Online Primal-Dual Mirror Descent under Stochastic Constraints

XIAOHAN WEI, Facebook, USA
HAO YU, Amazon, USA
MICHAEL J. NEELY, University of Southern California, USA

We consider online convex optimization with stochastic constraints where the objective functions are arbitrarily time-varying and the constraint functions are independent and identically distributed (i.i.d.) over time. Both the objective and constraint functions are revealed after the decision is made at each time slot. The best known expected regret for solving such a problem is $O(\sqrt{T})$, with a coefficient that is polynomial in the dimension of the decision variable and relies on the *Slater condition* (i.e. the existence of interior point assumption), which is restrictive and in particular precludes treating equality constraints. In this paper, we show that such Slater condition is in fact not needed. We propose a new *primal-dual mirror descent* algorithm and show that one can attain $O(\sqrt{T})$ regret and constraint violation under a much weaker Lagrange multiplier assumption, allowing general equality constraints and significantly relaxing the previous Slater conditions. Along the way, for the case where decisions are contained in a probability simplex, we reduce the coefficient to have only a logarithmic dependence on the decision variable dimension. Such a dependence has long been known in the literature on mirror descent but seems new in this new constrained online learning scenario. Simulation experiments on a data center server provision problem with real electricity price traces further demonstrate the performance of our proposed algorithm.

Additional Key Words and Phrases: Stochastic programming, Constrained programming, Online learning

ACM Reference Format:
Xiaohan Wei, Hao Yu, and Michael J. Neely. 2020. Online Primal-Dual Mirror Descent under Stochastic Constraints. *Proc. ACM Meas. Anal. Comput. Syst.* 4, 2, Article 50 (June 2020), 36 pages. https://doi.org/10.1145/3392157

1 INTRODUCTION

We consider an online convex optimization (OCO) problem with a sequence of arbitrarily varying convex objective functions $f^t(\mu),\ t = 0, 1, 2, \cdots,\ \mu \in \Delta \subseteq \mathbb{R}^d$ which are revealed per slot after the decision is made, and Δ is a closed bounded convex set. For a fixed time horizon T, define the regret of a sequence of decisions $\{\mu^0, \mu^1, \cdots, \mu^{T-1}\} \subseteq \Delta$ as

$$\sum_{t=0}^{T-1} f^t(\mu^t) - \min_{\mu\in\Delta} \sum_{t=0}^{T-1} f^t(\mu).$$

The goal of OCO is to choose the decision sequence so that the regret grows sublinearly with respect to T. OCO is a classical problem and has been considered in a number of previous works

Authors' addresses: Xiaohan Wei, Facebook, 1 Hacker Way, Menlo Park, CA, 94025, USA, xiaohanw@usc.edu; Hao Yu, Amazon, Houdini South,300 Boren Ave N, Seattle, WA, 98109, USA, yuhao@usc.edu; Michael J. Neely, University of Southern California, 3740 McClintock Ave. Los Angeles, CA, 90089, USA, mikejneely@gmail.com.

2476-1249/2020/6-ART50 $15.00
https://doi.org/10.1145/3392157

such as [4, 9, 10, 31]. In particular, it is known that for differentiable functions $f^t(\cdot)$, the projected gradient descent algorithm achieves an $O(\sqrt{T})$ regret which is also worst case optimal. When the set Δ is a probability simplex, the mirror descent algorithm further achieves an "almost dimension free" logarithmic dependency on the dimension d.

The framework considered in this paper builds upon the previous OCO model by incorporating a sequence of time varying constraint functions $g_i^t(\mu),\ i = 1, 2, \cdots, L$, which are also revealed at each time slot t after the decision is made. The goal of this constrained OCO is to choose the decision sequence $\left\{\mu^0, \mu^1, \cdots, \mu^{T-1}\right\} \subseteq \Delta$ so that both the regret and constraint violations grow sublinearly in T (i.e. $\sum_{t=0}^{T-1} g_i^t(\mu_t) \leq o(T)$) with respect to the best fixed decision in hindsight solving the following convex program:

$$\min_{\mu\in\Delta} \sum_{t=0}^{T-1} f^t(\mu),\ s.t.\ \sum_{t=0}^{T-1} g_i^t(\mu) \leq 0,\ i = 1, 2, \cdots, L. \tag{1}$$

The constrained OCO was first considered in the work [14] where the authors (somewhat surprisingly) show via a counterexample that even with only one constraint, it is not always possible to achieve the aforementioned goal if we allow both objective and constraint functions to vary arbitrarily. Such an impossibility result implies that if one wants to obtain meaningful results on constrained OCO, then more assumptions have to be posed.

The works [11, 13, 19] consider the scenario where the constraint functions are fixed (i.e. do not depend on the time index t) and propose primal-dual type methods whose analyses give $O(T^{\max\{\beta,1-\beta\}})$ regret and $O(T^{1-\beta/2})$ constraint violation, where $\beta \in [0, 1]$ is an algorithm parameter. This bound is improved in the work [27] where the authors show an $O(\sqrt{T})$ regret bound and finite constraint violations (i.e. $O(1)$ constraint violation) via Slater condition (i.e. There exists a $\mu \in \Delta$ such that $g_i(\mu) < 0,\ \forall i$). A more recent work [29] shows that one can get logarithm regret and $O(\sqrt{T})$ constraint violations if one assumes instead that all objective functions are strongly convex.

Constrained OCO with stochastic constraints, where $g_i^t(\mu) = g_i(\mu, \gamma^t)$ and $\{\gamma^t\}_{t=0}^{T-1}$ are i.i.d., is considered in the works such as [5, 12, 26], where a primal-dual proximal gradient algorithm is proposed and $O(\sqrt{T})$ expected regret and constraint violations are shown under the Slater condition (i.e. there exists a $\mu \in \Delta$ such that $\mathbb{E}\big(g_i(\mu, \omega^t)\big) < 0,\ \forall i$). Without Slater condition, the best known result is again $O(T^{\max\{\beta,1-\beta\}})$ regret and $O(T^{1-\beta/2})$ constraint violation as is shown in [25]. Also, to the best of our knowledge, previous bounds in constrained online learning fail to recover the "almost dimension free" phenomenon for the probability simplex decision set ubiquitous in unconstrained scenarios. In this paper, we make steps towards *removing the Slater condition while maintaining the worst case optimal $O(\sqrt{T})$ regret, constraint violations, and sharpening the dimension dependency on decision variables.*

Slater condition is assumed in the classical analysis of optimization algorithms for constrained convex programs such as the dual subgradient algorithm [15] and the interior point method [3]. A key implication of Slater condition, which is adopted in the $O(1/\sqrt{T})$ convergence rate analysis in [15], is that it implies the existence and boundedness of Lagrange multipliers. However, the reverse implication is in general untrue, as one can show that for many equality constrained convex programs, Lagrange multipliers do exist and are bounded [2]. This makes "Slater condition free" analysis an important topic in optimization theory and motivates series of improved primal-dual type algorithms and analysis for constrained convex programs with competitive convergence rate under the existence of Lagrange multipliers assumption [6, 16, 28, 30].

Replacing the Slater condition with Lagrangian type assumptions in online problems is highly non-trivial and does not follow from that of constrained convex programs. A key issue is that the objective function varies arbitrarily per slot, and so the definition of Lagrange multiplier is not

clear. A simple attempt is to look at in-hindsight problems such as (1) and see if the Lagrange multiplier of this problem helps with the regret analysis. However, since problem (1) sums the objectives across the horizon, it hardly gives any insight on the per slot dynamics for any practical algorithm considered. If we instead look at the per slot constrained problem, then, one might be able to conduct analysis and obtain per-slot multipliers, but it is not clear how to piece together the analysis for different slots.

1.1 Contributions

In this paper, we consider the stochastic constrained online learning problem and propose a new primal-dual online mirror descent framework, which simultaneously weakens the assumptions and improves the dimension factors in the previously known online proximal gradient type algorithms. We introduce a new *sequential existence of Lagrange multipliers* condition, which is shown to be *strictly weaker* than the Slater condition, allows for equality constraints and bridges the aforementioned dilemma between on-hindsight problem and per slot problem. We then show via a new analysis that under such an assumption, the proposed algorithm enjoys a matching $O(\sqrt{T})$ expected regret and constraint violations. For the case when decisions are contained in a probability simplex, we reduce the dimension dependency to have only a logarithmic factor. Conceptually, our analysis seems to be distinctive from the previous known methods in the sense that we look at the cumulative objectives over a specifically chosen time period (of length $\sqrt{T}$), and consider the following static constrained program starting from any time slot t: $\min_{\mu\in\Delta} \sum_{\tau=t}^{t+\sqrt{T}} \mathbb{E}(f^\tau(\mu)),\ s.t.\ \mathbb{E}\big(g_i(\mu,\omega^t)\big) \le 0,\ i = 1, 2, \cdots, L$. We demonstrate that the existence and boundedness of Lagrange multipliers for this problem provides certain weak error bound conditions for the dual function sufficient to bound the size of the dual variable process, leading to the desired results.

1.2 Notation

For any vector $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{v} \ge 0,\ \mathbf{v} = 0,\ \mathbf{v} \le 0$ means $\mathbf{v}$ is entrywise nonnegative, zero and nonpositive, respectively. The notation $[\mathbf{v}]_+$ denotes entrywise application of the function $\max(x, 0)$. The notation $\mathbb{R}^d_+$ stands for the positive orthant of $\mathbb{R}^d$. For any set $\mathcal{S} \subseteq \mathbb{R}^d$, let $\text{int}(\mathcal{S})$ be its interior. The norms $\|\mathbf{v}\|_1 := \sum_{i=1}^d |v(i)|$, $\|\mathbf{v}\|_2 := (\sum_{i=1}^d |v(i)|^2)^{1/2}$ and $\|\mathbf{v}\|_\infty := \max_i |v(i)|$. For any convex function $f : \mathbb{R}^d \to \mathbb{R}$, we use $\nabla f(\mathbf{v})$ to denote any one of the subgradients at $\mathbf{v}$ and use $\partial f(\mathbf{v})$ to denote the set of all subgradients at $\mathbf{v}$. For any function $g(\mathbf{v}, \xi)$ which is convex on the first argument $\mathbf{v}$, $\nabla g(\mathbf{v}, \xi)$ denotes the subgradient of g on $\mathbf{v}$ while fixing ξ. For any closed set $K \subseteq \mathbb{R}^d$ and any point $\mathbf{x} \in \mathbb{R}^d$, the distance of $\mathbf{x}$ to K is defined as $\text{dist}(\mathbf{x}, K) := \min_{\mathbf{y}\in K} \|\mathbf{x} - \mathbf{y}\|_2$.

2 PROBLEM FORMULATION AND ALGORITHMS

2.1 Basic definitions

Let $\|\cdot\|$ be a general norm in $\mathbb{R}^d$. Define the dual norm on any $x \in \mathbb{R}^d$ as $\|x\|_* := \sup_{\|y\|\le 1} \langle x, y\rangle$, where $\langle x, y\rangle = \sum_{i=1^d} x(i)y(i)$. Consider a convex set $C \subseteq \mathbb{R}^d$ (potentially being $\mathbb{R}^d$ itself) with a non-empty interior, i.e. $\text{int}(C) \neq \emptyset$. Let $\omega : C \to \mathbb{R}$ be a function that is continuously differentiable in the interior of C. Let $\Delta \subseteq C$ be a *compact convex* subset containing the origin and $\Delta^o := \Delta \cap \text{int}(C)$, which is non-empty. Define the *Bregman divergence* function $D : \Delta \times \Delta^o \to \mathbb{R}$ generated from $\omega(\cdot)$ as follows:

$$D(x, y) := \omega(x) - \omega(y) - \langle \nabla\omega(y), x - y\rangle .$$

The following is a key property of the Bregman divergence:

Lemma 2.1 (Pushback). *Let $f : C \to \mathbb{R}$ be a continuous convex function. Fix $\alpha > 0$, $y \in \Delta^o$. Suppose $x^* \in argmin_{x\in\Delta} f(x) + \alpha D(x, y)$ and $x^* \in \Delta^o$, then, for any $z \in \Delta$,*

$$f(x^*) + \alpha D(x^*, y) \leq f(z) + \alpha D(z, y) - \alpha D(z, x^*).$$

Remark 2.1. *For the case where f is a linear function and ω is convex, such a pushback result can be found, for example, in [17]. For results with f being on domain $\mathbb{R}^d$, the proof can be found in [20]. Our result generalizes previous results to arbitrary set Δ. It is proved in the Supplement (Section A.1)*

We say $\omega(\cdot)$ is a *distance generating function* if for any $x \in \text{int}(C)$, $\omega(\cdot)$ is a continuously differentiable and strongly convex with modulus β with respect to the primal norm $\|\cdot\|$, i.e. $\langle x - y, \nabla\omega(x) - \nabla\omega(y)\rangle \geq \beta\|x - y\|^2,\ \forall x, y \in \text{int}(C)$. It is easy to see if ω is a distance generating function, then, the corresponding $D(\cdot, \cdot)$ satisfies

$$D(x, y) \geq \frac{\beta}{2}\|x - y\|^2,\ \forall x, y \in \text{int}(C). \tag{2}$$

Note that $D(x, y)$ behaves asymmetrically on x and y over potentially different domains, which results from the (possible) non-differentiability of the distance generating function $\omega(\cdot)$ on the boundary of Δ. We provide two examples below:

(1) The set $\Delta = \{\mu \in \mathbb{R}^d :\ \|\mu\|_1 = 1,\ \mu \geq 0\}$ is a probability simplex, $C = \mathbb{R}^d_+$ with ℓ_1-norm $\|\cdot\|_1$, the function $\omega(\mu) = -\sum_{i=1}^d \mu(i)\log\mu(i)$ is the entropy function, and for any two distributions $\mu^a \in \Delta,\ \mu^b \in \Delta^o$,

$$D(\mu^a, \mu^b) = \sum_{i=1}^d \mu^a(i)\log\frac{\mu^a(i)}{\mu^b(i)}$$

is the well-known Kullback-Leibler (KL) divergence. Furthermore, by Pinsker's inequality, it is strongly convex with respect to $\|\cdot\|_1$ with the strongly convex modulus $\beta = 1$. The dual norm in this space is $\|\cdot\|_\infty$.

(2) The set Δ is in the Euclidean space $\mathbb{R}^d, C = \mathbb{R}^d$ with the usual ℓ_2-norm $\|\cdot\|_2$ and $\omega(x) = \frac{1}{2}\|x\|_2^2$, which is strongly convex with respect to $\|\cdot\|_2$, $D(x, y) = \|x - y\|_2^2$, and the dual norm is also $\|\cdot\|_2$.

2.2 Problem formulation

In this section, we set up the basic formulation of stochastic constrained online optimization. Let $\{\xi^t\}_{t=0}^\infty$ and $\{\gamma^t\}_{t=0}^\infty$ be two processes, where $\{\xi^t\}_{t=0}^\infty$ can be arbitrarily time varying (might be chosen based on the system history) and $\{\gamma^t\}_{t=0}^\infty$ are i.i.d. realizations of a random variable γ with a possibly unknown distribution. Let $f(\mu, \xi^t), g_i(\mu, \gamma^t), i \in \{1, 2, \ldots, L\}$ be deterministic functions which are continuous convex in the first component given the second component. Furthermore, let $\{h_j^t\}_{t=0}^\infty,\ j \in \{1, 2, \cdots, M\}$ be sequences of i.i.d. random vectors in $\mathbb{R}^d$. Throughout the paper, we assume ξ^t, γ^t, h_j^t are mutually independent for all t with system history up to time t as $\mathcal{F}_t := \{\xi^\tau, \gamma^\tau, h_j^\tau\}_{\tau=0}^{t-1}$. For any fixed $\mu \in \Delta$, we write $f^t(\mu) := f(\mu, \xi^t)$, $g_i^t(\mu) := g_i(\mu, \gamma^t)$, and $\overline{f}^t(\mu) = \mathbb{E}\big(f^t(\mu)|\mathcal{F}_t\big), \overline{g}_i(\mu) = \mathbb{E}\big(g_i^t(\mu)\big)$. We further define the vectorized notations

$$\begin{aligned}
\mathbf{g}^t(\mu) &= [g_1(\mu, \gamma^t), \ldots, g_L(\mu, \gamma^t)]^T \\
\overline{\mathbf{g}}(\mu) &= [\mathbb{E}(g_1(\mu, \gamma^t)), \ldots, \mathbb{E}(g_L(\mu, \gamma^t))]^T \\
\mathbf{h}^t(\mu) &= [\langle h_1^t, \mu\rangle, \ldots, \langle h_M^t, \mu\rangle]^T \\
\overline{\mathbf{h}}(\mu) &= [\langle \mathbb{E}(h_1^t), \mu\rangle, \ldots, \langle \mathbb{E}(h_M^t), \mu\rangle]^T.
\end{aligned}$$

It is also worth noting that our algorithms and analysis also apply to the special case where $\{\xi^t\}_{t=0}^{\infty}$ are also i.i.d. for which we have $\overline{f}^t(\mu) = \mathbb{E}(f(\mu, \xi^t)) := \overline{f}(\mu),\ \forall t$.

Define the benchmarking decision in-hindsight μ^* as a solution to the following static convex program:

$$\min_{\mu\in\Delta} \sum_{t=0}^{T-1} \overline{f}^t(\mu)\ s.t.\ \overline{\mathbf{g}}(\mu) \le 0,\ \overline{\mathbf{h}}(\mu) = \mathbf{b}, \tag{3}$$

where $\mathbf{b} = [b_1,\ b_2,\ \cdots,\ b_M]^T$ is a vector of constants. At the beginning of each time slot t, none of the objective function $f^t(\mu)$, constraint function $g_i^t(\mu)$ or random vector h_j^t is known. The decision maker is supposed to choose a vector $\mu^t \in \Delta$ first before observing these quantities. The goal is to make sequential (possibly randomized) decisions so that both the expected regret, defined as $\sum_{t=0}^{T-1} \mathbb{E}\big(f^t(\mu^t) - f^t(\mu^*)\big)$, and expected constraint violations, define as $\sum_{t=0}^{T-1} \mathbb{E}(g_i^t(\mu^t))$ and $\mathbb{E}|\sum_{t=0}^{T-1} h_j^t(\mu^t)|$, grow sublinearly with respect to the time horzon T. Throughout this paper, we make the following boundedness assumption:

Assumption 2.1 (Boundedness of objectives and constraint functions).

(1) *Objective functions $f^t(\mu)$ and constraint functions $g_i^t(\mu)$ have bounded subgradients on Δ, i.e. there exist constants $D_1 > 0$ and $D_2 > 0$ such that $\|\nabla f^t(\mu)\|_* \le D_1$, $\sum_{i=1}^{L} \|\nabla g_i^t(\mu)\|_*^2 \le D_2^2$, for all $\mu \in \Delta$, all $t \in \{0, 1, \ldots\}$, and all $i \in \{1, 2, \ldots, L\}$.*
(2) *There exist constants $F, G, H > 0$ such that $|f^t(\mu)| \le F,\ \forall t \in \{0, 1, 2, \cdots\}$, $\sum_{i=1}^{L} |g_i^t(\mu)|^2 \le G^2$ for all $\mu \in \Delta,\ t \in \{0, 1, 2, \cdots\}$, and $\sum_{j=1}^{M} \|h_j^t\|_*^2 \le H^2$, for all $j \in \{1, 2, \cdots, M\},\ t \in \{0, 1, \ldots\}$.*
(3) *The Bregman divergence $D(\cdot, \cdot)$ is generated from a distance generating $\omega(\cdot)$ and bounded on the set Δ, i.e. there exists a constant R such that $\sup_{x\in\Delta, y\in\Delta^o} D(x, y) \le R$.*

By strong convexity of the Bregman divergence (2), we have

$$\sup_{x\in\Delta, y\in\Delta^o} \|x - y\|^2 \le \frac{2R}{\beta}.$$

Note further that KL divergence does not satisfy Assumption 2.1(3), for which we will develop a separate new algorithm in Section 3.2.

2.3 Primal-dual online mirror descent

We are now in a position to introduce our new online mirror descent (Algorithm 1) for the stochastic constrained online learning. The algorithm computes the next decision μ^{t+1} by a proximal mirror map using μ^t, f^t and g_i^t, and control the constraint violations via dual multipliers $\mathbf{Q}(t)$ and $\mathbf{H}(t)$.

2.4 Sequential Existence of Lagrange Multipliers (SELM)

In this section, we introduce our Lagrange multiplier condition. A detailed comparison between such a condition and other constraint qualification conditions is delayed to the Supplementary (Section A.2). We start by defining a partial average function starting from any time slot t as:

$$\overline{f}^{t,k} := \frac{1}{k} \sum_{i=0}^{k-1} \overline{f}^{t+i}.$$

Consider the following optimization problem:

$$\min_{\mu\in\Delta} \overline{f}^{t,k}(\mu)\ s.t.\ \overline{\mathbf{g}}(\mu) \le 0,\ \overline{\mathbf{h}}(\mu) = \mathbf{b}, \tag{7}$$

ALGORITHM 1: : Let $\mu^0 = \mu^{-1} \in \Delta$. Let $V, \alpha > 0$ be some trade-off parameters. Let $Q_i(t),\ H_j(t)$ be sequences of dual multipliers such that $Q_i(0) = 0,\ H_j(0) = 0,\ \forall i, j$. For each slot $t \in \{0, 1, \cdots, T-1\}$:

- Choose μ^t as a solution to the following problem:

$$\min_{\mu\in\Delta} \left\langle V\nabla f^{t-1}(\mu^{t-1}) + \sum_{i=1}^{L} Q_i(t)\nabla g_i^{t-1}(\mu^{t-1}) + \sum_{j=1}^{M} H_i(t)h_i^{t-1}, \mu \right\rangle + \alpha D(\mu, \mu^{t-1}) \tag{4}$$

- Update dual multiplier $Q_i(t), H_j(t),\ i \in \{1, 2, \cdots, L\},\ j \in \{1, 2, \cdots, M\}$ via

$$Q_i(t+1) = \max\left\{Q_i(t) + \tilde{g}_i^{t-1}(\mu^t), 0\right\} \tag{5}$$

$$H_j(t+1) = H_j(t) + \left\langle h_j^{t-1}, \mu^t \right\rangle - b_j, \tag{6}$$

  where $\tilde{g}_i^t(\mu^t) := g_i^{t-1}(\mu^{t-1}) + \left\langle \nabla g_i^{t-1}(\mu^{t-1}), \mu^t - \mu^{t-1} \right\rangle$.
- Observe the objective function f^t and constraint functions $\{g_i^t\}_{i=1}^L,\ \{h_j^t\}_{j=1}^M$.

where $\overline{\mathbf{g}}(\mu),\ \overline{\mathbf{h}}(\mu)$ are defined in Section 2.2. Denote the solution to this program as $\overline{f}_*^{t,k}$. Define the Lagrangian dual function of (7) as

$$q^{(t,k)}(\lambda, \eta) := \min_{\mu\in\Delta} \overline{f}^{t,k}(\mu) + \sum_{i=1}^{L} \lambda_i \overline{g}_i(\mu) + \sum_{j=1}^{M} \eta_j(\overline{h}_j(\mu) - b_j), \tag{8}$$

where $\lambda \in \mathbb{R}_+^L$ and $\eta \in \mathbb{R}^M$ are dual variables. For simplicity of notation, we always enforce them to be row vectors. Now, we are ready to state our condition:

Assumption 2.2 (Sequential existence of Lagrange multipliers (SELM)). *For any time slot t and any time period $k \geq \sqrt{T}$, the set of primal optimal solution to (7) is non-empty. Also, the set of dual optimal solution, which is the set of Lagrange multipliers of (7) denoted as $\mathcal{V}_{t,k}^* := \text{argmax}_{\lambda\in\mathbb{R}_+^L,\ \eta\in\mathbb{R}^M} q^{(t,k)}(\lambda, \eta)$, is non-empty and bounded. Furthermore, let $B > 0$ be a constant such that for any $t \in \{0, 1, \cdots, T-1\}$ and $k = \sqrt{T}$, the dual optimal set $\mathcal{V}_{t,k}^*$ defined above satisfies $\max_{[\lambda,\mu]\in\mathcal{V}_{t,k}^*} \|[\lambda, \mu]\|_2 \leq B$.*

Remark 2.2. *SELM asserts the existence and boundedness of Lagrange multipliers on the set of subproblems (7) for any time epoch $t \in \{0, 1, 2, \cdots, T-1\}$ and any time duration $k \geq \sqrt{T}$. In the special case where the objectives are also i.i.d. functions, we have*

$$\overline{f}^{t,k}(\mu) = \frac{1}{k}\sum_{i=0}^{k-1} \overline{f}^{t+i}(\mu) = \frac{1}{k}\sum_{i=0}^{k-1} \mathbb{E}\big(f(\mu, \xi^{t+i})\big) := \overline{f}(\mu),\ \forall\ t, k$$

and SELM reduces to an existence and boundedness of Lagrange multipliers condition for a single constrained convex program:

$$\min_{\mu\in\Delta} \overline{f}(\mu)\ s.t.\ \overline{\mathbf{g}}(\mu) \leq 0,\ \overline{\mathbf{h}}(\mu) = \mathbf{b}.$$

Remark 2.3. *In Section A.2 of the Supplement, we show that SELM is implies by certain constraint qualification conditions and strictly weaker than the Slater conditions. In particular, we obtain the following simplifications in special cases:*

(1) *Lemma A.2 shows that Slater condition implies SELM.*
(2) *Corollary A.10 shows that when the interior of Δ is non-empty and there are only equality constraints, the linear independence of $\{\mathbb{E}(h_1^t),\ \mathbb{E}(h_2^t),\ \cdots,\ \mathbb{E}(h_M^t)\}$ implies SELM.*

(3) *Lemma A.5 shows that when Δ is the probability simplex and there are only equality constraints, the linear independence of $\{\mathbf{1},\ \mathbb{E}(h_1^t),\ \mathbb{E}(h_2^t),\ \cdots,\ \mathbb{E}(h_M^t)\}$ implies SELM.*

Detailed arguments are deferred to Section A.2 of the Supplement.

The motivation for SELM is as follows: whenever Lagrange multipliers exist and are bounded, we have the dual function deviates according to a certain curve related to the distance from the set of Lagrange multipliers, namely, the weak error bound condition (EBC).

Definition 2.2 (Weak error bound condition (EBC)). Let $F(\mathbf{x})$ be a concave function over $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X}$ is closed and convex. Suppose $\Lambda^* := \text{argmax}_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x})$ is non-empty. The function $F(\mathbf{x})$ satisfies the weak EBC if there exists constants $\ell_0,\ c_0 > 0$ such that for any $\mathbf{x} \in \mathcal{X}$ satisfying $\text{dist}(\mathbf{x}, \Lambda^*) \geq \ell_0$,

$$F(\mathbf{x}^*) - F(\mathbf{x}) \geq c_0 \cdot \text{dist}(\mathbf{x}, \Lambda^*),$$

where $\text{dist}(\mathbf{x}, \Lambda^*)$ is defined as:

$$\text{dist}(\mathbf{x}, \Lambda^*) = \inf_{\mathbf{y}\in\Lambda^*} \|\mathbf{x} - \mathbf{y}\|_2$$

Note that in Definition 2.2, Λ^* is a closed convex set. This follows from the fact that $F(\mathbf{x})$ is a convex function and thus all sub level sets are closed and convex. The following lemma shows SELM implies weak EBC on the dual function:

Lemma 2.3. *Fix $T \geq 1$. Suppose Assumption 2.2 holds, then for any $t \in \{0, 1, \cdots, T-1\}$ and $k = \sqrt{T}$, there exists constants $c_0, \ell_0 > 0$, such that the dual function $-q^{(t,k)}(\lambda, \eta)$ defined in (8) satisfies the weak EBC with parameter c_0, ℓ_0.*

This lemma is restated as Lemma A.13 with more explicit expressions on $c_0,\ \ell_0$ and the proof is in Supplement A.5. In the Supplement (Section A.2.3), we also compare this weak EBC with the classical EBC in optimization theory and show that classical EBC implies weak EBC with explicit constants.

3 MAIN RESULTS

3.1 Sets with bounded Bregman divergence

In this section, we present our main performance guarantee on Algorithm 1, when Assumption 2.1 and 2.2 hold under the general norm $\|\cdot\|$ setup in $\mathbb{R}^d$ as we described in Section 2.1. In particular, we assume that Assumption 2.1(3), i.e. the Bregman divergence is bounded, holds, which will be relaxed in Section 3.2.

Theorem 3.1. *Let μ^* be a solution to the in-hindsight optimization problem (3). Suppose Assumption 2.1 and 2.2 hold. Let $\bar{c},\ \bar{\ell} > 0$ be absolute constants such that $c_0 \geq \bar{c}$ and $\ell_0 \leq \bar{\ell}$ for all $c_0,\ \ell_0$ obtained in Lemma 2.3 over $t = 0, 1, 2, \cdots, T-1$ and $k = \sqrt{T}$. If we choose $\alpha = T, V = \sqrt{T}$ in Algorithm 1, then the expected regret and constraint violations satisfy:*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big(f^t(\mu^t) - f^t(\mu^*)\big) \leq \frac{C_0'}{\sqrt{T}},$$

$$\mathbb{E}\left\|\left[\frac{1}{T}\sum_{t=0}^{T-1} \overline{\mathbf{g}}(\mu^t)\right]_+\right\|_2 \leq \frac{C_1'}{\sqrt{T}},$$

$$\mathbb{E}\left\|\frac{1}{T}\sum_{t=0}^{T-1} \overline{\mathbf{h}}(\mu^t) - \mathbf{b}\right\|_2 \leq \frac{C_2'}{\sqrt{T}},$$

where C_0', C_1', C_2' are constants depending linearly on $D_1^2 + D_1 + D_2^2 + G^2 + H^2 + G + H + F$ and independent of T.

Note that throughout the paper, we always use Euclidean ℓ_2-norm $\|\cdot\|_2$ to measure the constraint violation, and it is irrelevant to what norm we choose on the primal space $C \subseteq \mathbb{R}^d$.

3.2 The probability simplex case

In this section, we deal with the probability simplex case where the decision set Δ is a d-dimensional probability simplex with huge d. While Algorithm 1 can be applied to solve such problems by choosing $D(\mu, \mu^{t-1})$ to be $\|\mu - \mu^{t-1}\|_2^2$, due to the dependencies on the D_1, D_2, G, H, F, the constant factors in Theorem 3.1 linearly depend on d. For mirror descent over a probability simplex, to improve the dimension dependence, people usually choose the Bregman divergence distance $D(\cdot,\cdot)$ to be the KL divergence. However, KL divergence fundamentally violates the third assumption in Assumption 2.1. We now present an alternative algorithm in Algorithm 2 and shows that it can achieve sublinear regret and constraint violations that logarithmically depends on d .

ALGORITHM 2: : Let $V, \alpha > 0$, $\theta \in [0,1)$ be some trade-off parameters. Let $D(\mu_1, \mu_2) = \sum_{i=1}^d \mu_1(i) \log \frac{\mu_1(i)}{\mu_2(i)}$. Let $Q_i(t)$, $H_j(t)$ be sequences of dual multipliers such that $Q_i(0) = 0$, $H_j(0) = 0$, $\forall i, j$. Let $\mu_0 = \mu_{-1} = \frac{1}{d}\mathbf{1}$. For any slot $t \in \{0, 1, \cdots, T-1\}$:

- Let $\tilde{\mu}^{t-1} = (1-\theta)\mu^{t-1} + \frac{\theta}{d}\mathbf{1}$.
- Choose μ^t as a solution to the following problem:

$$\min_{\mu \in \Delta} \left\langle V\nabla f^{t-1}(\mu^{t-1}) + \sum_{i=1}^L Q_i(t)\nabla g_i^{t-1}(\mu^{t-1}) + \sum_{j=1}^M H_i(t) h_i^{t-1}, \mu \right\rangle + \alpha D(\mu, \tilde{\mu}^{t-1}) \tag{9}$$

- Update each dual multiplier $Q_i(t), H_j(t)$ via (5) and (6).
- Observe the objective function f^t and constraint functions $\{g_i^t\}_{i=1}^L$, $\{h_j^t\}_{j=1}^M$.

Compared to Algorithm 1, Algorithm 2 uses the K-L divergence as the particular Bregman divergence and introduces a probability mixing step $\tilde{\mu}^{t-1} = (1-\theta)\mu^{t-1} + \frac{\theta}{d}\mathbf{1}$, which pushes the update away from the boundary, at each round. Furthermore, it is known that the problem (9) admits a closed form solution known as the exponential gradient update [10]. More specifically, define

$$\mathbf{p}^{t-1} := \alpha^{-1}\big(V\nabla f^{t-1}(\mu^{t-1}) + \sum_{i=1}^L Q_i(t)\nabla g_i^{t-1}(\mu^{t-1}) + \sum_{j=1}^M H_i(t) h_i^{t-1}\big).$$

Then, the update μ^t can simply be written as

$$\mu^t(i) = \frac{\tilde{\mu}^{t-1}(i)\exp(-p^{t-1}(i))}{\sum_{k=1}^d \tilde{\mu}^{t-1}(k)\exp(-p^{t-1}(k))}, \ i \in \{1, 2, \cdots, d\}. \tag{10}$$

We have the following performance bound on this algorithm whose proof is similar to Theorem 3.1 and delayed to the Supplement (Section A.4):

Theorem 3.2. *Suppose the first two in Assumption 2.1 (using $\|\cdot\| = \|\cdot\|_1$ and $\|\cdot\|_* = \|\cdot\|_\infty$) and Assumption 2.2 hold. Let $\bar{c}$, $\bar{\ell} > 0$ be absolute constants such that $c_0 \geq \bar{c}$ and $\ell_0 \leq \bar{\ell}$ for all c_0, ℓ_0 obtained in Lemma 2.3 over $t = 0, 1, 2, \cdots, T-1$ and $k = \sqrt{T}$. Choose $\alpha = T$, $V = \sqrt{T}$, $\theta = 1/T$ in*

Algorithm 2. The expected regret and constraint violations satisfy:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big(\overline{f}^t(\mu^t)-f_t(\mu^*)\Big)\le\frac{\hat{C}_0'}{\sqrt{T}}+\frac{\hat{C}_0'\log(d)}{\sqrt{T}}$$

$$\mathbb{E}\left\|\left[\frac{1}{T}\sum_{t=0}^{T-1}\overline{\mathbf{g}}(\mu^t)\right]_+\right\|_2\le\frac{\hat{C}_1'}{\sqrt{T}}+\frac{\hat{C}_1''\log(Td)}{\sqrt{T}},$$

$$\mathbb{E}\left\|\frac{1}{T}\sum_{t=0}^{T-1}\overline{\mathbf{h}}(\mu^t)-\mathbf{b}\right\|_2\le\frac{\hat{C}_2'}{\sqrt{T}}+\frac{\hat{C}_2''\log(Td)}{\sqrt{T}}.$$

where $\hat{C}_0', \hat{C}_1', \hat{C}_1'', \hat{C}_2', \hat{C}_2''$ are absolute constants depending linearly on $D_1^2+D_1+D_2^2+G^2+H^2+G+H+F$ and independent of d or T. (Note that D_1, D_2, G, H, F in Assumption 2.1 are independent of d when $\|\cdot\|_*=\|\cdot\|_\infty$.)

Remark 3.1. *As a comparison, previously known algorithms and performance bounds, when applying to this problem, yield worse dependencies on dimension d or time period T. For example, when assuming Slater condition, Theorem 1 of [26] gives $C\text{poly}(d)/\sqrt{T}$ regret bound and constraint violations. Without Slater condition, [12] shows $C_1\text{poly}(d)/\sqrt{T}$ regret bound and $C_2\text{poly}(d)/T^{1/4}$ constraint violation. Here poly(d) stands for polynomial dependency on d and C, C_1, C_2 are all absolute constants independent of d or T.*

4 SIMULATION EXPERIMENTS

We consider the problem of cost minimization under budget pacing constraints in data center service scheduling. More specifically, consider a geographically distributed data center consists of 5 server clusters serving one stream of incoming jobs arriving at a central controller. Each cluster contains 10 servers. The jobs are directed to different clusters for processing by controller with different per unit electricity costs. In the simulation, we use electricity market price (EMP) data traces from 5 zones of New York ISO open access pricing data (http://www.nyiso.com/). For example, Fig 1(a) depicts the per 5 min EMP data of zone DUNWOD between 05/01/2017 and 05/10/2017. The number of incoming jobs per 5 min is $\lambda(t)$, which is assumed to be poisson distributed with mean equals 1000. each server k can choose a power allocation option $\mu_k^t\in[0,30]$. This option determines the following over the 5 min slot:

(1) The electricity money spend of server k: $f_k^t(\mu_k^t)=c_k^t\cdot\mu_k^t$, where c_k^t is the per unit EMP of zone server k belongs to.
(2) The number of jobs served $g_k^t(\mu_k^t)$ which follows a Pareto distribution (a.k.a. power law, see [7]) of mean $8\log(1+4\mu_k^t)$.
(3) Internal budget consumptions $h_k^t\cdot\mu_k^t$, where h_k^t follows a Pareto distribution of mean 5 units.

In a typical online service system such as ads service, budget is a measure of internal resources [1]. The goal is to minimize total average electricity cost over $T=10000$ slots, i.e. $\sum_{t=1}^T\sum_{k=1}^{50}\mathbb{E}(c_k^t\cdot\mu_k^t)/T$, subject to the following two requirements: (1) The service rate supports the arrival rate: $\sum_{t=1}^T\sum_{k=1}^{50}\mathbb{E}(g_k^t(\mu_k^t))\ge\sum_{t=1}^T\mathbb{E}(\lambda(t))$. Note that since $g_k^t(\mu_k^t)$ is concave function for $\mu_k^t\ge0$, this is a convex inequality constraint. (2) The internal budget consumption is well-paced, i.e. each cluster consumes a fixed ratio of the total consumed budget in expectation. More specifically, in the simulation, let $\mathcal{I}_1,\cdots,\mathcal{I}_5$ be index sets of 5 clusters, then, it is required that $\sum_{t=1}^T\sum_{k\in\mathcal{I}_j}\mathbb{E}(h_k^t\cdot\mu_k^t)=\beta_j\cdot\sum_{t=1}^T\sum_{k=1}^{50}\mathbb{E}(h_k^t\cdot\mu_k^t)$, $j=1,2,3$ and $\sum_{t=1}^T\sum_{k\in\mathcal{I}_4\cup\mathcal{I}_5}\mathbb{E}(h_k^t\cdot\mu_k^t)=\beta_4\cdot\sum_{t=1}^T\sum_{k=1}^{50}\mathbb{E}(h_k^t\cdot\mu_k^t)$, where $[\beta_1,\beta_2,\beta_3,\beta_4]=[0.05,\ 0.10,\ 0.25,\ 0.60]$. In Fig 1, we compare our proposed algorithm with the best fixed solution in hindsight choosing the best fixed power allocation knowing all the data, and a benchmark Reac algorithm [7]. The Reac algorithm is adapted to our pacing scenario by estimating

the number of jobs in the next slot via the average of past 10 slots and assign the load according to the pacing ratio. For cluster 4 and cluster 5 (which take up a total ratio of 0.60), the Reac algorithm evenly distribute the workload between the two. Our algorithm achieves a similar electricity money spend with the best fixed solution which is better than Reac, while keeping the average number of unserved job low and achieving a fast budget pacing.

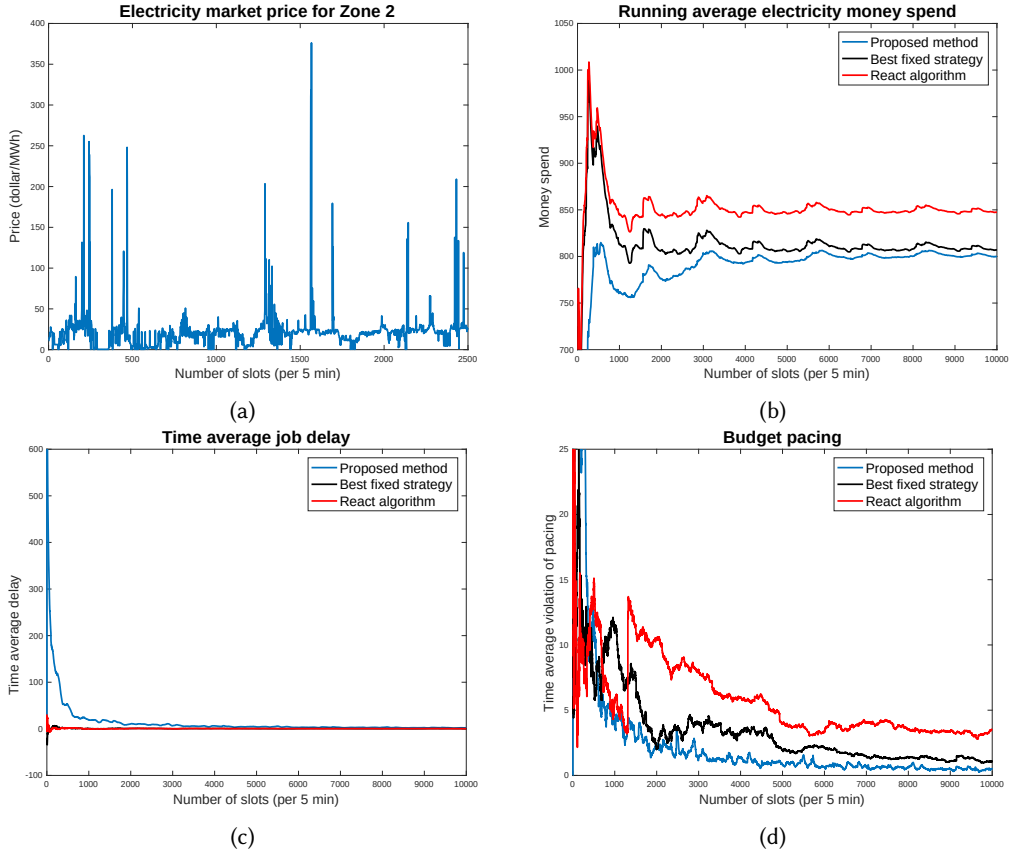

Fig. 1. (a) Electricity market prices at zone DUNWOD New York; (b) Average money spent buying electricity; (c) Average unserved jobs. (d) Average violation of pacing constraints.

5 PROOF OF MAIN RESULTS

In this section, we present the proof Theorem 3.1. The main lemmas as well as how they lead to the regret and constraint violation bounds will be presented in Section 5.1 and 5.2, respectively. The detailed proofs of those lemmas will be presented in Section 5.3. The idea of proving Theorem 3.2 is similar and the details will be delayed to the Supplement A.4.

5.1 Proof of regret bound

We start with the following key bound of a "drift-plus-penalty (DPP)" expression:

LEMMA 5.1. *Define the drift* $\Delta(t) := (\|\mathbf{Q}(t+1)\|_2^2 - \|\mathbf{Q}(t)\|_2^2)/2 + (\|\mathbf{H}(t+1)\|_2^2 - \|\mathbf{H}(t)\|_2^2)/2$. *Consider the following "drift-plus-penalty" (DPP) expression at time* t: $V\left\langle \nabla f^{t-1}(\mu^{t-1}), \mu^t - \mu^{t-1}\right\rangle + \Delta(t) +$

$\alpha D(\mu^t, \mu^{t-1})$. *Let* $M = 4RH^2/\beta + G^2 + 2RD_2^2/\beta$ *where* β *is in* (2)*, then, for any* $\mu \in \Delta$,

$$V\left\langle \nabla f^{t-1}(\mu^{t-1}), \mu^t - \mu^{t-1}\right\rangle + \Delta(t) + \alpha D(\mu^t, \mu^{t-1}) \\ \le V(f^{t-1}(\mu) - f^{t-1}(\mu^{t-1})) + \sum_{j=1}^{M} H_j(t)\left(\left\langle h_j^{t-1}, \mu\right\rangle - b_j\right) \\ + \sum_{i=1}^{L} Q_i(t) g_i^{t-1}(\mu) + \alpha D(\mu, \mu^{t-1}) - \alpha D(\mu, \mu^t) + M. \quad (11)$$

This lemma is proved via the property of Bregman divergence (Lemma 2.1). The details are deferred to Section 5.3.1. Now, for the DPP expression on the left hand side, we also have the following lower bound:

Lemma 5.2. *Our Algorithm 1 ensures*

$$V\left\langle \nabla f^{t-1}(\mu^{t-1}), \mu^t - \mu^{t-1}\right\rangle + \alpha D(\mu^t, \mu^{t-1}) \ge -V^2 D_1^2/2\alpha\beta. \quad (12)$$

This lemma is also proved in Section 5.3.1. Substituting this bound in to (11), taking $\mu = \mu^*$ which is the solution to the in-hindsight problem (3), and taking conditional expectations from both sides, we readily get:

$$-\frac{V^2}{2\alpha\beta}D_1^2 + \mathbb{E}(\Delta(t)|\mathcal{F}_{t-1}) \le V\mathbb{E}\Big(f^{t-1}(\mu^*) - f^{t-1}(\mu^{t-1})|\mathcal{F}_{t-1}\Big) \\ + \mathbb{E}\Big[\sum_{i=1}^{L} Q_i(t) g_i^{t-1}(\mu^*)\Big|\mathcal{F}_{t-1}\Big] + \mathbb{E}\Big[\sum_{j=1}^{M} H_j(t)\left(\left\langle h_j^{t-1}, \mu^*\right\rangle - b_j\right)\Big|\mathcal{F}_{t-1}\Big] \\ + \alpha\mathbb{E}\Big(D(\mu^*, \mu^{t-1}) - D(\mu^*, \mu^t)\,|\mathcal{F}_{t-1}\Big) + M. \quad (13)$$

Note that

$$\mathbb{E}\Big(\sum_{j=1}^{M} H_j(t)\left(\left\langle h_j^{t-1}, \mu^*\right\rangle - b_j\right)\Big|\mathcal{F}_{t-1}\Big) = \sum_{j=1}^{M} H_j(t)\mathbb{E}\Big(\left\langle h_j^{t-1}, \mu^*\right\rangle - b_j\Big) = 0,$$

$$\mathbb{E}\Big(\sum_{i=1}^{L} Q_i(t) g_i^{t-1}(\mu^*)\,\Big|\mathcal{F}_{t-1}\Big) = \sum_{i=1}^{L} Q_i(t)\mathbb{E}\Big(g_i^{t-1}(\mu^*)\Big) \le 0,$$

where, in both inequalities, the first step follows from the fact that h_j^t, g_i^t are i.i.d. and $H_j(t), Q_i(t)$ depend on $\mathcal{F}_{t-1}$, and the second step follows from μ^* being a solution to the in-hindsight optimization (3), thus, must be feasible, i.e. $\mathbb{E}(g_i^{t-1}(\mu^*)) \le 0$, $\mathbb{E}(\langle h_j^{t-1}, \mu^*\rangle) = 0$. Thus, taking full expectation from both sides of (13) gives

$$\mathbb{E}(\Delta(t)) + V\mathbb{E}\big(f^{t-1}(\mu^{t-1}) - f^{t-1}(\mu^*)\big) \le M + \frac{V^2 D_1^2}{2\alpha\beta} + \alpha\mathbb{E}\big(D(\mu^*, \mu^{t-1}) - D(\mu^*, \mu^t)\big).$$

Taking a telescoping sum on both sides from 0 to $T-1$ and dividing both sides by TV,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big(f^{t-1}(\mu^{t-1}) - f^{t-1}(\mu^*)\big) \le \frac{M}{V} + \frac{VD_1^2}{2\alpha\beta} + \frac{\alpha}{VT}D(\mu^*, \mu^0),$$

where we use the fact that since $Q_i(0) = 0$, $H_j(0) = 0$, and $\sum_{t=0}^{T-1}\Delta(t) = (\|\mathbf{Q}(T)\|_2^2 + \|\mathbf{H}(T)\|_2^2)/2 \ge 0$. Substituting $\alpha = T, V = \sqrt{T}$, and $D(\mu^*, \mu^0) \le R$ yields the desired result with $C_0' = RH^2/\beta + G^2 + 2RD_2^2/\beta + D_1^2/2\beta + R$.

5.2 Proof of constraint violations

In this section, we present the proof of constraint violations in Theorem 3.1. First, it is enough to bound dual multipliers via the following lemma:

LEMMA 5.3. *The updating rule (5) and (6) delivers the following constraint violation bounds:*

$$\mathbb{E}\left\|\left[\frac{1}{T}\sum_{t=0}^{T-1}\overline{\mathbf{g}}(\mu^t)\right]_+\right\|_2 \leq \frac{\mathbb{E}(\|\mathbf{Q}(t)\|_2)}{T}+\frac{VD_1D_2}{\alpha\beta}+\frac{1}{T}\sum_{t=1}^{T}\frac{D_2}{\alpha\beta}\left(D_2\mathbb{E}(\|\mathbf{Q}(t)\|_2)+H\mathbb{E}(\|\mathbf{H}(t)\|_2)\right),$$

$$\mathbb{E}\left\|\frac{1}{T}\sum_{t=0}^{T-1}\overline{\mathbf{h}}(\mu^t)-\mathbf{b}\right\|_2 \leq \frac{\mathbb{E}(\|\mathbf{H}(t)\|_2)}{T}+\frac{VD_1H}{\alpha\beta}+\frac{1}{T}\sum_{t=1}^{T}\frac{H}{\alpha\beta}\left(D_2\mathbb{E}(\|\mathbf{Q}(t)\|_2)+H\mathbb{E}(\|\mathbf{H}(t)\|_2)\right).$$

This lemma is proved in Section 5.3.2. To bound $\mathbb{E}(\|\mathbf{Q}(t)\|_2)$ and $\mathbb{E}(\|\mathbf{H}(t)\|_2)$, we have the following lemma whose proof can be found in Section 5.3.3:

LEMMA 5.4. *Define constant* $C_{V,\alpha,t_0} := 2\left(4RH^2/\beta+G^2+2RD_2^2/\beta+V^2D_1^2/(2\alpha\beta)+VF\right)t_0+2\left(3G^2/2+2RD_2^2/\beta+8RH^2/\beta\right)t_0^2+2\alpha R$. *Then, for any integer* $t_0\geq 1$, *we have the* t_0 *step drift satisfies*

$$\begin{aligned}&\mathbb{E}\left(\|\mathbf{Q}(t+t_0)\|_2^2+\|\mathbf{H}(t+t_0)\|_2^2\,|\mathcal{F}^{t-1}\right)-\|\mathbf{Q}(t)\|_2^2-\|\mathbf{H}(t)\|_2^2\\ \leq& 2Vt_0\mathbb{E}\left(q^{(t-1,t_0)}\left(\frac{\mathbf{Q}(t)}{V},\ \frac{\mathbf{H}(t)}{V}\right)\,\middle|\,\mathcal{F}^{t-1}\right)+C_{V,\alpha,t_0}.\end{aligned}\tag{14}$$

where the dual function $q^{(t-1,t_0)}$ *is defined in (8).*

This bound establishes the relation between dual multipliers and the dual function. Next, in view of (14), we would like to show that $\mathbb{E}\left(q^{(t-1,t_0)}\left(\frac{\mathbf{Q}(t)}{V},\ \frac{\mathbf{H}(t)}{V}\right)\,\middle|\,\mathcal{F}^{t-1}\right)$ is small. This is done via Lemma 2.3 that whenever $\left(\frac{\mathbf{Q}(t)}{V},\ \frac{\mathbf{H}(t)}{V}\right)$ is far away from the optimal set $\mathcal{V}^*_{t-1,t_0} := \operatorname{argmax}_{\lambda,\eta}q^{(t-1,t_0)}(\lambda,\ \eta)$, which is nonempty and bounded by Assumption 2.2, $\mathbb{E}\left(q^{(t-1,t_0)}\left(\frac{\mathbf{Q}(t)}{V},\ \frac{\mathbf{H}(t)}{V}\right)\,\middle|\,\mathcal{F}^{t-1}\right)$ becomes negative. In fact one can prove the following lemma:

LEMMA 5.5. *The dual function has the following bound:*

$$\mathbb{E}\left(q^{(t-1,t_0)}\left(\frac{\mathbf{Q}(t)}{V},\ \frac{\mathbf{H}(t)}{V}\right)\,|\mathcal{F}^{t-1}\right)\leq F+\overline{\ell}(G+\sqrt{2RH^2/\beta}+\overline{c})+\overline{c}B-\overline{c}\left\|\left(\frac{\mathbf{Q}(t)}{V},\ \frac{\mathbf{H}(t)}{V}\right)\right\|_2,$$

where B *is defined in Assumption 2.2.*

The detailed proof can be found in Section 5.3.4. Substituting the above lemma into (14) and using a known stochastic drift lemma, one can prove the following bound by setting $t_0=\sqrt{T}$, $V=\sqrt{T}$, $\alpha=T$. The details are in Section 5.3.5:

LEMMA 5.6. *The quantity* $\|(\mathbf{Q}(t),\ \mathbf{H}(t))\|_2$ *satisfies the following conditions:*

$$\mathbb{E}\left(\left\|\left(\mathbf{Q}(t),\ \mathbf{H}(t)\right)\right\|_2\right)\leq C'+C''\sqrt{T}\tag{15}$$

where $C':=\frac{2}{\overline{c}}\left(4RH^2/\beta+G^2+2RD_2^2/\beta+D_1^2/(2\beta)\right)$ *and* $C'':=\frac{2}{\overline{c}}\left(2F+3G^2/2+2RD_2^2/\beta+8RH^2/\beta+R+\overline{\ell}(G+\sqrt{8RH^2/\beta}+\overline{c})+\overline{c}B+4\left(2(G+\sqrt{2RD_2^2/\beta})+\sqrt{8RH^2/\beta}\right)^2\log\left(\frac{32}{\overline{c}^2}\left(2(G+\sqrt{2RD_2^2/\beta})+\sqrt{8RH^2/\beta}\right)^2\right)\right)$ *are absolute constants.*

Substituting the bound (15) into Lemma 5.3 with $\alpha=T$ and $V=\sqrt{T}$ gives the final constraint violation bounds.

5.3 Proof of Technical Lemmas

Throughout the section, we let $\mathcal{F}_t$ be the system history up to time t, which includes $\{g_i^\tau\}_{\tau=0}^{t-1}$, $\{h_i^\tau\}_{\tau=0}^{t-1}$, and $\{f^\tau\}_{\tau=0}^{t-1}$.

5.3.1 *Proof of Lemma 5.1 and 5.2.*

Proof of Lemma 5.1. Applying Lemma 2.1 by setting $y = \mu^{t-1}$, $x^* = \mu^t$, $f(x) = \langle x, p\rangle$ and

$$p = V\nabla f^{t-1}(\mu^{t-1}) + \sum_{i=1}^{L} Q_i(t)\nabla g_i^{t-1}(\mu^{t-1}) + \sum_{i=1}^{M} H_i(t)h_i^{t-1},$$

we have

$$\begin{aligned}&\left\langle V\nabla f^{t-1}(\mu^{t-1}) + \sum_{i=1}^{L} Q_i(t)\nabla g_i^{t-1}(\mu^{t-1}) + \sum_{i=1}^{M} H_i(t)h_i^{t-1}, \mu^t\right\rangle + \alpha D(\mu^t, \mu^{t-1})\\ &\le \left\langle V\nabla f^{t-1}(\mu^{t-1}) + \sum_{i=1}^{L} Q_i(t)\nabla g_i^{t-1}(\mu^{t-1}) + \sum_{i=1}^{M} H_i(t)h_i^{t-1}, \mu\right\rangle + \alpha\left(D(\mu, \mu^{t-1}) - D(\mu, \mu^t)\right) \quad (16)\end{aligned}$$

On the other hand, recall that we define

$$\tilde{g}_i^t(\mu^t) := g_i^t(\mu^{t-1}) + \left\langle \nabla g_i^{t-1}(\mu^{t-1}), \mu^t - \mu^{t-1}\right\rangle.$$

Using the updating rule (5), (6) and Holder's inequality that $\langle x, y\rangle \le \|x\|\|y\|_*$, we have

$$\begin{aligned}H_i(t+1)^2 - H_i(t)^2 =& 2H_i(t)(\left\langle h_i^{t-1}, \mu^t\right\rangle - b_i) + |\left\langle h_i^{t-1}, \mu^t\right\rangle - b_i|^2\\ \le& 2H_i(t)(\left\langle h_i^{t-1}, \mu^t\right\rangle - b_i) + \frac{8R}{\beta}\|h_i^{t-1}\|_*^2,\end{aligned}$$

where the inequality for $H_i(t+1)^2 - H_i(t)^2$ follows from

$$|\left\langle h_i^{t-1}, \mu^t\right\rangle - b_i|^2 \le 2|\left\langle h_i^{t-1}, \mu^t\right\rangle|^2 + 2|b_i|^2 = 2|\left\langle h_i^{t-1}, \mu^t\right\rangle|^2 + 2\left|\mathbb{E}\left(\left\langle h_i^{t-1}, \mu^*\right\rangle\right)\right|^2 \le 8R/\beta,$$

via Assumption 2.1(3) that $\sup_{\mu^a,\mu^b\in\Delta}\|\mu^a - \mu^b\|^2 \le 2R/\beta$ and $b_i = \mathbb{E}\left(\left\langle h_i^{t-1}, \mu^*\right\rangle\right)$. Also, we have

$$\begin{aligned}Q_i(t+1)^2 - Q_i(t)^2 =& \max\{Q_i(t) + \tilde{g}_i^t(\mu^t),\ 0\}^2 - Q_i(t)^2\\ \le& 2Q_i(t)\tilde{g}_i^t(\mu^t) + \tilde{g}_i^t(\mu^t)^2\\ \le& 2Q_i(t)\tilde{g}_i^t + 2(g_i^{t-1}(\mu^{t-1}))^2 + \frac{4R}{\beta}\|\nabla g_i^{t-1}(\mu^{t-1})\|_*^2,\end{aligned}$$

where the first inequality follows from the following fact: If $Q_i(t) + \tilde{g}_i^t(\mu^t) \ge 0$, then, the equality is attained and if $Q_i(t) + \tilde{g}_i^t(\mu^t) \le 0$, then, $\max\{Q_i(t) + \tilde{g}_i^t(\mu^t),\ 0\}^2 = 0$ and the inequality follows from $Q_i(t)^2 + 2Q_i(t)\tilde{g}_i^t(\mu^t) + \tilde{g}_i^t(\mu^t) \ge 0$. The second inequality follows from the assumption $\sup_{\mu^a,\mu^b\in\Delta}\|\mu^a - \mu^b\|^2 \le 2R/\beta$ and the definition of $\tilde{g}_i^t(\mu^t)$ in Algorithm 1 that

$$\begin{aligned}\tilde{g}_i^t(\mu^t)^2 =& \left(g_i^{t-1}(\mu^{t-1}) + \left\langle \nabla g_i^{t-1}(\mu^{t-1}), \mu^t - \mu^{t-1}\right\rangle\right)^2\\ \le& 2(g_i^{t-1}(\mu^{t-1}))^2 + 2\left(\left\langle \nabla g_i^{t-1}(\mu^{t-1}), \mu^t - \mu^{t-1}\right\rangle\right)^2\\ \le& 2(g_i^{t-1}(\mu^{t-1}))^2 + 2\|\nabla g_i^{t-1}(\mu^{t-1})\|_*^2\|\mu^t - \mu^{t-1}\|^2\\ \le& 2(g_i^{t-1}(\mu^{t-1}))^2 + \frac{4R}{\beta}\|\nabla g_i^{t-1}(\mu^{t-1})\|_*^2,\end{aligned}$$

where the last inequality follows from Assumption 2.1 and (2). Thus, we have

$$\begin{aligned}\Delta(t) \leq & \sum_{i=1}^{L} Q_i(t)\tilde{g}_i^t(\mu^t) + \sum_{i=1}^{M} H_i(t)\left(\left\langle h_i^t, \mu^t\right\rangle - b_i\right) + \frac{4R}{\beta}\sum_{i=1}^{M}\|h_i^t\|_*^2 \\ & + \sum_{i=1}^{L} g_i^{t-1}(\mu^{t-1})^2 + \frac{2R}{\beta}\sum_{i=1}^{L}\|\nabla g_i^{t-1}(\mu^{t-1})\|_*^2 \\ \leq & \sum_{i=1}^{L} Q_i(t)\tilde{g}_i^t(\mu^t) + \sum_{i=1}^{M} H_i(t)(\left\langle h_i^t, \mu^t\right\rangle - b_i) + \frac{4RH^2}{\beta} + G^2 + \frac{2RD_2^2}{\beta}, \end{aligned} \tag{17}$$

where the last inequality follows from Assumption 2.1(1). To this point, we consider the following drift-plus-penalty term, by (17),

$$\begin{aligned}& \Delta(t) + V\left\langle \nabla f^{t-1}(\mu^{t-1}), \mu^t - \mu^{t-1}\right\rangle + \alpha D(\mu^t, \mu^{t-1}) \\ \leq & \sum_{i=1}^{L} Q_i(t)\left(g_i^{t-1}(\mu^{t-1}) + \left\langle \nabla g_i^{t-1}(\mu^{t-1}), \mu^t - \mu^{t-1}\right\rangle\right) + \sum_{j=1}^{M} H_j(t)\left(\left\langle h_j^{t-1}, \mu^t\right\rangle - b_i\right) \\ & + \frac{4RH^2}{\beta} + G^2 + \frac{2RD_2^2}{\beta} + V\left\langle \nabla f^{t-1}(\mu^{t-1}), \mu^t - \mu^{t-1}\right\rangle + \alpha D(\mu^t, \mu^{t-1}).\end{aligned}$$

Now, by (16), we have for any $\mu \in \Delta$,

$$\begin{aligned}& \Delta(t) + V\left\langle \nabla f^{t-1}(\mu^{t-1}), \mu^t - \mu^{t-1}\right\rangle + \alpha D(\mu^t, \mu^{t-1}) \\ \leq & \sum_{i=1}^{L} Q_i(t)\left(g_i^{t-1}(\mu^{t-1}) + \left\langle \nabla g_i^{t-1}(\mu^{t-1}), \mu - \mu^{t-1}\right\rangle\right) + \sum_{j=1}^{M} H_j(t)\left(\left\langle h_j^{t-1}, \mu\right\rangle - b_i\right) \\ & + \frac{4RH^2}{\beta} + G^2 + \frac{2RD_2^2}{\beta} + \alpha D(\mu, \mu^{t-1}) - \alpha D(\mu, \mu^t) + V\left\langle \nabla f^{t-1}(\mu^{t-1}), \mu - \mu^{t-1}\right\rangle\end{aligned}$$

Note that by convexity, we have for any μ,

$$\begin{aligned} f^{t-1}(\mu) &\geq f^{t-1}(\mu^{t-1}) + \left\langle \nabla f^{t-1}(\mu^{t-1}), \mu - \mu^{t-1}\right\rangle, \\ g_i^{t-1}(\mu) &\geq g_i^{t-1}(\mu^{t-1}) + \left\langle \nabla g_i^{t-1}(\mu^{t-1}), \mu - \mu^{t-1}\right\rangle. \end{aligned}$$

Thus, it follows (11) holds. □

Proof of Lemma 5.2. We have

$$\begin{aligned}& V\left\langle \nabla f^{t-1}(\mu^{t-1}), \mu^t - \mu^{t-1}\right\rangle + \alpha D(\mu^t, \mu^{t-1}) \\ \geq & V\left\langle \nabla f^{t-1}(\mu^{t-1}), \mu^t - \mu^{t-1}\right\rangle + \frac{\alpha\beta}{2}\|\mu^t - \mu^{t-1}\|^2 \\ \geq & -V\|\nabla f^{t-1}(\mu^{t-1})\|_*\|\mu^t - \mu^{t-1}\| + \frac{\alpha\beta}{2}\|\mu^t - \mu^{t-1}\|^2 \\ \geq & -V\left(\frac{\alpha\beta}{2V}\|\mu^t - \mu^{t-1}\|^2 + \frac{V}{2\alpha\beta}\|\nabla f^{t-1}(\mu^{t-1})\|_*^2\right) + \frac{\alpha\beta}{2}\|\mu^t - \mu^{t-1}\|^2 \\ = & -\frac{V^2}{2\alpha\beta}\|\nabla f^{t-1}(\mu^{t-1})\|_*^2 \geq -\frac{V^2}{2\alpha\beta}D_1^2.\end{aligned}$$

where the first inequality follows from the strong convexity (2), the second inequality follows from Holder's inequality, the third inequality follows from the fact that $ab \leq \frac{a^2+b^2}{2}$, $\forall a, b$, and the last inequality follows from the bound $\|\nabla f^{t-1}(\mu^{t-1})\|_* \leq D_1$. □

5.3.2 Proof of Lemma 5.3. We start with a supporting lemma:

Lemma 5.7. *The updating rule (5) and (6) delivers the following constraint violation bounds:*

$$\mathbb{E}\left\|\left[\frac{1}{T}\sum_{t=0}^{T-1}\overline{\mathbf{g}}(\mu^t)\right]_+\right\|_2 \leq \frac{\mathbb{E}(\|\mathbf{Q}(t)\|_2)}{T} + \frac{D_2}{T}\sum_{t=0}^{T-1}\mathbb{E}\big(\|\mu^{t+1}-\mu^t\|\big)$$
$$\mathbb{E}\left\|\frac{1}{T}\sum_{t=0}^{T-1}\overline{\mathbf{h}}(\mu^t)-\mathbf{b}\right\|_2 \leq \frac{\mathbb{E}(\|\mathbf{H}(t)\|_2)}{T} + \frac{H}{T}\sum_{t=0}^{T-1}\mathbb{E}\big(\|\mu^{t+1}-\mu^t\|\big)$$

Proof of Lemma 5.7. We prove the first inequality and the second inequality is proved in the same way. Note by (5), we have

$$\begin{aligned}
Q_i(t+1) &= \max\{Q_i(t) + g_i^{t-1}(\mu^{t-1}) + \left\langle \nabla g_i^{t-1}(\mu^{t-1}), \mu^t - \mu^{t-1}\right\rangle,\ 0\}\\
&\geq \max\{Q_i(t) + g_i^{t-1}(\mu^{t-1}) - \|\nabla g_i^{t-1}(\mu^{t-1})\|_*\|\mu^t-\mu^{t-1}\|,\ 0\}\\
&\geq Q_i(t) + g_i^{t-1}(\mu^{t-1}) - \|\nabla g_i^{t-1}(\mu^{t-1})\|_*\|\mu^t-\mu^{t-1}\|.
\end{aligned}$$

Taking a telescoping sum from both sides from 0 to $T-1$,

$$Q_i(T) \geq \sum_{t=0}^{T-1} g_i^t(\mu^t) - \sum_{t=0}^{T-1}\|\nabla g_i^t(\mu^t)\|_*\|\mu^{t+1}-\mu^t\|.$$

Rearranging the terms and dividing both sides by T give

$$\frac{1}{T}\sum_{t=0}^{T-1} g_i^t(\mu^t) \leq \frac{Q_i(T)}{T} + \frac{1}{T}\sum_{t=0}^{T-1}\|\nabla g_i^t(\mu^{t-1})\|_*\|\mu^{t+1}-\mu^t\|.$$

Note that the right hand side is nonnegative due to $Q_i(T) \geq 0$, the inequality still holds if take the max with 0 from the left hand side, i.e. denote $[x]_+ := \max\{x, 0\}$, then,

$$\left[\frac{1}{T}\sum_{t=0}^{T-1} g_i^t(\mu^t)\right]_+ \leq \frac{Q_i(T)}{T} + \frac{1}{T}\sum_{t=0}^{T-1}\|\nabla g_i^t(\mu^{t-1})\|_*\|\mu^{t+1}-\mu^t\|.$$

Thus, we have

$$\begin{aligned}
\left\|\left[\frac{1}{T}\sum_{t=0}^{T-1}\overline{\mathbf{g}}(\mu^t)\right]_+\right\|_2 &\leq \frac{\|\mathbf{Q}(T)\|_2}{T} + \frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\sum_{i=1}^{L}\|\nabla g_i^t(\mu^t)\|_*^2}\|\mu^{t+1}-\mu^t\|\\
&\leq \frac{\|\mathbf{Q}(T)\|_2}{T} + \frac{D_2}{T}\sum_{t=0}^{T-1}\|\mu^{t+1}-\mu^t\|,
\end{aligned}$$

where the second inequality follows from Assumption 2.1. □

Proof of Lemma 5.3. It is enough to bound the difference term $\mathbb{E}(\|\mu^{t+1}-\mu^t\|)$. We start from the relation (16) by taking $\mu=\mu^{t-1}$,

$$\begin{aligned} V\left\langle \nabla f^{t-1}(\mu^{t-1}),\mu^t\right\rangle + \sum_{i=1}^{L} Q_i(t)\left\langle \nabla g_i^{t-1}(\mu^{t-1}),\mu^t\right\rangle + \sum_{j=1}^{M} H_i(t)\left\langle h_i^{t-1},\mu^t\right\rangle + \alpha D(\mu^t,\mu^{t-1}) \\ \leq V\left\langle \nabla f^{t-1}(\mu^{t-1}),\mu^{t-1}\right\rangle + \sum_{i=1}^{L} Q_i(t)\left\langle \nabla g_i^{t-1}(\mu^{t-1}),\mu^{t-1}\right\rangle \\ + \sum_{j=1}^{M} H_j(t)\left\langle h_i^{t-1},\mu^{t-1}\right\rangle - \alpha D(\mu^{t-1},\mu^t). \quad (18) \end{aligned}$$

Note that we have

$$\left\langle \nabla f^{t-1}(\mu^{t-1}),\mu^{t-1}-\mu^t\right\rangle \leq \|\nabla f^{t-1}(\mu^{t-1})\|_*\|\mu^{t-1}-\mu^t\| \leq D_1\|\mu^t-\mu^{t-1}\|.$$

On the other hand, we also have

$$\begin{aligned} \sum_{i=1}^{L} Q_i(t)\left\langle \nabla g_i^{t-1}(\mu^{t-1}),\mu^{t-1}-\mu^t\right\rangle \leq & \|\mathbf{Q}(t)\|_2\sqrt{\sum_{i=1}^{L}(\|\nabla g_i(\mu^{t-1})\|_*\|\mu^t-\mu^{t-1}\|)^2} \\ \leq & D_2\|\mathbf{Q}(t)\|_2\|\mu^t-\mu^{t-1}\|, \end{aligned}$$

and

$$\sum_{j=1}^{M} H_j(t)\left\langle h_i^{t-1},\mu^{t-1}-\mu^t\right\rangle \leq \|\mathbf{H}(t)\|_2\sqrt{\sum_{i=1}^{M}(\|h_i^{t-1}\|_*\|\mu^t-\mu^{t-1}\|)^2} \leq H\|\mathbf{H}(t)\|_2\|\mu^t-\mu^{t-1}\|.$$

Substituting the above three bounds into (18) gives

$$D(\mu^t,\mu^{t-1})+D(\mu^{t-1},\mu^t) \leq \frac{1}{\alpha}\left(VD_1+D_2\|\mathbf{Q}(t)\|_2+H\|\mathbf{H}(t)\|_2\right)\|\mu^t-\mu^{t-1}\|$$

By strong convexity (2), we have

$$D(\mu^t,\mu^{t-1})+D(\mu^{t-1},\mu^t) \geq \beta\|\mu^t-\mu^{t-1}\|^2$$

Thus, it follows,

$$\beta\|\mu^t-\mu^{t-1}\|^2 \leq \frac{1}{\alpha}\left(VD_1+D_2\|\mathbf{Q}(t)\|_2+H\|\mathbf{H}(t)\|_2\right)\|\mu^t-\mu^{t-1}\|.$$

Solving the above quadratic inequality yields

$$\|\mu^t-\mu^{t-1}\| \leq \frac{1}{\alpha\beta}\left(VD_1+D_2\|\mathbf{Q}(t)\|_2+H\|\mathbf{H}(t)\|_2\right).$$

Taking the expectation from both sides and subtracting this bound into Lemma 5.7 result in

$$\mathbb{E}\left\|\left[\frac{1}{T}\sum_{t=0}^{T-1}\overline{\mathbf{g}}(\mu^t)\right]_+\right\|_2 \leq \frac{\mathbb{E}(\|\mathbf{Q}(t)\|_2)}{T}+\frac{VD_1D_2}{\alpha\beta}+\frac{1}{T}\frac{D_2}{\alpha\beta}\sum_{t=1}^{T}\left(D_2\mathbb{E}(\|\mathbf{Q}(t)\|_2)+H\mathbb{E}(\|\mathbf{H}(t)\|_2)\right)$$

One can prove the bound on $\mathbb{E}\left\|\frac{1}{T}\sum_{t=0}^{T-1}\overline{\mathbf{h}}(\mu^t)\right\|_2$ with exactly the same computation and we omit the details. □

5.3.3 Proof of Lemma 5.4. For simplicity of notations, let constant $\bar{c}$ be the minimum over all c_0's and let $\bar{\ell}$ be the maximum over all ℓ_0's in Lemma 2.3 with $t = 0, 1, 2, \cdots, T-1$ and $k = \sqrt{T}$. We start with the following supporting lemma:

Lemma 5.8. *Consider the t_0 slots drift for some positive integer t_0, then we have*

$$\frac{\|\mathbf{Q}(t+t_0)\|_2^2 + \|\mathbf{H}(t+t_0)\|_2^2 - \|\mathbf{Q}(t)\|_2^2 - \|\mathbf{H}(t)\|_2^2}{2}$$
$$\leq V \sum_{\tau=t}^{t+t_0-1} f^{\tau-1}(\mu) + \sum_{i=1}^{L} Q_i(t) \sum_{\tau=t}^{t+t_0-1} g_i^{\tau-1}(\mu) + \sum_{j=1}^{M} H_j(t) \sum_{\tau=t}^{t+t_0-1} \left(\left\langle h_j^{\tau-1}, \mu \right\rangle - b_j \right) + \frac{1}{2} C_{V,\alpha,t_0}. \quad (19)$$

Proof of Lemma 5.8. We start from equation (11). Substituting (12), we have

$$\Delta(t) + V(f^{t-1}(\mu^{t-1}) - f^{t-1}(\mu)) \leq \sum_{i=1}^{L} Q_i(t) g_i^{t-1}(\mu) + \sum_{j=1}^{M} H_j(t) \left(\left\langle h_j^{t-1}, \mu \right\rangle - b_j \right)$$
$$+ \frac{4RH^2}{\beta} + G^2 + \frac{2RD_2^2}{\beta} + \frac{V^2 D_1^2}{2\alpha\beta} + \alpha D(\mu, \mu^{t-1}) - \alpha D(\mu, \mu^t).$$

Take the summation from both sides between t to $t + t_0 - 1$ for some t_0 to be determined later, we obtain

$$\frac{\|\mathbf{Q}(t+t_0)\|_2^2 + \|\mathbf{H}(t+t_0)\|_2^2 - \|\mathbf{Q}(t)\|_2^2 - \|\mathbf{H}(t)\|_2^2}{2}$$
$$\leq \sum_{\tau=t}^{t+t_0-1} \sum_{i=1}^{L} Q_i(\tau) g_i^{\tau-1}(\mu) + \sum_{\tau=t}^{t+t_0-1} \sum_{j=1}^{M} H_j(\tau) \left(\left\langle h_j^{\tau-1}, \mu \right\rangle - b_j \right)$$
$$+ \left(\frac{4RH^2}{\beta} + G^2 + \frac{2RD_2^2}{\beta} + \frac{V^2 D_1^2}{2\alpha\beta} \right) t_0 + \alpha D(\mu, \mu^{t-1}) - \alpha D(\mu, \mu^{t+t_0-1})$$
$$+ V \sum_{\tau=t}^{t+t_0-1} (f^{\tau-1}(\mu^{\tau-1}) - f^{\tau-1}(\mu)) \quad (20)$$

Using Assumption 2.1, we have $V \sum_{\tau=t}^{t+t_0-1} f^{\tau-1}(\mu^{\tau-1}) \leq VFt_0$. Recall that $Q_i(t+1) = \max\{Q_i(t) + \tilde{g}_i^t(\mu^t),\ 0\}$, where

$$\tilde{g}_i^t(\mu^t) := g_i^{t-1}(\mu^{t-1}) + \left\langle \nabla g_i^{t-1}(\mu^{t-1}), \mu^t - \mu^{t-1} \right\rangle,$$

and $H_j(t+1) = H_j(t) + \left\langle h_j^{t-1}, \mu^t \right\rangle - b_j$, we have

$$
\begin{aligned}
&\sum_{\tau=t}^{t+t_0-1}\sum_{i=1}^{L}(Q_i(\tau)-Q_i(t))\,g_i^{\tau-1}(\mu)\\
\le&\sum_{\tau=t+1}^{t+t_0-1}\sum_{i=1}^{L}(\sum_{\tau'=t}^{\tau-1}|\tilde{g}_i^{\tau'}(\mu^{\tau'})|)\cdot|g_i^{\tau-1}(\mu)|\\
\le&\sum_{\tau=t+1}^{t+t_0-1}\sum_{\tau'=t}^{\tau-1}\sum_{i=1}^{L}(|\tilde{g}_i^{\tau'}(\mu^{\tau'})|^2+|g_i^{\tau-1}(\mu)|^2)/2\\
\le&\sum_{\tau=t+1}^{t+t_0-1}\sum_{\tau'=t}^{\tau-1}\sum_{i=1}^{L}|g_i^{\tau'-1}(\mu^{\tau'-1})|^2+\|\nabla g_i^{\tau'-1}\|_*^2\frac{2R}{\beta}+|g_i^{\tau-1}(\mu)|^2/2\\
\le&t_0^2\left(\frac{3}{2}G^2+\frac{2RD_2^2}{\beta}\right),
\end{aligned}
\tag{21}
$$

where the second from the last inequality follows from $\|\mu^t-\mu^{t-1}\|^2\le 2R/\beta$, and the last inequality follows from Assumption 2.1. Similarly, we can show that

$$
\sum_{\tau=t}^{t+t_0-1}\sum_{j=1}^{M}\left(H_j(\tau)-H_j(t)\right)\left(\left\langle h_j^{\tau-1},\mu^{\tau}\right\rangle-b_j\right)\le t_0^2\frac{8RH^2}{\beta}. \tag{22}
$$

Substituting the above two bounds into (20), using the fact that $\alpha D(\mu,\mu^{t+t_0-1})\ge 0$ and $D(\mu,\mu^t)\le R$ and that $C_{V,\alpha,t_0}:=2\left(\frac{4RH^2}{\beta}+G^2+\frac{2RD_2^2}{\beta}+\frac{V^2D_1^2}{2\alpha\beta}+VF\right)t_0+2\left(\frac{3}{2}G^2+\frac{2RD_2^2}{\beta}+\frac{8RH^2}{\beta}\right)t_0^2+2\alpha R$ yields the desired result. □

Proof of Lemma 5.4. Taking a conditional expectation from both sides of (5.8) conditioned on $\mathcal{F}^{t-1}$, we get

$$
\begin{aligned}
&\mathbb{E}\big(\|\mathbf{Q}(t+t_0)\|_2^2+\|\mathbf{H}(t+t_0)\|_2^2\,|\mathcal{F}^{t-1}\big)-\|\mathbf{Q}(t)\|_2^2-\|\mathbf{H}(t)\|_2^2\\
&\le 2\mathbb{E}\left(V\sum_{\tau=t}^{t+t_0-1}f^{\tau-1}(\mu)\,|\mathcal{F}^{t-1}\right)+2\sum_{i=1}^{L}Q_i(t)\mathbb{E}\left(\sum_{\tau=t}^{t+t_0-1}g_i^{\tau-1}(\mu)\right)\\
&\quad+2\sum_{j=1}^{M}H_j(t)\mathbb{E}\left(\sum_{\tau=t}^{t+t_0-1}(\left\langle h_j^{\tau-1},\mu\right\rangle-b_j)\right)+C_{V,\alpha,t_0},
\end{aligned}
\tag{23}
$$

where we use the following two facts: (1) The multipliers $\mathbf{Q}(t),\ \mathbf{H}(t)\in\mathcal{F}^{t-1}$. (2) The functions g_i^τ and h_i^τ are independent of system history $\mathcal{F}^{t-1}$ and thus the conditional expectation equals the expectation.

Note that by definition, $f^t(\mu)=f(\mu,\xi^t)$, and according to the notation in (7),

$$
\mathbb{E}\left(V\sum_{\tau=t}^{t+t_0-1}f^{\tau-1}(\mu)\,|\mathcal{F}^{t-1}\right)=\mathbb{E}\left(V\mathbb{E}_\xi\left[\sum_{\tau=t}^{t+t_0-1}f^{\tau-1}(\mu)\right]\,\middle|\,\mathcal{F}^{t-1}\right)=Vt_0\mathbb{E}\left(\overline{f}^{(t,t_0)}(\mu)\,|\mathcal{F}^{t-1}\right).
$$

Furthermore,

$$
\mathbb{E}\left(\sum_{\tau=t}^{t+t_0-1}g_i^{\tau-1}(\mu)\right)=t_0\overline{g}_i(\mu),\ \mathbb{E}\left(\sum_{\tau=t}^{t+t_0-1}\left\langle h_j^{\tau-1},\mu\right\rangle\right)=t_0\overline{h}_j(\mu).
$$

Substituting these two relations into (23), we get

$$\mathbb{E}\big(\|\mathbf{Q}(t+t_0)\|_2^2 + \|\mathbf{H}(t+t_0)\|_2^2 \,|\mathcal{F}^{t-1}\big) - \|\mathbf{Q}(t)\|_2^2 - \|\mathbf{H}(t)\|_2^2$$
$$\leq 2Vt_0\mathbb{E}\left(\overline{f}^{(t,t_0)}(\mu) + \sum_{i=1}^{L}\frac{Q_i(t)}{V}\overline{g}_i(\mu) + \sum_{j=1}^{M}\frac{H_j(t)}{V}(\overline{h}_j(\mu)-b_j)\Big|\mathcal{F}^{t-1}\right) + C_{V,\alpha,t_0}. \quad (24)$$

The key, as is mentioned in the proof outline, is to realize that

$$q^{(t,t_0)}\big(\frac{\mathbf{Q}(t)}{V},\ \frac{\mathbf{H}(t)}{V}\big) = \min_{\mu\in\Delta}\ \overline{f}^{(t,t_0)}(\mu) + \sum_{i=1}^{L}\frac{Q_i(t)}{V}\overline{g}_i(\mu) + \sum_{j=1}^{M}\frac{H_j(t)}{V}(\overline{h}_j(\mu)-b_j),$$

where $q^{(t,t_0)}$ is the Lagrangian dual function defined in (8) with dual variables $(\frac{\mathbf{Q}(t)}{V},\ \frac{\mathbf{H}(t)}{V})$. This implies if we choose $\mu=\mu_0$ in (24) as one of the solutions to the above problem, then, we can transform the bound (24) to (14) and finish the proof. □

5.3.4 Proof of Lemma 5.5. We take $t_0=\sqrt{T}$ and by SELM (Assumption 2.2), there exists a solution to the maximization problem

$$\Lambda^* := \mathrm{argmax}_{\lambda,\eta}\ q^{(t,t_0)}(\lambda,\eta).$$

Let $(\lambda^*,\ \eta^*)$ be one of the solutions to this problem. Recall that we define $\overline{c}$ to be the minimum over all c_0's and define $\overline{\ell}$ to be the maximum over all ℓ_0's in Lemma 2.3 with $t=0,1,2,\cdots,T-1$ and $k=\sqrt{T}$. If $\mathrm{dist}\left((\frac{\mathbf{Q}(t)}{V},\ \frac{\mathbf{H}(t)}{V}),\ \Lambda^*\right)\geq\overline{\ell}$, then, by Lemma 2.3 we have

$$\begin{aligned}
&q^{(t,t_0)}\big(\frac{\mathbf{Q}(t)}{V},\ \frac{\mathbf{H}(t)}{V}\big)\\
=&q^{(t,t_0)}\big(\frac{\mathbf{Q}(t)}{V},\ \frac{\mathbf{H}(t)}{V}\big) - q^{(t,t_0)}(\lambda^*,\ \eta^*) + q^{(t,t_0)}(\lambda^*,\ \eta^*)\\
\leq&-\overline{c}\cdot\mathrm{dist}\left((\frac{\mathbf{Q}(t)}{V},\ \frac{\mathbf{H}(t)}{V}),\ \Lambda^*\right) + q^{(t,t_0)}(\lambda^*,\ \eta^*)\\
\leq&-\overline{c}\cdot\mathrm{dist}\left((\frac{\mathbf{Q}(t)}{V},\ \frac{\mathbf{H}(t)}{V}),\ \Lambda^*\right) + \overline{f}^{(t,t_0)}(\mu_0)\\
\leq&-\overline{c}\Big\|\big(\frac{\mathbf{Q}(t)}{V},\ \frac{\mathbf{H}(t)}{V}\big)\Big\|_2 + \overline{c}B + F,
\end{aligned}$$

where the first inequality follows from Lemma 2.3, the second inequality follows from choosing μ_0 as the solution to the following problem

$$\min_{\mu\in\Delta}\overline{f}^{t,t_0}(\mu)\ s.t.\ \overline{\mathbf{g}}(\mu)\leq 0,\ \overline{\mathbf{h}}(\mu)=\mathbf{b},$$

and using weak duality. The third inequality follows from triangle inequality and the boundedness of Lagrange multipliers $\max_{[\lambda,\mu]\in\mathcal{V}^*}\|[\lambda,\mu]\|_2\leq B$.

On the other hand, if dist $\Big(\big(\frac{\mathbf{Q}(t)}{V},\ \frac{\mathbf{H}(t)}{V}\big),\ \Lambda^*\Big) < \overline{\ell}$, then, one has

$$
\begin{aligned}
&q^{(t,t_0)}\big(\frac{\mathbf{Q}(t)}{V},\ \frac{\mathbf{H}(t)}{V}\big)\\
=&\min_{\mu\in\Delta}\ \overline{f}^{(t,t_0)}(\mu)+\sum_{i=1}^{L}\frac{Q_i(t)}{V}\overline{g}_i(\mu)+\sum_{j=1}^{M}\frac{H_j(t)}{V}(\overline{h}_j(\mu)-b_j)\\
=&\min_{\mu\in\Delta}\ \overline{f}^{(t,t_0)}(\mu)+\sum_{i=1}^{L}\Big(\lambda_i^*\overline{g}_i(\mu)+\Big\langle\frac{Q_i(t)}{V}-\lambda_i^*,\overline{g}_i(\mu)\Big\rangle\Big)+\sum_{j=1}^{M}\Big(\mu_j^*\overline{h}_j(\mu)-b_j+\Big\langle\frac{H_j(t)}{V}-\mu_j^*,\overline{h}_j(\mu)\Big\rangle\Big)\\
\le&q^{(t,t_0)}(\lambda^*,\ \eta^*)+\overline{\ell}\Big(G+\sqrt{\frac{2RH^2}{\beta}}\Big)\le F+\overline{\ell}\Big(G+\sqrt{\frac{2RH^2}{\beta}}\Big),
\end{aligned}
$$

where we choose (λ^*,μ^*) to be a point in Λ^* closest to $\big(\frac{\mathbf{Q}(t)}{V},\ \frac{\mathbf{H}(t)}{V}\big)$, the first inequality follows from

$$
\begin{aligned}
\sum_{i=1}^{L}\Big\langle\frac{Q_i(t)}{V}-\lambda_i^*,\overline{g}_i(\mu)\Big\rangle&\le\|\mathbf{Q}(t)/V-\lambda^*\|_2\|\overline{\mathbf{g}}(\mu)\|_2\le G\overline{\ell}\\
\sum_{j=1}^{M}\Big\langle\frac{H_j(t)}{V}-\mu_j^*,\overline{h}_j(\mu)\Big\rangle&\le\|\mathbf{H}(t)/V-\mu^*\|_2\|\overline{\mathbf{h}}(\mu)\|_2\le\sqrt{\frac{2RH^2}{\beta}}\overline{\ell},
\end{aligned}
$$

and the second inequality follows from weak duality. Overall, we finish the proof.

5.3.5 Proof of Lemma 5.6. The proof Lemma 5.6 is based on Lemma 5.5 and a general drift bound. First, substituting Lemma 5.5 into (14) in Lemma 5.4, we have

$$
\begin{aligned}
&\mathbb{E}\big(\|\mathbf{Q}(t+t_0)\|_2^2+\|\mathbf{H}(t+t_0)\|_2^2\ |\mathcal{F}^{t-1}\big)-\|\mathbf{Q}(t)\|_2^2-\|\mathbf{H}(t)\|_2^2\\
&\qquad\le C_{V,\alpha,t_0}+2\Big(F+\overline{\ell}(G+\sqrt{2RH^2/\beta}+\overline{c})+\overline{c}B\Big)Vt_0-2\overline{c}t_0\Big\|\big(\mathbf{Q}(t),\ \mathbf{H}(t)\big)\Big\|_2. \qquad (25)
\end{aligned}
$$

This bound is the key to our analysis. Intuitively, if $\|\big(\mathbf{Q}(t),\ \mathbf{H}(t)\big)\|_2$ is very large at certain time slot t, then, $\|\big(\mathbf{Q}(t+t_0),\ \mathbf{H}(t+t_0)\big)\|_2$ becomes very small. Since $\|\big(\mathbf{Q}(t+t_0),\ \mathbf{H}(t+t_0)\big)\|_2$ is nonnegative, this means $\|\big(\mathbf{Q}(t),\ \mathbf{H}(t)\big)\|_2$ cannot be too large to start with. To transform this intuition into a uniform bound on $\big(\mathbf{Q}(t),\ \mathbf{H}(t)\big)$ over all time slots, we invoke the following drift lemma:

Lemma 5.9 (Lemma 5 of [26]). *Let $\{Z(t), t\ge 1\}$ be a discrete time stochastic process adapted to a filtration $\{\mathcal{F}(t), t\ge 1\}$ with $Z(0)=0$ and $\mathcal{F}(0)=\{\emptyset,\Omega\}$. Suppose there exist integer $t_0>0$, real constants $\theta\in\mathbb{R}$, $\delta_{\max}>0$ and $0<\zeta\le\delta_{\max}$ such that*

$$
|Z(t+1)-Z(t)|\le\delta_{\max}, \qquad (26)
$$

$$
\mathbb{E}[Z(t+t_0)-Z(t)|\mathcal{F}(t)]\le\begin{cases}t_0\delta_{\max}, & \text{if } Z(t)<\theta\\ -t_0\zeta, & \text{if } Z(t)\ge\theta\end{cases}. \qquad (27)
$$

hold for all $t\in\{0,1,2,\ldots\}$. Then, $\mathbb{E}[Z(t)]\le\theta+t_0\frac{4\delta_{\max}^2}{\zeta}\log\big[\frac{8\delta_{\max}^2}{\zeta^2}\big]$, $\forall t\in\{0,1,2,\ldots\}$.

To apply this lemma, we set $Z(t)=\|\big(\mathbf{Q}(t),\ \mathbf{H}(t)\big)\|_2$ and check conditions (26) and (27), for which we detail below:

Proof of Lemma 5.6. For condition (26), we have

$$\begin{aligned}
&\left|\|\big(\mathbf{Q}(t+1),\ \mathbf{H}(t+1)\big)\|_2 - \|\big(\mathbf{Q}(t),\ \mathbf{H}(t)\big)\|_2\right| \\
\leq &\|\big(\mathbf{Q}(t+1)-\mathbf{Q}(t),\ \mathbf{H}(t+1)-\mathbf{H}(t)\big)\|_2 \\
\leq &\sqrt{\sum_{i=1}^{L}(\tilde{g}_i^t)^2} + \sqrt{\sum_{j=1}^{M}(\left\langle h_j^t, \mu^t\right\rangle - b_j)^2} \\
\leq &2\Big(G+\sqrt{\frac{2RD_2^2}{\beta}}\Big) + \sqrt{\frac{8RH^2}{\beta}}.
\end{aligned}$$

On the other hand, for condition (27) we start from (25). Suppose

$$\Big\|\big(\mathbf{Q}(t),\ \mathbf{H}(t)\big)\Big\|_2 \geq \frac{C_{V,\alpha,t_0} + 2\Big(F+\overline{\ell}(G+\sqrt{2RH^2/\beta}+\overline{c})+\overline{c}B\Big)Vt_0}{\overline{c}t_0},$$

then, we can derive from (25) that

$$\begin{aligned}
&\mathbb{E}\big(\|\mathbf{Q}(t+t_0)\|_2^2 + \|\mathbf{H}(t+t_0)\|_2^2\ |\mathcal{F}^{t-1}\big) - \|\mathbf{Q}(t)\|_2^2 - \|\mathbf{H}(t)\|_2^2 \\
\leq &-\overline{c}t_0\|\big(\mathbf{Q}(t),\ \mathbf{H}(t)\big)\|_2 \leq -\overline{c}t_0\|\big(\mathbf{Q}(t),\ \mathbf{H}(t)\big)\|_2 + \frac{\overline{c}^2t_0^2}{4},
\end{aligned}$$

which implies

$$\mathbb{E}\Big(\|\mathbf{Q}(t+t_0)\|_2^2 + \|\mathbf{H}(t+t_0)\|_2^2\ \Big|\mathcal{F}^{t-1}\Big) \leq \Big(\|\big(\mathbf{Q}(t),\ \mathbf{H}(t)\big)\|_2 - \frac{\overline{c}t_0}{2}\Big)^2.$$

Taking a square root from both sides and by Jensen's inequality,

$$\mathbb{E}\Big(\Big\|\big(\mathbf{Q}(t+t_0),\ \mathbf{H}(t+t_0)\big)\Big\|_2\ \Big|\mathcal{F}^{t-1}\Big) \leq \Big\|\big(\mathbf{Q}(t),\ \mathbf{H}(t)\big)\Big\|_2 - \frac{\overline{c}t_0}{2}.$$

Overall, by Lemma 5.9, we obtain

$$\begin{aligned}
\mathbb{E}\Big(\Big\|\big(\mathbf{Q}(t),\ \mathbf{H}(t)\big)\Big\|_2\Big) \leq &\frac{C_{V,\alpha,t_0} + 2\Big(F+\overline{\ell}(G+\sqrt{8RH^2/\beta}+\overline{c})+\overline{c}B\Big)Vt_0}{\overline{c}t_0} \\
&+ \frac{8t_0\Big(2(G+\sqrt{2RD_2^2/\beta})+\sqrt{8RH^2/\beta}\Big)^2}{\overline{c}} \cdot \log\left(\frac{32\left(2(G+\sqrt{\frac{2RD_2^2}{\beta}})+\sqrt{\frac{8RH^2}{\beta}}\right)^2}{\overline{c}^2}\right).
\end{aligned}$$

Taking $V=\sqrt{T}, \alpha = T$ and $t_0 = \sqrt{T}$ and recalling the definition of C_{V,α,t_0} yields:

$$\mathbb{E}\Big(\Big\|\big(\mathbf{Q}(t),\ \mathbf{H}(t)\big)\Big\|_2\Big) \leq C' + C''\sqrt{T}$$

where $C' := \frac{2}{\overline{c}}\left(\frac{4RH^2}{\beta} + G^2 + \frac{2RD_2^2}{\beta} + \frac{D_1^2}{2\beta}\right)$ and $C'' := \frac{2}{\overline{c}}\Big(2F + \frac{3}{2}G^2 + \frac{2RD_2^2}{\beta} + \frac{8RH^2}{\beta} + R + \overline{\ell}(G+\sqrt{8RH^2/\beta}+\overline{c}) + \overline{c}B + 4\big(2(G+\sqrt{2RD_2^2/\beta})+\sqrt{8RH^2/\beta}\big)^2 \log\big(\frac{32(2(G+\sqrt{\frac{2RD_2^2}{\beta}})+\sqrt{\frac{8RH^2}{\beta}})^2}{\overline{c}^2}\big)\Big)$ are constants. □

6 CONCLUSIONS

This paper proposes a new primal-dual online mirror descent framework for stochastic constrained online learning problem. We introduce a new sequential existence of Lagrange multipliers condition, which is shown to be strictly weaker than the Slater condition, and prove that the proposed algorithm enjoys a $O(\sqrt{T})$ expected regret and constraint violations. We also obtain an almost dimension free result in the special case when the decision set is a probability simplex. Simulation experiments demonstrate the performance of the proposed algorithm.

ACKNOWLEDGMENTS

Michael J. Neely was partially supported by the National Science Foundation under grant CCF-1718477.

REFERENCES

[1] Deepak Agarwal, Souvik Ghosh, Kai Wei, and Siyu You. 2014. Budget pacing for targeted online advertisements at LinkedIn. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1613–1619.
[2] Dimitri P Bertsekas. 1999. *Nonlinear programming*. Athena scientific Belmont.
[3] Stephen Boyd and Lieven Vandenberghe. 2004. *Convex optimization*. Cambridge university press.
[4] Nicolò Cesa-Bianchi, Philip M Long, and Manfred K Warmuth. 1996. Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Transactions on Neural Networks* 7, 3 (1996), 604–619.
[5] Tianyi Chen and Georgios B Giannakis. 2019. Bandit convex optimization for scalable and dynamic IoT management. *IEEE Internet of Things Journal* 6, 1 (2019), 1276–1286.
[6] Wei Deng, Ming-Jun Lai, Zhimin Peng, and Wotao Yin. 2017. Parallel multi-block ADMM with o (1/k) convergence. *Journal of Scientific Computing* 71, 2 (2017), 712–736.
[7] Anshul Gandhi, Mor Harchol-Balter, and Michael A Kozuch. 2012. Are sleep states effective in data centers?. In *2012 International Green Computing Conference (IGCC)*. IEEE, 1–10.
[8] Jacques Gauvin. 1977. A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming. *Mathematical Programming* 12, 1 (1977), 136–138.
[9] Geoffrey J Gordon. 1999. Regret bounds for prediction problems. In *Proceeding of Conference on Learning Theory (COLT)*.
[10] Elad Hazan. 2016. Introduction to online convex optimization. *Foundations and Trends in Optimization* 2, 3–4 (2016), 157–325.
[11] Rodolphe Jenatton, Jim Huang, and Cédric Archambeau. 2016. Adaptive Algorithms for Online Convex Optimization with Long-term Constraints. In *Proceedings of International Conference on Machine Learning (ICML)*.
[12] Nikolaos Liakopoulos, Apostolos Destounis, Georgios Paschos, Thrasyvoulos Spyropoulos, and Panayotis Mertikopoulos. 2019. Cautious Regret Minimization: Online Optimization with Long-Term Budget Constraints. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, Long Beach, California, USA, 3944–3952. http://proceedings.mlr.press/v97/liakopoulos19a.html
[13] Mehrdad Mahdavi, Rong Jin, and Tianbao Yang. 2012. Trading regret for efficiency: online convex optimization with long term constraints. *Journal of Machine Learning Research* 13, 1 (2012), 2503–2528.
[14] Shie Mannor, John N Tsitsiklis, and Jia Yuan Yu. 2009. Online learning with sample path constraints. *Journal of Machine Learning Research* 10 (March 2009), 569–590.
[15] Angelia Nedić and Asuman Ozdaglar. 2009. Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM Journal on Optimization* 19, 4 (2009), 1757–1780.
[16] Michael J Neely. 2014. A simple convergence time analysis of drift-plus-penalty for stochastic optimization and convex programs. *arXiv preprint arXiv:1412.0791* (2014).
[17] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. 2009. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization* 19, 4 (2009), 1574–1609.
[18] V Hien Nguyen, J-J Strodiot, and Robert Mifflin. 1980. On conditions to have bounded multipliers in locally Lipschitz programming. *Mathematical Programming* 18, 1 (1980), 100–106.
[19] Alexander A Titov, Fedor S Stonyakin, Alexander V Gasnikov, and Mohammad S Alkousa. 2018. Mirror Descent and Constrained Online Optimization Problems. In *International Conference on Optimization and Applications*. Springer, 64–78.

[20] Paul Tseng. 2005. On accelerated proximal gradient methods for convex-concave optimization. *MIT Technical Report* (2005).

[21] Paul Tseng. 2010. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming* 125, 2 (2010), 263–295.

[22] Xiaohan Wei, Hao Yu, Qing Ling, and Michael Neely. 2018. Solving Non-smooth Constrained Programs with Lower Complexity than $O(1/\varepsilon)$: A Primal-Dual Homotopy Smoothing Approach. In *Advances in Neural Information Processing Systems*. 3999–4009.

[23] Yi Xu, Mingrui Liu, Qihang Lin, and Tianbao Yang. 2017. ADMM without a Fixed Penalty Parameter: Faster Convergence with New Adaptive Penalization. In *Advances in Neural Information Processing Systems*. 1267–1277.

[24] Tianbao Yang and Qihang Lin. 2015. Rsg: Beating subgradient method without smoothness and strong convexity. *arXiv preprint arXiv:1512.03107* (2015).

[25] Xinlei Yi, Xiuxian Li, Lihua Xie, and Karl H Johansson. 2019. Distributed Online Convex Optimization with Time-Varying Coupled Inequality Constraints. *arXiv preprint arXiv:1903.04277* (2019).

[26] Hao Yu, Michael Neely, and Xiaohan Wei. 2017. Online convex optimization with stochastic constraints. In *Advances in Neural Information Processing Systems*. 1428–1438.

[27] Hao Yu and Michael J Neely. 2016. A Low Complexity Algorithm with $O(\sqrt{T})$ Regret and Finite Constraint Violations for Online Convex Optimization with Long Term Constraints. *arXiv preprint arXiv:1604.02218* (2016).

[28] Hao Yu and Michael J Neely. 2017. A Simple Parallel Algorithm with an $O(1/t)$ Convergence Rate for General Convex Programs. *SIAM Journal on Optimization* 27, 2 (2017), 759–783.

[29] Jianjun Yuan and Andrew Lamperski. 2018. Online convex optimization for cumulative constraints. In *Advances in Neural Information Processing Systems*. 6140–6149.

[30] Alp Yurtsever, Quoc Tran Dinh, and Volkan Cevher. 2015. A universal primal-dual convex optimization framework. In *Advances in Neural Information Processing Systems*. 3150–3158.

[31] Martin Zinkevich. 2003. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of International Conference on Machine Learning (ICML)*.

A SUPPLEMENT

A.1 The pushback property of Bregman divergences

In this section, we prove the following key property of the Bregman divergence:

LEMMA A.1. *Let $f : C \to \mathbb{R}$ be a convex function. Fix $\alpha > 0, y \in \Delta^o$. Suppose $x^* \in argmin_{x\in\Delta} f(x) + \alpha D(x, y)$ and $x^* \in \Delta^o$, then, for any $z \in \Delta$,*

$$f(x^*) + \alpha D(x^*, y) \le f(z) + \alpha D(z, y) - \alpha D(z, x^*).$$

PROOF OF LEMMA A.1. First of all, we recall the following known facts about convex functions and their subgradients whose proofs can be found, for example, in [2]:

- The set $\partial f(x)$ is non-empty for any $x \in \text{int}(C)$.
- For any bounded subset $\mathcal{X} \subseteq \text{int}(C)$, the union $\cup_{x\in\mathcal{X}}\partial f(x)$ is bounded.

By definition of Bregman divergence, we have for any $x,\ y \in \Delta^o$,

$$D(x, y) = \omega(x) - \omega(y) - \langle \nabla\omega(y), x - y\rangle,$$

and

$$\nabla_x D(x, y) = \nabla\omega(x) - \nabla\omega(y).$$

Now, we claim the following optimality condition:
Claim 1: For any $z \in \Delta$, there exists a $\nabla f(x^*) \in \partial f(x^*)$ such that following holds:

$$\langle \nabla f(x^*) + \alpha\nabla\omega(x^*) - \alpha\nabla\omega(y), z - x^*\rangle \ge 0.$$

PROOF OF CLAIM 1. Fix a constant $h \in (0, 1)$. Since Δ is a convex set, it follows $(1 - h)x^* + hz \in \Delta$. Thus, by the fact that x^* is a minimizer:

$$\begin{aligned}
&f(x^*) + \alpha D(x^*, y)\\
\le& f((1 - h)x^* + hz) + \alpha D((1 - h)x^* + hz, y)\\
=& f((1 - h)x^* + hz) + \alpha \left(D(x^*, y) + \langle \nabla D(x^*, y), h(z - x^*)\rangle + o(h)\right)\\
=& f((1 - h)x^* + hz) + \alpha D(x^*, y)\\
&+ \alpha \left(\langle \nabla\omega(x^*) - \nabla\omega(y), h(z - x^*)\rangle + o(h)\right),
\end{aligned}$$

where the first equality follows from the fact that $D(x, z)$ is continuously differentially on the first argument at $x = x^*$ with $o(h)$ representing a high order term such that $\lim_{h\to 0} o(h)/h = 0$, and the second equality follows from the definition of Bregman divergence. Canceling the common term $\alpha D(x^*, y)$ and rearranging the terms give

$$\frac{f((1 - h)x^* + hz) - f(x^*)}{h} \ge -\alpha \langle \nabla\omega(x^*) - \nabla\omega(y), z - x^*\rangle - o(\alpha h)/h. \tag{28}$$

Since f is convex and $(1 - h)x^* + hz \in \text{int}(C),\ \forall h < 1$, we have for any $\nabla f((1 - h)x^* + hz) \in \partial f((1 - h)x^* + hz)$.

$$f(x^*) \ge f((1 - h)x^* + hz) + \langle \nabla f((1 - h)x^* + hz), h(x^* - z)\rangle.$$

Substituting this bound into (28) gives

$$\langle \nabla f((1 - h)x^* + hz), z - x^*\rangle \ge -\alpha \langle \nabla\omega(x^*) - \nabla\omega(y), z - x^*\rangle - o(\alpha h)/h. \tag{29}$$

To this point, consider any sequence $\{h_k\}_{k\ge 0} \subseteq (0, 1)$ such that $\lim_{k\to\infty} h_k = 0$. By the aforementioned property of subgradient, we have the union $\cup_{k\ge 0}\partial f((1 - h_k)x^* + h_k z)$ is bounded. Thus, the sequence $\{\nabla f((1 - h_k)x^* + h_k z)\}_{k\ge 0}$ is bounded, and there exists a subsequence $\{\nabla f((1 - h_{k_\ell})x^* +$

$h_{k_\ell} z)\}_{\ell \geq 0}$ such that $\nabla f((1-h_{k_\ell})x^* + h_{k_\ell} z) \to d$. On the other hand, by definition of subgradient, we have for any $u \in C$,

$$f(u) \geq f((1-h_{k_\ell})x^* + h_{k_\ell} z) + \left\langle \nabla f((1-h_{k_\ell})x^* + h_{k_\ell} z), u - ((1-h_{k_\ell})x^* + h_{k_\ell} z) \right\rangle.$$

Taking the limit $\ell \to \infty$ gives

$$f(u) \geq f(x^*) + \langle d, u - x^* \rangle,$$

where we use the fact that a convex function must be continuous on the interior point x^* of C. This implies that $d \in \partial f(x^*)$. Substituting $\{h_{k_\ell}\}_{\ell \geq 0}$ into (29) and taking the limit finish the proof. □

Thus, by Claim 1, we have there exists a $\nabla f(x^*)$,

$$\begin{aligned}
&\alpha(D(z,y) - D(z,x^*)) \\
=&\alpha\left(\omega(z) - \omega(y) - \langle \nabla\omega(y), z-y\rangle\right) - \alpha\left(\omega(z) - \omega(x^*) - \langle \nabla\omega(x^*), z-x^*\rangle\right) \\
=&\alpha\left(\omega(x^*) - \omega(y) + \langle \nabla\omega(x^*), z-x^*\rangle - \langle \nabla\omega(y), z-y\rangle\right) \\
=&\alpha\left(\omega(x^*) - \omega(y) + \langle \nabla f(x^*)/\alpha + \nabla\omega(x^*) - \nabla\omega(y), z-x^*\rangle - \langle \nabla\omega(y), z-y\rangle\right) \\
&- \langle \nabla f(x^*), z-x^*\rangle + \alpha\langle \nabla\omega(y), z-x^*\rangle \\
\geq&\alpha\left(\omega(x^*) - \omega(y) - \langle \nabla\omega(y), x^*-y\rangle\right) - \langle \nabla f(x^*), z-x^*\rangle \\
=&\alpha D(x^*, y) - \langle \nabla f(x^*), z-x^*\rangle \\
\geq&\alpha D(x^*, y) + f(x^*) - f(z),
\end{aligned}$$

where third equality follows from adding and subtracting the term $\langle \nabla f(x^*), z-x^*\rangle - \alpha\langle \nabla\omega(y), z-x^*\rangle$, the first inequality follows from the aforementioned optimality condition and the last inequality follows from convexity that $f(z) \geq f(x^*) + \langle \nabla f(x^*), z-x^*\rangle$. Rearranging the terms yields the desired result. □

A.2 SELM and constraint qualifications

A.2.1 Slater condition implies SELM. The SELM assumption is actually implied by the Slater condition. More specifically, Slater condition considers the scenario where there is no equality constraint and there exists a $\mu \in \Delta$ such that $\overline{g}_i(\mu) < 0,\ \forall i \in \{1, 2, \cdots, L\}$. First of all, it is well-known that the Slater condition is sufficient for the existence of a dual optimal solution (see, for example, [2]). Furthermore, the following lemma, which is essentially the same as Lemma 1 of [15], implies that the set of dual optimal solutions is also bounded:

Lemma A.2. *Consider the convex program (7) without equality constraints $\overline{\mathbf{h}}(\mu) = 0$, and define the Lagrange dual function $q^{(t,k)}(\lambda) = \inf_{\mu\in\Delta}\left\{\overline{f}^{(t,k)}(\mu) + \sum_{i=1}^m \lambda_i \overline{g}_i(\mu)\right\}$. Suppose there exists $\widetilde{\mu} \in \Delta$ such that $\overline{g}_i(\widetilde{\mu}) \leq -\varepsilon,\ \forall i \in \{1, 2, \cdots, L\}$ for some positive constant $\varepsilon > 0$. Then, the level set $\mathcal{V}_{\bar{\lambda}} = \left\{\lambda_1, \lambda_2, \cdots, \lambda_L \geq 0,\ q^{(t,k)}(\lambda) \geq q^{(t,k)}(\bar{\lambda})\right\}$ is bounded for any nonnegative $\bar{\lambda}$. Furthermore, we have*

$$\max_{\lambda\in\mathcal{V}_{\bar{\lambda}}} \|\lambda\|_2 \leq \varepsilon^{-1}\left(\overline{f}^{(t,k)}(\widetilde{\mu}) - q^{(t,k)}(\bar{\lambda})\right).$$

Note that since $|f^t(\mu)|$ is bounded by some constant $F > 0$ as stated in Assumption 2.1. Taking $\bar{\lambda} = \lambda^*$ for any optimal dual solution λ^*, and notice that $\overline{f}^{(t,k)}(\widetilde{\mu}) \leq F$, and

$$q^{(t,k)}(\lambda^*) \geq \min_{\mu\in\Delta} \overline{f}^{(t,k)}(\mu) \geq -F,$$

the above lemma readily implies $\max_{\lambda\in\mathcal{V}^*} \|\lambda\|_2 \leq 2F/\varepsilon$. Thus, Slater condition implies the existence and boundedness of Lagrange multipliers.

A.2.2 SELM is implied by Mangasarian-Fromovitz constraint qualification (MFCQ). In this section, we show SELM is able to handle general equality constraints and thus strictly weaker than the Slater condition. In 1977, J. Gauvin [8] observed that for any constrained convex program, where both the objective and constraint functions are continuously differentiable, the Mangasarian-Fromovitz constraint qualification (MFCQ) condition is in fact equivalent to the boundedness of the KKT set.[1] More specifically, MFCQ is defined as follows:

Definition A.3 (Mangasarian-Fromovitz constraint qualification (MFCQ)). Consider a convex program:

$$\begin{aligned} &\text{minimize}_{x\in\mathbb{R}^d}\ f(x),\\ &\text{subject to}\quad g_i(x)\le 0,\ i\in\{1,2,\cdots,L\},\\ &\qquad\qquad \left\langle h_j, x\right\rangle = b_j,\ j\in\{1,2,\cdots,M\}. \end{aligned} \tag{30}$$

It satisfies MFCQ if (a) The solution to (30) exists. (b) The vectors $\{h_j\}_{j=1}^M$ are linearly independent. (c) For a solution x^* to the program, there exists some $y\in\mathbb{R}^d$ such that $\langle\nabla g_i(x^*), y\rangle < 0,\ \forall i\in I(x^*)$, where $I(x^*)=\{i\mid g_i(x^*)=0\}$.

Theorem A.4 ([8]). *Let x^* be a solution to (30). Consider the Karush-Kuhn-Tucker(KKT) set for the program (30), which is the set $K(x^*)$ of vectors $(\lambda,\eta)\in\mathbb{R}_+^L\times\mathbb{R}^M$ such that the following set of equations holds:*

$$\begin{aligned} -\nabla f(x^*) &= \sum_{i=1}^{L}\lambda_i\nabla g_i(x^*)+\sum_{j=1}^{M}\eta_j\mathbf{h}_j,\\ \lambda\ge 0,\ \lambda_i g_i(x^*)&=0,\ \forall i\in\{1,2,\cdots,M\}. \end{aligned}$$

Then, the set $K(x^*)$ is non-empty and bounded if and only if MFCQ is satisfied for (30).

Note that compared to (30) our program (7) has an extra set constraint $\mu\in\Delta$. The good news is that for the case where Δ is a probability simplex, i.e. it can be written explicitly as $\{\mu\in\mathbb{R}^d:\ \mu_i\ge 0,\ \forall i,\ \sum_{i=1}^d\mu_i=1\}$, applying Theorem A.4, we have the following lemma whose proof is delayed to Section A.5:

Lemma A.5. *Consider the optimization problem (7) for any soecific time slot t and any time period k where Δ is the probability simplex. Suppose (a) The vectors $\left\{\mathbf{1},\ \mathbb{E}\left(h_1^t\right),\ \mathbb{E}\left(h_2^t\right),\ \cdots,\ \mathbb{E}\left(h_M^t\right)\right\}$ are linearly independent. (b) There exists a solution to (7), denoted as μ^*, and a vector $y\in\mathbb{R}^d$ such that $\left\langle\nabla\overline{g}_i(\mu^*), y\right\rangle < 0,\ \forall i\in I(\mu^*)$, where $I(\mu^*)=\{i\mid\overline{g}_i(\mu^*)=0\}$. Then, the set of Lagrange multipliers $\mathcal{V}^*:=argmax_{\lambda\in\mathbb{R}_+^L,\ \eta\in\mathbb{R}^M}q^{(t,k)}(\lambda,\eta)$, where $q^{(t,k)}$ is defined in (8), is non-empty and bounded.*

Remark A.1. *In the case where there is no inequality constraints in (7), lemma A.5 gives a simple objective-irrelevant equivalence condition of SELM that $\left\{\mathbf{1},\ \mathbb{E}\left(h_1^t\right),\ \mathbb{E}\left(h_2^t\right),\ \cdots,\ \mathbb{E}\left(h_M^t\right)\right\}$ are linearly independent, which could be useful for online linear program.*

For general scenarios where Δ is just an arbitrary abstract convex set, we have the following definition of generalized MFCQ following [18]. First, we have the definitions of normal cones and tangent cones:

Definition A.6 (Normal cone). Consider any set $S\subseteq\mathbb{R}^d$, the normal cone of S at any $x\in S$ is

$$N(S,x):=\{g\in\mathbb{R}^d:\ \langle g, x-y\rangle\ge 0,\ \forall y\in\mathbb{R}^d\}.$$

[1] In fact, MFCQ does not require convexity of the constrained programs. Thus, the result in [8] even applies to non-convex programs.

Note that normal cone at $x \in S$ is the subgradient of the indicator function of S, namely $I_S(x)$. To see this, consider any $y \in \mathbb{R}^d$, then, we have g is a subgradient of $I_S(x)$ at x if

$$I_S(y) \geq I_S(x) + \langle g, y - x \rangle, \ \forall y \in \mathbb{R}^d.$$

Note that if $y \notin S$, then $I_S(y) = +\infty$, otherwise, $I_S(y) = I_S(x) = 0$. Thus, $\langle g, x - y \rangle \geq 0$.

Definition A.7 (Tangent cone). Consider any set $S \subseteq \mathbb{R}^d$, the tangent cone of S at any $x \in S$ is

$$T(S, x) := \text{cone}(S - x) = \{\lambda d : \ \lambda \geq 0, \ d \in S - x\},$$

and $S - x = \{y \in \mathbb{R}^d, y = z - x, \ \exists z \in S\}$.

Definition A.8 (Generalized MFCQ). Consider a convex program:

$$\begin{aligned} &\text{minimize}_{x \in S} \ f(x), \\ &\text{subject to} \quad g_i(x) \leq 0, \ i \in \{1, 2, \cdots, L\}, \\ &\qquad\qquad \langle h_j, x \rangle = b_j, \ j \in \{1, 2, \cdots, M\}. \end{aligned} \tag{31}$$

It satisfies the generalized MFCQ if (a) The vectors $\{h_j\}_{j=1}^M$ are linearly independent. (b) For a solution $\mathbf{x}^*$ to the above program, there exists some $y \in \text{int}(T(S, x^*))$ such that $\langle \nabla g_i(x^*), y \rangle < 0, \ \forall i \in I(x^*)$ and any subgradient $\nabla g_i(x^*)$, where $I(x^*) = \{i \mid g_i(\mathbf{x}^*) = 0\}$ and $\text{int}(T(S, x^*))$ denotes the interior of $T(S, x^*)$.

Note that this definition requires the interior of $T(S, x^*)$ to be non-empty, which *does not* work for the case where S is a probability simplex. This is why we have a separate lemma (Lemma A.5). When assuming the interior of $T(S, x^*)$ is non-empty, we have the following theorem:

Theorem A.9 ([18]). *Let x^* be a solution to (31). Consider the Karush-Kuhn-Tucker(KKT) set of the program (31), which is the set $K(x^*)$ of vectors $(\lambda, \eta) \in \mathbb{R}_+^L \times \mathbb{R}^M$ such that the following set of equations holds:*

$$\begin{aligned} &0 \in \partial f(x^*) + \sum_{i=1}^{L} \lambda_i \nabla g_i(x^*) + \sum_{j=1}^{M} \eta_j \mathbf{h}_j + N(S, x^*), \\ &\lambda \geq 0, \ \lambda_i g_i(x^*) = 0, \ \forall i \in \{1, 2, \cdots, M\}. \end{aligned}$$

Then, the set $K(x^*)$ is non-empty and bounded if and only if (30) satisfies the generalized MFCQ.

Applying the above theorem to (7) with $S = \Delta$, we readily get the equivalence condition for the existence and boundedness of Lagrange multipliers for (7) as follows

Corollary A.10. *Consider the optimization problem (7) for any time slot t and any time period k where Δ has an nonempty interior. Suppose (a) The vectors $\left\{\mathbb{E}\left(h_1^t\right), \ \mathbb{E}\left(h_2^t\right), \ \cdots, \ \mathbb{E}\left(h_M^t\right)\right\}$ are linearly independent. (b) There exists a solution to (7), denoted as μ^*, and a vector $y \in int(T(\Delta, \mu^*))$ such that $\left\langle \nabla \overline{g}_i(\mu^*), y \right\rangle < 0, \ \forall i \in I(\mu^*)$, where $I(\mu^*) = \{i \mid \overline{g}_i(\mu^*) = 0\}$. Then, the set of Lagrange multipliers $\mathcal{V}^* := argmax_{\lambda \in \mathbb{R}_+^L, \ \eta \in \mathbb{R}^M} q^{(t,k)}(\lambda, \eta)$, where $q^{(t,k)}$ is defined in (8), is non-empty and bounded.*

A.2.3 SELM implies weak EBC. In this section, we prove a key property of SELM, namely Lemma 2.3, which says SELM implies a weak EBC condition. We restate the lemma as follows, and for simplicity, we omit the subscript t, k on the set $\mathcal{V}^*$ for simplicity:

Lemma A.11. *Suppose Assumption 2.2 holds, then, there exists constants $c_0, \ell_0 > 0$ such that the dual function $q^{(t,k)}(\lambda, \eta)$ defined in (8) satisfies a weak error bound condition, namely, for any $(\lambda^*, \eta^*) \in \mathcal{V}^*$, $q^{(t,k)}(\lambda^*, \eta^*) - q^{(t,k)}(\lambda, \eta) \geq c_0 \cdot dist((\lambda, \eta), \mathcal{V}^*)$ for any (λ, η) such that $dist((\lambda, \eta), \mathcal{V}^*) \geq \ell_0$.*

Proof of Lemma A.11. Since $\mathcal{V}^*$ is bounded, there must exist $\ell_0 > 0$ such that $\mathcal{S}_1 := \{(\lambda, \eta) : dist((\lambda, \eta), \mathcal{V}^*) = l_0\} \neq \emptyset$. Define $\tilde{q} := \sup_{(\lambda,\eta)\in\mathcal{S}_1} q^{(t,k)}(\lambda, \eta)$. Then, since the set $\mathcal{S}_1$ is closed, there exists some constant $c_0 > 0$ such that $q^{(t,k)}(\lambda^*, \eta^*) - \tilde{q} \geq c_0 l_0$. Now, consider any (λ, η) such that $\text{dist}((\lambda, \eta), \mathcal{V}^*) \geq l_0$, and choose $(\lambda^*, \eta^*) \in \mathcal{V}^*$ such that

$$(\lambda^*, \eta^*) = \text{argmin}_{(\lambda_0,\eta_0)\in\mathcal{V}^*} \|(\lambda_0, \eta_0) - (\lambda, \eta)\|_2^2, \tag{32}$$

i.e. $\|(\lambda^*, \eta^*) - (\lambda, \eta)\|_2 = \text{dist}((\lambda, \eta), \mathcal{V}^*) \geq l_0$.

Choose $\theta := \frac{l_0}{\|(\lambda^*,\eta^*)-(\lambda,\eta)\|_2}$. Note that $0 < \theta \leq 1$. Let $(\tilde{\lambda}, \tilde{\eta}) := ((1-\theta)\lambda^* + \theta\lambda, (1-\theta)\eta^* + \theta\eta)$. The next claim shows that $(\tilde{\lambda}, \tilde{\eta}) \in \mathcal{S}_1$.

Claim 1: $(\tilde{\lambda}, \tilde{\eta}) \in \mathcal{S}_1$.

Proof. It is easy to verify that $\|(\tilde{\lambda}, \tilde{\eta}) - (\lambda^*, \eta^*)\|_2 = l_0$. To prove this claim, it suffices to show that

$$(\lambda^*, \eta^*) = \text{argmin}_{(\lambda_0,\eta_0)\in\mathcal{V}^*} \|(\tilde{\lambda}, \tilde{\eta}) - (\lambda_0, \eta_0)\|_2^2.$$

To see this, suppose on the contrary, there exists $(\overline{\lambda}, \overline{\eta}) \neq (\lambda^*, \eta^*)$ such that $(\overline{\lambda}, \overline{\eta})$ attains the above minimum, then, by the strong convexity of the square norm function and convexity of the set $\mathcal{V}^*$, the solution is unique, and it follows

$$\begin{aligned}\|(\overline{\lambda}, \overline{\eta}) - (\lambda, \eta)\|_2 &\leq \|(\overline{\lambda}, \overline{\eta}) - (\lambda', \eta')\|_2 + \|(\lambda', \eta') - (\lambda, \eta)\|_2 \\ &< \|(\lambda^*, \eta^*) - (\lambda', \eta')\|_2 + \|(\lambda', \eta') - (\lambda, \eta)\|_2 = \|(\lambda^*, \eta^*) - (\lambda, \eta)\|_2,\end{aligned}$$

where the strict inequality follows from the aforementioned strong convexity and the last equality follows from the fact that $(\lambda', \eta') \in \mathcal{L}$. However, this implies $\overline{\lambda}, \overline{\eta}$ is of smaller distance to (λ, η) contradicting (32). □

By the concavity of $q^{(t,k)}(\lambda, \eta)$, we have,

$$q^{(t,k)}((1-\theta)\lambda^* + \theta\lambda, (1-\theta)\eta^* + \theta\eta) \geq (1-\theta)q^{(t,k)}(\lambda^*, \eta^*) + \theta q^{(t,k)}(\lambda, \eta). \tag{33}$$

This further implies that

$$\begin{aligned}&q^{(t,k)}(\tilde{\lambda}, \tilde{\eta}) \geq (1-\theta)q^{(t,k)}(\lambda^*, \eta^*) + \theta q^{(t,k)}(\lambda, \eta) \\ \Rightarrow\ &q^{(t,k)}(\tilde{\lambda}, \tilde{\eta}) - q^{(t,k)}(\lambda^*, \eta^*) \geq \theta(q^{(t,k)}(\lambda, \eta) - q^{(t,k)}(\lambda^*, \eta^*)).\end{aligned}$$

Recalling the definition of $\tilde{q} = \sup_{(\lambda,\eta)\in\mathcal{S}_1} q^{(t,k)}(\lambda, \eta)$ and that $(\tilde{\lambda}, \tilde{\eta}) \in \mathcal{S}_1$ by Claim 1, we have

$$\begin{aligned}&\tilde{q} - q^{(t,k)}(\lambda^*, \eta^*) \geq \theta(q^{(t,k)}(\lambda, \eta) - q^{(t,k)}(\lambda^*, \eta^*)) \\ \Rightarrow\ &q^{(t,k)}(\lambda^*, \eta^*) - q^{(t,k)}(\lambda, \eta) \geq \frac{1}{\theta}(q^{(t,k)}(\lambda^*, \eta^*) - \tilde{q}).\end{aligned}$$

Recalling that $q^{(t,k)}(\lambda^*, \eta^*) - \tilde{q} \geq c_0 l_0$ and $\theta = \frac{l_0}{\|(\lambda^*,\eta^*)-(\lambda,\eta)\|_2}$, we have

$$q^{(t,k)}(\lambda^*, \eta^*) - q^{(t,k)}(\lambda, \eta) \geq c_0 \text{dist}((\lambda, \eta), \mathcal{V}^*),$$

and we finish the proof. □

A.3 On the relation between weak EBC and classical EBC

Recall that the classical EBC, which has been shown to accelerate the convergence rate solving unconstrained and constrained programs [21–24], is stated as follows:

Definition A.12. Let $F(\mathbf{x})$ be a convex function over $\mathbf{x} \in \mathcal{X}$. Suppose $\Lambda^* := \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x})$ is non-empty. The function $F(\mathbf{x})$ is said to satisfy the error bound condition (EBC) with parameters $\beta \in (0, 1]$, $\delta > 0$ and $C_\delta > 0$ if for any $\mathbf{x} \in \mathcal{S}_\delta$, the δ-sublevel set defined as $\{\mathbf{x} \in \mathcal{X} \mid F(\mathbf{x}) - F(\mathbf{x}^*) \leq \delta,\ \mathbf{x}^* \in \Lambda^*\}$,

$$\operatorname{dist}(\mathbf{x}, \Lambda^*) \leq C_\delta (F(\mathbf{x}) - F(\mathbf{x}^*))^\beta, \tag{34}$$

where C_δ is a positive constant possibly depending on δ. In particular, when $\beta = 1/2$, $F(\mathbf{x})$ is said to be locally quadratic and when $\beta = 1$, it is said to be locally linear.

The following lemma shows that if the dual function further satisfies classical EBC, then, we can show that weak EBC holds with computable constants $\ell_0, c_0 > 0$.

Lemma A.13. *Suppose Assumption 2.2 holds, the dual function $q^{(t,k)}(\lambda, \eta)$ is continuous and satisfies an EBC as is defined in Definition A.12, then, one has for any $(\lambda, \eta) \in \mathbb{R}^L_+ \times \mathbb{R}^M$ such that $dist((\lambda, \eta), \mathcal{V}^*) \geq C_\delta \delta^\beta$,*

$$dist((\lambda, \eta), \mathcal{V}^*) \leq C_\delta \delta^{\beta-1} (q^{(t,k)}(\lambda^*, \eta^*) - q^{(t,k)}(\lambda, \eta)),$$

for any $\lambda \in \mathbb{R}^L_+,\ \eta \in \mathbb{R}^M$.

The proof of this lemma is delayed to Section A.5.

A.4 Proof of Theorem 3.2

In this section, we present the proof for Theorem 3.2. The proof takes into account the fact that Δ is the probability simplex and the effect of pull-away operation $\tilde{\mu}^{t-1} = (1-\theta)\mu^{t-1} + \frac{\theta}{d}\mathbf{1}$. Note that in this probability simplex case, we have $\sup_{\mu_1,\mu_2\in\Delta} \|\mu_1 - \mu_2\|_1 \leq 1$, which will be used to replace the frequently used relation $\sup_{\mu_1,\mu_2\in\Delta} \|\mu_1 - \mu_2\| \leq \sqrt{\frac{2R}{\beta}}$ in the proof for general cases. Note further that when Δ is the probability simplex and $D(\mu_1, \mu_2)$ is chosen to be K-L divergence, we do not have a uniform bound R such that $\sup_{\mu_1,\mu_2\in\Delta} D(\mu_1, \mu_2) \leq R$. Fortunately, our analysis does not need such a uniform bound but instead uses a bound on $D(\mu_1, \tilde{\mu}_2)$ where $\tilde{\mu}_2$ is in the form of $\tilde{\mu}^t$ specified in Algorithm 2.

The following lemma bounds the difference between $D(\mu, \tilde{\mu}^{t-1})$ and $D(\mu, \mu^{t-1})$:

Lemma A.14. *Consider any $\mu_1, \mu_2 \in \Delta \subseteq \mathbb{R}^d$ such that $\mu_2(i) > 0,\ \forall i \in \{1, 2, \cdots, d\}$, and let $\tilde{\mu}_2 = (1-\theta)\mu_2 + \theta\frac{1}{d}\mathbf{1}$, for some $\theta \in (0, 1]$, then, it follows*

$$D(\mu_1, \tilde{\mu}_2) - D(\mu_1, \mu_2) \leq \theta \log d.$$

Furthermore, $D(\mu_1, \tilde{\mu}_2) \leq \log(d/\theta)$.

Proof of Lemma A.14. We have

$$\begin{aligned}
&D(\mu_1, \tilde{\mu}_2) - D(\mu_1, \mu_2)\\
=&\sum_{i=1}^{d} \mu_1(i) \left(\log \frac{\mu_1(i)}{\tilde{\mu}_2(i)} - \log \frac{\mu_1(i)}{\mu_2(i)} \right)\\
=&\sum_{i=1}^{d} \mu_1(i) \log \frac{\mu_2(i)}{\tilde{\mu}_2(i)}\\
=&\sum_{i=1}^{d} \mu_1(i) \left(\log \mu_2(i) - \log \left((1-\theta)\mu_2(i) + \theta \frac{1}{d} \mathbf{1} \right) \right)\\
\leq&\sum_{i=1}^{d} \mu_1(i) \left(\log \mu_2(i) - (1-\theta) \log \mu_2(i) - \theta \log \frac{1}{d} \right)\\
=&\theta \sum_{i=1}^{d} \mu_1(i) \left(\log \mu_2(i) + \log d \right)\\
\leq&\theta \log d,
\end{aligned}$$

where the first inequality follows from the concavity of log function. Furthermore, the second inequality follows from

$$\begin{aligned}
D(\mu_1, \tilde{\mu}_2) &= \sum_{i=1}^{d} \mu(i) \log \frac{\mu_1(i)}{\tilde{\mu}_2(i)}\\
&= \sum_{i=1}^{d} \mu(i) \log \frac{\mu_1(i)}{(1-\theta)\mu_2^{t-1}(i) + \theta/d}\\
&\leq - \sum_{i=1}^{d} \mu_1(i) \log((1-\theta)\mu_2^{t-1}(i) + \theta/d)\\
&\leq \log(d/\theta).
\end{aligned}$$

finishing the proof. □

A.4.1 Regret bound. First of all, by the same proof as that of Lemma 5.1 one can show the following:

$$\begin{aligned}
&V \left\langle \nabla f^{t-1}(\mu^{t-1}), \mu^t - \mu^{t-1} \right\rangle + \Delta(t) + \alpha D(\mu^t, \tilde{\mu}^{t-1})\\
&\leq V(f^{t-1}(\mu) - f^{t-1}(\mu^{t-1})) + \sum_{i=1}^{L} Q_i(t) g_i^{t-1}(\mu) + \sum_{j=1}^{M} H_j(t) \left(\left\langle h_j^{t-1}, \mu \right\rangle - b_j \right) + H^2 + G^2 + D_2^2\\
&\qquad + \alpha D(\mu, \tilde{\mu}^{t-1}) - \alpha D(\mu, \mu^t). \quad (35)
\end{aligned}$$

Furthermore, similar to that of Lemma 5.2, we have

$$\begin{aligned}
&V\left\langle \nabla f^{t-1}(\mu^{t-1}), \mu^t - \mu^{t-1}\right\rangle + \alpha D(\mu^t, \tilde{\mu}^{t-1}) \\
\geq& V\left\langle \nabla f^{t-1}(\mu^{t-1}), \mu^t - \tilde{\mu}^{t-1}\right\rangle + \alpha D(\mu^t, \tilde{\mu}^{t-1}) - V\theta\|\nabla f^{t-1}(\mu^{t-1})\|_\infty \\
\geq& V\left\langle \nabla f^{t-1}(\mu^{t-1}), \mu^t - \tilde{\mu}^{t-1}\right\rangle + \frac{\alpha}{2}\|\mu^t - \tilde{\mu}^{t-1}\|_1^2 - V\theta\|\nabla f^{t-1}(\mu^{t-1})\|_\infty \\
\geq& -V\|\nabla f^{t-1}(\mu^{t-1})\|_\infty\|\mu^t - \tilde{\mu}^{t-1}\|_1 + \frac{\alpha}{2}\|\mu^t - \tilde{\mu}^{t-1}\|_1^2 - V\theta\|\nabla f^{t-1}(\mu^{t-1})\|_\infty \\
\geq& -V\left(\frac{\alpha}{2V}\|\mu^t - \tilde{\mu}^{t-1}\|_1^2 + \frac{V}{2\alpha}\|\nabla f^{t-1}(\mu^{t-1})\|_\infty^2\right) + \frac{\alpha}{2}\|\mu^t - \tilde{\mu}^{t-1}\|_1^2 - V\theta\|\nabla f^{t-1}(\mu^{t-1})\|_\infty \\
=& -\left(\frac{V^2}{2\alpha}\|\nabla f^{t-1}(\mu^{t-1})\|_\infty^2 + V\theta\|\nabla f^{t-1}(\mu^{t-1})\|_\infty\right) \\
\geq& -\frac{VD_1^2}{2\alpha} - V\theta D_1. \qquad (36)
\end{aligned}$$

Substituting (36) into (35) gives

$$\begin{aligned}
&\Delta(t) + V(f^{t-1}(\mu^{t-1}) - f^{t-1}(\mu)) \\
&\quad\leq \sum_{i=1}^{L} Q_i(t) g_i^{t-1}(\mu) + \sum_{j=1}^{M} H_j(t)\left(\left\langle h_j^{t-1}, \mu\right\rangle - b_j\right) + H^2 + G^2 + D_2^2 + \frac{V^2}{2\alpha}D_1^2 + V\theta D_1 \\
&\qquad + \alpha D(\mu, \tilde{\mu}^{t-1}) - \alpha D(\mu, \mu^t). \qquad (37)
\end{aligned}$$

Using Lemma A.14, we get

$$\begin{aligned}
&\Delta(t) + V(f^{t-1}(\mu^{t-1}) - f^{t-1}(\mu)) \\
&\quad\leq \sum_{i=1}^{L} Q_i(t) g_i^{t-1}(\mu) + \sum_{j=1}^{M} H_j(t)(\left\langle h_j^{t-1}, \mu\right\rangle - b_j) + H^2 + G^2 + D_2^2 + \frac{V^2}{2\alpha}D_1^2 + V\theta D_1 \\
&\qquad + \alpha\theta\log d + \alpha D(\mu, \mu^{t-1}) - \alpha D(\mu, \mu^t).
\end{aligned}$$

The rest follows from the same argument as that of Section 5.1 after (13) and we omit the details for brevity.

A.4.2 Constraint violations. Similar as before, we start with the following lemma:

Lemma A.15. *The updating rule (5) and (6) delivers the following constraint violation bounds:*

$$\begin{aligned}
&\mathbb{E}\left\|\left[\frac{1}{T}\sum_{t=0}^{T-1}\overline{\mathbf{g}}(\mu^t)\right]_+\right\|_2 \leq \frac{\mathbb{E}(\|\mathbf{Q}(t)\|_2)}{T} + \frac{2D_2}{\alpha}\left(VD_1 + D_2\mathbb{E}(\|\mathbf{Q}(t)\|_2) + H\mathbb{E}(\|\mathbf{H}(t)\|_2)\right) + D_2\theta, \\
&\mathbb{E}\left\|\frac{1}{T}\sum_{t=0}^{T-1}\overline{\mathbf{h}}(\mu^t) - \mathbf{b}\right\|_2 \leq \frac{\mathbb{E}(\|\mathbf{H}(t)\|_2)}{T} + \frac{2H}{\alpha}\left(VD_1 + D_2\mathbb{E}(\|\mathbf{Q}(t)\|_2) + H\mathbb{E}(\|\mathbf{H}(t)\|_2)\right) + H\theta.
\end{aligned}$$

Proof of Lemma A.15. Using Lemma 5.7, it is enough to bound the difference $\mathbb{E}\left(\|\mu^{t+1} - \mu^t\|_1\right)$. For this, applying Lemma 2.1 by setting $y = \mu^{t-1}$, $x^* = \mu^t$, and $f(x) = \langle x, p\rangle$ with

$$p = V\nabla f^{t-1}(\mu^{t-1}) + \sum_{i=1}^{L} Q_i(t)\nabla g_i^{t-1}(\mu^{t-1}) + \sum_{j=1}^{M} H_j(t) h_j^{t-1},$$

we have

$$\left\langle V\nabla f^{t-1}(\mu^{t-1}) + \sum_{i=1}^{L} Q_i(t)\nabla g_i^{t-1}(\mu^{t-1}) + \sum_{j=1}^{M} H_i(t)h_j^{t-1}, \mu^t \right\rangle + \alpha D(\mu_t, \tilde{\mu}^{t-1})$$
$$\leq \left\langle V\nabla f^{t-1}(\mu^{t-1}) + \sum_{i=1}^{L} Q_i(t)\nabla g_i^{t-1}(\mu^{t-1}) + \sum_{j=1}^{M} H_j(t)h_j^{t-1}, \mu \right\rangle + \alpha(D(\mu, \tilde{\mu}^{t-1}) - D(\mu, \mu^t)). \quad (38)$$

Taking $\mu = \tilde{\mu}^{t-1}$ in (38) gives,

$$V\left\langle \nabla f^{t-1}(\mu^{t-1}), \mu^t \right\rangle + \sum_{i=1}^{L} Q_i(t)\left\langle \nabla g_i^{t-1}(\mu^{t-1}), \mu^t \right\rangle + \sum_{j=1}^{M} H_j(t)\left\langle h_j^{t-1}, \mu^t \right\rangle + \alpha D(\mu^t, \tilde{\mu}^{t-1})$$
$$\leq V\left\langle \nabla f^{t-1}(\mu^{t-1}), \tilde{\mu}^{t-1} \right\rangle + \sum_{i=1}^{L} Q_i(t)\left\langle \nabla g_i^{t-1}(\mu^{t-1}), \tilde{\mu}^{t-1} \right\rangle + \sum_{j=1}^{M} H_j(t)\left\langle h_j^{t-1}, \tilde{\mu}^{t-1} \right\rangle - \alpha D(\tilde{\mu}^{t-1}, \mu^t).$$

Note that we have

$$\left\langle \nabla f^{t-1}(\mu^{t-1}), \tilde{\mu}^{t-1} - \mu^t \right\rangle \leq \|\nabla f^{t-1}(\mu^{t-1})\|_\infty \|\mu^t - \tilde{\mu}^{t-1}\|_1 \leq D_1\|\mu^t - \tilde{\mu}^{t-1}\|_1.$$

Also, we have

$$\sum_{i=1}^{L} Q_i(t)\left\langle \nabla g_i^{t-1}(\mu^{t-1}), \tilde{\mu}^{t-1} - \mu^t \right\rangle \leq \|\mathbf{Q}(t)\|_2 \sqrt{\sum_{i=1}^{L} (\|\nabla g_i(\mu^{t-1})\|_\infty \|\mu^t - \tilde{\mu}^{t-1}\|_1)^2}$$
$$\leq D_2\|\mathbf{Q}(t)\|_2 + \|\mu^t - \tilde{\mu}^{t-1}\|_1,$$

and

$$\sum_{j=1}^{M} H_j(t)\left\langle h_j^{t-1}, \tilde{\mu}^{t-1} - \mu^t \right\rangle \leq \|\mathbf{H}(t)\|_2 \sqrt{\sum_{j=1}^{M} (\|h_j^{t-1}\|_\infty \|\mu^t - \tilde{\mu}^{t-1}\|_1)^2}$$
$$\leq H\|\mathbf{H}(t)\|_2 \|\mu^t - \tilde{\mu}^{t-1}\|_1.$$

Thus, it follows from the above three bounds,

$$D(\mu^t, \tilde{\mu}^{t-1}) + D(\tilde{\mu}^{t-1}, \mu^t) \leq \frac{1}{\alpha}\left(VD_1 + D_2\|\mathbf{Q}(t)\|_2 + H\|\mathbf{H}(t)\|_2\right)\|\mu^t - \tilde{\mu}^{t-1}\|_1.$$

By Pinsker's inequality, we have

$$D(\mu^t, \tilde{\mu}^{t-1}) + D(\tilde{\mu}^{t-1}, \mu^t) \geq \|\mu^t - \mu^{t-1}\|_1^2$$

Thus, it follows,

$$\|\mu^t - \tilde{\mu}^{t-1}\|_1^2 \leq 2\theta^2 + \frac{1}{\alpha}\left(VD_1 + D_2\|\mathbf{Q}(t)\|_2 + H\|\mathbf{H}(t)\|_2\right)\|\mu^t - \tilde{\mu}^{t-1}\|_1.$$

Solving the above quadratic inequality

$$\|\mu^t - \tilde{\mu}^{t-1}\|_1 \leq \frac{2}{\alpha}\left(VD_1 + D_2\|\mathbf{Q}(t)\|_2 + H\|\mathbf{H}(t)\|_2\right) + 2\theta,$$

which implies

$$\|\mu^t - \mu^{t-1}\|_1 \leq \frac{2}{\alpha}\left(VD_1 + D_2\|\mathbf{Q}(t)\|_2 + H\|\mathbf{H}(t)\|_2\right) + 3\theta,$$

Taking the expectation from both sides and subtracting this bound into Lemma 5.7 results in

$$\mathbb{E}\left\|\left[\frac{1}{T}\sum_{t=0}^{T-1}\overline{\mathbf{g}}(\mu^t)\right]_+\right\|_2 \leq \frac{\mathbb{E}(\|\mathbf{Q}(t)\|_2)}{T} + 3\theta D_2 + \frac{2VD_1D_2}{\alpha} + \frac{1}{T}\sum_{t=0}^{T-1}\frac{2D_2}{\alpha}\left(D_2\mathbb{E}(\|\mathbf{Q}(t)\|_2) + H\mathbb{E}(\|\mathbf{H}(t)\|_2)\right)$$

One can prove the bound on $\mathbb{E}\left\|\frac{1}{T}\sum_{t=0}^{T-1}\overline{\mathbf{h}}(\mu^t) - \mathbf{b}\right\|_2$ with exactly the same computation and we omit the proof. □

Now, by Lemma A.15 it is enough to bound $\mathbf{Q}(t)$ and $\mathbf{H}(t)$, for which we have the following lemma:

Lemma A.16. *Consider the t_0 slots drift for some positive integer t_0, then we have*

$$\frac{\|\mathbf{Q}(t+t_0)\|_2^2 + \|\mathbf{H}(t+t_0)\|_2^2 - \|\mathbf{Q}(t)\|_2^2 - \|\mathbf{H}(t)\|_2^2}{2} \leq V\sum_{\tau=t}^{t+t_0-1} f^{\tau-1}(\mu) + \sum_{i=1}^{L}Q_i(t)\sum_{\tau=t}^{t+t_0-1} g_i^{\tau-1}(\mu) + \sum_{j=1}^{M}H_j(t)\sum_{\tau=t}^{t+t_0-1}\left(\left\langle h_j^{\tau-1},\mu\right\rangle - b_j\right) + \frac{1}{2}\hat{C}_{V,\alpha,t_0}, \quad (39)$$

where

$$\hat{C}_{V,\alpha,t_0} := 2\left(H^2 + \frac{3}{2}G^2 + D_2^2\right)t_0^2 + 2\left(H^2 + G^2 + D_2^2 + \frac{V^2}{2\alpha}D_1^2 + V\theta D_1 + \alpha\theta\log d\right)t_0 + 2\alpha\log(d/\theta)$$

Proof of Lemma A.16. First of all, summing both sides of (37) from $\tau = t$ to $\tau = t + t_0 - 1$ gives

$$\begin{aligned}&\frac{\|\mathbf{Q}(t+t_0)\|_2^2 + \|\mathbf{H}(t+t_0)\|_2^2 - \|\mathbf{Q}(t)\|_2^2 - \|\mathbf{H}(t)\|_2^2}{2}\\ &\leq \sum_{\tau=t}^{t+t_0-1}\sum_{i=1}^{L}Q_i(\tau)g_i^{\tau-1}(\mu) + \sum_{\tau=t}^{t+t_0-1}\sum_{j=1}^{M}H_j(\tau)\left(\left\langle h_j^{\tau-1},\mu\right\rangle - b_j\right) + \left(H^2 + G^2 + 2D_2^2 + \frac{V^2}{2\alpha}D_1^2 + V\theta D_1\right)t_0\\ &\quad + V\sum_{\tau=t}^{t+t_0-1}\left(f^{\tau-1}(\mu^{\tau-1}) - f^{\tau-1}(\mu)\right) + \alpha D(\mu,\tilde{\mu}^{t-1}) - \alpha D(\mu,\mu^{t+t_0-1}) + \alpha\sum_{\tau=t+1}^{t+t_0-1}\left(D(\mu,\tilde{\mu}^{\tau-1}) - D(\mu,\mu^{\tau-1})\right).\end{aligned} \quad (40)$$

By Lemma A.14, one has

$$\alpha\sum_{\tau=t+1}^{t+t_0-1}\left(D(\mu,\tilde{\mu}^{\tau-1}) - D(\mu,\mu^{\tau-1})\right) \leq t_0\alpha\theta\log d.$$

and

$$\alpha D(\mu,\tilde{\mu}^{t-1}) \leq \alpha\log(d/\theta),$$

Thus, substituting these two bounds into (40) gives

$$\frac{\|\mathbf{Q}(t+t_0)\|_2^2+\|\mathbf{H}(t+t_0)\|_2^2-\|\mathbf{Q}(t)\|_2^2-\|\mathbf{H}(t)\|_2^2}{2}$$
$$\leq \sum_{\tau=t}^{t+t_0-1}\sum_{i=1}^{L}Q_i(\tau)g_i^{\tau-1}(\mu)+\sum_{\tau=t}^{t+t_0-1}\sum_{j=1}^{M}H_j(\tau)(\langle h_j^{\tau-1},\mu\rangle-b_j)+V\sum_{\tau=t}^{t+t_0-1}(f^{\tau-1}(\mu^{\tau-1})-f^{\tau-1}(\mu))$$
$$+\left(H^2+G^2+2D_2^2+\frac{V^2}{2\alpha}D_1^2+V\theta D_1+\alpha\theta\log d\right)t_0+\alpha\log(d/\theta). \quad (41)$$

Furthermore, following the steps to obtain (21) and (22) by invoking $\|\mu^t-\mu^{t-1}\|_1\leq 1$, we have

$$\sum_{\tau=t}^{t+t_0-1}\sum_{j=1}^{M}\left(H_i(\tau)-H_j(t)\right)\left\langle h_i^{\tau-1},\mu^{\tau}\right\rangle\leq t_0^2H^2.$$
$$\sum_{\tau=t}^{t+t_0-1}\sum_{i=1}^{L}\left(Q_i(\tau)-Q_i(t)\right)g_i^{\tau-1}(\mu)\leq t_0\left(\frac{3}{2G^2}+D_2^2\right),$$

and $V\sum_{\tau=t}^{t+t_0-1}f^{\tau-1}(\mu^{\tau-1})\leq t_0VF$. Substituting these three bounds into (41) and recall the definition of $\hat{C}_{V,\alpha,t_0}$ in the statement of the lemma give the final bound. □

Using the previous bound, one can prove the following lemma:

Lemma A.17. *If we take $V=\sqrt{T},\alpha=T,t_0=T,\theta=1/T$ in Algorithm 2, then the quantity $\|(\mathbf{Q}(t),\ \mathbf{H}(t))\|_2$ satisfies the following conditions:*

$$\mathbb{E}\left(\left\|(\mathbf{Q}(t),\ \mathbf{H}(t))\right\|_2\right)\leq\hat{C}'+\hat{C}''\sqrt{T}+\frac{2\log(d)}{\bar{c}}+\frac{2}{\bar{c}}\sqrt{T}\log Td, \quad (42)$$

where $\hat{C}'=\frac{2}{\bar{c}}\left(H^2+G^2+D_2^2+D_1^2/2+D_1\right)$ and $\hat{C}''=\frac{2}{\bar{c}}\left(H^2+\frac{3}{2}G^2+D_2^2+F+\bar{l}(G+H+\bar{c})+\bar{c}B+2(2(G+D_2)+H)^2\log(\frac{8(2(G+D_2)+H)^2}{\bar{c}^2})\right)$ are absolute constants independent of d or t.

Proof of Lemma A.17. Following the same arguments as those in Lemma 5.4, 5.5 and 5.6, we can show

$$\mathbb{E}\left(\left\|(\mathbf{Q}(t),\ \mathbf{H}(t))\right\|_2\right)\leq\frac{\hat{C}_{V,\alpha,t_0}+2\left(F+\bar{\ell}(G+H+\bar{c})+\bar{c}B\right)Vt_0}{\bar{c}t_0}$$
$$+\frac{4t_0\left(2(G+D_2)+H\right)^2}{\bar{c}}\log\left(\frac{8\left(2(G+D_2)+H\right)^2}{\bar{c}^2}\right).$$

Taking $V=\sqrt{T},\alpha=T,t_0=T,\theta=1/T$ and recalling the definition of $\hat{C}_{V,\alpha,t_0}$ yields

$$\mathbb{E}\left(\left\|(\mathbf{Q}(t),\ \mathbf{H}(t))\right\|_2\right)\leq\hat{C}'+\hat{C}''\sqrt{T}+\frac{2\log(d)}{\bar{c}}+\frac{2}{\bar{c}}\sqrt{T}\log Td,$$

where $\hat{C}'=\frac{2}{\bar{c}}\left(H^2+G^2+D_2^2+D_1^2/2+D_1\right)$ and $\hat{C}''=\frac{2}{\bar{c}}\left(H^2+\frac{3}{2}G^2+D_2^2+F+\bar{l}(G+H+\bar{c})+\bar{c}B+2(2(G+D_2)+H)^2\log(\frac{8(2(G+D_2)+H)^2}{\bar{c}^2})\right)$. □

The constraint violations in Theorem 3.2 then follows by combining Lemma A.15 and Lemma A.17.

A.5 Proof of other supporting lemmas

Proof of Lemma A.5. We expand the simplex constraints in (7) explicitly and the full dual function writes

$$q_0^{(t,k)}(\lambda,\ \eta,\ \mathbf{u},\ v) := \min_{\mu\in\mathbb{R}^d} \overline{f}^{t,k}(\mu) + \sum_{i=1}^{L}\lambda_i\nabla\overline{g}_i(\mu) + \sum_{j=1}^{M}\eta_j\left\langle\mathbb{E}\left(h_j^t\right),\mu\right\rangle - \sum_{i=1}^{d}u_i\mu_i + v(\sum_{i=1}^{d}\mu_i - 1).$$

Let $q_0^* = \max_{\lambda\geq 0,\ \eta\in\mathbb{R}^M,\ \mathbf{u}\geq 0, v\in\mathbb{R}} q_0^{(t,k)}(\lambda,\ \eta,\ \mathbf{u},\ v)$. By the assumption of lemma A.5 and Theorem A.4 we have the solution set $K(\mu^*)$ of vectors $(\lambda,\ \eta,\ \mathbf{u},\ v)$ of the following equations (KKT conditions) is non-empty and bounded:

$$\begin{aligned}
&\nabla\overline{f}^{t,k}(\mu^*) + \sum_{i=1}^{L}\lambda_i\nabla\overline{g}_i(\mu^*) + \sum_{j=1}^{M}\eta_j\mathbb{E}\left(h_j^t\right) - \sum_{i=1}^{d}u_i\mathbf{e}_i + v\mathbf{1} = 0,\\
&\lambda\geq 0,\ \mathbf{u}\geq 0,\\
&\lambda_i g_i(\mu^*) = 0,\ \forall i\in\{1,2,\cdots,M\},\\
&u_i\mu_i^* = 0,\ \forall i\in\{1,2,\cdots,d\}.
\end{aligned}\tag{43}$$

It can be verified that

$$K(\mu^*) = \text{argmax}_{\lambda\geq 0,\ \eta\in\mathbb{R}^M,\ \mathbf{u}\geq 0, v\in\mathbb{R}}\ q_0^{(t,k)}(\lambda,\ \eta,\ \mathbf{u},\ v)$$

and we have zero duality gap, i.e. $q_0^* = \overline{f}^{(t,k)}(\mu^*)$. Our goal is to show that the set $\mathcal{V}^*$, defined in the statement of the lemma, is equal to the set $\{(\lambda^*,\eta^*)\mid(\lambda^*,\eta^*,\mathbf{u}^*,v^*)\in K(\mu^*),\ \exists\,\mathbf{u}^*,v^*\}$.

First of all, for any $(\lambda^*,\eta^*,\mathbf{u}^*,v^*)\in K(\mu^*)$, we have $q^{(t,k)}(\lambda^*,\eta^*)\geq q_0^{(t,k)}(\lambda^*,\ \eta^*,\ \mathbf{u}^*,\ v^*) = q_0^*$. Since we have zero duality gap $q_0^* = \overline{f}^{(t,k)}(\mu^*)$ and one always has $q^{(t,k)}(\lambda,\eta)\leq\overline{f}^{(t,k)}(\mu^*),\ \forall\lambda\in\mathbb{R}_+^L,\ \eta\in\mathbb{R}^M$, it follows $q^{(t,k)}(\lambda^*,\eta^*) = \overline{f}^{(t,k)}(\mu^*)$. Thus, not only do we have a zero duality gap of $q^{(t,k)}(\lambda^*,\eta^*)$, we also have λ^*,η^* being the solution to the dual maximization problem $\max_{\lambda\in\mathbb{R}_+^L,\ \eta\in\mathbb{R}^M} q^{(t,k)}(\lambda,\eta)$, showing that $\mathcal{V}^*$ is non-empty and $\{(\lambda^*,\eta^*)\mid(\lambda^*,\eta^*,\mathbf{u}^*,v^*)\in K(\mu^*),\ \exists\,\mathbf{u}^*,v^*\}\subseteq\mathcal{V}^*$.

For the other direction, we pick any $(\lambda^*,\eta^*)\in\mathcal{V}^*$ and consider the following optimization problem:

$$q^{(t,k)}(\lambda^*,\eta^*) = \min_{\mu\in\Delta}\ \overline{f}^{t,k}(\mu) + \sum_{i=1}^{L}\lambda_i^*\overline{g}_i(\mu) + \sum_{j=1}^{M}\eta_j^*\overline{h}_j(\mu).\tag{44}$$

By zero duality gap, the solution to this optimization problem is equal to $\overline{f}^{(t,k)}(\mu^*)$. Thus μ^* must be one of the solution points of (44) such that the complementary slackness $\lambda_i^*\overline{g}_i(\mu^*) = 0,\ \forall i\in\{1,2,\cdots,L\}$ is satisfied.[2] Furthermore, it is obvious that MFCQ is also satisfied for (44) (we only need to check the simplex constraints satisfy MFCQ, which is obvious). Thus, by Theorem A.4, we have there exists $\mathbf{u}^*\geq 0, v^*\in\mathbb{R}$ such that the stationary condition (43) is satisfied, and $u_i\mu_i^* = 0,\ \forall i\in\{1,2,\cdots,d\}$. Combining with the previous complementary slackness $\lambda_i^*\overline{g}_i(\mu^*) = 0$, we arrive at the conclusion that $(\lambda^*,\eta^*,\mathbf{u}^*,v^*)\in K(\mu^*)$. This implies $\mathcal{V}^*\subseteq\{(\lambda^*,\eta^*)\mid(\lambda^*,\eta^*,\mathbf{u}^*,v^*)\in K(\mu^*),\ \exists\,\mathbf{u}^*,v^*\}$. Overall, we have the set $\mathcal{V}^*$ is also bounded and we finish the proof. □

[2] Suppose on the contrary $\lambda_i^*\overline{g}_i(\mu^*) < 0$ for some index i, then, this means taking μ^* gives smaller value of the objective than $\overline{f}^{(t,k)}(\mu^*)$, contradicting the fact that the minimum is $\overline{f}^{(t,k)}(\mu^*)$.

Proof of Lemma A.13. First of all, note that by the EBC, for any $(\lambda,\eta)\in\mathcal{S}_\delta$, one has $\text{dist}((\lambda,\eta),\mathcal{V}^*)\leq C_\delta\delta^\beta$, thus, for those (λ,η) such that $\text{dist}((\lambda,\eta),\mathcal{V}^*)\geq C_\delta\delta^\beta$, $(\lambda,\eta)\notin\mathcal{S}_\delta$. We then recall the following result:

Lemma A.18 ([24]). *Consider any convex function $F:\mathcal{X}\to\mathbb{R}$ such that the minimal set Λ^* is non-empty. Then, for any $\mathbf{x}\in\mathcal{X}$ and any $\varepsilon>0$,*

$$\|\mathbf{x}-\mathbf{x}_\varepsilon^\dagger\|\leq\frac{dist(\mathbf{x}_\varepsilon^\dagger,\Lambda^*)}{\varepsilon}\left(F(\mathbf{x})-F(\mathbf{x}_\varepsilon^\dagger)\right),$$

where $\mathbf{x}_\varepsilon^\dagger:=argmin_{\mathbf{x}_\varepsilon\in\mathcal{S}_\varepsilon}\|\mathbf{x}-\mathbf{x}_\varepsilon\|$, and $\mathcal{S}_\varepsilon$ is the ε-sublevel set defined in Lemma A.13.

Applying this lemma to our scenario, we define

$$(\lambda_\delta^\dagger,\eta_\delta^\dagger)=\operatorname*{argmin}_{(\lambda_\delta,\eta_\delta)\in\mathcal{S}_\delta}\|(\lambda_\delta,\eta_\delta)-(\lambda,\eta)\|_2$$

and take function to be $q^{(t,k)}(\lambda,\eta)$ and consider the δ-superlevel set S_δ. By lemma (A.18), we readily have

$$\begin{aligned}&\|(\lambda,\eta)-(\lambda_\delta^\dagger,\eta_\delta^\dagger)\|_2\\ \leq&\frac{\text{dist}((\lambda_\delta^\dagger,\eta_\delta^\dagger),\mathcal{V}^*)}{\delta}\left(q^{(t,k)}(\lambda_\delta^\dagger,\eta_\delta^\dagger)-q^{(t,k)}(\lambda,\eta)\right)\\ \leq&\frac{C_\delta\delta^\beta}{\delta}\left(q^{(t,k)}(\lambda_\delta^\dagger,\eta_\delta^\dagger)-q^{(t,k)}(\lambda,\eta)\right)\\ =&C_\delta\delta^{\beta-1}\left(q^{(t,k)}(\lambda_\delta^\dagger,\eta_\delta^\dagger)-q^{(t,k)}(\lambda,\eta)\right).\end{aligned}$$

On the other hand,

$$\text{dist}((\lambda_\delta^\dagger,\eta_\delta^\dagger),\mathcal{V}^*)\leq C_\delta\left(q^{(t,k)}(\lambda^*,\eta^*)-q^{(t,k)}(\lambda_\delta^\dagger,\eta_\delta^\dagger)\right)^\beta$$

Now, we claim that $q^{(t,k)}(\lambda^*,\eta^*)-q^{(t,k)}(\lambda_\delta^\dagger,\eta_\delta^\dagger)=\delta$. Indeed, suppose on the contrary, $q^{(t,k)}(\lambda^*,\eta^*)-q^{(t,k)}(\lambda_\delta^\dagger,\eta_\delta^\dagger)<\delta$, then, by the continuity of the function $q^{(t,k)}$, there exists $\alpha\in(0,1)$ and $(\lambda',\eta')=\alpha(\lambda_\delta^\dagger,\eta_\delta^\dagger)+(1-\alpha)(\lambda,\eta)$ such that $q^{(t,k)}(\lambda^*,\eta^*)-q^{(t,k)}(\lambda',\eta')=\delta$, i.e. $(\lambda',\eta')\in\mathcal{S}_\delta$, and $\|(\lambda,\eta)-(\lambda',\eta')\|_2=\alpha\|(\lambda,\eta)-(\lambda_\delta^\dagger,\eta_\delta^\dagger)\|_2<\|(\lambda,\eta)-(\lambda_\delta^\dagger,\eta_\delta^\dagger)\|_2$, contradicting the definition that $(\lambda_\delta^\dagger,\eta_\delta^\dagger)=\text{argmin}_{(\lambda_\delta,\eta_\delta)\in\mathcal{S}_\delta}\|(\lambda_\delta,\eta_\delta)-(\lambda,\eta)\|_2$.

Thus, we have

$$\text{dist}((\lambda_\delta^\dagger,\eta_\delta^\dagger),\mathcal{V}^*)\leq C_\delta\delta^{\beta-1}\left(q^{(t,k)}(\lambda^*,\eta^*)-q^{(t,k)}(\lambda_\delta^\dagger,\eta_\delta^\dagger)\right).$$

Overall, we have

$$\begin{aligned}\text{dist}((\lambda,\eta),\mathcal{V}^*)\leq&\text{dist}((\lambda_\delta^\dagger,\eta_\delta^\dagger),\mathcal{V}^*)+\|(\lambda,\eta)-(\lambda_\delta^\dagger,\eta_\delta^\dagger)\|_2\\ \leq&C_\delta\delta^{\beta-1}\left(q^{(t,k)}(\lambda^*,\eta^*)-q^{(t,k)}(\lambda,\eta)\right),\end{aligned}$$

and we finish the proof. □