Decision-Dependent Risk Minimization in Geometrically Decaying Dynamic Environments

Mitias Ray, Lillian J. Ratliff, Dmitriy Drusvyatskiy, Maryam Fazel

Abstract

This paper considers the problem of expected loss minimization given a data distribution that is dependent on the decision-maker’s action and evolves dynamically in time according to a geometric decay process. Motivated by practice, empirical information settings are considered: namely, the decision-maker either has oracle access, for a fixed batch size, to the empirical gradient of the loss (first order setting), or the empirical loss function (zero order setting). Novel algorithms for each of these settings are introduced, both of which operate on the same underlying principle: the decision-maker repeatedly deploys a fixed decision over a fixed length epoch, thereby allowing the dynamically changing environment to sufficiently mix before updating the decision. The proposed algorithms are shown to converge to the optimal point. Specifically, high-probability sample complexity guarantees are given, which depend exponentially on the epoch length and logarithmically on the batch size. The algorithms are evaluated on a “semi-synthetic” example using real world data from the SFpark dynamic pricing pilot study; it is shown that the announced prices result in an improvement for the institution’s objective (target occupancy), while achieving an overall reduction in parking rates.

1 Introduction

Traditionally, supervised machine learning algorithms are trained based on past data under the assumption that the past data is representative of the future. However, machine learning algorithms are increasingly being used in settings where the output of the algorithm changes the environment and hence, the data distribution. Indeed, loan procurement processes, online labor markets [Anagnostopoulos et al., 2018, Horton, 2010], predictive policing [Lum and Isaac, 2016], on-street parking [Pierce and Shoup, 2018, Dowling et al., 2020], and vehicle sharing markets [Banerjee et al., 2015] are all examples of real-world settings in which not only is algorithmic decision-making commonplace, but the algorithm’s decisions change the underlying data distribution in large part because the algorithm interacts with strategic users.

To address this problem, the machine learning community introduced the problem of performative prediction which models the data distribution as being decision-dependent thereby accounting for feedback induced distributional shift [Perdomo et al., 2020, Miller et al., 2021, Drusvyatskiy and Xiao, 2020, Brown et al., 2020, Mendler-Düünner et al., 2020]. With the exception of [Brown et al., 2020], this work has focused on static environments.

In many of the aforementioned application domains, however, the underlying data distribution also may have memory or even be changing dynamically in time. When a decision-making mechanism is announced it may take time to see the full effect of the decision as the environment and strategic data sources respond given their prior history or interactions.

For example, many municipalities announce quarterly a new quasi-static set of prices for on-street parking. In this setting, the institution may adjust parking rates for certain blocks in order to to achieve a desired occupancy range to reduce cruising phenomena and increase business district vitality [Fiez et al., 2018, Dowling et al., 2017, Pierce and Shoup, 2013, Shoup, 2006]. For instance, in high traffic areas, the institution may announce increased parking rates to free up parking spots and redistribute those drivers to less populated blocks. However, upon announcing a new price, the population may react slowly, whether it be from initially being unaware of the price change, to facing natural inconveniences from changing one’s parking routine. This introduces dynamics into our setting; hence, the data distribution takes time to equilibrate after the pricing change is made.
Motivated by such scenarios, we study the problem of decision-dependent risk minimization (or, synonymously, performative prediction) in dynamic settings wherein the underlying decision-dependent distribution evolves according to a geometrically decaying process. Taking into account the time it takes for a decision to have the full effect on the environment, we devise an algorithmic framework for finding the optimal solution in settings where the decision maker has access to different types of information—namely, first-order (gradient) feedback, and zeroth-order (or bandit) feedback.

For both information settings (gradient access and loss function access, via the appropriate oracle), the decision-maker deploys the current decision for a fixed number of steps (the length of an epoch) allowing the dynamically evolving distribution to approach the fixed point distribution for that announced decision. At the end of the epoch, the decision is updated using a first-order or zeroth-order oracle.

One interpretation of this procedure is that the environment is operating on a faster timescale compared to the update of the decision-maker’s action. For instance, consider the dynamically changing distribution as the data distribution corresponding to a population of strategic data sources. The phase during which the same decision is deployed for a fixed number of steps can be interpreted as the population of agents adapting at a faster rate than the update of the decision. As previously noted, this in fact occurs in many practical settings such as on-street parking, where prices and policies more generally are quasi-static, meaning they are updated infrequently relative to actual curbspace utilization or resource consumption.

1.1 Contributions

For the decision-dependent learning problem in geometrically decaying environments, we propose algorithms relying on a first-order or zeroth-order oracle that converge to the optimal point (under appropriate assumptions, which make the performative risk minimization problem strongly convex). We also characterize the sample complexity of the proposed algorithms.

- **First Order Oracle** (Algorithm 1, Section 3): with probability $1 - \rho$, for arbitrary error tolerance $\xi > 0$, epoch length $n \geq \Omega(\log \xi)$, time horizon $T$, and batch size $N \geq \Omega(q \log q)$, the sample complexity is
  $$O((1 - \alpha/M)^T + \xi + \sqrt{q \log(1/\rho) \log(N)/N}),$$
  where the loss is $\alpha$-strongly convex and $M$-smooth, and $q$ is the dimension of the decision space.

- **Zero Order Oracle** (Algorithm 2, Section 3): with probability $1 - \rho$, for $n \geq \Omega(\log \eta)$, $\eta > 1/(3\alpha T)$, and $N \geq \Omega(q \log q)$, the sample complexity is $O(T^{-1/3} + \sqrt{q \log(1/\rho) \log(N)/N})$.

The primary technical novelty arises from bounding the error between expected and empirical gradients, considering that the distribution with respect to which gradient information is available is changing dynamically over time. We show that with high probability, the sample complexity is nearly the optimal rate in the setting where the expected loss is available, and characterize how it depends on the batch size $N$.

The algorithms are applied to a set of semi-synthetic experiments using real data from the SFpark pilot study on the use of dynamic pricing to manage curbside parking (Section 5). We show that decision-dependent optimization of the price enables the target occupancy to be reached (70%) and reduces the overall price experienced by users as compared to the SFpark pilot outcomes. These results suggest that optimizing taking into consideration feedback-induced distribution shift even in a dynamic environment leads to the institution—and perhaps surprisingly, the user as well—experiencing lower expected cost. Moreover, there are important secondary effects of this improvement including increased access to parking—hence, business district vitality—and reduced circling for parking and congestion which not only saves users time but also reduces carbon emissions [Shoup, 2006].

1.2 Related work

**Dynamic Decision-Dependent Optimization.** As hinted above, dynamic decision-dependent optimization has been considered quite extensively in the stochastic optimization literature wherein the problem of recourse arises due to decision-makers being able to make a secondary decision after some information has been revealed [Jonsbråten et al., 1998, Goel and Grossmann, 2004, Varaiya and Wets, 1988]. In this problem, the goal of the institution is to solve a multi-stage stochastic program, in which the probability distribution
of the population is a function of the decision announced by the institution. This multi-stage procedure models a dynamic process. Unlike the setting considered in this paper, the institution has the ability to make a recourse decision upon observing full or partial information about the stochastic components.

**Reinforcement Learning.** Reinforcement learning is a more closely related problem in the sense that a decision is being made over time where the environment dynamically changes as a function of the state and the decision-maker’s actions [Sutton and Barto, 2018]. A subtle but important difference is that the setting we consider is such that the decision maker’s objective is to find the action $\theta^*$ which optimizes the decision-dependent expected risk at the fixed point distribution (cf. Definition 1, Section 2) induced by $\theta^*$ and the environment dynamics. This is in contrast to finding a policy which is a state-dependent distribution over actions given an accumulated cost over time. Our setting can be viewed as a special case of the general reinforcement learning problem, however with additional structure that is both practically well-motivated, and beneficial to exploit in the design and analysis of algorithms. More concretely, we crucially exploit the assumed model of environment dynamics (in this case, the geometric decay), the distribution dependence, and convexity to obtain strong convergence guarantees for the algorithms proposed herein.

**Performative prediction.** As alluded to in the introductory remarks, the most closely related body of literature is on performative prediction wherein the decision-maker or optimizer takes into consideration that the underlying data distribution depends on the decision. Performative prediction has been studied in both static and dynamic environments. In the static setting, as the distribution shifts, a naïve strategy is to re-train the model on this new distribution using heuristics to determine when to trigger the retraining process. Under the guise that if retraining is repeated, eventually the distribution will stabilize, early works on performative prediction—such as the works of Perdomo et al. [2020] and Mendler-Dünner et al. [2020]—studied this equilibrium notion, and called these points *performatively stable*. Mendler-Dünner et al. [2020] and Drusvyatskiy and Xiao [2020] study stochastic optimization algorithms applied to the performative prediction problem and recover optimal convergence guarantees to the performatively stable point. Yet, performatively stable points may differ from the optimal solution of the decision-dependent risk minimization problem as was shown in Perdomo et al. [2020]. Taking this gap between stable and optimal points into consideration, Miller et al. [2021] characterize when the performative prediction problem is strongly convex, and devise a two-stage algorithm for finding the so-called *performatively optimal* solution—that is, the optimal solution to the decision-dependent risk minimization problem—when the decision-dependent distribution is from the location-scale family.

None of the aforementioned works consider dynamic environments. Brown et al. [2020] is the first paper, to our knowledge, to investigate the dynamic setting for performative prediction. Assuming regularity properties of the dynamics, they show that classical retraining algorithms (repeated gradient descent and repeated risk minimization) converge to the performatively stable point of the expected risk at the corresponding fixed point distribution. Counter to this, in this paper we propose algorithms for the dynamic setting which target performatively optimal points.

## 2 Preliminaries

As noted in the introduction, the algorithms we propose proceed in epochs, wherein the decision-maker deploys the current decision for a fixed number of steps (the length of the epoch) allowing the dynamically evolving distribution to mix—in other words, enough time for the dynamics to sufficiently settle down. For such an approach to work, it needs to be the case that when the same decision is deployed repeatedly, the distribution converges to a fixed point. Further, convexity (in the decision variable) of the expected loss function given the fixed point distribution ensures the optimal solution is unique and obtainable via gradient-based learning. This motivates the assumptions we introduce in this section.

**Problem Formulation.** For a distribution $\mathcal{D}(\theta)$, the performative risk is the loss incurred by the institution (decision-maker), evaluated by a loss function $\ell : Z \times \Theta \to \mathbb{R}$, for announcing $\theta \in \Theta \subset \mathbb{R}^q$, where $Z \subset \mathbb{R}^m$ and we assume norm bounds $r \leq \|\theta\| \leq R$, $\forall \theta \in \Theta$.
Given a loss function $\ell$, the **performative risk** is defined as

$$L(\theta) = \mathbb{E}_{z \sim D(\theta)} [\ell(z, \theta)].$$

In the dynamic setting, the distribution evolves in time according to a geometrically decaying process. That is, letting $d_t$ denote the data distribution at time $t$, the distribution after a single deployment of the decision $\theta_t$ is given as $d_{t+1} = T(d_t, \theta_t)$, where

$$T(d, \theta) = \delta d + (1 - \delta) D(\theta),$$

(1)

and $\delta \in [0, 1)$ is the geometric decay rate. One interpretation of this transition map is that it captures the phenomenon that for each time, a $1 - \delta$ fraction of the population becomes aware of the machine learning model $\theta$ being used by the institution. Another interpretation is that the environment (and strategic data sources in the environment) has memory based on past interactions which is captured in the ‘state’ of the distribution, and the effects of the past decay geometrically at a rate of $\delta$.

Observe that given the geometrically decaying dynamics in (1), for any $\theta \in \Theta$, the distribution $D(\theta)$ is trivially a fixed point (i.e., $T(D(\theta), \theta) = D(\theta)$). Moreover, for any $\delta \in (0, 1)$ and fixed $\theta \in \Theta$, the distribution $d_t$ converges at an exponential rate to $D(\theta)$.

Performatively optimality is defined as follows.

**Definition 1** (Performatively optimal point). For a given distribution $D(\theta)$, the decision vector $\theta^* \in \Theta$ is a performatively optimal point if

$$\theta^* \in \arg\min_{\theta \in \Theta} L(\theta).$$

As is commonly assumed in the performative prediction literature [Miller et al., 2021, Perdomo et al., 2020], we assume that the distribution $D(\theta)$ is $\epsilon$-Lipschitz (commonly referred to as $\epsilon$-sensitive) with respect to the Wasserstein-1 distance denoted by $W_1$.

**Assumption 1** (Lipschitz Distributions). There exists an $\epsilon \geq 0$ satisfying $W_1(D(\theta), D(\theta')) \leq \epsilon \|\theta - \theta'\|$ for all $\theta, \theta' \in \Theta$.

We also assume that the loss is Lipschitz, and the expected loss is smooth (i.e., gradient Lipschitz).

**Assumption 2**. The loss function $\ell(z, \theta)$ is $G_z$-Lipschitz in $z$, and $G_\theta$-Lipschitz in $\theta$—i.e., for any $\theta \in \Theta$, we have that $|\ell(z, \theta) - \ell(z', \theta)| \leq G_z \|z - z'\|$, and for any $z \in Z$, $|\ell(z, \theta) - \ell(z, \theta')| \leq G_\theta \|\theta - \theta'\|$.

**Assumption 3**. The performative risk (expected loss) $L(\theta)$ is $M$-smooth—i.e., the gradient satisfies $\|\nabla L(\theta) - \nabla L(\theta')\| \leq M \|\theta - \theta'\|$, for all $\theta, \theta' \in \Theta$.

For a given distribution $D(\cdot)$, when $L(\theta)$ is strongly convex in $\theta$, the performatively optimal point is unique.

**Assumption 4**. The loss function $\ell$ satisfies the following:

a. The partial gradient of $\ell(z, \theta)$ with respect to $\theta$ is $\beta$-Lipschitz in $z$—i.e., for any $\theta$, the partial gradient satisfies $\|\nabla_\theta \ell(z, \theta) - \nabla_\theta \ell(z', \theta)\| \leq \beta \|z - z'\|$, for all $z, z' \in Z$.

b. The loss $\ell(z, \theta)$ is $\gamma$-strongly convex in $\theta$—i.e., for any $\theta, \theta' \in \Theta$, the following holds:

$$\nabla_\theta \ell(z, \theta) \top (\theta' - \theta) + \frac{\gamma}{2} \|\theta' - \theta\|^2 \leq \ell(z, \theta') - \ell(z, \theta).$$

The following assumption implies a convex ordering on the random variables on which the loss is dependent.

**Assumption 5** (Mixture Dominance). The distribution map $D$ and loss $\ell$ satisfy mixture dominance—i.e., for any $\theta \in \Theta$ and $s \in (0, 1)$,

$$\mathbb{E}_{z \sim D(s\theta' + (1 - s)\theta'')} [\ell(z, \theta)] \leq \mathbb{E}_{z \sim D(\theta') + (1 - s) D(\theta'')} [\ell(z, \theta)],$$

for all $\theta', \theta'' \in \Theta$. 

4
The following result from Miller et al. [2021] states that the performative risk is strongly convex.

**Theorem 1** (Theorem 3.1, [Miller et al., 2021]). Under Assumptions 1, 4 and 5, \( \mathbb{E}_{z \sim D(\theta)} [\ell(z, \theta)] \) is \( \alpha = (\gamma - 2\epsilon\beta) \)-strongly convex.

Strong convexity ensures that there is a unique performatively optimal point, and hence upon convergence, the algorithms we propose in Section 3 obtain precisely the unique performatively optimal point.

### 3 Algorithms and Oracles

As alluded to in the introduction, the algorithms we propose for each of the information settings are similar in spirit—namely, they each operate by holding fixed a decision for the length \( n \) of each epoch and sampling or querying the environment until the distribution dynamics have mixed sufficiently towards the fixed point distribution.

In practice, different information may be available to the decision maker. We consider the following oracles.

a. **First Order (Algorithm 1).** The decision-maker has access to an empirical estimate of its gradient at \( d_t \)—namely,

\[
\hat{g}_t = \frac{1}{N} \sum_{k=1}^{N} \nabla \ell(z_k(\theta_t), \theta_t)
\]

where \( z_k(\theta_t) \sim d_t = \delta^n d_{t-1} + (1 - \delta^n) D(\theta_t) \). The gradient \( \nabla \ell(z_k(\theta), \theta) \) is the total derivative of \( \ell \) differentiating through the dependence of \( z_k \) on \( \theta \).

b. **Zeroth Order (Algorithm 2).** The decision-maker has a zero-order estimate of gradient using samples of the loss at \( d_t \)—namely,

\[
\hat{g}_t = \frac{1}{N} \sum_{k=1}^{N} \frac{\partial}{\partial \theta} \ell(z_k(\theta_t + \mu v_t), \theta_t + \mu v_t) v_t,
\]

where \( v_t \) is a random vector of unit length, and \( z_k(\theta_t + \mu v_t) \sim d_t = \delta^n d_{t-1} + (1 - \delta^n) D(\theta_t + \mu v_t) \).

To characterize the sample complexity in both of these settings, we need to bound the error between the empirical gradient information and the expected gradient information. Towards this end, we need an additional assumption that now explicitly depends on the algorithm structure—namely, that it proceeds in epochs of length \( n \).

**Assumption 6.** Given an epoch length \( n \), for every \( d_t = \delta^n d_{t-1} + (1 - \delta^n) D(\theta) \) induced by the initial distribution and the mapping \( D(\cdot) \), the loss \( \ell(z(\theta), \theta) \) is \( G \)-Lipschitz and \( H \)-Hessian Lipschitz.

Observe that in Assumption 6 we explicitly denote the dependence of \( z \) on \( \theta \) by \( z(\theta) \). From here forward we will use the notation \( z(\theta) \) to indicate the \( \theta \) dependence and just \( z \) when referring to a specific instance in \( Z \). This is to make clear that the derivatives in the analysis are constructed using the chain rule taking into account the data dependent distribution.

The location-scale family of distributions satisfies Assumption 6. The following example shows what the value of \( G \) is (\( H \) can similarly be derived).
Algorithm 2: Epoch-Based Zeroth Order Algorithm

**Initialization:** epoch length $n$, stepsize $\eta$, initial point $\theta_1$, query radius $\mu$, shrinking radius $\tau = \mu/r$ (where $r$ is the lower bound on $||\theta||$), horizon $T$, batch size $N$, initial distribution $d_0$;

```plaintext
for $t = 1, 2, \ldots, T$ do
    // Step 1: Mixing
    Sample vector $v_t$ from the unit sphere;
    Run $\theta_t + \mu v_t$ for $n$ steps, so $d_t = T(\ldots T(\ldots T(d_{t-1}, \theta_t + \mu v_t), \theta_t + \mu v_t), \ldots, \theta_t + \mu v_t)$;
    // Step 2: Update
    Oracle reveals $\hat{g}_t = \frac{2}{\mu N} \sum_{k=1}^N \ell(z_k(\theta_t + \mu v_t), \theta_t + \mu v_t) v_t$;
    Update $\theta_{t+1} = \text{Proj}_{(1-\tau)\Theta}(\theta_t - \eta \hat{g}_t)$;
end
```

**Example 1** (Location-Scale Distributions). We say $D(\theta)$ belongs to the family of location-scale distributions if

$$z(\theta) \sim D(\theta) \iff z(\theta) = z_0 + \zeta_0 + \zeta^T \theta,$$

where the location depends on $\theta$, the base random variable $z_0 \sim D_0$ is a sample from a zero-mean distribution, and $\zeta$ is norm bounded. If $d_0$ is additionally a location-scale distribution, then $d_t$ is a location-scale family for all $t$, such that $\ell(z(\theta), \theta)$ is $G$-Lipschitz with $G = (1-\delta^n)G_z ||\zeta|| + G_\theta$.

This class encompasses a broad set of distributions that are commonplace in the performative prediction literature. This class of distributions is also $\epsilon$-sensitive and satisfies the mixture dominance condition when $\ell$ is convex, observations that appeared in Miller et al. [2021].

4 Sample Complexity Analysis

In this section, we present our main theoretical results on sample complexity for Algorithm 1 and 2. In the analysis of both algorithms, there are two key sources of additional sample complexity as compared to classical stochastic gradient descent with an unbiased estimator or even derivative free methods with an unbiased estimator of the smoothed loss. Firstly, due to the fact that players are following the an estimate of the gradient constructed from sampling, the sample complexity includes additional terms due to sampling bias. Secondly, since the environment is non-stationary and reacts to deployed actions $\theta$, there is an additional sampling complexity term dependent on the epoch horizon $n$ which captures the fact that the gradient information the agent is using to update its action is computed at $d_t$ as opposed to the fixed point distribution $D(\theta_t)$.

4.1 First Order Oracle

In this setting, we assume access to an oracle that reveals the empirical gradient of the performative risk at the current distribution—i.e.,

$$\hat{g}_t = \frac{1}{N} \sum_{k=1}^N \nabla \ell(z_k(\theta_t), \theta_t)$$

where $z_k(\theta_t)$ is the $k$-th sample from $d_t$. To analyze the performance of Algorithm 1, there are two crucial steps:

1. Obtain a bound on the difference between the gradient of the expected performative risk at the current distribution $d_t$ and the gradient of the expected risk at $D(\theta_t)$ (Lemma 1).

2. Obtain a bound on the difference between the empirical gradient and the expected gradient (Lemma 5, Appendix A).

We leave the proof of the latter to the appendix since it is a modification of Theorem 1 in Mei et al. [2016] to apply to norm-subGaussian random vectors.
Lemma 1. Under Assumptions 1–4, the gradient error satisfies
\[ \| \nabla E_{z(\theta_0) \sim d_t}[\ell(z(\theta_0), \theta_t)] - \nabla E_{z(\theta_0) \sim D(\theta_0)}[\ell(z(\theta_0), \theta_t)] \| \leq \delta^i \tilde{\beta} W_1(d_0, D(\theta_1)) + \epsilon(G_\theta + G_z) \frac{\beta \rho \delta^n}{1 - \delta^n} + \delta^n G_z \epsilon, \]
where \( d_t = \delta^n d_{t-1} + (1 - \delta^n) D(\theta_t). \)

We defer the proof to Appendix B.1. The first term \( \delta^i \tilde{\beta} W_1(d_0, D(\theta_1)) \) captures how the dependence on the initial decision \( \theta_1 \) in the gradient error exponentially decays to zero as \( t \) approaches infinity. The second term \( \beta \rho \delta^n (G_\theta + G_z) \delta^n / (1 - \delta^n) \) captures the difference between subsequent decisions of Algorithm 1. The third term \( \delta^n G_z \epsilon \) captures the effect of the decision on the loss function. Observe that as the epoch length \( n \) becomes large, the bound approaches zero, which is equivalent to the static setting—i.e., there is no past state dependence (equivalently, \( \delta = 0 \)), and hence, the gradient error is zero.

To bound the difference between the empirical gradient and the expected gradient we apply Lemma 5 (Appendix A) to the empirical loss
\[ L_i^{(N)}(\theta_t) = \frac{1}{N} \sum_{k=1}^{N} \ell(z_k(\theta_t), \theta_t) \]
and the expected loss
\[ L_t(\theta_t) = \mathbb{E}_{z(\theta_0) \sim d_t}[\ell(z(\theta_0), \theta_t)] \]
where the expectation is evaluated at iteration \( d_t \). Indeed, Assumptions 3 and 6 are enough so that the assumptions of Lemma 5 hold. Hence, with probability \( 1 - \rho \), we have that
\[ \sup_{\theta \in \Theta} \| \nabla L_i^{(N)}(\theta) - \nabla L_i(\theta) \| \leq \mathfrak{G} \left( \frac{C q \log N}{N} \right)^{\frac{1}{2}} \]
for \( N \geq \Omega(q \log q) \) and \( C \) depends on \( \rho \) as detailed in the following theorem which characterizes the sample complexity of Algorithm 1 given the two gradient error bounds.

Theorem 2. Suppose that Assumptions 1–6 hold. Fix arbitrary \( \xi > 0 \) and let \( \eta = 1/M \) and \( n \geq \log(\xi \alpha / \mathcal{C}) / \log(\delta) \) where \( \mathcal{C} = \beta W_1(D(\theta_1), d_0) + \epsilon (G_\theta + G_z) / M + \epsilon G_z + \xi \alpha \). There exists a universal constant \( C_0 \) such that with probability \( 1 - \rho \), Algorithm 1 returns iterates that satisfy
\[ \| \theta_{T+1} - \theta^* \| \leq \left( 1 - \frac{\alpha}{M} \right)^T \| \theta_1 - \theta^* \| + \xi + \left( \frac{\log \left( \frac{1}{\rho} \right) q \log N}{N} \right)^{\frac{1}{2}} \]
for \( N \geq C q \log q \) where \( C = C_0 \max\{c_h, \log(RG/\rho), 1\} \) with
\[ c_h = \frac{1}{\log q} \max \left\{ \log \left( \frac{4M}{G^2} \right), \log \left( \frac{8H}{G^3} \right) \right\} . \]

We defer the proof to Appendix B.2. The idea is to massage the problem into stochastic gradient descent with biased gradients where there are two sources of bias. While the size of the error \( \xi > 0 \) can be arbitrarily chosen, the lower bound on the number of samples \( n \) depends logarithmically on the choice of \( \xi \) and hence, there is a tradeoff between the error tolerance and epoch size. If access to an expected gradient oracle is available, then the convergence rate is analogous to Ajalloeian and Stich [2020, Theorem 6] since the batch-size \( N \) dependent term drops out.

### 4.2 Zero Order Oracle

In this setting, the decision-maker has access only to the empirical loss at the current distribution. This is a more realistic setting given that the form of the data distribution \( d_t \)—and more specifically, \( D(\cdot) \)—may be a priori unknown. For example, if the data is generated by strategic data sources having their own private utility functions and preferences, then the decision-maker does not necessarily have access to the distribution map \( D(\theta) \) in practice.
A decision-maker using Algorithm 2 updates with gradient estimates

$$\hat{g}_t = \frac{q}{\mu} \frac{1}{N} \sum_{k=1}^{N} \ell(z_k(\theta_t + \mu v_t), \theta_t + \mu v_t)v_t,$$

where \(v_t\) is a unit vector and \(z_k(\theta_t + \mu v_t)\) is the \(k\)-th sample from \(d_t\). This is a one-point gradient estimate of the expected loss at \(d_t\) (cf. Flaxman et al. [2004], Spall [1997]). As is well known, for a given function \(f : \mathbb{R}^d \rightarrow \mathbb{R}\) and query radius \(\mu > 0\),

$$\mathbb{E}_{\nu \sim \mathcal{B}}[f(x + \mu \nu)] = \frac{\mu}{q} \nabla \hat{f}(x),$$

where \(\hat{f}(x) = \mathbb{E}_{\nu \sim \mathcal{B}}[f(x + \mu \nu)]\) and \(\mathcal{B}\) and \(\mathcal{S}\) are the Euclidean unit ball and unit sphere, respectively, in dimension \(q\).

The zeroth order algorithm projects onto \((1 - \tau)\Theta\) for \(\tau \in (0, 1)\)—namely,

$$\theta_{t+1} = \text{Proj}_{(1 - \tau)\Theta}(\theta_t - \eta \hat{g}_t)$$

where \((1 - \tau)\Theta\) is shorthand for \(\{(1 - \tau)\theta : \theta \in \Theta\}\). The projection is made onto this set so that the random query points around \(\theta_t\) belong to \(\Theta\). In particular, for any \(\theta \in (1 - \tau)\Theta\) and any unit vector \(v\), it holds that \(\theta + \mu v \in \Theta\) for any \(\mu \in [0, \tau\eta]\) [Flaxman et al., 2004, Observation 2].

In the zero order setting, we also need the loss to be uniformly bounded.

**Assumption 7.** The absolute value of the loss function \(\ell(z, \theta)\) is uniformly bounded—i.e., for all \((z, \theta) \in \mathcal{Z} \times \Theta\) pairs, the absolute value of the loss function satisfies \(|\ell(z, \theta)| \leq \ell_*\).

Analogous to the analysis of Algorithm 1, in order to analyze Algorithm 2, we bound the difference in the expected gradients at \(z \sim d_t = \delta^n d_{t-1} + (1 - \delta^n)\mathcal{D}(\theta_t + \mu v_t)\) and \(z \sim \mathcal{D}(\theta_t + \mu v)\) (Lemma 2), and we bound the difference between the empirical gradient at \(d_t\) of the smoothed loss and the gradient of the smoothed loss at \(d_t\) (Lemma 5, Appendix A). Define the smoothed expected risk as follows:

$$\hat{L}(\theta) = \mathbb{E}_{\nu \sim \mathcal{B}}[\mathbb{E}_{z(\theta + \mu \nu) \sim \mathcal{D}(\theta + \mu v)}[\ell(z(\theta + \mu v), \theta + \mu v)]]$$

**Lemma 2.** Under Assumptions 1, 2, and 7, the error between the gradient smoothed loss at \(d_t\) and the gradient of the smoothed risk at the fixed point \(\mathcal{D}\) satisfies

$$||\nabla \mathbb{E}_{\nu \sim \mathcal{B}}[\mathbb{E}_{z(\theta + \mu \nu) \sim d_t}[\ell(z(\theta + \mu v), \theta + \mu v)]] - \nabla \hat{L}(\theta)|| \leq G_z \left(\delta^n W(d_0) + 2\delta^n \mu \epsilon + \ell_* \frac{C_q}{\mu} \frac{\delta^n}{1 - \delta^n}\right)$$

where \(d_t = \delta^n d_{t-1} + (1 - \delta^n)\mathcal{D}(\theta_t + \mu v_t)\), and \(W(d_0) = \max_{\theta \in \Theta} W_1(d_0, \mathcal{D}(\theta))\).

We defer the proof to Appendix C.1. Analogous to the first order setting, we apply Lemma 5 to get a bound on the error between the empirical gradient of the smoothed loss

$$\hat{L}_t^{(N)}(\theta_t) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}_{\nu \sim \mathcal{B}}[\ell(z(\theta_t + \mu v), \theta_t + \mu v)]$$

and the gradient of the smoothed loss

$$\hat{L}_t(\theta_t) = \mathbb{E}_{\nu \sim \mathcal{B}}[\mathbb{E}_{z(\theta + \mu \nu) \sim d_t}[\ell(z(\theta + \mu v), \theta + \mu v)]]$$

at the current distribution \(d_t\). Following a similar reasoning as in the preceding subsection, we have that with probability \(1 - \rho\), the error \(||\nabla \hat{L}_t(\theta) - \nabla \hat{L}_t^{(N)}(\theta)||\) uniformly converges—i.e.,

$$\sup_{\theta \in \Theta} ||\nabla \hat{L}_t(\theta) - \nabla \hat{L}_t^{(N)}(\theta)|| \leq \frac{G}{2} \sqrt{\frac{C_q \log N}{N}}$$

for \(N \geq \Omega(q \log q)\) and where \(C\) depends on \(\rho\) as detailed in the following theorem which characterizes the sample complexity of Algorithm 2. For a fixed time \(T\), let \(\mathcal{E}_T\) be the event that (3) for all \(t \in \{0, \ldots, T\}\).
Theorem 3. Suppose that Assumptions 1-7 hold. There exists a universal constant $C_0$ such that with probability $1 - \rho$, Algorithm 2 with $n \geq \Omega(\log \eta)$, $\eta = \eta_0/T$ with $\eta_0 > 1/(3\alpha)$, $\tau = \mu/r$ and $\mu = (q^2 \ell^2 \eta/(2M R))^{1/3}$, returns iterates satisfying

$$
E[\frac{1}{2}\|\theta_T - \theta^*\|^2|\mathcal{E}_T] \leq O \left( T^{-1/3} + \left( \log \left( \frac{1}{\rho} \right) \frac{q \log N}{N} \right)^{1/2} \right)
$$

for $N \geq Cq \log q$, $C = \max\{c_b, \log(RG/\rho), 1\}$ and

$$
c_b = \frac{1}{\log q} \max \left\{ \log \left( \frac{4M}{G^2} \right), \log \left( \frac{8H}{G^3} \right) \right\}.
$$

The proof is deferred to Appendix C.2. Unlike the standard constrained zero order setting for which it has been shown that the approach of Flaxman et al. [2004] obtains a $O(T^{-1/3})$ rate [Agarwal et al., 2010],\textsuperscript{1} we have an additional error term due to the environment dynamics that results in having to bound the error between the empirical and expected gradients at $d_t$ to their counterparts at $\mathcal{D}(\theta)$—i.e., the distribution which defines the performatively optimal point (cf. Definition 1).

5 Numerical Experiments

In this section, we apply our aforementioned algorithms to a semi-synthetic example based on real data from the dynamic pricing experiment—namely, SFpark\textsuperscript{2}—for on-street parking in San Francisco. Parking availability, location, and price are some of the most important factors when people choose whether or not to use a personal vehicle to make a trip [Shoup, 2006, 2021, Fiez and Ratliff, 2020]. The primary goal of the SFpark pilot project was to make it easy to find a parking space. To this end, SFpark targeted a range of 60–80% occupancy in order to ensure some availability at any given time, and devised a controlled experiment for demand responsive pricing. Meter operational hours are split into distinct rate periods, and rates are adjusted on a block-by-block basis, using occupancy data from parking sensors in on-street parking spaces in the pilot areas. We focus on weekdays in the numerical experiments; for weekdays, distinct rate periods are 900–1200, 1200–1500, and 1500–1800. Excluding special events, SFpark adjusted hourly rates as follows: a) 80–100% occupancy, rates are increased by $\theta$; b) 60–80% occupancy, no adjustment is made; c) 30–60% occupancy, rate is decreased by $\theta/3$; d) occupancy below 30%, rate is decreased by $\theta/50$. When a price change is deployed it takes time for users to become aware of the price change through signage and mobile payment apps [Pierce and Shoup, 2013].

Given the target occupancy, the dynamic decision-dependent loss (or performative risk) is given by

$$
\mathbb{E}_{z(\theta) \sim d_t} \left[ (\ell(z(\theta), \theta)) \right] = \mathbb{E}_{z(\theta) \sim d_t} \left[ (z(\theta) - 0.7)^2 + \frac{\nu}{2} \theta^2 \right],
$$

for each block, where $z(\theta)$ is the occupancy (which is between zero and one), $\theta$ is the change in price from the nominal price at the beginning of the SFpark study, and $\nu$ is the regularization parameter. For the initial distribution $d_0$, we sample from the data at the beginning of the pilot study where the price is at the nominal (or initial) price. The distribution $\mathcal{D}(\theta)$ is defined as follows:

$$
z \sim \mathcal{D}(\theta) \iff z = z_0 + m \theta
$$

where $z_0$ follows the same distribution as $d_0$ described above, and $m$ is a proxy for the price elasticity which is estimated by fitting a line to the final and initial occupancy and price (cf. Appendix D.1).\textsuperscript{3}

Comparing Performative Optimum to SFpark. We run Algorithms 1 and 2 (using parameters as dictated by Theorems 2 and 3, respectively) for Beach ST 600, a representative block in the Fisherman’s Wharf sub-area, in the time window of 1200–1500 as depicted in Figure 1. Beach ST is frequently visited by

\textsuperscript{1}The rate is $O(T^{-1/2})$ in unconstrained settings.

\textsuperscript{2}SFpark: tinyurl.com/dwtf7wwn

\textsuperscript{3}Price elasticity is the change in percentage occupancy for a given percentage change in price.
Figure 1: Results of Algorithm 1 (first and third plots) and Algorithm 2 (second and fourth plots) with different \((n,T)\) pairs for 600 Beach ST and time window 1200–1500. Each marker represents a price announcement, and the plots show the prices and corresponding predicted occupancies. The SFpark prices and occupancies are far from the target and performative optimal price, whereas the proposed algorithms obtain both points up to theoretical error bounds.

Figure 2: Final prices announced by first and zero order algorithms (Algorithms 1 and 2) run with \((n,T) = (8,15)\) and \((n,T) = (1,120)\), respectively, as compared to SFpark for streets depicted in the right graphic (color coded to the bar charts) during the 900–1200 time period. The center plot shows the corresponding predicted occupancies. The dotted lines represent performatively optimal price and target occupancy of 70%, in the left and center plots, respectively. The average price overall is lower for both proposed methods, the occupancy is better distributed, and the average occupancy closer to the desired range.

Figure 2: Final prices announced by first and zero order algorithms (Algorithms 1 and 2) run with \((n,T) = (8,15)\) and \((n,T) = (1,120)\), respectively, as compared to SFpark for streets depicted in the right graphic (color coded to the bar charts) during the 900–1200 time period. The center plot shows the corresponding predicted occupancies. The dotted lines represent performatively optimal price and target occupancy of 70%, in the left and center plots, respectively. The average price overall is lower for both proposed methods, the occupancy is better distributed, and the average occupancy closer to the desired range.

Moreover, the prices under the performatively optimal solution obtained by the proposed algorithms are lower than the SFpark solution for the entire trajectory, and the algorithms both reach the target occupancy while SFpark is far from it. The third and fourth plots of Figure 1 show the effect of the negative price...
elasticity on the occupancy; an increased price causes a decreased occupancy. An interesting observation is that for Algorithm 1, a larger choice of \( n \), and consequently a smaller choice of \( T \), allows for convergence closer to the performatively optimal price, but for Algorithm 2, a smaller choice of \( n \), and consequently, a larger choice of \( T \), allows for quicker (and with lower variance) convergence to the performatively optimal price. This is due to the randomness in the query direction for the gradient estimator used in Algorithm 2, meaning that a larger \( T \) is needed to converge quickly to the optimal solution. This suggests that in the more realistic case of zero order feedback, the institution should make more price announcements.

**Redistributing Parking Demand.** In this semi-synthetic experiment, we set \( \nu = 1e-3 \) and take \( \Theta = [-3.5, 4.5] \) since the base distribution for these blocks has a nominal price of $3.50. We also use the estimated \( \delta \) and \( m \) values (described in more detail in Appendix D.3). We run Algorithms 1 and 2 (using parameters as dictated by the corresponding sample complexity theorems) for a collection of blocks during the time period 900–1200 in a highly mixed use area (i.e., with tourist attractions, a residential building, restaurants and other businesses). The results are depicted in Figure 2.

Hawthorne ST 0 is a very high demand street; the occupancy is around 90% on average during the initial distribution and remains high for SFpark (cf. center, Figure 2). The performatively optimal point, on the other hand, reduces this occupancy to within the target range 60–80% for both the first and zeroth order methods. This occupancy can be seen as being redistributed to the Folsom ST 500–600 block, as depicted in Figure 2 (center) for our proposed methods: the SFpark occupancy is much below the 70% target average for these blocks, while both the decision-dependent algorithms lead to occupancy at the target average. Interestingly, this also comes at a lower price (not just on average, but for each block) than SFpark.

Hawthorne ST 100 is an interesting case in which both our approach and SFpark do not perform well. This is because the performatively optimal price in the unconstrained case is $9.50 an hour which is well above the maximum price of $8 in the constrained setting we consider. In addition, the price elasticity is positive for this block; together these facts explain the low occupancy. Potentially other control knobs available to SFpark, such as time limits, can be used in conjunction with price to manage occupancy; this is an interesting direction of future work.

6 Discussion and Future Directions

This work is an important step in understanding performative prediction in dynamic environments. Moving forward there are a number of interesting future directions. We consider a single class of dynamics in this paper. Another practically motivated class of dynamics are period dynamics; indeed, in many applications there is an external context which evolves periodically such as seasonality or other temporal effects. Devising algorithms for such cases is an interesting direction of future work. As compared to classical reinforcement learning problems, in this work, we exploit the structure of the dynamics along with convexity to devise convergent algorithms. However, we only considered general conditions on the class of distributions \( \mathcal{D}(\theta) \); it may be possible to exploit additional structure on \( \mathcal{D}(\theta) \) in improving the sample complexity of the proposed algorithms or devising more appropriate algorithms that leverage this structure. For instance, particular applications such as strategic classification (where agents or populations of agents have memory) may lend themselves to more precise modeling of the decision dependence in \( \mathcal{D}(\theta) \). This is another interesting direction of future work.

References


A Technical Lemmas and Notation

Notation. Throughout we will use the following derivative and partial derivative notation. For a given function \( \ell(z, \theta) \), the partial derivative of \( \ell \) with respect to \( z \) is denoted \( \nabla_z \ell(z, \theta) \) and the partial derivative with respect to \( \theta \) is denoted \( \nabla_\theta \ell(z, \theta) \). For the expected risk \( \mathbb{E}_{z \sim \mathcal{D}(\theta)}[\ell(z, \theta)] \), the total derivative with respect to \( \theta \) is denoted

\[
\nabla \mathbb{E}_{z \sim \mathcal{D}(\theta)}[\ell(z, \theta)] = \nabla \left( \int \ell(z, \theta) p_\theta(z) dz \right) = \mathbb{E}_{z \sim \mathcal{D}(\theta)}[\nabla_\theta \ell(z, \theta)] + \mathbb{E}_{z \sim \mathcal{D}(\theta)}[\ell(z, \theta) \nabla_\theta \log(p_\theta(z))]
\]

where \( p_\theta(z) \) is the density function for \( \mathcal{D}(\theta) \) and in the last equality we have applied the so-called ‘log trick’. Throughout, we use the notation \( \| \cdot \| \) for the Euclidean norm.

The following lemma is a direct consequence of dual form of the Wasserstein-1 distance.

Lemma 3. Let \( f : \mathbb{R}^n \to \mathbb{R}^n \) be \( \beta \)-Lipschitz, and let \( X, X' \) be random vectors with distributions \( p \) and \( p' \), respectively. Then,

\[
\| \mathbb{E}[f(X)] - \mathbb{E}[f(X')] \| \leq \beta W_1(p, p').
\]

Since it’s used in a few places, the following straightforward result is included as a lemma for ease of reference.

Lemma 4. For \( 0 \leq x \leq 1 \), and for any \( t > 0 \),

\[
\sum_{i=1}^{t-1} x^{i-1} \leq \frac{1}{1-x} \quad \text{and} \quad \sum_{i=1}^{t-1} x^{i-1} i \leq \left( \frac{1}{1-x} \right)^2.
\]

Technical Lemma on Uniform Convergence. For both the first and zero order information settings, we need a bound on the uniform convergence of the error in the empirical and expected gradient information at the current distribution \( \mathcal{D}_t \). Towards obtaining this result, we appeal to Theorem 1 in Mei et al. [2016] with a few slight modifications to Step 2 of the proof which we state below as a lemma.

Since we apply the lemma to two different settings, we state the lemma for a general expected risk \( R(\theta) \) and empirical risk \( R^{(N)}(\theta) \) which are defined by

\[
R(\theta) = \mathbb{E}[\ell(z(\theta), \theta)]
\]

and

\[
R^{(N)}(\theta) = \frac{1}{N} \sum_{k=1}^{N} \ell(z_k(\theta), \theta)
\]

respectively, where \( z_k(\theta) \) denotes the \( k \)-th sample from the decision dependent distribution. When we apply the lemma, the distribution will be \( \mathcal{D}_t \). Following Jin et al. [2019], we say a random vector \( Z \in \mathbb{R}^q \) is norm-sub-Gaussian (or nSG(\( \phi^2 \))) if there exists \( \phi \) such that

\[
\mathbb{P}(\|Z - \mathbb{E}[Z]\| \geq t) \leq 2 \exp(-t^2/(2\phi^2)), \quad \forall \ t \in \mathbb{R}.
\]

Assumption 8. The loss gradient of \( \ell(z(\theta), \theta) \) is nSG(\( \phi^2 \)).

Assumption 9. The loss \( \ell(z(\theta), \theta) \) is \( H \)-Hessian Lipschitz and the expected risk \( R(\theta) \) is \( M \)-smooth.

Recall that \( \Theta \subset \mathbb{B} \) where \( \mathbb{B} \) is the unit ball in dimension \( q \).

Lemma 5. Under Assumptions 8 and 9, there exists a universal constant \( C_0 \) such that the sample gradient \( \nabla R^{(N)} \) converges uniformly to the population gradient \( \nabla R \) in Euclidean norm—namely, if \( N \geq Cq \log q \), we have that

\[
\mathbb{P} \left( \sup_{\theta \in \Theta} \| \nabla R^{(N)}(\theta) - \nabla R(\theta) \| \leq \phi \sqrt{\frac{Cq \log N}{N}} \right) \geq 1 - \rho
\]
where \( C = C_0 \max \{ c_h, \log(\rho/\phi), 1 \} \) with
\[
c_h = \frac{1}{\log q} \max \left\{ \log \left( \frac{4M}{\phi^2} \right), \log \left( \frac{8H}{\phi^3} \right) \right\}.
\]

Proof. We proceed in a similar manner to the proof of Theorem 1 from Mei et al. [2016] where we simply modify Step 2. For completeness, we write the full proof of each of the steps.

**Step 1: Decompose the ‘bad’ events using \( \varepsilon \)-nets.** Let \( \mathcal{N}_\varepsilon \) be the \( \varepsilon \)-covering of \( \Theta \). Following arguments in Vershynin [2018, Section 4], we know that \( \mathcal{N}_\varepsilon \leq |\Theta + \frac{\varepsilon}{2} B^q|/\frac{\varepsilon}{2} B^q \) where \(| \cdot |\) denotes the volume of its argument and \( B^q \) is the unit Euclidean ball.\(^4\) We can further upper bound this by \( |(R + \frac{\varepsilon}{2}) B^q|/\frac{\varepsilon}{2} B^q | \) so that \( \log \mathcal{N}_\varepsilon \leq q \log(3R/\varepsilon) \). Let \( \Theta_\varepsilon = \{ \theta_1, \ldots, \theta_N \} \) be a corresponding \( \varepsilon \)-cover with \( m = \mathcal{N}_\varepsilon \) elements. For any \( \theta \in \Theta \), let \( j(\theta) = \arg \min_{j \in [m]} \| \theta - \theta_j \| \). Then \( \| \theta - \theta_{j(\theta)} \| \leq \varepsilon \) for all \( \theta \in \Theta \).

For any \( \theta \in \Theta \), we have
\[
\| \nabla R^{(N)}(\theta) - \nabla R(\theta) \| \leq \left\| \frac{1}{N} \sum_{k=1}^N \nabla \ell(z_k(\theta), \theta) - \nabla \ell(z_k(\theta_{j(\theta)}), \theta_{j(\theta)}) \right\| \\
+ \left\| \frac{1}{N} \sum_{k=1}^N \nabla \ell(z_k(\theta_{j(\theta)}), \theta_{j(\theta)}) - E[\nabla \ell(z(\theta_{j(\theta)}), \theta_{j(\theta)})] \right\| \\
+ \left\| E[\nabla \ell(z(\theta_{j(\theta)}), \theta_{j(\theta)})] - E[\nabla \ell(z(\theta), \theta)] \right\|.
\]
Hence,
\[
\Pr \left( \sup_{\theta \in \Theta} \| \nabla R^{(N)}(\theta) - \nabla R(\theta) \| \geq t \right) \leq \Pr(A_t) + \Pr(B_t) + \Pr(C_t), \tag{4}
\]
where the events are defined as
\[
A_t = \left\{ \sup_{\theta \in \Theta} \left\| \frac{1}{N} \sum_{k=1}^N \nabla \ell(z_k(\theta), \theta) - \nabla \ell(z_k(\theta_{j(\theta)}), \theta_{j(\theta)}) \right\| \geq \frac{t}{3} \right\},
\]
\[
B_t = \left\{ \sup_{\theta \in \Theta} \left\| \frac{1}{N} \sum_{k=1}^N \nabla \ell(z_k(\theta_{j(\theta)}), \theta_{j(\theta)}) - E[\nabla \ell(z(\theta_{j(\theta)}), \theta_{j(\theta)})] \right\| \geq \frac{t}{3} \right\},
\]
and
\[
C_t = \left\{ \sup_{\theta \in \Theta} \left\| E[\nabla \ell(z(\theta_{j(\theta)}), \theta_{j(\theta)})] - E[\nabla \ell(z(\theta), \theta)] \right\| \geq \frac{t}{3} \right\}.
\]
Next, we bound the probability of each of these events starting with \( B_t \) since this is the main part of the proof of Theorem 1 of Mei et al. [2016] that needs to be altered.

**Step 2: Upper bound on \( \Pr(B_t) \).** The triangle inequality implies that
\[
\left\| \frac{1}{N} \sum_{k=1}^N \nabla \ell(z_k(\theta_{j(\theta)}), \theta_{j(\theta)}) - E[\nabla \ell(z(\theta_{j(\theta)}), \theta_{j(\theta)})] \right\| = \frac{1}{N} \left\| \sum_{k=1}^N (\nabla \ell(z_k(\theta_{j(\theta)}), \theta_{j(\theta)}) - E[\nabla \ell(z(\theta_{j(\theta)}), \theta_{j(\theta)})]) \right\| \\
\leq \frac{1}{N} \sum_{k=1}^N \| \nabla \ell(z_k(\theta_{j(\theta)}), \theta_{j(\theta)}) - E[\nabla \ell(z(\theta_{j(\theta)}), \theta_{j(\theta)})] \|.
\]

\(^4\)In contrast to other sections of the paper, here we explicitly indicate the dimension of the ball in the superscript for clarity.
Taking a union bound over $\Theta_{\varepsilon}$ yields

$$\Pr(B_t) \leq \Pr \left( \sup_{j \in [m]} \left\{ \frac{1}{N} \sum_{k=1}^{n} \left\| \nabla \ell(z_k(\theta_{j(\theta)}), \theta_{j(\theta)}) - E[\nabla \ell(z(\theta_{j(\theta)}), \theta_{j(\theta)})] \right\| \geq \frac{t}{3} \right\} \right)$$

$$\leq e^{n \log(3R/\varepsilon)} \sup_{j \in [m]} \Pr \left( \frac{1}{N} \sum_{k=1}^{n} \left\| \nabla \ell(z_k(\theta_{j(\theta)}), \theta_{j(\theta)}) - E[\nabla \ell(z(\theta_{j(\theta)}), \theta_{j(\theta)})] \right\| \geq \frac{t}{3} \right)$$

Since $\|\nabla \ell(z(\theta), \theta) - E[\nabla \ell(z(\theta), \theta)]\|$ is $\phi^2$ norm sub-Gaussian, we have that

$$\frac{1}{N} \sum_{k=1}^{n} \left\| \nabla \ell(z_k(\theta_{j(\theta)}), \theta_{j(\theta)}) - E[\nabla \ell(z(\theta_{j(\theta)}), \theta_{j(\theta)})] \right\|$$

is $(\phi^2/N)$ norm sub-Gaussian. Hence,

$$\Pr \left( \frac{1}{N} \sum_{k=1}^{n} \left\| \nabla \ell(z_k(\theta_{j(\theta)}), \theta_{j(\theta)}) - E[\nabla \ell(z(\theta_{j(\theta)}), \theta_{j(\theta)})] \right\| \geq \frac{t}{3} \right) \leq e^{-\frac{N\phi^2 t^2}{36q^2}}$$

so that

$$\Pr(B_t) \leq \exp \left( - \frac{Nt^2}{36\phi^2} + q \log \left( \frac{3R}{\varepsilon} \right) \right).$$

Thus,

$$t > \sqrt{\frac{(\log(2/\rho) + q \log(3R/\varepsilon))36\phi^2}{N}}$$

ensures $\Pr(B_t) \leq \rho/2$.

**Step 3: Upper bound on $\Pr(A_t)$ and $\Pr(C_t)$.** Note that $C_t$ is a deterministic event and we have shown above that we satisfy the requisite assumptions for Theorem 1 of Mei et al. [2016]; hence, the same argument applies to show that $C_t$ will never occur. To bound the probability of event $A_t$ we also apply the same argument as in Theorem 1 of Mei et al. [2016] to show that $\Pr(A_t) \leq \rho/2$ when $t \geq 6\varepsilon D_*/\rho$ where

$$\mathbb{E} \left[ \sup_{\theta \in \Theta} \|\nabla^2 R(\theta)\|_{\text{op}} \right] \leq D_* = 2RH + M.$$

Indeed, Markov’s inequality implies that

$$\Pr(A_t) = \Pr \left( \sup_{\theta \in \Theta} \left\| \frac{1}{N} \sum_{k=1}^{N} \nabla \ell(z_k(\theta), \theta) - \nabla \ell(z_k(\theta_{j(\theta)}), \theta_{j(\theta)}) \right\| \geq \frac{t}{3} \right)$$

$$\leq \frac{3}{t} \mathbb{E} \left[ \sup_{\theta \in \Theta} \left\| \frac{1}{N} \sum_{k=1}^{N} \nabla \ell(z_k(\theta), \theta) - \nabla \ell(z_k(\theta_{j(\theta)}), \theta_{j(\theta)}) \right\| \right]$$

$$\leq \frac{3\varepsilon}{t} \mathbb{E} \left[ \sup_{\theta \in \Theta} \|\nabla^2 R(\theta)\|_{\text{op}} \right]$$

$$\leq \frac{3\varepsilon D_*}{t}.$$

**Step 4: Concluding the proof.** This step also follows exactly the same logic of Step 4 in the proof of Theorem 1 of Mei et al. [2016], however, with different constants due to the different bound obtained for $\Pr(B_t)$.

To ensure the probability of the bad event (which we upper bounded with the three other bad events $A_t$, $B_t$ and $C_t$) to be less than $\rho$, it is sufficient to take

$$\varepsilon = \frac{\rho\phi}{6(M + 2RH)Nq}.$$
This completes the proof.

\[ t \geq \max \left\{ \frac{\phi}{Nq}, 3\phi^2 \left( \frac{2}{3} + q \log \left( \frac{36R(M + 2RH)Nq}{\rho^2} \right) \right) \right\}. \]

Since \( M \leq \phi^2 q^k \) and \( H \leq \phi^3 q^k \), there exists a universal constant \( C_0 \) such that

\[ \Pr \left( \sup_{\theta \in \Theta} \| \nabla R^{(N)}(\theta) - \nabla R(\theta) \| \geq \phi \sqrt{\frac{Cq \log N}{N}} \right) \leq \rho \]

as long as \( N \geq Cq \log q \) where

\[ C = C_0 \max\{c_h, \log(R\phi/\rho), 1\}. \]

This completes the proof.

\[ \square \]

B Proofs for First Order Oracle Setting

B.1 Proof of Lemma 1

For convenience, we restate Lemma 1 here.

Lemma 1. Under Assumptions 1–4, the gradient error satisfies

\[ \| \nabla \mathbb{E}_{z(\theta_t) \sim d_t}[\ell(z(\theta_t), \theta_t)] - \nabla \mathbb{E}_{z(\theta_t) \sim D(\theta_t)}[\ell(z(\theta_t), \theta_t)] \| \leq \delta^n \beta \mathcal{W}_1(d_0, D(\theta_1)) + \epsilon (G_{\theta} + G_{\theta} \epsilon) \beta n \delta^n + \delta^n G_{\theta} \epsilon, \]

where \( d_t = \delta^n d_{t-1} + (1 - \delta^n) D(\theta_t) \).

Proof of Lemma 1. The organization of the proof is as follows. We first show that, for any \( \bar{\theta} \in \Theta \), the mapping \( \theta \mapsto \mathbb{E}_{z(\theta) \sim D(\theta)}[\ell(z(\theta), \bar{\theta})] \) is \( (G_{\theta} + G_{\theta} \epsilon) \)-Lipschitz. We then show that

\[ L_t(\theta) = \mathbb{E}_{z(\theta) \sim \mathbb{D}(\theta)}[\ell(z(\theta), \theta)] \]

is \( (G_{\theta} + G_{\theta} \epsilon) \)-Lipschitz. Using these facts, we construct the gradient error bound.

Proving Lipschitz bounds. We first show that \( \theta \mapsto \mathbb{E}_{z(\theta) \sim D(\theta)}[\ell(z(\theta), \bar{\theta})] \) is Lipschitz. Using Lemma 3, we have that

\[ \| \mathbb{E}_{z(\theta) \sim D(\theta)}[\ell(z(\theta), \bar{\theta})] - \mathbb{E}_{z(\theta') \sim D(\theta')}[\ell(z(\theta'), \bar{\theta})] \| \leq G_{\theta} \mathcal{W}_1(D(\theta), D(\theta')) \leq G_{\theta} \| \theta - \theta' \|. \] (5)

To show that \( L_t \) is Lipschitz, we expand the definition of \( L_t \) with a series of triangle inequalities. Indeed,

\[ |L_t(\theta) - L_t(\theta')| \leq |\mathbb{E}_{z \sim d_t}[\ell(z(\theta), \theta)] - \mathbb{E}_{z \sim d_t}[\ell(z(\theta'), \theta')]| \]

\[ \leq \delta^n |\mathbb{E}_{z \sim d_{t-1}}[\ell(z, \theta) - \ell(z, \theta')]| + (1 - \delta^n) |\mathbb{E}_{z \sim d_{t-1}}[\ell(z, \theta)] - \mathbb{E}_{z \sim d_{t-1}}[\ell(z, \theta')]| \]

\[ \leq \delta^n |\mathbb{E}_{z \sim d_{t-1}}[\ell(z, \theta) - \ell(z, \theta')]| + (1 - \delta^n) \left( |\mathbb{E}_{z(\theta) \sim D(\theta)}[\ell(z(\theta), \theta)] - \mathbb{E}_{z(\theta) \sim D(\theta)}[\ell(z(\theta), \theta')]| \right) \]

\[ + (1 - \delta^n) \left( |\mathbb{E}_{z(\theta') \sim D(\theta')}[\ell(z(\theta), \theta')] - \mathbb{E}_{z(\theta') \sim D(\theta')}[\ell(z(\theta'), \theta')]| \right). \]

For any distribution \( d \) and any \( \theta, \theta' \), we have

\[ \| \mathbb{E}_{z \sim d}[\ell(z, \theta) - \ell(z, \theta')] \| \leq \mathbb{E}_{z \sim d}[\| \ell(z, \theta) - \ell(z, \theta') \|] \leq G_{\theta} \| \theta - \theta' \|. \] (6)

Combining (5) and (6), we have that

\[ |L_t(\theta) - L_t(\theta')| \leq \delta^n G_{\theta} \| \theta - \theta' \| + (1 - \delta^n) G_{\theta} \epsilon \| \theta - \theta' \| \]

\[ \leq G_{\theta} \| \theta - \theta' \| + G_{\theta} \epsilon \| \theta - \theta' \| \]

\[ = (G_{\theta} + G_{\theta} \epsilon) \| \theta - \theta' \|, \]

which shows that \( L_t \) is \( (G_{\theta} + G_{\theta} \epsilon) \)-Lipschitz.
Proving the gradient error bound. Using the non-expansiveness of the projection operator, we have that

$$\|\theta_{t+1} - \theta_t\| = \|\text{Proj}_\Theta(\theta_t - \eta\nabla L_t(\theta_t)) - \text{Proj}_\Theta(\theta_t)\| \leq \eta\|\nabla L_t(\theta_t)\| \leq \eta(G_\Theta + G_\zeta\epsilon),$$

where the last inequality follows from the fact that Lipschitzness of $L_t$. Applying Lemma 3, for any $\theta_{t+1}, \theta_t \in \Theta$, where $\theta_{t+1} = \text{Proj}_\Theta(\theta_t - \eta\nabla L_t(\theta_t))$, we have that

$$\|E_{z(\theta_t) \sim D(\theta_t)}[\nabla \ell(z(\theta_t), \theta_t)] - E_{z(\theta_{t+1}) \sim D(\theta_{t+1})}[\nabla \ell(z(\theta_{t+1}), \theta_{t+1})]\| \leq \beta W_1(D(\theta_t), D(\theta_{t+1}))$$

$$\leq \beta\epsilon\|\theta_t - \theta_{t+1}\| \quad \text{(7)}$$

Applying the chain rule, we have that

$$\nabla E_{z(\theta) \sim D(\theta)}[\ell(z(\theta), \theta)] = E_{z(\theta) \sim D(\theta)}[\nabla \ell(z(\theta), \theta)] + E_{z(\theta) \sim D(\theta)}[\ell(z(\theta), \theta)\nabla \log(p_\theta(z(\theta)))].$$

(8)

Applying the relationship between the distribution at time $t$ and the distribution at time $t - 1$, i.e., $d_t = \delta^n d_{t-1} + (1 - \delta^n)D(\theta_t)$, and applying (8) with the triangle inequality we have that

$$\|\nabla E_{z(\theta_t) \sim d_t}[\ell(z(\theta_t), \theta_t)] - \nabla E_{z(\theta_t) \sim D(\theta_t)}[\ell(z(\theta_t), \theta_t)]\|$$

$$= \|\delta^n\nabla E_{z(\theta_t) \sim d_{t-1}}[\ell(z, \theta_t)] + (1 - \delta^n)\nabla E_{z(\theta_t) \sim D(\theta_t)}[\ell(z(\theta_t), \theta_t)] - \nabla E_{z(\theta_t) \sim D(\theta_t)}[\ell(z(\theta_t), \theta_t)]\|$$

$$= \delta^n\|\nabla E_{z(\theta_t) \sim d_{t-1}}[\ell(z, \theta_t)] - \nabla E_{z(\theta_t) \sim D(\theta_t)}[\ell(z(\theta_t), \theta_t)]\|$$

$$\leq \delta^n\|E_{z(\theta_t) \sim d_{t-1}}[\nabla \ell(z, \theta_t)] - E_{z(\theta_t) \sim D(\theta_t)}[\nabla \ell(z(\theta_t), \theta_t)]\| + \delta^n\|E_{z(\theta_t) \sim D(\theta_t)}[\ell(z(\theta_t), \theta_t)]\|. \quad \text{(9)}$$

Using (5), we have that

$$\|E_{z(\theta_t) \sim D(\theta_t)}[\ell(z(\theta_t), \theta_t)\nabla \log(p_\theta(z(\theta_t)))]\| \leq G_\zeta\epsilon,$$

and so it remains to bound

$$\|E_{z \sim d_{t-1}}[\nabla \ell(z, \theta_t)] - E_{z \sim D(\theta_t)}[\nabla \ell(z, \theta_t)]\|.$$  

Again, using the relationship between the current and previous distributions, and applying the triangle inequality, we have that

$$\|E_{z \sim d_{t-1}}[\nabla \ell(z, \theta_t)] - E_{z(\theta_t) \sim D(\theta_t)}[\nabla \ell(z(\theta_t), \theta_t)]\|$$

$$= \|\delta^n E_{z \sim d_{t-1}}[\nabla \ell(z, \theta_t)] + (1 - \delta^n)E_{z \sim D(\theta_t)}[\nabla \ell(z(\theta_t), \theta_t)] - E_{z(\theta_t) \sim D(\theta_t)}[\nabla \ell(z(\theta_t), \theta_t)]\|$$

$$\leq \delta^n\|E_{z \sim d_{t-1}}[\nabla \ell(z, \theta_t)] - E_{z \sim D(\theta_t)}[\nabla \ell(z, \theta_t)]\| + \|E_{z \sim D(\theta_t)}[\nabla \ell(z(\theta_t), \theta_t)] - E_{z(\theta_t) \sim D(\theta_t)}[\nabla \ell(z(\theta_t), \theta_t)]\|.$$  

Applying this bound recursively and applying (9), we obtain

$$\|\nabla E_{z \sim d_t}[\ell(z, \theta_t)] - \nabla E_{z(\theta_t) \sim D(\theta_t)}[\ell(z(\theta_t), \theta_t)]\|$$

$$\leq \delta^n\|E_{z \sim d_{t-1}}[\nabla \ell(z, \theta_t)] - E_{z(\theta_t) \sim D(\theta_t)}[\nabla \ell(z(\theta_t), \theta_t)]\| + \delta^n G_\zeta\epsilon$$

$$\leq \delta^{2n}\|E_{z \sim d_{t-2}}[\nabla \ell(z, \theta_t)] - E_{z \sim D(\theta_{t-1})}[\nabla \ell(z, \theta_t)]\| + \delta^n G_\zeta\epsilon$$

$$\leq \delta^{2n}\|E_{z \sim d_{t-2}}[\nabla \ell(z, \theta_t)] - E_{z \sim D(\theta_{t-1})}[\nabla \ell(z, \theta_t)]\| + \delta^n G_\zeta\epsilon,$$

where the final inequality follows from (7). Continuing the recursion, we have that

$$\|\nabla E_{z \sim d_t}[\ell(z, \theta_t)] - \nabla E_{z(\theta_t) \sim D(\theta_t)}[\ell(z(\theta_t), \theta_t)]\| \leq \delta^{tn}\beta W_1(d_0, D(\theta_1)) + \beta\epsilon(G_\Theta + G_\zeta\epsilon)\sum_{k=1}^{t-1} \delta^k \leq \delta^n + \delta^n G_\zeta\epsilon$$

$$\leq \delta^{tn}\beta W_1(d_0, D(\theta_1)) + \beta\epsilon(G_\Theta + G_\zeta\epsilon)\frac{\delta^n}{1 - \delta^n} + \delta^n G_\zeta\epsilon,$$

where the last inequality follows from the fact that $\sum_{k=1}^{t-1} \delta^k \leq \sum_{k=1}^{\infty} \delta^k = \frac{\delta^n}{1 - \delta^n}$. \qed
B.2 Proof of Theorem 2

Before proving Theorem 2, we present some results from prior work which we will use. Observing that in our setting we are simply performing project gradient descent, Proposition 3 from Schmidt et al. [2011] can be restated as follows.

**Proposition 1.** Suppose that $L(\theta)$ is $\alpha$-strongly convex and $M$-smooth, and that $L$ attains its minimum at some point $\theta^* \in \mathbb{R}^q$. Then, using the update rule of $\theta_t = \text{Proj}_\Omega(\theta_{t-1} - \eta(VL(\theta_{t-1}) + e_t))$, with $\eta = 1/M$, for all $t \geq 1$, we have

$$\|\theta_t - \theta^*\| \leq (1 - \alpha/M)^t (\|\theta_0 - \theta^*\| + \psi_t), \quad (10)$$

where

$$\psi_t = \frac{1}{M} \sum_{l=1}^t (1 - \alpha/M)^{-l} \|\epsilon_l\|, \quad (11)$$

and $\epsilon_l$ denotes the error in the calculation of the gradient.

To prove Theorem 2, we upper bound $\psi_t$ appearing in (11) in Proposition 1 as follows:

$$\psi_t = \frac{1}{M} \sum_{l=1}^t (1 - \alpha/M)^{-l} \|\epsilon_l\|$$

$$\leq \frac{1}{M} \left( \max_{1 \leq l \leq t} \|\epsilon_l\| \right) \sum_{l=1}^t (1 - \alpha/M)^{-l}$$

$$\leq \frac{1}{M} \left( \max_{1 \leq l \leq t} \|\epsilon_l\| \right) \frac{t}{(1 - \alpha/M)}$$

$$= \frac{1}{M} \left( \max_{1 \leq l \leq t} \|\epsilon_l\| \right) (1 - \alpha/M)^{-t}$$

$$= \frac{1}{\alpha} \left( \max_{1 \leq l \leq t} \|\epsilon_l\| \right) (1 - \alpha/M)^{-t}.$$ 

Recall that

$$L_t^{(N)}(\theta_t) = \frac{1}{N} \sum_{k=1}^N \ell(z_k(\theta_t), \theta_t)$$

is the empirical risk where $z_k(\theta_t)$ is sampled from the current distribution $d_t$, and recall also that

$$L_t(\theta_t) = \mathbb{E}_{z(\theta_t) \sim d_t} [\ell(z(\theta_t), \theta_t)]$$

is the expected risk. For the reader’s convenience, we restate Theorem 2 here.

**Theorem 2.** Suppose that Assumptions 1–6 hold. Fix arbitrary $\xi > 0$ and let $\eta = 1/M$ and $n \geq \log(\xi\alpha/C)/\log(\delta)$ where $\hat{C} = \beta W_1(D(\theta_1), d_0) + \epsilon \beta (G_{\theta} + G_z \epsilon)/M + \epsilon G_z + \xi \alpha$. There exists a universal constant $C_0$ such that with probability $1 - \rho$, Algorithm 1 returns iterates that satisfy

$$\|\theta_{T+1} - \theta^*\| \leq \left(1 - \frac{\alpha}{M}\right)^T \|\theta_1 - \theta^*\| + \xi + \left(\log \left(\frac{1}{\rho}\right) \frac{q \log N}{N}\right)^{1/2}$$

for $N \geq Cq \log q$ where $C = C_0 \max\{c_h, \log(RG/\rho), 1\}$ with

$$c_h = \frac{1}{\log q} \max \left\{ \log \left(\frac{4M}{G^2}\right), \log \left(\frac{8H}{G^3}\right) \right\}.$$

Proof of Theorem 2. Fix arbitrary $\xi > 0$ and let $n \geq \log(\xi\alpha/\hat{C})/\log(\delta)$ where

$$\hat{C} = \beta W_1(D(\theta_1), d_0) + \epsilon \beta (G_{\theta} + G_z \epsilon)/M + \epsilon G_z + \xi \alpha.$$


Then, using the upper bound for $\psi_t$, we have that

$$
\psi_t = \frac{1}{M} \sum_{l=1}^{t} \left(1 - \frac{\alpha}{M}\right)^{-l} \|e_l\| \leq \frac{1}{\alpha} \left(\max_{1 \leq l \leq t} \|e_l\|\right) \left(1 - \frac{\alpha}{M}\right)^{-l},
$$

where

$$
e_l = \nabla E_{z(\theta_l) \sim d_t} [\ell(z(\theta_l), \theta_t)] - \nabla E_{z(\theta_l) \sim D(\theta_t)} [\ell(z(\theta_l), \theta_t)] - \nabla E_{z(\theta_l) \sim d_t} [\ell(z(\theta_l), \theta_t)] + \nabla L_t^{(N)}(\theta_t)
$$

is the gradient error. We thus need bounds on

$$
\|\nabla E_{z \sim d_t} [\ell(z(\theta_l), \theta_t)] - \nabla E_{z(\theta_l) \sim D(\theta_t)} [\ell(z(\theta_l), \theta_t)]\|
$$

and

$$
\|\nabla E_{z(\theta_l) \sim d_t} [\ell(z(\theta_l), \theta_t)] - \nabla L_k^{(N)}(\theta_k)\|.
$$

Lemma 1 provides a bound for the former, and Lemma 5 provides a bound for the latter since $\ell(z(\theta), \theta)$ is $G$-Lipschitz—and hence, $\nabla \ell(z(\theta), \theta)$ is nSGD$((G/2)^2)$ by Lemma 1 in Jin et al. [2019]—and Assumption 3 implies that Assumption 9 holds. Indeed, applying Lemma 5, we get that there exists a universal constant $C_0$ such that the sample gradient $\nabla L_t^{(N)}$ converges uniformly to the population gradient $\nabla L_t$ in Euclidean norm—namely, if $N \geq Cq \log q$, we have that

$$
\Pr \left( \sup_{\theta \in \Theta} \|\nabla L_t^{(N)}(\theta) - \nabla L_t(\theta)\| \leq \phi \sqrt{\frac{Cq \log N}{N}} \right) \geq 1 - \rho
$$

where $\phi = G/2$ is the norm-sub-Gaussian parameter on the gradient of $\ell(z(\theta), \theta)$, and

$$
C = C_0 \max\{c_h, \log(R\phi/\rho), 1\}
$$

with

$$
c_h = \frac{1}{\log q} \max\left\{ \log \left(\frac{4M}{\phi^2}\right), \log \left(\frac{8H}{\phi^3}\right) \right\}.
$$

Hence, with probability $1 - \rho$, we have that

$$
\sup_{\theta \in \Theta} \|\nabla L_t^{(N)}(\theta) - \nabla L_t(\theta)\| \leq \phi \sqrt{\frac{Cq \log N}{N}}.
$$

Putting this together with the bound from Lemma 1, we have that with probability $1 - \rho$,

$$
\max_{1 \leq l \leq t} \|e_l\| \leq \max_{1 \leq l \leq t} \left(\delta^n \beta \mathcal{W}_1(d_0, D(\theta_1)) + \beta \eta \epsilon(G_\theta + G_\epsilon z) \frac{\delta^n}{1 - \delta^n} + \delta^n G_\epsilon z + \frac{G}{2} \sqrt{\frac{Cq \log N}{N}}\right)
$$

$$
\leq \delta^n \beta \mathcal{W}_1(d_0, D(\theta_1)) + \beta \eta \epsilon(G_\theta + G_\epsilon z) \frac{\delta^n}{1 - \delta^n} + \delta^n G_\epsilon z + \frac{G}{2} \sqrt{\frac{Cq \log N}{N}}
$$

where $N \geq Cq \log q$.

Setting $\eta = 1/M$, we need to verify that

$$
\frac{1}{\alpha} \left(\delta^n \beta \mathcal{W}_1(d_0, D(\theta_1)) + \frac{\beta}{M} \epsilon(G_\theta + G_\epsilon z) \frac{\delta^n}{1 - \delta^n} + \delta^n G_\epsilon z\right) \leq \xi
$$

in order to apply Proposition 1 to get the claimed convergence result. To simplify the notation, we define $C_1 = \beta \mathcal{W}_1(D(\theta_1), d_0), C_2 = \epsilon \beta (G_\theta + G_\epsilon z), \text{ and } C_3 = G_\epsilon z$. Observe that by the choice of $n$,

$$
n \geq \log \left(\frac{\xi \alpha}{C_1 + C_2/M + C_3 + \xi \alpha}\right) \frac{1}{\log(\delta)}.
$$
By rearranging, we have that
\[ \delta^n \leq \frac{\xi \alpha}{C_1 + C_2/M + C_3 + \xi \alpha}, \]
so that
\[ \frac{\delta^n}{1 - \delta^n} \frac{1}{\alpha} (C_1 + C_2/M + C_3) \leq \xi. \]
Now, since \( 1 - \delta^n \leq 1 \), we have that
\[ \frac{\delta^n}{\alpha} \left( C_1 + \frac{C_2}{M(1 - \delta^n)} + C_3 \right) \leq \frac{\delta^n}{1 - \delta^n} \frac{1}{\alpha} (C_1 + C_2/M + C_3) \leq \xi, \]
Hence, (13) holds for \( n \geq \log((\xi \alpha/\tilde{C})/\log(\delta)) \) as claimed.

Applying Proposition 1, we have that
\[ \|\theta_T - \theta^*\| \leq \left(1 - \frac{\alpha}{M}\right)^T \|\theta_0 - \theta^*\| + \xi + \phi \sqrt{\frac{Cq \log N}{N}} \]
with probability \( 1 - \rho \), which concludes the proof.

Comments on Expected Value Case. If the decision-maker has access to the gradient of the expected loss (i.e., \( \hat{g}_t = \nabla E_{z(\theta_t) \sim d_t}[\ell(z(\theta_t), \theta_t)] \)), then Algorithm 1 with \( \hat{g}_t = \nabla E_{z(\theta_t) \sim d_t}[\ell(z(\theta_t), \theta_t)] \) returns iterates that satisfy
\[ \|\theta_{T+1} - \theta^*\| \leq \left(1 - \frac{\alpha}{M}\right)^T \|\theta_1 - \theta^*\| + \xi. \]
Corollary 1 follows directly from applying the same proof technique as for Theorem 2 with the exception of the gradient error simply being the difference between the gradient of the expected loss at \( d_t \) and the fixed point distribution. Hence, we omit the details.

C Proofs for Zero Order Oracle Setting

C.1 Proof of Lemma 2
Recall that
\[ \hat{L}(\theta) = E_{\nu \sim \mathcal{Z}}[E_{z(\theta + \mu \nu) \sim D(\theta + \mu \nu)}[\ell(z(\theta + \mu \nu), \theta + \mu \nu)]]. \]
For convenience, we restate Lemma 2 here.

Lemma 2. Under Assumptions 1, 2, and 7, the error between the gradient smoothed loss at \( d_t \) and the gradient of the smoothed expected risk at the fixed point \( D \) satisfies
\[ \|\nabla E_{\nu \sim \mathcal{Z}}[E_{z(\theta_t + \mu \nu) \sim d_t}[\ell(z(\theta_t + \mu \nu), \theta_t + \mu \nu)]] - \nabla \hat{L}(\theta_t)\| \leq G_z \left( \delta^n W(d_0) + 2\delta^n \mu \epsilon + \ell_* \frac{Cq}{\mu} \frac{\delta^n}{1 - \delta^n} \right) \]
where \( d_t = \delta^n d_{t-1} + (1 - \delta^n)D(\theta_t + \mu \nu_t) \), and \( W(d_0) = \max_{\theta \in \Theta} W_1(d_0, D(\theta)) \).

Proof of Lemma 2. The proof of this lemma proceeds as follows. First, we give an upper bound on \( W_1(d_t, D(\theta_t + \mu \nu_t)) \) which is the Wasserstein-1 distance between the distribution at time \( t \) and the fixed point distribution for the query point \( \theta_t + \mu \nu_t \). This requires bounding each of the gaps between iterates \( \|\theta_i - \theta_{i-1}\| \) for \( i \in \{1, \ldots, t-1\} \). Then, using this upper bound, we bound the gradient error using Lipschitzness of \( \ell \) in \( z \).
Upper bound on $W_1(d_t, \mathcal{D}(\theta_t + \mu v_t))$. Using the fact that $d_t = \delta^n d_{t-1} + (1 - \delta^n) \mathcal{D}(\theta_t + \mu v_t)$, we expand $W_1(\mathcal{D}(\theta_t + \mu v_t), d_{t-1})$ as follows:

$$W_1(\mathcal{D}(\theta_t + \mu v_t), d_{t-1}) \leq \delta^n W_1(\mathcal{D}(\theta_t + \mu v_t), d_{t-2}) + (1 - \delta^n) \mathcal{D}(\theta_{t-1} + \mu v_{t-1})$$

where the first inequality comes from the triangle inequality, and the second inequality comes from Assumption 1. Continuing to unroll the recursion, we have that

$$W_1(\mathcal{D}(\theta_t + \mu v_t), d_{t-1}) \leq \delta^n W_1(\mathcal{D}(\theta_t + \mu v_t), d_{t-2} + (1 - \delta^n) \mathcal{D}(\theta_{t-1} + \mu v_{t-1}))$$

Hence, we need a bound on $\|\theta_t - \theta_{t-1}\|$ for each $i \in \{1, \ldots, t-1\}$. Using the fact that $\theta_t = \theta_{t-1} - \eta \mu n_1 W_{z-d_{t-1}}[\ell(z, \theta_{t-1} + \mu v_{t-1})]$, we have that

$$\|\theta_t - \theta_{t-1}\| = \|\theta_{t-1} - \eta \mu n_1 W_{z-d_{t-1}}[\ell(z, \theta_{t-1} + \mu v_{t-1})]v_{t-1} - \theta_{t-1}\|$$

$$= \|\theta_{t-2} - \eta \mu n_1 W_{z-d_{t-2}}[\ell(z, \theta_{t-2} + \mu v_{t-2})]v_{t-2} - \eta \mu n_1 W_{z-d_{t-1}}[\ell(z, \theta_{t-1} + \mu v_{t-1})]v_{t-1} - \theta_{t-1}\|$$

$$= \left\| \frac{q}{\mu} \sum_{j=t-i}^{t-1} \mathbb{E}_{z \sim d_j}[\ell(z, \theta_j + \mu v_j)]v_j \right\|$$

$$\leq \frac{q}{\mu} \sum_{j=t-i}^{t-1} \left\| \mathbb{E}_{z \sim d_j}[\ell(z, \theta_j + \mu v_j)] \right\| v_j$$

This immediately implies that

$$\|\theta_t + \mu v_t - \theta_{t-1} - \mu v_{t-1}\| \leq \|\theta_t - \theta_{t-1}\| + \mu \|v_t - v_{t-1}\| \leq \frac{q n}{\mu} \ell_i + 2\mu.$$

Returning to (14), we have that

$$W_1(\mathcal{D}(\theta_t + \mu v_t), d_{t-1}) \leq \delta^{(t-1)n} W_1(\mathcal{D}(\theta_t + \mu v_t), d_0) + \epsilon(1 - \delta^n) \sum_{i=1}^{t-1} \delta^{(i-1)n} \left( \frac{q n}{\mu} \ell_i + 2\mu \right).$$

so that by applying Lemma 4 with $x = \delta^n$,

$$W_1(\mathcal{D}(\theta_t + \mu v_t), d_{t-1}) \leq \delta^{(t-1)n} W_1(\mathcal{D}(\theta_t + \mu v_t), d_0) + 2\mu \epsilon + \epsilon \frac{q n}{\mu} \ell_i \frac{1}{1 - \delta^n}.$$

Since

$$W_1(\mathcal{D}(\theta_t + \mu v_t), d_t) = W_1(\mathcal{D}(\theta_t + \mu v_t), \delta^n d_{t-1} + (1 - \delta^n) \mathcal{D}(\theta_t + \mu v_t)) \leq \delta^n W_1(\mathcal{D}(\theta_t + \mu v_t), d_{t-1}),$$

we have the desired bound.
we have that
\[ W_1(d_t, D(\theta_t + \mu v_t)) \leq \delta^n W_1(d_0, D(\theta_t + \mu v_t)) + 2\delta^n \mu \epsilon + \frac{\eta \mu \ell \epsilon}{1 - \delta^n} \] (15)
as desired.

**Bounding gradient error.** Using Jensen’s inequality along with Lemma 3, and (15), we have that
\[
\|\nabla E_{v \sim B}[E_z(\theta_t + \mu v), \theta_t + \mu v)] - \nabla \hat{L}(\theta_t)\|^2 \\
\leq E_{v \sim B} \left[ \|\nabla E_{z(\theta_t + \mu v), \theta_t + \mu v)] - \nabla E_{z(\theta_t + \mu v), \mu v)}[\ell(z(\theta_t + \mu v), \theta_t + \mu v)]\|^2 \right] \\
\leq G^2 (W_1(d_t, D(\theta_t + \mu v)))^2 \\
\leq G^2 \left( \delta^n \bar{W}(d_0) + 2\delta^n \mu \epsilon + \frac{\eta \mu \ell \epsilon}{1 - \delta^n} \right)^2,
\]
Taking the square root of both sides completes the proof. \(\square\)

### C.2 Zero Order Empirical Oracle: Proof of Theorem 3

For convenience, we restate Theorem 3 here.

**Theorem 3.** Suppose that Assumptions 1-7 hold. There exists a universal constant \(C_0\) such that with probability \(1 - \rho\), Algorithm 2 with \(n \geq \Omega(\log \eta), \eta = \eta_0/T\) with \(\eta_0 > 1/(3\alpha), \tau = \mu/r\) and \(\mu = (q^2 \ell^2 \eta^2/(2MR))^{1/3}\), returns iterates satisfying

\[
E[\frac{1}{2} ||\theta_T - \theta^*||^2 | E_T] \leq O \left( T^{-1/3} + \left( \log \left( \frac{1}{\rho} \right) \frac{q \log N}{N} \right)^{1/2} \right)
\]

for \(N \geq Cq \log q, C = C_0 \max\{c_h, \log(RG/\rho), 1\}\) and

\[
c_h = \frac{1}{\log q} \max \left\{ \log \left( \frac{4M}{G^2} \right), \log \left( \frac{8H}{G^2} \right) \right\}.
\]

Recall that we assume that \(\Theta\) contains the ball of radius \(r\) centered at the origin and is contained in the ball of radius \(R\)—i.e., \(rB \subseteq \Theta \subseteq RB\).

The zeroth order algorithm projects onto \((1 - \tau)\Theta\) for \(\tau \in (0, 1)\)—namely,

\[
\theta_{t+1} = \text{Proj}_{(1 - \tau)\Theta}(\theta_t - \eta \hat{g}_t)
\]

where \((1 - \tau)\Theta\) is shorthand for \(\{(1 - \tau)\theta : \theta \in \Theta\}\). The projection is made onto this set so that the random query points around \(\theta_t\) belong to \(\Theta\). In particular, as noted in the main body, for any \(\theta \in (1 - \tau)\Theta\) and any unit vector \(v\), it holds that \(\theta + \mu v \in \Theta\) for any \(\mu \in [0, \tau r]\) [Flaxman et al., 2004, Observation 2].

Recall also that the smoothed expected risk at \(D(\theta)\) is denoted

\[
\hat{L}(\theta) = E_{v \sim B}[E_{z(\theta + \mu v), \theta + \mu v)}[\ell(z(\theta + \mu v), \theta + \mu v)],
\]

the expected risk is denoted

\[
L(\theta) = E_{z(\theta) \sim D(\theta)}[\ell(z(\theta), \theta)],
\]

and the smoothed expected risk at \(d_t\) is denoted

\[
L_s(\theta) = E_{v \sim B}[E_{z(\theta + \mu v), \theta + \mu v)}[\ell(z(\theta + \mu v), \theta + \mu v)].
\]

Before getting into the formal proof of Theorem 3, we start by decomposing the gradient to reveal additional terms in the bias that we must bound. To bound this additional source of bias we appeal to Theorem 1 of Mei et al. [2016], which we make slight modifications to for our use case. This leads to the technical lemma given in Lemma 5.
Decomposing zeroth-order gradient. As stated in Algorithm 2, the gradient the decision-maker is following is the empirical estimate of the zeroth order gradient estimate at $d_t$—i.e.,

$$
\hat{g}(\theta_t + \mu v_t) = \frac{q}{\mu} \frac{1}{N} \sum_{k=1}^{N} \ell(z_k(\theta_t + \mu v_t), \theta_t + \mu v_t) v_t,
$$

where $z_k(\theta_t + \mu v_t)$ is the $k$-th sample from $d_t$. We decompose the gradient $\hat{g}(\theta_t + \mu v_t)$ as follows:

$$
\hat{g}(\theta_t + \mu v_t) = \nabla L(\theta_t) + \hat{g}(\theta_t + \mu v_t) - \mathbb{E}_{v \sim \mathcal{B}} \left[ \frac{q}{\mu} \frac{1}{N} \sum_{k=1}^{N} \ell(z_k(\theta_t + \mu v_t), \theta_t + \mu v_v) v_t \right] \\
+ \nabla \mathbb{E}_{v \sim \mathcal{B}} \left[ \frac{1}{N} \sum_{k=1}^{N} \ell(z_k(\theta_t + \mu v_t), \theta_t + \mu v_t) \right] + \nabla L_t(\theta_t) - \nabla \hat{L}(\theta_t) - \nabla \hat{L}(\theta_t) - \nabla L(\theta)
$$

Hence,

$$
\hat{g}(\theta_t + \mu v_t) = \nabla L(\theta_t) + U_{t+1} + b_t
$$

where $U_{t+1} = \hat{g}(\theta_t + \mu v_t) - \mathbb{E}_{v \sim \mathcal{B}}[\hat{g}(\theta_t + \mu v)]$ is the noise process due to the zeroth order gradient estimate, and the term $b_t$ is the bias due to the zeroth order gradient estimate and the gap between the distributions $d_t$ and $\mathcal{D}(\theta_t + \mu v_t)$—i.e., the distribution at time $t$ versus the fixed point distribution.

Proof of Theorem 3. The proof proceeds as follows. Using the gradient decomposition above, we show that $V(\theta^*, \theta_t) = \frac{1}{2} \|\theta_t - \theta^*\|^2$ is decreasing along trajectories generated by the zeroth order algorithm (Algorithm 2). Finally, we optimize over the query radius $\mu$ and stepsize $\eta$ to get the claimed convergence rate.

Constructing a potential function. Let $V(\theta^*, \theta_t) = \frac{1}{2} \|\theta_t - \theta^*\|^2$. Using the non-expansiveness of the projection operator, we have that

$$
V(\theta^*, \text{Proj}_{(1-\tau)\Theta}(\theta_t - \eta \hat{g}(\theta_t + \mu v_t))) \leq V(\theta^*, \theta_t) - \eta \langle \nabla L(\theta_t) + U_{t+1} + b_t, \theta_t - \theta^* \rangle + \frac{\eta^2}{2} \|\hat{g}(\theta_t + \mu v_t)\|^2
$$

$$
= V(\theta^*, \theta_t) - \eta \langle \nabla L(\theta_t) + U_{t+1} + b_t, \theta_t - \theta^* \rangle + \frac{\eta^2}{2} \|\hat{g}(\theta_t + \mu v_t)\|^2
$$

$$
\leq V(\theta^*, \theta_t) - \eta \langle \nabla L(\theta_t), \theta_t - \theta^* \rangle - \eta \varphi_{t+1} - \eta r_t + \frac{\eta^2}{2} \|\hat{g}(\theta_t + \mu v_t)\|^2
$$

where $\varphi_{t+1} = \langle U_{t+1}, \theta_t - \theta^* \rangle$ and $r_t = \langle b_t, \theta_t - \theta^* \rangle$. Since the performative risk is $\alpha$-strongly convex at the fixed point distribution, we have that

$$
-\langle \nabla L(\theta_t), \theta_t - \theta^* \rangle \leq -\langle \nabla \hat{L}(\theta_t) - \nabla \hat{L}(\theta^*), \theta_t - \theta^* \rangle \leq -\frac{\alpha}{2} \|\theta_t - \theta^*\|^2 = -\alpha V(\theta^*, \theta_t).
$$

Hence,

$$
V(\theta^*, \text{Proj}_{(1-\tau)\Theta}(\theta_t - \eta \hat{g}(\theta_t + \mu v_t))) \leq V(\theta^*, \theta_t) - \eta \langle \nabla L(\theta), \theta_t - \theta^* \rangle - \eta \varphi_{t+1} - \eta r_t + \frac{\eta^2}{2} \|\hat{g}(\theta_t + \mu v_t)\|^2
$$

$$
\leq (1 - \alpha \eta) V(\theta^*, \theta_t) - \eta \varphi_{t+1} - \eta r_t + \frac{\eta^2}{2} \|\hat{g}(\theta_t + \mu v_t)\|^2,
$$

so that

$$
\mathbb{E}[V(\theta^*, \text{Proj}_{(1-\tau)\Theta}(\theta_t - \eta \hat{g}(\theta_t + \mu v_t)))] \leq (1 - \alpha \eta) \mathbb{E}[V(\theta^*, \theta_t)] - \eta \mathbb{E}[\varphi_{t+1}] - \eta \mathbb{E}[r_t] + \frac{\eta^2}{2} \mathbb{E}[\|\hat{g}(\theta_t + \mu v_t)\|^2]
$$

$$
\leq (1 - \alpha \eta) \mathbb{E}[V(\theta^*, \theta_t)] - \eta \mathbb{E}[r_t] + \frac{\eta^2 q^2 \ell^2}{2 \mu^2},
$$

since $\mathbb{E}_{v \in \mathcal{B}}[U_{t+1}] = 0$ implies that $\mathbb{E}[\varphi_{t+1}] = \mathbb{E}[U_{t+1}, \theta_t - \theta^*] = 0$. Now, by Cauchy-Schwartz, we have that

$$
-\eta r_t = \langle b_t, \theta^* - \theta_t \rangle \leq \|b_t\| \|\theta_t - \theta^*\| \leq 2R \|b_t\|.
$$
Bound on $b_t$. To simplify notation a bit, define

$$
\hat{L}_t^{(N)}(\theta_t) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}_{v \sim B}[\ell(z_k(\theta_t + \mu v), \theta_t + \mu v)]
$$

and

$$
\hat{L}_t(\theta_t) = \mathbb{E}_{v \sim B}[\mathbb{E}_{z(\theta_t + \mu v) \sim d_t}[\ell(z(\theta_t + \mu v), \theta_t + \mu v)],
$$

so that

$$
|b_t| \leq \|\nabla \hat{L}_t(\theta_t) - \nabla \hat{L}_t^{(N)}(\theta_t)\| + \|\nabla \hat{L}_t(\theta_t) - \nabla \hat{L}(\theta)\| + \|\nabla \hat{L}(\theta) - \nabla L(\theta)\|
$$

where the second inequality follows from the fact that

$$
\|\nabla \mathbb{E}_{v \sim B}[\mathbb{E}_{z(\theta_t + \mu v) \sim d_t}[\ell(z(\theta_t + \mu v), \theta_t + \mu v)] - \nabla \hat{L}(\theta)\| + \|\nabla \hat{L}(\theta) - \nabla L(\theta)\|
$$

$$
\leq G_z \left( \delta^n \overline{W}(d_0) + 2\delta^n \mu \epsilon + \frac{\eta}{\mu} q^e, \epsilon \frac{\delta^n}{1 - \delta^n} \right) + M \mu,
$$

since

$$
\|\nabla \hat{L}(\theta) - \nabla L(\theta)\| = \|\nabla \mathbb{E}_{v \sim B}[\mathbb{E}_{z \sim D(\theta + \mu v)}[\ell(z, \theta + \mu v)] - \nabla \mathbb{E}_{v \sim B}[\ell(z, \theta)]\| = 0,
$$

which follows from the fact that $L(\theta)$ is $M$-smooth. Hence, we simply need to bound

$$
\|\nabla \hat{L}_t(\theta_t) - \nabla \hat{L}_t^{(N)}(\theta_t)\|
$$

which follows from an application of Lemma 5. Indeed, by Lemma 5 with probability $1 - \rho$, we have that

$$
|b_t| \leq \frac{G_z}{2} \sqrt{\frac{C q \log N}{N}} + G_z \left( \delta^n \overline{W}(d_0) + 2\delta^n \mu \epsilon + \frac{\eta}{\mu} q^e, \epsilon \frac{\delta^n}{1 - \delta^n} \right) + M \mu
$$

as long as $N \geq C q \log q$.

Finishing the argument that $V(\theta^*, \theta_t)$ decreases on trajectories. With the above bound on $b_t$, we have that

$$
\mathbb{E}[V(\theta^*, \text{Proj}_{(1 - r)\Theta}(\theta_t - \eta \hat{g}(\theta_t + \mu v_t)))]
$$

$$
\leq (1 - \alpha \eta)\mathbb{E}[V(\theta^*, \theta_t)] + \eta \mathbb{E}[-r_t] + \frac{\eta^2 q^2 \ell_z^2}{2\mu^2}
$$

$$
\leq (1 - \alpha \eta)\mathbb{E}[V(\theta^*, \theta_t)] + 2R \left( \frac{G_z}{2} \sqrt{\frac{C q \log N}{N}} + G_z \left( \delta^n \overline{W}(d_0) + 2\delta^n \mu \epsilon + \frac{\eta}{\mu} q^e, \epsilon \frac{\delta^n}{1 - \delta^n} \right) + M \mu \right) + \frac{\eta^2 q^2 \ell_z^2}{2\mu^2}.
$$

The epoch horizon length $n$ is chosen such that

$$
G_z \left( \delta^n \overline{W}(d_0) + \mu \left( 2\delta^n \epsilon + \frac{M}{G_z} \right) + \frac{\eta}{\mu} q^e, \epsilon \frac{\delta^n}{1 - \delta^n} \right) \leq \frac{\eta q^2 \ell_z^2}{4R \mu^2}.
$$

Indeed, since $1 - \delta^n \leq 1$,

$$
\delta^n \overline{W}(d_0) + \mu \left( 2\delta^n \epsilon + \frac{M}{G_z} \right) + \frac{\eta}{\mu} q^e, \epsilon \frac{\delta^n}{1 - \delta^n} \leq \frac{\delta^n}{1 - \delta^n} \left( \overline{W}(d_0) + 2\mu \epsilon + \frac{\eta}{\mu} q^e, \epsilon \right) \leq \frac{\eta q^2 \ell_z^2}{4RG_z \mu^2}.
$$
or equivalently,
\[ n \geq \log \left( \frac{\eta^2 \ell^2}{4RG\mu^2} \right) \frac{1}{\log(\delta)} + \frac{\eta^2 q^2 \ell^2}{4RG\mu^2} + \frac{\eta q}{\mu} \epsilon + \frac{\eta^2 q^2 \ell^2}{4RG\mu^2}. \] (18)

Hence,
\[ \mathbb{E}[V(\theta^*, \text{Proj}_{(1-r)}(\theta_t - \eta \hat{g}(\theta_t + \mu v_t)))] \leq (1 - \alpha \eta) V(\theta^*, \theta_t) + \eta 2RM\mu + \frac{\eta^2 \ell^2}{\mu^2} + 2\eta \sqrt{\frac{Cq \log N}{N}}. \]

Iterating on this recursion, we have
\[ \mathbb{E}[V(\theta^*, \text{Proj}_{(1-r)}(\theta_T - \eta \hat{g}(\theta_T + \mu v_T)))]] \leq (1 - \alpha \eta)^T \mathbb{E}[V(\theta^*, \theta_0)] + \sum_{t=0}^{T} (1 - \alpha \eta)^t 2\eta \left( RM\mu + \frac{\eta^2 \ell^2}{\mu^2} + 2RG \sqrt{\frac{Cq \log N}{N}} \right). \]

Since \[ \sum_{t=0}^{T} (1 - a)^t \leq \sum_{t=0}^{\infty} (1 - a)^t = \frac{1}{a} \]
we have that
\[ \mathbb{E}[V(\theta^*, \text{Proj}_{(1-r)}(\theta_T - \eta \hat{g}(\theta_T + \mu v_T)))]] \leq (1 - \alpha \eta)^T \mathbb{E}[V(\theta^*, \theta_0)] + \frac{1}{a} \left( 2RM\mu + \frac{\eta^2 \ell^2}{\mu^2} + 2RG \sqrt{\frac{Cq \log N}{N}} \right). \]

**Optimizing over query radius and stepsize.** Optimizing over \( \mu \), we get
\[ \mu_* = \left( \frac{\ell^2 \eta q^2}{2MR} \right)^{1/3}. \]

Plugging in \( \mu_* \), we have that
\[ \mathbb{E}[V(\theta^*, \text{Proj}_{(1-r)}(\theta_T - \eta \hat{g}(\theta_T + \mu v_T)))]] \leq (1 - \alpha \eta)^T \mathbb{E}[V(\theta^*, \theta_0)] + \frac{2RG}{\alpha} \sqrt{\frac{Cq \log N}{N}} + 3(\ell_* qMR)^{2/3} \eta^{1/3}. \]

Now, suppose that we choose \( \eta = \eta_0 / T^p \) for some \( 0 < p \leq 1 \). Then, we have that
\[ \mathbb{E}[V(\theta^*, \text{Proj}_{(1-r)}(\theta_T - \eta \hat{g}(\theta_T + \mu v_T)))]] \leq \left( 1 - \frac{\alpha \eta_0}{T^p} \right)^T \mathbb{E}[V(\theta^*, \theta_0)] + \frac{2RG}{\alpha} \sqrt{\frac{Cq \log N}{N}} + 3(\ell_* qMR)^{2/3} \frac{\eta_0}{T^{p/3}}. \]

With \( p = 1 \) and \( \eta_0 \geq 1/(3\alpha) \), by Lemma D.2 of Bravo et al. [2018], we have that
\[ \mathbb{E}[V(\theta^*, \text{Proj}_{(1-r)}(\theta_T - \eta \hat{g}(\theta_T + \mu v_T)))]] \leq O \left( T^{-1/3} + \frac{1}{(\rho^2 \log N)} \right), \]
which completes the proof. \( \square \)

**Comments on Setting with Access to Expected Loss.** Analogous to the first order setting, if the decision maker has access to the expected loss at \( d_t \), then the convergence results of Theorem 3 improve accordingly.

**Corollary 2.** Suppose that Assumptions 1–4, and 7 hold. Algorithm 2 with oracle \( \hat{g}_t = \frac{1}{n} \mathbb{E}_{z(\theta_t + \mu v_t)} \mathbb{E}_{z(\theta_t + \mu v_t)} [f(z(\theta_t + \mu v_t), \theta_t + \mu v_t)] v_t \), \( n \geq \Omega(\log \eta) \), \( \eta = \eta_0 / T \) where \( \eta_0 > 1/(3M) \), \( \tau = \mu / r \) and \( \mu = (q^2 \ell^2 \eta / (MR))^{1/3} \), returns iterates satisfying
\[ \mathbb{E}[\|\theta_T - \theta^*\|^2] \leq O(T^{-1/3}). \]

This setting no longer requires the additional regularity assumptions (smoothness and Hessian Lipschitzness of \( f \)) required for the uniform convergence result in Lemma 5. We omit the proof of this corollary since it follows directly the same analysis as that of Theorem 3 sans the additional bound on the gap between the empirical zero order gradient estimate and the expected gradient of the smoothed loss at \( d_t \).

26
D Numerical Simulations

In this section, we start by describing the SFpark data and experiment set-up. Then we provide additional figures and details for each of the two experiments conducted in the main. Finally, we introduce a synthetic data example which abstracts strategic classification in settings where agents have memory.

D.1 SFpark Data Description

In this section, we provide more details on our data cleaning strategies and our model for the SFpark dataset.

Data cleaning. We start by discussing our data cleaning strategy. Of the many features in the dataset, the key ones of interest to us were the street name, district name, total time available (number of parking spots multiplied by number of seconds per hour), total time occupied, and rate. Many of the rates were unavailable in the original dataset, but the rate charged for the day before and day after were. If we encountered a missing rate, we replaced it with the rate before and after, if those rates were equal. We only worked with blocks where we could successfully fill in each of the missing rates. This process can be found in the accompanying code.

Estimating price sensitivity. The model we consider is explained in the main body. To provide more intuition and details, as an example, consider the 600 block of Beach Street for the time window between 1200–1500. The initial distribution, \( d_0 \), is sampled from the data at the initial price for parking along the 600 block of Beach ST (Beach ST 600), which in this case is \( \theta_0 = $3 \) per hour. As described in Section 5, we assume that for an announced price difference of \( \theta = \tilde{\theta} - \theta_0 \), \( \tilde{\theta} \) is the charged price and \( \theta \) is the variable of optimization. The occupancy follows a distribution of \( \zeta_0 + m(\tilde{\theta} - \theta_0) \), where \( \zeta_0 \) follows the same distribution as \( d_0 \).

The price sensitivity \( m \) is a proxy for the price elasticity, in that it provides us a relationship between the change in price and the change in occupancy mapped to a \((0, 1)\) scale. Indeed, recall that price elasticity is a change in the percentage occupancy for a given change in percentage price. Hence, price sensitivity as we have defined it has the same sign as price sensitivity except that it is in the right units of our mathematical abstraction for the problem, and is in this sense a proxy thereof. We compute \( m \) by considering the following:

a. The average occupancy for the initial price over every weekday in the beginning of the pilot study until the price is changed.

b. The average occupancy over every weekday in the final week of the last price announcement.

As an example, for the 600 block of Beach ST, the initial price was $3.00 per hour and the average occupancy before a new price was announced was approximately 60.6%, the final price announced during the pilot study was $4.25, and the average occupancy for the final week was approximately 41.1%. Therefore, for the 600 block of Beach ST, we estimate that

\[
m \approx \frac{0.411 - 0.606}{4.25 - 3} = -0.156,
\]

where occupancy percentage is mapped to the \([0, 1]\) scale. It was shown in Pierce and Shoup [2018] that price elasticity is in general a small negative number on average for the SFpark pilot study and experiment. This is consistent with prior studies on price elasticity for on-street parking where information about price and location plays a crucial role [Fiez and Ratliff, 2020, Glasnapp et al., 2014]. However, for the SFpark pilot study, the price elasticity also depends highly on the block and neighborhood.

Estimating geometric decay parameter \( \delta \). We also use this data to estimate the geometric decay rate, \( \delta \). As described in Section 5, when a new rate is posted, the effect on the occupancy is not immediate, and so the geometric decay rate, \( \delta \), in this context represents the speed at which this new announced price travels through the population (and consequently affects the parking occupancy). We group the occupancy data by day of week, in order to account for different traffic patterns on different weekdays. We assume that the week before a new price is announced is the fixed point distribution of the previous rate. For example, for
Figure 3: Results of Algorithm 1 (first and third plots of each row) and Algorithm 2 (second and fourth plots of each row) with different \((n, T)\) pairs for the 500, 700 and 800 blocks of Beach ST for time window 1200–1500. Each marker represents a price announcement, and the plots show the prices and corresponding predicted occupancies. The SFpark prices and occupancies are far from the target and performative optimal price, whereas the proposed algorithms obtain both points up to theoretical error bounds.

Figure 4: Map of Beach Street showing blocks 500 to 800. The tourist attractions Ghiradelli Square, Fisherman’s Wharf, and the Embarcadero are also depicted.

the 600 block of Beach ST, a rate of $3.50 per hour was announced on February 14, 2012, which means that we assumed that the occupancies on February 7-13, 2012 were the fixed point distributions of the previous rate $3.25. We now fix a day of the week (e.g., Monday), a block (e.g., Beach ST 600), and a time window (e.g., 1200–1500). Suppose the prices \(\{p_i\}\) are announced and \(D(p_i)\) represents the fixed point distribution of announcing \(p_i\), where the price \(p_i\) is in effect for \(K_i\) weeks. Then, for the \(k\)-th week after announcing \(p_i\), we assume that the occupancy is represented by \(\delta^k D(p_{i-1}) + (1 - \delta^k) D(p_i)\). For each week \(k\), and for price \(p_i\), the occupancy for the specified day is represented as \(y_{i,k}\). To find the value of \(\delta\), for the specified day
and block, we solve the following optimization problem:

$$\min_{\delta \in [0, 1]} \sum_{i} \sum_{k=1}^{K_i} (\delta^k D(p_{i-1}) + (1 - \delta^k) D(p_i) - y_{i,k})^2.$$  

We perform projected gradient descent to solve this problem. For the final value of \(\delta\) that we use for the specified block, we average the estimated values of delta for each day.

**D.2 Comparing Performative Optimum to SFpark**

Here, we provide experiments for other blocks on Beach Street (beyond just the 600 block in Section 5). Each row in Figure 3 shows prices and corresponding occupancies for Algorithm 1 and Algorithm 2 for the 500, 700, and 800 blocks of Beach ST, respectively. In each instance, we make similar observations to those in Section 5 for the 600 block on Beach ST, namely, that SFpark consistently overshot the price to reach the target occupancy, and that the choice of \(n = 8\) is reasonable, in that a time period of 8 weeks is sufficient for the population to equilibriate before announcing a new price.

An interesting observation from Figure 3 comes from the fact that the 500 block of Beach ST has a price sensitivity of \(m \approx -0.844\), and the 800 block of Beach ST has a price sensitivity of \(m \approx -0.424\). Since both of these values have large magnitudes, we observe that for a small price reduction, the estimated occupancy increases to 100%. Therefore, for blocks where the magnitude of the price sensitivity is large, our experiments suggest using a smaller choice of \(n\), and consequently a larger choice of \(T\), in order to reduce the variance for the price announcements to prevent having large fluctuations in occupancy. All four of the blocks on Beach Street have very similar estimated \(\delta\) values. Table 1 indicates that each block adjusts to new price announcements at similar rates. This makes sense given that each of the blocks are on the same street all next to each other as seen in Figure 4, and located near similar landmarks.

**D.3 Redistributing Parking Demand**

In this appendix subsection, we describe the details for the experiment on redistributing parking demand. The study includes the four connected blocks Hawthorne ST 0, Hawthorne ST 100, Folsom ST 500, and Folsom ST 600 because the blocks are adjacent to one another as shown in Figure 2. Thus, we wanted to investigate whether price changes would redistribute the traffic such that each block had an occupancy closer to the target of 70%. An interesting note is that while Folsom ST 500 and Folsom ST 600 both have negative price sensitivity values of \(m \approx -0.399\) and \(m \approx -0.284\) respectively, Hawthorne ST 0 and Hawthorne ST 100 have positive price sensitivity values of \(m \approx 0.454\) and \(m \approx 0.044\) respectively. Since Hawthorne ST has a very high initial average occupancy, SFpark should consider decreasing prices on this street in order to shift demand to the nearby streets. This is exactly what we see done by both Algorithms 1 and 2 so that both streets are closer to the target occupancy. Although the price sensitivity is very different for these blocks, the estimated \(\delta\) values are very similar. Hawthorne ST 0 has \(\delta \approx 0.853\), Hawthorne ST 100 has \(\delta \approx 0.979\), Folsom ST 500 has \(\delta \approx 0.996\), and Folsom ST 600 has \(\delta \approx 0.793\), so each block adjusts to new price announcements at similar rates.

**D.4 Synthetic Data: Strategic Classification in Dynamic Environments**

In this appendix subsection, we apply our algorithm to a synthetic strategic classification problem—which was considered in the dynamic setting in Brown et al. [2020] and in the static setting in Drusvyatskiy and
Xiao [2020], Miller et al. [2021], Perdomo et al. [2020], e.g.—where there is memory in the agent population. For simplicity (and to support visualization of the classifier performance), each data point contains a feature vector, \( x \in \mathbb{R}^2 \), and a corresponding label, \( y \in \{-1, 1\} \). The loss incurred by the institution is given by an \( \ell_2 \)-regularized logistic loss:

\[
\frac{1}{2} \sum_{i=1}^{m} -y_i \langle \theta, x_i \rangle + \log(1 + \exp(\langle \theta, x_i \rangle)) + \nu \|\theta\|^2,
\]

where \( m = 1000 \) is the number of agents. The agents are non-strategic (meaning they do not perturb their true feature vector \( x_i \)) if they have label \( y = 1 \), and otherwise ‘best respond’ to the announced classifier according to the model

\[
x_i = \arg \max_{x_i'} -\langle \theta, x_i' \rangle - \frac{1}{2\epsilon} \|x_i' - x_i\|^2 = x_i' - \epsilon \theta.
\]

We take \( \epsilon = 0.1 \), but the observations we make hold more generally with the exception of very large magnitude perturbations for which the problem (even in the static setting) becomes untenable. We randomly select a subset of the two features to treat as strategic. We also randomly generate a ground truth data set by drawing \( m \times 2 \) samples from a normal distribution, drawing the ground truth \( \theta_{gt} \) from a (2 dimensional) normal distribution and then assigning labels according to

\[
y_i = (\text{sign}(x_i^T \theta_{gt} + 0.1v) + 1)/2, \ v \sim \mathcal{N}(0, 1).
\]

Specifically, agents are allowed to perturb in the \( x_1 \) direction as can be seen in Figure 5. Moreover, we take the initial data distribution \( d_0 \) to be far from the base distribution for users’ true preferences \( x_i' \) even with performative effects; specifically, \( d_0 \) is a Gaussian distribution with a mean of 1.0 and scale (standard deviation) of 45. More details on the implementation can be found in the accompanying code.

We divide the data into a training and test set with a \( (2/3)–(1/3) \) split. We set the regularization parameter to \( \nu = 1/m_{\text{train}} \) where \( m_{\text{train}} \) is the size of the training data set and is also the batch size we use—namely, \( N = m_{\text{train}} \). The inner product can be interpreted as the utility of the agent and the norm difference as the cost of manipulation. We present results for a modest value of \( n = 20 \); similar or lower values are consistent with our observations and as our theory suggests, as \( n \to \infty \), the solution obtained by Algorithm 1 approaches the performatively optimal solution.

![Figure 5: Classifiers and losses for different values of \( \delta \) and \( n = 20 \). In order of appearance from left to right, the first three plots show the learned classifiers with the data at the distribution \( \mathcal{D}(\theta) \) induced by the learned classifier for \((\delta, n) = (0.5, 20), (\delta, n) = (0.95, 20), (\delta, n) = (0.99, 20)\). The fourth plot from the left is the ground truth data distribution without performative effects. The differences in the data distributions are subtle, but one can see that the different learned classifiers evoke different responses from the strategic users. The far right plot shows the losses as a function of iterations.](image_url)

We explore different values of \( \delta \) and \( n \)—i.e., the mixing parameter of the geometric dynamics and the epoch length of Algorithm 1—on not just convergence (the result of which is characterize by Theorem 2 in the main body) but also on accuracy. The observations we report actually lead to a number of interesting open questions for this field including how performative optimality relates to generalization. We find that depending on the skew of the data distribution and the strength of the perturbation power of the strategic agents—namely, \( \epsilon \)—that surprisingly, the performatively optimal point may not generalize very well as compared to the solution obtained by Algorithm 1 when the mixing parameter \( \delta \) is large. The latter has
better accuracy as can be seen in Figure 6; the loss value per iteration and the classifiers for different $\delta$ values are shown in Figure 5.

In other settings (e.g., with different ground truth data), the performatively optimal solution obtained by Algorithm 1, even with different values of $\delta$ and different choices of epoch length $n$, performs just as well as the performatively optimal solution as depicted in Figure 8, the data for which has original distribution depicted in Figure 7, which also contains the learned classifiers and losses per iteration for different $\delta$ values.

These observations about the generalization performance of the obtained solution under our proposed algorithm (for different values of the geometric process or mixing constant $\delta$) as compared to the performatively optimal point, while highly dependent on the underlying data distribution, open up a number of interesting directions for future work on understanding precisely when the performatively optimal point gives good generalization and robustness guarantees.

Figure 6: Accuracy of the classifiers (via confusion matrix) learned for the data distribution and setting shown in Figure 5. For this randomly sampled data distribution, $\delta$ plays a significant role on the generalization capability (as measured by accuracy on the test set). Surprisingly, accuracy improves as the mixing parameter $\delta$ increases (meaning longer time to mix) and this also has an impact on auxiliary but related metrics such as the false positive and false negative rates. This observation depends highly on the data distribution, but exposes interesting directions for future theoretical work on understanding how performative optimality translates to generalization and robustness guarantees.

Figure 7: Classifiers and losses for different values of $\delta$ and $n$, for the given original data distribution shown in the far right plot. (left) Different classifiers (as a function of $\delta$ and $n$) and the data distribution given the strategic best response at the performatively optimal point. (center) Losses for the different $(\delta,n)$ pairs as a function of iteration. (right) Original data distribution and ground truth classifier.

D.5 Semi-Synthetic Data: Strategic Classification in Dynamic Environments

As a point of comparison to the existing literature, we perform additional numerical experiments on a strategic classification simulator from the Kaggle *Give Me Some Credit* dataset discussed in Perdomo et al. [2020] and Brown et al. [2020]. In this dataset, each data point contains a feature vector, $x \in \mathbb{R}^q$, which represents historical information about an individual, and the label, $y \in \{0,1\}$, represents whether or not
In the second set of experiments, we compare our approach to an epoch based implementation of both risk minimization (RRM) [Perdomo et al., 2020, Brown et al., 2020], and repeated gradient descent (RGD) [Perdomo et al., 2020]—implemented for the dynamic environment which was not considered in Perdomo et al. [2020]—both of which, notably update $\theta$ at every iteration in $[0,nT]$ where as our approach (Algorithm 1) only updates at every $n$ steps in that same interval.

In the second set of experiments, we compare our approach to an epoch based implementation of both RRM and RGD where in these implementations the dynamics are also allowed to “mix” and the decision maker updates only every $n$ steps as in our method. These later experiments are more comparable even though the epoch based implementations of RRM and RGD have not been studied theoretically. For both experiments, we plot the $\ell_2$ distance to the performatively optimal point.

**Experiment 1: Comparison to Iteration-Based (Classical) RRM and RGD.** Figure 9 shows the results of the first set of experiments, for which we have taken $\delta = 0.9$, which is relatively large meaning that the mixing time for the geometric process is large. Neither RRM nor RGD target the performatively optimal point, but instead the performatively stable point, i.e., the point at which repeated retraining will
Figure 10: ‘Give Me Some Credit’ Experiment 2: Results of Algorithm 1 compared to epoch-based implementations of RRM and RGD—i.e., where in each epoch the dynamics are updated \( n \) times with the same classifier deployed—each called with \((n, T) \in \{(10, 1000), (100, 1000)\}\). Each marker represents a new \( \theta \) announcement, and the plots show the Euclidean distance from the performatively optimal point.

stabilize. As shown in Figure 9, a performatively stable point (the point RRM was shown to converge to in Brown et al. [2020]) may be far from the performatively optimal point. Interestingly, we also observe that for small values of \( \tilde{\varepsilon} \) (i.e. on the order of 1e-2), the performatively optimal point and the performatively stable point are very close, and so RGD behaves nearly identically to calling Algorithm 1 with \( n = 1 \). This seems to imply that when performative effects (i.e., size of \( \tilde{\varepsilon} \) in this set of experiments) are very low, the naive strategies of RRM or RGD suffice when trying to find the optimal point. On the other hand, for values of \( \tilde{\varepsilon} \) on the order of 1e-1 or larger, RRM and RGD do not converge to the performatively optimal point while Algorithm 1 does, albeit with worse iteration complexity to convergence to the stable point of the respective algorithm.

Experiment 2: Comparison to Epoch-Based RRM and RGD. Figure 10 shows the results of the second set of experiments. As noted above, in this set of experiments, we compare to epoch based implementations of RRM and RGD to Algorithm 1 which is also an epoch-based algorithm, the idea here being that these are more comparable algorithms in a sense. As can be seen in Figure 10, the observations are analogous to the first set of experiments. Epoch-based RRM and RGD converge to the performatively stable point (as defined in [Perdomo et al., 2020] and [Brown et al., 2020], for the dynamic setting). For \( \tilde{\varepsilon} \) on the order of 1e-2, the performatively stable point is close to the performatively optimal point (although still not equal to it), and for \( \tilde{\varepsilon} \) on the order of 1e-1 or larger, the performatively stable point is considerably farther away from the performatively optimal point. On the other hand, Algorithm 1 converges to the performatively optimal point for all shown values of \( \tilde{\varepsilon} \), the size of the strategic perturbation.

We note that we did not compare to the zero-th order method since it has different information than both the RRM and RGD and is thus less comparable. We expect the same observations about non-convergence of RRM and RGD for large \( \tilde{\varepsilon} \) to persist and Algorithm 2 will converge as the theory predicts, albeit at a much slower rate than Algorithm 1 due to the bandit feedback.