Abstract

Labeling cost is often expensive and is a fundamental limitation of supervised learning. In this paper, we study importance labeling problem, in which we are given many unlabeled data and select a limited number of data to be labeled from the unlabeled data, and then a learning algorithm is executed on the selected one. We propose a new importance labeling scheme that can effectively select an informative subset of unlabeled data in least squares regression in Reproducing Kernel Hilbert Spaces (RKHS). We analyze the generalization error of gradient descent combined with our labeling scheme and show that the proposed algorithm achieves the optimal rate of convergence in much wider settings and especially gives much better generalization ability in a small label noise setting than the usual uniform sampling scheme. Numerical experiments verify our theoretical findings.

1 Introduction

One of the most popular task in machine learning is supervised learning, in which we estimate a function that maps an input to its label based on finite labeled examples called training data. The goodness of the learned function is measured by the generalization ability, that is roughly the accuracy of the learned function for previously unseen data. Statistical learning theory is a powerful tool which gives a framework for analysing the generalization errors of learning algorithms [Vapnik and Vapnik, 1998]. Enormous learning algorithms have been proposed and their generalization abilities are analysed in various settings.

In spite of the great successes of supervised learning, it has a fundamental limitation due to the expensive cost for making training examples. Particularly, it is often the case that collecting input data is cheap but to give labels of them is limited or expensive and that is one of bottlenecks in supervised learning [Roh et al., 2019]. The dilemma is that the more labeled data, better generalization ability is guaranteed but the higher labeling cost is incurred.

In this limited situation, importance labeling problem naturally arises, which is a special case of active learning [Settles, 2009]. In the importance labeling settings, we first collect many unlabeled examples. Then we choose a limited number of examples to be labeled from unlabeled ones. The most naive selection of labeled examples is based on uniform subsampling from unlabeled data. What we expect here is that if we choose labeled samples effectively, then better generalization ability may be acquired.

Despite the significance of the problem, theoretical aspects of importance labeling is little known. The essential question is what importance labeling scheme surpasses the standard uniform labeling in what settings.

In this paper, we consider this quite general question in the context of least squares regression in Reproducing Kernel Hilbert Spaces (RKHS). Kernel method is classical and promising approach for learning nonlinear functions [Schölkopf et al., 2002]. In kernel method, input data is mapped to an (potentially) infinite dimensional feature space and then a linear predictor on the feature space is learned. The feature space is determined by the user-defined kernel function and numerous kernel functions are known, e.g., classical Gaussian kernel and more modern neural tangent kernel (NTK) [Jacot et al., 2018]. Least squares regression in RKHS has a long history and its generalization ability has been thoroughly studied in supervised learning set-
The comparison of theoretical generalization errors between stopping as implicit regularization has been theoretically less regression has been considered and the effect of early Rosasco and Villa (2015), gradient descent for kernel ridge-regression under suitable assumptions. In Yao et al. (2007); the minimax optimal generalization ability of kernel ridge 2018; Pillaud-Vivien et al., 2018; Jun et al., 2019). Capon- Rosasco, 2017; Lin and Rosasco, 2017; Carratino et al., 2007; Caponnetto and De Vito, 2007; Steinwart et al., 2009;  

Main Contributions

- We propose a new importance labeling scheme called CRED (Contribution Ratios to Effective Dimension), which employs so-called contribution ratio as the importance of each data point so that we can efficiently exploit information of input data. The contribution ratio measures how each data point contributes to the effective dimensionality of RKHS which plays the essential role for characterizing the estimation performance of kernel ridge regression.

- The generalization error of gradient descent on the labeled dataset selected by CRED is theoretically analyzed in the settings of kernel ridge regression. It is shown that our algorithm achieves wider optimality than existing methods in general settings and significantly better generalization ability particularly under low label noise (i.e., near interpolation) settings.

- The algorithm and the theoretical results are extended to random features settings and the potential computational intractability of CRED from infinite dimensionality of RKHS is resolved.

The comparison of theoretical generalization errors between our proposed algorithms with the most relevant existing methods is summarised in Table 1.

Related Work

Here, we briefly overview the most relevant research areas and methods to our work.

Supervised Learning. Supervised least squares regression in RKHS has been thoroughly studied (Yao et al. 2007; Caponnetto and De Vito 2007; Steinwart et al. 2009; Rosasco and Villa 2015; Dieuleveut et al. 2016; Rudi and Rosasco 2017). However, these papers do not consider the utilization of the unlabeled data and hence the derived theoretical generalization ability may be sub-optimal because the uniform labeling never captures the “importance” of each data point. This paper gives a novel sampling scheme from unlabeled data by defining the importance of each data point as the contribution ratio to effective dimension.

Semi-Supervised Learning. Semi-supervised learning has a close relation to importance labeling. In semi-supervised learning, we are given many unlabeled data and small number of labeled data. Typically the labeled data is uniformly selected from unlabeled data. Semi-supervised learning aims to get better generalization ability by the effective use of unlabeled examples typically under so-called cluster assumption (Balcan and Blum 2005; Rigollet 2007; Ben-David et al. 2008; Wasserman and Lafferty 2008). In contrast, the importance labeling scheme in this paper aims to get better generalization ability by the effective choice of labeled examples without the assumption. In Ji et al. (2012), a simple semi-supervised kernel regression algorithm called SSSR has been proposed and they have shown that the generalization ability surpasses the one of supervised learning when the true function is attainable and deterministic. Roughly speaking, the algorithm first computes eigen-system of covariance operator in the feature space using unlabeled data. Then, linear regression is executed on the principle eigen-functions as features. The theory of SSSR does not require the cluster assumption and is on the standard theoretical settings of kernel regression, but the generalization ability may be still sub-optimal.

Active Learning. Active learning is also a close concept to importance labeling. In active learning, we are given learned model on small labeled data and then select new labeled data from unlabeled one by utilizing the information of the learned model. In some sense, active learning is a generalized concept of important labeling. However,
2 Problem Settings and Assumptions

In this section, we provide the formal problem settings in this paper and theoretical assumptions for our analysis.

2.1 Kernel Regression with Importance Labeling

Let $Z_N = \{(x_j, y_j)\}_{j=1}^N$ be i.i.d. samples from some distribution $\rho_Z$, where $z_j = (x_j, y_j) \in X \times Y \subset \mathbb{R}^d \times \mathbb{R}$, and $X_N = \{x_j\}_{j=1}^N$, $Y_N = \{y_j\}_{j=1}^N$. We denote $p_X$ as the marginal distribution of $Z$ on $X$ and $p_{Y|X}$ as the conditional distribution of $Y$ with respect to $X$. We sub-sample $Z_n = \{(x_{j(i)}, y_{j(i)})\}_{i=1}^n (n < N)$ from $Z_N$ according to user-defined distribution $q$ on $Z_N$ and we denote $X_n = \{x_{j(i)}\}_{i=1}^n$, $y_n = \{y_{j(i)}\}_{i=1}^n$.

The objective of this paper is to minimize the excess risk $\mathcal{E}(f) = \inf_{f \in H} \mathcal{E}(f')$ only using the information of labeled observations $Z_n$, where $\mathcal{E}(w) = \int_2 \frac{1}{2} (y - w(x))^2 d\rho_Z (x, y)$ and $H \subset L^2(\rho_X)(\mathbb{R}^d)$ is some Reproducing Kernel Hilbert Space (RKHS) with inner product $\langle \cdot, \cdot \rangle_H : H \times H \to \mathbb{R}$ and kernel $K(\cdot, \cdot) : X \times X \to \mathbb{R}$.

**Notation** We denote by $\| \cdot \|_H$ the norm induced by $\langle \cdot, \cdot \rangle_H$ and $\| \cdot \|_2$ as the Euclidean norm. Let $\Sigma = S^*S : H \to H$ and $L = SS^* : L^2(\rho_X) \to L^2(\rho_X)$, where the operator $S$ is the natural embedding from $H$ to $L^2(\rho_X)$ and $S^*$ is the adjoint operator of $S$. We define $T_3 = T + \lambda I$ for operator $T$. For natural number $m$, We denote $\{1, \ldots, m\}$ by $[m]$. $K_x$ denotes the operator $K(x, \cdot) = K(\cdot, x) : X \to \mathbb{R}$ for $x \in X$. $K_x$ can be regard as a “feature” of input $x$.

2.2 Theoretical Assumptions

We make the following assumptions for our theoretical analysis. These are fairly standard in the literature of statistical learning theory for kernel methods (Steinwart et al., 2009; Dieuleveut et al., 2016; Lin and Rosasco, 2017; Pillaud-Vivien et al., 2018).

**Assumption 1** (Boundedness of feature). For some $\kappa > 0$, $\sup_{x \in \text{supp}(\rho_X)} \| K_x \|_H \leq \kappa$.

**Assumption 2** (Smoothness of true function). There exists $r \in (0, 1]$ such that $f_* = L^2(\rho_X)$ with $\|\phi\|_{L^2(\rho_X)} \leq R$ ($R > 0$). Here $f_*(\cdot) = \int_Y y d\rho_Y$, that is the regression function (or true function).

**Assumption 2** quantifies the complexity of true function $f_*$ in terms of the eigen-system of $L$. It is known that when $r \geq$...
1/2, \( \mathcal{L}'(L^2(\rho_x)) \) becomes a subset of \( H \) and particularly \( r = 1/2 \), it exactly matches to \( H \). Thus, we have \( f_\ast \in H \) whenever \( r \geq 1/2 \). However, when \( r < 1/2 \), generally \( f_\ast \notin H \). As \( r \to 0 \), roughly \( \mathcal{L}'(L^2(\rho_x)) \to L^2(\rho_x) \). This means that \( f_\ast \) can be more complex (or non-smooth) for smaller \( r \).

**Assumption 3** (Polynomial decay of eigenvalues). There exists \( \alpha > 1 \) such that \( \text{Tr}(\Sigma^{1/\alpha}) < \infty \).

Parameter \( \alpha \) characterizes the complexity of feature space \( H \). For larger \( \alpha \), the feature space becomes more simple and particularly when \( \alpha = \infty \), the feature space must have finite dimension. Note that even for feature spaces with finite dimensionality \( d \), discussions of the case \( \alpha < \infty \) are important because \( \text{Tr}(\Sigma^{1/\alpha}) \) can be much smaller than \( \text{Tr}(\Sigma^{1/\infty}) = d \) for some \( \alpha \in (1, \infty) \).

**Assumption 4** (Bounded variance and uniform boundedness of labels). There exists \( \sigma \geq 0 \) and \( M \geq 1 \) such that \( \mathbb{E}(y - f_\ast(x))^2 \leq \sigma^2 \) and \( |y| \leq M \) almost surely.

Generally label noise \( \sigma > 0 \), but we are particularly interested in the case \( \sigma \to 0 \).

### 3 Proposed Algorithm

In this section, first the behind ideas are described and then formal descriptions of the proposed algorithm are given.

**Behind Ideas.** Our proposed importance labeling scheme is based on the contribution ratios to effective dimension which plays the essential role for characterizing the estimation performance of kernel ridge regression [Zhang 2005]. First recall the notion of effective dimension \( \mathcal{N}_\infty(\lambda) = \mathbb{E}_x[\|\Sigma_x^{1/2}K_x\|_H^{-1}] \), that is roughly the mean of the squared Mahalanobis distances of the features if \( \mathbb{E}_x[K_x] = 0 \). The essential intuition of our scheme is that labeling input \( x \) that has a large contribution to effective dimension reduces the estimation variance. To realize this intuition, we construct an importance sampling distribution proportional to \( \|\Sigma_x^{-1/2}K_x\|_H^2 \) on unlabeled data samples. After sampling the data to be labeled, correcting the bias of the empirical risk caused by the importance labeling is needed. This situation is very similar to the one in the well-known importance sampling in the literature of classical Monte Carlo methods.

Next, for supporting the intuition and understanding how our sampling scheme works, we conduct simple synthetic experiments. We focus on a two dimensional feature space in \( \mathbb{R}^2 \). First we generated 100,000 unlabeled samples \( \{(x_1^{(i)}, x_2^{(i)})\}_{i=1}^{100000} \) according to \( X_1 \sim N(0, 1) \) and \( X_2 \sim N(0, 0.01) \) independently. For comparing our scheme with uniform labeling, we labeled 100 data samples from unlabeled one using two sampling scheme independently. Figure 1 shows the comparison of the labeled data by the two schemes. We can see that the data samples labeled by our proposed CRED covers a wider range of areas than uniform labeling. For making sure that CRED reduces the estimation variance, we conducted 1,000 runs of least square regression on randomly labeled 3 data samples using CRED and uniform labeling independently. We set true function \( f_\ast \) to \( f_\ast(x_1, x_2) = x_1 + x_2 \) and added Gaussian noise with mean zero and variance 0.01 for generating labels. Note that for each labeled sample we multiplied the inverse of the labeling probability of the sample to the correspondence loss and corrected the bias of the empirical risk caused by the importance labeling as in the standard importance sampling scheme. Figure 2 shows the comparison of the deviation of the estimated regression coefficients. We can see that CRED in fact significantly reduces the estimation variance.

**Concrete Algorithm.** Our proposed algorithm is illustrated in Algorithm 1. The algorithm consists of two blocks of importance labeling and optimization by gradient descent.

First we select a subset of the unlabeled data using a sam-
we need to correct the bias of the sampling by multiplying Algorithm 1. The proofs are found in Section B of the 
Computational Tractability
δ
poly(log(\cdot))\supplementary material. We use
4 Generalization Error Analysis
For the details, see Section 6.

is generally intractable due to the inapplicability of kernel 
computation of the contribution ratios to effective dimension.
RKHS can be efficiently executed even in infinite dimen-
Next, we run the standard gradient descent to minimize the 
empirical risk estimated by the labeled data, but each loss 
is weighed by the inverse labeling probability to guarantee 
the unbiasedness of the risk. Thus, the gradient of the bias 
corrected risk is used for updating the solution. Concretely, 
since gradient at g with respect to given single observation
(x, y) is \((K_x, y) - y)K_x = (K_x \otimes K_x)y - yK_x\), if the 
sampling probability of \((x, y)\) from \(N\) unlabeled data is \(q\), 
we need to correct the bias by multiplying a factor \(1/(Nq)\) to the 
gradient. Then all the gradient with respect to labeled data is averaged. 
The formal description of this procedure is given in Line 5-6. Note that when the 
labeling distribution is uniform, i.e., \(q = 1/N\), the algorithm 
matches to the standard gradient descent.

Remark (Computational Tractability). Gradient descent on 
RKHS can be efficiently executed even in infinite dimen-
isional feature spaces thanks to kernel trick. However the 
computation of the contribution ratios to effective dimension 
is generally intractable due to the inapplicability of kernel 
\cite{scholkopf2002learning}. This computational problem 
can be avoided by introducing random features technique. 
For the details, see Section B.

4 Generalization Error Analysis

Here, we give the main theoretical results of CRED-GD 
(Algorithm 1). The proofs are found in Section B of the 
 supplementary material. We use \(\bar{O}\) and \(\bar{\Omega}\) notation to hide 
extra \(\text{poly}(\log(n), \log(\delta^{-1}))\) factors for simplicity, where 
\(\delta\) is a confidence parameter for high probability bounds.

Our analysis starts from bias-variance decomposi-
tion \(\|Sg_t - f_\star\|_2^2 \leq 2\|Sf_t - f_\star\|_2^2 + 2\|S(g_t - f_t)\|_2^2\) 
where \(\{f_t\}_{t=1}^\infty\) is the ideal GD path on excess risk, i.e., 
\(f_t = f_{t-1} - \eta(\Sigma f_{t-1} - S^* f_\star) + f_{t-1} - \eta[E_x[K_x \otimes K_x] - E_x[yK_x]]\) with \(f_0 = 0\). The 
first term is called as bias and the second term is called as variance. 
The bias can be bounded by the following Proposition:

Proposition 4.1 (Bias bound, simplified version of Lemma [A.1]). Suppose that Assumptions 7 and 2 hold. Let \(\eta = O(1/\kappa^2)\) be sufficiently small. Then, for any \(t \in \mathbb{N}\),

\[\|Sf_t - f_\star\|_2^2 = O(R^2(\eta t)^{-2r}).\]

Lemma 4.1 shows that the bias converges to 0 as \(t \to \infty\). 
Moreover, the convergence speed is controlled by the smoothness of the true function.

Definition 4.1. We define \(\mathcal{N}_\infty(\lambda) = E_x[\Sigma^{-1/2}_x K_x]_2^2\) and 
\(\mathcal{F}_\infty(\lambda) = \sup_{x \in \mathcal{X}}[\Sigma^{-1/2}_x K_x]_2^2\).

These quantities play the essential roles for characterizing the estimation performance. 
We can bound these quantities as follows:

Lemma 4.2. Suppose that Assumption 4 holds. For any \(\lambda > 0\), \(\mathcal{F}_\infty(\lambda) \leq \kappa^2 \lambda^{-1}\).

Since \(\alpha > 1\), \(\mathcal{N}_\infty(\lambda)\) has a much tighter bound than \(\mathcal{F}_\infty(\lambda)\) for small \(\lambda\).

Now, we bound the second term, that is called as variance, 
using th following proposition:

Proposition 4.3 (Variance bound, simplified version 
of Proposition B.1). Suppose that \(\eta = O(1/\kappa^2)\) be sufficiently small. Let \(t \in \mathbb{N}\), \(\lambda = 1/(\eta t) \geq \lambda_0 = \Omega(\text{Tr}(\Sigma_0^1)/n)^{\alpha}\), 
\(\delta \in (0, 1)\) and \(n \geq \Omega(1 + \text{Tr}(\Sigma_0^1)/\lambda_0^{1-\alpha})\) and \(N \geq \Omega(1 + \kappa^2 \lambda_0^{-1})\). Then there exits event \(A\) with \(P(A) \geq 1 - \delta\) such that 

\[\mathbb{E}\left[\frac{\|S(g_t - f_t)\|_2^2}{n} \mid A\right] = \bar{O}\left(\frac{\sigma^2 + R^2 \lambda^{2r} \mathcal{N}_\infty(\lambda_0) + \lambda^{2r} + r_N}{n}\right),\]

where \(r_N = \mathcal{F}_\infty(\lambda_0)(\sigma^2 + R^2 \lambda^{2r} + (M^2 + \kappa^{4r-2} R^2 + R^2 \lambda^{-1+2r})/n)\) \(\to 0\) as \(N \to \infty\).

Proposition 4.3 shows that the variance diverges to \(\infty\) as \(t \to \infty\) (because \(\mathcal{N}_\infty(\lambda) = O(\lambda^{-1/\alpha}) \to \infty\) as \(t \to \infty\)), 
but is scale to \(1/n\). Thus, for moderate \(t\), the variance 
can still be small.

Remark. Proposition 4.3 is the main novelty of our analysis. In 
\cite{pillaud-vivien2018}, the variance bound of the standard GD is roughly \((\sigma^2 + \mathcal{N}_\infty(\lambda) + \lambda^{2r} \mathcal{F}_\infty(\lambda))/n\) in our settings. In contrast, our bound is roughly \((\sigma^2 + \mathcal{N}_\infty(\lambda) + \lambda^{2r} \mathcal{F}_\infty(\lambda))/n\).
λνr)N∞(λ)/n for λ ≈ λq and sufficiently large N (note that λνr can be ignored because it never dominates the bias term (see Proposition 4.1)). Since N∞(λ) ≤ F∞(λ) always holds, CRED-GD improves the variance bound of the standard GD when σ^2 is small. Later, we discuss the case N∞(λ) ≪ F∞(λ) (see Lemma 4.2 and Section 5).

Remark. In Pillai-Vivien et al. [2018], under Assumption 1 and additional assumption sup_{supp(x)} \|Σ^{1/2}−1/2 Kx\|_F = O(κ_p^2R^2μ) for some κ_p > 0 and μ ∈ [0, 1], the authors have shown that F∞(λ) = O(κ_p^2R^2μλ−r) (Lemma 13 in Pillai-Vivien et al. [2018]), which is a better bound than ours in Lemma 4.2 when μ < 1. However, in the worst case μ = 1 their bound matches ours in Lemma 4.2. For an example of this case, see Section 5.

For balancing the bias and variance term, we introduce a notion of the optimal number of iterations:

Definition 4.2 (Optimal number of iterations). Optimal number of iterations for CRED-GD t∗ = \lceil 1/(sλ) \rceil, where λ is defined as

\[
λ = O \left( \left( \frac{\sigma^2 T r(Σ)}{n} \right) \frac{\sqrt{\kappa + β}}{n} + \left( \frac{R^2 T r(Σ)}{n} \right)^{α/2} + \lambda_N \right)
\]

where \[λ_N = \begin{cases} κ^2σ^2 + κ^2M^2/N + κ^4R^2/\sqrt{nN} + γκ^2R^2/(\sqrt{nN}) + κR/(\sqrt{nN}) \rightarrow 0 as N → ∞. \]

Lemma 4.2 and Proposition 4.3 with λq = λ yields the following main theorem:

Theorem 4.4 (Generalization Error of CRED-GD). Suppose that Assumptions 2 and 4 hold. Let η = Θ(1/κ^2) be sufficiently small and δ ∈ (0, 1). Then setting λq = λ, T = Θ(t∗), there exists event A with P(A) ≥ 1 − δ such that CRED-GD satisfies E \[\|SGt\_t\_f - f\|^2_{L^2(ρ_X)}\] = \[\beta(λ^2_r, φ), \] where λ is defined in Definition 4.2.

From Theorem 4.4 we obtain the following observations:

Wider Optimality. When σ^2 = Θ(1), the generalization error of CRED-GD with sufficiently many unlabeled data becomes the optimal rate n−2r/α/(2r/α + 1). The same rate is also achieved by supervised GD or SGD but under restrictive condition r > (α − 1)/(2α). In contrast, supervised GD or SGD only achieves O(n−2r/α) in our theoretical settings and σ^2 → 0, and thus CRED-GD significantly improves the generalization ability of supervised methods. Semi-supervised method SSSL [Ji et al. 2012] only achieves n−(α−1)/2 when σ^2 = 0 and r = 1/2, which is worse than ours.

Remark (Equivalence to Kernel Ridge Regression with Importance Labeling). Using very similar arguments of our analysis, it can be shown that analytical kernel ridge regression solution (Σ^{−1}_nλ^−1)(S_n^{(q)})*y_n also achieves the generalization error bound in Theorem 4.4 (see Section C of supplementary material). When λ, is extremely small, the analytical solution is computationally cheap than gradient descent and sometimes useful.

5 Sufficient Condition for N∞(λ) ≪ F∞(λ)

In this section, we give a sufficient condition for N∞(λ) ≪ F∞(λ) and its simple example. The proofs are found in Section D of the supplementary material.

Proposition 5.1. Let \{(λi, φ_i)\}_{i=1} (d ∈ N \cup \{∞\}) be the eigen-system of Σ in L^2(ρ_X), where λ_1 ≥ λ_2 ≥ . . . ≥ 0. Assume that λ_i = Θ(i^−α) and ||φ_i||_L^2(ρ_X) = Ω(i^p/2) for any i for some α = 1 + Ω(1) and p ≥ 1. Moreover if d = ∞, we additionally assume ||φ_i||_L^2(ρ_X) = O((i^α)^−1−ε) for any i for some ε > 0. Then Assumption 1 is satisfied and for any λ ∈ (0, 1), F∞(λ) = Ω(λ^−(p+α) \wedge d^p).

Example. Let X = [−1, 1]^d and ρ_X = TN(0, σ_X^2) ⊗ . . . ⊗ TN(xd, σ_X^2), that is, the product measure of truncated normal distributions with mean 0 and scale parameter σ^2, i.e., independent normal distributions with mean 0 and variance σ^2 conditioned on [−1, 1]. Let σ^2 ≥ . . . ≥ σ^2. We denote σ_i as the variance of TN(0, σ^2_i) for i ∈ [d]. Note that for sufficiently small σ^2_i = Θ(1), we have \[σ_i^2 = Θ(σ^2_i) \] for any i ∈ [d]. Then we particularly consider linear kernel K and thus H = X. Since the covariance matrix is Σ = diag(σ^2_1, . . ., σ^2_d), the eigen-system of Σ in L^2([−1, 1]^d) is \{(σ^2_i, e_i/√(σ^2_i))^d, i∈[d] \} = 1, where e_i(x) = x_i for x ∈ [−1, 1]^d. Suppose that the polynomial decay of σ^2_i^d holds: σ^2_i = Θ(σ^2_i) = Θ(i^−α). Then from Lemma 4.2, N∞(λ) = O((log(d)/λ^−1/α) \wedge d). On the other hand, from Proposition 5.1 with p = α, we have F∞(λ) = Ω(λ^−1 \wedge d^α).

6 Extension to Random Features Settings

In this section, we discuss the application of random features technique to Algorithm 1 for computational tractability. Then we theoretically analyse the generalization error of the algorithm. The proofs are given in Section E of the supplementary material.

Suppose that kernel K has an integral representation K(x, x′) = \int w(x, ω)ψ(x, ω)ψ(x′, ω) dx for some ψ. Random features φ_m,x ∈ R^m is defined by \[m^{-1/2}\psi(x, ω_1), . . . , \psi(x, ω_m) \] , where ω_1, . . ., ω_m ∼ ω independent, is used for an approximation of K(x, x′) < φ_m,x, φ_m,x′ >. Here, the number of random features m ∈ N is a user-defined parameter and characterizes the goodness of the approximation. More details and concrete examples
of random features are found in Rudi and Rosasco (2017).

**Algorithm.** The random features version of CRED-GD is illustrated in Algorithm 2. The difference from Algorithm 1 is only the replacement of $K_x$ to random features $\phi_{x,\omega}$. Note that we can properly compute important labeling distribution $q$ using standard SVD solvers thanks to the finite dimensionality of the random features.

We need the following additional assumption about the boundedness of the random features for theoretical analysis:

**Assumption 5.** $\sup_{x \in \text{supp}(\rho_x), \omega \in \text{supp}(\pi)} |\psi(x, \omega)| \leq \kappa$ for some $\kappa > 0$.

For example, random features of Gaussian kernel satisfies this assumption (Rudi and Rosasco 2017).

We define $\hat{S} : \mathbb{R}^m \to L^2(\rho_x)$ by $\langle \hat{S}f, x \rangle = \langle \phi_{x,\omega}, f \rangle$ and $\hat{S}^* \text{ by the adjoint of } \hat{S}$. Then we denote $\hat{\Sigma} = \hat{S}^* \hat{S}$ and $L = \hat{S} \hat{S}^*$.

**Generalization Error Analysis.** We consider generalization error $\|\hat{S}f_t - f^*_x\|_{L^2(\rho_x)}^2$. We decompose the generalization error to bias and variance $\|\hat{S}f_t - f_x\|_{L^2(\rho_x)}^2 \leq 2\|\hat{S}f_t - \hat{S}f_x\|_{L^2(\rho_x)}^2 + 2\|\hat{S}f_t - \hat{S}f_t\|_{L^2(\rho_x)}^2$, where $\{f_t\}_{t=1}^\infty$ is the ideal path of GD with RF on excess risk, i.e., $f_t = f_t - \eta(\hat{S}f_{t-1} - \hat{S}^*f_x) = f_{t-1} - \eta(\mathbb{E}_x[\phi_{x,\omega}\phi_{x,\omega}^\top] - \mathbb{E}_{x,y}[\phi_{x,\omega}] - \mathbb{E}_{x,y}[\phi_{x,\omega}])$ with $f_0 = 0$. The bias term can be bounded similar to Proposition 4.1.

**Proposition 6.1 (Bias bound for RF setting, simplified version of Lemma 6.1).** Suppose that Assumptions 2 and 5 hold. Let $\eta = O(1/\kappa^2)$ be sufficiently small and $t \in \mathbb{N}$ such that $m = \Omega(1 + \kappa^2 \eta t)$. Then for any $\delta > 0$, with probability at least $1 - \delta$,

$$\|\hat{S}f_t - f_x\|_{L^2(\rho_x)}^2 = O(R^2(\eta t)^{-2r}).$$

**Remark.** Compared to Lemma 4.1, additional condition $m = \Omega(1 + \kappa^2 \eta t)$ is assumed. This implies that to make bias small, appropriately large number of random features $m$ is required.

The variance conditioned on random features $\{\omega_k\}_{k=1}^m$ can be bounded in a perfectly similar manner to the proof of Proposition 4.3 with replacing $\mathcal{N}_\infty(\lambda Q)$ and $\mathcal{F}_\infty(\lambda)$ by random features approximations $\mathcal{N}_\infty(\lambda q)$ and $\mathcal{F}_\infty(\lambda)$ respectively. $\mathcal{F}_\infty(\lambda)$ has a trivial bound $O(\lambda^{-1})$. The key lemma for bounding $\mathcal{N}_\infty(\lambda q)$ is the following:

**Lemma 6.2 (Proposition 10 in Rudi and Rosasco (2017)).** Suppose that Assumption 5 holds. We denote $\mathcal{N}_\infty(\lambda q) = \mathbb{E}_x[\Sigma_{\lambda q}^{-1/2} \phi_{x,\omega}]_2^2$ for $\lambda > 0$. For any $\delta \in (0, 1)$ and sufficiently small $\lambda = O(1)$, if $m = \Omega(1 + \kappa^2 \lambda^{-1})$, with probability at least $1 - \delta$ it holds that $\mathcal{N}_\infty(\lambda q) \leq 1.55 \mathcal{N}_\infty(\lambda)$.

Combining the bias and variance bounds with Lemma 6.2 yields the following theorem:

**Theorem 6.3 (Generalization error of CRED-GD with RF, simplified version of Theorem 6.3).** Suppose that Assumptions 2, 3, 4, and 5 hold. Let $\eta = O(1/\kappa^2)$ be sufficiently small, $\lambda_q = \lambda$, and $T = \Theta(t_0^2)$. For any $\delta \in (0, 1)$, if $m \geq \Omega(1 + \kappa^2 \lambda^{-1})$, there exists event $A$ with $P(A) \geq 1 - \delta$ such that RF-CRED-GD has the same generalization error bounds as CRED-GD in Theorem 4.4.

Theorem 6.3 ensures that Algorithm 2 achieves still the same generalization ability as Algorithm 1 when the number of random features $m$ is sufficiently large.

## 7 Numerical Experiments

In this section, numerical results are provided to empirically verify our theoretical findings.

**Experimental Settings.** In our experiments, the input data of public datasets MNIST and Fashion MNIST (Xiao et al. 2017) were used. First we randomly split each dataset into train (60,000) and test (10,000) and normalized input data by dividing 255. We conducted both linear regression (LR) and nonlinear regression (NLR) tasks. For linear tasks, we used the original inputs with bias as features. For nonlinear tasks, we used a randomly initialized three hidden layered fully connected ReLU network with width 500 without output layer as features. Here, the random weights were from i.i.d. standard normal distributions. Then we randomly generated true linear function on the feature spaces, whose regression coefficients were defined by $\sum a_i e_i / \sqrt{\lambda_i} \in \mathbb{R}^{500}$, where $a_i \sim N(0, 1)$ and $\{(e_i, \lambda_i)\}$ was the eigen-system of the covariance matrix in the correspondence feature space.
Finally, we generated noised labels based on them, where the noises were from i.i.d. normal distributions with mean 0 and variance $\sigma^2 \in \{10^{-6}, 10^{-4}, 10^{-2}, 1, 10^2\}$. We compared our proposed method with KRR (Kernel Ridge Regression), KTR$^3$ (Jun et al., 2019) and SSSR (Ji et al., 2012). The hyper-parameters were fairly and reasonably determined. The train data was used as unlabeled data and the labeled data was selected from it. The number of labeled data was ranged in $\{1000, 2000, 4000\}$. We independently ran each experiment five times and recorded the median of test RMSE on each setting.

**Results** Figure 3 shows the comparisons of test RMSE of our proposed method with previous methods. From these results, we make the following observations:

- When the label noise $\sigma^2$ was large, all the algorithms have similar performances.
- When the label noise $\sigma^2$ was small, CRED significantly outperformed the other methods overall. SSSR was always comparable to or better than KRR and KTR$^3$, but sometimes significantly worse than CRED.

These observations can be well-explained by the theoretical results that show our proposed CRED achieves much better generalization ability than the other methods when $\sigma^2 \to 0$ as described in Table 1.
Conclusion and Future Work

In this paper, we proposed a new importance labeling scheme called CRED, which employs the contribution ratio to the effective dimension of the feature space as the importance of each data point. The generalization error of GD with CRED was theoretically analysed and much better bound than previous methods was derived when label noise is small. Further, the algorithm and analysis were extended to random features settings and computational intractability of CRED was resolved. Finally, we provided numerical comparisons with existing methods. The numerical results showed empirical superiority to the other methods and verified our theoretical findings.

One direction of future work would be an application of our importance labeling idea to deep learning. Since the feature space of a deep neural network is updated in training time, our importance labeling scheme can be naturally extended to active learning settings. The theoretical and empirical study of the application to active learning of deep neural networks is a promising future work.

Acknowledgement

TS was partially supported by JSPS KAKENHI (18K19793, 18H03201, and 20H00576), Japan DigitalDesign, and JST CREST.

References

G. Alain, A. Lamb, C. Sankar, A. Courville, and Y. Bengio. Variance reduction in sgd by distributed importance sampling. *arXiv preprint arXiv:1511.06481*, 2015.

M.-F. Balcan and A. Blum. A pac-style model for learning from labeled and unlabeled data. In *International Conference on Computational Learning Theory*, pages 111–126. Springer, 2005.

S. Ben-David, T. Lu, and D. Pál. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *COLT*, pages 33–44, 2008.

K. Brinker. Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 59–66, 2003.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics, 7*(3):331–368, 2007.

L. Carratino, A. Rudi, and L. Rosasco. Learning with sgd and random features. In *Advances in Neural Information Processing Systems*, pages 10192–10203, 2018.

B. Chen, Y. Xu, and A. Shrivastava. Fast and accurate stochastic gradient estimation. In *Advances in Neural Information Processing Systems*, pages 12339–12349, 2019.

D. Csiba and P. Richtárik. Importance sampling for minibatches. *The Journal of Machine Learning Research, 19*(1):962–982, 2018.

S. Dasgupta. Analysis of a greedy active learning strategy. In *Advances in neural information processing systems*, pages 337–344, 2005.

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

Y. Gal, R. Islam, and Z. Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1183–1192. JMLR. org, 2017.

Y. Guo and D. Schuurmans. Discriminative batch mode active learning. In *Advances in neural information processing systems*, pages 593–600, 2008.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

M. Ji, T. Yang, B. Lin, R. Jin, and J. Han. A simple algorithm for semi-supervised learning with improved generalization error bound. *arXiv preprint arXiv:1206.6412*, 2012.

K.-S. Jun, A. Cutkosky, and F. Orabona. Kernel truncated randomized ridge regression: Optimal rates and low noise acceleration. In *Advances in Neural Information Processing Systems*, pages 15332–15341, 2019.

A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with gaussian processes for object categorization. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

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

D. Needell, R. Ward, and N. Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in neural information processing systems*, pages 1017–1025, 2014.

L. Pillaud-Vivien, A. Rudi, and F. Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, pages 8114–8124, 2018.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

P. Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research, 8*(Jul):1369–1392, 2007.
Y. Roh, G. Heo, and S. E. Whang. A survey on data collection for machine learning: a big data-ai integration perspective. IEEE Transactions on Knowledge and Data Engineering, 2019.

L. Rosasco and S. Villa. Learning with incremental iterative regularization. In Advances in Neural Information Processing Systems, pages 1630–1638, 2015.

A. Rudi and L. Rosasco. Generalization properties of learning with random features. In Advances in Neural Information Processing Systems, pages 3215–3225, 2017.

B. Schölkopf, A. J. Smola, F. Bach, et al. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2002.

O. Sener and S. Savarese. Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489, 2017.

B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

I. Steinwart, D. R. Hush, C. Scovel, et al. Optimal rates for regularized least squares regression. In COLT, pages 79–93, 2009.

V. Vapnik and V. Vapnik. Statistical learning theory wiley. New York, 1, 1998.

L. Wasserman and J. D. Lafferty. Statistical analysis of semi-supervised regression. In Advances in Neural Information Processing Systems, pages 801–808, 2008.

K. Wei, R. Iyer, and J. Bilmes. Submodularity in data subset selection and active learning. In International Conference on Machine Learning, pages 1954–1963, 2015.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017.

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. Constructive Approximation, 26(2):289–315, 2007.

K. Yu, J. Bi, and V. Tresp. Active learning via transductive experimental design. In Proceedings of the 23rd international conference on Machine learning, pages 1081–1088, 2006.

T. Zhang. Learning bounds for kernel regression using effective data dimensionality. Neural Computation, 17 (9):2077–2098, 2005.

P. Zhao and T. Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In international conference on machine learning, pages 1–9, 2015.
A Auxiliary Results

First we introduce GD path on the excess risk:

\[ f_t = f_{t-1} - \eta \nabla E(f_{t-1}) \]

\[ = f_{t-1} - \eta \int_Z ((f_{t-1}, K_x) H - y) K_x d\rho_Z(x, y) \]

\[ = f_{t-1} - \eta (\Sigma_{t-1} - S^* f_s) \]

with \( f_0 = 0 \in H \) for \( t \in \mathbb{N} \).

**Lemma A.1** (Proposition 2 and Extension of Lemma 16 in Lin and Rosasco (2017)). Suppose that Assumptions 1 and 2 hold. Let \( \eta = O(1/\kappa^2) \) be sufficiently small. Then, for any \( t \in \mathbb{N} \),

\[ \|Sf_t - f_s\|_{L^2(\rho_X)}^2 = O \left( R^2 \left( \eta t \right)^{-2r} \right). \]

Moreover for any \( \lambda > 0 \) and \( s \in [0, r] \)

\[ \|\Sigma^{-s}_\lambda f_t\|_H^2 \leq O \left( R^2 \left( \kappa^4 (r-s)^2 + (\eta t)^{1-2(r-s)} \right) \right). \]

**Proof.** The second statement is a straightforward extension of Lemma 16 in Lin and Rosasco (2017) and we omit it. \( \square \)

**Lemma A.2.** Suppose that Assumptions 2 and 3 hold. For any \( \lambda > 0 \),

\[ N_\infty(\lambda) \overset{\text{def}}{=} E_x \|\Sigma^{-1/2}_\lambda K_x\|_H^2 = \text{Tr}(\Sigma^{-1} \Sigma) \leq \text{Tr}(\Sigma^{-1} \lambda) \lambda^{-\frac{2}{\gamma}}. \]

**Proof.** \( E_x \|\Sigma^{-1/2}_\lambda K_x\|_H^2 = E_x \text{Tr}(\Sigma^{-1/2}_\lambda K_x \Sigma^{-1/2}_\lambda) = \text{Tr}(\Sigma^{-1} \Sigma). \) Observe that \( \text{Tr}(\Sigma^{-1} \Sigma) = \sum_{i=1}^{\infty} (\lambda_i/(\lambda_i + \lambda))^{1/\alpha} \leq \sum_{i=1}^{\infty} (\lambda_i/(\lambda_i + \lambda))^{1/\alpha} \leq \text{Tr}(\Sigma^{-1} \Sigma) \lambda^{-1/\alpha} \). This finishes the proof. \( \square \)

**Lemma A.3.** Suppose that Assumption 4 holds. For any \( \lambda > 0 \),

\[ F_\infty(\lambda) \overset{\text{def}}{=} \sup_{x \in \text{supp}(\rho_X)} \|\Sigma^{-1/2}_\lambda K_x\|_H^2 = \kappa^2 \lambda^{-1}. \]

**Proof.** From Assumptions 1 we immediately obtain the claim. \( \square \)

**Lemma A.4** (Spectral filters). Let \( p_t(x) = \eta \sum_{k=0}^{t} (1 - \eta x)^k \) for \( x \in [0, 1/\eta) \) and \( t \in \mathbb{N} \). Also we define \( r_t(x) = 1 - xp_t(x) \) \( x \in [0, 1/\eta) \) and \( t \in \mathbb{N} \). Then the following inequalities hold:

\[ \sup_{x \in [0, 1/\eta]} p_t(x) x \leq O(1) \]

for any \( t \in \mathbb{N} \) and

\[ \sup_{x \in [0, 1/\eta]} r_t(x) x^u \leq O(1)(\eta t)^{-u} \]

for any \( t \in \mathbb{N} \) and \( u \in [0, 1] \).

**Proof.** When \( x = 0 \), the inequalities always hold and so we assume \( x > 0 \). Note that \( p_t(x) = (1 - (1 - \eta x)^t)/x \). The first inequality is trivial because \( 1 - (1 - \eta x)^t \leq 1 \). We show the second inequality. Note that \( r_t(x) = (1 - \eta x)^t \). Observe that from elemental calculus, function \( (1 - \eta x)^t x^u \) for \( x \in (0, 1/\eta) \) is maximized at \( x = u/(\eta(u+t)) \) and has maximum value \( u^u/(\eta^u(u+t)^u)(1-u/(u+t))^t \leq u^u(\eta t)^{-u} \). This finishes the proof. \( \square \)

Recall that

\[ q_j = \frac{\|\Sigma^{-\frac{1}{2}_{N\lambda_q}} K_{x_j}\|_H^2 + \frac{1}{N} \sum_{i=1}^{N} \|\Sigma^{-\frac{1}{2}_{N\lambda_q}} K_{x_i}\|_H^2}{2 \sum_{j=1}^{N} \|\Sigma^{-\frac{1}{2}_{N\lambda_q}} K_{x_j}\|_H^2} \]

for \( j \in [N] \). \( \lambda_q \) will be set to \( \lambda_{q+} \), where \( 1/(\eta \lambda_{q+}) \) is the optimal number of iterations (see Definition 4.2 in the main paper). Then we define \( \Sigma^{(q)}_n = (1/n) \sum_{i=1}^{n} 1/(Nq_{j(i)}) K_{x_{j(i)}} \otimes K_{x_{j(i)}} \), where \( j(i) \) is uniformly random on \([N]\).
Lemma A.5. Suppose that Assumption [7] Let \( s \in [0, 1/2) \) and \( \delta \in (0, 1] \). Suppose that \( \lambda = \Omega((\text{Tr}(\Sigma^{1/\alpha})/n)^\alpha) \). When \( N = \Omega(1 + \kappa^2\lambda^{-1}\log(n/\delta)) \), with probability at least \( 1 - \delta \)

\[
\left\| \Sigma^{1/2-s}\Sigma_{N,\lambda}^{-1/2} \right\| = O(\lambda^{-s})
\]

holds.

Proof. For \( \lambda > \|\Sigma\| \), the claim is trivial. Note that since \( s \in [0, 1/2) \), \( \|\Sigma^{1/2-s}\Sigma_{N,\lambda}^{-1/2}\| \leq \|\Sigma^{1/2-s}\Sigma_{N,\lambda}^{-1/2}\| \leq \lambda^{-s}(1 + \|\Sigma^{1/2}\|) \). Then from Proposition 8 in [Rudi and Rosasco, 2017], we have \( \|\Sigma^{1/2}\Sigma_{N,\lambda}^{-1/2}\| \leq \{1 - \lambda\max(\Sigma^{-1/2}(\Sigma^{1/2} - \Sigma_{N})\Sigma_{N}^{-1/2})\}^{-1/2} \). Now from Proposition 6 in [Rudi and Rosasco, 2017], we have with probability \( 1 - \delta \),

\[
\lambda_{\max}(\Sigma_{N}^{-1/2}(\Sigma - \Sigma_{N})\Sigma_{N}^{-1/2}) = O(1)
\]

for any \( \lambda \in (0, \|\Sigma\|) \). Assume \( \lambda = \Omega((\text{Tr}(\Sigma^{1/\alpha})/n)^\alpha) \). By using Lemma A.3, we can see that \( \text{r.h.s of (1)} \) becomes smaller than 0.5 when \( N = \Omega(1 + \kappa^2\lambda^{-1}\log(n/\delta)) \). Then we obtain the desired result. \( \square \)

Lemma A.6. Suppose that Assumptions [7] and [8] hold. Let \( \delta \in (0, 1) \) and \( \lambda \geq \lambda_q = \Omega((\text{Tr}(\Sigma^{1/\alpha})/n)^\alpha) \). When \( n = \Omega(1 + \text{Tr}(\Sigma^{-1/\alpha})\lambda_q^{-1/\alpha}\log^2(n/\delta)) \) and \( N = \Omega(1 + \kappa^2\lambda_q^{-1}\log(n/\delta)) \), with probability \( 1 - 3\delta \)

\[
\left\| \Sigma_{N}^{-1/2}(\Sigma_{n,\lambda}^{(q)})^{-1/2} \right\| = O(1)
\]

holds.

Proof. The proof is similar to the one of Lemma A.5. Suppose that \( X_N \) is given. At first, for \( \lambda > \|\Sigma_N\| \), the claim is trivial. Next, from Proposition 8 in [Rudi and Rosasco, 2017], we have \( \|\Sigma_{N}^{-1/2}(\Sigma_{n,\lambda}^{(q)})^{-1/2}\| \leq \{1 - \lambda\max(\Sigma^{-1/2}(\Sigma_{N} - \Sigma_{n,\lambda}^{(q)})\Sigma_{n,\lambda}^{-1/2})\}^{-1/2} \). Recall that \( \mathbb{E}[\Sigma_{n,\lambda}^{(q)}|X_N] = \Sigma_N \). Observe that

\[
\left\| \frac{1}{\sqrt{Nq_{j(i)}}} K_{x_j(i)} \right\|_H^2 = \frac{1}{Nq_{j(i)}} \left\| \Sigma_{N,\lambda}^{-1/2} K_{x_j(i)} \right\|_H^2 \leq \frac{1}{Nq_{j(i)}} \left\| \Sigma_{N,\lambda}^{-1/2} K_{x_j(i)} \right\|_H^2 \leq \frac{2}{N} \sum_{j=1}^{N} \left\| \Sigma_{N,\lambda}^{-1/2} K_{x_j} \right\|_H^2 \leq \frac{2}{N} \sum_{j=1}^{N} \left\| \Sigma_{N,\lambda}^{-1/2} K_{x_j} \right\|_H^2 = O(1) \frac{1}{N} \sum_{j=1}^{N} \left\| \Sigma_{N,\lambda}^{-1/2} K_{x_j} \right\|_H^2 \]

\[\text{Proposition 6 in [Rudi and Rosasco, 2017], the logarithmic factor in (1) is replaced with } \log((\text{Tr}(\Sigma^{1/\alpha})/n)^\alpha) \text{. This is due to loose bound } \text{Tr}(\Sigma^{1/\alpha}) \leq \text{Tr}(\Sigma^{1/\alpha}) \text{ in their proof and we can improve the bound to } \text{Tr}(\Sigma^{1/\alpha}) \leq \text{Tr}(\Sigma^{1/\alpha}) \text{ from Lemma A.2. This improvement is important for extremely small } \lambda \text{. For example, when } \lambda = \text{Tr}(\Sigma^{1/\alpha})/n^{\alpha}, \text{ that is the lower bound of } \lambda \text{ in our theory, the loose log factor becomes } \alpha \log(n) \text{ rather than } \log(n), \text{ which goes to } \infty \text{ as } \alpha \to \infty.\]
where the second inequality holds from $\lambda \geq \lambda_q$ and the last inequality holds from Lemma A.5. Similar to the arguments in the proof of Lemma A.5 for any $\lambda \in (0, \|\Sigma_N\|)$, we have with probability at least $1 - 2\delta$,

$$\lambda_{\text{max}} \left( \Sigma_{\lambda}^{-\frac{1}{2}} (\Sigma_N - \Sigma_{n,\lambda}) \Sigma_{\lambda}^{-\frac{1}{2}} \right) = O(1) \left( \frac{\log(\text{Tr}(\Sigma_{\lambda}^{\frac{1}{2}})\lambda^{-\frac{1}{2}}\delta^{-1})}{n} \sum_{j=1}^{N} \|\Sigma_{\lambda_q}^{-\frac{1}{2}} K_{x_j}\|^2 + \frac{\log(\text{Tr}(\Sigma_{\lambda}^{\frac{1}{2}})\lambda^{-\frac{1}{2}}\delta^{-1})}{n} \right)$$

given $\{x_j\}_{j=1}^{N}$. Here we used that fact that $\text{Tr}(\Sigma_{\lambda}^{-1}\Sigma_N) = O(1)\text{Tr}(\Sigma_{\lambda}^{-1}\Sigma)$ with probability at least $1 - \delta$ if $N = \Omega(1 + \kappa^2 \lambda^{-1} \log(n/\delta))$ from the similar results to Proposition 10 in [Rudi and Rosasco, 2017].

Then, using standard Bernstein’s inequality for i.i.d. random variables $\{\|\Sigma_{\lambda}^{-\frac{1}{2}} K_{x_j}\|^2\}_{j=1}^{N}$, we have

$$\frac{1}{N} \sum_{j=1}^{N} \left\| \Sigma_{\lambda_q}^{-\frac{1}{2}} K_{x_j} \right\|^2 = O(1) \left( N_{\infty}(\lambda_q) + \frac{\log(\delta^{-1})N_{\infty}(\lambda_q)F_{\infty}(\lambda_q)}{N} + \frac{\log(\delta^{-1})F_{\infty}(\lambda_q)}{N} \right)$$

$$= O(1) \left( \text{Tr}(\Sigma_{\lambda}^{\frac{1}{2}})\lambda^{-\frac{1}{2}} + \frac{\log(\delta^{-1})\text{Tr}(\Sigma_{\lambda}^{\frac{1}{2}})\kappa^2 \lambda^{-\frac{1}{2}}}{N} + \frac{\log(\delta^{-1})\kappa^2 \lambda^{-1}}{N} \right)$$

$$= O(1) \left( \text{Tr}(\Sigma_{\lambda}^{\frac{1}{2}})\lambda^{-\frac{1}{2}}\log(\delta^{-1}) + \frac{\kappa^2 \lambda^{-1}\log(\delta^{-1})}{n} \right)$$

with probability at least $1 - \delta$. For the second inequality, we used Lemma A.2 and A.3. The last inequality holds due to inequality of arithmetic and geometric means.

Combining all the results, with probability at least $1 - 3\delta$, when $n = \Omega(1 + \text{Tr}(\Sigma_{\lambda}^{-1/\alpha})\lambda_q^{-1/\alpha}\log^2(n/\delta))$ and $N = \Omega(1 + \kappa^2 \lambda^{-1} \log(n/\delta))$, we obtain the claim of Lemma A.6.

The following lemma is essential for our analysis:

**Lemma A.7.** Suppose that Assumptions 1 and 4 hold. For $\delta \in (0, 1)$, $\lambda \geq \lambda_q$ and $t \in \mathbb{N}$, there exists event $A$ such that

$$\mathbb{E} \left[ \left\| \Sigma_{\lambda}^{-\frac{1}{2}} (\Sigma_n^{\frac{1}{2}} f_t - (S_n^{\frac{1}{2}})^* y_n - (\Sigma_{\lambda}^{-1/\alpha}) f_t - S^* f_s) \right\|_H^2 \right | A] = \tilde{O} \left( \frac{\text{Tr}(\Sigma_{\lambda}^{\frac{1}{2}})(\sigma^2 + R^2(\eta)^{-2r})(\lambda_q^{-\frac{1}{2}})}{n^2} + \frac{\kappa^2 \lambda^{-1}}{n^2N} \left( \sigma^2 + R^2(\eta)^{-2r} + \frac{M^2 + \kappa^{4r-2}R^2 + R^2(\eta)^{1-2r}}{N} \right) \right).$$

Here $\tilde{O}$ hides extra $\text{poly}(\delta^{-1})$ factors.

**Proof.** Let $\xi_i = \Sigma_{\lambda}^{-1/2} / (N q_{ij(i)}) (K_{x(i)} \otimes K_{x(i)} f_t - y_j(i) K_{x(i)}) = \Sigma_{\lambda}^{-1/2} / (N q_{ij(i)}) K_{x(i)} (f_t(x_j(i)) - y_j(i))$ for $i \in [n]$. Since $\{\xi_i\}_{i=1}^{N}$ is i.i.d. sequence and $\mathbb{E}[\xi_i | X_N] = \Sigma_{\lambda}^{-1/2} (\Sigma_N f_t - S^*_N f_s)$, we have

$$\mathbb{E} \left[ \left\| \Sigma_{\lambda}^{-\frac{1}{2}} (\Sigma_n^{\frac{1}{2}} f_t - (S_n^{\frac{1}{2}})^* y_n - (\Sigma_{\lambda}^{-1/\alpha}) f_t - S^* f_s) \right\|_H^2 \right] = \mathbb{E} \left[ \left\| \Sigma_{\lambda}^{-\frac{1}{2}} (\Sigma_n^{\frac{1}{2}} f_k - (S_n^{\frac{1}{2}})^* y_n - (\Sigma_{\lambda}^{-1/\alpha}) f_t - S^* f_s) \right\|_H^2 \right]$$

$$+ \mathbb{E} \left[ \left\| \Sigma_{\lambda}^{-\frac{1}{2}} (\Sigma_N f_k - S_N^* y_N) - (\Sigma_{\lambda}^{-1/\alpha}) f_t - S^* f_s) \right\|_H^2 \right]$$

$$\leq \frac{1}{n} \mathbb{E}[\xi_i^2]_H + \mathbb{E}[\Sigma_{\lambda}^{-\frac{1}{2}} (\Sigma_N f_t - S_N^* y_N - (\Sigma_{\lambda}^{-1/\alpha}) f_t - S^* f_s)]_H^2.$$
Observe that

\[ E[\|\xi_i\|_H^2 | X_N] = \left( \frac{1}{N} \sum_{i=1}^N \|\Sigma_{\lambda_q}^{-\frac{1}{2}} K_{x_i}\|_H^2 \right) \left( \frac{1}{N} \sum_{i=1}^N (y_i - f_t(x_i))^2 \right) \]

from the definition of \( q \) and Lemma [A.6] and [A.5]. Hence we have

\[ E[\|\xi_i\|_H^2] = E \left[ \left( \frac{1}{N} \sum_{i=1}^N \|\Sigma_{\lambda_q}^{-\frac{1}{2}} K_{x_i}\|_H^2 \right) \left( \frac{1}{N} \sum_{i=1}^N (y_i - f_t(x_i))^2 \right) \right] . \]

Now, similar to the arguments in the proof of Lemma [A.6] since

\[ \frac{1}{N} \sum_{j=1}^N \|\Sigma_{\lambda_q}^{-\frac{1}{2}} K_{x_j}\|^2 \leq \kappa^2 \lambda_q^{-1} \]

a.s. from Lemma [A.3], we have

\[ \frac{1}{N} \sum_{j=1}^N \|\Sigma_{\lambda_q}^{-\frac{1}{2}} K_{x_j}\|^2 \leq O \left( \text{Tr}(\Sigma^{\frac{1}{2}}) \lambda_q^{-\frac{1}{2}} \log(\delta^{-1}) + \frac{\kappa^2 \lambda_q^{-1} \log(\delta^{-1})}{N} \right) \quad (2) \]

with probability at least \( 1 - \delta \). Also, from Bernstein’s inequality and Assumption [4] with probability at least \( 1 - 1/\delta \) it holds that

\[ \frac{1}{N} \sum_{j=1}^N (y_i - f_t(x_i))^2 \leq O \left( \sigma^2 + R^2(\eta t)^{-2r} \log(\delta^{-1}) \right. \]

\[ \left. + \frac{M^2 + \kappa^2 R^2 + R^2(\eta t)^{1 - 2r} \log(\delta^{-1})}{N} \right) \quad (3) \]

because

\[ E_{(x_i, y_i)}[(y_i - f_t(x_i))^2] = \sigma^2, \]

\[ E_{x_i}[(f_t(x_i) - f_t(x_i))^2] \leq O(R^2(\eta t)^{-2r}), \]

\[ (y_i - f_t(x_i))^2 = M^2 \]

and

\[ (f_t(x_i) - f_t(x_i))^2 \leq O \left( M^2 + \|f_{k-1}\|_H^2 \right) \leq O \left( M^2 + \kappa^{4r-2} \kappa^{4r-2} + R^2(\eta t)^{1 - 2r} \right) \]

a.s. from Lemma [A.1].

Denote the event \( A \) that satisfies [2] and [3]. Then \( P(A) \geq 1 - 2\delta \) and

\[ E[\|\xi_i\|^2 | A] \leq O(\log^2(\delta^{-1})) \left( \text{Tr}(\Sigma^{\frac{1}{2}}) \lambda_q^{-\frac{1}{2}} + \frac{\lambda_q^{-1}}{N} \right) \left( \sigma^2 + (\eta k)^{-2r} + \frac{(M^2 + (\eta k)^{1 - 2r})}{N} \right) . \]

Finally, \( E[\Sigma_{\lambda_q}^{-1/2}(\Sigma N f_t - S_N y_N - (\Sigma f_t - S^* f_t))_H^2] \) can be bounded by

\[ O \left( \frac{\log(\delta^{-1})}{N} \left( \sigma^2 \text{Tr}(\Sigma^{\frac{1}{2}}) \lambda_q^{\frac{1}{2}} + \lambda^{-1}(\eta t)^{-2r} \right) \right) . \]

Combining these results, we obtain the desired inequality. \( \square \)

**Lemma A.8.** Let \( \eta = O(1/\kappa^2) \) be sufficiently small. For any \( t \in \mathbb{N} \),

\[ \|\Sigma_{\lambda_q}^{-\frac{1}{2}}(\Sigma f_t - S^* f_t)\|_H^2 = O(R^2(\eta t)^{-2r}) \]
Proof. We denote $\lambda = 1/(\eta t)$. Note that $\Sigma f_t - S f_s = r_t(\Sigma)S f_s = r_t(\Sigma)S L^t \phi$ for some $phi \in L^2(\rho_X)$. Then we have
\[
\|\Sigma^{-\frac{1}{2}}(\Sigma f_t - S f_s)\|_H \leq \|\Sigma^{-\frac{1}{2}} r_t(\Sigma)\|_\infty \|\Sigma^{-\frac{1}{2}} S\|_\infty \|\phi\|_{L^2(\rho_X)}.
\]
From Lemma A.4, we have
\[
\|\Sigma^{-\frac{1}{2}} r_t(\Sigma)\|_\infty \leq O(\lambda^\tau).
\]
Also, observe that
\[
\|\Sigma^{-\frac{1}{2}} S\|_\infty \leq 1.
\]
Combining these results finishes the proof.

B Proof of Main Results

First we decompose the error $\|\Sigma^{{1/2-s}}(g_t - f_t)\|_H^2$ to two terms:
\[
\|\Sigma^{{1/2-s}}(g_t - f_t)\|_H^2 \leq 2 \|\Sigma^{{1/2-s}}(g_t - p_t(\Sigma_n^{(q)}) \Sigma_n^{(q)} f_t)\|_H^2 + 2 \|\Sigma^{{1/2-s}}(p_t(\Sigma_n^{(q)}) \Sigma_n^{(q)} f_t - f_t)\|_H^2.
\]
The first term can be bounded as follows:
\[
\|\Sigma^{{1/2-s}}(g_t - p_t(\Sigma_n^{(q)}) \Sigma_n^{(q)} f_t)\|_H^2 = \|\Sigma^{{1/2-s}}(p_t(\Sigma_n^{(q)}) (S_n^{(q)})^* y_n - p_t(\Sigma_n^{(q)}) \Sigma_n^{(q)} f_t)\|_H^2 \leq \|\Sigma^{{1/2-s}}(S_n^{(q)})^* y_n - \Sigma_n^{(q)} f_t\|_H^2 \leq \|\Sigma^{{1/2}} \Sigma^{-\frac{1}{2}} \Sigma_n^{(q)} \|_\infty \|\Sigma^{{1/2}} (\Sigma_n^{(q)})^* y_n - \Sigma_n^{(q)} f_t\|_H^2
\]
for any $\lambda > 0$. The second term has following bound:
\[
\|\Sigma^{{1/2-s}}(p_t(\Sigma_n^{(q)}) \Sigma_n^{(q)} f_t - f_t)\|_H^2 \leq \|\Sigma^{{1/2-s}} \Sigma^{-\frac{1}{2}} \Sigma_n^{(q)} \|_\infty \|\Sigma^{{1/2}} \Sigma_n^{(q)} \|_\infty \|\Sigma^{{1/2}} \Sigma_n^{(q)} \|_\infty \|\Sigma_n^{(q)} f_t\|_H^2 \leq \|\Sigma^{{1/2}} \Sigma_n^{(q)} \|_\infty \|\Sigma_n^{(q)} f_t\|_H^2.
\]
for any $\lambda > 0$. We particularly set $\lambda = 1/(\eta t)$.

First, we consider inequality (4) which corresponds to the second term. For bounding $\|\Sigma^{{1/2-s}} \Sigma^{-\frac{1}{2}} \Sigma_n^{(q)} \|_\infty$ and $\|\Sigma^{-\frac{1}{2}} \Sigma_n^{(q)} \|_\infty$, we apply Lemma A.4. Similarly, we can bound $\|\Sigma^{{1/2-b}} \Sigma_n^{(q)} \|_\infty$ and $\|\Sigma^{-\frac{1}{2}} \Sigma_n^{(q)} \|_\infty$. Also, we can use Lemma A.4 for $\|\Sigma^{{1/2-b}} \Sigma_n^{(q)} \|_\infty$. Finally, $\|\Sigma^{-\frac{1}{2}} \Sigma_n^{(q)} \|_\infty$ can be bounded by Lemma A.1.

Next we focus on inequality (4). We can use Lemma A.5 for bounding $\|\Sigma^{{1/2-s}} \Sigma^{-\frac{1}{2}} \Sigma_n^{(q)} \|_\infty$ and $\|\Sigma^{-\frac{1}{2}} \Sigma_n^{(q)} \|_\infty$, with high probability. Also, for bounding $\|\Sigma^{{1/2-b}} \Sigma_n^{(q)} \|_\infty$, we...
Lemma A.7 and A.6 can be applied. For bounding \( \left\| \Sigma_{\lambda}^{-\frac{1}{2}}((S_n^q)^{\top}y_n - \Sigma_n^q f_t) \right\|_H^2 \), note that the decomposition
\[
\left\| \Sigma_{\lambda}^{-\frac{1}{2}}((S_n^q)^{\top}y_n - \Sigma_n^q f_t) \right\|_H^2 \leq 2 \left\| \Sigma_{\lambda}^{-\frac{1}{2}}((S_n^q)^{\top}y_n - \Sigma_n^q f_t) - \Sigma_{\lambda}^{-\frac{1}{2}}(f_t - S_n^{q^*}f_t) \right\|_H^2 + 2 \left\| \Sigma_{\lambda}^{-\frac{1}{2}}(f_t - S_n^{q^*}f_t) \right\|_H^2.
\]
The second term can be bounded by Lemma A.8. Also, \( \left\| (S_n^{q^*})_p (\Sigma_n^q) \right\|_2^2 \) can be bounded by Lemma A.4. Then, we set \( A \) to the event that all the aforementioned bounds hold with high probability and apply Lemma A.7. Combining the results leads to the following proposition:

**Proposition B.1.** Suppose that \( \eta = O(1/\kappa^2) \) be sufficiently small. Let \( t \in \mathbb{N}, \lambda = 1/(\eta t) \geq \lambda_q = \Omega((\text{Tr}(\Sigma^{1/\alpha})/n)^{\alpha}), \delta \in (0,1) \) and \( n \geq \Omega(1 + \text{Tr}(\Sigma^{1/\alpha})\lambda_q^{-1/\alpha}) \) and \( N \geq \Omega(1 + \kappa^2 \lambda_q^{-1}). \) Then there exists event \( A \) with \( P(A) \geq 1 - \delta \) such that

\[
\mathbb{E} \left[ \left\| g_t - f_t \right\|_{L_2(\mathcal{P},X)}^2 \mid A \right] = \tilde{O} \left( \frac{\text{Tr}(\Sigma_n^q)(\sigma^2 + R^2\lambda^2r)\lambda_q^{-\frac{1}{2}} + \lambda^2r + r_N}{n} \right),
\]

where

\[
r_N = \frac{\kappa^2 \lambda_q^{-1}}{nN} \left( \sigma^2 + R^2\lambda^2r + \frac{M^2 + \kappa^4 r^{-2} R^2 + R^2\lambda^{-1} + 2r}{N} \right).
\]

Here \( \tilde{O} \) hides extra \( \text{poly}(\log(n), \delta^{-1}) \) factors.

### C Equivalence of Gradient Descent Solution to Analytic Solution

Let \( g'_n = (\Sigma_n^{q^*})^{-1}(S_n^q)^{\top}y_n \in H \). We want to bound \( \left\| S(g'_n - f_t) \right\|_{L_2(\mathcal{P},X)}^2 \) for \( \eta = \Theta(1/\kappa^2) \), where \( \lambda_n \) is defined in Definition 4.2 in the main paper. First we decompose the error \( \left\| \Sigma_1^{1/2} - (g'_n - f_t) \right\|_H^2 \) to two terms:

\[
\left\| \Sigma_{\lambda}^{1/2} - (g'_n - f_t) \right\|_H^2 = 2 \left\| \Sigma_{\lambda}^{1/2} - (\Sigma_{n,\lambda}^{q^*})^{-1} \Sigma_n^q f_t \right\|_H^2 + 2 \left\| \Sigma_{\lambda}^{1/2} - (\Sigma_{n,\lambda}^{q^*})^{-1} \Sigma_n^q f_t - f_t \right\|_H^2.
\]

The first term can be bounded as follows:

\[
\left\| \Sigma_{\lambda}^{1/2} - (\Sigma_{n,\lambda}^{q^*})^{-1} \Sigma_n^q f_t \right\|_H^2 = \left\| \Sigma_{\lambda}^{1/2} - (\Sigma_{n,\lambda}^{q^*})^{-1} \right\|_H^2 \left\| (\Sigma_{n,\lambda}^{q^*})^{-1} \Sigma_n^q f_t \right\|_H^2.
\]

The second term has following bound:

\[
\left\| \Sigma_{\lambda}^{1/2} - (\Sigma_{n,\lambda}^{q^*})^{-1} \right\|_H^2 \left\| (\Sigma_{n,\lambda}^{q^*})^{-1} \Sigma_n^q f_t - f_t \right\|_H^2.
\]

We particularly set \( t = 1/(\eta \lambda_n) \). The only differences from the arguments in Section B are the replacements of \( \left\| (\Sigma_n^{q^*})_p (\Sigma_n^q) \right\|_2^2 \) (which has bound \( O(1) \)) with 1 and \( \left\| (\Sigma_n^q)^{1/2} \Sigma_n^q f_t \right\|_H^2 \) (which has a bound \( O(\lambda_1^{1/2}r) \)) with \( \lambda_1^{1/2}r \). Hence, we obtain the perfectly same variance bound as the one of gradient descent in Theorem B.1.
D  Sufficient Condition for $\mathcal{N}_\infty(\lambda) \ll \mathcal{F}_\infty(\lambda)$

**Proposition D.1.** Let $\{(\lambda_i, \phi_i)\}_{i=1}^d (d \in \mathbb{N} \cup \{\infty\})$ be the eigen-system of $\Sigma$ in $L^2(\rho_X)$, where $\lambda_1 \geq \lambda_2 \geq \ldots > 0$. Assume that $\lambda_i = \Theta(i^{-\alpha})$ and $\|\phi_i\|_{L^\infty(\rho_X)} = \Omega(i^{p/2})$ for any $i$ for some $\alpha = 1 + \Omega(1)$ and $p \geq 1$. Moreover if $d = \infty$, we additionally assume $\|\phi_i\|^2_{L^\infty(\rho_X)} = O(i^{\alpha-1-\varepsilon})$ for any $i$ for some $\varepsilon > 0$. Then Assumption $[\![1\!]$ is satisfied and for any $\lambda \in (0, 1)$,

$$\mathcal{F}_\infty(\lambda) = \Omega(\lambda^{-\frac{p}{2}} \wedge d^p).$$

**Proof.** First note that from Mercer’s theorem, we have $K(x, x') = \sum_{i=1}^d \lambda_i \phi_i(x)\phi_i(x')$. Assumption $[\![1\!]$ is always satisfied when $d < \infty$ and thus we consider the case $d = \infty$. Since $\|\phi_i\|^2_{L^\infty(\rho_X)} = O(i^{\alpha-1-\varepsilon})$, $\|K_x\|^2_{H}$ is uniformly bounded and thus Assumption $[\![1\!]$ is satisfied. Let $\lambda > 0$.

$$\mathcal{F}_\infty(\lambda) = \sup_{x \in \text{supp}(\rho_X)} \|\Sigma^{-\frac{1}{2}} K_x\|^2_H$$

$$= \sup_{x \in \text{supp}(\rho_X)} \left\|\sum_{i=1}^d \lambda_i \phi_i(x)(\lambda_i + \lambda)^{-\frac{1}{2}} \phi_i\right\|^2_H$$

$$= \sup_{x \in \text{supp}(\rho_X)} \sum_{i=1}^d \frac{\lambda_i \phi_i(x)^2}{\lambda_i + \lambda} \geq \frac{[\lambda^{-\frac{1}{2}} \wedge d]^{-\alpha + \varepsilon}}{[\lambda^{-\frac{1}{2}} \wedge d]^{-\alpha} + \lambda} = \Omega(\lambda^{-\frac{p}{2}} \wedge d^p)$$

$\square$

E  Extension to Random Features Settings

The following lemma is analogous to Lemma $[A.1$.

**Lemma E.1.** Suppose that Assumptions $[\![2\!]$ and $[\![5\!]$ hold. Let $\eta = O(1/\kappa^2)$ be sufficiently small and $t \in \mathbb{N}$ such that $m = \Omega(1 + \kappa^2 \eta t)$. Then for any $\delta > 0$, with probability at least $1 - \delta$,

$$\|\hat{S} f_t - f_\star\|^2_{L^2(\rho_X)} = O(R^2(\eta t)^{-2r})$$

and for any $\lambda > 0$ and $s \in [0, r]$

$$\|\Sigma^{-s} f_t\|^2_{H} = O(R^2(\kappa^{4(r-s)} + (\eta t)^{1-2(r-s)})).$$

**Proof.** Recall that $\hat{f}_t = \hat{f}_{t-1} - \eta(\hat{S} f_{t-1} - \hat{S} f_\star)$ and $\hat{f}_0 = 0$. Thus we have $\hat{S} f_t = \hat{S} f_{t-1} - \eta \hat{L}(\hat{S} f_{t-1} - f_\star)$. Hence it holds that $\hat{S} f_t - f_\star = (I - \eta \hat{L})(\hat{S} f_{t-1} - f_\star)$.

Therefore we get $\|\hat{S} f_t - f_\star\|^2_{L^2(\rho_X)} = \|(I - \eta \hat{L}) f_\star\|^2_{L^2(\rho_X)} = \|(I - \eta \hat{L})^t f_\star\|^2_{L^2(\rho_X)} = \Omega(R^2)\|(I - \eta \hat{L})^t L^r\|^2.$

Let $\lambda' > 0$. Observe that $\|(I - \eta \hat{L})^t L^r\|^2 \leq \|(I - \eta \hat{L})^t \hat{L}_\lambda^c L^r\|^2 \leq \|(I - \eta \hat{L})^t \hat{L}_\lambda^c \| L^r\|^2$. We have $\|(I - \eta \hat{L})^t \hat{L}_\lambda^c\|^2 \leq \|(I - \eta \hat{L})^t L^r\|^2 + \lambda^{2r} = O((\eta t)^{2r} + \lambda^{2r})$ from Lemma $[A.4$. Also similar to Lemma $[A.5$ with $s = 0$, with high probability we have $\|\hat{L}_\lambda^c L^r\|^2 = O(1)$ if $m = \Omega(1 + \kappa^2 \lambda'^{-1})$. Finally setting $\lambda' = (\eta t)^{-1}$ yields the first statement. The second statement can be easily proven in a very similar manner to the proof of Lemma 16 in Lin and Rosasco (2017) but we need to use the fact that $\|\hat{L}_\lambda^c L^r\|^2 = O(1)$ with high probability as in the proof of the first statement. This finishes the proof.

$\square$
Lemma E.2 (Proposition 10 in [Rudi and Rosasco 2017]). Suppose that Assumption 5 holds. We denote \( \tilde{N}_\infty(\lambda) = E_x \| \Sigma_0^{-1/2} \phi_{m,x} \|_2^2 \) for \( \lambda > 0 \). For any \( \delta \in (0, 1) \) and sufficiently small \( \lambda = O(1) \), if \( m = \Omega(1 + \kappa^2 \lambda^{-1}) \), with probability at least \( 1 - \delta \) it holds that

\[
\tilde{N}_\infty(\lambda) \leq 1.55 N_\infty(\lambda).
\]

Combining the bias and variance bounds with Lemma E.2 yields the following theorem:

Theorem E.3 (Generalization Error of CRED-GD with RF). Suppose that Assumptions 2, 3, 4 and 5 hold. Let \( \eta = \Theta(1/\kappa^2) \) be sufficiently small, \( \lambda_q = \lambda_* \) and \( T = \tilde{\Theta}(t^*_n) \). For any \( \delta \in (0, 1) \), if \( m \geq \tilde{\Omega}(1 + \kappa^2 \lambda_*^{-1}) \), there exists event \( A \) with \( P(A) \geq 1 - \delta \) such that RF-CRED-GD satisfies

\[
E \left[ \left\| \hat{S}_{T} - f_* \right\|_{L^2(\rho_X)}^2 \mid A \right] = \tilde{O} \left( \left( \frac{\sigma^2 \text{Tr}(\Sigma_0^2)}{n} \right)^{\frac{2\alpha}{2\alpha + 1}} + \left( \frac{R^2 \text{Tr}(\Sigma_0^2)}{n} \right)^{\frac{2\alpha}{2\alpha + 1}} + \lambda_N^2 \right),
\]

where \( \lambda_N \) is defined in Definition 4.2 in Section 4 of the main paper.