Smoothing for nonsmooth optimization

Yuxin Chen
Princeton University, Fall 2019
Outline

- Smoothing
- Smooth approximation
- Algorithm and convergence analysis
Nonsmooth optimization

\[
\begin{align*}
\text{minimize}_{x \in \mathbb{R}^n} & \quad f(x) \\
\text{where } f \text{ is convex but not always differentiable} & \\
\text{• subgradient methods yield } \varepsilon\text{-accuracy in} & \\
& O \left( \frac{1}{\varepsilon^2} \right) \text{ iterations} \\
\text{• in contrast, if } f \text{ is smooth, then accelerated GD yields} & \\
& \varepsilon\text{-accuracy in} \\
& O \left( \frac{1}{\sqrt{\varepsilon}} \right) \text{ iterations} \\
& \quad \text{— significantly better than the nonsmooth case}
\end{align*}
\]
Lower bound

— Nemirovski & Yudin ’83

If one only has access to the first-order oracle (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
A convex function $f$ is called $(\alpha, \beta)$-smoothable if, for any $\mu > 0$, $\exists$ convex function $f_\mu$ s.t.

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

— $\mu$: tradeoff between approximation accuracy and smoothness

Here, $f_\mu$ is called a $\frac{1}{\mu}$-smooth approximation of $f$ with parameters $(\alpha, \beta)$
Example: $\ell_1$ norm

Consider the Huber function

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

which satisfies

$$h_\mu(z) \leq |z| \leq h_\mu(z) + \mu/2$$

and $h_\mu(z)$ is $\frac{1}{\mu}$-smooth.
Example: $\ell_1$ norm

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

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

$$\implies \|\cdot\|_1 \text{ is } (1, n/2)-\text{smoothable}$$
Example: $\ell_2$ norm

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

$$f_\mu(x) \leq (\|x\|_2 + \mu) - \mu = \|x\|_2$$

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

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

Therefore, $\| \cdot \|_2$ is $(1,1)$-smoothable
Consider $f_\mu(x) := \mu \log \left( \sum_{i=1}^{n} e^{x_i/\mu} \right) - \mu \log n$, then $\forall \mu > 0$ and $\forall x \in \mathbb{R}^n$,

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

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

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

- $f_{\mu,1}$ is a $\frac{1}{\mu}$-smooth approximation of $f_1$ with parameters $(\alpha_1, \beta_1)$
- $f_{\mu,2}$ is a $\frac{1}{\mu}$-smooth approximation of $f_2$ with parameters $(\alpha_2, \beta_2)$

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

- $h_\mu$ is a $\frac{1}{\mu}$-smooth approximation of $h$ with parameters $(\alpha, \beta)$
- $f(x) := h(Ax + b)$

$\implies h_\mu(Ax + b)$ is a $\frac{1}{\mu}$-smooth approximation of $f$ with parameters $(\alpha \|A\|^2, \beta)$
Example: $\|Ax + b\|_2$

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

One can use the basic rule to show that

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

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

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

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

Recall that $\mu \log (e^{x_1/\mu} + e^{x_2/\mu}) - \mu \log 2$ 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 $(\|A\|^2, \log 2) = (2, \log 2)$
The Moreau envelope (or Moreau-Yosida regularization) of a convex function \( f \) with parameter \( \mu > 0 \) is defined as

\[
M_{\mu f}(x) := \inf_z \left\{ f(z) + \frac{1}{2\mu} \|x - z\|^2 \right\}
\]

- \( M_{\mu f} \) is a smoothed or regularized form of \( f \)
- minimizers of \( f = \) minimizers of \( M_f \)
  \[ \implies \]  minimizing \( f \) and minimizing \( M_f \) are equivalent
Connection with the proximal operator

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

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

- $M_f$ is continuously differentiable with gradients (homework)

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

This means

$$\text{prox}_\mu f(x) = x - \mu \nabla M_\mu f(x)$$

$\text{prox}_\mu f(x)$ is the gradient step for minimizing $M_\mu f$
Properties of the Moreau envelope

\[ M_{\mu f}(x) := \inf_z \left\{ f(z) + \frac{1}{2\mu} \| x - z \|^2 \right\} \]

- \( M_{\mu f} \) 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 \( (1, L^2_f / 2) \)
Proof of smoothability

To begin with,

\[ M_{\mu f}(x) \leq f(x) + \frac{1}{2\mu} \|x - x\|^2 = f(x) \]

In addition, let \( g_x \in \partial f(x) \), which obeys \( \|g_x\|_2 \leq L_f \). Hence,

\[ M_{\mu f}(x) - f(x) = \inf_z \left\{ f(z) - f(x) + \frac{1}{2\mu} \|z - x\|^2 \right\} \]

\[ \geq \inf_z \left\{ \langle g_x, z - x \rangle + \frac{1}{2\mu} \|z - x\|^2 \right\} \]

\[ = -\frac{\mu}{2} \|g_x\|^2 \geq -\frac{L_f^2}{2\mu} \]

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

\[
    f(x) = \sup_z \{ \langle z, x \rangle - g(z) \}
\]

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

\[
    f_\mu(x) = \sup_z \{ \langle z, x \rangle - g(z) - \mu d(z) \} = (g + \mu d)^*(x)
\]

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

2 properties:

• \( g + \mu d \) is \( \mu \)-strongly convex \( \implies \) \( f_\mu \) is \( \frac{1}{\mu} \)-smooth

• \( f_\mu(x) \leq f(x) \leq f_\mu(x) + \mu D \) with \( D := \sup_x d(x) \)

\( \implies \) \( f_\mu \) is a \( \frac{1}{\mu} \)-smooth approximation of \( f \) with parameters \((1, D)\)
Recall that

\[ |x| = \sup_{|z| \leq 1} zx \]

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

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

which is exactly the Huber function.
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

\[ f_\mu(x) = \mu \log (\cosh(x/\mu)) \]

where \( \cosh x = \frac{e^x + e^{-x}}{2} \)
Example: norm

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

\[
f_\mu(x) = \sup_{\|z\|_* \leq 1} \{ \langle z, x \rangle - \mu d(z) \}
\]
Algorithm and convergence analysis
Algorithm

\[
\text{minimize}_{x} \quad F(x) = f(x) + h(x)
\]

- \(f\) is convex and \((\alpha, \beta)\)-smoothable
- \(h\) is convex but may not be differentiable
Algorithm

Build $f_{\mu} \leftarrow \frac{1}{\mu}$-smooth approximation of $f$ with parameters $(\alpha, \beta)$

\[
x^{t+1} = \text{prox}_{\eta_t h}(y^t - \eta_t \nabla f_{\mu}(y^t))
\]

\[
y^{t+1} = x^{t+1} + \frac{\theta_t - 1}{\theta_{t+1}}(x^{t+1} - x^t)
\]

where $y^0 = x^0$, $\theta_0 = 1$ and $\theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2}$
Theorem 8.1 (informal)

Take \( \mu = \frac{\varepsilon}{2\beta} \). Then one has \( F(x^t) - F^{\text{opt}} \leq \varepsilon \) 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 $F_\mu(x) := f_\mu(x) + h(x)$, one needs $O\left(\sqrt{\frac{\alpha}{\mu}} \cdot \frac{1}{\sqrt{\varepsilon}}\right)$ iterations

• approximation error: set $\beta_\mu = \frac{\varepsilon}{2}$ to ensure $|f(x) - f_\mu(x)| \leq \frac{\varepsilon}{2}$

• since $F(x^t) - F(x^{opt}) \leq \left| f(x^t) - f_\mu(x^t) \right| + \left( F_\mu(x^t) - F^{opt}_\mu \right)$, 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


