Is Lasso Minimax Optimal?

Navid Ardeshir (na2844)

Statistics Department

December 2020

Abstract

Numerous work have been devoted to come up with estimators in the context of sparse linear regression with satisfactorial statistical properties. Based on the minimax framework one can compare these methods at the hope of introducing an optimal estimation procedure. Though, a family of optimal procedures are known but they suffer from computational intractability. This article will mainly focus on the minimax rates of sparse linear regression with fixed design and further tries to demonstrate that Lasso is optimal among the family of computationally efficient procedures.

Suppose a statistician is tasked to do a regression on a problem equipped with the knowledge that the true underlying model is at most k-sparse. Due to curse of dimensionality the more inactive covariates one keeps in the regression the less powerful the inference would become. Hence, the statistician ought to select a model among many, trading off between lesser power and selecting a wrong model. Consider the usual setting $y = X\beta^* + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$ and columns of X are normalized to have ℓ_2-norm equal to one. Further, σ is assumed to be known and β^* to be inside $\Theta \subseteq \mathbb{R}^d$ which represents the prior knowledge over the family of parameters.[1] One of many successful procedures for model selection, introduced in [Birge, Massart 2001] is the following:

$$\hat{\beta} \in \arg\min_{\beta\in\Theta} \|y - X\beta\|_2^2 + \sigma^2 \mathrm{Pen}(\beta) \tag{1}$$

where Pen:$\mathbb{R}^d \mapsto \mathbb{R}_+$ is a penalty over the family of models. In fact, [Birge, Massart 2001] showed the obtained estimator achieves minimax optimality for some family of penalties which only depends on the support of its input. More precisely, optimality is assessed through the worst case an estimator achieves over the set of k-sparse vectors $B_0(k)$:[2]

$$\mathcal{M}_\Theta(\hat{\beta}, X) \coloneqq \sup_{\beta\in\Theta} \mathbb{E}[\ \frac{\|X\hat{\beta} - X\beta\|_2^2}{n}\ |\ X, \beta]$$

Under the mild assumption on the sparsity $k \leq d/5$ the following minimax bound which relies heavily on the quality of the design matrix holds for any estimator (see [Verzelen 2012, Raskutti et al. 2011]):

$$\inf_{\hat{\beta}} \mathcal{M}_{\mathbb{B}_0(k)}(\hat{\beta}, X) \gtrsim k\frac{\log(\frac{d}{k})}{n}\sigma^2(\frac{\underline{\phi}_{2k}(X)}{\overline{\phi}_{2k}(X)})^2 \tag{2}$$

where $\underline{\phi}_k(X) \coloneqq \inf_{\beta\in\mathbb{B}_0(k)} \|X\beta\|_2/\|\beta\|_2 \leq \sup_{\beta\in\mathbb{B}_0(k)} \|X\beta\|_2/\|\beta\|_2 =: \overline{\phi}_k(X)$ are the restricted eigenvalues corresponding to the design matrix.[3] This dependence on design matrix is not ideal since even if X is comprised from orthogonal columns but a duplicate in the last column then the lower bound essentially

[1] Here we may assume that Θ is the family of parameters with at most k non zeros.

[2] Note that $\mathbb{B}_q(r) \coloneqq \{\beta \in \mathbb{R}^d \mid \|\beta\|_{\ell_q} \leq r\}$ is the ball with radius r induced by ℓ_q norm

[3] Note that $\gtrsim$ means greater than equal up to a constant independent from variables in the problem.

becomes zero! Nonetheless, there exists matrices (e.g. entries drawn from Gaussian) for which the ratio is close to one, hence, it immediately implies:

$$\sup_X \inf_{\hat{\beta}} \mathcal{M}_{\mathbb{B}_0(k)}(\hat{\beta}, X) \gtrsim k\frac{\log(\frac{d}{k})}{n}\sigma^2$$

Moreover, In order to demonstrate that the lower bound is sharp [Birge, Massart 2001] constructed estimators that matches with this lower bound uniformly for any design matrix. For instance, they considered the estimator $\hat{\beta}^{GS}$ (GS stands for Gaussian Selection) to be the solution of (1) with $\text{Pen}(\beta) := |m|(2 + \log(\frac{d}{|m|}))$ where m is the support of β and proved the following:

$$\sup_X \mathcal{M}_{\mathbb{B}_0(k)}(\hat{\beta}^{GS}, X) \lesssim k\frac{\log(\frac{d}{k})}{n}\sigma^2$$

To reiterate, this bound holds for any design matrix. Another interesting special case is $\text{Pen}(\beta) := \lambda\|\beta\|_0$ which can be rewritten as the following:

$$\hat{\beta}_k^{(0)} \in \arg\min \|y - X\beta\|_2 \equiv \arg\min \|y - X\beta\|_2^2 + \lambda\|\beta\|_0 \equiv \arg\min \|\beta\|_0 \quad \text{s.t. } \|\beta\|_0 \leq k \qquad \text{s.t. } \|y - X\beta\|_2 \leq \epsilon \tag{3}$$

Above the relation between λ, ϵ, and k is assumed to be implicit. The ℓ_0-based estimator also known to satisfy a uniform upper bound with the same rate [Bunea, et al. 2007]:

$$\sup_X \mathcal{M}_{\mathbb{B}_0(k)}(\hat{\beta}_k^{(0)}, X) \lesssim k\frac{\log(\frac{d}{k})}{n}\sigma^2$$

Although both of these penalized estimators $\hat{\beta}_k^{(0)}, \hat{\beta}^{GS}$ are optimal regardless of the design matrix, however, they are computationally expensive which suggests existence of a trade-off between optimality and computational efficiency. More precisely, if an upper bound on the sparsity is available to the statistician then these methods require a search among $\binom{d}{k}$ models, however, sparsity is usually unknown and the search becomes exponential.

It has been established in [Natarajan 1995] that finding the sparsest solution while the empirical prediction error (residuals) are in a certain ball is in general intractable. More precisely, given a tolerance level ϵ the right-most optimization problem in (3) can be reduced as an instance of "exact 3 cover problem" which is famously known as X3C to be NP-complete. It is also noteworthy that the reduction relies on a specific design matrix X, which provides more evidence that the dependence on the design matrix in lower bounds is unavoidable. In fact, alternative methods which are computationally tractable mostly rely on some sort of condition on X in order to derive oracle bounds as opposed to optimal methods mentioned earlier; In particular, Lasso [Tibshirani, 1996] and Dantzig Selector [Candes, Tao 2007] are of such where both can be computed via convex optimization problems. Suppose we denote the estimator obtained via either Lasso or Dantzig Selector as $\hat{\beta}_\lambda^{(1)}$ for $\lambda = \Omega(\sqrt{2\log(p)})$ then [Geer, Bühlmann 2009, Bickel, et al. 2009] shows an upper bound based on a modified restricted eigenvalue quantity:[4]

$$\mathcal{M}_{\mathbb{B}_0(k)}(\hat{\beta}^{(1)}, X) \lesssim k\frac{\log(\frac{d}{k})}{n}\sigma^2(\frac{\overline{\phi}_{2k}(X)}{\gamma_k(X)})^2 \quad , \quad \gamma_k(X) := \inf_{\beta \in \cup_{|S|\leq k}\mathbb{C}(S)} \frac{\|X\beta\|_2}{\|\beta\|_2}$$

Where $\mathbb{C}(S) = \{\beta \in \mathbb{R}^d \mid \|\beta_{S^c}\|_1 \leq 3\|\beta_S\|_1\}$ is a cone that any vector in it is approximately supported on S. Unfortunately, one can not get rid of this ratio if they stick to computationally efficient methods! Surprisingly, [Zhang, et al. 2014] proves that there exists a family of "bad" design matrices for which any estimator that can be computed in polynomial time satisfies the following lower bound:

$$\inf_{\hat{\beta}} \mathcal{M}_{\mathbb{B}_0(k)}(\hat{\beta}, X) \gtrsim k\frac{\log(\frac{d}{k})}{n}\sigma^2(\frac{\overline{\phi}_{2k}(X)}{\gamma_k(X)})^2$$

[4]Note that the estimator is not necessarily k-sparse so they take a threshold version of it which contains the k greatest elements.

where the infimum is over all estimators with polynomial algorithmic complexity. In other words, this result simply shows that there is a fundamental gap between optimal methods and computationally efficient ones among which Lasso and Dantzig Selector achieve the optimal rate. Indeed, these results relies on the assumption that variance is known which does not seem to be practical. However, in the case of unknown variance there aren't such definite answers [Verzelen 2012]. It is worth mentioning that this line of work does not recognize Lasso as a silver bullet! Many other metrics might be of interest in practice which changes the notion of optimal.

References

[Giraud, 2014] Giraud, C., 2014. Introduction to high-dimensional statistics (Vol. 138). CRC Press.

[Natarajan 1995] Natarajan, B.K., 1995. Sparse approximate solutions to linear systems. SIAM journal on computing, 24(2), pp.227-234.

[Birge, Massart 2001] Birgé, L. and Massart, P., 2001. Gaussian model selection. Journal of the European Mathematical Society, 3(3), pp.203-268.

[Bunea, et al. 2007] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H., 2007. Aggregation for Gaussian regression. The Annals of Statistics, 35(4), pp.1674-1697.

[Verzelen 2012] Verzelen, N., 2012. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. Electronic Journal of Statistics, 6, pp.38-90.

[Raskutti et al. 2011] Raskutti, G., Wainwright, M.J. and Yu, B., 2011. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. IEEE transactions on information theory, 57(10), pp.6976-6994.

[Ye, Zhang 2010] Ye, F. and Zhang, C.H., 2010. Rate minimaxity of the Lasso and Dantzig selector for the lq loss in lr balls. The Journal of Machine Learning Research, 11, pp.3519-3540.

[Bickel, et al. 2009] Bickel, P.J., Ritov, Y.A. and Tsybakov, A.B., 2009. Simultaneous analysis of Lasso and Dantzig selector. The Annals of statistics, 37(4), pp.1705-1732.

[Geer, Bühlmann 2009] Van De Geer, S.A. and Bühlmann, P., 2009. On the conditions used to prove oracle results for the Lasso. Electronic Journal of Statistics, 3, pp.1360-1392.

[Zhang, et al. 2014] Zhang, Y., Wainwright, M.J. and Jordan, M.I., 2014, May. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In Conference on Learning Theory (pp. 921-948).

[Tibshirani, 1996] Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), pp.267-288.

[Candes, Tao 2007] Candes, E. and Tao, T., 2007. The Dantzig selector: Statistical estimation when p is much larger than n. The annals of Statistics, 35(6), pp.2313-2351.