Differential Privacy via Randomized Sketching

Yuheng Ma
Renmin University of China
yma@ruc.edu.cn

April 22, 2024

Abstract

This is a failed attempt towards achieving differential privacy with randomized sketched kernel ridge regression. The rest of this paper is organized as follows. After the introduction of some notations, we give thorough background knowledge in empirical risk minimization and randomized sketching in Section 1. In section 2 and 3 are the theoretical findings involving utility and privacy. All technical proof is postponed to appendix and our code is available at 5.

1 Methodology and Related Work

1.1 Notations

We will use $\|\cdot\|_2, \|\cdot\|_\infty$, and $\|\cdot\|_{\mathcal{H}}$ to denote the ℓ_2-norm, the infinity norm, and norm in a Hilbert space $\mathcal{H}$, respectively. For matrices, we will use the notation $\|\cdot\|_2$ and $\|\cdot\|_F$ to denote the 2-norm and frobenius norm, respectively. $A^\top$ is the transpose of matrix/vector A. Let I_n denote an $n \times n$ identity matrix. We use $|\cdot|$ to denote the determinant. Let $\mathrm{col}(A)$ be the column space of matrix A. With a slight abuse of notation, 0 represents a constant/vector/matrix with the desired shape whose entries are all zero. Let $A_{:,i:j}$ be the $i:j$ column slicing of A, i.e. the sub-matrix consists of $i, i+1, \cdots, j$-th columns of A. Analogously, let $A_{i:j,:}$ be the $i:j$ row slicing of A, i.e. the sub-matrix consists of $i, i+1, \cdots, j$-th rows of A. We use $[n]$ to denote $\{1, 2, \cdots, n\}$. Let $\mathrm{span}(a_1, \cdots, a_n)$ be the linear space spanned by $a_1, \cdots, a_n$. For a matrix A, let $\lambda_i(A)$ be the i-th largest eigenvalue of A. Without special notification, let A^{-1} denote the pseudo inverse of matrix A. In the sequel, the notation $a_n = \mathcal{O}(b_n)$ denotes that there exists some positive constant c such that $a_n \leq cb_n$. Let $A \triangle B$ denote the difference set of sets A and B. We use $\cong$ to indicate two algebraic structures are isomorphic. Let $a \vee b$ denote $\max(a, b)$. We use V_d to denote the volume of the unit ball in d dimensional space.

1.2 Empirical Risk Minimization

We consider classical regularized empirical risk minimization problem, in which we are given a dataset $\mathcal{D} := \{(x_1, y_1), \ldots, (x_n, y_n)\}$ consisting of n i.i.d. samples drawn from an unknown probability measure $\mathrm{P}_{X,Y}$ on $\mathcal{X} \times \mathcal{Y}$. Throughout this paper, we assume that $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$ are compact and non-empty. To be more specific, we have $\|x\|_2 \leq 1$ for all $x \in \mathcal{X}$ and $|y| \leq 1$ for all $y \in \mathcal{Y}$. We desire to find a predictor $f : \mathcal{X} \to \mathcal{Y}$. f is constraint in a Hilbert hypothesis function class $\mathcal{H}$ with norm $\|f\|_{\mathcal{H}}$. We measure the quality of our predictor on the training data via a nonnegative loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Moreover, we denote the Bayesian risk as $\ell^* := \min_{f \in \mathcal{H}} \mathbb{E}[\ell(f(x), y)]$ and its minimizer is f^*. Based on the definitions, the *regularized empirical risk minimization* problem intends to minimize

$$L(f) + \lambda \|f\|_{\mathcal{H}}^2 = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2 \tag{1}$$

over $f \in \mathcal{H}$. $L(f)$ aims to find an accurate predictor to fit D_n while the regularizer $\|f\|_{\mathcal{H}}$ prevents over-fitting. Hyper-parameter λ, which is fixed beforehand by the user and possibly depending on n, balances the trade-off between two parts.

In this paper, we consider $\mathcal{H}$ to be linear functionals in a Reproducing Kernel Hilbert Space (RKHS) which corresponds to a positive semidefinite kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. For each positive semidefinite kernel k, we have $k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle$. Here $\phi : \mathcal{X} \to \mathcal{H}$ is called the feature map. $\langle \cdot, \cdot \rangle$ is the inner product in Hilbert space $\mathcal{H}$. The dimension of $\mathcal{H}$ is denoted as h which is potentially infinity. Then, optimization problem (1) turns into

$$L(f_w) + \lambda \|w\|_2^2 = \frac{1}{n} \sum_{i=1}^{n} \ell(\phi(x_i)^\top w, y_i) + \lambda \|w\|_2^2 \tag{2}$$

where $w \in \mathcal{H}$ stands as a linear functional in $\mathcal{H}$ and $f_w = \phi(x)^\top w$. Representer theorem yields that, given training dataset $\mathcal{D}$, the optimization problem is equivalent to find $\beta \in \mathbb{R}^n$ and $w = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \beta_i y_i \phi(x_i)$ that minimize (2). If mild assumptions such as convexity and smoothness are posed on ℓ, the optimization problem (2) is smooth and convex. Thus, there exists an unique minimizer $\overline{\beta}$ and $\overline{w} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \overline{\beta}_i y_i \phi(x_i)$. For notation simplicity, we define empirical kernel matrix $K = \big(k(x_i, x_j)\big)_{n \times n}$ and feature map matrix $\Phi = \big(\phi(x_1), \cdots, \phi(x_n)\big)_{h \times n}$.

1.3 Randomized Sketching

In this section, we consider optimization of (3). Though there exist fast optimization methods for certain ℓ such as hinge loss and logistic loss, no dual scheme to achieve time complexity lower than $O(n^3 T)$ has been proposed for general loss function. Here, T is the iteration required by second order methods which is of an ignorable double log order. Due to its smoothness and strict convexity of (3), we adopt second order methods for numerical optimization. In this case, we use sketching technique [Dri+11; YPW17] to obtain approximate solution and thus faster computation $O(n^2 m T)$.

Instead of optimizing original parameter, we consider an approximation based on limiting the β in $\mathbb{R}^n$ to an m-dimensional subspace of $\mathbb{R}^n$, where $m \ll n$ is a predetermined projection dimension. The approximation is defined via a sketch matrix $S \in \mathbb{R}^{n\times m}$ and the m-dimensional subspace is generated by the column span of S. To be more specific, we substitute the n dimensional β by $S\alpha$ and optimize m dimensional α. Together, we reformulate optimization object (1) as follows.

$$L(f_\alpha) + \lambda \|f_\alpha\|_{\mathcal{H}} = \frac{1}{n}\sum_{i=1}^{n} \ell(\frac{\phi(x_i)^\top \Phi S\alpha}{\sqrt{n}}, y_i) + \frac{\lambda}{n}\alpha^\top S^\top \Phi^\top \Phi S\alpha \tag{3}$$

where $f_\alpha(x) = \phi(x)^\top \Phi S\alpha/\sqrt{n}$.

Vast choice has been designed for the choice of S. Gaussian sketching, i.e., matrices $S \in \mathbb{R}^{m\times n}$ with i.i.d. Gaussian entries $\mathcal{N}(0, 1/m)$ are a classical random projection whose spectral and subspace embedding properties are tightly characterized. Though the cost of forming the sketch $K \cdot S$ for kernel matrix K requires $\mathcal{O}(n^2 m)$, parallel computation can significantly accelerate this process. In this paper, we only consider Gaussian sketching, while we discuss potential usage of other sketching methods in Section 4.

So far, we have presented the necessary notations and background knowledge. We now present our algorithm which follows the straightforward logic of the Newton–Raphson method.

Algorithm 1: Sketching Empirical Risk Minimization

Data: Training data $\mathcal{D} := \{(x_1, y_1), \ldots, (x_n, y_n)\}$; Query x.

Parameters: Sketching dimension m; Sketching matrix S; Regularization λ.

Initialization: $\alpha_0 \sim \mathcal{N}(0, I_m)$.

for $t = 1$ ***to*** T **do**

  Compute $\nabla\big(L(\alpha_{t-1}) + \lambda\|f_{\alpha_{t-1}}\|_{\mathcal{H}}\big)$ and $\nabla^2\big(L(\alpha_{t-1}) + \lambda\|f_{\alpha_{t-1}}\|_{\mathcal{H}}\big)$ for (3);

  Update $\alpha_t = \alpha_{t-1} - \big[\nabla^2\big(L(\alpha_{t-1}) + \lambda\|f_{\alpha_{t-1}}\|_{\mathcal{H}}\big)\big]^{-1}\nabla\big(L(\alpha_{t-1}) + \lambda\|f_{\alpha_{t-1}}\|_{\mathcal{H}}\big)$.

end

Result: The prediction $f_{\alpha_T}(x) = \phi^\top(x)\Phi S\alpha_T/\sqrt{n}$.

1.4 Related Work

Our work also lies close to the studies of DP kernel learning. By adding Laplace noise to linear functional in RKHS, [Rub+12; JT13] build DP kernel machines with utility guarantees. The utility bound of [Rub+12] measures the difference between private and non-private functionals, which is not comparable to ours. Our problem setting belongs to the non-interactive case of [JT13] while our method is more straightforward and our analysis is more detailed. [HRW13] choose the noise level added onto the output by measuring the sensitivity of the function in the RKHS norm. However, their approach takes cubic computation time.

2 Utility

In this section, we consider the utility guarantee for ERM with randomized sketching. Utility demands that the solution to the perturbed problem must achieve small excess risk. We first introduce necessary restrictions on the loss functions and kernels in Section 2.1 and present the main result in Section 2.2. An error analysis of the convergence rate is conducted in Section 2.3.

2.1 Basic Assumptions

To facilitate theoretical analysis, we adopt mild assumptions on ℓ which are satisfied by frequently used ℓ such as logistic loss and least square loss.

Assumption 1 (**Lipschitz Continuity**). *Assume that the loss function $\ell(t, y)$ is second order Lipschitz with respect to t, i.e.*

$$|\nabla\ell(t, y) - \nabla\ell(t', y)| \leq c_L|t - t'|, \text{ for } t, t' \in [-1, 1].$$

for some universal positive constant c_L.

Assumption 2 (**Strong Convexity**). *Assume that $\ell(t, y)$ is Δ-strongly convex with respect to t, meaning that*

$$\ell(t', y) \geq \ell(t, y) + \frac{\partial\ell(t, y)}{\partial t}(t' - t) + \frac{\Delta}{2}(t' - t)^2$$

for $t, t' \in [-1, 1]$ and some universal positive constant Δ.

We also require the kernel k to satisfy the following conditions.

Assumption 3 (**Bounded Kernel Function**). *For any $x_1, x_2 \in \mathcal{X}$, we have $k(x_1, x_2) \leq \kappa$ for constant $\kappa > 0$.*

Assumption 4 (**Eigenvalues of Kernel Matrix**). *Assume that the dimension of feature space h, which potentially depends on sample size n, is finite. Moreover, for all eigenvalues of $\Phi^\top\Phi$ denoted as $\lambda_1, \cdots, \lambda_h$, there holds $\lambda_h \leq \cdots \lambda_1 \leq 1$.*

Note that we require $\mathcal{H}$ to be finite-dimensional. Though all kernels do not admit this property, a great many of them can be well approximated as argued in [RR07]. For translation invariant kernels, a random h-dimensional RKHS, which is constructed by the Fourier transform of the kernel function, can uniformly approximate the original ones. Thus, we can use the approximated version of these kernels, such as the commonly used RBF kernel and Laplace kernel. In fact, the Assumption 1-4 are commonly required for DP-ERM analysis [CMS11; Rub+12; BST14; WYX17; Bas+19].

2.2 Main Result

Now we present the convergence bound of the excess risk of the estimator returned by Algorithm 1.

Theorem 5 (Utility). *Suppose Assumption 1 - 4 hold. Let f_{α_T} be the output of Algorithm 1 with $T \asymp \log\log n$. Φ is the feature map matrix of i.i.d. sample $\mathcal{D} \sim P_{X,Y}$. For some constant $C \in (0,1)$, if we choose $(C \vee 2/3)h \leq m \leq h$ and $\lambda \asymp \sqrt{\sum_{i=\lfloor Ch \rfloor+1}^{h} \lambda_i}$, then with probability $1 - 2e^{-2m} - 2/n^2$, we have*

$$\mathbb{E}[\ell(f_{\alpha_T}(x), y)] - \ell^* \leq \frac{1}{\sqrt{n}} + \sqrt{\sum_{i=\lfloor Ch \rfloor+1}^{h} \lambda_i}$$

where λ_i is the i-th singular value of Φ for $i = 1, \cdots, h$.

The theorem states that, if we properly choose an RKHS dimension h and a sketching dimension m, the excess risk of the estimation f_{α_T} can be controlled by the tail eigenvalues $\lambda_{\lfloor Ch \rfloor+1}, \cdots, \lambda_h$. As illustrated in [Bra05], these quantities will converge to zero as h grows to infinity with n. The convergence rate is decided by the decay speed of the kernel's tail eigenvalues. For instance, the polynomial kernel with order h has eigenvalues that decay quickly. Thus, its excess risk can achieve convergence rate $\mathcal{O}(e^{-h})$. In contrast, for kernels with polynomial decay, such as the Sobolev kernel, each of its eigenvalues λ_i is of order $\mathcal{O}(j^{-p})$ with $p = 2$. Thus, the upper bound of excess risk will be of order $\mathcal{O}(h^{-1/2})$, which is relatively slow. For the Cauchy kernel, which also admits polynomial decay but with $p = 1$, the upper bound even fails to converge.

2.3 Error Analysis

In this section, we conduct our error analysis for Theorem 5 by decomposing the excess risk into several parts that are intuitively associated with approximation error and optimization error. We first define two instrumental quantities used in the decomposition. Let $\widehat{\alpha}$ be the minimizer of (3), i.e.

$$L(\widehat{f}) + \lambda \|\widehat{f}\|_{\mathcal{H}} = \min_{\alpha} L(f_\alpha) + \lambda \|f_\alpha\|_{\mathcal{H}}. \tag{4}$$

Also, let $\overline{f}$ be the optimal solution to the original problem (1), i.e.

$$L(\overline{f}) + \lambda \|\overline{f}\|_{\mathcal{H}} = \min_{f \in \mathcal{H}} L(f) + \lambda \|f\|_{\mathcal{H}}. \tag{5}$$

We rely on the following decomposition.

$$\begin{aligned}
\mathbb{E}[\ell(f_{\alpha_T}(x), y)] - \ell^* = & \underbrace{\mathbb{E}[\ell(f_{\alpha_T}(x), y)] - L(f_{\alpha_T})}_{\mathbf{A}} + \underbrace{L(f_{\alpha_T}) + \lambda \|f_{\alpha_T}\|_{\mathcal{H}} - L(\widehat{f}) - \lambda \|\widehat{f}\|_{\mathcal{H}}}_{\mathbf{B}} \\
& + \underbrace{L(\widehat{f}) + \lambda \|\widehat{f}\|_{\mathcal{H}} - L(\overline{f})}_{\mathbf{C}} + \underbrace{L(\overline{f}) + \lambda \|\overline{f}\|_{\mathcal{H}} - L(f^*) - \lambda \|f^*\|_{\mathcal{H}}}_{\leq 0} \\
& + \underbrace{L(f^*) - \ell^*}_{\mathbf{D}} + \lambda \|f^*\|_{\mathcal{H}} - \lambda \|f_{\alpha_T}\|_{\mathcal{H}} - \lambda \|\overline{f}\|_{\mathcal{H}}
\end{aligned} \tag{6}$$

Besides the remaining terms, each part of the decomposition has its interpretation. The term **A** and **D** represent the discrepancy between population level loss and empirical level loss. They arise because we turned the problem of bounding the population excess risk into bounding their empirical counterparts. The existence of term **B** is due to the fact that we can not optimize problem (3) exactly. Thus, we need to control the difference between the output of Algorithm 1 and the global minimizer (4). It is referred to as the optimization error. The term **C** is called the approximation error. It depicts the difference between the solution of the original problem, i.e. (1) and the sketched problem, i.e. (3). The next term is controlled naturally via the definition of $\overline{f}$ in 5. In what follows, we control these terms separately.

2.3.1 Control Term A and D

Lemma 6. *Suppose Assumption 1 - 3 hold. Let $\hat{f}$ be defined in (4). Then we have*

$$\mathbb{E}[\ell(f_{\alpha_T}(x), y)] - L(f_{\alpha_T}) \lesssim \sqrt{\frac{\log n}{n}}$$

with probability $1 - 1/n^2$.

Lemma 7. *Suppose Assumption 1 - 3 hold. Let $f^* \in \mathcal{H}$ be the minimizer for ℓ^*. Then we have*

$$\mathbb{E}[\ell(f^*(x), y)] - L(f^*) \lesssim \sqrt{\frac{\log n}{n}}$$

with probability $1 - 1/n^2$.

2.3.2 Control Term B

Lemma 8. *Suppose Assumption 1 - 3 hold. Let $f_{\alpha_T}(x)$ be the output of Algorithm 1. Let $\hat{f}$ be defined in (4). If we choose $T \asymp \log\log n$, then we have*

$$L(f_{\alpha_T}) + \lambda \|f_{\alpha_T}\|_{\mathcal{H}} - L(\widehat{f}) - \lambda \|\widehat{f}\|_{\mathcal{H}} \lesssim \frac{1}{n}.$$

2.3.3 Control Term C

Lemma 9. *Suppose Assumption 1 - 3 hold. Let $\hat{f}$ and $\overline{f}$ be defined in (4) and (5), respectively. For some constant $C < 1$, let $(C \vee 2/3)h \leq m \leq h$. Then we have*

$$L(\widehat{f}) - L(\overline{f}) \lesssim \lambda + \sqrt{\sum_{i=\lfloor Ch \rfloor + 1}^{h} \lambda_i}$$

with probability at least $1 - 2e^{-2m}$.

2.4 Comments

2.4.1 Comments on the Convergence Rate

The convergence rate are dominated by two terms which are $\sqrt{\sum_{i=\lfloor Ch\rfloor+1}^{h}\lambda_i}$ and $\frac{1}{\sqrt{n}}$. The former term comes from approximation error. For a Gaussian sketching matrix S, $S^\top S$ can approximate I_h, i.e. $\lfloor Ch\rfloor$ of its eigenvalues are close to 1 with high probability. As a result, the solution to (1) and (3) holds identical on a subspace with dimension $\lfloor Ch\rfloor$. The difference on the remaining $h-\lfloor Ch\rfloor$ is controlled by the smallest eigenvalues of Φ and thus forms $\sqrt{\sum_{i=\lfloor Ch\rfloor+1}^{h}\lambda_i}$. Note that we can not further improve this result for general ℓ. For least square loss, whose first-order term of Taylor expansion with respect to $f(x)$ vanishes, this rate can be improved to $\sum_{i=\lfloor Ch\rfloor+1}^{h}\lambda_i$, as is derived in [YPW17].

The term $\frac{1}{\sqrt{n}}$ comes from Lemma 6 and 7, where we used theoretical tools from empirical process [VVW96; Wai19]. To be specific, a Talagrand-type inequality [Tal94] is established such that the difference between the population loss and empirical loss is controlled by the complexity of the potential function space $\mathcal{H}$. In fact, this rate can also be improved by using advanced theoretical tools such as local Rademacher complexity [BBM05]. As illustrated in [BBM05; YPW17], the optimal rate is the "critical radius" of kernel k, which is generally smaller than $\frac{1}{\sqrt{n}}$. Besides maintaining the clarity of our proof, the reason that we do not adopt these techniques is that $\frac{1}{\sqrt{n}}$ is usually dominated by $\sqrt{\sum_{i=\lfloor Ch\rfloor+1}^{h}\lambda_i}$. Thus, it is unnecessary to make this rate faster.

Note that in [YPW17], the above-mentioned rates meet. Our results do not achieve a balanced trade-off between different error terms. We are capable of achieving the best trade-off by choosing large m and h. However, doing so yields a privacy guarantee with order $\mathcal{O}(1)$, as illustrated in the next section. To provide a privacy guarantee that is able to converge to zero, we must sacrifice the utility property.

3 Privacy

In this section, we provide a detailed analysis of the privacy guarantee of our algorithm. We first provide the main results in Section 3.1 and some comments in Section **??**. In Section 3.2, we intuitively illustrate how our mechanism brings differential privacy. In Section 3.3 and 3.4, we introduce the Grassman manifold and its properties which will be the mathematical foundation of our analysis for privacy. The road map of its proof is presented in Section 3.5.

3.1 Main Results

To be self-contained, we give a definition of approximate differential privacy as follows. We say two data sets $\mathcal{D}$ and $\mathcal{D}'$ are neighboring data sets if they differ in one entry (that is, $|\mathcal{D}\triangle\mathcal{D}'|=2$).

Definition 10 (Approximate Differential Privacy [Dwo+06a; Dwo+06b]). *An algorithm $\mathcal{A}:\mathcal{X}^n\to\mathcal{Y}$ is (ε,δ)-differentially private (i.e., it satisfies approximate differential privacy) if, for all*

neighboring databases $X, X' \in \mathcal{X}^n$, *and all* $T \subseteq \mathcal{Y}$

$$\mathbb{P}[\mathcal{A}(X) \in T] \leq e^{\varepsilon}\mathbb{P}\left[\mathcal{A}\left(X'\right) \in T\right] + \delta$$

To facilitate theoretical analysis, we adopt mild assumptions on the optimization object. Note that for general loss function ℓ such as logistic loss and least square loss, the assumption is naturally satisfied.

Assumption 11 (**Gradient Convexity**). *We assume that* $\|\nabla_w J(f_w, \mathcal{D})\|_2^2$ *is a* $\tilde{\Delta}$*- strongly convex function with respect to* w*, meaning that the smallest eigenvalue of its Hessian matrix is larger than the positive constant* $\tilde{\Delta}$*.*

With the additional assumption, we now present our main result for privacy guarantee.

Theorem 12 (**Privacy**). *Suppose Assumption 1- 4 and 11 hold. Given a dataset* $\mathcal{D}$ *and query* x*, the output of Algorithm 1, denoted as* $f_{\alpha_T}(x)$ *is* (ϵ, δ) *differentially private, i.e. for any measurable set* $\mathcal{T} \in \mathbb{R}$*, we have*

$$\mathbb{P}[f_{\alpha_T}(x) \in \mathcal{T}|\mathcal{D}] \leq e^{\epsilon}\mathbb{P}[f_{\alpha_T}(x) \in \mathcal{T}|\mathcal{D}'] + \delta$$

with ϵ *and* δ *specified as*

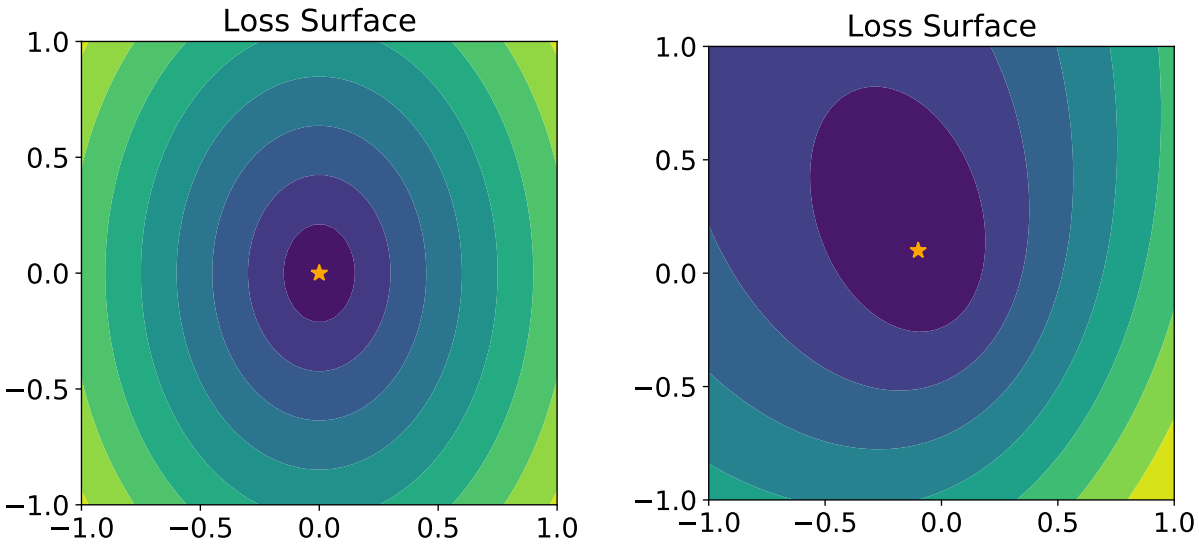
$$\epsilon = (64c_L\sqrt{\kappa})C_h^{-1}(h-m)mn^{-1/2+\nu} + (32c_L\sqrt{\kappa} + 2\sqrt{\kappa})C_h^{-2}n^{-1/2+2\nu}$$

$$\delta = V_d 4^d C^{-d} m^{h(m-1)/2}(2\pi)^{-h/2} n^{\nu(m-h)m/2-d/2}$$

3.2 Perturbation Mechanism

As mentioned in the first section, there are three types of privacy mechanisms, that are gradient perturbation, output perturbation, and objective perturbation. Our proposed mechanism is closer to objective perturbation. Both mechanisms distort the non-private optimization object by a randomly generated matrix (vector) in order to shift the solution around the original solution. The perturbation matrix (vector) is carefully designed. On one hand, the distortion should be mild such that the utility of the shifted solution is kept. On the other hand, regarding the shifted solution as a random function of the perturbation matrix, the distribution of the solution should be adequately spread out to provide a privacy guarantee. In [CM08], the optimization object $\frac{1}{n}\sum_{i=1}^n \ell(\phi(x_i)^\top w, y_i) + \lambda\|w\|_2^2$ is added with term $w \cdot b$ where b is a vector consisting of independent Laplace random variables. In other words, the gradient of the object for all w is added with a fixed vector b. An illustrative example is provided to show the privacy mechanism of classical objective perturbation in [CMS11].

As shown in Figure 1, a subtle shift occurs to the loss surface due to the perturbation on the optimization target, which brings a slight move to the optimal solution. To ensure both privacy and utility, the variance of Laplace perturbation is carefully designed such that the perturbed optimal



(a) Loss surface for the original problem.

(b) Loss surface for the perturbed problem.

Figure 1: The comparison of contour plots for classical objective perturbation. The yellow star is the minimum of the corresponding surface.

solutions will not be too far from the original solution while the migration is enough to confuse outside attackers.

The mechanism in Algorithm 1 follows an analogous logic. However, our mechanism is intrinsically different from theirs. The object in (2) search for solution $\hat{w}$ in $\{\Phi\beta|\beta \in \mathbb{R}^n\}$ with dimension n. However, (3) search for solution in the space $\{\Phi S\alpha|\alpha \in \mathbb{R}^m\}$. In other words, the set of feasible solutions to (3) is a subset of those to (2). Thus, the solution to (3) is shifted by the randomly generated subspace associated with S and its privacy is preserved. The intuition for our mechanism is illustrated in Figure 2. In Figure 2, two pink lines are induced by different random matrices S. Thus, the corresponding constraint optimal solutions, which are marked with yellow stars, are different.

The distribution of subspace, as well as the optimal solution in it, are decided completely by the random Gaussian matrix S. Therefore, the main difficulty in showing privacy is computing the distribution of subspaces.

3.3 Random Projected Subspace

In this section, we introduce theoretical tools that arise from differential geometry. The topological space of all linear p-dimensional subspaces in $\mathbb{R}^q$ for $p \leq q$, denoted as $\mathrm{Gr}(p, q)$, is known as *Grassmannian* or *Grassman Manifold* [MS; KRB97]. The Grassman Manifold is naturally endowed with a Haar measure [Haa33] which is translation invariant under unitary transformation in $\mathbb{R}^q$. [CC03] developed a thorough analysis for distributions on Grassmannian under different sampling schemes.

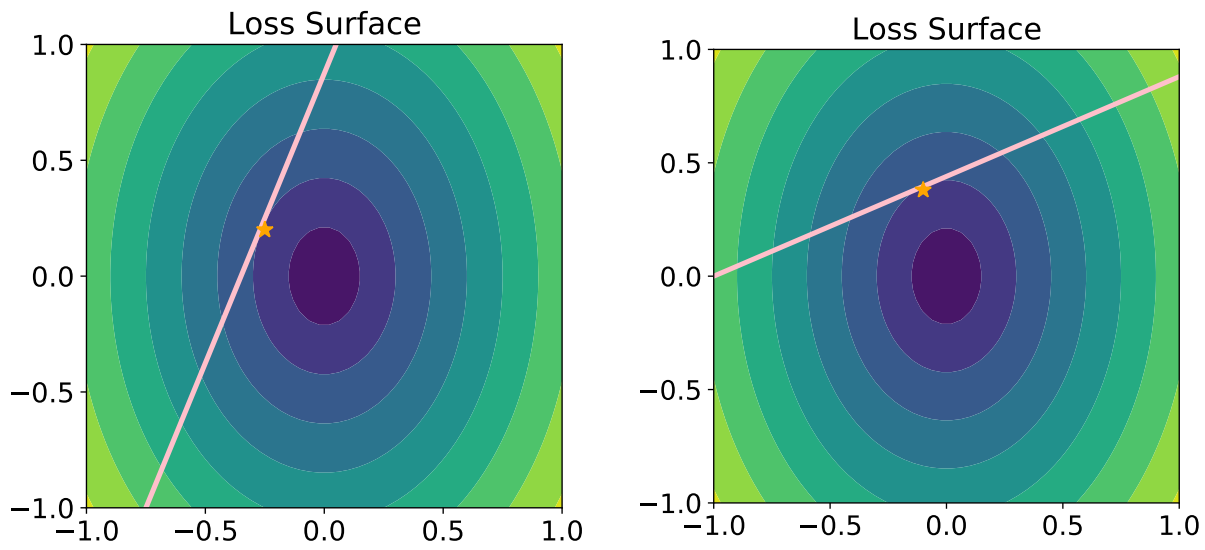


Figure 2: Comparison of the subspaces of feasible solutions. On the contour plot of the loss surface, the pink line represents a one-dimensional subspace of feasible solutions to problem (3). Subspaces on the left and right are induced by different S. The yellow star stands for the optimal solution in the corresponding subspace.

In our case, the subspace in which α is optimized within is the span of columns of ΦS. The distribution of ΦS is called Rectangular Matrix-Variate Distribution [CC03, Section 2.3], the matrix version of Gaussian distribution. Grassmannian serves as a perfect tool for quantifying the distribution of subspaces. However, as illustrated in Figure 2, we would like to deal with affine spaces, i.e. subspaces with intercept. Thus, we need to generalize Grassmannian to affine subspaces.

The topological space of all affine p-dimensional subspaces in $\mathbb{R}^q$, denoted as $\mathrm{Graff}(p, q)$, is called *Affine Grassmannian* [KRB97; LWY21]. Affine Grassmannian naturally inherits most merits of Grassimannian including its Haar measure [LWY21]. Let $g^{p,q}$ denote an element in $\mathrm{Graff}(p, q)$. Let $\gamma^{p,q}$ be the Haar measure on $\mathrm{Graff}(p, q)$ which is unique up to a constant. Without loss of generality, let $\int_{\mathrm{Graff}(p,q)} d\gamma^{p,q} = 1$.

It is desirable to uniquely represent elements of $\mathrm{Graff}(p, q)$ as actual matrices instead of an abstract set of affine spaces. Especially, an explicit representation in Euclidean space is a prerequisite for discussing probability distributions on $\mathrm{Graff}(p, q)$. We follow the well-known representation results [Nic20; LWY21] to write the topology spaces as the set of rank-p orthogonal projection matrices and a vector orthogonal to the column space of the matrix:

$$\mathrm{Graff}(p, q) \cong \left\{ [P, b] \in \mathbb{R}^{q\times(q+1)} : P^\top = P^2 = P, \mathrm{tr}(P) = p, Pb = 0 \right\}.$$

Though named as orthogonal, P is not an orthogonal matrix except for $P = I_n$. Note that $\mathrm{rank}(P) = \mathrm{tr}(P)$ for an orthogonal projection matrix P. With this definition, we use $[P, b]$, which is called

projection affine coordinates, to represent points on $\text{Graff}(p,q)$. Let $g_{P,b}^{p,q}$ denote the element in $\text{Graff}(p,q)$ corresponding to $[P,b]$.

To make the space we are optimizing α an affine space, we need an intercept. Thus, we define an augmented problem. To be specific, let S be a $n \times (m+1)$ Gaussian sketching matrix and let $\alpha \in \mathbb{R}^m$ have one fixed entry. Without loss of generality, we fix $\alpha_{m+1} = 1$. In other words, we solve the optimization problem

$$\min_{\alpha \in \mathbb{R}^{m+1}, \alpha_{m+1}=1} L(f_\alpha) + \lambda \|f_\alpha\|_{\mathcal{H}}.$$

For notation simplicity, we denote the optimization object for any function f and training dataset $\mathcal{D}$ as $J(f, \mathcal{D}) = L(f) + \lambda \|f\|_{\mathcal{H}}$. Thus, the above optimization problem can be rewritten as

$$\min_{\alpha \in \mathbb{R}^{m+1}, \alpha_{m+1}=1} J(f_\alpha, \mathcal{D}). \tag{7}$$

Note that setting $\alpha_{m+1} = 1$ does not affect any conclusion about f_α derived before since the approximation power of kernel functions is not related to the intercept term. See [SC08] for discussions of intercept in kernel machines. With these preparations, the following result depicts the probability distribution of the subspaces $g_{P,b}^{m,h}$.

Lemma 13. *Given dataset $\mathcal{D}$, let S be a Gaussian sketching matrix. Then, the density of element $g_{P,b}^{m,h}$ is given by*

$$p(g_{P,b}^{m,h}) = \frac{m^{hm/2} \prod_{\lambda_i > 0} \lambda_i(\Pi)}{(2\pi)^{h/2} \left|\Sigma - (\Sigma - I_h)P\right|^{m/2}} \cdot \exp\big(- b^\top \Pi \, b \big).$$

Here, Σ is the $h \times h$ diagonal matrix with singular value of $\frac{1}{\sqrt{n}}\Phi$ ordered decreasingly on its diagonal. Π is the pseudo inverse of $(I_h - P)\Sigma(I_h - P)$.

3.4 Bijection Between Subspaces and Optimization Solution

One may notice that there is no bijection $g_{P,b}^{m,h}$ and w since they have different cardinalities. In fact, the cardinality of $g_{P,b}^{m,h}$ can be explicitly calculated by $\text{Graff}(m,h) \sim \mathbb{R}^{(h-m)\cdot(m+2)}$ ([LWY21]). An illustration is provided in Figure 3. The figure considers optimization in one-dimensional subspaces in $\mathbb{R}^3$. The colorful surface parameterized by $z = x^2 + y^2$ represents an equipotential surface. Both black lines pass $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ and are perpendicular to the normal vector of the surface at this point. Due to the convexity of the surface, both black lines induce the same optimal solution $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ but they associate to different $g_{P,b}^{1,3}$.

Thus, we need to identify the quotient space of $\text{Graff}(m,h)$ by $\mathbb{R}^h$, i.e. a collection of $g_{P,b}^{m,h}$ that induce the same w. We formally define this class of $g_{P,b}^{m,h}$ as follows.

Definition 14 (w-Optimal). *For $g_{P,b}^{m,h} \in \text{Graff}(m,h)$ and dataset $\mathcal{D}$, we say it is w-optimal if*

$$J(f_w, \mathcal{D}) = \min_{\tilde{w} \in g_{P,b}^{m,h}} J(f_{\tilde{w}}, \mathcal{D}).$$

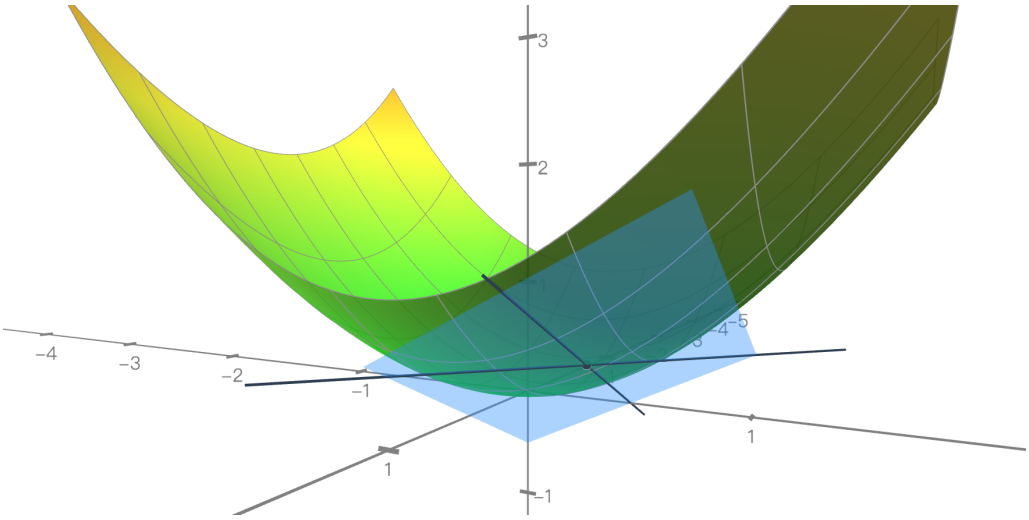


Figure 3: Illustration of different affine spaces that induce the same solution.

We depict the $g_{P,b}^{m,h}$ that potentially induces w by the following property.

Definition 15 (**w-Inducible**). *For $g_{P,b}^{m,h} \in \mathrm{Graff}(m,h)$ and dataset $\mathcal{D}$, we say it is w-inducible if*

$$P\nabla_w J(f_w, \mathcal{D}) = \mathbf{0} \quad \textit{and} \quad (I_h - P)w = b.$$

The first condition in Definition 14 requires that the direction of $g_{P,b}^{m,h}$ belongs to the tangent space of the loss surface at w. The second condition identifies the intercept term. The following lemma reveals that these definitions are actually equivalent.

Lemma 16 (**Equivalence of w-Optimal and w-Inducible**). *Suppose Assumption 1 - 3 hold. Then, for each $w \in \mathbb{R}^h$, a subspace $g_{P,b}^{m,h}$ is w-Optimal if and only if it is w-Inducible.*

As a result, we denote collection of all w-optimal $g_{P,b}^{m,h}$, equivalently all w-inducible $g_{P,b}^{m,h}$, as $G_w(\mathcal{D})$ for each $w \in \mathbb{R}^h$. This can be written rigorously as

$$G_w(\mathcal{D}) := \left\{ g_{P,b}^{m,h} \mid J(f_w, \mathcal{D}) = \min_{\tilde{w} \in g_{P,b}^{m,h}} J(f_{\tilde{w}}, \mathcal{D}) \right\}. \tag{8}$$

Lemma 16 implies that optimizing within the affine subspace $g_{P,b}^{m,h} \in G_w(\mathcal{D})$ leads to optimal solution w. Meanwhile, for each w, all m dimensional affine subspaces that lead to the final output of w belong to $G_w(\mathcal{D})$.

The definition of $G_w(\mathcal{D})$ is beneficial from two perspectives. On one hand, Definition 14 explains $G_w(\mathcal{D})$ such that it is the collection of subspaces that induce optimal solution w. This definition is conceptually straightforward. On the other hand, Definition 15 provides explicit conditions that we can verify for each pair of w and $g_{P,b}^{m,h}$. This definition is computationally convenient. Lemma 16 unifies these two properties and facilitates the identification of the $g_{P,b}^{m,h}$s that induce w. Intuitively, the density of w is the integral of the density of $g_{P,b}^{m,h}$ contained in $G_w(\mathcal{D})$. Thus, this identification is an important procedure for evaluating $p(w|\mathcal{D})$.

3.5 Error Analysis

In this section, we provide useful lemmas for proof of Theorem 12. We first construct a region Δ_n belonging to $\mathcal{H}$ and show that the probability of the solution of (7) falling outside of Δ_n is low, as will be shown in Lemma 17. For any $w \in \Delta_n$, it is privacy-preserving as will be argued in Lemma 18.

Without loss of generality, it suffices to consider $\mathcal{X}_w = \{w \| \|w\|_2 \leq 1\}$. The region Δ_n is specified as the collection of w with a large gradient, namely

$$\Delta_n := \left\{ w \;\middle|\; \|\nabla_w J(f_w, \mathcal{D})\|_2 \geq \frac{1}{\sqrt{n}} \right\} \cap \mathcal{X}_w \tag{9}$$

and consequently $\Delta_n^c = \mathcal{X}_w / \Delta_n$.

Lemma 17. *Suppose Assumption 1- 4 and 11 hold. Let Δ_n be defined in (9). Then for dataset $\mathcal{D}$, there holds*

$$\mathbb{P}[w \in \Delta_n^c | \mathcal{D}] \leq \delta$$

where δ is specified as $\delta = V_d 4^d C^{-d} m^{h(m-1)/2} (2\pi)^{-h/2} n^{\nu(m-h)m/2 - d/2}$.

Lemma 18. *Suppose Assumption 1- 4 hold. Let Δ_n be defined in (9). Let $\mathcal{D}$ and $\mathcal{D}'$ be neighboring datasets. Then for any $\mathcal{T}$ which is a subset of Δ_n, we have*

$$\mathbb{P}[w \in \mathcal{T} | \mathcal{D}] \leq e^{\epsilon} \mathbb{P}[w \in \mathcal{T} | \mathcal{D}']$$

where ϵ is specified as $\epsilon = (64 c_L \sqrt{\kappa}) \lambda_h^{-1} (h-m) m n^{-1/2} + (32 c_L \sqrt{\kappa} + 2\sqrt{\kappa}) \lambda_h^{-2} n^{-1/2}$.

4 Discussion

We typically consider Gaussian sketching, while many other sketching methods, such as SRHT and count sketch, also have strong subspace embedding properties and are faster to compute. As a special type of sketching, sub-sampling methods have favorable computation and space complexity while they may perform poorly under certain assumptions. Both sub-sampling and Gaussian sketching are shown to induce DP. Moreover, [Bas+17] showed that the noisy count sketch can be used for private frequency estimation. Thus, a prospective future work would be investigating DP induced by SRHT and count sketch.

5 Proofs

5.1 Utility Proofs Related to Section 2

5.1.1 Technical Lemmas

In this section, we provide technical lemmas that are necessary for the subsequent analysis.

Lemma 19. *Suppose Assumption 1- Assumption 3 hold. the SVD decomposition of matrix $\Phi/\sqrt{n}$ is $Q_1\Sigma^{1/2}Q_2$, where $\Sigma^{1/2}$ is a $h \times h$ matrix with eigenvalues of Φ decreasingly sorted on its diagonal, namely $\lambda_1^{1/2}, \lambda_2^{1/2}, \cdots, \lambda_h^{1/2}$. Q_1 and Q_2 are $h \times h$ and $n \times n$ orthogonal matrix, respectively. For some constant C which will be specified in proof, let $\Phi_A = \left(Q_1^\top \Phi\right)_{1:\lfloor Ch \rfloor,:}$ be the first $\lfloor Ch \rfloor$ rows of $Q_1^\top \Phi$ and $\Phi_B = \left(Q_1^\top \Phi\right)_{(\lfloor Ch \rfloor+1):h,:}$ be the rest of the rows of $Q_1^\top \Phi$. Then if we take $(Ch \vee 2h/3) \leq m \leq h$, there holds*

$$\|(\Phi_A\Phi_A^\top)^{-1/2}\Phi_A SS^\top \Phi_A^\top (\Phi_A\Phi_A^\top)^{-1/2} - I_{\lfloor Ch \rfloor}\|_2 \leq 1/2, \quad \|\Phi_B S\|_2 \leq \lambda_{\lfloor Ch \rfloor+1}^{1/2}$$

with probability $1 - 2e^{2m}$.

To conduct our analysis, we need to recall the definitions of Rademacher complexity which is broadly used in theoretical machine learning [VVW96; Wai19].

Definition 20 (Rademacher Complexity). *Given function class $\mathcal{F}$, the Rademacher complexity of $\mathcal{F}$ is defined as*

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right| \mid x_1, \cdots, x_n \right]$$

where $\{\epsilon_i\}_{i=1}^n$ are i.i.d. Rademacher variables.

Rademacher complexity measures the richness of a class of real-valued functions with respect to a probability distribution. Given a set of data, it picks $f \in \mathcal{F}$ to maximize the product for each group of $\epsilon_i, i = 1, \cdots, n$. When the function class is rich enough, it contains functions that can appropriately adapt for each group of Rademacher random variables. When the function class is rather restricted, it can not shatter the data and thus have small $\mathcal{R}_n(\mathcal{F})$. In our case, our goal is to bound the complexity of the kernel class, which is formalized as follows.

Lemma 21 (Rademacher Complexity of Kernel Class). *Suppose Assumption 3 holds. Let $\mathcal{F}$ be the function class of kernel, i.e. $\mathcal{F} = \{f(x) | f(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \beta_i k(x_i, x)\}$. Then*

$$\mathcal{R}_n(\mathcal{F}) \leq \frac{\sqrt{\kappa} \sup_{f \in \mathcal{F}} \|f\|_{\mathcal{F}}}{\sqrt{n}}.$$

Lemma 22 (Rademacher Complexity of Sketched Kernel Class). *Suppose Assumption 3 holds. Let $\tilde{\mathcal{F}}$ be the sketched kernel class, i.e. $\tilde{\mathcal{F}} = \{f(x) | f(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (S\alpha)_i k(x_i, x)\}$ with Gaussian sketching matrix S. Then*

$$\mathcal{R}_n(\tilde{\mathcal{F}}) \leq \frac{\kappa \sup_{f \in \mathcal{F}} \|f\|_{\tilde{\mathcal{F}}}}{\sqrt{n}}.$$

Lemma 23 ([LT91])**.** *Let g be Lipschitz, namely $|g(x) - g(y)| \leq L|x - y|$. Then, for every class $\mathcal{F}$,*

$$\mathcal{R}_n g \circ \mathcal{F} \leq L\mathcal{R}_n \mathcal{F},$$

where $g \circ \mathcal{F} := \{g \circ f : f \in \mathcal{F}\}$.

Lemma 24 (**Error Bound by Rademacher Complexity**). *Let $\mathcal{F}$ be a class of functions that map $\mathcal{X}$ into $[a,b]$. Assume that there is some $r>0$ such that for every $f\in\mathcal{F}, \mathbb{E}\left[f(x)^2\right]\leq r$. Then, with probability at least $1-1/n^2$,*

$$\sup_{f\in\mathcal{F}}\left(\mathbb{E}[f(x)]-\sum_{i=1}^{n}f(x_i)\right)\leq 4\mathcal{R}_n\mathcal{F}+\sqrt{\frac{2r\log n}{n}}+2(b-a)\frac{\log n}{n}.$$

The opposite direction is also true, i.e.

$$\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{n}f(x_i)-\mathbb{E}[f(x)]\right)\leq 4\mathcal{R}_n\mathcal{F}+\sqrt{\frac{2r\log n}{n}}+2(b-a)\frac{\log n}{n}.$$

5.1.2 Proofs Related to Section 2.3

Without loss of generality, we assume that

$$\|f^*\|_{\mathcal{H}}=\frac{\beta^{*\top}\Phi^\top\Phi\beta^*}{\sqrt{n}}\leq 1.$$

Since we have

$$|f^*(x)|\leq\frac{|\phi^\top(x)\Phi\overline{\beta}|}{\sqrt{n}}\leq\frac{\|\phi(x)\|_2\|\Phi\beta^*\|_2}{\sqrt{n}}\leq\kappa,$$

it is fair to assume that there exists a constant B such that $|\overline{f}(x)|\leq B$ and $|\widehat{f}(x)|\leq B$. This means if the ground truth is bounded, any fitting that is bounded by B is enough. In the subsequent proof, we define several $\alpha^\dagger$ and $f^\dagger(x)=\phi^\dagger(x)\Phi S\alpha^\dagger/\sqrt{n}$ for temporary technical usage. The meaning of each pair of $(\alpha^\dagger,f^\dagger)$ may vary between proofs.

Proof of Lemma 6. Since $\|f\|_\infty\leq B$, we have $\mathbb{E}\left[\ell(f(x),y)^2\right]\leq c_L^2B^2$ by Assumption 1. Then, applying the first argument of Lemma 24 to the function class $\mathcal{F}_\ell=\{\ell(f(x),y)|f(x)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\beta_ik(x_i,x)\}$ yields

$$L(f^*)-\ell^*\lesssim\mathcal{R}_n(\mathcal{F}_\ell)+\sqrt{\frac{\log n}{n}}.$$

Lemma 22 and 23 together leads to

$$\mathcal{R}_n(\mathcal{F}_\ell)\leq c_L\mathcal{R}_n(\mathcal{F})\lesssim\frac{1}{\sqrt{n}}$$

which completes the proof. Here we used Assumption 1.

□

Proof of Lemma 7. Since $\|f\|_\infty\leq B$, we have $\mathbb{E}\left[\ell(f(x),y)^2\right]\leq c_L^2B^2$ by Assumption 1. Then, applying the second argument of Lemma 24 to the function class $\mathcal{F}_\ell=\{\ell(f(x),y)|f(x)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\beta_ik(x_i,x)\}$ yields

$$L(f^*)-\ell^*\lesssim\mathcal{R}_n(\mathcal{F}_\ell)+\sqrt{\frac{\log n}{n}}.$$

Lemma 21 and 23 together leads to

$$\mathcal{R}_n(\mathcal{F}_\ell) \le c_L \mathcal{R}_n(\mathcal{F}) \lesssim \frac{1}{\sqrt{n}}$$

which completes the proof. Here we used Assumption 1. □

Proof of Lemma 8. Since the optimization object (3) is strongly convex, the classical analysis of Newton's method yields that, with $\mathcal{O}(\log\log \epsilon^{-1})$ iterations, there holds

$$L(f_{\alpha_T}) + \lambda \|f_{\alpha_T}\|_{\mathcal{H}} - L(\widehat{f}) - \lambda \|\widehat{f}\|_{\mathcal{H}} \le \epsilon.$$

Taking $\epsilon = \frac{1}{n}$ brings the desired result. □

Proof of Lemma 9. We first decompose ϕ into two parts: one is associated with the principal eigenvalues and the other is associated with the remainder. To be specific, the SVD decomposition of matrix $\Phi/\sqrt{n}$ is $Q_1 \Sigma^{1/2} Q_2$, where $\Sigma^{1/2}$ is a $h \times h$ matrix with eigenvalues of Φ decreasingly sorted on its diagonal, namely $\lambda_1^{1/2}, \lambda_2^{1/2}, \cdots, \lambda_h^{1/2}$. Q_1 and Q_2 are $h \times h$ and $n \times n$ orthogonal matrix, respectively. Then, we define $\Phi_A = \left(Q_1^\top \Phi\right)_{1:\lfloor Ch \rfloor,:}$ be the first $\lfloor Ch \rfloor$ rows of $Q_1^\top \Phi$ and $\Phi_B = \left(Q_1^\top \Phi\right)_{(\lfloor Ch \rfloor + 1):h,:}$ be the rest of the rows of $Q_1^\top \Phi$. ϕ_A and ϕ_B are defined analogously by $\left(Q_1^\top \phi\right)_{1:\lfloor Ch \rfloor}$ and $\left(Q_1^\top \phi\right)_{(\lfloor Ch \rfloor + 1):h}$.

There exists a unique $\overline{\beta}$ such that $\overline{f}(x) = \phi^\top(x)\Phi\overline{\beta}/\sqrt{n}$ is the minimizer of (1). Define

$$\alpha^\dagger = S^\top \Phi_A^\top (\Phi_A S S^\top \Phi_A^\top)^{-1} \Phi_A \overline{\beta}$$

where we are almost surely guaranteed with invertibility of $\Phi_A S S^\top \Phi_A^\top$ if $m > \lfloor Ch \rfloor$. Let $f^\dagger(x) = \phi^\top(x)\Phi S \alpha^\dagger/\sqrt{n}$. Note that the RKHS norm of $f^\dagger$ has

$$\|f^\dagger\|_{\mathcal{H}} = \alpha^{\dagger T} S^\top \Phi^\top \Phi S \alpha^\dagger = \frac{\beta^{*\top} \Phi_A^\top \Phi_A \overline{\beta}}{n} \le \frac{\|\Phi\overline{\beta}\|_2^2}{n} \le 1.$$

Then we have

$$\begin{aligned}
|f^\dagger(x) - f^*(x)| =& \frac{1}{\sqrt{n}} |\phi_A^\top(x)\Phi_A S\alpha^\dagger + \phi_B^\top(x)\Phi_B S\alpha^\dagger - \phi_A^\top(x)\Phi_A\overline{\beta} - \phi_B^\top(x)\Phi_B\overline{\beta}| \\
\le& \frac{1}{\sqrt{n}} |\phi_B^\top(x)\Phi_B S\alpha^\dagger| + \frac{1}{\sqrt{n}} |\phi_B^\top(x)\Phi_B\overline{\beta}|
\end{aligned}$$

We control the two terms separately. The latter term is controlled by Cauchy inequality as

$$\frac{1}{\sqrt{n}} \phi_B^\top(x)\Phi_B\overline{\beta} \le \frac{1}{\sqrt{n}} \|\phi_B(x)\|_2 \|\Phi_B\overline{\beta}\|_2 \le \frac{1}{\sqrt{n}} \|\phi_B(x)\|_2 \|\Phi\overline{\beta}\|_2 \le \|\phi_B(x)\|_2.$$

The former term is controlled as follows.

$$\begin{aligned}
|\phi_B^\top(x)\Phi_B S\alpha^\dagger| &= \left|\phi_B(x)\Phi_B S S^\top \Phi_A^\top (\Phi_A S S^\top \Phi_A^\top)^{-1} \Phi_A \overline{\beta}\right| \\
&\le \|\phi_B(x)\|_2 \|\Phi_B S\|_2 \|S^\top \Phi_A^\top (\Phi_A \Phi_A^\top)^{-1/2}\|_2 \\
&\quad \cdot \left\|\left[(\Phi_A\Phi_A^\top)^{-1/2} \Phi_A S S^\top \Phi_A^\top (\Phi_A\Phi_A^\top)^{-1/2}\right]^{-1}\right\|_2 \|(\Phi_A^\top\Phi_A)^{-1}\|_2 \|\Phi_A\overline{\beta}\|_2 \\
&\le \|\phi_B(x)\|_2 \cdot \lambda_{\lfloor Ch \rfloor}^{1/2} \cdot \frac{3}{2} \cdot 2 \cdot \lambda_{\lfloor Ch \rfloor}^{-1/2} \cdot \sqrt{n}
\end{aligned}$$

where in the last inequality we used Lemma 19. Note that the above inequality holds with probability at least $1-2e^{-2m}$. Thus, we have

$$|f^\dagger(x)-\overline{f}(x)| \lesssim \|\phi_B(x)\|_2.$$

Next, we prove the intended result. By Assumption 1, we have

$$\left|\frac{1}{n}\sum_{i=1}^n \ell(f^\dagger(x),y_i)-\frac{1}{n}\sum_{i=1}^n \ell(\overline{f}(x),y_i)\right| \lesssim \frac{1}{n}\sum_{i=1}^n |f^\dagger(x_i)-\overline{f}(x_i)| \lesssim \frac{1}{n}\sum_{i=1}^n \|\phi_B(x_i)\|_2.$$

Then, there holds

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n \ell(f^\dagger(x),y_i)-\frac{1}{n}\sum_{i=1}^n \ell(\overline{f}(x),y_i)\right| \lesssim & \sqrt{\frac{1}{n}\sum_{i=1}^n \|\phi_B(x_i)\|_2^2} \\ = & \sqrt{\mathrm{tr}(\Phi_B^\top \Phi_B)/n} = \sqrt{\sum_{i=\lfloor Ch\rfloor+1}^h \lambda_i} \end{aligned} \tag{10}$$

By definition of $\hat{\alpha}$, we have

$$\frac{1}{n}\sum_{i=1}^n \ell(\widehat{f}(x_i),y)+\lambda\|\widehat{f}\|_{\mathcal{H}} \le \frac{1}{n}\sum_{i=1}^n \ell(f^\dagger(x_i),y)+\lambda\|f^\dagger\|_{\mathcal{H}}$$

which, together with (10), yields

$$\frac{1}{n}\sum_{i=1}^n \ell(\widehat{f}(x_i),y)-\frac{1}{n}\sum_{i=1}^n \ell(\overline{f}(x_i),y)+\lambda\|\widehat{f}\|_{\mathcal{H}} \lesssim \lambda+\sqrt{\sum_{i=\lfloor Ch\rfloor+1}^h \lambda_i}.$$

□

5.1.3 Proof of the Main Result

Proof of Theorem 5. Remind the error decomposition (6). Combining Lemma 6, 7, 8, and 9, there holds

$$\begin{aligned}\mathbb{E}[\ell(f_{\alpha_T}(x),y)]-\ell^* \lesssim & \sqrt{\frac{\log n}{n}}+\frac{1}{n}+\lambda+\sqrt{\sum_{i=\lfloor Ch\rfloor+1}^h \lambda_i}+\lambda\|f^*\|_{\mathcal{H}}-\lambda\|f_{\alpha_T}\|_{\mathcal{H}}-\lambda\|\overline{f}\|_{\mathcal{H}} \\ \lesssim & \sqrt{\frac{\log n}{n}}+\lambda+\sqrt{\sum_{i=\lfloor Ch\rfloor+1}^h \lambda_i}\end{aligned}$$

with probability $1-2e^{-2m}-2/n^2$. □

5.2 Privacy Proofs Related to Section 3

5.2.1 Technical Lemmas

In this section, we provide technical lemmas that are necessary for the subsequent analysis.

Lemma 25. *Suppose Assumption 3 and 4 hold. For dataset $\mathcal{D}$, let its feature map matrix be Φ and Σ be defined in Theorem 12. Also, let the orthogonal projection matrix P be defined in Definition 15 for any w. For neighboring dataset $\mathcal{D}'$ of $\mathcal{D}$, let Φ', Σ, and P' be defined analogously. Then we have*

$$\|\Sigma - \Sigma'\|_2 \leq \frac{2\sqrt{\kappa}}{\sqrt{n}}$$

and consequently

$$\lambda_i(\Sigma - (\Sigma - I_h)P) - \lambda_i(\Sigma' - (\Sigma' - I_h)P') \leq \frac{4\sqrt{\kappa}}{\sqrt{n}} + 2\|P - P'\|_2$$

for $i = 1, \cdots, h$. Here $\lambda_i(\Sigma - (\Sigma - I_h)P)$ is the i-th eigenvalue of $\Sigma - (\Sigma - I_h)P$.

Lemma 26. *Suppose Assumption 3 and 4 hold. For orthogonal projection matrix P, we have*

$$\frac{\lambda_h(\Sigma)}{2} \leq \lambda_i(\Sigma - (\Sigma - I_h)P) \leq 2$$

for $i = 1, \cdots, h$. Moreover, there are m of eigenvalues of $\Sigma - (\Sigma - I_h)P$ that are exactly 1.

Lemma 27. *Suppose Assumption 1 and 3 hold. Then for neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, we have*

$$\|\nabla_w J(w, \mathcal{D}) - \nabla_w J(w, \mathcal{D}')\|_2 \leq \frac{2c_L\sqrt{\kappa}}{n}.$$

Lemma 28. *Suppose Assumption 1 and 3 hold. For neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, let $G_w(\mathcal{D})$ and $G_w(\mathcal{D}')$ be defined in (8). Then, there exists a bijection $\mathcal{U}$ between $G_w(\mathcal{D})$ and $G_w(\mathcal{D}')$ such that, for any $g_{P,b}^{m,h} \in G_w(\mathcal{D})$ and its image $g_{P',b'}^{m,h} = \mathcal{U}(g_{P,b}^{m,h})$, there holds*

$$\|P - P'\|_2 \leq \frac{16c_L\sqrt{\kappa}}{\sqrt{n}}.$$

5.2.2 Proofs Related to Section 3.3

Proof of Lemma 13. By Bayes rule, there holds

$$p(g_{P,b}^{p,q}|\mathcal{D}) = p(P, b|\mathcal{D}) = p(P|\mathcal{D}) \cdot p(b|P, \mathcal{D}).$$

We deal with two terms separately. Since Φ is a $h \times n$ matrix with $h < n$, the SVD decomposition of Φ can be written as $U\Lambda V$ where U and V are $h \times h$ and $n \times n$ orthogonal matrix, respectively. Λ is a $h \times n$ matrix whose diagonal is filled with eigenvalues of Φ decreasingly. Notice that entries in the $h + 1, \cdots, n$-th rows of Λ are all zero. Let $\Sigma = \Lambda^\top\Lambda$. Σ is thus a $h \times h$ diagonal square

matrix with singular values of Φ, which are supposed to be sorted decreasingly. Since V is an orthogonal square matrix, $VS_{1:m}$ is distributed identically to $S_{1:m}$ as rectangular matrix-variate normal $\mathcal{N}_{n,m}(\mathbf{0}, I_n/m, I_m)$ [CC03, Section 1.5.3]. The density function of $p \times q$ matrix Y with the rectangular matrix-variate normal distribution $\mathcal{N}_{p,q}(M, \Sigma_1, \Sigma_2)$ is given by

$$p\left(Y|M, \Sigma_1, \Sigma_2\right) = \frac{|\Sigma_1|^{-q/2}\,|\Sigma_2|^{-p/2}}{(2\pi)^{pq/2}} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Sigma_2^{-1/2}(Y-M)^\top(Y-M)\Sigma_2^{-1/2}\right)\right]$$

for positive definite Σ_1 and Σ_2. Then $\Phi S_{1:m} = U\Lambda V S_{1:m}$ is distributed identically to $\mathcal{N}_{h,m}(\mathbf{0}, \Lambda^\top U^\top U\Lambda/m, I_m) = \mathcal{N}_{h,m}(\mathbf{0}, \Sigma/m, I_m)$. With this distribution of $\Phi S_{1:S}$, the corresponding density of P can be explicitly computed. According to [CC03, Remark 2.4.11], the density of associated rank m orthogonal projection matrix P is given by

$$p(P|\mathcal{D}) = m^{hm/2}|\Sigma|^{-m/2}|I_h - (I_h - \Sigma^{-1})P|^{-m/2}. \tag{11}$$

Then we turn to the second term $p(b|P, \mathcal{D})$. The distribution of b can be considered as a projection from $b_0 \sim \mathcal{N}(0, \Sigma)$ onto the kernel space of P, i.e. $\{b|Pb = 0\}$. The projection can be written as $(I_h - P)b_0$. Thus the conditional density is given by $b \sim \mathcal{N}(0, (I_h - P)^\top\Sigma(I_h - P))$. Note that the rank of $I_h - P$ is $h - m$. That is to say, the distribution of b is a singular Gaussian on the kernel space of P. The density function is given by

$$p(b|P, \mathcal{D}) = \frac{\prod_{\lambda_i>0}\lambda_i(\Pi)}{(2\pi)^{h/2}} \exp\big(-b^\top\Pi\, b\big)$$

where Π is the pseudo inverse of $(I_h - P)^\top\Sigma(I_h - P)$. This together with (11) yields

$$p(g_{P,b}^{m,h}) = \frac{m^{hm/2}\prod_{\lambda_i>0}\lambda_i(\Pi)}{(2\pi)^{h/2}\big|\Sigma - (\Sigma - I_h)P\big|^{m/2}} \cdot \exp\big(-b^\top\Pi\, b\big).$$

□

5.2.3 Proofs Related to Section 3.4

Proof of Lemma 16. (i) We first show that w-inducible implies w-optimal. According to [CC03, Section 2.4.4], the set $g_{P,b}^{m,h}$ is the same as $\{P\beta + w \mid \beta \in \mathbb{R}^h\}$. Thus, the original optimization problem $\min_{\tilde{w}\in g_{P,b}^{m,h}} J(f_{\tilde{w}}, \mathcal{D})$ is identical to solving the following optimization problem

$$\min_{\beta\in\mathbb{R}^h} J(f_{P\beta+b}, \mathcal{D}).$$

By definition 15, $b = (I_h - P)w$. Thus, we have $P\beta + b = P(\beta - w) + w$, which converts the problem into

$$\min_{\beta\in\mathbb{R}^h} J(f_{P\beta+w}, \mathcal{D})$$

Since both $\ell(f(x), y)$ and operator $\|f\|_{\mathcal{H}}$ is strictly convex with respect to f, the object $J(\cdot, \mathcal{D})$ is strictly convex. As a result, there holds

$$J(f_{P\beta+w}, \mathcal{D}) > J(f_w, \mathcal{D}) + (\nabla_\beta J(f_{P\beta+w}, \mathcal{D}) \mid_{\beta=0})^\top (\beta - \mathbf{0}).$$

Moreover, let $\tilde{w} = P\beta + w$. By the chain rule, there holds

$$\nabla_\beta J(f_{P\beta+w}, \mathcal{D}) = \nabla_w \tilde{w}^\top \nabla_{\tilde{w}} J(f_{\tilde{w}}, \mathcal{D}) = P^\top \nabla_w J(f_w, \mathcal{D}).$$

w-inducible implies that $P^\top \nabla_w J(f_w, \mathcal{D}) = \mathbf{0}$. Therefore, for all $\beta \in \mathbb{R}^h$, we have

$$J(f_{P\beta+w}, \mathcal{D}) > J(f_w, \mathcal{D})$$

except for $\beta = \mathbf{0}$. Thus, the optimal solution to the optimization problem in Definition 14 is w, i.e. $g_{P,b}^{m,h}$ is w-optimal.

(ii) For the opposite direction, if $g_{P,b}^{m,h}$ leads to w being the optimal solution, it is clear that $w \in g_{P,b}^{m,h}$. Again we can write problem $\min_{\tilde{w} \in g_{P,b}^{m,h}} J(f_{\tilde{w}}, \mathcal{D})$ as

$$\min_{\beta \in \mathbb{R}^h} J(f_{P\beta+w}, \mathcal{D})$$

with $\beta = \mathbf{0}$ being its optimal solution. By KKT necessary condition [BBV04], there holds

$$\mathbf{0} = \nabla_\beta J(f_{P\beta+w}, \mathcal{D}) \mid_{\beta=\prime} = P^\top \nabla_{\tilde{w}} J(f_{\tilde{w}}, \mathcal{D}).$$

For the second statement, note that $P^\top b = \mathbf{0}$. Thus, b is the projection on the kernel space of $\mathrm{col}(P)$, namely $(I_h - P)w$. □

5.2.4 Proofs Related to Section 3.5

Proof of Lemma 17. Let $\overline{w}$ be the optimal solution of $\min_{w \in \mathbb{R}^h} J(f_w, \mathcal{D})$. Consequently, $\nabla_w J(f_w, \mathcal{D})|_{w=\overline{w}} = \mathbf{0}$. By Assumption 11, $\|\nabla_w J(f_w, \mathcal{D})\|_2^2$ is strictly positive definite by some constant $\tilde{\Delta}$. Thus we have

$$\begin{aligned}\|\nabla_w J(f_w, \mathcal{D})\|_2^2 &\geq 0 + \left(\nabla_w^\top \|\nabla_w J(f_w, \mathcal{D})\|_2^2\right)\Big|_{w=\overline{w}} \cdot (w - \overline{w}) + \frac{\tilde{\Delta}}{4}(w - \overline{w})^\top \cdot (w - \overline{w}) \\ &= \frac{\tilde{\Delta}}{4}(w - \overline{w})^\top \cdot (w - \overline{w}).\end{aligned}$$

Here we used the fact that

$$\nabla_w \|\nabla_w J(f_w, \mathcal{D})\|_2^2\big|_{w=\overline{w}} = \left(2 \cdot \nabla_w^2 J(f_w, \mathcal{D}) \cdot \nabla_w J(f_w, \mathcal{D})\right)\big|_{w=\overline{w}} = \mathbf{0}.$$

Then the region $\{w \mid \|\nabla_w J(f_w, \mathcal{D})\|_2 \leq 1/\sqrt{n}\}$ is an subset of the ellipse

$$\left\{ w \,\middle|\, \frac{\tilde{\Delta}}{4}(w - \overline{w})^\top \cdot (w - \overline{w}) \leq \frac{1}{n} \right\}$$

whose volume is $V_d\ (4\tilde{\Delta}^{-1}n^{-1})^{d/2}$. Reminding the density function in Lemma 13

$$p(g_{P,b}^{m,h}) = \frac{m^{hm/2}\prod_{\lambda_i>0}\lambda_i(\Pi)}{(2\pi)^{h/2}\left|\Sigma-(\Sigma-I_h)P\right|^{m/2}}\cdot\exp\big(-b^\top\Pi\, b\big)$$

where $\Pi=(I_h-P)^\top\Sigma(I_h-P)$. By Lemma 26, we have

$$\left|\Sigma-(\Sigma-I_h)P\right|\geq\lambda_h(\Sigma-(\Sigma-I_h)P)^{(h-m)}\geq\lambda_h(\Sigma)^{(h-m)}.$$

Also, by Assumption 4, we have

$$\prod_{\lambda_i>0}\lambda_i(\Pi)\leq\lambda_h(\Sigma)^{-(h-m)}.$$

since I_h-P is a rank-$(h-m)$ matrix. Also, since Π is semi-positive definite, we have

$$\exp\big(-b^\top\Pi\, b\big)\leq 0.$$

Combining these pieces, we can bound $p(g_{P,b}^{m,h})$ by

$$p(g_{P,b}^{m,h})\leq\frac{m^{hm/2}\lambda_h^{-(h-m)(m+2)/2}}{(2\pi)^{h/2}}.$$

As a consequence, the probability density of $G_w(\mathcal{D})$ is bounded by

$$p(G_w(\mathcal{D}))=\int_{g\in G_w(\mathcal{D})}p(g)\frac{d\gamma^{m,h}}{dw}$$

Thus, the density of each w, equivalently of each $G_w(\mathcal{D})$, is also uniformly bounded by the quantity of $m^{h(m-1)/2}(2\pi)^{-h/2}n^{\nu(m-h)m/2}$. Together, these yields

$$P(\{G_w(\mathcal{D}),w\in\Delta_n^c\}|\mathcal{D})\leq V_d4^dC^{-d}m^{h(m-1)/2}(2\pi)^{-h/2}n^{\nu(m-h)m/2-d/2}.$$

□

Proof of Lemma 18. It suffices to show that, for each w, we have

$$\frac{p(g_{P,b}^{m,h}|\mathcal{D})}{p(g_{P',b'}^{m,h}|\mathcal{D}')}\leq e^\epsilon$$

for $g_{P,b}^{m,h}$ and $g_{P',b'}^{m,h}$ given in Lemma 28. If so, there holds

$$\frac{\mathbb{P}[w\in\mathcal{T}|\mathcal{D}]}{\mathbb{P}[w\in\mathcal{T}|\mathcal{D}']}=\frac{\int_{g\in G_w(\mathcal{D}),w\in\mathcal{T}}p(g)d\gamma^{m,h}}{\int_{g'\in G_w(\mathcal{D}'),w\in\mathcal{T}}p(g')d\gamma^{m,h}}\leq\sup\frac{p(g_{P,b}^{m,h}|\mathcal{D})}{p(g_{P',b'}^{m,h}|\mathcal{D}')}\leq e^\epsilon.$$

The density function $p(g_{P,b}^{m,h}|\mathcal{D})$ has three parts, and we deal with their ratios separately.

(i) We first bound the determinant part, which is

$$\frac{|\Sigma-(\Sigma-I_h)P|}{|\Sigma'-(\Sigma'-I_h)P'|}=\frac{\prod_{i=1}^h\lambda_i(\Sigma-(\Sigma-I_h)P)}{\prod_{i=1}^h\lambda_i(\Sigma'-(\Sigma'-I_h)P')}. \tag{12}$$

By Lemma 26, there are m eigenvalues that are exactly 1. For the rest eigenvalues, we provide an upper bound for

$$\frac{\lambda_i(\Sigma-(\Sigma-I_h)P)}{\lambda_i(\Sigma'-(\Sigma'-I_h)P')}$$

for each $i=1,\cdots,h$. By Lemma 25, we have

$$\begin{aligned}\frac{\lambda_i(\Sigma-(\Sigma-I_h)P)}{\lambda_i(\Sigma'-(\Sigma'-I_h)P')}\leq&1+\frac{\lambda_i(\Sigma-(\Sigma-I_h)P)-\lambda_i(\Sigma'-(\Sigma'-I_h)P')}{\lambda_i(\Sigma'-(\Sigma'-I_h)P')}\\ \leq&1+\frac{4\sqrt{\kappa}/\sqrt{n}+2\|P-P'\|_2}{\lambda_i(\Sigma'-(\Sigma'-I_h)P')}.\end{aligned}$$

The lower bound of λ_i in Lemma 26 yields

$$\frac{\lambda_i(\Sigma-(\Sigma-I_h)P)}{\lambda_i(\Sigma'-(\Sigma'-I_h)P')}\leq 1+\frac{8\sqrt{\kappa}}{\lambda_h\sqrt{n}}+\frac{4\|P-P'\|_2}{\lambda_h}.$$

Combining Lemma 28, we have

$$\frac{\lambda_i(\Sigma-(\Sigma-I_h)P)}{\lambda_i(\Sigma'-(\Sigma'-I_h)P')}\leq 1+\frac{8\sqrt{\kappa}}{\lambda_h\sqrt{n}}+\frac{64c_L\sqrt{\kappa}}{\lambda_h\sqrt{n}}.$$

Bring this into (12), we have

$$\frac{|\Sigma-(\Sigma-I_h)P|}{|\Sigma'-(\Sigma'-I_h)P'|}\leq\left(1+\frac{(8+64c_L)\sqrt{\kappa}}{\lambda_h\sqrt{n}}\right)^{h-m}$$

and consequently

$$\begin{aligned}\frac{|\Sigma-(\Sigma-I_h)P|^{m/2}}{|\Sigma'-(\Sigma'-I_h)P'|^{m/2}}\leq&\left(1+\frac{(8+64c_L)\sqrt{\kappa}}{\lambda_h\sqrt{n}}\right)^{(h-m)m/2}\\ \leq&\exp\left(\frac{(8+64c_L)\sqrt{\kappa}}{\lambda_h\sqrt{n}}\frac{(h-m)m}{2}\right).\end{aligned} \tag{13}$$

(ii) Next, we bound the exponential part

$$\frac{\exp\big(-b^\top\big[(I_h-P)^\top\Sigma(I_h-P)\big]^{-1}b\big)}{\exp\big(-b'^\top\big[(I_h-P')^\top\Sigma'(I_h-P')\big]^{-1}b'\big)}.$$

We first give a bound on the matrix norm

$$\Big\|\big[(I_h-P)^\top\Sigma(I_h-P)\big]^{-1}-\big[(I_h-P')^\top\Sigma'(I_h-P')\big]^{-1}\Big\|_2.$$

Basic matrix inversion leads to

$$\begin{aligned}
&\left\|\left[(I_h-P)^\top\Sigma(I_h-P)\right]^{-1}-\left[(I_h-P')^\top\Sigma'(I_h-P')\right]^{-1}\right\|_2\\
\le&\left\|\left[(I_h-P)^\top\Sigma(I_h-P)\right]^{-1}\right\|_2\cdot\left\|\left[(I_h-P')^\top\Sigma'(I_h-P')\right]^{-1}\right\|_2\\
&\cdot\left\|(I_h-P)^\top\Sigma(I_h-P)-(I_h-P')^\top\Sigma'(I_h-P')\right\|_2
\end{aligned}\tag{14}$$

By Lemma 26, we have

$$\left\|\left[(I_h-P)^\top\Sigma(I_h-P)\right]^{-1}\right\|_2\cdot\left\|\left[(I_h-P')^\top\Sigma'(I_h-P')\right]^{-1}\right\|_2\le\frac{4}{\lambda_h^2}.\tag{15}$$

For the last term, we use the following decomposition

$$\begin{aligned}
&(I_h-P)^\top\Sigma(I_h-P)-(I_h-P')^\top\Sigma'(I_h-P')\\
=&(I_h-P)^\top\Sigma(I_h-P)-(I_h-P')^\top\Sigma(I_h-P)\\
+&(I_h-P')^\top\Sigma(I_h-P)-(I_h-P')^\top\Sigma'(I_h-P)\\
+&(I_h-P')^\top\Sigma'(I_h-P)-(I_h-P')^\top\Sigma'(I_h-P')
\end{aligned}$$

to get

$$\begin{aligned}
&\left\|(I_h-P)^\top\Sigma(I_h-P)-(I_h-P')^\top\Sigma'(I_h-P')\right\|_2\\
\le&\|P-P'\|_2\|\Sigma(I_h-P)\|_2+\|I_h-P'\|_2\|\Sigma-\Sigma'\|_2\|I_h-P\|_2+\|(I_h-P')\Sigma'\|_2\|P-P'\|_2\\
\le&2\|P-P'\|_2+\|\Sigma-\Sigma'\|_2
\end{aligned}$$

Combining Lemma 25 and Lemma 28, this becomes

$$\left\|(I_h-P)^\top\Sigma(I_h-P)-(I_h-P')^\top\Sigma'(I_h-P')\right\|_2\le\frac{32c_L\sqrt{\kappa}+3\sqrt{\kappa}}{\sqrt{n}}.\tag{16}$$

Bring (15) and (16) into (14), we have

$$\left\|\left[(I_h-P)^\top\Sigma(I_h-P)\right]^{-1}-\left[(I_h-P')^\top\Sigma'(I_h-P')\right]^{-1}\right\|_2\le\frac{128c_L\sqrt{\kappa}+8\sqrt{\kappa}}{\lambda_h^2\sqrt{n}}.$$

Since $b=(I_h-P)w$ and $b'=(I_h-P')w$, we have

$$\|b-b'\|_2\le\|P-P'\|_2\|w\|_2\le\frac{16c_L\sqrt{\kappa}}{\sqrt{n}}.$$

Also, we have apparently $\|b\|_2\le1$ since we only consider $\|w\|_2\le1$. Together, we have

$$\begin{aligned}
&b^\top\left[(I_h-P)^\top\Sigma(I_h-P)\right]^{-1}b-b'^\top\left[(I_h-P')^\top\Sigma'(I_h-P')\right]^{-1}b'\\
\le&\|b-b'\|_2\left\|\left[(I_h-P)^\top\Sigma(I_h-P)\right]^{-1}\right\|_2\left(\|b\|_2+\|b'\|_2\right)\\
&+\left\|\left[(I_h-P)^\top\Sigma(I_h-P)\right]^{-1}-\left[(I_h-P')^\top\Sigma'(I_h-P')\right]^{-1}\right\|_2\|b\|_2\|b'\|_2\\
\le&\frac{64c_L\sqrt{\kappa}}{\lambda_h\sqrt{n}}+\frac{128c_L\sqrt{\kappa}+8\sqrt{\kappa}}{\lambda_h^2\sqrt{n}}.
\end{aligned}\tag{17}$$

where in the last inequality we used Lemma 26.

(*iii*) For the ratio of $|\Pi|$ part, we analogously bound for

$$\frac{\lambda_i((I_h - P)^\top \Sigma (I_h - P))}{\lambda_i((I_h - P')^\top \Sigma' (I_h - P'))}$$

for $i = 1, \cdots, h$. By Weyl's Theorem [HJ12, Theorem 4.3.1], (16) yields that

$$\begin{aligned}
&\lambda_i((I_h - P)^\top \Sigma (I_h - P)) - \lambda_i((I_h - P')^\top \Sigma' (I_h - P')) \\
&\leq \left\| ((I_h - P)^\top \Sigma (I_h - P)) - ((I_h - P')^\top \Sigma' (I_h - P')) \right\|_2 \leq \frac{32 c_L \sqrt{\kappa} + 3\sqrt{\kappa}}{\sqrt{n}}.
\end{aligned}$$

Thus combining Lemma 26, we have

$$\begin{aligned}
\frac{\lambda_i((I_h - P)^\top \Sigma (I_h - P))}{\lambda_i((I_h - P')^\top \Sigma' (I_h - P'))} &= 1 + \frac{\left\| ((I_h - P)^\top \Sigma (I_h - P)) - ((I_h - P')^\top \Sigma' (I_h - P')) \right\|_2}{\lambda_i((I_h - P')^\top \Sigma' (I_h - P'))} \\
&\leq 1 + \frac{64 c_L \sqrt{\kappa} + 6\sqrt{\kappa}}{\lambda_h \sqrt{n}}
\end{aligned}$$

The same argument in the proof of Lemma 26 yields only $h - m$ eigenvalues that are non-zero. Thus, we have

$$\begin{aligned}
\frac{\prod_{\lambda_i > 0} \lambda_i((I_h - P)^\top \Sigma (I_h - P))}{\prod_{\lambda_i > 0} \lambda_i((I_h - P')^\top \Sigma' (I_h - P'))} &\leq (1 + \frac{64 c_L \sqrt{\kappa} + 6\sqrt{\kappa}}{\lambda_h \sqrt{n}})^{h-m} \\
&\leq \exp\left(\frac{(64 c_L \sqrt{\kappa} + 6\sqrt{\kappa})(h - m)}{\lambda_h \sqrt{n}} \right). \qquad (18)
\end{aligned}$$

With these conclusions, (13), (17) and (18) together yield the desired result that

$$p(g_{P,b}^{m,h} | \mathcal{D}) / p(g_{P',b'}^{m,h} | \mathcal{D}') \leq e^\epsilon$$

where we have

$$\epsilon = \frac{(8 + 64 c_L)\sqrt{\kappa}}{C_h} \frac{(h - m)m}{2} n^{-1/2+\nu} + \frac{32 c_L \sqrt{\kappa}}{C_h} n^{-1/2+\nu} + + \frac{32 c_L \sqrt{\kappa} + 2\sqrt{\kappa}}{C_h^2} n^{-1/2+2\nu}.$$

This value is dominated by

$$\epsilon \leq \frac{64 c_L \sqrt{\kappa}}{C_h} (h - m) m n^{-1/2+\nu} + \frac{32 c_L \sqrt{\kappa} + 2\sqrt{\kappa}}{C_h^2} n^{-1/2+2\nu}$$

for sufficiently large n.

□

5.2.5 Proof of the Main Result

Proof of Theorem 12. Consider w the final output of Algorithm 1. We know from Lemma 16 that, given a fixed dataset $\mathcal{D}$, there is a bijective between w and $G_w(\mathcal{D})$. Let $G_w(\mathcal{D})$ and $G_w(\mathcal{D})'$ be the w-optimal set conditioned on $\mathcal{D}$ and $\mathcal{D}'$ respectively. To show (ϵ, δ) differential privacy, we need to compute the ratio of probability of outputting $w \in \mathcal{T}$ conditioned on dataset $\mathcal{D}$ and $\mathcal{D}'$, where $\mathcal{T}$ is any set in $\mathbb{R}^h$. $P(w \in \mathcal{T}|\mathcal{D})$ can be written as

$$P(w \in \mathcal{T}|\mathcal{D}) = P(\{G_w(\mathcal{D}), w \in \mathcal{T}\}|\mathcal{D}). \tag{19}$$

By Lemma 18, we have

$$P(G_w(\mathcal{D})|\mathcal{D}) \leq e^{\epsilon} P(G_w(\mathcal{D})|\mathcal{D}') \tag{20}$$

for each $w \in \Delta_n$ where Δ_n is defined in (9). Moreover, we have $P(\{G_w(\mathcal{D}), w \in \Delta_n^c\}|\mathcal{D}) \leq \delta$ by Lemma 17. This together with (20) yield

$$P(\{G_w(\mathcal{D}), w \in \mathcal{T}\}|\mathcal{D}) \leq e^{\epsilon} P(\{G_w(\mathcal{D}), w \in \mathcal{T}\}|\mathcal{D}') + \delta.$$

This exactly gives (ϵ, δ) privacy in Definition 10 by bringing in (19). □

References

[Bas+17] Raef Bassily et al. "Practical locally private heavy hitters". In: *Advances in Neural Information Processing Systems* 30 (2017).

[Bas+19] Raef Bassily et al. "Private stochastic convex optimization with optimal rates". In: *Advances in neural information processing systems* 32 (2019).

[BBM05] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. "Local rademacher complexities". In: *The Annals of Statistics* 33.4 (2005), pp. 1497–1537.

[BBV04] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[Bra05] Mikio Ludwig Braun. "Spectral properties of the kernel matrix and their relation to kernel methods in machine learning". PhD thesis. Universitäts-und Landesbibliothek Bonn, 2005.

[BST14] Raef Bassily, Adam Smith, and Abhradeep Thakurta. "Private empirical risk minimization: Efficient algorithms and tight error bounds". In: *2014 IEEE 55th annual symposium on foundations of computer science*. IEEE. 2014, pp. 464–473.

[CC03] Yasuko Chikuse and Y Chikuse. *Statistics on special manifolds*. Vol. 1. Springer, 2003.

[CM08] Kamalika Chaudhuri and Claire Monteleoni. "Privacy-preserving logistic regression". In: *Advances in neural information processing systems* 21 (2008).

[CMS11] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. "Differentially private empirical risk minimization." In: *Journal of Machine Learning Research* 12.3 (2011).

[Dri+11] Petros Drineas et al. "Faster least squares approximation". In: *Numerische mathematik* 117.2 (2011), pp. 219–249.

[Dwo+06a] Cynthia Dwork et al. "Calibrating noise to sensitivity in private data analysis". In: *Theory of cryptography conference*. Springer. 2006, pp. 265–284.

[Dwo+06b] Cynthia Dwork et al. "Our data, ourselves: Privacy via distributed noise generation". In: *Annual international conference on the theory and applications of cryptographic techniques*. Springer. 2006, pp. 486–503.

[Haa33] Alfred Haar. "Der Massbegriff in der Theorie der kontinuierlichen Gruppen". In: *Annals of mathematics* (1933), pp. 147–169.

[HJ12] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[HRW13] Rob Hall, Alessandro Rinaldo, and Larry Wasserman. "Differential privacy for functions and functional data". In: *Journal of Machine Learning Research* 14.Feb (2013), pp. 703–727.

[JT13] Prateek Jain and Abhradeep Thakurta. "Differentially private learning with kernels". In: *International conference on machine learning*. PMLR. 2013, pp. 118–126.

[KRB97] D.A. Klain, G.C. Rota, and L.A.R. di Brozolo. *Introduction to Geometric Probability*. Lezioni Lincee. Cambridge University Press, 1997. ISBN: 9780521596541. URL: https://books.google.co.in/books?id=Q1ytkNM6BtAC.

[LT91] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Vol. 23. Springer Science & Business Media, 1991.

[LWY21] Lek-Heng Lim, Ken Sze-Wai Wong, and Ke Ye. "The Grassmannian of affine subspaces". In: *Foundations of Computational Mathematics* 21.2 (2021), pp. 537–574.

[MS] John Milnor and JAMES STASHEFF. "Characteristic Classes.(AM-76)". In: ().

[Nic20] Liviu I Nicolaescu. *Lectures on the Geometry of Manifolds*. World Scientific, 2020.

[Pic09] Clifford A Pickover. *The math book: from Pythagoras to the 57th dimension, 250 milestones in the history of mathematics*. Sterling Publishing Company, Inc., 2009.

[RR07] Ali Rahimi and Benjamin Recht. "Random features for large-scale kernel machines". In: *Advances in neural information processing systems* 20 (2007).

[Rub+12] Benjamin IP Rubinstein et al. "Learning in a Large Function Space: Privacy-Preserving Mechanisms for SVM Learning". In: *Journal of Privacy and Confidentiality* 4.1 (2012), pp. 65–100.

[SC08] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[Tal94] Michel Talagrand. "Sharper bounds for Gaussian and empirical processes". In: *The Annals of Probability* (1994), pp. 28–76.

[VVW96] AW van der Vaart, Aad van der Vaart, and Jon Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media, 1996.

[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019.

[WYX17] Di Wang, Minwei Ye, and Jinhui Xu. "Differentially private empirical risk minimization revisited: Faster and more general". In: *Advances in Neural Information Processing Systems* 30 (2017).

[YPW17] Yun Yang, Mert Pilanci, and Martin J. Wainwright. "Randomized sketches for kernels: Fast and optimal nonparametric regression". In: *The Annals of Statistics* 45.3 (2017), pp. 991–1023. DOI: 10.1214/16-AOS1472. URL: https://doi.org/10.1214/16-AOS1472.

A Technical Proofs

A.1 Proofs of Results in Appendix 5.1.1

Proof of Lemma 19. We prove the two arguments separately.

(i) First, note that $\Phi_A\Phi_A^\top$ is symmetric and thus $(\Phi_A\Phi_A^\top)^{-1/2}$ is also symmetric. Then, there holds

$$(\Phi_A\Phi_A^\top)^{-1/2}\Phi_A\Phi_A^\top\big((\Phi_A\Phi_A^\top)^{-1/2}\big)^\top = (\Phi_A\Phi_A^\top)^{-1/2}\Phi_A\Phi_A^\top(\Phi_A\Phi_A^\top)^{-1/2} = I_{\lfloor Ch\rfloor}$$

which yields that $(\Phi_A\Phi_A^\top)^{-1/2}\Phi_A$ has orthogonal rows. With this, we begin to bound the operator norm of the matrix $Q = (\Phi_A\Phi_A^\top)^{-1/2}\Phi_A SS^\top\Phi_A^\top(\Phi_A\Phi_A^\top)^{-1/2} - I_{\lfloor Ch\rfloor}$. Let $\{v^1,\ldots,v^N\}$ be a 1/2-cover of the Euclidean sphere $\mathcal{S}^{\lfloor ch\rfloor-1}$; by standard arguments [Wai19], we can find such a set with $N \le e^{2\lfloor Ch\rfloor}$ elements. Using this cover, a straightforward discretization argument yields

$$\|Q\|_{\rm op} \le 4\max_{j,k=1,\ldots,N} v^{j\top}Qv^k = 4\max_{j,k=1,\ldots,N}\widetilde{v}^{j\top}(S^\top S - I_n)\widetilde{v}^k,$$

where $\widetilde{v}^j = \Phi_A^\top(\Phi_A\Phi_A^\top)^{-1/2}v^j \in \mathcal{S}^{n-1}$. Since each entry of S is an i.i.d. Gaussian, we can apply the concentration result by using sub-exponential bounds [Wai19, Proposition 2.9] to obtain

$$\mathbb{P}\left[\widetilde{v}^{j\top}(S^\top S - I_n)\widetilde{v}^k \ge 1/8\right] \le c_1e^{-c_2m}, \quad j,k = 1,\cdots,N$$

for some constant c_1, c_2. Consequently, by the union bound, for any constant $C \le c_2/8$, we have

$$\mathbb{P}\left[\|Q\|_{\rm op} \ge 1/2\right] \le c_1e^{-c_2m+4\lfloor Ch\rfloor} \le c_1e^{-8Cm+4Ch} \le c_1e^{-2Cm} \le e^{-2m}$$

for sufficiently large m. Here we used the assume lower bound $m \ge 2h/3$.

(ii) For notation simplicity, we use $\overline{\Phi}_B$ to denote the augmented feature matrix $(0,\Phi_B^\top)^\top$. For the second argument, we want to bound the operator norm

$$\|\Phi_B S\|_2 = \|\overline{\Phi}_B S\|_2 = \sup_{u\in\mathcal{S}^{m-1},v\in\mathcal{E}} v^\top Su.$$

where we define $\mathcal{E} = \left\{\overline{\Phi}_B^\top w \mid \|\|w\|_2 \le 1\right\}$, and $\mathcal{S}^{m-1} = \{u\in\mathbb{R}^m \mid \|u\|_2 = 1\}$. Again with standard discretization arguments [Wai19], we can find a 1/2-cover $\{u^1,\ldots,u^N\}$ of the set $\mathcal{S}^{m-1}$ of the set with $N \le e^{2m}$ elements that guarantees

$$\|\Phi_B S\|_2 \le 2\max_{j\in[N]}\sup_{v\in\mathcal{E}} v^\top Su^j$$

For each fixed $u^j \in \mathcal{S}^{m-1}$, consider the random variable $\sup_{v\in\mathcal{E}} v^\top Su^j$ for $j = 1,\cdots,N$. It is equal in distribution to the random variable $V(g) = \frac{1}{\sqrt{m}}\sup_{v\in\mathcal{E}} g^\top v$, where $g\in\mathbb{R}^n$ is a standard Gaussian vector. For $g, g'\in\mathbb{R}^n$, we have

$$\left|V(g) - V\left(g'\right)\right| \le \frac{2}{\sqrt{m}}\sup_{v\in\mathcal{E}}(g-g')^\top v \le \frac{2}{\sqrt{m}}\|g-g'\|_2\|v\|_2.$$

Definition of $\mathcal{E}$ yields that $\|v\|_2 \leq \|\Phi_B\|_2 \leq \sqrt{\lambda_{\lfloor Ch \rfloor+1}}$. Consequently, by the concentration of measure for Lipschitz functions of Gaussian random variables [Wai19, Theorem 2.26], we have

$$\mathbb{P}[V(g) \geq \mathbb{E}[V(g)] + t] \leq \exp\left(- \frac{mt^2}{8\lambda_{\lfloor Ch \rfloor+1}} \right).$$

Turning to the expectation we have

$$\mathbb{E}[V(g)] = \frac{1}{\sqrt{m}} \mathbb{E}[\sup_{v \in \mathcal{E}} g^\top v] \leq \frac{1}{\sqrt{m}} \mathbb{E}\left[\sqrt{g^\top \overline{\Phi}_B^\top \overline{\Phi}_B g} \right].$$

By Jensen inequality, this leads to

$$\mathbb{E}[V(g)] \leq \frac{1}{\sqrt{m}} \sqrt{\mathbb{E}[g^\top \overline{\Phi}_B^\top \overline{\Phi}_B g]} = \frac{1}{\sqrt{m}} \sqrt{\sum_{i=\lfloor Ch \rfloor+1}^{h} \lambda_i} \leq \sqrt{\frac{(1-C)h}{m} \lambda_{\lfloor Ch \rfloor+1}}.$$

Since $m \geq 2h/3$, we get $\mathbb{E}[V(g)] \lesssim \lambda_{\lfloor Ch \rfloor+1}^{1/2}$. Combining the pieces, we have shown that

$$\mathbb{P}\left[\sup_{v \in \mathcal{E}} v^\top S u^j \geq c' \lambda_{\lfloor Ch \rfloor+1}^{1/2} + t \right] \leq e^{-\frac{mt^2}{8\lambda_{\lfloor Ch \rfloor+1}}}$$

for each $j = 1, \ldots, N$. Then, we take $t = 4\sqrt{2}\lambda_{\lfloor Ch \rfloor+1}^{1/2}$ and take the union bound over all $j \in [N]$. This leads to

$$\mathbb{P}\left[\|\Phi_B S\|_2 \geq c'' \lambda_{\lfloor Ch \rfloor+1}^{1/2} \right] \leq e^{-4m+2m} = e^{-2m}.$$

which completes the proof.

□

Proof of Lemma 21. For the kernel class, we have

$$\begin{aligned}
\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right] &= \frac{1}{n} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i \langle f, k(\cdot, x_i) \rangle \right] \\
&= \frac{1}{n} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left\langle f, \sum_{i=1}^n \epsilon_i k(\cdot, x_i) \right\rangle \right].
\end{aligned}$$

By the representer theorem, the maximizer of the above quantity can be explicitly formulated as

$$f = \sup_{f \in \mathcal{F}} \|f\|_{\mathcal{F}} \frac{\sum_{i=1}^n \epsilon_i k(\cdot, X_i)}{\|\sum_{i=1}^n \epsilon_i k(\cdot, X_i)\|_{\mathcal{H}}}.$$

Thus, there holds

$$\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right] = \frac{\sup_{f \in \mathcal{F}} \|f\|_{\mathcal{F}}}{n} \mathbb{E}_\epsilon \left[\left\| \sum_{i=1}^n \epsilon_i k(\cdot, x_i) \right\|_{\mathcal{H}} \right].$$

Reformulate kernel function as feature map yields

$$\mathbb{E}_\epsilon\left[\left\|\sum_{i=1}^n \epsilon_i k\left(\cdot, x_i\right)\right\|_{\mathcal{H}}\right]=\mathbb{E}_\epsilon\left[\sqrt{\left\|\sum_{i=1}^n \epsilon_i \phi(x_i)\right\|_2}\right]=\mathbb{E}_\epsilon\left[\sqrt{\sum_{i,j=1}^n \epsilon_i \epsilon_j \phi(x_i)^\top \phi(x_j)}\right].$$

Then, by Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}_\epsilon\left[\sqrt{\sum_{i,j=1}^n \epsilon_i \epsilon_j \phi(x_i)^\top \phi(x_j)}\right] \leq & \sqrt{\mathbb{E}_\epsilon\left[\sum_{i,j=1}^n \epsilon_i \epsilon_j \phi(x_i)^\top \phi(x_j)\right]} \\ = & \sqrt{\sum_{i=1}^n \phi(x_i)^\top \phi(x_i)} \leq \sqrt{n\kappa}\end{aligned}$$

which completes the proof. □

Proof of Lemma 22. The lemma is a direct corollary of Lemma 21 since $\{S\alpha|\alpha \in \mathbb{R}^m\}$ is a subset of $\mathbb{R}^n$. □

Proof of Lemma 23. Note that $\tilde{\mathcal{F}}$ is a subset of $\mathcal{F}$. Thus, by monotonicity of local Rademacher complexity with respect to the function class, the conclusion holds.

□

Proof of Lemma 24. The lemma is a direct corollary of Theorem 2.1 in [BBM05]. We let $x = 2\log n$ and $\alpha = 2/3$. □

A.2 Proofs of Results in Appendix 5.1.1

Proof of Lemma 25. By Weyl's Theorem [HJ12, Theorem 4.3.1] , we have

$$\lambda_i(\Sigma-(\Sigma-I_h)P)-\lambda_i(\Sigma'-(\Sigma'-I_h)P') \leq \|\Sigma-(\Sigma-I_h)P-\Sigma'-(\Sigma'-I_h)P'\|_2.$$

Then, by triangle inequality, there holds

$$\begin{aligned}&\|\Sigma-(\Sigma-I_h)P-\Sigma'-(\Sigma'-I_h)P'\|_2 \\ \leq&\|\Sigma-\Sigma'\|_2+\|P-P'\|_2+\|P'\|_2\|\Sigma-\Sigma'\|_2+\|\Sigma\|_2\|P-P'\|_2 \\ \leq&2\big(\|\Sigma-\Sigma'\|_2+\|P-P'\|_2\big)\end{aligned}$$

where the last inequality holds because $\|P'\|_2 = 1$ by definition and $\lambda_i(\Sigma) \leq 1$ as in Assumption 4. It remains to show that $\|\Sigma - \Sigma'\|_2 \leq 2\sqrt{\kappa}/\sqrt{n}$. Since $\mathcal{D}$ and $\mathcal{D}'$ are neighboring data sets, Φ and Φ' only differ in their first column. Then, $\Phi^\top\Phi$ and $\Phi'^\top\Phi'$ only differ in their first row and first column, which consist of $2n-1$ entries in total. As a result, we have

$$\|\Phi^\top\Phi-\Phi'^\top\Phi'\|_2 \leq \|\Phi^\top\Phi-\Phi'^\top\Phi'\|_F \leq \sqrt{2\kappa(2n-1)}.$$

The last inequality holds since the (i,j)-th element of $\Phi^\top\Phi - \Phi'^\top\Phi'$ is $k(x_i,x_j) - k(x_i',x_j')$, which is absolutely bounded by 2κ and is non-zero only if $i=1$ or $j=1$. Consequently, we have

$$\|\Sigma - \Sigma'\|_2 \le \frac{1}{n}\|\Phi^\top\Phi - \Phi'^\top\Phi'\|_2 \le \frac{2\sqrt{\kappa}}{\sqrt{n}}$$

which yields the desired result.

□

Proof of Lemma 26. Remind that, by definition, m of the eigenvalues of P is 1 and the others are 0. Since P is symmetric, its eigenspaces are mutually orthogonal. Let u be an eigenvector of $\Sigma - (\Sigma - I_h)P$. u has a unique orthogonal decomposition $u^0 + u^1$ where u^i is the projection onto the eigenspace of P with eigenvalue i for $i = 0, 1$. Then,

$$\lambda u = \big(\Sigma - (\Sigma - I_h)P\big)u = \Sigma u - (\Sigma - I_h)u^1 = \Sigma u^0 + u^1. \tag{21}$$

Thus, for each eigenvector $\tilde{u}^1$ of P with eigenvalue 1, it is also an eigenvector of $\Sigma - (\Sigma - I_h)P$ with eigenvalue 1. This is to say, m of eigenvalues of $\Sigma - (\Sigma - I_h)P$ are 1. Assume that $\|u\|_2$ is 1. Multiply (21) by u, we have

$$\begin{aligned}\lambda = u^1\Sigma u^0 + u^0\Sigma u^0 + u^{1\top}u^1 \ge& -\frac{1}{2}u^0\Sigma u^0 - \frac{1}{2}u^1\Sigma u^1 + u^0\Sigma u^0 + u^{1\top}u^1 \\ =& \frac{1}{2}u^0\Sigma u^0 - \frac{1}{2}u^1\Sigma u^1 + u^{1\top}u^1.\end{aligned}$$

Reminding Assumption 4 that $\lambda_h(\Sigma) \le \lambda_i(\Sigma) \le 1$, we have

$$\frac{1}{2}u^0\Sigma u^0 - \frac{1}{2}u^1\Sigma u^1 + u^{1\top}u^1 \ge \frac{1}{2}\big(u^0\Sigma u^0 + u^1\Sigma u^1\big) \ge \frac{\lambda_h(\Sigma)}{2}\big(\|u^0\|_2^2 + \|u^1\|_2^2\big) = \frac{\lambda_h(\Sigma)}{2}.$$

Obviously, we also have

$$\lambda = \|\lambda u\|_2 \le \|\Sigma u^0\|_2 + \|u^1\|_2 \le 2.$$

□

Proof of Lemma 27. Remind the definition of $J(f_w, \mathcal{D})$, taking derivative yields

$$\nabla_w J(f_w,\mathcal{D}) = \frac{1}{n}\sum_{i=1}^n \ell'(\phi(x_i)^\top w, y_i)\phi(x_i) + 2\lambda w.$$

Since $\mathcal{D}$ and $\mathcal{D}'$ only differs at x_1, we have

$$\begin{aligned}\|\nabla_w J(f_w,\mathcal{D}) - \nabla_w J(f_w,\mathcal{D}')\|_2 =& \|\frac{1}{n}\ell'(\phi(x_1)^\top w, y_1)\phi(x_1) - \frac{1}{n}\ell'(\phi(x_1')^\top w, y_1')\phi(x_1')\|_2 \\ \le& \frac{2c_L\sqrt{\kappa}}{n}\end{aligned}$$

where we used Assumption 1 and 3 in the last step. □

Proof of Lemma 28. We first give explicit expression of a bijection between $G_w(\mathcal{D})$ and $G_w(\mathcal{D}')$ by construction. For each w, let $e_1, \cdots e_h$ be a set of standard orthogonal basis in $\mathbb{R}^h$ satisfying $e_1, e_2 \in \text{span}\{\nabla_w J(f_w, \mathcal{D}), \nabla_w J(f_w, \mathcal{D}')\}$ and $e_3, \cdots e_h \in \text{span}\{\nabla_w J(f_w, \mathcal{D}), \nabla_w J(f_w, \mathcal{D}')\}^\perp$. A rotation matrix U in $\mathbb{R}^h$ satisfies

$$U \in \mathbb{R}^{h\times h}, U^{-1} = U^\top, \det U = 1.$$

Define the rotation matrix $U = U_{w,\mathcal{D},\mathcal{D}'}$ from $\nabla_w J(f_w, \mathcal{D})$ to $\nabla_w J(f_w, \mathcal{D}')$ by

$$U\frac{\nabla_w J(f_w, \mathcal{D})}{\|\nabla_w J(f_w, \mathcal{D})\|_2} = \frac{\nabla_w J(f_w, \mathcal{D}')}{\|\nabla_w J(f_w, \mathcal{D}')\|_2} \text{ and } \quad Ue_i = e_i, \quad \text{for} \quad i = 2, \cdots, h. \tag{22}$$

Condition (22) specifies a unique U for each pair of $w, \mathcal{D}$ and $\mathcal{D}'$. The rotation matrix only operates on the two-dimensional subspace $\text{span}\{\nabla_w J(f_w, \mathcal{D}), \nabla_w J(f_w, \mathcal{D}')\}$. Then we claim that for each $g_{P,b}^{m,h} \in G_w(\mathcal{D})$, we have $g_{P',b'}^{m,h} \in G_w(\mathcal{D}')$ where $P' = UPU^\top$ and $b' = (I_h - P')w$. To see this, recall the formulation of $G_w(\mathcal{D})$ in Definition 14 that

$$P\nabla_w J(f_w, \mathcal{D}) = \mathbf{0} \quad \text{and} \quad (I_h - P)w = b.$$

Then,

$$\begin{aligned} P'\nabla_w J(f_w, \mathcal{D}) = UPU^\top \nabla_w J(f_w, \mathcal{D}') =& UPU^\top U \nabla_w J(f_w, \mathcal{D}) \frac{\|\nabla_w J(f_w, \mathcal{D}')\|}{\|\nabla_w J(f_w, \mathcal{D})\|} \\ =& UP\nabla_w J(f_w, \mathcal{D}) \frac{\|\nabla_w J(f_w, \mathcal{D}')\|}{\|\nabla_w J(f_w, \mathcal{D})\|} = \mathbf{0}. \end{aligned}$$

Also, $b' = (I_h - P')w$ follows from definition and thus $g_{P',b'}^{m,h} \in G_w(\mathcal{D}')$. The map $\mathcal{U}$ is defined as the map from $g_{P,b}^{m,h}$ to $g_{P',b'}^{m,h}$. Note that this is a bijection since U is invertible.

Next, we show that $\|P - P'\|_2 \le 16 c_L \sqrt{\kappa} n^{-1/2}$ for $w \in \Delta_n$. Recall the definition of U in (22), the operator norm of $U - I_h$ satisfies

$$\|U - I_h\|_2 \le \|\big(U - I_h\big)\frac{\nabla_w J(f_w, \mathcal{D})}{\|\nabla_w J(f_w, \mathcal{D})\|_2}\|_2 = \|\frac{\nabla_w J(f_w, \mathcal{D})}{\|\nabla_w J(f_w, \mathcal{D})\|_2} - \frac{\nabla_w J(f_w, \mathcal{D}')}{\|\nabla_w J(f_w, \mathcal{D}')\|_2}\|_2.$$

Note that both $\frac{\nabla_w J(f_w, \mathcal{D})}{\|\nabla_w J(f_w, \mathcal{D})\|_2}$ and $\frac{\nabla_w J(f_w, \mathcal{D}')}{\|\nabla_w J(f_w, \mathcal{D}')\|_2}$ are unit length vector. By the law of cosine, see for instance [Pic09], there holds

$$\|\frac{\nabla_w J(f_w, \mathcal{D})}{\|\nabla_w J(f_w, \mathcal{D})\|_2} - \frac{\nabla_w J(f_w, \mathcal{D}')}{\|\nabla_w J(f_w, \mathcal{D}')\|_2}\|_2 = 2 - 2\frac{\nabla_w J(f_w, \mathcal{D}) \cdot \nabla_w J(f_w, \mathcal{D}')}{\|\nabla_w J(f_w, \mathcal{D})\|_2 \|\nabla_w J(f_w, \mathcal{D}')\|_2}.$$

We can further decompose this as

$$\begin{aligned} & 2 - 2\frac{\nabla_w J(f_w, \mathcal{D}) \cdot \big(\nabla_w J(f_w, \mathcal{D}) + \nabla_w J(f_w, \mathcal{D}') - \nabla_w J(f_w, \mathcal{D})\big)}{\|\nabla_w J(f_w, \mathcal{D})\|_2 \|\nabla_w J(f_w, \mathcal{D}')\|_2} \\ =& 2 - 2\frac{\|\nabla_w J(f_w, \mathcal{D})\|_2}{\|\nabla_w J(f_w, \mathcal{D}')\|_2} - 2\frac{\nabla_w J(f_w, \mathcal{D}) \cdot \big(\nabla_w J(f_w, \mathcal{D}') - \nabla_w J(f_w, \mathcal{D})\big)}{\|\nabla_w J(f_w, \mathcal{D})\|_2 \|\nabla_w J(f_w, \mathcal{D}')\|_2} \end{aligned} \tag{23}$$

By Lemma 27, $\|\nabla_w J(f_w, \mathcal{D}) - \nabla_w J(f_w, \mathcal{D}')\|_2 \leq 2c_L\sqrt{\kappa}/n$. Then we have

$$\begin{aligned}\frac{\|\nabla_w J(f_w, \mathcal{D})\|_2}{\|\nabla_w J(f_w, \mathcal{D}')\|_2} \geq & \frac{\|\nabla_w J(f_w, \mathcal{D}')\|_2 - \|\nabla_w J(f_w, \mathcal{D}) - \nabla_w J(f_w, \mathcal{D}')\|_2}{\|\nabla_w J(f_w, \mathcal{D}')\|_2} \\ = & 1 - \frac{2c_L\sqrt{\kappa}}{n\|\nabla_w J(f_w, \mathcal{D}')\|_2}.\end{aligned}$$

For $w \in \Delta_n$, this yields

$$\frac{\|\nabla_w J(f_w, \mathcal{D})\|_2}{\|\nabla_w J(f_w, \mathcal{D}')\|_2} \geq 1 - \frac{2c_L\sqrt{\kappa}}{\sqrt{n}}. \tag{24}$$

Analogously, we also have

$$\begin{aligned}&\frac{\nabla_w J(f_w, \mathcal{D}) \cdot \big(\nabla_w J(f_w, \mathcal{D}') - \nabla_w J(f_w, \mathcal{D})\big)}{\|\nabla_w J(f_w, \mathcal{D})\|_2 \|\nabla_w J(f_w, \mathcal{D}')\|_2} \\ \geq & -\frac{\|\nabla_w J(f_w, \mathcal{D})\|_2 \|\nabla_w J(f_w, \mathcal{D}') - \nabla_w J(f_w, \mathcal{D})\|}{\|\nabla_w J(f_w, \mathcal{D})\|_2 \|\nabla_w J(f_w, \mathcal{D}')\|_2} \geq -\frac{2c_L\sqrt{\kappa}}{\sqrt{n}}.\end{aligned} \tag{25}$$

Combining (24) and (25), (23) yields

$$\|U - I_h\|_2 \leq \frac{8c_L\sqrt{\kappa}}{\sqrt{n}}.$$

Then, by definition of P', we have

$$\begin{aligned}\|P - P'\|_2 = \|P - UPU^\top\|_2 = \|PU - UP\|_2 \leq & \|PU - P\|_2 + \|P - UP\|_2 \\ \leq & 2\|P\|_2\|U - I_h\|_2 \leq \frac{16c_L\sqrt{\kappa}}{\sqrt{n}}.\end{aligned}$$

□