Two-timescale Derivative Free Optimization for Performative Prediction with Markovian Data

Haitong Liu [1] **Qiang Li** [2] **Hoi-To Wai** [2]

Abstract

This paper studies the performative prediction problem where a learner aims to minimize the expected loss with a decision-dependent data distribution. Such setting is motivated when outcomes can be affected by the prediction model, e.g., strategic classification. We consider a state-dependent setting where the data distribution evolves according to a controlled Markov chain. We focus on stochastic derivative free optimization (DFO) where the learner is given access to a loss function evaluation oracle with the above Markovian data. We propose a two-timescale DFO(λ) algorithm that features (i) a sample accumulation mechanism that utilizes every observed sample to estimate the gradient of performative risk, (ii) a two-timescale diminishing step size that balances the rates of DFO updates and bias reduction. Under a non-convex optimization setting, we show that DFO(λ) requires $\mathcal{O}(1/\epsilon^3)$ samples (up to a log factor) to attain a near-stationary solution with expected squared gradient norm less than ϵ. Numerical experiments verify our analysis.

1. Introduction

Consider the following stochastic optimization problem with decision-dependent data:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{Z\sim\Pi_{\boldsymbol{\theta}}}\big[\ell(\boldsymbol{\theta};Z)\big]. \tag{1}$$

Notice that the decision variable $\boldsymbol{\theta}$ appears in both the loss function $\ell(\boldsymbol{\theta};Z)$ and the data distribution $\Pi_{\boldsymbol{\theta}}$ (denoted by $D(\boldsymbol{\theta})$ in some prior literature) supported on Z. The overall loss function $\mathcal{L}(\boldsymbol{\theta})$ is known as the *performative risk* which captures the distributional shift due to changes in the deployed model. This setting is motivated by the recent studies on *performative prediction* (Perdomo et al., 2020), which considers outcomes Z that are supported by the deployed model $\boldsymbol{\theta}$. For example, it covers strategic classification (Hardt et al., 2016; Dong et al., 2018) in economics and financial practices such as with the training of loan classifier for customers who may react to the deployed model $\boldsymbol{\theta}$ to maximize their gains; or in price promotion mechanism (Zhang et al., 2018) where customers react to prices with the aim of gaining a lower price; or in ride sharing business (Narang et al., 2022) with customers who adjust their demand according to prices set by the platform.

Due to the effects of $\boldsymbol{\theta}$ on both the loss function and distribution, the objective function $\mathcal{L}(\boldsymbol{\theta})$ is non-convex in general. Numerous efforts have focused on characterizing and finding the so-called *performative stable* solution which is a fixed point to the repeated risk minimization (RRM) process (Perdomo et al., 2020; Mendler-Dünner et al., 2020; Brown et al., 2022; Li & Wai, 2022; Roy et al., 2022; Drusvyatskiy & Xiao, 2022). While RRM might be a natural algorithm for scenarios when the learner is agnostic to the performative effects in the dynamic data distribution, the obtained solution maybe far from being optimal or stationary to (1).

On the other hand, recent works have studied how to approximate *performative optimal* solutions that minimizes (1). This is challenging due to the non-convexity of $\mathcal{L}(\boldsymbol{\theta})$ and more importantly, the absence of knowledge of $\Pi_{\boldsymbol{\theta}}$. In fact, evaluating $\nabla\mathcal{L}(\boldsymbol{\theta})$ or its stochastic gradient estimate would require learning the distribution $\Pi_{\boldsymbol{\theta}}$ *a-priori* (Izzo et al., 2021). To design a tractable procedure, prior works have assumed additional structures for (1) such as approximating $\Pi_{\boldsymbol{\theta}}$ by Gaussian mixture (Izzo et al., 2021), $\Pi_{\boldsymbol{\theta}}$ depends linearly on $\boldsymbol{\theta}$ (Narang et al., 2022), etc., combined with a two-phase algorithm that separately learns $\Pi_{\boldsymbol{\theta}}$ and optimizes $\boldsymbol{\theta}$. Other works have assumed a *mixture dominance* structure (Miller et al., 2021) on the combined effect of $\Pi_{\boldsymbol{\theta}}$ and $\ell(\cdot)$ on $\mathcal{L}(\boldsymbol{\theta})$, which in turn implies that $\mathcal{L}(\boldsymbol{\theta})$ is convex. Based on this assumption, a derivative free optimization (DFO) algorithm was analyzed in (Ray et al., 2022); also see the variants of this condition when used in different settings (Wood & Dall'Anese, 2023; Zhu et al., 2023). In addition,

[1]Department of Computer Science, ETH Zurich. The research was primarily conducted as a student at The Chinese University of Hong Kong. [2]Department of System Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong, China. Correspondence to: Hoi-To Wai <htwai@se.cuhk.edu.hk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Literature	Cvx-ℓ	Cvx-$\mathcal{L}$	Dec-dep	State-dep	Oracle	Rate	θ_∞-Type
(Ghadimi & Lan, 2013)	✗	✗	✗	✗	0th	$\mathcal{O}(T^{-\frac{1}{2}})$	Stationary
(Miller et al., 2021)	✓	✓	✓	✗	$1^{st\dagger}$	$\mathcal{O}(T^{-1})$	Perf. Opt.
(Ray et al., 2022)	✓	✓	✓	✓(Geo)	0th	$\mathcal{O}(T^{-\frac{1}{2}})$	Perf. Opt.
(Mendler-Dünner et al., 2020)	✓	✗	✓	✗	1st	$\mathcal{O}(T^{-1})$	Perf. Stable
(Brown et al., 2022)	✓	✗	✓	✓(Geo)	1st	$\mathcal{O}(T^{-1})$	Perf. Stable
(Li & Wai, 2022)	✓	✗	✓	✓(Mkv)	1st	$\mathcal{O}(T^{-1})$	Perf. Stable
(Izzo et al., 2021)	✗	✗	✓	✓(Geo)	$1^{st\dagger}$	$\mathcal{O}(T^{-\frac{1}{5}})$	Stationary
(Roy et al., 2022)	✗	✗	✓	✓(Mkv)	$1^{st\ddagger}$	$\mathcal{O}(T^{-\frac{2}{5}})$	Stationary
This Work	✗	✗	✓	✓(Mkv)	0th	$\mathcal{O}(T^{-\frac{1}{3}})$	Stationary

Table 1. **Comparison to Existing Literature**. 'θ_∞-Type' describes the convergent points – *Perf. stable*: Def. 2 of Mendler et al., *Perf. opt.*: solves $\min_\theta \mathcal{L}(\theta)$, *Stationary*: $\|\nabla\mathcal{L}(\theta)\| = 0$. 'Geo'/'Mkv': geometric decay/Markov. The rates sometimes omit the logarithmic terms. †uses two-stages to estimate Π_θ with linear distribution model. ‡needs asymp. unbiased estimate of $\nabla\mathcal{L}(\theta)$ [cf. (3)].

(Kim & Perdomo, 2023) considered reaching performative optimal solution where performativity is only modeled at the label, several works have initiated the analysis of performative prediction with general non-convex loss (Dong et al., 2023; Mofakhami et al., 2023).

This paper focuses on approximating the *performative optimal* solution without relying on additional condition on the distribution $\Pi_{\boldsymbol{\theta}}$ and/or using a two-phase algorithm. We concentrate on stochastic DFO algorithms (Ghadimi & Lan, 2013) which do not involve first order information (i.e., gradient) about $\mathcal{L}(\boldsymbol{\theta})$. An advantage of these algorithms is that they avoid the need for estimating $\Pi_{\boldsymbol{\theta}}$ nor making structural assumption on the latter. Instead, the learner only requires access to the loss function evaluation oracle $\ell(\boldsymbol{\theta}; Z)$ and receive data samples from a controlled Markov chain. Note that the latter models the *stateful* and *strategic* agent setting considered in (Ray et al., 2022; Roy et al., 2022; Li & Wai, 2022; Brown et al., 2022). Such setting is motivated when the actual data distribution adapts slowly to the decision model, which is to be deployed/announced by the learner during the stochastic optimization process.

The proposed $\texttt{DFO}\,(\lambda)$ algorithm features (i) two-timescale step sizes design to control the bias-variance tradeoff in the derivative-free gradient estimates, and (ii) a sample accumulation mechanism with forgetting factor λ that aggregates all observed samples to control the amount of error in gradient estimates. Our findings are summarized as:

- Under the Markovian data setting, we show in Theorem 3.6 that the $\texttt{DFO}\,(\lambda)$ algorithm finds a near-stationary solution $\bar{\boldsymbol{\theta}}$ with $\mathbb{E}[\|\nabla\mathcal{L}(\bar{\boldsymbol{\theta}})\|^2] \leq \epsilon$ using $\mathcal{O}(\frac{d^2}{\epsilon^3}\log 1/\epsilon)$ samples and iterations. Compared to prior works, our analysis does not require structural assumption on the distribution $\Pi_{\boldsymbol{\theta}}$ or convexity condition on the performative risk (Izzo et al., 2021; Miller et al., 2021; Ray et al., 2022).
- Our analysis demonstrates the trade-off induced by the forgetting factor λ. We identify the desiderata for the optimal value(s) of λ. We show that increasing λ allows to reduce the number of samples required by the algorithm if the performative risk gradient has a small Lipschitz constant.

For the rest of this paper, §2 describes the problem setup and the $\texttt{DFO}\,(\lambda)$ algorithm, §3 presents the main results, §4 outlines the proofs. Finally, we provide numerical results to verify our findings in §5.

Finally, as displayed in Table 1, for stationary points convergence of (1), our stochastic DFO under *decision dependent* (and Markovian) samples has a convergence rate of $\mathcal{O}(1/T^{\frac{1}{3}})$ towards an ϵ-stationary point, while other known gradient-based methods, e.g., (Izzo et al., 2021) has a convergence rate of $\mathcal{O}(1/T^{\frac{1}{5}})$, and (Roy et al., 2022) has a convergence rate of $\mathcal{O}(1/T^{\frac{2}{5}})$. We also include other relevant results in the table.

We remark that our rate is worse than works on the decision independent (i.e., non-performative) setting, e.g., $\mathcal{O}(1/\sqrt{T})$ in (Ghadimi & Lan, 2013). We believe the rate degradation is due to a fundamental limit for DFO-type algorithms when tackling problems with decision-dependent sample due to the challenges in designing a low variance gradient estimator (including 2-point estimators); see §4.1.

Related Works: The idea of DFO dates back to (Nemirovskiĭ, 1983), and has been extensively studied thereafter (Flaxman et al., 2005; Agarwal et al., 2010; Nesterov & Spokoiny, 2017; Ghadimi & Lan, 2013). Results on matching lower bound were established in (Jamieson et al., 2012). While a similar DFO framework is adopted in the current paper for performative prediction, our algorithm is limited to using a special design in the gradient estimator to avoid introducing unwanted biases.

Only a few works have considered the Markovian data setting in performative prediction. (Brown et al., 2022) is the

first paper to study the dynamic settings, where the response of agents to learner's deployed classifier is modeled as a function of classifier and the current distribution of the population; also see (Izzo et al., 2022). On the other hand, Li & Wai (2022); Roy et al. (2022) model the unforgetful nature and the reliance on past experiences of *single/batch* agent(s) via controlled Markov Chain. Lastly, Ray et al. (2022) investigated the state-dependent framework where agents' response may be driven to best response at a geometric rate. The current paper considers more relaxed conditions than Ray et al. (2022). See Appendix A for a detailed discussion.

Notations: Let $\mathbb{R}^d$ be the d-dimensional Euclidean space equipped with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\|x\| = \sqrt{\langle x, x \rangle}$. Let Z be a (measurable) sample space, $\mathcal{Z}$ be the Borel σ-algebra generated by Z, and μ, ν are two probability measures on $\mathcal{Z}$. Then, we use $\boldsymbol{\delta}_{\text{TV}}(\mu, \nu) := \sup_{A \subset \mathsf{Z}} \mu(A) - \nu(A)$ to denote the total variation distance between μ and ν. Denote $\mathbb{T}_{\boldsymbol{\theta}}(\cdot, \cdot)$ as the state-dependent Markov kernel and its stationary distribution is $\Pi_{\boldsymbol{\theta}}(\cdot)$. Let $\mathbb{B}^d$ and $\mathbb{S}^{d-1}$ be the unit ball and its boundary (i.e., a unit sphere) centered around the origin in d-dimensional Euclidean space, respectively, and correspondingly, the ball and sphere of radius $r > 0$ are $r\mathbb{B}^d$ and $r\mathbb{S}^{d-1}$.

2. Problem Setup, Algorithm Design

In this section, we develop the $\texttt{DFO}(\lambda)$ algorithm for tackling (1) and describe the problem setup. Assume that $\mathcal{L}(\boldsymbol{\theta})$ is differentiable, we focus on finding an ϵ-*stationary* solution, $\boldsymbol{\theta}$, which satisfies

$$\|\nabla \mathcal{L}(\boldsymbol{\theta})\|^2 \leq \epsilon. \tag{2}$$

With the goal of reaching (2), there are two key challenges in our stochastic algorithm design: (i) to estimate the gradient $\nabla \mathcal{L}(\boldsymbol{\theta})$ without prior knowledge of $\Pi_{\boldsymbol{\theta}}$, and (ii) to handle the *stateful* setting where one cannot draw samples directly from the distribution $\Pi_{\boldsymbol{\theta}}$. We shall discuss how the proposed $\texttt{DFO}(\lambda)$ algorithm, which is summarized in Algorithm 1, tackles the above issues through utilizing two ingredients: (a) two-timescales step sizes, and (b) sample accumulation with the forgetting factor $\lambda \in [0, 1)$.

Estimating $\nabla \mathcal{L}(\boldsymbol{\theta})$ via Two-timescales DFO. First notice that the gradient of $\mathcal{L}(\cdot)$ can be derived as

$$\nabla \mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{Z \sim \Pi_{\boldsymbol{\theta}}}[\nabla \ell(\boldsymbol{\theta}; Z) + \ell(\boldsymbol{\theta}; Z) \nabla_{\boldsymbol{\theta}} \log \Pi_{\boldsymbol{\theta}}(Z)]. \tag{3}$$

As a result, constructing stochastic estimates of $\nabla \mathcal{L}(\boldsymbol{\theta})$ typically requires knowledge of $\Pi_{\boldsymbol{\theta}}(\cdot)$ which is not known a-priori unless a separate estimation procedure is applied; see e.g., (Izzo et al., 2021). To avoid the need for direct evaluations of $\nabla_{\boldsymbol{\theta}} \log \Pi_{\boldsymbol{\theta}}(Z)$, we consider an alternative design via zero-th order optimization (Ghadimi & Lan, 2013). The intuition comes from observing that with $\delta \to 0^+$,

Algorithm 1 $\texttt{DFO}(\lambda)$ Algorithm

Input: Constants $\delta_0, \eta_0, \tau_0, \alpha, \beta$, maximum epochs T, forgetting factor λ, loss function $\ell(\cdot; \cdot)$.
Initialization: Set initial $\boldsymbol{\theta}_0$ and sample Z_0.
for $k = 0$ **to** $T - 1$ **do**
  $\delta_k \leftarrow \delta_0/(1+k)^{\beta}$, $\eta_k \leftarrow \eta_0/(1+k)^{\alpha}$, $\tau_k \leftarrow \max\{1, \tau_0 \log(1+k)\}$
  Let $\boldsymbol{\theta}_k^{(1)} \leftarrow \boldsymbol{\theta}_k$, $Z_k^{(0)} \leftarrow Z_k$, $\boldsymbol{u}_k \sim \text{Unif}(\mathbb{S}^{\text{d}-1})$
  for $m = 1, 2, \cdots, \tau_k$ **do**
    Deploy the model $\check{\boldsymbol{\theta}}_k^{(m)} = \boldsymbol{\theta}_k^{(m)} + \delta_k \boldsymbol{u}_k$
    Draw $Z_k^{(m)} \sim \mathbb{T}_{\check{\boldsymbol{\theta}}_k^{(m)}}(Z_k^{(m-1)}, \cdot)$
    Update $\boldsymbol{\theta}_k^{(m)}$ as

$$\boldsymbol{g}_k^{(m)} = \tfrac{d}{\delta_k} \ell\big(\check{\boldsymbol{\theta}}_k^{(m)}; Z_k^{(m)}\big) \boldsymbol{u}_k,$$
$$\boldsymbol{\theta}_k^{(m+1)} = \boldsymbol{\theta}_k^{(m)} - \eta_k \lambda^{\tau_k - m} \boldsymbol{g}_k^{(m)}.$$

  end for
  $Z_{k+1} \leftarrow Z_k^{(\tau_k)}$, $\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k^{(\tau_k+1)}$.
end for
Draw $s \sim \text{Uniform}(\{0, 1, \ldots, T\})$
Output: Average iterate $\boldsymbol{\theta}_s$,

$(\mathcal{L}(\boldsymbol{\theta} + \delta \boldsymbol{u}) - \mathcal{L}(\boldsymbol{\theta}))/\delta$ is an approximation of the directional derivative of $\mathcal{L}$ along $\boldsymbol{u}$. This suggests that an estimate for $\nabla \mathcal{L}(\boldsymbol{\theta})$ can be constructed using the *objective function values* of $\ell(\boldsymbol{\theta}; Z)$ only.

Inspired by the above, we aim to construct a gradient estimate by querying $\ell(\cdot)$ at randomly perturbed points. Formally, given the current iterate $\boldsymbol{\theta} \in \mathbb{R}^d$ and a query radius $\delta > 0$, we sample a vector $\boldsymbol{u} \in \mathbb{R}^d$ uniformly from the sphere $\mathbb{S}^{d-1}$. The zero-th order gradient estimator for $\mathcal{L}(\boldsymbol{\theta})$ is then defined as

$$g_\delta(\boldsymbol{\theta}; \boldsymbol{u}, Z) := \frac{d}{\delta} \ell(\check{\boldsymbol{\theta}}; Z)\, \boldsymbol{u}, \text{ with } \check{\boldsymbol{\theta}} := \boldsymbol{\theta} + \delta \boldsymbol{u}, \tag{4}$$

and $Z \sim \Pi_{\check{\boldsymbol{\theta}}}(\cdot)$. In fact, as $\boldsymbol{u}$ is zero-mean, $g_\delta(\boldsymbol{\theta}; \boldsymbol{u}, Z)$ is an unbiased estimator for $\nabla \mathcal{L}_\delta(\boldsymbol{\theta})$. Here, $\mathcal{L}_\delta(\boldsymbol{\theta})$ is a smooth approximation of $\mathcal{L}(\boldsymbol{\theta})$ (Flaxman et al., 2005; Nesterov & Spokoiny, 2017) defined as

$$\mathcal{L}_\delta(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{u}}[\mathcal{L}(\check{\boldsymbol{\theta}})] = \mathbb{E}_{\boldsymbol{u}}[\mathbb{E}_{Z \sim \Pi_{\check{\boldsymbol{\theta}}}}[\ell(\check{\boldsymbol{\theta}}; Z)]]. \tag{5}$$

It is known that under mild condition [cf. Assumption 3.1 to be discussed later], $\|\nabla \mathcal{L}_\delta(\boldsymbol{\theta}) - \nabla \mathcal{L}(\boldsymbol{\theta})\| = \mathcal{O}(\delta)$ and thus (4) is an $\mathcal{O}(\delta)$-biased estimate for $\nabla \mathcal{L}(\boldsymbol{\theta})$.

We remark that the gradient estimator in (4) differs from the one used in classical works on DFO such as (Ghadimi & Lan, 2013). The latter takes the form of $\frac{d}{\delta}(\ell(\check{\boldsymbol{\theta}}; Z) - \ell(\boldsymbol{\theta}; Z))\, \boldsymbol{u}$. Under the setting of standard stochastic optimization where the sample Z is drawn *independently* of $\boldsymbol{u}$ and Lipschitz

continuous $\ell(\cdot; Z)$, the said estimator in (Ghadimi & Lan, 2013) is shown to have constant variance while it remains $\mathcal{O}(\delta)$-biased. Such properties *cannot* be transferred to (4) since Z is drawn from a distribution dependent on $\boldsymbol{u}$ via $\check{\boldsymbol{\theta}} = \boldsymbol{\theta} + \delta \boldsymbol{u}$. In this case, the two-point gradient estimator would become biased; see §4.1.

However, we note that the variance of (4) scales as $\mathcal{O}(1/\delta^2)$ when $\delta \to 0$, thus the parameter δ yields a bias-variance trade off in the estimator design. To remedy for the increase of variance, the DFO (λ) algorithm incorporates a *two-timescale step size* design for generating gradient estimates (δ_k) and updating models (η_k), respectively. Our design principle is such that the models are updated at a *slower timescale* to adapt to the gradient estimator with $\mathcal{O}(1/\delta^2)$ variance. Particularly, we will set $\eta_{k+1}/\delta_{k+1} \to 0$ to handle the bias-variance trade off, e.g., by setting $\alpha > \beta$ in line 4 of Algorithm 1.

Markovian Data and Sample Accumulation. We consider a setting where the sample/data distribution observed by the DFO (λ) algorithm evolves according to a *controlled Markov chain (MC)*. Notice that this describes a stateful agent(s) scenario such that the deployed models ($\boldsymbol{\theta}$) would require a certain amount of time to manifest their influence on the samples obtained; see (Li & Wai, 2022; Roy et al., 2022; Brown et al., 2022; Ray et al., 2022; Izzo et al., 2022).

To formally describe the setting, we denote $\mathbb{T}_{\boldsymbol{\theta}} : \mathsf{Z} \times \mathcal{Z} \to \mathbb{R}_+$ as a Markov kernel controlled by a deployed model $\boldsymbol{\theta}$. For a given $\boldsymbol{\theta}$, the kernel has a unique stationary distribution $\Pi_{\boldsymbol{\theta}}(\cdot)$ which cannot be conveniently accessed. Under this setting, suppose that the previous state/sample is Z, the next sample follows the distribution $Z' \sim \mathbb{T}_{\boldsymbol{\theta}}(Z, \cdot)$ which is not necessarily the same as $\Pi_{\boldsymbol{\theta}}(\cdot)$. As a consequence, the gradient estimator (4) is a biased estimator of $\nabla \mathcal{L}_\delta(\boldsymbol{\theta})$.

A common strategy in settling the above issue is to allow a *burn-in* phase in the algorithm as in (Ray et al., 2022); also commonly found in MCMC methods (Robert et al., 1999). Using the fact that $\mathbb{T}_{\boldsymbol{\theta}}$ admits the stationary distribution $\Pi_{\boldsymbol{\theta}}$, if one can wait a sufficiently long time before applying the current sample, i.e., consider initializing with the previous sample $Z^{(0)} = Z$, the procedure

$$Z^{(m)} \sim \mathbb{T}_{\boldsymbol{\theta}}(Z^{(m-1)}, \cdot),\ m = 1, \ldots, \tau, \tag{6}$$

would yield a sample $Z^+ = Z^{(\tau)}$ that admits a distribution close to $\Pi_{\boldsymbol{\theta}}$ provided that $\tau \gg 1$ is sufficiently large compared to the mixing time of $\mathbb{T}_{\boldsymbol{\theta}}$.

Intuitively, the procedure (6) can be inefficient as a number of samples $Z^{(1)}, Z^{(2)}, \ldots, Z^{(\tau-1)}$ will be completely ignored at the end of each iteration. As a remedy, the DFO (λ) algorithm incorporates a sample accumulation mechanism which gathers the gradient estimates generated from possibly non-stationary samples via a forgetting factor of $\lambda \in [0, 1)$. Following (4), $\nabla \mathcal{L}(\boldsymbol{\theta})$ is estimated by

$$\boldsymbol{g} = \tfrac{d}{\delta} \textstyle\sum_{m=1}^{\tau} \lambda^{\tau-m} \ell(\boldsymbol{\theta}^{(m)} + \delta \boldsymbol{u}; Z^{(m)})\, \boldsymbol{u}, \tag{7}$$

with $Z^{(m)} \sim \mathbb{T}_{\boldsymbol{\theta}^{(m)}+\delta \boldsymbol{u}}(Z^{(m-1)}, \cdot)$ being generated in a recursive manner according to the MC. At a high level, the mechanism works by assigning large weights to samples that are close to the end of an epoch (which are less biased). Moreover, $\boldsymbol{\theta}^{(m)}$ is *simultaneously updated* within the epoch to obtain an online algorithm that gradually improves the objective value of (1). Note that with $\lambda = 0$, the DFO(0) algorithm reduces into one that utilizes *burn-in* (6). We remark that from the implementation perspective for performative prediction, Algorithm 1 corresponds to a *greedy deployment* scheme (Mendler-Dünner et al., 2020) as the latest model $\boldsymbol{\theta}_k^{(m)} + \delta_k \boldsymbol{u}_k$ is deployed at every sampling step. Line 6–10 of Algorithm 1 details the above procedure.

Lastly, we note that recent works have analyzed stochastic algorithms that rely on a *single trajectory* of samples taken from a Markov Chain, e.g., (Sun et al., 2018; Karimi et al., 2019; Doan, 2022), that are based on stochastic gradient. (Sun & Li, 2019) considered a DFO algorithm for general optimization problems but the MC studied is not controlled by $\boldsymbol{\theta}$.

3. Main Results

This section studies the convergence of the DFO (λ) algorithm and demonstrates that the latter finds an ϵ-stationary solution [cf. (2)] to (1). We first state the assumptions required for our analysis:

Assumption 3.1. (Smoothness) $\mathcal{L}(\boldsymbol{\theta})$ is differentiable, and there exists a constant $L > 0$ such that

$$\|\nabla \mathcal{L}(\boldsymbol{\theta}) - \nabla \mathcal{L}(\boldsymbol{\theta}')\| \le L \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|,\ \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d.$$

Assumption 3.2. (Bounded Loss) There exists a constant $G > 0$ such that

$$|\ell(\boldsymbol{\theta}; z)| \le G,\ \forall\, \boldsymbol{\theta} \in \mathbb{R}^d,\ \forall\, z \in \mathsf{Z}.$$

Assumption 3.3. (Lipschitz Distribution Map) There exists a constant $L_1 > 0$ such that

$$\boldsymbol{\delta}_{\mathrm{TV}}\left(\Pi_{\boldsymbol{\theta}_1}, \Pi_{\boldsymbol{\theta}_2}\right) \le L_1 \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \quad \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d.$$

Informally, the conditions above state that the gradient of the performative risk is Lipschitz continuous and the state-dependent distribution vary smoothly w.r.t. $\boldsymbol{\theta}$. Note that Assumption 3.1 and Assumption 3.2 are both regularity conditions that can also be found in (Izzo et al., 2021; Ray et al., 2022). Assumption 3.3 is slightly strengthened from the Wasserstein-1 distance bound in (Perdomo et al., 2020), and it gives tighter control for distribution shift in our Markovian data setting. Note that this particular condition can be slightly relaxed, see Lemma F.7 in the appendix.

Next, we consider the assumptions about the controlled Markov chain induced by $\mathbb{T}_{\boldsymbol{\theta}}$:

Assumption 3.4. (Geometric Mixing) Let $\{Z_k\}_{k\geq 0}$ denote a Markov Chain on the state space Z with transition kernel $\mathbb{T}_{\boldsymbol{\theta}}$ and stationary measure $\Pi_{\boldsymbol{\theta}}$. There exist constants $\rho \in [0,1)$, $M \geq 0$, such that for any $k \geq 0$, $z \in \mathsf{Z}$,

$$\boldsymbol{\delta}_{\text{TV}}\left(\mathbb{P}_{\boldsymbol{\theta}}(Z_k \in \cdot | Z_0 = z), \Pi_{\boldsymbol{\theta}}\right) \leq M\rho^k.$$

Assumption 3.5. (Smoothness of Markov Kernel) There exists a constant $L_2 \geq 0$ such that

$$\boldsymbol{\delta}_{\text{TV}}\left(\mathbb{T}_{\boldsymbol{\theta}_1}(z,\cdot), \mathbb{T}_{\boldsymbol{\theta}_2}(z,\cdot)\right) \leq L_2 \left\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\right\|,$$

holds for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d,\ z \in \mathsf{Z}$.

Assumption 3.4 is a standard condition on the mixing time of the Markov chain induced by $\mathbb{T}_{\boldsymbol{\theta}}$; Assumption 3.5 imposes a smoothness condition on the Markov transition kernel $\mathbb{T}_{\boldsymbol{\theta}}$ with respect to $\boldsymbol{\theta}$. The geometric dynamically environment in (Ray et al., 2022) constitutes a special case which satisfies the above conditions, yet Assumption 3.4 is strictly more general, e.g., it covers cases when $(Z_k)_{k\geq 0}$ follows an autoregressive (AR) process.

Unlike (Ray et al., 2022; Izzo et al., 2021; Miller et al., 2021), other than Assumption 3.3, we do not impose any additional assumption on $\Pi_{\boldsymbol{\theta}}$ such as mixture dominance. As a result, (1) remains an 'unstructured' non-convex stochastic optimization problem with decision-dependent distribution. Our main result on the convergence of the $\texttt{DFO}(\lambda)$ algorithm towards a near-stationary solution of (1) is summarized below:

Theorem 3.6. *Suppose Assumptions 3.1-3.5 hold, step size sequence $\{\eta_k\}_{k\geq 1}$, and query radius sequence $\{\delta_k\}_{k\geq 1}$ satisfy the following conditions,*

$$\begin{aligned} &\eta_k = d^{-2/3}\cdot(1+k)^{-2/3}, \quad \delta_k = d^{1/3}\cdot(1+k)^{-1/6}, \\ &\tau_k = \max\{1, \frac{2}{\log 1/\max\{\rho,\lambda\}}\log(1+k)\} \quad \forall k \geq 0. \end{aligned} \tag{8}$$

Then, there exists constants t_0, c_5, c_6, c_7, such that for any $T \geq t_0$, the iterates $\{\boldsymbol{\theta}_k\}_{k\geq 0}$ generated by $\texttt{DFO}(\lambda)$ satisfy the following inequality,

$$\begin{aligned} &\frac{1}{1+T}\sum_{k=0}^{T}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\|^2 \\ &\leq 12\max\left\{c_5(1-\lambda), c_6, \frac{c_7}{1-\lambda}\right\}\frac{d^{2/3}}{(T+1)^{1/3}}. \end{aligned} \tag{9}$$

We have defined the following quantities and constants: $c_5 = 2G$,

$$c_6 = \frac{\max\{L^2, G^2(1-\beta)\}}{1-2\beta},\ c_7 = \frac{LG^2}{2\beta-\alpha+1}, \tag{10}$$

with $\alpha = \frac{2}{3}, \beta = \frac{1}{6}$. Observe the following corollary on the iteration complexity of $\texttt{DFO}(\lambda)$ algorithm:

Corollary 3.7. *(ϵ-stationarity) Suppose that the Assumptions of Theorem 3.6 hold. Fix any $\epsilon > 0$, the condition $\frac{1}{1+T}\sum_{k=0}^{T}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\|^2 \leq \epsilon$ holds whenever*

$$T \geq \left(12\max\left\{c_5(1-\lambda), c_6, \frac{c_7}{1-\lambda}\right\}\right)^3\frac{d^2}{\epsilon^3}. \tag{11}$$

In the corollary above, the lower bound on T is expressed in terms of the number of epochs that Algorithm 1 needs to achieve the target accuracy. Consequently, the total number of samples required (i.e., the number of inner iterations taken in Line 6–9 of Algorithm 1 across all epochs) is:

$$\mathsf{S}_\epsilon = \sum_{k=1}^{T}\tau_k = \mathcal{O}\left(\frac{d^2}{\epsilon^3}\log(1/\epsilon)\right). \tag{12}$$

We remark that due to the decision-dependent properties of the samples, the $\texttt{DFO}(\lambda)$ algorithm exhibits a worse sampling complexity (12) than prior works in stochastic DFO algorithm, e.g., (Ghadimi & Lan, 2013) which shows a rate of $\mathcal{O}(d/\epsilon^2)$ on non-convex smooth objective functions. In particular, the adopted one-point gradient estimator in (4) admits a variance that can only be controlled by a time varying δ; see the discussions in §4.1.

Achieving the desired convergence rate requires setting $\eta_k = \Theta(k^{-2/3})$, $\delta_k = \Theta(k^{-1/6})$, i.e., yielding a two-timescale step sizes design with $\eta_k/\delta_k \to 0$. Notice that the influence of forgetting factor λ are reflected in the constant factor of (9). Particularly, if $c_5 > c_7$ and $c_5 \geq c_6$, the optimal choice is $\lambda = 1 - \sqrt{\frac{c_7}{c_5}}$, otherwise the optimal choice is $\lambda \in [0, 1 - c_7/c_6]$. Informally, this indicates that when the performative risk is smoother (i.e. its gradient has a small Lipschitz constant), a large λ can speed up the convergence of the algorithm; otherwise a smaller λ is preferable.

4. Proof of Main Results

This section outlines the key steps in proving Theorem 3.6. Notice that analyzing the $\texttt{DFO}(\lambda)$ algorithm is challenging due to the two-timescales step sizes and Markov chain samples with time varying kernel. Our analysis departs from prior works such as (Ray et al., 2022; Izzo et al., 2021; Brown et al., 2022; Li & Wai, 2022) to handle the above technicalities.

Let $\mathcal{F}^k = \sigma(\boldsymbol{\theta}_0, Z_s^{(m)}, u_s, 0 \leq s \leq k, 0 \leq m \leq \tau_k)$ be the filtration. Our first step is to exploit the smoothness of $\mathcal{L}(\boldsymbol{\theta})$ to bound the squared norms of gradient. Observe that:

Lemma 4.1. (Decomposition) *Under Assumption 3.1, it holds that*

$$\sum_{k=0}^{t}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\|^2 \leq \mathbf{I}_1(t) + \mathbf{I}_2(t) + \mathbf{I}_3(t) + \mathbf{I}_4(t), \tag{13}$$

for any $t \geq 1$, where

$$\mathbf{I}_1(t) := \sum_{k=1}^{t} \frac{1-\lambda}{\eta_k} \left(\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_k)\right] - \mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{k+1})\right]\right)$$

$$\mathbf{I}_2(t) := -\sum_{k=1}^{t} \mathbb{E}\Big\langle \nabla\mathcal{L}(\boldsymbol{\theta}_k)\Big|(1-\lambda)\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m} \cdot \left(g_k^{(m)} - \mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}\left[g_{\delta_k}(\boldsymbol{\theta}_k;u_k,Z)\right]\right)\Big\rangle$$

$$\mathbf{I}_3(t) := -\sum_{k=1}^{t} \mathbb{E}\Big\langle \nabla\mathcal{L}(\boldsymbol{\theta}_k)\Big| (1-\lambda)\left(\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\nabla\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)\right) - \nabla\mathcal{L}(\boldsymbol{\theta}_k)\Big\rangle$$

$$\mathbf{I}_4(t) := \frac{L(1-\lambda)}{2}\sum_{k=1}^{t}\eta_k\mathbb{E}\left\|\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}g_k^{(m)}\right\|^2$$

The lemma is achieved through the standard descent lemma implied by Assumption 3.1 and decomposing the upper bound on $||\nabla\mathcal{L}(\boldsymbol{\theta}_k)||^2$ into respectful terms; see the proof in Appendix B. Among the terms on the right hand side of (13), $\mathbf{I}_1(t), \mathbf{I}_3(t)$ and $\mathbf{I}_4(t)$ arises directly from Assumption 3.1, while $\mathbf{I}_2(t)$ comes from bounding the noise terms due to Markovian data.

We bound the four components in Lemma 4.1 as follows. For simplicity, we denote $\mathcal{A}(t) := \frac{1}{1+t}\sum_{k=0}^{t}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\|^2$. Among the four terms, we highlight that the main challenge lies on obtaining a tight bound for $\mathbf{I}_2(t)$. Observe that

$$\frac{\mathbf{I}_2(t)}{1-\lambda} \leq \mathbb{E}\left[\sum_{k=0}^{t}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|\left\|\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\Delta_{k,m}\right\|\right], \tag{14}$$
$$\Delta_{k,m} \stackrel{\text{def}}{=} \mathbb{E}_{\mathcal{F}^{k-1}}[g_k^{(m)} - \mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}} g_k(\boldsymbol{\theta}_k;u_k,Z)].$$

There are two sources of bias in $\Delta_{k,m}$: one is the noise induced by drifting of decision variable within each epoch, the other is the bias that depends on the mixing time of Markov kernel.

To control these biases, we introduce a reference Markov chain $\tilde{Z}_k^{(\ell)}$, $\ell = 0, ..., \tau_k$, whose decision variables remains fixed for a period of length τ_k and is initialized with $\tilde{Z}_k^{(0)} = Z_k^{(0)}$:

$$\tilde{Z}_k^{(0)} \xrightarrow{\check{\boldsymbol{\theta}}_k} \tilde{Z}_k^{(1)} \xrightarrow{\check{\boldsymbol{\theta}}_k} \tilde{Z}_k^{(2)} \xrightarrow{\check{\boldsymbol{\theta}}_k} \tilde{Z}_k^{(3)} \cdots \xrightarrow{\check{\boldsymbol{\theta}}_k} \tilde{Z}_k^{(\tau_k)} \tag{15}$$

and we recall that the actual chain in the algorithm evolves as

$$Z_k^{(0)} \xrightarrow{\check{\boldsymbol{\theta}}_{k+1}^{(0)}} Z_k^{(1)} \xrightarrow{\check{\boldsymbol{\theta}}_{k+1}^{(1)}} Z_k^{(2)} \cdots \xrightarrow{\check{\boldsymbol{\theta}}_{k+1}^{(\tau_k-1)}} Z_k^{(\tau_k)}. \tag{16}$$

Note the reference Markov chain idea is inspired by (Wu et al., 2020, Theorem 4.7) which studied the convergence of a strongly convex subproblem.

With the help of the reference chain, we decompose the conditional expectation $\Delta_{k,m}$ as:

$$\begin{aligned}
\Delta_{k,m} &= \mathbb{E}_{\mathcal{F}^{k-1}}\Bigg[\frac{d}{\delta_k}\Bigg(\mathbb{E}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};Z_k^{(m)})|\check{\boldsymbol{\theta}}_k^{(m)},Z_k^{(0)}] \\
&\qquad - \mathbb{E}_{\tilde{Z}_k^{(m)}}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};\tilde{Z}_k^{(m)})|\check{\boldsymbol{\theta}}_k^{(m)},\tilde{Z}_k^{(0)}]\Bigg)\boldsymbol{u}_k\Bigg] \\
&+ \mathbb{E}_{\mathcal{F}^{k-1}}\Bigg[\frac{d}{\delta_k}\Bigg(\mathbb{E}_{\tilde{Z}_k^{(m)}}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};\tilde{Z}_k^{(m)})|\check{\boldsymbol{\theta}}_k^{(m)},\tilde{Z}_k^{(0)}] \\
&\qquad - \mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};Z)|\check{\boldsymbol{\theta}}_k^{(m)}]\Bigg)\boldsymbol{u}_k\Bigg] \\
&+ \mathop{\mathbb{E}}_{\mathcal{F}^{k-1}}\Bigg[\frac{d}{\delta_k}\mathop{\mathbb{E}}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}\left[\ell(\check{\boldsymbol{\theta}}_k^{(m)};Z) - \ell(\check{\boldsymbol{\theta}}_k;Z)|\check{\boldsymbol{\theta}}_k^{(m)},\check{\boldsymbol{\theta}}_k\right]\boldsymbol{u}_k\Bigg] \\
&:= A_1 + A_2 + A_3
\end{aligned}$$

We remark that A_1 reflects the drift of (16) from initial sample $Z_k^{(0)}$ driven by varying $\check{\boldsymbol{\theta}}_k^{(m)}$, A_2 captures the statistical discrepancy between above two Markov chains (16) and (15) at same step m, and A_3 captures the drifting gap between $\check{\boldsymbol{\theta}}_k$ and $\check{\boldsymbol{\theta}}_k^{(m)}$. Applying Assumption 3.3, A_1 and A_2 can be upper bounded with the smoothness and geometric mixing property of Markov kernel. In addition, A_3 can be upper bounded using Lipschitz condition on (stationary) distribution map $\Pi_{\boldsymbol{\theta}}$. Finally, the forgetting factor λ helps to control $\|\check{\boldsymbol{\theta}}_k^{(\cdot)} - \check{\boldsymbol{\theta}}_k\|$ to be at the same order of a single update. Therefore, $\|\Delta_{k,m}\|$ can be controlled by an upper bound relying on λ, ρ, L.

The following lemma summarizes the above results as well as the bounds on the other terms:

Lemma 4.2. *Under Assumption 3.2, 3.3, 3.4 and 3.5, with $\eta_{t+1} = \eta_0(1+t)^{-\alpha}$, $\delta_{t+1} = \delta_0(1+t)^{-\beta}$ and $\alpha \in (0,1)$, $\beta \in (0, \frac{1}{2})$. Suppose $0 < 2\alpha - 4\beta < 1$ and*

$$\tau_k \geq \frac{1}{\log 1/\max\{\rho,\lambda\}}\left(\log(1+k) + \max\{\log\frac{\delta_0}{d}, 0\}\right).$$

Then, it holds that $\forall\, t \geq \max\{t_1, t_2\}$

$$\mathbf{I}_1(t) \leq c_1(1-\lambda)(1+t)^{\alpha}, \tag{17}$$

$$\mathbf{I}_2(t) \leq \frac{c_2 d^{5/2}}{(1-\lambda)^2}\mathcal{A}(t)^{\frac{1}{2}}(1+t)^{1-(\alpha-2\beta)}, \tag{18}$$

$$\mathbf{I}_3(t) \leq c_3\mathcal{A}(t)^{\frac{1}{2}}(1+t)^{1-\beta}, \tag{19}$$

$$\mathbf{I}_4(t) \leq \frac{c_4 d^2}{1-\lambda}(1+t)^{1-(\alpha-2\beta)}, \tag{20}$$

where t_1, t_2 are defined in (25), (26), and c_1, c_2, c_3, c_4 are constants defined as follows:

$$c_1 := 2G/\eta_0, \quad c_2 := \frac{\eta_0}{\delta_0^2}\frac{6\cdot(L_1G^2 + L_2G^2 + \sqrt{L}G^{3/2})}{\sqrt{1-2\alpha+4\beta}},$$

$$c_3 := \frac{2}{\sqrt{1-2\beta}}\max\{L\delta_0, G\sqrt{1-\beta}\},$$

$$c_4 := \frac{\eta_0}{\delta_0^2} \cdot \frac{LG^2}{2\beta - \alpha + 1}.$$

See Appendix C for the proof.

We comment that the bound for $\mathbf{I}_4(t)$ cannot be improved. As a concrete example, consider the constant function $\ell(\boldsymbol{\theta}; z) = \mathrm{c} \neq 0$ for all $z \in \mathsf{Z}$, it can be shown that $\|g_k^{(m)}\|^2 = \mathrm{c}^2$ and consequently $\mathbf{I}_4(t) = \Omega(\eta_k/\delta_k^2) = \Omega(t^{1-(\alpha-2\beta)})$, which matches (20).

Finally, plugging Lemma 4.2 into Lemma 4.1 gives:

$$\mathcal{A}(t) \leq \frac{c_1(1-\lambda)}{(1+t)^{1-\alpha}} + \frac{c_2 d^{5/2}}{(1-\lambda)^2} \frac{\mathcal{A}(t)^{\frac{1}{2}}}{(1+t)^{\alpha-2\beta}} + c_3 \frac{\mathcal{A}(t)^{\frac{1}{2}}}{(1+t)^{\beta}} + c_4 \frac{d^2}{1-\lambda} \frac{1}{(1+t)^{\alpha-2\beta}}.$$

Since $\mathcal{A}(t) \geq 0$, the above is a quadratic inequality that implies the following bound:

Lemma 4.3. *Under Assumption 3.1–3.5, with the step sizes* $\eta_{t+1} = \eta_0(1+t)^{-\alpha}$, $\delta_{t+1} = \delta_0(1+t)^{-\beta}, \tau_k \geq \frac{1}{\log 1/\max\{\rho,\lambda\}}\left(\log(1+k) + \max\{\log \frac{\delta_0}{d}, 0\}\right)$, $\eta_0 = d^{-2/3}, \delta_0 = d^{1/3}$, $\alpha \in (0,1)$, $\beta \in (0, \frac{1}{2})$. *If* $2\alpha - 4\beta < 1$, *then there exists a constant* t_0 *such that the iterates* $\{\boldsymbol{\theta}_k\}_{k\geq 0}$ *satisfies the following inequality for all* $T \geq t_0$,

$$\frac{1}{1+T}\sum_{k=0}^{T} \mathbb{E}\,\|\nabla \mathcal{L}(\boldsymbol{\theta}_k)\|^2 \leq 12 \max\left\{c_5(1-\lambda), c_6, \frac{c_7}{1-\lambda}\right\} d^{2/3} T^{-\min\{2\beta, 1-\alpha, \alpha-2\beta\}}.$$

Optimizing the step size exponents α, β in the above concludes the proof of Theorem 3.6.

4.1. Discussions

We conclude by discussing two alternative zero-th order gradient estimators to (4), and argue that they do not improve over the sample complexity in the proposed $\texttt{DFO}(\lambda)$ algorithm. We study:

$$\boldsymbol{g}_{\mathsf{2pt-I}} := \frac{d}{\delta}\left[\ell(\boldsymbol{\theta} + \delta \boldsymbol{u}; Z) - \ell(\boldsymbol{\theta}; Z)\right]\boldsymbol{u},$$
$$\boldsymbol{g}_{\mathsf{2pt-II}} := \frac{d}{\delta}\left[\ell(\boldsymbol{\theta} + \delta \boldsymbol{u}; Z_1) - \ell(\boldsymbol{\theta}; Z_2)\right]\boldsymbol{u},$$

where $\boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. For ease of illustration, we assume that the samples Z, Z_1, Z_2 are drawn directly from the stationary distributions $Z \sim \Pi_{\boldsymbol{\theta}+\delta u}$, $Z_1 \sim \Pi_{\boldsymbol{\theta}+\delta \boldsymbol{u}}$, $Z_2 \sim \Pi_{\boldsymbol{\theta}}$.

We recall from §2 that the estimator $\boldsymbol{g}_{\mathsf{2pt-I}}$ is a finite difference approximation of the directional derivative of objective function along the randomized direction $\boldsymbol{u}$[1], as proposed in (Nesterov & Spokoiny, 2017; Ghadimi & Lan, 2013). For non-convex stochastic optimization with decision independent sample distribution, i.e., $\Pi_{\boldsymbol{\theta}} \equiv \bar{\Pi}$ for all $\boldsymbol{\theta}$, the DFO algorithm based on $\boldsymbol{g}_{\mathsf{2pt-I}}$ is known to admit an optimal sample complexity of $\mathcal{O}(1/\epsilon^2)$ (Jamieson et al., 2012). Note that $\mathbb{E}_{\boldsymbol{u}\sim \mathrm{Unif}(\mathbb{S}^{d-1}), Z\sim\bar{\Pi}}[\ell(\boldsymbol{\theta}; Z)\boldsymbol{u}] = \mathbf{0}$.

However, in the case of decision-dependent sample distribution as in (1), $\boldsymbol{g}_{\mathsf{2pt-I}}$ would become a *biased* estimator since the sample Z is drawn from $\Pi_{\boldsymbol{\theta}+\delta \boldsymbol{u}}$ which depends on $\boldsymbol{u}$. The DFO algorithm based on $\boldsymbol{g}_{\mathsf{2pt-I}}$ may not converge to a stationary solution of (1).

A remedy to handle the above issues is to consider the estimator $\boldsymbol{g}_{\mathsf{2pt-II}}$ which utilizes *two samples* Z_1, Z_2, each independently drawn at a different decision variable, to form the gradient estimate. In fact, it can be shown that $\mathbb{E}[\boldsymbol{g}_{\mathsf{2pt-II}}] = \nabla \mathcal{L}_\delta(\boldsymbol{\theta})$ yields an unbiased gradient estimator. However, due to the decoupled random samples Z_1, Z_2, we have

$$\begin{aligned}
&\mathbb{E}\,\|\boldsymbol{g}_{\mathsf{2pt-II}}\|^2 \\
&= \frac{d^2}{\delta^2}\mathbb{E}\left(\ell(\boldsymbol{\theta} + \delta \boldsymbol{u}; Z_1) - \ell(\boldsymbol{\theta}; Z_1) + \ell(\boldsymbol{\theta}; Z_1) - \ell(\boldsymbol{\theta}; Z_2)\right)^2 \\
&\overset{(a)}{\geq} \frac{d^2}{\delta^2}\mathbb{E}\Big[\frac{3}{4}\left(\ell(\boldsymbol{\theta}; Z_1) - \ell(\boldsymbol{\theta}; Z_2)\right)^2 \\
&\qquad - 3\left(\ell(\boldsymbol{\theta} + \delta \boldsymbol{u}; Z_1) - \ell(\boldsymbol{\theta}; Z_1)\right)^2\Big] \\
&= \frac{3d^2}{\delta^2}\left(\frac{\mathsf{Var}[\ell(\boldsymbol{\theta}; Z)]}{2} - \mathbb{E}\left[\left(\ell(\boldsymbol{\theta} + \delta \boldsymbol{u}; Z_1) - \ell(\boldsymbol{\theta}; Z_1)\right)^2\right]\right) \\
&\overset{(b)}{\geq} \frac{3}{2}\frac{\sigma^2 d^2}{\delta^2} - 3\mu^2 d^2 = \Omega(d^2/\delta^2).
\end{aligned}$$

where in (a) we use the fact that $(x+y)^2 \geq \frac{3}{4}x^2 - 3y^2$, in (b) we assume $\mathsf{Var}[\ell(\boldsymbol{\theta}; Z)] := \mathbb{E}\left(\ell(\boldsymbol{\theta}; Z) - \mathcal{L}(\boldsymbol{\theta})\right)^2 \geq \sigma^2 > 0$ and $\ell(\boldsymbol{\theta}; z)$ is μ-Lipschitz in $\boldsymbol{\theta}$. As such, this two-point gradient estimator does not significantly reduce the variance when compared with (4). We remark that a two-sample estimator also incurs additional sampling overhead with Markovian data.

5. Numerical Experiments

We examine the efficacy of the $\texttt{DFO}(\lambda)$ algorithm on a few examples by comparing $\texttt{DFO}(\lambda)$ to several benchmarks. Unless otherwise specified, we use the step size choices in (8) for $\texttt{DFO}(\lambda)$. All experiments are conducted on a server with an Intel Xeon 6318 CPU using Python 3.7. The expected performance are estimated using at least 10 trials.

1-Dimensional Case: Quartic Loss. The first example

[1] In (Nesterov & Spokoiny, 2017; Ghadimi & Lan, 2013), $\boldsymbol{u}$ is drawn from the Gaussian distribution.

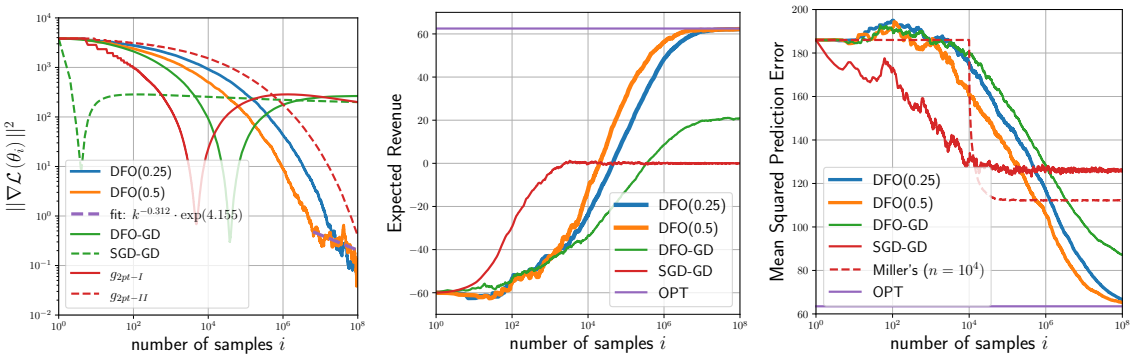

Figure 1. (*left*) One Dimension Quartic Minimization problem with samples generated by AR distribution model where regressive parameter $\gamma = 0.5$. (*middle*) Markovian Pricing Problem with $d = 5$ dimension. (*right*) Linear Regression problem based on AR distribution model ($\gamma = 0.5$).

considers a scalar polynomial loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ defined by $\ell(\boldsymbol{\theta}; z) = \frac{1}{12} z\boldsymbol{\theta}(3\boldsymbol{\theta}^2 - 8\boldsymbol{\theta} - 48)$. To simulate the controlled Markov chain scenario, the samples are generated dynamically according to an auto-regressive (AR) process $Z_{t+1} = (1-\gamma)Z_t + \gamma\bar{Z}_{t+1}$ with $\bar{Z}_{t+1} \sim \mathcal{N}(\boldsymbol{\theta}, \frac{(2-\gamma)}{\gamma}\sigma^2)$ with parameter $\gamma \in (0, 1)$. Note that the stationary distribution of the AR process is $\Pi_{\boldsymbol{\theta}} = \mathcal{N}(\boldsymbol{\theta}, \sigma^2)$. As such, the performative risk function in this case is $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{Z\sim\Pi_{\boldsymbol{\theta}}}[\ell(\boldsymbol{\theta}; Z)] = \frac{\boldsymbol{\theta}^2}{12}(\boldsymbol{\theta}^2 - 8\boldsymbol{\theta} - 48)$, which is quartic in $\boldsymbol{\theta}$. Note that $\mathcal{L}(\boldsymbol{\theta})$ is not convex in $\boldsymbol{\theta}$ and the set of stationary solutions is $\{\boldsymbol{\theta} : \nabla\mathcal{L}(\boldsymbol{\theta}) = 0\} = \{-2, 0, 4\}$, among which the optimal solution is $\boldsymbol{\theta}_{PO} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = 4$.

In our experiments below, all the algorithms are initialized by $\boldsymbol{\theta}_0 = 6$. In Figure 1 (left), we compare the norms of the gradient for performative risk with `DFO-GD` (no burn-in phase), the `DFO`(λ) algorithm, stochastic gradient descent with greedy deployment scheme (SGD-GD), and two 2-point estimators, $\boldsymbol{g}_{\mathsf{2pt-I}}, \boldsymbol{g}_{\mathsf{2pt-II}}$ (with a burn-in phase) as described in Section 4.1 against the number of samples observed by the algorithms. We first observe from Figure 1 (left) that `DFO-GD` and `SGD-GD` methods do not converge to a stationary point to $\mathcal{L}(\boldsymbol{\theta})$ even after more samples are observed. On the other hand, `DFO`(λ) converges to a stationary point of $\mathcal{L}(\boldsymbol{\theta})$ at the rate of $\|\nabla\mathcal{L}(\boldsymbol{\theta})\|^2 = \mathcal{O}(1/S^{0.312})$, matching Theorem 3.6 that predicts a rate of $\mathcal{O}(1/S^{\frac{1}{3}})$ (up to a logarithmic factor), where S is the total number of samples observed.

Besides, with $\lambda = 0.5$, `DFO`(λ) converges at a faster rate at the beginning (i.e., transient phase), but the fluctuation increases at the steady phase, as compared to a smaller λ (e.g., $\lambda = 0.25$). The estimator $\boldsymbol{g}_{\mathsf{2pt-I}}$ is indeed biased, and converges to a non-stationary point to $\mathcal{L}(\boldsymbol{\theta})$. On the other hand, $\boldsymbol{g}_{\mathsf{2pt-II}}$ converges to a stationary point, but uses twice as many samples as our one-point estimator for each update.

Higher Dimension Case: Markovian Pricing. The second example examines a multi-dimensional ($d = 5$) pricing problem similar to (Izzo et al., 2021, Sec. 5.2). The decision variable $\boldsymbol{\theta} \in \mathbb{R}^5$ denotes the prices of $d = 5$ goods and κ is a drifting parameter for the prices.

Our goal is to maximize the *Expected Revenue*, which is the opposite of performative risk $\mathbb{E}_{Z\sim\Pi_{\boldsymbol{\theta}}}[\ell(\boldsymbol{\theta}; Z)]$ with $\ell(\boldsymbol{\theta}; z) = -\langle \boldsymbol{\theta} \,|\, z\rangle$, where $\Pi_{\boldsymbol{\theta}} \equiv \mathcal{N}(\boldsymbol{\mu}_0 - \kappa\boldsymbol{\theta}, \sigma^2\boldsymbol{I})$ is the unique stationary distribution of the Markov process

$$Z_{t+1} = (1-\gamma)Z_t + \gamma\bar{Z}_{t+1} \; \bar{Z}_{t+1} \sim \mathcal{N}(\boldsymbol{\mu}_0 - \kappa\boldsymbol{\theta}, \tfrac{2-\gamma}{\gamma}\sigma^2\boldsymbol{I}).$$

Note that in this case, the performative optimal solution is $\boldsymbol{\theta}_{PO} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \boldsymbol{\mu}_0/(2\kappa)$. We set $\gamma = 0.1, \sigma = 1$, drifting parameter $\kappa = 0.5$, initial mean of non-shifted distribution $\boldsymbol{\mu}_0 = [5, -5, -5, 5, -5]^\top$. All the algorithms are initialized by $\boldsymbol{\theta}_0 = [12, -12, 12, -12, 12]^\top$.

The performance of different algorithms in terms of expected revenue ($-\mathbb{E}_{\Pi_{\boldsymbol{\theta}}}[\ell(\boldsymbol{\theta}; Z)]$) are contrasted in Figure 1 (middle), where `OPT` denotes the optimal objective value. Observe that `DFO`(λ) algorithms converge to the highest expected reward, while the benchmarks `SGD-GD` and `DFO-GD` fail to find a solution with comparable performance. This is expected as the latter algorithms are at best guaranteed to converge to a performative stable solution.

Markovian Performative Regression. The last example considers linear regression problem in (Nagaraj et al., 2020) which is a prototype problem for studying stochastic optimization with Markovian data. Such problems are rarely studied under both performativity and Markovianity.

Unlike the previous examples, this problem involves a pair of correlated r.v.s that follows a decision-dependent joint distribution. We adopt a setting similar to the regression example in (Izzo et al., 2021), where $(X, Y) \sim \Pi_{\boldsymbol{\theta}}$ with $X \sim \mathcal{N}(0, \sigma_1^2\boldsymbol{I}), Y|X \sim \mathcal{N}\left(\beta(\boldsymbol{\theta}, X), \sigma_2^2\right)$, $\beta(\boldsymbol{\theta}, x) = \langle x + \kappa\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_0\rangle$. The loss function is the usual quadratic loss $\ell(\boldsymbol{\theta}; x, y) = (\langle x \,|\, \boldsymbol{\theta}\rangle - y)^2$. In this case, we define the *Mean Squared Prediction Error* as the following performative risk:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\Pi_{\boldsymbol{\theta}}}[\ell(\boldsymbol{\theta}; X, Y)] = (\sigma_1^2 + \kappa^2 \|\boldsymbol{\theta}_0\|^2)\|\boldsymbol{\theta}\|^2 - 2\sigma_1^2\langle \boldsymbol{\theta} \,|\, \boldsymbol{\theta}_0\rangle + \sigma_1^2\|\boldsymbol{\theta}_0\|^2 + \sigma_2^2,$$

In this experiment, we consider Markovian samples $(\tilde{X}_t, \tilde{Y}_t)_{t=1}^T$ drawn from the following AR process:

$$(\tilde{X}_t, \tilde{Y}_t) = (1-\gamma)(\tilde{X}_{t-1}, \tilde{Y}_{t-1}) + \gamma(X_t, Y_t),$$
$$X_t \sim \mathcal{N}(0, \tfrac{2-\gamma}{\gamma}\sigma_1^2 I),\ Y_t|X_t \sim \mathcal{N}(\beta(\boldsymbol{\theta}_{t-1}, X_t), \tfrac{2-\gamma}{\gamma}\sigma_2^2),$$

for any $t \geq 1$. We set $d = 5$, $\boldsymbol{\theta}_0 = [5, -5, 5, -5, 5]^\top$, $\kappa = 1/\|\boldsymbol{\theta}_0\|, \sigma_1^2 = \sigma_2^2 = 1$, mixing parameter $\gamma = 0.25$. Figure 1 (right) shows the result of the simulation. Similar to the previous examples, we observe that DFO-GD and SGD fail to find a stationary solution to $\mathcal{L}(\boldsymbol{\theta})$.

Moreover, as a benchmark, we included the two-phase algorithm from (Miller et al., 2021), which is slightly modified as we change the minimization in the second phase to an SGD update. From the figure, we observe that it does not find the desired optimal objective value with $n = 10^4$ (Markovian) samples gathered in the first phase. In contrast, DFO (λ) converges to a near-optimal solution after a reasonable number of samples are observed.

6. Conclusions

We have studied a derivative-free optimization approach for finding a stationary point of the performative risk function. In particular, we consider a non-i.i.d. data setting with samples generated from a controlled Markov chain and propose a two-timescale step sizes approach in constructing the gradient estimator. The proposed DFO (λ) algorithm is shown to converge to a stationary point of the performative risk function at the rate of $\mathcal{O}(1/T^{1/3})$.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, A., Dekel, O., and Xiao, L. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Annual Conference Computational Learning Theory*, 2010.

Brown, G., Hod, S., and Kalemaj, I. Performative prediction in a stateful world. In *International Conference on Artificial Intelligence and Statistics*, pp. 6045–6061. PMLR, 2022.

Doan, T. T. Finite-time analysis of markov gradient descent. *IEEE Transactions on Automatic Control*, 2022.

Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 55–70, 2018.

Dong, R., Zhang, H., and Ratliff, L. Approximate regions of attraction in learning with decision-dependent distributions. In *International Conference on Artificial Intelligence and Statistics*, pp. 11172–11184. PMLR, 2023.

Drusvyatskiy, D. and Xiao, L. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 2022.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, pp. 385–394, USA, 2005. Society for Industrial and Applied Mathematics. ISBN 0898715857.

Ghadimi, S. and Lan, G. Stochastic first- and zeroth-order methods for nonconvex stochastic programming, 2013. URL https://arxiv.org/abs/1309.5549.

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pp. 111–122, 2016.

Izzo, Z., Ying, L., and Zou, J. How to learn when data reacts to your model: Performative gradient descent. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4641–4650. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/izzo21a.html.

Izzo, Z., Zou, J., and Ying, L. How to learn when data gradually reacts to your model. In *International Conference on Artificial Intelligence and Statistics*, pp. 3998–4035. PMLR, 2022.

Jamieson, K. G., Nowak, R., and Recht, B. Query complexity of derivative-free optimization. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/e6d8545daa42d5ced125a4bf747b3688-Paper.pdf.

Karimi, B., Miasojedow, B., Moulines, E., and Wai, H.-T. Non-asymptotic analysis of biased stochastic approximation scheme. In *Conference on Learning Theory*, pp. 1944–1974. PMLR, 2019.

Kim, M. P. and Perdomo, J. C. Making decisions under outcome performativity. In *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2023.

Li, Q. and Wai, H.-T. State dependent performative prediction with stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3164–3186. PMLR, 2022.

Mendler-Dünner, C., Perdomo, J., Zrnic, T., and Hardt, M. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33: 4929–4939, 2020.

Miller, J., Perdomo, J. C., and Zrnic, T. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, 2021.

Mofakhami, M., Mitliagkas, I., and Gidel, G. Performative prediction with neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 11079–11093. PMLR, 2023.

Nagaraj, D., Wu, X., Bresler, G., Jain, P., and Netrapalli, P. Least squares regression with markovian data: Fundamental limits and algorithms. *Advances in neural information processing systems*, 33:16666–16676, 2020.

Narang, A., Faulkner, E., Drusvyatskiy, D., Fazel, M., and Ratliff, L. J. Multiplayer performative prediction: Learning in decision-dependent games. *arXiv preprint arXiv:2201.03398*, 2022.

Nemirovskiĭ, A. S. *Problem complexity and method efficiency in optimization*. Wiley series in discrete mathematics. Wiley, Chichester, 1983.

Nesterov, Y. and Spokoiny, V. G. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

Ray, M., Ratliff, L. J., Drusvyatskiy, D., and Fazel, M. Decision-dependent risk minimization in geometrically decaying dynamic environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8081–8088, 2022.

Robert, C. P., Casella, G., and Casella, G. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

Roy, A., Balasubramanian, K., and Ghadimi, S. Projection-free constrained stochastic nonconvex optimization with state-dependent markov data. In *Advances in neural information processing systems*, 2022.

Sun, T. and Li, D. Decentralized markov chain gradient descent. *arXiv preprint arXiv:1909.10238*, 2019.

Sun, T., Sun, Y., and Yin, W. On markov chain gradient descent. *Advances in neural information processing systems*, 31, 2018.

Wood, K. and Dall'Anese, E. Stochastic saddle point problems with decision-dependent distributions. *SIAM Journal on Optimization*, 33(3):1943–1967, 2023.

Wu, Y. F., Zhang, W., Xu, P., and Gu, Q. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628, 2020.

Zhang, D. J., Dai, H., Dong, L., Qi, F., Zhang, N., Liu, X., Liu, Z., and Yang, J. How do price promotions affect customer behavior on retailing platforms? evidence from a large randomized experiment on alibaba. *Production and Operations Management*, 2018.

Zhu, Z., Fang, E. X., and Yang, Z. Online performative gradient descent for learning nash equilibria in decision-dependent games. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Supplementary Materials: Two-timescale Derivative Free Optimization for Performative Prediction with Markovian Data

A. Comparison to Related Works

This section provides a detailed comparison to related works on performative prediction under the stateful agent setting. This setting is relevant as the influences of the updated $\boldsymbol{\theta}$ on the agent may not be manifested immediately due to the unforgetful nature of the agent. The recent works can be grouped into two categories in terms of the sought solution to (1): (i) finding the *performative stable* solution satisfying $\boldsymbol{\theta}_{PS} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathbb{E}_{Z\sim\Pi_{\boldsymbol{\theta}_{PS}}}[\ell(\boldsymbol{\theta}; Z)]$, (ii) finding or approximating the *performative optimal* solution that tackles (1) directly.

For seeking the *performative stable* solution, (Brown et al., 2022) is the first to study population-based algorithms where the stateful agent updates the state-dependent distribution iteratively towards $\Pi_{\boldsymbol{\theta}}$. The authors proved that under a special case when k groups that form the mixture distribution for $\Pi_{\boldsymbol{\theta}}$ respond slowly, then classical retraining algorithms converge to the performative stable solution. The follow-up works (Li & Wai, 2022; Roy et al., 2022) focus on more sophisticated stateful agents and the reliance on past experiences of agents via controlled Markov Chain. In (Li & Wai, 2022), the authors developed gradient-type state-dependent stochastic approximation algorithm to achieve performative stable solution. In (Roy et al., 2022), the authors proposed a stochastic conditional gradient-type algorithm with state-dependent Markovian data to tackle constrained nonconvex performative prediction problem.

The search for (approximate) *performative optimal* solution is challenging due to the non-convexity of (1). Izzo et al. (2022) assumes that the transient distribution is parameterized by a low-dimensional vector and the distribution converges to $\Pi_{\boldsymbol{\theta}}$ geometrically. Under these settings, the authors proposed to learn the distribution as a linear model to form an unbiased estimate of $\nabla\mathcal{L}(\boldsymbol{\theta})$. The resultant algorithm follows a two-phases update approach: it first estimate the gradient correction term [cf. second term in (3)], followed by stochastic gradient update steps. Such approach has two main drawbacks: (i) estimating the gradient correction term requires strong prior assumptions on the distribution map (see e.g. Assumption 2 and 3 in (Izzo et al., 2022)), which limits its applicability, (ii) the estimation phase gathers a substantial amount of potentially sensitive information from data reaction patterns, which may incur privacy concern. Furthermore, it is noted that such procedure has a convergence rate of $\mathcal{O}(T^{-1/5})$ to stationary solution of $\mathcal{L}(\boldsymbol{\theta})$, which is outperformed by the approximation scheme proposed in the current paper.

As mentioned in the main paper, adopting the DFO setting avoids the need to estimate the gradient correction term, which may necessitate additional assumptions on $\Pi_{\boldsymbol{\theta}}$ as seen in (Izzo et al., 2022). To this end, one of the first works to address performative optimal points with DFO method is (Ray et al., 2022) in the stateful agent setting. Notably, the analysis in (Ray et al., 2022) relies on (i) a mixture dominance assumption on $\Pi_{\boldsymbol{\theta}}$, and (ii) a geometric decay environment assumption on the stateful agent. In addition to relaxing the mixture dominance assumption, we remark that Assumption 3.4 is relaxed from the geometric decay environment condition in (Ray et al., 2022). For example, our setting also covers general MDP models and the controlled AR(1) model, see (Li & Wai, 2022, Appendix A.1).

B. Proof of Lemma 4.1

Proof. Throughout this section, we let $\check{\boldsymbol{\theta}}_k := \boldsymbol{\theta}_k + \delta_k u_k$, $g_k(\boldsymbol{\theta}; u, z) := g_{\delta_k}(\boldsymbol{\theta}; u, z)$ and $\mathcal{L}_k(\boldsymbol{\theta}) := \mathcal{L}_{\delta_k}(\boldsymbol{\theta})$ for simplicity. We begin our analysis from Assumption 3.1 and the observation that $\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k = -\eta_k \sum_{m=1}^{\tau_k} \lambda^{\tau_k - m} g_k^{(m)}$. Recall that $g_k^{(m)} = \frac{d}{\delta_k}\ell\left(\check{\boldsymbol{\theta}}_k^{(m)}; z_k^{(m)}\right) u_k$ and $\check{\boldsymbol{\theta}}_k^{(m)} = \boldsymbol{\theta}_k^{(m)} + \delta_k u_k$, we have

$$\mathcal{L}(\boldsymbol{\theta}_{k+1}) - \mathcal{L}(\boldsymbol{\theta}_k) + \eta_k \left\langle \nabla\mathcal{L}(\boldsymbol{\theta}_k) \,|\, \sum_{m=1}^{\tau_k} \lambda^{\tau_k - m} g_k^{(m)} \right\rangle \leq \frac{L}{2}\eta_k^2 \left\| \sum_{m=1}^{\tau_k} \lambda^{\tau_k - m} g_k^{(m)} \right\|^2,$$

Rearranging terms and adding $\frac{\eta_k}{1-\lambda}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|^2$ on the both sides lead to

$$\frac{\eta_k}{1-\lambda}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|^2 \leq \mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}_{k+1}) - \frac{\eta_k}{1-\lambda}\left\langle \nabla\mathcal{L}(\boldsymbol{\theta}_k) \,|\, (1-\lambda)\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m} g_k^{(m)} - \nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\rangle + \frac{L}{2}\eta_k^2\left\|\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m} g_k^{(m)}\right\|^2$$

Let $\mathcal{F}^k = \sigma(\boldsymbol{\theta}_0, Z_s^{(m)}, u_s, 0\leq s\leq k, 0\leq m\leq \tau_k)$ be the filtration of random variables. Taking expectation conditioned on $\mathcal{F}^{k-1}$ gives

$$\begin{aligned}\frac{\eta_k}{1-\lambda}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|^2 \leq& \mathbb{E}_{\mathcal{F}^{k-1}}\left[\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}_{k+1})\right] \\ &- \frac{\eta_k}{1-\lambda}\left\langle \nabla\mathcal{L}(\boldsymbol{\theta}_k) \,|\, (1-\lambda)\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\mathbb{E}_{\mathcal{F}^{k-1}}\left[g_k^{(m)}\right] - \nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\rangle \\ &+ \frac{L}{2}\eta_k^2\mathbb{E}_{\mathcal{F}^{k-1}}\left\|\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m} g_k^{(m)}\right\|^2,\end{aligned}$$

By adding and subtracting, we obtain

$$\begin{aligned}&\frac{\eta_k}{1-\lambda}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|^2 \leq \mathbb{E}_{\mathcal{F}^{k-1}}\left[\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}_{k+1})\right] \\ &- \frac{\eta_k}{1-\lambda}\left\langle \nabla\mathcal{L}(\boldsymbol{\theta}_k) \,|\, (1-\lambda)\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\left(\mathbb{E}_{\mathcal{F}^{k-1}}\left[g_k^{(m)}\right] - \mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k},\mathcal{F}^{k-1}}\left[g_k(\boldsymbol{\theta}_k; u_k, Z)\right]\right)\right\rangle \\ &- \frac{\eta_k}{1-\lambda}\left\langle \nabla\mathcal{L}(\boldsymbol{\theta}_k) \,|\, (1-\lambda)\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k},\mathcal{F}^{k-1}}\left[g_k(\boldsymbol{\theta}_k; u_k, Z)\right] - \nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\rangle \\ &+ \frac{L}{2}\eta_k^2\mathbb{E}_{\mathcal{F}^{k-1}}\left\|\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m} g_k^{(m)}\right\|^2\end{aligned}$$

By Lemma F.2, the conditional expectation evaluates to $\mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}\left[g_k(\boldsymbol{\theta}_k; u_k, Z)\right] = \nabla\mathcal{L}_k(\boldsymbol{\theta}_k)$. Dividing $\frac{\eta_k}{1-\lambda}$ derive that

$$\begin{aligned}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|^2 \leq& \frac{1-\lambda}{\eta_k}\mathbb{E}_{\mathcal{F}^{k-1}}\left(\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}_{k+1})\right) \\ &- \left\langle \nabla\mathcal{L}(\boldsymbol{\theta}_k) \,|\, (1-\lambda)\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\left(\mathbb{E}_{\mathcal{F}^{k-1}}\left[g_k^{(m)}\right] - \mathbb{E}_{\mathcal{F}^{k-1}}\mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}\left[g_k(\boldsymbol{\theta}_k; u_k, Z)|u_k\right]\right)\right\rangle \\ &- \left\langle \nabla\mathcal{L}(\boldsymbol{\theta}_k) \,|\, (1-\lambda)\left(\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\nabla\mathcal{L}_k(\boldsymbol{\theta}_k)\right) - \nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\rangle \\ &+ \frac{L(1-\lambda)}{2}\eta_k\mathbb{E}_{\mathcal{F}^{k-1}}\left\|\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m} g_k^{(m)}\right\|^2\end{aligned}$$

Summing over k from 0 to t, indeed we obtain

$$\begin{aligned}&\sum_{k=0}^{t}\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|^2 \\ &\leq \sum_{k=0}^{t}\frac{1-\lambda}{\eta_k}\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}_{k+1})\right]\end{aligned}$$

$$
\begin{aligned}
&- \sum_{k=0}^{t} \mathbb{E}\left\langle \nabla \mathcal{L}(\boldsymbol{\theta}_k) \,|\, (1-\lambda) \sum_{m=1}^{\tau_k} \lambda^{\tau_k - m} \left(\mathbb{E}_{\mathcal{F}^{k-1}}\left[g_k^{(m)} \right] - \mathbb{E}_{\mathcal{F}^{k-1}} \mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}} \left[g_k(\boldsymbol{\theta}_k; u_k, Z) | u_k \right] \right) \right\rangle \\
&- \sum_{k=0}^{t} \mathbb{E}\left\langle \nabla \mathcal{L}(\boldsymbol{\theta}_k) \,|\, (1-\lambda) \left(\sum_{m=1}^{\tau_k} \lambda^{\tau_k - m} \nabla \mathcal{L}_k(\boldsymbol{\theta}_k) \right) - \nabla \mathcal{L}(\boldsymbol{\theta}_k) \right\rangle \\
&+ \frac{L(1-\lambda)}{2} \sum_{k=0}^{t} \eta_k \mathbb{E} \left\| \sum_{m=1}^{\tau_k} \lambda^{\tau_k - m} g_k^{(m)} \right\|^2 := \mathbf{I}_1(t) + \mathbf{I}_2(t) + \mathbf{I}_3(t) + \mathbf{I}_4(t)
\end{aligned}
$$

□

C. Proof of Lemma 4.2

Lemma C.1. *Under Assumption 3.2 and step size $\eta_t = \eta_0(1+t)^{-\alpha}$, it holds that*

$$
\mathbf{I}_1(t) \le c_1(1-\lambda)(1+t)^{\alpha} \tag{21}
$$

where constant $c_1 = \frac{2G}{\eta_0}$.

Proof. We observe the following chain

$$
\begin{aligned}
\mathbf{I}_1(t) &= \sum_{k=0}^{t} \frac{1-\lambda}{\eta_k} \left(\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_k)\right] - \mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{k+1})\right] \right) \\
&= (1-\lambda) \sum_{k=0}^{t} \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}_k)]/\eta_k - \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}_{k+1})]/\eta_{k+1} + \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}_{k+1})]/\eta_{k+1} - \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}_{k+1})]/\eta_k \\
&\overset{(a)}{=} (1-\lambda) \left[\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}_0)/\eta_0] - \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}_{t+1})/\eta_{t+1}] + \sum_{k=0}^{t} \left(\frac{1}{\eta_{k+1}} - \frac{1}{\eta_k}\right) \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}_{k+1})] \right] \\
&\le (1-\lambda) \max_k |\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}_k)]| \left(\frac{1}{\eta_0} + \frac{1}{\eta_{t+1}} + \frac{1}{\eta_{t+1}} - \frac{1}{\eta_0} \right)
\end{aligned}
$$

where equality (a) is obtained using the fact that step size $\eta_k > 0$ is a decreasing sequence. Applying assumption 3.2 to the last inequality leads to

$$
\mathbf{I}_1(t) \le (1-\lambda) G \frac{2}{\eta_{t+1}} \le c_1(1-\lambda)(1+t)^{\alpha}
$$

where the constant $c_1 = \frac{2G}{\eta_0}$.

□

Lemma C.2. *Under Assumption 3.1, 3.2, 3.3, 3.4, 3.5, and constraint $0 < 2\alpha - 4\beta < 1$, and for all $k \ge 0$, $\tau_k \ge \frac{1}{\log 1/\max\{\rho,\lambda\}} \log(1+k)$, then there exists universal constants $t_1, t_2 > 0$ such that*

$$
\mathbf{I}_2(t) \le c_2 \frac{d^2}{(1-\lambda)^2} \mathcal{A}(t)^{1/2} (1+t)^{1-(\alpha-2\beta)} \quad \forall t \ge \max\{t_1, t_2\} \tag{22}
$$

where $\mathcal{A}(t) := \frac{1}{1+t} \sum_{k=0}^{t} \mathbb{E} \left\| \nabla \mathcal{L}(\boldsymbol{\theta}_k) \right\|^2$ and $c_2 := \frac{\eta_0}{\delta_0^2} \frac{6\cdot(L_1 G^2 + L_2 G^2 + \sqrt{L} G^{3/2})}{\sqrt{1-2\alpha+4\beta}}$ is a constant.

Proof. Fix $k > 0$, and recall $\check{\boldsymbol{\theta}}_k := \boldsymbol{\theta}_k + \delta_k u_k, \check{\boldsymbol{\theta}}_k^{(\ell)} := \boldsymbol{\theta}_k^{(\ell)} + \delta_k u_k$, then consider the following pair of Markov chains:

$$
Z_k = Z_k^{(0)} \xrightarrow{\check{\boldsymbol{\theta}}_k^{(1)}} Z_k^{(1)} \xrightarrow{\check{\boldsymbol{\theta}}_k^{(2)}} Z_k^{(2)} \xrightarrow{\check{\boldsymbol{\theta}}_k^{(3)}} Z_k^{(3)} \cdots \xrightarrow{\check{\boldsymbol{\theta}}_k^{(\tau_k)}} Z_k^{(\tau_k)} = Z_{k+1} \tag{23}
$$

$$
Z_k = \tilde{Z}_k^{(0)} \xrightarrow{\check{\boldsymbol{\theta}}_k} \tilde{Z}_k^{(1)} \xrightarrow{\check{\boldsymbol{\theta}}_k} \tilde{Z}_k^{(2)} \xrightarrow{\check{\boldsymbol{\theta}}_k} \tilde{Z}_k^{(3)} \cdots \xrightarrow{\check{\boldsymbol{\theta}}_k} \tilde{Z}_k^{(\tau_k)} \tag{24}
$$

where the arrow associated with $\boldsymbol{\theta}$ represents the transition kernel $\mathbb{T}_{\boldsymbol{\theta}}(\cdot,\cdot)$.

Note that Chain 23 is the trajectory of $\mathsf{DFO}(\lambda)$ algorithm at iteration k, while Chain 24 describes the trajectory of the same length generated by a reference Markov chain with fixed transition kernel $\mathbb{T}_{\check{\boldsymbol{\theta}}_k}(\cdot,\cdot)$. Since $Z_k = Z_k^{(0)} = \tilde{Z}_k^{(0)}$, we shall use them interchangeably.

Define $\Delta_{k,m} := \mathbb{E}_{\mathcal{F}^{k-1}}\left[g_k^{(m)} - \mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}\left[g_k(\boldsymbol{\theta}_k;u_k,Z)\right]\right]$, then $\mathbf{I}_2(t)$ can be reformed as

$$\begin{aligned}\mathbf{I}_2(t) &= -(1-\lambda)\mathbb{E}\sum_{k=0}^{t}\left\langle \nabla\mathcal{L}(\boldsymbol{\theta}_k) \mid \sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\Delta_{k,m}\right\rangle \\ &\le (1-\lambda)\mathbb{E}\sum_{k=0}^{t}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|\cdot\left\|\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\Delta_{k,m}\right\|\end{aligned}$$

Next, observe that each $\Delta_{k,m}$ can be decomposed into 3 bias terms as follows

$$\begin{aligned}\Delta_{k,m} &= \mathbb{E}_{\mathcal{F}^{k-1}}\left[\frac{d}{\delta_k}\left(\mathbb{E}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};Z_k^{(m)})|\check{\boldsymbol{\theta}}_k^{(m)},Z_k^{(0)}] - \mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}[\ell(\check{\boldsymbol{\theta}}_k;Z)|\check{\boldsymbol{\theta}}_k]\right)u_k\right] \\ &= \mathbb{E}_{\mathcal{F}^{k-1}}\left[\frac{d}{\delta_k}\left(\mathbb{E}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};Z_k^{(m)})|\check{\boldsymbol{\theta}}_k^{(m)},Z_k^{(0)}] - \mathbb{E}_{\tilde{Z}_k^{(m)}}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};\tilde{Z}_k^{(m)})|\check{\boldsymbol{\theta}}_k^{(m)},\tilde{Z}_k^{(0)}]\right)u_k\right] \\ &+ \mathbb{E}_{\mathcal{F}^{k-1}}\left[\frac{d}{\delta_k}\left(\mathbb{E}_{\tilde{Z}_k^{(m)}}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};\tilde{Z}_k^{(m)})|\check{\boldsymbol{\theta}}_k^{(m)},\tilde{Z}_k^{(0)}] - \mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};Z)|\check{\boldsymbol{\theta}}_k^{(m)}]\right)u_k\right] \\ &+ \mathbb{E}_{\mathcal{F}^{k-1}}\frac{d}{\delta_k}\underbrace{\mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}\left[\ell(\check{\boldsymbol{\theta}}_k^{(m)};Z) - \ell(\check{\boldsymbol{\theta}}_k;Z)|\check{\boldsymbol{\theta}}_k^{(m)},\check{\boldsymbol{\theta}}_k\right]}_{\le c_8\left\|\check{\boldsymbol{\theta}}_k^{(m)}-\check{\boldsymbol{\theta}}_k\right\|+\frac{L}{2}\left\|\check{\boldsymbol{\theta}}_k^{(m)}-\check{\boldsymbol{\theta}}_k\right\|^2}u_k\end{aligned}$$

where we use Lemma F.4 in the last inequality and $c_8 := 2\left(\sqrt{LG}+GL_1\right)$.

Here we bound these three parts separately. For the first term, it holds that

$$\begin{aligned}&\left|\mathbb{E}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};Z_k^{(m)})|\check{\boldsymbol{\theta}}_k^{(m)},Z_k^{(0)}] - \mathbb{E}_{\tilde{Z}_k^{(m)}}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};\tilde{Z}_k^{(m)})|\check{\boldsymbol{\theta}}_k^{(m)},\tilde{Z}_k^{(0)}]\right| \\ =&\left|\int_{\mathsf{Z}}\ell(\check{\boldsymbol{\theta}}_k^{(m)};z)\mathbb{P}(Z_k^{(m)}=z|Z_k^{(0)}) - \ell(\check{\boldsymbol{\theta}}_k^{(m)};z)\mathbb{P}(\tilde{Z}_k^{(m)}=z|\tilde{Z}_k^{(0)})dz\right| \\ \le& G\int_{\mathsf{Z}}\left|\mathbb{P}(Z_k^{(m)}=z|Z_k^{(0)}) - \mathbb{P}(\tilde{Z}_k^{(m)}=z|\tilde{Z}_k^{(0)})\right|dz \\ =&2G\boldsymbol{\delta}_{\mathrm{TV}}\left(\mathbb{P}(z_k^{(m)}\in\cdot|Z_k^{(0)}),\mathbb{P}(\tilde{Z}_k^{(m)}\in\cdot|Z_k^{(0)})\right) \\ \le&2GL_2\sum_{\ell=1}^{m-1}\left\|\check{\boldsymbol{\theta}}_k^{(\ell)}-\check{\boldsymbol{\theta}}_k\right\| = 2GL_2\sum_{\ell=1}^{m-1}\left\|\boldsymbol{\theta}_k^{(\ell)}-\boldsymbol{\theta}_k\right\|\end{aligned}$$

where the first inequality is due to Assumption 3.2, the second inequality is due to Lemma F.5.

For the second term, we have

$$\begin{aligned}&\left|\mathbb{E}_{\tilde{Z}_k^{(m)}}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};Z_k^{(m)})] - \mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}[\ell(\check{\boldsymbol{\theta}}_k;Z)]\right| \\ =&\left|\int_{\mathsf{Z}}\ell(\check{\boldsymbol{\theta}}_k^{(m)};z)\mathbb{P}(\tilde{Z}_k^{(m)}=z|\tilde{Z}_k^{(0)}) - \ell(\check{\boldsymbol{\theta}}_k^{(m)};z)\Pi_{\check{\boldsymbol{\theta}}_k}(z)dz\right| \\ \overset{(a)}{\le}& G\int_{\mathsf{Z}}|\mathbb{P}(\tilde{Z}_k^{(m)}=z|\tilde{Z}_k^{(0)}) - \Pi_{\check{\boldsymbol{\theta}}_k}(z))|dz \\ =&2G\boldsymbol{\delta}_{\mathrm{TV}}\left(\mathbb{P}(\tilde{Z}_k^{(m)}\in\cdot|\tilde{Z}_k^{(0)}),\Pi_{\check{\boldsymbol{\theta}}_k}\right) \\ \overset{(b)}{\le}&2GM\rho^m\end{aligned}$$

where we use Assumption 3.2 in inequality (a) and Assumptions 3.4 in inequality (b). Combining three upper bounds, we obtain that

$$\begin{aligned}
\|\Delta_{k,m}\| &\leq \mathbb{E}_{\mathcal{F}^{k-1}} \frac{d}{\delta_k}\left(2GL_2\sum_{\ell=1}^{m-1}\left[\left\|\boldsymbol{\theta}_k^{(\ell)}-\boldsymbol{\theta}_k\right\|\right]+2GM\rho^m+c_8\left\|\check{\boldsymbol{\theta}}_k^{(m)}-\check{\boldsymbol{\theta}}_k\right\|+\frac{L}{2}\left\|\check{\boldsymbol{\theta}}_k^{(m)}-\check{\boldsymbol{\theta}}_k\right\|^2\right)\\
&\leq \frac{d}{\delta_k}\left(2L_2G\sum_{\ell=1}^{m-1}\sum_{j=1}^{\ell-1}\eta_k\lambda^{\tau_k-j}\frac{dG}{\delta_k}+2GM\rho^m+c_8\sum_{j=1}^{m-1}\eta_k\lambda^{\tau_k-j}\frac{dG}{\delta_k}\right)\\
&+\frac{d}{\delta_k}\frac{L}{2}\left(\sum_{j=1}^{m-1}\eta_k\lambda^{\tau_k-j}\frac{dG}{\delta_k}\right)^2\\
&< \frac{d}{(1-\lambda)^2}\left(2L_2G^2d+c_8Gd\right)\lambda^{\tau_k-m+1}\frac{\eta_k}{\delta_k^2}+\frac{LG^2d^3}{2(1-\lambda)^2}\lambda^{2(\tau_k-m+1)}\frac{\eta_k^2}{\delta_k^3}+2GMd\frac{\rho^m}{\delta_k}
\end{aligned}$$

Then it holds that

$$\begin{aligned}
\left\|\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\Delta_{k,m}\right\| &\leq \sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\|\Delta_{k,m}\|\\
&\leq \frac{d}{(1-\lambda)^2}\left(2L_2G^2d+c_8dG\right)\frac{\eta_k}{\delta_k^2}\sum_{m=1}^{\tau_k}\lambda^{2(\tau_k-m)}\lambda\\
&+\frac{LG^2}{2(1-\lambda)^2}d^3\frac{\eta_k^2}{\delta_k^3}\sum_{m=1}^{\tau_k}\lambda^{3(\tau_k-m)}\lambda^2\\
&+2GMd\delta_k^{-1}\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\rho^m\\
&\leq (2L_2G^2d+c_8Gd)\frac{d\lambda}{(1-\lambda)^3}\frac{\eta_k}{\delta_k^2}\\
&+\frac{LG^2}{2}\frac{d^3\lambda^2}{1-\lambda}\frac{\eta_k^2}{\delta_k^3}+2GMd\delta_k^{-1}\tau_k\max\{\rho,\lambda\}^{\tau_k}
\end{aligned}$$

Finally, provided $\tau_k \geq \log_{\max\{\rho,\lambda\}}(1+k)^{-1}$ and $0<2\alpha-4\beta<1$, we can bound $\mathbf{I}_2(t)$ as follows:

$$\begin{aligned}
\mathbf{I}_2(t) &\leq (1-\lambda)\mathbb{E}\sum_{k=0}^{t}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|\cdot\left\|\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\Delta_{k,m}\right\|\\
&\leq (1-\lambda)\mathbb{E}\sum_{k=0}^{t}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|\left[(2L_2G^2d+c_8Gd)\frac{d\lambda}{(1-\lambda)^3}\frac{\eta_k}{\delta_k^2}+\frac{LG^2}{2}\frac{d^3\lambda^2}{1-\lambda}\frac{\eta_k^2}{\delta_k^3}+2GMd\frac{\tau_k}{\delta_k(1+k)}\right]\\
&\leq \frac{d\lambda}{(1-\lambda)^2}(2L_2G^2d+c_8Gd)\left(\sum_{k=0}^{t}\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|^2\right)^{1/2}\left(\sum_{k=0}^{t}\frac{\eta_k^2}{\delta_k^4}\right)^{1/2}\\
&+d^3\lambda^2\frac{LG^2}{2}\left(\sum_{k=0}^{t}\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|^2\right)^{1/2}\left(\sum_{k=0}^{t}\frac{\eta_k^4}{\delta_k^6}\right)^{1/2}\\
&+\frac{d}{1-\lambda}GM\left(\sum_{k=0}^{t}\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|^2\right)^{1/2}\left(\sum_{k=0}^{t}\frac{\tau_k^2}{\delta_k^2(1+k)^2}\right)^{1/2}\\
&\overset{(b)}{\leq} \frac{d^2\lambda}{(1-\lambda)^2}6(L_2G^2+\sqrt{L}G^{3/2}+L_1G^2)\left(\sum_{k=0}^{t}\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|^2\right)^{1/2}\left(\sum_{k=0}^{t}\frac{\eta_k^2}{\delta_k^4}\right)^{1/2}\\
&\leq c_2\frac{d^2}{(1-\lambda)^2}\left(\frac{1}{1+t}\sum_{k=0}^{t}\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\|^2\right)^{1/2}\cdot(1+t)^{1-(\alpha-2\beta)}\quad\forall t\geq\max\{t_1,t_2\}
\end{aligned}$$

where $c_2 := \frac{\eta_0}{\delta_0^2}\frac{6\cdot(L_1G^2+L_2G^2+\sqrt{L}G^{3/2})}{\sqrt{1-2\alpha+4\beta}}$. The inequality (b) holds since $\tau_k = \Theta(\log k)$, $4\alpha - 6\beta > 2\alpha - 4\beta$ and $2-2\beta > 2\alpha - 4\beta$, so there exist constants

$$t_1 := \inf_t \left\{ t \geq 0 \,|\, \frac{d^6\lambda^4L^2G^4}{4}\sum_{k=0}^{t}\frac{\eta_k^4}{\delta_k^6} \leq \frac{d^2\lambda^2(2L_2G^2d + c_8Gd)^2}{(1-\lambda)^4}\sum_{k=0}^{t}\frac{\eta_k^2}{\delta_k^4}\right\} \tag{25}$$

$$t_2 := \inf_t \left\{ t \geq 0 \,|\, d^2G^2M^2\sum_{k=0}^{t}\frac{\tau_k^2}{\delta_k^2(1+k)^2} \leq \frac{d^2\lambda^2(2L_2G^2d + c_8Gd)^2}{(1-\lambda)^4}\sum_{k=0}^{t}\frac{\eta_k^2}{\delta_k^4}\right\} \tag{26}$$

In brief, we have

$$\mathbf{I}_2(t) \leq c_2\frac{d^2}{(1-\lambda)^2}\left(\frac{1}{1+t}\sum_{k=0}^{t}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\|^2\right)^{1/2}\cdot(1+t)^{1-(\alpha-2\beta)} \quad \forall t \geq \max\{t_1, t_2\}$$

□

Lemma C.3. *Under Assumption 3.1, 3.2 and $0 < \beta < 1/2$, with $\tau_k \geq \frac{1}{\log 1/\max\{\rho,\lambda\}}\left(\log(1+k) + \log\frac{d}{\delta_0}\right)$, it holds that*

$$\mathbf{I}_3(t) \leq c_3\mathcal{A}(t)^{\frac{1}{2}}(1+t)^{1-\beta} \tag{27}$$

where $\mathcal{A}(t) := \frac{1}{1+t}\sum_{k=0}^{t}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\|^2$ and constant $c_3 = \frac{1}{\sqrt{1-2\beta}}\max\{2^{1-\beta}L\delta_0, 2^\beta G\sqrt{1-\beta}\}$.

Proof. Recall that $g_k(\boldsymbol{\theta}; u, z) := g_{\delta_k}(\boldsymbol{\theta}; u, z)$ and $\mathcal{L}_k(\boldsymbol{\theta}) := \mathcal{L}_{\delta_k}(\boldsymbol{\theta})$.

$$\begin{aligned}
\mathbf{I}_3(t) &= -\sum_{k=0}^{t}\mathbb{E}\left\langle \nabla\mathcal{L}(\boldsymbol{\theta}_k) \,|\, (1-\lambda)\left(\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\nabla\mathcal{L}_k(\boldsymbol{\theta}_k)\right) - \nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\rangle \\
&= -\sum_{k=0}^{t}\mathbb{E}\left\langle \nabla\mathcal{L}(\boldsymbol{\theta}_k) \,|\, \left((1-\lambda)\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\right)\nabla\mathcal{L}_k(\boldsymbol{\theta}_k) - \nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\rangle \\
&= -\sum_{k=0}^{t}\mathbb{E}\left\langle \nabla\mathcal{L}(\boldsymbol{\theta}_k) \,|\, \nabla\mathcal{L}_k(\boldsymbol{\theta}_k) - \nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\rangle - \lambda^{\tau_k}\mathbb{E}\left\langle \nabla\mathcal{L}(\boldsymbol{\theta}_k) \,|\, \mathbb{E}_{Z\sim\Pi_{\hat{\boldsymbol{\theta}}_k}}\left[g_k(\boldsymbol{\theta}_k; u_k, Z)\right]\right\rangle
\end{aligned}$$

where we apply Lemma F.1 at the last equality.

By triangle inequality, Cauchy-Schwarz inequality and Assumption 3.2, we obtain

$$\mathbf{I}_3(t) \leq \sum_{k=0}^{t}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\|\cdot\left\|\nabla\mathcal{L}_k(\boldsymbol{\theta}_k) - \nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\| + \sum_{k=0}^{t}\lambda^{\tau_k}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\|\frac{dG}{\delta_k}$$

Provided $\tau_k \geq \frac{\log(1+k)+\log\frac{d}{\delta_0}}{\log 1/\max\{\rho,\lambda\}} \geq \frac{\log\delta_0/d(1+k)^{-1}}{\log\max\{\rho,\lambda\}} = \log_{\max\{\rho,\lambda\}}\frac{\delta_0}{d}(1+k)^{-1} \geq \log_\lambda\frac{\delta_0}{d}(1+k)^{-1}$, with Lemma F.2 as a consequence of Assumption 3.1, we have

$$\begin{aligned}
\mathbf{I}_3(t) &\leq \sum_{k=0}^{t}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\|\cdot L\delta_k + \sum_{k=0}^{t}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\|\frac{\delta_0}{d}\frac{dG}{\delta_0}(1+k)^{\beta-1} \\
&= \sum_{k=0}^{t}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\|\cdot L\delta_k + G\sum_{k=0}^{t}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\|(1+k)^{\beta-1} \\
&\leq L\left(\sum_{k=0}^{t}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\|^2\right)^{1/2}\left(\sum_{k=0}^{t}\delta_k^2\right)^{1/2} + G\left(\sum_{k=0}^{t}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\|^2\right)^{1/2}\left(\sum_{k=0}^{t}(1+k)^{2(\beta-1)}\right)^{1/2}
\end{aligned}$$

Since $\beta < 1/2$, it holds that

$$\sum_{k=0}^{t} \delta_k^2 = \sum_{k=0}^{t} \frac{\delta_0^2}{(1+k)^{2\beta}} \leq \frac{\delta_0^2}{1-2\beta}\left[1-2\beta+(1+t)^{1-2\beta}-1\right] \leq \frac{\delta_0^2}{1-2\beta}(1+t)^{1-2\beta}$$
$$\sum_{k=0}^{t}(1+k)^{2(\beta-1)} \leq 1+\int_0^t (x+1)^{2(\beta-1)}dx < 1+\frac{1}{1-2\beta}$$

Then we can conclude

$$\mathbf{I}_3(t) \leq c_3\left(\frac{1}{1+t}\sum_{k=0}^{t}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\|^2\right)^{1/2}\cdot(1+t)^{1-\beta}$$

where $c_3 := \frac{2}{\sqrt{1-2\beta}}\max\{L\delta_0, G\sqrt{1-\beta}\}$. □

Lemma C.4. *Under assumption 3.2 and constraint $0<\alpha<1$, it holds that*

$$\mathbf{I}_4(t) \leq c_4\frac{d^2}{1-\lambda}(1+t)^{1-(\alpha-2\beta)} \tag{28}$$

where constant $c_4 = \frac{\eta_0 LG^2}{\delta_0^2(2\beta-\alpha+1)}$.

Proof.

$$\begin{aligned}
\mathbf{I}_4(t) &= \frac{(1-\lambda)L}{2}\sum_{k=0}^{t}\eta_k\mathbb{E}\left\|\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}g_k^{(m)}\right\|^2 \\
&\leq \frac{(1-\lambda)L}{2}\sum_{k=0}^{t}\eta_k\mathbb{E}\left(\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\left\|g_k^{(m)}\right\|\right)^2 \\
&\leq \frac{(1-\lambda)L}{2}\sum_{k=0}^{t}\eta_k\left(\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\right)^2\frac{(dG)^2}{\delta_k^2} \\
&\leq \frac{(1-\lambda)Ld^2G^2}{2}\sum_{k=0}^{t}\left(\frac{1-\lambda^{\tau_k}}{1-\lambda}\right)^2\frac{\eta_k}{\delta_k^2} \\
&< \frac{d^2LG^2}{2(1-\lambda)}\sum_{k=0}^{t}\frac{\eta_k}{\delta_k^2}
\end{aligned}$$

Recall that $\eta_k = \frac{\eta_0}{(k+1)^\alpha}$, $\delta_k = \frac{\delta_0}{(1+k)^\beta}$ and $\alpha<1, \beta\geq 0$, it is clear that $\alpha-2\beta<1$, so it holds that

$$\begin{aligned}
\sum_{k=0}^{t}\frac{\eta_k}{\delta_k^2} &= \frac{\eta_0}{\delta_0^2}\sum_{k=0}^{t}(1+k)^{2\beta-\alpha} \leq \frac{\eta_0}{\delta_0^2}\left(1+\int_0^t(1+x)^{2\beta-\alpha}dx\right) \\
&\leq \frac{\eta_0}{\delta_0^2(2\beta-\alpha+1)}\left[(1+t)^{2\beta-\alpha+1}-\alpha+2\beta\right] \leq \frac{2\eta_0}{\delta_0^2(2\beta-\alpha+1)}(1+t)^{2\beta-\alpha+1}
\end{aligned}$$

In conclusion, we obtain that

$$\mathbf{I}_4(t) \leq d^2\frac{LG^2}{1-\lambda}\frac{\eta_0}{\delta_0^2(2\beta-\alpha+1)}\cdot(1+t)^{2\beta-\alpha+1} = c_4\frac{d^2}{1-\lambda}(1+t)^{1-(\alpha-2\beta)}$$

where $c_4 := \frac{\eta_0}{\delta_0^2}\cdot\frac{LG^2}{2\beta-\alpha+1}$. □

D. Proof of Lemma 4.3

Proof. Combining Lemmas 4.1 and 4.2, subject to the constraints $0 < \alpha < 1, 0 < \beta \leq 1/2, 0 < 2\alpha - 4\beta \leq 1$, it holds that for any $t \geq \max\{t_1, t_2\}$,

$$\begin{aligned}
&\sum_{k=0}^{t} \mathbb{E}\left\|\nabla \mathcal{L}(\boldsymbol{\theta}_k)\right\|^2 \\
&\leq \mathbf{I}_1(t) + \mathbf{I}_2(t) + \mathbf{I}_3(t) + \mathbf{I}_4(t) \\
&\leq c_1(1-\lambda)(1+t)^{\alpha} + c_2 \frac{d^{5/2}}{(1-\lambda)^2}(1+t)^{1-(\alpha-2\beta)}\mathcal{A}(t)^{1/2} \\
&\quad + c_3(1+t)^{1-\beta}\mathcal{A}(t)^{1/2} + c_4 \frac{d^2}{1-\lambda}(1+t)^{1-(\alpha-2\beta)}
\end{aligned}$$

Recall $\mathcal{A}(t) := \frac{1}{1+t}\sum_{k=0}^{t} \mathbb{E}\left\|\nabla \mathcal{L}(\boldsymbol{\theta}_k)\right\|^2$, above inequality can be rewritten as

$$\begin{aligned}
\mathcal{A}(t) &\leq \frac{1}{1+t}\Bigg[c_2 \frac{d^{5/2}}{(1-\lambda)^2}(1+t)^{1-(\alpha-2\beta)}\mathcal{A}(t)^{1/2} \\
&\quad + c_3(1+t)^{1-\beta}\mathcal{A}(t)^{1/2} + c_1(1-\lambda)(1+t)^{\alpha} + c_4 \frac{d^2}{1-\lambda}(1+t)^{1-(\alpha-2\beta)}\Bigg] \\
&= \left(c_2 \frac{d^{5/2}}{(1-\lambda)^2}(1+t)^{-(\alpha-2\beta)} + c_3(1+t)^{-\beta}\right)\mathcal{A}(t)^{1/2} + c_1(1-\lambda)(1+t)^{-(1-\alpha)} \\
&\quad + c_4 \frac{d^2}{1-\lambda}(1+t)^{-(\alpha-2\beta)}
\end{aligned}$$

which is a quadratic inequality in $\mathcal{A}(t)^{1/2}$.

Let $x = \mathcal{A}(t)^{1/2}, a = c_2 \frac{d^{5/2}}{(1-\lambda)^2}(1+t)^{-(\alpha-2\beta)} + c_3 t^{-\beta}, b = c_1(1-\lambda)(1+t)^{-(1-\alpha)} + c_4 \frac{d^2}{1-\lambda}(1+t)^{-(\alpha-2\beta)}$, we have $x^2 - ax - b \leq 0$. Since $a, b > 0$, the quadratic has two real roots, denoted as x_1, x_2 respectively, and $x_1 < 0 < x_2$. Moreover, we must have $x \leq x_2$, which implies $x \leq \frac{a+\sqrt{a^2+4b}}{2} \leq \frac{a+a+2\sqrt{b}}{2} = a+\sqrt{b}$. Therefore, $\mathcal{A}(t) = x^2 \leq (a+\sqrt{b})^2 \leq 2(a^2+b)$. Substituting a, b back leads to

$$\begin{aligned}
\mathcal{A}(t) &\leq 2\left(c_2 \frac{d^{5/2}}{(1-\lambda)^2}(1+t)^{-(\alpha-2\beta)} + c_3(1+t)^{-\beta}\right)^2 + 2c_1(1-\lambda)(1+t)^{-(1-\alpha)} \\
&\quad + 2c_4 \frac{d^2}{1-\lambda}(1+t)^{-(\alpha-2\beta)} \\
&\overset{(a)}{\leq} 4c_2^2 \frac{d^5}{(1-\lambda)^4}(1+t)^{-2(\alpha-2\beta)} + 4c_3^2(1+t)^{-2\beta} + 2c_1(1-\lambda)(1+t)^{-(1-\alpha)} \\
&\quad + 2c_4 \frac{d^2}{1-\lambda}(1+t)^{-(\alpha-2\beta)} \\
&\leq 4c_3^2(1+t)^{-2\beta} + 2c_1(1-\lambda)(1+t)^{-(1-\alpha)} + 4c_4 \frac{d^2}{1-\lambda}(1+t)^{-(\alpha-2\beta)},
\end{aligned}$$

where inequality (a) is due to the fact $(x+y)^2 \leq 2(x^2+y^2)$, the last inequality holds because there exists sufficiently large constant t_3 such that, $4c_2^2 \frac{d^5}{(1-\lambda)^4}(1+t)^{-2(\alpha-2\beta)} \leq 2c_4 \frac{d^2}{1-\lambda}(1+t)^{-(\alpha-2\beta)} \forall t \geq t_3$ given $\alpha > 2\beta$. Therefore, set $t_0 := \max\{t_1, t_2, t_3\}$, then for all $t \geq t_0$, we have

$$\begin{aligned}
\mathcal{A}(t) &\leq 4\max\{c_1(1-\lambda), c_3^2, c_4 \frac{d^2}{1-\lambda}\} \cdot \left((1+t)^{-2\beta} + (1+t)^{-(1-\alpha)} + (1+t)^{-(\alpha-2\beta)}\right) \\
&\leq 12\max\{c_1(1-\lambda), c_3^2, c_4 \frac{d^2}{1-\lambda}\}(1+t)^{-\min\{2\beta, 1-\alpha, \alpha-2\beta\}}
\end{aligned}$$

Recall that constant c_1 contains $1/\eta_0$, c_3 contains δ_0, c_4 contains η_0/δ_0^2, , thus we can set $\delta_0 = d^{1/3}, \eta_0 = d^{-2/3}$, which yields

$$\mathcal{A}(t) \leq 12\max\{c_5(1-\lambda), c_6, \frac{c_7}{1-\lambda}\}d^{2/3}(1+t)^{-\min\{2\beta,1-\alpha,\alpha-2\beta\}}$$

where constants

$$c_5 = 2G, \quad c_6 = \frac{4\max\{L^2, G^2(1-\beta)\}}{1-2\beta}, \quad c_7 = \frac{LG^2}{2\beta-\alpha+1}$$

do not contain η_0 and δ_0. Moreover, note that $\max_{\alpha,\beta}\min\{2\beta, 1-\alpha, \alpha-2\beta\} = \frac{1}{3}$, thus it holds

$$\frac{1}{1+T}\sum_{k=0}^{T}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{\theta})_k\right\|^2 \leq 12\max\{c_5(1-\lambda), c_6, \frac{c_7}{1-\lambda}\}d^{2/3}(1+T)^{-1/3}$$

where the rate $\mathcal{O}(1/T^{1/3})$ can be attained by choosing $\alpha = \frac{2}{3}$, $\beta = \frac{1}{6}$. This immediately leads to Theorem 3.6. □

E. Non-smooth Analysis

In this section, we aim to apply our algorithm to non-smooth performative risk optimization problem and analyze its convergence rate. Before presenting the theorem, we need the following Lipschitz loss assumption E.1.

Assumption E.1. (Lipschitz Loss) There exists constant $L_0 > 0$ such that

$$|\ell(\boldsymbol{\theta}_1; z) - \ell(\boldsymbol{\theta}_2; z)| \leq L_0\left\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\right\|, \ \forall\, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d, \ \forall\, z \in \mathsf{Z}$$

Under Assumption E.1 and some other regularity conditions, one can show that the performative risk is also Lipschitz continuous. Formally, this can be stated as follows.

Lemma E.2. *Under Assumption E.1, 3.2, 3.3, the performative risk $\mathcal{L}(\boldsymbol{\theta})$ is $(L_0 + 2L_1G)$-Lipschitz continuous.*

Under non-smooth settings, the convergence behavior can be characterized in both squared gradient norm and proximity gap. Now, we are ready to show the following theorem:

Theorem E.3. (`DFO`(λ) for Non-smooth Optimization) *Under Assumption E.1, 3.2, 3.3, 3.4, 3.5, with two time-scale step sizes $\eta_k = \eta_0(1+k)^{-\alpha}, \delta_k = d(1+k)^{-\beta}, \tau_k \geq \frac{\log(1+k)}{\log 1/\max\{\rho,\lambda\}}$, where α, β satisfies $0 < 3\beta < \alpha < 1$, there exists a constant t_4 such that, the iterates $\{\boldsymbol{\theta}_k\}_{k\geq 1}$ satisfies for all $T \geq t_4$*

$$\frac{1}{1+T}\sum_{k=0}^{T}\mathbb{E}\left\|\nabla\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)\right\|^2 = \mathcal{O}(T^{-\min\{1-\alpha,\alpha-3\beta\}})$$

and the following error estimate holds for all $T > 0$ and $\boldsymbol{\theta} \in \mathbb{R}^d$

$$\frac{1}{1+T}\sum_{k=0}^{T}\mathbb{E}|\mathcal{L}_{\delta_k}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta})| = \mathcal{O}(T^{-\beta})$$

Corollary E.4. *(ϵ-stationarity, μ-proximity) Suppose Assumptions of Theorem E.3 hold. Fix any $\epsilon, \mu > 0$, for $T = \max\{\mathcal{O}(1/\epsilon^4), \mathcal{O}(1/\mu^6)\}$, the following estimates hold simultaneously*

$$\frac{1}{1+T}\sum_{k=0}^{T}\mathbb{E}\left\|\nabla\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)\right\|^2 \leq \epsilon$$

$$\frac{1}{1+T}\sum_{k=0}^{T}\mathbb{E}\left|\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k) - \mathcal{L}(\boldsymbol{\theta}_k)\right| \leq \mu$$

Next, we present the proof of Theorem E.3.

Proof. This proof resembles the proof of Lemma 4.2, where we reinterpret $\sum_{k=0}^{t}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_k)\right\|^2$ as $\sum_{k=0}^{t}\mathbb{E}\left\|\nabla\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)\right\|^2$, and $\mathcal{L}(\boldsymbol{\theta}_k)$ as $\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)$, with additional bias terms that, as we shall prove, are not dominant.

Due to Lemma E.2, $\mathcal{L}(\boldsymbol{\theta})$ is (L_0+2L_1G)-Lipschitz. Then by Lemma F.1, $\mathcal{L}_\delta(\boldsymbol{\theta})$ is $\frac{d}{\delta}(L_0+2L_1G)$-smooth for all $\delta>0$. Similar to Lemma 4.1, we have

$$\begin{aligned}
&\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_{k+1})-\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)+\frac{\eta_k}{1-\lambda}\left\langle\nabla\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)\,|\,(1-\lambda)\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}g_k^{(m)}\right\rangle\\
&\le\frac{d(L_0+2L_1G)}{2\delta_k}\eta_k^2\left\|\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}g_k^{(m)}\right\|^2
\end{aligned}$$

By adding, subtracting and rearranging terms, after taking conditional expectation on $\mathcal{F}^{k-1}$, it holds that

$$\begin{aligned}
\frac{\eta_k}{1-\lambda}\left\|\nabla\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)\right\|^2\le{}&\mathbb{E}_{\mathcal{F}^{k-1}}\left[\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)-\mathcal{L}_{\delta_{k+1}}(\boldsymbol{\theta}_{k+1})+\mathcal{L}_{\delta_{k+1}}(\boldsymbol{\theta}_{k+1})-\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_{k+1})\right]\\
&+\frac{\eta_k}{1-\lambda}\mathbb{E}_{\mathcal{F}^{k-1}}\left\langle\nabla\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)\,|\,\nabla\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)-(1-\lambda)\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}g_k^{(m)}\right\rangle\\
&+\frac{d}{2\delta_k}(L_0+2L_1G)\eta_k^2\mathbb{E}_{\mathcal{F}^{k-1}}\left\|\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}g_k^{(m)}\right\|^2
\end{aligned}$$

By Lemma F.1, we have $\mathbb{E}_{Z\sim\Pi_{\hat{\boldsymbol{\theta}}_k},u_k}[g_{\delta_k}(\boldsymbol{\theta}_k;u_k,Z)]=\nabla\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)$, then by dividing and summing over k, it holds that

$$\begin{aligned}
&(1-\lambda)\sum_{k=0}^{t}\mathbb{E}\left\|\nabla\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)\right\|^2\\
&\le\sum_{k=0}^{t}\frac{1-\lambda}{\eta_k}\mathbb{E}\left[\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)-\mathcal{L}_{\delta_{k+1}}(\boldsymbol{\theta}_{k+1})+\mathcal{L}_{\delta_{k+1}}(\boldsymbol{\theta}_{k+1})-\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_{k+1})\right]\\
&\quad+(1-\lambda)\sum_{k=0}^{t}\mathbb{E}\left\langle\nabla\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)\,|\,\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\left(\mathbb{E}_{Z\sim\Pi_{\hat{\boldsymbol{\theta}}_k}}[g_{\delta_k}(\boldsymbol{\theta}_k;u_k,Z)]-g_k^{(m)}\right)\right\rangle\\
&\quad+\sum_{k=0}^{t}\lambda^{\tau_k}\mathbb{E}\left\|\nabla\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)\right\|^2\\
&\quad+\frac{d(L_0+2L_1G)(1-\lambda)}{2}\sum_{k=0}^{t}\frac{\eta_k}{\delta_k}\mathbb{E}\left\|\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}g_k^{(m)}\right\|^2\\
&:=\mathbf{I}_5(t)+\mathbf{I}_6(t)+\mathbf{I}_7(t)+\mathbf{I}_8(t)
\end{aligned}$$

After splitting RHS into $\mathbf{I}_5(t),\mathbf{I}_6(t),\mathbf{I}_7(t),\mathbf{I}_8(t)$, we can bound them separately.

Under Assumption 3.2 and the estimate $\delta_k-\delta_{k+1}=\Theta(k^{-\beta-1})$, it holds that

$$\begin{aligned}
\mathbf{I}_5(t)&=(1-\lambda)\sum_{k=0}^{t}\frac{1}{\eta_k}\mathbb{E}\left[\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)-\mathcal{L}_{\delta_{k+1}}(\boldsymbol{\theta}_{k+1})\right]+(1-\lambda)\sum_{k=0}^{t}\frac{1}{\eta_k}\mathbb{E}\left[\mathcal{L}_{\delta_{k+1}}(\boldsymbol{\theta}_{k+1})-\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_{k+1})\right]\\
&\overset{(a)}{\le}(1-\lambda)G\frac{2}{\eta_{t+1}}+(1-\lambda)\sum_{k=0}^{t}\mathbb{E}\frac{\mathcal{L}_{\delta_{k+1}}(\boldsymbol{\theta}_k)-\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)}{\eta_k}\\
&\overset{(b)}{\le}(1-\lambda)G\frac{2}{\eta_{t+1}}+(1-\lambda)(L_0+2L_1G)\sum_{k=0}^{t}\frac{\delta_k-\delta_{k+1}}{\eta_k}\\
&=\mathcal{O}\left((1+t)^{\alpha}+(1+t)^{\alpha-\beta}\right)=\mathcal{O}\left((1+t)^{\alpha}\right)
\end{aligned}$$

where we apply the summation by part in inequality (a) as in Lemma C.1, and use the fact $|\mathcal{L}_{\delta_1}(\boldsymbol{\theta}) - \mathcal{L}_{\delta_2}(\boldsymbol{\theta})| \leq \mathbb{E}_w|\mathcal{L}(\boldsymbol{\theta} + \delta_1 w) - \mathcal{L}(\boldsymbol{\theta} + \delta_2 w)| \leq (L_0 + 2L_1 G)|\delta_1 - \delta_2|$ in inequality (e), as a consequence of Lipschitz continuity.

As for $\mathbf{I}_6(t)$, if we let $\mathcal{B}(t) := \frac{1}{1+t}\sum_{k=0}^{t}\mathbb{E}\left\|\nabla\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)\right\|^2$, by definition of $g_k^{(m)}$, we can split the term as follows

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{F}^{k-1}}\frac{d}{\delta_k}\left(\mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}[\ell(\check{\boldsymbol{\theta}}_k;Z)|\check{\boldsymbol{\theta}}_k] - \mathbb{E}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};Z_k^{(m)})|\check{\boldsymbol{\theta}}_k^{(m)},Z_k^{(0)}]\right)\\
&= \mathbb{E}_{\mathcal{F}^{k-1}}\frac{d}{\delta_k}\mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}\left[\ell(\check{\boldsymbol{\theta}}_k;Z) - \ell(\check{\boldsymbol{\theta}}_k^{(m)};Z)|\check{\boldsymbol{\theta}}_k^{(m)},\check{\boldsymbol{\theta}}_k\right]\\
&\quad+\mathbb{E}_{\mathcal{F}^{k-1}}\frac{d}{\delta_k}\left(\mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};Z)|\check{\boldsymbol{\theta}}_k^{(m)}] - \mathbb{E}_{\tilde{Z}_k^{(m)}}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};\tilde{Z}_k^{(m)})|\check{\boldsymbol{\theta}}_k^{(m)},\tilde{Z}_k^{(0)}]\right)\\
&\quad+\mathbb{E}_{\mathcal{F}^{k-1}}\frac{d}{\delta_k}\left(\mathbb{E}_{\tilde{Z}_k^{(m)}}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};\tilde{Z}_k^{(m)})|\check{\boldsymbol{\theta}}_k^{(m)},\tilde{Z}_k^{(0)}] - \mathbb{E}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};Z_k^{(m)})|\check{\boldsymbol{\theta}}_k^{(m)},Z_k^{(0)}]\right)
\end{aligned}
$$

By applying Jensen's inequality and triangle inequality according to the above splitting, it holds that

$$
\begin{aligned}
&\left\|\mathbb{E}_{\mathcal{F}^{k-1}}\mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}[g_{\delta_k}(\boldsymbol{\theta}_k;u_k,Z)] - g_k^{(m)}\right\|\\
=&\left|\frac{d}{\delta_k}|\mathbb{E}_{\mathcal{F}^{k-1}}\mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}[\ell(\check{\boldsymbol{\theta}}_k;Z)|\check{\boldsymbol{\theta}}_k] - \mathbb{E}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};Z_k^{(m)})|\check{\boldsymbol{\theta}}_k^{(m)},Z_k^{(0)}]\right|\\
\leq&\mathbb{E}_{\mathcal{F}^{k-1}}\frac{d}{\delta_k}\left|\mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}[\ell(\check{\boldsymbol{\theta}}_k;Z)|\check{\boldsymbol{\theta}}_k] - \mathbb{E}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};Z_k^{(m)})|\check{\boldsymbol{\theta}}_k^{(m)},Z_k^{(0)}]\right|\\
\leq&\mathbb{E}_{\mathcal{F}^{k-1}}\frac{d}{\delta_k}\left|\mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}\left[\ell(\check{\boldsymbol{\theta}}_k;Z) - \ell(\check{\boldsymbol{\theta}}_k^{(m)};Z)|\check{\boldsymbol{\theta}}_k^{(m)},\check{\boldsymbol{\theta}}_k\right]\right|\\
+&\mathbb{E}_{\mathcal{F}^{k-1}}\frac{d}{\delta_k}\left|\mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};Z)|\check{\boldsymbol{\theta}}_k^{(m)}] - \mathbb{E}_{\tilde{Z}_k^{(m)}}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};\tilde{Z}_k^{(m)})|\check{\boldsymbol{\theta}}_k^{(m)},\tilde{Z}_k^{(0)}]\right|\\
+&\mathbb{E}_{\mathcal{F}^{k-1}}\frac{d}{\delta_k}\left|\mathbb{E}_{\tilde{Z}_k^{(m)}}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};\tilde{Z}_k^{(m)})|\check{\boldsymbol{\theta}}_k^{(m)},\tilde{Z}_k^{(0)}] - \mathbb{E}[\ell(\check{\boldsymbol{\theta}}_k^{(m)};Z_k^{(m)})|\check{\boldsymbol{\theta}}_k^{(m)},Z_k^{(0)}]\right|\\
\overset{(c)}{\leq}&\frac{d}{\delta_k}\mathbb{E}_{\mathcal{F}^{k-1}}L_0\left\|\check{\boldsymbol{\theta}}_k^{(m)} - \check{\boldsymbol{\theta}}_k\right\|\\
+&\frac{2dG}{\delta_k}\mathbb{E}_{\mathcal{F}^{k-1}}\boldsymbol{\delta}_{\mathrm{TV}}\left(\Pi_{\boldsymbol{\theta}_k},\mathbb{P}(\hat{Z}_k^{(m)}\in\cdot|\check{\boldsymbol{\theta}}_k^{(0)},\hat{Z}_k^{(0)})\right)\\
+&\frac{2dG}{\delta_k}\mathbb{E}_{\mathcal{F}^{k-1}}\boldsymbol{\delta}_{\mathrm{TV}}\left(\mathbb{P}(\hat{Z}_k^{(m)}\in\cdot|\check{\boldsymbol{\theta}}_k^{(0)},\hat{Z}_k^{(0)}),\mathbb{P}(Z_k^{(m)}\in\cdot|\check{\boldsymbol{\theta}}_k^{(0)},Z_k^{(0)})\right)\\
\overset{(d)}{\leq}&\frac{dL_0}{\delta_k}\mathbb{E}_{\mathcal{F}^{k-1}}\left\|\check{\boldsymbol{\theta}}_k^{(m)} - \check{\boldsymbol{\theta}}_k\right\| + \frac{2dG}{\delta_k}M\rho^m + \frac{2dL_2G}{\delta_k}\mathbb{E}_{\mathcal{F}^{k-1}}\sum_{\ell=1}^{m-1}\left\|\check{\boldsymbol{\theta}}_k^{(\ell)} - \check{\boldsymbol{\theta}}_k\right\|\\
\leq&\frac{dL_0}{\delta_k}dG\sum_{j=1}^{m-1}\lambda^{\tau_k-j}\frac{\eta_k}{\delta_k} + \frac{2dGM}{\delta_k}\rho^m + \frac{2dL_2G}{\delta_k}dG\sum_{\ell=1}^{m-1}\sum_{j=1}^{\ell-1}\lambda^{\tau_k-j}\frac{\eta_k}{\delta_k}\\
<&d^2L_0G\frac{\eta_k}{\delta_k^2}\frac{\lambda^{\tau_k-m+1}}{1-\lambda} + \frac{2dGM}{\delta_k}\rho^m + 2d^2L_2G^2\frac{\eta_k}{\delta_k^2}\frac{\lambda^{\tau_k-m+2}}{(1-\lambda)^2}
\end{aligned}
$$

where inequality (c) is due to Lipschitzness of decoupled risk, inequality (d) is due to Assumption 3.4 and Lemma F.5 (a consequence of Assumption 3.5). Given $\tau_k \geq \frac{\log(1+k)}{\log 1/\max\{\rho,\lambda\}}$, then the following deterministic bound holds for all $k > 0$,

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{F}^{k-1}}\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\left\|\mathbb{E}_{Z\sim\Pi_{\check{\boldsymbol{\theta}}_k}}[g_{\delta_k}(\boldsymbol{\theta}_k;u_k,Z)] - g_k^{(m)}\right\|\\
\leq&d^2L_0G\frac{\eta_k}{\delta_k^2}\frac{\lambda}{1-\lambda}\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m} + 2d^2L_2G^2\frac{\eta_k}{\delta_k^2}\frac{\lambda^2}{1-\lambda}\sum_{m=1}^{\tau_k}\lambda^{\tau_k-m}\\
+&2dGM/\delta_k\sum_{m=1}^{\tau_k}\rho^m\lambda^{\tau_k-m}
\end{aligned}
$$

$$
\begin{aligned}
&< d^2 \frac{\lambda}{(1-\lambda)^2} L_0 G \frac{\eta_k}{\delta_k^2} + 2d^2 \frac{\lambda^2}{(1-\lambda)^2} L_2 G^2 \frac{\eta_k}{\delta_k^2} + 2dGM/\delta_k \sum_{m=1}^{\tau_k} \max\{\rho, \lambda\}^{\tau_k} \\
&\leq d^2 \frac{\lambda}{(1-\lambda)^2} L_0 G \frac{\eta_k}{\delta_k^2} + 2d^2 \frac{\lambda^2}{(1-\lambda)^2} L_2 G^2 \frac{\eta_k}{\delta_k^2} + 2dGM \frac{\tau_k}{(1+k)\delta_k}
\end{aligned}
$$

So for sufficiently large t, it holds that

$$
\begin{aligned}
\mathbf{I}_6(t) &\leq (1-\lambda) \sum_{k=0}^{t} \mathbb{E} \left\| \nabla \mathcal{L}(\boldsymbol{\theta}_k) \right\| \mathbb{E}_{\mathcal{F}^{k-1}} \left\| \sum_{m=1}^{\tau_k} \lambda^{\tau_k - m} \mathbb{E}_{Z \sim \Pi_{\hat{\boldsymbol{\theta}}_k}} [g_{\delta_k}(\boldsymbol{\theta}_k; u_k, Z)] - g_k^{(m)} \right\| \\
&\leq (1-\lambda) \sum_{k=0}^{t} \mathbb{E} \left\| \nabla \mathcal{L}(\boldsymbol{\theta}_k) \right\| \sum_{m=1}^{\tau_k} \lambda^{\tau_k - m} \mathbb{E}_{\mathcal{F}^{k-1}} \left\| \mathbb{E}_{Z \sim \Pi_{\hat{\boldsymbol{\theta}}_k}} [g_{\delta_k}(\boldsymbol{\theta}_k; u_k, Z)] - g_k^{(m)} \right\| \\
&\leq \sum_{k=0}^{t} \mathbb{E} \left\| \nabla \mathcal{L}(\boldsymbol{\theta}_k) \right\| d^2 \frac{\lambda}{1-\lambda} ((L_0 + 2L_1 G)G + 2L_1 G + 2\lambda L_2 G^2) \frac{\eta_k}{\delta_k^2} \\
&+ \sum_{k=0}^{t} \mathbb{E} \left\| \nabla \mathcal{L}(\boldsymbol{\theta}_k) \right\| 2dGM \frac{\tau_k}{(1+k)\delta_k} \\
&\leq \sum_{k=0}^{t} \mathbb{E} \left\| \nabla \mathcal{L}(\boldsymbol{\theta}_k) \right\| d^2 \frac{2\lambda}{(1-\lambda)^2} (L_0 G + 2L_1 G + 2\lambda L_2 G^2) \frac{\eta_k}{\delta_k^2} \\
&= d^2 \frac{2\lambda}{(1-\lambda)^2} (L_0 G + 2L_1 G + 2\lambda L_2 G^2) \sum_{k=0}^{t} \mathbb{E} \left\| \nabla \mathcal{L}(\boldsymbol{\theta}_k) \right\| \frac{\eta_k}{\delta_k^2} \\
&\leq d^2 \frac{2\lambda}{(1-\lambda)^2} (L_0 G + 2L_1 G + 2\lambda L_2 G^2) \Big(\sum_{k=0}^{t} \mathbb{E} \left\| \nabla \mathcal{L}(\boldsymbol{\theta}_k) \right\|^2 \Big)^{1/2} \Big(\sum_{k=0}^{t} \frac{\eta_k^2}{\delta_k^4} \Big)^{1/2} \\
&\leq c_9 d^2 \mathcal{B}(t)^{1/2} (1+t)^{\frac{1}{2} + \frac{1}{2} - (\alpha - 2\beta)}
\end{aligned}
$$

Therefore, there exists a constant $c_9 > 0$ such that

$$
\mathbf{I}_6(t) \leq c_9 d^2 \mathcal{B}(t)^{1/2} (1+t)^{1-(\alpha-2\beta)}
$$

where there is an extra $\beta/2$ in exponent because the L in c_2 is now a variable $d(L_0 + 2L_1 G)/\delta_k$.

For $(L_0 + 2L_1 G)$-Lipschitz continuous $\mathcal{L}(\boldsymbol{\theta})$, for all $\delta > 0$ it holds that $\|\nabla \mathcal{L}_\delta(\boldsymbol{\theta})\| \leq (L_0 + 2L_1 G)$. Given $\tau_k \geq \frac{\log(1+k)}{\log 1/\max\{\rho,\lambda\}}$, it holds that $\lambda^{\tau_k} \mathbb{E} \left\| \nabla \mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k) \right\|^2 \leq \frac{dL^2}{\delta_0(1+k)}$, then $\mathbf{I}_7(t)$ can be bounded as follows

$$
\mathbf{I}_7(t) \leq \frac{dL^2}{\delta_0} \sum_{k=0}^{t} (1+k)^{-1} = \mathcal{O}(\log(1+t))
$$

$\mathbf{I}_8(t)$ is similar to $\mathbf{I}_4(t)$. For all $0 \leq k \leq t, 1 \leq m \leq \tau_k$, it holds that $\left\| g_k^{(m)} \right\| \leq \frac{dG}{\delta_k}$, which implies

$$
\begin{aligned}
\mathbf{I}_8(t) &\leq (1-\lambda) \frac{d(L_0 + 2L_1 G)}{2} \sum_{k=0}^{t} \frac{\eta_k}{\delta_k} \mathbb{E} \left\| \sum_{m=1}^{\tau_k} \lambda^{\tau_k - m} g_k^{(m)} \right\|^2 \\
&\leq (1-\lambda) \frac{d(L_0 + 2L_1 G)}{2} \sum_{k=0}^{t} \frac{\eta_k}{\delta_k} \mathbb{E} \Big(\sum_{m=1}^{\tau_k} \lambda^{\tau_k - m} \left\| g_k^{(m)} \right\| \Big)^2 \\
&\leq (1-\lambda) \frac{d(L_0 + 2L_1 G)}{2} \sum_{k=0}^{t} \frac{\eta_k}{\delta_k} \Big(\sum_{m=1}^{\tau_k} \lambda^{\tau_k - m} \frac{dG}{\delta_k} \Big)^2 \\
&= (1-\lambda) \frac{d^3 (L_0 + 2L_1 G) G^2}{2} \sum_{k=0}^{t} \frac{\eta_k}{\delta_k^3} \Big(\sum_{m=1}^{\tau_k} \lambda^{\tau_k - m} \Big)^2
\end{aligned}
$$

$$\leq \frac{d^3(L_0+2L_1G)G^2}{2(1-\lambda)}\sum_{k=0}^{t}\frac{\eta_k}{\delta_k^3} \leq c_{10}(1+t)^{1-(\alpha-3\beta)}$$

where $c_{10}>0$ is a constant hiding the factor $\frac{\eta_0}{\delta_0^3}$.

Applying quadratic technique in Lemma 4.3, and for all α,β satisfying $0<3\beta<\alpha<1$, it is clear that only $\mathbf{I}_5(t)$ and $\mathbf{I}_8(t)$ contribute to the asymptotic rate, so for all $t\geq t_4$ (for some constant $t_4>0$), we have

$$\frac{1}{1+T}\sum_{k=0}^{T}\mathbb{E}\left\|\nabla\mathcal{L}_{\delta_k}(\boldsymbol{\theta}_k)\right\|^2 = \mathcal{O}(T^{-\min\{1-\alpha,\alpha-3\beta\}})$$

The error estimate directly follows from Lemma E.2. □

Remark E.5. Note that Corrollary E.4 follows directly from Theorem E.3 by setting $\alpha=3/4$ and $\beta=1/6$.

F. Auxiliary Lemmas

Lemma F.1. (Smoothing) *For continuous $\mathcal{L}(\boldsymbol{\theta}):\mathbb{R}^d\to\mathbb{R}$, its smoothed approximation $\mathcal{L}_\delta(\boldsymbol{\theta}):=\mathbb{E}_{w\sim Unif(\mathbb{B}^d)}[\mathcal{L}(\boldsymbol{\theta}+\delta w)]$ is differentiable, and it holds that*

$$\mathbb{E}_{\substack{u\sim Unif(\mathbb{S}^{d-1}),\\ Z\sim\Pi_{\boldsymbol{\theta}+\delta u}}}[g_\delta(\boldsymbol{\theta};u,Z)] = \nabla\mathcal{L}_\delta(\boldsymbol{\theta})$$

Moreover, if $\mathcal{L}(\boldsymbol{\theta})$ is $\bar{L}$-Lipschitz continuous, then $\mathcal{L}_\delta(\boldsymbol{\theta})$ is $\frac{d}{\delta}\bar{L}$-smooth.

Proof. The first fact follows from (generalized) Stoke's theorem. Given continuous $\mathcal{L}(\boldsymbol{\theta})$, it holds that

$$\nabla\int_{\delta\mathbb{B}^d}\mathcal{L}(\boldsymbol{\theta}+v)dv = \int_{\delta\mathbb{S}^{d-1}}\mathcal{L}(\boldsymbol{\theta}+r)\frac{r}{\|r\|}dr \tag{29}$$

Observe that the RHS of Equation (29) is continuous in $\boldsymbol{\theta}$, which implies $\mathcal{L}_\delta(\boldsymbol{\theta})=\frac{1}{\mathsf{vol}(\delta\mathbb{B}^d)}\int_{\delta\mathbb{B}^d}\mathcal{L}(\boldsymbol{\theta}+v)dv$ is differentiable. Note that the volume to surface area ratio of $\delta\mathbb{B}^d$ is δ/d, so it follows from Equation (29) that

$$\begin{aligned}\nabla\mathcal{L}_\delta(\boldsymbol{\theta}) &= \frac{\mathsf{vol}(\delta\mathbb{S}^{d-1})}{\mathsf{vol}(\delta\mathbb{B}^d)}\int_{\delta\mathbb{S}^{d-1}}\mathcal{L}(\boldsymbol{\theta}+r)\frac{r}{\mathsf{vol}(\delta\mathbb{S}^{d-1})\,\|r\|}dr = \frac{d}{\delta}\mathbb{E}_{u\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[\mathcal{L}(\boldsymbol{\theta}+\delta u)u]\\ &= \mathbb{E}_{u\sim\mathrm{Unif}(\mathbb{S}^{d-1})}\mathbb{E}_{Z\sim\pi_{\boldsymbol{\theta}+\delta u}}[\frac{d}{\delta}\ell(\boldsymbol{\theta}+\delta u;Z)u] = \mathbb{E}_{\substack{u\sim\mathrm{Unif}(\mathbb{S}^{d-1}),\\ Z\sim\Pi_{\boldsymbol{\theta}+\delta u}}}[g_\delta(\boldsymbol{\theta};u,Z)]\end{aligned}$$

where we use the definition of $g_\delta(\boldsymbol{\theta};u,z)$ in the last equality.

If further assuming $\mathcal{L}(\boldsymbol{\theta})$ is $\bar{L}$-Lipschitz continuous, then we obtain

$$\begin{aligned}\|\nabla\mathcal{L}_\delta(\boldsymbol{\theta}_1)-\nabla\mathcal{L}_\delta(\boldsymbol{\theta}_2)\| &= \frac{d}{\delta}\cdot\left\|\frac{1}{\mathsf{vol}(\mathbb{S}^{d-1})}\int_{\mathbb{S}^{d-1}}\left[\mathcal{L}(\boldsymbol{\theta}_1+\delta u)-\mathcal{L}(\boldsymbol{\theta}_2+\delta u)\right]udu\right\|\\ &\leq \frac{d}{\delta}\cdot\bar{L}\left\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\right\|.\end{aligned}$$

□

Lemma F.2. ($\mathcal{O}(\delta)$-Biased Gradient Estimation)

Under Assumptions 3.1, fix a proximity parameter $\delta>0$, it holds that

$$\left\|\mathbb{E}_{\substack{u\sim Unif(\mathbb{S}^{d-1}),\\ Z\sim\Pi_{\boldsymbol{\theta}+\delta u}}}[g_\delta(\boldsymbol{\theta};u,Z)]-\nabla\mathcal{L}(\boldsymbol{\theta})\right\| = \|\nabla\mathcal{L}_\delta(\boldsymbol{\theta})-\nabla\mathcal{L}(\boldsymbol{\theta})\|\leq\delta L$$

Proof. By Lemma F.1, we have

$$\mathbb{E}_{\substack{u\sim\text{Unif}(\mathbb{S}^{d-1}),\\ Z\sim\Pi_{\boldsymbol{\theta}+\delta u}}}[g(\boldsymbol{\theta};u,Z)] = \nabla\mathcal{L}_\delta(\boldsymbol{\theta})$$

Note that when $\mathcal{L}(\boldsymbol{\theta})$ is differentiable, we have

$$\nabla\mathcal{L}_\delta(\boldsymbol{\theta}) = \nabla\left[\mathbb{E}_{w\sim\text{Unif}(\mathbb{B}^d)}\mathcal{L}(\boldsymbol{\theta}+\delta w)\right] = \mathbb{E}_{w\sim\text{Unif}(\mathbb{B}^d)}\nabla\mathcal{L}(\boldsymbol{\theta}+\delta w)$$

Then under Assumption 3.1, by linearity of expectation and Jensen's inequality, it holds that

$$\|\nabla\mathcal{L}_\delta(\boldsymbol{\theta}) - \nabla\mathcal{L}(\boldsymbol{\theta})\| = \left\|\mathbb{E}_{w\sim\text{Unif}(\mathbb{B}^d)}[\nabla\mathcal{L}(\boldsymbol{\theta}+\delta w) - \nabla\mathcal{L}(\boldsymbol{\theta})]\right\| \le \delta L.$$

□

Corollary F.3. *Under Assumption 3.1 and 3.2, for all $\boldsymbol{\theta}\in\mathbb{R}^d$, it holds that*

$$\|\nabla\mathcal{L}(\boldsymbol{\theta})\| \le 2\sqrt{LG}$$

Proof. Omitted. □

Lemma F.4. (Lipschitz Continuity of Decoupled Risk) *Under Assumption 3.1, 3.2 and 3.3, it holds that*

$$\left|\mathbb{E}_{Z\sim\Pi_{\boldsymbol{\theta}_2}}\left[\ell(\boldsymbol{\theta}_1;Z) - \ell(\boldsymbol{\theta}_2;Z)\right]\right| \le 2(GL_1+\sqrt{LG})\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\| + \frac{L}{2}\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|^2$$

Proof. Let $\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) := \mathbb{E}_{Z\sim\Pi_{\boldsymbol{\theta}_2}}\ell(\boldsymbol{\theta}_1;Z)$ denote the decoupled performative risk, then we have

$$\begin{aligned}
\text{LHS} &= |\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) - \mathcal{L}(\boldsymbol{\theta}_2,\boldsymbol{\theta}_2)| \\
&\le |\mathcal{L}(\boldsymbol{\theta}_1) - \mathcal{L}(\boldsymbol{\theta}_2)| + |\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) - \mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_1)| \\
&\le |\mathcal{L}(\boldsymbol{\theta}_1) - \mathcal{L}(\boldsymbol{\theta}_2) - \langle\nabla\mathcal{L}(\boldsymbol{\theta}_2)\,|\,\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\rangle| + |\langle\nabla\mathcal{L}(\boldsymbol{\theta}_2)\,|\,\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\rangle| + |\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_1) - \mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)| \\
&\overset{(a)}{\le} \frac{L}{2}\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|^2 + |\langle\nabla\mathcal{L}(\boldsymbol{\theta}_2)\,|\,\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\rangle| + |\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_1) - \mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)| \\
&\overset{(b)}{\le} \frac{L}{2}\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|^2 + 2\sqrt{LG}\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\| + |\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_1) - \mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)| \\
&= \frac{L}{2}\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|^2 + 2\sqrt{LG}\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\| + \left|\int \ell(\boldsymbol{\theta}_1;z)\left(\Pi_{\boldsymbol{\theta}_1}(z) - \Pi_{\boldsymbol{\theta}_2}(z)\right)dz\right| \\
&\overset{(c)}{\le} \frac{L}{2}\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|^2 + 2\sqrt{LG}\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\| + 2G\boldsymbol{\delta}_{\text{TV}}\left(\Pi_{\boldsymbol{\theta}_1},\Pi_{\boldsymbol{\theta}_2}\right) \\
&\le \frac{L}{2}\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|^2 + 2\sqrt{LG}\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\| + 2GL_1\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\| \\
&= 2\left(\sqrt{LG}+GL_1\right)\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\| + \frac{L}{2}\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|^2
\end{aligned}$$

where we use Assumption 3.1 in inequality (a), Corollary F.3 in inequality (b), Assumption 3.2 in inequality (c), and Assumption 3.3 in the last inequality. □

Lemma F.5. *Under Assumption 3.5, it holds that for all $0\le\ell\le m$, $m\ge 1$*

$$\boldsymbol{\delta}_{TV}\left(\mathbb{P}(Z_k^{(\ell+1)}\in\cdot|Z_k^{(0)}),\mathbb{P}(\tilde{Z}_k^{(\ell+1)}\in\cdot|Z_k^{(0)})\right) \le L_2\left\|\check{\boldsymbol{\theta}}_k^{(\ell)} - \check{\boldsymbol{\theta}}_k\right\| + \boldsymbol{\delta}_{TV}\left(\mathbb{P}(Z_k^{(\ell)}\in\cdot|Z_k^{(0)}),\mathbb{P}(\tilde{Z}_k^{(\ell)}\in\cdot|Z_k^{(0)})\right)$$

Unfold above recursion leads to the following inequality,

$$\boldsymbol{\delta}_{TV}\left(\mathbb{P}(Z_k^{(m)}\in\cdot|Z_k^{(0)}),\mathbb{P}(\tilde{Z}_k^{(m)}\in\cdot|Z_k^{(0)})\right) \le L_2\sum_{\ell=1}^{m-1}\left\|\check{\boldsymbol{\theta}}_k^{(\ell)} - \check{\boldsymbol{\theta}}_k\right\|, \quad \forall m\ge 1.$$

Proof. Recall the notation $\check{\boldsymbol{\theta}}_k^{(\ell)} = \check{\boldsymbol{\theta}}_k^{(\ell)} + \delta_k u_k, \check{\boldsymbol{\theta}}_k = \check{\boldsymbol{\theta}}_k + \delta_k u_k$, and the fact that $Z_k = Z_k^{(0)} = \tilde{Z}_k^{(0)}$, we have

$$
\begin{aligned}
2 \cdot \text{LHS} &= \int_{\mathsf{Z}} \left| \mathbb{P}(Z_k^{(\ell+1)} = z | Z_k^{(0)}) - \mathbb{P}(\tilde{Z}_k^{(\ell+1)} = z | Z_k^{(0)}) \right| dz \\
&= \int_{\mathsf{Z}} \left| \int_{\mathsf{Z}} \mathbb{P}(Z_k^{(\ell)} = z', Z_k^{(\ell+1)} = z | Z_k^{(0)}) - \mathbb{P}(\tilde{Z}_k^{(\ell)} = z', \tilde{Z}_k^{(\ell+1)} = z | Z_k^{(0)}) dz' \right| dz \\
&\leq \int_{\mathsf{Z}} \int_{\mathsf{Z}} \left| \mathbb{T}_{\check{\boldsymbol{\theta}}_k^{(\ell)}}(z', z) \mathbb{P}(Z_k^{(\ell)} = z' | Z_k^{(0)}) - \mathbb{T}_{\check{\boldsymbol{\theta}}_k}(z', z) \mathbb{P}(\tilde{Z}_k^{(\ell)} = z' | Z_k^{(0)}) \right| dz' dz \\
&\leq \int_{\mathsf{Z}} \int_{\mathsf{Z}} \left| \mathbb{T}_{\check{\boldsymbol{\theta}}_k^{(l\ell}}(z', z) \mathbb{P}(Z_k^{(\ell)} = z' | Z_k^{(0)}) - \mathbb{T}_{\check{\boldsymbol{\theta}}_k}(z', z) \mathbb{P}(Z_k^{(\ell)} = z' | Z_k^{(0)}) \right| dz' dz \\
&\quad + \int_{\mathsf{Z}} \int_{\mathsf{Z}} \left| \mathbb{T}_{\check{\boldsymbol{\theta}}_k}(z', z) \mathbb{P}(Z_k^{(\ell)} = z' | Z_k^{(0)}) - \mathbb{T}_{\check{\boldsymbol{\theta}}_k}(z', z) \mathbb{P}(\tilde{Z}_k^{(\ell)} = z' | Z_k^{(0)}) \right| dz' dz \\
&\overset{(a)}{=} \int_{\mathsf{Z}} \mathbb{P}(Z_k^{\ell} = z' | Z_k^{(0)}) \int_{\mathsf{Z}} \left| \mathbb{T}_{\check{\boldsymbol{\theta}}_k}(z', z) - \mathbb{T}_{\check{\boldsymbol{\theta}}_k^{(\ell)}}(z', z) \right| dz dz' \\
&\quad + \int_{\mathsf{Z}} \left[\int_{\mathsf{Z}} \mathbb{T}_{\check{\boldsymbol{\theta}}_k}(z', z) dz \right] \left| \mathbb{P}(Z_k^{(\ell)} = z' | Z_k^{(0)}) - \mathbb{P}(\tilde{Z}_k^{(\ell)} = z' | Z_k^{(0)}) \right| dz' \\
&\leq \int_{\mathsf{Z}} \mathbb{P}(Z_k = z' | Z_k^{(0)}) \cdot 2\boldsymbol{\delta}_{\text{TV}} \left(\mathbb{T}_{\check{\boldsymbol{\theta}}_k}(z', \cdot), \mathbb{T}_{\check{\boldsymbol{\theta}}_k}(z', \cdot) \right) dz' + 2\boldsymbol{\delta}_{\text{TV}} \left(\mathbb{P}(Z_k^{(\ell)} \in \cdot | Z_k^{(0)}), \mathbb{P}(\tilde{Z}_k^{(\ell)} \in \cdot | Z_k^{(0)}) \right) \\
&\leq 2 \int_{\mathsf{Z}} \mathbb{P}(Z_k^{(\ell)} = z' | Z_k^{(0)}) dz' \cdot L_2 \left\| \check{\boldsymbol{\theta}}_k^{(\ell)} - \check{\boldsymbol{\theta}}_k \right\| + 2\boldsymbol{\delta}_{\text{TV}} \left(\mathbb{P}(Z_k^{(\ell)} \in \cdot | Z_k^{(0)}), \mathbb{P}(\tilde{Z}_k^{(\ell)} \in \cdot | Z_k^{(0)}) \right) \\
&= 2 \left[L_2 \left\| \check{\boldsymbol{\theta}}_k^{(\ell)} - \check{\boldsymbol{\theta}}_k \right\| + \boldsymbol{\delta}_{\text{TV}} \left(\mathbb{P}(Z_k^{(\ell)} \in \cdot | Z_k^{(0)}), \mathbb{P}(\tilde{Z}_k^{(\ell)} \in \cdot | Z_k^{(0)}) \right) \right] = 2 \cdot \text{RHS}
\end{aligned}
$$

where inequality (a) holds due to the (absolutely) integrable condition (which automatically holds for probability density functions and kernels), and Assumption 3.5 is used in the last inequality. □

Assumption F.6. Assume that there exists constants L_0, L_1 such that:

(i) $|\ell(\boldsymbol{\theta}; z) - \ell(\boldsymbol{\theta}; z')| \leq L_0 \|z - z'\|$ for any $\boldsymbol{\theta} \in \mathbb{R}^d$, $z, z' \in \mathsf{Z}$,

(ii) $W_1(\Pi_{\boldsymbol{\theta}}, \Pi_{\boldsymbol{\theta}'}) \leq L_1 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$ for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$, where $W_1(\Pi, \Pi')$ denotes the Wasserstein-1 distance between the distributions Π, Π'.

We observe that a similar result to Lemma F.4 can be proven by replacing Assumption 3.3 with Assumption F.6:

Lemma F.7. (Lipschitz Continuity of Decoupled Risk, Alternative Condition based on Wasserstein-1 distance.) *Under Assumption 3.1, 3.2, F.6. Then, for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$, it holds that*

$$
\left| \mathbb{E}_{Z \sim \Pi_{\boldsymbol{\theta}_2}} \left[\ell(\boldsymbol{\theta}_1; Z) - \ell(\boldsymbol{\theta}_2; Z) \right] \right| \leq (L_0 L_1 + 2\sqrt{LG}) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \frac{L}{2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 .
$$

Proof. Observe that

$$
\begin{aligned}
\text{LHS} &= |\mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) - \mathcal{L}(\boldsymbol{\theta}_2, \boldsymbol{\theta}_2)| \\
&\leq |\mathcal{L}(\boldsymbol{\theta}_1) - \mathcal{L}(\boldsymbol{\theta}_2)| + |\mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) - \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1)| \\
&\leq |\mathcal{L}(\boldsymbol{\theta}_1) - \mathcal{L}(\boldsymbol{\theta}_2) - \langle \nabla \mathcal{L}(\boldsymbol{\theta}_2) \,|\, \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle| + |\langle \nabla \mathcal{L}(\boldsymbol{\theta}_2) \,|\, \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle| + |\mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1) - \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)| \\
&\leq \frac{L}{2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 + |\langle \nabla \mathcal{L}(\boldsymbol{\theta}_2) \,|\, \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \rangle| + |\mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1) - \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)| \\
&\leq \frac{L}{2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 + 2\sqrt{LG} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + |\mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1) - \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)| \\
&= \frac{L}{2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 + 2\sqrt{LG} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + \left| \mathbb{E}_{Z \sim \Pi_{\boldsymbol{\theta}_1}, Z' \sim \Pi_{\boldsymbol{\theta}_2}} [\ell(\boldsymbol{\theta}_1; Z) - \ell(\boldsymbol{\theta}_1; Z')] \right| \\
&\overset{(a)}{\leq} \frac{L}{2} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 + 2\sqrt{LG} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| + L_0 W_1(\Pi_{\boldsymbol{\theta}_1}, \Pi_{\boldsymbol{\theta}_2})
\end{aligned}
$$

$$\overset{(b)}{\leq} \frac{L}{2}\left\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\right\|^2+2\sqrt{LG}\left\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\right\|+L_0L_1\left\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\right\|$$
$$=\left(L_0L_1+2\sqrt{LG}\right)\left\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\right\|+\frac{L}{2}\left\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\right\|^2$$

where the inequality (a) is due to [Lemma D.4, (Perdomo et al., 2020)] and the alternative assumption (i), the inequality (b) is due to the alternative assumption (ii) (i.e., Lipschitz condition on distribution map $\Pi_{\boldsymbol{\theta}}$ in Wasserstein-1 metric). □

Consequently, the conclusions in Theorem 3.6 hold (with slightly different constants) when Assumption 3.3 is replaced by Assumption F.6. We observe that the former assumption is only used in ensuring the bound in Lemma F.4; cf. the proof of Lemma C.2.