Multi-Armed Sampling Problem and the End of Exploration

Mohammad Pedramfar **Siamak Ravanbakhsh**
Mila - Quebec AI Institute, McGill University
{mohammad.pedramfar, siamak.ravanbakhsh}@mila.quebec

Abstract

This paper introduces the framework of multi-armed sampling, which serves as the sampling counterpart to the optimization problem of multi-armed bandits. Our primary motivation is to rigorously examine the exploration-exploitation trade-off in the context of sampling. We systematically define plausible notions of regret for this framework and establish corresponding lower bounds. We then propose a simple algorithm that achieves near-optimal regret bounds. Our theoretical results suggest that, in contrast to optimization, *sampling barely requires any exploration.* To further connect our findings with those of multi-armed bandits, we define a continuous family of problems and associated regret measures that smoothly interpolate and unify multi-armed sampling and multi-armed bandit problems using a temperature parameter. We believe that the multi-armed sampling framework and our findings in this setting can play a foundational role in the study of sampling, including recent neural samplers, much like the role of multi-armed bandits in reinforcement learning. In particular, our work sheds light on the role of exploration (or lack thereof) and the convergence properties of algorithms for entropy-regularized reinforcement learning, fine-tuning of pretrained models and reinforcement learning with human feedback (RLHF).

1 INTRODUCTION

Efficiently sampling from a computationally expensive target density is a key challenge in sciences with applications ranging from drug and material discovery to ecology and climate science. The problem involves sampling from a target distribution with density p which is described by an unnormalized energy model $p(x) = \exp(-\mathcal{E}(x))/Z$ with normalizing constant $Z = \int \exp(-\mathcal{E}(x))dx$. To make the notation similar to reinforcement learning, we refer to $r(x) = -\mathcal{E}(x)$ as the reward and assume that we have access to an unbiased estimate of r, but not the normalizing constant Z or samples from p. Obtaining estimates of the reward function often involves expensive and potentially noisy simulations or experiments, which motivates the development of efficient sampling techniques. Examples in high-impact application areas include energy estimates based on approximate quantum mechanical calculations for new drug and material discovery, or large-scale hydrodynamical simulations in cosmology and astrophysics. This problem is ubiquitous in Bayesian statistics and machine learning and has been an object of study for decades, with Monte-Carlo methods (Duane et al., 1987; Roberts and Tweedie, 1996; Hoffman et al., 2014; Leimkuhler et al., 2014; Lemos et al., 2023) recently being complemented by deep generative models (Albergo et al., 2019; Noé et al., 2019; Gabrié et al., 2021; Midgley et al., 2022; Akhound-Sadegh et al., 2024). Models such as GFlowNets (Bengio et al., 2021a; Malkin et al., 2022; Jain et al., 2022, 2023b; Kim et al., 2024; Jain et al., 2023a) also directly address this problem.

It is important to note that there are many problems that are not commonly formulated as sampling problems but are equivalent to one. An important example is entropy-regularized RL. It is a well-known fact that the optimization objective of entropy regularized RL could be formulated as the minimization of KL-divergence between the policy and a certain form of target distribution function. In particular, in the bandit case where there is only a single state and the Q-function is the same as the reward function, the target density becomes $\exp(r(x))/Z$, which is the sampling target that is the focus of our work. (See Ziebart (2010); Haarnoja et al. (2017) and Theorem 7.) In fact, it has been shown that, in many cases, the task of learning a

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

GFlowNet may be formulated as entropy-regularized RL (Tiapkin et al., 2024).

An important application of entropy-regularized RL is in RLHF and fine-tuning of large pretrained models Jaques et al. (2019); Stiennon et al. (2020); Bai et al. (2022); Ouyang et al. (2022); Azar et al. (2024), such as LLMs (e.g. Rafailov et al. (2023)) and diffusion models (e.g. Fan et al. (2023)). In LLMs, the equivalence of the DPO objective with a minimization of reverse-KL divergence could be seen in Rafailov et al. (2023). At each time step, the learner generates a sample and observes the associated reward. The DPO objective is equivalent to minimizing $D_{KL}(p_\theta||p)$, where p_θ is the model that is being tuned, and p is the target density defined by

$$p(y|x) := \frac{1}{Z} p_{\text{ref}}(y|x) \exp(r(y|x)),$$

where x is the context, e.g., prompt, y is a sample, $p(y|x)$ is the likelihood of generating y given x in the optimal target model, p_{ref} is the pretrained model and Z is the normalization constant. This formulation is adopted in nearly all RLHF methods for LLM alignment (Bai et al., 2022; Ouyang et al., 2022; Touvron et al., 2023). *It is important to note that, unlike the classical RL objective, the goal here is not to find "the best" sample. Instead, the goal is to learn a policy that generates samples according to the target distribution.*

Despite all the progress in this field, the exploration-exploitation trade-off between sampling from identified modes and exploration to find new modes has not been addressed so far. The optimization counterpart of the sampling problem, which aims to minimize the energy, or maximize the reward corresponds to established areas of machine learning, namely Bayesian Optimization (BO; Snoek et al., 2012) and Reinforcement Learning (Sutton, 2018). While the importance of exploration-exploitation trade-off and experimental design in the optimization paradigms is well-studied and understood, this trade-off in the context of sampling has gone unnoticed.

Motivated by the defining role of multi-arm bandit algorithms (Berry and Fristedt, 1985; Lattimore and Szepesvári, 2020) in addressing this trade-off in the optimization setting, we set out to investigate an analogous problem of multi-arm sampling. In this problem, at each timestep, the learner generates a sample and observes the reward, just as in a classical multi-armed bandit. In order to rigorously address the exploration-exploitation tradeoff in this setting, we first need a uniform way to measure the performance of an algorithm. In classical RL theory, this is done using the notion of regret. However, as we mentioned earlier, the goal of sampling is not to maximize a scalar, but to match a target distribution. Thus, we start with a systematic study of what "regret" could plausibly mean and how different notions of regret relate to each other. Once the notions of regret are established, we investigate the greedy algorithm and an algorithm with minor exploration, to show that, at least in the multi-armed case, unlike RL, sampling barely requires any exploration.

While previous works consider similar problems, they are limited to special cases and either obtain suboptimal regret bounds or require significant exploration. In particular, (Lu et al., 2018) considered this problem with a notion of regret that we refer to as forward-KL policy-level regret and obtained $\tilde{O}(T^{1/2})$ regret bound using of UCB-type exploration. Similarly, (Zhao et al., 2025b) also considered UCB-type exploration, but obtained logarithmic reverse-KL policy-level regret. Our work is the first to show that the performance of the greedy algorithm is nearly optimal, and the nature of the exploration-exploitation tradeoff is different from that of RL. Among prior works (Xiong et al., 2024; Zhao et al., 2025a,b) consider a more general setting where the number of arms is not finite. In contrast, our focus on the more basic multi-armed setting enables us to provide a more comprehensive framework and analysis that also shows the redundancy of exploration for a range of plausible settings. For a comprehensive literature review, please see Appendix A.

In addition to its potential role for understanding the exploration-exploitation trade-off in sampling, the multi-armed sampling problem can have direct applications of its own, in areas where multi-armed bandits are found useful; examples include online advertising, recommender systems, and adaptive routing. The benefit lies in avoiding repetition and providing diverse actions, choices, or recommendations. For example, suggesting different ads to the same visitor based on their click probability can be preferred to recommending the same "optimal" choice in every visit that ends with no interaction.

1.1 Outline and Contributions

The main contributions of our work is listed below.

1. We introduce the multi-armed sampling framework as a theoretical counterpart to multi-armed bandits, tailored for sampling tasks.

2. We systematically define various potential notions of regret specific to multi-armed sampling and discuss their relations with each other.

3. We establish lower bounds for all notions of regret discussed earlier.

4. We propose a simple algorithm, called Greedy Active Sampling with Warm-up, that achieves near-optimal regret bounds with respect to all notions of regret with little to no exploration. This is a surprising result, given that sampling reduces to optimization at zero temperature limit, and optimization does require explicit exploration.

5. We then generalize our regret bound to all algorithms that behave within the confidence bound (See Theorem 6), thus showing that a similar regret bound holds algorithms using optimism, Thompson sampling, or even pessimism.

6. We present a continuous family of problems that interpolates between multi-armed sampling and multi-armed bandit problems, offering a unified theoretical perspective.

2 PROBLEM SETUP

A *multi-armed bandit sampling* problem, or simply *multi-armed sampling* problem, is specified with a choice of regret (as we will discuss in Section 3) and a multi-armed bandit environment. The interaction between the policy and the environment is identical to that of multi-armed bandits; meaning that, at each time-step, the policy selects an action based on the history so far and observes a noisy sample of the reward. The core difference is the goal of the policy, encoded within the notion of regret.

In this section, we focus on the environment. Formally, the environment is a collection of distributions $\nu = (P_i)_{1\leq i\leq k}$ where $k > 1$ is the number of actions. The agent and the environment interact sequentially over T time-steps. In each time-step $1 \leq t \leq T$, the agent chooses an arm $\mathbf{a}_t \in [k] := \{1, \cdots, k\}$. Then the environment samples a return value $\mathbf{x}_t$ from the distribution $P_{\mathbf{a}_t}$ and reveals $\mathbf{x}_t$ to the agent. We define the *unnormalized log-probability*, *negative energy*, or *reward* as $r := (r_i)_{i=1}^k$ where r_i is the mean of the distribution P_i, for $1 \leq i \leq k$. The target distribution of ν is defined as $p := \mathrm{softmax}(r)$, i.e., $p_i = e^{r_i} / \left(\sum_{j=1}^k e^{r_j}\right)$, for $1 \leq i \leq k$. We also define $\mathbf{n}_{t,i} := \sum_{t'=1}^t \mathbf{1}_{\mathbf{a}_{t'}=i}$ and $\mathbf{q}_{t,i} := \mathbf{n}_{t,i}/t$. In other words, $\mathbf{q}_t = (\mathbf{q}_{t,1}, \cdots, \mathbf{q}_{t,k})$ is the empirical probability distribution of the actions selected up to and including time-step t. For $1 \leq t \leq T+1$, let $\mathbf{h}_t = (\mathbf{a}_1, \mathbf{x}_1, \cdots, \mathbf{a}_{t-1}, \mathbf{x}_{t-1})$ denote the history of actions and observation up to, but not including, time-step t and let $\mathcal{H}_t := ([k] \times \mathbb{R})^t$ denote the space of all possible histories. A policy π is collection of functions $(\pi_t)_{t=1}^T$ and a probability space Ω^π. Here the probability space Ω^π captures all the randomness in the policy. Before taking the first action, random variable ω^π is sampled from Ω^π and remains constant throughout the "game". Then, at time-step t, the function $\pi_t : \Omega^\pi \times \mathcal{H}_t \to [k]$ maps $(\omega^\pi, \mathbf{h}_t) \in \Omega^\pi \times \mathcal{H}_t$ to the next action $\mathbf{a}_t$. We say a policy is *deterministic* if Ω^π only contains a single point, or equivalently, if $\mathbf{a}_t$ is a deterministic function of the history. We will use the notation $\hat{\pi}_t := (\mathbb{P}(\mathbf{a}_t = 1|\mathbf{h}_t), \cdots, \mathbb{P}(\mathbf{a}_t = k|\mathbf{h}_t)) \in \mathbb{R}^k$ to denote the distribution of $\mathbf{a}_t$ conditioned on the history. We will also use $\omega^\nu \in \Omega^\nu := \mathbb{R}^{T\times k}$ to denote a random matrix that captures all of the randomness of the environment ν. This matrix, which we call a *realization of* ν, specifies the reward at each arm and each time-step, i.e., if at time-step t the arm i is sampled, then the observed reward will be $(\omega^\nu)_{t,i}$.

For $\sigma > 0$ we define $\mathcal{E}^k_{\mathrm{SG}}(\sigma^2)$ as the class of environments $\nu = (P_i)_{1\leq i\leq k}$ where each P_i is σ-subgaussian. Similarly, we define $\mathcal{E}^k_{\mathrm{B}}(\sigma)$ to be the class of environments where the absolute value of the noise of each arm is at most σ. It follows from Hoeffding's lemma that $\mathcal{E}^k_{\mathrm{B}}(\sigma) \subseteq \mathcal{E}^k_{\mathrm{SG}}(\sigma^2)$.

We refer to Appendix B for a discussion on the necessity of noise and Appendix C for the role of softmax in the problem formulation.

3 NOTIONS OF REGRET

The goal of the agent is the multi-armed sampling problem is to match the target distribution p. To formalize this idea, we need to decide exactly what distribution should be close to p and how we measure this closeness. We assume a *performance distribution*, $f(\nu, \pi, \omega^\nu, \omega^\pi)$, representing some notion of agent's sampling distribution, which is a deterministic function of $(\nu, \pi, \omega^\nu, \omega^\pi)$, and a statistical distance d between this performance distribution and the target distribution. The (f, d)-*regret* is then defined as

$$\mathcal{R}^{f,d}(\pi, \nu, \omega^\pi, \omega^\nu) := d\left(f(\pi, \nu, \omega^\pi, \omega^\nu), p\right).$$

Hence, an algorithm that tries to minimize (f, d)-regret is attempting to make the distribution f close to p with respect to the statistical distance d. We will simply use the notation $\mathcal{R}^{f,d}(\pi, \nu)$ or $\mathcal{R}^{f,d}$ when there is no ambiguity. Since f is allowed to depend on ω^ν and ω^π, it is a random variable and therefore (f, d)-regret is also a random variable. The *expected* (f, d)*-regret* of policy π over environment ν is defined as

$$\mathbb{E}\left[\mathcal{R}^{f,d}\right] := \mathbb{E}_{\omega^\nu,\omega^\pi}\left[d\left(f(\nu, \pi, \omega^\nu, \omega^\pi), p\right)\right],$$

where the expectation is taken with respect to the randomness of ω^ν and ω^π. More generally, if $l \geq 1$ and $f = (f_1, \cdots, f_l)$ is a family of performance distributions as above, we define (f, d)-regret and *expected* (f, d)-

Table 1: Comparison of MAS algorithms

Regret type			Reference	Regret	Lower bound
TV	$\mathcal{P}$	$\mathcal{S}$	Theorem 5	$\tilde{O}(T^{-1/2})$	$\Omega(T^{-1/2})$
		$\mathcal{C}$	Theorem 5	$\tilde{O}(T^{1/2})$	$\Omega(T^{1/2})$
	$\mathcal{A}$	$\mathcal{S}$	Theorem 5	$\tilde{O}(T^{-1/2})$	$\Omega(T^{-1/2})$
		$\mathcal{C}$	Theorem 5	$\tilde{O}(T^{1/2})$	$\Omega(T^{1/2})$
r-KL	$\mathcal{P}$	$\mathcal{S}$	Online Iterative GSHF Xiong et al. (2024)	$\tilde{O}(T^{-1/2})$	$\Omega(T^{-1})$
			Two-Stage Mixed-Policy Sampling Zhao et al. (2025a)	$\tilde{O}(T^{-1})$	
			KL-UCB Zhao et al. (2025b)	$\tilde{O}(T^{-1})$	
			Theorem 5	$\tilde{O}(T^{-1})$	
		$\mathcal{C}$	KL-UCB Zhao et al. (2025b)	$\tilde{O}(1)$	$\Omega(1)$
			Theorem 5	$\tilde{O}(1)$	
	$\mathcal{A}$	$\mathcal{S}$	Theorem 5	$\tilde{O}(T^{-1})$	$\Omega(T^{-1})$
		$\mathcal{C}$	Theorem 5	$\tilde{O}(1)$	$\Omega(1)$
f-KL †	$\mathcal{P}$	$\mathcal{S}$	Theorem 5	$\tilde{O}(T^{-1})$	$\Omega(T^{-1})$
		$\mathcal{C}$	DAISEE Lu et al. (2018)	$\tilde{O}(T^{1/2})$	$\Omega(1)$
			Theorem 5	$\tilde{O}(1)$	
	$\mathcal{A}$	$\mathcal{S}$	Theorem 5	$\tilde{O}(T^{-1})$	$\Omega(T^{-1})$
		$\mathcal{C}$	Theorem 5	$\tilde{O}(1)$	$\Omega(1)$

Lower-bounds that are matched by the regret bound of our simple algorithm that forgoes exploration. Here, regret types vary along three axes: 1) choice statistical distance; 2) policy-level ($\mathcal{P}$) vs. action-level ($\mathcal{A}$), and 3) simple ($\mathcal{S}$) vs. cumulative ($\mathcal{C}$) regret.
Note that the problem considered in Xiong et al. (2024); Zhao et al. (2025a,b) is KL-regularized bandit, in contextual setting with continuous arms. Moreover, the setting of Xiong et al. (2024) uses preference feedback model, which is slightly different than the ground-truth reward model used in other two papers. As we further discuss in Section 7, KL-regularized bandit is equivalent to sampling with policy level reverse-KL regret. Even though the setting we consider is finite-armed setting and therefore more limited, it allows us to provide a more comprehensive formulation and analysis. As we see, the results of Xiong et al. (2024); Lu et al. (2018) are sub-optimal. The algorithm proposed in Zhao et al. (2025a) uses no exploration but the proposed algorithm is not a simple greedy algorithm and the analysis is limited to simple policy-level reverse-KL regret. The algorithm in Zhao et al. (2025b) use UCB-style exploration to obtain near optimal bounds for simple and cumulative policy-level reverse-KL regret. We note that Lu et al. (2018) also uses a form of UCB-style exploration, but the regret analysis provides sub-optimal bounds for cumulative policy-level forward-KL regret. In our result, we consider two algorithms, one pure greedy and another with a small amount of exploration (much lower than that of optimism-based approaches), and show that the all forms of regret considered are near-optimal. In fact, as we show in Theorem 6, our main result also directly applies to KL-UCB and show that it also enjoys near-optimal regret bound for all forms of regret.
†: In the case of forward-KL distance, we assume that environment has bounded noise. Also see Remark 1 about the notion of regret in this case.

regret as their corresponding summation

$$\mathcal{R}^{f,d} := \sum_{s=1}^{l} d\left(f_s(\nu, \pi, \omega^\nu, \omega^\pi), p\right),$$

$$\mathbb{E}\left[\mathcal{R}^{f,d}\right] := \mathbb{E}\left[\sum_{s=1}^{l} d\left(f_s(\nu, \pi, \omega^\nu, \omega^\pi), p\right)\right].$$

This is a general notion meant to capture many plausible formalizations of the idea of regret in multi-armed sampling problems. In the remainder of this section, we consider some possible options for f and d.

3.1 The statistical distance

We need a notion of distance in order to compare distributions. In this work, we mostly focus on three notions: total variation distance, reverse KL distance, and forward KL distance. Specifically, given a distribution $\hat{p}$ and the target distribution p, we define $d^{\text{r-KL}}(\hat{p}, p) := D_{\text{KL}}(\hat{p}||p) = \sum_{i=1}^{k} \hat{p}_i \log(\hat{p}_i/p_i)$ and $d^{\text{f-KL}}(\hat{p}, p) := D_{\text{KL}}(p||\hat{p}) = \sum_{i=1}^{k} p_i \log(p_i/\hat{p}_i)$, where we use log to denote the natural logarithm and $d^{\text{TV}}(\hat{p}, p) := \sup_{A \subseteq [k]} |\hat{p}(A) - p(A)| = \frac{1}{2}\sum_{i=1}^{k} |\hat{p}_i - p_i|$. We simply use d and drop the superscript when the type of distance is either irrelevant or clear from the context.

3.2 Action-level vs. policy-level regret

Next we need to decide the performance function, i.e., the distribution that is being compared to the target distribution.

When the performance function is $\hat{\pi}_t$, we refer to the regret as the *policy-level regret* denoted by $\mathcal{PR} := d(\hat{\pi}_t, p)$. Minimizing this notion of regret means that we require $\hat{\pi}_t$ to remain close to p in every time-step.

This is a strong requirement and there are useful algorithms for which this regret is not minimized. For example, one could use Metropolis-Hasting algorithm to sample from a target p using a proposal distribution for local moves, which means $\hat{\pi}_t$ will also have a

local support. In fact, the chain distribution in Markov Chain Monte Carlo methods is the average of $\hat{\pi}_t$ over the randomness of the policy, i.e., $\mathbb{E}_{\omega^\pi}[\hat{\pi}_t]$. Thus, in order to handle such scenarios, we define *environment policy-level regret* by $\mathcal{PR}^{\text{env}} := d\left(\mathbb{E}_{\omega^\pi}[\hat{\pi}_t], p\right)$. In this case, the notion of regret is no longer defined for any specific run of the algorithm, but for a all runs of the algorithm in a given realization of the environment, i.e., we average over all values of ω^π while keeping ω^ν fixed. Thus, the two types of expected policy-level regrets are given by

$$\mathbb{E}\left[\mathcal{PR}\right] = \mathbb{E}_{\omega^\nu,\omega^\pi}\left[d\left(\hat{\pi}_t, p\right)\right],$$
$$\mathbb{E}\left[\mathcal{PR}^{\text{env}}\right] = \mathbb{E}_{\omega^\nu}\left[d\left(\mathbb{E}_{\omega^\pi}\left[\hat{\pi}_t\right], p\right)\right].$$

One problem with policy-level regret is that it is not always meaningful. For example, assume the environment has two arms and consider the policy that chooses arm 1 in odd time steps and arm 2 in even time steps. If the target distribution is $p = (1/2, 1/2)$, then the policy is behaving well with respect to the target, but its policy-level regret is the statistical distance between p and $(1, 0)$ or $(0, 1)$, which is high.

When the performance function is $\mathbf{q}_t$, we refer to the regret $\mathcal{AR}(\pi) := d\left(\mathbf{q}_t, p\right)$ as the *action-level regret*. Note that $\mathbf{q}_t$ is a deterministic function of $\mathbf{h}_{t+1}$ and therefore it is a random variable. Thus the *expected action-level regret* is given by $\mathbb{E}_{\omega^\nu,\omega^\pi}\left[d\left(\mathbf{q}_t, p\right)\right]$. Similar to above, we define *environment action-level regret* by $\mathcal{AR}^{\text{env}} := d\left(\mathbb{E}_{\omega^\pi}[\mathbf{q}_t], p\right)$. Thus, the two types of expected action-level regrets are given by

$$\mathbb{E}\left[\mathcal{AR}\right] = \mathbb{E}_{\omega^\nu,\omega^\pi}\left[d\left(\mathbf{q}_t, p\right)\right],$$
$$\mathbb{E}\left[\mathcal{AR}^{\text{env}}\right] = \mathbb{E}_{\omega^\nu}\left[d\left(\mathbb{E}_{\omega^\pi}\left[\mathbf{q}_t\right], p\right)\right].$$

3.3 Simple vs. cumulative regret

In bandit literature, simple regret refers to the cost of the action at the last time-step relative to the optimal action. Inspired by this idea, we define the *simple action-level regret* and *simple policy-level regret* of policy π, up to time-step T as

$$\mathcal{SAR}_T(\pi) := d(\mathbf{q}_T, p), \quad \mathcal{SPR}_T(\pi) := d(\hat{\pi}_T, p).$$

We use superscript to specify the type of distance, e.g., we use $\mathcal{SAR}_T^{\text{TV}}$ to denote simple action-level total variation regret.

In the bandit setting, selecting suboptimal arms at any time-step other than the last does not contribute to simple regret. Hence algorithms that are designed to minimize simple regret are free to explore as much as possible and do not need to consider the exploration-exploitation trade-off. This is why optimizing for simple regret in bandit settings are sometimes referred to as pure exploration problems. This is in contrast with the problem of minimizing cumulative regret which takes the entire history of actions into account.

In our setting, minimizing simple policy-level regret could be considered a pure exploration problem. For example, consider the algorithm that samples each arm $(T-1)/k$ times deterministically and then samples the last arm based on the sample mean of the observations. Such an algorithm would have low simple policy-level regret at time T but not earlier.

On the other hand, minimizing simple action-level regret is not pure exploration problem as it takes the history of actions into account. However, it is still mostly focused on the end result and not the process. For example, if there are only two arms, $p = (1/2, 1/2)$ and $T = 1000$, the algorithm that select the first arm for the first 500 steps and the second arm for the remaining steps achieves zero simple action-level regret.

In order to formalize and measure the performance of algorithms its history, we define *cumulative regret* as the sum of simple regret over all time-steps.

$$\mathcal{CAR}_T(\pi) := \sum_{t=1}^{T} d(\mathbf{q}_t, p), \quad \mathcal{CPR}_T(\pi) := \sum_{t=1}^{T} d(\hat{\pi}_t, p).$$

Remark 1. When considering forward-KL regret, one issue is that if the output of the performance function does not have full support, then the regret will be infinity. This is particularly problematic if we consider policy-level forward-KL regret. If at any time-step, the action taken by the policy is deterministic or it is stochastic but not supported on all arms, then the simple regret at that time-step and the cumulative regret for any of the following time-steps is infinite. Thus, in Algorithm 1, we only consider simple policy-level forward-KL regret at times $t > Mk$ and we define $\mathcal{CPR}_T^{\text{f-KL}}(\pi) := \sum_{t=Mk+1}^{T} d^{\text{f-KL}}(\hat{\pi}_t, p)$. Note that Lu et al. (2018) which considered this notion of regret also skips the first k time-steps in their algorithm, i.e., DAISEE, when computing their regret.

Given all these possible notions of regret, it is natural to ask if one is preferable to others. The above remark suggests not using policy-level forward-KL regret. We postpone a complete answer to this question until Section 7.

4 RELATION BETWEEN DIFFERENT NOTIONS OF REGRET

Statistical distance We start by comparing the effect of statistical distance over different notions of regret. The core idea here is to use Pinsker's inequality

and inverse Pinsker's inequality to relate total variation and KL-divergences.

Theorem 1. *For any performance distribution f, we have*

$$2\left(\mathcal{R}^{f,d^{\mathrm{TV}}}\right)^2 \le \mathcal{R}^{f,d^{\text{r-KL}}} \le \frac{2}{\alpha}\left(\mathcal{R}^{f,d^{\mathrm{TV}}}\right)^2,$$
$$2\left(\mathcal{R}^{f,d^{\mathrm{TV}}}\right)^2 \le \mathcal{R}^{f,d^{\text{f-KL}}} \le \frac{2}{\alpha'}\left(\mathcal{R}^{f,d^{\mathrm{TV}}}\right)^2,$$

where $\alpha = \min_i p_i$ and $\alpha' = \min_i f_i$. Moreover, we also have

$$2\mathbb{E}\left[\mathcal{R}^{f,d^{\mathrm{TV}}}\right]^2 \le \min\left\{\mathbb{E}\left[\mathcal{R}^{f,d^{\text{r-KL}}}\right], \mathbb{E}\left[\mathcal{R}^{f,d^{\text{f-KL}}}\right]\right\}$$

See Appendix E for the proof.

Regret vs environment regret In Section 3.2, we mentioned that environment regret is weaker than regret. This claim is formalized in the following theorem.

Theorem 2. *If the map $q \mapsto d(q,p)$ is convex, then we have*

$$\mathbb{E}\left[\mathcal{PR}\right] \ge \mathbb{E}\left[\mathcal{PR}^{\mathrm{env}}\right], \quad \mathbb{E}\left[\mathcal{AR}\right] \ge \mathbb{E}\left[\mathcal{AR}^{\mathrm{env}}\right],$$

for both simple and cumulative regret. In particular, this statement holds for $d \in \{d^{\mathrm{TV}}, d^{\text{r-KL}}, d^{\text{f-KL}}\}$.

Proof. Since $q \mapsto d(q,p)$ is convex, we may use Jensen's inequality to see that

$$\mathbb{E}_{\omega^\nu,\omega^\pi}\left[d\left(\hat{\pi}_t, p\right)\right] \ge \mathbb{E}_{\omega^\nu}\left[d\left(\mathbb{E}_{\omega^\pi}\left[\hat{\pi}_t\right], p\right)\right],$$

for any $1 \le t \le T$. The claim follows from the definition of the notions of regret mentioned. The proof for action-level regret is obtained similarly by replacing $\hat{\pi}_t$ with $\mathbf{q}_t$. □

Action-level vs policy-level regret Finally, we consider the relation between action-level vs policy-level regrets. As mentioned in Section 3.2, a deterministic algorithm may have low action-level regret but high policy-level regret. Thus, in general, we can not bound policy-level regret using action-level regret. Instead, we try to bound action-level regret by policy-level regret. Specifically, we show that that simple action-level regret is bounded by cumulative policy-level regret up to an additive term. We will use this idea in the proof of Theorem 5 to bound the action-level regret of Algorithm 1.

Theorem 3. *If the map $q \mapsto d(q,p)$ is convex, then we have*

$$\mathcal{SAR}_T - \frac{1}{T}\mathcal{CPR}_T \le d\left(\mathbf{q}_T, p\right) - d\left(\frac{1}{T}\left(\sum_{t=1}^{T}\hat{\pi}_t\right), p\right).$$

Proof. Since d is convex in its first argument, we have

$$\mathcal{SAR}_T - \frac{1}{T}\mathcal{CPR}_T = d\left(\mathbf{q}_T, p\right) - \frac{1}{T}\sum_{t=1}^{T} d\left(\hat{\pi}_t, p\right)$$
$$\le d\left(\mathbf{q}_T, p\right) - d\left(\frac{1}{T}\left(\sum_{t=1}^{T}\hat{\pi}_t\right), p\right).$$ □

Remark 2. Note that the expression in the right-hand side is not necessarily negligable. For example, in the case of a two-armed sampling problem with target distribution $(1/2, 1/2)$ and a policy that selects uniformly randomly at each time step, all types of policy-level regret are zero. However, simple action-level total-variation regret would be the difference between the average number of times the first arm is selected and $1/2$, which has a variance of $O(T^{-1/2})$, thus contributing $O(T^{-1/2})$ to the expected regret.

5 LOWER BOUNDS

In this section, we discuss lower bounds for regret bounds of multi-armed sampling problems.

First, we define a class of performance functions that do not have an unreasonable amount of access to the environment. Specifically, we say a performance function is *environment agnostic* if it can be written as a deterministic function of $(\pi, \omega^\pi, \omega^\nu)$. Such a performance function has no direct access to the distribution of the environment and instead only depends on the realization of it, i.e., ω^ν. As such, if two environments are sufficiently similar so that they produce similar realizations with high probability, then such a performance function would have a difficult time distinguishing between the environments. This is the core idea we use to obtain lower bounds. Note that *all performance functions considered in this paper are environment-agnostic.*

Next, we state the main lemma which provides a lower bound on the sum of the regret of a single policy in two different environments based on the difference between the environments. To put it simply, if the environments are sufficiently different, then a policy that performs well in one environment can not perform well in the other. See Appendix F for the proof.

Lemma 1. *Let ν and ν' denote multi-armed sampling problems with reward functions $r : [k] \to \mathbb{R}$ and $r' : [k] \to \mathbb{R}$ and unit normal noise and let $\epsilon := \max_i |r_i - r'_i|$, $p := \mathrm{softmax}(r)$ and $p' := \mathrm{softmax}(r')$. Then, for any policy π, statistical distance d, and environment agnostic performance function f as defined above, we have*

$$\mathbb{E}\left[\mathcal{R}^{f,d}(\pi,\nu)\right] + \mathbb{E}\left[\mathcal{R}^{f,d}(\pi,\nu')\right]$$
$$\ge \frac{1}{2}\exp\left(-\frac{1}{2}T\epsilon^2 - \epsilon\sqrt{2T}\right) d'(p,p'),$$

where $d'(p,p') := \inf_q d(q,p) + d(q,p')$.

Using the above lemma, we obtain lower bounds for all forms of regret discussed earlier.

Theorem 4. *Let $k > 1$ and $T > 4$ be the number of arms and the horizon respectively, and let f be any environment-agnostic performance function as defined above. Then, for any policy π, we have*

$$\sup_{\nu \in \mathcal{E}^k_{\mathrm{SG}}(1)} \mathbb{E}\left[\mathcal{R}^{f,d^{\mathrm{TV}}}(\pi,\nu)\right] \geq \frac{1}{126\sqrt{Tk}},$$

$$\sup_{\nu \in \mathcal{E}^k_{\mathrm{SG}}(1)} \mathbb{E}\left[\mathcal{R}^{f,d}(\pi,\nu)\right] \geq \frac{1}{567Tk},$$

for $d \in \{d^{\text{r-KL}}, d^{\text{f-KL}}\}$. Thus, these lower bounds apply to all notions of simple regret considered in this work. Moreover, by multiplying the bounds by T, we obtain corresponding lower bounds for cumulative regret.

See Appendix G for the proof.

6 AN OPTIMAL ALGORITHM

In this section, we introduce the Greedy Active Sampling with Warm-up (GASW) algorithm, which is a sampling version of the classical Explore-Then-Commit algorithm in multi-armed bandit problems. In the case where there is no exploration, i.e. exploration factor M is equal to 1, *the algorithm becomes greedy* and we simply refer to it as Greedy Active Sampling (GAS). As seen in Algorithm 1, GAS performs random exploration for MK steps, and then it samples each arm according to its reward estimate. The algorithm then continuously updates these estimates during sampling.

Algorithm 1: Greedy Active Sampling (with Warm-up) (GAS / GASW)

Input : Horizon T, number of arms k, and exploration factor $1 \leq M \leq T/k$

```
for t = 1 to T do
    if t ≤ Mk then
        Play a_t ← (t mod k) + 1
    else
        Sample 1 ≤ a_t ≤ k according to p̂_t and play a_t
    end
    for 1 ≤ i ≤ k do
        r̂_{t,i} ← (1/n_{t,i}) Σ_{t'=1}^{t} x_{t'} 1_{a_{t'}=i}
    end
    p̂_{t+1} ← softmax(r̂_t)
end
```

The following theorem demonstrates that, unlike bandits, there is barely any need for exploration in multi-armed sampling problems. Specifically, we see that with a negligible exploration of $O(\log(T))$, we obtain near-optimal regret bounds that differ from our lower bounds up to logarithmic terms. Moreover, if there is no exploration at all, then the regret bounds would be multiplied by a factor that grows faster than $\log(T)$, but slower than T^{ε}, for any $\varepsilon > 0$. Note that with the same amount of exploration, the Explore-Then-Commit algorithms suffer a linear regret in multi-armed bandit problems.

In the following theorem, we use the notation $\hat{\mathcal{O}}^{1/2}$ to hide $\frac{1}{2}$-quasi-polynomial terms. [1]

Theorem 5. *In Algorithm 1, if $\nu \in \mathcal{E}^k_{\mathrm{SG}}(\sigma^2)$ and $M \geq 36\sigma^2\log(T)$ and $d = d^{\mathrm{TV}}$, then the regrets $\mathcal{SAR}^d_T$, $\mathcal{SAR}^{d,\mathrm{env}}_T$, $\mathcal{SPR}^d_T$, and $\mathcal{SPR}^{d,\mathrm{env}}_T$ are all bounded by $\tilde{\mathcal{O}}(T^{-1/2})$ while the corresponding cumulative regrets are bounded by $\tilde{\mathcal{O}}(T^{1/2})$. On the other hand, if $d = d^{\text{r-KL}}$ or $d = d^{\text{f-KL}}$ and we also have $\nu \in \mathcal{E}^k_{\mathrm{B}}(\sigma)$, then the simple regrets are bounded by $\tilde{\mathcal{O}}(T^{-1})$ while the corresponding cumulative regrets are $\tilde{\mathcal{O}}(1)$.*
Moreover, if we have no exploration, i.e., $M = 1$, then the same bounds hold if we replace $\tilde{\mathcal{O}}$ with $\hat{\mathcal{O}}^{1/2}$.

In fact, a more general result holds.

Theorem 6. *Let $C \geq 0$ be a hyperparameter and let $\mathcal{A}$ be an algorithm that behaves similar to Algorithm 1, with the only difference being that $\mathcal{A}$ selects $\hat{\mathbf{r}}_{t,i}$ such that $|\hat{\mathbf{r}}_{t,i} - \bar{\mathbf{x}}_{t,i}| \leq 2C\sigma\sqrt{\frac{\log(T)}{\mathbf{n}_{t,i}}}$, where $\bar{\mathbf{x}}_{t,i} := \frac{1}{\mathbf{n}_{t,i}}\sum_{t'=1}^{t}\mathbf{x}_{t'}\mathbf{1}_{\mathbf{a}_{t'}=i}$. Then regret bounds similar to those of Theorem 5 hold for $\mathcal{A}$ as well.*

See Appendix H for the exact expressions of the regret bounds and the proof for both theorems.

Note that, when $C = 0$, the algorithm becomes greedy (with warm-up) and Theorem 6 reduces to Theorem 5. On the other hand, Theorem 6 covers UCB-type exploration (as in KL-UCB Zhao et al. (2025b)) by choosing $\hat{\mathbf{r}}_{t,i} = \bar{\mathbf{x}}_{t,i} + 2C\sigma\sqrt{\frac{\log(T)}{\mathbf{n}_{t,i}}}$. Similarly, this result also covers pessimism, when we choose $\hat{\mathbf{r}}_{t,i} = \bar{\mathbf{x}}_{t,i} - 2C\sigma\sqrt{\frac{\log(T)}{\mathbf{n}_{t,i}}}$. One approach to obtain regret bounds for Thompson Sampling (TS) is to use the well-known connection between TS and UCB to convert the regret bounds of UCB into regret bounds for TS. (Russo and Van Roy, 2014; Russo et al., 2018). Thus, a similar bound also applies to many forms of Thompson sampling algorithms.

These regret bounds clearly demonstrate that the nature of exploration-exploitation tradeoff is different in MAB and MAS. Specifically, when using optimism, Thompson sampling, or even pessimism, we still obtain near optimal regret bounds.

[1] Specifically, we say $f(T) = \hat{\mathcal{O}}^{1/2}(g(T))$ if $f(T) = \mathcal{O}(g(T)h(T))$ for some h that satisfies $h(T) = \exp\left(\mathcal{O}\left(\log(T)^{1/2}\right)\right)$, which implies that $h(T) = \mathcal{O}(T^{\varepsilon})$ for all $\varepsilon > 0$.

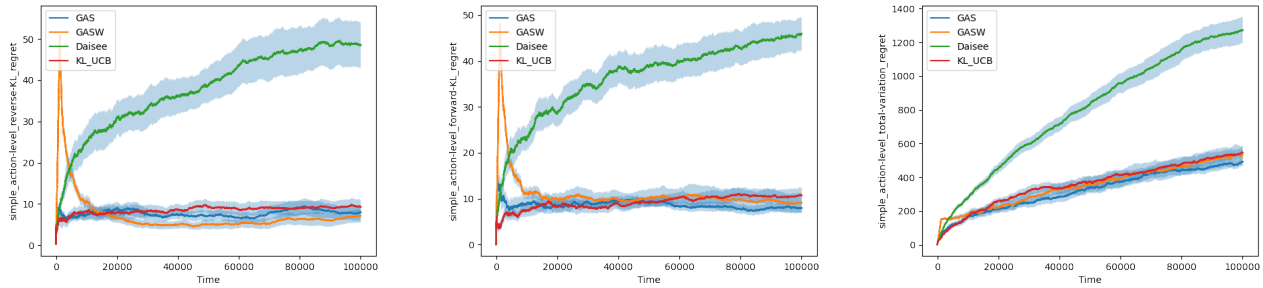

Figure 1: Comparison between regret, specifically $T.\mathcal{SAR}_T$, of GAS, GSAW with logarithmic exploration, DAISEE, and KL-UCB with different statistical distances. From left to right, the distances are reverse-KL, forward-KL and total variation. Experiments for other notions are regret are included in the appendix. The initial increase in regret in the orange lines corresponds to the exploration phase (and its effect on action history). Environments have unit normal noise and the reward are sampled from $[0, 1]$. The average over 10 runs are shown. See Appendix K for details.

7 FROM SAMPLING TO OPTIMIZATION

Consider the multi-armed sampling problem with arm distribution ν as defined in Section 2. We define $\mathcal{R}_T^{\text{MAB}} := T \max_{1\le i\le k} r_i - \sum_{i=1}^k \mathbf{n}_{T,i} r_i$, i.e., the regret if the reward are distributed according to ν

Next we let $\beta \in [0, \infty]$ and define $p^\beta := \text{softmax}(\beta r)$ for $\beta < \infty$ and $p^\infty := \lim_{\beta\to\infty} p^\beta$ where the limit is taken as vectors in $\mathbb{R}^k$. β is often referred to as the inverse-temperature. It follows that p^∞ is simply the distribution that assigns the same weight to all optimal arms $\text{argmax}_{1\le i\le k} r_i$ while assigning zero weight to any other arm. In other words, sampling according to p^∞ is the same as maximizing reward. Thus, it is natural to expect that, at the limit of $\beta \to \infty$, the multi-armed sampling problem becomes similar to the multi-armed bandit problem. The following theorem makes this connection precise for the notion of reverse-KL regret.

First we define the *multi-armed sampling action-level β-regret, simple policy level β-regret and policy level β-regret*, as [2]

$$\mathcal{AR}_T^\beta := \begin{cases} \frac{T}{\beta} d^{\text{r-KL}}(\mathbf{q}_T, p^\beta) & \text{if } \beta < \infty, \\ \mathcal{R}_T^{\text{MAB}} & \text{if } \beta = \infty. \end{cases}, \tag{1}$$

$$\mathcal{SPR}_T^\beta := \begin{cases} \frac{T}{\beta} d^{\text{r-KL}}(\hat{\pi}_T, p^\beta) & \text{if } \beta < \infty, \\ \mathcal{R}_T^{\text{MAB}} & \text{if } \beta = \infty. \end{cases} \tag{2}$$

and $\mathcal{PR}_T^\beta := \sum_{t=1}^T \mathcal{SPR}_t^\beta$, respectively.

[2]Here we do not consider the cumulative version of action-level regret, since the simple version already captures the main ideas.

Theorem 7. *We have*

$$\mathcal{AR}_T^\beta = \mathcal{R}_T^{\text{MAB}} - \frac{T}{\beta} H(\mathbf{q}_T) - \frac{T}{\beta}\log(p_*^\beta),$$

$$\mathcal{PR}_T^\beta = \sum_{t=1}^T \mathbb{E}\left[\mathcal{SR}_t^{\text{MAB}} - \frac{1}{\beta} H(\hat{\pi}_t)\right] - \frac{T}{\beta}\log(p_*^\beta),$$

$$\mathcal{SPR}_T^\beta = \mathbb{E}\left[\mathcal{SR}_t^{\text{MAB}} - \frac{1}{\beta} H(\hat{\pi}_t)\right] - \frac{1}{\beta}\log(p_*^\beta).$$

In particular, as $\beta \to \infty$, both $\mathbb{E}\left[\mathcal{AR}_T^\beta\right]$ and $\mathbb{E}\left[\mathcal{PR}_T^\beta\right]$ tend to $\mathbb{E}\left[\mathcal{R}_T^{\text{MAB}}\right]$ and $\mathbb{E}\left[\mathcal{SPR}_T^\beta\right]$ tends to $\mathbb{E}\left[\mathcal{SR}_t^{\text{MAB}}\right]$.

See Appendix I for the proof.

Note that, for any fixed problem and β, the value of p_*^β is constant. Therefore, this theorem demonstrates that the notion of policy-level β-regret defined here is equivalent to entropy-regularized bandits, in the sense that is commonly considered in the literature. Also note that, in the case of $\beta = 1$, we have $\mathcal{AR}_T^\beta = T.\mathcal{SAR}_T^{\text{r-KL}}$ and $\mathcal{SPR}_T^\beta = \mathcal{SPR}_T^{\text{r-KL}}$. Thus the above theorem also demonstrates that, as $\beta \to \infty$, these notions of regret in sampling converge and the limit is the multi-armed bandit regret. It is not surprising that, for any finite β, minimizing β-regret is not fundamentally different from minimizing 1-regret. As we will discuss in Appendix J, for Algorithm 1, replacing the step $\hat{p}_{t+1} \leftarrow \text{softmax}(\hat{r}_t)$ with $\hat{p}_{t+1} \leftarrow \text{softmax}(\beta\hat{r}_t)$ results in an algorithm that achieves low β-regret. We refer to Appendix J for more details.

Acknowledgments

This research is in part supported by CIFAR AI Chairs program, NSERC Discovery and IVADO IAR3.

References

Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, pages 39.1–39.26. PMLR.

Akhound-Sadegh, T., Rector-Brooks, J., Bose, A. J., Mittal, S., Lemos, P., Liu, C.-H., Sendera, M., Ravanbakhsh, S., Gidel, G., Bengio, Y., Malkin, N., and Tong, A. (2024). Iterated denoising energy matching for sampling from Boltzmann densities. *International Conference on Machine Learning (ICML)*.

Albergo, M. S., Kanwar, G., and Shanahan, P. E. (2019). Flow-based generative models for Markov chain Monte Carlo in lattice field theory. *Physical Review D*, 100(3):034515.

Auer, P. (2003). Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3:397–422.

Azar, M. G., Guo, Z. D., Piot, B., Munos, R., Rowland, M., Valko, M., and Calandriello, D. (2024). A general theoretical paradigm to understand learning from human preferences. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback.

Bengio, E., Jain, M., Korablyov, M., Precup, D., and Bengio, Y. (2021a). Flow network based generative models for non-iterative diverse candidate generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27381–27394.

Bengio, E., Jain, M., Korablyov, M., Precup, D., and Bengio, Y. (2021b). Flow network based generative models for non-iterative diverse candidate generation. *Advances in Neural Information Processing Systems*, 34:27381–27394.

Bengio, Y., Lahlou, S., Deleu, T., Hu, E. J., Tiwari, M., and Bengio, E. (2023). Gflownet foundations. *The Journal of Machine Learning Research*, 24(1):10006–10060.

Berry, D. A. and Fristedt, B. (1985). Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall*, 5(71-87):7–7.

Cappé, O., Douc, R., Guillin, A., Marin, J.-M., and Robert, C. P. (2008). Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18:447–459.

Duane, S., Kennedy, A., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222.

Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., and Lee, K. (2023). DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885.

Fox, R., Pakman, A., and Tishby, N. (2015). Taming the noise in reinforcement learning via soft updates. *arXiv preprint arXiv:1512.08562*.

Gabrié, M., Rotskoff, G. M., and Vanden-Eijnden, E. (2021). Efficient Bayesian sampling using normalizing flows to assist Markov chain Monte Carlo methods. *arXiv preprint arXiv:2107.08001*.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. (2017). Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pages 1352–1361. PMLR.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Hoffman, M. D., Gelman, A., et al. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research (JMLR)*, 15(1):1593–1623.

Jain, M., Bengio, E., Hernandez-Garcia, A., Rector-Brooks, J., Dossou, B. F. P., Ekbote, C. A., Fu, J., Zhang, T., Kilgour, M., Zhang, D., Simine, L., Das, P., and Bengio, Y. (2022). Biological sequence design with GFlowNets. In *Proceedings of the 39th International Conference on Machine Learning*, pages 9786–9801. PMLR.

Jain, M., Deleu, T., Hartford, J., Liu, C.-H., Hernandez-Garcia, A., and Bengio, Y. (2023a). GFlowNets for AI-driven scientific discovery. *Digital Discovery*, 2(3):557–577.

Jain, M., Raparthy, S. C., Hernandez-Garcia, A., Rector-Brooks, J., Bengio, Y., Miret, S., and Bengio, E. (2023b). Multi-objective GFlowNets. In *Proceedings of the 40th International Conference on Machine Learning*, pages 14631–14653. PMLR.

Jaques, N., Ghandeharioun, A., Shen, J. H., Ferguson, C., Lapedriza, A., Jones, N., Gu, S., and Picard, R. (2019). Way off-policy batch deep reinforcement learning of implicit human preferences in dialog.

Kim, H., Kim, M., Choi, S., and Park, J. (2024). Genetic-guided GFlowNets for sample efficient molecular optimization. *Advances in Neural Information Processing Systems*, 37:42618–42648.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Leimkuhler, B. J., Matthews, C., and Tretyakov, M. V. (2014). On the long-time integration of stochastic gradient systems. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 470.

Lemos, P., Malkin, N., Handley, W., Bengio, Y., Hezaveh, Y., and Perreault-Levasseur, L. (2023). Improving gradient-guided nested sampling for posterior inference. *International Conference on Machine Learning (ICML)*.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Lu, X., Rainforth, T., Zhou, Y., van de Meent, J.-W., and Teh, Y. W. (2018). On exploration, exploitation and learning in adaptive importance sampling. *arXiv preprint arXiv:1810.13296*.

Malkin, N., Jain, M., Bengio, E., Sun, C., and Bengio, Y. (2022). Trajectory balance: Improved credit assignment in GFlowNets. *Advances in Neural Information Processing Systems*, 35:5955–5967.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Midgley, L. I., Stimper, V., Simm, G. N., Schölkopf, B., and Hernández-Lobato, J. M. (2022). Flow annealed importance sampling bootstrap. *arXiv preprint arXiv:2208.01893*.

Noé, F., Olsson, S., Köhler, J., and Wu, H. (2019). Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Rector-Brooks, J., Madan, K., Jain, M., Korablyov, M., Liu, C.-H., Chandar, S., Malkin, N., and Bengio, Y. (2023). Thompson sampling for improved exploration in gflownets. *arXiv preprint arXiv:2306.17693*.

Robbins, H. E. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363.

Russo, D. and Van Roy, B. (2014). Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243.

Russo, D. J., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. (2018). A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96.

Sason, I. and Verdú, S. (2015). Upper bounds on the relative entropy and rényi divergence as a function of total variation distance for finite alphabets. In *2015 IEEE Information Theory Workshop - Fall (ITW)*, pages 214–218.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. (2020). Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Sutton, R. S. (2018). Reinforcement learning: An introduction. *A Bradford Book*.

Thompson, W. R. (1933). On the likelihood hat one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294.

Tiapkin, D., Morozov, N., Naumov, A., and Vetrov, D. P. (2024). Generative flow networks as entropy-regularized rl. In *International Conference on Artificial Intelligence and Statistics*, pages 4213–4221. PMLR.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V.,

Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models.

Vargas, F., Grathwohl, W., and Doucet, A. (2023). Denoising diffusion samplers. *arXiv preprint arXiv:2302.13834*.

Xiong, W., Dong, H., Ye, C., Wang, Z., Zhong, H., Ji, H., Jiang, N., and Zhang, T. (2024). Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *Proceedings of the 41st International Conference on Machine Learning*, pages 54715–54754. PMLR.

Zhang, Q. and Chen, Y. (2021). Path integral sampler: a stochastic control approach for sampling. *arXiv preprint arXiv:2111.15141*.

Zhao, H., Ye, C., Gu, Q., and Zhang, T. (2025a). Sharp analysis for KL-regularized contextual bandits and RLHF.

Zhao, H., Ye, C., Xiong, W., Gu, Q., and Zhang, T. (2025b). Logarithmic regret for online KL-regularized reinforcement learning.

Ziebart, B. D. (2010). Modeling purposeful adaptive behavior with the principle of maximum causal entropy. *PhD thesis, Carnegie Mellon University*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes. The description in Appendix K is enough to reproduce the toy experiments in this paper.]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A RELATED WORKS

We discussed some of the related works, specifically those related to the fine-tuning of pretrained models in the introduction. In the following, we discuss other related works.

A.1 Multi-armed bandits

Originating from the work of (Robbins, 1952), the multi-armed bandit problem captures the fundamental trade-off between exploration—gathering information about the potential rewards of each arm—and exploitation—leveraging the current knowledge to maximize immediate rewards. Over the decades, numerous algorithms have been developed to effectively address this problem. Notable among these are the *Upper Confidence Bound (UCB)* algorithms introduced by (Auer, 2003), which select arms based on optimistic estimates of their potential rewards, and *Thompson Sampling* (Thompson, 1933; Agrawal and Goyal, 2012), a Bayesian approach that balances exploration and exploitation by sampling from the posterior distributions of each arm's rewards. Another classical algorithm is *Explore-Then-Commit (ETC)* algorithm, which is based on 'certainty equivalence with forcing' algorithm of (Robbins, 1952). ETC initially explores all arms for a predetermined number of rounds before committing to the arm with the highest estimated reward, and it has been widely studied for its simplicity and effectiveness (Lattimore and Szepesvári, 2020). The distinct separation of exploration and exploitation makes ETC an excellent candidate for us to adapt to the multi-armed sampling problem, as it enables the study of the exploration-exploitation trade-off in one of the most simple scenarios. Note that ETC is a suboptimal algorithm that requires $\Omega(T^{2/3})$ exploration to guarantee regret of $O(T^{2/3})$ while the optimal minmax regret bound is $O(T^{1/2})$.

A.2 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) methods are a class of algorithms widely employed for sampling from complex probability distributions (Metropolis et al., 1953; Hastings, 1970; Geman and Geman, 1984). MCMC approaches construct a Markov chain whose stationary distribution matches the target distribution, enabling the generation of representative samples through iterative stochastic processes.

Convergence of Markov chains is typically measured based on their mixing time. The fundamental theorem of Markov chains states that, under suitable conditions, the total variation distance of the chain distribution to the target distribution converges exponentially; see Levin and Peres (2017). As discussed in Section 3.2, the chain distribution corresponds to $\mathbb{E}_{\omega^\pi}[\hat{\pi}_t]$ in our notation. Therefore the fundamental theorem of Markov provides bounds for $\mathcal{SPR}^{\text{env},\text{TV}}$. At first glance, this seems to contradict the lower bound in Theorem 4. However, Theorem 4 can only be applied when the environment is stochastic. In a deterministic case, typically assumed by MCMC methods, we only need to sample each arm once to find the exact target distribution. Thus, after sampling each of the k arms once, we may simply sample according to the exact target distribution which results in $\mathcal{SPR}_t^{\text{env},d} = 0$, for any statistical distance d and all $t \geq k+1$.

A.3 Adaptive importance sampling and DAISEE

Adaptive Importance Sampling is a Monte Carlo technique in which proposal distributions used for sampling are iteratively updated based on the result of the evaluation of previously sampled points Cappé et al. (2008).

Closely related to our work is (Lu et al., 2018) which frames AIS as an online learning algorithm. Their formulation could be seen as a form of multi-armed sampling problem with cumulative policy-level forward-KL regret. They propose an algorithm, namely DAISEE, which uses the *optimism principle* as an explicit exploration mechanism, similar to that of UCB in multi-armed bandit problems. They obtain a regret bound of $\tilde{O}(T^{1/2})$. In contrast, for the same notion of regret, we obtain a regret bound of $\tilde{O}(1)$ with minor exploration and $\hat{O}^{1/2}(1)$ with no exploration. It is important to note that the exploration in DAISEE is done by adding the confidence bound to unnormalized probabilities. This is in contrast with our algorithm and also KL-UCB (Zhao et al., 2025b) where the the optimism term is added to unnormalized *log*-probabilities. Figure 1 compares GAS, GASW with logarithmic exploration, and KL-UCB with DAISEE, where other algorithms seem to outperform DAISEE.

A.4 Neural Samplers

Neural networks are increasingly used in the context of sampling through diverse means. Some are inspired by ideas from deep generative models, such as diffusion (e.g., Zhang and Chen, 2021; Vargas et al., 2023) and normalizing flows (e.g., Noé et al., 2019; Midgley et al., 2022), and some are inspired by ideas from RL, including generative flow networks (Bengio et al., 2021b, 2023) and entropy-regularized RL (Fox et al., 2015; Haarnoja et al., 2017) (and its applications in fine-tuning of pretrained models as we have discussed in the introduction); see also (Tiapkin et al., 2024). The fact that exploration does not seem to become an issue in these sampling counterparts to RL is consistent with our findings. An exception is Rector-Brooks et al. (2023) which uses Thompson sampling to encourage exploration for sampling in generative flow networks.

B THE NECESSITY OF NOISE

Given that we are working in the finite-arm setting, it is necessary to assume that observations are noisy. Indeed, in both MAS and MAB, if observations were noise-free, we could simply try each arm once to determine its true reward, eliminating any exploration–exploitation tradeoff.

An alternative way to interpret noise is through the lens of approximating a continuous (or otherwise infinite) arm space with a finite one. Specifically, we can partition the arms of the original continuous problem into a finite number of super-arms and define a new problem over these aggregated arms. To complete this formulation, each super-arm must be assigned a reward. If we denote a super-arm by $\mathbf{x}$, corresponding to a subset of arms $a \in A_{\mathbf{x}}$, a natural approach is to define a probability distribution $\rho_{\mathbf{x}}$ over $A_{\mathbf{x}}$ and set $r(\mathbf{x}) = r(a)$, where a is drawn according to $\rho_{\mathbf{x}}$.

This approach is commonly adopted in the reinforcement learning and sampling literature. For instance, many algorithms for Lipschitz bandit problems rely on this idea, and Lu et al. (2018) apply it to extend their MAS algorithm (DAISEE) to continuous settings. Consequently, even if the reward function in the original continuous problem is deterministic, it remains reasonable to model noise in the corresponding finite-arm version. Introducing stochasticity in this way helps preserve the essential exploration–exploitation dynamics of the original problem, ensuring that analysis in the finite setting provides meaningful insights into exploration–exploitation tradeoffs in more general, continuous environments.

C THE ROLE OF SOFTMAX

As discussed earlier, it is important in our setting that observations are noisy. Consider first the case where we observe noisy samples of the unnormalized probabilities of the target distribution. Since unnormalized probabilities are always non-negative, additive noise is not necessarily a meaningful modeling choice. A more appropriate alternative is to introduce *multiplicative* noise. Taking the logarithm of these noisy values then yields *unnormalized log-probabilities* with *additive* noise.

Recall that our goal is to estimate or match the target distribution. When working with unnormalized log-probabilities, i.e., $r_i = \log(p_i) + C$ for some constant $C \in \mathbb{R}$, applying the softmax function recovers the original distribution p. Equivalently, if we had instead modeled unnormalized probabilities with multiplicative noise, there would be no need to include the softmax transformation explicitly.

As noted in Appendix A.3, DAISEE performs exploration by adding a confidence bound directly to unnormalized probabilities. This contrasts with our algorithm, as well as KL-UCB, where the optimism term is added in the space of unnormalized *log*-probabilities. Both theoretical regret bounds and empirical results from toy experiments suggest that applying additive noise in the space of unnormalized probabilities is suboptimal.

Given that additive noise in the log-probability space is both mathematically convenient and more consistent with the structure of multi-armed bandit problems, we adopt the formulation where the observation at each time step corresponds to unnormalized log-probabilities with additive noise.

D USEFUL LEMMAS

Lemma 2. *Let $r = (r_i)_{i=1}^k$ and $r' = (r'_i)_{i=1}^k$ be two vectors in $\mathbb{R}^d$ and let $p = \mathrm{softmax}(r)$ and $p' = \mathrm{softmax}(r')$. Also let $\epsilon := \max_{1\le i\le k} |r_i - r'_i|$. Then we have*

$$p_i \exp(-2\epsilon) \le p'_i \le p_i \exp(2\epsilon) \quad \text{and} \quad |p_i - p'_i| \le 2\epsilon,$$

for all $1 \le i \le k$. As a consequence, we have $d^{\mathrm{TV}}(p, p') \le 2\epsilon$.

Proof. To prove the first claim, we note that

$$\begin{aligned}\log\left(\frac{p'_i}{p_i}\right) &= \log\left(\frac{\mathrm{softmax}(r')_i}{\mathrm{softmax}(r)_i}\right) = \log\left(\frac{\exp(r'_i)}{\exp(r_i)}\right) + \log\left(\frac{\sum_j \exp(r_j)}{\sum_j \exp(r'_j)}\right)\\ &\le \log\left(\frac{\exp(r_i+\epsilon)}{\exp(r_i)}\right) + \log\left(\frac{\sum_j \exp(r'_j+\epsilon)}{\sum_j \exp(r'_j)}\right) = \epsilon + \epsilon = 2\epsilon.\end{aligned}$$

The inequality in the other direction is proven similarly.

Using the above bounds, we see that

$$p'_i - p_i \le p'_i - p'_i \exp(-2\epsilon) = p'_i(1 - \exp(-2\epsilon)) \le 2\epsilon p'_i \le 2\epsilon,$$

where the last inequality follows from the fact that $1 - \exp(-x) \le x$ for all $x \in \mathbb{R}$. Similarly we see that $p'_i - p_i \ge -2\epsilon$.

To complete the proof of the lemma, we note that

$$d^{\mathrm{TV}}(p, p') = \sum_{i=1}^k \max\{0, p'_i - p_i\} \le \sum_{i=1}^k 2\epsilon p'_i = 2\epsilon. \qquad \square$$

Lemma 3. *Let d be a statistical distance and define $d'(p, p') := \inf_q d(q, p) + d(q, p')$. Then the following are true.*

(i) If d is a metric, e.g. d^{TV}, then $d' = d$.

(ii) If $d(p, p') = D_{\mathrm{KL}}(p||p')$ or $d(p, p') = D_{\mathrm{KL}}(p'||p)$, then we have $d'(p, p') \ge d^{\mathrm{TV}}(p, p')^2$.

Proof. **(i)** We note that

$$d'(p, p') = \inf_q d(q, p) + d(q, p') \le d(p, p) + d(p, p') = d(p, p').$$

On the other hand, using the triangle inequality, we have

$$d'(p, p') = \inf_q d(q, p) + d(q, p') \ge \inf_q d(p, p') = d(p, p').$$

.

(ii) Assume $d(p, p') = D_{\mathrm{KL}}(p||p')$. Using Pinsker's inequality, we have

$$\begin{aligned}d(q, p) + d(q, p') &\ge 2\left(d^{\mathrm{TV}}(q, p)^2 + d^{\mathrm{TV}}(q, p')^2\right)\\ &\ge \left(d^{\mathrm{TV}}(q, p) + d^{\mathrm{TV}}(q, p')\right)^2 \ge d^{\mathrm{TV}}(p, p')^2.\end{aligned}$$

The same argument applies to the case $d(p, p') = D_{\mathrm{KL}}(p'||p)$. $\square$

E PROOF OF THEOREM 1

Proof. We write the proof for $d^{\text{r-KL}}$. The proof for $d^{\text{f-KL}}$ is similar. Using Pinsker's inequality, we have

$$\mathcal{R}^{f,d^{\text{r-KL}}} = D_{\text{KL}}(f\|p) \geq 2d^{\text{TV}}(f,p)^2 = 2\left(\mathcal{R}^{f,d^{\text{TV}}}\right)^2.$$

To prove the other side of the inequality for $d^{\text{r-KL}}$, we use inverse Pinsker's inequality[3]

$$\mathcal{R}^{f,d^{\text{r-KL}}} = D_{\text{KL}}(f\|p) \leq \frac{2}{\alpha} d^{\text{TV}}(f,p)^2 = \frac{2}{\alpha}\left(\mathcal{R}^{f,d^{\text{TV}}}\right)^2.$$

Finally, to prove the inequality with expected values, we note that

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}^{f,d^{\text{r-KL}}}\right] &= \mathbb{E}\left[D_{\text{KL}}(f\|p)\right] \geq \mathbb{E}\left[2d^{\text{TV}}(f,p)^2\right] \\ &\geq 2\mathbb{E}\left[d^{\text{TV}}(f,p)\right]^2 = 2\mathbb{E}\left[\mathcal{R}^{f,d^{\text{TV}}}\right]^2.\end{aligned}$$

□

F PROOF OF LEMMA 1

Proof. For a multi-armed sampling environment ν, let A_ν be the $T \times k$ matrix where the (t,i) element is r_i. We define a *noise matrix* for ν as a random matrix $\varepsilon = (\varepsilon_{t,i})_{1\leq t\leq T, 1\leq i\leq k} = \omega^\nu - A_\nu$. Given a policy π and a noise matrix ε, we may run the policy where the observation $\mathbf{x}_t$ at time-step t is given by $r_{\mathbf{a}_t} + \varepsilon_{t,\mathbf{a}_t}$. Hence a noise matrix ε captures all the randomness of the environment and we may use ε and ω^ν interchangeably. Note for any given realization of the noise matrix, $\mathbf{a}_t$ is still a random variable since the policy π is not necessarily deterministic.

Let $\mathbf{x}_t$ and $\mathbf{x}'_t$ denote the sequence of observations when running the policy π against ν and ν' with noise matrices ε and ε'. Also let $A_{\nu,\nu'} := A_\nu - A_{\nu'}$ and $\varepsilon' = A_{\nu,\nu'} + \varepsilon$. The key property of this matrix is that it if $\varepsilon' = A_{\nu,\nu'} + \varepsilon$, then we have

$$\mathbf{x}'_t = r'_{\mathbf{a}_t} + \varepsilon'_{t,\mathbf{a}_t} = r'_{\mathbf{a}_t} + [A_{\nu,\nu'}]_{t,\mathbf{a}_t} + \varepsilon_{t,\mathbf{a}_t} = r_{\mathbf{a}_t} + \varepsilon_{t,\mathbf{a}_t} = \mathbf{x}_t.$$

In other words, if $\varepsilon' = A_{\nu,\nu'} + \varepsilon$, then the sequence of actions and observations is identical between (ν,π,ε) and (ν',π,ε'), and therefore the value of the performance function f is the same between the two cases.

Let Φ denote the probability density function of $N(0, I_{Tk})$, the Tk-dimensional unit normal distribution and let

$$g(\pi,\nu,\varepsilon) := \mathbb{E}_{\omega^\pi}\left[d(f(\pi,\omega^\pi,\varepsilon + A_\nu),p)\right].$$

Hence we have

$$\mathcal{R}^{f,d}(\pi,\nu) = \mathbb{E}_\varepsilon[g(\pi,\nu,\varepsilon)] = \int g(\pi,\nu,\varepsilon)\Phi(\varepsilon)d\varepsilon.$$

If $\varepsilon' = A_{\nu,\nu'} + \varepsilon$, then using the fact that $|r_i - r'_i| \leq \epsilon$, we have

$$\begin{aligned}\frac{\Phi(\varepsilon)}{\Phi(\varepsilon')} &= \prod_{i=1}^k \prod_{t=1}^T \frac{\Phi(\varepsilon_{t,i})}{\Phi(\varepsilon'_{t,i})} \\ &= \prod_{i=1}^k \prod_{t=1}^T \exp\left(-\frac{1}{2}\varepsilon_{t,i}^2 + \frac{1}{2}(\varepsilon'_{t,i})^2\right) \\ &= \prod_{i=1}^k \prod_{t=1}^T \exp\left(-\frac{1}{2}(\varepsilon'_{t,i} + r'_i - r_i)^2 + \frac{1}{2}(\varepsilon'_{t,i})^2\right) \\ &= \exp\left(-\frac{1}{2}T\sum_{i=1}^T (r'_i - r_i)^2 - \sum_{i=1}^k \sum_{t=1}^T \varepsilon'_{t,i}(r'_i - r_i)\right) \\ &\geq \exp\left(-\frac{1}{2}Tk\epsilon^2 - \epsilon\sum_{i=1}^k \sum_{t=1}^T \varepsilon'_{t,i}\right).\end{aligned}$$

[3] See Remark 1 in Sason and Verdú (2015)

Let $E' := \left\{\varepsilon' \mid |\sum_{i=1}^{k}\sum_{t=1}^{T}\varepsilon'_t| < \sqrt{2Tk}\right\}$. It follows that, for any $\varepsilon \in E'$, we have

$$\frac{\Phi(\varepsilon)}{\Phi(\varepsilon')} \geq \exp\left(-\frac{1}{2}Tk\epsilon^2 - \epsilon\sum_{i=1}^{k}\sum_{t=1}^{T}\varepsilon'_{t,i}\right) \geq \exp\left(-\frac{1}{2}Tk\epsilon^2 - \epsilon\sqrt{2Tk}\right).$$

Using the definition of d', we have

$$d(f,p) + d(f,p') \geq d'(p,p'),$$

where we use f to denote the output of the function f. Therefore, if $\varepsilon' = \varepsilon + A_{\nu,\nu'} \in E'$, then we have

$$\begin{aligned}
d(f,p)\Phi(\varepsilon) &\geq \max\{0, d'(p,p') - d(f,p')\}\,\Phi(\varepsilon) \\
&= \max\{0, d'(p,p') - d(f,p')\}\,\Phi(\varepsilon')\frac{\Phi(\varepsilon)}{\Phi(\varepsilon')} \\
&\geq \max\{0, d'(p,p') - d(f,p')\}\,\Phi(\varepsilon')\exp\left(-\frac{1}{2}Tk\epsilon^2 - \epsilon\sqrt{2Tk}\right) \\
&\geq (d'(p,p') - d(f,p'))\,\Phi(\varepsilon')\exp\left(-\frac{1}{2}Tk\epsilon^2 - \epsilon\sqrt{2Tk}\right),
\end{aligned}$$

which implies that

$$\begin{aligned}
g(\pi,\nu,\varepsilon)\Phi(\varepsilon) &\geq \mathbb{E}_{\omega^\pi}\left[(d'(p,p') - d(f,p'))\,\Phi(\varepsilon')\exp\left(-\frac{1}{2}Tk\epsilon^2 - \epsilon\sqrt{2Tk}\right)\right] \\
&= \exp\left(-\frac{1}{2}Tk\epsilon^2 - \epsilon\sqrt{2Tk}\right)\mathbb{E}_{\omega^\pi}\left[d'(p,p') - d(f,p')\right]\Phi(\varepsilon') \\
&= \exp\left(-\frac{1}{2}Tk\epsilon^2 - \epsilon\sqrt{2Tk}\right)(d'(p,p') - g(\pi,\nu',\varepsilon'))\,\Phi(\varepsilon')
\end{aligned}$$

Therefore

$$\begin{aligned}
\int g(\pi,\nu,\varepsilon)\Phi(\varepsilon)d\varepsilon &\geq \int_{E'-A_{\nu,\nu'}} g(\pi,\nu,\varepsilon)\Phi(\varepsilon)d\varepsilon = \int_{E'} g(\pi,\nu,\varepsilon)\Phi(\varepsilon)d\varepsilon' \\
&\geq \exp\left(-\frac{1}{2}Tk\epsilon^2 - \epsilon\sqrt{2Tk}\right)\int_{E'}(d'(p,p') - g(\pi,\nu',\varepsilon'))\,\Phi(\varepsilon')d\varepsilon' \\
&= \exp\left(-\frac{1}{2}Tk\epsilon^2 - \epsilon\sqrt{2Tk}\right)\left(d'(p,p')\mathbb{P}(E') - \int_{E'} g(\pi,\nu',\varepsilon')\Phi(\varepsilon')d\varepsilon'\right) \\
&\geq \exp\left(-\frac{1}{2}Tk\epsilon^2 - \epsilon\sqrt{2Tk}\right)\left(d'(p,p')\mathbb{P}(E') - \int g(\pi,\nu',\varepsilon')\Phi(\varepsilon')d\varepsilon'\right).
\end{aligned}$$

According to Hoeffding's inequality, we have $\mathbb{P}(\varepsilon' \in E') \geq 1 - 1/e > 1/2$. Therefore we have

$$\begin{aligned}
\int g(\pi,\nu,\varepsilon)\Phi(\varepsilon)d\varepsilon &\geq \exp\left(-\frac{1}{2}Tk\epsilon^2 - \epsilon\sqrt{2Tk}\right)\left(d'(p,p')\mathbb{P}(E') - \int g(\pi,\nu',\varepsilon')\Phi(\varepsilon')d\varepsilon'\right) \\
&\geq \frac{1}{2}\exp\left(-\frac{1}{2}Tk\epsilon^2 - \epsilon\sqrt{2Tk}\right)d'(p,p') - \int g(\pi,\nu',\varepsilon')\Phi(\varepsilon')d\varepsilon',
\end{aligned}$$

which completes the proof. □

G PROOF OF THEOREM 4

Proof. If $\epsilon = \frac{1}{\sqrt{Tk}}$, then we have

$$\frac{1}{2}\exp\left(-\frac{1}{2}Tk\epsilon^2 - \epsilon\sqrt{2Tk}\right) = \frac{1}{2}\exp(-\frac{1}{2} - \sqrt{2}) > \frac{1}{14},$$

and

$$\exp(2\epsilon) = \exp\left(\frac{2}{\sqrt{Tk}}\right) \le \exp\left(\frac{2}{\sqrt{5\cdot 2}}\right) < \exp\left(\frac{2}{3}\right) < 2.$$

We want to choose r and r' in Lemma 1 such that $|r_i - r'_i| \le \epsilon$. Let

$$\forall 1 \le i \le k, \quad r_i = \epsilon, \quad r'_i = \begin{cases} 2\epsilon & \text{if } i < (k+1)/2; \\ \log\left(\frac{\exp(2\epsilon)+1}{2}\right) & \text{if } k \text{ is odd and } i = (k+1)/2; \\ 0 & \text{if } i > (k+1)/2. \end{cases}$$

It follows that

$$\forall 1 \le i \le k, \quad p_i = \frac{1}{k}, \quad p'_i = \begin{cases} \frac{2\exp(2\epsilon)}{k(\exp(2\epsilon)+1)} & \text{if } i < (k+1)/2; \\ \frac{1}{k} & \text{if } k \text{ is odd and } i = (k+1)/2; \\ \frac{2}{k(\exp(2\epsilon)+1)} & \text{if } i > (k+1)/2. \end{cases}$$

Hence we have

$$\begin{aligned} d(p', p) = \frac{1}{2}\sum_{i=1}^{k} |p'_i - p_i| &= \begin{cases} \frac{\exp(2\epsilon)-1}{2(\exp(2\epsilon)+1)} & \text{if } k \text{ is even} \\ \frac{k-1}{k}\cdot\frac{\exp(2\epsilon)-1}{2(\exp(2\epsilon)+1)} & \text{if } k \text{ is odd} \end{cases} \\ &\ge \frac{2}{3}\cdot\frac{\exp(2\epsilon)-1}{2(\exp(2\epsilon)+1)} \ge \frac{2}{3}\cdot\frac{2\epsilon}{2(\exp(2\epsilon)+1)} \ge \frac{2\epsilon}{9} = \frac{2}{9\sqrt{Tk}}, \end{aligned}$$

where the second inequality follows from the fact that $\exp(x) \ge 1 + x$ for all $x \in \mathbb{R}$ and the last inequality follows from the fact that $\exp(2\epsilon) < 2$. Now we may use Lemma 1 to see that

$$\begin{aligned} \sup_{\nu} \mathbb{E}\left[\mathcal{R}^{f,d^{\mathrm{TV}}}(\pi,\nu)\right] &\ge \frac{1}{2}\left(\mathbb{E}\left[\mathcal{R}^{f,d^{\mathrm{TV}}}(\pi,\nu)\right] + \mathbb{E}\left[\mathcal{R}^{f,d^{\mathrm{TV}}}(\pi,\nu')\right]\right) \\ &\ge \frac{1}{28} d^{\mathrm{TV}}(p', p) \ge \frac{1}{126\sqrt{Tk}}. \end{aligned}$$

For $d^{\text{r-KL}}$, we have

$$\begin{aligned} \sup_{\nu} \mathbb{E}\left[\mathcal{R}^{f,d^{\text{r-KL}}}(\pi,\nu)\right] &\ge \frac{1}{2}\left(\mathbb{E}\left[\mathcal{R}^{f,d^{\text{r-KL}}}(\pi,\nu)\right] + \mathbb{E}\left[\mathcal{R}^{f,d^{\text{r-KL}}}(\pi,\nu')\right]\right) \ge \frac{1}{28} d'(p', p) \\ &\ge \frac{1}{28} d(p', p)^2 \ge \frac{1}{28}\left(\frac{2}{9\sqrt{Tk}}\right)^2 = \frac{1}{567Tk}. \end{aligned}$$

The proof for $d^{\text{f-KL}}$ is identical. □

H PROOF OF THEOREMS 5 and 6

In this section, we will provide the regret bounds and proofs for Theorems 5 and 6. Algorithm 2 provides a pseudo-code for the family of algorithms described in Theorem 6. In particular, we can immediately see that setting $C = 0$, reduces Algorithm 2 to Algorithm 1.

We start with some notations and definitions. Let $\delta = T^{-2}$ and $\sigma_* = (1 + C)\sigma$. Also let

$$\begin{aligned} \epsilon &:= \sqrt{\frac{2(1+C)\sigma^2\log(1/\delta)}{M}} = \sqrt{\frac{2\sigma_*^2\log(1/\delta)}{M}}, \\ T_0 &:= \max\left\{\min\left\{Mk, \frac{k}{1-\exp(-2\epsilon)}\right\}, \frac{8}{\alpha^2}\exp(4\epsilon)\log(1/\delta)\right\}, \end{aligned}$$

where $\alpha = \min_i p_i$. Note that if $M \ge 18\sigma_*^2\log(1/\delta) = 36\sigma_*^2\log(T)$, then

$$\epsilon = \sqrt{\frac{2\sigma_*^2\log(1/\delta)}{M}} \le \sqrt{\frac{2\sigma_*^2\log(1/\delta)}{18\sigma_*^2\log(1/\delta)}} = \frac{1}{3},$$

Algorithm 2: Active Sampling with Warm-up within Confidence Bounds (ASWCB)

Input : Horizon T, number of arms k, exploration factor $1 \leq M \leq T/k$, confidence factor $C \geq 0$

```
for t = 1 to T do
    if t ≤ Mk then
        Play a_t ← (t mod k) + 1
    else
        Sample 1 ≤ a_t ≤ k according to p̂_t and play a_t
    end
    for 1 ≤ i ≤ k do
        Choose r̂_{t,i} such that |r̂_{t,i} − (1/n_{t,i}) Σ_{t'=1}^{t} x_{t'} 1_{a_{t'}=i}| ≤ 2Cσ √(log(T)/n_{t,i})
    end
    p̂_{t+1} ← softmax(r̂_t)
end
```

which implies that $\exp(\epsilon) \leq \frac{3}{2}$. More generally, if $M = \Omega(\log(T))$, then $\epsilon = O(1)$ and $T_0 = O(\log(T))$. On the other hand, if $M = 1$, then $\epsilon = O((\log T)^{1/2})$ and therefore $\exp(A\epsilon) = \hat{O}^{1/2}(1)$ for any $A > 0$. Thus we also have $T_0 = O(\exp(4\epsilon)\log(T)) = \tilde{O}(\exp(4\epsilon)) = \hat{O}^{1/2}(1)$.

Recall that $\mathbf{n}_{t,i} = \sum_{t'=1}^{t} \mathbf{1}_{\mathbf{a}_{t'}=i}$. We define $\mathbf{n}'_{t,i} = \mathbf{n}_{t,i} - \sum_{t'=1}^{t} \hat{\pi}_{t',i}$. We also define $\mathcal{E}'_t$ to be the event that the inequality

$$|\mathbf{n}'_{t,i} - \mathbf{n}'_{0,i}| \leq \sqrt{2t\log(1/\delta)}$$

holds for all $1 \leq i \leq k$. We have

$$\begin{aligned}\mathbb{E}[\mathbf{n}'_{t+1,i}|\mathbf{n}'_{1,i}, \cdots, \mathbf{n}'_{t,i}] &= \mathbb{E}[\mathbf{1}_{\mathbf{a}_{t+1}=i} - \hat{\pi}_{t+1,i} + \mathbf{n}'_{t,i}|\mathbf{n}'_{1,i}, \cdots, \mathbf{n}'_{t,i}] \\ &= \mathbb{E}[\mathbf{1}_{\mathbf{a}_{t+1}=i} - \hat{\pi}_{t+1,i}|\mathbf{n}'_{1,i}, \cdots, \mathbf{n}'_{t,i}] + \mathbf{n}'_{t,i} = \mathbf{n}'_{t,i}.\end{aligned}$$

Therefore, $\{\mathbf{n}'_{t,i}\}_{t=0}^{T}$ is a martingale where

$$|\mathbf{n}'_{t+1,i} - \mathbf{n}'_{t,i}| = |\mathbf{1}_{\mathbf{a}_{t+1}=i} - \hat{\pi}_{t+1,i}| \leq 1.$$

Hence we may use Azuma-Hoeffding's inequality to see that, for each $1 \leq i \leq k$, the inequality

$$\mathbf{n}'_{t,i} = |\mathbf{n}'_{t,i} - \mathbf{n}'_{0,i}| \leq \sqrt{2t\log(1/\delta)} \tag{3}$$

occurs with probability $1 - \delta$. Hence $\mathbb{P}(\mathcal{E}'_t) \geq 1 - k\delta$. On the other hand, according to Hoeffding's inequality, with probability $1 - \delta$, we have

$$\left|\frac{1}{\mathbf{n}_{t,i}}\sum_{t'=1}^{t}\mathbf{x}_{t'}\mathbf{1}_{\mathbf{a}_{t'}=i} - r_i\right| \leq \sqrt{\frac{2\sigma^2\log(1/\delta)}{\mathbf{n}_{t,i}}}$$

and therefore, since $\delta = T^{-2}$, we see that

$$\begin{aligned}|\hat{\mathbf{r}}_{t,i} - r_i| &\leq \left|\frac{1}{\mathbf{n}_{t,i}}\sum_{t'=1}^{t}\mathbf{x}_{t'}\mathbf{1}_{\mathbf{a}_{t'}=i} - r_i\right| + \left|\hat{\mathbf{r}}_{t,i} - \frac{1}{\mathbf{n}_{t,i}}\sum_{t'=1}^{t}\mathbf{x}_{t'}\mathbf{1}_{\mathbf{a}_{t'}=i}\right| \\ &\leq (1+C)\sqrt{\frac{2\sigma^2\log(1/\delta)}{\mathbf{n}_{t,i}}} = \sqrt{\frac{2\sigma_*^2\log(1/\delta)}{\mathbf{n}_{t,i}}}\end{aligned}$$

Let $\mathcal{E}_t$ be the event where the above inequality holds for all $1 \leq i \leq k$. It follows that $\mathbb{P}(\mathcal{E}_t) \geq 1 - k\delta$. Finally, we define $\mathcal{E}$ to be the event $\bigcap_{t'=Mk}^{T}(\mathcal{E}_{t'} \cap \mathcal{E}'_{t'})$. Thus $\mathbb{P}(\mathcal{E}) \geq 1 - 2kT\delta$.

As can be seen in Algorithm 2, we have $\hat{\pi}_t = \hat{\mathbf{p}}_t$ for all $t \geq Mk + 1$. On the other hand, for $1 \leq t \leq Mk$, $\hat{\pi}_t$ is the Dirac distribution with mass at $(t \mod k) + 1$. Therefore, for all $t \geq Mk$, we have

$$\mathbf{n}_{t,i} = \sum_{t'=1}^{t}\hat{\pi}_{t',i} + \mathbf{n}'_{t,i} = M + \sum_{t'=Mk+1}^{t}\hat{\mathbf{p}}_{t',i} + \mathbf{n}'_{t,i}.$$

We provide the proof in a series of lemmas.

Lemma 4. *If $T_0 \le t \le T$, and the event $\mathcal{E}$ occurs, then*

$$d^{\mathrm{TV}}(\hat{\mathbf{p}}_t, p) \le 4\sigma_* \exp(\epsilon)\sqrt{\frac{\log(1/\delta)}{\alpha t}}. \tag{4}$$

Proof. For $t' \ge Mk$, the event $\mathcal{E}_{t'}$ occurs and therefore we have

$$|\hat{\mathbf{r}}_{t',i} - r_i| \le \sqrt{\frac{2\sigma_*^2 \log(1/\delta)}{\mathbf{n}_{t',i}}} \le \sqrt{\frac{2\sigma_*^2 \log(1/\delta)}{M}} = \epsilon.$$

If $t \ge Mk+1$, using Lemma 2 and the fact that $\alpha \le 1/k$, we have

$$\begin{aligned}\sum_{t'=1}^{t} \hat{\pi}_{t',i} = M + \sum_{t'=Mk+1}^{t} \hat{\mathbf{p}}_{t',i} &\ge M + \exp(-2\epsilon) \sum_{t'=Mk+1}^{t} p_i = M + \exp(-2\epsilon)(t - Mk)p_i \\ &\ge M + \exp(-2\epsilon)(t - Mk)\alpha \ge \exp(-2\epsilon)\alpha t.\end{aligned} \tag{5}$$

On the other hand, if $T_0 \le t \le Mk$, then by definition of T_0 we have $t \ge \frac{k}{1-\exp(-2\epsilon)}$ and therefore

$$\sum_{t'=1}^{t} \hat{\pi}_{t',i} = \lfloor t/k \rfloor \ge t/k - 1 \ge \exp(-2\epsilon)t/k \ge \exp(-2\epsilon)\alpha t. \tag{6}$$

Thus, since $\mathcal{E}'_t$ occurs, we see that

$$\mathbf{n}_{t,i} = \sum_{t'=1}^{t} \hat{\pi}_{t',i} + \mathbf{n}'_{t,i} \ge \exp(-2\epsilon)\alpha t - \sqrt{2t\log(1/\delta)}. \tag{7}$$

The inequality $t \ge T_0 \ge \frac{8}{\alpha^2}\exp(4\epsilon)\log(1/\delta)$ implies that

$$\sqrt{2t\log(1/\delta)} \le \sqrt{\frac{2t^2\log(1/\delta)}{\frac{8}{\alpha^2}\exp(4\epsilon)\log(1/\delta)}} = \frac{1}{2}\exp(-2\epsilon)\alpha t.$$

Plugging this into Inequality 7, we see that

$$\mathbf{n}_{t,i} \ge \exp(-2\epsilon)\alpha t - \frac{1}{2}\exp(-2\epsilon)\alpha t = \frac{1}{2}\exp(-2\epsilon)\alpha t.$$

Hence, using the fact that $\mathcal{E}_t$ holds, we see that

$$|\hat{\mathbf{r}}_{t,i} - r_i| \le \sqrt{\frac{2\sigma_*^2\log(1/\delta)}{\mathbf{n}_{t,i}}} \le \sqrt{\frac{4\exp(2\epsilon)\sigma_*^2\log(1/\delta)}{\alpha t}} = 2\sigma_*\exp(\epsilon)\sqrt{\frac{\log(1/\delta)}{\alpha t}}.$$

Therefore, Lemma 2 implies that

$$d^{\mathrm{TV}}(\hat{\mathbf{p}}_t, p) \le 2\max_{1\le i\le k}|\hat{\mathbf{r}}_{t,i} - r_i| \le 4\sigma_*\exp(\epsilon)\sqrt{\frac{\log(1/\delta)}{\alpha t}},$$

which completes the proof. □

Lemma 5 (Policy-level total-variation regret)**.** *If $T_0 \le t \le T$, then we have*

$$\begin{aligned}\mathbb{E}\left[\mathcal{SPR}_t^{\mathrm{TV}}\right] &\le 4\sigma_*\exp(\epsilon)\sqrt{\frac{\log(1/\delta)}{\alpha t}} + 2kT\delta, \\ \mathbb{E}\left[\mathcal{CPR}_T^{\mathrm{TV}}\right] &\le 8\sigma_*\exp(\epsilon)\sqrt{\frac{\log(1/\delta)T}{\alpha}} + T_0 + 2kT\delta.\end{aligned}$$

Thus, $\mathbb{E}\left[\mathcal{SPR}_T^{\mathrm{TV}}\right] = \tilde{O}(\exp(\epsilon)T^{-1/2})$ *and* $\mathbb{E}\left[\mathcal{CPR}_T^{\mathrm{TV}}\right] = \tilde{O}(\exp(4\epsilon)T^{1/2})$.

Proof. Assume $T_0 \leq t \leq T$. Using Lemma 4 we see that

$$\mathbb{E}\left[\mathcal{SPR}_t^{\mathrm{TV}}|\mathcal{E}\right] = \mathbb{E}\left[d^{\mathrm{TV}}(\hat{\mathbf{p}}_t, p)|\mathcal{E}\right] \leq 4\sigma_* \exp(\epsilon)\sqrt{\frac{\log(1/\delta)}{\alpha t}}.$$

Note that $\mathcal{SPR}_t^{\mathrm{TV}} = d^{\mathrm{TV}}(\hat{\pi}_t \| p) \leq 1$, for all $1 \leq t \leq T$. Therefore, for $T_0 \leq t \leq T$, we have

$$\begin{aligned}
\mathbb{E}\left[\mathcal{SPR}_t^{\mathrm{TV}}\right] &= \mathbb{P}(\mathcal{E})\mathbb{E}\left[\mathcal{SPR}_t^{\mathrm{TV}}|\mathcal{E}\right] + \mathbb{P}(\mathcal{E}^c)\mathbb{E}\left[\mathcal{SPR}_t^{\mathrm{TV}}|\mathcal{E}^c\right] \\
&\leq \mathbb{E}\left[\mathcal{SPR}_t^{\mathrm{TV}}|\mathcal{E}\right] + \mathbb{P}(\mathcal{E}^c) \\
&\leq 4\sigma_* \exp(\epsilon)\sqrt{\frac{\log(1/\delta)}{\alpha t}} + 2kT\delta.
\end{aligned}$$

To bound cumulative regret, we see that when $\mathcal{E}$ occurs,

$$\begin{aligned}
\mathcal{CPR}_T^{\mathrm{TV}} &= \sum_{t=1}^{T} \mathcal{SPR}_t^{\mathrm{TV}} \\
&= \sum_{t=1}^{T_0-1} \mathcal{SPR}_t^{\mathrm{TV}} + \sum_{t=T_0}^{T} \mathbb{E}\mathcal{SPR}_t^{\mathrm{TV}} \\
&\leq T_0 + \sum_{t=T_0}^{T} \left(4\sigma_* \exp(\epsilon)\sqrt{\frac{\log(1/\delta)}{\alpha t}}\right) \\
&\leq T_0 + 8\sigma_* \exp(\epsilon)\sqrt{\frac{\log(1/\delta)T}{\alpha}},
\end{aligned} \tag{8}$$

where we used the fact that $\sum_{n=1}^{N} \frac{1}{\sqrt{n}} < 2\sqrt{N}$ in the last inequality. Therefore

$$\begin{aligned}
\mathbb{E}\left[\mathcal{CPR}_T^{\mathrm{TV}}\right] &= \mathbb{P}(\mathcal{E})\mathbb{E}\left[\mathcal{CPR}_T^{\mathrm{TV}}|\mathcal{E}\right] + \mathbb{P}(\mathcal{E}^c)\mathbb{E}\left[\mathcal{CPR}_T^{\mathrm{TV}}|\mathcal{E}^c\right] \\
&\leq \mathbb{E}\left[\mathcal{CPR}_T^{\mathrm{TV}}|\mathcal{E}\right] + \mathbb{P}(\mathcal{E}^c) \\
&\leq 8\sigma_* \exp(\epsilon)\sqrt{\frac{\log(1/\delta)T}{\alpha}} + T_0 + 2kT\delta.
\end{aligned}$$

□

Lemma 6 (Action-level total-variation regret)**.** *If $T_0 \leq t \leq T$, then we have*

$$\mathbb{E}\left[\mathcal{SAR}_t^{\mathrm{TV}}\right] \leq 8\sigma_* \exp(\epsilon)\sqrt{\frac{\log(1/\delta)}{\alpha t}} + \frac{T_0}{t} + k\sqrt{\frac{\log(1/\delta)}{2t}} + 2kT\delta, \tag{9}$$

$$\mathbb{E}\left[\mathcal{CAR}_T^{\mathrm{TV}}\right] \leq 8\sigma_* \exp(\epsilon)\sqrt{\frac{\log(1/\delta)T}{\alpha}}\log T + T_0(1+\log T) + 2k\sqrt{T\log T} + 2kT^2\delta. \tag{10}$$

Thus, $\mathbb{E}\left[\mathcal{SAR}_T^{\mathrm{TV}}\right] = \tilde{O}(\exp(4\epsilon)T^{-1/2})$ *and* $\mathbb{E}\left[\mathcal{CAR}_T^{\mathrm{TV}}\right] = \tilde{O}(\exp(4\epsilon)T^{1/2})$.

Proof. We have $\mathbf{n}_{T,i} = \sum_{t=1}^{T} \hat{\pi}_{t,i} + \mathbf{n'}_{T,i}$, for $1 \leq i \leq k$, which implies that $\mathbf{q}_T = \frac{1}{T}\left(\sum_{t=1}^{T} \hat{\pi}_t\right) + \frac{\mathbf{n'}_T}{T}$. Assume $d = d^{\mathrm{TV}}$. Using Theorem 3 and the fact that d^{TV} is a metric, we have

$$\mathcal{SAR}_t^{\mathrm{TV}} - \frac{1}{t}\mathcal{CPR}_t^{\mathrm{TV}} \leq d^{\mathrm{TV}}\left(\mathbf{q}_t, p\right) - d^{\mathrm{TV}}\left(\mathbf{q}_t - \frac{\mathbf{n'}_t}{t}, p\right) \leq d^{\mathrm{TV}}\left(\mathbf{q}_t, \mathbf{q}_t - \frac{\mathbf{n'}_t}{t}\right) = \frac{1}{2t}\sum_{i=1}^{k} \left|\mathbf{n'}_t\right|.$$

Therefore we may use Inequality 3 to see that

$$\begin{aligned}
\mathbb{E}\left[\mathcal{SAR}_t^{\mathrm{TV}} \mid \mathcal{E}\right] - \mathbb{E}\left[\frac{1}{t}\mathcal{CPR}_t^{\mathrm{TV}} \mid \mathcal{E}\right] &\leq \mathbb{E}\left[\mathcal{SAR}_t^{\mathrm{TV}} \mid \mathcal{E}\right] - \mathbb{E}\left[\frac{1}{t}\mathcal{CPR}_t^{\mathrm{TV}} \mid \mathcal{E}\right] \\
&\leq \mathbb{E}\left[\frac{1}{2t}\sum_{i=1}^{k}|\mathbf{n'}_t| \mid \mathcal{E}\right] \\
&\leq \frac{1}{2t}\sum_{i=1}^{k}\mathbb{E}\left[|\mathbf{n'}_t|\right] \\
&\leq k\sqrt{\frac{\log(1/\delta)}{2t}}.
\end{aligned}$$

Thus, using the fact that d is bounded by 1, we see that

$$\begin{aligned}
\mathbb{E}\left[\mathcal{SAR}_t^{\mathrm{TV}}\right] &= \mathbb{P}(\mathcal{E})\mathbb{E}\left[\mathcal{SAR}_t^{\mathrm{TV}} \mid \mathcal{E}\right] + \mathbb{P}(\mathcal{E}^c)\mathbb{E}\left[\mathcal{SAR}_t^{\mathrm{TV}} \mid \mathcal{E}^c\right] \\
&\leq \mathbb{E}\left[\mathcal{SAR}_t^{\mathrm{TV}} \mid \mathcal{E}\right] + \mathbb{P}(\mathcal{E}^c) \\
&\leq \frac{1}{t}\mathbb{E}\left[\mathcal{CPR}_t^{\mathrm{TV}} \mid \mathcal{E}\right] + k\sqrt{\frac{\log(1/\delta)}{2t}} + 2kT\delta \\
&\leq 8\sigma_* \exp(\epsilon)\sqrt{\frac{\log(1/\delta)}{\alpha t}} + \frac{T_0}{t} + k\sqrt{\frac{\log(1/\delta)}{2t}} + 2kT\delta.
\end{aligned}$$

where we use Inequality 8 in the last line.

Similarly, to bound the cumulative action-level regret, we note that

$$\begin{aligned}
\mathbb{E}\left[\mathcal{CAR}_T^{\mathrm{TV}}\right] &= \mathbb{P}(\mathcal{E})\mathbb{E}\left[\mathcal{CAR}_T^{\mathrm{TV}} \mid \mathcal{E}\right] + \mathbb{P}(\mathcal{E}^c)\mathbb{E}\left[\mathcal{CAR}_T^{\mathrm{TV}} \mid \mathcal{E}^c\right] \\
&\leq \mathbb{E}\left[\mathcal{CAR}_T^{\mathrm{TV}} \mid \mathcal{E}\right] + \mathbb{P}(\mathcal{E}^c)T \\
&= \sum_{t=1}^{T}\left(\mathbb{E}\left[\mathcal{SAR}_T^{\mathrm{TV}} \mid \mathcal{E}\right]\right) + \mathbb{P}(\mathcal{E}^c)T \\
&= \sum_{t=1}^{T_0-1}\left(\mathbb{E}\left[\mathcal{SAR}_T^{\mathrm{TV}} \mid \mathcal{E}\right]\right) + \sum_{t=T_0}^{T}\left(\mathbb{E}\left[\mathcal{SAR}_T^{\mathrm{TV}} \mid \mathcal{E}\right]\right) + \mathbb{P}(\mathcal{E}^c)T \\
&\leq T_0 + \sum_{t=T_0}^{T}\left(\frac{1}{t}\mathbb{E}\left[\mathcal{CPR}_t \mid \mathcal{E}\right] + k\sqrt{\frac{\log(1/\delta)}{2t}}\right) + 2kT^2\delta \\
&\leq T_0 + \log(T)\mathbb{E}\left[\mathcal{CPR}_T^{\mathrm{TV}}\right] + 2k\sqrt{T\log T} + 2kT^2\delta \\
&\leq 8\sigma_* \exp(\epsilon)\sqrt{\frac{\log(1/\delta)T}{\alpha}}\log T + T_0(1+\log T) + 2k\sqrt{T\log T} + 2kT^2\delta.
\end{aligned}$$

□

Lemma 7 (Policy-level reverse-KL regret)**.** *If $T_0 \leq t \leq T$, then we have*

$$\begin{aligned}
\mathbb{E}\left[\mathcal{SPR}_t^{\text{r-KL}}\right] &\leq \frac{32\sigma_*^2\exp(2\epsilon)\log(1/\delta)}{\alpha^2 t} + \frac{4kT\delta}{\alpha}, \\
\mathbb{E}\left[\mathcal{CPR}_T^{\text{r-KL}}\right] &\leq \frac{32\sigma_*^2\exp(2\epsilon)\log(1/\delta)\log(T)}{\alpha^2} + \frac{2T_0}{\alpha} + \frac{4kT^2\delta}{\alpha}.
\end{aligned}$$

Thus, $\mathbb{E}\left[\mathcal{SPR}_T^{\text{r-KL}}\right] = \tilde{O}(\exp(2\epsilon)T^{-1})$ *and* $\mathbb{E}\left[\mathcal{CPR}_T^{\text{r-KL}}\right] = \tilde{O}(\exp(4\epsilon))$.

Proof. Using Theorem 1 and Lemma 4 we see that

$$\mathbb{E}\left[\mathcal{SPR}_t^{\text{r-KL}}|\mathcal{E}\right] = \mathbb{E}\left[\frac{2}{\alpha}\left(\mathcal{SPR}_t^{\mathrm{TV}}\right)^2|\mathcal{E}\right] = \mathbb{E}\left[\frac{2}{\alpha}\left(d^{\mathrm{TV}}(\hat{\mathbf{p}}_t, p)\right)^2|\mathcal{E}\right] \leq \frac{32\sigma_*^2\exp(2\epsilon)\log(1/\delta)}{\alpha^2 t}.$$

Note that, using inverse Pinsker's inequality, we have $\mathcal{SPR}_{t'}^{\text{r-KL}} = D_{\text{KL}}(\hat{\pi}_{t'}\|p) \leq \frac{2}{\alpha}$ for all $1 \leq t' \leq T$. Therefore

$$\begin{aligned}\mathbb{E}\left[\mathcal{SPR}_t^{\text{r-KL}}\right] &= \mathbb{P}(\mathcal{E})\mathbb{E}\left[\mathcal{SPR}_t^{\text{r-KL}}|\mathcal{E}\right] + \mathbb{P}(\mathcal{E}^c)\mathbb{E}\left[\mathcal{SPR}_t^{\text{r-KL}}|\mathcal{E}^c\right] \\ &\leq \mathbb{E}\left[\mathcal{SPR}_t^{\text{r-KL}}|\mathcal{E}\right] + \frac{2}{\alpha}\mathbb{P}(\mathcal{E}^c) \\ &\leq \frac{32\sigma_*^2\exp(2\epsilon)\log(1/\delta)}{\alpha^2 t} + \frac{4kT\delta}{\alpha}.\end{aligned}$$

To bound the cumulative action-level regret, we note that

$$\begin{aligned}\mathbb{E}\left[\mathcal{CPR}_T^{\text{r-KL}}|\mathcal{E}\right] &= \sum_{t=1}^{T}\mathbb{E}\left[\mathcal{SPR}_t^{\text{r-KL}}|\mathcal{E}\right] \\ &= \sum_{t=1}^{T_0-1}\mathbb{E}\left[\mathcal{SDR}_t^{\text{r-KL}}|\mathcal{E}\right] + \sum_{t=T_0}^{T}\mathbb{E}\left[\mathcal{SDR}_t^{\text{r-KL}}|\mathcal{E}\right] \\ &\leq \frac{2T_0}{\alpha} + \sum_{t=T_0}^{T}\left(\frac{32\sigma_*^2\exp(2\epsilon)\log(1/\delta)}{\alpha^2 t}\right) \\ &\leq \frac{2T_0}{\alpha} + \frac{32\sigma_*^2\exp(2\epsilon)\log(1/\delta)\log(T)}{\alpha^2}.\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}\left[\mathcal{CPR}_T^{\text{r-KL}}\right] &= \mathbb{P}(\mathcal{E})\mathbb{E}\left[\mathcal{CPR}_T^{\text{r-KL}}|\mathcal{E}\right] + \mathbb{P}(\mathcal{E}^c)\mathbb{E}\left[\mathcal{CPR}_T^{\text{r-KL}}|\mathcal{E}^c\right] \\ &\leq \mathbb{E}\left[\mathcal{CPR}_T^{\text{r-KL}}|\mathcal{E}\right] + \frac{2T}{\alpha}\mathbb{P}(\mathcal{E}^c) \\ &\leq \frac{32\sigma_*^2\exp(2\epsilon)\log(1/\delta)\log(T)}{\alpha^2} + \frac{2T_0}{\alpha} + \frac{4kT^2\delta}{\alpha}.\end{aligned}$$

□

Lemma 8 (Action-level reverse-KL regret)**.** *We have*

$$\begin{aligned}\mathbb{E}\left[\mathcal{SAR}_T^{\text{r-KL}}\right] \leq{}& \frac{32\sigma_*^2\exp(2\epsilon)\log(1/\delta)\log(T)}{\alpha^2 T} + \frac{2T_0}{\alpha T} + \frac{\exp(2\epsilon)k^2\log(1/\delta)}{\alpha T} \\ &+ \frac{8\sqrt{2}k\sigma_*\exp(3\epsilon)\log(1/\delta)}{\alpha^{3/2}T} + \frac{k\exp(2\epsilon)\sqrt{2\log(1/\delta)}}{\alpha T^{3/2}}T_0 + \frac{4kT\delta}{\alpha}.\end{aligned}$$

Thus, $\mathbb{E}\left[\mathcal{SAR}_T^{\text{r-KL}}\right] = \tilde{O}(\exp(6\epsilon)T^{-1})$ *and* $\mathbb{E}\left[\mathcal{CAR}_T^{\text{r-KL}}\right] = \tilde{O}(\exp(6\epsilon))$.

Proof. We only provide the proof for $\mathbb{E}\left[\mathcal{SAR}_T^{\text{r-KL}}\right]$ for brevity. The bound for the cumulative regret may be obtained from the simple regret similar to previous lemmas.

Using Theorem 3 with $d = d^{\text{r-KL}}$, we have

$$\begin{aligned}\mathcal{SAR}_T^{\text{r-KL}} - \frac{1}{T}\mathcal{CPR}_T^{\text{r-KL}} &\leq d^{\text{r-KL}}\left(\mathbf{q}_T, p\right) - d^{\text{r-KL}}\left(\frac{1}{T}\left(\sum_{t=1}^{T}\hat{\pi}_t\right), p\right) \\ &= d^{\text{r-KL}}\left(\mathbf{q}_T, p\right) - d^{\text{r-KL}}\left(\mathbf{q}_T - \frac{\mathbf{n}'_T}{T}, p\right) \\ &= \sum_{i=1}^{k}\mathbf{q}_{T,i}\log\left(\frac{\mathbf{q}_{T,i}}{p_i}\right) - \sum_{i=1}^{k}\left(\mathbf{q}_{T,i} - \frac{\mathbf{n}'_{T,i}}{T}\right)\log\left(\frac{\mathbf{q}_{T,i} - \frac{\mathbf{n}'_{T,i}}{T}}{p_i}\right) \\ &= \sum_{i=1}^{k}\mathbf{q}_{T,i}\log\left(\frac{\mathbf{q}_{T,i}}{p_i}\right) - \sum_{i=1}^{k}\mathbf{q}_{T,i}\log\left(\frac{\mathbf{q}_{T,i} - \frac{\mathbf{n}'_{T,i}}{T}}{p_i}\right) + \sum_{i=1}^{k}\frac{\mathbf{n}'_{T,i}}{T}\log\left(\frac{\mathbf{q}_{T,i} - \frac{\mathbf{n}'_T}{T}}{p_i}\right) \\ &= \sum_{i=1}^{k}\mathbf{q}_{T,i}\log\left(\frac{\mathbf{q}_{T,i}}{\mathbf{q}_{T,i} - \frac{\mathbf{n}'_{T,i}}{T}}\right) + \sum_{i=1}^{k}\frac{\mathbf{n}'_{T,i}}{T}\log\left(\frac{\mathbf{q}_{T,i} - \frac{\mathbf{n}'_{T,i}}{T}}{p_i}\right) \\ &= D_{\text{KL}}\left(\mathbf{q}_T\|\mathbf{q}_T - \frac{\mathbf{n}'_{T,i}}{T}\right) + \sum_{i=1}^{k}\frac{\mathbf{n}'_{T,i}}{T}\log\left(\frac{\mathbf{q}_{T,i} - \frac{\mathbf{n}'_{T,i}}{T}}{p_i}\right).\end{aligned} \tag{11}$$

Using inverse Pinsker's inequality, the first term may be bounded as

$$\begin{aligned}
D_{\mathrm{KL}}\left(\mathbf{q}_T \| \mathbf{q}_T - \frac{\mathbf{n}'_T}{T}\right) &\leq \frac{2}{\min_i \left\{\mathbf{q}_{T,i} - \frac{\mathbf{n}'_{T,i}}{T}\right\}} d^{\mathrm{TV}}\left(\mathbf{q}_T, \mathbf{q}_T - \frac{\mathbf{n}'_T}{T}\right)^2 \\
&\leq \frac{2}{\min_i \left\{\mathbf{q}_{T,i} - \frac{\mathbf{n}'_{T,i}}{T}\right\}} \left(\frac{1}{2T}\sum_{i=1}^k |\mathbf{n}'_T|\right)^2 \\
&= \frac{2}{\min_i \left\{\frac{1}{T}\sum_{t=1}^T \hat{\pi}_{t,i}\right\}} \left(\frac{1}{2T}\sum_{i=1}^k |\mathbf{n}'_T|\right)^2 .
\end{aligned}$$

If we assume that $\mathcal{E}$ occurs, using Inequalities 5 and 6, we may further bound it as

$$\begin{aligned}
D_{\mathrm{KL}}\left(\mathbf{q}_T \| \mathbf{q}_T - \frac{\mathbf{n}'_T}{T}\right) &\leq \frac{2}{\min_i \left\{\frac{1}{T}\sum_{t=1}^T \hat{\pi}_{t,i}\right\}} \left(\frac{1}{2T}\sum_{i=1}^k |\mathbf{n}'_T|\right)^2 \\
&\leq \frac{2}{\exp(-2\epsilon)\alpha} \left(\frac{1}{2T}\sum_{i=1}^k |\mathbf{n}'_T|\right)^2 \\
&\leq \frac{2}{\exp(-2\epsilon)\alpha} \left(\frac{1}{2T}\sum_{i=1}^k \sqrt{2T\log(1/\delta)}\right)^2 \\
&= \frac{\exp(2\epsilon)k^2\log(1/\delta)}{\alpha T}. \qquad (12)
\end{aligned}$$

On the other hand, using the fact that $|\log(x)| \leq \max\{|x-1|, |1/x-1|\}$ for all $x > 0$, we have

$$\begin{aligned}
\sum_{i=1}^k \frac{\mathbf{n}'_{T,i}}{T} \log\left(\frac{\mathbf{q}_{T,i} - \frac{\mathbf{n}'_{T,i}}{T}}{p_i}\right) &= \sum_{i=1}^k \frac{\mathbf{n}'_{T,i}}{T} \log\left(\frac{\frac{1}{T}\sum_{t=1}^T \hat{\pi}_{t,i}}{p_i}\right) \\
&\leq \sum_{i=1}^k \frac{|\mathbf{n}'_{T,i}|}{T} \max\left\{\left|\frac{\frac{1}{T}\sum_{t=1}^T \hat{\pi}_{t,i}}{p_i} - 1\right|, \left|\frac{p_i}{\frac{1}{T}\sum_{t=1}^T \hat{\pi}_{t,i}} - 1\right|\right\} \\
&= \sum_{i=1}^k \frac{|\mathbf{n}'_{T,i}|}{T} \left|\frac{\frac{1}{T}\sum_{t=1}^T \hat{\pi}_{t,i} - p_i}{\min\left\{p_i, \frac{1}{T}\sum_{t=1}^T \hat{\pi}_{t,i}\right\}}\right|
\end{aligned}$$

If we assume that $\mathcal{E}$ occurs, using Inequalities 5 and 6, we may further bound it as

$$\begin{aligned}
\sum_{i=1}^{k} \frac{\mathbf{n}'_{T,i}}{T} \log\left(\frac{\mathbf{q}_{T,i} - \frac{\mathbf{n}'_{T,i}}{T}}{p_i}\right) &\leq \sum_{i=1}^{k} \frac{|\mathbf{n}'_{T,i}|}{T} \left| \frac{\frac{1}{T}\sum_{t=1}^{T} \hat{\pi}_{t,i} - p_i}{\min\left\{p_i, \frac{1}{T}\sum_{t=1}^{T} \hat{\pi}_{t,i}\right\}} \right| \\
&\leq \sum_{i=1}^{k} \frac{|\mathbf{n}'_{T,i}|}{T} \left| \frac{\frac{1}{T}\sum_{t=1}^{T} \hat{\pi}_{t,i} - p_i}{\min\left\{p_i, \exp(-2\epsilon)\alpha\right\}} \right| \\
&\leq \sum_{i=1}^{k} \frac{|\mathbf{n}'_{T,i}|}{\exp(-2\epsilon)\alpha T^2} \sum_{t=1}^{T} |\hat{\pi}_{t,i} - p_i| \\
&\leq \sum_{i=1}^{k} \frac{|\mathbf{n}'_{T,i}|}{\exp(-2\epsilon)\alpha T^2} \sum_{t=1}^{T} d^{\mathrm{TV}}\left(\hat{\pi}_t, p_i\right) \\
&\leq \sum_{i=1}^{k} \frac{|\mathbf{n}'_{T,i}|}{\exp(-2\epsilon)\alpha T^2} \sum_{t=1}^{T} d^{\mathrm{TV}}\left(\hat{\pi}_t, p_i\right) \\
&= \sum_{i=1}^{k} \frac{|\mathbf{n}'_{T,i}|}{\exp(-2\epsilon)\alpha T^2} \mathcal{CPR}_T^{\mathrm{TV}} \\
&\leq \frac{k\sqrt{2T\log(1/\delta)}}{\exp(-2\epsilon)\alpha T^2} \left(T_0 + 8\sigma_* \exp(\epsilon)\sqrt{\frac{\log(1/\delta)T}{\alpha}}\right),
\end{aligned}$$

where we used Inequalities 3 and 8 in the last inequality.

Thus, using Inequalities 11 and 12, when $\mathcal{E}$ occurs, we have

$$\begin{aligned}
\mathcal{SAR}_T^{\text{r-KL}} - \frac{1}{T}\mathcal{CPR}_T^{\text{r-KL}} &\leq D_{\mathrm{KL}}\left(\mathbf{q}_T \| \mathbf{q}_T - \frac{\mathbf{n}'_{T,i}}{T}\right) + \sum_{i=1}^{k} \frac{\mathbf{n}'_{T,i}}{T} \log\left(\frac{\mathbf{q}_{T,i} - \frac{\mathbf{n}'_{T,i}}{T}}{p_i}\right) \\
&\leq \frac{\exp(2\epsilon)k^2\log(1/\delta)}{\alpha T} + \frac{k\sqrt{2T\log(1/\delta)}}{\exp(-2\epsilon)\alpha T^2}\left(8\sigma_*\exp(\epsilon)\sqrt{\frac{\log(1/\delta)T}{\alpha}} + T_0\right) \\
&= \frac{\exp(2\epsilon)k^2\log(1/\delta)}{\alpha T} + \frac{8\sqrt{2}k\sigma_*\exp(3\epsilon)\log(1/\delta)}{\alpha^{3/2}T} + \frac{k\exp(2\epsilon)\sqrt{2\log(1/\delta)}}{\alpha T^{3/2}} T_0.
\end{aligned}$$

Therefore, using Lemma 7, we have

$$\begin{aligned}
\mathbb{E}\left[\mathcal{SAR}_T^{\text{r-KL}}\right] &= \mathbb{P}(\mathcal{E})\mathbb{E}\left[\mathcal{SAR}_T^{\text{r-KL}}|\mathcal{E}\right] + \mathbb{P}(\mathcal{E}^c)\mathbb{E}\left[\mathcal{SAR}_T^{\text{r-KL}}|\mathcal{E}^c\right] \\
&\leq \mathbb{E}\left[\mathcal{SAR}_T^{\text{r-KL}}|\mathcal{E}\right] + \frac{2}{\alpha}\mathbb{P}(\mathcal{E}^c) \\
&\leq \mathbb{E}\left[\mathcal{SAR}_T^{\text{r-KL}}|\mathcal{E}\right] + \frac{4kT\delta}{\alpha} \\
&\leq \frac{1}{T}\mathbb{E}\left[\mathcal{CPR}_T^{\text{r-KL}} \mid \mathcal{E}\right] + \frac{\exp(2\epsilon)k^2\log(1/\delta)}{\alpha T} \\
&\qquad + \frac{8\sqrt{2}k\sigma_*\exp(3\epsilon)\log(1/\delta)}{\alpha^{3/2}T} + \frac{k\exp(2\epsilon)\sqrt{2\log(1/\delta)}}{\alpha T^{3/2}} T_0 + \frac{4kT\delta}{\alpha} \\
&\leq \frac{1}{T}\left(\frac{32\sigma_*^2\exp(2\epsilon)\log(1/\delta)\log(T)}{\alpha^2} + \frac{2T_0}{\alpha}\right) + \frac{\exp(2\epsilon)k^2\log(1/\delta)}{\alpha T} \\
&\qquad + \frac{8\sqrt{2}k\sigma_*\exp(3\epsilon)\log(1/\delta)}{\alpha^{3/2}T} + \frac{k\exp(2\epsilon)\sqrt{2\log(1/\delta)}}{\alpha T^{3/2}} T_0 + \frac{4kT\delta}{\alpha}.
\end{aligned}$$

□

Lemma 9 (Policy-level forward-KL regret)**.** *If we further assume that $\nu \in \mathcal{E}_{\mathrm{B}}^k(\sigma)$, $C = 0$ and $T_0 \leq t \leq T$, then*

we have

$$\mathbb{E}\left[\mathcal{SPR}_t^{\text{f-KL}}\right] \leq \frac{32\sigma^2 \exp(4\epsilon)\log(1/\delta)}{\alpha^2 t} + \frac{4\exp(2\sigma)kT\delta}{\alpha},$$
$$\mathbb{E}\left[\mathcal{CPR}_T^{\text{f-KL}}\right] \leq \frac{32\sigma^2 \exp(4\epsilon)\log(1/\delta)\log(T)}{\alpha^2} + \frac{2\exp(2\sigma)T_0}{\alpha} + \frac{4\exp(2\sigma)kT^2\delta}{\alpha},$$

where cumulative regret is defined as the sum from time-step $Mk+1$, as discussed in Remark 1. Thus, $\mathbb{E}\left[\mathcal{SPR}_T^{\text{f-KL}}\right] = \tilde{O}(\exp(4\epsilon)T^{-1})$ *and* $\mathbb{E}\left[\mathcal{CPR}_T^{\text{f-KL}}\right] = \tilde{O}(\exp(4\epsilon))$.

Proof. Recall that $\mathcal{E}_{\text{B}}^k(\sigma) \subseteq \mathcal{E}_{\text{SG}}^k(\sigma^2)$ by Hoeffding's lemma and therefore all the previous lemmas hold. Using Theorem 1 and Lemma 4 we see that

$$\begin{aligned}
\mathbb{E}\left[\mathcal{SPR}_t^{\text{f-KL}}|\mathcal{E}\right] &= \mathbb{E}\left[\frac{2}{\min_i \hat{\mathbf{p}}_{t,i}}\left(\mathcal{SPR}_t^{\text{TV}}\right)^2|\mathcal{E}\right] \\
&= \mathbb{E}\left[\frac{2}{\min_i \hat{\mathbf{p}}_{t,i}}\left(d^{\text{TV}}(\hat{\mathbf{p}}_t, p)\right)^2|\mathcal{E}\right] \\
&\leq \mathbb{E}\left[\frac{2}{\exp(-2\epsilon)\alpha}\left(d^{\text{TV}}(\hat{\mathbf{p}}_t, p)\right)^2|\mathcal{E}\right] \\
&\leq \frac{32\sigma^2\exp(4\epsilon)\log(1/\delta)}{\alpha^2 t}.
\end{aligned}$$

Since $\nu \in \mathcal{E}_{\text{B}}^k(\sigma)$, all observation are within σ of the true reward for all arms. Thus, since $C = 0$, we have

$$|\hat{\mathbf{r}}_{t,i} - r_i| = \left|\frac{1}{\mathbf{n}_{t,i}}\sum_{t'=1}^{t}\mathbf{x}_{t'}\mathbf{1}_{\mathbf{a}_{t'}=i} - r_i\right| \leq \sigma.$$

Hence we may use Lemma 2 to see that $\hat{p}_{t',i} \geq p_i \exp(-2\sigma)$. Therefore we have

$$\mathcal{SPR}_{t'}^{\text{f-KL}} = D_{\text{KL}}(p\|\hat{\pi}_{t'}) \leq \frac{2}{\min_i \hat{\pi}_{t',i}} = \frac{2}{\min_i \hat{\mathbf{p}}_{t',i}} \leq \frac{2}{\exp(-2\sigma)\alpha},$$

for all $Mk+1 \leq t' \leq T$. Thus

$$\begin{aligned}
\mathbb{E}\left[\mathcal{SPR}_t^{\text{f-KL}}\right] &= \mathbb{P}(\mathcal{E})\mathbb{E}\left[\mathcal{SPR}_t^{\text{f-KL}}|\mathcal{E}\right] + \mathbb{P}(\mathcal{E}^c)\mathbb{E}\left[\mathcal{SPR}_t^{\text{f-KL}}|\mathcal{E}^c\right] \\
&\leq \mathbb{E}\left[\mathcal{SPR}_t^{\text{f-KL}}|\mathcal{E}\right] + \frac{2\exp(2\sigma)}{\alpha}\mathbb{P}(\mathcal{E}^c) \\
&\leq \frac{32\sigma^2\exp(4\epsilon)\log(1/\delta)}{\alpha^2 t} + \frac{4\exp(2\sigma)kT\delta}{\alpha}.
\end{aligned}$$

To bound the cumulative action-level regret, we note that

$$\begin{aligned}
\mathbb{E}\left[\mathcal{CPR}_T^{\text{f-KL}}|\mathcal{E}\right] &= \sum_{t=Mk+1}^{T}\mathbb{E}\left[\mathcal{SPR}_t^{\text{f-KL}}|\mathcal{E}\right] \\
&= \sum_{t=Mk+1}^{T_0-1}\mathbb{E}\left[\mathcal{SDR}_t^{\text{f-KL}}|\mathcal{E}\right] + \sum_{t=T_0}^{T}\mathbb{E}\left[\mathcal{SDR}_t^{\text{f-KL}}|\mathcal{E}\right] \\
&\leq \frac{2\exp(2\sigma)T_0}{\alpha} + \sum_{t=T_0}^{T}\left(\frac{32\sigma^2\exp(4\epsilon)\log(1/\delta)}{\alpha^2 t}\right) \\
&\leq \frac{2\exp(2\sigma)T_0}{\alpha} + \frac{32\sigma^2\exp(4\epsilon)\log(1/\delta)\log(T)}{\alpha^2}.
\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{E}\left[\mathcal{CPR}_T^{\text{f-KL}}\right] &= \mathbb{P}(\mathcal{E})\mathbb{E}\left[\mathcal{CPR}_T^{\text{f-KL}}|\mathcal{E}\right] + \mathbb{P}(\mathcal{E}^c)\mathbb{E}\left[\mathcal{CPR}_T^{\text{f-KL}}|\mathcal{E}^c\right] \\ &\leq \mathbb{E}\left[\mathcal{CPR}_T^{\text{f-KL}}|\mathcal{E}\right] + \frac{2\exp(2\sigma)T}{\alpha}\mathbb{P}(\mathcal{E}^c) \\ &\leq \frac{32\sigma^2\exp(4\epsilon)\log(1/\delta)\log(T)}{\alpha^2} + \frac{2\exp(2\sigma)T_0}{\alpha} + \frac{4\exp(2\sigma)kT^2\delta}{\alpha}.\end{aligned}$$

□

Lemma 10 (Action-level forward-KL regret)**.** *Under the assumptions of Lemma 9, we have*

$$\mathbb{E}\left[\mathcal{SAR}_T^{\text{f-KL}}\right] = \tilde{O}(\exp(6\epsilon)T^{-1}), \quad \textit{and} \quad \mathbb{E}\left[\mathcal{CAR}_T^{\text{f-KL}}\right] = \tilde{O}(\exp(6\epsilon)).$$

Proof. The proof is similar to the previous cases. □

I PROOF OF THEOREM 7

Proof. For $0 < \beta < \infty$, let $Z^\beta = \sum_{j=1}^k e^{\beta r_j}$. It follows that $p^\beta = \text{softmax}(\beta r) = \frac{e^{\beta r_i}}{Z^\beta}$. Therefore, if we define $r_* = \max_{1\leq i\leq k} r_i$ and $p_*^\beta = \max_{1\leq i\leq k} p_i^\beta$, then we have $\log(p_*^\beta/p_i^\beta) = \beta(r_* - r_i)$ for $1 \leq i \leq k$. Hence

$$\begin{aligned}d^{\text{r-KL}}(\mathbf{q}_T, p^\beta) = D_{\text{KL}}(\mathbf{q}_T\|p^\beta) &= \sum_{i=1}^k \mathbf{q}_{T,i}\log\left(\frac{\mathbf{q}_{T,i}}{p_i^\beta}\right) \\ &= \sum_{i=1}^k \mathbf{q}_{T,i}\log\left(\frac{p_*^\beta}{p_i^\beta}\right) - \log\left(p_*^\beta\right) + \sum_{i=1}^k \mathbf{q}_{T,i}\log\left(\mathbf{q}_{T,i}\right) \\ &= \frac{\beta}{T}\sum_{i=1}^k \mathbf{n}_{T,i}(r_* - r_i) - \log(p_*^\beta) - H(\mathbf{q}_T) \\ &= \frac{\beta}{T}\mathcal{R}_T^{\text{MAB}} - \log(p_*^\beta) - H(\mathbf{q}_T).\end{aligned}$$

Similarly

$$\begin{aligned}d^{\text{r-KL}}(\hat{\pi}_t, p^\beta) = D_{\text{KL}}(\hat{\pi}_t\|p^\beta) &= \sum_{i=1}^k \hat{\pi}_{t,i}\log\left(\frac{\hat{\pi}_{t,i}}{p_i^\beta}\right) \\ &= \sum_{i=1}^k \hat{\pi}_{t,i}\log\left(\frac{p_*^\beta}{p_i^\beta}\right) - \log\left(p_*^\beta\right) + \sum_{i=1}^k \hat{\pi}_{t,i}\log\left(\hat{\pi}_{t,i}\right) \\ &= \beta\sum_{i=1}^k \hat{\pi}_{t,i}(r_* - r_i) - \log(p_*^\beta) - H(\hat{\pi}_t) \\ &= \beta\mathbb{E}[\mathcal{SR}_t^{\text{MAB}}|\mathbf{h}_t] - \log(p_*^\beta) - H(\hat{\pi}_t).\end{aligned}$$

Thus

$$\mathbb{E}[d^{\text{r-KL}}(\hat{\pi}_t, p^\beta)] = \mathbb{E}[\beta\mathcal{SR}_t^{\text{MAB}} - H(\hat{\pi}_t)] - \log(p_*^\beta),$$

which implies that

$$\mathbb{E}\left[\sum_{t=1}^T d^{\text{r-KL}}(\hat{\pi}_t, p^\beta)\right] = \sum_{t=1}^T \mathbb{E}\left[\beta\mathcal{SR}_t^{\text{MAB}} - H(\hat{\pi}_t)\right] - T\log(p_*^\beta).$$

The entropy terms $H(\mathbf{q}_T)$ and $H(\hat{\pi}_T)$ are always bounded by 0 and $\log(k)$ and $\log(p_*^\beta) = \max_i \log(p_i^\beta) \geq \log(1/k)$

is bounded between $-\log(k)$ and 0. It follows that

$$\lim_{\beta\to\infty} \frac{T}{\beta} d^{\text{r-KL}}(\mathbf{q}_T, p^\beta) = \mathcal{R}_T^{\text{MAB}},$$
$$\lim_{\beta\to\infty} \frac{1}{\beta} \mathbb{E}[d^{\text{r-KL}}(\hat{\pi}_t, p^\beta)] = \mathbb{E}[\mathcal{S}\mathcal{R}_t^{\text{MAB}}],$$
$$\lim_{\beta\to\infty} \frac{1}{\beta} \sum_{t=1}^{T} \mathbb{E}[d^{\text{r-KL}}(\hat{\pi}_t, p^\beta)] = \mathbb{E}[\mathcal{R}_T^{\text{MAB}}].$$

□

J MINIMIZING β-REGRET

If ν is an environment and $c \in \mathbb{R}$, we define $c\nu$ to be the environment where each observation and reward is c times the corresponding value in ν. On the other hand, for any policy π, we define π^c to be the policy that first multiplies any received observation by c and then applies π.

Theorem 8. *For any policy π, environment ν, and $0 < \beta < \infty$, we have*

$$\mathcal{R}_T^\beta(\pi^\beta, \nu) = \frac{1}{\beta}\mathcal{R}_T^1(\pi, \beta\nu).$$

To see why this holds, we simply note that the game played between π^β and ν is identical to one played between π and $\beta\nu$. We only need to divide by β to account for the effect of division by β in the definition of β-regret.

Using this theorem, we see that the regret bounds in Theorem 5 could be used to obtain β-regret bounds for π^β for all $0 < \beta < \infty$. Thus demonstrating that π^β is optimal for all such β.

At this point, we may see a contradiction by putting together the fact that (1) sampling does not require exploration and (2) sampling from the zero temperature limit of a distribution corresponds to optimization; therefore one should be able to optimize without exploration, a conclusion that contradicts the fact that exploration is necessary for multi-armed bandits. This apparent contradiction is resolved by noting that optimality is measured only in terms of time and other constants are ignored. When the temperature goes to zero, i.e., β goes to infinity, the role of these constants slowly grows and in the limit, they dominate. Thus resulting in linear regret of Algorithm 1 in the bandit setting.

To make this claim more precise, first, we note that π^∞ is effectively Explore-Then-Commit; see Appendix A.1. The reason for this claim is that π^β is equivalent to running π, but replacing the line $\hat{p}_{t+1} \leftarrow \text{softmax}(\hat{r}_t)$ with $\hat{p}_{t+1} \leftarrow \text{softmax}(\beta\hat{r}_t)$. In the limit of $\beta \to \infty$, the expression $\text{softmax}(\beta\hat{r}_t)$ tends to the distribution that has equal weights on arms with $\hat{r}_{t,i} = \max_j \hat{r}_{t,j}$ and zero weight elsewhere. Thus π^β, in some sense, tends to Explore-Then-Commit. [4] Now, if we take the limit of both sides in Theorem 8, we see that

$$\mathcal{R}_T^{\text{MAB}}(\text{ETC}, \nu) = \lim_{\beta\to\infty} \mathcal{R}_T^\beta(\pi^\beta, \nu) = \lim_{\beta\to\infty} \frac{1}{\beta}\mathcal{R}_T^1(\pi, \beta\nu).$$

Note that this is not a rigorous mathematical statement as we have not formally defined the notion of limit for algorithms. The expression of regret bound in Lemma 8 contains two terms that are affected by replacing ν with $\beta\nu$. The variance σ^2 gets multiplied by β^2 and the term $\alpha = \min_i p_i^\beta$ goes to zero exponentially fast with β. Thus we see that, even though the dependence of the regret bound on time is small, its dependence on β results in vacuous regret bound in the multi-armed bandit limit.

K EXPERIMENTS

In Figure 2, we plot action-level regret of GAS, GASW (with logarithmic exploration), DAISEE, and KL-UCB. In the first row we plot simple action-level regret multiplied by time-step to make the plots easier to see. The

[4] In ETC algorithm, the estimates for arms are not updated after the exploration phase. This is not the case in this limiting algorithm, where the estimated means will be updated at every time-step. Thus, it would be more accurate to say that the limiting algorithm is a small variant of the classical ETC algorithm. However, this variation does not have a significant effect on the performance of the algorithm. Specifically, it is trivial to see that just like the classical ETC, without significant exploration, this algorithm still suffers a linear regret.

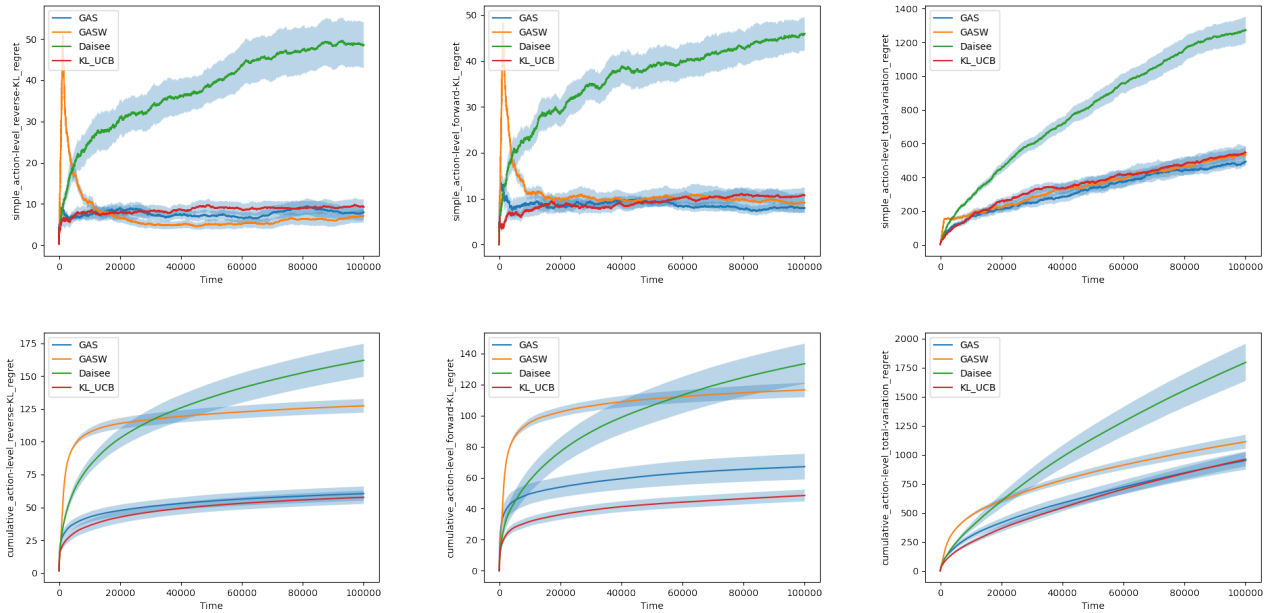

Figure 2: Action-level regret

second row demonstrates the cumulative action-level regret. From left to right, the distances are reverse-KL, forward-KL and total variation. The initial increase in regret in the orange lines corresponds to the exploration phase (and its effect on action history).

In Figure 3, we plot policy-level regret of the same algorithms. Similar to the previous figure, the first row we plot simple policy-level regret multiplied by time-step to make the plots easier to see. Here the effect of the exploration phase is more pronounced. Thus, for simple regret that does not rely on history and only depends on the current time-step, we assign zero to the GASW algorithm during the exploration phase. We include the effect of the exploration phase in the cumulative regret.

Following the discussion in Remark 1, we have assigned zero regret to the exploration phase for forward-KL policy-level regret, corresponding to the second column in Figure 3. This hides the effect of the exploration phase for this choice of statistical distance.

In these experiments, there are 10 arms, the environments have unit normal noise and the reward are sampled from $[0, 1]$. The average over 10 runs are shown and one standard error is highlighted.

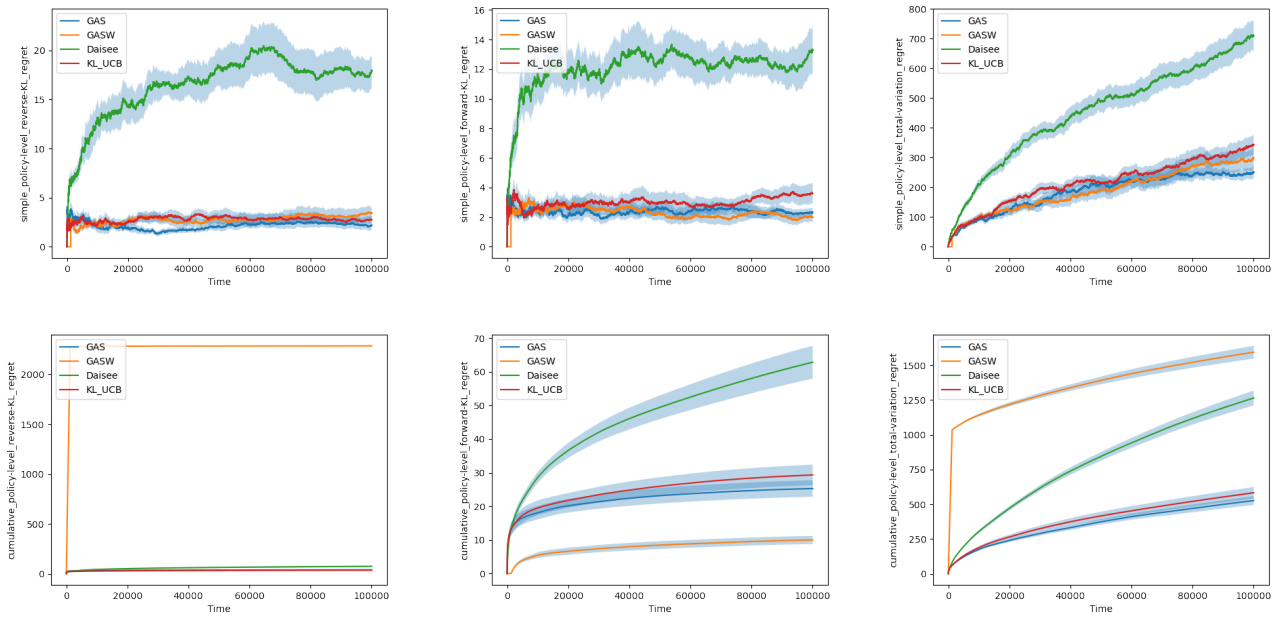

Figure 3: Policy-level regret