7. Proximal gradient methods

- proximal gradient method
- convergence analysis
- estimate sequence
- accelerated proximal gradient method

Proximal mapping

in last lecture we discussed

- def of proximal mapping, examples, properties
- prox operator for norms, distances (to points, sets)
- projections onto 'simple' sets (hyperplanes, halfspaces, boxes, balls, cones,. . .)
- calculus rules for prox

Minimization of composite objective

problem class

$$\underset{x \in \mathbf{R}^n}{\text{minimize}} \quad \Big\{ F(x) \triangleq f(x) + \Psi(x) \Big\}$$

- f is convex and smooth (having Lipschitz-continuous gradient)
- Ψ is convex and closed, but may be nondifferentiable

assumption: Ψ is *simple*, meaning that its proximal mapping

$$\mathbf{prox}_{\Psi}(x) = \underset{u}{\text{argmin}} \left\{ \Psi(u) + \frac{1}{2}\|u - x\|_2^2 \right\}$$

has closed-form solution or is inexpensive to compute

Proximal gradient method

to minimize composite objective $f(x) + \Psi(x)$: choose $x^{(0)}$ and repeat

$$x^{(k+1)} = \mathbf{prox}_{t_k \Psi}\left(x^{(k)} - t_k \nabla f(x^{(k)})\right), \qquad k = 0, 1, 2, \ldots$$

where $t_k > 0$ is step size, constant or determined by line search

Interpretation: from definition of proximal operator

$$\begin{aligned} x^{(k+1)} &= \operatorname*{argmin}_u \left\{ \Psi(u) + \frac{1}{2t_k} \left\| u - x^{(k)} + t \nabla f(x^{(k)}) \right\|_2^2 \right\} \\ &= \operatorname*{argmin}_u \left\{ \Psi(u) + f(x^{(k)}) + \nabla f(x^{(k)})^T (u - x^{(k)}) + \frac{1}{2t_k} \|u - x^{(k)}\|_2^2 \right\} \end{aligned}$$

$x^{(k+1)}$ minimizes $\Psi(u)$ plus a simple quadratic model of $f(u)$ around $x^{(k)}$

Examples of proximal gradient method

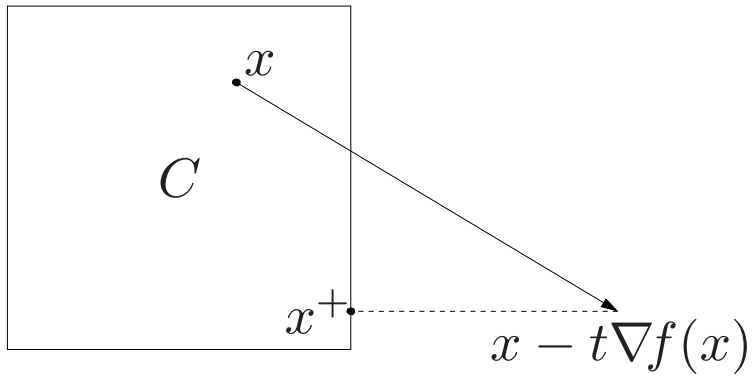
$$x^+ = \mathbf{prox}_{t\Psi}\left(x - t\nabla f(x)\right)$$

gradient method: $\Psi(x) = 0$, *i.e.*, minimize $f(x)$

$$x^+ = x - t\,\nabla f(x)$$

gradient projection method: $\Psi(x) = I_C(x)$, *i.e.*, $\underset{x\in C}{\text{minimize}}\ f(x)$

$$x^+ = P_C\left(x - t\nabla f(x)\right)$$

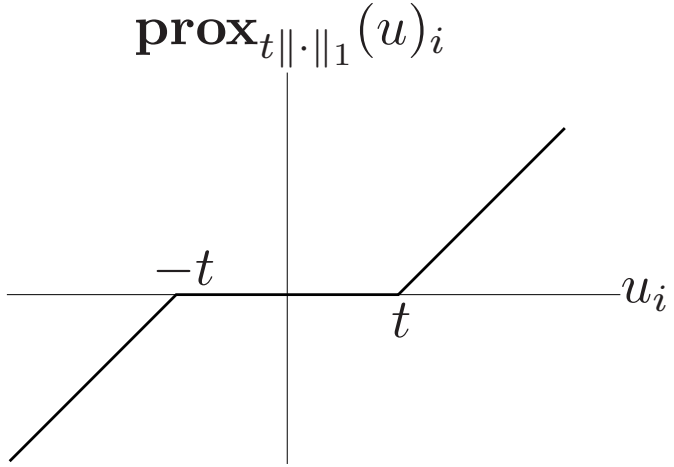


soft-thresholding: $\Psi(x) = \|x\|_1$, *i.e.*, minimize $f(x) + \|x\|_1$

$$x^+ = \mathbf{prox}_{t\|\cdot\|_1}\left(x - t\,\nabla f(x)\right)$$

where

$$\mathbf{prox}_{t\|\cdot\|_1}(u)_i = \begin{cases} u_i - t & u_i \geq t \\ 0 & |u_i| \leq t \\ u_i + t & u_i \leq -t \end{cases}$$



Convergence analysis of proximal gradient method

to minimize $F(x) = f(x) + \Psi(x)$, choose $x^{(0)}$ and repeat

$$x^{(k+1)} = \mathbf{prox}_{t_k \Psi}\left(x^{(k)} - t_k \nabla f(x^{(k)})\right), \qquad k = 0, 1, 2, \ldots$$

assumptions

- f is convex and $\nabla f(x)$ is Lipschitz continuous with constant L:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2, \quad \forall\, x, y \in \mathbf{dom}\, f$$

- Ψ is closed and convex ($\mathbf{prox}_{\Psi}$ is well defined), and $\mathbf{dom}\, \Psi \subseteq \mathbf{dom}\, f$
- optimal value $F^\star$ is finite and attained at $x^\star$ (not necessarily unique)

convergence rate: $O(1/k)$ with constant t_k or backtracking line search

Composite gradient mapping

an analog of gradient for the composite objective $F(x) = f(x) + \Psi(x)$

$$G_t(x) = \frac{1}{t}\big(x - \mathbf{prox}_{t\Psi}(x - t\nabla f(x))\big)$$

such that the proximal gradient iteration can be written as

$$x^+ = \mathbf{prox}_{t\Psi}\left(x - t\nabla f(x)\right) = x - tG_t(x)$$

- if $\Psi \equiv 0$, then $G_t(x) = \nabla F(x) = \nabla f(x)$ for any $t > 0$
- in general, $G_t(x)$ is *not* a gradient or subgradient of $F(x)$, instead

$$G_t(x) \in \nabla f(x) + \partial\Psi\left(x - tG_t(x)\right) \qquad \text{("forward looking" in } \Psi\text{)}$$

- $G_t(x) = 0$ if and only if x minimizes $F(x) = f(x) + \Psi(x)$

Consequences of smoothness assumption

recall quadratic upper bound derived from Lipschitz continuous gradient

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|_2^2, \quad \forall\, x, y \in \mathbf{dom}\, f$$

- substitute $y = x - tG_t(x)$:

$$f(x - tG_t(x)) \leq f(x) - t\nabla f(x)^T G_t(x) + \frac{t^2 L}{2}\|G_t(x)\|_2^2$$

- if $0 < t \leq 1/L$, then

$$f(x - tG_t(x)) \leq f(x) - t\nabla f(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 \qquad (1)$$

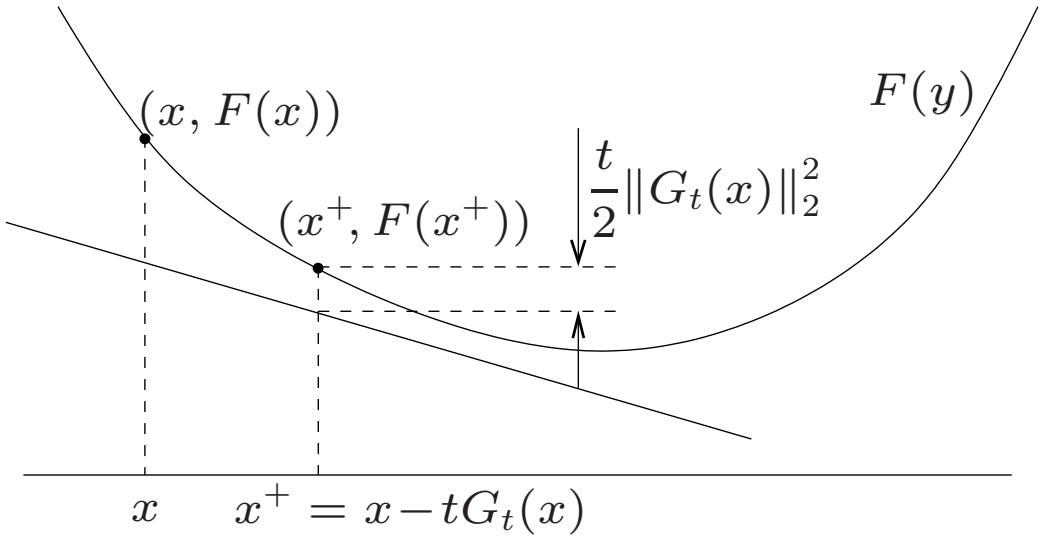
A forward-looking global lower bound

whenever the inequality (1) holds, then for all $y \in \mathbf{dom}\, F$,

$$F(y) \;\geq\; F(x - tG_t(x)) + G_t(x)^T(y - x) + \frac{t}{2}\|G_t(x)\|_2^2$$

forward-looking interpretation:

$$F(y) \;\geq\; F(x - tG_t(x)) + G_t(x)^T\big(y - (x - tG_t(x))\big) - \frac{t}{2}\|G_t(x)\|_2^2$$



proof:

use (1), convexity of f and Ψ, and $G_t(x) - \nabla f(x) \in \partial\Psi(x - tG_t(x))$

$$\begin{aligned}
F(x - tG_t(x)) &\leq f(x) - t\nabla f(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 + \Psi(x - tG_t(x)) \\
&\leq f(y) + \nabla f(x)^T(x - y) - t\nabla f(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 \\
&\quad + \Psi(y) + (G_t(x) - \nabla f(x))^T(x - y - tG_t(x)) \\
&= F(y) + G_t(x)^T(x - y) - \frac{t}{2}\|G_t(x)\|_2^2
\end{aligned}$$

therefore

$$F(y) \geq F(x - tG_t(x)) + G_t(x)^T(y - x) + \frac{t}{2}\|G_t(x)\|_2^2 \qquad (2)$$

Progress in one iteration

$$x^+ = x - tG_t(x) = \mathbf{prox}_{t\Psi}(x - t\nabla f(x))$$

- inequality (2) with $y = x$ shows the alg is a descent method:

$$F(x^+) \;\leq\; F(x) - \frac{t}{2}\|G_t(x)\|_2^2$$

- inequality (2) with $y = x^\star$ yields

$$\begin{aligned} F(x^+) - F^\star &\leq G_t(x)^T(x - x^\star) - \frac{t}{2}\|G_t(x)\|_2^2 \\ &= \frac{1}{2t}\left(\|x - x^\star\|_2^2 - \|x - x^\star - tG_t(x)\|_2^2\right) \\ &= \frac{1}{2t}\left(\|x - x^\star\|_2^2 - \|x^+ - x^\star\|_2^2\right) \end{aligned} \tag{3}$$

(hence, $\|x^+ - x^\star\|_2 \leq \|x - x^\star\|_2$, i.e., distance to optimal set decreases)

Analysis for fixed step size

add inequalities (3) for $x = x^{(i-1)}$, $x^+ = x^{(i)}$, $t_i = t \leq 1/L$

$$\begin{aligned}
\sum_{i=1}^{k} \left(F(x^{(i)}) - F^\star \right) &\leq \frac{1}{2t} \sum_{i=1}^{k} \left(\|x^{(i-1)} - x^\star\|_2^2 - \|x^{(i)} - x^\star\|_2^2 \right) \\
&= \frac{1}{2t} \left(\|x^{(0)} - x^\star\|_2^2 - \|x^{(k)} - x^\star\|_2^2 \right) \\
&\leq \frac{1}{2t} \|x^{(0)} - x^\star\|_2^2
\end{aligned}$$

since $F(x^{(i)})$ is non-increasing,

$$F(x^{(k)}) - F^\star \;\leq\; \frac{1}{k} \sum_{i=1}^{k} \left(F(x^{(i)}) - F^\star \right) \;\leq\; \frac{1}{2kt} \|x^{(0)} - x^\star\|_2^2$$

conclusion: $F(x^{(k)}) - F^\star \leq \epsilon$ after $O(1/\epsilon)$ iterations

Line search

to determine step size t in

$$x^+ = x - tG_t(x)$$

start at some $t := \hat{t}$ and repeat $t := \beta t$ (with $0 < \beta < 1$) until

$$f(x - tG_t(x)) \leq f(x) - t\nabla f(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2$$

guaranteed to hold if $t \leq 1/L$ (consequence of quadratic upper bound)

- selected step size t satisfies $t \geq t_{\min} \triangleq \min\{\hat{t}, \beta/L\}$
- requires one $\mathbf{prox}$ evaluation per line search iteration (backtracking)
- several other types of line search also work

Analysis with line search

from page 7–12, if line search stopping (same as ineq. (1)) holds in iteration i, then $F(x^{(i)}) < F(x^{(i-1)})$ and

$$\begin{aligned} F(x^{(i)}) - F^\star &\leq \frac{1}{2t_i}\left(\|x^{(i-1)} - x^\star\|_2^2 - \|x^{(i)} - x^\star\|_2^2\right) \\ &\leq \frac{1}{2t_{\min}}\left(\|x^{(i-1)} - x^\star\|_2^2 - \|x^{(i)} - x^\star\|_2^2\right) \end{aligned}$$

- adding inequalities for $i = 1$ to $i = k$ gives

$$\sum_{i=1}^{k}\left(F(x^{(i)}) - F^\star\right) \leq \frac{1}{2t_{\min}}\|x^{(0)} - x^\star\|_2^2$$

- since $F(x^{(i)})$ is non-increasing, obtain similar $1/k$ convergence bound

as for fixed t_i:

$$F(x^{(k)}) - F^\star \;\leq\; \frac{1}{2kt_{\min}}\|x^{(0)} - x^\star\|_2^2$$

Convergence analysis for strongly convex functions

$$\text{minimize} \quad \left\{ F(x) \triangleq f(x) + \Psi(x) \right\}$$

assumptions

- f is convex and $\nabla f(x)$ is Lipschitz continuous with constant L:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \, \|x - y\|_2, \quad \forall \, x, y \in \mathbf{dom} \, f$$

- Ψ is closed and convex ($\mathbf{prox}_\Psi$ is well defined), and $\mathbf{dom} \, \Psi \subseteq \mathbf{dom} \, f$
- f and Ψ have strong convexity parameters μ_f and μ_Ψ, respectively, and

$$\mu_F = \mu_f + \mu_\Psi > 0$$

proximal gradient method

$$x^{(k+1)} = \mathbf{prox}_{t_k \Psi} \left(x^{(k)} - t_k \nabla f(x^{(k)}) \right), \qquad k = 0, 1, 2, \ldots$$

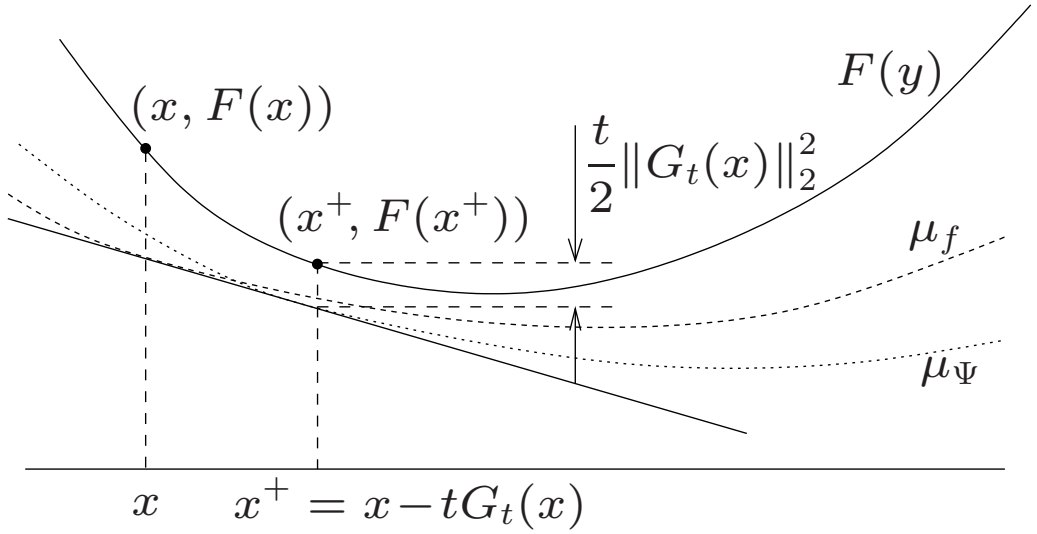
Forward-looking global lower bound

denote $x^+ = x - tG_t(x)$ and assume (1) holds, then

$$F(y) \geq F(x^+) + G_t(x)^T(y-x) + \frac{t}{2}\|G_t(x)\|_2^2 + \frac{\mu_f}{2}\|y-x\|_2^2 + \frac{\mu_\Psi}{2}\|y-x^+\|_2^2$$

forward-looking interpretation:

$$F(y) \geq F(x^+) + G_t(x)^T(y-x^+) - \frac{t}{2}\|G_t(x)\|_2^2 + \frac{\mu_f}{2}\|y-x\|_2^2 + \frac{\mu_\Psi}{2}\|y-x^+\|_2^2$$



proof:

use (1), strong convexity of f and Ψ, and $G_t(x) - \nabla f(x) \in \partial\Psi(x^+)$

$$\begin{aligned}
F(x^+) &\leq f(x) - t\nabla f(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 + \Psi(x^+) \\
&\leq f(y) + \nabla f(x)^T(x-y) - \frac{\mu_f}{2}\|y-x\|_2^2 \\
&\quad -t\nabla f(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 \\
&\quad + \Psi(y) + (G_t(x) - \nabla f(x))^T(x^+ - y) - \frac{\mu_\Psi}{2}\|y - x^+\|_2^2 \\
&= F(y) + G_t(x)^T(x-y) - \frac{t}{2}\|G_t(x)\|_2^2 \\
&\quad -\frac{\mu_f}{2}\|y-x\|_2^2 - \frac{\mu_\Psi}{2}\|y-x^+\|_2^2
\end{aligned}$$

therefore

$$F(y) \geq F(x^+) + G_t(x)^T(y-x) + \frac{t}{2}\|G_t(x)\|_2^2 + \frac{\mu_f}{2}\|y-x\|_2^2 + \frac{\mu_\Psi}{2}\|y-x^+\|_2^2$$

Two inequalities

- forward-looking lower bound with $y = x$:

$$F(x^+) \;\leq\; F(x) - \frac{t(1+t\mu_\Psi)}{2}\|G_t(x)\|_2^2$$

- forward-looking lower bound with $y = x^\star$:

$$G_t(x)^T(x - x^\star) \;\geq\; \frac{t}{2}\|G_t(x)\|_2^2 + \frac{\mu_f}{2}\|x - x^\star\|_2^2 + \frac{\mu_\Psi}{2}\|x^+ - x^\star\|_2^2$$

compare with two fundamental inequalities for smooth functions

$$\begin{aligned} f\left(x - \frac{1}{L}\nabla f(x)\right) &\leq f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2 \\ \nabla f(x)^T(x - x^\star) &\geq \frac{1}{L}\|\nabla f(x)\|_2^2 \end{aligned}$$

Analysis for constant step size

$$x^{(k+1)} = \mathbf{prox}_{t\Psi}\left(x^{(k)} - t\nabla f(x^{(k)})\right) = x^{(k)} - tG_t(x^{(k)})$$

theorem: suppose $\mu_f + \mu_\Psi > 0$ and $t = 1/L$, then for any $k \geq 0$

$$\|x^{(k)} - x^\star\|_2^2 \leq \left(\frac{L - \mu_f}{L + \mu_\Psi}\right)^k \|x^{(0)} - x^\star\|_2^2$$

proof:

$$\begin{aligned}
\|x^+ - x^\star\|_2^2 &= \|x - x^\star - tG_t(x)\|_2^2 \\
&= \|x - x^\star\|_2^2 - 2tG_t(x)^T(x - x^\star) + t^2\|G_t(x)\|_2^2 \\
&\leq \|x - x^\star\|_2^2 - t\mu_f\|x - x^\star\|_2^2 - t\mu_\Psi\|x^+ - x^\star\|_2^2
\end{aligned}$$

therefore

$$\|x^+ - x^\star\|_2^2 \leq \frac{1 - t\mu_f}{1 + t\mu_\Psi}\|x - x^\star\|_2^2$$

Outline

- proximal gradient method
- **estimate sequence**
- accelerated proximal gradient methods

Fast proximal gradient methods history

- Nesterov '83, '88, '05: three gradient projection methods with rate $1/k^2$
- Beck and Teboulle '08: FISTA, prox-grad version of Nesterov's '83 method
- Nesterov '04 (book), Tseng '08: overview & unified analysis of fast gradient and prox-grad methods
- several recent extensions

Estimate sequence (Nesterov)

a pair of sequences $\{\lambda_k, \phi_k(x)\}_{k=0}^{\infty}$ is called *estimate sequence* of $F(x)$ if

- $\lambda_k \to 0$
- $\phi_k(x) \leq (1-\lambda_k)F(x) + \lambda_k\phi_0(x)$ for any $x \in \mathbf{R}^n$ and all $k > 0$

lemma: if a sequence $\{x^{(k)}\}$ satisfies $F(x^{(k)}) \leq \min_{x\in\mathbf{R}^n} \phi_k(x)$, then

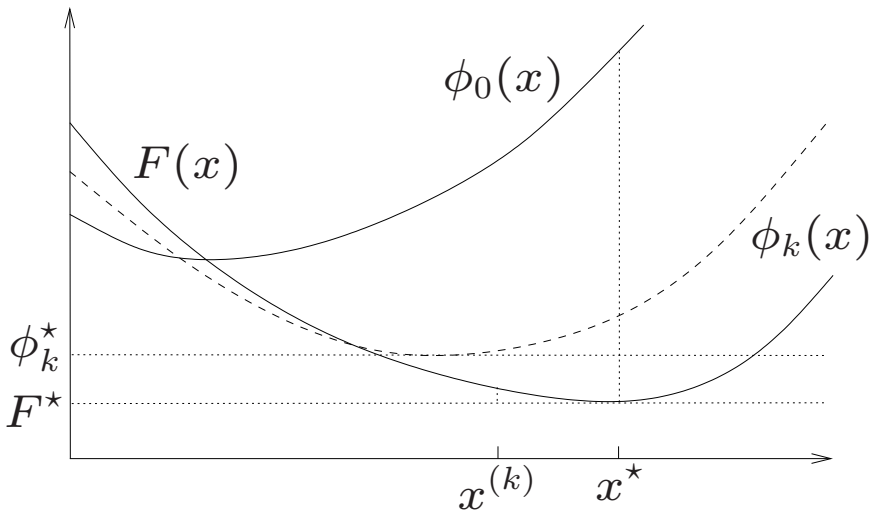
$$F(x^{(k)}) - F^\star \;\leq\; \lambda_k\,(\phi_0(x^\star) - F^\star) \;\;\to 0$$

proof:

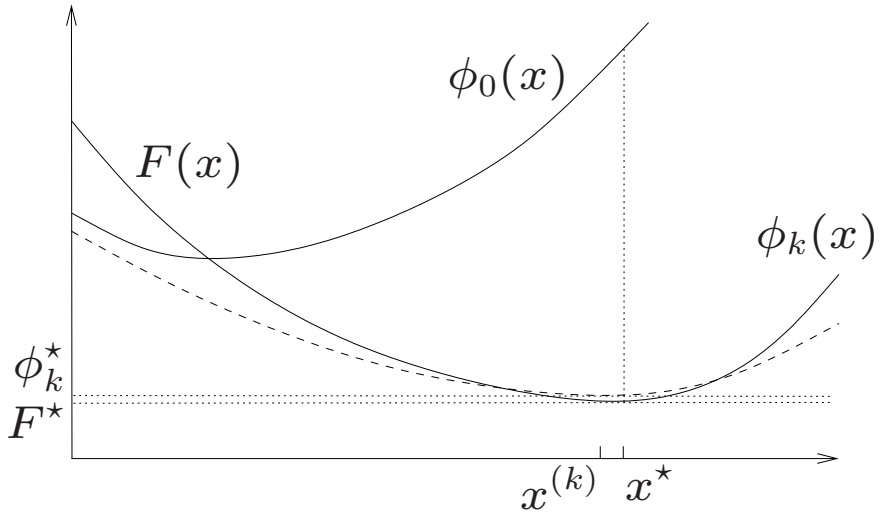
$$\begin{aligned} F(x^{(k)}) &\leq \min_{x\in\mathbf{R}^n} \phi_k(x) \;\leq\; \min_{x\in\mathbf{R}^n} \big\{(1-\lambda_k)F(x) + \lambda_k\phi_0(x)\big\} \\ &\leq (1-\lambda_k)F(x^\star) + \lambda_k\phi_0(x^\star) \\ &= F(x^\star) + \lambda_k\,(\phi_0(x^\star) - F(x^\star)) \end{aligned}$$

estimate sequence: pair of sequences $\{\lambda_k, \phi_k(x)\}_{k=0}^{\infty}$ such that

- $\lambda_k \to 0$
- $\phi_k(x) \leq (1-\lambda_k)F(x) + \lambda_k\phi_0(x)$ for any $x \in \mathbf{R}^n$ and all $k > 0$



k small



k large

questions:

- how to form the estimate sequence?
- how can we ensure $F(x^{(k)}) \leq \phi_k^\star \triangleq \min_{x \in \mathbf{R}^n} \phi_k(x)$?

lemma: suppose $f \in \mathcal{S}_{\mu,L}(\mathbf{R}^n)$, and the sequence $\{\alpha_k\}_{k=0}^{\infty}$ satisfies

$$\alpha_k \in (0,1), \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

then for any function $\phi_0(x)$ and a sequence of functions $\{\psi(y^{(k)}; x)\}$ such that $\psi(y^{(k)}; x) \leq F(x)$ for all k, the following pair is an estimate sequence

$$\lambda_{k+1} = (1-\alpha_k)\lambda_k, \quad \text{with } \lambda_0 = 1$$
$$\phi_{k+1}(x) = (1-\alpha_k)\phi_k(x) + \alpha_k\, \psi(y^{(k)}; x)$$

proof: first, $\phi_0(x) \leq (1-\lambda_0)F(x) + \lambda_0\phi_0(x) = \phi_0(x)$; then

$$\begin{aligned}
\phi_{k+1}(x) &\leq (1-\alpha_k)\phi_k(x) + \alpha_k F(x) \\
&= \big(1-(1-\alpha_k)\lambda_k\big)F(x) + (1-\alpha_k)\big(\phi_k(x) - (1-\lambda_k)F(x)\big) \\
&\leq \big(1-(1-\alpha_k)\lambda_k\big)F(x) + (1-\alpha_k)\lambda_k\phi_0(x) \\
&= (1-\lambda_{k+1})F(x) + \lambda_{k+1}\phi_0(x)
\end{aligned}$$

Variants of Nesterov's accelerated methods

in Lecture 3 ($\Psi(x) \equiv 0$), estimate sequence constructed using the following

$$\begin{aligned} \phi_0(x) &= f(x^{(0)}) + \frac{\gamma_0}{2}\|x - x^{(0)}\|_2^2 \\ \psi(y; x) &= f(y) + \nabla f(y)^T(x - y) + \frac{\mu}{2}\|x - y\|_2^2 \end{aligned}$$

three variants for nontrivial $\Psi(x)$

- Nesterov's 1st method (1983, 2004); see also Beck & Teboulle (2009)

$$\begin{aligned} \phi_0(x) &= F(x^{(0)}) + \frac{\gamma_0}{2}\|x - x^{(0)}\|_2^2 \\ \psi(y; x) &= F\big(y - tG_t(y)\big) + G_t(y)^T(x - y) + \frac{t}{2}\|G_t(y)\|_2^2 + \frac{\mu}{2}\|x - y\|_2^2 \end{aligned}$$

  i.e., use composite gradient mapping and forward-looking lower bound

- Nesterov's 2nd method (1988); see also Tseng (2008)

$$\begin{aligned}\phi_0(x) &= f(x^{(0)}) + \frac{\gamma_0}{2}\|x - x^{(0)}\|_2^2 + \Psi(x)\\ \psi(y;x) &= f(y) + \nabla f(y)^T(x-y) + \frac{\mu}{2}\|x-y\|_2^2 + \Psi(x)\end{aligned}$$

  - use lower bound for f but keep Ψ untouched
  - update $\phi_k(x)$ iteratively (use $\mathbf{prox}_\Psi$ on the new gradient at $y^{(k)}$)

- Nesterov's 3rd method (2005, 2007)

$$\begin{aligned}\phi_0(x) &= F(x^{(0)}) + \frac{\gamma_0}{2}\|x - x^{(0)}\|_2^2\\ \psi(y;x) &= f(y) + \nabla f(y)^T(x-y) + \frac{\mu}{2}\|x-y\|_2^2 + \Psi(x)\end{aligned}$$

  - use lower bound for f but keep Ψ untouched
  - update $\phi_k(x)$ in a batch mode (use $\mathbf{prox}_\Psi$ on accumulated gradient)

all three variants recover the same method when $\Psi(x) \equiv 0$

Nesterov's 1st method

let $\phi_0(x) = F(v^{(0)}) + \frac{\gamma_0}{2}\|x - v_0\|^2$, then $\{\phi_k(x)\}$ can be written as

$$\phi_k(x) = \phi_k^\star + \frac{\gamma_k}{2}\|x - v^{(k)}\|^2,$$

where

$$\begin{aligned}
\gamma_{k+1} &= (1-\alpha_k)\gamma_k + \alpha_k\mu \\
v^{(k+1)} &= \frac{1}{\gamma_{k+1}}\Big((1-\alpha_k)\gamma_k v^{(k)} + \alpha_k\mu y^{(k)} - \alpha_k G_{1/L}(y^{(k)})\Big) \\
\phi_{k+1}^\star &= (1-\alpha_k)\phi_k^\star + \alpha_k F(y^{(k)+}) + \left(\frac{\alpha_k}{2L} - \frac{\alpha_k^2}{2\gamma_{k+1}}\right)\|G_{1/L}(y^{(k)})\|^2 \\
&\quad + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\left\langle G_{1/L}(y^{(k)}), v^{(k)} - y^{(k)}\right\rangle + \frac{\mu}{2}\|y^{(k)} - v^{(k)}\|^2\right)
\end{aligned}$$

notation: $y^{(k)+} = y^{(k)} - (1/L)G_{1/L}(y^{(k)})$

construction via induction: assume $\phi_k^\star \geq F(x^{(k)})$ is true, then

$$\begin{aligned}\phi_{k+1}^\star \quad &\geq \quad (1-\alpha_k)F(x^{(k)}) + \alpha_k F(y^{(k)+}) + \left(\frac{\alpha_k}{2L} - \frac{\alpha_k^2}{2\gamma_{k+1}}\right)\|G_{1/L}(y^{(k)})\|^2 \\ &\qquad + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left\langle G_{1/L}(y^{(k)}), v^{(k)} - y^{(k)}\right\rangle\end{aligned}$$

use forward-looking lower bound $F(x^{(k)}) \geq F(y^{(k)+}) + \cdots$

$$\begin{aligned}\phi_{k+1}^\star \quad &\geq \quad F(y^{(k)+}) + \left(\frac{1}{2L} - \frac{\alpha_k^2}{2\gamma_{k+1}}\right)\|G_{1/L}(y^{(k)})\|^2 \\ &\qquad + (1-\alpha_k)\left\langle G_{1/L}(y^{(k)}),\ \frac{\alpha_k\gamma_k}{\gamma_{k+1}}(v^{(k)} - y^{(k)}) + x^{(k)} - y^{(k)}\right\rangle\end{aligned}$$

finally, in order to make $\phi_{k+1}^\star \geq F(x^{(k+1)})$

- choose $x^{(k+1)} = y^{(k)+}$ and choose α_k such that $\left(\frac{1}{2L} - \frac{\alpha_k^2}{2\gamma_{k+1}}\right) = 0$
- choose $y^{(k)}$ so that $\frac{\alpha_k\gamma_k}{\gamma_{k+1}}(v^{(k)} - y^{(k)}) + x^{(k)} - y^{(k)} = 0$

Choose $\{\alpha_k\}$, $\{y^{(k)}\}$ and $\{x^{(k+1)}\}$

- choose α_k by solving the equation $\frac{\alpha_k^2}{\gamma_{k+1}} = \frac{1}{L}$, that is

$$L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\mu \tag{4}$$

- choose $y^{(k)}$ to eliminate inner-product term

$$y^{(k)} = \frac{1}{\gamma_k + \alpha_k\mu}(\alpha_k\gamma_k v^{(k)} + \gamma_{k+1}x^{(k)})$$

- choose $x^{(k+1)} = y^{(k)+}$, that is

$$x^{(k+1)} = y^{(k)} - \frac{1}{L}G_{\frac{1}{L}}(y^{(k)}) = \mathbf{prox}_{\frac{1}{L}\Psi}\left(y^{(k)} - \frac{1}{L}\nabla f(y^{(k)})\right)$$

Accelerated proximal gradient method

- choose $x_0 \in \mathbf{R}^n$ and $\gamma_0 > 0$, and set $v_0 = x_0$
- for $k = 0, 1, 2, \ldots$, repeat

  1. find $\alpha_k \in (0, 1)$ that satisfies the equation

  $$L\alpha_k^2 = (1 - \alpha_k)\gamma_k + \alpha_k\mu$$

  and let $\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \alpha_k\mu$

  2. choose

  $$y^{(k)} = \frac{1}{\gamma_k + \alpha_k\mu}(\alpha_k\gamma_k v^{(k)} + \gamma_{k+1}x^{(k)})$$

  3. proximal gradient step

  $$x^{(k+1)} = \mathbf{prox}_{\frac{1}{L}\Psi}\left(y^{(k)} - \frac{1}{L}\nabla f(y^{(k)})\right)$$

  4. set

  $$v^{(k+1)} = \frac{1}{\gamma_{k+1}}\Big((1 - \alpha_k)\gamma_k v^{(k)} + \alpha_k\mu y^{(k)} - \alpha_k G_{1/L}(y^{(k)})\Big)$$

Rate of convergence

only need to bound the growth of λ_k, same as analysis on page 3-17

lemma: if $\gamma_0 \geq \mu$ in the accelerated scheme on page 7–31, then

$$\lambda_k \;=\; \prod_{i=0}^{k-1}(1-\alpha_i) \;\leq\; \min\left\{\left(1-\sqrt{\frac{\mu}{L}}\right)^k,\; \frac{4L}{(2\sqrt{L}+k\sqrt{\gamma_0})^2}\right\}$$

theorem: let $\gamma_0 = L$, then the method on page 7–31 generates $\{x^{(k)}\}_{k=0}^{\infty}$ such that

$$f(x^{(k)}) - f^\star \leq \min\left\{\left(1-\sqrt{\frac{\mu}{L}}\right)^k,\; \frac{4}{(k+2)^2}\right\} L\|x_0 - x^\star\|^2$$

Variant of Nesterov's 1st method

can eliminate $\{v^{(k)}\}$ and $\{\gamma_k\}$, and use constant step size $t = 1/L$

- choose $x^{(0)} \in \mathbf{R}^n$ and $\alpha_0 \in (\sqrt{\frac{\mu}{L}}, 1)$, set $y^{(0)} = x^{(0)}$ and $q = \mu/L$
- for $k = 0, 1, 2, \ldots$, repeat
  1. compute $f(y^{(k)})$ and $\nabla f(y^{(k)})$, set

$$x^{(k+1)} = \mathbf{prox}_{\frac{1}{L}\Psi}\left(y^{(k)} - \frac{1}{L}\nabla f(y^{(k)})\right)$$

  2. compute $\alpha_{k+1} \in (0, 1)$ from equation

$$\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 + q\alpha_{k+1}$$

  and set $\beta_k = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2+\alpha_{k+1}}$ and

$$y^{(k+1)} \;=\; x^{(k+1)} + \beta_k(x^{(k+1)} - x^{(k)})$$

A simpler variant when $\mu > 0$

if $\alpha_0 = \sqrt{\frac{\mu}{L}}$, which corresponds to $\gamma_0 = \mu$, then

$$\alpha_k = \sqrt{\frac{\mu}{L}}, \qquad \beta_k = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$$

this leads to the following simple method

- choose $y^{(0)} = x^{(0)} \in \mathbf{R}^n$
- for $k = 0, 1, 2, \ldots$, repeat

$$\begin{aligned} x^{(k+1)} &= \mathbf{prox}_{\frac{1}{L}\Psi}\left(y^{(k)} - \frac{1}{L}\nabla f(y^{(k)})\right) \\ y^{(k+1)} &= x^{(k+1)} + \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}(x^{(k+1)} - x^{(k)}) \end{aligned}$$

Choosing $\{\alpha_k\}$ when $\mu = 0$ or unknown

when $\mu = 0$ or unknown, can replace equation (4) with

$$L\alpha_{k+1}^2 \;\geq\; (1-\alpha_{k+1})\gamma_{k+1} \;=\; (1-\alpha_{k+1})L\alpha_k^2$$

this leads to the condition

$$\frac{1-\alpha_{k+1}}{\alpha_{k+1}^2} \leq \frac{1}{\alpha_k^2}, \qquad k \geq 0 \tag{5}$$

which guarantees $O(1/k^2)$ convergence rate (see page 7–32)

- can use a fixed sequence that satisfies (5): e.g, $\alpha_k = \frac{2}{k+2}$
- or use solution with equality: $\alpha_{k+1} = \frac{1}{2}\left(\sqrt{\alpha_k^4 + 4\alpha_k^2} - \alpha_k^2\right)$ with $\alpha_0 = 1$
- from page 7–33, $\beta_k = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2+\alpha_{k+1}}$; with equality solution, $\beta_k = \frac{\alpha_{k+1}(1-\alpha_k)}{\alpha_k}$

A simple variant when $\mu = 0$ or unknown

- choose $y^{(0)} = x^{(0)} \in \mathbf{R}^n$, and set $\alpha_0 = 1$
- for $k = 0, 1, 2, \ldots$, repeat

$$\begin{aligned} x^{(k+1)} &= \mathbf{prox}_{\frac{1}{L}\Psi}\left(y^{(k)} - \frac{1}{L}\nabla f(y^{(k)})\right) \\ \alpha_{k+1} &= \frac{1}{2}\left(\sqrt{\alpha_k^4 + 4\alpha_k^2} - \alpha_k^2\right) \\ y^{(k+1)} &= x^{(k+1)} + \frac{\alpha_{k+1}(1-\alpha_k)}{\alpha_k}(x^{(k+1)} - x^{(k)}) \end{aligned}$$

let $\eta_k = 1/\alpha_k$, then last two steps become (FISTA)

$$\begin{aligned} \eta_{k+1} &= \frac{1 + \sqrt{1 + 4\eta_k^2}}{2} \\ y^{(k+1)} &= x^{(k+1)} + \frac{\eta_k - 1}{\eta_{k+1}}(x^{(k+1)} - x^{(k)}) \end{aligned}$$

A even simpler variant

use predefined sequence

$$\alpha_k = \frac{2}{k+2}, \qquad \beta_k = \frac{\alpha_{k+1}(1-\alpha_k)}{\alpha_k}$$

to arrive at

- choose $y^{(0)} = x^{(0)} \in \mathbf{R}^n$
- for $k = 0, 1, 2, \ldots$, repeat

$$\begin{aligned} x^{(k+1)} &= \mathbf{prox}_{\frac{1}{L}\Psi}\left(y^{(k)} - \frac{1}{L}\nabla f(y^{(k)})\right) \\ y^{(k+1)} &= x^{(k+1)} + \frac{k}{k+3}(x^{(k+1)} - x^{(k)}) \end{aligned}$$

Nesterov's 2nd method

algorithm: choose $x^{(0)} = v^{(0)}$ and repeat the following for $k = 0, 1, 2, \ldots$

$$\begin{aligned} y^{(k)} &= (1-\alpha_k)x^{(k)} + \alpha_k v^{(k)} \\ v^{(k+1)} &= \mathbf{prox}_{\frac{1}{L\alpha_k}\Psi}\left(v^{(k)} - \frac{1}{L\alpha_k}\nabla f(y^{(k)})\right) \\ x^{(k+1)} &= (1-\alpha_k)x^{(k)} + \alpha_k v^{(k+1)} \end{aligned}$$

- $O(1/k^2)$ convergence guaranteed by same condition (5), repeated here:

$$\frac{1-\alpha_{k+1}}{\alpha_{k+1}^2} \leq \frac{1}{\alpha_k^2}, \qquad k \geq 0$$

- can use $\alpha_{k+1} = \frac{1}{2}\left(\sqrt{\alpha_k^4 + 4\alpha_k^2} - \alpha_k^2\right)$ or pre-fixed series like $\alpha_k = \frac{2}{k+2}$
- unlike in the 1st method, all three sequences are feasible (in $\mathbf{dom}\,\Psi$)

Nesterov's 3rd method

algorithm: choose $x^{(0)} = v^{(0)}$ and repeat the following for $k = 0, 1, 2, \ldots$

$$\begin{aligned} y^{(k)} &= (1-\alpha_k)x^{(k)} + \alpha_k v^{(k)} \\ v^{(k+1)} &= \mathbf{prox}_{\left(\frac{1}{L}\sum_{i=1}^{k}\frac{1}{\alpha_i}\right)\Psi}\left(v^{(0)} - \frac{1}{L}\sum_{i=1}^{k}\frac{1}{\alpha_i}\nabla f(y^{(i)})\right) \\ x^{(k+1)} &= (1-\alpha_k)x^{(k)} + \alpha_k v^{(k+1)} \end{aligned}$$

- $O(1/k^2)$ convergence guaranteed by same condition (5), repeated here:

$$\frac{1-\alpha_{k+1}}{\alpha_{k+1}^2} \leq \frac{1}{\alpha_k^2}, \qquad k \geq 0$$

- can use $\alpha_{k+1} = \frac{1}{2}\left(\sqrt{\alpha_k^4 + 4\alpha_k^2} - \alpha_k^2\right)$ or pre-fixed series like $\alpha_k = \frac{2}{k+2}$
- 2nd and 3rd methods are variants as presented by Tseng (2008)

References

- L. Vandenberghe, *Lecture notes for EE236C - Optimization Methods for Large-Scale Systems* (Spring 2012), UCLA.

- Yu. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course* (2004), Section 2.2.

- Yu. Nesterov, *Smooth minimizatin of non-smooth functions*, Mathematical Programming (2005).

- P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization* (2008).

- A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences (2009)

- N. Parikh and S. Boyd, *Proximal algorithms*, 2013.