Proximal Algorithms

Idealistic Process

Optimization Problem:

$$\min_{x \in \mathbb{R}^d} F(x)$$

where $F: \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is closed. [We'll assume more later]

Proximal Point Method (PPA):

$$x_{t+1} = \arg\min_{x \in \mathbb{R}^d} \left\{ F(x) + \frac{1}{2\alpha} \|x - x_t\|^2 \right\}$$

(Moreau '63, Rockafellar '76, Martinet '70)

Goal of the lecture

- proximal (stochastic sub)gradient
- (stochastic) prox-point
- (stochastic) Gauss-Newton

are all approximate PPA

Convex Setting

(assume F is convex)

Three viewpoints

1) Proximal Map:

$$\text{prox}_{\alpha F}(x) := \underset{z}{\text{argmin}} \left\{ F(z) + \frac{1}{2\alpha}\|z - x\|^2 \right\}$$

$\Rightarrow$ PPA is the fixed point iteration:

$$\boxed{x_{t+1} = \text{prox}_{\alpha F}(x_t)}$$

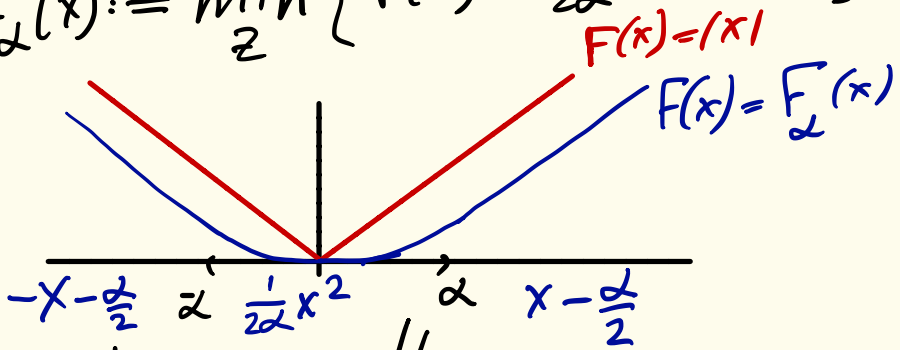
2) Discretization of $\dot{x}(t) \in -\partial F(x(t))$

- PPA: $\alpha^{-1}(x_t - x_{t+1}) \in \partial F(x_{t+1})$ [Implicit]
- Subgradient Method: $\alpha^{-1}(x_t - x_{t+1}) \in \partial F(x_t)$ [Explicit]

Three viewpoints

3) Moreau Envelope:

$$F_\alpha(x) := \min_z \left\{ F(z) + \frac{1}{2\alpha}\|z-x\|^2 \right\}$$



Thm: F_α is C^1-smooth with

$$\nabla F_\alpha(x) = \alpha^{-1}\left[x - \text{prox}_{\alpha F}(x)\right]$$

$$\Rightarrow x_{t+1} = x_t - \alpha\left[\alpha^{-1}(x_t - x_{t+1})\right] = x_t - \alpha \nabla F_\alpha(x_t)$$

$\therefore$ PPA is gradient descent on F_α!

Iteration Complexity

	Convex	μ-Strongly convex
PPA	$\frac{\alpha^{-1}}{\epsilon}$ ①	$\frac{\alpha^{-1}}{\mu} \cdot \ln\left(\frac{1}{\epsilon}\right)$ ②
Accelerated PPA	$\sqrt{\frac{\alpha^{-1}}{\epsilon}}$	$\sqrt{\frac{\alpha^{-1}}{\mu}} \cdot \ln\left(\frac{1}{\epsilon}\right)$

- # iterations to get $F(x_t) - F^* \leq \epsilon$

(Güler '92, Nesterov '83)

pf: ① $F(x_{t+1}) + \frac{1}{2\alpha}\|x_{t+1} - x_t\|^2 + \frac{1}{2\alpha}\|x^* - x_{t+1}\|^2 \le F(x^*) + \frac{1}{2\alpha}\|x^* - x_t\|^2$

$$\Rightarrow F(x_{t+1}) - F(x^*) \le \frac{1}{2\alpha}\left[\|x^* - x_t\|^2 - \|x^* - x_{t+1}\|^2\right]$$

$$\therefore F(x_{T+1}) - F^* \le \frac{1}{T}\sum_{t=0}^{T}\left[F(x_{t+1}) - F^*\right] \le \frac{\|x^* - x_0\|^2}{2\alpha T}$$ ∎

② $\text{prox}_{\alpha F}(\cdot)$ is $\frac{1}{1+\mu\alpha}$ - contractive:

$$\|x_{t+1} - x^*\| = \|\text{prox}_{\alpha F}(x_t) - \text{prox}_{\alpha F}(x^*)\| \le \frac{1}{1+\mu\alpha}\|x_t - x^*\|$$ ∎

Accelerated Proximal Gradient

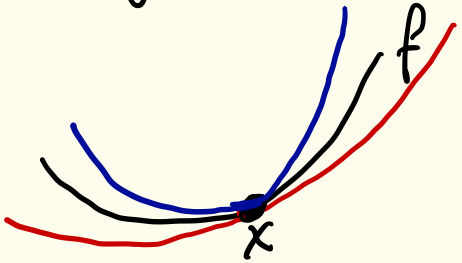
(Beck-Teboulle '08, Nesterov '10)

Smooth convex + convex

Problem Class: $\min_{x \in \mathbb{R}^d} F(x) = f(x) + r(x)$ where

- f is μ-strongly convex and L-smooth:

$$\left\{ \begin{aligned} f(y) &\le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \\ f(y) &\ge f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2 \end{aligned} \right\} \forall x, y$$



$$\text{condition \#} = \frac{L}{\mu}$$

- r is convex

Examples

- **LASSO:** $\min_x \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1$
- **$\|\cdot\|_1$-logistic regression** on $\{(w_i, b_i)\}_{i=1}^m \subset \mathbb{R}^d \times \{-1, 1\}$

$$\min_x \frac{1}{m}\sum_{i=1}^m \log\left(1 + e^{-b_i\langle x, w_i\rangle}\right) + \lambda\|x\|_1$$

- **Group Sparsity:**

$$\min_x \frac{1}{2}\|Ax-b\|_2^2 + \lambda\sum_{i=1}^k \|x_i\|_2$$

- **Low-rank matrix completion/sensing:**

$$\min_{X \in \mathbb{R}^{m\times n}} \frac{1}{2}\|\mathcal{A}(X)-b\|_2^2 + \lambda\|X\|_*$$

Proximal Gradient

PPA: $x_{t+1} = \arg\min_x \left\{ f(x) + r(x) + \frac{L}{2}\|x - x_t\|^2 \right\}$

Prox-gradient:

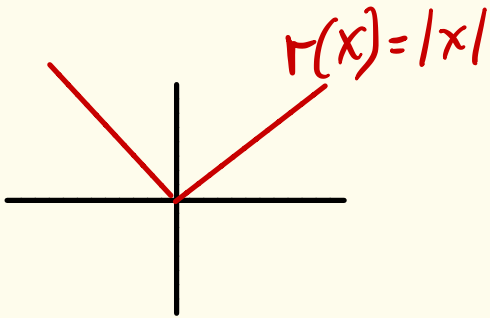
$$x_{t+1} = \arg\min_x \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + r(x) + \frac{L}{2}\|x - x_t\|^2 \right\}$$

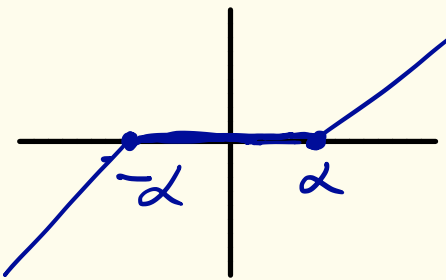
$$= \text{prox}_{r/\beta}\left(x_t - \frac{1}{L}\nabla f(x_t)\right)$$

Accelerated prox-gradient:

$$\begin{cases} x_{t+1} = \text{prox}_{r/\beta}\left(y_t - \frac{1}{L}\nabla f(y_t)\right) \\ y_{t+1} = x_{t+1} + \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\left(x_{t+1} - x_t\right) \end{cases}$$

Exercise: $r(x) = \|x\|_1, \|x\|_\infty, \delta_{B_1}, \delta_{B_\infty}, \delta_{\mathbb{R}^d_+}, \delta_{\text{unit simplex}}$ have easy $\text{prox}_{\alpha r}(\cdot)$





$\text{prox}_{\alpha|\cdot|}(\cdot)$ is soft thresholding

Harder Example: $\|X\|_{op} = \max_i \{|\sigma_i(X)|\}$, $\|X\|_* = \|\sigma(X)\|_1$

Thm: Suppose $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is signed-permutation invariant. Define $\hat{f} : \mathbb{R}^{d \times n} \to \mathbb{R} \cup \{\infty\}$ by

$$\hat{f}(X) = f(\sigma(X))$$

($\sigma(X)$: singular values)

Then

$$\text{prox}_{\alpha \hat{f}}(X) = U \left[\text{Diag}\left[\text{prox}_{\alpha f}(\sigma(X)) \right] \right] V^T$$

for any $U \in \mathbb{O}^d$, $V \in \mathbb{O}^m$ with

$$X = U [\text{Diag}\ \sigma(X)] V^T$$

	convex	μ-strongly convex
Grad descent	$\frac{L}{\varepsilon}$ ①	$\frac{L}{\mu}\cdot\ln\left(\frac{1}{\varepsilon}\right)$
Accelerated Grad descent	$\sqrt{\frac{L}{\varepsilon}}$	$\sqrt{\frac{L}{\mu}}\cdot\ln\left(\frac{1}{\varepsilon}\right)$

pf: ① Set $f_x(y) = f(x) + \langle \nabla f(x), y - x\rangle$. Then

$$F(x_{t+1}) \le f_{x_t}(x_{t+1}) + r(x_{t+1}) + \frac{L}{2}\|x_{t+1} - x_t\|^2$$

$$\le f_{x_t}(x^*) + r(x^*) + \frac{L}{2}\|x^* - x_t\|^2 - \frac{L}{2}\|x^* - x_{t+1}\|^2 \le F^* + \frac{L}{2}\left[\|x^* - x_t\|^2 - \|x^* - x_{t+1}\|^2\right]$$

$$\Rightarrow F(x_{T+1}) - F^* \le \frac{1}{T}\sum_{t=0}^{T} F(x_{t+1}) - F^* \le \frac{L\|x^* - x_0\|^2}{2T}$$

∎

Nonconvex and Nonsmooth optimization

Weakly Convex Functions

Recall: PPA

$$x_{t+1} = \operatorname{argmin}_x \left\{ F(x) + \frac{1}{2\alpha}\|x - x_t\|^2 \right\}$$

convex?

Defn: F is ρ-weakly convex if

$$x \mapsto F(x) + \frac{\rho}{2}\|x\|^2 \text{ is convex}$$

Fact: The following are equivalent:

1) F is ρ-weakly convex
2) $F(\lambda x + (1-\lambda)y) \le \lambda F(x) + (1-\lambda)F(y) + \frac{\rho\lambda(1-\lambda)}{2}\|x-y\|^2$ $\quad \forall x, y \in \mathbb{R}^d$, $\lambda \in [0,1]$
3) $\forall x \; \exists v_x \in \mathbb{R}^d$ s.t.
$f(y) \ge f(x) + \langle v_x, y - x\rangle - \frac{\rho}{2}\|x-y\|^2$ $\quad \forall y$

Examples

- If $\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\| \quad \forall x,y$
  then f is L-weakly convex.

- C^2-smooth f is ρ-weakly convex $\iff \nabla^2 f(x) \succeq -\rho I \quad \forall x$

- Suppose $f(x) = h(c(x))$
  (h: convex, ℓ-Lipschitz; c: C^1-smooth with $\|\nabla c(x) - \nabla c(y)\|_{op} \le L\|x-y\| \quad \forall x,y$)
  Then f is ℓL-weakly convex.

pf: $f(y) = h(c(y)) \geq h\left(c(x) + \nabla c(x)(y-x)\right) - \frac{\ell L}{2}\|x-y\|^2$

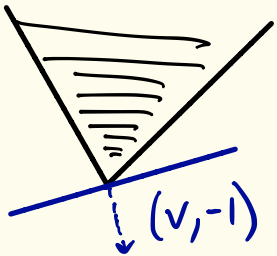
$\geq h(c(x)) + \langle v, \nabla c(x)(y-x)\rangle - \frac{\ell L}{2}\|x-y\|^2$

$= f(x) + \langle \nabla c(x)^T v, y-x\rangle - \frac{\ell L}{2}\|x-y\|^2$

for any $v \in \partial h(c(x))$ ☐

Reminder:

$v \in \partial h(z) \Leftrightarrow h(\tilde{z}) \geq h(z) + \langle v, \tilde{z} - z\rangle \quad \forall \tilde{z}$



$$\partial |\cdot|(x) = \begin{cases} +1, & x > 0 \\ [-1,1], & x = 0 \\ -1, & x < 0 \end{cases}$$

Examples

- Sparse Dictionary Learning

$$\min_{D \in \mathbb{R}^{d \times n},\, r_i \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} h_i(x_i - D r_i) + \lambda \|r_i\|_1 \quad \text{s.t.} \quad \begin{array}{l} D \geq 0 \\ r_i \geq 0 \ \forall i \\ \|D_j\|_2 \leq 1 \ \forall j \end{array}$$

- Sparse Phase Retrieval

$$\min_x \frac{1}{m} \sum_{i=1}^{m} h_i(\langle a_i, x \rangle^2 - b_i) + \lambda \|x\|_1$$

- Covariance Matrix Estimation

$$\min_{\Sigma} \frac{1}{m} \sum_{i=1}^{m} h_i(\|\Sigma a_i\|^2 - b_i)$$

- Robust PCA

$$\min_{U,V} \|UV - M\|_1$$

Stochastic Algos for weakly cvx problems

Problem Class: [Eg. $f(x) = \frac{1}{m}\sum_{i=1}^{m} h_i(c_i(x))$]

$$\min_{x \in \mathbb{R}^d} F(x) = f(x) + r(x) \quad \text{where}$$

where $f(\cdot)$ is weakly convex and only stochastically available

Main Example: $f(x) = \frac{1}{m}\sum_{i=1}^{m} h_i(c_i(x))$

Stochastic one-sided model

Assumption:

- Access to f is through a stochastic model

$$(x, y, \omega) \mapsto f_x(y, \omega)$$

satisfying

1) $\mathbb{E}_\omega[f_x(x, \omega)] = f(x)$ and $\mathbb{E}_\omega[f_x(y, \omega) - f(y)] \le \frac{\gamma}{2}\|y - x\|^2$
2) $f_x(\cdot, \omega) + r(\cdot)$ is ρ-weakly convex
3) $f_x(\cdot, \omega)$ is γ-Lipschitz.

Algorithm:

$$\begin{cases} \text{Sample } \omega_t \sim P \\ x_{t+1} = \operatorname{argmin}_x \left\{ f_{x_t}(x, \omega_t) + r(x) + \frac{1}{2\alpha_t}\|x - x_t\|^2 \right\} \end{cases}$$

Examples

Problem: $\min_x F(x) = f(x) + r(x)$

where $f(x) = \mathbb{E}_\omega[h(c(x,\omega),\omega)]$

[eg. $f(x) = \frac{1}{m}\sum_{i=1}^{m} h_i(c_i(x))$]

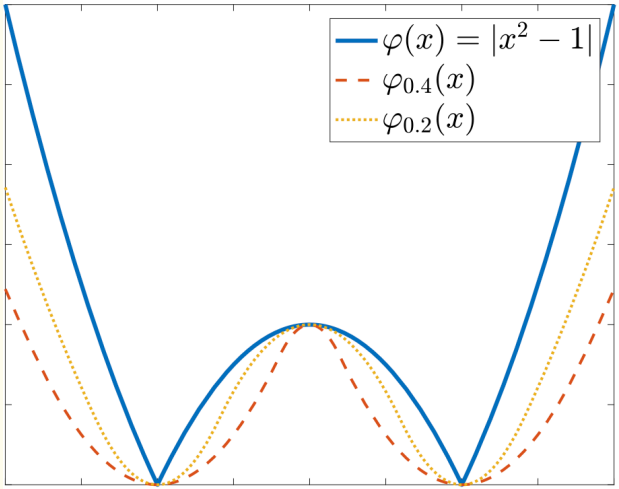
Stochastic PPA: $f_x(y,\omega) = h(c(x,\omega),\omega)$

Stochastic Gauss-Newton: $f_x(y,\omega) = h(c(x,\omega) + \nabla c(x,\omega)(y-x),\omega)$

Stochastic Subgradient: $f_x(y,\omega) = h(c(x,\omega),\omega) + \langle \nabla c(x,\omega)^T v, y-x\rangle$
for any $v \in \partial h(c(x,\omega),\omega)$

Convergence Rate ???

Moreau Envelope: $F_\alpha(x) := \min_z \left\{ F(z) + \frac{1}{2\alpha}\|z-x\|^2 \right\}$



Thm: F is ρ-weakly convex and $\alpha < \rho^{-1}$

$\Longrightarrow$ $\nabla F_\alpha(x) = \alpha^{-1}\left[x - \text{prox}_{\alpha F}(x)\right]$

Thm: With appropriate choice of α_t, get

$$\mathbb{E}\left[\|\nabla F_{1/2\rho}(x_{t^*})\|^2\right] \leq O\left(\frac{1}{T^{1/4}}\right)$$

where t^* is chosen uniformly from $\{0, 1, \ldots, T\}$

Key Estimate:

$$\mathbb{E}\left[F_{1/2\rho}(x_{t+1})\right] \leq \mathbb{E}\left[F_{1/2\rho}(x_t)\right] - c_1 \alpha_t \mathbb{E}\left[\|\nabla F_{1/2\rho}(x_t)\|^2\right] + c_2 \alpha_t^2$$

∴ Stochastic PPA, subgradient, and Gauss-Newton are approximate gradient methods on $F_{1/2\rho}$!

(Davis-D '18)

Further reading and references:

- Awesome survey:
  "A simplified view of first order methods for optimization", Marc Teboulle, Math. Prog. Ser. B 170(1): 67-96, 2018
- Short News Article:
  "The proximal point method revisited", SIAG/OPT Views and News, 26(1): 1-7, 2018