Outline for Week 8 Presentation on Fast Rates and Sparse Linear Prediction

Tomas Vaškevičius

Setup

Consider a fixed-design linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}^{\star} + \boldsymbol{\xi},$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times d}, \boldsymbol{w}^{\star} \in \mathbb{R}^{d}, \boldsymbol{y} \in \mathbb{R}^{n}$ and $\boldsymbol{\xi} \sim N(0, \sigma^2 I_{n \times n})$. We assume that the following is true:

1. The ℓ_2 norms of columns of $\boldsymbol{X}/\sqrt{n}$ are bounded by 1;
2. The "true" parameter $\boldsymbol{w}^{\star} \in \mathbb{R}^{d}$ is s-sparse, meaning that only s entries of $\boldsymbol{w}^{\star}$ are non-zero.

We measure an estimator's $\widehat{\boldsymbol{w}} = \widehat{\boldsymbol{w}}(\boldsymbol{X}, \boldsymbol{y})$ statistical performance via its *in-sample prediction error* defined as

$$\mathcal{E}(\widehat{\boldsymbol{w}}) = \mathbf{E}_{\boldsymbol{\xi}}\left[\frac{1}{n}\|\boldsymbol{X}\widehat{\boldsymbol{w}} - \boldsymbol{X}\boldsymbol{w}^{\star}\|_2^2\right].$$

The Problem

Under the two assumptions stated above, the ℓ_0 norm constrained estimator

$$\widehat{\boldsymbol{w}}_{\ell_0} \in \operatorname{argmin}_{\{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\|_0 \leq s\}} \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$$

is known to attain the minimax optimal rate (see [Raskutti, Wainwright, and Yu, 2011])

$$\mathcal{E}(\widehat{\boldsymbol{w}}) \lesssim \frac{s\sigma^2 \log{(d/s)}}{n},$$

where the notation $\lesssim$ hides multiplicative universal constants independent of problem parameters.

The computational intractability of the $\widehat{\boldsymbol{w}}_{\ell_0}$ estimator has motivated the development of computationally-efficient estimators, most notably the lasso, which replaces the ℓ_0 constraint via an ℓ_1 penalty term as follows:

$$\widehat{\boldsymbol{w}}_{\ell_1}^{\lambda} \in \operatorname{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2+\lambda\|\boldsymbol{w}\|_1.$$

The above problem is convex and from the computational point of view it can be solved efficiently. However, the lasso is only known to satisfy the fast rate $O(1/n)$ bound on the in-sample prediction error $\mathcal{E}(\widehat{\boldsymbol{w}}_{\ell_1}^{\lambda})$ provided that the design matrix $\boldsymbol{X}$ satisfies additional regularity assumptions. If $\boldsymbol{X}$ is *only* assumed to satisfy the normalized columns assumption, several authors have proved "slow rate" $\Omega(1/\sqrt{n})$ lower-bounds on $\mathcal{E}(\widehat{w}_{\ell_1}^{\lambda})$, thus establishing *statistical* sub-optimality of the lasso in comparison with the ℓ_0 estimator $\widehat{\boldsymbol{w}}_{\ell_0}$ (see, e.g., [Candès and Plan, 2009, Foygel and Srebro, 2011, Dalalyan, Hebiri, and Lederer, 2017]). More broadly, Zhang, Wainwright, and Jordan [2014] show that the "slow rate" is intrinsic for any polynomial time algorithm that is constrained to output a *sparse* output vector $\widehat{\boldsymbol{w}}$.

Main Result

It remains unknown whether there exists a polynomial-time algorithm that attains the minimax-optimal "fast rate" on the in-sample prediction error. In the reading group meeting, we are going to review a result due to Zhang, Wainwright, and Jordan [2017], which establishes that the "slow rate" is unavoidable to a large class of M-estimators $\widehat{\boldsymbol{w}}_{\rho}^{\lambda}$ defined as follows:

$$\widehat{\boldsymbol{w}}_{\rho}^{\lambda} \in \mathcal{W}_{\rho}^{\lambda} = \left\{\boldsymbol{w}\in\mathbb{R}^d : \boldsymbol{w} \text{ is a local minimum of the function } \boldsymbol{w}\mapsto \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2+\lambda\rho(\boldsymbol{w})\right\}.$$

The penalty ρ is assumed to satisfy the following conditions:

1. The function ρ is coordinate-separable: i.e., there exist univariate functions ρ_i for $i=1,\dots,d$ such that $\rho(\boldsymbol{w})=\sum_{i=1}^d \rho_i(w_i)$;
2. For each i, we have $\rho_i(0)=0$ and for all $w\in\mathbb{R}$ we have $\rho_i(w)=\rho_i(-w)$;
3. Each function ρ_i is non-decreasing on $[0,\infty)$.

We remark that ρ is not assumed to be a convex. The below result contains a slightly simplified statement of the one presented in [Zhang, Wainwright, and Jordan, 2017].

Theorem (Theorem 1 in [Zhang, Wainwright, and Jordan, 2017])**.** *For any $d\geq n$ and large enough n, there exists a design matrix $\boldsymbol{X}$ (that satisfies the column-normalization condition) such that for any penalty ρ that satisfies the three conditions above, there exists a 2-sparse vector $\boldsymbol{w}^\star=\boldsymbol{w}^\star(\rho)$, such that the following holds:*

$$\mathbf{E}_{\boldsymbol{\xi}}\left[\inf_{\lambda\geq 0}\sup_{\boldsymbol{w}\in\mathcal{W}_{\rho}^{\lambda}}\frac{1}{n}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{X}\boldsymbol{w}^\star\|_2^2\right]\geq\Omega\left(1/\sqrt{n}\right).$$

Presentation Outline

Below is the plan for the presentation:

1. First, we are going to review some basic properties of the lasso. In particular, we will review the assumption-free "slow rate" upper-bound on $\mathcal{E}(\widehat{\boldsymbol{w}}_{\ell_1}^{\lambda})$ and the "fast rate" upper-bound that holds under the restricted eigenvalue condition.

2. We will then construct a design matrix $\boldsymbol{X}$ for which the lasso (with the classical regularization parameter $\lambda \asymp \sigma\sqrt{\log d}/\sqrt{n}$) incurs a suboptimal in-sample prediction error. This part of the presentation will be based on Section 2 in [Candès and Plan, 2009], which already contains the key ideas used in the proof of the Theorem of Zhang, Wainwright, and Jordan [2017].

3. Using the proof presented in [Zhang, Wainwright, and Jordan, 2017], we will extend the above sub-optimality result for the lasso for all scales of the parameter λ.

4. Finally, if time permits, we will discuss how the proof for the lasso extends to general penalties ρ satisfying the three conditions stated above.

References

Emmanuel J Candès and Yaniv Plan. Near-ideal model selection by 1 minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.

Arnak S Dalalyan, Mohamed Hebiri, and Johannes Lederer. On the prediction performance of the lasso. *Bernoulli*, 23(1):552–581, 2017.

Rina Foygel and Nathan Srebro. Fast-rate and optimistic-rate error bounds for l1-regularized regression. *arXiv preprint arXiv:1108.0373*, 2011.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over lq-balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.

Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948, 2014.

Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Optimal prediction for sparse linear models? lower bounds for coordinate-separable m-estimators. *Electronic Journal of Statistics*, 11(1):752–799, 2017.