arXiv: arXiv:1601.05584

SUPPLEMENTARY MATERIAL TO REGULARIZATION AND THE SMALL-BALL METHOD I: SPARSE RECOVERY

By Guillaume Lecué[*,‡] and Shahar Mendelson[†,§]

CREST, CNRS, Université Paris Saclay.[‡]
Technion I.I.T. and Mathematical Sciences Institute, ANU.[§]
E-mail: guillaume.lecue@ensae.fr
E-mail: shahar@tx.technion.ac.il

An inspection of Theorem 3.2 reveals no mention of an isotropicity assumption. There is no choice of a Euclidean structure, and in fact, the statement itself is not even finite dimensional. All that isotropicity has been used for was to bound the "complexity function" $r(\cdot)$ and the "sparsity function" $\Delta(\cdot)$ in the three applications — the LASSO (in Theorem 1.4), SLOPE (in Theorem 1.6) and the trace norm regularization (in Theorem 1.7). We may apply Theorem 3.2 to situations that do not involve an isotropic vector and here we give an example of how this may be done.

To simplify our presentation we will only consider ℓ_1 and SLOPE regularization, which may both be written as

$$\Psi(t) = \sum_{j=1}^{d} \beta_j t_j^\sharp,$$

where $\beta_1 \geq \cdots \geq \beta_d > 0$ and $t_1^\sharp \geq \cdots \geq t_d^\sharp \geq 0$ is the nondecreasing rearrangement of $(|t_j|)$. As mentioned previously, the LASSO case is recovered for $\beta_1 = \cdots = \beta_d = 1$ and the SLOPE norm is obtained for $\beta_j = C\sqrt{\log(ed/j)}$ for some constant C. We also denote by B_Ψ (resp. S_Ψ) the unit ball (resp. sphere) associated with the Ψ-norm.

Let $\Sigma \in \mathbb{R}^{d\times d}$ be the covariance matrix of X and set $D = \{x \in \mathbb{R}^d : \left\|\Sigma^{1/2}x\right\|_2 \leq 1\}$ to be the corresponding ellipsoid. Naturally, if X is not isotropic than Σ is not the identity matrix.

*Supported by Labex ECODEC (ANR - 11-LABEX-0047) and by the "Chaire Economie et Gestion des Nouvelles Données", under the auspices of Institut Louis Bachelier, Havas-Media and Paris-Dauphine.

†Supported by the Israel Science Foundation.

In order to apply Theorem 3.2, we need to bound from above the expectation of the supremum of the Gaussian process indexed by $\rho B_\Psi \cap rD$:

$$\ell_*\left(\rho B_\Psi \cap rD\right) = \mathbb{E} \sup_{w \in \rho B_\Psi \cap rD} \left\langle \Sigma^{1/2} G, w \right\rangle \tag{0.1}$$

where G is a standard Gaussian vector in $\mathbb{R}^d$.

We also need to solve the "sparsity equation"—that is, find $\rho^* > 0$ for which $\Delta(\rho^*) \geq 4\rho^*/5$ where, for every $\rho > 0$,

$$\Delta(\rho) = \inf_{h \in \rho S_\Psi \cap rD} \sup_{g \in \Gamma_{t^*}(\rho)} \left\langle h, g \right\rangle$$

and $\Gamma_{t^*}(\rho)$ is the collection of all subgradients of Ψ of vectors in $t^* + (\rho/20)B_\Psi$.

We will show that the same results that have been obtained for the LASSO and SLOPE in Theorem 1.4 and Theorem 1.6 actually hold under the following assumption.

ASSUMPTION 0.1. *Let $j \in \{1,...,d\}$ and denote by $\Sigma^{1/2}_{j\bullet}$ the j-th row of $\Sigma^{1/2}$. Let $s \in \{1,\dots,d\}$ and set $\mathcal{B}_s = \sum_{j=1}^s \beta_j/\sqrt{j}$.*

1. *There exists $\sigma > 0$ such that for all $j \in \{1,\dots,d\}$, $\left\|\Sigma^{1/2}_{j\bullet}\right\|_2 \leq \sigma$.*
2. *For all $x \in (20\mathcal{B}_s S_\Psi) \cap D$, $2\left\|\Sigma^{1/2}x\right\|_2 \geq \sup_{|J|\leq s} \|x_J\|_2$.*

SLOPE.

We first control the Gaussian mean width in (0.1) when $\Psi(\cdot)$ is the SLOPE norm.

LEMMA 0.1. *Set $\beta_j = C\sqrt{\log(ed/j)}$ and let Σ be a $d\times d$ symmetric non-negative matrix for which $\max_j \left\|\Sigma^{1/2}_{j\bullet}\right\|_2 \leq \sigma$. If $D = \{x \in \mathbb{R}^d : \left\|\Sigma^{1/2}x\right\|_2 \leq 1\}$ then*

$$\mathbb{E} \sup_{w \in \rho B_\Psi \cap rD} \left\langle \Sigma^{1/2}G, w\right\rangle \leq \min\left\{\frac{\rho}{C}\left(\frac{3\sqrt{6}\sigma}{8} + r\sqrt{\frac{\pi}{2}}\right), r\sqrt{d}\right\}$$

PROOF. Note that

$$\mathbb{E} \sup_{w \in \rho B_\Psi \cap rD} \left\langle \Sigma^{1/2}G, w\right\rangle \leq r\mathbb{E}\sup_{w\in D}\left\langle G, \Sigma^{1/2}w\right\rangle = r\mathbb{E}\left\|G\right\|_2 \leq r\left(\mathbb{E}\left\|G\right\|_2^2\right)^{1/2} \leq r\sqrt{d}.$$

Next, let $H : \mathbb{R}^d \to \mathbb{R}$ be defined by $H(u) = \sup\left(\left\langle \Sigma^{1/2}u, w\right\rangle : w \in \rho B_\Psi \cap rD\right)$ and recall that G is the standard Gaussian vector in $\mathbb{R}^d$. It is straightforward to verify that

$$H(G) \le \sup_{w \in \rho B_\Psi} \left\langle \Sigma^{1/2}G, w\right\rangle \le \frac{\rho}{C} \max_{1 \le j \le d} \frac{\xi_j^\sharp}{\sqrt{\log(ed/j)}}$$

where we set $(\xi_j)_{j=1}^d = \Sigma^{1/2}G$ and $(\xi_j^\sharp)_{j=1}^d$ is the non-increasing rearrangement of $(|\xi_j|)_{j=1}^d$. Observe that for $u, v \in \mathbb{R}^d$,

$$|H(u) - H(v)| \le \sup_{w \in \rho B_\Psi \cap rD} |\left\langle \Sigma^{1/2}(u-v), w\right\rangle| \le r \sup_{w \in B_2^d} |\left\langle u-v, w\right\rangle| = r\left\|u-v\right\|_2,$$

implying that H is a Lipschitz function with constant r; thus, it follows from p. 21 in Chapter 1 of [1] that

$$\mathbb{E}H(G) \le \mathrm{Med}(H(G)) + r\sqrt{\frac{\pi}{2}}, \tag{0.2}$$

where $\mathrm{Med}(H(G))$ is the median of $H(G)$.

Hence, to obtain the claimed bound on $\mathbb{E}\sup_{w \in \rho B_\Psi} \left\langle \Sigma^{1/2}G, w\right\rangle$ it suffices to establish a suitable upper estimate on the median of $\max_{1 \le j \le d} \frac{\xi_j^\sharp}{\sqrt{\log(ed/j)}}$. With that in mind, let $\xi_1, \ldots, \xi_N$ be mean-zero Gaussian variables and assume that for every $j = 1, \ldots, d$,

$$\mathbb{E}\exp(\xi_j^2/L^2) \le e \tag{0.3}$$

for some $L > 0$. Note that in our case, $\xi_1, \ldots, \xi_N$ satisfying (0.3) for $L = 3\sigma/8$.

By Jensen's inequality,

$$\mathbb{E}\exp\left(\frac{1}{j}\sum_{k=1}^{j}\frac{(\xi_k^\sharp)^2}{L^2}\right) \le \frac{1}{j}\sum_{k=1}^{j}\mathbb{E}\exp\left(\frac{(\xi_k^\sharp)^2}{L^2}\right) \le \frac{1}{j}\sum_{k=1}^{d}\mathbb{E}\exp\left(\frac{\xi_k^2}{L^2}\right) \le \frac{ed}{j};$$

hence,

$$Pr\left(\frac{1}{j}\sum_{k=1}^{j}\frac{(\xi_k^\sharp)^2}{L^2} \ge 2\log\left(ed/j\right)\right) \le \exp\left(-\log\left(ed/j\right)\right) = \frac{j}{ed}. \tag{0.4}$$

Let $q \ge 0$ be the integer that satisfies $2^q \le d < 2^{q+1}$. It follows from (0.4) that with probability at least

$$1 - \sum_{\ell=0}^{q-1}\frac{2^\ell}{ed} = 1 - \frac{2^q - 1}{ed} > \frac{1}{2},$$

for every $\ell = 0, \cdots, q-1$,

$$(\xi^\sharp_{2^\ell})^2 \le \frac{1}{2^\ell}\sum_{j=1}^{2^\ell}(\xi^\sharp_j)^2 \le 2L^2\log(ed/2^\ell).$$

Moreover, for $2^\ell \le j < 2^{\ell+1}$, we have $\xi^\sharp_j \le \xi^\sharp_{2^\ell}$ and $\log(ed/2^\ell) \le 2\log(ed/j)$; also for $2^q \le j \le d$, we have $\xi^\sharp_j \le \xi^\sharp_{2^{q-1}}$ and $\log(ed/2^\ell) \le 3\log(ed/j)$. Therefore,

$$Pr\left(\max_{1\le j\le d}\frac{\xi^\sharp_j}{\sqrt{\log(ed/j)}} \le \sqrt{6}L\right) > \frac{1}{2}, \tag{0.5}$$

proving the requested bound on $\mathrm{Med}(H(G))$. □

Observe that up to constant σ, we actually recover the same result as in Equation (5.2); therefore, one may choose the same “complexity function” $r(\cdot)$ as in the proof of Theorem 1.6.

Let us turn to a lower bound on the “sparsity function”.

Lemma 0.2. *There exists an absolute constant $0 < c < 80$ for which the following holds. Let $s \in \{1,\ldots,d\}$ and set $\mathcal{B}_s = \sum_{j\le s}\beta_j/\sqrt{j}$. Assume that for every $x \in (80\mathcal{B}_s S_\psi)\cap D$ one has $\left\|\Sigma^{1/2}x\right\|_2 \ge (1/2)\sup_{|J|\le s}\|x_J\|_2$. Let $\rho > 0$ and assume further that there is a s-sparse vector in $t^* + (\rho/20)B_\Psi$. If $80\mathcal{B}_s \le \rho/r(\rho)$ then*

$$\Delta(\rho) = \inf_{h\in\rho S_\Psi\cap r(\rho)D}\ \sup_{g\in\Gamma_{t^*}(\rho)}\big\langle h, g\big\rangle \ge \frac{4\rho}{5}.$$

Proof. Let $h \in \rho S_\Psi \cap r(\rho)D$ and denote by $(h^\sharp_j)$ the non-increasing rearrangement of $(|h_j|)$. It follows from the proof of Lemma 4.2 that

$$\sup_{g\in\Gamma_{t^*}(\rho)}\big\langle h, g\big\rangle \ge \frac{17\rho}{20} - 2\sum_{j\le s}\beta_j h^\sharp_j.$$

Let $h^{\sharp,s}$ be the s-sparse vector with coordinates given by $h^\sharp_j$ for $1 \le j \le s$ and 0 otherwise. We have

$$\frac{h}{\rho} \in S_\Psi \cap \left(\frac{r(\rho)}{\rho}\right)D \subset S_\Psi \cap \left(\frac{1}{80\mathcal{B}_s}\right)D$$

implying that $2\left\|\Sigma^{1/2}h\right\|_2 \geq \left\|h^{\sharp,s}\right\|_2$. Furthermore, since $h_j^\sharp \leq \left\|h^\sharp\right\|_2/\sqrt{j}$ for every $1 \leq j \leq s$, we have

$$\sum_{j\leq s}\beta_j h_j^\sharp \leq \left\|h^{\sharp,s}\right\|_2 \mathcal{B}_s \leq 2\mathcal{B}_s\left\|\Sigma^{1/2}h\right\|_2 \leq 2\mathcal{B}_s r(\rho).$$

Hence, if $\rho \geq 80r(\rho)\mathcal{B}_s$ then $\Delta(\rho) \geq 4\rho/5$. $\square$

We thus recover the same condition as in Lemma 4.3, implying that Theorem 1.6 actually holds under the weaker Assumption 0.1: let X be an L-subgaussian random vector whose covariance matrix satisfies Assumption 0.1. The SLOPE estimator with regularization parameter $\lambda \sim \|\xi\|_{L_q}/\sqrt{N}$ satisfies, with probability at least $1-\delta-\exp(-c_0NL^8)$,

$$\Psi(\hat{t}-t^*) \leq c_3\|\xi\|_{L_q}\frac{s}{\sqrt{N}}\log\left(\frac{ed}{s}\right) \quad \text{and} \quad \left\|\Sigma^{1/2}(\hat{t}-t^*)\right\|_2^2 \leq c_3\|\xi\|_{L_q}^2\frac{s}{N}\log\left(\frac{ed}{s}\right)$$

when $N \geq c_4 s\log(ed/s)$ and when there is a s-sparse vector close enough to t^*.

The LASSO.

Here, for every $1 \leq j \leq d$, $\beta_j = 1$; $B_\Psi = B_1^d$; and $\mathcal{B}_s = \sum_{j=1}^s 1/\sqrt{j} \leq 2\sqrt{s}$.

Lemma 0.3. *Let Σ be a $d\times d$ symmetric nonnegative matrix for which $\max_j\left\|\Sigma_{j\bullet}^{1/2}\right\|_2 \leq \sigma$. If $D = \{x\in\mathbb{R}^d : \left\|\Sigma^{1/2}x\right\|_2 \leq 1\}$ then every $\rho > 0$ and $r > 0$,*

$$\mathbb{E}\sup_{w\in\rho B_1^d\cap rD}\left\langle\Sigma^{1/2}G, w\right\rangle \leq \min\left\{r\sqrt{d}, \rho\sigma\sqrt{\log(ed)}\right\}.$$

Proof. Note that

$$\mathbb{E}\sup_{w\in\rho B_1^d\cap rD}\left\langle\Sigma^{1/2}G, w\right\rangle \leq r\mathbb{E}\sup_{w\in D}\left\langle G, \Sigma^{1/2}w\right\rangle = r\mathbb{E}\left\|G\right\|_2 \leq r\left(\mathbb{E}\left\|G\right\|_2^2\right)^{1/2} \leq r\sqrt{d}.$$

Next, set $(\xi_j)_{j=1}^d = \Sigma^{1/2}G$ and let $(\xi_j^\sharp)_{j=1}^d$ be the non-increasing rearrangement of $(|\xi_j|)$. Therefore, $\xi_1,\ldots,\xi_d$ are mean-zero Gaussian variables and satisfy $\mathbb{E}\exp(\xi_j^2/L^2) \leq e$ for $L = 3\sigma/8$. It is evident that

$$\mathbb{E}\left(\frac{(\xi_1^\sharp)^2}{L^2}\right) \leq \log\left(\mathbb{E}\exp\left(\frac{(\xi_1^\sharp)^2}{L^2}\right)\right) \leq \log\left(\sum_{j=1}^d\mathbb{E}\exp\left(\frac{\xi_j^2}{L^2}\right)\right) \leq \log\left(ed\right),$$

and therefore,

$$\mathbb{E}\sup_{w\in\rho B_1^d\cap rD}\left\langle\Sigma^{1/2}G, w\right\rangle \leq \mathbb{E}\sup_{w\in\rho B_1^d}\left\langle\Sigma^{1/2}G, w\right\rangle \leq \rho\mathbb{E}\left((\xi_1^\sharp)^2\right)^{1/2} \leq \frac{3\rho\sigma L}{8}\sqrt{\log\left(ed\right)}.$$

$\square$

Lemma 0.3 leads to a slightly different result than in the isotropic case (Lemma 5.3), and as a consequence, $r(\cdot)$ has to be slightly modified. A straightforward computation shows that

$$r_M^2(\rho) \lesssim_{L,q,\delta} \min\left(\frac{\|\xi\|_{L_q} d}{N}, \rho\sigma \|\xi\|_{L_q} \sqrt{\frac{\log(ed)}{N}}\right)$$

and

$$r_Q^2(\rho) \lesssim_L \begin{cases} 0 & \text{if } N \gtrsim_L d \\ \frac{\rho^2\sigma^2}{N} \log\left(\frac{c(L)d}{N}\right) & \text{otherwise.} \end{cases}$$

and still $r(\rho) = \max\{r_M(\rho), r_Q(\rho)\}$.

Finally, let us prove the sparsity condition.

LEMMA 0.4. *There exists an absolute constant $0 < c < 80$ for which the following holds. Let $s \in \{1, \dots, d\}$ and set $\mathcal{B}_s = \sum_{j \le s} 1/\sqrt{j}$. Assume that for every $x \in \left(20\mathcal{B}_s S(\ell_1^d)\right) \cap D$ one has $\left\|\Sigma^{1/2}x\right\|_2 \ge (1/2) \sup_{|J| \le s} \|x_J\|_2$. Let $\rho > 0$ and assume further that there is a s-sparse vector in $t^* + (\rho/20)B_1^d$. If $20\mathcal{B}_s \le \rho/r(\rho)$ then*

$$\Delta(\rho) = \inf_{h \in \rho S(\ell_1^d) \cap r(\rho)D} \sup_{g \in \Gamma_{t^*}(\rho)} \langle h, g \rangle \ge \frac{4\rho}{5}.$$

PROOF. Let $h \in \rho S(\ell_1^d) \cap r(\rho)D$ and denote by $(h_j^\sharp)$ the non-increasing rearrangement of $(|h_j|)$. It follows from the proof of Lemma 4.2 that

$$\sup_{g \in \Gamma_{t^*}(\rho)} \langle h, g \rangle \ge \rho - 2\sum_{j \le s} h_j^\sharp.$$

Let $h^{\sharp,s}$ be the s-sparse vector with coordinates given by $h_j^\sharp$ for $1 \le j \le s$ and 0 otherwise. Observe that

$$\frac{h}{\rho} \in S(\ell_1^d) \cap \left(\frac{r(\rho)}{\rho}\right) D \subset S(\ell_1^d) \cap \left(\frac{1}{40\mathcal{B}_s}\right) D,$$

and therefore $2\left\|\Sigma^{1/2}h\right\|_2 \ge \left\|h^{\sharp,s}\right\|_2$.

It follows that

$$\sum_{j \le s} h_j^\sharp \le \sqrt{s}\left\|h^\sharp\right\|_2 \le \mathcal{B}_s \left\|h^\sharp\right\|_2 \le 2\mathcal{B}_s \left\|\Sigma^{1/2}h\right\|_2 \le 2\mathcal{B}_s r(\rho),$$

and in particular, if $\rho \ge 20r(\rho)\mathcal{B}_s$ then $\Delta(\rho) \ge 4\rho/5$. □

Using the estimate on $r(\cdot)$ and Lemma 0.4, it is evident that when $N \gtrsim_{L,q,\delta} s\sigma^2 \log(ed)$, one has $\Delta(\rho^*) \geq 4\rho^*/5$ for

$$\rho^* \sim_{L,q,\delta} \|\xi\|_{L_q} s\sqrt{\frac{\log(ed)}{N}}$$

and if there is a s-sparse vector in $t^* + (\rho^*/20)B_1^d$.

Finally, one may choose the regularization parameter by setting

$$\lambda \sim \frac{r^2(\rho^*)}{\rho^*} \sim_{L,q,\delta} \|\xi\|_{L_q} s\sqrt{\frac{\log(ed)}{N}}.$$

It follows that if X is an L-subgaussian random vector that satisfies Assumption 0.1 then with probability larger than $1-\delta-2\exp(-c_0 N L^8)$,

$$\left\|\hat{t}-t^*\right\|_1 \leq \rho^* = c_1(\delta)\,\|\xi\|_{L_q} s\sqrt{\frac{\log(ed)}{N}}$$

and

$$\left\|\Sigma^{1/2}(\hat{t}-t^*)\right\|_2 \leq r(\rho^*) = c_2(L,\delta)\,\|\xi\|_{L_q}\sqrt{\frac{s\log(ed)}{N}}.$$

References.

[1] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI, 2001.

ENSAE, 3, avenue Pierre Larousse,
92245 MALAKOFF. France.

Department of Mathematics,
Technion, I.I.T., Haifa, Israel