No-Regret Learning for Fair Multi-Agent Social Welfare Optimization

Mengxiao Zhang
University of Iowa
mengxiao-zhang@uiowa.edu

Ramiro Deo-Campo Vuong
Cornell University
ramdcv@cs.cornell.edu

Haipeng Luo
University of Southern California
haipengl@usc.edu

Abstract

We consider the problem of online multi-agent Nash social welfare (NSW) maximization. While previous works of Hossain et al. [2021], Jones et al. [2023] study similar problems in stochastic multi-agent multi-armed bandits and show that $\sqrt{T}$-regret is possible after T rounds, their fairness measure is the product of all agents' rewards, instead of their NSW (that is, their geometric mean). Given the fundamental role of NSW in the fairness literature, it is more than natural to ask whether no-regret fair learning with NSW as the objective is possible. In this work, we provide a complete answer to this question in various settings. Specifically, in stochastic N-agent K-armed bandits, we develop an algorithm with $\widetilde{\mathcal{O}}(K^{\frac{2}{N}}T^{\frac{N-1}{N}})$ regret and prove that the dependence on T is tight, making it a sharp contrast to the $\sqrt{T}$-regret bounds of Hossain et al. [2021], Jones et al. [2023]. We then consider a more challenging version of the problem with adversarial rewards. Somewhat surprisingly, despite NSW being a concave function, we prove that no algorithm can achieve sublinear regret. To circumvent such negative results, we further consider a setting with full-information feedback and design two algorithms with $\sqrt{T}$-regret: the first one has no dependence on N at all and is applicable to not just NSW but a broad class of welfare functions, while the second one has better dependence on K and is preferable when N is small. Finally, we also show that logarithmic regret is possible whenever there exists one agent who is indifferent about different arms.

1 Introduction

In this paper, we study online multi-agent Nash social welfare (NSW) maximization, which is a generalization of the classic multi-armed bandit (MAB) problem [Thompson, 1933, Lai and Robbins, 1985]. Different from MAB, in which the learner makes her decisions sequentially in order to maximize her own reward, in online multi-agent NSW maximization, the learner's decision affects multiple agents and the goal is to maximize the NSW over all the agents. Specifically, NSW is defined as the geometric mean of the expected utilities over all agents [Moulin, 2004], which can be viewed as a measure of fairness among the agents. This problem includes many important real-life applications such as resource allocation [Jones et al., 2023], where the learner needs to guarantee fair allocations among multiple agents. We refer the readers to [Hossain et al., 2021, Jones et al., 2023] for more applications of NSW maximization.

Recent work by Hossain et al. [2021], Jones et al. [2023] studies a similar problem but with NSW_{prod} as the objective, a variant of NSW that is defined as the product of the utilities over agents instead of

their geometric mean. While the optimal strategy is the same if the utility for each agent is stationary, this is not the case with a non-stationary environment. Moreover, NSW_{prod} is homogeneous of degree N instead of degree 1, where N is the number of agents, meaning that NSW_{prod} is more sensitive to the scale of the utility. Specifically, if the utilities of each agent are scaled by 2, then NSW is scaled by 2 as well, but NSW_{prod} is scaled by 2^N. Therefore, it is arguably more reasonable to consider regret with respect to NSW, which has not been studied before (to our knowledge) and is the main objective of our work.

From a technical perspective, however, due to the lack of Lipschitzness, NSW poses much more challenges in regret minimization compared to NSW_{prod}. For example, one cannot directly apply the algorithm for Lipschitz bandits [Kleinberg et al., 2019] to our problem, while it is directly applicable to NSW_{prod} as mentioned in [Hossain et al., 2021, Jones et al., 2023]. Despite such challenges, we manage to provide complete answers to this problem in various setting. Specifically, our contributions are listed below (where T, N, and K denote the number of rounds, agents, and arms/actions respectively):

- (Section 3) We first study the stochastic bandit setting, where the utility matrix at each round is i.i.d. drawn from an unknown distribution, and the learner can only observe the utilities (for different agents) of the action she picked. In this case, we develop an algorithm with $\widetilde{\mathcal{O}}(K^{\frac{2}{N}}T^{\frac{N-1}{N}})$ regret.[1] While our algorithm is also naturally based on the Upper Confidence Bound (UCB) algorithm as in Hossain et al. [2021], Jones et al. [2023], we show that a novel analysis with Bernstein-type confidence intervals is important for handling the lack of Lipschitzness of NSW. Moreover, we prove a lower bound of order $\widetilde{\Omega}(\frac{1}{N^3} \cdot K^{\frac{1}{N}}T^{\frac{N-1}{N}})$, showing that the dependence on T is tight. This is in sharp contrast to the $\sqrt{T}$-regret bound of Hossain et al. [2021], Jones et al. [2023] and demonstrates the difficulty of learning with NSW compared to NSW_{prod}.

- (Section 4.1) We then consider a more challenging setting where the utility matrix at each round can be adversarially chosen by the environment. Somewhat surprisingly, we show that no algorithm can achieve sublinear regret in this case, despite NSW being concave and the vast literature on bandit online maximization with concave utility functions (the subtlety lies in the slightly different feedback model). In fact, the same impossibility result also holds for NSW_{prod} as we show.

- (Section 4.2) To bypass such impossibility, we further consider this adversarial setting under richer feedback, where the learner observes the full utility matrix after her decision (the so-called full-information feedback). Contrary to the bandit feedback setting, learning is not only possible now but can also be much faster despite having adversarial utilities. Specifically, we design two different algorithms with $\sqrt{T}$-regret. The first algorithm is based on Follow-the-Regularized-Leader (FTRL) with the log-barrier regularizer, which achieves $\mathcal{O}(\sqrt{KT \log T})$ regret (Section 4.2.1). Notably, this algorithm does not have any dependence on the number of agents N and can also be generalized to a broader class of social welfare functions. The second algorithm is based on FTRL with a Tsallis entropy regularizer, which achieves $\widetilde{\mathcal{O}}(K^{\frac{1}{2}-\frac{1}{N}}\sqrt{NT})$ regret and is thus more favorable when K is much larger than N (Section 4.2.2). Finally, we also show that improved logarithmic regret is possible as long as at each round there exists at least one agent who is indifferent about the learner's choice of arm (Section 4.2.3).

1.1 Related Work

Hossain et al. [2021], Jones et al. [2023] are most related to our work. Hossain et al. [2021] is the first to consider designing no-regret algorithms under NSW_{prod} for the stochastic multi-agent multi-armed bandit problem. Specifically, they propose two algorithms. The first one is based on ε-greedy and achieves $\mathcal{O}(T^{\frac{2}{3}})$ regret efficiently, and the second one is based on UCB and achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ regret inefficiently. Jones et al. [2023] improves these results by providing a better UCB-based algorithm that is efficient and achieves the same $\widetilde{\mathcal{O}}(\sqrt{T})$ regret. To the best of our knowledge, there are no previous results for regret minimization over NSW under this particular setup.

However, several other models of fairness have been introduced in (single-agent or multi-agent) multi-armed bandits, some using NSW as well. These models differ in whether they aim to be fair among different objectives, different arms, different agents, different rounds, or others. Most

---

[1]The notation $\widetilde{\mathcal{O}}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ hide logarithmic dependence on K, N, and T.

related to this paper is multi-objective bandits, in which the learner tries to increase different and possibly competing objectives in a fair manner. For example, Drugan and Nowe [2013] introduces the multi-objective stochastic bandit problem and offers a regret measure to explore Pareto Optimal solutions, and Busa-Fekete et al. [2017] investigates the same setting using the Generalized Gini Index in their regret measure to promote fairness over objectives. Their regret measure closely resembles the one we use, except they apply some social welfare function (SWF) to the cumulative expected utility of agents over all rounds as opposed to the expected utility of agents each round. On the other hand, some other works study fairness among different rounds which incentivizes the learner to perform well consistently over all rounds [Barman et al., 2023, Sawarni et al., 2024]. Besides, there are other models that measure fairness in different ways, including how often each arm is pulled [Joseph et al., 2016, Liu et al., 2017, Gillen et al., 2018, Chen et al., 2020] and how the regret is allocated across different groups [Baek and Farias, 2021].

Kaneko and Nakamura [1979] axiomatically derives the NSW function. It is a fundamental and widely-adopted fairness measure and is especially popular for the task of fairly allocating goods. Caragiannis et al. [2019] justifies the fairness of NSW by showing that its maximum solution ensures some desirable envy-free property. This result prompted the design of approximation algorithms for the problem of allocating indivisible goods by maximizing NSW, which is known to be NP-hard even for simple valuation functions [Barman et al., 2018, Cole and Gkatzelis, 2015, Garg et al., 2023, Li and Vondrák, 2021].

There is a vast literature on the multi-armed bandit problem; see the book by Lattimore and Szepesvári [2020] for extensive discussions. The standard algorithm for the stochastic setting is UCB [Lai and Robbins, 1985, Auer et al., 2002a], while the standard algorithm for the adversarial setting is FTRL or the closely related Online Mirror Descent (OMD) algorithm [Auer et al., 2002b, Audibert and Bubeck, 2010, Abernethy et al., 2015]. For FTRL/OMD, the log-barrier or Tsallis entropy regularizers have been extensively studied in recent years due to some of their surprising properties (e.g., [Foster et al., 2016, Wei and Luo, 2018, Zimmert and Seldin, 2019, Lee et al., 2020]). They are rarely used in the full-information setting as far as we know, but our analysis reveals that they are useful even in such settings, especially for dealing with the lack of Lipschitzness of NSW.

2 Preliminaries

General Notation. Throughout this paper, we denote the set $\{1, 2, \ldots, n\}$ by $[n]$ for any positive integer n. For a matrix $M \in \mathbb{R}^{m \times n}$, we denote the i-th row vector of M by $M_{i,:} \in \mathbb{R}^n$, the j-th column vector of M by $M_{:,j} \in \mathbb{R}^m$, and the (i, j)-th entry of M by $M_{i,j}$. We say $M \succeq 0$ if M is a positive semi-definite matrix. The $(K - 1)$-dimensional simplex is denoted as Δ_K, and its clipped version with a parameter $\delta > 0$ is denoted as $\Delta_{K,\delta} = \{p \in \Delta_K \mid p_i \geq \delta, \forall i \in [K]\}$. We use $\mathbf{0}$ and $\mathbf{1}$ to denote the all-zero and all-one vector in an appropriate dimension. For two random variables X and Y, we use $X \stackrel{d}{=} Y$ to say X is equivalent to Y in distribution.

For a twice differentiable function f, we use $\nabla f(\cdot)$ and $\nabla^2 f(\cdot)$ to denote its gradient and Hessian. For concave functions that are not differentiable, $\nabla f(\cdot)$ denotes a super-gradient. Throughout the paper, we study functions of the form $f(u^\top p)$ for $u \in [0, 1]^{m \times n}$ and $p \in \Delta_m$. In such cases, the gradient, super-gradient, or hessian are all with respect to p unless denoted otherwise (for example, we write $\nabla_u f(u^\top p)$, with an explicit subscript u, to denote the gradient with respect to u).

Social Welfare Functions A social welfare function (SWF) $f : [0, 1]^N \to [0, 1]$ measures the desirability of the agents' expected utilities. Specifically, for two different vectors of expected utilities $\mu, \mu' \in [0, 1]^N$, $f(\mu) > f(\mu')$ means that μ is a fairer alternative than μ'. In each setting we explore, each action by the learner yields some expected utility for each of the N agents, and the learner's goal is maximize some SWF applied to these N expected utilities.

Nash Social Welfare (NSW) For the majority of this paper, we focus on a specific type of SWF, namely the Nash Social Welfare (NSW) function [Nash, 1950, Kaneko and Nakamura, 1979]. Specifically, for $\mu \in [0, 1]^N$, NSW is defined as the geometric mean of the N coordinates:

$$\text{NSW}(\mu) = \prod_{n \in [N]} \mu_n^{1/N}. \tag{1}$$

As mentioned, Hossain et al. [2021], Jones et al. [2023] considered a closely related variant that is simply the product of the coordinates: $\text{NSW}_{\text{prod}}(\mu) = \prod_{n \in [N]} \mu_n$. It is clear that NSW has a better scaling property since it is homogeneous: scaling each μ_n by a constant c also scales $\text{NSW}(\mu)$ by c, but it scales $\text{NSW}_{\text{prod}}(\mu)$ by c^N. This makes NSW_{prod} an unnatural learning objective, which motivates us to use NSW as our choice of SWF. Learning with NSW, however, brings extra challenges since it is not Lipschitz in the small-utility regime (while NSW_{prod} is Lipschitz over the entire $[0, 1]^N$). We shall see in subsequent sections how we address such challenges.

We remark that while our main focus is regret minimization with respect to NSW, some of our results also apply to NSW_{prod} or more general classes of SWFs (as will become clear later).

Problem Setup. The N-agent K-armed social welfare optimization problem we consider is defined as follows (with $N \geq 2$ and $K \geq 2$ throughout). Ahead of time, with the knowledge of the learner's algorithm, the environment decides T utility matrices $u_1, \ldots, u_T \in [0, 1]^{K \times N}$, where $u_{t,i,n}$ is the utility of agent n if arm/action i is selected at round t. Then, the learner interacts with the environment for T rounds: at each round t, the learner decides a distribution $p_t \in \Delta_K$ and then samples an action $i_t \sim p_t$. In the full-information feedback setting, the learner observes the full utility matrix u_t after her decision, and in the bandit feedback setting, the learner only observes $u_{t,i_t,n}$ for each agent $n \in [N]$, that is, the utilities of the selected action.

We consider two different types of environments, the *stochastic* one and the *adversarial* one, with a slight difference in their regret definition. In the stochastic environment, there exists a mean utility matrix $u \in [0, 1]^{K \times N}$ such that at each round t, u_t is an i.i.d. random variable with mean u. Fix an SWF f. The social welfare of a strategy $p \in \Delta_K$ is defined as $f(u^\top p)$, which is with respect to the agents' expected utilities over the randomness of both the learner's and the environment's. The regret is then defined as follows:

$$\text{Reg}_{\text{sto}} = T \cdot \max_{p \in \Delta_K} f(u^\top p) - \mathbb{E}\left[\sum_{t=1}^{T} f(u^\top p_t)\right], \tag{2}$$

which is the difference between the total social welfare of the optimal strategy and that of the learner. When f is chosen to be NSW_{prod}, Eq. (2) reduces to the regret notion considered in Hossain et al. [2021], Jones et al. [2023].

On the other hand, in the adversarial environment, we do not make any distributional assumption on the utility matrices and allow them to be selected arbitrarily. The social welfare of a strategy $p \in \Delta_K$ for time t is defined as $f(u_t^\top p)$, and the overall regret of the learner is correspondingly defined as:

$$\text{Reg}_{\text{adv}} = \max_{p \in \Delta_K} \sum_{t=1}^{T} f(u_t^\top p) - \mathbb{E}\left[\sum_{t=1}^{T} f(u_t^\top p_t)\right]. \tag{3}$$

In both Eq. (2) and Eq. (3), the expectation is taken with respect to the randomness of the algorithm.

Social welfare of expected utilities versus expected social welfare of realized utilities. One might wonder why we measure fairness using the social welfare of expected utilities (e.g., $f(u^\top p)$), instead of the expected social welfare of realized utilities (e.g., $\mathbb{E}_{i \sim p}[f(u^\top e_i)]$). This is because the former is arguably more meaningful as a fairness measure. To see this, consider $f = \text{NSW}$ or $f = \text{NSW}_{\text{prod}}$ and a setting with 2 agents, 2 arms, and u being the identity matrix. Then, in terms of $f(u^\top p)$, the uniform distribution is the best policy (which makes sense from a fairness viewpoint), while in terms of $\mathbb{E}_{i \sim p}[f(u^\top e_i)]$, all distributions achieve the same value of 0, implying that all polices are as fair, which is clearly undesired.

Connection to Bandit Convex optimization. When taking $f = \text{NSW}$ (our main focus) and considering the bandit feedback setting, our problem is seemingly an instance of the heavily-studied Bandit Convex optimization (BCO) problem, since $-\text{NSW}$ is convex. However, there is a slight but critical difference in the feedback model: a BCO algorithm would require observing $f(u_t^\top p_t)$, or equivalently $u_t^\top p_t$, at the end of each round t, while in our problem the learner only observes $u_{t,i_t,:}$, a much more realistic scenario. Even though they have the same expectation, due to the non-linearity of NSW, this slight difference in the feedback turns out to cause a huge difference in terms of learning — the minimax regret for BCO is known to be $\Theta(\sqrt{T})$, while in our problem (with bandit feedback),

Algorithm 1 UCB for N-agent K-armed NSW maximization

Input: warm-up phase length $N_0 > 0$.
Initialization: $\widehat{u}_{1,i,n} = 1$ for all $n \in [N]$, $i \in [K]$. $N_{1,i} = 0$ for all $i \in [K]$.
for $t = 1, 2, \ldots, T$ **do**

    if $t \le KN_0$ **then** select $i_t = \lceil \frac{t}{N_0} \rceil$;

    else calculate $p_t = \operatorname{argmax}_{p \in \Delta_K} \operatorname{NSW}(\widehat{u}_t^\top p)$ and select $i_t \sim p_t$;

    Observe $u_{t,i_t,n}$ for all $n \in [N]$.

    Update counters $N_{t+1,i_t} = N_{t,i_t} + 1$ and $N_{t+1,i} = N_{t,i}$ for $i \ne i_t$.

    Update upper confidence utility matrix:

$$\widehat{u}_{t+1,i,n} = \bar{u}_{t,i,n} + 4\sqrt{\frac{\bar{u}_{t,i,n} \log(NKT^2)}{N_{t+1,i}}} + \frac{8\log(NKT^2)}{N_{t+1,i}}, \tag{4}$$

    for all $n \in [N]$ and $i \in [K]$ where $\bar{u}_{t,i,n} = \frac{1}{N_{t+1,i}} \sum_{\tau \le t} u_{\tau,i,n} \mathbb{1}\{i_\tau = i\}$.

end

---

as we will soon show, the regret is either $\Theta(T^{\frac{N-1}{N}})$ in the stochastic setting or even $\Omega(T)$ in the adversarial setting. Therefore, in a sense our problem is much more difficult than BCO. For more details on BCO, we refer the reader to a recent survey by Lattimore [2024].

3 Stochastic Environments with Bandit Feedback

In this section, we consider regret minimization over $f = \operatorname{NSW}$ with bandit feedback in the stochastic setting, where the utility matrix u_t at each round $t \in [T]$ is i.i.d. drawn from a distribution with mean u. Again, this is the same setup as Hossain et al. [2021], Jones et al. [2023] except that $\operatorname{NSW}_{\operatorname{prod}}$ is replaced with NSW.

3.1 Upper Bound: a Refined Analysis of UCB with a Bernstein-Type Confidence Set

We start by describing our algorithm and its regret guarantee, followed by discussion on what the key ideas are and how the algorithm/analysis is different from previous work. Specifically, our algorithm, shown in Algorithm 1, is based on the classic UCB algorithm. It starts by picking each action for $N_0 = \widetilde{\mathcal{O}}(1)$ rounds. After this warm-up phase, at each time t, the algorithm picks the optimal strategy p_t that maximizes the NSW with respect to some upper confidence utility matrix $\widehat{u}_t$. After sampling an action $i_t \sim p_t$, the algorithm observes the utility of each agent for action i_t and then updates the upper confidence utility matrix $\widehat{u}_{t+1}$ as the empirical average utility plus a certain Bernstein-type confidence width (Eq. (4)).

The following theorem shows that Algorithm 1 guarantees $\widetilde{\mathcal{O}}(K^{\frac{2}{N}} T^{\frac{K-1}{K}})$ expected regret (with f, in the definition of $\operatorname{Reg}_{\operatorname{sto}}$, set to NSW; the same below unless stated otherwise). The full proof can be found in Appendix A.

Theorem 3.1. *With $N_0 = 1 + 18 \log KT$, Algorithm 1 guarantees $\mathbb{E}[\operatorname{Reg}_{\operatorname{sto}}] = \widetilde{\mathcal{O}}(K^{\frac{2}{N}} T^{\frac{N-1}{N}} + K)$.*

Other than replacing $\operatorname{NSW}_{\operatorname{prod}}$ with NSW, our algorithm differs from that of Jones et al. [2023] in the form of the confidence width, and the analysis sketch below explains why we need this change. Specifically, for either $f = \operatorname{NSW}_{\operatorname{prod}}$ or $f = \operatorname{NSW}$, standard analysis of UCB states that the regret is bounded by $\sum_{t=1}^T \left| f(\widehat{u}_t^\top p_t) - f(u^\top p_t) \right|$. When f is $\operatorname{NSW}_{\operatorname{prod}}$, a Lipschitz function, Hossain et al. [2021, Lemma 3] shows

$$\left| \operatorname{NSW}_{\operatorname{prod}}(\widehat{u}_t^\top p_t) - \operatorname{NSW}_{\operatorname{prod}}(u^\top p_t) \right| \le \sum_{n=1}^N \sum_{i=1}^K p_{t,i} \left| \widehat{u}_{t,i,n} - u_{i,n} \right|, \tag{5}$$

and the rest of the analysis follows by direct calculations. However, when f is NSW, a non-Lipschitz function, we cannot expect something similar to Eq. (5) anymore. Indeed, direct calculation shows that the Lipschitz constant of $\operatorname{NSW}(u^\top p)$ with respect to $u_{:,n}$ equals to $\Theta\left(\sum_{n=1}^N \langle p, u_{:,n} \rangle^{-\frac{N-1}{N}} \right)$, which can be arbitrarily large when $\langle p, u_{:,n} \rangle$ is close to 0 for some $n \in [N]$ and $N \ge 2$.

5

To handle this issue, we require a more careful analysis. Specifically, using Freedman's inequality, we know that with a high probability,

$$\widehat{u}_{t,i,n} \in \left[u_{i,n}, u_{i,n} + 8\sqrt{\frac{u_{i,n}\log(NKT^2)}{N_{t,i}}} + \widetilde{\mathcal{O}}\left(1/N_{t,i}\right) \right] \subseteq \left[u_{i,n}, 2u_{i,n} + \widetilde{\mathcal{O}}\left(1/N_{t,i}\right) \right]. \quad (6)$$

With the help of Eq. (6), we consider two different cases at each round t. The first case is that there exists certain $n \in [N]$ such that $\langle p_t, u_{:,n} \rangle \leq \sigma$ for some $\sigma > 0$ to be chosen later. In this case, we use Eq. (6) to show

$$\left| \mathrm{NSW}(\widehat{u}_t^\top p_t) - \mathrm{NSW}(u^\top p_t) \right| \leq \mathcal{O}\left(\mathrm{NSW}(u^\top p_t) \right) + \widetilde{\mathcal{O}}\left(\left(\sum_{i=1}^{K} \frac{p_{t,i}}{N_{t,i}} \right)^{\frac{1}{N}} \right)$$

$$\leq \sigma^{\frac{1}{N}} + \widetilde{\mathcal{O}}\left(\left(\sum_{i=1}^{K} \frac{p_{t,i}}{N_{t,i}} \right)^{\frac{1}{N}} \right), \quad (7)$$

where the first inequality uses Eq. (6) and the second inequality is because $\mathrm{NSW}(u^\top p_t) \leq \langle p_t, u_n \rangle^{\frac{1}{N}}$ for any $n \in [N]$. For the second term in Eq. (7), a standard analysis shows that it is upper bounded by $\widetilde{\mathcal{O}}\left(K^{\frac{1}{N}} T^{\frac{N-1}{N}} \right)$.

Now we consider the case where $\langle p_t, u_{:,n} \rangle \geq \sigma$ for all $n \in [N]$. In this case, via a decomposition lemma (Lemma C.1), we show that

$$\left| \mathrm{NSW}(\widehat{u}_t^\top p_t) - \mathrm{NSW}(u^\top p_t) \right| \leq \sum_{n=1}^{N} \left[\langle p_t, \widehat{u}_{t,:,n} \rangle^{\frac{1}{N}} - \langle p_t, u_{:,n} \rangle^{\frac{1}{N}} \right] = \mathcal{O}\left(\sum_{n=1}^{N} \frac{\langle p_t, \widehat{u}_{t,:,n} - u_{:,n} \rangle}{N \langle p_t, u_{:,n} \rangle^{\frac{N-1}{N}}} \right). \quad (8)$$

To bound Eq. (8), we use Eq. (6) again:

$$\frac{\langle p_t, \widehat{u}_{t,:,n} - u_{:,n} \rangle}{\langle p_t, u_{:,n} \rangle^{\frac{N-1}{N}}} \leq \mathcal{O}\left(\frac{1}{\langle p_t, u_{:,n} \rangle^{\frac{N-1}{N} - \frac{1}{2}}} \sum_{i=1}^{K} \sqrt{\frac{p_{t,i}}{N_{t,i}}} \right) \leq \mathcal{O}\left(\sigma^{\frac{1}{2} - \frac{N-1}{N}} \sum_{i=1}^{K} \sqrt{\frac{p_{t,i}}{N_{t,i}}} \right), \quad (9)$$

where the last inequality is due to the condition $\langle p_t, u_{:,n} \rangle \geq \sigma$ for all $n \in [N]$. Finally, combining Eq. (7), Eq. (8), Eq. (9), followed by direct calculations, we show that

$$\mathbb{E}\left[\mathrm{Reg}_{\mathrm{sto}} \right] \leq \sum_{t=1}^{T} \left| \mathrm{NSW}(\widehat{u}_t^\top p_t) - \mathrm{NSW}(u^\top p_t) \right| \leq \widetilde{\mathcal{O}}\left(T\sigma^{\frac{1}{N}} + K^{\frac{1}{N}} T^{\frac{N-1}{N}} + \sigma^{\frac{1}{2} - \frac{N-1}{N}} K\sqrt{T} \right).$$

Picking the optimal choice of σ finishes the proof.

We now highlight the importance of using a Bernstein-type confidence width in Eq. (4): if the standard Hoeffding-type confidence width is used instead, then one can only obtain $\widehat{u}_{t,i,n} - u_{i,n} \leq \mathcal{O}(\sqrt{\frac{1}{N_{t,i}}})$, and consequently, Eq. (8) can only be bounded by $\mathcal{O}\left(\sigma^{-\frac{N-1}{N}} \sqrt{KT} \right)$ after taking summation over $t \in [T]$. This eventually leads to a worse regret bound of $\widetilde{\mathcal{O}}(K^{\frac{1}{2N}} T^{\frac{2N-1}{2N}})$.

3.2 Lower Bound

Next, we prove an $\widetilde{\Omega}(T^{\frac{N-1}{N}})$ lower bound for this setting. This not only shows that the regret bound we achieve via Algorithm 1 is tight in T, but also highlights the difference and difficulty of learning with NSW compared to learning with $\mathrm{NSW}_{\mathrm{prod}}$, since in the latter case, $\Theta(\sqrt{T})$ regret is minimax optimal [Hossain et al., 2021, Jones et al., 2023].

Theorem 3.2. *In the bandit feedback setting, for any algorithm, there exists a stochastic environment in which the expected regret (with respect to NSW) of this algorithm is $\Omega\left(\frac{(\log K)^3}{N^3} \cdot K^{\frac{1}{N}} T^{\frac{N-1}{N}} \right)$ for $N \geq \log K$ and sufficiently large T.*

We defer the full proof to Appendix A.2 and discuss the hard instance used in the proof below. First, the mean utility vector $u_{:,n}$ for each agent $n \geq 2$ is a constant vector $\mathbf{1}$. This makes the problem

equivalent to a one-agent problem, but with $\langle p, u_{:,1}\rangle^{1/N}$ as the reward, instead of $\langle p, u_{:,1}\rangle$ as in standard stochastic K-armed bandits.

Then, for the first agent, different from the standard K-armed bandits, where the hardest instance is to hide one arm with a slightly better expected reward of $\frac{1}{2} + \sqrt{K/T}$ among other $K - 1$ arms with expected reward of exactly $\frac{1}{2}$,[2] we hide one arm with expected reward K/T among other $K - 1$ arms with exactly 0 reward (so overall the rewards are shifted towards 0 but with a smaller gap between the best arm and the others). By standard information theory arguments, within T rounds the learner cannot distinguish the best arm from the others. Therefore, the best strategy she can apply is to pick a uniform distribution over actions, suffering $\Omega((1 - K^{-\frac{1}{N}}) \cdot (K/T)^{\frac{1}{N}}) = \widetilde{\Omega}(K^{\frac{1}{N}}T^{-\frac{1}{N}})$ regret per round and leading to $\widetilde{\Omega}(K^{\frac{1}{N}}T^{\frac{N-1}{N}})$ regret overall.

4 Adversarial Environments

Now that we have a complete answer for the stochastic setting, we move on to consider the adversarial case where each u_t is chosen arbitrarily, a multi-agent generalization of the expert problem (full-information feedback) [Freund and Schapire, 1997] and the adversarial multi-armed bandit problem (bandit feedback) [Auer et al., 2002b]. There are no prior studies on this problem, be it with $f = \text{NSW}$ or $f = \text{NSW}_{\text{prod}}$, as far as we know.

4.1 Impossibility Results with Bandit Feedback

We start by considering the bandit feedback setting. As mentioned in Section 2, even though NSW is a concave function, our problem is not an instance of Bandit Convex Optimization, since we can only observe $u_{t,i_t,:}$ instead of $\text{NSW}(u_t^\top p_t)$ at the end of round t. Somewhat surprisingly, this slight difference in the feedback in fact makes a sharp separation in learnability — while $\mathcal{O}(\sqrt{T})$ regret is achievable in BCO, we prove that $o(T)$ regret is impossible in our problem.

Before showing the theorem and its proof, we first give high level ideas on the construction of the hard environments. Specifically, we consider the environment with 2 agents, 2 arms, and binary utility matrix $u_t \in \{0, 1\}^{2\times 2}$. Similar to the hard instance in the stochastic environment, we set $u_{t,:,2} = \mathbf{1}$, reducing the problem to a single-agent one. For the first agent, we let $u_{t,:,1}$ at each round t be i.i.d. drawn from a stationary distribution over the 4 binary utility vectors $\{(0,0), (0,1), (1,0), (1,1)\}$. Then, we construct two different distributions, $\mathcal{E}$ and $\mathcal{E}'$, over these 4 binary utility vectors satisfying that: 1) the distribution of the learner's observation is identical for $\mathcal{E}$ and $\mathcal{E}'$; 2) the optimal strategy for $\mathcal{E}$ and $\mathcal{E}'$ are significantly different. The first property guarantees that no algorithm can distinguish these two environments, while the second property ensures that there is no one single strategy that can perform well in both environments. Formally, we prove the following theorem.

Theorem 4.1. *In the bandit feedback setting, for any algorithm, there exists an adversarial environment such that* $\mathbb{E}[\text{Reg}_{\text{adv}}] = \Omega(T)$ *for* $f = \text{NSW}$.

Proof. As sketched earlier, we consider two different environments with 2 agents, 2 arms, and binary utility matrices $u_t \in \{0, 1\}^{2\times 2}$, $t \in [T]$. In both environments, we have $u_{t,:,2} = \mathbf{1}$. Next, we construct two different distributions from which $u_{t,:,1}$ is potentially drawn from, $\mathcal{E}$ and $\mathcal{E}'$, over $\{(0,0), (0,1), (1,0), (1,1)\}$. Specifically, $\mathcal{E}$ is characterized by $(q_{00}, q_{01}, q_{10}, q_{11}) = (\frac{4}{10}, \frac{2}{10}, \frac{1}{10}, \frac{3}{10})$, where q_{xy} is the probability of the vector (x, y) in $\mathcal{E}$; $\mathcal{E}'$ is characterized by $(q'_{00}, q'_{01}, q'_{10}, q'_{11}) = (\frac{3}{10}, \frac{3}{10}, \frac{2}{10}, \frac{2}{10})$, where q'_{xy} is the probability of vector (x, y) in $\mathcal{E}'$. With a slight abuse of notation, we write $u \sim \mathcal{E}$ for a matrix $u \in \{0, 1\}^{2\times 2}$ if $u_{:,1}$ is drawn from $\mathcal{E}$ and $u_{:,2} = \mathbf{1}$; the same for $\mathcal{E}'$.

We argue that the learner's observations are equivalent in distribution in $\mathcal{E}$ and $\mathcal{E}'$, since the marginal distributions of the utility of each action are the same. Specifically,

- When action 1 is chosen, the distributions of the learner's observation in both $\mathcal{E}$ and $\mathcal{E}'$ are a Bernoulli random variable with mean $q_{10} + q_{11} = q'_{10} + q'_{11} = \frac{4}{10}$;

---

[2]One can show that $\Theta(\sqrt{T})$ regret is possible in this environment, thus not suitable for our purpose.

- When action 2 is chosen, the distributions of the learner's observation in both $\mathcal{E}$ and $\mathcal{E}'$ are a Bernoulli random variable with mean $q_{01} + q_{11} = q'_{01} + q'_{11} = \frac{5}{10}$.

Direct calculation shows $p_\star = \operatorname{argmax}_{p \in \Delta_K} \mathbb{E}_{u \sim \mathcal{E}} \left[\mathrm{NSW}(u^\top p) \right] = \left(\frac{q_{10}^2}{q_{01}^2 + q_{10}^2}, \frac{q_{01}^2}{q_{01}^2 + q_{10}^2} \right) = (0.2, 0.8)$ and $p'_\star = \operatorname{argmax}_{p \in \Delta_K} \mathbb{E}_{u \sim \mathcal{E}'} \left[\mathrm{NSW}(u^\top p) \right] = \left(\frac{q_{10}^{'2}}{q_{10}^{'2} + q_{01}^{'2}}, \frac{q_{01}^{'2}}{q_{10}^{'2} + q_{01}^{'2}} \right) = \left(\frac{4}{13}, \frac{9}{13} \right)$, which are constant apart from each other. Pick a threshold value $\theta = \frac{33}{130} \in (0.2, \frac{4}{13})$. Direct calculation shows that for a strategy p with $p_1 \geq \theta$, we have $\mathbb{E}_{u \sim \mathcal{E}}[\mathrm{NSW}(u^\top p_\star) - \mathrm{NSW}(u^\top p)] \geq \Delta$ where $\Delta = \frac{1}{500}$; similarly, for a strategy p with $p_1 < \theta$, we have $\mathbb{E}_{u \sim \mathcal{E}'}[\mathrm{NSW}(u^\top p_\star) - \mathrm{NSW}(u^\top p)] \geq \Delta$ as well. Now, given an algorithm, let $\alpha_\mathcal{E}$ be the probability that the number of rounds $p_{t,1} \geq \theta$ is larger than $\frac{T}{2}$ under environment $\mathcal{E}$, and $\bar{\alpha}_{\mathcal{E}'}$ be the probability of the complement of this event under environment $\mathcal{E}'$. We have,

$$\mathbb{E}_\mathcal{E}[\mathrm{Reg}_{\mathrm{adv}}] \geq \mathbb{E}_\mathcal{E} \left[\sum_{t=1}^T \mathrm{NSW}(u_t^\top p_\star) - \sum_{t=1}^T \mathrm{NSW}(u_t^\top p_t) \right] \geq \frac{\alpha_\mathcal{E} T \Delta}{2},$$

$$\mathbb{E}_{\mathcal{E}'}[\mathrm{Reg}_{\mathrm{adv}}] \geq \mathbb{E}_{\mathcal{E}'} \left[\sum_{t=1}^T \mathrm{NSW}(u_t^\top p'_\star) - \sum_{t=1}^T \mathrm{NSW}(u_t^\top p_t) \right] \geq \frac{\bar{\alpha}_{\mathcal{E}'} T \Delta}{2}.$$

Finally, since the feedback for the algorithm is the same in distribution in these two environments, we know $\alpha_\mathcal{E} + \bar{\alpha}_{\mathcal{E}'} = 1$, and thus

$$\max\{\mathbb{E}_\mathcal{E}[\mathrm{Reg}_{\mathrm{adv}}], \mathbb{E}_{\mathcal{E}'}[\mathrm{Reg}_{\mathrm{adv}}]\} \geq \frac{\mathbb{E}_\mathcal{E}[\mathrm{Reg}_{\mathrm{adv}}] + \mathbb{E}_{\mathcal{E}'}[\mathrm{Reg}_{\mathrm{adv}}]}{2} \geq \frac{(\alpha_\mathcal{E} + \bar{\alpha}_{\mathcal{E}'}) T \Delta}{4} = \Omega(T),$$

which finishes the proof. $\qquad \square$

In fact, by a similar but more involved construction (that actually requires using two agents in a non-trivial way), the same impossibility result also holds for $f = \mathrm{NSW}_{\mathrm{prod}}$; see Appendix B.1. We remark that non-linearity of f in these results plays an important role in the hard instance construction, since otherwise, the optimal strategy for $\mathcal{E}$ and $\mathcal{E}'$ will be the same as they both induce the same marginal distributions.

4.2 Full-Information Feedback

To sidestep the impossibility result due to the bandit feedback, we shift our focus to the full-information feedback model, where the learner observes the entirety of the utility matrix u_t at the end of round t. As mentioned, this corresponds to a multi-agent generalization of the well-known expert problem [Freund and Schapire, 1997]. We propose several algorithms for this setting, showing that the richer feedback not only makes learning possible but also leads to much lower regret.

4.2.1 FTRL with Log-Barrier Regularizer

When f is concave, our problem is in fact also an instance of the well-known Online Convex Optimization (OCO) [Zinkevich, 2003]. However, standard OCO algorithms such as Online Gradient Descent, an instance of the more general Follow-the-Regularized-Leader algorithm with a ℓ_2 regularizer, require the utility function to also be Lipschitz and thus cannot be directly applied to learning NSW. Nevertheless, we will show that using a different regularizer that induces more stability than the ℓ_2 regularizer can resolve this issue.

More specifically, the FTRL algorithm is shown in Algorithm 2, which predicts at time t the distribution $p_t = \operatorname{argmin}_{p \in \Delta_K} \langle p, -\sum_{s=1}^{t-1} \nabla f(u_s^\top p_s) \rangle + \frac{1}{\eta} \psi(p)$ for some learning rate η and some strongly convex regularizer ψ. Standard analysis shows that the regret of FTRL contains two terms: the regularization penalty term that is of order $1/\eta$ and the stability term that is of order $\eta \sum_t \| \nabla f(u_t^\top p_t) \|_{\nabla^{-2} \psi(p_t)}^2$ where we use the notation $\|a\|_M = \sqrt{a^\top M a}$. To deal with the lack the Lipschitzness, that is, the potentially large $\nabla f(u_t^\top p_t)$, we need to find a regularizer ψ so that the induced local norm $\| \nabla f(u_t^\top p_t) \|_{\nabla^{-2} \psi(p_t)}$ is always reasonably small despite $\nabla f(u_t^\top p_t)$ being large (in ℓ_2 norm for example).

Algorithm 2 FTRL for N-agent K-armed SWF maximization with full-info feedback
---
Inputs: a SWF f, a learning rate $\eta > 0$, and a strongly convex regularizer $\psi : \Delta_K \to \mathbb{R}$.
for $t = 1, 2, \ldots, T$ **do**
    Play $p_t = \operatorname{argmin}_{p \in \Delta_K} \langle p, -\sum_{s=1}^{t-1} \nabla f(u_s^\top p_s) \rangle + \frac{1}{\eta} \psi(p)$.
    Observe u_t.
end

It turns out that the log-barrier regularizer, $\psi(p) = -\sum_{i=1}^{K} \log p_i$, ensures such a property. In fact, it induces a small local norm not just for NSW, but also for a broad family of SWFs as long as they are concave and *Pareto optimal* — an SWF $f : [0,1]^N \to [0,1]$ is Pareto optimal if for two utility vectors x and y such that $x_n \geq y_n$ for all $i \in [N]$, we have $f(x) \geq f(y)$. NSW is clearly in this family, and there are many other standard fairness measures that fall into this class; see Appendix B.2.1. For any SWF in this family, we prove the following regret bound, which remarkably has no dependence on the number of agents N at all.

Theorem 4.2. *For any $f : [0,1]^N \to [0,1]$ that is concave and Pareto optimal, Algorithm 2 with the log-barrier regularizer $\psi(p) = -\sum_{i=1}^{K} \log p_i$ and $\eta = \sqrt{\frac{K \log T}{T}}$ guarantees $\mathrm{Reg}_{\mathrm{adv}} = \mathcal{O}(\sqrt{KT \log T})$.*

Proof Sketch. Using the concrete form of ψ, it is clear that the local norm $\|\nabla f(u_t^\top p_t)\|_{\nabla^{-2}\psi(p_t)}^2$ simplifies to $\sum_{i=1}^{K} p_{t,i}^2 [\nabla f(u_t^\top p_t)]_i^2 \leq \langle p_t, \nabla f(u_t^\top p_t) \rangle^2$, where the inequality is due to $[\nabla f(u_t^\top p_t)]_i \geq 0$ implied by Pareto optimality. Furthermore, by concavity, we have $\langle p_t, \nabla f(u_t^\top p_t) \rangle \leq f(u_t^\top p_t) - f(0) \leq 1$, and thus the local norm at most 1. The rest of the proof is by direct calculation. $\qquad\square$

4.2.2 FTRL with Tsallis Entropy Regularizer

In fact, when $f = \mathrm{NSW}$, using the special structure of the welfare function, we find yet another regularizer that ensures a small $\mathcal{O}(N)$ local norm, with the benefit of having smaller dependence on K for the penalty term.

Theorem 4.3. *For $f = \mathrm{NSW}$, Algorithm 2 with the Tsallis entropy regularizer $\psi(p) = \frac{1 - \sum_{i=1}^{K} p_i^\beta}{1 - \beta}$, $\beta = \frac{2}{N}$, and the optimal choice of η guarantees $\mathrm{Reg}_{\mathrm{adv}} = \widetilde{\mathcal{O}}(K^{\frac{1}{2} - \frac{1}{N}} \sqrt{NT})$.*

The proof is more involved and is deferred to Appendix B.2.3. While the regret in Theorem 4.3 suffers polynomial dependence on N, it has better dependence on K compared to Theorem 4.2, and is thus more preferable when K is much larger than N.

4.2.3 Logarithmic Regret for a Special Case

Finally, we discuss a special case with $f = \mathrm{NSW}$ where logarithmic regret is possible. This is based on a simple observation that when there is one agent who is indifferent about the learner's choice (that is, the agent's utility is the same for all arms for this round), then $-\mathrm{NSW}$ is not only convex, but also exp-concave, a stronger curvature property. Therefore, by applying known results, specifically the EWOO algorithm [Hazan et al., 2007], we achieve the following result.

Theorem 4.4. *Fix $f = \mathrm{NSW}$. Suppose that for each time t, there is a set of agents $A_t \subseteq [N]$ such that $|A_t| \geq M$ and $u_{t,:,n} = c_{t,n} \mathbf{1}$ with $c_{t,n} \geq 0$ for each agent $n \in A_t$. Then the EWOO algorithm guarantees $\mathrm{Reg}_{\mathrm{adv}} = \mathcal{O}\left(\frac{N-M}{M} \cdot K \log T\right)$.*

The proof, which verifies the exp-concavity of $-\mathrm{NSW}$ in this special case, can be found in Appendix B.2.4. We note that the reason that we apply EWOO instead of Online Newton Step, another algorithm discussed in [Hazan et al., 2007] for exp-concave losses, is that the latter requires Lipschizness (which, again, is not satisfied by NSW).

9

5 Conclusion

In this work, motivated by recent research on social welfare maximization for the problem of multi-agent multi-armed bandits, we consider a variant with the arguably more natural version of Nash social welfare as the objective function, and develop multiple algorithms and regret upper/lower bounds in different settings (stochastic versus adversarial and full-information versus bandit feedback). Our results show a sharp separation between our problem and previous settings, including the heavily studied Bandit Convex Optimization problem.

There are many interesting future directions. First, in the stochastic bandit setting, we have only shown the tight dependence on T, so what about K and N? Second, is there a more general strategy/analysis that works for different social welfare functions (similar to our result in Theorem 4.2)? Taking one step further, similar to the recent research on "omniprediction" [Gopalan et al., 2022], is there one single algorithm that works for a class of social welfare functions simultaneously?

Acknowledgments and Disclosure of Funding

References

Jacob D Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

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

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *The Journal of Machine Learning Research*, 11:2785–2836, 2010.

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

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

Jackie Baek and Vivek Farias. Fair exploration via axiomatic bargaining. *Advances in Neural Information Processing Systems*, 34:22034–22045, 2021.

Siddharth Barman, Sanath Kumar Krishnamurthy, and Rohit Vaish. Finding fair and efficient allocations. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 557–574, 2018.

Siddharth Barman, Arindam Khan, Arnab Maiti, and Ayush Sawarni. Fairness and welfare quantification for regret in multi-armed bandits. AAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i6.25829.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop and Conference Proceedings, 2011.

Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, and Shie Mannor. Multi-objective bandits: Optimizing the generalized gini index. In *International Conference on Machine Learning*, pages 625–634. PMLR, 2017.

Ioannis Caragiannis, David Kurokawa, Hervé Moulin, Ariel D. Procaccia, Nisarg Shah, and Junxing Wang. The unreasonable fairness of maximum nash welfare. *ACM Trans. Econ. Comput.*, 7(3), sep 2019. ISSN 2167-8375. doi: 10.1145/3355902.

Violet Xinying Chen and J.N. Hooker. A guide to formulating fairness in an optimization model. *Annals of Operation Research*, 326:581–619, 2023.

Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, pages 181–190. PMLR, 2020.

Richard Cole and Vasilis Gkatzelis. Approximating the nash social welfare with indivisible items. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 371–380, 2015.

Madalina M Drugan and Ann Nowe. Designing multi-objective multi-armed bandits algorithms: A study. In *The 2013 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2013.

Dylan J Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, and Eva Tardos. Learning in games: Robustness of fast convergence. *Advances in Neural Information Processing Systems*, 29, 2016.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Jugal Garg, Pooja Kulkarni, and Rucha Kulkarni. Approximating nash social welfare under submodular valuations through (un) matchings. *ACM Transactions on Algorithms*, 19(4):1–25, 2023.

Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. *Advances in neural information processing systems*, 31, 2018.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, volume 215, pages 79:1–79:21, 2022.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Safwan Hossain, Evi Micha, and Nisarg Shah. Fair algorithms for multi-agent multi-armed bandits. *Advances in Neural Information Processing Systems*, 34:24005–24017, 2021.

Matthew Jones, Huy Nguyen, and Thy Nguyen. An efficient algorithm for fair multi-agent multi-armed bandit with low regret. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8159–8167, 2023.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Mamoru Kaneko and Kenjiro Nakamura. The nash social welfare function. *Econometrica*, 47:423–435, 1979.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. *Journal of the ACM (JACM)*, 66(4):1–77, 2019.

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

Tor Lattimore. Bandit convex optimisation. *arXiv preprint arXiv:2402.06535*, 2024.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

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

Wenzheng Li and Jan Vondrák. Estimating the nash social welfare for coverage and other submodular valuations. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1119–1130. SIAM, 2021.

Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*, 2017.

Haipeng Luo. Lecture note 13, Introduction to Online Learning. 2017. URL https://haipeng-luo.net/courses/CSCI699/lecture13.pdf.

Hervé Moulin. *Fair division and collective welfare*. MIT press, 2004.

John F Nash. The bargaining problem. *Econometrica*, 18(2):155–162, 1950.

Ayush Sawarni, Soumyabrata Pal, and Siddharth Barman. Nash regret guarantees for linear bandits. *Advances in Neural Information Processing Systems*, 36, 2024.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

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

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

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

A   Omitted Details in Section 3

In this section, we provide the omitted proofs for the results in Section 3. In Appendix A.1, we provide the proof for Theorem 3.1 and in Appendix A.2, we provide the proof for Theorem 3.2.

A.1   Proof of Theorem 3.1

To prove Theorem 3.1, we first consider the following two events.

Event 1. *For all $t \in \{KN_0 + 1, \ldots, T\}$ and $i \in [K]$,*

$$N_{t,i} \geq N_0 + \frac{1}{2} \sum_{\tau = KN_0}^{t} p_{\tau,i} - 18 \log(KT),$$

where $N_{t,i}$'s and $p_{t,i}$'s are defined in Algorithm 1.

Event 2. *For all $t \in \{KN_0 + 1, \ldots, T\}$, $i \in [K]$, and $n \in [N]$,*

$$u_{i,n} \leq \widehat{u}_{t,i,n} \leq u_{i,n} + 8\sqrt{\frac{u_{i,n} \log(NKT^2)}{N_{t,i}}} + \frac{15 \log(NKT^2)}{N_{t,i}},$$

where $N_{t,i}$'s are defined in Algorithm 1.

As we prove in Lemma A.1 and Lemma A.2, Event 1 and Event 2 hold with probability at least $1 - \frac{1}{T}$. Now we prove Theorem 3.1. For convenience, we restate the theorem as follows.

Theorem 3.1. *With $N_0 = 1 + 18 \log KT$, Algorithm 1 guarantees $\mathbb{E}[\text{Reg}_{\text{sto}}] = \widetilde{\mathcal{O}}(K^{\frac{2}{N}} T^{\frac{N-1}{N}} + K)$.*

Proof. Let $p^\star = \operatorname{argmax}_{p \in \Delta_K} \text{NSW}(u^\top p)$. According to a standard regret decomposition for UCB-type algorithms, we know that Reg_{sto} can be upper bounded as follows:

$$\mathbb{E}[\text{Reg}_{\text{sto}}]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} \left(\text{NSW}(u^\top p^\star) - \text{NSW}(u^\top p_t)\right)\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} \left(\text{NSW}(u^\top p^\star) - \text{NSW}(u^\top p_t)\right) \;\middle|\; \text{Event 1 and Event 2 hold}\right] + 2$$

$$\text{(according to Lemma A.1 and Lemma A.2)}$$

$$\leq \mathbb{E}\left[\sum_{t=KN_0+1}^{T} \left(\text{NSW}(u^\top p^\star) - \text{NSW}(\widehat{u}_t^\top p^\star)\right) \;\middle|\; \text{Event 1 and Event 2 hold}\right] + KN_0 + 2$$

$$+ \mathbb{E}\left[\sum_{t=KN_0+1}^{T} \left(\text{NSW}(\widehat{u}_t^\top p^\star) - \text{NSW}(\widehat{u}_t^\top p_t)\right) \;\middle|\; \text{Event 1 and Event 2 hold}\right]$$

$$+ \mathbb{E}\left[\sum_{t=KN_0+1}^{T} \left(\text{NSW}(\widehat{u}_t^\top p_t) - \text{NSW}(u^\top p_t)\right) \;\middle|\; \text{Event 1 and Event 2 hold}\right]$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T} \left(\text{NSW}(\widehat{u}_t^\top p^\star) - \text{NSW}(\widehat{u}_t^\top p_t)\right)\right] + \mathbb{E}\left[\sum_{t=1}^{T} \left(\text{NSW}(\widehat{u}_t^\top p_t) - \text{NSW}(u^\top p_t)\right)\right] + KN_0 + 2$$

$$\text{(based on Event 2)}$$

$$\leq \mathbb{E}\left[\sum_{t=KN_0+1}^{T} \left(\text{NSW}(\widehat{u}_t^\top p_t) - \text{NSW}(u^\top p_t)\right) \;\middle|\; \text{Event 1 and Event 2 hold}\right] + KN_0 + 2.$$

$$\text{(based on the definition of } p_t)$$

In the following, we bound the first term

$$\mathbb{E}\left[\sum_{t=KN_0+1}^{T}\left(\text{NSW}(\widehat{u}_t^\top p_t) - \text{NSW}(u^\top p_t)\right) \,\middle|\, \text{Event 1 and Event 2 hold}\right].$$

As discussed in Section 3.1, we consider two cases. First, consider the set of rounds $\mathcal{T}_\sigma$ such that for all $t \in \mathcal{T}_\sigma$ there exists at least one $n \in [N]$ such that $\langle p_t, u_{:,n}\rangle \leq \sigma$ for some σ that we will specify later. Denote such n to be n_t (if there are multiple such n's, we pick an arbitrary one). According to Event 2, we know that for all $i \in [K]$,

$$u_{i,n_t} \leq \widehat{u}_{t,i,n_t} \leq u_{i,n_t} + 8\sqrt{\frac{u_{i,n_t}\log(NKT^2)}{N_{t,i}}} + \frac{15\log(NKT^2)}{N_{t,i}} \leq 2u_{i,n_t} + \mathcal{O}\left(\frac{\log(NKT^2)}{N_{t,i}}\right),$$

where the last inequality is because of AM-GM inequality. Therefore, we know that

$$\langle p_t, \widehat{u}_{t,:,n_t}\rangle \leq 2\langle p_t, u_{:,n_t}\rangle + \widetilde{\mathcal{O}}\left(\sum_{j=1}^{K}\frac{p_{t,j}}{N_{t,j}}\right) \leq 2\sigma + \widetilde{\mathcal{O}}\left(\sum_{j=1}^{K}\frac{p_{t,j}}{N_{t,j}}\right).$$

Now consider $\sum_{t\in\mathcal{T}_\sigma}\left(\text{NSW}(\widehat{u}_t^\top p_t) - \text{NSW}(u^\top p_t)\right)$. Direct calculation shows that

$$\sum_{t\in\mathcal{T}_\sigma}\left(\text{NSW}(\widehat{u}_t^\top p_t) - \text{NSW}(u^\top p_t)\right)$$

$$\leq \sum_{t\in\mathcal{T}_\sigma}\text{NSW}(\widehat{u}_t^\top p_t) \qquad\qquad (\text{since } \text{NSW}(u^\top p_t) \geq 0)$$

$$\leq \sum_{t\in\mathcal{T}_\sigma}\langle p_t, \widehat{u}_{t,:,n_t}\rangle^{\frac{1}{N}}$$

$$\leq 2|\mathcal{T}_\sigma|\cdot\sigma^{\frac{1}{N}} + \sum_{t=1}^{T}\widetilde{\mathcal{O}}\left(\left(\sum_{j=1}^{K}\frac{p_{t,j}}{N_{t,j}}\right)^{\frac{1}{N}}\right) \qquad (\text{since } (a+b)^{\frac{1}{N}} \leq a^{\frac{1}{N}} + b^{\frac{1}{N}})$$

$$\leq 2|T_\sigma|\cdot\sigma^{\frac{1}{N}} + T^{\frac{N-1}{N}}\widetilde{\mathcal{O}}\left(\left(\sum_{t=1}^{T}\sum_{j=1}^{K}\frac{p_{t,j}}{N_{t,j}}\right)^{\frac{1}{N}}\right) \qquad (\text{Hölder's inequality})$$

$$\leq 2T\cdot\sigma^{\frac{1}{N}} + \widetilde{\mathcal{O}}\left(K^{\frac{1}{N}}\cdot T^{\frac{N-1}{N}}\right). \qquad\qquad (\text{using Lemma A.3})$$

Now consider the regret within $t \in \{NK_0 + 1, \ldots, T\}\backslash\mathcal{T}_\sigma$, in which we have $\langle p_t, u_{:,n}\rangle \geq \sigma$ for all $n \in [N]$. In this case, we bound $\sum_{t\notin\mathcal{T}_\sigma}\left(\text{NSW}(\widehat{u}_t^\top p_t) - \text{NSW}(u^\top p_t)\right)$ as follows:

$$\sum_{t\notin\mathcal{T}_\sigma}\left(\text{NSW}(\widehat{u}_t^\top p_t) - \text{NSW}(u^\top p_t)\right)$$

$$\leq \sum_{t\notin\mathcal{T}_\sigma}\sum_{n\in[N]}\left[\langle p_t, \widehat{u}_{t,:,n}\rangle^{\frac{1}{N}} - \langle p_t, u_{:,n}\rangle^{\frac{1}{N}}\right] \qquad (\text{using Lemma C.1 and Event 2})$$

$$= \sum_{t\notin\mathcal{T}_\sigma}\sum_{n\in[N]}\frac{\langle p_t, \widehat{u}_{t,:,n} - u_{:,n}\rangle}{\sum_{k=0}^{N-1}\langle p_t, \widehat{u}_{t,:,n}\rangle^{\frac{k}{N}}\langle p_t, u_{:,n}\rangle^{\frac{N-1-k}{N}}}$$

$$\leq \sum_{t\notin\mathcal{T}_\sigma}\sum_{n\in[N]}\frac{\langle p_t, \widehat{u}_{t,:,n} - u_{:,n}\rangle}{N\langle p_t, u_{:,n}\rangle^{\frac{N-1}{N}}} \qquad (\text{since } \widehat{u}_{t,i,n} \geq u_{i,n} \text{ for all } i, t, n \text{ based on Event 2})$$

$$\leq \sum_{t\notin\mathcal{T}_\sigma}\sum_{n\in[N]}\frac{\sum_{j=1}^{K}p_{t,j}\left(8\sqrt{\frac{u_{n,j}\log(NKT^2)}{N_{t,j}}} + \frac{8\log(NKT^2)}{N_{t,j}}\right)}{N\langle p_t, u_{:,n}\rangle^{\frac{N-1}{N}}} \qquad (\text{using Event 2})$$

14

$$\leq \sum_{t \notin \mathcal{T}_\sigma} \sum_{n \in [N]} \left(\frac{\sum_{j=1}^{K} 8\sqrt{\frac{p_{t,j} \log(NKT^2)}{N_{t,j}}}}{N \langle p_t, u_{:,n} \rangle^{\frac{N-1}{N} - \frac{1}{2}}} + \frac{\sum_{j=1}^{K} \frac{8 p_{t,j} \log(NKT^2)}{N_{t,j}}}{N \langle p_t, u_{:,n} \rangle^{\frac{N-1}{N}}} \right)$$

$$\text{(since } \sqrt{\langle p_t, u_{:,n} \rangle} \geq \sqrt{p_{t,i} u_{i,n}} \text{ for all } i \in [K])$$

$$\leq \widetilde{\mathcal{O}} \left(\frac{1}{N \sigma^{\frac{N-1}{N} - \frac{1}{2}}} \sum_{t \in T} \sum_{n \in [N]} \sum_{j=1}^{K} \sqrt{\frac{p_{t,j}}{N_{t,j}}} + \frac{1}{N \sigma^{\frac{N-1}{N}}} \sum_{t=1}^{T} \sum_{n \in [N]} \sum_{j=1}^{K} \frac{p_{t,j}}{N_{t,j}} \right)$$

$$\leq \widetilde{\mathcal{O}} \left(\sigma^{\frac{1}{2} - \frac{N-1}{N}} K \sqrt{T} + K \cdot \sigma^{-\frac{N-1}{N}} \right),$$

where the last inequality is because Lemma A.3. Combining the regret for both parts, we know that

$$\mathbb{E}[\text{Reg}_{\text{sto}}] \leq \widetilde{\mathcal{O}} \left(K^{\frac{1}{N}} T^{\frac{N-1}{N}} + T \cdot \sigma^{\frac{1}{N}} + \sigma^{\frac{1}{2} - \frac{N-1}{N}} K \sqrt{T} + K \sigma^{-\frac{N-1}{N}} + K \right).$$

Picking the optimal σ leads to the expected regret bounded by $\mathbb{E}[\text{Reg}_{\text{sto}}] \leq \widetilde{\mathcal{O}} \left(K^{\frac{2}{N}} T^{\frac{N-1}{N}} + K \right)$.

$\square$

Lemma A.1. *Event 1 happens with probability at least* $1 - \frac{1}{T}$.

Proof. According to Algorithm 1, we know that $N_{(KN_0+1),i} = N_0$ for each $i \in [K]$. Consider the case when $t \geq KN_0 + 1$. According to Freedman's inequality (Lemma C.3), we have with probability at least $1 - \delta$, for a fixed $t \geq KN_0 + 1$,

$$\sum_{\tau=KN_0+1}^{t} \mathbb{1}\{i_\tau = i\} \geq \sum_{\tau=KN_0+1}^{t} p_{\tau,i} - 2\sqrt{\sum_{\tau=KN_0}^{t} p_{\tau,i} \log(1/\delta)} - \log(1/\delta)$$

$$\geq \frac{1}{2} \sum_{\tau=KN_0+1}^{t} p_{\tau,i} - 9 \log(1/\delta).$$

Therefore, we know that with probability at least $1 - \delta$, for a fixed $t \geq KN_0 + 1$,

$$N_{t,i} = N_0 + \sum_{\tau=KN_0+1}^{t} \mathbb{1}\{i_\tau = i\} \geq N_0 + \frac{1}{2} \sum_{\tau=KN_0+1}^{t} p_{\tau,i} - 9 \log(1/\delta).$$

Picking $\delta = \frac{1}{KT^2}$ and taking a union bound over all $i \in [K]$ and $KN_0 + 1 \leq t \leq T$ reach the result. $\square$

Lemma A.2. *Event 2 happens with probability at least* $1 - \frac{1}{T}$.

Proof. According to Freedman's inequality Lemma C.3, applying a union bound over $t \in [T]$, $i \in [K]$, and $n \in [N]$, we know that with probability at least $1 - \delta$, for all $t \in [T]$, $i \in [K]$, and $n \in [N]$,

$$|\bar{u}_{t,i,n} - u_{i,n}| \leq 2\sqrt{\frac{u_{i,n} \log(NKT/\delta)}{N_{t,i}}} + \frac{\log(NKT/\delta)}{N_{t,i}}. \tag{10}$$

Solving the inequality with respect to $u_{i,n}$, we know that

$$\sqrt{u_{i,n}} \leq \sqrt{\frac{\log(NKT/\delta)}{N_{t,i}}} + \sqrt{\bar{u}_{t,i,n} + \frac{2\log(NKT/\delta)}{N_{t,i}}}$$

$$\leq \sqrt{2\bar{u}_{t,i,n} + \frac{6\log(NKT/\delta)}{N_{t,i}}}, \qquad \text{(using AM-GM inequality)}$$

$$\bar{u}_{t,i,n} \leq \left(\sqrt{u_{i,n}} + \sqrt{\frac{\log(NKT/\delta)}{N_{t,i}}} \right)^2$$

$$\leq 2u_{i,n} + \frac{2\log(NKT/\delta)}{N_{t,i}}. \qquad \text{(using AM-GM inequality)}$$

Using the above inequality and picking $\delta = \frac{1}{T}$, we can lower bound $\widehat{u}_{t,i,n}$ as follows:

$$\widehat{u}_{t,i,n} = \bar{u}_{t,i,n} + 4\sqrt{\frac{\bar{u}_{t,i,n}\log(NKT^2)}{N_{t,i}}} + \frac{8\log(NKT^2)}{N_{t,i}}$$

$$\geq \bar{u}_{t,i,n} + 4\sqrt{\frac{\log(NKT^2)}{N_{t,i}}\left(\frac{u_{i,n}}{2} - \frac{3\log(NKT^2)}{N_{t,i}}\right)} + \frac{8\log(NKT^2)}{N_{t,i}}$$

$$\geq \bar{u}_{t,i,n} + 4\sqrt{\frac{u_{i,n}\log(NKT^2)}{2N_{t,i}}} + \frac{(8 - 4\sqrt{3})\log(NKT^2)}{N_{t,i}}$$

$$\text{(using } \sqrt{a-b} \geq \sqrt{a} - \sqrt{b} \text{ for } a,b \geq 0)$$

$$\geq \bar{u}_{t,i,n} + 2\sqrt{\frac{u_{i,n}\log(NKT^2)}{N_{t,i}}} + \frac{\log(NKT^2)}{N_{t,i}}$$

$$\geq u_{i,n},$$

where the last inequality uses Eq. (10). To upper bound $\widehat{u}_{t,i,n}$, we have

$$\widehat{u}_{t,i,n} = \bar{u}_{t,i,n} + 4\sqrt{\frac{\bar{u}_{t,i,n}\log(NKT^2)}{N_{t,i}}} + \frac{8\log(NKT^2)}{N_{t,i}}$$

$$\leq u_{i,n} + 2\sqrt{\frac{u_{i,n}\log(NKT^2)}{N_{t,i}}} + \frac{\log(NKT^2)}{N_{t,i}} \qquad \text{(using Eq. (10))}$$

$$+ 4\sqrt{\frac{\log(NKT^2)}{N_{t,i}}\left(2u_{i,n} + \frac{2\log(NKT^2)}{N_{t,i}}\right)} + \frac{8\log(NKT^2)}{N_{t,i}}$$

$$\leq u_{i,n} + 8\sqrt{\frac{u_{i,n}\log(NKT^2)}{N_{t,i}}} + \frac{15\log(NKT^2)}{N_{t,i}}. \qquad \text{(using AM-GM inequality)}$$

Combining the lower and the upper bound finishes the proof. $\qquad \square$

Lemma A.3. *Under Event 1, Algorithm 1 guarantees that*

$$\sum_{\tau=KN_0+1}^{t} \frac{p_{\tau,i}}{N_{\tau,i}} \leq \mathcal{O}\left(\log T\right),$$

$$\sum_{\tau=KN_0+1}^{t} \sqrt{\frac{p_{\tau,i}}{N_{\tau,i}}} \leq \mathcal{O}\left(\sqrt{T\log T}\right),$$

for all $i \in [K]$ and $KN_0 + 1 \leq t \leq T$.

Proof. Since Event 1 holds, we know that $N_{t,i} \geq \frac{1}{2}\sum_{\tau=KN_0+1}^{t} p_{\tau,i} + 1$ holds for all $i \in [K]$ and $\tau \geq KN_0 + 1$ based on the choice of $N_0 = 18\log KT + 1$. Therefore, we know that for each $t \geq KN_0 + 1$,

$$\sum_{\tau=KN_0+1}^{t} \frac{p_{\tau,i}}{N_{\tau,i}} \leq \sum_{\tau=KN_0+1}^{t} \frac{2p_{\tau,i}}{\sum_{\tau'=KN_0+1}^{\tau} p_{\tau',i} + 2}$$

$$\leq 2\int_{0}^{\sum_{\tau=KN_0+1}^{t} p_{\tau,i}} \frac{1}{x+2}dx \leq 2\log(T+2).$$

As for the term $\sum_{\tau=KN_0+1}^{t} \sqrt{\frac{p_{\tau,i}}{N_{\tau,i}}}$, we have

$$\sum_{\tau=KN_0+1}^{t} \sqrt{\frac{p_{\tau,i}}{N_{\tau,i}}} \leq \sqrt{(t - KN_0)\sum_{\tau=KN_0+1}^{t} \frac{p_{\tau,i}}{N_{\tau,i}}} \qquad \text{(Cauchy-Schwarz inequality)}$$

$$\leq \mathcal{O}(\sqrt{T \log T}).$$

$\square$

A.2 Omitted Details in Section 3.2

Theorem 3.2. *In the bandit feedback setting, for any algorithm, there exists a stochastic environment in which the expected regret (with respect to* NSW*) of this algorithm is* $\Omega\left(\frac{(\log K)^3}{N^3} \cdot K^{\frac{1}{N}} T^{\frac{N-1}{N}}\right)$ *for* $N \geq \log K$ *and sufficiently large* T.

Proof. Consider the environment $\mathcal{E}$ that picks u uniformly from $\{u^{(1)}, \ldots, u^{(K)}\}$, where $u_{:,1}^{(i)} = \varepsilon \cdot \mathbf{e}_i \in \mathbb{R}^K$ and $u_{:,j}^{(i)} = \mathbf{1}$ for all $j \in \{2, 3, \ldots, N\}$. Here, $\varepsilon \in (0, \frac{1}{9}]$ is some constant to be specified later. Denote $u^{(0)}$ to be the environment where $u_{:,n} = \mathbf{0}$ for all $n \in [N]$. At each round t, $u_{t,i,n}$ is an i.i.d. Bernoulli random variable with mean $u_{i,n}$. For notational convenience, we use $\mathbb{E}_i[\cdot]$ when we take expectation over the environment $u^{(i)}$ for $i \in \{0\} \cup [K]$. Let n_i be the number of rounds that action i is selected over the total horizon T for all $i \in [K]$. Therefore, the expected regret with respect to environment $\mathcal{E}$ (a uniform distribution over $u^{(i)}$, $i \in [K]$) is lower bounded as follows:

$$\mathbb{E}_{\mathcal{E}}[\text{Reg}] = \frac{1}{K} \sum_{i=1}^{K} \mathbb{E}_i \left[\varepsilon^{\frac{1}{N}} \sum_{t=1}^{T} (1 - p_{t,i}^{\frac{1}{N}}) \right]$$

$$\geq T\varepsilon^{\frac{1}{N}} - \frac{\varepsilon^{\frac{1}{N}} T^{\frac{N-1}{N}}}{K} \sum_{i=1}^{K} \mathbb{E}_i \left[\left(\sum_{t=1}^{T} p_{t,i} \right)^{\frac{1}{N}} \right] \qquad \text{(Hölder's inequality)}$$

$$\geq T\varepsilon^{\frac{1}{N}} - \frac{\varepsilon^{\frac{1}{N}} T^{\frac{N-1}{N}}}{K} \sum_{i=1}^{K} \left(\mathbb{E}_i \left[\sum_{t=1}^{T} p_{t,i} \right] \right)^{\frac{1}{N}} \qquad \text{(Jensen's inequality)}$$

$$\geq T\varepsilon^{\frac{1}{N}} - K^{-\frac{1}{N}} \varepsilon^{\frac{1}{N}} T^{\frac{N-1}{N}} \left(\sum_{i=1}^{K} \mathbb{E}_i \left[\sum_{t=1}^{T} p_{t,i} \right] \right)^{\frac{1}{N}}, \qquad (11)$$

where the last inequality is again due to Hölder's inequality. Let $\text{Ber}(\alpha)$ be the Bernoulli distribution with mean α. Combining Exercise 14.4 of [Lattimore and Szepesvári, 2020], Pinsker's inequality, and Lemma 15.1 of [Lattimore and Szepesvári, 2020], we have

$$\mathbb{E}_i \left[\sum_{t=1}^{T} p_{t,i} \right] = \mathbb{E}_i[n_i] \leq \mathbb{E}_0[n_i] + T\sqrt{\frac{1}{2} \mathbb{E}_0[n_i] \text{KL}(\text{Ber}(0)|\text{Ber}(\varepsilon))}$$

$$\leq \mathbb{E}_0[n_i] + T\sqrt{\frac{1}{2} \mathbb{E}_0[n_i] \log\left(1 + \frac{\varepsilon}{1-\varepsilon}\right)}$$

$$\leq \mathbb{E}_0[n_i] + T\sqrt{\mathbb{E}_0[n_i] \frac{\varepsilon}{2(1-\varepsilon)}} \qquad \text{(using } \log(1+x) \leq x)$$

$$\leq \mathbb{E}_0[n_i] + \frac{3T}{4}\sqrt{\mathbb{E}_0[n_i]\varepsilon}$$

$$= \mathbb{E}_0 \left[\sum_{t=1}^{T} p_{t,i} \right] + \frac{3T}{4}\sqrt{\mathbb{E}_0\left[\sum_{t=1}^{T} p_{t,i}\right] \varepsilon}. \qquad (12)$$

where the last inequality is because $\varepsilon \leq \frac{1}{9}$.

Taking summation over all $i \in [K]$, we obtain that

$$\sum_{i \in [K]} \mathbb{E}_i \left[\sum_{t=1}^{T} p_{t,i} \right]$$

17

$$\leq \sum_{i \in [K]} \mathbb{E}_0 \left[\sum_{t=1}^{T} p_{t,i} \right] + \frac{3}{4} T \sum_{i=1}^{K} \sqrt{\mathbb{E}_0 \left[\sum_{t=1}^{T} p_{t,i} \right]} \varepsilon$$

$$\leq T + \frac{3T}{4} \sqrt{K \varepsilon \mathbb{E}_0 \left[\sum_{i=1}^{K} \sum_{t=1}^{T} p_{t,i} \right]}$$

$$= T + \frac{3T}{4} \sqrt{KT\varepsilon} \tag{13}$$

where the second inequality is due to Cauchy-Schwarz inequality.

Applying Eq. (13) to Eq. (11), we obtain that

$$\mathbb{E}_{\mathcal{E}}[\text{Reg}]$$

$$\geq T\varepsilon^{\frac{1}{N}} - K^{-\frac{1}{N}} \varepsilon^{\frac{1}{N}} T^{\frac{N-1}{N}} \left(T + \frac{3T}{4} \sqrt{KT\varepsilon} \right)^{\frac{1}{N}}$$

$$\geq \left(1 - K^{-\frac{1}{N}} \right) T\varepsilon^{\frac{1}{N}} - K^{-\frac{1}{2N}} \varepsilon^{\frac{3}{2N}} T^{\frac{2N+1}{2N}} \qquad \text{(using } (a+b)^{\frac{1}{N}} \leq a^{\frac{1}{N}} + b^{\frac{1}{N}})$$

$$\geq \frac{\log K}{2N} T\varepsilon^{\frac{1}{N}} - K^{-\frac{1}{2N}} \varepsilon^{\frac{3}{2N}} T^{\frac{2N+1}{2N}},$$

where the third inequality is according to Lemma C.2 with $x = \frac{1}{N}$ and $\alpha = \frac{1}{K}$, meaning that $N \left(1 - \frac{1}{K}^{\frac{1}{N}} \right) \geq \frac{\log K}{2}$.

Picking $\varepsilon = \frac{(\log K)^{2N} \cdot K}{(4N)^{2N} T}$, we know that

$$K^{-\frac{1}{2N}} \varepsilon^{\frac{3}{2N}} T^{\frac{2N+1}{2N}} = \frac{\varepsilon^{\frac{1}{N}} T \log K}{4N} = \Omega \left(\frac{(\log K)^3}{N^3} \cdot K^{\frac{1}{N}} T^{\frac{N-1}{N}} \right),$$

Combining the above all together, we know that $\mathbb{E}_{\mathcal{E}}[\text{Reg}] \geq \Omega \left(\frac{(\log K)^3}{N^3} \cdot K^{\frac{1}{N}} T^{\frac{N-1}{N}} \right)$. Therefore, there exists one environment among $u^{(i)}, i \in [K]$ such that $\mathbb{E}_i[\text{Reg}] \geq \Omega \left(\frac{(\log K)^3}{N^3} \cdot K^{\frac{1}{N}} T^{\frac{N-1}{N}} \right)$, which finishes the proof. $\square$

B Omitted Details in Section 4

B.1 Omitted Details in Section 4.1

In this section, we prove that that in the adversarial environment, it is also impossible to achieve sublinear regret when $f = \text{NSW}_{\text{prod}}$. The hard instance construction shares a similar spirit to the one for $f = \text{NSW}$ shown in Theorem 4.1.

Theorem B.1. *In the bandit feedback setting, for any algorithm, there exists an adversarial environment such that $\mathbb{E}[\text{Reg}_{\text{adv}}] = \Omega(T)$ for $f = \text{NSW}_{\text{prod}}$.*

Proof. We consider the learning environment with two agents and two arms. The agents utilities are binary, meaning that $u \in \{0,1\}^{2 \times 2}$. We construct two distributions $\mathcal{E}$ and $\mathcal{E}'$ with support $\{0,1\}^{2 \times 2}$. To define environment $\mathcal{E}$, we use q_{wxyz} for any $w, x, y, z \in \{0,1\}$ to denote the probability that $u_{1,:} = (w, x)$ and $u_{2,:} = (y, z)$ when $u \sim \mathcal{E}$. For simplicity of notation, the binary number $wxyz$ will be written in decimal form (i.e. $q_8 = \Pr_{u \sim \mathcal{E}}[u_{1,:} = (1,0), u_{2,:} = (0,0)]$). For environment $\mathcal{E}$, we assign the probabilities

$$q_i = \frac{1}{16} \quad \text{for } i \in \{0, \ldots, 15\} \setminus \{0, 2, 4, 6\} \qquad (q_0, q_2, q_4, q_6) = \left(\frac{1}{8}, 0, 0, \frac{1}{8} \right)$$

Similarly, for environment $\mathcal{E}'$, we use q'_{wxyz} for any $w, x, z, y \in \{0,1\}$ to denote the probability that $u'_{1,:} = (w, x)$ and $u'_{2,:} = (y, z)$ when $u' \sim \mathcal{E}'$. Again, we will write the binary number $wxyz$ in decimal form for ease of notation. To environment $\mathcal{E}'$ we assign probabilities

$$q'_i = \frac{1}{16} \quad \text{for } i \in \{0, \ldots, 15\} \setminus \{1, 3, 5, 7\} \qquad (q'_1, q'_3, q'_5, q'_7) = \left(0, \frac{1}{8}, \frac{1}{8}, 0 \right)$$

18

Next, we argue that the learner's observations are equivalent in distribution in $\mathcal{E}$ and $\mathcal{E}'$, since the marginal distribution of every possible observation of each action is the same. Specifically,

- When action 1 is played, the learner's observation $u_{1,:}$ in both $\mathcal{E}$ and $\mathcal{E}'$ are given by the following marginal distribution.
  - The probability of observation $(0,0)$ is $q_0 + q_1 + q_2 + q_3 = q'_0 + q'_1 + q'_2 + q'_3 = \frac{1}{4}$
  - The probability of observation $(0,1)$ is $q_4 + q_5 + q_6 + q_7 = q'_4 + q'_5 + q'_6 + q'_7 = \frac{1}{4}$
  - The probability of observation $(1,0)$ is $q_8 + q_9 + q_{10} + q_{11} = q'_8 + q'_9 + q'_{10} + q'_{11} = \frac{1}{4}$
  - The probability of observation $(1,1)$ is $q_{12} + q_{13} + q_{14} + q_{15} = q'_{12} + q'_{13} + q'_{14} + q'_{15} = \frac{1}{4}$

- When action 2 is played, the learner's observation $u_{2,:}$ in both $\mathcal{E}$ and $\mathcal{E}'$ are given by the following marginal distribution.
  - The probability of observation $(0,0)$ is $q_0 + q_4 + q_8 + q_{12} = q'_0 + q'_4 + q'_8 + q'_{12} = \frac{1}{4}$
  - The probability of observation $(0,1)$ is $q_1 + q_5 + q_9 + q_{13} = q'_1 + q'_5 + q'_9 + q'_{13} = \frac{1}{4}$
  - The probability of observation $(1,0)$ is $q_2 + q_6 + q_{10} + q_{14} = q'_2 + q'_6 + q'_{10} + q'_{14} = \frac{1}{4}$
  - The probability of observation $(1,1)$ is $q_3 + q_7 + q_{11} + q_{15} = q'_3 + q'_7 + q'_{11} + q'_{15} = \frac{1}{4}$

Direct calculation shows that

$$\mathbb{E}_{u\sim\mathcal{E}}\left[\mathrm{NSW}_{\mathrm{prod}}(u^\top p)\right]$$
$$= (q_5 - q_6 - q_9 + q_{10})p_1^2 + (q_6 - q_7 - 2q_5 + q_9 + q_{11} - q_{13} + q_{14})p_1 + (q_5 + q_7 + q_{13} + q_{15})$$
$$= -\frac{1}{16}p_1^2 + \frac{1}{16}p_1 + \frac{1}{4},$$
$$\mathbb{E}_{u\sim\mathcal{E}'}\left[\mathrm{NSW}_{\mathrm{prod}}(u^\top p)\right]$$
$$= (q'_5 - q'_6 - q'_9 + q'_{10})p_1^2 + (q'_6 - q'_7 - 2q'_5 + q'_9 + q'_{11} - q'_{13} + q'_{14})p_1 + (q'_5 + q'_7 + q'_{13} + q'_{15})$$
$$= \frac{1}{16}p_1^2 - \frac{1}{16}p_1 + \frac{1}{4}.$$

Therefore, we compute the learner's best strategy for environments $\mathcal{E}$ and $\mathcal{E}'$ by direct calculation:

$$p_\star = \underset{p\in\Delta_2}{\mathrm{argmax}}\,\mathbb{E}_{u\sim\mathcal{E}}\left[\mathrm{NSW}_{\mathrm{prod}}(u^\top p)\right] = \underset{p\in\Delta_2}{\mathrm{argmax}}\left[\frac{1}{4} + \frac{1}{16}p_1 - \frac{1}{16}p_1^2\right] = \left(\frac{1}{2}, \frac{1}{2}\right)$$

$$p'_\star = \underset{p\in\Delta_2}{\mathrm{argmax}}\,\mathbb{E}_{u\sim\mathcal{E}'}\left[\mathrm{NSW}_{\mathrm{prod}}(u^\top p)\right] = \underset{p\in\Delta_2}{\mathrm{argmax}}\left[\frac{1}{4} - \frac{1}{16}p_1 + \frac{1}{16}p_1^2\right] = \{(0,1),(1,0)\}$$

Next, consider a distribution $p \in \Delta_2$ such that $p_1 \in \left[0, \frac{1}{4}\right] \cup \left[\frac{3}{4}, 1\right]$. For such p, direct calculation shows that $\mathbb{E}_{u\sim\mathcal{E}}\left[\mathrm{NSW}_{\mathrm{prod}}(u^\top p_\star) - \mathrm{NSW}_{\mathrm{prod}}(u^\top p)\right] \geq \Delta$, where $\Delta = \frac{1}{256}$. On the other hand, for a strategy $p \in \Delta_2$ with $p_1 \in \left(\frac{1}{4}, \frac{3}{4}\right)$, we have $\mathbb{E}_{u\sim\mathcal{E}'}\left[\mathrm{NSW}_{\mathrm{prod}}(u^\top p'_\star) - \mathrm{NSW}_{\mathrm{prod}}(u^\top p)\right] > \Delta$ as well. Given any algorithm, let $\alpha_\mathcal{E}$ be the probability that the number of rounds $p_{t,1} \in \left[0, \frac{1}{4}\right] \cup \left[\frac{3}{4}, 1\right]$ is larger than $\frac{T}{2}$ under environment $\mathcal{E}$. Let $\bar{\alpha}_{\mathcal{E}'}$ be the probability of the complement of this event under environment $\mathcal{E}'$. By definition

$$\mathbb{E}_\mathcal{E}[\mathrm{Reg}_{\mathrm{adv}}] \geq \mathbb{E}_\mathcal{E}\left[\sum_{t=1}^{T}\mathrm{NSW}_{\mathrm{prod}}(u_t^\top p_\star) - \sum_{t=1}^{T}\mathrm{NSW}_{\mathrm{prod}}(u_t^\top p_t)\right] \geq \frac{\alpha_\mathcal{E} T\Delta}{2}$$

$$\mathbb{E}_{\mathcal{E}'}[\mathrm{Reg}_{\mathrm{adv}}] \geq \mathbb{E}_{\mathcal{E}'}\left[\sum_{t=1}^{T}\mathrm{NSW}_{\mathrm{prod}}(u_t^\top p'_\star) - \sum_{t=1}^{T}\mathrm{NSW}_{\mathrm{prod}}(u_t^\top p_t)\right] \geq \frac{\bar{\alpha}_{\mathcal{E}'} T\Delta}{2}$$

Since the feedback for the algorithm is the same in distribution, we have $\alpha_\mathcal{E} + \bar{\alpha}_{\mathcal{E}'} = 1$. Thus, we have

$$\max\{\mathbb{E}_\mathcal{E}[\mathrm{Reg}_{\mathrm{adv}}], \mathbb{E}_{\mathcal{E}'}[\mathrm{Reg}_{\mathrm{adv}}]\} \geq \frac{\mathbb{E}_\mathcal{E}[\mathrm{Reg}_{\mathrm{adv}}] + \mathbb{E}_{\mathcal{E}'}[\mathrm{Reg}_{adv}]}{2} \geq \frac{(\alpha_\mathcal{E} + \bar{\alpha}_{\mathcal{E}'})T\Delta}{4} = \Omega(T).$$

$\square$

B.2 Omitted Details in Section 4.2

B.2.1 Concave and Pareto Optimal SWFs

Formally, for a function $f : [0, 1]^N \mapsto [0, 1]$, concavity and Pareto Optimality are defined as:

- Concavity: $f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y)$ for any $\alpha \in [0, 1]$ and $x, y \in [0, 1]^N$.
- Pareto optimality: for any $x, y \in [0, 1]^N$, $x_n \geq y_n$ for all $n \in [N]$ implies $f(x) \geq f(y)$.

Pareto optimality is a "fundamental property" in social choice theory because it ensures that a SWF prefers alternatives that strictly more efficient: everyone is no worse off [Kaneko and Nakamura, 1979]. Concavity appears less in social choice literature. However, it promotes equity by modeling diminishing levels of desirability with the increase of a single agent's utility.

In the following, we provide examples of SWFs that satisfy concavity and Pareto optimality. Each of the following SWFs are parameterized by fixed weights $w \in \Delta_K$.

- Utilitarian SWF: $f(u) = \langle w, u \rangle$;
- Generalized Gini Index (GGI): $f(u) = \min_{\pi \in \mathbb{S}_N} \langle w_\pi, u \rangle$, where $\mathbb{S}_N$ is the set of permutations over N items and w_π is weights w permuted according to $\pi \in \mathbb{S}_N$;
- Weighted NSW: $f(u) = \prod_{n \in N} u_n^{w_n}$.

The last notable fact about the class of concave and Pareto Optimal SWFs is that it is closed under convex combinations. Specifically, for two concave and Pareto Optimal functions $f, g : [0, 1]^N \to [0, 1]$, the function $h(\cdot) = \lambda f(\cdot) + (1 - \lambda)g(\cdot)$ for any $\lambda \in [0, 1]$ is concave and Pareto Optimal. [Chen and Hooker, 2023] discusses how such convex combinations can be used to combine a SWF prioritizing efficiency and another prioritizing equity to derive a different SWF that prioritizes a balance between efficiency and equity.

B.2.2 Omitted Details in Section 4.2.1

In this section, we prove Theorem 4.2, which shows that $\mathcal{O}(\sqrt{KT \log T})$ regret is achievable for all concave and Pareto optimal SWFs.

Theorem 4.2. *For any $f : [0, 1]^N \to [0, 1]$ that is concave and Pareto optimal, Algorithm 2 with the log-barrier regularizer $\psi(p) = -\sum_{i=1}^K \log p_i$ and $\eta = \sqrt{\frac{K \log T}{T}}$ guarantees $\mathrm{Reg}_{\mathrm{adv}} = \mathcal{O}(\sqrt{KT \log T})$.*

Proof. Using the concavity of f, we can upper bound $\mathrm{Reg}_{\mathrm{adv}}$ as follows:

$$\mathrm{Reg}_{\mathrm{adv}} = \max_{p \in \Delta_K} \sum_{t=1}^T f(u_t^\top p) - \sum_{t=1}^T f(u_t^\top p_t)$$

$$\leq \underbrace{\max_{p \in \Delta_{K, \frac{1}{KT}}} \sum_{t=1}^T \langle -\nabla f(u_t^\top p_t), p_t - p \rangle}_{\texttt{Term (1)}} + \underbrace{\max_{p \in \Delta_K} \sum_{t=1}^T f(u_t^\top p) - \max_{p \in \Delta_{K, \frac{1}{KT}}} \sum_{t=1}^T f(u_t^\top p)}_{\texttt{Term (2)}},$$

where $\Delta_{K, \frac{1}{KT}} = \{p \in \Delta_K \mid p_i \geq \frac{1}{KT}, \forall i \in [K]\}$.

To bound `Term (1)`, according to a standard analysis of FTRL/OMD with log-barrier regularizer (e.g. Lemma 12 of [Agarwal et al., 2017]), we know that:

$$\texttt{Term (1)} \leq \max_{p \in \Delta_{K, \frac{1}{KT}}} \frac{D_\psi(p, p_1)}{\eta} + \eta \sum_{t=1}^T \sum_{i=1}^K p_{t,i}^2 \cdot \left[\nabla f(u_t^\top p_t)\right]_i^2, \tag{14}$$

where $D_\psi(p, q) \triangleq \psi(p) - \psi(q) - \langle \nabla \psi(q), p - q \rangle$ is the Bregman divergence between p and q with respect to ψ. To further bound the right-hand side, note that $p_1 = \frac{1}{K} \cdot \mathbf{1}$. Direct calculation shows that for any $p \in \Delta_{K, \frac{1}{KT}}$,

$$D_\psi(p, p_1) = \sum_{i=1}^K \left(\frac{p_i}{p_{1,i}} - 1 - \log\left(\frac{p_i}{p_{1,i}}\right)\right)$$

20

$$= \sum_{i=1}^{K} \log \left(\frac{1}{K p_i} \right)$$

$$\le K \log \left(\frac{1}{K \cdot \frac{1}{KT}} \right) \qquad \text{(since } p_i \ge \frac{1}{KT} \text{ for all } i \in [K])$$

$$\le K \log T. \tag{15}$$

Using the Pareto optimality property of f and the positivity of the utility matrix u_t, we know that $[\nabla f(u_t^\top p)]_i \ge 0$, meaning that $\sum_{i=1}^{K} p_{t,i}^2 \left[\nabla f(u_t^\top p_t) \right]_i^2 \le \langle p_t, \nabla f(u_t^\top p_t) \rangle^2$.

Moreover, using the concavity property of f, we know that $\langle p_t, \nabla f(u_t^\top p_t) \rangle \le f(u_t^\top p_t) - f(u_t^\top \mathbf{0}) = f(u_t^\top p_t) \le 1$. Combining the above two inequalities means that

$$\sum_{t=1}^{T} \sum_{i=1}^{K} p_{t,i}^2 \left[\nabla f(u_t^\top p_t) \right]_i^2 \le T. \tag{16}$$

Combining Eq. (15) and Eq. (16), we can upper bound `Term` (1) as follows:

$$\texttt{Term } (1) \le \frac{K \log T}{\eta} + \eta T. \tag{17}$$

Denote the optimal distribution $p^\star = \text{argmax}_{p \in \Delta_K} \sum_{t=1}^{T} f(u_t^\top p)$. Recall that $p_1 = \frac{1}{K} \cdot \mathbf{1}$. Now we upper bound `Term` (2) as follows:

$$\texttt{Term } (2) = \sum_{t=1}^{T} f(u_t^\top p^\star) - \max_{p \in \Delta_{K, \frac{1}{KT}}} \sum_{t=1}^{T} f(u_t^\top p)$$

$$\le \sum_{t=1}^{T} f(u_t^\top p^\star) - \sum_{t=1}^{T} f \left(u_t^\top \left(\left(1 - \frac{1}{T}\right) p^\star + \frac{1}{T} \cdot p_1 \right) \right)$$

$$((1 - \tfrac{1}{T})p^\star + \tfrac{1}{T} p_1 \in \Delta_{K, \frac{1}{KT}})$$

$$\le \sum_{t=1}^{T} f(u_t^\top p^\star) - \sum_{t=1}^{T} \left[\left(1 - \frac{1}{T}\right) \cdot f(u_t^\top p^\star) + \frac{1}{T} \cdot f(u_t^\top p_1) \right] \qquad \text{(Concavity)}$$

$$\le \frac{1}{T} \cdot \sum_{t=1}^{T} f(u_t^\top p^\star) \qquad \text{(since } f(u_t^\top p_1) \ge 0)$$

$$\le 1. \tag{18}$$

Combining Eq. (17) and Eq. (18), and choosing $\eta = \sqrt{\frac{K \log T}{T}}$ finishes the proof. $\qquad \square$

B.2.3   Omitted Details in Section 4.2.2

In this section, we present the omitted proof for Theorem 4.3, which shows a better dependency on K compared with Theorem 4.2.

Theorem 4.3. *For $f = \text{NSW}$, Algorithm 2 with the Tsallis entropy regularizer $\psi(p) = \frac{1 - \sum_{i=1}^{K} p_i^\beta}{1 - \beta}$, $\beta = \frac{2}{N}$, and the optimal choice of η guarantees $\text{Reg}_{\text{adv}} = \widetilde{\mathcal{O}}(K^{\frac{1}{2} - \frac{1}{N}} \sqrt{NT})$.*

Proof. Consider the case when $N \ge 3$. Direct calculation shows that

$$\left[\nabla f(u^\top p) \right]_i \le \frac{1}{N} \sum_{n=1}^{N} \frac{u_{i,n}}{\langle p, u_{:,n} \rangle^{1 - \frac{1}{N}}}. \tag{19}$$

Using the concavity of f and a standard analysis of FTRL with Tsallis entropy (e.g., [Luo, 2017, Theorem 1]), we know that

$$\text{Reg}_{\text{adv}} = \sum_{t=1}^{T} \left(f(u_t^\top p^\star) - f(u_t^\top p_t) \right)$$

21

$$\leq \langle \nabla f(u_t^\top p_t), p^\star - p_t \rangle \qquad\qquad \text{(concavity of } f\text{)}$$

$$\leq \frac{K^{1-\beta}-1}{\eta(1-\beta)} + \frac{\eta}{\beta}\sum_{t=1}^{T}\sum_{i=1}^{K} p_{t,i}^{2-\beta}\left[\nabla f(u_t^\top p_t)\right]_i^2$$

$$= \frac{K^{1-\beta}-1}{\eta(1-\beta)} + \frac{\eta}{N^2\beta}\sum_{t=1}^{T}\sum_{i=1}^{K}\left(\sum_{n=1}^{N}\frac{p_{t,i}^{1-\frac{\beta}{2}}u_{t,i,n}}{(\sum_{j=1}^{K}u_{t,j,n}\cdot p_{t,j})^{1-\frac{1}{N}}}\right)^2 \qquad \text{(using Eq. (19))}$$

$$\leq \frac{K^{1-\beta}-1}{\eta(1-\beta)} + \frac{\eta}{N\beta}\sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{n=1}^{N}\frac{p_{t,i}^{2-\beta}u_{t,i,n}^2}{(\sum_{j=1}^{K}u_{t,j,n}\cdot p_{t,j})^{2-\frac{2}{N}}} \qquad \text{(Cauchy-Schwarz inequality)}$$

$$\leq \frac{K^{1-\beta}-1}{\eta(1-\beta)} + \frac{\eta}{N\beta}\sum_{t=1}^{T}\sum_{n=1}^{N}\frac{\sum_{i=1}^{K}p_{t,i}^{2-\beta}u_{t,i,n}^2}{\sum_{j=1}^{K}p_{t,j}^{2-\frac{2}{N}}u_{t,j,n}^{2-\frac{2}{N}}} \qquad \left(\text{since } (\sum_i x_i)^\alpha \geq \sum_i x_i^\alpha \text{ for } \alpha \geq 1\right)$$

$$\leq \frac{K^{1-\beta}}{\eta(1-\beta)} + \frac{\eta}{N\beta}\sum_{t=1}^{T}\sum_{n=1}^{N}\frac{\sum_{i=1}^{K}p_{t,i}^{2-\beta}u_{t,i,n}^{2-\frac{2}{N}}}{\sum_{j=1}^{K}p_{t,j}^{2-\frac{2}{N}}u_{t,j,n}^{2-\frac{2}{N}}}. \qquad \text{(since } u_{t,i,n}\in[0,1]\text{)}$$

Picking $\beta = \frac{2}{N}$, the first term can be upper bounded by $\frac{3K^{1-\frac{2}{N}}}{\eta}$ and the second term can be upper bounded by $\frac{\eta T}{\beta} = \frac{\eta N T}{2}$. Further picking the optimal choice of η finishes the proof.

When $N=2$ and $\beta = \frac{2}{N} = 1$, the regularizer $\psi(p) = \frac{1-\sum_{i=1}^{K}p_i^\beta}{1-\beta}$ becomes the negative Shannon entropy $\psi(p) = \sum_{i=1}^{K}p_i \log p_i$. Using the concavity of f and following a standard analysis of FTRL with Shannon entropy regularizer (e.g., [Hazan et al., 2016, Theorem 5.2]), we obtain that

$$\text{Reg}_{\text{adv}} = \sum_{t=1}^{T}\left(f(u_t^\top p^\star) - f_t(u_t^\top p_t)\right)$$

$$\leq \sum_{t=1}^{T}\langle \nabla f(u_t^\top p_t), p^\star - p_t \rangle \qquad\qquad \text{(using the concavity of } f\text{)}$$

$$\leq \frac{\psi(p^\star)-\psi(p_1)}{\eta} + 2\eta\sum_{t=1}^{T}\sum_{i=1}^{K}p_{t,i}\left[\nabla f(u_t^\top p_t)\right]_i^2 \quad \text{(by [Hazan et al., 2016, Theorem 5.2])}$$

$$= \frac{\log K}{\eta} + 2\eta\sum_{t=1}^{T}\sum_{i=1}^{K}p_{t,i}\left(\frac{u_{t,i,1}}{2\sqrt{\langle p_t, u_{t,:,1}\rangle}} + \frac{u_{t,i,2}}{2\sqrt{\langle p_t, u_{t,:,2}\rangle}}\right)^2$$

$$\leq \frac{\log K}{\eta} + \eta\sum_{t=1}^{T}\left(\frac{\sum_{i=1}^{K}p_{t,i}u_{t,i,1}^2}{\sum_{i=1}^{K}p_{t,i}u_{t,i,1}} + \frac{\sum_{i=1}^{K}p_{t,i}u_{t,i,2}^2}{\sum_{i=1}^{K}p_{t,i}u_{t,i,2}}\right) \qquad \text{(AM-GM inequality)}$$

$$\leq \frac{\log K}{\eta} + 2\eta T. \qquad\qquad \text{(since } u_{t,i,n}\in[0,1]\text{ for all } t,i,n\text{)}$$

Picking $\eta = \sqrt{\frac{\log K}{T}}$ shows that $\text{Reg}_{\text{adv}} = \mathcal{O}\left(\sqrt{T\log K}\right) = \widetilde{\mathcal{O}}(\sqrt{T})$ for $N=2$. $\qquad\square$

B.2.4 Omitted Details in Section 4.2.3

In this section, we show the proof for Theorem 4.4, which shows that logarithmic regret is achievable when there is at least one agent who is indifferent about the learner's choice.

Theorem 4.4. *Fix $f = \text{NSW}$. Suppose that for each time t, there is a set of agents $A_t \subseteq [N]$ such that $|A_t| \geq M$ and $u_{t,:,n} = c_{t,n}\mathbf{1}$ with $c_{t,n} \geq 0$ for each agent $n \in A_t$. Then the EWOO algorithm guarantees $\text{Reg}_{\text{adv}} = \mathcal{O}\left(\frac{N-M}{M}\cdot K\log T\right)$.*

Proof. To show that EWOO algorithm achieves logarithmic regret, we need to show that $f_t(p) \triangleq -\text{NSW}(u_t^\top p)$ is α-exp-concave for some $\alpha > 0$ for all $t \in [T]$, meaning that

$$\nabla^2 f_t(p) - \alpha \nabla f_t(p)\nabla f_t(p)^\top \succeq 0.$$

Let $A_t \subseteq [N]$ be the set of agents with $u_{t,:,n} = c_{t,n} \cdot \mathbf{1}$ for all $n \in A$ on round $t \in [T]$. It is guaranteed that $|A_t| \geq M$ for all $t \in [T]$. Denote $B_t = [N] \setminus A_t$. Direct calculation shows that

$$\nabla f_t(p) = \frac{\Pi_{m \in A_t} c_{t,m}^{\frac{1}{N}}}{N} \sum_{n \in B_t} \frac{f_t(p)}{\langle p, u_{t,:,n} \rangle} u_{t,:,n},$$

$$\nabla^2 f_t(p) = \frac{\Pi_{m \in A_t} c_{t,m}^{\frac{1}{N}}}{N} \sum_{i \in B_t} \frac{u_{t,:,n} \nabla f_t(p)^\top \langle p, u_{t,:,n} \rangle - f_t(p) u_{t,:,n} u_{t,:,n}^\top}{\langle p, u_{t,:,n} \rangle^2}$$

$$= \frac{f_t(p) \Pi_{m \in A_t} c_{t,m}^{\frac{2}{N}}}{N^2} \left(\sum_{i \in B_t} \frac{u_{t,:,n}}{\langle p, u_{t,:,n} \rangle} \right) \left(\sum_{n \in B_t} \frac{u_{t,:,n}}{\langle p, u_{t,:,n} \rangle} \right)^\top$$

$$- \frac{f_t(p) \Pi_{m \in A_t} c_{t,m}^{\frac{1}{N}}}{N} \sum_{i \in B_t} \frac{u_{t,:,n} u_{t,:,n}^\top}{\langle p, u_{t,:,n} \rangle^2}.$$

For notational convenience, let $\lambda_t = \Pi_{m \in A_t} c_{t,m}^{\frac{1}{N}} \leq 1$. Picking $\alpha = \frac{M}{N-M}$, we know that

$$\nabla^2 f_t(p) - \alpha \nabla f_t(p) \nabla f_t(p)^\top$$

$$= \frac{\lambda_t^2 f_t(p)}{N^2} \left(\sum_{n \in B_t} \frac{u_{t,:,n}}{\langle p, u_{t,:,n} \rangle} \right) \left(\sum_{n \in B_t} \frac{u_{t,:,n}}{\langle p, u_{t,:,n} \rangle} \right)^\top - \frac{\lambda_t f_t(p)}{N} \sum_{i \in B_t} \frac{u_{t,:,n} u_{t,:,n}^\top}{\langle p, u_{t,:,n} \rangle^2}$$

$$- \frac{\alpha \lambda_t^2 f_t(p)^2}{N^2} \left(\sum_{n \in B_t} \frac{u_{t,:,n}}{\langle p, u_{t,:,n} \rangle} \right) \left(\sum_{n \in B_t} \frac{u_{t,:,n}}{\langle p, u_{t,:,n} \rangle} \right)^\top$$

$$= \frac{-\lambda_t f_t(p)}{N} \left[\sum_{n \in B_t} \frac{u_{t,:,n} u_{t,:,n}^\top}{\langle p, u_{t,:,n} \rangle^2} - \frac{\lambda_t (1 - \alpha f_t(p))}{N} \left(\sum_{n \in B_t} \frac{u_{t,:,n}}{\langle p, u_{t,:,n} \rangle} \right) \left(\sum_{n \in B_t} \frac{u_{t,:,n}}{\langle p, u_{t,:,n} \rangle} \right)^\top \right]$$

$$\succeq \frac{-\lambda_t f_t(p)}{N} \left[\sum_{n \in B_t} \frac{u_{t,:,n} u_{t,:,n}^\top}{\langle p, u_{t,:,n} \rangle^2} - \frac{1 - \alpha f_t(p)}{N} \left(\sum_{n \in B_t} \frac{u_{t,:,n}}{\langle p, u_{t,:,n} \rangle} \right) \left(\sum_{n \in B_t} \frac{u_{t,:,n}}{\langle p, u_{t,:,n} \rangle} \right)^\top \right]$$

$$\succeq \frac{-\lambda_t f_t(p)}{N} \left[\sum_{n \in B_t} \frac{u_{t,:,n} u_{t,:,n}^\top}{\langle p, u_{t,:,n} \rangle^2} - \frac{1 + \alpha}{N} \left(\sum_{n \in B_t} \frac{u_{t,:,n}}{\langle p, u_{t,:,n} \rangle} \right) \left(\sum_{n \in B_t} \frac{u_{t,:,n}}{\langle p, u_{t,:,n} \rangle} \right)^\top \right]$$

$$= \frac{-\lambda_t f_t(p)}{N} \left[\sum_{n \in B_t} \frac{u_{t,:,n} u_{t,:,n}^\top}{\langle p, u_{t,:,n} \rangle^2} - \frac{1}{N - M} \left(\sum_{n \in B_t} \frac{u_{t,:,n}}{\langle p, u_{t,:,n} \rangle} \right) \left(\sum_{n \in B_t} \frac{u_{t,:,n}}{\langle p, u_{t,:,n} \rangle} \right)^\top \right]$$

$$\succeq \frac{-\lambda_t f_t(p)}{N} \left[\sum_{n \in B_t} \frac{u_{t,:,n} u_{t,:,n}^\top}{\langle p, u_{t,:,n} \rangle^2} - \frac{1}{|B_t|} \left(\sum_{n \in B_t} \frac{u_{t,:,n}}{\langle p, u_{t,:,n} \rangle} \right) \left(\sum_{n \in B_t} \frac{u_{t,:,n}}{\langle p, u_{t,:,n} \rangle} \right)^\top \right]$$

$$\succeq 0, \qquad \qquad \text{(Cauchy-Schwarz inequality)}$$

where the first inequality is because $\lambda_t \leq 1$ and $f_t(p) \leq 0$; the second inequality is because $f_t(p) \geq -1$; the third inequality is because $|B_t| \leq \frac{1}{N-M}$. This shows that the $f_t(p)$ is $\frac{M}{(N-M)}$-exp-concave. Therefore, according to Theorem 4.4 of [Hazan et al., 2016], we know that the EWOO algorithm guarantees that

$$\sum_{t=1}^{T} (f_t(p_t) - f_t(p^\star)) \leq \left(\frac{N - M}{M} \right) \cdot K \log T + \frac{2(N - M)}{M}.$$

$\square$

C   Auxiliary Lemmas

In this section, we include several auxiliary lemmas that are useful in the analysis.

Lemma C.1. *Let $a_1, \ldots, a_n, b_1, \ldots b_n \in [0,1]$, where $a_i \geq b_i$ for all $i \in [n]$. Then, $\prod_{i=1}^n a_i - \prod_{i=1}^n b_i \leq \sum_{i=1}^n (a_i - b_i)$.*

Proof. Direct calculation shows that

$$\prod_{i=1}^n a_i - \prod_{i=1}^n b_i = \sum_{j=1}^n \left(\prod_{i=1}^j a_i \prod_{i=j+1}^n b_i - \prod_{i=1}^{j-1} a_i \prod_{i=j}^n b_i \right)$$

$$= \sum_{j=1}^n \left((a_j - b_j) \prod_{i=1}^{j-1} a_i \prod_{i=j+1}^n b_i \right)$$

$$\leq \sum_{j=1}^n (a_j - b_j).$$

$\square$

Lemma C.2. *For all $x \in (0,1)$ and $\alpha \in (0,1)$ satisfying $1 + x \log \alpha \geq 0$, we have $\frac{1-\alpha^x}{x} \geq -\frac{\log \alpha}{2}$.*

Proof. Let $y = \log \alpha < 0$. We know that

$$\frac{1-\alpha^x}{x} \geq -\frac{\log \alpha}{2}$$

$$\Longleftrightarrow 1 - e^{xy} \geq -\frac{xy}{2}$$

$$\Longleftrightarrow e^{xy} \leq 1 + \frac{xy}{2},$$

which is true since $e^u \leq 1 + \frac{u}{2}$ for all $u \in [-1, 0]$ and $xy = x \log \alpha \geq -1$. $\square$

Lemma C.3 (Theorem 1 in [Beygelzimer et al., 2011]). *Let $X_1, \ldots, X_T \in [-B, B]$ for some $B > 0$ be a martingale difference sequence and with $\sum_{t=1}^T \mathbb{E}_t[X_t^2] \leq V$ for some fixed quantity $V > 0$. We have for all $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$\sum_{t=1}^T X_t \leq \min_{\lambda \in [0, 1/B]} \left(\lambda V + \frac{\log(1/\delta)}{\lambda} \right) \leq 2\sqrt{V \log(1/\delta)} + B \log(1/\delta).$$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See abstract and Section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 1.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: See Section 3, Section 4, and the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The authors have reviewed the NeurIPS Code of Ethics. The research conducted in this paper conforms with it in every respect.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This work is mostly theoretical, and we do not foresee any negative ethical or societal outcomes.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: This paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.