Convex functions, subdifferentials, and the L.a.s.s.o.

MATH 697 AM:ST

September 26, 2017

Warmup

Let us consider the absolute value function in $\mathbb{R}^p$

$$u(x) = \sqrt{(x, x)} = \sqrt{x_1^2 + \ldots + x_p^2}$$

Evidently, u is differentiable in $\mathbb{R}^p \setminus \{0\}$ and

$$\nabla u(x) = \frac{x}{|x|} \text{ for } x \neq 0$$

Warmup

What happens with this function for $x = 0$?

Two observations:
1) The function u is definitely **not** differentiable at $x = 0$

2) For every "slope" $y \in \mathbb{R}^p$ with $|y| \leq 1$ we have

$$u(x) \geq (x, y)$$

It would seem that any $y \in B_1(0) \subset \mathbb{R}^p$ represents a **tangent** to the graph of u at the origin.

Warmup

From these observations we see the following:

For **any** $x \in \mathbb{R}^p$ we have

$$u(x) = \sup_{y \in B_1(0)} (x, y)$$

If $x \neq 0$ and $y \in B_1(0)$ is such that

$$u(x) = (x, y)$$

Then y has to be equal to $x/|x|$.

Warmup

Now, this function has a global minimum at $x = 0$.

What happens if we add a smooth function, say, a linear function?

$$u_y(x) = |x| + (x, y), \quad y \in \mathbb{R}^p.$$

What happens to the global minimum?

If $|y| < 1$, the global minimum remains at $x = 0$.
As soon as $|y| = 1$, we start getting lots of minima.
For $|y| > 1$, there aren't global minima at all.

A Rapid Course On Convex functions

In our early childhood, we were taught that a function f of the real variable x is said to be **convex** in the interval $[a, b]$ if

$$f(ty + (1-t)x) \leq tf(y) + (1-t)f(x)$$

for all $t \in [0, 1]$ and any $x, y \in [a, b]$.

This definition extends in an obvious manner, to functions defined in convex domains $\Omega \subset \mathbb{R}^p$ for all dimensions $p \geq 1$.

A Rapid Course On Convex functions

An alternative way of writing the convexity condition is

$$f(y) - f(x) \geq \frac{f(x + t(y - x)) - f(x)}{t}$$

By letting $t \to 0$, it can be shown there is at least one number m such that

$$f(y) \geq f(x) + m(y - x) \;\; \forall \, y \in [a, b]$$

This being for any $x \in [a, b]$. In other words, if f is convex, then its graph has a tangent line at every point, touching from below.

A Rapid Course On Convex functions

Therefore, we arrive an equivalent formulation of convexity, one in terms of **envelopes**: a function $f : [a, b] \mapsto \mathbb{R}$ is said to be convex if it can be expressed as

$$f(x) = \max_{m \in M} \{mx - c(m)\}$$

More generally, if $\Omega \subset \mathbb{R}^p$ is a convex set, then f defined in Ω is said to be convex if there is some set $\bar{\Omega} \subset \mathbb{R}^p$ such that

$$f(x) = \max_{y \in \bar{\Omega}} \{(x, y) - c(y)\}$$

for some scalar function $c(y)$ defined in $\bar{\Omega}$.

A Rapid Course On Convex functions

This leads to the notion of the **Legendre dual** of a function.

Given a function $u : \Omega \mapsto \mathbb{R}$, it's dual, denoted u^*, is defined by

$$u^*(y) = \sup_{x \in \Omega} \{(x, y) - u(x)\}$$

Note: we are being purposefully vague about the domain of definition for $u^*(y)$, in principle, it is all of $\mathbb{R}^p$, but if one wants to avoid $u^*(y) = \infty$ one may want to restrict to a smaller set, depending on u.

A Rapid Course On Convex functions

One way then of saying a function is convex is that it must be equal to the Legendre dual of its own Legendre dual

$$u(x) = \sup_y \{(x, y) - u^*(y)\}$$

A pair of convex functions $u(x)$ and $v(y)$ such that $v = u^*$ and $u = v^*$ are said to be a **Legendre pair**.

A Rapid Course On Convex functions

If $u(x)$ and $v(y)$ are Legendre pairs, then we have what is known as Young's inequality

$$u(x) + v(y) \geq (x, y) \;\; \forall \, x, y \in \mathbb{R}^p.$$

Example

Let $u(x) = \frac{1}{a}|x|^a$, where $a > 1$. Let b be defined by the relation

$$\frac{1}{a} + \frac{1}{b} = 1$$

(one says a is the dual exponent to b). Then, we have

$$u^*(y) = \sup_{x \in \mathbb{R}^p} \left\{ (x, y) - \frac{1}{a}|x|^a \right\} = \frac{1}{b}|y|^b$$

A Rapid Course On Convex functions

Example (continued)

In this instance, Young's inequality becomes

$$\tfrac{1}{a}|x|^a + \tfrac{1}{b}|y|^b \geq (x, y) \;\; \forall\, x, y \in \mathbb{R}^p.$$

For $a = b = 2$, this is nothing but the **arithmetic-geometric mean inequality**

$$2(x, y) \leq |x|^2 + |y|^2$$

A Rapid Course On Convex functions

Some nomenclature before going forward

We have seen a convex function is but an envelope of affine functions (convex sets = intersection of half spaces).

An affine function $\ell(x)$ is one of the form

$$\ell(x) = (x, y) + c$$

where $c \in \mathbb{R}$ and $y \in \mathbb{R}^p$, the latter referred as the **slope** of ℓ.

Let $u : \Omega \mapsto \mathbb{R}$ be a convex function, $\Omega \subset \mathbb{R}^p$ a convex domain, and $x_0 \in \Omega^0$ (that is, x_0 an interior point).

An affine function ℓ is said to be **supporting** to u at x_0 if

$$u(x) \geq \ell(x) \text{ for all } x \in \Omega$$
$$u(x_0) = \ell(x_0)$$

A Rapid Course On Convex functions

The Subdifferential

Let Ω be some convex set, and $u : \Omega \mapsto \mathbb{R}$ a convex function.

The **subdifferential** of u at $x \in \Omega$ is the set

$$\partial u(x) = \{\text{slopes of } \ell\text{'s supporting to } u \text{ at } x\}.$$

The following is a key fact: if u is convex in Ω, then

$$\partial u(x) \neq \emptyset \; \forall \, x \in \Omega$$

If u is not just convex, but also differentiable in Ω, then

$$\partial u(x) = \{\nabla u(x)\} \;\; \forall\, x \in \Omega.$$

Thus, the set-valued function $\partial u(x)$ generalizes the gradient to convex, not necessarily smooth, functions.

We shall see $\partial u(x)$ shares many properties with $\nabla u(x)$, with the added bonus that $\partial u(x)$ is defined even when u fails to be differentiable.

Example

Let $u(x) = |x| = \sqrt{(x,x)}$, then

$$\partial u(0) = B_1(0)$$

Example

Let $u(x) = |x| = \sqrt{(x,x)}$, then

$$\partial u(0) = B_1(0)$$

Example

Let $u(x) = |x| = \sqrt{(x,x)}$, then

$$\partial u(0) = B_1(0)$$

Meanwhile, if $x \neq 0$, then $\partial u(x)$ has a single element

$$\partial u(x) = \left\{ \frac{x}{|x|} \right\}$$

Further examples are given by any other norm.

Example

Let $u(x) = \|u\|$ for *some norm* $\|\cdot\|$. We consider the unit ball in this metric:

$$B_1^{\|\cdot\|}(0) := \{x \in \mathbb{R}^p \ | \ \|x\| \leq 1\}$$

Then u is convex and

$$\partial u(0) = B_1^{\|\cdot\|}(0)$$

where $\|\cdot\|_*$ denotes the norm dual to $\|\cdot\|$.

A particularly interesting example is given by the ℓ^1-norm.

Example

Let $u(x) = |x_1| + \ldots + |x_p|$ then

$$\partial u(0) = [-1, 1]^d$$

A particularly interesting example is given by the ℓ^1-norm.

Example

Let $u(x) = |x|_{\ell^1} = |x_1| + \ldots + |x_p|$ then

$$\partial u(0) = [-1, 1]^d = B_1^{\ell^\infty}(0)$$

A particularly interesting example is given by the ℓ^1-norm.

Example

Let $u(x) = |x|_{\ell^1} = |x_1| + \ldots + |x_p|$ then

$$\partial u(0) = [-1, 1]^d = B_1^{\ell^\infty}(0)$$

Now, for instance, if $x = (0, \ldots, 0, x_p)$, where $x_p \neq 0$, then

$$\partial u(x) = [-1, 1] \times [-1, 1] \times \ldots \times \{\text{sign}(x_p)\}$$

A particularly interesting example is given by the ℓ^1-norm.

Example

Let $u(x) = |x|_{\ell^1} = |x_1| + \ldots + |x_p|$ then

$$\partial u(0) = [-1,1]^d = B_1^{\ell^\infty}(0)$$

Further, if $x = (0, x_2, \ldots, x_p)$, with $x_i \neq 0$ for $i \neq 1$, then

$$\partial u(x) = [-1,1] \times \{\text{sign}(x_2)\} \times \ldots \times \{\text{sign}(x_p)\}$$

A particularly interesting example is given by the ℓ^1-norm.

Example

Let $u(x) = |x|_{\ell^1} = |x_1| + \ldots + |x_p|$ then

$$\partial u(0) = [-1, 1]^d = B_1^{\ell^\infty}(0)$$

If all the x_i are $\neq 0$, then

$$\partial u(x) = \{\nabla u(x)\} = \{\text{sign}(x)\}$$

where, for $x = (x_1, \ldots, x_p)$, we already defined

$$\text{sign}(x) = (\text{sign}(x_1), \ldots, \text{sign}(x_p))$$

Example

Lastly, consider

$$u(x) = u_1(x) + u_2(x)$$

where u_1, u_2 are convex and u_1 differentiable for all x, then

$$\begin{aligned} \partial u(x) &= \nabla u_1(x) + \partial u_2(x) \\ &= \{y \mid y = \nabla u_1(x) + y' \text{ for some } y' \in \partial u_2(x)\} \end{aligned}$$

Proposition

Let $u : \Omega \mapsto \mathbb{R}$ be convex in a convex domain Ω.

If $x_0 \in \Omega^0$, the minimum of u is achieved at Ω if and only if

$$0 \in \partial u(x_0).$$

Proof of the Proposition.

If u achieves it's minimum at x_0, then

$$u(x) \geq u(x_0) \ \ \forall\, x \in \Omega,$$

which means that $0 \in \partial u(x_0)$, since 0 lies in the interior of Ω.

Conversely, if $0 \in \partial u(x_0)$, then

$$\begin{aligned} u(x) &\geq u(x_0) + (0, x - x_0) \\ &= u(x_0) \ \ \forall\, x \in \Omega, \end{aligned}$$

which means u achieves its minimum at x_0. □

Example

A good example for this proposition is given by functions of the form $u_1 + u_2$ with u_1 differentiable and $u_2(x) = \lambda|x|_{\ell^2}$ or $\lambda|x|_{\ell^1}$.

In the first case, $\partial u(x)$ is given by

$$\{\nabla u_1(x) + \lambda \frac{x}{|x|}\} \quad \text{if } x \neq 0$$

$$B_\lambda^{\ell^2}(\nabla u_1(0)) \quad \text{if } x = 0.$$

Example

A good example for this proposition is given by functions of the form $u_1 + u_2$ with u_1 differentiable and $u_2(x) = \lambda|x|_{\ell^2}$ or $\lambda|x|_{\ell^1}$.

…while in second case,

$$\begin{array}{rl} \{\nabla u_1(x) + \lambda \text{sign}(x)\} & \text{if } x_i \neq 0 \; \forall i \\ B_\lambda^{\ell^1}(\nabla u_1(0)) & \text{if } x = 0. \end{array}$$

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

Let us return to the Lasso functional.

$$J(\beta) = \frac{1}{2}|\mathbf{X}\beta - \mathbf{y}|^2 + \lambda|\beta|_{\ell^1}$$

Where $\mathbf{X}$ and $\mathbf{y}$ are the usual suspects, and $\lambda > 0$.
(assumed to be deterministic and centered)

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

...Last class

We observed that for β's such that $\beta_j \neq 0 \; \forall \; j$

$$\nabla|\beta|_{\ell^1} = \text{sign}(\beta) := (\text{sign}(\beta_1), \ldots, \text{sign}(\beta_p))$$

Then, for such β, we have

$$\nabla J(\beta) = \mathbf{X}^t(\mathbf{X}\beta - \mathbf{y}) + \lambda \text{sign}(\beta)$$

...which we led us to conclude

Trying to solve $\nabla J(\beta) = 0$ is not as straightforward now as in least squares! The resulting equation is not linear and discontinuous whenever any of the β_j vanishes.

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

In light of the theory for convex functions, we conclude

$$\hat{\beta}^{\mathrm{L}} \text{ is characterized by the condition } 0 \in \partial J(\hat{\beta}^{\mathrm{L}})$$

This, in turn, becomes

$$-\mathbf{X}^t(\mathbf{X}\beta^{\mathrm{L}} - \mathbf{y}) \in \lambda \partial(\| \cdot \|_{\ell^1})(\hat{\beta}^{\mathrm{L}})$$

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

In light of the theory for convex functions, we conclude

$$\hat{\beta}^{\mathrm{L}} \text{ is characterized by the condition } 0 \in \partial J(\hat{\beta}^{\mathrm{L}})$$

This, in turn, becomes

$$-\mathbf{X}^t(\mathbf{X}\beta^{\mathrm{L}} - \mathbf{y}) \in \lambda \partial(\| \cdot \|_{\ell^1})(\hat{\beta}^{\mathrm{L}})$$

A good theoretical characterization, but still not enough to compute $\hat{\beta}^{\mathrm{L}}$ in practice!.

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

Example

Consider the case $p = 1$ and with data $x_1, \ldots, x_N$ such that

$$x_1^2 + \ldots + x_N^2 = 1$$

Then, we consider the function of the real variable β

$$J(\beta) = \frac{1}{2} \sum_{i=1}^{N} |x_i \beta - y_i|^2 + \lambda |\beta|$$

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

Example

As it turns out, the minimizer for $J(\beta)$ is given by

$$\text{sign}(\hat{\beta})(|\hat{\beta}| - \lambda)_+$$

where $\hat{\beta}$ is the corresponding least squares solution

$$\hat{\beta} = \sum_{i=1}^{N} x_i y_i$$

Example

Let us see why this is so. First, expand $J(\beta)$

$$J(\beta) = \frac{1}{2}\sum_{i=1}^{N}\left(x_i^2\beta^2 - 2\beta x_i y_i + y_i^2\right) + \lambda|\beta|$$

Differentiating, we have

$$J'(\beta) = \begin{cases} \beta - (\hat{\beta} - \lambda) & \text{if } \beta > 0 \\ \beta - (\hat{\beta} + \lambda) & \text{if } \beta < 0 \end{cases}$$

Example

If $\hat{\beta} \in [-\lambda, \lambda]$, then

$$J'(\beta) \geq 0 \text{ if } \beta > 0, \;\; J'(\beta) \leq 0 \text{ if } \beta < 0$$

In which case, J is minimized by $\beta = 0$.

Example

If $\hat{\beta} \not\in \in [-\lambda, \lambda]$, then, assuming that $\hat{\beta} > 0$

$$J'(\beta) \geq 0 \text{ if } \beta > \hat{\beta} - \lambda, \;\; J'(\beta) \leq 0 \text{ if } \beta < \hat{\beta} - \lambda, \; \beta \neq 0.$$

In other words, J is decreasing in $(-\infty, \hat{\beta} - \lambda)$ and increasing in $(\hat{\beta} - \lambda, +\infty)$. Therefore, J is minimized at $\hat{\beta} - \lambda$.

If $\hat{\beta} < 0$, an analogous argument shows J is minimized at $\hat{\beta} + \lambda$.

Example

There is popular, succint notation for this relation between $\hat{\beta}$ and $\hat{\beta}^{\mathrm{L}}$. If we define the "shrinking operator" of order λ,

$$\mathcal{S}_\lambda(\beta) = \operatorname{sign}(\beta)(|\beta| - \lambda)_+$$

then $\hat{\beta}^{\mathrm{L}} = \mathcal{S}_\lambda(\hat{\beta})$.

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

What about $p > 1$? Let $\beta = (\beta_1, \ldots, \beta_p)$, we have

$$J(\beta) = \frac{1}{2}\sum_{i=1}^{N}(y_i - (x_i, \beta))^2 + \lambda|\beta|_{\ell^1}$$

Let us see that, at least if the inputs are **orthogonal**, things are as simple as in one dimension.

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

Let us expand the quadratic part

$$\sum_{i=1}^{N} \left\{ \left(\sum_{j=1}^{N} x_{ij}\beta_j \right)^2 - 2\sum_{j=1}^{N} x_{ij}\beta_j y_i + y_i^2 \right\}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{\ell=1}^{N} x_{ij}\beta_j x_{i\ell}\beta_\ell - 2\sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij}\beta_j y_i + \sum_{i=1}^{N} y_i^2$$

The inputs x_i being **orthogonal** refers to the condition

$$\sum_{i=1}^{N} x_{ij}x_{il} = \delta_{j\ell}$$

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

Therefore, the quadratic part is

$$\sum_{j=1}^{N} \beta_j^2 - 2 \sum_{i=1}^{N} \sum_{j=1}^{N} x_{ij} \beta_j y_i + \sum_{i=1}^{N} y_i^2$$

and the full functional may be written as

$$\sum_{j=1}^{N} \left\{ \tfrac{1}{2} \beta_j^2 - \beta_j \left(\sum_{i=1}^{N} x_{ij} y_i \right) + \lambda |\beta_j| \right\} + \sum_{i=1}^{N} y_i^2$$

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

By comparing

$$\sum_{j=1}^{N}\left\{\tfrac{1}{2}\beta_j^2-\beta_j\left(\sum_{i=1}^{N}x_{ij}y_i\right)+\lambda|\beta_j|\right\}+\sum_{i=1}^{N}y_i^2$$

with the expansion for the case $p=1$,

$$J(\beta)=\left(\sum_{i=1}^{N}x_i^2\right)\tfrac{1}{2}\beta^2-\beta\left(\sum_{i=1}^{N}x_iy_i\right)+\lambda|\beta|+\sum_{i=1}^{N}y_i^2$$

We conclude that each β_j is solving a one dimensional problem, separate from all the other coefficients.

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

Therefore, we see that the Lasso works, at least for orthogonal data, according to

$$\hat{\beta}^{\mathrm{L}} = \mathcal{S}_\lambda(\hat{\beta})$$

where the muldimensional shrinking operator $\mathcal{S}_\lambda$ is defined by

$$\mathcal{S}_\lambda(\beta) = (\mathcal{S}_\lambda(\beta_1), \ldots, \mathcal{S}_\lambda(\beta_p))$$

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

The orthogonality assumption is, of course, **too restrictive for practical purposes**. A change of variables to normalize $\mathbf{X}^t\mathbf{X}$ is often problematic too.

Additionally, it is worth noting that ℓ^1 is not rotationally invariant, so changing the Cartesian system of coordinates can have **dramatic** effects on the outcome!.

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

The orthogonality assumption is, of course, **too restrictive for practical purposes**. A change of variables to normalize $\mathbf{X}^t\mathbf{X}$ is often problematic too.

Additionally, it is worth noting that ℓ^1 is not rotationally invariant, so changing the Cartesian system of coordinates can have **dramatic** effects on the outcome!.

As it turns out, however, the Lasso can be cast as a quadratic optimization problem with linear constraints.

More on the Lasso, a bit on Dantzig

MATH 697 AM:ST

September 28, 2017

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

Originally (Tibshirani 1996)) the Lasso was set up as follows:

Fix $t > 0$. Then, under the constraint $|\beta|_{\ell^1} \leq t$, minimize

$$\frac{1}{2}\sum_{i=1}^{N}(y_i - (x_i, \beta))^2$$

This is a convex optimization problem with finitely many linear constraints. Indeed, the set of β's such that $|\beta|_{\ell^1} \leq t$ corresponds to the intersection of a finite number of half-spaces.

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

If t is sufficiently large, then this problem has the same solution as standard least squares:

Let $\hat{\beta}$ denote the usual least squares estimator. Then, trivially

$$|\mathbf{X}\beta - \mathbf{y}|^2 \geq |\mathbf{X}\hat{\beta} - \mathbf{y}|^2 \;\; \forall\, \beta \in \mathbb{R}^p.$$

In particular, if t is such that

$$|\hat{\beta}|_{\ell^1} \leq t$$

then, $\hat{\beta}^{\mathrm{L}} = \hat{\beta}$, the least squares and Lasso solutions coincide.

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

If instead t is such that

$$|\hat{\beta}|_{\ell^1} > t$$

Then $\hat{\beta}^{\mathrm{L}}$ will be different, and will be such that $|\hat{\beta}^{\mathrm{L}}|_{\ell^1} = t$.

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

This means that for $\beta = \hat{\beta}^{\mathrm{L}}$ there is $\lambda > 0$ such that

$$-\nabla \frac{1}{2} |\mathbf{X}\beta - \mathbf{y}|^2 \in \lambda \partial u(\beta)$$

where $u(\beta) = |\beta|_{\ell^1}$. **See: Karush-Kuhn-Tucker conditions.**

We see then, that the parameter λ seen in the first formulation corresponds to a Lagrange multiplier in the second formulation.

Thinking in the Lasso formulation

$$\text{Minimize } \frac{1}{2}\sum_{i=1}^{N}(y_i - (x_i, \beta))^2 \text{ with the constraint } |\beta|_{\ell^1} \leq t$$

Then, for $p = 3$, for instance, one sees that

$$\hat{\beta}^{\text{Lasso}} \text{ lies on a vertex } = \text{ two zero components}$$
$$\hat{\beta}^{\text{Lasso}} \text{ lies on an edge } = \text{ one zero components}$$
$$\hat{\beta}^{\text{Lasso}} \text{ lies on a face } = \text{ no non-zero components}$$

The Lasso

The Simplex Method and Coordinate Descent

The importance of the Lasso being a **quadratic program** with finitely many linear constraints is that there it allows one to apply the classical **simplex method** to approximate the solution efficiently.

Another algorithm that is popular in practice (and the one used by ML libraries for instance, in python) is the **coordinate descent** algorithm, which in a sense reduces things to lots of one dimensional problems.

The fourth week, in one slide

1. For convex functions, the subdifferential is a set valued map which serves as a good replacement for the gradient for non-differentiable functions.
2. The subdifferential yields the criterium $0 \in \partial J(\hat{\beta})$ for global minimizers $\hat{\beta}$ of a convex functional J (such as the least squares or Lasso functional).
3. We learned that the outcome of the Lasso often reduces to the application of a **soft thresholding** operator to the standar least squares estimator.
4. The Lasso can be recast as a quadratic optimization problem with finitely many constraints, making it amenable to treatment via tools like the simplex method.