Bandits in Flux: Adversarial Constraints in Dynamic Environments

Tareq Si Salem
Paris Research Center, Huawei Technologies, Boulogne-Billancourt, France

Abstract

We investigate the challenging problem of adversarial multi-armed bandits operating under time-varying constraints, a scenario motivated by numerous real-world applications. To address this complex setting, we propose a novel primal-dual algorithm that extends online mirror descent through the incorporation of suitable gradient estimators and effective constraint handling. We provide theoretical guarantees establishing sublinear dynamic regret and sublinear constraint violation for our proposed policy. Our algorithm achieves state-of-the-art performance in terms of both regret and constraint violation. Empirical evaluations demonstrate the superiority of our approach.

1 INTRODUCTION

The multi-armed bandit (MAB) problem is a canonical framework within the domain of online decision-making. At each time step t, a decision-maker selects an action a_t from a finite action set $\mathcal{A}$. Subsequently, the environment reveals the corresponding cost $f_t(a_t)$. The objective is to minimize the cumulative incurred cost over time. A fundamental challenge in the MAB setting is the exploration-exploitation trade-off, which requires balancing the acquisition of new information about the environment with the exploitation of current knowledge to make optimal decisions.

While the foundational multi-armed bandit problem has been extensively explored for decades (Auer et al., 2002a; Katehakis and Veinott Jr, 1987; Bubeck et al., 2012), its relevance persists in contemporary computing systems. Its applications encompass a wide range of domains, including recommendation systems (Li et al., 2010, 2011), auction design (Avadhanula et al., 2021; Nguyen, 2020; Gao et al., 2021), resource allocation (Fu et al., 2022; Verma et al., 2019), and network routing and scheduling (György et al., 2007; Li et al., 2020; Xu et al., 2024; Huang et al., 2023).

In this work, we focus on a variant of the MAB problem characterized by additional soft constraints. We introduce an additional challenge by imposing soft constraints on the decision-maker. Specifically, when selecting an action at time t, the environment reveals a corresponding constraint $g_t(a_t)$. While momentary constraint violations are permissible, we require long-term constraint satisfaction, ensuring vanishing time-averaged constraint violation.

This formulation is motivated by a variety of real-world applications (Shi and Eryilmaz, 2022; Zhou and Ji, 2022; Deng et al., 2022a). For instance, in the optimization of machine learning model hyperparameters (Joy et al., 2016; Zhou and Ji, 2022), we often aim to minimize validation error while simultaneously ensuring a sufficiently short prediction time for the learned model. Another example arises in wireless communications, where maximizing throughput under energy constraints necessitates the management of a cumulative soft energy consumption limit (Orhan et al., 2012). Beyond these, related constrained bandit formulations arise in optimizing user engagement under fairness constraints (Hadiji et al., 2024), energy-aware client selection in federated learning systems (Zhu et al., 2024), budget-constrained ad allocation (Li and Stoltz, 2022), and fair resource distribution across competing tasks (Wang et al., 2024). In each of these domains, even rare constraint violations can have significant negative impacts, motivating algorithms that explicitly control long-run feasibility while learning from partial feedback.

We make the following contributions:

- We provide a methodology for leveraging the online mirror descent framework for a new variant of the non-stationary MAB problem with soft time-varying constraints.

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

- To our knowledge, the proposed algorithm achieves the tightest known theoretical guarantees for dynamic regret and soft constraint violation in the context of MAB with soft constraints. Specifically, we derive an improved dynamic regret bound of $\tilde{\mathcal{O}}\left(\min\left\{\sqrt{P_T T}, V_T^{1/3} T^{2/3}\right\}\right)$ and constraint violation of $\tilde{\mathcal{O}}\left(\sqrt{T}\right)$, surpassing the state-of-the-art results in (Deng et al., 2022a) of $\tilde{\mathcal{O}}\left(V_T^{1/4} T^{3/4}\right)$ and $\tilde{\mathcal{O}}\left(V_T^{1/4} T^{3/4}\right)$, respectively.
- We empirically validate the superiority of our approach through extensive experiments.

The remainder of the paper is organized as follows. We present related work and problem formulation in Section 2 and Section 3, respectively. We state our main results in Section 4. Our experimental results are in Section 6; we conclude in Section 7.

2 RELATED WORK

Online Convex Optimization and Dynamic Comparators. The full-information counterpart to the MAB problem is the expert problem (Littlestone and Warmuth, 1994). In the expert problem, a decision-maker aims to select an expert from a pool of options, but unlike the MAB setting, the costs associated with all experts are revealed, even for those not selected. This problem is subsumed by the more general framework of Online Convex Optimization (OCO), which considers a full-information online setting where reward functions are convex (rather than linear) and decisions are chosen from a compact convex set (instead of the probability simplex). First introduced by Zinkevich (2003), who showed that projected gradient descent attains sublinear regret. OCO generalizes several online problems, and has exerted a profound influence on the machine learning community (Hazan, 2016; Shalev-Shwartz, 2012; McMahan, 2017). A variant of OCO closely related to our work focuses on dynamic regret (Zinkevich, 2003; Besbes et al., 2015; Jadbabaie et al., 2015; Mokhtari et al., 2016), where performance is evaluated against an optimal time-varying sequence rather than a static optimal decision, wherein the regret is evaluated w.r.t. an optimal comparator sequence instead of an optimal static decision in hindsight.

Deriving worst-case bounds for dynamic regret without additional assumptions is unattainable (Jadbabaie et al., 2015). Therefore, following established methodologies (Jadbabaie et al., 2015), we impose regularity conditions on the comparator sequence, such as assuming slow growth in its *path length* or *temporal variability* of costs, to derive meaningful performance guarantees.

Non-Stationary MAB. The seminal work of Auer et al. (2002b) introduced the concept of modeling non-stationarity in the costs of MABs as an adversarial process. A significant contribution was the proposal of measuring regret against an arbitrary baseline policy, characterized by a difficulty metric based on the frequency of arm switches. This concept was subsequently formalized as switching regret in later studies (Yu and Mannor, 2009; Garivier and Moulines, 2011), equivalent to the comparator sequence's path length in the present work. While other research has focused on regret relative to the absolute variation of arm costs (Slivkins and Upfal, 2008; Besbes et al., 2014), aligning with the temporal variability inherent to our treatment of the problem. This work adheres to the OCO framework for consistency in terminology regarding path length and temporal variability regularity conditions.

Non-Stationary MAB with Constraints. The difficulty of non-stationary MABs is substantially amplified in realistic deployments, where decision-makers must cope not only with fluctuating costs but also with time-varying soft constraints. While several studies (Shi and Eryilmaz, 2022; Zhou and Ji, 2022; Deng et al., 2022a) address this setting under static regret, to the best of our knowledge only Deng et al. (2022a) proposes algorithmic approaches for the more challenging dynamic regret regime with jointly non-stationary costs and constraints. Their method models both costs and constraints using Gaussian processes (GPs) and employs an Upper Confidence Bound (UCB)-style exploration rule; non-stationarity is handled via forgetting mechanisms such as sliding windows and restarts (Cheung et al., 2019; Russac et al., 2019; Zhou and Shroff, 2021; Deng et al., 2022b), which effectively construct time-local GP posteriors and UCB statistics. A key limitation of forgetting-based approaches is that performance hinges on committing to a particular forgetting schedule (e.g., a window length or discount factor), often tuned using a known or assumed rate of temporal change and fixed once the algorithm begins; consequently, they can degrade when variation is irregular, adversarial, or unknown. In contrast, our work develops an algorithm with bandit feedback that adapts to non-stationarity *without* requiring such prior tuning, and our analysis provides order-wise optimal dynamic regret and cumulative constraint violation guarantees under standard measures of temporal complexity. A close alternative, Cao and Liu (2018), captures only comparator-drift-type complexity (cf. Zinkevich (2003)), remains within a Euclidean framework with two-point feedback, and uses coupled

Paper	Regularity used	Regret (Asm. 3)	Violation (Asm. 3)	Regret (Agnostic)	Violation (Agnostic)
Static Unconstrained MAB					
(Auer et al., 2002a)	none	$\mathcal{O}(\sqrt{T})$	N/A	N/A	N/A
Dynamic Unconstrained MAB					
(Zhao et al., 2021)	P_T	$T^{3/4}(1+P_T)^{1/2}$	N/A	$T^{3/4}(1+P_T)^{1/2}$	N/A
(Deng et al., 2022b)	V_T	$V_T^{1/4}T^{3/4}$	N/A	$V_T^{1/4}T^{3/4}$	N/A
(Kim and Yun, 2025)	P_T	×	N/A	$P_T\sqrt{T}$	N/A
(Chen et al., 2025)	both	$\min\left\{V_T^{1/3}T^{2/3}, \sqrt{P_T T}\right\}$	N/A	×	N/A
Dynamic Constrained MAB					
(Cao and Liu, 2018)	P_T	$\sqrt{P_T T}$	$(P_T)^{1/4}T^{3/4}$	×	×
(Deng et al., 2022a)	V_T	$V_T^{1/4}T^{3/4}$	$V_T^{1/4}T^{3/4}$	$V_T T^{3/4}$	$V_T T^{3/4}$
This work	both	$\min\left\{V_T^{1/3}T^{2/3}, \sqrt{P_T T}\right\}$	$\sqrt{T}$	$\min\left\{\sqrt{P_T}T^{2/3}, V_T^{1/3}T^{7/9}\right\}$	$\sqrt{T}$

Table 1: Comparison of regret and violation bounds under path length (P_T) and variation (V_T) measures. The "Regularity used" column specifies whether the algorithm requires prior knowledge of path length, variation, both, or neither. We consider three settings: static unconstrained, dynamic unconstrained, and dynamic constrained MABs. In the unconstrained case, our algorithm attains the same regret rate as Chen et al. (2025). To the best of our knowledge, this work establishes the tightest known regret and violation guarantees for non-stationary MABs across both constrained and unconstrained settings. Furthermore, the rates are order-optimal, matching the lower bound in Proposition 1. When regularity assumptions are removed, we also provide the best known regret and violation bounds in the constrained setting.

loss and constraint learning rates that yield feasibility bounds scaling with the amount of drift; by contrast, our *single-point*, *non-Euclidean* primal-dual updates decouple the two effects and provide guarantees that also accommodate changing environments.

3 PROBLEM STATEMENT

We consider a system operating over $T \in \mathbb{N}$ time slots. In each slot, a decision-maker selects an action a_t from a set of $n \in \mathbb{N} \setminus \{1\}$ possible actions $\mathcal{A} = \{1, 2, \dots, n\}$. Subsequently, the environment discloses the corresponding cost f_{t,a_t} and constraint g_{t,a_t} for the selected action a_t. This setting represents a multi-armed bandit problem with bandit feedback. An oblivious adversary determines the losses and constraints by selecting loss vectors $\boldsymbol{f}_t \triangleq (f_{t,a})_{a\in A}$ and constraint vectors $\boldsymbol{g}_t \triangleq (g_{t,a})_{a\in A}$ for each time slot $t \in [\![T]\!]$ at the beginning of the game. We assume bounded losses and constraints. Formally,

Assumption 1. *The loss vectors and constraint vectors are generated by an oblivious adversary and are assumed to be uniformly bounded, with $\boldsymbol{f}_t \in [0,1]^n$ and $\boldsymbol{g}_t \in [-1,1]^n$ for all $t \in [\![T]\!]$.*

This boundedness assumption is necessary to obtain meaningful guarantees: if losses were unbounded, an adversary could force an arbitrarily large loss in the first round. Once boundedness is imposed, we normalize both the losses and the constraints solely for notational convenience; *this entails no loss of generality.*

At each time step t, the decision-maker selects an action according to a distribution $\boldsymbol{x}_t$ over the action space $\mathcal{A}$. This distribution is represented by a probability vector $\boldsymbol{x}_t$ belonging to the n-dimensional simplex:

$$\triangle_n \triangleq \{\boldsymbol{x} \in [0,1]^n : \|\boldsymbol{x}\|_1 = 1\}. \tag{1}$$

The expected loss and constraint violation are defined as $f_t(\boldsymbol{x}) \triangleq \boldsymbol{f}_t \cdot \boldsymbol{x}$ and $g_t(\boldsymbol{x}) \triangleq \boldsymbol{g}_t \cdot \boldsymbol{x}$ for every $\boldsymbol{x} \in \triangle_n$ and $t \in [\![T]\!]$. For notational convenience, we adopt the shorthand $f_t(a)$ and $g_t(a)$ to denote $f_t(\boldsymbol{e}_a) = f_{t,a}$ and $g_t(\boldsymbol{e}_a) = g_{t,a}$, respectively, where $\boldsymbol{e}_a$ is the unit basis vector with a 1 in the a-th position.

We assume the existence of strictly feasible decisions, a condition commonly referred to as Slater's condition in the literature (Boyd and Vandenberghe, 2004). This assumption is standard in the context of time-varying constraints (Valls et al., 2020; Deng et al., 2022a). Formally, we posit the following:

Assumption 2. *There exists a point $\boldsymbol{x} \in \triangle_n$ and a positive constant $1 \geq \rho > 0$ such that $g_t(\boldsymbol{x}) \leq -\rho$ for $\forall t \in [\![T]\!]$.*

Policies. A *policy* $\boldsymbol{\mathcal{P}}$ is the rule for selecting actions based on the history of previously chosen actions together with the corresponding realized costs and constraint values. Formally, we define a policy as a sequence of randomized mappings

$$a_{t+1} = \mathcal{P}_t\left(\{a_s, f_s(a_s), g_s(a_s)\}_{s=1}^{t}\right), \tag{2}$$

where each $\mathcal{P}_t$ maps the trajectory up to round t to a distribution over feasible actions and samples an action a_{t+1} from it. The initial action is drawn according to an initial distribution $a_1 = \mathcal{P}_1(\varnothing) \sim \boldsymbol{x}_1$.

Optimal Comparator Sequence. We select the optimal comparator sequence after T rounds to minimize cumulative loss while adhering to time-varying constraints. For each timeslot $t \in [\![T]\!]$, we define the feasible subset $\triangle_{n,t} \subseteq \triangle_n$ as the set of feasible points within the simplex: $\triangle_{n,t} \triangleq \{\boldsymbol{x} \in \triangle_n : g_t(\boldsymbol{x}) \leq 0\}$. We aim to determine the best comparator sequence, which represents a solution to the following problem:

$$\{\boldsymbol{x}_t^\star\}_{t=1}^T \in \underset{\{\boldsymbol{x}_t\}_{t=1}^T \in \times_{t=1}^T \triangle_{n,t}}{\operatorname{argmin}} \left\{ \sum_{t=1}^T f_t(\boldsymbol{x}_t) \right\}, \quad (3)$$

Performance Metrics. The goal of a policy is to minimize the cumulative loss relative to the comparator decisions while satisfying the constraints. The regret of the policy and the constraint violations are defined as follows, respectively:

$$\mathfrak{R}_T(\boldsymbol{\mathcal{P}}) \triangleq \mathbb{E}_{\boldsymbol{\mathcal{P}}} \left[\sum_{t=1}^T f_t(a_t) \right] - \sum_{t=1}^T f_t(\boldsymbol{x}_t^\star), \quad (4)$$

$$\mathfrak{V}_T(\boldsymbol{\mathcal{P}}) \triangleq \mathbb{E}_{\boldsymbol{\mathcal{P}}} \left[\sum_{t=1}^T g_t(a_t) \right]. \quad (5)$$

The policy $\boldsymbol{\mathcal{P}}$ aims to achieve both sublinear regret $\mathfrak{R}_T(\boldsymbol{\mathcal{P}}) = o(T)$ and sublinear violation $\mathfrak{V}_T(\boldsymbol{\mathcal{P}}) = o(T)$. This implies that the policy's performance gradually approaches that of the best comparator sequence, while remaining feasible on average over time.

Regularity Assumptions. As demonstrated by Jadbabaie et al. (2015), deriving worst-case bounds for dynamic regret without imposing additional assumptions is infeasible. Consequently, we introduce regularity conditions on the non-stationarity of the problem.

We employ two distinct metrics to quantify non-stationarity: the path length, denoted by $P_T \in \mathbb{R}_{\geq 0}$, which indirectly measures fluctuations in the cost functions through the comparator sequence $\{\boldsymbol{x}_t^\star\}_{t=1}^T$, and the cost function temporal variation, denoted by $V_T \in \mathbb{R}_{\geq 0}$, which directly quantifies changes in the cost functions themselves. We formally define these measures as follows:

$$P_T \triangleq \sum_{t=1}^T \|\boldsymbol{x}_t^\star - \boldsymbol{x}_{t+1}^\star\|_1, \quad \text{(path length)} \quad (6)$$

$$V_T \triangleq \sum_{t=1}^T \|\boldsymbol{f}_t - \boldsymbol{f}_{t+1}\|_\infty, \quad \text{(temporal variation)} \quad (7)$$

and we adopt the following assumption.

Assumption 3. *The policy $\boldsymbol{\mathcal{P}}$ has access to the regularity measures P_T and V_T.*

This assumption makes explicit the side information used to tune the algorithm. In Section 5.3 we show how to remove this requirement.

Incomparability of Regularity Assumptions. The two non-stationarity measures, path length P_T and temporal variation V_T, are complementary, each capturing a different aspect of the non-stationarity of the cost functions. Path length quantifies the cumulative variability of the cost functions along a comparator sequence, whereas temporal variation directly measures fluctuations of the cost functions themselves. These measures are inherently incomparable: there exist cases where V_T grows linearly with time ($V_T = \Omega(T)$) while P_T remains constant ($P_T = \mathcal{O}(1)$), as well as cases where V_T is bounded ($V_T = \mathcal{O}(1)$) but P_T grows linearly ($P_T = \Omega(T)$). Concrete examples are provided in Appendix G

Fundamental Limits. Under this problem, the constrained setting strictly generalizes the unconstrained one. Therefore, the known lower bound for unconstrained dynamic regret directly transfers to this setting:

Proposition 1. *((Chen et al., 2025), (Besbes et al., 2015)) For given T, n, V_T, P_T satisfying $T \geq n \geq 2$ and $V_T \leq T/n$, the regret for any policy $\boldsymbol{\mathcal{P}}$ is lower bounded as*

$$\mathfrak{R}(\boldsymbol{\mathcal{P}}) = \Omega\left(\min\left\{V_T^{1/3} T^{2/3}, \sqrt{P_T T}\right\}\right). \quad (8)$$

The $\Omega\left(\sqrt{P_T T}\right)$ term is established by (Chen et al., 2025), while the $\Omega\left((V_T)^{1/3} T^{2/3}\right)$ term follows from the analysis of (Besbes et al., 2015). Combining the two yields the stated lower bound.

4 BCOMD POLICY

Our policy incorporates a Lagrangian function defined as $\Psi(\boldsymbol{x}, \lambda) \triangleq f_t(\boldsymbol{x}) + \lambda g_t(\boldsymbol{x})$. The first term minimizes the cost function, while the second term penalizes soft constraint violations through the Lagrangian multiplier, which serves as an importance weight. Operating in a challenging bandit setting, we estimate the gradient of this function with respect to the distribution $\boldsymbol{x} \in \triangle_n$, but compute the gradient with respect to the dual variable λ exactly.

4.1 Gradient Estimators

In the bandit feedback setting, we construct unbiased estimates of the gradients of the expected loss function $f_t(\boldsymbol{x})$ and constraint $g_t(\boldsymbol{x})$ using the following estimators:

$$\tilde{\boldsymbol{f}}_t = f(a_t) x_{t,a_t}^{-1} \boldsymbol{e}_{a_t}, \qquad \tilde{\boldsymbol{g}}_t = g(a_t) x_{t,a_t}^{-1} \boldsymbol{e}_{a_t}. \quad (9)$$

Here, a_t denotes the action selected at time step t, and $\boldsymbol{e}_{a_t}$ represents a unit basis vector aligned with action

Algorithm 1 BCOMD: Bandit-Feedback Mirror Descent with Time-Varying Constraints.

Require: Initial distribution $\boldsymbol{x}_1 = \mathbf{1}/n$, Initial dual variable $\lambda_1 = 0$, Mirror map $\Phi : \mathcal{D} \subseteq \mathbb{R}^n \to \mathbb{R}$, Learning rate $\eta \in \mathbb{R}_{>0}$, Shift parameter $\gamma \in [0, 1/n]$

1: Assign Ω as in Eq. (17).
2: **for** $t = 1, \ldots, T$ **do**
3: Sample action $a_t \sim \boldsymbol{x}_t$
4: Incur $f_t(a_t)$ and $g_t(a_t)$ ▷ Bandit-feedback costs and constraints
5: Construct estimates $\tilde{\boldsymbol{f}}_t$ and $\tilde{\boldsymbol{g}}_t$ (9)
6: $\tilde{\boldsymbol{\omega}}_t \leftarrow \frac{\Omega}{x_{t,a_t}} \boldsymbol{e}_{a_t}$ ▷ Stabilizer to ensure non-negative pseudo-costs
7: $\tilde{\boldsymbol{b}}_t \leftarrow \tilde{\boldsymbol{\omega}}_t + \tilde{\boldsymbol{f}}_t + \lambda_t \tilde{\boldsymbol{g}}_t$ ▷ Gradient estimate of $\Psi(\cdot, \lambda_t)$ at $\boldsymbol{x}_t$
8: $\boldsymbol{y}_{t+1} \leftarrow (\nabla\Phi)^{-1}(\nabla\Phi(\boldsymbol{x}_t) - \eta \tilde{\boldsymbol{b}}_t)$ ▷ Adapt primal variable
9: $\boldsymbol{x}_{t+1} \leftarrow \Pi^{\Phi}_{\triangle_{n,\gamma} \cap \mathcal{D}}(\boldsymbol{y}_{t+1})$ ▷ Bregman projection onto $\triangle_{n,\gamma}$
10: $\lambda_{t+1} \leftarrow (\lambda_t + \mu g_t(a_t))_+$ ▷ Adapt dual variable
11: **end for**

a_t. The gradient estimates are unbiased, i.e.:

$$\mathbb{E}_{a \sim \boldsymbol{x}} \left[f(a_t) x_a^{-1} \boldsymbol{e}_a \right] = \nabla f_t(\boldsymbol{x}), \tag{10}$$

The same property holds for the constraint gradient estimates $\tilde{\boldsymbol{g}}_t$.

Online Gradient Descent (OGD) methods are standard tools in online learning (Hazan, 2016; Shalev-Shwartz, 2012; McMahan, 2017). However, they are not applicable in our setting: their $\mathcal{O}(\sqrt{T})$ regret guarantees hinge on a bounded second moment of the (possibly stochastic) gradient estimator. Under bandit feedback over the simplex this condition fails—the second moment is unbounded,

$$\sup \left\{ \mathbb{E}_{a \sim \boldsymbol{x}} \left[\|\tilde{\boldsymbol{f}}_t\|_2^2 \right] : \boldsymbol{x} \in \triangle_n \right\} = \infty \tag{11}$$

whenever $f_t(a) > 0$ for some a, since taking $x_a \to 0$ makes the estimator explode. In this regime, the Euclidean-norm variance terms that appear in OGD analyses (Hazan, 2016) can diverge in the bandit regime, and consequently no meaningful regret bounds can be established. While two-point (or multi-point) bandit schemes can control this variance (Chen and Giannakis, 2018), such feedback is unavailable in our setting.

To address this challenge, we propose leveraging the online mirror descent (OMD) framework. OMD allows us to employ the same gradient estimates while departing from Euclidean geometry. A notable example is the EXP-3 policy for adversarial bandits under static regret and no constraints setting (Hazan and Levy, 2014) constitutes a specific instance of OMD that employs an entropic setup and the gradient estimate defined in Equation (9), achieving sublinear regret bounds.

4.2 Mirror Descent Policies

We propose a policy that employs OMD framework to address the challenge of controlling estimator variance. The core principle of mirror descent is the distinction between two spaces: the primal space for variables and the dual space for supergradients. A mirror map connects these spaces. Unlike standard online gradient descent, updates are computed in the dual space before being mapped back to the primal space using the mirror map. For several constrained optimization problems of interest, OMD exhibits faster convergence compared to OGD (Bubeck, 2011, Section 4.3). Within our specific context, we will establish that appropriately configured OMD yields tighter regret bounds with gradient estimators where OGD fails. The OMD policy is well-defined under the following assumption.

Assumption 4. *Let $\Phi : \mathcal{D} \to \mathbb{R}$ be a mirror map, where $\mathcal{D} \subset \mathbb{R}^n$ is an open, convex set. We assume that Φ is compatible with the simplex geometry and satisfies the following standard regularity conditions:*

1. *The probability simplex $\triangle_n$ is contained in the closure of the domain, i.e., $\triangle_n \subset \text{closure}(\mathcal{D})$, and it has a nonempty intersection with the interior domain, $\triangle_n \cap \mathcal{D} \neq \varnothing$.*
2. *The function Φ is strictly convex and continuously differentiable on $\mathcal{D}$, namely $\Phi \in C^1(\mathcal{D})$.*
3. *The gradient mapping induced by Φ covers the entire dual space, i.e., $\nabla\Phi(\mathcal{D}) = \mathbb{R}^n$.*
4. *The gradient norm diverges as one approaches the boundary of the domain: $|\nabla\Phi(x)| \to \infty$ as $x \to \partial\mathcal{D}$.*

A map Φ satisfying these properties is termed a *mirror map*. In particular, the above assumption implies the existence of the inverse of the mapping induced by gradient $\nabla\Phi$, and the uniqueness of the Bregman projection step in Line 9 Algorithm 1. Formally, defined as

Definition 1. *Let $\Phi : \mathcal{D} \to \mathbb{R}$ be a mirror map. The Bregman projection associated to a map Φ onto a convex set $\mathcal{S}$ is denoted by $\Pi^{\Phi}_{\mathcal{S}} : \mathbb{R}^n \to \mathcal{S}$, is defined as*

$$\Pi^{\Phi}_{\mathcal{S}}(\boldsymbol{y}') \triangleq \arg\min_{\boldsymbol{y} \in \mathcal{S}} D_{\Phi}(\boldsymbol{y}, \boldsymbol{y}'), \tag{12}$$

where $D_{\Phi}(\boldsymbol{y}, \boldsymbol{y}') \triangleq \Phi(\boldsymbol{y}) - \Phi(\boldsymbol{y}') - \nabla\Phi(\boldsymbol{y}') \cdot (\boldsymbol{y} - \boldsymbol{y}')$ is the Bregman divergence $D_{\Phi} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}_{\geq 0}$ associated to the map Φ.

Our policy, detailed in Algorithm 1, iteratively updates the action distribution. This update is informed by an OMD step that incorporates a gradient estimate of the cost function and a weighted gradient estimate of the constraint function (Lines 8–9, Algorithm 1). Crucially, we dynamically adjust the weights applied to the constraint function's gradients based on a cumulative measure of constraint violation (Line 10, Algorithm 1).

4.3 Entropic Instantiation

We derive our primary results for the entropic OMD algorithm. This specific instance employs the negative entropy mirror map $\Phi : \mathcal{D} \to \mathbb{R}$ defined as

$$\Phi(\boldsymbol{x}) \triangleq \sum_{i=1}^{n} x_i \log(x_i) \qquad \text{for } \boldsymbol{x} \in \mathcal{D} \triangleq \mathbb{R}^n_{>0}. \tag{13}$$

It is readily verified that this mapping satisfies Assumption 4.

We define the algorithm's domain as the restricted simplex $\triangle_{n,\gamma} = \triangle_n \cap [\gamma, 1]^n$. This construction ensures that the probability of selecting any action remains above a predetermined threshold, γ. This parameter is crucial for establishing the algorithm's theoretical guarantees. Intuitively, γ controls the algorithm's exploration, which is essential for adapting to distribution shifts and rapidly identifying optimal actions. We present our primary regret bound achieved by Algorithm 1 when configured with entropic OMD. Formally,

Theorem 1. *Let* $c_T = \min\left\{\sqrt{P_T}, V_T^{1/3} T^{1/6}\right\}$, $\mu = 1/(M\sqrt{T})$, $\eta = \max\{1, c_T\}/(M\sqrt{T})$, $\gamma = \Theta(T^{-1/2})$, *and* $M = 4\left(\frac{3n+2}{\rho} + 1\right)^2$. *Under Assumptions 1, 2, and 3, the policy* $\boldsymbol{\mathcal{P}}$ *produced by Algorithm 1 satisfies*

$$\begin{aligned} \mathfrak{R}_T(\texttt{BCOMD}) &= \tilde{\mathcal{O}}\left(\min\left\{\sqrt{P_T T}, V_T^{1/3} T^{2/3}\right\}\right), \\ \mathfrak{V}_T(\texttt{BCOMD}) &= \tilde{\mathcal{O}}\left(\sqrt{T}\right). \end{aligned} \tag{14}$$

We employ the operator $\tilde{\mathcal{O}}(\cdot)$ as a variant of the standard big-O notation, $\mathcal{O}(\cdot)$, where logarithmic factors are omitted. A sketch of the proof is provided in the following section. The full proof is deferred to the Appendix.

When decision-makers possess limited knowledge of optimal path length (P_T) and temporal variation (V_T), we propose a meta-algorithm designed to learn these parameters in Section 5.3.

5 FORMAL ANALYSIS OF PERFORMANCE GUARANTEE

We start by presenting various technical lemmas that support the main proof.

5.1 Supporting Lemmas

The history $\mathcal{H}_t$ of Algorithm 1 is fully characterized by its random actions and received feedback up to time t given by:

$$\mathcal{H}_t \triangleq \{(a_s, f_s(a_s), g_s(a_s))\}_{s \in [\![t]\!]}. \tag{15}$$

Lemma 1. *Under the negative entropy mirror map (13), the primal iterates of Algorithm 1 satisfy*

$$\mathbb{E}\left[D_\Phi(\boldsymbol{x}_t, \boldsymbol{y}_{t+1}) \middle| \mathcal{H}_{t-1}\right] \leq 1.5\eta^2 n(1 + \lambda_t^2 + \Omega^2).$$

Proof sketch. We obtain the lemma by instantiating the Bregman divergence induced by the negative-entropy mirror map. Using an appropriate upper bound, we show that the expected discrepancy between the iterates $\boldsymbol{x}_t$ and $\boldsymbol{y}_{t+1}$, measured in this divergence, remains bounded. This guarantee is specific to the mirror-descent geometry and does not extend to OGD, where the analogous term reduces to the Euclidean norm of the (estimated) gradients and need not be bounded. □

The full proof is deferred to Appendix C.1.

Lemma 2. *Under the negative entropy mirror map (13), for any* $t \in [\![T]\!]$ *the variables* λ_t *and* $\boldsymbol{x}_t$ *in Algorithm 1 satisfy the following inequality:*

$$\begin{aligned} &\mathbb{E}\left[\Psi_t(\boldsymbol{x}_t, \lambda) - \Psi_t(\boldsymbol{x}, \lambda_t)\right] \\ &\leq \mathbb{E}\left[\eta^{-1}\left(D_\Phi(\boldsymbol{x}, \boldsymbol{x}_t) - D_\Phi(\boldsymbol{x}, \boldsymbol{x}_{t+1})\right)\right] \\ &+ \mathbb{E}\left[(2\mu)^{-1}\left((\lambda_t - \lambda)^2 - (\lambda_{t+1} - \lambda)^2\right)\right] \\ &+ \mathbb{E}\left[\left(\eta^{-1} D_\Phi(\boldsymbol{x}_t, \boldsymbol{y}_{t+1}) + \mu/2\right)\right], \text{ for } \boldsymbol{x} \in \triangle_n,\ \lambda \geq 0. \end{aligned} \tag{16}$$

Proof sketch. We derive the following lemma by exploiting the linearity of the function Ψ with respect to its arguments $\boldsymbol{x}$ and λ. Since both iterates produced by Algorithm 1 are driven by gradient algorithms, with carfeul analysis we can bound the above difference with respect to the update rule of OMD over the distributions and standard OGD over the constraint weights. □

The full proof is deferred to Appendix C.2.

Lemma 3. *Under the negative entropy mirror map (13), the dual variables* λ_t *for* $t \in [\![T]\!]$ *in Algorithm 1 are bounded by*

$$\Omega \triangleq \frac{\log\left(\gamma^{-1}\right)}{\rho}\left(\frac{\mu}{\eta}\right) + \frac{3n}{2\rho}\eta + \frac{1}{2\rho}\mu + \frac{3n}{\rho} + \frac{2}{\rho} + 1, \tag{17}$$

for $\eta \leq \Omega^{-2}$ *and* $\mu \in \mathbb{R}_{>0}$.

Proof sketch. We establish the following Lemma by contradiction. Assuming the existence of a $t = t_0$ such that $\lambda_t > \Omega$, we demonstrate, in conjunction with Lemma 2, that this assumption is untenable. □

The full proof is deferred to Appendix C.4.

Lemma 4. *Let $\gamma \in \mathbb{R}_{>0}$ and $\{\boldsymbol{x}^\star_{t,\gamma}\}_{t=1}^T$ be the projection of the comparator sequence (3) onto $\triangle_{n,\gamma}$. Under the negative entropy mirror map (13), the primal decisions $\{\boldsymbol{x}_t\}_{t=1}^T$ of Algorithm 1 satisfy the following:*

$$\sum_{t=1}^{T} \left(D_\Phi(\boldsymbol{x}^\star_{t,\gamma}, \boldsymbol{x}_t) - D_\Phi(\boldsymbol{x}^\star_{t,\gamma}, \boldsymbol{x}_{t+1})\right) \tag{18}$$

$$\leq \log(1/\gamma)\left(1 + 2\sum_{t=1}^{T-1} \|\boldsymbol{x}^\star_{t+1} - \boldsymbol{x}^\star_t\|\right). \tag{19}$$

Proof sketch. We extend the standard mirror descent analysis of Bubeck (2011) to establish the above relationship. Specifically, for fixed comparator sequences, we derive telescoping terms upper-bounded by the range term $\log(n)$, coinciding with the diameter of our set under the Bregman divergence associated with the negative entropy mirror map. In contrast to the classical approach, our analysis for non-fixed comparator sequences highlights the indispensable role of the parameter $\gamma > 0$ in deriving regret guarantees within the dynamic setting. □

The full proof is deferred to Appendix C.6.

Lemma 5. *Consider a comparator sequence $\{\boldsymbol{u}_t\}_{t=1}^T \in \triangle_n^T$. Under the negative entropy mirror map (13), Algorithm 1 has the following regret guarantee under Asms. 1 and 2 for $\beta = n(3/2 + 3\Omega^2)$, $\eta = \frac{\eta_0}{\sqrt{\beta T}}$ and $\eta_0 > \log(n)/\sqrt{2}$:*

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(a_t) - f_t(\boldsymbol{u}_t)\right] \tag{20}$$

$$\leq \frac{\left(\log(n) + 2\log(1/\gamma)\sum_{t=1}^{T-1}\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_1\right)}{\eta_0}\sqrt{\beta T}$$

$$+ \eta_0\sqrt{\beta T} + 2\gamma T + \frac{\mu T}{2}. \tag{21}$$

Proof sketch. We employ Lemmas 1 through 4 for this proof. Specifically, summing Lemma 2 from $t = 1$ to T and setting $\lambda = 0$ completes the argument. □

The full proof is deferred to Appendix D.1.

Lemma 6. *Assume that $\lambda_t \leq \Omega$ for all $t \in [T]$. Fix an arbitrary comparator sequence $\{\boldsymbol{u}_t\}_{t=1}^T \in \triangle_n^T$. We run Algorithm 1 with the negative-entropy mirror map in (13). Let $\beta = n(3/2 + 3\Omega^2)$ and set $\eta = \eta_0/\sqrt{\beta T}$ with $\eta_0 > \log(n)/\sqrt{2}$. Under these choices, we obtain the following regret guarantee.*

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(a_t) - f_t(\boldsymbol{u}_t)\right] \leq 2\eta_0\sqrt{\beta T} + 2\gamma T + \frac{\mu T}{2}$$

$$+ \frac{4\log(1/\gamma)TV_T}{\eta_0^2 - \log(n)/2}\mathbb{1}\left(E_T\right). \tag{22}$$

where $E_T \triangleq \left(\sum_{t=1}^{T-1}\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_1 > \frac{\eta_0^2 - \log(n)/2}{\log(1/\gamma)}\right)$.

Proof sketch. We extend the argument of Lemma 2 in Jadbabaie et al. (2015) to complete the proof. □

The full proof is deferred to Appendix D.2.

5.2 Proof of Theorem 1

Proof. Combining Lemmas 5 and 6, we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(a_t) - f_t(\boldsymbol{u}_t^\star)\right] \tag{23}$$

$$= \min\left\{2\eta_0\sqrt{\beta T} + \frac{4\log(1/\gamma)TV_T}{\eta_0^2 - \log(n)/2}, \tag{24}$$

$$\left(\frac{\left(\log(n)/2 + \log(\frac{1}{\gamma})P_T\right)}{\eta_0} + \eta_0\right)\sqrt{\beta T}\right\} + 2\gamma T + \frac{\mu T}{2}. \tag{25}$$

We next set $\eta = \eta_0/\sqrt{\beta T}$ with $\eta_0 = \min\{\sqrt{P_T}, V_T^{1/3}T^{1/6}\}/M$, choose $\mu = 1/(M\sqrt{T})$, and take $\gamma = \Theta(T^{-1/2})$. Substituting these parameter choices into the preceding bound yields

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(a_t) - f_t(\boldsymbol{u}_t^\star)\right]$$

$$= \mathcal{O}\left(\min\left\{\sqrt{P_T T}, V_T^{1/3}T^{2/3}\right\}\right). \tag{26}$$

For the cumulative constraint violation, note that the dual update implies

$$\mu^{-1}\mathbb{E}\left[\lambda_{t+1} - \lambda_t \middle| \mathcal{H}_{t-1}\right] \geq g_t(\boldsymbol{x}_t). \tag{27}$$

Therefore, for $\mu = \Theta(T^{-1/2})$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} g_t(a_t)\right] = \mathbb{E}\left[\sum_{t=1}^{T} g_t(\boldsymbol{x}_t)\right] \tag{28}$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T} \mu^{-1}\mathbb{E}\left[\lambda_{t+1} - \lambda_t \middle| \mathcal{H}_{t-1}\right]\right] \tag{29}$$

$$= \mu^{-1}\mathbb{E}[\lambda_{T+1}] \leq \mu^{-1}\Omega = \mathcal{O}\left(\sqrt{T}\right), \tag{30}$$

where the penultimate inequality uses $\mathbb{E}[\lambda_{T+1}] \leq \Omega$.

Finally, our choice of M ensures that $\eta \leq \Omega^{-2}$ (Appendix C.5); hence, Lemma 3 guarantees that the dual iterates remain bounded by Ω, which justifies the bound in (30). This completes the proof. □

5.3 Meta-Algorithm for Learning Path Length and Temporal Variation

We eliminate the need for prior knowledge of P_T or V_T by employing a *meta-algorithm* (MBCOMD in Algorithm 2 in the Appendix). The time horizon is partitioned into geometrically growing phases $\mathcal{I}_m$ of length $L_m = 2^{m-1}$, for $m = 1, \dots, M = \lceil \log_2 T \rceil$. In each phase $\mathcal{I}_m$, we run $K_m = \lceil \log_2 L_m \rceil$ instances of BCOMD from a geometric grid. At the beginning of every phase, the dual variable is reset. The meta-learner maintains a mixture distribution over these experts and aggregates their predictions. Bandit feedback is broadcast to all experts, enabling them to update via entropic mirror descent with appropriate expert-level losses (see Algorithm 2 in the Appendix).

Bandit with Expert Advice (Meta-Bandits). Formally, let $\mathcal{A}$ be the action set as before and $\mathcal{E}$ the expert set with $|\mathcal{E}| = K$. At each round t: (1) each expert $k \in \mathcal{E}$ outputs a distribution $\boldsymbol{x}_t^{(k)} \in \triangle_n$; (2) the learner samples and plays action $a_t \sim \boldsymbol{x}_t^{\text{meta}} \in \triangle_K$; (3) the learner observes the bandit feedback for a_t and constructs unbiased estimates of cost and constraint; (4) the learner updates $\boldsymbol{x}_t^{\text{meta}}$ and the internal states of the experts. The regret in this setting satisfies $\widetilde{\mathcal{O}}\left(\sqrt{nL \log K}\right)$ over any block of length L (see Lemma 21 in Appendix), i.e., sublinear in the number of actions n and only logarithmic in the number of experts K. In our setting, $K_m = \Theta(\log L_m)$, so the per-phase overhead is $\mathcal{O}(\sqrt{nL_m \log K_m})$. Summing across phases with the doubling schedule gives a global overhead $\mathcal{O}(\sqrt{nT \log \log T})$, which is absorbed into $\widetilde{\mathcal{O}}(\cdot)$.

Theorem 2 (Adaptive regret and violation via doubling)**.** *Under Assumptions 1 and 2,* MBCOMD *(Algorithm 2 run across phases $m = 1, \dots, \lceil \log_2 T \rceil$) achieves*

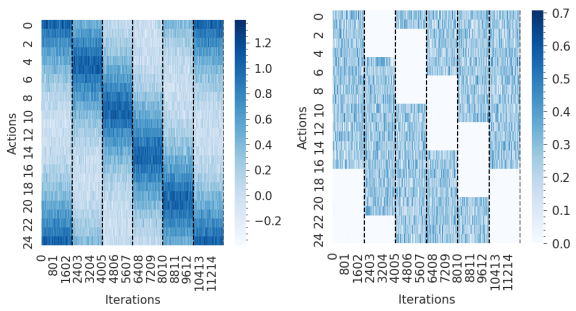
$$\mathfrak{R}_T(\texttt{MBCOMD}) = \tilde{\mathcal{O}}\left(\min\left\{\sqrt{P_T} T^{2/3},\, V_T^{1/3} T^{7/9}\right\}\right),$$
$$\mathfrak{V}_T(\texttt{MBCOMD}) = \tilde{\mathcal{O}}(\sqrt{T}). \tag{31}$$

The full proof is provided in the Appendix E.

6 NUMERICAL EVALUATION

We conclude with a numerical evaluation demonstrating the effectiveness of our proposed policy.

Policies and Hyperparameters. We implement our proposed policy, BCOMD, in Algorithm 1 under the entropic setup. We configure the hyperparameters of BCOMD as follows: $\eta \in [10^{-3}, 4 \times 10^{-2}]$, $\gamma \in [10^{-5}, 10^{-3}]$, and $\mu = \eta/2$, and $\Omega = 0$. Additionally, we implement the R-GP-UCB policy (Deng et al., 2022a) as a benchmark. This policy employs



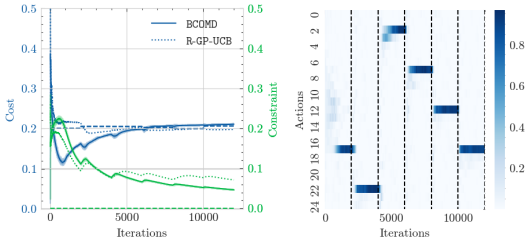
(a) Costs (b) Constraints

Figure 1: Non-stationary Trace Generation Setup. The underlying function changes every 2×10^3 timeslots for six times.

Gaussian Processes (GP) coupled with an Upper Confidence Bound (UCB)-inspired exploration strategy. For R-GP-UCB, we adopt a windowed restarting approach. For R-GP-UCB, we tune the five hyperparameters $\lambda \in [5 \times 10^{-2}, 10^{-1}]$, $W \in [2 \times 10^3, 8 \times 10^3]$, $\delta \in [10^{-4}, 5 \times 10^{-3}]$, $R \in [1, 2]$, $\tau \in [10^{-3}, 10^{-2}]$. To ensure fairness, we maintain a consistent resolution across all hyperparameter grids during the search process. We run the experiment on Intel(R) Xeon(R) Platinum 8164 CPU (104 cores).

Non-Stationary Setup. We construct a demonstrative example with $n = 25$ arms. A synthetic environment was constructed with fixed cost and constraint functions as initial conditions; each arm is associated with costs $f(a) = 1 + \sin(\pi(a/(n-1)))$ and $g(a) = 0.5\mathbb{1}(a \leq n/1.5) - 1/4$. To introduce non-stationarity, the cost and constraint values were cyclically shifted by five arms within each time window of length $W = 2 \times 10^2$, repeating this process six times. The constraints are truncated to values within $[-10^3, \infty)$. Gaussian noise $\xi \sim N(0, 0.1)$ was added independently to the function values at each timeslot to simulate real-world variability. The trace is depicted in Figure 1.

Discussion. We execute both BCOMD and R-GP-UCB policies on the generated trace. Figure 2 (a) presents the simulation results corresponding to the optimal hyperparameter values identified through grid search. The figure clearly demonstrates the superiority of our approach by achieving concurrently lower cumulative costs and constraint violation. Furthermore, we showcase in Figure 2 (b) the robustness of BCOMD by visualizing the learned distribution over arms across time, which reveals the policy's ability to identify optimal and feasible actions in diverse time slots.



(a) (b)

Figure 2: Subfigure (a) illustrates the cumulative costs and constraint satisfaction of both `BCOMD` and `R-GP-UCB` policies under a non-stationary environment. Subfigure (b) presents the action distribution $\mathcal{A}$ acquired by the `BCOMD` algorithm.

7 CONCLUSION

In this work, we present a novel primal-dual algorithm that extends online mirror descent by incorporating tailored gradient estimators and robust constraint handling. We provide rigorous theoretical analysis establishing sublinear dynamic regret and sublinear constraint violation for our proposed policy. Our empirical findings corroborate our theoretical results, demonstrating state-of-the-art performance in terms of both regret and constraint violation.

As avenues for future research, we propose exploring the impact of different mirror maps on algorithm performance. In particular, we are interested in investigating the Tsallis entropy map as a potential alternative to the standard entropy map. This exploration is motivated by the promising results reported in the literature (Zimmert and Seldin, 2021), where the Tsallis entropy map has been shown to achieve optimal regret bounds in both adversarial and stationary stochastic settings without requiring involved concentration reasoning. Secondly, we intend to assess the adaptability of our proposed methods to a wider range of bandit problem formulations, such as combinatorial bandits and bandit-feedback submodular maximization. By incorporating time-varying constraints, we anticipate addressing a broader spectrum of real-world applications.

References

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77.

Avadhanula, V., Colini Baldeschi, R., Leonardi, S., Sankararaman, K. A., and Schrijvers, O. (2021). Stochastic bandits for multi-platform budget optimization in online advertising. In *Proceedings of the Web Conference 2021*, pages 2805–2817.

Besbes, O., Gur, Y., and Zeevi, A. (2014). Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 27.

Besbes, O., Gur, Y., and Zeevi, A. (2015). Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Bubeck, S. (2011). Introduction to online optimization. *Lecture notes*.

Bubeck, S., Cesa-Bianchi, N., et al. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

Bubeck, S. et al. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357.

Cao, X. and Liu, K. R. (2018). Online convex optimization with time-varying constraints and bandit feedback. *IEEE Transactions on automatic control*, 64(7):2665–2680.

Chen, N., Yang, S., and Zhang, H. (2025). Bridging adversarial and nonstationary multi-armed bandit. *Production and Operations Management*, 34(8):2218–2231.

Chen, T. and Giannakis, G. B. (2018). Bandit convex optimization for scalable and dynamic iot management. *IEEE Internet of Things Journal*, 6(1):1276–1286.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. (2019). Learning to optimize under non-stationarity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1079–1087. PMLR.

Deng, Y., Zhou, X., Ghosh, A., Gupta, A., and Shroff, N. B. (2022a). Interference constrained beam alignment for time-varying channels via kernelized bandits. In *2022 20th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt)*, pages 25–32. IEEE.

Deng, Y., Zhou, X., Kim, B., Tewari, A., Gupta, A., and Shroff, N. (2022b). Weighted gaussian process bandits for non-stationary environments. In *International Conference on Artificial Intelligence and Statistics*, pages 6909–6932. PMLR.

Fu, J., Moran, B., and Taylor, P. G. (2022). A restless bandit model for resource allocation, competition, and reservation. *Operations Research*, 70(1):416–431.

Gao, G., Huang, H., Xiao, M., Wu, J., Sun, Y.-E., and Zhang, S. (2021). Auction-based combinatorial multi-armed bandit mechanisms with strategic arms. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10. IEEE.

Garivier, A. and Moulines, E. (2011). On upper-confidence bound policies for switching bandit problems. In *International conference on algorithmic learning theory*, pages 174–188. Springer.

Gupta, S., Goemans, M., and Jaillet, P. (2016). Solving combinatorial games using products, projections and lexicographically optimal bases. *arXiv preprint arXiv:1603.00522*.

György, A., Linder, T., Lugosi, G., and Ottucsák, G. (2007). The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8(10).

Hadiji, H., Gerchinovitz, S., Loubes, J.-M., and Stoltz, G. (2024). Diversity-preserving k-armed bandits, revisited. *Transactions on Machine Learning Research Journal*.

Hazan, E. (2016). Introduction to online convex optimization. *Foundations and Trends® in Optimization*.

Hazan, E. and Levy, K. (2014). Bandit convex optimization: Towards tight bounds. *NeurIPS*.

Huang, J., Golubchik, L., and Huang, L. (2023). Queue scheduling with adversarial bandit learning. *arXiv preprint arXiv:2303.01745*.

Jadbabaie, A., Rakhlin, A., Shahrampour, S., and Sridharan, K. (2015). Online optimization: Competing with dynamic comparators. In *AISTATS*.

Joy, T. T., Rana, S., Gupta, S., and Venkatesh, S. (2016). Hyperparameter tuning for big data using bayesian optimisation. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2574–2579. IEEE.

Katehakis, M. N. and Veinott Jr, A. F. (1987). The multi-armed bandit problem: decomposition and computation. *Mathematics of Operations Research*, 12(2):262–268.

Kim, J. and Yun, S. (2025). Adversarial bandits against arbitrary strategies.

Li, F., Yu, D., Yang, H., Yu, J., Karl, H., and Cheng, X. (2020). Multi-armed-bandit-based spectrum scheduling algorithms in wireless networks: A survey. *IEEE Wireless Communications*, 27(1):24–30.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670.

Li, L., Chu, W., Langford, J., and Wang, X. (2011). Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306.

Li, Z. and Stoltz, G. (2022). Contextual bandits with knapsacks for a conversion model. *Advances in Neural Information Processing Systems*, 35:35590–35602.

Littlestone, N. and Warmuth, M. K. (1994). The Weighted Majority Algorithm. *Information and computation*.

McMahan, H. B. (2017). A survey of algorithms and analysis for adaptive online learning. *JMLR*.

Mokhtari, A., Shahrampour, S., Jadbabaie, A., and Ribeiro, A. (2016). Online optimization in dynamic environments: Improved regret rates for strongly convex problems. In *CDC*.

Nguyen, K. T. (2020). A bandit learning algorithm and applications to auction design. *Advances in Neural Information Processing Systems*, 33:12070–12079.

Orhan, O., Gündüz, D., and Erkip, E. (2012). Throughput maximization for an energy harvesting communication system with processing cost. In *2012 IEEE Information Theory Workshop*, pages 84–88. IEEE.

Russac, Y., Vernade, C., and Cappé, O. (2019). Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems*, 32.

Shalev-Shwartz, S. (2012). Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*.

Shi, Z. and Eryilmaz, A. (2022). A bayesian approach for stochastic continuum-armed bandit with long-term constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 8370–8391. PMLR.

Slivkins, A. and Upfal, E. (2008). Adapting to a changing environment: the brownian restless bandits. In *COLT*, pages 343–354.

Valls, V., Iosifidis, G., Leith, D., and Tassiulas, L. (2020). Online convex optimization with perturbed constraints: Optimal rates against stronger benchmarks. In *International Conference on Artificial Intelligence and Statistics*, pages 2885–2895. PMLR.

Verma, A., Hanawal, M., Rajkumar, A., and Sankaran, R. (2019). Censored semi-bandits: A framework for resource allocation with censored

feedback. *Advances in Neural Information Processing Systems*, 32.

Wang, S., Xiong, G., and Li, J. (2024). Online restless multi-armed bandits with long-term fairness constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15616–15624.

Xu, Y., Wang, S., Guo, H., Liu, X., and Shao, Z. (2024). Learning to schedule online tasks with bandit feedback. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, page 1975–1983, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Yu, J. Y. and Mannor, S. (2009). Piecewise-stationary bandit problems with side observations. In *Proceedings of the 26th annual international conference on machine learning*, pages 1177–1184.

Zhao, P., Wang, G., Zhang, L., and Zhou, Z.-H. (2021). Bandit convex optimization in non-stationary environments. *Journal of Machine Learning Research*, 22(125):1–45.

Zhou, X. and Ji, B. (2022). On kernelized multi-armed bandits with constraints. *Advances in neural information processing systems*, 35:14–26.

Zhou, X. and Shroff, N. (2021). No-regret algorithms for time-varying bayesian optimization. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE.

Zhu, K., Zhang, F., Jiao, L., Xue, B., and Zhang, L. (2024). Client selection for federated learning using combinatorial multi-armed bandit under long-term energy constraint. *Computer Networks*, 250:110512.

Zimmert, J. and Seldin, Y. (2021). Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *ICML*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**] – Problem formulation and algorithm design are presented in Sections 3 and 4.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**] – Theoretical bounds on regret and constraint violation are derived.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Yes**] – The code is provided in the Supplementary Material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**] – Explicit assumptions are given alongside theorem statements in Section 4.
   (b) Complete proofs of all theoretical results. [**Yes**] – Full proofs are included in the appendix and proof sketches in the main paper.
   (c) Clear explanations of any assumptions. [**Yes**] – Formal statements of assumptions and theorems are discussed and provided.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**] – Experimental section reports setup and parameter choices.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**] – Performance measures and averaging across runs are described in Section 6.
   (d) A description of the computing infrastructure used (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**]

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

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [**Not Applicable**]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [**Not Applicable**]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Not Applicable**]

Technical Supplement

This appendix provides supporting results and deferred proofs. Appendix A reviews standard preliminaries on Bregman divergences and projections and derives a one-step inequality for Online Mirror Descent. Appendix B specializes these tools to the negative-entropy mirror map, establishing KL-diameter and quadratic entropic bounds. Appendix C develops problem-specific lemmas for the bandit primal–dual analysis, including control of the conditional divergence term, a saddle-point inequality, and bounds on the dual iterates. Appendix D proves the main regret guarantees, including bounds in terms of comparator path length P_T and temporal variation V_T. Appendix E presents the adaptive meta-algorithm and its analysis. Finally, Appendix F discusses computational complexity, and Appendix G provides examples showing that P_T and V_T are generally incomparable.

A Standard Preliminaries: Bregman Divergences and Projections

A.1 Three-Point ("Pythagorean") Identity for Bregman Divergences

Lemma 7. *Let $\Phi : \mathcal{D} \to \mathbb{R}$ be differentiable and let D_Φ be the associated Bregman divergence (12). Then for any $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{D}$,*

$$D_\Phi(\boldsymbol{x}, \boldsymbol{y}) + D_\Phi(\boldsymbol{z}, \boldsymbol{x}) - D_\Phi(\boldsymbol{z}, \boldsymbol{y}) = (\nabla\Phi(\boldsymbol{x}) - \nabla\Phi(\boldsymbol{y})) \cdot (\boldsymbol{x} - \boldsymbol{z}). \tag{32}$$

Proof. Expand the l.h.s. expression via the definition of Bregman divergences in Eq. (12):

$$\begin{aligned} &D_\Phi(\boldsymbol{x}, \boldsymbol{y}) + D_\Phi(\boldsymbol{z}, \boldsymbol{x}) - D_\Phi(\boldsymbol{z}, \boldsymbol{y}) && (33)\\ &= (\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{y}) - \nabla\Phi(\boldsymbol{y}) \cdot (\boldsymbol{x} - \boldsymbol{y})) + (\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}) - \nabla\Phi(\boldsymbol{x}) \cdot (\boldsymbol{z} - \boldsymbol{x})) \\ &- (\Phi(\boldsymbol{z}) - \Phi(\boldsymbol{y}) - \nabla\Phi(\boldsymbol{y}) \cdot (\boldsymbol{z} - \boldsymbol{y}) && (34)\\ &= (\nabla\Phi(\boldsymbol{x}) - \nabla\Phi(\boldsymbol{y})) \cdot (\boldsymbol{x} - \boldsymbol{z}). && (35) \end{aligned}$$

We conclude the proof. □

A.2 Bregman Projection Inequality

Lemma 8. *Fix $\boldsymbol{y} \in \mathcal{D}$ and let $\boldsymbol{x} = \Pi^\Phi_{\mathcal{X}}(\boldsymbol{y})$ be the Bregman projection of $\boldsymbol{y}$ onto $\mathcal{X}$. Then for all $\boldsymbol{x}' \in \mathcal{X}$,*

$$D_\Phi(\boldsymbol{x}', \boldsymbol{y}) \geq D_\Phi(\boldsymbol{x}', \boldsymbol{x}). \tag{36}$$

Proof. First, we show that the following holds

$$(\nabla\Phi(\boldsymbol{y}) - \nabla\Phi(\boldsymbol{x})) \cdot (\boldsymbol{x}' - \boldsymbol{x}) \leq 0 \qquad \forall \boldsymbol{x}' \in \mathcal{X}. \tag{37}$$

Since $\boldsymbol{x} = \Pi^\Phi_{\mathcal{X}}(\boldsymbol{y}) = \operatorname{argmin}_{\boldsymbol{x}' \in \mathcal{X}} D_\Phi(\boldsymbol{x}', \boldsymbol{y})$. The first order optimality condition Bubeck et al. (2015) gives:

$$\nabla_{\boldsymbol{x}} D_\Phi(\boldsymbol{x}, \boldsymbol{y}) \cdot (\boldsymbol{x} - \boldsymbol{x}') \leq 0, \qquad \forall \boldsymbol{x}' \in \mathcal{X}. \tag{38}$$

By taking $\nabla_{\boldsymbol{x}} D_\Phi(\boldsymbol{x}, \boldsymbol{y}) = \nabla\Phi(\boldsymbol{x}) - \nabla\Phi(\boldsymbol{y})$ completes the proof of the first part. Second, Lemma 7 gives

$$D_\Phi(\boldsymbol{x}, \boldsymbol{y}) + D_\Phi(\boldsymbol{x}', \boldsymbol{x}) - D_\Phi(\boldsymbol{x}', \boldsymbol{y}) = (\nabla\Phi(\boldsymbol{x}) - \nabla\Phi(\boldsymbol{y})) \cdot (\boldsymbol{x} - \boldsymbol{x}') \leq 0, \qquad \forall \boldsymbol{x}' \in \mathcal{X}. \tag{39}$$

Where the inequality is obtained from the preceding part. This concludes the proof. □

A.3 One-Step Online Mirror Descent Inequality

Lemma 9. *Let $\boldsymbol{x} \in \mathcal{X}$ be arbitrary. At time t, given a gradient (or subgradient) $\boldsymbol{b}_t$, the OMD update rule (1) with state $\boldsymbol{x}_t$ satisfies*

$$\boldsymbol{b}_t \cdot (\boldsymbol{x}_t - \boldsymbol{x}) \leq \frac{1}{\eta} \left(D_\Phi(\boldsymbol{x}, \boldsymbol{x}_t) - D_\Phi(\boldsymbol{x}, \boldsymbol{x}_{t+1}) + D_\Phi(\boldsymbol{x}_t, \boldsymbol{y}_{t+1}) \right). \tag{40}$$

Proof. The OMD algorithm satisfies:

$$\begin{aligned} \boldsymbol{b}_t \cdot (\boldsymbol{x}_t - \boldsymbol{x}) &= \frac{1}{\eta} (\nabla\Phi(\boldsymbol{x}_t) - \nabla\Phi(\boldsymbol{y}_{t+1})) \cdot (\boldsymbol{x}_t - \boldsymbol{x}) && (41)\\ &\leq \frac{1}{\eta} (D_\Phi(\boldsymbol{x}, \boldsymbol{x}_t) - D_\Phi(\boldsymbol{x}, \boldsymbol{y}_{t+1}) + D_\Phi(\boldsymbol{x}_t, \boldsymbol{x}_{t+1})) && \text{(Lemma 7)} \quad (42)\\ &\leq \frac{1}{\eta} \left(D_\Phi(\boldsymbol{x}, \boldsymbol{x}_t) - D_\Phi(\boldsymbol{x}, \boldsymbol{x}_{t+1}) + D_\Phi(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) \right). && \text{(Lemma 8)} \quad (43) \end{aligned}$$

This concludes the proof. □

B Specialization to Negative Entropy

B.1 Negative Entropy Mirror-Map Regularity

The mirror map $\Phi(\boldsymbol{x}) = \sum_{i=1}^{n} x_i \log x_i$ defined on $\mathcal{D} = \mathbb{R}^n_{>0}$ satisfies each of the required conditions in Assumption 4, as detailed below.

- The probability simplex $\triangle_n$ is contained in the closure closure $(\mathcal{D})$ and intersects $\mathcal{D}$ non-trivially ($\triangle_n \cap \mathcal{D} \neq \varnothing$).
- The function Φ is strictly convex and continuously differentiable on $\mathcal{D}$. This follows since, for each coordinate, the mapping $t \mapsto t \log t$ has strictly positive second derivative on $(0, \infty)$.
- The gradient of Φ is given component-wise by $\nabla\Phi(\boldsymbol{x}) = \left(1 + \log x_1, \ldots, 1 + \log x_n\right)^\top$. Consequently, $\nabla\Phi$ maps $\mathcal{D}$ onto $\mathbb{R}^n$: for any $\boldsymbol{y} \in \mathbb{R}^n$, choosing $x_i = e^{y_i - 1}$ yields $\nabla\Phi(\boldsymbol{x}) = \boldsymbol{y}$.
- As $\boldsymbol{x}$ approaches the boundary of $\mathcal{D}$, at least one coordinate satisfies $x_i \to 0^+$, which implies $\log x_i \to -\infty$; hence, $\|\nabla\Phi(x)\|$ diverges as $\boldsymbol{x}$ approaches $\partial\mathcal{D}$.

B.2 Negative-Entropy Mirror Map and Multiplicative Update

Lemma 10. *The iterates generated by Algorithm 1 and negative entropy mirror map (4) on cost functions $\{f_s\}_{t=1}^t$ and constraints $\{g_s\}_{t=1}^t$ satisfy the following:*

$$\boldsymbol{y}_{t+1} = \left(x_{t,a} \exp\left(-\eta \tilde{b}_{t,a}\right)\right)_{a \in \mathcal{A}}, t \in [\![T]\!] \tag{44}$$

Proof. The gradient of the map $\Phi(\boldsymbol{x}) = \sum_{a \in \mathcal{A}} x_a \log(x_a)$ is given by $\nabla\Phi(\boldsymbol{x}) = (\log(x_a) + 1)_{a \in \mathcal{A}}$. Thus, the inverse mapping of $\nabla\Phi : \mathbb{R}^n_{\geq 0} \to \mathbb{R}^n$ is given by

$$(\nabla)^{-1} \Phi(\boldsymbol{x}) = (\exp(x_a - 1))_{a \in \mathcal{A}}\,. \tag{45}$$

Thus, Line 8 in Algorithm 1 yields the following update rule

$$\boldsymbol{y}_{t+1} = \left(\exp((\log(x_{t,a}) + 1) - \eta \tilde{b}_{t,a} - 1)\right)_{a \in \mathcal{A}} = \left(\exp(\log(x_{t,a}) - \eta \tilde{b}_{t,a})\right)_{a \in \mathcal{A}} = \left(x_{t,a} \exp\left(-\eta \tilde{b}_{t,a}\right)\right)_{a \in \mathcal{A}}. \tag{46}$$

This concludes the proof. □

B.3 KL-Diameter Bounds on $\triangle_{n,\gamma}$

Lemma 11. *Consider $\boldsymbol{x}_1 = \mathbf{1}/n \in \triangle_n$. The Bregman divergence $\mathcal{D}_\Phi$ under the negative entropy mirror map Φ (13) has the following upper bounds*

$$D_\Phi(\boldsymbol{x}, \boldsymbol{x}') \leq \log(1/\gamma)\,, \qquad \text{for } \boldsymbol{x}, \boldsymbol{x}' \in \triangle_n. \tag{47}$$

$$D_\Phi(\boldsymbol{x}, \boldsymbol{x}_1) \leq \log(n), \qquad \text{for } \boldsymbol{x} \in \triangle_n. \tag{48}$$

Proof. **First upper-bound.** Take $\boldsymbol{x}, \boldsymbol{x}' \in \triangle_{n,\gamma}$. The Bregman divergence under the negative entropy over the simplex is the KL divergence:

$$D_\Phi(\boldsymbol{x}, \boldsymbol{x}') = \sum_{a \in \mathcal{A}} x_a \log\left(\frac{x_a}{x'_a}\right) = \sum_{a \in \mathcal{A}} x_a \log(x_a) + \sum_{a \in \mathcal{A}} \log\left(\frac{1}{x'_a}\right) \tag{49}$$

$$\leq \sum_{a \in \mathcal{A}} x_a \log(x_a) + \sum_{a \in \mathcal{A}} x_a \log\left(\frac{1}{\min\{x'_{a'} : a' \in \mathcal{A}\}}\right) \tag{50}$$

$$= \sum_{a \in \mathcal{A}} x_a \log(x_a) + \log\left(\frac{1}{\min\{x'_{a'} : a' \in \mathcal{A}\}}\right) \tag{51}$$

$$\leq \log\left(\frac{1}{\gamma}\right). \tag{52}$$

The last inequality is obtained considering that the negative entropy is non-positive over the simplex and $\min\{x'_{a'} : a' \in \mathcal{A}\} = \gamma$ for any $\boldsymbol{x}' \in \triangle_{n,\gamma}$. Recall that $\triangle_{n,\gamma} = \triangle_n \cap [\gamma, 1]^{\mathcal{A}}$. The equality holds for a sparse point $\boldsymbol{x} = \boldsymbol{e}_a$ and a near sparse point $\boldsymbol{x}' = \sum_{a \in A \setminus \{a'\}} \gamma + (n-1)\gamma \boldsymbol{e}_{a'}$ for some $a' \neq a \in \mathcal{A}$.

Second upper-bound. The decision $\boldsymbol{x}_1$ minimizes the negative entropy Φ, corresponding to a maximum entropy scenario. By the first-order condition Bubeck et al. (2015), we have $-\nabla\Phi(\boldsymbol{x}_1)(\boldsymbol{x} - \boldsymbol{x}_1) \leq 0$ as $\boldsymbol{x}_1$ minimizes a convex function.

Consequently, we derive the following upper bound for the Bregman divergence:

$$D_\Phi(\boldsymbol{x}, \boldsymbol{x}_1) \leq \sup_{\boldsymbol{x} \in \triangle_n} \Phi(\boldsymbol{x}) + \log(n) \leq \log(n). \tag{53}$$

This inequality holds due to the non-positivity of $\Phi(\boldsymbol{x})$. The equality holds for sparse points in $\triangle_n \cap \{0,1\}^n$. □

B.4 Quadratic Upper Bound for Entropic Bregman Divergence

Lemma 12. *Let the map Φ be the negative entropy (13). The Bregman divergence $D_\Phi(\boldsymbol{x}, \boldsymbol{y})$ of variables $\boldsymbol{x} \in \triangle_{n,\gamma}$ and $y_a \in x_a e^{b_a}$ for $b_a \in \mathbb{R}_{\leq 0}$ and $a \in \mathcal{A}$ is upper bounded as follows.*

$$D_\Phi(\boldsymbol{x}, \boldsymbol{y}) \leq \frac{1}{2} \sum_{a \in \mathcal{A}} x_a (b_a)^2, \quad \text{for } \eta \in \mathbb{R}^n_{\geq 0}. \tag{54}$$

Proof. From the definition of Bregman divergence (12)

$$D_\Phi(\boldsymbol{x}, \boldsymbol{y}) = \sum_{a \in \mathcal{A}} x_a \log(x_a) + x_a e^{b_a}(\log(x_a) + b_a) - \sum_{a \in \mathcal{A}} (\log(x_a) + b_a + 1) \cdot \left(x_a - x_a e^{b_a}\right) \tag{55}$$

$$= \sum_{a \in \mathcal{A}} x_a \left(e^{b_a} - b_a - 1\right) = \sum_{a \in \mathcal{A}} x_a \xi(b_a), \tag{56}$$

where $\xi(s) = \exp(s) - s - 1$. Considering that $\xi(s) \leq \frac{1}{2}s^2$ for $s \in \mathbb{R}_{\leq 0}$ concludes the proof. □

C Specialized Bounds for the Bandit Primal-Dual Analysis

C.1 Conditional Bregman Divergence Bound

Lemma 13. *Under the negative entropy mirror map (13), the primal iterates of Algorithm 1 satisfy*

$$\mathbb{E}\left[D_\Phi(\boldsymbol{x}_t, \boldsymbol{y}_{t+1}) \middle| \mathcal{H}_{t-1}\right] \leq \frac{3\eta^2 n}{2}(1 + \lambda_t^2 + \Omega^2). \tag{57}$$

Proof. At each timeslot t, we sample an action a from the action space A with probability $x + t, a$. The update rule for OMD is provided in Lemma 10. We subsequently compute the expectation of the bound presented in Lemma 12.

$$\mathbb{E}\left[D_\Phi(\boldsymbol{x}_t, \boldsymbol{y}_{t+1}) \middle| \mathcal{H}_{t-1}\right] \leq \frac{1}{2}\mathbb{E}\left[\sum_{a \in \mathcal{A}} x_{t,a}\left(\eta \tilde{\boldsymbol{b}}_t \cdot \boldsymbol{e}_a\right)^2 \middle| \mathcal{H}_{t-1}\right] \leq \frac{\eta^2}{2} \sum_{a \in \mathcal{A}} x_{t,a}^2 \left(\frac{3\Omega^2}{x_{t,a_t}^2} + \frac{3f_{t,a}^2}{x_{t,a}^2} + \lambda_t^2 \frac{3g_{t,a}^2}{x_{t,a}^2}\right)$$

$$\leq \frac{3\eta^2 n}{2}(1 + \lambda_t^2 + \Omega^2). \tag{58}$$

We used in the second inequality the fact that $(x + y + z)^2 \leq 3x + 3y + 3z$. This concludes the proof. □

C.2 Primal–Dual Saddle-Point Inequality

Lemma 14. *Under the negative entropy mirror map (13), for any $t \in [T]$ the variables λ_t and $\boldsymbol{x}_t$ in Algorithm 1 satisfy the following inequality:*

$$\mathbb{E}\left[\Psi_t(\boldsymbol{x}_t, \lambda) - \Psi_t(\boldsymbol{x}, \lambda_t)\right] \leq \mathbb{E}\left[\frac{1}{\eta}\left(D_\Phi(\boldsymbol{x}, \boldsymbol{x}_t) - D_\Phi(\boldsymbol{x}, \boldsymbol{x}_{t+1})\right) + \frac{1}{2\mu}\left((\lambda_t - \lambda)^2 - (\lambda_{t+1} - \lambda)^2\right)\right] + \mathbb{E}\left[\left(\frac{D_\Phi(\boldsymbol{x}_t, \boldsymbol{y}_{t+1})}{\eta} + \frac{\mu}{2}\right)\right], \qquad \text{for } \boldsymbol{x} \in \triangle_n,\ \lambda \geq 0. \tag{59}$$

Proof. Given any $\lambda_t \geq 0$ the function $\Psi_t(\,\cdot\,, \lambda_t)$ is linear, so the following holds

$$\Psi_t(\boldsymbol{x}, \lambda_t) - \Psi_t(\boldsymbol{x}_t, \lambda_t) = \nabla_{\boldsymbol{x}} \Psi_t(\boldsymbol{x}_t, \lambda_t) \cdot (\boldsymbol{x} - \boldsymbol{x}_t). \tag{60}$$

Given any $\boldsymbol{x}_t \in \triangle_n$ the function $\Psi_t(\boldsymbol{x}_t, \,\cdot\,)$ is linear, so the following holds

$$\Psi_t(\boldsymbol{x}_t, \lambda) - \Psi_t(\boldsymbol{x}_t, \lambda_t) = \frac{\partial \Psi_t(\boldsymbol{x}_t, \lambda_t)}{\partial \lambda} \cdot (\lambda - \lambda_t). \tag{61}$$

Combine the preceding equations to obtain:

$$\Psi_t(\boldsymbol{x}, \lambda_t) - \Psi_t(\boldsymbol{x}_t, \lambda) = \nabla_{\boldsymbol{x}} \Psi_t(\boldsymbol{x}_t, \lambda_t) \cdot (\boldsymbol{x} - \boldsymbol{x}_t) - \frac{\partial \Psi_t(\boldsymbol{x}_t, \lambda_t)}{\partial \lambda} \cdot (\lambda - \lambda_t) \tag{62}$$

Note that from the dual update rule we have

$$(\lambda - \lambda_{t+1})^2 \leq (\lambda - \lambda_t - \mu g_t(a_t))^2 \leq (\lambda - \lambda_t)^2 + \mu^2 (g_t(a_t))^2 - 2\mu g_t(a_t)(\lambda - \lambda_t). \tag{63}$$

This gives

$$\frac{\partial \Psi_t(\boldsymbol{x}_t, \lambda_t)}{\partial \lambda}(\lambda - \lambda_t) = g_t(\boldsymbol{x})(\lambda - \lambda_t) = \mathbb{E}\left[g_t(a_t)(\lambda - \lambda_t) \middle| \mathcal{H}_{t-1}\right] \tag{64}$$

$$\leq \mathbb{E}\left[\frac{1}{2\mu}\left((\lambda - \lambda_t)^2 - (\lambda - \lambda_{t+1})^2\right) + \frac{\mu}{2}(g_t(a_t))^2 \middle| \mathcal{H}_{t-1}\right] \tag{65}$$

$$\leq \mathbb{E}\left[\frac{1}{2\mu}\left((\lambda - \lambda_t)^2 - (\lambda - \lambda_{t+1})^2\right) + \frac{\mu}{2} \middle| \mathcal{H}_{t-1}\right]. \tag{66}$$

Consider the following:

$$\mathbb{E}\left[\Psi_t(\boldsymbol{x}, \lambda_t) - \Psi_t(\boldsymbol{x}, \lambda)\right] = \mathbb{E}\left[\nabla_{\boldsymbol{x}} \Psi_t(\boldsymbol{x}_t, \lambda_t) \cdot (\boldsymbol{x} - \boldsymbol{x}_t) - \frac{\partial \Psi_t(\boldsymbol{x}_t, \lambda_t)}{\partial \lambda} \cdot (\lambda - \lambda_t)\right] \tag{67}$$

$$= \mathbb{E}\left[\tilde{\boldsymbol{b}}_t \cdot (\boldsymbol{x} - \boldsymbol{x}_t) - g_t(a_t) \cdot (\lambda - \lambda_t)\right] \tag{68}$$

$$\leq \mathbb{E}\left[\frac{1}{2\mu}\left((\lambda - \lambda_t)^2 - (\lambda - \lambda_{t+1})^2\right) + \frac{\mu}{2}\right] + \mathbb{E}\left[\frac{1}{\eta}\left(D_\Phi(\boldsymbol{x}, \boldsymbol{x}_t) - D_\Phi(\boldsymbol{x}, \boldsymbol{x}_{t+1}) + D_\Phi(\boldsymbol{x}_t, \boldsymbol{y}_{t+1})\right)\right]. \tag{69}$$

The first equality is obtained from the Eq. (62). The second equation holds since $\tilde{\boldsymbol{b}}_t$ and $g_t(a_t)$ are unbiased estimators of the gradients $\nabla_{\boldsymbol{x}} \Psi_t(\boldsymbol{x}_t, \lambda_t)$ and $\frac{\partial \psi(\boldsymbol{x}_t, \lambda_t)}{\partial \lambda}$, respectively. The inequality follows from Lemma 9. This concludes the proof.

□

C.3 Dual Drift Inequality

Lemma 15. *Under the negative-entropy mirror map (13), the iterates of Algorithm 1 satisfy, for some $\boldsymbol{x}^\star \in \triangle_n$ guaranteed by Assumption 2,*

$$\frac{1}{\mu}\mathbb{E}\left[\lambda_{t+1}^2 - \lambda_t^2\right] \leq 2 - \rho\mathbb{E}[\lambda_t] + \frac{1}{\eta}\left(\mathbb{E}\left[D_\Phi(\boldsymbol{x}^\star, \boldsymbol{x}_t) - D_\Phi(\boldsymbol{x}^\star, \boldsymbol{x}_{t+1})\right]\right) + \left(\frac{3\eta n}{2}(1 + \mathbb{E}\left[\lambda_t^2\right] + \Omega^2) + \frac{\mu}{2}\right). \tag{70}$$

Proof. From Lemma 2, we have for any $\boldsymbol{x} \in \triangle_n$:

$$\mathbb{E}\left[\Psi_t(\boldsymbol{x}_t, 0) - \Psi_t(\boldsymbol{x}, \lambda_t)\right] \leq \mathbb{E}\left[\frac{1}{\eta}\left(D_\Phi(\boldsymbol{x}, \boldsymbol{x}_t) - D_\Phi(\boldsymbol{x}, \boldsymbol{x}_{t+1})\right) + \frac{1}{2\mu}\left(\lambda_t^2 - \lambda_{t+1}^2\right)\right] + \mathbb{E}\left[\left(\frac{D_\Phi(\boldsymbol{x}_t, \boldsymbol{y}_{t+1})}{\eta} + \frac{\mu}{2}\right)\right] \tag{71}$$

Rearranging:

$$\frac{1}{\mu}\mathbb{E}\left[\lambda_{t+1}^2 - \lambda_t^2\right] + \mathbb{E}\left[\Psi_t(\boldsymbol{x}_t, 0) - \Psi_t(\boldsymbol{x}, \lambda_t)\right] \tag{72}$$

$$\leq \mathbb{E}\left[\frac{1}{\eta}\left(D_\Phi(\boldsymbol{x}, \boldsymbol{x}_t) - D_\Phi(\boldsymbol{x}, \boldsymbol{x}_{t+1})\right) + \left(\frac{D_\Phi(\boldsymbol{x}_t, \boldsymbol{y}_{t+1})}{\eta} + \frac{\mu}{2}\right)^2\right] \tag{73}$$

$$\leq \mathbb{E}\left[\frac{1}{\eta}\left(D_\Phi(\boldsymbol{x}, \boldsymbol{x}_t) - D_\Phi(\boldsymbol{x}, \boldsymbol{x}_{t+1})\right) + \left(\frac{3\eta n}{2}(1 + \mathbb{E}\left[\lambda_t^2\right] + \Omega^2) + \frac{\mu}{2}\right)\right]. \tag{74}$$

Consider a feasible action $\boldsymbol{x}^\star$ which always exists for all $t \in [T]$ from Assumption 2, we note that $g_t(\boldsymbol{x}^\star) \leq -\rho$.

$$\frac{1}{\mu}\mathbb{E}\left[\lambda_{t+1}^2 - \lambda_t^2\right] \tag{75}$$

$$\leq \mathbb{E}\left[f_t(\boldsymbol{x}) + \lambda_t g_t(\boldsymbol{x}_t^\star) - f_t(\boldsymbol{x}_t) + \frac{1}{\eta}\left(D_\Phi(\boldsymbol{x}^\star, \boldsymbol{x}_t) - D_\Phi(\boldsymbol{x}^\star, \boldsymbol{x}_{t+1})\right) + \left(\frac{3\eta n}{2}(1 + \lambda_t^2 + \Omega^2) + \frac{\mu}{2}\right)\right] \tag{76}$$

$$\leq \mathbb{E}\left[2 - \rho\lambda_t + \frac{1}{\eta}\left(D_\Phi(\boldsymbol{x}^\star, \boldsymbol{x}_t) - D_\Phi(\boldsymbol{x}^\star, \boldsymbol{x}_{t+1})\right) + \left(\frac{3\eta n}{2}(1 + \lambda_t^2 + \Omega^2) + \frac{\mu}{2}\right)\right]. \tag{77}$$

This concludes the proof. □

C.4 Uniform Bound on the Dual Iterates

Lemma 16. *Under the negative entropy mirror map (13), the dual variables λ_t for $t \in [T]$ in Algorithm 1 are bounded by*

$$\Omega \triangleq \frac{\log\left(\gamma^{-1}\right)}{\rho}\left(\frac{\mu}{\eta}\right) + \frac{3n}{2\rho}\eta + \frac{1}{2\rho}\mu + \frac{3n}{\rho} + \frac{2}{\rho} + 1, \tag{78}$$

for $\eta \leq \Omega^{-2}$ and $\mu \in \mathbb{R}_{>0}$.

Proof. Our proof adopts a similar logical structure to that presented by Chen and Giannakis (2018) for the Euclidean setting.

First case ($t \leq 1/\mu$). Consider that $1 \leq t \leq \frac{1}{\mu}$, then the following holds

$$\lambda_t \leq \lambda_{t-1} + \mu \leq \mu t \leq 1 \leq \Omega. \tag{79}$$

Second case ($1/\mu \leq t \leq T$). Assume that $\frac{1}{\mu} \leq t \leq T$, we prove the claim by contradiction. Assume that T_0 is the first timeslot for which $\lambda_t > \Omega$. Therefore, it holds that $\lambda_{T_0 - \frac{1}{\mu}} \leq \Omega < \lambda_{T_0}$. This yields

$$\frac{1}{\mu}\sum_{t=T_0-\frac{1}{\mu}}^{T_0-1}\mathbb{E}\left[\lambda_{t+1}^2 - \lambda_t^2\right] = \frac{1}{\mu}\mathbb{E}\left[\lambda_{T_0}^2 - \lambda_{T_0-\frac{1}{\mu}}^2\right] > \Omega^2 - \Omega^2 > 0. \tag{80}$$

Use Lemma 15 and Lemma 18 and sum from $T_0 - \frac{1}{\mu}$ to $T_0 - 1$ for the points $\boldsymbol{x}^\star$ satisfying the Slater's condition 2 for $t \in [T_0 - \frac{1}{\mu}, T_0]$ to obtain

$$\frac{1}{\mu}\sum_{t=T_0-\frac{1}{\mu}}^{T_0-1}\mathbb{E}\left[\lambda_{t+1}^2 - \lambda_t^2\right] \leq \frac{2}{\mu} - \rho\sum_{t=T_0-\frac{1}{\mu}}^{T_0-1}\mathbb{E}\left[\lambda_t\right] + \frac{\log\left(\gamma^{-1}\right)}{\eta} + \frac{3\eta n}{2\mu} + \frac{3\eta n}{2}\sum_{t=T_0-\frac{1}{\mu}}^{T_0-1}\mathbb{E}\left[\lambda_t^2\right] + \frac{3\eta n}{2}\Omega^2 + \frac{1}{2}. \tag{81}$$

Note that since $\lambda_{T_0} > \Omega$ then we have

$$\lambda_{T_0 - s} > \Omega - s\mu, \tag{82}$$

where $s \leq \frac{1}{\mu}$. This gives:

$$\sum_{t=T_0-\frac{1}{\mu}}^{T_0-1} \mathbb{E}\left[\lambda_t\right] > \frac{\Omega}{\mu} - \mu \sum_{t=T_0-\frac{1}{\mu}}^{T_0-1} s \geq \frac{\Omega}{\mu} - \frac{1}{\mu}. \tag{83}$$

Thus,

$$\frac{1}{\mu} \sum_{t=T_0-\frac{1}{\mu}}^{T_0-1} \mathbb{E}\left[\lambda_{t+1}^2 - \lambda_t^2\right] \leq \frac{2}{\mu} - \frac{\rho\Omega}{\mu} + \frac{\rho}{\mu} + \frac{\log\left(\gamma^{-1}\right)}{\eta} + \frac{3\eta n}{2\mu} + \frac{3\eta n\Omega^2}{\mu} + \frac{1}{2}. \tag{84}$$

From Eq. (80) and consider $\eta \leq \Omega^{-2}$

$$0 < \frac{2}{\mu} - \frac{\rho\Omega}{\mu} + \frac{\rho}{\mu} + \frac{\log\left(\gamma^{-1}\right)}{\eta} + \frac{3\eta n}{2\mu} + \frac{3\eta n\Omega^2}{\mu} + \frac{1}{2} \tag{85}$$

$$\leq \frac{2}{\mu} - \frac{\rho\Omega}{\mu} + \frac{\rho}{\mu} + \frac{\log\left(\gamma^{-1}\right)}{\eta} + \frac{3\eta n}{2\mu} + \frac{3n}{\mu} + \frac{1}{2}. \tag{86}$$

Multiply each side by $\frac{\mu}{\rho}$

$$0 < \frac{2}{\rho} - \Omega + 1 + \frac{\mu \log\left(\gamma^{-1}\right)}{\rho\eta} + \frac{3\eta n}{2\rho} + \frac{3n}{\rho} + \frac{\mu}{2\rho}. \tag{87}$$

This gives

$$\Omega < \frac{\log\left(\gamma^{-1}\right)}{\rho}\left(\frac{\mu}{\eta}\right) + \frac{3n}{2\rho}\eta + \frac{1}{2\rho}\mu + \frac{3n}{\rho} + \frac{2}{\rho} + 1 = \Omega, \tag{88}$$

which is a contradiction. Then, such $T_0 \leq T$ for which $\lambda_t > \Omega$ does not exist, so $\lambda_t \leq \Omega, \forall t \leq T$. We conclude the proof.

□

C.5 Stability of the Dual Iterates

Lemma 17. *Let $n \in \mathbb{N}$, $\rho > 0$, and $T \geq 1$. Define $M = 4\left(\frac{3n+2}{\rho} + 1\right)^2$. For any scaling constant $c_T \in [1, \sqrt{T}]$, set the learning rate $\eta = \frac{c_T}{M\sqrt{T}}$, the dual learning rate $\mu = \frac{1}{M\sqrt{T}}$, and the restriction threshold $\gamma = T^{-1/2}$. Let Ω be defined as in Eq. (78):*

$$\Omega = \frac{\log\left(\gamma^{-1}\right)}{\rho}\left(\frac{\mu}{\eta}\right) + \frac{3n}{2\rho}\eta + \frac{1}{2\rho}\mu + \frac{3n}{\rho} + \frac{2}{\rho} + 1.$$

Then, for all $T \geq 1$ and all $c_T \in [1, \sqrt{T}]$, the following condition holds:

$$\Omega^2 \leq \frac{1}{\eta}. \tag{89}$$

Proof. Let $K = \left(\frac{3n+2}{\rho} + 1\right)$ so that $M = 4K^2$. Substituting $\gamma = T^{-1/2}$, $\eta = \frac{c_T}{M\sqrt{T}}$, and $\mu = \frac{1}{M\sqrt{T}}$ into the definition of Ω yields

$$\begin{aligned}
\Omega &= \frac{\log\left(\gamma^{-1}\right)}{\rho}\left(\frac{\mu}{\eta}\right) \cdot \frac{3n}{2\rho}\eta + \frac{1}{2\rho}\mu + \frac{3n}{\rho} + \frac{2}{\rho} + 1 \\
&= \frac{\log(T)}{2\rho} \cdot \frac{1}{c_T} \cdot \left(\frac{3n}{\rho} + \frac{2}{\rho} + 1\right) \cdot \frac{3nc_T + 1}{2\rho M\sqrt{T}} \\
&= \underbrace{\left(\frac{\log(T)}{2\rho c_T} + K\right)}\cdot A(c_T) + \underbrace{\frac{3nc_T + 1}{2\rho\sqrt{T}}} \cdot B(c_T) \cdot \frac{1}{M}.
\end{aligned}$$

Since $c_T \in [1, \sqrt{T}]$, we have $B(c_T) \le B_{\max}$ with

$$B_{\max} \triangleq \frac{3n\sqrt{T}+1}{2\rho\sqrt{T}}. \tag{90}$$

Define the upper envelope

$$\Omega_{\max}(c_T) \triangleq \frac{\log(T)}{2\rho c_T} + K + \frac{B_{\max}}{M}, \tag{91}$$

so that $\Omega \le \Omega_{\max}(c_T)$ for all admissible c_T. Therefore, it suffices to show

$$\Omega_{\max}(c_T)^2 \le \frac{1}{\eta} = \frac{M\sqrt{T}}{c_T}. \tag{92}$$

Write $\Omega_{\max}(c_T)$ as $\alpha + \beta/c_T$ where

$$\alpha \triangleq K + \frac{B_{\max}}{M} = K + \frac{3n\sqrt{T}+1}{2\rho\sqrt{T}M}, \qquad \beta \triangleq \frac{\log(T)}{2\rho}. \tag{93}$$

Then (92) is equivalent to

$$\phi(c_T) \triangleq c_T \left(\alpha + \frac{\beta}{c_T}\right)^2 \le M\sqrt{T}. \tag{94}$$

Expanding,

$$\phi(c) = \alpha^2 c + 2\alpha\beta + \frac{\beta^2}{c}, \qquad c > 0. \tag{95}$$

The function ϕ is strictly convex on $(0, \infty)$ because it is the sum of an affine term in c and the strictly convex term $c \mapsto \beta^2/c$. Hence, over the compact interval $[1, \sqrt{T}]$, its maximum is attained at an endpoint:

$$\max_{c \in [1, \sqrt{T}]} \phi(c) = \max\left\{\phi(1), \phi(\sqrt{T})\right\}. \tag{96}$$

Thus, it suffices to verify (94) for $c_T \in \left\{1, \sqrt{T}\right\}$.

Case 1: $c_T = \sqrt{T}$**.** Using $\log(T)/\sqrt{T} \le 1$ for $T \ge 1$ and $M = 4K^2$, we obtain

$$\begin{aligned}\Omega_{\max}(\sqrt{T}) &= K + \frac{\log(T)}{2\rho\sqrt{T}} + \frac{3n\sqrt{T}+1}{2\rho\sqrt{T}M} \\ &\le K + \frac{1}{2\rho} + \frac{3n+1}{2\rho M} = K + \frac{1}{2\rho} + \frac{3n+1}{8\rho K^2}.\end{aligned} \tag{97}$$

To show $\Omega_{\max}(\sqrt{T})^2 \le M = 4K^2$, it is enough to prove $\Omega_{\max}(\sqrt{T}) \le 2K$. By (97), it suffices that

$$K + \frac{1}{2\rho} + \frac{3n+1}{8\rho K^2} \le 2K \quad \Longleftrightarrow \quad \frac{1}{2\rho} + \frac{3n+1}{8\rho K^2} \le K. \tag{98}$$

Multiplying (98) by $\rho > 0$ gives

$$\frac{1}{2} + \frac{3n+1}{8K^2} \le \rho K. \tag{99}$$

Since $\rho K = 3n + 2 + \rho \ge 3n + 2$ and $K^2 \ge 1$, we have

$$\frac{1}{2} + \frac{3n+1}{8K^2} \le \frac{1}{2} + \frac{3n+1}{8} = \frac{3n+5}{8} \le 3n + 2 \le \rho K, \tag{100}$$

where $\frac{3n+5}{8} \leq 3n+2$ holds for all $n \in \mathbb{N}$. Hence $\Omega_{\max}(\sqrt{T}) \leq 2K$ and thus $\phi(\sqrt{T}) \leq M\sqrt{T}$.

Case 2: $c_T = 1$**.** Here $\eta = \frac{1}{M\sqrt{T}}$, so (94) becomes $\Omega_{\max}(1)^2 \leq M\sqrt{T}$, equivalently $\Omega_{\max}(1) \leq 2K, T^{1/4}$ since $M = 4K^2$. Using $M = 4K^2$ and $T \geq 1$,

$$\begin{aligned}\Omega_{\max}(1) &= K + \frac{\log(T)}{2\rho} + \frac{3n\sqrt{T}+1}{2\rho\sqrt{T}M} \\ &\leq K + \frac{\log(T)}{2\rho} + \frac{3n+1}{2\rho M} = K + \frac{\log(T)}{2\rho} + \frac{3n+1}{8\rho K^2}. \end{aligned} \tag{101}$$

Moreover, $\frac{3n+1}{8\rho K^2} \leq \frac{3n+1}{8\rho} \leq K$ (since $K \geq \frac{3n+2}{\rho}$), so from (101),

$$\Omega_{\max}(1) \leq \frac{\log(T)}{2\rho} + 2K. \tag{102}$$

Using $\log(T) \leq 4T^{1/4}$ for $T \geq 1$, we obtain

$$\Omega_{\max}(1) \leq \frac{2}{\rho}T^{1/4} + 2K. \tag{103}$$

Finally, $K = \frac{3n+2}{\rho} + 1 > \frac{2}{\rho}$ implies $\frac{2}{\rho}T^{1/4} \leq KT^{1/4}$, and also $2K \leq 2KT^{1/4}$ since $T^{1/4} \geq 1$. Applying both bounds to (103) yields

$$\Omega_{\max}(1) \leq 2KT^{1/4}, \tag{104}$$

which implies $\phi(1) \leq M\sqrt{T}$.

Since $\phi(c_T) \leq \max\left\{\phi(1), \phi(\sqrt{T})\right\} \leq M\sqrt{T}$ for all $c_T \in [1, \sqrt{T}]$, we conclude $\Omega^2 \leq 1/\eta$ for all valid T and c_T. □

C.6 Telescoping Bound for Drifting Comparators

Lemma 18. *Let $\gamma \in \mathbb{R}_{>0}$ and $\{\boldsymbol{x}^\star_{t,\gamma}\}_{t=1}^T$ be the projection of the comparator sequence (3) onto $\triangle_{n,\gamma}$. Under the negative entropy mirror map (13), the primal decisions $\{\boldsymbol{x}_t\}_{t=1}^T$ of Algorithm 1 satisfy the following:*

$$\sum_{t=t_0}^{T} \left(D_\Phi(\boldsymbol{x}^\star_{t,\gamma}, \boldsymbol{x}_t) - D_\Phi(\boldsymbol{x}^\star_{t,\gamma}, \boldsymbol{x}_{t+1})\right) \leq \xi + 2\log(1/\gamma) \sum_{t=t_0}^{T-1} \|\boldsymbol{x}^\star_{t+1} - \boldsymbol{x}^\star_t\|, \tag{105}$$

where $\xi = \log(n)$ when $t_0 = 1$ and $\xi = \log(1/\gamma)$ for $t_0 > 1$.

Proof. Consider the following:

$$\sum_{t=t_0}^{T} \big(D_\Phi(\boldsymbol{x}^\star_{t,\gamma}, \boldsymbol{x}_t) - D_\Phi(\boldsymbol{x}^\star_{t,\gamma}, \boldsymbol{x}_{t+1})\big) \tag{106}$$

$$\leq \underbrace{D_\Phi(\boldsymbol{x}^\star_{1,\gamma}, \boldsymbol{x}_{T'})}_{\leq \xi} - \underbrace{D_\Phi(\boldsymbol{x}^\star_{T,\gamma}, \boldsymbol{x}_{T+1})}_{\leq 0} + \sum_{t=t_0}^{T-1} \big(D_\Phi(\boldsymbol{x}^\star_{t+1,\gamma}, \boldsymbol{x}_{t+1}) - D_\Phi(\boldsymbol{x}^\star_{t,\gamma}, \boldsymbol{x}_{t+1})\big) \tag{107}$$

$$\leq \xi + \sum_{t=t_0}^{T-1} \big(\nabla\Phi(\boldsymbol{x}^\star_{t+1,\gamma}) - \nabla\Phi(\boldsymbol{x}_{t+1})\big) \cdot \big(\boldsymbol{x}^\star_{t+1,\gamma} - \boldsymbol{x}^\star_{t,\gamma}\big) - \underbrace{D_\Phi(\boldsymbol{x}^\star_{t,\gamma}, \boldsymbol{x}^\star_{t+1,\gamma})}_{\geq 0} \tag{108}$$

$$\leq \xi + \sum_{t=t_0}^{T-1} \underbrace{\big(\nabla\Phi(\boldsymbol{x}^\star_{t+1,\gamma}) - \nabla\Phi(\boldsymbol{x}_{t+1})\big) \cdot \big(\boldsymbol{x}^\star_{t+1,\gamma} - \boldsymbol{x}^\star_{t,\gamma}\big)}_{\text{Cauchy-Schwarz's Ineq.}} - \underbrace{D_\Phi(\boldsymbol{x}^\star_{t,\gamma}, \boldsymbol{x}^\star_{t+1,\gamma})}_{\geq 0} \tag{109}$$

$$\leq \xi + \sum_{t=t_0}^{T-1} \underbrace{\|\nabla\Phi(\boldsymbol{x}^\star_{t+1,\gamma}) - \nabla\Phi(\boldsymbol{x}_{t+1})\|_\infty}_{\leq 2\log(1/\gamma)} \|\boldsymbol{x}^\star_{t+1,\gamma} - \boldsymbol{x}^\star_{t,\gamma}\|_1 \tag{110}$$

$$\leq \xi + 2\log(1/\gamma) \sum_{t=t_0}^{T-1} \|\boldsymbol{x}^\star_{t+1,\gamma} - \boldsymbol{x}^\star_{t,\gamma}\|_1 \leq \xi + 2\log(1/\gamma) \sum_{t=t_0}^{T-1} \|\boldsymbol{x}^\star_{t+1} - \boldsymbol{x}^\star_{t,\gamma}\|_1. \tag{111}$$

This concludes the proof. □

D Regret Guarantees

D.1 Regret Bound in Terms of Comparator Path Length P_T

Lemma 19. *We assume that $\lambda_t \leq \Omega, \forall t \in [T]$. Given a comparator sequence $\{\boldsymbol{u}_t\}_{t=1}^T$ in the simplex $\triangle_n^T$, we apply Algorithm 1 under the negative entropy mirror map (13). For $\beta = n(3/2 + 3\Omega^2)$, $\eta = \frac{\eta_0}{\sqrt{\beta T}}$ and $\eta_0 > \log(n)/\sqrt{2}$, we establish the following regret guarantee:*

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(a_t) - f_t(\boldsymbol{u}_t)\right] \leq \left(\frac{\Big(\log(n) + 2\log(1/\gamma)\sum_{t=1}^{T-1}\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_1\Big)}{\eta_0} + \eta_0\right)\sqrt{\beta T} + 2(1+\Omega)\gamma T + \frac{\mu T}{2}. \tag{112}$$

Proof. Consider the following:

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(a_t) - \sum_{t=1}^{T} f_t(\boldsymbol{x}^\star_t)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[f_t(a_t)\middle|\mathcal{H}_{t-1}\right] - \sum_{t=1}^{T} f_t(\boldsymbol{x}^\star_t)\right] = \mathbb{E}\left[\sum_{t=1}^{T} f_t(\boldsymbol{x}_t) - \sum_{t=1}^{T} f_t(\boldsymbol{x}^\star_t)\right] \tag{113}$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T} f_t(\boldsymbol{x}_t) - \sum_{t=1}^{T} f_t(\boldsymbol{x}^\star_{\gamma,t})\right] + 2(1+\Omega)\gamma T. \tag{114}$$

It holds from Lemma 2 and Lemma 18

$$\sum_{t=1}^{T} \mathbb{E}\left[\Psi_t(\boldsymbol{e}_{a_t}, \lambda) - \Psi_t(\boldsymbol{x}^\star_{\gamma,t}, \lambda_t)\right] = \sum_{t=1}^{T} \mathbb{E}\left[\Psi_t(\boldsymbol{x}_t, \lambda) - \Psi_t(\boldsymbol{x}^\star_{\gamma,t}, \lambda_t)\right] \tag{115}$$

$$\leq \frac{1}{\eta}\left(\log(n) + 2\log(1/\gamma)\sum_{t=1}^{T-1}\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_1 + \lambda^2\right) + \frac{\mu T}{2} + \sum_{t=1}^{T} \mathbb{E}\left[\eta^{-1} D_\Phi(\boldsymbol{x}_t, \boldsymbol{y}_{t+1})\right] \tag{116}$$

It remains to bound the expectation of $D_\Phi(\boldsymbol{x}_t, \boldsymbol{y}_{t+1})$. Lemma 1 and Lemma 3 gives:

$$\mathbb{E}\left[D_\Phi(\boldsymbol{x}_t, \boldsymbol{y}_{t+1})\middle|\mathcal{H}_{t-1}\right] \leq \frac{3\eta^2 n}{2}(1 + 2\Omega^2). \tag{117}$$

Thus, it holds

$$\sum_{t=1}^{T} \mathbb{E}\left[\Psi_t(\boldsymbol{e}_{a_t}, \lambda) - \Psi_t(\boldsymbol{x}^\star_{\gamma,t}, \lambda_t)\right] \tag{118}$$

$$\leq \frac{1}{\eta}\left(\log(n) + 2\log(1/\gamma)\sum_{t=1}^{T-1}\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_1 + \lambda^2\right) + \frac{\mu T}{2} + \eta n(3/2 + 3\Omega^2)T. \tag{119}$$

Substitute the definition of the Lagrangian with for $\lambda = 0$ to obtain the following

$$\sum_{t=1}^{T} \mathbb{E}\left[\Psi_t(\boldsymbol{e}_{a_t}, \lambda) - \Psi_t(\boldsymbol{x}^\star_{\gamma,t}, \lambda_t)|\mathcal{H}_{t-1}\right] = \sum_{t=1}^{T} \mathbb{E}\left[\Psi_t(\boldsymbol{x}_t, \lambda) - \Psi_t(\boldsymbol{x}^\star_{\gamma,t}, \lambda_t)|\mathcal{H}_{t-1}\right] \tag{120}$$

$$= \sum_{t=1}^{T} f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}^\star_{\gamma,t}) - \sum_{t=1}^{T} \lambda_t g_t(\boldsymbol{x}^\star_{\gamma,t}) \tag{121}$$

The comparator point $\boldsymbol{x}^\star_t \in \triangle_{n,t}$, satisfies $g_t(\boldsymbol{x}^\star_t) \leq 0$, so the projected point $\boldsymbol{x}^\star_{t,\gamma}$ satisfies $g_t(\boldsymbol{x}^\star_{t,\gamma}) \leq 2\gamma n$. This gives

$$\sum_{t=1}^{T} f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}^\star_{\gamma,t}) \leq \frac{1}{\eta}\left(\log(n) + 2\log(1/\gamma)\sum_{t=1}^{T-1}\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_1\right) + \frac{\mu T}{2} + \eta n(3/2 + 3\Omega^2)T + 2\gamma n\Omega T. \tag{122}$$

Replace η with $\eta = \frac{\eta_0}{\sqrt{\beta T}}$ in the preceding equation to conclude the proof. □

D.2 Regret Bound in Terms of Temporal Variation V_T

Lemma 20. *Assume that $\lambda_t \leq \Omega$ for all $t \in [T]$. Fix an arbitrary comparator sequence $\{\boldsymbol{u}_t\}_{t=1}^T \in \triangle_n^T$. We run Algorithm 1 with the negative-entropy mirror map in (13). Let $\beta = n(3/2 + 3\Omega^2)$ and set $\eta = \eta_0/\sqrt{\beta T}$ with $\eta_0 > \log(n)/\sqrt{2}$. Under these choices, we obtain the following regret guarantee.*

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(a_t) - f_t(\boldsymbol{u}_t)\right] \leq 2\eta_0\sqrt{\beta T} + \frac{4\log(1/\gamma)TV_T}{\eta_0^2 - \log(n)/2}\mathbb{1}\left(\sum_{t=1}^{T-1}\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_1 > \frac{\eta_0^2 - \log(n)/2}{\log(1/\gamma)}\right)$$
$$+ 2(1+\Omega)\gamma T + \frac{\mu T}{2}. \tag{123}$$

Proof. We adapt the proof of Lemma 2 in Jadbabaie et al. (2015) for our analysis. To begin, we define the following set. Define the following set

$$\mathcal{U}_T \triangleq \left\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T \in \triangle_n : \sum_{t=1}^{T-1}\|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_1 \leq \frac{\eta_0^2 - \log(n)/2}{\log(1/\gamma)}, g_t(\boldsymbol{u}_t) \leq 0, t \in [T]\right\} \tag{124}$$

and

$$\{\boldsymbol{u}_t^\star\}_{t=1}^T \in \operatorname{argmin}_{\{\boldsymbol{u}_t\}_{t=1}^T \in \mathcal{U}_T} \sum_{t=1}^{T} f_t(\boldsymbol{u}_t). \tag{125}$$

The choice of $\eta_0 > \log(n)/\sqrt{2}$ indicates that the fixed comparator with zero path length is in $\mathcal{U}_T$ so $\mathcal{U}_T \neq \varnothing$, and in term the sequence $\{\boldsymbol{u}_t^\star\}_{t=1}^T$ exists.

Lemma 5 yields the following regret bound:

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(a_t) - f_t(\boldsymbol{u}_t)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} f_t(a_t) - f_t(\boldsymbol{u}_t^\star)\right] + \sum_{t=1}^{T} f_t(\boldsymbol{u}_t^\star) - f_t(\boldsymbol{u}_t) \tag{126}$$

$$\leq 3\eta_0\sqrt{\beta T} + \sum_{t=1}^{T} f_t(\boldsymbol{u}_t^\star) - f_t(\boldsymbol{u}_t) + 2(1+\Omega)\gamma T + \frac{\mu T}{2} \tag{127}$$

$$\leq 3\eta_0\sqrt{\beta T} + \left(\sum_{t=1}^{T} f_t(\boldsymbol{u}_t^\star) - f_t(\boldsymbol{u}_t)\right) \mathbb{1}\left(\sum_{t=1}^{T-1} \|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_1 > \frac{\eta_0^2 - \log(n)/2}{\log(1/\gamma)}\right) + 2(1+\Omega)\gamma T + \frac{\mu T}{2}, \tag{128}$$

where the last step follows from the fact that

$$\sum_{t=1}^{T} f_t(\boldsymbol{u}_t^\star) - f_t(\boldsymbol{u}_t) \leq 0 \qquad \text{if } (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T) \in \mathcal{U}_T. \tag{129}$$

Divide the time-horizon to B batches of equal lengths, and pick a fixed comparator within each batch. This yields:

$$\sum_{t=1}^{T} \|\boldsymbol{u}_t - \boldsymbol{u}_{t+1}\|_1 \leq 2B. \tag{130}$$

Take $B = \frac{\eta_0^2 - \log(n)/2}{2\log(1/\gamma)}$, $\boldsymbol{x}_t^\star \in \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}: g_t(\boldsymbol{x})\leq 0} f_t(\boldsymbol{x})$, and any fixed $t_k \in \mathcal{T}_k \triangleq [(k-1)(T/B)+1, k(T/B)]$ we have

$$\sum_{t=1}^{T} f_t(\boldsymbol{u}_t^\star) - f_t(\boldsymbol{u}_t) \leq \sum_{t=1}^{T} f_t(\boldsymbol{u}_t^\star) - f_t(\boldsymbol{x}_t^\star) \tag{131}$$

$$= \sum_{k=1}^{B} \sum_{t\in\mathcal{T}_k} f_t(\boldsymbol{u}_t^\star) - f_t(\boldsymbol{x}_t^\star) \tag{132}$$

$$\leq \sum_{k=1}^{B} \sum_{t\in\mathcal{T}_k} f_t(\boldsymbol{x}_{t_k}^\star) - f_t(\boldsymbol{x}_t^\star) \tag{133}$$

$$\leq \left(\frac{T}{B}\right) \sum_{k=1}^{B} \max_{t\in\mathcal{T}_k} \left\{ f_t(\boldsymbol{x}_{t_k}^\star) - f_t(\boldsymbol{x}_t^\star) \right\}. \tag{134}$$

We now claim that for any $t \in \mathcal{T}_k$, we have

$$f_t(\boldsymbol{x}_{t_k}^\star) - f_t(\boldsymbol{x}_t^\star) \leq 2 \sum_{s\in\mathcal{T}_k} \sup_{\boldsymbol{x}\in\mathbb{\Delta}_n} |f_s(\boldsymbol{x}) - f_{s-1}(\boldsymbol{x})| \tag{135}$$

We prove the claim by contradiction. Assume otherwise then there must be $\hat{t}_k$ such that

$$f_{\hat{t}_k}(\boldsymbol{x}_{t_k}^\star) - f_{\hat{t}_k}(\boldsymbol{x}_{\hat{t}_k}^\star) > 2 \sum_{s\in\mathcal{T}_k} \sup_{\boldsymbol{x}\in\mathbb{\Delta}_n} |f_s(\boldsymbol{x}) - f_{s-1}(\boldsymbol{x})| \tag{136}$$

which gives

$$f_t(\boldsymbol{x}_{t_k}^\star) \leq f_{\hat{t}_k}(\boldsymbol{x}_{\hat{t}_k}^\star) + \sum_{s\in\mathcal{T}_k} \sup_{\boldsymbol{x}\in\mathbb{\Delta}_n} |f_s(\boldsymbol{x}) - f_{s-1}(\boldsymbol{x})| \tag{137}$$

$$< f_{\hat{t}_k}(\boldsymbol{x}_{t_k}^\star) - \sum_{s\in\mathcal{T}_k} \sup_{\boldsymbol{x}\in\mathbb{\Delta}_n} |f_s(\boldsymbol{x}) - f_{s-1}(\boldsymbol{x})| \leq f_t(\boldsymbol{x}_{t_k}^\star). \tag{138}$$

Algorithm 2 MBCOMD: Meta (Agnostic) Bandit-Feedback Mirror Descent with Time-Varying Constraints.

Require: $K_m = \lceil \log_2 L_m \rceil$, grid for $k \in [\![K_m]\!]$: $c_k = 2^k$, $\mu = \frac{1}{M_k\sqrt{L_m}}$, $\eta^{(k)} = \frac{c_k}{M_k\sqrt{L_m}}$, $\gamma = \Theta(L_m^{-1})$, and $\gamma_m^{\text{meta}} = \Theta(L_m^{-1/3})$, $\eta_m^{\text{meta}} = \Theta(L_m^{-1/2})$ (M_k is selected as in Theorem 1)

Initialize $\boldsymbol{x}_{t_m}^{\text{meta}} = \frac{1}{K_m}\mathbf{1}$; for all $k \in [\![K_m]\!]$: $\boldsymbol{x}_{t_m}^{(k)} = \frac{1}{n}\mathbf{1}$; $\lambda_{t_m} = 0$. ▷ phase $\mathcal{I}_m$ with $L_m = |\mathcal{I}_m|$ and $t_m = \min(\mathcal{I}_m)$

for $t = t_m, \dots, t_m + L_m - 1$ **do**

  Each expert outputs $\boldsymbol{x}_t^{(k)} \in \triangle_{n,\gamma}$.

  $\boldsymbol{x}_t^{\text{meta}} = \sum_k x_t^{\text{meta},(k)} \boldsymbol{x}_t^{(k)}$; sample $a_t \sim \boldsymbol{x}_t^{\text{meta}}$.

  Observe $f_t(a_t), g_t(a_t)$; set $\hat{\boldsymbol{f}}_t = \frac{f_t(a_t)}{\boldsymbol{x}_{t,a_t}^{\text{meta}}}\boldsymbol{e}_{a_t}$, $\hat{\boldsymbol{g}}_t = \frac{g_t(a_t)}{\boldsymbol{x}_{t,a_t}^{\text{meta}}}\boldsymbol{e}_{a_t}$.

  For each k: BCOMD$^{(k)}$ (Algorithm 1) step with estimators $\hat{\boldsymbol{f}}_t$, $\hat{\boldsymbol{g}}_t$ and step $\eta^{(k)}$; project onto $\triangle_{n,\gamma}$.

  Meta update: $\tilde{m}_t^{(k)} = \frac{x_{t,a_t}^{(k)}}{x_{t,a_t}^{\text{meta}}} f_t(a_t)$; $\quad y_{t+1}^{\text{meta},(k)} = x_t^{\text{meta},(k)} \exp\big(-\eta_m^{\text{meta}} \tilde{m}_t^{(k)}\big)$, $\boldsymbol{x}_{t+1}^{\text{meta}} = \Pi_{\triangle_{n,\gamma_m^{\text{meta}}}}(\boldsymbol{y}_{t+1}^{\text{meta}})$ (KL projection onto $\triangle_{n,\gamma_m^{\text{meta}}}$).

end for

The preceding relation for $t = t_k$ violates the optimality of definition of $f_t(\boldsymbol{x}_{t_k}^\star)$. Thus, we have

$$\sum_{t=1}^{T} f_t(\boldsymbol{u}_t^\star) - f_t(\boldsymbol{u}_t) \le \left(\frac{T}{B}\right) \sum_{k=1}^{B} \max_{t \in \mathcal{T}_k} \left\{ f_t(\boldsymbol{x}_{t_k}^\star) - f_t(\boldsymbol{x}_t^\star) \right\} \tag{139}$$

$$\le 2\left(\frac{T}{B}\right) \sum_{k=1}^{B} \sum_{s \in \mathcal{T}_k} \sup_{\boldsymbol{x} \in \triangle_n} |f_s(\boldsymbol{x}) - f_{s-1}(\boldsymbol{x})| \tag{140}$$

$$= 2\left(\frac{T}{B}\right) V_T = \frac{4\log(1/\gamma) T V_T}{\eta_0^2 - \log(n)/2}. \tag{141}$$

We conclude the proof. □

E Adaptive Meta-Algorithm

E.1 Meta algorithm MBCOMD

We eliminate the need for prior knowledge of P_T or V_T by employing a *meta-algorithm*. The time horizon is partitioned into geometrically growing phases $\mathcal{I}_m$ of length $L_m = 2^{m-1}$, for $m = 1, \dots, M = \lceil \log_2 T \rceil$. In each phase $\mathcal{I}_m$, we run $K_m = \lceil \log_2 L_m \rceil$ BCOMD from a geometric grid. At the beginning of every phase, the dual variable is reset. The meta-learner maintains a mixture distribution over these experts and aggregates their predictions. Bandit feedback is broadcast to all experts, enabling them to update via entropic mirror descent with appropriate expert-level losses (see Algorithm 2). Formally,

Theorem 3 (Adaptive regret and violation via doubling)**.** *Under Assumptions 1 and 2,* MBCOMD *(Algorithm 2 run across phases $m = 1, \dots, \lceil \log_2 T \rceil$) achieves*

$$\mathfrak{R}_T(\texttt{MBCOMD}) = \tilde{\mathcal{O}}\Big(\min\Big\{\sqrt{P_T} T^{2/3},\, V_T^{1/3} T^{7/9}\Big\}\Big), \tag{142}$$

$$\mathfrak{V}_T(\texttt{MBCOMD}) = \tilde{\mathcal{O}}(\sqrt{T}). \tag{143}$$

Before providing the full proof, we establish the following Lemma for the meta learner.

E.2 Meta-Learner Regret Within a Phase

Lemma 21. *Within a phase $\mathcal{I}$ of length L, Algorithm 2 satisfies*

$$\mathbb{E}\left[\sum_{t \in \mathcal{I}} f_t(a_t)\right] - \min_{k \in [K]} \sum_{t \in \mathcal{I}} \boldsymbol{x}_t^{(k)} \cdot \boldsymbol{f}_t = \mathcal{O}\left(\sqrt{n \log(K) L}\right), \tag{144}$$

where the expectation is taken over the algorithm's randomness.

Proof. We provide the proof in five main steps.

Main inequality. Applying Lemma 2 to the unconstrained instance ($\lambda = \lambda_t = \mu = 0$) gives, for any $\boldsymbol{x}^{\mathrm{meta},\star} \in \triangle_K$,

$$\mathbb{E}\big[\Psi_t(\boldsymbol{x}_t^{\mathrm{meta}}, 0) - \Psi_t(\boldsymbol{x}^{\mathrm{meta},\star}, 0)\big] \le \underbrace{\mathbb{E}\Big[\tfrac{1}{\eta^{\mathrm{meta}}}\Big(D_\Phi(\boldsymbol{x}^{\mathrm{meta},\star}, \boldsymbol{x}_t^{\mathrm{meta}}) - D_\Phi(\boldsymbol{x}^{\mathrm{meta},\star}, \boldsymbol{x}_{t+1}^{\mathrm{meta}})\Big)\Big]}_{(1)} \tag{145}$$

$$+ \underbrace{\tfrac{1}{\eta^{\mathrm{meta}}}\mathbb{E}\big[D_\Phi(\boldsymbol{x}_t^{\mathrm{meta}}, \boldsymbol{y}_{t+1}^{\mathrm{meta}})\big]}_{(2)}. \tag{146}$$

Bounding the divergence term (1)**.** The comparator sequence $\{\boldsymbol{x}^{\mathrm{meta},\star}\}_{t\in\mathcal{I}}$ is fixed. Then by Lemma 18, the telescoping structure implies

$$\sum_{t\in\mathcal{I}} \mathbb{E}[(1)] \;\le\; \frac{1}{\eta^{\mathrm{meta}}}\log(K). \tag{147}$$

Bounding the variance term (2)**.** From Lemma 1, we have

$$\mathbb{E}\Big[D_\Phi(\boldsymbol{x}_t^{\mathrm{meta}}, \boldsymbol{y}_{t+1}^{\mathrm{meta}}) \,\Big|\, \mathcal{H}_{t-1}\Big] \le \tfrac{1}{2}\,(\eta^{\mathrm{meta}})^2 \mathbb{E}\left[\sum_k x_{t,k}^{\mathrm{meta}}\, \tilde{m}_{t,k}^2 \,\middle|\, \mathcal{H}_{t-1}\right]. \tag{148}$$

For expert k, the estimated loss is $\tilde{m}_{t,k} = x_{t,a_t}^{(k)}\, \hat{f}_t(a_t)$, where $\hat{f}_t(a) = \frac{f_t(a_t)}{x_{t,a_t}^{\mathrm{meta}}}\,\mathbb{1}\{a = a_t\}$ is the importance-weighted estimator. Thus, $\tilde{m}_{t,k} = \boldsymbol{x}_t^{(k)} \cdot \hat{\boldsymbol{f}}_t$ and

$$(\tilde{m}_{t,k})^2 = \Big(\sum_a x_{t,a}^{(k)}\, \hat{f}_t(a)\Big)^2 \le \sum_a x_{t,a}^{(k)}\, \hat{f}_t(a)^2. \tag{Jensen}$$

Averaging over experts:

$$\sum_k x_{t,k}^{\mathrm{meta}}\,(\tilde{m}_{t,k})^2 \le \sum_k x_{t,k}^{\mathrm{meta}} \sum_a x_{t,a}^{(k)}\, \hat{f}_t(a)^2 = \sum_a \Big(\sum_k x_{t,k}^{\mathrm{meta}} x_{t,a}^{(k)}\Big)\hat{f}_t(a)^2 = \sum_a x_{t,a}^{\mathrm{meta}}\, \hat{f}_t(a)^2, \tag{149}$$

since $\boldsymbol{x}_t^{\mathrm{meta}} = \sum_k x_{t,k}^{\mathrm{meta}} \boldsymbol{x}_t^{(k)}$.

Taking expectation over $a_t \sim \boldsymbol{x}_t^{\mathrm{meta}}$:

$$\mathbb{E}\left[\sum_a x_{t,a}^{\mathrm{meta}}\, \hat{f}_t(a)^2 \,\middle|\, \mathcal{H}_{t-1}\right] = \mathbb{E}\left[x_{t,a_t}^{\mathrm{meta}}\, \frac{f_t(a_t)^2}{(x_{t,a_t}^{\mathrm{meta}})^2} \,\middle|\, \mathcal{H}_{t-1}\right] = \mathbb{E}\left[\frac{f_t(a_t)^2}{x_{t,a_t}^{\mathrm{meta}}} \,\middle|\, \mathcal{H}_{t-1}\right] \tag{150}$$

$$= \sum_a x_{t,a}^{\mathrm{meta}}\, \frac{f_t(a)^2}{x_{t,a}^{\mathrm{meta}}} = \sum_a f_t(a)^2 \;\le\; n. \tag{151}$$

As shown earlier, the variance factor satisfies

$$\mathbb{E}\left[\sum_k x_{t,k}^{\mathrm{meta}}\, \tilde{m}_{t,k}^2 \,\middle|\, \mathcal{H}_{t-1}\right] \;\le\; n. \tag{152}$$

Hence,

$$\sum_{t\in\mathcal{I}} \mathbb{E}[(2)] \;\le\; \eta^{\mathrm{meta}} n L. \tag{153}$$

Combining bounds. Summing over $t \in \mathcal{I}$, we obtain

$$\sum_{t\in\mathcal{I}} \mathbb{E}\big[\Psi_t(\boldsymbol{x}_t^{\mathrm{meta}}, 0) - \Psi_t(\boldsymbol{x}^{\mathrm{meta},\star}, 0)\big] \;\le\; \frac{1}{\eta^{\mathrm{meta}}}\log(K) + \eta^{\mathrm{meta}} n L. \tag{154}$$

Unrolling the definition of Ψ_t gives

$$\sum_{t\in\mathcal{I}} f_t(\boldsymbol{x}_t^{\text{meta}}) - f_t(\boldsymbol{x}^{\text{meta},\star}) \;\le\; \frac{1}{\eta^{\text{meta}}}\log(K) + \eta^{\text{meta}} nL. \tag{155}$$

Optimizing the learning rate. Choosing $\eta^{\text{meta}} = \Theta\Big(\sqrt{\frac{\log(K)}{nL}}\Big)$ yields

$$\sum_{t\in\mathcal{I}} f_t(\boldsymbol{x}_t^{\text{meta}}) - f_t(\boldsymbol{x}^{\text{meta},\star}) = \mathcal{O}\left(\sqrt{n\log(K)L}\right). \tag{156}$$

This concludes the proof. □

E.3 Proof of Theorem 3

Proof. We begin by analyzing the regret of a fixed base learner, then quantify the additional regret introduced by the meta algorithm in searching for the best such learner, and finally combine these results to obtain the overall regret over the horizon T.

Fixed Base Learner (Phase $\mathcal{I}$). For a fixed base learner $k^\star \in [K]$ with a fractional state $\boldsymbol{x}_t$ and intermediate (yet unprojected) state $\boldsymbol{y}_t$ the following holds

$$\mathbb{E}\left[D_\Phi(\boldsymbol{x}_t, \boldsymbol{y}_{t+1})\Big|\mathcal{H}_{t-1}\right] \le \frac{1}{2}\mathbb{E}\left[\sum_{a\in\mathcal{A}} x_{t,a}\left(\eta\tilde{\boldsymbol{b}}_t\cdot\boldsymbol{e}_a\right)^2\Big|\mathcal{H}_{t-1}\right] \le$$
$$\frac{\eta^2}{2}\sum_{a\in\mathcal{A}} x_{t,a}^{\text{meta}} x_{t,a}\left(\frac{3\Omega^2}{\left(x_{t,a}^{\text{meta}}\right)^2} + \frac{3f_{t,a}^2}{\left(x_{t,a}^{\text{meta}}\right)^2} + \lambda_t^2\frac{3g_{t,a}^2}{\left(x_{t,a}^{\text{meta}}\right)^2}\right) \le \eta^2 n(3/2+3\Omega^2)\left(\frac{1}{\gamma^{\text{meta}}}\right). \tag{157}$$

The factor $\frac{1}{\gamma^{\text{meta}}}$ represents the additional cost introduced by the meta algorithm at this stage, which propagates through the analysis and ultimately worsens the regret bounds. From Lemmas 5 and 6 we have the following regret bound on the base algorithm:

$$\mathfrak{R}_T\left(\texttt{BCOMD}^{(k^\star)}\right) = \tilde{\mathcal{O}}\left(\min\left\{\frac{P_\mathcal{I}}{\eta}, \frac{V_\mathcal{I} L}{\eta_0^2}\right\} + \frac{\eta L}{\gamma^{\text{meta}}}\right). \tag{158}$$

The learning rate is selected to be $\eta = \eta_0/c_\mathcal{I}$, then we have

$$\mathfrak{R}_T\left(\texttt{BCOMD}^{(k^\star)}\right) = \tilde{\mathcal{O}}\left(\min\left\{\frac{P_\mathcal{I}}{\eta_0}c_\mathcal{I}, \frac{V_\mathcal{I} L}{\eta_0^2}\right\} + \frac{\eta_0 L}{\gamma^{\text{meta}} c_\mathcal{I}}\right). \tag{159}$$

Take $c_\mathcal{I} = \Theta(L^{2/3})$, $\eta_0 = \Theta\left(\min\left\{\sqrt{P_\mathcal{I}}, V_\mathcal{I}^{1/3}L^{1/9}\right\}\right)$, and $\gamma^{\text{meta}} = \Theta(T^{-1/3})$.

$$\mathfrak{R}_T\left(\texttt{BCOMD}^{(k^\star)}\right) = \tilde{\mathcal{O}}\left(\min\left\{\frac{P_\mathcal{I}}{\eta_0}c_\mathcal{I}, \frac{V_\mathcal{I} L}{\eta_0^2}\right\} + \frac{\eta_0 L}{\gamma^{\text{meta}} c_\mathcal{I}}\right) = \tilde{\mathcal{O}}\left(\min\left\{\sqrt{P_\mathcal{I}}L^{2/3}, V_\mathcal{I}^{1/3}L^{7/9}\right\}\right). \tag{160}$$

Meta- and Base-Regret Decomposition and Overall Regret. By Lemma 21 and the previous result (160), within $\mathcal{I}_m$ we have

$$\mathbb{E}\left[\sum_{t\in\mathcal{I}_m} f_t(a_t)\right] \le \min_k \sum_{t\in\mathcal{I}_m} \boldsymbol{x}_t^{(k)}\cdot\boldsymbol{f}_t + C\sqrt{L_m\log K_m} \tag{161}$$
$$\le \sum_{t\in\mathcal{I}_m} f_t(\boldsymbol{x}_t^\star) + \tilde{\mathcal{O}}\left(\min\left\{\sqrt{P_\mathcal{I}}L^{2/3}, V_\mathcal{I}^{1/3}L^{7/9}\right\}\right) + C\sqrt{L_m\log K_m}. \tag{162}$$

Summing over $m = 0,\dots,M-1$ and noting that the comparator sequences concatenate across phases, we get

$$\mathbb{E}\Big[\sum_{t=1}^T f_t(a_t)\Big] - \sum_{t=1}^T f_t(\boldsymbol{x}_t^\star) \le \sum_{m=0}^{M-1}\tilde{\mathcal{O}}\left(\min\left\{\sqrt{P_{\mathcal{I}_m}}, L_m^{2/3}, V_{\mathcal{I}_m}^{1/3}L_m^{7/9}\right\}\right) + C\sum_{m=0}^{M-1}\sqrt{L_m\log K_m}. \tag{163}$$

We bound the two sums separately.

Path-length branch. By Cauchy–Schwarz,

$$\sum_{m=0}^{M-1} \sqrt{P_{\mathcal{I}_m}}\, L_m^{2/3} \le \sqrt{\Big(\sum_m P_{\mathcal{I}_m}\Big)\Big(\sum_m L_m^{4/3}\Big)} = \sqrt{P_T \sum_m L_m^{4/3}} = \mathcal{O}\big(\sqrt{P_T}\, T^{2/3}\big). \tag{164}$$

where we used $L_m = 2^m$ and $T = \sum_m L_m = \Theta(2^M)$, and $\sum_m L_m^{4/3} = \Theta\big((2^M)^{4/3}\big) = \Theta(T^{4/3})$.

Variation branch. Apply Hölder with $(p, q) = (3, \frac{3}{2})$:

$$\sum_{m=0}^{M-1} V_{\mathcal{I}_m}^{1/3} L_m^{7/9} \le \Big(\sum_m V_{\mathcal{I}_m}\Big)^{1/3} \Big(\sum_m L_m^{(7/9)\cdot(3/2)}\Big)^{2/3} = V_T^{1/3} \Big(\sum_m L_m^{7/6}\Big)^{2/3}. \tag{165}$$

Again with $L_m = 2^m$, $\sum_m L_m^{7/6} = \Theta\big((2^M)^{7/6}\big) = \Theta(T^{7/6})$, so

$$\sum_m V_{\mathcal{I}_m}^{1/3} L_m^{7/9} = \mathcal{O}\left(V_T^{1/3} T^{7/9}\right). \tag{166}$$

Therefore,

$$\sum_{m=0}^{M-1} \tilde{\mathcal{O}}\Big(\min\Big\{\sqrt{P_{\mathcal{I}_m}} L_m^{2/3},\, V_{\mathcal{I}_m}^{1/3} L_m^{7/9}\Big\}\Big) \le \tilde{\mathcal{O}}\Big(\min\Big\{\sqrt{P_T} T^{2/3},\, V_T^{1/3} T^{7/9}\Big\}\Big), \tag{167}$$

since $\sum_m \min\{a_m, b_m\} \le \min\{\sum_m a_m, \sum_m b_m\}$.

Meta overhead. With $K_m = \lceil \log_2 L_m \rceil$ and $L_m = 2^m$,

$$\sum_{m=0}^{M-1} \sqrt{L_m \log K_m} \le \sum_{m=0}^{M-1} \sqrt{2^m \log(m+1)} = \Theta\big(2^{M/2}\sqrt{\log M}\big) = \Theta\big(\sqrt{T \log\log T}\big), \tag{168}$$

which is absorbed into $\tilde{\mathcal{O}}(\cdot)$ and is dominated by the main terms for nontrivial T.

4. Plugging (167) and (168) into (163) gives the stated regret bound

$$\mathbb{E}\Big[\sum_{t=1}^{T} f_t(a_t)\Big] - \min_{\boldsymbol{x}_t^\star} \sum_{t=1}^{T} f_t(\boldsymbol{x}_t^\star) \;\le\; \tilde{\mathcal{O}}\Big(\min\Big\{\sqrt{P_T} T^{2/3},\, V_T^{1/3} T^{7/9}\Big\}\Big) \tag{169}$$

and the $\sqrt{T \log\log T}$ meta overhead is suppressed in $\tilde{\mathcal{O}}$.

For the cumulative violation, the per-phase guarantee remains $\tilde{\mathcal{O}}(\sqrt{L_m})$ by linearity of $g_t(\cdot)$ and Theorem 1, hence

$$\mathbb{E}\left[\sum_{t=1}^{T} g_t(a_t)\right] = \sum_{m=0}^{M-1} \tilde{\mathcal{O}}(\sqrt{L_m}) = \tilde{\mathcal{O}}(\sqrt{T}). \tag{170}$$

This completes the proof. □

F Computational Complexity

F.1 Time Complexity and Projection Implementation Details

Algorithm 1 instantiated with the negative-entropy mirror map can be implemented in polynomial time. In particular, Lines 5–8 and 10 consist of elementary vector operations whose runtime is linear in the number of actions, $|\mathcal{A}| = n$. The dominant computational cost arises from the projection step onto the truncated simplex $\triangle_{n,\gamma}$. Given a membership oracle for $\triangle_{n,\gamma}$, such a projection can be carried out in polynomial time; see, e.g., Hazan (2016). Moreover, since the projection problem is convex, it admits efficient iterative solvers that achieve

arbitrary accuracy. Stronger guarantees are also available: there exist strongly polynomial-time algorithms for this projection problem; see Gupta et al. (2016, Theorem 3).

In our setting, the projection can often be performed in overall linear time by first normalizing $\boldsymbol{y}$ as $\boldsymbol{y}' = \boldsymbol{y}/\|\boldsymbol{y}\|_1$ and checking whether $\boldsymbol{y}' \in \triangle_{n,\gamma}$, i.e., whether $y'_a \geq \gamma$ for all $a \in \mathcal{A}$. If this condition holds, then $\boldsymbol{y}'$ already equals the projected point and no further computation is required.

G Comparing Variation Measures

G.1 Incomparability of P_T and V_T

In this section, we adapt an illustrative construction from Jadbabaie et al. (2015) to show that V_T and P_T are not, in general, comparable quantities, even when all losses lie in $[0, 1]$.

We first give an example in which $V_T \ll P_T$. For each round $t \geq 1$, define the loss vector $\boldsymbol{f}_t \in [0, 1]^n$ by $\boldsymbol{f}_t = (0, \frac{1}{T}, 1, \ldots, 1)$ if t is even, and $\boldsymbol{f}_t = (\frac{1}{T}, 0, 1, \ldots, 1)$ if t is odd. Let $\boldsymbol{x}_t^\star \in \arg\min_{x \in \triangle_n} \langle \boldsymbol{f}_t, x \rangle$ be an optimal comparator at round t. Since the identity of the minimizer alternates between the first two vertices of the simplex, the path-length satisfies $P_T = \Omega(T)$, whereas the per-round change in the loss vectors is only on the order of $1/T$, implying $V_T = \mathcal{O}(1)$.

Conversely, consider a two-action instance ($n = 2$) with losses in $[0, 1]^2$ given by $\boldsymbol{f}_t = (0, 1)$ if t is even and $\boldsymbol{f}_t = (0, \frac{1}{2})$ if t is odd. Here, action 1 is optimal at every round, so one may choose $\boldsymbol{x}_t^\star$ constant over time and obtain $P_T = \mathcal{O}(1)$, while the loss sequence changes by a constant amount each round, yielding $V_T = \Omega(T)$.

Together, these examples show that neither V_T nor P_T admits a general dominance relationship over the other.