Penalty method is not effective for nonnegative least sqaure

Softmax function and nonnegative penalty

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be Homepage: angms.science

First draft : December 31, 2017
Last update : February 19, 2020

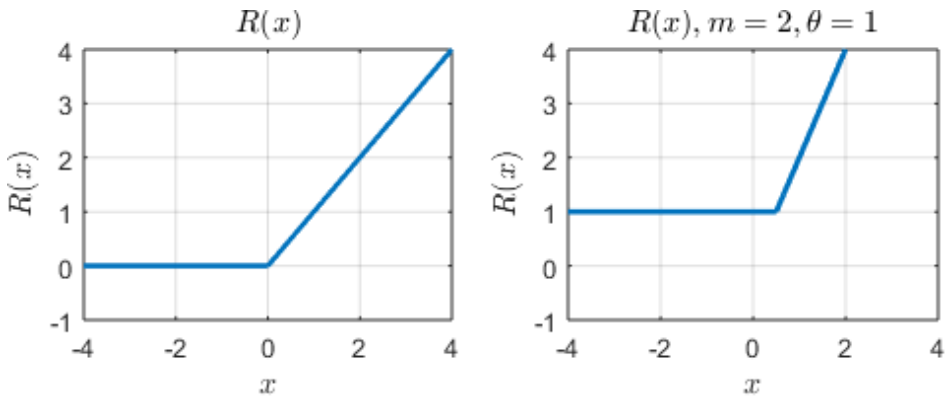
The ramp function

The definition of ramp function

$$R(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

With shift θ and slope m, the generalized ramp function is

$$R(x; \theta, m) = \begin{cases} mx & \text{if } x \geq \theta \\ \theta & \text{if } x < \theta \end{cases}$$



The max function can be seen as a ramp function with $m = 1$ and $\theta = 0$.

$$\max(x, 0) = R(x; 0, 0) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}.$$

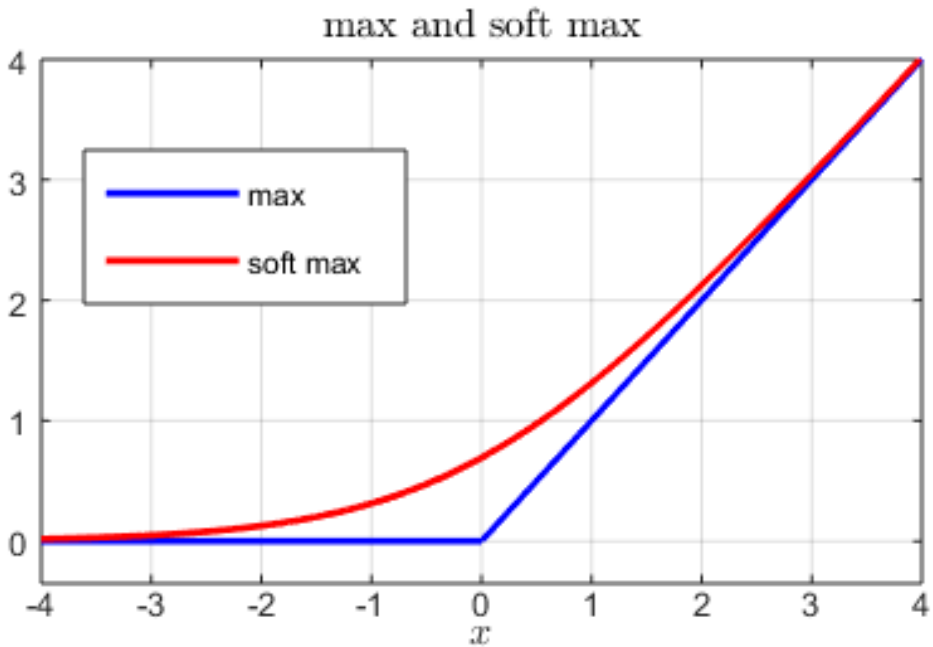
Softmax

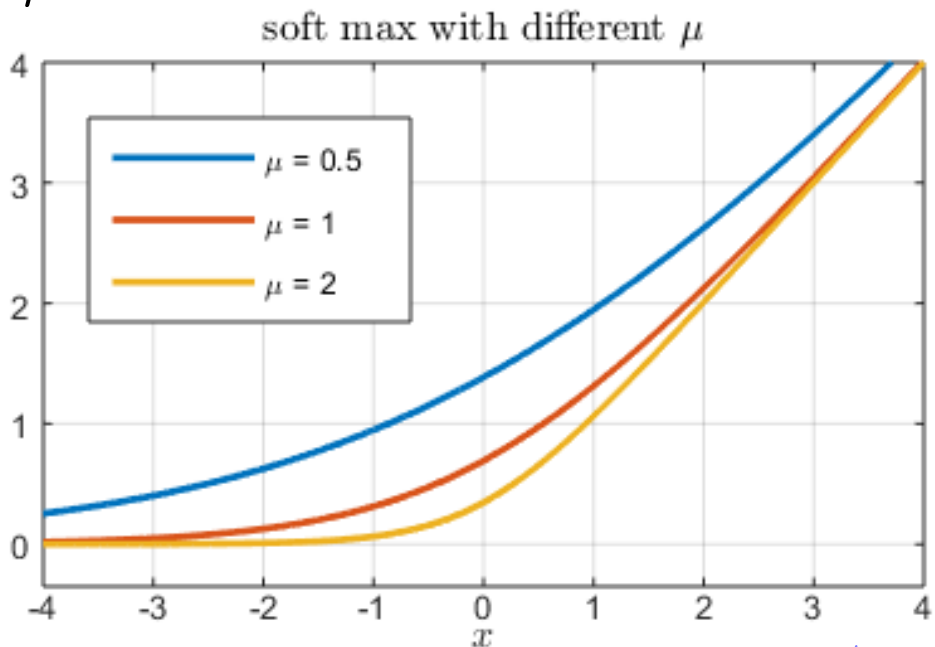
Max function is *singular* : it is not differentiable at 0.
We can approximate max by *softmax* which is smooth (differentiable).
Together with parameters θ, m, softmax have one more parameter μ

$$\text{softmax}(x;\mu,\theta,m) = \frac{1}{\mu}\log(e^{\mu m x} + e^{\mu\theta});$$

When $m = 1$, $\theta = 0$, we have

$$\text{softmax}(x;\mu) = \frac{1}{\mu}\log(e^{\mu x} + 1);$$





The derivative of softmax

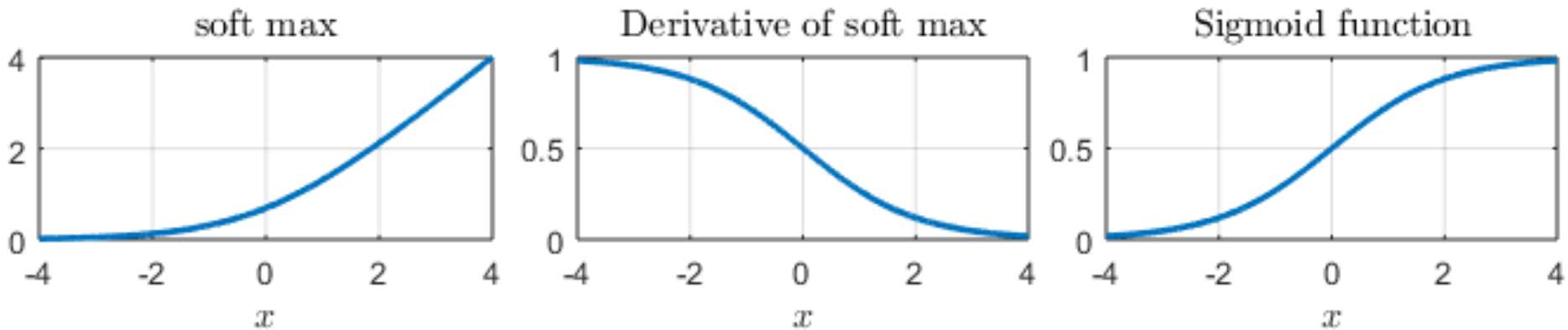
The softmax function

$$\text{softmax}(x;\mu) = \frac{1}{\mu}\log(e^{\mu x}+1);$$

is differentiable. Using chain rule, we have

$$\frac{d}{dx}\text{softmax}(x;\mu) = \frac{1}{\mu}\frac{d}{dx}\log(e^{\mu x}+1) = \frac{1}{\mu}\frac{\mu}{e^{\mu x}+1} = \frac{1}{e^{\mu x}+1},$$

which is the *inverted sigmoid function*.



Application to non-negativity constraint

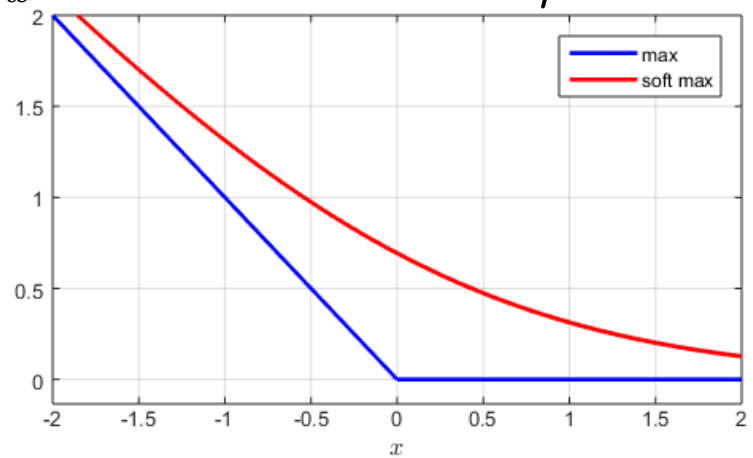
$$x = \underset{x \geq 0}{\operatorname{argmin}}\ f(x).$$

Such problem can be approximate by an unconstrained problem that put the constraint into the objective function as penalty

$$x = \underset{x}{\operatorname{argmin}}\ f_\lambda(x) = f(x) + \lambda g(x), \quad g(x) = \max\{-x, 0\}.$$

Penalty term $\max\{-x, 0\}$ means when x violates the constraint (x is negative), then $f_\lambda(x)$ increases. Here the amount of penalty is directly proportional to the size of violation scaled by the penalty parameter λ. Replacing the non-differentiable max by smooth softmax we have

$$x = \underset{x}{\operatorname{argmin}}\ f_\lambda(x) = f(x) + \frac{\lambda}{\mu} \log(e^{-\mu x} + 1).$$



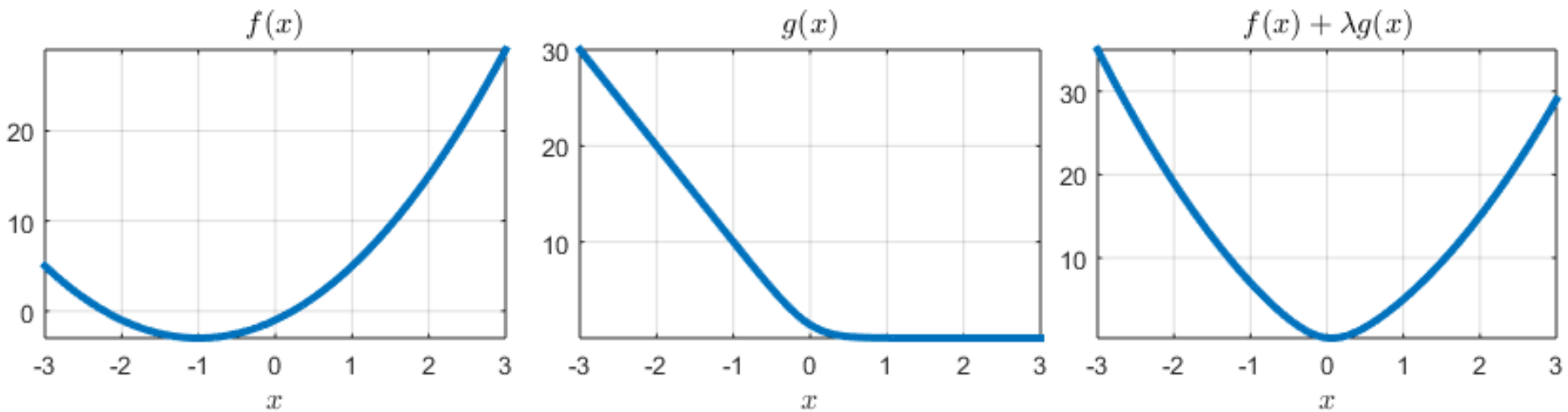
Example

Let $f(x) = 2x^2 + 4x - 1$ with minimizer at $x = -1$.
The original problem is $\min_{x \geq 0} f(x)$. Plot shows the optimal x is at $x = 0$.

Let $g(x)$ be the penalty function, which is the softmax, also set $\lambda = \mu = 5$.
The penalized objective function is

$$f_\lambda(x) = f(x) + \lambda g(x) = 2x^2 + 4x - 1 + \log(e^{-5x} + 1).$$

If we minimize the penalized objective function, we get optimal x at $x = 0.06$, which is close to the true solution of the original problem.



Application to nonnegative least squares

Non**n**egative **L**east **S**quares (NNLS) : given $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$, find $\mathbf{x} \in \mathbb{R}^n_+$ by solving

$$(\mathcal{P}) \ : \ \operatorname*{argmin}_{\mathbf{x}\geq 0} f(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2.$$

Using softmax penalty, we have the penalized problem

$$(\mathcal{P}') \ : \ \operatorname*{argmin}_{\mathbf{x}} f_{\lambda_i,\mu_i}(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + \sum_{i=1}^n \frac{\lambda_i}{\mu_i}\log(e^{-\mu_i x_i}+1).$$

If we let all λ_i equal to the same constant, and let μ_i equal to the same constant, we have

$$(\mathcal{P}') \ : \ \operatorname*{argmin}_{\mathbf{x}} f_{\lambda,\mu}(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + \frac{\lambda}{\mu}\sum_{i=1}^n \log(e^{-\mu_i x_i}+1).$$

Solving the softmax penalized least squares

$$(\mathcal{P}') \ : \ \operatorname*{argmin}_{\mathbf{x}} \ f_{\lambda,\mu}(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + \frac{\lambda}{\mu}\sum_{i=1}^{n}\log(e^{-\mu_i x_i}+1).$$

Denote the column of $\mathbf{A}$ as $\mathbf{a}_i$, then we have

$$f_{\lambda,\mu}(\mathbf{x}) = \frac{1}{2}\|\sum_{i=1}^{n} x_i\mathbf{a}_i - \mathbf{b}\|_2^2 + \frac{\lambda}{\mu}\sum_{i=1}^{n}\log(e^{-\mu_i x_i}+1).$$

Focusing on the i^{th} term, we have

$$f_{\lambda}(x_i) = \frac{1}{2}\,\|x_i\mathbf{a}_i - \mathbf{b}_{-i}\|_2^2 + \frac{\lambda}{\mu}\log(e^{-\mu x_i}+1) + c.$$

where $\mathbf{b}_{-i} = \sum_{j\neq i} x_j\mathbf{a}_j - \mathbf{b}$. Expand it we get

$$f_{\lambda}(x_i) = \frac{1}{2}\|\mathbf{a}_i\|_2^2 x_i^2 - \mathbf{a}_i^\top\mathbf{b}_{-i}x_i + \frac{\lambda}{\mu}\log(e^{-\mu x_i}+1) + c.$$

Solving the softmax penalized least squares

We now arrive at a coordinate descent with the following componentwise subproblem

$$x_i = \operatorname*{argmin}_x \ f_{\lambda,\mu}(x) = \frac{1}{2}\|\mathbf{a}_i\|_2^2 x^2 - \mathbf{a}_i^\top \mathbf{b}_{-i} x + \frac{\lambda}{\mu}\log(e^{-\mu x} + 1),$$

in which the objective function is highly non-linear.

To solve it, we can solve $\frac{\partial}{\partial x} f_{\lambda,\mu}(x) = 0$. The derivative is

$$\frac{\partial}{\partial x} f_{\lambda,\mu}(x) = \|\mathbf{a}_i\|_2^2 x_i - \mathbf{a}_i^\top \mathbf{b}_{-i} + \frac{\lambda}{\mu}\frac{1}{e^{-\mu x_i} + 1}.$$

One can see that solving $\frac{\partial}{\partial x} f_{\lambda,\mu}(x) = 0$ requires to find the root of a highly non-linear equation, we have to run an numerical solver to solve it. The cost for such solver may costs more than a simple projection step in PGD-based NNLS.

In other words, penalty method on NNLS may not be a good choice.

Last page - summary

- Ramp function.
- Max function.
- Softmax function.
- Softmax as penalty term on nonnegative constrained problem.
- Softmax penalty method may not be effective on nonnegative constrained least squares.

End of document