Pure Exploration via Frank–Wolfe Self-Play

Xinyu Liu
Department of Industrial Engineering and Decision Analytics
The Hong Kong University of Science and Technology
xliufn@connect.ust.hk

Chao Qin
Stanford Graduate School of Business
Stanford University
chaoqin@stanford.edu

Wei You
Department of Industrial Engineering and Decision Analytics
The Hong Kong University of Science and Technology
weiyou@ust.hk

Extended Abstract. A central focus of modern science is effectively and efficiently identifying the correct hypothesis from a finite set of alternatives. Canonical examples include: (i) identifying the single best alternative (best-arm identification, BAI; Audibert et al. 2010); (ii) identifying all alternatives that exceed a given threshold (thresholding bandits; Lai and Liao 2012); and (iii) identifying all Pareto-optimal alternatives under multiple objectives (Pareto set identification; Kone et al. 2023). Although rooted in Chernoff (1959), recent years have witnessed a surge of interest in characterizing fundamental performance limits and in developing algorithms that attain them.

Interestingly, these asymptotic analyses often reduce to a maximin optimization that admits a two-player zero-sum game interpretation between an experimenter and a skeptic; the game's equilibrium characterizes both the fundamental limits and the structure of optimal algorithms. The experimenter aims to answer a question of interest about the true state of nature denoted by $\boldsymbol{\theta} \in \mathbb{R}^d$. To achieve their goal, the experimenter gather convincing evidence by selecting a probability vector $\boldsymbol{p} \in \Delta_K$ that specifies the allocation of sampling effort across the K alternatives, where $\Delta_K \subset \mathbb{R}^K$ denotes the $(K-1)$-dimensional probability simplex. The skeptic, in turn, chooses an alternative state $\boldsymbol{\vartheta} \in \mathrm{Alt}(\boldsymbol{\theta})$, represents a instance with a different correct answer, in an attempt to mislead the experimenter's evidence collection. Given a pair of strategies $(\boldsymbol{p}, \boldsymbol{\vartheta})$, the experimenter's payoff is $\Gamma(\boldsymbol{p}, \boldsymbol{\vartheta}) \triangleq \sum_{i\in[K]} p_i \mathrm{KL}(P_{i,\boldsymbol{\theta}} \| P_{i,\boldsymbol{\vartheta}})$, where $P_{i,\boldsymbol{\theta}}$ and $P_{i,\boldsymbol{\vartheta}}$ denote the observation distributions of arm i under $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$, respectively, and $\mathrm{KL}(\cdot\|\cdot)$ is the KL divergence between two distributions. This quantity measures the expected amount of discriminative information against the skeptic's alternative $\boldsymbol{\vartheta}$ weighted by the experimenter's allocation $\boldsymbol{p}$. Because the game is zero-sum, the experimenter seeks to maximize $\Gamma(\boldsymbol{p}, \boldsymbol{\vartheta})$ while the skeptic seeks to minimize it, yielding the maximin formulation that underpins attainable performance limits for many pure-exploration problems:

$$\max_{\boldsymbol{p}\in\Delta_K} \inf_{\boldsymbol{\vartheta}\in\mathrm{Alt}(\boldsymbol{\theta})} \Gamma(\boldsymbol{p}, \boldsymbol{\vartheta}).$$

As in Wang et al. (2021), the skeptic's decision space $\mathrm{Alt}(\boldsymbol{\theta})$ can be expressed as as the union of a finite set of $\mathrm{Alt}_x(\boldsymbol{\theta})$, where each x represents a confusing scenario that can mislead the identification of the true correct answer. The set $\mathrm{Alt}_x(\boldsymbol{\theta})$ then consists of all instances in which this confusing scenario occurs. For instance, in BAI, a confusing scenario may be one in which a suboptimal arm x appears to outperform the best arm $I^*(\boldsymbol{\theta})$ under an alternative instance. The maximin problem can

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

then be reformulated as

$$\max_{\boldsymbol{p}\in\Delta_K} \min_{x\in\mathcal{X}} D(\boldsymbol{p}, x), \quad \text{where} \quad D(\boldsymbol{p}, x) \triangleq \inf_{\boldsymbol{\vartheta}\in \mathrm{Alt}_x(\boldsymbol{\theta})} \Gamma(\boldsymbol{p}, \boldsymbol{\vartheta}), \tag{1}$$

where $D(\boldsymbol{p}, x)$ quantifies the discriminative information the experimenter can obtain against the *most challenging* alternative in $\mathrm{Alt}_x(\boldsymbol{\theta})$. Although asymptotic in nature, this maximin formulation both motivates and guides the design of practical algorithms. For example, in BAI, a confusing scenario arises when a suboptimal arm $x \neq I^*(\boldsymbol{\theta})$ appears superior to the true best arm $I^*(\boldsymbol{\theta})$. The well-known (top-two) Thompson sampling allocates measurement effort so as to gather equal evidence to rule out *each* such scenario, thereby enabling it to quickly identify the best arm (Russo 2020). This corresponds to the structural requirement of the optimal allocation, known as the *(discriminative) information-balance* condition: in (1), $D(\boldsymbol{p}, x) = D(\boldsymbol{p}, x')$ for all $x, x' \neq I^*(\boldsymbol{\theta})$ (Russo 2020, Garivier and Kaufmann 2016). This indicates that the skeptic is indifferent to which of the most challenging alternatives—each corresponding to a scenario—is chosen.

However, this information balance property can fail to hold for structured bandits (such as linear bandits), suggesting that a naive modification of Thompson sampling may be inadequate. In fact, adapting optimism-based bandit algorithms (such as Thompson sampling) to such problems can be fundamentally inefficient, as they inherently prioritize exploring high-performing arms over efficiently uncovering the most informative ones. For illustration, we presents a simple linear-bandit instance, where the optimal allocation assigns zero sampling to the highest-performing arm, even though the objective is to identify it. Thompson Sampling and its variants nonetheless allocate substantial effort to this arm, thereby incurring markedly suboptimal performance. This example highlights the cost of ignoring the problem's information structure—an effect also observed in regret minimization; see Lattimore and Szepesvari (2017).

An optimization expert might ask: if the maximin formulation characterizes performance limits, why not treat it as a standard optimization problem and derive pure-exploration algorithms by solving it directly? A natural first approach is to view (1) as a maximization over $\boldsymbol{p}$ alone—maximize $\min_{x\in\mathcal{X}} D(\boldsymbol{p}, x)$ with respect to $\boldsymbol{p}$—and apply the Frank–Wolfe algorithm. When the feasible set for the allocation vector is the simplex, the linear minimization oracle returns a vertex (a one-hot vector), which aligns naturally with the one-arm-at-a-time sampling paradigm of bandit algorithms. However, even for unstructured problems such as BAI, numerical evidence indicates that Frank–Wolfe fails to attain asymptotic optimality (Ménard 2019, Degenne et al. 2020, Wang et al. 2021). A key reason is *nonsmoothness*: each $D(\boldsymbol{p}, x)$ in (1) is concave, so the objective is concave but typically nondifferentiable in $\boldsymbol{p}$, violating the bounded-curvature assumption for classical Frank–Wolfe guarantees (Jaggi 2013).

The nonsmoothness that breaks vanilla Frank–Wolfe comes from the skeptic's exact best responses—given $\boldsymbol{p}$, the skeptic solves $\min_{x\in\mathcal{X}} D(\boldsymbol{p}, x)$. But what if they move simultaneously? A natural question one may ask is whether the original maximin formulation in (1) admits a Nash equilibrium. Its existence is crucial: it anchors simultaneous-play algorithms and provides primal–dual optimality certificates. Interestingly, Qin and Russo (2024) show that even in the case of BAI, a strict minimax inequality holds. To guarantee the existence of a Nash equilibrium, we allow the skeptic to play a mixed strategy $\boldsymbol{\mu} \in \Delta_{|\mathcal{X}|}$ over scenarios $x \in \mathcal{X}$. Under mixed strategy $(\boldsymbol{p}, \boldsymbol{\mu})$, the payoff function is

$$F(\boldsymbol{p}, \boldsymbol{\mu}) \triangleq \mathbb{E}_{x\sim\boldsymbol{\mu}}[D(\boldsymbol{p}, x)]. \tag{2}$$

Since the simplices are compact and convex and F is continuous and concave–convex, Sion's minimax theorem implies existence of Nash equilibrium and

$$\max_{\boldsymbol{p}\in\Delta_K} \min_{\boldsymbol{\mu}\in\Delta_{|\mathcal{X}|}} F(\boldsymbol{p}, \boldsymbol{\mu}) = \min_{\boldsymbol{\mu}\in\Delta_{|\mathcal{X}|}} \max_{\boldsymbol{p}\in\Delta_K} F(\boldsymbol{p}, \boldsymbol{\mu}). \tag{3}$$

A perhaps surprising consequence of the mixed-strategy view is:

Simple Frank–Wolfe updates on both sides drive the players to a Nash equilibrium.

For payoff $F(\boldsymbol{p}, \boldsymbol{\mu})$ with simplex constraints, each Frank–Wolfe subproblem admits a closed form. In particular, both players react greedily to the linearized objective at $(\boldsymbol{p}, \boldsymbol{\mu})$:

$$\begin{cases} \boldsymbol{\mu} \leftarrow \frac{1}{n+1}(n\boldsymbol{\mu} + \boldsymbol{e}_x), & x \in \mathrm{argmin}_{x'\in\mathcal{X}} \left[\nabla_{\boldsymbol{\mu}} F(\boldsymbol{p}, \boldsymbol{\mu})\right]_{x'}, \\ \boldsymbol{p} \leftarrow \frac{1}{n+1}(n\boldsymbol{p} + \boldsymbol{e}_i), & i \in \mathrm{argmax}_{i'\in[K]} \left[\nabla_{\boldsymbol{p}} F(\boldsymbol{p}, \boldsymbol{\mu})\right]_{i'}, \end{cases} \tag{4}$$

breaking tie arbitrarily. Here, $\boldsymbol{p}$ denotes the current proportional allocation across all arms, making $1/(n+1)$ a natural choice of step size. Taken together, this yields a simple modification of classical Frank–Wolfe, which we call *Frank–Wolfe Self-Play* (FWSP): in each round, each player treats the opponent's current mixed strategy as fixed and takes a greedy Frank–Wolfe step with respect to the linearized payoff. Remarkably, under minimal regularity assumptions this idea extends to a broad class of pure-exploration problems and bandit models.

A Case Study on Linear Bandits. Consider a linear bandit with unknown parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$ and design matrix $A \in \mathbb{R}^{K\times d}$ whose i-th row is the arm feature vector $\boldsymbol{a}_i^\top$, so the mean reward for arm i is $m_i \triangleq \boldsymbol{a}_i^\top \boldsymbol{\theta}$. Pulling arm i yields a reward $Y \sim \mathcal{N}(m_i, \sigma_i^2)$; rewards are independent across pulls (and across arms). We instantiate (1) for identifying the best arm in linear bandits. Let $I^* = \operatorname{argmax}_{i\in[K]} m_i$ denote the (assumed unique) best arm. For each $x \in [K]\backslash\{I^*\}$, the alternative set is $\mathrm{Alt}_x(\boldsymbol{\theta}) = \left\{\boldsymbol{\vartheta} \in \mathbb{R}^d : \boldsymbol{a}_{I^*}^\top \boldsymbol{\vartheta} < \boldsymbol{a}_x^\top \boldsymbol{\vartheta}\right\}$, and furthermore,

$$D(\boldsymbol{p}, x) = \frac{(\mu_{I^*} - \mu_x)^2}{2\|\boldsymbol{a}_{I^*} - \boldsymbol{a}_x\|^2_{V_{\boldsymbol{p}}^{-1}}}, \quad \text{where} \quad \boldsymbol{p} \in \Delta_K, V_{\boldsymbol{p}} \triangleq \sum_{i=1}^K p_i \sigma_i^{-2} \boldsymbol{a}_i \boldsymbol{a}_i^\top, \text{ and } \|\boldsymbol{v}\|^2_M \triangleq \boldsymbol{v}^\top M \boldsymbol{v}.$$

Consider linear bandits with unit variances, unknown parameter $\boldsymbol{\theta} = \boldsymbol{e}_1 \in \mathbb{R}^2$, and three arms specified by $\boldsymbol{a}_1 = \boldsymbol{e}_1, \boldsymbol{a}_2 = \boldsymbol{e}_2$, and $\boldsymbol{a}_3 = -2\boldsymbol{e}_1$. We have $I^* = 1$, and $j \in \{2, 3\}$. Note that $D(\boldsymbol{p}, 2) < D(\boldsymbol{p}, 3)$ for all $\boldsymbol{p}$. Consequently, the maximin value in (1) is

$$\frac{1}{2} \max_{\boldsymbol{p}\in\mathcal{S}_3} \frac{p_2(p_1 + 4p_3)}{p_1 + p_2 + 4p_3}.$$

There is a unique optimal allocation $\boldsymbol{p}^* = (0, 2/3, 1/3)$. This example contrasts with the unstructured bandit setting: (i) no information balance condition holds as $D(\boldsymbol{p}, 2) < D(\boldsymbol{p}, 3)$, and (ii) the optimal allocation may assign zero mass to some arms, including the best arm.

Thompson sampling draws posterior samples and acts greedily with respect to the sampled values, whereas the top-two algorithm (Russo 2020) always selects the empirical best arm as the top candidate.[1] For both algorithms, arm 1 is selected too often, which is not ideal because arm 3 is more informative than arm 1 since both arms are colinear with $\boldsymbol{\theta}$ but $\boldsymbol{a}_3$ has a larger norm.

On the contrary, FWSP samples greedily according to the linearized objective. Since $D(\boldsymbol{p}, 2) < D(\boldsymbol{p}, 3)$, $\boldsymbol{\mu}$ quickly converges to $\boldsymbol{\mu}^* = (\mu_2^*, \mu_3^*) = (1, 0)$ and hence $\nabla_{\boldsymbol{p}} F(\boldsymbol{p}, \boldsymbol{\mu})$ converges to $\nabla_{\boldsymbol{p}} D(\boldsymbol{p}, 2) = (p_2^2, (p_1 + 4p_3)^2, 4p_2^2)/[2(p_1 + p_2 + 4p_3)^2]$. FWSP always samples the arm with the largest component of $\nabla_{\boldsymbol{p}} F$, which implies zero allocation[2] to arm 1 and the balancing condition $(p_1 + 4p_3)^2 = 4p_2^2$, leading to the optimal solution $\boldsymbol{p}^* = (0, 2/3, 1/3)$.

Main Result. Our main result is that the discrete updates (4) in FWSP converges in game value. Let $\mathrm{int}(\cdot)$ denote interior and let $F^* = \max_{\boldsymbol{p}\in\Delta_K} \min_{\boldsymbol{\mu}\in\Delta_{|\mathcal{X}|}} F(\boldsymbol{p}, \boldsymbol{\mu})$.

Theorem 1. *Let $(\boldsymbol{p}_n, \boldsymbol{\mu}_n)_{n\geq 0}$ be the sequence generated by the discrete-time updates in* (4) *with initial condition $(\boldsymbol{p}_0, \boldsymbol{\mu}_0) \in \mathrm{int}(\Delta_K) \times \Delta_{|\mathcal{X}|}$. Then $\lim_{n\to\infty} F(\boldsymbol{p}_n, \boldsymbol{\mu}_n) = F^*$.*

Learning Algorithm. Our FWSP algorithm serves as the backbone for learning variants. In particular, we propose a posterior-sampling-based pure-exploration algorithm. At each round $t \in \mathbb{N}_0 \triangleq \{0, 1, \dots\}$, the experimenter selects an arm i_t and observes a random reward Y_{t+1,i_t} with mean μ_{i_t}. We assume access to a posterior updater that, given the new data (i_t, Y_{t+1,i_t}) and the current posterior Π_t, returns the updated posterior Π_{t+1} of $\boldsymbol{\theta}$.

Our learning algorithm samples $\widehat{\boldsymbol{\theta}}_t \sim \Pi_t$ and treats it as ground truth when computing the gradients of $F(\boldsymbol{p}, \boldsymbol{\mu}; \widehat{\boldsymbol{\theta}}_t)$, where we make the dependence of F on $\widehat{\boldsymbol{\theta}}_t$ explicit (recall that the definition of F in (2) implicitly depends on $\boldsymbol{\theta}$). Instead of directly using the FW update on the skeptic's side, we adopt a posterior-sampling-based update similar to top-two Thompson sampling Russo (2020). We summarize our algorithm in Algorithm 1.

[1] Even when the top-two algorithm sets the optimal tuning parameter $\beta = p_1^* = 0$, it fails, since its asymptotic allocation is $(0, 1, 0)$.

[2] Since $p_2^2 < 4p_2^2$ for any $p_2 > 0$. Even if $p_2(0) = 0$, $(p_1 + 4p_3)^2$ is largest and hence drive p_2 positive.

Algorithm 1 Posterior-sample based FWSP

Input: Uninformative prior Π_0, an oracle to update posterior distribution.

1: **for** $t = 0, 1, \ldots, T$ **do**
2: Call the oracle to obtain the posterior Π_t and sample $\widehat{\boldsymbol{\theta}}_t \sim \Pi_t$.
3: Repeatedly sample $\widetilde{\boldsymbol{\theta}}_t \sim \Pi_t$ until $\widetilde{\boldsymbol{\theta}}_t \in \mathrm{Alt}(\widehat{\boldsymbol{\theta}}_t)$.
4: Choose any $x_t \in \{x' \in \mathcal{X}(\widehat{\boldsymbol{\theta}}_t) : \widetilde{\boldsymbol{\theta}}_t \in \mathrm{Alt}_{x'}(\widehat{\boldsymbol{\theta}}_t)\}$ and update $\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \frac{1}{t+1}\big(\boldsymbol{e}_{x_t} - \boldsymbol{\mu}_t\big)$
5: Choose any $i_t \in \mathrm{argmax}_{i\in[K]}\big[\nabla_{\boldsymbol{p}} F(\boldsymbol{p}_t, \boldsymbol{\mu}_{t+1}; \widehat{\boldsymbol{\theta}}_t)\big]_i$ and update $\boldsymbol{p}_{t+1} = \boldsymbol{p}_t + \frac{1}{t+1}\big(\boldsymbol{e}_{i_t} - \boldsymbol{p}_t\big)$
6: Play arm i_t and receive observation Y_{t+1,i_t}
7: **end for**

A Novel Proof Technique. We present what appears to be the first convergence proof for pure exploration via continuous-time dynamics and stochastic approximation. We show exponential convergence of the limiting differential inclusion (DI) via a Lyapunov function; embed the discrete-time FWSP iterates into an interpolated path that follows a mildly perturbed DI; and then use stochastic approximation to prove the iterates track the dynamics and converge. Careful boundary handling is required due to nondifferentiability on the simplex boundary. We illustrate the idea by presenting the continuous-time version and its convergence analysis.

Definition 1. *Define the Linear minimization oracle (LMO) correspondences*[3] *as*

$$\begin{aligned}
&\texttt{LMO}_{\boldsymbol{p}} : \mathrm{int}(\Delta_K) \times \Delta_{|\mathcal{X}|} \rightrightarrows \Delta_K, && \texttt{LMO}_{\boldsymbol{p}}(\boldsymbol{p}, \boldsymbol{\mu}) = \operatorname*{argmax}_{\boldsymbol{q}\in\Delta_K} \boldsymbol{q}^\top \nabla_{\boldsymbol{p}} F(\boldsymbol{p}, \boldsymbol{\mu}),\\
&\texttt{LMO}_{\boldsymbol{\mu}} : \mathrm{int}(\Delta_K) \times \Delta_{|\mathcal{X}|} \rightrightarrows \Delta_{|\mathcal{X}|}, && \texttt{LMO}_{\boldsymbol{\mu}}(\boldsymbol{p}, \boldsymbol{\mu}) = \operatorname*{argmin}_{\boldsymbol{\nu}\in\Delta_{|\mathcal{X}|}} \boldsymbol{\nu}^\top \nabla_{\boldsymbol{\mu}} F(\boldsymbol{p}, \boldsymbol{\mu}).
\end{aligned}$$

We analyze the discrete-time dynamics through its limiting continuous-time dynamics, as the time index n goes to infinity. Specifically, we consider the following natural continuous version of (4):

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{p}(t) \in \texttt{LMO}_{\boldsymbol{p}}(\boldsymbol{p}(t), \boldsymbol{\mu}(t)) - \boldsymbol{p}(t), \quad \boldsymbol{p}(0) = \boldsymbol{p}_0 \in \mathrm{int}(\Delta_K),\\
&\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\mu}(t) \in \texttt{LMO}_{\boldsymbol{\mu}}(\boldsymbol{p}(t), \boldsymbol{\mu}(t)) - \boldsymbol{\mu}(t), \quad \boldsymbol{\mu}(0) = \boldsymbol{\mu}_0 \in \Delta_{|\mathcal{X}|},
\end{aligned} \tag{5}$$

This is a *differential inclusion* (DI) as the LMO correspondences are not necessarily single-valued.

By classical results in differential inclusion theory, for any initial condition $(\boldsymbol{p}_0, \boldsymbol{\mu}_0) \in \mathrm{int}(\Delta_K) \times \Delta_{|\mathcal{X}|}$, there exists an absolutely continuous solution $(\boldsymbol{p}(t), \boldsymbol{\mu}(t))$ defined on $[0, \infty)$ to the continuous-time dynamics (5). Moreover, for any such solution, there exist measurable selections $t \mapsto \boldsymbol{q}(t) \in \Delta_K$ and $t \mapsto \boldsymbol{\nu}(t) \in \Delta_{|\mathcal{X}|}$ such that

$$\boldsymbol{q}(t) \in \texttt{LMO}_{\boldsymbol{p}}(\boldsymbol{p}(t), \boldsymbol{\mu}(t)), \quad \boldsymbol{\nu}(t) \in \texttt{LMO}_{\boldsymbol{\mu}}(\boldsymbol{p}(t), \boldsymbol{\mu}(t)), \quad \dot{\boldsymbol{p}}(t) = \boldsymbol{q}(t) - \boldsymbol{p}(t), \quad \dot{\boldsymbol{\mu}}(t) = \boldsymbol{\nu}(t) - \boldsymbol{\mu}(t).$$

Define the candidate Lyapunov function

$$V(\boldsymbol{p}, \boldsymbol{\mu}) = \max_{\boldsymbol{q}'\in\mathcal{S}_K} (\boldsymbol{q}' - \boldsymbol{p})^\top \nabla_{\boldsymbol{p}} F(\boldsymbol{p}, \boldsymbol{\mu}) - \min_{\boldsymbol{\nu}'\in\mathcal{S}_{|\mathcal{X}|}} (\boldsymbol{\nu}' - \boldsymbol{\mu})^\top \nabla_{\boldsymbol{\mu}} F(\boldsymbol{p}, \boldsymbol{\mu}) \tag{6}$$

Theorem 2 (Exponential convergence)**.** *For $(\boldsymbol{p}(0), \boldsymbol{\mu}(0)) \in \mathrm{int}(\Delta_K) \times \Delta_{|\mathcal{X}|}$, we have*

$$\frac{\mathrm{d}}{\mathrm{d}t} V(\boldsymbol{p}(t), \boldsymbol{\mu}(t)) \le -V(\boldsymbol{p}(t), \boldsymbol{\mu}(t)), \quad \textit{for almost all } t \ge 0.$$

The following theorem establishes that the duality gap vanishes along the trajectory of the continuous-time dynamics and that the objective value converges to the optimal value.

Theorem 3 (Convergence of game value)**.** *Let $(\boldsymbol{p}(t), \boldsymbol{\mu}(t))$ be a solution to the differential inclusion in* (5) *with initial condition $(\boldsymbol{p}_0, \boldsymbol{\mu}_0) \in \mathrm{int}(\Delta_K) \times \Delta_{|\mathcal{X}|}$. Then, as $t \to \infty$,*

$$\mathrm{Gap}(t) \triangleq \mathrm{Gap}(\boldsymbol{p}(t), \boldsymbol{\mu}(t)) = \max_{\boldsymbol{q}\in\Delta_K} F(\boldsymbol{q}, \boldsymbol{\mu}(t)) - \min_{\boldsymbol{\nu}\in\Delta_{|\mathcal{X}|}} F(\boldsymbol{p}(t), \boldsymbol{\nu}) \to 0.$$

Consequently, $F(\boldsymbol{p}(t), \boldsymbol{\mu}(t)) \to F^*$.

[3] A correspondence is a set-valued function.

References

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT-23th Conference on learning theory-2010*, pages 13–p, 2010.

Herman Chernoff. Sequential design of experiments. *Annals of Mathematical Statistics*, 30(3):755–770, 1959.

Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR, 2020.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International conference on machine learning*, pages 427–435. PMLR, 2013.

Cyrille Kone, Emilie Kaufmann, and Laura Richert. Adaptive algorithms for relaxed Pareto set identification. *Advances in Neural Information Processing Systems*, 36:35190–35201, 2023.

Tze Leung Lai and Olivia Yueh-Wen Liao. Efficient adaptive randomization and stopping rules in multi-arm clinical trials for testing a new treatment. *Sequential analysis*, 31(4):441–457, 2012.

Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.

Pierre Ménard. Gradient ascent for active exploration in bandit problems. *arXiv preprint arXiv:1905.08165*, 2019.

Chao Qin and Daniel Russo. Optimizing adaptive experiments: A unified approach to regret minimization and best-arm identification. *arXiv preprint arXiv:2402.10592*, 2024.

Daniel Russo. Simple bayesian algorithms for best-arm identification. *Operations Research*, 68(6):1625–1647, 2020.

Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere. Fast pure exploration via Frank-Wolfe. *Advances in Neural Information Processing Systems*, 34:5810–5821, 2021.