term*

$$
\mathit{stab}_t := \sup_{\alpha \in \Delta(\mathcal{X} \cup \{\widehat{x}\})} \langle \mu_{p_t} - \mu_\alpha, \widehat{\ell}_t \rangle - \frac{1}{\eta_t} D(\alpha, p_t) \,,
$$

satisfies

$$
\mathbb{E}_t[\mathit{stab}_t] = O((1-p_{t,\widehat{x}})\eta_t d) \,.
$$

Proof We have

$$\mu_{p_t} - \mu_\alpha = (p_{t,\widehat{x}} - \alpha_{\widehat{x}})\widehat{x} + (1-p_{t,\widehat{x}})m_{p_t} - (1-\alpha_{\widehat{x}})m_\alpha$$
$$= (p_{t,\widehat{x}} - \alpha_{\widehat{x}})(\widehat{x} - m_{\widetilde{p}_t} + m_{\widetilde{p}_t} - m_{p_t}) + (1-\alpha_{\widehat{x}})(m_{\widetilde{p}_t} - m_\alpha)\,.$$

Hence for $A_t = y \in \mathcal{X}$:

$$\begin{aligned}
\langle \mu_{p_t} - \mu_\alpha, \widehat{\ell}_t\rangle &= \langle \mu_{p_t} - \mu_\alpha, [H^{\mathbb{V}}_{\widetilde{p}_t}]^{-1}(y-\mu_{\widetilde{p}_t})\rangle \ell_{t,A_t} \\
&= (p_{t,\widehat{x}} - \alpha_{\widehat{x}})\langle \widehat{x} - m_{\widetilde{p}_t}, [H^{\mathbb{V}}_{\widetilde{p}_t}]^{-1}(y-\mu_{\widetilde{p}_t})\rangle \ell_{t,A_t} \\
&\quad + (p_{t,\widehat{x}} - \alpha_{\widehat{x}})\langle m_{\widetilde{p}_t} - m_{p_t}, [H^{\mathbb{V}}_{\widetilde{p}_t}]^{-1}(y-\mu_{\widetilde{p}_t})\rangle \ell_{t,A_t} \\
&\quad + (1-\alpha_{\widehat{x}})\langle m_{\widetilde{p}_t} - m_\alpha, [H^{\mathbb{V}}_{\widetilde{p}_t}]^{-1}(y-\mu_{\widetilde{p}_t})\rangle \ell_{t,A_t} \\
&= (p_{t,\widehat{x}} - \alpha_{\widehat{x}})\langle \widehat{x} - m_{\widetilde{p}_t}, [H^{\mathbb{V}}_{\widetilde{p}_t}]^{-1}(m_{\widetilde{p}_t}-\mu_{\widetilde{p}_t})\rangle \ell_{t,A_t} && \text{(By Eq. (12))} \\
&\quad + (p_{t,\widehat{x}} - \alpha_{\widehat{x}})\langle m_{\widetilde{p}_t} - m_{p_t}, [H^{\mathbb{V}}_{\widetilde{p}_t}]^{-1}(y-m_{\widetilde{p}_t})\rangle \ell_{t,A_t} \\
&\quad + (1-\alpha_{\widehat{x}})\langle m_{\widetilde{p}_t} - m_\alpha, [H^{\mathbb{V}}_{\widetilde{p}_t}]^{-1}(y-m_{\widetilde{p}_t})\rangle \ell_{t,A_t} \\
&\le \frac{|\alpha_{\widehat{x}} - p_{t,\widehat{x}}|}{1-p_{t,\widehat{x}}} + \frac{|p_{t,\widehat{x}} - \alpha_{\widehat{x}}|}{1-p_{t,\widehat{x}}}|\langle m_{\widetilde{p}_t} - m_{p_t}, [H^{\mathbb{V}}_{\widetilde{p}_t}]^{-1}(y-m_{\widetilde{p}_t})\rangle| \\
&+ \quad \frac{1-\alpha_{\widehat{x}}}{1-p_{t,\widehat{x}}}|\langle m_{\widetilde{p}_t} - m_\alpha, [G^{\mathbb{V}}_{\widetilde{p}_t}]^{+}(y-m_{\widetilde{p}_t})\rangle| && \text{(By Eq. (13) and Eq. (11))} \\
&\le \frac{|p_{t,\widehat{x}} - \alpha_{\widehat{x}}|}{1-p_{t,\widehat{x}}}(1+2\sqrt{d}) \\
&\quad + \frac{4}{3}\times\frac{1-\alpha_{\widehat{x}}}{1-p_{t,\widehat{x}}}\left\|m_{p_t} - m_\alpha\right\|_{[G^{\mathbb{V}}_{p_t}]^+}\left\|y - m_{\widetilde{p}_t})\right\|_{[G^{\mathbb{V}}_{\widetilde{p}_t}]^+} \\
& && \text{(Cauchy-Schwarz and Lemma 37, Lemma 36)}
\end{aligned}$$

Equivalently for $A_t = \widehat{x}$,

$$\begin{aligned}
\langle \mu_{p_t} - \mu_\alpha, \widehat{\ell}_t\rangle &= \langle \mu_{\widetilde{p}_t} - \mu_\alpha, [H^{\mathbb{V}}_{\widetilde{p}_t}]^{-1}(\widehat{x}-\mu_{\widetilde{p}_t})\rangle \ell_{t,A_t} \\
&= (p_{t,\widehat{x}} - \alpha_{\widehat{x}})\langle \widehat{x} - m_{\widetilde{p}_t}, [H^{\mathbb{V}}_{\widetilde{p}_t}]^{-1}(\widehat{x}-\mu_{\widetilde{p}_t})\rangle \widehat{\ell}_{t,A_t} \\
&\le \frac{|p_{t,\widehat{x}} - \alpha_{\widehat{x}}|}{p_{t,\widehat{x}}}\,.
\end{aligned}$$

Hence the stability term for $A_t \in \mathcal{X}$, is bounded by

$$\begin{aligned}
&\sup_{\alpha\in\Delta(\mathcal{X}\cup\{\widehat{x}\})} \langle \mu_{p_t} - \mu_\alpha, \widehat{\ell}_t\rangle - \frac{1}{\eta_t}D(\alpha, p_t) \\
&\le \sup_{\alpha\in\Delta(\mathcal{X}\cup\{\widehat{x}\})} \frac{|\alpha_{\widehat{x}} - p_{t,\widehat{x}}|}{1-p_{t,\widehat{x}}}(1+2\sqrt{d}) - \frac{1}{\eta_t}D(1-\alpha_{\widehat{x}}, 1-p_{t,\widehat{x}}) \\
&\quad + \frac{1-\alpha_{\widehat{x}}}{1-p_{t,\widehat{x}}}\left(\frac{4}{3}\left\|m_\alpha - m_{\widetilde{p}_t}\right\|_{[G^{\mathbb{V}}_{\widetilde{p}_t}]^+}\left\|y - m_{\widetilde{p}_t}\right\|_{[G^{\mathbb{V}}_{\widetilde{p}_t}]^+} - \frac{1}{\eta_t}\left\|m_\alpha - m_{\widetilde{p}_t}\right\|^2_{[G^{\mathbb{V}}_{\widetilde{p}_t}]^+}\right) \\
&\le \eta_t\left\|y - m_{\widetilde{p}_t}\right\|^2_{[G^{\mathbb{V}}_{\widetilde{p}_t}]^+} + \sup_{\alpha_{\widehat{x}}\in[0,1]} \frac{|\alpha_{\widehat{x}} - p_{t,\widehat{x}}|}{1-p_{t,\widehat{x}}}(1+2\sqrt{d})
\end{aligned}$$

$$
\begin{aligned}
&+ \eta_t \frac{|\alpha_{\widehat{x}} - p_{t,\widehat{x}}|}{1 - p_{t,\widehat{x}}} \left\|y - m_{\widetilde{p}_t}\right\|^2_{[G^{\mathbb{V}}_{\widetilde{p}_t}]^+} - \frac{1}{\eta_t} D(1 - \alpha_{\widehat{x}}, 1 - p_{t,\widehat{x}}) && \text{(AM-GM inequality)} \\
\leq\ & \eta_t \left\|y - m_{\widetilde{p}_t}\right\|^2_{[G^{\mathbb{V}}_{\widetilde{p}_t}]^+} \\
&+ \sup_{\alpha_{\widehat{x}} \in [0,1]} \frac{|\alpha_{\widehat{x}} - p_{t,\widehat{x}}|}{1 - p_{t,\widehat{x}}} (5 + 2\sqrt{d}) - \frac{1}{\eta_t} D(1 - \alpha_{\widehat{x}}, 1 - p_{t,\widehat{x}}) && \text{(Lemma 37)} \\
\leq\ & \eta_t \left\|y - m_{\widetilde{p}_t}\right\|^2_{[G^{\mathbb{V}}_{\widetilde{p}_t}]^+} + O(\eta_t d)\,. && \text{(Lemma 39)}
\end{aligned}
$$

Taking the expectation over $y \sim \widetilde{p}_{t,\mathcal{X}}$ leads to

$$
\mathbb{E}_{A_t \sim \widetilde{p}_{t,\mathcal{X}}}\left[\mathit{stab}_t\right] = O(\eta_t d)
$$

For $A_t = \widehat{x}$ we have two cases. If $p_{t,\widehat{x}} \geq \frac{1}{2}$, then by Lemma 39,

$$
\mathit{stab}_t \leq \sup_{\alpha_{\widehat{x}} \in [0,1]} \frac{|p_{t,\widehat{x}} - \alpha_{\widehat{x}}|}{p_{t,\widehat{x}}} - \frac{1}{\eta_t} D_{\log}(1 - \alpha_{\widehat{x}}, 1 - p_{t,\widehat{x}}) \leq O\left(\eta_t (1 - p_{t,\widehat{x}})^2 \times \frac{1}{p^2_{t,\widehat{x}}}\right) \leq O\left(\eta_t (1 - p_{t,\widehat{x}})\right)\,.
$$

Otherwise If $1 - p_{t,\widehat{x}} \geq \frac{1}{2}$, then by Lemma 39,

$$
\mathit{stab}_t \leq \sup_{\alpha_{\widehat{x}} \in [0,1]} \frac{|p_{t,\widehat{x}} - \alpha_{\widehat{x}}|}{p_{t,\widehat{x}}} - \frac{1}{\eta_t} D_{\log}(\alpha_{\widehat{x}}, p_{t,\widehat{x}}) \leq O\left(\eta_t p^2_{t,\widehat{x}} \times \frac{1}{p^2_{t,\widehat{x}}}\right) \leq O(\eta_t (1 - p_{t,\widehat{x}}))\,.
$$

Finally we have

$$
\mathbb{E}_t\left[\mathit{stab}_t\right] = (1 - p_{t,\widehat{x}})\mathbb{E}_{A_t \sim p_{t,\mathcal{X}}}[\mathit{stab}_t] + p_{t,\widehat{x}} \mathbb{E}_{A_t \sim \pi_{\widehat{x}}}[\mathit{stab}_t] = O((1 - p_{t,\widehat{x}})\eta_t d)
$$

■

Proof [Proof of Lemma 5] By standard analysis of FTRL (Lemma 27) and Lemma 40, for any τ and x

$$
\sum_{t=1}^{\tau} \mathbb{E}_t\left[\langle p_t, \widehat{\ell}_t\rangle - \widehat{\ell}_t(x)\right] \leq \frac{d \log T}{\eta_\tau} + \sum_{t=1}^{\tau} O(\eta_t (1 - p_{t,\widehat{x}}) d) \leq O\left(\sqrt{d \log T \sum_{t=1}^{\tau} (1 - \widetilde{p}_{t,\widehat{x}})}\right)\,,
$$

where the last inequality follows from the definition of learning rate and $\widetilde{p}_{t,\widehat{x}} = p_{t,\widehat{x}}$. Additionally, we have

$$
\mathbb{E}_t\left[\langle \widetilde{p}_t - p_t, \widehat{\ell}_t\rangle\right] = \langle \widetilde{p}_t - p_t, \ell_t\rangle \leq \eta_t d (1 - p_{t,\widehat{x}})\,,
$$

so that

$$
\sum_{t=1}^{\tau} \mathbb{E}_t\left[\langle \widetilde{p}_t, \widehat{\ell}_t\rangle - \widehat{\ell}_t(x)\right] = O\left(\sqrt{d \log T \sum_{t=1}^{\tau} (1 - \widetilde{p}_{t,\widehat{x}})}\right)\,.
$$

Taking expectations on both sides finishes the proof. ■

Appendix D. Proof of Proposition 2

Proof [Proof of Proposition 2] The guarantee in the adversarial regime is directly by Definition 1. In the corrupted stochastic regime, if $x^\star \neq \text{argmin}_{u\in\mathcal{X}} \mathbb{E}[\sum_{t=1}^T \ell_{t,u}]$, then it holds that $T\Delta - C = \sum_{t=1}^T(\Delta - C_t) \leq \sum_{t=1}^T \mathbb{E}[\ell_{t,u} - \ell_{t,x^\star}] < 0$ for some $u \neq x^\star$. Therefore, by Definition 1, we have that for any $u \in \mathcal{X}$,

$$\mathbb{E}\left[\sum_{t=1}^T(\ell_{t,A_t} - \ell_{t,u})\right] \leq (c_1 \log T)^{1-\alpha}T^\alpha + c_2 \log T \leq (c_1 \log T)^{1-\alpha}\left(C\Delta^{-1}\right)^\alpha + c_2 \log T.$$

Now, assume that $x^\star = \text{argmin}_{u\in\mathcal{X}} \mathbb{E}[\sum_{t=1}^T \ell_{t,u}]$. By Definition 1, we have

$$\mathbb{E}\left[\sum_{t=1}^T(\ell_{t,A_t} - \ell_{t,x^\star})\right] \leq (c_1 \log T)^{1-\alpha}\mathbb{E}\left[\sum_{t=1}^T(1 - p_{t,x^\star})\right]^\alpha + c_2 \log T.$$

On the other hand,

$$\mathbb{E}\left[\sum_{t=1}^T(\ell_{t,A_t} - \ell_{t,x^\star})\right] \geq \Delta\mathbb{E}\left[\sum_{t=1}^T(1 - p_{t,x^\star})\right] - C.$$

Combining the two inequalities, we get

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T(\ell_{t,A_t} - \ell_{t,x^\star})\right] \\
&= \sup_{\lambda\in[0,1]} (1+\lambda)\mathbb{E}\left[\sum_{t=1}^T(\ell_{t,A_t} - \ell_{t,x^\star})\right] - \lambda\mathbb{E}\left[\sum_{t=1}^T(\ell_{t,A_t} - \ell_{t,x^\star})\right] \\
&\leq \sup_{\lambda\in[0,1]} 2(c_1 \log T)^{1-\alpha}\mathbb{E}\left[\sum_{t=1}^T(1 - p_{t,x^\star})\right]^\alpha + 2c_2 \log T - \lambda\left(\Delta\mathbb{E}\left[\sum_{t=1}^T(1 - p_{t,x^\star})\right] - C\right) \\
&\leq O\left(c_1 \log(T)\Delta^{-\frac{\alpha}{1-\alpha}} + (c_1 \log T)^{1-\alpha}(C\Delta^{-1})^\alpha + c_2 \log T\right). \qquad \text{(using Lemma 41)}
\end{aligned}$$

■

Appendix E. Analysis for the First Reduction

E.1. BOBW to LSB (Algorithm 1 / Theorem 6)

Proof [Proof of Theorem 6] We use $\tau_k = T_{k+1} - T_k$ to denote the length of epoch k, and let n be the last epoch (define $T_{n+1} = T$). Also, we use $\mathbb{E}_t[\cdot]$ to denote expectation conditioned on all history up to time t. We first consider the adversarial regime. By the definition of local-self-bounding algorithms, we have for any u,

$$\mathbb{E}_{T_k}\left[\sum_{t=T_k+1}^{T_{k+1}}(\ell_{t,A_t} - \ell_{t,u})\right] \leq c_0^{1-\alpha}\mathbb{E}_{T_k}[\tau_k]^\alpha + c_2 \log(T),$$

which implies

$$\mathbb{E}\left[\sum_{t=T_k+1}^{T_{k+1}} (\ell_{t,A_t} - \ell_{t,u})\right] \le c_0^{1-\alpha}\mathbb{E}[\tau_k]^\alpha + c_2 \log(T)$$

using the property $\mathbb{E}[x^\alpha] \le \mathbb{E}[x]^\alpha$ for $x \in \mathbb{R}_+$ and $0 < \alpha < 1$. Summing the bounds over $k = 1, 2, \ldots, n$, and using the fact that $\tau_k \ge 2\tau_{k-1}$ for all $k < n$, we get

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} (\ell_{t,A_t} - \ell_{t,u})\right] \\
&\le c_0^{1-\alpha}\left(\mathbb{E}\left[\tau_n\right]^\alpha + \mathbb{E}\left[\tau_{n-1}\right]^\alpha\left(1 + \frac{1}{2^\alpha} + \frac{1}{2^{2\alpha}} + \cdots\right)\right) + c_2\log(T)\log_2\left(\frac{T}{c_2\log(T)}\right) \\
&\le O\left(c_0^{1-\alpha}T^\alpha + c_2\log^2(T)\right). \qquad (14)
\end{aligned}$$

The same analysis also gives

$$\mathbb{E}\left[\sum_{t=1}^{T} (\ell_{t,A_t} - \ell_{t,u})\right] \le O\left((c_1 \log T)^{1-\alpha}T^\alpha + c_2\log^2(T)\right). \qquad (15)$$

Next, consider the corrupted stochastic regime. We first argue that it suffices to consider the regret comparator $x^\star$. This is because if $x^\star \neq \operatorname{argmin}_{u\in\mathcal{X}} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,u}\right]$, then it holds that $C \ge T\Delta$. Then the right-hand side of (15) is further upper bounded by

$$O\left((c_1 \log T)^{1-\alpha}(C\Delta^{-1})^\alpha + c_2 \log(T)\log(C\Delta^{-1})\right),$$

which fulfills the requirement for the stochastic regime. Below, we focus on bounding the pseudo-regret with respect to $x^\star$.

Let $m = \max\{k \in \mathbb{N} \,|\, \widehat{x}_k \neq x^\star\}$. Notice that m is either the last epoch (i.e., $m = n$) or the second last (i.e., $m = n-1$), because for any two consecutive epochs, at least one of them must have $\widehat{x}_k \neq x^\star$. Below we show that $|\{t \in [T_{m+1}] \,|\, A_t \neq x^\star\}| \ge \frac{T_{m+1}}{8} - 2\tau_0$.

If $\tau_m > 2\tau_{m-1}$, by the fact that the termination condition was not triggered one round earlier, the number of plays $N_m(x^\star)$ in epoch m is at most $\frac{\tau_m - 1}{2} + 1 = \frac{\tau_m + 1}{2}$; in other words, the number of times $\sum_{t=T_m+1}^{T_{m+1}} \mathbb{I}\{A_t \neq x^\star\}$ is at least $\frac{\tau_m - 1}{2}$. Further notice that because $\tau_m > 2\tau_{m-1} \ge 4\tau_{m-2} > \cdots$, we have $\frac{\tau_m - 1}{2} \ge \frac{1}{4}\sum_{k=1}^{m} \tau_k - \frac{1}{2} = \frac{T_{m+1}}{4} - \frac{1}{2}$.

Now consider the case $m > 1$ and $\tau_m \le 2\tau_{m-1}$. Recall that $\widehat{x}_m$ is the action with $N_{m-1}(\widehat{x}_m) \ge \frac{\tau_{m-1}}{2}$. This implies that $\sum_{t=T_{m-1}+1}^{T_m} \mathbb{I}\{A_t \neq x^\star\} \ge \frac{\tau_{m-1}}{2} \ge \frac{1}{2}\max\left\{\frac{\tau_m}{2}, \frac{1}{2}\sum_{k=1}^{m-1} \tau_k\right\} \ge \frac{1}{8}\sum_{k=1}^{m} \tau_k = \frac{T_{m+1}}{8}$.

Finally, consider the case $m = 1$ and $\tau_1 \le 2\tau_0$, then we have $T_2 - 2\tau_0 \le 0$ and the statement holds trivially.

The regret up to and including epoch m can be lower and upper bounded using the self-bounding technique:

$$\mathbb{E}\left[\sum_{t=1}^{T_{m+1}}(\ell_{t,A_t}-\ell_{t,x^\star})\right]=(1+\lambda)\mathbb{E}\left[\sum_{t=1}^{T_{m+1}}(\ell_{t,A_t}-\ell_{t,x^\star})\right]-\lambda\mathbb{E}\left[\sum_{t=1}^{T_{m+1}}(\ell_{t,A_t}-\ell_{t,x^\star})\right]$$

(for $0\le\lambda\le 1$)

$$\le O\left((c_1\log T)^{1-\alpha}\mathbb{E}\left[T_{m+1}\right]^{\alpha}+c_2\log(T)\log\left(\frac{\mathbb{E}[T_{m+1}]}{c_2\log(T)}\right)\right)-\lambda\left(\left(\frac{1}{8}\mathbb{E}\left[T_{m+1}\right]-c_2\log(T)\right)\Delta-C\right)$$

(the first term is by a similar calculation as (14), but replacing T by T_m and c_0 by $c_1\log T$)

$$\le O\left(c_1\log(T)\Delta^{-\frac{\alpha}{1-\alpha}}+(c_1\log T)^{1-\alpha}\left(C\Delta^{-1}\right)^{\alpha}+c_2\log(T)\log(C\Delta^{-1})\right)$$

where in the last inequality we use Lemma 41.

If m is not the last epoch, then it holds that $\widehat{x}_n=x^\star$ for the final epoch n. In this case, the regret in the final epoch is

$$\mathbb{E}\left[\sum_{t=T_n+1}^{T}(\ell_{t,A_t}-\ell_{t,x^\star})\right]=(1+\lambda)\mathbb{E}\left[\sum_{t=T_n+1}^{T}(\ell_{t,A_t}-\ell_{t,x^\star})\right]-\lambda\mathbb{E}\left[\sum_{t=T_n+1}^{T}(\ell_{t,A_t}-\ell_{t,x^\star})\right]$$

$$\le O\left((c_1\log T)^{1-\alpha}\mathbb{E}\left[\sum_{t=T_n+1}^{T}(1-p_{t,x^\star})\right]^{\alpha}\right)-\lambda\left(\mathbb{E}\left[\sum_{t=T_n+1}^{T}(1-p_{t,x^\star})\right]\Delta-C\right)+c_2\log(T)$$

$$\le O\left(c_1\log(T)\Delta^{-\frac{\alpha}{1-\alpha}}+(c_1\log T)^{1-\alpha}\left(C\Delta^{-1}\right)^{\alpha}\right)+c_2\log(T). \qquad \text{(Lemma 41)}$$

■

Lemma 41 *For $\Delta\in(0,1]$ and $c,c',X\ge 1$ and $C\ge 0$, we have*

$$\min_{\lambda\in[0,1]}\left\{\frac{1}{2}c^{1-\alpha}X^{\alpha}+\frac{1}{2}c'\log X-\lambda(X\Delta-C)\right\}\le c\Delta^{-\frac{\alpha}{1-\alpha}}+2c^{1-\alpha}\left(\frac{C}{\Delta}\right)^{\alpha}+2c'\log\left(1+\frac{c'+C}{\Delta}\right)$$

Proof If $c^{1-\alpha}X^{\alpha}\ge c'\log T$, we have

$$\frac{1}{2}c^{1-\alpha}X^{\alpha}+\frac{1}{2}c'\log X\le c^{1-\alpha}X^{\alpha}\le\lambda X\Delta+c\lambda^{-\frac{\alpha}{1-\alpha}}\Delta^{-\frac{\alpha}{1-\alpha}}.$$

where the last inequality is by the weighted AM-GM inequality. Therefore,

$$\min_{\lambda\in[0,1]}\left\{\frac{1}{2}c^{1-\alpha}X^{\alpha}+\frac{1}{2}c'\log X-\lambda(X\Delta-C)\right\}\le\min_{\lambda\in[0,1]}c\lambda^{-\frac{\alpha}{1-\alpha}}\Delta^{-\frac{\alpha}{1-\alpha}}+C\lambda.$$

Choosing $\lambda=\min\{1,c^{1-\alpha}C^{-(1-\alpha)}\Delta^{-\alpha}\}$, we bound the last expression by

$$c\max\left\{1,\left(c^{1-\alpha}C^{-(1-\alpha)}\Delta^{-\alpha}\right)^{-\frac{\alpha}{1-\alpha}}\right\}\Delta^{-\frac{\alpha}{1-\alpha}}+c^{1-\alpha}C^{\alpha}\Delta^{-\alpha}$$
$$\le c\max\left\{1,c^{-\alpha}C^{\alpha}\Delta^{\frac{\alpha^2}{1-\alpha}}\right\}\Delta^{-\frac{\alpha}{1-\alpha}}+c^{1-\alpha}C^{\alpha}\Delta^{-\alpha}$$
$$\le c\Delta^{-\frac{\alpha}{1-\alpha}}+2c^{1-\alpha}C^{\alpha}\Delta^{-\alpha}.$$

If $c^{1-\alpha}X^\alpha \le c' \log T$, we have

$$\frac{1}{2}c^{1-\alpha}X^\alpha + \frac{1}{2}c'\log X - \lambda(X\Delta - C) \le c'\log X - \lambda X\Delta + \lambda C \le c'\log\left(1 + \frac{c'^2}{\lambda^2\Delta^2}\right) + \lambda C$$

where the last inequality is because if $X \ge \frac{c'^2}{\lambda^2\Delta^2}$ then $c'\log X - \lambda X\Delta \le c'\log X - c'\sqrt{X} < 0$. Choosing $\lambda = \min\{1, \frac{c'}{C}\}$, we bound the last expression by $c'\log(1+(c'^2+C^2)/\Delta^2) \le 2c'\log(1+(c'+C)/\Delta)$. Combining cases finishes the proof. ■

E.2. dd-BOBW to dd-LSB (Algorithm 1 / Theorem 23)

Proof [Proof of Theorem 23] In the adversarial regime, we have that the regret in each phase k is bounded by

$$\mathbb{E}_{T_k}\left[\sum_{t=T_k+1}^{T_{k+1}}(\ell_{t,A_t} - \ell_{t,u})\right] \le \sqrt{c_1\log(T)\mathbb{E}_{T_k}\left[\sum_{t=T_k+1}^{T_{k+1}}\sum_x p_{t,x}\xi_{t,x}\right]} + c_2\log T\,.$$

We have maximally $\log T$ episodes, since the length doubles every time. Via Cauchy-Schwarz, we get

$$\sum_{k=1}^{k_{\max}}\mathbb{E}_{T_k}\left[\sum_{t=T_k+1}^{T_{k+1}}(\ell_{t,A_t} - \ell_{t,u})\right] \le \sqrt{c_1\log T\sum_{k=1}^{k_{\max}}\mathbb{E}_{T_k}\left[\sum_{t=T_k+1}^{T_{k+1}}\sum_x p_{t,x}\xi_{t,x}\right]}\sqrt{\log T} + c_2\log^2 T\,.$$

Taking the expectation on both sides and the tower rule of expectations finishes the bound for the adversarial regime. For the stochastic regime, note that $\xi_{t,x} \le 1$ and hence

$$\begin{aligned}
&\sum_x p_{t,x}\xi_{t,x} - \mathbb{I}\{\boldsymbol{u} = \widehat{\boldsymbol{x}}\}p_{t,u}^2\xi_{t,u} \\
&\le 1 - \mathbb{I}\{\boldsymbol{u} = \widehat{\boldsymbol{x}}\}p_{t,u}^2 \\
&= (1 - \mathbb{I}\{\boldsymbol{u} = \widehat{\boldsymbol{x}}\}p_{t,u})(1 + \mathbb{I}\{\boldsymbol{u} = \widehat{\boldsymbol{x}}\}p_{t,u}) \le 2(1 - \mathbb{I}\{\boldsymbol{u} = \widehat{\boldsymbol{x}}\}p_{t,u})\,.
\end{aligned}$$

dd-LSB implies regular LSB (up to a factor of 2) and hence the stochastic bound of regular LSB applies. ■

Appendix F. Analysis for the Second Reduction

F.1. $\frac{1}{2}$-LSB to $\frac{1}{2}$-iw-stable (Algorithm 2 / Theorem 11)

Proof [Proof of Theorem 11] The per-step bonus $b_t = B_t - B_{t-1}$ is the sum of two terms:

$$b_t^{\mathsf{Ts}} = \sqrt{c_1\sum_{\tau=1}^{t}\frac{1}{q_{\tau,2}}} - \sqrt{c_1\sum_{\tau=1}^{t-1}\frac{1}{q_{\tau,2}}} \le \frac{\frac{c_1}{q_{t,2}}}{\sqrt{c_1\sum_{\tau=1}^{t}\frac{1}{q_{\tau,2}}}} \le \sqrt{\frac{c_1}{q_{t,2}}},$$

$$b_t^{\mathsf{Lo}} = \frac{c_2}{\min_{\tau\le t} q_{\tau,2}} - \frac{c_2}{\min_{\tau\le t-1} q_{\tau,2}} = \frac{c_2}{q_{t,2}}\left(1 - \frac{\min_{\tau\le t} q_{\tau,2}}{\min_{\tau\le t-1} q_{\tau,2}}\right).$$

Since $\frac{\bar{q}_{t,2}}{q_{t,2}} \leq 2$, using the inequalities above, we have

$$\eta_t\sqrt{\bar{q}_{t,2}}b_t^{\mathsf{Ts}} \leq \eta_t\sqrt{\frac{\bar{q}_{t,2}}{q_{t,2}}c_1} \leq \eta_t\sqrt{2c_1} \leq \frac{1}{4}. \tag{16}$$

$$\beta\bar{q}_{t,2}b_t^{\mathsf{Lo}} \leq \beta\frac{\bar{q}_{t,2}}{q_{t,2}}c_2 \leq 2\beta c_2 \leq \frac{1}{4}. \tag{17}$$

By Lemma 28 and that $\frac{\bar{q}_{t,2}}{q_{t,2}} \leq 2$, we have for any u,

$$\sum_{t=1}^{t'}\langle q_t - u, z_t\rangle \leq \underbrace{\sum_{t=1}^{t'}\langle q_t - \bar{q}_t, z_t\rangle}_{\textbf{term}_1}$$
$$+ O\Bigg(\sqrt{c_1} + \underbrace{\sum_{t=1}^{t'}\frac{\sqrt{q_{t,2}}}{\sqrt{t}}}_{\textbf{term}_2} + \underbrace{\frac{\log t'}{\beta}}_{\textbf{term}_3} + \underbrace{\sum_{t=1}^{t'}\eta_t\min_{|\theta_t|\leq 1} q_{t,i}^{\frac{3}{2}}(z_{t,i}-\theta_t)^2}_{\textbf{term}_4} + \underbrace{\sum_{t=1}^{t'} q_{t,2}b_t^{\mathsf{Ts}}}_{\textbf{term}_5} + \underbrace{\sum_{t=1}^{t'} q_{t,2}b_t^{\mathsf{Lo}}}_{\textbf{term}_6}\Bigg) - u_2\sum_{t=1}^{t'} b_t. \tag{18}$$

We bound individual terms below:

$$\mathbb{E}[\textbf{term}_1] = \mathbb{E}\left[\sum_{t=1}^{t'}\langle q_t - \bar{q}_t, z_t\rangle\right] \leq O\left(\sum_{t=1}^{t'}\frac{1}{t^2}\right) = O(1).$$

$$\textbf{term}_2 \leq O\left(\min\left\{\sqrt{t'},\ \sqrt{\sum_{t=1}^{t'} q_{t,2}\log T}\right\}\right).$$
$$\textbf{term}_3 = \frac{\log t'}{\beta} \leq O\left(c_2\log T\right).$$

$$\begin{aligned}
\mathbb{E}\left[\textbf{term}_4\right] &= \mathbb{E}\left[\sum_{t=1}^{t'}\eta_t\min_{\theta_t\in[-1,1]} q_{t,i}^{\frac{3}{2}}(z_{t,i}-\theta_t)^2\right]\\
&\leq \mathbb{E}\left[\sum_{t=1}^{t'}\eta_t\sum_{i=1}^{2} q_{t,i}^{\frac{3}{2}}\left(z_{t,i}-\ell_{t,A_t}\right)^2\right]\\
&= \mathbb{E}\left[\sum_{t=1}^{t'}\eta_t\sum_{i=1}^{2}\frac{1}{\sqrt{q_{t,i}}}\left(\mathbb{I}[i_t=i]\ell_{t,A_t} - q_{t,i}\ell_{t,A_t}\right)^2\right]\\
&\leq \mathbb{E}\left[\sum_{t=1}^{t'}\eta_t\sum_{i=1}^{2}\left(\sqrt{q_{t,i}}(1-q_{t,i})^2 + (1-q_{t,i})q_{t,i}^{\frac{3}{2}}\right)\right]\\
&\leq O\left(\mathbb{E}\left[\sum_{t=1}^{t'}\eta_t\sqrt{q_{t,2}}\right]\right) \leq O\left(\mathbb{E}[\textbf{term}_2]\right).
\end{aligned}$$

$$\begin{aligned}\textbf{term}_5 = \sum_{t=1}^{t'} q_{t,2} b_t^{\mathsf{Ts}} &\le \sum_{t=1}^{t'} q_{t,2} \left(\frac{\frac{c_1}{q_{t,2}}}{\sqrt{c_1 \sum_{\tau=1}^{t} \frac{1}{q_{\tau,2}}}} \right) \\ &= \sqrt{c_1} \sum_{t=1}^{t'} \frac{\frac{1}{\sqrt{q_{t,2}}}}{\sqrt{\sum_{\tau=1}^{t} \frac{1}{q_{\tau,2}}}} \times \sqrt{q_{t,2}} \qquad (19) \\ &\le \sqrt{c_1} \sqrt{\sum_{t=1}^{t'} \frac{\frac{1}{q_{t,2}}}{\sum_{\tau=1}^{t} \frac{1}{q_{\tau,2}}}} \sqrt{\sum_{t=1}^{t'} q_{t,2}} \qquad \text{(Cauchy-Schwarz)} \\ &\le \sqrt{c_1} \sqrt{1 + \log\left(\sum_{t=1}^{t'} \frac{1}{q_{t,2}} \right)} \sqrt{\sum_{t=1}^{t'} q_{t,2}} \\ &\le O\left(\sqrt{c_1 \sum_{t=1}^{t'} q_{t,2} \log T} \right).\end{aligned}$$

Continuing from (19), we also have $\textbf{term}_5 \le \sum_{t=1}^{t'} q_{t,2} b_t^{\mathsf{Ts}} \le \sqrt{c_1} \sum_{t=1}^{t'} \frac{1}{\sqrt{t}} \le 2\sqrt{c_1 t'}$ because $q_{\tau,2} \le 1$.

$$\textbf{term}_6 = \sum_{t=1}^{t'} q_{t,2} b_t^{\mathsf{Lo}} = c_2 \sum_{t=1}^{t'} \left(1 - \frac{\min_{\tau \le t} q_{\tau,2}}{\min_{\tau \le t-1} q_{\tau,2}} \right) \le c_2 \sum_{t=1}^{t'} \log \left(\frac{\min_{\tau \le t-1} q_{\tau,2}}{\min_{\tau \le t} q_{\tau,2}} \right) \le O\left(c_2 \log T \right).$$

Using all bounds above in (18), we can bound $\mathbb{E}\left[\sum_{t=1}^{t'} \langle q_t - u, z_t \rangle \right]$ by

$$\underbrace{O\left(\min\left\{ \sqrt{c_1 \mathbb{E}[t']},\ \sqrt{c_1 \left[\sum_{t=1}^{t'} q_{t,2} \right] \log T} \right\} + c_2 \log T \right)}_{\textbf{pos-term}} - u_2 \underbrace{\mathbb{E}\left[\sqrt{c_1 \sum_{t=1}^{t'} \frac{1}{q_{t,2}}} + \frac{c_2}{\min_{t \le t'} q_{t,2}} \right]}_{\textbf{neg-term}}.$$

For comparator $\widehat{x}$, we choose $u = \mathbf{e}_1$ and bound $\mathbb{E}\left[\sum_{t=1}^{t'} (\ell_{t,A_t} - \ell_{t,\widehat{x}}) \right]$ by the **pos-term** above. For comparator $x \ne \widehat{x}$, we first choose $u = \mathbf{e}_2$ and upper bound $\mathbb{E}\left[\sum_{t=1}^{t'} (\ell_{t,A_t} - \ell_{t,\widetilde{A}_t}) \right]$ by **pos-term**$-$**neg-term**. On the other hand, by the $\frac{1}{2}$-iw-stable assumption, $\mathbb{E}\left[\sum_{t=1}^{t'} (\ell_{t,\widetilde{A}_t} - \ell_{t,x}) \right] \le$ **neg-term**. Combining them, we get that for all $x \ne \widehat{x}$, we also have $\mathbb{E}\left[\sum_{t=1}^{t'} (\ell_{t,A_t} - \ell_{t,x}) \right] \le$ **pos-term**. Comparing the coefficients in **pos-term** with those in Definition 4, we see that Algorithm 2 satisfies $\frac{1}{2}$-LSB with constants (c_0', c_1', c_2') where $c_0' = c_1' = O(c_1)$ and $c_2' = O(c_2)$. ■

F.2. $\frac{2}{3}$-LSB to $\frac{2}{3}$-iw-stable (Algorithm 5 / Theorem 18)

Algorithm 5 LSB via Corral (for $\alpha = \frac{2}{3}$)

Input: candidate action $\widehat{x}$, $\frac{2}{3}$-iw-stable algorithm $\mathcal{B}$ over $\mathcal{X}\backslash\{\widehat{x}\}$ with constants c_1, c_2.

Define: $\psi_t(q) = -\frac{3}{\eta_t}\sum_{i=1}^{2} q_i^{\frac{2}{3}} + \frac{1}{\beta}\sum_{i=1}^{2}\ln\frac{1}{q_i}$.

for $t = 1, 2, \dots$ **do**

Let $\mathcal{B}$ generate an action $\widetilde{A}_t$.

Let

$$\bar{q}_t = \operatorname*{argmin}_{q\in\Delta_2}\left\{\left\langle q, \sum_{\tau=1}^{t-1} z_\tau - \begin{bmatrix} 0 \\ B_{t-1}\end{bmatrix}\right\rangle + \psi_t(q)\right\}, \quad q_t = (1-\gamma_t)\bar{q}_t,$$

$$\text{where} \quad \eta_t = \frac{1}{t^{\frac{2}{3}} + 8c_1^{\frac{1}{3}}},\ \beta = \frac{1}{8c_2}, \text{ and } \gamma_t = \max\left\{\sqrt{\eta_t} q_{t,2}^{\frac{2}{3}}, \eta_t q_{t,2}^{\frac{1}{3}}\right\}.$$

Sample $i_t \sim \bar{q}_t$.

if $i_t = 1$ **then** set $\bar{A}_t = \widehat{x}$;

else set $\bar{A}_t = \widetilde{A}_t$;

Sample $j_t \sim \gamma_t$.

if $j_t = 1$ **then** draw a revealing action of $\bar{A}_t$ and observe $\ell_{t,\bar{A}_t}$;

else draw $A_t = \bar{A}_t$;

Define $z_{t,i} = \frac{\ell_{t,\bar{A}_t}\mathbb{I}\{i_t=i\}\mathbb{I}\{j_t=1\}}{\gamma_t}$ and

$$B_t = c_1^{\frac{1}{3}}\left(\sum_{\tau=1}^{t}\frac{1}{\sqrt{q_{\tau,2}}}\right)^{\frac{2}{3}} + \frac{c_2}{\min_{\tau\le t} q_{\tau,2}}.$$

end

Proof [Proof of Theorem 18] The per-step bonus $b_t = B_t - B_{t-1}$ is the sum of two terms:

$$b_t^{\mathsf{Ts}} = c_1^{\frac{1}{3}}\left(\left(\sum_{\tau=1}^{t}\frac{1}{\sqrt{q_{\tau,2}}}\right)^{\frac{2}{3}} - \left(\sum_{\tau=1}^{t-1}\frac{1}{\sqrt{q_{\tau,2}}}\right)^{\frac{2}{3}}\right) \le c_1^{\frac{1}{3}}\frac{\frac{1}{\sqrt{q_{t,2}}}}{\left(\sum_{\tau=1}^{t}\frac{1}{\sqrt{q_{\tau,2}}}\right)^{\frac{1}{3}}} \le \left(\frac{c_1}{q_{t,2}}\right)^{\frac{1}{3}},$$

$$b_t^{\mathsf{Lo}} = \frac{c_2}{\min_{\tau\le t} q_{\tau,2}} - \frac{c_2}{\min_{\tau\le t-1} q_{\tau,2}} = \frac{c_2}{q_{t,2}}\left(1 - \frac{\min_{\tau\le t} q_{\tau,2}}{\min_{\tau\le t-1} q_{\tau,2}}\right).$$

Since $\frac{\bar{q}_{t,2}}{q_{t,2}} \le 2$, using the inequalities above, we have

$$\eta_t\bar{q}_{t,2}^{\frac{1}{3}} b_t^{\mathsf{Ts}} \le \eta_t\left(\frac{\bar{q}_{t,2}c_1}{q_{t,2}}\right)^{\frac{1}{3}} \le \eta_t(2c_1)^{\frac{1}{3}} \le \frac{1}{4}. \tag{20}$$

$$\beta\bar{q}_{t,2} b_t^{\mathsf{Lo}} \le \beta\frac{\bar{q}_{t,2}}{q_{t,2}}c_2 \le 2\beta c_2 \le \frac{1}{4}. \tag{21}$$

By Lemma 28 and that $\frac{\bar{q}_{t,2}}{q_{t,2}} \leq 2$, we have for any u,

$$\sum_{t=1}^{t'} \langle q_t - u, z_t \rangle \leq \underbrace{\sum_{t=1}^{t'} \langle q_t - \bar{q}_t, z_t \rangle}_{\textbf{term}_1}$$

$$+ O\Bigg(c_1^{\frac{1}{3}} + \underbrace{\sum_{t=1}^{t'} \frac{q_{t,2}^{\frac{2}{3}}}{t^{\frac{1}{3}}}}_{\textbf{term}_2} + \underbrace{\frac{\log t'}{\beta}}_{\textbf{term}_3} + \underbrace{\sum_{t=1}^{t'} \eta_t \min_{|\theta_t| \leq 1} q_{t,i}^{\frac{4}{3}} (z_{t,i} - \theta_t)^2}_{\textbf{term}_4} + \underbrace{\sum_{t=1}^{t'} q_{t,2} b_t^{\mathsf{Ts}}}_{\textbf{term}_5} + \underbrace{\sum_{t=1}^{t'} q_{t,2} b_t^{\mathsf{Lo}}}_{\textbf{term}_6} \Bigg) - u_2 \sum_{t=1}^{t'} b_t \tag{22}$$

We bound individual terms below:

$$\mathbb{E}[\textbf{term}_1] = \mathbb{E}\left[\sum_{t=1}^{t'} \langle q_t - \bar{q}_t, z_t \rangle\right] \leq O\left(\sum_{t=1}^{t'} \gamma_t\right) = O\left(\sum_{t=1}^{t'} \frac{q_{t,2}^{\frac{2}{3}}}{t^{\frac{1}{3}}} + \frac{q_{t,2}^{\frac{1}{3}}}{t^{\frac{2}{3}}}\right)$$
$$= O\left(\min\left\{ {t'}^{\frac{2}{3}}, \left(\sum_{t=1}^{t'} q_t\right)^{\frac{2}{3}} (\log T)^{\frac{1}{3}} + \log T \right\}\right).$$

$$\textbf{term}_2 \leq O\left(\textbf{term}_1\right)$$
$$\textbf{term}_3 = \frac{\log t'}{\beta} \leq O\left(c_2 \log T\right).$$

$$\mathbb{E}\left[\textbf{term}_4\right] = \mathbb{E}\left[\sum_{t=1}^{t'} \eta_t q_{t,2}^{\frac{4}{3}} (z_{t,2} - z_{t,1})^2\right]$$
$$\leq \mathbb{E}\left[\sum_{t=1}^{t'} \eta_t \frac{q_{t,i}^{\frac{4}{3}}}{\gamma_t}\right] \leq \mathbb{E}\left[\sum_{t=1}^{t'} \frac{\gamma_t^2}{\gamma_t}\right] = \mathbb{E}\left[O(\textbf{term}_1)\right].$$

$$\textbf{term}_5 = \sum_{t=1}^{t'} q_{t,2} b_t^{\mathsf{Ts}} \leq \sum_{t=1}^{t'} q_{t,2} \left(c_1^{\frac{1}{3}} \frac{\frac{1}{\sqrt{q_{t,2}}}}{\left(\sum_{\tau=1}^{t} \frac{1}{\sqrt{q_{\tau,2}}}\right)^{\frac{1}{3}}} \right)$$
$$= c_1^{\frac{1}{3}} \sum_{t=1}^{t'} \frac{q_{t,2}^{-\frac{1}{6}}}{\left(\sum_{\tau=1}^{t} \frac{1}{\sqrt{q_{\tau,2}}}\right)^{\frac{1}{3}}} \times q_{t,2}^{\frac{2}{3}} \tag{23}$$
$$\leq c_1^{\frac{1}{3}} \left(\sum_{t=1}^{t'} \frac{\frac{1}{\sqrt{q_{t,2}}}}{\sum_{\tau=1}^{t} \frac{1}{\sqrt{q_{\tau,2}}}}\right)^{\frac{1}{3}} \left(\sum_{t=1}^{t'} q_{t,2}\right)^{\frac{2}{3}} \qquad \text{(Cauchy-Schwarz)}$$

$$\leq c_1^{\frac{1}{3}} \left(1 + \log\left(\sum_{t=1}^{t'} \frac{1}{q_{t,2}}\right)\right)^{\frac{1}{3}} \left(\sum_{t=1}^{t'} q_{t,2}\right)^{\frac{2}{3}}$$
$$\leq O\left(c_1^{\frac{1}{3}} \left(\sum_{t=1}^{t'} q_{t,2}\right)^{\frac{2}{3}} (\log T)^{\frac{1}{3}}\right).$$

Continuing from (23), we also have $\textbf{term}_5 \leq \sum_{t=1}^{t'} q_{t,2} b_t^{\mathsf{Ts}} \leq c_1^{\frac{1}{3}} \sum_{t=1}^{t'} \frac{1}{t^{\frac{1}{3}}} \leq O(c_1^{\frac{1}{3}} t'^{\frac{2}{3}})$ because $q_{\tau,2} \leq 1$.

$$\textbf{term}_6 = \sum_{t=1}^{t'} q_{t,2} b_t^{\mathsf{Lo}} = c_2 \sum_{t=1}^{t'} \left(1 - \frac{\min_{\tau\leq t} q_{\tau,2}}{\min_{\tau\leq t-1} q_{\tau,2}}\right) \leq c_2 \sum_{t=1}^{t'} \log\left(\frac{\min_{\tau\leq t-1} q_{\tau,2}}{\min_{\tau\leq t} q_{\tau,2}}\right) \leq O\left(c_2 \log T\right).$$

Using all bounds above in (22), we can bound $\mathbb{E}\left[\sum_{t=1}^{t'} \langle q_t - u, z_t\rangle\right]$ by

$$\underbrace{O\left(\min\left\{c_1^{\frac{1}{3}} \mathbb{E}[t']^{\frac{2}{3}},\ (c_1 \log T)^{\frac{1}{3}} \left[\sum_{t=1}^{t'} q_{t,2}\right]^{\frac{2}{3}}\right\} + c_2 \log T\right)}_{\textbf{pos-term}} \underbrace{- u_2\, \mathbb{E}\left[c_1^{\frac{1}{3}} \left(\sum_{t=1}^{t'} \frac{1}{q_{t,2}}\right)^{\frac{2}{3}} + \frac{c_2}{\min_{t\leq t'} q_{t,2}}\right]}_{\textbf{neg-term}}.$$

For comparator $\widehat{x}$, we choose $u = \mathbf{e}_1$ and bound $\mathbb{E}\left[\sum_{t=1}^{t'} (\ell_{t,A_t} - \ell_{t,\widehat{x}})\right]$ by the **pos-term** above. For comparator $x \neq \widehat{x}$, we first choose $u = \mathbf{e}_2$ and upper bound $\mathbb{E}\left[\sum_{t=1}^{t'} (\ell_{t,A_t} - \ell_{t,\widetilde{A}_t})\right]$ by **pos-term**$-$**neg-term**. On the other hand, by the $\frac{1}{2}$-iw-stable assumption, $\mathbb{E}\left[\sum_{t=1}^{t'} (\ell_{t,\widetilde{A}_t} - \ell_{t,x})\right] \leq$ **neg-term**. Combining them, we get that for all $x \neq \widehat{x}$, we also have $\mathbb{E}\left[\sum_{t=1}^{t'} (\ell_{t,A_t} - \ell_{t,x})\right] \leq$ **pos-term**.

Finally, notice that $\mathbb{E}_t[\mathbb{I}\{A_t \neq \widehat{x}\}] \geq \bar{q}_{t,2}(1 - \gamma_t) = q_{t,2},$. This implies that Algorithm 5 is $\frac{2}{3}$-LSB with coefficient (c_0', c_1', c_2') with $c_0' = c_1' = O(c_1)$, and $c_2' = O(c_1^{\frac{1}{3}} + c_2)$.

■

F.3. $\frac{1}{2}$-dd-LSB to $\frac{1}{2}$-dd-iw-stable (Algorithm 6 / Theorem 22)

Proof [Proof of Theorem 22] Define $b_t = B_t - B_{t-1}$. Notice that we have

$$\eta_t \bar{q}_{t,2} b_t \leq 2\eta_t q_{t,2} b_t$$
$$\leq 2\eta_t q_{t,2} \left(\frac{\frac{c_1 \xi_{t,A_t} \mathbb{I}[i_t=2]}{q_{t,2}^2}}{\sqrt{c_1 \sum_{\tau=1}^{t-1} \frac{\xi_{\tau,A_\tau} \mathbb{I}[i_\tau=2]}{q_{\tau,2}^2}}}\right) + 2c_2 \left(\frac{1}{\min_{\tau\leq t} q_{\tau,2}} - \frac{1}{\min_{\tau\leq t-1} q_{\tau,2}}\right)$$
$$\leq 2\eta_t \sqrt{c_1} + 2\eta_t c_2 \left(1 - \frac{\min_{\tau\leq t} q_{\tau,2}}{\min_{\tau\leq t-1} q_{\tau,2}}\right)$$
$$\leq \frac{1}{4}. \tag{24}$$

Algorithm 6 dd-LSB via Corral (for $\alpha = \frac{1}{2}$)

Input: candidate action $\widehat{x}$, $\frac{1}{2}$-iw-stable algorithm $\mathcal{B}$ over $\mathcal{X}\backslash\{\widehat{x}\}$ with constant c.
Define: $\psi(q) = \sum_{i=1}^{2} \ln \frac{1}{q_i}$. $B_0 = 0$.
Define: For first-order bound, $\xi_{t,x} = \ell_{t,x}$ and $m_{t,x} = 0$; for second-order bound, $\xi_{t,x} = (\ell_{t,x} - m_{t,x})^2$ where $m_{t,x}$ is the loss predictor.
for $t = 1, 2, \dots$ **do**

Let $\mathcal{B}$ generate an action $\widetilde{A}_t$ (which is the action to be chosen if $\mathcal{B}$ is selected in this round).
Receive prediction $m_{t,x}$ for all $x \in \mathcal{X}$, and set $y_{t,1} = m_{t,\widehat{x}}$ and $y_{t,2} = m_{t,\widetilde{A}_t}$.
Let

$$\bar{q}_t = \operatorname*{argmin}_{q \in \Delta_2} \left\{ \left\langle q, \sum_{\tau=1}^{t-1} z_\tau + y_t - \begin{bmatrix} 0 \\ B_{t-1} \end{bmatrix} \right\rangle + \frac{1}{\eta_t}\psi(q) \right\}, \quad q_t = \left(1 - \frac{1}{2t^2}\right)\bar{q}_t + \frac{1}{4t^2}\mathbf{1},$$

$$\text{where} \quad \eta_t = \frac{1}{4}(\log T)^{\frac{1}{2}} \left(\sum_{\tau=1}^{t-1} (\mathbb{I}[i_\tau = i] - q_{\tau,i})^2 \xi_{\tau,A_\tau} + (c_1 + c_2^2)\log T \right)^{-\frac{1}{2}}.$$

Sample $i_t \sim q_t$.
if $i_t = 1$ **then** draw $A_t = \widehat{x}$ and observe ℓ_{t,A_t} ;
else draw $A_t = \widetilde{A}_t$ and observe ℓ_{t,A_t} ;
Define $z_{t,i} = \frac{(\ell_{t,A_t} - y_{t,i})\mathbb{I}\{i_t = i\}}{q_{t,i}} + y_{t,i}$ and

$$B_t = \sqrt{c_1 \sum_{\tau=1}^{t} \frac{\xi_{t,A_t}\mathbb{I}[i_\tau = 2]}{q_{\tau,2}^2}} + \frac{c_2}{\min_{\tau \le t} q_{\tau,2}}.$$

end

By Lemma 29 and that $\frac{\bar{q}_{t,2}}{q_{t,2}} \le 2$, we have for any u,

$$\sum_{t=1}^{t'} \langle q_t - u, z_t \rangle \le O\Bigg(\underbrace{\frac{\log T}{\eta_{t'}}}_{\textbf{term}_1} + \underbrace{\sum_{t=1}^{t'} \eta_t \min_{|\theta| \le 1} \sum_{i=1}^{2} q_{t,i}^2 (z_{t,i} - y_{t,i} - \theta_t)^2}_{\textbf{term}_2}$$
$$+ \underbrace{\sum_{t=1}^{t'} \langle q_t - \bar{q}_t, z_t \rangle}_{\textbf{term}_3} + \underbrace{\sum_{t=1}^{t'} q_{t,2} b_t}_{\textbf{term}_4} \Bigg) - \sum_{t=1}^{t'} u_2 b_t$$

$$\textbf{term}_1 \leq O\left(\sqrt{\frac{\sum_{t=1}^{t'-1}\sum_{i=1}^{2}(\mathbb{I}[i_t=i]-q_{t,i})^2\xi_{t,A_t}+(c_1+c_2^2)\log T}{\log T}}\log T\right)$$
$$\leq O\left(\sqrt{\sum_{t=1}^{t'}\sum_{i=1}^{2}(\mathbb{I}[i_t=i]-q_{t,i})^2\xi_{t,A_t}\log T}+(\sqrt{c_1}+c_2)\log T\right).$$

$$\mathbb{E}\left[\textbf{term}_2\right] \leq \mathbb{E}\left[\sum_{t=1}^{t'}\eta_t\min_\theta\sum_{i=1}^{2}q_{t,i}^2\left(\frac{(\ell_{t,A_t}-m_{t,A_t})\mathbb{I}[i_t=i]}{q_{t,i}}-\theta\right)^2\right]$$
$$\leq \mathbb{E}\left[\sum_{t=1}^{t'}\eta_t\sum_{i=1}^{2}(\mathbb{I}[i_t=i]-q_{t,i})^2(\ell_{t,A_t}-m_{t,A_t})^2\right] \qquad \text{(choosing } \theta=\ell_{t,A_t}-m_{t,A_t}\text{)}$$
$$\leq \mathbb{E}\left[\sqrt{\sum_{t=1}^{t'}\sum_{i=1}^{2}(\mathbb{I}[i_t=i]-q_{t,i})^2\xi_{t,A_t}\log T}\right].$$

$$\mathbb{E}\left[\textbf{term}_3\right] = \mathbb{E}\left[\sum_{t=1}^{t'}\langle q_t-\bar{q}_t,\mathbb{E}_t[z_t]\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{t'}\left\langle -\frac{1}{2t^2}\bar{q}_t+\frac{1}{4t^2}\mathbf{1},\mathbb{E}_t[z_t]\right\rangle\right] \leq O(1).$$

$$\textbf{term}_4 \leq \sum_{t=1}^{t'}q_{t,2}b_t$$
$$= \sum_{t=1}^{t'}q_{t,2}\left(\frac{\frac{c_1\xi_{t,A_t}\mathbb{I}[i_t=2]}{q_{t,2}^2}}{\sqrt{c_1\sum_{\tau=1}^{t}\frac{\xi_{\tau,A_\tau}\mathbb{I}[i_\tau=2]}{q_{\tau,2}^2}}}+c_2\left(\frac{1}{\min_{\tau\leq t}q_{\tau,2}}-\frac{1}{\min_{\tau\leq t-1}q_{\tau,2}}\right)\right)$$
$$\leq \sqrt{c_1}\sum_{t=1}^{t'}\sqrt{\frac{\frac{\xi_{t,A_t}\mathbb{I}[i_t=2]}{q_{t,2}^2}}{\sum_{\tau=1}^{t}\frac{\xi_{\tau,A_\tau}\mathbf{1}[i_\tau=2]}{q_{\tau,2}^2}}}\times\sqrt{\xi_{t,A_t}\mathbb{I}[i_t=2]}+c_2\sum_{t=1}^{t'}\left(1-\frac{\min_{\tau\leq t}q_{\tau,2}}{\min_{\tau\leq t-1}q_{\tau,2}}\right)$$
$$\leq \sqrt{c_1}\sqrt{\sum_{t=1}^{t'}\frac{\frac{\xi_{t,A_t}\mathbb{I}[i_t=2]}{q_{t,2}^2}}{\sum_{\tau=1}^{t}\frac{\xi_{\tau,A_\tau}\mathbf{1}[i_\tau=2]}{q_{\tau,2}^2}}}\times\sqrt{\sum_{t=1}^{t'}\xi_{t,A_t}\mathbb{I}[i_t=2]}+c_2\sum_{t=1}^{t'}\log\left(\frac{\min_{\tau\leq t-1}q_{\tau,2}}{\min_{\tau\leq t}q_{\tau,2}}\right)$$
(Cauchy-Schwarz)
$$\leq \sqrt{c_1}\sqrt{1+\log\left(\sum_{t=1}^{t'}\frac{\xi_{t,A_t}\mathbb{I}[i_t=2]}{q_{t,2}^2}\right)}\sqrt{\sum_{t=1}^{t'}\xi_{t,A_t}\mathbb{I}[i_t=2]}+c_2\log\frac{1}{\min_{\tau\leq t'}q_{\tau,2}}$$
$$\leq O\left(\sqrt{c_1\sum_{t=1}^{t'}\xi_{t,A_t}\mathbb{I}[i_t=2]\log T}+c_2\log T\right).$$

$$\textbf{term}_5 = -u_2 B_{t'} = -u_2 \left(\sqrt{c_1 \sum_{t=1}^{t'} \frac{\xi_{t,A_t}\mathbb{I}[i_t = 2]}{q_{t,2}^2}} + \frac{c_2}{\min_{t\le t'} q_{t,2}} \right).$$

Combining all inequalities above, we can bound $\mathbb{E}\left[\sum_{t=1}^{t'}\langle q_t - u, z_t\rangle\right]$ by

$$\underbrace{O\left(\mathbb{E}\left[\sqrt{c_1 \sum_{t=1}^{t'}\sum_{i=1}^{2}(\mathbb{I}[i_t = i] - q_{t,i})^2 \xi_{t,A_t}\log T} + \sqrt{c_1\sum_{t=1}^{t'}\xi_{t,A_t}\mathbb{I}[i_t=2]\log T}\right] + (\sqrt{c_1}+c_2)\log T\right)}_{\textbf{pos-term}} \tag{25}$$

$$- u_2 \underbrace{\left(\sqrt{c_1 \sum_{t=1}^{t'} \frac{\xi_{t,A_t}\mathbb{I}[i_t = 2]}{q_{t,2}^2}} + \frac{c_2}{\min_{t\le t'} q_{t,2}} \right)}_{\textbf{neg-term}}.$$

Similar to the arguments in the proofs of Theorem 11 and Theorem 18, we end up bounding $\mathbb{E}\left[\sum_{t=1}^{t'}(\ell_{t,A_t} - \ell_{t,x})\right]$ by the **pos-term** above for all $x \in \mathcal{X}$. Finally, we process **pos-term**. Observe that

$$\begin{aligned}
&\mathbb{E}_t\left[\sum_{i=1}^{2}(\mathbb{I}[i_t = i] - q_{t,i})^2\xi_{t,A_t}\right] \\
&= \mathbb{E}_t\left[q_{t,1}(1-q_{t,1})^2\xi_{t,\widehat{x}} + (1-q_{t,1})q_{t,1}^2\xi_{t,\widetilde{A}_t} + q_{t,2}(1-q_{t,2})^2\xi_{t,\widetilde{A}_t} + (1-q_{t,2})q_{t,2}^2\xi_{t,\widehat{x}}\right] \\
&= \mathbb{E}_t\left[2q_{t,1}q_{t,2}^2\xi_{t,\widehat{x}} + 2q_{t,1}^2 q_{t,2}\xi_{t,\widetilde{A}_t}\right] \\
&= 2q_{t,1}q_{t,2}^2\xi_{t,\widehat{x}} + 2q_{t,1}^2\left(\sum_{x\neq\widehat{x}} p_{t,x}\xi_{t,x}\right) \\
&\le 2p_{t,\widehat{x}}(1-p_{t,\widehat{x}})\xi_{t,\widehat{x}} + 2\left(\sum_{x\neq\widehat{x}} p_{t,x}\xi_{t,x}\right) \\
&= 2\left(\sum_{x} p_{t,x}\xi_{t,x} - p_{t,\widehat{x}}^2\xi_{t,\widehat{x}}\right),
\end{aligned}$$

and that

$$\mathbb{E}_t\left[\xi_{t,\widetilde{A}_t}\mathbb{I}[i_t = 2]\right] = \sum_{x\neq\widehat{x}} p_{t,\widehat{x}}\xi_{t,x} \le \sum_{x} p_{t,x}\xi_{t,x} - p_{t,\widehat{x}}^2\xi_{t,\widehat{x}}$$

Thus, for any $u \in \mathcal{X}$,

$$\mathbb{E}\left[\sum_{t=1}^{t'}(\ell_{t,A_t} - \ell_{t,u})\right]$$

$$\leq \mathbb{E}\left[\textbf{pos-term}\right] \leq O\left(\sqrt{c_1\mathbb{E}\left[\sum_{t=1}^{t'}\left(\sum_x p_{t,x}\xi_{t,x} - p_{t,\widehat{x}}^2\xi_{t,\widehat{x}}\right)\right]\log T} + (\sqrt{c_1}+c_2)\log T\right),$$

which implies that the algorithm satisfies $\frac{1}{2}$-dd-LSB with constants $(O(c_1), O(\sqrt{c_1}+c_2))$. ■

F.4. $\frac{1}{2}$-LSB to $\frac{1}{2}$-strongly-iw-stable (Algorithm 7 / Theorem 25)

Algorithm 7 LSB via Corral (for $\alpha = \frac{1}{2}$, using a $\frac{1}{2}$-strongly-iw-stable algorithm)

Input: candidate action $\widehat{x}$, $\frac{1}{2}$-iw-stable algorithm $\mathcal{B}$ over $\mathcal{X}$ with constant c.
Define: $\psi_t(q) = \frac{-2}{\eta_t}\sum_{i=1}^2 \sqrt{q_i} + \frac{1}{\beta}\sum_{i=1}^2 \ln\frac{1}{q_i}$.
$B_0 = 0$.
for $t = 1, 2, \dots$ **do**

Let $\mathcal{B}$ generate an action $\widetilde{A}_t$ (which is the action to be chosen if $\mathcal{B}$ is selected in this round).
Let

$$\bar{q}_t = \operatorname*{argmin}_{q\in\Delta_2}\left\{\left\langle q, \sum_{\tau=1}^{t-1} z_\tau - \begin{bmatrix}0\\ B_{t-1}\end{bmatrix}\right\rangle + \psi_t(q)\right\}, \quad q_t = \left(1-\frac{1}{2t^2}\right)\bar{q}_t + \frac{1}{4t^2}\mathbf{1},$$

$$\text{where} \quad \eta_t = \frac{1}{\sqrt{\sum_{\tau=1}^t \mathbb{I}\{\widetilde{A}_\tau \neq \widehat{x}\} + 8\sqrt{c_1}}}, \quad \beta = \frac{1}{8c_2}.$$

Sample $i_t \sim q_t$.
if $i_t = 1$ **then** draw $A_t = \widehat{x}$ and observe ℓ_{t,A_t} ;
else draw $A_t = \widetilde{A}_t$ and observe ℓ_{t,A_t} ;
Define $z_{t,i} = \frac{\ell_{t,A_t}\mathbb{I}\{i_t=i\}}{q_{t,i}}\mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}$ and

$$B_t = \sqrt{c_1\sum_{\tau=1}^t \frac{\mathbb{I}\left\{\widetilde{A}_t \neq \widehat{x}\right\}}{q_{\tau,2}}} + c_2\max_{\tau\leq t}\frac{1}{q_{\tau,2}}.$$

end

Proof [Proof of Theorem 25] The proof of this theorem mostly follows that of Theorem 11. The difference is that the regret of the base algorithm $\mathcal{B}$ is now bounded by

$$\sqrt{c_1\sum_{t=1}^{t'}\frac{1}{q_{t,2}+q_{t,1}\mathbb{I}\{\widetilde{A}_t = \widehat{x}\}}} + \frac{c_2\log T}{\min_{t\leq t'} q_{t,2}} \leq \sqrt{c_1\sum_{t=1}^{t'}\frac{\mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}}{q_{t,2}}} + \sqrt{c_1 t'} + \frac{c_2\log T}{\min_{t\leq t'} q_{t,2}} \tag{26}$$

because when $\widetilde{A}_t = \widehat{x}$, the base algorithm $\mathcal{B}$ is able to receive feedback no matter which side the Corral algorithm chooses. The goal of adding bonus is now only to cancel the first term and the third term on the right-hand side of (26).

Similar to Eq. (18), we have

$$\sum_{t=1}^{t'}\langle q_t - u, z_t\rangle \le \underbrace{\sum_{t=1}^{t'}\langle q_t - \bar{q}_t, z_t\rangle}_{\textbf{term}_1}$$
$$+ O\Bigg(\sqrt{c_1} + \underbrace{\sum_{t=1}^{t'}\frac{\sqrt{q_{t,2}}\mathbb{I}\{\widetilde{A}_t \ne \widehat{x}\}}{\sqrt{\sum_{\tau=1}^{t}\mathbb{I}\{\widetilde{A}_\tau \ne \widehat{x}\}}}}_{\textbf{term}_2} + \underbrace{\frac{\log T}{\beta}}_{\textbf{term}_3} + \underbrace{\sum_{t=1}^{t'}\eta_t \min_{|\theta_t|\le 1} q_{t,i}^{\frac{3}{2}}(z_{t,i}-\theta_t)^2}_{\textbf{term}_4} + \underbrace{\sum_{t=1}^{t'} q_{t,2}b_t^{\textsf{Ts}}}_{\textbf{term}_5} + \underbrace{\sum_{t=1}^{t'} q_{t,2}b_t^{\textsf{Lo}}}_{\textbf{term}_6}\Bigg)$$
$$- u_2\sum_{t=1}^{t'} b_t \tag{27}$$

where

$$b_t^{\textsf{Ts}} = \sqrt{c_1\sum_{\tau=1}^{t}\frac{\mathbb{I}\{\widetilde{A}_\tau \ne \widehat{x}\}}{q_{\tau,2}}} - \sqrt{c_1\sum_{\tau=1}^{t-1}\frac{\mathbb{I}\{\widetilde{A}_\tau \ne \widehat{x}\}}{q_{\tau,2}}} \le \frac{\frac{c_1\mathbb{I}\{\widetilde{A}_t\ne\widehat{x}\}}{q_{t,2}}}{\sqrt{c_1\sum_{\tau=1}^{t}\frac{\mathbb{I}\{\widetilde{A}_\tau\ne\widehat{x}\}}{q_{\tau,2}}}} \le \sqrt{\frac{c_1}{q_{t,2}}},$$
$$b_t^{\textsf{Lo}} = \frac{c_2}{\min_{\tau\le t} q_{\tau,2}} - \frac{c_2}{\min_{\tau\le t-1} q_{\tau,2}} = \frac{c_2}{q_{t,2}}\left(1 - \frac{\min_{\tau\le t} q_{\tau,2}}{\min_{\tau\le t-1} q_{\tau,2}}\right)$$

satisfying $\eta_t\sqrt{\bar{q}_{t,2}}b_t^{\textsf{Ts}} \le \frac{1}{2}$ and $\beta\bar{q}_{t,2}b_t^{\textsf{Lo}} \le \frac{1}{2}$ as in (20) and (21).

Below, we bound $\textbf{term}_1, \ldots, \textbf{term}_6$.

$$\mathbb{E}[\textbf{term}_1] = \mathbb{E}\left[\sum_{t=1}^{t'}\langle q_t - \bar{q}_t, z_t\rangle\right] \le O\left(\sum_{t=1}^{t'}\frac{1}{t^2}\right) = O(1).$$

$$\textbf{term}_2 \le O\left(\min\left\{\sqrt{\sum_{t=1}^{t'}\mathbb{I}\{\widetilde{A}_t \ne \widehat{x}\}},\ \sqrt{\sum_{t=1}^{t'} q_{t,2}\mathbb{I}\{\widetilde{A}_t \ne \widehat{x}\}\log T}\right\}\right)$$
$$\textbf{term}_3 = \frac{\log T}{\beta} \le O\left(c_2\log T\right).$$

$$\begin{aligned}
\mathbb{E}\left[\mathbf{term}_4\right] &= \mathbb{E}\left[\sum_{t=1}^{t'} \eta_t \min_{\theta_t \in [-1,1]} q_{t,i}^{\frac{3}{2}}(z_{t,i} - \theta_t)^2\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{t'} \eta_t \sum_{i=1}^{2} q_{t,i}^{\frac{3}{2}} \left(z_{t,i} - \ell_{t,A_t}\right)^2 \mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}\right] \qquad \text{(when } \widetilde{A}_t = \widehat{x},\, z_{t,1} = z_{t,2} = 0\text{)} \\
&= \mathbb{E}\left[\sum_{t=1}^{t'} \eta_t \sum_{i=1}^{2} \frac{1}{\sqrt{q_{t,i}}} \left(\mathbb{I}[i_t = i]\ell_{t,A_t} - q_{t,i}\ell_{t,A_t}\right)^2 \mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{t'} \eta_t \sum_{i=1}^{2} \left(\sqrt{q_{t,i}}(1-q_{t,i})^2 + (1-q_{t,i})q_{t,i}^{\frac{3}{2}}\right) \mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{t'} \eta_t \sqrt{q_{t,2}\mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}}\right] \le O\left(\mathbb{E}[\mathbf{term}_2]\right).
\end{aligned}$$

$$\begin{aligned}
\mathbf{term}_5 + \mathbf{term}_6 &= \sum_{t=1}^{t'} q_{t,2}(b_t^{\mathsf{Ts}} + b_t^{\mathsf{Lo}}) \\
&= \sum_{t=1}^{t'} q_{t,2}\left(\frac{\frac{c_1\mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}}{q_{t,2}}}{\sqrt{c_1\sum_{\tau=1}^{t}\frac{\mathbb{I}\{\widetilde{A}_\tau \neq \widehat{x}\}}{q_{\tau,2}}}} + c_2\left(\frac{1}{\min_{\tau\le t} q_{\tau,2}} - \frac{1}{\min_{\tau \le t-1} q_{\tau,2}}\right)\right) \\
&= \sqrt{c_1}\sum_{t=1}^{t'} \frac{\frac{\mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}}{\sqrt{q_{t,2}}}}{\sqrt{\sum_{\tau=1}^{t}\frac{\mathbb{I}\{\widetilde{A}_\tau \neq \widehat{x}\}}{q_{\tau,2}}}} \times \sqrt{q_{t,2}\mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}} + c_2\sum_{t=1}^{t'}\left(1 - \frac{\min_{\tau\le t} q_{\tau,2}}{\min_{\tau\le t-1} q_{\tau,2}}\right) \qquad (28) \\
&\le \sqrt{c_1}\sqrt{\sum_{t=1}^{t'} \frac{\frac{\mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}}{q_{t,2}}}{\sum_{\tau=1}^{t}\frac{\mathbb{I}\{\widetilde{A}_\tau \neq \widehat{x}\}}{q_{\tau,2}}}}\sqrt{\sum_{t=1}^{t'} q_{t,2}\mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}} + c_2\sum_{t=1}^{t'}\log\left(\frac{\min_{\tau\le t-1} q_{\tau,2}}{\min_{\tau \le t} q_{\tau,2}}\right) \qquad \text{(Cauchy-Schwarz)} \\
&\le \sqrt{c_1}\sqrt{1 + \log\left(\sum_{t=1}^{t'}\frac{\mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}}{q_{t,2}}\right)}\sqrt{\sum_{t=1}^{t'} q_{t,2}\mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}} + c_2\log\frac{1}{\min_{\tau\le t'} q_{\tau,2}} \\
&\le O\left(\sqrt{c_1\sum_{t=1}^{t'} q_{t,2}\mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}\log T} + c_2\log T\right). \qquad (29)
\end{aligned}$$

Continuing from (28), we also have $\mathbf{term}_5 \le \sqrt{c_1}\sum_{t=1}^{t'}\frac{\mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}}{\sum_{\tau=1}^{t}\mathbb{I}\{\widetilde{A}_\tau \neq \widehat{x}\}} \le 2\sqrt{c_1 t'}$.

Using all the inequalities above in (27), we can bound $\mathbb{E}\left[\sum_{t=1}^{t'}\langle q_t - u, z_t\rangle\right]$ by

$$\underbrace{O\left(\min\left\{\sqrt{c_1\mathbb{E}[t']},\ \sqrt{c_1\left[\sum_{t=1}^{t'} q_{t,2}\mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}\right]\log T}\right\} + (\sqrt{c_1}+c_2)\log T\right)}_{\textbf{pos-term}}$$
$$- u_2\,\underbrace{\mathbb{E}\left[\sqrt{c_1\sum_{t=1}^{t'}\frac{\mathbb{I}\{\widetilde{A}_t\neq\widehat{x}\}}{q_{t,2}}} + \frac{c_2}{\min_{t\leq t'} q_{t,2}}\right]}_{\textbf{neg-term}}.$$

For comparator $\widehat{x}$, we set $u = \mathbf{e}_1$. Then we have $\mathbb{E}[\sum_{t=1}^{t'}(\ell_{t,A_t} - \ell_{t,\widehat{x}})] \leq$ **pos-term**. Observe that $\mathbb{E}\left[\sum_{t=1}^{t'} q_{t,2}\mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}\right] = \mathbb{E}\left[\sum_{t=1}^{t'}(1-p_{t,\widehat{x}})\right]$, so this gives the desired property of $\frac{1}{2}$-LSB for action $\widehat{x}$. For $x \neq \widehat{x}$, we set $u = \mathbf{e}_2$ and get $\mathbb{E}\left[\sum_{t=1}^{t'}(\ell_{t,A_t} - \ell_{t,x})\right] = \mathbb{E}\left[\sum_{t=1}^{t'}(\ell_{t,A_t} - \ell_{t,\widetilde{A}_t})\right] + \mathbb{E}\left[\sum_{t=1}^{t'}(\ell_{t,\widetilde{A}_t} - \ell_{t,x})\right] \leq$ (**pos-term** $-$ **neg-term**) $+$ (**neg-term** $+ \sqrt{c_1\mathbb{E}[t']}$) where the additional $\sqrt{c_1 t'}$ term comes from the second term on the right-hand side of (26), which is not cancelled by the negative term. Note, however, that this still satisfies the requirement of $\frac{1}{2}$-LSB for $x \neq \widehat{x}$ because for these actions, we only require the worst-case bound to hold. Overall, we have justified that Algorithm 7 satisfies $\frac{1}{2}$-LSB with constants (c_0', c_1', c_2') where $c_0' = c_1' = O(c_1)$ and $c_2 = (\sqrt{c_1} + c_2)$. ■

F.5. $\frac{1}{2}$-dd-LSB to $\frac{1}{2}$-dd-strongly-iw-stable (Algorithm 8 / Lemma 42)

Lemma 42 *Let $\mathcal{B}$ be an algorithm with the following stability guarantee: given an adaptive sequence of weights $q_1, q_2, \dots \in (0,1]^{\mathcal{X}}$ such that the feedback in round t is observed with probability $q_t(x)$ if x is chosen, and an adaptive sequence $\{m_{t,x}\}_{x\in\mathcal{X}}$ available at the beginning of round t, it obtains the following pseudo regret guarantee for any stopping time t':*

$$\mathbb{E}\left[\sum_{t=1}^{t'}(\ell_{t,A_t} - \ell_{t,u})\right] \leq \mathbb{E}\left[\sqrt{c_1\sum_{t=1}^{t'}\frac{\mathsf{upd}_t\cdot\xi_{t,A_t}}{q_t(A_t)^2}} + \frac{c_2}{\min_{t\leq t'}\min_x q_t(x)}\right],$$

where $\mathsf{upd}_t = 1$ if feedback is observed in round t and $\mathsf{upd}_t = 0$ otherwise. $\xi_{t,x} = (\ell_{t,x} - m_{t,x})^2$ in the second-order bound case, and $\xi_{t,x} = \ell_{t,x}$ in the first-order bound case. Then Algorithm 8 with $\mathcal{B}$ as input satisfies $\frac{1}{2}$-dd-LSB.

Proof The proof of this lemma is a combination of the elements in Theorem 22 (data-dependent iw-stable) and Theorem 25 (strongly-iw-stable), so we omit the details here and only provide a sketch.

Algorithm 8 dd-LSB via Corral (for $\alpha = \frac{1}{2}$, using a $\frac{1}{2}$-dd-strongly-iw-stable algorithm)

Input: candidate action $\widehat{x}$, $\frac{1}{2}$-iw-stable algorithm $\mathcal{B}$ over $\mathcal{X}$ with constant (c_1, c_2).
Define: $\psi(q) = \sum_{i=1}^{2} \ln \frac{1}{q_i}$. $B_0 = 0$.
Define: For first-order bound, $\xi_{t,x} = \ell_{t,x}$ and $m_{t,x} = 0$; for second-order bound, $\xi_{t,x} = (\ell_{t,x} - m_{t,x})^2$ where $m_{t,x}$ is the loss predictor.
for $t = 1, 2, \dots$ **do**

Receive prediction $m_{t,x}$ for all $x \in \mathcal{X}$, and set $y_{t,1} = m_{t,\widehat{x}}$ and $y_{t,2} = m_{t,\widetilde{A}_t}$.
Let $\mathcal{B}$ generate an action $\widetilde{A}_t$ (which is the action to be chosen if $\mathcal{B}$ is selected in this round).
Let

$$\bar{q}_t = \operatorname*{argmin}_{q \in \Delta_2} \left\{ \left\langle q, \sum_{\tau=1}^{t-1} z_\tau + y_t - \begin{bmatrix} 0 \\ B_{t-1} \end{bmatrix} \right\rangle + \frac{1}{\eta_t}\psi(q) \right\}, \quad q_t = \left(1 - \frac{1}{2t^2}\right)\bar{q}_t + \frac{1}{4t^2}\mathbf{1},$$

$$\text{where} \quad \eta_t = \frac{1}{4}(\log T)^{\frac{1}{2}} \left(\sum_{\tau=1}^{t-1} (\mathbb{I}[i_\tau = i] - q_{\tau,i})^2 \xi_{\tau,A_\tau} \mathbb{I}[\widetilde{A}_\tau \neq \widehat{x}] + (c_1 + c_2^2)\log T \right)^{-\frac{1}{2}}.$$

Sample $i_t \sim q_t$.
if $i_t = 1$ **then** draw $A_t = \widehat{x}$ and observe ℓ_{t,A_t} ;
else draw $A_t = \widetilde{A}_t$ and observe ℓ_{t,A_t} ;
Define $z_{t,i} = \left(\frac{(\ell_{t,A_t} - y_{t,i})\mathbb{I}\{i_t = i\}}{q_{t,i}} + y_{t,i} \right) \mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}$ and

$$B_t = \sqrt{c_1 \sum_{\tau=1}^{t} \frac{\xi_{t,A_t}\mathbb{I}\{i_\tau = 2\}\mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}}{q_{\tau,2}^2}} + \frac{c_2}{\min_{\tau \le t} q_{\tau,2}}. \tag{30}$$

end

In Algorithm 8, since $\mathcal{B}$ is $\frac{1}{2}$-dd-strongly-iw-stable, its regret is upper bounded by the order of

$$\sqrt{c_1 \sum_{t=1}^{t'} \left(\frac{\mathbb{I}\{\widetilde{A}_t = \widehat{x}\} \cdot \xi_{t,A_t}}{1} + \frac{\mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}\mathbb{I}\{i_t = 2\}\xi_{t,A_t}}{q_{t,2}^2} \right)} + \frac{c_2}{\min_{t \le t'} \min_x q_t(x)}$$
$$\le \sqrt{c_1 \sum_{t=1}^{t'} \frac{\mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\}\mathbb{I}\{i_t = 2\} \cdot \xi_{t,A_t}}{q_{t,2}^2}} + \sqrt{c_1 \sum_{t=1}^{t'} \xi_{t,A_t}} + \frac{c_2}{\min_{t \le t'} \min_x q_t(x)} \tag{31}$$

because if $\mathcal{B}$ chooses $\widehat{x}$, then the probability of observing the feedback is 1, and is $q_{t,2}$ otherwise. This motivates the choice of the bonus in (30). Then we can follow the proof of Theorem 22 step-by-step, and show that the regret compared to $\widehat{x}$ is upper bounded by the order of

$$\mathbb{E}\left[\sqrt{c_1 \sum_{t=1}^{t'} \sum_{i=1}^{2} (\mathbb{I}[i_t = i] - q_{t,i})^2 \xi_{t,A_t} \mathbb{I}\{\widetilde{A}_t \neq \widehat{x}\} \log T} \right]$$

$$+\mathbb{E}\left[\sqrt{c_1\sum_{t=1}^{t'}\xi_{t,A_t}\mathbb{I}[i_t=2]\mathbb{I}\{\widetilde{A}_t\neq\widehat{x}\}\log T}\right]+(\sqrt{c_1}+c_2)\log T$$

$$\leq\sqrt{c_1\mathbb{E}\left[\sum_{t=1}^{t'}\left(2q_{t,1}q_{t,2}^2\xi_{t,\widehat{x}}+2q_{t,1}^2q_{t,2}\xi_{t,\widetilde{A}_t}\right)\mathbb{I}\{\widetilde{A}_t\neq\widehat{x}\}\right]}$$

$$+\sqrt{c_1\mathbb{E}\left[\sum_{t=1}^{t'}\xi_{t,\widetilde{A}_t}q_{t,2}\mathbb{I}\{\widetilde{A}_t\neq\widehat{x}\}\right]\log T}+(\sqrt{c_1}+c_2)\log T$$

(taking expectation over i_t and following the calculation in the proof of Theorem 22)

$$=\sqrt{c_1\mathbb{E}\left[\sum_{t=1}^{t'}\left(2q_{t,1}q_{t,2}^2\xi_{t,\widehat{x}}\Pr[\widetilde{A}_t\neq\widehat{x}]+2q_{t,1}^2q_{t,2}\sum_{x\neq\widehat{x}}\Pr[\widetilde{A}_t=x]\xi_{t,x}\right)\right]}$$

$$+\sqrt{c_1\mathbb{E}\left[\sum_{t=1}^{t'}q_{t,2}\sum_{x\neq\widehat{x}}\Pr[\widetilde{A}_t=x]\xi_{t,x}\right]\log T}+(\sqrt{c_1}+c_2)\log T$$

(taking expectation over $\widetilde{A}_t$)

$$\leq\sqrt{c_1\mathbb{E}\left[\sum_{t=1}^{t'}\left(2q_{t,1}(1-p_{t,\widehat{x}})\xi_{t,\widehat{x}}+2q_{t,1}^2\sum_{x\neq\widehat{x}}p_{t,x}\xi_{t,x}\right)\right]}$$

$$+\sqrt{c_1\mathbb{E}\left[\sum_{t=1}^{t'}\sum_{x\neq\widehat{x}}p_{t,x}\xi_{t,x}\right]\log T}+(\sqrt{c_1}+c_2)\log T$$

(using the property that for $x\neq\widehat{x}$, $p_{t,x}=\Pr[\widetilde{A}_t=x]q_{t,2}$ and thus $1-p_{t,\widehat{x}}=\Pr[\widetilde{A}_t\neq\widehat{x}]q_{t,2}$)

$$\leq\sqrt{c_1\mathbb{E}\left[\sum_{t=1}^{t'}\left(2p_{t,\widehat{x}}(1-p_{t,\widehat{x}})\xi_{t,\widehat{x}}+2\sum_{x\neq\widehat{x}}p_{t,x}\xi_{t,x}\right)\right]}\qquad(q_{t,1}\leq p_{t,\widehat{x}}\leq 1)$$

$$+\sqrt{c_1\mathbb{E}\left[\sum_{t=1}^{t'}\sum_{x\neq\widehat{x}}p_{t,x}\xi_{t,x}\right]\log T}+(\sqrt{c_1}+c_2)\log T$$

$$\leq O\left(\sqrt{c_1\mathbb{E}\left[\sum_{t=1}^{t'}\left(\sum_{x}p_{t,x}\xi_{t,x}-p_{t,\widehat{x}}^2\xi_{t,\widehat{x}}\right)\right]\log T}\right)+(\sqrt{c_1}+c_2)\log T$$

which satisfies the requirement of $\frac{1}{2}$-dd-LSB for the regret against $\widehat{x}$. For the regret against $x\neq\widehat{x}$, similar to the proof of Theorem 25, an extra positive regret comes from the second term in (31). Therefore, the regret against $x\neq\widehat{x}$, can be upper bounded by

$$O\left(\sqrt{c_1\mathbb{E}\left[\sum_{t=1}^{t'}\sum_{x}p_{t,x}\xi_{t,x}\right]\log T}\right)+(\sqrt{c_1}+c_2)\log T$$

which also satisfies the requirement of $\frac{1}{2}$-dd-LSB for $x \neq \widehat{x}$. ∎

Appendix G. Analysis for IW-Stable Algorithms

Algorithm 9 EXP2

Input: $\mathcal{X}$.

for $t = 1, 2, \ldots$ **do**

Receive update probability q_t.

Let

$$\eta_t = \min\left\{\sqrt{\frac{\ln|\mathcal{X}|}{d\sum_{\tau=1}^{t}\frac{1}{q_\tau}}}, \frac{1}{2d}\min_{\tau\leq t} q_\tau\right\}, \qquad P_t(a) \propto \exp\left(-\eta_t \sum_{\tau=1}^{t-1}\widehat{\ell}_\tau(a)\right).$$

Sample an action $a_t \sim p_t = (1 - \frac{d\eta_t}{q_t})P_t + \frac{d\eta_t}{q_t}\nu$, where ν is John's exploration.

With probability q_t, receive $\ell_t(a_t) = \langle a_t, \theta_t\rangle + \text{noise}$

(in this case, set $\mathsf{upd}_t = 1$; otherwise, set $\mathsf{upd}_t = 0$).

Construct loss estimator:

$$\widehat{\ell}_t(a) = \frac{\mathsf{upd}_t}{q_t} a^\top \left(\mathbb{E}_{b\sim p_t}[bb^\top]\right)^{-1} a_t \ell_t(a_t)$$

end

G.1. EXP2 (Algorithm 9 / Lemma 9)

Proof [Proof of Lemma 9] Consider the EXP2 algorithm (Algorithm 9) which corresponds to FTRL with negentropy potential. By standard analysis of FTRL (Lemma 27), for any τ and any $a^\star$,

$$\sum_{t=1}^{\tau}\mathbb{E}_t\left[\sum_a P_t(a)\widehat{\ell}_t(a)\right] - \sum_{t=1}^{\tau}\mathbb{E}_t\left[\widehat{\ell}_t(a^\star)\right] \leq \frac{\ln|\mathcal{X}|}{\eta_\tau} + \sum_{t=1}^{\tau}\mathbb{E}_t\left[\max_P\left(\langle P_t - P, \ell_t\rangle - \frac{1}{\eta_t}D_\psi(P, P_t)\right)\right].$$

To apply Lemma 31, we need to show that $\eta_t\widehat{\ell}_t(a) \geq -1$. We have by Cauchy Schwarz $|a^\top\left(\mathbb{E}_{b\sim p_t}[bb^\top]\right)^{-1} a_t| \leq \sqrt{a^\top\left(\mathbb{E}_{b\sim p_t}[bb^\top]\right)^{-1} a}\sqrt{a_t^\top\left(\mathbb{E}_{b\sim p_t}[bb^\top]\right)^{-1} a_t}$. For each term, we have

$$a^\top\left(\mathbb{E}_{b\sim p_t}[bb^\top]\right)^{-1} a \leq \frac{q_t}{d\eta_t} a^\top\left(\mathbb{E}_{b\sim \nu}[bb^\top]\right)^{-1} a = \frac{q_t}{\eta_t},$$

due to the properties of John's exploration. Hence $|\eta_t\widehat{\ell}_t(a)| \leq |\ell_t(a)| \leq 1$. We can apply Lemma 31 for the stability term, resulting in

$$\begin{aligned}
&\mathbb{E}_t\left[\max_P\left(\langle P_t - P, \ell_t\rangle - \frac{1}{\eta_t}D_\psi(P, P_t)\right)\right] \\
&\leq \eta_t\mathbb{E}_t\left[\sum_a P_t(a)\frac{a^\top\left(\mathbb{E}_{b\sim p_t}[bb^\top]\right)^{-1}a_t a_t^\top\left(\mathbb{E}_{b\sim p_t}[bb^\top]\right)^{-1}a}{q_t^2}\right] \\
&= \eta_t\sum_a P_t(a)\frac{a^\top\left(\mathbb{E}_{b\sim p_t}[bb^\top]\right)^{-1}a}{q_t} \\
&= 2\eta_t\sum_a P_t(a)\frac{a^\top\left(\mathbb{E}_{b\sim P_t}[bb^\top]\right)^{-1}a}{q_t} = \frac{2\eta_t d}{q_t}\,.
\end{aligned}$$

By $|\ell_t(a)| \leq 1$ we have furthermore

$$\sum_{t=1}^{\tau}\mathbb{E}_t\left[\sum_a(P_t(a) - p_t(a))\widehat{\ell}_t(a)\right] = \sum_{t=1}^{\tau}\mathbb{E}_t\left[\sum_a(P_t(a) - p_t(a))\ell_t(a)\right] \leq \sum_{t=1}^{\tau}\frac{d\eta_t}{q_t}\,.$$

Combining everything and taking the expectation on both sides leads to

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{\tau}\sum_a p_t(a)\ell_t(a) - \ell_t(a^\star)\right] &\leq \mathbb{E}\left[\frac{\log|\mathcal{X}|}{\eta_\tau} + \sum_{t=1}^{\tau}\frac{3d\eta_t}{q_t}\right] \\
&\leq \mathbb{E}\left[7\sqrt{d\log|\mathcal{X}|\sum_{t=1}^{\tau}\frac{1}{q_t}} + 2\log|\mathcal{X}|d\frac{1}{\min_{t\leq\tau}q_t}\right]\,.
\end{aligned}$$

■

G.2. EXP4 (Algorithm 10 / Lemma 10)

In this section, we use the more standard notation Π to denote the policy class.

Algorithm 10 EXP4

Input: Π (policy class), K (number of arms)

for $t = 1, 2, \ldots$ **do**

Receive context x_t.

Receive update probability q_t.

Let

$$\eta_t = \sqrt{\frac{\ln|\Pi|}{K\sum_{\tau=1}^{t}\frac{1}{q_\tau}}}, \qquad P_t(\pi) \propto \exp\left(-\eta_t \sum_{\tau=1}^{t-1} \widehat{\ell}_\tau\left(\pi(x_\tau)\right)\right).$$

Sample an arm $a_t \sim p_t$ where $p_t(a) = \sum_{\pi:\pi(x_t)=a} P_t(\pi)\pi(x_t)$.

With probability q_t, receive $\ell_t(a_t)$ (in this case, set $\mathsf{upd}_t = 1$; otherwise, set $\mathsf{upd}_t = 0$).

Construct loss estimator:

$$\widehat{\ell}_t(a) = \frac{\mathsf{upd}_t \mathbb{I}[a_t = a]\ell_t(a)}{q_t p_t(a)}.$$

end

Proof [Proof of Lemma 10] Consider the EXP4 algorithm with adaptive stepsize (Algorithm 10), which corresponds to FTRL with negentropy regularization. By standard analysis of FTRL (Lemma 27) and Lemma 31, for any τ and any $\pi^\star$

$$\begin{aligned}
\sum_{t=1}^{\tau}\mathbb{E}_t\left[\sum_\pi P_t(\pi)\widehat{\ell}_t(\pi(x_t))\right] - \sum_{t=1}^{\tau}\mathbb{E}_t\left[\widehat{\ell}_t(\pi^\star(x_t))\right] &\le \frac{\ln|\Pi|}{\eta_\tau} + \sum_{t=1}^{\tau}\frac{\eta_t}{2}\mathbb{E}_t\left[\sum_\pi P_t(\pi)\widehat{\ell}_t^2(\pi(x_t))\right] \\
&\le \frac{\ln|\Pi|}{\eta_\tau} + \sum_{t=1}^{\tau}\frac{\eta_t}{2}\sum_\pi P_t(\pi)\frac{\ell_t(\pi(x_t))}{q_t p_t(\pi(x_t))} \\
&\le \frac{\ln|\Pi|}{\eta_\tau} + K\frac{\eta_t}{2q_t} \le 2\sqrt{K\ln|\Pi|\sum_{t=1}^{\tau}\frac{1}{q_t}}.
\end{aligned}$$

Taking expectation on both sides finishes the proof. ∎

G.3. $(1 - 1/\log(K))$-Tsallis-INF (Algorithm 11 / Lemma 14)

Algorithm 11 $(1 - 1/\log(K))$-Tsallis-Inf (for strongly observable graphs)

Input: $\mathcal{G} \setminus \widehat{x}$.

Define: $\psi(x) = \sum_{i=1}^{K} \frac{x_i^\beta}{\beta(1-\beta)}$, where $\beta = 1 - 1/\log(K)$.

for $t = 1, 2, \ldots$ **do**

Receive update probability q_t.

Let

$$p_t = \operatorname*{argmin}_{x \in \Delta([K])} \left\{ \sum_{\tau=1}^{t-1} \langle x, \widehat{\ell}_\tau \rangle + \frac{1}{\eta_t} \psi(x) \right\}$$

$$\text{where} \quad \eta_t = \sqrt{\frac{\log K}{\sum_{s=1}^{t} \frac{1}{q_s}(1 + \min\{\widetilde{\alpha}, \alpha \log K\})}}.$$

Sample $A_t \sim p_t$.

With probability q_t receive $\ell_{t,i}$ for all $A_t \in \mathcal{N}(i)$ (in this case, set $\mathsf{upd}_t = 1$; otherwise, set $\mathsf{upd}_t = 0$).

Define

$$\widehat{\ell}_{t,i} = \mathsf{upd}_t \left(\frac{(\ell_{t,i} - \mathbb{I}\{i \in \mathcal{I}_t\}) \mathbb{I}\{A_t \in \mathcal{N}(i)\}}{\sum_{j \in \mathcal{N}(i)} p_{t,j} q_t} + \mathbb{I}\{i \in \mathcal{I}_t\} \right),$$

$$\text{where } \mathcal{I}_t = \left\{ i \in [K] : i \notin \mathcal{N}(i) \wedge p_{t,i} > \frac{1}{2} \right\} \text{ and } \mathcal{N}(i) = \{ j \in \mathcal{G} \setminus \widehat{x} \,|\, (j, i) \in E \}.$$

end

We require the following graph theoretic Lemmas.

Lemma 43 *Let $\mathcal{G} = (V, E)$ be a directed graph with independence number α and vertex weights $p_i > 0$, then*

$$\exists i \in V : \frac{p_i}{p_i + \sum_{j:(j,i) \in E} p_j} \leq \frac{2 p_i \alpha}{\sum_i p_i}.$$

Proof Without loss of generality, assume that $p \in \Delta(V)$, since the statement is scale invariant. The statement is equivalent to

$$\min_i p_i + \sum_{j:(j,i) \in E} p_j \geq \frac{1}{2\alpha}.$$

Let

$$\min_{p \in \Delta(V)} \min_i \left(p_i + \sum_{j:(j,i) \in E} p_j \right) \leq \min_{p \in \Delta(V)} \sum_i \left(p_i^2 + \sum_{j:(j,i) \in E} p_i p_j \right).$$

Via K.K.T. conditions, there exists $\lambda \in \mathbb{R}$ such that for an optimal solution $p^\star$ it holds

$$\forall i \in V : \text{either } \; 2p_i^\star + \sum_{j:(j,i)\in E} p_j^\star + \sum_{j:(i,j)\in E} p_j^\star = \lambda$$
$$\text{or } \; 2p_i^\star + \sum_{j:(j,i)\in E} p_j^\star + \sum_{j:(i,j)\in E} p_j^\star \geq \lambda \text{ and } p_i^\star = 0\,.$$

Next we bound λ. Take the sub-graph over $V_+ = \{i : \, p_i^\star > 0\}$ and take a maximally independence set S over V_+ ($|S| \leq \alpha$, since $V_+ \subset V$). We have

$$1 \leq \underbrace{\sum_{j \notin V_+} p_j^\star}_{=0} + \sum_{i \in S}\left(2p_i^\star + \sum_{j:(j,i)\in E} p_j^\star + \sum_{j:(i,j)\in E} p_j^\star\right) = |S|\lambda \leq \alpha\lambda.$$

Hence $\lambda \geq \frac{1}{\alpha}$. Finally,

$$\sum_i \left((p_i^\star)^2 + \sum_{j:(j,i)\in E} p_i^\star p_j^\star\right) = \frac{1}{2}\sum_i \left(2(p_i^\star)^2 + \sum_{j:(j,i)\in E} p_i^\star p_j^\star + \sum_{j:(i,j)\in E} p_i^\star p_j^\star\right) \geq \sum_i \frac{p_i^\star \lambda}{2} \geq \frac{1}{2\alpha}\,.$$

■

Lemma 44 *(Lemma 10 Alon et al. (2013)) let $\mathcal{G} = (V, E)$ be a directed graph. Then, for any distribution $p \in \Delta(V)$ we have:*

$$\sum_i \frac{p_i}{p_i + \sum_{j:(j,i)\in E} p_j} \leq \mathrm{mas}(\mathcal{G})\,,$$

where $\mathrm{mas}(\mathcal{G}) \leq \widetilde{\alpha}$ is the maximal acyclic sub-graph.

With these Lemmas, we are ready to proof the iw-stability.

Proof [Proof of Lemma 14] Consider the $(1 - 1/\log(K))$-Tsallis-INF algorithm with adaptive stepsize (Algorithm 11). Applying Lemma 27, we have for any stopping time τ and $a^\star \in [K]$:

$$\sum_{t=1}^{\tau} \mathbb{E}_t[\langle p_t, \widehat{\ell}_t\rangle - \widehat{\ell}_{t,a^\star}] \leq \frac{2e\log K}{\eta_\tau} + \sum_{t=1}^{\tau}\mathbb{E}_t\left[\max_p \langle p - p_t, \widehat{\ell}_t\rangle - \frac{1}{\eta_t}D_\psi(p, p_t)\right]$$
$$= \frac{2e\log K}{\eta_\tau} + \sum_{t=1}^{\tau}\mathbb{E}_t\left[\max_p \langle p - p_t, \widehat{\ell}_t + c_t\mathbf{1}\rangle - \frac{1}{\eta_t}D_\psi(p, p_t)\right]\,,$$

where $c_t = -\mathsf{upd}_t \frac{(\ell_{t,\mathcal{I}_t} - 1)\mathbb{I}\{A_t \in \mathcal{N}(\mathcal{I}_t)\}}{\sum_{j\in\mathcal{N}(\mathcal{I}_t)} p_{t,j} q_t}$, which is well defined because $\mathcal{I}_t$ contains by definition maximally one arm.

We can apply Lemma 32, since the losses are strictly positive.

$$\mathbb{E}_t\left[\max_p \langle p - p_t, \widehat{\ell}_t + c_t\mathbf{1}\rangle - \frac{1}{\eta_t}D_\psi(p, p_t)\right] \leq \frac{\eta_t}{2}\mathbb{E}_t\left[\sum_i p_{t,i}^{1+\frac{1}{\log K}}(\widehat{\ell}_{t,i} + c_t)^2\right]\,.$$

We split the vertices in three sets. Let $M_1 = \{i \,|\, i \in \mathcal{N}(i)\}$ the nodes with self-loops. Let $M_2 = \{i \,|\, i \notin M_1, \notin \mathcal{I}_t\}$ and $M_3 = \mathcal{I}_t$. we have

$$\frac{\eta_t}{2}\mathbb{E}_t\left[\sum_i p_{t,i}^{1+\frac{1}{\log K}}(\widehat{\ell}_{t,i}+c_t)^2\right]$$
$$\leq \eta_t\mathbb{E}_t\left[\mathsf{upd}_t\left(\sum_{i\in M_1\cup M_2}P_{t,i}c_t^2+\sum_{i\in M_1}\frac{p_{t,i}^{1+\frac{1}{\log K}}}{(\sum_{j\in\mathcal{N}(i)}p_{t,j}q_t)^2}+\sum_{i\in M_2}\frac{4p_{t,i}}{q_t^2}+p_{t,\mathcal{I}_t}\right)\right],$$

since for any $i \in M_2$, we have $\sum_{j\in\mathcal{N}(i)}p_{t,j} > \frac{1}{2}$. The first term is in expectation $\mathbb{E}_t\left[\sum_{i\in M_1\cup M_2}p_{t,i}c_t^2\right] \leq \frac{1}{q_t}$, the third term is in expectation $\mathbb{E}_t\left[\sum_{i\in M_2}\frac{4p_{t,i}\mathsf{upd}_t}{q_t^2}\right] \leq \frac{4}{q_t}$, while the second is

$$\mathbb{E}_t\left[\sum_{i\in M_1}\frac{p_{t,i}^{1+\frac{1}{\log K}}\mathsf{upd}_t}{(\sum_{j\in\mathcal{N}(i)}p_{t,j}q_t)^2}\right] \leq \frac{2}{q_t}\sum_{i\in M_1}\frac{p_{t,i}^{1+\frac{1}{\log K}}}{\sum_{j\in\mathcal{N}(i)}p_{t,j}}.$$

The subgraph consisting of M_1 contains sef-loops, hence we can apply Lemma 44 to bound this term by $\frac{2\widetilde{\alpha}}{q_t}$. If $\alpha\log K < \widetilde{\alpha}$, we instead take a permutation $\widetilde{p}$ of p, such that M_1 is in position $1,..,|M_1|$ and for any $i\in[|M_1|]$

$$\frac{\widetilde{p}_{t,i}}{\sum_{j\in\mathcal{N}(i)}\widetilde{p}_{t,j}} \leq \frac{\widetilde{p}_{t,i}}{\sum_{j\in\mathcal{N}(i)\cap[i]}\widetilde{p}_{t,j}} \leq \frac{2\alpha\widetilde{p}_{t,i}}{\sum_{j=1}^{i}\widetilde{p}_{t,j}}.$$

The existence of such a permutation is guaranteed by Lemma 43: Apply Lemma 43 on the graph M_1 to select $\tilde{p}_{|M_1|}$. Recursively remove the i-th vertex from the graph and apply Lemma 43 on the resulting sub-graph to pick $\tilde{p}_{i-1}$. Applying this permutation yields

$$\frac{2}{q_t}\sum_{i\in M_1}\frac{p_{t,i}^{1+\frac{1}{\log K}}}{\sum_{j\in\mathcal{N}(i)}p_{t,j}} \leq \frac{4\alpha}{q_t}\sum_{i=1}^{M_1}\frac{\widetilde{p}_i^{1+\frac{1}{\log K}}}{\sum_{j=1}^{i}\widetilde{p}_j} \leq \frac{4\alpha}{q_t}\sum_{i=1}^{M_1}\frac{\widetilde{p}_i}{\left(\sum_{j=1}^{i}\widetilde{p}_j\right)^{1-\frac{1}{\log K}}} \leq \frac{4\alpha\log K}{q_t}.$$

Combining everything

$$\mathbb{E}_t\left[\max_p\langle p-p_t,\widehat{\ell}_t+c_t\mathbf{1}\rangle-\frac{1}{\eta_t}D_\psi(p,p_t)\right] \leq \frac{\eta_t}{q_t}\left(6+4\min\{\widetilde{\alpha},\alpha\log K\}\right).$$

By the definition of the learning rate, we obtain

$$\sum_{t=1}^{\tau}\mathbb{E}_t[\langle p_t,\widehat{\ell}_t\rangle-\widehat{\ell}_{t,a^\star}] \leq \frac{2e\log K}{\eta_\tau}+\sum_{t=1}^{\tau}\frac{\eta_t}{q_t}\left(6+4\min\{\widetilde{\alpha},\alpha\log K\}\right)$$
$$\leq 18\sqrt{(1+\min\{\widetilde{\alpha},\alpha\log K\})\sum_{t=1}^{\tau}\frac{1}{q_t}\log K}.$$

■

G.4. EXP3 for weakly observable graphs (Algorithm 12 / Lemma 17)

Algorithm 12 EXP3 (for weakly observable graphs)

Input: $\mathcal{G} \setminus \widehat{x}$, dominating set D (potentially including $\widehat{x}$).
Define: $\psi(x) = \sum_{i=1}^{K} x_i \log(x_i)$. ν_D is the uniform distribution over D.
for $t = 1, 2, \dots$ **do**

Receive update probability q_t.
Let

$$P_t = \operatorname*{argmin}_{x \in \Delta([K])} \left\{ \sum_{\tau=1}^{t-1} \langle x, \widehat{\ell}_\tau \rangle + \frac{1}{\eta_t} \psi(x) \right\}, p_t = (1 - \gamma_t) P_t + \gamma_t \nu_D$$

$$\text{where} \quad \eta_t = \left(\left(\frac{\sqrt{\delta} \sum_{s=1}^{t} \frac{1}{\sqrt{q_s}}}{\log(K)} \right)^{\frac{2}{3}} + \frac{4\delta}{\min_{s \le t} q_s} \right)^{-1} \quad \text{and } \gamma_t = \sqrt{\frac{\eta_t \delta}{q_t}}.$$

Sample $A_t \sim p_t$.
With probability q_t receive $\ell_{t,i}$ for all $A_t \in \mathcal{N}(i)$ (in this case, set $\mathsf{upd}_t = 1$; otherwise, set $\mathsf{upd}_t = 0$).
Define

$$\widehat{\ell}_{t,i} = \frac{\ell_{t,i} \mathsf{upd}_t \mathbb{I}\{A_t \in \mathcal{N}(i)\}}{\sum_{j \in \mathcal{N}(i)} p_{t,j} q_t}.$$

end

Proof [Proof of Lemma 17] Consider the EXP3 algorithm (Algorithm 12). By standard analysis of FTRL (Lemma 27) and Lemma 31 by the non-negativity of all losses, for any τ and any $a^\star$,

$$\begin{aligned}
&\sum_{t=1}^{\tau} \mathbb{E}_t \left[\sum_a P_t(a) \widehat{\ell}_t(a) \right] - \sum_{t=1}^{\tau} \mathbb{E}_t \left[\widehat{\ell}_t(a^\star) \right] \\
&\le \frac{\log K}{\eta_\tau} + \sum_{t=1}^{\tau} \frac{\eta_t}{2} \mathbb{E}_t \left[\sum_i P_{t,i} \widehat{\ell}_{t,i}^2 \right] \\
&\le \frac{\log K}{\eta_\tau} + \sum_{t=1}^{\tau} \frac{\eta_t}{2} \sum_i P_{t,i} \frac{1}{\sum_{j \in \mathcal{N}(i)} p_{t,j} q_t} \\
&\le \frac{\log K}{\eta_\tau} + \sum_{t=1}^{\tau} \sqrt{\frac{\eta_t \delta}{4 q_t}}
\end{aligned}$$

where in the last inequality we use $\frac{\eta_t}{2}\sum_i P_{t,i}\times\frac{\delta}{\gamma_t}\times\frac{1}{q_t} = \frac{1}{2}\sqrt{\frac{\eta_t\delta}{q_t}}$. We have due to $\mathbb{E}_t[\langle P_t - p_t, \widehat{\ell}_t\rangle] \le \gamma_t$

$$\sum_{t=1}^{\tau}\mathbb{E}_t\left[\sum_a p_t(a)\widehat{\ell}_t(a)\right] - \sum_{t=1}^{\tau}\mathbb{E}_t\left[\widehat{\ell}_t(a^\star)\right] \le \frac{\log K}{\eta_\tau} + \sum_{t=1}^{\tau}\sqrt{9\frac{\eta_t\delta}{4q_t}}$$
$$\le 6(\delta\log(K))^{\frac{1}{3}}\left(\sum_{t=1}^{\tau}\frac{1}{\sqrt{q_t}}\right)^{\frac{2}{3}} + \frac{4\delta\log(K)}{\min_{t\le T} q_t}.$$

Taking expectation on both sides finishes the proof. ■

Appendix H. Surrogate Loss for Strongly Observable Graph Problems

When there exist arms such that $i \notin \mathcal{N}(i)$, we cannot directly apply Algorithm 2. To make the algorithm applicable to all strongly observable graphs, define the surrogate losses $\tilde{\ell}_t$ in the following way:

$$\tilde{\ell}_{t,\widehat{x}} = \ell_{t,\widehat{x}}\mathbb{I}\{(\widehat{x},\widehat{x})\in E\} - \sum_{j\in[K]\setminus\{\widehat{x}\}} p_{t,j}\ell_{t,j}\mathbb{I}\{(j,j)\notin E\}$$
$$\forall j\in[K]\setminus\{\widehat{x}\}: \tilde{\ell}_{t,j} = \ell_{t,j}\mathbb{I}\{(j,j)\in E\} - \ell_{t,\widehat{x}}\mathbb{I}\{(\widehat{x},\widehat{x})\notin E\},$$

where p_t is the distribution of the base algorithm $\mathcal{B}$ over $[K]\setminus\{\widehat{x}\}$ at round t. By construction and the definition of strongly observable graphs, $\tilde{\ell}_{t,j}$ is observed when playing arm j. (When the player does not have access to p_t, one can also sample one action from the current distribution for an unbiased estimate of $\ell_{t,\widehat{x}}$.) The losses $\tilde{\ell}_t$ are in range $[-1,1]$ instead of $[0,1]$ and can further be shifted to be strictly non-negative. Finally observe that

$$\mathbb{E}[\tilde{\ell}_{t,A_t} - \tilde{\ell}_{t,\widehat{x}}] = q_{t,2}\left(\sum_{j\in[K]\setminus\{\widehat{x}\}} p_{t,j}\tilde{\ell}_{t,j} - \tilde{\ell}_{t,\widehat{x}}\right)$$
$$= q_{t,2}\left(\sum_{j\in[K]\setminus\{\widehat{x}\}} p_{t,j}\ell_{t,j} - \ell_{t,\widehat{x}}\right)$$
$$= \mathbb{E}[\ell_{t,A_t} - \ell_{t,\widehat{x}}],$$

as well as

$$\mathbb{E}\left[\tilde{\ell}_{t,A_t} - \sum_{j\in[K]\setminus\{\widehat{x}\}} p_{t,j}\tilde{\ell}_{t,j}\right] = q_{t,1}\left(\tilde{\ell}_{t,\widehat{x}} - \sum_{j\in[K]\setminus\{\widehat{x}\}} p_{t,j}\tilde{\ell}_{t,j}\right)$$
$$= q_{t,1}\left(\tilde{\ell}_{t,\widehat{x}} - \sum_{j\in[K]\setminus\{\widehat{x}\}} p_{t,j}\ell_{t,j}\right)$$
$$= \mathbb{E}\left[\ell_{t,A_t} - \sum_{j\in[K]\setminus\{\widehat{x}\}} p_{t,j}\ell_{t,j}\right].$$

That means running Algorithm 2, replacing ℓ_t by $\tilde{\ell}_t$ allows to apply Theorem 11 to strongly observable graphs where not every arm receives its own loss as feedback.

Appendix I. Tabular MDP (Theorem 26)

In this section, we consider using the UOB-Log-barrier Policy Search algorithm in Lee et al. (2020) (their Algorithm 4) as our base algorithm. To this end, we need to show that it satisfies a dd-strongly-iw-stable condition specified in Lemma 42. We consider a variant of their algorithm by incorporating the feedback probability $q_t' = q_t + (1-q_t)\mathbb{I}[\pi_t = \widehat{\pi}]$. The algorithm is Algorithm 13.

We refer the readers to Section C.3 of Lee et al. (2020) for the setting descriptions and notations, since we will follow them tightly in this section.

I.1. Base algorithm

Lemma 45 *Algorithm 13 ensures for any $u^\star$,*

$$\mathbb{E}\left[\sum_{t=1}^T \langle w_t - u^\star, \ell_t\rangle\right] \le O\left(\sqrt{HS^2A\iota^2\mathbb{E}\left[\sum_{t=1}^T\sum_{s,a}\frac{\textit{upd}_t\mathbb{I}_t(s,a)\ell_t(s,a)}{q_t'^2}\right]} + \mathbb{E}\left[\frac{S^5A^2\iota^2}{\min_t q_t'}\right]\right).$$

Proof The regret is decomposed as the following (similar to Section C.3 in Lee et al. (2020)):

$$\sum_{t=1}^T\langle w_t - u^\star, \ell_t\rangle = \underbrace{\sum_{t=1}^T\langle w_t - \widehat{w}_t, \ell_t\rangle}_{\text{Error}} + \underbrace{\sum_{t=1}^T\langle \widehat{w}_t, \ell_t - \widehat{\ell}_t\rangle}_{\text{Bias-1}} + \underbrace{\sum_{t=1}^T\langle \widehat{w}_t - u, \widehat{\ell}_t\rangle}_{\text{Reg-term}}$$
$$+ \underbrace{\sum_{t=1}^T\langle u, \widehat{\ell}_t - \ell_t\rangle}_{\text{Bias-2}} + \underbrace{\sum_{t=1}^T\langle u - u^\star, \ell_t\rangle}_{\text{Bias-3}},$$

with the same definition of u as in (24) of Lee et al. (2020). Bias-3 $\le H$ trivially (see (25) of Lee et al. (2020)). For the remaining four terms, we use Lemma 49, Lemma 50, Lemma 51, and Lemma 52, respectively. Combining all terms finishes the proof. ■

Lemma 46 (Lemma C.2 of Lee et al. (2020)) *With probability $1 - O(\delta)$, for all t, s, a, s',*

$$\left|P(s'|s,a) - \overline{P}_t(s'|s,a)\right| \le \frac{\epsilon_t(s'|s,a)}{2}.$$

Lemma 47 (*cf.* Lemma C.3 of Lee et al. (2020)) *With probability at least $1-\delta$, for all h,*

$$\sum_{t=1}^T\sum_{s\in\mathcal{S}_h, a\in\mathcal{A}}\frac{q_t'\cdot w_t(s,a)}{\max\{1, N_t(s,a)\}} = O\left(|\mathcal{S}_h|A\log(T) + \ln(H/\delta)\right)$$

Proof The proof of this lemma is identical to the original one — only need to notice that in our case, the probability of obtaining a sample of (s,a) is $q_t'\cdot w_t(s,a)$. ■

Algorithm 13 UOB-Log-Barrier Policy Search

Input: state space $\mathcal{S}$, action space $\mathcal{A}$, candidate policy $\widehat{\pi}$.

Define: $\Omega = \left\{\widehat{w} : \ \widehat{w}(s,a,s') \geq \frac{1}{T^3 S^2 A}\right\}$, $\psi(w) = \sum_h \sum_{(s,a,s')\in\mathcal{S}_h\times\mathcal{A}\times\mathcal{S}_{h+1}} \ln \frac{1}{w(s,a,s')}$.

$\delta = \frac{1}{T^5 S^3 A}$.

Initialization:

$$\widehat{w}_1(s,a,s') = \frac{1}{|\mathcal{S}_h||\mathcal{A}||\mathcal{S}_{h+1}|}, \qquad \pi_1 = \pi^{\widehat{w}_1}, \qquad t^\star \leftarrow 1.$$

for $t = 1, 2, \ldots$ **do**

If $\pi_t = \widehat{\pi}$, $\mathsf{upd}_t = 1$; otherwise, $\mathsf{upd}_t = 1$ w.p. q_t and $\mathsf{upd}_t = 0$ w.p. $1 - q_t$.

If $\mathsf{upd}_t = 1$, obtain the trajectory $s_h, a_h, \ell_t(s_h, a_h)$ for all $h = 1, \ldots, H$.

Construct loss estimators

$$\widehat{\ell}_t(s,a) = \frac{\mathsf{upd}_t}{q'_t} \cdot \frac{\ell_t(s,a)\mathbb{I}_t(s,a)}{\phi_t(s,a)}, \quad \text{where } \mathbb{I}_t(s,a) = \mathbb{I}\{(s_{h(s)}, a_{h(s)}) = (s,a)\}$$

where $q'_t = q_t + (1-q_t)\mathbf{1}[\pi_t = \widehat{\pi}]$.

Update counters: for all s, a, s',

$$N_{t+1}(s,a) \leftarrow N_t(s,a) + \mathbb{I}_t(s,a), \quad N_{t+1}(s,a,s') \leftarrow N_t(s,a,s') + \mathbb{I}_t(s,a)\mathbb{I}\{s_{h(s)+1} = s'\}.$$

Compute confidence set

$$\mathcal{P}_{t+1} = \left\{\hat{P} : \ \left|\hat{P}(s'|s,a) - \overline{P}_{t+1}(s'|s,a)\right| \leq \epsilon_{t+1}(s'|s,a), \ \forall (s,a,s')\right\}$$

where $\overline{P}_{t+1}(s'|s,a) = \frac{N_{t+1}(s,a,s')}{\max\{1, N_{t+1}(s,a)\}}$ and

$$\epsilon_{t+1}(s'|s,a) = 4\sqrt{\frac{\overline{P}_{t+1}(s'|s,a)\ln(SAT/\delta)}{\max\{1, N_{t+1}(s,a)\}}} + \frac{28\ln(SAT/\delta)}{3\max\{1, N_{t+1}(s,a)\}}$$

if $\sum_{\tau=t^\star}^{t}\sum_{s,a} \frac{\mathsf{upd}_\tau \mathbb{I}_\tau(s,a)\ell_\tau(s,a)^2}{q_t'^2} + \max_{\tau\leq t}\frac{H}{q_\tau'^2} \geq \frac{S^2 A\ln(SAT)}{\eta_t^2}$ **then**

$\eta_{t+1} \leftarrow \frac{\eta_t}{2}$

$\widehat{w}_{t+1} = \mathrm{argmin}_{w\in\Delta(\mathcal{P}_{t+1})\cap\Omega}\,\psi(w)$

$t^\star \leftarrow t+1$.

end

else

$\eta_{t+1} = \eta_t$

$\widehat{w}_{t+1} = \mathrm{argmin}_{w\in\Delta(\mathcal{P}_{t+1})\cap\Omega}\left\{\langle w, \widehat{\ell}_t\rangle + \frac{1}{\eta_t}D_\psi(w, \widehat{w}_t)\right\}$.

end

Update policy $\pi_{t+1} = \pi^{\widehat{w}_{t+1}}$.

end

Lemma 48 (*cf.* Lemma C.6 of Lee et al. (2020)) *With probability at least $1 - O(\delta)$, for any t and any collection of transition functions $\{P_t^s\}_{s\in\mathcal{S}}$ such that $P_t^s \in \mathcal{P}_t$ for all s, we have*

$$\sum_{t=1}^{T}\sum_{s\in\mathcal{S},a\in\mathcal{A}} \left|w^{P_t^s,\pi_t}(s,a) - w_t(s,a)\right| \ell_t(s,a) = O\left(S\sqrt{HA\sum_{t=1}^{T}\frac{\langle w_t, \ell_t^2\rangle}{q_t'}\iota^2} + \frac{S^5A^2\iota^2}{\min_t q_t'}\right).$$

where $\iota \triangleq \log(SAT/\delta)$ and $\ell_t^2(s,a) \triangleq (\ell_t(s,a))^2$.

Proof Following the proof of Lemma C.6 in Lee et al. (2020), we have

$$\sum_{t=1}^{T}\sum_{s\in\mathcal{S},a\in\mathcal{A}} \left|w^{P_t^s,\pi_t}(s,a) - w_t(s,a)\right| \ell_t(s,a) \le B_1 + SB_2$$

where

$$B_2 \le O\left(\sum_{t=1}^{T}\sum_{s,a,s'}\sum_{x,y,x'}\sqrt{\frac{P(s'|s,a)\iota}{\max\{1,N_t(s,a)\}}}w_t(s,a)\sqrt{\frac{P(x'|x,y)\iota}{\max\{1,N_t(x,y)\}}}w_t(x,y|s')\right) + O\left(\sum_{t=1}^{T}\sum_{s,a,s'}\sum_{x,y,x'}\frac{w_t(s,a)\iota}{\max\{1,N_t(s,a)\}}\right) \quad (32)$$

The first term in (32) can be upper bounded by the order of

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{s,a,s'}\sum_{x,y,x'}\sqrt{\frac{w_t(s,a)P(x'|x,y)w_t(x,y|s')\iota}{\max\{1,N_t(s,a)\}}}\sqrt{\frac{w_t(s,a)P(s'|s,a)w_t(x,y|s')\iota}{\max\{1,N_t(x,y)\}}}\\
&\le \sum_{t=1}^{T}\sqrt{\sum_{s,a,s'}\sum_{x,y,x'}\frac{w_t(s,a)P(x'|x,y)w_t(x,y|s')\iota}{\max\{1,N_t(s,a)\}}}\sqrt{\sum_{s,a,s'}\sum_{x,y,x'}\frac{w_t(s,a)P(s'|s,a)w_t(x,y|s')\iota}{\max\{1,N_t(x,y)\}}}\\
&\le \sum_{t=1}^{T}\sqrt{H\sum_{s,a,s'}\frac{w_t(s,a)\iota}{\max\{1,N_t(s,a)\}}}\sqrt{H\sum_{x,y,x'}\frac{w_t(x,y)\iota}{\max\{1,N_t(x,y)\}}}\\
&\le HS\sum_{t=1}^{T}\sum_{s,a,s'}\frac{w_t(s,a)\iota}{\max\{1,N_t(s,a)\}}\\
&\le O\left(\frac{HS^2\iota}{\min_t q_t'}\left(SA\ln(T) + H\log(H/\delta)\right)\right). && \text{(by Lemma 47)}
\end{aligned}$$

The second term in (32) can be upper bounded by the order of

$$S^2A\sum_{t=1}^{T}\sum_{s,a,s'}\frac{w_t(s,a)\iota}{\max\{1,N_t(s,a)\}} = O\left(\frac{S^3A\iota}{\min_t q_t'}\left(SA\ln(T) + H\log(H/\delta)\right)\right). \quad \text{(by Lemma 47)}$$

Combining the two parts, we get

$$B_2 \le O\left(\frac{1}{\min_t q_t'}\left(S^4A^2\ln(T)\iota + HS^3A\iota\log(H/\delta)\right)\right) \le O\left(\frac{S^4A^2\iota^2}{\min_t q_t'}\right).$$

Next, we bound B_1. By the same calculation as in Lemma C.6 of Lee et al. (2020),

$$B_1 \le O\left(\sum_{t=1}^{T}\sum_{s,a,s'}\sum_{x,y} w_t(s,a)\sqrt{\frac{P(s'|s,a)\iota}{\max\{1,N_t(s,a)\}}}w_t(x,y|s')\ell_t(x,y) + H\sum_{t=1}^{T}\sum_{s,a,s'}\frac{w_t(x,a)\iota}{\max\{1,N_t(s,a)\}}\right)$$
$$\le O\left(\sum_{t=1}^{T}\sum_{s,a,s'}\sum_{x,y} w_t(s,a)\sqrt{\frac{P(s'|s,a)\iota}{\max\{1,N_t(s,a)\}}}w_t(x,y|s')\ell_t(x,y) + \frac{HS^2A\iota^2}{\min_t q'_t}\right)$$

For the first term above, we consider the summation over $(s,a,s') \in \mathcal{T}_h \triangleq \mathcal{S}_h \times \mathcal{A} \times \mathcal{S}_{h+1}$. We continue to bound it by the order of

$$\sum_{t=1}^{T}\sum_{s,a,s'\in\mathcal{T}_h}\sum_{x,y} w_t(s,a)\sqrt{\frac{P(s'|s,a)\iota}{\max\{1,N_t(s,a)\}}}w_t(x,y|s')\ell_t(x,y)$$
$$\le \alpha\sum_{t=1}^{T}\frac{1}{q'_t}\sum_{s,a,s'\in\mathcal{T}_h}\sum_{x,y} w_t(s,a)P(s'|s,a)w_t(x,y|s')\ell_t(x,y)^2$$
$$+ \frac{1}{\alpha}\sum_{t=1}^{T}\sum_{s,a,s'\in\mathcal{T}_h}\sum_{x,y} q'_t w_t(s,a)w_t(x,y|s')\cdot\frac{\iota}{\max\{1,N_t(s,a)\}}$$
(by AM-GM, holds for any $\alpha > 0$)
$$\le \alpha\sum_{t=1}^{T}\frac{1}{q'_t}\sum_{x,y} w_t(x,y)\ell_t(x,y)^2 + \frac{H|\mathcal{S}_{h+1}|}{\alpha}\sum_{t=1}^{T}\sum_{s,a\in\mathcal{S}_h\times\mathcal{A}}\frac{q'_t w_t(s,a)\iota}{N_t(s,a)}$$
$$\le \alpha\sum_{t=1}^{T}\frac{\langle w_t,\ell_t^2\rangle}{q'_t} + \frac{H|\mathcal{S}_{h+1}||\mathcal{S}_h|A\iota^2}{\alpha} \quad \text{(by Lemma 47)}$$
$$\le O\left(\sqrt{H|\mathcal{S}_h||\mathcal{S}_{h+1}|A\sum_{t=1}^{T}\frac{\langle w_t,\ell_t^2\rangle\iota^2}{q'_t}}\right) \quad \text{(choose the optimal } \alpha)$$
$$\le O\left((|\mathcal{S}_h|+|\mathcal{S}_{h+1}|)\sqrt{HA\sum_{t=1}^{T}\frac{\langle w_t,\ell_t^2\rangle\iota^2}{q'_t}}\right)$$

Therefore,

$$B_1 \le O\left(\sum_h(|\mathcal{S}_h|+|\mathcal{S}_{h+1}|)\sqrt{HA\sum_{t=1}^{T}\frac{\langle w_t,\ell_t^2\rangle\iota^2}{q'_t}} + \frac{HS^2A\iota^2}{\min_t q'_t}\right)$$
$$\le O\left(S\sqrt{HA\sum_{t=1}^{T}\frac{\langle w_t,\ell_t^2\rangle\iota^2}{q'_t}} + \frac{HS^2A\iota^2}{\min_t q'_t}\right).$$

Combining the bounds finishes the proof. ■

Lemma 49 (*cf.* Lemma C.7 of Lee et al. (2020))

$$\mathbb{E}\left[\text{Error}\right] = \mathbb{E}\left[\sum_{t=1}^{T}\langle \widehat{w}_t - w_t, \ell_t\rangle\right] \leq O\left(\mathbb{E}\left[S\sqrt{HA\sum_{t=1}^{T}\frac{\langle w_t, \ell_t^2\rangle \iota^2}{q_t'}} + \frac{S^5A^2\iota^2}{\min_t q_t'}\right]\right).$$

Proof By Lemma 48, with probability at least $1 - O(\delta)$,

$$\begin{aligned}\text{Error} &= \sum_{t=1}^{T}\langle \widehat{w}_t - w_t, \ell_t\rangle \leq \sum_{t=1}^{T}\sum_{s,a}|\widehat{w}_t(s,a) - w_t(s,a)|\ell_t(s,a)\\ &\leq O\left(S\sqrt{HA\sum_{t=1}^{T}\frac{\langle w_t, \ell_t^2\rangle \iota^2}{q_t'}} + \frac{S^5A^2\iota^2}{\min_t q_t'}\right)\end{aligned}$$

Furthermore, $|\text{Error}| \leq O(SAT)$ with probability 1. Therefore,

$$\begin{aligned}\mathbb{E}[\text{Error}] &\leq O\left(\mathbb{E}\left[S\sqrt{HA\sum_{t=1}^{T}\frac{\langle w_t, \ell_t^2\rangle \iota^2}{q_t'}} + \frac{S^5A^2\iota^2}{\min_t q_t'}\right] + \delta SAT\right)\\ &\leq O\left(\mathbb{E}\left[S\sqrt{HA\sum_{t=1}^{T}\frac{\langle w_t, \ell_t^2\rangle \iota^2}{q_t'}} + \frac{S^5A^2\iota^2}{\min_t q_t'}\right]\right)\end{aligned}$$

by our choice of δ. ∎

Lemma 50 (*cf.* Lemma C.8 of Lee et al. (2020))

$$\mathbb{E}[\text{Bias-1}] = \mathbb{E}\left[\sum_{t=1}^{T}\langle \widehat{w}_t, \ell_t - \widehat{\ell}_t\rangle\right] \leq O\left(\mathbb{E}\left[S\sqrt{HA\sum_{t=1}^{T}\frac{\langle w_t, \ell_t\rangle}{q_t'}} + \frac{S^5A^2\iota^2}{\min_t q_t'}\right]\right).$$

Proof Let $\mathcal{E}_t$ be the event that $P \in \mathcal{P}_\tau$ for all $\tau \leq t$.

$$\begin{aligned}\mathbb{E}_t\left[\langle \widehat{w}_t, \ell_t - \widehat{\ell}_t\rangle \,\middle|\, \mathcal{E}_t\right] &= \sum_{s,a}\widehat{w}_t(s,a)\ell_t(s,a)\left(1 - \frac{w_t(s,a)}{\phi_t(s,a)}\right)\\ &\leq \sum_{s,a}|\phi_t(s,a) - w_t(s,a)|\ell_t(s,a)\end{aligned}$$

Thus,

$$\mathbb{E}_t\left[\langle \widehat{w}_t, \ell_t - \widehat{\ell}_t\rangle\right] \leq \sum_{s,a}|\phi_t(s,a) - w_t(s,a)|\ell_t(s,a) + O(H\mathbb{I}[\overline{\mathcal{E}}_t])$$

Summing this over t,

$$\sum_{t=1}^{T}\mathbb{E}_t\left[\langle \widehat{w}_t, \ell_t - \widehat{\ell}_t\rangle\right] \leq \underbrace{\sum_{t=1}^{T}\sum_{s,a}|\phi_t(s,a) - w_t(s,a)|\ell_t(s,a)}_{(\star)} + O\left(H\sum_{t=1}^{T}\mathbb{I}[\overline{\mathcal{E}}_t]\right)$$

By Lemma 48, $(\star)$ is upper bounded by $O\left(S\sqrt{HA\sum_{t=1}^{T}\frac{\langle w_t,\ell_t^2\rangle}{q'_t}\iota^2+\frac{S^5A^2\iota^2}{\min_t q'_t}}\right)$ with probability $1-O(\delta)$. Taking expectations on both sides and using that $\Pr[\overline{\mathcal{E}}_t]\leq\delta$, we get

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle\widehat{w}_t,\ell_t-\widehat{\ell}_t\rangle\right]\leq O\left(\mathbb{E}\left[S\sqrt{HA\sum_{t=1}^{T}\frac{\langle w_t,\ell_t^2\rangle}{q'_t}\iota^2+\frac{S^5A^2\iota^2}{\min_t q'_t}}\right]\right)+O\left(\delta SAT\right)$$
$$\leq O\left(\mathbb{E}\left[S\sqrt{HA\sum_{t=1}^{T}\frac{\langle w_t,\ell_t^2\rangle}{q'_t}\iota^2+\frac{S^5A^2\iota^2}{\min_t q'_t}}\right]\right).$$

■

Lemma 51 (*cf.* Lemma C.9 of Lee et al. (2020))

$$\mathbb{E}[\text{Bias-2}]=\mathbb{E}\left[\sum_{t=1}^{T}\langle u,\widehat{\ell}_t-\ell_t\rangle\right]\leq O(1).$$

Proof Let $\mathcal{E}_t$ be the event that $P\in\mathcal{P}_\tau$ for all $\tau\leq t$. By the construction of the loss estimator, we have

$$\mathbb{E}_t\left[\langle u,\widehat{\ell}_t-\ell_t\rangle\;\middle|\;\mathcal{E}_t\right]\leq 0$$

and thus

$$\mathbb{E}_t\left[\langle u,\widehat{\ell}_t-\ell_t\rangle\right]\leq\mathbb{I}[\overline{\mathcal{E}}_t]\cdot H\cdot T^3S^2A\quad\text{(by Lemma C.5 of Lee et al. (2020), }\widehat{\ell}_t(s,a)\leq T^3S^2A)$$

and

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle u,\widehat{\ell}_t-\ell_t\rangle\right]\leq HT^4S^2A\,\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}[\overline{\mathcal{E}}_t]\right]\leq O(\delta HT^5S^2A)\leq O(1),$$

where the last inequality is by our choice of δ. ■

Lemma 52 (*cf.* Lemma C.10 of Lee et al. (2020))

$$\mathbb{E}[\text{Reg-term}]\leq O\left(\sqrt{S^2A\ln(SAT)\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s,a}\frac{\mathsf{upd}_t\mathbb{I}_t(s,a)\ell_t(s,a)^2}{q_t'^2}+\max_t\frac{H}{q_t'^2}\right]}\right).$$

Proof By the same calculation as in the proof of Lemma C.10 in Lee et al. (2020),

$$\langle\widehat{w}_t-u,\widehat{\ell}_t\rangle\leq\frac{D_\psi(u,\widehat{w}_t)-D_\psi(u,\widehat{w}_{t+1})}{\eta_t}+\eta_t\sum_{s,a}\widehat{w}_t(s,a)^2\widehat{\ell}_t(s,a)^2$$
$$\leq\frac{D_\psi(u,\widehat{w}_t)-D_\psi(u,\widehat{w}_{t+1})}{\eta_t}+\eta_t\sum_{s,a}\frac{\mathsf{upd}_t\mathbb{I}_t(s,a)\ell_t(s,a)^2}{q_t'^2}.$$

Let $t_1, t_2, \ldots$ be the time indices when $\eta_t = \frac{\eta_{t-1}}{2}$, and let $t_{i^\star}$ be the last time this happens. Summing the inequalities above and using telescoping, we get

$$\begin{aligned}
\sum_{t=1}^{T}\langle \widehat{w}_t - u, \widehat{\ell}_t\rangle &\le \sum_i \left[\frac{D_\psi(u,\widehat{w}_{t_i})}{\eta_{t_i}} + \eta_{t_i}\sum_{t=t_i}^{t_{i+1}-1}\sum_{s,a}\frac{\mathsf{upd}_t \mathbb{I}_t(s,a)\ell_t(s,a)^2}{q_t'^2}\right] \\
&\le \sum_i \left[\frac{O(S^2A\ln(SAT))}{\eta_{t_i}} + \eta_{t_i}\sum_{t=t_i}^{t_{i+1}-1}\sum_{s,a}\frac{\mathsf{upd}_t \mathbb{I}_t(s,a)\ell_t(s,a)^2}{q_t'^2}\right] \\
&\qquad \text{(computed in the proof of Lemma C.10 in Lee et al. (2020))} \\
&\le \sum_i \frac{O(S^2A\ln(SAT))}{\eta_{t_i}} \qquad \text{(by the timing we halve the learning rate)} \\
&= O\left(\frac{S^2A\ln(SAT)}{\eta_{t_{i^\star}}}\right) \\
&\le O\left(\sqrt{S^2A\ln(SAT)\left(\sum_{t=1}^{T}\sum_{s,a}\frac{\mathsf{upd}_t \mathbb{I}_t(s,a)\ell_t(s,a)^2}{q_t'^2} + \max_t \frac{H}{q_t'^2}\right)}\right).
\end{aligned}$$

■

I.2. Corraling

We use Algorithm 8 as the corral algorithm for the MDP setting, with Algorithm 13 being the base algorithm. The guarantee of Algorithm 8 is provided in Lemma 42.

Proof [Proof of Theorem 26] To apply Lemma 42, we have to perform re-scaling on the loss because for MDPs, the loss of a policy in one round is H. Therefore, scale down all losses by a factor of $\frac{1}{H}$. Then by Lemma 45, the base algorithm satisfies the condition in Lemma 42 with $c_1 = S^2A\iota^2$ and $c_2 = S^5A^2\iota^2/H$ where $\xi_{t,\pi}$ is defined as $\ell'_{t,\pi} = \ell_{t,\pi}/H$, the expected loss of policy π after scaling. Therefore, by Lemma 42, the we can transform it to an $\frac{1}{2}$-dd-LSB algorithm with $c_1 = S^2A\iota^2$ and $c_2 = S^5A^2\iota^2/H$. Finally, using Theorem 23 and scaling back the loss range, we can get

$$O\left(\sqrt{S^2AHL_\star\log^2(T)\iota^2} + S^5A^2\log^2(T)\iota^2\right)$$

regret in the adversarial regime, and

$$O\left(\frac{H^2S^2A\iota^2\log T}{\Delta} + \sqrt{\frac{H^2S^2A\iota^2\log T}{\Delta}C} + S^5A^2\iota^2\log(T)\log(C\Delta^{-1})\right)$$

regret in the corrupted stochastic regime. ■