The Dantzig Selector

Hannes Gernandt

June 2016

1 Introduction

This talk is about a method proposed by Candès and Tao in [2] for the reconstruction of sparse x_0 from the equation

$$b = Ax_0 + e, \quad b \in \mathbb{R}^n, A \in \mathbb{R}^{n\times m}, \quad m >> n$$

and e is Gaussian white noise with iid entries $e_i \sim \mathcal{N}(0, \sigma^2)$. A natural approach for the solution would be

ε proportional to $\sqrt{n}\sigma$

$$(P_0^\varepsilon) \quad \min_x \|x\|_0, \quad \text{s.t.} \quad \|Ax - b\|_2 \leq \varepsilon$$

which can be made convex by

$$(P_1^\varepsilon) \quad \min_x \|x\|_1, \quad \text{s.t.} \quad \|Ax - b\|_2 \leq \varepsilon$$

This is just the "Basis pursuit denoising". Now the problem with BPDN is, that the constraint does not describe a white noise, because it merely restricts the norm of the residual. Therefore Candès and Tao replaced the constraint and formulated the Dantzig selector (DS) in the following way, depending on a parameter λ,

$$(P_{DS}^\lambda) \quad x_{DS}^\lambda = \arg\min_x \|x\|_1, \quad \text{s.t.} \quad \|A^T(b - Ax)\|_\infty \leq \lambda\sigma.$$

This new constraint may seem surprising at first, but the idea is simple: For $n \leq m$ the least squares estimator of the linear regression is given by the solution of the normal equation $A^TAx_0 = A^Tb$. Hence the constraint in DS can be viewed as the ℓ^∞-relaxation of the corresponding normal equation. Another idea to illustrate that the new constraint uses the fact that e is white noise more than BPDN does, is given below. For possible reconstruction x of x_0 with support different than x_0, say, not containing the i-th column, we would have that $a_i^T(b - Ax) = a_i^T(b - Ax_0) + x_0^{(i)}\|a_i\|^2 = a_i^Te + x_0^{(i)}$ is white noise which implies that a_i^Te should be small, because the average of e is expected to be zero. Therefore requiring $\|A^T(b-Ax)\|_\infty = \sup_i |a_i^T(b-Ax)|$ to be small means that all dominant columns in the creation of $b = Ax_0 + e$ are selected.

A first property of Dantzig selector is, that it can be reformulated as a LP problem by adding the variable $u \geq |x|$.

Lemma 1. *The Problem (P_{DS}^{λ}) is equivalent to the following linear program*

$$\min_{u,x} 1^T u \quad s.t. \qquad -u \leq x \leq u, \quad -\sigma\lambda 1 \leq A^T(b - Ax) \leq \sigma\lambda 1$$

Therefore one can reconstruct just by linear programming, e.g. by applying the simplex algorithm, and this is the reason, why it is named after Dantzig

Because of the additional information on the noise used in DS, the hope is that the performance is better.

Another reason might be, that Dantzig passed away during the creation of [2]

2 RIP and ROP

For the performance analysis of the Dantzig selector we first recall the "Restricted Isometry Property", denoted by RIP.

Definition 2. *For $A \in \mathbb{R}^{n\times m}$ with ℓ^2 normalized columns and $s \in \mathbb{N}$ consider $A_s \in \mathbb{R}^{n\times s}$ containing s columns of A. Define δ_s as the smallest quantity such that for all $c \in \mathbb{R}^n$*

$$(1 - \delta_s)\|c\|_2^2 \leq \|A_s c\|_2^2 \leq (1 + \delta_s)\|c\|_2^2$$

holds for any choice of s columns.

Next, we introduce the "Restricted Orthogonality Property" (ROP).

Definition 3. *For $A \in \mathbb{R}^{n\times m}$ with ℓ^2 normalized columns and $s_1, s_2 \in \mathbb{N}$ consider two disjoint subsets of columns of A denoted as A_{s_1} and A_{s_2}. Define θ_{s_1,s_2} as the smallest quantity such that for all $c_1 \in \mathbb{R}^{s_1}$ and all $c_2 \in \mathbb{R}^{s_2}$*

$$| < A_{s_1} c_1, A_{s_2} c_2 > | \leq \theta_{s_1,s_2}\|c_1\|_2\|c_2\|_2$$

holds for any choice of s_1, s_2 columns.

The measurements δ_s and θ_{s_1,s_2} are impossible to evaluate for general A, but easy to evaluate for stochastic matrices, see Section 3 in [1].

3 Performance Guarantee

Next we state the main result of this talk which can be found in [2, Theorem 1.1] or [3, Theorem 8.1].

Theorem 4. *Given (A, b, σ) and $\lambda = \sqrt{2(1+a)\log m}$. Suppose that $x_0 \in \mathbb{R}^m$ satisfies $\|x_0\|_0 = s$ with $\delta_{2s} + \theta_{2s,s} < 1$ and $b = Ax_0 + e$ where e is zero-mean Gaussian random noise with iid entries having variance σ^2.*

Then, the solution x_{DS}^{λ} of (P_{DS}^{λ}) obeys

$$\|x_{DS}^{\lambda} - x_0\|_2^2 \leq \frac{32(1+a)\log m}{(1 - \delta_{2s} - \theta_{2s,s})^2} s\sigma^2,$$

with probability exceeding $1 - (\sqrt{\pi(1+a)\log m} m^a)^{-1}$.

Here a is a constant appearing in the proof which must be large enough

Proof. We basically follow the proof from [3], but in a different order.

We consider two candidate solutions x^λ_{DS}, x_0 and estimate the norm of $d := x^\lambda_{DS} - x_0$.

Step 1: Since x_0 must not be feasible, we estimate the probability that it is feasible. The feasibility of x_0 is equivalent to

$$\|A^T(b - Ax_0)\|_\infty = \|A^T e\|_\infty \leq \lambda\sigma$$

Since the m entries of $A^T e$ are again $\mathcal{N}(0, \sigma)$ (Recall the affine transformation of normal distribution has $Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$) we have by the union bound

By assumption A is a normalized row vector

$$P(\sigma^{-1}\|A^T e\|_\infty \leq \lambda) \geq 1 - m\frac{2}{\sqrt{2\pi}}\int_\lambda^\infty e^{-x^2/2}dx.$$

Taking derivatives one sees $\int_u^\infty e^{-x^2/2}dx \leq e^{-u^2/2}/u$ and with $\lambda = \sqrt{2(1+a)\log m}$ we obtain

$$P(\sigma^{-1}\|A^T e\|_\infty \leq \lambda) \geq 1 - \frac{2m}{\sqrt{2\pi}\lambda}e^{-\lambda^2/2} = 1 - (\sqrt{\pi(1+a)\log m}m^a)^{-1}.$$

Hence for large a the probability gets arbitrarily close to 1. Hence in the majority of the cases x_0 is feasible.

Step 2: We show the estimate $\|d\|_2^2 \leq 2\|d\|_{2,q}^2 := 2\sum_{i\in q} d_i^2$ where q is the index set of size $2s$ which contains the support s of x_0 and additionally the s entries with the next largest absolute values. The inequality

sloppy notation: s is both, the set and its cardinality

Same methods as in the proof of BPDN

$$\|x_0\|_{1,s} = \|x_0\|_1 \geq \|\underbrace{x_0 + d}_{=x^\lambda_{DS}}\|_1 = \underbrace{\|x_0 + d\|_{1,s}}_{\geq \|x_0\|_{1,s} - \|d\|_{1,s}} + \underbrace{\|x_0 + d\|_{1,s^c}}_{=\|d\|_{1,s^c}}$$

Rearranging yields

$$\|d\|_{1,s} \geq \|d\|_{1,s^c}. \tag{1}$$

Next, we decompose d into a set of vectors each of length m. Each d_j contains s non-zeros, where d_0 has the support s, d_1 has a support $q \setminus s$, d_2 has support on the remaining s largest entries and so on. This construction implies that entry of d_j is smaller than arithmetic mean of entries of d_{j-1} which means

This construction appeared in a previous talk!

$$\sqrt{\sum_i (d_j^{(i)})^2} = \|d_j\|_2 \leq \sqrt{\sum_i \frac{(\sum_k |d_{j-1}^{(k)}|)^2}{s^2}} = \|d_{j-1}\|_1/\sqrt{s}$$

and this yields

$$\|d\|_{2,q^c}^2 = \sum_{j>1}\|d_j\|_2^2 \leq \frac{1}{s}\sum_{j>0}\|d_j\|_1^2 = \frac{1}{s}\|d\|_{1,s^c}^2.$$

This, (1) and the arithmetic-quadratic mean inequality $\|d\|_{1,s} \leq \sqrt{s}\|d\|_{2,s}$ imply

$$\|d\|_{2,q^c}^2 \leq \frac{1}{s}\|d\|_{1,s^c}^2 \leq \frac{1}{s}\|d\|_{1,s}^2 \leq \|d\|_{2,s}^2 \leq \|d\|_{2,q}^2.$$

Hence

$$\|d\|_2^2 = \|d\|_{2,q}^2 + \|d\|_{2,q^c}^2 \leq 2\|d\|_{2,q}^2.$$

Step 3: In the final step we estimate $\|d\|_{2,q}^2$ from above. The main idea is to use the orthogonal projection onto the space spanned by the columns of A_q given by $P_q = A_q(A_q^T A_q)^{-1}A_q^T$. For this one only need to check symmetry and $P_q = P_q^2$. Note that RIP implies $\|A_q v\|_2 \geq \sqrt{1-\delta_{2s}}\|v\|_2$ from $\delta_{2s} < 1$ we see that $A_q^T A_q$ is invertible.

This something new!

it has by RIP no kernel

Next, we prove the upper bound

$$\|P_q Ad\|_2^2 \leq \frac{2s(2\lambda\sigma)^2}{1-\delta_{2s}}. \tag{2}$$

For the proof one rewrites the RIP with the help of a weight matrix $Q = A_q^T A_q$ as

$$(1-\delta_{2s})w^T Q^{-1} w \leq w^T w = \|w\|_2^2, \quad w = A_q^T Ad$$

to see

$$(1-\delta_{2s})\|P_q Ad\|_2^2 \leq \|A_q^T Ad\|_2^2 \leq 2s\|A_q^T Ad\|_\infty^2. \tag{3}$$

the multiplication with A_q^T gives at most $2s$ values

Now, one needs the feasibility of x_0, because this implies

$$\|A^T(b - Ax_0)\|_\infty \leq \lambda\sigma, \quad \|A^T(b - Ax_{DS}^\lambda)\|_\infty \leq \lambda\sigma$$

and therefore by triangle inequality

$$\|A^T Ad\|_\infty = \|A^T(b - Ax_0) - A^T(b - Ax_{DS}^\lambda)\|_\infty \leq 2\lambda\sigma. \tag{4}$$

Plugging (4) in (3) proves (2).

Next we prove an lower bound for $\|P_q Ad\|_2$. Using the decomposition $\{d_j\}_j$ with $d = \sum_{j\geq 0} d_j$ from Step 2 we see that

$$P_q Ad = P_q Ad_0 + P_q Ad_1 + \sum_{j>1} P_q Ad_j = A_q d_q + \sum_{j>1} P_q Ad_j.$$

The reverse triangle inequality yields

$$\|P_q Ad\|_2 \geq \|A_q d_q\|_2 - \sum_{j>1}\|P_q Ad_j\|_2 \overset{RIP}{\geq} \sqrt{1-\delta_s}\|d\|_{2,q} - \sum_{j>1}\|P_q Ad_j\|_2$$

Here d_q means d restricted to entries with index in q

With the ROP we can estimate the second summand

$$\|P_q Ad_j\|_2^2 = < P_q Ad_j, Ad_j > \overset{ROP}{\leq} \theta_{2s,s}\|P_q Ad_j\|_2\|Ad_j\|_2 \overset{RIP}{\leq} \theta_{2s,s}\|P_q Ad_j\|_2\sqrt{1+\delta_s}\|d_j\|_2$$

With the inequalities $\sqrt{1+\delta_s} \leq (\sqrt{1-\delta_s})^{-1}$ and $\delta_s \leq \delta_{2s}$ this shows

$$\|P_q A d_j\|_2 \leq \frac{\theta_{2s,s}}{\sqrt{1-\delta_{2s}}} \|d_j\|_2.$$

As in Step 2 we can now estimate with $\|d\|_{1,s^c} \leq \|d\|_{1,s} \leq \sqrt{s}\|d\|_{2,s}$

$$\begin{aligned}\sum_{j>1} \|P_q A d_j\|_2 \leq \frac{\theta_{2s,s}}{\sqrt{1-\delta_{2s}}} \sum_{j>1} \|d_j\|_2 &\leq \frac{\theta_{2s,s}}{\sqrt{s}\sqrt{1-\delta_{2s}}} \sum_{j>0} \|d_j\|_1 \\ &= \frac{\theta_{2s,s}}{\sqrt{s}\sqrt{1-\delta_{2s}}} \|d\|_{1,s^c} \\ &\leq \frac{\theta_{2s,s}}{\sqrt{1-\delta_{2s}}} \|d\|_{2,q}\end{aligned}$$

This altogether yields

$$\|P_q A d\|_2 \geq \sqrt{1-\delta_{2s}}\|d\|_{2,q} - \frac{\theta_{2s,s}}{\sqrt{1-\delta_{2s}}}\|d\|_{2,q} = \frac{1-\delta_{2s}-\theta_{2s,s}}{\sqrt{1-\delta_{2s}}}\|d\|_{2,q} \qquad (5)$$

Combining the inequalities $\|d\|_2^2 \leq 2\|d\|_{2,q}^2$ and (2) and (5) we obtain

$$\|d\|_2^2 \leq 2\|d\|_{2,q}^2 \leq 2\frac{2s4\lambda^2\sigma^2}{(1-\delta_{2s}-\theta_{2s,s})^2} = \frac{16(1+a)\log m}{(1-\delta_{2s}-\theta_{2s,s})} s\sigma^2$$

which holds with the probability exceeding $1-(\sqrt{\pi(1+a)\log m}m^a)^{-1}$ since we used that x_0 is feasible. This proves the Theorem. □

4 Discussion

(a) The upper bound given in Theorem 8.1 from [3] is wrong, first of all it does not depend on the constant a which is strange for many reasons. The author is mixing up the two cases $a = 0$ and $a \geq 0$ of the correctly stated Theorem 1.1 in [2].

(b) For unitary A the problems (P_0^ε) and (P_1^ε) admit closed form solutions. But also the constraint $\|A^T(b-Ax)\|_\infty$ can be rewritten with $A^TA = I$ as $\|A^Tb - x\|_\infty \leq \lambda\sigma$. Therefore (P_{DS}^λ) can also be solved explicitly, see Section 8.2 in [3].

(c) The Dantzig selector for Random matrices can be found in [4].

(d) The Dantzig selector is surprisingly good in the following sense: Assume that we know from an oracle the support s of x_0. Then the estimation of x_0 is just the solution of the least squares problem

$$z^{opt} = \arg\min_z \|A_s z - b\|_2^2 = (A_s^T A_s)^{-1} A_s^T b.$$

Restricting the vector x_0 on its support s we obtain as an error estimate

Surprise: The Ortho projector from before!

$$\begin{aligned} E(\|z^{opt} - x_0^s\|_2^2) &= E(\|(A_s^T A_s)^{-1} A_s^T b - x_0^s\|_2^2) \\ &= E(\|(A_s^T A_s)^{-1} A_s^T (A_s x_0^2 + e) - x_0^s\|_2^2) \\ &= E(\|(A_s^T A_s)^{-1} A_s^T e\|_2^2) \\ &\stackrel{iid}{=} \mathrm{Tr}\left\{(\mathrm{A_s^T A_s})^{-1} \mathrm{A_s^T E(ee^T) A_s} (\mathrm{A_s^T A_s})^{-1}\right\} \\ &= \sigma^2 \mathrm{Tr}\left\{(\mathrm{A_s^T A_s})^{-1}\right\} \geq \frac{\mathrm{s}\sigma^2}{1+\delta_\mathrm{s}} \end{aligned}$$

with $E(ee^T) = \sigma^2 I$ and we used that the eigenvalues of $A_s^T A_s$ are due to RIP in the range $[1-\delta_s, 1+\delta_s]$. Therefore the Dantzig selector is close to the best possible error bound without knowing the support s of x which is surprising.

Tr is the trace

References

[1] E. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. on Information Theory*, 51:4203–4215, 2005.

[2] E. Candès and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313–2351, 2007.

[3] M. Elad. *Sparse and redundant representations*. Springer, 2010.

[4] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, 2013.