EECS 559
Optimization Methods for SIPML

Lecture 7 – Smoothing & Proximal Methods

Instructor: Prof. Qing Qu (qingqu@umich.edu)

Lecture Agenda

- Smoothing for Nonsmooth Problems
- Moreau Envelope and Proximal Operator
- Proximal Gradient Methods

Lecture Agenda

- **Smoothing for Nonsmooth Problems**
- Moreau Envelope and Proximal Operator
- Proximal Gradient Methods

Nonsmooth Optimization

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; f(\boldsymbol{x})$$

Can the convergence gap be closed?

- **For convex and smooth problems**, (accelerated) gradient descent (GD) converges to ε-accuracy in

$$O\left(\frac{1}{\sqrt{\varepsilon}}\right) \text{ iterations, or } O\left(\frac{1}{k^2}\right) \text{convergence rate.}$$

- **For convex but non-smooth problems,** subgradient method converges to ε-accuracy in

$$O\left(\frac{1}{\varepsilon^2}\right) \text{ iterations, or } O\left(\frac{1}{\sqrt{k}}\right) \text{convergence rate.}$$

Smooth Approximation

Definition. A convex function f is called (α, β)-smoothable, if for any $\mu > 0$, there exists a convex function f_μ, such that

- $f_\mu(\boldsymbol{x}) \leq f(\boldsymbol{x}) \leq f_\mu(\boldsymbol{x}) + \beta\mu, \quad \forall \boldsymbol{x}$
- f_μ is $\frac{\alpha}{\mu}$-smooth, with gradient:

$$\|\nabla f_\mu(\boldsymbol{x}) - \nabla f_\mu(\boldsymbol{x}')\|_2 \leq \frac{\alpha}{\mu}\|\boldsymbol{x} - \boldsymbol{x}'\|_2, \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n$$

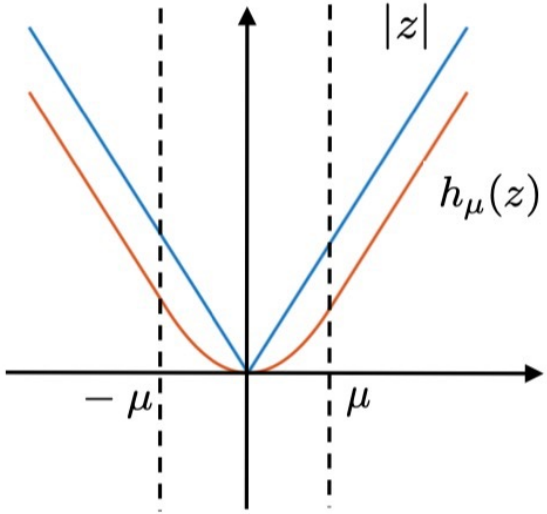
f_μ is called a $\frac{1}{\mu}$–smooth approximation of with parameters (α, β).
μ : tradeoff between approximation accuracy ϵ and smoothness

Example: ℓ_1-Norm

Consider the Huber function

$$f_\mu(\boldsymbol{x}) \;:=\; \sum_{i=1}^{n} h_\mu(x_i),$$

$$h_\mu(x) \;=\; \begin{cases} |x| - \frac{\mu}{2} & |x| \geq \mu \\ \frac{x^2}{2\mu} & |x| < \mu \end{cases}$$



which satisfies

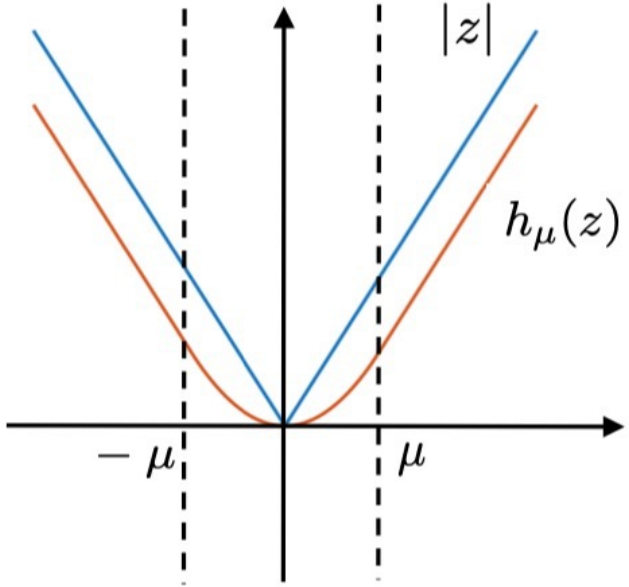
$$h_\mu(x) \;\leq\; |x| \;\leq\; h_\mu(x) + \mu/2, \text{ and } h_\mu(x) \text{ is } \tfrac{1}{\mu}\text{-smooth.}$$

Example: ℓ_1-Norm

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

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

so that $\|\boldsymbol{x}\|_1$ is $(1, n/2)$-smoothable.

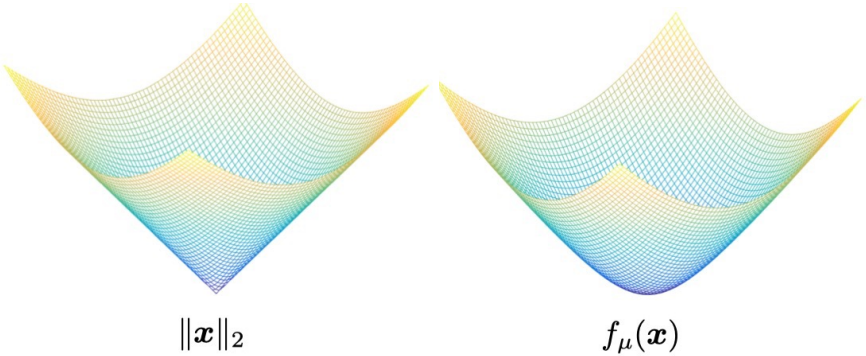


Example: ℓ_2-Norm

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

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

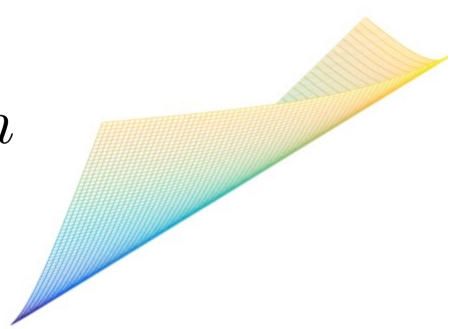
$$\|\boldsymbol{x}\|_2 \leq \sqrt{\|\boldsymbol{x}\|_2^2 + \mu^2} = f_\mu(\boldsymbol{x}) + \mu$$



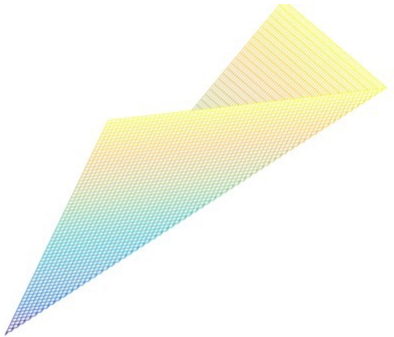
Additionally, we can show that $f_\mu(\boldsymbol{x})$ is $\frac{1}{\mu}$-smooth.
Therefore, $\|\boldsymbol{x}\|_2$ is (1,1)-smoothable.

Example: Max Function

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



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



$\max\{x_1, x_2\}$

For any $\mu > 0$ and any $\boldsymbol{x} \in \mathbb{R}^n$,

$$f_\mu(\boldsymbol{x}) \;\leq\; f(\boldsymbol{x}) \;=\; \max_i x_i \;\leq\; f_\mu(\boldsymbol{x}) + \mu \log n.$$

Additionally, we can show that $f_\mu(\boldsymbol{x})$ is $\frac{1}{\mu}$-smooth.
Therefore, $f(\boldsymbol{x}) = \max_{1 \leq i \leq n} x_i$ is $(1, \log n)$-smoothable.

Basic Composition Rules

- **Addition:**

Let $f_{\mu,1}$ be a $\frac{1}{\mu}$-smooth approximation of f_1 with (α_1, β_1), and let $f_{\mu,2}$ be a $\frac{1}{\mu}$-smooth approximation of f_2 with (α_2, β_2). Then

$\lambda_1 f_{\mu,1} + \lambda_2 f_{\mu,2}$ $(\lambda_1, \lambda_2 > 0)$ 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 Composition Rules

- **Affine Transformation:**

h_μ is a $\frac{1}{\mu}$-smooth approximation of h with parameters (α, β) and let $f(\boldsymbol{x}) := h(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$. Then $h_\mu(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})$ is a $\frac{1}{\mu}$-smooth approximation of f with parameters $\left(\alpha \|\boldsymbol{A}\|^2, \beta\right)$.

Example: $f(x) = |x|$

Claim. $f_\mu(x) := \mu \log\left(e^{x/\mu} + e^{-x/\mu}\right) - \mu \log 2$ is a $\frac{1}{\mu}$-smooth approximation of $f(x) = |x|$ with $(2, \log 2)$.

Example: $f(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}\|_2$

Claim. $f_\mu(\boldsymbol{x}) = \sqrt{\|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}\|_2^2 + \mu^2} - \mu$ is a $\frac{1}{\mu}$-smooth approximation of $\|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}\|_2$ with $(\|\boldsymbol{A}\|_2^2, 1)$.

Lecture Agenda

- Smoothing for Nonsmooth Problems
- **Moreau Envelope and Proximal Operator**
- Proximal Gradient Methods

Minimizing a Nonsmooth Function?

Suppose we want to minimize

$$\min_{\boldsymbol{x}} \quad \underbrace{f(\boldsymbol{x})}_{\textbf{nonsmooth}}, \quad \text{s.t.} \quad \boldsymbol{x} \in \mathbb{R}^n.$$

- We know that subgradient method converges slowly;
- Can we improve the convergence by minimizing a smooth surrogate of f?

Smoothing via the Moreau Envelope

Definition. The Moreau envelope 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 approximation of f;
- Minimizers of f = minimizers of $M_{\mu f}$.

Smoothing via the Moreau Envelope

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

Show the following (homework):

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

Smoothing via the Moreau Envelope

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

Therefore, we can minimize f by minimizing $M_{\mu f}(\boldsymbol{x})$

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - \tau_k \nabla M_{\mu f}(\boldsymbol{x}_k) \\ &= \operatorname{prox}_{\mu f}(\boldsymbol{x}_k), \quad \text{if } \tau_k = \mu \end{aligned}$$

Proximal Operator & Connections

Definition. The proximal operator for a convex function f is defined as

$$\text{prox}_{\mu f}(\boldsymbol{x}) = \arg\inf_{\boldsymbol{z}} \left(f(\boldsymbol{z}) + \frac{1}{2\mu} \|\boldsymbol{z} - \boldsymbol{x}\|_2^2 \right).$$

Therefore, we have

$$M_{\mu f}(\boldsymbol{x}) \;=\; f(\text{prox}_{\mu f}(\boldsymbol{x})) + \frac{1}{2\mu} \left\|\boldsymbol{x} - \text{prox}_{\mu f}(\boldsymbol{x})\right\|_2^2$$

Proximal Operator

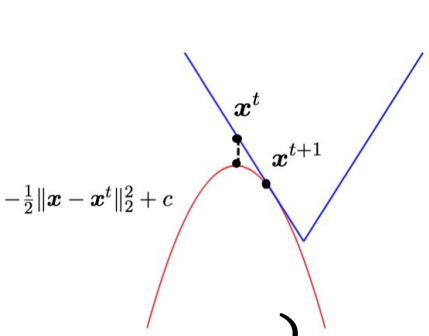
Definition. The proximal operator for a convex function f at the point $\boldsymbol{x}$ is defined as

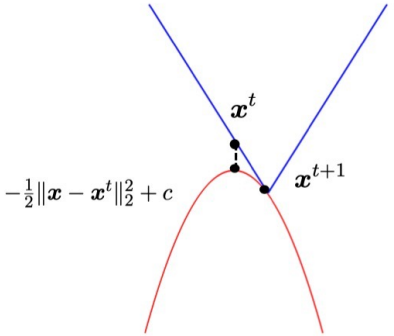
$$\operatorname{prox}_{\mu f}(\boldsymbol{x}) = \arg\inf_{\boldsymbol{z}} \left(f(\boldsymbol{z}) + \frac{1}{2\mu} \|\boldsymbol{z} - \boldsymbol{x}\|_2^2 \right).$$

- well-defined under very general conditions;
- can be evaluated efficiently for many widely used functions;
- the abstraction is conceptually & mathematically simple, covers many well-known optimization methods.

Proximal Operator: ℓ_1-Norm

If $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ with





$$\text{prox}_{\mu f}(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2\mu} \|\boldsymbol{z} - \boldsymbol{x}\|_2^2 + \|\boldsymbol{z}\|_1 \right\}.$$

Then the proximal operator is the *soft-thresholding operator*

$$\text{prox}_{\mu f}(\boldsymbol{x}) = \text{S}_\mu(\boldsymbol{x}), \qquad [\text{S}_\mu(\boldsymbol{x})]_i = \begin{cases} x_i - \mu & \text{if } x_i > \mu, \\ 0 & \text{if } |x_i| \leq \mu, \\ x_i + \mu & \text{if } x_i \leq -\mu. \end{cases}$$

Proximal Operator: Nuclear Norm

If $f(\boldsymbol{X}) = \|\boldsymbol{X}\|_*$ with $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, and

$$\text{prox}_{\mu f}(\boldsymbol{X}) \;=\; \arg\min_{\boldsymbol{Z}} \left\{ \frac{1}{2\mu} \|\boldsymbol{Z} - \boldsymbol{X}\|_F^2 + \|\boldsymbol{Z}\|_* \right\}.$$

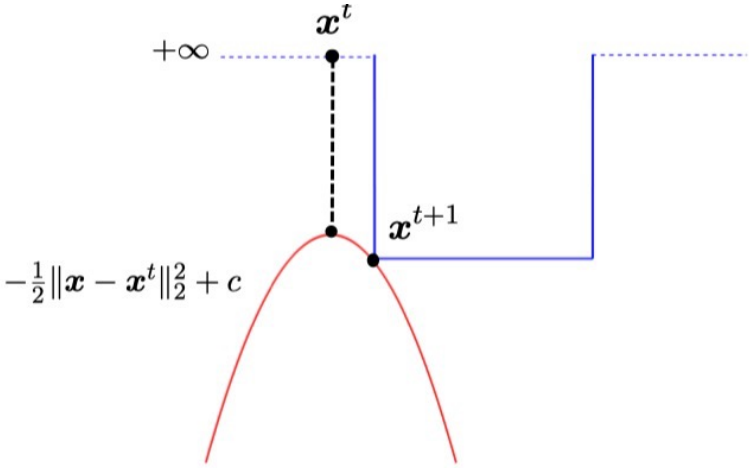
Then the proximal operator is the *singular value thresholding operator*

$$\text{prox}_{\mu f}(\boldsymbol{X}) \;=\; \boldsymbol{U}\text{S}_\mu(\boldsymbol{\Sigma})\boldsymbol{V}^\top.$$

Proximal Operator: Indicator Function

If $f(\boldsymbol{x}) = \mathbb{1}_{\boldsymbol{x} \in \mathcal{C}}$ with

$$f(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \boldsymbol{x} \in \mathcal{C} \\ \infty & \text{else.} \end{cases}$$

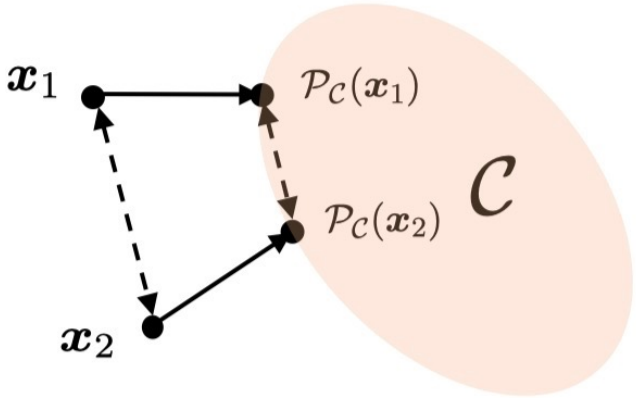


The proximal operator is the *Euclidean projection*:

$$\text{prox}_{\mu f}(\boldsymbol{x}) = \arg\min_{\boldsymbol{z} \in \mathcal{C}} \|\boldsymbol{z} - \boldsymbol{x}\|_2$$

Nonexapansiveness of Proximal Operators

$$f(\boldsymbol{x}) = \mathbb{1}_{\boldsymbol{x}\in\mathcal{C}}$$

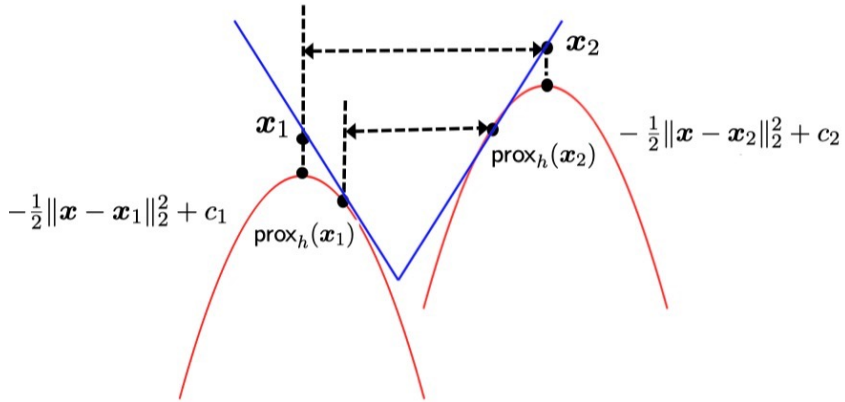


When $\text{prox}_{\mu f}(\boldsymbol{x})$ is the Euclidean projection onto $\mathcal{C}$, it is *nonexpansive* for a *convex* $\mathcal{C}$:

$$\|\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}_1) - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}_2)\|_2 \;\leq\; \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 \,.$$

Nonexapansiveness of Proximal Operators

Proximal operator behaves
like the projection operator.



- **Nonexpansiveness** (homework):

$$\left\|\text{prox}_{\mu f}(\boldsymbol{x}_1)-\text{prox}_{\mu f}(\boldsymbol{x}_2)\right\|_2 \;\leq\; \|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2\,.$$

- **Firm nonexpansiveness** (homework):

$$\left\langle \text{prox}_{\mu f}(\boldsymbol{x}_1)-\text{prox}_{\mu f}(\boldsymbol{x}_2),\boldsymbol{x}_1-\boldsymbol{x}_2\right\rangle \;\geq\; \left\|\text{prox}_{\mu f}(\boldsymbol{x}_1)-\text{prox}_{\mu f}(\boldsymbol{x}_2)\right\|_2^2$$

Proximal Operator & Moreau Envelope

Theorem. (Gradient of Moreau envelope)
The Moreau envelope

$$M_{\mu f}(\boldsymbol{x}) \;=\; f(\mathrm{prox}_{\mu f}(\boldsymbol{x})) + \frac{1}{2\mu}\left\|\boldsymbol{x} - \mathrm{prox}_{\mu f}(\boldsymbol{x})\right\|_2^2$$

is continuously differentiable with gradient

$$\nabla M_{\mu f}(\boldsymbol{x}) \;=\; \frac{1}{\mu}\left(\boldsymbol{x} - \mathrm{prox}_{\mu f}(\boldsymbol{x})\right).$$

- Proximal operator is a gradient step for minimizing $M_{\mu f}$:

$$\mathrm{prox}_{\mu f}(\boldsymbol{x}) \;=\; \boldsymbol{x} - \mu \cdot \nabla M_{\mu f}(\boldsymbol{x}).$$

Proximal Operator & Connections

Proof. First of all, let us introduce the conjugate of a given function f, which is defined by

$$f^*(\boldsymbol{x}) := \sup_{\boldsymbol{z} \in \mathrm{dom} f} \boldsymbol{z}^\top \boldsymbol{x} - f(\boldsymbol{z})$$

and we can show that

$$\nabla f^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{z}} \left\{ \boldsymbol{z}^\top \boldsymbol{x} - f(\boldsymbol{z}) \right\}$$

Proximal Operator & Connections

For our case, we can rearrange terms in $M_{\mu f}(\boldsymbol{x})$ as

$$\begin{aligned} M_{\mu f}(\boldsymbol{x}) &= \frac{1}{2\mu}\|\boldsymbol{x}\|_2^2 - \frac{1}{\mu}\sup_{\boldsymbol{z}}\left\{\boldsymbol{z}^\top\boldsymbol{x} - \mu f(\boldsymbol{z}) - \frac{1}{2}\|\boldsymbol{z}\|_2^2\right\} \\ &= \frac{1}{2\mu}\|\boldsymbol{x}\|_2^2 - \frac{1}{\mu}\left(\mu f + \frac{1}{2}\|\cdot\|_2^2\right)^*(\boldsymbol{x}) \end{aligned}$$

Therefore, we have

$$\nabla M_{\mu f}(\boldsymbol{x}) = \frac{\boldsymbol{x}}{\mu} - \frac{1}{\mu}\arg\max_{\boldsymbol{z}}\left\{\boldsymbol{z}^\top\boldsymbol{x} - \mu f(\boldsymbol{z}) - \frac{1}{2}\|\boldsymbol{z}\|_2^2\right\} = \frac{1}{\mu}\left(\boldsymbol{x} - \mathrm{prox}_{\mu f}(\boldsymbol{x})\right)$$

which implies that $\mathrm{prox}_{\mu f}(\boldsymbol{x}) = \boldsymbol{x} - \mu\cdot\nabla M_{\mu f}(\boldsymbol{x})$.

Further Readings

- *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. John Wright, Yi Ma. (**Chapter 8.2**)

- *Proximal Algorithms, Foundations & Trends in Optimization*. Neal Parikh, Stephen Boyd, 2014. https://web.stanford.edu/~boyd/papers/prox_algs.html

Lecture Agenda

- Smoothing for Nonsmooth Problems
- Moreau Envelope and Proximal Operator
- **Proximal Gradient Methods**

Nonsmooth Composite Problems

$$\min_{\boldsymbol{x}} \; F(\boldsymbol{x}) \;=\; f(\boldsymbol{x}) \;+\; g(\boldsymbol{x})$$

- f : convex, continuously differentiable, and L-smooth,

$$\|\nabla f(\boldsymbol{x}') - \nabla f(\boldsymbol{x})\|_2 \;\leq\; L\,\|\boldsymbol{x}' - \boldsymbol{x}\|_2\,, \quad \forall \boldsymbol{x},\, \boldsymbol{x}' \in \mathbb{R}^n.$$

- g : convex but may *not* be differentiable.

Example I: ℓ_1-Regularized Problems

$$\min_{\boldsymbol{x}} \; f(\boldsymbol{x}) \; + \; \underbrace{\lambda \cdot \|\boldsymbol{x}\|_1}_{g(\boldsymbol{x})}$$

- **Lasso problem:** $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$
- **Sparse logistic regression:**

$$f(\boldsymbol{x}) \; = \; \sum_{i=1}^{m} \left(-y_i \boldsymbol{x}^\top \boldsymbol{z}_i + \log\left(1 + e^{y_i \boldsymbol{x}^\top \boldsymbol{z}_i}\right)\right)$$

Example II: Nuclear Norm Regularization

- **Stable low-rank matrix recovery:**

$$\min_{\boldsymbol{X}} \; F(\boldsymbol{X}) \;=\; \underbrace{\frac{1}{2}\left\|\boldsymbol{y}-\mathcal{A}(\boldsymbol{X})\right\|_F^2}_{f(\boldsymbol{X})} \;+\; \underbrace{\lambda\left\|\boldsymbol{X}\right\|_*}_{g(\boldsymbol{X})}.$$

- **Stable principal component pursuit (PCP):**

$$\min_{\boldsymbol{L},\boldsymbol{S}} \; F(\boldsymbol{L},\boldsymbol{S}) \;=\; \underbrace{\frac{1}{2}\left\|\boldsymbol{Y}-\boldsymbol{L}-\boldsymbol{S}\right\|_F^2}_{f(\boldsymbol{L},\boldsymbol{S})} \;+\; \underbrace{\lambda\left\|\boldsymbol{L}\right\|_* + \mu\left\|\boldsymbol{S}\right\|_1}_{g(\boldsymbol{L},\boldsymbol{S})}.$$

Closing the Gap?

$$\min_{\boldsymbol{x}} \; F(\boldsymbol{x}) \;=\; f(\boldsymbol{x}) \;+\; g(\boldsymbol{x})$$

- **For convex and smooth problems**, (accelerated) gradient descent (GD) converges to ε-accuracy in

$$O\left(\frac{1}{\sqrt{\varepsilon}}\right) \text{iterations, or } O\left(\frac{1}{k^2}\right) \text{convergence rate.}$$

- **For convex but non-smooth problems,** subgradient method converges to ε-accuracy in

$$O\left(\frac{1}{\varepsilon^2}\right) \text{ iterations, or } O\left(\frac{1}{\sqrt{k}}\right) \text{convergence rate.}$$

Taylor Approximation Theory I

Theorem (first-order approximation).

Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be continuously differentiable, and $\nabla f(\boldsymbol{x})$ is $\gamma_L(\boldsymbol{x}_0)$-Lipschitz continuous at $\boldsymbol{x}_0$, then

$$\left| f(\boldsymbol{x}) - \widehat{f}_L(\boldsymbol{x}; \boldsymbol{x}_0) \right| \leq \frac{\gamma_L(\boldsymbol{x}_0)}{2} \|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2 ,$$

where we define

$$\widehat{f}_L(\boldsymbol{x}; \boldsymbol{x}_0) := f(\boldsymbol{x}_0) + \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{x} - \boldsymbol{x}_0 \rangle .$$

Proximal Gradient Method

Consider optimizing the following *quadratic upper bound*

$$\widehat{F}_\mu(\boldsymbol{x}, \boldsymbol{x}_k) = \underbrace{f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k \rangle + g(\boldsymbol{x})}_{=: \ell_F(\boldsymbol{x}, \boldsymbol{x}_k)} + \frac{1}{2\mu} \|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2$$

$$\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{x}} \widehat{F}_\mu(\boldsymbol{x}, \boldsymbol{x}_k).$$

Proximal Gradient Method

$$\widehat{F}_\mu(\boldsymbol{x}, \boldsymbol{x}_k) = g(\boldsymbol{x}) + \frac{1}{2\mu} \|\boldsymbol{x} - (\boldsymbol{x}_k - \mu \nabla f(\boldsymbol{x}_k))\|_2^2 + f(\boldsymbol{x}_k) - \frac{\mu}{2} \|\nabla f(\boldsymbol{x}_k)\|_2^2$$

Let $\boldsymbol{w}_k = \boldsymbol{x}_k - \mu \nabla f(\boldsymbol{x}_k)$, then

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \arg\min_{\boldsymbol{x}} \left\{ g(\boldsymbol{x}) + \frac{1}{2\mu} \|\boldsymbol{x} - \boldsymbol{w}_k\|_2^2 \right\} \\ &= \operatorname{prox}_{\mu g}(\boldsymbol{w}_k). \end{aligned}$$

Proximal Gradient Method

Proximal Gradient (PG)

Problem Class:

$$\min_{\boldsymbol{x}} F(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$$

$f, g : \mathbb{R}^n \to \mathbb{R}$ are convex, ∇f L-Lipschitz and g (maybe) non-smooth.

Basic Iteration: set $\boldsymbol{x}_0 \in \mathbb{R}^n$.

Repeat:

$$\boldsymbol{w}_k \leftarrow \boldsymbol{x}_k - \frac{1}{L}\nabla f(\boldsymbol{x}_k),$$
$$\boldsymbol{x}_{k+1} \leftarrow \text{prox}_{g/L}[\boldsymbol{w}_k].$$

Convergence Guarantee:

$F(\boldsymbol{x}_k) - F(\boldsymbol{x}_\star)$ converges at a rate of $O(1/k)$.

Proximal Gradient Method

In summary, one step of PG is

$$\boldsymbol{x}_{k+1} \;=\; \mathrm{prox}_{\mu g}\left(\boldsymbol{x}_k - \mu \nabla f(\boldsymbol{x}_k)\right)$$

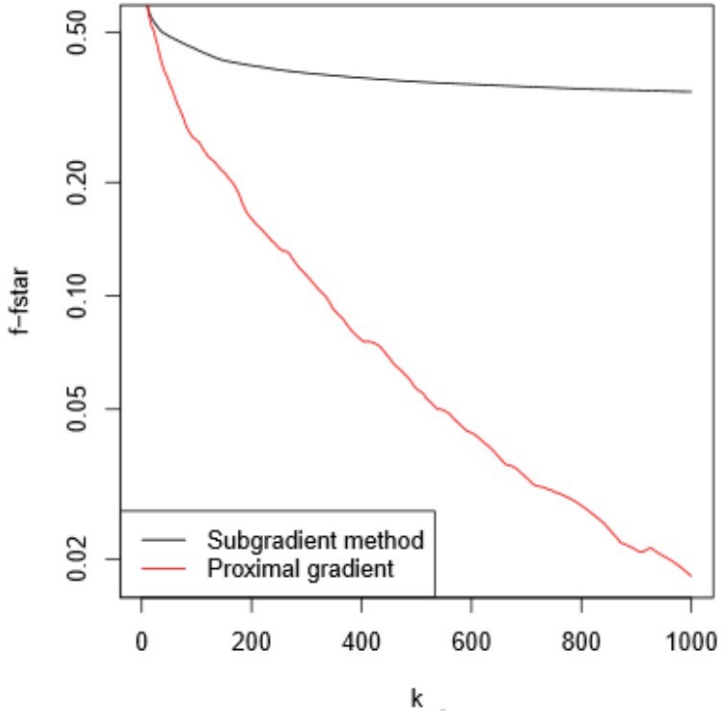
where $\mu \in (0, 1/L]$.

Let $G_\mu(\boldsymbol{x}) := \dfrac{1}{\mu}\left(\boldsymbol{x} - \mathrm{prox}_{\mu g}(\boldsymbol{x} - \mu \nabla f(\boldsymbol{x})\right)$, then

$$\boldsymbol{x}_{k+1} \;=\; \boldsymbol{x}_k - \mu \cdot \underbrace{G_\mu(\boldsymbol{x}_k)}_{\text{proximal gradient}} .$$

Comparison with Subgradient

$$\min_{\boldsymbol{x}} \; F(\boldsymbol{x}) = \underbrace{\frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\right\|_2^2}_{f(\boldsymbol{x})} + \underbrace{\lambda\left\|\boldsymbol{x}\right\|_1}_{g(\boldsymbol{x})}$$



Convergence of Proximal Gradient

Theorem. Let $F(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$, with

- f is convex, differentiable, and L-smooth;
- g is convex, but possibly nonsmooth.

Consider the sequence of proximal gradient iterates

$$\boldsymbol{x}_{k+1} = \operatorname{prox}_{\mu_k g}\left(\boldsymbol{x}_k - \mu_k \nabla f(\boldsymbol{x}_k)\right), \quad \mu_k \equiv 1/L.$$

Let $\boldsymbol{x}_\star$ be a minimizer of F. Then for any $k \geq 1$,

$$F(\boldsymbol{x}_k) - F(\boldsymbol{x}_\star) \leq \frac{L \left\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\right\|_2^2}{2k}.$$

Convergence of Proximal Gradient

Proof. By convexity of f and g, we have

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k \rangle$$
$$g(\boldsymbol{x}) \geq g(\boldsymbol{x}_{k+1}) + \langle \boldsymbol{\gamma}, \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle \quad \boldsymbol{\gamma} \in \partial g$$

Thus, we obtain

$$\begin{aligned} F(\boldsymbol{x}) &= f(\boldsymbol{x}) + g(\boldsymbol{x}) \\ &\geq f(\boldsymbol{x}_k) + g(\boldsymbol{x}_{k+1}) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k \rangle + \langle \boldsymbol{\gamma}, \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle \end{aligned}$$

On the other hand, for $\mu \in (0, 1/L]$ we have

$$\widehat{F}_\mu(\boldsymbol{x}, \boldsymbol{x}_k) = f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k \rangle + g(\boldsymbol{x}) + \frac{1}{2\mu} \|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2 \geq F(\boldsymbol{x})$$

$$\Rightarrow \quad \widehat{F}_\mu(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) \geq F(\boldsymbol{x}_{k+1})$$

Convergence of Proximal Gradient

$$
\begin{aligned}
F(\boldsymbol{x}) - F(\boldsymbol{x}_{k+1}) &\geq F(\boldsymbol{x}) - \widehat{F}_{\mu}(\boldsymbol{x}_{k+1}, \boldsymbol{x}_k) \\
&\geq f(\boldsymbol{x}_k) + g(\boldsymbol{x}_{k+1}) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k \rangle + \langle \boldsymbol{\gamma}, \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle \\
&- \left(f(\boldsymbol{x}_k) + g(\boldsymbol{x}_{k+1}) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle + \frac{L}{2} \left\| \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \right\|_2^2 \right) \\
&= \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k \rangle + \langle \boldsymbol{\gamma}, \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle - \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle \\
&- \frac{L}{2} \left\| \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \right\|_2^2 \\
&= \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle + \langle \boldsymbol{\gamma}, \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle - \frac{L}{2} \left\| \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \right\|_2^2 \\
&= \langle \nabla f(\boldsymbol{x}_k) + \boldsymbol{\gamma}, \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle - \frac{L}{2} \left\| \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \right\|_2^2
\end{aligned}
$$

Convergence of Proximal Gradient

On the other hand, we have the following

$$\boldsymbol{x}_{k+1} = \min_{\boldsymbol{x}} \left\{ g(\boldsymbol{x}) + \frac{L}{2} \left\| \boldsymbol{x} - \left(\boldsymbol{x}_k - \frac{1}{L} \nabla f(\boldsymbol{x}_k) \right) \right\|_2^2 \right\} = \min_{\boldsymbol{x}} h(\boldsymbol{x})$$

Thus, we have

$$\mathbf{0} \in \partial h(\boldsymbol{x}_{k+1}) \Longrightarrow \boldsymbol{\gamma} + \nabla f(\boldsymbol{x}_k) + L(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) = \mathbf{0}, \quad \boldsymbol{\gamma} \in \partial g$$

Therefore, we have

$$F(\boldsymbol{x}) - F(\boldsymbol{x}_{k+1}) \geq \langle \nabla f(\boldsymbol{x}_k) + \boldsymbol{\gamma}, \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle - \frac{L}{2} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_2^2$$

$$\geq -\frac{L}{2} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_2^2 + \langle \nabla f(\boldsymbol{x}_k) + \boldsymbol{\gamma}, \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle$$

Convergence of Proximal Gradient

Therefore, we have

$$F(\boldsymbol{x}) - F(\boldsymbol{x}_{k+1}) \geq \langle \nabla f(\boldsymbol{x}_k) + \boldsymbol{\gamma}, \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle - \frac{L}{2} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_2^2$$

$$= -\frac{L}{2} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_2^2 + L \langle \boldsymbol{x}_k - \boldsymbol{x}_{k+1}, \boldsymbol{x} - \boldsymbol{x}_{k+1} \rangle - \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{x}_{k+1}\|_2^2$$

$$+ \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{x}_{k+1}\|_2^2$$

$$= \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{x}_{k+1}\|_2^2 - \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2$$

If $\boldsymbol{x} = \boldsymbol{x}_\star$, we have

$$0 \geq F(\boldsymbol{x}_\star) - F(\boldsymbol{x}_{k+1}) \geq \frac{L}{2} \|\boldsymbol{x}_\star - \boldsymbol{x}_{k+1}\|_2^2 - \frac{L}{2} \|\boldsymbol{x}_\star - \boldsymbol{x}_k\|_2^2$$

Convergence of Proximal Gradient

If $\boldsymbol{x} = \boldsymbol{x}_\star$, we have

$$0 \geq F(\boldsymbol{x}_\star) - F(\boldsymbol{x}_{k+1}) \geq \frac{L}{2} \|\boldsymbol{x}_\star - \boldsymbol{x}_{k+1}\|_2^2 - \frac{L}{2} \|\boldsymbol{x}_\star - \boldsymbol{x}_k\|_2^2$$

$$\Rightarrow \quad \|\boldsymbol{x}_\star - \boldsymbol{x}_{k+1}\|_2^2 \leq \|\boldsymbol{x}_\star - \boldsymbol{x}_k\|_2^2$$

Finally, we obtain

$$k\left(F(\boldsymbol{x}_k) - F(\boldsymbol{x}_\star)\right) \leq \sum_{i=0}^{k-1} \left(F(\boldsymbol{x}_i) - F(\boldsymbol{x}_\star)\right) \leq \frac{L}{2} \sum_{i=0}^{k} \left(\|\boldsymbol{x}_\star - \boldsymbol{x}_{i+1}\|_2^2 - \|\boldsymbol{x}_\star - \boldsymbol{x}_i\|_2^2 \right)$$

$$\leq \frac{L}{2} \|\boldsymbol{x}_\star - \boldsymbol{x}_0\|_2^2$$

So that $F(\boldsymbol{x}_k) - F(\boldsymbol{x}_\star) \leq \frac{L}{2k} \|\boldsymbol{x}_0 - \boldsymbol{x}_k\|_2^2$

PG Optimizes Smooth FBE Envelope

Definition. (Forward-backward envelope (FBE) $F(\boldsymbol{x})$ of)

$$Q_\mu(\boldsymbol{x}) := \min_{\boldsymbol{z}} \widehat{F}_\mu(\boldsymbol{z}, \boldsymbol{x})$$

$$Q_\mu(\boldsymbol{x}) = \underbrace{\min_{\boldsymbol{z}} \left\{ g(\boldsymbol{z}) + \frac{1}{2\mu} \|\boldsymbol{z} - (\boldsymbol{x} - \mu \nabla f(\boldsymbol{x}))\|_2^2 \right\}}_{\text{Moreau envelope } M_{\mu g}(\boldsymbol{x} - \mu \nabla f(\boldsymbol{x})) \text{ of } g(\boldsymbol{x})}$$

$$+ f(\boldsymbol{x}) - \frac{\mu}{2} \|\nabla f(\boldsymbol{x})\|_2^2$$

PG Optimizes Smooth FBE Envelope

Theorem. (Differentiability of FBE $Q_\mu(\boldsymbol{x})$)
Suppose f is twice continuously differentiable and L-smooth. Then $Q_\mu(\boldsymbol{x})$ is *continuously differentiable* with

$$\nabla Q_\mu(\boldsymbol{x}) \;=\; \big(\boldsymbol{I} - \mu\nabla^2 f(\boldsymbol{x})\big)\cdot G_\mu(\boldsymbol{x}).$$

If $\mu \in (0, 1/L)$, $G_\mu(\boldsymbol{x})$ is a descent direction on $Q_\mu(\boldsymbol{x})$, and the set of stationary points of $Q_\mu(\boldsymbol{x})$ equals

$$\operatorname{zer} \partial F \;=\; \{\boldsymbol{x} \in \mathbb{R}^n \;\mid\; 0 \in \partial F(\boldsymbol{x})\}.$$

PG Optimizes Smooth FBE Envelope

Whenever $\mu \in (0, 1/L)$, $Q_\mu(\boldsymbol{x})$ is continuously differentiable with

$$\langle G_\mu(\boldsymbol{x}), \nabla Q_\mu(\boldsymbol{x}) \rangle \; > \; 0$$

The proximal gradient

$$\boldsymbol{x}_{k+1} \; = \; \boldsymbol{x}_k - \mu \cdot G_\mu(\boldsymbol{x}_k)$$

can be viewed as a descent method on the smooth envelope $Q_\mu(\boldsymbol{x})$ of $F(\boldsymbol{x})$.

Convergence for Strongly Convex Functions

Theorem. Let $F(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$, with

- f is γ-strongly convex, differentiable, and L-smooth;
- g is convex, but possibly nonsmooth.

Consider the sequence of proximal gradient iterates

$$\boldsymbol{x}_{k+1} \;=\; \mathrm{prox}_{\mu_k g}\left(\boldsymbol{x}_k - \mu_k \nabla f(\boldsymbol{x}_k)\right), \quad \mu_k \equiv 1/L.$$

Let $\boldsymbol{x}_\star$ be a minimizer of F. Then for any $k \geq 1$,

$$\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2^2 \;\leq\; \left(1 - \frac{\gamma}{L}\right)^k \cdot \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2^2 .$$

Rate of Convergence

Methods	Convex Nonsmooth	Convex smooth	Strongly Convex
Subgradient Method	$O(1/\sqrt{k})$		
Gradient Descent		$O(1/k)$	Linear
Proximal Method	$O(1/k)$		Linear

Backtracking Linesearch for L

- **Proximal gradient:**

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{1}{L}\mathcal{G}_\mu(\boldsymbol{x}_k)$$

In many cases, the Lipschitz constant L is hard to compute

- **Example:**

$$F(\boldsymbol{x}) = \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\right\|_2^2 + \lambda\left\|\boldsymbol{x}\right\|_1$$

Backtracking Linesearch for L

Smooth case: checking Armijo condition for gradient descent, to guarantee sufficient decrease

$$f(\boldsymbol{x}_k - \tau_k \nabla f(\boldsymbol{x}_k)) \;\leq\; f(\boldsymbol{x}_k) - \frac{\tau_k}{2} \|\nabla f(\boldsymbol{x}_k)\|_2^2 .$$

which is equivalent to updating $\tau_k = 1/L_k$ until

$$f(\underbrace{\boldsymbol{x}_k - \tau_k \nabla f(\boldsymbol{x}_k)}_{\boldsymbol{x}_{k+1}}) \;\leq\; f(\boldsymbol{x}_k) - \frac{1}{L_k} \langle \nabla f(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_k) \rangle + \frac{1}{2L_k} \|\nabla f(\boldsymbol{x}_k)\|_2^2$$

$$\leq\; f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle + \frac{L_k}{2} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_2^2 .$$

Backtracking Stepsize Rule

- Similarly, for PG we have

$$\boldsymbol{x}_{k+1} \;=\; T_{\mu_k}(\boldsymbol{x}_k) \;:=\; \boldsymbol{x}_k - \mu_k G_{\mu_k}(\boldsymbol{x}_k).$$

- Thus, we would like to select μ_k such that

$$f\left(T_{\mu_k}(\boldsymbol{x}_k)\right) \;\leq\; f(\boldsymbol{x}_k) - \langle \nabla f(\boldsymbol{x}_k), \mu_k G_{\mu_k}(\boldsymbol{x}_k)\rangle + \frac{\mu_k}{2}\left\|G_{\mu_k}(\boldsymbol{x}_k)\right\|_2^2.$$

Backtracking Stepsize Rule

Algorithm 1 Backtracking Stepsize Rule for PG

Initialize $\mu_k = 1$, $\beta \in (0,1)$
while $f\left(T_{\mu_k}(\boldsymbol{x}_k)\right) \geq f(\boldsymbol{x}_k) - \mu_k \left\langle \nabla f(\boldsymbol{x}_k), G_{\mu_k}(\boldsymbol{x}_k) \right\rangle + \frac{\mu_k}{2} \left\| G_{\mu_k}(\boldsymbol{x}_k) \right\|_2^2$ **do**
  $\mu_k \leftarrow \beta \cdot \mu_k$
end while

In practice, we often do *warm restart* for μ_k to accelerate the speed

$$\mu_k = \alpha \cdot \mu_{k-1}, \quad \alpha > 1.$$

Further Readings

- *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications.* John Wright, Yi Ma. (**Chapter 8.2**)

- *Proximal Algorithms, Foundations & Trends in Optimization.* Neal Parikh, Stephen Boyd, 2014. https://web.stanford.edu/~boyd/papers/prox_algs.html

- *A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems.* Amir Beck and Marc Teboulle, 2009. *https://epubs.siam.org/doi/abs/10.1137/080716542?mobileUi=0*