Online Performative Gradient Descent for Learning Nash Equilibria in Decision-Dependent Games

Zihan Zhu* Ethan X. Fang† Zhuoran Yang‡

Abstract

We study multi-agent games within the innovative framework of decision-dependent games, which establishes a feedback mechanism that population data reacts to agents' actions and further characterizes the strategic interactions among agents. We focus on finding the Nash equilibrium of decision-dependent games. However, gradients of reward functions are unknown due to the strategic interactions between agents, and classical gradient-based methods are infeasible. To overcome this challenge, we model the strategic interactions by a general parametric model and propose a novel online algorithm, Online Performative Gradient Descent (`OPGD`), which leverages the ideas of online stochastic approximation and projected gradient descent to learn the Nash equilibrium in the context of function approximation for the unknown gradient. In particular, under mild assumptions on the function classes defined in the parametric model, we prove that the `OPGD` algorithm finds the Nash equilibrium efficiently for strongly monotone decision-dependent games. Synthetic numerical experiments validate our theory.

1 Introduction

The classical theory of learning and prediction fundamentally relies on the assumption that data follows a static distribution. This assumption, however, does not account for many dynamic real-world scenarios where decisions can influence the data involved. Recent literature on performative classification (Hardt et al., 2016; Dong et al., 2018; Miller et al., 2020) and performative prediction (Perdomo et al., 2020) offers a variety of examples where agents are strategic, and data is performative. For instance, in the ride-sharing market, both passengers and drivers engage with multiple platforms using various strategies such as "price shopping". Consequently, these platforms observe performative demands, and the pricing policy becomes strategically coupled.

In this paper, we explore the multi-agent performative prediction problem, specifically, the multi-agent decision-dependent games, as proposed by Narang et al. (2022). We aim to develop online algorithms to find Nash equilibria with the first-order oracle, and further extend it to the bandit feedback setting. In this scenario, agents can only access their utility functions instead of gradients through the oracle. Finding Nash equilibria in decision-dependent games is a challenging

*Department of Statistical Science, Duke University. zihan.zhu@duke.edu

†Department of Biostatistics and Bioinformatics, Duke University. xingyuan.fang@duke.edu

‡Department of Statistics and Data Science, Yale University. zhuoran.yang@yale.edu

task. Most existing works primarily focus on finding performative stable equilibria within the single-agent setting, an approach that approximates the Nash equilibrium and is relatively straightforward to compute (Mendler-Dünner et al., 2020; Wood et al., 2021; Drusvyatskiy and Xiao, 2022; Brown et al., 2022; Li and Wai, 2022).

There are two major challenges associated with this problem: (i) the distribution shift induced by performative data, and (ii) the lack of first-order information for the performative gradient due to the strategic interaction between agents. To address these two challenges, we propose a novel online gradient-based algorithm, Online Performative Gradient Descent (`OPGD`). In particular, our algorithm employs a general parametric framework to model the decision-dependent distribution, which provides an unbiased estimator for the unknown gradient, and leverages online stochastic approximation methods to estimate the parametric functions.

1.1 Major Contributions

Our work provides new fundamental understandings of decision-dependent games. Expanding upon the linear parametric assumption in Narang et al. (2022), we propose a more comprehensive parametric framework that models decision-dependent distributions of the observed data. We also derive sufficient conditions under this parametric framework that guarantee a strongly monotone decision-dependent game, thereby ensuring a unique Nash equilibrium.

From the algorithmic perspective, we propose `OPGD` , the first online algorithm to find the Nash equilibrium under linear and kernel parametric models. While the existing algorithm only handles the linear case and cannot be extended to the non-linear parametric model (Section 3), and `OPGD` uses an essentially different method to learn the strategic interaction between agents. Under the proposed parametric framework, learning the Nash equilibrium in decision-dependent games can be formulated as a bilevel problem, where the lower level is learning the strategic model and the upper level is finding equilibria. We acknowledge this learning framework bridges online optimization and statistical learning with time-varying models. Besides, the `OPGD` algorithm leverages the ideas of online stochastic approximation for the lower problem and projected gradient descent to learn the Nash equilibrium. Further, we extend `OPGD` into the bandit feedback setting by the analogous idea.

We further prove that under mild assumptions, `OPGD` converges to the Nash equilibrium. Given the first-order oracle, `OPGD` achieves a convergence rate of $\mathcal{O}(t^{-1})$ under the linear parametric model. This rate matches the optimal rate of SGD in the strongly convex setting. For the kernel function class $\mathcal{H}$ that associated with a bounded kernel K, we posit that the parametric functions reside within the power space $\mathcal{H}^\beta$ and evaluate the approximation error of `OPGD` under the α-power norm, where α represents the minimal value that ensures the power space $\mathcal{H}^\alpha$ possesses a bounded kernel. We present the first analysis for online stochastic approximation under the power norm (Lemma 4.18), in contrast to the classical RKHS norm (Tarres and Yao, 2014; Pillaud-Vivien et al., 2018; Lei et al., 2021). The difference between the RKHS $\mathcal{H}$ and the power space $\mathcal{H}^\beta$ makes the standard techniques fail under the power norm, and we use novel proof steps to obtain the estimation error bound. We demonstrate that `OPGD` leverages the embedding property of the kernel K to accelerate convergence and achieves the rate of $\mathcal{O}(t^{-\frac{\beta-\alpha}{\beta-\alpha+2}})$. Moreover, `OPGD` can handle the challenging scenario, where parametric functions are outside the RKHS. We further extend the analysis into the bandit feedback setting and obtain an analogous convergence rate. See Section 4 for more details.

1.2 Related Work

Performative prediction. The multi-agent decision-dependent game in this paper is inspired by the performative prediction framework (Perdomo et al., 2020). This framework builds upon the pioneering works of strategic classification (Hardt et al., 2016; Dong et al., 2018; Miller et al., 2020), and extends the classical statistical theory of risk minimization to incorporate the performativity of data. Perdomo et al. (2020); Mendler-Dünner et al. (2020); Miller et al. (2021) introduce the concepts of performative optimality and stability, demonstrating that repeated retraining and stochastic gradient methods converge to the performatively stable point. Miller et al. (2021), in pursuit of the performatively optimal point, model the decision-dependent distribution using location families and propose a two-stage algorithm. Similarly, Izzo et al. (2021) develop algorithms to estimate the unknown gradient using finite difference methods. More recently, Narang et al. (2022); Piliouras and Yu (2022) expand the performative prediction to the multi-agent setting, deriving algorithms to find the performatively optimal point.

Learning in continuous games. Our work aligns closely with optimization in continuous games. Rosen (1965) lays the groundwork, deriving sufficient conditions for a unique Nash equilibrium in convex games. For strongly monotone games, Bravo et al. (2018); Mertikopoulos and Zhou (2019); Lin et al. (2021) achieve the convergence rate and iteration complexity of stochastic and derivative-free gradient methods. For monotone games, the convergence of such methods is established by Tatarenko and Kamgarpour (2019, 2020). Additional with bandit feedback settings, zeroth-order methods (or derivative-free methods) achieve convergence (Bravo et al., 2018; Lin et al., 2021; Drusvyatskiy et al., 2022; Narang et al., 2022), albeit with slow convergence rates (Shamir, 2013; Lin et al., 2021; Narang et al., 2022). Relaxing the convex assumption, Ratliff et al. (2016); Agarwal et al. (2019); Cotter et al. (2019) study non-convex continuous games in various settings.

Learning with kernels. Our proposed algorithm closely relies on stochastic approximation, utilizing online kernel regression for the RKHS function class. Prior research investigates the generalization capability of least squares and ridge regression in RKHS De Vito et al. (2005); Caponnetto and De Vito (2007); Smale and Zhou (2007); Rosasco et al. (2010); Mendelson and Neeman (2010). Meanwhile, extensive works study algorithms for kernel regression. For instance, Yao et al. (2007); Dieuleveut and Bach (2016); Pillaud-Vivien et al. (2018); Lin and Rosasco (2017); Lei et al. (2021) propose offline algorithms with optimal convergence rates under the RKHS norm and L^2 norm using early stopping and stochastic gradient descent methods, while Ying and Pontil (2008); Tarres and Yao (2014); Dieuleveut and Bach (2016) design online algorithms with optimal convergence rates. The convergence of kernel regression in power norm (or Sobolev norm) is studied in Steinwart et al. (2009); Fischer and Steinwart (2020); Liu and Li (2020); Lu et al. (2022), with offline spectral filter algorithms achieving the statistical optimal rate under the power norm (Pillaud-Vivien et al., 2018; Blanchard and Mücke, 2018; Lin and Cevher, 2020; Lu et al., 2022).

Notation. We introduce some useful notation before proceeding. Throughout this paper, we denote the set $1, 2, \cdots, n$ by $[n]$ for any positive integer n. For two positive sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, we write $a_n = \mathcal{O}(b_n)$ or $a_n \lesssim b_n$ if there exists a positive constant C such that $a_n \le C \cdot b_n$. For any integer d, we denote the d-dimensional Euclidean space by $\mathbb{R}^d$, with inner produce $\langle\cdot,\cdot\rangle$ and the induced norm $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$. For a Hilbert space $\mathcal{H}$, let $\|\cdot\|_{\mathcal{H}}$ be the associated Hilbert norm. For a set $\mathcal{X}$ and a probability measure $\rho_{\mathcal{X}}$ on $\mathcal{X}$, let $\mathcal{L}^2_{\rho_{\mathcal{X}}}$ be the L^2 space on $\mathcal{X}$ induced by the measure $\rho_{\mathcal{X}}$, equipped with inner product $\langle\cdot,\cdot\rangle_{\rho_{\mathcal{X}}}$ and L^2 norm $\|\cdot\|_{\rho_{\mathcal{X}}} = \sqrt{\langle\cdot,\cdot\rangle_{\rho_{\mathcal{X}}}}$. For any matrix $A = (a_{ij})$, the Frobenius norm and the operator norm (or spectral norm) of A are $\|A\|_F = (\sum_{i,j} a_{ij}^2)^{1/2}$ and $\|A\|_{\text{op}} = \sigma_1(A)$, where $\sigma_1(A)$ stands for the largest singular value of A. For any square matrix $A = (a_{ij})$, denote its trace by $\text{tr}(A) = \sum_i a_{ii}$. For any $y \in \mathbb{R}^d$, we denote its projection onto a set $\mathcal{X} \subset \mathbb{R}^d$ by $\text{proj}_{\mathcal{X}}(y) = \arg\min_{x\in\mathcal{X}}\|x - y\|$. The set denoted by $N_{\mathcal{X}}(x)$ represents the normal cone to a convex set $\mathcal{X}$ at $x \in \mathcal{X}$, namely, $N_{\mathcal{X}}(x) = \{v \in \mathbb{R}^d : \langle v, y - x\rangle \le 0,\ \forall y \in \mathcal{X}\}$. For any metric space $\mathcal{Z}$ with metric $d(\cdot,\cdot)$, the symbol $\mathbb{P}(\mathcal{Z})$ will denote the set of Radon probability measures μ on $\mathcal{Z}$ with a finite first moment $\mathbb{E}_{z\sim\mu}[d(z, z_0)] < \infty$ for some $z_0 \in \mathcal{Z}$.

2 Preliminaries and Problem Formulation

We briefly introduce the formulation of n-agent decision-dependent games based on Narang et al. (2022). In this setting, each agent $i \in [n]$ takes the action $x_i \in \mathcal{X}_i$ from an action set $\mathcal{X}_i \subset \mathbb{R}^{d_i}$. Define the joint action $x \coloneqq (x_1, x_2, \cdots, x_n) \in \mathcal{X}$ and the joint action set $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \subset \mathbb{R}^d$, where $d \coloneqq \sum_{i=1}^n d_i$. For all $i \in [n]$, we write $x = (x_i, x_{-i})$, where x_{-i} denotes the vector of all coordinates except x_i. Let $\mathcal{L}_i : \mathcal{X} \to \mathbb{R}$ be the utility function of agent i. In the game, each agent i seeks to solve the problem

$$\min_{x_i\in\mathcal{X}_i} \mathcal{L}_i(x_i, x_{-i}), \qquad \text{where} \qquad \mathcal{L}_i(x) \coloneqq \mathop{\mathbb{E}}_{z_i\sim\mathcal{D}_i(x)} \ell_i(x, z_i). \tag{2.1}$$

Here $z_i \in \mathcal{Z}_i$ represents the data observed by agent i, where the sample space $\mathcal{Z}_i$ is assumed to be $\mathcal{Z}_i = \mathbb{R}^p$ with $p \in \mathbb{N}$ throughout this paper. Moreover, $\mathcal{D}_i : \mathcal{X} \to \mathbb{P}(\mathcal{Z}_i)$ is the distribution map, and $\ell_i : \mathbb{R}^d \times \mathcal{Z}_i \to \mathbb{R}$ denotes the loss function. During play, each agent i performs an action x_i and observes performative data $z_i \sim \mathcal{D}_i(x)$, where the performativity is modeled by the decision-dependent distribution $\mathcal{D}_i(x)$. In the round t, the agent i only has access to $z_i^1, \cdots, z_i^{t-1}$ as well as $x^1, \cdots, x^{t-1}$ and seeks to solve the ERM version of (2.1). We assume the access to the first-order oracle, namely, loss functions ℓ_i are known to agents but distribution maps $\mathcal{D}_i$ are unknown.

Definition 2.1. (Nash equilibrium). In the game (2.1), a joint action $x^* = (x_1^*, x_2^*, \cdots, x_n^*)$ is a Nash equilibrium (Nash Jr, 1996) if all agents play the best response against other agents, namely,

$$x_i^* = \mathop{\arg\min}_{x_i\in\mathcal{X}_i} \mathcal{L}_i(x_i, x_{-i}^*) = \mathop{\arg\min}_{x_i\in\mathcal{X}_i} \mathop{\mathbb{E}}_{z_i\sim\mathcal{D}_i(x_i,x_{-i}^*)} \ell_i(x_i, x_{-i}^*, z_i), \qquad \forall i \in [n]. \tag{2.2}$$

In general continuous games, Nash equilibria may not exist or there might be multiple Nash equilibria (Fudenberg and Tirole, 1991). The existence and uniqueness of a Nash equilibrium in a continuous game depend on the game's structure and property. In general, finding the unique Nash equilibrium is only possible for convex and strongly monotone games (Debreu, 1952).

Definition 2.2. (Convex game). Game (2.1) is a convex game if sets $\mathcal{X}_i$ are non-empty, compact, convex and utility functions $\mathcal{L}_i(x_i, x_{-i})$ are convex in x_i when x_{-i} are fixed.

Suppose that utility functions $\mathcal{L}_i$ are differentiable, we use $\nabla_i \mathcal{L}_i(x)$ to denote the gradient of $\mathcal{L}_i(x)$ with respect to x_i (the i-th individual gradient). We say the game (2.1) is C^1-smooth if the gradient $\nabla_i \mathcal{L}_i(x)$ exists and is continuous for all $i \in [n]$. Using this notation, we define the gradient $H(x)$ comprised of individual gradients

$$H(x) := (\nabla_1 \mathcal{L}_1(x), \cdots, \nabla_n \mathcal{L}_n(x)).$$

Definition 2.3. (Strongly monotone game). For a constant $\tau \geq 0$, a C^1-smooth convex game (2.1) is called τ-strongly monotone if it satisfies

$$\langle H(x) - H(x'), x - x' \rangle \geq \tau \|x - x'\|^2, \quad \text{for all } x, x' \in \mathcal{X}.$$

Note that a τ-strongly monotone game ($\tau > 0$) over a compact and convex action set $\mathcal{X}$ admits a unique Nash equilibrium (Rosen, 1965). According to the optimal conditions in convex optimization (Boyd et al., 2004), this Nash equilibrium x^* is characterized by the variational inequality

$$0 \in H(x^*) + N_{\mathcal{X}}(x^*). \tag{2.3}$$

The agents in the game (2.1) are strategically coupled in two ways. First, the data z_i seen by agents is influenced by the joint action x, since each of them follows a decision-dependent distribution $\mathcal{D}_i(x)$. Second, the loss functions ℓ_i depend on the joint action x and the observed data z_i. Note that the decision-dependence in distributions $\mathcal{D}_i(x)$ may involve the reaction of strategic users in a population to the announced joint action x. This interaction structure between the decision-maker and the strategic users induces a game in the environment, which is known as a Stackelberg game [Von Stackelberg (2010)]. In the game (2.1), we aggregate the strategic interaction between strategic agents and strategic users in distributions $\mathcal{D}_i(x)$.

Next, we use a real-world example of the ride-share market [Hardt et al. (2016); Perdomo et al. (2020); Narang et al. (2022)] to digest the decision-dependent game (2.1).

Example 2.4. (Revenue Maximization via Demand Forecasting). In the ride-sharing market, several platforms act as strategic agents (suppose there are n platforms), predicting ride demands of strategic users in a city to maximize revenue. Typically, both drivers and passengers, regarded as strategic users, engage with multiple platforms by employing tactics such as "price shopping". To elaborate, users call the ride in multiple platforms, and each platform i presents its price and time cost (action x_i) for users. Strategic users compare prices and time costs among these platforms and choose the best one. Consequently, the forecasted ride demand z_i for platform i, which is generated by the strategic users, relies on the platform's own decision x_i as well as the choices of competitors x_{-i}, thereby shaping the distributions $z_i \sim \mathcal{D}_i(x)$.

Example 2.5. (University Admissions). Multiple universities, acting as strategic agents, evaluate applications to decide on admissions. Each applicant, considered a strategic user, tailors their application to meet the criteria of universities. Every university i evaluates numerous applications, represented by data z_i (might contain GPA and other related grades), and formulates a rule x_i to decide which candidates are admitted. Each university's goal is to accept qualified students, and applicants may apply to various universities. To elaborate, students might compare different programs by assessing their admission rules and selecting several universities that match their qualifications. Consequently, the predicted applications z_i received by the university i are shaped by the joint rule x, thus formulating the decision-dependent distribution $z_i \sim \mathcal{D}_i(x)$. Furthermore, every university assesses the quality of students using a loss function, denoted as $\ell_i(x, z_i)$, and subsequently forms a decision-dependent game.

2.1 A Peek into Decision-Dependent Game: Why Challenging?

We briefly talk about the challenges and our idea of designing the algorithm. In decision-dependent games, the classical theory of risk minimization does not work. [Perdomo et al. (2020); Narang et al. (2022)] propose the repeated retraining method, or repeated risk minimization algorithm for the game (2.1). The idea is to decouple the effects of joint action x on loss functions ℓ_i and distributions $\mathcal{D}_i(x)$. This method repeatedly minimizes the utility function with the distribution map $\mathcal{D}_i$ fixed at the result of the previous iteration:

$$x^{t+1} = \arg\min_{x \in \mathcal{X}} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(x^t)} \ell_i(x, z_i). \tag{2.4}$$

In each iteration, distribution $\mathcal{D}_i(x^t)$ is fixed and (2.4) is a regular optimization problem. We can derive the corresponding repeated gradient descent algorithm

$$x^{t+1} = \mathrm{proj}_{\mathcal{X}} \left(x^t - \eta \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(x^t)} \nabla_i \ell(x^t, z_i) \right). \tag{2.5}$$

Repeated retraining is numerically feasible but it fails to find the Nash equilibrium. In fact, the update rule (2.5) is a biased gradient descent because it only uses the term $P_i(x)$ rather than the full gradient $\nabla_i \mathcal{L}_i(x)$. As a result, this algorithm converges to the so-called performatively stable equilibrium instead of the Nash equilibrium.

The primary obstacles to finding the Nash equilibrium in the game (2.1) include: (i) the distribution shift induced by performative data, and (ii) the lack of first-order information for the performative gradient. To make it clear, standard methods, such as gradient-based algorithms, necessitate the gradient $H(x)$. However, $H(x)$ is unknown since distributions $\mathcal{D}_i$ are unknown, and estimating $H(x)$ is complex due to the dependency between $\mathcal{D}_i(x)$ and x. Mathematically, assuming C^1-smoothness, the chain rule directly yields the following expression for the gradient

$$\nabla_i \mathcal{L}_i(x) = \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(x)} \nabla_i \ell_i(x_i, x_{-i}, z_i) + \frac{d}{du_i} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(u_i, x_{-i})} \ell_i(x_i, x_{-i}, z_i) \Big|_{u_i = x_i}, \tag{2.6}$$

where $\nabla_i \ell_i(x, z_i)$ denotes the gradient of $\ell_i(x, z_i)$ with respect to x_i. The main difficulty is estimating the second term in (2.6) due to the absence of closed-form expressions.

To estimate the unknown gradient $H(x)$, we impose a parametric assumption on the observed data z_i and model the distribution maps $\mathcal{D}_i$ using parametric functions. Note that the linear parametric assumption was first proposed in Narang et al. (2022). In this paper, we extend this assumption to a general framework and show that under the parametric assumption, the gradient $H(x)$ has a closed-form expression, which yields an unbiased estimator for $H(x)$.

Assumption 2.6. (Parametric assumption). Suppose there exists a function class $\mathscr{F}$ and p-dimensional functions $f_i : \mathcal{X} \to \mathbb{R}^p$ over the joint action set $\mathcal{X}$ such that $f_i \in \mathscr{F}^p$ (i.e. each coordinate of f_i is in the function class $\mathscr{F}$) and

$$z_i \sim \mathcal{D}_i(x) \Longleftrightarrow z_i = f_i(x) + \epsilon_i, \quad \forall i \in [n],$$

where $\epsilon_i \in \mathbb{R}^p$ are zero-mean noise terms with finite variance σ^2, namely, $\mathbb{E}\epsilon_i = 0$ and $\mathbb{E}\|\epsilon_i\|^2 \leq \sigma^2$.

Under Assumption 2.6, assuming that f_i are differentiable and letting $\mathcal{P}_i$ be the distribution of the noise term ϵ_i, we derive the following expression for the utility functions $\mathcal{L}_i(x) = \mathbb{E}_{z_i \sim \mathcal{D}_i(x)} \ell_i(x, z_i) = \mathbb{E}_{\epsilon_i \sim \mathcal{P}_i} \ell_i(x, f_i(x) + \epsilon_i)$. Then the individual gradient would be $\nabla_i \mathcal{L}_i(x) = \nabla_i \mathbb{E}_{z_i \sim \mathcal{D}_i(x)} \ell_i(x, z_i) = \nabla_i [\mathbb{E}_{\epsilon_i \sim \mathcal{P}_i} \ell_i(x, f_i(x) + \epsilon_i)]$. Consequently, the chain rule directly implies the following expression

$$\nabla_i \mathcal{L}_i(x) = \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(x)} \nabla_i \ell_i(x, z_i) + \left(\frac{\partial f_i(x)}{\partial x_i} \right)^\top \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(x)} \nabla_{z_i} \ell_i(x, z_i), \tag{2.7}$$

where $\nabla_{z_i} \ell_i(x, z_i)$ denotes the gradient of $\ell_i(x, z_i)$ with respect to z_i. Given a joint action x, each agent i observes data $z_i \sim \mathcal{D}_i(x)$. Equation (2.7) suggests the following unbiased estimator for $H(x)$:

$$\widehat{H}(x) \coloneqq \left(\widehat{\nabla}_i \mathcal{L}_i(x) \right)_{i \in [n]} = \left(\nabla_i \ell_i(x, z_i) + \left(\frac{\partial f_i(x)}{\partial x_i} \right)^\top \nabla_{z_i} \ell_i(x, z_i) \right)_{i \in [n]}. \tag{2.8}$$

However, direct computation of $\widehat{H}(x)$ is infeasible because f_i are unknown. To overcome this challenge, we approximate the unknown functions f_i with the function class $\mathscr{F}^p$. In fact, the estimation of f_i can be formed as a non-parametric regression problem, namely,

$$\widehat{f}_i = \mathop{\arg\min}_{f \in \mathscr{F}^p} \int_{\mathcal{X} \times \mathcal{Z}_i} \|z_i - f(x)\|^2 d\rho_i, \quad \forall i \in [n], \tag{2.9}$$

where ρ_i is the joint distribution of (x, z_i) induced by $x \sim \rho_{\mathcal{X}}$ and $z_i \sim \mathcal{D}_i(x)$. Here $\rho_{\mathcal{X}}$ is a user-specified sampling distribution on $\mathcal{X}$ and has full support.

3 The `OPGD` Algorithm

In this section, we derive gradient-based online algorithms to find the Nash equilibrium in the game (2.1), namely, the Online Performative Gradient Descent (`OPGD`). In Section 3.1, we formulate the problem into bi-level optimization under Assumption 2.6. In Section 3.2, we consider $\mathscr{F}$ to be the linear and kernel function classes and derive the `OPGD` under first-order oracle. In Section 3.3, we extend the kernel `OPGD` into the bandit feedback setting.

3.1 Bi-Level Formulation

To derive gradient-based algorithms, the first task is to estimate the unknown gradient $H(x)$. Recalling the unbiased estimator $\widehat{H}(x)$ defined in (2.8), a natural method is estimating unknown functions f_i and using the estimation to compute the estimator $\widehat{H}(x)$. In more detail, the estimation of f_i can be formed as a non-parametric regression problem (2.9. In each iteration t, the algorithm gets a point $u_i^t \sim \rho_{\mathcal{X}}$ in the joint action set $\mathcal{X}$ following a fixed distribution $\rho_{\mathcal{X}}$ and draws a sample $y_i^t \sim \mathcal{D}_i(u_i^t)$. Then for any iteration T, $\{(u_i^k, y_i^k)\}_{k\in[T]}$ are i.i.d. random variables. One might minimize the empirical risk of (2.9) as estimations for f_i, namely,

$$\widehat{f}_i = \arg\min_{f\in\mathscr{F}^p} \frac{1}{T}\sum_{k=1}^{T} \|y_i^k - f(u_i^k)\|^2, \qquad \forall i \in [n]. \tag{3.1}$$

Thus, (2.8) and (3.1) together yield an estimator for the gradient:

$$\widehat{\nabla}_i \mathcal{L}_i(x) \coloneqq \nabla_i \ell_i(x, z_i) + \left(\frac{\partial \widehat{f}_i(x)}{\partial x_i}\right)^{\top} \nabla_{z_i} \ell_i(x, z_i). \tag{3.2}$$

In each iteration t, assuming that $x^t \coloneqq (x_1^t, \cdots, x_n^t)$ is the output of the previous iteration, `OPGD` performs the following update for all $i \in [n]$:

(i) (Estimation update). Update the estimation of f_i by online stochastic approximation for (2.9).

(ii) (Individual gradient update). Compute the estimator (2.8) and perform projected gradient steps

$$x_i^{t+1} = \mathrm{proj}_{\mathcal{X}_i}(x_i^t - \eta_t \widehat{\nabla}_i \mathcal{L}_i(x^t)), \qquad \forall i \in [n].$$

In fact, we formulate the learning of Nash equilibria into a bi-level optimization problem. The lower-level problem is learning the parametric model, and the upper-level problem is finding the Nash equilibrium. Moreover, step (i) solves the lower-level problem by stochastic approximation, and step (ii) solves the upper-level problem by projected gradient descent.

3.2 Learning with First-Order Oracle

In this section, we derive the `OPGD` for both linear and kernel parametric models given the first-order oracle. While the linear `OPGD` performs exactly the same as steps (i) and (ii), the kernel `OPGD` adds a dynamic regularization term due to the infinite dimension of RKHS.

Linear Function Class. Let $\mathscr{F}$ be the linear function class, namely, $f_i(x) = A_i x$ for $i \in [n]$, where $A_i \in \mathbb{R}^{p\times d}$ are unknown matrices. Then (2.9) becomes the least square problem $A_i = \arg\min_{A\in\mathbb{R}^{p\times d}} \mathbb{E}_{(u_i,y_i)\sim\rho_i} \|y_i - A_i u_i\|^2$ with random variables $u_i \sim \rho_{\mathcal{X}}, y_i \sim \mathcal{D}_i(u_i)$. We use the gradient of the least square objective $\|y_i - A_i u_i\|^2$ to derive the online least square update: $A^{\mathrm{new}} \leftarrow A - \nu(Au_i - y_i)u_i^{\top}$ (Dieuleveut et al., 2017; Narang et al., 2022). In each iteration t, we suppose that

A_i^{t-1} is the estimation of A_i from the previous iteration, `OPGD` samples $u_i^t \sim \rho_{\mathcal{X}}$ and $y_i^t \sim \mathcal{D}_i(u_i^t)$ and performs the following estimation update:

$$\text{(i)} \quad A_i^t = A_i^{t-1} - \nu_t \left(A_i^{t-1} u_i^t - y_i^t\right) (u_i^t)^\top. \tag{3.3}$$

Recalling (2.7), the individual gradient is $\nabla_i \mathcal{L}_i(x) = \mathbb{E}_{z_i \sim \mathcal{D}_i(x)} \left[\nabla_i \ell_i(x, z_i) + A_{ii}^\top \nabla_{z_i} \ell_i(x, z_i)\right]$, where $A_{ii} = \partial f_i(x)/\partial x_i \in \mathbb{R}^{p \times d_i}$ denotes the submatrix of A_i whose columns are indexed by the agent i. After step (i), `OPGD` draws a sample $z_i^t \sim \mathcal{D}_i(x^t)$ and compute the estimator (2.8) to perform the projected gradient step:

$$\text{(ii)} \quad x_i^{t+1} = \text{proj}_{\mathcal{X}_i} \left(x_i^t - \eta_t \left(\nabla_i \ell_i(x^t, z_i^t) + (A_{ii}^t)^\top \nabla_{z_i} \ell_i(x^t, z_i^t)\right)\right). \tag{3.4}$$

Kernel Function Class Now we consider $\mathscr{F}$ as the kernel function class, namely, we suppose $f_i \in (\mathcal{H})^p$, where $\mathcal{H}$ is an RKHS induced by a Mercer kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and a user-specified probability measure $\rho_{\mathcal{X}}$. By the reproducing property of $\mathcal{H}$, f_i can be represented as $f_i(x) = \langle f_i, \phi_x \rangle_{\mathcal{H}}$, where $\phi : \mathcal{X} \to \mathcal{H}$ is the feature map, i.e. $\phi_x \coloneqq K(\cdot, x) \in \mathcal{H}$ for any $x \in \mathcal{X}$. Therefore, (2.9) becomes the kernel regression $\arg\min_{f \in \mathscr{F}^p} \mathbb{E}_{(u_i, y_i) \sim \rho_i} \|y_i - \langle f, \phi_{u_i} \rangle_{\mathcal{H}}\|^2$.

The extension of linear `OPGD` into the non-linear case is nontrivial. The major difficulty in the RKHS case is that $\mathcal{H}$ generally has infinite dimensions, and solving the ERM version of (3.1) leads to ill-posed solutions. Consequently, we consider the regularized kernel ridge regression $\arg\min_{f \in \mathscr{F}^p} \mathbb{E}_{(u_i, y_i) \sim \rho_i} \|y_i - \langle f, \phi_{u_i} \rangle_{\mathcal{H}}\|^2/2 + \lambda_t \|f\|_{\mathcal{H}}^2$. In each iteration t, we suppose that f_i^{t-1} is the estimation of f_i from the previous iteration, the `OPGD` algorithm samples $u_i^t \sim \rho_{\mathcal{X}}, y_i^t \sim \mathcal{D}_i(u_i^t)$ and takes gradient steps on the kernel ridge objective $\|y_i^t - \langle f, \phi_{u_i^t} \rangle_{\mathcal{H}}\|^2/2 + \lambda_t \|f\|_{\mathcal{H}}^2$, i.e. it takes the online kernel ridge update (Tarres and Yao, 2014; Dieuleveut and Bach, 2016):

$$\text{(i)} \quad f_i^t = f_i^{t-1} - \nu_t \left[\left(f_i^{t-1}(u_i^t) - y_i^t\right) \phi_{u_i^t} + \lambda_t f_i^{t-1}\right]. \tag{3.5}$$

Since the kernel ridge regression $\arg\min_{f \in \mathscr{F}^p} \mathbb{E}_{(u_i, y_i) \sim \rho_i} \|y_i - \langle f, \phi_{u_i} \rangle_{\mathcal{H}}\|^2/2 + \lambda_t \|f\|_{\mathcal{H}}^2$ has a biased solution f_{i,λ_t}, we let λ_t shrink to 0 gradually to ensure $f_{i,\lambda_t} \to f_i$. We remark that the change of λ_t will bring drift error $f_{i,\lambda_t} - f_{i,\lambda_{t-1}}$, which is closely nested with the estimation error $f_i^t - f_i$. We choose ν_t and λ_t carefully to let $f_{i,\lambda_t} - f_i$ and $f_i^t - f_{i,\lambda_t}$ converge simultaneously (Theorem 4.18).

We suppose that the kernel K is 2-differentiable, i.e., $K \in C^2(\mathcal{X}, \mathcal{X})$. Define $\partial_i \phi : \mathcal{X} \to \mathcal{H}$ as the partial derivative of the feature map ϕ with respect to x_i, namely, $\partial_i \phi_x = \partial_i K(x, \cdot) = \partial K(x, \cdot)/\partial x_i$. Steinwart and Christmann (2008, Lemma 4.34) shows that $\partial_i \phi_x$ exists, continuous and $\partial_i \phi_x \in \mathcal{H}$. By the reproducing property $\partial f_i(x)/\partial x_i = \partial \langle f_i, \phi_x \rangle_{\mathcal{H}}/\partial x_i = \langle f_i, \partial_i \phi_x \rangle_{\mathcal{H}}$, the individual gradient $\nabla_i \mathcal{L}_i(x)$ has the form $\nabla_i \mathcal{L}_i(x) = \mathbb{E}_{z_i \sim \mathcal{D}_i(x)}[\nabla_i \ell_i(x, z_i) + (\langle f_i, \partial_i \phi_x \rangle_{\mathcal{H}})^\top \nabla_{z_i} \ell_i(x, z_i)]$. After step (i), `OPGD` draws a sample $z_i^t \sim \mathcal{D}_i(x^t)$ and performs the projected gradient step:

$$\text{(ii)} \quad x_i^{t+1} \leftarrow \text{proj}_{\mathcal{X}_i} \left(x_i^t - \eta_t \left(\nabla_i \ell_i(x^t, z_i^t) + (\langle f_i^t, \partial_i \phi_{x^t} \rangle_{\mathcal{H}})^\top \nabla_{z_i} \ell_i(x^t, z_i^t)\right)\right). \tag{3.6}$$

We remark that the gradient steps η_t, ν_t and regularization terms λ_t should be chosen carefully to ensure convergence (see Theorem 4.19). Specifically, the regularization terms λ_t must shift to 0 gradually. If λ_t is a constant, f_i^t in (3.5) converges to the solution of a regularized kernel ridge regression, which is a biased estimator of f_i. Thus (3.6) fails to converge because the gradient estimation has a constant bias. We present the pseudocode of `OPGD` for the linear setting as Algorithm 1 and for the RKHS setting as Algorithm 2 in Appendix A.

Comparison with Narang et al. (2022). We clarify the difference between `OPGD` and the Adaptive Gradient Method (AGM) proposed in Narang et al. (2022). To elaborate, AGM samples $z_i^t \sim \mathcal{D}_i(x^t)$ at current the action and let agents play again with an injected noise u^t to obtain $q_i^t \sim \mathcal{D}_i(x^t+u^t)$. The algorithm is based on the fact that $\mathbb{E}[q_i^t - z_i^t | u^t, x^t] = A_i u^t$, which is not related to x^t. Thus, A_i can be estimated by online least squares. We remark that $\mathbb{E}[q_i^t - z_i^t | u^t, x^t]$ depends on agents' actions in the non-linear (RKHS) cases, because $\mathbb{E}[q_i^t - z_i^t | u^t, x^t] = f_i(x^t + u^t) - f_i(x^t) = \langle f_i, \phi_{x^t+u^t} - \phi_{x^t} \rangle_{\mathcal{H}}$. Thus, the change of action will bring additional error that makes the estimation fail to converge. In contrast, `OPGD` lets agents play $u_i^t \sim \rho_{\mathcal{X}}$ to explore the action space and learn the strategic behavior of other agents. `OPGD` estimates the parametric function by solving the ERM version of (2.9) using online stochastic approximation (3.3) and (3.5). This learning framework can be applied to RKHS and potentially beyond that, such as overparameterized neural networks using the technique of neural tangent kernel (Allen-Zhu et al., 2019).

3.3 Learning in the Bandit Feedback Setting

In this section, we extend the kernel `OPGD` into the bandit feedback setting, which is common in the real-world application. Given a joint action x and data z_i, we only observe the loss $\ell_i(x, z_i)$ without access to the first-order oracle (i.e. the gradient of ℓ_i is unknown). Therefore, to compute the performative gradient and conduct the projected gradient steps (3.4) and (3.6), we need to estimate the unknown gradients $\nabla_i \ell_i$ and $\nabla_{z_i} \ell_i$ from the observed loss $\ell_i(x, z_i)$.

To estimate the gradients, we leverage the similar idea of estimating the decision-dependent distribution $\mathcal{D}_i$. Suppose loss functions ℓ_i are in an RKHS $\mathcal{B}$ associated with the feature map φ, we estimate ℓ_i by online regression. The reproducibility of the RKHS implies that $\nabla \ell_i = \nabla \langle \ell_i, \varphi \rangle_{\mathcal{B}} = \langle \ell_i, \partial\varphi \rangle_{\mathcal{B}}$. Thus, in order to obtain an estimation of the gradient, it is enough to estimate the loss function ℓ_i.

Assumption 3.1. Suppose there exists a function class $\mathcal{B}$ such that $\ell_i \in \mathcal{B}$ for all $i \in [n]$. Specifically, we assume $\mathcal{Z}_i = \mathcal{Z}_j$ for all $i, j \in [n]$, and $\mathcal{B}$ is an RKHS on $\mathcal{Y} := \mathcal{X} \times \mathcal{Z}_1$ induced by a Mercer kernel $R : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ associated with a measure $\rho_{\mathcal{Y}}$ on $\mathcal{Y}$ with full support and a feature map $\varphi : \mathcal{Y} \to \mathcal{B}$.

Next, to estimate the loss function ℓ_i, we consider the following kernel ridge regression. $\mathcal{Y}$.

$$\widehat{\ell}_i = \operatorname*{arg\,min}_{\ell \in \mathcal{B}} \int_{(x,z)\sim\rho_{\mathcal{Y}}} (\ell_i(x, z) - \langle \varphi(x, z), \ell \rangle_{\mathcal{B}})^2 + \iota \|\ell\|_{\mathcal{B}}^2.$$

We remark that the full support of $\rho_{\mathcal{Y}}$ in Assumption 3.1 is crucial, it ensures that the sampling distribution $\rho_{\mathcal{Y}}$ is non-degenerate and the sampling strategy can sufficiently explore $\mathcal{Y}$. Intuitively, the full support assumption implies the noise term ϵ_i has the full support and we leverage the noise term to explore. Given this intuition, we solve the ERM version of this ridge regression and consider the following online gradient steps. To elaborate, in each iteration t, `OPGD` samples $u_i^t \sim \rho_{\mathcal{X}}, y_i^t \sim \mathcal{D}_i(u_i^t)$ and obtains w_i^t as the corresponding loss (i.e. $w_i^t = \ell_i(u_i^t, y_i^t)$). Firstly, `OPGD` performs the gradient step (3.5) to learn the parametric function f_i. Next, suppose that ℓ_i^{t-1} is the estimation of ℓ_i from the previous iteration, `OPGD` performs the following estimation update:

$$\ell_i^t = \ell_i^{t-1} - s_t \left[\left(\ell_i^{t-1}(u_i^t, y_i^t) - w_i^t \right) \varphi_{u_i^t, y_i^t} + \iota_t \ell_i^{t-1} \right]. \tag{3.7}$$

Here ι_t is the dynamic regularization term analogous to λ_t in (3.5). Given the estimated parametric model and the estimated loss functions, `OPGD` performs the following projected gradient step:

$$x_i^{t+1} \leftarrow \mathrm{proj}_{\mathcal{X}_i} \left(x_i^t - \eta_t \left(\langle \nabla_i \varphi(x^t, z_i^t), \ell_i^t \rangle_{\mathcal{B}} + (\langle f_i^t, \partial_i \phi_{x^t} \rangle_{\mathcal{H}})^\top \langle \nabla_{z_i} \varphi(x^t, z_i^t), \ell_i^t \rangle_{\mathcal{B}} \right) \right). \tag{3.8}$$

In summary, `OPGD` performs three gradient steps in each iteration: first updates the estimation of the parametric function f_i, then updates the estimation of the loss function ℓ_i, finally performs the projected gradient steps leveraging the estimated loss functions and parametric functions.

4 Theoretical Results

We provide theoretical guarantees for `OPGD` in both linear and RKHS settings. We first impose some mild assumptions. Similar assumptions are adopted in Mendler-Dünner et al. (2020); Izzo et al. (2021); Narang et al. (2022); Cutler et al. (2022).

Assumption 4.1. (τ-strongly monotone). The game (2.1) is τ-strongly monotone.

Assumption 4.2. (Smoothness). $H(x)$ is L-Lipschitz continuous:

$$H(x_1) - H(x_2) \le L\|x_1 - x_2\|, \qquad \forall x_1, x_2 \in \mathcal{X}.$$

Assumption 4.3. (Lipschitz continuity in z). Define $\mathcal{D} = \mathcal{D}_1 \times \mathcal{D}_2 \times \cdots \times \mathcal{D}_n : \mathcal{X} \to \mathbb{P}(\mathcal{Z})$, where $\mathcal{Z}$ is the sample space $\mathcal{Z} = \mathcal{Z}_1 \times \mathcal{Z}_2 \times \cdots \times \mathcal{Z}_n$. For all $i \in [n], x \in \mathcal{X}$, there exists a constant $\delta > 0$,

$$\mathop{\mathbb{E}}_{z \sim \mathcal{D}(x)} \sqrt{\sum_{i=1}^n \|\nabla_{z_i} \ell_i(x, z_i)\|^2} \le \delta.$$

Assumption 4.4. (Finite variance). There exists a constant $\zeta > 0$,

$$\mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(x)} \|\nabla_{i,z_i} \ell_i(x, z_i) - \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(x)} \nabla_{i,z_i} \ell_i(x, z_i)\|^2 \le \zeta^2, \qquad \forall i \in [n], \forall x \in \mathcal{X},$$

where $\nabla_{i,z_i} \ell_i$ denotes the gradient of $\ell_i(x, z_i)$ with respect to x_i and z_i.

We remark that Assumption 4.2 is the standard smoothness assumption for the utility functions $\mathcal{L}_i(x)$ (Boyd et al., 2004; Nesterov et al., 2018). Since $\mathcal{X}$ is a compact set within $\mathbb{R}^d$, Assumption 4.3 holds if $\ell_i(x, z_i)$ is Lipschitz continuous in z_i and the gradient $\nabla_{z_i} \ell_i(x, z_i)$ is continuous in x, and Assumption 4.4 holds if $\ell_i(x, z_i)$ is Lipschitz in x and z_i (thus $\nabla_{i,z_i} \ell_i(x, z_i)$ has a bounded norm). Assumption 4.4 implies that the variances of $\nabla_i \ell_i(x, z_i)$ and $\nabla_{z_i} \ell_i(x, z_i)$ are both bounded by ζ^2 for any $x \in \mathcal{X}$ and $z_i \sim \mathcal{D}_i(x)$. Besides, we present sufficient conditions for the game (2.1) to be strongly monotone.

Proposition 4.5. (Sufficient conditions for Assumption 4.1). Suppose action sets $\mathcal{X}_i$ are compact and convex and loss functions ℓ_i are C^1-smooth in x, z_i. Suppose Assumption 2.6 holds and parametric functions f_i are differentiable in x. If there exist positive constants S, L_i, R_i such that $S > 2\sqrt{\sum_{i=1}^n (L_i R_i)^2}$ and the following properties hold for all $i \in [n]$:

(i) $f_i(x)$ is L_i-Lipschitz continuous in $x \in \mathcal{X}$.

(ii) The map $z_i \to \nabla_i \ell_i(x, z_i)$ is R_i-Lipschitz continuous and the map $u \to \mathbb{E}_{z_i \sim \mathcal{D}_i(u, x_{-i})} \ell_i(x, z_i)$ is monotone in $u \in \mathcal{X}_i$ for any fixed $x \in \mathcal{X}$.

(iii) The static game (4.1) is S-strongly monotone for any $y \in \mathcal{X}$:

$$\min_{x_i \in \mathcal{X}_i} \mathcal{L}_i^y(x_i, x_{-i}), \quad \text{where} \quad \mathcal{L}_i^y(x) := \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(y)} \ell_i(x, z_i). \tag{4.1}$$

Then Assumption 4.1 holds for $\tau = S - 2\sqrt{\sum_{i=1}^n (L_i R_i)^2}$.

We refer the reader to Appendix C.4 for complete proof. Next, we propose the convergence guarantee given the estimation error of the gradient of the loss function ℓ_i and the parametric function f_i.

Theorem 4.6. (General convergence guarantee). Suppose that Assumptions 4.1, 4.2, 4.3, and 4.4 hold. Suppose that there is an algorithm outputs ∂f_i^t and $\nabla \ell_i$ in iteration t, and performs the following projected gradient step to find the Nash equilibrium

$$x_i^{t+1} \leftarrow \mathrm{proj}_{\mathcal{X}_i}\left(x_i^t - \eta_t \left(\nabla_i \ell_i^t(x^t, z_i^t) + (\partial f_i^t(x^t)/\partial x_i)^\top \nabla_{z_i} \ell_i^t(x^t, z_i^t)\right)\right), \tag{4.2}$$

where x^t is the output of iteration $t-1$ and $z_i^t \sim \mathcal{D}_i(x^t)$. Suppose that there exists some positive constants a_1 and a_2, such that the estimation error of gradients holds for all $x \in \mathcal{X}, z_i \in \mathcal{Z}_i$ ($i \in [n]$) and each iteration t:

$$\|\partial f_i^t(x)/\partial x_i - \partial f_i(x)/\partial x_i\|_F \lesssim \mathcal{O}(t^{-a_1}) \quad \text{and} \quad \|\nabla \ell_i(x, z_i) - \nabla \ell_i^t(x, z_i)\| \lesssim \mathcal{O}(t^{-a_2}). \tag{4.3}$$

For all $t \geq 1$, set $\eta_t = (1 + (1 \wedge 2a_1 \wedge 2a_2))/(\tau(t + t_0))$ where t_0 is a positive integer, then the x^t generated by this algorithm satisfies

$$\mathbb{E}\|x^t - x^*\|^2 \lesssim \mathcal{O}(t^{-(1 \wedge 2a_1 \wedge 2a_2)}).$$

We refer the reader to Appendix C.1 for complete proof. Intuitively, Theorem 4.6 presents a general convergence framework for `OPGD`, where it approximates the loss function ℓ_i and the parametric function f_i by the linear or kernel function class with polynomial rate.

4.1 Convergence Rate in the Linear Setting

In this section, we derive the convergence rate of the linear `OPGD` with first-order oracle. We introduce two assumptions necessary to derive theoretical guarantees for the linear function class.

Assumption 4.7. (Linear assumption). Suppose that the parametric assumption holds (Assumption 2.6) and $f_i(x) = A_i x$ for $i \in [n]$, where $A_i \in \mathbb{R}^{p\times d}$ are unknown matrices.

Assumption 4.8. (Sufficiently isotropic). There exists constants $l_1, l_2, R > 0$ such that

$$l_1 I \preceq \mathbb{E}_{u\sim\rho_{\mathcal{X}}} uu^\top, \quad \mathbb{E}_{u\sim\rho_{\mathcal{X}}} \|u\|^2 \le l_2, \quad \mathbb{E}_{u\sim\rho_{\mathcal{X}}} \left[\|u\|^2 uu^\top\right] \preceq R\mathbb{E}_{u\sim\rho_{\mathcal{X}}} uu^\top.$$

Assumption 4.8 has been studied in the literature on online least squares regression (Dieuleveut et al., 2017; Narang et al., 2022). Essentially, this requires the distribution $\rho_{\mathcal{X}}$ to be sufficiently isotropic and non-singular, and it ensures the random variable $u_i^t \sim \rho_{\mathcal{X}}$ in the online estimation update step (3.3) can explore all the "directions" of $\mathbb{R}^p$. A simple example that satisfies Assumption 4.8 is the uniform distribution $\rho_{\mathcal{X}} = \mathcal{U}[0, 1]$, in which case $l_1 = l_2 = 1/3, R = 3/5$.

The next theorem provides the convergence rate of `OPGD` under the linear setting.

Theorem 4.9. (Convergence in the linear setting). Suppose that Assumptions 4.1, 4.2, 4.3, 4.4, 4.7, and 4.8 hold. Set $\eta_t = 2/(\tau(t + t_0))$, $\nu_t = 2/(l_1(t + t_0))$, where t_0 is a constant that satisfies $t_0 \ge 2l_2R/l_1^2$. For all iterations $t \ge 1$, the x^t generated by the `OPGD` algorithm in Section 3 for linear function class satisfies

$$\mathbb{E}\|x^t - x^*\|^2 \le \frac{(4D_1 + 2D_2(t_0 + 1)\tau)(t_0 + 2)^2/(t_0 + 1)^2}{\tau^2(t + t_0)} + \frac{(t_0 + 1)^2\|x^1 - x^*\|^2}{(t + t_0)^2}, \tag{4.4}$$

where D_1 and D_2 are constants that

$$D_1 := 4\zeta^2(1+2(M/(t_0+1)+\sup_{i\in[n]}\|A_i\|_F^2)), \quad D_2 := 2\delta^2 M, \quad M := \frac{2t_0^4 \sum_{i=1}^n \|A_i^0 - A_i\|_F^2}{(t_0 + 1)^3} + \frac{8nl_2\sigma^2(t_0 + 2)^2}{l_1^2(t_0 + 1)^2}.$$

Sketch of the Proof. The proof has three steps. First, we derive estimation error bounds of (3.3):

Lemma 4.10. (Estimation error). Suppose Assumptions 4.7, 4.8 hold and set $\nu_t = 2/(l_1(t + t_0))$, where t_0 is a constant satisfies $t_0 \ge 2l_2R/l_1^2$. Then the matrix A_i^t generated by `OPGD` satisfies

$$\mathbb{E}\|A_i^t - A_i\|_F^2 \le \frac{\frac{2t_0^4}{(t_0+1)^3}\|A_i^0 - A_i\|_F^2 + \frac{8l_2\sigma^2}{l_1^2}\left(\frac{t_0+2}{t_0+1}\right)^2}{t + t_0}. \tag{4.5}$$

Second, we prove that projected gradient steps (3.4) satisfy the stochastic framework (Assumption B.1 in Appendix B.1). In more detail, we derive bias and variance bounds for gradient estimators $\widehat{H}(x)$:

$$\begin{aligned}
&\text{(Bias)} \quad \|((A_{ii}^t - A_{ii})^\top \mathbb{E}_t \nabla_{z_i}\ell_i(x^t, z_i^t))_{i\in[n]}\|,\\
&\text{(Variance)} \quad \mathbb{E}_t\|([\nabla_i\ell_i(x^t, z_i^t) - \mathbb{E}_t\nabla_i\ell_i(x^t, z_i^t)] + (A_{ii}^t)^\top[\nabla_{z_i}\ell_i(x^t, z_i^t) - \mathbb{E}_t\nabla_{z_i}\ell_i(x^t, z_i^t)])_{i\in[n]}\|^2.
\end{aligned}$$

Using Lemma B.2 in Appendix B.1, we derive the following one-step error bound:

Lemma 4.11. (One-step error). Suppose Assumptions 4.1, 4.2, 4.3, 4.4, and 4.7 hold. Let $\mathcal{G} = \{\mathcal{G}_t\}_{t\in\mathbb{N}}$ be the filtration $\mathcal{G}_t = \sigma\{\{x^j\}_{j\in[T]} \cup (u_i^t, y_i^t)\}$ and define $\mathbb{E}_t[\cdot] = E[\cdot|\mathcal{G}_t]$. For any gradient steps $\eta_t \le \tau/(4L^2)$, the x^t generated by `OPGD` satisfies

$$\mathbb{E}_t\|x^{t+1} - x^*\|^2 \le \frac{1}{1+\eta_t\tau}\|x^t - x^*\|^2 + \frac{4\eta_t^2\zeta^2(1+\sup_{i\in[n]}\|A_i^t\|_F^2)}{1+\eta_t\tau} + \frac{2\eta_t\delta^2\sup_{i\in[n]}\|A_i^t - A_i\|_F^2}{\tau(1+\eta_t\tau)}. \tag{4.6}$$

Finally, putting the estimation error and the one-step error bounds together, we have that

$$\mathbb{E}\|x^{t+1} - x^*\|^2 \le \frac{1}{1+\eta_t\tau}\mathbb{E}\|x^t - x^*\|^2 + \frac{D_1\eta_t^2}{1+\eta_t\tau} + \frac{D_2\eta_t/t}{1+\eta_t\tau}, \tag{4.7}$$

which further leads to (4.4). See Appendix Sections C.2 and D for the detailed proofs. □

We illustrate the parameters involved in Theorem 4.9: τ is the strongly monotone parameter of the game (2.1), l_1, l_2, R are intrinsic parameters describing the isotropy of the distribution $\rho_{\mathcal{X}}$ (Assumption 4.8), σ^2 is the variance of the noise term ϵ_i defined in Assumption 2.6, ζ and δ describe the continuity of ℓ_i (Assumption 4.3, 4.4), t_0 is a sufficiently large value, A_i^0 is the initial estimation of A_i, x^1 is the initial input. Theorem 4.9 is a combination of Lemma 4.10 and Lemma 4.11, where Lemma 4.10 is the statistical error of the online approximation step (3.3) and Lemma 4.11 is the one-step optimization error of the projected gradient step (3.4). Theorem 4.9 implies the convergence rate of `OPGD` in the linear setting is $\mathcal{O}(t^{-1})$, which matches the optimal rate of stochastic gradient descent in the strongly-convex setting.

4.2 Convergence Rate in the RKHS Setting

In this section, we derive the convergence rate of the kernel `OPGD` with first-order oracle. Suppose that $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is a continuous Mercer kernel and $\rho_{\mathcal{X}}$ has full support. Define the integral operator $L_K : \mathcal{L}^2_{\rho_{\mathcal{X}}} \to \mathcal{H}$ by the integral transformation:

$$L_K(f)(x) := \int_{\mathcal{X}} K(x,t)f(t)d\rho_{\mathcal{X}}(t), \qquad \forall f \in \mathcal{H}, \forall x \in \mathcal{X}.$$

By Mercer's theorem, K has the spectral representation $K = \sum_{i=1}^\infty \mu_i e_i \otimes e_i$, where $\otimes$ denotes the tensor product, $\{\mu_i\}_{i=1}^\infty$ are eigenvalues and $\{e_i\}_{i=1}^\infty$ are eigenfunctions with respect to the operator L_K. Moreover, $\{e_i\}_{i=1}^\infty$ is an orthogonal basis of $\mathcal{L}^2_{\rho_{\mathcal{X}}}$ and $\{\mu_i^{1/2}e_i\}_{i=1}^\infty$ is the orthogonal basis of $\mathcal{H}$, which induces the representation $\mathcal{H} = \{\sum_{i=1}^\infty a_i\mu_i^{1/2}e_i : \{a_i\}_{i=1}^\infty \in \ell^2\}$.

Definition 4.12. (Power space). For a constant $\alpha \ge 0$, the α-power space of an RKHS $\mathcal{H}$ is defined by

$$\mathcal{H}^\alpha = \left\{\sum_{i=1}^\infty a_i\mu_i^{\alpha/2}e_i : \{a_i\}_{i=1}^\infty \in \ell^2\right\},$$

equipped with the α-power norm $\|\cdot\|_\alpha$ and inner product $\langle\cdot,\cdot\rangle_\alpha$, where $\|\sum_{i=1}^\infty a_i\mu_i^{\alpha/2}e_i\|_\alpha := \left(\sum_{i=1}^\infty a_i^2\right)^{1/2}$ and $\langle\sum_{i=1}^\infty a_i\mu_i^{\alpha/2}e_i, \sum_{i=1}^\infty b_i\mu_i^{\alpha/2}e_i\rangle_\alpha = \sum_{i=1}^\infty a_ib_i$.

We remark that: (i) $\mathcal{H}^1 = \mathcal{H}$ and $\mathcal{H}^\alpha \subset \mathcal{H}^\beta$ for any $\alpha > \beta$, (ii) $\|\cdot\|_1 = \|\cdot\|_\mathcal{H}$ and $\|\cdot\|_0 = \|\cdot\|_{\rho_\mathcal{X}}$, and (iii) $\mathcal{H}^\alpha$ is an RKHS on $\mathcal{X}$ with kernel $K^\alpha \coloneqq \sum_{i=1}^\infty \mu_i^\alpha e_i \otimes e_i$ and measure $\rho_\mathcal{X}$. We review more properties of RKHS and power spaces in Appendix Sections B.2 and B.3.

We present assumptions on the kernel function class, similar assumptions can be found in the literature on kernel regression and stochastic approximation (Caponnetto and De Vito, 2007; Steinwart et al., 2009; Dicker et al., 2017; Pillaud-Vivien et al., 2018; Fischer and Steinwart, 2020).

Assumption 4.13. (Source condition). Suppose Assumption 2.6 holds and there exists an RKHS, $\mathcal{H}$, with a bounded differentiable Mercer kernel, K, and constants $\beta, \kappa > 0$ such that $\sup_{x\in\mathcal{X}} K(x,x) \leq \kappa^2$ and $f_i \in \left(\mathcal{H}^\beta\right)^p$ for all $i \in [n]$.

Assumption 4.14. (Embedding property). There exist constants $\alpha \in (0,1], A > 0$ such that $K^\alpha(x,x) = \sum_{i=1}^\infty \mu_i^\alpha e_i^2(x) \leq A^2$, for all $x \in \mathcal{X}$.

Assumption 4.15. (Lipschitz kernel). Suppose Assumption 4.14 holds and there exists $\xi > 0$ such that $\|\partial_i \phi_x^\alpha\|_\alpha \leq \xi$ for any $i \in [n]$ and $x \in \mathcal{X}$, where $\phi_x^\alpha : \mathcal{X} \to \mathcal{H}^\alpha$ is the feature map of the kernel K^α.

Assumption 4.13 holds when K is bounded, differentiable, and each coordinate of parametric functions f_i lies in the power space $\mathcal{H}^\beta$. When $\beta < 1$, Assumption 4.13 includes the challenging scenario, namely, $f_i \notin (\mathcal{H})^p$. Assumption 4.14 holds if there exists a power space $\mathcal{H}^\alpha$ such that the kernel K^α is bounded. Thus, Assumption 4.14 holds with $\alpha = 1$ for any bounded kernel K. We further propose Proposition 4.16 as sufficient conditions for the embedding property following Mendelson and Neeman (2010). Recalling the definition of partial derivative $\partial_i \phi^\alpha : \mathcal{X} \to \mathcal{H}^\alpha$ (Section 3), Assumption 4.15 holds if $\partial_i \partial_{i+d} K^\alpha(x,x) = \|\partial_i \phi_x^\alpha\|_\alpha^2 \leq \xi^2$ for any $x \in \mathcal{X}$, i.e. it holds for any Lipschitz kernel K^α.

Proposition 4.16. (Sufficient conditions for Assumption 4.14) Suppose there exist constants $C, D, p > 0$ and $q \in (0,1)$ such that

$$\sup_{i\in\mathbb{N}} \mu_i^p \|e_i\|_\infty \leq C \quad \text{and} \quad \mu_i \leq D i^{-1/q},$$

where $\|\cdot\|_\infty$ denotes the L^∞ norm. Then Assumption 4.14 holds for any $\alpha > 2p + q$.

Proposition 4.16 follows from the inequality: $\sup_{x\in\mathcal{X}} K^\alpha(x,x) = \sup_{x\in\mathcal{X}} \sum_{i=1}^\infty (\mu_i^p e_i(x))^2 \mu_i^{\alpha-2p} \leq C^2 D^{\alpha-2p} \sum_{i=1}^\infty i^{-(\alpha-2p)/q} < \infty$. Now we present an example that satisfies these assumptions.

Example 4.17. (Splines on the Circle). Let $\mathcal{X} = [0,1]$ associated with the measure $\mathcal{U}[0,1]$. For any $m \in \mathbb{N}$, let $\mathcal{H}$ be the collection of all zero-mean periodic functions f on $[0,1]$ of the form $f : t \to \sqrt{2}\sum_{i=1}^\infty a_i(f)\cos(2\pi i t) + \sqrt{2}\sum_{i=1}^\infty b_i(f)\sin(2\pi i t)$, associated with the norm $\|f\|_\mathcal{H}^2 = \sum_{i=1}^\infty (a_i(f)^2 + b_i(f)^2)(2\pi i)^{2m}$ and the inner product $\langle f, g\rangle_\mathcal{H} = \sum_{i=1}^\infty (2\pi i)^{2m}(a_i(f)a_i(g) + b_i(f)b_i(g))$. Following Wahba (1990), $\mathcal{H}$ is an RKHS with the kernel $R_m(x,y)$:

$$R_m(x,y) = \sum_{i=1}^\infty \frac{2}{(2\pi i)^{2m}} \cos(2\pi i(x-y)) = \frac{(-1)^{m-1}}{(2m)!} B_{2m}(\{x-y\}),$$

where $B_{2m}(\cdot)$ denotes the $2m$-th Bernoulli polynomial and $\{x-y\}$ is the fractional part of $x-y$. It is easy to check that $(\sqrt{2}\cos(2\pi ix), \sqrt{2}\sin(2\pi ix))_{i\in\mathbb{N}}$ are eigenfunctions with eigenvalues $\mu_i = (2\pi i)^{-2m}$ (Dieuleveut and Bach, 2016). Thus, Proposition 4.16 holds for any $p > 0$ and $q \geq 1/(2m)$ because the eigenfunctions are uniformly bounded and eigenvalues $\mu_i \sim i^{-2m}$, which implies that Assumption 4.14 holds for any $\alpha > 1/(2m)$. Moreover, for any $k \in \mathbb{N}$, let each coordinate of the parametric function $f_i(x)$ be $B_k(x)$, then Assumption 4.13 holds for $\beta = (2k-1)/(2m)$ because $B_k(x) = -2k!\sum_{i=1}^{\infty}\frac{\cos(2\pi ix - k\pi/2)}{(2\pi i)^k}$ (Abramowitz et al., 1988). For any $m > 1$, the kernel $R_{2m}(x,y)$ is bounded and differentiable, thus Assumption 4.15 holds.

We then provide the convergence of the proposed algorithm under the RKHS setting. Specifically, we present the guarantees for the online estimation error (Lemma 4.18) as well as the rate of convergence to the Nash equilibrium (Theorem 4.19).

Lemma 4.18. (Estimation error of f_i^t). Suppose Assumption 4.13 holds for some $\beta \in (0,2]$, Assumptions 4.14, 4.15 hold for some $\alpha \in (0,1]$ and $\alpha < \beta$. For all iterations t and positive constant a, define $\bar{t} = t + t_0$, where t_0 is a constant satisfies $t_0 \geq (a\kappa^2+1)^2$. For a constant $\gamma \in [\alpha, \beta)$ and $\gamma \leq 1$, set the gradient steps and regularization terms as

$$\nu_t = a\left(\frac{1}{\bar{t}}\right)^{\frac{\beta-\gamma+1}{\beta-\gamma+2}}, \qquad \lambda_t = \frac{1}{a}\left(\frac{1}{\bar{t}}\right)^{\frac{1}{\beta-\gamma+2}}.$$

If $a < \sqrt{(\beta-\gamma+2)/(\beta-\gamma)}(t_0+1)/(t_0+2)\kappa^{\gamma-2}A^{-1}$, the f_i^t generated by `OPGD` (Algorithm 2) with input kernel K satisfies

$$\mathbb{E}\|f_i^t - f_i\|_\gamma^2 \lesssim \mathcal{O}(t^{-\frac{\beta-\gamma}{\beta-\gamma+2}}). \tag{4.8}$$

We refer the reader to Appendix E.1 for the complete proof. Lemma 4.18 presents the error bound of online stochastic approximation (3.5) under the power norm $\|\cdot\|_\gamma$. Our result includes the classical theory under the RKHS norm $\|\cdot\|_{\mathcal{H}}$ and extends it on a continuous scale. In more detail, for any $\gamma \in [\alpha,\beta)$ and $\gamma \leq 1$, Lemma 4.18 describes how to choose the step-sizes ν_t and regularization term λ_t properly to insure convergence under $\|\cdot\|_\gamma$ with the convergence rate $\mathcal{O}(t^{-(\beta-\gamma)/(\beta-\gamma+2)})$. For $\beta > 1$ and $\gamma = 1$, this rate would be $\mathcal{O}(t^{-(\beta-1)/(\beta+1)})$ and matches the optimal rate under the RKHS norm Ying and Pontil (2008); Tarres and Yao (2014). If the embedding property (Assumption 4.14) holds for some $\alpha < 1$, we choose $\gamma = \alpha$ to achieve a faster rate $\mathcal{O}(t^{-(\beta-\alpha)/(\beta-\alpha+2)})$ (which further leads to Theorem 4.19). Besides, while classical theory assumes Assumption 4.13 holds for $\beta > 1$ (i.e. $f_i \in (\mathcal{H}^\beta)^p \subset (\mathcal{H})^p$), our result relaxes this assumption to $\beta > \alpha$ and allows $\beta \leq 1$. Intuitively, this implies that the online stochastic approximation can address the misspecification case $f_i \notin (\mathcal{H})^p$ if $\alpha < 1$.

Here we briefly talk about the technical challenges to obtain the power norm bound. Intuitively, the main challenge to derive power norm bounds for iteration (3.5) (i.e. f_t) under the norm $\|\cdot\|_\gamma$ arises from the differing properties between the power space $\mathcal{H}^\gamma$ and the RKHS $\mathcal{H}$. To elaborate, the standard method to derive error bounds under $\|\cdot\|_{\mathcal{H}}$ decomposes the error $f_i^t - f_i$ by the operator $I - \nu_t(L_t + \lambda_t I)$ where $L_t := \phi_t^*\phi_t$ (refer to (E.2)), and the analysis is based on the fact that $I - \nu_t(L_t + \lambda_t I)$ is a contraction map on $\mathcal{H}$. This is because the sampling operator

L_t is compact, self-adjoint, and positive-semidefinite on $\mathcal{H}$, thus, the spectral theorem implies $\|I - \nu_t(L_t + \lambda_t I)\|_{\mathcal{H}\to\mathcal{H}} \le 1 - \nu_k\lambda_k$ where $\|\cdot\|_{\mathcal{H}\to\mathcal{H}}$ denotes the spectral norm. However, this operator does not exhibit the same behavior on the power space $\mathcal{H}^\gamma$. By the definition that $L_t = \phi_t^*\phi_t$, for any $h_1, h_2 \in \mathcal{H}^\gamma$, we have $L_t(h_i) = h_i(x^t)\phi_t = \langle h_i, \phi_t^\gamma\rangle_\gamma \phi_t$ and $\langle L_t h_1, h_2\rangle_\gamma \ne \langle h_1, L_t h_2\rangle_\gamma$ (here we lift the domain of L_t from $\mathcal{H}$ to $\mathcal{H}^\gamma$). Thus, L_t is not self-adjoint or positive-definite on $\mathcal{H}^\gamma$ and the spectral norm $\|I - \nu_t(L_t + \lambda_t I)\|_{\mathcal{H}^\gamma\to\mathcal{H}^\gamma}$ might larger than 1.

To overcome the aforementioned difficulty, we propose a series of novel proof steps. The main technical innovation is that our analysis decouples power norm bounds by RKHS norm by considering semi-population iteration and recursive decomposition, such methods can be applied to derive power norm bounds for other online algorithms (refer to "Technical contributions" paragraph in Appendix E.1 to a detail explanation for our innovation).

Theorem 4.19. (Convergence in the RKHS setting). Suppose that Assumptions 4.1, 4.2, 4.3, 4.4 hold, Assumption 4.13 holds for some $\beta \in (0, 2]$, and Assumptions 4.14, 4.15 hold for some $\alpha \in (0, 1]$ and $\alpha < \beta$. For all iterations $t \ge 1$ and positive constant a, define $\bar{t} = t + t_0$, where t_0 is a constant that satisfies $t_0 \ge (a\kappa^2 + 1)^2$. Set the gradient steps and regularization terms as

$$\eta_t = (\tau\bar{t})^{-1}, \qquad \nu_t = a \cdot \bar{t}^{-\frac{\beta-\alpha+1}{\beta-\alpha+2}}, \qquad \lambda_t = a^{-1} \cdot \bar{t}^{-\frac{1}{\beta-\alpha+2}}.$$

If $a < \sqrt{(\beta-\alpha+2)/(\beta-\alpha)}(t_0+1)/(t_0+2)\kappa^{\alpha-2}A^{-1}$, the x^t generated by the `OPGD` algorithm in Section 3 using kernel K for online estimation steps (3.5) and projected gradient steps (3.6) satisfies

$$\mathbb{E}\|x^t - x^*\|^2 \lesssim \mathcal{O}(t^{-\frac{\beta-\alpha}{\beta-\alpha+2}}). \tag{4.9}$$

The proof strategy for Theorem 4.19 is similar to that of Theorem 4.9, namely, we derive the estimation error (setting $\gamma = \alpha$ in Lemma 4.18) as well as the one-step error (Lemma C.1) and combine them to obtain the result. We refer readers to Appendix C.3 for the complete proof of these results.

We demonstrate the parameters involved in Theorem 4.19. Parameters $\alpha, \beta, \kappa, \tau, A$ are intrinsic: β, κ, A are determined by source condition (Assumption 4.13), α is determined by embedding property (Assumption 4.14), and τ is the strongly monotone parameter. Parameters a, t_0 are user-specified: t_0 is a sufficiently large value, a is characterized by the inequality $a < \sqrt{(\beta-\alpha+2)/(\beta-\alpha)}(t_0+1)/(t_0+2)\kappa^{\alpha-2}A^{-1}$ when t_0 is determined, a smaller a leads to a larger constant term in (4.9).

Theorem 4.19 implies that `OPGD` leverages the embedding property (Assumption 4.14) to obtain better convergence rates. For any bounded kernel, Assumption 4.14 holds for $\alpha = 1$, thus Theorem 4.19 guarantees the rate $\mathcal{O}(t^{-\frac{\beta-1}{\beta+1}})$. Moreover, suppose that the kernel satisfies some good embedding property, that is, $\alpha < 1$, since larger $\beta - \alpha$ leads to faster convergence rates. In that case, we obtain a better rate $\mathcal{O}(t^{-\frac{\beta-\alpha}{\beta-\alpha+2}})$ by setting the gradient steps and regularization terms ν_t, λ_t corresponding to α, β. Besides, `OPGD` can handle the challenging scenario ($f_i \notin (\mathcal{H})^p$ if $\beta < 1$) when the embedding property of kernel holds for $\alpha < \beta$. Furthermore, one could extend Theorem 4.19 to the situation where $\beta - \alpha \ge 2$, using the same proof steps, but the convergence rate will fix at $\mathcal{O}(t^{-1/2})$, this saturation phenomenon has been studied in Dieuleveut and Bach (2016); Lin and Cevher (2020); Li et al. (2023).

4.3 Convergence rate in the bandit feedback setting

In this section, we derive the convergence rate of the kernel OPGD in the bandit feedback setting. Compared with learning under the first-order oracle, the OPGD performs the additional gradient step (3.7) to estimate the unknown loss function ℓ_i leveraging a kernel function class $\mathcal{B}$. Suppose $\mathcal{B}$ is an RKHS associated with a continuous Mercer kernel $R : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and a measure $\rho_{\mathcal{Y}}$ (has the full support). Therefore, it is sufficient to derive the estimation error $\ell_i^t - \ell_i$, and the remaining steps would be analogous to Section 4.2.

For simplicity, we consider $\ell_i^t - \ell_i$ with respect to the RKHS norm $\|\cdot\|_{\mathcal{B}}$ instead of the power norm, and present the following assumption on the kernel function class $\mathcal{B}$, which is similar to Assumptions 4.13, 4.14, 4.15 for the kernel function class $\mathcal{H}$ and its power space. We remark that Assumption 4.20 is analogous Assumptions 4.13,4.14,4.15, where β corresponds to β', $\kappa = A$ corresponds to κ', $\alpha = 1$, and ξ' corresponds to ξ. The embedding property is included in the source condition since R is assumed to be a bounded kernel ($\sup_{y\in\mathcal{Y}} R(y,y) \le {\kappa'}^2$).

Assumption 4.20. (Assumptions on kernel).

1. (Source condition). Suppose Assumption 3.1 holds and there exists an RKHS, $\mathcal{B}$, with a bounded differentiable Mercer kernel, R, and constants $\beta' > 1, \kappa' > 0$ such that $\sup_{y\in\mathcal{Y}} R(y,y) \le {\kappa'}^2$ and $\ell_i \in \mathcal{B}^{\beta'}$ for all $i \in [n]$.

2. (Lipschitz kernel). There exists $\xi' > 0$ such that $\|\partial_i \varphi_y\|_{\mathcal{B}} \le \xi'$ for any $i \in [n]$ and $y \in \mathcal{Y}$, where $\varphi_y : \mathcal{X} \to \mathcal{B}$ is the feature map of the kernel R.

Now we derive the convergence of the kernel OPGD in the bandit feedback setting. As aforementioned, the key step would be bounding the estimation error $\|\ell_i^t - \ell_i\|_{\mathcal{B}}$.

Lemma 4.21. (Estimation error of ℓ_i^t). Suppose Assumption 4.20 holds. For all iterations t and positive constant a, define $\bar{t} = t + t_0$, where t_0 is a constant satisfies $t_0 \ge (a{\kappa'}^2 + 1)^2$. Set the gradient steps and regularization terms as

$$s_t = a\left(\frac{1}{\bar{t}}\right)^{\frac{\beta'}{\beta'+1}}, \qquad \iota_t = \frac{1}{a}\left(\frac{1}{\bar{t}}\right)^{\frac{1}{\beta'+1}}.$$

If $a < \sqrt{(\beta'+1)/\beta'} \cdot (t_0+1)/(t_0+2) \cdot {\kappa'}^{-2}$, the ℓ_i^t generated by OPGD using the gradient step (3.7) and kernel R satisfies

$$\mathbb{E}\|\ell_i^t - \ell_i\|_{\mathcal{B}}^2 \lesssim \mathcal{O}(t^{-\frac{\beta'-1}{\beta'+1}}). \tag{4.10}$$

Lemma 4.10 is a direct corollary of Lemma 4.18. In fact, the proofs are the same if we set $\beta = \beta'$, $\alpha = \gamma = 1$, and $\kappa = \kappa'$. Next, we combine the estimation error $\ell_i^t - \ell_i$ (Lemma 4.21) and $f_i^t - f_i$ (Lemma 4.18) to derive the convergence result.

Theorem 4.22. (Convergence in the bandit feedback setting). Suppose the assumptions in Theorem 4.19 and Lemma 4.21 hold. For all iterations $t \ge 1$ and positive constant a, define $\bar{t} = t + t_0$, where t_0 is a constant that satisfies $t_0 \geq (a\kappa^2+1)^2 \vee (a\kappa'^2+1)^2$. Set the gradient steps and regularization terms as

$$\eta_t = (\tau \bar{t})^{-1}, \quad \nu_t = a \cdot \bar{t}^{-\frac{\beta-\alpha+1}{\beta-\alpha+2}}, \quad \lambda_t = a^{-1} \cdot \bar{t}^{-\frac{1}{\beta-\alpha+2}}, \quad s_t = a \cdot \bar{t}^{-\frac{\beta'}{\beta'+1}}, \quad \iota_t = a^{-1} \cdot \bar{t}^{-\frac{1}{\beta'+1}}.$$

If $a < (t_0+1)/(t_0+2) \cdot \left(\sqrt{(\beta-\alpha+2)/(\beta-\alpha)}\kappa^{\alpha-2}A^{-1}\right) \wedge \left(\sqrt{(\beta'+1)/\beta'} \cdot \kappa'^{-2}\right)$, the x^t generated by the `OPGD` algorithm in Section 3 leveraging online estimation steps (3.5), (3.8) (use kernels K and R) and projected gradient steps (3.6) satisfies

$$\mathbb{E}\|x^t - x^*\|^2 \lesssim \mathcal{O}\left(t^{-\left(\frac{\beta-\alpha}{\beta-\alpha+2}\right) \wedge \left(\frac{\beta'-1}{\beta'+1}\right)}\right). \tag{4.11}$$

Given Lemma 4.18 and Lemma 4.21, Theorem 4.22 is a corollary of Theorem 4.6. We remark that the parameter t_0 is a sufficiently large value, a is a sufficiently small value, and other parameters are determined in the assumptions involved in the statement. The convergence rate $\mathcal{O}(t^{-(\frac{\beta-\alpha}{\beta-\alpha+2}) \wedge (\frac{\beta'-1}{\beta'+1})})$ is the minimum of the estimation errors $\mathbb{E}\|f_i^t - f_i\|_\gamma^2$ and $\mathbb{E}\|\ell_i^t - \ell_i\|_{\mathcal{B}}^2$. Since $(\frac{\beta-\alpha}{\beta-\alpha+2}) \wedge (\frac{\beta'-1}{\beta'+1}) = \frac{(\beta-\alpha)\wedge(\beta'-1)}{(\beta-\alpha)\wedge(\beta'-1)+2}$, (4.11) implies the convergence rate of `OPGD` is mostly determined by the regularity of the loss function ℓ_i and the parametric function f_i, namely, a larger $\beta - \alpha$ or $\beta' - 1$ might leads to a faster convergence rate.

5 Numerical Experiments

In this section, we conduct experiments on multi-agent decision-dependent games in both the linear and the RKHS settings to verify our theory. All experiments are conducted with Python on a laptop using 14 threads of a 12th Gen Intel(R) Core(TM) i7-12700H CPU.

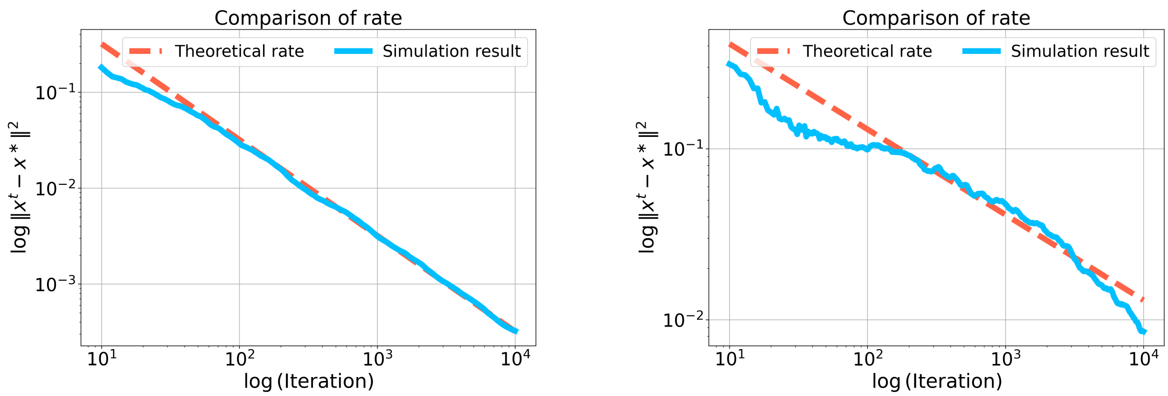
5.1 Convergence Rate Analysis

Basic Setup. We consider two-agent decision-dependent games with 1-dimensional actions. Namely, for all $i \in [2]$, define the game

$$\min_{x \in \mathcal{X}} \mathcal{L}_i(x), \quad \text{where} \quad \mathcal{L}_i(x) \coloneqq \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(x)} \ell_i(x, z_i), \tag{5.1}$$

where $\mathcal{X} = [0,1] \times [0,1]$, $x \in \mathcal{X}$, $z_i \in \mathbb{R}$, and $\ell_i(x, z_i)$ is the loss function to be determined. Let the distribution map be $\mathcal{D}_i(x) \sim \mathcal{N}(f_i(x), 0.2)$, where f_i is the parametric function determined by the specific function class. Then the game (5.1) follows the parametric assumption (Assumption 2.6) with $z_i = f_i(x) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 0.2)$ the independent Gaussian noise term.

Linear parametric model. Let the loss function be $\ell_i(x, z_i) = -z_i + x_i^2$ and set the linear parametric function as $f_1(x) = x_1$ and $f_2(x) = 2x_2$, namely, the parametric model is $z_i = A_i x + \epsilon_i$ where $A_1 = [1\ 0]$ and $A_2 = [0\ 2]$. The the game (5.1) has the gradient $H(x) = (2x_1 - 1, 2x_2 - 2)$, therefore, the game (5.1) is convex, C^1-smooth, 1-strongly monotone and the Nash equilibrium is $x^* = (1/2, 1)$. We conduct a simulation based on this model to check the convergence rate. We



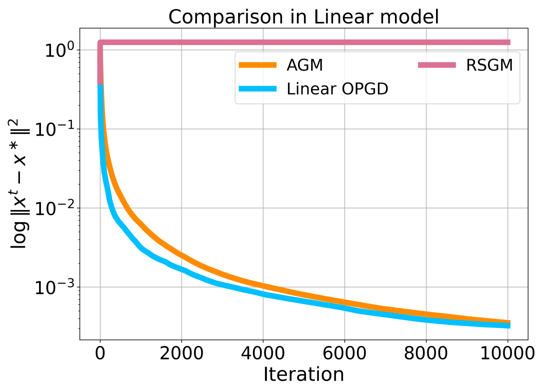
(a) (b)

Figure 1: **(a) Linear setting:** The X-axis represents the iteration from 1 to 10,000, while the Y-axis represents the norm-squared error of x^t to the Nash equilibrium $x^* = (1/2, 1)$, averaged over 20 random seeds. Both axes are on a log scale. The blue solid line represents the output of `OPGD` and the orange dashed line represents the theoretical rate $\mathcal{O}(t^{-1})$. **(b) RKHS setting:** The X-axis represents the iteration from 1 to 10,000, while the Y-axis represents the norm-squared error to the Nash equilibrium $x^* = (1/2, 1/2)$, averaged over 400 random seeds. Both axes are on a log scale. The blue solid line represents the output of `OPGD` and the orange dashed line represents the theoretical rate $\mathcal{O}(t^{-1/2})$.

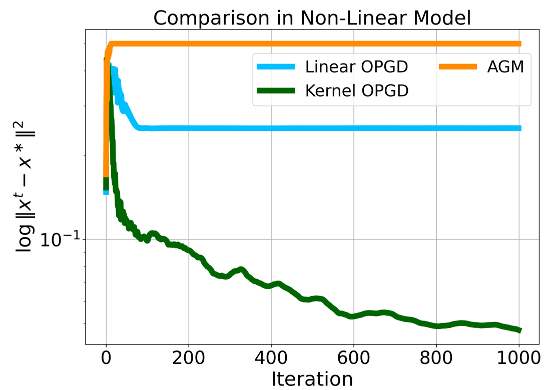
set the sampling distribution as $\rho_{\mathcal{X}} = \mathcal{U}[0,1] \times \mathcal{U}[0,1]$, the initial point as $x^0 \sim \rho_{\mathcal{X}}$, and the initial estimation as zero. Moreover, letting $t_0 = 10$, we set the gradient step sizes as $\eta_t = 6/(t + t_0), \nu_t = 6/(t + t_0)$.

Kernel parametric model. Following Example 4.17, we set $\mathcal{X} = [0,1] \times [0,1]$, $\rho_{\mathcal{X}} = \mathcal{U}[0,1] \times \mathcal{U}[0,1]$, and define the kernel $Q((x_1, x_2), (y_1, y_2)) = K(x_1, y_1) \cdot K(x_2, y_2)$ as the product kernel of $K(x, y) = 40B_4(\{x - y\})$ (i.e. $K(x, y) = 960R_{2m}(x, y)$ where $m = 2$). Suppose that $\mathcal{H}$ is the RKHS on $\mathcal{X}$ induced by the kernel Q and the distribution $\rho_{\mathcal{X}}$. Set the parametric function as the product of two 3-order Bernoulli polynomials, namely, $f(x_1, x_2) = B_3(x_1) \cdot B_3(x_2) = (x_1^3 - 3x_1^2/2 + x_1/2) \cdot (x_2^3 - 3x_2^2/2 + x_2/2)$. Since $\cos(2\pi i x)$ and $\sin(2\pi i x)$ are eigenfunctions of the kernel K for $i \in \mathbb{N}^+$, by the property of product kernel, $\sin(2\pi i x_1)\sin(2\pi j x_2), \cos(2\pi i x_1)\sin(2\pi j x_2), \cos(2\pi i x_1)\cos(2\pi j x_2)$ are eigenfunctions of the kernel Q for all $i, j \in \mathbb{N}^+$. Therefore, we set $\ell_i(x, z_i) = -z_i + \cos(2\pi x_1)\cos(2\pi x_2) - x_i + x_i^2$ and let $f_i(x) = \cos(2\pi x_1)\cos(2\pi x_2)$ for $i \in [2]$. Then the gradient of this game is $H(x) = (2x_1 - 1, 2x_2 - 1)$, thus, this game is convex, C^1-smooth, 1-strongly monotone and the Nash equilibrium is $x^* = (0.5, 0.5)$.

Following Example 4.17, Assumption 4.13, 4.14, 4.15 hold for any $\beta > 1$ and any $\alpha > 1/4$. Set $t_0 = 10$ and $a = 7$, we set the gradient step sizes as $\eta_t = 6/(t + t_0), \nu_t = a/(t + t_0)^{3/4}, \lambda_t = 1/(a(t + t_0)^{1/4})$ (because $(\beta - \alpha) \wedge 2 = 2$, which corresponds to $(\beta - \alpha + 1)/(\beta - \alpha + 2) = 3/4$). Thus, following Theorem 4.19, the convergence rate is $\mathcal{O}(t^{-1/2})$.



(a)



(b)

Figure 2: **(a) Linear setting:** Comparison of Linear OPGD averaged by 400 seeds (blue), AGM averaged by 20 seeds (orange), and RSGM averaged by 20 seeds (red) in the proposed linear model. The Y-axe takes the log scale. **(b) RKHS setting:** Comparison of Kernel OPGD averaged by 400 seeds (green), Linear OPGD averaged by 20 seeds (blue), AGM averaged by 20 seeds (orange) in the proposed non-linear model. The Y-axe takes the log scale.

Results. We perform experiments for both parametric settings to verify the convergence rates and compare the theoretical and simulated rates, as shown in Figure 1. Since both X and Y axes take the log scale in Figure 1, the slopes of these lines denote the convergence rates. Figure 1(a) shows the converge rate of the linear setting within $10,000$ iterations, the simulated rate matches our prediction, i.e. it is close to $\mathcal{O}(t^{-1})$. Figure 1(b) shows the convergence rate of the RKHS setting within $10,000$ iterations, it implies that the simulated rate is close to the theoretical rate $\mathcal{O}(t^{-1/2})$ when the iteration t is larger than $1,000$. These results validate Theorems 4.9 and 4.19.

5.2 Comparison with Baseline Algorithms

In this section, we compare the performance of Linear OPGD (Algorithm 1), Kernel OPGD (Algorithm 2), AGM (Algorithm 1 in Narang et al. (2022)), and the baseline algorithm in performative prediction RSGM (Section 4.3, Repeated Stochastic Gradient Method in Narang et al. (2022)).

Figure 2(a) compares the performance of the linear OPGD, adaptive gradient methods (AGM), and Repeated Stochastic Gradient Method (RSGM) on the two-player decision-dependent game with the linear parametric model as previously described. Letting $t_0 = 10$, for AGM, we set the injective noise as $\mathcal{N}(0, 0.32)$, and the gradient steps are the same as linear OPGD. For RSGM, we set the gradient steps as $\eta_t = 5/(t + t_0)$. Figure 2(a) shows that for the decision-dependent game with linear parametric function, linear OPGD and AGM converge to the Nash equilibrium $(1/2, 1)$ with the same rate $\mathcal{O}(1/t)$, while RSGM fails to find the Nash equilibrium. This is because RSGM only uses the term $\nabla_i \ell_i(x, z_i)$ of the performative gradient (2.6) for gradient descent, and ignores the dependence between the distribution $\mathcal{D}_i$, consequently, it cannot characterize the decision-dependent distribution.

Figure 2(b) compares the performance of the Kernel OPGD, the linear OPGD, and AGM on the game with the aforementioned non-linear parametric model. We set the gradient steps of linear OPGD and AGM as $\eta_t = 4/(t+t_0), \nu_t = 4/(t+t_0)$ where $t_0 = 10$, and set the injective noise of AGM as $\mathcal{N}(0, 0.32)$. The gradient steps and regularization terms of the kernel OPGD are $\eta_t = 6/(t+t_0), \nu_t = 7/(t+t_0)^{3/4}, \lambda_t = 7/(t+t_0)^{1/4}$. Figure 2(b) shows that for proposed the non-linear parametric function, the kernel OPGD converges to the Nash equilibrium $(1/2, 1/2)$, while both the linear OPGD and AGM fail to find the NE. In fact, the linear OPGD and AGM approximate the parametric function by linear models and have large estimation errors, thus, the estimated performative gradient (2.8) has a constant bias and makes the projected gradient descent fails to converge.

5.3 Semi-Synthetic Simulation: Revenue Maximization

In this section, we conduct a semi-synthetic simulation for revenue maximization of the rideshare market (Example 2.4). We study the rideshare market in Boston from November 26, 2018, to December 18, 2018. To elaborate, we consider a multi-agent decision-dependent game, where the strategic agents are Uber and Lyft. [1] We perform both linear OPGD and kernel OPGD on this game and calculate the corresponding revenue.

Game construction. We set up the decision-dependent game analogous to Example 2.4. In more detail, we consider a ride-share market with two companies as strategic agents. Each company i sets the price $x_i \in \mathbb{R}$ and seeks to maximize its revenue $z_i x_i$, where $z_i \in \mathbb{R}$ is the demand generated by the strategic users. We remark that the demand z_i is decision-dependent because users compare the prices x_i among all the companies. Suppose that company i's loss function ℓ_i is defined by

$$\ell_i(x, z_i) = -z_i x_i + \lambda_i x_i^2,$$

where $\lambda_i \geq 0$ is the some regularization parameter and $x = (x_i)_{i\in[n]}$ is the joint action that represents the prices of all the companies. Intuitively, this loss function is the negative revenue plus some regularization term to ensure the game is strongly monotone. Next, we learn the decision-dependent distributions $\mathcal{D}_i$ from the aforementioned dataset of the rideshare market in Boston. To elaborate, this dataset contains the prices for 72 different routes of Uber and Lyft. Consequently, there are two strategic users and we use the record for the route starting from Back Bay and ending at Boston University to learn $\mathcal{D}_i$ for all $i \in [2]$.

We model these decision-dependent distributions following the parametric assumption, namely, $z_i = f_i(x) + \epsilon_i$, where f_i is the unknown parametric function and $\epsilon_i \sim \mathcal{N}(0, 10)$ is the independent Gaussian noise. We remark the prices in this dataset are rounded to one decimal place, specifically to the nearest .5. Thus, we count the number of prices from 5 to 25 and approximate the parametric function f_i by the period kernel (Example 4.17) using kernel ridge regression. Note that here we fit a non-linear parametric model, which is more flexible. Specifically, we set $\mathcal{X} = [0,1] \times [0,1]$,

[1] The data used in this paper is publicly available (https://www.kaggle.com/datasets/brllrb/uber-and-lyft-dataset-boston-ma).

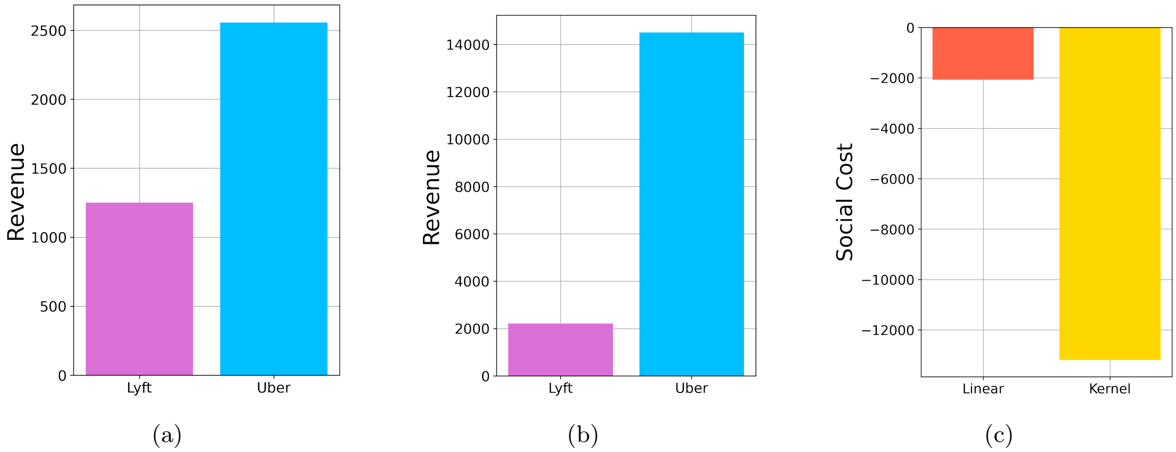


Figure 3: **(a) Linear OPGD:** Revenue corresponds with the Nash equilibrium $(10.610581, 7.802483)$ obtained by the linear OPGD, where the revenue of Lyft is 1251.447 and of Uber is 2556.880. **(b) Kernel OPGD:** Revenue corresponds with the Nash equilibrium $(10.776331, 15.376748)$ obtained by the kernel OPGD, where the revenue of Lyft is 2208.374 and of Uber is 14512.855. **(c) Social costs:** Social costs for Lyft and Uber (i.e. the sum of loss functions for these two agents) obtained by the linear and kernel OPGD, the social cost for the linear case is -2073.695 and for the kernel case is -13195.492.

$\rho_{\mathcal{X}} = \mathcal{U}[0,1] \times \mathcal{U}[0,1]$, and the kernel as $Q((x_1, x_2), (y_1, y_2)) = K(x_1, y_1) \cdot K(x_2, y_2)$, where $K(x, y) = 65B_4(\{x - y\})$. We use outputs of this kernel regression, which are estimations for f_1 and f_2, to generate the demand z_i synthetically for OPGD.

Revenue and social cost. We use the linear OPGD and kernel OPGD for the aforementioned two-agent decision-dependent game to find the Nash equilibrium. For the linear OPGD, we set gradient step sizes as $t_0 = 1e5, \eta_t = 0.5/(t + t_0)$, and $\nu_t = 0.5/(t + t_0)$. For the kernel OPGD, we set the gradient step and regularization terms as $t_0 = 2e5, \eta_t = 0.1/(t + t_0), \nu_t = 7/(t + t_0)^{3/4}$, and $\lambda_t = 1/(7(t + t_0)^{1/4})$. We run the linear OPGD for 10 different seeds with 10000 iterations and compute the average as the Nash equilibrium, similarly, we run the kernel OPGD for 10 different seeds with 10000 iterations and take the average. The results show that the linear OPGD converges to $(10.610581, 7.802483)$ and the kernel OPGD converges to $(10.776331, 15.376748)$. Next, we calculate the loss functions $\mathcal{L}_i$ for both Uber and Lyft as well as their revenue. Moreover, to evaluate the efficiency of the Nash equilibrium, we compute the social cost for each equilibrium, which is defined as the sum of all the agents' individual rewards:

$$\mathcal{S}(x) \coloneqq \sum_{i=1}^{2} \mathcal{L}_i(x).$$

We use the aforementioned estimated decision-dependent distribution $\mathcal{D}_i$ to calculate the revenue as well as social cost for the Nash equilibria obtained by the linear OPGD and the kernel OPGD. Figure 3(a) presents the revenue for both Lyft and Uber corresponding with the Nash equilibrium

(10.610581, 7.802483) calculated by the linear `OPGD` and Figure 3(b) presents the revenue for these companies obtained by the kernel `OPGD`. These plots show that for both the linear and non-linear methods, Uber has a larger revenue compared with Lyft, especially for the Nash equilibrium obtained by the kernel `OPGD`. Moreover, Figure 3(c) presents the social costs for the equilibria obtained by both the linear and kernel `OPGD`, it shows that the social cost obtain by the kernel algorithm is much smaller than the linear algorithm, which further implies that the equilibrium obtained by the kernel `OPGD` has better social efficiency.

6 Conclusion and Discussion

In this paper, we study the problem of learning Nash equilibria in multi-agent decision-dependent games. We propose a parametric assumption to handle the distribution shift and develop a novel online algorithm `OPGD` in both the linear and RKHS settings. We derive sufficient conditions to ensure the decision-dependent game is strongly monotone under the parametric assumption. Given the first-order oracle, we show that `OPGD` converges to the Nash equilibrium at a rate of $\mathcal{O}(t^{-1})$ in the linear setting and $\mathcal{O}(t^{-\frac{\beta-\alpha}{\beta-\alpha+2}})$ in the RKHS setting. We further extend the algorithm into the bandit feedback setting and derive the corresponding convergence rate.

References

Abramowitz, M., Stegun, I. A. and Romer, R. H. (1988). Handbook of mathematical functions with formulas, graphs, and mathematical tables.

Adams, R. A. and Fournier, J. J. (2003). *Sobolev spaces*. Elsevier.

Agarwal, N., Gonen, A. and Hazan, E. (2019). Learning in non-convex games with an optimization oracle. In *Conference on Learning Theory*. PMLR.

Allen-Zhu, Z., Li, Y. and Liang, Y. (2019). Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, **32**.

Bach, F. and Moulines, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). *Advances in neural information processing systems*, **26**.

Blanchard, G. and Mücke, N. (2018). Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, **18** 971–1013.

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

Bravo, M., Leslie, D. and Mertikopoulos, P. (2018). Bandit learning in concave n-person games. *Advances in Neural Information Processing Systems*, **31**.

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

Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, **7** 331–368.

Cotter, A., Jiang, H. and Sridharan, K. (2019). Two-player games for efficient non-convex constrained optimization. In *Algorithmic Learning Theory*. PMLR.

Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of the American mathematical society*, **39** 1–49.

Cutler, J., Díaz, M. and Drusvyatskiy, D. (2022). Stochastic approximation with decision-dependent distributions: asymptotic normality and optimality. *arXiv preprint arXiv:2207.04173*.

De Vito, E., Rosasco, L., Caponnetto, A., De Giovannini, U., Odone, F. and Bartlett, P. (2005). Learning from examples as an inverse problem. *Journal of Machine Learning Research*, **6**.

Debreu, G. (1952). A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences*, **38** 886–893.

Dicker, L. H., Foster, D. P. and Hsu, D. (2017). Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators. *Electronic Journal of Statistics*, **11** 1022–1047.

Dieuleveut, A. and Bach, F. (2016). Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, **44** 1363 – 1399.

Dieuleveut, A., Flammarion, N. and Bach, F. (2017). Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, **18** 3520–3570.

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

Drusvyatskiy, D., Fazel, M. and Ratliff, L. J. (2022). Improved rates for derivative free gradient play in strongly monotone games. In *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE.

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

Fischer, S. and Steinwart, I. (2020). Sobolev norm learning rates for regularized least-squares algorithms. *The Journal of Machine Learning Research*, **21** 8464–8501.

Fudenberg, D. and Tirole, J. (1991). *Game theory*. MIT press.

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

Izzo, Z., Ying, L. and Zou, J. (2021). How to learn when data reacts to your model: performative gradient descent. In *International Conference on Machine Learning*. PMLR.

Lei, Y., Hu, T. and Tang, K. (2021). Generalization performance of multi-pass stochastic gradient descent with convex loss functions. *The Journal of Machine Learning Research*, **22** 1145–1185.

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

Li, Y., Zhang, H. and Lin, Q. (2023). On the saturation effect of kernel ridge regression. In *The Eleventh International Conference on Learning Representations*.

Lin, J. and Cevher, V. (2020). Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *The Journal of Machine Learning Research*, **21** 5852–5914.

Lin, J. and Rosasco, L. (2017). Optimal rates for multi-pass stochastic gradient methods. *The Journal of Machine Learning Research*, **18** 3375–3421.

Lin, T., Zhou, Z., Ba, W. and Zhang, J. (2021). Optimal no-regret learning in strongly monotone games with bandit feedback. *arXiv preprint arXiv:2112.02856*.

Liu, Z. and Li, M. (2020). On the estimation of derivatives using plug-in krr estimators. *arXiv preprint arXiv:2006.01350*.

Lu, Y., Blanchet, J. and Ying, L. (2022). Sobolev acceleration and statistical optimality for learning elliptic equations via gradient descent. *arXiv preprint arXiv:2205.07331*.

Mendelson, S. and Neeman, J. (2010). Regularization in kernel learning. *Annals of statistics*, **38** 526–565.

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

Mertikopoulos, P. and Zhou, Z. (2019). Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, **173** 465–507.

Miller, J., Milli, S. and Hardt, M. (2020). Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*. PMLR.

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

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

Nash Jr, J. (1996). Non-cooperative games. In *Essays on Game Theory*. Edward Elgar Publishing, 22–33.

Nesterov, Y. et al. (2018). *Lectures on convex optimization*, vol. 137. Springer.

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

Piliouras, G. and Yu, F.-Y. (2022). Multi-agent performative prediction: From global stability and optimality to chaos. *arXiv preprint arXiv:2201.10483*.

Pillaud-Vivien, L., Rudi, A. and Bach, F. (2018). Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems*, **31**.

Ratliff, L. J., Burden, S. A. and Sastry, S. S. (2016). On the characterization of local nash equilibria in continuous games. *IEEE transactions on automatic control*, **61** 2301–2307.

Rosasco, L., Belkin, M. and De Vito, E. (2010). On learning with integral operators. *Journal of Machine Learning Research*, **11**.

Rosen, J. B. (1965). Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society* 520–534.

Shamir, O. (2013). On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*. PMLR.

Smale, S. and Zhou, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive approximation*, **26** 153–172.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Steinwart, I., Hush, D. R., Scovel, C. et al. (2009). Optimal rates for regularized least squares regression. In *COLT*.

Steinwart, I. and Scovel, C. (2012). Mercer's theorem on general domains: On the interaction between measures, kernels, and rkhss. *Constructive Approximation*, **35** 363–417.

Tarres, P. and Yao, Y. (2014). Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. *IEEE Transactions on Information Theory*, **60** 5716–5735.

Tatarenko, T. and Kamgarpour, M. (2019). Learning nash equilibria in monotone games. In *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE.

Tatarenko, T. and Kamgarpour, M. (2020). Bandit online learning of nash equilibria in monotone games. *arXiv preprint arXiv:2009.04258*.

Triebel, H. (1995). Interpolation theory, function spaces. *Differential Operators*.

Triebel, H. (2010). *Theory of Function Spaces*. Modern Birkhäuser Classics, Springer Basel.

Von Stackelberg, H. (2010). *Market structure and equilibrium*. Springer Science & Business Media.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Wood, K., Bianchin, G. and Dall'Anese, E. (2021). Online projected gradient descent for stochastic optimization with decision-dependent distributions. *IEEE Control Systems Letters*, **6** 1646–1651.

Yao, Y., Rosasco, L. and Caponnetto, A. (2007). On early stopping in gradient descent learning. *Constructive Approximation*, **26** 289–315.

Ying, Y. and Pontil, M. (2008). Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, **8** 561–596.

A Pseudocode

Algorithm 1: `OPGD` in the Linear setting

Input: Initial x^1, initial $A_i^0 = 0$, step sizes of GD $\{\eta_t\}_{t\in\mathbb{N}}$, step sizes of stochastic approximation $\{\nu_t\}_{t\in\mathbb{N}}$, sampling distribution $\rho_{\mathcal{X}}$

for $t \in \mathbb{N}$ **do**
  for $i \in [n]$ **do**
    1. **Random sampling:** Draw sample $u_i^t \sim \rho_{\mathcal{X}}$;
    2. **Query the environment:** Draw sample $y_i^t \sim \mathcal{D}_i(u_i^t)$;
    3. **Estimation update:** $A_i^t \leftarrow A_i^{t-1} - \nu_t \left(A_i^{t-1} u_i^t - y_i^t\right) (u_i^t)^\top$;
    4. **Query the environment:** Draw sample $z_i^t \sim \mathcal{D}_i(x^t)$;
    5. **Individual gradient update:**
    $x_i^{t+1} \leftarrow \mathrm{proj}_{\mathcal{X}_i} \left(x_i^t - \eta_t \left(\nabla_i \ell_i(x^t, z_i^t) + (A_{ii}^t)^\top \nabla_{z_i} \ell_i(x^t, z_i^t)\right)\right)$;
  end
end

Algorithm 2: `OPGD` in the RKHS setting

Input: Kernel K, initial x^1, initial $f_i^0 = 0$, step sizes of GD $\{\eta_t\}_{t\in\mathbb{N}}$, step sizes of stochastic approximation $\{\nu_t\}_{t\in\mathbb{N}}$, regularization term $\{\lambda_t\}_{t\in\mathbb{N}}$, sampling distribution $\rho_{\mathcal{X}}$

for $t \in \mathbb{N}$ **do**
  for $i \in [n]$ **do**
    1. **Random sampling:** Draw sample $u_i^t \sim \rho_{\mathcal{X}}$;
    2. **Query the environment:** Draw sample $y_i^t \sim \mathcal{D}_i(u_i^t)$;
    3. **Estimation update:** $f_i^t \leftarrow f_i^{t-1} - \nu_t \left[\left(f_i^{t-1}(u_i^t) - y_i^t\right) \phi_{u_i^t} + \lambda_t f_i^{t-1}\right]$;
    4. **Query the environment:** Draw sample $z_i^t \sim \mathcal{D}_i(x^t)$;
    5. **Individual gradient update:**
    $x_i^{t+1} \leftarrow \mathrm{proj}_{\mathcal{X}_i} \left(x_i^t - \eta_t \left(\nabla_i \ell_i(x^t, z_i^t) + \left(\langle f_i^t, \partial_i \phi_{x^t} \rangle_{\mathcal{H}}\right)^\top \nabla_{z_i} \ell_i(x^t, z_i^t)\right)\right)$;
  end
end

B Preliminary

B.1 Lemma for Stochastic Gradient Methods

In this section, we present a lemma (Lemma B.2) for the projected stochastic gradient method with bias following Narang et al. (2022). This lemma plays an important role in the convergence analysis of Algorithm 1 and Algorithm 2, specifically, in the proofs of Lemma 4.10 and Lemma C.1.

We consider the variational inequality

$$0 \in H(x) + N_{\mathcal{X}}(x), \tag{B.1}$$

where $\mathcal{X} \subset \mathbb{R}^d$ the compact and convex joint action set and $H : \mathbb{R}^d \to \mathbb{R}^d$ is an L-Lipschitz continuous and τ-strongly monotone map. Recalling (2.3), this inequality characterizes the Nash equilibrium x^* of the game (2.1):

$$x_i^* = \arg\min_{x_i \in \mathcal{X}_i} \mathcal{L}_i(x_i, x_{-i}^*).$$

Suppose we use projected stochastic gradient descent to find x^* and perform the following update in each iteration:

$$x^{t+1} = \mathrm{proj}_{\mathcal{X}}(x^t - \eta h^t), \tag{B.2}$$

where η is the gradient step size and h^t is the estimator of $H(x^t)$. Mathematically, we make the following assumption on the randomness of the estimator h_t.

Assumption B.1. (Stochastic framework). Suppose that there exist a filtered probability space $(\Omega, \mathbb{P})$ with filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\in\mathbb{N}}$ such that $\mathcal{F}_0 = \{\varnothing, \Omega\}$. Suppose that h^t is $\mathcal{F}_{t+1}$-measurable and there exist constants $U, V \geq 0$ and $\mathcal{F}_t$-measurable random variables $m_t, \sigma_t \geq 0$ such that the following inequalities hold

$$\begin{aligned} &\text{(Bias)} \quad && \|\mathbb{E}_t h^t - H(x^t)\| \leq m_t + U\|x^t - x^*\|, \\ &\text{(Variance)} \quad && \mathbb{E}_t\|h^t - \mathbb{E}_t h^t\|^2 \leq \sigma_t^2 + V^2\|x^t - x^*\|, \end{aligned} \tag{B.3}$$

where $\mathbb{E}_t = \mathbb{E}[\cdot|\mathcal{F}_t]$ denotes the conditional expectation with $\mathcal{F}_t$.

If Assumption B.1 holds, then the following one-step error bound holds for the projected stochastic gradient descent when constant $U = 0$. In this case, x^t converges to a neighborhood of x^* and radius of the neighborhood is depend on $\{m_t\}_{t\in\mathbb{N}}$,

Lemma B.2. (One-step error). Suppose $U = 0$ and $\eta \leq \tau/(4L^2)$, then the iterates x^t generated by the projected stochastic gradient method satisfy the inequality

$$\mathbb{E}_t\|x^{t+1} - x^*\|^2 \leq \frac{1 + 2\eta^2 V^2}{1 + \eta\tau}\|x^t - x^*\|^2 + \frac{2\eta^2\sigma_t^2}{1 + \eta\tau} + \frac{2\eta m_t^2}{\tau(1 + \eta\tau)}. \tag{B.4}$$

Proof. See Narang et al. (2022, Theorem 8) for a detailed proof. □

B.2 Basic of RKHS

In this section, we summarize the basic properties of RKHS. We refer readers to Cucker and Smale (2002); Smale and Zhou (2007); Steinwart and Christmann (2008); Steinwart and Scovel (2012); Fischer and Steinwart (2020) for the mathematical foundations of RKHS.

Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$ and let $\rho_{\mathcal{X}}$ be a probability measure on $\mathcal{X}$. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous Mercer kernel, namely, a continuous, symmetric, and positive semi-definite real function. The Mercer kernel K and measure $\rho_{\mathcal{X}}$ induce a unique RKHS $\mathcal{H}$. Let $\phi : \mathcal{X} \to \mathcal{H}$ be the feature map of the kernel K, namely, $\phi_x := K(\cdot, x) \in \mathcal{H}$ for all $x \in \mathcal{X}$. The most important property of RKHS is the reproducing property: $f(x) = \langle f, \phi_x \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$. Define the integral operator $L_K : \mathcal{L}^2_{\rho_{\mathcal{X}}} \to \mathcal{H}$ by the following integral transform

$$L_K(f)(x) := \int_{\mathcal{X}} K(x,t) f(t) d\rho_{\mathcal{X}}(t), \quad \forall f \in \mathcal{H}, \forall x \in \mathcal{X}.$$

For any $x \in \mathcal{X}$, let $\phi_x^* : \mathcal{H} \to \mathbb{R}$ be the dual of ϕ_x, which satisfies $\phi_x^*(f) = \langle f, \phi_x \rangle_{\mathcal{H}} = f(x)$ for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$. With this notation, we define the operator $L_x : \mathcal{H} \to \mathcal{H}$

$$L_x(f) := (\phi_x^* \phi_x)(f) = f(x)\phi_x, \quad \forall f \in \mathcal{H}, \forall x \in \mathcal{X}.$$

This definition implies that L_x is a compact, self-adjoint, and positive-semidefinite operator on $\mathcal{H}$. Using the reproducing property, the covariance operator $L_K|_{\mathcal{H}} : \mathcal{H} \to \mathcal{H}$ is the expectation of L_x:

$$L_K|_{\mathcal{H}}(\cdot) = \mathbb{E}_{x \sim \rho_{\mathcal{X}}}[\langle \cdot, K_x \rangle_{\mathcal{H}} K_x] = \mathbb{E}_{x \sim \rho_{\mathcal{X}}} \phi_x^* \phi_x.$$

Define the natural inclusion $I_{\rho_{\mathcal{X}}} : \mathcal{H} \to \mathcal{L}^2_{\rho_{\mathcal{X}}}$, mapping a function $f \in \mathcal{H}$ to $\mathcal{L}^2_{\rho_{\mathcal{X}}}$ (since K is a Mercer kernel, any $f \in \mathcal{H}$ is square-integrable). According to Steinwart and Christmann (2008); Fischer and Steinwart (2020), the operator $I_{\rho_{\mathcal{X}}}$ is well-defined, Hilbert-Schmidt, the Hilbert-Schmidt norm is finite and satisfies

$$\|I_{\rho_{\mathcal{X}}}\|_{\mathrm{HS}} = \left(\int_{\mathcal{X}} k(x,x) d\rho_{\mathcal{X}}(x) \right)^{1/2} < \infty.$$

Moreover, the adjoint operator of $I_{\rho_{\mathcal{X}}}$ is the integral operator L_K, namely, $I^*_{\rho_{\mathcal{X}}} = L_K$. Therefore, the covariance operator $L_K|_{\mathcal{H}}$ is compact, self-adjoint, and positive semi-definite:

$$L_K|_{\mathcal{H}} = L_K I_{\rho_{\mathcal{X}}} = I^*_{\rho_{\mathcal{X}}} I_{\rho_{\mathcal{X}}}.$$

By Mercer's theorem, the kernel K has the spectral decomposition $K = \sum_{i=1}^{\infty} \mu_i e_i \otimes e_i$, where $\otimes$ denotes the tensor product, $\{\mu_i\}_{i=1}^{\infty}$ are eigenvalues and $\{e_i\}_{i=1}^{\infty}$ are eigenfunctions of the operator L_K. Here $\{e_i\}_{i=1}^{\infty}$ is the orthogonal basis of $\mathcal{L}^2_{\rho_{\mathcal{X}}}$ and $\{\mu_i^{1/2} e_i\}_{i=1}^{\infty}$ is the orthogonal basis of $\mathcal{H}$, namely, $\mathcal{L}^2_{\rho_{\mathcal{X}}} = \{\sum_{i=1}^{\infty} a_i e_i : \{a_i\}_{i=1}^{\infty} \in \ell^2\}, \mathcal{H} = \{\sum_{i=1}^{\infty} a_i \mu_i^{1/2} e_i : \{a_i\}_{i=1}^{\infty} \in \ell^2\}$.

Following Steinwart and Scovel (2012, Theorem 2.11), L_K has the spectral representation

$$L_K = \sum_{i=1}^{\infty} \mu_i^{1/2} \langle e_i, \cdot \rangle_{\mathcal{L}_{\rho^2_{\mathcal{X}}}} \mu_i^{1/2} e_i = \sum_{i=1}^{\infty} \mu_i \langle \mu_i^{1/2} e_i, \cdot \rangle_{\mathcal{H}} \mu_i^{1/2} e_i. \tag{B.5}$$

We further define the operator $L_K^{\alpha} : \mathcal{L}^2_{\rho_{\mathcal{X}}} \to \mathcal{L}^2_{\rho_{\mathcal{X}}}$ such that

$$L_K^{\alpha} := \sum_{i=1}^{\infty} \mu_i^{\alpha} \langle e_i, \cdot \rangle_{\mathcal{L}_{\rho^2_{\mathcal{X}}}} e_i,$$

namely, $L_K^{\alpha}(f) = \sum_{i=1}^{\infty} a_i \mu_i^{\alpha} e_i$ for all $f = \sum_{i=1}^{\infty} a_i e_i \in \mathcal{L}^2_{\rho_{\mathcal{X}}}$. It is well-known that $L_K^{1/2} : \mathcal{L}^2_{\rho_{\mathcal{X}}} \to \mathcal{H}$ is an isometric isomorphism between $\mathcal{L}^2_{\rho_{\mathcal{X}}}$ and $\mathcal{H}$.

B.3 Power Spaces

Given the Mercer kernel K and measure $\rho_{\mathcal{X}}$, suppose that μ_i, e_i are eigenvalues and eigenfunctions. For any $\alpha > 0$, define the α-power space: $\mathcal{H}^\alpha = \left\{\sum_{i=1}^\infty a_i \mu_i^{\alpha/2} e_i : \{a_i\}_{i=1}^\infty \in \ell^2\right\}$, equipped with the α-power norm $\|\cdot\|_\alpha$ and inner product $\langle\cdot,\cdot\rangle_\alpha$, where $\|\sum_{i=1}^\infty a_i \mu_i^{\alpha/2} e_i\|_\alpha := \left(\sum_{i=1}^\infty a_i^2\right)^{1/2}$ and $\langle\sum_{i=1}^\infty a_i \mu_i^{\alpha/2} e_i, \sum_{i=1}^\infty b_i \mu_i^{\alpha/2} e_i\rangle_\alpha = \sum_{i=1}^\infty a_i b_i$. We summarize the basic properties of power spaces and refer readers to Fischer and Steinwart (2020) for a detailed review.

(i) $\mathcal{H}^1 = \mathcal{H}$, $\mathcal{H}^0 = \mathcal{L}^2_{\rho_{\mathcal{X}}}$, and $\mathcal{H}^\alpha \subset \mathcal{H}^\beta \subset \mathcal{L}^2_{\rho_{\mathcal{X}}}$ for any $\alpha > \beta > 0$.

(ii) $\|\cdot\|_1 = \|\cdot\|_{\mathcal{H}}$ and $\|\cdot\|_0 = \|\cdot\|_{\rho_{\mathcal{X}}}$.

(iii) $\mathcal{H}^\alpha$ is an RKHS on $\mathcal{X}$ induced by the kernel $K^\alpha := \sum_{i=1}^\infty \mu_i^\alpha e_i \otimes e_i$ and measure $\rho_{\mathcal{X}}$ if the kernel K^α is bounded, namely, the embedding property (Assumption 4.14) holds for α.

(iv) $\{\mu_i^\alpha\}_{i=1}^\infty$ are eigenvalues and $\{e_i\}_{i=1}^\infty$ are eigenfunctions of the kernel K.

(v) $\{\mu_i^{\alpha/2} e_i\}_{i=1}^\infty$ is an orthogonal basis of the power space $\mathcal{H}^\alpha$.

(vi) $\mathcal{H}^\alpha = L_K^{\alpha/2}(\mathcal{L}^2_{\rho_{\mathcal{X}}}) = L_K^{(\alpha-1)/2}(\mathcal{H})$.

Moreover, for any $\alpha \in (0, 1]$, the α-power space is characterized by the interpolation spaces of the real method, namely, $\mathcal{H}^\alpha \cong [\mathcal{L}^2_{\rho_{\mathcal{X}}}, \mathcal{H}]_{\alpha,2}$ (Triebel, 1995; Steinwart and Scovel, 2012). Given this interpolation property, we present another example that satisfies Assumptions 4.13, 4.14, 4.15 following Fischer and Steinwart (2020).

Example B.3. (Besov RKHS). For $d \in \mathbb{N}$, let $\mathcal{X} \subset \mathbb{R}^d$ be a non-empty, open, connected, and bounded set with a C_∞ boundary, equipped with the Lebesgue-Borel σ-algebra and measure μ. Let $\mathcal{L}_2(\mathcal{X}) := \mathcal{L}_2(\mu)$ denote the corresponding L^2 space. For $m \in \mathbb{N}$, denote the Sobolev space with smoothness m by $\mathcal{W}_m(\mathcal{X}) := \mathcal{W}_{m,2}(\mathcal{X})$, and for $r > 0$ the Besov space $\mathcal{B}^r_{2,2}(\mathcal{X})$ is defined by means of the real interpolation $\mathcal{B}^r_{2,2}(\mathcal{X}) := [\mathcal{L}_2(\mathcal{X}), \mathcal{W}_m(\mathcal{X})]_{r,m/2}$, where $m = \min\{k \in \mathbb{N} : k > r\}$ Adams and Fournier (2003). By the theory of interpolation space, the Besov spaces $\mathcal{B}^r_{2,2}(\mathcal{X})$ are separable Hilbert space and satisfy

$$\mathcal{B}^r_{2,2}(\mathcal{X}) \cong [\mathcal{L}_2(\mathcal{X}), \mathcal{B}^t_{2,2}(\mathcal{X})]_{r/t,2}$$

for all $t > r > 0$. Define the Besov RKHS as

$$\mathcal{H}_r(\mathcal{X}) := \{f \in C_0(\mathcal{X}) : [f]_\mu \in \mathcal{B}^r_{2,2}(\mathcal{X}))\}$$

equipped with the norm $\|f\|_{\mathcal{H}_r(\mathcal{X})} := \|[f]_\mu\|_{\mathcal{B}^r_{2,2}(\mathcal{X})}$, where $[f]_\mu$ denotes the μ-equivalent class of f. Let each coordinate of the parametric function $f_i \in \mathcal{B}^s_{2,2}(\mathcal{X})$ for some $0 < s < r$, then the interpolation property of α-power space

$$\mathcal{H}^\alpha_r(\mathcal{X}) \cong [\mathcal{L}^2(\mathcal{X}), \mathcal{H}_r(\mathcal{X})]_{\alpha,2} \cong \mathcal{B}^{\alpha r}_{2,2}(\mathcal{X}) \tag{B.6}$$

implies that Assumption 4.13 holds for $\beta = s/r$. Moreover, the Sobolev embedding theorem for Besov Spaces indicates that the mapping of a μ-equivalence class to its continuous representative is linear and continuous (Triebel, 2010), i.e. for $r > j + d/2$, $\mathcal{B}^r_{2,2}(\mathcal{X})$ is continuously embedded into $C_j(\mathcal{X})$:

$$\mathcal{B}^r_{2,2}(\mathcal{X}) \to C_j(\mathcal{X}) \to \mathcal{L}^\infty(\mathcal{X}), \tag{B.7}$$

thus, Assumption 4.15 holds for $j \geq 1$. Combining (B.6) and (B.7), the embedding property (Assumption 4.14) holds for all $\alpha \in (\frac{d+2j}{2r}, 1]$, see Fischer and Steinwart (2020, Section 4) for more details.

C Proofs of Main Theorems

C.1 Proof of Theorem 4.6

We use Lemma B.2 to derive the one-step error bound. It is sufficient to check that the estimator for the gradient $H(x)$ satisfies Assumption B.1. Recalling the gradient step (4.2), the gradient estimator at iteration t is

$$h^t = (\nabla_i \ell_i^t(x^t, z_i^t) + (\partial f_i^t(x^t)/\partial x_i)^\top \nabla_{z_i} \ell_i^t(x^t, z_i^t))_{i\in[n]}. \tag{C.1}$$

Recalling (2.8), the true gradient is

$$H(x^t) = (\nabla_i \mathcal{L}_i(x^t))_{i\in[n]} = \left(\mathbb{E}_t \left[\nabla_i \ell_i(x^t, z_i^t) + \left(\partial f_i(x^t)/\partial x_i\right)^\top \nabla_{z_i} \ell_i(x^t, z_i^t)\right]\right)_{i\in[n]}. \tag{C.2}$$

Let us prove that the gradient estimator h^t satisfies the Assumption B.1. To do this, we compute the bias and variance terms, respectively.

Bias. Combining (C.1) and (C.2), we have

$$\begin{aligned}\|\mathbb{E}_t h^t - H(x^t)\|^2 \leq 2\sum_{i=1}^n \Bigg(& \underbrace{\|\mathbb{E}_t[\nabla_i(\ell_i^t - \ell_i)(x^t, z_i^t)]\|^2}_{\text{(I)}} + \underbrace{\|\mathbb{E}_t \left[(\partial f_i^t(x^t)/\partial x_i - \partial f_i(x^t)/\partial x_i)^\top \nabla_{z_i} \ell_i(x^t, z_i^t)\right]\|^2}_{\text{(II)}} \\ & + \underbrace{\|\mathbb{E}_t \left[(\partial f_i^t(x^t)/\partial x_i)^\top \nabla_{z_i}(\ell_i^t - \ell_i)(x^t, z_i^t)\right]\|^2}_{\text{(III)}} \Bigg).\end{aligned} \tag{C.3}$$

Recall (4.3), we obtain the bounds (I) $\lesssim \mathcal{O}(t^{-2a_2})$, (III) $\lesssim \mathcal{O}(t^{-2a_2})$, and the bound for (II):

$$\text{(II)} \leq \sum_{i=1}^n \|\partial f_i^t(x^t)/\partial x_i - \partial f_i(x^t)/\partial x_i\|_F^2 \|\mathbb{E}_t \nabla_{z_i} \ell_i(x^t, z_i^t)\|^2 \lesssim \mathcal{O}(t^{-2a_1}),$$

where we use the fact that $\|\mathbb{E}_t(\nabla_{z_i} \ell_i(x^t, z_i^t))_{i\in[n]}\| \leq \delta$ (following Assumption 4.3). Plugging these bounds into (C.3), we obtain

$$\|\mathbb{E}_t h^t - H(x^t)\| \leq \mathcal{O}(t^{-a_1 \wedge a_2}). \tag{C.4}$$

Comparing (C.4) with (B.3), we have $m_t = \mathcal{O}(t^{-a_1\wedge a_2})$ and $U = 0$.

Variance. Recalling (C.1), define $A^t := (\nabla_i \ell_i^t(x^t, z_i^t))_{i\in[n]}$ and $B^t := ((\partial f_i^t(x^t)/\partial x_i)^\top \nabla_{z_i} \ell_i^t(x^t, z_i^t))_{i\in[n]}$, then $h^t = A^t + B^t$. We compute the variance of h^t

$$\begin{aligned}\mathbb{E}_t\|h^t - \mathbb{E}_t h^t\|^2 &= \mathbb{E}_t\|(A^t - \mathbb{E}_t A^t) + (B^t - \mathbb{E}_t B^t)\|^2 \\ &\le 2\left(\mathbb{E}_t\|A^t - \mathbb{E}_t A^t\|^2 + \mathbb{E}_t\|B^t - \mathbb{E}_t B^t\|^2\right).\end{aligned} \tag{C.5}$$

Now we derive the upper bounds for last the two terms of (C.5), respectively.

Upper bound of $\mathbb{E}_t\|A^t - \mathbb{E}_t A^t\|^2$. By the definition of A^t,

$$\begin{aligned}\mathbb{E}_t\|A^t - \mathbb{E}_t A^t\|^2 &\lesssim \mathbb{E}_t\|(\nabla_i(\ell_i - \ell_i^t)(x^t, z_i^t))_{i\in[n]}\|^2 + \mathbb{E}_t\|(\nabla_i \ell_i(x^t, z_i^t) - \mathbb{E}_t \nabla_i \ell_i(x^t, z_i^t))_{i\in[n]}\|^2 \\ &\quad + \|\mathbb{E}_t(\nabla_i(\ell_i - \ell_i^t)(x^t, z_i^t))_{i\in[n]}\|^2 \\ &\lesssim \mathcal{O}(t^{-2a_2}) + \zeta^2 \lesssim \mathcal{O}(1),\end{aligned} \tag{C.6}$$

where we use the fact $\mathbb{E}_t\|(\nabla_i \ell_i(x^t, z_i^t) - \mathbb{E}_t \nabla_i \ell_i(x^t, z_i^t))_{i\in[n]}\|^2 \le \zeta^2$ (following Assumption 4.4).

Upper bound of $\mathbb{E}_t\|B^t - \mathbb{E}_t B^t\|^2$. We similar decomposition, we have

$$\begin{aligned}\mathbb{E}_t\|B^t - \mathbb{E}_t B^t\|^2 &\lesssim \mathcal{O}(t^{-2a_1}) + \mathbb{E}_t\|((\partial f_i^t(x^t)/\partial x_i)^\top \left[\nabla_{z_i} \ell_i(x^t, z_i^t) - \mathbb{E}_t \nabla_{z_i} \ell_i(x^t, z_i^t)\right])_{i\in[n]}\|^2 \\ &\lesssim \mathcal{O}(t^{-2a_1}) + \sup_{i\in[n]} \|\partial f_i^t(x^t)/\partial x_i\|_F^2 \mathbb{E}_t\|(\left[\nabla_{z_i} \ell_i(x^t, z_i^t) - \mathbb{E}_t \nabla_{z_i} \ell_i(x^t, z_i^t)\right])_{i\in[n]}\|^2.\end{aligned}$$

Again, Assumption 4.4 implies that

$$\mathbb{E}_t\|(\left[\nabla_{z_i} \ell_i(x^t, z_i^t) - \mathbb{E}_t \nabla_{z_i} \ell_i(x^t, z_i^t)\right])_{i\in[n]}\|^2 \le \zeta^2.$$

Therefore,

$$\mathbb{E}_t\|B^t - \mathbb{E}_t B^t\|^2 \lesssim \mathcal{O}(1), \tag{C.7}$$

Combining (C.6) and (C.7), we have

$$\mathbb{E}_t\|h^t - \mathbb{E}_t h^t\|^2 \lesssim \mathcal{O}(1). \tag{C.8}$$

Comparing (C.8) with (B.3), we have $\sigma_t^2 = \mathcal{O}(1)$ and $V = 0$. Using Lemma B.2, we obtain the one-step error

$$\begin{aligned}\mathbb{E}_t\|x^{t+1} - x^*\|^2 &\lesssim \frac{1}{1+\eta_t\tau}\|x^t - x^*\|^2 + \eta_t^2 \cdot \mathcal{O}(1) + \eta_t \cdot \mathcal{O}(t^{-2(a_1\wedge a_2)}) \\ &\lesssim \frac{1}{1+\eta_t\tau}\|x^t - x^*\|^2 + \mathcal{O}(t^{-1-(1\wedge 2a_1\wedge 2a_2)}).\end{aligned} \tag{C.9}$$

Using Lemma G.1, we have

$$\mathbb{E}_t\|x^{t+1} - x^*\|^2 \lesssim \mathcal{O}(t^{-(1\wedge 2a_1\wedge 2a_2)}).$$

C.2 Proof of Theorem 4.9

Recalling Lemma 4.11. Taking expectation on both sides of (4.6), we have

$$\mathbb{E}\|x^{t+1}-x^*\|^2 \leq \frac{1}{1+\eta_t\tau}\mathbb{E}\|x^t-x^*\|^2+\frac{4\eta_t^2\zeta^2(1+\mathbb{E}\left[\sup_{i\in[n]}\|A_i^t\|_F^2\right])}{1+\eta_t\tau}+\frac{2\eta_t\delta^2\mathbb{E}\left[\sup_{i\in[n]}\|A_i^t-A_i\|_F^2\right]}{\tau(1+\eta_t\tau)}, \tag{C.10}$$

where $\eta_t = 2/(\tau(t+t_0))$ denotes the gradient step size. Let us consider the two terms involving A_i^t in the RHS of (C.10), i.e. $\mathbb{E}\left[\sup_{i\in[n]}\|A_i^t-A_i\|_F^2\right]$ and $\mathbb{E}\left[\sup_{i\in[n]}\|A_i^t\|_F^2\right]$.

Upper bound of $\mathbb{E}\left[\sup_{i\in[n]}\|A_i^t-A_i\|_F^2\right]$. Using Lemma 4.10, we have

$$\mathbb{E}\left[\sup_{i\in[n]}\|A_i^t-A_i\|_F^2\right] \leq \mathbb{E}\left[\sum_{i=1}^n\|A_i^t-A_i\|_F^2\right] \leq \frac{M}{t+t_0}, \tag{C.11}$$

where $M := 2t_0^4\sum_{i=1}^n\|A_i^0-A_i\|_F^2/(t_0+1)^3 + 8nl_2\sigma^2(t_0+2)^2/(l_1^2(t_0+1)^2)$ is a constant.

Upper bound of $\mathbb{E}\left[\sup_{i\in[n]}\|A_i^t\|_F^2\right]$. Using (C.11), we have

$$\begin{aligned}\mathbb{E}\left[\sup_{i\in[n]}\|A_i^t\|_F^2\right] &\leq 2\mathbb{E}\left[\sup_{i\in[n]}\|A_i^t-A_i\|_F^2+\sup_{i\in[n]}\|A_i\|_F^2\right]\\ &\leq 2(M(t+t_0)^{-1}+\sup_{i\in[n]}\|A_i\|_F^2).\end{aligned} \tag{C.12}$$

Plugging (C.11) and (C.12) into (C.10), we obtain the upper bound for the RHS of (C.10), i.e.

$$\mathbb{E}\|x^{t+1}-x^*\|^2 \leq \frac{1}{1+\eta_t\tau}\mathbb{E}\|x^t-x^*\|^2+\frac{4\eta_t^2\zeta^2(1+2(M/(t+t_0)+\sup_{i\in[n]}\|A_i\|_F^2))}{1+\eta_t\tau}+\frac{2\eta_t\delta^2M/(t+t_0)}{\tau(1+\eta_t\tau)}.$$

Define constants D_1 and D_2 as follow

$$D_1 := 4\zeta^2(1+2(M/(t+t_0)+\sup_{i\in[n]}\|A_i\|_F^2)), \qquad D_2 := 2\delta^2M,$$

then we have

$$\begin{aligned}\mathbb{E}\|x^{t+1}-x^*\|^2 &\leq \frac{1}{1+\eta_t\tau}\mathbb{E}\|x^t-x^*\|^2+\frac{D_1\eta_t^2}{1+\eta_t\tau}+\frac{D_2\eta_t/t}{1+\eta_t\tau}\\ &\leq (1-\eta_t\tau)\mathbb{E}\|x^t-x^*\|^2+D_1\eta_t^2+D_2\eta_t t^{-1}.\end{aligned} \tag{C.13}$$

Since $\eta_t = 2/(\tau(t+t_0))$, (C.13) implies that

$$\begin{aligned}\mathbb{E}\|x^{t+1}-x^*\|^2 &\leq \left(1-\frac{2}{t+t_0}\right)\mathbb{E}\|x^t-x^*\|^2+\frac{4D_1}{\tau^2(t+t_0)^2}+\frac{2D_2(t_0+1)}{\tau(t+t_0)^2}\\ &= \left(1-\frac{2}{t+t_0}\right)\mathbb{E}\|x^t-x^*\|^2+\frac{4D_1/\tau^2+2D_2(t_0+1)/\tau}{(t+t_0)^2}\\ &= \prod_{i=1}^t\left(1-\frac{2}{i+t_0}\right)\|x^1-x^*\|^2+\sum_{i=1}^t\frac{4D_1/\tau^2+2D_2(t_0+1)/\tau}{(i+t_0)^2}\prod_{j=i+1}^t\left(1-\frac{2}{j+t_0}\right).\end{aligned} \tag{C.14}$$

Moreover, note that

$$\prod_{j=s}^{t}\left(1-\frac{2}{j+t_0}\right) \le e^{-2\sum_{j=s}^{t}(j+t_0)^{-1}} \le e^{-2\sum_{j=s}^{t}\int_j^{j+1}(x+t_0)^{-1}dx} = \left(\frac{s+t_0}{t+1+t_0}\right)^2. \tag{C.15}$$

Plugging (C.15) into (C.14), we finish the proof

$$\begin{aligned}\mathbb{E}\|x^{t+1}-x^*\|^2 &\le \left(\frac{t_0+1}{t+1+t_0}\right)^2\|x^1-x^*\|^2 + \sum_{i=1}^{t}\frac{4D_1/\tau^2+2D_2(t_0+1)/\tau}{(i+t_0)^2}\left(\frac{i+1+t_0}{t+1+t_0}\right)^2 \\ &\le \left(\frac{t_0+1}{t+1+t_0}\right)^2\|x^1-x^*\|^2 + \frac{(4D_1/\tau^2+2D_2(t_0+1)/\tau)(t_0+2)^2/(t_0+1)^2}{t+1+t_0}.\end{aligned}$$

C.3 Proof of Theorem 4.19

We prove the following generalized version of Theorem 4.19.

Theorem 4.19′ . Suppose that all the assumptions of Theorem 4.19 hold. For all iterations t and positive constant a, define $\bar{t} = t + t_0$, where t_0 is a constant satisfies $t_0 \ge (a\kappa^2+1)^2$. Set the gradient steps and regularization terms as

$$\eta_t = \frac{1}{\tau\bar{t}}, \quad \nu_t = a\left(\frac{1}{\bar{t}}\right)^{\frac{\beta-\gamma+1}{\beta-\gamma+2}}, \quad \lambda_t = \frac{1}{a}\left(\frac{1}{\bar{t}}\right)^{\frac{1}{\beta-\gamma+2}}, \tag{C.16}$$

where γ takes values in $[\alpha,\beta)$ and $\gamma \le 1$. If $a < \sqrt{(\beta-\gamma+2)/(\beta-\gamma)}(t_0+1)/(t_0+2)\kappa^{\gamma-2}A^{-1}$, the x^t generated by the `OPGD` algorithm in Section 3 for kernel function class (Algorithm 2) with input kernel K satisfies

$$\mathbb{E}\|x^{t+1}-x^*\|^2 \lesssim \mathcal{O}(t^{-\frac{\beta-\gamma}{\beta-\gamma+2}}). \tag{C.17}$$

We use Lemmas 4.18 and C.1 to derive the convergence rate of x^t. See Appendix E for the proof of these claims.

Lemma C.1. (One-step error). Suppose that Assumptions 4.1, 4.2, 4.3, 4.4 hold and all the assumptions in Lemma 4.18 hold. Let $\mathcal{G} = \{\mathcal{G}_t\}_{t\in\mathbb{N}}$ be the filtration $\mathcal{G}_t = \sigma\{\{x^j\}_{j\in[T]} \cup (u_i^t, y_i^t)\}$ and define $\mathbb{E}_t[\cdot] = E[\cdot|\mathcal{G}_t]$. For any gradient steps $\eta_t \le \tau/(4L^2)$, the iterates generated by Algorithm 2 satisfies

$$\mathbb{E}_t\|x^{t+1}-x^*\|^2 \le \frac{1}{1+\eta_t\tau}\|x^t-x^*\|^2 + \frac{4\eta_t^2\zeta^2(1+\xi^2\sup_{i\in[n]}\|f_i^t\|_\gamma^2)}{1+\eta_t\tau} + \frac{2\eta_t\xi^2\delta^2\sup_{i\in[n]}\|f_i^t-f_i\|_\gamma^2}{\tau(1+\eta_t\tau)}, \tag{C.18}$$

where L is the Lipschitz constant of individual gradient $H(x)$ defined in Section 2, δ is defined in Assumption 4.3, and ζ is defined in Assumption 4.4.

To begin with, using Lemma C.1 and taking expectation on both sides of (C.18), we have

$$\mathbb{E}\|x^{t+1}-x^*\|^2 \le \frac{1}{1+\eta_t\tau}\mathbb{E}\|x^t-x^*\|^2 + \frac{4\eta_t^2\zeta^2(1+\xi^2\mathbb{E}\left[\sup_{i\in[n]}\|f_i^t\|_\gamma^2\right])}{1+\eta_t\tau} + \frac{2\eta_t\xi^2\delta^2\mathbb{E}\left[\sup_{i\in[n]}\|f_i^t-f_i\|_\gamma^2\right]}{\tau(1+\eta_t\tau)}, \tag{C.19}$$

where $\gamma \in [\alpha,\beta)$ and $\eta_t = 1/(\tau(t+t_0))$ is the gradient step size. Let us consider the two terms involving f_i^t in the RHS of (C.19), namely, $\mathbb{E}\left[\sup_{i\in[n]}\|f_i^t-f_i\|_\gamma^2\right]$ and $\mathbb{E}\left[\sup_{i\in[n]}\|f_i^t\|_\gamma^2\right]$.

Upper bound of $\mathbb{E}\left[\sup_{i\in[n]}\|f_i^t - f_i\|_\gamma^2\right]$. Lemma 4.18 implies that for any $\gamma \in [\alpha, \beta)$ and $i \in [n]$, if we set the parameters as (C.16), then the estimation error $\mathbb{E}\|f_i^t - f_i\|_\gamma$ is bounded by $\mathcal{O}(t^{-(\beta-\gamma)/(\beta-\gamma+2)})$, i.e. there exist constants $\{M_i(\gamma)\}_{i\in[n]}$ such that

$$\mathbb{E}\|f_i^t - f_i\|_\gamma \le M_i(\gamma) t^{-\frac{\beta-\gamma}{\beta-\gamma+2}},$$

where the constant $M_i(\gamma)$ only depends on γ. Define $M(\gamma) \coloneqq \sum_{i=1}^n M_i(\gamma)$, we have

$$\mathbb{E}\left[\sup_{i\in[n]}\|f_i^t - f_i\|_\gamma^2\right] \le \mathbb{E}\left[\sum_{i=1}^n \|f_i^t - f_i\|_\gamma^2\right] \le M(\gamma) t^{-\frac{\beta-\gamma}{\beta-\gamma+2}}. \tag{C.20}$$

Upper bound of $\mathbb{E}\left[\sup_{i\in[n]}\|f_i^t\|_\gamma^2\right]$. Using Lemma 4.18 again, we have

$$\begin{aligned}\mathbb{E}\left[\sup_{i\in[n]}\|f_i^t\|_\gamma^2\right] &\le 2\mathbb{E}\left[\sup_{i\in[n]}\|f_i^t - f_i\|_\gamma^2 + \sup_{i\in[n]}\|f_i\|_\gamma^2\right] \\ &\overset{(a)}{\le} 2(M(\gamma)t^{-\frac{\beta-\gamma}{\beta-\gamma+2}} + \sup_{i\in[n]}\|f_i\|_\gamma^2).\end{aligned} \tag{C.21}$$

where (a) uses (C.20).

Plugging (C.20) and (C.21) into (C.19), we obtain the upper bound for the RHS of (C.19), i.e.

$$\begin{aligned}\mathbb{E}\|x^{t+1} - x^*\|^2 \le& \frac{1}{1+\eta_t\tau}\mathbb{E}\|x^t - x^*\|^2 + \frac{4\eta_t^2\zeta^2(1 + 2\xi^2(M(\gamma)t^{-\frac{\beta-\gamma}{\beta-\gamma+2}} + \sup_{i\in[n]}\|f_i\|_\gamma^2))}{1+\eta_t\tau} \\ &+ \frac{2\eta_t\xi^2\delta^2 M(\gamma)t^{-\frac{\beta-\gamma}{\beta-\gamma+2}}}{\tau(1+\eta_t\tau)}.\end{aligned}$$

Define constants D_1 and D_2

$$D_1 \coloneqq 4\zeta^2(1 + 2\xi^2(M(\gamma) + \sup_{i\in[n]}\|f_i\|_\gamma^2)), \qquad D_2 \coloneqq 2\xi^2\delta^2 M(\gamma)/\tau,$$

then we have

$$\begin{aligned}\mathbb{E}\|x^{t+1} - x^*\|^2 &\le \frac{1}{1+\eta_t\tau}\mathbb{E}\|x^t - x^*\|^2 + \frac{D_1\eta_t^2}{1+\eta_t\tau} + \frac{D_2\eta_t t^{-\frac{\beta-\gamma}{\beta-\gamma+2}}}{1+\eta_t\tau} \\ &\le (1-\eta_t\tau)\mathbb{E}\|x^t - x^*\|^2 + D_1\eta_t^2 + D_2\eta_t t^{-\frac{\beta-\gamma}{\beta-\gamma+2}}.\end{aligned}$$

Note that $\eta_t = 1/(\tau(t+t_0))$, therefore

$$\mathbb{E}\|x^{t+1} - x^*\|^2 \le \left(1 - \frac{1}{t+t_0}\right)\mathbb{E}\|x^t - x^*\|^2 + \frac{D_1}{\tau^2(t+t_0)^2} + \frac{D_2(t_0+1)^{\frac{\beta-\gamma}{\beta-\gamma+2}}}{\tau(t+t_0)^{1+\frac{\beta-\gamma}{\beta-\gamma+2}}}.$$

Moreover, since $\beta - \gamma < 2$, we have $2 > 1 + (\beta-\gamma)/(\beta-\gamma+2)$. As a result,

$$\mathbb{E}\|x^{t+1} - x^*\|^2 \le \left(1 - \frac{1}{t+t_0}\right)\mathbb{E}\|x^t - x^*\|^2 + \frac{D_1/\tau^2 + D_2/\tau(t_0+1)^{(\beta-\gamma)/(\beta-\gamma+2)}}{(t+t_0)^{1+\frac{\beta-\gamma}{\beta-\gamma+2}}}. \tag{C.22}$$

Using Lemma G.1, comparing (C.22) with (G.1), we have $a = 1+\frac{\beta-\gamma}{\beta-\gamma+2}$ and $b = D_1/\tau^2+D_2/\tau(t_0+1)^{(\beta-\gamma)/(\beta-\gamma+2)}$. Therefore, by Lemma G.1, we have

$$\mathbb{E}\|x^t - x^*\|^2 \lesssim \mathcal{O}(t^{-\frac{\beta-\gamma}{\beta-\gamma+2}}).$$

Note that $\gamma \in [\alpha, \beta)$, to get the best convergence rate, we set $\gamma = \alpha$, then we obtain Theorem 4.19.

C.4 Proof of Proposition 4.5

Recalling Definition 2.3, we are going to prove that: (i) $\mathcal{L}_i(x)$ are convex in x_i when x_{-i} are fixed for all $i \in [n]$; (ii) the gradient $H(x)$ is a strongly monotone map with respect to x.

Convexity of $\mathcal{L}_i(x)$ in x_i. Let us prove the following inequality, which is enough to show the convexity:

$$\langle \nabla_i \mathcal{L}_i(x_i, x_{-i}) - \nabla_i \mathcal{L}_i(x_i', x_{-i}), x_i - x_i' \rangle \geq 0, \quad \forall x_i, x_i' \in \mathcal{X}_i, x_{-i} \in \mathcal{X}_{-i}.$$

Recalling (2.6), we have

$$\nabla_i \mathcal{L}_i(x_i, x_{-i}) = P_i(x_i, x_{-i}) + Q_i(x_i, x_{-i}), \tag{C.23}$$

where

$$P_i(x_i, x_{-i}) = \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(x_i, x_{-i})} \nabla_i \ell_i(x_i, x_{-i}, z_i) \quad \text{and} \quad Q_i(x_i, x_{-i}) = \frac{d}{du_i} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(u_i, x_{-i})} \ell_i(x_i, x_{-i}, z_i)\Big|_{u_i = x_i}.$$

Thus, the difference between individual gradients at (x_i, x_{-i}) and (x_i', x_{-i}) is

$$\nabla_i \mathcal{L}_i(x_i, x_{-i}) - \nabla_i \mathcal{L}_i(x_i', x_{-i}) = \underbrace{P_i(x_i, x_{-i}) - P_i(x_i', x_{-i})}_{\text{(I)}} + \underbrace{Q_i(x_i, x_{-i}) - Q_i(x_i', x_{-i})}_{\text{(II)}}. \tag{C.24}$$

- **Analysis of (I).**

Let us derive the lower bound of $\langle P_i(x_i, x_{-i}) - P_i(x_i', x_{-i}), x_i - x_i' \rangle$. Under the parametric assumption 2.6, we have the following decomposition

$$\begin{aligned}
P_i(x_i, x_{-i}) - P_i(x_i', x_{-i}) &= \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(x_i, x_{-i})} \nabla_i \ell_i(x_i, x_{-i}, z_i) - \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(x_i', x_{-i})} \nabla_i \ell_i(x_i', x_{-i}, z_i) \\
&= \mathop{\mathbb{E}}_{\epsilon_i \sim \mathcal{P}_i} \nabla_i \ell_i(x_i, x_{-i}, f_i(x_i, x_{-i}) + \epsilon_i) - \mathop{\mathbb{E}}_{\epsilon_i \sim \mathcal{P}_i} \nabla_i \ell_i(x_i', x_{-i}, f_i(x_i', x_{-i}) + \epsilon_i) \\
&= \mathop{\mathbb{E}}_{\epsilon_i \sim \mathcal{P}_i} \left[\nabla_i \ell_i(x_i, x_{-i}, f_i(x_i, x_{-i}) + \epsilon_i) - \nabla_i \ell_i(x_i', x_{-i}, f_i(x_i', x_{-i}) + \epsilon_i)\right].
\end{aligned}$$

Thus, we can further decompose the above equation as

$$\begin{aligned}
P_i(x_i, x_{-i}) - P_i(x_i', x_{-i}) = \mathop{\mathbb{E}}_{\epsilon_i \sim \mathcal{P}_i} \big[&\nabla_i \ell_i(x_i, x_{-i}, f_i(x_i, x_{-i}) + \epsilon_i) - \nabla_i \ell_i(x_i', x_{-i}, f_i(x_i, x_{-i}) + \epsilon_i) \\
&\nabla_i \ell_i(x_i', x_{-i}, f_i(x_i, x_{-i}) + \epsilon_i) - \nabla_i \ell_i(x_i', x_{-i}, f_i(x_i', x_{-i}) + \epsilon_i)\big].
\end{aligned} \tag{C.25}$$

We analyze two rows in RHS of (C.25), respectively. Since the game (4.1) is S-strongly monotone for $y = (x_i, x_{-i})$, then for the first row of (C.25), we have

$$\mathop{\mathbb{E}}_{\epsilon_i \sim \mathcal{P}_i} \langle \nabla_i \ell_i(x_i, x_{-i}, f_i(x_i, x_{-i}) + \epsilon_i) - \nabla_i \ell_i(x_i', x_{-i}, f_i(x_i, x_{-i}) + \epsilon_i), x_i - x_i' \rangle \geq S \|x_i - x_i'\|^2. \quad \text{(C.26)}$$

Also, using the R_i-Lipschitz continuity of $\nabla_i \ell_i(x_i, x_{-i}, z_i)$ with respect to z_i and the L_i-Lipschitz continuity of $f_i(x_i, x_{-i})$ with respect to (x_i, x_{-i}). For the second row of (C.25), we obtain

$$\begin{aligned} &\langle \nabla_i \ell_i(x_i', x_{-i}, f_i(x_i, x_{-i}) + \epsilon_i) - \nabla_i \ell_i(x_i', x_{-i}, f_i(x_i', x_{-i}) + \epsilon_i), x_i - x_i' \rangle \\ &\qquad \geq -R_i \|f_i(x_i, x_{-i}) - f_i(x_i', x_{-i})\| \|x_i - x_i'\| \geq -L_i R_i \|x_i - x_i'\|^2. \end{aligned} \quad \text{(C.27)}$$

Plugging (C.26) and (C.27) into (C.25), we have

$$\langle P_i(x_i, x_{-i}) - P_i(x_i', x_{-i}), x_i - x_i' \rangle \geq (S - L_i R_i) \|x_i - x_i'\|^2. \quad \text{(C.28)}$$

- **Analysis of (II).**

Let us derive the lower bound of $\langle Q_i(x_i, x_{-i}) - Q_i(x_i', x_{-i}), x_i - x_i' \rangle$. The same as (I), we propose the following decomposition:

$$\begin{aligned} Q_i(x_i, x_{-i}) - Q_i(x_i', x_{-i}) = &\frac{d}{du_i} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(u_i, x_{-i})} \ell_i(x_i, x_{-i}, z_i)\Big|_{u_i = x_i} - \frac{d}{du_i} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(u_i, x_{-i})} \ell_i(x_i', x_{-i}, z_i)\Big|_{u_i = x_i} \\ &+ \frac{d}{du_i} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(u_i, x_{-i})} \ell_i(x_i', x_{-i}, z_i)\Big|_{u_i = x_i} - \frac{d}{du_i} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(u_i, x_{-i})} \ell_i(x_i', x_{-i}, z_i)\Big|_{u_i = x_i'}. \end{aligned} \quad \text{(C.29)}$$

We analyze two rows in RHS of (C.29), respectively. For the first row of (C.29), we have

$$\begin{aligned} &\langle \frac{d}{du_i} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(u_i, x_{-i})} \ell_i(x_i, x_{-i}, z_i)\Big|_{u_i = x_i} - \frac{d}{du_i} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(u_i, x_{-i})} \ell_i(x_i', x_{-i}, z_i)\Big|_{u_i = x_i}, x_i - x_i' \rangle \\ &= \langle \frac{d}{du_i} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(u_i, x_{-i})} \left[\ell_i(x_i, x_{-i}, z_i) - \ell_i(x_i', x_{-i}, z_i)\right] \Big|_{u_i = x_i}, x_i - x_i' \rangle \end{aligned} \quad \text{(C.30)}$$

Using the fundamental theorem of calculus, we have

$$\ell_i(x_i, x_{-i}, z_i) - \ell_i(x_i', x_{-i}, z_i) = \int_0^1 \langle \nabla_i \ell_i(x_i' + s(x_i - x_i'), x_{-i}, z_i), x_i - x_i' \rangle ds.$$

Plugging this into (C.30) and using Cauchy–Schwarz inequality, we obtain

$$\begin{aligned} &\langle \frac{d}{du_i} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(u_i, x_{-i})} \ell_i(x_i, x_{-i}, z_i)\Big|_{u_i = x_i} - \frac{d}{du_i} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(u_i, x_{-i})} \ell_i(x_i', x_{-i}, z_i)\Big|_{u_i = x_i}, x_i - x_i' \rangle \\ &\geq -\|\frac{d}{du_i} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(u_i, x_{-i})} \left[\ell_i(x_i, x_{-i}, z_i) - \ell_i(x_i', x_{-i}, z_i)\right] \Big|_{u_i = x_i} \| \|x_i - x_i'\| \\ &\overset{(a)}{=} -\int_0^1 \|\frac{d}{du_i} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(u_i, x_{-i})} \nabla_i \ell_i(x_i' + s(x_i - x_i'), x_{-i}, z_i)\Big|_{u_i = x_i} \| ds * \|x_i - x_i'\|^2, \end{aligned} \quad \text{(C.31)}$$

here (a) holds because the integral $\int_0^1$ is exchangeable with the expectation and differential operators and further induced by Jensen's inequality. Now we consider the differentiable map $u \to \mathbb{E}_{z_i \sim \mathcal{D}_i(u, x_{-i})} \nabla_i \ell_i(x, z_i)$, it is easy to check that this map is $L_i R_i$-Lipschitz:

$$\| \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(u, x_{-i})} \nabla_i \ell_i(x, z_i) - \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(u', x_{-i})} \nabla_i \ell_i(x, z_i) \| \leq L_i R_i \|u - u'\|, \quad \forall x \in \mathcal{X}.$$

Therefore, the gradient of this map is bounded by $L_i R_i$. Plugging this into (C.31), we have

$$\langle \frac{d}{du_i} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(u_i, x_{-i})} \left[\ell_i(x_i, x_{-i}, z_i) - \ell_i(x_i', x_{-i}, z_i) \right] \Big|_{u_i = x_i}, x_i - x_i' \rangle \geq -L_i R_i \|x_i - x_i'\|^2. \tag{C.32}$$

For the second row of (C.29), since the map $u \to \mathbb{E}_{z_i \sim \mathcal{D}_i(u, z_{-i})} \ell_i(x, z_i)$ is monotone (by the assumption (ii) in Proposition 4.5), therefore,

$$\langle \frac{d}{du_i} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(u_i, x_{-i})} \ell_i(x_i', x_{-i}, z_i) \Big|_{u_i = x_i} - \frac{d}{du_i} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(u_i, x_{-i})} \ell_i(x_i', x_{-i}, z_i) \Big|_{u_i = x_i'}, x_i - x_i' \rangle \geq 0. \tag{C.33}$$

Plugging (C.32) and (C.33) into (C.29), we have

$$\langle Q_i(x_i, x_{-i}) - Q_i(x_i', x_{-i}), x_i - x_i' \rangle \geq -L_i R_i \|x_i - x_i'\|^2. \tag{C.34}$$

Combining (C.28) and (C.34), we obtain the convexity of $\mathcal{L}_i(x)$ in x_i:

$$\langle \nabla_i \mathcal{L}_i(x_i, x_{-i}) - \nabla_i \mathcal{L}_i(x_i', x_{-i}), x_i - x_i' \rangle \geq (S - 2L_i R_i) \|x_i - x_i'\|^2 \geq 0.$$

Strong monotonicity of $H(x)$. The steps of proofs for $H(x)$ are exactly the same as the proof of the individual gradient $\nabla_i \mathcal{L}_i(x)$. Therefore, we omit the proof here. One can use the same steps and technic the check that $H(x)$ is $S - 2\sqrt{\sum_{i=1}^n (L_i R_i)^2}$-strongly monotone, namely,

$$\langle H(x) - H(x'), x - x' \rangle \geq \left(S - 2\sqrt{\sum_{i=1}^n (L_i R_i)^2} \right) \|x_i - x_i'\|^2.$$

D Proofs of Auxiliary Lemmas for Theorem 4.9

D.1 Proof of Lemma 4.10

Recalling the online estimation step for A_i^t in Algorithm 1. For any $i \in [n], t \in \mathbb{N}$,

$$A_i^t = A_i^{t-1} - \nu_t \left(A_i^{t-1} u_i^t - y_i^t \right) (u_i^t)^\top, \tag{D.1}$$

where ν_t is the step size, u_i^t, y_i^t are random variables obtained from the environment following the distribution $u_i^t \sim \rho_{\mathcal{X}}, y_i^t \sim \mathcal{D}_i(u_i^t)$. Note that (D.1) can be rewritten as the one-step decomposition

$$\begin{aligned} A_i^t - A_i &= A_i^{t-1} - A_i - \nu_t \left(A_i^{t-1} u_i^t - y_i^t \right) (u_i^t)^\top \\ &= (A_i^{t-1} - A_i)(I - \nu_t u_i^t {u_i^t}^\top) - \nu_t (A_i u_i^t - y_i^t)(u_i^t)^\top. \end{aligned}$$

Using this one-step error decomposition recursively, we obtain

$$A_i^t - A_i = (A_i^0 - A_i)\prod_{j=1}^{t}(I - \nu_j u_i^j {u_i^j}^\top) - \sum_{j=1}^{t}\nu_j(A_i u_i^j - y_i^j)(u_i^j)^\top \prod_{k=j+1}^{t}(I - \nu_k u_i^k {u_i^k}^\top). \tag{D.2}$$

We remark that $\prod_{j=1}^{t}(I - \nu_j u_i^j {u_i^j}^\top) = (I - \nu_1 u_i^1 {u_i^1}^\top)(I - \nu_2 u_i^2 {u_i^2}^\top)\cdots(I - \nu_t u_i^t {u_i^t}^\top)$. Notably, for any distinct a and b, the matrices $I - \nu_a u_i^a {u_i^a}^\top$ and $I - \nu_b u_i^b {u_i^b}^\top$ are not commutative.

Now we use (D.2) to derive the upper bound of $\mathbb{E}\|A_i^t - A_i\|_F^2$. By Cauchy's inequality, we have

$$\mathbb{E}\|A_i^t - A_i\|_F^2 \le 2(\underbrace{\mathbb{E}\|(A_i^0 - A_i)\prod_{j=1}^{t}(I - \nu_j u_i^j {u_i^j}^\top)\|_F^2}_{\text{(I)}} + \underbrace{\mathbb{E}\|\sum_{j=1}^{t}\nu_j(A_i u_i^j - y_i^j)(u_i^j)^\top \prod_{k=j+1}^{t}(I - \nu_k u_i^k {u_i^k}^\top)\|_F^2}_{\text{(II)}}). \tag{D.3}$$

- **Upper bound of (I).**

By the definition of matrix operator norm, $\|A(I - \nu_j u_i^j {u_i^j}^\top)\|_F \le \|A\|_F\|(I - \nu_j u_i^j {u_i^j}^\top)\|_{\text{op}}$, thus

$$\mathbb{E}\|(A_i^0 - A_i)\prod_{j=1}^{t}(I - \nu_j u_i^j {u_i^j}^\top)\|_F^2 \le \mathbb{E}\prod_{j=1}^{t}\|I - \nu_j u_i^j {u_i^j}^\top\|_{\text{op}}^2\|A_i^0 - A_i\|_F^2. \tag{D.4}$$

For any $i \in [n]$ and $j \in \mathbb{N}$, given that u_i^j are independent and identically distributed (i.i.d.) random variables drawn from the distribution $\rho_{\mathcal{X}}$, we can interchange the order of expectation and multiplication

$$\mathbb{E}\prod_{j=1}^{t}\|I - \nu_j u_i^j {u_i^j}^\top\|_{\text{op}}^2 = \prod_{j=1}^{t}\mathbb{E}\|I - \nu_j u_i^j {u_i^j}^\top\|_{\text{op}}^2. \tag{D.5}$$

Therefore, it is sufficient to derive the upper bound for $\mathbb{E}\|I - \nu_j u_i^j {u_i^j}^\top\|_{\text{op}}$

$$\begin{aligned}\|I - \nu_j u_i^j {u_i^j}^\top\|_{\text{op}}^2 &= \sup_{x\in\mathbb{R}^d}\frac{\|(I - \nu_j u_i^j {u_i^j}^\top)x\|^2}{\|x\|^2} \\ &= \sup_{x\in\mathbb{R}^d}\frac{\|x\|^2 + \nu_j^2\|u_i^j {u_i^j}^\top x\|^2 - 2\nu_j\langle x, u_i^j {u_i^j}^\top x\rangle}{\|x\|^2}.\end{aligned} \tag{D.6}$$

Moreover, the matrix inner product can be rewritten as

$$\langle x, u_i^j {u_i^j}^\top x\rangle = \text{tr}(x^\top u_i^j {u_i^j}^\top x), \qquad \|u_i^j {u_i^j}^\top x\|_F^2 = \text{tr}(\|u_i^j\|^2 x^\top u_i^j {u_i^j}^\top x).$$

Therefore, we can (D.6) has the following epxression

$$\mathbb{E}\|I - \nu_j u_i^j {u_i^j}^\top\|_{\text{op}}^2 = \sup_{x\in\mathbb{R}^d}\frac{\|x\|^2 + \nu_j^2\text{tr}(\mathbb{E}\|u_i^j\|^2 x^\top u_i^j {u_i^j}^\top x) - 2\nu_j\text{tr}(x^\top\mathbb{E}u_i^j {u_i^j}^\top x)}{\|x\|^2}. \tag{D.7}$$

Recalling Assumption 4.8, we have

$$\mathrm{tr}(\mathbb{E}\|u_i^j\|^2 x^\top u_i^j {u_i^j}^\top x) \le R * \mathrm{tr}(x^\top \mathbb{E} u_i^j {u_i^j}^\top x) \le l_2 R\|x\|^2, \quad \mathrm{tr}(x^\top \mathbb{E} u_i^j {u_i^j}^\top x) \ge l_1\|x\|^2.$$

Plugging these inequalities into (D.7), we obtain

$$\begin{aligned} \mathbb{E}\|I - \nu_j u_i^j {u_i^j}^\top\|_{\mathrm{op}}^2 &\le 1 - \nu_j(2l_1 - \nu_j l_2 R) \le 1 - \nu_j(2l_1 - \nu_0 l_2 R) \\ &\overset{(a)}{\le} 1 - l_1\nu_j, \end{aligned} \tag{D.8}$$

where (a) uses the fact that $\nu_t = 2/(l_1(t+t_0))$ and $t_0 \ge 2l_2R/l_1^2$. Now combining (D.4), (D.5), and (D.8), we obtain the upper bound for (I):

$$\mathbb{E}\|(A_i^0 - A_i)\prod_{j=1}^t (I - \nu_j u_i^j {u_i^j}^\top)\|_F^2 \le \prod_{j=1}^t (1 - l_1\nu_j)^2 \|A_i^0 - A_i\|_F^2 \overset{(a)}{\le} \left(\frac{t_0}{t+t_0}\right)^4 \|A_i^0 - A_i\|_F^2, \tag{D.9}$$

where (a) uses the inequality (C.15).

- **Upper bound of (II).**

Define $\psi_i^j = (A_i u_i^j - y_i^j)(u_i^j)^\top$, then (II) has the following form

$$\begin{aligned} &\mathbb{E}\|\sum_{j=1}^t \nu_j (A_i u_i^j - y_i^j)(u_i^j)^\top \prod_{k=j+1}^t (I - \nu_k u_i^k {u_i^k}^\top)\|_F^2 = \mathbb{E}\|\sum_{j=1}^t \nu_j \psi_i^j \prod_{k=j+1}^t (I - \nu_k u_i^k {u_i^k}^\top)\|_F^2 \\ &\quad = \mathbb{E}\sum_{j=1}^t \|\nu_j \psi_i^j \prod_{k=j+1}^t (I - \nu_k u_i^k {u_i^k}^\top)\|_F^2 + 2\mathbb{E}\sum_{s<j} \langle \nu_s \psi_i^s \prod_{k=s+1}^t (I - \nu_k u_i^k {u_i^k}^\top), \nu_j \psi_i^j \prod_{k=j+1}^t (I - \nu_k u_i^k {u_i^k}^\top)\rangle_F. \end{aligned} \tag{D.10}$$

Let $\mathcal{F} = \{\mathcal{F}_t\}_{t\in\mathbb{N}}$ be the filtration $\mathcal{F}_t = \sigma\{\{x^j\}_{j\in[t]}\}$ and define $\mathbb{E}_t = E[\cdot|\mathcal{F}_i^t]$. Since $u_i^j \overset{i.i.d}{\sim} \rho_{\mathcal{X}}$ and $y_i^j \sim \mathcal{D}(u_i^j)$, then for any $i \in [n]$ and $j \in \mathbb{N}$, the following equation holds

$$\mathbb{E}_j \psi_i^j = \mathbb{E}[\psi_i^j|\mathcal{F}_j] = \mathbb{E}[(A_i u_i^j - y_i^j)(u_i^j)^\top|\mathcal{F}_j] = 0, \tag{D.11}$$

here we use the linear parametric assumption (Assumption 4.7) and the independency between u_i^j and the noise term ϵ_i^j to obtain

$$\mathbb{E}[(A_i u_i^j - y_i^j)(u_i^j)^\top|\mathcal{F}_j] = \mathbb{E}[(A_i u_i^j - (A_i u_i^j + \epsilon_i^j))(u_i^j)^\top|\mathcal{F}_j] = 0.$$

Equation (D.11) shows that for a fixed $i \in [n]$, $\{\psi_i^j\}_{j\in\mathbb{N}}$ is a martingale difference sequence with filtration $\mathcal{F}$. Therefore, the second term in the last equation of (D.10) is zero, i.e.

$$\begin{aligned} &\mathbb{E}\sum_{s<j} \langle \nu_s \psi_i^s \prod_{k=s+1}^t (I - \nu_k u_i^k {u_i^k}^\top), \nu_j \psi_i^j \prod_{k=j+1}^t (I - \nu_k u_i^k {u_i^k}^\top)\rangle_F \\ &= \mathbb{E}\left[\sum_{s<j} \mathbb{E}_s\left[\langle \nu_s \psi_i^s \prod_{k=s+1}^t (I - \nu_k u_i^k {u_i^k}^\top), \nu_j \psi_i^j \prod_{k=j+1}^t (I - \nu_k u_i^k {u_i^k}^\top)\rangle_F\right]\right] \\ &= \mathbb{E}\sum_{s<j} \langle \nu_s \mathbb{E}_s[\psi_i^s] \prod_{k=s+1}^t (I - \nu_k u_i^k {u_i^k}^\top), \nu_j \psi_i^j \prod_{k=j+1}^t (I - \nu_k u_i^k {u_i^k}^\top)\rangle_F \\ &= 0. \end{aligned}$$

Plugging this equation into (D.10), we have

$$\begin{aligned}\mathbb{E}\|\sum_{j=1}^{t}\nu_j(A_iu_i^j-y_i^j)(u_i^j)^\top\prod_{k=j+1}^{t}(I-\nu_ku_i^ku_i^{k\top})\|_F^2&=\mathbb{E}\sum_{j=1}^{t}\|\nu_j\psi_i^j\prod_{k=j+1}^{t}(I-\nu_ku_i^ku_i^{k\top})\|_F^2\\&\le\sum_{j=1}^{t}\nu_j^2\mathbb{E}\left[\prod_{k=j+1}^{t}\|I-\nu_ku_i^ku_i^{k\top}\|_{\mathrm{op}}^2\mathbb{E}_j\|\psi_i^j\|_F^2\right].\end{aligned}\tag{D.12}$$

Moreover, Assumption 4.7 and Assumption 4.8 together imply that

$$\begin{aligned}\mathbb{E}_j\|\psi_i^j\|_F^2&=\mathbb{E}_j\|(A_iu_i^j-y_i^j)(u_i^j)^\top\|_F^2=\mathbb{E}_j\|\epsilon_i^j(u_i^j)^\top\|_F^2\\&=\mathbb{E}_j\|\epsilon_i^j\|^2\mathbb{E}_j\|u_i^j\|^2\le\sigma^2l_2.\end{aligned}$$

Plugging this inequality into (D.12) and using (D.8), we have

$$\begin{aligned}\mathbb{E}\|\sum_{j=1}^{t}\nu_j(A_iu_i^j-y_i^j)(u_i^j)^\top\prod_{k=j+1}^{t}(I-\nu_ku_i^ku_i^{k\top})\|_F^2&\le\sigma^2l_2\sum_{j=1}^{t}\nu_j^2\prod_{k=j+1}^{t}\mathbb{E}\|I-\nu_ku_i^ku_i^{k\top}\|_{\mathrm{op}}^2\\&\le\sigma^2l_2\sum_{j=1}^{t}\nu_j^2\prod_{k=j+1}^{t}(1-l_1\nu_k)^2.\end{aligned}$$

Using (C.15), we obtain the upper bound for (II)

$$\begin{aligned}\mathbb{E}\|\sum_{j=1}^{t}\nu_j(A_iu_i^j-y_i^j)(u_i^j)^\top\prod_{k=j+1}^{t}(I-\nu_ku_i^ku_i^{k\top})\|_F^2&\le\sigma^2l_2\sum_{j=1}^{t}\nu_j^2\prod_{k=j+1}^{t}(1-l_1\nu_k)^2\\&\le\sigma^2l_2\sum_{j=1}^{t}\nu_j^2\prod_{k=j+1}^{t}(1-l_1\nu_k)^2\le\sigma^2l_2\sum_{j=1}^{t}\left(\frac{2}{l_1(j+t_0)}\right)^2\left(\frac{j+1+t_0}{t+1+t_0}\right)^2\\&\le\frac{4l_2\sigma^2}{l_1^2}\left(\frac{t_0+2}{t_0+1}\right)^2(t+t_0)^{-1}.\end{aligned}\tag{D.13}$$

Plugging (D.9) and (D.13) into (D.3), we finish the proof

$$\begin{aligned}\mathbb{E}\|A_i^t-A_i\|_F^2&\le2\left(\left(\frac{t_0}{t+t_0}\right)^4\|A_i^0-A_i\|_F^2+\frac{4l_2\sigma^2}{l_1^2}\left(\frac{t_0+2}{t_0+1}\right)^2(t+t_0)^{-1}\right)\\&\le2\left(\frac{t_0^4}{(t_0+1)^3}\|A_i^0-A_i\|_F^2+\frac{4l_2\sigma^2}{l_1^2}\left(\frac{t_0+2}{t_0+1}\right)^2\right)(t+t_0)^{-1}.\end{aligned}$$

D.2 Proof of Lemma 4.11

We use Lemma B.2 to derive the one-step error bound, it is sufficient to check that the estimator for the gradient $H(x)$ satisfies Assumption B.1. Recalling Algorithm 1, the gradient estimator at iteration t is

$$h^t=(\nabla_i\ell_i(x^t,z_i^t)+(A_{ii}^t)^\top\nabla_{z_i}\ell_i(x^t,z_i^t))_{i\in[n]}.\tag{D.14}$$

Recalling (2.7), the true gradient is

$$H(x^t) = (\nabla_i \mathcal{L}_i(x^t))_{i\in[n]} = (\mathbb{E}_t \left[\nabla_i \ell_i(x^t, z_i^t) + (A_{ii})^\top \nabla_{z_i} \ell_i(x^t, z_i^t)\right])_{i\in[n]}. \tag{D.15}$$

We remark that $\{(x^j, u_i^j, y_i^j, z_i^j)\}_{i\in[n], j\in[t-1]}$ and (u_i^t, y_i^t) are deterministic with respect to the conditional expectation $\mathbb{E}_t[\cdot]$ (recalling the definition in Lemma 4.11), therefore, this expectation is equivalent to $\mathbb{E}_{z_i^t\sim\mathcal{D}_i(x^t)}[\cdot]$.

Let us prove that the gradient estimator h^t satisfies the Assumption B.1. To do this, we compute the bias and variance terms, respectively.

- **Bias.**

Using (D.14) and (D.15), we have

$$\|\mathbb{E}_t h^t - H(x^t)\|^2 = \sum_{i=1}^n \|\mathbb{E}_t \left[(A_{ii}^t - A_{ii})^\top \nabla_{z_i} \ell_i(x^t, z_i^t)\right]\|^2.$$

Since A_{ii}^t is deterministic with respect to the conditional expectation $\mathbb{E}_t[\cdot]$, we obtain

$$\begin{aligned}\|\mathbb{E}_t h^t - H(x^t)\|^2 &= \sum_{i=1}^n \|(A_{ii}^t - A_{ii})^\top \mathbb{E}_t \nabla_{z_i} \ell_i(x^t, z_i^t)\|^2 \\ &\le \sum_{i=1}^n \|A_{ii}^t - A_{ii}\|_F^2 \|\mathbb{E}_t \nabla_{z_i} \ell_i(x^t, z_i^t)\|^2.\end{aligned}$$

Assumption 4.3 implies that $\|\mathbb{E}_t(\nabla_{z_i} \ell_i(x^t, z_i^t))_{i\in[n]}\| \le \delta$. Plugging in this inequality, we obtain

$$\|\mathbb{E}_t h^t - H(x^t)\| \le \delta \sup_{i\in[n]} \|A_{ii}^t - A_{ii}\|_F \le \delta \sup_{i\in[n]} \|A_i^t - A_i\|_F. \tag{D.16}$$

Comparing (D.16) with (B.3), we have $m_t = \delta \sup_{i\in[n]} \|A_i^t - A_i\|_F$ and $U = 0$.

- **Variance.**

Recalling (D.14), define $A^t := (\nabla_i \ell_i(x^t, z_i^t))_{i\in[n]}$ and $B^t := ((A_{ii}^t)^\top \nabla_{z_i} \ell_i(x^t, z_i^t))_{i\in[n]}$, then $h^t = A^t + B^t$. We compute the variance of h^t

$$\begin{aligned}\mathbb{E}_t\|h^t - \mathbb{E}_t h^t\|^2 &= \mathbb{E}_t\|(A^t - \mathbb{E}_t A^t) + (B^t - \mathbb{E}_t B^t)\|^2 \\ &\le 2\left(\mathbb{E}_t\|A^t - \mathbb{E}_t A^t\|^2 + \mathbb{E}_t\|B^t - \mathbb{E}_t B^t\|^2\right).\end{aligned} \tag{D.17}$$

Now we derive the upper bounds for last the two terms of (D.17), respectively.

Upper bound of $\mathbb{E}_t\|A^t - \mathbb{E}_t A^t\|^2$. By the definition of A^t,

$$\mathbb{E}_t\|A^t - \mathbb{E}_t A^t\|^2 = \mathbb{E}_t\|(\nabla_i \ell_i(x^t, z_i^t) - \mathbb{E}_t \nabla_i \ell_i(x^t, z_i^t))_{i\in[n]}\|^2.$$

Assumption 4.4 implies that

$$\mathbb{E}_t\|A^t - \mathbb{E}_t A^t\|^2 \le \zeta^2. \tag{D.18}$$

Upper bound of $\mathbb{E}_t\|B^t - \mathbb{E}_t B^t\|^2$. By the definition of B^t, we have

$$\begin{aligned}\mathbb{E}_t\|B^t - \mathbb{E}_t B^t\|^2 &= \mathbb{E}_t\|((A_{ii}^t)^\top \left[\nabla_{z_i}\ell_i(x^t, z_i^t) - \mathbb{E}_t\nabla_{z_i}\ell_i(x^t, z_i^t)\right])_{i\in[n]}\|^2 \\ &\leq \sup_{i\in[n]}\|A_{ii}^t\|_F^2 \mathbb{E}_t\|(\left[\nabla_{z_i}\ell_i(x^t, z_i^t) - \mathbb{E}_t\nabla_{z_i}\ell_i(x^t, z_i^t)\right])_{i\in[n]}\|^2.\end{aligned}$$

Again, Assumption 4.4 implies that

$$\mathbb{E}_t\|(\left[\nabla_{z_i}\ell_i(x^t, z_i^t) - \mathbb{E}_t\nabla_{z_i}\ell_i(x^t, z_i^t)\right])_{i\in[n]}\|^2 \leq \zeta^2.$$

Therefore,

$$\mathbb{E}_t\|B^t - \mathbb{E}_t B^t\|^2 \leq \zeta^2 \sup_{i\in[n]}\|A_i^t\|_F^2, \tag{D.19}$$

Combining (D.18) and (D.19), we have

$$\mathbb{E}_t\|h^t - \mathbb{E}_t h^t\|^2 \leq 2\zeta^2(1 + \sup_{i\in[n]}\|A_i^t\|_F^2). \tag{D.20}$$

Comparing (D.20) with (B.3), we have $\sigma_t^2 = 2\zeta^2(1 + \sup_{i\in[n]}\|A_i^t\|_F^2)$ and $V = 0$.

Now we have proved that the stochastic gradient estimator h^t satisfies the stochastic framework (Assumption B.1) with $U = V = 0$, $m_t = \delta \sup_{i\in[n]}\|A_i^t - A_i\|_F$, and $\sigma_t^2 = 2\zeta^2(1 + \sup_{i\in[n]}\|A_i^t\|_F^2)$. Using Lemma B.2, we obtain the one-step error

$$\mathbb{E}_t\|x^{t+1} - x^*\|^2 \leq \frac{1}{1+\eta_t\tau}\|x^t - x^*\|^2 + \frac{4\eta_t^2\zeta^2(1 + \sup_{i\in[n]}\|A_i^t\|_F^2)}{1+\eta_t\tau} + \frac{2\eta_t\delta^2\sup_{i\in[n]}\|A_i^t - A_i\|_F^2}{\tau(1+\eta_t\tau)}. \tag{D.21}$$

E Proofs of Auxiliary Lemmas for Theorem 4.19

E.1 Proof of Lemma 4.18

Roadmap To prove Lemma 4.18, we begin with the symmetric of estimation updates, we show that these updates are independent between agents and have the same structure, thus, it is sufficient to study the iteration of any single agent. Then we propose the basic decomposition of the estimation error (Lemma E.2), which decomposes the error into three terms. Lemma E.4, Lemma E.6, and Lemma E.11 provide upper bounds for the three terms in Lemma E.2 and together finish the proof.

Symmetric of the estimation update. Recalling the estimation update in Algorithm 2

$$f_i^t = f_i^{t-1} - \nu_t\left[\left(f_i^{t-1}(u_i^t) - y_i^t\right)\phi_{u_i^t} + \lambda_t f_i^{t-1}\right]. \tag{E.1}$$

We show that for any $i, j \in [n]$, $i \neq j$, and any iteration t, the stochastic estimators f_i^t, f_j^t are independent and have the same iteration structure.

- **Independency.** Note that for any iteration t and $i \in [n]$, we have $u_i^t \overset{i.i.d}{\sim} \rho_{\mathcal{X}}$ and $y_i^t = f_i(u_i^t) + \epsilon_i^t$, where the noise terms ϵ_i^t are independent of u_i^t for any i and t (Assumption 2.6). Thus, for any $i \neq j$, (u_i^t, y_i^t) and (u_j^t, y_j^t) are independent. Moreover, recalling (E.1), f_i^t is random variables determined by $(u_i^1, y_i^1), (u_i^2, y_i^2), \cdots, (u_i^t, y_i^t)$. As a result, the stochastic estimators for different agents are independent, i.e. for any $i \neq j$, f_i^t and f_j^t are independent.

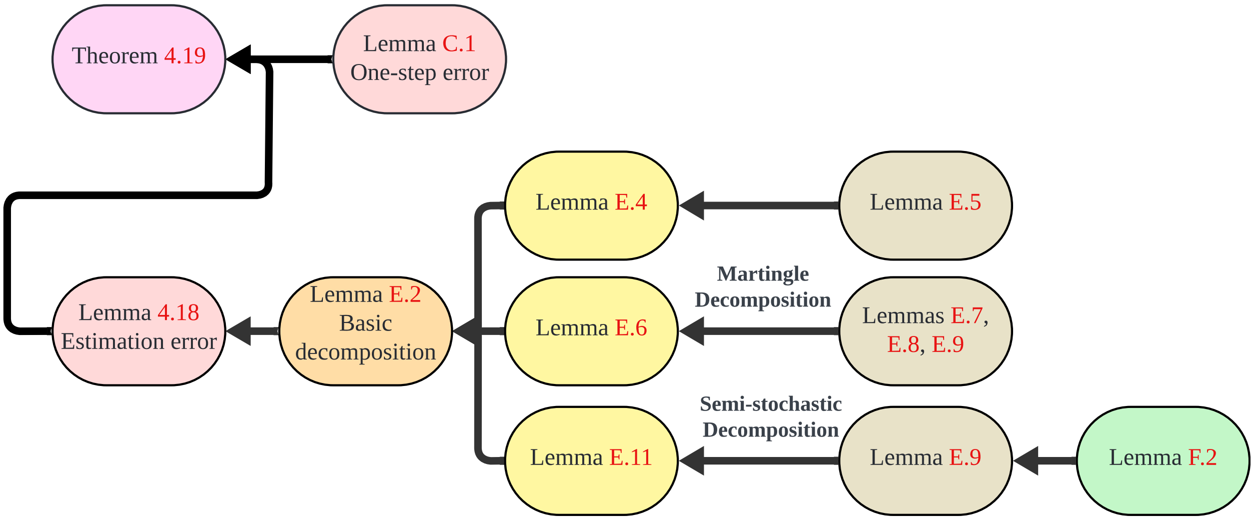


Figure 4: Proof sketch

- **Symmetric.** For any $i \in [n]$ and iteration t, f_i^t is a linear combination of $\{\phi_{u_i^k}\}_{k\in[t]}$, thus, $f_i^t \in \mathcal{H}$. Using the reproducing property of $\mathcal{H}$, $f_i^{t-1}(u_i^t) = \langle f_i^{t-1}, \phi_{u_i^t}\rangle_{\mathcal{H}}$. Also, recalling that $y_i^t = f_i(u_i^t) + \epsilon_i^t$. The iteration (E.1) can be rewritten as

$$f_i^t = (I - \nu_t(L_{u_i^t} + \lambda_t I))f_i^{t-1} + \nu_t(f_i(u_i^t) + \epsilon_i^t)\phi_{u_i^t}, \tag{E.2}$$

  where $L_{u_i^t} = \phi_{u_i^t}^*\phi_{u_i^t} : \mathcal{H} \to \mathcal{H}$ is a compact, self-adjoint and positive-semidefinite operator. Moreover, since $u_i^t \overset{i.i.d}{\sim} \rho_{\mathcal{X}}$ follow the same distribution for any $i \in [n]$, the induced operators $L_{u_i^t}$ and $\phi_{u_i^t}$ also follow the same distribution for all agents. Consequently, all the f_i^t have the same iteration structure

$$f_t = (I - \nu_t(L_t + \lambda_t I))f_{t-1} + \nu_t y_t \phi_{u_t},$$

  where $u_t \sim \rho_{\mathcal{X}}$, L_t is the operator induced by u_t, and $y_t \sim \mathcal{D}_i(u_t)$ is the observed data. (Note that the distribution of y_t is different for different agents.)

Simplified notation As previously stated, it is enough to analyze the iteration of a single agent. In the rest of this section, we study the iteration of agent 1 and define the following simplified notation. We drop the lower index i of terms in (E.2) and define

$$f := f_1, \quad f_t := f_1^t, \quad u_t := u_1^t, \quad y_t := y_1^t, \quad \phi_t := \phi_{u_1^t},$$

where f denotes the true function f_1, and f_t denotes the estimated parametric function f_1^t. Moreover, we drop the lower index i of sample spaces $\mathcal{Z}_i$ and decision-dependent distribution maps $\mathcal{D}_i$, define

$$\mathcal{Z} = \mathcal{Z}_1, \quad \mathcal{D} = \mathcal{D}_1,$$

note that $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ still denotes the joint action set of n agents.

With the new notation, the gradient step (E.2) would be

$$f_t = (I - \nu_t(L_t + \lambda_t I))f_{t-1} + \nu_t y_t \phi_t, \tag{E.3}$$

where $L_t = \phi_t^* \phi_t : \mathcal{H} \to \mathcal{H}$ is a compact, self-adjoint and positive-semidefinite operator.

Technical contributions We consider the semi-stochastic population iteration g_t (E.4) and decompose $f_t - f$ into three terms $f_{\lambda_t} - f$, $f_{\lambda_t} - g_t$, and $f_t - g_t$ (Lemma E.2). While the analysis for $f_{\lambda_t} - f$ (Lemma E.4) uses standard spectral decomposition techniques, the proofs for $f_{\lambda_t} - g_t$ and $f_t - g_t$ (Lemma E.6 and Lemma E.11) are novel. The proofs are based on important observations that $\|h\|_\gamma = \|L_K^{(1-\gamma)/2} h\|_{\mathcal{H}}$ for any $h \in \mathcal{H}^\gamma$ and $L_K^{(1-\gamma)/2}$ is commutative with the operator $I - \nu_t(L_K + \lambda_t I)$. For instance, in (F.17), the commutativity between $L_K^{(1-\gamma)/2}$ and $\Pi_i^t = \prod_{j=i}^t (I - \nu_j(L_K + \lambda_j I))$ allow us to leverage the commutativity to decouple the power norm by the operator's RKHS spectral norm $\|\Pi_i^t\|_{\mathcal{H}\to\mathcal{H}}$ and the RKHS norm $\|L_K^{(1-\gamma)/2}(f_{\lambda_i} - f_{\lambda_{i-1}})\|_{\mathcal{H}}$. This method is involved in the proofs for Lemmas E.6, E.7, E.8, E.9, E.11 and plays a central role in our analysis.

Another contribution is the recursion decomposition for $f_t - g_t$ (E.14), where we decompose $f_t - g_t$ by a sequence of sampling noise iteration $r_t^{(k)}$ and derive the power norm bound $\|r_t^{(k)}\|_\gamma$ (Lemma F.1). Although an analogous idea was proposed in non-strongly-convex SGD (Bach and Moulines, 2013), the analysis is essentially difference (because we study this decomposition under the power norm) and we find that the sampling error $f_t - g_t$ can be decomposed as a finite sum of the noise process (Lemma F.2), instead of the infinite sum in Bach and Moulines (2013). This might potentially extend the method to a broader class of problems, particularly in situations where $\|r_t^{(k)}\|_\gamma$, is not constrained by geometrization sequences.

Basic decomposition Now let us derive the upper bound for $\mathbb{E}\|f_t - f\|_\gamma^2$. We begin with the basic decomposition (Lemma E.2), which decomposes $\mathbb{E}\|f_t - f\|_\gamma^2$ into three terms. First, define the semi-stochastic population g_t,

$$g_0 = f_0, \quad g_t = (I - \nu_t(L_K + \lambda_t I))g_{t-1} + \nu_t y_t \phi_t, \tag{E.4}$$

where $L_K : \mathcal{L}^2_{\rho_\mathcal{X}} \to \mathcal{H}$ is the integral operator and its limitation on $\mathcal{H}$ is the covariance operator, i.e. $L_K|_\mathcal{H} = \mathbb{E}_{x\sim\rho_\mathcal{X}} L_x = \mathbb{E}_{x\sim\rho_\mathcal{X}} \phi_x^* \phi_x$. Moreover, $L_K|_\mathcal{H}$ is a compact, self-adjoint, and positive-semidefinite operator on $\mathcal{H}$.

Remark E.1. Compare (E.3) with (E.4), the semi-stochastic iteration g_t replaces the stochastic operator L_t with its expectation L_K, and remains the second stochastic term $y_t \phi_t$, thus, g_t can be viewed as a population iteration of f_t with half-randomness. We show that the deterministic operator L_K has good properties under γ-norm and plays a key role in the following proof.

Lemma E.2. (Basic decomposition). For any iteration t,

$$(\mathbb{E}\|f_t - f\|_\gamma^2)^{1/2} \le \|f - f_{\lambda_t}\|_\gamma + (\mathbb{E}\|f_{\lambda_t} - g_t\|_\gamma^2)^{1/2} + (\mathbb{E}\|g_t - f_t\|_\gamma^2)^{1/2}, \tag{E.5}$$

where $f_{\lambda_t} = (L_K + \lambda_t I)^{-1} L_K f$ is the solution of the kernel ridge regression

$$\arg\min_{h\in\mathcal{H}} \int_{\mathcal{X}\times\mathcal{Z}} (y - h(u))^2 d\rho + \lambda_t \|h\|_{\mathcal{H}}^2.$$

Here, the random variable $(u, y) \in \mathcal{X} \times \mathcal{Z}$ follows the distribution ρ, where ρ is induced by $u \sim \rho_{\mathcal{X}}$ and $y \sim \mathcal{D}(u)$.

Proof. By Minkowski's inequality,

$$\begin{aligned}(\mathbb{E}\|f_t - f\|_\gamma^2)^{1/2} &= (\mathbb{E}\|f - f_{\lambda_t} + f_{\lambda_t} - g_t + g_t - f_t\|_\gamma^2)^{1/2} \\ &\le \|f - f_{\lambda_t}\|_\gamma + (\mathbb{E}\|f_{\lambda_t} - g_t\|_\gamma^2)^{1/2} + (\mathbb{E}\|g_t - f_t\|_\gamma^2)^{1/2}.\end{aligned}$$

□

Remark E.3. We explain the three terms in Lemma E.2.

- The first term $\|f - f_{\lambda_t}\|_\gamma^2$ is deterministic and independent from the gradient step (E.3), it only related to the property of Hilbert space $\mathcal{H}$ and kernel K, and the choice of regularization terms λ_t. If the source condition (Assumption 4.13) holds with $\beta > \gamma$, we will prove that $\|f - f_{\lambda_t}\|_\gamma^2$ converges to zero if λ_t goes to zero gradually.
- The second term $\mathbb{E}\|f_{\lambda_t} - g_t\|_\gamma^2$ describes the gap between the semi-stochastic population iteration and the regularization path f_{λ_t}. As aforementioned in Section 4.2, we should choose gradient step sizes $\{\nu_t\}_{t\in\mathbb{N}}$ and regularization sequence $\{\lambda_t\}_{t\in\mathbb{N}}$ properly to control this error. The convergence of $\mathbb{E}\|f_{\lambda_t} - g_t\|_\gamma^2$ is guaranteed by the source condition (Assumption 4.13) and the embedding property (Assumption 4.14).
- The third term $\mathbb{E}\|g_t - f_t\|_\gamma^2$ shows the difference between the population iteration g_t and true iteration f_t. Recalling (E.3) and (E.4), this error mainly comes from the difference between the stochastic operator L_t and its expectation L_K, we show that $\mathbb{E}\|g_t - f_t\|_\gamma^2$ goes to zero if the expectation of the spectral norm $\mathbb{E}\|L_t - L_K\|_{\mathcal{H}^\gamma\to\mathcal{H}^\gamma}$ is small. The convergence of $\mathbb{E}\|g_t - f_t\|_\gamma^2$ is guaranteed by the source condition (Assumption 4.13) and the embedding property (Assumption 4.14).

We set the gradient step size and regularization sequence in the following forms

$$\nu_t = a(t + t_0)^{-\theta}, \qquad \lambda_t = \frac{1}{a}(t + t_0)^{-(1-\theta)}, \tag{E.6}$$

where a and t_0 are positive constants, and θ is a constant to be determined. We derive the convergence rate for $\mathbb{E}\|f_t - f\|_\gamma^2$ with respect to $\theta \in (1/2, 1]$, and choose θ to match the best convergence rate.

In the rest of this section, we are going to derive the upper bounds for three terms in Lemma E.2

$$\begin{aligned}&\text{Lemma E.4} \qquad \|f - f_{\lambda_t}\|_\gamma^2 \lesssim \mathcal{O}\left(t^{-(1-\theta)(\beta-\gamma)}\right), \\ &\text{Lemma E.6} \qquad \mathbb{E}\|f_{\lambda_t} - g_t\|_\gamma^2 \lesssim \mathcal{O}\left(t^{-((1-\theta)(\beta-\gamma)\wedge(2\theta-1))}\right), \\ &\text{Lemma E.11} \qquad \mathbb{E}\|g_t - f_t\|_\gamma^2 \lesssim \mathcal{O}\left(t^{-(2\theta-1)}\right).\end{aligned}$$

First term of Lemma E.2 Let us derive the upper bound for $\|f - f_{\lambda_t}\|_\gamma^2$.

Lemma E.4. Suppose that the source condition (Assumption 4.13) holds with some $\beta \in (0, 2]$. For any $\gamma \in (0, 1]$ and $\beta > \gamma$, let $\lambda_t = a^{-1} \cdot (t + t_0)^{-(1-\theta)}$, where $a, t_0 > 0$ are constants. The following bound holds for all $t \in \mathbb{N}$:

$$\|f - f_{\lambda_t}\|_\gamma^2 \lesssim \mathcal{O}\left(t^{-(1-\theta)(\beta-\gamma)}\right). \tag{E.7}$$

Proof. We make use of the following lemma.

Lemma E.5. Suppose that the assumptions of Lemma E.4 hold, the following bound holds for all $\lambda > 0$:

$$\|f - f_\lambda\|_\gamma^2 \le \frac{(\beta-\gamma)^{\beta-\gamma}(2-\beta+\gamma)^{2-\beta+\gamma}}{4}\|f\|_\beta^2 \lambda^{\beta-\gamma}, \tag{E.8}$$

where $f_\lambda = (L_K + \lambda I)^{-1} L_K f$.

Proof. See Appendix F.1 for the proof of this claim. □

Using Lemma E.5, for any $t \in \mathbb{N}$, let $\lambda = \lambda_t = 1/(a(t + t_0)^{1-\theta})$, we have

$$\begin{aligned}
\|f - f_{\lambda_t}\|_\gamma^2 &\le \frac{(\beta-\gamma)^{\beta-\gamma}(2-\beta+\gamma)^{2-\beta+\gamma}}{4}\|f\|_\beta^2 \lambda_t^{\beta-\gamma} \\
&= \frac{(\beta-\gamma)^{\beta-\gamma}(2-\beta+\gamma)^{2-\beta+\gamma}}{4a^{\beta-\gamma}}\|f\|_\beta^2 (t+t_0)^{-(1-\theta)(\beta-\gamma)} \\
&\lesssim \mathcal{O}\left(t^{-(1-\theta)(\beta-\gamma)}\right).
\end{aligned}$$

□

Second term of Lemma E.2 Now we are going to bound the second term in Lemma E.2, i.e. $\mathbb{E}\|f_{\lambda_t} - g_t\|_\gamma^2$. Using the martingale decomposition for $g_t - f_{\lambda_t}$ (see Appendix F.6 and Lemma F.1 for explanations):

$$g_t - f_{\lambda_t} = \underbrace{\Pi_1^t(g_0 - f_{\lambda_0})}_{\text{(I)}} + \underbrace{\sum_{i=1}^t \nu_i \Pi_{i+1}^t (y_i \phi_i - L_K f)}_{\text{(II)}} - \underbrace{\sum_{i=1}^t \Pi_i^t (f_{\lambda_i} - f_{\lambda_{i-1}})}_{\text{(III)}}, \tag{E.9}$$

where $\Pi_i^j = \prod_{k=i}^j (I - \nu_k(L_K + \lambda_k I))$ is an operator on $\mathcal{H}$. We derive the γ-norm bound for (I), (II), (III), respectively. We propose the intuitive explanation for the three terms in (E.9):

(I) is the initial error caused by the deviation of initialization $g_0 - f_{\lambda_0}$.

(II) is the sampling error induced by the randomness of sampling, i.e. the difference between the random variable $y_i \phi_i$ and its expectation $L_K f$.

(III) is the drift error caused by the changing of the regularization term λ_t. Lemma E.9 shows that $f_{\lambda_t} - f_{\lambda_{t-1}}$ is mostly determined by the gap between λ_{t-1} and λ_t, i.e. $\lambda_t - \lambda_{t-1}$.

Note that (I) and (III) are non-random and independent from the gradient step (E.4), they are determined by the gradient step size $\{\nu_t\}_{t\in\mathbb{N}}$ and the regularization sequence $\{\lambda_t\}_{t\in\mathbb{N}}$, (II) is stochastic and determined by the sampling data $(u_1, y_1), (u_2, y_2), \cdots, (u_t, y_t)$.

Lemma E.6. Suppose that the assumptions of Lemma E.4 hold, let $\nu_t = a(t+t_0)^{-\theta}$ and $\lambda_t = 1/(a(t+t_0)^{1-\theta})$, where $a, t_0 > 0$ are constants and $t_0 \geq (a\kappa^2+1)^2$. Suppose the embedding property (Assumption 4.14) holds for some $\alpha \in (0, \gamma]$. For any $\theta \in (1/2, 1]$ and any iteration t, the inequality holds

$$\mathbb{E}\|g_t - f_{\lambda_t}\|_\gamma^2 \lesssim \mathcal{O}\left(t^{-((1-\theta)(\beta-\gamma)\wedge(2\theta-1))}\right). \tag{E.10}$$

Proof. The following lemmas provide the upper bound for (I), (II), and (III) defined in (E.9), respectively.

Lemma E.7. (Upper bound for (I)). Suppose that the assumptions of Lemma E.6 hold. For any $\theta \in (1/2, 1]$ and any iteration t, the following inequality holds

$$\|\Pi_1^t(g_0 - f_{\lambda_0})\|_\gamma^2 \lesssim \mathcal{O}(t^{-2}). \tag{E.11}$$

Proof. See Appendix F.2 for the proof of this claim. □

Lemma E.8. (Upper bound for (II)). Suppose that the assumptions of Lemma E.6 hold. For any $\theta \in (1/2, 1]$ and any iteration t, the inequality holds

$$\mathbb{E}\|\sum_{i=1}^t \nu_i \Pi_{i+1}^t (y_i\phi_i - L_K f)\|_\gamma^2 \lesssim \mathcal{O}(t^{1-2\theta}). \tag{E.12}$$

Proof. See Appendix F.3 for the proof of this claim. □

Lemma E.9. (Upper bound for (III)). Suppose that the assumptions of Lemma E.6 hold. For any $\theta \in (1/2, 1]$ and any iteration t, the inequality holds

$$\|\sum_{i=1}^t \Pi_i^t (f_{\lambda_i} - f_{\lambda_{i-1}})\|_\gamma^2 \lesssim \mathcal{O}(t^{-(1-\theta)(\beta-\gamma)}). \tag{E.13}$$

Proof. See Appendix F.4 for the proof of this claim. □

Given the martingale decomposition (E.9), Lemma E.7, Lemma E.8, and Lemma E.9 together yield the upper bound for $\mathbb{E}\|g_t - f_{\lambda_t}\|_\gamma^2$. Using equation (E.9) and Minkowski's inequality, we obtain

$$(\mathbb{E}\|g_t - f_{\lambda_t}\|_\gamma^2)^{1/2} \leq (\|\Pi_1^t(g_0 - f_{\lambda_0})\|_\gamma^2)^{1/2} + (\mathbb{E}\|\sum_{i=1}^t \nu_i \Pi_{i+1}^t (y_i\phi_i - L_K f)\|_\gamma^2)^{1/2} + (\|\sum_{i=1}^t \Pi_i^t (f_{\lambda_i} - f_{\lambda_{i-1}})\|_\gamma^2)^{1/2}$$

Plugging in Lemma E.7, Lemma E.8, and Lemma E.9, we have

$$\begin{aligned}(\mathbb{E}\|g_t - f_{\lambda_t}\|_\gamma^2)^{1/2} &\lesssim \mathcal{O}(t^{-1}) + \mathcal{O}(t^{(1-2\theta)/2}) + \mathcal{O}(t^{-(1-\theta)(\beta-\gamma)/2}) \\ &\lesssim \mathcal{O}\left(t^{-((1-\theta)(\beta-\gamma)\wedge(2\theta-1))/2}\right),\end{aligned}$$

i.e. $\mathbb{E}\|g_t - f_{\lambda_t}\|_\gamma^2 \lesssim \mathcal{O}\left(t^{-((1-\theta)(\beta-\gamma)\wedge(2\theta-1))}\right)$. □

Third term of Lemma E.2 Now we are going to analyze the third term in the basic decomposition (E.5), i.e. $\mathbb{E}\|f_t - g_t\|_\gamma^2$. This error describes the deviation between the true iteration f_t and the semi-stochastic population iteration g_t. We decompose $f_t - g_t$ by a sequence of semi-stochastic noise process $\{r_t^{(k)}\}_{k\in\mathbb{N}_0}$ ($\mathbb{N}_0$ denotes $\mathbb{N}\cup 0$). Lemma F.2 shows that, for a fixed iteration t, $f_t - g_t$ is a finite sum of the noise process. Thus, $\mathbb{E}\|f_t - g_t\|_\gamma^2$ is bounded by the upper bounds of the sequence $\{\mathbb{E}\|r_t^{(k)}\|_\gamma^2\}_{k\in\mathbb{N}_0}$. As a result, to derive the convergence rate of $\mathbb{E}\|f_t - g_t\|_\gamma^2$, it is sufficient to derive upper bounds for th γ-norm of the noise process $\{r_t^{(k)}\}_{k\in\mathbb{N}_0}$.

Let us define the noise process $\{r_t^{(k)}\}_{k\in\mathbb{N}_0}$. For any $t \in \mathbb{N}$, define

$$\begin{aligned}
r_0^{(0)} &= 0, \quad r_t^{(0)} = (I - \nu_t(L_K + \lambda_t I))r_{t-1}^{(0)} + \nu_t(L_K - L_t)g_{t-1},\\
r_0^{(1)} &= 0, \quad r_t^{(1)} = (I - \nu_t(L_K + \lambda_t I))r_{t-1}^{(1)} + \nu_t(L_K - L_t)r_{t-1}^{(0)},\\
r_0^{(2)} &= 0, \quad r_t^{(2)} = (I - \nu_t(L_K + \lambda_t I))r_{t-1}^{(2)} + \nu_t(L_K - L_t)r_{t-1}^{(1)},\\
&\vdots\\
r_0^{(k)} &= 0, \quad r_t^{(k)} = (I - \nu_t(L_K + \lambda_t I))r_{t-1}^{(k)} + \nu_t(L_K - L_t)r_{t-1}^{(k-1)},\\
&\vdots
\end{aligned} \tag{E.14}$$

Intuitive explanation We briefly illustrate the noise process. (Appendix F.6 provides a detailed explanation and visualizes the values of this noise process sequence.) The aim is to derive the upper bound for $\mathbb{E}\|f_t - g_t\|_\gamma^2$, to do this, the noise process $r_t^{(k)}$ decomposes $f_t - g_t$ with a finite sum. Recalling (E.3) and (E.4), we have

$$f_0 - g_0 = 0, \quad f_t - g_t = (I - \nu_t(L_t + \lambda_t I))(f_{t-1} - g_{t-1}) + \nu_t(L_K - L_t)g_{t-1}. \tag{E.15}$$

This iteration of $f_t - g_t$ involves the stochastic operator L_t, which is hard to handle. Therefore, we replace L_t with its expectation L_K and consider the iteration

$$r_0^{(0)} = 0, \quad r_t^{(0)} = (I - \nu_t(L_K + \lambda_t I))r_{t-1}^{(0)} + \nu_t(L_K - L_t)g_{t-1}.$$

Then $r_t^{(0)}$ is the semi-stochastic population iteration of $f_t - g_t$, and we can decompose $f_t - g_t$ as

$$f_t - g_t = (f_t - g_t - r_t^{(0)}) + r_t^{(0)}.$$

Note that first term $f_t - g_t - r_t^{(0)}$ has the same recursion structure as (E.15)

$$f_0 - g_0 - r_0^{(0)} = 0, \quad f_t - g_t - r_t^{(0)} = (I - \nu_t(L_t + \lambda_t I))(f_{t-1} - g_{t-1} - r_{t-1}^{(0)}) + \nu_t(L_K - L_t)r_{t-1}^{(0)}.$$

Again, we consider the semi-stochastic population iteration for $f_0 - g_0 - r_0^{(0)}$ and define $r_t^{(1)}$. We obtain the error decomposition for $f_t - g_t$ as follows

$$f_t - g_t = (f_t - g_t - r_t^{(0)} - r_t^{(1)}) + r_t^{(0)} + r_t^{(1)}.$$

Moreover, we can repeat this procedure and yield the sequence $\{r_t^{(k)}\}_{k\in\mathbb{N}_0}$, the corresponding error decomposition would be

$$f_t - g_t = (f_t - g_t - \sum_{j=0}^{k} r_t^{(j)}) + \sum_{j=0}^{k} r_t^{(j)}.$$

Remark E.10. The noise process $\{r_t^{(k)}\}_{k\in\mathbb{N}_0}$ plays a key role in the following proof, Lemma F.2 shows that $f_t - g_t$ can be decomposed as a finite sum of the noise process, i.e.

$$f_t - g_t = \sum_{j=0}^{(t-1)\vee 0} r_t^{(j)}.$$

Using Minkowski's inequality, $(\mathbb{E}\|f_t-g_t\|_\gamma^2)^{1/2} \le \sum_{j=0}^{t-1}(\mathbb{E}\|r_t^{(j)}\|_\gamma^2)^{1/2}$. Therefore, to derive the upper bound for $\mathbb{E}\|f_t - g_t\|_\gamma^2$, it is sufficient to derive the γ-norm bound for $r_t^{(k)}$ (Lemma E.12).

Lemma E.11. Suppose that the assumptions of Lemma E.6 hold. If $\sqrt{2\theta-1}(t_0+2)/(t_0+1)a\kappa^{2-\gamma}A < 1$, for all iteration t, the inequality holds

$$\mathbb{E}\|f_t - g_t\|_\gamma^2 \lesssim \mathcal{O}(t^{1-2\theta}). \tag{E.16}$$

Proof. WLOG, suppose t is sufficient large ($t > \lfloor\frac{2}{2\theta-1}\rfloor$). Using Lemma F.2, $f_t - g_t$ is a finite sum of the noise process

$$f_t - g_t = \sum_{k=0}^{t-1} r_t^{(k)}.$$

Using Minkowski's inequality, we have

$$(\mathbb{E}\|f_t - g_t\|_\gamma^2)^{1/2} \le \sum_{k=0}^{t-1}(\mathbb{E}\|r_t^{(k)}\|_\gamma^2)^{1/2} \tag{E.17}$$

We make use of the following lemma, which provides the upper bounds for $\mathbb{E}\|r_t^{(k)}\|_\gamma^2$.

Lemma E.12. Suppose that the assumptions of Lemma E.6 hold. For any iteration t, the inequality holds

$$\mathbb{E}\|r_t^{(k)}\|_\gamma^2 \le \begin{cases} C_k(t+1)^{(k+1)(1-2\theta)}, & k < \lfloor\frac{2}{2\theta-1}\rfloor, \\ C_k(t+1)^{-2}, & k \ge \lfloor\frac{2}{2\theta-1}\rfloor. \end{cases} \tag{E.18}$$

where C_k is a constant:

$$C_k = \begin{cases} 2M\left(\frac{t_0+2}{t_0+1}a\kappa^{2-\gamma}A\right)^{2(k+1)}\prod_{i=0}^{k}(2+(i+1)(1-2\theta))^{-1}, & k < \lfloor\frac{2}{2\theta-1}\rfloor, \\ 2M\left(\frac{t_0+2}{t_0+1}a\kappa^{2-\gamma}A\right)^{2(k+1)}(2\theta-1)^{k-\lfloor\frac{2}{2\theta-1}\rfloor}\prod_{i=0}^{\lfloor\frac{2}{2\theta-1}\rfloor}|(2+(i+1)(1-2\theta))^{-1}|, & k \ge \lfloor\frac{2}{2\theta-1}\rfloor. \end{cases} \tag{E.19}$$

Proof. See Appendix F.5 for the proof of this claim. □

Remark E.13. (Intuitive explanation of C_k).

- If we fix the iteration t and increase k. Lemma E.12 shows that

$$C_{k+1} = \left(\frac{t_0+2}{t_0+1}a\kappa^{2-\gamma}A\right)^2 (2+(k+1)(1-2\theta))^{-1}C_k, \quad k < \lfloor\frac{2}{2\theta-1}\rfloor,$$
$$C_{k+1} = \left(\frac{t_0+2}{t_0+1}a\kappa^{2-\gamma}A\right)^2 (2\theta-1)C_k, \quad k \geq \lfloor\frac{2}{2\theta-1}\rfloor,$$

which implies that when k is sufficient large, C_k is roughly a geometric sequence with ratio $((t_0+2)(t_0+1)a\kappa^{2-\gamma}A)^2(2\theta-1)$. Moreover, if this ratio is smaller than 1, the sum of the geometric sequence is always finite.

- If we fix k and increase the iteration t. Lemma E.12 shows that when $k < \lfloor\frac{2}{2\theta-1}\rfloor$, the convergence rate of $\mathbb{E}\|r_t^{(k)}\|_\gamma^2$ is $\mathcal{O}(t^{(k+1)(1-2\theta)})$. Then if k increases (and $k < \lfloor\frac{2}{2\theta-1}\rfloor$), the convergence rate $(k+1)(1-2\theta)$ increases linearly with k. When $k \geq \lfloor\frac{2}{2\theta-1}\rfloor$, the convergence is fixed as $\mathcal{O}(t^{-2})$ and unrelated with k. This indicates that the convergence rate for the $\{\mathbb{E}\|r_t^{(k)}\|_\gamma^2\}_{k\in\mathbb{N}_0}$ is saturated at $\mathcal{O}(t^{-2})$ when $k \geq \lfloor\frac{2}{2\theta-1}\rfloor$.

Plugging Lemma E.12 into (E.17), we obtain

$$\begin{aligned}\sum_{k=0}^{t-1}(\mathbb{E}\|r_t^{(k)}\|_\gamma^2)^{1/2} &\leq \sum_{k=0}^{\lfloor\frac{2}{2\theta-1}\rfloor-1}(C_k(t+1+t_0)^{(k+1)(1-2\theta)})^{1/2} + \sum_{k=\lfloor\frac{2}{2\theta-1}\rfloor}^{t-1}(C_k(t+1+t_0)^{-2})^{1/2} \\ &\overset{(a)}{\leq} \left(\sum_{k=0}^{t-1}C_k^{1/2}\right)(t+1+t_0)^{(1-2\theta)/2}.\end{aligned} \tag{E.20}$$

where (a) uses the fact that $(t+1+t_0)^{(k+1)(1-2\theta)} \leq (t+1+t_0)^{-2})^{1/2}$ when $k < \lfloor\frac{2}{2\theta-1}\rfloor$. Given (E.19), the last step of the proof is showing that $\sum_{k=0}^{t-1}C_k^{1/2}$ is uniformly bounded. Recalling the definition of constants C_k , we have

$$\begin{aligned}\sum_{k=0}^{t-1}C_k^{1/2} &= \sum_{k=0}^{\lfloor\frac{2}{2\theta-1}\rfloor-1}\sqrt{2M\left(\frac{t_0+2}{t_0+1}a\kappa^{2-\gamma}A\right)^{2(k+1)}\prod_{i=0}^{k}(2+(i+1)(1-2\theta))^{-1}} \\ &+ \sum_{k=\lfloor\frac{2}{2\theta-1}\rfloor}^{t-1}\sqrt{2M\left(\frac{t_0+2}{t_0+1}a\kappa^{2-\gamma}A\right)^{2(k+1)}(2\theta-1)^{k-\lfloor\frac{2}{2\theta-1}\rfloor}\prod_{i=0}^{\lfloor\frac{2}{2\theta-1}\rfloor}(2+(i+1)(1-2\theta))^{-1}} \\ &\leq \underbrace{\sqrt{2M(2\theta-1)^{-\lfloor\frac{2}{2\theta-1}\rfloor-1}\prod_{i=0}^{\lfloor\frac{2}{2\theta-1}\rfloor}|(2+(i+1)(1-2\theta))^{-1}|}}_{:=E}\sum_{k=0}^{t-1}\left(\frac{t_0+2}{t_0+1}a\kappa^{2-\gamma}A\sqrt{2\theta-1}\right)^{k+1}.\end{aligned} \tag{E.21}$$

Note that $(t_0+2)/(t_0+1)a\kappa^{2-\gamma}A\sqrt{2\theta-1} < 1$. Then the summation of geometric sequence

$\sum_{k=0}^{t-1} \left((t_0+2)/(t_0+1)a\kappa^{2-\gamma}A\sqrt{2\theta-1}\right)^{k+1}$ is uniformly bounded. We obtain

$$\begin{aligned}\sum_{k=0}^{t-1} C_k^{1/2} &= \frac{E(\frac{t_0+2}{t_0+1}a\kappa^{2-\gamma}A\sqrt{2\theta-1})(1-(\frac{t_0+2}{t_0+1}a\kappa^{2-\gamma}A\sqrt{2\theta-1})^{t-1})}{1-\frac{t_0+2}{t_0+1}a\kappa^{2-\gamma}A\sqrt{2\theta-1}} \\ &\le \frac{E\frac{t_0+2}{t_0+1}a\kappa^{2-\gamma}A\sqrt{2\theta-1}}{1-\frac{t_0+2}{t_0+1}a\kappa^{2-\gamma}A\sqrt{2\theta-1}}\end{aligned} \tag{E.22}$$

Combining (E.17), (E.20), and (E.22), we have

$$\mathbb{E}\|f_t-g_t\|_\gamma^2 \le \left(\frac{E\frac{t_0+2}{t_0+1}a\kappa^{2-\gamma}A\sqrt{2\theta-1}}{1-\frac{t_0+2}{t_0+1}a\kappa^{2-\gamma}A\sqrt{2\theta-1}}\right)^2 (t+1+t_0)^{1-2\theta} \lesssim \mathcal{O}(t^{1-2\theta}).$$

□

Putting Lemma E.4, Lemma E.6, and Lemma E.11 together, we finish the proof for Lemma 4.18. Recalling the basic decomposition (Lemma E.2),

$$(\mathbb{E}\|f_t-f\|_\gamma^2)^{1/2} \le \|f-f_{\lambda_t}\|_\gamma + (\mathbb{E}\|f_{\lambda_t}-g_t\|_\gamma^2)^{1/2} + (\mathbb{E}\|g_t-f_t\|_\gamma^2)^{1/2}.$$

Lemma E.4, Lemma E.6, and Lemma E.11 provide the upper bounds for the three terms in this inequality, respectively. Therefore, we obtain

$$\begin{aligned}(\mathbb{E}\|f_t-f\|_\gamma^2)^{1/2} &\le \mathcal{O}\left(t^{-(1-\theta)(\beta-\gamma-2)/2}\right) + \mathcal{O}\left(t^{-((1-\theta)(\beta-\gamma)\wedge(2\theta-1))/2}\right) + \mathcal{O}(t^{(1-2\theta)/2}) \\ &= \mathcal{O}\left(t^{-((1-\theta)(\beta-\gamma)\wedge(2\theta-1))/2}\right).\end{aligned} \tag{E.23}$$

To get the best convergence rate, let $(1-\theta)(\beta-\gamma)=2\theta-1$, i.e. set $\theta=(\beta-\gamma+1)/(\beta-\gamma+2)$. Plugging this value into (E.23), we have

$$\mathbb{E}\|f_t-f\|_\gamma^2 \lesssim \mathcal{O}\left(t^{-\frac{\beta-\gamma}{\beta-\gamma+2}}\right).$$

E.2 Proof of Lemma C.1

Similar to the proof of Lemma 4.11, we use Lemma B.2 to derive the one-step error bound and it is sufficient to check that the estimator for the gradient $H(x)$ satisfies Assumption B.1. Recalling Algorithm 2, the gradient estimator at iteration t is

$$h^t = (\nabla_i \ell_i(x^t, z_i^t) + \langle f_i^t, \partial_i K_{x^t}\rangle_{\mathcal{H}} \nabla_{z_i}\ell_i(x^t, z_i^t))_{i\in[n]}. \tag{E.24}$$

Recalling (2.7), the true gradient is

$$H(x^t) = (\nabla_i \mathcal{L}_i(x^t))_{i\in[n]} = (\mathbb{E}_t\left[\nabla_i \ell_i(x^t, z_i^t) + \partial_i f_i(x^t)\nabla_{z_i}\ell_i(x^t, z_i^t)\right])_{i\in[n]}, \tag{E.25}$$

We remark that $\{(x^j, u_i^j, y_i^j, z_i^j)\}_{i\in[n], j\in[t-1]}$ and (u_i^t, y_i^t) are deterministic with respect to the conditional expectation $\mathbb{E}_t[\cdot]$ (recalling the definition in Lemma C.1), therefore, this expectation is equivalent to $\mathbb{E}_{z_i^t\sim\mathcal{D}_i(x^t)}[\cdot]$.

The source condition (Assumption 4.13) implies that for any $i \in [n]$, $f_i \in (\mathcal{H}^\beta)^p$. Therefore, for any $\gamma < \beta$, using the fact that $\mathcal{H}^\beta \subset \mathcal{H}^\gamma$, we have $f_i \in (\mathcal{H}^\gamma)^p$. Moreover, let K^γ be the kernel associated with the RKHS $\mathcal{H}^\gamma$, the reproducing property of $\mathcal{H}^\gamma$ implies that $f_i(x^t) = \langle f_i, K^\gamma_{x^t}\rangle_\gamma$. Thus, $\partial_i f_i(x^t) = \langle f_i, \partial_i K^\gamma_{x^t}\rangle_\gamma = \langle f_i, \partial_i \phi^\gamma_{x^t}\rangle_\gamma$ and we can rewrite the true gradient (E.25) as follow

$$H(x^t) = (\mathbb{E}_t \left[\nabla_i \ell_i(x^t, z_i^t) + \langle f_i, \partial_i \phi^\gamma_{x^t}\rangle_\gamma \nabla_{z_i} \ell_i(x^t, z_i^t)\right])_{i\in[n]}. \tag{E.26}$$

Besides, since $f_i^t \in (\mathcal{H})^p$ and $\gamma \leq 1$, we have $f_i^t \in (\mathcal{H}^\gamma)^p$ and $\langle f_i^t, \partial_i K_{x^t}\rangle_{\mathcal{H}} = \partial_i f_i^t(x^t) = \langle f_i^t, \partial_i^\gamma K_{x^t}\rangle_\gamma = \langle f_i^t, \partial_i \phi^\gamma_{x^t}\rangle_\gamma$. Rewrite the gradient estimator (E.24)

$$h^t = (\nabla_i \ell_i(x^t, z_i^t) + \langle f_i^t, \partial_i \phi^\gamma_{x^t}\rangle_\gamma \nabla_{z_i} \ell_i(x^t, z_i^t))_{i\in[n]}. \tag{E.27}$$

Now let us check that the gradient estimator h^t satisfies the Assumption B.1. We compute the bias term and variance term, respectively.

- **Bias.**

Using (E.26) and (E.27), we have

$$\begin{aligned}\|\mathbb{E}_t h^t - H(x^t)\|^2 &= \|(\mathbb{E}_t \left[(\langle f_i^t, \partial_i \phi^\gamma_{x^t}\rangle_\gamma - \langle f_i, \partial_i \phi^\gamma_{x^t}\rangle_\gamma)\nabla_{z_i} \ell_i(x^t, z_i^t)\right])_{i\in[n]}\|^2 \\ &= \sum_{i=1}^n \|\mathbb{E}_t \left[\langle f_i^t - f_i, \partial_i \phi^\gamma_{x^t}\rangle_\gamma \nabla_{z_i} \ell_i(x^t, z_i^t)\right]\|^2.\end{aligned}$$

Note that f_i^t is deterministic with the filtration $\mathcal{G}_t$, then

$$\mathbb{E}_t \left[\langle f_i^t - f_i, \partial_i \phi^\gamma_{x^t}\rangle_\gamma \nabla_{z_i} \ell_i(x^t, z_i^t)\right] = \langle f_i^t - f_i, \partial_i \phi^\gamma_{x^t}\rangle_\gamma \mathbb{E}_t \nabla_{z_i} \ell_i(x^t, z_i^t).$$

Plugging in this equation, we obtain

$$\begin{aligned}\|\mathbb{E}_t h^t - H(x^t)\|^2 &= \sum_{i=1}^n \|\langle f_i^t - f_i, \partial_i \phi^\gamma_{x^t}\rangle_\gamma \mathbb{E}_t \nabla_{z_i} \ell_i(x^t, z_i^t)\|^2 \\ &\leq \sum_{i=1}^n \|f_i^t - f_i\|_\gamma^2 \|\partial_i \phi^\gamma_{x^t}\|_\gamma^2 \|\mathbb{E}_t \nabla_{z_i} \ell_i(x^t, z_i^t)\|^2.\end{aligned}$$

Assumption 4.15 implies that $\|\partial_i \phi^\gamma_x\|_\gamma^2$ are uniformly bounded by ξ^2, therefore

$$\begin{aligned}\|\mathbb{E}_t h^t - H(x^t)\|^2 &\leq \xi^2 \sup_{i\in[n]} \|f_i^t - f_i\|_\gamma^2 \sum_{i=1}^n \|\mathbb{E}_t \nabla_{z_i} \ell_i(x^t, z_i^t)\|^2 \\ &\overset{(a)}{\leq} \left(\xi \sup_{i\in[n]} \|f_i^t - f_i\|_\gamma \|\mathbb{E}_t(\nabla_{z_i} \ell_i(x^t, z_i^t))_{i\in[n]}\|\right)^2,\end{aligned}$$

where (a) uses that $\sum_{i=1}^n \|\mathbb{E}_t \nabla_{z_i} \ell_i(x^t, z_i^t)\|^2 = \|\mathbb{E}_t(\nabla_{z_i} \ell_i(x^t, z_i^t))_{i\in[n]}\|^2$. Moreover, Assumption 4.3 implies that $\|\mathbb{E}_t(\nabla_{z_i} \ell_i(x^t, z_i^t))_{i\in[n]}\| \leq \delta$. Plugging in this inequality, we obtain

$$\|\mathbb{E}_t h^t - H(x^t)\| \leq \xi\delta \sup_{i\in[n]} \|f_i^t - f_i\|_\gamma. \tag{E.28}$$

Comparing (E.28) with (B.3), we have $m_t = \xi\delta \sup_{i\in[n]} \|f_i^t - f_i\|_\gamma$ and $U = 0$.

- **Variance.**

Recalling (E.27), let $A^t := (\nabla_i \ell_i(x^t, z_i^t))_{i\in[n]}$ and $B^t := (\langle f_i^t, \partial_i \phi_{x^t}^\gamma\rangle_\gamma \nabla_{z_i}\ell_i(x^t, z_i^t))_{i\in[n]}$, then $h^t = A^t + B^t$. We compute the variance of h^t

$$\begin{aligned}\mathbb{E}_t\|h^t - \mathbb{E}_t h^t\|^2 &= \mathbb{E}_t\|(A^t - \mathbb{E}_t A^t) + (B^t - \mathbb{E}_t B^t)\|^2 \\ &\le 2\left(\mathbb{E}_t\|A^t - \mathbb{E}_t A^t\|^2 + \mathbb{E}_t\|B^t - \mathbb{E}_t B^t\|^2\right).\end{aligned} \tag{E.29}$$

Now we derive the upper bounds for last the two terms of (E.29), respectively.

Upper bound of $\mathbb{E}_t\|A^t - \mathbb{E}_t A^t\|^2$. By the definition of A^t,

$$\mathbb{E}_t\|A^t - \mathbb{E}_t A^t\|^2 = \mathbb{E}_t\|(\nabla_i \ell_i(x^t, z_i^t) - \mathbb{E}_t \nabla_i \ell_i(x^t, z_i^t))_{i\in[n]}\|^2.$$

Then Assumption 4.4 implies that

$$\mathbb{E}_t\|A^t - \mathbb{E}_t A^t\|^2 \le \zeta^2. \tag{E.30}$$

Upper bound of $\mathbb{E}_t\|B^t - \mathbb{E}_t B^t\|^2$. By the definition of B^t, we have

$$\begin{aligned}\mathbb{E}_t\|B^t - \mathbb{E}_t B^t\|^2 &= \mathbb{E}_t\|(\langle f_i^t, \partial_i \phi_{x^t}^\gamma\rangle_\gamma \left[\nabla_{z_i}\ell_i(x^t, z_i^t) - \mathbb{E}_t \nabla_{z_i}\ell_i(x^t, z_i^t)\right])_{i\in[n]}\|^2 \\ &\le \|f_i^t\|_\gamma^2 \|\partial_i \phi_{x^t}^\gamma\|_\gamma^2 \mathbb{E}_t\|(\left[\nabla_{z_i}\ell_i(x^t, z_i^t) - \mathbb{E}_t \nabla_{z_i}\ell_i(x^t, z_i^t)\right])_{i\in[n]}\|^2.\end{aligned}$$

Again, Assumption 4.4 implies that

$$\mathbb{E}_t\|(\left[\nabla_{z_i}\ell_i(x^t, z_i^t) - \mathbb{E}_t \nabla_{z_i}\ell_i(x^t, z_i^t)\right])_{i\in[n]}\|^2 \le \zeta^2.$$

Moreover, Assumption 4.15 implies that $\|\partial_i \phi_{x^t}^\gamma\|_\gamma^2 \le \xi^2$. Therefore, we obtain

$$\mathbb{E}_t\|B^t - \mathbb{E}_t B^t\|^2 \le \xi^2\zeta^2 \sup_{i\in[n]} \|f_i^t\|_\gamma^2. \tag{E.31}$$

Plugging (E.30) and (E.31) into (E.29), we have

$$\mathbb{E}_t\|h^t - \mathbb{E}_t h^t\|^2 \le 2\zeta^2(1 + \xi^2 \sup_{i\in[n]} \|f_i^t\|_\gamma^2). \tag{E.32}$$

Comparing this inequality with (B.3), we have $\sigma_t^2 = 2\zeta^2(1 + \xi^2 \sup_{i\in[n]}\|f_i^t\|_\gamma^2)$ and $V = 0$.

Now we have proved that the stochastic gradient estimator h^t satisfies the stochastic framework (Assumption B.1) with $U = V = 0$, $m_t = \xi\delta \sup_{i\in[n]}\|f_i^t - f_i\|_\gamma$, and $\sigma_t^2 = 2\zeta^2(1 + \xi^2 \sup_{i\in[n]}\|f_i^t\|_\gamma^2)$. Using Lemma B.2, we obtain the one-step error

$$\mathbb{E}_t\|x^{t+1} - x^*\|^2 \le \frac{1}{1+\eta_t\tau}\|x^t - x^*\|^2 + \frac{4\eta_t^2\zeta^2(1 + \xi^2 \sup_{i\in[n]}\|f_i^t\|_\gamma^2)}{1+\eta_t\tau} + \frac{2\eta_t\xi^2\delta^2 \sup_{i\in[n]}\|f_i^t - f_i\|_\gamma^2}{\tau(1+\eta_t\tau)}. \tag{E.33}$$

F Omitted Proofs in Lemma 4.18

F.1 Proof of Lemma E.5

To begin with, we derive the spectral representation of f, f_λ, and use these representations to derive the upper bound for $\|f - f_\lambda\|_\gamma^2$.

- **Spectral representation of f.**

Let $\{\mu_i^{\beta/2} e_i\}_{i\in\mathbb{N}}$ be an orthogonal basis of $\mathcal{H}^\beta$, the source condition implies that $f \in \mathcal{H}^\beta$. Therefore, we assume that

$$f = \sum_{i=1}^{\infty} a_i \mu_i^{\beta/2} e_i, \qquad \{a_i\}_{i=1}^{\infty} \in \ell^2. \tag{F.1}$$

- **Spectral representation of L_K.**

Using the spectral representation of the integral operator (B.5), we have

$$\begin{aligned} L_K &= \sum_{i=1}^{\infty} \mu_i^{1/2} \langle e_i, \cdot \rangle_{\mathcal{L}^2_{\rho_\mathcal{X}}} \mu_i^{1/2} e_i = \sum_{i=1}^{\infty} \mu_i \langle \mu_i^{1/2} e_i, \cdot \rangle_{\mathcal{H}} \mu_i^{1/2} e_i, \\ (L_K + \lambda I)^{-1} &= \sum_{i=1}^{\infty} (\mu_i + \lambda)^{-1} \langle \mu_i^{1/2} e_i, \cdot \rangle_{\mathcal{H}} \mu_i^{1/2} e_i, \end{aligned} \tag{F.2}$$

where $\|\cdot\|_{\rho_\mathcal{X}}$ is the norm on $\mathcal{L}^2_{\rho_\mathcal{X}}$ induced by the measure $\rho_\mathcal{X}$. Note that $\{\mu_i^{1/2} e_i\}_{i\in\mathbb{N}}$ is an orthogonal basis of $\mathcal{H}$ and $\{e_i\}_{i\in\mathbb{N}}$ is an orthogonal basis of $\mathcal{L}^2_{\rho_\mathcal{X}}$, thus, $\langle e_i, e_j \rangle_{\mathcal{L}^2_{\rho_\mathcal{X}}} = \langle \mu_i^{1/2} e_i, \mu_j^{1/2} e_j \rangle_{\mathcal{H}} = \delta_{ij}$.

- **Spectral representation of f_λ.**

Recalling that $f_\lambda = (L_K + \lambda I)^{-1} L_K f$. Using (F.1) and the first equation in (F.2), we have

$$\begin{aligned} L_K f &= \sum_{i=1}^{\infty} \mu_i^{1/2} \langle e_i, f \rangle_{\mathcal{L}^2_{\rho_\mathcal{X}}} \mu_i^{1/2} e_i \\ &= \sum_{i=1}^{\infty} \mu_i^{1/2} \langle e_i, \sum_{j=1}^{\infty} a_j \mu_j^{\beta/2} e_j \rangle_{\mathcal{L}^2_{\rho_\mathcal{X}}} \mu_i^{1/2} e_i \\ &\overset{(a)}{=} \sum_{i=1}^{\infty} \mu_i^{1/2} \langle e_i, a_i \mu_i^{\beta/2} e_i \rangle_{\mathcal{L}^2_{\rho_\mathcal{X}}} \mu_i^{1/2} e_i \\ &\overset{(b)}{=} \sum_{i=1}^{\infty} a_i \mu_i^{(1+\beta)/2} \mu_i^{1/2} e_i, \end{aligned} \tag{F.3}$$

where (a) uses the fact that $\langle e_i, e_j \rangle_{\mathcal{L}^2_{\rho_\mathcal{X}}} = \delta_{ij}$. Now plugging in the spectral representation of $(L_K + \lambda I)^{-1}$ (i.e. the second equation of (F.2)), we have

$$\begin{aligned} f_\lambda &= (L_K + \lambda I)^{-1} L_K f = (L_K + \lambda I)^{-1} \sum_{i=1}^{\infty} a_i \mu_i^{(1+\beta)/2} \mu_i^{1/2} e_i \\ &= \sum_{i=1}^{\infty} (\mu_i + \lambda)^{-1} \langle \mu_i^{1/2} e_i, \sum_{j=1}^{\infty} a_j \mu_j^{(1+\beta)/2} \mu_j^{1/2} e_j \rangle_{\mathcal{H}} \mu_i^{1/2} e_i. \end{aligned}$$

Using the fact that $\langle \mu_i^{1/2} e_i, \mu_i^{1/2} e_j \rangle_{\mathcal{H}} = \delta_{ij}$, we obtain the spectral decomposition for f_λ

$$\begin{aligned} f_\lambda &= \sum_{i=1}^{\infty} (\mu_i + \lambda)^{-1} \langle \mu_i^{1/2} e_i, a_i \mu_i^{(1+\beta)/2} \mu_i^{1/2} e_i \rangle_{\mathcal{H}} \mu_i^{1/2} e_i \\ &= \sum_{i=1}^{\infty} (\mu_i + \lambda)^{-1} a_i \mu_i^{(1+\beta)/2} \mu_i^{1/2} e_i \\ &= \sum_{i=1}^{\infty} \frac{\mu_i^{(1+\beta)/2}}{\mu_i + \lambda} a_i \mu_i^{1/2} e_i. \end{aligned} \tag{F.4}$$

Combining (F.1) and (F.4), we obtain

$$\begin{aligned} f - f_\lambda &= \sum_{i=1}^{\infty} \left(\mu_i^{\beta/2} - \frac{\mu_i^{(1+\beta)/2}}{\mu_i + \lambda} \mu_i^{1/2} \right) a_i e_i = \sum_{i=1}^{\infty} \left(1 - \frac{\mu_i}{\mu_i + \lambda} \right) a_i \mu_i^{\beta/2} e_i \\ &= \sum_{i=1}^{\infty} \frac{\lambda}{\mu_i + \lambda} a_i \mu_i^{\beta/2} e_i. \end{aligned} \tag{F.5}$$

Therefore, the γ-norm has the following expression

$$\begin{aligned} \|f - f_\lambda\|_\gamma^2 &= \|\sum_{i=1}^{\infty} \frac{\lambda}{\mu_i + \lambda} a_i \mu_i^{\beta/2} e_i\|_\gamma^2 = \|\sum_{i=1}^{\infty} \frac{\lambda}{\mu_i + \lambda} \mu_i^{(\beta-\gamma)/2} a_i \mu_i^{\gamma/2} e_i\|_\gamma^2 \\ &\overset{(a)}{=} \sum_{i=1}^{\infty} \left(\frac{\lambda}{\mu_i + \lambda} \mu_i^{(\beta-\gamma)/2} a_i \right)^2, \end{aligned}$$

where (a) uses the fact that $\{\mu_i^{\gamma/2} e_i\}_{i \in \mathbb{N}}$ is an orthogonal basis of the γ-pwoer space $\mathcal{H}^\gamma$ (which further indicates $\langle \mu_i^{\gamma/2} e_i, \mu_j^{\gamma/2} e_j \rangle_\gamma = \delta_{ij}$). Given the equation above, we obtain the following bound

$$\begin{aligned} \|f - f_\lambda\|_\gamma^2 &= \sum_{i=1}^{\infty} \left(\frac{\lambda}{\mu_i + \lambda} \mu_i^{(\beta-\gamma)/2} a_i \right)^2 \leq \lambda^2 \sup_{i \in \mathbb{N}} \left(\frac{\mu_i^{(\beta-\gamma)/2}}{\mu_i + \lambda} \right)^2 \sum_{i=1}^{\infty} a_i^2 \\ &= \lambda^2 \sup_{i \in \mathbb{N}} \left(\frac{\mu_i^{(\beta-\gamma)/2}}{\mu_i + \lambda} \right)^2 \|f\|_\beta^2. \end{aligned} \tag{F.6}$$

We compute the last equation of (F.6), define function $h(x) = x^{\beta-\gamma}/(x+\lambda)^2$, for all $x \in \mathbb{R}^+$, we have

$$h'(x) = \frac{x^{\beta-\gamma-1}((\beta-\gamma)(x+\lambda) - 2x)}{(x+\lambda)^3}.$$

Therefore, $h(x)$ takes the maximum at $x = (\beta-\gamma)\lambda/(2-\beta+\gamma)$. Plugging in this maximum, we obtain

$$\begin{aligned} \|f - f_\lambda\|_\gamma^2 &\leq \lambda^2 \sup_{x \in \mathbb{R}} \left(\frac{x^{(\beta-\gamma)/2}}{x + \lambda} \right)^2 \|f\|_\beta^2 = \lambda^2 \sup_{x \in \mathbb{R}} h(x) \|f\|_\beta^2 \\ &\leq \frac{(\beta-\gamma)^{\beta-\gamma}(2-\beta+\gamma)^{2-\beta+\gamma}}{4} \|f\|_\beta^2 \lambda^{\beta-\gamma}. \end{aligned}$$

F.2 Proof of Lemma E.7

The aim is to derive the upper bound for $\|\Pi_1^t(g_0 - f_{\lambda_0})\|_\gamma^2$. Recalling the definition of iteration g_t (E.4), we have $g_0 = f_0$, therefore

$$\|\Pi_1^t(g_0 - f_{\lambda_0})\|_\gamma^2 = \|\Pi_1^t(f_0 - f_{\lambda_0})\|_\gamma^2.$$

Using Lemma G.2, we have

$$\|\Pi_1^t(f_0 - f_{\lambda_0})\|_\gamma^2 = \|L_K^{(1-\gamma)/2}\Pi_1^t(f_0 - f_{\lambda_0})\|_{\mathcal{H}}^2. \tag{F.7}$$

Note that the operator $\Pi_1^t = \prod_{i=1}^t (I - \nu_i(L_K + \lambda_i I))$ is the product of linear combinations of the integral operator L_K and the identity operator I. Therefore, operator Π_1^t is commutative with operator $L_K^{(1-\gamma)/2}$, thus, we can rewrite (F.7)

$$\begin{aligned}\|\Pi_1^t(f_0 - f_{\lambda_0})\|_\gamma^2 &= \|L_K^{(1-\gamma)/2}\Pi_1^t(f_0 - f_{\lambda_0})\|_{\mathcal{H}}^2 = \|\Pi_1^t L_K^{(1-\gamma)/2}(f_0 - f_{\lambda_0})\|_{\mathcal{H}}^2 \\ &\le \|\Pi_1^t\|_{\mathcal{H}\to\mathcal{H}}^2 \|L_K^{(1-\gamma)/2}(f_0 - f_{\lambda_0})\|_{\mathcal{H}}^2,\end{aligned}$$

where $\|\Pi_1^t\|_{\mathcal{H}\to\mathcal{H}}^2$ is the spectral norm of the operator Π_1^t. Using Lemma G.3 (all the assumptions of Lemma G.3 are satisfied, since $t_0 \ge (a\kappa^2 + 1)^2$ and $\theta \in (1/2, 1]$), this spectral norm satisfies the inequality

$$\|\Pi_1^t\|_{\mathcal{H}\to\mathcal{H}}^2 \le \prod_{i=1}^t (1 - \nu_i\lambda_i)^2.$$

Plugging in this inequality, we obtain

$$\|\Pi_1^t(g_0 - f_{\lambda_0})\|_\gamma^2 \le \|\Pi_1^t\|_{\mathcal{H}\to\mathcal{H}}^2 \|L_K^{(1-\gamma)/2}(f_0 - f_{\lambda_0})\|_{\mathcal{H}}^2 \le \prod_{i=1}^t (1 - \nu_i\lambda_i)^2 \|f_0 - f_{\lambda_0}\|_\gamma^2.$$

Note that we set $\nu_i = a(i + t_0)^{-\theta}$ and $\lambda_i = 1/(a(i + t_0)^{1-\theta})$, therefore,

$$\begin{aligned}\|\Pi_1^t(g_0 - f_{\lambda_0})\|_\gamma^2 &\le \prod_{i=1}^t (1 - (i + t_0)^{-1})^2 \|f_0 - f_{\lambda_0}\|_\gamma^2 = \prod_{i=1}^t \left(\frac{i + t_0 - 1}{i + t_0}\right)^2 \|f_0 - f_{\lambda_0}\|_\gamma^2 \\ &= \left(\frac{t_0}{t + t_0}\right)^2 \|f_0 - f_{\lambda_0}\|_\gamma^2 \lesssim \mathcal{O}(t^{-2}).\end{aligned}$$

F.3 Proof of Lemma E.8

The aim is to derive the upper bound for $\mathbb{E}\|\sum_{i=1}^t \nu_i \Pi_{i+1}^t (y_i\phi_i - L_K f)\|_\gamma^2$. Define $\omega_i = y_i\phi_i - L_K f$, then we have

$$\begin{aligned}\mathbb{E}\|\sum_{i=1}^t \nu_i \Pi_{i+1}^t (y_i\phi_i - L_K f)\|_\gamma^2 &= \mathbb{E}\|\sum_{i=1}^t \nu_i \Pi_{i+1}^t \omega_i\|_\gamma^2 = \mathbb{E}\sum_{i,j=1}^t \langle \nu_i \Pi_{i+1}^t \omega_i, \nu_j \Pi_{j+1}^t \omega_j\rangle_\gamma \\ &= \mathbb{E}\sum_{i=1}^t \|\nu_i \Pi_{i+1}^t \omega_i\|_\gamma^2 + 2\mathbb{E}\sum_{i<j} \langle \nu_i \Pi_{i+1}^t \omega_i, \nu_j \Pi_{j+1}^t \omega_j\rangle_\gamma.\end{aligned} \tag{F.8}$$

Let $\mathcal{F} = \{\mathcal{F}_t\}_{t\in\mathbb{N}}$ be the filtration $\mathcal{F}_t = \sigma\{\{x^j\}_{j\in[t]}\}$ and define $\mathbb{E}_t = E[\cdot|\mathcal{F}_t]$. Since $u_i \overset{i.i.d}{\sim} \rho_{\mathcal{X}}$ and $y_i \sim \mathcal{D}(u_i)$, for any $i \in \mathbb{N}$, the following equation holds

$$\mathbb{E}_{i-1}\omega_i = \mathbb{E}[\omega_i|\mathcal{F}_{i-1}] = \mathbb{E}[y_i\phi_i - L_K f|\mathcal{F}_{i-1}] = 0, \tag{F.9}$$

here we use the fact that

$$\mathbb{E}[y_i\phi_i - L_K f|\mathcal{F}_{i-1}] \overset{(a)}{=} \mathbb{E}[(f(u_i)+\epsilon_i)\phi_i - L_K f|\mathcal{F}_{i-1}] \overset{(b)}{=} \mathbb{E}[f(u_i)\phi_i - L_K f|\mathcal{F}_{i-1}],$$

where (a) is guaranteed by the parametric assumption (Assumption 4.13) and (b) is induced by the following equation (recalling the definition of L_K in Appendix B.2)

$$\mathbb{E}_{u_i\sim\rho_{\mathcal{X}}} f(u_i)\phi_i(x) = \int_{\mathcal{X}} K(u_i,x)f(u_i)d\rho_{\mathcal{X}}(u_i) = L_K(f)(x), \quad \forall x \in \mathcal{X}.$$

Equation (F.9) shows that $\{\omega_i\}_{i\in\mathbb{N}}$ is a martingale difference sequence with filtration $\mathcal{F}$. Therefore, using the tower rule, for any $i < j$ the second term in (F.8) is zero

$$\begin{aligned}\mathbb{E}\langle \nu_i\Pi_{i+1}^t\omega_i, \nu_j\Pi_{j+1}^t\omega_j\rangle_\gamma &= \mathbb{E}\left[\mathbb{E}\left[\langle \nu_i\Pi_{i+1}^t\omega_i, \nu_j\Pi_{j+1}^t\omega_j\rangle_\gamma \middle| \mathcal{F}_{j-1}\right]\right] \\ &= \mathbb{E}\left[\langle \nu_i\Pi_{i+1}^t\omega_i, \mathbb{E}_{j-1}[\nu_j\Pi_{j+1}^t\omega_j]\rangle_\gamma\right] \\ &= \mathbb{E}\left[\langle \nu_i\Pi_{i+1}^t\omega_i, 0\rangle_\gamma\right] = 0.\end{aligned} \tag{F.10}$$

Combining (F.8) and (F.10), we obtain

$$\mathbb{E}\|\sum_{i=1}^t \nu_i\Pi_{i+1}^t(y_i\phi_i - L_K f)\|_\gamma^2 = \mathbb{E}\sum_{i=1}^t \|\nu_i\Pi_{i+1}^t\omega_i\|_\gamma^2. \tag{F.11}$$

Moreover, for any $i, t \in \mathbb{N}$ and $i \le t$, using Lemma G.2, we have

$$\begin{aligned}\mathbb{E}\|\nu_i\Pi_{i+1}^t\omega_i\|_\gamma^2 &= \mathbb{E}\|L_K^{(1-\gamma)/2}\nu_i\Pi_{i+1}^t\omega_i\|_{\mathcal{H}}^2 = \mathbb{E}\|\nu_i\Pi_{i+1}^t L_K^{(1-\gamma)/2}\omega_i\|_{\mathcal{H}}^2 \\ &\le \nu_i^2\|\Pi_{i+1}^t\|_{\mathcal{H}\to\mathcal{H}}^2\mathbb{E}\|L_K^{(1-\gamma)/2}\omega_i\|_{\mathcal{H}}^2.\end{aligned}$$

Applying Lemma G.3 to $\|\Pi_{i+1}^t\|_{\mathcal{H}\to\mathcal{H}}^2$, we obtain

$$\begin{aligned}\mathbb{E}\|\nu_i\Pi_{i+1}^t\omega_i\|_\gamma^2 &\le \nu_i^2\prod_{j=i+1}^t(1-\nu_j\lambda_j)^2\mathbb{E}\|\omega_i\|_\gamma^2 = \nu_i^2\prod_{j=i+1}^t\left(\frac{j+t_0-1}{j+t_0}\right)^2\mathbb{E}\|\omega_i\|_\gamma^2 \\ &= \nu_i^2\left(\frac{i+t_0}{t+t_0}\right)^2\mathbb{E}\|\omega_i\|_\gamma^2.\end{aligned} \tag{F.12}$$

Combine (F.11) and (F.12), we obtain

$$\mathbb{E}\|\sum_{i=1}^t \nu_i\Pi_{i+1}^t(y_i\phi_i - L_K f)\|_\gamma^2 \le \sum_{i=1}^t \nu_i^2\left(\frac{i+t_0}{t+t_0}\right)^2\mathbb{E}\|\omega_i\|_\gamma^2. \tag{F.13}$$

Let us derive the upper bound for the last term of (F.13). We are going to prove that ω_i has a uniform bound under the γ-norm and derive the upper bound for $\sum_{i=1}^t \nu_i^2((i+t_0)/(t+t_0))^2$.

- **Uniform bound of $\mathbb{E}\|\omega_i\|_\gamma^2$.**

For any random variable z in a Hilbert space with norm $\|\cdot\|$, it is well known that $\mathrm{Var}(z) \le \mathbb{E}\|z\|^2$, therefore

$$\mathbb{E}\|\omega_i\|_\gamma^2 = \mathbb{E}\|y_i\phi_i - L_K f\|_\gamma^2 = \mathrm{Var}(y_i\phi_i) \le \mathbb{E}\|y_i\phi_i\|_\gamma^2.$$

Using Lemma G.5, we have

$$\mathbb{E}\|\omega_i\|_\gamma^2 \le \mathbb{E}\|y_i\phi_i\|_\gamma^2 \le \kappa^{2(2-\gamma)}(A^2\|f\|_\beta^2 + \sigma^2). \tag{F.14}$$

- **Upper bound of $\sum_{i=1}^t \nu_i^2((i+t_0)/(t+t_0))^2$.**

$$\begin{aligned}\sum_{i=1}^t \nu_i^2 \left(\frac{i+t_0}{t+t_0}\right)^2 &= \sum_{i=1}^t a^2(i+t_0)^{-2\theta}\left(\frac{i+t_0}{t+t_0}\right)^2 = a^2(t+t_0)^{-2}\sum_{i=1}^t (i+t_0)^{2-2\theta} \\ &\le a^2(t+t_0)^{-2}\sum_{i=1}^t \int_i^{i+1}(x+t_0)^{2-2\theta}dx \le a^2(t+t_0)^{-2}\int_1^{t+1}(x+t_0)^{2-2\theta}dx\end{aligned} \tag{F.15}$$

Calculating the last term of (F.15):

$$\begin{aligned}(t+t_0)^{-2}\int_1^{t+1}(x+t_0)^{2-2\theta}dx &= \frac{1}{3-2\theta}(t+t_0)^{-2}\Big((t+t_0+1)^{3-2\theta} - (t_0+1)^{3-2\theta}\Big) \\ &\le \frac{1}{3-2\theta}(t+t_0)^{-2}(t+t_0+1)^{3-2\theta} \le \frac{(t_0+2)^{3-2\theta}}{(3-2\theta)(t_0+1)^{2-2\theta}}(t+t_0)^{1-2\theta}.\end{aligned} \tag{F.16}$$

Combining (F.13), (F.14), (F.15), and (F.16), we obtain

$$\begin{aligned}\mathbb{E}\|\sum_{i=1}^t \nu_i \Pi_{i+1}^t(y_i\phi_i - L_K f)\|_\gamma^2 &\le a^2\kappa^{2(2-\gamma)}(A^2\|f\|_\beta^2+\sigma^2)\frac{(t_0+2)^{3-2\theta}}{(3-2\theta)(t_0+1)^{2-2\theta}}(t+t_0)^{1-2\theta} \\ &\lesssim \mathcal{O}(t^{1-2\theta}).\end{aligned}$$

F.4 Proof of Lemma E.9

The aim is to derive the upper bound for $\|\sum_{i=1}^t \Pi_i^t(f_{\lambda_i} - f_{\lambda_{i-1}})\|_\gamma^2$. Using Lemma G.2 and Lemma G.3, we have

$$\begin{aligned}\|\sum_{i=1}^t \Pi_i^t(f_{\lambda_i} - f_{\lambda_{i-1}})\|_\gamma &= \|L_K^{(1-\gamma)/2}\sum_{i=1}^t \Pi_i^t(f_{\lambda_i} - f_{\lambda_{i-1}})\|_{\mathcal{H}} = \|\sum_{i=1}^t \Pi_i^t L_K^{(1-\gamma)/2}(f_{\lambda_i} - f_{\lambda_{i-1}})\|_{\mathcal{H}} \\ &\le \sum_{i=1}^t \|\Pi_i^t\|_{\mathcal{H}\to\mathcal{H}}\|L_K^{(1-\gamma)/2}(f_{\lambda_i} - f_{\lambda_{i-1}})\|_{\mathcal{H}} \\ &\le \sum_{i=1}^t \prod_{j=i}^t (1-\nu_j\lambda_j)\|L_K^{(1-\gamma)/2}(f_{\lambda_i} - f_{\lambda_{i-1}})\|_{\mathcal{H}}.\end{aligned} \tag{F.17}$$

We first derive the upper bound of $\|L_K^{(1-\gamma)/2}(f_{\lambda_i} - f_{\lambda_{i-1}})\|_{\mathcal{H}}$ and then obtain the upper bound for the last term of (F.17), i.e. $\sum_{i=1}^t \prod_{j=i}^t (1-\nu_j\lambda_j)\|L_K^{(1-\gamma)/2}(f_{\lambda_i} - f_{\lambda_{i-1}})\|_{\mathcal{H}}$.

- **Upper bound of $\|L_K^{(1-\gamma)/2}(f_{\lambda_i} - f_{\lambda_{i-1}})\|_{\mathcal{H}}$.**

The same as the proof of Lemma E.5, let us derive the spectral representation for $L_K^{(1-\gamma)/2}(f_{\lambda_i} - f_{\lambda_{i-1}})$. Suppose $f = \sum_{i=1}^{\infty} a_i \mu_i^{\beta/2} e_i$. Using (F.2), we have

$$f_{\lambda_i} = (L_K + \lambda_i I)^{-1} L_K f = \sum_{j=1}^{\infty} \frac{\mu_j^{(1+\beta)/2}}{\mu_j + \lambda_i} a_j \mu_j^{1/2} e_j.$$

Thus,

$$\begin{aligned} f_{\lambda_i} - f_{\lambda_{i-1}} &= \sum_{j=1}^{\infty} \frac{\mu_j^{(1+\beta)/2}}{\mu_j + \lambda_i} a_j \mu_j^{1/2} e_j - \sum_{j=1}^{\infty} \frac{\mu_j^{(1+\beta)/2}}{\mu_j + \lambda_{i-1}} a_j \mu_j^{1/2} e_j \\ &= \sum_{j=1}^{\infty} \frac{\mu_j^{(1+\beta)/2}(\lambda_{i-1} - \lambda_i)}{(\mu_j + \lambda_i)(\mu_j + \lambda_{i-1})} a_j \mu_j^{1/2} e_j. \end{aligned}$$

Using the fact that $L_K^{(1-\gamma)/2} e_i = \mu_i^{(1-\gamma)/2} e_i$, we obtain

$$L_K^{(1-\gamma)/2}(f_{\lambda_i} - f_{\lambda_{i-1}}) = \sum_{j=1}^{\infty} \mu_i^{(1-\gamma)/2} \frac{\mu_j^{(1+\beta)/2}(\lambda_{i-1} - \lambda_i)}{(\mu_j + \lambda_i)(\mu_j + \lambda_{i-1})} a_j \mu_j^{1/2} e_j. \tag{F.18}$$

Therefore, using the spectral representation (F.18) and the fact that $\langle \mu_i^{1/2} e_i, \mu_j^{1/2} e_j \rangle_{\mathcal{H}} = \delta_{ij}$, we obtain the spectral representation of $\|L_K^{(1-\gamma)/2}(f_{\lambda_i} - f_{\lambda_{i-1}})\|_{\mathcal{H}}^2$:

$$\begin{aligned} \|L_K^{(1-\gamma)/2}(f_{\lambda_i} - f_{\lambda_{i-1}})\|_{\mathcal{H}}^2 &= \|\sum_{j=1}^{\infty} \mu_i^{(1-\gamma)/2} \frac{\mu_j^{(1+\beta)/2}(\lambda_{i-1} - \lambda_i)}{(\mu_j + \lambda_i)(\mu_j + \lambda_{i-1})} a_j \mu_j^{1/2} e_j\|_{\mathcal{H}}^2 \\ &= \sum_{j=1}^{\infty} \left(\frac{\mu_j^{(2-\gamma+\beta)/2}(\lambda_{i-1} - \lambda_i)}{(\mu_j + \lambda_i)(\mu_j + \lambda_{i-1})} \right)^2 a_j^2 \le \sup_{j \in \mathbb{N}} \left(\frac{\mu_j^{(2-\gamma+\beta)/2}(\lambda_{i-1} - \lambda_i)}{(\mu_j + \lambda_i)(\mu_j + \lambda_{i-1})} \right)^2 \sum_{j=1}^{\infty} a_j^2 \\ &= \sup_{j \in \mathbb{N}} \frac{\mu_j^{2-\gamma+\beta}(\lambda_{i-1} - \lambda_i)^2}{(\mu_j + \lambda_i)^2(\mu_j + \lambda_{i-1})^2} \|f\|_\beta^2. \end{aligned}$$

Recalling that $\lambda_i = 1/(a(i+t_0)^{1-\theta})$, we have $\lambda_i \le \lambda_{i-1}$ and

$$\|L_K^{(1-\gamma)/2}(f_{\lambda_i} - f_{\lambda_{i-1}})\|_{\mathcal{H}}^2 = \sup_{j \in \mathbb{N}} \frac{\mu_j^{2-\gamma+\beta}(\lambda_{i-1} - \lambda_i)^2}{(\mu_j + \lambda_i)^2(\mu_j + \lambda_{i-1})^2} \|f\|_\beta^2 \le \sup_{j \in \mathbb{N}} \frac{\mu_j^{2-\gamma+\beta}(\lambda_{i-1} - \lambda_i)^2}{(\mu_j + \lambda_i)^4} \|f\|_\beta^2. \tag{F.19}$$

To upper bound the last term of (F.19), define function $h(x) = x^{2-\gamma+\beta}(\lambda_{i-1} - \lambda_i)^2/(x+\lambda_i)^4$, for all $x \in \mathbb{R}^+$, we have

$$h'(x) = \frac{x^{1-\gamma+\beta}((2-\gamma+\beta)(x+\lambda) - 4x)}{(x+\lambda_i)^5}.$$

Therefore, $h(x)$ takes the maximum at $x = \lambda_i(2-\gamma+\beta)/(2+\gamma-\beta)$. Combining this maximum with (F.19), we obtain

$$\begin{aligned}\|L_K^{(1-\gamma)/2}(f_{\lambda_i} - f_{\lambda_{i-1}})\|_{\mathcal{H}}^2 &\le \sup_{j\in\mathbb{N}} \frac{\mu_j^{2-\gamma+\beta}(\lambda_{i-1}-\lambda_i)^2}{(\mu_j+\lambda_i)^4}\|f\|_\beta^2 \\ &\le \frac{(2-\gamma+\beta)^{2-\gamma+\beta}(2+\gamma-\beta)^{2+\gamma-\beta}}{256}\|f\|_\beta^2\lambda_i^{\beta-\gamma-2}(\lambda_{i-1}-\lambda_i)^2.\end{aligned} \tag{F.20}$$

Since $\lambda_i = 1/(a(i+t_0)^{1-\theta})$, we have

$$\lambda_i^{\beta-\gamma-2}(\lambda_{i-1}-\lambda_i)^2 = a^{\gamma-\beta}(i+t_0)^{-(1-\theta)(\beta-\gamma-2)}((i-1+t_0)^{-(1-\theta)} - (i+t_0)^{-(1-\theta)})^2. \tag{F.21}$$

- **Upper bound of $\sum_{i=1}^t \prod_{j=i}^t (1-\nu_j\lambda_j)\|L_K^{(1-\gamma)/2}(f_{\lambda_i} - f_{\lambda_{i-1}})\|_{\mathcal{H}}$.**

Now let us find derive the upper bound for the last term of (F.17). Using (F.20), we have

$$\begin{aligned}\sum_{i=1}^t\prod_{j=i}^t(1-\nu_j\lambda_j)\|L_K^{(1-\gamma)/2}(f_{\lambda_i} - f_{\lambda_{i-1}})\|_{\mathcal{H}} &\le \sum_{i=1}^t\prod_{j=i}^t(1-\nu_j\lambda_j)\sqrt{\sup_{j\in\mathbb{N}} \frac{\mu_j^{2-\gamma+\beta}(\lambda_{i-1}-\lambda_i)^2}{(\mu_j+\lambda_i)^4}\|f\|_\beta^2} \\ &\le \sqrt{\frac{(2-\gamma+\beta)^{2-\gamma+\beta}(2+\gamma-\beta)^{2+\gamma-\beta}}{256}\|f\|_\beta^2}\sum_{i=1}^t\prod_{j=i}^t(1-\nu_j\lambda_j)\lambda_i^{(\beta-\gamma-2)2}(\lambda_{i-1}-\lambda_i).\end{aligned}$$

Plugging (F.21) into the last inequality above, we obtain

$$\begin{aligned}&\sum_{i=1}^t\prod_{j=i}^t(1-\nu_j\lambda_j)\|L_K^{(1-\gamma)/2}(f_{\lambda_i} - f_{\lambda_{i-1}})\|_{\mathcal{H}} \\ &\qquad\le C\underbrace{\sum_{i=1}^t\prod_{j=i}^t(1-(j+t_0)^{-1})(i+t_0)^{-(1-\theta)(\beta-\gamma-2)/2}((i-1+t_0)^{-(1-\theta)} - (i+t_0)^{-(1-\theta)})}_{\text{(I)}},\end{aligned} \tag{F.22}$$

where $C = \sqrt{\frac{(2-\gamma+\beta)^{2-\gamma+\beta}(2+\gamma-\beta)^{2+\gamma-\beta}}{256a^{\beta-\gamma}}\|f\|_\beta^2}$ is a constant.

- **Upper bound of (I)**

By some calculation, we obtain the following inequality of (I)

$$\begin{aligned}\text{(I)} &= \sum_{i=1}^t\left(\frac{i-1+t_0}{t+t_0}\right)(i+t_0)^{-(1-\theta)(\beta-\gamma-2)/2}((i-1+t_0)^{-(1-\theta)} - (i+t_0)^{-(1-\theta)}) \\ &\le (1-\theta)\sum_{i=1}^t\frac{(i+t_0)^{1-(1-\theta)(\beta-\gamma-2)/2}}{t+t_0}\int_{i-1}^i(x+t_0)^{-(1-\theta)-1}dx \\ &\le (1-\theta)\sum_{i=1}^t\frac{(i+t_0)^{-(1-\theta)(\beta-\gamma-2)/2}}{t+t_0}.\end{aligned}$$

Since $(i+t_0)^{-(1-\theta)(\beta-\gamma-2)/2} \le \int_{i-1}^{i}(x+t_0)^{-(1-\theta)(\beta-\gamma-2)/2}dx$, we have

$$\begin{aligned}
(\mathrm{I}) &\le (1-\theta)\sum_{i=1}^{t}\frac{\int_{i-1}^{i}(x+t_0)^{-(1-\theta)(\beta-\gamma-2)/2}dx}{t+t_0} \\
&\le \frac{2(1-\theta)}{2-(1-\theta)(\beta-\gamma)}(t+t_0)^{-(1-\theta)(\beta-\gamma-2)/2}.
\end{aligned}$$

Plugging this inequality into (F.22) and using (F.17), we obtain

$$\begin{aligned}
\|\sum_{i=1}^{t}\Pi_i^t(f_{\lambda_i}-f_{\lambda_{i-1}})\|_\gamma^2 \le C^2\cdot(\mathrm{I})^2 &\le \left(\frac{2(1-\theta)C}{2-(1-\theta)(\beta-\gamma)}\right)^2(t+t_0)^{-(1-\theta)(\beta-\gamma)} \\
&\lesssim \mathcal{O}(t^{-(1-\theta)(\beta-\gamma)}).
\end{aligned}$$

F.5 Proof of Lemma E.12

We use mathematical induction to derive the upper bound for $\mathbb{E}\|r_t^{(k)}\|_\gamma^2$. We start with $r_t^{(0)}$ (**Initial case I**) and use induction to derive the upper bound of $r_t^{(k)}$ with $k < \lfloor\frac{2}{2\theta-1}\rfloor$ (**Induction step I**). Then we consider the upper bound of $r_t^{(k)}$ with $k = \lfloor\frac{2}{2\theta-1}\rfloor$ (**Initial case II**) and use induction to derive the upper bound of $r_t^{(k)}$ for all $k > \lfloor\frac{2}{2\theta-1}\rfloor$ (**Induction step II**).

- **Initial case I: For $k = 0$.**

Recalling (F.39), $r_t^{(0)}$ can be rewritten as

$$\begin{aligned}
r_t^{(0)} &= (I-\nu_t(L_K+\lambda_t I))r_{t-1}^{(0)} + \nu_t(L_K-L_t)g_{t-1} \\
&= \Pi_1^t r_0^{(0)} + \sum_{i=1}^{t}\nu_i\Pi_{i+1}^t(L_K-L_i)g_{i-1},
\end{aligned}$$

where $\Pi_i^j = \prod_{k=i}^{j}(I-\nu_k(L_K+\lambda_k I))$. Since $r_0^{(0)} = 0$, we have

$$r_t^{(0)} = \sum_{i=1}^{t}\nu_i\Pi_{i+1}^t(L_K-L_i)g_{i-1}.$$

Moreover, $\{(L_K-L_t)g_{t-1}\}_{t\in\mathbb{N}}$ is a martingale difference sequence with filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\in\mathbb{N}}$, where $\mathcal{F}_t = \sigma\{\{x^j\}_{j\in[t]}\}$, because

$$\mathbb{E}_t(L_K-L_t)g_{t-1} = \mathbb{E}[(L_K-L_t)g_{t-1}|\mathcal{F}_{t-1}] = \mathbb{E}[(L_K-L_t)|\mathcal{F}_{t-1}]g_{t-1} = 0.$$

Therefore, $\mathbb{E}\langle(L_K-L_i)g_{i-1},(L_K-L_j)g_{j-1}\rangle_{\mathcal{H}} = 0$ for any $i \ne j$. We further have

$$\begin{aligned}
\mathbb{E}\|r_t^{(0)}\|_\gamma^2 = \mathbb{E}\|r_t^{(0)}\|_\gamma^2 &= \mathbb{E}\|\sum_{i=1}^{t}\nu_i\Pi_{i+1}^t(L_K-L_i)g_{i-1}\|_\gamma^2 \\
&= \sum_{i=1}^{t}\mathbb{E}\|\nu_i\Pi_{i+1}^t(L_K-L_i)g_{i-1}\|_\gamma^2
\end{aligned}$$

Using Lemma G.2 and commute $L_K^{(1-\gamma)/2}$ with Π_{i+1}^t, we obtain

$$\begin{aligned}\mathbb{E}\|r_t^{(0)}\|_\gamma^2 &= \sum_{i=1}^t \mathbb{E}\|L_K^{(1-\gamma)/2}\nu_i\Pi_{i+1}^t(L_K - L_i)g_{i-1}\|_{\mathcal{H}}^2 \\ &\le \sum_{i=1}^t \mathbb{E}\nu_i^2\|\Pi_{i+1}^t\|_{\mathcal{H}\to\mathcal{H}}^2\|L_K^{(1-\gamma)/2}(L_K - L_i)g_{i-1}\|_{\mathcal{H}}^2 \\ &= \sum_{i=1}^t \nu_i^2\|\Pi_{i+1}^t\|_{\mathcal{H}\to\mathcal{H}}^2\mathbb{E}\|(L_K - L_i)g_{i-1}\|_\gamma^2.\end{aligned} \tag{F.23}$$

Using the tower rule for thee expectation $\mathbb{E}\|(L_K - L_i)g_{i-1}\|_\gamma^2$, we have

$$\mathbb{E}\|(L_K - L_i)g_{i-1}\|_\gamma^2 = \mathbb{E}[\mathbb{E}[\|(L_K - L_i)g_{i-1}\|_\gamma^2|\mathcal{F}_{i-1}]].$$

Since $\mathbb{E}[L_ig_{i-1}|\mathcal{F}_{i-1}] = L_Kg_{i-1}$, the above expectation is actually a variance of L_ig_{i-1}, namely, $\mathbb{E}[\|(L_K - L_i)g_{i-1}\|_\gamma^2|\mathcal{F}_{i-1}] = \mathrm{Var}(L_ig_{i-1}|\mathcal{F}_{i-1})$. Using the fact that for any random variable $z \in \mathcal{H}^\gamma$, $\mathrm{Var}(z) \le \mathbb{E}\|z\|_\gamma^2$, we obtain

$$\mathbb{E}\|(L_K - L_i)g_{i-1}\|_\gamma^2 \le \mathbb{E}\|L_ig_{i-1}\|_\gamma^2 \le \mathbb{E}[\|L_i\|_{\mathcal{H}_\gamma\to\mathcal{H}_\gamma}^2\|g_{i-1}\|_\gamma^2]. \tag{F.24}$$

Now we are going to derive the uniform upper bound for the spectral norm $\|L_i\|_{\mathcal{H}_\gamma\to\mathcal{H}_\gamma}^2$. For any $h \in \mathcal{H}^\gamma$, suppose $h = \sum_{j=1}^\infty a_j\mu_j^{\gamma/2}e_j$, we have

$$\begin{aligned}\|L_i\|_{\mathcal{H}_\gamma\to\mathcal{H}_\gamma}^2 &= \sup_{h\in\mathcal{H}^\gamma}\frac{\|L_ih\|_\gamma^2}{\|h\|_\gamma^2} = \sup_{\{a_j\}\in\ell^2}\frac{h^2(u_i)\|\phi_i\|_\gamma^2}{\sum_{j=1}^\infty a_j^2} \\ &= \|\phi_i\|_\gamma^2\sup_{\{a_j\}\in\ell^2}\frac{\left(\sum_{j=1}^\infty a_j\mu_j^{\gamma/2}e_j(u_i)\right)^2}{\sum_{j=1}^\infty a_j^2} \le \|\phi_i\|_\gamma^2\sup_{\{a_j\}\in\ell^2}\frac{\left(\sum_{j=1}^\infty a_j^2\right)\left(\sum_{j=1}^\infty \mu_j^\gamma e_j^2(u_i)\right)}{\sum_{j=1}^\infty a_j^2} \\ &= \|\phi_i\|_\gamma^2\left(\sum_{j=1}^\infty \mu_j^\gamma e_j^2(u_i)\right),\end{aligned}$$

Using Lemma G.4, we obtain the upper bound for the spectral norm $\|L_i\|_{\mathcal{H}_\gamma\to\mathcal{H}_\gamma}^2$

$$\|L_i\|_{\mathcal{H}_\gamma\to\mathcal{H}_\gamma}^2 \le \|\phi_i\|_\gamma^2\left(\sum_{j=1}^\infty \mu_j^\gamma e_j^2(u_i)\right) \le \kappa^{4-2\gamma}A^2 \tag{F.25}$$

Combining (F.23), (F.24), and (F.25), we have

$$\mathbb{E}\|r_t^{(0)}\|_\gamma^2 \le \kappa^{4-2\gamma}A^2\sum_{i=1}^t \nu_i^2\|\Pi_{i+1}^t\|_{\mathcal{H}\to\mathcal{H}}^2\mathbb{E}\|g_{i-1}\|_\gamma^2.$$

Using the triangle inequality,

$$\begin{aligned}\mathbb{E}\|r_t^{(0)}\|_\gamma^2 &\le 2\kappa^{4-2\gamma}A^2\sum_{i=1}^t \nu_i^2\|\Pi_{i+1}^t\|_{\mathcal{H}\to\mathcal{H}}^2\left(\mathbb{E}\|g_{i-1} - f_{\lambda_{i-1}}\|_\gamma^2 + \|f_{\lambda_{i-1}}\|_\gamma^2\right) \\ &\le 2\kappa^{4-2\gamma}A^2\sum_{i=1}^t \nu_i^2\|\Pi_{i+1}^t\|_{\mathcal{H}\to\mathcal{H}}^2\left(\mathbb{E}\|g_{i-1} - f_{\lambda_{i-1}}\|_\gamma^2 + 2(\|f_{\lambda_{i-1}} - f\|_\gamma^2 + \|f\|_\gamma^2)\right).\end{aligned} \tag{F.26}$$

Using Lemma E.4 and Lemma E.6, we have

$$\mathbb{E}\|g_{i-1} - f_{\lambda_{i-1}}\|_\gamma^2 \lesssim \mathcal{O}\left(i^{-((1-\theta)(\beta-\gamma)\wedge(2\theta-1))}\right), \qquad \|f_{\lambda_{i-1}} - f\|_\gamma^2 \lesssim \mathcal{O}\left(i^{-(1-\theta)(\beta-\gamma)}\right).$$

Therefore, there exist a constant $M > 0$, such that

$$\mathbb{E}\|g_{i-1} - f_{\lambda_{i-1}}\|_\gamma^2 + 2(\|f_{\lambda_{i-1}} - f\|_\gamma^2 + \|f\|_\gamma^2) \leq M\|f\|_\gamma^2.$$

Plugging this inequality into (F.26) and using Lemma G.3, we obtain

$$\begin{aligned}
\mathbb{E}\|r_t^{(0)}\|_\gamma^2 &\leq 2\kappa^{4-2\gamma}A^2M\|f\|_\gamma^2\sum_{i=1}^t \nu_i^2\|\Pi_{i+1}^t\|_{\mathcal{H}\to\mathcal{H}}^2 \leq 2\kappa^{4-2\gamma}A^2M\|f\|_\gamma^2\sum_{i=1}^t \nu_i^2\prod_{j=i+1}^t(1-\nu_j\lambda_j)^2 \\
&= 2\kappa^{4-2\gamma}A^2M\|f\|_\gamma^2\sum_{i=1}^t a^2(i+t_0)^{-2\theta}\prod_{j=i+1}^t(1-(j+t_0)^{-1})^2 \\
&= 2a^2\kappa^{4-2\gamma}A^2M\|f\|_\gamma^2\frac{\sum_{i=1}^t(i+t_0)^{2-2\theta}}{(t+t_0)^2}.
\end{aligned}$$

By some calculation, we obtain

$$\begin{aligned}
\mathbb{E}\|r_t^{(0)}\|_\gamma^2 &\leq 2a^2\kappa^{4-2\gamma}A^2M\|f\|_\gamma^2\frac{\sum_{i=1}^t(i+t_0)^{2-2\theta}}{(t+t_0)^2} \\
&\leq 2a^2\kappa^{4-2\gamma}A^2M\|f\|_\gamma^2\frac{\int_1^{t+1}(x+t_0)^{2-2\theta}dx}{(t+t_0)^2} \\
&\leq \frac{2(t_0+2)^2}{(3-2\theta)(t_0+1)^2}a^2\kappa^{4-2\gamma}A^2M\|f\|_\gamma^2(t+1+t_0)^{1-2\theta}.
\end{aligned}$$

Therefore, the inequality (E.18) for the initial case ($k=0$) holds.

- **Induction step I: For $k > 0$ and $k < \lfloor\frac{2}{2\theta-1}\rfloor$.**

Suppose the inequality for $k-1$ holds, i.e.

$$\mathbb{E}\|r_t^{(k-1)}\|_\gamma^2 \leq C_{k-1}(t+1+t_0)^{k(1-2\theta)}. \tag{F.27}$$

Recalling the definition of $r_t^{(k)}$ (F.40), we have $r_0^{(k)} = 0$ and

$$\begin{aligned}
r_t^{(k)} &= (I - \nu_t(L_K + \lambda_t I))r_{t-1}^{(k)} + \nu_t(L_K - L_t)r_{t-1}^{(k-1)} \\
&= \Pi_1^t r_0^{(k)} + \sum_{i=1}^t \nu_i\Pi_{i+1}^t(L_K - L_i)r_{i-1}^{(k-1)} \\
&= \sum_{i=1}^t \nu_i\Pi_{i+1}^t(L_K - L_i)r_{i-1}^{(k-1)}.
\end{aligned}$$

Again, $\{(L_K - L_t)r_{t-1}^{(k-1)}\}_{t\in\mathbb{N}}$ is a martingale difference sequence with filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\in\mathbb{N}}$ and we can apply the same analysis on $\sum_{i=1}^t \nu_i \Pi_{i+1}^t (L_K - L_i) r_{i-1}^{(k-1)}$ as (F.23) and (F.24) of the **Initial case I**.

$$\begin{aligned}\mathbb{E}\|r_t^{(k)}\|_\gamma^2 &= \sum_{i=1}^t \mathbb{E}\|\nu_i \Pi_{i+1}^t (L_K - L_i) r_{i-1}^{(k-1)}\|_\gamma^2 \\ &\le \sum_{i=1}^t \nu_i^2 \|\Pi_{i+1}^t\|_{\mathcal{H}\to\mathcal{H}}^2 \mathbb{E}\|(L_K - L_i) r_{i-1}^{(k-1)}\|_\gamma^2 \\ &\le \sum_{i=1}^t \nu_i^2 \|\Pi_{i+1}^t\|_{\mathcal{H}\to\mathcal{H}}^2 \mathbb{E}\|L_i r_{i-1}^{(k-1)}\|_\gamma^2\end{aligned}$$

Using (F.25) and Lemma G.3, we have

$$\begin{aligned}\mathbb{E}\|r_t^{(k)}\|_\gamma^2 &\le \sum_{i=1}^t \nu_i^2 \|\Pi_{i+1}^t\|_{\mathcal{H}\to\mathcal{H}}^2 \mathbb{E}[\|L_i\|_{\mathcal{H}_\gamma\to\mathcal{H}_\gamma}^2 \|r_{i-1}^{(k-1)}\|_\gamma^2] \\ &\le \kappa^{4-2\gamma} A^2 \sum_{i=1}^t \nu_i^2 \prod_{j=i+1}^t (1-\nu_j\lambda_j)^2 \mathbb{E}\|r_{i-1}^{(k-1)}\|_\gamma^2.\end{aligned} \tag{F.28}$$

Plugging (F.27) into (F.28), we obtain

$$\begin{aligned}\mathbb{E}\|r_t^{(k)}\|_\gamma^2 &\le \kappa^{4-2\gamma} A^2 \sum_{i=1}^t \nu_i^2 \prod_{j=i+1}^t (1-\nu_j\lambda_j)^2 C_{k-1}(i+t_0)^{k(1-2\theta)} \\ &= \kappa^{4-2\gamma} A^2 C_{k-1} \sum_{i=1}^t a^2 (i+t_0)^{-2\theta} \prod_{j=i+1}^t (1-(j+t_0)^{-1})^2 (i+t_0)^{k(1-2\theta)} \\ &= a^2 \kappa^{4-2\gamma} A^2 C_{k-1} \frac{\sum_{i=1}^t (i+t_0)^{1+(k+1)(1-2\theta)}}{(t+t_0)^2}.\end{aligned} \tag{F.29}$$

Condition $1°$**:** If $k < \lfloor \frac{1}{2\theta-1} \rfloor$, then $1+(k+1)(1-2\theta) \ge 0$, we have

$$\begin{aligned}\sum_{i=1}^t (i+t_0)^{1+(k+1)(1-2\theta)} &\le \sum_{i=1}^t \int_i^{i+1} (x+t_0)^{1+(k+1)(1-2\theta)} dx = \int_1^{t+1} (x+t_0)^{1+(k+1)(1-2\theta)} dx \\ &\le \frac{(t+1+t_0)^{2+(k+1)(1-2\theta)}}{2+(k+1)(1-2\theta)}.\end{aligned}$$

Combining this inequality with (F.29), we obtain

$$\begin{aligned}\mathbb{E}\|r_t^{(k)}\|_\gamma^2 &\le a^2 \kappa^{4-2\gamma} A^2 C_{k-1} \frac{\sum_{i=1}^t (i+t_0)^{1+(k+1)(1-2\theta)}}{(t+t_0)^2} \\ &\le \frac{(t_0+2)^2}{(2+(k+1)(1-2\theta))(t_0+1)^2} a^2 \kappa^{4-2\gamma} A^2 C_{k-1} (t+1+t_0)^{(k+1)(1-2\theta)}.\end{aligned}$$

Thus, the inequality for $r_t^{(k)}$ (E.18) holds for $k > 0$ and $k < \lfloor \frac{1}{2\theta-1} \rfloor$.

Condition $2°$**:** If $k \geq \lfloor \frac{1}{2\theta-1} \rfloor$ and $k < \lfloor \frac{2}{2\theta-1} \rfloor$, then $1+(k+1)(1-2\theta) \in [-1, 0)$, we have

$$\begin{aligned}\sum_{i=1}^{t}(i+t_0)^{1+(k+1)(1-2\theta)} &\leq \sum_{i=1}^{t}\int_{i-1}^{i}(x+t_0)^{1+(k+1)(1-2\theta)}dx = \int_{0}^{t}(x+t_0)^{1+(k+1)(1-2\theta)}dx \\ &\leq \frac{(t+1+t_0)^{2+(k+1)(1-2\theta)}}{2+(k+1)(1-2\theta)}.\end{aligned}$$

The same as **Condition** $1°$, the inequality for $r_t^{(k)}$ (E.18) holds for $k \geq \lfloor \frac{1}{2\theta-1} \rfloor$ and $k < \lfloor \frac{2}{2\theta-1} \rfloor$.

- **Initial case II: For** $k = \lfloor \frac{2}{2\theta-1} \rfloor$**.**

Suppose the inequality for $k-1$ (i.e. $\lfloor \frac{2}{2\theta-1} \rfloor - 1$) holds

$$\mathbb{E}\|r_t^{(k-1)}\|_\gamma^2 \leq C_{k-1}(t+1+t_0)^{k(1-2\theta)}. \tag{F.30}$$

The same as (F.28), the following inequality is guaranteed by (F.30) (we omit some calculation here because the analysis is exactly the same as **Induction step I**)

$$\begin{aligned}\mathbb{E}\|r_t^{(k)}\|_\gamma^2 &\leq \kappa^{4-2\gamma}A^2\sum_{i=1}^{t}\nu_i^2\prod_{j=i+1}^{t}(1-\nu_j\lambda_j)^2\mathbb{E}\|r_{i-1}^{(k-1)}\|_\gamma^2 \\ &\leq \kappa^{4-2\gamma}A^2C_{k-1}\sum_{i=1}^{t}a^2(i+t_0)^{-2\theta}\prod_{j=i+1}^{t}(1-(j+t_0)^{-1})^2(i+t_0)^{k(1-2\theta)} \\ &= a^2\kappa^{4-2\gamma}A^2C_{k-1}\frac{\sum_{i=1}^{t}(i+t_0)^{1+(k+1)(1-2\theta)}}{(t+t_0)^2}.\end{aligned} \tag{F.31}$$

Since $k = \lfloor \frac{2}{2\theta-1} \rfloor$, we have $2+(k+1)(1-2\theta) < 0$. Therefore,

$$(i+t_0)^{1+(k+1)(1-2\theta)} \leq \int_{i-1}^{i}(x+t_0)^{1+(k+1)(1-2\theta)}dx = \frac{(i-1)^{2+(k+1)(1-2\theta)} - i^{2+(k+1)(1-2\theta)}}{(k+1)(2\theta-1)-2}$$

and we have

$$\begin{aligned}\mathbb{E}\|r_t^{(k)}\|_\gamma^2 &\leq a^2\kappa^{4-2\gamma}A^2C_{k-1}\frac{\sum_{i=1}^{t}\int_{i-1}^{i}(x+t_0)^{1+(k+1)(1-2\theta)}dx}{(t+t_0)^2} \\ &\leq a^2\kappa^{4-2\gamma}A^2C_{k-1}\frac{t_0^{2+(k+1)(1-2\theta)} - (t+t_0)^{2+(k+1)(1-2\theta)}}{(k+1)(2\theta-1)-2}(t+t_0)^{-2} \\ &\leq a^2\kappa^{4-2\gamma}A^2C_{k-1}\frac{(t_0+2)^2}{((k+1)(2\theta-1)-2)(t_0+1)^2}(t+1+t_0)^{-2}.\end{aligned}$$

Thus, the inequality for $r_t^{(k)}$ (E.18) holds for $k = \lfloor \frac{2}{2\theta-1} \rfloor$.

- **Induction step II: For** $k > \lfloor \frac{2}{2\theta-1} \rfloor$**.**

Suppose the inequality for $k-1$ holds, i.e.

$$\mathbb{E}\|r_t^{(k-1)}\|_\gamma^2 \le C_{k-1}(t+1+t_0)^{-2}. \tag{F.32}$$

The same as **Initial case II**, plugging in (F.32), the following inequality holds,

$$\begin{aligned}
\mathbb{E}\|r_t^{(k)}\|_\gamma^2 &\le \kappa^{4-2\gamma}A^2 \sum_{i=1}^{t} \nu_i^2 \prod_{j=i+1}^{t} (1-\nu_j\lambda_j)^2 \mathbb{E}\|r_{i-1}^{(k-1)}\|_\gamma^2 \\
&\le \kappa^{4-2\gamma}A^2 C_{k-1} \sum_{i=1}^{t} a^2(i+t_0)^{-2\theta} \prod_{j=i+1}^{t} (1-(j+t_0)^{-1})^2 (i+t_0)^{-2} \\
&\le a^2\kappa^{4-2\gamma}A^2 C_{k-1} \frac{t_0^{1-2\theta}-(t+t_0)^{1-2\theta}}{2\theta-1}(t+t_0)^{-2} \\
&\le a^2\kappa^{4-2\gamma}A^2 C_{k-1} \frac{(t_0+2)^2}{(2\theta-1)(t_0+1)^2}(t+1+t_0)^{-2}.
\end{aligned}$$

Thus, the inequality for $r_t^{(k)}$ (E.18) holds for $k > \lfloor \frac{2}{2\theta-1} \rfloor$.

F.6 Error Decomposition in the Proof of Lemma 4.18

In this section, we study the error decomposition of stochastic approximation sequences f_t generated by the Robbins-Monro algorithm (E.3):

$$f_t = f_{t-1} - \nu_t(A_t f_{t-1} - b_t), \tag{F.33}$$

where $A_t := L_t + \lambda_t I$ are stochastic operators and $b_t := y_t \phi_{u_t}$ are random variables in $\mathcal{H}$. Define their expectation

$$\bar{A}_t := \mathbb{E}A_t, \quad \bar{b} := \mathbb{E}b_t.$$

We introduce the basic decomposition for $f_t - f$ (corresponding to Lemma E.2), and demonstrate the martingale decomposition and the semi-stochastic decomposition involved in the proof of Lemma 4.18 (especially Lemma E.11).

F.6.1 Martingale Decomposition

The aim is to study the error decomposition for $f_t - f$. To do this, we first define the semi-stochastic population iteration g_t:

$$g_0 = f_0, \quad g_t = (I - \nu_t \bar{A}_t) g_{t-1} + \nu_t b_t. \tag{F.34}$$

Comparing (F.34) with (F.33), the semi-stochastic iteration replaces A_t in f_t by its expectation $\bar{A}_t$ and remains the second stochastic term b_t. This semi-stochastic iteration eliminates the randomness of A_t, and thus can be viewed as a population iteration of (F.33).

With this definition, the basic decomposition decomposes the error $f_t - f$ into three terms

$$f_t - f = (f_t - g_t) + (g_t - f_{\lambda_t}) + (f_{\lambda_t} - f), \tag{F.35}$$

where $f_t - g_t$ characterize the sampling error of the stochastic operator A_t, $g_t - f_{\lambda_t}$ denotes the difference between the online estimation function f_t of the kernel ridge regression $\arg\min_{h\in(\mathcal{H})^p} \mathbb{E}_{(u,y)\sim\rho}\|y - f(u)\|^2 + \lambda_t\|h\|_{\mathcal{H}}^2$ and the true solution f_{λ_t}, and $f_{\lambda_t} - f$ characterizes the error between the solution for the regularized kernel ridge regression and the true parametric function f (i.e. the solution for the unregularized kernel regression).

Lemma F.1. (Martingale decomposition) For all $s, t \in \mathbb{N}$, $t \geq s$,

$$g_t - f_{\lambda_t} = \Pi_{s+1}^t (g_s - f_{\lambda_s}) + \sum_{i=s+1}^{t} \nu_i \Pi_{i+1}^t (b_i - \bar{b}) - \sum_{i=s+1}^{t} \Pi_i^t (f_{\lambda_i} - f_{\lambda_{i-1}}), \tag{F.36}$$

where $f_\lambda = (L_K + \lambda I)^{-1} L_K f$ is the solution of kernel ridge regression and

$$\Pi_s^t = \prod_{i=s}^{t} (I - \nu_i \bar{A}_i), \qquad \bar{A}_i = L_K + \lambda_i I.$$

Specifically, when $s = 0$,

$$g_t - f_{\lambda_t} = \Pi_1^t (g_0 - f_{\lambda_0}) + \sum_{i=1}^{t} \nu_i \Pi_{i+1}^t (b_i - \bar{b}) - \sum_{i=1}^{t} \Pi_i^t (f_{\lambda_i} - f_{\lambda_{i-1}}). \tag{F.37}$$

Proof. The proof is direct and we omit it here. □

The operator Π_s^t in decomposition (F.37) is deterministic and $\{b_i - \bar{b}\}_{t\in\mathbb{N}}$ is a sequence of i.i.d random variables with zero mean, namely, $\mathbb{E}(b_i - \bar{b}) = 0$. Consequently, $\Pi_1^t(g_0 - f_{\lambda_0})$ and $\sum_{i=1}^t \Pi_i^t (f_{\lambda_i} - f_{\lambda_{i-1}})$ are deterministic and the randomness is contained in $\sum_{i=1}^t \nu_i \Pi_{i+1}^t (b_i - \bar{b})$. We have seem that positive-semidefiniteness and non-randomness of the operator Π_s^t play a key role in the proof of Lemma 4.18.

F.6.2 Semi-Stochastic Decomposition

To decompose $f_t - g_t$, we consider a sequence of semi-stochastic noise processes (Dieuleveut and Bach, 2016). To begin with, we observe that $f_t - g_t$ has a recursion structure

$$\begin{aligned} &f_0 - g_0 = 0, \\ &f_t - g_t = (I - \nu_t A_t)(f_{t-1} - g_{t-1}) + \nu_t (\bar{A}_t - A_t) g_{t-1}. \end{aligned} \tag{F.38}$$

Define a noise process $r_0^{(0)}$ as the semi-stochastic iteration of the above recursion

$$\begin{aligned} &r_0^{(0)} = f_0 - g_0, \\ &r_t^{(0)} = (I - \nu_t \bar{A}_t) r_{t-1}^{(0)} + \nu_t (\bar{A}_t - A_t) g_{t-1}, \end{aligned} \tag{F.39}$$

the error $f_t - g_t - r_t^{(0)}$ also has a recursion structure

$$\begin{aligned} &f_0 - g_0 - r_0^{(0)} = 0, \\ &f_t - g_t - r_t^{(0)} = (I - \nu_t A_t)(f_{t-1} - g_{t-1} - r_{t-1}^{(0)}) + \nu_t (\bar{A}_t - A_t) r_{t-1}^{(0)}. \end{aligned}$$

Repeat this procedure to define a sequence of noise process $\{r_t^{(k)}\}_{t\in\mathbb{N},k\in\mathbb{N}_0}$

$$\begin{aligned}
r_0^{(1)} &= 0, \quad r_t^{(1)} = (I-\nu_t\bar{A}_t)r_{t-1}^{(1)} + \nu_t(\bar{A}_t - A_t)r_{t-1}^{(0)},\\
r_0^{(2)} &= 0, \quad r_t^{(2)} = (I-\nu_t\bar{A}_t)r_{t-1}^{(2)} + \nu_t(\bar{A}_t - A_t)r_{t-1}^{(1)},\\
&\vdots\\
r_0^{(k)} &= 0, \quad r_t^{(k)} = (I-\nu_t\bar{A}_t)r_{t-1}^{(k)} + \nu_t(\bar{A}_t - A_t)r_{t-1}^{(k-1)},\\
&\vdots
\end{aligned} \tag{F.40}$$

the error $f_t - g_t - \sum_{j=1}^k r_t^{(j)}$ has a recursion structure

$$\begin{aligned}
&f_0 - g_0 - \sum_{j=0}^{k} r_0^{(j)} = 0,\\
&f_t - g_t - \sum_{j=0}^{k} r_t^{(j)} = (I-\nu_t A_t)(f_{t-1} - g_{t-1} - \sum_{j=0}^{k} r_{t-1}^{(j)}) + \nu_t(\bar{A}_t - A_t)r_{t-1}^{(k)}.
\end{aligned} \tag{F.41}$$

Lemma F.2. (Noise decomposition) For all $t\in\mathbb{N}$ and $i\le t$,

$$r_i^{(t)} = 0 \quad \text{and} \quad f_t - g_t = \sum_{j=0}^{(t-1)\vee 0} r_t^{(j)} \tag{F.42}$$

Proof. We use mathematical induction. For $t=0$, given $r_0^{(0)} = f_0 - g_0$ and $g_0 = f_0$, we have

$$r_0^{(0)} = f_0 - g_0 = 0.$$

For $t=1$, given $r_0^{(0)} = 0$ and $r_0^{(1)} = 0$, (F.40) indicates that

$$r_1^{(1)} = (I-\nu_1\bar{A}_1)r_0^{(1)} + \nu_1(\bar{A}_1 - A_1)r_0^{(0)}.$$

Moreover, using (F.41) for $k=0$, we have

$$f_1 - g_1 - r_1^{(0)} = (I-\nu_1 A_1)(f_0 - g_0 - r_0^{(0)}) + \nu_1(\bar{A}_1 - A_1)r_0^{(0)} = 0.$$

Therefore, we obtain $f_1 - g_1 = r_1^{(0)}$.

Assuming that the statements

$$f_t - g_t = \sum_{j=0}^{t-1} r_t^{(j)} \quad \text{and} \quad r_i^{(t)} = 0 \tag{F.43}$$

hold for any $i\le t$, where $t>1$. We are going to prove that (F.43) holds for $f_{t+1} - g_{t+1}$ and $r_i^{(t+1)}$.

Plugging $k = t$ into (F.41) and using (F.43), we have

$$
\begin{aligned}
f_{t+1} - g_{t+1} - \sum_{j=0}^{t} r_{t+1}^{(j)} &= (I - \nu_{t+1}A_{t+1})(f_t - g_t - \sum_{j=0}^{t} r_t^{(j)}) + \nu_{t+1}(\bar{A}_{t+1} - A_{t+1})r_t^{(t)} \\
&= (I - \nu_{t+1}A_{t+1})(f_t - g_t - \sum_{j=0}^{t-1} r_t^{(j)} - r_t^{(t)}) + \nu_{t+1}(\bar{A}_{t+1} - A_{t+1})r_t^{(t)} \\
&= (I - \nu_{t+1}A_{t+1})(0 - r_t^{(t)}) + \nu_{t+1}(\bar{A}_{t+1} - A_{t+1})r_t^{(t)} = 0.
\end{aligned}
$$

Moreover, plugging $k = t + 1$ into (F.40) and using (F.43), for any $i \le t + 1$, we obtain

$$
\begin{aligned}
r_i^{(t+1)} &= (I - \nu_i\bar{A}_i)r_{i-1}^{(t+1)} + \nu_i(\bar{A}_i - A_i)r_{i-1}^{(t)} \\
&= \sum_{j=1}^{i} \nu_j \prod_{s=j+1}^{i} (I - \nu_s\bar{A}_s)(\bar{A}_j - A_j)r_{j-1}^{(t)} \\
&= \sum_{j=1}^{i} \nu_j \prod_{s=j+1}^{i} (I - \nu_s\bar{A}_s)(\bar{A}_j - A_j)0 = 0.
\end{aligned}
$$

Thus, (F.43) holds for any $t \in \mathbb{N}$ and we finish the proof. □

Remark F.3. Lemma F.2 indicates that the error between the true iteration f_t and the semi-stochastic population iteration g_t can be decomposed by a finite sum of the noise process $r_t^{(k)}$. We visualize this theorem in the following tables. For Table 1, the i-th column denotes the outputs of semi-stochastic iteration and noise processes at iteration i, namely, g_i and $r_i^{(k)}$, the k-th row denotes that output of the noise processes $r_i^{(k-1)}$. For any $t \in \mathbb{N}, k \in \mathbb{N}_0$, define $C(t, -1) \coloneqq g_t$ and $C(t, k) \coloneqq r_t^{(k)}$ for convenience, then the (i, j) entry of Table 1 denotes $r_{j-1}^{(i-2)} = C(j-1, i-2)$. Recalling recalling (F.40), we have

$$
\begin{aligned}
C(t,k) = r_t^{(k)} &= (I - \nu_t\bar{A}_t)r_{t-1}^{(k)} + \nu_t(\bar{A}_t - A_t)r_{t-1}^{(k-1)} \\
&= \sum_{j=1}^{t} \nu_j \prod_{i=j+1}^{t} (I - \nu_i\bar{A}_i)(\bar{A}_j - A_j)r_{j-1}^{(k-1)} = \sum_{j=1}^{t} \nu_j \Pi_{j+1}^{t}(\bar{A}_j - A_j)r_{j-1}^{(k-1)}.
\end{aligned} \tag{F.44}
$$

Equation (F.44) implies that each entry $C(i, j) = r_i^{(j)}$ is comprised of entries in its upper left region ($r_t^{(k)}$ such that $t \le i$ and $j - i + t \le k \le j$) and $r_0^{(k)} = 0$ for all $k \in \mathbb{N}_0$. Therefore, the entries in the lower triangle of Table 1 are zero and for all iteration t, the first equation of Lemma F.2 holds, namely, $r_t^{(t)} = 0$.

The same as Table 1, we visualize $f_t - g_t - \sum_{j=0}^{k} r_t^{(j)}$ in Table 2. For any $t \in \mathbb{N}, k \in \mathbb{N}_0$, define $D(t, -2) \coloneqq g_t$, $D(t, -1) \coloneqq f_t - g_t$, and $D(t, k) \coloneqq f_t - g_t - \sum_{j=0}^{k} r_t^{(j)}$ for convenience, then the (i, j)

Table 1: Visualization of $r_i^{(k)}$

$i \in \mathbb{N}$	0	1	2	3	$\cdots$	t	$t+1$	$\cdots$
g_i	$C(0,-1)$	$C(1,-1)$	$C(2,-1)$	$C(3,-1)$	$\cdots$	$C(t,-1)$	$C(t+1,-1)$	$\cdots$
$r_i^{(0)}$	0	$C(1,0)$	$C(2,0)$	$C(3,0)$	$\cdots$	$C(t,0)$	$C(t+1,0)$	$\cdots$
$r_i^{(1)}$	0	0	$C(2,1)$	$C(3,1)$	$\cdots$	$C(t,1)$	$C(t+1,1)$	$\cdots$
$r_i^{(2)}$	0	0	0	$C(3,2)$	$\cdots$	$C(t,2)$	$C(t+1,2)$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
$r_i^{(t-1)}$	0	0	0	0	$\cdots$	$C(t,t-1)$	$C(t+1,t-1)$	$\cdots$
$r_i^{(t)}$	0	0	0	0	$\cdots$	0	$C(t+1,t)$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\ddots$

entry of Table 2 denotes $f_{j-1} - g_{j-1} - \sum_{s=0}^{i-3} r_{j-1}^{(s)} = D(j-1, i-3)$. Recalling (F.41), we have

$$\begin{aligned} D(t,k) &= f_t - g_t - \sum_{j=0}^{k} r_t^{(j)} \\ &= (I - \nu_t A_t)(f_{t-1} - g_{t-1} - \sum_{j=0}^{k} r_{t-1}^{(j)}) + \nu_t(\bar{A}_t - A_t) r_{t-1}^{(k)} \qquad \text{(F.45)} \\ &= \sum_{j=1}^{t} \nu_j \prod_{i=j+1}^{t} (I - \nu_i A_i)(\bar{A}_j - A_j) r_{j-1}^{(k)}. \end{aligned}$$

Again, each entry is comprised of entries in its upper left region and the entries in the lower triangle of Table 2 are zero. Therefore, for all iteration t, the second equation of Lemma F.2 holds. Namely, $f_t - g_t - \sum_{j=0}^{(t-1)\vee 0} r_t^{(j)} = 0$ for all $t \in \mathbb{N}$.

G Technical Lemmas

Lemma G.1. Consider a sequence $G_t \geq 0$ for $t \in \mathbb{N}$. Suppose that for some constants $a \in (1,2), b > 0$ and constant $t_0 \geq 0$, the sequence G_t satisfies

$$G_{t+1} \leq (1 - (t+t_0)^{-1}) G_t + b(t+t_0)^{-a}, \qquad \text{(G.1)}$$

then the estimate holds for all $t \in \mathbb{N}$

$$G_t \lesssim \mathcal{O}(t^{1-a}). \qquad \text{(G.2)}$$

Proof. Using (G.1) recursively, we have

$$\begin{aligned} G_{t+1} &\leq (1 - (t+t_0)^{-1}) G_t + b(t+t_0)^{-a} \\ &= \prod_{i=0}^{t} (1 - (i+t_0)^{-1}) G_0 + \sum_{i=0}^{t} b \prod_{j=i+1}^{t} (1 - (j+t_0)^{-1})(i+t_0)^{-a}. \end{aligned}$$

Table 2: Visualization of $f_i - g_i - \sum_{j=0}^{k} r_i^{(j)}$

$i \in \mathbb{N}$	0	1	2	3	$\cdots$	t	$t+1$	$\cdots$
g_i	$D(0,-2)$	$D(1,-2)$	$D(2,-2)$	$D(3,-2)$	$\cdots$	$D(t,-2)$	$D(t+1,-2)$	$\cdots$
$f_i - g_i$	0	$D(1,-1)$	$D(2,-1)$	$D(3,-1)$	$\cdots$	$D(t,-1)$	$D(t+1,-1)$	$\cdots$
$f_i - g_i - r_i^{(0)}$	0	0	$D(2,0)$	$D(3,0)$	$\cdots$	$D(t,0)$	$D(t+1,0)$	$\cdots$
$f_i - g_i - \sum_{j=0}^{1} r_i^{(j)}$	0	0	0	$D(3,1)$	$\cdots$	$D(t,1)$	$D(t+1,1)$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
$f_i - g_i - \sum_{j=0}^{t-2} r_i^{(j)}$	0	0	0	0	$\cdots$	$D(t,t-2)$	$D(t+1,t-2)$	$\cdots$
$f_i - g_i - \sum_{j=0}^{t-1} r_i^{(j)}$	0	0	0	0	$\cdots$	0	$D(t+1,t-1)$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\ddots$

Define $X_i^j = \prod_{k=i}^{j}(1-(k+t_0)^{-1})$ and rewrite the inequality above

$$G_{t+1} \leq X_0^t G_0 + \sum_{i=0}^{t} b X_{i+1}^t (i+t_0)^{-a}. \tag{G.3}$$

- **Upper bound of X_i^j.**

Using the inequality $1+x \leq e^x$, we obtain

$$X_i^j = \prod_{k=i}^{j}(1-(k+t_0)^{-1}) \leq e^{-\sum_{k=i}^{j}(k+t_0)^{-1}}.$$

Moreover, using the fact that $(k+t_0)^{-1} \leq \int_{k-1+t_0}^{k+t_0} x^{-1}dx$ for any $k \geq 1$, we have

$$X_i^j \leq e^{-\sum_{k=i}^{j}\int_{k-1+t_0}^{k+t_0} x^{-1}dx} = e^{-(\log(j+t_0)-\log(i-1+t_0))} = \frac{i-1+t_0}{j+t_0}.$$

Plugging this bound into (G.3),

$$\begin{aligned} G_{t+1} &\leq \frac{t_0-1}{t+t_0}G_0 + \sum_{i=0}^{t} b\frac{i+t_0}{t+t_0}(i+t_0)^{-a} \\ &= (t+t_0)^{-1}\left((t_0-1)G_0 + \sum_{i=0}^{t} b(i+t_0)^{1-a}\right). \end{aligned} \tag{G.4}$$

Since $1-a \in [-1,0)$, therefore $(i+t_0)^{1-a} \leq \int_{i-1}^{i}(x+t_0)^{1-a}dx$, we further obtain

$$\begin{aligned} G_{t+1} &\leq (t+t_0)^{-1}\left((t_0-1)G_0 + b\sum_{i=0}^{t}\int_{i-1}^{i}(x+t_0)^{1-a}dx\right) \\ &= (t+t_0)^{-1}\left((t_0-1)G_0 + b\frac{(t+t_0)^{2-a}-(t_0-1)^{2-a}}{2-a}\right) \\ &\lesssim \mathcal{O}(t^{1-a}). \end{aligned}$$

□

Lemma G.2. Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$, let $\mathcal{H}$ be an RKHS on $\mathcal{X}$ with respect to a bounded Mercer kernel K (for all $x \in \mathcal{X}$, $K(x,x) \le \kappa^2$) and a probability measure $\rho_{\mathcal{X}}$. For any $\gamma \ge 0$ and $f \in \mathcal{H}$, the equality holds

$$\|f\|_\gamma^2 = \|L_K^{(1-\gamma)/2} f\|_{\mathcal{H}}^2. \tag{G.5}$$

Proof. For all $f = \sum_{i=1}^\infty a_i \mu_i^{1/2} e_i \in \mathcal{H}$, the γ-norm is

$$\|f\|_\gamma^2 = \|\sum_{i=1}^\infty a_i \mu_i^{1/2} e_i\|_\gamma^2 = \sum_{i=1}^\infty a_i^2 \mu_i^{1-\gamma}.$$

Besides, since $L_K^{(1-\gamma)/2} f = \sum_{i=1}^\infty a_i \mu_i^{(2-\gamma)/2} e_i$, we obtain

$$\|L_K^{(1-\gamma)/2} f\|_{\mathcal{H}}^2 = \|\sum_{i=1}^\infty a_i \mu_i^{(2-\gamma)/2} e_i\|_{\mathcal{H}}^2 = \sum_{i=1}^\infty a_i \mu_i^{1-\gamma} = \|f\|_\gamma^2.$$

□

Lemma G.3. Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$, let $\mathcal{H}$ be an RKHS on $\mathcal{X}$ with respect to a bounded Mercer kernel K (for all $x \in \mathcal{X}$, $K(x,x) \le \kappa^2$) and a probability measure $\rho_{\mathcal{X}}$. Let $\{\nu_t\}_{t\in\mathbb{N}}$ and $\{\lambda_t\}_{t\in\mathbb{N}}$ be positive sequences satisfied $\lim_{t\to\infty} \nu_t = 0, \lim_{t\to\infty} \lambda_t = 0$ and

$$\frac{1-\nu_t\lambda_t}{\nu_t} \ge \kappa^2$$

holds for all $t \in \mathbb{N}$. For any $i, j \in \mathbb{N}$, the following inequality holds

$$\|\prod_{k=i}^{j}(I - \nu_k(L_K + \lambda_k I))\|_{\mathcal{H}\to\mathcal{H}} \le \prod_{k=i}^{j}(1 - \nu_k\lambda_k). \tag{G.6}$$

Proof. By the definition of spectral norm,

$$\|\prod_{k=i}^{j}(I - \nu_k(L_K + \lambda_k I))\|_{\mathcal{H}\to\mathcal{H}} \le \prod_{k=i}^{j}\|(I - \nu_k(L_K + \lambda_k I))\|_{\mathcal{H}\to\mathcal{H}}.$$

Note that $L_K|_{\mathcal{H}}$ and I are compact, self-adjoint, and positive-semidefinite operators. By the spectral representation of the compact operator, we obtain

$$L_K = \sum_{i=1}^\infty \mu_i^{1/2} \langle e_i, \cdot\rangle_{\mathcal{L}_{\rho_{\mathcal{X}}}^2} \mu_i^{1/2} e_i = \sum_{i=1}^\infty \mu_i \langle \mu_i^{1/2} e_i, \cdot\rangle_{\mathcal{H}} \mu_i^{1/2} e_i$$

and

$$I = \sum_{i=1}^\infty \langle e_i, \cdot\rangle_{\mathcal{L}_{\rho_{\mathcal{X}}}^2} e_i = \sum_{i=1}^\infty \langle \mu_i^{1/2} e_i, \cdot\rangle_{\mathcal{H}} \mu_i^{1/2} e_i.$$

Therefore, for any $k \in \mathbb{N}$,

$$\begin{aligned}
\|(I - \nu_k(L_K + \lambda_k I))\|_{\mathcal{H}\to\mathcal{H}} &= \|(1 - \nu_k\lambda_k)I - \nu_k L_K\|_{\mathcal{H}\to\mathcal{H}} \\
&= \|(1 - \nu_k\lambda_k)\sum_{i=1}^{\infty}\langle \mu_i^{1/2} e_i, \cdot\rangle_{\mathcal{H}}\mu_i^{1/2} e_i - \nu_k \sum_{i=1}^{\infty}\mu_i\langle \mu_i^{1/2} e_i, \cdot\rangle_{\mathcal{H}}\mu_i^{1/2} e_i\|_{\mathcal{H}\to\mathcal{H}} \\
&= \|\sum_{i=1}^{\infty}(1 - \nu_k(\lambda_k + \mu_i))\langle \mu_i^{1/2} e_i, \cdot\rangle_{\mathcal{H}}\mu_i^{1/2} e_i\|_{\mathcal{H}\to\mathcal{H}} \\
&= \sup_{i\in\mathbb{N}} |1 - \nu_k(\lambda_k + \mu_i)|.
\end{aligned}$$

Given the spectral representation $K = \sum_{i=1}^{\infty} \mu_i e_i \otimes e_i$, we have

$$\mu_i \le \sum_{i=1}^{\infty}\mu_i = \sum_{i=1}^{\infty}\int_{\mathcal{X}} \mu_i e_i^2(x) d\rho_{\mathcal{X}} \le \int_{\mathcal{X}}\sum_{i=1}^{\infty}\mu_i e_i^2(x) d\rho_{\mathcal{X}} \le \sup_{x\in\mathcal{X}}\sum_{i=1}^{\infty}\mu_i e_i^2(x) \le \sup_{x\in\mathcal{X}} K(x, x)$$

Therefore,

$$\mu_i \le \sup_{x\in\mathcal{X}} K(x, x) \le \kappa^2, \tag{G.7}$$

combining with the fact that $\kappa^2 \le (1 - \nu_t\lambda_t)/\nu_t$, we obtain

$$\mu_i \le \kappa^2 \le \frac{1 - \nu_t\lambda_t}{\nu_t},$$

namely, $1 - \nu_k(\lambda_k + \mu_i)$ is non-negative. Moreover, since L_K is positive-semidefinite, its eigenvalues $\mu_i \ge 0$ holds for any $i \in \mathbb{N}$. As a result, we have

$$\sup_{i\in\mathbb{N}} |1 - \nu_k(\lambda_k + \mu_i)| = \sup_{i\in\mathbb{N}}(1 - \nu_k(\lambda_k + \mu_i)) \le 1 - \nu_k\lambda_k.$$

Finally, we obtain the spectral norm bound

$$\|\prod_{k=i}^{j}(I - \nu_k(L_K + \lambda_k I))\|_{\mathcal{H}\to\mathcal{H}} \le \prod_{k=i}^{j}(1 - \nu_k\lambda_k).$$

□

Lemma G.4. Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$, let $\mathcal{H}$ be an RKHS on $\mathcal{X}$ with respect to a bounded Mercer kernel K (for all $x \in \mathcal{X}$, $K(x, x) \le \kappa^2$) and a probability measure $\rho_{\mathcal{X}}$. Let ϕ be the feature map of K, for all $\gamma \in [0, 1]$ and $x \in \mathcal{X}$, the γ-norm of ϕ_x is bounded

$$\|\phi_x\|_\gamma \le \kappa^{2-\gamma}. \tag{G.8}$$

Suppose that $K^\gamma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the kernel associated with $\mathcal{H}^\gamma$ and let $\phi^\gamma : \mathcal{X} \to \mathcal{H}^\gamma$ be the feature map of K^γ. If the embedding property (Assumption 4.14) holds for $\alpha \le \gamma$ (i.e. $K^\gamma(x, x) \le A^2, \forall x \in \mathcal{X}$). For any $x \in \mathcal{X}$,

$$\|\phi_x^\gamma\|_\gamma \le A. \tag{G.9}$$

Proof. Using the spectral representation for kernel K, we obtain

$$K = \sum_{i=1}^{\infty} \mu_i e_i \otimes e_i, \qquad \phi_x = \sum_{i=1}^{\infty} \mu_i^{1/2} e_i(x) \mu_i^{1/2} e_i.$$

Therefore, the γ-norm of ϕ_x has the expression

$$\begin{aligned}\|\phi_x\|_\gamma^2 &= \|\sum_{i=1}^{\infty} \mu_i^{1/2} e_i(x) \mu_i^{1/2} e_i\|_\gamma^2 = \sum_{i=1}^{\infty} \mu_i^{2-\gamma} e_i^2(x) \\ &\leq \sup_{i\in\mathbb{N}} \mu_i^{1-\gamma} \sum_{i=1}^{\infty} \mu_i e_i^2(x) = \sup_{i\in\mathbb{N}} \mu_i^{1-\gamma} K(x,x) \\ &\leq \sup_{i\in\mathbb{N}} \mu_i^{1-\gamma} \kappa^2.\end{aligned}$$

Plugging in (G.7), we have

$$\|\phi_x\|_\gamma^2 \leq \sup_{i\in\mathbb{N}} \mu_i^{1-\gamma} \kappa^2 \leq \kappa^{2(2-\gamma)}.$$

Again, using the spectral representation for kernel K^γ, we obtain

$$K^\gamma = \sum_{i=1}^{\infty} \mu_i^\gamma e_i \otimes e_i, \qquad \phi_x^\gamma = \sum_{i=1}^{\infty} \mu_i^{\gamma/2} e_i(x) \mu_i^{\gamma/2} e_i.$$

Using the fact $\sup_{x\in\mathcal{X}} K^\gamma(x,x) \leq A^2$, we have

$$\begin{aligned}\|\phi_x^\gamma\|_\gamma^2 &= \|\sum_{i=1}^{\infty} \mu_i^{\gamma/2} e_i(x) \mu_i^{\gamma/2} e_i\|_\gamma^2 = \sum_{i=1}^{\infty} \mu_i^\gamma e_i^2(x) \\ &= K^\gamma(x,x) \leq A^2.\end{aligned}$$

□

Lemma G.5. Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$, let $\mathcal{H}$ be an RKHS on $\mathcal{X}$ with respect to a bounded Mercer kernel K (for all $x \in \mathcal{X}$, $K(x,x) \leq \kappa^2$) and a probability measure $\rho_\mathcal{X}$. For any independent random variables sequence $\{u_t\}_{t\in\mathbb{N}}$, let

$$\phi_t = \phi_{u_t} \quad \text{and} \quad y_t = f(u_t) + \epsilon_t,$$

where $\{\epsilon_t\}_{t\in\mathbb{N}}$ are independent noise term (and independent of $\{u_t\}_{t\in\mathbb{N}}$) with uniform norm bound σ^2 and $f \in \mathcal{H}^\beta$. If the embedding property (Assumption 4.14) holds for $\alpha < \beta$ (i.e. $K^\beta(x,x) \leq A^2, \forall x \in \mathcal{X}$). For any $\gamma \in [\alpha, \beta)$, the following uniform γ-norm bound for sequence $\{y_t\phi_t\}_{t\in\mathbb{N}}$ holds

$$\mathbb{E}\|y_t\phi_t\|_\gamma^2 \leq \kappa^{2(2-\gamma)}(A^2\|f\|_\beta^2 + \sigma^2). \tag{G.10}$$

Proof. Using Cauchy inequality,

$$\begin{aligned}\mathbb{E}\|y_t\phi_t\|_\gamma^2 &= \mathbb{E}(|y_t|\|\phi_t\|_\gamma)^2 \\ &\leq (\mathbb{E}y_t^2)(\mathbb{E}\|\phi_t\|_\gamma^2) = (\mathbb{E}(f(u_t)+\epsilon_t)^2)(\mathbb{E}\|\phi_t\|_\gamma^2) \\ &\overset{(a)}{=} \underbrace{(\mathbb{E}(f^2(u_t)+\epsilon_t^2))}_{\text{(I)}} \underbrace{(\mathbb{E}\|\phi_t\|_\gamma^2)}_{\text{(II)}},\end{aligned} \tag{G.11}$$

where (a) makes use of the independence between u_t and the noise term ϵ_t, and uses the fact that $\mathbb{E}\epsilon_t = 0$.

- **Upper bound of (I).**

For any $t \in \mathbb{N}$, define $\phi_t^\beta = \phi_{u_t}^\beta \in \mathcal{H}^\beta$. Since $f \in \mathcal{H}^\beta$, by the reproducing property, we have $f(u_t) = \langle f, \phi_t^\beta \rangle_\beta$ and

$$\mathbb{E}f^2(u_i) = \mathbb{E}\langle f, \phi_t^\beta \rangle_\beta^2 \leq \|f\|_\beta^2 \mathbb{E}\|\phi_t^\beta\|_\beta^2 \leq \|f\|_\beta^2 \sup_{u_t \in \mathcal{X}} \|\phi_t^\beta\|_\beta^2.$$

The embedding property implies that $\sup_{u_t \in \mathcal{X}} \|\phi_t^\beta\|_\beta^2 \leq A^2$, therefore,

$$\mathbb{E}f^2(u_i) \leq A^2 \|f\|_\beta^2.$$

We further obtain

$$(\mathrm{I}) = \mathbb{E}(f^2(u_t) + \epsilon^2) \leq A^2 \|f\|_\beta^2 + \sigma^2.$$

- **Upper bound of (II).**

Using Lemma G.4, we have

$$(\mathrm{II}) = \mathbb{E}\|\phi_t\|_\gamma^2 \leq \sup_{u_t \in \mathcal{X}} \|\phi_t\|_\gamma^2 \leq \kappa^{2(2-\gamma)}.$$

Combining the upper bounds for (I) and (II), we obtain the upper bound for (G.11)

$$\mathbb{E}\|y_t \phi_t\|_\gamma^2 \leq \kappa^{2(2-\gamma)}(A^2 \|f\|_\beta^2 + \sigma^2).$$

□

Table 3: Table of Notation

n	number of agents ($i \in [n]$)
x_i	action of agent i
x	joint action $x = (x_1, x_2, \cdots, x_n)$
x_{-i}	joint action of all agents except i
$\mathcal{X}_i$	action set of agent i
$\mathcal{X}$	joint action set $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$
d_i	dimension of the action set $\mathcal{X}_i$
d	dimension of the joint action set $\mathcal{X}$
$\mathcal{L}_i(x)$	utility function of agent i
$\ell_i(x, z_i)$	loss function of agent i
z_i	data observed by agent i
$\mathcal{D}_i(x)$	decision-dependent distribution of agent i
$\mathcal{Z}_i$	sample space of z_i
p	dimension of $\mathcal{Z}_i$
$\nabla_i \mathcal{L}_i(x)$	individual gradient of agent i
$H(x)$	gradient of the decision-dependent game
x^*	Nash equilibrium $x^* = (x_1^*, x_2^*, \cdots, x_n^*)$
τ	strongly monotone parameter
$\mathscr{F}$	function class of parametric model
$f_i(x)$	parametric function of agent i
ϵ_i	noise term the parametric model of agent i
σ^2	variance bound of ϵ_i
$\mathcal{P}_i$	distribution of ϵ_i
$\nabla_i \ell_i(x, z_i)$	gradient of $\ell_i(x, z_i)$ to x_i
$\nabla_{z_i} \ell_i(x, z_i)$	gradient of $\ell_i(x, z_i)$ to z_i
$\widehat{\nabla}_i \mathcal{L}_i(x)$	unbiased estimator of $\nabla_i \mathcal{L}_i(x)$
$\widehat{H}(x)$	unbiased estimator of $H(x)$
$\rho_{\mathcal{X}}$	sampling distribution on $\mathcal{X}$
ρ_i	distribution on $\mathcal{X} \times \mathcal{Z}_i$ induced by $x \sim \rho_{\mathcal{X}}$ and $z_i \sim \mathcal{D}_i(x)$
x_i^t	action of agent i at iteration t
x^t	joint action $x = (x_1^t, x_2^t, \cdots, x_n^t)$ at iteration t
x_{-i}^t	joint action of all agents except i at iteration t
A_i	parametric function of agent i in the linear setting
A_i^t	estimation of A_i at iteration t
A_{ii}, A_{ii}^t	submatrix of A_i, A_i^t with columns indexed by agent i
u_i^t, y_i^t	samples for estimation update at iteration t
z_i^t	samples for projected gradient step at iteration t
ν_t	gradient size of estimation update at iteration t
η_t	gradient size of projected gradient step at iteration t
K	Mercer kernel on $\mathcal{X} \times \mathcal{X}$
$\mathcal{H}$	RKHS induced by kernel K and measure $\rho_{\mathcal{X}}$
ϕ	feature map of K ($\phi_x = K(\cdot, x)$)
λ_t	regularization term of estimation update at iteration t
$\partial_i \phi$	partial derivative of ϕ to x_i
L	Lipschitz parameter of $H(x)$
δ, ζ	parameters about Lipschitz continuity defined in Assumptions 4.3, 4.4
l_1, l_2, R	parameters about isotropic defined in Assumption 4.8
t_0	sufficiently large constant to set the gradient steps
μ_i	eigenvalues of kernel K
e_i	eigenfunctions of kernel K
$\mathcal{H}^\alpha$	α-power space of $\mathcal{H}$
K^α	kernel of $\mathcal{H}^\alpha$
ϕ^α	feature map of K^α
κ^2, A^2	upper bound of K, K^α
α, β	parameters about source condition and embedding property defined in Assumptions 4.14, 4.15
ξ	upper bound of $\partial_i \phi^\alpha$ in α-power norm