Least-Squares Regression on Sparse Spaces

Yuri Grinberg, Mahdi Milani Fard, Joelle Pineau
School of Computer Science
McGill University
Montreal, Canada
{ygrinb,mmilan1,jpineau}@cs.mcgill.ca

1 Introduction

Compressed sampling has been studied in the context of regression theory from two prospectives. One is when given a training set, we aim to compress the set into a smaller size by combining training instances using random projections (see e.g. [1]). Such method is useful, for instance, when the training set is too large or one has to handle privacy issues.

Another application is when one uses random projections to project each input vector into a lower dimensional space, and then train a predictor in the new compressed space (compression on the feature space). As is typical of dimensionality reduction techniques, this will reduce the variance of most predictors at the expense of introducing some bias. Random projections on the feature space, along with least-squares predictors are studied in [2], and the method is shown to reduces the estimation error at the price of a controlled approximation error. The analysis in [2] provides on-sample error bounds and extends them to bounds on the sampling measure, assuming an i.i.d. sampling strategy.

This paper includes the bias–variance analysis of regression in compressed spaces when random projections are applied on sparse input signals. We show that the sparsity assumption let us work with arbitrary non i.i.d. sampling strategies and we derive a worst-case bound on the entire space. Such a bound can be used to select the optimal size of projection, such as to minimize the sum of expected estimation and prediction errors. It also provides the means to compare the error of linear predictors in the original and compressed spaces.

2 Notations and Sparsity Assumption

Throughout this paper, column vectors are represented by lower case bold letters, and matrices are represented by bold capital letters. $|.|$ denotes the size of a set, and $\|.\|_0$ is Donoho's zero "norm" indicating the number of non-zero elements in a vector. $\|.\|$ denotes the L^2 norm for vectors and the operator norm for matrices: $\|\mathbf{M}\| = \sup_{\mathbf{v}} \|\mathbf{M}\mathbf{v}\|/\|\mathbf{v}\|$. Also, we denote the Moore-Penrose pseudo-inverse of a matrix $\mathbf{M}$ with $\mathbf{M}^\dagger$ and the smallest singular value of $\mathbf{M}$ by $\sigma_{\min}^{(M)}$.

We will be working in sparse input spaces for our prediction task. Our input is represented by a vector $\mathbf{x} \in \mathcal{X}$ of D features, having $\|\mathbf{x}\| \leq 1$. We assume that $\mathbf{x}$ is k-sparse in some known or unknown basis $\mathbf{\Psi}$, implying that $\mathcal{X} \triangleq \{\mathbf{\Psi}\mathbf{z}, \text{ s.t. } \|\mathbf{z}\|_0 \leq k \text{ and } \|\mathbf{z}\| \leq 1\}$.

3 Random Projections and Inner Product

It is well known that random projections of appropriate sizes preserve enough information for exact reconstruction with high probability (see e.g. [3, 4]). In this section, we show that a function (almost-)linear in the original space is almost linear in the projected space, when we have random projections of appropriate sizes.

There are several types of random projection matrices that can be used. In this work, we assume that each entry in a projection $\mathbf{\Phi}^{D\times d}$ is an i.i.d. sample from a Gaussian [1]:

$$\phi_{i,j} = \mathcal{N}(0, 1/d). \tag{1}$$

We build our work on the following (based on theorem 4.1 from [3]), which shows that for a finite set of points, inner product with a fixed vector is almost preserved after a random projection.

Theorem 1. *Let $\mathbf{\Phi}^{D\times d}$ be a random projection according to Eqn 1. Let S be a finite set of points in $\mathbb{R}^D$. Then for any fixed $\mathbf{w} \in \mathbb{R}^D$ and $\epsilon > 0$:*

$$\forall \mathbf{s} \in S : \left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{s}\rangle - \langle \mathbf{w}, \mathbf{s}\rangle\right| \le \epsilon\|\mathbf{w}\|\|\mathbf{s}\|, \tag{2}$$

fails with probability less than $(4|S| + 2)e^{-d\epsilon^2/48}$.

The above theorem is based on the well-known Johnson–Lindenstrauss lemma (see [3]), which considers random projections of finite sets of points. We derive the corresponding theorem for sparse feature spaces.

Theorem 2. *Let $\mathbf{\Phi}^{D\times d}$ be a random projection according to Eqn 1. Let $\mathcal{X}$ be a D-dimensional k-sparse space. Then for any fixed $\mathbf{w}$ and $\epsilon > 0$:*

$$\forall \mathbf{x} \in \mathcal{X} : \left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{x}\rangle - \langle \mathbf{w}, \mathbf{x}\rangle\right| \le \epsilon\|\mathbf{w}\|\|\mathbf{x}\|, \tag{3}$$

fails with probability less than:

$$(eD/k)^k(4(12/\epsilon)^k + 2)e^{-d\epsilon^2/192} \le e^{k\log(12eD/\epsilon k) - d\epsilon^2/192 + \log 5}.$$

Note that the above theorem does not require w to be in the sparse space, and thus is different from guarantees on the preservation of inner product between vectors in the sparse space.

Proof of Theorem 2. The proof follows the steps of the proof of theorem 5.2 from [5]. Because $\mathbf{\Phi}$ is a linear transformation, we only need to prove the theorem when $\|\mathbf{w}\| = \|\mathbf{x}\| = 1$.

Denote $\mathbf{\Psi}$ to be the basis with respect to which $\mathcal{X}$ is sparse. Let $T \subset \{1, 2, \ldots, D\}$ be any set of k indexes. For each set of indexes T, we define a k-dimensional hyperplane in the D-dimensional input space: $\mathcal{X}_T \triangleq \{\mathbf{\Psi}\mathbf{z},\ \text{s.t. } \mathbf{z} \text{ is zero outside } T \text{ and } \|\mathbf{z}\| \le 1\}$. By definition we have $\mathcal{X} = \cup_T \mathcal{X}_T$. We first show that Eqn 3 holds for each $\mathcal{X}_T$ and then use the union bound to prove the theorem.

For any given T, we choose a set $S \subset \mathcal{X}_T$ such that we have:

$$\forall \mathbf{x} \in \mathcal{X}_T : \min_{\mathbf{s}\in\mathbf{S}} \|\mathbf{x} - \mathbf{s}\| \le \epsilon/4. \tag{4}$$

It is easy to prove (see e.g. Chapter 13 of [6]) that these conditions can be satisfied by choosing a grid of size $|S| \le (12/\epsilon)^k$, since $\mathcal{X}_T$ is a k-dimensional hyperplane in $\mathbb{R}^n$ (S fills up the space within $\epsilon/4$ distance). Now applying Theorem 1, and with $\|\mathbf{w}\| = 1$ we have that:

$$\forall \mathbf{s} \in \mathcal{S} : \left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{s}\rangle - \langle \mathbf{w}, \mathbf{s}\rangle\right| \le \frac{\epsilon}{2}\|\mathbf{s}\|, \tag{5}$$

fails with probability less than $(4(12/\epsilon)^k + 2)e^{-d\epsilon^2/192}$.

Let a be the smallest number such that:

$$\forall \mathbf{x} \in \mathcal{X}_T : \left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{x}\rangle - \langle \mathbf{w}, \mathbf{x}\rangle\right| \le a\|\mathbf{x}\|, \tag{6}$$

holds when Eqn 5 holds. The goal is to show that $a \le \epsilon$. For any given $\mathbf{x} \in \mathcal{X}_T$, we choose an $\mathbf{s} \in S$ for which $\|\mathbf{x} - \mathbf{s}\| \le \epsilon/4$. Therefore we have:

$$\begin{aligned}
\left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{x}\rangle - \langle \mathbf{w}, \mathbf{x}\rangle\right| &\le \left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{x}\rangle - \langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{s}\rangle - \langle \mathbf{w}, \mathbf{x}\rangle + \langle \mathbf{w}, \mathbf{s}\rangle\right| + && (7)\\
&\quad \left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{s}\rangle - \langle \mathbf{w}, \mathbf{s}\rangle\right| && (8)\\
&\le \left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T(\mathbf{x} - \mathbf{s})\rangle - \langle \mathbf{w}, (\mathbf{x} - \mathbf{s})\rangle\right| + && (9)\\
&\quad \left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{s}\rangle - \langle \mathbf{w}, \mathbf{s}\rangle\right| && (10)\\
&\le a\epsilon/4 + \epsilon/2. && (11)
\end{aligned}$$

[1] The elements of the projection are typically taken to be distributed with $\mathcal{N}(0, 1/D)$, but we scale them by $\sqrt{D/d}$, so that we avoid scaling the projected values (see e.g. [3]).

The last line is by the definition of a, and by applying Eqn 5 (with high probability). Because of the definition of a, there is an $\mathbf{x} \in \mathcal{X}_T$ (and by scaling, one with size 1), for which Eqn 6 is tight. Therefore we have $a \leq a\epsilon/4 + \epsilon/2$, which proves $a \leq \epsilon$ for any choice of $\epsilon < 1$.

Note that there are $\binom{D}{k}$ possible sets T. Since $\binom{D}{k} \leq (eD/k)^k$ and $\mathcal{X} = \cup_T \mathcal{X}_T$, the union bound gives us that the theorem fails with probability less than $(eD/k)^k(4(12/\epsilon)^k + 2)e^{-d\epsilon^2/192}$. □

4 Bias–Variance Analysis of Ordinary Least-Squares

In this section, we analyze the worst case prediction error made by the ordinary least-squares (OLS) solution. For completeness, we provide bounds on OLS in the original space (which is partly a classical result in linear prediction theory). Then, we proceed to the main result of this paper, which is the bias–variance analysis of OLS in the projected space.

We seek to predict a signal f that is assumed to be a (near-)linear function of $\mathbf{x} \in \mathcal{X}$:

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + b_f(\mathbf{x}), \text{ where } |b_f(\mathbf{x})| \leq \epsilon_f, \tag{12}$$

for some $\epsilon_f > 0$, where we assume $\|\mathbf{w}\| \leq 1$. We are given a training set of n input–output pairs, consisting of a full-rank input matrix $\mathbf{X}^{n\times D}$, along with noisy observations of f:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{b}_f + \eta, \tag{13}$$

where for the additive bias term (overloading the notation) $\mathbf{b}_{f,i} = b_f(\mathbf{x}_i)$; and we assume a homoscedastic noise term η to be a vector of i.i.d. random variables distributed as $\mathcal{N}(0, \sigma_\eta^2)$.

Given the above, we seek to find a predictor that for any query $\mathbf{x} \in \mathcal{X}$ predicts the target signal $f(\mathbf{x})$. The following lemma provides a bound over worst-case error of the ordinary least-squares predictor.

Lemma 3. *Let $\mathbf{w}_{ols}$ be the OLS solution of Eqn 13 with additive bias bounded by ϵ_f and i.i.d. noise with variance σ_η^2. Then for any $0 < \delta_{var} \leq \sqrt{2/e\pi}$, for all $\mathbf{x} \in \mathcal{X}$, with probability no less than $1 - \delta_{var}$ the error in the OLS prediction follows this bound:*

$$|f(\mathbf{x}) - \mathbf{x}^T\mathbf{w}_{ols}| \leq \|\mathbf{x}\|\|\mathbf{X}^\dagger\| \left(\epsilon_f\sqrt{n} + \sigma_\eta\sqrt{\log(2/\pi\delta_{var}^2)}\right) + \epsilon_f. \tag{14}$$

Proof of Lemma 3. For the OLS solution of Eqn 13 we have:

$$\mathbf{w}_{\text{ols}} \quad = \quad \mathbf{X}^\dagger\mathbf{y} = \mathbf{X}^\dagger(\mathbf{X}\mathbf{w} + \mathbf{b}_f + \eta) = \mathbf{w} + \mathbf{X}^\dagger\mathbf{b}_f + \mathbf{X}^\dagger\eta. \tag{15}$$

Therefore for all $\mathbf{x} \in \mathcal{X}$ we have the error:

$$\begin{aligned} |f(\mathbf{x}) - \mathbf{x}^T\mathbf{w}_{\text{ols}}| &\leq |\mathbf{x}^T\mathbf{w}_{\text{ols}} - \mathbf{x}^T\mathbf{w}| + \epsilon_f & (16)\\ &\leq |\mathbf{x}^T\mathbf{X}^\dagger\mathbf{b}_f| + |\mathbf{x}^T\mathbf{X}^\dagger\eta| + \epsilon_f. & (17) \end{aligned}$$

For the first term (part of prediction bias) on the right hand side, we have:

$$|\mathbf{x}^T\mathbf{X}^\dagger\mathbf{b}_f| \quad \leq \quad \|\mathbf{x}^T\|\|\mathbf{X}^\dagger\|\|\mathbf{b}_f\| \leq \|\mathbf{x}\|\|\mathbf{X}^\dagger\|\epsilon_f\sqrt{n}. \tag{18}$$

For the second term in line 17 (prediction variance), we have that the expectation of $\mathbf{x}^T\mathbf{X}^\dagger\eta$ is 0, as η is independent of data and its expectation is zero. We also know that it is a weighted sum of normally distributed random variables, and thus is normal with the variance:

$$\begin{aligned} \text{Var}[\mathbf{x}^T\mathbf{X}^\dagger\eta] &= \mathbb{E}[\mathbf{x}^T\mathbf{X}^\dagger\eta\eta^T(\mathbf{X}^\dagger)^T\mathbf{x}] & (19)\\ &= \sigma_\eta^2\mathbf{x}^T\mathbf{X}^\dagger(\mathbf{X}^\dagger)^T\mathbf{x} & (20)\\ &\leq \sigma_\eta^2\|\mathbf{x}^T\|\|\mathbf{X}^\dagger\|\|(\mathbf{X}^\dagger)^T\|\|\mathbf{x}\| & (21)\\ &\leq \sigma_\eta^2\|\mathbf{x}\|^2\|\mathbf{X}^\dagger\|^2, & (22) \end{aligned}$$

where in line 20 we used the i.i.d. assumption on the noise. Thereby we can bound $|\mathbf{x}^T\mathbf{X}^\dagger\eta|$ by the tail probability of the normal distribution as needed. Using an standard upper bound on the tail probability of normals, when $0 < \delta_{\text{var}} \leq \sqrt{2/e\pi}$, with probability no less than $1 - \delta_{\text{var}}$:

$$|\mathbf{x}^T\mathbf{X}^\dagger\eta| \quad \leq \quad \sigma_\eta\|\mathbf{x}\|\|\mathbf{X}^\dagger\|\sqrt{\log(2/\pi\delta_{\text{var}}^2)}. \tag{23}$$

Adding up the bias and the variance term gives us the bound in the lemma. □

5 Compressed Ordinary Least-Squares

We are now ready to study an upper bound for the worst-case error of the OLS predictor in a compressed space. In this setting, we will first project the inputs into a lower dimensional space using random projections, and then use the OLS estimator on the compressed input signals.

Theorem 4. *Let $\mathbf{\Phi}^{D\times d}$ be a random projection according to Eqn 1 and $\mathbf{w}_{ols}^{(\Phi)}$ be the OLS solution in the compressed space induced by the projection. Assume an additive bias in the original space bounded by some $\epsilon_f > 0$ and i.i.d. noise with variance σ_η^2. Choose any $0 < \delta_{prj}, \delta_\Phi < 1$ and $0 < \delta_{var} \leq \sqrt{2/e\pi}$. Then, with probability no less than $1 - (\delta_{prj} + \delta_\Phi)$, we have $\forall \mathbf{x} \in \mathcal{X}$ with probability no less than $1 - \delta_{var}$:*

$$|f(\mathbf{x}) - \mathbf{x}^T\mathbf{\Phi}\mathbf{w}_{ols}^{(\Phi)}| \quad \leq \quad \frac{(\epsilon_f + \epsilon_{prj})\sqrt{n} + \sigma_\eta\sqrt{\log(2/\pi\delta_{var}^2)}}{\sigma_{\min}^{(X)}\left(\sqrt{\frac{D}{d}} - \sqrt{\frac{-2\log\delta_\Phi}{d}} - 1\right)}\left(1 + \sqrt{\frac{12}{d}\log\frac{2}{\delta_{prj}}}\right) \tag{24}$$

$$+\epsilon_f + \epsilon_{prj}, \tag{25}$$

where,

$$\epsilon_{prj} = c\sqrt{\frac{k\log d\log(12eD/k\delta_{prj})}{d}}.$$

Proof of Theorem 4. Using Theorem 2, the following holds with probability no less than $1 - \delta_{\text{prj}}$:

$$f(\mathbf{x}) \quad = \quad (\mathbf{\Phi}^T\mathbf{x})^T(\mathbf{\Phi}^T\mathbf{w}) + b_f(\mathbf{x}) + b_{\text{prj}}(\mathbf{x}), \tag{26}$$

where $|b_f(\mathbf{x})| \leq \epsilon_f$, $|b_{\text{prj}}(\mathbf{x})| \leq \epsilon_{\text{prj}}$.

Note that $\|(\mathbf{X\Phi})^\dagger\| \leq \|\mathbf{X}^\dagger\|\|\mathbf{\Phi}^\dagger\| \leq 1/(\sigma_{\min}^{(X)}\sigma_{\min}^{(\Phi)})$. Using the bound discussed in [7], we have with probability $1 - \delta_\Phi$:

$$\sigma_{\min}^{(\Phi)} \leq \sqrt{\frac{D}{d}} - \sqrt{\frac{-2\log\delta_\Phi}{d}} - 1.$$

Now, using Lemma 3 with the form of a function described in Eqn 26, we have:

$$|f(\mathbf{x}) - \mathbf{x}^T\mathbf{\Phi}\mathbf{w}_{\text{ols}}^{(\Phi)}| \leq \|\mathbf{x}^T\mathbf{\Phi}\|\|(\mathbf{X\Phi})^\dagger\|\left((\epsilon_f + \epsilon_{\text{prj}})\sqrt{n} + \sigma_\eta\sqrt{\log(2/\pi\delta_{\text{var}}^2)}\right) + \epsilon_f + \epsilon_{\text{prj}}, \tag{27}$$

which yields the theorem after the substitution of ϵ_{prj} and matrix norms.

□

Assuming that $\epsilon_f = 0$ (for simplification) we can rewrite the bound as:

$$|f(\mathbf{x}) - \mathbf{x}^T\mathbf{\Phi}\mathbf{w}_{\text{ols}}^{(\Phi)}| \quad \leq \quad \tilde{O}\left(\sqrt{k\log d\log(D/k)}\left(\frac{1}{\sqrt{d}} + \frac{\sqrt{n}}{\sigma_{\min}^{(X)}\sqrt{D}}\right)\right) + \tilde{O}\left(\frac{\sigma_\eta}{\sigma_{\min}^{(X)}}\sqrt{d/D}\right).$$

The first $\tilde{O}$ term of the RHS is a part of a bias due to the projection. The second $\tilde{O}$ term is the variance term. This bound is particularly useful when $n > D$. With that assumption, in order to illustrate a more clear bias–variance trade-off, assume that $\mathbf{X}_{i,j}$ comes from $\mathcal{N}(0, \frac{1}{D})$. Then we have $\sigma_{\min}^{(X)} \approx \sqrt{\frac{n}{D}}$. Fixing the values for δ's and ignoring the $\sqrt{\log d}$ term (slow growing function of d), we get that the error is bounded by:

$$c_0 + c_1\sqrt{k\log(D/k)}\left(\frac{1}{\sqrt{d}} + 1\right) + c_2\frac{\sigma_\eta}{\sqrt{n}}\sqrt{d},$$

in which case we clearly observe the trade-off with respect to the compressed dimension d. Now if $d < n < D$ with $\mathbf{X}_{i,j} \sim \mathcal{N}(0, \frac{1}{D})$, we have $\sigma_{\min}^{(X)} \approx 1 - \sqrt{\frac{n}{D}}$ which gives us the following bound:

$$c_0 + c_1\sqrt{k\log(D/k)}\left(\frac{1}{\sqrt{d}} + \frac{\sqrt{n}}{\sqrt{D} - \sqrt{n}}\right) + c_2\frac{\sigma_\eta}{\sqrt{D} - \sqrt{n}}\sqrt{d}.$$

This bound is, however, counter-intuitive, as the error grows when n is increased. This is due to the fact that the bound $\|(\mathbf{X\Phi})^\dagger\| \leq \|\mathbf{X}^\dagger\|\|\mathbf{\Phi}^\dagger\|$ gets looser as n gets close to D. When n is close to D, we might not have a tight closed-form bound over the $\|(\mathbf{X\Phi})^\dagger\|$ term, but we can still calculate it empirically for any given value of d by sampling a specific projection of the corresponding size.

References

[1] S. Zhou, J. Lafferty, and L. Wasserman. Compressed regression. In *Proceedings of Advances in neural information processing systems*, 2007.

[2] O.A. Maillard and R. Munos. Compressed least-squares regression. In *Proceedings of Advances in neural information processing systems*, 2009.

[3] M.A. Davenport, M.B. Wakin, and R.G. Baraniuk. Detection and estimation with compressive measurements. *Dept. of ECE, Rice University, Tech. Rep*, 2006.

[4] E.J. Candès and M.B. Wakin. An introduction to compressive sampling. *Signal Processing Magazine, IEEE*, 25(2):21–30, 2008.

[5] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. The Johnson–Lindenstrauss lemma meets compressed sensing. *Constructive Approximation*, 2007.

[6] G.G. Lorentz, M. von Golitschek, and Y. Makovoz. *Constructive approximation: advanced problems*, volume 304. Springer Berlin, 1996.

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