Online Convex Optimization with Stochastic Constraints: Zero Constraint Violation and Bandit Feedback

Yeongjong Kim [1] Dabeen Lee [2] *

January 27, 2023

Abstract

This paper studies online convex optimization with stochastic constraints. We propose a variant of the drift-plus-penalty algorithm that guarantees $O(\sqrt{T})$ expected regret and zero constraint violation, after a fixed number of iterations, which improves the vanilla drift-plus-penalty method with $O(\sqrt{T})$ constraint violation. Our algorithm is oblivious to the length of the time horizon T, in contrast to the vanilla drift-plus-penalty method. This is based on our novel drift lemma that provides time-varying bounds on the virtual queue drift and, as a result, leads to time-varying bounds on the expected virtual queue length. Moreover, we extend our framework to stochastic-constrained online convex optimization under two-point bandit feedback. We show that by adapting our algorithmic framework to the bandit feedback setting, we may still achieve $O(\sqrt{T})$ expected regret and zero constraint violation, improving upon the previous work for the case of identical constraint functions. Numerical results demonstrate our theoretical results.

1 Introduction

Online convex optimization (OCO) is a general framework for modeling decision-making problems under uncertainty. OCO can be viewed as a repeated game between a learner and an adversarial environment as follows. At each iteration, the learner selects a decision without the knowledge of the convex loss function chosen by the environment, after which the learner receives the loss associated with the decision. Based on the repeated interactions, the learner adapts to the environment in real-time to minimize cumulative loss. The OCO framework is well-suited for optimizing a large-scale complex system, as the problem is often tackled by decomposing it into small optimization problems and solving each piece with limited information to improve tractability. Therefore, OCO is applied to portfolio management [2], routing [5], display learning [10], recommendation systems [13], binary classification [8], etc. For a comprehensive account of OCO, we refer the reader to [12, 19] and references therein.

[1] Department of Mathematical Sciences, KAIST, Daejeon 34126, Republic of Korea

[2] Department of Industrial and Systems Engineering, KAIST, Daejeon 34126, Republic of Korea

*Correspondence to <dabeenl@kaist.ac.kr>

OCO with *long-term constraints* is an extension of OCO to deal with complex functional constraints [16] and long-term budget restrictions [15]. When the set of available decisions is given by some complicated functions, projection onto the feasible set can be difficult. Under such a scenario, finding a feasible solution at each iteration can be expensive, although we may hope for achieving feasibility on average in the long run. For a resource planning scenario, it may be possible to pool the budget for multiple periods, over which the budget allocation is flexible. Motivated by this, we aggregate the constraint functions over the time periods and require satisfying the long-term constraint aggregation. More precisely, we consider

$$\min_{\boldsymbol{x_1},\ldots,\boldsymbol{x_T}\in\mathcal{X}} \quad \sum_{t=1}^{T} f_t(\boldsymbol{x_t}) \quad \text{s.t.} \quad \sum_{t=1}^{T} g_t(\boldsymbol{x_t}) \leq 0 \tag{1}$$

where $\{f_t\}_{t=1}^T$ and $\{g_t\}_{t=1}^T$ are the loss and constraint functions chosen by the environment over T time steps while $\{\boldsymbol{x_t}\}_{t=1}^T$ is the sequence of decisions selected from a domain $\mathcal{X}$ by the learner. Here, the learner selects $\boldsymbol{x_t}$ based on the history up to time step t before observing f_t and g_t. Applications of OCO with long-term constraints include online routing in wireless networks [17], multi-objective classification [6], and multi-armed bandits with knapsack constraints [3].

As the standard OCO framework, the performance of the learner can be measured by the notion of *regret*, defined as

$$\text{regret}(T) = \sum_{t=1}^{T} f_t(\boldsymbol{x_t}) - \sum_{t=1}^{T} f_t(\boldsymbol{x}^*)$$

against some fixed benchmark decision $\boldsymbol{x}^*$ and the *long-term constraint violation* given by

$$\text{violation}(T) = \sum_{t=1}^{T} g_t(\boldsymbol{x_t}).$$

Basically, the learner's goal is to perform as well as the benchmark while satisfying the constraints in the long run by minimizing the regret measure and trying to keep the constraint violation expression below zero. However, [17] found an example where $\{g_t\}_{t=1}^T$ is adversarially chosen and the benchmark $\boldsymbol{x}^*$ is set to an optimal fixed solution of (1), for which it is impossible to simultaneously bound the regret and constraint violation by a sublinear function in T. Given this impossibility result, several works have studied some structured special cases.

[16] considered the special case where $g_t = g$ for some fixed function g for all $t \in [T]$, for which they developed an augmented-Lagrangian-based method that achieves $O(\sqrt{T})$ regret and $O(T^{3/4})$ constraint violation. [14] generalized this result to $O(T^{\max\{\beta,1-\beta\}})$ regret and $O(T^{1-\beta/2})$ constraint violation by an adaptive variant of the augmented Lagrangian algorithm parameterized by $\beta \in (0,1)$. Later, [24] gave an algorithm with $O(\sqrt{T})$ regret and $O(1)$ constraint violation, but it is not a first-order method.

[25] studied the case where g_t's are time-varying but are independent and identically distributed (i.i.d.) with an unknown probability distribution, which subsumes the identical constraint function case discussed above. They set the benchmark $\boldsymbol{x}^*$ to be the best fixed solution minimizing the cumulative loss among the ones satisfying the constraint in expectation, and they showed that their drift-plus-penalty (DPP) algorithm achieves $O(\sqrt{T})$ expected regret and $O(\sqrt{T})$ expected constraint violation under Slater's condition. Later, [22] proposed a mirror-descent-type variant of DPP that achieves the same asymptotic performance under slightly more general settings.

Besides these results on structured cases, there are more works regarding constrained OCO. [15, 18, 21] also consider problem (1) with adversarially chosen constraint functions but they work over different benchmarks with more restrictions to guarantee sublinear regret and constraint violation simultaneously. [11, 23, 26] studied the notion of *cumulative constraint violation*, given by $\sum_{t=1}^{T}\left[g_t(\boldsymbol{x_t})\right]_+$ where $[a]_+ = \max\{a, 0\}$ over $a \in \mathbb{R}$, instead of long-term constraints.

Note that algorithms for OCO use the information about loss and constraint functions and their gradients to proceed. However, for some applications, we observe only the function values of the decisions and do not have access to the functions and their gradients. This setting is called OCO with *bandit feedback* [9]. [1] introduced the *two-point bandit feedback* setting, in which we can observe the function values at (at least) two points per iteration. For the two-point feedback setting, we can achieve $O(\sqrt{T})$ regret [1, 20]. [7] studied long-term constrained OCO with identical constraint functions under two-point bandit feedback and provided an algorithm with $O(T^{1/2}\tilde{\Delta}(T)^{1/2})$ regret and $O(T^{3/4}\tilde{\Delta}(T)^{1/4})$ constraint violation where $\tilde{\Delta}(T)$ is some sublinear function in T.

Our contributions In this paper, we focus on OCO with stochastic constraints under the full information setting and the two-point bandit feedback setting. We improve upon the results of [22, 25] in three aspects.

1. We develop a variant of the drift-plus-penalty algorithm achieving $O(\sqrt{T})$ expected regret and zero expected constraint violation, after a fixed number of iterations.
2. Our algorithm is completely oblivious to the length of the time horizon T, in contrast to the vanilla drift-plus-penalty algorithm by [22, 25] which sets the penalty parameter $V = \sqrt{T}$ and the step size parameter $\alpha = T$. This development is based on our novel drift lemma that provides time-varying bounds on the drift and leads to time-varying bounds on the expected virtual queue size.
3. We adapt our algorithm to the two-point bandit feedback setting and show that we may still guarantee $O(\sqrt{T})$ expected regret and zero expected constraint violation.

The result on stochastic-constrained OCO under two-point bandit feedback improves upon the work of [7] because our algorithm guarantees better bounds on regret and constraint violation for a strictly more general setting.

2 OCO with Stochastic Constraints

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a known fixed compact convex set. Let $f_1, \ldots, f_T : \mathbb{R}^d \to \mathbb{R}$ be a sequence of arbitrary convex functions. Let $\bar{g}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\omega}}\left[g(\boldsymbol{x}, \boldsymbol{\omega})\right] : \mathbb{R}^d \to \mathbb{R}$ be a function where $g(\boldsymbol{x}, \boldsymbol{\omega})$ is convex with respect to $\boldsymbol{x} \in \mathcal{X}$ and the expectation is taken with $\boldsymbol{\omega} \in \Omega$ from an unknown distribution. Constraint functions $g_1, \ldots, g_T : \mathbb{R}^d \to \mathbb{R}$ are given by $g_t(\boldsymbol{x}) := g(\boldsymbol{x}, \boldsymbol{\omega_t})$ for $t \in [T]$ where $\boldsymbol{\omega_1}, \ldots, \boldsymbol{\omega_T}$ are i.i.d. samples of $\boldsymbol{\omega}$. We assume that f_t is independent of $\boldsymbol{\omega_s}$ for $s \geq t+1$.

As in [25], we take the benchmark decision $\boldsymbol{x}^*$ defined as an optimal solution to

$$\min_{\boldsymbol{x} \in \mathcal{X}} \sum_{t=1}^{T} f_t(\boldsymbol{x}) \quad \text{s.t.} \quad \bar{g}(\boldsymbol{x}) = \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} g_t(\boldsymbol{x})\right] \leq 0. \tag{2}$$

Then the goal is to design an algorithm that guarantees a sublinear regret against the benchmark $\boldsymbol{x}^*$ and a sublinear constraint violation at the same time. We focus on the single constraint setting for simplicity, but our framework easily extends to multiple constraints.

Henceforth, we work over a norm $\|\cdot\|$ in $\mathbb{R}^d$ and its dual norm $\|\cdot\|_*$. Let $\Phi : \mathcal{C} \to \mathbb{R}$ be a mirror map over an open convex set $\mathcal{C}$. We assume that $\mathcal{X} \subseteq \mathcal{C}$ and consider the corresponding *Bregman divergence* defined as $D(\boldsymbol{x}, \boldsymbol{y}) = \Phi(\boldsymbol{x}) - \Phi(\boldsymbol{y}) - \Phi(\boldsymbol{y})^\top(\boldsymbol{x} - \boldsymbol{y})$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$.

Assumption 1 (Basic assumptions)**.** *There exist positive constants D_f, D_g, G, R satisfying the following.*

- $\|\nabla f_t(\boldsymbol{x})\|_* \leq D_f$ *for all* $t \in [T]$ *and* $\boldsymbol{x} \in \mathcal{X}$.
- $\|\nabla g(\boldsymbol{x}, \boldsymbol{\omega})\|_* \leq D_g$, *and* $|g(\boldsymbol{x}, \boldsymbol{\omega})| \leq G$ *for all* $\boldsymbol{x} \in \mathcal{X}$ *and* $\boldsymbol{\omega} \in \Omega$.
- $D(\boldsymbol{x}, \boldsymbol{y}) \leq R^2$ *for all* $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$.

For convenience, we assume that Φ is 2-strongly convex with respect to the norm $\|\cdot\|$, which implies that

$$\|\boldsymbol{x} - \boldsymbol{y}\| \leq \sqrt{D(\boldsymbol{x}, \boldsymbol{y})} \leq R \tag{3}$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$.

Assumption 2 (Slater's condition)**.** *There exist a solution $\hat{\boldsymbol{x}} \in \mathcal{X}$ and $\epsilon > 0$ with $\bar{g}(\hat{\boldsymbol{x}}) = \mathbb{E}_{\boldsymbol{\omega}}[g(\hat{\boldsymbol{x}}, \boldsymbol{\omega})] \leq -\epsilon$.*

The framework of OCO with stochastic constraints has various applications in optimization and learning theory. Here we discuss a few that are direct applications of the framework. As discussed before, a special case is when the constraints functions are not time-varying, i.e., g_t is given by $g_t = g$ for some fixed function g for all t. Another direct application is *stochastic constrained stochastic optimization*, that is formulated as the following optimization problem.

$$\min_{\boldsymbol{x} \in \mathcal{X}} \quad \bar{f}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\omega}}\left[f(\boldsymbol{x}, \boldsymbol{\omega})\right] \quad \text{s.t.} \quad \bar{g}(\boldsymbol{x}) \leq 0.$$

An iterative algorithm would obtain an i.i.d. sample $\boldsymbol{\omega_t}$ of $\boldsymbol{\omega}$ at each iteration t and consider $f_t = f_t(\cdot, \boldsymbol{\omega_t})$ and $g_t = g_t(\cdot, \boldsymbol{\omega_t})$. Given $\{\boldsymbol{x_t}\}_{t=1}^T$, we may obtain $\bar{\boldsymbol{x}}_{\boldsymbol{T}} = (1/T)\sum_{t=1}^T \boldsymbol{x_t}$. By Jensen's inequality, the optimality gap of $\bar{\boldsymbol{x}}_{\boldsymbol{T}}$ satisfies $\mathbb{E}\left[\bar{f}(\bar{\boldsymbol{x}}_{\boldsymbol{T}})\right] - \bar{f}(\boldsymbol{x}^*) \leq \mathbb{E}\left[\text{regret}(T)\right]/T$, and the constraint violation is given by $\bar{g}(\bar{\boldsymbol{x}}_{\boldsymbol{T}}) = T \cdot \mathbb{E}\left[\text{violation}(T)\right]$.

3 Conservative Drift-Plus-Penalty

3.1 Algorithm Descriptions and Intuitions

We deduce our algorithm by combining the drift-plus-penalty algorithm in [22, 25] for OCO with stochastic constraints and the *conservative* optimization method in [4] achieving zero constraint violation for stochastic constrained stochastic optimization.

Algorithm 1, which we refer to as Conservative Drift-Plus-Penalty (CDPP), follows the basic outline of the vanilla DPP algorithm, while there are two main distinctions.

Algorithm 1 Conservative Drift-Plus-Penalty

```
Initialize: Initial iterates $\boldsymbol{x_1} \in \mathcal{X}$ and $Q_1 = 0$, step size parameters $\{\alpha_t\}_{t=1}^T$, penalty parameters $\{V_t\}_{t=1}^T$, and conservatism parameters $\{\gamma_t\}_{t=1}^T$.
for $t = 1$ to $T$ do
    Observe $f_t$ and $g_t$.
    Primal update: set $\boldsymbol{x_{t+1}}$ to be a solution to
        $\min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \left(V_t \nabla f_t(\boldsymbol{x_t}) + Q_t \nabla g_t(\boldsymbol{x_t})\right)^\top \boldsymbol{x} + \alpha_t D(\boldsymbol{x}, \boldsymbol{x_t}) \right\}$
    Dual update: set $Q_{t+1}$ to
        $\left[Q_t + g_t(\boldsymbol{x_t}) + \nabla g_t(\boldsymbol{x_t})^\top (\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t\right]_+$
end for
```

1. For the primal update, we allow penalty parameter V_t and step size parameter α_t to be time-varying. In contrast, [25] used fixed parameters given by $V_t = \sqrt{T}$ and $\alpha_t = T$ for all t. Instead, we set for $t \geq 1$,

$$V_t = \sqrt{t} \quad \text{and} \quad \alpha_t = t.$$

2. For the dual update, we use a new parameter γ_t and add it to the current virtual queue Q_t in each step t. We will set w_t to

$$\gamma_t = \frac{C}{\sqrt{t}}$$

for some constant C. The dual update of the vanilla DPP method is the one with $C = 0$.

Note that our choice of parameters V_t and α_t for primal updates is oblivious to the length of time horizon T, providing flexibility for deciding when to stop the online learning procedure. To demonstrate that the time-varying penalty and step size parameters still guarantee the desired performance, we refine and improve the analysis of the DPP algorithm.

For the dual update, we introduce the new parameter γ_t to better control the (expected) long-term constraint violation. The basic idea is to consider function $g_t(\cdot) + \gamma_t$ instead of $g_t(\cdot)$. Then, as the algorithm aims to control the corresponding constraint violation $\sum_{t=1}^T g_t(\boldsymbol{x_t}) + \sum_{t=1}^T \gamma_t$, this would encourage a further reduction in the actual constraint violation $\sum_{t=1}^T g_t(\boldsymbol{x_t})$. We call γ_t the conservatism parameter, indicating that we add extra conservatism toward satisfying the long-term constraint. This idea of conservative optimization was used in [4].

We will show that the introduction of parameter γ_t leads to a reduction of $O(C\sqrt{T})$ in the (expected) long-term constraint violation while incurring an additional (expected) regret of $O(C\sqrt{T})$. As the vanilla DPP bounds the long-term constraint violation and the regret by $O(\sqrt{T})$, we may properly choose a value for C to achieve zero long-term constraint violation while still bounding the regret by $O(\sqrt{T})$.

Let us also mention that our algorithm, as well as the original DPP method, has a close connection to the online primal-dual gradient method for the saddle-point problem. Consider

$$L_t(\boldsymbol{x}, \lambda) = f_t(\boldsymbol{x}) + \lambda(g_t(\boldsymbol{x}) + \gamma_t)$$

for $\boldsymbol{x} \in \mathcal{X}$ and $\lambda \geq 0$ and the corresponding online primal-dual gradient update with step size $1/\sqrt{2T}$ given by

$$\begin{aligned}\boldsymbol{x_{t+1}} &= \mathcal{P}_{\mathcal{X}}\left[\boldsymbol{x_t} - \frac{1}{\sqrt{2T}}\left(\nabla f_t(\boldsymbol{x_t}) + \lambda_t \nabla g_t(\boldsymbol{x_t})\right)\right],\\ \lambda_{t+1} &= \left[\lambda_t + \frac{1}{\sqrt{2T}}(g_t(\boldsymbol{x_t}) + \gamma_t)\right]_+\end{aligned}$$

where $\mathcal{P}_{\mathcal{X}}$ denote the Euclidean projection operator onto $\mathcal{X}$. In fact, setting $Q_t = \lambda_t\sqrt{2T}$, the primal update of online primal-dual gradient is equivalent to that of our algorithm with $V_t = \sqrt{2T}$, $\alpha_t = T$, and $\Phi(\cdot) = \|\cdot\|_2^2$, in which case $D(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$. On the other hand, the dual update of online primal-dual gradient translates to

$$Q_{t+1} = \left[Q_t + g_t(\boldsymbol{x_t}) + \gamma_t\right]_+.$$

Notice that our dual update has the additional term $\nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t})$, which is the distinctive component of the drift-plus-penalty algorithm. Next, let us briefly elaborate on the intuition behind adding the term $\nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t})$ to the dual update.

Following [25], we may regard Q_t as the size of a *virtual queue* at time t. Then we consider the associated *quadratic Lyapunov term* $L_t = Q_t^2/2$ and study the corresponding *drift* given by $\Delta_t = L_{t+1} - L_t = (Q_{t+1}^2 - Q_t^2)/2$.

Lemma 3.1. *For* $t \geq 1$*,*

$$\Delta_t \leq Q_t\left(g_t(\boldsymbol{x_t}) + \nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t\right) + (G + D_g R + \gamma_t)^2.$$

The upper bound on the drift Δ_t provided by Lemma 3.1 has the (only) term $Q_t\nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t})$ that depends on the next iterate $\boldsymbol{x_{t+1}}$. Hence, by choosing $\boldsymbol{x_{t+1}}$ that minimizes $Q_t\nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t})$, we may attempt to control the drift. In fact, the primal update of CDPP sets $\boldsymbol{x_{t+1}}$ to be the minimizer of

$$\underbrace{Q_t\nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x} - \boldsymbol{x_t})}_{\text{drift}} + \underbrace{V_t\nabla f_t(\boldsymbol{x_t})^\top(\boldsymbol{x} - \boldsymbol{x_t}) + \alpha_t D(\boldsymbol{x}, \boldsymbol{x_t})}_{\text{penalty}}$$

over $\mathcal{X}$. Consequently, at each iteration, we get to choose a solution that minimizes the drift term Δ_t and a penalty term for controlling the objective simultaneously.

3.2 Performance of Conservative Drift-Plus-Penalty

In this section, we show that the expected regret and constraint violation under CDPP are bounded by $O(\sqrt{T})$ and 0, respectively.

The following two lemmas provide upper bounds on the expected regret and constraint violation under CDPP. The bounds exhibit the dependence on the conservative parameter C and the expected virtual queue length $\mathbb{E}\left[Q_t\right]$.

Lemma 3.2. *Under CDPP (Algorithm 1), the expected regret* $\mathbb{E}\left[\mathrm{regret}(T)\right]$ *is at most*

$$\left(\frac{C(G+D_gR+C)}{\epsilon}\right)^2 + K_1\sqrt{T} + C\sum_{t=1}^{T}\frac{\mathbb{E}[Q_t]}{t}$$

for some positive constant K_1 *that depends only on* G, D_g, D_f, R, ϵ.

Lemma 3.3. *Under CDPP (Algorithm 1), the expected constraint violation* $\mathbb{E}\left[\mathrm{violation}(T)\right]$ *is at most*

$$D_fD_g\sqrt{T} + \mathbb{E}[Q_{T+1}] + \frac{D_g^2}{2}\sum_{t=1}^{T}\frac{\mathbb{E}[Q_t]}{t} - C\sqrt{T}.$$

Given Lemmas 3.2 and 3.3, what remains is to bound the expected virtual queue length $\mathbb{E}\left[Q_t\right]$, which is carried out based on the following lemma. Before stating the lemma, let us define filtration $\{\mathcal{F}_t:\ t\geq 0\}$ where $\mathcal{F}_0=\{\emptyset,\Omega\}$ and $\mathcal{F}_t=\sigma(\boldsymbol{\omega_1},\ldots,\boldsymbol{\omega_t})$ being the σ-algebra generated by the set of random samples $\{\boldsymbol{\omega_1},\ldots,\boldsymbol{\omega_t}\}$. Note that $\boldsymbol{x_t}$ and Q_t are $\mathcal{F}_{t-1}$-measurable for all $t\geq 1$.

Lemma 3.4. *Let* $t\geq 1$, *and let* $1\leq\tau\leq\min\left\{\sqrt{T}, t+1\right\}$. *In addition, let* $\theta_t(\tau)$ *be defined as*

$$2(G+D_gR)\tau + \frac{8R^2\alpha_t}{\epsilon\tau} + \frac{8D_fRV_t}{\epsilon} + \frac{4(G+D_gR+\epsilon)^2}{\epsilon}.$$

Then the following statements hold.

(a) For any $t\geq 1$, $Q_t\leq C(2C/\epsilon)^2+(G+\epsilon)t$.

(b) For $t\geq(2C/\epsilon)^2$, $Q_{t+1}-Q_t\leq G+\epsilon$.

(c) For $t\geq(2C/\epsilon)^2$,

$$\mathbb{E}\left[Q_{t+\tau}-Q_t\mid\mathcal{F}_{t-1}\right]\leq\begin{cases}(G+\epsilon)\tau, & \text{if } Q_t\leq\theta_t(\tau)\\ -(\epsilon/4)\tau, & \text{if } Q_t>\theta_t(\tau)\end{cases}.$$

Here, as α_t and V_t increase with t, so does $\theta_t(\tau)$. Note that Lemma 3.4 is a refinement of the drift lemma [25, Lemma 7]. The parameter τ can be any number less than or equal to $\sqrt{T}$ and $t+1$, in contrast to the previous work where τ is fixed at $\sqrt{T}$. Furthermore, we use the drift radius $\theta_t(\tau)$ that varies over time. Based on Lemma 3.4, we show the following lemma that provides a time-varying bound on the expected virtual queue size.

Lemma 3.5. *For any* $t\geq 1$,

$$\mathbb{E}\left[Q_t\right]\leq\theta_t(\lceil\sqrt{t}\rceil)+4(G+\epsilon)\sqrt{t}+\log\frac{128(G+\epsilon)^2}{\epsilon^2}+C\left(2C/\epsilon\right)^2+(G+\epsilon)\left(9+\left(2C/\epsilon\right)^4\right).$$

Plugging in this bound on $\mathbb{E}\left[Q_t\right]$ to Lemmas 3.2 and 3.3, we deduce the proposed bounds on the regret and constraint violation under CDPP.

Theorem 3.6. *For any constant* C *greater than or equal to*

$$D_fD_g+(4D_g^2+8)\left(2G+D_gR+\epsilon+\frac{2R^2+2D_fR}{\epsilon}\right)+1,$$

there exists a constant T_1 that depends only on D_f, D_g, G, R, ϵ such that for any $T \geq T_1$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} g_t(\boldsymbol{x_t})\right] \leq 0$$

under CDPP (Algorithm 1).

Theorem 3.7. *For any choice of constant C,*

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(\boldsymbol{x_t}) - \sum_{t=1}^{T} f_t(\boldsymbol{x}^*)\right] = O(\sqrt{T})$$

under CDPP (Algorithm 1).

4 Two-Point Bandit Feedback

4.1 Problem Setting and Bandit Drift-Plus-Penalty

As an extension of stochastic constrained OCO studied in Section 3, we consider the *two-point bandit optimization* setting proposed in [1]. The extension still follows the basic setup described in Section 2. In particular, we obtain an adversarial sequence of functions $f_1, \ldots, f_T$ while the constraint functions $g_1, \ldots, g_T$ are i.i.d. realizations of $g(\cdot, \boldsymbol{\omega})$. Moreover, we use the same notions of regret and constraint violation.

In the previous setup, referred to as the full information setting, we observe not only the function values $f_t(\boldsymbol{x_t})$ and $g_t(\boldsymbol{x_t})$ but also the gradients $\nabla f_t(\boldsymbol{x_t})$ and $\nabla g_t(\boldsymbol{x_t})$. In contrast, under the bandit optimization setting, we do not have access to the gradients. Instead, we may take two points $\boldsymbol{y_t}, \boldsymbol{z_t} \in \mathcal{X}$ and receive their function values $f_t(\boldsymbol{y_t}), g_t(\boldsymbol{y_t})$ and $f_t(\boldsymbol{z_t}), g_t(\boldsymbol{z_t})$ in each time slot t. From these *bandit* feedback, we estimate the gradients $\nabla f_t(\boldsymbol{x_t})$ and $\nabla g_t(\boldsymbol{x_t})$.

Our algorithm for the bandit setting, which we call Bandit Drift-Plus-Penalty (BDPP, Algorithm 2), is a modification of CDPP (Algorithm 1) based on the framework of [20]. For a simpler presentation of our results, we assume that we may observe $g_t(\boldsymbol{x_t})$ for $t \in [T]$ as well, but we may replace $g_t(\boldsymbol{x_t})$ with $g_t(\boldsymbol{x_t} + \delta_t \boldsymbol{u_t})$ or $g_t(\boldsymbol{x_t} - \delta_t \boldsymbol{u_t})$ and still obtain the same asymptotic performance guarantees.

In Algorithm 2, $\tilde{\nabla} f_t$ and $\tilde{\nabla} g_t$ are the estimates of $\nabla f_t(\boldsymbol{x_t})$ and $\nabla g_t(\boldsymbol{x_t})$ computed using the two-point feedback on f_t and g_t, respectively. The basic outline of Algorithm 2 is the same as that of Algorithm 1 with parameters $V_t = \sqrt{t}$, $\alpha_t = t$, and $\gamma_t = C/\sqrt{t}$, although we would need to set a different value for the constant C. In addition, we will set the radius parameter as

$$\delta_t = \frac{1}{\sqrt{t}}.$$

On top of Assumptions 1 and 2, we assume that functions f_t, g_t are Lipschitz continuous in the ℓ_2-norm and that the fourth moment of $\boldsymbol{u_t}$ is bounded, as in [20].

Assumption 3. *There exist positive constants L_f, L_g, p_* satisfying the following.*

- $\|\nabla f_t(\boldsymbol{x})\|_2 \leq L_f$ *for all* $t \in [T]$ *and* $\boldsymbol{x} \in \mathcal{X}$.

Algorithm 2 Bandit Drift-Plus-Penalty

```
Initialize: Initial iterates $\boldsymbol{x_1} \in \mathcal{X}$, $Q_1 = 0$, step size parameters $\{\alpha_t\}_{t=1}^T$, penalty parameters $\{V_t\}_{t=1}^T$, conservatism parameters $\{\gamma_t\}_{t=1}^T$, and radius parameters $\{\delta_t\}_{t=1}^T$.
for $t = 1$ to $T$ do
    Sample $\boldsymbol{u_t}$ from $\left\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_2 = 1\right\}$ uniformly at random.
    Observe $f_t(\boldsymbol{x})$ and $g_t(\boldsymbol{x})$ for $\boldsymbol{x} \in \{\boldsymbol{x_t} \pm \delta_t \boldsymbol{u_t}\}$ and $g_t(\boldsymbol{x_t})$.
    Set $\tilde{\nabla} f_t$ and $\tilde{\nabla} g_t$ as
        $\tilde{\nabla} f_t := \frac{d}{2\delta_t}\left(f_t(\boldsymbol{x_t} + \delta_t \boldsymbol{u_t}) - f_t(\boldsymbol{x_t} - \delta_t \boldsymbol{u_t})\right)\boldsymbol{u_t}$
        $\tilde{\nabla} g_t := \frac{d}{2\delta_t}\left(g_t(\boldsymbol{x_t} + \delta_t \boldsymbol{u_t}) - g_t(\boldsymbol{x_t} - \delta_t \boldsymbol{u_t})\right)\boldsymbol{u_t}$
    Primal update: set $\boldsymbol{x_{t+1}}$ to be a solution to
        $\min_{\boldsymbol{x} \in \mathcal{X}} \left\{ \left(V_t \tilde{\nabla} f_t + Q_t \tilde{\nabla} g_t\right)^\top \boldsymbol{x} + \alpha_t D(\boldsymbol{x}, \boldsymbol{x_t}) \right\}$
    Dual update: set $Q_{t+1}$ to
        $\left[Q_t + g_t(\boldsymbol{x_t}) + \tilde{\nabla} g_t^\top (\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t\right]_+$
end for
```

- $\|\nabla g(\boldsymbol{x}, \boldsymbol{\omega})\|_2 \leq L_g$ *for all* $\boldsymbol{x} \in \mathcal{X}$ *and* $\boldsymbol{\omega} \in \Omega$.
- $\mathbb{E}_{\boldsymbol{u_t}}\left[\|\boldsymbol{u_t}\|_*^4\right] \leq p_*^4$ *for all* $t \in [T]$.

When $\|\cdot\|$ is the ℓ_2-norm, we may set $L_f = D_f$, $L_g = D_g$, and $p_* = 1$. If $\|\cdot\|$ is the ℓ_1-norm, we may set $L_f = D_f\sqrt{d}$ and $L_g = D_g\sqrt{d}$. Moreover, p_* can be set $p\sqrt{\log d/d}$ for some constant p when $\|\cdot\|$ is the ℓ_1-norm [20, Lemma 4].

4.2 Performance of Bandit Drift-Plus-Penalty

The distinction between Algorithm 1 and Algorithm 2 is that the latter uses stochastic estimates $\tilde{\nabla} f_t$ and $\tilde{\nabla} g_t$ instead of the true gradients $\nabla f_t(\boldsymbol{x_t})$ and $\nabla g_t(\boldsymbol{x_t})$. To adapt the analysis of CDPP to the bandit setting, the first step is to bound $\|\tilde{\nabla} f_t\|_*$ and $\|\tilde{\nabla} g_t\|_*$. Regarding this, the following lemma provides a deterministic bound on $\|\tilde{\nabla} g_t\|_*$. We derive the lemma based on the assumption that g_t is L_g-Lipschitz continuous in the ℓ_2-norm.

Lemma 4.1. *For any* $t \geq 1$, $\|\tilde{\nabla} g_t\|_* \leq L_g \ell$ *where* ℓ *is defined as* $\ell = d \cdot \sup\{\|\boldsymbol{u}\|_* : \|\boldsymbol{u}\|_2 \leq 1\}$.

Note that $L_g\ell = D_g d$ if $\|\cdot\|$ is the ℓ_2-norm and $L_g\ell = D_g d\sqrt{d}$ if $\|\cdot\|$ is the ℓ_1-norm. Lemma 4.1 implies a deterministic bound on $\|\tilde{\nabla} g_t\|_*^2$, that is, $\|\tilde{\nabla} g_t\|_*^2 \leq L_g^2 \ell^2$, but this bound has a (super)quadratic dependence on the dimension d. The next lemma shows that in expectation, $\|\tilde{\nabla} g_t\|_*^2$ can be bounded by a function that has a strictly lower dependence on the dimension d than $L_g^2\ell^2$. Before we state the lemma, let us define filtration $\{\mathcal{H}_t^0 : t \geq 0\}$ where $\mathcal{H}_0^0 = \{\emptyset, \Omega\}$ and $\mathcal{H}_t^0$ is the σ-algebra generated by the set of random samples $\{\boldsymbol{\omega_1}, \ldots, \boldsymbol{\omega_t}\} \cup$

$\{\boldsymbol{u_1},\ldots,\boldsymbol{u_{t-1}}\}$.

Lemma 4.2. *[20, Lemma 9] There is a positive constant q that does not depend on d such that for any $t \geq 1$,*

$$\mathbb{E}\left[\|\tilde{\nabla} f_t\|_*^2 \mid \mathcal{H}_t^0\right] \leq qdp_*^2 L_f^2 \quad \textit{and} \quad \mathbb{E}\left[\|\tilde{\nabla} g_t\|_*^2 \mid \mathcal{H}_t^0\right] \leq qdp_*^2 L_g^2.$$

In particular, when $\|\cdot\|$ is the ℓ_2 norm, the expected bounds in Lemma 4.2 are $O(D_f^2 d)$ and $O(D_g^2 d)$. When $\|\cdot\|$ is the ℓ_1 norm, the expected bounds are $O(D_f^2 d\log d)$ and $O(D_g^2 d\log d)$. Hereinafter, we treat the dimension d as a constant to focus on the dependence on the time length T.

Next, as for CDPP, we prove upper bounds on the expected regret and constraint violation under Bandit Drift-Plus-Penalty that exhibit the dependence on the parameter C and the expected virtual queue length $\mathbb{E}\left[Q_t\right]$.

Lemma 4.3. *Under BDPP (Algorithm 2), the expected regret $\mathbb{E}\left[\mathrm{regret}(T)\right]$ is at most*

$$\left(\frac{C(G+C)}{\epsilon}\right)^2 + K_2\sqrt{T} + (C+2L_g)\sum_{t=1}^{T}\frac{\mathbb{E}[Q_t]}{t}$$

for some positive constant K_2 that depends only on G, D_g, D_f, R, ϵ and L_g, L_f, q, d, p_*.

Lemma 4.4. *Under BDPP (Algorithm 2),*

$$\sum_{t=1}^{T} g_t(\boldsymbol{x_t}) \leq Q_{T+1} + \sum_{t=1}^{T}\frac{\|\tilde{\nabla} g_t\|_*^2 Q_t}{2t} + \sum_{t=1}^{T}\frac{\|\tilde{\nabla} g_t\|_*^2 + \|\tilde{\nabla} f_t\|_*^2}{4\sqrt{t}} - C\sqrt{T}.$$

Note that the expectation of the right-hand side of the inequality in Lemma 4.4 can be bounded using Lemma 4.2, providing an upper bound on the expected constraint violation under BDPP. Given Lemmas 4.3 and 4.4, the remaining task is to bound the expected virtual queue length $\mathbb{E}\left[Q_t\right]$.

As Lemma 3.4, we prove a drift lemma for the bandit setting, which shows time-varying bounds on the drift. In addition, we define filtration $\{\mathcal{H}_t : t \geq 0\}$ where $\mathcal{H}_0 = \{\emptyset, \Omega\}$ and $\mathcal{H}_t$ is the σ-algebra generated by the set of random samples $\{\boldsymbol{\omega_1},\ldots,\boldsymbol{\omega_t}\} \cup \{\boldsymbol{u_1},\ldots,\boldsymbol{u_t}\}$.

Lemma 4.5. *Let $t \geq 1$, and let $1 \leq \tau \leq \min\left\{\sqrt{T}, t+1\right\}$. In addition, let $\theta_t(\tau)$ be defined as*

$$(2G + R^2 + qdp_*^2 L_g^2)\tau + \frac{8R^2\alpha_t}{\epsilon\tau} + \frac{4(R^2 + qdp_*^2 L_f^2)V_t}{\epsilon} + \frac{4\left((G+\epsilon)^2 + R^2 qdp_*^2 L_g^2\right)}{\epsilon}.$$

Then the following statements hold.

(a) For any $t \geq 1$,

$$Q_t \leq C(2(C+2L_g)/\epsilon)^2 + (G + RL_g\ell + \epsilon)t.$$

(b) For $t \geq (2(C+2L_g)/\epsilon)^2$,

$$Q_{t+1} - Q_t \leq G + RL_g\ell + \epsilon.$$

(c) For $t \geq (2(C+2L_g)/\epsilon)^2$,

$$\mathbb{E}\left[Q_{t+\tau} - Q_t \mid \mathcal{H}_{t-1}\right] \leq \begin{cases} (G + R^2/2 + qdp_*^2 L_g^2/2 + \epsilon)\tau, & \textit{if } Q_t \leq \theta_t(\tau) \\ -(\epsilon/4)\tau, & \textit{if } Q_t > \theta_t(\tau) \end{cases}.$$

Note that the drift radius and the drift upper bounds have a dependence on the dimension d, which is due to the usage of stochastic estimates $\tilde{\nabla} f_t$ and $\tilde{\nabla} g_t$. In particular, the deterministic bounds (parts (a) and (b)) rely on Lemma 4.1, while the stochastic drift bound (part (c)) uses Lemma 4.2.

Based on the drift lemma (Lemma 4.5), we deduce a time-varying bound on the expected virtual queue length $\mathbb{E}[Q_t]$.

Lemma 4.6. *For any $t \geq 1$, $\mathbb{E}[Q_t]$ is bounded above by*

$$\theta_t(\lceil\sqrt{t}\rceil) + 4(G + RL_g\ell + \epsilon)\sqrt{t} + \log\frac{128(G + RL_g\ell + \epsilon)^2}{\epsilon^2}$$
$$+ C\left(2(C + 2L_g)/\epsilon\right)^2 + (G + RL_g\ell + \epsilon)\left(9 + (2(C + 2L_g)/\epsilon)^4\right).$$

In comparison to the full information setting (Lemma 3.5), the bound on $\mathbb{E}[Q_t]$ grows as the dimension d gets large, and the growth rate is given by ℓ.

Theorem 4.7. *For any constant C greater than or equal to*

$$(2qdp_*^2L_g^2 + 4)\left(4G + R^2 + qdp_*^2L_g^2 + 2\epsilon + 2RL_g\ell + \frac{12R^2 + 4qdp_*^2L_f^2}{\epsilon}\right) + \frac{qdp_*^2(L_f^2 + L_g^2)}{2} + 1.$$

Then there exists a constant T_2 that depends only on D_f, D_g, G, R, ϵ and $L_f, L_g, q, p_*, d, \ell$ such that for any $T \geq T_2$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} g_t(\boldsymbol{x_t})\right] \leq 0$$

under BDPP (Algorithm 2).

Theorem 4.8. *For any choice of constant C,*

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(\boldsymbol{x_t}) - \sum_{t=1}^{T} f_t(\boldsymbol{x}^*)\right] = O(\sqrt{T})$$

under BDPP (Algorithm 2).

5 Numerical Experiments

We test the performance of our conservative drift-plus-penalty algorithm (Algorithm 1). We consider an online scheduling problem with loss and constraint functions are given by

$$f_t(\boldsymbol{x}) = \boldsymbol{c_t}^\top \boldsymbol{x} \quad \text{and} \quad g_t(\boldsymbol{x}) = n_t - \sum_{i=1}^{15} h(x_i)$$

where $\boldsymbol{x} \in \mathbb{R}^{15}$ represents the vector of power assigned to 15 locations, $\boldsymbol{c_t} \in \mathbb{R}_+^{15}$ is the vector of per unit power electricity costs at time t, n_t is the i.i.d. random number of jobs given at time slot t, and h is a concave function representing the number of jobs that can be done by power allocation $\boldsymbol{x}$. We compare the following methods.

- DPP: the vanilla drift-plus-penalty algorithm due to [25] with $V_t = \sqrt{T}$, $\alpha_t = T$, and $\gamma_t = 0$,

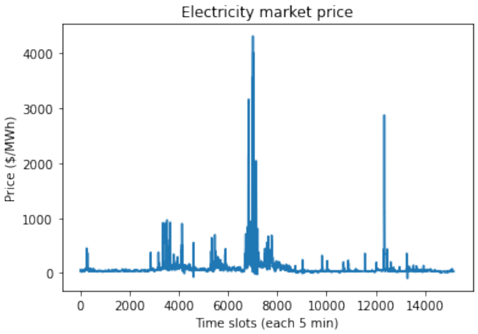
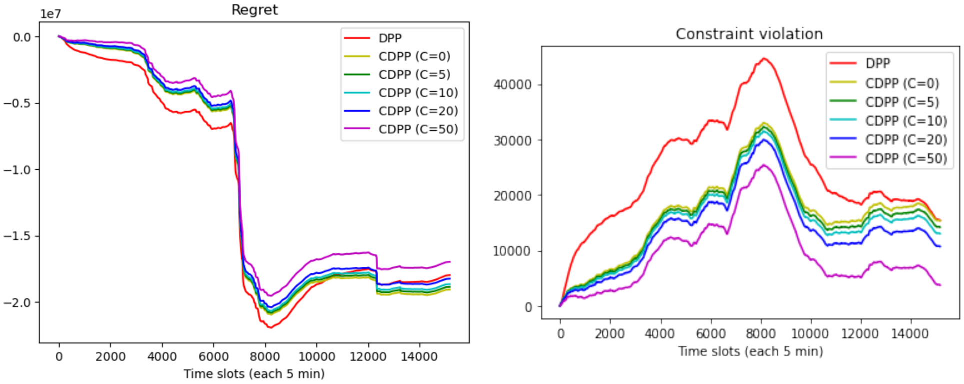


Figure 1: Average electricity market price



(a) Regret

(b) Constraint violation

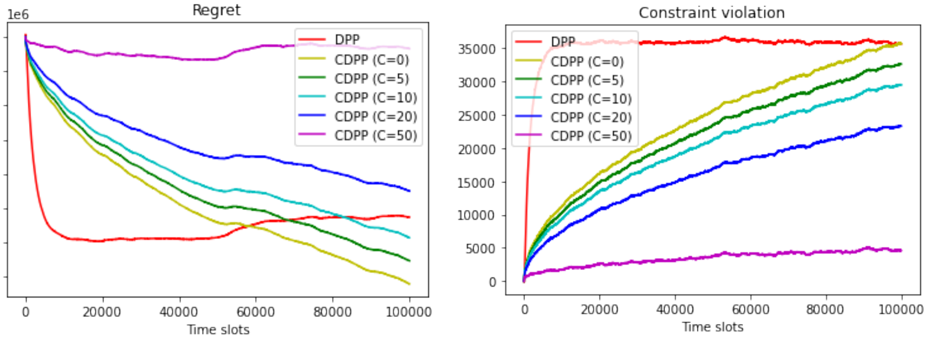
Figure 2: Regret and constraint violation under DPP and CDPP with various values of C for the electricity market price data

- CDPP: our conservative drift-plus-penalty algorithm (Algorithm 1) with $V_t = \sqrt{t}$, $\alpha_t = t$, and $\gamma_t = C/\sqrt{t}$. We test various values for $C \in \{0, 5, 10, 20, 50\}$.

For the first set of experiments, we use the electricity cost data for vectors $\{\boldsymbol{c_t}\}_{t=1}^{T}$ from New York ISO (`http://www.nyiso.com/`), following the experiment setup in [25]. The data consist of electricity costs at 15 different zones every 5 minutes during 2022/12/01-2023/01/20. This corresponds to $T \simeq 15,000$. The average price across 15 zones during this period is shown in Fig. 1.

For the second set of experiments, we synthetically generate random cost vectors. We obtain $T = 100,000$ cost vectors whose coordinates are sampled from the normal distribution with a mean of 30 and standard deviation of 10, followed by projection onto $[0, 60]$. Under this setting, as cost vectors $\boldsymbol{c_t}$ are i.i.d., the loss functions are not adversarial but stochastic.

We further assume that the power allocation vector $\boldsymbol{x}$ is contained in the set $\mathcal{X} = [0, 10]^{15}$ due to some hardware restriction and that $h(x) = \log(1 + x)$. In addition, n_t is sampled from the Poisson distribution with parameter 20 followed by projection onto $[0, 35]$. We use the ℓ_2-norm, in which case $D(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$.



(a) Regret

(b) Constraint violation

Figure 3: Regret and constraint violation under DPP and CDPP with various values of C for synthetic data

Results from the experiments are summarized in Fig. 2 and 3. Fig. 2 show the results from the experiments with the real data, while Fig. 3 is for the second set of experiments with synthetic data.

Fig. 2a and 3a show that increasing C results in a higher regret from CDPP, conforming to the theory. Comparing DPP and CDPP, we observe that DPP incurs the smallest regret at the beginning but the regret of CDPP with a small C value (0,5,10) later becomes smaller than that of DPP.

We observe from Fig. 2b that increasing the conservatism parameter C reduces constraint violation. In fact, CDPP with $C = 0$ incurs a smaller amount of constraint violation than DPP, which perhaps indicates that using time-varing penalty parameter V_t and step size parameter α_t is helpful for decreasing constraint violation. As in the real data setting, Fig. 3b for the synthetic data setting shows that DPP incurs the largest constraint violation while CDPP with a higher C has a smaller constraint violation.

Looking at Fig. 2a and 3a, it may seem strange at first glance that the regret goes below zero. This is because the benchmark $\boldsymbol{x}^*$ is an optimal solution of (2) not (1). On the other hand, the regret and constraint violation under CDPP with $C = 50$ stay close to zero. This is perhaps because, when C gets large, the decisions made by CDPP more likely to satisfy the constraint in (2).

Lastly, we briefly discuss the non-uniform growth pattern in Fig. 2b. This is perhaps due to the irregular changes in cost vectors, shown in Fig. 1, resulting in abrupt changes in the loss functions. More precisely, when $\boldsymbol{c_t}$ soars at t, the algorithms would set $\boldsymbol{x_{t+1}}$ small accordingly because $\boldsymbol{c_t} = \nabla f_t(\boldsymbol{x_t})$. Then $g_{t+1}(\boldsymbol{x_{t+1}})$ would be large in response.

6 Conclusion

This paper studies online convex optimization with stochastic constraints and develops a varaint of the drift-plus-penalty algorithm that achieves zero constraint violation in expectation while still bounding the expected regret by $O(\sqrt{T})$. The algorithm is oblivious to the time horizon length T, in contrast to the vanilla drift-plus-penalty algorithm. The numerical experiments show that using the time-varying algorithm parameters and the conservatism parameter leads to a significant reduction in constraint violation, without much sacrifice in regret. We also consider the two-point bandit feedback setting, for which we deduce the same asymptotic bounds on the expected regret and constraint violation.

References

[1] Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Conference on Learning Theory (COLT)*, pages 28–40, 12 2010.

[2] Amit Agarwal, Elad Hazan, Satyen Kale, and Robert E. Schapire. Algorithms for portfolio management based on the newton method. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 9–16, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143846. URL https://doi.org/10.1145/1143844.1143846.

[3] Shipra Agrawal and Nikhil R. Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, page 989–1006, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450325653. doi: 10.1145/2600057.2602844. URL https://doi.org/10.1145/2600057.2602844.

[4] Zeeshan Akhtar, Amrit Singh Bedi, and Ketan Rajawat. Conservative stochastic optimization with expectation constraints. *IEEE Transactions on Signal Processing*, 69:3190–3205, 2021. doi: 10.1109/TSP.2021.3082467.

[5] Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, 2008. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2007.04.016. URL https://www.sciencedirect.com/science/article/pii/S0022000007000621. Learning Theory 2004.

[6] Andrey Bernstein, Shie Mannor, and Nahum Shimkin. Online classification with specificity constraints. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/9cfdf10e8fc047a44b08ed031e1f0ed1-Paper.pdf.

[7] Xuanyu Cao and K. J. Ray Liu. Online convex optimization with time-varying constraints and bandit feedback. *IEEE Transactions on Automatic Control*, 64(7):2665–2680, 2019. doi: 10.1109/TAC.2018.2884653.

[8] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *J. Mach. Learn. Res.*, 7:551–585, dec 2006. ISSN 1532-4435.

[9] Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, page 385–394, USA, 2005. Society for Industrial and Applied Mathematics. ISBN 0898715857.

[10] Avi Goldfarb and Catherine Tucker. Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3):389–404, 2011. doi: 10.1287/mksc.1100.0583. URL https://doi.org/10.1287/

mksc.1100.0583.

[11] Hengquan Guo, Xin Liu, Honghao Wei, and Lei Ying. Online convex optimization with hard constraints: Towards the best of two worlds and beyond. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=rwdpFgfVpvN.

[12] Elad Hazan. Introduction to online convex optimization. *Found. Trends Optim.*, 2(3–4):157–325, aug 2016. ISSN 2167-3888. doi: 10.1561/2400000013. URL https://doi.org/10.1561/2400000013.

[13] Elad Hazan and Satyen Kale. Projection-free online learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1843–1850, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

[14] Rodolphe Jenatton, Jim Huang, and Cedric Archambeau. Adaptive algorithms for online convex optimization with long-term constraints. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 402–411, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/jenatton16.html.

[15] Nikolaos Liakopoulos, Apostolos Destounis, Georgios Paschos, Thrasyvoulos Spyropoulos, and Panayotis Mertikopoulos. Cautious regret minimization: Online optimization with long-term budget constraints. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3944–3952. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/liakopoulos19a.html.

[16] Mehrdad Mahdavi, Rong Jin, and Tianbao Yang. Trading regret for efficiency: Online convex optimization with long term constraints. *Journal of Machine Learning Research*, 13(81):2503–2528, 2012. URL http://jmlr.org/papers/v13/mahdavi12a.html.

[17] Shie Mannor, John N. Tsitsiklis, and Jia Yuan Yu. Online learning with sample path constraints. *Journal of Machine Learning Research*, 10(20):569–590, 2009. URL http://jmlr.org/papers/v10/mannor09a.html.

[18] Michael J. Neely and Hao Yu. Online convex optimization with time-varying constraints, 2017. URL https://arxiv.org/abs/1702.04783.

[19] Shai Shalev-Shwartz. Online learning and online convex optimization. *Found. Trends Mach. Learn.*, 4(2): 107–194, feb 2012. ISSN 1935-8237. doi: 10.1561/2200000018. URL https://doi.org/10.1561/2200000018.

[20] Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017. URL http://jmlr.org/papers/v18/16-632.html.

[21] Victor Valls, George Iosifidis, Douglas Leith, and Leandros Tassiulas. Online convex optimization with

perturbed constraints: Optimal rates against stronger benchmarks. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2885–2895. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/valls20a.html.

[22] Xiaohan Wei, Hao Yu, and Michael J. Neely. Online primal-dual mirror descent under stochastic constraints. In *Abstracts of the 2020 SIGMETRICS/Performance Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '20, page 3–4, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379854. doi: 10.1145/3393691.3394209. URL https://doi.org/10.1145/3393691.3394209.

[23] Xinlei Yi, Xiuxian Li, Tao Yang, Lihua Xie, Tianyou Chai, and Karl Johansson. Regret and cumulative constraint violation analysis for online convex optimization with long term constraints. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11998–12008. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/yi21b.html.

[24] Hao Yu and Michael J. Neely. A low complexity algorithm with o(â^št) regret and o(1) constraint violations for online convex optimization with long term constraints. *Journal of Machine Learning Research*, 21(1):1–24, 2020. URL http://jmlr.org/papers/v21/16-494.html.

[25] Hao Yu, Michael Neely, and Xiaohan Wei. Online convex optimization with stochastic constraints. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/da0d1111d2dc5d489242e60ebcbaf988-Paper.pdf.

[26] Jianjun Yuan and Andrew Lamperski. Online convex optimization for cumulative constraints. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/9cb9ed4f35cf7c2f295cc2bc6f732a84-Paper.pdf.

A Performance Analysis of Conservative Drift-Plus-Penalty (Algorithm 1)

A.1 Proof of Lemma 3.1: Basic Upper Bound on the Drift

As $Q_{t+1} = \max\left\{Q_t + g_t(\boldsymbol{x_t}) + \nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t, 0\right\}$,

$$Q_{t+1}^2 \leq \left(Q_t + g_t(\boldsymbol{x_t}) + \nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t\right)^2.$$

Expanding the right-hand side, we obtain

$$\Delta_t = \frac{Q_{t+1}^2}{2} - \frac{Q_t^2}{2} \leq Q_t(g_t(\boldsymbol{x_t}) + \nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t) + \frac{1}{2}\left(g_t(\boldsymbol{x_t}) + \nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t\right)^2.$$

Here,

$$\begin{aligned}\left(g_t(\boldsymbol{x_t}) + \nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t\right)^2 &\leq (|g_t(\boldsymbol{x_t})| + \|\nabla g_t(\boldsymbol{x_t})\|_* \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\| + \gamma_t)^2 \\ &\leq (G + D_g R + \gamma_t)^2\end{aligned}$$

where the second inequality is due to Assumption 1. Therefore,

$$\Delta_t \leq Q_t(g_t(\boldsymbol{x_t}) + \nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t) + (G + D_g R + \gamma_t)^2,$$

as required.

A.2 Proof of Lemma 3.2: Providing an Upper Bound on the Expected Regret

We will use the following useful lemma to prove Lemma 3.2.

Lemma A.1. *[22, Equation (22)] For any $\boldsymbol{x} \in \mathcal{X}$ and $t \geq 1$,*

$$\begin{aligned}&(V_t \nabla f_t(\boldsymbol{x_t}) + Q_t \nabla g_t(\boldsymbol{x_t}))^\top (\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \alpha_t D(\boldsymbol{x_{t+1}}, \boldsymbol{x_t}) \\ &\leq (V_t \nabla f_t(\boldsymbol{x_t}) + Q_t \nabla g_t(\boldsymbol{x_t}))^\top (\boldsymbol{x} - \boldsymbol{x_t}) + \alpha_t D(\boldsymbol{x}, \boldsymbol{x_t}) - \alpha_t D(\boldsymbol{x}, \boldsymbol{x_{t+1}}).\end{aligned}$$

Using Lemma A.1, we will show that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T f_t(\boldsymbol{x_t}) - \sum_{t=1}^T f_t(\boldsymbol{x^*})\right] \leq \frac{\alpha_T}{V_T}R^2 + \sum_{t=1}^T \frac{V_t D_f^2}{4\alpha_t} + \sum_{t=1}^T \frac{\gamma_t}{V_t}\mathbb{E}[Q_t] \\ + \left(\frac{C(G + D_g R + C)}{\epsilon}\right)^2 + \sum_{t=1}^T \frac{(G + D_g R + \epsilon)^2}{V_t}\end{aligned} \tag{4}$$

holds. We know from basic calculus that

$$\sqrt{T} \leq \sum_{t=1}^T \frac{1}{\sqrt{t}} \leq 2\sqrt{T},$$

and therefore, (4) implies that

$$\mathbb{E}\left[\sum_{t=1}^T f_t(\boldsymbol{x_t}) - \sum_{t=1}^T f_t(\boldsymbol{x^*})\right] \leq \left(\frac{C(G + D_g R + C)}{\epsilon}\right)^2 + K\sqrt{T} + C\sum_{t=1}^T \frac{\mathbb{E}[Q_t]}{t}$$

where

$$K = R^2 + \frac{D_f^2}{2} + 2(G + D_g R + \epsilon)^2.$$

Now let us prove that (4) holds. Adding $V_t f_t(\boldsymbol{x_t}) + Q_t g_t(\boldsymbol{x_t})$ to both sides of the inequality given in Lemma A.1, we obtain

$$\begin{aligned} &V_t f_t(\boldsymbol{x_t}) + Q_t g_t(\boldsymbol{x_t}) + \left(V_t \nabla f_t(\boldsymbol{x_t}) + Q_t \nabla g_t(\boldsymbol{x_t})\right)^\top (\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \alpha_t D(\boldsymbol{x_{t+1}}, \boldsymbol{x_t}) \\ &\le V_t f_t(\boldsymbol{x_t}) + Q_t g_t(\boldsymbol{x_t}) + \left(V_t \nabla f_t(\boldsymbol{x_t}) + Q_t \nabla g_t(\boldsymbol{x_t})\right)^\top (\boldsymbol{x} - \boldsymbol{x_t}) + \alpha_t D(\boldsymbol{x}, \boldsymbol{x_t}) - \alpha_t D(\boldsymbol{x}, \boldsymbol{x_{t+1}}). \end{aligned} \tag{5}$$

Here, as f_t and g_t are convex, we have $f_t(\boldsymbol{x_t}) + \nabla f_t(\boldsymbol{x_t})^\top (\boldsymbol{x} - \boldsymbol{x_t}) \le f_t(\boldsymbol{x})$ and $g_t(\boldsymbol{x_t}) + \nabla g_t(\boldsymbol{x_t})^\top (\boldsymbol{x} - \boldsymbol{x_t}) \le g_t(\boldsymbol{x})$. Then the right-hand side of (5) is bounded above by

$$V_t f_t(\boldsymbol{x}) + Q_t g_t(\boldsymbol{x}) + \alpha_t D(\boldsymbol{x}, \boldsymbol{x_t}) - \alpha_t D(\boldsymbol{x}, \boldsymbol{x_{t+1}}).$$

Moreover, by Lemma 3.1, the left-hand side of (5) is greater than or equal to

$$V_t f_t(\boldsymbol{x_t}) + V_t \nabla f_t(\boldsymbol{x_t})^\top (\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \alpha_t D(\boldsymbol{x_{t+1}}, \boldsymbol{x_t}) + \Delta_t - Q_t \gamma_t - (G + D_g R + \gamma_t)^2.$$

Therefore, it follows that

$$\begin{aligned} V_t f_t(\boldsymbol{x_t}) \le\ & V_t f_t(\boldsymbol{x}) + Q_t g_t(\boldsymbol{x}) + \alpha_t D(\boldsymbol{x}, \boldsymbol{x_t}) - \alpha_t D(\boldsymbol{x}, \boldsymbol{x_{t+1}}) \\ & - V_t \nabla f_t(\boldsymbol{x_t})^\top (\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) - \alpha_t D(\boldsymbol{x_{t+1}}, \boldsymbol{x_t}) - \Delta_t + Q_t \gamma_t + (G + D_g R + \gamma_t)^2. \end{aligned} \tag{6}$$

In the right-hand side of (6), the part $-V_t \nabla f_t(\boldsymbol{x_t})^\top (\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) - \alpha_t D(\boldsymbol{x_{t+1}}, \boldsymbol{x_t})$ can be bounded as follows.

$$\begin{aligned} -V_t \nabla f_t(\boldsymbol{x_t})^\top (\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) - \alpha_t D(\boldsymbol{x_{t+1}}, \boldsymbol{x_t}) &\le V_t \|\nabla f_t(\boldsymbol{x_t})\|_* \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\| - \alpha_t \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|^2 \\ &\le V_t D_f \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\| - \alpha_t \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|^2 \\ &= \frac{V_t^2 D_f^2}{4\alpha_t} - \alpha_t \left(\frac{V_t D_f}{2\alpha_t} - \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\| \right)^2 \\ &\le \frac{V_t^2 D_f^2}{4\alpha_t} \end{aligned} \tag{7}$$

where the first inequality holds due to $\boldsymbol{u}^\top \boldsymbol{v} \le \|\boldsymbol{u}\|_* \|\boldsymbol{v}\|$ and (3) while the second inequality is from Assumption 1. Then (6) and (7) imply that

$$V_t f_t(\boldsymbol{x_t}) \le V_t f_t(\boldsymbol{x}) + Q_t g_t(\boldsymbol{x}) + \alpha_t D(\boldsymbol{x}, \boldsymbol{x_t}) - \alpha_t D(\boldsymbol{x}, \boldsymbol{x_{t+1}}) + \frac{V_t^2 D_f^2}{4\alpha_t} - \Delta_t + Q_t \gamma_t + (G + D_g R + \gamma_t)^2. \tag{8}$$

Dividing both sides of (8) and summing the resulting inequality for $t = 1, \ldots, T$, we obtain the following inequality.

$$\begin{aligned} \sum_{t=1}^T f_t(\boldsymbol{x_t}) \le \sum_{t=1}^T f_t(\boldsymbol{x}) + \sum_{t=1}^T \frac{Q_t}{V_t} g_t(\boldsymbol{x}) + \sum_{t=1}^T \frac{\alpha_t}{V_t} \left(D(\boldsymbol{x}, \boldsymbol{x_t}) - D(\boldsymbol{x}, \boldsymbol{x_{t+1}}) \right) \\ + \sum_{t=1}^T \frac{V_t D_f^2}{4\alpha_t} - \sum_{t=1}^T \frac{\Delta_t}{V_t} + \sum_{t=1}^T \frac{Q_t \gamma_t}{V_t} + \sum_{t=1}^T \frac{(G + D_g R + \gamma_t)^2}{V_t}. \end{aligned} \tag{9}$$

What remains is to derive the desired upper bound on the right-hand side of (9). First,

$$\mathbb{E}\left[\frac{Q_t}{V_t}g_t(\boldsymbol{x}^*)\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{Q_t}{V_t}g_t(\boldsymbol{x}^*) \mid \mathcal{F}_{t-1}\right]\right] = \mathbb{E}\left[\frac{Q_t}{V_t}\mathbb{E}\left[g_t(\boldsymbol{x}^*) \mid \mathcal{F}_{t-1}\right]\right] = \mathbb{E}\left[\frac{Q_t}{V_t}\bar{g}(\boldsymbol{x}^*)\right] \leq 0 \tag{10}$$

where the first equality comes from the tower rule, the second equality is because Q_t is $\mathcal{F}_{t-1}$-measurable, and the last equality holds as $\boldsymbol{\omega_t}$ is independent of $\{\boldsymbol{\omega_1}, \ldots \boldsymbol{\omega_{t-1}}\}$. Next,

$$\begin{aligned}\sum_{t=1}^{T}\frac{\alpha_t}{V_t}\left(D(\boldsymbol{x},\boldsymbol{x_t}) - D(\boldsymbol{x},\boldsymbol{x_{t+1}})\right) &= \frac{\alpha_1}{V_1}D(\boldsymbol{x},\boldsymbol{x_1}) + \sum_{t=2}^{T}D(\boldsymbol{x},\boldsymbol{x_t})\left(\frac{\alpha_t}{V_t} - \frac{\alpha_{t-1}}{V_{t-1}}\right) - \frac{\alpha_T}{V_T}D(\boldsymbol{x},\boldsymbol{x_{T+1}}) \\ &\leq \frac{\alpha_1}{V_1}R^2 + \sum_{t=2}^{T}R^2\left(\frac{\alpha_t}{V_t} - \frac{\alpha_{t-1}}{V_{t-1}}\right) \\ &= \frac{\alpha_T}{V_T}R^2\end{aligned} \tag{11}$$

where the second inequality holds because $\alpha_t/V_t - \alpha_{t-1}/V_{t-1} = \sqrt{t} - \sqrt{t-1} > 0$ and $D(\boldsymbol{x},\boldsymbol{x_t}) \leq R^2$. Furthermore,

$$\begin{aligned}\sum_{t=1}^{T}\frac{\Delta_t}{V_t} &= \frac{1}{2}\sum_{t=1}^{T}\frac{1}{V_t}(Q_{t+1}^2 - Q_t^2) \\ &= -\frac{1}{2V_1}Q_1^2 + \frac{1}{2V_T}Q_{T+1}^2 + \frac{1}{2}\sum_{t=2}^{T}Q_t^2\left(\frac{1}{V_{t-1}} - \frac{1}{V_t}\right) \\ &\geq -\frac{1}{2V_1}Q_1^2 \\ &= 0\end{aligned} \tag{12}$$

where the inequality holds because $Q_t \geq 0$ for all $t \geq 0$ and $1/V_{t-1} - 1/V_t = 1/\sqrt{t-1} - 1/\sqrt{t} \geq 0$. Lastly,

$$\begin{aligned}\sum_{t=1}^{T}\frac{(G + D_gR + \gamma_t)^2}{V_t} &= \sum_{t=1}^{\lfloor (C/\epsilon)^2\rfloor}\frac{(G + D_gR + \gamma_t)^2}{V_t} + \sum_{t=\lceil (C/\epsilon)^2\rceil}^{T}\frac{(G + D_gR + \gamma_t)^2}{V_t} \\ &\leq \sum_{t=1}^{\lfloor (C/\epsilon)^2\rfloor}(G + D_gR + C)^2 + \sum_{t=\lceil (C/\epsilon)^2\rceil}^{T}\frac{(G + D_gR + \epsilon)^2}{V_t} \\ &\leq (G + D_gR + C)^2(C/\epsilon)^2 + \sum_{t=1}^{T}\frac{(G + D_gR + \epsilon)^2}{V_t}\end{aligned} \tag{13}$$

where the first inequality holds because for $t \geq (C/\epsilon)^2$,

$$\gamma_t = \frac{C}{\sqrt{t}} \leq \epsilon$$

and $V_t \geq 1$ and $\gamma_t \leq C$ for $t \geq 1$. Taking the expectation of both sides of (9) with $\boldsymbol{x} = \boldsymbol{x}^*$ and using the bounds (10)–(13),

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}f_t(\boldsymbol{x_t})\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}f_t(\boldsymbol{x}^*)\right] + \frac{\alpha_T}{V_T}R^2 + \sum_{t=1}^{T}\frac{V_tD_f^2}{4\alpha_t} + \sum_{t=1}^{T}\frac{\gamma_t}{V_t}\mathbb{E}[Q_t] \\ + \left(\frac{C(G + D_gR + C)}{\epsilon}\right)^2 + \sum_{t=1}^{T}\frac{(G + D_gR + \epsilon)^2}{V_t},\end{aligned}$$

which proves (4), as required.

A.3 Proof of Lemma 3.3: Giving an Upper Bound on the Expected Constraint Violation

We will show that

$$\sum_{t=1}^{T} g_t(\boldsymbol{x_t}) \leq Q_{T+1} + \sum_{t=1}^{T} \frac{D_f D_g V_t + D_g^2 Q_t}{2\alpha_t} - \sum_{t=1}^{T} \gamma_t \tag{14}$$

holds. This would imply that

$$\mathbb{E}\left[\sum_{t=1}^{T} g_t(\boldsymbol{x_t})\right] \leq \mathbb{E}\left[Q_{T+1}\right] + D_f D_g \sqrt{T} + \frac{D_g^2}{2}\sum_{t=1}^{T}\frac{\mathbb{E}\left[Q_t\right]}{t} - C\sqrt{T},$$

as required. Therefore, it suffices to show that (14) holds. Since $Q_{t+1} \geq Q_t + g_t(\boldsymbol{x_t}) + \nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t$, we have

$$\begin{aligned} g_t(\boldsymbol{x_t}) &\leq Q_{t+1} - Q_t - \nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) - \gamma_t \\ &\leq Q_{t+1} - Q_t + \|\nabla g_t(\boldsymbol{x_t})\|_* \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\| - \gamma_t \\ &\leq Q_{t+1} - Q_t + D_g \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\| - \gamma_t \end{aligned} \tag{15}$$

where the second inequality is due to the fact that $\boldsymbol{u}^\top \boldsymbol{v} \leq \|\boldsymbol{u}\|_* \|\boldsymbol{v}\|$ for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ and the last inequality is by Assumption 1. Summing (15) over $t = 1, \ldots, T$, we obtain

$$\sum_{t=1}^{T} g_t(\boldsymbol{x_t}) \leq Q_{T+1} - Q_1 + D_g \sum_{t=1}^{T} \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\| - \sum_{t=1}^{T} \gamma_t. \tag{16}$$

To provide an upper bound on the term $\|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|$, we use the inequality in Lemma A.1 with $\boldsymbol{x} = \boldsymbol{x_t}$, which implies that

$$\alpha_t D(\boldsymbol{x_{t+1}}, \boldsymbol{x_t}) + \alpha_t D(\boldsymbol{x_t}, \boldsymbol{x_{t+1}}) \leq \left(V_t \nabla f_t(\boldsymbol{x_t}) + Q_t \nabla g_t(\boldsymbol{x_t})\right)^\top \left(\boldsymbol{x_t} - \boldsymbol{x_{t+1}}\right). \tag{17}$$

The left-hand side of (17) is greater than or equal to $2\alpha_t\|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|^2$ because $D(\boldsymbol{x_{t+1}}, \boldsymbol{x_t}) \geq \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|^2$ and $D(\boldsymbol{x_t}, \boldsymbol{x_{t+1}}) \geq \|\boldsymbol{x_t} - \boldsymbol{x_{t+1}}\|^2$ by (3). The right-hand side can be bounded by

$$\left\|V_t \nabla f_t(\boldsymbol{x_t}) + Q_t \nabla g_t(\boldsymbol{x_t})\right\|_* \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\| \leq (D_f V_t + D_g Q_t)\|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|$$

where the inequality is due to Assumption 1. Then we obtain from (17) that

$$2\alpha_t\|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|^2 \leq (D_f V_t + D_g Q_t)\|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|,$$

implying in turn that

$$\|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\| \leq \frac{1}{2\alpha_t}(D_f V_t + D_g Q_t). \tag{18}$$

Then (16) and (18) together with $Q_1 = 0$ imply

$$\sum_{t=1}^{T} g_t(\boldsymbol{x_t}) \leq Q_{T+1} + \frac{D_f D_g}{2}\sum_{t=1}^{T}\frac{V_t}{\alpha_t} + \frac{D_g^2}{2}\sum_{t=1}^{T}\frac{Q_t}{\alpha_t} - \sum_{t=1}^{T}\gamma_t,$$

so (14) holds, as required.

A.4 Proof of Lemma 3.4: Time-Varying Drift Lemma

Recall that $\hat{\boldsymbol{x}} \in \mathcal{X}$ is a solution satisfying Slater's condition (Assumption 2), i.e., $\mathbb{E}_{\boldsymbol{\omega}}[g(\hat{\boldsymbol{x}}, \boldsymbol{\omega})] \leq -\epsilon$. Assume that $T \geq (2C/\epsilon)^4$. Then for any $t \geq \lceil \sqrt{T} \rceil$,

$$\gamma_t = \frac{C}{\sqrt{t}} \leq \frac{C}{T^{1/4}} \leq \frac{\epsilon}{2}.$$

Lemma A.2. *For any $t \geq (2C/\epsilon)^2$ and $s \geq 0$,*

$$\mathbb{E}\left[Q_{t+s}\left(g_{t+s}(\hat{\boldsymbol{x}}) + \gamma_{t+s}\right) \mid \mathcal{F}_{t-1}\right] \leq -\frac{\epsilon}{2} \cdot \mathbb{E}\left[Q_{t+s} \mid \mathcal{F}_{t-1}\right]$$

Proof. Since $t \geq (2C/\epsilon)^2$, we have

$$\gamma_t = \frac{C}{\sqrt{t}} \leq \frac{\epsilon}{2}.$$

Note that

$$\begin{aligned}
\mathbb{E}\left[Q_{t+s}\left(g_{t+s}(\hat{\boldsymbol{x}}) + \gamma_{t+s}\right) \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\mathbb{E}\left[Q_{t+s}\left(g_{t+s}(\hat{\boldsymbol{x}}) + \gamma_{t+s}\right) \mid \mathcal{F}_{t+s-1}\right] \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[Q_{t+s} \cdot \mathbb{E}\left[g_{t+s}(\hat{\boldsymbol{x}}) + \gamma_{t+s} \mid \mathcal{F}_{t+s-1}\right] \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[Q_{t+s}\left(\bar{g}(\hat{\boldsymbol{x}}) + \gamma_{t+s}\right) \mid \mathcal{F}_{t-1}\right] \\
&= \left(\bar{g}(\hat{\boldsymbol{x}}) + \gamma_{t+s}\right) \cdot \mathbb{E}\left[Q_{t+s} \mid \mathcal{F}_{t-1}\right]
\end{aligned}$$

where the first equality is from the tower rule, the second equality holds because Q_{t+s} is $\mathcal{F}_{t+s-1}$-measurable, the third equality holds because $\boldsymbol{\omega_{t+s}}$ is independent of $\{\boldsymbol{\omega_1}, \ldots, \boldsymbol{\omega_{t+s-1}}\}$, and the last equality is because $\bar{g}(\hat{\boldsymbol{x}}) + \gamma_{t+s}$ is constant. By Slater's condition (Assumption 2), we have $\bar{g}(\hat{\boldsymbol{x}}) \leq -\epsilon$. As $\gamma_t \leq \epsilon/2$ for any $t \geq (C/2\epsilon)^2$, it follows that $\bar{g}(\hat{\boldsymbol{x}}) + \gamma_{t+s} \leq -\epsilon/2$. Since Q_{t+s} is always nonnegative,

$$\mathbb{E}\left[Q_{t+s}\left(g_{t+s}(\hat{\boldsymbol{x}}) + \gamma_{t+s}\right) \mid \mathcal{F}_{t-1}\right] = \left(\bar{g}(\hat{\boldsymbol{x}}) + \gamma_{t+s}\right) \cdot \mathbb{E}\left[Q_{t+s} \mid \mathcal{F}_{t-1}\right] \leq -\frac{\epsilon}{2} \cdot \mathbb{E}\left[Q_{t+s} \mid \mathcal{F}_{t-1}\right],$$

as required. □

Lemma A.3. *For $t \geq 1$,*

$$-G - D_g R \leq Q_{t+1} - Q_t \leq G + \gamma_t$$

Proof. Let us first show the upper bound on $Q_{t+1} - Q_t$. As $Q_{t+1} = \max\left\{Q_t + g_t(\boldsymbol{x_t}) + \nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t, 0\right\}$ and $g_t(\boldsymbol{x_t}) + \nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) \leq g_t(\boldsymbol{x_{t+1}}) \leq G$ by convexity, $Q_{t+1} \leq \max\left\{Q_t + G + \gamma_t, 0\right\}$ holds. Then it follows that $Q_{t+1}^2 \leq (Q_t + G + \gamma_t)^2$, implying in turn that $Q_{t+1} \leq Q_t + G + \gamma_t$. For the lower bound, we deduce from $Q_{t+1} = \max\left\{Q_t + g_t(\boldsymbol{x_t}) + \nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t, 0\right\}$ that

$$Q_{t+1} \geq Q_t + g_t(\boldsymbol{x_t}) + \nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t \geq Q_t - G - D_g R$$

where the second inequality is comes from $g_t(\boldsymbol{x_t}) \geq -G$, $\nabla g_t(\boldsymbol{x_t})^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) \geq -\|g_t(\boldsymbol{x_t})\|_* \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\| \geq -D_g R$, and $\gamma_t \geq 0$. Therefore, $Q_{t+1} - Q_t \geq -G - D_g R$ holds as required. □

Proof of Lemma 3.4. **Part (a).** Recall that $Q_1 = 0$. Let $t \geq 2$. By Lemma A.3, we know that $Q_{s+1} - Q_s \leq G + \gamma_s$

holds for $s \geq 1$. Summing this inequality over $s = 1, \ldots, t-1$ and using $Q_1 = 0$,

$$Q_t = Q_t - Q_1 = \sum_{s=1}^{t-1}(Q_{s+1} - Q_s) \leq \sum_{s=1}^{t-1}(G + \gamma_s).$$

Recall that $\gamma_s \leq \epsilon/2$ for $s \geq (2C/\epsilon)^2$. Moreover, $\gamma_s \leq C$ for any $s \geq 1$. Then it follows that

$$\begin{aligned}
\sum_{s=1}^{t-1}(G + \gamma_s) &= \sum_{s=1}^{\lfloor (2C/\epsilon)^2 \rfloor}(G + \gamma_s) + \sum_{s=\lfloor (2C/\epsilon)^2 \rfloor + 1}^{t-1}(G + \gamma_s)\\
&= (t-1)G + \sum_{s=1}^{\lfloor (2C/\epsilon)^2 \rfloor}\gamma_s + \sum_{s=\lfloor (2C/\epsilon)^2 \rfloor + 1}^{t-1}\gamma_s\\
&\leq (t-1)G + \sum_{s=1}^{\lfloor (2C/\epsilon)^2 \rfloor} C + \sum_{s=\lfloor (2C/\epsilon)^2 \rfloor + 1}^{t-1}\frac{\epsilon}{2}\\
&\leq (t-1)G + C(2C/\epsilon)^2 + (t-1)(\epsilon/2)
\end{aligned}$$

Since $t - 1 \leq t$, this inequality implies that

$$Q_t \leq \sum_{s=1}^{t-1}(G + \gamma_s) \leq C(2C/\epsilon)^2 + (G + \epsilon/2)t,$$

as required.

Part (b). Since $t \geq (2C/\epsilon)^2$, we have $\gamma_t \leq \epsilon/2$. Then the upper bound $Q_{t+1} - Q_t \leq G + \gamma_t$ from Lemma A.3 implies that $Q_{t+1} - Q_t \leq G + \epsilon/2$.

Part (c). Let $t \geq (2C/\epsilon)^2$. By part (b), it is straightforward that for any t and τ,

$$Q_{t+\tau} - Q_t = \sum_{s=t}^{t+\tau-1}(Q_{s+1} - Q_s) \leq (G + \epsilon)\tau$$

Now suppose that $Q_t \geq \theta_t(\tau)$. Recall that $1 \leq \tau \leq t+1$. We will show that when $Q_t \geq \theta_t(\tau)$,

$$\mathbb{E}\left[Q_{t+\tau}^2 \mid \mathcal{F}_{t-1}\right] \leq \left(Q_t - \frac{\epsilon}{4}\tau\right)^2. \tag{19}$$

If (19) holds, as $\sqrt{\cdot}$ is a concave function over $\mathbb{R}_+$, we deduce from Jensen's inequality that

$$\mathbb{E}\left[Q_{t+\tau} \mid \mathcal{F}_{t-1}\right] \leq \sqrt{\mathbb{E}\left[Q_{t+\tau}^2 \mid \mathcal{F}_{t-1}\right]} \leq Q_t - \frac{\epsilon}{4}\tau.$$

Therefore, it is sufficient to show that (19) holds true. Note that

$$Q_{t+\tau}^2 = Q_t^2 + 2\sum_{s=t}^{t+\tau-1}\Delta_s,$$

so to provide the desired bound on $Q_{t+\tau}^2$, we analyze the drift terms. Lemma 3.1 and the observation that $\gamma_t \leq \epsilon/2$ for any $t \geq (2C/\epsilon)^2$ imply that

$$\Delta_s \leq Q_s\left(g_s(\boldsymbol{x_s}) + \nabla g_s(\boldsymbol{x_s})^\top(\boldsymbol{x_{s+1}} - \boldsymbol{x_s}) + \gamma_s\right) + (G + D_g R + \epsilon)^2 \tag{20}$$

for $s = t, \ldots, t+\tau-1$. Moreover,

$$\begin{aligned}
&Q_s(g_s(\boldsymbol{x_s}) + \nabla g_s(\boldsymbol{x_s})^\top(\boldsymbol{x_{s+1}} - \boldsymbol{x_s}) + \gamma_s) \\
&\le Q_s(g_s(\boldsymbol{x_s}) + \nabla g_s(\boldsymbol{x_s})^\top(\hat{\boldsymbol{x}} - \boldsymbol{x_s}) + \gamma_s) \\
&\quad + V_s \nabla f_s(\boldsymbol{x_s})^\top(\hat{\boldsymbol{x}} - \boldsymbol{x_{s+1}}) + \alpha_s\left(D(\hat{\boldsymbol{x}}, \boldsymbol{x_s}) - D(\hat{\boldsymbol{x}}, \boldsymbol{x_{s+1}}) - D(\boldsymbol{x_{s+1}}, \boldsymbol{x_s})\right) \\
&\le Q_s(g_s(\hat{x}) + \gamma_s) + D_f R V_s + \alpha_s\left(D(\hat{\boldsymbol{x}}, \boldsymbol{x_s}) - D(\hat{\boldsymbol{x}}, \boldsymbol{x_{s+1}})\right)
\end{aligned} \tag{21}$$

where the first inequality comes from the inequality given in Lemma A.1 with $\boldsymbol{x}$ set to $\hat{\boldsymbol{x}}$ and the second inequality is because g_s is convex, $V_s \nabla f_s(\boldsymbol{x_s})^\top(\hat{\boldsymbol{x}} - \boldsymbol{x_{s+1}}) \le V_s \|\nabla f_s(\boldsymbol{x_s})\|_* \|\hat{\boldsymbol{x}} - \boldsymbol{x_{s+1}}\| \le V_s D_f R$, and $D(\boldsymbol{x_{s+1}}, \boldsymbol{x_s}) \ge 0$. Combining (20) and (21), we may deduce the following.

$$\begin{aligned}
\mathbb{E}\left[\Delta_s \mid \mathcal{F}_{t-1}\right] &\le \mathbb{E}\left[Q_s(g_s(\hat{x}) + \gamma_s) \mid \mathcal{F}_{t-1}\right] + \alpha_s \mathbb{E}\left[D(\hat{\boldsymbol{x}}, \boldsymbol{x_s}) - D(\hat{\boldsymbol{x}}, \boldsymbol{x_{s+1}}) \mid \mathcal{F}_{t-1}\right] \\
&\quad + D_f R V_s + (G + D_g R + \epsilon)^2 \\
&\le -\frac{\epsilon}{2} \cdot \mathbb{E}\left[Q_s \mid \mathcal{F}_{t-1}\right] + \alpha_s \mathbb{E}\left[D(\hat{\boldsymbol{x}}, \boldsymbol{x_s}) - D(\hat{\boldsymbol{x}}, \boldsymbol{x_{s+1}}) \mid \mathcal{F}_{t-1}\right] + D_f R V_s + (G + D_g R + \epsilon)^2
\end{aligned} \tag{22}$$

where we used Lemma A.2 to obtain the second inequality. Summing (22) over $s = t, \ldots, t+\tau-1$, we get the following.

$$\begin{aligned}
&\sum_{s=t}^{t+\tau-1} \mathbb{E}\left[\Delta_s \mid \mathcal{F}_{t-1}\right] \\
&\le -\frac{\epsilon}{2} \sum_{s=t}^{t+\tau-1} \mathbb{E}\left[Q_s \mid \mathcal{F}_{t-1}\right] + \mathbb{E}\left[\alpha_t D(\hat{\boldsymbol{x}}, \boldsymbol{x_t}) + \sum_{s=t+1}^{t+\tau-1} D(\hat{\boldsymbol{x}}, \boldsymbol{x_s})\left(\alpha_s - \alpha_{s-1}\right) - \alpha_{t+\tau-1} D(\hat{\boldsymbol{x}}, \boldsymbol{x_{t+\tau}}) \mid \mathcal{F}_{t-1}\right] \\
&\quad + D_f R \sum_{s=t}^{t+\tau-1} V_s + (G + D_g R + \epsilon)^2 \tau \\
&\le -\frac{\epsilon}{2} \sum_{s=t}^{t+\tau-1} \mathbb{E}\left[Q_s \mid \mathcal{F}_{t-1}\right] + R^2 \alpha_{t+\tau-1} + D_f R \sum_{s=t}^{t+\tau-1} V_s + (G + D_g R + \epsilon)^2 \tau
\end{aligned} \tag{23}$$

where the second inequality holds because $\{\alpha_t\}_{t=1}^T$ is an increasing sequence and $0 \le D(\boldsymbol{x}, \boldsymbol{y}) \le R^2$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$. Since $\tau \le t+1$, we have $t+\tau-1 \le 2t$. Moreover, as $\alpha_t = t$ and $V_t = \sqrt{t}$,

$$\alpha_{t+\tau-1} \le \alpha_{2t} = 2\alpha_t \quad \text{and} \quad \sum_{s=t}^{t+\tau-1} V_s \le \tau V_{t+\tau-1} \le \tau V_{2t} \le 2\tau V_t. \tag{24}$$

Then we deduce from (23) and (24) the following.

$$\sum_{s=t}^{t+\tau-1} \mathbb{E}\left[\Delta_s \mid \mathcal{F}_{t-1}\right] \le -\frac{\epsilon}{2} \sum_{s=t}^{t+\tau-1} \mathbb{E}\left[Q_s \mid \mathcal{F}_{t-1}\right] + 2R^2 \alpha_t + 2 D_f R \tau V_t + (G + D_g R + \epsilon)^2 \tau. \tag{25}$$

Here, by Lemma A.3,

$$\begin{aligned}
\sum_{s=t}^{t+\tau-1} \mathbb{E}\left[Q_s \mid \mathcal{F}_{t-1}\right] &\ge \sum_{s=t}^{t+\tau-1} \mathbb{E}\left[Q_t - (G + D_g R)(s-t) \mid \mathcal{F}_{t-1}\right] \\
&\ge \tau \mathbb{E}\left[Q_t \mid \mathcal{F}_{t-1}\right] - (G + D_g R)\tau^2 \\
&= \tau Q_t - (G + D_g R)\tau^2
\end{aligned} \tag{26}$$

where the equality holds because Q_t is $\mathcal{F}_{t-1}$-measurable. Then, by (25) and (26),

$$\begin{aligned}\mathbb{E}\left[Q_{t+\tau}^2 \mid \mathcal{F}_{t-1}\right] &= Q_t^2 + 2\sum_{s=t}^{t+\tau-1}\mathbb{E}\left[\Delta_s \mid \mathcal{F}_{t-1}\right] \\ &\leq Q_t^2 - \epsilon\tau Q_t + \epsilon(G+D_gR)\tau^2 + 4R^2\alpha_t + 4D_fR\tau V_t + 2(G+D_gR+\epsilon)^2\tau.\end{aligned} \tag{27}$$

Recall that

$$\begin{aligned}\theta_t(\tau) &= 2(G+D_gR)\tau + \frac{8R^2\alpha_t}{\epsilon\tau} + \frac{8D_fRV_t}{\epsilon} + \frac{4(G+D_gR+\epsilon)^2}{\epsilon} \\ &= \frac{2}{\epsilon\tau}\left(\epsilon(G+D_gR)\tau^2 + 4R^2\alpha_t + 4D_fR\tau V_t + 2(G+D_gR+\epsilon)^2\tau\right).\end{aligned}$$

Therefore, if $Q_t \geq \theta_t(\tau)$, then it follows from (27) that

$$\mathbb{E}\left[Q_{t+\tau}^2 \mid \mathcal{F}_{t-1}\right] \leq Q_t^2 - \frac{\epsilon\tau}{2}Q_t \leq \left(Q_t - \frac{\epsilon\tau}{4}\right)^2,$$

which proves (19), as required. □

A.5 Proof of Lemma 3.5: Time-Varying Bound on the Expected Virtual Queue Length

Let $\delta = (G+\epsilon)$ and $\xi = \epsilon/4$. Moreover, let r and ρ be two constants defined as

$$r = \frac{\xi}{4\lceil\sqrt{T}\rceil\delta^2} \quad \text{and} \quad \rho = 1 - \frac{\xi^2}{8\delta^2}.$$

Since $\xi \leq \delta$, we know that $0 < \rho < 1$.

Lemma A.4. *Let $t \geq 1$. For any $1 \leq \tau \leq \sqrt{T}$ satisfying $t - \tau \geq (2C/\epsilon)^2$ and $t \geq 2\tau - 1$, we have*

$$\mathbb{E}\left[e^{rQ_t}\right] \leq \rho\mathbb{E}\left[e^{rQ_{t-\tau}}\right] + e^{r\delta\tau}e^{r\theta_{t-\tau}}.$$

Proof. As $t \geq 2\tau - 1$, we have $t - \tau \geq \tau - 1$. Henceforth, we use the notation $w_t = Q_t - Q_{t-\tau}$. Note that

$$w_t = \sum_{s=t-\tau}^{t-1} Q_{s+1} - Q_s \leq \tau\delta \tag{28}$$

holds due to Lemma 3.4(b) because $t - \tau \geq (2C/\epsilon)^2$. Furthermore, as $\tau \leq \sqrt{T}$,

$$rw_t \leq r\tau\delta \leq \frac{\xi}{4\delta} \leq 1. \tag{29}$$

Next we observe that $e^x \leq 1 + x + 2x^2$ for any $x \leq 1$. It was already observed in [25, Appendix A] that $e^x \leq 1 + x + 2x^2$ holds for any $|x| \leq 1$. When $x \leq -1$, we know that $e^x \leq 1$ and $x + 2x^2 \geq 0$, which indicates that $e^x \leq 1 + x + 2x^2$. Then

$$e^{rw_t} \leq 1 + rw_t + 2r^2w_t^2 \leq 1 + rw_t + 2r^2\tau^2\delta^2 \leq 1 + rw_t + \frac{1}{2}r\tau\xi$$

where the first inequality is from (29) and the observation that $e^x \leq 1 + x + 2x^2$ holds for any $x \leq 1$ while the second inequality is from (28). This inequality implies that

$$e^{rQ_t} = e^{r(Q_{t-\tau}+w_t)} \leq e^{rQ_{t-\tau}}\left(1 + rw_t + \frac{1}{2}r\tau\xi\right). \tag{30}$$

Let us first consider the case $Q_{t-\tau} > \theta_{t-\tau}(\tau)$. For ease of notation, we use $\theta_{t-\tau}$ to denote $\theta_{t-\tau}(\tau)$. As $t-\tau \geq \tau - 1$ and $t - \tau \geq (2C/\epsilon)^2$, it follows from Lemma 3.4(c) that $\mathbb{E}\left[w_t \mid \mathcal{F}_{t-\tau-1}\right] \leq -\xi\tau$. In this case, we obtain the following based on (30).

$$\begin{aligned}\mathbb{E}\left[e^{rQ_t} \mid Q_{t-\tau} > \theta_{t-\tau}\right] &\leq \mathbb{E}\left[e^{rQ_{t-\tau}}\left(1 + rw_t + \frac{1}{2}r\xi\tau\right) \mid Q_{t-\tau} > \theta_{t-\tau}\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[e^{rQ_{t-\tau}}\left(1 + rw_t + \frac{1}{2}r\xi\tau\right) \mid \mathcal{F}_{t-\tau-1}\right] \mid Q_{t-\tau} > \theta_{t-\tau}\right]\\ &\leq \mathbb{E}\left[e^{rQ_{t-\tau}}\left(1 - r\xi\tau + \frac{1}{2}r\xi\tau\right) \mid Q_{t-\tau} > \theta_{t-\tau}\right]\\ &= \mathbb{E}\left[\rho e^{rQ_{t-\tau}} \mid Q_{t-\tau} > \theta_{t-\tau}\right]\end{aligned} \tag{31}$$

where the first inequality is from (30), the first equality is from the tower rule, and the second inequality holds due to $\mathbb{E}\left[w_t \mid \mathcal{F}_{t-\tau-1}\right] \leq -\xi\tau$. If $Q_{t-\tau} \leq \theta_{t-\tau}$, we may deduce the following based on (28).

$$\begin{aligned}\mathbb{E}\left[e^{rQ_t} \mid Q_{t-\tau} \leq \theta_{t-\tau}\right] &= \mathbb{E}\left[e^{rw_t}e^{rQ_{t-\tau}} \mid Q_{t-\tau} \leq \theta_{t-\tau}\right]\\ &\leq \mathbb{E}\left[e^{r\delta\tau}e^{rQ_{t-\tau}} \mid Q_{t-\tau} \leq \theta_{t-\tau}\right].\end{aligned} \tag{32}$$

Note that

$$\begin{aligned}\mathbb{E}\left[e^{rQ_t}\right] &= \mathbb{P}\left[Q_{t-\tau} > \theta_{t-\tau}\right] \cdot \mathbb{E}\left[e^{rQ_t} \mid Q_{t-\tau} > \theta_{t-\tau}\right] + \mathbb{P}\left[Q_{t-\tau} \leq \theta_{t-\tau}\right] \cdot \mathbb{E}\left[e^{rQ_t} \mid Q_{t-\tau} \leq \theta_{t-\tau}\right]\\ &\leq \rho\mathbb{E}\left[e^{rQ_{t-\tau}} \mid Q_{t-\tau} > \theta_{t-\tau}\right] \cdot \mathbb{P}\left[Q_{t-\tau} > \theta_{t-\tau}\right] + e^{r\delta\tau}\mathbb{E}\left[e^{rQ_{t-\tau}} \mid Q_{t-\tau} \leq \theta_{t-\tau}\right] \cdot \mathbb{P}\left[Q_{t-\tau} \leq \theta_{t-\tau}\right]\\ &= \rho\mathbb{E}\left[e^{rQ_{t-\tau}}\right] + \left(e^{r\delta\tau} - \rho\right)\mathbb{E}\left[e^{rQ_{t-\tau}} \mid Q_{t-\tau} \leq \theta_{t-\tau}\right] \cdot \mathbb{P}\left[Q_{t-\tau} \leq \theta_{t-\tau}\right]\\ &\leq \rho\mathbb{E}\left[e^{rQ_{t-\tau}}\right] + e^{r\delta\tau}e^{r\theta_{t-\tau}}\end{aligned} \tag{33}$$

where the first inequality is deduced by (31) and (32) and the second inequality holds because $e^{r\delta\tau} - \rho \leq e^{r\delta\tau}$ and $\mathbb{P}\left[Q_{t-\tau} \leq \theta_{t-\tau}\right] \leq 1$. □

Proof of Lemma 3.5. Let $t \geq 1$, and let $\tau = \lceil\sqrt{t}\rceil$. For any s, we use θ_s to denote $\theta_s(\tau)$. We first consider the case where

$$t \geq \max\left\{9, \left(\frac{2C}{\epsilon}\right)^4\right\}.$$

Then it follows that $t \geq 2\tau$ and $\tau \geq (2C/\epsilon)^2$. Moreover, $\lfloor t/\tau \rfloor = k$ for some $k \geq 2$, in which case

$$t - (k-1)\tau \geq \tau \geq (2C/\epsilon)^2.$$

Then, we may apply Lemma A.4 for $s = t, t-\tau, \ldots, t-(k-2)\tau$. In particular, we obtain

$$\begin{aligned}\mathbb{E}\left[e^{rQ_t}\right] &\leq \rho\mathbb{E}\left[e^{rQ_{t-\tau}}\right] + e^{r\delta\tau}e^{r\theta_{t-\tau}}\\ &\leq \rho^{k-1}\mathbb{E}\left[e^{rQ_{t-(k-1)\tau}}\right] + e^{r\delta\tau}\sum_{i=1}^{k-1}\rho^{i-1}e^{r\theta_{t-i\tau}}.\end{aligned} \tag{34}$$

As θ_t increases as t increases, $\theta_{t-i\tau} \leq \theta_t$ for any $i \geq 1$. In addition, $r\delta\tau \leq 2r\delta\tau$ and $e^{rC(2C/\epsilon)^2} \geq 1$. Hence, we deduce from (34) that

$$\mathbb{E}\left[e^{rQ_t}\right] \leq \rho^{k-1}\mathbb{E}\left[e^{rQ_{t-(k-1)\tau}}\right] + e^{rC(2C/\epsilon)^2}e^{2r\delta\tau}e^{r\theta_t}\sum_{i=0}^{k-2}\rho^i. \tag{35}$$

Moreover, note that $t-(k-1)\tau < 2\tau$. As $Q_{t-(k-1)\tau} \leq C(2C/\epsilon)^2 + 2\delta\tau$ by Lemma 3.4(a),

$$e^{rQ_{t-(k-1)\tau}} \leq e^{rC(2C/\epsilon)^2} e^{2r\delta\tau} \leq e^{rC(2C/\epsilon)^2} e^{2r\delta\tau} \left(\frac{1}{1-\rho} e^{r\theta_t}\right) \tag{36}$$

where the second inequality holds because $0 < \rho < 1$ and $e^{r\theta_t} \geq 1$. Combining (35) and (36),

$$\mathbb{E}\left[e^{rQ_t}\right] \leq e^{rC(2C/\epsilon)^2} e^{2r\delta\tau} e^{r\theta_t} \left(\frac{\rho^{k-1}}{1-\rho} + \sum_{i=0}^{k-2} \rho^i\right) = \frac{1}{1-\rho} e^{rC(2C/\epsilon)^2} e^{2r\delta\tau} e^{r\theta_t}. \tag{37}$$

As $p(x) = e^{rx}$ is convex over $x \in \mathbb{R}$, Jensen's inequality implies that

$$e^{r\mathbb{E}[Q_t]} \leq \mathbb{E}\left[e^{rQ_t}\right].$$

Then, by (37), we obtain

$$\begin{aligned}
E\left[Q_t\right] &\leq \theta_t + 2\delta\tau + C\left(\frac{2C}{\epsilon}\right)^2 + \log\frac{1}{1-\rho} \\
&= \theta_t + 2(G+\epsilon)\tau + \log\frac{128(G+\epsilon)^2}{\epsilon^2} + C\left(\frac{2C}{\epsilon}\right)^2 \\
&\leq \theta_t + 4(G+\epsilon)\sqrt{t} + \log\frac{128(G+\epsilon)^2}{\epsilon^2} + C\left(\frac{2C}{\epsilon}\right)^2 \\
&\leq \theta_t + 4(G+\epsilon)\sqrt{t} + \log\frac{128(G+\epsilon)^2}{\epsilon^2} + C\left(\frac{2C}{\epsilon}\right)^2 + (G+\epsilon)\left(9 + \left(\frac{2C}{\epsilon}\right)^4\right)
\end{aligned}$$

where the second inequality holds because $\tau \leq 2\sqrt{t}$.

Now consider the case $t < \max\left\{9, (2C/\epsilon)^4\right\}$. Then $t < 9 + (2C/\epsilon)^4$. Note that

$$\begin{aligned}
Q_t &\leq C\left(\frac{2C}{\epsilon}\right)^2 + (G+\epsilon)t \\
&\leq \theta_t + 4(G+\epsilon)\sqrt{t} + \log\frac{128(G+\epsilon)^2}{\epsilon^2} + C\left(\frac{2C}{\epsilon}\right)^2 + (G+\epsilon)\left(9 + \left(\frac{2C}{\epsilon}\right)^4\right)
\end{aligned}$$

where the first inequality comes from Lemma 3.4(a) and the second inequality holds because $t \leq 9 + (2C/\epsilon)^4$. □

A.6 Completing the Proofs of Theorem 3.6 (Constraint Violation) and Theorem 3.7 (Regret)

Plugging in the formula of θ_t to Lemma 3.5, it follows that for any $t \geq 1$,

$$\begin{aligned}
\mathbb{E}\left[Q_t\right] \leq\ & 4\left(2G + D_g R + \epsilon + \frac{2R^2 + 2D_f R}{\epsilon}\right)\sqrt{t} + \frac{4(G + D_g R + \epsilon)^2}{\epsilon} \\
&+ \log\frac{128(G+\epsilon)^2}{\epsilon^2} + C\left(\frac{2C}{\epsilon}\right)^2 + (G+\epsilon)\left(9 + \left(\frac{2C}{\epsilon}\right)^4\right).
\end{aligned} \tag{38}$$

It follows from basic calculus that

$$\log(T+1) \leq \sum_{t=1}^{T} \frac{1}{t} \leq 1 + \log T.$$

Proof of Theorem 3.6. Note that $\sqrt{T+1} \leq 2\sqrt{T}$ for any $T \geq 1$. Based on Lemma 3.3 and the expected queue size bound (38), we deduce the following.

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} g_t(\boldsymbol{x_t})\right] \\
&\leq \mathbb{E}\left[Q_{T+1}\right] + \sum_{t=1}^{T} \frac{D_f D_g V_t + D_g^2 \mathbb{E}\left[Q_t\right]}{2\alpha_t} - \sum_{t=1}^{T} \gamma_t \\
&\leq \left(D_f D_g + (4D_g^2 + 8)\left(2G + D_g R + \epsilon + \frac{2R^2 + 2D_f R}{\epsilon}\right)\right)\sqrt{T} - C\sqrt{T} \\
&\quad + \left(\frac{4(G + D_g R + \epsilon)^2}{\epsilon} + \log\frac{128(G+\epsilon)^2}{\epsilon^2} + C\left(\frac{2C}{\epsilon}\right)^2 + (G+\epsilon)\left(9 + \left(\frac{2C}{\epsilon}\right)^4\right)\right)\left(1 + \frac{D_g^2(1+\log T)}{2}\right).
\end{aligned}
$$

Since we set C as

$$
C = D_f D_g + (4D_g^2 + 8)\left(2G + D_g R + \epsilon + \frac{2R^2 + 2D_f R}{\epsilon}\right) + 1,
$$

we have

$$
\mathbb{E}\left[\sum_{t=1}^{T} g_t(\boldsymbol{x_t})\right] \leq -\sqrt{T} + \left(\frac{4(G + D_g R + \epsilon)^2}{\epsilon} + \log\frac{128(G+\epsilon)^2}{\epsilon^2} + C\left(\frac{2C}{\epsilon}\right)^2\right)\left(1 + \frac{D_g^2(1+\log T)}{2}\right).
$$

Therefore, there exists some positive constant T_1 such that for any $T \geq T_1$,

$$
\mathbb{E}\left[\sum_{t=1}^{T} g_t(\boldsymbol{x_t})\right] \leq 0,
$$

as required. □

Proof of Theorem 3.7. If $C = O(1)$, then $\mathbb{E}[Q_t] \leq B\sqrt{t}$ for some constant B that depends only on G, D_g, D_f, R, ϵ by (38). By Lemma 3.2, it follows that

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} f_t(\boldsymbol{x_t}) - \sum_{t=1}^{T} f_t(\boldsymbol{x}^*)\right] &\leq \left(\frac{C(G + D_g R + C)}{\epsilon}\right)^2 + K_1\sqrt{T} + C\sum_{t=1}^{T}\frac{\mathbb{E}[Q_t]}{t} \\
&\leq \left(\frac{C(G + D_g R + C)}{\epsilon}\right)^2 + K_1\sqrt{T} + CB\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \\
&= O(\sqrt{T}),
\end{aligned}
$$

as required. □

B Performance Analysis of Bandit Drift-Plus-Penalty (Algorithm 2)

B.1 Bounds on the Gradient Estimates

Proof of Lemma 4.1. By the definition of $\tilde{\nabla} g_t$, it follows that

$$\begin{aligned}\|\tilde{\nabla} g_t\|_* &= \frac{d}{2\delta_t}\left|g_t(\boldsymbol{x_t}+\delta_t\boldsymbol{u_t}) - g_t(\boldsymbol{x_t}-\delta_t\boldsymbol{u_t})\right| \cdot \|\boldsymbol{u_t}\|_* && (39)\\ &\leq \frac{d}{2\delta_t} L_g \|2\delta_t \boldsymbol{u_t}\|_2 \|\boldsymbol{u_t}\|_* && (40)\\ &= dL_g \|\boldsymbol{u_t}\|_* && (41)\end{aligned}$$

where the inequality comes from Assumption 3. Since $\boldsymbol{u_t}$ is a point in $\left\{\boldsymbol{u} \in \mathbb{R}^d : \ \|\boldsymbol{u}\|_2 \leq 1\right\}$. □

Following [20], let us define functions $\hat{f}_t$ and $\hat{g}_t$ as follows. For $t \in [T]$ and $\boldsymbol{x} \in \mathcal{X}$,

$$\hat{f}_t(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{u_t}}\left[f_t(\boldsymbol{x}+\delta_t\boldsymbol{u_t})\right] \quad \text{and} \quad \hat{g}_t(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{u_t}}\left[g_t(\boldsymbol{x}+\delta_t\boldsymbol{u_t})\right]$$

where $\boldsymbol{u_t}$ is sampled from the Euclidean unit sphere $\left\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_2 \leq 1\right\}$ uniformly at random. We will use the following two lemmas shown in [20].

Lemma B.1. *[20, Lemma 8] For any $t \geq 1$, $\hat{f}_t$ and $\hat{g}_t$ are convex and satisfy*

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \left|\hat{f}_t(\boldsymbol{x}) - f_t(\boldsymbol{x})\right| \leq L_f \delta_t \quad \textit{and} \quad \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\hat{g}_t(\boldsymbol{x}) - g_t(\boldsymbol{x})\right| \leq L_g \delta_t.$$

Moreover, $\hat{f}_t$ and $\hat{g}_t$ are differentiable, and

$$\nabla \hat{f}_t(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{u_t}}\left[\frac{d}{\delta_t} f_t(\boldsymbol{x}+\delta_t\boldsymbol{u_t})\boldsymbol{u_t}\right] \quad \textit{and} \quad \nabla \hat{g}_t(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{u_t}}\left[\frac{d}{\delta_t} g_t(\boldsymbol{x}+\delta_t\boldsymbol{u_t})\boldsymbol{u_t}\right].$$

Lemma B.2. *[20, Lemma 9] For any $t \geq 1$,*

$$\mathbb{E}\left[\tilde{\nabla} f_t \mid \mathcal{H}_t^0\right] = \nabla \hat{f}_t(\boldsymbol{x_t}) \ \textit{and} \ \mathbb{E}\left[\tilde{\nabla} g_t \mid \mathcal{H}_t^0\right] = \nabla \hat{g}_t(\boldsymbol{x_t}).$$

B.2 Proof of Lemma 4.3: Upper Bound on the Expected Regret

The following lemma is analogous to Lemma 3.1.

Lemma B.3. *For $t \geq 1$,*

$$\begin{aligned}\Delta_t \leq Q_t &\left(g_t(\boldsymbol{x_t}) + \tilde{\nabla} g_t^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t\right)\\ &+ (G+\gamma_t)^2 + R^2 \|\tilde{\nabla} g_t\|_*^2.\end{aligned}$$

Proof. As $Q_{t+1} = \max\left\{Q_t + g_t(\boldsymbol{x_t}) + \tilde{\nabla} g_t^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t, 0\right\}$,

$$Q_{t+1}^2 \leq \left(Q_t + g_t(\boldsymbol{x_t}) + \tilde{\nabla} g_t^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t\right)^2.$$

Expanding the right-hand side, we obtain

$$\Delta_t = \frac{Q_{t+1}^2}{2} - \frac{Q_t^2}{2} \leq Q_t(g_t(\boldsymbol{x_t}) + \tilde{\nabla} g_t^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t) + \frac{1}{2}\left(g_t(\boldsymbol{x_t}) + \tilde{\nabla} g_t^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t\right)^2.$$

Here,

$$\begin{aligned}\left(g_t(\boldsymbol{x_t}) + \tilde{\nabla} g_t^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t\right)^2 &\leq \left(|g_t(\boldsymbol{x_t})| + \|\tilde{\nabla} g_t\|_* \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\| + \gamma_t\right)^2 \\ &\leq 2\left(|g_t(\boldsymbol{x_t})| + \gamma_t\right)^2 + 2\|\tilde{\nabla} g_t\|_*^2 \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|^2 \\ &\leq 2(G+\gamma_t)^2 + 2R^2\|\tilde{\nabla} g_t\|_*^2\end{aligned}$$

where the third inequality is due to Assumption 1. Therefore,

$$\Delta_t \leq Q_t(g_t(\boldsymbol{x_t}) + \tilde{\nabla} g_t^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t) + (G+\gamma_t)^2 + R^2\|\tilde{\nabla} g_t\|_*^2,$$

as required. □

The following lemma is a modification of Lemma A.1 which can be derived based on the fact that

$$\left(V_t \tilde{\nabla} f_t + Q_t \tilde{\nabla} g_t\right)^\top (\boldsymbol{x} - \boldsymbol{x_t}) + \alpha_t D(\boldsymbol{x}, \boldsymbol{x_t})$$

is $2\alpha_t$-strongly convex with respect to the norm $\|\cdot\|$ and that $\boldsymbol{x_{t+1}}$ is its minimizer over $\mathcal{X}$.

Lemma B.4. *For any $\boldsymbol{x} \in \mathcal{X}$ and $t \geq 1$,*

$$\begin{aligned}&\left(V_t \tilde{\nabla} f_t + Q_t \tilde{\nabla} g_t\right)^\top (\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \alpha_t D(\boldsymbol{x_{t+1}}, \boldsymbol{x_t}) \\ &\leq \left(V_t \tilde{\nabla} f_t + Q_t \tilde{\nabla} g_t\right)^\top (\boldsymbol{x} - \boldsymbol{x_t}) + \alpha_t D(\boldsymbol{x}, \boldsymbol{x_t}) - \alpha_t D(\boldsymbol{x}, \boldsymbol{x_{t+1}})\end{aligned}$$

Proof of Lemma 4.3. We will show that

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}\left[f_t(\boldsymbol{x_t}) - f_t(\boldsymbol{x})\right] \leq \sum_{t=1}^T 2L_f\delta_t + \sum_{t=1}^T \frac{2L_g\delta_t + \gamma_t}{V_t}\mathbb{E}\left[Q_t\right] + \frac{\alpha_T}{V_T}R^2 \\ + \sum_{t=1}^T \frac{V_t}{4\alpha_t} qdp_*^2 L_f^2 + \frac{C^2(G+C)^2}{\epsilon^2} + \sum_{T=1}^T \frac{(G+\epsilon)^2 + R^2 qdp_*^2 L_g^2}{V_t}\end{aligned}$$

holds.

First, let us add $V_t \hat{f}_t(\boldsymbol{x_t}) + Q_t \hat{g}_t(\boldsymbol{x_t})$ to both sides of the inequality given in Lemma B.4. Then we obtain

$$\begin{aligned}&V_t \hat{f}_t(\boldsymbol{x_t}) + Q_t \hat{g}_t(\boldsymbol{x_t}) + \left(V_t \tilde{\nabla} f_t + Q_t \tilde{\nabla} g_t\right)^\top (\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \alpha_t D(\boldsymbol{x_{t+1}}, \boldsymbol{x_t}) \\ &\leq V_t \hat{f}_t(\boldsymbol{x_t}) + Q_t \hat{g}_t(\boldsymbol{x_t}) + \left(V_t \tilde{\nabla} f_t + Q_t \tilde{\nabla} g_t\right)^\top (\boldsymbol{x} - \boldsymbol{x_t}) + \alpha_t D(\boldsymbol{x}, \boldsymbol{x_t}) - \alpha_t D(\boldsymbol{x}, \boldsymbol{x_{t+1}}).\end{aligned} \tag{42}$$

Based on Lemma 3.1, we may observe that the left-hand side of (42) is greater than or equal to

$$V_t \hat{f}_t(\boldsymbol{x_t}) + V_t \tilde{\nabla} f_t^\top (\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \alpha_t D(\boldsymbol{x_{t+1}}, \boldsymbol{x_t}) + \Delta_t - Q_t\gamma_t - Q_t(g_t(\boldsymbol{x_t}) - \hat{g}_t(\boldsymbol{x_t})) - (G+\gamma_t)^2 - R^2\|\tilde{\nabla} g_t\|_*^2. \tag{43}$$

Moreover, we can find a lower bound on the term $V_t \tilde{\nabla} f_t^\top (\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \alpha_t D(\boldsymbol{x_{t+1}}, \boldsymbol{x_t})$ as follows.

$$\begin{aligned}V_t \tilde{\nabla} f_t^\top (\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \alpha_t D(\boldsymbol{x_{t+1}}, \boldsymbol{x_t}) &\geq -V_t\|\tilde{\nabla} f_t\|_* \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\| + \alpha_t\|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|^2 \\ &\geq -\frac{V_t^2}{4\alpha_t}\|\tilde{\nabla} f_t\|_*^2 - \alpha_t\|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|^2 + \alpha_t\|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|^2 \\ &= -\frac{V_t^2}{4\alpha_t}\|\tilde{\nabla} f_t\|_*^2\end{aligned} \tag{44}$$

where the first inequality holds due to $\boldsymbol{u}^\top \boldsymbol{v} \leq \|\boldsymbol{u}\|_* \|\boldsymbol{v}\|$ and (3) while the second inequality is from $pq \leq p^2/(4\alpha_t) + \alpha_t q^2$. Combining (42), (43), and (44), it follows that

$$\begin{aligned} V_t \hat{f}_t(\boldsymbol{x_t}) \leq V_t \left(\hat{f}_t(\boldsymbol{x_t}) + \tilde{\nabla} f_t^\top (\boldsymbol{x} - \boldsymbol{x_t}) \right) + Q_t \left(\hat{g}_t(\boldsymbol{x_t}) + \tilde{\nabla} g_t^\top (\boldsymbol{x} - \boldsymbol{x_t}) \right) + \alpha_t D(\boldsymbol{x}, \boldsymbol{x_t}) - \alpha_t D(\boldsymbol{x}, \boldsymbol{x_{t+1}}) \\ + \frac{V_t^2}{4\alpha_t} \|\tilde{\nabla} f_t\|_*^2 - \alpha_t D(\boldsymbol{x_{t+1}}, \boldsymbol{x_t}) - \Delta_t + Q_t \gamma_t + Q_t(g_t(\boldsymbol{x_t}) - \hat{g}_t(\boldsymbol{x_t})) + (G + \gamma_t)^2 + R^2 \|\tilde{\nabla} g_t\|_*^2. \end{aligned} \tag{45}$$

Note that

$$Q_t(g_t(\boldsymbol{x_t}) - \hat{g}_t(\boldsymbol{x_t})) \leq Q_t |g_t(\boldsymbol{x_t}) - \hat{g}_t(\boldsymbol{x_t})| \leq Q_t L_g \delta_t \tag{46}$$

by Lemma B.1. Next, we consider the terms $V_t \left(\hat{f}_t(\boldsymbol{x_t}) + \tilde{\nabla} f_t^\top (\boldsymbol{x} - \boldsymbol{x_t}) \right)$ and $Q_t \left(\hat{g}_t(\boldsymbol{x_t}) + \tilde{\nabla} g_t^\top (\boldsymbol{x} - \boldsymbol{x_t}) \right)$. Note that

$$\begin{aligned} \mathbb{E}\left[V_t \left(\hat{f}_t(\boldsymbol{x_t}) + \tilde{\nabla} f_t^\top (\boldsymbol{x} - \boldsymbol{x_t}) \right) \right] &= V_t \cdot \mathbb{E}\left[\hat{f}_t(\boldsymbol{x_t}) + \mathbb{E}\left[\tilde{\nabla} f_t^\top (\boldsymbol{x} - \boldsymbol{x_t}) \mid \mathcal{H}_t^0 \right] \right] \\ &= V_t \cdot \mathbb{E}\left[\hat{f}_t(\boldsymbol{x_t}) + (\boldsymbol{x} - \boldsymbol{x_t})^\top \mathbb{E}\left[\tilde{\nabla} f_t \mid \mathcal{H}_t^0 \right] \right] \\ &= V_t \cdot \mathbb{E}\left[\hat{f}_t(\boldsymbol{x_t}) + (\boldsymbol{x} - \boldsymbol{x_t})^\top \nabla \hat{f}_t(\boldsymbol{x_t}) \right] \\ &\leq V_t \cdot \mathbb{E}\left[\hat{f}_t(\boldsymbol{x}) \right] \end{aligned} \tag{47}$$

where the first equality is from the linearity of expectation and the tower rule, the second equality holds because $\boldsymbol{x} - \boldsymbol{x_t}$ is $\mathcal{H}_t^0$-measurable, the third equality is by Lemma B.2, and the inequality is due to the convexity of $\hat{f}_t$ (Lemma B.1). Moreover,

$$\begin{aligned} \mathbb{E}\left[Q_t \left(\hat{g}_t(\boldsymbol{x_t}) + \tilde{\nabla} g_t^\top (\boldsymbol{x} - \boldsymbol{x_t}) \right) \right] &= \mathbb{E}\left[Q_t \hat{g}_t(\boldsymbol{x_t}) + \mathbb{E}\left[Q_t \tilde{\nabla} g_t^\top (\boldsymbol{x} - \boldsymbol{x_t}) \mid \mathcal{H}_t^0 \right] \right] \\ &= \mathbb{E}\left[Q_t \hat{g}_t(\boldsymbol{x_t}) + Q_t (\boldsymbol{x} - \boldsymbol{x_t})^\top \mathbb{E}\left[\tilde{\nabla} g_t \mid \mathcal{H}_t^0 \right] \right] \\ &= \mathbb{E}\left[Q_t \hat{g}_t(\boldsymbol{x_t}) + Q_t (\boldsymbol{x} - \boldsymbol{x_t})^\top \nabla \hat{g}_t(\boldsymbol{x_t}) \right] \\ &\leq \mathbb{E}\left[Q_t \hat{g}_t(\boldsymbol{x}) \right] \end{aligned} \tag{48}$$

where the first equality is from the linearity of expectation and the tower rule, the second equality holds because $Q_t(\boldsymbol{x} - \boldsymbol{x_t})$ is $\mathcal{H}_t^0$-measurable, the third equality is by Lemma B.2, and the inequality is due to the convexity of $\hat{g}_t$ (Lemma B.1). Furthermore, the terms involving $\|\tilde{\nabla} f_t\|_*^2$ and $\|\tilde{\nabla} g_t\|_*^2$ on the right-hand side of (45) can be dealt with based on Lemma 4.2. Taking the expectation of both sides of (45), the resulting inequality implies the following. Using the bounds (46), (47), and (48),

$$\begin{aligned} V_t \mathbb{E}\left[\hat{f}_t(\boldsymbol{x_t}) \right] \leq V_t \mathbb{E}\left[\hat{f}_t(\boldsymbol{x}) \right] + \mathbb{E}\left[Q_t \hat{g}_t(\boldsymbol{x}) \right] + \mathbb{E}\left[\alpha_t D(\boldsymbol{x}, \boldsymbol{x_t}) - \alpha_t D(\boldsymbol{x}, \boldsymbol{x_{t+1}}) \right] \\ + \frac{V_t^2}{4\alpha_t} q d p_*^2 L_f^2 - \mathbb{E}\left[\Delta_t \right] + \gamma_t \mathbb{E}\left[Q_t \right] + Q_t L_g \delta_t + (G + \gamma_t)^2 + R^2 q d p_*^2 L_g^2. \end{aligned} \tag{49}$$

Next, by Lemma B.1, we know that $|\hat{f}_t(\boldsymbol{x}) - f_t(\boldsymbol{x})| \leq L_f \delta_t$ and $|\hat{g}_t(\boldsymbol{x}) - g_t(\boldsymbol{x})| \leq L_g \delta_t$, applying which to (49), we obtain

$$\begin{aligned} V_t \mathbb{E}\left[f_t(\boldsymbol{x_t}) - f_t(\boldsymbol{x}) \right] \leq 2 L_f \delta_t V_t + 2 L_g \delta_t Q_t + \mathbb{E}\left[Q_t g_t(\boldsymbol{x}) \right] + \mathbb{E}\left[\alpha_t D(\boldsymbol{x}, \boldsymbol{x_t}) - \alpha_t D(\boldsymbol{x}, \boldsymbol{x_{t+1}}) \right] \\ + \frac{V_t^2}{4\alpha_t} q d p_*^2 L_f^2 - \mathbb{E}\left[\Delta_t \right] + \gamma_t \mathbb{E}\left[Q_t \right] + (G + \gamma_t)^2 + R^2 q d p_*^2 L_g^2. \end{aligned} \tag{50}$$

Dividing both sides of (50) and summing the resulting inequality for $t = 1, \ldots, T$, we obtain the following inequality.

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[f_t(\boldsymbol{x_t}) - f_t(\boldsymbol{x})\right] \\
&\leq \sum_{t=1}^{T} 2L_f\delta_t + \sum_{t=1}^{T} \frac{2L_g\delta_t}{V_t}\mathbb{E}\left[Q_t\right] + \sum_{t=1}^{T} \frac{1}{V_t}\mathbb{E}\left[Q_t g_t(\boldsymbol{x})\right] + \sum_{t=1}^{T} \frac{\alpha_t}{V_t}\mathbb{E}\left[D(\boldsymbol{x}, \boldsymbol{x_t}) - D(\boldsymbol{x}, \boldsymbol{x_{t+1}})\right] \qquad (51) \\
&\quad + \sum_{t=1}^{T} \frac{V_t}{4\alpha_t} qdp_*^2 L_f^2 - \sum_{t=1}^{T} \frac{1}{V_t}\mathbb{E}\left[\Delta_t\right] + \sum_{t=1}^{T} \frac{\gamma_t}{V_t}\mathbb{E}\left[Q_t\right] + \sum_{T=1}^{T} \frac{(G+\gamma_t)^2}{V_t} + \sum_{T=1}^{T} \frac{R^2 qdp_*^2 L_g^2}{V_t}.
\end{aligned}$$

As in the proof of Lemma 3.2, we can argue the following inequalities bounding some terms on the right-hand side of (51).

$$\mathbb{E}\left[Q_t g_t(\boldsymbol{x^*})\right] = \mathbb{E}\left[Q_t \bar{g}(\boldsymbol{x^*})\right] \leq 0, \qquad (52)$$

$$\sum_{t=1}^{T} \frac{\alpha_t}{V_t} E\left[D(\boldsymbol{x}, \boldsymbol{x_t}) - D(\boldsymbol{x}, \boldsymbol{x_{t+1}})\right] \leq \mathbb{E}\left[\frac{\alpha_1}{V_1}R^2 + \sum_{t=2}^{T} R^2 \left(\frac{\alpha_t}{V_t} - \frac{\alpha_{t-1}}{V_{t-1}}\right)\right] = \frac{\alpha_T}{V_T}R^2, \qquad (53)$$

$$\sum_{t=1}^{T} \frac{1}{V_t}\mathbb{E}\left[\Delta_t\right] = \mathbb{E}\left[-\frac{1}{2V_1}Q_1^2 + \frac{1}{2V_T}Q_{T+1}^2 + \frac{1}{2}\sum_{t=2}^{T} Q_t^2 \left(\frac{1}{V_{t-1}} - \frac{1}{V_t}\right)\right] \geq -\frac{1}{2V_1}Q_1^2 = 0, \qquad (54)$$

$$\sum_{T=1}^{T} \frac{(G+\gamma_t)^2}{V_t} \leq (G+C)^2 (C/\epsilon)^2 + \sum_{t=1}^{T} \frac{(G+\epsilon)^2}{V_t}. \qquad (55)$$

Applying the bounds (52), (53), (54), and (55) to (51) with $\boldsymbol{x} = \boldsymbol{x^*}$, it follows that

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\left[f_t(\boldsymbol{x_t}) - f_t(\boldsymbol{x})\right] &\leq \sum_{t=1}^{T} 2L_f\delta_t + \sum_{t=1}^{T} \frac{2L_g\delta_t + \gamma_t}{V_t}\mathbb{E}\left[Q_t\right] + \frac{\alpha_T}{V_T}R^2 \\
&\quad + \sum_{t=1}^{T} \frac{V_t}{4\alpha_t} qdp_*^2 L_f^2 + \frac{C^2(G+C)^2}{\epsilon^2} + \sum_{T=1}^{T} \frac{(G+\epsilon)^2 + R^2 qdp_*^2 L_g^2}{V_t},
\end{aligned}$$

as required. □

B.3 Proof of Lemma 4.4: Upper Bound on the Constraint Violation

We will show that

$$\sum_{t=1}^{T} g_t(\boldsymbol{x_t}) \leq Q_{T+1} + \sum_{t=1}^{T} \frac{V_t}{4\alpha_t}\left(\|\tilde{\nabla} g_t\|_*^2 + \|\tilde{\nabla} f_t\|_*^2\right) + \sum_{t=1}^{T} \frac{Q_t}{2\alpha_t}\|\tilde{\nabla} g_t\|_*^2 - \sum_{t=1}^{T} \gamma_t$$

holds.

Since $Q_{t+1} \geq Q_t + g_t(\boldsymbol{x_t}) + \tilde{\nabla} g_t^\top (\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t$, we have

$$\begin{aligned}
g_t(\boldsymbol{x_t}) &\leq Q_{t+1} - Q_t - \tilde{\nabla} g_t^\top (\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) - \gamma_t \\
&\leq Q_{t+1} - Q_t + \|\tilde{\nabla} g_t\|_* \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\| - \gamma_t
\end{aligned} \qquad (56)$$

where the second inequality is due to the fact that $\boldsymbol{u}^\top \boldsymbol{v} \leq \|\boldsymbol{u}\|_* \|\boldsymbol{v}\|$ for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$. To provide an upper bound on the term $\|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|$, we use the inequality in Lemma B.4 with $\boldsymbol{x} = \boldsymbol{x_t}$, which implies that

$$\alpha_t D(\boldsymbol{x_{t+1}}, \boldsymbol{x_t}) + \alpha_t D(\boldsymbol{x_t}, \boldsymbol{x_{t+1}}) \leq \left(V_t \tilde{\nabla} f_t + Q_t \tilde{\nabla} g_t\right)^\top (\boldsymbol{x_t} - \boldsymbol{x_{t+1}}). \tag{57}$$

The left-hand side of (57) is greater than or equal to $2\alpha_t \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|^2$ because $D(\boldsymbol{x_{t+1}}, \boldsymbol{x_t}) \geq \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|^2$ and $D(\boldsymbol{x_t}, \boldsymbol{x_{t+1}}) \geq \|\boldsymbol{x_t} - \boldsymbol{x_{t+1}}\|^2$ by (3). The right-hand side can be bounded by

$$\left\|V_t \tilde{\nabla} f_t + Q_t \tilde{\nabla} g_t\right\|_* \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\| \leq \left(V_t \left\|\tilde{\nabla} f_t\right\|_* + Q_t \left\|\tilde{\nabla} g_t\right\|_*\right) \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|.$$

Then we obtain from (57) that

$$2\alpha_t \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|^2 \leq \left(V_t \left\|\tilde{\nabla} f_t\right\|_* + Q_t \left\|\tilde{\nabla} g_t\right\|_*\right) \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\|,$$

implying in turn that

$$\|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\| \leq \frac{1}{2\alpha_t} \left(V_t \left\|\tilde{\nabla} f_t\right\|_* + Q_t \left\|\tilde{\nabla} g_t\right\|_*\right). \tag{58}$$

Then

$$\begin{aligned} g_t(\boldsymbol{x_t}) &\leq Q_{t+1} - Q_t + \frac{1}{2\alpha_t} \left(V_t \|\tilde{\nabla} f_t\|_* \|\tilde{\nabla} g_t\|_* + Q_t \|\tilde{\nabla} g_t\|_*^2\right) - \gamma_t \\ &\leq Q_{t+1} - Q_t + \frac{1}{4\alpha_t} \left(V_t \left(\|\tilde{\nabla} f_t\|_*^2 + \|\tilde{\nabla} g_t\|_*^2\right) + 2Q_t \|\tilde{\nabla} g_t\|_*^2\right) - \gamma_t \end{aligned} \tag{59}$$

where the first inequality is obtained by (56) and (58) and the second inequality holds because $2pq \leq p^2 + q^2$ for any p, q. Summing (59) over $t = 1, \ldots, T$, we obtain

$$\sum_{t=1}^T g_t(\boldsymbol{x_t}) \leq Q_{T+1} - Q_1 + \sum_{t=1}^T \frac{V_t}{4\alpha_t} \|\tilde{\nabla} g_t\|_*^2 \left(\|\tilde{\nabla} f_t\|_*^2 + \|\tilde{\nabla} g_t\|_*^2\right) + \sum_{t=1}^T \frac{Q_t}{2\alpha_t} \|\tilde{\nabla} g_t\|_*^2 - \sum_{t=1}^T \gamma_t. \tag{60}$$

Finally, as $Q_1 = 0$, (60) completes the proof of this lemma.

B.4 Proof of Lemma 4.5: Time-Varying Drift Lemma for the Bandit Setting

Lemma B.5. *For $t \geq 1$,*

$$-G - RL_g \ell \leq Q_{t+1} - Q_t \leq G + RL_g \ell + \gamma_t.$$

Moreover, for any $s \geq t$,

$$-G - \frac{R^2}{2} - \frac{1}{2} qdp_*^2 L_g^2 \leq \mathbb{E}\left[Q_{s+1} - Q_s \mid \mathcal{H}_{t-1}\right] \leq G + \frac{R^2}{2} + \frac{1}{2} qdp_*^2 L_g^2 + \gamma_t.$$

Proof. Let us first argue that

$$-G - R\|\tilde{\nabla} g_t\|_* \leq Q_{t+1} - Q_t \leq G + R\|\tilde{\nabla} g_t\|_* + \gamma_t \tag{61}$$

holds. Here, for the upper bound, note that $Q_{t+1} = \max\left\{Q_t + g_t(\boldsymbol{x_t}) + \tilde{\nabla} g_t^\top (\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t, 0\right\}$ and

$$g_t(\boldsymbol{x_t}) + \tilde{\nabla} g_t^\top (\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) \leq g_t(\boldsymbol{x_{t+1}}) \leq |g_t(\boldsymbol{x_t})| + \|\tilde{\nabla} g_t\|_* \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\| \leq G + R\|\tilde{\nabla} g_t\|_*$$

by Assumption 1. Then $Q_{t+1} \leq \max\left\{Q_t + G + R\|\tilde{\nabla} g_t\|_* + \gamma_t, 0\right\}$ holds. Since $Q_t + G + R\|\tilde{\nabla} g_t\|_* + \gamma_t$ is nonnegative, it follows that $Q_{t+1} \leq Q_t + G + R\|\tilde{\nabla} g_t\|_* + \gamma_t$. For the lower bound, we deduce from $Q_{t+1} = \max\left\{Q_t + g_t(\boldsymbol{x_t}) + \tilde{\nabla} g_t^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t, 0\right\}$ that

$$Q_{t+1} \geq Q_t + g_t(\boldsymbol{x_t}) + \tilde{\nabla} g_t^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) + \gamma_t \geq Q_t - G - R\|\tilde{\nabla} g_t\|_*$$

where the second inequality is comes from $g_t(\boldsymbol{x_t}) \geq -G$, $\tilde{\nabla} g_t^\top(\boldsymbol{x_{t+1}} - \boldsymbol{x_t}) \geq -\|\tilde{\nabla} g_t\|_* \|\boldsymbol{x_{t+1}} - \boldsymbol{x_t}\| \geq -R\|\tilde{\nabla} g_t\|_*$, and $\gamma_t \geq 0$. From (61), we obtain the desired bounds on $Q_{t+1} - Q_t$ (the first part) by applying the bound on $\|\tilde{\nabla} g_t\|_*$ shown in Lemma 4.1.

For the second part, it is sufficient to consider the following.

$$\begin{aligned}
\mathbb{E}\left[R\|\tilde{\nabla} g_s\|_* \mid \mathcal{H}_{t-1}\right] &\leq \mathbb{E}\left[\frac{R^2}{4} + \|\tilde{\nabla} g_s\|_*^2 \mid \mathcal{H}_{t-1}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{R^2}{2} + \frac{1}{2}\|\tilde{\nabla} g_s\|_*^2 \mid \mathcal{H}_s^0\right] \mid \mathcal{H}_{t-1}\right] \\
&= \mathbb{E}\left[\frac{R^2}{2} + \frac{1}{2} q d p_*^2 L_g^2 \mid \mathcal{H}_{t-1}\right] \\
&= \frac{R^2}{2} + \frac{1}{2} q d p_*^2 L_g^2,
\end{aligned}$$

where the first inequality holds because $2pq \leq p^2 + q^2$. □

Proof of Lemma 4.5. **Parts (a) and (b)** Note that $Q_{t+1} - Q_t \leq G + \gamma_t$ for the full information setting (Lemma A.3) whereas the upper bound on $Q_{t+1} - Q_t$ due to Lemma B.5 has an additional term $RL_g\ell$. Following the same proof of Lemma 3.4, we deduce that $Q_t \leq C(2(C+2L_g)/\epsilon)^2 + (G + RL_g\ell + \epsilon)t$ for $t \geq 1$ and $Q_{t+1} - Q_t \leq G + RL_g\ell + \epsilon$, as required.

Part (c). Let $t \geq (2(C + 2L_g)/\epsilon)^2$. The case when $Q_t \leq \theta_t(\tau)$ is clear thanks to Lemma B.5 and part (b). Now suppose that $Q_t \geq \theta_t(\tau)$. Recall that $1 \leq \tau \leq t + 1$. We will show that

$$\mathbb{E}\left[Q_{t+\tau}^2 \mid \mathcal{H}_{t-1}\right] \leq \left(Q_t - \frac{\epsilon}{4}\tau\right)^2 - \frac{\epsilon^2}{16}\tau^2. \tag{62}$$

If (62) holds, as $\sqrt{\cdot}$ is a concave function over $\mathbb{R}_+$, we deduce from Jensen's inequality that

$$\mathbb{E}\left[Q_{t+\tau} \mid \mathcal{H}_{t-1}\right] \leq \sqrt{\mathbb{E}\left[Q_{t+\tau}^2 \mid \mathcal{H}_{t-1}\right]} \leq Q_t - \frac{\epsilon}{4}\tau.$$

Therefore, it is sufficient to show that (62) holds true. Note that

$$Q_{t+\tau}^2 = Q_t^2 + 2\sum_{s=t}^{t+\tau-1} \Delta_s.$$

As in the proof of Lemma 3.4(c), we analyze the drift terms to provide the desired bound on $Q_{t+\tau}^2$. By Lemma B.3 and the observation that $\gamma_t \leq \epsilon/2$ for any $t \geq (2(C + 2L_g)/\epsilon)^2$,

$$\Delta_s \leq Q_s\left(g_s(\boldsymbol{x_s}) + \tilde{\nabla} g_s^\top(\boldsymbol{x_{s+1}} - \boldsymbol{x_s}) + \gamma_s\right) + (G + \epsilon)^2 + R^2\|\tilde{\nabla} g_s\|_*^2 \tag{63}$$

for $s = t, \dots, t+\tau-1$. Moreover,

$$\begin{aligned}
&Q_s(g_s(\boldsymbol{x_s}) + \tilde{\nabla} g_s^\top(\boldsymbol{x_{s+1}} - \boldsymbol{x_s}) + \gamma_s) \\
&\le Q_s(g_s(\boldsymbol{x_s}) + \tilde{\nabla} g_s^\top(\hat{\boldsymbol{x}} - \boldsymbol{x_s}) + \gamma_s) \\
&\quad + V_s \tilde{\nabla} f_s^\top(\hat{\boldsymbol{x}} - \boldsymbol{x_{s+1}}) + \alpha_s\left(D(\hat{\boldsymbol{x}}, \boldsymbol{x_s}) - D(\hat{\boldsymbol{x}}, \boldsymbol{x_{s+1}}) - D(\boldsymbol{x_{s+1}}, \boldsymbol{x_s})\right) \\
&\le Q_s(g_s(\boldsymbol{x_s}) + \tilde{\nabla} g_s^\top(\hat{\boldsymbol{x}} - \boldsymbol{x_s}) + \gamma_s) + \frac{V_s}{2}(R^2 + \|\tilde{\nabla} f_s\|_*^2) + \alpha_s\left(D(\hat{\boldsymbol{x}}, \boldsymbol{x_s}) - D(\hat{\boldsymbol{x}}, \boldsymbol{x_{s+1}})\right)
\end{aligned} \tag{64}$$

where the first inequality comes from the inequality given in Lemma A.1 with $\boldsymbol{x}$ set to $\hat{\boldsymbol{x}}$ and the second inequality is because

$$V_s \tilde{\nabla} f_s^\top(\hat{\boldsymbol{x}} - \boldsymbol{x_{s+1}}) \le V_s \|\tilde{\nabla} f_s\|_* \|\hat{\boldsymbol{x}} - \boldsymbol{x_{s+1}}\| \le \frac{V_s}{2}\left(\|\tilde{\nabla} f_s\|_*^2 + \|\hat{\boldsymbol{x}} - \boldsymbol{x_{s+1}}\|^2\right).$$

Moreover, as in the proof of Lemma 4.3, we may argue the following.

$$\begin{aligned}
&\mathbb{E}\left[Q_s\left(g_s(\boldsymbol{x_s}) + \tilde{\nabla} g_s^\top(\hat{\boldsymbol{x}} - \boldsymbol{x_s}) + \gamma_s\right) \mid \mathcal{H}_{t-1}\right] \\
&= \mathbb{E}\left[Q_s(g_s(\boldsymbol{x_s}) + \gamma_s) + \mathbb{E}\left[\tilde{\nabla} g_s^\top(\hat{\boldsymbol{x}} - \boldsymbol{x_s}) \mid \mathcal{H}_s^0\right] \mid \mathcal{H}_{t-1}\right] \\
&= \mathbb{E}\left[Q_s(g_s(\boldsymbol{x_s}) + \gamma_s) + \nabla \hat{g}_s(\boldsymbol{x_s})^\top(\hat{\boldsymbol{x}} - \boldsymbol{x_s}) \mid \mathcal{H}_{t-1}\right] \\
&\le \mathbb{E}\left[Q_s(g_s(\boldsymbol{x_s}) - \hat{g}_s(\boldsymbol{x_s})) \mid \mathcal{H}_{t-1}\right] + \mathbb{E}\left[Q_t \hat{g}_s(\hat{\boldsymbol{x}}) \mid \mathcal{H}_{t-1}\right] \\
&= \mathbb{E}\left[Q_s(g_s(\boldsymbol{x_s}) - \hat{g}_s(\boldsymbol{x_s})) \mid \mathcal{H}_{t-1}\right] + \mathbb{E}\left[Q_s(\hat{g}_s(\hat{\boldsymbol{x}}) - g_s(\hat{\boldsymbol{x}})) \mid \mathcal{H}_{t-1}\right] + \mathbb{E}\left[Q_s g_s(\hat{\boldsymbol{x}}) \mid \mathcal{H}_{t-1}\right] \\
&= \mathbb{E}\left[Q_s(g_s(\boldsymbol{x_s}) - \hat{g}_s(\boldsymbol{x_s})) \mid \mathcal{H}_{t-1}\right] + \mathbb{E}\left[Q_s(\hat{g}_s(\hat{\boldsymbol{x}}) - g_s(\hat{\boldsymbol{x}})) \mid \mathcal{H}_{t-1}\right] + \mathbb{E}\left[\mathbb{E}\left[Q_s g_s(\hat{\boldsymbol{x}}) \mid \mathcal{H}_{s-1}\right] \mid \mathcal{H}_{t-1}\right] \\
&\le \mathbb{E}\left[Q_s\left(\bar{g}(\hat{\boldsymbol{x}}) + 2\delta_s L_g + \gamma_s\right) \mid \mathcal{H}_{t-1}\right]
\end{aligned} \tag{65}$$

where the equalities are due to the linearity of expectation and the tower rule. Furthermore, we have

$$2\delta_s L_g + \gamma_s \le 2\delta_t L_g + \gamma_t = \frac{2L_g + C}{\sqrt{t}} \le \frac{\epsilon}{2}. \tag{66}$$

Lastly,

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{\nabla} f_s\|_*^2 \mid \mathcal{H}_{t-1}\right] &= \mathbb{E}\left[\mathbb{E}\left[\|\tilde{\nabla} f_s\|_*^2 \mid \mathcal{H}_s^0\right] \mid \mathcal{H}_{t-1}\right] \le q d p_*^2 L_f^2, \\
\mathbb{E}\left[\|\tilde{\nabla} g_s\|_*^2 \mid \mathcal{H}_{t-1}\right] &= \mathbb{E}\left[\mathbb{E}\left[\|\tilde{\nabla} g_s\|_*^2 \mid \mathcal{H}_s^0\right] \mid \mathcal{H}_{t-1}\right] \le q d p_*^2 L_g^2.
\end{aligned} \tag{67}$$

Taking the expectation of both sides of (64), applying the bounds (63) and (65), (66), and (67), and using Slater's condition (Assumption 2), we deduce the following.

$$\begin{aligned}
&\mathbb{E}\left[\Delta_s \mid \mathcal{H}_{t-1}\right] \\
&\le -\frac{\epsilon}{2} \cdot \mathbb{E}\left[Q_s \mid \mathcal{H}_{t-1}\right] + \alpha_s \mathbb{E}\left[D(\hat{\boldsymbol{x}}, \boldsymbol{x_s}) - D(\hat{\boldsymbol{x}}, \boldsymbol{x_{s+1}}) \mid \mathcal{H}_{t-1}\right] \\
&\quad + \frac{V_s}{2}\left(R^2 + q d p_*^2 L_f^2\right) + (G+\epsilon)^2 + R^2 q d p_*^2 L_g^2.
\end{aligned} \tag{68}$$

Then, following the argument of the proof of Lemma 3.4, we obtain

$$\begin{aligned}
\sum_{s=t}^{t+\tau-1} \mathbb{E}\left[\Delta_s \mid \mathcal{H}_{t-1}\right] \le -\frac{\epsilon}{2}\tau Q_t + \frac{\epsilon}{2}\left(G + \frac{R^2}{2} + \frac{1}{2} q d p_*^2 L_g^2\right)\tau^2 \\
+ 2R^2 \alpha_t + \left(R^2 + q d p_*^2 L_f^2\right)\tau V_t + \left((G+\epsilon)^2 + R^2 q d p_*^2 L_g^2\right)\tau.
\end{aligned} \tag{69}$$

Then it follows from (69) that

$$\begin{aligned}\mathbb{E}\left[Q_{t+\tau}^2 \mid \mathcal{H}_{t-1}\right] &= Q_t^2 + 2\sum_{s=t}^{t+\tau-1} \mathbb{E}\left[\Delta_s \mid \mathcal{H}_{t-1}\right] \\ &\leq Q_t^2 - \epsilon\tau Q_t + \epsilon\left(G + \frac{R^2}{2} + \frac{1}{2}qdp_*^2L_g^2\right)\tau^2 \\ &\quad + 4R^2\alpha_t + 2\left(R^2 + qdp_*^2L_f^2\right)\tau V_t + 2\left((G+\epsilon)^2 + R^2qdp_*^2L_g^2\right)\tau.\end{aligned} \tag{70}$$

Recall that

$$\begin{aligned}\theta_t(\tau) &= (2G + R^2 + qdp_*^2L_g^2)\tau + \frac{8R^2\alpha_t}{\epsilon\tau} + \frac{4(R^2 + qdp_*^2L_f^2)V_t}{\epsilon} + \frac{4\left((G+\epsilon)^2 + R^2qdp_*^2L_g^2\right)}{\epsilon} \\ &= \frac{2}{\epsilon\tau}\left(\epsilon\left(G + \frac{R^2}{2} + \frac{1}{2}qdp_*^2L_g^2\right)\tau^2 + 4R^2\alpha_t + 2\left(R^2 + qdp_*^2L_f^2\right)\tau V_t + 2\left((G+\epsilon)^2 + R^2qdp_*^2L_g^2\right)\tau\right).\end{aligned}$$

Therefore, if $Q_t \geq \theta_t(\tau)$, then it follows from (70) that

$$\mathbb{E}\left[Q_{t+\tau}^2 \mid \mathcal{H}_{t-1}\right] \leq Q_t^2 - \frac{\epsilon\tau}{2}Q_t = \left(Q_t - \frac{\epsilon\tau}{4}\right)^2 - \frac{\epsilon^2\tau^2}{16},$$

which proves (62), as required. □

B.5 Proof of Lemma 4.6: Time-Varying Bound on the Expected Virtual Queue Size for the Bandit Setting

As for the full information setting, let $\delta = (G + \epsilon + RL_g\ell)$ and $\xi = \epsilon/4$. Moreover, let r and ρ be two constants defined as

$$r = \frac{\xi}{4\lceil\sqrt{T}\rceil\delta^2} \quad \text{and} \quad \rho = 1 - \frac{\xi^2}{8\delta^2}.$$

Since $\xi \leq \delta$, we know that $0 < \rho < 1$.

Lemma B.6. *Let $t \geq 1$. For any $1 \leq \tau \leq \sqrt{T}$ satisfying $t - \tau \geq (2(C + 2L_g)/\epsilon)^2$ and $t \geq 2\tau - 1$, we have*

$$\mathbb{E}\left[e^{rQ_t}\right] \leq \rho\mathbb{E}\left[e^{rQ_{t-\tau}}\right] + e^{r\delta\tau}e^{r\theta_{t-\tau}}.$$

Proof. As $t \geq 2\tau - 1$, we have $t - \tau \geq \tau - 1$. Henceforth, we use the notation $w_t = Q_t - Q_{t-\tau}$. Note that

$$w_t = \sum_{s=t-\tau}^{t-1} Q_{s+1} - Q_s \leq \tau\delta$$

holds due to Lemma 4.5(b) because $t - \tau \geq (2(C + 2L_g)/\epsilon)^2$. Furthermore, as $\tau \leq \sqrt{T}$,

$$rw_t \leq r\tau\delta \leq \frac{\xi}{4\delta} \leq 1.$$

The rest of the argument is the same as the proof of Lemma A.4. □

Proof of Lemma 4.6. Let $t \geq 1$, and let $\tau = \lceil\sqrt{t}\rceil$. For any s, we use θ_s to denote $\theta_s(\tau)$. As in the proof of Lemma 3.5, we first consider the case where

$$t \geq \max\left\{9, \left(\frac{2(C + 2L_g)}{\epsilon}\right)^4\right\}.$$

Then $t \geq 2\tau$ and $\tau \geq (2(C+2L_g)/\epsilon)^2$. Moreover, $\lfloor t/\tau \rfloor = k$ for some $k \geq 2$, in which case

$$t-(k-1)\tau \geq \tau \geq (2(C+2L_g)/\epsilon)^2.$$

Following the same proof argument of Lemma 3.5, we deduce the following.

$$\mathbb{E}\left[e^{rQ_t}\right] \leq e^{rC(2(C+2L_g)/\epsilon)^2} e^{2r\delta\tau} e^{r\theta_t}\left(\frac{\rho^{k-1}}{1-\rho}+\sum_{i=0}^{k-2}\rho^i\right) = \frac{1}{1-\rho} e^{rC(2(C+2L_g)/\epsilon)^2} e^{2r\delta\tau} e^{r\theta_t}. \tag{71}$$

As $p(x) = e^{rx}$ is convex over $x \in \mathbb{R}$, Jensen's inequality implies that

$$e^{r\mathbb{E}[Q_t]} \leq \mathbb{E}\left[e^{rQ_t}\right].$$

Then, by (71), we obtain

$$\begin{aligned} E\left[Q_t\right] &\leq \theta_t + 2\delta\tau + C\left(\frac{2(C+2L_g)}{\epsilon}\right)^2 + \log\frac{1}{1-\rho} \\ &= \theta_t + 2(G+RL_g\ell+\epsilon)\tau + \log\frac{128(G+RL_g\ell+\epsilon)^2}{\epsilon^2} + C\left(\frac{2(C+2L_g)}{\epsilon}\right)^2 \\ &\leq \theta_t + 4(G+RL_g\ell+\epsilon)\sqrt{t} + \log\frac{128(G+RL_g\ell+\epsilon)^2}{\epsilon^2} + C\left(\frac{2(C+2L_g)}{\epsilon}\right)^2 \\ &\leq \theta_t + 4(G+RL_g\ell+\epsilon)\sqrt{t} + \log\frac{128(G+RL_g\ell+\epsilon)^2}{\epsilon^2} + C\left(\frac{2(C+2L_g)}{\epsilon}\right)^2 \\ &\quad + (G+RL_g\ell+\epsilon)\left(9+\left(\frac{2(C+2L_g)}{\epsilon}\right)^4\right) \end{aligned}$$

where the second inequality holds because $\tau \leq 2\sqrt{t}$. We can also show that the inequality holds even when $t < \max\{9, (2(C+2L_g)/\epsilon)^4\}$, as required. □

B.6 Completing the Proofs of Theorem 4.7 (Constraint Violation) and Theorem 4.8 (Regret)

Plugging in the formula of θ_t to Lemma 4.6, it follows that for any $t \geq 1$,

$$\begin{aligned} \mathbb{E}\left[Q_t\right] \leq\ & 2\left(4G + R^2 + qdp_*^2L_g^2 + 2\epsilon + 2RL_g\ell + \frac{12R^2+4qdp_*^2L_f^2}{\epsilon}\right)\sqrt{t} \\ &+ \frac{4(G+\epsilon)^2+4R^2qdp_*^2L_g^2}{\epsilon} + \log\frac{128(G+RL_g\ell+\epsilon)^2}{\epsilon^2} \\ &+ C\left(\frac{2(C+2L_g)}{\epsilon}\right)^2 + (G+RL_g\ell+\epsilon)\left(9+\left(\frac{2(C+2L_g)}{\epsilon}\right)^4\right). \end{aligned} \tag{72}$$

Proof of Theorem 4.7. By Lemma 4.4 and the expected queue size bound (72), we obtain the following.

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T g_t(\boldsymbol{x_t})\right]\\
&\le \mathbb{E}\left[Q_{T+1}\right] + qdp_*^2(L_f^2+L_g^2)\sum_{t=1}^T \frac{V_t}{4\alpha_t} + qdp_*^2L_g^2\sum_{t=1}^T\frac{\mathbb{E}[Q_t]}{2\alpha_t} - \sum_{t=1}^T \gamma_t\\
&\le \left(\frac{qdp_*^2(L_f^2+L_g^2)}{2} + (2qdp_*^2L_g^2+4)\left(4G+R^2+qdp_*^2L_g^2+2\epsilon+2RL_g\ell+\frac{12R^2+4qdp_*^2L_f^2}{\epsilon}\right)\right)\sqrt{T}\\
&\quad - C\sqrt{T}\\
&\quad + \left(\frac{4(G+\epsilon)^2+4R^2qdp_*^2L_g^2}{\epsilon} + \log\frac{128(G+RL_g\ell+\epsilon)^2}{\epsilon^2}\right)\left(1+\frac{qdp_*^2L_g^2(1+\log T)}{2}\right)\\
&\quad + \left(C\left(\frac{2(C+2L_g)}{\epsilon}\right)^2 + (G+RL_g\ell+\epsilon)\left(9+\left(\frac{2(C+2L_g)}{\epsilon}\right)^4\right)\right)\left(1+\frac{qdp_*^2L_g^2(1+\log T)}{2}\right).
\end{aligned} \tag{73}$$

Since we set C as

$$C = \frac{qdp_*^2(L_f^2+L_g^2)}{2} + (2qdp_*^2L_g^2+4)\left(4G+R^2+qdp_*^2L_g^2+2\epsilon+2RL_g\ell+\frac{12R^2+4qdp_*^2L_f^2}{\epsilon}\right)+1,$$

we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T g_t(\boldsymbol{x_t})\right]\\
&\le -\sqrt{T}\\
&\quad + \left(\frac{4(G+\epsilon)^2+4R^2qdp_*^2L_g^2}{\epsilon} + \log\frac{128(G+RL_g\ell+\epsilon)^2}{\epsilon^2}\right)\left(1+\frac{qdp_*^2L_g^2(1+\log T)}{2}\right)\\
&\quad + \left(C\left(\frac{2(C+2L_g)}{\epsilon}\right)^2 + (G+RL_g\ell+\epsilon)\left(9+\left(\frac{2(C+2L_g)}{\epsilon}\right)^4\right)\right)\left(1+\frac{qdp_*^2L_g^2(1+\log T)}{2}\right).
\end{aligned}$$

Therefore, there exists some positive constant B such that for any $T \ge B$,

$$\mathbb{E}\left[\sum_{t=1}^T g_t(\boldsymbol{x_t})\right] \le 0,$$

as required. □

Proof of Theorem 4.8. If $C = O(1)$, then $\mathbb{E}[Q_t] \le B\sqrt{t}$ for some constant B that depends only on G, D_g, D_f, R, ϵ and $L_f, L_g, q, p_*, d, \ell$ by (72). By Lemma 4.3, it follows that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T f_t(\boldsymbol{x_t}) - \sum_{t=1}^T f_t(\boldsymbol{x}^*)\right] &\le \left(\frac{C(G+C)}{\epsilon}\right)^2 + K_2\sqrt{T} + (C+2L_g)\sum_{t=1}^T\frac{\mathbb{E}[Q_t]}{t}\\
&\le \left(\frac{C(G+C)}{\epsilon}\right)^2 + K_2\sqrt{T} + (C+2L_g)B\sum_{t=1}^T\frac{1}{\sqrt{t}}\\
&= O(\sqrt{T}),
\end{aligned}$$

as required. □