Closing the Computational-Statistical Gap in Best Arm Identification for Combinatorial Semi-bandits

Ruo-Chun Tzeng
EECS
KTH, Stockholm, Sweden
rctzeng@kth.se

Po-An Wang
EECS
KTH, Stockholm, Sweden
wang9@kth.se

Alexandre Proutiere
EECS and Digital Futures
KTH, Stockholm, Sweden
alepro@kth.se

Chi-Jen Lu
Institute of Information Science
Academia Sinica, Taiwan
cjlu@iis.sinica.edu.tw

Abstract

We study the best arm identification problem in combinatorial semi-bandits in the fixed confidence setting. We present Perturbed Frank-Wolfe Sampling (P-FWS), an algorithm that (i) runs in polynomial time, (ii) achieves the instance-specific minimal sample complexity in the high confidence regime, and (iii) enjoys polynomial sample complexity guarantees in the moderate confidence regime. To the best of our knowledge, even for the vanilla bandit problems, no algorithm was able to achieve (ii) and (iii) simultaneously. With P-FWS, we close the computational-statistical gap in best arm identification in combinatorial semi-bandits. The design of P-FWS starts from the optimization problem that defines the information-theoretical and instance-specific sample complexity lower bound. P-FWS solves this problem in an online manner using, in each round, a single iteration of the Frank-Wolfe algorithm. Structural properties of the problem are leveraged to make the P-FWS successive updates computationally efficient. In turn, P-FWS only relies on a simple linear maximization oracle.

1 Introduction

An efficient method to design statistically optimal algorithms solving active learning tasks (e.g., regret minimization or pure exploration in bandits and reinforcement learning) consists in the following two-step procedure. The first step amounts to deriving, through change-of-measure arguments, tight information-theoretical fundamental limits satisfied by a wide class of learning algorithms. These limits are often expressed as the solution of an optimization problem, referred in this paper to as the lower-bound problem. Interestingly, this solution specifies the instance-specific optimal exploration process: it characterizes the limiting behavior of the adaptive sampling rule any statistically optimal algorithm should implement. In the second step, the learning algorithm is designed so that its exploration process approaches the solution of the lower-bound problem. This design yields statistically optimal algorithms, but typically requires to repeatedly solve the lower-bound problem. This method has worked remarkably well for simple learning tasks such as regret minimization or best-arm identification with fixed confidence in classical stochastic bandits [Lai87, GC11, GK16], but also in bandits whose arm-to-average reward function satisfies simple structural properties (e.g., Lipschitz, unimodal) [MCPT14, WTP21].

The method also provides a natural way of studying the computational-statistical gap [KLLM22] for active learning tasks. Indeed, if solving the lower-bound problem in polynomial time is possible, one
may hope to devise learning algorithms that are both statistically optimal and computationally efficient. As of now, however, the computational complexity of the lower-bound problem remains largely unexplored, except for simple learning tasks. For example, in the case of best policy identification in tabular Markov Decision Processes, the lower-bound problem is non-convex [AMP21] and its complexity and approximability are unclear.

In this paper, we leverage the aforementioned two-step procedure to assess the computational-statistical gap for the best arm identification in combinatorial semi-bandits in the fixed confidence setting. We establish that, essentially, this gap does not exist (a result that was conjectured in [JMKK21]). Specifically, we present an algorithm that enjoys the following three properties: (i) it runs in polynomial time, (ii) its sample complexity matches the fundamental limits asymptotically in the high confidence regime, and (iii) its sample complexity is at most polynomial in the moderate confidence regime. Next, after formally introducing combinatorial semi-bandits, we describe our contributions and techniques in detail.

**Best arm identification in combinatorial semi-bandits.** In combinatorial semi-bandits [CBL12, CTMSP+15], the learner sequentially selects an action from a combinatorial set \(X \subset \{0, 1\}^K\). When in round \(t\), the action \(x(t) = (x_1(t), \ldots, x_K(t)) \in X\) is chosen, the environment samples a \(K\)-dimensional vector \(y(t)\) whose distribution is assumed to be Gaussian \(N(\mu, I)\). The learner then receives the detailed reward vector \(x(t) \odot y(t)\) where \(\odot\) denotes the element-wise product (in other words, the learner observes the individual reward \(y_k(t)\) of the arm \(k\) if and only if this arm is selected in round \(t\), i.e., \(x_k(t) = 1\)). The parameter \(\mu\) characterizing the average rewards of the various arms is initially unknown. The goal of a learner is to identify the best action \(\hat{i}^*(\mu) = \arg\max_{x \in X} \langle x, \mu \rangle\) with a given level of confidence \(1 - \delta\), for some \(\delta > 0\) while minimizing the expected number of rounds needed. We assume that the best action is unique and denote by \(\Lambda = \{\mu \in \mathbb{R}^K : \langle \hat{i}^*(\mu) \rangle = 1\}\) the set of parameters satisfying this assumption. The learner strategy is defined by three components: (i) a sampling rule dictating the sequence of the selected actions; (ii) a stopping time \(\tau\) defining the last round where the learner interacts with the environment; (iii) a decision rule specifying the action \(\hat{i}\) believed to be optimal based on the data gathered until \(\tau\).

**The sample complexity lower-bound problem.** Consider the set of \(\delta\)-PAC algorithms such that for any \(\mu \in \Lambda\), the best action is identified correctly with probability at least \(1 - \delta\). We wish to find a \(\delta\)-PAC algorithm with minimal expected sample complexity \(E_\mu[\tau]\). To this aim, using classical change-of-measure arguments [GK16], we may derive a lower bound on the expected sample complexity satisfied by any \(\delta\)-PAC algorithm. This lower bound is given by \(\mathbb{E}_\mu[\tau] \geq T^*(\mu) k(\delta, 1 - \delta)\).

The characteristic time \(T^*(\mu)\) is defined as the value of the following problem

\[
T^*(\mu)^{-1} = \sup_{\omega \in \Sigma} \inf_{\lambda \in \text{Alt}(\mu)} \left\langle \omega, \frac{(\mu - \lambda)^2}{2} \right\rangle,
\]

where \(\Sigma = \{\{x \in X : w_x \cdot x : w \in \Sigma_{\{|X|\}}\}\}, \text{kl}(a, b)\) is the KL-divergence between two Bernoulli distributions with respective means \(a\) and \(b\), and \(\text{Alt}(\mu) = \{\lambda \in \Lambda : \hat{i}^*(\lambda) \neq \hat{i}^*(\mu)\}\) is the set of confusing parameters. As it turns out (see Lemma [I]), \(T^*(\mu)\) is at most quadratic in \(K\) and hence the sample complexity lower bound is polynomial \([\text{T}1]\) is a concave program over \(\Sigma\) [WTP21], and a point \(\omega^*\) in its solution set corresponds to an optimal allocation of action draws: an algorithm sampling actions according to \(\omega^*\) and equipped with an appropriate stopping rule would yield a sample complexity matching the lower bound. In this paper, we provide computationally efficient algorithms to solve \([\text{T}1]\) and show how these can be used to devise a \(\delta\)-PAC best action identification algorithm with minimal sample complexity and running in polynomial time. We only assume that we have access to a computationally efficient Oracle, referred to as the LM (Linear Maximization) Oracle, identifying the best action should \(\mu\) be known (but for any possible \(\mu\)). This assumption, made in all previous work on combinatorial semi-bandits (see e.g. [JMKK21, PBVP20]), is crucial as indeed, if there is no computationally efficient algorithm solving the offline problem \(\arg\max_{x \in X} \langle x, \mu \rangle\) with known \(\mu\), there is no hope to solve its online version with unknown \(\mu\) in a computationally efficient manner. The assumption holds for a large array of combinatorial sets of actions \([S^2]\), including \(m\)-sets, matchings, (source–destination)-paths, spanning trees, matroids (refer to [CCG21e] for a thorough discussion).

\[1\] We present proof in Appendix K for completeness – see also [JMKK21].

\[2\] \(\Sigma_N\) denotes the \((N - 1)\) dimensional simplex.
The Most-Confusing-Parameter (MCP) algorithm. The difficulty of solving \((1)\) lies in the inner optimization problem, i.e., in evaluating the objective function:

\[
F_\mu(\omega) = \inf_{\lambda \in \mathcal{A}(\mu)} \left\langle \omega, \frac{(\mu - \lambda)^2}{2} \right\rangle = \min_{x \neq v(\mu)} f_x(\omega, \mu)
\]

where \(f_x(\omega, \mu) = \inf_{\lambda \in C_x(\omega)} \langle \omega, \frac{(\mu - \lambda)^2}{2}\rangle\) and \(C_x = \{\lambda \in \mathbb{R}^K : \langle \lambda, i^*(\mu) - x \rangle < 0\}\). Evaluating \(F_\mu(\omega)\) is required to solve \((1)\), but also in the design of an efficient stopping rule. Our first contribution is MCP (Most-Confusing-Parameter), a polynomial time algorithm able to approximate \(F_\mu(\omega)\) for any given \(\mu\) and \(\omega\). The algorithm’s name refers to the fact that by computing \(F_\mu(\omega)\), we implicitly identify the most confusing parameter \(\lambda^* \in \arg \min_{\lambda \in \mathcal{A}(\mu)} \langle \omega, \frac{(\mu - \lambda)^2}{2}\rangle\). The design of MCP leverages a Lagrangian relaxation of the optimization problem defining \(f_x(\omega, \mu)\) and exploits the fact that the Lagrange dual function linearly depends on \(x\). In turn, this linearity allows us to make use of the LM Oracle. From these observations, we show that computing \(F_\mu(\omega)\) boils down to solving a two-player game, for which one of the players can simply update her strategy using the LM Oracle. We prove the following informally stated theorem quantifying the performance of the MCP algorithm (see Theorem 3 for a more precise statement).

**Theorem 1.** For any \((\omega, \mu)\), the MCP algorithm with precision \(\epsilon\) and certainty parameter \(\theta\) returns \(\hat{F}\) and \(\hat{x}\) satisfying \(\mathbb{P}_{\mu}[F_\mu(\omega) \leq \hat{F} \leq (1+\epsilon)F_\mu(\omega)] \geq 1-\theta\) and \(\hat{F} = f_\mu(\omega, \mu)\). The number of calls to the LM Oracle is, almost surely, at most polynomial in \(K\), \(\epsilon^{-1}\), and \(\ln \theta^{-1}\).

The Perturbed Frank-Wolfe Sampling (P-FWS) algorithm. The MCP algorithm allows us to solve the lower-bound problem \((1)\) for any given \(\mu\). The latter is initially unknown, but could be estimated. A Track-and-Stop algorithm \([\text{GK}16]\) solving \((1)\) with this plug-in estimator in each round would yield asymptotically minimal sample complexity. It would however be computationally expensive. To circumvent this difficulty, as in \([\text{WTP21}]\), our algorithm, P-FWS, performs a single iteration of the Frank-Wolfe algorithm for the program \((1)\) instantiated with an estimator of \(\mu\). To apply the Frank-Wolfe algorithm, P-FWS uses stochastic smoothing techniques to approximate the non-differentiable objective function \(f_\mu\) by a smooth function. To estimate the gradient of the latter, P-FWS leverages both the LM Oracle and the MCP algorithm (more specifically its second output \(\hat{x}\)). Finally, P-FWS stopping rule takes the form of a classical Generalized Likelihood Ratio Test (GLRT) where the estimated objective function is compared to a time-dependent threshold. Hence the stopping rule also requires the MCP algorithm. We analyze the sample and computational complexities of P-FWS. Our main results are summarized in the following theorem (refer to Theorem 4 for details).

**Theorem 2.** For any \(\delta \in (0, 1)\), P-FWS is \(\delta\)-PAC, and for any \((\epsilon, \tilde{\epsilon}) \in (0, 1)\) small enough, its sample complexity satisfies:

\[
\mathbb{E}_\mu[\tau] \leq \frac{(1+\tilde{\epsilon})^2}{T^*(\mu)^{-1} - \epsilon} \times \left( \frac{1}{\delta} \epsilon \right) + \Psi(\epsilon, \tilde{\epsilon}),
\]

where \(H(x) = \ln(x) + \ln(\ln(x))\), \(c > 0\) is a universal constant, and \(\Psi(\epsilon, \tilde{\epsilon})\) is polynomial in \(\epsilon^{-1}, \tilde{\epsilon}^{-1}, K, \|\mu\|_{\infty}, \|x\|_{\infty}, \text{ and } \Delta^{-1}_{\min}\) where \(\Delta_{\min} = \min_{x \neq x^*}(\epsilon^*(\mu) - x, \mu)\). Under P-FWS, the number of LM Oracle calls per round is at most polynomial in \(\ln \delta^{-1}\) and \(K\). The total expected number of these calls is also polynomial.

To the best of our knowledge, P-FWS is the first polynomial time best action identification algorithm with minimal sample complexity in the high confidence regime (when \(\delta\) tends to 0). Its sample complexity is also polynomial in \(K\) in the moderate confidence regime.

## 2 Preliminaries

We start by introducing some notation. We use bold lowercase letters (e.g., \(x\)) for vectors, and bold uppercase letters (e.g., \(A\)) for matrices. \(\odot\) (resp. \(\oplus\)) denotes the element-wise product (resp. sum) over \(\mathbb{Z}_2\). For \(x \in \mathbb{R}^K\), \(i \in \mathbb{N}\), \(x^i = (x^i_k)_{k \in [K]}\) is the \(i\)-th element-wise power of \(x\). \(D = \max_{x \in X} \|x\|_1\) denotes the maximum number of arms part of an action. For any \(\mu \in \Lambda\), we define the sub-optimality gap of \(x \in X\) as \(\Delta_x(\mu) = (i^*(\mu) - x, \mu)\), and the minimal gap as \(\Delta_{\min}(\mu) = \min_{x \neq \mu}(i^*(\mu) - x, \mu)\). \(\mathbb{P}_\mu\) (resp. \(\mathbb{E}_\mu\)) denotes the probability measure (resp. expectation) when the arm rewards are parametrized by \(\mu\). Whenever it is clear from the context, we will drop \(\mu\) for simplicity, e.g. \(\hat{i}^* = i^*(\mu)\), \(\hat{\Delta}_x = \Delta_x(\mu)\), and \(\hat{\Delta}_{\min} = \min_{x \neq \mu} \Delta_x(\mu)\). Refer to Appendix A for an exhaustive table of notation.
2.1 The lower-bound problem

Classical change-of-measure arguments lead to the asymptotic sample complexity lower bound $\mathbb{E}_\mu[T] \geq T^*(\mu)k(\delta, 1 - \delta)$ where the characteristic time is defined in $[1]$. To have a chance to develop a computationally efficient best action identification algorithm, we need that the sample complexity lower bound grows at most polynomially in $K$. This is indeed the case as stated in the following lemma, whose proof is provided in Appendix [K].

**Lemma 1.** For any $\mu \in \Delta$, $T^*(\mu) \leq 4KD\Delta_{\min}(\mu)^{-2}$.

We will use first-order methods to solve the lower-bound problem, and to this aim, we will need to evaluate the gradient w.r.t. $\omega$ of $f_\omega(\omega, \mu)$. We can apply the envelop theorem $[WTP21]$ to show that for $(\omega, \mu) \in \Sigma_+ \times \Lambda$,

$$\nabla_\omega f_\omega(\omega, \mu) = \frac{(\mu - \lambda_{\omega, \mu}^*(x))^2}{2},$$

where $\Sigma_+ = \Sigma \cap \mathbb{R}_0^K$, $\lambda_{\omega, \mu}^*(x) = \arg\min_{\lambda \in \mathcal{C}_x} \langle \omega, \frac{\mu - \lambda}{2} \rangle$ and $\mathcal{C}_x$ is the closure of $\mathcal{C}_x$ (refer to Lemma $[19]$ in Appendix $[G, J]$).

2.2 The Linear Maximization Oracle

As mentioned earlier, we assume that we have access to a computationally efficient Oracle, referred to as the LM (Linear Maximization) Oracle, identifying the best action if $\mu$ is known. More precisely, as in existing works in combinatorial semi-bandits $[KWA+14, PPV19, PBVP20]$, we make the following assumption.

**Assumption 1.** (i) There exists a polynomial-time algorithm identifying $i^*(v)$ for any $v \in \mathbb{R}^K$; (ii) $\mathcal{X}$ is inclusion-wise maximal, i.e., there is no $x, x' \in \mathcal{X}$ s.t. $x < x'$; (iii) for each $k \in [K]$, there exists $x \in \mathcal{X}$ such that $x_k = 1$; (iv) $|\mathcal{X}| \geq 2$.

Assumption $[1]$ holds for combinatorial sets including $m$-sets, spanning forests, bipartite matching, $s$-$t$ paths. For completeness, we verify the assumption for these action sets in Appendix $[J]$. In the design of our MCP algorithm, we will actually need to solve for some $v \in \mathbb{R}^K$ the linear maximization problem $\max \{x, v\}$ over $\mathcal{X} \setminus \{i^*(\mu)\}$; in other words, we will probably need to identify the second best action. Fortunately, this can be done in a computationally efficient manner under Assumption $[1]$. The following lemma formalizes this observation. Its proof, presented in Appendix $[J]$ is inspired by Lawler’s $m$-best algorithm $[Law72]$.

**Lemma 2.** Let $v \in \mathbb{R}^K$ and $x \in \mathcal{X}$. Under Assumption $[1]$ there exists an algorithm that solves $\max_{x' \in \mathcal{X} : x' \neq x} \langle v, x' \rangle$ by only making at most $D$ queries to the LM Oracle.

3 Solving the lower bound problem: the MCP algorithm

Solving the lower bound problem first requires to evaluate its objective function $F_\mu(\omega)$. A naive approach, enumerating $f_\omega(\omega, \mu)$ for all $x \in \mathcal{X} \setminus \{i^*\}$, would be computationally infeasible. In this section, we present and analyze MCP, an algorithm that approximates $F_\mu(\omega)$ by calling the LM Oracle a number of times growing at most polynomially in $K$.

3.1 Lagrangian relaxation

The first step towards the design of MCP consists in considering the Lagrangian relaxation of the optimization problem defining $f_\omega(\omega, \mu) = \inf_{\lambda \in \mathcal{C}_x} \langle \omega, \frac{\mu - \lambda}{2} \rangle$ (see e.g., $[BV04, Vis21]$). For any $(\omega, \mu) \in \Sigma_+ \times \Lambda$ and $x \neq i^*$, the Lagrangian $\mathcal{L}_{\omega, \mu}$ and Lagrange dual function $g_{\omega, \mu}$ of this problem are defined as, for all $\alpha \geq 0$,

$$\mathcal{L}_{\omega, \mu}(\lambda, x, \alpha) = \langle \omega, \frac{\mu - \lambda}{2} \rangle + \alpha \langle i^* - x, \lambda \rangle \quad \text{and} \quad g_{\omega, \mu}(x, \alpha) = \inf_{\lambda \in \mathbb{R}^K} \mathcal{L}_{\omega, \mu}(\lambda, x, \alpha),$$

respectively. The following proposition, proved in Appendix $[C, J]$ provides useful properties of $g_{\omega, \mu}$:

**Proposition 1.** Let $(\omega, \mu) \in \Sigma_+ \times \Lambda$ and $x \in \mathcal{X} \setminus \{i^*(\mu)\}$.

(a) The Lagrange dual function is linear in $x$. More precisely, $g_{\omega, \mu}(x, \alpha) = c_{\omega, \mu}(\alpha) + (\mathcal{L}_{\omega, \mu}(\alpha), x)$.
where $c_{\omega, \mu}(\alpha) = \alpha \langle \mu - \frac{\alpha}{2} \omega^{-1}, i^* (\mu) \rangle$ and $\ell_{\omega, \mu}(\alpha) = -\alpha \langle \mu + \frac{\alpha}{2} \omega^{-1}, (1 - 2i^* (\mu)) \rangle$.

(b) $g_{\omega, \mu}(x, \cdot)$ is strictly concave (for any fixed $x$).

c) $f_{x}(\omega, \mu) = \max_{\alpha \geq 0} g_{\omega, \mu}(x, \alpha)$ is attained by $\alpha^*_x = \frac{\Delta_{\omega}(\mu)}{\langle x, i^* (\mu), \omega^{-1} \rangle}$.

d) $\|\ell_{\omega, \mu}(\alpha^*_x)\|_1 \leq L_{\omega, \mu} = 4D^2 K \|\mu\|_\infty \|\omega^{-1}\|_\infty$.

From Proposition 1(c), strong duality holds for the program defining $f_x(\omega, \mu)$, and we conclude:

$$F_\mu(\omega) = \min_{\eta \neq i^*} \max_{\alpha \geq 0} g_{\omega, \mu}(x, \alpha).$$

$F_\mu(\omega)$ can hence be seen as the value in a two-player game. The aforementioned properties of the Lagrange dual function will help to compute this value.

### 3.2 Solving the two-player game with no regret

There is a rich and growing literature on solving zero-sum games using no-regret algorithms, see for example [RS13, ALL18, FEG21, ZOD21]. Our game has the particularity that the $x$-player has a discrete combinatorial action set whereas the $\omega$-player has a convex action set. Importantly, for this game, we wish not only to estimate its value $F_\mu(\omega)$ but also an equilibrium action $x_\omega$ such that $F_\mu(\omega) = \max_{\alpha \geq 0} g_{\omega, \mu}(x_\omega, \alpha)$. Indeed, an estimate of $x_\omega$ will be needed when implementing the Frank-Wolfe algorithm and more specifically when estimating the gradient of $F_\mu(\omega)$. To return such an estimate, one could think of leveraging results from the recent literature on last-iterate convergence, see e.g. [DP19, GPDO20, LNP21, WILZ21, APPS22, AAS23]. However, most of these results concern saddle-point problems only, and are not applicable in our setting. Here, we adopt a much simpler solution, and take advantage of the properties of the Lagrange dual function $g_{\omega, \mu}(x, \alpha)$ to design an iterative procedure directly leading to estimates of $(F_\mu(\omega), x_\omega)$. In this procedure, the two players successively update their actions until a stopping criterion is met, say up to the $N$-th iteration.

The procedure generates a sequence $\{(x^{(n)}, \alpha^{(n)})\}_{1 \leq n \leq N}$, and from this sequence, estimates $(\hat{F}, \hat{x})$ of $(F_\mu(\omega), x_\omega)$. The details of the resulting MCP algorithm are presented in Algorithm 1.

**x-player.** We use a variant of the Follow-the-Perturbed-Leader (FTPL) algorithm [Han57, KV05]. The $x$-player updates her action as follows:

$$x^{(n)} \in \arg\min_{x \neq i^*} \left( \sum_{m=1}^{n-1} g_{\omega, \mu}(x^{(m)} + \frac{Z_n}{\eta_n}, x) \right) = \arg\min_{x \neq i^*} \left( \sum_{m=1}^{n-1} \ell_{\omega, \mu}(\alpha^{(m)}) + \frac{Z_n}{\eta_n}, x \right),$$

where $Z_n$ is a random vector, exponentially distributed and with unit mean ($\{Z_n\}_{n \geq 1}$ are i.i.d.). Compared to the standard FTPL algorithm, we vary learning rate $\eta_n$ over time to get anytime guarantees (as we do not know a priori when the iterative procedure will stop). This kind of time-varying learning rate was also used in [Neu15] with a similar motivation. Note that thanks to the linearity of $g_{\omega, \mu}$ and Lemma 2 the $x$-player update can be computed using at most $D$ calls to the LM Oracle.

**$\alpha$-player and MCP outputs.** From Proposition 1, $f_x(\omega, \mu) = \max_{\alpha \geq 0} g_{\omega, \mu}(x, \alpha)$. This suggests that the $\alpha$-player can just implement a best-response strategy: after the $x$-player action $x^{(n)}$ is selected, the $\alpha$-player chooses $\alpha^{(n)} = \alpha^*_x = \frac{\Delta_{\omega}(\mu)}{\langle x^{(n)}, i^* (\mu), \omega^{-1} \rangle}$. This choice ensures that $f_{x^{(n)}}(\omega, \mu) = g_{\omega, \mu}(x^{(n)}, \alpha^{(n)})$, and suggests natural outputs for MCP: should it stops after $N$ iterations, it can return $\hat{F} = \min_{n \in [N]} g_{\omega, \mu}(x^{(n)}, \alpha^{(n)})$ and $\hat{x} \in \arg\min_{n \in [N]} g_{\omega, \mu}(x^{(n)}, \alpha^{(n)})$.

**Stopping criterion.** The design of the MCP stopping criterion relies on the convergence analysis and regret from the $x$-player perspective of the above iterative procedure, which we present in the next subsection. This convergence will be controlled by $\ell_{\omega, \mu}(\alpha^*_x)$ and its upper bound $L_{\omega, \mu}$ derived in Proposition 1. Introducing $c_\omega = L_{\omega, \mu}(4 \sqrt{K(\ln K + 1)} + \sqrt{\ln(\theta^{-1})}/2)$, the MCP stopping criterion is: $\sqrt{\theta} > c_\theta(1 + \epsilon)/(\epsilon \hat{F})$. Since $\sqrt{\theta}$ strictly increases with $n$ and since $\hat{F} \geq F_\mu(\omega)$, this criterion ensures that the algorithm terminates in a finite number of iterates. Moreover, as shown in the next subsection, it also ensures that $\hat{F}$ returned by MCP is an $(1 + \epsilon)$-approximation of $F_\mu(\omega)$ with probability at least $1 - \theta$. 

5
We start the analysis by quantifying the regret from Algorithm 1:

\[ \text{Theorem 3. Let } \alpha \text{ and } \beta. \]

To identify an optimal sampling strategy, rather than solving the lower-bound problem in each round \( \alpha \geq 0 \), we devise a Track-and-Stop algorithm \[ (1+\epsilon) \text{P} \] algorithm outputs \( (\hat{F}, \hat{x}) \) satisfying \( \mathbb{P}[F_\mu(\omega) \leq \hat{F} \leq (1+\epsilon)F_\mu(\omega)] \geq 1 - \theta \) and \( \hat{F} = \max_{\alpha \geq 0} g_{\omega, \mu}(\hat{x}, \alpha). \) Moreover, the number of \( \text{LM} \) oracle calls the algorithm does is almost surely at most

\[ \left\lceil \frac{c_0 + 1 - \theta}{c \epsilon^2} \right\rceil \]

4 The Perturbed Frank-Wolfe Sampling (\( \text{P-FWS} \)) algorithm

To identify an optimal sampling strategy, rather than solving the lower-bound problem in each round as a Track-and-Stop algorithm would \[ \epsilon, \theta \in [0, 1]. \] Under Assumption \[ \epsilon \text{, } \theta \in [0, 1]. \]
decision sets, $P$-FW$S$ applies more standard stochastic smoothing techniques as described in the next subsection. All the ingredients of $P$-FW$S$ are gathered in §4.2. By design, the algorithm just leverages the MCP algorithm as a subroutine, and hence only requires the LM Oracle. In §4.3, we analyze the performance of $P$-FW$S$.

4.1 Smoothing the objective function $F_\mu$

Here, we present and analyze a standard stochastic technique to smooth a function $\Phi$. In $P$-FW$S$, this technique will be applied to the objective function $\Phi = F_\mu$. Let $\Phi : \mathbb{R}_{\geq 0}^K \mapsto \mathbb{R}$ be a concave and $\ell$-Lipschitz function. Assume that the set of points where $\Phi$ is not differentiable is of Lebesgue-measure zero. To smooth $\Phi$, we can take its average value in a neighborhood of the point considered, see e.g. [FKM05]. Formally, we define the stochastic smoothed approximate of $\Phi$ as:

$$\tilde{\Phi}_\eta(\omega) = \mathbb{E}_{\omega \sim \text{Uniform}(B_2)}[\Phi(\omega + \eta Z)],$$

where $B_2 = \{\omega \in \mathbb{R}^K : \|\omega\|_2 \leq 1\}$ and $\eta \in (0, \min_k \|\omega_k\|)$. The following proposition lists several properties of this smoothed function, and gathers together some of the results from [DBW12], see Appendix A for more details.

**Proposition 2.** For any $\omega \in \Sigma_+$ and $\eta \in (0, \min_k \|\omega_k\|)$, $\tilde{\Phi}_\eta$ satisfies: (i) $\Phi(\omega) - \ell \Omega \leq \tilde{\Phi}_\eta(\omega)$ and $\Phi(\omega)$; (ii) $\tilde{\Phi}_\eta(\omega)$, $\omega$; (iii) $\tilde{\Phi}_\eta(\omega)$; (iv) $\tilde{\Phi}_\eta(\omega)$.

Note that with (i), we may control the approximation error between $\tilde{\Phi}_\eta$ and $\Phi$ by $\eta$. (ii) and (iii) ensure the differentiability and smoothness of $\Phi$, respectively. (iv) is equivalent to $\Phi(\omega') \leq \Phi(\omega) + \nabla \Phi(\omega) \cdot (\omega' - \omega) + \frac{\ell \Omega}{\|\omega\|} \|\omega - \omega'\|_2^2$, $\forall \omega, \omega' \in \Sigma_+$. Finally, (iv) stems from the concavity of $\Phi$, and implies that the value $\tilde{\Phi}_\eta(\omega)$ monotonously increases while $\eta$ decreases, and it is upper bounded by $\Phi(\omega)$ thanks to (i). The above results hold for $\Phi = F_\mu$. Indeed, first it is clear that the definition (2) of $F_\mu$ can be extended to $\mathbb{R}^K$; then, it can be shown that $F_\mu$ is Lipschitz-continuous and almost-everywhere differentiable – refer to Appendices A and B for formal proofs.

4.2 The algorithm

Before presenting $P$-FW$S$, we need to introduce the following notation. For $t \geq 1$, $k \in [K]$, we define $N_k(t) = \sum_{s=1}^t \mathbb{1} \{x_k(s) = 1\}$, $\hat{\omega}_k(t) = N_k(t) / t$, and $\hat{\mu}(t) = \sum_{k=1}^K y_k(s) / N_k(t)$ when $N_k(t) > 0$.

**Sampling rule.** The design of the sampling rule is driven by the following objectives: (i) the empirical allocation should converge to the solution of the lower-bound problem [1], and (ii) the number of calls to the LM Oracle should be controlled. To meet the first objective, we need in the Frank-Wolfe updates to plug an accurate estimator of $\mu$ in. The accuracy of our estimator will be guaranteed by alternating between forced exploration and FW update sampling phases. Now for the second objective, we also use forced exploration phases when in a Frank-Wolfe update, the required number of calls to the LM Oracle predicted by the upper bound presented in Theorem 5 is too large. In view of Lemma 1 and Theorem 3, this happens in round $t$ if $\|\hat{\mu}(t-1)\|_\infty$ or $\Delta_{\min}(\hat{\mu}(t-1))^{-1}$ is too large. Next, we describe the forced exploration and Frank-Wolfe update phases in detail.

**Forced exploration.** Initially, $P$-FW$S$ applies the LM Oracle to compute the forced exploration set $X_0 = \{i^*(e_k) : k \in [K]\}$, where $e_k$ is the $K$-dimensional vector whose $k$-th component is equal to one and zero elsewhere. $P$-FW$S$ then selects each action in $X_0$ once. Note that Assumption 1 (iii) ensures that the $k$-th component of $i^*(e_k)$ is equal to one. In turn, this ensures that $X_0$ is a $[K]$-covering set, and that playing actions from $X_0$ is enough to estimate $\mu$. $P$-FW$S$ starts an exploration phase at rounds $t$ such that $\sqrt{t/|X_0|}$ is an integer or such that the maximum of $\Delta_{\min}(\hat{\mu}(t-1))^{-1}$ and $\|\hat{\mu}(t-1)\|_\infty$ is larger than $\sqrt{t-1}$. Whenever this happens, $P$-FW$S$ pulls each $x \in X_0$ once.

**Frank-Wolfe updates.** When in round $t$, the algorithm is not in a forced exploration phase, it implements an iteration of the Frank-Wolfe algorithm applied to maximize the smoothed function $\tilde{F}_\mu(t-1), \eta(t) \omega(t-1) = \mathbb{E}_{\omega \sim \text{Uniform}(B_2)}[\tilde{F}_\mu(t-1)(\omega(t-1) + \eta(t) Z)]$. The sequence of parameters $\{\eta_t\}_{t \geq 1}$ is chosen to ensure that $\eta_t$ chosen in $(0, \min_k \hat{\omega}_k(t))$, and hence $\omega(t-1) + \eta(t) Z \in \mathbb{R}_{\geq 0}$. Also note that in a round $t$ where the algorithm is not in a forced exploration phase,
Algorithm 2: P-\text{FW} (\{(\epsilon_t, \eta_t, n_t, \rho_t, \theta_t)\}_t)

\textbf{initialization:}

\begin{align*}
\text{for } k = 1, \ldots, K \text{ do} \\
& X_0 \leftarrow \arg\max_{x \in X} \langle e_k, x \rangle \text{ (tie broken arbitrarily)} \\
\text{end}
\end{align*}

Sample \( x \in X_0 \) in a round-robin manner for \( t |X_0| \) rounds; update \( \hat{\mu}(4|X_0|) \) and \( \hat{\omega}(4|X_0|) \);

\begin{align*}
\text{for } t = 4|X_0| + 1, \ldots \text{ do} \\
\text{if } \sqrt{t/|X_0|} \notin \mathbb{N} \text{ or } \max\{\Delta_{\min}(\hat{\mu}(t-1))^{-1}, \|\hat{\mu}(t-1)\|_{\infty}\} > \sqrt{t-1} \text{ then} \\
& \text{Sample each } x \in X_0 \text{ once, update } \hat{\mu}(t) \text{ and } \hat{\omega}(t), \text{ and } t \leftarrow t + |X_0| - 1; \\
\text{else} \\
& \text{Compute } \nabla F_{\hat{\mu}(t-1), \eta, n_t}(\hat{\omega}(t-1)) \text{ by } (\rho_t, \theta_t)\text{-MCP algorithm;} \\
& x(t) \leftarrow i^∗(\nabla F_{\hat{\mu}(t-1), \eta, n_t}(\hat{\omega}(t-1))); \\
& \text{Sample } x(t) \text{ and update } \hat{\mu}(t) \text{ and } \hat{\omega}(t); \\
\text{end} \\
\text{if } \max\{\Delta_{\min}(\hat{\mu}(t))^{-1}, \|\hat{\mu}(t)\|_{\infty}\} \leq \sqrt{t} \text{ then} \\
& \hat{F}_t \leftarrow \left(\epsilon_t, \delta/t^2\right)\text{-MCP}(\hat{\omega}(t), \hat{\mu}(t)); \\
& \text{if } t\hat{F}_t > (1 + \epsilon_t) \beta \left(1, 1 - \frac{1}{4|X_0|}\right)\text{ then break;}
\end{align*}

return \( i = i^∗(\hat{\mu}(t)) \);

by definition \( \Delta_{\min}(\hat{\mu}(t-1)) > 0 \). This implies that \( \hat{\mu}(t-1) \in \Lambda \) and that \( F_{\hat{\mu}(t-1)} \) are well-defined. Now an ideal FW update would consist in playing an action \( i^∗(\nabla F_{\hat{\mu}(t-1), \eta, n_t}(\hat{\omega}(t-1))) = \arg\max_{x \in X} \langle \nabla F_{\hat{\mu}(t-1), \eta, n_t}(\hat{\omega}(t-1)), x \rangle \), see e.g. [Jag13]. Unfortunately, we do not have access to \( \nabla F_{\hat{\mu}(t-1), \eta, n_t}(\hat{\omega}(t-1)) \). But the latter can be approximated, as suggested in Proposition 2(ii), by \( \nabla F_{\hat{\mu}(t-1), \eta, n_t}(\hat{\omega}(t-1)) = \frac{1}{n_t} \sum_{m=1}^{n_t} \nabla f_{\hat{x}_m}(\hat{\omega}(t-1) + \eta_t Z_m, \hat{\mu}(t-1)) \). \( \text{P-FW} \) uses this approximation and the \( \text{LM Oracle} \) to select the action: \( x(t) \in i^∗(\nabla F_{\hat{\mu}(t-1), \eta, n_t}(\hat{\omega}(t-1))) \). The choices of the parameters \( \eta_t, n_t, \rho_t \) and \( \theta_t \) do matter. \( \eta_t \) impacts the sample complexity and should converge to 0 as \( t \to \infty \) so that \( F_{\eta_t, n_t}(\hat{\omega}) \to F_{\hat{\mu}}(\hat{\omega}) \) at any point \( \hat{\omega} \in \Sigma_+ \) (this is a consequence of Proposition 2(iii)). \( \eta_t \) should not decay too fast however as it would alter the smoothness of \( F_{\hat{\mu}, \eta_t} \). We will show that \( \eta_t \) should actually decay as \( 1/\sqrt{t} \). \( (n_t, \rho_t, \theta_t) \) impact the trade-off between the sample complexity and the computational complexity of the algorithm. We let \( n_t \to \infty \) and \( (\rho_t, \theta_t) \to 0 \) as \( t \to \infty \) so that \( \langle \nabla F_{\hat{\mu}, \eta_t, n_t}(\hat{\omega}) - \nabla F_{\hat{\mu}, \eta_t, \hat{\omega}}(x), x \rangle \to 0 \) for any \( (\hat{\omega}, x) \in \Sigma_+ \times \mathcal{X} \).

\textbf{Stopping and decision rule.} As often in best arm identification algorithms, the \( \text{P-FW} \) stopping rule takes the form of a GLRT:

\[ \tau = \inf \left\{ t > 4|X_0| : \frac{t\hat{F}_t}{1 + \epsilon_t} > \beta \left(1, 1 - \frac{1}{4|X_0|}\right)\delta, \max\{\Delta_{\min}(\hat{\mu}(t))^{-1}, \|\hat{\mu}(t)\|_{\infty}\} \leq \sqrt{t} \right\}, \]

where \( \epsilon_t \in \mathbb{R}_{>0}, \hat{F}_t \) is returned by the \( (\epsilon_t, \delta/t^2)\text{-MCP}(\hat{\omega}(t), \hat{\mu}(t)) \) algorithm. The function \( \beta \) satisfies

\[ \begin{align*}
& \forall t \geq 1, \quad t\hat{F}_t(\hat{\omega}(t)) \geq \beta(t, \delta) \quad \Rightarrow \quad (\text{P-}i^∗(\hat{\mu}(t)) \neq i^∗(\hat{\mu})) \leq \delta, \\
& \exists c_1, c_2 > 0 : \quad \forall t \geq c_1, \beta(t, \delta) \leq \frac{c_2 t}{\delta}. \tag{9}
\end{align*} \]

Examples of function \( \beta \) satisfying the above conditions can be found in [GKT16, JP20, KK21]. The condition \( (8) \) will ensure the \( \delta \)-correctness of \( \text{P-FW} \), whereas \( (9) \) will control its sample complexity. Finally, the action returned by \( \text{P-FW} \) is simply defined as \( i = i^∗(\hat{\mu}(\tau)) \). The complete pseudo-code of \( \text{P-FW} \) is presented in Algorithm 2.

\footnote{Our Julia implementation could be found at \url{https://github.com/rctzeng/NeurIPS2023-PerturbedFWS}}
4.3 Non-asymptotic performance analysis of \(P-FWS\)

The following theorem provides an upper bound of the sample complexity of \(P-FWS\) valid for any confidence level \(\delta\), as well as the computational complexity of the algorithm.

**Theorem 4.** Let \(\mu \in \Lambda\) and \(\delta \in (0, 1)\). If \(P-FWS\) is parametrized using

\[
(\epsilon, \eta, n, \rho, \theta, \theta) = \left( t^{-\frac{1}{2}}, \frac{1}{4\sqrt{t} \mid X_0 \mid}, \left[ t^{\frac{1}{2}} \right], \frac{1}{16tD^2 \mid X_0 \mid}, \frac{1}{t^{\frac{1}{2}}e^{\sqrt{t}}} \right),
\]

then (i) the algorithm finishes in finite time almost surely and \(\mathbb{P}_\mu[i \neq \hat{i}(\mu)] \leq \delta\); (ii) its sample complexity satisfies \(\mathbb{P}_\mu[\|i \neq \hat{i}(\mu)\| \leq \delta]\) = 1 and for any \(\epsilon, \tilde{\epsilon} \in (0, 1)\) with \(\epsilon < \min\{1, 2D^2 \Delta_{\min}^2, D^2 \|\mu\|_2^2\}\),

\[
\mathbb{E}_\mu[\tau] \leq \frac{(1 + \tilde{\epsilon})^2}{T^*(\mu)^{-1} - 6\epsilon} \times H\left( \frac{1}{\delta} \cdot rac{4c_2}{3} \cdot \frac{(1 + \tilde{\epsilon})^2}{T^*(\mu)^{-1} - 6\epsilon} \right) + \Psi(\epsilon, \tilde{\epsilon}),
\]

where \(H(x) = \ln x + \ln \ln x + 1\) and \(\Psi(\epsilon, \tilde{\epsilon})\) (refer to (34) for a detailed expression) is polynomial in \(e^{-1}, \tilde{e}^{-1}, K, \|\mu\|_\infty\), and \(\Delta_{\min}(\mu)^{-1}\); (iii) the expected number of \(LM\) Oracle calls is upper bounded by a polynomial in \(\ln \delta^{-1}, K, \|\mu\|_\infty\), and \(\Delta_{\min}(\mu)^{-1}\).

The above theorem establishes the statistical asymptotic optimality of \(P-FWS\) since it implies that \(\lim_{\delta \to 0} \mathbb{E}_\mu[\tau] / \ln(1/\delta) \leq \left(1 + \frac{1}{\epsilon}\right)^2/(T^*(\mu)^{-1} - 6\epsilon)\). This upper bound matches the sample complexity lower bound \([1]\) when \(\epsilon \to 0\) and \(\tilde{\epsilon} \to 0\).

**Proof sketch.** The complete proof of Theorem 4 is presented in Appendix D.

(i) **Correctness.** To establish the \(\delta\)-correctness of the algorithm, we introduce the event \(\mathcal{G}\) under which \(\hat{F}_t\), computed by \((\epsilon, \eta, \|\mathcal{F}\|)\)-MCP \((\hat{\omega}(t), \hat{\mu}(t))\), is an \((1 + \epsilon\delta)\)-approximation of \(F^*_{\mu}(t)\) for each round \(t \geq 4|X_0| + 1\). From Theorem 3, we deduce that \(\mathbb{P}_\mu[\mathcal{G}^c] \leq \sum_{t=4|X_0|+1}^{\infty} \frac{\delta}{|X_0|} \leq \delta / 4|X_0|\). In view of (38), this implies that \(\mathbb{P}_\mu[i \neq \hat{i}(\mu)] \leq \delta\).

(ii) **Non-asymptotic sample complexity upper bound.**

Step 1. (Concentration and certainty equivalence) We first define two good events, \(\mathcal{E}_t^{(1)}\) and \(\mathcal{E}_t^{(2)}\). \(\mathcal{E}_t^{(1)}\) corresponds to the event where \(\hat{\mu}(t)\) is close to \(\mu\), and its occurrence probability can be controlled using the forced exploration rounds and concentration inequalities. Under \(\mathcal{E}_t^{(1)}\), the selected action \(x(t)\) is close to the ideal FW update. Again using concentration results and the performance guarantees of MCP given in Theorem 3, we can control the occurrence probability of \(\mathcal{E}_t^{(2)}\). Overall, we show that \(\sum_{t=1}^{\infty} \mathbb{P}_\mu[\mathcal{E}_t^{(1)} \cap \mathcal{E}_t^{(2)}] < \infty\). To this aim, we derive several important continuity results presented in Appendix G. These results essentially allow us to study the convergence of the smoothed FW updates as if the certainty equivalence principle held, i.e., as if \(\hat{\mu}(t) = \mu\).

Step 2. (Convergence of the smoothed FW updates) We study the convergence assuming that \((\mathcal{E}_t^{(1)} \cap \mathcal{E}_t^{(2)})\) holds. We first show that \(\Phi_{\mu, \eta, \xi} = \ell\)-Lipschitz and smooth for \(\ell = 2D^2 \|\mu\|_2^2\), see Appendices H and I. Then, in Appendix E, we establish that the dynamics of \(\phi_t = \max_{\omega \in \Sigma} \Phi_{\mu}(\omega) - F^*_{\mu}(\omega(t))\) satisfy \(t \phi_t \leq (t - 1) \phi_{t-1} + \ell (\eta_t + \frac{1}{2} \phi_t^2)\). Observe that, as mentioned earlier, \(1/\sqrt{t}\) is indeed the optimal scaling choice for \(\eta_t\). We deduce that after a certain finite number \(T_1\) of rounds, \(\phi_t\) is sufficiently small and \(\max\{\Delta_{\min}(\mu(t))^{-1}, \|\mu(t)\|_\infty\} \leq \sqrt{t}\).

Step 3. Finally, we observe that \(\mathbb{E}_\mu[\tau] \leq T_1 + \sum_{t=T_1}^{\infty} \mathbb{P}_\mu[t \Phi_t \leq \beta \left( t, (1 - \frac{2}{\pi \|X_0\|}) \delta \right) + \frac{\sum_{t=T_1}^{\infty} \mathbb{P}_\mu[\mathcal{E}_t^{(1)} \cap \mathcal{E}_t^{(2)}]}{\mathcal{E}_t^{(1)} \cap \mathcal{E}_t^{(2)}}] , \] and show that the second term in the r.h.s. in this inequality is equivalent to \(T^*(\mu) \ln(1/\delta)\) as \(\delta \to 0\) using the property of the function \(\beta\) defining the stopping threshold and similar arguments as those used in [GK16, WTP21].

(iii) **Expected number of \(LM\) Oracle calls.** The MCP algorithm is called to compute \(\hat{F}_t\) and to perform the FW update only in rounds \(t\) such that \(\max\{\Delta_{\min}(\mu(t))^{-1}, \|\mu(t)\|_\infty\} \leq \sqrt{t}\). Thus, from Theorem 3 and Lemma 1, the number of \(LM\) Oracle calls per-round is a polynomial in \(t\) and \(K\). As the \(\mathbb{E}_\mu[\tau]\) is polynomial in \(\ln \delta^{-1}, K, \|\mu\|_\infty\) and \(\Delta_{\min}^{-1}\), the expected number of \(LM\) Oracle calls is also polynomial in the same variables.
5 Related Work

We provide an exhaustive survey of the related literature in Appendix B. To summarize, to the best of our knowledge, CombGame [JMKK21] is the state-of-the-art algorithm for BAI in combinatorial semi-bandits in the high confidence regimes. A complete comparison to P-FWS is presented in Appendix B. CombGame was initially introduced in [DKM19] for classical bandit problems. There, the lower-bound problem is casted as a two-player game and the authors propose to use no-regret algorithms for each player to solve it. [JMKK21] adapts the algorithm for combinatorial semi-bandits, and provides a non-asymptotic sample complexity upper bound matching (1) asymptotically. However, the resulting algorithm requires to call an oracle solving the Most-Confusing-Parameter problem as our MCP algorithm. The authors of [JMKK21] conjectured the existence of such a computationally efficient oracle, and we establish this result here.

6 Conclusion

In this paper, we have presented P-FWS, the first computationally efficient and statistically optimal algorithm for the best arm identification problem in combinatorial semi-bandits. For this problem, we have studied the computational-statistical trade-off through the analysis of the optimization problem leading to instance-specific sample complexity lower bounds. This approach can be extended to study the computational-statistical gap in other learning tasks. Of particular interest are problems with an underlying structure (e.g. linear bandits [DMSV20, JP20], or RL in linear / low rank MDPs [AKKS20]). Most results on these problems are concerned with statistical efficiency, and ignore computational issues.

Acknowledgments and Disclosure of Funding

We thank Aristides Gionis and the anonymous reviewers for their valuable feedback. The research is funded by ERC Advanced Grant REBOUND (834862), the Wallenberg AI, Autonomous Systems and Software Program (WASP), and Digital Futures.

References


Contents

1 Introduction 1

2 Preliminaries 3
   2.1 The lower-bound problem 4
   2.2 The Linear Maximization Oracle 4

3 Solving the lower bound problem: the MCP algorithm 4
   3.1 Lagrangian relaxation 4
   3.2 Solving the two-player game with no regret 5
   3.3 Performance analysis of the MCP algorithm 6

4 The Perturbed Frank-Wolfe Sampling (P-FWS) algorithm 6
   4.1 Smoothing the objective function $\mathcal{F}_\mu$ 7
   4.2 The algorithm 7
   4.3 Non-asymptotic performance analysis of P-FWS 9

5 Related Work 10

6 Conclusion 10

A Table of Notation 16

B Further related work 17

C Results related to our $(\epsilon, \theta)$-MCP algorithm 19
   C.1 Properties of Lagrangian dual of $f_x$ 19
   C.2 Analysis of MCP 20
   C.3 Regret analysis of Follow-the-Perturbed-Leader 21

D Analysis of P-FWS 24
   D.1 $\delta$-correctness (Theorem 4(i)) 24
   D.2 Almost-sure upper bound (Theorem 4(ii)) 24
   D.3 Non-asymptotic sample complexity (Theorem 4(ii)) 25
      D.3.1 Good events 26
      D.3.2 Proof of non-asymptotic sample complexity 27
      D.3.3 Technical lemmas 28
   D.4 Computational complexity (Theorem 4(iii)) 29

E Convergence of P-FWS under the good events 31

F Upper bound of $\sum_{T=M+1}^{\infty} \mathbb{P}_\mu[(E_{1,\epsilon}(T) \cap E_{2,\epsilon}(T))^c]$ under P-FWS 34
   F.1 Lemmas for bounding $\mathbb{P}_\mu[E_{1,\epsilon}(T)^c]$ 35
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>F.2</td>
<td>Lemmas for bounding $\mathbb{P}<em>\mu[\mathcal{E}</em>{2,\epsilon}(T)]$</td>
<td>36</td>
</tr>
<tr>
<td>F.3</td>
<td>Technical lemmas</td>
<td>36</td>
</tr>
<tr>
<td>G</td>
<td>Continuity arguments</td>
<td>38</td>
</tr>
<tr>
<td>G.1</td>
<td>An application of the maximum theorem</td>
<td>39</td>
</tr>
<tr>
<td>G.2</td>
<td>The length of gradients</td>
<td>40</td>
</tr>
<tr>
<td>H</td>
<td>Stochastic smoothing</td>
<td>43</td>
</tr>
<tr>
<td>I</td>
<td>Lipschitzness of $F_\mu$ and boundness of $F_\mu$ on $\Sigma_K \cap \mathbb{R}^K_{&gt;0}$</td>
<td>44</td>
</tr>
<tr>
<td>J</td>
<td>Proofs related to combinatorial sets</td>
<td>45</td>
</tr>
<tr>
<td>K</td>
<td>Sample complexity lower bound</td>
<td>47</td>
</tr>
<tr>
<td>L</td>
<td>Extension to the transductive setting</td>
<td>48</td>
</tr>
</tbody>
</table>
# A Table of Notation

## Problem setting

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$</td>
<td>Number of arms</td>
</tr>
<tr>
<td>$[m]$ for any $m \in \mathbb{N}$</td>
<td>The set ${1, 2, \ldots, m}$</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Required uncertainty</td>
</tr>
<tr>
<td>$\mu \in \mathbb{R}^K$</td>
<td>Vector of the expected rewards of the various arms</td>
</tr>
<tr>
<td>$E_{\mu}$ and $P_{\mu}$</td>
<td>The expectation and probability measure corresponding to $\mu$</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>${\mu \in \mathbb{R}^K :</td>
</tr>
<tr>
<td>$i^*(\mu)$</td>
<td>Best arm under parameter $\mu$</td>
</tr>
<tr>
<td>$\mathcal{X}$</td>
<td>Set of actions in ${0, 1}^K$</td>
</tr>
<tr>
<td>$D$</td>
<td>$\max_{x \in X} |x|_1$</td>
</tr>
<tr>
<td>$\Delta_x(\mu)$</td>
<td>$(i^*(\mu) \ominus x, \mu)$</td>
</tr>
<tr>
<td>$\Delta_{\min}(\mu)$</td>
<td>$\min_{x \neq i^*(\mu)} \Delta_x(\mu)$</td>
</tr>
</tbody>
</table>

## Notation related to a given algorithm

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_k(t)$</td>
<td>Number of pulls of arm $k$ up to time $t$</td>
</tr>
<tr>
<td>$\hat{\omega}_k(t)$</td>
<td>$N_k(t)/t$</td>
</tr>
<tr>
<td>$x(t)$</td>
<td>The action taken in time $t$</td>
</tr>
<tr>
<td>$y_k(t)$</td>
<td>Random reward received if $x_k(t) = 1$</td>
</tr>
<tr>
<td>$\hat{\mu}_k(t)$</td>
<td>$\sum_{s=1}^t y_k(s) \mathbb{1}{x_k(s) = 1}/N_k(t)$</td>
</tr>
<tr>
<td>$\tau$</td>
<td>Stopping time</td>
</tr>
<tr>
<td>$\hat{i}$</td>
<td>Recommended action</td>
</tr>
</tbody>
</table>

## Notation used for sets and vectors

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\odot$</td>
<td>Elementwise product</td>
</tr>
<tr>
<td>$\oplus$</td>
<td>Elementwise sum over $\mathbb{Z}_2$</td>
</tr>
<tr>
<td>$x^i$</td>
<td>The $i$-th elementwise power of $x \in \mathbb{R}^K$, i.e., $(x^i_k)_{k \in [K]}$</td>
</tr>
<tr>
<td>$\text{cl} (S)$</td>
<td>The closure of set $S$</td>
</tr>
<tr>
<td>$e_k$</td>
<td>The $K$-dimensional vector whose $k$-th component is equal to one and zero elsewhere</td>
</tr>
</tbody>
</table>

## Properties for lower bound

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d(\mu, \mu')$</td>
<td>KL divergence between the distributions parametrized by $\mu$ and $\mu'$</td>
</tr>
<tr>
<td>$kl(a, b)$</td>
<td>KL divergence between two Bernoulli distributions of means $a$ and $b$</td>
</tr>
<tr>
<td>$\text{Alt} (\mu)$</td>
<td>${\lambda : i^<em>(\lambda) \neq i^</em>(\mu)}$</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>${\sum_{x \in \mathbb{X}} w_x x : w_x \in \Sigma, x \in \mathbb{X}}$ where $\Sigma_N$ is a $(N-1)$-dimensional simplex</td>
</tr>
</tbody>
</table>

## Notation for MCP

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{\mu}(\omega)$</td>
<td>$\min_{x \neq i^*(\mu)} f_x(\omega, \mu)$</td>
</tr>
<tr>
<td>$f_x(\omega, \mu)$</td>
<td>$\inf_{\lambda \in C_x(\omega, \frac{\mu - \lambda}{2})}$, where $C_x = {\lambda : (\lambda, i^*(\mu) - x) &lt; 0}$</td>
</tr>
<tr>
<td>$L_{\omega, \mu}(x, \alpha)$</td>
<td>$\langle \omega, \frac{\mu - \lambda}{2}\rangle + \alpha (i^*(\mu) - x, \lambda)$</td>
</tr>
<tr>
<td>$\psi_{\omega, \mu}(x, \alpha)$</td>
<td>$\inf_{\lambda \in C_x(\omega, \frac{\mu - \lambda}{2})} L_{\omega, \mu}(x, \alpha)$</td>
</tr>
</tbody>
</table>

## Notation for $P$-FWS

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{X}_0$</td>
<td>A $[K]$-covering set</td>
</tr>
<tr>
<td>$\hat{F}_t$</td>
<td>$P$-FWS-approximated value of $F_{\mu(t)}(\hat{\omega}(t))$ for stopping rule</td>
</tr>
<tr>
<td>$F_{\mu, \eta,n}(\cdot)$</td>
<td>$\mathbb{E}<em>{\mathcal{Z} \sim \text{Uniform}(B_2)}[\nabla F</em>{\mu}(\cdot + \eta \mathcal{Z})]$ where $B_2 = {v \in \mathbb{R}^K : |v|_2 \leq 1}$</td>
</tr>
<tr>
<td>$\hat{F}_{\mu, n}$</td>
<td>The empirical $n$-sample estimate of $F_{\mu, \eta}$</td>
</tr>
<tr>
<td>$\ell$</td>
<td>Lipschitz constant of $F_{\mu}$</td>
</tr>
</tbody>
</table>
B Further related work

Combinatorial semi-bandits [CBL12] have found numerous applications including online ranking [DKC21], network routing [CLK+14, KWA+14], loan assignment [KWA+14], path planning problem [JMKK21], and influence marketing [Per22]). We do not discuss these applications here, but rather focus the literature that is the most relevant to our analysis and results.

Solving the lower-bound problem in combinatorial semi-bandits. We are not aware of any computationally efficient algorithm to solve the lower-bound problem, or to compute its objective function. To the best of our knowledge, MCP is the first algorithm to do so. A work closed to ours is [CCG21a] for combinatorial semi-bandits but in the regret minimization. Regret minimization yields a different lower-bound problem. There exits a statistically optimal algorithm [CMP17], called OSSB, that matches the regret lower bound by [CTMSP+15]. OSSB requires to solve the lower-bound problem in each round, and the authors [CCG21a] are the first to investigate whether this is at all possible in a computationally efficient way. They establish that if budgeted-linear maximization (BLM) [RG96, BBGS11] can be solved within an $\epsilon$-approximation factor for the combinatorial set $X$, then the lower-bound problem can be approximately solved with a precision depending on $\epsilon$. As a consequence, the approach leads to an algorithm with asymptotically minimal regret only if one has access to an exact BLM solver. This is the case for $m$-sets and $s$-$t$ paths but this is not the case for spanning trees and perfect matchings. For the latter case, as mentioned [CCG21a], an algorithm using an approximately correct BLM solver would not be statistically optimal.

Best arm identification in combinatorial semi-bandits. Many tasks related to combinatorial best arm identification are formulated in the transductive setting [JMKK21], where the set $A \subseteq \{0, 1\}^K$ available for exploration is not necessarily the same as the set $X \subseteq \{0, 1\}^K$ for decision. The minimal sample complexity in the transductive setting is exactly $\mathbb{I}$ with $\Sigma$ replaced with $\{\sum_{x \in A} w_x x : \omega \in \Sigma_{[A]}\}$ - see (58) in Appendix L for details. Two most studied tasks are combinatorial multi-arm bandit (C-MB) where $A = \{e_A\}_{A \subseteq [K]}$ and the best action identification (C-BAI) where $A = X$. The former is arguably simpler than the latter if we compare the corresponding minimal sample complexities (note that $\Sigma_K \supseteq \Sigma$). We note that our results for C-BAI can be easily generalized to the transductive setting (see Appendix L).

Prior works mainly focus on the C-MB task. UCB-based [KTAS12, CLK+14] and elimination-based [CGL16, CGL+17, KSJJ+20] approaches are popular. Among these, EfficientGapElim [CGL+17] achieves the lowest sample complexity $\mathcal{O}(T^*(\mu)(\ln \delta^{-1} + \ln^2 \Delta^{-1})\ln \ln \Delta^{-1} + \ln |X|))$ with high probability, but its computational complexity is hard to analyze. Peace [KSJJ+20], another elimination-based approach by experimental design, requires with high probability a polynomial number of the $\mathcal{L}$M Oracle calls in total. The sample complexity of Peace has a $\delta$-dependent term (scaling as $K)^T(\mu) \ln \delta^{-1}$) worse than EfficientGapElim. Overall, none of these are statistically optimal when $\delta \to 0$. Note that algorithms for linear best-arm identification [DMSV20, WTP21] are applicable to C-MB but not to C-BAI and the general transductive setting.

For the task of C-BAI, we are only aware of two works: GCB-PE [DKC21] and CombGame [JMKK21]. GCB-PE is a UCB-based algorithm with guarantees on the sample complexity and computational complexity valid with high probability only. CombGame [JMKK21] is proposed for the transductive setting, and its design inherits from [DKM19] that interprets the lower-bound problem and more precisely $T^*(\mu)^{-1}$ as the value of a two-player game (a $\omega$-player and a $\lambda$-player). Assuming that an MCP oracle is available, CombGame leverages Frank-Wolfe algorithms, namely OFW [HK12] and LLOO [GH16], for the $\omega$-player and the MCP algorithm for the $\lambda$-player. [JMKK21] leaves the existence of such an oracle running in polynomial time as an open problem. Our MCP algorithm resolves this issue. CombGame is statistically optimal in the high confidence regime but has no clear guarantees in the moderate regime [BGK22].

We wish to finally mention an algorithm that has inspired the design of $\mathcal{P}$–FWS. This algorithm is referred to as Frank-Wolfe Sampling (FWS) [WTP21]. FWS is optimal in high confidence regime

---

4In Section 4.5 in [CGL+17], the authors provide a lemma stating that: if parallel simulation is additionally allowed, then any high-probability sample complexity upper bound can be converted to an upper bound in expectation.

5Note that this two-player game is different than the two-player game involved in our algorithm MCP.
but is not computationally efficient for combinatorial semi-bandits. For example, to deal with the non-smoothness issue of the objective function $F_{\mu}$, FW-S needs to construct the so-called $r$-subdifferentiable spaces and to optimize a linear function on these spaces. Unfortunately, these spaces can be generated by a number of vectors exponentially increasing with $K$ in combinatorial semi-bandits. Moreover, in moderate confidence regime, the sample complexity upper bound derived in [WTP21] has an exponential dependence in $K$.

All the relevant algorithms, their sample complexity guarantees and computational complexity are summarized in Table 1.

Table 1: Algorithms for best-arm identification in combinatorial semi-bandits with fixed confidence and their performance.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Task</th>
<th>Instance-specific Sample Complexity</th>
<th>Computational Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peace</td>
<td>C-MB</td>
<td>$\text{poly}(K, \Delta_{\text{min}}^{-1}, \ln \delta^{-1})$ w.h.p.</td>
<td>$\times$ LP solver (√) $\text{poly}(K, \Delta_{\text{min}}^{-1}, \delta^{-1})$ w.h.p.</td>
</tr>
<tr>
<td>GCB-PE</td>
<td>C-BAI</td>
<td>$\text{poly}(K, \Delta_{\text{min}}^{-1}, \ln \delta^{-1})$ w.h.p.</td>
<td>$\times$ - $\text{poly}(K, \Delta_{\text{min}}^{-1}, \ln \delta^{-1})$ w.h.p.</td>
</tr>
<tr>
<td>CombGame</td>
<td>Trans.</td>
<td>$\times$ (incomparable)</td>
<td>√ MCP (√)</td>
</tr>
<tr>
<td>P-FWS</td>
<td>Trans.</td>
<td>$\text{poly}(K, \Delta_{\text{min}}^{-1}, \ln \delta^{-1})$</td>
<td>√ MCP (√) $\text{poly}(K, \Delta_{\text{min}}^{-1}, \ln \delta^{-1})$</td>
</tr>
</tbody>
</table>
C Results related to our \((\epsilon, \theta)\)-MCP algorithm

C.1 Properties of Lagrangian dual of \(f_x\)

**Proposition** \([1]\) Let \((\omega, \mu) \in \Sigma_+ \times \Lambda\) and \(x \in \mathcal{X} \setminus \{i^* (\mu)\}\).

(a) The Lagrange dual function is linear in \(x\). More precisely, \(g_{\omega, \mu}(x, \alpha) = c_{\omega, \mu}(\alpha) + \ell_{\omega, \mu}(\alpha, x)\) where \(c_{\omega, \mu}(\alpha) = \alpha \langle \mu - \frac{\omega}{2} \omega^{-1}, i^*(\mu) \rangle\) and \(\ell_{\omega, \mu}(\alpha) = -\alpha \langle \mu + \frac{\omega}{2} \omega^{-1} \circ (1_K - 2i^*(\mu)) \rangle\).

(b) \(g_{\omega, \mu}(x, \cdot)\) is strictly concave (for any fixed \(x\)).

(c) \(f_x(\omega, \mu) = \max_{\alpha \geq 0} g_{\omega, \mu}(x, \alpha)\) is attained by \(\alpha^*_x = \frac{\Delta_x(\mu)}{\langle x \oplus i^*(\mu), \omega^{-1} \rangle}\).

(d) \(\|\ell_{\omega, \mu}(\alpha^*_x)\|_1 \leq L_{\omega, \mu} = 4D^2 K \|\mu\|_\infty \|\omega^{-1}\|_\infty\).

**Proof** Fix any \((\omega, \mu) \in \Sigma_+ \times \Lambda\) and let \(i^* = i^*(\mu)\) for short. For convenience, the definition of \(L_{\omega, \mu}\) and \(g_{\omega, \mu}\) are restated:

\[
L_{\omega, \mu}(\lambda, x, \alpha) = \left\langle \omega, \frac{(\mu - \lambda)^2}{2} \right\rangle + \alpha \langle i^* - x, \lambda \rangle \quad \text{and} \quad g_{\omega, \mu}(x, \alpha) = \inf_{\lambda \in \mathbb{R}^K} L_{\omega, \mu}(\lambda, x, \alpha).
\]

Proof of (a): linearity of \(g_{\omega, \mu}(\cdot, \cdot)\): Let \(\lambda_{x, \mu}(x, \alpha) = \arg \inf_{\lambda \in \mathbb{R}^K} L_{\omega, \mu}(\lambda, x, \alpha)\). The first-order condition implies that \(0 = \nabla_{\lambda} L_{\omega, \mu}(\lambda_{x, \mu}(x, \alpha), x, \alpha) = \omega \odot (\lambda_{x, \mu}(x, \alpha) - \mu) + \alpha (i^* - x)\), which directly yields (as \(\omega > 0\))

\[
\lambda_{x, \mu}(x, \alpha) = \mu + \alpha \omega^{-1} \odot (x - i^*).
\]

We plug \([11]\) into \(L_{\omega, \mu}(\lambda_{x, \mu}(x, \alpha), x, \alpha)\) and directly obtain that

\[
g_{\omega, \mu}(x, \alpha) = \left\langle \omega, \frac{\alpha^2}{2} \omega^{-2} \odot (x - i^*)^2 \right\rangle + \alpha \langle \mu, i^* - x \rangle - \alpha^2 \langle \omega^{-1}, (x - i^*)^2 \rangle
\]

\[
= \alpha \langle \mu, i^* - x \rangle - \frac{\alpha^2}{2} \langle \omega^{-1}, (x - i^*)^2 \rangle
\]

\[
= c_{\omega, \mu}(\alpha) + \langle \ell_{\omega, \mu}(\alpha), x \rangle,
\]

where \([13]\) follows from the fact that \((x - i^*)^2 = i^* - 2x \odot i^* + x = i^* + x \odot (1_K - 2i^*)\).

Proof of (b): strict concavity of \(g_{\omega, \mu}(x, \cdot)\): This is trivial from \([12]\).

Proof of (c): \(f_x(\omega, \mu) = \max_{\alpha \geq 0} g_{\omega, \mu}(x, \alpha)\) is attained by \(\alpha^*_x = \frac{\Delta_x(\mu)}{\langle x \oplus i^*(\mu), \omega^{-1} \rangle}\): For a fixed \(x \neq i^*\), by the first-order condition of \([12]\), we find that the maximum of \(g_{\omega, \mu}(x, \cdot)\) is reached at

\[
\alpha^*_x = \frac{\Delta_x(\mu)}{\langle \omega^{-1}, (x - i^*)^2 \rangle} = \frac{\Delta_x(\mu)}{\langle x \oplus i^*(\mu), \omega^{-1} \rangle},
\]

where for the second equality, we use the assumption that \(i^*\) and \(x\) are binary vectors and hence \((x - i^*)^2 = x \oplus i^*\). We now verify that \((\alpha^*_x, \lambda_{x, \mu}^*)\), where \(\lambda_{x, \mu}^* = \lambda_{x, \mu}(x, \alpha^*_x)\) (see \([11]\)), satisfies KKT conditions, which is equivalent to strong duality (refer to \([\text{Vis} \& \text{BV}04]\)) under Slater’s condition (there exists a \(\lambda \in \mathbb{R}^K\) such that the constraint is strict). Since \(x\) is a suboptimal action, \(\Delta_x(\mu)\) is positive, so is \(\alpha^*_x\) (dual feasibility). To verify \(\langle \lambda_{x, \mu}^*, i^* - x \rangle \leq 0\) (primal feasibility), the definition of \(\lambda_{x, \mu}^*(\cdot, \cdot)\), \([11]\), yields

\[
\langle \lambda_{x, \mu}^*, i^* - x \rangle = \Delta_x(\mu) + \alpha_x^* \langle \omega^{-1} \odot (x - i^*), (i^* - x) \rangle
\]

\[
= \Delta_x(\mu) - \alpha_x^* \langle \omega^{-1}, x \oplus i^* \rangle = 0,
\]

which implies that \(\alpha_x^* (i^* - x, \lambda_{x, \mu}^*) = 0\) (complementary slackness). Finally, stationarity holds automatically as \(\nabla_{\lambda} L_{\omega, \mu}(\lambda_{x, \mu}^*, x, \alpha) = 0\) for all \(\alpha\).

Proof of (d): \(\|\ell_{\omega, \mu}(\alpha^*_x)\|_1 \leq L_{\omega, \mu} = 4D^2 K \|\mu\|_\infty \|\omega^{-1}\|_\infty\): Following from the expression of \(\ell_{\omega, \mu}(\alpha)\), we have \(\ell_{\omega, \mu}(\alpha^*_x) = -\alpha^*_x \mu + \alpha^*_x \omega^{-1} \odot (1_K - 2i^*)\). Observe that \(\|\mu\|_1 \leq K \|\mu\|_\infty \leq
By Proposition 1 (a), and the coordinate of $1_K - 2i^*$ is either 1 or −1, a simple application of triangle inequality leads to

$$\|\ell_{\omega, \mu}(\alpha_x^*)\|_1 \leq K\|\omega^{-1}\|_\infty \left(\|\mu\|_\infty + \frac{\alpha_x^*}{2}\right)\alpha_x^*.$$  

As for $\alpha_1^*$ (see [14]), $\triangle_x(\mu) \leq 2D\|\mu\|_\infty$ and $\langle \omega^{-1} , x \oplus i^* \rangle \geq \min_k \omega^{-1}_k \geq 1$, hence we conclude that $\|\ell_{\omega, \mu}(\alpha_x^*)\|_1 \leq 2D(D+1)K\|\omega^{-1}\|_\infty \leq L_{\omega, \mu}$.

C.2 Analysis of MCP

**Theorem 3** Let $\epsilon, \theta \in (0, 1)$. Under Assumption[7] for any $(\omega, \mu) \in \Sigma_+ \times \Lambda$, the $(\epsilon, \theta)$-MCP $(\omega, \mu)$ algorithm outputs $(\hat{F}, \hat{x})$ satisfying

$$P\left[ F_{\mu}(\omega) \leq \hat{F} \leq (1 + \epsilon)F_{\mu}(\omega) \right] \geq 1 - \theta \quad \text{and} \quad \hat{F} = \max_{\alpha \geq 0} g_{\omega, \mu}(\hat{x}, \alpha).$$

Moreover, the number of LM Oracle calls the algorithm does is almost surely at most

$$\left[ \frac{c_0^2(1 + \epsilon)^2}{c^2F_{\mu}(\omega)^2} \right] = \mathcal{O}\left(\frac{\|\omega^{-1}\|_\infty^4 K^2D^5 \ln K \ln \theta^{-1}}{c^2F_{\mu}(\omega)^2}\right).$$

**Proof** Fix any $(\omega, \mu) \in \Sigma_+ \times \Lambda$ and denote by $i^* = i^*(\mu)$. Suppose Algorithm[1] reaches the stopping criterion at the $N$-th iteration.

Guarantees on the outputs of MCP: By Proposition[1](a),

$$\sum_{n=1}^{N} g_{\omega, \mu}(x^{(n)}, \alpha^{(n)}) - \min_{x \neq i^*} \sum_{n=1}^{N} g_{\omega, \mu}(x, \alpha^{(n)}) = \sum_{n=1}^{N} \langle \ell_{\omega, \mu}(\alpha^{(n)}), x^{(n)} \rangle - \min_{x \neq i^*} \sum_{n=1}^{N} \langle \ell_{\omega, \mu}(\alpha^{(n)}), x \rangle.$$  

The regret of $x$-player can be bounded by applying Lemma[3] resulting in:

$$P\left[ \sum_{n=1}^{N} g_{\omega, \mu}(x^{(n)}, \alpha^{(n)}) - \min_{x \neq i^*} \sum_{n=1}^{N} g_{\omega, \mu}(x, \alpha^{(n)}) \leq c_0\sqrt{N} \right] \geq 1 - \theta. \quad (15)$$

To relate $F_{\mu}(\omega)$ with (15), let $x_c$ be the minimizer attaining $F_{\mu}(\omega) = f_{x_c}(\omega, \mu)$. Then,

$$\min_{x \neq i^*} \sum_{n=1}^{N} g_{\omega, \mu}(x, \alpha^{(n)}) \leq \sum_{n=1}^{N} g_{\omega, \mu}(x_c, \alpha^{(n)}) \leq N \max_{\alpha \geq 0} g_{\omega, \mu}(x_c, \alpha) = NF_{\mu}(\omega). \quad (16)$$

Recall that $\alpha^{(n)}$ is chosen as the best response $\max_{\alpha \geq 0} g_{\omega, \mu}(x^{(n)}, \alpha) = g_{\omega, \mu}(x^{(n)}, \alpha^{(n)})$ and that $\hat{F} = \min_{n \in [N]} g_{\omega, \mu}(x^{(n)}, \alpha^{(n)})$. These together with (16) imply that

$$N(\hat{F} - F_{\mu}(\omega)) \leq \sum_{n=1}^{N} g_{\omega, \mu}(x^{(n)}, \alpha^{(n)}) - \min_{x \neq i^*} \sum_{n=1}^{N} g_{\omega, \mu}(x, \alpha^{(n)}). \quad (17)$$

A simple rearrangement on (15) and (17) implies that: with probability at least 1 − $\theta$,

$$F_{\mu}(\omega) \leq \frac{c_0}{\sqrt{N}} \leq \epsilon(\hat{F} - \frac{c_0}{\sqrt{N}}) \leq \epsilon F_{\mu}(\omega),$$

where the second inequality follows from the stopping criterion that $\sqrt{N} > c_0(1 + \epsilon)/\epsilon \hat{F}$, and the last inequality simply comes from the rearrangement of the first inequality.

**Computational cost:** From the stopping criterion of MCP, we know that

$$N = \left[ \frac{c_0^2(1 + \epsilon)^2}{c^2F_{\mu}(\omega)^2} \right] \leq \left[ \frac{c_0^2(1 + \epsilon^{-1})^2}{F_{\mu}(\omega)^2} \right] = \mathcal{O}\left(\frac{L_{\omega, \mu}^2\left(\sqrt{K \ln K} + \sqrt{\ln \theta^{-1}}\right)^2}{c^2F_{\mu}(\omega)^2}\right).$$
since \( \hat{F} \geq F_\mu(\omega) \) and \( c_\theta = L_{\omega,\mu} \left( 4K \ln K + 1 + \sqrt{\ln(\theta^{-1})/2} \right) \). Finally, as computing each \( x^{(n)} \) takes at most \( D \) calls to \( L_{M,\omega} \) Oracle, the total number of \( L_{M,\omega} \) Oracle calls is
\[
\mathcal{O} \left( \frac{L_{\omega,\mu} D \left( \sqrt{K \ln K} + \sqrt{\ln(\theta^{-1})} \right)^2}{\epsilon^2 F_\mu(\omega)^2} \right) = \mathcal{O} \left( \frac{\|\mu\|_\infty^2 \|\omega^{-1}\|_\infty^2 K^3 D^5 \ln K \ln \theta^{-1}}{\epsilon^2 F_\mu(\omega)^2} \right)
\]
by recalling \( L_{\omega,\mu} = 4D^2 K \|\mu\|_2^2 \|\omega^{-1}\|_\infty \) from Proposition [1](d) and \( (\sqrt{K} + \sqrt{\theta^{-1}})^2 = \mathcal{O}(\ln K \ln \theta^{-1}) \).

C.3 Regret analysis of Follow-the-Perturbed-Leader

In this subsection, we aim at proving Lemma [3] which is a direct consequence of Lemma [4]. One can find similar proofs in e.g. [KV05, Neu15, SN20]. However, the parameter \( \eta_n \) in our MCP algorithm is varying and carefully chosen (without the knowledge of the last round), which makes the proof slightly more complicated.

**Lemma 3.** Let \( N \in \mathbb{N} \). Under \((\epsilon, \theta)\)-MCP(\( \omega, \mu \)), then
\[
\mathbb{P} \left[ 1/N \sum_{n=1}^N g_{\omega,\mu}(x^{(n)}, \alpha^{(n)}) - 1/N \min_{x \neq \hat{x}} \sum_{n=1}^N g_{\omega,\mu}(x, \alpha^{(n)}) \leq \frac{c_\theta}{\sqrt{N}} \right] \geq 1 - \theta.
\]

**Lemma 4.** Let \( \theta \in (0, 1) \) and \( \mathcal{M} \subseteq \{0, 1\}^K \). Given an arbitrary sequence \( \{\ell_n\}_{n \geq 1} \) of vectors in \( \mathbb{R}^K \) whose length \( \|\ell_n\|_1 \) is bounded by \( L > 0 \) for all \( n \in \mathbb{N} \). Suppose \( \{x^{(n)}\}_{n \geq 1} \) is generated by
\[
x^{(n)} \in \text{argmin}_{x \in \mathcal{M}} \left( \sum_{n=1}^{N-1} \langle \ell_m, x \rangle + \frac{\langle Z_n, x \rangle}{\eta_n} \right),
\]
where \( Z_n = (Z_{1,n}, \cdots, Z_{K,n}) \) is a random vector with uncorrelated exponentially distributed (with unit mean) components, and \( \eta_n = \sqrt{K (\ln K + 1)/(4nL^2)} \). Then, for any \( N \in \mathbb{N} \),
\[
\mathbb{P} \left[ \sum_{n=1}^N \langle \ell_n, x^{(n)} \rangle - \min_{x \in \mathcal{M}} \sum_{n=1}^N \langle \ell_n, x \rangle \leq L \sqrt{N} \left( 4K (\ln K + 1) + \sqrt{\frac{\ln \theta^{-1}}{2}} \right) \right] \geq 1 - \theta.
\]

**Proof of Lemma 4:** We will prove this lemma as if \( \{\ell_n\}_{n} \) is chosen in advance since there exists a standard technique for extending regret against oblivious player to the one against nonoblivious one (see Lemma 4.1 in [CBL06]). For convenience, we introduce the following notation. Let \( m^* (\cdot) = \text{argmin}_{x \in \mathcal{M}} \langle \cdot, x \rangle \). Finally, further define global minimizer \( x_* = m^* \left( \sum_{n=1}^N \ell_n \right) \) and an auxiliary vector \( b^{(n)} = m^* \left( \sum_{n=1}^N \ell_n + Z_1/\eta_n \right) \).

It suffices to show the expected regret bound [18].
\[
\mathbb{E} \left[ \sum_{n=1}^N \langle \ell_n, x^{(n)} \rangle \right] - \min_{x \in \mathcal{M}} \sum_{n=1}^N \langle \ell_n, x \rangle \leq 4L \sqrt{NK (\ln K + 1)}.
\]

This is because \( \{\langle \ell_n, x^{(n)} \rangle - \mathbb{E} \left[ \langle \ell_n, x^{(n)} \rangle \right] \}_{n} \) forms a sequence of bounded martingale difference, so an application of a concentration inequality (Lemma [6]) with \( V_n = \langle \ell_n, x^{(n)} \rangle - \mathbb{E} \left[ \langle \ell_n, x^{(n)} \rangle \right] \), \( r_n = L \) for \( n \in [N] \), and \( s = L \sqrt{N \ln \theta^{-1}/2} \) gives that
\[
\mathbb{P} \left[ \sum_{n=1}^N \langle \ell_n, x^{(n)} \rangle - \mathbb{E} \left[ \sum_{n=1}^N \langle \ell_n, x^{(n)} \rangle \right] > L \sqrt{\frac{N \ln \theta^{-1}}{2}} \right] \leq \theta
\]
and combining this with [18] completes the proof.
Proof of (18): We decompose the regret into two terms:

\[ \sum_{n=1}^{N} \langle \ell_n, x^{(n)} - x* \rangle = \sum_{n=1}^{N} \langle b^{(n)} - x*, \ell_n \rangle + \sum_{n=1}^{N} \langle x^{(n)} - b^{(n)}, \ell_n \rangle. \]

(i). We show that \( \mathbb{E} \left[ \sum_{n=1}^{N} \langle b^{(n)} - x*, \ell_n \rangle \right] \leq K(\ln K + 1) \). Invoking Lemma 5 with \( x = x* \) results in

\[
\mathbb{E} \left[ \sum_{n=1}^{N} \langle b^{(n)} - x*, \ell_n \rangle \right] \leq \mathbb{E} \left[ \left\langle x* - \left( \frac{b^{(1)}}{\eta_1} + \sum_{n=2}^{N} \left( \frac{1}{\eta_n} - \frac{1}{\eta_{n-1}} \right) b^{(n)} \right), Z_1 \right\rangle \right].
\]

where the last inequality uses Hölder’s inequality. As all the components of \( x* \) and \( \frac{\eta_n b^{(n)}}{\eta_1} + \sum_{n=2}^{N} \eta_n \left( \frac{1}{\eta_n} - \frac{1}{\eta_{n-1}} \right) b^{(n)} \) are nonnegative and bounded by 1, the 1-norm of their difference is bounded by \( K \). It remains to show

\[
\mathbb{E}[\|Z_1\|_\infty] = \int_0^\infty \mathbb{P} \left[ \max_i Z_{1,i} \geq x \right] dx \leq \int_0^{\ln K} \mathbb{P} \left[ \max_i Z_{1,i} \geq x \right] dx + \int_0^\infty Ke^{-x} dx \leq \ln K + 1.
\]

(ii). We show that \( \mathbb{E} \left[ \sum_{n=1}^{N} \langle x^{(n)} - b^{(n)}, \ell_n \rangle \right] \leq 2L^2 \sum_{n=1}^{N} \eta_n \). Let the pdf of \( \exp(1) \) be \( \pi(\cdot) = e^{-\|\cdot\|_1} \).

\[
\mathbb{E} \left[ \langle b^{(n)}, \ell_n \rangle \right] = \int_{\mathbb{R}^K} m^* \left( \eta_n \sum_{m=1}^{n} \ell_m + z, \ell_n \right) d\pi(z)
\]

\[
= \int_{\mathbb{R}^K} m^* \left( \eta_n \sum_{m=1}^{n} \ell_m + y, \ell_n \right) d\pi(y)
\]

Notice that the triangular inequality implies \( \|y - \eta_n \ell_n\|_1 + \|y\|_1 \leq \|\eta_n \ell_n\|_1 \leq \eta_n L \) and \( e^x \leq 1 + 2x \) for all \( x \in (0,1) \) (Taylor expansion), so recalling \( x^{(n)} = m^* \left( \eta_n \sum_{m=1}^{n-1} \ell_m + Z_1 \right) \), we deduce that

\[
\sum_{n=1}^{N} \mathbb{E} \left[ \langle x^{(n)} - b^{(n)}, \ell_n \rangle \right] \leq \sum_{n=1}^{N} 2\eta_n L \int_{\mathbb{R}^K} m^* \left( \eta_n \sum_{m=1}^{n} \ell_m + z, \ell_n \right) d\pi(z)
\]

\[
\leq \sum_{n=1}^{N} 2\eta_n L \int_{\mathbb{R}^K} \left\| m^* \left( \eta_n \sum_{m=1}^{n} \ell_m + z \right) \right\|_\infty \|\ell_n\|_1 d\pi(z) \leq 2L^2 \sum_{n=1}^{N} \eta_n.
\]

Finally, plugging \( \eta_n = \sqrt{\frac{K(\ln K + 1)}{4nL^2}} \) into (i), and (ii). directly concludes the proof.

The following lemma is a result that can be found in [CBL06, H+16], we rewrite it here for completeness.

Lemma 5. According to \( b^{(n)} = m^* \left( \eta_n \sum_{m=1}^{n} \ell_m + Z_1 \right) \), we can have

\[
\forall x \in M, \quad \sum_{n=1}^{N} \langle b^{(n)} - x, \ell_n \rangle \leq \left\langle \frac{x}{\eta N}, Z_1 \right\rangle - \left\langle \frac{b^{(1)}}{\eta_1} + \sum_{n=2}^{N} \left( \frac{1}{\eta_n} - \frac{1}{\eta_{n-1}} \right) b^{(n)}, Z_1 \right\rangle. \quad (19)
\]
Proof. This is done by induction. For the base case, \( N = 1 \), as \( b^{(1)} = m^* (\ell_1 + Z_1/\eta_1) \)

\[
\langle b^{(1)}, \ell_1 + Z_1/\eta_1 \rangle \leq \langle x, \ell_1 + Z_1/\eta_1 \rangle
\]

for any \( x \in \mathcal{M} \). A simple rearrangement yields (19). While considering \( N + 1 \), we suppose (19) holds for all integers smaller than \( N + 1 \). For an arbitrary \( x \in \mathcal{M} \), the fact \( b^{(N + 1)} = m^* \left( \sum_{n=1}^{N+1} \ell_n + Z_1/\eta_{N+1} \right) \) directly implies that

\[
\langle x, \sum_{n=1}^{N+1} \ell_n + Z_1/\eta_{N+1} \rangle \geq \langle b^{(N+1)}, \sum_{n=1}^{N+1} \ell_n + Z_1/\eta_{N+1} \rangle
\]

\[
= \langle b^{(N+1)}, \ell_{N+1} + \left( \frac{1}{\eta_{N+1}} - \frac{1}{\eta_N} \right) Z_1 \rangle + \langle b^{(N+1)}, \sum_{n=1}^{N} \ell_n + Z_1/\eta_N \rangle
\]

\[
\geq \sum_{n=1}^{N+1} \langle b^{(n)}, \ell_n \rangle + \left( \frac{b^{(1)}}{\eta_1} + \sum_{n=2}^{N+1} \left( \frac{1}{\eta_n} - \frac{1}{\eta_{n-1}} \right) b^{(n)}, Z_1 \right),
\]

where the last inequality comes from applying the hypothesis (19) with \( x = b^{(N+1)} \) on the second inner product. Rearrange the above inequality, our induction is completed. \( \square \)

Lemma 6 (Hoeffding-Azuma). Let \( N \in \mathbb{N} \), \( V_1, V_2, \ldots, V_N \) be a bounded martingale difference sequence w.r.t. \( X_1, X_2, \ldots, X_N \) such that for any \( n \in [N] \) \( V_n \in [A_n, A_n + r_n] \) for some random variable \( A_n \), measurable w.r.t. \( X_1, \ldots, X_{n-1} \) and a positive constant \( r_n \). Then, for any \( s > 0 \),

\[
\mathbb{P} \left[ \sum_{n \in [N]} V_n > s \right] \leq \exp \left( -\frac{2s^2}{\sum_{n \in [N]} r_n^2} \right) \quad \text{and} \quad \mathbb{P} \left[ \sum_{n \in [N]} V_n < -s \right] \leq \exp \left( -\frac{2s^2}{\sum_{n \in [N]} c_n^2} \right).
\]
D Analysis of $\text{P-FWS}$

In this appendix, we prove our main theorem.

**Theorem 4** Let $\mu \in \Lambda$ and $\delta \in (0,1)$. If $\text{P-FWS}$ is parametrized using
\[
(\epsilon_t, \eta_t, n_t, \rho_t, \theta_t) = \left( t^{-\frac{1}{2}}, \frac{1}{4t|X_0|}, \left\lfloor \frac{t}{\epsilon} \right\rfloor, \frac{1}{16tD^2|X_0|}, \frac{1}{t^{\frac{1}{2}}} \right),
\]
then (i) the algorithm finishes in finite time almost surely and $\mathbb{P}_\mu[i \neq i^*(\mu)] \leq \delta$; (ii) its sample complexity satisfies $\mathbb{P}_\mu[N_{\text{MCP}} \leq T^*(\mu)] = 1$ and for any $\epsilon, \delta \in (0,1)$ with $\epsilon < \min\{1, \frac{2D^2\alpha^2}{K}, \frac{D^2\|\mu\|_2^2}{3}\}$,
\[
\mathbb{E}_\mu[\tau] \leq \frac{(1+\delta)^2}{T^*(\mu)^{-1} - 6\epsilon} \times H\left( \frac{1}{\delta} \cdot \frac{4c_2}{3} \cdot \frac{(1+\delta)^2}{T^*(\mu)^{-1} - 6\epsilon} \right) + \Psi(\epsilon, \delta),
\]
where $H(x) = \ln x + \ln\ln x + 1$ and $\Psi(\epsilon, \delta)$ (refer to (34) in Appendix F), (refer to (34) for a detailed expression) is polynomial in $\epsilon^{-1}, \delta^{-1}, K, \|\mu\|_\infty$ and $\Delta_{\min}(\mu)$; (iii) the expected number of MCP Oracle calls is upper bounded by a polynomial in $\ln \delta^{-1}, K, \|\mu\|_\infty$ and $\Delta_{\min}(\mu)^{-1}$.

D.1 $\delta$-correctness (Theorem 4(i))

Recall that $\text{P-FWS}$ stopping rule is:
\[
\tau = \inf \left\{ t > 4|X_0| : \frac{t\hat{F}_t}{1 + \epsilon_t} > \beta \left( t, \left( 1 - \frac{1}{4|X_0|} \right) \frac{\delta}{\|\mu(t)\|_\infty} \right), \max \left\{ \frac{1}{\Delta_{\min}(\hat{\mu}(t)), \|\hat{\mu}(t)\|_\infty} \right\} \leq \sqrt{t} \right\},
\]
where $\hat{F}_t$ is computed by $(\epsilon_t, \delta/\epsilon^2)$-MCP($\hat{\mu}(t), \hat{\mu}(t)$). Let $\hat{i} = \hat{i}^*(\hat{\mu}(\tau))$ be the output of $\text{P-FWS}$. Define the good event $G = \bigcap_{t=4|X_0|+1}^{\infty} \{ \hat{F}_t \leq (1 + \epsilon_t)F_{\hat{\mu}(t)}(\hat{\omega}(t)) \}$. Hence, it follows from the guarantee of $(\epsilon_t, \delta/\epsilon^2)$-MCP algorithm that
\[
\mathbb{P}_\mu[G] \leq \delta \sum_{t=4|X_0|+1}^{\infty} t^{-2} \leq \delta \int_{4|X_0|}^{\infty} x^{-2} dx \leq \frac{\delta}{4|X_0|}.
\]
Besides, under the event $G$,
\[
(1 + \epsilon_t)\tau F_{\hat{\mu}(\tau)}(\hat{\omega}(\tau)) \geq \tau \hat{F}_\tau \geq (1 + \epsilon_t)\beta \left( \tau, \left( 1 - \frac{1}{4|X_0|} \right) \frac{\delta}{\|\mu(t)\|_\infty} \right)
\]
holds, implying that $\tau F_{\hat{\mu}(\tau)}(\hat{\omega}(\tau)) \geq \beta \left( \tau, \left( 1 - \frac{1}{4|X_0|} \right) \frac{\delta}{\|\mu(t)\|_\infty} \right)$. So, by (38) and (9), $\hat{i} = \hat{i}^*(\hat{\mu}(t))$ satisfies:
\[
\mathbb{P}_\mu[i \neq \hat{i}^*(\mu), G] \leq \left( 1 - \frac{1}{4|X_0|} \right) \frac{\delta}{\|\mu(t)\|_\infty}.
\]
and thus $\mathbb{P}_\mu[i \neq \hat{i}^*(\mu)] \leq \mathbb{P}_\mu[i \neq \hat{i}^*(\mu), G] + \mathbb{P}_\mu[G^c] \leq \delta$.

D.2 Almost-sure upper bound (Theorem 4(ii))

In this section, we show Theorem 4(ii) an almost-sure upper bound on the sample complexity for $\text{P-FWS}$. Our proof is based on the continuity of $F_\mu$ in $\mu$ (as in [GK16], [WTP21]) and also on the following observations:

(a) $\{ \hat{\mu}(t) \xrightarrow{t} \mu \}$ and $\nabla \hat{F}_{\hat{\mu}(t), n_t}(\hat{\omega}) \xrightarrow{t} \nabla \hat{F}_{\hat{\mu}(t)}(\hat{\omega})$, $\forall \hat{\omega} \in \Sigma_+$ happen almost surely,
(b) $\hat{F}_t \geq F_{\hat{\mu}(t)}(\hat{\omega}(t))$.

For (a), by the law of large numbers, $\hat{\mu}(t) \xrightarrow{t} \mu$ as $N_k(t) \xrightarrow{t} \infty$ for all $k \in [K]$ yielded by forced exploration rounds involved in $\text{P-FWS}$ (Lemma 14 in Appendix F), $\nabla \hat{F}_{\hat{\mu}(t), n_t}(\omega) \xrightarrow{t} \infty$
\[ \nabla F_{\tilde{\mu}, \gamma}(\omega), \forall \omega \in \Sigma_+ \text{ is a direct consequence that } n_t \xrightarrow{t \to \infty} \infty. \] (b) is immediately derived from the definition of \( \tilde{F}_t \) as \( \tilde{F}_t = f_\gamma(\omega(t), \tilde{\mu}(t)) \) for some action \( \tilde{x} \neq i^*(\tilde{\mu}(t)) \) and \( F_{\mu(t)}(\omega(t)) = \min_{x \in X \setminus i^*(\mu(t))} f_\gamma(\omega(t), \mu(t)). \)

Introduce the event
\[ \mathcal{E} = \left\{ F_{\mu}(\omega(t)) \xrightarrow{t \to \infty} \max_{\omega \in \Sigma} F_{\mu}(\omega) \text{ and } \tilde{\mu}(t) \xrightarrow{t \to \infty} \mu \right\}. \]

Because of (a), Theorem 5 in Appendix E ensures that \( \mathbb{P}_\mu[\mathcal{E}] = 1 \). Also, by the uniform continuity of \( F_{\mu}(\omega) \) in \( \mu \) for an arbitrary \( \omega \in \Sigma_+ \) (Lemma 7 in D.3.3),
\[
\max_{\omega \in \Sigma_+} |F_{\mu(t)}(\omega) - F_{\mu}(\omega)| \xrightarrow{t \to \infty} 0
\]
almost surely, and hence by the triangle inequality, this implies that
\[
\mathbb{P}_\mu \left[ F_{\mu(t)}(\omega(t)) \xrightarrow{t \to \infty} \max_{\omega \in \Sigma} F_{\mu}(\omega) \right] = 1.
\]

For any \( \epsilon \in (0, 1) \), under \( \mathcal{E} \), there exists a positive integer \( T_\epsilon > \max \{ c_1, 4|\lambda_0| \} \) such that for any \( t \geq T_\epsilon \), we have
\[
F_{\mu(t)}(\omega(t)) \geq (1 - \epsilon) \max_{\omega \in \Sigma} F_{\mu}(\omega), \quad \max \left\{ \frac{1}{\Delta_{\min}(\mu(t))}, ||\mu(t)||_\infty \right\} \leq \sqrt{t}, \quad \text{and } \epsilon_t \leq \epsilon,
\]
where the second inequality is due to (a) and the third is because \( \epsilon_t \to 0 \). So, the stopping time (7) can be upper bounded by
\[
\tau \leq T_\epsilon + \inf \left\{ t > T_\epsilon : t \tilde{F}_t > (1 + \epsilon) \beta \left( t, \frac{4|\lambda_0| - 1}{4|\lambda_0|} \right) \right\}
\]
\[
\leq T_\epsilon + \inf \left\{ t > T_\epsilon : tF_{\mu(t)}(\omega(t)) > (1 + \epsilon) \beta \left( t, \frac{4|\lambda_0| - 1}{4|\lambda_0|} \right) \right\}
\]
\[
\leq T_\epsilon + \inf \left\{ t > T_\epsilon : t(1 - \epsilon) \max_{\omega \in \Sigma} F_{\mu}(\omega) > (1 + \epsilon) \beta \left( t, \frac{4|\lambda_0| - 1}{4|\lambda_0|} \right) \right\}
\]
\[
\leq T_\epsilon + \inf \left\{ t > T_\epsilon : \frac{(1 - \epsilon)t}{(1 + \epsilon)T^*(\mu)} > \ln \left( \frac{c_2 t}{\delta} \cdot \frac{4|\lambda_0|}{4|\lambda_0| - 1} \right) \right\}
\]
\[
\leq 2T_\epsilon + \left( 1 + \frac{\epsilon}{1 - \epsilon} \right) T^*(\mu) H \left( \frac{1}{\delta} \cdot \frac{8c_2}{7} \left( \frac{1 + \epsilon}{1 - \epsilon} \right) T^*(\mu) \right).
\]

where the first inequality uses the last two inequalities of (20), the second inequality uses (b), the third inequality is based on the first inequality of (20), the fourth uses \( T^*(\mu)^{-1} = \max_{\omega \in \Sigma} F_{\mu}(\omega) \) and (9), and the last inequality results from \( 4|\lambda_0| \leq 8/7 \) (as \( |\lambda_0| \geq 2 \)), and an application of Lemma 2 with
\[
\alpha = 1, \quad b_1 = \frac{1 - \epsilon}{1 + \epsilon} \cdot \frac{1}{T^*(\mu)} \quad \text{and} \quad b_2 = \frac{8c_2}{7} \cdot \frac{1}{\delta}.
\]

Finally, as \( \epsilon \in (0, 1) \) can be arbitrarily small, (21) implies that
\[
\mathbb{P}_\mu \left[ \limsup_{\delta \to 0} \frac{\tau}{\ln \delta^{-1}} \leq T^*(\mu) \right] = 1.
\]

D.3 Non-asymptotic sample complexity (Theorem 4(ii))

We establish the following non-asymptotic upper bound on \( \mathbb{E}_\mu[\tau] \): for any \( \tilde{\epsilon}, \epsilon \in (0, 1) \) small enough,
\[
\mathbb{E}_\mu[\tau] \leq \frac{(1 + \tilde{\epsilon})^2}{T^*(\mu)^{-1} - 6\epsilon} H \left( \frac{1}{\delta} \cdot \frac{8c_2}{7} \cdot \frac{(1 + \tilde{\epsilon})^2}{T^*(\mu)^{-1} - 6\epsilon} \right) + \Psi(\epsilon, \tilde{\epsilon}),
\]
where \( H(x) = \ln x + \ln \ln x + 1 \) and \( \Psi(\epsilon, \tilde{\epsilon}) \) is defined in (34).
Note that this directly implies the asymptotic optimality. Indeed, when $\delta \to 0$, we get:

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\mu}[r]}{\ln \delta^{-1}} \leq \frac{(1 + \bar{\epsilon})^2}{T^*(\mu)^{-1} - 6\epsilon}.$$

As $\epsilon, \bar{\epsilon}$ can be set arbitrarily small and $\text{kl}(\delta, 1 - \delta) \approx \ln \delta^{-1}$ as $\delta \to 0$, it matches the sample complexity lower bound \[\text{[1]}\] (Theorem 7 in Appendix K) asymptotically.

Throughout this section, we assume $\mu \in \Lambda$ is given and take any $\epsilon \in (0, 1)$ satisfying the following:

$$\epsilon < \min \left\{1, \frac{2D^2 \Delta_{\min}^2}{K}, \frac{1}{6T^*(\mu)} \right\} \leq \min \left\{1, \frac{2D^2 \Delta_{\min}^2}{K}, \frac{D^2 \|\mu\|_\infty^2}{3} \right\},$$

where the second inequality is because $T^*(\mu)^{-1} \leq \ell = 2D^2 \|\mu\|_\infty^2$ by Lemma 22 in Appendix I. The assumption of $\epsilon < \min \left\{1, \frac{2D^2 \Delta_{\min}^2}{K}, \frac{D^2 \|\mu\|_\infty^2}{3} \right\}$ is used to define the good events introduced in \[\text{D.3.1}\] as well as to derive several necessary technical lemmas summarized in \[\text{D.3.3}\].

**D.3.1 Good events**

Since in early rounds, the estimation of $\hat{\mu}(t)$ is noisy, we introduce two threshold functions, $\underline{h}$ and $\bar{h}$, on the round index $T$:

$$\begin{align*}
\underline{h}(T) &= \min \{t \in \mathbb{N} : t \geq T^0, \sqrt{t/|X_0|} \in \mathbb{N}\}, \\
\bar{h}(T) &= \min \{t \in \mathbb{N} : t \geq T^0 \underline{h}(T), \sqrt{t/|X_0|} \in \mathbb{N}\},
\end{align*}$$

where $a, b \in (0, 1)$ and $a + b < 1$ will be explained later in \[\text{(27)}\]. Now, we define our good events:

$$\mathcal{E}_{1, \epsilon}(T) = \bigcap_{t=\underline{h}(T)}^{T} \mathcal{E}^{(t)}_{1, \epsilon} \quad \text{and} \quad \mathcal{E}_{2, \epsilon}(T) = \bigcap_{t=\underline{h}(T)}^{T} \mathcal{E}^{(t)}_{2, \epsilon},$$

where $\mathcal{E}^{(t)}_{1, \epsilon} = \{\langle \nabla F\hat{\mu}(t-1), \eta_1(\hat{\omega}(t-1)), x(t) \rangle \geq \max_{x \in X} \langle \nabla \bar{F}\hat{\mu}(t-1), \eta_1(\hat{\omega}(t-1)), x \rangle - \epsilon \}$ and $\mathcal{E}^{(t)}_{2, \epsilon} = \{\|\hat{\mu}(t-1) - \mu\|_\infty^2 < \frac{\|\mu\|_\infty^2}{M} \}$.

$\mathcal{E}_{1, \epsilon}$ is the event when the solution of FW update is bounded by at most $\epsilon$, and $\mathcal{E}^{(t)}_{2, \epsilon}$ is the event when the empirical estimate of $\mu$ is sufficiently accurate. Under $\mathcal{E}_{2, \epsilon}$, the uniform continuity shown in Lemma 22 in \[\text{D.3.3}\] ensures that:

$$|F\hat{\mu}(t-1) - F\hat{\mu}(\omega)| < \epsilon, \quad \forall \omega \in \Sigma_+,$$

$$|\langle \nabla F\hat{\mu}(t-1), \eta_1(\hat{\omega}(t-1)) \rangle - \langle \nabla F\hat{\mu}(t-1), \eta_1(\omega(2)), \omega \rangle | < \epsilon, \quad \forall (\omega, \omega) \in \Sigma_+ \times X, \forall \eta \in (0, \min_{k \in [K]} \omega_k).$$

The second inequality enables the duality gap of FW algorithm to be controlled, leading to the convergence of $\hat{F}$. Let

$$M = \max \left\{ \frac{4K^2}{\epsilon^2 D^2 |X_0|}, \left( \frac{2}{\Delta_{\min}(\mu)} \right)^{\frac{1}{2}}, \left( \frac{3 \|\mu\|_\infty}{2} \right)^{\frac{1}{2}} \right\}$$

$$+ \max \left\{ \left( \frac{\ell}{\epsilon} \right)^{\frac{1}{2}}, \left( \frac{5\ell K^2}{\epsilon \sqrt{|X_0|}} \right)^{\frac{1}{2}} \right\},$$

then overall, we have (Theorem 5 in \[\text{D.3.3}\]): for any $t \geq \bar{h}(M)$,

$$\max_{\omega \in X} F\hat{\mu}(\omega) - F\hat{\mu}(\hat{\omega}(t)) \leq 5\epsilon, \quad \Delta_{\min}(\hat{\mu}(t)) \geq \frac{\Delta_{\min}(\mu)}{2}, \quad \text{and} \quad \|\hat{\mu}(t)\|_\infty^2 \leq \frac{3 \|\mu\|_\infty^2}{2}. \quad \text{(26)}$$

Finally, the values of $a, b$ are set to the following:

$$a = \frac{7}{9} \quad \text{and} \quad b = \frac{1}{9}. \quad \text{(27)}$$

This choices will balance the leading order between $\epsilon^{-1}$ and $\bar{\epsilon}^{-1}$ in the $\delta$-independent terms \[\text{(34)}\] of the non-asymptotic upper bound (which will be shown later).
D.3.2 Proof of non-asymptotic sample complexity

Let $\delta \in (0, 1)$. We claim that:

$$
E_\mu[\tau] \leq \sum_{T=1}^{\infty} \mathbb{P}_\mu[\tau \geq T] \leq T_0(\delta) + \sum_{T=M+1}^{\infty} \mathbb{P}_\mu[(E_{1,\epsilon}(T) \cap E_{2,\epsilon}(T))^c],
$$

(28)

where $T_0(\delta) = \inf \left\{ T \geq M : \bar{h}(T) + \frac{(1+\epsilon)^2}{4\epsilon} \beta \left( T, \frac{1}{4} \right) \leq T \right\}$. The proof is completed by bounding each term in the right-hand side of (28).

Proof of (28): Suppose $T \geq M$ and $E_{1,\epsilon}(T) \cap E_{2,\epsilon}(T)$ holds. Observe that

$$
\min\{\tau, T\} \leq \bar{h}(T) + \sum_{t=\lceil \bar{h}(T) \rceil}^{T} I\{\tau > t\}.
$$

To derive an upper bound of $\sum_{t=\lceil \bar{h}(T) \rceil}^{T} I\{\tau > t\}$, recall the stopping rule (7) that

$$
\tau = \inf \left\{ t > 4|\mathcal{X}_0| : \frac{t - (1+\epsilon))}{1+\epsilon} \beta \left( t, \frac{4|\mathcal{X}_0| - 1}{4|\mathcal{X}_0|} \right), \max \left\{ \frac{1}{\Delta_{\min}(\mu(t))}, \|\mu(t)\|_\infty \right\} \leq \sqrt{t} \right\}
$$

$$
\leq \inf \left\{ t \geq \bar{h}(M) : t \geq (1+\epsilon) \beta \left( t, \frac{4|\mathcal{X}_0| - 1}{4|\mathcal{X}_0|} \right) \right\}
$$

$$
\leq \inf \left\{ t \geq \bar{h}(M) : t(T^*(\mu)^{-1} - 6\epsilon) \geq (1 + \epsilon) \beta \left( t, \frac{4|\mathcal{X}_0| - 1}{4|\mathcal{X}_0|} \right) \right\},
$$

where the first inequality uses (26), and the second follows from Lemma 7 and Theorem 5 in D.3.3 \( |F_{\mu(t-1)}(\omega(t-1)) - F_{\mu}(\omega(t-1))| < \epsilon \) and \( T^*(\mu)^{-1} - F_{\mu}(\omega(t)) \leq 5\epsilon \), (29) and the fact that $\tilde{F}_t \geq F_{\mu(t)}(\omega(t)).$ Hence, $\sum_{t=\lceil \bar{h}(M) \rceil}^{T} I\{\tau > t\}$ is upper bounded by

$$
\sum_{t=\lceil \bar{h}(T) \rceil}^{T} I\{t(T^*(\mu)^{-1} - 6\epsilon) \geq (1 + \epsilon) \beta \left( t, \frac{4|\mathcal{X}_0| - 1}{4|\mathcal{X}_0|} \right) \} \leq \frac{(1+\epsilon)^2}{4\epsilon} \beta \left( T, \frac{1}{4} \right) = \frac{T^*(\mu)^{-1} - 6\epsilon}{1+\epsilon} \beta \left( T, \frac{1}{4} \right).
$$

By defining $T_0(\delta)$ as done in (28), we get (28), i.e.,

$$
E_\mu[\tau] \leq \sum_{T=1}^{\infty} \mathbb{P}_\mu[\tau \geq T] \leq T_0(\delta) + \sum_{T=M+1}^{\infty} \mathbb{P}_\mu[(E_{1,\epsilon}(T) \cap E_{2,\epsilon}(T))^c]
$$

because $E_{1,\epsilon}(T) \cap E_{2,\epsilon}(T) \subseteq \{\tau \leq T\}$ for any $T \geq T_0(\delta)$.

Now, we proceed with the proof by upper-bounding each term in the right-hand side of (28).

Bounding $T_0(\delta)$: Introduce $\bar{\epsilon} \in (0, 1)$ that can be chosen arbitrarily small. Notice that

$$
T - \bar{h}(T) = T - T^a b^{\alpha} \geq \frac{T}{1 + \bar{\epsilon}} \quad \text{when} \quad T \geq \left( \frac{1 + \bar{\epsilon}}{\bar{\epsilon}} \right)^{-\frac{1}{\alpha + \beta}},
$$

(30)

$$
\bar{\epsilon} = T^{-\frac{1}{\alpha}} \leq \bar{\epsilon} \quad \text{when} \quad T \geq \left( \frac{1}{\bar{\epsilon}} \right)^{-\frac{1}{\beta}},
$$

(31)

where the first inequality results from a simple rearrangement, and the second substitutes $\epsilon_t = t^{-1/9}$. Then, it follows from (9) that:

$$
T_0(\delta) \leq \inf \left\{ T \geq \max \left\{ M, (1 + \bar{\epsilon})^{-\frac{1}{\alpha + \beta}} \right\} : \left( \frac{1 + \bar{\epsilon}}{\bar{\epsilon}} \right) \beta \left( \frac{4c_1T}{3\delta} \right) \leq \frac{T}{1 + \bar{\epsilon}} \right\}
$$

$$
\leq \inf \left\{ T \geq \max \left\{ M, (1 + \bar{\epsilon})^{-\frac{1}{\alpha + \beta}} \right\} : \ln \left( \frac{4c_1T}{3\delta} \right) \leq \frac{T^*(\mu)^{-1} - 6\epsilon}{(1 + \bar{\epsilon})^2} \right\}
$$

$$
\leq \max \left\{ M, (1 + \bar{\epsilon})^{-\frac{1}{\alpha + \beta}} \right\} + \frac{(1 + \bar{\epsilon})^2}{T^*(\mu)^{-1} - 6\epsilon} \times H \left( \frac{4c_1}{3\delta} \cdot \frac{(1 + \bar{\epsilon})^2}{T^*(\mu)^{-1} - 6\epsilon} \right),
$$

(32)
where the first inequality uses (30)-(31) and \(\frac{4|x_0|-1}{4|x_0|} \geq \frac{3}{2}\) (as \(|x_0| \geq 2\) is shown in Lemma 23 in Appendix [I], the second inequality is due to (9), and the last results from an application of Lemma 9 in Appendix [D.3.3] with
\[
\alpha = 1, \quad b_1 = \frac{T^*(\mu)^{-1} - 6\varepsilon}{(1 + \varepsilon)^2}, \quad \text{and} \quad b_2 = \frac{4c_2}{3\delta}.
\]
Bounding \(\sum_{T=M+1}^{\infty} \mathbb{P}_{\mu}[(E_{1,\varepsilon}(T) \cap E_{2,\varepsilon}(T))^c]\) By Lemma 8 in Appendix [D.3.3] it is upper bounded by
\[
2K \left( \min\left(1, \frac{3}{829D^2\|X\|^2} \right)^{2 + \frac{2}{3}} + 2 \left( \frac{2304D^6 \|\mu\|_\infty \sqrt{|x_0|}}{\delta^2} \right)^{2 + \frac{2}{3}} \right) \Gamma \left( 2 + \frac{2}{a} \right).
\]
Putting things together: Finally, substituting \((a, b) = (\frac{7}{\varepsilon}, \frac{1}{\varepsilon})\) into (25)-(32)-(33) yields that:
\[
\begin{align*}
&\cdot T_0(\delta) \leq M + \left(1 + \frac{1}{\varepsilon}\right)^9 + \left(\frac{1}{\varepsilon}\right)^9 + c_1 + \frac{(1 + \varepsilon)^2}{T^*(\mu)^{-1} - 6\varepsilon} \cdot H \left( \frac{4c_2}{3\delta} \cdot \frac{(1 + \varepsilon)^2}{T^*(\mu)^{-1} - 6\varepsilon} \right) \\
&\cdot M \leq \max\left\{\left(\frac{|4\|X_0|\|}{\varepsilon^2D^2\|X_0\|}\right)^2, \left(\frac{4K^2}{\varepsilon^2D^2}\|X_0\|\right)^{\frac{2}{3}}, \left(\frac{4}{\delta}\right), \left(\frac{\delta}{4}\right)^{\frac{2}{3}}, \left(\frac{\delta}{4}\right)^{\frac{2}{3}}\right\} + \max\left\{\left(\frac{\delta}{4}\right)^{\frac{2}{3}}, \left(\frac{\delta}{4}\right)^{\frac{2}{3}}\right\} + \max\left\{\left(\frac{\delta}{4}\right)^{\frac{2}{3}}, \left(\frac{\delta}{4}\right)^{\frac{2}{3}}\right\} + \max\left\{\left(\frac{\delta}{4}\right)^{\frac{2}{3}}, \left(\frac{\delta}{4}\right)^{\frac{2}{3}}\right\}
\end{align*}
\]
where simplifications are obtained remarking that \(\Gamma(2 + \frac{2}{a}) \leq 13\) and \(2 + \frac{2}{a} < 5\). Therefore, substituting \(\ell = 2D^2 \|\mu\|_\infty\) (defined in Appendix [I] and \(78 < 2^7, 3^9 \leq 2^{15}\), and \(4^9/7 < 6\) lead to:
\[
\mathbb{E}_{\mu}[\tau] \leq \frac{(1 + \varepsilon)^2}{T^*(\mu)^{-1} - 6\varepsilon} \cdot H \left( \frac{4c_2}{3\delta} \cdot \frac{(1 + \varepsilon)^2}{T^*(\mu)^{-1} - 6\varepsilon} \right) + \Psi(\varepsilon, \varepsilon),
\]
where
\[
\Psi(\varepsilon, \varepsilon) = \max\left\{\frac{|X_0|}{\varepsilon^2D^2\|X_0\|}, \left(\frac{1}{\varepsilon^2D^2\|X_0\|}\right)^{\frac{1}{2}}, \left(\frac{1}{\varepsilon^2D^2\|X_0\|}\right)^{\frac{2}{3}}, \frac{\|\mu\|_\infty^{\frac{1}{3}}}{\varepsilon^3}, \frac{\|\mu\|_\infty^{\frac{1}{2}}}{\varepsilon^{\frac{2}{3}}} \right\}
\]
\[
+ \left(1 + \frac{1}{\varepsilon}\right)^9 + \left(\frac{1}{\varepsilon}\right)^9 + c_1 + \frac{232KD^3\|\mu\|_\infty^{10}}{\varepsilon^{10}} \left(\frac{K^5\|\mu\|_\infty^{10} + 2^{15}\|X_0\|^{2.5}}{(1 + \varepsilon)^2} \right).
\]

D.3.3 Technical lemmas

The most important step in Theorem 4 is to bound the term \(\sum_{T=M+1}^{\infty} \mathbb{P}_{\mu}[(\mathcal{E}_{1,\varepsilon}(T) \cap \mathcal{E}_{2,\varepsilon}(T))^c]\) in (28) explicitly in terms of \(K, \|\mu\|_\infty\) and \(\varepsilon\). For this purpose, inspired by Assumption 3 in [WTP21], we developed Proposition 4 (see Appendix [G.2] for the proof) and combine it with the mean-valued theorem to derive our main continuity results in Lemma 7 (see Appendix [G] for the proof). Throughout this section, we fix \(\mu \in \mathcal{X}\) and denote \(\Delta_{\min}(\mu)\) by \(\Delta_{\min}\).

**Lemma 7.** Let \(\varepsilon \in (0, \frac{2D^2\Delta_{\min}^2}{K})\). Then, any \(\pi \in \mathbb{R}^K\) with \(\|\pi - \mu\|_\infty \leq \frac{\varepsilon}{24D^2\|\mu\|_\infty}\) satisfies the following:
\[
|F_\mu(\omega) - F_\pi(\omega)| < \varepsilon, \quad \forall \omega \in \Sigma_+ \\
|\langle \nabla F_{\pi,\eta}(\omega) - \nabla F_{\mu,\eta}(\omega), x - \omega \rangle| < \varepsilon, \quad \forall (\omega, x) \in \Sigma_+ \times \mathcal{X}, \forall \eta \in (0, \min_{k \in [K]} \omega_k).
\]

Our main concentration result with error specified explicitly in terms of \(\varepsilon\) is (see Appendix [F] for the proof):

**Lemma 8.** Let \(\varepsilon \in (0, \frac{2D^2\Delta_{\min}^2}{K})\) and \(M\) be defined as in (25). Then,
\[
\sum_{T=M+1}^{\infty} \mathbb{P}_{\mu}[(E_{1,\varepsilon}(T) \cap E_{2,\varepsilon}(T))^c] < 2K \left( \frac{3}{A_1(\varepsilon)^{2 + \frac{2}{3}}} + \frac{2}{A_2(\varepsilon)^{2 + \frac{2}{3}}} \right) \frac{\Gamma \left( 2 + \frac{2}{a} \right)}{\varepsilon^{10}}
\]
where \(A_1(\varepsilon) = \min\{1, \frac{\varepsilon^2}{829D^2\|X_0\|}\}\), \(A_2(\varepsilon) = \frac{\varepsilon^2}{2304D^6\|\mu\|_\infty \sqrt{|x_0|}}\), and \(\Gamma\) denotes the gamma function \(\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt\) for any \(z > 0\).
We remark that Lemma 8 sharpens a similar result, Lemma 2 in [WIP21], by a factor of $e^K$ after performing a more careful analysis.

Under good events $\mathcal{E}_{1,\epsilon}(T) \cap \mathcal{E}_{2,\epsilon}(T)$, we show Theorem 5 the convergence of $P^\epsilon\text{-FWS}$ when $\hat{\mu}(t)$ is replaced with $\mu$. As shown in Appendix D.3.2, the extra error due to this replacement is controlled, thanks to Lemma 7

**Theorem 5.** Let $\epsilon \in (0, \min\{1, 2D^2\Delta_{\min}^{-2}/K\})$ and $T$ be an integer at least larger than

$$\max \left\{ \left( \frac{4|\chi_0|}{2} \right)^{\frac{1}{2}}, \left( \frac{2}{\Delta_{\min}} \right)^{\frac{1}{2}}, \left( \frac{3\|\mu\|_{\infty}}{2} \right)^{\frac{1}{2}} \right\} + \max \left\{ \left( \frac{\epsilon}{\|\mu\|_{\infty}} \right)^{\frac{1}{2}}, \left( \frac{5\|\mu\|_{\infty}^2}{\epsilon} \right)^{\frac{1}{2}} \right\}. $$

Under $\mathcal{E}_{1,\epsilon}(T) \cap \mathcal{E}_{2,\epsilon}(T)$, Algorithm 2 with (10) satisfies that: for any $t = T(n(T)), T(n(T) + 1, \ldots, T$,

(i) $\max_{\omega \in \Sigma} F_\mu(\omega) - F_\mu(\hat{\omega}(t)) \leq 5\epsilon$, (ii) $\Delta_{\min}(\hat{\mu}(t)) \geq \frac{\Delta_{\min}}{2}$, and (iii) $\|\hat{\mu}(t)\|_{\infty} \leq \frac{3\|\mu\|_{\infty}}{2}$.

The proof of Theorem 5 is given in Appendix E.

Finally, the last ingredient is Lemma 9

**Lemma 9 (Lemma 18 in [GK16]).** Let $\alpha \in \left[1, \frac{2}{3}\right]$ and $b_1, b_2 > 0$. Then,

$$x = \frac{1}{b_1} \left( \ln \frac{b_2}{b_1} + \ln \left( \frac{b_2}{b_1} \right) \right)$$

satisfies $b_1x \geq \ln(b_2x^\alpha)$.

### D.4 Computational complexity (Theorem 3 (iii))

In this section, we analyze the computational complexity of $P^\epsilon\text{-FWS}$ running with (10) in terms of the number of calls to $\text{LM Oracle}$. We will show that the expected number of $\text{LM Oracle}$ calls is upper bounded by a polynomial in $\ln \delta^{-1}$, $K$, $\|\mu\|_{\infty}$ and $\Delta_{\min}(\mu)^{-1}$.

**Proof** The construction of $\chi_0$ and computation of $\hat{t}$ merely takes $O(KD)$ calls to $\text{LM Oracle}$. The overall complexity is dominated by the $\text{LM Oracle}$ calls performed from $4|\chi_0| + 1$ to round $\tau$, analyzed as follows.

Per-round complexity: Fix $t \in \{4|\chi_0| + 1, \cdots, \tau\}$. Recall from $P^\epsilon\text{-FWS}$ that the FW update in round $\hat{t}$ and the stopping rule in round $t - 1$ are computed only if:

$$\max \{ \Delta_{\min}(\hat{\mu}(t - 1)^{-1}, \|\hat{\mu}(t - 1)\|_{\infty} \} \leq \sqrt{t - 1}. \quad (37)$$

Otherwise, forced-exploration procedure is invoked. Verifying (37) takes at most $D + 1$ calls. The computation of $F_{\mu(t-1)}$ and $\hat{t}^*(\nabla F_{\mu(t-1)}, n_t, (\hat{\omega}(t - 1)))$ by Theorem 3 in Appendix C.2 takes at most

$$O \left( \frac{D + \|\hat{\mu}(t - 1)\|_{\infty}^4 \|\hat{\omega}(t - 1)\|_{\infty}^2 K^3 D^3 \ln K}{F_{\mu(t-1)}(\hat{\omega}(t - 1))^2} \left( \frac{\ln(t^2\delta^{-1}) + n_t \ln \theta^{-1}}{\epsilon_t^2} \right) \right) \quad (38)$$

calls to $\text{LM Oracle}$. To evaluate (38), we need a lower bound on $F_{\mu(t-1)}(\hat{\omega}(t - 1))$. By Proposition 1 (c) in Appendix C.1 one evaluates $E_{\mu(t-1)}(\hat{\omega}(t - 1))$ in closed-form:

$$F_{\mu(t-1)}(\hat{\omega}(t - 1)) = \min_{x \neq \hat{t}^*(\hat{\mu}(t - 1))} \frac{\Delta_{\min}(\hat{\mu}(t - 1)^2}{4D(t - 1)} \geq \frac{\min_{k \in [K]} \hat{\omega}_k(t - 1)}{4D(t - 1)},$$

where the inequality results from (37) that $\Delta_{\min}(\hat{\mu}(t - 1)) \geq \frac{1}{\sqrt{t - 1}}$, and $\|x \pm x'\|_1 \leq 2D$ for any $x, x' \in \chi$, and $\langle y, z \rangle \leq \|y\|_1 \|z\|_{\infty}$ for any $y, z \in \mathbb{R}^K$. Further, combining with Lemma 14 which states $\min_{k \in [K]} \hat{\omega}_k(t - 1) \geq \frac{1}{2\sqrt{|\chi_0|} \delta}$ in Appendix F yields

$$F_{\mu(t-1)}(\hat{\omega}(t - 1)) \geq \frac{1}{8D \sqrt{|\chi_0|} (t - 1)^{1.5}}. \quad (39)$$

---

For any $\pi \in \Lambda$, $\Delta_{\min}(\pi)$ requires to compute $\hat{t}^*(\pi)$ and and solve $\max_{x \neq \hat{t}^*(\pi)} \langle \pi, x \rangle$, where the latter requires at most $D$ calls to the $\text{LM Oracle}$ by Lemma 2 in §2.2
From (39), \( \| \hat{\mu}(t - 1) \|_\infty \leq \sqrt{t - 1} \), Lemma [14], and substituting the parameters (10) into (38), we know that the number of \( \text{LM Oracle} \) calls performed at any round \( t \geq 4|\mathcal{X}_0| + 1 \) is at most

\[
\mathcal{O}\left( t^6|\mathcal{X}_0|^2K^3D^7\ln K \left( \frac{\ln(t^2\delta^{-1})}{\epsilon_t^2} + \frac{n_t \ln \theta_t^{-1}}{\rho_t^2} \right) \right) \\
= \mathcal{O}\left( t^6|\mathcal{X}_0|^2K^3D^7\ln K \left( \frac{t^2 \ln \left( \frac{t}{\delta} \right) + t^{2.75}D^4|\mathcal{X}_0|^2}{\delta} \right) \right) \\
= \mathcal{O}\left( t^{8.75} \ln \left( \frac{t}{\delta} \right) |\mathcal{X}_0|^4D^{11}K^3 \ln K \right).
\]

(40)

Overall complexity: Invoking Theorem 4 in D.3 with \( \hat{\epsilon} = 0.1 \) and \( \epsilon = \frac{1}{12T^\star(\mu)} \) results in

\[
E_{\mu}[\tau] = \mathcal{O}\left( T^\star(\mu) \ln \left( \frac{T^\star(\mu)}{\delta} \right) + \frac{1}{\Delta^2_{\min}} + K^26D^{40} \| \mu \|_\infty^{20} T^\star(\mu)^{10} \right)
\]

which after using \( T^\star(\mu) \leq 4KD/\Delta^2_{\min} \) (Lemma [1] in §2.1) becomes

\[
E_{\mu}[\tau] = \mathcal{O}\left( \frac{KD}{\Delta^2_{\min}} \ln \left( \frac{KD}{\delta\Delta^2_{\min}} \right) + \frac{K^26D^{40} \| \mu \|_\infty^{20}}{\Delta^2_{\min}} \right).
\]

(41)

Hence, by a summation of (41) over \( t = 4|\mathcal{X}_0| + 1 \) to \( E_{\mu}[\tau] \), the expected total number of the \( \text{LM Oracle} \) calls is upper bounded by

\[
\mathcal{O}\left( \mathbb{E}_{\mu}[\tau]^{9.75} \ln \left( \frac{\mathbb{E}_{\mu}[\tau]}{\delta} \right) |\mathcal{X}_0|^4D^{11}K^3 \ln K \right),
\]

(42)

where the inequality uses integral by parts for \( \int t^{8.75} \ln t dt = \mathcal{O}(t^{9.75} \ln t) \). Remind that \( \max\{D, |\mathcal{X}_0|\} \leq K \). Thus, we conclude that (42) is bounded by a polynomial function in \( \ln \delta^{-1}, \| \mu \|_\infty, \Delta^2_{\min}, \) and \( K \) (due to (41), \( E_{\mu}[\tau] \) is bounded by a polynomial function in the same variables).
E Convergence of $P$-FWS under the good events

Throughout this section, we assume that $\mu$ is fixed and drop $\mu$ from the notation, e.g., $F = F_\mu$. $\bar{F}_n, \tilde{F}_n, \tilde{F}_n = \tilde{F}_n, \tilde{F}_n, \Delta_{\min} = \Delta_{\min}(\mu)$. Also, we will use $\omega^* \in \arg\max_{\omega \in \Sigma} F(\omega)$ to denote any optimal allocation and let $i^* = i^*(\mu)$. Recall that $\bar{h}(T) \geq T^a$ and $\tilde{h}(T) \geq T^{a+b}$ is defined in (23) in Appendix D.3.1 for some $a, b \in (0, 1)$.

**Theorem.** Let $\epsilon \in (0, \min\{1, 2D^2\Delta^2_{\min}/K\})$ and $T$ be an integer at least larger than

$$\max\left\{\left(\frac{4K^2}{\epsilon^2 D^2 |X_0|}\right)^{\frac{1}{2}}, \left(\frac{2}{\Delta_{\min}}\right)^{\frac{1}{2}}, \left(\frac{3\|\mu\|_{\infty}}{2}\right)^{\frac{1}{2}}\right\} + \max\left\{\left(\frac{\ell}{\epsilon}\right)^{\frac{1}{2}}, \left(\frac{5\ell K^2}{\epsilon \sqrt{|X_0|}}\right)^{\frac{1}{2}}\right\}.$$

Under $\mathcal{E}_{1,\epsilon}(T) \cap \mathcal{E}_{2,\epsilon}(T)$, Algorithm 2 with (10) satisfies: for any $t = \bar{h}(T), \tilde{h}(T) + 1, \cdots, T$,

(i) $F(\omega^*) - F(\hat{\omega}(t)) \leq 5\epsilon$, (ii) $\Delta_{\min}(\mu(t)) \geq \frac{\Delta_{\min}}{2}$, and (iii) $\|\mu(t)\|_{\infty} \leq \frac{3\|\mu\|_{\infty}}{2}$.

**Proof.** Fix arbitrary $\epsilon$ and $T$ that satisfy the conditions in the statement, and suppose $\mathcal{E}_{1,\epsilon}(T) \cap \mathcal{E}_{2,\epsilon}(T)$ holds. (ii)(iii) directly follows from Lemma [10] (where one can verified that its assumption of Lemma [10] on $T$ is satisfied). With (ii)(iii), the analysis of FW convergence will be greatly simplified as (ii)(iii) ensure that

$$\max\left\{\frac{1}{\Delta_{\min}(\hat{\mu}(t-1))}, \|\mu(t-1)\|\right\} \leq \sqrt{t-1}.$$

This means that the forced-exploration procedure will only be invoked by the condition of $\sqrt{t/\|X_0\|}$ when $t \geq \hat{h}(T) = T^a$.

Proof of (i) $F(\omega^*) - F(\hat{\omega}(t)) \leq 5\epsilon$: Fix $t \geq \hat{h}(T)$. As mentioned above, for such $t$, the forced-exploration procedure will be invoked only when $\sqrt{t/\|X_0\|} \in \mathbb{N}$. To specify the rounds performing FW updates, introduce $s(t) = \lceil \sqrt{t/\|X_0\|} \rceil - 1$ and define

$$p(t) = (s(t)^2 + 1)|X_0| \quad \text{and} \quad q(t) = (s(t) + 1)^2|X_0| - 1. $$

Notice that $p(t)$ and $q(t)$ are respectively the starting (including) and the ending (including) round of a successive FW update rounds with no forced exploration in between. Let $\phi_t = F(\omega^*) - \tilde{F}_n(\hat{\omega}(t))$ be the error. By a careful analysis, we derive a recursive relationship satisfied by $\phi_t$ (Lemma [11]):

$$\begin{cases} t\phi_t \leq (t - |X_0|)\phi_{t-|X_0|} + 2\ell\sqrt{D}|X_0|^2 & \text{if } t = p(t) - 1, \\ \phi_t \leq (t - 1)\phi_{t-1} + 3\epsilon + \ell \left(\eta_{t-1} + K^2/2\theta t\right) & \text{if } t \in [p(t), q(t)]. \\ \end{cases}$$

(43)

The first case (in round $t = p(t) - 1$) is exactly the ending round of a forced-exploration procedure (from $t - |X_0|, \cdots, t$), and the second case (in round $t \in [p(t), q(t)]$) is a FW-update round. By repeatedly applying (43), we have

$$\bar{h}(T) \phi_{\bar{h}(T)} \leq \bar{h}(T) \phi_{\bar{h}(T)} + 2\ell \sqrt{D}|X_0|^2 \left(s(\bar{h}(T)) - s(\bar{h}(T))\right) + \sum_{t=\bar{h}(T)}^{\bar{h}(T)} \left(\frac{\ell K^2}{\sqrt{|X_0|}} + \epsilon\right)$$

$$\leq \bar{h}(T)\ell \left(2\sqrt{D}|X_0|^{1.5} + \frac{3K^2}{\sqrt{|X_0|}}\right) \left(\sqrt{\bar{h}(T) - \bar{h}(T)}\right) + 3\epsilon \left(\bar{h}(T) - \bar{h}(T)\right),$$

where the second inequality follows from $\phi_{\bar{h}(T)} \leq \max_{\omega \in \Sigma} F_{\mu}(\omega) \leq \ell$ (Lemma [22] in Appendix I, $s(\bar{h}(T)) - s(\bar{h}(T)) \leq \sqrt{\bar{h}(T) - \bar{h}(T)}$ and $\sum_{t=\bar{h}(T)}^{\bar{h}(T)} \frac{1}{\sqrt{t}} \leq \sqrt{\bar{h}(T) - \bar{h}(T)}$. Substituting $\bar{h}(T)$ and $\tilde{h}(T)$ from (23) and simplifying the terms, we get:

$$F(\omega^*) - F(\hat{\omega}(t)) \leq \phi_{\bar{h}(T)} \leq \ell T^{-b} + \frac{5\ell K^2}{\sqrt{|X_0|}} T^{-a+b} + 3\epsilon (1 - T^{-b}) \leq 5\epsilon$$

31
where the first inequality is due to $F(\hat{\omega}(t)) \geq \hat{F}_{\eta_1}(\hat{\omega}(t))$ by Proposition 2(i) in §4.1.

**Lemma 10.** Let $\epsilon \in (0, \min\{1, 2D^2 \Delta_{\text{min}}^2/K\})$ and $T$ be a positive integer s.t.

$$T \geq \max \left\{ \left( 4|\lambda_0| \right)^{\frac{1}{2}}, \left( \frac{4K^2}{\epsilon^2 D^2 |\lambda_0|} \right)^{\frac{1}{2}}, \left( \frac{3\|\mu\|_{\infty}}{2} \right)^{\frac{2}{3}} \right\} + \max \left\{ \left( \frac{\ell}{\epsilon} \right)^{\frac{1}{2}}, \left( \frac{5\ell K^2}{\epsilon \sqrt{|\lambda_0|}} \right)^{\frac{2}{3}} \right\},$$

where the first inequality is due to

when

$$T \geq \max \left\{ \left( 4|\lambda_0| \right)^{\frac{1}{2}}, \left( \frac{4K^2}{\epsilon^2 D^2 |\lambda_0|} \right)^{\frac{1}{2}}, \left( \frac{3\|\mu\|_{\infty}}{2} \right)^{\frac{2}{3}} \right\} + \max \left\{ \left( \frac{\ell}{\epsilon} \right)^{\frac{1}{2}}, \left( \frac{5\ell K^2}{\epsilon \sqrt{|\lambda_0|}} \right)^{\frac{2}{3}} \right\},$$

where the first inequality is due to $F(\hat{\omega}(t)) \geq \hat{F}_{\eta_1}(\hat{\omega}(t))$ by Proposition 2(i) in §4.1.

**Lemma 10.** Let $\epsilon \in (0, \min\{1, 2D^2 \Delta_{\text{min}}^2/K\})$ and $T$ be a positive integer s.t.

$$T \geq \max \left\{ \left( 4|\lambda_0| \right)^{\frac{1}{2}}, \left( \frac{4K^2}{\epsilon^2 D^2 |\lambda_0|} \right)^{\frac{1}{2}}, \left( \frac{3\|\mu\|_{\infty}}{2} \right)^{\frac{2}{3}} \right\} + \max \left\{ \left( \frac{\ell}{\epsilon} \right)^{\frac{1}{2}}, \left( \frac{5\ell K^2}{\epsilon \sqrt{|\lambda_0|}} \right)^{\frac{2}{3}} \right\}.$$

Suppose $E_{1,\epsilon}(T) \cap E_{2,\epsilon}(T)$ holds. Then, for any $t \geq h(T)$,

$$\Delta_{\text{min}}(\hat{\mu}(t)) \geq \frac{\Delta_{\text{min}}}{2} \quad \text{and} \quad \|\hat{\mu}(t)\|_{\infty} \leq \frac{3\|\mu\|_\infty}{2}.$$  \hspace{1cm} (45)

**Proof** Fix any $T$ satisfying (45) and suppose $E_{1,\epsilon}(T) \cap E_{2,\epsilon}(T)$ holds. Consider any $t \geq T^\star$ and hence $t \geq 4|\lambda_0|$. To show the first inequality of (45), from $E_{2,\epsilon}(T)$ and $\epsilon < 2D^2 \Delta_{\text{min}}^2/K$, we have

$$\Delta_{\text{min}}(\hat{\mu}(t - 1)) \geq \Delta_{\text{min}} - \frac{2\epsilon}{24D^3 \|\mu\|_{\infty}} \geq \Delta_{\text{min}} - \frac{\Delta_{\text{min}}^2}{12K \|\mu\|_{\infty}} > \frac{\Delta_{\text{min}}}{2},$$

where the last inequality is because $\Delta_{\text{min}} \leq 2D \|\mu\|_{\infty}$ and $D \leq K$. To show the second inequality of (45), observe that

$$\|\hat{\mu}(t)\|_{\infty} \leq \|\mu\|_{\infty} + \epsilon \frac{24D^3 \|\mu\|_{\infty}}{\epsilon^2 D^2 |\lambda_0|} < \|\mu\|_{\infty} + \frac{\Delta_{\text{min}}^2}{12KD \|\mu\|_{\infty}} < \frac{3\|\mu\|_\infty}{2},$$

where the first inequality is because of $E_{2,\epsilon}(T)$, the second is due to $\epsilon < 2D^2 \Delta_{\text{min}}^2/K$, and the last uses $\Delta_{\text{min}} \leq 2D \|\mu\|_{\infty}$. \hspace{1cm} □

**Lemma 11.** Let $\epsilon > 0$ and $t \in \mathbb{N}$ be such that (45) holds. Then, under the event $E_{1,\epsilon}(t) \cap E_{2,\epsilon}$,

$$\begin{cases}
t\phi_t \leq (t - |\lambda_0|)\phi_t - |\lambda_0| + 2\ell\sqrt{D} |\lambda_0|^2 & \text{if } t \neq p(t) - 1 \\
t\phi_t \leq (t - 1)\phi_{t - 1} + 3\ell + \ell \left( \eta_{t - 1} + \frac{K^2}{2\ell \eta_t} \right) & \text{if } t \in [p(t), q(t)]
\end{cases}$$

where $p(t) = (s(t)^2 + 1)|\lambda_0|$, $q(t) = (s(t) + 1)^2|\lambda_0| - 1$, and $s(t) = \sqrt{t}/|\lambda_0| - 1$.

**Proof** The first case basically follows from the Lipschitzness of $F$ (Appendix I), whereas the second relies on results on stochastic smoothing (Appendix II).

Case $t = p(t) - 1$: In this case, round $t$ is exactly the end (including) round of a forced-exploration procedure. By $\ell$-Lipschitzness of $F$ (Lemma 21 in Appendix I),

$$F(\hat{\omega}(t)) - F(\hat{\omega}(t - |\lambda_0|)) \geq -\ell \|\hat{\omega}(t) - \hat{\omega}(t - |\lambda_0|)\|_2 \geq -\frac{\ell \sqrt{D} |\lambda_0|^2}{t},$$

where the second inequality stems from $\hat{\omega}(t) = \ell \left( \frac{|\lambda_0|}{t} \right) \hat{\omega}(t - |\lambda_0|) + \frac{|\lambda_0|}{t} \sum_{x \in X_0} x$ after performing the forced exploration. It then follows that $\|\hat{\omega}(t) - \hat{\omega}(t - |\lambda_0|)\|_2 \geq \ell \sqrt{D} |\lambda_0|^2$. By $\max_{\omega \in \Omega} F(\omega) \leq \ell$ (Lemma 22 in Appendix I) and a rearrangement of the above yields

$$t\phi_t \leq t\phi_{t - |\lambda_0|} + \ell \sqrt{D} |\lambda_0|^2 \leq (t - |\lambda_0|)\phi_t - |\lambda_0| + \ell |\lambda_0|(\sqrt{D} |\lambda_0| + 1).$$

The proof is completed after simplifying the terms.

Case: $t \in [p(t), q(t)]$: In this case, round $t$ performs a FW update. For brevity, let $z = \hat{\omega}(t)$ and $y = \hat{\omega}(t - 1)$. By $\frac{K^2}{\eta_t}$-smoothness of $\hat{F}_{\eta_1}$ (Proposition 2(iii) in §4.1) and $z - y = \frac{1}{\ell}(x(t) - y)$,

$$\hat{F}_{\eta_1}(z) \geq (\ast) - \ell K \frac{\eta_t}{2\eta_t} \|z - y\|_2^2 \geq (\ast) - \frac{\ell K^2}{2\ell^2 \eta_t} \|x(t) - y\|_2^2 \geq (\ast) - \frac{\ell K^2}{2\ell^2 \eta_t}.$$
where \((\ast) = \bar{F}_{\eta_t}(y) + \langle \nabla \bar{F}_{\eta_t}(y), z - y \rangle = \bar{F}_{\eta_t}(y) + \frac{1}{t} \langle \nabla \bar{F}_{\eta_t}(y), x(t) - y \rangle\). It follows from \(\mathcal{E}_{1, \epsilon}^{(t)} \cap \mathcal{E}_{2, \epsilon}^{(t)}\) and the continuity argument (Lemma 7 in Appendix G.1) that

\[
\langle \nabla \bar{F}_{\eta_t}(y), x(t) - y \rangle \geq \langle \nabla \bar{F}_{\mu(t-1), \eta_t}(y), x(t) - y \rangle - \epsilon \\
\geq \max_{\omega \in \Sigma} \langle \nabla \bar{F}_{\mu(t-1), \eta_t}(y), \omega - y \rangle - 2 \epsilon \geq \max_{\omega \in \Sigma} \langle \nabla \bar{F}_{\eta_t}(y), \omega - y \rangle - 3 \epsilon.
\]

Then, the duality gap [Jag13] and the \(\ell\)-Lipschitzness of \(F\) (Lemma 21 in Appendix I) yield

\[
\max_{x \in X} \langle \nabla \bar{F}_{\eta_t}(y), \omega - y \rangle \geq \max_{\omega \in \Sigma} \bar{F}_{\eta_t}(\omega) - \bar{F}_{\eta_t}(y) \\
\geq (F(\omega^*) - \eta_t \ell) - (\bar{F}_{\eta_t-1}(y) + \ell(\eta_t - \eta_{t-1})) = \phi_{t-1} - \ell \eta_{t-1}.
\]

Therefore, \(\bar{F}_{\eta_t}(z) \geq \bar{F}_{\eta_t}(y) + \frac{\phi_{t-1} - \ell \eta_{t-1} - 3 \epsilon}{t - \ell K^2 2t^2 \eta_t} \) and subtracting \(F(\omega^*)\) on both sides,

\[
\phi_t = F(\omega^*) - \bar{F}_{\eta_t}(x) \\
\leq (F(\omega^*) - \bar{F}_{\eta_t}(y)) + \frac{\phi_{t-1} + \ell \eta_{t-1} + 3 \epsilon}{t} + \frac{\ell K^2}{2t^2 \eta_t} \\
= \frac{t - 1}{t} \phi_{t-1} + \frac{1}{t} \left(3 \epsilon + \ell \left(\eta_{t-1} + \frac{K^2}{2t \eta_t}\right)\right)
\]

which completes the proof. \(\Box\)
Upper bound of $\sum_{T=M+1}^{\infty} \Pr_{\mu}[(E_{1,\epsilon}(T) \cap E_{2,\epsilon}(T))^c]$ under $P$-FWS

Recall from (24) that $E_{1,\epsilon}(T) = \cap_{t=\hat{T}(T) + 1}^{T} E_{1,\epsilon}^{(t)}$ and $E_{2,\epsilon}(T) = \cap_{t=\hat{T}(T) + 1}^{T} E_{2,\epsilon}^{(t)}$, where

$$E_{1,\epsilon}^{(t)} = \left\{ \langle \nabla \hat{F}_{\mu}(t-1), n_t \rangle (\hat{\omega}(t-1)), x(t) \rangle \geq \max_{x \in X} \langle \nabla \hat{F}_{\mu}(t-1), n_t \rangle (\hat{\omega}(t-1)), x \rangle - \epsilon \right\},$$

$$E_{2,\epsilon}^{(t)} = \left\{ \| \hat{\mu}(t-1) - \mu \|_* < \frac{\epsilon}{24D^3 \| \mu \|_*} \right\},$$

$T \geq M$ and $M$ is defined in (25). Also, recall $x(t) \in \arg\max_{x \in X} \langle \nabla \hat{F}_{\mu}(t-1), n_t \rangle (\hat{\omega}(t-1)), x \rangle$, where $\nabla \hat{F}_{\mu}(t-1), n_t, \hat{\omega}(t-1))$ is computed by $(\rho_t, \theta_t)$-MCP algorithm with

$$\left( \eta_t, \eta_t, \rho_t, \theta_t \right) = \left( \frac{1}{4 \sqrt{t / |X_0|}}, \left\lfloor t^{\frac{1}{2}} \right\rfloor, \frac{1}{16t D^2 |X_0|}, \frac{1}{4 t e^{\sqrt{t}}} \right). \quad (10)$$

Our main result Lemma 8 is built by bounding $\Pr_{\mu}[E_{1,\epsilon}(T)]$ and $\Pr_{\mu}[E_{2,\epsilon}(T)]$ separately with Lemma 12 in F.1 and Lemma 14 in F.2.

**Lemma 8.** Let $\epsilon \in (0, 2D^2 \Delta_{\text{min}}^2 / K)$ and $M$ be defined as in (25) Then,

$$\sum_{T=M+1}^{\infty} \Pr_{\mu}[(E_{1,\epsilon}(T) \cap E_{2,\epsilon}(T))^c] \leq 2K \left( \frac{3}{A_1(\epsilon)^{2+\frac{2}{\alpha}}} + \frac{2}{A_2(\epsilon)^{2+\frac{2}{\alpha}}} \right) \Gamma \left( 2 + \frac{2}{\alpha} \right),$$

where $A_1(\epsilon) = \min \{ 1, \frac{\epsilon^2}{2 |X_0|^2 D^2} \}$, $A_2(\epsilon) = \frac{\epsilon^2}{2048 D^6 |X_0|^{2+\frac{2}{\alpha}}}$, and $\Gamma$ denotes the gamma function $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ for any $z > 0$.

**Proof.** Fix $\epsilon > 0$. For all $T \geq M$, we have

$$\Pr_{\mu}[(E_{1,\epsilon}(T) \cap E_{2,\epsilon}(T))^c] \leq \Pr_{\mu}[E_{1,\epsilon}(T)^c] + \Pr_{\mu}[E_{2,\epsilon}(T)^c].$$

Bounding $\Pr_{\mu}[E_{1,\epsilon}(T)^c]$: This is done by using Lemma 12 with $v = \hat{\omega}(t)$ and $(\eta_t, n_t) = (\eta_t, n_t)$. Before applying Lemma 12, we verify that our chosen $\rho_t$ in (10) satisfies the assumption of Lemma 12

$$\frac{(\min_{k \in [K]} \hat{\omega}_k(t) - \eta_t)}{D^2} \geq \frac{1}{D^2} \left( \frac{1}{2 \sqrt{t / |X_0|}} - \frac{1}{4 \sqrt{t / |X_0|}} \right)^2 = \frac{1}{16t D^2 |X_0|} = \rho_t,$$

where the inequality is because of $\min_{k \in [K]} \hat{\omega}_k(t) \geq \frac{1}{2 \sqrt{t / |X_0|}}$ (Lemma 14) and that $\eta_t = \frac{1}{4 t e^{\sqrt{t}}}$ in (10). Then, applying Lemma 12 with $v = \hat{\omega}(t), (v, \eta, n) = (\frac{1}{2 \sqrt{t / |X_0|}}, \frac{1}{4 \sqrt{t / |X_0|}}, \left\lfloor t^{\frac{1}{2}} \right\rfloor)$ yields that:

$$\Pr_{\mu}[E_{1,\epsilon}(T)^c] \leq \sum_{t=\hat{T}(T)}^{T} K \left( 2 \exp \left( -\frac{c^2 \sqrt{t}}{8t^2 K^2 D^2} \right) + \exp \left( -\sqrt{t} \right) \right) \leq \sum_{t=\hat{T}(T)}^{T} 3K \exp \left( -\sqrt{t} A_1(\epsilon) \right),$$

where $A_1(\epsilon) = \min \{ 1, \frac{\epsilon^2}{2 |X_0|^2 D^2} \}$.

Bounding $\Pr_{\mu}[E_{2,\epsilon}(T)^c]$: As Lemma 14 provides a lower bound on the number of pulls, $\min_{k \in [K]} N_k(t) \geq \frac{t}{2 \sqrt{t / |X_0|}}$, for all arms, using this lower bound of $N_k(t)$ as the number of i.i.d. samples in the application of Chernoff bound leads to:

$$\Pr_{\mu} \left[ |\hat{\mu}_k(t) - \mu| \geq \frac{\epsilon}{24D^3 \| \mu \|_*} \right] \leq 2 \exp \left( -\sqrt{t} A_2(\epsilon) \right).$$

Hence, $\Pr_{\mu}[E_{2,\epsilon}(T)^c] \leq 2K \sum_{t=\hat{T}(T)}^{T} \exp \left( -\sqrt{t} A_2(\epsilon) \right)$. Then, we have

$$\sum_{T=M+1}^{\infty} \Pr_{\mu}[(E_{1,\epsilon}(T) \cap E_{2,\epsilon}(T))^c] \leq 2K \int_{M+1}^{\infty} \int_{\hat{T}(T)}^{T} \left( 3Ke^{-\sqrt{t} A_1(\epsilon)} + 2Ke^{-\sqrt{t} A_2(\epsilon)} \right) dt dT \leq 2K \left( \frac{3}{A_1(\epsilon)^{2+\frac{2}{\alpha}}} + \frac{2}{A_2(\epsilon)^{2+\frac{2}{\alpha}}} \right) \Gamma \left( 2 + \frac{2}{\alpha} \right),$$

where the second inequality uses Lemma 15. \qed
F.1 Lemmas for bounding $\mathbb{P}_\mu [E_{1,\epsilon}(T)^r]$ 

The following lemma is a result of two concentration inequalities, one bounds how much the empirical average deviates from the expectation (Proposition 3), and the other bounds the error incurred by MCP (Lemma 13).

**Lemma 12.** Let $(\pi, \omega, \theta) \in \Lambda \times \Sigma_+ \times (0,1)$, $v \in (0, \min_{k \in [K]} \omega_k)$, and $\eta \in (0,v)$. Then, $\forall \epsilon \in (0,4K(v-\eta)/D)$,

$$
\mathbb{P} \left[ \mathbb{E}_{x \sim \mathcal{A}} \left[ \langle \nabla \tilde{F}_{\pi,\omega}(\omega), x^* - \hat{x} \rangle \right] \geq \max_{x \in \mathcal{A}} \langle \nabla \tilde{F}_{\pi,\omega}(\omega), x - \omega \rangle - \epsilon \right] \geq 1 - K \left( 2 \exp \left( - \frac{\epsilon^2 n^2}{8 \epsilon^2 K^3 D^2} \right) + n \theta \right),
$$

where $\nabla \tilde{F}_{\pi,\eta,n}(\omega)$ is computed by $(\frac{(v-\eta)^2}{D^2}, \theta) \cdot $MCP, and $\hat{x}^* \in \arg\max_{x \in \mathcal{A}} \langle \nabla \tilde{F}_{\pi,\eta,n}(\omega), x \rangle$.

**Proof.** Let $x^* \in \arg\max_{x \in \mathcal{A}} \langle \nabla \tilde{F}_{\pi,\omega}(\omega), x \rangle$. From $\hat{x}^* \in \arg\max_{x \in \mathcal{A}} \langle \nabla \tilde{F}_{\pi,\eta,n}(\omega), x \rangle$,

$$
\langle \nabla \tilde{F}_{\pi,\eta}(\omega), x^* - \hat{x}^* \rangle \leq \langle \nabla \tilde{F}_{\pi,\eta}(\omega), x^* - \hat{x}^* \rangle + \langle \nabla \tilde{F}_{\pi,\eta}(\omega), \hat{x}^* - x^* \rangle
$$

$$
= \langle \nabla \tilde{F}_{\pi,\eta}(\omega) - \nabla \tilde{F}_{\pi,\eta,n}(\omega), x^* - \hat{x}^* \rangle.
$$

Fix $\epsilon > 0$. Recall that $\nabla \tilde{F}_{\pi,\eta,n}(\omega) = \frac{1}{\eta} \sum_{m=1}^{\eta} \nabla \omega f_{x_m}(\omega + \eta \mathcal{Z}_m, \pi)$ where each $\hat{x}_m$ is computed by $(\frac{(v-\eta)^2}{D^2}, \eta) \cdot $MCP $(\omega + \eta \mathcal{Z}_m, \pi)$, and each $\mathcal{Z}_m$ is independently sampled from Uniform($B_2$). Now, consider any fixed $x = e_k$ for any $k \in [K]$. Invoking Proposition 3 with $\epsilon = \frac{\epsilon}{4K}$ and $x = e_k$, we get:

$$
\mathbb{P} \left[ \left| \left\langle \nabla \tilde{F}_{\pi,\eta}(\omega) - \frac{1}{\eta} \sum_{m=1}^{\eta} \nabla \omega f_{\pi}(\omega + \eta \mathcal{Z}_m, e_k), e_k \right\rangle \right| \geq \frac{\epsilon}{4K} \right] \leq 2 \exp \left( - \frac{\epsilon^2 n^2}{8 \epsilon^2 K^3 D^2} \right).
$$

Also, for $\nabla \tilde{F}_{\pi,\eta,n}(\omega)$ computed by the $(\frac{(v-\eta)^2}{D^2}, \theta) \cdot $MCP algorithm, Lemma 13 with $x = e_k$, and $\theta = \theta$ and the assumption that $\epsilon \in (0,4K(v-\eta)/D)$ implies that:

$$
\mathbb{P} \left[ \left| \left\langle \frac{1}{\eta} \sum_{m=1}^{\eta} \nabla \omega f_{\pi}(\omega + \eta \mathcal{Z}_m) - \nabla \tilde{F}_{\pi,\eta,n}(\omega), e_k \right\rangle \right| \geq \frac{\epsilon}{4K} \right] \leq n \theta.
$$

Combining the two inequalities leads to:

$$
\mathbb{P} \left[ \left| \left\langle \nabla \tilde{F}_{\pi,\eta}(\omega) - \nabla \tilde{F}_{\pi,\eta,n}(\omega), e_k \right\rangle \right| \leq \frac{\epsilon}{2K} \right] \geq 1 - \left( 2 \exp \left( - \frac{\epsilon^2 n^2}{8 \epsilon^2 K^3 D^2} \right) + n \theta \right).
$$

Then, an application of a union bound over all $\{e_k\}_{k \in [K]}$ gives

$$
\mathbb{P} \left[ \left| \left\langle \nabla \tilde{F}_{\pi,\eta}(\omega) - \nabla \tilde{F}_{\pi,\eta,n}(\omega), x^* - \hat{x}^* \right\rangle \right| \leq \epsilon \right] \geq 1 - K \left( 2 \exp \left( - \frac{\epsilon^2 n^2}{8 \epsilon^2 K^3 D^2} \right) + n \theta \right). \quad (46)
$$

Observe $\langle -\nabla \tilde{F}_{\pi,\eta,n}(\omega), x^* - \hat{x}^* \rangle \geq 0$ implies

$$
\left\{ \langle \nabla \tilde{F}_{\pi,\eta}(\omega) - \nabla \tilde{F}_{\pi,\eta,n}(\omega), x^* - \hat{x}^* \rangle \leq \epsilon \right\} \subseteq \left\{ \langle \nabla \tilde{F}_{\pi,\eta}(\omega), x^* - \hat{x}^* \rangle \leq \epsilon \right\}. \quad (47)
$$

From (46)-(47), we conclude that the r.h.s. of (47) happens with probability at least $1 - K \left( 2 \exp \left( - \frac{\epsilon^2 n^2}{8 \epsilon^2 K^3 D^2} \right) + n \theta \right)$. The proof is completed by simply rearranging the r.h.s. of (47).

\[ \square \]

**Lemma 13.** Let $(\pi, \omega, x, \theta) \in \Lambda \times \Sigma_+ \times \{0, 1\}^K \times \{0, 1\}$ with $\|x\|_1 \leq D$ and $v \in (0, \min_{k \in [K]} \omega_k)$.

$\forall (\eta, z) \in (0,v) \times B_2$, \hspace{1cm} $\mathbb{P} \left[ \left| \langle \nabla \omega f_{x}(\omega + \eta z) - \nabla \omega f_{\hat{x}}(\omega + \eta z), x \rangle \right| \leq \frac{v - \eta}{D} \right] \geq 1 - \theta$,

where $x$, is some action satisfying $f_{x}(\omega + \eta z) = F_{\pi}(\omega + \eta z)$, and $\hat{x}$ is the returned action of $(\frac{(v-\eta)^2}{D^2}, \theta) \cdot $MCP $(\omega + \eta z, \pi)$.
Proof. This basically follows from a direct calculation. Let \( \epsilon > 0 \) and fix any \( (\pi, \omega, x) \in \Lambda \times \Sigma_+ \times \{0,1\}^K \), \( \|x\|_1 \leq D \), and any \( (\eta, z) \in (0, v) \times B_2 \). Then, for \( \hat{x} \) computed by \((\rho, \theta)\)-MCP\((\omega + \eta z, \pi)\) with \( \rho = (v - \eta)^2 / D^2 \), we have with probability at least \( 1 - \theta \)
\[
\rho \geq |(\nabla \omega f_x, (\omega + \eta z) - \nabla \omega f_{\hat{x}}(\omega + \eta z), \omega + \eta z)| \\
\geq \min_{k \in [K]} (\omega + \eta z)_k \| \nabla \omega f_x, (\omega + \eta z) - \nabla \omega f_{\hat{x}}(\omega + \eta z) \|_\infty.
\]
Hence, remarking that \( \min_{k \in [K]} (\omega + \eta z)_k \geq v - \eta > 0 \), we get: with probability at least \( 1 - \theta \),
\[
| (\nabla \omega f_x, (\omega + \eta z) - \nabla \omega f_{\hat{x}}(\omega + \eta z), x) | \leq \frac{\rho D}{v - \eta} = \frac{v - \eta}{D},
\]
where we used the fact that \( \|x\|_1 \leq D \) and Hölder’s inequality. \( \square \)

Proposition 3. Let \((\pi, \omega, x) \in \Lambda \times \Sigma_+ \times \{0,1\}^K \), \( \eta \in (0, \min_{k \in [K]} \omega_k) \), and \( \|x\|_1 \leq D \). Then,
\[
\forall \epsilon > 0, \quad \mathbb{P} \left[ \left| \left\langle \nabla \bar{F}_{\pi, \eta}(\omega) - \frac{1}{n} \sum_{m=1}^n \nabla F_{\pi}(\omega + \eta z_m), x \right\rangle \right| \geq \epsilon \right] \leq 2 \exp \left( - \frac{2\epsilon^2 n^2}{\ell^2 K D^2} \right),
\]
where \( Z_1, \ldots, Z_n \) are independently sampled from Uniform\((B_2)\).

Proof. Fix \((\pi, \omega, x) \in \Lambda \times \Sigma_+ \times \{0,1\}^K \) where \( \|x\|_1 \leq D \), and fix \( \eta \in (0, \min_{k \in [K]} \omega_k) \). Define
\[
\phi(z_1, \ldots, z_t) = \left\langle \nabla \bar{F}_{\pi, \eta}(\omega) - \frac{1}{n} \sum_{m=1}^n \nabla F_{\pi}(\omega + \eta z_m), x \right\rangle.
\]
Note that \( \mathbb{E}_{z_1, \ldots, z_n} [\phi(Z_1, \ldots, Z_n)] = 0 \) by definition. Now we also observe that:
\[
\max_{z_1, \ldots, z_n, x \in B_2, m \in [n]} |\phi(z_1, \ldots, z_n) - \phi(z_1, \ldots, z_m, z_{m+1}, \ldots, z_n)| \leq \frac{\ell D}{n}
\]
due to the \( \ell \)-Lipschitzness of \( F_{\pi} \) (Lemma 21 in Appendix I) and \( \max_{x \in X} \|x\|_1 \leq D \). Hence it follows from McDiarmid’s inequality (Lemma 16 in F.3) that
\[
\forall \epsilon > 0, \quad \mathbb{P} [\|\phi(Z_1, \ldots, Z_n)\| \geq \epsilon] \leq 2 \exp \left( - \frac{2\epsilon^2 n^2}{K (\ell D)^2} \right) = 2 \exp \left( - \frac{2\epsilon^2 n^2}{\ell^2 K D^2} \right).
\]
\( \square \)

F.2 Lemmas for bounding \( \mathbb{P}_\mu [E_2, \epsilon(T)^c] \)

Lemma 14 (forced exploration). Let \( \mathcal{X}_0 \subseteq \mathcal{X} \) be any set covering all arms \([K]\) and \( t \geq 4 |\mathcal{X}_0| \). Any algorithm with forced-exploration procedure satisfies
\[
\hat{\omega}(t) \in \sqrt{\frac{1}{|\mathcal{X}_0|}} - \frac{1}{k} \leq \sqrt{\frac{1}{|\mathcal{X}_0|}} - \frac{1}{2} \sqrt{\frac{t}{|\mathcal{X}_0|}}, \quad \forall t \geq 4 |\mathcal{X}_0|.
\]

Proof. Fix any \( k \in [K] \). By merely counting the rounds before \( t \) performing forced exploration,
\[
N_k(t) \geq \sum_{s \in [t] : \sqrt{\frac{1}{|\mathcal{X}_0|}} - \frac{1}{k} \leq \sqrt{\frac{t}{|\mathcal{X}_0|} + \frac{1}{2} \sqrt{\frac{t}{|\mathcal{X}_0|}}} - \frac{1}{2} \sqrt{\frac{t}{|\mathcal{X}_0|}},
\]
where the last inequality holds for any \( t \geq 4 |\mathcal{X}_0| \). \( \square \)

F.3 Technical lemmas

Lemma 15 ([WTP21]). Let \( \alpha \in (0,1) \) and \( A, \beta > 0 \). Then,
\[
\int_0^\infty \left( \int_T^\infty e^{-At} \, dt \right) \, dT = \frac{\Gamma(\frac{1}{\alpha \beta} + \frac{1}{\beta})}{\beta A^{\frac{1}{\alpha}} + \frac{1}{\beta}}.
\]

36
The result of Lemma 5 [WTP21] is stated for $\alpha, \beta \in (0, 1)$ but it actually applies for the case of $\beta > 0$ as well. Here we provide a proof for completeness.

$$\int_0^{\infty} \left( \int_{-t}^{\infty} e^{-At} dt \right) dT = \int_0^{\infty} \alpha T^\alpha e^{-AT} dT = \frac{1}{\beta} \int_0^{\infty} x^{1/\beta - \frac{1}{2}} e^{-Ax} dx = \frac{\Gamma\left(\frac{1}{\alpha\beta} + \frac{1}{\beta}\right)}{\beta A^{1/\beta + 1/\beta}}.$$ 

□

The below Lemma 16 also known as bounded different inequality, can be found in many textbooks, e.g., Theorem 6.2 in [BLM13].

**Lemma 16** (McDiarmid’s inequality). Let $Z = (Z_1, \ldots, Z_n)$ be independent random variables, and $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$ be a measurable function. Suppose $\phi(z)$ changes by at most $c_i > 0$ under an arbitrary change of the $i$-th coordinate. Then,

$$\forall \epsilon > 0, \quad \mathbb{P}[|\phi(Z) - \mathbb{E}[\phi(Z)]| \geq \epsilon] \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$
G Continuity arguments

In this section, we establish the continuity of $F_\pi(\omega)$ and $\nabla F_{\pi,\eta}(\omega)$ in $\pi$ for any fixed $\omega \in \Sigma_+$, where $\nabla F_{\pi,\eta}(\omega)$ denotes the gradient $\nabla_\omega F_{\pi,\eta}(\omega)$ taken w.r.t. the input $\omega$. As the consequence of the continuity of $F_\pi$ and $\nabla F_{\pi,\eta}$ in $\pi$, we can show the point-wise convergence of $F_{\mu(t)} \to F_{\mu}$ and $\nabla F_{\mu(t),\eta} \to \nabla F_{\mu,\eta}$ given that $\mu(t) \to \mu$ almost surely.

Notation. Throughout this section, we define $\nabla F_{\pi,\eta}(\omega) = 0_K$ if $\eta \geq \min_{k \in [K]}\omega_k$ for any $(\pi, \omega) \in \Lambda \times \Sigma_+$. Moreover, for any $(\nu, \omega) \in \mathbb{R}^K \times \Sigma_+$, we will use $\nabla_\pi F_{\nu}(\omega)$ (resp. $\nabla_\pi \left( \frac{\partial F_{\omega}(\omega)}{\partial \omega_k} \right)$) to denote the gradient of the function $\pi \mapsto F_{\nu}(\omega)$ (resp. $\pi \mapsto \frac{\partial F_{\omega}(\omega)}{\partial \omega_k}$) evaluated at the point $\nu$.

The main result in this section, Lemma 7 is derived based on Lemma 17 in Appendix G.1 which asserts the continuity of the function $\psi_{\omega,x,\eta}(\pi) = \langle \nabla F_{\pi,\eta}(\omega), x - \omega \rangle$ on $\mathbb{R}^K$ and Proposition 4 in Appendix G.2 (which upper bounds the length of $\nabla f_\pi(\omega, \mu)$).

**Lemma 7** Let $\mu \in \Lambda$ and $\epsilon \in (0, 2D^2 \Delta_{\min}(\mu)^2)$. Then, any $\pi \in \mathbb{R}^K$ with $\|\pi - \mu\|_\infty < \frac{\epsilon}{24D^3\|\mu\|_\infty}$ satisfies the following:

$$|F_\mu(\omega) - F_\pi(\omega)| < \epsilon, \quad \forall \omega \in \Sigma_+$$

$$|\langle \nabla F_{\pi,\eta}(\omega), x - \omega \rangle| < \epsilon, \quad \forall (\omega, x) \in \Sigma_+ \times \mathcal{X}, \forall \eta \in (0, \min_{k \in [K]} \omega_k).$$

**Proof** Inspired by Lemma 14 in [WTP21], we prove this lemma using Proposition 4 and applying the mean-value theorem to $\psi_{\omega,x,\eta}$.

Fix $(\omega, \mu) \in \Sigma_+ \times \Lambda$, and let $i^* = i^*(\mu)$ and $\Delta_x = \triangle_x(\mu)$ for any $x \in \mathcal{X} \setminus \{i^*\}$. Fix $\epsilon \in (0, 2D^2 \Delta_{\min})$ and $\pi \in \mathbb{R}^K$ such that $\|\pi - \mu\|_\infty < \frac{\epsilon}{24D^3\|\mu\|_\infty}$. One may check that this $\pi$ satisfies the assumption of Proposition 4 as

$$\|\pi - \mu\|_\infty < \frac{\epsilon}{24D^3\|\mu\|_\infty} < \frac{2D^2 \Delta_{\min}^2}{24KD \|\mu\|_\infty} = \frac{\Delta_x^2}{12KD \|\mu\|_\infty} \leq \frac{\Delta_{\min}}{6K} < \frac{\Delta_{\min}}{\sqrt{2KD}},$$

where the second inequality stems from the choice of $\epsilon$ and the second last is because $\Delta_{\min} \leq 2D \|\mu\|_\infty$. In what follows, we will be applying the mean-value theorem to $\psi_{\omega,x,\eta}$ (whose continuity is stated in Lemma 17). For convenience, introduce the function $r(\beta) = (1 - \beta)\mu + \beta \pi$ for any $\beta \in (0, 1)$.

Proof of (35). For any $x \in \mathcal{X} \setminus \{i^*\}$, by the mean-value theorem, there exists a $\beta \in (0, 1)$ such that

$$|f_x(\omega, \pi) - f_x(\omega, \mu)| = |\langle \nabla_\pi f_x(\omega, r(\beta)), \pi - \mu \rangle|$$

$$= \left| \sum_{k \in [K]} \omega_k \left( \nabla_\pi \left( \frac{\partial f_x(\omega, r(\beta))}{\partial \omega_k} \right), \pi - \mu \right) \right|$$

$$\leq \sum_{k \in [K]} \omega_k \left\| \nabla_\pi \left( \frac{\partial f_x(\omega, r(\beta))}{\partial \omega_k} \right) \right\|_1 \|\pi - \mu\|_\infty < \epsilon,$$

where the last inequality uses $\omega \in \Sigma_+$, $\|\pi - \mu\|_\infty < \frac{\epsilon}{24D^3\|\mu\|_\infty}$, $\left\| \nabla_\pi \left( \frac{\partial f_x(\omega, r(\beta))}{\partial \omega_k} \right) \right\|_1 \leq 12D^2 \|\mu\|_\infty$ (Proposition 4). Hence, from a substitution of $\omega$ in (48) with $x \in \arg\min_{x \neq i^*} f_x(\omega, \mu)$ and the fact that $F_\pi(\omega) \leq f_{x_0}(\omega, \pi)$, we derive

$$F_\mu(\omega) - F_\pi(\omega) \leq f_{x_0}(\omega, \pi) - f_{x_0}(\omega, \mu) < \epsilon.$$  

The other inequality of $F_\mu(\omega) - F_\pi(\omega) < \epsilon$ can be derived similarly. This proves (35).

Proof of (36). Recall that $\psi_{\omega,x,\eta}(\pi) = \langle \nabla F_{\pi,\eta}(\omega), x - \omega \rangle$ is continuous on $\mathbb{R}^K$ (Lemma 17). By the mean-value theorem, there exists $\beta \in (0, 1)$ such that

$$|\psi_{\omega,x,\eta}(\pi) - \psi_{\omega,x,\eta}(\mu)| = |\langle \nabla_\pi \psi_{\omega,x,\eta}(r(\beta)), \pi - \mu \rangle|$$

$$\leq \left\| \nabla_\pi \psi_{\omega,x,\eta}(r(\beta)) \right\|_1 \|\pi - \mu\|_\infty.$$

(49)
To bound $\|\nabla \psi \cdot \psi(x, x, \mu(r(x)))\|_1$, we write

$$
\nabla \psi \cdot \psi(x, x, \mu(r(x))) = \sum_{k \in \mathcal{K}} \nabla \left( \frac{\partial \tilde{F}_{r(\beta), \eta}(\omega)}{\partial \omega_k} \right) (x_k - \omega_k).
$$

Then it follows from the fundamental theorem of calculus that: the gradient $\nabla$ and the expectation operators are exchangeable, i.e.,

$$
\forall k \in \mathcal{K}, \quad \nabla \left( \frac{\partial \tilde{F}_{r(\beta), \eta}(\omega)}{\partial \omega_k} \right) = \mathbb{E}_{Z \sim \text{Uniform}(B_2)} \left[ \nabla \left( \frac{\partial \tilde{F}_{r(\beta)}(\omega + \eta Z)}{\partial \omega_k} \right) \right].
$$

As shown in Appendix H, $\frac{\partial \tilde{F}_{r(\beta)}(\omega + \eta Z)}{\partial \omega_k}$ exists almost surely. When such gradient exists, Proposition 4 bounds its 1-norm length by

$$
\left\| \nabla \left( \frac{\partial \tilde{F}_{r(\beta)}(\omega + \eta Z)}{\partial \omega_k} \right) \right\|_1 \leq 12D^2 \| \mu \|_\infty,
$$

so it follows that $\left\| \nabla \left( \frac{\partial \tilde{F}_{r(\beta), \eta}(\omega)}{\partial \omega_k} \right) \right\|_1 \leq 12D^2 \| \mu \|_\infty$ as well. Hence, substituting the above back to $\nabla \psi \cdot \psi(x, x, \mu(r(x)))$ yields:

$$
\left\| \nabla \psi \cdot \psi(x, x, \mu(r(x))) \right\|_1 \leq \max_{k \in \mathcal{K}} \nabla \left( \frac{\partial \tilde{F}_{r(\beta), \eta}(\omega)}{\partial \omega_k} \right) \left\| x - \omega \right\|_1 \leq 24D^3 \| \mu \|_\infty,
$$

where the first inequality use Hölder’s inequality. Finally, plugging the above into (49) and recalling that $\| \pi - \mu \|_\infty < \frac{1}{24D^3 \| \mu \|_\infty}$, we have

$$
|\psi(x, x, \mu) - \psi(x, x, \mu)| < \epsilon.
$$

This concludes the proof. \(\square\)

### G.1 An application of the maximum theorem

Recall that $\psi(x, x, \eta) = \langle \nabla \tilde{F}_{\pi, \eta}(\omega), x - \omega \rangle$.

**Lemma 17.** For any $\epsilon > 0$, there exists a constant $\xi > 0$ such that if $\| \pi - \mu \|_\infty < \xi$, then

$$
|\psi(x, x, \eta) - \psi(x, x, \mu)| < \epsilon, \quad \forall (\omega, x) \in \Sigma_+ \times \mathcal{X}, \quad \forall \eta \in (0, \min_{k \in \mathcal{K}} \omega_k).
$$

The proof of Lemma 17 relies on the celebrated maximum theorem [FKV14], which is introduced below. After that, we then show its proof.

**Maximum Theorem:** Here we briefly introduce the maximum theorem and Lemma 17 will be proved at the end of this section. The definitions and results are taken from [FKV14] (see also Appendix K.1 of [WTP21]).

**Definition 1.** Let $U \neq \emptyset$ be a subset of a topological space and $h : U \mapsto \mathbb{R}$ be a function. Define the level sets of $h$ for $y \in \mathbb{R}$ as

$$
L_h(y, U) = \{ x \in U : h(x) \leq y \} \quad \text{and} \quad L_h^+(y, U) = \{ x \in U : h(x) < y \}.
$$

The function $h$ is said to be lower semi-continuous (resp. upper semi-continuous) on $U$ if $L_h(y, U)$ are closed (resp. $L_h^+(y, U)$ are open) for all $y \in \mathbb{R}$; $h$ is said to be inf-compact on $U$ if $L_h(y, U)$ and $L_h^+(y, U)$ are compact for all $y \in \mathbb{R}$.

**Definition 2.** Let $\mathcal{X}$ and $\mathcal{Y}$ be Hausdorff topological spaces and $\Phi : \mathcal{X} \rightrightarrows \mathcal{Y}$ be a set-valued function, where $\mathcal{S}(\mathcal{Y})$ is the set of non-empty subsets of $\mathcal{Y}$. Define

$$
Gr_U(\Phi) = \{(x, y) \in U \times \mathcal{Y} : y \in \Phi(x)\}
$$

as the graph of $\Phi$ restricted to $U$. The function $u : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ is said to be $K$-inf-compact on $Gr_U(\Phi)$ if for all non-empty compact subset $C$ of $\mathcal{X}$, $u$ is inf-compact on $Gr_U(\Phi)$.

**Theorem 6 (Maximum theorem).** Suppose $\mathcal{X}$ is compactly generated, $\Phi : \mathcal{X} \rightrightarrows \mathcal{S}(\mathcal{Y})$ is lower hemicontinuous, and $u : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ is $K$-inf-compact and upper semi-continuous on $Gr_U(\Phi)$. Then, the function $v(x) = \inf_{y \in \Phi(x)} u(x, y)$ is continuous and the set of its optimal solutions $\Phi^*(x) = \{ y \in \Phi(x) : u(x, y) = v(x) \}$ is upper hemicontinuous and compact-valued.
Proof of Lemma 17: Fix any \( \mu \in \Lambda \) and let \( i^* = i^*(\mu) \). The goal is to show that for any \( \epsilon > 0 \), there exists a constant \( \xi_\epsilon > 0 \) such that if \( \| \pi - \mu \|_\infty < \xi_\epsilon \), then

\[
|\psi_{\omega,x,\eta}(\pi) - \psi_{\omega,x,\eta}(\mu)| < \epsilon, \quad \forall (\omega, x) \in \Sigma_+ \times X, \forall \eta \in (0, \min_{k \in [K]} \omega_k),
\]

where \( \psi_{\omega,x,\eta}(\pi) = \langle \nabla F_{\pi,\eta}(\omega), x - \omega \rangle \). In what follows, we will use \( p \) to denote the probability distribution of \( \text{Uniform}(B_2) \). We will first show that \( \psi_{\omega,x,\eta} \) is continuous for each fixed \( (\omega, x, \eta) \in \Sigma_+ \times X \times (0,1) \), and then use Theorem 6 to show (50).

Continuity of \( \psi_{\omega,x,\eta} \): Fix \( (\omega, x, \eta) \in \Sigma_+ \times X \times (0,1) \). Let \( U_\eta = \{ z \in B_2 : |\partial F_\mu(\omega + \eta z)| > 1 \} \) which is a measure-zero set under \( \mu \) (Lemma 20 in Appendix H). For its complement set \( B_2 \setminus U_\eta \), we split \( B_2 \setminus U_\eta \) into possibly overlapping sets \( B_\eta(y) = \{ z \in B_2 \setminus U_\eta : \nabla F_{\pi,\eta}(\omega + \eta z) = \nabla \omega f_y(\omega + \eta z, \pi) \} \), and define \( \psi_{\omega,x,\eta}(y, \cdot) = \int_{z \in B_\eta(y)} \langle \nabla \omega f_y(\omega + \eta z, \cdot), x - \omega \rangle \, dp(z) \) on each of these sets \( B_\eta(y) \). Observe that for any \( \pi \in \mathbb{R}^K \), we have

\[
\psi_{\omega,x,\eta}(\pi) = \int_{z \in B_2 \setminus U_\eta} \langle \nabla F_{\pi,\eta}(\omega + \eta z), x - \omega \rangle \, dp(z) = \sum_{y \neq i^*} \psi_{\omega,x,\eta}(y, \pi).
\]

To show the continuity of \( \psi_{\omega,x,\eta}(\pi) \), it suffices to show that each \( \psi_{\omega,x,\eta}(y, \cdot) \) is continuous. Fix \( y \in X \setminus \{i^*\} \) and any sequence \( \{ \pi_n \}_{n=1}^\infty \) converging to \( \mu \). Then, for any \( \forall z \in B_2 \), we have

\[
\begin{align*}
&\text{(i)} \quad \langle \nabla \omega f_y(\omega + \eta z, \pi_n), x - \omega \rangle \leq \| \nabla \omega f_y(\omega + \eta z, \pi_n) \|_\infty \| x - \omega \| \leq 2D\ell, \\
&\text{(ii)} \quad \lim_{n \to \infty} \langle \nabla \omega f_y(\omega + \eta z, \pi_n), x - \omega \rangle = \langle \nabla \omega f_y(\omega + \eta z, \mu), x - \omega \rangle.
\end{align*}
\]

This is because

\[
\nabla \omega f_y(\omega + \eta z, \cdot) = \frac{(i^*-\eta z)^2(i^*+\eta z)\circ (\omega + \eta z) - 2(i^*+\eta z)\circ (\omega + \eta z)}{2(i^*+\eta z)}
\]

by Lemma 19 and Proposition 1 (Appendix C) is obviously continuous and that function composition preserves continuity.

From (i) and (ii), the dominated convergence theorem implies that

\[
\psi_{\omega,x,\eta}(y, \mu) = \lim_{n \to \infty} \int_{z \in B_\eta(y)} \langle \nabla \omega f_y(\omega + \eta z, \pi_n), x - \omega \rangle \, dp(z).
\]

This shows the continuity of \( \psi_{\omega,x,\eta}(y, \cdot) \) for each \( y \neq i^* \), and thus \( \psi_{\omega,x,\eta} \) is continuous.

Application of the maximum theorem (Theorem 6): For this part, we take the approach similar to Lemma 6 in [WP2]. Define

\[
\phi(\pi) = \min \left\{ -\| \psi_{\omega,x,\eta}(\pi) - \psi_{\omega,x,\eta}(\mu) \| : (\omega, x, \eta) \in \Sigma_+ \times X \times (0,1) \right\}.
\]

We prove the continuity of \( \phi \) on \( S = \mathbb{R}^K \setminus \text{cl} \left( \text{Alt}(\mu) \right) \) by invoking Theorem 6 with the following substitutions:

- \( \Sigma = S \), \( \Phi = \Sigma_+ \times X \times (0,1) \), \( u(\pi, \omega, x, \eta) = -\| \psi_{\omega,x,\eta}(\pi) - \psi_{\omega,x,\eta}(\mu) \| \).

Here we verify that the assumptions of Theorem 6 are satisfied. \( \Sigma \) is compactly generated as \( S \) is a metric space; \( \Phi \) is continuous as it is a constant map; \( u \) is continuous due to the continuity of \( \psi_{\omega,x,\eta} \).

G.2 The length of gradients

Throughout this subsection, we fix \( \mu \in \Lambda \) and denote \( i^* = i^*(\mu) \), \( \triangle_x = \triangle_x(\mu) \), and \( \triangle_{\min}(\mu) = \triangle_{\min} \) for short. Here we aim to present Proposition 4 in which (i) quantifies how close an estimate \( \pi \) of \( \mu \) should be such that \( i^*(\pi) = i^* \), and (ii) asserts the continuity of any component of \( \nabla \omega f_x(\omega, \pi) \) in \( \pi \), and that its gradient with respect to \( \pi \) is bounded.

Proposition 4. Any \( \pi \in \mathbb{R}^K \) such that \( \| \pi - \mu \|_\infty < \frac{\triangle_{\min}}{2K\ell} \) satisfies
(i) \( i^*(\pi) = i^* \).

(ii) \( \forall x \in \mathcal{X}\setminus\{i^*\} \) and all \( k \in [K] \), \( \frac{\partial f_k(\omega, \pi)}{\partial \omega_k} \) is continuous in \( \pi \) and
\[
\left\| \nabla_{\pi} \left( \frac{\partial f_k(\omega, \pi)}{\partial \omega_k} \right) \right\|_1 \leq 12D^2 \| \mu \|_{\infty}.
\]

**Proof**  
Proof of (i): Lemma 18 is equivalent to that: any \( \pi \in \mathbb{R}^K \) satisfying \( \| \mu - \pi \|_{\infty} < \frac{\Delta_{\min}}{\sqrt{2KD}} \) implies that \( \pi \notin \text{cl} (\text{Alt}(\mu)) \). As closure of finite union equals union of closures,
\[
\mathbb{R}^K \setminus \text{cl} (\text{Alt}(\mu)) = \mathbb{R}^K \setminus \left( \cup_{x \neq i} \text{cl} (\{ \lambda \in \mathbb{R}^K : \langle i^* - x, \lambda \rangle < 0 \}) \right)
\]

\[
= \mathbb{R}^K \setminus \left( \cup_{x \neq i} \{ \lambda \in \mathbb{R}^K : \langle i^* - x, \lambda \rangle \leq 0 \} \right)
\]

\[
= \{ \lambda \in \mathbb{R}^K : i^*(\lambda) = i^* \}.
\]

Thus, \( \pi \notin \text{cl} (\text{Alt}(\mu)) \) is equivalent to \( i^*(\pi) = i^* \). This concludes the proof of (i).

Proof of (ii): Fix any \( \pi \in \mathbb{R}^K \) satisfying \( \| \mu - \pi \|_{\infty} < \frac{\Delta_{\min}}{\sqrt{2KD}} \). By Lemma 19 and \( i^*(\pi) = i^* \),
\[
\forall k \in [K], \quad \frac{\partial f_k(\omega, \pi)}{\partial \omega_k} = \frac{\langle i^* - x, \pi \rangle^2 (x_k + i_k^*)}{2 (x + i^*, \omega^{-1})^2 \omega_k^2}.
\]

Fix \( k \in [K] \). Note that the function \( \pi \mapsto \frac{\partial f_k(\omega, \pi)}{\partial \omega_k} \) is continuous and differentiable since it consists of inner products, element-wise products, and since its denominator is always positive. For its derivative,
\[
\left\| \nabla_{\pi} \left( \frac{\partial f_k(\omega, \pi)}{\partial \omega_k} \right) \right\|_1 = \left\| \frac{\langle i^* - x, \pi \rangle (x_k + i_k^*)^2}{(x + i^*, \omega^{-1})^2 \omega_k^2} \right\|_1
\]

\[
\leq \| i^* - x \|_1 \| i^* - x, \pi \rangle \|_1 \leq 4D^2 \| \pi \|_{\infty} \leq 12D^2 \| \mu \|_{\infty},
\]
where the first inequality is because \( \langle x + i^*, \omega^{-1} \rangle \omega_k \geq 1 \) if \( (x_k + i_k^*) = 1 \); the second is because \( x_k + i_k^* \leq 1 \); the third uses \( \| i^* - x \|_1 \leq 2D \) and \( \| i^* - x, \pi \rangle \leq \| i^* - x \|_1 \| \pi \|_{\infty} \); the last uses the triangle inequality:
\[
\| \pi \|_{\infty} \leq \| \mu \|_{\infty} + \| \mu - \pi \|_{\infty} \leq \| \mu \|_{\infty} + \frac{\Delta_{\min}}{\sqrt{2KD}} \leq 3 \| \mu \|_{\infty},
\]
where the last inequality is due to an application of Hölder’s inequality to
\[
\Delta_{\min} \leq \min_{x \neq i} \| i^* - x \|_1 \| \mu \|_{\infty} \leq 2D \| \mu \|_{\infty}.
\]

**Lemma 18.** \( \inf_{\lambda \in \text{Alt}(\mu)} \| \mu - \lambda \|_{\infty} \geq \frac{\Delta_{\min}}{\sqrt{2KD}} \).

**Proof**  
We claim that
\[
\inf_{\lambda \in \Lambda : \langle \lambda, i^* - x \rangle < 0} \| \mu - \lambda \|_{2}^2 = \frac{\Delta_{\min}}{\| i^* + x \|_2^2}, \quad \forall x \neq i^*.
\]

(51)

Observe that the proof immediately follows from (51) because the facts that \( \| y \|_2 \leq \sqrt{K} \| y \|_{\infty} \) for any \( y \in \mathbb{R}^K \), \( \text{Alt}(\mu) = \cup_{x \neq i} \{ \lambda \in \Lambda : \langle \lambda, i^* - x \rangle < 0 \} \), and \( \| i^* + x \|_2 \leq 2D \).

Proof of (51): By solving the stationary conditions, i.e., \( \nabla_{\lambda} \mathcal{L}_x(x^*, \alpha^*) = 2(\mu - \lambda^*_x) + \alpha^*(i^* - x) = 0 \) and \( \nabla_{\alpha} \mathcal{L}_x(x^*, \alpha^*) = (\lambda^*_x, i^* - x) = 0 \), we find
\[
\lambda^*_x = \mu - \frac{\Delta_{\min} \odot (i^* - x)}{\| i^* + x \|_2},
\]

is a minimizer for \( \inf_{\lambda \in \Lambda : \langle \lambda, i^* - x \rangle < 0} \| \mu - \lambda \|_{2} \). (51) follows by plugging \( \lambda^*_x \) into \( \| \mu - \lambda \|_{2} \).  

Remind that \( \nabla_{\omega} f_x \) can be evaluated by the following Lemma 19.

41
Lemma 19 (Envelope theorem). Let $(\omega, \mu) \in \Sigma_+ \times \Lambda$ and $x \in X \setminus \{i^*\}$. Define $\lambda^*_{\omega, \mu}(x) = \arg\min_{\lambda \in \text{cl}(C_x)} \left\langle \omega, \frac{(\mu - \lambda)^2}{2} \right\rangle$. Then,

$$\nabla_{\omega} f_x(\omega, \mu) = \frac{(\mu - \lambda^*_{\omega, \mu}(x))^2}{2} = \frac{\Delta_\omega(\mu)^2(x \oplus i^*) \odot \omega^{-2}}{2 \langle x \oplus i^*, \omega^{-1} \rangle^2}.$$

Proof The first equality is an application of Lemma 6 and Proposition 1 of [WTP21] with $\mathcal{I} = X$, $\mathcal{J}_x = \{x\}$, $\Sigma = \Sigma_K$, $S_x = \{\lambda \in \Lambda : i^*(\lambda) = x\}$ (see Appendix K.2 and Appendix K.4 in [WTP21] for more details). The second equality substitutes $\lambda^*_{\omega, \mu}(x) = \mu + \frac{\Delta_\omega(\mu)(x \oplus i^*) \odot \omega^{-1}}{\langle x \oplus i^*, \omega^{-1} \rangle}$ by using (11)-(14). \qed
H  Stochastic smoothing

This section is devoted to present Proposition 2 and verify the assumptions required for applying Proposition 2 to our objective \( F_{\mu} \).

Stochastic smoothing \[\text{FKM05, DBW12}\] is a well-studied technique and has been widely applied to online convex non-smooth optimization \[\text{HKT12, H+16}\]. Proposition 2 is a restatement of existing results. In particular, Proposition 2(i), (ii) and (iii) directly follow from Lemma E.2 in \[\text{DBW12}\] with \( (\lambda_0, u) = (\ell, \eta) \), \( f = -\Phi \) and \( f_u = -\Phi_{\eta}(\cdot) \), and Proposition 2(iv) can be established by Jensen’s inequality as done in the proof of Theorem 2.1 \[\text{DBW12}\].

**Proposition 2** Assume that \( \Phi : \mathbb{R}^K_{\geq 0} \to \mathbb{R} \) is concave, \( \ell \)-Lipschitz, and differentiable almost everywhere. Let \( B_2 = \{ v \in \mathbb{R}^K : \| v \|_2 \leq 1 \} \). For any \( \omega \in \Sigma^+ \) and \( \eta \in (0, \min_{k \in [K]} \omega_k) \), define
\[
\tilde{\Phi}_{\eta}(\omega) = \mathbb{E}_{\mathbb{Z} \sim \text{Uniform}(B_2)}[\Phi(\omega + \eta \mathbb{Z})].
\]

Then, \( \tilde{\Phi}_{\eta}(\omega) \) satisfies that:

1. \( \tilde{\Phi}_{\eta}(\omega) \leq \Phi(\omega) \leq \tilde{\Phi}_{\eta}(\omega) + \eta f(\omega) \)
2. \( \nabla \tilde{\Phi}_{\mu, \eta}(\omega) = \mathbb{E}_{\mathbb{Z} \sim \text{Uniform}(B_2)}[\nabla \Phi(\omega + \eta \mathbb{Z})] \)
3. \( \tilde{\Phi}_{\eta} \) is \( \frac{\ell K}{\eta} \)-smooth
4. If \( \eta > \eta' > 0 \), then \( \tilde{\Phi}_{\eta'}(\omega) \geq \tilde{\Phi}_{\eta}(\omega) \)

Now, we validate assumptions of Proposition 2 on \( F_{\mu} \). The concavity of \( F_{\mu} \), which is shown by \[\text{WTP21}\], follows from the facts that each \( f_x(\cdot, \mu) \) is concave and that \( F_{\mu} \) is a minimum of these functions \( f_x(\cdot, \mu) \) over all possible \( x \). The Lipschitzness of \( F_{\mu} \) is shown in Lemma 21 in Appendix I. Hence, it remains to show the almost-everywhere differentiability of \( F_{\mu} \). To show that the set of non-differentiable points of \( F_{\mu} \), i.e.,
\[
\bigcup_{x, x' \in \mathcal{X} \setminus \{ \hat{i}^{*}(\mu) \}, x \neq x'} \{ z \in B_2 : f_x((\omega + \eta z, \mu) = f_{x'}((\omega + \eta z, \mu) \},
\]
is measure-zero under \( \text{Uniform}(B_2) \), it suffices to show the following lemma.

**Lemma 20.** Let \( \mu \in \Lambda \) and \( x_1, x_2 \) be distinct actions in \( \mathcal{X} \setminus \{ \hat{i}^{*}(\mu) \} \). Then under the probability measure of \( \text{Uniform}(B_2) \),
\[
\{ z \in B_2 : f_{x_1}(\omega + \eta z, \mu) = f_{x_2}(\omega + \eta z, \mu) \}
\]
is a measure-zero set.

**Proof** To simplify the notation, let \( \hat{i}^{*} = \hat{i}^{*}(\mu) \) and \( \Delta_x = \Delta_x(\mu) \). Thanks to the close-form expressions of \( f_{x_1} \) and \( f_{x_2} \), \( z \in B_2 \) such that \( f_{x_1}(\omega + \eta z, \mu) = f_{x_2}(\omega + \eta z, \mu) \) are the points satisfying that:
\[
\frac{\Delta_{x_{1}}^{2}}{2(x_1 \oplus i^{*}, (\omega + \eta z)^{-1})} = \frac{\Delta_{x_2}^{2}}{2(x_2 \oplus i^{*}, (\omega + \eta z)^{-1})}.
\]
In other words, the set of interests is
\[
\left\{ z \in B_2 : \sum_{k=1}^{K} a_k \prod_{k' \neq k} (\omega_{k'} + \eta z_{k'}) = 0 \right\}, \tag{52}
\]
where \( a_k = (x_2 \oplus i^{*})_k \Delta_{x_1}^{2} - (x_1 \oplus i^{*})_k \Delta_{x_2}^{2} \) for all \( k \in [K] \). We claim that \( \alpha \) is a non-zero vector. Otherwise, \( \alpha_k = 0, \forall k \in [K] \), which together with the fact that \( \Delta_{x_{1}}^{2}, \Delta_{x_2}^{2} > 0 \) directly imply \( (x_2 \oplus i^{*})_k = 0 \) if and only if \( (x_1 \oplus i^{*})_k = 0 \). That means \( x_1 = x_2 \), but this becomes a contradiction. Therefore, the set in (52) are the roots of a non-zero polynomial inside \( B_2 \), and hence it is a measure-zero set (see e.g. Lemma in [Oka73]). \( \square \)
I  Lipschitzness of $F_\mu$ and boundness of $F_\mu$ on $\Sigma_K \cap \mathbb{R}^K_{>0}$

In this section, we show the Lipschitzness of $F_\mu(v) = \min_{x \neq \mu^*} f_\mu(x, v, \mu)$ for $v \in \mathbb{R}^K_{>0}$. Let $x_\mu$ be an equilibrium action such that $F_\mu(v) = f_\mu(x_\mu, v, \mu)$. We will use the envelope theorem (Lemma 19 in Appendix G.2) to evaluate $\nabla f_\mu(x_\mu, v, \mu)$ in closed-form, and then bound its length. We will also derive an upper bound of $F_\mu(v)$ valid for any positive vector $v$ in the $(K-1)$-dimensional simplex $\Sigma_K$. In what below, we denote $i^* = i^*(\mu)$ and $\Delta_x = \Delta_x(\mu)$ for any $x \neq i^*$ for short.

**Lemma 21.** Let $\mu \in \Lambda$ and $\ell = 2D^2 \|\mu\|_\infty^2$. Then, $F_\mu$ is $\ell$-Lipschitz with respect to $\|\cdot\|_\infty$ on $\mathbb{R}^K_{>0}$.

**Proof.** Let $v \in \mathbb{R}^K_{>0}$. Recall that $F_\mu(v) = \min_{x \neq \mu^*} f_\mu(x, v, \mu)$, and each $f_\mu(x, v, \mu)$ is differentiable (proven in Lemma 19 in Appendix G.2). Hence if $x$ is the action such that $F_\mu(v) = f_\mu(x, v, \mu)$, the concavity of $F_\mu(v)$ and the fact that $\nabla f_\mu(x, v, \mu)$ is the subdifferential of $F_\mu$ on $v$ yield that

$$\forall v' \in \mathbb{R}^K_{>0}, |F_\mu(v) - F_\mu(v')| \leq \|\nabla f_\mu(x, v, \mu), v - v'\|_1 \leq \|\nabla f_\mu(x, v, \mu)\|_1 \|v - v'\|_\infty,$$

where the last inequality stems from Hölder’s inequality. From the above, the $\ell$-Lipschitz can be derived by upper bounding $\|\nabla f_\mu(x, v, \mu)\|_1$ by $\ell$. Now applying Lemma 19 in Appendix G.2 yields

$$\|\nabla f_\mu(x, v, \mu)\|_1 = \left\|\frac{(\mu - \mu^* x, \mu^*)^2}{2} \right\|_1 = \frac{\|v - (x \oplus i^*)\|_1^2}{2 \|x \oplus i^*, v^{-1}\|_1^2}. \quad (53)$$

To simplify the above, we observe that

$$\langle x \oplus i^*, v^{-1} \rangle^2 = \left(\sum_{k=1}^K v_k^{-1} 1\{x_k \neq i_k^*\}\right)^2 \geq \sum_{k=1}^K v_k^{-2} 1\{x_k \neq i_k^*\} = \|v^{-2} \circ (x \oplus i^*)\|_1. \quad (54)$$

where the inequality uses the fact that $v_k > 0$ for all $k \in [K]$. Also,

$$\Delta_x = \|i^* - x, \mu\|_1 \leq \|i^* - x\|_1 \|\mu\|_\infty \leq 2D \|\mu\|_\infty. \quad (55)$$

Thus, (53), (54), and (55) yields that $\|\nabla f_\mu(x, v, \mu)\|_1 \leq 2D^2 \|\mu\|_\infty^2$. \hfill \Box

**Lemma 22.** Let $\mu \in \Lambda$ and $\ell = 2D^2 \|\mu\|_\infty^2$. Then, $\max_{\omega \in \Sigma_K \cap \mathbb{R}^K_{>0}} F_\mu(\omega) \leq \ell$.

**Proof.** Observe that $f_\mu(x, v, \mu) = \langle \omega, \nabla f_\mu(x, v, \mu) \rangle$ for any $x \neq i^*$. Combining this observation with the fact that $\Delta_x \leq 2D \|\mu\|_\infty$ (as argued in (53) in proof of Lemma 21) implies:

$$F_\mu(v) = \min_{x \neq i^*} \frac{\Delta_x^2}{2 \langle x \oplus i^*, v^{-1} \rangle} \leq \frac{(2D \|\mu\|_\infty)^2}{2} = \ell,$$

where the first inequality is because $\langle x \oplus i^*, v^{-1} \rangle \geq \min_{k \in [K]} v_k^{-1} \geq 1$ (as $v \in \Sigma_K$ and $v_k > 0$ for all $k \in [K]$). The proof is completed since $v$ is taken arbitrarily. \hfill \Box
J Proofs related to combinatorial sets

Assumption 1 (i) There exists a polynomial-time algorithm identifying \( i^*(v) \) for any \( v \in \mathbb{R}^K \); (ii) \( \mathcal{X} \) is inclusion-wise maximal, i.e., there is no \( x, x' \in \mathcal{X} \) s.t. \( x < x' \); (iii) for each \( k \in [K] \), there exists \( x \in \mathcal{X} \) such that \( x_k = 1 \); (iv) \( |\mathcal{X}| \geq 2 \).

As claimed in §2.2, Assumption 1 holds for the following combinatorial sets:

- \( m \)-sets: \( \mathcal{X} = \{ x \in \{0, 1\}^K : \|x\|_1 = m \} \)
- spanning forests: \( \mathcal{X} \) is a set of all spanning forests in a given graph
- bipartite matchings: \( \mathcal{X} \) is a set of all maximal matchings in a given bipartite graph
- \( s \)-\( t \) paths: \( \mathcal{X} \) is the set of all source-destination paths in a directed acyclic graph

In what below, we present a simple proof for the above examples.

Proof Suppose (iii) (iv) hold (as we can always achieve (iii) by removing arms not covered by \( \mathcal{X} \) and (iv) holds for non-trivial sets). For (i), it is well-known that a polynomial-time \( \text{LM Oracle} \), i.e., \( i^*(\cdot) \), exists for each of the discussed combinatorial structures. For example, see Chapter 39 in [S03] for the greedy algorithm for matroids (applicable to \( m \)-set and spanning forests), Chapter 41 in [S03] for the augmentation-based algorithm for 2-matroid intersection (applicable to bipartite matchings), and algorithms such as Dijkstra’s algorithm for \( s \)-\( t \) paths.

It remains to verify (ii) the inclusion-wise maximal property of \( \mathcal{X} \). For \( \mathcal{X} \) as \( m \)-sets, the inclusion-wise maximal property clearly holds because any binary vector \( x' > x \) (resp. \( x' < x \)) for some \( x \in \mathcal{X} \) must have \( \sum_{k \in [K]} x'_k > m \) (resp. \( < m \)) and thus \( x' \notin \mathcal{X} \). The case is similar for \( \mathcal{X} \) as spanning forests since the number of edges of any spanning forests of a graph is the same. For \( \mathcal{X} \) as maximal matchings in which the term ‘maximal’ exactly refers to being inclusion-wise maximal, (ii) directly follows from the definition. For \( \mathcal{X} \) as the set of all source-destination paths in an acyclic graph, if there exists any source-destination path \( x' > x \) for some \( x \in \mathcal{X} \) then \( x' \) must contain a cycle, so inclusion-wise maximal property holds. \( \square \)

Lemma 2 Let \( v \in \mathbb{R}^K \) and \( x \in \mathcal{X} \). Under Assumption 1 there exists an algorithm that solves \( \max_{x' \in \mathcal{X}, x' \neq x} \langle v, x' \rangle \) by only making at most \( D \) queries to the \( \text{LM Oracle} \).

Proof Fix \( x \in \mathcal{X} \). Assume \( v \neq 0_K \) (as otherwise, any \( x' \neq x \) is a second-best action). Inspired by Lawler-Murty’s \( m \)-best algorithm [Law72], we will prove this lemma by considering the algorithm described as follows. It first computes \( i^*(v) \) by the \( \text{LM Oracle} \), and returns it as the output if \( i^*(v) = x \). Otherwise, we identify the second-best action by the program below:

\[
\max_{k \in [K] : x_k = 1} \langle v, i^*(v^{(k)}) \rangle, \quad \text{where} \quad v^{(k)}_i = \begin{cases} -3 \|v\|_1 & \text{if } i = k \\ v_i & \text{otherwise.} \end{cases} \tag{56}
\]

Intuitively, for each arm \( k \) of \( x \), the action \( i^*(v^{(k)}) \) represents the best one among all actions without \( k \) (we have a strong negative weight on the \( k \)-th component of \( v^{(k)} \)). In the following, we will show that at least one of \( \{i^*(v^{(k)}) : k \in [K], x_k = 1 \} \) is the second-best action.

More precisely, we will show that for any maximizer \( a \in [K] \) to (56), \( i^*(v^{(a)}) \) is a second-best action. Consider if \( (i^*(v^{(a)}))_a = 0 \), then the claim follows from the fact that \( i^*(v^{(a)}) \) is the best among all actions without \( a \) and also the best in \( \{i^*(v^{(k)}) : k \in [K], x_k = 1 \} \). It suffices to show that \( (i^*(v^{(a)}))_a = 1 \) cannot happen. If \( (i^*(v^{(a)}))_a = 1 \), then it follows from Assumption 1 (iv) \( |\mathcal{X}| \geq 2 \) and (ii) the inclusion-wise maximality of \( \mathcal{X} \) that there is another action \( x' \) such that \( x'_k = 0 \) but \( x_k = 1 \) for some \( k \in [K] \). So, by \( i^*(v^{(a)})_a = 1, v \neq 0_K \) and the definition of \( v^{(a)} \), we get

\[
\langle v, i^*(v^{(a)}) \rangle = \sum_{j \in [K] : i^*(v^{(a)})} v_j - 3 \|v\|_1 \leq -2 \|v\|_1 < \langle v, x' \rangle \leq \langle v, i^*(v^{(k)}) \rangle,
\]

which contradicts the optimality of \( a \) (as it would imply that \( i^*(v^{(k)}) \) is better).
Finally, as $\|x\|_1 \leq D$, the number of LM Oracle calls required for solving (56) is at most $D$. □

Finally, we present the property of $X_0$ briefly argued in § 4.2.

**Lemma 23.** Let $e_k$ is the $k$-th column of an identity matrix. Under Assumption $\square$, $X_0$ is a $[K]$-covering set and $|X_0| \geq 2$.

**Proof** Showing that $X_0$ covers $[K]$: Assumption $\square$(iii) ensures $\{x \in X : x_k = 1\} \neq \emptyset$, and it follows that $\max_{x \in X} \langle x, e_k \rangle = 1$, i.e., $(i^*(e_k))_k = 1$. As $(i^*(e_k))_k = 1$ holds for all $k$, the proof is completed.

Showing that $|X_0| \geq 2$: Suppose on the contrary, $|X_0| = 1$. Thanks to Assumption $\square$(iv) $|X| \geq 2$, there exists $x \in X$ such that $x_k = \langle e_k, x \rangle \geq \langle e_k, x' \rangle = x'_k$ for all $k \in [K]$, $x' \neq x$. Together with Assumption $\square$(iii), one easily deduce that $x_k = 1$ for all $k \in [K]$. However, this implies $x' < x$ for any $x' \neq x$ and hence contradicts to Assumption $\square$(ii) that $X$ is inclusion-wise maximal. □
K Sample complexity lower bound

In this section, we assume \( \mu \in \Lambda \) and \( \delta \in (0, 1) \) is fixed, and show Theorem 7 by adapting Lemma 19 in [KCG16].

Lemma 24 ([KCG16]). Any \( \delta \)-PAC algorithm satisfies

\[
\forall \lambda \in \text{Alt}(\mu), \quad \sum_{k \in [K]} \sum_{x \in \mathcal{X} : x_k = 1} \mathbb{E}_\mu[N_x(\tau)] \frac{(\mu_k - \lambda_k)^2}{2} \geq \text{kl}(\delta, 1 - \delta).
\]

(57)

Theorem 7. Any \( \delta \)-PAC strategy satisfies

\[
\mathbb{E}_\mu[\tau] \geq T^*(\mu)\text{kl}(\delta, 1 - \delta) \quad \text{with} \quad T^*(\mu)^{-1} = \sup_{\omega \in \Sigma} \inf_{\lambda \in \text{Alt}(\mu)} \left\langle \omega, \frac{\mu - \lambda}{2} \right\rangle,
\]

where \( \Sigma = \{ \sum_{x \in \mathcal{X}} w_x : w \in \Sigma_{|\mathcal{X}|} \} \) and \( \text{Alt}(\mu) = \{ \lambda \in \Lambda : i^*(\lambda) \neq i^*(\mu) \} \).

Proof We have: under any algorithm,

\[
\sup_{\omega \in \Sigma} \inf_{\lambda \in \text{Alt}(\mu)} \sum_{k \in [K]} \omega_k \frac{(\mu_k - \lambda_k)^2}{2} \geq \inf_{\lambda \in \text{Alt}(\mu)} \sum_{k \in [K]} \sum_{x \in \mathcal{X} : x_k = 1} \mathbb{E}_\mu[N_x(\tau)] \frac{(\mu_k - \lambda_k)^2}{2},
\]

Hence if the algorithm is \( \delta \)-PAC, by Lemma 24

\[
\mathbb{E}_\mu[\tau] \sup_{\omega \in \Sigma} \inf_{\lambda \in \text{Alt}(\mu)} \sum_{k \in [K]} \omega_k \frac{(\mu_k - \lambda_k)^2}{2} \geq \inf_{\lambda \in \text{Alt}(\mu)} \sum_{k \in [K]} \sum_{x \in \mathcal{X} : x_k = 1} \mathbb{E}_\mu[N_x(\tau)] \frac{(\mu_k - \lambda_k)^2}{2}
\]

\[
\geq \text{kl}(\delta, 1 - \delta).
\]

\[\square\]

Lemma 1. For any \( \mu \in \Lambda \), \( T^*(\mu) \leq 4KD\Delta_{\text{min}}(\mu)^{-2} \).

Proof Take \( \omega_0 = \sum_{x \in \mathcal{X}_0} x / |\mathcal{X}_0| \in \Sigma \), where \( \mathcal{X}_0 = \{ i^*(e_k) : k \in [K] \} \). Observe that \( \omega_0 \geq 1/K \) by Lemma 23 (which leads to \( \sum_{x \in \mathcal{X}_0} x \geq 1/K \) and \( 1/|\mathcal{X}_0| \geq 1/K \)). Thus,

\[
F_\mu(\omega_0) = \min_{x \neq i^*(\mu)} \frac{\Delta_x(\mu)^2}{2} \frac{\text{kl}(\delta, 1 - \delta)}{2} \geq \frac{\Delta_{\text{min}}(\mu)^2}{4KD},
\]

where we used Proposition 1 in §3.1 to obtain the equality, and the last inequality is because

\[
\langle x \oplus i^*(\mu), \omega_0^{-1} \rangle \leq \|
\[
\frac{2D}{\min_{k \in [K]} (\omega_0)_k} \leq 2KD.
\]

As \( T^*(\mu)^{-1} = \max_{\omega \in \Sigma} F_\mu(\omega) \geq F_\mu(\omega_0) \), we then have \( T^*(\mu) \leq \frac{4KD}{\Delta_{\text{min}}(\mu)^2} \).

\[\square\]
I  Extension to the transductive setting

In this section, we extend our results to the transductive combinatorial semi-bandits. In transductive best-arm identification with fixed confidence with semi-bandit feedback [JMKK21], the decision maker is given an exploration set \( \mathcal{A} \subseteq \{0,1\}^K \) and a decision set \( \mathcal{X} \subseteq \{0,1\}^K \) (\( \mathcal{A} \) might differ from \( \mathcal{X} \)), and at each round, she selects an action in \( \mathcal{A} \) to receive a semi-bandit feedback. Her goal is to identify the best action in \( \mathcal{X} \) using as few samples as possible.

**Notation.** Let \( \mathcal{M} \subseteq \{0,1\}^K \) be any set of actions. We use \( \mathbf{i}_{\mathcal{M}}^* (\mu) \) to denote any maximizer in \( \mathcal{M} \) of the linear maximization \( \max_{x \in \mathcal{M}} \langle x, \mu \rangle \). We also use \( \Sigma_\mathcal{M} = \{ \sum_{x \in \mathcal{M}} w_x : w \in \Sigma_{|\mathcal{M}|} \} \).

**Sample complexity lower bound.** The generalization of Theorem 7 to the transductive setting has been made in [JMKK21]: any \( \delta \)-PAC algorithm satisfies

\[
E_{\mu}[\tau] \geq T^* (\mu) k d(\delta, 1 - \delta) \quad \text{with} \quad T^* (\mu) = \frac{1}{2} \sup_{\omega \in \Sigma_{\mathcal{A}}} \inf_{\lambda \in \Lambda(\mu)} \left( \omega, \frac{(\mu - \lambda)^2}{2} \right).
\]

The inner optimization is still with respect to \( \mathcal{X} \) while the outer optimization is with respect to the exploration set \( \mathcal{A} \). Refer to Appendix C in [JMKK21] for the proof.

**Transductive \( P-FWS \) algorithm.** Assumption [7] has to be extended. It now needs to ensure that \( \mathbf{i}_{\mathcal{A}}^* (v) \) for any \( v \in \mathbb{R}^K \) can be computed in polynomial-time. The \( P-FWS \) algorithm also needs to be adapted to the transductive setting. This is done by the following two modifications:

- \([K]-\text{covering set}: \mathcal{X}_0 \leftarrow \{ \mathbf{i}_{\mathcal{A}}^* (e_k) : k \in [K] \} \)
- \( \text{FW update}: \mathbf{x}(t) \leftarrow \mathbf{i}_{\mathcal{A}}^* \left( \nabla F_{\hat{\mu}(t-1), n_t, n_t (\hat{\omega}(t-1))} \right) \)

**Analysis of \( P-FWS \).** Let \( D_\mathcal{A} = \max_{x \in \mathcal{A}} \| x \|_1 \). The analysis is easily extended by replacing \((D, \mathcal{X})\) with \((D_\mathcal{A}, \mathcal{A})\) in Appendix D, Appendix E, Appendix F and Appendix G whenever the context is subject to the exploration set rather than the decision set.