Compressed Least-Squares Regression

Maillard Odalric, Rémi Munos

Compressed Least-Squares Regression

Odalric-Ambrym Maillard and Rémi Munos
SequeL Project, INRIA Lille - Nord Europe, France
{odalric.maillard, remi.munos}@inria.fr

Abstract

We consider the problem of learning, from K input data, a regression function in a function space of high dimension N using projections onto a random subspace of lower dimension M. From any linear approximation algorithm using empirical risk minimization (possibly penalized), we provide bounds on the excess risk of the estimate computed in the projected subspace (compressed domain) in terms of the excess risk of the estimate built in the high-dimensional space (initial domain). We apply the analysis to the ordinary Least-Squares regression and show that by choosing $M = O(\sqrt{K})$, the estimation error (for the quadratic loss) of the "Compressed Least Squares Regression" is $O(1/\sqrt{K})$ up to logarithmic factors. We also discuss the numerical complexity of several algorithms (both in initial and compressed domains) as a function of N, K, and M.

1 Problem setting

We consider a regression problem where we observe data $\mathcal{D}_K = (\{x_k, y_k\}_{k\leq K})$ (where $x_k \in \mathcal{X}$ and $y_k \in \mathbb{R}$) are assumed to be independently and identically distributed (i.i.d.) from some distribution P, where $x_k \sim P_x$ and

$$y_k = f^*(x_k) + \eta_k(x_k),$$

with f^* being the (unknown) target function, and η_k a centered noise of variance $\sigma^2(x_k)$.

For a given class of functions $\mathcal{F}$, and $f \in \mathcal{F}$, we define the empirical (quadratic) error

$$L_K(f) \stackrel{\text{def}}{=} \frac{1}{K}\sum_{k=1}^{K}[y_k - f(x_k)]^2,$$

and the generalization (quadratic) error

$$L(f) \stackrel{\text{def}}{=} \mathbb{E}_{(X,Y)\sim P}[(Y - f(X))^2].$$

Our goal is to return a regression function $\widehat{f} \in \mathcal{F}$ with lowest possible generalization error $L(\widehat{f})$.

Notations: *In the sequel we will make use of the following notations about norms: for $h : \mathcal{X} \mapsto \mathbb{R}$, we write $||h||_P$ for the L_2 norm of h with respect to (w.r.t.) the measure P, $||h||_{P_K}$ for the L_2 norm of h w.r.t. the empirical measure P_K, and for $u \in \mathbb{R}^n$, $||u||$ denotes by default $\left(\sum_{i=1}^n u_i^2\right)^{1/2}$.*

The measurable function minimizing the prediction error is f^*, but it may be the case that $f^* \notin \mathcal{F}$. For any regression function f, the **excess risk**

$$L(f) - L(f^*) = ||f - f^*||_P^2,$$

decomposes as the sum of the **estimation error** $L(f) - \inf_{f\in\mathcal{F}} L(f)$ (which is our main concern in this paper) and the **approximation error** $\inf_{f\in\mathcal{F}} L(f) - L(f^*) = \inf_{f\in\mathcal{F}} ||f - f^*||_P^2$ which measures the distance between f^* and the considered function space.

In this paper we consider a class of functions $\mathcal{F}_N$ defined as the vector space spanned by a set of N functions $\{\varphi_n\}_{1\leq n\leq N}$ called *features*. Thus: $\mathcal{F}_N \stackrel{\text{def}}{=} \{f_\alpha \stackrel{\text{def}}{=} \sum_{n=1}^N \alpha_n\varphi_n,\ \alpha \in \mathbb{R}^N\}$.

Case $K > N$: When the number of data K is larger than the number of features N, the ordinary Least-Squares Regression (LSR) provides the LS solution $f_{\widehat{\alpha}}$ which is, in this case, defined as the minimizer of the empirical risk $L_K(f)$ in $\mathcal{F}_N$, i.e. $\widehat{\alpha} = \operatorname{argmin}_{\alpha \in \mathbb{R}^N} L_K(f_\alpha)$ with $L_K(f_\alpha) = ||\Phi\alpha - Y||_K$, where we introduce Φ to be the $K \times N$ matrix with elements $(\varphi_n(x_k))_{1 \leq n \leq N, 1 \leq k \leq K}$ and Y the K-vector with components $(y_k)_{1 \leq k \leq K}$.

Usual results provide estimation error bounds in terms of the capacity of the function space and the number of data. In the case of linear approximation, the capacity measures (such as covering numbers [23] or the pseudo-dimension [16]) depend on the number of features (for example the pseudo-dimension is at most $N + 1$). For example, let $f_{\widehat{\alpha}}$ be a LS estimate (minimizer of L_K in $\mathcal{F}_N$), then (a more precise statement will be stated later in subsection 3.2) the expected estimation error is bounded as:

$$\mathbb{E}\big[L(f_{\widehat{\alpha}}) - \inf_{f \in \mathcal{F}_N} L(f)\big] \leq c\sigma^2 \frac{N \log K}{K}, \tag{1}$$

where c is a universal constant, $\sigma \stackrel{\text{def}}{=} \sup_{x \in \mathcal{X}} \sigma(x)$, and the expectation is taken with respect to P.

Now, given that the excess risk is the sum of this estimation error and the approximation error $\inf_{f \in \mathcal{F}_N} ||f - f^*||_P$ of the class $\mathcal{F}_N$, and since the later usually decreases when the number of features N increases [12] (e.g. when $\bigcup_N \mathcal{F}_N$ is dense in $L_2(P)$), we wish to consider problems with large N, which is the setting considered in this paper.

Case $N > K$: Now if N is larger than K then we face the problem of overfitting since there are more parameters than actual data (more variables than constraints) and (1) provides poor information about the generalization ability of any LS regressor. In addition, there are many minimizers (in fact a vector space of same dimension as the null space of $\Phi^T\Phi$) of the empirical risk.

Several approaches have been proposed in the literature:

- **LS solution with minimal norm**: The solution is the minimizer of the empirical risk with minimal (l_1 or l_2)-norm:

  $$\widehat{\alpha} = \arg \min_{\alpha \in \operatorname{argmin}_{\alpha'} L_K(f'_\alpha)} ||\alpha||_{1 \text{ or } 2}$$

  The choice of ℓ_2-norm yields the ordinary LS solution. The choice of ℓ_1-norm has been used for generating sparse solutions, and assuming that the target function admits a sparse decomposition, the field of Compressed Sensing [9, 21] provides sufficient conditions for recovering the exact solution. However, such conditions (e.g. that Φ possesses the Restricted Isometric Property (RIP)) does not hold in general in this regression setting. In addition, when N is large, solving these problems (both for l_1 or l_2-norm) may be computationally expensive.

- **Regularization**. The solution is the minimizer of the empirical risk plus a penalty term, for example

  $$\widehat{f} = \arg \min_{f \in \mathcal{F}_N} L_K(f) + \lambda ||f||^2_{1 \text{ or } 2}.$$

  where λ is a parameter and usual choices for the norm are ℓ_2 (ridge-regression [20]) and ℓ_1 (LASSO [19]). A close alternative is the Dantzig selector [8, 5] which solves:

  $$\widehat{\alpha} = \arg \min_{||\alpha||_1 \leq \lambda} ||\Phi^T(Y - \Phi\alpha)||_\infty.$$

  The numerical complexity and generalization bounds of those methods depend on the sparsity of the target function decomposition in $\mathcal{F}_N$.

Now if we possess a sequence of function classes $(\mathcal{F}_N)_{N \geq 1}$ with increasing capacity, we may perform **structural risk minimization** [22] by solving in each model the empirical risk penalized by a term that depends on the size of the model:

$$\widehat{f_N} = \arg \min_{f \in \mathcal{F}_N, N \geq 1} L_K(f) + \text{pen}(N, K),$$

where the penalty term measures the capacity of the function space.

In this paper we follow another approach.

Our contribution: We consider a set of M random linear combinations of the initial N features and perform our favorite LS regression algorithm (possibly regularized) using those "compressed features". This is equivalent to projecting the K points from the initial domain (of size N) onto a random subspace of dimension M, and then performing the regression in the "compressed domain" (i.e. span of the compressed features). This is made possible because random projections approximately preserve inner products between vectors (by a variant of the Johnson-Lindenstrauss Lemma stated in Section 3.1).

Our main result is a bound on the excess risk of a linear estimator built in the compressed domain in terms of the excess risk of the linear estimator built in the initial domain (Section 3.1).

We further detail the case of ordinary Least-Squares Regression (Section 3.2) and discuss, in terms of M, N, K, the different tradeoffs concerning:

- *The excess risk*: the smaller estimation error in the compressed domain versus the additional approximation error introduced by the random projection,
- *The numerical complexity*: the reduced complexity of solving the LSR in the compressed domain versus the additional load of performing the projection.

As a consequence, we show that by choosing $M = O(\sqrt{K})$ projections and using LSR in the compressed domain, we define a **Compressed Least-Squares Regression** which uses $O(NK^{3/2})$ elementary operations to compute a regression function with estimation error (relatively to the initial function space $\mathcal{F}_N$) of order $\sqrt{\frac{\log K}{K}}$. This is competitive with the best methods up to our knowledge.

Related works: Using compression and random projections in various learning areas has received increasing interest over the past few years, especially with the success of Compressed Sensing in signal processing. In [7], the authors use a SVM algorithm in a compressed space for the purpose of classification and show that their resulting algorithm has good generalization properties. In [25], the authors consider a notion of compressed linear regression. For data $Y = X\beta + \varepsilon$, where β is the target and ε a standard noise, they use compression of the set of data, thus considering $AY = AX\beta + A\varepsilon$, where A has a Restricted Isometric Property. They provide an analysis of the LASSO estimator built from these compressed data, and discuss a property called sparsistency, i.e. the number of random projections needed to recover β (with high probability) when it is sparse. These works differ from our approach in the fact that we do not consider a compressed (input and/or output) data space but a compressed feature space instead.

In [10], the authors discuss how compressed measurements may be useful to solve many detection, classification and estimation problems without having to reconstruct the signal ever. Interestingly, they make no assumption about the signal being sparse, like in our work. In [6, 17], the authors show how to map a kernel $k(x, y) = \varphi(x) \cdot \varphi(y)$ into a low dimensional space, while still approximately preserving the inner products. Thus they build a low-dimensional feature space specific for (translation invariant) kernels.

2 Variant of the Johnson-Lindenstrauss Lemma

We start by stating a simple consequence of the Johnson-Lindenstrauss Lemma which shows that inner-product are approximately preserved through random projections.

Let A be a $M \times N$ matrix of i.i.d. elements drawn for some distribution ρ. Examples of distributions are:

- Gaussian random variables $\mathcal{N}(0, \frac{1}{M})$,
- $\pm$ Bernoulli distributions, i.e. which takes values $\pm\frac{1}{\sqrt{M}}$ with equal probability $1/2$,
- Other related distributions such as taking values $\pm\sqrt{\frac{3}{M}}$ with probability $1/6$ and 0 with probability $2/3$.

Proposition 1 *Let $(u_k)_{1\leq k\leq K}$ and v be vectors of $\mathbb{R}^N$. Let A be a $M\times N$ matrix of i.i.d. elements drawn for one of the previously defined distributions. For any $\varepsilon>0$, $\delta>0$, for $M\geq \frac{1}{\frac{\varepsilon^2}{4}-\frac{\varepsilon^3}{6}}\ln\frac{4K}{\delta}$, we have, with probability at least $1-\delta$, for all $k\leq K$,*

$$|Au_k\cdot Av-u_k\cdot v|\leq\varepsilon||u_k||\,||v||.$$

Proof: We make use of the following lemma, which states that the random (with respect to the choice of the matrix A) variable $||Au||^2$ concentrates around its expectation $||u||$ when M is large. The proof uses concentration inequalities (Cramer's large deviation Theorem) and may be found e.g. in [1].

Lemma 1 *For any vector u in $\mathbb{R}^N$ and any $\varepsilon\in(0,1)$, we have*

$$\begin{aligned}\mathbb{P}\Big(||Au||^2\geq(1+\varepsilon)||u||^2\Big) &\leq e^{-M(\varepsilon^2/4-\varepsilon^3/6)}\\ \mathbb{P}\Big(||Au||^2\leq(1-\varepsilon)||u||^2\Big) &\leq e^{-M(\varepsilon^2/4-\varepsilon^3/6)}\end{aligned}$$

To prove the proposition, we apply the lemma to any couple of vectors $u+w$ and $u-w$, where u and w are vectors of norm 1. Thus we have that

$$\begin{aligned}4Au\cdot Aw &= ||Au+Aw||^2-||Au-Aw||^2\\ &\leq (1+\varepsilon)||u+w||^2-(1-\varepsilon)||u-w||^2\\ &= 4u\cdot w+\varepsilon(||u+w||^2+||u-w||^2)\\ &= 4u\cdot w+2\varepsilon(||u||^2+||w||^2)=4u\cdot w+4\varepsilon.\end{aligned}$$

fails with probability $2e^{-M(\varepsilon^2/4-\varepsilon^3/6)}$ (we applied the previous lemma twice at line 2).

Thus for each $k\leq K$, we have with same probability:

$$Au_k\cdot Av\leq u_k\cdot v+\varepsilon||u_k||\,||v||.$$

Now the symmetric inequality holds with the same probability, and using a union bound for considering all $(u_k)_{k\leq K}$, we have that

$$|Au_k\cdot Av-u_k\cdot v|\leq\varepsilon||u_k||\,||v||,$$

holds for all $k\leq K$, with probability $1-4Ke^{-M(\varepsilon^2/4-\varepsilon^3/6)}$, and the proposition follows. □

3 Compressed Least Square Regression

We remind that the initial set of features is $\{\varphi_n:\mathcal{X}\mapsto\mathbb{R},1\leq n\leq N\}$ and the initial domain $\mathcal{F}_N\stackrel{\text{def}}{=}\{f_\alpha=\sum_{n=1}^N\alpha_n\varphi_n,\,\alpha\in\mathbb{R}^N\}$ is the span of those features. We write $\varphi(x)$ the N-vector of components $(\varphi_n(x))_{n\leq N}$.

Let A be a $M\times N$ matrix satisfying the property of Proposition 1. We now introduce the set of M **compressed features** $(\psi_m)_{1\leq m\leq M}$ such that $\psi_m(x)\stackrel{\text{def}}{=}\sum_{n=1}^N A_{m,n}\varphi_n(x)$. We also write $\psi(x)$ the M-vector of components $(\psi_m(x))_{m\leq M}$. Thus $\psi(x)=A\varphi(x)$.

We define the **compressed domain** $\mathcal{G}_M\stackrel{\text{def}}{=}\{g_\beta=\sum_{m=1}^M\beta_m\psi_m,\,\beta\in\mathbb{R}^M\}$ the span of the compressed features (vector space of dimension at most M).

We call a learning algorithm $\mathcal{A}$ a procedure that takes as input a set of data $\mathcal{D}$ and a function space $\mathcal{F}$ and returns a regression function $\widehat{f}\in\mathcal{F}$. For any learning algorithm $\mathcal{A}$ applied to $(\mathcal{D}_K,\mathcal{F}_N)$, where $\mathcal{F}_N$ a linear function space, we define the **compressed algorithm** $\mathcal{A}$ as the algorithm that outputs $\widehat{g}=\mathcal{A}(\mathcal{D}_K,\mathcal{G}_M)\in\mathcal{G}_M$.

The next subsection provides bounds on the excess risk of compressed $\mathcal{A}$ in terms of the excess risk of $\mathcal{A}$ and the number of projections M. This result applies to the algorithms mentioned in the introduction, e.g. the ordinary LS regression (analyzed in details in Section 3.2), the LASSO, the ridge regression, etc.

3.1 General result

Definition of $\kappa_{\mathcal{A}}, \kappa'_{\mathcal{A}}, c_{\mathcal{A}}, c'_{\mathcal{A}}$: For a learning algorithm $\mathcal{A}$, we define the functions $\kappa_{\mathcal{A}}$ and $\kappa'_{\mathcal{A}}$ and the constants $c_{\mathcal{A}}$ and $c'_{\mathcal{A}}$ such that for any dataset $\mathcal{D} = ((X_1, Y_1), \ldots, (X_K, Y_K))$ i.i.d. from the probability measure P, any function space $\mathcal{F}$, the estimation function $\widehat{f} \in \mathcal{F}$ returned by $\mathcal{A}(\mathcal{D}, \mathcal{F})$ satisfies:

$$\begin{aligned} \mathbb{E}(||\widehat{f} - f^*||_P^2) &\leq \kappa_{\mathcal{A}}(\mathcal{D}, \mathcal{F}, P) + c_{\mathcal{A}} \inf_{f \in \mathcal{F}} ||f - f^*||_P^2 \\ \text{and:} \quad \mathbb{E}_Y(||\widehat{f} - f^*||_{P_K}^2) &\leq \kappa'_{\mathcal{A}}(\mathcal{D}, \mathcal{F}, P_K) + c'_{\mathcal{A}} \inf_{f \in \mathcal{F}} ||f - f^*||_{P_K}^2 \end{aligned}$$

where $\mathbb{E}_Y$ denotes the expectation conditionally on the input samples $\{X_1, ..., X_K\}$ and P_K the empirical measure.

Theorem 1 *For any $\varepsilon > 0, \delta > 0$, let A be a random $M \times N$ matrix satisfying the property of Proposition 1 with $M \geq \frac{1}{\frac{\varepsilon^2}{4} - \frac{\varepsilon^3}{6}} \log(8K/\delta)$, and $\mathcal{G}_M$ be the compressed domain resulting from this choice of A. Let $\widehat{g} \in \mathcal{G}_M$ be the output of the compressed algorithm $\mathcal{A}$ (i.e. $\widehat{g} = \mathcal{A}(\mathcal{D}_K, \mathcal{G}_M)$). Then with probability at least $1 - \delta$,*

$$\mathbb{E}(||\widehat{g} - f^*||_P^2) \leq \kappa_{\mathcal{A}}(\mathcal{D}_K, \mathcal{G}_M, P) + c_{\mathcal{A}}\Big(\varepsilon^2 ||\alpha^+||^2 \sup_{x \in \mathcal{X}} ||\varphi(x)||^2 \big(1 + \frac{\log 4/\delta}{K}\big) + \inf_{f \in \mathcal{F}_N} ||f - f^*||_P^2\Big). \quad (2)$$

and the similar result holds with $\mathbb{E}, \kappa_{\mathcal{A}}, c_{\mathcal{A}}, ||\cdot||_P$ replaced (respectively) by $\mathbb{E}_Y, \kappa'_{\mathcal{A}}, c'_{\mathcal{A}}, ||\cdot||_Y$, where $||h||_Y \stackrel{\text{def}}{=} (\mathbb{E}_Y(h^2))^{1/2}$.

The theorem links the excess risk bound (2) of the estimator $\widehat{g}$ obtained in the compressed domain to that of the estimator $\widehat{f}$ obtained in the initial domain:

$$\mathbb{E}(||\widehat{f} - f^*||_P^2) \leq \kappa_{\mathcal{A}}(\mathcal{D}, \mathcal{F}_N, P) + c_{\mathcal{A}} \inf_{f \in \mathcal{F}_N} ||f - f^*||_P^2$$

This shows the tradeoff between the reduced estimation term ($\kappa_{\mathcal{A}}(\mathcal{D}_K, \mathcal{G}_M, P)$ is usually smaller than $\kappa_{\mathcal{A}}(\mathcal{D}, \mathcal{F}_N, P)$ when $M < N$) and the increased (but controlled) approximation term due to the compression.

Proof: We write for convenience $f^+ = f_{\alpha^+} = \arg\min_{f \in \mathcal{F}_N} ||f - f^*||_P$. Thus $\alpha^+ = \arg\min_{\alpha \in \mathbb{R}^N} ||f_\alpha - f^*||_P$. Write also $\beta^+ = A\alpha^+$ and $g^+ = g_{\beta^+}$. By the definition of $\kappa_{\mathcal{A}}$ and $c_{\mathcal{A}}$, we have:

$$\mathbb{E}\big(||\widehat{g} - f^*||_P^2\big) \leq \kappa_{\mathcal{A}}(\mathcal{D}_K, \mathcal{G}_M, P) + c_{\mathcal{A}} \inf_{g \in \mathcal{G}_M} ||g - f^*||_P^2. \quad (3)$$

We have:

$$\inf_{g \in \mathcal{G}_M} ||g - f^*||_P^2 \leq ||g^+ - f^*||_P^2 = ||g^+ - f^+||_P^2 + ||f^+ - f^*||_P^2 \quad (4)$$

Note that the two functions $(g^+ - f^+)$ and $(f^+ - f^*)$ are orthogonal since g^+ belongs to $\mathcal{F}_N$.

We now bound $||g^+ - f^+||_P^2$ by using concentration inequalities applied to the random variables $Z_k = A\alpha^+ \cdot A\varphi(x_k) - \alpha^+ \cdot \varphi(x_k)$. From Proposition 1 we have that for any $\varepsilon > 0$, on an event $\mathcal{E}$ of probability more than $1 - \delta/2 = 1 - 4K\exp(-M(\varepsilon^2/4 - \varepsilon^3/6))$, we have for all $k \leq K$,$|Z_k| \leq \varepsilon ||\alpha^+|| \, ||\varphi(x_k)|| \leq \varepsilon ||\alpha^+|| \sup_{x \in \mathcal{X}} ||\varphi(x)|| \stackrel{\text{def}}{=} C$.

By setting $a = ||g^+ - f^+||_{P_K}^2 = \frac{1}{K}\sum_{k=1}^K Z_k^2$ and $b = ||g^+ - f^+||_P^2 = \frac{1}{K}\sum_{k=1}^K \mathbb{E}(Z_k^2)$, we deduce that conditionally on the event $\mathcal{E}$:

$$\begin{aligned} \mathbb{P}\Big(\big|||g^+ - f^+||_{P_K} - ||g^+ - f^+||_P\big| \geq t\Big) &= \mathbb{P}(|\sqrt{a} - \sqrt{b}| \geq t) \leq \mathbb{P}(|a - b| \geq t\sqrt{a + b}) \\ &\leq 2\exp\Big[-2t^2K^2 \frac{\frac{1}{K}\sum_{k=1}^K Z_k^2 + \mathbb{E}(Z_k^2)}{\sum_{k=1}^K [Z_k^2 - \mathbb{E}(Z_k^2)]^2}\Big] \\ &\leq 2\exp\Big[-2t^2K \frac{\sum_{k=1}^K Z_k^2 + \mathbb{E}(Z_k^2)}{C^2 \sum_{k=1}^K [Z_k^2 + \mathbb{E}(Z_k^2)]}\Big] \\ &= 2\exp(-\frac{2t^2K}{C^2}), \end{aligned}$$

where we used the property that $|\sqrt{a}-\sqrt{b}| = \frac{|a-b|}{\sqrt{a}+\sqrt{b}} \leq \frac{|a-b|}{\sqrt{a+b}}$ in the first line, the Chernoff-Hoeffding bound in the second line, the fact that $|u^2-v^2|^2 \leq u^4+v^4$ and the definition of C in the third line. Thus, conditionally on $\mathcal{E}$, we have that with probability at least $1-\delta'$, $\big|||g^+-f^+||_P - ||g^+-f^+||_{P_K}\big| \leq C\sqrt{\frac{\log 2/\delta'}{2K}}$, thus also

$$\begin{aligned} ||g^+-f^+||_P^2 &\leq ||g^+-f^+||_{P_K}^2 + 2\varepsilon^2||\alpha^+||^2 \sup_x ||\varphi(x)||^2 \frac{\log 2/\delta'}{2K} \\ &\leq \varepsilon^2||\alpha^+||^2 \sup_x ||\varphi(x)||^2 \Big(1+\frac{\log 2/\delta'}{K}\Big) \end{aligned}$$

since under $\mathcal{E}$, we have $||g^+-f^+||_{P_K}^2 = \frac{1}{K}\sum_{k=1}^K Z_k^2 \leq C^2$.

Combining with (4) we deduce that by setting $\delta' = \delta/2$, with probability at least $(1-\delta/2)(1-\delta') \geq 1-\delta$,

$$\inf_{g\in\mathcal{G}_M} ||g-f^*||_P^2 \leq \Big(\varepsilon||\alpha^+|| \sup_x ||\varphi(x)|| \sqrt{1+\frac{\log 4/\delta}{K}}\Big)^2 + ||f^+-f^*||_P^2 \tag{5}$$

and (2) follows.

For the second part of the theorem, we simply note the dependency of $\widehat{\alpha}$ with Y and $(X_1,\ldots,X_K)$ and replace $\mathbb{E}, \kappa_{\mathcal{A}}, c_{\mathcal{A}}, ||\cdot||_P$ by (respectively) $\mathbb{E}_Y, \kappa'_{\mathcal{A}}, c'_{\mathcal{A}}, ||\cdot||_Y$. □

Computational issues: We now discuss the relative computational costs of a given algorithm applied either in the initial domain or in the compressed domain. Let us write $\mathrm{Cx}(\mathcal{D}_K, \mathcal{F}_N, P)$ the complexity (e.g. number of elementary operations) of an algorithm $\mathcal{A}$ to compute the regression function $\widehat{f}$ when provided with the data $\mathcal{D}_K$ and function space $\mathcal{F}_N$.

We plot in the table below, both for the initial and the compressed versions of the algorithm $\mathcal{A}$, the order of complexity for (i) the cost of building the feature matrix, (ii) the cost for computing the estimator, (iii) the cost of making one prediction (i.e. computing $\widehat{f}(x)$ for any x):

	Initial domain	Compressed domain
Construction of the feature matrix	NK	NKM
Computing the regression function	$\mathrm{Cx}(\mathcal{D}_K, \mathcal{F}_N, P)$	$\mathrm{Cx}(\mathcal{D}_K, \mathcal{G}_M, P)$
Making one prediction	N	NM

Note that the values mentioned for the compressed domain are upper-bounds on the real complexity and do not take into account the possible sparsity of the projection matrix A (which would speed up matrix computations, see e.g. [2, 1]).

In light of these complexity results as well as the bounds for the excess risk given by Theorem 1, one can decide whether to apply an algorithm in the initial domain or in the compressed domain. We now analyze the specific case of the ordinary Least-Squares Regression.

3.2 Application to ordinary Least-Squares Regression

The ordinary LS regression provides the regression function $f_{\widehat{\alpha}}$ where

$$\widehat{\alpha} = \operatorname*{argmin}_{\alpha\in\operatorname{argmin}_{\alpha'\in\mathbb{R}^N}||Y-\Phi\alpha'||} ||\alpha||.$$

Note that we have $\Phi\Phi^T\widehat{\alpha} = \Phi^T Y$, hence $\widehat{\alpha} = \Phi^\dagger Y \in \mathbb{R}^N$ where $\Phi^\dagger$ is the Penrose pseudo-inverse of Φ[1].

Following [12] we truncate the prediction to the level $\pm L$ where L is a bound (assumed to be known) on $||f^*||_\infty$. More precisely, our final predictor is : $\widehat{f}_L(x) \stackrel{\text{def}}{=} T_L[f_{\widehat{\alpha}}(x)]$, where

$$T_L(u) \stackrel{\text{def}}{=} \begin{cases} u & \text{if } |u| \leq L, \\ L\ \mathrm{sign}(u) & \text{otherwise} \end{cases}$$

[1] In the full rank case, $\Phi^\dagger = (\Phi^T\Phi)^{-1}\Phi^T$ when $K \geq N$ and $\Phi^\dagger = \Phi^T(\Phi\Phi^T)^{-1}$ when $K \leq N$

Truncation after the computation of the parameter $\widehat{\alpha} \in \mathbb{R}^N$, which is the solution of an unconstrained optimization problem, is easier than solving an optimization problem under the constraint that $||\alpha||$ is small (which is the approach followed in [23]) and allows for consistency results and prediction bounds (see [12]).

Indeed, using the notation of previous section, we have

$$\kappa_{\mathcal{A}}(\mathcal{D}_K, \mathcal{F}_N, P) = c' \max\{\sigma^2, L^2\} \frac{1 + \log K}{K} N, \quad \text{and} \quad c_{\mathcal{A}} = 8. \tag{6}$$

A bound on c' is 9216 (see [12]) and the $\log K$ term in $\kappa_{\mathcal{A}}$ is due to an analysis using the $\log$ entropy and covering numbers. Note that it seems possible to remove the $\log$ term using properties of the quadratic loss function, see [3], or by deriving tight bounds on the Rademacher complexity [13], but this goes beyond the scope of this paper.

However $\kappa'_{\mathcal{A}}$ is much nicer (which justifies its introduction). We have

$$\kappa'_{\mathcal{A}}(\mathcal{D}_K, \mathcal{F}_N, P) = \sigma^2 \frac{N}{K}, \quad \text{and} \quad c'_{\mathcal{A}} = 1.$$

Compressed Least-Squares Regression (CLSR): We use ordinary LSR in the compressed domain and define $\widehat{\beta} = \Psi^{\dagger} Y \in \mathbb{R}^M$, where Ψ is the $K \times M$ matrix with elements $(\psi_m(x_k))_{1 \le m \le M, 1 \le k \le K}$. Our CLSR predictor is defined as $\widehat{g}_L(x) \stackrel{\text{def}}{=} T_L[g_{\widehat{\beta}}(x)]$.

We deduce from Theorem 1 the excess risk of the CLSR estimator.

Corollary 1 *Let $M = \frac{||\alpha^+|| \sup_x ||\varphi(x)||}{\max(\sigma, L)} \sqrt{\frac{K}{\log K}}$. Then for any $\delta > 0$, with probability at least $1 - \delta$, we have*

$$\mathbb{E}(||\widehat{g}_L - f^*||_P^2) = O\Big(\max\{\sigma, L\} ||\alpha^+|| \sup_x ||\varphi(x)|| \frac{\log K/\delta}{\sqrt{K}} + \inf_{f \in \mathcal{F}_N} ||f - f^*||_P^2 \Big) \tag{7}$$

and by setting $M = \frac{||\alpha^+|| \sup_x ||\varphi(x)||}{\sigma} \sqrt{K}$, we have with probability at least $1 - \delta$,

$$\mathbb{E}_Y(||\widehat{g}_L - f^*||_Y^2) = O\Big(\max\{\sigma, L\} ||\alpha^+|| \sup_x ||\varphi(x)|| \sqrt{\frac{\log K/\delta}{K}} + \inf_{f \in \mathcal{F}_N} ||f - f^*||_Y^2 \Big)$$

Proof: Theorem 1 applied with the $\kappa_{\mathcal{A}}$ and $c_{\mathcal{A}}$ values of LSR in (6) gives:

$$\begin{aligned} \mathbb{E}(||\widehat{g}_L - f^*||_P^2) \quad \le \quad & c' \max\{\sigma^2, L^2\} \frac{1 + \log K}{K} M \\ & + 16 \Big(\varepsilon^2 ||\alpha^+||^2 \sup_x ||\varphi(x)||^2 \big(2 + \frac{\log 4/\delta}{K}\big) + \inf_{f \in \mathcal{F}_N} ||f - f^*||_P^2 \Big). \end{aligned}$$

By setting $M = \frac{||\alpha^+|| \sup_x ||\varphi(x)||}{\max(\sigma, L)} \sqrt{\frac{K}{\log K}}$, we deduce (7). Similarly the second result uses the definition of $\kappa'_{\mathcal{A}}$ and $c'_{\mathcal{A}}$, which gives

$$\mathbb{E}_Y(||\widehat{g}_L - f^*||_Y^2) \le \sigma^2 \frac{M}{K} + 2 \Big(\varepsilon^2 ||\alpha^+||^2 \sup_x ||\varphi(x)||^2 \big(2 + \frac{\log 4/\delta}{K}\big) + \inf_{f \in \mathcal{F}_N} ||f - f^*||_Y^2 \Big).$$

and we finally optimize on M. □

Remark 1 *Notice that the choice of M in the previous corollary depends on $||\alpha^+||$ which is a priori unknown. If we set M independently of $||\alpha^+||$, then an additional multiplicative factor of $||\alpha^+||$ appears in the bound.*

Complexity of CLSR: The complexity of LSR for computing the regression function in the compressed domain only depends on M and K, and is (see e.g. [4]) $\text{Cx}(\mathcal{D}_K, \mathcal{G}_M, P) = O(MK^2) = O(K^{5/2})$ for $M = O(\sqrt{K})$. However the leading term when using CLSR is the cost for building the Ψ matrix: $O(NK^{3/2})$.

4 Discussion

From Corollary 1, the estimation error of CLSR is $O(\log K/\sqrt{K})$. It is clear that whenever $N > \sqrt{K}$ (which is the case of interest here), this is better than the ordinary LSR in the initial domain, whose estimation error is $O(N \log K/K)$.

It is difficult to compare our results with LASSO (or the Dantzig selector that has similar properties [5]) for which the main concern is to design sparse regression functions or to recover a solution assumed to be sparse. From [11, 15, 24] one deduces that under some assumptions, the estimation error of LASSO is of order $S\frac{\log N}{K}$ where S is the sparsity of the best regressor f^+ in $\mathcal{F}_N$. If the sparsity $S < \sqrt{K}$ then this is more interesting than CLSR. However our method does not make any assumption about the sparsity of f^+ and our goal is not to recover f^+ but to make good predictions.

Now in terms of complexity issues, CLSR requires $O(NK^{3/2})$ operations to build the matrix and compute the regression function, whereas according to [18], the (heuristical) complexity of the LASSO algorithm is $O(NK^2)$ in the best cases (assuming that the number of step required for convergence is $O(K)$, which is not proved theoretically). Thus CLSR seems to be a good and simple competitor to LASSO.

Corollary 1 says that N appears in the bound only through the product of $||\alpha^+||$ and $\sup_x ||\varphi(x)||$. A natural question is whether this product can be made small for appropriate choices of features. We now show that this is actually the case for some wavelets where the dependency w.r.t. N actually vanishes.

Indeed consider a infinite orthogonal family of wavelets $(\varphi_n) = (\varphi_{h,l})$ (indexed by $n \geq 1$ or equivalently by the scale $h \geq 0$ and translation $0 \leq l \leq 2^h - 1$) where $\varphi_{h,l}(x) = c_h \varphi_0(2^h x - l)$, for some c_h coefficients, where φ_0 is the mother wavelet. Then we have the following result:

Proposition 2 *Assume that the mother wavelet has compact support and has at least p null moments, where $p > 0$, and that the function $f^+ = f_{\alpha^+}$ is p-Lipschitz with constant L. Then, setting $c_h = 2^{h(1-p)/2}$, we have that $||\alpha^+||$ and $\sup_x ||\varphi(x)||$ are finite and $||\alpha^+|| \sup_x ||\varphi(x)|| = L(\int \varphi_0) 2^p/(1 - 2^{-p})$.*

Proof: We use Theorem 6.3 of [14] to bound $||\alpha^+||$ and scale appropriately the wavelets using $c_h = 2^{h(1-p)/2}$ to minimize the product. □

For illustration, considering the Haar walevets (which are orthogonal, symmetrical wavelets with compact support, have $p = 1$ null moment, and are defined by the mother wavelet, $\varphi_0 = \mathbb{I}_{\{0 \leq x < 1/2\}} - \mathbb{I}_{\{1/2 \leq x < 1\}}$), if f^* is Lipschitz with constant L, then we have: $||\alpha^+|| \sup_x ||\varphi(x)|| \leq 4L$.

5 Conclusion

The general methodology described in Section 3.1 enables to use any algorithm (using linear function spaces) in the compressed domain, thus reducing the estimation error though at the price of a (controlled) increase of the approximation error. The numerical complexity of computing the regression function in the compressed domain is reduced, but additional projection costs appear.

We showed in Section 3.2 that a very interesting tradeoff is obtained in the case of ordinary LS regression. The excess risk of CLSR is bounded by $O(\sqrt{\log K/K} + \inf_{f \in \mathcal{F}_N} ||f - f^*||_P^2)$, which is comparable with the best available methods, and its numerical complexity is $O(NK^{3/2})$.

References

[1] Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687, June 2003.

[2] Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *STOC '06: Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563, New York, NY, USA, 2006. ACM.

[3] Jean-Yves Audibert and Olivier Catoni. Risk bounds in linear regression through pac-bayesian truncation. Technical Report HAL : hal-00360268, 2009.

[4] David Bau III and Lloyd N. Trefethen. *Numerical linear algebra*. Philadelphia: Society for Industrial and Applied Mathematics, 1997.

[5] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *To appear in Annals of Statistics*, 2008.

[6] Avrim Blum. Random projection, margins, kernels, and feature-selection. *Subspace, Latent Structure and Feature Selection*, pages 52–68, 2006.

[7] Robert Calderbank, Sina Jafarpour, and Robert Schapire. Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. *Technical Report*, 2009.

[8] Emmanuel Candes and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313, 2007.

[9] Emmanuel J. Candes and Justin K. Romberg. Signal recovery from random projections. volume 5674, pages 76–86. SPIE, 2005.

[10] Mark A. Davenport, Michael B. Wakin, and Richard G. Baraniuk. Detection and estimation with compressive measurements. Technical Report TREE 0610, Department of Electrical and Computer Engineering, Rice University, 2006.

[11] E. Greenshtein and Y. Ritov. Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli*, 10:971–988, 2004.

[12] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer-Verlag, 2002.

[13] Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Leon Bottou, editors, *Neural Information Processing Systems*, pages 793–800. MIT Press, 2008.

[14] Stephane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[15] Yuval Nardi and Alessandro Rinaldo. On the asymptotic properties of the group Lasso estimator for linear models. *Electron. J. Statist.*, 2:605–633, 2008.

[16] D. Pollard. *Convergence of Stochastic Processes*. Springer Verlag, New York, 1984.

[17] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Neural Information Processing Systems*, 2007.

[18] Saharon Rosset and Ji Zhu. Piecewise linear regularized solution paths. *Annals of Statistics*, 35:1012, 2007.

[19] Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[20] A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math Dokl 4*, pages 1035–1038, 1963.

[21] Yaakov Tsaig and David L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52:1289–1306, 2006.

[22] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[23] Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

[24] Tong Zhang. Some sharp performance bounds for least squares regression with L1 regularization. *To appear in Annals of Statistics*, 2009.

[25] Shuheng Zhou, John D. Lafferty, and Larry A. Wasserman. Compressed regression. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *Neural Information Processing Systems*. MIT Press, 2007.