

Proximal Algorithms

Lecturer: Dmitry Drusvyatskiy

Idealistic Process

Optimization Problem:

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

where $F: \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$ 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:

$$\text{prox}_{\lambda F}(x) := \arg \min_z \left\{ F(z) + \frac{1}{2\lambda} \|z - x\|^2 \right\}$$

\Rightarrow PPA is the fixed point iteration:

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

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

• PPA : $\lambda^{-1}(x_t - x_{t+1}) \in \partial F(x_{t+1})$ [Implicit]

• Subgradient Method :

$$\lambda^{-1}(x_t - x_{t+1}) \in \partial F(x_t)$$

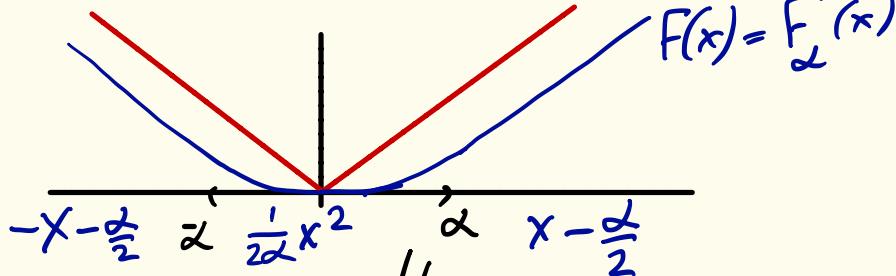
[Explicit]

Three viewpoints

3) Moreau Envelope:

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

$F(x) = \langle x \rangle$



Thm: F_α is C' -smooth with

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

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

$\therefore \text{PPA}$ 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)

$$\begin{aligned}
 & \text{Pf: } \textcircled{1} \quad F(x_{t+1}) + \frac{1}{2\alpha} \|x_{t+1} - x_t\|^2 + \frac{1}{2\alpha} \|x^* - x_{t+1}\|^2 \leq F(x^*) + \frac{1}{2\alpha} \|x^* - x_t\|^2 \\
 \Rightarrow & \quad F(x_{t+1}) - F(x^*) \leq \frac{1}{2\alpha} [\|x^* - x_t\|^2 - \|x^* - x_{t+1}\|^2] \\
 \therefore & \quad F(x_{T+1}) - F^* \leq \frac{1}{T} \sum_{t=0}^T [F(x_{t+1}) - F^*] \leq \frac{\|x^* - x_0\|^2}{2\alpha T} \quad \square
 \end{aligned}$$

$$\begin{aligned}
 & \textcircled{2} \quad \text{prox}_{\alpha F}(\cdot) \text{ is } \frac{1}{1+\alpha\lambda} - \text{contractive:} \\
 & \|x_{t+1} - x^*\| = \|\text{prox}_{\alpha F}(x_t) - \text{prox}_{\alpha F}(x^*)\| \leq \frac{1}{1+\alpha\lambda} \|x_t - x^*\| \quad \square
 \end{aligned}$$

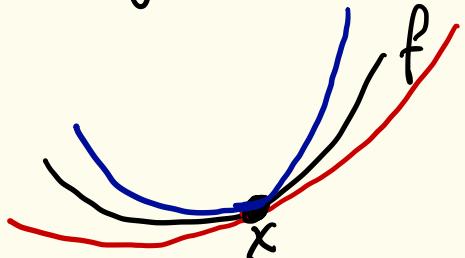
Accelerated Proximal Gradient
(Beck-Teboulle '08, Nesterov '10)

Smooth convex + convex

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

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

$$\left\{ \begin{array}{l} f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 \\ f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2} \|y - x\|^2 \end{array} \right\} \quad t_{x,y}$$



condition $\# = \frac{L}{M}$

- Γ is convex

Examples

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

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

$$\min_x \frac{1}{2} \|Ax - b\|_2^2 + \lambda \sum_{i=1}^k \|x_i\|_2$$
- Low-rank matrix completion/sensing:

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

Proximal Gradient

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

Prox-gradient:

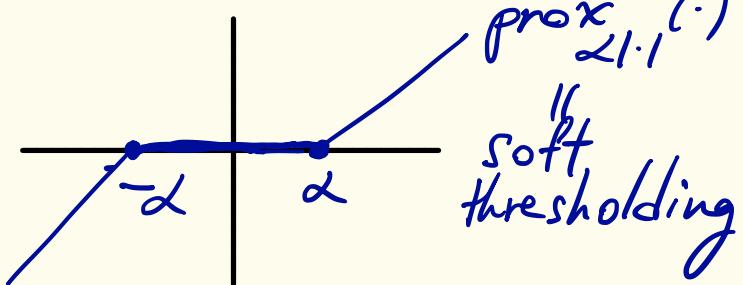
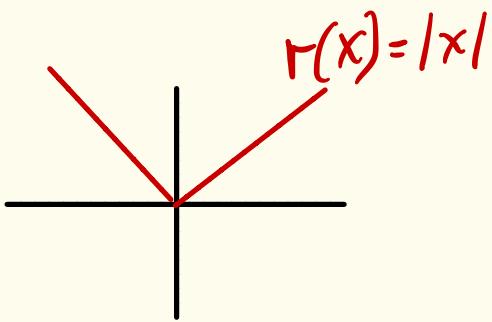
$$x_{t+1} = \arg \min_x \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + r(x) + \frac{L}{2} \|x - x_t\|^2 \right\}$$
$$= \text{prox}_{r/L} \left(x_t - \frac{1}{L} \nabla f(x_t) \right)$$

Accelerated prox-gradient:

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

Exercise:

$r(x) = \|x\|_1, \|x\|_\infty, \delta_B, \delta_{B_\infty}, \delta_{R^d}, \delta$
have easy $\text{prox}_{\lambda r}(\cdot)$ unit simplex



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

Thm: Suppose $\Gamma: \mathbb{R}^d \rightarrow \mathbb{R}V\{\infty\}$ is signed-permutation invariant. Define $\hat{\Gamma}: \mathbb{R}^{d+n} \rightarrow \mathbb{R}V\{\infty\}$ by $\hat{\Gamma}(\underline{X}) = \Gamma(\underline{\sigma}(\underline{X}))$ singular values

Then

$$\text{prox}_{2\hat{\Gamma}}(\underline{X}) = U \left[\text{Diag}[\text{prox}_{2\Gamma}(\sigma(\underline{X}))] \right] V^T$$

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

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

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

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

$$\begin{aligned}
 F(x_{t+1}) &\leq f_{x_t}(x_{t+1}) + \Gamma(x_{t+1}) + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\
 &\leq f_{x_t}(x^*) + \Gamma(x^*) + \frac{L}{2} \|x^* - x_t\|^2 - \frac{L}{2} \|x^* - x_{t+1}\|^2 \leq F^* + \frac{L}{2} \left[\|x^* - x_t\|^2 - \|x^* - x_{t+1}\|^2 \right] \\
 \implies F(x_{T+1}) - F^* &\leq \frac{1}{T} \sum_{t=0}^T F(x_{t+1}) - F^* \leq \frac{L \|x^* - x_0\|^2}{2T}
 \end{aligned}$$

Nonconvex and Nonsmooth
optimization

Weakly Convex Functions

Recall //: PPA

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

convex?

Defn: F is β -weakly convex if

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

Fact: The following are equivalent:

- 1) F is β -weakly convex
- 2) $F(\lambda x + (1-\lambda)y) \leq \lambda F(x) + (1-\lambda)F(y) + \frac{\beta\lambda(1-\lambda)}{2} \|x-y\|^2$
- 3) $\forall x \exists v_x \in \mathbb{R}^d$ s.t.
 $f(y) \geq f(x) + \langle v_x, y-x \rangle - \frac{\beta}{2} \|x-y\|^2$

$x, y \in \mathbb{R}^d$
 $\lambda \in [0, 1]$

v_y

Examples

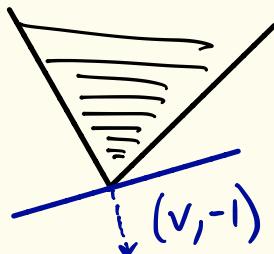
- If $\|Df(x) - Df(y)\| \leq L \|x-y\| \quad \forall x, y$
then f is L -weakly convex.
- C^2 -smooth f is ρ -weakly convex $\Leftrightarrow D^2f(x) \succcurlyeq \rho I \quad \forall x$
- Suppose $f(x) = h(c(x))$
 $\begin{array}{c} \text{convex} \\ \text{L-Lipschitz} \end{array} \quad \begin{array}{c} C^1 \text{-smooth with} \\ \|Dc(x) - Dc(y)\|_{op} \leq L \|x-y\| \quad \forall x, y \end{array}$
 Then f is L -weakly convex.

$$\begin{aligned}
 \text{pf: } f(y) &= h(c(y)) \geq h\left(c(x) + \nabla c(x)(y-x)\right) - \frac{\ell L}{2} \|x-y\|^2 \\
 &\geq h(c(x)) + \langle \nabla c(x)(y-x), y-x \rangle - \frac{\ell L}{2} \|x-y\|^2 \\
 &= f(x) + \langle \nabla c(x)^T V, y-x \rangle - \frac{\ell L}{2} \|x-y\|^2
 \end{aligned}$$

for any $v \in \partial h(c(x))$ \blacksquare

Reminder:

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



$$\partial I_1(x) = \begin{cases} +1 & x > 0 \\ [-1, 1] & x = 0 \\ -1 & x < 0 \end{cases}$$

Examples

- Sparse Dictionary Learning

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

- 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 crx problems

Problem Class:

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

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

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) $E_{\omega}[f_x(x, \omega)] = f(x)$ and $E_{\omega}[f_x(y, \omega) - f(y)] \leq \frac{\gamma}{2} \|y - x\|^2$
- 2) $f_x(\cdot, \omega) + r(\cdot)$ is β -weakly convex
- 3) $f_x(\cdot, \omega)$ is $\delta(\omega)$ -Lipschitz with $E_{\omega}[\delta(\omega)^2] < \infty$

Algorithm:

$$\left\{ \begin{array}{l} \text{Sample } \omega_t \sim P \\ x_{t+1} = \arg \min_x \left\{ f_{x_t}(x, \omega_t) + r(x) + \frac{1}{2\omega_t} \|x - x_t\|^2 \right\} \end{array} \right\}$$

Examples

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

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

$$\left[\text{e.g. } f(x) = \frac{1}{m} \sum_{i=1}^m h_i(c_i(x)) \right]$$

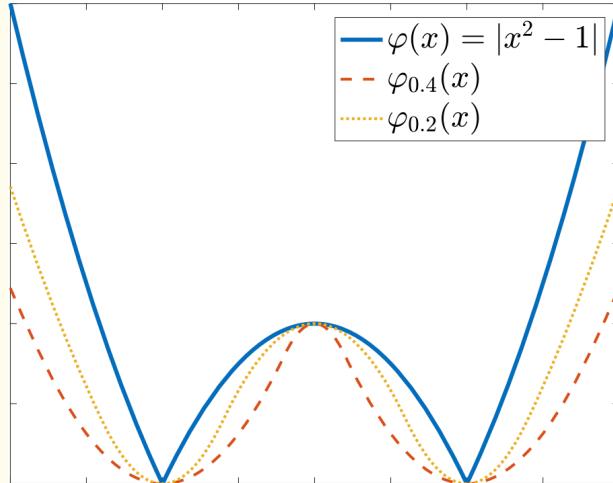
Stochastic PPA: $f_x(y, \omega) = h(c(x, \omega), \omega)$

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



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

\implies

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

Thm: With appropriate choice of α_t , get

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

Where t^* is chosen uniformly from $\{0, 1, \dots, T\}$

Key Estimate:

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

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