Error Prediction and Model Selection via Unbalanced Expander Graphs

Yohann de Castro*

October 13, 2010

Abstract

This article investigates *deterministic* design matrices X for the fundamental problems of error prediction and model selection given observations $y = X\beta + z$, where z is a stochastic error term. We are interested in the so-called '$p \gg n$' setup where the number p of predictors is far more important than the number n of observations. Our deterministic design matrices are constructed from *unbalanced expander graphs*, and we wonder if it is possible to accurately estimate $X\beta$ and the support of β using *computationally tractable algorithms*.

We show that for any adjacency matrix of an unbalanced expander graph and any target vector $\beta^\star$, the lasso (ℓ_1-penalized least squares) and the Dantzig selector (ℓ_∞-penalized basis pursuit) satisfy oracle inequalities in error prediction and model selection involving the s largest (in magnitude) coefficients of $\beta^\star$, i.e. upper bounds in term of the *best sparse approximation*. Our oracle inequalities allow error prediction with an accuracy which is the best, up to a logarithmic factor, one could expect knowing the support of the target $\beta^\star$.

From a practical standpoint, these estimators can be computed by solving, either a simple quadratic program for the lasso, or a linear program for the Dantzig selector. Our results are non-asymptotic and describe the performance one can expect in all cases.

1 Introduction

This article focuses on the problem of processing high-dimensional data. Our framework is broadly the *compressive sensing* where one seeks to acquire the main information of a signal directly from a minimum of measurements. The field of applications is wide and encompasses compressive imaging, MRI (magnetic resonance imaging), NMR (Nuclear Magnetic resonance) spectroscopy, radar design, real-number error correction, communications and high-speed analog-to-digital conversions [Can06].

Beyond the wide spectrum of applications, a fundamental question is to find efficient design matrices for common estimators. Unlike the traditional approach that looks for random matrices, our goal is to find deterministic design matrices. Our present work is based on *unbalanced expander graphs* [BI08, JXHC09] that give outstanding explicit design matrices.

*Institut de Mathématiques de Toulouse CNRS UMR 5219, Équipe de Statistique et Probabilités, Université Paul Sabatier, 31062 Toulouse cedex 9, France.
E-mail: yohann.decastro@math.univ-toulouse.fr

1.1 The Deterministic Design Matrix

It emerged recently that compressive sensing and *coding theory* share similar properties. In 2007, B. Hassibi and W. Xu [HX07] gave a generalization of *expander codes* [SS96] (which are linear error-correcting codes derived from expander graphs) to compressive sensing. Furthermore, Berinde *et al.* [BGI$^+$08] pointed out that unbalanced expander graphs satisfy a restricted isometry property appeared in compressive sensing.

We recall that, in their fundamental article [CRT06], E. Candès, J. Romberg, and T. Tao showed that the standard RIP_2 property is a sufficient condition that enables compressive sensing using random projections. Intuitively, it says that the design matrix preserves the ℓ_2-norm of sparse vectors (i.e. it is an almost isometry on the space of sparse vectors). This property implies that recovery using ℓ_1 minimization (i.e. *basis pursuit*) is possible. In 2008, Berinde *et al.* showed that the adjacency matrix X of an expander graph satisfies a very similar property called the restricted isometry property in the ℓ_1-norm (RIP_1). They used this property to show that basis pursuit is still possible in this case. They proved a useful uncertainty principle connecting the mass on a small subset S, namely $\|\gamma_S\|_1$, to the whole mass $\|\gamma\|_1$. We use this last property to obtain oracle inequalities in error prediction and model selection.

Adjacency Matrix of a Bipartite Graph

We consider a *bipartite* graph $G = (A, B, E)$, where A is the set of the left vertices, B the set of the right vertices, and E the set of the edges between A and B. Denote p and n respectively the cardinality of A and B.

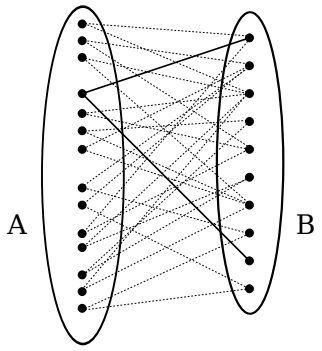

Figure 1: A bipartite graph G with regular left degree d. Each vertex in A has exactly d neighbors in B (here $d = 2$).

A bipartite graph is said to have regular left degree d if every vertex in A has exactly d neighbors in B, see Figure 1. Suppose that G has regular left degree d, then the *renormalized adjacency matrix* X is

$$X_{ij} = \begin{cases} 1/d & \text{if } i \text{ is connected to } j\,, \\ 0 & \text{otherwise}\,, \end{cases} \tag{1}$$

where $i \in [1, p]$ and $j \in [1, n]$. In the following, the design matrix X will always refer to the renormalized adjacency matrix of an unbalanced expander graph.

Recently, Guruswami *et al.* [GUV09] proved that there exist *explicit* unbalanced expander graphs which are very close, in terms of asymptotic upper bounds on the number of right

vertices n as p tends to $+\infty$, to the 'optimum' expanders build by random constructions. Quantitatively speaking, the optimum number of right vertices is such that

$$n = \underset{p\to+\infty}{\mathcal{O}}\left(s\log\left(\frac{p}{s}\right)\right),$$

where s is a parameter of the graph that can be interpreted as the number of largest coefficients that we want to recover, and the $\mathcal{O}(.)$ notation does not depend on s.

Uncertainty Principle

The main property exploited in this article is an uncertainty principle shown in [BGI+08]. For suitable parameters, it holds

$$\forall\gamma\in\mathbb{R}^p,\quad \|\gamma_S\|_1 \le 2\,\|X\gamma\|_1 + \frac{1}{2}\,\|\gamma_{S^c}\|_1\,, \tag{2}$$

where γ_S is the vector that is equal to γ on S and zero elsewhere. In the last inequality, the cardinality of S is upper bounded by a parameter s derived from the expansion property. From a practical view point, we are interested in sparse vectors such that $s \ll p$, but we will see in the next section that we can consider whatever value for the parameter s. The property (2) is derived form the fact that X preserves the ℓ_1-norm of vectors with small support (RIP_1 property). In particular, if γ belongs to the kernel of X it yields

$$\|\gamma_S\|_1 \le \frac{1}{2}\,\|\gamma_{S^c}\|_1\,. \tag{3}$$

This last inequality means that the vectors of the kernel can not be concentrated on small subsets. In fact, the inequality (3) is a *necessary and sufficient* condition for the basis pursuit estimator

$$\beta^{bp} = \arg\min_{\beta\in\mathbb{R}^p}\|\beta\|_1 \quad \text{such that } X\beta^{bp} = X\beta^\star\,,$$

exactly recovers the target vector $\beta^\star$. Thus we can see (2) as a *sufficient* condition for the lasso and the Dantzig selector which generalizes the condition (3).

1.2 Error Prediction and Model Selection

Two of the most common problems in statistics are to estimate the response $X\beta^\star$ (error prediction) and the support (model selection) of $\beta^\star$ from the data $y\in\mathbb{R}^n$ and the linear model

$$y = X\beta^\star + z\,,$$

where X is a design matrix, and $z\in\mathbb{R}^n$ a noise vector. We assume that z is a *Gaussian white noise*, and we show, using Lemma 3, that we can consider any *correlated* Gaussian white noise. This means that the z_i's have same Gaussian law but they could be correlated, with $z = (z_1,\dots,z_n)$.

We introduce a so-called *bound on the noise* keeping in mind that it allows us to set the threshold of our tuning parameter λ of our estimators. Denote

$$\Lambda = 2\,\sigma\,\sqrt{\log n}\,,$$

where σ is the variance of the noise.

The Lasso

In his fundamental article [Tib96] R. Tibshirani pointed out that the geometry of the ℓ_1-norm produces coefficients that are exactly 0. The *lasso estimator* is

$$\beta^l = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} ,$$

where λ is a tuning parameter. Intuitively, the lasso estimator will be at the point of contact of this smooth residual sum of squares function and convex, piecewise-flat constraint surface. This point of contact is very likely to belong to a k-face (i.e. the k-simplex generated by k extremal points) of a ball ℓ_1. Thus it is very likely to have a lot of coefficients that are exactly 0, see Figure 2.

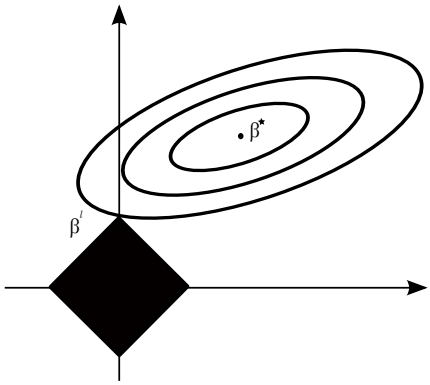

Figure 2: The lasso estimator produces coefficients that are are exactly 0. The black square represents a ball in the ℓ_1-norm, while the ellipses represent the level sets of the quadratic criterion $\|y - X\beta\|_2^2$. For simplicity, this figure is derived from the noiseless case where $z = 0$. In this case, $\beta^\star$ is at the center of the quadratic criterion $\|y - X\beta\|_2^2$. In the noisy case, one has to replace $\beta^\star$ by $\beta^\star + \zeta$ where ζ is such that $X\zeta$ is the orthogonal projection of the noise z onto the subspace spanned by the columns of X.

In this paper, we prove that, with high probability, for any $\lambda \geq 6\Lambda$ and any target vector $\beta^\star$,

$$\left\| X\beta^\star - X\beta^l \right\|_2^2 + (\lambda - 6\Lambda) \left\| \beta^l_{S^c} - \beta^\star_{S^c} \right\|_1 \leq 4\lambda \left(2\lambda n + \|\beta^\star_{S^c}\|_1 \right) , \tag{4}$$

where n is the number of measurements (i.e. number of lines of the matrix X). Remark that we do not suppose that the cardinality of the support of $\beta^\star$ is upper bounded. In fact, the inequality (4) stands for **all** target vectors in $\mathbb{R}^p$.

If the target vector $\beta^\star$ is *sparse*, denote S its support (the set of all nonzero entries) and suppose that $|S| \leq s$. In this case, with high probability, it holds

$$\left\| X\beta^\star - X\beta^l \right\|_2^2 + (\lambda - 6\Lambda) \left\| \beta^l_{S^c} \right\| \leq 8\lambda^2 n \, .$$

If one takes $\lambda = 6\Lambda$, the last inequality yields

$$\left\| X\beta^\star - X\beta^l \right\|_2 \leq 24\sqrt{2}\,\sigma\sqrt{n \log n} \, , \tag{5}$$

for any s-sparse target vector $\beta^\star$. The inequality (5) shows that we can estimate $X\beta^\star$ with nearly the same precision as if one knew in advance the support of $\beta^\star$. Indeed, consider the *ordinary least squares*:

$$\beta^{ols} = \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2 \;\text{ such that } \operatorname{supp}\left(\beta^{ols}\right) = S\,,$$

where $\operatorname{supp}\left(\beta^{ols}\right)$ denotes the support of β^{ols}. Observe that this estimator uses a prior knowledge on the support of $\beta^\star$. For this reason, we can say that this estimator is *optimal*. We claim that $X\beta^{ols} - X\beta^\star$ is the orthogonal projection of z on the subspace spanned by the X_i's with $i \in S$. Hence, a simple calculation gives

$$\mathbb{E}\left\|X\beta^{ols} - X\beta^\star\right\|_2 \leq \sigma\sqrt{s}\,.$$

Moreover we know that $n = \mathcal{O}\left(s\log\left(\frac{p}{s}\right)\right)$ for optimum unbalanced expander graphs. In this case, we deduce that the inequality (5) is optimal within the square root of logarithmic factors. Namely, it holds

$$\left\|X\beta^\star - X\beta^l\right\|_2 \leq C\,.\sqrt{\log\left(\frac{p}{s}\right)\log(n)}\,.\sigma\sqrt{s}\,, \tag{6}$$

where C is some positive numerical constant. Since $'n \ll p'$, the $\log n$ term is not large compared to $\log p$.

In 2007, E. J. Candès and Y. Plan obtained a remarkable estimate in error prediction via the lasso. They used a so-called *coherence property* following the work of D.L. Donoho *et al.* [DET06]. They showed (Theorem 1.2 in [CP09]) that, with high probability, for every design matrix satisfying the coherence property, it holds

$$\left\|X\beta^\star - X\beta^l\right\|_2 \leq C'\,.\sqrt{\log(p)}\,.\sigma\sqrt{s}\,, \tag{7}$$

where C' is some positive numerical constant. Note that the upper bounds (6) and (7) are similar. The coherence is the maximum correlation between pairs of predictors. This property is fundamental and allows to deal with random design matrices. We do not use this property here, though we get the same accuracy and we extend their error prediction result to *deterministic* design matrices.

The Dantzig Selector

In 2005, E. Candès and T. Tao [CT07a] gave a new estimator, the Dantzig selector. This estimator is the solution to the ℓ_1-regularization problem

$$\beta^d = \arg\min_{\beta \in \mathbb{R}^p} \|\beta\|_1 \;\text{ s.t. }\; \left\|X^T(y - X\beta)\right\|_\infty \leq \lambda\,,$$

where $\|.\|_\infty$ is the ℓ_∞-norm and λ a tuning parameter. We consider tuning parameters λ such that $\lambda \geq \Lambda$. This last inequality involves that $\beta^\star$ is *feasible* with high probability. We prove that, with high probability, for any target vector $\beta^\star$ and $\lambda \geq \Lambda$,

$$\left\|X\beta^\star - X\beta^d\right\|_2^2 \leq 4(\lambda+\Lambda)\left(16(\lambda+\Lambda)n + 3\left\|\beta^\star_{S^c}\right\|_1\right)\,.$$

In the case where the target vector $\beta^\star$ is sparse then, with high probability, it holds

$$\left\| X\beta^\star - X\beta^d \right\|_2 \leq 8(\lambda + \Lambda)\sqrt{n}\,.$$

If $\lambda = \Lambda$, we derive the error prediction:

$$\left\| X\beta^\star - X\beta^l \right\|_2 \leq 32\,\sigma\sqrt{n \log n}\,. \tag{8}$$

As for (5), the inequality (8) shows that we can estimate $X\beta^\star$ with nearly the same accuracy one would get if he knew in advance the support of $\beta^\star$.

We provide an upper bound on the error of the Dantzig selector in model selection. Consider a s-sparse target vector $\beta^\star$ with support S. We show that, with high probability, for $\lambda \geq \Lambda$,

$$\left\| \beta^d_{S^c} \right\|_1 \leq 32(\lambda + \Lambda)n\,.$$

In the case $\lambda = \Lambda$, it holds

$$\left\| \beta^d_{S^c} \right\|_1 \leq 128\,\sigma\, n\sqrt{\log n}\,.$$

In the $'n \ll p'$ setup, observe that the error vector $\beta^d_{S^c}$ has a size almost equal to p, whereas the upper bound is *much smaller* than p. This last inequality estimates the error of the Dantzig selector in model selection.

1.3 Organization of the paper

The outline of the paper is as follows. The second section presents unbalanced expander graphs and recalls the uncertainty principle of Berinde *et al.*. The third section introduces the parameter Λ and gives, with high probability, upper bounds on the ℓ_∞-norm of the noisy covariance. The fourth section studies the lasso estimator and gives oracle inequalities in term of the best sparse approximation, error prediction, and model selection. Finally the fifth section presents the Dantzig selector and gives upper bounds in error prediction and model selection.

2 Uncertainty Principle

In this section we introduce unbalanced expander graphs and recall the main results shown by Berinde *et al.*. The main property of expander graphs is a property of *expansion*. In the case of unbalanced expander graphs, this property controls the neighborhood J of any sufficiently small subset I of vertices on the left. Let $G = (A, B, E)$ be a bipartite graph with A the set of left vertices, B the set of right vertices, and E the set of edges between A and B. We recall that p and n denote respectively the cardinality of A and B. The size n may depend on p and others parameters of the graph. Suppose that G has regular left degree d. Hence, every subset $I \subset A$ has at most $d\,|I|$ neighbors. The expansion property states that the neighborhood of I is 'almost' $d\,|I|$ as soon as $|I| \leq s$, where s is a parameter of the graph that can be as large as desired, see Figure 3. The formal definition of unbalanced expander graph is as follows.

Definition 1 ((s,ε)-unbalanced expander) — *An (s,ε)-unbalanced expander is a bipartite simple graph $G = (A, B, E)$ with left degree d such that for any $I \subset A$ with $|I| \leq s$, the set of neighbors J of I has size*

$$|J| \geq (1 - \varepsilon)\, d\,|I|\,. \tag{9}$$

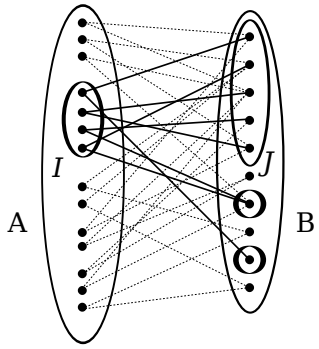

Figure 3: The expansion property of an unbalanced expander graph: any sufficiently small subset I on the left has a neighborhood J of size at least $(1-\varepsilon)\, d\, |I|$.

We recall that X is the renormalized adjacency matrix of an unbalanced expander graph. The reader may find a definition of X in (1). The parameter s can be as large as possible. However in the $n \ll p$ setup, we deal with the values for which $s \ll p$. Indeed we recall that n is of the order of s within a logarithmic factor. Subsequently we consider a parameter ε such that $\varepsilon \leq 1/8$. Remark that ε is fixed and **does not** depend on others parameters. In particular, we do not require that ε goes to zero as p goes to the infinity. We call ε the *expansion constant*. Using the expansion property (9), Berinde *et al.* showed the fundamental theorem:

Theorem 1 (Restricted Isometry Property) — *Let X be the renormalized adjacency matrix of an (s,ε)-unbalanced expander. Then X satisfies the following RIP_1 property:*

$$\forall \gamma \in \mathbb{R}^p, \;\; (1-2\varepsilon)\, \|\gamma_S\|_1 \leq \|X\gamma_S\|_1 \leq \|\gamma_S\|_1 \;,$$

where S is any subset of $[1,p]$ of size less than s, and γ_S the vector with coefficients equal to the coefficients of γ in S and zero outside.

In their article [BGI$^+$08] (Lemma 16 and Theorem 17), Berinde *et al.* derive a useful lemma which is a consequence of the RIP_1 property. In fact, this lemma can seen as an uncertainty principle and we show in the next sections that it is a sufficient condition for error prediction and model selection.

Lemma 1 (Uncertainty Principle) — *Let X be the renormalized adjacency matrix of an (s,ε)-unbalanced expander with $\varepsilon < 1/4$. Then X satisfies the following uncertainty principle:*

$$\forall \gamma \in \mathbb{R}^p \,,\; \forall S \subset [1,p] \text{ s.t. } |S| \leq s \,,\;\; (1-4\varepsilon)\, \|\gamma_S\|_1 \leq \|X\gamma\|_1 + 2\varepsilon\, \|\gamma_{S^c}\|_1 \;.$$

In particular for $\varepsilon \leq 1/8$, it yields

$$\forall \gamma \in \mathbb{R}^p \,,\; \forall S \subset [1,p] \text{ s.t. } |S| \leq s \,,\;\; \|\gamma_S\|_1 \leq 2\, \|X\gamma\|_1 + \frac{1}{2}\, \|\gamma_{S^c}\|_1 \;. \tag{10}$$

As mentioned in the introduction, this uncertainty principle can be seen as a sufficient condition for the lasso and the Dantzig selector that generalizes the condition (3) of the basis pursuit.

We conclude this section by introducing the work of Guruswami *et al.* on the explicit construction of unbalanced expander graphs. They recently proved [GUV09], based on the *Parvaresh-Vardy codes* [PV05], the theorem:

Theorem 2 (Explicit Construction) — *For any $\alpha > 0$ and any p, s, $\varepsilon > 0$, there exists an (s, ε)-unbalanced expander $G = (A, B, E)$ with $|A| = p$, left degree*

$$d = \underset{p \to +\infty}{\mathcal{O}} \left((\log p)^{1+\frac{1}{\alpha}} \right) ,$$

and number of right side vertices (namely $n = |B|$),

$$n = \underset{p \to +\infty}{\mathcal{O}} \left(s^{1+\alpha} (\log p)^{2+\frac{2}{\alpha}} \right) ,$$

where the $\mathcal{O}(.)$ notation does not depend on s but on ε.

The bounds may depend on ε, however our parameter ε is fixed and does not depend on p. In a probabilistic framework, the following proposition can be shown using *Chernoff Bounds* [HX07].

Proposition 1 (Probabilistic Construction) — *Consider $\varepsilon > 0$ and $p/2 \geq s$. Then, with a positive probability, there exists an (s, ε)-unbalanced expander $G = (A, B, E)$ with $|A| = p$, left degree*

$$d = \underset{p \to +\infty}{\mathcal{O}} \left(\log \left(\frac{p}{s} \right) \right) ,$$

and number of right side vertices (namely $n = |B|$),

$$n = \underset{p \to +\infty}{\mathcal{O}} \left(s \log \left(\frac{p}{s} \right) \right) ,$$

where the $\mathcal{O}(.)$ notation does not depend on s but on ε.

In the following discussion, we denote by n the number of measurement (i.e. the size of B). These theorems show that it is possible to construct explicit unbalanced expander graphs close, in terms of the bound on n, to the optimum graphs obtained probabilistically.

3 Bound on the Noise

In this section, we present the notation of our linear model and we give an upper bound on the noise amplification $\left\| X^T z \right\|_\infty$. We seek to reconstruct a high dimensional vector $\beta^\star \in \mathbb{R}^p$ from noisy observation $y \in \mathbb{R}^n$. We consider a linear model

$$y = X\beta^\star + z \,, \tag{11}$$

where X is the renormalized adjacency matrix of an unbalanced expander graph, and $z \in \mathbb{R}^n$ a *Gaussian white noise* with variance σ^2. We start with a lemma which shows that the noise is not amplified by the graph.

Lemma 2 (Non-Amplification) — *It holds*

$$\forall z \in \mathbb{R}^n \,, \ \left\| X^T z \right\|_\infty \leq \|z\|_\infty \,.$$

Proof — Let $\gamma \in \mathbb{R}^p$ such that $\|\gamma\|_1 = 1$. Since the graph has left degree d (see (1)),

$$\|X\gamma\|_1 \leq \|\gamma\|_1 .$$

Remark that this inequality stands for all vectors, not only sparse vectors. Next we dualize,

$$\left\langle X^T z, \gamma \right\rangle = \langle z, X\gamma \rangle \leq \|z\|_\infty \|X\gamma\|_1 \leq \|z\|_\infty \|\gamma\|_1 \leq \|z\|_\infty ,$$

where $\langle \, , \, \rangle$ is the standard Euclidean product. This last inequality ends the proof. □

In order to upper bound $\left\|X^T z\right\|_\infty$ it is enough to estimate $\|z\|_\infty$, which allows us to reduce the dimension of the ambient space from p to n.

Lemma 3 (Bound on the Noise) — *Suppose that z is a white centered Gaussian noise with variance σ (i.e. $z = (z_i)_{i=1\ldots n}$ with z_i i.i.d. $\mathcal{N}\left(0, \sigma^2\right)$-distributed). Then, for*

$$\Lambda = 2\,\sigma\sqrt{\log n}\,,$$

$$\mathbb{P}\left(\left\|X^T z\right\|_\infty \leq \Lambda\right) \geq 1 - \frac{1}{\sqrt{2\pi}\, n \sqrt{\log n}}\,.$$

Proof —Denote $(z_i)_{i=1\ldots n}$ the coefficients of z. The Lemma 2 gives

$$\mathbb{P}\left(\left\|X^T z\right\|_\infty \leq \Lambda\right) \geq \mathbb{P}\left(\|z\|_\infty \leq \Lambda\right) = \prod_{i=1}^{n} \mathbb{P}\left(|z_i| \leq \Lambda\right) , \tag{12}$$

because the z_i's are i.i.d.. Denote Φ and φ respectively the *cumulative distribution function* and the probability density function of the standard normal. Set $\delta = 2\sqrt{\log n}$. It holds

$$\begin{aligned} \prod_{i=1}^{n} \mathbb{P}\left(|z_i| \leq \Lambda\right) &= \mathbb{P}\left(|z_1| \leq \Lambda\right)^n , \\ &= \left(2\Phi(\delta) - 1\right)^n , \\ &> \left(1 - 2\,\frac{\varphi(\delta)}{\delta}\right)^n , \end{aligned}$$

where we used an integration by parts to show that

$$1 - \Phi(\delta) = \int_\delta^{+\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\,\mathrm{d}t < \frac{\varphi(\delta)}{\delta}\,.$$

We complete the proof with

$$\mathbb{P}\left(\left\|X^T z\right\|_\infty \leq \Lambda\right) \geq \left(1 - 2\,\frac{\varphi(\delta)}{\delta}\right)^n \geq 1 - 2n\frac{\varphi(\delta)}{\delta} = 1 - \frac{1}{\sqrt{2\pi}\, n \sqrt{\log n}}\,.$$

□

Using Šidák's inequality in (12), we can dispense with the assumption of independence. Indeed, even for correlated z_i's, it holds [Šid68]:

$$\mathbb{P}\left(\|z\|_\infty \leq \Lambda\right) \geq \mathbb{P}\left(\|\widetilde{z}\|_\infty \leq \Lambda\right) = \prod_{i=1}^{n} \mathbb{P}\left(|\widetilde{z}_i| \leq \Lambda\right) ,$$

where the $\widetilde{z}_i$'s are independent and have the same law as the z_i's. The lemma extends to correlated z_i's as long as they have the same Gaussian law. Hence we can consider any *correlated* Gaussian white noise in our linear model (11). This upper bound is valuable to give oracle inequalities, as we shall see in subsequent sections. For readability sake, denote

$$\eta_n = \frac{1}{\sqrt{2\pi}\, n \sqrt{\log n}}\,.$$

All the probabilities appearing in our theorems are of the form $1-\eta_n$. Since n denote the number of observations, η_n is very small (less than $1/1000$ for most common problems). Furthermore, by repeating the same argument as in Lemma 3, we have the next proposition.

Proposition 2 — *Suppose that z is a correlated white Gaussian noise with variance σ (i.e. $z = (z_i)_{i=1\dots n}$ where the z_i's are $\mathcal{N}\left(0,\sigma^2\right)$-distributed). Then, for $\alpha \geq 1$ and*

$$\Lambda_\alpha = (1+\alpha)\,\sigma\sqrt{\log n}\,,$$

$$\mathbb{P}\left(\left\|X^T z\right\|_\infty \leq \Lambda_\alpha\right) \geq 1 - \frac{\sqrt{2}}{(1+\alpha)\sqrt{\pi \log n}\, n^{\frac{(1+\alpha)^2}{2}-1}}\,. \tag{13}$$

By replacing Λ by Λ_α in the statements of our theorems, it is possible to replace all the probabilities of the form $1-\eta_n$ by probabilities of the form (13) . Observe that these probabilities can be as small as desired.

4 Oracle Inequalities for the Lasso

The lasso estimator is

$$\beta^l = \arg\min_{\beta\in\mathbb{R}^p}\left\{\|y - X\beta\|_2^2 + \lambda\,\|\beta\|_1\right\}\,,$$

where λ is a tuning parameter. In their article [vdGB09], P. Bühlmann and S. van de Geer give the *weakest* condition on X, a so-called *compatibility condition*, to have oracle inequalities for the lasso. We do not use this prominent approach here, although the uncertainty principle (10) makes it possible to verify the compatibility condition. We did not choose this approach because it provides upper bounds far greater than what we obtain by directly using the uncertainty principle. We recall that $\Lambda = 2\sigma\sqrt{\log n}$.

Theorem 3 — *Let X be the renormalized adjacency matrix of an (s,ε)-unbalanced expander with expansion constant $\varepsilon \leq 1/8$. Let $\beta^\star$ be any vector of $\mathbb{R}^p$ and S its s largest (in magnitude) coefficients. Take $\lambda \geq 6\Lambda$ then it holds*

$$\left\|X\beta^\star - X\beta^l\right\|_2^2 + (\lambda - 6\Lambda)\left\|\beta^l_{S^c} - \beta^\star_{S^c}\right\|_1 \leq 4\lambda\left(2\lambda n + \|\beta^\star_{S^c}\|_1\right)\,,$$

with probability at least $1-\eta_n$.

We show this theorem using the upper bound on the noise given by Lemma 3, and invoking the uncertainty principle given by Lemma 1.

Proof — Set $\gamma = \beta^\star - \beta^l$. On the event $\left\{\left\|X^T z\right\|_\infty \leq \Lambda\right\}$, it holds

$$\begin{aligned}
\|X\gamma\|_2^2 + \lambda \left\|\beta^l\right\|_1 &= \left\|y - z - X\beta^l\right\|_2^2 + \lambda \left\|\beta^l\right\|_1 , \\
&= \left\|y - X\beta^l\right\|_2^2 - 2\, z^T \left(y - X\beta^l\right) + \|z\|_2^2 + \lambda \left\|\beta^l\right\|_1 , \\
&= \left\|y - X\beta^l\right\|_2^2 - 2 \left(X^T z\right)^T \gamma - \|z\|_2^2 + \lambda \left\|\beta^l\right\|_1 , \\
&\leq \left\|y - X\beta^l\right\|_2^2 + 2\Lambda \|\gamma\|_1 - \|z\|_2^2 + \lambda \left\|\beta^l\right\|_1 , \qquad (14) \\
&\leq \|y - X\beta^\star\|_2^2 + 2\Lambda \|\gamma\|_1 - \|z\|_2^2 + \lambda \|\beta^\star\|_1 , \qquad (15) \\
&= 2\Lambda \|\gamma\|_1 + \lambda \|\beta^\star\|_1 ,
\end{aligned}$$

using the definition of the lasso estimator in the inequality (15) and the event $\left\{\left\|X^T z\right\|_\infty \leq \Lambda\right\}$ in the inequality (14). It follows that

$$\begin{aligned}
\|X\gamma\|_2^2 + \lambda \left\|\beta^l_{S^c}\right\|_1 - 2\Lambda \|\gamma_{S^c}\|_1 &\leq 2\Lambda \|\gamma_S\|_1 + \lambda \left(\|\beta^\star_S\|_1 - \left\|\beta^l_S\right\|_1\right) + \lambda \|\beta^\star_{S^c}\|_1 , \\
&\leq (\lambda + 2\Lambda) \|\gamma_S\|_1 + \lambda \|\beta^\star_{S^c}\|_1 .
\end{aligned}$$

Hence we get

$$\|X\gamma\|_2^2 + (\lambda - 2\Lambda) \|\gamma_{S^c}\|_1 \leq (\lambda + 2\Lambda) \|\gamma_S\|_1 + 2\lambda \|\beta^\star_{S^c}\|_1 .$$

We recall the uncertainty principle (10) shown by Berinde *et al.*:

$$\forall \gamma \in \mathbb{R}^p, \quad \|\gamma_S\|_1 \leq 2 \|X\gamma\|_1 + \frac{1}{2} \|\gamma_{S^c}\|_1 .$$

Combining the two last inequalities, it holds

$$\begin{aligned}
\|X\gamma\|_2^2 + \frac{\lambda - 6\Lambda}{2} \|\gamma_{S^c}\|_1 &\leq 2 (\lambda + 2\Lambda) \|X\gamma\|_1 + 2\lambda \|\beta^\star_{S^c}\|_1 , \\
&\leq 2 (\lambda + 2\Lambda) \sqrt{n} \|X\gamma\|_2 + 2\lambda \|\beta^\star_{S^c}\|_1 ,
\end{aligned}$$

We deduce the inequality:

$$\|X\gamma\|_2^2 + (\lambda - 6\Lambda) \|\gamma_{S^c}\|_1 \leq 4 (\lambda + 2\Lambda)^2 n + 4\lambda \|\beta^\star_{S^c}\|_1 ,$$

Since $\lambda \geq 6\Lambda$, we get

$$\left\|X\beta^\star - X\beta^l\right\|_2^2 + (\lambda - 6\Lambda) \left\|\beta^l_{S^c} - \beta^\star_{S^c}\right\| \leq 4\lambda \left(2\lambda n + \|\beta^\star_{S^c}\|_1\right) .$$

Using Lemma 3, we pretend that the event $\left\{\left\|X^T z\right\|_\infty \leq \Lambda\right\}$ has probability at least $1 - \eta_n$. This concludes the proof. □

If $\beta^\star$ is *s-sparse* (i.e. it has at most s nonzero coefficients), we derive the next result.

Proposition 3 (Sparse Lasso) — *Let X be the renormalized adjacency matrix of an (s, ε)-unbalanced expander with expansion constant $\varepsilon \leq 1/8$. Let $\beta^\star$ be a s-sparse vector (i.e. with only s nonzero entries) and S its support. Take $\lambda \geq 6\Lambda$ then it holds*

$$\left\|X\beta^\star - X\beta^l\right\|_2^2 + (\lambda - 6\Lambda) \left\|\beta^l_{S^c}\right\|_1 \leq 8\lambda^2 n ,$$

with probability at least $1-\eta_n$. *In the case* $\lambda = 6\Lambda$, *we derive the error prediction:*

$$\left\| X\beta^\star - X\beta^l \right\|_2 \leq 24\sqrt{2}\,\sigma\sqrt{n\log n}\,, \tag{16}$$

with probability at least $1-\eta_n$, *and* σ *the variance of the noise.*

This proposition is a direct consequence of Theorem 3. Our oracle inequalities give the error of prediction $\left\| X\beta^\star - X\beta^l \right\|_2$ and selection $\left\| \beta^l_{S^c} \right\|_1$ based on the best spare approximation $\left\| \beta^\star_{S^c} \right\|_1$. Moreover, as mentioned in the introduction, the inequality (16) is *optimal* within the square root of logarithmic factors.

5 Prediction and Selection via the Dantzig Selector

In their article [CT07a] E. Candès and T. Tao introduced a new estimator, the *Dantzig Selector*. It is defined by

$$\beta^d = \arg\min_{\beta\in\mathbb{R}^p} \|\beta\|_1 \quad \text{s.t.} \quad \left\| X^T(y - X\beta) \right\|_\infty \leq \lambda\,,$$

where λ is a tuning parameter. Using the Lemma 3, observe that for $\lambda \geq \Lambda$ the vector $\beta^\star$ is *feasible* with high probability (i.e. $\beta^\star$ satisfies the inequality $\left\| X^T(y - X\beta) \right\|_\infty \leq \lambda$). Using the uncertainty principle (10) we prove the next theorem.

Theorem 4 — *Let* X *be the renormalized adjacency matrix of an* (s,ε)*-unbalanced expander with expansion constant* $\varepsilon \leq 1/8$. *Let* $\beta^\star$ *be any vector of* $\mathbb{R}^p$ *and* S *its* s *largest* (*in magnitude*) *coefficients. For* $\lambda \geq \Lambda$, *it holds*

$$\left\| X\beta^\star - X\beta^d \right\|_2^2 \leq 4(\lambda+\Lambda)\left(16(\lambda+\Lambda)n + 3\left\|\beta^\star_{S^c}\right\|_1\right)\,.$$

with probability at least $1-\eta_n$.

Proof — Set $\gamma = \beta^\star - \beta^d$. On the event $\left\{ \left\| X^T z \right\|_\infty \leq \Lambda \right\}$, it yields

$$\begin{aligned} \|X\gamma\|_2^2 &\leq \left\| X^T X\gamma \right\|_\infty \|\gamma\|_1 \\ &= \left\| X^T\left(y - X\beta^d\right) + X^T\left(X\beta^\star - y\right) \right\|_\infty \|\gamma\|_1 \\ &\leq (\lambda+\Lambda)\,\|\gamma\|_1\,. \end{aligned}$$

Hence we get

$$\|X\gamma\|_2^2 - (\lambda+\Lambda)\,\|\gamma_{S^c}\|_1 \leq (\lambda+\Lambda)\,\|\gamma_S\|_1\,. \tag{17}$$

Moreover, using the fact that $\beta^\star$ is feasible, it holds

$$\left\| \beta^d \right\|_1 \leq \|\beta^\star\|_1\,.$$

Thus,

$$\begin{aligned} \left\| \beta^d_{S^c} \right\|_1 &\leq \left(\|\beta^\star_S\|_1 - \left\| \beta^d_S \right\|_1 \right) + \|\beta^\star_{S^c}\|_1 \\ &\leq \|\gamma_S\|_1 + \|\beta^\star_{S^c}\|_1 \end{aligned}$$

Since $\|\gamma_{S^c}\|_1 \leq \left\|\beta^d_{S^c}\right\|_1 + \|\beta^\star_{S^c}\|_1$, it yields

$$\|\gamma_{S^c}\|_1 \leq \|\gamma_S\|_1 + 2\,\|\beta^\star_{S^c}\|_1 \,. \tag{18}$$

Combining $(17) + 3(\lambda+\Lambda)(18)$, we get

$$\|X\gamma\|_2^2 + 2(\lambda+\Lambda)\,\|\gamma_{S^c}\|_1 \leq 4(\lambda+\Lambda)\,\|\gamma_S\|_1 + 6(\lambda+\Lambda)\,\|\beta^\star_{S^c}\|_1 \,.$$

We recall the uncertainty principle (10) shown by Berinde *et al.*:

$$\forall \gamma \in \mathbb{R}^p, \;\; \|\gamma_S\|_1 \leq 2\,\|X\gamma\|_1 + \frac{1}{2}\,\|\gamma_{S^c}\|_1 \,.$$

Using the two last inequalities,

$$\begin{aligned} \|X\gamma\|_2^2 &\leq 8(\lambda+\Lambda)\,\|X\gamma\|_1 + 6(\lambda+\Lambda)\,\|\beta^\star_{S^c}\|_1 \,, \\ &\leq 8(\lambda+\Lambda)\sqrt{n}\,\|X\gamma\|_2 + 6(\lambda+\Lambda)\,\|\beta^\star_{S^c}\|_1 \,. \end{aligned}$$

We deduce the inequality:

$$\|X\gamma\|_2^2 \leq 64(\lambda+\Lambda)^2 n + 12(\lambda+\Lambda)\,\|\beta^\star_{S^c}\|_1 \,.$$

Finally, it holds

$$\left\|X\beta^\star - X\beta^d\right\|_2^2 \leq 4(\lambda+\Lambda)\left(16(\lambda+\Lambda)n + 3\,\|\beta^\star_{S^c}\|_1\right) \,.$$

Using Lemma 3, we pretend that the event $\left\{\left\|X^T z\right\|_\infty \leq \Lambda\right\}$ has probability at least $1-\eta_n$. This concludes the proof. □

If $\beta^\star$ is s-sparse, we derive the next result.

Proposition 4 — *Let X be the renormalized adjacency matrix of an (s,ε)-unbalanced expander with expansion constant $\varepsilon \leq 1/8$. Let $\beta^\star$ be a s-sparse vector. Then, for $\lambda \geq \Lambda$,*

$$\left\|X\beta^\star - X\beta^d\right\|_2 \leq 8(\lambda+\Lambda)\sqrt{n}\,, \tag{19}$$

with probability at least $1-\eta_n$. In the case $\lambda = \Lambda$, we derive the error prediction:

$$\left\|X\beta^\star - X\beta^l\right\|_2 \leq 32\,\sigma\sqrt{n\log n}\,,$$

with probability at least $1-\eta_n$, and σ the variance of the noise.

This proposition is a direct consequence of Theorem 4. As mentioned in the introduction, our result is optimal within the square root of logarithmic factors. In fact, we achieve nearly the same accuracy that one would get if he knew a head in advance the support of $\beta^\star$. By repeating the proof of the Theorem 4, we derive a result in model selection.

Proposition 5 — *Let X be the renormalized adjacency matrix of an (s,ε)-unbalanced expander with expansion constant $\varepsilon \leq 1/8$. Let $\beta^\star$ be a s-sparse vector and S be its support. Then, for $\lambda \geq \Lambda$,*

$$\left\|\beta^d_{S^c}\right\|_1 \leq 32(\lambda+\Lambda)n\,,$$

with probability at least $1-\eta_n$. In the case $\lambda = \Lambda$, we derive the model selection:

$$\left\|\beta^d_{S^c}\right\|_1 \leq 128\,\sigma\, n\sqrt{\log n}\,, \tag{20}$$

with probability at least $1-\eta_n$, and σ the variance of the noise.

Proof — Set $\gamma = \beta^\star - \beta^d$. On the event $\left\{\left\|X^T z\right\|_\infty \leq \Lambda\right\}$, the inequality (18) holds. Since $\beta^\star$ is s-sparse, this inequality yields

$$\|\gamma_{S^c}\|_1 \leq \|\gamma_S\|_1 \ .$$

Using the uncertainty principle (10), we deduce that

$$\begin{aligned} \|\gamma_{S^c}\|_1 &\leq 4\,\|X\gamma\|_1 \ , \\ &\leq 4\sqrt{n}\,\|X\gamma\|_2 \ . \end{aligned}$$

We conclude invoking (19). □

This result allows us to estimate the performance of our estimator in model selection. Observe that, in the $'n \ll p'$ setup, the set S^c has a size almost equal to p. Moreover the upper bound in (20) is much smaller than p. Thus the inequality (20) controls the error vector $\beta^d_{S^c}$ by an upper bound much smaller than its size.

Acknowledgments — The author wishes to thank especially Jean-Marc Azaïs and Franck Barthe for their unconditional support throughout this paper.

References

[BGI+08] R. Berinde, A. Gilbert, P. Indyk, H. Karloff, and M. Strauss. Combining geometry and combinatorics: a unified approach to sparse signal recovery. 2008.

[BI08] R. Berinde and P. Indyk. Sparse recovery using sparse random matrices. 2008.

[Can06] Emmanuel J. Candès. Compressive sampling. In *International Congress of Mathematicians. Vol. III*, pages 1433–1452. Eur. Math. Soc., Zürich, 2006.

[CP09] Emmanuel J. Candès and Yaniv Plan. Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.*, 37(5A):2145–2177, 2009.

[CRT06] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2006.

[CT07a] Emmanuel Candès and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[CT07b] Emmanuel Candès and Terence Tao. Rejoinder: "The Dantzig selector: statistical estimation when p is much larger than n" [Ann. Statist. **35** (2007), no. 6, 2313–2351; mr2382644]. *Ann. Statist.*, 35(6):2392–2404, 2007.

[DET06] David L. Donoho, Michael Elad, and Vladimir N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1):6–18, 2006.

[GUV09] Venkatesan Guruswami, Christopher Umans, and Salil Vadhan. Unbalanced expanders and randomness extractors from Parvaresh-Vardy codes. *J. ACM*, 56(4):Art. 20, 34, 2009.

[HX07] B. Hassibi and W. Xu. Further results on performance analysis for compressive sensing analysis for compressive sensing using expander graphs. *Conf. Rec. 41st Asilomar Conf. Signals, Systems and Computers (ACSSC 2007)*, pages 621–625, 2007.

[JXHC09] Sina Jafarpour, Weiyu Xu, Babak Hassibi, and Robert Calderbank. Efficient and robust compressed sensing using optimized expander graphs. *IEEE Trans. Inform. Theory*, 55(9):4299–4308, 2009.

[PV05] F. Parvaresh and A. Vardy. Correcting errors beyond the guruswami- sudan radius in polynomial time. pages 285–294, 2005.

[Šid68] Zbyněk Šidák. On multivariate normal probabilities of rectangles: Their dependence on correlations. *Ann. Math. Statist.*, 39:1425–1434, 1968.

[SS96] Michael Sipser and Daniel A. Spielman. Expander codes. *IEEE Trans. Inform. Theory*, 42(6, part 1):1710–1722, 1996. Codes and complexity.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[vdGB09] Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392, 2009.