Optimization

Presenter: Pushpendre Rastogi

Some content taken from
http://www.ece.rice.edu/~fk1/classes/ELEC697/Lec_8n9_LinearClass2.ppt
http://www.cse.msu.edu/~rongjin/adv_ml/slides/boydsection2KeyurDesai.ppt
http://www.cse.bgu.ac.il/common/download.asp?FileName=Lectur3.ppt
http://networks.cs.ucdavis.edu/opt_review/appendix.ppt
http://www.cs475.org/spring2017/download/type=lectures/name=cs475sp17-lecture8-handout.pdf
http://www.cs.nyu.edu/~mohri/mlu/mlu_lecture_8.pdf

Problem Formulation

- Let $w \in \mathbb{R}^d$ and $S \subset \mathbb{R}^d$ and $f_0(w), ..., f_m(w)$ be real-valued functions.
- The standard optimization formulation is

$$\begin{aligned} \underset{w \in S}{\text{minimize}} \quad & f_0(w) \\ \text{subject to} \quad & f_i(w) \leq 0, \ i = 1, \dots, m. \end{aligned}$$

- f_0 is the objective function, f_i; $i = 1, \dots, m$ the constraint functions.
- **Optimal Solution:** w^* - has the smallest value of f_0 among all the vectors that satisfy the constraints.

Reminder: Our Goal (1)

Prove the following:

Can rewrite the optimization problem

$$\min_{\mathbf{w}} \sum_{i=1}^{N} (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2 + \lambda \sum_{j=1}^{m} w_j^2$$

in the proper objective/constraint form:

$$\min_{\mathbf{w}} \sum_{i=1}^{N} (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2$$

subject to $\sum_{j=1}^{m} w_j^2 \leq t$

Our Goal (2)

Similarly for Lasso

$$\mathbf{w}^*_{\text{lasso}} = \arg\max_{\mathbf{w}} \left\{ -\sum_{i=1}^{N} (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2 - \lambda \sum_{j=1}^{m} |w_j| \right\}$$

$$\min_{\mathbf{w}} \sum_{i=1}^{N} (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2$$

subject to $\sum_{j=1}^{m} |w_j| \leq t$

Motivation

- Recall that the Ridge and Lasso objectives arose from considerations like:
  - Model Complexity
  - MLE with Gaussian/Laplacian Priors
- An objective of the adjacent form is much more direct.
  - Easier to interpret
  - Leads to optimization algorithms

$$\min_{\mathbf{w}} \sum_{i=1}^{N} (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2$$

$$\text{subject to} \sum_{j=1}^{m} |w_j| \leq t$$

Topics

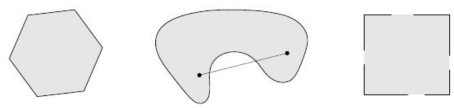
- Convex Sets
  – Intrinsic, Extrinsic Description
- Convex Functions
  – Equivalent Definitions
  – Algebra of Convex Functions
- Lagrangian Duality
  – Lagrange Dual Function ($\neq$ Lagrangian)

Convex Sets

- A convex set contains a segment between any two points in the set

$$w_1, w_2 \in S \implies \lambda w_1 + (1-\lambda) w_2 \in S$$

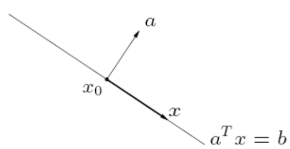
where $\lambda \in [0, 1]$.



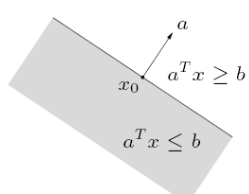
- Is $\{w \mid \|w\|_p \leq 1\}$ convex for $p = 1$? $p = 2$?
- What about $p = 0.5$?
- How about sets of the form $\{w \mid w^T x \leq b\}$?

Examples : Hyperplanes and Halfspaces

hyperplane: set of the form $\{x \mid a^T x = b\}$ $(a \neq 0)$



halfspace: set of the form $\{x \mid a^T x \leq b\}$ $(a \neq 0)$



- a is the normal vector
- hyperplanes are affine and convex; halfspaces are convex

Examples: Norm balls and norm cones

norm: a function $\|\cdot\|$ that satisfies

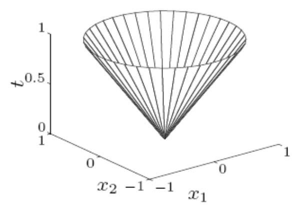
- $\|x\| \geq 0$; $\|x\| = 0$ if and only if $x = 0$
- $\|tx\| = |t| \, \|x\|$ for $t \in \mathbf{R}$
- $\|x + y\| \leq \|x\| + \|y\|$

notation: $\|\cdot\|$ is general (unspecified) norm; $\|\cdot\|_{\text{symb}}$ is particular norm

norm ball with center x_c and radius r: $\{x \mid \|x - x_c\| \leq r\}$

norm cone: $\{(x, t) \mid \|x\| \leq t\}$

Euclidean norm cone is called second-order cone

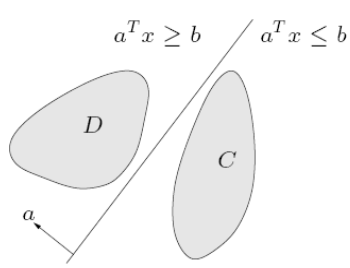


norm balls and cones are convex

Separating Hyperplane Theorem (Fundamental theorem of convex sets)

if C and D are disjoint convex sets, then there exists $a \neq 0$, b such that

$$a^T x \leq b \text{ for } x \in C, \qquad a^T x \geq b \text{ for } x \in D$$

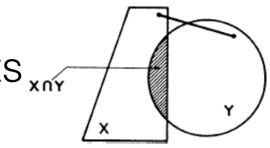


the hyperplane $\{x \mid a^T x = b\}$ separates C and D

strict separation requires additional assumptions (*e.g.*, C is closed, D is a singleton)

Algebra of Convex Sets

- The intersection of convex sets is a convex set



- The affine function of a convex set is convex.

suppose $f : \mathbf{R}^n \to \mathbf{R}^m$ is affine ($f(x) = Ax + b$ with $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$)

- the image of a convex set under f is convex

$$S \subseteq \mathbf{R}^n \text{ convex} \quad \Longrightarrow \quad f(S) = \{f(x) \mid x \in S\} \text{ convex}$$

- the inverse image $f^{-1}(C)$ of a convex set under f is convex

$$C \subseteq \mathbf{R}^m \text{ convex} \quad \Longrightarrow \quad f^{-1}(C) = \{x \in \mathbf{R}^n \mid f(x) \in C\} \text{ convex}$$

Convex Function

A function is convex on a convex set D iff.
For any two points $x_1, x_2 \in D$ and $0 \leq \lambda \leq 1$
$f[\lambda x_1 + (1-\lambda)x_2] \leq \lambda f(x_1) + (1-\lambda)\, x_2$

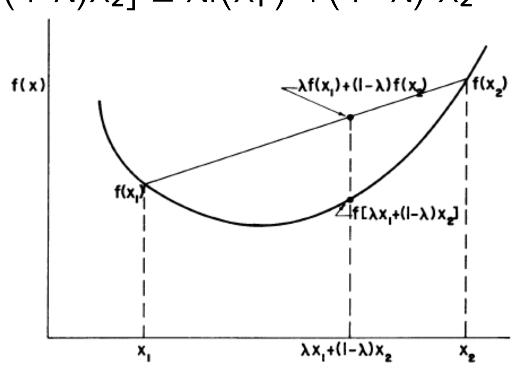


Figure B.1. Convex function.

Convex Function

A function is convex on a convex set D iff. its epigraph is a convex set.

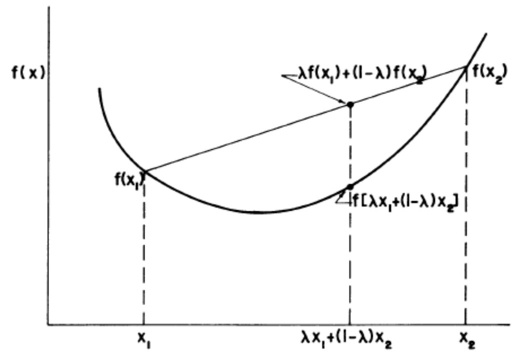


Figure B.1. Convex function.

Convex Function

A fnc. is (closed) convex on convex set D iff. it has a linear underestimator at every point.

$\forall x \in dom(f)\ \exists g_x : f(y) \geq f(x)+g_x^T(y-x)$ for all y

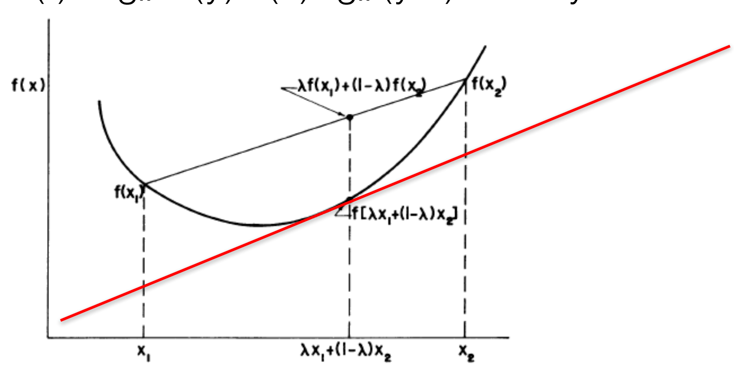


Figure B.1. Convex function.

More Characterizations of Convex Functions

1. First order characterization: Let f be a differentiable convex function.
   I. The first order directional derivative is always nondecreasing.
   II. $f(y) \geq f(x) + \nabla f(x)(y-x)$ $\forall$ $y, x \in \text{dom } f$ (When ∇f exists)
2. Second order characterization: The Hessian of a convex function (when it exists) is always positive semi-definite.

Local Optima $\Rightarrow$ Global Optima, since gradient can not be zero at two disconnected points.

Examples

- Affine functions : $w^T x + b$
- $\|w\|_p$ for $p \geq 1$

- logistic loss: $\log(1 + e^{-yw^T x})$ (why?)
- If $A \succeq 0$ then $\lambda_{\max}(A)$ (why?)
- If f and g are convex so is $\max\{f(x), g(x)\}$
- Is $e^{f(x)}$ convex if $f(x)$ is convex?

Operations that preserve convexity

1. Nonnegative weighted sum
2. Pre-composition with affine function
   – $g(x) = f(Ax + b)$ is convex if f is convex
3. Pointwise maximum and supremum
4. Post composition with monotonic increasing convex function

Subgradients and Subdifferential

- Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a convex function. Let $x \in \text{domain}(f)$, the normal of a linear under-estimator at x is a subgradient at x. g_x is a subgradient at x if $f(y) \geq f(x) + g_x^T(y-x)$
- The collection of all subgradients at x is called the subdifferential at x, and it is denoted $\partial f(x)$
  $\partial f(x) = \{g \in \mathbb{R}^d : f(y) \geq f(x) + g^T(y-x) \ \forall y \in \mathbb{R}^d \}$

Examples

- Let f(x) = ||x||₁ , ∂f(x) is a vector

$$[\partial f(x)]_i = \begin{cases} \{1\} \text{ if } x_i > 0 \\ [-1, 1] \text{ if } x_i = 0 \\ \{-1\} \text{if } x_i < 0 \end{cases}$$

$\partial f([-2, 0, 2]) = \{[a, b, c] \mid a \in \{-1\}, b \in [-1, 1], c \in \{1\}\}$

- Let $[x]_+$ denote the function f(x) = max(0, x)
  What is $\partial f[x]_+$?
- Let $f(x) = -[v - x]_+^2$, What is ∂f(x)?

Subdifferential Calculus

Theorem 3.60. Subdifferential calculus. The following are all true.

1. Let $f_1, f_2 : \mathbb{R}^d \to \mathbb{R}$ be convex functions and let $t_1, t_2 \geq 0$. Then
$$\partial(t_1 f_1 + t_2 f_2)(\mathbf{x}) = t_1 \partial f_1(\mathbf{x}) + t_2 \partial f_2(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathbb{R}^d.$$

2. Let $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$ and let $T(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ be the corresponding affine map from $\mathbb{R}^d \to \mathbb{R}^m$ and let $g : \mathbb{R}^m \to \mathbb{R}$ be a convex function. Then
$$\partial(g \circ T)(\mathbf{x}) = A^T \partial g(A\mathbf{x}) \text{ for all } \mathbf{x} \in \mathbb{R}^d.$$

3. Let $f_j : \mathbb{R}^d \to \mathbb{R}$, $j \in J$ be convex functions for some (possibly infinite) index set J, and let $f = \sup_{j \in J} f_j$. Then
$$\text{cl}(\text{conv}(\cup_{j \in J(\mathbf{x})} \partial f_j(\mathbf{x}))) \subseteq \partial f(\mathbf{x}),$$
where $J(\mathbf{x})$ is the set of indices j such that $f_j(\mathbf{x}) = f(\mathbf{x})$. Moreover, equality holds in the above relation, if one can impose a topology on J such that $J(\mathbf{x})$ is a compact set.

Unconstrained Optimization

(Fermat, 1629)

- Theorem: let $f\colon X \subseteq \mathbb{R}^N \to \mathbb{R}$ be a differentiable function. If f admits a local extremum at $x^* \in X$, then

$$\nabla f(x^*) = 0.$$

  - x^* is a stationary point.
  - a local minimum is a global minimum if the function is convex.

Unconstrained Optimization (Convex, Nondifferentiable)

- Let f : $\mathbb{R}^d \to \mathbb{R}$ be a convex function, x minimizes f if $0 \in \partial f(x)$

Constrained Optimization Problem

- Problem: Let $X \subseteq \mathbb{R}^N$ and $f, g_i : X \to \mathbb{R}$, $i \in [1, m]$. A constrained optimization problem has the form:

$$\min_{\mathbf{x} \in X} \quad f(\mathbf{x})$$
$$\text{subject to: } \quad g_i(\mathbf{x}) \leq 0, i \in [1, m].$$

  - no convexity assumption.
  - can be augmented with equality constraints.
  - primal problem.
  - optimal value p^*.

Lagrangian/Lagrange Function

- Definition: the Lagrange function or Lagrangian associated to a constraint problem is the function defined by:

$$\forall \mathbf{x} \in X, \forall \boldsymbol{\alpha} \geq 0, L(\mathbf{x}, \boldsymbol{\alpha}) = f(\mathbf{x}) + \sum_{i=1}^{m} \alpha_i g_i(\mathbf{x}).$$

  - α_is are called Lagrange or dual variables.

Lagrange Dual Function

- Definition: the (Lagrange) dual function associated to the constraint optimization problem is defined by

$$\forall \boldsymbol{\alpha} \geq 0, F(\boldsymbol{\alpha}) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \boldsymbol{\alpha})$$
$$= \inf_{\mathbf{x} \in X} f(\mathbf{x}) + \sum_{i=1}^{m} \alpha_i g_i(\mathbf{x}).$$

  - F is always concave: Lagrangian is linear with respect to α and $\inf$ preserves concavity.
  - $\forall \boldsymbol{\alpha} \geq 0, F(\boldsymbol{\alpha}) \leq p^*$: for a feasible $\mathbf{x}$,

$$f(\mathbf{x}) + \sum_{i=1}^{m} \alpha_i g_i(\mathbf{x}) \leq f(\mathbf{x}).$$

Dual Optimization Problem

- Definition: the dual (optimization) problem associated to the constraint optimization is

$$\max_{\boldsymbol{\alpha}} \quad F(\boldsymbol{\alpha})$$
$$\text{subject to:} \quad \boldsymbol{\alpha} \geq 0.$$

  - always a convex optimization problem.
  - optimal value d^*.

Weak and Strong Duality

- Weak duality: $d^* \leq p^*$.
  - always holds (clear from previous observations).
- Strong duality: $d^* = p^*$.
  - does not hold in general.
  - holds for convex problems with constraint qualifications.

Weak and Strong Duality

- Weak duality: $d^* \leq p^*$.
  - always holds (clear from previous observations).
- Strong duality: $d^* = p^*$.
  - does not hold in general.
  - holds for convex problems with constraint qualifications.

Proof Out of Scope

Putting It All Together

Let us start the proof that we

Can rewrite the optimization problem

$$\min_{\mathbf{w}} \sum_{i=1}^{N} (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2 + \lambda \sum_{j=1}^{m} w_j^2$$

in the proper objective/constraint form:

$$\min_{\mathbf{w}} \sum_{i=1}^{N} (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2$$

subject to $\sum_{j=1}^{m} w_j^2 \leq t$

Proof for Ridge Regression

Proof For Lasso Regression

Equivalence of Constrained and Unconstrained Forms of Lasso and Ridge Regression

Pushpendre Rastogi

February 24, 2017

The Lasso and Ridge regression problems can be stated in primal (constrained) and dual (unconstrained) forms.

	Primal (Constrained)	**Dual (Unconstrained)**
Lasso	$\mathrm{L}^{\mathrm{P}}(t) = \begin{array}{r} \arg\min\limits_{\beta} \frac{1}{2}\|\|y - X\beta\|\|_2^2 \\ \text{subject to} \|\|\beta\|\|_2 \leq t \end{array}$	$\mathrm{L}^{\mathrm{D}}(\lambda) = \arg\min_{\beta} \frac{1}{2}\|\|y - X\beta\|\|_2^2 + \lambda\beta$
Ridge	$\mathrm{R}^{\mathrm{P}}(t) = \begin{array}{r} \arg\min\limits_{\beta} \|\|y - X\beta\|\|_2^2 \\ \text{subject to } \|\|\beta\|\|_2^2 \leq t^2 \end{array}$	$\mathrm{R}^{\mathrm{D}}(\lambda) = \|\|y - X\beta\|\|_2^2 + \lambda\|\|\beta\|\|_2^2$

Table 1: Optimization Problems

In the following notes we will abuse notation and use the symbols

$\mathrm{L}^{\mathrm{P}}(t)$ etc. to refer to the optimization problem and the solution of the optimization problem. The meaning should be clear from context. We will prove the following 4 statements

1. Let $\lambda > 0$, $\exists t : \mathrm{R}^{\mathrm{P}}(t) = \mathrm{R}^{\mathrm{D}}(\lambda)$
2. Let $t > 0$, $\exists \lambda : \mathrm{R}^{\mathrm{P}}(t) = \mathrm{R}^{\mathrm{D}}(\lambda)$
3. Let $\lambda > 0$, $\exists t : \mathrm{L}^{\mathrm{P}}(t) = \mathrm{L}^{\mathrm{D}}(\lambda)$
4. Let $t > 0$, $\exists \lambda : \mathrm{L}^{\mathrm{P}}(t) = \mathrm{L}^{\mathrm{D}}(\lambda)$

In other words, for both lasso and ridge regression we will prove that for every constrained primal problem there exists an equivalent unconstrained dual problem and vice versa.

Proof 1— Let $\beta^* = \mathrm{R}^{\mathrm{D}}(\lambda)$. Since $\mathrm{R}^{\mathrm{D}}(\lambda)$ is an unconstrained convex minimization problem therefore $\left.\frac{d\mathrm{R}^{\mathrm{D}}(\lambda)}{d\beta}\right|_{\beta^*} = 0$.

$$\left.\frac{d\mathrm{R}^{\mathrm{D}}(\lambda)}{d\beta}\right|_{\beta^*} = 2(X\beta^* - y)^T X + \lambda \beta^{*T} = 0 \tag{1}$$

$$\implies \beta^{*T}(X^T X + \lambda I) = y^T x \tag{2}$$

$$\implies \mathrm{R}^{\mathrm{D}}(\lambda) = \beta^* = (X^T X + \lambda I)^{-1} X^T y \tag{3}$$

Now consider the problem $\mathrm{R}^{\mathrm{P}}(t)$ for a general value of t. Since $\mathrm{R}^{\mathrm{P}}(t)$ is a constrained minimization problem with a convex constraint, therefore assuming some *constraint qualifications*, the optimization problem $\mathrm{R}^{\mathrm{P}}(t)$ should have zero duality gap, therefore we can solve this problem by maximizing the lagrange dual function.

Let $\mathcal{L}_t^{\mathrm{R}}(\beta,\gamma)$ denote the lagrangian of $\mathrm{R}^{\mathrm{P}}(t)$, i.e.

$$\mathcal{L}_t^{\mathrm{R}}(\beta,\gamma) = ||y - X\beta||_2^2 + \gamma(||\beta||_2^2 - t^2)$$

Let $\mathcal{G}_t^{\mathrm{R}}(\gamma)$ denote the lagrangian dual function, i.e.

$$\mathcal{G}_t^{\mathrm{R}}(\gamma) = \inf_{\beta} ||y - X\beta||_2^2 + \gamma(||\beta||_2^2 - t^2)$$

We can simplify the above expression to get $\mathcal{G}_t^{\mathrm{R}}(\gamma) = \mathcal{L}_t^{\mathrm{R}}(\mathrm{R}^{\mathrm{D}}(\lambda),\gamma)$. Now we must maximize the dual function $\mathcal{G}_t^{\mathrm{R}}(\gamma)$ with respect to γ. Note that

$$\left.\frac{d\mathrm{R}^{\mathrm{D}}(\gamma)}{d\gamma}\right|_{\gamma^*} = -(X^TX + \gamma^* I)^{-1}\mathrm{R}^{\mathrm{D}}(\gamma^*)$$

We equate $\left.\frac{d\mathcal{G}_t^{\mathrm{R}}(\gamma)}{d\gamma}\right|_{\gamma^*}$ with 0 to get

$$\begin{aligned}
0 &= 2(X\mathrm{R}^{\mathrm{D}}(\gamma^*) - y)^T X \left.\frac{d\mathrm{R}^{\mathrm{D}}(\gamma)}{d\gamma}\right|_{\gamma^*} + ||\mathrm{R}^{\mathrm{D}}(\gamma^*)||_2^2 - t^2 \\
&\quad + \gamma^*(2\mathrm{R}^{\mathrm{D}}(\gamma^*)^T \left.\frac{d\mathrm{R}^{\mathrm{D}}(\gamma)}{d\gamma}\right|_{\gamma^*}) && (4)\\
&= 2(X^T(X\mathrm{R}^{\mathrm{D}}(\gamma^*) - y) + \gamma^*\mathrm{R}^{\mathrm{D}}(\gamma^*))^T \left.\frac{d\mathrm{R}^{\mathrm{D}}(\gamma)}{d\gamma}\right|_{\gamma^*} + ||\mathrm{R}^{\mathrm{D}}(\gamma^*)||_2^2 - t^2 && (5)\\
&= 2((X^TX + \gamma^* I)\mathrm{R}^{\mathrm{D}}(\gamma^*) - X^Ty)^T \left.\frac{d\mathrm{R}^{\mathrm{D}}(\gamma)}{d\gamma}\right|_{\gamma^*} + ||\mathrm{R}^{\mathrm{D}}(\gamma^*)||_2^2 - t^2 && (6)\\
&= ||\mathrm{R}^{\mathrm{D}}(\gamma^*)||_2^2 - t^2 && (7)\\
\implies t &= ||\mathrm{R}^{\mathrm{D}}(\gamma^*)||_2 && (8)
\end{aligned}$$

Now for a given value of λ we know the values of $\mathrm{R}^{\mathrm{D}}(\lambda)$. If we let $t = ||\beta^*||_2 = \mathrm{R}^{\mathrm{D}}(\lambda)$ then $\mathrm{R}^{\mathrm{D}}(\lambda) = \mathrm{R}^{\mathrm{D}}(\gamma^*) \implies \lambda = \gamma^*$. This means that the solution of the primal problem will be equal to $\mathrm{R}^{\mathrm{D}}(\lambda)$. $\square$

Proof 2— Let $\beta^* = \mathrm{R}^{\mathrm{P}}(t)$, Chose λ such that $||\mathrm{R}^{\mathrm{D}}(\lambda)||_2 = t$. Then $\mathrm{R}^{\mathrm{D}}(\lambda) = \beta^*$.

Proof 3— Let $\beta^* = \mathrm{L}^{\mathrm{D}}(\lambda)$. Since $\mathrm{L}^{\mathrm{D}}(\lambda)$ is the solution of an unconstrained, convex, but non-differentiable minimization problem, therefore at the solution, the subgradient of $||y - X\beta||_2 + \lambda||\beta||_1$ contains zero.

The subgradient of $\mathrm{L}^{\mathrm{D}}(\lambda)$ is

$$(X\beta - y)^T X + \lambda \frac{\partial||\beta||_1}{\partial\beta}$$

In order to proceed with the proof we make the assumption that $X^T X = I$. This implies that the subgradient of

$$\mathrm{L}^{\mathrm{D}}(\lambda) = (\beta^T - y^T X) + \lambda \frac{\partial||\beta||_1}{\partial\beta}$$

Let $v = y^T X \implies v \in \beta^T + \lambda \frac{\partial||\beta||_1}{\partial\beta}$.

Now consider the i^{th} component of v. $v_i \in \beta_i + \lambda[\frac{\partial||\beta||_1}{\partial\beta}]_i$.

This implies that

$$\begin{cases} \beta_i > 0 & \implies v_i > \lambda \\ \beta_i = 0 & \implies v_i \in [-\lambda, \lambda] \\ \beta_i < 0 & \implies v_i < -\lambda \end{cases} \tag{9}$$

We can rewrite the above analysis as

$$\begin{cases} v_i > \lambda & \implies \beta_i = v_i - \lambda \\ v_i \in [-\lambda, \lambda] & \implies \beta_i = 0 \\ v_i < -\lambda & \implies \beta_i = v_i + \lambda \end{cases} \tag{10}$$

This function is called the shrinkage operator.

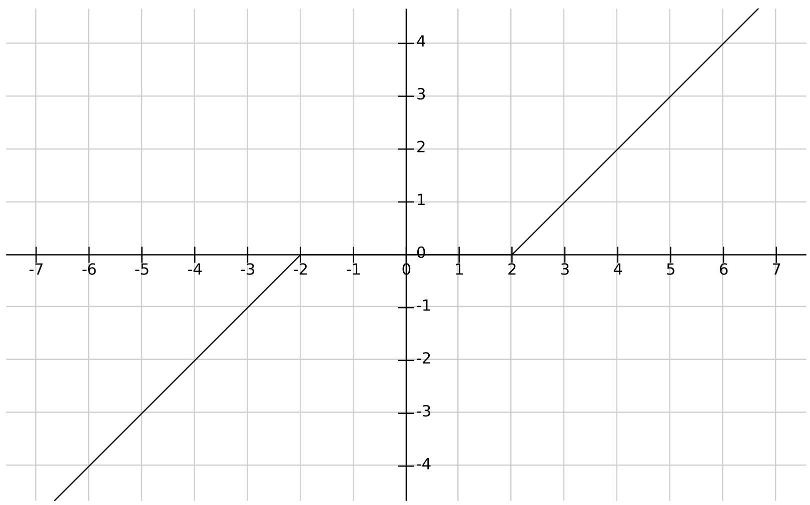


Figure 1: Plot of the $\text{shrink}_2(x)$ function.

$$\mathrm{L}^{\mathrm{D}}(\lambda) = \beta^* = \text{shrink}_\lambda(X^T y) = \text{shrink}_\lambda(v^T)$$

Now consider the lagrangian $\mathcal{L}_t(\beta, \gamma)$ of the primal problem $\mathrm{L}^{\mathrm{P}}(t)$.

$$\mathcal{L}_t(\beta, \gamma) = \frac{1}{2}||y - X\beta||_2 + \gamma(||\beta||_1 - t) \tag{11}$$
$$= \frac{1}{2}(y^t y + \sum \beta_i^2 - \sum v_i\beta_i) + \gamma \sum |\beta_i| - \gamma t \tag{12}$$
$$= \frac{y^t y}{2} - \gamma t + \frac{1}{2}\sum_i (\beta_i^2 - 2v_i\beta_i + 2\gamma|\beta_i|) \tag{13}$$

Let $\mathcal{G}_t(\gamma)$ be the lagrangian dual function. Its value is

$$\mathcal{G}_t(\gamma) = \inf_\beta \frac{1}{2}||y - X\beta||_2 + \gamma(||\beta||_1 - t)$$

.

Using similar argument as earlier, the value of the dual function is $\mathcal{L}_t(\mathrm{L}^{\mathrm{D}}(\gamma), \gamma) = \mathcal{L}_t(\mathrm{shrink}_\gamma(v), \gamma)$.

$$\mathcal{L}_t(\mathrm{shrink}_\gamma(v), \gamma) = \frac{y^t y}{2} - \gamma t + \frac{1}{2}\sum_i ((|v_i| - \gamma)_+^2 - 2|v_i|(|v_i| - \gamma)_+ + 2\gamma(|v_i| - \gamma)_+) \tag{14}$$
$$= \frac{y^t y}{2} - \gamma t + \frac{1}{2}\sum_i ((|v_i| - \gamma)_+^2 - 2(|v_i| - \gamma)(|v_i| - \gamma)_+ \tag{15}$$
$$= \frac{y^t y}{2} - \gamma t - \frac{1}{2}\sum_i (|v_i| - \gamma)_+^2 \tag{16}$$

This function is **DIFFERENTIABLE** with respect to γ. Now the derivative of $\mathcal{L}_t(\mathrm{shrink}_\gamma(v), \gamma)$ with respect to γ is

$$\frac{d\mathcal{L}_t(\mathrm{shrink}_\gamma(v), \gamma)}{d\gamma} = -t + \sum_i (|v_i| - \gamma \text{ if } |v_i| > \gamma \text{ else } 0) \tag{17}$$

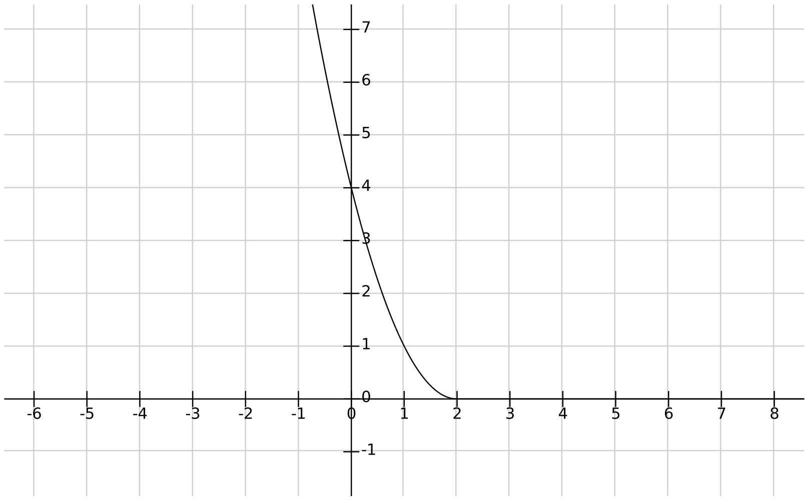


Figure 2: The differentiable function

This derivative should be zero, which gives us the value of γ as a function of t. Now if $t = ||\text{shrink}_\lambda(v)||_1$ then $\gamma = \lambda$. Therefore we have constructed a primal problem, $\mathrm{L^P}(t)$, corresponding to $\mathrm{L^D}(\lambda)$. $\square$