Proximal Gradient Method

Pontus Giselsson

Outline

- **Introducing proximal gradient method and examples**
- Solving composite problem – Fixed-points and convergence
- Application to primal and dual problems

Composite optimization problems

- We have introduced the composite optimization problem

$$\underset{x}{\text{minimize}}\, f(Lx) + g(x)$$

- Need an algorithm that solves it - proximal gradient method
- We will consider the simpler composite optimization problem

$$\underset{x}{\text{minimize}}\, f(x) + g(x)$$

  that gives the former by letting $f \to f \circ L$

Problem assumptions

- Proximal gradient method works, e.g., for problems that satisfy
  - f is β-smooth $f : \mathbb{R}^n \to \mathbb{R}$ (not necessarily convex)
  - g is closed convex
- Recall that if β-smoothness implies that f satisfies

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{\beta}{2}\|y-x\|_2^2$$
$$f(y) \geq f(x) + \nabla f(x)^T(y-x) - \frac{\beta}{2}\|y-x\|_2^2$$

  it has convex quadratic upper and concave quadratic lower bounds

- If f in addition is convex, we instead have

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{\beta}{2}\|y-x\|_2^2$$
$$f(y) \geq f(x) + \nabla f(x)^T(y-x)$$

  where the concave quadratic lower bound is replaced by affine

Minimizing upper bound

- Due to β-smoothness of f, we have

$$f(y) + g(y) \le f(x) + \nabla f(x)^T (y - x) + \tfrac{\beta}{2}\|y - x\|_2^2 + g(y)$$

for all $x, y \in \mathbb{R}^n$, i.e., r.h.s. is upper bound to l.h.s.

- Minimizing in every iteration the r.h.s. w.r.t. y for given x gives

$$\begin{aligned} v &= \operatorname*{argmin}_y \Big(f(x) + \nabla f(x)^T (y - x) + \tfrac{\beta}{2}\|y - x\|_2^2 + g(y) \Big) \\ &= \operatorname*{argmin}_y \Big(g(y) + \tfrac{\beta}{2}\|y - (x - \beta^{-1}\nabla f(x))\|_2^2 \Big) \\ &= \operatorname{prox}_{\beta^{-1} g}(x - \beta^{-1}\nabla f(x)) \end{aligned}$$

Proximal gradient method

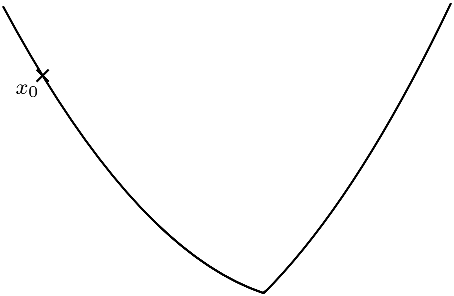
- Let us replace β by γ_k^{-1}, x by x_k, and v by x_{k+1} to get:

$$\begin{aligned} x_{k+1} &= \underset{y}{\operatorname{argmin}} \left(f(x_k) + \nabla f(x_k)^T (y - x_k) + \tfrac{1}{2\gamma_k} \|y - x_k\|_2^2 + g(y) \right) \\ &= \underset{y}{\operatorname{argmin}} \left(g(y) + \tfrac{1}{2\gamma_k} \|y - (x_k - \gamma_k \nabla f(x_k))\|_2^2 \right) \\ &= \operatorname{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k)) \end{aligned}$$

- This is exactly the proximal gradient method
- The method replaces f by quadratic approximation and minimizes
- (Note that we need an initial guess x_0 to start the iteration)

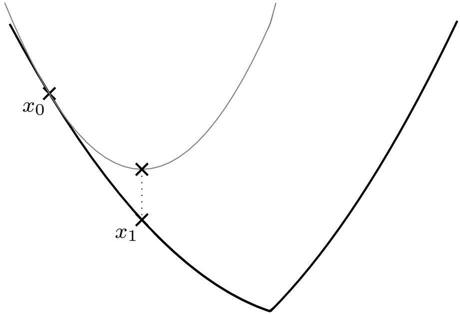
Proximal gradient – Example

- Proximal gradient iterations for problem $\underset{x}{\text{minimize}}\, \frac{1}{2}(x-a)^2 + |x|$
- $f(x) = \frac{1}{2}(x-a)^2$ is smooth term and $g(x) = |x|$ is nonsmooth
- Iteration: $x_{k+1} = \text{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k))$
- Note: convergence in finite number of iterations (not always)



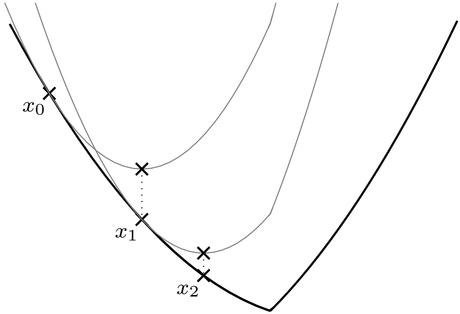
Proximal gradient – Example

- Proximal gradient iterations for problem $\underset{x}{\text{minimize}}\, \frac{1}{2}(x-a)^2 + |x|$
- $f(x) = \frac{1}{2}(x-a)^2$ is smooth term and $g(x) = |x|$ is nonsmooth
- Iteration: $x_{k+1} = \text{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k))$
- Note: convergence in finite number of iterations (not always)



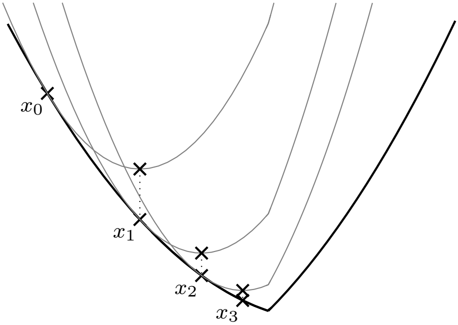
Proximal gradient – Example

- Proximal gradient iterations for problem $\underset{x}{\text{minimize}}\, \frac{1}{2}(x-a)^2 + |x|$
- $f(x) = \frac{1}{2}(x-a)^2$ is smooth term and $g(x) = |x|$ is nonsmooth
- Iteration: $x_{k+1} = \text{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k))$
- Note: convergence in finite number of iterations (not always)



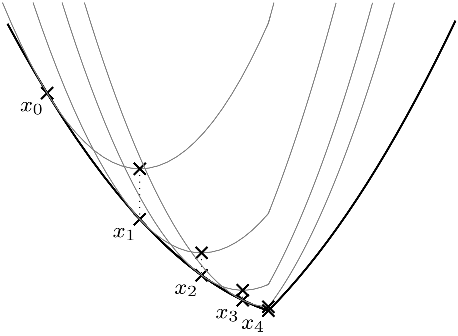
Proximal gradient – Example

- Proximal gradient iterations for problem $\underset{x}{\text{minimize}}\, \frac{1}{2}(x-a)^2 + |x|$
- $f(x) = \frac{1}{2}(x-a)^2$ is smooth term and $g(x) = |x|$ is nonsmooth
- Iteration: $x_{k+1} = \text{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k))$
- Note: convergence in finite number of iterations (not always)



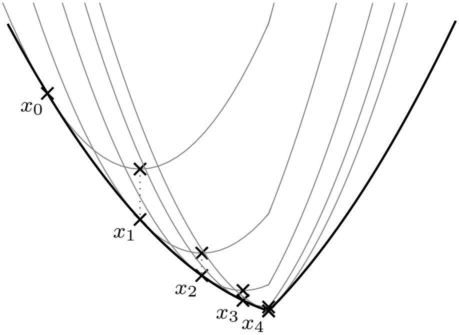
Proximal gradient – Example

- Proximal gradient iterations for problem $\underset{x}{\text{minimize}}\, \frac{1}{2}(x-a)^2 + |x|$
- $f(x) = \frac{1}{2}(x-a)^2$ is smooth term and $g(x) = |x|$ is nonsmooth
- Iteration: $x_{k+1} = \text{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k))$
- Note: convergence in finite number of iterations (not always)



Proximal gradient – Example

- Proximal gradient iterations for problem $\underset{x}{\text{minimize}}\, \frac{1}{2}(x-a)^2 + |x|$
- $f(x) = \frac{1}{2}(x-a)^2$ is smooth term and $g(x) = |x|$ is nonsmooth
- Iteration: $x_{k+1} = \text{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k))$
- Note: convergence in finite number of iterations (not always)



Proximal gradient – Special cases

- Proximal gradient method:
  - solves $\underset{x}{\text{minimize}}(f(x) + g(x))$
  - iteration: $x_{k+1} = \text{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))$
- Proximal gradient method with $g = 0$:
  - solves $\underset{x}{\text{minimize}}(f(x))$
  - $\text{prox}_{\gamma_k g}(z) = \text{argmin}_x(0 + \frac{1}{2\gamma}\|x - z\|_2^2) = z$
  - iteration: $x_{k+1} = \text{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k)) = x_k - \gamma_k \nabla f(x_k)$
  - reduces to gradient method
- Proximal gradient method with $f = 0$:
  - solves $\underset{x}{\text{minimize}}(g(x))$
  - $\nabla f(x) = 0$
  - iteration: $x_{k+1} = \text{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k)) = \text{prox}_{\gamma_k g}(x_k)$
  - reduces to *proximal point method* (which is not very useful)

Outline

- Introducing proximal gradient method and examples
- **Solving composite problem – Fixed-points and convergence**
- Application to primal and dual problems

Proximal gradient method – Fixed-point set

- Proximal gradient step

$$x_{k+1} = \text{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))$$

- If $x_{k+1} = x_k$, they are in *proximal gradient fixed-point set*

$$\{x : x = \text{prox}_{\gamma g}(x - \gamma \nabla f(x))\}$$

- Under some assumptions, algorithm will satisfy $x_{k+1} - x_k \to 0$
  - this means that fixed-point equation will be satisfied in limit
  - what does it mean for x to be a fixed-point?

Proximal gradient – Optimality condition

- Proximal gradient step:

$$v = \mathrm{prox}_{\gamma g}(x - \gamma\nabla f(x)) = \underset{y}{\mathrm{argmin}}(g(y) + \underbrace{\tfrac{1}{2\gamma}\|y - (x - \gamma\nabla f(x))\|_2^2}_{h(y)})$$

  where v is unique due to strong convexity of h

- Fermat's rule (since CQ holds) gives $v = \mathrm{prox}_{\gamma g}(x - \gamma\nabla f(x))$ iff:

$$\begin{aligned} 0 &\in \partial g(v) + \partial h(v) \\ &= \partial g(v) + \gamma^{-1}(v - (x - \gamma\nabla f(x))) \\ &= \partial g(v) + \nabla f(x) + \gamma^{-1}(v - x) \end{aligned}$$

  since h differentiable

Proximal gradient – Fixed-point characterization

For $\gamma > 0$, we have that

$$\bar{x} = \operatorname{prox}_{\gamma g}(\bar{x} - \gamma \nabla f(\bar{x})) \quad \text{if and only if} \quad 0 \in \partial g(\bar{x}) + \nabla f(\bar{x})$$

- Proof: the proximal step equivalence

$$v = \operatorname{prox}_{\gamma g}(x - \gamma \nabla f(x)) \quad \Leftrightarrow \quad 0 \in \partial g(v) + \nabla f(x) + \gamma^{-1}(v - x)$$

evaluated at a fixed-point $x = v = \bar{x}$ reads

$$\bar{x} = \operatorname{prox}_{\gamma g}(\bar{x} - \gamma \nabla f(\bar{x})) \quad \Leftrightarrow \quad 0 \in \partial g(\bar{x}) + \nabla f(\bar{x})$$

- We call inclusion $0 \in \partial g(\bar{x}) + \nabla f(\bar{x})$ *fixed-point characterization*

Meaning of fixed-point characterization

- What does fixed-point characterization $0 \in \partial g(\bar{x}) + \nabla f(\bar{x})$ mean?
- For convex differentiable f, subdifferential $\partial f(x) = \{\nabla f(x)\}$ and

$$0 \in \partial f(\bar{x}) + \partial g(\bar{x}) = \partial(f+g)(\bar{x})$$

  (subdifferential sum rule holds), i.e., fixed-points solve problem
- For nonconvex differentiable f, we might have $\partial f(\bar{x}) = \emptyset$
  - Fixed-point are not in general global solutions
  - Points $\bar{x}$ that satisfy $0 \in \partial g(\bar{x}) + \nabla f(\bar{x})$ are called *critical points*
  - If $g = 0$, the condition is $\nabla f(\bar{x}) = 0$, i.e., a *stationary point*
- Quality of fixed-points differs between convex and nonconvex f

Conditions on γ_k for convergence

- We replace in proximal gradient method $f(y)$ by

$$f(x_k) + \nabla f(x_k)^T(y - x_k) + \tfrac{1}{2\gamma_k}\|y - x_k\|_2^2$$

and minimize this plus $g(y)$ over y to get the next iterate

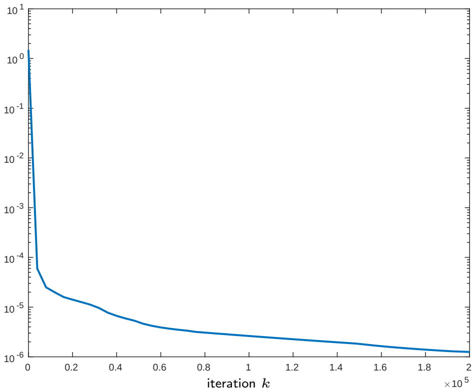
- We know from β-smoothness of f that for all x, y

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \tfrac{\beta}{2}\|y - x\|_2^2$$

- If $\gamma_k \in [\epsilon, \frac{1}{\beta}]$ with $\epsilon > 0$, an upper bound is minimized
- Can use $\gamma_k \in [\epsilon, \frac{2}{\beta} - \epsilon]$ and show convergence of some quantity

Practical convergence – Example

- Logarithmic y axis of quantity that should go to 0 for convergence
- Linear x axis with iteration number



- Fast convergence to medium accuracy, slow from medium to high
- Many iterations may be required

Stopping conditions

- For β-smooth $f : \mathbb{R}^n \to \mathbb{R}$, we can stop algorithm when

$$\tfrac{1}{\beta}u_k := \tfrac{1}{\beta}(\gamma_k^{-1}(x_k - x_{k+1}) + \nabla f(x_{k+1}) - \nabla f(x_k))$$

  is small (notation and reason will be motivated in future lecture)
- This is the plotted quantity on the previous slide
- We can use absolute or relative stopping conditions:
  - absolute stopping conditions with small $\epsilon_{\rm abs} > 0$

$$\tfrac{1}{\beta}\|u_k\|_2 \leq \epsilon_{\rm abs} \qquad \text{or} \qquad \tfrac{1}{\beta}\|u_k\|_2 \leq \epsilon_{\rm abs}\sqrt{n}$$

  - relative stopping condition with small $\epsilon_{\rm rel}, \epsilon > 0$:

$$\tfrac{1}{\beta}\tfrac{\|u_k\|_2}{\|x_k\|_2 + \beta^{-1}\|\nabla f(x_k)\|_2 + \epsilon} \leq \epsilon_{\rm rel}$$

- Problem considered solved to optimality if, say, $\frac{1}{\beta}\|u_k\|_2 \leq 10^{-6}$
- Often lower accuracy of 10^{-3} or 10^{-4} is enough

Outline

- Introducing proximal gradient method and examples
- Solving composite problem – Fixed-points and convergence
- **Application to primal and dual problems**

Applying proximal gradient to primal problems

Problem $\underset{x}{\text{minimize}}\, f(x) + g(x)$:

- Assumptions:
  - f smooth
  - g closed convex and prox friendly[1]
- Algorithm: $x_{k+1} = \text{prox}_{\gamma_k g}(x_k - \gamma_k \nabla f(x_k))$

Problem $\underset{x}{\text{minimize}}\, f(Lx) + g(x)$:

- Assumptions:
  - f smooth (implies $f \circ L$ smooth)
  - g closed convex and prox friendly[1]
- Gradient $\nabla(f \circ L)(x) = L^T \nabla f(Lx)$
- Algorithm: $x_{k+1} = \text{prox}_{\gamma_k g}(x_k - \gamma_k L^T \nabla f(Lx_k))$

---

[1] Prox friendly: proximal operator cheap to evaluate, e.g., g separable

Applying proximal gradient to dual problem

- Let us apply the proximal gradient method to the dual problem

$$\underset{\mu}{\text{minimize}}\, f^*(\mu) + g^*(-L^T\mu)$$

- Assumptions:
  - f: closed convex and prox friendly
  - g: σ-strongly convex
- Why these assumptions?
  - f^*: closed convex and prox friendly
  - $g^* \circ -L^T$: $\frac{\|L\|_2^2}{\sigma}$-smooth and convex
- Algorithm:

$$\mu_{k+1} = \text{prox}_{\gamma_k f^*}(\mu_k - \gamma_k \nabla(g^* \circ -L^T)(\mu_k))$$

Dual proximal gradient method – Explicit version 1

- We will make the dual proximal gradient method more explicit

$$\mu_{k+1} = \mathrm{prox}_{\gamma_k f^*}(\mu_k - \gamma_k \nabla(g^* \circ -L^T)(\mu_k))$$

- Use $\nabla(g^* \circ -L^T)(\mu) = -L\nabla g^*(-L^T\mu)$ to get

$$\begin{aligned} x_k &= \nabla g^*(-L^T\mu_k) \\ \mu_{k+1} &= \mathrm{prox}_{\gamma_k f^*}(\mu_k + \gamma_k L x_k) \end{aligned}$$

Dual proximal gradient method – Explicit version 2

- Restating the previous formulation

$$x_k = \nabla g^*(-L^T\mu_k)$$
$$\mu_{k+1} = \mathrm{prox}_{\gamma_k f^*}(\mu_k + \gamma_k L x_k)$$

- Use Moreau decomposition for prox:

$$\mathrm{prox}_{\gamma f^*}(v) = v - \gamma \mathrm{prox}_{\gamma^{-1} f}(\gamma^{-1} v)$$

to get

$$x_k = \nabla g^*(-L^T\mu_k)$$
$$v_k = \mu_k + \gamma_k L x_k$$
$$\mu_{k+1} = v_k - \gamma_k \mathrm{prox}_{\gamma_k^{-1} f}(\gamma_k^{-1} v_k)$$

Dual proximal gradient method – Explicit version 3

- Restating the previous formulation

$$
\begin{aligned}
x_k &= \nabla g^*(-L^T\mu_k)\\
v_k &= \mu_k + \gamma_k L x_k\\
\mu_{k+1} &= v_k - \gamma_k \mathrm{prox}_{\gamma_k^{-1}f}(\gamma_k^{-1}v_k)
\end{aligned}
$$

- Use subdifferential formula, since g^* differentiable:

$$
\nabla g^*(\nu) = \operatorname*{argmax}_x(\nu^T x - g(x)) = \operatorname*{argmin}_x(g(x) - \nu^T x)
$$

with $\nu = -L^T\mu_k$ to get

$$
\begin{aligned}
x_k &= \operatorname*{argmin}_x(g(x) + (\mu_k)^T L x)\\
v_k &= \mu_k + \gamma_k L x_k\\
\mu_{k+1} &= v_k - \gamma_k \mathrm{prox}_{\gamma_k^{-1}f}(\gamma_k^{-1}v_k)
\end{aligned}
$$

- Can implement method without computing conjugate functions

Dual proximal gradient method – Primal recovery

- Can we recover a primal solution from dual prox grad method?
- Let us use explicit version 1

$$x_k = \nabla g^*(-L^T\mu_k)$$
$$\mu_{k+1} = \mathrm{prox}_{\gamma_k f^*}(\mu_k + \gamma_k L x_k)$$

and assume we have found fixed-point $(\bar{x}, \bar{\mu})$: for some $\bar{\gamma} > 0$,

$$\bar{x} = \nabla g^*(-L^T\bar{\mu})$$
$$\bar{\mu} = \mathrm{prox}_{\bar{\gamma} f^*}(\bar{\mu} + \bar{\gamma} L\bar{x})$$

- Fermat's rule for proximal step

$$0 \in \partial f^*(\bar{\mu}) + \bar{\gamma}^{-1}(\bar{\mu} - (\bar{\mu} + \bar{\gamma}L\bar{x})) = \partial f^*(\bar{\mu}) - L\bar{x}$$

is with $\bar{x} = \nabla g^*(-L^T\bar{\mu})$ a primal-dual optimality condition
- So x_k will solve primal problem if algorithm converges

Problems that prox-grad cannot solve

- Problem $\underset{x}{\text{minimize}}\, f(x) + g(x)$
- Assumptions: f and g convex but nondifferentiable
- No term differentiable, another method must be used:
  - Subgradient method
  - Douglas-Rachford splitting
  - Primal-dual methods

Problems that prox-grad cannot solve efficiently

- Problem $\underset{x}{\text{minimize}}\, f(x) + g(Lx)$
- Assumptions:
  - f smooth
  - g nonsmooth convex
  - L arbitrary structured matrix
- Can apply proximal gradient method

$$x_{k+1} = \underset{y}{\text{argmin}}(g(Ly) + \tfrac{1}{2\gamma_k}\|y - (x_k - \gamma_k \nabla f(x_k))\|_2^2)$$

but proximal operator of $g \circ L$

$$\text{prox}_{\gamma(g \circ L)}(z) = \underset{x}{\text{argmin}}(g(Lx) + \tfrac{1}{2\gamma}\|x - z\|_2^2)$$

often not "prox friendly", i.e., it is expensive to evaluate