

4. Gradient methods for nonsmooth problems

- motivation
- example: 1-norm regularization
- gradient mapping
- gradient method
- Nesterov's gradient method
- examples

4-1

Motivation

complexity results from previous lectures

bounds on number of iterations to reach accuracy $f(x) - f^* \leq \epsilon$:

- subgradient method: $O(1/\epsilon^2)$
- gradient method: $O(1/\epsilon)$
- Nesterov's optimal gradient method: $O(1/\sqrt{\epsilon})$

can the faster gradient methods be extended to nonsmooth problems?

- no, if we consider the problem class and the (oracle) algorithm model for which the subgradient method is known to be optimal
- yes, if we can take advantage of additional structure in the problem

Interpretation of gradient update

recall the gradient update for convex differentiable f :

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

interpretation

$$x^+ = \operatorname{argmin}_z \left(f(x) + \nabla f(x)^T(z - x) + \frac{1}{2t}\|z - x\|_2^2 \right)$$

x^+ minimizes a quadratic approximation of f , consisting of

- the first-order linearization $f(x) + \nabla f(x)^T(z - x)$ of $f(z)$ at x
- a proximity term $\|z - x\|_2^2$ with weight $1/(2t)$

Extension to nondifferentiable problems

split f in a smooth and a nonsmooth component:

$$\text{minimize } f(x) = g(x) + h(x)$$

g convex, differentiable; h convex, nondifferentiable

generalized gradient update

$$x^+ = \operatorname{argmin}_z \left(g(x) + \nabla g(x)^T(z - x) + \frac{1}{2t}\|z - x\|_2^2 + h(z) \right)$$

- we make a quadratic approximation to g only
- complexity of computing x^+ depends on structure of h

repeating the update provides a 'gradient method' for minimizing f

Example: 1-norm regularization

$$\text{minimize } f(x) = g(x) + \|x\|_1$$

generalized gradient update

$$\begin{aligned} x^+ &= \operatorname{argmin}_z \left(g(x) + \nabla g(x)^T(z - x) + \frac{1}{2t} \|z - x\|_2^2 + \|z\|_1 \right) \\ &= \operatorname{argmin}_z \left(\frac{1}{2t} \|z - x + t \nabla g(x)\|_2^2 + \|z\|_1 \right) \\ &= S_t(x - t \nabla g(x)) \end{aligned}$$

where

$$S_t(y) \triangleq \operatorname{argmin}_z \left(\frac{1}{2t} \|z - y\|_2^2 + \|z\|_1 \right)$$

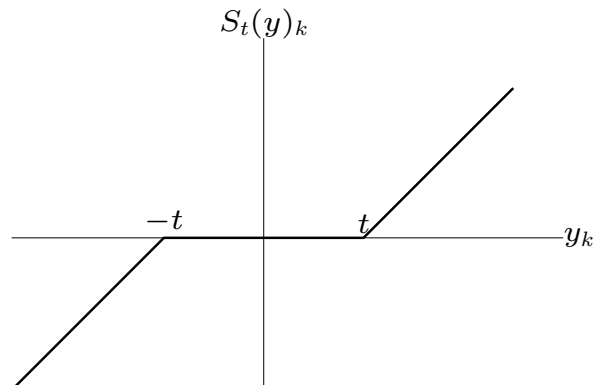
computing S_t : solve a simple separable problem in $z \in \mathbf{R}^n$

$$\text{minimize } \sum_{k=1}^n \left(\frac{1}{2t} (z_k - y_k)^2 + |z_k| \right)$$

solution:

$$S_t(y)_k = \begin{cases} y_k - t & y_k \geq t \\ 0 & -t \leq y_k \leq t \\ y_k + t & y_k \leq -t \end{cases}$$

S_t is often called the *shrinkage* or *soft thresholding* operator



Generalized thresholding operator

for convex h with $\text{dom } h = \mathbf{R}^n$, and $t > 0$, define

$$S_t(y) = \underset{z}{\operatorname{argmin}} \left(\frac{1}{2t} \|z - y\|_2^2 + h(z) \right)$$

properties

- $S_t(y)$ exists and is unique for all y
- optimality condition: $S_t(y)$ satisfies

$$\frac{1}{t}(S_t(y) - y) + v = 0, \quad v \in \partial h(S_t(y))$$

(i.e., v is a subgradient of h at $S_t(y)$)

Gradient update with thresholding

to minimize $g(x) + h(x)$, repeat the update

$$x^+ = S_t(x - t\nabla g(x))$$

- gradient update for smooth component, followed by thresholding
- interpretation

$$\begin{aligned} & S_t(x - t\nabla g(x)) \\ &= \underset{z}{\operatorname{argmin}} \left(g(x) + \nabla g(x)^T(z - x) + \frac{1}{2t} \|z - x\|_2^2 + h(z) \right) \\ &= \underset{z}{\operatorname{argmin}} \left(\frac{1}{2t} \|z - x + t\nabla g(x)\|_2^2 + h(z) \right) \end{aligned}$$

Gradient map

for $t > 0$, define gradient map for $f(x) = g(x) + h(x)$ as

$$G_t(x) = \frac{1}{t}(x - S_t(x - t\nabla g(x)))$$

- allows us to write gradient update as $S_t(x - t\nabla g(x)) = x - tG_t(x)$
- when $h = 0$, reduces to $G_t(x) = \nabla f(x)$
- from optimality condition on page 4–7:

$$G_t(x) = \nabla g(x) + v, \quad v \in \partial h(x - tG_t(x))$$

we will see that the gradient map has many of the properties of a gradient

Optimality condition

$$\text{minimize } f(x) = g(x) + h(x)$$

optimality condition: x is optimal if and only if

$$G_t(x) = 0$$

(for any $t > 0$)

- proof: from the previous page, $G_t(x) = 0$ is equivalent to

$$\nabla g(x) + v = 0, \quad v \in \partial h(x)$$

this is the optimality condition for minimizing f

- reduces to $\nabla f(x) = 0$ when $h = 0$

Gradient method

to minimize $f(x) = g(x) + h(x)$: choose $x^{(0)}$ and repeat

$$x^{(k)} = x^{(k-1)} - t_k G_{t_k}(x^{(k-1)}), \quad k = 1, 2, \dots$$

step size rules

- fixed: t_k constant
- backtracking line search: start at $t := \hat{t}$, repeat $t := \beta t$ until

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

can use the same \hat{t} in each iteration, or use $\hat{t} = t_{k-1}$ in iteration k

Assumptions

- f has finite optimal value f^* , minimizer x^*
- g and h are convex with domains \mathbf{R}^n
- $\nabla g(x)$ is Lipschitz continuous with constant $L > 0$

$$\|\nabla g(x) - \nabla g(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y$$

Upper bound in negative gradient direction

for all x, y

$$f(x - tG_t(x)) \leq f(y) + G_t(x)^T(x - y) - t\left(1 - \frac{tL}{2}\right)\|G_t(x)\|_2^2$$

- if $y = x$, this reduces to

$$f(x - tG_t(x)) \leq f(x) - t\left(1 - \frac{tL}{2}\right)\|G_t(x)\|_2^2$$

- these inequalities (with $G_t(x) = \nabla f(x)$) were the basis of the analysis of the gradient method (page 7 of lecture 1)

proof. define $v = G_t(x) - \nabla g(x)$

- from convexity of g ,

$$\begin{aligned} g(x) &\leq g(y) + \nabla g(x)^T(x - y) \\ &= g(y) + G_t(x)^T(x - y) - v^T(x - y) \end{aligned}$$

- from the fact that v is a subgradient of h at $x - tG_t(x)$,

$$h(x - tG_t(x)) \leq h(y) + v^T(x - tG_t(x) - y)$$

- from the quadratic bound for g on page 5 of lecture 1,

$$\begin{aligned} f(x - tG_t(x)) &\leq g(x) - t\nabla g(x)^T G_t(x) + \frac{Lt^2}{2}\|G_t(x)\|_2^2 + h(x - tG_t(x)) \\ &= g(x) - t\left(1 - \frac{tL}{2}\right)\|G_t(x)\|_2^2 + h(x - tG_t(x)) + tv^T G_t(x) \\ &\leq g(y) + G_t(x)^T(x - y) - t\left(1 - \frac{tL}{2}\right)\|G_t(x)\|_2^2 + h(y) \end{aligned}$$

Convergence analysis

if $x^+ = x - tG_t(x)$ and $0 < t \leq 1/L$, then from the bound on page 4–13

$$\begin{aligned} f(x^+) &\leq f^* + G_t(x)^T(x - x^*) - \frac{t}{2}\|G_t(x)\|_2^2 \\ &= f^* + \frac{1}{2t}(\|x - x^*\|_2^2 - \|x - x^* - tG_t(x)\|_2^2)^2 \\ &= f^* + \frac{1}{2t}(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2)^2 \end{aligned}$$

as on page 8–10 of lecture 1, this implies

$$f(x^{(k)}) - f^* \leq \frac{1}{2kt}\|x^{(0)} - x^*\|_2^2, \quad f(x^{(k)}) - f^* \leq \frac{1}{2kt_{\min}}\|x^{(0)} - x^*\|_2^2$$

for constant step size, resp., backtracking line search

conclusion: #iterations to reach $f(x^{(k)}) - f^* \leq \epsilon$ is $O(1/\epsilon)$

Nesterov's method

choose $x^{(0)}$ and set $y^{(0)} = x^{(0)}$

repeat for $k = 1, 2, \dots$

$$\begin{aligned} x^{(k)} &= y^{(k-1)} - t_k G_{t_k}(y^{(k-1)}) \\ y^{(k)} &= x^{(k)} + \frac{k-1}{k+2}(x^{(k)} - x^{(k-1)}) \end{aligned}$$

step size rules

- fixed: t_k constant
- backtracking line search: start at $t = \hat{t}$, repeat $t := \beta t$ until

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

use $\hat{t} = t_{k-1}$ in iteration k

Convergence analysis

almost identical to the proof in lecture 1

- on page 1-15 we replace $\nabla f(y)$ with $G_t(y)$
- the 2nd inequality on p. 1-15 follows from p. 4-13 with $0 < t \leq 1/L$:

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

and

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

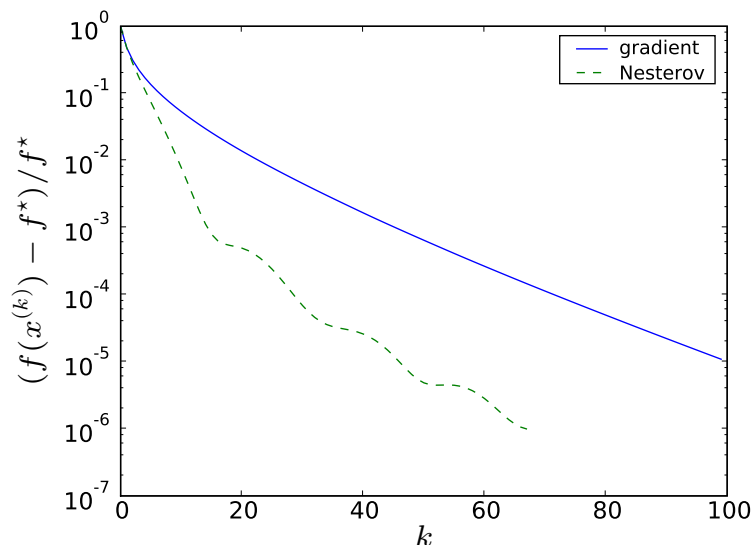
making a convex combination gives

$$f(x^+) \leq (1 - \theta)f(x) + \theta f^* + G_t(y)^T(y - (1 - \theta)x - \theta x^*) - \frac{t}{2} \|G_t(y)\|_2^2$$

conclusion: #iterations to reach $f(x^{(k)}) - f^* \leq \epsilon$ is $O(1/\sqrt{\epsilon})$

1-norm regularized least-squares

$$\text{minimize } \frac{1}{2} \|Ax - b\|_2^2 + \|x\|_1$$



randomly generated $A \in \mathbf{R}^{2000 \times 1000}$; step $t_k = 1/L$ with $L = \lambda_{\max}(A^T A)$

Example: ∞ -norm regularization

gradient update for minimizing $g(x) + \|x\|_\infty$ is $x^+ = S_t(x - t\nabla g(x))$ where

$$S_t(y) = \operatorname{argmin}_z \left(\frac{1}{2t} \|z - y\|_2^2 + \|z\|_\infty \right)$$

$S_t(y) = 0$ if $\|y\|_1 \leq t$; otherwise, solve

$$\sum_{k=1}^n \max\{|y_k| - \lambda, 0\} = t$$

for λ , and take

$$S_t(y)_k = \begin{cases} \lambda & y_k \geq \lambda \\ y_k & |y_k| \leq \lambda \\ -\lambda & y_k \leq -\lambda \end{cases}$$

Example: nuclear norm regularization

$$\text{minimize } g(X) + \|X\|_*$$

g is smooth and convex; variable $X \in \mathbf{R}^{m \times n}$ (with $m \geq n$)

nuclear norm

$$\|X\|_* = \sum_i \sigma_i(X)$$

- $\sigma_1(X) \geq \sigma_2(X) \geq \dots$ are the singular values of X
- also known as trace norm, Ky Fan norm, . . .
- the dual norm of the matrix norm $\|\cdot\|$ (maximum singular value)
- for diagonal X , reduces to the 1-norm of $\mathbf{diag}(X)$
- popular as penalty function that promotes low rank

gradient method (and Nesterov's extension) are based on the update

$$X^+ = S_t(X - t\nabla g(X))$$

where S_t is the thresholding operator

$$S_t(Y) = \operatorname{argmin}_Z \left(\frac{1}{2t} \|Z - Y\|_F^2 + \|Z\|_* \right)$$

- take singular value decomposition $Y = P \mathbf{diag}(\sigma_1, \dots, \sigma_n) Q^T$
- apply thresholding to singular values:

$$S_t(Y) = P \mathbf{diag}(\hat{\sigma}_1, \dots, \hat{\sigma}_n) Q^T$$

where

$$\hat{\sigma}_k = \begin{cases} \sigma_k - t & \sigma_k \geq t \\ 0 & -t \leq \sigma_k \leq t \\ \sigma_k + t & \sigma_k \leq -t \end{cases}$$

Approximate low-rank completion

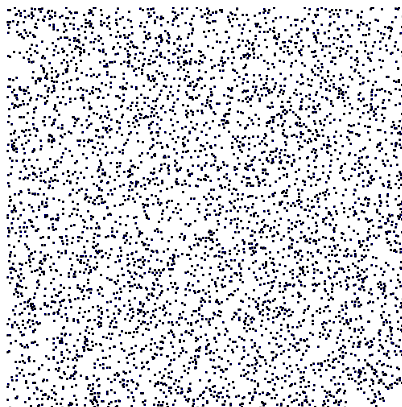
$$\text{minimize } \sum_{(i,j) \in N} (X_{ij} - A_{ij})^2 + \gamma \|X\|_*$$

- entries $(i, j) \in N$ are approximately specified ($X_{ij} \approx A_{ij}$); rest is free
- nuclear norm regularization added to obtain low rank X

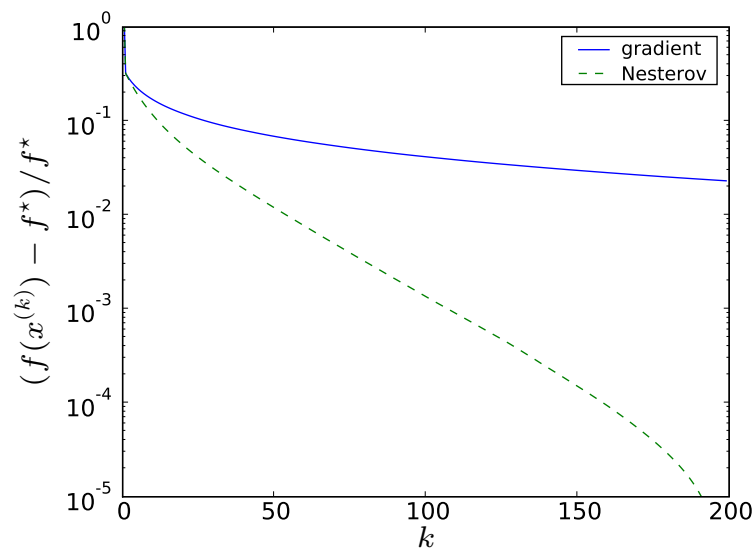
example

$m = n = 500$

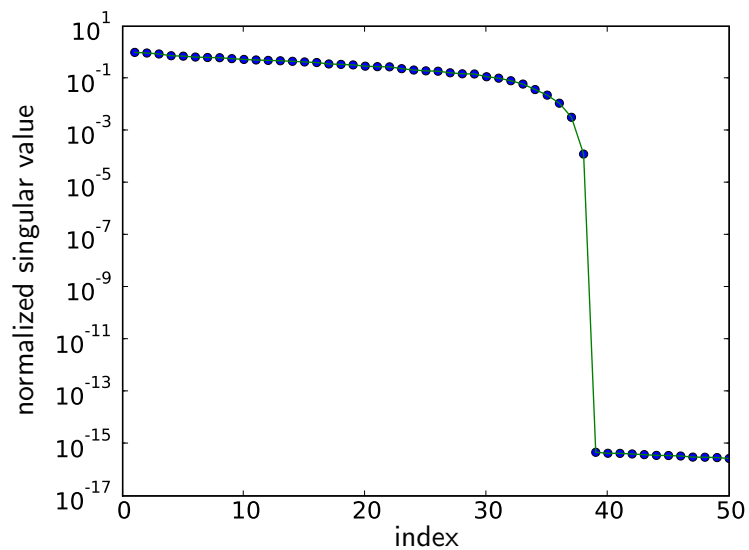
5000 specified entries



convergence (fixed step size $t = 1/L$)



result



optimal X has rank 38; relative error in specified entries is 9%

Reference

A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, to appear in *SIAM J. Imaging Sciences*

- the algorithm on page 4–16 is essentially the FISTA algorithm in the paper
- the convergence analysis on page 4–15 and 4–17