Appl. Math. J. Chinese Univ.
2024, 39(1): 1-23

Differentially private SGD with random features

WANG Yi-guang[1] GUO Zheng-chu[2,*]

Abstract. In the realm of large-scale machine learning, it is crucial to explore methods for reducing computational complexity and memory demands while maintaining generalization performance. Additionally, since the collected data may contain some sensitive information, it is also of great significance to study privacy-preserving machine learning algorithms. This paper focuses on the performance of the differentially private stochastic gradient descent (SGD) algorithm based on random features. To begin, the algorithm maps the original data into a low-dimensional space, thereby avoiding the traditional kernel method for large-scale data storage requirement. Subsequently, the algorithm iteratively optimizes parameters using the stochastic gradient descent approach. Lastly, the output perturbation mechanism is employed to introduce random noise, ensuring algorithmic privacy. We prove that the proposed algorithm satisfies the differential privacy while achieving fast convergence rates under some mild conditions.

§1 Introduction

The stochastic gradient descent (SGD) method is one of the most popular methods to handle large scale datasets. Compared to the gradient descent method, it demonstrates comparable performance while significantly reducing the computational burden at each iteration. The reduction in computational burden is achieved by computing the gradient using only a single training example instead of traversing through all training examples. Recently, it has gained significant attention and popularity due to its wide applications in training neural networks [4, 27]. There is a large literature in learning theory on the analysis of performance of SGD, e.g., see [4, 7, 9, 15, 17–19, 32] and references therein. Moreover, since the training examples may contain some sensitive information, such as medical data, financial data and web search histories. Many machine learning models inadvertently reveal sensitive information during the training process, thus violating privacy even if private details are removed from the data. For instance, even after the removal of names, genders, and addresses, re-identification remains possible since the remaining features may still create a unique signature.

Therefore, it is also of great significance to study privacy-preserving machine learning algorithms. In this paper, we use (ϵ, δ_p)-differential privacy to measure the privacy which is

Received: 2023-06-21. Revised: 2023-10-08.
MR Subject Classification: 68Q32, 68T05.
Keywords: learning theory, differential privacy, stochastic gradient descent, random features, reproducing kernel Hilbert spaces.
Digital Object Identifier(DOI): https://doi.org/10.1007/s11766-024-5037-0.
The work is supported by Zhejiang Provincial Natural Science Foundation of China (LR20A010001) and National Natural Science Foundation of China(12271473 and U21A20426).
*Corresponding author.

proposed in [10] and has recently received a significant amount of attention due to its resilience against known attacks and broad applicability. The mathematical description of differential privacy tells us that a statistical procedure satisfies (ϵ, δ_p)-differential privacy if changing a single data point does not affect the output distribution too much. This property makes it difficult for adversaries to infer the value of any specific data point from the algorithm's output [6, 11].

In this paper, we are interested in studying differentially private SGD based on random features. Random features [21] are proposed to overcome the memory bottleneck that prevents large scale applications of kernel methods [1, 23]. This breakthrough has paved the way for the widespread adoption of kernel methods in a variety of large-scale learning tasks. We consider the least square regression problem, which aims at learning a functional relation f from training examples that can make predictions for new observations. Let $\mathcal{X}$ be a compact metric space and $\mathcal{Y} \subset \mathbb{R}$, ρ is a Borel probability distribution on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. For a function $f : \mathcal{X} \mapsto \mathcal{Y}$, and $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $f(x)$ represents the prediction of y based on x, the prediction error is measured by the least square error $(f(x) - y)^2$. The regression problem aims at estimating an ideal function which minimizes the following *generalization error*

$$\mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} (f(x) - y)^2 d\rho \tag{1}$$

where the minimization is intended over all measurable functions. The *regression function* is the minimizer of the generalization error $\mathcal{E}(f)$ and is given by

$$f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x), \quad x \in \mathcal{X} \tag{2}$$

where $\rho(\cdot|x)$ is the conditional distribution of ρ at x. Since ρ is unknown, we learn a function f from samples $(x_i, y_i)_{i=1}^n$ drawn from ρ independently to approximate the target function f_ρ. In this paper, we are interested in functions of the form

$$f(x) = \langle w, \phi_M(x)\rangle, \quad \forall x \in \mathcal{X}, \tag{3}$$

here $\langle\cdot,\cdot\rangle$ denotes the inner product in $\mathbb{R}^M$, $w \in \mathbb{R}^M$, $M \in \mathbb{N}$ and the *random feature map* $\phi_M : \mathcal{X} \mapsto \mathbb{R}^M$ is defined as

$$\phi_M(x) = \frac{1}{\sqrt{M}}(\psi(x, \nu_1), \cdots, \psi(x, \nu_M))^\top, \quad \forall x \in \mathcal{X} \tag{4}$$

where $\nu_1, \cdots, \nu_M \in \Omega$ are drawn independently according to some distribution π. We assume the function $\psi : \mathcal{X} \times \Omega \to \mathbb{R}$ is continuous and there exists constant $\kappa \geq 1$ such that $|\psi(x, \nu)| \leq \kappa$ for any $x \in \mathcal{X}, \nu \in \Omega$.

The coefficient w can be learned based on the examples $\{(x_i, y_i)\}_{i=1}^n$ by the following SGD method, $\hat{w}_0 = 1$, and for $1 \leq t \leq T$,

$$\hat{w}_{t+1} = \hat{w}_t - \eta(\langle \hat{w}_t, \phi_M(x_{i_t})\rangle - y_{i_t})\phi_M(x_{i_t}), \tag{5}$$

here $\eta > 0$ is the step size, and i_t is drawn uniformly from $\{1, 2, \cdots, n\}$.

In this paper, we develop SGD based on random features for approximating f_ρ while guaranteeing differential privacy. To this end, we use the output perturbation mechanism [6, 10, 11] based on the sensitivity method to achieve the differential privacy, in which random noise is added to the SGD output iterates $\hat{w}_{T+1}$, therefore, the differentially private estimator is of the form $\hat{f}_{priv}(\cdot) = \langle \hat{w}_{priv}, \phi_M(\cdot)\rangle$ after T iterations. See details in algorithm 1. And the performance of algorithm 1 can be measured by the following *excess generalization error*

$$\mathcal{E}(\hat{f}_{priv}) - \mathcal{E}(f_\rho). \tag{6}$$

For a randomized learning algorithm $\mathcal{A} : \mathcal{Z}^n \mapsto \mathbb{R}^d$, let $\mathcal{A}(\mathcal{S})$ denotes the model produced by running $\mathcal{A}$ over the training dataset $\mathcal{S}$. Two datasets $\mathcal{S}$ and $\mathcal{S}'$ are called neighboring datasets

Algorithm 1 Differentially private SGD based on random features

Input: Data $\{(x_i, y_i)\}_{i=1}^n$, sampling function $\pi(\nu)$, random feature map ϕ_M, parameters $M, \epsilon, \delta_p, T, \eta$

Output: The predictor $\hat{f}_{priv}$

Draw $\nu_j, j = 1, 2, \cdots, M$ according to $\pi(\nu)$.
Set $\phi_M(x_i) = \sqrt{1/M}[\psi(x_i, \nu_1), \psi(x_i, \nu_2), \cdots, \psi(x_i, \nu_M)]^\top$ for each i
Set $\hat{w}_1 = 0$
for $t = 1$ to T **do**
  Sample $i_t \sim Unif[n]$
  $\hat{w}_{t+1} = \hat{w}_t - \eta(\langle \hat{w}_t, \phi_M(x_{i_t})\rangle - y_{i_t})\phi_M(x_{i_t})$
end for
Let $\Delta = \Delta_{SGD}(\delta/2)$
Compute $\sigma^2 = \frac{2\log(2.5/\delta)\Delta^2}{\epsilon^2}$
$\hat{w}_{priv} = \hat{w}_{T+1} + b$, where $b \sim \mathcal{N}(0, \sigma^2 I_M)$
$\hat{f}_{priv}(x) = \langle \hat{w}_{priv}, \phi_M(x)\rangle, \phi_M(x) = \sqrt{1/M}[\psi(x, \nu_1), \psi(x, \nu_2), \cdots, \psi(x, \nu_M)]^\top$

if they differ by a single datum, that is, $\mathcal{S}$ and $\mathcal{S}'$ have $n-1$ points (x_i, y_i) in common which is denoted by $\mathcal{S} \simeq \mathcal{S}'$. The privacy measurement used in our paper is the (ϵ, δ_p)-differential privacy, which defines a notion of privacy for a randomized algorithm $\mathcal{A}(\mathcal{S})$. In this paper, $\mathcal{A}$ is the SGD algorithm (5), and $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$, then $\mathcal{A}(\mathcal{S}) = \hat{w}_{T+1}$.

Definition 1 (Differential Privacy(DP), [10]). *We say a randomized algorithm $\mathcal{A}$ satisfies (ϵ, δ_p)-DP if, for any two neighboring datasets $\mathcal{S}$ and $\mathcal{S}'$ and any event E in the output space of $\mathcal{A}$, there holds*

$$\mathbb{P}(\mathcal{A}(\mathcal{S}) \in E) \leq e^\epsilon \mathbb{P}(\mathcal{A}(\mathcal{S}') \in E) + \delta_p \tag{7}$$

In particular, we call it satisfies ϵ-DP if $\delta_p = 0$.

The output perturbation mechanism is conducted by adding noise with a particular distribution to the output of $\mathcal{A}(\mathcal{S})$, which has the effect of masking the effect of any kind of particular data point [6]. This is usually called the sensitivity method. The ℓ_2-sensitivity of an algorithm (function) $\mathcal{A}$ is defined as follows.

Definition 2 (ℓ_2-sensitivity, [29]). *The ℓ_2-sensitivity of an algorithm (function) $\mathcal{A} : \mathcal{Z}^n \mapsto \mathbb{R}^d$ is defined as $\Delta = \sup_{\mathcal{S}\simeq\mathcal{S}'} \|\mathcal{A}(\mathcal{S}) - \mathcal{A}(\mathcal{S}')\|_2$, where $\|\cdot\|_2$ denotes the Euclidean norm, $\mathcal{S}$ and $\mathcal{S}'$ are neighboring datasets.*

The following Gaussian mechanism [29] generates a (ϵ, δ_p)-DP by adding a random noise from a Gaussian distribution $\mathcal{N}(0, \sigma^2 I_d)$ to the output of algorithm (function) $\mathcal{A}$, where σ is proportional to the ℓ_2-sensitivity of $\mathcal{A}$.

Lemma 1.1 ([11]). *Given an algorithm (function) $\mathcal{A} : \mathcal{Z}^n \mapsto \mathbb{R}^d$ with the ℓ_2-sensitivity Δ and a dataset $\mathcal{S} \subset \mathcal{Z}^n$, and assume that $\sigma \geq \frac{\sqrt{2\log(1.25/\delta_p)}\Delta}{\epsilon}$. Then the following Gaussian mechanism yields (ϵ, δ_p)-DP*

$$\mathcal{G}(\mathcal{S}, \sigma) := \mathcal{A}(\mathcal{S}) + b, \quad b \sim \mathcal{N}(0, \sigma^2 I_d), \tag{8}$$

where I_d is the identity matrix in $\mathbb{R}^{d\times d}$.

The goal of this paper is to prove algorithm 1 satisfies the (ϵ, δ_p)-differential privacy and establish fast excess generalization error bounds under some mild conditions. The rest of the

paper is organized as follows. The main results are given in Section 2. To prove our main results, we first introduce our error decomposition and obtain some technical estimates in Section 3. Then we provide proofs of convergence rates and privacy guarantee in Section 4.

§2 Main results

Throughout the paper, we assume that the output y is uniformly bounded, i.e., for some constant $c > 0$, $|y| \leq c$ almost surely.

By the definition (4) of random feature map ϕ_M, we have

$$\langle \phi_M(x), \phi_M(x') \rangle = \frac{1}{M} \sum_{j=1}^{M} \psi(x, \nu_j)\psi(x', \nu_j), \quad \forall x, x' \in \mathcal{X} \tag{9}$$

which can be shown to converge to the following symmetric and positive semi-definite kernel $K(x, x')$

$$K(x, x') = \int \psi(x, \nu)\psi(x', \nu) d\pi(\nu), \quad \forall x, x' \in \mathcal{X}.$$

as the number of random features M tends to infinity [5, 21, 22]. And the positive semi-definite kernel K can be expressed as $K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{F}}$ with feature map $\phi : \mathcal{X} \mapsto \mathcal{F}$, and unlike the random feature map $\phi_M : \mathcal{X} \mapsto \mathbb{R}^M$, here feature space $\mathcal{F}$ can be infinite dimensional. Then we have

$$K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{F}} \approx \langle \phi_M(x), \phi_M(x') \rangle = K_M(x, x') \tag{10}$$

The Gaussian kernel provides a basic example [21], more examples we refer the interested readers to [21, 22] and reference therein.

Example 2.1 (Random Fourier features [21]). *If we write the Gaussian kernel as $K(x, x') = G(x - x')$, with $G(z) = e^{-\frac{\|z\|^2}{2\sigma^2}}$ for $\sigma > 0$, then since the inverse Fourier transform of G is a Gaussian and using a basic symmetry argument, it is easy to show that*

$$G(x - x') = \frac{1}{2\pi Q} \int \int_0^{2\pi} \sqrt{2} cos(w^\top x + b)\sqrt{2}(w^\top x' + b) e^{-\frac{\sigma^2}{2}\|w\|^2} dw db \tag{11}$$

where Q is a normalizing factor. Then, the Gaussian kernel has an approximation of the form with $\phi_M(x) = \frac{1}{\sqrt{M}}(\sqrt{2}cos(w_1^\top x + b_1), \cdots, \sqrt{2}cos(w_M^\top x + b_M))^\top$ and $w_1, w_2, \cdots, w_M$ and $b_1, b_2, \cdots, b_M$ sampled independently from $\frac{1}{Q} e^{-\frac{\sigma^2\|w\|^2}{2}}$ and uniformly in $[0, 2\pi]$ respectively.

The reproducing kernel Hilbert space $\mathcal{H}_K$ associated with the kernel K is the completion of the span $\{K_x = K(\cdot, x) : x \in \mathcal{X}\}$ with respect to the inner product $\langle \cdot, \cdot \rangle_K$ given by $\langle K_x, K_{x'} \rangle_K = K(x, x')$. And the reproducing property shows that

$$f(x) = \langle f, K_x \rangle_K, \quad , \forall x \in \mathcal{X}, f \in \mathcal{H}_K. \tag{12}$$

The integral operator $L : L^2_{\rho_\mathcal{X}} \mapsto L^2_{\rho_\mathcal{X}}$ associated with the kernel K and the marginal distribution $\rho_\mathcal{X}$ of ρ is defined by

$$Lf(\cdot) = \int_\mathcal{X} K(\cdot, x) f(x) d\rho_\mathcal{X}(x), \quad \forall f \in L^2_{\rho_\mathcal{X}} \tag{13}$$

here $L^2_{\rho_\mathcal{X}}$ denotes the square-integrable Hilbert space.

Our main results are stated based on two key assumptions. The first assumption is about the regularity condition of the regression function. We assume the target function f_ρ satisfies some smoothness condition(regularity condition) which is standard in learning theory [8, 25].

Assumption 2.1. *There exist $r \geq 1/2$ and $g_\rho \in L^2_{\rho_\mathcal{X}}$ such that*

$$f_\rho = L^r g_\rho, \quad \|g_\rho\|_\rho \leq R. \tag{14}$$

Here the $\|\cdot\|_\rho$ denotes the norm in $L^2_{\rho_\mathcal{X}}$ induced by the inner product $\langle h, g\rangle_\rho = \int_\mathcal{X} f(x)g(x)\, d\rho_\mathcal{X}(x)$ for $h, g \in L^2_{\rho_\mathcal{X}}$. This assumption implies that f_ρ belongs to the range space of L^r. Moreover, since $\rho_\mathcal{X}$ is non-degenerate, from Theorem 4.12 in [8], $L^{1/2}$ is an isomorphism from $\overline{\mathcal{H}_K}$, the closure of $\mathcal{H}_K$ in $L^2_{\rho_\mathcal{X}}$, to $\mathcal{H}_K$, i.e., for each $f \in \overline{\mathcal{H}_K}$, $L^{1/2} f \in \mathcal{H}_K$ and $\|f\|_\rho = \|L^{1/2} f\|_K$, Therefore, $L^{1/2}(L^2_{\rho_\mathcal{X}}) = \mathcal{H}_K$, and when $r \geq \frac{1}{2}$, condition (14) implies $f_\rho \in \mathcal{H}_K$.

We next introduce the second key assumption on the capacity of the RKHS $\mathcal{H}_K$. In this paper, we measure the capacity of the hypothesis space $\mathcal{H}_K$ by the effective dimension $\mathcal{N}(\lambda) = \mathrm{Tr}(L(L+\lambda I))$, here $\mathrm{Tr}(A)$ denotes the trace of an operator A of trace class.

Assumption 2.2. *We assume that*

$$\mathcal{N}(\lambda) \leq C_0 \lambda^{-\alpha} \text{ for some } C_0 > 0 \text{ and } 0 < \alpha < 1. \tag{15}$$

Since L is a trace class operator satisfying $\mathrm{Tr}(L) = \sum_{k\geq 1} \sigma_k = \int_\mathcal{X} K(x,x) d\rho_\mathcal{X} \leq \kappa^2$, Assumption 2.2 holds trivially with $\alpha = 1$.

Theorem 1. *Let $\hat{f}_{priv}$ be defined by algorithm 1 and $0 < \delta < 1$. Under Assumption 2.1 with $\frac{1}{2} < r \leq 1$ and Assumption 2.2 with $\frac{4-4r}{3-2r} < \alpha < 1$, let step size $0 < \eta < \frac{1}{\kappa^2(\log T+1)}$, and the number of random features $M \simeq n^{\frac{1+\alpha(2r-1)}{2r+\alpha}} \log \frac{n}{\delta}$. If $\eta \simeq n^{-1}, T = n^{\frac{2r+\alpha+1}{2r+\alpha}}$; or $\eta \simeq n^{-\frac{2r}{2r+\alpha}}, T = n^{\frac{2r+1}{2r+\alpha}}$, then with probability at least $1-\delta$, there holds*

$$\mathbb{E}_{\mathbf{I}_T}\left(\mathcal{E}(\hat{f}_{priv}) - \mathcal{E}(f_\rho)\right) \lesssim n^{-\frac{4r+3\alpha-4-2r\alpha}{2r+\alpha}} \frac{1}{\epsilon^2} \log\frac{n}{\delta} \log\frac{2n}{\delta_p} \log\frac{2.5}{\delta_p} + n^{-\frac{2r}{2r+\alpha}} \log^2 n \log^2 \frac{18}{\delta}. \tag{16}$$

Here we denote by $\mathbb{E}_{\mathbf{I}_T}$ the expectation with expectation to the set $\{i_1, i_2, \cdots, i_T\}$. The symbols $\simeq$ and $\lesssim$ mean that the inequality holds up to a multiplicative constant that depends on various parameters appearing in the assumptions, but not on the sample size n or the number of random features M. The proof of Theorem 1 will be given in subsection 4.1.

Theorem 2. *Let $\{\hat{w}_t\}$ be defined by (5), Let $c_{\gamma,T} = \max\left\{\sqrt{\frac{3\log(n/\gamma)}{T/n}}, \frac{3\log(n/\gamma)}{T/n}\right\}$, then for any $0 < \gamma < 1$, there holds*

$$\mathbb{P}\left(\sup_{\mathcal{S}\simeq\mathcal{S}} \delta_\mathcal{A}(\mathcal{S},\mathcal{S}') = \|\mathcal{A}(\mathcal{S}) - \mathcal{A}(\mathcal{S}')\|_2 \geq \Delta_{SGD}(\gamma)\right) \leq \gamma, \tag{17}$$

where $\Delta_{SGD}(\gamma) = 4e\eta^2\left(c\kappa + \kappa^2\sqrt{\eta T}\right)^2 \frac{T}{n}(1+c_{\gamma,T})\left(1+\frac{T}{n}(1+c_{\gamma,T})\right)$.

Theorem 3 (Privacy guarantee)**.** *The algorithm 1 satisfies (ϵ, δ_p)-DP.*

The proof of Theorem 2 and Theorem 3 will be given in subsection 4.2.

Our convergence rate (16) consists of two terms, the first term $n^{-\frac{4r+3\alpha-4-2r\alpha}{2r+\alpha}} \frac{1}{\epsilon^2} \log\frac{n}{\delta} \log\frac{n}{\delta_p} \log\frac{1}{\delta_p}$ depends on regularity parameter r, capacity parameter α, the sample size n and the parameters (ϵ, δ_p) of differential privacy, while the second term $n^{-\frac{2r}{2r+\alpha}} \log^2 n$ is optimal in the mini-max sense(up to a logarithmic term $\log n$) which matches the results of SGD with random features and mini-batches [5] and ridge regression with random features [22] under the same assumptions. There is a large literature on the analysis on privacy preserving machine learning algorithms [2,6,12,29,30], random features [5,16,21,22,22,26] and SGD in the setting of stochastic convex optimization [4, 18, 24] or in reproducing kernel Hilbert space [7, 9, 13–15, 17, 31, 32].

To the best of our knowledge, this is the first result that combines the advantages of SGD, random features and differential privacy. Here we only review some results on differentially private SGD with output perturbation for the sake of scale of the paper. For more results, we refer the readers to the work mentioned in this paper and the reference therein. Most of the differentially private SGD algorithms are considered in the setting of stochastic convex optimization, where the parameter domain $\mathcal{W} \subset \mathbb{R}^d$ is convex. Differentially private SGD with output perturbation is studied in [30], where the loss is assumed to be Lipschitz continuous and strongly smooth, and $\mathcal{W}$ is assumed to be uniformly bounded, then it showed an excess risk rate $\mathcal{O}\Big(\frac{(d\log(1/\delta_p))^{1/4}}{\sqrt{n\epsilon}}\Big)$ with linear gradient complexity. Recently, private SGD with more general α-Hölder smooth ($\alpha \in [0,1]$) loss function ($\alpha = 0$ corresponds to Lipschitz continuous and $\alpha = 1$ means strong smooth) are considered in [29], where privacy guarantee and generalization bounds are established for both output and gradient perturbations. For the unbounded domain $\mathcal{W}$, it shows in [29] that the private SGD with output perturbation attains the excess generalization error $\mathcal{O}\Big(\frac{\sqrt{d\log(1/\delta_p)}\log(n/\delta_p)}{n^{2/(3+\alpha)}\epsilon} + \frac{\log(n/\delta_p)}{n^{1/(3+\alpha)}}\Big)$. In addition to the output perturbation mechanism, another popular mechanism to achieve differential privacy is called gradient perturbation, which adds random noise to the stochastic gradient at each step. Under the same assumption on the loss functions as [30], it shows in [3] that the differentially private SGD with gradient perturbation achieves an optimal excess error rate $\mathcal{O}\Big(\frac{\sqrt{d\log(1/\delta_p)}}{n\epsilon} + \frac{1}{\sqrt{n}}\Big)$ but with computationally inefficient gradient complexity, then the gradient complexity of the algorithm have been improved by [12] and [2]. And optimal excess error rate (up to some logarithmic terms) is also established in [29] for private SGD with α-Hölder smooth function using gradient perturbation.

§3 Preliminaries

Before proving the main results, we give some notations [5] and useful lemmas in this section. Let $\mathcal{F}$ be the feature space corresponding to the kernel K. Given the feature map $\phi : \mathcal{X} \to \mathcal{F}$, we can define the operator $S : \mathcal{F} \to L^2_{\rho_\mathcal{X}}$ as

$$(Sw)(\cdot) = \langle w, \phi(\cdot)\rangle_\mathcal{F}, \quad \forall w \in \mathcal{F}, \tag{18}$$

and define $C : \mathcal{F} \to \mathcal{F}$ as $C = S^*S$, where $S^* : L^2_{\rho_\mathcal{X}} \to \mathcal{F}$ is the adjoint operator of S, and C is given by

$$C = \int_\mathcal{X} \phi(x) \otimes \phi(x) d\rho_X(x) = \int_\mathcal{X} \langle \phi(x), \cdot\rangle_\mathcal{F} \phi(x) d\rho_X(x). \tag{19}$$

And the integral operator L defined by (13) can be represented as $L = SS^*$. Now we define similar operators where we use the random feature map ϕ_M instead of feature map ϕ. We define $S_M : \mathbb{R}^M \to L^2_{\rho_\mathcal{X}}$ as

$$(S_M v)(\cdot) = \langle v, \phi_M(\cdot)\rangle, \quad \forall v \in \mathbb{R}^M, \tag{20}$$

And we define $C_M : \mathbb{R}^M \to \mathbb{R}^M$ and $L_M : L^2_{\rho_\mathcal{X}} \to L^2_{\rho_\mathcal{X}}$ as $C_M = S_M^* S_M$ and $L_M = S_M S_M^*$ respectively, where S_M^* is the adjoint operator of S_M.

We also define some operators with respect to training samples. The operator $\hat{S}_M : \mathbb{R}^M \to \mathbb{R}^n$ is defined as

$$\hat{S}_M^\top = \frac{1}{\sqrt{n}}(\phi_M(x_1), \cdots, \phi_M(x_n)). \tag{21}$$

And $\hat{C}_M : \mathbb{R}^M \to \mathbb{R}^M$ and $\hat{L}_M : \mathbb{R}^n \to \mathbb{R}^n$ are defined as $\hat{C}_M = \hat{S}_M^\top \hat{S}_M$ and $\hat{L}_M = \hat{S}_M \hat{S}_M^\top$ respectively.

To sum up, the operators associated with the random feature map ϕ_M defined above are as follows:

$$\begin{aligned}
S_M &: \mathbb{R}^M \to L^2_{\rho\mathcal{X}}. \quad (S_M v)(\cdot) = \langle v, \phi_M(\cdot)\rangle, \forall v \in \mathbb{R}^M \\
C_M &: \mathbb{R}^M \to \mathbb{R}^M. \quad C_M = S_M^* S_M. \quad C_M = \int_{\mathcal{X}} \phi_M(x)\phi_M(x)^\top d\rho_{\mathcal{X}}(x) \\
L_M &: L^2_{\rho\mathcal{X}} \to L^2_{\rho\mathcal{X}}. \quad L_M = S_M S_M^*. \\
&\quad (L_M g)(\cdot) = \int_{\mathcal{X}} K_M(\cdot, z) g(z) d\rho_{\mathcal{X}}(z) \quad (K_M(x,y) = \langle \phi_M(x), \phi_M(y)\rangle) \\
\hat{S}_M &: \mathbb{R}^M \to \mathbb{R}^n. \quad \hat{S}_M^\top = \frac{1}{\sqrt{n}}(\phi_M(x_1), \cdots, \phi_M(x_n)) \\
\hat{C}_M &: \mathbb{R}^M \to \mathbb{R}^M. \quad \hat{C}_M = \hat{S}_M^\top \hat{S}_M. \quad \hat{C}_M = \frac{1}{n}\sum_{i=1}^n \phi_M(x_i)\phi_M(x_i)^\top \\
\hat{L}_M &: \mathbb{R}^n \to \mathbb{R}^n. \quad \hat{L}_M = \hat{S}_M \hat{S}_M^\top \\
\hat{y} &: \hat{y} = \frac{1}{\sqrt{n}}(y_1, y_2, \cdots, y_n)^\top
\end{aligned}$$

3.1 Error decomposition

It is well known that the excess generalization error $\mathcal{E}(\hat{f}_{priv}) - \mathcal{E}(f_\rho)$ can be expressed as

$$\mathcal{E}(\hat{f}_{priv}) - \mathcal{E}(f_\rho) = \int_{\mathcal{X}\times\mathcal{Y}} (\hat{f}_{priv}(x) - y)^2 d\rho - \int_{\mathcal{X}\times\mathcal{Y}} (f_\rho(x) - y)^2 d\rho = \left\| \hat{f}_{priv} - f_\rho \right\|_\rho^2 \tag{22}$$

and $\hat{f}_{priv} - f_\rho$ can be decomposed into the following six terms [5]

$$\hat{f}_{priv} - f_\rho =$$
$$[f_{priv} - \hat{f}_{T+1}] + [\hat{f}_{T+1} - \hat{g}_{T+1}] + [\hat{g}_{T+1} - \tilde{g}_{T+1}] + [\tilde{g}_{T+1} - \tilde{g}_\lambda] + [\tilde{g}_\lambda - g_\lambda] + [g_\lambda - f_\rho]$$

where $\hat{f}_t = \langle \hat{w}_t, \phi_M(\cdot)\rangle,\ \hat{g}_t = \langle \hat{v}_t, \phi_M(\cdot)\rangle,\ \tilde{g}_t = \langle \tilde{v}_t, \phi_M(\cdot)\rangle,\ \tilde{g}_\lambda = \langle \tilde{u}_\lambda, \phi_M(\cdot)\rangle,\ g_\lambda = \langle u_\lambda, \phi(\cdot)\rangle_{\mathcal{F}}$, $\forall 1 \le t \le T$ and $\hat{w}_t, \hat{v}_t, \tilde{v}_t, \tilde{u}_\lambda, u_\lambda$ are defined as follows [5]

$$\hat{w}_1 = 0; \quad \hat{w}_{t+1} = \hat{w}_t - \eta(\langle \hat{w}_t, \phi_M(x_{i_t})\rangle - y_{i_t})\phi_M(x_{i_t}),\ \forall 1 \le t \le T \tag{23}$$

$$\hat{v}_1 = 0; \quad \hat{v}_{t+1} = \hat{v}_t - \eta\frac{1}{n}\sum_{i=1}^n (\langle \hat{v}_t, \phi_M(x_i)\rangle - y_i)\phi_M(x_i),\ \forall 1 \le t \le T \tag{24}$$

$$\tilde{v}_1 = 0; \quad \tilde{v}_{t+1} = \tilde{v}_t - \eta\int_{\mathcal{X}} (\langle \tilde{v}_t, \phi_M(x)\rangle - y)\phi_M(x) d\rho(x,y),\ \forall 1 \le t \le T \tag{25}$$

$$\tilde{u}_\lambda = \arg\min_{u\in\mathbb{R}^M} \int_{\mathcal{X}} (\langle u, \phi_M(x)\rangle - y)^2 d\rho_{\mathcal{X}}(x) + \lambda\|u\|^2,\ \lambda > 0 \tag{26}$$

$$u_\lambda = \arg\min_{u\in\mathcal{F}} \int_{\mathcal{X}} (\langle u, \phi(x)\rangle_{\mathcal{F}} - y)^2 d\rho(x,y) + \lambda\|u\|_{\mathcal{F}}^2,\ \lambda > 0 \tag{27}$$

With the operators introduced in the beginning of Section 3, we can rewrite $\hat{v}_{T+1}, \tilde{v}_{T+1}, \tilde{u}_\lambda$ and u_λ respectively as

$$\hat{v}_{T+1} = \sum_{t=1}^T \eta \prod_{k=t+1}^T (I - \eta\hat{C}_M)\hat{S}_M^*\hat{y} = \sum_{t=1}^T \eta(I - \eta\hat{C}_M)^{T-t}\hat{S}_M^*\hat{y}, \tag{28}$$

$$\tilde{v}_{T+1} = \sum_{t=1}^T \eta \prod_{k=t+1}^T (I - \eta C_M)S_M^* f_\rho = \sum_{t=1}^T \eta(I - \eta C_M)^{T-t}S_M^* f_\rho, \tag{29}$$

$$\tilde{u}_\lambda = S_M^*(L_M + \lambda I)^{-1} f_\rho, \tag{30}$$
$$u_\lambda = S^*(L + \lambda I)^{-1} f_\rho. \tag{31}$$

It follows that

$f_{priv} = S_M \hat{w}_{priv}, \quad \hat{f}_{T+1} = S_M \hat{w}_{T+1}, \quad \hat{g}_{T+1} = S_M \hat{v}_{T+1}, \quad \tilde{g}_{T+1} = S_M \tilde{v}_{T+1},$
$\tilde{g}_\lambda = S_M S_M^*(L_M + \lambda I)^{-1} f_\rho = L_M(L_M + \lambda I)^{-1} f_\rho, \quad g_\lambda = SS^*(L + \lambda I)^{-1} f_\rho = L(L + \lambda I)^{-1} f_\rho$

Let

$$\begin{aligned}
G_1 &:= \mathbb{E}_{\mathbf{I}_T}\left[\|S_M \hat{w}_{priv} - S_M \hat{w}_{T+1}\|_\rho^2\right], \\
G_2 &:= \mathbb{E}_{\mathbf{I}_T}\left[\|S_M \hat{w}_{T+1} - S_M \hat{v}_{T+1}\|_\rho^2\right], \\
G_3 &:= \|S_M \hat{v}_{T+1} - S_M \tilde{v}_{T+1}\|_\rho^2, \\
G_4 &:= \left\|S_M \tilde{v}_{T+1} - L_M(L_M + \lambda I)^{-1} f_\rho\right\|_\rho^2, \\
G_5 &:= \left\|L_M(L_M + \lambda I)^{-1} f_\rho - L(L + \lambda I)^{-1} f_\rho\right\|_\rho^2, \\
G_6 &:= \left\|L(L + \lambda I)^{-1} f_\rho - f_\rho\right\|_\rho^2,
\end{aligned} \tag{32}$$

then the expected excess generalization error can be divided into the following six terms

$$\mathbb{E}_{\mathbf{I}_T}\left(\mathcal{E}(\hat{f}_{priv}) - \mathcal{E}(f_\rho)\right) = \mathbb{E}_{\mathbf{I}_T}\left[\left\|\hat{f}_{priv} - f_\rho\right\|_\rho^2\right] \le 6(G_1 + G_2 + G_3 + G_4 + G_5 + G_6). \tag{33}$$

We will bound the six terms of the right hand side of (33) respectively in Section 4.

3.2 Technical estimates

In order to prove our main results in Section 4, we need some preliminary lemmas. For any $w \in \mathbb{R}^M$, we define empirical risk as

$$\mathcal{E}_Z(w) = \frac{1}{n}\sum_{i=1}^n \left(\langle w, \phi_M(x_i)\rangle - y_i\right)^2. \tag{34}$$

Lemma 3.1. *Let $T \ge 3$, the step size η satisfies*

$$0 < \eta < \frac{1}{\kappa^2(\log T + 1)}, \tag{35}$$

then for $2 \le t \le T$, we have

$$\mathbb{E}_{\mathbf{I}_{t-1}}[\mathcal{E}_Z(\hat{w}_t)] \le \frac{c^2}{(1 - \eta\kappa^2)(1 - \eta\kappa^2 - \eta\kappa^2 \log T)}. \tag{36}$$

Proof. We borrow some ideas from [19, 24] to prove this lemma. Let $\{\hat{w}_t\}_t$ be defined by (5), for $k = 1, \cdots, t-1$, we have

$$\begin{aligned}
&\frac{1}{k}\sum_{l=t-k+1}^{t} \mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)] - \frac{1}{k+1}\sum_{l=t-k}^{t} \mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)] \\
&= \frac{1}{k(k+1)}\left((k+1)\sum_{l=t-k+1}^{t} \mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)] - k\sum_{l=t-k}^{t} \mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)]\right) \\
&= \frac{1}{k(k+1)}\sum_{l=t-k+1}^{t} \mathbb{E}_{\mathbf{I}_{l-1}}\left[\mathcal{E}_Z(\hat{w}_l) - \mathcal{E}_Z(\hat{w}_{t-k})\right].
\end{aligned}$$

Using this inequality repeatedly and by summing over $k = 1, 2, \cdots, t-1$, we have

$$\mathbb{E}_{\mathbf{I}_{t-1}}[\mathcal{E}_Z(\hat{w}_t)] = \frac{1}{t}\sum_{l=1}^{t}\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)] + \sum_{k=1}^{t-1}\frac{1}{k(k+1)}\sum_{l=t-k+1}^{t}\mathbb{E}_{\mathbf{I}_{l-1}}\left[\mathcal{E}_Z(\hat{w}_l) - \mathcal{E}_Z(\hat{w}_{t-k})\right]. \quad (37)$$

Next we estimate the two terms of the right hand side of (37) respectively. For the first term $\frac{1}{t}\sum_{l=1}^{t}\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)]$. By the definition of $\{\hat{w}_l\}$, we have

$$\begin{aligned}\|\hat{w}_{l+1}\|_2^2 &= \|\hat{w}_l\|_2^2 - 2\eta(\langle\hat{w}_l, \phi_M(x_{i_l})\rangle - y_{i_l})\langle\hat{w}_l, \phi_M(x_{i_l})\rangle + \eta^2(\langle\hat{w}_l, \phi_M(x_{i_l})\rangle - y_{i_l})^2\|\phi_M(x_{i_l})\|_2^2\\ &\le \|\hat{w}_l\|_2^2 + \eta\left((y_{i_l})^2 - (\langle\hat{w}_l, \phi_M(x_{i_l})\rangle - y_{i_l})^2\right) + \eta^2\kappa^2\left(\langle\hat{w}_l, \phi_M(x_{i_l})\rangle - y_{i_l}\right)^2.\end{aligned}$$

Then rearrange terms and by the definition (34) of $\mathcal{E}_Z(w)$, we have

$$\eta\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_l(\hat{w}_l) - \mathcal{E}_Z(0)] \le \mathbb{E}_{\mathbf{I}_l}[\|\hat{w}_l\|_2^2 - \|\hat{w}_{l+1}\|_2^2] + \eta^2\kappa^2\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)],$$

and

$$\eta(1-\eta\kappa^2)\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)] \le \mathbb{E}_{\mathbf{I}_l}[\|\hat{w}_l\|_2^2 - \|\hat{w}_{l+1}\|_2^2] + \eta\mathcal{E}_Z(0).$$

Since $\eta\kappa^2 < 1$, it follows that

$$\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)] \le \frac{1}{\eta(1-\eta\kappa^2)}\mathbb{E}_{\mathbf{I}_l}[\|\hat{w}_l\|_2^2 - \|\hat{w}_{l+1}\|_2^2] + \frac{1}{1-\eta\kappa^2}\mathcal{E}_Z(0)$$

and

$$\frac{1}{t}\sum_{l=1}^{t}\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)] \le \frac{1}{t\eta(1-\eta\kappa^2)}\mathbb{E}_{\mathbf{I}_t}[\|\hat{w}_1\|_2^2 - \|\hat{w}_{t+1}\|_2^2] + \frac{1}{1-\eta\kappa^2}\mathcal{E}_Z(0) \le \frac{1}{1-\eta\kappa^2}\mathcal{E}_Z(0).$$

Now we turn to estimate the second term $\mathbb{E}_{\mathbf{I}_{l-1}}\left[\mathcal{E}_Z(\hat{w}_l) - \mathcal{E}_Z(\hat{w}_{t-k})\right]$ for $t-k+1 \le l \le T$. By the definition of $\hat{w}_{l+1}$, we have

$$\hat{w}_{l+1} - \hat{w}_{t-k} = \hat{w}_l - \hat{w}_{t-k} - \eta(\langle\hat{w}_l, \phi_M(x_{i_l})\rangle - y_{i_l})\phi_M(x_{i_l}), \quad (38)$$

then

$$\begin{aligned}\|\hat{w}_{l+1} - \hat{w}_{t-k}\|_2^2 &= \|\hat{w}_l - \hat{w}_{t-k}\|_2^2 - 2\eta(\langle\hat{w}_l, \phi_M(x_{i_l})\rangle - y_{i_l})\langle\hat{w}_l - \hat{w}_{t-k}, \phi_M(x_{i_l})\rangle\\ &\quad + \eta^2(\langle\hat{w}_l, \phi_M(x_{i_l})\rangle - y_{i_l})^2\|\phi_M(x_{i_l})\|_2^2\\ &\le \|\hat{w}_l - \hat{w}_{t-k}\|_2^2 + \eta\left((\langle\hat{w}_{t-k}, \phi_M(x_{i_l})\rangle - y_{i_l})^2 - (\langle\hat{w}_l, \phi_M(x_{i_l})\rangle - y_{i_l})^2\right)\\ &\quad + \eta^2\kappa^2\left(\langle\hat{w}_l, \phi_M(x_{i_l})\rangle - y_{i_l}\right)^2.\end{aligned}$$

It follows that

$$\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l) - \mathcal{E}_Z(\hat{w}_{t-k})] \le \frac{1}{\eta}\mathbb{E}_{\mathbf{I}_{l-1}}[\|\hat{w}_l - \hat{w}_{t-k}\|_2^2 - \|\hat{w}_{l+1} - \hat{w}_{t-k}\|_2^2] + \eta\kappa^2\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)].$$

Applying this relationship iteratively from $l = t-k+1$ to $l = t$,

$$\begin{aligned}&\sum_{l=t-k+1}^{t}\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l) - \mathcal{E}_Z(\hat{w}_{t-k})] = \sum_{l=t-k}^{t}\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l) - \mathcal{E}_Z(\hat{w}_{t-k})]\\ &\le \sum_{l=t-k}^{t}\frac{1}{\eta}\mathbb{E}_{\mathbf{I}_{l-1}}[\|\hat{w}_l - \hat{w}_{t-k}\|_2^2 - \|\hat{w}_{l+1} - \hat{w}_{t-k}\|_2^2] + \eta\kappa^2\sum_{l=t-k}^{t}\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)]\\ &\le \frac{1}{\eta}\mathbb{E}_{\mathbf{I}_{t-k-1}}[\|\hat{w}_{t-k} - \hat{w}_{t-k}\|_2^2 - \|\hat{w}_{t+1} - \hat{w}_{t-k}\|_2^2] + \eta\kappa^2\sum_{l=t-k}^{t}\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)]\\ &\le \eta\kappa^2\sum_{l=t-k}^{t}\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)].\end{aligned}$$

Then the second term of the right hand side of (37) can be estimated as

$$\begin{aligned}
&\sum_{k=1}^{t-1}\frac{1}{k(k+1)}\sum_{l=t-k+1}^{t}\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)-\mathcal{E}_Z(\hat{w}_{t-k})]\\
&\le\sum_{k=1}^{t-1}\frac{\eta\kappa^2}{k(k+1)}\sum_{l=t-k}^{t}\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)]\\
&=\sum_{k=1}^{t-1}\frac{\eta\kappa^2}{k(k+1)}\sum_{l=t-k}^{t-1}\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)]+\sum_{k=1}^{t-1}\frac{\eta\kappa^2}{k(k+1)}\mathbb{E}_{\mathbf{I}_{t-1}}[\mathcal{E}_Z(\hat{w}_t)]\\
&\le\sum_{k=1}^{t-1}\frac{\eta\kappa^2}{k(k+1)}\sum_{l=t-k}^{t-1}\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)]+\eta\kappa^2\mathbb{E}_{\mathbf{I}_{t-1}}[\mathcal{E}_Z(\hat{w}_t)].
\end{aligned}\tag{39}$$

Then putting (38) and (39) back into (37), we get

$$\begin{aligned}
\mathbb{E}_{\mathbf{I}_{t-1}}[\mathcal{E}_Z(\hat{w}_t)]&\le\frac{1}{(1-\eta\kappa^2)^2}\mathcal{E}_Z(0)+\frac{\eta\kappa^2}{1-\eta\kappa^2}\sum_{k=1}^{t-1}\frac{1}{k(k+1)}\sum_{l=t-k}^{t-1}\sup_{1\le l\le t-1}\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)]\\
&\le\frac{1}{(1-\eta\kappa^2)^2}\mathcal{E}_Z(0)+\frac{\eta\kappa^2\log t}{1-\eta\kappa^2}\sup_{1\le l\le t-1}\mathbb{E}_{\mathbf{I}_{l-1}}[\mathcal{E}_Z(\hat{w}_l)].
\end{aligned}\tag{40}$$

Let $A=\frac{1}{(1-\eta\kappa^2)^2}\mathcal{E}_Z(0)$, $B_t=\frac{\eta\kappa^2\log t}{1-\eta\kappa^2}$, $C_t=\mathbb{E}_{\mathbf{I}_{t-1}}[\mathcal{E}_Z(\hat{w}_t)]$. Since $\eta\le\frac{1}{\kappa^2(\log T+1)}$, there holds

$$B_t\le\frac{\eta\kappa^2\log T}{1-\eta\kappa^2}:=B\le 1.$$

Then for $t\ge 2$, we can get from (40) that

$$C_t\le A+B_t\sup_{1\le l\le t-1}C_l\le A+B\sup_{1\le l\le t-1}C_l,$$

then it follows that

$$\sup_{2\le l\le t-1}C_l\le A+B\sup_{1\le l\le t-1}C_l.$$

And for C_1, by $\eta\le\frac{1}{\kappa^2(\log T+1)}$, it is obvious that $C_1\le A+BC_1$. Then we have

$$\begin{aligned}
\sup_{1\le l\le t-1}C_l&\le A+B\sup_{1\le l\le t-1}C_l,\\
\sup_{1\le l\le t-1}C_l&\le\frac{A}{1-B}.
\end{aligned}$$

Therefore, for $t\ge 2$, we have

$$C_t\le A+B_t\sup_{1\le l\le t-1}C_l\le A+B_t\frac{A}{1-B}\le\frac{A}{1-B}.$$

That is, for $2\le t\le T$, there holds

$$\mathbb{E}_{\mathbf{I}_{t-1}}[\mathcal{E}_Z(\hat{w}_t)]\le\frac{1}{(1-\frac{\eta\kappa^2\log T}{1-\eta\kappa^2})(1-\eta\kappa^2)^2}\mathcal{E}_Z(0)\le\frac{c^2}{(1-\eta\kappa^2)(1-\eta\kappa^2-\eta\kappa^2\log T)},\tag{41}$$

where the last inequality holds due to $|y|\le c$ almost surely. □

We also need the following bound for the sequence $\{\tilde{v}_t\}_{t\ge1}$.

Lemma 3.2. *Let $\{\tilde{v}_t\}$ be defined by (25), $\delta\in(0,1)$, $\hat{\lambda}\ge\frac{9\kappa^2}{M}\log\frac{M}{\delta}$, and the target function f_ρ satisfies Assumption 2.1 with $1/2<r\le 1$, then*

$$\|\tilde{v}_{t+1}\|_2\le 2R\kappa^{2r-1}\left(1+\eta\hat{\lambda}t\right),\quad\forall 1\le t\le T,\tag{42}$$

holds with probability at least $1-\delta$.

Proof. Let $\{\tilde{v}_t\}$ be defined by (25), and $f_\rho = L^r g_\rho$ with $\frac{1}{2} < r \leq 1$ and $\|g_\rho\|_\rho \leq R$, then for $\hat{\lambda} > 0$, we have

$$\begin{aligned}
\|\tilde{v}_{t+1}\|_2 &= \left\|\sum_{k=1}^{t} \eta(I-\eta C_M)^{t-k} S_M^* f_\rho\right\|_2 = \left\|\sum_{k=1}^{t} \eta S_M^*(I-\eta L_M)^{t-k} f_\rho\right\|_2 \\
&= \left\|\sum_{k=1}^{t} \eta L_M^{\frac{1}{2}}(I-\eta L_M)^{t-k} f_\rho\right\|_\rho \\
&= \left\|\sum_{k=1}^{t} \eta L_M^{\frac{1}{2}}(I-\eta L_M)^{t-k}(L_M+\hat{\lambda}I)^{\frac{1}{2}}(L_M+\hat{\lambda}I)^{-\frac{1}{2}} L^{\frac{1}{2}} L^{r-\frac{1}{2}} g_\rho\right\|_\rho \\
&\leq \left\|\sum_{k=1}^{t} \eta L_M^{\frac{1}{2}}(I-\eta L_M)^{t-k}(L_M+\hat{\lambda}I)^{\frac{1}{2}}\right\| \cdot \left\|(L_M+\hat{\lambda}I)^{-\frac{1}{2}} L^{\frac{1}{2}}\right\| \cdot \left\|L^{r-\frac{1}{2}}\right\| \|g_\rho\|_\rho \\
&\leq 2R\kappa^{2r-1} \left\|\sum_{k=1}^{t} \eta(I-\eta L_M)^{t-k}(L_M+\hat{\lambda}I)\right\|,
\end{aligned}$$

the last inequality holds since $\left\|(L_M+\hat{\lambda}I)^{-\frac{1}{2}} L^{\frac{1}{2}}\right\| \leq 2$ with probability at least $1-\delta$ for $\hat{\lambda} \geq \frac{9\kappa^2}{M}\log\frac{M}{\delta}$ [5], $\|g_\rho\|_\rho \leq R$ and $\|L\| \leq \kappa^2$. Now we consider the term $\left\|\sum_{k=1}^{t} \eta(I-\eta L_M)^{t-k}(L_M+\hat{\lambda}I)\right\|$,

$$\begin{aligned}
\left\|\sum_{k=1}^{t} \eta(I-\eta L_M)^{t-k}(L_M+\hat{\lambda}I)\right\| &\leq \left\|\sum_{k=1}^{t} \eta(I-\eta L_M)^{t-k} L_M\right\| + \hat{\lambda}\left\|\sum_{k=1}^{t} \eta(I-\eta L_M)^{t-k}\right\| \\
&= \left\|-\sum_{k=1}^{t}(I-\eta L_M)^{t-k}(I-\eta L_M) + \sum_{k=1}^{t}(I-\eta L_M)^{t-k}\right\| + \hat{\lambda}\eta\left\|\sum_{k=1}^{t}(I-\eta L_M)^{t-k}\right\| \\
&\leq \|I-(I-\eta L_M)^t\| + \eta\hat{\lambda}t \\
&\leq 1+\eta\hat{\lambda}t.
\end{aligned}$$

Then the desired result holds. □

The following Bernstein inequality also plays a crucial role in the proof of the main results.

Lemma 3.3 ([20]). *For a random variable* ξ *on* (Z,ρ) *with values in a Hilbert space* $(H,\|\cdot\|_H)$ *satisfying* $\|\xi\|_H \leq \tilde{M} < \infty$ *almost surely, and a random sample* $\{z_i\}_{i=1}^s$ *independently drawn according to* ρ*, there holds with confidence* $1-\delta$*,*

$$\left\|\frac{1}{s}\sum_{i=1}^{s}[\xi(z_i)-\mathbb{E}(\xi)]\right\|_H \leq \frac{2\tilde{M}\log(2/\delta)}{s} + \sqrt{\frac{2\mathbb{E}(\|\xi\|_H^2)\log(2/\delta)}{s}}. \tag{43}$$

Lemma 3.4. *Let* $\mathcal{A}$ *be a compact positive semi-definite operator on a separable Hilbert space, for any* $a>0$*,* $b>0$ *and* $\eta\|\mathcal{A}\| < 1$*, we have*

$$\left\|(I-\eta\mathcal{A})^a \mathcal{A}^b\right\| \leq \left(\frac{b}{ea\eta}\right)^b. \tag{44}$$

Proof. By the elementary inequality $1-x \le \exp(-x)$ for $x \ge 0$, then we have

$$\left\|(I-\eta\mathcal{A})^a\mathcal{A}^b\right\| \le \sup_{x\ge 0}\left((1-\eta x)^a x^b\right) \le \sup_{x\ge 0}\left(\exp(-\eta x a)x^b\right) \le \left(\frac{b}{ea\eta}\right)^b. \tag{45}$$

the last inequality holds since the function $\exp(-\eta x a)x^b$ attains its maximum at $x=\frac{b}{a\eta}$. □

Lemma 3.5. *([28]) Let $X_1,\cdots,X_d$ be i.i.d. standard Gaussian random variables, and $\mathbf{X}=(X_1,\cdots,X_d)^\top \in \mathbb{R}^d$. Then for any $t\in(0,1)$, with probability at least $1-\exp(-dt^2/8)$, there holds $\|\mathbf{X}\|_2^2 \le d(1+t)$.*

§4 Proof of main results

We prove our main results in this section. In the first part, we estimate the six terms G_1, G_2, G_3, G_4, G_5, G_6 in (32) respectively for the proof of Theorem 1. While the second part focuses on the privacy guarantee of algorithm 1.

4.1 Convergence analysis

In this subsection, we will prove the convergence rates of algorithm 1.

Proposition 4.1. *Let $c_{\delta_p,T}=\max\{\sqrt{3n\log(2n/\delta_p)/T}, 3n\log(2n/\delta_p)/T\}$, and*

$$\sigma^2 = \frac{8e\eta^2\log\frac{2.5}{\delta_p}}{\epsilon^2}\left(c\kappa+c\kappa^2\sqrt{\eta T}\right)^2\frac{T}{n}(1+c_{\delta_p,T})\left(1+\frac{T}{n}(1+c_{\delta_p,T})\right),$$

then with probability at least $1-\delta$, there holds

$$\begin{aligned} G_1 &= \mathbb{E}_{\mathbf{I}_T}\left[\|S_M\hat{w}_{priv}-S_M\hat{w}_{T+1}\|_\rho^2\right] \\ &\le \frac{8M\kappa^4 e\eta^2\log\frac{2.5}{\delta_p}}{\epsilon^2}\left(c+c\kappa\sqrt{\eta T}\right)^2\frac{T}{n}(1+c_{\delta_p,T})\left(1+\frac{T}{n}(1+c_{\delta_p,T})\right)\cdot\left(1+\sqrt{\frac{8\log(1/\delta)}{M}}\right). \end{aligned}$$

Proof. By the definition of $\hat{w}_{priv}$, we have

$$\|S_M\hat{w}_{priv}-S_M\hat{w}_{T+1}\|_\rho^2 = \|S_M(\hat{w}_{T+1}+b)-S_M\hat{w}_{T+1}\|_\rho^2 = \|S_Mb\|_\rho^2 \le \kappa^2\|b\|^2. \tag{46}$$

Since $b\sim\mathcal{N}(0,\sigma^2I_M)$, then for $\delta\in(0,1)$, Lemma 3.5 implies with probability at least $1-\delta$,

$$\|b\|^2 \le M\sigma^2\left(1+\sqrt{\frac{8\log(1/\delta)}{M}}\right) \tag{47}$$

Then the proof is completed by putting the above bound and σ^2 back into (46). □

Proposition 4.2. *Let $\delta\in(0,1), T\ge 3$, $\tilde{\lambda}\ge\frac{9\kappa^2}{M}\log\frac{M}{\delta}$ and the step size η satisfies*

$$0<\eta<\frac{1}{\kappa^2(\log T+1)}, \tag{48}$$

then with probability at least $1-\delta$, there holds

$$G_2=\mathbb{E}_{\mathbf{I}_T}\left[\|S_M(\hat{w}_{T+1}-\hat{v}_{T+1})\|_\rho^2\right] \le \frac{\kappa^2(3+4T\tilde{\lambda}\eta)}{(1-\eta\kappa^2)(1-\eta\kappa^2-\eta\kappa^2\log T)}\eta\log T, \quad \forall T\ge 3. \tag{49}$$

Proof. We borrow some ideas from [19,31] to prove this proposition. By the definition of $\{\hat{w}_t\}$ and $\{\hat{v}_t\}$, for any $1\le t\le T$, we have

$$\begin{aligned} \hat{w}_{t+1}-\hat{v}_{t+1} &= \hat{w}_t-\hat{v}_t+\eta(\hat{C}_M\hat{v}_t-\hat{S}_M^*\hat{y}-(\langle\hat{w}_t,\phi_M(x_{i_t}\rangle-y_{i_t}))\phi_M(x_{i_t})) \\ &= (I-\eta\hat{C}_M)(\hat{w}_t-\hat{v}_t)+\eta(\hat{C}_M\hat{w}_t-\hat{S}_M^*\hat{y}-(\langle\hat{w}_t,\phi_M(x_{i_t}\rangle-y_{i_t}))\phi_M(x_{i_t})). \end{aligned} \tag{50}$$

Let

$$\mathcal{M}_t := \hat{C}_M \hat{w}_t - \hat{S}_M^* \hat{y} - (\langle \hat{w}_t, \phi_M(x_{i_t}\rangle - y_{i_t}))\phi_M(x_{i_t}). \tag{51}$$

Then one can easily see that $\mathbb{E}_{i_t}[\mathcal{M}_t] = 0$ since $\hat{w}_t$ depends only on $\{z_{i_1}, z_{i_2}, \cdots, z_{i_{t-1}}\}$. Applying the relationship (50) iteratively, we have

$$\hat{w}_{T+1} - \hat{v}_{T+1} = (I - \eta\hat{C}_M)^T(\hat{w}_1 - \hat{v}_1) + \sum_{t=1}^{T} \eta \prod_{l=t+1}^{T} (I - \eta\hat{C}_M)\mathcal{M}_t = \sum_{t=1}^{T} \eta(I - \eta\hat{C}_M)^{T-t}\mathcal{M}_t.$$

Then we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{I}_T}\left[\|S_M\hat{w}_{T+1} - S_M\hat{v}_{T+1}\|_\rho^2\right] &= \mathbb{E}_{\mathbf{I}_T}\left[\left\|S_M \sum_{t=1}^{T} \eta(I - \eta\hat{C}_M)^{T-t}\mathcal{M}_t\right\|_\rho^2\right] \\
&= \sum_{t=1}^{T} \eta^2 \mathbb{E}_{\mathbf{I}_t}\left[\left\|S_M(I - \eta\hat{C}_M)^{T-t}\mathcal{M}_t\right\|_\rho^2\right] \\
&+ \sum_{t\neq t'} \mathbb{E}_{\mathbf{I}_{t'}} \langle {}_\rho S_M(I - \eta\hat{C}_M)^{T-t}\mathcal{M}_t, S_M(I - \eta\hat{C}_M)^{T-t'}\mathcal{M}_{t'}\rangle
\end{aligned}$$

Without loss of generality, let $t < t'$, then there holds

$$\begin{aligned}
&\mathbb{E}_{\mathbf{I}_{t'}} \langle S_M(I - \eta\hat{C}_M)^{T-t}\mathcal{M}_t, S_M(I - \eta\hat{C}_M)^{T-t'}\mathcal{M}_{t'}\rangle_\rho \\
&= \mathbb{E}_{\mathbf{I}_{t'-1}} \langle S_M(I - \eta\hat{C}_M)^{T-t}\mathcal{M}_t, S_M(I - \eta\hat{C}_M)^{T-t'}\mathbb{E}_{i_{t'}}[\mathcal{M}_{t'}]\rangle_\rho = 0.
\end{aligned}$$

When $t > t'$, we also have $\mathbb{E}_{\mathbf{I}_T}\langle S_M(I-\eta\hat{C}_M)^{T-t}\mathcal{M}_t, S_M(I-\eta\hat{C}_M)^{T-t'}\mathcal{M}_{t'}\rangle_\rho = 0$, which implies

$$\begin{aligned}
\mathbb{E}_{\mathbf{I}_T}\left[\|S_M\hat{w}_{T+1} - S_M\hat{v}_{T+1}\|_\rho^2\right] &= \sum_{t=1}^{T} \eta^2 \mathbb{E}_{\mathbf{I}_t}\left\|S_M(I - \eta\hat{C}_M)^{T-t}\mathcal{M}_t\right\|_\rho^2 \\
&\leq \sum_{t=1}^{T} \eta^2 \left\|C_M^{\frac{1}{2}}(I - \eta\hat{C}_M)^{T-t}\right\|^2 \mathbb{E}_{\mathbf{I}_t}\|\mathcal{M}_t\|_\rho^2.
\end{aligned} \tag{52}$$

In the following, we estimate the two terms $\left\|C_M^{\frac{1}{2}}(I - \eta\hat{C}_M)^{T-t}\right\|^2$ and $\mathbb{E}_{\mathbf{I}_t}\|\mathcal{M}_t\|_\rho^2$ respectively. First, by the elementary inequality $\mathbb{E}[\|\xi - \mathbb{E}\xi\|^2] \leq \mathbb{E}\|\xi\|^2$ for random variable ξ, for any $1 \leq t \leq T$, we have the following uniform bound

$$\begin{aligned}
\mathbb{E}_{\mathbf{I}_t}\left[\|\mathcal{M}_t\|_\rho^2\right] &\leq \mathbb{E}_{\mathbf{I}_t}\|(\langle \hat{w}_t, \phi_M(x_{i_t}\rangle - y_{i_t}))\phi_M(x_{i_t})\|_\rho^2 \leq \kappa^2 \mathbb{E}_{\mathbf{I}_t}\left[(\langle \hat{w}_t, \phi_M(x_{i_t})\rangle - y_{i_t})^2\right] \\
&= \kappa^2 \mathbb{E}_{\mathbf{I}_{t-1}}[\mathcal{E}_Z(\hat{w}_t)] \leq \kappa^2 \frac{c^2}{(1 - \eta\kappa^2)(1 - \eta\kappa^2 - \eta\kappa^2 \log T)}.
\end{aligned} \tag{53}$$

where the last inequality holds due to the bound for $\mathbb{E}_{\mathbf{I}_{t-1}}[\mathcal{E}_Z(\hat{w}_t)]$ from Lemma 3.1. Now we turn to consider the term $\left\|C_M^{\frac{1}{2}}(I - \eta\hat{C}_M)^{T-t}\right\|^2$ for $1 \leq t \leq T$. For any $\tilde{\lambda} > 0$,

$$\begin{aligned}
&\left\|C_M^{\frac{1}{2}}(I - \eta\hat{C}_M)^{T-t}\right\|^2 \\
&\leq \left\|C_M^{\frac{1}{2}}(\hat{C}_M + \tilde{\lambda}I)^{-\frac{1}{2}}\right\| \cdot \left\|(\hat{C}_M + \tilde{\lambda}I)^{\frac{1}{2}}(I - \eta\hat{C}_M)^{T-t}\right\| \\
&\leq \left\|(C_M + \tilde{\lambda}I)^{\frac{1}{2}}(\hat{C}_M + \tilde{\lambda}I)^{-\frac{1}{2}}\right\| \cdot \left\|(\hat{C}_M + \tilde{\lambda}I)^{\frac{1}{2}}(I - \eta\hat{C}_M)^{T-t}\right\| \\
&\leq 4\left\|(\hat{C}_M + \tilde{\lambda}I)(I - \eta\hat{C}_M)^{2(T-t)}\right\|,
\end{aligned}$$

where the last inequality holds due to $\left\|(C_M + \tilde{\lambda}I)^{\frac{1}{2}}(\hat{C}_M + \tilde{\lambda}I)^{-\frac{1}{2}}\right\| \leq 2$ for $\tilde{\lambda} \geq \frac{9\kappa^2}{M}\log\frac{M}{\delta}$ with probability at least $1 - \delta$ [5]. And for any $\tilde{\lambda} > 0$, and $1 \leq t \leq T-1$, by taking $\mathcal{A} = \hat{C}_M$, $b = 1$

and $a = 2(T-t)$ in Lemma 3.4, there holds

$$\left\|(\hat{C}_M + \tilde{\lambda}I)(I - \eta\hat{C}_M)^{2(T-t)}\right\| \le \left\|\hat{C}_M(I - \eta\hat{C}_M)^{2(T-t)}\right\| + \tilde{\lambda} \le \frac{1}{2e\eta(T-t)} + \tilde{\lambda}.$$

It follows that

$$\begin{aligned}\sum_{t=1}^{T}\left\|C_M^{\frac{1}{2}}(I-\eta\hat{C}_M)^{T-t}\right\|^2 &= \sum_{t=1}^{T-1}\left\|C_M^{\frac{1}{2}}(I-\eta\hat{C}_M)^{T-t}\right\|^2 + \left\|C_M^{\frac{1}{2}}\right\|^2 \\ &\le \sum_{t=1}^{T-1}\frac{2}{e\eta(T-t)} + 4(T-1)\tilde{\lambda} + \kappa^2 \le \frac{1+\log T}{\eta} + 4T\tilde{\lambda} + \kappa^2 \qquad (54)\\ &\le \frac{1}{\eta}(1 + \log T + 4T\tilde{\lambda}\eta + \eta\kappa^2).\end{aligned}$$

Then putting the bounds (53) and (54) back into (52) yields the desired result. □

We borrow some ideas from [5, 19] to get the following bound for G_3.

Proposition 4.3. *Under Assumption 2.1 and 2.2, let* $\delta \in (0, 1/5)$, $\hat{\lambda} \ge \frac{9\kappa^2}{M}\log\frac{M}{\delta}$, *and*

$$M \ge \left(4 + \frac{18\kappa^2}{\hat{\lambda}}\right)\log\frac{12\kappa^2}{\hat{\lambda}\delta} \qquad (55)$$

then with probability at least $1 - 5\delta$,

$$G_3 = \|S_M\hat{v}_{T+1} - S_M\tilde{v}_{T+1}\|_\rho^2 \le C_3\left(\frac{C_1}{\sqrt{\lambda}n} + \sqrt{\frac{C_2\mathcal{N}(\lambda)}{n}}\right)^2\log^2 T\log^2\frac{2}{\delta}, \qquad (56)$$

where $C_1 = \max(c,\kappa)^2$, $C_2 = 2C_1^2\max(2.55, \frac{2\kappa^2}{\|L\|})$, $C_3 = \left(4/e + 4 + \eta\kappa^2\right)^2\left(1 + 4R\kappa^{2r}\right)^2$.

Proof. By the definition of $\{\hat{v}_t\}$ and $\{\tilde{v}_t\}$, we have

$$\begin{aligned}\hat{v}_{T+1} - \tilde{v}_{T+1} &= \hat{v}_T - \tilde{v}_T + \eta[(C_M\tilde{v}_T - S_M^*f_\rho) - (\hat{C}_M\hat{v}_T - \hat{S}_M^*\hat{y})]\\ &= (I-\eta\hat{C}_M)(\hat{v}_T - \tilde{v}_T) + \eta[(C_M\tilde{v}_T - S_M^*f_\rho) - (\hat{C}_M\tilde{v}_T - \hat{S}_M^*\hat{y})]\\ &= (I-\eta\hat{C}_M)(\hat{v}_T - \tilde{v}_T) + \eta N_T\\ &= (I-\eta\hat{C}_M)^T(\hat{v}_1 - \tilde{v}_1) + \sum_{t=1}^{T}\eta\prod_{i=t+1}^{T}(I-\eta\hat{C}_M)N_t\\ &= \sum_{t=1}^{T}\eta(I-\eta\hat{C}_M)^{T-t}N_t,\end{aligned}$$

where we denote $N_t := (C_M\tilde{v}_t - S_M^*f_\rho) - (\hat{C}_M\tilde{v}_t - \hat{S}_M^*\hat{y})$ for $1 \le t \le T$. Then we have

$$\begin{aligned}\|S_M(\hat{v}_{T+1} - \tilde{v}_{T+1})\|_\rho &= \left\|C_M^{\frac{1}{2}}\sum_{t=1}^{T}\eta(I-\eta\hat{C}_M)^{T-t}N_t\right\|_\rho\\ &\le \sum_{t=1}^{T}\eta\left\|C_M^{\frac{1}{2}}(I-\eta\hat{C}_M)^{T-t}N_t\right\|_\rho + \eta\left\|C_M^{\frac{1}{2}}N_t\right\|_\rho.\end{aligned}$$

For $1 \le t \le T-1$, and any $\lambda^* > 0$,

$$\begin{aligned}
&\left\|C_M^{\frac{1}{2}}(I-\eta\hat{C}_M)^{T-t}N_t\right\|_\rho = \left\|C_M^{\frac{1}{2}}(\hat{C}_M+\lambda^*I)^{-\frac{1}{2}}(\hat{C}_M+\lambda^*I)^{\frac{1}{2}}\right.\\
&\left.(I-\eta\hat{C}_M)^{T-t}(\hat{C}_M+\lambda^*I)^{\frac{1}{2}}(\hat{C}_M+\hat{\lambda}I)^{-\frac{1}{2}}(C_M+\hat{\lambda}I)^{\frac{1}{2}}(C_M+\lambda^*I)^{-\frac{1}{2}}N_t\right\|_\rho\\
&\le \left\|(C_M+\hat{\lambda}I)^{\frac{1}{2}}(\hat{C}_M+\lambda^*I)^{-\frac{1}{2}}\right\|\cdot\left\|(\hat{C}_M+\lambda^*I)^{\frac{1}{2}}(I-\eta\hat{C}_M)^{T-t}(\hat{C}_M+\lambda^*I)^{\frac{1}{2}}\right\|\\
&\quad\cdot\left\|(\hat{C}_M+\hat{\lambda}I)^{-\frac{1}{2}}(C_M+\lambda^*I)^{\frac{1}{2}}\right\|\cdot\left\|(C_M+\lambda^*I)^{-\frac{1}{2}}N_t\right\|_\rho.
\end{aligned}$$

By Lemma 3 of [5], $\left\|(C_M+\lambda^*I)^{\frac{1}{2}}(\hat{C}_M+\hat{\lambda}I)^{-\frac{1}{2}}\right\| = \left\|(\hat{C}_M+\lambda^*I)^{-\frac{1}{2}}(C_M+\lambda^*I)^{\frac{1}{2}}\right\| \le 2$ holds with probability at least $1-\delta$ for $\lambda^* \ge \frac{9\kappa^2}{M}\log\frac{M}{\delta}$. Applying Lemma 3.4 with $\mathcal{A}=\hat{C}_M$, $\alpha=1$ and $\beta=T-t$, then for $1\le t<T-1$, we have

$$\begin{aligned}
&\left\|(\hat{C}_M+\hat{\lambda}I)^{\frac{1}{2}}(I-\eta\hat{C}_M)^{T-t}(\hat{C}_M+\lambda^*I)^{\frac{1}{2}}\right\| \le \left\|(\hat{C}_M+\lambda^*I)(I-\eta\hat{C}_M)^{T-t}\right\|\\
&\le \left\|\hat{C}_M(I-\eta\hat{C}_M)^{T-t}\right\| + \lambda^*\left\|(I-\eta\hat{C}_M)^{T-t}\right\|\\
&\le \frac{1}{e\eta(T-t)} + \lambda^*.
\end{aligned}$$

When $t=T$,

$$\begin{aligned}
&\left\|C_M^{\frac{1}{2}}(I-\eta\hat{C}_M)^{T-t}N_t\right\|_\rho = \left\|C_M^{\frac{1}{2}}N_T\right\|_\rho \le \left\|C_M^{\frac{1}{2}}(C_M+\lambda^*I)^{\frac{1}{2}}\right\|\cdot\left\|(C_M+\lambda^*I)^{-\frac{1}{2}}N_T\right\|_\rho\\
&\le \|C_M\|+\lambda^*\|\left\|(C_M+\hat{\lambda}I)^{-\frac{1}{2}}N_T\right\| \le (\kappa^2+\lambda^*)\left\|(C_M+\hat{\lambda}I)^{-\frac{1}{2}}N_T\right\|_\rho.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\|S_M(\hat{v}_{T+1}-\tilde{v}_{T+1})\|_\rho \le \sum_{t=1}^{T-1} 4\left(\frac{1}{e(T-t)}+\hat{\lambda}\eta\right)\left\|(C_M+\lambda^*I)^{-\frac{1}{2}}N_t\right\|_\rho\\
+\eta(\kappa^2+\lambda^*)\left\|(C_M+\lambda^*I)^{-\frac{1}{2}}N_T\right\|_\rho.
\end{aligned} \tag{57}$$

Now we consider the term $\left\|(C_M+\lambda^*I)^{-\frac{1}{2}}N_t\right\|_\rho$ for $1\le t\le T$,

$$\begin{aligned}
\left\|(C_M+\lambda^*I)^{-\frac{1}{2}}N_t\right\|_\rho &= \left\|(C_M+\lambda^*I)^{-\frac{1}{2}}(\hat{S}_M^*\hat{y}-S_M^*f_\rho+C_M\tilde{v}_t-\hat{C}_M\tilde{v}_t)\right\|_\rho\\
&\le \left\|(C_M+\lambda^*I)^{-\frac{1}{2}}(\hat{S}_M^*\hat{y}-S_M^*f_\rho)\right\|_\rho\\
&\quad+\left\|(C_M+\lambda^*I)^{-\frac{1}{2}}(C_M-\hat{C}_M)\right\|\cdot\|\tilde{v}_t\|.
\end{aligned} \tag{58}$$

In the following, we will estimate the three terms of the right hand side of (58) respectively. For the first term $\left\|(C_M+\lambda^*I)^{-\frac{1}{2}}(\hat{S}_M^*\hat{y}-S_M^*f_\rho)\right\|_\rho$, let $\xi(z_i)=(C_M+\lambda^*I)^{-\frac{1}{2}}\phi_M(x_i)y_i$, then

$$\begin{aligned}
\mathbb{E}\xi(z_i) &= (C_M+\lambda^*I)^{-\frac{1}{2}}\int\phi_M(x)yd\rho(x,y) = (C_M+\lambda^*I)^{-\frac{1}{2}}\int yd\rho(y|x)\phi_M(x)d\rho_{\mathcal{X}}(x)\\
&= (C_M+\lambda^*I)^{-\frac{1}{2}}\int\phi_M(x)f_\rho(x)d\rho_{\mathcal{X}}(x) = (C_M+\lambda^*I)^{-\frac{1}{2}}S_M^*f_\rho,
\end{aligned}$$

and $(C_M+\lambda^*I)^{-\frac{1}{2}}(\hat{S}_M^*\hat{y}-S_M^*f_\rho) = \frac{1}{n}\sum_{i=1}^n\xi(z_i)-\mathbb{E}[\xi]$. Apply Lemma 3.3 to the random variable ξ with $\tilde{M}=\frac{c\kappa}{\sqrt{\lambda^*}}$ and $\mathbb{E}\|\xi\|_\rho^2 \le c^2\mathcal{N}_M(\lambda^*)$, we know by (43) that with confidence at

least $1-\delta$,

$$\left\|(C_M+\lambda^* I)^{-\frac{1}{2}}(\hat{S}_M^*\hat{y}-S_M^* f_\rho)\right\|_\rho \le \frac{2c\kappa}{\sqrt{\lambda^* n}}\log\frac{2}{\delta}+\sqrt{\frac{2c^2\mathcal{N}_M(\lambda^*)\log\frac{2}{\delta}}{n}}.$$

The second term $\left\|(C_M+\lambda^* I)^{-\frac{1}{2}}(C_M-\hat{C}_M)\right\|$ can also be estimated by the same method as

$$\left\|(C_M+\lambda^* I)^{-\frac{1}{2}}(C_M-\hat{C}_M)\right\| \le \frac{2\kappa^2}{\sqrt{\lambda^* n}}\log\frac{2}{\delta}+\sqrt{\frac{2\kappa^2\mathcal{N}_M(\lambda^*)\log\frac{2}{\delta}}{n}}.$$

By Lemma 3.2, for $1\le t\le T$ we have

$$\|\tilde{v}_{t+1}\|_2 \le 2R\kappa^{2r-1}\left(1+\eta\lambda^* t\right).$$

Putting the above bounds back into (58) yields that

$$\begin{aligned}&\left\|(C_M+\lambda^* I)^{-\frac{1}{2}}N_t\right\|_\rho = \left\|(C_M+\lambda^* I)^{-\frac{1}{2}}(\hat{S}_M^*\hat{y}-S_M^* f_\rho+C_M\tilde{v}_t-\hat{C}_M\tilde{v}_t)\right\|_\rho\\ &\le \log\frac{2}{\delta}\left(\frac{2c\kappa}{\sqrt{\lambda^* n}}+\sqrt{\frac{2c^2\mathcal{N}_M(\lambda^*)}{n}}\right)+2R\kappa^{2r}\log\frac{2}{\delta}\left(\frac{2\kappa^2}{\sqrt{\lambda^* n}}+\sqrt{\frac{2\kappa^2\mathcal{N}_M(\lambda^*)}{n}}\right).\end{aligned}$$

holds at least $1-2\delta$. Putting the above bound back into (57), we have

$$\begin{aligned}&\|S_M\hat{v}_{T+1}-S_M\tilde{v}_{T+1}\|_\rho\\ &\le \sum_{t=1}^{T-1}4\left(\frac{1}{e(T-t)}+\lambda^*\eta\right)\left\|(C_M+\lambda^* I)^{-\frac{1}{2}}N_t\right\|_\rho+\eta(\kappa^2+\lambda^*)\left\|(C_M+\lambda^* I)^{-\frac{1}{2}}N_T\right\|_\rho\\ &\le \left(\sum_{t=1}^{T-1}4\left(\frac{1}{e(T-t)}+\lambda^*\eta\right)+\eta(\kappa^2+\lambda^*)\right)\sup_{1\le t\le T}\left\|(C_M+\lambda^* I)^{-\frac{1}{2}}N_t\right\|_\rho\\ &\le \left(4\log T/e+4\lambda^* T\eta+\eta\kappa^2\right)\left(\left(\frac{2c\kappa}{\sqrt{\lambda^* n}}+\sqrt{\frac{2c^2\mathcal{N}_M(\lambda^*)}{n}}\right)\right.\\ &\left.+2R\kappa^{2r}\left(\frac{2\kappa^2}{\sqrt{\lambda^* n}}+\sqrt{\frac{2\kappa^2\mathcal{N}_M(\lambda^*)}{n}}\right)\right)\log\frac{2}{\delta}.\end{aligned}$$

Moreover, we know from Lemma 4 of [5] that with probability at least $1-\delta$, there holds

$$\mathcal{N}_M(\lambda^*)\le\max\left\{2.55,\frac{2\kappa^2}{\|L\|}\right\}\mathcal{N}(\lambda^*)$$

when $M\ge(4+\frac{18\kappa^2}{\lambda^*})\log\frac{12\kappa^2}{\lambda^*\delta}$. The proof is completed by taking $C_1=\max(c,\kappa)^2$ and $C_2=2C_1^2\max\left\{2.55,\frac{2\kappa^2}{\|L\|}\right\}$. □

Proposition 4.4. *Let $\{\tilde{v}_t\}$ be defined by (25), then*

$$G_4=\left\|S_M\tilde{v}_{T+1}-L_M(L_M+\lambda I)^{-1}f_\rho\right\|_\rho^2\le 2(1+(\lambda\eta T)^{-1})^2(G_5+G_6).$$

Proof. Recall the expression (29) of $\{\tilde{v}_{T+1}\}$, i.e., $\tilde{v}_{T+1}=\sum_{t=1}^T\eta(I-\eta C_M)^{T-t}S_M^* f_\rho$, then

$$\begin{aligned}S_M\tilde{v}_{T+1}&=\sum_{t=1}^T\eta S_M(I-\eta C_M)^{T-t}S_M^* f_\rho\\ &=\sum_{t=1}^T\eta S_M S_M^*(I-\eta L_M)^{T-t}f_\rho=\sum_{t=1}^T\eta L_M(I-\eta L_M)^{T-t}f_\rho.\end{aligned}$$

Then we have

$$
\begin{aligned}
&\left\|S_M\tilde{v}_{T+1} - L_M(L_M+\lambda I)^{-1}f_\rho\right\|_\rho \\
&= \left\|\sum_{t=1}^{T}\eta(I-\eta L_M)^{T-t}(L_M+\lambda I)L_M(L_M+\lambda I)^{-1}f_\rho - L_M(L_M+\lambda I)^{-1}f_\rho\right\|_\rho \\
&= \left\|\left(\sum_{t=1}^{T}\eta(I-\eta L_M)^{T-t}(L_M+\lambda I) - I\right)L_M(L_M+\lambda I)^{-1}f_\rho\right\|_\rho \\
&= \left\|\left(\sum_{t=1}^{T}\eta(I-\eta L_M)^{T-t}(L_M+\lambda I) - I\right)L_M(L_M+\lambda I)^{-1}f_\rho\right\|_\rho \\
&= \left\|\left(\left(\sum_{t=1}^{T}\eta(I-\eta L_M)^{T-t}L_M - I\right)L_M + \lambda\sum_{t=1}^{T}\eta(I-\eta L_M)^{T-t}L_M\right)(L_M+\lambda I)^{-1}f_\rho\right\|_\rho \\
&\le \left(\left\|\left(\sum_{t=1}^{T}\eta(I-\eta L_M)^{T-t}L_M - I\right)L_M\right\| + \lambda\left\|\sum_{t=1}^{T}\eta(I-\eta L_M)^{T-t}L_M\right\|\right)\left\|(L_M+\lambda I)^{-1}f_\rho\right\|_\rho.
\end{aligned}
$$

Now we estimate the three parts of the right hand side of the above inequality respectively. For the first term, we have

$$
\begin{aligned}
&\left\|\left(\sum_{t=1}^{T}\eta(I-\eta L_M)^{T-t}L_M - I\right)L_M\right\| \\
&= \left\|\left(-\sum_{t=1}^{T}(I-\eta L_M)^{T-t}(I-\eta L_M) + \sum_{t=1}^{T}(I-\eta L_M)^{T-t} - I\right)L_M\right\| \\
&= \left\|\left(\sum_{t=1}^{T}(I-\eta L_M)^{T-t} - \sum_{t=1}^{T}(I-\eta L_M)^{T-t+1} - I\right)L_M\right\| \\
&= \left\|-(I-\eta L_M)^T L_M\right\| \le \frac{1}{\eta T}.
\end{aligned}
$$

For the second term, we have

$$
\begin{aligned}
\lambda\left\|\sum_{t=1}^{T}\eta(I-\eta L_M)^{T-t}L_M\right\| &= \lambda\left\|-\sum_{t=1}^{T}\eta(I-\eta L_M)^{T-t}(I-\eta L_M) + \sum_{t=1}^{T}\eta(I-\eta L_M)^{T-t}\right\| \\
&= \lambda\left\|I-(I-\eta L_M)^T\right\| \le \lambda.
\end{aligned}
$$

And for the third term, there holds

$$
\begin{aligned}
\lambda\left\|(L_M+\lambda I)^{-1}f_\rho\right\|_\rho &= \|(\lambda(L_M+\lambda I)^{-1} - \lambda(L+\lambda I)^{-1} + \lambda(L+\lambda I)^{-1})f_\rho\|_\rho \\
&\le \|(\lambda(L_M+\lambda I)^{-1} - \lambda(L+\lambda I)^{-1})f_\rho\| + \|\lambda(L+\lambda I)^{-1}f_\rho\|_\rho \\
&= \|L_M(L_M+\lambda I)^{-1}f_\rho - L(L+\lambda I)^{-1}f_\rho\|_\rho + \|L(L+\lambda I)^{-1}f_\rho - f_\rho\|_\rho
\end{aligned}
$$

where the last inequality holds due to $\lambda(A+\lambda I) = I - A(A+\lambda I)^{-1}$ for any bounded symmetric operator A and $\lambda > 0$. Combining the above three bounds together yields

$$
\begin{aligned}
G_4 &= \left\|S_M\tilde{v}_{T+1} - L_M(L_M+\lambda I)^{-1}f_\rho\right\|_\rho^2 \\
&\le (1+(\lambda\eta T)^{-1})^2\left(\left\|L_M(L_M+\lambda I)^{-1}f_\rho - L(L+\lambda I)^{-1}f_\rho\right\|_\rho + \left\|L(L+\lambda I)^{-1}f_\rho - f_\rho\right\|_\rho\right)^2.
\end{aligned}
$$

Then the desired result holds due to the definition of G_5 and G_6. □

Proposition 4.5. *Under Assumption 2.1 with* $1/2 < r \le 1$, *let* $0 < \delta < 1/2$, $\lambda \ge \frac{9\kappa^2}{M}\log\frac{M}{\delta}$, *and*

$$M \ge \left(4 + \frac{18\kappa^2}{\lambda}\right)\log\frac{12\kappa^2}{\lambda\delta} \tag{59}$$

then with probability at least $1-2\delta$, *there holds*

$$\begin{aligned} G_5 &= \left\|L_M(L_M+\lambda I)^{-1}f_\rho - L(L+\lambda I)^{-1}f_\rho\right\|_\rho^2 \\ &\le 128R^2\kappa^{4r-2}\left(\frac{\log^2\frac{2}{\delta}}{M^{2r}} + \frac{\lambda^{2r-1}(\mathcal{N}(\lambda))^{2r-1}\log\frac{2}{\delta}}{M}\right)\left(\log\frac{11\kappa^2}{\lambda}\right)^{2-2r}. \end{aligned}$$

Proof. By the identity $A(A+\lambda I)^{-1} = I - \lambda(A+\lambda I)^{-1}$ for any bounded positive operator A, and $A^{-1}-B^{-1} = A^{-1}(B-A)B^{-1} = B^{-1}(B-A)A^{-1}$ for any invertible bounded operators A and B, we have

$$\begin{aligned} L_M(L_M+\lambda I)^{-1} - L(L+\lambda I)^{-1} &= \lambda((L+\lambda I)^{-1} - (L_M+\lambda I)^{-1}) \\ &= \lambda(L_M+\lambda I)^{-1}(L_M - L)(L+\lambda I)^{-1}. \end{aligned}$$

By the regularity assumption (14) with $\frac{1}{2} < r \le 1$, i.e., $f_\rho = L^r g_\rho$ with $g_\rho \in L^2_{\rho_{\mathcal{X}}}$ and $\|g_\rho\|_\rho \le R$, we have the following decomposition

$$\begin{aligned} &\left\|(L_M(L_M+\lambda I)^{-1} - L(L+\lambda I)^{-1})f_\rho\right\|_\rho = \left\|\lambda(L_M+\lambda I)^{-1}(L_M-L)(L+\lambda I)^{-1}f_\rho\right\|_\rho \\ &= \left\|\sqrt{\lambda}\sqrt{\lambda}(L_M+\lambda I)^{-\frac{1}{2}}(L_M+\lambda I)^{-\frac{1}{2}}(L+\lambda I)^{\frac{1}{2}}(L+\lambda I)^{-\frac{1}{2}}(L-L_M)(L+\lambda I)^{r-1}(L+\lambda I)^{-r}\right. \\ &\quad \left.\cdot L^r g_\rho\right\|_\rho \\ &\le \sqrt{\lambda}\left\|\sqrt{\lambda}(L_M+\lambda I)^{-\frac{1}{2}}\right\|\left\|(L_M+\lambda I)^{-\frac{1}{2}}(L+\lambda I)^{\frac{1}{2}}\right\|\left\|(L+\lambda I)^{-\frac{1}{2}}(L-L_M)(L+\lambda I)^{-(1-r)}\right\| \\ &\quad \cdot \left\|(L+\lambda I)^{-r}L^r\right\|\|g_\rho\|_\rho \\ &\le 2R\sqrt{\lambda}\left\|(L+\lambda I)^{-\frac{1}{2}}(L-L_M)\right\|^{2r-1}\left\|(L+\lambda I)^{-\frac{1}{2}}(L-L_M)(L+\lambda I)^{-\frac{1}{2}}\right\|^{2-2r}, \end{aligned} \tag{60}$$

the last inequality holds due to $\left\|\sqrt{\lambda}(L_M+\lambda I)^{-\frac{1}{2}}\right\| \le 1$, $\|(L+\lambda I)^{-r}L^r\| \le 1$ for any $\lambda > 0$, $\left\|(L_M+\lambda I)^{-\frac{1}{2}}(L+\lambda I)^{\frac{1}{2}}\right\| \le 2$ for any $\lambda \ge \frac{9\kappa^2}{M}\log\frac{M}{\delta}$ with probability at least $1-\delta$, $\|g_\rho\|_\rho \le R$, and $\left\|(L+\lambda I)^{-\frac{1}{2}}(L-L_M)(L+\lambda I)^{-(1-r)}\right\| \le \left\|(L+\lambda I)^{-\frac{1}{2}}(L-L_M)\right\|^{2r-1}\cdot\left\|(L+\lambda I)^{-\frac{1}{2}}(L-L_M)(L+\lambda I)^{-\frac{1}{2}}\right\|^{2-2r}$ by Proposition 9 of [22]. Moreover, by Lemma 8 of [22], for $M \ge \left(4+\frac{18\kappa^2}{\lambda}\right)\log\frac{12\kappa^2}{\lambda\delta}$, with confidence at least $1-2\delta$, there holds

$$\begin{aligned} &\sqrt{\lambda}\left\|(L+\lambda I)^{-\frac{1}{2}}(L-L_M)\right\|^{2r-1}\left\|(L+\lambda I)^{-\frac{1}{2}}(L-L_M)(L+\lambda I)^{-\frac{1}{2}}\right\|^{2-2r} \\ &\le 4\kappa^{2r-1}\left(\frac{\log\frac{2}{\delta}}{M^r} + \sqrt{\frac{\lambda^{2r-1}\mathcal{N}(\lambda)^{2r-1}\log\frac{2}{\delta}}{M}}\right)\left(\log\frac{11\kappa^2}{\lambda}\right)^{1-r}. \end{aligned} \tag{61}$$

Then the proof is finished by putting the above bound back into (60). □

Proposition 4.6. *Under Assumption 2.1 with* $1/2 < r \le 1$, *we have*

$$G_6 = \left\|L(L+\lambda I)^{-1}f_\rho - f_\rho\right\|_\rho^2 \le R^2\lambda^{2r}. \tag{62}$$

Proof. By the identity $A(A+\lambda I)^{-1} = I - \lambda(A+\lambda I)^{-1}$ for any $\lambda > 0$ and any bounded self-

adjoint positive operator A, and the assumption $f_\rho = L^r g_\rho$ with $1/2 < r \leq 1$, and $g_\rho \in L^2_{\rho_\mathcal{X}}$ and $\|g_\rho\|_\rho \leq R$, then we have

$$\begin{aligned}\left\|(L(L+\lambda I)^{-1} - I)f_\rho\right\|_\rho &= \left\|-\lambda(L+\lambda I)^{-1} f_\rho\right\|_\rho = \left\|\lambda(L+\lambda I)^{-1} L^r g_\rho\right\|_\rho \\ &= \left\|\lambda^r (\lambda^{1-r}(L+\lambda I)^{-(1-r)})((L+\lambda I)^{-r} L^r) g_\rho\right\|_\rho \leq R\lambda^r\end{aligned}$$

where the last inequality holds due to $\left\|\lambda^{1-r}(L+\lambda I)^{-(1-r)}\right\| \leq 1$, $\|(L+\lambda I)^{-r} L^r\| \leq 1$, and $\|g_\rho\|_\rho \leq R$. □

Now we are ready to prove our convergence rates of algorithm 1.

Proof of Theorem 1. Under Assumption 2.1 with $1/2 < r \leq 1$, let $\tilde{\lambda} = \hat{\lambda} = \lambda^* = \lambda \simeq \frac{1}{\eta T} \log \frac{n}{\delta}$, and Propositions 4.1, 4.2, 4.3, 4.4, 4.5,4.6 to the error decomposition (33), then with probability at least $1 - 9\delta$, there holds

$$\begin{aligned}&\mathbb{E}_{\mathbf{I}_T}\left(\mathcal{E}(\hat{f}_{priv}) - \mathcal{E}(f_\rho)\right) = \mathbb{E}_{\mathbf{I}_T}\left[\|f_{priv} - f_\rho\|_\rho^2\right] \\ &\leq \frac{48M\kappa^4 e\eta^2 \log\frac{2.5}{\delta_p}}{\epsilon^2}\left(c + c\kappa\sqrt{\eta T}\right)^2 \frac{T}{n}(1 + c_{\delta_p,T})\left(1 + \frac{T}{n}(1 + c_{\delta_p,T})\right)\left(1 + \sqrt{\frac{8\log(1/\delta)}{M}}\right) \\ &+ \frac{42\kappa^2}{(1-\eta\kappa^2)(1-\eta\kappa^2 - \eta\kappa^2\log T)}\eta\log T + 6C_3\left(\frac{C_1}{\sqrt{\lambda}n} + \sqrt{\frac{C_2\mathcal{N}(\lambda)}{n}}\right)^2 \log^2 T \log^2\frac{2}{\delta} \\ &+ 6912R^2\kappa^{4r-2}\left(\frac{\log^2\frac{2}{\delta}}{M^{2r}} + \frac{\lambda^{2r-1}(\mathcal{N}(\lambda))^{2r-1}\log\frac{2}{\delta}}{M}\right)\left(\log\frac{11\kappa^2}{\lambda}\right)^{2-2r} + 54R(\eta T)^{-2r}.\end{aligned}$$

Then under Assumption 2.2 with $0 < \alpha < 1$, i.e., $\mathcal{N}(\lambda) \leq C_0\lambda^{-\alpha}$ with $0 < \alpha < 1$, and we take $M \simeq n^{\frac{1+\alpha(2r-1)}{2r+\alpha}} \log\frac{n}{\delta}$, and $\eta = n^{-\frac{2r}{2r+\alpha}}$ and $T = n^{\frac{2r+1}{2r+\alpha}}$; or $\eta = n^{-1}$ and $T = n^{\frac{2r+\alpha+1}{2r+\alpha}}$, in both cases, we have $T > n$, $\eta T = n^{\frac{1}{2r+\alpha}}$ $1 + c_{\delta_p,T} = 1 + \max\{\sqrt{3n\log(2n/\delta_p)/T}, 3n\log(2n/\delta_p)/T\} \leq 1 + 3\log(2n/\delta_p) \leq 4\log(2n/\delta_p)$, and when $\alpha > (4-4r)/(3-2r)$, then with confidence at least $1 - 9\delta$, there holds

$$\begin{aligned}&\mathbb{E}_{\mathbf{I}_T}\left(\mathcal{E}(\hat{f}_{priv}) - \mathcal{E}(f_\rho)\right) \\ &\leq \mathcal{O}\left(n^{-\frac{4r+3\alpha-4-2r\alpha}{2r+\alpha}}\frac{1}{\epsilon^2}\log\frac{n}{\delta}\log\frac{2n}{\delta_p}\log\frac{2.5}{\delta_p}\sqrt{\log\frac{2}{\delta}} + n^{-\frac{2r}{2r+\alpha}}\log^2 n\log^2\frac{2}{\delta}\right).\end{aligned}$$

The proof is completed by scaling 9δ to δ. □

4.2 Analysis of the performance on privacy protection

In this subsection, we borrow some ideas from [29] to prove the differential privacy of algorithm 1.

Proof of Theorem 2. Assume that $\mathcal{S}$ and $\mathcal{S}'$ differ by the i-th datum, i.e.,$z_i \neq z_i'$. Let $\{\hat{w}_t\}_{t=1}^T$ and $\{\hat{w}_t'\}_{t=1}^T$ be the sequence produced by SGD update (5) based on $\mathcal{S}$ and $\mathcal{S}'$ respectively. For any $1 \leq t \leq T$, we consider the following two cases.

Case 1: $i_t \neq i$. We have

$$\begin{aligned}&\|\hat{w}_{t+1} - \hat{w}_{t+1}'\|_2^2 \\ &= \|\hat{w}_t - \eta(\langle\hat{w}_t, \phi_M(x_{i_t})\rangle - y_{i_t})\phi_M(x_{i_t}) - \hat{w}_t' + \eta(\langle\hat{w}_t', \phi_M(x_{i_t})\rangle - y_{i_t})\phi_M(x_{i_t})\|_2^2 \\ &= \|\hat{w}_t - \hat{w}_t' - \eta\phi_M(x_{i_t})\langle\hat{w}_t - \hat{w}_t', \phi_M(x_{i_t})\rangle\|_2^2 \\ &= \|\hat{w}_t - \hat{w}_t'\|_2^2 + \eta^2\|\phi_M(x_{i_t})\langle\hat{w}_t - \hat{w}_t', \phi_M(x_{i_t})\rangle\|_2^2 - 2\eta|\langle\hat{w}_t - \hat{w}_t', \phi_M(x_{i_t})\rangle|^2\end{aligned}$$

$$\leq \|\hat{w}_t - \hat{w}'_t\|_2^2 + \eta^2\kappa^2|\langle \hat{w}_t - \hat{w}'_t, \phi_M(x_{i_t})\rangle|^2 - 2\eta|\langle \hat{w}_t - \hat{w}'_t, \phi_M(x_{i_t})\rangle|^2$$
$$\leq \|\hat{w}_t - \hat{w}'_t\|_2^2$$

where the last inequality holds due to $\eta\kappa^2 < 1$.

Case 2: $i_t = i$.

From the elementary inequality $(a+b)^2 \leq (1+p)a^2 + \left(1+\frac{1}{p}\right)b^2$ for any $p > 0$ and $a, b \in \mathbb{R}$, we have

$$\|\hat{w}_{t+1} - \hat{w}'_{t+1}\|_2^2$$
$$= \|\hat{w}_t - \eta(\langle \hat{w}_t, \phi_M(x_{i_t})\rangle - y_{i_t})\phi_M(x_{i_t}) - \hat{w}'_t + \eta(\langle \hat{w}'_t, \phi_M(x'_{i_t})\rangle - y'_{i_t})\phi_M(x'_{i_t})\|_2^2$$
$$\leq (1+p)\|\hat{w}_t - \hat{w}'_t\|_2^2$$
$$+ \left(1+\frac{1}{p}\right)\eta^2\|\phi_M(x_{i_t})(\langle \hat{w}_t, \phi_M(x_{i_t})\rangle - y_{i_t}) - \phi_M(x'_{i_t})(\langle \hat{w}'_t, \phi_M(x'_{i_t})\rangle - y'_{i_t})\|_2^2$$

and

$$\|\phi_M(x_{i_t})(\langle \hat{w}_t, \phi_M(x_{i_t})\rangle - y_{i_t})\|_2 \leq \kappa^2\|\hat{w}_t\|_2 + c\kappa. \tag{63}$$

Now we consider the bound of $\|\hat{w}_t\|_2$, from the definition of $\hat{w}_t$, we have

$$\|\hat{w}_{t+1}\|_2^2 = \|\hat{w}_t - \eta(\langle \hat{w}_t, \phi_M(x_{i_t})\rangle - y_{i_t})\phi_M(x_{i_t})\|_2^2$$
$$= \|\hat{w}_t\|_2^2 + \eta^2\|(\langle \hat{w}_t, \phi_M(x_{i_t})\rangle - y_{i_t})\phi_M(x_{i_t})\|_2^2 - 2\eta\langle \hat{w}_t, (\langle \hat{w}_t, \phi_M(x_{i_t})\rangle - y_{i_t})\phi_M(x_{i_t})\rangle.$$

One can easily see that

$$-2\langle \hat{w}_t, (\langle \hat{w}_t, \phi_M(x_{i_t})\rangle - y_{i_t})\phi_M(x_{i_t})\rangle = 2\langle 0 - \hat{w}_t, (\langle \hat{w}_t, \phi_M(x_{i_t})\rangle - y_{i_t})\phi_M(x_{i_t})\rangle$$
$$\leq (y_{i_t})^2 - (\hat{w}_t, \phi_M(x_{i_t})\rangle - y_{i_t})^2$$

and

$$\eta^2\|(\langle \hat{w}_t, \phi_M(x_{i_t})\rangle - y_{i_t})\phi_M(x_{i_t})\|_2^2 \leq \eta^2\kappa^2(\langle \hat{w}_t, \phi_M(x_{i_t})\rangle - y_{i_t})^2$$

and since $\eta\kappa^2 < 1$ we have

$$\|\hat{w}_{t+1}\|_2^2 \leq \|\hat{w}_t\|_2^2 + (\eta^2\kappa^2 - \eta)\left(\langle \hat{w}_t, \phi_M(x_{i_t})\rangle - y_{i_t}\right)^2 + \eta c^2 \leq \eta t c^2$$

so there holds

$$\|\phi_M(x_{i_t})(\langle \hat{w}_t, \phi_M(x_{i_t})\rangle - y_{i_t})\|_2 \leq c\kappa + c\kappa^2\sqrt{\eta t},$$

then

$$\|\hat{w}_{t+1} - \hat{w}'_{t+1}\|_2^2 \leq (1+p)\|\hat{w}_t - \hat{w}'_t\|_2^2 + 4\left(1+\frac{1}{p}\right)\eta^2(c\kappa + c\kappa^2\sqrt{\eta t})^2 \tag{64}$$

Combining case 1 and case 2, we have

$$\|\hat{w}_{t+1} - \hat{w}'_{t+1}\|_2^2 \leq (1+p)^{I(i_t=i)}\|\hat{w}_t - \hat{w}'_t\|_2^2 + 4\left(1+\frac{1}{p}\right)I(i_t = i)\eta^2(c\kappa + c\kappa^2\sqrt{\eta t})^2$$
$$\leq \prod_{k=1}^{t}(1+p)^{I(i_k=i)}\|\hat{w}_1 - \hat{w}'_1\|_2^2$$
$$+ 4\left(1+\frac{1}{p}\right)\eta^2(c\kappa + c\kappa^2\sqrt{\eta t})^2\sum_{k=1}^{t}I(i_k = i)\prod_{j=k+1}^{t}(1+p)^{I(i_j=i)}$$
$$\leq 4\left(1+\frac{1}{p}\right)\eta^2(c\kappa + c\kappa^2\sqrt{\eta t})^2(1+p)^{\sum_{j=2}^{t}I(i_j=i)}\sum_{k=1}^{t}I(i_k = i).$$

Let $X_j = I(i_j = i)$ and $X = \sum_{j=1}^{t}X_j$, using Chernoff bound for Bernoulli random variable [28],

for any $\exp(-t/3n) \le \gamma \le 1$, with probability at least $1-\frac{\gamma}{n}$, there holds

$$\sum_{j=1}^{t} I(i_j = i) \le \frac{t}{n}\left(1+\sqrt{\frac{3\log(n/\gamma)}{t/n}}\right).$$

For any $0 < \gamma < \exp(-t/3n)$, with probability at least $1-\frac{\gamma}{n}$, there holds

$$\sum_{j=1}^{t} I(i_j = i) \le \frac{t}{n}\left(1+\frac{3\log(n/\gamma)}{t/n}\right).$$

Let $c_{\gamma,t} = \max\left\{\sqrt{\frac{3\log(n/\gamma)}{t/n}}, \frac{3\log(n/\gamma)}{t/n}\right\}$, then with probability at least $1-\frac{\gamma}{n}$, there holds

$$\|\hat{w}_{t+1} - \hat{w}'_{t+1}\|_2^2 \le 4\left(1+\frac{1}{p}\right)\eta^2(c\kappa + c\kappa^2\sqrt{\eta t})^2(1+p)^{\frac{t}{n}(1+c_{\gamma,t})}\frac{t}{n}(1+c_{\gamma,t}) \tag{65}$$

Let $p = \frac{1}{\frac{t}{n}(1+c_{\gamma,t})}$ then $(1+p)^{\frac{t}{n}(1+c_{\gamma,t})} \le e$ so

$$\|\hat{w}_{t+1} - \hat{w}'_{t+1}\|_2^2 \le 4e\eta^2(c\kappa + c\kappa^2\sqrt{\eta t})^2\frac{t}{n}(1+c_{\gamma,t})\left(1+\frac{t}{n}(1+c_{\gamma,t})\right)$$

By taking a union bound of probabilities over $i = 1, 2, \cdots, n$, with probability at least $1-\gamma$, there holds

$$\sup_{\mathcal{S}\simeq\mathcal{S}'} \|\hat{w}_{T+1} - \hat{w}'_{T+1}\|_2^2 \le 4e\eta^2\left(c\kappa + c\kappa^2\sqrt{\eta T}\right)^2\frac{T}{n}(1+c_{\gamma,T})\left(1+\frac{T}{n}(1+c_{\gamma,T})\right) := \Delta^2_{SGD}(\gamma)$$

The proof of the theorem is complete. □

Lemma 4.1 (Post-Processing [11]). *Let $\mathcal{A}: \mathcal{Z}^n \to \Omega$ be a randomized algorithm that is (ϵ, δ_p)-differentially private. Let $f: \Omega \to \mathcal{R}$ be an arbitrary randomized mapping. Then $f \circ \mathcal{A}: \mathcal{Z}^n \to \mathcal{R}$ is (ϵ, δ_p)-differentially private.*

Now we are in a position to prove Theorem 3.

Proof of Theorem 3. Let $\mathbf{I}_T = \{i_1, i_2, \cdots, i_T\}$ and $\delta_{\mathcal{A}}(\mathcal{S}, \mathcal{S}') = \|\mathcal{A}(\mathcal{S}) - \mathcal{A}(\mathcal{S}')\|_2$. Define

$$\mathcal{B} = \{\mathbf{I}_T : \sup_{\mathcal{S}\simeq\mathcal{S}'} \delta_{\mathcal{A}}(\mathcal{S}, \mathcal{S}') \le \Delta_{SGD}(\delta_p/2)\},$$

then Theorem 2 tells us that $\mathbb{P}(\mathbf{I}_T \in \mathcal{B}) \ge 1 - \frac{\delta_p}{2}$, then we have

$$\begin{aligned}
\mathbb{P}(\hat{w}_{priv} \in E) &= \mathbb{P}(\hat{w}_{priv} \in E \cap \mathbf{I}_T \in \mathcal{B}) + P(\hat{w}_{priv} \in E \cap \mathbf{I}_T \in \mathcal{B}^c) \\
&\le \mathbb{P}(\hat{w}_{priv} \in E | \mathbf{I}_T \in \mathcal{B})\mathbb{P}(\mathbf{I}_T \in \mathcal{B}) + \frac{\delta_p}{2} \\
&\le \left(e^{\epsilon}\mathbb{P}(\hat{w}'_{priv} \in E | \mathbf{I}_T \in \mathcal{B}) + \frac{\delta_p}{2}\right)\mathbb{P}(\mathbf{I}_T \in \mathcal{B}) + \frac{\delta_p}{2} \\
&\le e^{\epsilon}\mathbb{P}(\hat{w}'_{priv} \in E \cap \mathbf{I}_T \in \mathcal{B}) + \delta_p \\
&\le e^{\epsilon}\mathbb{P}(\hat{w}'_{priv} \in E) + \delta_p.
\end{aligned}$$

The proof is completed by applying Lemma 4.1 with $f = \hat{f}_{priv}(\cdot) = \langle \hat{w}_{priv}, \phi_M(\cdot)\rangle$. □

Declarations

Conflict of interest The authors declare no conflict of interest.

References

[1] N Aronszajn. *Theory of reproducing kernels*, Transactions of the American mathematical society, 1950, 68(3): 337-404.

[2] R Bassily, V Feldman, C Guzmán, K Talwar. *Stability of stochastic gradient descent on nonsmooth convex losses*, Advances in Neural Information Processing Systems, 2020, 33: 4381-4391.

[3] R Bassily, V Feldman, K Talwar, A Thakurta. *Private stochastic convex optimization with optimal rates*, Advances in Neural Information Processing Systems, 2019, 32.

[4] L Bottou, O Bousquet. *The tradeoffs of large scale learning*, Advances in Neural Information Processing Systems, 2007, 20.

[5] L Carratino, A Rudi, L Rosasco. *Learning with sgd and random features*, Advances in Neural Information Processing Systems, 2018, 31.

[6] K Chaudhuri, C Monteleoni, A D Sarwate. *Differentially private empirical risk minimization*, Journal of Machine Learning Research, 2011, 12(3): 1069-1109.

[7] X Chen, B Tang, J Fan, X Guo. *Online gradient descent algorithms for functional data learning*, Journal of Complexity, 2022, 70: 101635.

[8] F Cucker, D X Zhou. *Learning theory: an approximation theory viewpoint*, Cambridge University Press, 2007.

[9] A Dieuleveut, F Bach. *Nonparametric stochastic approximation with large step-sizes*, The Annals of Statistics, 2016, 44(4): 1363-1399.

[10] C Dwork, F McSherry, K Nissim, A Smith. *Calibrating noise to sensitivity in private data analysis*, In Theory of Cryptography Conference, 2006, 265-284.

[11] C Dwork, A Roth. *The algorithmic foundations of differential privacy*, Foundations and Trends® in Theoretical Computer Science, 2014, 9(3-4): 211-407.

[12] V Feldman, T Koren, K Talwar. *Private stochastic convex optimization: optimal rates in linear time*, Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, 2020, 439-449.

[13] X Guo, Z C Guo, L Shi. *Capacity dependent analysis for functional online learning algorithms*, Applied and Computational Harmonic Analysis, 2023, 67: 101567.

[14] Z C Guo, A Christmann, L Shi. *Optimality of robust online learning*, Foundations of Computational Mathematics, 2023.

[15] Z C Guo, L Shi. *Fast and strong convergence of online learning algorithms*, Advances in Computational Mathematics, 2019, 45: 2745-2770.

[16] P Jain, A Thakurta. *Differentially private learning with kernels*, In International Conference on Machine Learning, 2013, 118-126.

[17] Y Lei, L Shi, Z C Guo. *Convergence of unregularized online learning algorithms*, The Journal of Machine Learning Research, 2017, 18(1): 6269-6301.

[18] Y Lei, Y Ying. *Fine-grained analysis of stability and generalization for stochastic gradient descent*, In International Conference on Machine Learning, 2020, 5809-5819.

[19] J Lin, L Rosasco. *Optimal rates for multi-pass stochastic gradient methods*, The Journal of Machine Learning Research, 2017, 18(1): 3375-3421.

[20] I Pinelis. *Optimum bounds for the distributions of martingales in banach spaces*, The Annals of Probability, 1994, 1679-1706.

[21] A Rahimi, B Recht. *Random features for large-scale kernel machines*, Advances in Neural Information Processing Systems, 2007, 20.

[22] A Rudi, L Rosasco. *Generalization properties of learning with random features*, Advances in Neural Information Processing Systems, 2017, 30.

[23] B Schölkopf, A J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT press, 2002.

[24] O Shamir, T Zhang. *Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes*, In International Conference on Machine Learning, 2013, 71-79.

[25] S Smale, D X Zhou. *Learning theory estimates via integral operators and their approximations*, Constructive approximation, 2007, 26(2): 153-172.

[26] B Sriperumbudur, Z Szabó. *Optimal rates for random fourier features*, Advances in Neural Information Processing Systems, 2015, 28.

[27] I Sutskever, J Martens, G Dahl, G Hinton. *On the importance of initialization and momentum in deep learning*, In International Conference on Machine Learning, 2013, 1139-1147.

[28] M J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, Cambridge University Press, 2019.

[29] P Wang, Y Lei, Y Ying, H Zhang. *Differentially private sgd with non-smooth losses*, Applied and Computational Harmonic Analysis, 2022, 56: 306-336.

[30] X Wu, F Li, A Kumar, K Chaudhuri, S Jha, J Naughton. *Bolt-on differential privacy for scalable stochastic gradient descent-based analytics*, In Proceedings of the 2017 ACM International Conference on Management of Data, 2017, 1307-1322.

[31] Y Ying, M Pontil. *Online gradient descent learning algorithms*, Foundations of Computational Mathematics, 2008, 8: 561-596.

[32] Y Ying, D X Zhou. *Online regularized classification algorithms*, IEEE Transactions on Information Theory, 2006, 52(11): 4775-4788.

[1]Polytechnic Institute of Zhejiang University, Zhejiang University, Hangzhou 310015, China.
Email: yiguang@zju.edu.cn
[2]School of Mathematical Sciences, Zhejiang University, Hangzhou 310058, China.
Email: guozhengchu@zju.edu.cn