Least squares with sparsity

Notes by Parker Knight

April 18, 2022

Abstract

It is well known that, in the low-dimensional setting, the OLS estimator enjoys several nice properties, including consistency. However, in the high-dimensional regime, the aspect ratio stays bounded away from 0, and the estimator may not converge to the true parameter value. We may ameliorate this shortcoming by imposing additional structure on the true parameter; in particular, we assume that the parameter vector is sparse.

This note set will study nonasymptotic bounds on estimation and prediction error for the LASSO, the most well-known variant of sparse regression. To do so, we first develop a theory of tail bounds and sub-Gaussian random variables. After studying the LASSO, we give a brief treatment of general regularized M-estimators.

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.

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

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 $t \geq 0$*

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

Proof. By direct calculation:

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

□

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 $t \geq 0$*

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

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

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

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

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

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 Sparse regression

3 General M-estimators