Least squares with sparsity and generalizations

Notes by Parker Knight

May 18, 2022

1 Preliminaries

1.1 Basic tail bounds

The development of a nonasymptotic theory involves understanding the extreme behavior of random variables; more specifically, we seek to study how random variables fluctuate around their mean. Our primary tool to do so is known as *Markov's inequality*, which is stated and proven below.

Theorem 1 (Markov's inequality)**.** *Let X be a random variable and let $g(.)$ be a nonnegative function. Then for $t \geq 0$*

$$\mathbb{P}\left\{g(X) \geq t\right\} \leq \frac{\mathbb{E}\left[g(X)\right]}{t}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[g(X)\right] &= \int_{\mathbb{R}} g(x)f(x)dx \\
&\geq \int_{\{x:g(x)\geq t\}} g(x)f(x)dx \quad \text{(using nonnegativity of } g\text{)} \\
&\geq t\int_{\{x:g(x)\geq t} f(x)dx \\
&= t\mathbb{P}\left\{g(X) \geq t\right\}
\end{aligned}$$

Rearranging terms gives the result. □

Through a careful choice of function $g(.)$, we can control the tails of X rather elegantly.

Corollary 1 (Chebyshev's inequality)**.** *Let X be a random variable with a finite second moment. Then for $t \geq 0$*

$$\mathbb{P}\left\{|X - \mathbb{E}\left[X\right]| \geq t\right\} \leq \frac{var(X)}{t^2}$$

Proof. By direct calculation:

$$\begin{aligned}\mathbb{P}\left\{|X - \mathbb{E}[X]| \geq t\right\} &= \mathbb{P}\left\{(X - \mathbb{E}[X])^2 \geq t^2\right\} \\ &\leq \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}{t^2} \quad \text{(by Markov)} \\ &= \frac{\text{var}(X)}{t^2}\end{aligned}$$

□

Chebyshev's inequality requires only the existence of a second moment, but in many cases, can be quite loose. We can obtain tighter bounds under more stringent conditions on X.

Corollary 2 (Chernoff bound)**.** *Let X be a random variable with a moment-generating function that exists at all $\lambda \in \mathbb{R}$. Then for $t \geq 0$*

$$\mathbb{P}\left\{X \geq t\right\} \leq \inf_{\lambda \in \mathbb{R}} e^{-t\lambda} \mathbb{E}\left[e^{\lambda X}\right]$$

Proof. Apply Markov with the function $g(x) = e^{\lambda x}$. □

The Chernoff bound allows us to control the tails of X with its moment generating function. Often, this can give us much tighter bounds than those obtained by Chebyshev.

For example, let $X \sim N(0, \sigma^2)$. A simple calculation yields $\mathbb{E}\left[e^{\lambda X}\right] = e^{\sigma^2 \lambda^2 / 2}$ for all $\lambda \in \mathbb{R}$. The Chernoff bound yields

$$\mathbb{P}\left\{X \geq t\right\} \leq \inf_{\lambda \in \mathbb{R}} e^{\sigma^2 \lambda^2 / 2 - \lambda t}$$

Some calculus reveals that this infimum is attained at $\lambda = t/\sigma^2$, yielding an upper bound of

$$\mathbb{P}\left\{X \geq t\right\} \leq \exp\left[-\frac{t^2}{2\sigma^2}\right]$$

Importantly, the form of the normal MGF leads to very fast decay in the tail. It is natural to wonder whether other random variables exhibit similar rates. This motivates the following definition.

Definition 1 (sub-Gaussian random variable)**.** *Let X be a mean-zero random variable taking values in $\mathbb{R}$. We say X is sub-Gaussian with parameter σ^2 if for all $\lambda \in \mathbb{R}$:*

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\lambda^2 \sigma^2 / 2}$$

We write $X \in \mathsf{subG}(\sigma^2)$.

Clearly, any sub-Gaussian random variable with achieve the same tail rate as the corresponding normal.

Definition 2 (sub-Gaussian random vector)**.** *Let X be a mean zero random vector taking values in $\mathbb{R}^d$. We say X is σ^2 sub-Gaussian if $X^T u \in \mathsf{subG}(\sigma^2)$ for any unit vector $u \in \mathbb{R}^d$.*

Sub-gaussians have many useful properties. We will be particularly interested in the maximum (or ∞-norm) of sub-Gaussian vectors. The key lemma is stated below (proof follows by a union bound).

Lemma 1. *Let $X_1, ..., X_n \in \mathsf{subG}(\sigma^2)$. Then*

$$\max_{i=1...n} |X_i| \lesssim \sigma\sqrt{\log n}$$

with high probability.

2 The LASSO

Suppose a linear model

$$Y = X\theta^* + \varepsilon$$

with $Y \in \mathbb{R}^n, \mathbb{X} \in R^{n\times d}, \theta^* \in \mathbb{R}^d$ and $\varepsilon \in \mathbb{R}^n$. Furthermore, suppose that each element of ε are independent random variables, with a distribution to be specified later. Our goal is to estimate the vector θ^*. A naive approach (which uses no further assumptions) is least squares, which will achieve a convergence rate of $\frac{d}{n}$. In the high-dimensional setting when $\frac{d}{n} \to \gamma > 0$, this approach is not good enough. Even as the sample size tends to infinity, the error will stay bounded away from zero. We get around this by placing further assumptions on θ^*.

In particular, assume that θ^* is s-sparse, supported on an index set S. A natural estimator is the LASSO, which induces sparsity in the estimate by way of an ℓ_1 penalty:

$$\hat{\theta} \in \underset{\theta\in\mathbb{R}^d}{\arg\min} \left\{ \frac{1}{2n} \|Y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right\}$$

We now aim to study the convergence rate of the LASSO estimator, given our knowledge of the support of θ^*. First, we need a couple of definitions. For any index set $S \subset [d]$ and $\alpha \geq 1$, define

$$\mathcal{C}_\alpha(S) = \left\{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_S\|_1\right\}$$

The set defined abvoe constains the set of 'good' vectors with respect to S; i.e., those for which the ℓ_1 norm on S is greater than the ℓ_1 norm off of S. Clearly, every vector supported exactly on S lies in $\mathcal{C}_\alpha(S)$. Importantly, if $\Delta \in \mathcal{C}_\alpha(S)$, then

$$\begin{aligned}\|\Delta\|_1 &= \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \\ &\leq (\alpha+1)\|\Delta_S\|_1 \\ &\leq (\alpha+1)\sqrt{s}\|\Delta\|_2\end{aligned}$$

where $s = |S|$. The first equality in the computation is a result of the decomposability of the ℓ_1 norm over an index set S, and the final inequality follows by Cauchy-Schwarz. We will return to this notion of decomposability more broadly in the next section; for now, it will serve as a useful tool in the computations that follow. The result above tells us that when $s << d$, inclusion in $\mathcal{C}_\alpha(S)$ grants much tighter control over the ℓ_1 norm than the usual bound $\|x\|_1 \leq \sqrt{d}\|x\|_2$.

We say that the matrix X satisfies the *restricted eigenvalue* condition over S with paramters (k, α) if

$$\frac{1}{n}\|X\Delta\|_2^2 \geq k\|\Delta\|_2^2 \quad \forall \Delta \in \mathcal{C}_\alpha(S)$$

This condition is closely related to the curvature of the loss function. Observe that the Hessian of the loss function is $\frac{1}{2n}X^TX$: we want this matrix to be positive definite so that our optimization problem is strictly convex. However, this is impossible in the high-dimensional setting, so we only require positive eigenvalues in the 'good' directions, defined by the set $\mathcal{C}_\alpha(S)$.

Theorem 2 (LASSO Convergence)**.** *Assume that θ^* has support $S \subset [d]$ with size $|S| = s$, and suppose that X satisfies the restricted eigenvalue condition over S with parameters $(k, 3)$. Then any LASSO solution $\hat{\theta}$ computed with tuning paramter $\lambda \geq 2\left\|\frac{1}{n}X^T\varepsilon\right\|_\infty$ satisfies*

$$\left\|\hat{\theta} - \theta^*\right\|_2 \lesssim \frac{\sqrt{s}}{k}\lambda$$

To prove Theorem 2, we state and prove two Lemmas that will yield the result.

Lemma 2 (Deviation inequalities)**.** *For any $\theta^* \in \mathbb{R}^d$ supported on S and any $\Delta \in \mathbb{R}^d$, we have*

$$\|\theta^* + \Delta\|_1 - \|\theta^*\|_1 \geq \|\Delta_{S^c}\|_1 - \|\Delta_S\|_1$$

Furthermore, if $\lambda \geq 2\left\|\frac{1}{n}X^T\varepsilon\right\|_\infty$, we have

$$\frac{1}{2n}\|Y - X(\theta^* + \Delta)\|_2^2 - \frac{1}{2n}\|Y - X\theta^*\|_2^2 \geq -\frac{\lambda}{2}\left(\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1\right)$$

Proof. To prove the first claim, observe

$$\begin{aligned}
\|\theta^* + \Delta\|_1 &= \|\theta^*_S + \theta^*_{S^c} + \Delta_S + \Delta_{S^c}\| \\
&= \|\theta^*_S + \Delta_S + \Delta_{S^c}\| \quad \text{(since } \theta^* \text{ is supported on } S\text{)} \\
&\geq \|\theta^*_S + \Delta_{S^c}\|_1 - \|\Delta_S\|_1 \quad \text{(triangle inequality)} \\
&= \|\theta^*_S\|_1 + \|\Delta_{S^c}\|_1 - \|\Delta_S\|_1 \quad \text{(decomposability)} \\
&= \|\theta^*\|_1 + \|\Delta_{S^c}\|_1 - \|\Delta_S\|_1 \quad \text{(support of } \theta^*\text{)}
\end{aligned}$$

Rearranging terms gives the result. To prove the second claim, we first introduce the notation

$$\mathcal{L}(\theta) = \frac{1}{2n} \|Y - X\theta\|_2^2$$

Observe that $\mathcal{L}(.)$ is a convex function. Using this, we have

$$\begin{aligned}
\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) &\geq \nabla\mathcal{L}(\theta^*)^T\Delta \quad \text{(first order condition for convex functions)} \\
&\geq -|\nabla\mathcal{L}(\theta^*)^T\Delta| \\
&\geq -\|\nabla\mathcal{L}(\theta^*)\|_\infty \|\Delta\|_1 \quad \text{(Holder)} \\
&= -\left\|\frac{1}{n}X^T\varepsilon\right\|_\infty (\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1) \\
&\geq -\frac{\lambda}{2}(\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1)
\end{aligned}$$

as desired. □

Lemma 3 (Control of error vector)**.** *Let $\hat{\theta}$ be the solution to the LASSO program with tuning parameter $\lambda \geq 2\left\|\frac{1}{n}X^T\varepsilon\right\|_\infty$. Then $\hat{\Delta} = \hat{\theta} - \theta^* \in \mathcal{C}_3(S)$.*

Proof. Optimality of $\hat{\theta}$ tells us

$$\frac{1}{2n}\left\|Y - X\hat{\theta}\right\|_2^2 + \lambda\left\|\hat{\theta}\right\|_1 \leq \frac{1}{2n}\|Y - X\theta^*\|_2^2 + \lambda\|\theta^*\|_1$$

Using $\hat{\theta} = \hat{\Delta} + \theta^*$ and rearranging terms gives

$$0 \geq \frac{1}{2n}\left\|Y - X(\theta^* + \hat{\Delta})\right\|_2^2 - \frac{1}{2n}\|Y - X\theta^*\|_2^2 + \lambda(\left\|\hat{\Delta} + \theta^*\right\|_1 - \|\theta^*\|_1)$$

Applying Lemma 2 to the terms in the above inequality gives

$$0 \geq \lambda\left(\left\|\hat{\Delta}_{S^c}\right\|_1 - \left\|\hat{\Delta}_S\right\|_1 - \frac{1}{2}\left\|\hat{\Delta}_S\right\|_1 - \frac{1}{2}\left\|\hat{\Delta}_{S^c}\right\|_1\right)$$

Rearranging gives

$$3\left\|\hat{\Delta}_S\right\|_1 \geq \left\|\hat{\Delta}_{S^c}\right\|$$

which shows $\hat{\Delta} \in \mathcal{C}_3(S)$.

□

Now we can prove Theorem 2.

Proof of the Theorem. □

3 Regularized M-estimators

Here we present a general theory for estimation under a family of regularized M-estimators. This content is guided by Negahban et al. 2012.

Let $Z_1, ..., Z_n \overset{\text{iid}}{\sim} \mathbb{P}$ be random variables taking values in a set $\mathcal{Z}$. Let $\theta \in \Theta$ be an unknown parameter of the marginal distribution $\mathbb{P}$, where Θ is a vector space. Let $\mathcal{L} : \Theta \times \mathcal{Z}^n \to \mathbb{R}$ be a convex and differentiable loss function[1], and define

$$\theta^* := \arg\min_{\theta \in \Theta} \mathbb{E}\left[\mathcal{L}(\theta)\right]$$

where expectation is taken with respect to $\mathbb{P}$. We aim to estimate θ^* by the solution to the convex program

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} \left\{\mathcal{L}(\theta) + \lambda \Phi(\theta)\right\}$$

where $\lambda > 0$ is a tuning parameter chosen by the analyst, and Φ is a regularizer from Θ to $\mathbb{R}_+$ (typically a norm). Let $\|.\|$ be a norm on Θ induced by inner product $\langle ., . \rangle$. Our aim is to provide bounds on $\left\|\hat{\theta} - \theta^*\right\|$ in a high-dimensional setting. To do so, we need to assume certain conditions on both $\mathcal{L}$ and Φ.

3.1 Decomposability

First we introduce the notion of *decomposable* regularizers.

Definition 3.

3.2 Restricted Strong Convexity

3.3 A General Result

3.4 Revisiting LASSO regression

[1] We write $\mathcal{L}(\theta, Z_1, ..., Z_n) = \mathcal{L}(\theta)$ for brevity, but the dependence on the data maintains.