A Sharper Generalization Bound for Divide-and-Conquer Ridge Regression

Shusen Wang
Department of Computer Science, Stevens Institute of Technology
shusen.wang@stevens.edu

Abstract

We study the distributed machine learning problem where the n feature-response pairs are partitioned among m machines uniformly at random. The goal is to approximately solve an empirical risk minimization (ERM) problem with the minimum amount of communication. The divide-and-conquer (DC) method, which was proposed several years ago, lets every worker machine independently solve the same ERM problem using its local feature-response pairs and the driver machine combine the solutions. This approach is in one-shot and thereby extremely communication-efficient. Although the DC method has been studied by many prior works, reasonable generalization bound has not been established before this work.

For the ridge regression problem, we show that the prediction error of the DC method on unseen test samples is at most ϵ times larger than the optimal. There have been constant-factor bounds in the prior works, their sample complexities have a quadratic dependence on d, which does not match the setting of most real-world problems. In contrast, our bounds are much stronger. First, our $1+\epsilon$ error bound is much better than their constant-factor bounds. Second, our sample complexity is merely linear with d.

Introduction

We study linear regression, a fundamental problem in machine learning (ML), and conduct a statistical analysis of the divide-and-conquer (DC) method (Zhang, Duchi, and Wainwright 2013; 2015) for solving ridge regression. We first formally define the problem. Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times d}$ be the training feature matrix and $\mathbf{y} = [y_1, \cdots, y_n] \in \mathbb{R}^n$ be the training response vector. Assume the observed response y_i is the sum of a linear function $\mathbf{w}_0^T \mathbf{x}_i$ and unknown random noise. The goal is to estimate $\mathbf{w}_0$ based on the training data and make a prediction for any unseen test feature vector $\mathbf{x}' \in \mathbb{R}^d$.

A principled approach to the estimation of $\mathbf{w}_0$ is the regularized empirical risk minimization (ERM), including the ridge regression, LASSO (Tibshirani 1996), and the elastic net (Zou and Hastie 2005). Because of the simplicity, we focus on the ridge regression model

$$\mathbf{w}_{\text{erm}} = \operatorname*{argmin}_{\mathbf{w} \in \mathbb{R}^d} \tfrac{1}{2n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \tfrac{\gamma}{2} \|\mathbf{w}\|_2^2, \qquad (1)$$

where γ is the regularization parameter trading off the bias and variance. It requires $\mathcal{O}(nd)$ memory and $\mathcal{O}(nd^2)$ time (via the QR decomposition) or $\mathcal{O}(nd\sqrt{\kappa}\log\frac{d}{\epsilon})$ time (via the conjugate gradient method), where κ is the condition number of the Hessian matrix $\frac{1}{n}\mathbf{X}^T\mathbf{X} + \gamma\mathbf{I}$ and ϵ is the error tolerance. For big-data or high-dimensional problems, the feature matrix $\mathbf{X}$ may not fit in the memory of any single machine, and the computation may be too expensive.

Different methods have been proposed to address the computational challenges. One approach is the randomized approximation which sacrifices some accuracy for a significant reduction in the time and space complexities (Lu et al. 2013; Avron, Clarkson, and Woodruff 2016; Wang, Gittens, and Mahoney 2018; Wang et al. 2017; Derezinski and Warmuth 2018). The key idea is the matrix sketching, which finds a smaller matrix which preserves some task-specific information in the original matrix and does the computation on the smaller matrix (Drineas and Mahoney 2016; Mahoney 2011; Woodruff 2014). Although the theories in (Avron, Clarkson, and Woodruff 2016; Clarkson and Woodruff 2013; Drineas, Mahoney, and Muthukrishnan 2006; Drineas et al. 2011) showed that this approximate solution, denote $\tilde{\mathbf{w}}$, is not far from the ERM solution $\mathbf{w}_{\text{erm}}$ in the ℓ_2 norm sense, it magnifies either the bias or the variance for all the commonly used matrix sketching (Raskutti and Mahoney 2015; Wang, Gittens, and Mahoney 2018). Such a pessimistic result indicates that in ML applications, matrix sketching leads to much worse statistical risk.

Another popular approach to large-scale linear regression is distributed computing which uses massive computation power to handle big data, e.g., the works by (Avron, Maymounkov, and Toledo 2010; Meng, Saunders, and Mahoney 2014). Although it significantly reduces the local computation and memory costs, distributed computing inevitably incurs communication across the computer network. Numerical methods such as conjugate gradient are highly iterative, and each iteration requires communication and synchronization. When implemented using distributed computing systems such as MapReduce (Dean and Ghemawat 2008), Apache Spark (Zaharia et al. 2010; Meng et al. 2016), and Parameter Server (Li et al. 2014), communication and/or synchronization can often be the bottleneck of such numerical methods. Avoiding or reducing communication can make distributed computing much more efficient.

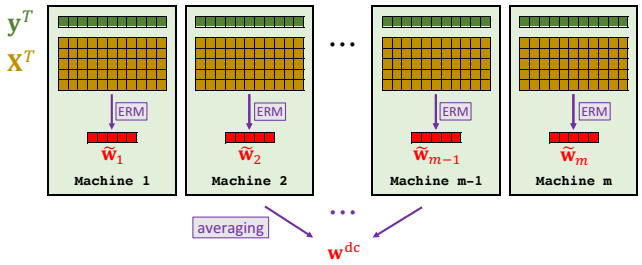

Figure 1: Illustration of the divide-and-conquer (DC) method for distributed computing. The DC method has only one round of communication.

In the federated learning framework (Konecnỳ et al. 2016a; 2016b; Bonawitz et al. 2017; McMahan et al. 2017; Smith et al. 2017), the communication and synchronization are much more expensive than computation. Federated learning assumes the data are generated by or distributed over a network across nodes that enjoy reasonable computational resources, e.g., mobile phones, wearable devices, and smart homes. As the network has limited bandwidth and high latency, the communication between the central server and the nodes (e.g., mobile phones) is slow and may cost money. In such settings, avoiding or reducing the communication is not only preferable but also necessary.

In recent years, many communication-efficient methods, e.g., (Mahajan et al. 2013; Zhang, Duchi, and Wainwright 2013; Shamir, Srebro, and Zhang 2014; Smith et al. 2016; Wang et al. 2018), have been proposed to make distributed computing more efficient. Among the communication-efficient methods, the divide-and-conquer (DC) method (illustrated in Figure 1) reduces the communication to the extreme. The DC method lets every worker machine solve the same empirical risk minimization (ERM) problem using its local data and averages the local solutions. Thus the DC method has only one round of communication.

Zhang, Duchi, and Wainwright (2013) showed that the DC solution $\mathbf{w}_{\mathrm{dc}}$ converges to the ERM solution $\mathbf{w}_{\mathrm{erm}}$ in terms of the ℓ_2 norm distance $\|\mathbf{w}_{\mathrm{dc}} - \mathbf{w}_{\mathrm{erm}}\|_2$. However, such a result does not imply good training or test error. Wang, Gittens, and Mahoney (2017) showed that for the ridge regression problem, the DC solution is close to the ERM solution in terms of the in-sample statistical risk, which is a more interesting result because it is relevant to the ML objective. Empirical studies in (Wang, Gittens, and Mahoney 2018; Zhang, Duchi, and Wainwright 2013) have shown that the DC solution indeed generalizes under some conditions. Thus it is interesting to provide theoretical justification for the empirical observations.

Main Results

Compared to the ℓ_2 distance and the in-sample risk (training error), the machine learning (ML) community is more interested in the prediction error on the unseen test set. For this reason, we consider this question: *Does the divide-and-conquer (DC) solution generalize to unseen test samples?*

To answer this question, we study the out-of-sample prediction errors. Let $\mathbf{w}_0$ be the unknown ground truth; assume $\mathbf{y}$ is $\mathbf{X}\mathbf{w}_0$ plus random noise. For the empirical risk minimization (ERM), we define the prediction error

$$P_{\mathrm{erm}}(\mathbf{X}) \triangleq \mathbb{E}_{\mathbf{x}',\mathbf{y}}\left[\left(\mathbf{w}_0^T\mathbf{x}' - \mathbf{w}_{\mathrm{erm}}^T\mathbf{x}'\right)^2\right], \tag{2}$$

where $\mathbf{x}'$ is an unseen test feature vector and the expectation is taken w.r.t. the randomness in $\mathbf{x}'$ and $\mathbf{y}$. We analogously define $P_{\mathrm{dc}}(\mathbf{X})$ by replacing $\mathbf{w}_{\mathrm{erm}}$ by $\mathbf{w}_{\mathrm{dc}}$.

Knowing $P_{\mathrm{dc}}(\mathbf{X})$ alone does not provide much information about how well the DC solution $\mathbf{w}_{\mathrm{dc}}$ generalizes. The ERM solution $\mathbf{w}_{\mathrm{erm}}$ is the "optimal" we can hope for with the available training set $(\mathbf{X}, \mathbf{y})$. Thus we contrast $P_{\mathrm{dc}}(\mathbf{X})$ with the in-sample (training) risk of the ERM solution, denote $R_{\mathrm{erm}}(\mathbf{X})$. For the ERM, the in-sample statistical risk be defined as

$$R_{\mathrm{erm}}(\mathbf{X}) \triangleq \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_{\mathbf{y}}\left[\left(\mathbf{w}_0^T\mathbf{x}_j - \mathbf{w}_{\mathrm{erm}}^T\mathbf{x}_j\right)^2\right], \tag{3}$$

where the expectation is taken w.r.t. the randomness in the training response $\mathbf{y}$. We analogously define $R_{\mathrm{dc}}(\mathbf{X})$.

Under mild assuptions, we show that the prediction error of the DC solution, denote $P_{\mathrm{dc}}(\mathbf{X})$, converges to $R_{\mathrm{erm}}(\mathbf{X})$ at a rate of $\frac{1}{\sqrt{n}}$. Because $R_{\mathrm{erm}}(\mathbf{X})$ is the best we can hope for with the available training data, our result indicates the generalization of the DC solution.

Our Contributions

The prior works (Zhang, Duchi, and Wainwright 2015; Lin, Guo, and Zhou 2017) established generalization bounds for the divide-and-conquer (DC) ridge regression. Their results have two major limitations. First, they required a high sample complexity: the number of training samples, n, has a high-order dependence on the number of features, d, which is unrealistic in real-world problems. Second, they showed mere constant-factor bounds, which means that $P_{\mathrm{dc}}(\mathbf{X})$ is at most constant times larger than $R_{\mathrm{erm}}(\mathbf{X})$.

In contrast, our results are much stronger. First, our sample complexity has at most linear dependence on d. Second, we establish $1+\epsilon$ bounds, which are much stronger than the constant-factor bounds. More specifically, we show that $P_{\mathrm{dc}}(\mathbf{X})$ is ϵ times worse than $R_{\mathrm{erm}}(\mathbf{X})$ and that ϵ vanishes at a rate of $\frac{1}{\sqrt{n}}$.

Paper Organization

The rest of this paper is organized as follows. We first define the notation and model assumptions. We then summarize our main result and compare it with the prior works. The two subsequent sections prove the main result. Since extensive empirical studies of the DC method have been conducted by (Zhang, Duchi, and Wainwright 2013; Wang, Gittens, and Mahoney 2018), we focus on the theory without conducting experiments.

Notation and Model Assumptions

We define the notation and model assumptions which are used throughout this paper. In Table 1, we list some commonly used notation.

Table 1: Commonly used notation.

Notation	Definition
n	total number of training samples
d	number of features (attributes)
m	number of partitions
d_{eff}^{γ}	γ-effective dimension of $\mathbf{X}$ ($d_{\text{eff}}^{\gamma} \leq d$)
μ^{γ}	γ-row coherence of $\mathbf{X}$
$\mathbf{w}_{\text{erm}}$	the empirical risk minimization (ERM) solution
$\mathbf{w}_{\text{dc}}$	the divide-and-conquer (DC) solution

Notation

Let $[n]$ be the set $\{1, 2, \cdots, \}$. Let $\mathbf{I}_n$ be the $n \times n$ identity matrix. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the training feature matrix, $r = \text{rank}(\mathbf{X})$, and $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^T$ be its singular value decomposition (SVD). The Moore-Penrose inverse of $\mathbf{X}$ is defined by $\mathbf{X}^{\dagger} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T$.

The row γ-ridge leverage score (for $\gamma \geq 0$) of $\mathbf{X}$ is defined by

$$\ell_i^{\gamma} \;=\; \mathbf{x}_i^T \big(\mathbf{X}^T\mathbf{X} + n\gamma \mathbf{I}_d\big)^{\dagger} \mathbf{x}_i \;=\; \sum_{j=1}^{d} \frac{\sigma_j^2}{\sigma_j^2 + n\gamma} u_{ij}^2,$$

for $i = 1, \cdots, n$.

The γ-effective dimension of $\mathbf{X}$ is

$$d_{\text{eff}}^{\gamma}(\mathbf{X}) \;=\; \sum_{i=1}^{n} \ell_i^{\gamma} \;=\; \sum_{j=1}^{d} \frac{\sigma_j^2}{\sigma_j^2 + n\gamma} \;\leq\; \text{rank}(\mathbf{X}).$$

The effective dimension is small when $n\gamma$ is substantially larger than the tail singular values of $\mathbf{X}^T\mathbf{X}$. In particular, (Cohen, Musco, and Musco 2015) showed that for any positive k, if $n\gamma \geq \frac{1}{k}\sum_{j=k+1}^{d} \sigma_j^2$, then $d_{\text{eff}}^{\gamma} \leq 2k$. In the worst case where $\gamma = 0$, the effective dimension equals d.

The γ-row coherence of $\mathbf{X}$ is

$$\mu^{\gamma} \;=\; \frac{n}{d_{\text{eff}}^{\gamma}} \max_{j \in [n]} \ell_j^{\gamma}.$$

Low coherence means the information in $\mathbf{X}$ are spread out rather than concentrated to a small number of rows, and vice versa. The standard row coherence of $\mathbf{X}$ can be expressed as

$$\mu^0 \;=\; \frac{n}{d} \max_{j \in [n]} \ell_j^0 \;=\; \frac{n}{d} \max_{j \in [n]} \mathbf{x}_j^T (\mathbf{X}^T\mathbf{X})^{\dagger} \mathbf{x}_j,$$

which is widely used in compress sensing (Candes and Tao 2006) and matrix completion (Candes and Recht 2009).

Model Assumptions

We consider the following random design model. Assume that the training and test feature vectors are independently sampled and that they have the same second moment:

$$\mathbb{E}[\mathbf{x}\mathbf{x}^T] \;=\; \mathbf{M}, \tag{4}$$

where $\mathbf{x} \in \mathbb{R}^d$ is either a training or test feature vector. Let $\mathbf{w}_0 \in \mathbb{R}^d$ be the true and unknown model. The observed response associated with $\mathbf{x}$ is

$$y \;=\; \mathbf{w}_0^T \mathbf{x} + \xi, \tag{5}$$

where ξ captures the random noise. We assume

$$\mathbb{E}[\xi] \;=\; 0 \quad \text{and} \quad \mathbb{E}[\xi^2] \;=\; \sigma^2$$

and each copy of ξ is independently drawn.

As for the divide-and-conquer (DC) method, we assume the n training feature-response pairs, $\{(\mathbf{x}_j, y_j)\}_{j=1}^{n}$, are randomly and uniformly partitioned among m machines.

Main Results and Comparisons

First, we formally define the divide-and-conquer (DC) method for ridge regression. Then, we present our main technical result. Last, we compare our result with the prior works.

Divide-and-Conquer Method for Ridge Regression

Let the rows of $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$ be randomly and uniformly partitioned to m parts, denote $\mathbf{X}_{[1]}, \cdots, \mathbf{X}_{[m]} \in \mathbb{R}^{\frac{n}{m} \times d}$ and $\mathbf{y}_{[1]}, \cdots, \mathbf{y}_{[m]} \in \mathbb{R}^{\frac{n}{m}}$. The i-th worker machine holds $\mathbf{X}_{[i]}$ and $\mathbf{y}_{[i]}$ and locally solves the empirical risk minimization (EMR) problem

$$\widetilde{\mathbf{w}}_i \;=\; \underset{\mathbf{w}}{\text{argmin}} \; \tfrac{m}{n} \big\|\mathbf{X}_{[i]}\mathbf{w} - \mathbf{y}_{[i]}\big\|_2^2 + \gamma \|\mathbf{w}\|_2^2. \tag{6}$$

The driver computes the divide-and-conquer (DC) solution

$$\mathbf{w}_{\text{dc}} \;=\; \frac{1}{m} \sum_{i=1}^{m} \widetilde{\mathbf{w}}_i,$$

which takes only one round of communication. The DC method is illustrated in Figure 1.

Summary of Main Results

Let $\epsilon, \delta \in (0, 1)$ be arbitrary, μ^{γ} be the γ-ridge coherence of $\mathbf{X}$, τ ($\approx \mu^0$) is another coherence to be defined, and d_{eff}^{γ} ($\leq d$) be the γ-effective dimension of $\mathbf{X}$. For the sample complexity

$$n = \tilde{\Theta}\Big(\tfrac{\mu^{\gamma} d_{\text{eff}}^{\gamma} m^2}{\epsilon^2} + \mu^0 dm + \tfrac{\tau d}{\epsilon^2}\Big), \tag{7}$$

it holds with high probability that

$$P_{\text{dc}}(\mathbf{X}) \;\leq\; (1 + \epsilon)\, R_{\text{erm}}(\mathbf{X}). \tag{8}$$

Here $\tilde{\Theta}$ hides the logarithms; the failure probability is from (1) the random distribution of $\mathbf{X}$ and $\mathbf{x}'$, (2) random noise in $\mathbf{y}$, and (3) the random partition of $\mathbf{X}$ and $\mathbf{y}$. This is our main technical result and can be proved by combining Theorems 6 and 9.

It was established by (Wang, Gittens, and Mahoney 2018) that under the model assumptions,

$$R_{\text{erm}}(\mathbf{X}) \;=\; \mathsf{bias}_{\text{erm}}^2(\mathbf{X}) + \mathsf{var}_{\text{erm}}(\mathbf{X}), \tag{9}$$

where

$$\begin{aligned} \mathsf{bias}_{\text{erm}}^2(\mathbf{X}) &= \gamma^2 n \big\|(\mathbf{I}_r + n\gamma\mathbf{\Sigma}^{-2})^{-1}\mathbf{V}^T\mathbf{w}_0\big\|_2^2, \\ \mathsf{var}_{\text{erm}}(\mathbf{X}) &= \tfrac{\sigma^2}{n} \big\|\big(\mathbf{I}_r + n\gamma\mathbf{\Sigma}^{-2}\big)^{-1}\big\|_F^2. \end{aligned}$$

Here $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the SVD and $r = \text{rank}(\mathbf{X})$. By setting the regularization parameter $\gamma \propto \frac{1}{\sqrt{n}}$, $R_{\text{erm}}(\mathbf{X})$ converges to zero at a rate of $\frac{1}{\sqrt{n}}$. This is why we contrast $P_{\text{dc}}(\mathbf{X})$ with $R_{\text{erm}}(\mathbf{X})$. Since $R_{\text{erm}}(\mathbf{X})$ is the best in-sample risk one can hope for with the training data and converges to zeros with the increase of n, Eqns (7) and (8) together show that the DC solution is a very practical choice when n is big compared to $m^2 d_{\text{eff}}^{\gamma}$ and d.

Related Works

The divide-and-conquer (DC) method was proposed by Zhang, Duchi, and Wainwright (2013) for solving a class of convex optimization problems. Their paper established a bound on the ℓ_2 norm error $\|\mathbf{w}_{\text{erm}} - \mathbf{w}_{\text{dc}}\|_2$, where $\mathbf{w}_{\text{erm}}$ and $\mathbf{w}_{\text{dc}}$ are the ERM and DC solutions, respectively. However, such a bound is irrelevant to machine learning. As pointed out by (Wang, Gittens, and Mahoney 2018), a small $\|\mathbf{w}_{\text{erm}} - \mathbf{w}_{\text{dc}}\|_2$ does not necessarily imply a good training or test error.

Wang, Gittens, and Mahoney (2017) analyzed the in-sample statistical risk $R_{\text{dc}}(\mathbf{X})$ (defined in (3)) and compared it to the ERM solution. They showed that for the sample complexity $n = \tilde{\Theta}(dm^2/\epsilon^2)$, where $\tilde{\Theta}$ hides logarithms and coherence, $R_{\text{dc}}(\mathbf{X})$ is ϵ times worse than $R_{\text{erm}}(\mathbf{X})$, where d and m are the number of features and the number of partitions, respectively. Their results have two limitations. First, whether the DC solution generalizes to unseen test samples was unknown. Second, their bound requires a complexity of $n = \tilde{\Theta}(dm^2/\epsilon^2)$, and thus the bound does not apply to high-dimensional data (i.e., d is large).

The generalization of the DC solution has been shown by (Zhang, Duchi, and Wainwright 2015; Lin, Guo, and Zhou 2017). Zhang, Duchi, and Wainwright (2015) established a constant-factor bound (i.e., $\mathcal{O}(1)$ times larger than the optimal generalization error), assuming the sample complexity is $n = \tilde{\Theta}(md^2)$. Unfortunately, the $\tilde{\Theta}(d^2)$ dependence makes their theory not applicable to most real-world problems. Lin, Guo, and Zhou (2017) showed different guarantees for the DC solution; however, because the kernel function $K(\mathbf{x}, \mathbf{x}) = \mathbf{x}^T\mathbf{x} = \Theta(d)$ for the linear kernel, their results have high-order dependence on d.

This work provides a stronger generalization bound than (Zhang, Duchi, and Wainwright 2015; Lin, Guo, and Zhou 2017). First, our result is $1+\epsilon$ bound, which is much stronger than their constant-factor bounds. Second, we improve the sample complexity to $n = \tilde{\Theta}(d_{\text{eff}}^{\gamma} m^2/\epsilon^2 + dm + d/\epsilon^2)$, where d_{eff}^{γ} ($\leq d$) is the effective dimension and typically very small. Our sample complexity is meaningful even for high-dimensional data. Our analysis techniques are reminiscent of (Wang, Gittens, and Mahoney 2018) but totally different from (Lin, Guo, and Zhou 2017; Zhang, Duchi, and Wainwright 2015). Our main results are simple and clear, and our proof techniques are easy to follow.

Analysis of the In-Sample Risk

In this section, we analyze the in-sample risk of the divide-and-conquer (DC) solution $\mathbf{w}_{\text{dc}}$ and compare it to the empirical risk minimization (ERM) solution $\mathbf{w}_{\text{erm}}$. Theorem 6 is the main theorem of this section. We prove Theorem 6 using random matrix theories and the bias-variance decomposition.

Analysis via Random Matrix Theories

We call $\mathbf{S} \in \mathbb{R}^{n\times s}$ a *column selection matrix* if each column has exactly one non-zero entry whose value is $\sqrt{n/s}$ and whose position indicates the selected column. Applying $\mathbf{S}$ to $\mathbf{X}^T \in \mathbb{R}^{d\times n}$, the result $\mathbf{X}^T\mathbf{S} \in \mathbb{R}^{d\times s}$ contains s selected and scaled feature vectors. Uniform sampling can be captured by a column selection matrix; we call the random matrix *the uniform sampling matrix*.

The model assumption that the n samples are randomly and uniformly partitioned among m machines can be expressed using the notation of uniform sampling matrix. Let $\mathbf{X}_{[1]}, \cdots, \mathbf{X}_{[m]} \in \mathbb{R}^{s\times d}$ ($s = \frac{n}{m}$) be the partition of $\mathbf{X}$, $\mathbf{y}_{[1]}, \cdots, \mathbf{y}_{[m]} \in \mathbb{R}^s$ be the partition of $\mathbf{y}$, and $\mathbf{S}_1, \cdots, \mathbf{S}_m \in \mathbb{R}^{n\times s}$ be the corresponding uniform sampling matrices. (The m subsets are not mutually or even pairwisely independent, but we do not need such independence in our analysis.) Then

$$\mathbf{X}_{[i]} = \tfrac{1}{\sqrt{m}}\mathbf{S}_i^T\mathbf{X} \quad \text{and} \quad \mathbf{y}_{[i]} = \tfrac{1}{\sqrt{m}}\mathbf{S}_i^T\mathbf{y},$$

for $i = 1, \cdots, m$. Eqn. 6 can be equivalently written as

$$\begin{aligned}\widetilde{\mathbf{w}}_i &= \operatorname*{argmin}_{\mathbf{w}} \tfrac{1}{n}\big\|\mathbf{S}_i^T\mathbf{X}\mathbf{w} - \mathbf{S}_i^T\mathbf{y}\big\|_2^2 + \gamma\|\mathbf{w}\|_2^2 \\ &= \big(\mathbf{X}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{X} + n\gamma\mathbf{I}_d\big)^{\dagger}\big(\mathbf{X}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{y}\big).\end{aligned}$$

With the uniform sample matrix notation, we are able to analyze the DC solution using random matrix theories. Lemmas 1 and 2 are used in proving Theorem 6.

Lemma 1 shows that when s is large compared to d_{eff}^{γ}, we can form a spectral approximation to the Hessian matrix of the ridge regression problem 1 by uniform sampling. A very similar result was previously established by (Cohen, Musco, and Musco 2015).

Lemma 1. *The Hessian matrix of the ridge regression problem* (1) *is* $\mathbf{H} = \mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d$. *Let* $\mathbf{S} \in \mathbb{R}^{n\times s}$ *be a uniform sampling matrix and* $\widetilde{\mathbf{H}} = \mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X} + n\gamma\mathbf{I}_d$. *Let* $\epsilon, \delta \in (0,1)$ *be arbitrary. When*

$$s = \Theta\big(\tfrac{\mu^{\gamma} d_{\text{eff}}^{\gamma}}{\epsilon^2}\log\tfrac{d_{\text{eff}}^{\gamma}}{\delta}\big),$$

the spectral approximation holds with probability at least $1-\delta$*:*

$$(1-\epsilon)\mathbf{H} \preceq \widetilde{\mathbf{H}} \preceq (1+\epsilon)\mathbf{H}.$$

Lemma 2, known as the subspace embedding property, was established by (Wang, Luo, and Zhang 2016; Woodruff 2014). It ensures that all the singular values of $\mathbf{U}^T\mathbf{S} \in \mathbb{R}^{r\times s}$ are close to one.

Lemma 2. *Let* $\mathbf{S} \in \mathbb{R}^{n\times s}$ *be a uniform sampling matrix and* $\epsilon, \delta \in (0,1)$ *be arbitrary. When*

$$s = \Theta\big(\tfrac{\mu^0 d}{\epsilon^2}\log\tfrac{d}{\delta}\big),$$

it holds with probability at least $1-\delta$ *that*

$$1-\epsilon \leq \big\|\mathbf{U}^T\mathbf{S}\big\|_2^2 \leq 1+\epsilon.$$

Analysis via Bias-Variance Decomposition

Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be the training feature matrix, $r = \operatorname{rank}(\mathbf{X})$, and $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the SVD. For the DC solution, the in-sample statistical risk can be decomposed in the following way. The lemma was established by (Wang, Gittens, and Mahoney 2018).

Lemma 3 (Bias-Variance Decomposition)**.** *The in-sample risk defined of the DC solution can be decomposed as*

$$R(\mathbf{w}_{dc}) \quad = \quad \mathsf{bias}^2(\mathbf{w}_{dc}) + \mathsf{var}(\mathbf{w}_{dc}).$$

The bias and variance terms are

$$\mathsf{bias}_{dc}(\mathbf{X}) = \gamma\sqrt{n}\Big\|\frac{1}{m}\sum_{i=1}^{m}\big(\mathbf{\Sigma}\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U}\mathbf{\Sigma} + n\gamma\mathbf{I}_r\big)^\dagger\mathbf{\Sigma}\mathbf{V}^T\mathbf{w}_0\Big\|_2,$$

$$\mathsf{var}_{dc}(\mathbf{X}) = \frac{\sigma^2}{n}\Big\|\frac{1}{m}\sum_{i=1}^{m}\big(\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\Big\|_F^2,$$

where $\mathbf{S}_1, \cdots, \mathbf{S}_m$ are the uniform sampling matrices capturing the random partition of training samples.

Theorem 4 shows that when the local sample size, $s \triangleq \frac{n}{m}$, is sufficiently large, the bias of the DC solution is comparable to the ERM solution.

Theorem 4 (Analysis of Bias)**.** *Let d_{eff}^γ be the γ-effective dimension of $\mathbf{X}$. Let μ^γ be the γ-row coherence of $\mathbf{X}$. Assume*

$$s \;\triangleq\; \tfrac{n}{m} \;=\; \Theta\Big(\tfrac{\mu^\gamma d_{\mathit{eff}}^\gamma}{\epsilon^2}\log\tfrac{m d_{\mathit{eff}}^\gamma}{\delta}\Big)$$

for some parameters $\epsilon, \delta \in (0,1)$. Then

$$\mathsf{bias}_{dc}(\mathbf{X}) \;\le\; \tfrac{1}{1-\epsilon}\,\mathsf{bias}_{\mathit{erm}}(\mathbf{X}).$$

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

Proof. Let $r = \mathrm{rank}(\mathbf{X})$ and $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the SVD. The bias can be bounded by

$$\begin{aligned}
&\Big\|\frac{1}{m}\sum_{i=1}^{m}\big(\mathbf{\Sigma}\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U}\mathbf{\Sigma} + n\gamma\mathbf{I}_r\big)^\dagger\mathbf{\Sigma}\mathbf{V}^T\mathbf{w}_0\Big\|_2^2\\
&\le\; \frac{1}{m}\sum_{i=1}^{m}\Big\|\big(\mathbf{\Sigma}\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U}\mathbf{\Sigma} + n\gamma\mathbf{I}_r\big)^\dagger\mathbf{\Sigma}\mathbf{V}^T\mathbf{w}_0\Big\|_2^2\\
&=\; \frac{1}{m}\sum_{i=1}^{m}\mathbf{w}_0^T\mathbf{V}\mathbf{\Sigma}\big(\mathbf{\Sigma}\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U}\mathbf{\Sigma} + n\gamma\mathbf{I}_r\big)^{\dagger 2}\mathbf{\Sigma}\mathbf{V}^T\mathbf{w}_0.
\end{aligned}$$

Since $s = \Theta\big(\frac{\mu^\gamma d_{\mathrm{eff}}^\gamma}{\epsilon^2}\log\frac{m d_{\mathrm{eff}}^\gamma}{\delta}\big)$, Lemma 1 ensures that with probability at least $1-\frac{\delta}{m}$,

$$\begin{aligned}
&(1-\epsilon)\big(\mathbf{\Sigma}^2 + n\gamma\mathbf{I}_r\big)\\
&\preceq \mathbf{\Sigma}\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U}\mathbf{\Sigma} + n\gamma\mathbf{I}_r \preceq (1+\epsilon)\big(\mathbf{\Sigma}^2 + n\gamma\mathbf{I}_r\big),
\end{aligned}$$

for any $i \in [m]$. It follows that with probability at least $1-\delta$,

$$\begin{aligned}
&\Big\|\frac{1}{m}\sum_{i=1}^{m}\big(\mathbf{\Sigma}\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U}\mathbf{\Sigma} + n\gamma\mathbf{I}_r\big)^\dagger\mathbf{\Sigma}\mathbf{V}^T\mathbf{w}_0\Big\|_2^2\\
&\le\; \frac{1}{(1-\epsilon)^2}\frac{1}{m}\sum_{i=1}^{m}\mathbf{w}_0^T\mathbf{V}\mathbf{\Sigma}\big(\mathbf{\Sigma}^2 + n\gamma\mathbf{I}_r\big)^{-2}\mathbf{\Sigma}\mathbf{V}^T\mathbf{w}_0\\
&\le\; \frac{1}{(1-\epsilon)^2}\Big\|\big(\mathbf{\Sigma}^2 + n\gamma\mathbf{I}_r\big)^{-1}\mathbf{\Sigma}\mathbf{V}^T\mathbf{w}_0\Big\|_2^2.
\end{aligned}$$

It follows from the definition of $\mathsf{bias}_{\mathrm{dc}}(\mathbf{X})$ that with probability at least $1-\delta$,

$$\begin{aligned}
\mathsf{bias}_{\mathrm{dc}}(\mathbf{X}) \;&\le\; \tfrac{\gamma\sqrt{n}}{1-\epsilon}\Big\|\big(\mathbf{\Sigma}^2 + n\gamma\mathbf{I}_r\big)^{-1}\mathbf{\Sigma}\mathbf{V}^T\mathbf{w}_0\Big\|_2\\
&=\; \tfrac{1}{1-\epsilon}\,\mathsf{bias}_{\mathrm{erm}}(\mathbf{X}),
\end{aligned}$$

by which the theorem follows. □

Theorem 5 establishes a relative-error bound on the variance of the DC solution. It states that if the local sample size, $s = \frac{n}{m}$, is sufficiently large compared to the effective dimension, d_{eff}^γ, then $\mathsf{var}(\mathbf{w}_{\mathrm{dc}})$ is comparable to $\mathsf{var}(\mathbf{w}_{\mathrm{erm}})$. The required sample size weakly depends on d, which may not be an issue when $n > md$.

Theorem 5 (Analysis of Variance)**.** *Let d_{eff}^γ be the γ-effective dimension of $\mathbf{X}$. Let μ^γ be the γ-row coherence of $\mathbf{X}$. Assume*

$$s \;\triangleq\; \tfrac{n}{m} \;=\; \Theta\Big(\tfrac{\mu^\gamma d_{\mathit{eff}}^\gamma}{\epsilon^2}\log\tfrac{m d_{\mathit{eff}}^\gamma}{\delta} \;+\; \mu^0 d\log\tfrac{md}{\delta}\Big)$$

for some parameters $\epsilon, \delta \in (0,1)$. Then

$$\mathsf{var}_{dc}(\mathbf{X}) \;\le\; \big(1+\epsilon\sqrt{m}\big)^2 \;\cdot\; \mathsf{var}_{\mathit{erm}}(\mathbf{X})$$

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

Proof. Let $r = \mathrm{rank}(\mathbf{X})$ and $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the SVD. We define $\mathbf{\Delta}_i = \big(\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger - \big(\mathbf{I}_r + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger$. The variance can be bounded by

$$\begin{aligned}
\tfrac{\sqrt{n}}{\sigma}\sqrt{\mathsf{var}_{\mathrm{dc}}(\mathbf{X})} \;&=\; \Big\|\frac{1}{m}\sum_{i=1}^{m}\big(\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\Big\|_F\\
&=\; \Big\|\frac{1}{m}\sum_{i=1}^{m}\big(\mathbf{I}_r + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T \;+\; \frac{1}{m}\sum_{i=1}^{m}\mathbf{\Delta}_i\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\Big\|_F\\
&\le\; \Big\|\big(\mathbf{I}_r + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\Big\|_F\Big\|\frac{1}{m}\sum_{i=1}^{m}\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\Big\|_2\\
&\qquad + \frac{1}{m}\sum_{i=1}^{m}\big\|\mathbf{\Delta}_i\mathbf{U}^T\mathbf{S}_i\big\|_F\big\|\mathbf{S}_i\big\|_2.
\end{aligned}$$

Define $\mathbf{S} \triangleq \frac{1}{m}[\mathbf{S}_1, \cdots, \mathbf{S}_m] \in \mathbb{R}^{n\times n}$ be the scaled concatenation of all the uniform sampling matrices. It can be verified that $\mathbf{S}$ is obtained by permuting the columns of the identity matrix $\mathbf{I}_n$. Thus

$$\Big\|\frac{1}{m}\sum_{i=1}^{m}\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\Big\|_2 \;=\; \big\|\mathbf{U}^T\mathbf{S}\mathbf{S}^T\big\|_2 \;=\; \big\|\mathbf{U}\big\|_2 \;=\; 1.$$

Because $\mathbf{S}_i$ is uniform sampling matrix, each of its nonzero entry equals to $\sqrt{\frac{n}{s}} = \sqrt{m}$, and thus $\|\mathbf{S}_i\|_2 = \sqrt{m}$. It follows that

$$\begin{aligned}
&\tfrac{\sqrt{n}}{\sigma}\sqrt{\mathsf{var}_{\mathrm{dc}}(\mathbf{X})}\\
&\le\; \Big\|\big(\mathbf{I}_r + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\Big\|_F + \frac{1}{m}\sum_{i=1}^{m}\sqrt{m}\big\|\mathbf{\Delta}_i\mathbf{U}^T\mathbf{S}_i\big\|_F. \qquad (10)
\end{aligned}$$

Since $s = \Theta\big(\frac{\mu^\gamma d_{\mathrm{eff}}^\gamma}{\epsilon^2}\log\frac{n d_{\mathrm{eff}}^\gamma}{\delta}\big)$, Lemma 1 guarantees that with probability at least $1-\frac{\delta}{m}$, $\mathbf{X}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{X} + n\gamma\mathbf{I}_d$ is within $(1\pm\epsilon)(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)$ for any i. By writing $\mathbf{X}$ as its SVD form $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, we have that $\big(\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger$ is within $\frac{1}{1\mp\epsilon}\big(\mathbf{I}_r + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger$, and thus

$$\tfrac{\epsilon}{1+\epsilon}\big(\mathbf{I}_r + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger \;\preceq\; \mathbf{\Delta}_i \;\preceq\; \tfrac{\epsilon}{1-\epsilon}\big(\mathbf{I}_r + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger$$

with probability at least $1-\frac{\delta}{m}$. Thus

$$\begin{aligned}
\big\|\mathbf{\Delta}_i\mathbf{U}^T\mathbf{S}_i\big\|_F^2 \;&=\; \mathrm{tr}\big(\mathbf{S}_i^T\mathbf{U}\mathbf{\Delta}_i^2\mathbf{U}^T\mathbf{S}_i\big)\\
&\le\; \big(\tfrac{\epsilon}{1-\epsilon}\big)^2\mathrm{tr}\big[\mathbf{S}_i^T\mathbf{U}\big(\big(\mathbf{I}_r + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\big)^2\mathbf{U}^T\mathbf{S}_i\big]\\
&=\; \big(\tfrac{\epsilon}{1-\epsilon}\big)^2\big\|\big(\mathbf{I}_r + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\mathbf{U}^T\mathbf{S}_i\big\|_F^2
\end{aligned}$$

holds with probability at least $1 - \frac{\delta}{m}$. Lemma 2 show that when $s = \Theta\big(\mu^0 d \log \frac{md}{\delta}\big)$, it holds with probability at least $1 - \frac{\delta}{m}$ that $\|\mathbf{U}^T \mathbf{S}_i\|_2 \leq 2$ for any i. It follows that

$$\big\|\mathbf{\Delta}_i \mathbf{U}^T \mathbf{S}_i\big\|_F \;\leq\; \tfrac{2\epsilon}{1-\epsilon} \big\|\big(\mathbf{I}_r + n\gamma \mathbf{\Sigma}^{-2}\big)^\dagger\big\|_F.$$

It follows from (10) that

$$\tfrac{\sqrt{n}}{\sigma} \sqrt{\mathsf{var}_{\text{dc}}(\mathbf{X})} \;\leq\; \big(1 + \tfrac{2\epsilon\sqrt{m}}{1-\epsilon}\big) \big\|\big(\mathbf{I}_r + n\gamma \mathbf{\Sigma}^{-2}\big)^\dagger\big\|_F$$

holds with probability at least $1 - 2\delta$. The theorem follows by the definition of $\mathsf{var}_{\text{erm}}(\mathbf{X})$. □

Main Theorem

Theorem 6 shows that if each partition of the data is sufficiently large, then $R_{\text{dc}}(\mathbf{X})$ is comparable to the optimal in-sample risk $R_{\text{dc}}(\mathbf{X})$.

Theorem 6 (In-Sample Risk)**.** *The in-sample risk (defined in* (3)*) of the divide-and-conquer ridge regression can be decomposed as*

$$R_{dc}(\mathbf{X}) \quad = \quad \mathsf{bias}_{dc}^2(\mathbf{X}) + \mathsf{var}_{dc}(\mathbf{X}).$$

Let $\epsilon, \delta \in (0,1)$ be arbitrary constants. Assume the n samples are randomly and uniformly partitioned among m machines and

$$s \;\triangleq\; \tfrac{n}{m} \;=\; \Theta\Big(\tfrac{\mu^\gamma d_{\mathit{eff}}^\gamma m}{\epsilon^2} \log \tfrac{m d_{\mathit{eff}}^\gamma}{\delta} + \mu^0 d \log \tfrac{md}{\delta}\Big)$$

Then with probability at least $1 - \delta$,

$$\begin{aligned} \mathsf{bias}_{dc}(\mathbf{X}) \quad &\leq \quad (1+\epsilon) \cdot \mathsf{bias}_{\mathit{erm}}(\mathbf{X}), \\ \mathsf{var}_{dc}(\mathbf{X}) \quad &\leq \quad \big(1+\epsilon\big)^2 \cdot \mathsf{var}_{\mathit{erm}}(\mathbf{X}). \end{aligned}$$

Here $\mathsf{bias}_{\mathit{erm}}(\mathbf{X})$ and $\mathsf{var}_{\mathit{erm}}(\mathbf{X})$ are defined in (9)*.*

Proof. The theorem is the combination of Lemma 3 and Theorems 4 and 5. □

A similar but weaker result was shown in (Wang, Gittens, and Mahoney 2018). To get the same bound on $\mathsf{bias}_{\text{dc}}(\mathbf{X})$ and $\mathsf{var}_{\text{dc}}(\mathbf{X})$ as Theorem 6, they require a sample complexity of

$$s \;\triangleq\; \tfrac{n}{m} \;=\; \Theta\big(\tfrac{\mu^0 d}{\epsilon^2} \log \tfrac{md}{\delta}\big).$$

However, in high-dimensional problems (i.e., d is large), n may not be sufficiently larger than md, and thus their result is not very meaningful.

In contrast, our sample complexity mainly depends on the effective dimension, d_{eff}^γ, which is a small constant if the regularization parameter γ is not too small. even if md is comparable to n, our bound ensures that $R_{\text{dc}}(\mathbf{X})$ is close to $R_{\text{erm}}(\mathbf{X})$.

Analysis of Generalization

In this section, we analyze the generalization. To be specific, we show the gap between the in-sample risk R and the out-of-sample prediction error P. Theorem 9 is the main theorem of this section. We prove the theorem using random matrix theories. Our analysis of generalization is reminiscent of (Hsu, Kakade, and Zhang 2014).

Convergence of the Second Moments

Let $\mathbf{x}_1, \cdots, \mathbf{x}_n$ be the training feature vectors and $\mathbf{M} = \mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^T] \in \mathbb{R}^{d\times d}$ be the second moment. Here we show that the empirical second moment

$$\widehat{\mathbf{M}} \;=\; \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$$

converges to $\mathbf{M}$ at a rate of $\frac{1}{\sqrt{n}}$. Before presenting our result, we define the coherence of probabilistic distribution.

Definition 1 (Coherence of Distribution)**.** *Let the training and test samples have the same probability density function p. Let $\mathbf{M} = \mathbb{E}_{\mathbf{x}\sim p}[\mathbf{x}\mathbf{x}^T] \in \mathbb{R}^{d\times d}$ be the second moment and $\mathcal{D}$ be the set where p has nonzero measure. Let $\tau = \frac{1}{d} \sup_{\mathbf{x}\in\mathcal{D}} \mathbf{x}^T \mathbf{M}^{-1} \mathbf{x}$ be the coherence of p.*

The coherence of distribution is a new notation made by this paper for analyzing generalization. It is different from but analogous to the standard coherence which is defined by

$$\mu^0 \;=\; \tfrac{1}{d} \sup_{j\in[n]} \mathbf{x}_j^T (\tfrac{1}{n}\mathbf{X}^T\mathbf{X})^\dagger \mathbf{x}_j.$$

The standard coherence μ^0 converges to τ as $n \to \infty$.

Theorem 7. *Let τ be the coherence in Definition 1. For any $\eta, \delta \in (0,1)$, assume the number of samples, n, is sufficiently large:*

$$n \;\geq\; \tfrac{10\tau d}{3\eta^2} \log \tfrac{d}{\delta}.$$

Then $(1-\eta)\mathbf{M} \preceq \widehat{\mathbf{M}} \preceq (1+\eta)\mathbf{M}$ holds with probability at least $1 - \delta$.

Proof. Let $\mathbf{M} = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^T$ be the eigenvalue decomposition of $\mathbf{M} \in \mathbb{R}^{d\times d}$. Assume every training or test sample is generated by $\mathbf{x} = \mathbf{V}\mathbf{\Lambda}\mathbf{u}$ where $\mathbf{u} \in \mathbb{R}^d$ is random. Obviously, to enforce $\mathbb{E}[\mathbf{x}\mathbf{x}^T] = \mathbf{M}$, we assume $\mathbb{E}[\mathbf{u}\mathbf{u}^T] = \mathbf{I}_d$. The coherence can be equivalently written as $\tau = \frac{1}{d} \sup_{\mathbf{u}} \|\mathbf{u}\|_2^2$.

We let $\mathbf{Z}_i = \frac{1}{n}(\mathbf{u}_i\mathbf{u}_i^T - \mathbf{I}_d)$. The spectral norm of $\mathbf{Z}_i$ can be bounded by

$$\big\|\mathbf{Z}_i\big\|_2 \;\leq\; \tfrac{1}{n}\big(\|\mathbf{u}_i\|_2^2 + 1\big) \;\leq\; \tfrac{\tau d + 1}{n} \;\triangleq\; L.$$

The second moment of $\mathbf{Z}_i$ can be bounded by

$$\begin{aligned} \mathbb{E}\big[\mathbf{Z}_i^2\big] \quad &= \quad \tfrac{1}{n^2}\mathbb{E}\big[\mathbf{u}_i\mathbf{u}_i^T\mathbf{u}_i\mathbf{u}_i^T - \mathbf{I}_d\big] \\ &\preceq \quad \tfrac{1}{n^2}\mathbb{E}\big[\|\mathbf{u}_i\|_2^2 \mathbf{u}_i\mathbf{u}_i^T\big] \;\preceq\; \tfrac{\tau d}{n^2}\mathbf{I}_d. \end{aligned}$$

Let $\mathbf{Y} = \sum_{i=1}^n \mathbf{Z}_i$. Then the second moment of $\mathbf{Y}_i$ can be bounded by

$$\Big\|\mathbb{E}\big[\mathbf{Y}^2\big]\Big\|_2 \;=\; \Big\|\sum_{i=1}^n \mathbb{E}\big[\mathbf{Z}_i^2\big]\Big\|_2 \;\leq\; \frac{\tau d}{n} \;\triangleq\; v(\mathbf{Y}).$$

We can bound the spectral norm of $\mathbf{Y}$ using the matrix Bernstein. Elementary proof of the matrix Bernstein can be found in (Tropp and others 2015).

Lemma 8 (Matrix Bernstein)**.** *Consider a finite sequence $\{\mathbf{Z}_i\}$ of independent, random, Hermitian matrices with dimension d. Assume that*

$$\mathbb{E}\mathbf{Z}_i = \mathbf{0} \quad \textit{and} \quad \max_i \|\mathbf{Z}_i\|_2 \leq L$$

for each index i. Introduce the random matrix $\mathbf{Y} = \sum_i \mathbf{Z}_i$. Let $v(\mathbf{Y})$ be the matrix variance statistics of the sum:

$$v(\mathbf{Y}) \;=\; \left\|\mathbb{E}\mathbf{Y}^2\right\|_2 \;=\; \left\|\sum_i \mathbb{E}\mathbf{Z}_i^2\right\|_2.$$

Then

$$\mathbb{P}\{\lambda_{\max}(\mathbf{Y}) \geq t\} \quad \leq \quad d \cdot \exp\left(\tfrac{-t^2/2}{v(\mathbf{Y})+Lt/3}\right).$$

It follows the matrix Bernstein that

$$\begin{aligned}\mathbb{P}\{\lambda_{\max}(\mathbf{Y}) \geq \eta\} \quad &\leq \quad d \cdot \exp\left(\tfrac{-\eta^2/2}{v+L\eta/3}\right) \\ &= \quad d \cdot \exp\left(\tfrac{-n}{\tau d(\frac{2}{\eta^2}+\frac{4}{3\eta})}\right) \;\triangleq\; \delta.\end{aligned}$$

Therefore, for $n \geq \frac{10\tau d}{3\eta^2}\log\frac{d}{\delta}$, the spectral norm of $\mathbf{Y}$ is bounded by η. Because $\mathbf{Y} = \sum_{i=1}^n \mathbf{Z}_i = \frac{1}{n}\sum_{i=1}^n \mathbf{u}_i\mathbf{u}_i^T - \mathbf{I}_d$, we have

$$(1-\eta)\mathbf{I}_d \;\preceq\; \frac{1}{n}\sum_{i=1}^n \mathbf{u}_i\mathbf{u}_i^T \;\preceq\; (1+\eta)\mathbf{I}_d.$$

Since $\mathbf{x}_i = \mathbf{V}\mathbf{\Lambda}\mathbf{u}_i$, we have $\mathbf{x}_i\mathbf{x}_i^T = \mathbf{V}\mathbf{\Lambda}\mathbf{u}_i\mathbf{u}_i^T\mathbf{\Lambda}\mathbf{Z}^T$, and thus

$$(1-\eta)\mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^T \;\preceq\; \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T \;\preceq\; (1+\eta)\mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^T.$$

The theorem follows by the above inequality and the definition $\mathbf{M} = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^T$. □

Main Theorem

Theorem 9 is the main theorem of this section. It shows that the gap between the in-sample risk and the out-of-sample prediction error vanishes at a rate of $\frac{1}{\sqrt{n}}$.

Theorem 9. *Let τ be the coherence in Definition 1. Let $\eta, \delta \in (0,1)$ be arbitrary constants. Assume $n \geq \frac{10\tau d}{3\eta^2}\log\frac{d}{\delta}$. Then $P_{erm}(\mathbf{X}) \leq \frac{1}{1-\eta}R_{erm}(\mathbf{X})$ and $P_{dc}(\mathbf{X}) \leq \frac{1}{1-\eta}R_{dc}(\mathbf{X})$ both hold with probability at least $1-\delta$.*

Proof. The in-sample statistical risk can be written as

$$\begin{aligned}R_{\text{erm}}(\mathbf{X}) \quad &= \quad \tfrac{1}{n}\mathbb{E}_{\mathbf{y}}\left\|\mathbf{X}\mathbf{w}_{\text{erm}} - \mathbf{X}\mathbf{w}_0\right\|_2^2 \\ &= \quad \mathbb{E}_{\mathbf{y}}\left[(\mathbf{w}_{\text{erm}} - \mathbf{w}_0)^T\widehat{\mathbf{M}}(\mathbf{w}_{\text{erm}} - \mathbf{w}_0)\right],\end{aligned}$$

where the expectation is taken w.r.t. the random noise in $\mathbf{y}$. The out-of-sample prediction error can be written as

$$\begin{aligned}P_{\text{erm}}(\mathbf{X}) \quad &= \quad \mathbb{E}_{\mathbf{x}',\mathbf{y}}\left[\left(\mathbf{w}_{\text{erm}}^T\mathbf{x}' - \mathbf{w}_0^T\mathbf{x}'\right)^2\right] \\ &= \quad \mathbb{E}_{\mathbf{y}}\left[(\mathbf{w}_{\text{erm}} - \mathbf{w}_0)^T\mathbf{M}(\mathbf{w}_{\text{erm}} - \mathbf{w}_0)\right],\end{aligned}$$

where the expectation is taken w.r.t. the randomness in the test feature vector $\mathbf{x}'$. Theorem 7 shows that $(1-\eta)\mathbf{M} \preceq \widehat{\mathbf{M}}$. Hence $(1-\eta)P_{\text{erm}}(\mathbf{X}) \leq R_{\text{erm}}(\mathbf{X})$. We can prove $(1-\eta)P_{\text{dc}}(\mathbf{X}) \leq R_{\text{dc}}(\mathbf{X})$ in the same way. □

Conclusions and Future Work

We studied the divide-and-conquer (DC) method for ridge regression and established a strong generalization bound. If the total number of samples is $n = \tilde{\Theta}(m^2 d_{\text{eff}}^{\gamma}/\epsilon^2 + md + d/\epsilon^2)$, where m is the number of partitions, d_{eff}^{γ} ($\leq d$) is the effective dimension, and $\tilde{\Theta}$ hides logarithms and coherence parameters, then the out-of-sample prediction error of the DC solution is ϵ times worse than the optimal in-sample error. In contrast, the prior works (Zhang, Duchi, and Wainwright 2015; Lin, Guo, and Zhou 2017) established constant-factor bounds which are worse than our $1+\epsilon$ bound. In addition, their sample complexities are much worse than ours: they require n to be at least quadratic with the number of features, d.

Our result is not directly applicable to the divide-and-conquer kernel ridge regression (DC-KRR) (Zhang, Duchi, and Wainwright 2015; Lin, Guo, and Zhou 2017). Because the feature space of kernel method is high-dimenional or even infinite-dimensional, directly following our theory will result in a too high sample complexity. Developing elegant and strong generalization bound for DC-KRR without making uncheckable assumptions will be the future work.

Acknowledgment

The author thanks the very helpful suggestions given by the three anonymous reviewers and Miles Lopes.

References

Avron, H.; Clarkson, K. L.; and Woodruff, D. P. 2016. Sharper bounds for regression and low-rank approximation with regularization. *CoRR, abs/1611.03225*.

Avron, H.; Maymounkov, P.; and Toledo, S. 2010. Blendenpik: Supercharging LAPACK's least-squares solver. *SIAM Journal on Scientific Computing* 32(3):1217–1236.

Bonawitz, K.; Ivanov, V.; Kreuter, B.; Marcedone, A.; McMahan, H. B.; Patel, S.; Ramage, D.; Segal, A.; and Seth, K. 2017. Practical secure aggregation for privacy preserving machine learning. *IACR Cryptology ePrint Archive* 2017:281.

Candes, E. J., and Recht, B. 2009. Exact matrix completion via convex optimization. *Foundations of Computational mathematics* 9(6):717.

Candes, E. J., and Tao, T. 2006. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory* 52(12):5406–5425.

Clarkson, K. L., and Woodruff, D. P. 2013. Low rank approximation and regression in input sparsity time. In *Annual ACM Symposium on Theory of Computing (STOC)*.

Cohen, M. B.; Musco, C.; and Musco, C. 2015. Ridge leverage scores for low-rank approximation. *arXiv preprint arXiv:1511.07263* 6.

Dean, J., and Ghemawat, S. 2008. MapReduce: simplified data processing on large clusters. *Communications of the ACM* 51(1):107–113.

Derezinski, M., and Warmuth, M. K. 2018. Subsampling for ridge regression via regularized volume sampling. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Drineas, P., and Mahoney, M. W. 2016. RandNLA: randomized numerical linear algebra. *Communications of the ACM* 59(6):80–90.

Drineas, P.; Mahoney, M. W.; Muthukrishnan, S.; and Sarlós, T. 2011. Faster least squares approximation. *Numerische Mathematik* 117(2):219–249.

Drineas, P.; Mahoney, M. W.; and Muthukrishnan, S. 2006. Sampling algorithms for ℓ_2 regression and applications. In *Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*.

Hsu, D.; Kakade, S.; and Zhang, T. 2014. Random design analysis of ridge regression. *Foundations of Computational Mathematics* 14(3).

Konecnỳ, J.; McMahan, H. B.; Ramage, D.; and Richtárik, P. 2016a. Federated optimization: distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*.

Konecnỳ, J.; McMahan, H. B.; Yu, F. X.; Richtárik, P.; Suresh, A. T.; and Bacon, D. 2016b. Federated learning: strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Li, M.; Andersen, D. G.; Park, J. W.; Smola, A. J.; Ahmed, A.; Josifovski, V.; Long, J.; Shekita, E. J.; and Su, B.-Y. 2014. Scaling distributed machine learning with the parameter server. In *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*.

Lin, S.-B.; Guo, X.; and Zhou, D.-X. 2017. Distributed learning with regularized least squares. *Journal of Machine Learning Research* 18(1):3202–3232.

Lu, Y.; Dhillon, P.; Foster, D. P.; and Ungar, L. 2013. Faster ridge regression via the subsampled randomized hadamard transform. In *Advances in Neural Information Processing Systems (NIPS)*.

Mahajan, D.; Agrawal, N.; Keerthi, S. S.; Sundararajan, S.; and Bottou, L. 2013. An efficient distributed learning algorithm based on effective local functional approximations. *arXiv preprint arXiv:1310.8418*.

Mahoney, M. W. 2011. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning* 3(2):123–224.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Meng, X.; Bradley, J.; Yavuz, B.; Sparks, E.; Venkataraman, S.; Liu, D.; Freeman, J.; Tsai, D.; Amde, M.; Owen, S.; Xin, D.; Xin, R.; Franklin, M. J.; Zadeh, R.; Zaharia, M.; and Talwalkar, A. 2016. MLlib: Machine Learning in Apache Spark. *Journal of Machine Learning Research* 17(34):1–7.

Meng, X.; Saunders, M. A.; and Mahoney, M. W. 2014. LSRN: A parallel iterative solver for strongly over-or underdetermined systems. *SIAM Journal on Scientific Computing* 36(2):C95–C118.

Raskutti, G., and Mahoney, M. W. 2015. Statistical and algorithmic perspectives on randomized sketching for ordinary least-squares. In *International Conference on Machine Learning (ICML)*.

Shamir, O.; Srebro, N.; and Zhang, T. 2014. Communication-efficient distributed optimization using an approximate Newton-type method. In *International conference on machine learning (ICML)*.

Smith, V.; Forte, S.; Ma, C.; Takac, M.; Jordan, M. I.; and Jaggi, M. 2016. CoCoA: A general framework for communication-efficient distributed optimization. *arXiv preprint arXiv:1611.02189*.

Smith, V.; Chiang, C.-K.; Sanjabi, M.; and Talwalkar, A. 2017. Federated multi-task learning. *arXiv preprint arXiv:1705.10467*.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Tropp, J. A., et al. 2015. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning* 8(1-2):1–230.

Wang, J.; Lee, J. D.; Mahdavi, M.; Kolar, M.; and Srebro, N. 2017. Sketching Meets Random Projection in the Dual: a Provable Recovery Algorithm for Big and High-Dimensional Data. *Electronic Journal of Statistics* 11(2):4896–4944.

Wang, S.; Roosta-Khorasani, F.; Xu, P.; and Mahoney, M. W. 2018. GIANT: Globally improved approximate Newton method for distributed optimization. In *Conference on Neural Information Processing Systems (NIPS)*.

Wang, S.; Gittens, A.; and Mahoney, M. W. 2018. Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. *Journal of Machine Learning Research* 18(218):1–50.

Wang, S.; Luo, L.; and Zhang, Z. 2016. SPSD matrix approximation vis column selection: Theories, algorithms, and extensions. *Journal of Machine Learning Research* 17(49):1–49.

Woodruff, D. P. 2014. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science* 10(1–2):1–157.

Zaharia, M.; Chowdhury, M.; Franklin, M. J.; Shenker, S.; and Stoica, I. 2010. Spark: Cluster computing with working sets. *HotCloud* 10(10-10):95.

Zhang, Y.; Duchi, J. C.; and Wainwright, M. J. 2013. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research* 14:3321–3363.

Zhang, Y.; Duchi, J.; and Wainwright, M. 2015. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research* 16:3299–3340.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301–320.