Sparse Regression

In these notes we consider the problem of sparse regression, where the goal is to obtain a linear model that only depends on a small number of features. In Section 1 we motivate the use of the ℓ_1 norm for this purpose. In Section 2 we introduce the lasso, which combines least squares with ℓ_1 norm regularization. In Section 3 we describe convex functions and some of their properties. In Section 4 we present a geometric characterization of convex functions based on subgradients. Finally, in Section 5 we analyze the lasso cost function for a simple example with two features.

1 Promoting sparsity

In these notes we describe how to promote sparsity by minimizing the ℓ_1 norm. We will start with a motivating example that illustrates the advantages of this approach. Consider a vector, parametrized by a single real variable $t \in \mathbb{R}$,

$$v_t := \begin{bmatrix} t \\ t-1 \\ t-1 \end{bmatrix}. \tag{1}$$

Our objective is to fit t so that v_t is as sparse as possible. This amounts to minimizing the number of nonzeros of v_t, which is sometimes known as the ℓ_0 "norm" of v_t. Note, however, that this is not a valid norm because it is not homogeneous: for any x $||2x||_0 = ||x||_0 \neq 2\,||x||_0$. The graph of $||v_t||_0$ is depicted in Figure 1. Locating the global minimum of the function is very difficult. It is constant except at two isolated points, so we essentially have to evaluate the function everywhere to find the global minimum. This is a problem, because trying out all possible options is computationally infeasible. Let us consider an alternative approach: minimizing

$$f(t) := ||v_t|| \tag{2}$$

where $||\cdot||$ is a norm. As a reminder, the norm of a vector is a generalization of the concept of *length*.

Definition 1.1 (Norm)**.** *Let $\mathcal{V}$ be a vector space, a norm is a function $||\cdot||$ from $\mathcal{V}$ to $\mathbb{R}$ that satisfies the following conditions.*

- *It is homogeneous. For any scalar α and any $x \in \mathcal{V}$*

$$||\alpha\, x|| = |\alpha|\, ||x||\,. \tag{3}$$

- *It satisfies the triangle inequality*

$$||x + y|| \leq ||x|| + ||y||\,. \tag{4}$$

Carlos Fernandez-Granda, Courant Institute of Mathematical Sciences and Center for Data Science, NYU

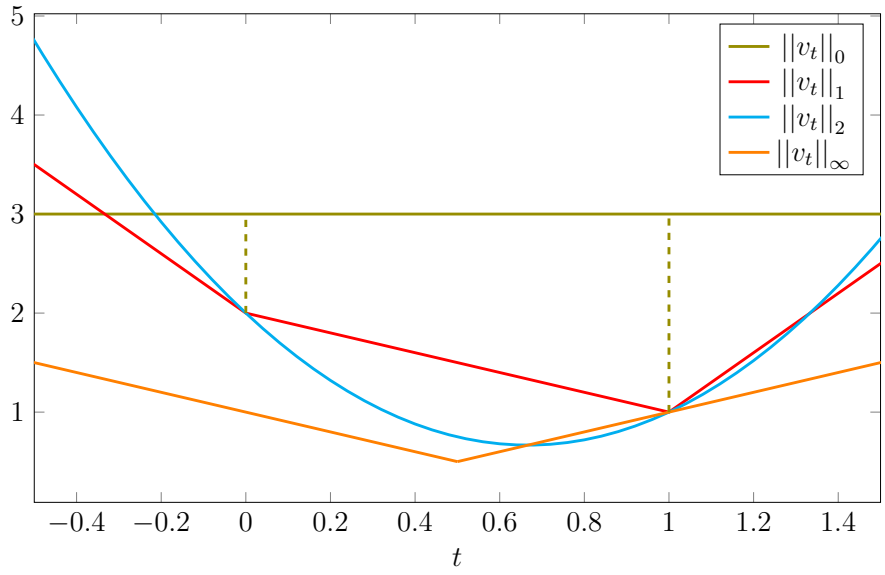

Figure 1: Graph of the function in Eq. (2) for different norms and for the nonconvex ℓ_0 "norm".

- $||x|| = 0$ *implies that* x *is the zero vector* $\vec{0}$.

Intuitively, norms represent how large a vector is, so perhaps performing the minimization might lead to a sparse solution. Some of the most commonly used norms are the ℓ_1, ℓ_2, and the ℓ_∞ norms (check that they satisfy the properties!). For a vector $x \in \mathbb{R}^d$, they are defined respectively, as

$$||x||_1 := \sum_{i=1}^{d} |x_i|\,, \tag{5}$$

$$||x||_2 := \sqrt{\sum_{i=1}^{d} x_i^2}, \tag{6}$$

$$||x||_\infty := \max_{1 \le i \le d} |x_i|\,. \tag{7}$$

As shown in Figure 1, minimizing the ℓ_2 or ℓ_∞ norms of v_t in Eq. (2) do not result in sparse solutions, but the ℓ_1 norm does: its global minimum is at the same location as the minimum ℓ_0 "norm" solution. This is not a coincidence: minimizing the ℓ_1 norm tends to promote sparsity. When compared to the ℓ_2 norm, it penalizes small entries more (ϵ^2 is much smaller than $|\epsilon|$ for small ϵ), as a result it tends to produce solutions that contain a small number of nonzero entries. In contrast to the ℓ_0 "norm", the ℓ_1 norm can be minimized by local descent methods. Following the direction in which the function decreases eventually reaches the global minimum. As suggested by this simple example, minimizing the ℓ_1 norm is a computationally-tractable way to promote sparsity.

2 The lasso

In the notes on linear regression, we observed that the performance of linear regression degrades when the number of features is close to the number of training data. This makes sense: the number of parameters of a model should be significantly smaller than the number of measurements used to fit it. However, when the number of features is very large, it is often possible to achieve accurate prediction using only a subset of them. Selecting a useful subset of features is a crucial problem in statistics, known as model selection. Consider a linear regression problem with p features associated to a coefficient vector $\beta \in \mathbb{R}^p$. The goal of model selection is to find a set of indices $\mathcal{I} \subset \{1, \ldots, p\}$, such that the response $y \in \mathbb{R}$ is well approximated by the corresponding features,

$$y \approx \sum_{i \in \mathcal{I}} \beta[i] x[i], \tag{8}$$

where we assume that the feature and the response are centered so we don't need an intercept. Equivalently, we would like to find a sparse coefficient vector $\beta \in \mathbb{R}^p$ such that

$$y \approx \langle x, \beta \rangle . \tag{9}$$

The problem of finding sparse coefficients that achieve a good fit to the data is called sparse regression. In these notes, we study the lasso, a popular sparse-regression method based on regularization.

When fitting a sparse linear model we have two objectives:

- Achieving a good fit to the data; i.e. minimizing $\|X\beta - y\|_2^2$.
- Using a small number of features; i.e. making β as sparse as possible.

This suggests minimizing a cost function that simultaneously minimizes the fitting error and maximizes the sparsity of the coefficients. In the lecture notes on linear regression we describe ridge regression, where an ℓ_2-norm regularization term penalizes coefficients that are too large to reduce overfitting. Here we would like to penalize the number of nonzeros in the support of the coefficient, i.e. its ℓ_0 "norm" instead. As explained in Section 1, this "norm" is intractable to minimize, whereas the ℓ_1 norm is convex and promotes sparsity. This suggests leveraging it as a regularizer. In statistics, the solution to an ℓ_1-norm-regularized least-squares problem is called the *lasso* estimator, introduced in [3] (see also [1]).

Definition 2.1 (The lasso)**.** *For $X \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{R}^p$, the lasso estimate is the minimizer of the optimization problem*

$$\beta_{\text{lasso}} := \arg\min_{\beta} \frac{1}{2} \|y - X^T \beta\|_2^2 + \lambda \|\beta\|_1, \tag{10}$$

where $\lambda > 0$ is a regularization parameter.

The following example with real data illustrates that the lasso can be quite effective.

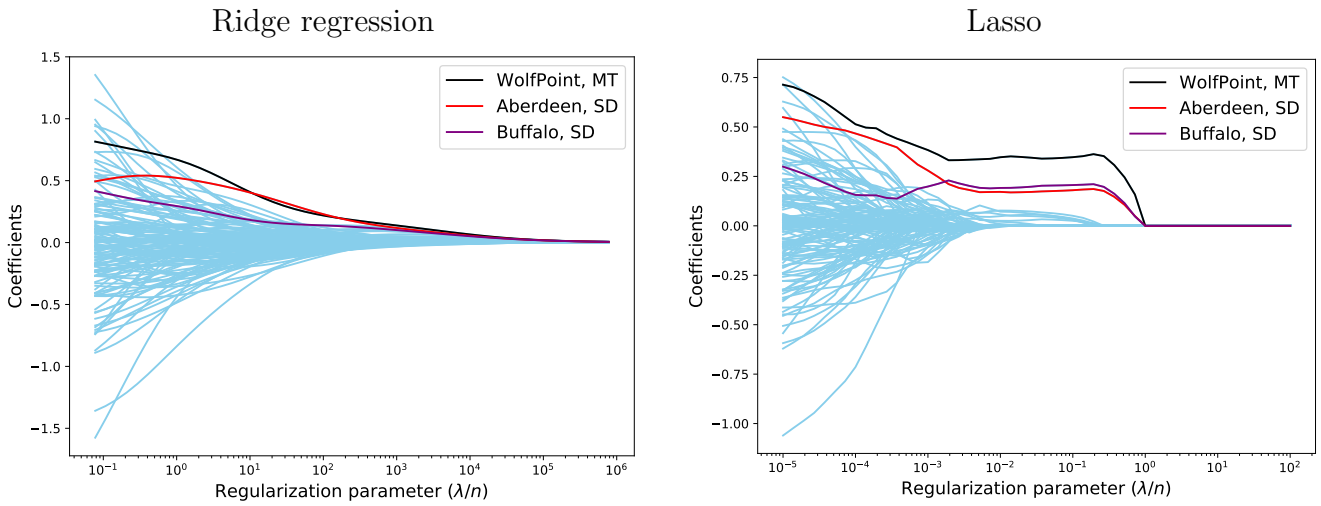

Figure 2: Coefficients of the ridge-regression (left) and lasso (right) estimators computed from the data described in Example 2.2 for different values of the regularization parameter λ. The number of training data is fixed to $n := 135$ training data. All coefficients are depicted in light blue except the three that have the largest magnitudes for large n.

Example 2.2 (Temperature prediction via sparse regression)**.** We consider the same dataset of hourly temperatures measured at weather stations all over the United States that we used in the lecture notes on linear regression. Our goal is to design a model that can be used to estimate the temperature in Jamestown (North Dakota) from the temperatures of 133 other stations. We perform estimation by fitting a linear model where the response is the temperature in Jamestown and the features are the rest of the temperatures ($p = 133$). We use 10^3 measurements from 2015 as a test set, and train a ridge-regression and a lasso model using a variable number of training data also from 2015 but disjoint from the test data. In addition, we test both models on data from 2016.

Figure 2 compares the coefficients obtained by the lasso and ridge regression for different values of the regularization parameter λ when the number of data n equals 135. Since the number of features is just 133, the least-squares estimator severely overfits the training data. This is evident in the large magnitude of the coefficients for small values of λ. As λ increases the coefficients of the ridge-regression and lasso estimators shrink, limiting the overfitting effect. However, there is a striking difference: the ridge-regression coefficients all shrink simultaneously, whereas the lasso coefficients shrink sequentially yielding increasingly sparse models. The left image in Figure 3 shows that this shrinkage indeed controls overfitting and results in an improved validation error for the lasso. The right image shows the values of λ that minimize validation error for different values of n. As expected, larger values of λ are more useful for smaller values of n, where we need regularization to avoid overfitting.

Figure 4 shows the ridge-regression and lasso coefficients corresponding to the optimal λ for different values of n. The lasso results in a much sparser linear model, except for very large values of n where both estimators approach the least-squares solution. Finally, Figure 5 shows the performance of the lasso and ridge-regression estimators on a held-out test set, as well as on data from

The data are available at http://www1.ncdc.noaa.gov/pub/data/uscrn/products

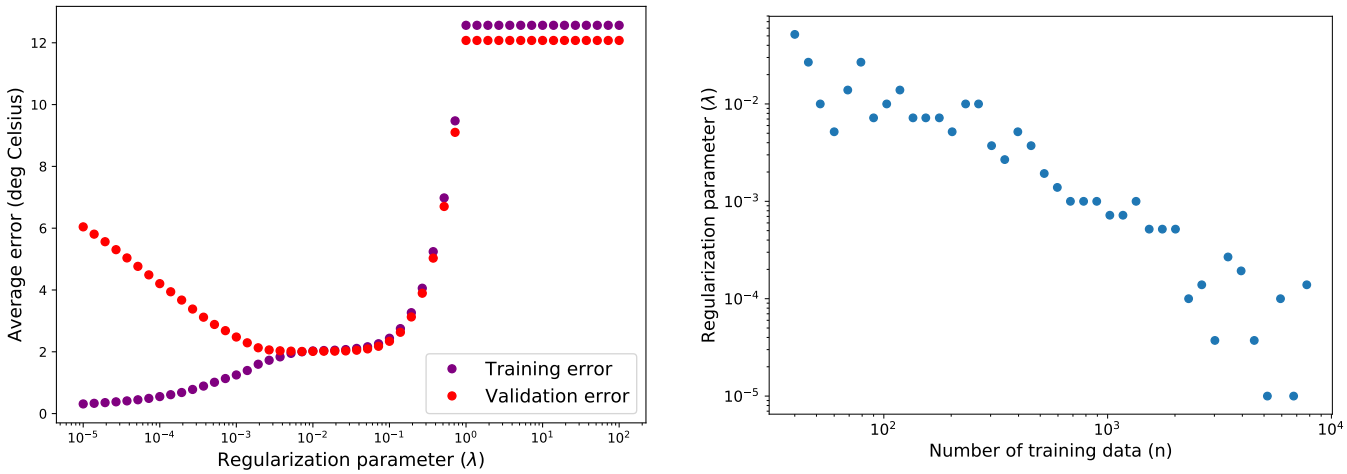

Figure 3: The left image shows the training and validation RMSE of the lasso estimator on the temperature data described in Example 2.2 when the number of training data is fixed to $n := 135$ training data. The right graph shows the values of λ selected from a validation dataset of size 100 for different values of n.

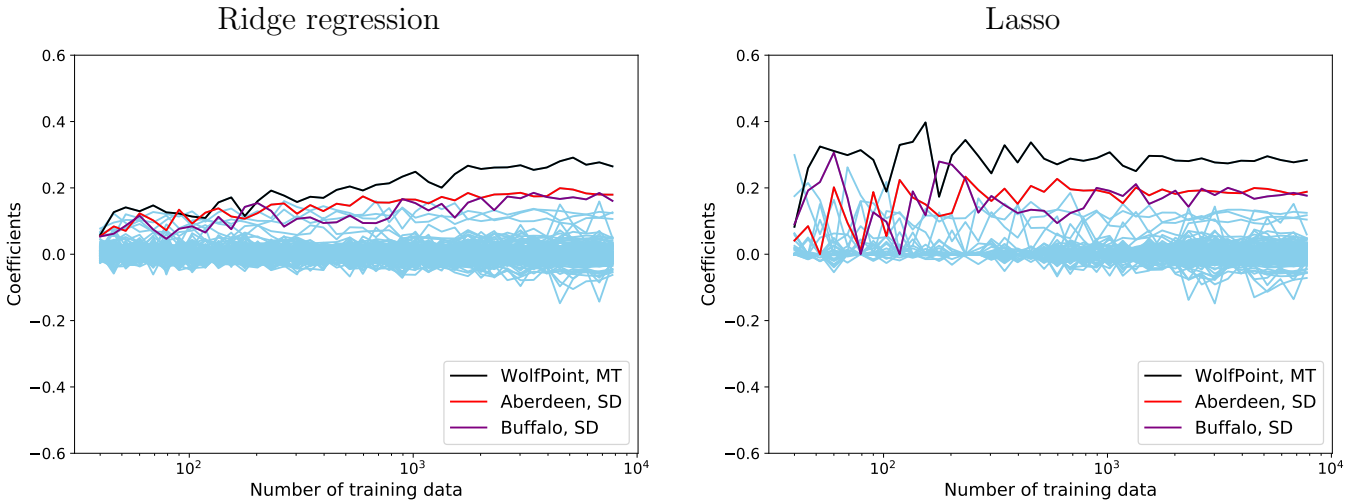

Figure 4: Coefficients of the ridge-regression (left) and lasso (right) estimators computed from the data described in Example 2.2 for different sizes of the training dataset. The coefficients to a value of λ chosen via cross validation. All coefficients are depicted in light blue except the three that have the largest magnitudes for large n.

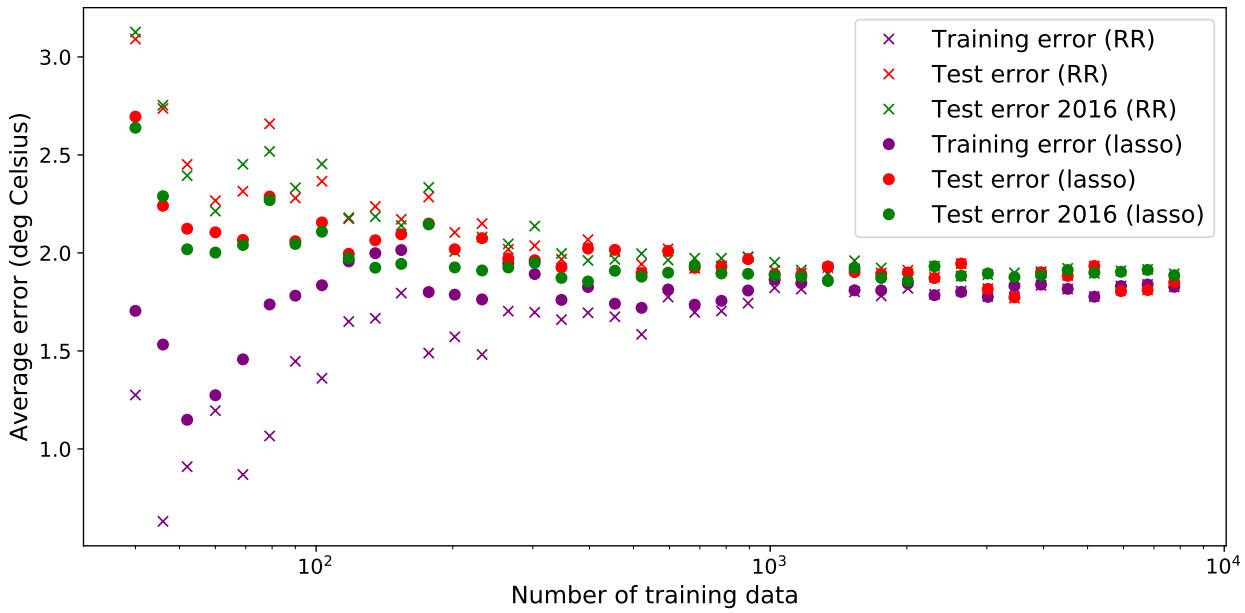

Figure 5: Performance of the lasso estimator on the temperature data described in Example 2.2. The graph shows the RMSE achieved by the models on the training and test sets, and on the 2016 data, for different number of training data and compares it to the RMSE achieved by ridge regression.

another year. The lasso achieves better prediction for all values of n, up until the point where the estimators approach the least-squares solution. △

3 Convexity

Convex functions are of crucial importance in data analysis because they can be efficiently minimized. In this section we introduce the concept of convexity and then discuss norms, which are often used to design convex cost functions when fitting models to data. A function is convex if and only if its curve lies below any chord joining two of its points. Figure 6 shows a 1D convex function.

Definition 3.1 (Convex function)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for any $x, y \in \mathbb{R}^n$ and any $\theta \in (0, 1)$,*

$$\theta f(x) + (1 - \theta) f(y) \geq f(\theta x + (1 - \theta) y). \tag{11}$$

The function is strictly convex if the inequality is always strict, i.e. if $x \neq y$ implies that

$$\theta f(x) + (1 - \theta) f(y) > f(\theta x + (1 - \theta) y). \tag{12}$$

The following lemmas illustrate some simple cases.

Lemma 3.2. *Linear functions are convex but not strictly convex.*

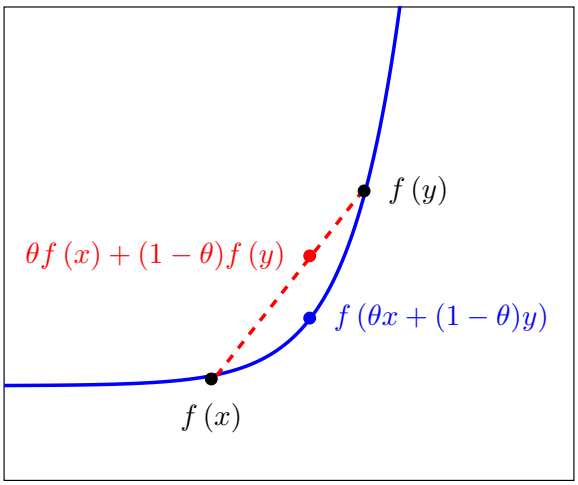

Figure 6: Illustration of condition (11) in Definition 3.1. The curve corresponding to the function must lie below any chord joining two of its points.

Proof. If f is linear, for any $x, y \in \mathbb{R}^n$ and any $\theta \in (0,1)$,

$$f(\theta x + (1-\theta) y) = \theta f(x) + (1-\theta) f(y). \tag{13}$$

□

Lemma 3.3. *Positive semidefinite quadratic forms are convex. Positive definite quadratic functions are strictly convex.*

Proof. If the following expression is nonnegative, then the function is convex. If it is positive, the function is strictly convex.

$$\begin{aligned}
&\theta f(x) + (1-\theta) f(y) - f(\theta x + (1-\theta) y) \\
&= \theta x^T A x + (1-\theta) y^T A y - (\theta x + (1-\theta) y)^T A (\theta x + (1-\theta) y) \\
&= (\theta - \theta^2) x^T A x + (1 - \theta - (1-\theta)^2) y^T A y - 2\theta(1-\theta) x^T A y \\
&= \theta(1-\theta) x^T A x + \theta(1-\theta) y^T A y - 2\theta(1-\theta) x^T A y \\
&= \theta(1-\theta)(x-y)^T A (x-y).
\end{aligned}$$

Therefore, the expression is nonnegative if the quadratic form is positive semidefinite, and positive if the quadratic form is positive definite. □

Lemma 3.4. *Adding two convex functions results in a convex function. If at least one of them is strictly convex, then the result is also strictly convex.*

Lemma 3.5. *Scaling a convex function by a nonnegative factor results in a convex function.*

We omit the proof of the two last results because they directly follow from the definition of convexity.

All norms including the ℓ_1 norm are convex.

Lemma 3.6 (Norms are convex)**.** *Any valid norm $||\cdot||$ is a convex function.*

Proof. By the triangle inequality inequality and homogeneity of the norm, for any $x, y \in \mathbb{R}^n$ and any $\theta \in (0, 1)$

$$||\theta x + (1-\theta)\, y|| \leq ||\theta x|| + ||(1-\theta)\, y|| \tag{14}$$
$$= \theta\, ||x|| + (1-\theta)\, ||y||\,. \tag{15}$$

□

However, the ℓ_0 "norm" is not convex, which corroborates what we observed in the toy example of Section 1.

Lemma 3.7. *The ℓ_0 "norm" defined as the number of nonzero entries in a vector is not convex.*

Proof. We provide a simple counterexample with vectors in $\mathbb{R}^2$ that can be easily extended to vectors in $\mathbb{R}^n$. Let $x := \left(\begin{smallmatrix}1\\0\end{smallmatrix}\right)$ and $y := \left(\begin{smallmatrix}0\\1\end{smallmatrix}\right)$, then for any $\theta \in (0, 1)$

$$||\theta x + (1-\theta)\, y||_0 = 2 > 1 = \theta\, ||x||_0 + (1-\theta)\, ||y||_0\,. \tag{16}$$

□

We conclude that the cost function of the lasso is a convex function.

Lemma 3.8. *The cost function of the lasso,*

$$f(\beta) := \frac{1}{2}\|y - X^T\beta\|_2^2 + \lambda\|\beta\|_1, \tag{17}$$

where $X \in \mathbb{R}^{p\times n}$ and $y \in \mathbb{R}^n$, is a convex function. If $p < n$ and X is full rank then it is strictly convex.

Proof. If we expand the square ℓ_2-norm term, the function is the sum of a positive semidefinite quadratic, a linear term, a constant, and a nonnegative scaling of the ℓ_1 norm. By Lemmas 3.2, 3.3, 3.4, and 3.5 the function is convex. If $p < n$ and X then the quadratic term is positive definite, and hence strictly convex, which implies that f is strictly convex. □

A crucial property of convex functions is that they cannot have suboptimal local minima.

Theorem 3.9 (Local minima are global)**.** *Any local minimum of a convex function is also a global minimum.*

Proof. We prove the result by contradiction. Let x_{loc} be a local minimum and x_{glob} a global minimum such that $f\left(x_{\text{glob}}\right) < f\left(x_{\text{loc}}\right)$. Since x_{loc} is a local minimum, there exists $\gamma > 0$ for which $f\left(x_{\text{loc}}\right) \leq f\left(x\right)$ for all $x \in \mathbb{R}^n$ such that $||x - x_{\text{loc}}||_2 \leq \gamma$. If we choose $\theta \in (0, 1)$ small enough, $x_\theta := \theta x_{\text{loc}} + (1-\theta)\, x_{\text{glob}}$ satisfies $||x_\theta - x_{\text{loc}}||_2 \leq \gamma$ and therefore

$$f\left(x_{\text{loc}}\right) \leq f\left(x_\theta\right) \tag{18}$$
$$\leq \theta f\left(x_{\text{loc}}\right) + (1-\theta)\, f\left(x_{\text{glob}}\right) \quad \text{by convexity of } f \tag{19}$$
$$< f\left(x_{\text{loc}}\right) \quad \text{because } f\left(x_{\text{glob}}\right) < f\left(x_{\text{loc}}\right). \tag{20}$$

□

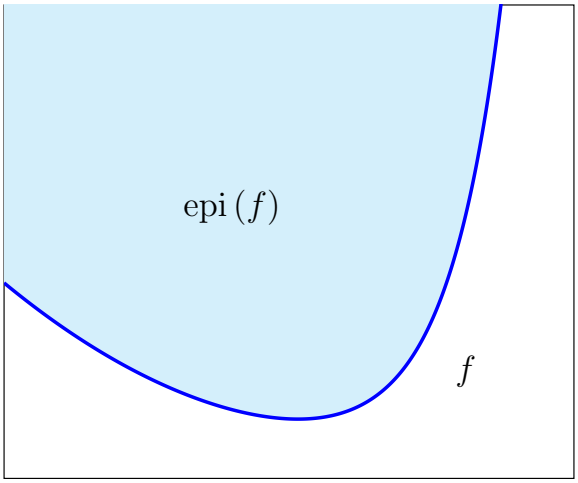

Figure 7: Epigraph of a function.

If a function is strictly convex then any local minimum is both global and unique: every other point is guaranteed to yield a larger value.

Corollary 3.10. *Strictly convex functions have at most one global minimum, and no other local minima.*

Proof. By Theorem 3.9 all local minima of the function are global minima and hence have the same value $v_{\min} := f(x) = f(y)$. Let x and y be two such minima. By strict convexity

$$f\left(0.5x + 0.5y\right) < 0.5f(x) + 0.5f(y) \tag{21}$$

$$= v_{\min}, \tag{22}$$

which contradicts the assumption that x and y are global minima. □

4 Subgradients

We now provide a geometric characterization of convex functions that will be useful to analyze the lasso. We define the epigraph of a function as the subset of $\mathbb{R}^{n+1}$ that lies above its graph. The graph is the set of vectors in $\mathbb{R}^{n+1}$ obtained by concatenating $x \in \mathbb{R}^n$ and $f\left(x\right)$ for every $x \in \mathbb{R}^n$. Figure 7 shows the epigraph of a convex function.

Definition 4.1 (Epigraph)**.** *The epigraph of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the set*

$$\text{epi}\left(f\right) := \left\{ x \mid f\left(\begin{bmatrix} x[1] \\ \cdots \\ x[n] \end{bmatrix} \right) \leq x[n+1] \right\}. \tag{23}$$

A supporting hyperplane is a hyperplane that touches a set but does not cut through it.

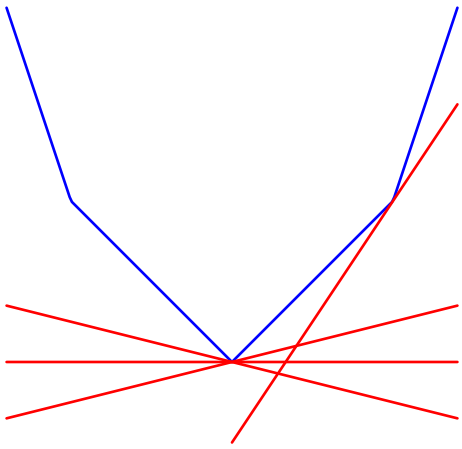

Figure 8: A nondifferentiable convex function (blue). The red lines are supporting hyperplanes (lines since the function is univariate) of its epigraph.

Definition 4.2 (Supporting hyperplane)**.** *A hyperplane $\mathcal{H}$ is a supporting hyperplane of a set $\mathcal{S}$ at x if*

- *$\mathcal{H}$ and $\mathcal{S}$ intersect at x,*
- *$\mathcal{S}$ is contained in one of the half-spaces bounded by $\mathcal{H}$.*

As established by Theorem 4.6 below, convex functions have supporting hyperplanes at any point of their epigraph, i.e. for any convex function f and any point x we can find a linear function that has the same value as f at x and lies under f otherwise. Moreover, any function that satisfies this property is convex. Figure 8 shows a 1D example.

The vector g that determines a supporting hyperplane of a convex function is called a subgradient.

Definition 4.3 (Subgradient)**.** *The subgradient of a function $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$ is a vector $g \in \mathbb{R}^n$ such that*

$$f(y) \geq f(x) + g^T (y - x), \quad \text{for all } y \in \mathbb{R}^n. \tag{24}$$

The set of all subgradients is called the subdifferential of the function at x.

Lemma 4.4. *Let $g \in \mathbb{R}^n$ be the subgradient of a function $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$. The hyperplane*

$$\mathcal{H}_g := \left\{ y \mid y[n+1] = f(x) + g^T \left(\begin{bmatrix} y[1] \\ \cdots \\ y[n] \end{bmatrix} - x \right) \right\}, \tag{25}$$

which belongs to $\mathbb{R}^{n+1}$, is a supporting hyperplane of the epigraph of f at $\begin{bmatrix} x \\ f(x) \end{bmatrix}$.

Proof. First, the two sets intersect at $\begin{bmatrix} x \\ f(x) \end{bmatrix}$. This point is clearly in the epigraph of f because it satisfies Eq. (23) (in fact it is part of the graph of the function). It also belongs to the hyperplane,

because if $y := \begin{bmatrix} x \\ f(x) \end{bmatrix}$,

$$f(x) + g^T \left(\begin{bmatrix} y[1] \\ \cdots \\ y[n] \end{bmatrix} - x \right) = f(x) = y[n+1]. \tag{26}$$

Second, any point y in the epigraph satisfies

$$y[n+1] \geq f \left(\begin{bmatrix} y[1] \\ \cdots \\ y[n] \end{bmatrix} \right) \quad \text{by definition of epigraph} \tag{27}$$

$$\geq f(x) + g^T \left(\begin{bmatrix} y[1] \\ \cdots \\ y[n] \end{bmatrix} - x \right) \quad \text{by definition of subgradient,} \tag{28}$$

so it is in the half-space

$$\left\{ y \mid y[n+1] \geq f(x) + g^T \left(\begin{bmatrix} y[1] \\ \cdots \\ y[n] \end{bmatrix} - x \right) \right\} \tag{29}$$

bounded by $\mathcal{H}_g$. □

If a function is differentiable at a given point, then the gradient is the only subgradient at that point.

Theorem 4.5 (Subdifferential of differentiable functions)**.** *If a convex function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $x \in \mathbb{R}^n$, then its subdifferential at x only contains $\nabla f(x)$.*

Proof. By the definition of subgradient, for any $1 \leq i \leq n$

$$f(x + \alpha e_i) \geq f(x) + g^T \alpha e_i \tag{30}$$

$$= f(x) + g[i]\,\alpha, \tag{31}$$

$$f(x) \leq f(x - \alpha e_i) + g^T \alpha e_i \tag{32}$$

$$= f(x - \alpha e_i) + g[i]\,\alpha, \tag{33}$$

where e_i is the ith vector in the standard basis (all its entries are equal to zero, except the ith entry which is equal to one). Combining both inequalities

$$\frac{f(x) - f(x - \alpha e_i)}{\alpha} \leq g[i] \leq \frac{f(x + \alpha e_i) - f(x)}{\alpha}. \tag{34}$$

If we let $\alpha \to 0$, this implies $g[i] = \frac{\partial f(x)}{\partial x[i]}$. Consequently, $g = \nabla f(x)$. □

Theorem 4.6. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for all $x \in \mathbb{R}^n$ there exists a subgradient $g \in \mathbb{R}^n$ such that*

$$f(y) \geq f(x) + g^T (y - x), \quad \textit{for all } y \in \mathbb{R}^n. \tag{35}$$

The inequality is always strict if and only if the function is strictly convex.

Proof. One can prove that g exists by applying the Supporting-Hyperplane Theorem, we refer the interested reader to Section 3.1.5 in [2].

To prove the converse statement, assume that for any $x, y \in \mathbb{R}^n$ and $\theta \in (0,1)$ there exists such a g. This implies

$$f(y) \geq f(\theta x + (1-\theta) y) + g^T (y - \theta x - (1-\theta) y) \tag{36}$$
$$= f(\theta x + (1-\theta) y) + \theta g^T (y - x), \tag{37}$$
$$f(x) \geq f(\theta x + (1-\theta) y) + g^T (x - \theta x - (1-\theta) y) \tag{38}$$
$$= f(\theta x + (1-\theta) y) - (1-\theta) g^T (y - x). \tag{39}$$

Multiplying Eq. (37) by $1-\theta$ and Eq. (39) by θ and adding them together yields

$$\theta f(x) + (1-\theta) f(y) \geq f(\theta x + (1-\theta) y). \tag{40}$$

□

Subgradients are useful to characterize the minima of nondifferentiable convex functions.

Theorem 4.7 (Optimality condition)**.** *A convex function attains its minimum value at a vector x if and only if the zero vector is a subgradient of f at x. If the function is strictly convex, then the minimum is unique.*

Proof. By the definition of subgradient, if $g := 0$ is a subgradient at x, then for any $y \in \mathbb{R}^n$

$$f(y) \geq f(x) + g^T (y - x) = f(x), \tag{41}$$

which is equivalent to x being a global minimum of the function. If the function is strictly convex, then the inequality is strict for all $y \neq x$. □

This optimality condition has a very intuitive geometric interpretation in terms of the supporting hyperplane associated to the subgradient. If $g = 0$ then the hyperplane is horizontal. Since the graph of the function lies above the hyperplane, the point at which they intersect must be a minimum of the function.

The sum of subgradients of two or more functions is a subgradient of their sum.

Lemma 4.8 (Sum of subgradients)**.** *Let g_1 and g_2 be subgradients at $x \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$ and $f_2 : \mathbb{R}^n \to \mathbb{R}$ respectively. Then $g := g_1 + g_2$ is a subgradient of $f := f_1 + f_2$ at x.*

Proof. For any $y \in \mathbb{R}^n$

$$f(y) = f_1(y) + f_2(y) \tag{42}$$
$$\geq f_1(x) + g_1^T (y - x) + f_2(y) + g_2^T (y - x) \tag{43}$$
$$\geq f(x) + g^T (y - x). \tag{44}$$

□

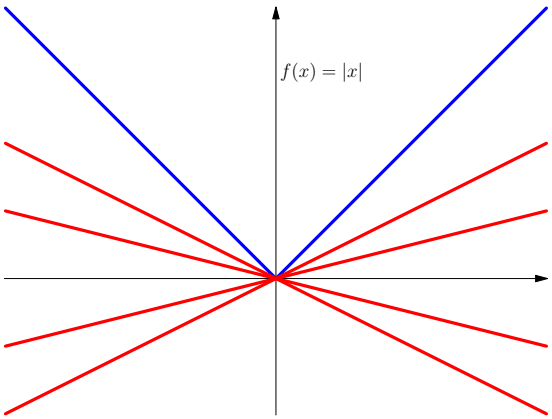

Figure 9: Examples of supporting lines of the absolute value function at the origin. The subgradients at the origin determine the slope of the lines.

The subgradient of a function scaled by a constant can be obtained by scaling the subgradient.

Lemma 4.9 (Subgradient of scaled function)**.** *Let g_1 be a subgradient at $x \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$. Then for any nonnegative $\alpha \in \mathbb{R}$ $g_2 := \alpha g_1$ is a subgradient of $f_2 := \alpha f_1$ at x.*

Proof. For any $y \in \mathbb{R}^n$

$$
\begin{aligned}
f_2(y) &= \alpha f_1(y) && (45)\\
&\geq \alpha \left(f_1(x) + g_1^T (y - x) \right) && (46)\\
&\geq f_2(x) + g_2^T (y - x). && (47)
\end{aligned}
$$

□

Finally. we derive the subdifferential of the ℓ_1 norm, which is crucial to analyze the effect of ℓ_1-norm regularization. We begin by characterizing the subdifferential in one dimension, where the ℓ_1-norm is just the absolute-value function. Figure 9 shows the supporting hyperplanes (in this case 1D lines) corresponding to a few subgradients.

Lemma 4.10 (Subdifferential of absolute value)**.** *The subdifferential of the absolute value function $|\cdot| : \mathbb{R} \to \mathbb{R}$ at x is equal to $\{\text{sign}(x)\}$ if $x \neq 0$ and to $\{g \in \mathbb{R} \mid |g| \leq 1\}$ if $x = 0$.*

Proof. If $x \neq 0$ the function is differentiable and the only subgradient is equal to the derivative by Theorem 4.5. At $x = 0$, we need

$$
\begin{aligned}
|y| &= f(0 + y) && (48)\\
&\geq f(0) + g(y - 0) && (49)\\
&\geq gy && (50)
\end{aligned}
$$

for all $y \in \mathbb{R}$, which holds if and only if $|g| \leq 1$. □

The following theorem characterizes the subdifferential of the ℓ_1 norm. Figure 10 shows several examples, as well as a supporting hyperplane corresponding to a specific subgradient.

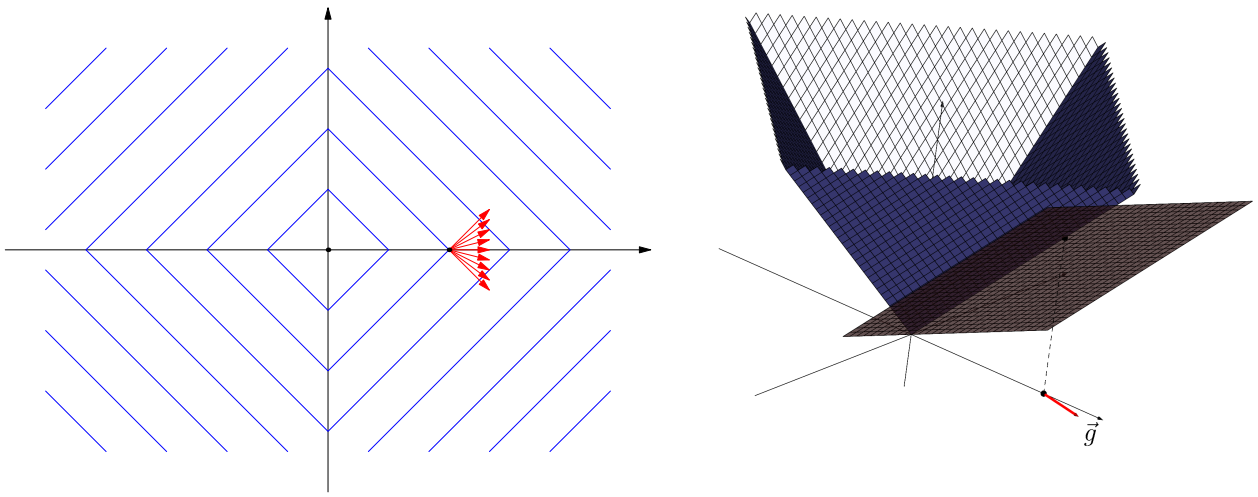

Figure 10: On the left the blue lines are contour lines of the ℓ_1 norm in $\mathbb{R}^2$. The red arrows correspond to subgradients at a point where the function is nondifferentiable. On the right, the graph of the function is shown in blue, and the supporting hyperplane corresponding to one of the subgradients (plotted as a red line with the label g) is shown in brown.

Theorem 4.11 (Subdifferential of ℓ_1 norm)**.** *The subdifferential of the ℓ_1 norm at $x \in \mathbb{R}^n$ is the set of vectors $g \in \mathbb{R}^n$ that satisfy*

$$g[i] = \text{sign}\left(x[i]\right) \quad \text{if } x[i] \neq 0, \tag{51}$$

$$|g[i]| \leq 1 \quad \text{if } x[i] = 0. \tag{52}$$

The result is a direct consequence of Lemma 4.10 and the following result.

Lemma 4.12. *The vector $g \in \mathbb{R}^n$ is a subgradient of $||\cdot||_1 : \mathbb{R}^n \to \mathbb{R}$ at x if and only if $g[i]$ is a subgradient of $|\cdot| : \mathbb{R} \to \mathbb{R}$ at $x[i]$ for all $1 \leq i \leq n$.*

Proof. If g is a subgradient of $||\cdot||_1$ at x then for any $y \in \mathbb{R}$

$$|y| = ||ye_i||_1 \tag{53}$$

$$= |x[i]| + ||x + (y - x[i])\, e_i||_1 - ||x||_1 \tag{54}$$

$$\geq |x[i]| + ||x||_1 + g^T (y - x[i])\, e_i - ||x||_1 \tag{55}$$

$$= |x[i]| + g[i]\,(y - x[i])\,, \tag{56}$$

so $g[i]$ is a subgradient of $|\cdot|$ at $|x[i]|$ for any $1 \leq i \leq n$.

If $g[i]$ is a subgradient of $|\cdot|$ at $|x[i]|$ for $1 \leq i \leq n$ then for any $y \in \mathbb{R}^n$

$$||y||_1 = \sum_{i=1}^{n} |y\,[i]| \tag{57}$$

$$\geq \sum_{i=1}^{n} |x[i]| + g[i]\,(y\,[i] - x[i]) \tag{58}$$

$$= ||x||_1 + g^T (y - x) \tag{59}$$

so g is a subgradient of $||\cdot||_1$ at x. $\square$

5 Analysis of the lasso for a simple example

In order to gain some intuition about why the lasso promotes sparsity, we consider a simple sparse regression problem where the response only depends on one feature,

$$\tilde{y} := x_{\text{true}} + \tilde{z}, \tag{60}$$

where $\tilde{y} \in \mathbb{R}^n$ is a random response vector, $x_{\text{true}} \in \mathbb{R}^n$ contains the corresponding feature values and $\tilde{z} \in \mathbb{R}^n$ is a random vector containing additive noise. In our dataset, there is an additional feature $x_{\text{other}} \in \mathbb{R}^n$, which is not part of the generative model. However, we don't know this a priori, so we use the feature matrix

$$X := \begin{bmatrix} x_{\text{true}} & x_{\text{other}} \end{bmatrix}^T, \tag{61}$$

to fit a linear-regression model with both features. Notice that each column of the matrix X corresponds to a two-dimensional feature example, where the first entry is the true feature, and the second entry is the irrelevant feature. The lasso cost function can be decomposed into a deterministic quadratic form centered at the true coefficients, a random linear function that depends on the noise, and the weighted ℓ_1 norm term (see Eq. (74) in the notes on linear regression):

$$\arg\min_{\beta} \|\tilde{y}_{\text{train}} - X^T\beta\|_2^2 + \lambda\,||\beta||_1 = \arg\min_{\beta} (\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) + \lambda\,||\beta||_1 - 2\tilde{z}_{\text{train}}^T X^T \beta.$$

where the true coefficients are

$$\beta_{\text{true}} := \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{62}$$

Figure 11 shows the different components for a specific example. The ℓ_1-norm regularization term pulls the minimum of the deterministic component towards the origin along the horizontal axis, which preserves the sparsity pattern of the minimum. Even after incorporating the noise component, the minimum of the cost function is still sparse. This is in contrast to the OLS coefficient estimate, which would not be sparse at all. Figure 12 shows the landscape of the cost function and the corresponding minimum for different noise realizations. Again, the solutions to the lasso are sparse. Figure 13 shows a scatterplot of the minima corresponding to 200 noise realizations for different values of the regularization parameter λ. For small λ the minima are spread out following the eigenvectors of the covariance matrix of the features, as in least squares. As λ increases, more and more of the minima become sparse. When λ is very large the minima shrink on the horizontal axis towards zero.

As opposed to the OLS or the ridge-regression coefficient estimate, there is no closed-form expression for the minimizer of the lasso cost function. However, we can still characterize it using subgradients. The following lemma derives the lasso solution for our simple example with two features. Intuitively, the analysis boils down to proving the existence of a horizontal supporting hyperplane at a sparse solution under certain conditions.

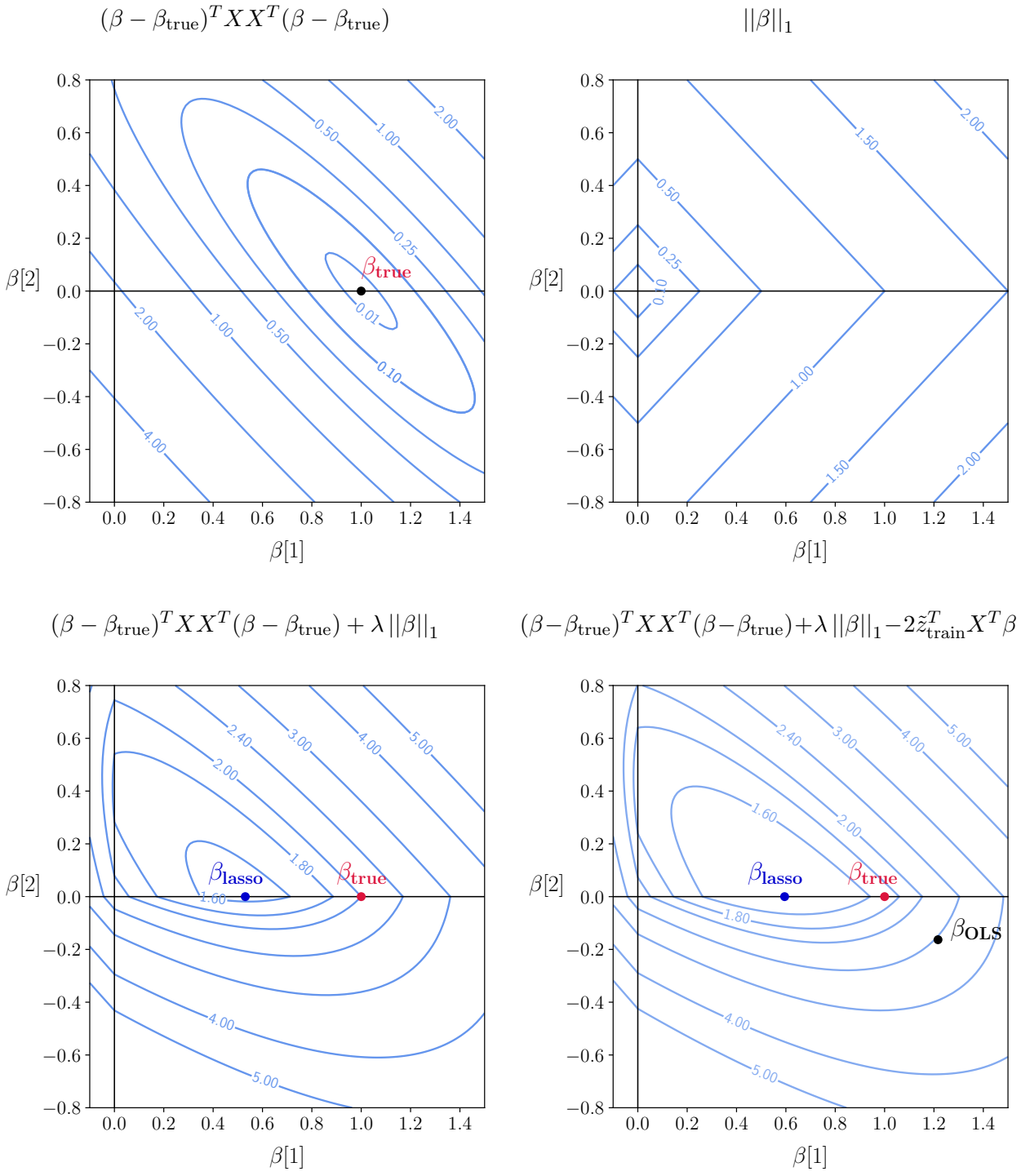

Figure 11: Visualization of the different components of the lasso cost function for a specific example with two features. The regularization parameter is set to $\lambda := 2$. The top row shows the two deterministic quadratic forms cost function: the least square component (left) and the regularization component (right). The bottom left plot shows the combination of the quadratic component and the ℓ_1-norm component. The resulting function has a sparse minimum. Finally, the bottom right plot shows a realization of the lasso cost function obtained by adding the deterministic terms with the random linear component that depends on the noise. The minimum of the resulting cost function is denoted by β_{lasso}. For comparison, we also include the minimum of the OLS cost function β_{OLS}.

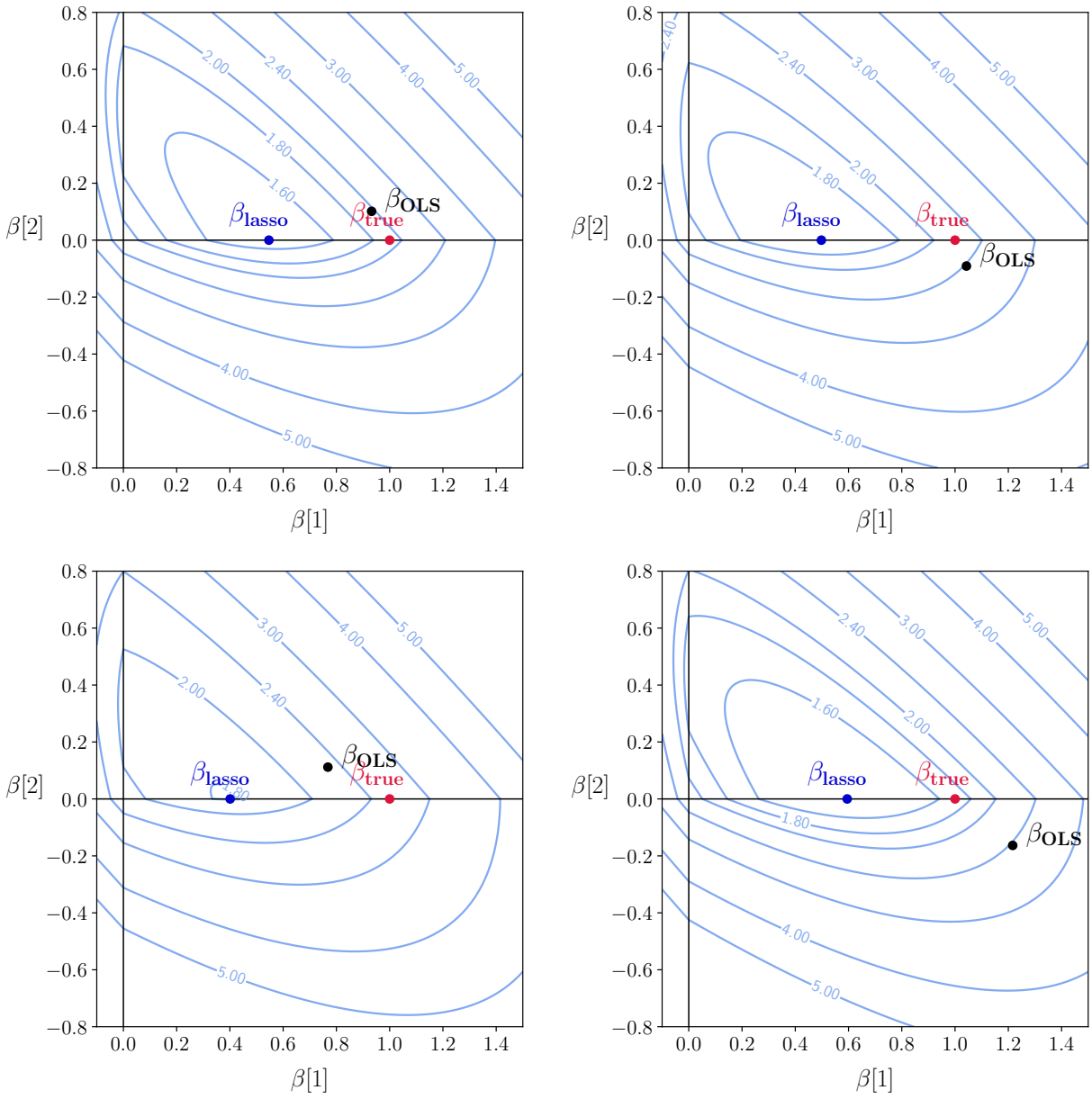

Figure 12: Different realizations of the lasso cost function corresponding to different realizations of the noise (the true coefficients and the feature matrix remain the same) for the example in Figure 11. The regularization parameter is set to $\lambda := 2$.

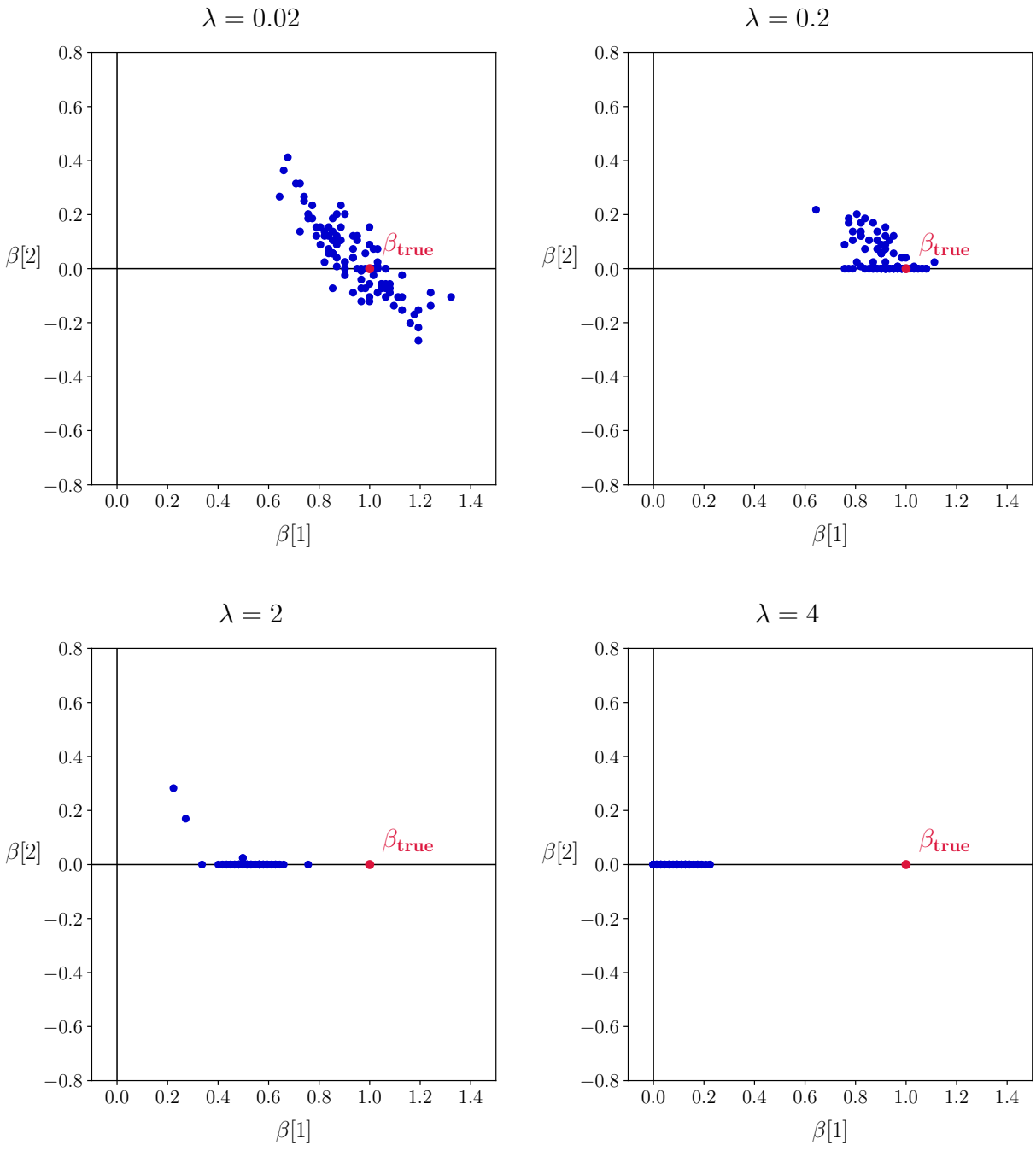

Figure 13: Scatterplot of the lasso estimate corresponding to 200 noise realizations of the example in Figure 12. Each image corresponds to a different choice of the regularization parameter λ.

Lemma 5.1 (Sparse regression with two features)**.** *Let $x_{\text{true}} \in \mathbb{R}^n$ be a fixed vector with n examples of a feature. Assume that the response is equal to the feature corrupted by some additive noise $z \in \mathbb{R}^n$, $y := x_{\text{true}} + z$. We consider a regression model that incorporates an additional feature vector $x_{\text{other}} \in \mathbb{R}^n$. Both features are centered and normalized, in particular $||x_{\text{true}}||_2 = ||x_{\text{other}}||_2 = 1$. If the regularization parameter λ of the lasso satisfies*

$$\frac{\left|x_{\text{other}}^T z - \rho x_{\text{true}}^T z\right|}{1 - |\rho|} \leq \lambda \leq 1 + x_{\text{true}}^T z, \tag{63}$$

then the lasso coefficient estimate equals

$$\beta_{\text{lasso}} = \begin{bmatrix} 1 + x_{\text{true}}^T z - \lambda \\ 0 \end{bmatrix} \tag{64}$$

where $\rho := x_{\text{true}}^T x_{\text{other}}$.

Proof. The lasso cost function is strictly convex if $n \geq 2$ and the matrix X is full rank (i.e. $|\rho| \neq 1$) by Lemma 3.8. By Theorem 4.7, to establish that β_{lasso} is the unique minimizer it suffices to prove that the zero vector is a subgradient of the cost function at β_{lasso}. Geometrically, this amounts to showing that there is a horizontal hyperplane supporting the graph of the function at β_{lasso}.

The gradient of the quadratic term

$$q(\beta) := \frac{1}{2} \left|\left|X^T \beta - y\right|\right|_2^2 \tag{65}$$

at β_{lasso} equals

$$\nabla q(\beta_{\text{lasso}}) = X\left(X^T \beta_{\text{lasso}} - y\right). \tag{66}$$

By Theorem 4.11, if only the first entry of β_{lasso} is nonzero and nonnegative, then

$$g_{\ell_1} := \begin{bmatrix} 1 \\ \gamma \end{bmatrix} \tag{67}$$

is a subgradient of the ℓ_1 norm at β_{lasso} for any $\gamma \in \mathbb{R}$ such that $|\gamma| \leq 1$. By Lemmas 4.8 and 4.9, the sum of $\nabla q(\beta_{\text{lasso}})$ and λg_{ℓ_1} is a subgradient of the lasso cost function at β_{lasso}. If only the first entry of β_{lasso} is nonzero, this subgradient equals

$$g_{\text{lasso}} := X\left(X^T \beta_{\text{lasso}} - y\right) + \lambda \begin{bmatrix} 1 \\ \gamma \end{bmatrix} \tag{68}$$

$$= X\left(\beta_{\text{lasso}}[1] x_{\text{true}} - x_{\text{true}} - z\right) + \lambda \begin{bmatrix} 1 \\ \gamma \end{bmatrix} \tag{69}$$

$$= \begin{bmatrix} x_{\text{true}}^T\left((\beta_{\text{lasso}}[1] - 1) x_{\text{true}} - z\right) + \lambda \\ x_{\text{other}}^T\left((\beta_{\text{lasso}}[1] - 1) x_{\text{true}} - z\right) + \lambda\gamma \end{bmatrix} \tag{70}$$

$$= \begin{bmatrix} \beta_{\text{lasso}}[1] - 1 - x_{\text{true}}^T z + \lambda \\ \rho(\beta_{\text{lasso}}[1] - 1) - x_{\text{other}}^T z + \lambda\gamma \end{bmatrix}. \tag{71}$$

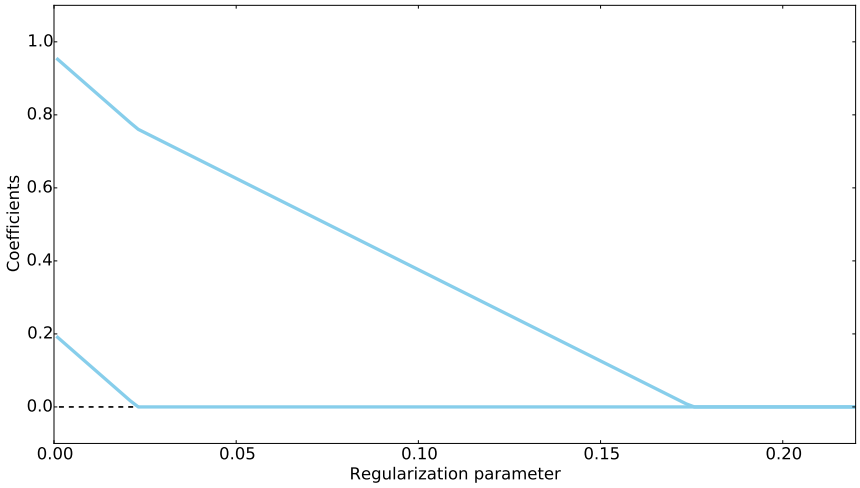

Figure 14: Coefficients of the lasso estimates for different values of the regularization parameter λ for an example where $n = 5$ and $\rho := -0.43$.

Setting g_{lasso} equal to zero we obtain

$$\beta_{\text{lasso}}[1] = 1 - \lambda + x_{\text{true}}^T z, \tag{72}$$

$$\gamma = \frac{\rho + x_{\text{other}}^T z - \rho \beta_{\text{lasso}}[1]}{\lambda} \tag{73}$$

$$= \frac{x_{\text{other}}^T z - \rho x_{\text{true}}^T z}{\lambda} + \rho. \tag{74}$$

In order to ensure that g_{lasso} is a valid subgradient, we need to check that (1) $\beta_{\text{lasso}}[1]$ is indeed nonnegative, which is the case if λ satisfies Eq. (63), and (2) that $|\gamma| \leq 1$. By the triangle inequality

$$|\gamma| \leq \left| \frac{w^T z - \rho x^T z}{\lambda} \right| + |\rho| \tag{75}$$

$$\leq 1, \tag{76}$$

as long as λ satisfies Eq. (63). We conclude that 0 is a subgradient of the cost function at β_{lasso}, which establishes that β_{lasso} as given by Eq. (64) is the unique solution to the optimization problem. □

The lemma establishes that the lasso estimator detects the relevant feature vector, setting the coefficient of the irrelevant feature vector to zero, for a certain range of λ. Within that range the coefficient corresponding to the relevant predictor scales linearly with λ. This is confirmed in Figure 14.

References

[1] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC Press, 2015.

[2] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014.

[3] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.