Gradient Descent in RKHS with Importance Labeling

Tomoya Murata
NTT DATA Mathematical Systems Inc.
The University of Tokyo[1]
murata@msi.co.jp

Taiji Suzuki
The University of Tokyo[1]
RIKEN AIP[2]
taiji@mist.i.u-tokyo.ac.jp

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 settings (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

[1]Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo
[2]Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

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*.

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 analysed 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; Lin and Rosasco, 2017; Carratino et al., 2018; Pillaud-Vivien et al., 2018; Jun et al., 2019). Caponnetto and De Vito (2007); Steinwart et al. (2009) have shown the minimax optimal generalization ability of kernel ridge regression under suitable assumptions. In Yao et al. (2007); Rosasco and Villa (2015), gradient descent for kernel ridgeless regression has been considered and the effect of early stopping as implicit regularization has been theoretically justified. The analysis has been further improved with additional assumption about eigenvalues decay of the covariance operator of the feature space (Lin and Rosasco, 2017). Online stochastic gradient descent (SGD) has been studied in (Dieuleveut et al., 2016) and the minimax optimal rate has been established when the true function is (nearly) attainable. Recently the authors of (Pillaud-Vivien et al., 2018) have considered Multi-Pass SGD and shown its optimality without attainability of the true function under additional assumption about the capacity of the feature space in terms of infinity norm. Random features technique (Rahimi and Recht, 2008) can be applicable to kernel regression and reduces the computational time. The generalization ability of kernel regression with random features has been studied in Rudi and Rosasco (2017); Carratino et al. (2018) and it has been shown that random features technique doesn't hurt the generalization ability when the number of random features is sufficiently large and the true function is attainable. More recently, in (Jun et al., 2019), low label noise cases have been particularly discussed and their proposed Kernel Truncated Randomized Ridge Regression (KTR^3) achieves an improved rate when the label noise is low. However, these papers do not consider the utilization of the unlabeled data and hence the generalization ability may be sub-optimal in the importance labeling settings considered in this paper.

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, in active learning, how to select the initially labeled data is out-of-scope and typically assumed to be uniform selection. Enormous active learning strategies have been proposed (Brinker, 2003; Dasgupta, 2005; Yu et al., 2006; Kapoor

Method	Generalization Error	Additional Assumptions
(S)GD (Pillaud-Vivien et al., 2018)	$\left(\frac{C}{n}\right)^{\frac{2r}{\mu}} + \left(\frac{\sigma^2 \mathrm{Tr}(\Sigma^{\frac{1}{\alpha}})}{n}\right)^{\frac{2r\alpha}{2r\alpha+1}}$	$\exists \mu \in [\frac{1}{\alpha}, 1] : \|\Sigma^{\frac{\mu}{2}-\frac{1}{2}} K_x\|_H^2 \leq C$ a.e. x
KTR^3 (Jun et al., 2019)	$\left(n^{-2r} + \left(\frac{\sigma^2}{n}\right)^{\frac{2r}{2r+1}}\right) \wedge \left(\frac{M^2 \mathrm{Tr}(\Sigma^{\frac{1}{\alpha}})}{n}\right)^{\frac{2r\alpha}{2r\alpha+1}}$	None
SSSL (Ji et al., 2012)	$n^{-\frac{(\alpha-1)}{2}}$	$r \geq 0.5$, $\sigma^2 = 0$ sufficiently large N
CRED-GD (this paper)	$\left(\frac{\mathrm{Tr}(\Sigma^{\frac{1}{\alpha}})}{n}\right)^{2r\alpha} + \left(\frac{\sigma^2 \mathrm{Tr}(\Sigma^{\frac{1}{\alpha}})}{n}\right)^{\frac{2r\alpha}{2r\alpha+1}}$	sufficiently large N
RF-KRLS (Rudi and Rosasco, 2017)	$n^{-2r} + \left(\frac{\sigma^2 \mathrm{Tr}(\Sigma^{\frac{1}{\alpha}})}{n}\right)^{\frac{2r\alpha}{2r\alpha+1}}$	$r \geq 0.5$, sufficiently large m
RF-CRED-GD (this paper)	$\left(\frac{\mathrm{Tr}(\Sigma^{\frac{1}{\alpha}})}{n}\right)^{2r\alpha} + \left(\frac{\sigma^2 \mathrm{Tr}(\Sigma^{\frac{1}{\alpha}})}{n}\right)^{\frac{2r\alpha}{2r\alpha+1}}$	sufficiently large m, N

Table 1: Comparison of theoretical generalization errors between our proposed algorithms and most relevant existing methods (The bottom two methods use approximation by m random features). n is the number of labeled data, σ^2 is the variance of label noise, M is the uniform upper bound of labels, $r \in [0, 1]$ represents the smoothness of the target function and $\alpha > 1$ captures the simplicity of the feature space. In column "Additional Assumptions," N means the number of unlabeled data. Please refer to Section 2 for the detailed definitions of these parameters. Extra log factors $\mathrm{poly}(\log(n), \log(\delta^{-1}))$ are hided for simplicity, where δ is confidence parameter for high probability bounds.

et al., 2007; Guo and Schuurmans, 2008; Wei et al., 2015; Gal et al., 2017; Sener and Savarese, 2017) ((Settles, 2009) for extensive survey) and empirically studied their performances but their theoretical aspects are little known at least in our kernel regression setting.

Importance Sampling. Importance sampling is a general technique to reduce the variance of estimations and typically used in Monte Carlo methods and stochastic optimization (Needell et al., 2014; Zhao and Zhang, 2015; Alain et al., 2015; Csiba and Richtárik, 2018; Chen et al., 2019). The behind idea is that if the realizations that potentially cause large variance are more frequently sampled, the variance of a bias-corrected estimator can be reduced. However, the definition of importance is strongly problem-dependent and to the best of our knowledge, any algorithms for importance labeling problem have not been proposed so far.

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 $\rho_{\mathcal{Z}}$, where $z_j = (x_j, y_j) \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$, and $X_N = \{x_j\}_{j=1}^N$, $\boldsymbol{y}_N = \{y_j\}_{j=1}^N$. We denote $\rho_{\mathcal{X}}$ as the marginal distribution of $\mathcal{Z}$ on $\mathcal{X}$ and $\rho_{\mathcal{Y}|x}$ as the conditional distribution of $\mathcal{Y}$ with respect to $x \in \mathcal{X}$. We subsample $Z_n = \{(x_{j(i)}, y_{j(i)})\}_{i=1}^n$ $(n < N)$ from Z_N according to user-defined distribution q on Z_N and we denote $X_n = \{x_{j(i)}\}_{i=1}^n$, $\boldsymbol{y}_n = \{y_{j(i)}\}_{i=1}^n$.

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 $\mathcal{E}(w) = \int_{\mathcal{Z}} \frac{1}{2}(y - w(x))^2 d\rho_{\mathcal{Z}}(x, y)$ and $H \subset L^2(\rho_{\mathcal{X}}) (\subset \mathbb{R}^{\mathcal{X}})$ is some Reproducing Kernel Hilbert Space (RKHS) with inner product $\langle \cdot, \cdot \rangle_H : H \times H \to \mathbb{R}$ and kernel $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

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 $\mathcal{L} = SS^* : L^2(\rho_{\mathcal{X}}) \to L^2(\rho_{\mathcal{X}})$, where the operator S is the natural embedding from H to $L^2(\rho_{\mathcal{X}})$ and S^* is the adjoint operator of S. We define T_λ as $T + \lambda I$ for operator T. For natural number m, We denote $\{1, \ldots, m\}$ by $[m]$. K_x denotes the operator $K(x, \cdot) = K(\cdot, x) : \mathcal{X} \to \mathbb{R} \in H$ for $x \in \mathcal{X}$. 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 $\kappa > 0$, $\sup_{x \in \mathrm{supp}(\rho_{\mathcal{X}})} \|K_x\|_H \leq \kappa$.

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

Assumption 2 quantifies the complexity of true function f_* in terms of the eigen-system of $\mathcal{L}$. It is known that when $r \geq 1/2$, $\mathcal{L}^r(L^2(\rho_{\mathcal{X}}))$ becomes a subset of H and particularly $r = 1/2$, it exactly matches to H. Thus, we have $f_* \in H$ whenever $r \geq 1/2$. However, when $r < 1/2$, generally $f_* \notin H$. As $r \to 0$, roughly $\mathcal{L}^r(L^2(\rho_{\mathcal{X}})) \to L^2(\rho_{\mathcal{X}})$. This

means that f_* can be more complex (or non-smooth) for smaller r.

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

Parameter α characterizes the complexity of feature space H. For larger α, 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 $\mathrm{Tr}(\Sigma^{1/\alpha})$ can be much smaller than $\mathrm{Tr}(\Sigma^{1/\infty}) = d$ for some $\alpha \in (1, \infty)$.

Assumption 4 (Bounded variance and uniform bouneded-ness of labels)**.** There exists $\sigma \geq 0$ and $M \geq 1$ such that $\mathbb{E}(y - f_*(x))^2 \leq \sigma^2$ 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 $\mathcal{N}_\infty(\lambda) = \mathbb{E}_x \|\Sigma_\lambda^{-1/2} K_x\|_H^2$, 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 $\|\Sigma_\lambda^{-1/2} K_x\|_H^2$ 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_* to

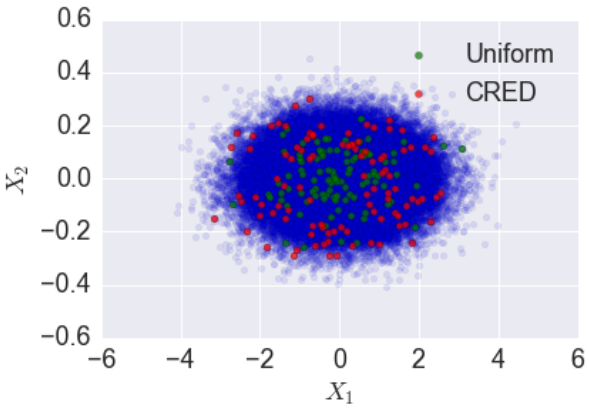


Figure 1: Comparision of the selected (labeled) data samples. (Blue) Unlabeled data ($100,000$ points). (Green) Labeled data selected by uniform sampling (100 points). (Red) Labeled data samples selected by CRED (100 points).

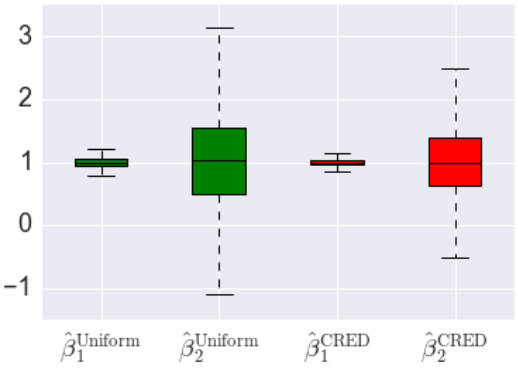


Figure 2: Comparison of the deviation of the estimated regression coefficients using 3 labeled points selected by uniform sampling and CRED ($1,000$ independent trials). The true coefficients were $(\beta_1^*, \beta_2^*) = (1, 1)$.

$f_*(x_1, x_2) = x_1 + x_2$ 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 sampling distribution proportional to $\|\Sigma_\lambda^{-1/2} K_x\|_H^2$ on unlabeled data x, that can be regard as contribution ratio to effective dimension. For stability of sampling, we add the

Algorithm 1: CRED-GD(η, λ_q, T)

1: Set $q_j = \frac{\left\|\Sigma_{N,\lambda_q}^{-\frac{1}{2}} K_{x_j}\right\|_H^2 + \frac{1}{N}\sum_{i=1}^N \left\|\Sigma_{N,\lambda_q}^{-\frac{1}{2}} K_{x_j}\right\|_H^2}{2\sum_{j=1}^N \left\|\Sigma_{N,\lambda_q}^{-\frac{1}{2}} K_{x_j}\right\|_H^2}$ for $j \in [N]$.
2: Sample $\{x_{j(i)}\}_{i=1}^n$ independently according to q and get their labels $\{y_{j(i)}\}_{i=1}^n$.
3: Set $g_0 = 0$.
4: **for** $t = 1$ to T **do**
5: $\quad A = \frac{1}{n}\sum_{i=1}^n \frac{1}{Nq_{j(i)}}(K_{x_{j(i)}} \otimes K_{x_{j(i)}})$, $b = \frac{1}{n}\sum_{i=1}^n \frac{1}{Nq_{j(i)}} y_{x_{j(i)}} K_{x_{j(i)}}$.
6: $\quad g_t = g_{t-1} - \eta\left(Ag_{t-1} - b\right)$.
7: **end for**
8: **return** g_T.

mean of the contribution ratios to it. Finally, since covariance operator Σ is unknown, we replace it by empirical covariance operator $\Sigma_{N,\lambda}$ using N unlabeled data. Line 1 in Algorithm 2 gives the formal description of this procedure.

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 $(\langle K_x, g\rangle_H - y)K_x = (K_x \otimes K_x)g - yK_x$, if the sampling probability of (x, y) from N unlabeled data is q, we need to correct the bias of the sampling 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 dimensional 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 trick (Schölkopf et al., 2002). This computational problem can be avoided by introducing random features technique. For the details, see Section 6.

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 $\widetilde{O}$ and $\widetilde{\Omega}$ notation to hide extra $\mathrm{poly}(\log(n), \log(\delta^{-1}))$ factors for simplicity, where δ is a confidence parameter for high probability bounds.

Our analysis starts from bias-variance decomposition $\|Sg_t - f_*\|_{L^2(\rho_\mathcal{X})}^2 \le 2\|Sf_t - f_*\|_{L^2(\rho_\mathcal{X})}^2 + 2\|S(g_t - f_t)\|_{L^2(\rho_\mathcal{X})}^2$, where $\{f_t\}_{t=1}^\infty$ is the ideal GD path on excess risk, i.e., $f_t = f_{t-1} - \eta(\Sigma f_{t-1} - S^* f_*) = f_{t-1} - \eta(\mathbb{E}_x[K_x \otimes K_x] - \mathbb{E}_{x,y}[yK_x])$ 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 1 and 2 hold. Let $\eta = O(1/\kappa^2)$ be sufficiently small. Then, for any $t \in \mathbb{N}$,*

$$\|Sf_t - f_*\|_{L^2(\rho_\mathcal{X})}^2 = O\left(R^2(\eta t)^{-2r}\right).$$

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 $\mathcal{N}_\infty(\lambda) = \mathbb{E}_x\|\Sigma_\lambda^{-1/2}K_x\|_H^2$ and $\mathcal{F}_\infty(\lambda) = \sup_{x\in\rho(\mathcal{X})}\|\Sigma_\lambda^{-1/2}K_x\|_H^2$.

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

Lemma 4.2. *Suppose that Assumption 1 holds. For any $\lambda > 0$, $\mathcal{F}_\infty(\lambda) \le \kappa^2\lambda^{-1}$. Additionally, under Assumption 3, for any $\lambda > 0$, $\mathcal{N}_\infty(\lambda) \le \mathrm{Tr}(\Sigma^{1/\alpha})\lambda^{-1/\alpha}$.*

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

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 $t \in \mathbb{N}$, $\lambda = 1/(\eta t) \ge \lambda_q = \Omega((\mathrm{Tr}(\Sigma^{1/\alpha})/n)^\alpha)$, $\delta \in (0, 1)$ and $n \ge \widetilde{\Omega}(1 + \mathrm{Tr}(\Sigma^{1/\alpha})\lambda_q^{-1/\alpha})$ and $N \ge \widetilde{\Omega}(1 + \kappa^2\lambda_q^{-1})$. Then there exits event A with $P(A) \ge 1 - \delta$ such that*

$$\mathbb{E}\left[\|S(g_t - f_t)\|_{L^2(\rho_\mathcal{X})}^2 \mid A\right] = \widetilde{O}\left(\frac{(\sigma^2 + R^2\lambda^{2r})\mathcal{N}_\infty(\lambda_q)}{n} + \lambda^{2r} + r_N\right),$$

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

Proposition 4.3 shows that the variance diverges to ∞ 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 (Pillaud-Vivien et al., 2018), 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 $(\sigma^2 + \lambda^{2r})\mathcal{N}_\infty(\lambda)/n$ for $\lambda \approx \lambda_q$ and sufficiently large N (note that λ^{2r} can be ignored because it never dominates the bias

term (see Proposition 4.1)). Since $\mathcal{N}_\infty(\lambda) \leq \mathcal{F}_\infty(\lambda)$ always holds, CRED-GD improves the variance bound of the standard GD when σ^2 is small. Later, we discuss the case $\mathcal{N}_\infty(\lambda) \ll \mathcal{F}_\infty(\lambda)$ (see Lemma 4.2 and Section 5).

Remark. In Pillaud-Vivien et al. (2018), under Assumption 1 and additional assumption $\sup_{\mathrm{supp}(\rho_\mathcal{X})} \|\Sigma^{\mu/2-1/2} K_x\|_H^2 = O(\kappa_\mu^2 R^{2\mu})$ for some $\kappa_\mu > 0$ and $\mu \in [0,1]$, the authors have shown that $\mathcal{F}_\infty(\lambda) = O(\kappa_\mu^2 R^{2\mu} \lambda^{-\mu})$ (Lemma 13 in (Pillaud-Vivien et al., 2018)), which is a better bound than ours in Lemma 4.2 when $\mu < 1$. However, in the worst case $\mu = 1$ their bound matches to 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_η^* is defined by $t_\eta^* = \lceil 1/(\eta \lambda_*) \rceil$, where λ_* is defined as

$$\lambda_* = \widetilde{O}\left(\left(\frac{\sigma^2 \mathrm{Tr}(\Sigma^{\frac{1}{\alpha}})}{n} \right)^{\frac{\alpha}{2r\alpha+1}} + \left(\frac{R^2 \mathrm{Tr}(\Sigma^{\frac{1}{\alpha}})}{n} \right)^{\alpha} + \lambda_N \right),$$

where $\lambda_N = \{\kappa^2\sigma^2 + \kappa^2 M^2/N + \kappa^{4r} R^2/N)/(nN)\}^{1/(1+2r)} + \kappa^2 R^2/(nN) + \kappa R/(\sqrt{n}N) \to 0$ as $N \to \infty$.

Lemma 4.2 and Proposition 4.3 with $\lambda_q = \lambda_*$ yields the following main theorem:

Theorem 4.4 (Generalization Error of CRED-GD)**.** *Suppose that Assumptions 1, 2, 3 and 4 hold. Let $\eta = \Theta(1/\kappa^2)$ be sufficiently small and $\delta \in (0,1)$. Then setting $\lambda_q = \lambda_*$, $T = \widetilde{\Theta}(t_\eta^*)$, there exists event A with $P(A) \geq 1 - \delta$ such that CRED-GD satisfies $\mathbb{E}\left[\|S g_T - f_*\|_{L^2(\rho_\mathcal{X})}^2 \mid A \right] = \widetilde{O}(\lambda_*^{2r})$, where λ_* is defined in Definition 4.2.*

From Theorem 4.4, we obtain the following observations:
Wider Optimality. When $\sigma^2 = \Theta(1)$, the generalization error of CRED-GD with sufficiently many unlabeled data becomes the optimal rate $n^{-2r\alpha/(2r\alpha+1)}$. The same rate is also achieved by supervised GD or SGD but under restrictive condition $r > (\alpha - 1)/(2\alpha)$ in our theoretical settings (Dieuleveut et al., 2016; Pillaud-Vivien et al., 2018), which is not necessary for CRED-GD.

Low Noise Acceleration. When $\sigma^2 \to 0$, the rate of CRED-GD with sufficiently many unlabeled data becomes $n^{-2r\alpha}$. In contrast, supervised GD or SGD only achieves $O(n^{-2r})$ in our theoretical settings when $\sigma^2 \to 0$, and thus CRED-GD significantly improves the generalization ability of supervised methods. Semi-supervised method SSSL (Ji et al., 2012) only achieves $n^{-(\alpha-1)/2}$ when $\sigma^2 = 0$ 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 $(\Sigma_{n,\lambda_*}^{(q)})^{-1} (S_n^{(q)})^* \boldsymbol{y}_n$ 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 $\mathcal{N}_\infty(\lambda) \ll \mathcal{F}_\infty(\lambda)$

In this section, we give a sufficient condition for $\mathcal{N}_\infty(\lambda) \ll \mathcal{F}_\infty(\lambda)$ and its simple example. The proofs are found in Section D of the supplementary material.

Proposition 5.1. *Let $\{(\lambda_i, \phi_i)\}_{i=1}^d$ ($d \in \mathbb{N} \cup \{\infty\}$) be the eigen-system of Σ in $L^2(\rho_\mathcal{X})$, where $\lambda_1 \geq \lambda_2 \geq \ldots > 0$. Assume that $\lambda_i = \Theta(i^{-\alpha})$ and $\|\phi_i\|_{L^\infty(\rho_\mathcal{X})} = \Omega(i^{p/2})$ for any i for some $\alpha = 1 + \Omega(1)$ and $p \geq 1$. Moreover if $d = \infty$, we additionally assume $\|\phi_i\|_{L^\infty(\rho_\mathcal{X})}^2 = O(i^{\alpha-1-\varepsilon})$ for any i for some $\varepsilon > 0$. Then Assumption 1 is satisfied and for any $\lambda \in (0,1)$, $\mathcal{F}_\infty(\lambda) = \Omega\left(\lambda^{-\frac{p}{\alpha}} \wedge d^p\right)$.*

Example. *Let $\mathcal{X} = [-1,1]^d$ and $\rho_\mathcal{X} = TN(0,\sigma_1^2) \otimes \cdots \otimes TN(x_d, \sigma_d^2)$, that is the product measure of truncated normal distributions with mean 0 and scale parameter σ_i^2, i.e., independent normal distributions with mean 0 and variance σ_i^2 conditioned on $[-1,1]$. Let $\sigma_1^2 \geq \cdots \geq \sigma_d^2$. We denote $\widetilde{\sigma}_i$ as the variance of $TN(0,\sigma_i^2)$ for $i \in [d]$. Note that for sufficiently small $\sigma_1^2 = \Theta(1)$, we have $\widetilde{\sigma}_i^2 = \Theta(\sigma_i^2)$ for any $i \in [d]$. Then we particularly consider linear kernel K and thus $H = \mathcal{X}$. Since the covariance matrix is $\Sigma = \mathrm{diag}(\widetilde{\sigma}_1^2, \ldots, \widetilde{\sigma}_d^2)$, the eigen-system of Σ in $L^2([-1,1]^d)$ is $\{(\widetilde{\sigma}_i^2, e_i/\sqrt{\widetilde{\sigma}_i^2})\}_{i=1}^d$, where $e_i(x) = x_i$ for $x \in [-1,1]^d$. Suppose that the polynomial decay of $\{\sigma_i^2\}_{i=1}^d$ holds: $\sigma_i^2 = \Theta(\widetilde{\sigma}_i^2) = \Theta(i^{-\alpha})$. Then from Lemma 4.2, $\mathcal{N}_\infty(\lambda) = O(\log(d)\lambda^{-1/\alpha} \wedge d)$. On the other hand, from Proposition 5.1 with $p \leftarrow \alpha$, we have $\mathcal{F}_\infty(\lambda) = \Omega(\lambda^{-1} \wedge d^\alpha)$.*

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 $K(x,x') = \mathbb{E}_{\omega \sim \pi}[\psi(x,\omega)\psi(x',\omega)]$ for $x, x' \in \mathcal{X}$ for some ψ. Random features $\phi_{m,x} \in \mathbb{R}^m$ is defined by $m^{-1/2}(\psi(x,\omega_1), \ldots, \psi(x,\omega_m))$, where $\omega_1, \ldots, \omega_m \sim \pi$ independently, is used for an approximation of $K(x,x')$ by $\langle \phi_{m,x}, \phi_{m,x'} \rangle$. Here, the number of random features $m \in \mathbb{N}$ 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

Algorithm 2: RF-CRED-GD(η, λ_q, m, T)

1: Sample $\omega_1, \dots, \omega_m \sim \pi$ independently.
2: Set $\phi_m(x_j) = m^{-\frac{1}{2}}(\phi(x_j, \omega_1), \dots, \phi(x_j, \omega_m))$ for $j \in [N]$.
3: Set $q_j = \frac{\left\|\hat{\Sigma}_{N,\lambda_q}^{-\frac{1}{2}}\phi_{m,x_j}\right\|_2^2 + \frac{1}{N}\sum_{i=1}^N \left\|\hat{\Sigma}_{N,\lambda_q}^{-\frac{1}{2}}\phi_{m,x_j}\right\|_2^2}{2\sum_{j=1}^N \left\|\hat{\Sigma}_{N,\lambda_q}^{-\frac{1}{2}}\phi_{m,x_j}\right\|_2^2}$ for $j \in [N]$, where $\hat{\Sigma}_{N,\lambda_q} = \frac{1}{N}\sum_{j=1}^N \phi_{m,x_j}\phi_{m,x_j}^\top + \lambda_q I$.
4: Sample $\{x_{j(i)}\}_{i=1}^n$ independently according to q and get their labels $\{y_{j(i)}\}_{i=1}^n$.
5: Set $\hat{g}_0 = 0$.
6: **for** $t = 1$ to T **do**
7: $\quad A = \frac{1}{n}\sum_{i=1}^n \frac{1}{Nq_{j(i)}}(\phi_{m,x_{j(i)}}\phi_{m,x_{j(i)}}^\top)$, $b = \frac{1}{n}\sum_{i=1}^n \frac{1}{Nq_{j(i)}} y_{x_{j(i)}}\phi_{m,x_{j(i)}}$.
8: $\quad \hat{g}_t = \hat{g}_{t-1} - \eta\,(A\hat{g}_{t-1} - b)$.
9: **end for**
10: **return** $\hat{g}_T$.

illustrated in Algorithm 2. The difference from Algorithm 1 is only the replacement of K_x to random features $\phi_{M,x}$. 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 abound the boundedness of the random features for theoretical analysis:

Assumption 5. $\sup_{x\in\mathrm{supp}(\rho_\mathcal{X}),\omega\in\mathrm{supp}(\pi)}|\psi(x,\omega)| \le \kappa$ for some $\kappa > 0$.

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

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

Generalization Error Analysis. We consider generalization error $\|\hat{S}\hat{g}_t - f_*\|_{L^2(\rho_\mathcal{X})}^2$. We decompose the generalization error to bias and variance $\|\hat{S}\hat{g}_t - f_*\|_{L^2(\rho_\mathcal{X})}^2 \le 2\|\hat{S}\hat{f}_t - f_*\|_{L^2(\rho_\mathcal{X})}^2 + 2\|\hat{S}\hat{g}_t - \hat{S}\hat{f}_t\|_{L^2(\rho_\mathcal{X})}^2$, where where $\{\hat{f}_t\}_{t=1}^\infty$ is the ideal path of GD with RF on excess risk, i.e., $\hat{f}_t = \hat{f}_{t-1} - \eta(\hat{\Sigma}f_{t-1} - \hat{S}^*f_*) = \hat{f}_{t-1} - \eta(\mathbb{E}_x[\phi_{m,x}\phi_{m,x}^\top] - \mathbb{E}_{x,y}[y\phi_{m,x}])$ with $\hat{f}_0 = 0$. The bias term can be bounded similar to Proposition 4.1:

Proposition 6.1 (Bias bound for RF setting, simplified version of Lemma 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 $m = \widetilde{\Omega}(1+\kappa^2\eta t)$. Then for any $\delta > 0$, with probability at least $1-\delta$,*

$$\left\|\hat{S}\hat{f}_t - f_*\right\|_{L^2(\rho_\mathcal{X})}^2 = O\left(R^2(\eta t)^{-2r}\right).$$

Remark. Compared to Lemma 4.1, additional condition $m = \widetilde{\Omega}(1+\kappa^2\eta t)$ 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 $\hat{\mathcal{N}}_\infty(\lambda_q)$ and $\hat{\mathcal{F}}_\infty(\lambda)$ respectively. $\hat{\mathcal{F}}_\infty(\lambda)$ has a trivial bound $O(\lambda^{-1})$. The key lemma for bounding $\hat{\mathcal{N}}_\infty(\lambda_q)$ is the following:

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

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

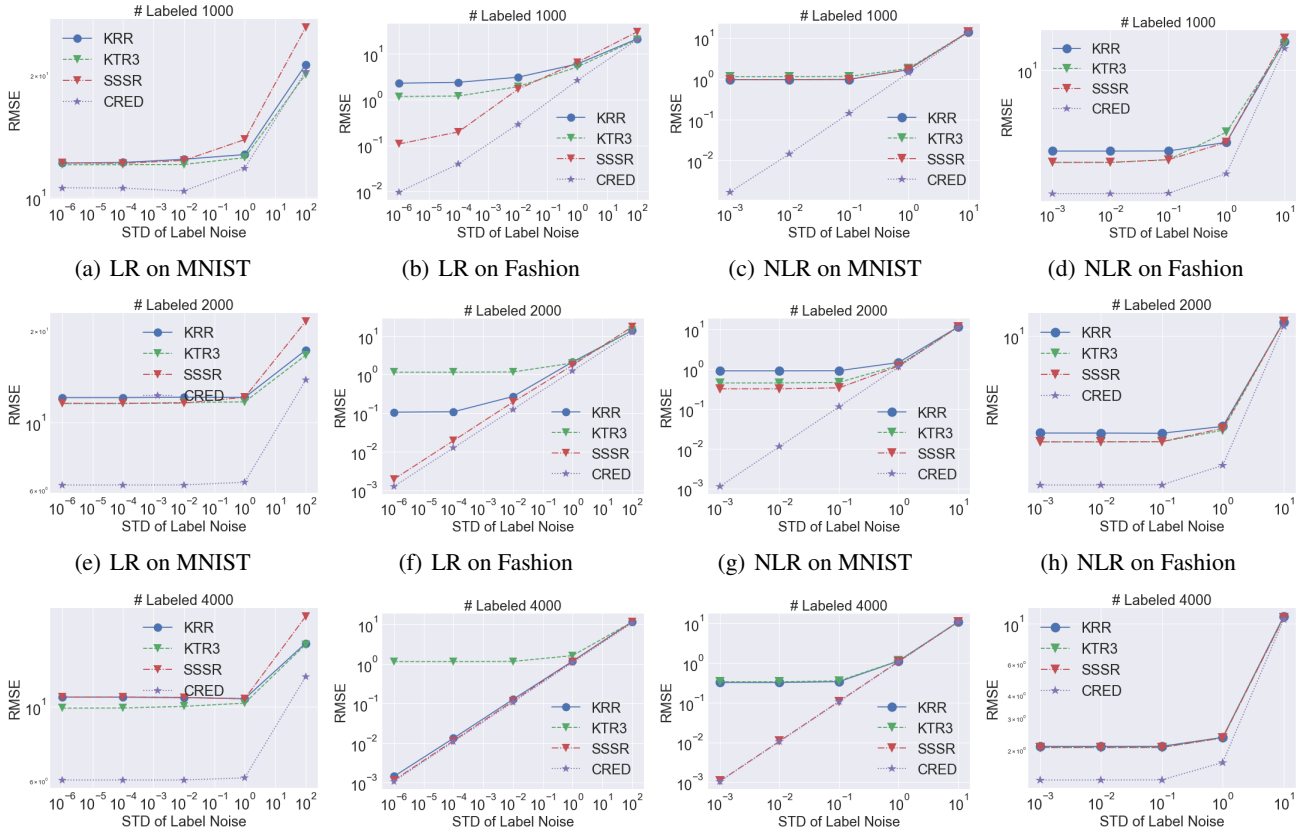
Theorem 6.3 (Generalization error of CRED-GD with RF, simplified version of Theorem E.3)**.** *Suppose that Assumptions 2, 3, 4 and 5 hold. Let $\eta = \Theta(1/\kappa^2)$ be sufficiently small, $\lambda_q = \lambda_*$ and $T = \widetilde{\Theta}(t_\eta^*)$. For any $\delta \in (0,1)$, if $m \ge \widetilde{O}(1+\kappa^2\lambda_*^{-1})$, there exists event A with $P(A) \ge 1-\delta$ 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 $\sum_i a_i e_i/\sqrt{\lambda_i} \in \mathbb{R}^{500}$, where $a_i \overset{\text{i.i.d}}{\sim} N(0,1)$ 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[3]with KRR (Kernel Ridge Re-



(a) LR on MNIST (b) LR on Fashion (c) NLR on MNIST (d) NLR on Fashion

(e) LR on MNIST (f) LR on Fashion (g) NLR on MNIST (h) NLR on Fashion

(i) LR on MNIST (j) LR on Fashion (k) NLR on MNIST (l) NLR on Fashion

Figure 3: Comparisons of test RMSE of our method with existing methods on linear and nonlinear regression tasks. The first column depicts the results of linear regression task on MNIST. The second column does the ones of linear regression on Fashion MNIST. The third column does the ones of nonlinear regression on MNIST. The last column does the ones of nonlinear regression on Fashion MNIST. From top to bottom, the number of labeled data increases from 1000 to 4000.

gression), KTR[3] (Jun et al., 2019) and SSSR (Ji et al., 2012).

The hyper-parameters were fairly and reasonably determined.[4] 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 σ^2 was large, all the algorithms have similar performances.
- When the label noise σ^2 was small, CRED significantly outperformed the other methods overall. SSSR was always comparable to or better than KRR and KTR3, 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

[3] As we mentioned before, we used very small synthetic label noise in some experiments and then the convergence speed of gradient descent was sometimes quite slow. Hence we decided that optimization methods were replaced with analytical methods. As we pointed out in the end of Section 4, the same generalization error bound is guaranteed for the analytical solution.

[4] CRED has hyper-parameter λ_q and selecting best one requires additional labeling. In our experiments, we recorded the best test error by trying λ_q in $\{10^i \mid i \in \{-12, \ldots, -3\}\}$. This potentially violates the fair comparison with the other methods because CRED implicitly uses ten patterns of labeled data. Hence we decided that the other methods were ran ten times with independent uniform labeling and then the best test error was recorded as one experimental trial.

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.