Optimization Methods in Machine Learning Lectures 13-14

Katya Scheinberg

Lehigh University

katyas@lehigh.edu

First Order Methods

First-order proximal gradient methods

- Consider: $$\min_x f(x)$$

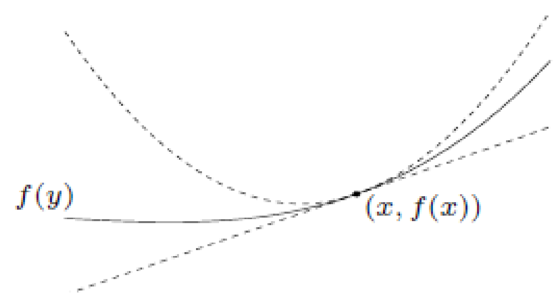
$$|\nabla f(x) - \nabla f(y)| \leq L||x - y||$$

- Linear lower approximation

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x)$$

- Quadratic upper approximation

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2\mu}||y - x||^2 = Q_{f,\mu}(x, y)$$



$$f(y) \leq f(x) + \frac{1}{2\mu}||x - \mu\nabla f(x)^\top - y||^2 = Q_{f,\mu}(x, y)$$

First-order proximal gradient method

$$\min_x f(x)$$

- Minimize quadratic upper approximation on each iteration

$$x^{k+1} = \mathrm{argmin}_y Q_{f,\mu}(x^k, y)$$

$$\Updownarrow$$

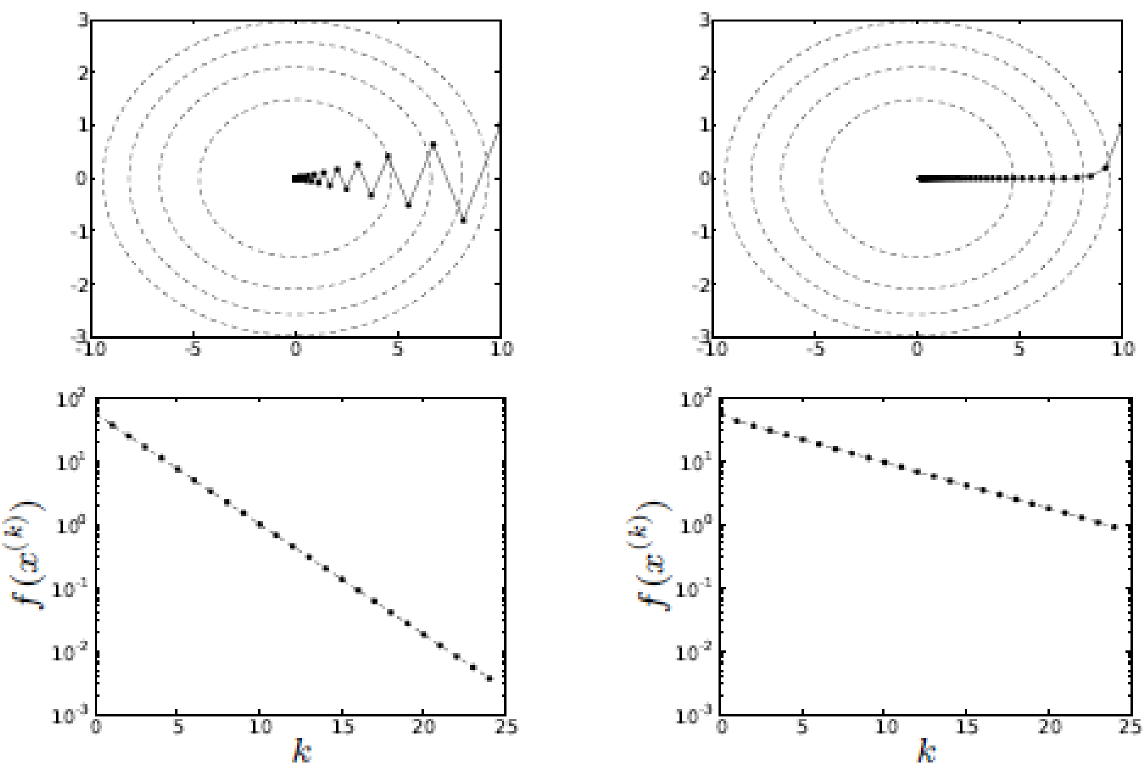
$$x^{k+1} = x^k - \mu \nabla f(x^k)$$

- If $\mu \leq$ *1/L* then

$$f(x^{k+1}) \leq f(x^k) + \frac{1}{2\mu} ||x^k - \mu \nabla f(x^k)^\top - x^{k+1}||^2 = Q_{f,\mu}(x^k, x^{k+1})$$

Quadratic example

$f(x_1, x_2) = (x_1^2 + Lx_2^2)/2$; left: $\mu = 1.8/L$; right: $\mu = 0.8/L$



Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

Complexity bound derivation outline

$$f(x^{k+1}) \leq f(x^k) + \frac{1}{2\mu}||x^k - \mu \nabla f(x^k)^\top - x^{k+1}||^2 = Q_{f,\mu}(x^k, x^{k+1})$$

⇓

$$f(x^{k+1}) - f(x^*) \leq \frac{1}{2\mu}(\|x^k - x^*\| - \|x^{k+1} - x^*\|)$$

⇓

$$f(x^k) - f(x^*) \leq \frac{L\|x^0 - x^*\|}{2k}$$

Complexity of proximal gradient method

- Minimize quadratic upper approximation on each iteration

$$x^{k+1} = \text{argmin}_y Q_{f,\mu}(x^k, y)$$

$$\Updownarrow$$

$$x^{k+1} = x^k - \mu \nabla f(x^k)$$

- If $\mu \leq$ *1/L* then in *O(L||x^0-x^*||/ϵ)* iterations finds solution

$$x^k : \; f(x^k) \leq f(x^*) + \epsilon$$

Compare to ***O(log(L/ϵ))*** of interior point methods.

Can we do better?

Accelerated first-order method

Nesterov, ’83, ‘00s,

Beck&Teboulle ‘09

$$\min_x f(x)$$

- Minimize upper approximation at an intermediate point.

$$x^{k+1} = y^k - \mu \nabla f(y^k)$$

$$y^{k+1} := \quad x^k + \tfrac{k-1}{k+2}[x^k - x^{k-1}]$$

- If $\mu \leq$ *1/L* then

$$f(x^k) - f(x^*) \leq \frac{L\|x^0 - x^*\|}{2k^2}$$

Complexity of accelerated first-order method

Nesterov, ’83, ‘00s,

Beck&Teboulle ‘09

$$\min_x f(x)$$

- Minimize upper approximation at an intermediate point.

$$x^{k+1} = y^k - \mu \nabla f(y^k)$$

$$y^{k+1} := \quad x^k + \tfrac{k-1}{k+2}[x^k - x^{k-1}]$$

- If $\mu \leq$ *1/L* then in $O(\sqrt{\frac{L\|x^0 - x^*\|}{\epsilon}})$ iterations finds solution

$$\bar{x} : \; f(\bar{x}) \leq f(x^*) + \epsilon$$

This method is optimal if only gradient information is used.

Optimality of Nesterov's method

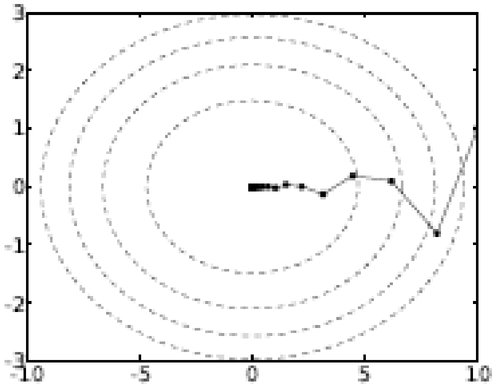
define a **first order method** as any iterative algorithm that selects $x^{(k)}$ in

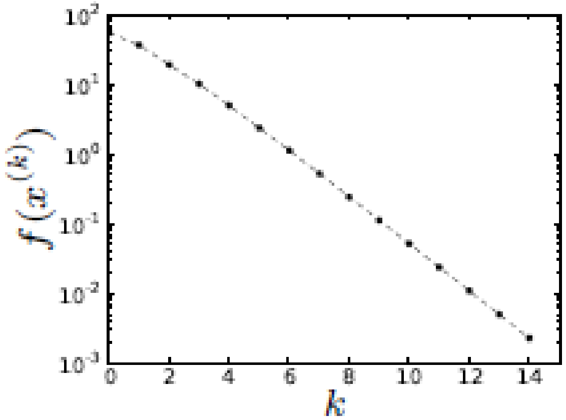
$$x^{(0)} + \operatorname{span}\{\nabla f(x^{(0)}), \nabla f(x^{(1)}), \ldots, \nabla f(x^{(k-1)})\}$$

optimality

no first-order method improves the $1/k^2$ convergence rate (uniformly, over all convex functions with Lipschitz continuous gradients)

quadratic example $\mu = 1.8/L,\ s_k = 0.3$





Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

FISTA method

Beck&Teboulle ‘09

$$\min_x f(x)$$

- Minimize upper approximation at an intermediate point.

$$x^{k+1} = \text{argmin}_y Q_{f,\mu}(y^k, y)$$

$$t_{k+1} := (1 + \sqrt{1 + 4t_k^2})/2$$

$$y^{k+1} := x^k + \frac{t_k - 1}{t_{k+1}}[x^k - x^{k-1}]$$

- If $\mu \leq$ *1/L* then in $O(\sqrt{L/\epsilon})$ iterations finds solution

$$\bar{x} : \; f(\bar{x}) \leq f(x^*) + \epsilon$$

Nondifferentiable optimization by smoothing

for nondifferentiable f that cannot be handled by proximal gradient method

- replace f with differentiable approximation f_μ (parametrized by μ)
- minimize f_μ by (fast) gradient method

μ is not a prox parameter here

complexity: #iterations for (fast) gradient method depends on L_μ/ϵ_μ

- L_μ is Lipschitz constant of ∇f_μ
- ϵ_μ is accuracy with which the smooth problem is solved

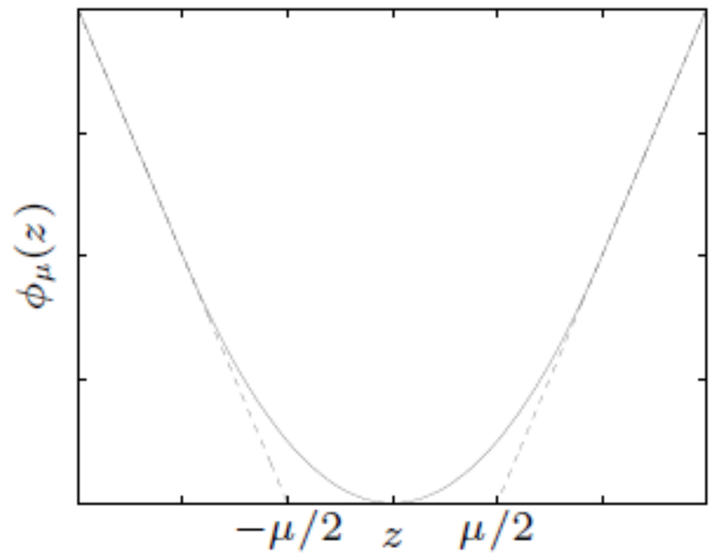
trade-off in amount of smoothing (choice of μ)

- large L_μ (less smoothing) gives more accurate approximation
- small L_μ (more smoothing) gives faster convergence

Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

Example: Huber penalty as smoothed absolute value

$$\phi_\mu(z) = \begin{cases} z^2/(2\mu) & |z| \le \mu \\ |z| - \mu/2 & |z| \ge \mu \end{cases}$$



μ controls accuracy and smoothness

- accuracy

$$|z| - \frac{\mu}{2} \le \phi_\mu(z) \le |z|$$

- smoothness

$$\phi''_\mu(z) \le \frac{1}{\mu}$$

Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

Huber penalty approximation of 1-norm minimization

$$f(x) = \|Ax - b\|_1, \qquad f_\mu(x) = \sum_{i=1}^{m} \phi_\mu(a_i^T x - b_i)$$

- accuracy: from $f(x) - m\mu/2 \le f_\mu(x) \le f(x)$,

$$f(x) - f^\star \le f_\mu(x) - f_\mu^\star + \frac{m\mu}{2}$$

  to achieve $f(x) - f^\star \le \epsilon$ we need $f_\mu(x) - f_\mu^\star \le \epsilon_\mu$ with $\epsilon_\mu = \epsilon - m\mu/2$

- Lipschitz constant of f_μ is $L_\mu = \|A\|_2^2/\mu$

complexity: for $\mu = \epsilon/m$

$$\frac{L_\mu}{\epsilon_\mu} = \frac{\|A\|_2^2}{\mu(\epsilon - m\mu/2)} = \frac{2m\|A\|^2}{\epsilon^2}$$

i.e., $O(\sqrt{L_\mu/\epsilon_\mu}) = O(1/\epsilon)$ iteration complexity for fast gradient method

Slides from L. Vandenberghe
http://www.ee.ucla.edu/~vandenbe/ee236c.html

Unconstrained formulation of the SVM problem

Given a training set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, $x_i \in \mathbf{R}^d$, $y \in \{+1, -1\}$

$$\min_w f(w) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} \ell(w, (x_i, y_i))$$

where

$$\ell(w, (x, y)) = \max\{0, 1 - y(w^\top x)\}$$

Find $f(w) \leq f(w^*) + \epsilon$ - ϵ-optimal solution.

SVM problem using Huber loss function

Given a training set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$,
$x_i \in \mathbf{R}^d$, $y \in \{+1, -1\}$

$$\min_w f(w) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} \phi_\mu(w, (x_i, y_i))$$

where

$$\phi_\mu(w, (x, y)) = \begin{cases} 0 & y(w^\top x) \geq 1 \\ \frac{(y(w^\top x)-1)^2}{2\mu} & 1 - \mu < y_i(w^\top x) < 1 \\ 1 - y(w^\top x) - \frac{\mu}{2} & y(w^\top x) \leq 1 - \mu \end{cases}$$

Find $f(w) \leq f(w^*) + \epsilon$ - ϵ-optimal solution in $O(\frac{1}{\epsilon})$ iterations

First order methods for composite functions

Examples

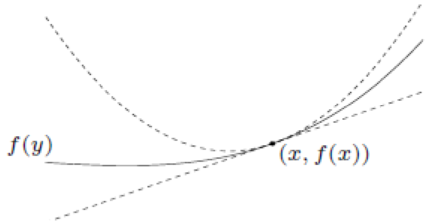
- Lasso or CS: $$\min_x \quad \frac{1}{2}||Ax-b||^2+\lambda||x||_1$$
- Group Lasso or MMV $$\min_x \quad \frac{1}{2}||Ax-b||^2+\lambda\sum_{j\in J}||x_j||$$
- Matrix Completion $$\min_{X\in R^{n\times m}} \lambda\sum_{(i,j)\in I}(X_{ij}-M_{ij})^2+||X||_*$$
- Robust PCA $$\min_{X\in R^{n\times m}} \lambda\|X_{ij}-M_{ij}\|_1+||X||_*$$
- SICS $$\max_X \frac{m}{2}(\log\det X - Tr(AX)) - \lambda||X||_1$$

Prox method with nonsmooth term

- Consider: $\min_x F(x) = f(x) + g(x)$

$$|\nabla f(x) - \nabla f(y)| \leq L||x - y||$$

- Quadratic upper approximation



$$f(y) + g(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2\mu}||y - x||^2 + g(y) = Q_{f,\mu}(x, y)$$

$$F(y) \leq f(x) + \frac{1}{2\mu}||x - \mu \nabla f(x)^\top - y||^2 + g(y) = Q_{f,\mu}(x, y)$$

Assume that *g(y)* is such that the above function is easy to optimize over *y*

Example 1 (Lasso and SICS)

$$\min_x f(x) + ||x||_1$$

- Minimize upper approximation function $Q_{f,\mu}(x,y)$ on each iteration

$$\min_y Q_{f,\mu}(x,y) = \min_y f(x) + \frac{1}{2\mu}||x - \mu\nabla f(x)^\top - y||^2 + ||y||_1$$

$$\Updownarrow$$

$$\sum_i \min_{y_i}\left[\frac{1}{2\mu}(y_i - r_i)^2 + |y_i|\right]$$

$$\Updownarrow$$

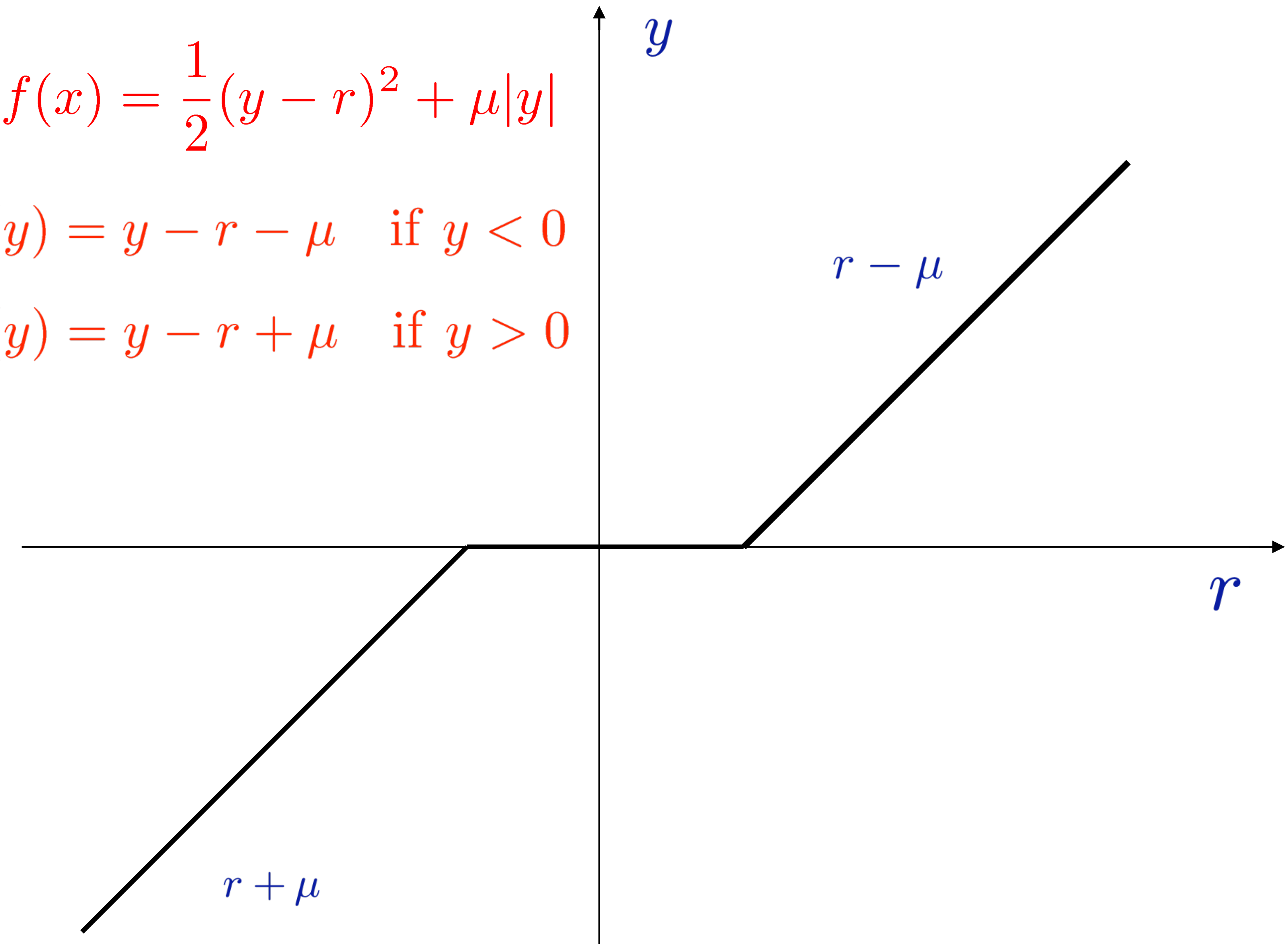
$$\min_{y_i} \frac{1}{2}(y_i - r_i)^2 + \mu|y_i| \rightarrow y_i^* = \begin{cases} r_i - \mu & \text{if } r_i > \mu \\ 0 & \text{if } -\lambda \le r_i \le \mu \\ r_i + \mu & \text{if } r_i < -\mu \end{cases}$$

Closed form solution!
O(n) effort

$$f(x) = \frac{1}{2}(y-r)^2 + \mu|y|$$

$$f'(y) = y - r - \mu \quad \text{if } y < 0$$

$$f'(y) = y - r + \mu \quad \text{if } y > 0$$



Example 2 (Group Lasso)

$$\min_x f(x) + \sum_i ||x_i||, \;\; x_i \in \mathrm{R}^{n_i}$$

Very similar to the previous case, but with ||.|| instead of |.|

$$\sum_i \min_{y_i \in \mathrm{R}^{n_i}} \left[\frac{1}{2\mu}(y_i - r_i)^2 + \|y_i\| \right]$$

$$\Updownarrow$$

$$y_i^* = \frac{r_i}{\|r_i\|} \max(0, \|r_i\| - \mu)$$

Closed form solution! *O(n)* effort

Example 3 (Collaborative Prediction)

$$\min_{X \in \mathrm{R}^{n \times m}} f(X) + ||X||_*$$

$$\min_Y Q_f(X, Y)$$

$$\Updownarrow$$

$$\min_Y \left[\frac{1}{2\mu} \|Y - Z\|_F^2 + \|Y\|_* \right]$$

$$\Updownarrow$$

$$Z = P \mathrm{diag}\left\{\sigma_1, \sigma_2, \ldots, \sigma_n\right\} Q^\top$$

Closed form solution! *O(n^3)* effort

$$Y^* = P \mathrm{diag}\left\{\sigma_1^*, \sigma_2^*, \ldots, \sigma_n^*\right\} Q^\top, \; \sigma_i^* = \begin{cases} \sigma_i - \mu & \text{if } \sigma_i > \mu \\ 0 & \text{if } -\mu \le \sigma_i \le \mu \\ \sigma_i + \mu & \text{if } \sigma_i < -\mu \end{cases}$$

ISTA/Gradient prox method

$$\min_x F(x) = f(x) + g(x)$$

- Minimize quadratic upper approximation on each iteration

$$x^{k+1} = \text{argmin}_y Q_f(x^k, y)$$

$$Q_{f,\mu}(x, y) = f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2\mu}||y - x||^2 + g(y)$$

- If $\mu \leq 1/L$ then in $O(L/\epsilon)$ iterations finds solution

$$\bar{x} : \; F(\bar{x}) \leq F(x^*) + \epsilon$$

Fast first-order method

Nesterov, Beck & Teboulle

$$\min_x F(x) = f(x) + g(x)$$

- Minimize upper approximation at an "accelerated" point.

$$x^k = \mathrm{argmin}_y Q_f(y^k, y)$$

$$t_{k+1} := \quad (1 + \sqrt{1 + 4t_k^2})/2$$
$$y^{k+1} := \quad x^k + \frac{t_k - 1}{t_{k+1}}[x^k - x^{k-1}]$$

- If $\mu \leq$ *1/L* then in $O(\sqrt{L/\epsilon})$ iterations finds solution

$$\bar{x} : \; F(\bar{x}) \leq F(x^*) + \epsilon$$

Practical first order algorithms using backtracking search

Iterative Shrinkage Threshholding Algorithm (ISTA)

$$\min_x F(x) = f(x) + g(x)$$

- Minimize quadratic upper relaxation on each iteration

$$x^{k+1} = \text{argmin}_y Q_f(x^k, y) = f(x^k) + \frac{1}{2\mu_k}||x^k - \mu_k \nabla f(x^k)^\top - y||^2 + g(y)$$

- Using line search find μ_k such that

$$F(x^{k+1}) \leq Q_f(x^k, x^{k+1})$$

- In $O(1/\mu_{min}\epsilon)$ iterations finds ϵ-optimal solution (in practice better)

Nesterov, 07
Beck&Teboulle, Tseng,
Auslender&Teboulle, 08

Fast Iterative Shrinkage Threshholding Algorithm (FISTA)

$$\min_x F(x) = f(x) + g(x)$$

- Minimize quadratic upper relaxation on each iteration

$$x^k = \operatorname{argmin}_y Q_f(y^k, y) = f(y^k) + \frac{1}{2\mu_k}||y^k - \mu_k \nabla f(y^k)^\top - y||^2 + g(y)$$

- Using line search find $\mu_k \leq \mu_{k-1}$ such that

Very restrictive

$$F(x^k) \leq Q_f(y^k, x^k)$$

$$t_{k+1} := (1 + \sqrt{1 + 4t_k^2})/2$$

$$y^{k+1} := x^k + \frac{t_k - 1}{t_{k+1}}[x^k - x^{k-1}]$$

- In $O(\sqrt{1/\mu_{min}\epsilon})$ iterations finds ϵ-optimal solution

Nesterov, Beck&Teboulle, Tseng

FISTA with line search

Goldfarb and S. 2010

- ISTA's complexity is O(L/ϵ) while FISTA's is $O(\sqrt{L/\epsilon})$
- However, FISTA's condition $\mu_k \leq \mu_{k-1}$ often slows down practical performance and simply ignoring the condition does not help.

$$F(x^{k+1}) \leq Q_f(y^k, x^{k+1})$$

$$\begin{aligned} t_{k+1} &:= (1 + \sqrt{1 + 4\theta_k t_k^2})/2 \\ y^{k+1} &:= x^k + \tfrac{t_k - 1}{t_{k+1}}[x^k - x^{k-1}] \end{aligned}$$

- We want to modify FISTA algorithm to relax $\mu_k \leq \mu_{k-1}$, while maintaining $O(\sqrt{L/\epsilon})$ complexity bound or maybe even improving it

Find $\mu_k \leq \mu_{k-1}$ such that

Cycle to find μ_k

$$x^k = \text{argmin}_y Q_f(y^k, y)$$

$$F(x^k) \leq Q_f(y^k, x^k)$$

$$t_{k+1} := \quad (1 + \sqrt{1 + 4t_k^2})/2$$

$$y^{k+1} := \quad x^k + \frac{t_k - 1}{t_{k+1}}[x^k - x^{k-1}]$$

Convergence rate: $$F(x^k) - F(x^*) \leq \frac{2L\|x^0 - x^*\|^2}{k^2}$$

ϵ

Find μ_k such that

$$x^k = \mathrm{argmin}_y Q_f(y^k, y)$$

$$F(x^k) \le Q_f(y^k, x^k)$$

This condition….

$$\mu_k t_k^2 \ge \mu_{k+1} t_{k+1}(t_{k+1} - 1)$$

$$y^{k+1} := \quad x^k + \frac{t_k - 1}{t_{k+1}}[x^k - x^{k-1}]$$

$$F(x^k) - F(x^*) \le \frac{\|x^0 - x^*\|^2}{2\mu_k t_k^2}$$

… gives this bound on the error

Goldfarb & S. 2011

FISTA with full line search

Find μ_k such that

Cycle to find μ and t

$$x^k = \text{argmin}_y Q_f(y^k, y)$$

$$F(x^k) \le Q_f(y^k, x^k)$$

$$\mu_k t_k^2 = \mu_{k+1} t_{k+1}(t_{k+1} - 1)$$

$$y^{k+1} := x^k + \frac{t_k - 1}{t_{k+1}}[x^k - x^{k-1}]$$

$$\Downarrow$$

$$\mu_k t_k^2 \ge \left(\sum_{i=1}^k \sqrt{\mu_i}/2\right)^2 \ge \frac{k^2}{4L}$$

$$F(x^k) - F(x^*) \le \frac{\|x^0 - x^*\|^2}{(2\sum_{i=1}^k \sqrt{\mu_i}/2)^2}$$

Goldfarb & S. 2011