08 High-dimensional regression & sparsity

For smooth nonparametric regression, the best possible rate of convergence is

$$n^{\frac{-1}{2+d/s}}$$

which can be slow when d is large relative to the assumed number of bounded derivatives s (e.g., when $d = 20$ and $s = 2$, this gives a $n^{-1/12}$ rate). Additive and partially linear models can yield faster rates but the required assumptions may be too strict and could be violated, resulting in non-trivial bias or approximation error. Here we discuss sparsity and the lasso, which are particularly useful when the dimension of the features d is greater than sample size n.

1 Motivation

Consider the linear model

$$\mu(x) \equiv \mathbb{E}(Y \mid X = x) = x^{\mathrm{T}}\beta.$$

For low-dimensional X linearity is typically viewed as a strong assumption, but things are different here since we allow X to be high-dimensional, i.e., $X \in \mathbb{R}^d$ with $d \geq n$. For example we could let $X = b(V)$ be a high-dimensional basis vector transformation of some lower-dimensional covariates V. In the high-dimensional setting we often take $d \asymp n^{\alpha}$ for some $\alpha > 1$.

The classic estimator in low dimensions with $d < n$ is the least squares estimator

$$\underset{b \in \mathbb{R}^d}{\arg\min}\ \mathbb{P}_n\Big\{(Y - X^{\mathrm{T}}b)^2\Big\}.$$

We studied this estimator in the series estimation context where the dimension $d = k < n$ but can scale with n. However, when $d > n$, least squares breaks down, for example, giving infinitely many solutions. For example, $\widehat{\beta} + v$ is a solution for any v in the null space of the design matrix $\mathbf{X}$, for any solution $\widehat{\beta}$. This is not just a deficiency of least squares: Dicker (2012) shows that without any further assumptions beyond $\|\beta\|_2 \leq C$, the minimax rate for estimating $\mu(x) = x^{\mathrm{T}}\beta$ in MSE is d/n, so consistent estimation is indeed impossible when $d > n$.

One natural way to proceed is to impose extra structure on the size of the coefficients β, i.e., sparsity. For example, one can assume *hard sparsity* by imposing a bound on the ℓ_0 norm

$$\|\beta\|_0 \equiv \sum_{j=1}^{d} \mathbb{1}(\beta_j \neq 0) \leq s$$

(though note this is not technically a norm). This says at most s of the coefficients in β are non-zero and the rest do not matter. Alternatively one can assume *weak sparsity* by imposing a bound on the ℓ_1 norm

$$\|\beta\|_1 \equiv \sum_{j=1}^{d} |\beta_j| \leq s.$$

This allows all the coefficients to be non-zero, but restricts their magnitude to decay quickly enough with d. Note if each $|\beta_j| \leq C$ then $\|\beta\|_1 \leq C\|\beta\|_0$, so a bound on the ℓ_0 norm implies the same bound on the ℓ_1 norm, up to a constant factor.

2 Best subset, the lasso, and ridge regression

Given the imposed structure on the ℓ_0 or ℓ_1 norms of β, it is natural to consider constrained or penalized versions of least squares. For example, one could use the estimator

$$\widehat{\beta}_{\text{bs}} = \underset{b \in \mathbb{R}^d}{\arg\min} \ \mathbb{P}_n\Big\{(Y - X^{\mathrm{T}} b)^2\Big\} \ \text{ subject to } \ \|b\|_0 \leq t$$

for some t (e.g., $t = s$ if s were known, otherwise t can be viewed as a tuning parameter), which is called *best subset selection*. Similarly, one could use the estimator

$$\widehat{\beta}_{\text{lasso}} = \underset{b \in \mathbb{R}^d}{\arg\min} \ \mathbb{P}_n\Big\{(Y - X^{\mathrm{T}} b)^2\Big\} \ \text{ subject to } \ \|b\|_1 \leq t,$$

which is called the *lasso* (for "Least Absolute Shrinkage and Selection Operator"). The lasso is one of the most popular regression methods, and is attributed to Tibshirani [1996], "Regression shrinkage and selection via the lasso". Note the similarity of the above to the estimator

$$\widehat{\beta}_{\text{ridge}} = \underset{b \in \mathbb{R}^d}{\arg\min} \ \mathbb{P}_n\Big\{(Y - X^{\mathrm{T}} b)^2\Big\} \ \text{ subject to } \ \|b\|_2^2 \leq t$$

for the ℓ_2 norm $\|b\|_2^2 = \sum_j b_j^2$, which is called *ridge regression* [Hoerl and Kennard, 1970]. One can also use analogous penalized estimators

$$\begin{aligned}
&\underset{b \in \mathbb{R}^d}{\arg\min} \ \mathbb{P}_n\Big\{(Y - X^{\mathrm{T}} b)^2\Big\} + \lambda \|b\|_0 \\
&\underset{b \in \mathbb{R}^d}{\arg\min} \ \mathbb{P}_n\Big\{(Y - X^{\mathrm{T}} b)^2\Big\} + \lambda \|b\|_1 \\
&\underset{b \in \mathbb{R}^d}{\arg\min} \ \mathbb{P}_n\Big\{(Y - X^{\mathrm{T}} b)^2\Big\} + \lambda \|b\|_2^2.
\end{aligned}$$

In fact, the constrained and penalized versions are equivalent for the lasso and the ridge, in the sense that for any $t \geq 0$ and resulting constrained estimator $\widehat{\beta}$ there is a corresponding $\lambda \geq 0$ such that $\widehat{\beta}$ also equals the penalized estimator. For this reason we mostly treat the constrained/penalized formulations as equivalent, for the lasso and ridge. (Note though that the exact correspondence between t and λ depends on the data, so in practice selecting λ and t via cross-validation can lead to different estimators depending on whether one uses the constrained or penalized approaches).

However, the constrained and penalized versions of the best subset selection estimators are not equivalent. For every $\lambda \geq 0$ and resulting penalized estimator $\widehat{\beta}$, there is a corresponding value $t \geq 0$ for which $\widehat{\beta}$ also equals the constrained estimator, but not vice versa. Importantly, best subset selection is also nonconvex and NP hard, while lasso and ridge are convex optimization problems. This makes the latter two estimates much easier to compute in practice (in fact there is a closed-form expression for the ridge regression estimator).

The following figure (taken from these notes) illustrates an example lasso path, tracing out estimates $\widehat{\beta}_j$ across a range of λ values. When $\lambda = 0$ we do not penalize/constrain at all, and recover the least squares estimates if feasible, while as λ grows we penalize/constrain more, and push more coefficient estimates to zero.

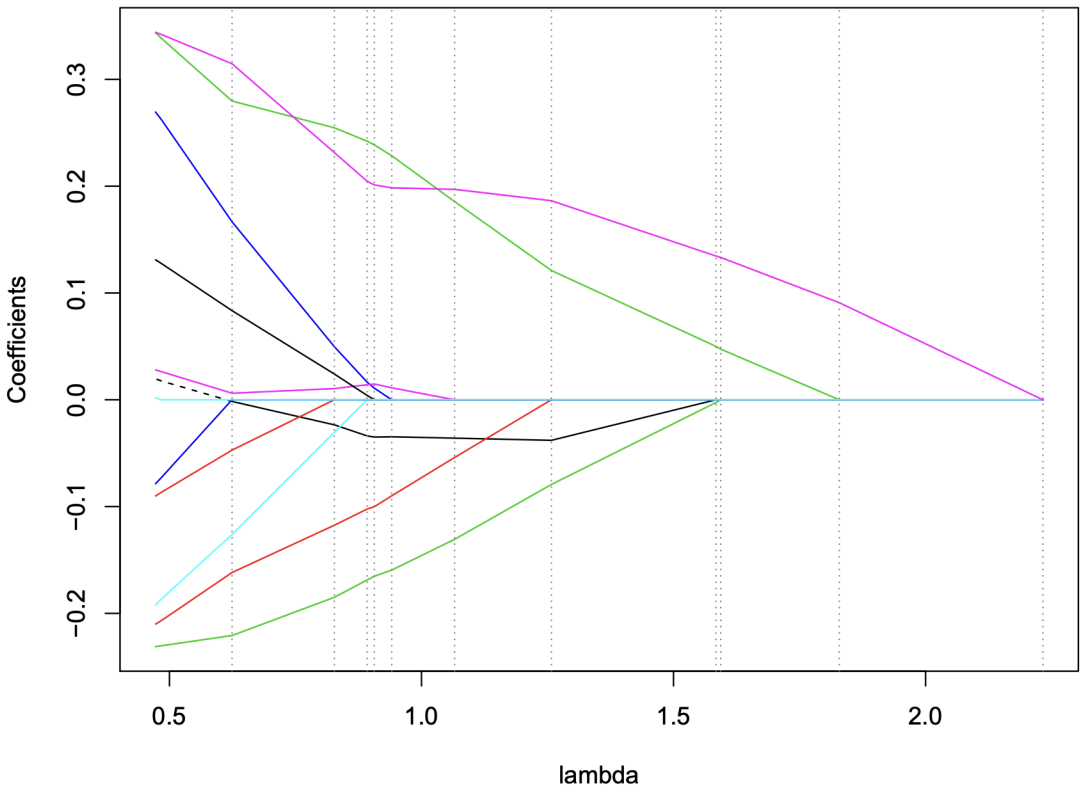


Figure 3: *An example of the lasso path. Each colored line denotes a component of the lasso solution* $\hat{\beta}_j(\lambda)$*,* $j = 1, \ldots, p$ *as a function of* λ*. The gray dotted vertical lines mark the knots* $\lambda_1 \geq \lambda_2 \geq \ldots$

The lasso path is a piecewise linear and continuous function of λ. The "knots" in the figure above are the points at which the "active set" $\text{supp}(\widehat{\beta})$ changes. Although typically knots correspond to variables entering the active set as λ decreases, somewhat surprisingly, variables can also leave the active set as λ decreases.

Crucially, best subset selection and the lasso yield sparse solutions, i.e., many $\widehat{\beta}_j$ are exactly equal to zero (depending on the choice of t and λ); this is not true for ridge regression, which generically has all $\widehat{\beta}_j \neq 0$ regardless of t or λ. It is clear that constraining the ℓ_0 norm will impose sparsity; but why is this true for ℓ_1 but not ℓ_2? Figure 3.11 from Hastie et al. [2009] illustrates why. In this plot there are only two features, and the full least squares estimate $\widehat{\beta}$ is shown as a black dot, with contours of the least squares error shown in red. We minimize the error constrained to the diamond $|\beta_1| + |\beta_2| \leq t$ for the lasso, and to the disk $\beta_1^2 + \beta_2^2 \leq t$ for ridge regression. As Hastie et al. [2009] note:

> Unlike the disk, the diamond has corners; if the solution occurs at a corner, then it has one parameter equal to zero. When $d > 2$, the diamond becomes a rhomboid, and has many corners, flat edges and faces; there are many more opportunities for the estimated parameters to be zero.

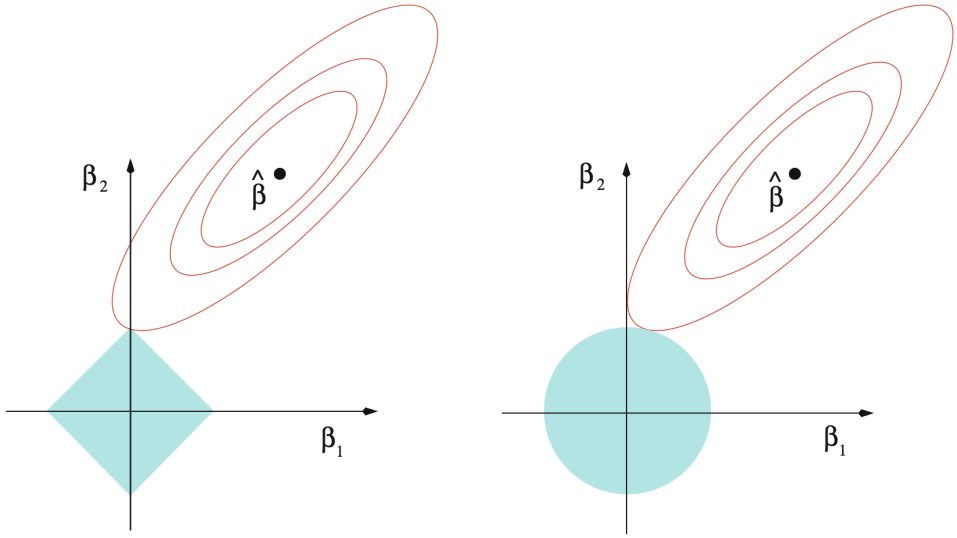


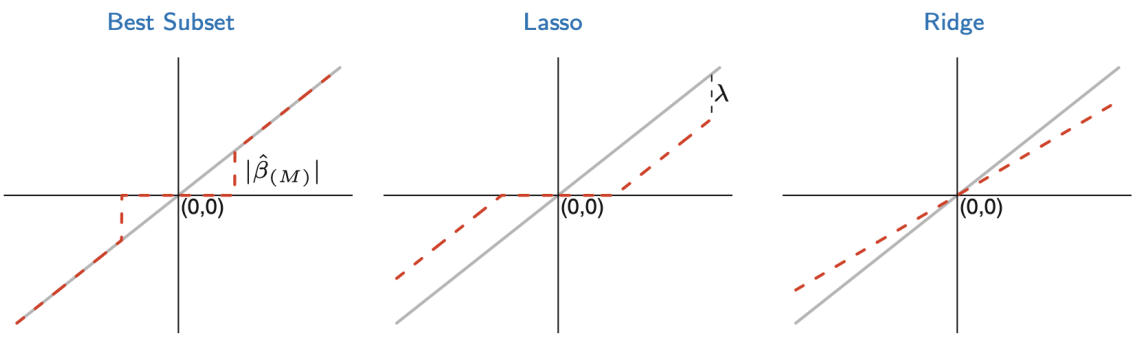
FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

3 Orthogonal design

In the orthogonal design setting where $\mathbb{P}_n(XX^{\mathrm{T}}) = I_d$, the best subset, lasso, and ridge estimators all have closed form expressions (and note that ridge does more generally). Here the least squares estimates are unique and equal $\widehat{\beta}_{\mathrm{ls}} = \mathbb{P}_n(XY)$, and

$$\begin{aligned}
\widehat{\beta}_{\mathrm{bs},j} &= \widehat{\beta}_j \mathbb{1}(|\widehat{\beta}_j| \geq |\widehat{\beta}_{(m)}|) \\
\widehat{\beta}_{\mathrm{lasso},j} &= \mathrm{sign}(\widehat{\beta}_j)(|\widehat{\beta}_j| - \lambda)_+ \\
\widehat{\beta}_{\mathrm{ridge},j} &= \widehat{\beta}_j/(1+\lambda)
\end{aligned}$$

where $\widehat{\beta}_{(m)}$ denotes the m^{th} largest coefficient, for some m depending on λ. This shows how best subset selection does a form of "hard thresholding", lasso does "soft thresholding", and ridge does simple shrinkage. This is illustrated below.



4 Analysis of the lasso

4.1 Slow rates

Our first result follows the logic of Chatterjee [2013].

Theorem 1. *Assume* $\max(|Y - \mathbb{E}(Y \mid X)|, |X_j|) \leq B$ *and*

$$\mathbb{E}(Y \mid X = x) = x^{\mathrm{T}}\beta \quad \textit{with} \quad \|\beta\|_1 \leq s.$$

Then the lasso estimator $\widehat{\mu}(x) = x^{\mathrm{T}}\widehat{\beta}$ *with* $\widehat{\beta} = \arg\min_{b:\|b\|_1 \leq s} \mathbb{P}_n\{(Y - X^{\mathrm{T}}b)^2\}$ *satisfies*

$$\mathbb{E}\|\widehat{\mu} - \mu\|^2 \leq (8B^2)\Big(2\sqrt{2}s + 1\Big)s\sqrt{\frac{\log d}{n}}.$$

Proof. We want to bound the expectation of

$$\|\widehat{\mu} - \mu\|^2 = \mathbb{P}\Big\{\Big(X^{\mathrm{T}}(\widehat{\beta} - \beta)\Big)^2\Big\} = (\mathbb{P} - \mathbb{P}_n)\Big\{\Big(X^{\mathrm{T}}(\widehat{\beta} - \beta)\Big)^2\Big\} + \mathbb{P}_n\Big\{\Big(X^{\mathrm{T}}(\widehat{\beta} - \beta)\Big)^2\Big\}$$

where $\mathbb{P}f(X) = \int f(x)\, d\mathbb{P}(x)$. Consider the second term first. Note we have

$$\mathbb{P}_n\left\{\left(Y - X^{\mathrm{T}}\widehat{\beta}\right)^2\right\} \leq \mathbb{P}_n\left\{\left(Y - X^{\mathrm{T}}\beta\right)^2\right\}$$

by definition of $\widehat{\beta}$ as a minimizer and since $\|\beta\|_1 \leq s$, which implies after rearranging that

$$\begin{aligned}
&\mathbb{P}_n\left\{-2YX^{\mathrm{T}}\widehat{\beta} + (X^{\mathrm{T}}\widehat{\beta})^2\right\} \leq \mathbb{P}_n\left\{-2YX^{\mathrm{T}}\beta + (X^{\mathrm{T}}\beta)^2\right\} \\
&\implies \mathbb{P}_n\left\{(X^{\mathrm{T}}\beta)^2 - 2(X^{\mathrm{T}}\beta X^{\mathrm{T}}\widehat{\beta}) + (X^{\mathrm{T}}\widehat{\beta})^2\right\} \leq \mathbb{P}_n\left\{2YX^{\mathrm{T}}(\widehat{\beta} - \beta) + 2(X^{\mathrm{T}}\beta)^2 - 2(X^{\mathrm{T}}\beta X^{\mathrm{T}}\widehat{\beta})\right\} \\
&\implies \mathbb{P}_n\left\{(X^{\mathrm{T}}(\widehat{\beta} - \beta))^2\right\} \leq 2\mathbb{P}_n\left\{(Y - X^{\mathrm{T}}\beta)(X^{\mathrm{T}}(\widehat{\beta} - \beta))\right\}.
\end{aligned}$$

This is called the *basic inequality* for the lasso, equivalently $\|\mathbf{X}(\widehat{\beta} - \beta)\|_2^2 \leq 2\langle \epsilon, \mathbf{X}(\widehat{\beta} - \beta)\rangle$ for $\mathbf{X}$ the design matrix with $\mathbf{X}^{\mathrm{T}}\mathbf{X} = n\mathbb{P}_n(XX^{\mathrm{T}})$, and ϵ the stacked residuals $Y - X^{\mathrm{T}}\beta$. Therefore

$$\begin{aligned}
\mathbb{P}_n\left\{(X^{\mathrm{T}}(\widehat{\beta} - \beta))^2\right\} \leq 2\mathbb{P}_n\left\{(Y - X^{\mathrm{T}}\beta)(X^{\mathrm{T}}(\widehat{\beta} - \beta))\right\} &= 2\sum_{j=1}^{d} \mathbb{P}_n\left\{(Y - X^{\mathrm{T}}\beta)X_j\right\}(\widehat{\beta}_j - \beta_j) \\
&\leq 2\max_j \left|\mathbb{P}_n\left\{(Y - X^{\mathrm{T}}\beta)X_j\right\}\right| \|\widehat{\beta} - \beta\|_1 \leq (4s/n)\max_j \left|\sum_i \{(Y_i - X_i^{\mathrm{T}}\beta)X_{ji}\}\right|.
\end{aligned}$$

To bound the expectation of this maximum, we use the following lemma.

Lemma 1. *Assume* Z_j, $j = 1, ..., d$ *are mean-zero random variables, possibly dependent, but sub-Gaussian with parameter* σ, *i.e.,* $\mathbb{E}\{\exp(tZ_j)\} \leq \exp(t^2\sigma^2/2)$ *for all* t. *Then*

$$\mathbb{E}\left(\max_j |Z_j|\right) \leq 2\sigma\sqrt{\log d}.$$

Proof. Exercise 2.12 of Wainwright [2019]. □

Now note each $Z_j = \sum_i (Y_i - X_i^{\mathrm{T}}\beta)X_{ji}$ is mean-zero and sub-Gaussian with parameter $\sqrt{n}B^2$, since each independent $(Y - X^{\mathrm{T}}\beta)X_j \in [-B^2, B^2]$ is sub-Gaussian with parameter B^2. Therefore

$$\mathbb{E}\left[\mathbb{P}_n\left\{(X^{\mathrm{T}}(\widehat{\beta} - \beta))^2\right\}\right] \leq \frac{4s}{n}(2B^2\sqrt{n})\sqrt{\log d} = (8B^2)s\sqrt{\frac{\log d}{n}}$$

by Lemma 1.

Now we consider the second term in the initial decomposition

$$(\mathbb{P} - \mathbb{P}_n)\left\{\left(X^{\mathrm{T}}\beta - X^{\mathrm{T}}\widehat{\beta}\right)^2\right\}.$$

For it note that

$$\mathbb{P}\left\{\left(X^{\mathrm{T}}(\beta - \widehat{\beta})\right)^2\right\} = (\beta - \widehat{\beta})^{\mathrm{T}}\mathbb{E}(XX^{\mathrm{T}})(\beta - \widehat{\beta}) = \sum_j \sum_k (\beta_j - \widehat{\beta}_j)\mathbb{E}(X_jX_k)(\beta_k - \widehat{\beta}_k)$$

$$\mathbb{P}_n\left\{\left(X^{\mathrm{T}}(\beta - \widehat{\beta})\right)^2\right\} = (\beta - \widehat{\beta})^{\mathrm{T}}\mathbb{P}_n(XX^{\mathrm{T}})(\beta - \widehat{\beta}) = \sum_j \sum_k (\beta_j - \widehat{\beta}_j)\mathbb{P}_n(X_jX_k)(\beta_k - \widehat{\beta}_k)$$

so subtracting and using the fact that $\max(\|\widehat{\beta}\|_1, \|\beta\|_1) \leq s$ gives

$$\begin{aligned}
(\mathbb{P} - \mathbb{P}_n)\left\{\left(X^{\mathrm{T}}\beta - X^{\mathrm{T}}\widehat{\beta}\right)^2\right\} &= \sum_j \sum_k (\beta_j - \widehat{\beta}_j)(\mathbb{P} - \mathbb{P}_n)(X_jX_k)(\beta_k - \widehat{\beta}_k) \\
&\leq \max_{j,k} |(\mathbb{P} - \mathbb{P}_n)(X_jX_k)| \sum_j |\beta_j - \widehat{\beta}_j| \sum_k |\beta_k - \widehat{\beta}_k| \\
&\leq 4s^2 \max_{j,k} |(\mathbb{P} - \mathbb{P}_n)(X_jX_k)|.
\end{aligned}$$

Therefore

$$\mathbb{E}\left[(\mathbb{P} - \mathbb{P}_n)\left\{\left(X^{\mathrm{T}}\beta - X^{\mathrm{T}}\widehat{\beta}\right)^2\right\}\right] \leq \frac{4s^2}{n}(4\sqrt{n}B^2)\sqrt{\log d^2} = 16B^2s^2\sqrt{\frac{2\log d}{n}}$$

where the last line used Lemma 1, together with the fact that each $\sum_i (X_{ji}X_{ki} - \mathbb{E}(X_{ji}X_{ki}))$ is sub-Gaussian with parameter at most $\sqrt{n}2B^2$ since $(X_jX_k - \mathbb{E}(X_jX_k)) \in [-2B^2, 2B^2]$ is sub-Gaussian with parameter $2B^2$, and we have d^2 such variables for every $j, k = 1, ..., d$. □

Crucially the above shows $\mathbb{E}\|\widehat{\mu} - \mu\|^2 \lesssim s^2\sqrt{\frac{\log d}{n}}$, i.e., the mean $L_2(\mathbb{P})$ error only has a logarithmic dependence on the dimension d. Therefore even when dimension $d \asymp n^\alpha$ is very large relative to sample size, we get squared error rates of the form

$$s^2\sqrt{\frac{\alpha \log n}{n}}$$

paying only a constant factor and logarithmic penalty in sample size. This also shows the lasso is consistent as long as $s = o\left((n/\log d)^{1/4}\right)$, i.e., if $s \ll n^{1/4}$, ignoring log factors.

4.2 Fast rates

There are two basic flavors of results appearing in the lasso literature. The previous subsection is of the first flavor, which employs mild assumptions (e.g., mere boundedness), and obtains "slow" rates, of the order in the above result. In what follows we discuss the second flavor, which employs stronger assumptions (e.g., that features X_j are unrelated in various ways), and obtains "fast" rates, typically of the order $s \log(d/s)/n$. The assumptions on the predictors include so-called "restricted eigenvalue", "incoherence", and "restricted isometry" conditions, for example, and essentially limit how correlated the features can be.

Pairwise incoherence is perhaps the easiest such assumption to understand, albeit strongest. For sparsity s we say the Gram matrix $\mathbb{P}_n(XX^{\mathrm{T}})$ is pairwise incoherent if

$$\max_{j,k} \left| \mathbb{P}_n(X_j X_k) - \mathbb{1}(j=k) \right| = \max_{j,k} \left| \Big(\mathbb{P}_n(XX^{\mathrm{T}}) - I_d \Big)_{j,k} \right| \leq C/s.$$

This simply restricts how far any row-column element of the Gram matrix is from the identity. Restricted isometry is similar, but instead restricts the operator norm of submatrices of $\mathbb{P}_n(XX^{\mathrm{T}}) - I_d$. These conditions were developed in the compressed sensing literature, and are perhaps better suited to noiseless settings where features are under the control of the analyst.

A weaker and perhaps more appropriate condition in standard statistical settings is the following "restricted eigenvalue" condition. We say the design matrix $\mathbf{X}$ satisfies the restricted eigenvalue (RE) condition over support S with parameters (κ, α) if

$$\|\mathbf{X}v\|_2^2/n \geq \kappa \|v\|_2^2 \quad \text{for all vectors } v \;\; \text{s.t. } \|v_{S^c}\|_1 \leq \alpha \|v_S\|_1$$

where for a set $T \subseteq \{1, ..., d\}$, we let $T^c = \{1, ..., d\} \setminus T$, and $v_T = \{v_j : j \in T\}$. To get some intuition for this condition, note that if the eigenvalues of the Gram matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}/n$ were bounded below away from zero by some $C > 0$ (which is a condition we used for series estimation, for example) then by the min-max theorem

$$\|\mathbf{X}v\|_2^2/n \geq C\|v\|_2^2 \quad \text{for all vectors } v \neq 0.$$

When $d > n$ this bounded eigenvalue condition cannot hold; the smallest eigenvalue is zero. Therefore we can view the RE condition as a relaxation, where we do not need the above lower bound to hold *for all* vectors v, but instead only for some vectors, and specifically those that are smaller (in ℓ_1 norm) off the support S than on it.

To start to see why this condition can help attain faster rates, recall earlier we showed the inequality $\mathbb{P}_n\{(X^{\mathrm{T}}(\widehat{\beta} - \beta))^2\} \leq 2 \max_j |\mathbb{P}_n(\epsilon X_j)| \|\widehat{\beta} - \beta\|_1$. The first term in this product is bounded above by $C\sqrt{\log d/n}$ with high probability (recall we used an analogous in-expectation result), and the second term can be shown to be bounded above by $Cs\sqrt{\log d/n}$ with high probability. This follows from the fact that $\|\widehat{\beta} - \beta\|_1 \leq 2\sqrt{s}\|\widehat{\beta} - \beta\|_2$ and from applying the RE condition to $v = \widehat{\beta} - \beta$ to obtain

$$\kappa \|\widehat{\beta} - \beta\|_2^2 \leq \left\| \mathbf{X}(\widehat{\beta} - \beta) \right\|_2^2 / n = \mathbb{P}_n \left\{ (X^{\mathrm{T}}(\widehat{\beta} - \beta))^2 \right\} \leq 2 \max_j |\mathbb{P}_n(\epsilon X_j)| \left(2\sqrt{s} \|\widehat{\beta} - \beta\|_2 \right),$$

which implies

$$\|\widehat{\beta} - \beta\|_2 \leq \frac{4\sqrt{s}}{\kappa} \max_j |\mathbb{P}_n(\epsilon X_j)| \implies \|\widehat{\beta} - \beta\|_1 \leq \frac{8s}{\kappa} \max_j |\mathbb{P}_n(\epsilon X_j)|$$

so that $\mathbb{P}_n\{(X^{\mathrm{T}}(\widehat{\beta} - \beta))^2\} \leq \frac{16s}{\kappa} \{\max_j |\mathbb{P}_n(\epsilon X_j)|\}^2$; see Theorem 7.13 of Wainwright [2019].

5 Minimax rates

Are the above lasso rates optimal? What are the best rates possible under different sparsity structures? These and related questions are resolved by Raskutti et al. [2011]. They consider the general hard and soft sparsity models where $\beta \in \mathbb{B}_q(s)$ for

$$\mathbb{B}_q(s) \equiv \{\beta \in \mathbb{R}^d : \|\beta\|_q^q = \sum_j |\beta_j|^q \leq s\}$$

with the limiting case the ℓ_0-ball $\mathbb{B}_0(s) = \{\beta \in \mathbb{R}^d : \|\beta\|_0 = \sum_j \mathbb{1}(\beta_j \neq 0) \leq s\}$. They show that, under RE-type conditions on the Gram matrix, the minimax rate for in-sample prediction error under soft ℓ_1-sparsity constraints is lower bounded as

$$\inf_{\widehat{\beta}} \sup_{\beta \in \mathbb{B}_1(s)} \mathbb{E}\left(\mathbb{P}_n\left[\{X^{\mathrm{T}}(\widehat{\beta} - \beta)\}^2\right]\right) \;\geq\; Cs\sqrt{\frac{\log d}{n}}$$

and for hard ℓ_0-sparsity constraints as

$$\inf_{\widehat{\beta}} \sup_{\beta \in \mathbb{B}_0(s)} \mathbb{E}\left(\mathbb{P}_n\left[\{X^{\mathrm{T}}(\widehat{\beta} - \beta)\}^2\right]\right) \;\geq\; C\frac{s\log(d/s)}{n}.$$

Raskutti et al. [2011] go on to show that, *without any conditions on the Gram matrix (beyond normalization when $q > 0$)*, these rates are attainable by least squares estimators

$$\widehat{\beta} = \operatorname*{arg\,min}_{\beta \in \mathbb{B}_q(s)} \mathbb{P}_n\Big\{(Y - X^{\mathrm{T}}b)^2\Big\}.$$

Importantly, when $q = 1$ this corresponds to the constrained version of the lasso; but when $q = 0$ this estimator is akin to best subset selection. Therefore the lasso is optimal under soft ℓ_1-sparsity under weak conditions, but seemingly not under hard ℓ_0-sparsity, since current analyses seem to require RE-type conditions; to optimally exploit hard ℓ_0-sparsity, it seems that more computationally expensive methods akin to best subset selection might be required. As far as I am aware, it is an open problem whether the lasso necessarily requires RE-type conditions to attain optimal rates under hard sparsity, or whether current analyses simply need to be improved. Similarly, it is also apparently open to determine more generally whether other more computationally tractable procedures are optimal under hard sparsity, beyond best subset selection.

Remark 1. Note the *lower bound* on the minimax rate requires conditions on the Gram matrix, but not the upper bound. This is because, for example, if all the covariates are exactly equal to each other then parametric $1/n$ rates are attainable, since the problem is essentially one-dimensional. Interestingly, although the minimax rates are the same for parameter estimation, where the error metric is $\mathbb{E}\|\widehat{\beta} - \beta\|_2^2$, the situation regarding assumptions on the Gram matrix is completely reversed. For parameter estimation the *lower bound* does not require any conditions on the Gram matrix, while *upper bounds do*. This is because the easiest cases for prediction are in some sense the hardest for parameter estimation; for example, when all the covariates are equal, the parameter estimation problem becomes nonidentifiable, while as mentioned earlier parametric rates are attainable for prediction error.

6 Further topics

The lasso/high-dimensional literature is huge; there are many additional topics not covered here. For example, we have focused on prediction/estimation of the regression function μ (with error metric $\mathbb{P}_n\{(\widehat{\mu} - \mu)^2\}$ or $\|\widehat{\mu} - \mu\|^2$); however, it also of interest to study parameter estimation (with error metric $\|\widehat{\beta} - \beta\|_r$ for some $r \geq 1$) and/or variable selection (i.e., is $\text{supp}(\widehat{\beta}) = \text{supp}(\beta)$?). In terms of assumptions, variable selection requires more than parameter estimation, which requires more than prediction/regression estimation.

For textbook treatments of these and other topics, Bühlmann and Van De Geer [2011] and Wainwright [2019] are excellent resources. See below for a sampling of topics of interest and corresponding important/representative papers/textbooks.

- stepwise methods (ESL 2009)
- adaptive lasso (Zou 2006)
- Dantzig selector (Candes & Tao 2007)
- group lasso (Bakin 1999, Yuan & Lin 2006)
- intrinsic dimension (Bickel & Li 2008)
- sparse additive models (Ravikumar, Lafferty, Liu, & Wasserman 2009)
- square-root lasso (Belloni, Chernozhukov, & Wang 2011)
- minimax optimality (Raskutti, Wainwright, & Yu 2011)
- variable selection & support recovery (Buhlmann & van de Geer 2011)
- connections to aggregation (Rigollet & Tsybakov 2011)
- variance estimation (Dicker 2014)
- computational feasibility of best subset selection (Bertsimas, King, & Mazumder 2016)
- confidence intervals (Cai & Guo 2017)
- signal-to-noise ratio estimation (Verzelen & Gassiat 2018)
- sparsity in causal inference (Bradic, Wager, & Zhu 2019, Liu & Wang 2023)

References

P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

S. Chatterjee. Assumptionless consistency of the lasso. *arXiv preprint arXiv:1303.5817*, 2013.

T. J. Hastie, R. J. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*, volume 2. Springer, 2009.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

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

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.