ECE 830
Note on Lasso

This is a short note on the compressed sensing analysis in the tutorial paper "Estimation Theory in High Dimensions" by Roman Vershynin.

Let $\boldsymbol{x}^\star$ be a s-sparse vector and suppose that we observe

$$\boldsymbol{y} \;=\; \boldsymbol{A}\boldsymbol{x}^\star \;+\; \boldsymbol{\nu}\;,$$

where $\boldsymbol{A}$ is a known $m \times n$ matrix with entries $A_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$ and $\boldsymbol{\nu}$ is an unknown error vector satisfying $\frac{1}{m}\sum_{i=1}^m |\nu_i| \leq \varepsilon$, for some known $\varepsilon > 0$. Let $\widehat{\boldsymbol{x}}$ be the solution to the lasso optimization

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{x}\|_1 \;\text{ subject to }\; \frac{1}{m}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_1 \leq \varepsilon\;.$$

We will show that $\widehat{\boldsymbol{x}} \approx \boldsymbol{x}^\star$. More precisely, we will show that

$$\mathbb{E}[\|\widehat{\boldsymbol{x}}-\boldsymbol{x}^\star\|_2] \;\leq\; C\,\|\boldsymbol{x}^\star\|_2\sqrt{\frac{s\log n}{m}}\;,$$

where $C > 0$ is a small constant.

Why would such a bound be true? There are two key facts/constraints at our disposal:

1. $\boldsymbol{x}^\star$ lies in a known set $\{\boldsymbol{x} : \boldsymbol{x} \text{ has } s \text{ nonzeros }\}$

2. $\boldsymbol{x}^\star$ lies (approximately) in a known random affine subspace $\{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} \approx \boldsymbol{y}\}$

So our first thought should be to find a vector that lies in the intersection of these two sets. In fact, that's essentially all we need to do, because the intersection turns out to be a very small set (due to the randomness in $\boldsymbol{A}$ and the sparsity assumption). The lasso optimization is a convex relaxation of this search problem.

First of all, let's focus discuss the convex relaxation. Consider the set of all vectors $\boldsymbol{x}$ with s nonzeros and $\|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}^\star\|_2$. This is a non-convex set. However, it is easy to see that every such vector must satisfy $\|\boldsymbol{x}\|_1 \leq \sqrt{s}\|\boldsymbol{x}^\star\|_2$. So we will optimize over the *convex* set

$$\mathcal{C} := \{\boldsymbol{x} : \|\boldsymbol{x}\|_1 \leq \sqrt{s}\|\boldsymbol{x}^\star\|_2\}\;.$$

Now suppose we find a vector $\widehat{\boldsymbol{x}}$ that satisfies the following two conditions:

1. $\widehat{\boldsymbol{x}} \in \mathcal{C}$

2. $\frac{1}{m}\|\boldsymbol{y}-\boldsymbol{A}\widehat{\boldsymbol{x}}\|_1 \leq \varepsilon$

Note that $\widehat{\boldsymbol{x}}-\boldsymbol{x}^\star \in \mathcal{C}-\mathcal{C}$. Let $\mathcal{T} = \mathcal{C}-\mathcal{C}$. Let $\boldsymbol{a}_i^T$ denote the ith row of $\boldsymbol{A}$ and observe that for any $\boldsymbol{u}$

$$\mathbb{E}[|\boldsymbol{a}_i^T\boldsymbol{u}|] = \sqrt{\frac{2}{\pi}}\,\|\boldsymbol{u}\|_2\;.$$

This follows by the isotropy of the multivariate Gaussian distribution. This suggests that by averaging such measurements we have

$$\frac{1}{m}\sum_{i=1}^m |\boldsymbol{a}_i^T\boldsymbol{u}| \;\approx\; \sqrt{\frac{2}{\pi}}\,\|\boldsymbol{u}\|_2\;.$$

In particular, we wish to show that this approximate holds for $\boldsymbol{u} = \widehat{\boldsymbol{x}} - \boldsymbol{x}^\star$. This is a bit more tricky since $\widehat{\boldsymbol{x}}$ depends on $\boldsymbol{A}$.

Using standard concentration inequalities it can be show that

$$\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathcal{T}}\left|\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{u}| - \sqrt{\frac{2}{\pi}}\,\|\boldsymbol{u}\|_2\right|\right] \;\leq\; \frac{4}{\sqrt{m}}\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathcal{T}}|\boldsymbol{g}^T\boldsymbol{u}|\right] \;,$$

where $\boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$. Since this holds for all $\boldsymbol{u}\in\mathcal{T}$ it must hold for $\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star$, even though $\widehat{\boldsymbol{x}}$ depends on $\boldsymbol{A}$. Re-arranging this yields

$$\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathcal{T}}\|\boldsymbol{u}\|_2\right] \;\leq\; \sqrt{\frac{8\pi}{m}}\,\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathcal{T}}|\boldsymbol{g}^T\boldsymbol{u}|\right] \;+\; \sup_{\boldsymbol{u}\in\mathcal{T}}\sqrt{\frac{\pi}{2}}\,\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{u}| \;. \tag{1}$$

Now recall that we are interested only in $\widehat{\boldsymbol{x}}$ that satisfy $\frac{1}{m}\|\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{x}}\|_1 \leq \varepsilon$. Also recall that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^\star + \boldsymbol{\nu}$. Therefore, we have

$$\frac{1}{m}\|\boldsymbol{A}(\boldsymbol{x}^\star - \widehat{\boldsymbol{x}})\|_1 \;=\; \frac{1}{m}\|-\boldsymbol{\nu} + \boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{x}}\|_1 \;\leq\; \frac{1}{m}\|\boldsymbol{\nu}\|_1 + \frac{1}{m}\|\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{x}}\|_1 \;\leq\; 2\varepsilon \;.$$

So we can restrict our attention to the subset of $\boldsymbol{u}\in\mathcal{T}$ that satisfy the additional constraint $\frac{1}{m}\|\boldsymbol{A}\boldsymbol{u}\|_1 \leq 2\varepsilon$. Let $\mathcal{T}'$ denote this subset. The bound in (1) remains valid of we restrict the $\boldsymbol{u}$ in the sup to $\mathcal{T}'$. So we can write

$$\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathcal{T}'}\|\boldsymbol{u}\|_2\right] \;\leq\; \sqrt{\frac{8\pi}{m}}\,\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathcal{T}}|\boldsymbol{g}^T\boldsymbol{u}|\right] \;+\; \sup_{\boldsymbol{u}\in\mathcal{T}'}\sqrt{\frac{\pi}{2}}\,\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{u}| \;.$$

Note that we have kept $\mathcal{T}$ in the first supremum on the right hand side. The second term on the right is bounded by recalling that we just showed that $\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{u}| \leq 2\varepsilon$ for all $\boldsymbol{u}\in\mathcal{T}'$. So we now have

$$\mathbb{E}[\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star\|_2] \;\leq\; \mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathcal{T}'}\|\boldsymbol{u}\|_2\right] \;\leq\; \sqrt{\frac{8\pi}{m}}\,\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathcal{T}}|\boldsymbol{g}^T\boldsymbol{u}|\right] \;+\; \sqrt{2\pi}\varepsilon \;.$$

Finally, we must bound $\mathbb{E}\left[\sup_{\boldsymbol{u}\in\mathcal{T}}|\boldsymbol{g}^T\boldsymbol{u}|\right]$. Recall that $\boldsymbol{g}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$. In general, the quantity $|\boldsymbol{g}^T\boldsymbol{u}|$ is maximized by concentrating $\boldsymbol{u}$ on the index where $\boldsymbol{g}$ achieves its maximum absolute value. Recall that every $\boldsymbol{x}\in\mathcal{C}$ satisfies $\|\boldsymbol{x}\|_1 \leq \sqrt{s}\|\boldsymbol{x}^\star\|_2$. Therefore, $\|\boldsymbol{u}\|_1 \leq \|\widehat{\boldsymbol{x}}\|_1 + \|\boldsymbol{x}^\star\|_1 \leq 2\sqrt{s}\|\boldsymbol{x}^\star\|_2$. This implies that

$$\sup_{\boldsymbol{u}\in\mathcal{T}}|\boldsymbol{g}^T\boldsymbol{u}| \;\leq\; 2\sqrt{s}\|\boldsymbol{x}^\star\|_2\max_i|g_i| \;.$$

Next we need to bound $\mathbb{E}[\max_i|g_i|]$. Recall from our discussions of multiple testing (and the Gaussian tail bound) that the maximum of n independent $\mathcal{N}(0,1)$ variables is bounded by about $\sqrt{2\log n}$. Using this, we can easily obtain the bound

$$\mathbb{E}[\max_i|g_i|] \;\leq\; \sqrt{2\log n} + 1 \;.$$

Putting everything together, we have shown that for $n > 1$ any vector $\widehat{\boldsymbol{x}}$ satisfying

$$\|\widehat{\boldsymbol{x}}\|_1 \;\leq\; \sqrt{s}\|\boldsymbol{x}^\star\|_2 \;\text{ and }\; \frac{1}{m}\|\boldsymbol{y} - \boldsymbol{A}\widehat{\boldsymbol{x}}\|_1 \;\leq\; \varepsilon \tag{2}$$

satisfies the bound

$$\mathbb{E}[\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star\|_2] \;\leq\; 8\sqrt{\pi}\sqrt{\frac{s\log n}{m}} \;+\; \sqrt{2\pi}\varepsilon\ .$$

It is also easy to show that any solution $\widehat{\boldsymbol{x}}$ to the optimization

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{x}\|_1 \text{ subject to } \frac{1}{m}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_1 \leq \varepsilon$$

also satisfies the bound above. This is verified by showing that the solution to the optimization will meet the conditions in (2) above. The important thing to notice is that the error is small as long as $m = O(s\log n)$.