Sparsity, the Lasso, and Friends

Statistical Machine Learning, Spring 2017
Ryan Tibshirani (with Larry Wasserman)

1 Introduction

1.1 Basic setup

- Consider i.i.d. samples $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$ from the linear model

$$y_i = x_i^T \beta_0 + \epsilon_i, \quad i = 1, \ldots, n \tag{1}$$

  where $\beta_0 \in \mathbb{R}^p$ is an unknown coefficient vector, and ϵ_i, $i = 1, \ldots, n$ are random errors with mean zero. Here and throughout, without a loss of generality, we'll ignore the intercept term. We can more succintly express this data model as

$$y = X\beta_0 + \epsilon, \tag{2}$$

  where $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ is the vector of responses, $X \in \mathbb{R}^{n \times p}$ is the matrix of predictor variables, with ith row x_i, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$ is the vector of errors

- In the above, we have assumed that $\mathbb{E}(y_i|x_i)$ is a linear function of x_i. This itself could be a strong assumption, depending on the situation. Of course, now here comes all the additional assumptions that make our lives easier, but are worth being explicit about (just as discussed in our nonparametric regression notes):
  - Typically we think of x_i, $i = 1, \ldots, n$ as fixed, and that ϵ_i, $i = 1, \ldots, n$ are i.i.d.
  - This is is equivalent to conditioning on x_i, $i = 1, \ldots, n$, and then assuming that these are independent of $\epsilon_i = y_i - \mathbb{E}(y_i|x_i)$, $i = 1, \ldots, n$.
  - These are strong assumptions. They preclude, e.g., heteroskedasticity, omitted variable bias, etc.
  - Even on top of all this, we typically assume the distribution of the errors ϵ_i, $i = 1, \ldots, n$ to be Gaussian, or sub-Gaussian.
- It certainly does sound like we assume a lot, but we're also going to consider a very difficult problem: *high-dimensional* regression, where the dimension p of the predictors is comparable or even possibly much larger than the sample size n! Also, some of these assumptions can be relaxed (even the independence of X, ϵ), at the expense of a more compicated analysis

1.2 The risk of least squares

- Let's remind ourselves of the risk properties of least squares regression. Let $X_1, \ldots, X_p \in \mathbb{R}^n$ be the columns of the predictor matrix X. The least squares coefficients can be defined as the solution of the optimization problem

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (y_i - x_i^T \beta)^2 \iff \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \Big(y_i - \sum_{j=1}^p \beta_j X_j \Big)^2 \iff \min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2. \tag{3}$$

If $\text{rank}(X) = p$, i.e, the predictors $X_1, \ldots, X_p$ are linearly independent, then the above least squares problem has a unique solution, which (as you of course must have memorized by now) is $\hat{\beta} = (X^T X)^{-1} X^T y$

- The fitted values are $X\hat{\beta} = P_X y$, where $P_X = X(X^T X)^{-1} X^T$ is the projection matrix onto the column space of X. These are the predictions at the sample points x_i, $i = 1, \ldots, n$. To make a prediction new point $x_0 \in \mathbb{R}^p$, we would use $x_0^T \hat{\beta}$

- It is not hard to see that such least squares predictions are unbiased. Given x_0, the point at which we want to make a prediction, we can condition on X, x_0, we compute:

$$\mathbb{E}(x_0^T \hat{\beta} \,|\, X, x_0) = x_0^T (X^T X)^{-1} X^T \mathbb{E}(y|X) = x_0^T \beta_0.$$

  Hence the bias will still be zero after integrating out over X, x_0. Note that this unbiasedness doesn't actually require the strong assumption of X, ϵ being independent

- The *in-sample risk* or simply *risk* of the least squares estimator is defined as

$$\frac{1}{n}\mathbb{E}\|X\hat{\beta} - X\beta_0\|_2^2 = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(x_i^T \hat{\beta} - x_i^T \beta_0)^2 = \mathbb{E}(x_1^T \hat{\beta} - x_1^T \beta_0)^2,$$

  where X recall has rows x_i, $i = 1, \ldots, n$ and β_0 are the underlying regression coefficients as in (1), (2). The expectation here is over the randomness in the i.i.d. pairs (x_i, y_i), $i = 1, \ldots, n$, and we will we assume that X, ϵ are independent, as well as $\epsilon \sim N(0, \sigma^2 I)$. To compute it, as usual, we condition on X:

$$\frac{1}{n}\mathbb{E}\big(\|X\hat{\beta} - X\beta_0\|_2^2 \,\big|\, X\big) = \frac{1}{n}\text{tr}\big(\text{Cov}(X\hat{\beta} \,|\, X)\big) = \frac{1}{n}\text{tr}(\sigma^2 P_X) = \sigma^2 \frac{p}{n}.$$

  Therefore, integrating out over X, we get that the in-sample risk is again

$$\frac{1}{n}\mathbb{E}\|X\hat{\beta} - X\beta_0\|_2^2 = \sigma^2 \frac{p}{n}$$

- The *out-of-sample risk* or *predictive risk* of the least squares estimator is defined as

$$\mathbb{E}(x_0^T \hat{\beta} - x_0^T \beta_0)^2,$$

  where x_0 is a new independent draw from the predictor distribution. To compute it, we again condition on X, x_0:

$$\mathbb{E}(x_0^T \hat{\beta} - x_0^T \beta_0 \,|\, X, x_0)^2 = \text{Var}(x_0^T \hat{\beta} \,|\, X, x_0) = \sigma^2 x_0^T (X^T X)^{-1} x_0,$$

  then integrating out over X, x_0:

$$\mathbb{E}(x_0^T \hat{\beta} - x_0^T \beta_0)^2 = \sigma^2 \mathbb{E}\big[\text{tr}\big(x_0 x_0^T (X^T X)^{-1}\big)\big] = \sigma^2 \text{tr}\Big(\mathbb{E}(x_0 x_0^T)\mathbb{E}\big[(X^T X)^{-1}\big]\Big),$$

  where we have used the independence of X, x_0. An exact formula will not be possible in full generality here, since as we can see the out-of-sample risk depends on the distribution of the predictors. Contrast this with the in-sample risk, which did not

- In general, as shown in Groves & Rothenberg (1969), $\mathbb{E}[(X^T X)^{-1}] - [\mathbb{E}(X^T X)]^{-1}$ is positive semidefinite, so writing Σ for the covariance of the predictor distribution,

$$\mathbb{E}(x_0^T \hat{\beta} - x_0^T \beta_0)^2 = \sigma^2 \text{tr}\Big(\mathbb{E}(x_0 x_0^T)\mathbb{E}\big[(X^T X)^{-1}\big]\Big) \geq \sigma^2 \text{tr}\left(\Sigma \frac{\Sigma^{-1}}{n}\right) = \sigma^2 \frac{p}{n}.$$

  Thus, out-of-sample risk is always larger than in-sample risk, which makes sense, since intuitively, actual (out-of-sample) prediction is harder

- When the predictor distribution is, e.g., $N(0, \Sigma)$, we can compute the out-of-sample risk exactly. It holds that $X^T X \sim W(\Sigma, n)$, a Wishart distribution, and $(X^T X)^{-1} \sim W^{-1}(\Sigma^{-1}, n)$, an inverse Wishart distribution, so

$$\mathbb{E}(x_0^T \hat{\beta} - x_0^T \beta_0)^2 = \sigma^2 \mathrm{tr}\bigg(\Sigma \frac{\Sigma^{-1}}{n-p-1}\bigg) = \sigma^2 \frac{p}{n-p-1}$$

1.3 The failure of least squares in high dimensions

- When $\mathrm{rank}(X) < p$, e.g., this happens when $p > n$, there are infinitely many solutions in the least squares problem (3). Given one solution $\hat{\beta}$, the quantity $\hat{\beta} + \eta$ is also a solution for any $\eta \in \mathrm{null}(X)$. Furthermore, this type of nonuniqueness makes interpretation of solutions meaningless: for at least one $j \in \{1, \ldots, p\}$, we will have $\hat{\beta}_j > 0$ at one solution $\hat{\beta}$, but $\tilde{\beta}_j < 0$ at another solution $\tilde{\beta}$

- The fitted values from least squares regression are always unique; that is, $X\hat{\beta} = X\tilde{\beta}$, for any two solutions $\hat{\beta}, \tilde{\beta}$, no matter the column rank of X. This is because we can always write the fitted values as $P_X y$, where P_X is the projection matrix onto the column space of X; recall $P_X = X(X^T X)^+ X^T$ where $(X^T X)^+$ is the pseudoinverse of $X^T X$ (and the projection matrix to $P_X = X(X^T X)^{-1} X^T$ when X has full column rank)

- But in terms of actual predictions, at say a new point $x_0 \in \mathbb{R}^p$, it will not generally be the case that $x_0^T \hat{\beta} = x_0^T \tilde{\beta}$ for two solutions $\hat{\beta}, \tilde{\beta}$ (because the solutions need not be equal)

- So both interpretation and actual predictions are impossible with least squares when $p > n$, which is a pretty serious failure

- Even when $\mathrm{rank}(X) = p$, so that a unique least squares solution exists, we still may not want to use least squares if p is moderately close to n, because its risk could be quite poor (i.e., $\sigma^2 p/n$ in-sample risk, which will be poor if p is an appreciable fraction of n

- How do we deal with such issues? The short answer is *regularization*. In our present setting, we would modify the least squares estimator in one of two forms:

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \;\; \text{subject to} \;\; \beta \in C \qquad \text{(Constrained form)}$$

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + P(\beta) \qquad \text{(Penalized form)}$$

  where C is some (typically convex) set, and $P(\cdot)$ is some (typically convex) penalty function

- At its core, regularization provides us with a way of navigating the bias-variance tradeoff: we (hopefully greatly) reduce the variance at the expense of introducing some bias

1.4 What we cover here

- The goal is to introduce you to some important developments in methodology and theory in high-dimensional regression. Perhaps biasedly, we will focus on the lasso and related methods. High-dimensional statistics is both an enormous and enormously fast-paced field, so of course we will have to leave a lot out. E.g., a lot of what we say carries over in some way to high-dimensional generalized linear models, but we will not discuss these problems

- There are several great books on high-dimensional estimation, and here are a few:
  - Great general reference: Hastie, Tibshirani & Wainwright (2015)
  - Great theoretical references: Buhlmann & van de Geer (2011), Wainwright (2017)

2 Best subset selection, ridge regression, and the lasso

2.1 Three norms: ℓ_0, ℓ_1, ℓ_2

- In terms of regularization, we typically choose the constraint set C to be a sublevel set of a norm (or seminorm), and equivalently, the penalty function $P(\cdot)$ to be a multiple of a norm (or seminorm)

- Let's consider three canonical choices: the ℓ_0, ℓ_1, and ℓ_2 norms:

$$\|\beta\|_0 = \sum_{j=1}^p 1\{\beta_j \neq 0\}, \quad \|\beta\|_1 = \sum_{j=1}^p |\beta_j|, \quad \|\beta\|_2 = \bigg(\sum_{j=1}^p \beta_j^2\bigg)^{1/2}.$$

  (Truthfully, calling it "the ℓ_0 norm" is a misnomer, since it is not a norm: it does not satisfy positive homogeneity, i.e., $\|a\beta\|_0 \neq a\|\beta\|_0$ whenever $a \neq 0, 1$.)

- In constrained form, this gives rise to the problems:

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \;\text{ subject to }\; \|\beta\|_0 \leq k \qquad \text{(Best subset selection)} \tag{4}$$

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \;\text{ subject to }\; \|\beta\|_1 \leq t \qquad \text{(Lasso regression)} \tag{5}$$

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 \;\text{ subject to }\; \|\beta\|_2^2 \leq t \qquad \text{(Ridge regession)} \tag{6}$$

  where $k, t \geq 0$ are tuning parameters. Note that it makes sense to restrict k to be an integer; in best subset selection, we are quite literally finding the best subset of variables of size k, in terms of the achieved training error

- Though it is likely the case that these ideas were around earlier in other contexts, in statistics we typically subset selection to Beale et al. (1967), Hocking & Leslie (1967), ridge regression to Hoerl & Kennard (1970), and the lasso to Tibshirani (1996), Chen et al. (1998)

- In penalized form, the use of ℓ_0, ℓ_1, ℓ_2 norms gives rise to the problems:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_0 \qquad \text{(Best subset selection)} \tag{7}$$

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \qquad \text{(Lasso regression)} \tag{8}$$

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \qquad \text{(Ridge regression)} \tag{9}$$

  with $\lambda \geq 0$ the tuning parameter. In fact, problems (5), (8) are equivalent. By this, we mean that for any $t \geq 0$ and solution $\hat{\beta}$ in (5), there is a value of $\lambda \geq 0$ such that $\hat{\beta}$ also solves (8), and vice versa. The same equivalence holds for (6), (9). (The factors of $1/2$ multiplying the squared loss above are inconsequential, and just for convenience)

- It means, roughly speaking, that computing solutions of (5) over a sequence of t values and performing cross-validation (to select an estimate) should be basically the same as computing solutions of (8) over some sequence of λ values and performing cross-validation (to select an estimate). Strictly speaking, this isn't quite true, because the precise correspondence between equivalent t, λ depends on the data X, y

- Notably, problems (4), (7) are *not equivalent*. For every value of $\lambda \geq 0$ and solution $\hat{\beta}$ in (7), there is a value of $t \geq 0$ such that $\hat{\beta}$ also solves (4), but the converse is not true

2.2 One of these problems is not like the others: sparsity

- The best subset selection and the lasso estimators have a special, useful property: their solutions are *sparse*, i.e., at a solution $\hat{\beta}$ we will have $\hat{\beta}_j = 0$ for many components $j \in \{1, \ldots, p\}$. In problem (4), this is obviously true, where $k \geq 0$ controls the sparsity level. In problem (5), it is less obviously true, but we get a higher degree of sparsity the smaller the value of $t \geq 0$. In the penalized forms, (7), (8), we get more sparsity the larger the value of $\lambda \geq 0$

- This is not true of ridge regression, i.e., the solution of (6) or (9) generically has all nonzero components, no matter the value of t or λ. Note that sparsity is desirable, for two reasons: (i) it corresponds to performing variable selection in the constructed linear model, and (ii) it provides a level of interpretability (beyond sheer accuracy)

- That the ℓ_0 norm induces sparsity is obvious. But, why does the ℓ_1 norm induce sparsity and not the ℓ_2 norm? There are different ways to look at it; let's stick with intuition from the constrained problem forms (5), (8). Figure 1 shows the "classic" picture, contrasting the way the contours of the squared error loss hit the two constraint sets, the ℓ_1 and ℓ_2 balls. As the ℓ_1 ball has sharp corners (aligned with the coordinate axes), we get sparse solutions

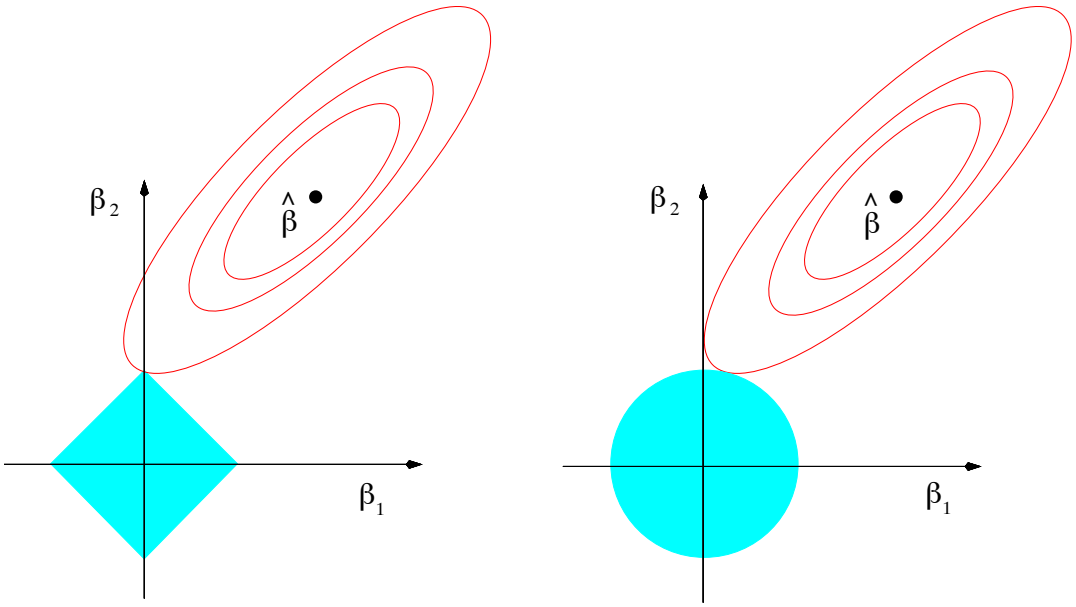

Figure 1: *The "classic" illustration comparing lasso and ridge constraints. From Chapter 3 of Hastie et al. (2009)*

- Intuition can also be drawn from the orthogonal case. When X is orthogonal, it is not hard to show that the solutions of the penalized problems (7), (8), (9) are

$$\hat{\beta}^{\text{subset}} = H_{\sqrt{2\lambda}}(X^T y), \quad \hat{\beta}^{\text{lasso}} = S_\lambda(X^T y), \quad \hat{\beta}^{\text{ridge}} = \frac{X^T y}{1 + 2\lambda}$$

  respectively, where $H_t(\cdot), S_t(\cdot)$ are the componentwise hard- and soft-thresholding functions at the level t. We see several revealing properties: subset selection and lasso solutions exhibit sparsity when the componentwise least squares coefficients (inner products $X^T y$) are small enough; the lasso solution exihibits shrinkage, in that large enough least squares coefficients

are shrunken towards zero by λ; the ridge regression solution is never sparse and compared to the lasso, preferentially shrinkage the larger least squares coefficients even more

2.3 One of these problems is not like the others: convexity

- The lasso and ridge regression problems (5), (6) have another very important property: they are convex optimization problems. Best subset selection (4) is not, in fact it is very far from being convex
- It is convexity that allows to equate (5), (8), and (6), (9) (and yes, the penalized forms are convex problems too). It is also convexity that allows us to both efficiently solve, and in some sense, precisely understand the nature of the lasso and ridge regression solutions
- Here is a (far too quick) refresher/introduction to basic convex analysis and convex optimization. Recall that a set $C \subseteq \mathbb{R}^n$ is called *convex* if for any $x, y \in C$ and $t \in [0, 1]$, we have

$$tx + (1-t)y \in C,$$

i.e., the line segment joining x, y lies entirely in C. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *convex* if its domain $\mathrm{dom}(f)$ is convex, and for any $x, y \in \mathrm{dom}(f)$ and $t \in [0, 1]$,

$$f\big(tx + (1-t)y\big) \leq tf(x) + (1-t)f(y),$$

i.e., the function lies below the line segment joining its evaluations at x and y. A function is called *strictly convex* if this same inequality holds strictly for $x \neq y$ and $t \in (0, 1)$

- E.g., lines, rays, line segments, linear spaces, affine spaces, hyperplans, halfspaces, polyhedra, norm balls are all convex sets
- E.g., affine functions $a^T x + b$ are convex and concave, quadratic functions $x^T Q x + b^T x + c$ are convex if $Q \succeq 0$ and strictly convex if $Q \succ 0$, norms are convex
- Formally, an *optimization problem* is of the form

$$\begin{aligned} \min_{x \in D} \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0, \;\; i = 1, \ldots m \\ & \ell_j(x) = 0, \;\; j = 1, \ldots r \end{aligned}$$

Here $D = \mathrm{dom}(f) \cap \bigcap_{i=1}^m \mathrm{dom}(h_i) \cap \bigcap_{j=1}^r \mathrm{dom}(\ell_j)$ is the common domain of all functions. A *convex optimization problem* is an optimization problem in which all functions $f, h_1, \ldots h_m$ are convex, and all functions $\ell_1, \ldots \ell_r$ are affine. (Think: why affine?) Hence, we can express it as

$$\begin{aligned} \min_{x \in D} \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0, \;\; i = 1, \ldots m \\ & Ax = b \end{aligned}$$

- Why is a convex optimization problem so special? The short answer: because *any local minimizer is a global minimizer*. To see this, suppose that x is feasible for the convex problem formulation above and there exists some $R > 0$ such that

$$f(x) \leq f(y) \quad \text{for all feasible } y \text{ with } \|x - y\|_2 \leq R.$$

Such a point x is called a local minimizer. For the sake of contradiction, suppose that x was not a global minimizer, i.e., there exists some feasible z such that $f(z) < f(x)$. By convexity

of the constraints (and the domain D), the point $tz + (1-t)x$ is feasible for any $0 \leq t \leq 1$. Furthermore, by convexity of f,

$$f\big(tz + (1-t)x\big) \leq tf(z) + (1-t)f(x) < f(x)$$

for any $0 < t < 1$. Lastly, we can choose $t > 0$ small enough so that $\|x - (tz + (1-t)x)\|_2 = t\|x - z\|_2 \leq R$, and we obtain a contradiction

- Algorithmically, this is a very useful property, because it means if we keep "going downhill", i.e., reducing the achieved criterion value, and we stop when we can't do so anymore, then we've hit the global solution

- Convex optimization problems are also special because they come with a beautiful theory of beautiful convex duality and optimality, which gives us a way of understanding the solutions. We won't have time to cover any of this, but we'll mention what subgradient optimality looks like for the lasso

- Just based on the definitions, it is not hard to see that (5), (6), (8), (9) are convex problems, but (4), (7) are not. In fact, the latter two problems are known to be NP-hard, so they are in a sense even the worst kind of nonconvex problem

2.4 Some theoretical backing for subset selection

- Despite its computational intractability, best subset selection has some attractive risk properties. A classic result is due to Foster & George (1994), on the in-sample risk of best subset selection in penalized form (7), which we will paraphrase here. First, we raise a very simple point: if A denotes the support (also called the active set) of the subset selection solution $\hat{\beta}$ in (7)—meaning that $\hat{\beta}_j = 0$ for all $j \notin A$, and denoted $A = \mathrm{supp}(\hat{\beta})$—then we have

$$\begin{aligned} \hat{\beta}_A &= (X_A^T X_A)^{-1} X_A^T y, \\ \hat{\beta}_{-A} &= 0. \end{aligned} \tag{10}$$

  Here and throughout we write X_A for the columns of matrix X in a set A, and x_A for the components of a vector x in A. We will also use X_{-A} and x_{-A} for the columns or components not in A. The observation in (10) follows from the fact that, given the support set A, the ℓ_0 penalty term in the subset selection criterion doesn't depend on the actual magnitudes of the coefficients (it contributes a constant factor), so the problem reduces to least squares

- Now, consider a standard linear model as in (2), with X fixed, and $\epsilon \sim N(0, \sigma^2 I)$. Suppose that the underlying coefficients have support $S = \mathrm{supp}(\beta_0)$, and $s_0 = |S|$. Then, the estimator given by least squares on S, i.e.,

$$\begin{aligned} \hat{\beta}_S^{\mathrm{oracle}} &= (X_S^T X_S)^{-1} X_S^T y, \\ \hat{\beta}_{-S}^{\mathrm{oracle}} &= 0. \end{aligned}$$

  is is called *oracle estimator*, and as we know from our previous calculations, has in-sample risk

$$\frac{1}{n}\|X\hat{\beta}^{\mathrm{oracle}} - X\beta_0\|_2^2 = \sigma^2 \frac{s_0}{n}$$

- Foster & George (1994) consider this setup, and compare the risk of the best subset selection estimator $\hat{\beta}$ in (7) to the oracle risk of $\sigma^2 s_0/n$. They show that, if we choose $\lambda \asymp \sigma^2 \log p$, then the best subset selection estimator satisfies

$$\frac{\mathbb{E}\|X\hat{\beta} - X\beta_0\|_2^2/n}{\sigma^2 s_0/n} \leq 4 \log p + 2 + o(1), \tag{11}$$

as $n, p \to \infty$. This holds without any conditions on the predictor matrix X. Moreover, they prove the lower bound

$$\inf_{\hat{\beta}} \sup_{X, \beta_0} \frac{\mathbb{E}\|X\hat{\beta} - X\beta_0\|_2^2/n}{\sigma^2 s_0/n} \geq 2\log p - o(\log p),$$

where the infimum is over all estimators $\hat{\beta}$, and the supremum is over all predictor matrices X and underlying coefficients with $\|\beta_0\|_0 = s_0$. Hence, in terms of rate, best subset selection achieves the optimal risk inflation over the oracle risk

- Returning to what was said above, the kicker is that we can't really compute the best subset selection estimator for even moderately-sized problems. As we will in the following, the lasso provides a similar risk inflation guarantee, though under considerably stronger assumptions

- Lastly, it is worth remarking that even if we *could* compute the subset selection estimator at scale, it's not at all clear that we would want to use this in place of the lasso. (Many people assume that we would.) We must remind ourselves that theory provides us an understanding of the performance of various estimators under typically idealized conditions, and it doesn't tell the complete story. It could be the case that the lack of shrinkage in the subset selection coefficients ends up being harmful in practical situations, in a signal-to-noise regime, and yet the lasso could still perform favorably in such settings

- **Update.** Some nice recent work in optimization (Bertsimas et al. 2016) shows that we can cast best subset selection as a mixed integer quadratic program, and proposes to solve it (in general this means approximately, though with a certified bound on the duality gap) with an industry-standard mixed integer optimization package like Gurobi. If we have time, we'll discuss this at the end and make some comparisons between subset selection and the lasso

3 Basic properties and geometry of the lasso

3.1 Ridge regression and the elastic net

- A quick refresher: the ridge regression problem (9) is always strictly convex (assuming $\lambda > 0$), due to the presense of the squared ℓ_2 penalty $\|\beta\|_2^2$. To be clear, this is true regardless of X, and so the ridge regression solution is always well-defined, and is in fact given in closed-form by $\hat{\beta} = (X^T X + 2\lambda I)^{-1} X^T y$

- In contrast, the lasso problem is not always strictly convex and hence by standard convexity theory, it need not have a unique solution (more on this shortly). However, we can define a modified probblem that it always strictly convex, via the *elastic net* (Zou & Hastie 2005):

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 + \delta\|\beta\|_2^2, \tag{12}$$

where now both $\lambda, \delta \geq 0$ are tuning parameters. Aside from guaranteeing uniqueness for all X, the elastic net combines some of the desirable predictive properties of ridge regression with the sparsity properties of the lasso

3.2 Nonuniqueness, sign patterns, and active sets

- A few basic observations on the lasso problem in (8):

  1. There need not always be a unique solution $\hat{\beta}$ in (8), because the criterion is not strictly convex when $X^T X$ is singular (which, e.g., happens when $p > n$).

2. There is however always a unique fitted value $X\hat{\beta}$ in (8), because the squared error loss is strictly convex in $X\beta$.

The first observation is worrisome; of course, it would be very bad if we encountered the same problem with interpretation that we did in ordinary least squares. We will see shortly that there is really nothing to worry about. The second observation is standard (it is also true in least squares), but will be helpful

- Now we turn to subgradient optimality (sometimes called the KKT conditions) for the lasso problem in (8). They tell us that any lasso solution $\hat{\beta}$ must satisfy

$$X^T(y - X\hat{\beta}) = \lambda s, \tag{13}$$

where $s \in \partial\|\hat{\beta}\|_1$, a subgradient of the ℓ_1 norm evaluated at $\hat{\beta}$. Precisely, this means that

$$s_j \in \begin{cases} \{+1\} & \hat{\beta}_j > 0 \\ \{-1\} & \hat{\beta}_j < 0 \\ [-1, 1] & \hat{\beta}_j = 0, \end{cases} \quad j = 1, \ldots, p \tag{14}$$

- From (13) we can read off a straightforward but important fact: even though the solution $\hat{\beta}$ may not be uniquely determined, the optimal subgradient s is a function of the unique fitted value $X\hat{\beta}$ (assuming $\lambda > 0$), and hence is itself unique

- Now from (14), note that the uniqueness of s implies that any two lasso solutions must have the same signs on the overlap of their supports. That is, it cannot happen that we find two different lasso solutions $\hat{\beta}$ and $\tilde{\beta}$ with $\hat{\beta}_j > 0$ but $\tilde{\beta}_j < 0$ for some j, and hence we have no problem interpretating the signs of components of lasso solutions

- Aside from possible interpretation issues, recall, nonuniqueness also means that actual (out-of-sample) prediction is not well-defined, which is also a big deal. In the next subsection, we'll see we also don't have to worry about this, for almost all lasso problems we might consider

- Before this, let's discuss active sets of lasso solutions. Define the *equicorrelation set*

$$E = \big\{j \in \{1, \ldots, p\} : |X_j^T(y - X\hat{\beta})| = \lambda\big\}.$$

This is the set of variables that achieves the maximum absolute inner product (or, correlation for standard predictors) with the lasso residual vector. Assuming $\lambda > 0$, this is the same as

$$E = \big\{j \in \{1, \ldots, p\} : |s_j| = 1\big\}.$$

This is a uniquely determined set (since $X\hat{\beta}, s$ are unique)

- Importantly, the set E contains the active set $A = \mathrm{supp}(\hat{\beta})$ of any lasso solution $\hat{\beta}$, because for $j \notin E$, we have $|s_j| < 1$, which implies that $\hat{\beta}_j = 0$

- Also importantly, the set E is the active set of a particular lasso solution, namely, the lasso solution with the smallest ℓ_2 norm, call it $\hat{\beta}^{\mathrm{lasso},\ell_2}$. The lasso solution with the smallest ℓ_2 norm is (perhaps not surprisingly) on the limiting end of the elastic net solution path (12) as the ridge penalty parameter goes to 0:

$$\hat{\beta}^{\mathrm{lasso},\ell_2} = \lim_{\delta \to 0} \left\{ \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 + \delta\|\beta\|_2^2 \right\}$$

3.3 Uniqueness and saturation

- Fortunately, the lasso solution in (8) is unique under very general conditions, specifically it is unique if X has columns in *general position* (Tibshirani 2013). We say that $X_1, \ldots, X_p \in \mathbb{R}^n$ are in general position provided that for any $k < \min\{n, p\}$, indices $i_1, \ldots, i_{k+1} \in \{1, \ldots p\}$, and signs $\sigma_1, \ldots, \sigma_{k+1} \in \{-1, +1\}$, the affine span of $\sigma_1 X_{i_1}, \ldots, \sigma_{k+1} X_{i_{k+1}}$ does not contain any element of $\{\pm X_i : i \neq i_1, \ldots, i_{k+1}\}$. This is equivalent to the following statement: no k-dimensional subspace $L \subseteq \mathbb{R}^n$, for $k < \min\{n, p\}$, contains more that $k+1$ points of $\{\pm X_1, \ldots \pm X_p\}$, excluding antipodal pairs (i.e., $+X_i$ and $-X_i$)

- This is a very weak condition on X, and it can hold no matter the (relative) sizes of n and p. It is straightforward to show that if the elements X_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$ have any continuous joint distribution (i.e., that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{np}$), then X has columns in general position almost surely

- Moreover, general position of X implies the following fact: for any $\lambda > 0$, the submatrix X_A of active predictor variables always has full column rank. This means that $|A| \leq \min\{n, p\}$, or in words, no matter where we are on the regularization path, the (unique) lasso solution never has more than $\min\{n, p\}$ nonzero components

- The above property is called *saturation* of the lasso solution, which is not necessarily a good property. If we have, e.g., 100,000 continuously distributed variables and 100 samples, then we will never form a working linear model with more than 100 selected variables with the lasso

- Note that the elastic net (12) was proposed as a means of overcoming this saturation problem (it does not have the same property); it also has a grouping effect, where it tends to pull in variables with similar effects into the active set together

3.4 Form of solutions

- Let's assume henceforth that the columns of X are in general position (and we are looking at a nontrivial end of the path, with $\lambda > 0$), so the lasso solution $\hat{\beta}$ is unique. Let $A = \mathrm{supp}(\hat{\beta})$ be the lasso active set, and let $s_A = \mathrm{sign}(\hat{\beta}_A)$ be the signs of active coefficients. From the subgradient conditions (13), (14), we know that

$$X_A^T(y - X_A\hat{\beta}_A) = \lambda s_A,$$

and solving for $\hat{\beta}_A$ gives

$$\begin{aligned} \hat{\beta}_A &= (X_A^T X_A)^{-1}(X_A^T y - \lambda s_A), \\ \hat{\beta}_{-A} &= 0 \end{aligned} \tag{15}$$

(where recall we know that $X_A^T X_A$ is invertible because X has columns in general position). We see that the active coefficients $\hat{\beta}_A$ are given by taking the least squares coefficients on X_A, $(X_A^T X_A)^{-1} X_A^T y$, and shrinking them by an amount $\lambda(X_A^T X_A)^{-1} s_A$. Contrast this to, e.g., the subset selection solution in (10), where there is no such shrinkage

- Now, how about this so-called shrinkage term $(X_A^T X_A)^{-1} X_A^T y$? Does it always act by moving each one of the least squares coefficients $(X_A^T X_A)^{-1} X_A^T y$ towards zero? Indeed, this is not always the case, and one can find empirical examples where a lasso coefficient is actually larger (in magnitude) than the corresponding least squares coefficient on the active set. Of course, we also know that this is due to the correlations between active variables, because when X is orthogonal, as we've already seen, this never happens

- On the other hand, it is always the case that the lasso solution has a strictly smaller ℓ_1 norm than the least squares solution on the active set, and in this sense, we are (perhaps) justified in always referring to $(X_A^T X_A)^{-1} X_A^T y$ as a shrinkage term. We can see this as

$$\|\hat{\beta}\|_1 = s_A^T (X_A^T X_A)^{-1} X_A^T y - \lambda s_A^T (X_A^T X_A)^{-1} s_A < \|(X_A^T X_A)^{-1} X_A^T y\|_1.$$

  The first term is less than or equal to $\|(X_A^T X_A)^{-1} X_A^T y\|_1$, and the term we are subtracting is strictly negative (because $(X_A^T X_A)^{-1}$ is positive definite)

3.5 Geometry of solutions

- One undesirable feature of the best subset selection solution (10) is the fact that it behaves discontinuously with y. As we change y, the active set A must change at some point, and the coefficients will jump discontinuously, because we are just doing least squares onto the active set

- So, does the same thing happen with the lasso solution (15)? The answer it not immediately clear. Again, as we change y, the active set A must change at some point; but if the shrinkage term were defined "just right", then perhaps the coefficients of variables to leave the active set would gracefully and continously drop to zero, and coefficients of variables to enter the active set would continuously move form zero. This would make whole the lasso solution continuous

- Fortuitously, this is indeed the case, and the lasso solution $\hat{\beta}$ is continuous as a function of y. It might seem a daunting task to prove this, but a certain perspective using convex geometry provides a very simple proof. The geometric perspective in fact proves that the lasso fit $X\hat{\beta}$ is nonexpansive in y, i.e., 1-Lipschitz continuous, which is a very strong form of continuity

- Define the convex polyhedron $C = \{u : \|X^T u\|_\infty \leq \lambda\} \subseteq \mathbb{R}^n$. Some simple manipulations of the KKT conditions show that the lasso fit is given by

$$X\hat{\beta} = (I - P_C)(y),$$

  the residual from projecting y onto C. A picture to show this (just look at the left panel for now) is given in Figure 2

- The projection onto any convex set is nonexpansive, i.e., $\|P_C(y) - P_C(y')\|_2 \leq \|y - y'\|_2$ for any y, y'. This should be visually clear from the picture. Actually, the same is true with the residual map: $I - P_C$ is also nonexpansive, and hence the lasso fit is 1-Lipschitz continuous

- Viewing the lasso fit as the residual from projection onto a convex polyhedron is actually an even more fruitful perspective. Write this polyhedron as

$$C = (X^T)^{-1}\{v : \|v\|_\infty \leq \lambda\},$$

  where $(X^T)^{-1}$ denotes the preimage operator under the linear map X^T. The set $\{v : \|v\|_\infty \leq \lambda\}$ is a hypercube in $\mathbb{R}^p$. Every face of this cube corresponds to a subset $A \subseteq \{1, \ldots p\}$ of dimensions (that achieve the maximum value $|\lambda|$) and signs $s_A \in \{-1, 1\}^{|A|}$ (that tell which side of the cube the face will lie on, for each dimension). Now, the faces of C are just faces of $\{v : \|v\|_\infty \leq \lambda\}$ run through the (linear) preimage transformation, so each face of C can also indexed by a set $A \subseteq \{1, \ldots p\}$ and signs $s_A \in \{-1, 1\}^{|A|}$. The picture in Figure 2 attempts to convey this relationship with the colored black face in each of the panels

- Now imagine projecting y onto C; it will land on some face. We have just argued that this face corresponds to a set A and signs s_A. One can show that this set A is exactly the active set of the lasso solution at y, and s_A are exactly the active signs. The size of the active set $|A|$ is the co-dimension of the face

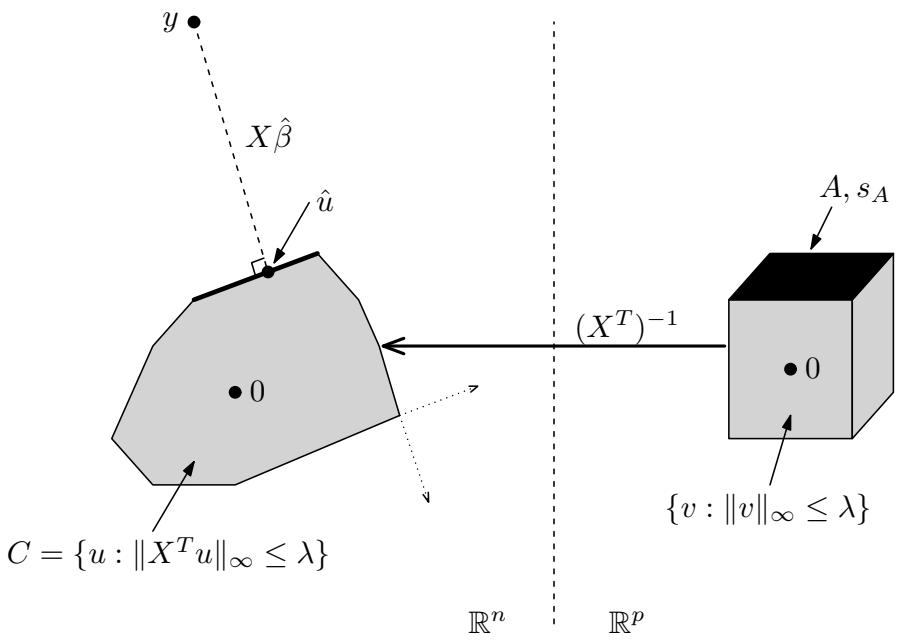

Figure 2: *A geometric picture of the lasso solution. The left panel shows the polyhedron underlying all lasso fits, where each face corresponds to a particular combination of active set A and signs s; the right panel displays the "inverse" polyhedron, where the dual solutions live*

- Looking at the picture: we can that see that as we wiggle y around, it will project to the same face. From the correspondence between faces and active set and signs of lasso solutions, this means that A, s_A do not change as we perturb y, i.e., they are locally constant
- But this isn't true for all points y, e.g., if y lies on one of the rays emanating from the lower right corner of the polyhedron in the picture, then we can see that small perturbations of y do actually change the face that it projects to, which invariably changes the active set and signs of the lasso solution. However, this is somewhat of an exceptional case, in that such points can be form a of Lebesgue measure zero, and therefore we can assure ourselves that the active set and signs A, s_A are locally constant for almost every y

3.6 Piecewise linear solution path

- From the lasso KKT conditions (13), (14), it is possible to compute the lasso solution in (8) as a function of λ, which we will write as $\hat{\beta}(\lambda)$, for all values of the tuning parameter $\lambda \in [0, \infty]$. This is called the *regularization path* or *solution path* of the problem (8)
- Path algorithms like the one we will describe below are not always possible; the reason that this ends up being feasible for the lasso problem (8) is that the solution path $\hat{\beta}(\lambda)$, $\lambda \in [0, \infty]$ turns out to be a piecewise linear, continuous function of λ. Hence, we only need to compute and store the *knots* in this path, which we will denote by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \geq 0$, and the lasso solution at these knots. From this information, we can then compute the lasso solution at any value of λ by linear interpolation
- The knots $\lambda_1 \geq \ldots \geq \lambda_r$ in the solution path correspond to λ values at which the active set

$A(\lambda) = \text{supp}(\hat{\beta}(\lambda))$ changes. As we decrease λ from ∞ to 0, the knots typically correspond to the point at which a variable enters the active set; this connects the lasso to an incremental variable selection procedure like forward stepwise regression. Interestingly though, as we decrease λ, a knot in the lasso path can also correspond to the point at which a variables leaves the active set. See Figure 3

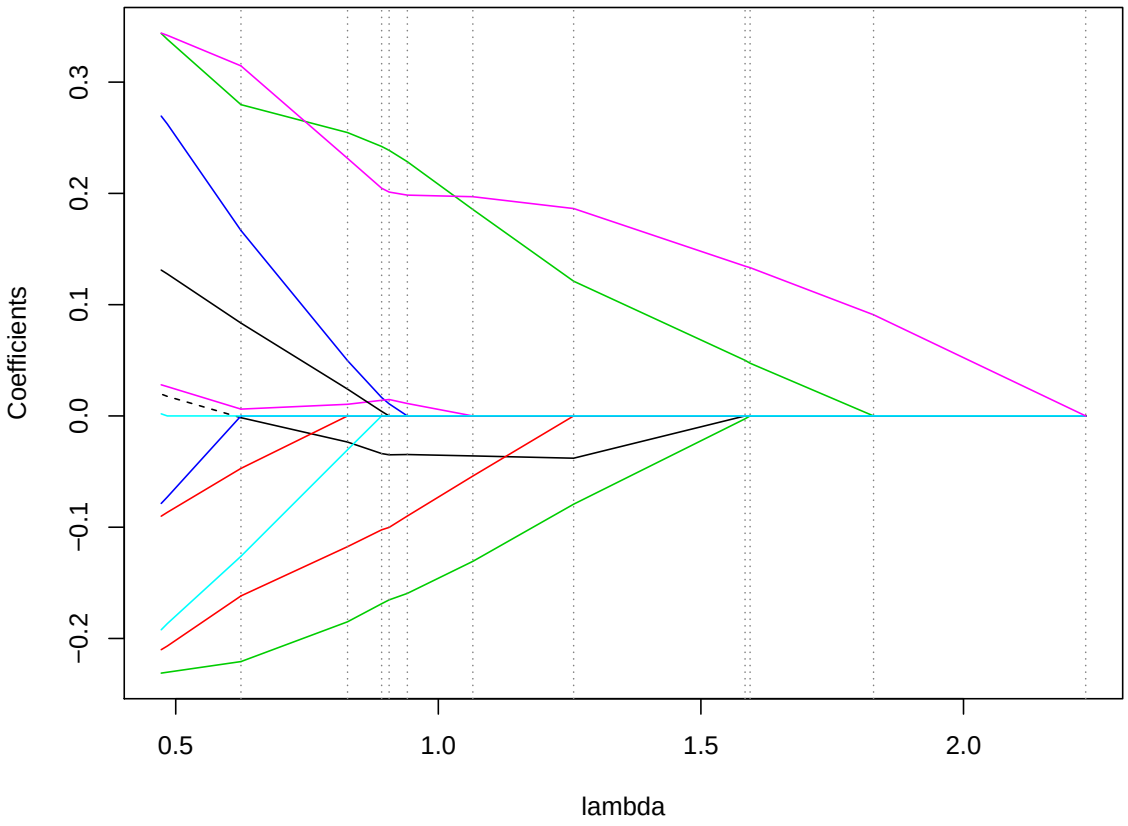

Figure 3: *An example of the lasso path. Each colored line denotes a component of the lasso solution* $\hat{\beta}_j(\lambda)$, $j = 1, \ldots, p$ *as a function of* λ. *The gray dotted vertical lines mark the knots* $\lambda_1 \geq \lambda_2 \geq \ldots$

- The lasso solution path was described by Osborne et al. (2000*a*,*b*), Efron et al. (2004). Like the construction of all other solution paths that followed these seminal works, the lasso path is essentially given by an iterative or inductive verification of the KKT conditions; if we can maintain that the KKT conditions holds as we decrease λ, then we know we have a solution. The trick is to start at a value of λ at which the solution is trivial; for the lasso, this is $\lambda = \infty$, at which case we know the solution must be $\hat{\beta}(\infty) = 0$

- Why would the path be piecewise linear? The construction of the path from the KKT conditions is actually rather technical (not difficult conceptually, but somewhat tedious), and doesn't shed insight onto this matter. But we can actually see it clearly from the projection picture in Figure 2

  As λ decreases from ∞ to 0, we are shrinking (by a multiplicative factor λ) the polyhedron onto which y is projected; let's write $C_\lambda = \{u : \|X^T u\|_\infty \leq \lambda\} = \lambda C_1$ to make this clear. Now suppose that y projects onto the relative interior of a certain face F of C_λ, corresponding to

an active set A and signs s_A. As λ decreases, the point on the boundary of C_λ onto which y projects, call it $\hat{u}(\lambda) = P_{C_\lambda}(y)$, will move along the face F, and change linearly in λ (because we are equivalently just tracking the projection of y onto an affine space that is being scaled by λ). Thus, the lasso fit $X\hat{\beta}(\lambda) = y - \hat{u}(\lambda)$ will also behave linearly in λ

Eventually, as we continue to decrease λ, the projected point $\hat{u}(\lambda)$ will move to the relative boundary of the face F; then, decreasing λ further, it will lie on a different, neighboring face F'. This face will correspond to an active set A' and signs $s_{A'}$ that (each) differ by only one element to A and s_A, respectively. It will then move linearly across F', and so on

- Now we will walk through the technical derivation of the lasso path, starting at $\lambda = \infty$ and $\hat{\beta}(\infty) = 0$, as indicated above. Consider decreasing λ from ∞, and continuing to set $\hat{\beta}(\lambda) = 0$ as the lasso solution. The KKT conditions (13) read

$$X^T y = \lambda s,$$

where s is a subgradient of the ℓ_1 norm evaluated at 0, i.e., $s_j \in [-1, 1]$ for every $j = 1, \ldots, p$. For large enough values of λ, this is satisfied, as we can choose $s = X^T y/\lambda$. But this ceases to be a valid subgradient if we decrease λ past the point at which $\lambda = |X_j^T y|$ for some variable $j = 1, \ldots, p$. In short, $\hat{\beta}(\lambda) = 0$ is the lasso solution for all $\lambda \geq \lambda_1$, where

$$\lambda_1 = \max_{j=1,\ldots,p} |X_j^T y|. \tag{16}$$

What happens next? As we decrease λ from λ_1, we know that we're going to have to change $\hat{\beta}(\lambda)$ from 0 so that the KKT conditions remain satisfied. Let j_1 denote the variable that achieves the maximum in (16). Since the subgradient was $|s_{j_1}| = 1$ at $\lambda = \lambda_1$, we see that we are "allowed" to make $\hat{\beta}_{j_1}(\lambda)$ nonzero. Consider setting

$$\begin{aligned} \hat{\beta}_{j_1}(\lambda) &= (X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1}) \\ \hat{\beta}_j(\lambda) &= 0, \quad \text{for all } j \neq j_1, \end{aligned} \tag{17}$$

as λ decreases from λ_1, where $s_{j_1} = \text{sign}(X_{j_1}^T y)$. Note that this makes $\hat{\beta}(\lambda)$ a piecewise linear and continuous function of λ, so far. The KKT conditions are then

$$X_{j_1}^T \Big(y - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1}) \Big) = \lambda s_{j_1},$$

which can be checked with simple algebra, and

$$\Big| X_j^T \Big(y - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1}) \Big) \Big| \leq \lambda,$$

for all $j \neq j_1$. Recall that the above held with strict inequality at $\lambda = \lambda_1$ for all $j \neq j_1$, and by continuity of the constructed solution $\hat{\beta}(\lambda)$, it should continue to hold as we decrease λ for at least a little while. In fact, it will hold until one of the piecewise linear paths

$$X_j^T (y - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1})), \quad j \neq j_1$$

becomes equal to $\pm\lambda$, at which point we have to modify the solution because otherwise the implicit subgradient

$$s_j = \frac{X_j^T (y - X_{j_1}(X_{j_1}^T X_{j_1})^{-1}(X_{j_1}^T y - \lambda s_{j_1}))}{\lambda}$$

will cease to be in $[-1, 1]$. It helps to draw yourself a picture of this

Thanks to linearity, we can compute the critical "hitting time" explicitly; a short calculation shows that, the lasso solution continues to be given by (17) for all $\lambda_1 \geq \lambda \geq \lambda_2$, where

$$\lambda_2 = \max_{j \neq j_1,\, s_j \in \{-1,1\}}{}^{+} \frac{X_j^T (I - X_{j_1}(X_{j_1}^T X_{j_1})^{-1} X_{j_1}) y}{s_j - X_j^T X_{j_1}(X_{j_1}^T X_{j_1})^{-1} s_{j_1}}, \tag{18}$$

and $\max^+$ denotes the maximum over all of its arguments that are $< \lambda_1$

To keep going: let j_2, s_2 achieve the maximum in (18). Let $A = \{j_1, j_2\}$, $s_A = (s_{j_1}, s_{j_2})$, and consider setting

$$\begin{aligned} \hat{\beta}_A(\lambda) &= (X_A^T X_A)^{-1}(X_A^T y - \lambda s_A) \\ \hat{\beta}_{-A}(\lambda) &= 0, \end{aligned} \tag{19}$$

as λ decreases from λ_2. Again, we can verify the KKT conditions for a stretch of decreasing λ, but will have to stop when one of

$$X_j^T (y - X_A(X_A^T X_A)^{-1}(X_A^T y - \lambda s_A), \quad j \notin A$$

becomes equal to $\pm\lambda$. By linearity, we can compute this next "hitting time" explicitly, just as before. Furthermore, though, we will have to check whether the active components of the computed solution in (19) are going to cross through zero, because past such a point, s_A will no longer be a proper subgradient over the active components. We can again compute this next "crossing time" explicitly, due to linearity. Therefore, we maintain that (19) is the lasso solution for all $\lambda_2 \geq \lambda \geq \lambda_3$, where λ_3 is the maximum of the next hitting time and the next crossing time. For convenience, the lasso path algorithm is summarized below

Algorithm 1 (**Lasso path algorithm**).

Given y and X.

- *Start with the iteration counter $k = 0$, regularization parameter $\lambda_0 = \infty$, active set $A = \emptyset$, and active signs $s_A = \emptyset$*
- *While $\lambda_k > 0$:*
  1. *Compute the lasso solution as λ decreases from λ_k by*

$$\begin{aligned} \hat{\beta}_A(\lambda) &= (X_A^T X_A)^{-1}(X_A^T y - \lambda s_A) \\ \hat{\beta}_{-A}(\lambda) &= 0 \end{aligned}$$

  2. *Compute the next hitting time (where $\max^+$ denotes the maximum of its arguments $< \lambda_k$),*

$$\lambda_{k+1}^{\text{hit}} = \max_{j \notin A,\, s_j \in \{-1,1\}}{}^{+} \frac{X_j^T (I - X_A(X_A^T X_A)^{-1} X_A^T) y}{s_j - X_j^T X_A (X_A^T X_A)^{-1} s_A}$$

  3. *Compute the next crossing time (where $\max^+$ denotes the maximum of its arguments $< \lambda_k$),*

$$\lambda_{k+1}^{\text{cross}} = \max_{j \in A}{}^{+} \frac{[(X_A^T X_A)^{-1} X_A^T y]_j}{[(X_A^T X_A)^{-1} s_A]_j},$$

  4. *Decrease λ until λ_{k+1}, defined by*

$$\lambda_{k+1} = \max\{\lambda_{k+1}^{\text{hit}}, \lambda_{k+1}^{\text{cross}}\}$$

  5. *If $\lambda_{k+1}^{\text{hit}} > \lambda_{k+1}^{\text{cross}}$, then add the hitting variable to A and its sign to s_A; otherwise, remove the crossing variable from A and its sign from s_A. Update $k = k + 1$*

- As we decrease λ from a knot λ_k, we can rewrite the lasso coefficient update in Step 1 as

$$\begin{aligned} \hat{\beta}_A(\lambda) &= \hat{\beta}_A(\lambda_k) + (\lambda_k - \lambda)(X_A^T X_A)^{-1} s_A, \\ \hat{\beta}_{-A}(\lambda) &= 0. \end{aligned} \tag{20}$$

  We can see that we are moving the active coefficients in the direction $(\lambda_k - \lambda)(X_A^T X_A)^{-1} s_A$ for decreasing λ. In other words, the lasso fitted values proceed as

$$X\hat{\beta}(\lambda) = X\hat{\beta}(\lambda_k) + (\lambda_k - \lambda) X_A (X_A^T X_A)^{-1} s_A,$$

  for decreasing λ. Efron et al. (2004) call $X_A(X_A^T X_A)^{-1} s_A$ the *equiangular direction*, because this direction, in $\mathbb{R}^n$, takes an equal angle with all $X_j \in \mathbb{R}^n$, $j \in A$

- For this reason, the lasso path algorithm in Algorithm 1 is also often referred to as the *least angle regression* path algorithm in "lasso mode", though we have not mentioned this yet to avoid confusion. Least angle regression is considered as another algorithm by itself, where we skip Step 3 altogether. In words, Step 3 disallows any component path to cross through zero. The left side of the plot in Figure 3 visualizes the distinction between least angle regression and lasso estimates: the dotted black line displays the least angle regression component path, crossing through zero, while the lasso component path remains at zero

- Lastly, an alternative expression for the coefficient update in (20) (the update in Step 1) is

$$\begin{aligned} \hat{\beta}_A(\lambda) &= \hat{\beta}_A(\lambda_k) + \frac{\lambda_k - \lambda}{\lambda_k} (X_A^T X_A)^{-1} X_A^T r(\lambda_k), \\ \hat{\beta}_{-A}(\lambda) &= 0, \end{aligned} \tag{21}$$

  where $r(\lambda_k) = y - X_A \hat{\beta}_A(\lambda_k)$ is the residual (from the fitted lasso model) at λ_k. This follows because, recall, $\lambda_k s_A$ are simply the inner products of the active variables with the residual at λ_k, i.e., $\lambda_k s_A = X_A^T(y - X_A \hat{\beta}_A(\lambda_k))$. In words, we can see that the update for the active lasso coefficients in (21) is in the direction of the least squares coefficients of the residual $r(\lambda_k)$ on the active variables X_A

4 Theoretical analysis of the lasso

4.1 Slow rates

- Recently, there has been an enormous amount theoretical work analyzing the performance of the lasso. Some references (warning: a highly incomplete list) are Greenshtein & Ritov (2004), Fuchs (2005), Donoho (2006), Candes & Tao (2006), Meinshausen & Buhlmann (2006), Zhao & Yu (2006), Candes & Plan (2009), Wainwright (2009); a helpful text for these kind of results is Buhlmann & van de Geer (2011)

- We begin by stating what are called *slow rates* for the lasso estimator. Most of the proofs are simple enough that they are given below. These results don't place any real assumptions on the predictor matrix X, but deliver slow(er) rates for the risk of the lasso estimator than what we would get under more assumptions, hence their name

- We will assume the standard linear model (2), with X fixed, and $\epsilon \sim N(0, \sigma^2)$. We will also assume that $\|X_j\|_2^2 \leq n$, for $j = 1, \ldots, p$. That the errors are Gaussian can be easily relaxed to sub-Gaussianity. That X is fixed (or equivalently, it is random, but we condition on it and assume it is independent of ϵ) is more difficult to relax, but can be done as in Greenshtein & Ritov (2004). This makes the proofs more complicated, so we don't consider it here

- **Bound form.** The lasso estimator in bound form (5) is particularly easy to analyze. Suppose that we choose $t = \|\beta_0\|_1$ as the tuning parameter. Then, simply by virtue of optimality of the solution $\hat{\beta}$ in (5), we find that

$$\|y - X\hat{\beta}\|_2^2 \leq \|y - X\beta_0\|_2^2,$$

or, expanding and rearranging,

$$\|X\hat{\beta} - X\beta_0\|_2^2 \leq 2\langle \epsilon, X\hat{\beta} - X\beta_0 \rangle.$$

Here we denote $\langle a, b \rangle = a^T b$. The above is sometimes called the *basic inequality* (for the lasso in bound form). Now, rearranging the inner product, using Holder's inequality, and recalling the choice of bound parameter:

$$\|X\hat{\beta} - X\beta_0\|_2^2 \leq 2\langle X^T\epsilon, \hat{\beta} - \beta_0 \rangle \leq 4\|\beta_0\|_1 \|X^T\epsilon\|_\infty.$$

Notice that $\|X^T\epsilon\|_\infty = \max_{j=1,\ldots,p} |X_j^T\epsilon|$ is a maximum of p Gaussians, each with mean zero and variance upper bounded by $\sigma^2 n$. By a standard maximal inequality for Gaussians, for any $\delta > 0$,

$$\max_{j=1,\ldots,p} |X_j^T\epsilon| \leq \sigma\sqrt{2n\log(ep/\delta)},$$

with probability at least $1 - \delta$. Plugging this to the second-to-last display and dividing by n, we get the finite-sample result for the lasso estimator

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \leq 4\sigma\|\beta_0\|_1\sqrt{\frac{2\log(ep/\delta)}{n}}, \tag{22}$$

with probability at least $1 - \delta$

- The high-probability result (22) implies an in-sample risk bound of

$$\frac{1}{n}\mathbb{E}\|X\hat{\beta} - X\beta_0\|_2^2 \lesssim \|\beta_0\|_1\sqrt{\frac{\log p}{n}}.$$

Compare to this with the risk bound (11) for best subset selection, which is on the (optimal) order of $s_0 \log p/n$ when β_0 has s_0 nonzero components. If each of the nonzero components here has constant magnitude, then above risk bound for the lasso estimator is on the order of $s_0\sqrt{\log p/n}$, which is much slower

- **Bound form, predictive risk.** Instead of in-sample risk, we might also be interested in out-of-sample risk, as after all that reflects actual (out-of-sample) predictions. In least squares, recall, we saw that out-of-sample risk was generally higher than in-sample risk. The same is true for the lasso

Chatterjee (2013) gives a nice, simple analysis of out-of-sample risk for the lasso. He assumes that x_0, x_i, $i = 1, \ldots, n$ are i.i.d. from an arbitrary distribution supported on a compact set in $\mathbb{R}^p$, and shows that the lasso estimator in bound form (5) with $t = \|\beta_0\|_1$ has out-of-sample risk satisfying

$$\mathbb{E}(x_0^T\hat{\beta} - x_0^T\beta)^2 \lesssim \|\beta_0\|_1^2\sqrt{\frac{\log p}{n}}.$$

The proof is not much more complicated than the above, for the in-sample risk, and reduces to a clever application of Hoeffding's inequality, though we omit it for brevity. Note here the dependence on $\|\beta_0\|_1^2$, rather than $\|\beta_0\|_1$ as in the in-sample risk

- **Penalized form.** The analysis of the lasso estimator in penalized form (8) is similar to that in bound form, but only slightly more complicated. From the exact same steps leading up to the basic inequality for the bound for estimator, we have the basic inequality for the penalized form lasso estimator

$$\|X\hat{\beta} - X\beta_0\|_2^2 \leq 2\langle X^T\epsilon, \hat{\beta} - \beta_0\rangle + 2\lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1).$$

Now by Holder's inequality again, and the maximal inequality for Gaussians, we have for any $\delta > 0$,

$$\|X\hat{\beta} - X\beta_0\|_2^2 \leq 2\sigma\sqrt{2n\log(ep/\delta)}\|\hat{\beta} - \beta_0\|_1 + 2\lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1),$$

and if we choose $\lambda \geq \sigma\sqrt{2n\log(ep/\delta)}$, then by the triangle inequality,

$$\|X\hat{\beta} - X\beta_0\|_2^2 \leq 2\lambda\|\hat{\beta} - \beta_0\|_1 + 2\lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1) \leq 4\lambda\|\beta_0\|_1,$$

To recap, for any $\delta > 0$ and choice of tuning parameter $\lambda \geq \sigma\sqrt{2n\log(ep/\delta)}$, we have shown the finite-sample bound

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \leq \frac{4\lambda\|\beta_0\|_1}{n},$$

and in particular for $\lambda = \sigma\sqrt{2n\log(ep/\delta)}$,

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \leq 4\sigma\|\beta_0\|_1\sqrt{\frac{2n\log(ep/\delta)}{n}}.$$

This is the same bound as we established for the lasso estimator in bound form

- **Oracle inequality.** If we don't want to assume linearity of the mean in (2), then we can still derive an *oracle inequality* that characterizes the risk of the lasso estimator in excess of the risk of the best linear predictor. For this part only, assume the more general model

$$y = \mu(X) + \epsilon,$$

with an arbitrary mean function $\mu(X)$, and normal errors $\epsilon \sim N(0, \sigma^2)$. We will analyze the bound form lasso estimator (5) for simplicity. By optimality of $\hat{\beta}$, for any other $\tilde{\beta}$ feasible for the lasso problem in (5), it holds that

$$\langle X^T(y - X\hat{\beta}), \tilde{\beta} - \hat{\beta}\rangle \leq 0.$$

Rearranging gives

$$\langle \mu(X) - X\hat{\beta}, X\tilde{\beta} - X\hat{\beta}\rangle \leq \langle X^T\epsilon, \hat{\beta} - \tilde{\beta}\rangle.$$

Now using the polarization identity $\|a\|_2^2 + \|b\|_2^2 - \|a - b\|_2^2 = 2\langle a, b\rangle$,

$$\|X\hat{\beta} - \mu(X)\|_2^2 + \|X\hat{\beta} - X\tilde{\beta}\|_2^2 \leq \|X\tilde{\beta} - \mu(X)\|_2^2 + 2\langle X^T\epsilon, \hat{\beta} - \tilde{\beta}\rangle,$$

and from the exact same arguments as before, it holds that

$$\frac{1}{n}\|X\hat{\beta} - \mu(X)\|_2^2 + \frac{1}{n}\|X\hat{\beta} - X\tilde{\beta}\|_2^2 \leq \frac{1}{n}\|X\tilde{\beta} - \mu(X)\|_2^2 + 4\sigma t\sqrt{\frac{2\log(ep/\delta)}{n}},$$

with probability at least $1 - \delta$. This holds simultaneously over all $\tilde{\beta}$ with $\|\tilde{\beta}\|_1 \leq t$. Thus, we may write, with probability $1 - \delta$,

$$\frac{1}{n}\|X\hat{\beta} - \mu(X)\|_2^2 \leq \left\{\inf_{\|\tilde{\beta}\|_1 \leq t} \frac{1}{n}\|X\tilde{\beta} - \mu(X)\|_2^2\right\} + 4\sigma t\sqrt{\frac{2\log(ep/\delta)}{n}}.$$

Also if we write $X\tilde{\beta}^{\text{best}}$ as the best linear that predictor of ℓ_1 at most t, achieving the infimum on the right-hand side (which we know exists, as we are minimizing a continuous function over a compact set), then

$$\frac{1}{n}\|X\hat{\beta} - X\tilde{\beta}^{\text{best}}\|_2^2 \leq 4\sigma t\sqrt{\frac{2\log(ep/\delta)}{n}},$$

with probability at least $1 - \delta$

4.2 Fast rates

- Now we cover so-called *fast rates* for the lasso, which assume more about the predictors X—specifically, assume some kind of low-correlation assumption—and then provide a risk bound on the order of $s_0 \log p/n$, just as we saw for subset selection. These strong assumptions also allow us to "invert" an error bound on the fitted values into one on the coefficients

- As before, assume the linear model in (2), with X fixed, such that $\|X_j\|_2^2 \leq n$, $j = 1, \ldots, p$, and $\epsilon \sim N(0, \sigma^2)$. Denote the underlying support set by $S = \mathrm{supp}(\beta_0)$, with size $s_0 = |S|$

- There are many flavors of fast rates, and the conditions required are all very closely related. van de Geer & Buhlmann (2009) provides a nice review and discussion. Here we just discuss two such results, for simplicity

- **Compatibility result.** Assume that X satisfies the *compatibility condition* with respect to the true support set S, i.e., for some compatibility constant $\phi_0 > 0$,

$$\frac{1}{n}\|Xv\|_2^2 \geq \frac{\phi_0^2}{s_0}\|v_S\|_1^2 \quad \text{for all } v \in \mathbb{R}^p \text{ such that } \|v_{-S}\|_1 \leq 3\|v_S\|_1. \tag{23}$$

  While this may look like an odd condition, we will see it being useful in the proof below, and we will also have some help interpreting it when we discuss the restricted eigenvalue condition shortly. Roughly, it means the (truly active) predictors can't be too correlated

  Recall from our previous analysis for the lasso estimator in penalized form (8), we showed on an event E_δ of probability at least $1 - \delta$,

$$\|X\hat{\beta} - X\beta_0\|_2^2 \leq 2\sigma\sqrt{2n\log(ep/\delta)}\|\hat{\beta} - \beta_0\|_1 + 2\lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1).$$

  Choosing λ large enough and applying the triangle inequality then gave us the slow rate we derived before. Now we choose λ just slightly larger (by a factor of 2): $\lambda \geq 2\sigma\sqrt{2n\log(ep/\delta)}$. The remainder of the analysis will be performed on the event E_δ and we will no longer make this explicit until the very end. Then

$$\begin{aligned}
\|X\hat{\beta} - X\beta_0\|_2^2 &\leq \lambda\|\hat{\beta} - \beta_0\|_1 + 2\lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1) \\
&\leq \lambda\|\hat{\beta}_S - \beta_{0,S}\|_1 + \lambda\|\hat{\beta}_{-S}\|_1 + 2\lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1) \\
&\leq \lambda\|\hat{\beta}_S - \beta_{0,S}\|_1 + \lambda\|\hat{\beta}_{-S}\|_1 + 2\lambda(\|\beta_{0,S} - \hat{\beta}_S\|_1 - \|\hat{\beta}_{-S}\|_1) \\
&= 3\lambda\|\hat{\beta}_S - \beta_{0,S}\|_1 - \lambda\|\hat{\beta}_{-S}\|_1,
\end{aligned}$$

  where the two inequalities both followed from the triangle inequality, one application for each of the two terms, and we have used that $\hat{\beta}_{0,-S} = 0$. As $\|X\hat{\beta} - X\beta_0\|_2^2 \geq 0$, we have shown

$$\|\hat{\beta}_{-S} - \hat{\beta}_{0,-S}\|_1 \leq 3\|\hat{\beta}_S - \beta_{0,S}\|_1,$$

  and thus we may apply the compatibility condition (23) to the vector $v = \hat{\beta} - \beta_0$. This gives us two bounds: one on the fitted values, and the other on the coefficients. Both start with the key inequality (from the second-to-last display)

$$\|X\hat{\beta} - X\beta_0\|_2^2 \leq 3\lambda\|\hat{\beta}_S - \beta_{0,S}\|_1. \tag{24}$$

  For the fitted values, we upper bound the right-hand side of the key inequality (24),

$$\|X\hat{\beta} - X\beta_0\|_2^2 \leq 3\lambda\sqrt{\frac{s_0}{n\phi_0^2}}\|X\hat{\beta} - X\beta_0\|_2,$$

or dividing through both sides by $\|X\hat{\beta} - X\beta_0\|_2$, then squaring both sides, and dividing by n,

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \leq \frac{9s_0\lambda^2}{n^2\phi_0^2}.$$

Plugging in $\lambda = 2\sigma\sqrt{2n\log(ep/\delta)}$, we have shown that

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \leq \frac{72\sigma^2 s_0 \log(ep/\delta)}{n\phi_0^2}, \tag{25}$$

with probability at least $1-\delta$. Notice the similarity between (25) and (11): both provide us in-sample risk bounds on the order of $s_0 \log p/n$, but the bound for the lasso requires a strong compability assumption on the predictor matrix X, which roughly means the predictors can't be too correlated

For the coefficients, we lower bound the left-hand side of the key inequality (24),

$$\frac{n\phi_0^2}{s_0}\|\hat{\beta}_S - \beta_{0,S}\|_1^2 \leq 3\lambda\|\hat{\beta}_S - \beta_{0,S}\|_1,$$

so dividing through both sides by $\|\hat{\beta}_S - \beta_{0,S}\|_1$, and recalling $\|\hat{\beta}_{-S}\|_1 \leq 3\|\hat{\beta}_S - \beta_{0,S}\|_1$, which implies by the triangle inequality that $\|\hat{\beta} - \beta_0\|_1 \leq 4\|\hat{\beta}_S - \beta_{0,S}\|_1$,

$$\|\hat{\beta} - \beta_0\|_1 \leq \frac{12s_0\lambda}{n\phi_0^2}.$$

Plugging in $\lambda = 2\sigma\sqrt{2n\log(ep/\delta)}$, we have shown that

$$\|\hat{\beta} - \beta_0\|_1 \leq \frac{24\sigma s_0}{\phi_0^2}\sqrt{\frac{2\log(ep/\delta)}{n}}, \tag{26}$$

with probability at least $1-\delta$. This is a error bound on the order of $s_0\sqrt{\log p/n}$ for the lasso coefficients (in ℓ_1 norm)

- **Restricted eigenvalue result.** Instead of compatibility, we may assume that X satisfies the *restricted eigenvalue condition* with constant $\phi_0 > 0$, i.e.,

$$\frac{1}{n}\|Xv\|_2^2 \geq \phi_0^2\|v\|_2^2 \quad \text{for all subsets } J \subseteq \{1,\ldots,p\} \text{ such that } |J| = s_0 \\ \text{and all } v \in \mathbb{R}^p \text{ such that } \|v_{J^c}\|_1 \leq 3\|v_J\|_1. \tag{27}$$

This produces essentially the same results as in (25), (26), but additionally, in the ℓ_2 norm,

$$\|\hat{\beta} - \beta_0\|_2^2 \lesssim \frac{s_0 \log p}{n\phi_0^2}$$

with probability tending to 1

Note the similarity between (27) and the compatibility condition (23). The former is actually stronger, i.e., it implies the latter, because $\|\beta\|_2^2 \geq \|\beta_J\|_2^2 \geq \|\beta_J\|_1^2/s_0$. We may interpret the restricted eigenvalue condition roughly as follows: the requirement $(1/n)\|Xv\|_2^2 \geq \phi_0^2\|v\|_2^2$ for all $v \in \mathbb{R}^n$ would be a lower bound of ϕ_0^2 on the smallest eigenvalue of $(1/n)X^TX$; we don't require this (as this would of course mean that X was full column rank, and couldn't happen when $p > n$), but instead that require that the same inequality hold for v that are "mostly" supported on small subsets J of variables, with $|J| = s_0$

4.3 Support recovery

- Here we discuss results on support recovery of the lasso estimator. There are a few versions of support recovery results and again Buhlmann & van de Geer (2011) is a good place to look for a thorough coverage. Here we describe a result due to Wainwright (2009), who introduced a proof technique called the *primal-dual witness method*

- Again we assume a standard linear model (2), with X fixed, subject to the scaling $\|X_j\|_2^2 \leq n$, for $j = 1, \ldots, p$, and $\epsilon \sim N(0, \sigma^2)$. Denote by $S = \text{supp}(\beta_0)$ the true support set, and $s_0 = |S|$. Assume that X_S has full column rank

- We aim to show that, at some value of λ, the lasso solution $\hat{\beta}$ in (8) has an active set that exactly equals the true support set,

$$A = \text{supp}(\hat{\beta}) = S,$$

  with high probability. We actually aim to show that the signs also match,

$$\text{sign}(\hat{\beta}_S) = \text{sign}(\beta_{0,S}),$$

  with high probability. The primal-dual witness method basically plugs in the true support S into the KKT conditions for the lasso (13), (14), and checks when they can be verified

- We start by breaking up (13) into two blocks, over S and S^c. Suppose that $\text{supp}(\hat{\beta}) = S$ at a solution $\hat{\beta}$. Then the KKT conditions become

$$X_S^T(y - X_S\hat{\beta}_S) = \lambda s_S \tag{28}$$
$$X_{-S}^T(y - X_S\hat{\beta}_S) = \lambda s_{-S}. \tag{29}$$

  Hence, if we can satisfy the two conditions (28), (29) with a proper subgradient s, such that

$$s_S = \text{sign}(\beta_{0,S}) \quad \text{and} \quad \|s_{-S}\|_\infty = \max_{j \notin S} |s_j| < 1,$$

  then we have met our goal: we have recovered a (unique) lasso solution whose active set is S, and whose active signs are $\text{sign}(\beta_{0,S})$

  So, let's solve for $\hat{\beta}_S$ in the first block (28). Just as we did in the work on basic properties of the lasso estimator, this yields

$$\hat{\beta}_S = (X_S^T X_S)^{-1}\big(X_S^T y - \lambda \text{sign}(\beta_{0,S})\big), \tag{30}$$

  where we have substituted $s_S = \text{sign}(\beta_{0,S})$. From (29), this implies that s_{-S} must satisfy

$$s_{-S} = \frac{1}{\lambda} X_{-S}^T\big(I - X_S(X_S^T X_S)^{-1}X_S^T\big)y + X_{-S}^T X_S(X_S^T X_S)^{-1}\text{sign}(\beta_{0,S}). \tag{31}$$

  To lay it out, for concreteness, the primal-dual witness method proceeds as follows:

  1. Solve for the lasso solution over the S components, $\hat{\beta}_S$, as in (30), and set $\hat{\beta}_{-S} = 0$
  2. Solve for the subgradient over the S^c components, s_{-S}, as in (31)
  3. Check that $\text{sign}(\hat{\beta}_S) = \text{sign}(\beta_{0,S})$, and that $\|s_{-S}\|_\infty < 1$. If these two checks pass, then we have certified there is a (unique) lasso solution that exactly recovers the true support and signs

The success of the primal-dual witness method hinges on Step 3. We can plug in $y = X\beta_0 + \epsilon$, and rewrite the required conditions, $\text{sign}(\hat{\beta}_S) = \text{sign}(\beta_{0,S})$ and $\|s_{-S}\|_\infty < 1$, as

$$\text{sign}(\beta_{0,j} + \Delta_j) = \text{sign}(\beta_{0,j}), \text{ where}$$
$$\Delta_j = e_j^T (X_S^T X_S)^{-1} \big(X_S^T \epsilon - \lambda \text{sign}(\beta_{0,S}) \big), \text{ for all } j \in S, \quad (32)$$

and

$$\Big\| \frac{1}{\lambda} X_{-S}^T \big(I - X_S (X_S^T X_S)^{-1} X_S^T \big) \epsilon + X_{-S}^T X_S (X_S^T X_S)^{-1} \text{sign}(\beta_{0,S}) \Big\|_\infty < 1. \quad (33)$$

As $\epsilon \sim N(0, \sigma^2 I)$, we see that the two required conditions have been reduced to statements about Gaussian random variables. The arguments we need to check these conditions actually are quite simply, but we will need to make assumptions on X and β_0. These are:

- *Mutual incoherence:* for some $\gamma > 0$, we have

$$\|(X_S^T X_S)^{-1} X_S^T X_j\|_1 \leq 1 - \gamma, \quad \text{for } j \notin S,$$

- *Minimum eigenvalue:* for some $C > 0$, we have

$$\Lambda_{\min} \left(\frac{1}{n} X_S^T X_S \right) \geq C,$$

  where $\Lambda_{\min}(A)$ denotes the minimum eigenvalue of a matrix A

- *Minimum signal:*

$$\beta_{0,\min} = \min_{j \in S} |\beta_{0,j}| \geq \lambda \|(X_S^T X_S)^{-1}\|_\infty + \frac{4\gamma\lambda}{\sqrt{C}},$$

  where $\|A\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^q |A_{ij}|$ denotes the ℓ_∞ norm of an $m \times q$ matrix A

With these assumptions in place on X and β_0, let's first consider verifying (32), and examine Δ_S, whose components Δ_j, $j \in S$ are as defined in (32). We have

$$\|\Delta_S\|_\infty \leq \|(X_S^T X_S)^{-1} X_S^T \epsilon\|_\infty + \lambda \|(X_S^T X_S)^{-1}\|_\infty.$$

Note that $w = (X_S^T X_S)^{-1} X_S^T \epsilon$ is Gaussian with mean zero and covariance $\sigma^2 (X_S^T X_S)^{-1}$, so the variances of components of w are bounded by

$$\sigma^2 \Lambda_{\max} \Big((X_S^T X_S)^{-1} \Big) \leq \frac{\sigma^2 n}{C},$$

where we have used the minimum eigenvalue assumption. By a standard result on the maximum of Gaussians, for any $\delta > 0$, it holds with probability at least $1 - \delta$ that

$$\begin{aligned} \|\Delta_S\|_\infty &\leq \frac{\sigma}{\sqrt{C}} \sqrt{2n \log(es_0/\delta)} + \lambda \|(X_S^T X_S)^{-1}\|_\infty \\ &\leq \beta_{0,\min} + \frac{\gamma}{\sqrt{C}} \underbrace{\left(\frac{\sigma}{\gamma} \sqrt{2n \log(es_0/\delta)} - 4\lambda \right)}_{a}. \end{aligned}$$

where in the second line we used the minimum signal condition. As long as $a < 0$, we can see that the sign condition (32) is verified

Now, let's consider verifying (33). Using the mutual incoherence condition, we have

$$\Big\| \frac{1}{\lambda} X_{-S}^T \big(I - X_S (X_S^T X_S)^{-1} X_S^T \big) \epsilon + X_{-S}^T X_S (X_S^T X_S)^{-1} \text{sign}(\beta_{0,S}) \Big\|_\infty \leq \|z\|_\infty + (1 - \gamma),$$

where $z = (1/\lambda)X_{-S}^T(I - X_S(X_S^TX_S)^{-1}X_S^T)\epsilon = (1/\lambda)X_{-S}^TP_{X_S}\epsilon$, with P_{X_S} the projection matrix onto the column space of X_S. Notice that z is Gaussian with mean zero and covariance $(\sigma^2/\lambda^2)X_{-S}^TP_{X_S}X_{-S}$, so the components of z have variances bounded by

$$\frac{\sigma^2 n}{\lambda^2}\Lambda_{\max}(P_{X_S}) \leq \frac{\sigma^2 n}{\lambda^2}.$$

Therefore, again by the maximal Gaussian inequality, for any $\delta > 0$, it holds with probability at least $1 - \delta$ that

$$\begin{aligned}\Big\|\frac{1}{\lambda}X_{-S}^T\big(I - X_S(X_S^TX_S)^{-1}X_S^T\big)\epsilon + X_{-S}^TX_S(X_S^TX_S)^{-1}\mathrm{sign}(\beta_{0,S})\Big\|_\infty \\ \leq \frac{\sigma}{\lambda}\sqrt{2n\log\left(e(p-s_0)/\delta\right)} + (1-\gamma) \\ = 1 + \underbrace{\left(\frac{\sigma}{\lambda}\sqrt{2n\log\left(e(p-s_0)/\delta\right)} - \gamma\right)}_{b},\end{aligned}$$

Thus as long as $b < 0$, we can see that the subgradient condition (33) is verified

So it remains to choose λ so that $a, b < 0$. For $\lambda \geq (2\sigma/\gamma)\sqrt{2n\log(ep/\delta)}$, we can see that

$$a \leq 2\lambda - 4\lambda < 0, \quad b \leq \gamma/2 - \gamma < 0,$$

so (32), (33) are verified—and hence lasso estimator recovers the correct support and signs—with probability at least $1 - 2\delta$

4.4 A note on the conditions

- As we moved from the slow rates, to fast rates, to support recovery, the assumptions we used just got stronger and stronger. For the slow rates, we essentially assumed nothing about the predictor matrix X except for column normalization. For the fast rates, we had to additionally assume a compatibility or restricted eigenvalue condition, which roughly speaking, limited the correlations of the predictor variables (particularly concentrated over the underlying support S). For support recovery, we still needed whole lot more. The minimum eigenvalue condition on $(1/n)(X_S^TX_S)^{-1}$ is somewhat like the restricted eigenvalue condition on X. But the mutual incoherence condition is even stronger; it requires the regression coefficients

$$\eta_j(S) = (X_S^TX_S)^{-1}X_S^TX_j,$$

given by regressing each X_j on the truly active variables X_S, to be small (in ℓ_1 norm) for all $j \notin S$. In other words, no truly inactive variables can be highly correlated (or well-explained, in a linear projection sense) by any of the truly active variables. Finally, this minimum signal condition ensures that the nonzero entries of the true coefficient vector β_0 are big enough to detect. This is quite restrictive and is not needed for risk bounds, but it is crucial to support recovery

4.5 Minimax bounds

- Under the data model (2) with X fixed, subject to the scaling $\|X_j\|_2^2 \leq n$, for $j = 1, \ldots, p$, and $\epsilon \sim N(0, \sigma^2)$, Raskutti et al. (2011) derive upper and lower bounds on the minimax prediction error

$$M(s_0, n, p) = \inf_{\hat{\beta}} \sup_{\|\beta_0\|_0 \leq s_0} \frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2.$$

(Their analysis is acutally considerably more broad than this and covers the coefficient error $\|\hat{\beta} - \beta_0\|_2$, as well ℓ_q constraints on β_0, for $q \in [0, 1]$.) They prove that, under no additional assumptions on X,

$$M(s_0, n, p) \lesssim \frac{s_0 \log(p/s_0)}{n},$$

with probability tending to 1

- They also prove that, under a type of restricted eigenvalue condition in which

$$c_0 \leq \frac{(1/n)\|Xv\|_2^2}{\|v\|_2^2} \leq c_1 \text{for all } v \in \mathbb{R}^p \text{ such that } \|v\|_0 \leq 2s_0,$$

for some constants $c_0 > 0$ and $c_1 < \infty$, it holds that

$$M(s_0, n, p) \gtrsim \frac{s_0 \log(p/s_0)}{n},$$

with probability at least $1/2$

- The implication is that, for some X, minimax optimal prediction may be able to be performed at a faster rate than $s_0 \log(p/s_0)/n$; but for low correlations, this is the rate we should expect. (This is consistent with the worst-case-X analysis of Foster & George (1994), who actually show the worst-case behavior is attained in the orthogonal X case)

5 Friends, enemies, extensions (you decide which is which)

5.1 Stepwise regression

- *Forward stepwise regression* is an old method that dates back to Efroymson (1966), Draper & Smith (1966), if not earlier. Unlike the lasso, ridge regression, or best subset selection, the forward stepwise regression estimator is defined directly by an iterative algorithm, instead of by (the solution of) an optimization problem. We begin with an empty active model $A_0 = \emptyset$ and an estimate $\hat{\beta}^{(0)} = 0$ of the regression coefficients. Then for $k = 1, 2, 3, \ldots$, we repeat the following steps:

  1. Find
  $$j_k = \underset{j \notin A_{k-1}}{\text{argmax}} \frac{X_j^T P_{A_{k-1}}^{\perp} y}{\|P_{A_{k-1}}^{\perp} X_j\|_2}, \tag{34}$$
  where $P_{A_{k-1}}^{\perp} = I - P_{A_{k-1}}$, and $P_{A_{k-1}}$ is shorthand for $P_{X_{A_{k-1}}}$, the projection onto the column space of $X_{A_{k-1}}$
  2. Update $A_k = A_{k-1} \cup \{j_k\}$, and
  $$\begin{aligned} \beta_{A_k}^{(k)} &= (X_{A_k}^T X_{A_k})^{-1} X_{A_k}^T y, \\ \beta_{-A_k}^{(k)} &= 0 \end{aligned} \tag{35}$$

- The variable j_k in (34), to enter the active set, maximizes the absolute correlation with the residual from step $k-1$. Equivalently, this is the variable that minimizes the training error at step k, among all variables that we could have added to A_{k-1}; i.e., an equivalent definition is

$$j_k = \underset{j \notin A_{k-1}}{\text{argmin}} \|P_{A_{k-1} \cup \{j\}}^{\perp} y\|_2^2$$

- The k-step forward stepwise estimator (35), like the subset selection estimator (10), just performs least squares on the active set A_k, and does not apply shrinkage. Unlike best subset selection, the set A_k is chosen sequentially and is not general optimally (in the sense of minimizing the training error over *all* active sets of size k)

- There are several related stepwise methods from the signal processing literature, such as *orthogonal matching pursuit*. This algorithm replaces the definition of j_k in (34) with

$$j_k = \underset{j \notin A_{k-1}}{\operatorname{argmax}} \; X_j^T P_{A_{k-1}}^{\perp} y,$$

  hence it looks at inner products with the residual, rather than correlations with the residual, as its entry criterion

- Theory for stepwise regression (or orthogonal matching pursuit) is generally more complicated than theory is for the lasso, but several comparable results certainly exist in the statistics (or signal processing) literature

5.2 Stagewise regression

- *Forward stagewise regression* is similar to forward stepwise regression, but much less greedy. As with stepwise, we begin with $\beta^{(0)} = 0$. Now repeat for $k = 1, 2, 3, \ldots$, the following steps:

  1. Find
  $$j_k = \underset{j=1,\ldots p}{\operatorname{argmax}} \; |X_j^T (y - X\beta^{(k-1)})| \tag{36}$$
  2. Update
  $$\beta^{(k)} = \beta^{(k-1)} + \epsilon \cdot \operatorname{sign}\big(X_{j_k}^T (y - X\beta^{(k-1)})\big) \cdot e_{j_k} \tag{37}$$

  Above, $\epsilon > 0$ is a small constant (e.g., $\epsilon = 0.01$), called the step size or learning rate, and e_j denotes the jth standard basis vector in $\mathbb{R}^p$

- Once it has selected a variable j_k, as in (36), forward stagewise regression only increments the coefficient of X_{j_k} by ϵ. This "slow learning" property is a key difference between forward stagewise regression and forward stepwise regression. (The latter performs a full least squares fit after each time it selects a variable.) While both are greedy algorithms, stepwise is much greedier; after k iterations, it produces a model with exactly k active variables; on the other hand, stagewise typically requires many iterations to produce estimates of reasonable interest

- According to Hastie et al. (2009) forward stagewise was historically dismissed by statisticians as being "inefficient" and hence less useful than methods like forward or backward stepwise. This is perhaps understandable, if we keep in mind the limited computational resources of the time. From a modern perspective, however, we now appreciate that "slow learning" is a form of regularization which can of course present considerable statistical benefits

- Furthermore, by modern standards, forward stagewise is computationally cheap: to trace out a path of regularized estimates, we repeat very simple iterations, each one requiring (at most) p inner products, computations that could be trivially parallelized

- Unlike forward stepwise, whose estimates usually deviative substantially from lasso estimates, forward stagewise estimates are often surprisingly close to those from the lasso solution path. See Figure 4 for an example

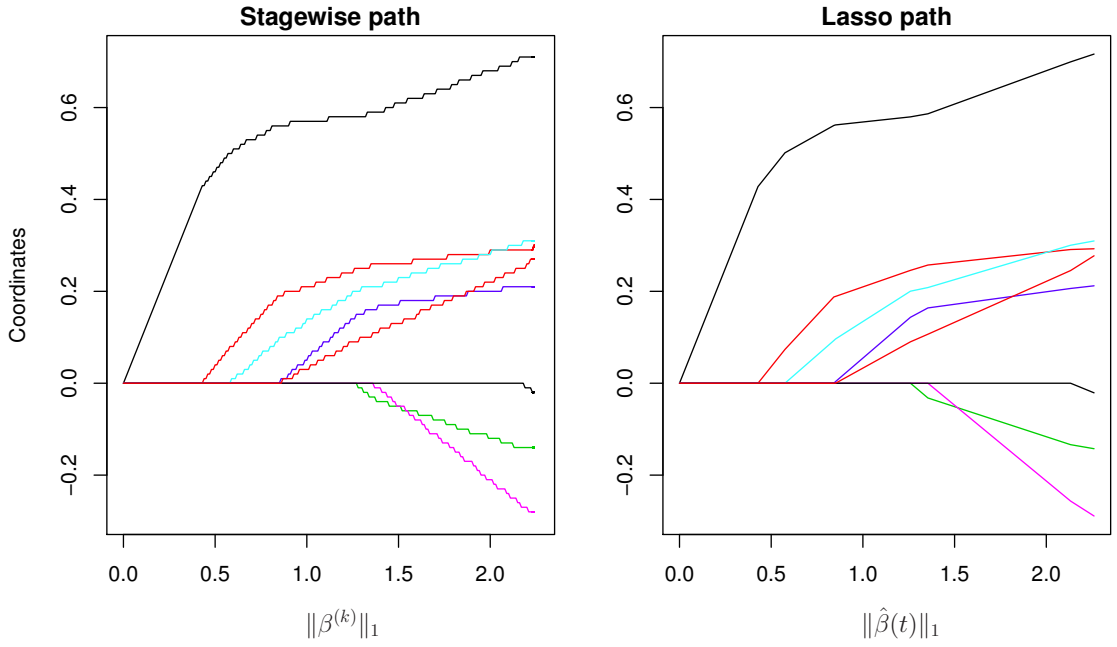

Figure 4: *A simple example comparing forward stagewise regression (with $\epsilon = 0.01$) to the lasso path*

- This connection is explained by the seminal work of Efron et al. (2004), who showed that the infintesimal forward stagewise path (i.e., the limit of the stagewise path as $\epsilon \to 0$) can be computed by modifying Algorithm 1, which as we saw, computes the lasso path. Recall that the update to the lasso coefficients in Step 1 can be alternatively written as in (21), where we move the lasso coefficients along the direction of the least squares fit of the residual $r(\lambda_k)$ on X_A. For infintesimal stagewise, we change this update to

$$\hat{\beta}_A(\lambda) = \hat{\beta}_A(\lambda_k) + \frac{\lambda_k - \lambda}{\lambda_k} v_A,$$
$$\hat{\beta}_{-A}(\lambda) = 0,$$

  where the direction v_A is defined by the nonnegative least squares problem

$$v_A = \underset{v \in \mathbb{R}^{|A|}}{\operatorname{argmin}} \ \|r(\lambda_k) - X_A v\|_2^2 \ \text{ subject to } \ v_A s_A \geq 0,$$

  ie., it is the least squares coefficients of the current residual $r(\lambda_k)$ on X_A, where we constrain the coefficients to have signs matching $s_A = \operatorname{sign}(X_A^T r(\lambda_k))$. With this modification, and also dropping Step 3 altogether, we get the infintesimal stagewise path

- This connection can be also explained more simply, by rewriting the stagewise steps in (36), (37) as

$$\beta^{(k)} = \beta^{(k-1)} + \Delta^{(k)},$$
$$\text{where } \Delta^{(k)} = \underset{z \in \mathbb{R}^p}{\operatorname{argmin}} \ \langle \nabla f(\beta^{(k-1)}), z \rangle \ \text{ subject to } \ \|z\|_1 \leq \epsilon,$$

  with $f(\beta) = (1/2)\|y - X\beta\|_2^2$ denoting the least squares loss function. Thus, at each iteration, forward stagewise moves in a direction that minimizes the inner product with the gradient of f, among all directions constrained to have a small ℓ_1 norm, and so the sequence of stagewise

estimates balances (small) decreases in the loss f with (small) increases in the ℓ_1 norm, which is just like the lasso solution path. See Tibshirani (2015) for more development of this idea

- Stagewise regression is sometimes called ϵ-boosting, and is closely related to gradient boosting where the weak learners are the variables X_j, $j = 1, \ldots, p$ themselves (instead of trees built from the variables, the typical choice). Another closely related method is *least squares boosting*, which simply replaces the stagewise update in (37) with

$$\beta^{(k)} = \beta^{(k-1)} + \epsilon \cdot \big(X_{j_k}^T (y - X\beta^{(k-1)})\big) \cdot e_{j_k},$$

  i.e., which uses the value of the inner product itself (rather than simply its sign), serving as a kind of automatic step size tuning

5.3 Relaxed lasso

- As we have discussed periodically throughout, and seen explicitly in (15), the lasso coefficients are shrunken towards zero. Depending on the scenario, such bias in the coefficient estimates may be significant and undesirable

- The *relaxed lasso* Meinshausen (2007) is an attempt to fix this, and is defined with two tuning parameters: $\lambda \geq 0$ and $\alpha \in [0, 1]$. Having computed the lasso solution $\hat{\beta}^{\text{lasso}}$ at λ, with active set $A = \text{supp}(\hat{\beta}^{\text{lasso}})$, the relaxed lasso estimate $\hat{\beta}^{\text{relaxed}}$ at λ, α solves the problem

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \alpha\lambda\|\beta\|_1 \quad \text{subject to} \quad \beta_{-A} = 0, \tag{38}$$

  i.e., we solve a reduced lasso problem, in which we allows ourselves only to fit on the original active variables X_A, but relax the amount of shrinkage by taking the penalty parameter to be $\alpha\lambda < \lambda$, when $\alpha < 1$

- Note that when $\alpha = 0$, the solution in (38) is simply $\hat{\beta}^{\text{relaxed}} = (X_A^T X_A)^{-1} X_A^T y$. This is often referred to (somewhat confusingly) as the relaxed lasso. It is also sometimes called the debiased lasso, or least squares after lasso. From our previous discussion on uniqueness, we know that when the columns of X are in general position, this least squares estimate will be well-defined (as X_A will always have full column rank)

- In general, the active sets of the relaxed lasso solutions in (38) could depart from A, as we take $\alpha < 1$, i.e., we could in principle select fewer variables than the original lasso estimate. This may be a good or bad thing. A simpler (alternative) definition for the relaxed lasso, that doesn't share this property, is

$$\begin{aligned}
\tilde{\beta}_A^{\text{relaxed}} &= \alpha\hat{\beta}_A^{\text{lasso}} + (1 - \alpha)(X_A^T X_A)^{-1} X_A^T y, \\
\tilde{\beta}_{-A}^{\text{relaxed}} &= 0.
\end{aligned}$$

  This is just linear interpolation between the lasso coefficients and the least squares coefficients on the active set. Using the form of the lasso solution in (15), we can also express this as

$$\begin{aligned}
\tilde{\beta}_A^{\text{relaxed}} &= (X_A^T X_A)^{-1} X_A^T y - \alpha\lambda(X_A^T X_A)^{-1} s_A, \\
\tilde{\beta}_{-A}^{\text{relaxed}} &= 0,
\end{aligned}$$

  so we are just "undoing" the lasso shrinkage term $\lambda(X_A^T X_A)^{-1} s_A$, as we are multiplying it by a factor $\alpha < 1$

5.4 Adaptive lasso

- Another way to reduce the bias in the lasso estimator is to weight the ℓ_1 penalty so that the coefficients we expect to be large (in magnitude) receive a smaller penalty. Zou (2006) defined the *adaptive lasso* to do just this, as the solution $\hat{\beta}^{\text{adapt}}$ of the weighted lasso problem

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \sum_{j=1}^p w_j |\beta_j|, \tag{39}$$

  for weights $w_j = 1/|\hat{\beta}_j^{\text{init}}|^\gamma$, $j = 1, \ldots, p$, where $\hat{\beta}^{\text{init}}$ is some initial estimate of the regression coefficients, and $\gamma > 0$ is another tuning parameter

- The initial estimate $\hat{\beta}^{\text{init}}$ could come from (say) ridge regression or the lasso itself. Note that when the initial estimate $\hat{\beta}^{\text{init}}$ has zero components, this makes some weights infinite, which we would formally handle by introducing equality constraints into the problem (39). Hence, since the active set of the adaptive lasso solution $\hat{\beta}^{\text{adapt}}$ is always a subset of that of $\hat{\beta}^{\text{init}}$, we would typically avoid making the initial estimate $\hat{\beta}^{\text{init}}$ super sparse; e.g., if $\hat{\beta}^{\text{init}}$ is fit via the lasso, then we might want to use a bit less regularization in the initial lasso problem

- The original work of Zou (2006) considered the predictor dimension p fixed, and showed that if $\hat{\beta}^{\text{init}}$ was chosen to be a $\sqrt{n}$-consistent estimator of the regression coefficients (which could be given simply, e.g., by least squares), then the adaptive lasso enjoys what is known as the *oracle property* for variable selection and estimation procedures

- This property is defined as follows: assuming the data model (2), with X fixed, $\epsilon \sim N(0, \sigma^2)$, and $S = \text{supp}(\beta_0)$, an estimator $\hat{\beta}$ is said to have the oracle property provided

  (i) it selects the correct variables, $\text{supp}(\hat{\beta}) = S$, with probability tending to 1, and

  (ii) the estimate $\hat{\beta}_S$ over the support is such that $\sqrt{n}(\hat{\beta}_S - \beta_{0,S})$ converges in distribution to a centered normal variate with the "right" covariance matrix, $\sigma^2(X_S^T X_S)^{-1}$ (which is the same as what the oracle estimator, least squares on X_S, would give us)

- Even when p is fixed, the Knight & Fu (2000), Zou (2006) showed that the lasso fails to meet both of the two properties simultaneously. In order for the rate of convergence of $\hat{\beta} - \beta_0$ to be $\sqrt{n}$, they showed that the tuning parameter must scale as $\lambda \asymp \sqrt{n}$, yet in this case it selects incorrect variables with positive asymptotic probability. Zou (2006) then established that the adaptive lasso remedies this issue, so long as we take $\hat{\beta}^{\text{init}}$ to be itself a $\sqrt{n}$-consistent of β_0, or much more broadly, satisfy $a_n(\hat{\beta}^{\text{init}} - \beta_0) = O_{\mathbb{P}}(1)$ for a sequence $a_n \to \infty$

- Work since has extended the theory to the high-dimensional case, in which p diverges with n, and has connected the adaptive lasso closely to the nonconvex SCAD estimator

5.5 Nonconvex penalties

- Yet another way to improve on the bias inherent to the lasso coefficients is to replace the ℓ_1 penalty in (8) for a *nonconvex penalty* defined around a nonconvex function $P : [0, \infty) \to \mathbb{R}$, and then instead solve a nonconvex optimization problem

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \sum_{j=1}^p P(|\beta_j|). \tag{40}$$

  Of course, the ℓ_0 norm, in which $P(t) = 1\{t \neq 0\}$, already fit this framework, but it presented (extreme) computational difficulties

- One might think that the natural candidates for nonconvex penalties are the ℓ_γ, $\gamma < 1$ norms, in which $P(t) = t^\gamma$, $\gamma < 1$ (these are also called bridge or power penalties). But these actually present greater computational difficulties when compared to other choices such as SCAD (Fan & Li 2001) and MC+ (Zhang 2010), defined as

$$\lambda P(t) = \int_0^t \left(1\{t \le \lambda\} + \frac{\gamma\lambda - x}{(\gamma - 1)\lambda} 1\{t > \lambda\} \right) dx \qquad \text{(SCAD)}$$

$$\lambda P(t) = \int_0^t \left(1 - \frac{x}{\gamma\lambda} \right)_+ dx \qquad \text{(MC+)}$$

respectively, where $\gamma > 2$ for SCAD, and $\gamma > 1$ for MC+. Plots of the ℓ_γ, SCAD, and MC+ penalties, along with their thresholding functions (which determine the solution in (40) when X is orthogonal) are given in Figure 5

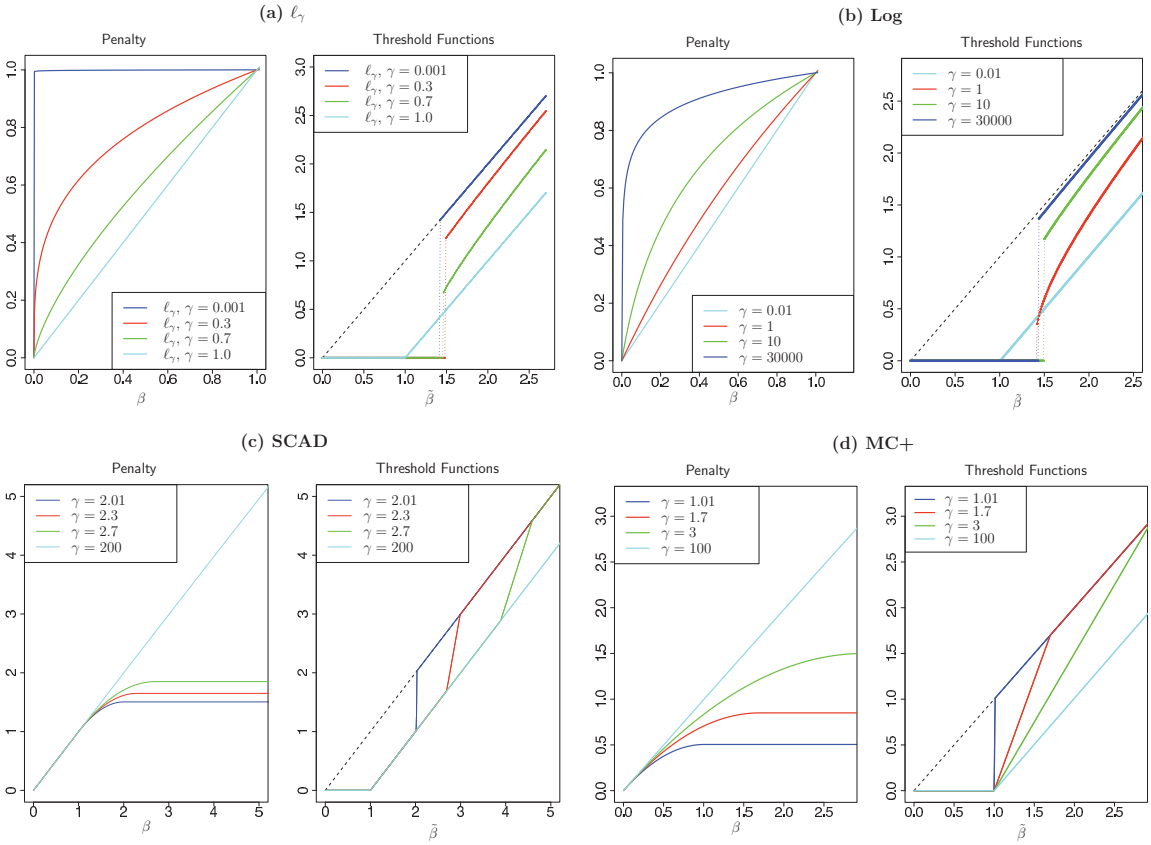

Figure 5: *Plots of the ℓ_γ, log, SCAD, MC+ nonconvex families of penalties, and their threhsolding functions, taken from Mazumder et al. (2011)*

- SCAD and MC+ have some theoretical advantages, such as the oracle property (as defined in the last subsection), over the lasso. Of course, the caveat is that we can't generically compute a global minimizer of their respective nonconvex problems; but the theory has developed to accomodate this, to varying degrees, depending on the specific theoretical statement. E.g., in some instances, statistical properties can be proved for the local minimizers found by simple iterative algorithms. A neat connection: performing one step of a local linearization algorithm

to compute the SCAD estimator is actually just the adaptive lasso, where $\hat{\beta}^{\text{init}}$ is itself (a kind of) lasso estimate

- There are many other formulations for nonconvex regularization (in regression and beyond). The literature on this just as large (it seems) as that on ℓ_1 regularization, and so there is a lot more out there and this was just a very brief introduction

5.6 Group lasso

- If we are interested in selecting groups of variables rather than individual variables to form a working linear model, then we can use the *group lasso* (Bakin 1999, Yuan & Lin 2006), which is defined as the solution of the convex problem

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \sum_{g=1}^{G} \sqrt{p_g}\|\beta_g\|_2. \tag{41}$$

  Here $\beta = (\beta_1, \ldots, \beta_G)$ denotes a block partition of the parameter β, where each block corresponds to a grouping of the variables, and p_g is the length of the gth block, for $g = 1, \ldots, G$. The ℓ_2 penalty induces group sparsity at the solution, i.e., it will be the case that $\hat{\beta}_g = 0$ for many groups g, and more so for larger λ

- A highly related problem is that of *multi-task learning* (Argyriou et al. 2006, Obozinski et al. 2010), where the same predictor variables are used across multiple regression problems (i.e., with multiple responses), and we want to perform variable selection in a uniform way across the problem. To do so, we could define groups based on the coefficients corresponding to the same variable across the multiple regressions, and solve a very similar problem to (41)

- Another useful application of the group lasso is sparse additive modeling (Lin & Zhang 2006, Ravikumar et al. 2009). In the most basic form, here we construct a basis matrix $H_j \in \mathbb{R}^{n \times N_j}$ for each dimension j of the predictor space, having entries

$$(H_j)_{i\ell} = h_{j\ell}(x_{ij}), \quad i = 1, \ldots, n, \ \ell = 1, \ldots, N_j,$$

  where N_j is a number of basis functions (say, to be chosen by the user), for $j = 1, \ldots, p$. We then solve (41) with $X = [H_1, \ldots, H_p]$, the columnwise concatenation of $H_1, \ldots, H_d$, and we define groups $\beta = (\beta_1, \ldots, \beta_G)$ just based on the coefficient blocks for each one of these basis matrices. That is, the group lasso problem (41) becomes

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\Big\|y - \sum_{j=1}^{p} H_j\beta_j\Big\|_2^2 + \lambda \sum_{j=1}^{p} \sqrt{N_j}\|\beta_j\|_2.$$

  The estimated additive function is then given by the basis expansion

$$\hat{f}(x_1, \ldots, x_p) = \sum_{j=1}^{p} \sum_{\ell=1}^{N_j} \hat{\beta}_{j\ell} h_{j\ell}(x_j),$$

  and group sparsity of $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$ means that $\hat{f}$ will depend on a selectively small number of the dimensions $j = 1, \ldots, p$

5.7 Lasso with missing data

- Consider once again the standard lasso problem (8), but suppose now that the entries of the predictor matrix X are missing at random, independently, each one missing with probability $\rho \in [0, 1)$. How should we proceed in forming a sparse working linear model, with something like the lasso? It first helps to rewrite (8) as

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\beta^T \hat{\Sigma} \beta - \hat{\alpha}^T y + n\lambda \|\beta\|_1, \tag{42}$$

  where we define $\hat{\Sigma} = X^T X/n$ and $\hat{\alpha} = X^T y/n$. We can view these quantities as plug-in estimates of $\Sigma = \mathbb{E}(x_0 x_0^T)$ and $\alpha = \mathbb{E}(x_0 y_0)$, for (x_0, y_0) drawn from the same joint distribution as the i.i.d. training pairs (x_i, y_i), $= 1, \ldots, n$. Assuming w.l.o.g. that $\mathbb{E}(x_0) = 0$, notice that Σ is the sample covariance matrix of the predictors, and α is the sample covariance between the predictors and the response

- When X is not fully observed, we can still form unbiased estimates of Σ and α. Let us define a matrix Z to have entries

$$Z_{ij} = \begin{cases} X_{ij} & \text{if } X_{ij} \text{ is observed} \\ 0 & \text{otherwise,} \end{cases} \quad i = 1, \ldots, n, \; j = 1, \ldots, p,$$

  and define $\tilde{Z} = Z/(1 - \rho)$. It is not hard to see that

$$\tilde{\Sigma} = \frac{\tilde{Z}^T \tilde{Z}}{n} - \rho \cdot \mathrm{diag}\left(\frac{\tilde{Z}^T \tilde{Z}}{n}\right) \quad \text{and} \quad \tilde{\alpha} = \frac{\tilde{Z}^T y}{n}$$

  are unbiased estimates of Σ and α, respectively. Thus the natural thing to do, it seems, is to replace $\hat{\Sigma}, \hat{\alpha}$ in the lasso problem (42) with $\tilde{\Sigma}, \tilde{\alpha}$, i.e., to solve

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\beta^T \tilde{\Sigma} \beta - \tilde{\alpha}^T y + n\lambda \|\beta\|_1 \tag{43}$$

- However, unlike the lasso problem (42), the missing-data-lasso problem (43) is not necessarily convex, because $\tilde{\Sigma}$ is not necessarily positive definite; indeed, for a large enough probability ρ of missingness, it will have negative eigenvalues

- Loh & Wainwright (2012) show that, provided that we use an additional ℓ_1 constraint in (43) (important since, if $\tilde{\Sigma}$ has negative eigenvalues, then the criterion in (43) will be unbounded from below), the missing-data-lasso problem has appealing theoretical and computational properties, despite its nonconvexity. Under conditions similar to those we discussed previously in the derivation of fast rates for the lasso, these authors prove that the missing-data-lasso estimator has risk properties in line with those in the fullly observed case. Further, they prove it is good enough to use a simple proximal gradient descent algorithm to obtain a local minimizer of (43), as it will share the same statistical properties as the global minimizer

References

Argyriou, A., Evgeniou, T. & Pontil, M. (2006), 'Multi-task feature learning', *Advances in Neural Information Processing Systems* **19**.

Bakin, S. (1999), Adaptive regression and model selection in data mining problems, PhD thesis, School of Mathematical Sciences, Australian National University.

Beale, E. M. L., Kendall, M. G. & Mann, D. W. (1967), 'The discarding of variables in multivariate analysis', *Biometrika* **54**(3/4), 357–366.

Bertsimas, D., King, A. & Mazumder, R. (2016), 'Best subset selection via a modern optimization lens', *The Annals of Statistics* **44**(2), 813–852.

Buhlmann, P. & van de Geer, S. (2011), *Statistics for High-Dimensional Data*, Springer.

Candes, E. J. & Plan, Y. (2009), 'Near ideal model selection by ℓ_1 minimization', *Annals of Statistics* **37**(5), 2145–2177.

Candes, E. J. & Tao, T. (2006), 'Near optimal signal recovery from random projections: Universal encoding strategies?', *IEEE Transactions on Information Theory* **52**(12), 5406–5425.

Chatterjee, S. (2013), Assumptionless consistency of the lasso. arXiv: 1303.5817.

Chen, S., Donoho, D. L. & Saunders, M. (1998), 'Atomic decomposition for basis pursuit', *SIAM Journal on Scientific Computing* **20**(1), 33–61.

Donoho, D. L. (2006), 'Compressed sensing', *IEEE Transactions on Information Theory* **52**(12), 1289–1306.

Draper, N. & Smith, H. (1966), *Applied Regression Analysis*, Wiley.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), 'Least angle regression', *Annals of Statistics* **32**(2), 407–499.

Efroymson, M. (1966), 'Stepwise regression—a backward and forward look', *Eastern Regional Meetings of the Institute of Mathematical Statistics* .

Fan, J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American Statistical Association* **96**(456), 1348–1360.

Foster, D. & George, E. (1994), 'The risk inflation criterion for multiple regression', *The Annals of Statistics* **22**(4), 1947–1975.

Fuchs, J. J. (2005), 'Recovery of exact sparse representations in the presense of bounded noise', *IEEE Transactions on Information Theory* **51**(10), 3601–3608.

Greenshtein, E. & Ritov, Y. (2004), 'Persistence in high-dimensional linear predictor selection and the virtue of overparametrization', *Bernoulli* **10**(6), 971–988.

Groves, T. & Rothenberg, T. (1969), 'A note on the expected value of an inverse matrix', *Biometrika* **56**(3), 690–691.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer. Second edition.

Hastie, T., Tibshirani, R. & Wainwright, M. (2015), *Statistical Learning with Sparsity: the Lasso and Generalizations*, Chapman & Hall.

Hocking, R. R. & Leslie, R. N. (1967), 'Selection of the best subset in regression analysis', *Technometrics* **9**(4), 531–540.

Hoerl, A. & Kennard, R. (1970), 'Ridge regression: biased estimation for nonorthogonal problems', *Technometrics* **12**(1), 55–67.

Knight, K. & Fu, W. (2000), 'Asymptotics for lasso-type estimators', *The Annals of Statistics* **28**(5), 1356–1378.

Lin, Y. & Zhang, H. H. (2006), 'Component selection and smoothing in multivariate nonparametric regression', *Annals of Statistics* **34**(5), 2272–2297.

Loh, P.-L. & Wainwright, M. J. (2012), 'High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity', *The Annals of Statistics* **40**(3), 1637–1664.

Mazumder, R., Friedman, J. & Hastie, T. (2011), 'SparseNet: Coordinate descent with nonconvex penalties', *Journal of the American Statistical Association* **106**(495), 1125–1138.

Meinshausen, N. (2007), 'Relaxed lasso', *Computational Statistics & Data Analysis* **52**, 374–393.

Meinshausen, N. & Buhlmann, P. (2006), 'High-dimensional graphs and variable selection with the lasso', *The Annals of Statistics* **34**(3), 1436–1462.

Obozinski, G., Taskar, B. & Jordan, M. (2010), 'Joint covariate selection and joint subspace selection for multiple classification problems', *Statistics and Computing* **20**(2), 231–252.

Osborne, M., Presnell, B. & Turlach, B. (2000*a*), 'A new approach to variable selection in least squares problems', *IMA Journal of Numerical Analysis* **20**(3), 389–404.

Osborne, M., Presnell, B. & Turlach, B. (2000*b*), 'On the lasso and its dual', *Journal of Computational and Graphical Statistics* **9**(2), 319–337.

Raskutti, G., Wainwright, M. J. & Yu, B. (2011), 'Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls', *IEEE Transactions on Information Theory* **57**(10), 6976–6994.

Ravikumar, P., Lafferty, J., Liu, H. & Wasserman, L. (2009), 'Sparse additive models', *Journal of the Royal Statistical Society: Series B* **71**(5), 1009–1030.

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

Tibshirani, R. J. (2013), 'The lasso problem and uniqueness', *Electronic Journal of Statistics* **7**, 1456–1490.

Tibshirani, R. J. (2015), 'A general framework for fast stagewise algortihms', *Journal of Machine Learning Research* **16**, 2543–2588.

van de Geer, S. & Buhlmann, P. (2009), 'On the conditions used to prove oracle results for the lasso', *Electronic Journal of Statistics* **3**, 1360–1392.

Wainwright, M. (2017), *High-Dimensional Statistics: A Non-Asymptotic View*, Cambridge University Press. To appear.

Wainwright, M. J. (2009), 'Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso)', *IEEE Transactions on Information Theory* **55**(5), 2183–2202.

Yuan, M. & Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', *Journal of the Royal Statistical Society: Series B* **68**(1), 49–67.

Zhang, C.-H. (2010), 'Nearly unbiased variable selection under minimax concave penalty', *The Annals of Statistics* **38**(2), 894–942.

Zhao, P. & Yu, B. (2006), 'On model selection consistency of lasso', *Journal of Machine Learning Research* **7**, 2541–2564.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American Statistical Association* **101**(476), 1418–1429.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B* **67**(2), 301–320.