Smoothing for nonsmooth optimization

Yuxin Chen

Princeton University, Fall 2019

Outline

- Smoothing
- Smooth approximation
- Algorithm and convergence analysis

Nonsmooth optimization

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x})$$

where f is convex but not always differentiable

- subgradient methods yield ε-accuracy in

$$O\left(\frac{1}{\varepsilon^2}\right) \text{ iterations}$$

- in contrast, if f is smooth, then accelerated GD yields ε-accuracy in

$$O\left(\frac{1}{\sqrt{\varepsilon}}\right) \text{ iterations}$$

— *significantly better than the nonsmooth case*

Lower bound

— *Nemirovski & Yudin '83*

If one only has access to $\underbrace{\text{the first-order oracle}}_{\text{black box model}}$ (which takes as inputs a point x and outputs a subgradient of f at x), then one cannot improve upon $O\left(\frac{1}{\varepsilon^2}\right)$ in general

Nesterov's smoothing idea

Practically, we rarely meet pure black box models; rather, we know something about the structure of the underlying problems

One possible strategy is:

1. approximate the nonsmooth objective by a smooth function
2. optimize the smooth approximation instead (using, e.g., Nesterov's accelerated method)

Smooth approximation

Smooth approximation

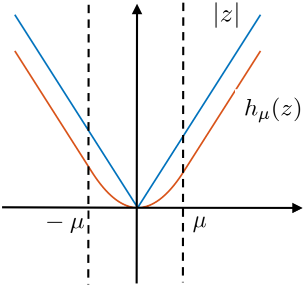
A convex function f is called (α, β)-smoothable if, for any $\mu > 0$, $\exists$ convex function f_μ s.t.

- $f_\mu(\boldsymbol{x}) \leq f(\boldsymbol{x}) \leq f_\mu(\boldsymbol{x}) + \beta\mu$, $\forall \boldsymbol{x}$ (approximation accuracy)
- f_μ is $\frac{\alpha}{\mu}$-smooth (smoothness)

— *μ: tradeoff between approximation accuracy and smoothness*

Here, f_μ is called a $\frac{1}{\mu}$-smooth approximation of f with parameters (α, β)

Example: ℓ_1 norm



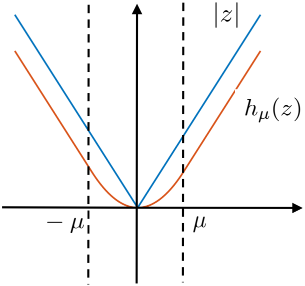
Consider the Huber function

$$h_\mu(z) = \begin{cases} z^2/2\mu, & \text{if } |z| \leq \mu \\ |z| - \mu/2, & \text{else} \end{cases}$$

which satisfies

$$h_\mu(z) \leq |z| \leq h_\mu(z) + \mu/2 \qquad \text{and} \qquad h_\mu(z) \text{ is } \frac{1}{\mu}\text{-smooth}$$

Example: ℓ_1 norm

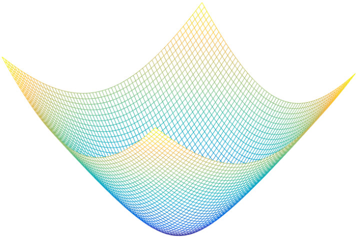


Therefore, $f_\mu(\boldsymbol{x}) := \sum_{i=1}^{n} h_\mu(x_i)$ is $\frac{1}{\mu}$-smooth and obeys

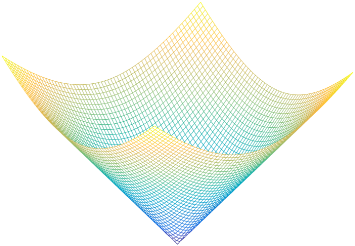
$$f_\mu(\boldsymbol{x}) \leq \|\boldsymbol{x}\|_1 \leq f_\mu(\boldsymbol{x}) + \frac{n\mu}{2}$$

$$\Longrightarrow \quad \|\cdot\|_1 \text{ is } (1, n/2)\text{-smoothable}$$

Example: ℓ_2 norm



$f_\mu(\boldsymbol{x})$



$\|\boldsymbol{x}\|_2$

Consider $f_\mu(\boldsymbol{x}) := \sqrt{\|\boldsymbol{x}\|_2^2 + \mu^2} - \mu$, then for any $\mu > 0$ and any $\boldsymbol{x} \in \mathbb{R}^n$,

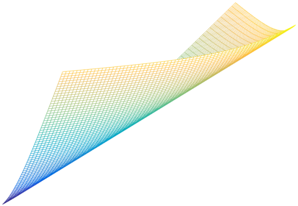
$$f_\mu(\boldsymbol{x}) \leq \big(\|\boldsymbol{x}\|_2 + \mu\big) - \mu = \|\boldsymbol{x}\|_2$$

$$\|\boldsymbol{x}\|_2 \leq \sqrt{\|\boldsymbol{x}\|_2^2 + \mu^2} = f_\mu(\boldsymbol{x}) + \mu$$

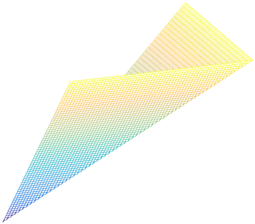
In addition, $f_\mu(\boldsymbol{x})$ is $\frac{1}{\mu}$-smooth (exercise)

Therefore, $\|\cdot\|_2$ is (1,1)-smoothable

Example: max function



$f_\mu(\boldsymbol{x})$



$\max\{x_1, x_2\}$

Consider $f_\mu(\boldsymbol{x}) := \mu \log\left(\sum_{i=1}^n e^{x_i/\mu}\right) - \mu \log n$, then $\forall \mu > 0$ and $\forall \boldsymbol{x} \in \mathbb{R}^n$,

$$f_\mu(\boldsymbol{x}) \le \mu \log\left(n \max_i e^{x_i/\mu}\right) - \mu \log n = \max_i x_i$$

$$\max_i x_i \le \mu \log\left(\sum_{i=1}^n e^{x_i/\mu}\right) = f_\mu(\boldsymbol{x}) + \mu \log n$$

In addition, $f_\mu(\boldsymbol{x})$ is $\frac{1}{\mu}$-smooth (exercise). Therefore, $\max_{1\le i\le n} x_i$ is (1,$\log n$)-smoothable

Basic rules: addition

- $f_{\mu,1}$ is a $\frac{1}{\mu}$-smooth approximation of f_1 with parameters (α_1, β_1)
- $f_{\mu,2}$ is a $\frac{1}{\mu}$-smooth approximation of f_2 with parameters (α_2, β_2)

$\Longrightarrow \quad \lambda_1 f_{\mu,1} + \lambda_2 f_{\mu,2}$ $(\lambda_1, \lambda_2 > 0)$ is a $\frac{1}{\mu}$-smooth approximation of $\lambda_1 f_1 + \lambda_2 f_2$ with parameters $(\lambda_1\alpha_1 + \lambda_2\alpha_2, \lambda_1\beta_1 + \lambda_2\beta_2)$

Basic rules: affine transformation

- h_μ is a $\frac{1}{\mu}$-smooth approximation of h with parameters (α, β)
- $f(\boldsymbol{x}) := h(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$

$\Longrightarrow$ $h_\mu(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ is a $\frac{1}{\mu}$-smooth approximation of f with parameters $(\alpha\|\boldsymbol{A}\|^2, \beta)$

Example: $\|\boldsymbol{Ax}+\boldsymbol{b}\|_2$

Recall that $\sqrt{\|\boldsymbol{x}\|_2^2+\mu^2}-\mu$ is a $\frac{1}{\mu}$-smooth approximation of $\|\boldsymbol{x}\|_2$ with parameters $(1,1)$

One can use the basic rule to show that

$$f_\mu(\boldsymbol{x})=\sqrt{\|\boldsymbol{Ax}+\boldsymbol{b}\|_2^2+\mu^2}-\mu$$

is a $\frac{1}{\mu}$-smooth approximation of $\|\boldsymbol{Ax}+\boldsymbol{b}\|_2$ with parameters $(\|\boldsymbol{A}\|^2,1)$

Example: $|x|$

Rewrite $|x| = \max\{x, -x\}$, or equivalently,

$$|x| = \max\{\boldsymbol{A}x\} \qquad \text{with } \boldsymbol{A} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Recall that $\mu \log\left(e^{x_1/\mu} + e^{x_2/\mu}\right) - \mu \log 2$ is a $\frac{1}{\mu}$-smooth approximation of $\max\{x_1, x_2\}$ with parameters $(1, \log 2)$

One can then invoke the basic rule to show that

$$f_\mu(x) := \mu \log\left(e^{x/\mu} + e^{-x/\mu}\right) - \mu \log 2$$

is $\frac{1}{\mu}$-smooth approximation of $|x|$ with parameters $(\|\boldsymbol{A}\|^2, \log 2) = (2, \log 2)$

Smoothing via the Moreau envelope

The Moreau envelope (or Moreau-Yosida regularization) of a convex function f with parameter $\mu > 0$ is defined as

$$M_{\mu f}(\boldsymbol{x}) := \inf_{\boldsymbol{z}} \left\{ f(\boldsymbol{z}) + \frac{1}{2\mu} \|\boldsymbol{x} - \boldsymbol{z}\|_2^2 \right\}$$

- $M_{\mu f}$ is a smoothed or regularized form of f
- minimizers of f = minimizers of M_f

  $\Longrightarrow$ minimizing f and minimizing M_f are equivalent

Connection with the proximal operator

- $\mathsf{prox}_f(\boldsymbol{x})$ is the unique point that achieves the infimum that defines M_f, i.e.

$$M_f(\boldsymbol{x}) = f\big(\mathsf{prox}_f(\boldsymbol{x})\big) + \frac{1}{2}\|\boldsymbol{x} - \mathsf{prox}_f(\boldsymbol{x})\|_2^2$$

- M_f is continuously differentiable with gradients (homework)

$$\nabla M_{\mu f}(\boldsymbol{x}) = \frac{1}{\mu}\big(\boldsymbol{x} - \mathsf{prox}_{\mu f}(\boldsymbol{x})\big)$$

This means

$$\underbrace{\mathsf{prox}_{\mu f}(\boldsymbol{x}) = \boldsymbol{x} - \mu \nabla M_{\mu f}(\boldsymbol{x})}_{\mathsf{prox}_{\mu f}(\boldsymbol{x}) \text{ is the gradient step for minimizing } M_{\mu f}}$$

Properties of the Moreau envelope

$$M_{\mu f}(\boldsymbol{x}) := \inf_{\boldsymbol{z}} \left\{ f(\boldsymbol{z}) + \frac{1}{2\mu} \|\boldsymbol{x} - \boldsymbol{z}\|_2^2 \right\}$$

- $M_{\mu f}$ is convex (homework)
- $M_{\mu f}$ is $\frac{1}{\mu}$-smooth (homework)
- If f is L_f-Lipschitz, then $M_{\mu f}$ is a $\frac{1}{\mu}$-smooth approximation of f with parameters $(1, L_f^2/2)$

Proof of smoothability

To begin with,

$$M_{\mu f}(\boldsymbol{x}) \leq f(\boldsymbol{x}) + \frac{1}{2\mu}\|\boldsymbol{x}-\boldsymbol{x}\|_2^2 = f(\boldsymbol{x})$$

In addition, let $\boldsymbol{g}_{\boldsymbol{x}} \in \partial f(\boldsymbol{x})$, which obeys $\|\boldsymbol{g}_{\boldsymbol{x}}\|_2 \leq L_f$. Hence,

$$\begin{aligned} M_{\mu f}(\boldsymbol{x}) - f(\boldsymbol{x}) &= \inf_{\boldsymbol{z}} \left\{ f(\boldsymbol{z}) - f(\boldsymbol{x}) + \frac{1}{2\mu}\|\boldsymbol{z}-\boldsymbol{x}\|_2^2 \right\} \\ &\geq \inf_{\boldsymbol{z}} \left\{ \langle \boldsymbol{g}_{\boldsymbol{x}}, \boldsymbol{z}-\boldsymbol{x} \rangle + \frac{1}{2\mu}\|\boldsymbol{z}-\boldsymbol{x}\|_2^2 \right\} \\ &= -\frac{\mu}{2}\|\boldsymbol{g}_{\boldsymbol{x}}\|_2^2 \geq -\frac{L_f^2}{2}\mu \end{aligned}$$

These together with the smoothness condition of M_f demonstrate that M_f is a $\frac{1}{\mu}$-smooth approximation of f with parameters $(1, L_f^2/2)$

Smoothing via conjugation

Suppose $f = g^*$, namely,

$$f(\boldsymbol{x}) = \sup_{\boldsymbol{z}} \{\langle \boldsymbol{z}, \boldsymbol{x} \rangle - g(\boldsymbol{z})\}$$

One can build a smooth approximation of f by adding a strongly convex component to its dual, namely,

$$f_\mu(\boldsymbol{x}) = \sup_{\boldsymbol{z}} \{\langle \boldsymbol{z}, \boldsymbol{x} \rangle - g(\boldsymbol{z}) - \mu d(\boldsymbol{z})\} = (g + \mu d)^* (\boldsymbol{x})$$

for some 1-strongly convex and continuous function $d \geq 0$ (called proximity function)

Smoothing via conjugation

2 properties:

- $g + \mu d$ is μ-strongly convex $\implies$ f_μ is $\frac{1}{\mu}$-smooth
- $f_\mu(\boldsymbol{x}) \le f(\boldsymbol{x}) \le f_\mu(\boldsymbol{x}) + \mu D$ with $D := \sup_{\boldsymbol{x}} d(\boldsymbol{x})$

$$\Longrightarrow \quad f_\mu \text{ is a } \frac{1}{\mu}\text{-smooth approximation of } f \text{ with parameters } (1, D)$$

Example: $|x|$

Recall that

$$|x| = \sup_{|z|\le 1} zx$$

If we take $d(z) = \frac{1}{2}z^2$, then smoothing via conjugation gives

$$f_\mu(x) = \sup_{|z|\le 1} \left\{ zx - \frac{\mu}{2}z^2 \right\} = \begin{cases} x^2/2\mu, & |x| \le \mu \\ |x| - \mu/2, & \text{else} \end{cases}$$

which is exactly the Huber function

Example: $|x|$

Another way of conjugation:

$$|x| = \sup_{z_1, z_2 \geq 0, z_1 + z_2 = 1} (z_1 - z_2)x$$

If we take $d(z) = z_1 \log z_1 + z_2 \log z_2 + \log 2$, then smoothing via conjugation gives

$$f_\mu(x) = \mu \log\big(\cosh(x/\mu)\big)$$

where $\cosh x = \frac{e^x + e^{-x}}{2}$

Example: norm

Consider $\|\boldsymbol{x}\| = \sup_{\|\boldsymbol{z}\|_* \leq 1} \langle \boldsymbol{z}, \boldsymbol{x} \rangle$, then smoothing via conjugation gives

$$f_\mu(\boldsymbol{x}) = \sup_{\|\boldsymbol{z}\|_* \leq 1} \left\{ \langle \boldsymbol{z}, \boldsymbol{x} \rangle - \mu d(\boldsymbol{z}) \right\}$$

Algorithm and convergence analysis

Algorithm

$$\text{minimize}_{\boldsymbol{x}} \quad F(\boldsymbol{x}) = f(\boldsymbol{x}) + h(\boldsymbol{x})$$

- f is convex and (α, β)-smoothable
- h is convex but may not be differentiable

Algorithm

Build f_μ — $\frac{1}{\mu}$-smooth approximation of f with parameters (α, β)

$$\boldsymbol{x}^{t+1} = \mathsf{prox}_{\eta_t h}(\boldsymbol{y}^t - \eta_t \nabla f_\mu(\boldsymbol{y}^t))$$
$$\boldsymbol{y}^{t+1} = \boldsymbol{x}^{t+1} + \frac{\theta_t - 1}{\theta_{t+1}}(\boldsymbol{x}^{t+1} - \boldsymbol{x}^t)$$

where $\boldsymbol{y}^0 = \boldsymbol{x}^0$, $\theta_0 = 1$ and $\theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2}$

Convergence

Theorem 8.1 (informal)

Take $\mu = \frac{\varepsilon}{2\beta}$. *Then one has* $F(\boldsymbol{x}^t) - F^{\mathsf{opt}} \leq \varepsilon$ *for any*

$$t \gtrsim \frac{\sqrt{\alpha\beta}}{\varepsilon}$$

- iteration complexity: $O(1/\varepsilon)$, which improves upon that of subgradient methods $O(1/\varepsilon^2)$

Proof sketch

- **convergence rate for smooth problem:** to attain $\frac{\varepsilon}{2}$-accuracy for minimizing $F_\mu(\boldsymbol{x}) := f_\mu(\boldsymbol{x}) + h(\boldsymbol{x})$, one needs $O\Big(\sqrt{\frac{\alpha}{\mu}} \cdot \frac{1}{\sqrt{\varepsilon}}\Big)$ iterations

- **approximation error:** set $\beta\mu = \frac{\varepsilon}{2}$ to ensure $|f(\boldsymbol{x}) - f_\mu(\boldsymbol{x})| \le \frac{\varepsilon}{2}$

- since $F(\boldsymbol{x}^t) - F(\boldsymbol{x}^{\mathsf{opt}}) \le \underbrace{\big|f(\boldsymbol{x}^t) - f_\mu(\boldsymbol{x}^t)\big|}_{\le \varepsilon/2} + \underbrace{\big(F_\mu(\boldsymbol{x}^t) - F_\mu^{\mathsf{opt}}\big)}_{\le \varepsilon/2}$,

  the iteration complexity is

$$O\left(\sqrt{\frac{\alpha}{\mu}} \cdot \frac{1}{\sqrt{\varepsilon}}\right) = O\left(\sqrt{\frac{\alpha\beta}{\varepsilon}} \cdot \frac{1}{\sqrt{\varepsilon}}\right) = O\left(\frac{\sqrt{\alpha\beta}}{\varepsilon}\right)$$

Reference

[1] "*Smooth minimization of non-smooth functions*," Y. Nesterov, *Mathematical programming*, 2005.

[2] "*First-order methods in optimization*," A. Beck, Vol. 25, SIAM, 2017.

[3] "*Optimization methods for large-scale systems, EE236C lecture notes*," L. Vandenberghe, UCLA.

[4] "*Mathematical optimization, MATH301 lecture notes*," E. Candes, Stanford.

[5] "*Smoothing and first order methods: A unified framework*," A. Beck, M. Teboulle, *SIAM Journal on Optimization*, 2012.

[6] "*Problem complexity and method efficiency in optimization*," A. Nemirovski, D. Yudin, Wiley, 1983.