Compressed Least-Squares Regression on Sparse Spaces

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

Abstract

Recent advances in the area of compressed sensing suggest that it is possible to reconstruct high-dimensional sparse signals from a small number of random projections. Domains in which the sparsity assumption is applicable also offer many interesting large-scale machine learning prediction tasks. It is therefore important to study the effect of random projections as a dimensionality reduction method under such sparsity assumptions.

In this paper we develop the bias–variance analysis of a least-squares regression estimator in compressed spaces when random projections are applied on sparse input signals. Leveraging the sparsity assumption, we are able to work with arbitrary non i.i.d. sampling strategies and derive a worst-case bound on the entire space. Empirical results on synthetic and real-world datasets shows how the choice of the projection size affects the performance of regression on compressed spaces, and highlights a range of problems where the method is useful.

Modern machine learning methods have to deal with overwhelmingly large datasets, e.g. for text, sound, image and video processing, as well as for time series prediction and analysis. Much of this data contains very high numbers of *features* or attributes, sometimes exceeding the number of labelled instances available for training. Even though learning from such data may seem hopeless, in reality, the data often contains structure which can facilitate the development of learning algorithms. In this paper, we focus on a very common type of structure, in which the instances are *sparse*, in the sense that a very small percentage of the features in each instance is non-zero. For example, a text may be encoded as a very large feature vector (millions of dimensions) with each feature being 1 if a corresponding dictionary word is present in the text, and zero otherwise. Hence, in each document, a very small number of features will be non-zero.

Several algorithms have been designed to deal with this setting (which we discuss in detail at the end of the paper). Here, we focus on a new class of methods for learning in large, sparse feature sets: random projections (Davenport, Wakin, and Baraniuk 2006; Baraniuk and Wakin 2009). Random projections have originated recently in the signal processing literature (Candès and Tao 2006; Candès and Wakin 2008). The idea was motivated by the need to sample and store very efficiently large datasets (such as images and video). The basic idea is that if the signal is generated as a linear combination of a small set of functions (chosen from a much larger set), then it can be reconstructed very well from a small, fixed number of *randomized* measurements. A solid theoretical foundation has been established for compressed sensing methods, showing that as the number of random measurements increases, the error in the reconstruction decreases at a nearly-optimal rate (Donoho 2006).

Compressed sampling has been studied in the context of machine learning from two points of view. One idea is to use random projections to compress the dataset, by combining training instances using random projections (see e.g. Zhou, Lafferty, and Wasserman (2007)). Such methods are useful, for instance, when the training set is too large or one has to handle privacy issues. Another idea is 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 Maillard and Munos (2009), and their analysis shows a bias–variance trade-off with respect to on-sample error bounds, which is further extended to bounds on the sampling measure, assuming an i.i.d. sampling strategy.

In this paper, we provide a bias–variance analysis of ordinary least-squares (OLS) regression in compressed spaces, referred to as COLS, when random projections are applied on sparse input feature vectors. We show that the sparsity assumption allows us to work with arbitrary non i.i.d. sampling strategies and we derive a worst-case bound on the entire space. The fact that we can work with non-i.i.d. data makes our results applicable to a large range of problems, including video, sound processing, music and time series data. The results allow us to make predictions about the generalization power of the random projection method, outside of the training data. The bound can be used to select the optimal size of the projection, such as to minimize the sum of expected approximation (bias) and estimation (variance) errors. It also provides the means to compare the error of linear predictors in the original and compressed spaces.

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 $\mathcal{X} \triangleq \{\mathbf{\Psi}\mathbf{z}, \text{ s.t. } \|\mathbf{z}\|_0 \leq k \text{ and } \|\mathbf{z}\| \leq 1\}$. For a concrete example, the signals can be natural images and Ψ can represent these signals in the frequency domain (e.g., see Olshausen, Sallee, and Lewicki (2001)).

The *on-sample* error of a regressor is the expected error when the input is drawn from the empirical distribution (the expected error when the input is chosen uniformly from the training set), and the *off-sample* error is the error on a measure other than the empirical one.

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. Davenport, Wakin, and Baraniuk (2006), Candès and Wakin (2008)). 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 Davenport, Wakin, and Baraniuk (2006)), which shows that for a finite set of points, inner product with a fixed vector is almost preserved after a random projection.

Theorem 1. *(Davenport, Wakin, and Baraniuk (2006)) 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 $\epsilon > 0$:*

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

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

We derive the corresponding theorem for sparse feature spaces.

[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. Davenport, Wakin, and Baraniuk (2006)).

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 $\epsilon > 0$:*

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

fails with probability less than:

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

The proof is attached in the appendix.

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

Bias–Variance Analysis of Ordinary Least-Squares

In this section, we analyze the worst case prediction error of the OLS solution. 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 $\mathbf{x} \in \mathcal{X}$:

$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + b_f(\mathbf{x}), \text{ where } |b_f(\mathbf{x})| \leq \epsilon_f, \tag{4}$$

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:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{b}_f + \eta, \tag{5}$$

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 on the worst-case error of the ordinary least-squares predictor. This lemma is partly a classical result in linear prediction theory and given here with a proof mainly for completeness.

Lemma 3. *Let $\mathbf{w}_{ols}$ be the OLS solution of Eqn 5 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 any $\mathbf{x} \in \mathcal{X}$, with probability no less than $1 - \delta_{var}$ the error in the OLS prediction follows this bound:*

$$|f(\mathbf{x}) - \mathbf{x}^T\mathbf{w}_{ols}| \leq \|\mathbf{x}\|\|\mathbf{X}^\dagger\|\left(\epsilon_f\sqrt{n} + \sigma_\eta\sqrt{\log(2/\pi\delta_{var}^2)}\right) + \epsilon_f. \tag{6}$$

The proof is attached in the appendix.

Compressed Ordinary Least-Squares

We are now ready to derive 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, 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 $0 < \delta_{prj} < 1$ and $0 < \delta_{var} \leq \sqrt{2/e\pi}$. Then, with probability no less than $1-\delta_{prj}$, we have for any $\mathbf{x} \in \mathcal{X}$ with probability no less than $1 - \delta_{var}$:*

$$|f(\mathbf{x}) - \mathbf{x}^T\mathbf{\Phi}\mathbf{w}_{ols}^{(\Phi)}| \leq \|\mathbf{x}^T\mathbf{\Phi}\|\|(\mathbf{X}\mathbf{\Phi})^\dagger\| \left((\epsilon_f + \epsilon_{prj})\sqrt{n} + \sigma_\eta\sqrt{\log(2/\pi\delta_{var}^2)}\right) + \epsilon_f + \epsilon_{prj}, \quad (7)$$

where,

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

The proof is included in the appendix.

Note that because we use random projections of the type defined in Eqn 1, the norm of $\mathbf{\Phi}$ can be bounded using the bound discussed in (Candès and Tao 2006); we have with probability $1 - \delta_\Phi$:

$$\|\mathbf{\Phi}\| \leq \sqrt{D/d} + \sqrt{(2\log(2/\delta_\Phi))/d} + 1 \quad \text{and}$$
$$\|\mathbf{\Phi}^\dagger\| \leq \left[\sqrt{D/d} - \sqrt{(2\log(2/\delta_\Phi))/d} - 1\right]^{-1}.$$

Similarly, when $n > D$, and the observed features are distributed as $\mathcal{N}(0, 1/\sqrt{D})$, we have that $\|\mathbf{X}^\dagger\|$ is of order $\tilde{O}(\sqrt{D/n})$. Thus $\|(\mathbf{X}\mathbf{\Phi})^\dagger\|$ is of order $\tilde{O}(\sqrt{d/n})$. In a more general case, when the training samples are sufficiently spread out, and even when the sample size is small ($d < n < D$), we expect the same behavior for the $\|(\mathbf{X}\mathbf{\Phi})^\dagger\|$ term[2]. Assuming that $\epsilon_f = 0$ (for simplification), and ignoring the terms constant in d, we can rewrite the bound on the error up to logarithmic terms as:

$$\tilde{O}\left(\sqrt{k\log(D/k)}\frac{1}{\sqrt{d}}\right) + \tilde{O}\left(\frac{\sigma_\eta}{\sqrt{n}}\sqrt{d}\right).$$

The first $\tilde{O}$ term is a part of the bias due to the projection (excess approximation error). The second $\tilde{O}$ term is the variance term that shrinks with larger training sets (estimation error). We clearly observe the trade-off with respect to the compressed dimension d. With the assumptions discussed above, the optimal projection size is thus of order $\tilde{O}(\sqrt{kn})$, which resembles the suggested size of $\tilde{O}(\sqrt{n})$ for on-measure error minimization discussed in Maillard and Munos (2009).

Empirical Analysis

In this section, we aim to elucidate the conditions under which random projections are useful in sparse regression problems. We start with synthetic regression datasets, in which the input vectors are sampled from a ($D = 1000$)-dimensional feature space, and at most 5% of the features are non-zero in any particular instance ($k = 50$). The target function is linear in the input features. The weight vector $\mathbf{w}$ of the target function is generated randomly from a normal distribution with diagonal covariance matrix.

Bias–Variance Trade-off

In order to be faithful to the theoretical analysis, we study the bias–variance trade-off using the maximum squared-error on the training set (on-sample error) and testing set (off-sample error), as a function of the projection size, in different noise settings. To do so, we generate weight vectors for the target function, $\mathbf{w}$, in which each component is drawn independently from a normal distribution. Ten of the weights are generated using a larger variance, such that 80% of the norm of $\mathbf{w}$ is on those elements. This relatively sparse choice of $\mathbf{w}$ helps to illustrate the trade-off, as we discuss later. The features of the training and testing instances are also independently normally distributed on k randomly chosen elements and are 0 elsewhere. Both the input values and the weight vector are of norm $\simeq 1$.

We consider two settings. In the first setting, the training set is smaller than number of original dimensions ($n < D$), while in the second, there are more training examples than the number of dimensions ($n > D$). In each setting, we generate different levels of noise, quantified by the ratio between the standard deviation of the noise and that of the target function on the training set (similar to the concept of signal-to-noise ratio from signal processing). We plot the maximum squared error as a function of the dataset size.

Figure 1 summarizes the trade-off, averaged over 100 runs, when the training and testing sets are of size $n = 800$. Even though OLS is not well defined in the original space when $n < D$, we can still use the pseudo-inverse of the input matrix to find the linear fit, among many possible solutions, that has the smallest norm for the weight vector (for convenience, we call this the OLS solution here). Because we are using a relatively sparse $\mathbf{w}$, we might get a fit that is better than the baseline constant predictor (i.e. projection of size 0). Figure 2 shows a similar analysis for training sets slightly larger than the number of dimensions ($n = 2000$).

For small noise levels, OLS in the original space works well and the plot shows a constant decrease in the error of the COLS predictor as we use larger random projections. In these cases, one should use as many features as possible (depending on the computational cost) for regression. As we increase the noise level (middle panel of Figure 1 and Figure 2), the bias–variance trade-off becomes apparent. Here we see that it is better to use random projections of intermediate size to optimize the trade-off. Finding the optimal projection size is a challenge. Error bounds such as the ones presented in this paper give clues on the existence and values of such optimal sizes; or, one can use cross-validation to find the optimal projection size, as illustrated in these experiments. Higher noise levels, presented in the right panel of Figure 1 and Figure 2, make the prediction problem impossible to solve, in which case the best regression is the constant baseline predictor. Note that the OLS solution in this case is significantly worse, so we do not plot it in order to keep the panels on the same scale.

[2]This behavior is observed empirically in different experiments. The technical proof is a subject of future work.

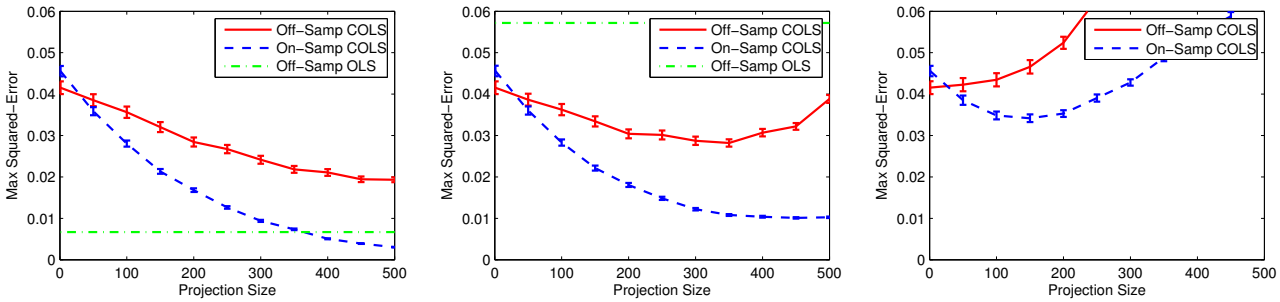

Figure 1: Error vs. projection size when $n = 800$ is less than $D = 1000$.
The noise ratio is (0.3), (1.1) and (3.0) from left to right. The OLS error for the right-most plot is $\simeq 0.29$.

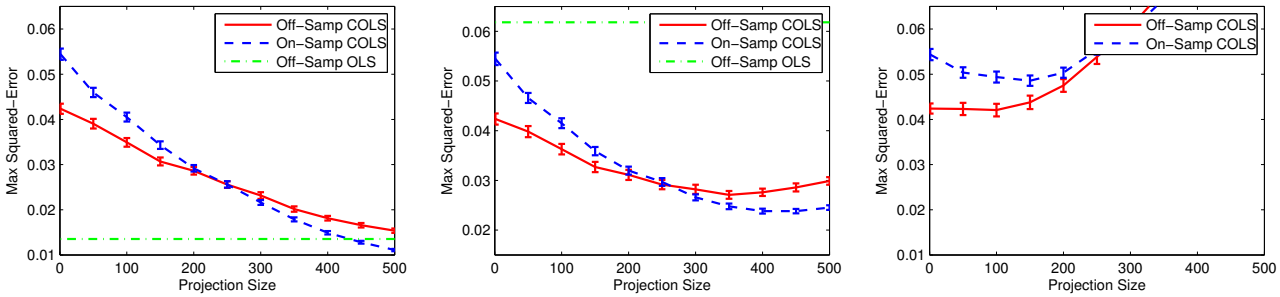

Figure 2: Error vs. projection size when $n = 2000$ is greater than $D = 1000$.
The noise ratio is (1.1), (2.4) and (5.0) from left to right. The OLS error for the right-most plot is $\simeq 0.40$.

We can conclude from this experiment that when the problem is not too easy (almost noiseless), or too difficult (large noise), random projections of sizes smaller than the number of dimensions provide the optimal worst-case error rates. Even in the noiseless setting, if D is very large, we might still want to use random projection as a form of feature selection/extraction, instead of L^1 regularization methods, which are significantly more expensive computationally (Efron et al. 2004).

Sparsity of Linear Coefficients

Next, we analyze the effect of the sparsity of the linear weight vector (rather than the sparsity of the inputs) on the worst-case prediction error. In this experiment, we fix the noise ratio and observe the bias–variance trade-off as we change the level of sparsity in the linear coefficients. We use a similar setting as described in the previous experiment. We generate a $\mathbf{w}$ by sampling from a Gaussian distribution with mean zero and unit diagonal covariance matrix, and then scale 10 of its elements such that they account for 50%, 70% and 90% of the norm of $\mathbf{w}$. The trade-off on the worst-case testing error and the error of the OLS predictor on the original and compressed space are shown in Figure 3.

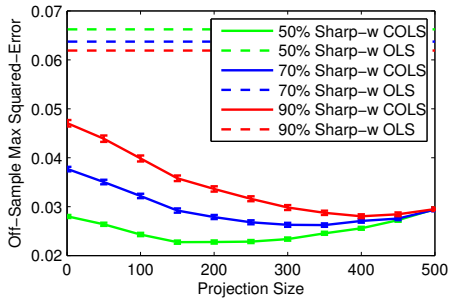

Figure 3: Error vs. projection size for different concentration of w, when $n = 2000$, $D = 1000$ and the noise ratio is (2.4).

The results indicate that the trade-off is present even when we use less concentrated weight vectors. However, random projections seem to have higher effectiveness in the reduction of prediction error for highly concentrated weight vectors. This could be due to the fact that in the current setup, the inner-product between $\mathbf{w}$ and points in a sparse space is better preserved when $\mathbf{w}$ is sparse itself. Nevertheless, for a problem having large enough number of samples and features, the trade-off is expected to be apparent even for non-concentrated weight vectors, as the theory suggests.

Music Similarity Prediction

To assess the effectiveness of random projections as means of feature extraction for regression in high-dimensional spaces, we experiment with a music dataset for a similarity prediction task. The task is to predict the similarity between tracks of classical music using audio analysis. Last.fm (2012) provides a similarity measure between different tracks in its music database, which is calculated using the listening patterns from the users and the co-presence

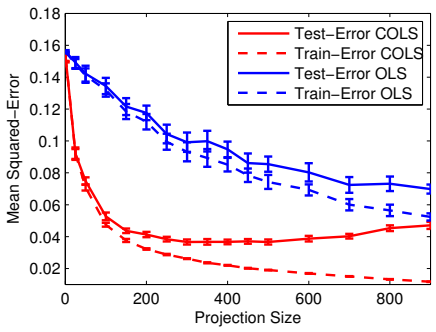

Figure 4: Error vs. number of features used for random projection and naive feature selection on the music dataset.

of tracks in playlists. For newly added songs, however, this value is unknown, and thus a predictor might be very useful for music recommendation or automatic playlist generation.

The regression problem is constructed as follows. We analyze 30 second audio samples of music tracks and apply frequency analysis on each 200ms segment. Scanning the entire dataset, we extract the top 50 common *chords*[3]. Using those chords, we then extract the top 1000 common *chord progressions*[4] of up to 8 segments. Having this set, we can check whether or not each of these progressions is present in a music track.

Our training set consists of pairs of music tracks along with a similarity score between 0 and 1 which is provided by Last.fm. For each pair of tracks in our training set, we construct a set of binary features corresponding to the simultaneous presence of a certain progression in track one, and another progression in track two. Therefore, there are 1 million features for each training example. Using random projections, we reduce the dimension and apply OLS in the compressed space. We compare this with a naive baseline which randomly chooses features from the original 1 million dimensional space and applies OLS on the chosen features.

Figure 4 compares the mean squared-error of the discussed methods for the training and testing sets, using ten-fold cross-validation on 2000 training samples. We use the average here instead of max, as we do not have access to the true target value and the worst case error contains a noise term.

The random projection method significantly outperforms the naive feature selection method, even with relatively small projection sizes. We expect the error of the naive OLS predictor to get minimized with a much larger number of selected features; however, solving the OLS for these larger problems would become computationally intractable.

This experiment shows that random projections can be useful in real world prediction tasks with very large feature spaces, while incurring relatively small computational cost. However, a more thorough analysis is required to compare this approach with other commonly used feature extraction methods.

[3]Combinations of notes present in each segment.

[4]Patterns of chords happening in a sequence.

Discussion

In this work, we analyze the worst-case prediction error of the OLS estimator built on the space induced by random projections from sparse spaces. We prove that random projections preserve the inner-product between sparse features and any fixed vector. This shows that near-linearity is preserved after the projection into smaller spaces, even for points that are never observed in the training set.

Leveraging the sparsity assumption in the input space, unlike previous analysis of random projection for regression tasks (Maillard and Munos 2009), we provide worst-case bounds that hold regardless of the distribution of the training or testing sets. Most noticeably, we do not require i.i.d. sampling or similarity between the training and testing sets. This is particularly useful when the regressor is expected to perform well on data that is not sampled from the same distribution as the one on which it is trained (e.g. time series analysis and domain adaptation settings).

Under mild assumptions on the distribution of the training data, our bound reduces to a bias–variance trade-off on the prediction error, as a function of the projection size. Increasing the number of projected features reduces the approximation error of the OLS estimator on the induced space, but at the same time introduces some estimation error as the number of learning parameters increases.

The optimal choice for the error trade-off depends on the structure of the target function (sparsity of $\mathbf{w}$), and the noise level. Our analysis suggests an optimal projection size of $\tilde{O}(\sqrt{nk})$ for worst-case error minimization, resembling the suggested size of $\tilde{O}(\sqrt{n})$ for on-measure error minimization (Maillard and Munos 2009). Depending on the noise level, the minimizer of expected error might be out of the feasible range of $1 \leq d \leq D$. This is manifested in our empirical evaluation of the method. With small noise, it is sometimes better to apply OLS on the original space (or if not possible, use projections as large as computationally feasible). For large noise levels and relatively small sample sizes, we are often in a situation where the best predictor is a constant one. Nevertheless, there is an important range of problems for which OLS prediction based on random projections vastly outperforms the prediction in the original space.

Our theoretical and empirical analysis of compressed OLS estimators provides some level of understanding of the usefulness of random projections in regression problems. This work focuses on sparse input space, which are typical in many fields of machine learning. There are many areas of future work in this domain. While L^1 regularized regression is not applicable in domains with large feature spaces due to its computational complexity, other types of linear estimators (e.g. L^2 regularized regression) should be analyzed in the settings we examine.

Since linear predictors are the building blocks of many learning algorithms, we expect random projections to be effective means of feature extraction when working with high dimensional data in many other fields of machine learning. These include the use of random projections in audio, video and time-series analysis, or with LSTD-type algorithms for high-dimensional reinforcement learn-

ing (Lazaric, Ghavamzadeh, and Munos 2010). These remain interesting subjects of future work.

Appendix

Proof of Theorem 2. The proof follows the steps of the proof of theorem 5.2 from Baraniuk et al. (2007). 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 indices. For each set of indices T, we define a k-dimensional hyperplane in the D-dimensional input space: $\mathcal{X}_T \triangleq \{\mathbf{\Psi}\mathbf{z}, \text{ s.t. } \mathbf{z} \text{ is zero outside } T \text{ and } \|\mathbf{z}\| \leq 1\}$. 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:

$$\forall \mathbf{x} \in \mathcal{X}_T : \min_{\mathbf{s} \in \mathbf{S}} \|\mathbf{x} - \mathbf{s}\| \leq \epsilon/4. \tag{8}$$

It is easy to prove (see e.g. Chapter 13 of Lorentz, von Golitschek, and Makovoz (1996)) that these conditions can be satisfied by choosing a grid of size $|S| \leq (12/\epsilon)^k$, 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:

$$\forall \mathbf{s} \in \mathcal{S} : \left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{s}\rangle - \langle \mathbf{w}, \mathbf{s}\rangle\right| \leq \frac{\epsilon}{2}\|\mathbf{s}\|, \tag{9}$$

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

Let a be the smallest number such that:

$$\forall \mathbf{x} \in \mathcal{X}_T : \left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{x}\rangle - \langle \mathbf{w}, \mathbf{x}\rangle\right| \leq a\|\mathbf{x}\|, \tag{10}$$

holds when Eqn 9 holds. The goal is to show that $a \leq \epsilon$. For any given $\mathbf{x} \in \mathcal{X}_T$, we choose an $\mathbf{s} \in S$ for which $\|\mathbf{x} - \mathbf{s}\| \leq \epsilon/4$. Therefore we have:

$$\begin{aligned}
&\left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{x}\rangle - \langle \mathbf{w}, \mathbf{x}\rangle\right| \leq \\
&\left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{x}\rangle - \langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{s}\rangle - \langle \mathbf{w}, \mathbf{x}\rangle + \langle \mathbf{w}, \mathbf{s}\rangle\right| + \\
&\left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{s}\rangle - \langle \mathbf{w}, \mathbf{s}\rangle\right| \\
&\leq \left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T(\mathbf{x} - \mathbf{s})\rangle - \langle \mathbf{w}, (\mathbf{x} - \mathbf{s})\rangle\right| + \\
&\left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{s}\rangle - \langle \mathbf{w}, \mathbf{s}\rangle\right| \\
&\leq a\epsilon/4 + \epsilon/2.
\end{aligned}$$

The last line is by the definition of a, and by applying Eqn 9 (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 10 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 $(eD/k)^k(4(12/\epsilon)^k + 2)e^{-d\epsilon^2/192}$. □

The above Theorem requires setting the magnitude error ϵ to obtain a probability bound that is a function thereof. While this result has the same form as Theorem 1, we need to use it the other way around, by setting the probability of error and obtaining a corresponding error magnitude. The following Corollary resolves this translation (the proof is straightforward by substitution).

Corollary 5. *Let $\mathbf{\Phi}^{D\times d}$ be a random projection according to Eqn 1. Let $\mathcal{X}$ be a D-dimensional k-sparse space. Fix $\mathbf{w} \in \mathbb{R}^D$ and $1 > \delta > 0$. Then, with probability $> 1 - \delta$*

$$\forall \mathbf{x} \in \mathcal{X} : \left|\langle \mathbf{\Phi}^T\mathbf{w}, \mathbf{\Phi}^T\mathbf{x}\rangle - \langle \mathbf{w}, \mathbf{x}\rangle\right| \leq \epsilon\|\mathbf{w}\|\|\mathbf{x}\|, \tag{11}$$

where $\epsilon = c\sqrt{\frac{k \log d \log(12eD/k\delta)}{d}}$.

Proof of Lemma 3. For the OLS solution of Eqn 5 we have:

$$\begin{aligned}
\mathbf{w}_{\text{ols}} &= \mathbf{X}^\dagger\mathbf{y} = \mathbf{X}^\dagger(\mathbf{X}\mathbf{w} + \mathbf{b}_f + \eta) \\
&= \mathbf{w} + \mathbf{X}^\dagger\mathbf{b}_f + \mathbf{X}^\dagger\eta. \qquad (12)
\end{aligned}$$

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

$$\begin{aligned}
|f(\mathbf{x}) - \mathbf{x}^T\mathbf{w}_{\text{ols}}| &\leq |\mathbf{x}^T\mathbf{w}_{\text{ols}} - \mathbf{x}^T\mathbf{w}| + \epsilon_f \qquad (13)\\
&\leq |\mathbf{x}^T\mathbf{X}^\dagger\mathbf{b}_f| + |\mathbf{x}^T\mathbf{X}^\dagger\eta| + \epsilon_f. \qquad (14)
\end{aligned}$$

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

$$|\mathbf{x}^T\mathbf{X}^\dagger\mathbf{b}_f| \leq \|\mathbf{x}^T\|\|\mathbf{X}^\dagger\|\|\mathbf{b}_f\| \leq \|\mathbf{x}\|\|\mathbf{X}^\dagger\|\epsilon_f\sqrt{n}. \tag{15}$$

For the second term in line 14 (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:

$$\begin{aligned}
\text{Var}[\mathbf{x}^T\mathbf{X}^\dagger\eta] &= \mathbb{E}[\mathbf{x}^T\mathbf{X}^\dagger\eta\eta^T(\mathbf{X}^\dagger)^T\mathbf{x}] \qquad (16)\\
&= \sigma_\eta^2\mathbf{x}^T\mathbf{X}^\dagger(\mathbf{X}^\dagger)^T\mathbf{x} \qquad (17)\\
&\leq \sigma_\eta^2\|\mathbf{x}^T\|\|\mathbf{X}^\dagger\|\|(\mathbf{X}^\dagger)^T\|\|\mathbf{x}\| \qquad (18)\\
&\leq \sigma_\eta^2\|\mathbf{x}\|^2\|\mathbf{X}^\dagger\|^2, \qquad (19)
\end{aligned}$$

where in line 17 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 $1 - \delta_{\text{var}}$:

$$|\mathbf{x}^T\mathbf{X}^\dagger\eta| \leq \sigma_\eta\|\mathbf{x}\|\|\mathbf{X}^\dagger\|\sqrt{\log(2/\pi\delta_{\text{var}}^2)}. \tag{20}$$

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

Proof of Theorem 4. Using Corollary 5, the following holds with probability no less than $1 - \delta_{\text{prj}}$:

$$f(\mathbf{x}) = (\mathbf{\Phi}^T\mathbf{x})^T(\mathbf{\Phi}^T\mathbf{w}) + b_f(\mathbf{x}) + b_{\text{prj}}(\mathbf{x}), \tag{21}$$

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

Now, using Lemma 3 with the form of a function described in Eqn 21, we have:

$$\begin{aligned}
&|f(\mathbf{x}) - \mathbf{x}^T\mathbf{\Phi}\mathbf{w}_{\text{ols}}^{(\Phi)}| \leq \\
&\|\mathbf{x}^T\mathbf{\Phi}\|\|(\mathbf{X}\mathbf{\Phi})^\dagger\|\left((\epsilon_f + \epsilon_{\text{prj}})\sqrt{n} + \sigma_\eta\sqrt{\log(2/\pi\delta_{\text{var}}^2)}\right) \\
&+ \epsilon_f + \epsilon_{\text{prj}}, \qquad (22)
\end{aligned}$$

which yields the theorem.

□

References

Baraniuk, R., and Wakin, M. 2009. Random projections of smooth manifolds. *Foundations of Computational Mathematics* 9(1):51–77.

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

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

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

Davenport, M.; Wakin, M.; and Baraniuk, R. 2006. Detection and estimation with compressive measurements. *Dept. of ECE, Rice University, Tech. Rep*.

Donoho, D. 2006. Compressed sensing. *Information Theory, IEEE Transactions on* 52(4):1289–1306.

Efron, B.; Hastie, T.; Johnstone, I.; and Tibshirani, R. 2004. Least angle regression. *The Annals of statistics* 32(2):407–499.

Last.fm. 2012. Web API. `http://www.last.fm/api`.

Lazaric, A.; Ghavamzadeh, M.; and Munos, R. 2010. Finite-Sample Analysis of LSTD. In *Proceedings of the international conference on machine learning*.

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

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

Olshausen, B.; Sallee, P.; and Lewicki, M. 2001. Learning sparse image codes using a wavelet pyramid architecture. In *Proceedings of Advances in neural information processing systems*.

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