Fast Optimization Methods for L1 Regularization: A Comparative Study and Two New Approaches Suplemental Material

Mark Schmidt [1], Glenn Fung[2], Romer Rosales[2]

[1] Department of Computer Science University of British Columbia,
[2] IKM, Siemens Medical Solutions, USA

1 Smooth L1-norm aproximation

In order to deal with the non-differentiable penalty, we propose a smooth approximation to the L1 penalty based on the following:

(i) $|x| = (x)_+ + (-x)_+$, where the *plus* function is $(x)_+ = max\{x, 0\}$
(ii) The plus function can be approximated (smoothly)[2], by the integral to a smooth approximation of the sigmoid function:

$$(x)_+ \approx p(x, \alpha) = x + \frac{1}{\alpha}\log(1 + \exp(-\alpha x)) \tag{1}$$

Combinining these, we arrive to the following smooth approximation for the absolute value function consisting of the sum of the integral of two sigmoid functions (Fig. 1 plots this approximation near 0 for different values of α):

$$\begin{aligned} |x| &= (x)_+ + (-x)_+ \approx p(x, \alpha) + p(-x, \alpha) \\ &= \tfrac{1}{\alpha}\left[\log(1 + \exp(-\alpha x)) + \log(1 + \exp(\alpha x))\right] \\ &= |x|_\alpha \end{aligned} \tag{2}$$

In practice, $\alpha = 10^6$ yields results that are within some small tolerance of the results produced by (optimal) constrained optimization methods. As opposed to the L1-penalty, this approximation is amenable to standard unconstrained optimization methods since it is twice-differentiable:

$$\nabla(|x|) \approx (1 + \exp(-\alpha x))^{-1} - (1 + \exp(\alpha x))^{-1} \tag{3}$$

$$\nabla^2(|x|) \approx 2\alpha \exp(\alpha x)/(1 + \exp(\alpha x))^2 \tag{4}$$

This approximation can be used in conjuction with any general likelihood or loss functions. Next we will show that for optimization problems derived from learning methods with L1 regularization, the solutions of the smooth approximated problems approach the solution to the original problems when α approaches infinity. We will start by proposing and proving a simple lemma similar to one proposed in [1] for the plus function. This gives us a bound relating $|x|$ and $|x|_\alpha$.

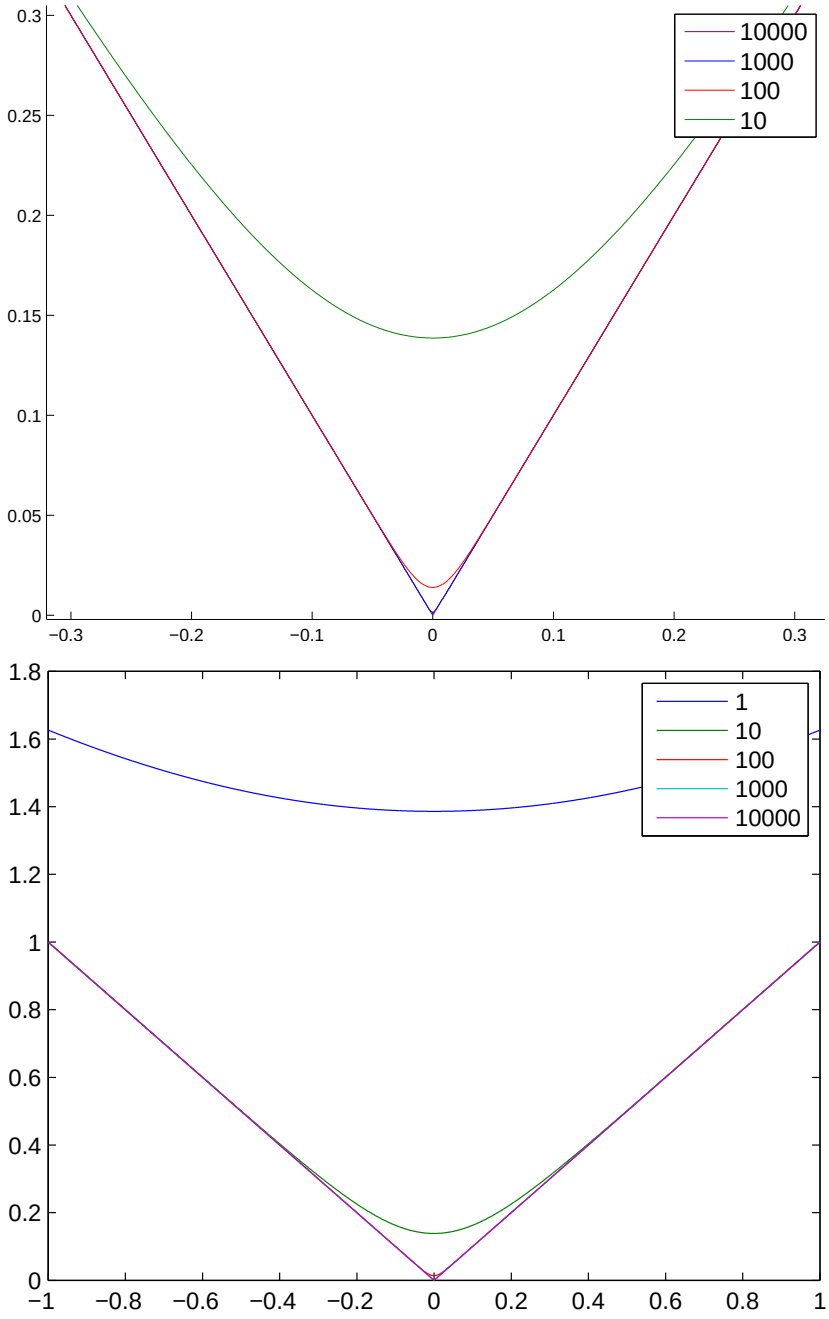

Fig. 1. Approximation near 0 for different settings of the parameter α

Lemma 11 Approximation bound *For any $x \in \Re$ and for any $\alpha > 0$:*

$$||x| - |x|_\alpha| \leq 2\frac{\log 2}{\alpha} \tag{5}$$

Proof:
Lets consider two cases. For $x > 0$,

$$\begin{aligned} p(x,\alpha) - (x)_+ &= x + \frac{1}{\alpha}\log(1+\exp(-\alpha x)) - x \\ &= \frac{1}{\alpha}\log(1+\exp(-\alpha x)) \leq \frac{\log 2}{\alpha} \end{aligned} \tag{6}$$

For $x \leq 0$,

$$0 \leq p(x,\alpha) - (x)_+ = p(x,\alpha) \leq p(0,\alpha) = \frac{\log 2}{\alpha}, \tag{7}$$

where the last inequality derives from the fact that p is a monotonically incresing function. Hence, from equations (6) and (7) and because $p(x,\alpha)$ always dominates $(x)_+$, we conclude that:

$$|p(x,\alpha) - (x)_+| \leq \frac{\log 2}{\alpha} \tag{8}$$

Since $|x| = (x)_+ + (-x)_+$, using equation (1) we have:

$$\begin{aligned} ||x| - |x|_\alpha| &= |p(x,\alpha) + p(-x,\alpha) - ((x)_+ + (-x)_+)| \\ &\leq |p(x,\alpha) - (x)_+| + |p(-x,\alpha) - (-x)_+| \\ &\leq \tfrac{\log 2}{\alpha} + \tfrac{\log 2}{\alpha} = 2\tfrac{\log 2}{\alpha}.\square \end{aligned} \tag{9}$$

Let us now define $\|x\|_{(1,\alpha)}$ as a smooth approximation to the 1-norm function $\|x\|_1$ for a vector $x \in \Re^n$ in the following way: $\|x\|_{(1,\alpha)} = \sum_i^n |x_i|_\alpha$.

Then, using Lemma 11 we have that

$$\left|\|x\|_{(1,\alpha)} - \|x\|_1\right| \leq 2n\tfrac{\log 2}{\alpha} \tag{10}$$

Hence, we can conclude that:

$$\lim_{\alpha\to\infty} \|x\|_{(1,\alpha)} = \|x\|_1 \quad \forall x \in \Re^n \tag{11}$$

Letting $L : \Re^n \to \Re$ be any continuous loss function and defining $f(x) = L(x) + \|x\|_1$ and $f_\alpha(x) = L(x) + \|x\|_{(1,\alpha)}$. Let us also define $\bar{x} = \arg\min_x f(x)$ and $\bar{x_\alpha} = \arg\min_x f_\alpha(x)$. By the definition of f and f_α and by equation (11), it follows that:

$$\lim_{\alpha\to\infty} f_\alpha(x) = f(x) \quad \forall x \in \Re^n \tag{12}$$

In adition, we know that $f(\bar{x}) \leq f(x)$, $\forall x$. In particular $f(\bar{x}) \leq f(\bar{x_\alpha})$, then:

$$\begin{aligned} f(\bar{x}) \leq f(\bar{x_\alpha}) &= L(\bar{x_\alpha}) + \|\bar{x_\alpha}\|_1 \\ &= L(\bar{x_\alpha}) + \|\bar{x_\alpha}\|_1 + \|\bar{x_\alpha}\|_{(1,\alpha)} - \|\bar{x_\alpha}\|_{(1,\alpha)} \\ &= \left(L(\bar{x_\alpha}) + \|\bar{x_\alpha}\|_{(1,\alpha)}\right) + \left(\|\bar{x_\alpha}\|_1 - \|\bar{x_\alpha}\|_{(1,\alpha)}\right) \\ &= f_\alpha(\bar{x_\alpha}) + \left(\|\bar{x_\alpha}\|_1 - \|\bar{x_\alpha}\|_{(1,\alpha)}\right) \end{aligned} \tag{13}$$

This implies $f(\bar{x}) - f_\alpha(\bar{x_\alpha}) \geq -2n\frac{\log 2}{\alpha}$ (using equation (10)). Similarly, we can prove that $f(\bar{x}) - f_\alpha(\bar{x_\alpha}) \leq 2n\frac{\log 2}{\alpha}$, hence: $\lim_{\alpha\to\infty} f_\alpha(\bar{x_\alpha}) = f(\bar{x})$.
Furthermore:

$$\begin{aligned} |f(\bar{x_\alpha}) - f(\bar{x})| &= |f(\bar{x_\alpha}) - f(\bar{x}) - f_\alpha(\bar{x_\alpha}) + f_\alpha(\bar{x_\alpha})| \\ &\leq |f(\bar{x_\alpha}) - f_\alpha(\bar{x_\alpha})| + |f_\alpha(\bar{x_\alpha}) - f(\bar{x})| \end{aligned} \tag{14}$$

This implies that: $\lim_{\alpha\to\infty} f(\bar{x_\alpha}) = f(\bar{x})$. Moreover, if L is stricly convex, it is easy to proove that: $\lim_{\alpha\to\infty} x_\alpha = \bar{x}$

References

1. Chunhui Chen and O. L. Mangasarian. A class of smoothing functions for nonlinear and mixed complementarity problems. *Computational Optimization and Applications*, 5(2):97–138, 1996.
2. Y.-J. Lee and O. L. Mangasarian. SSVM: A smooth support vector machine. *Comp. Opt. and App.*, 20:5–22, 2001.