

14. Generalized proximal gradient method

- proximal gradient method with Bregman distance
- accelerated proximal gradient method

Generalized proximal gradient method

- we extend the proximal gradient method of lecture 4 to Bregman distances
- the method applies to convex optimization problems with differentiable term g :

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

Algorithm: start at $x_0 \in \text{dom } f \cap \text{int}(\text{dom } \phi)$ and repeat

$$\begin{aligned} x_{k+1} &= \underset{x}{\text{argmin}} \left(g(x_k) + \nabla g(x_k)^T (x - x_k) + h(x) + \frac{1}{t_k} d(x, x_k) \right) \\ &= \text{prox}_{t_k h}^d(x_k, t_k \nabla g(x_k)) \end{aligned}$$

t_k is a positive step size, fixed or selected by line search

Assumptions

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

- h is convex and prox_{th}^d is well defined: for every $x \in \text{int}(\text{dom } \phi)$ and every a ,

$$\text{minimize } h(u) + a^T u + \frac{1}{t}d(u, x)$$

has a unique solution $\text{prox}_{th}^d(x, ta) \in \text{int}(\text{dom } \phi)$

- g is convex and differentiable with $\text{dom } \phi \subseteq \text{dom } g$
- the function $L\phi - g$ is convex, for some $L > 0$; equivalently,

$$g(x) \leq g(y) + \nabla g(y)^T(x - y) + Ld(x, y) \quad \text{for all } (x, y) \in \text{dom } d \quad (1)$$

this is sometimes called *relative smoothness*

- the optimal value f^\star is finite and attained at $x^\star \in \text{dom } \phi$

Consequence of relative smoothness

- the following inequality holds if $0 < t_k \leq 1/L$:

$$g(x_{k+1}) \leq g(x_k) + \nabla g(x_k)^T (x_{k+1} - x_k) + \frac{1}{t_k} d(x_{k+1}, x_k) \quad (2)$$

- if this inequality holds, then for all $x \in \text{dom } f \cap \text{dom } \phi$,

$$\begin{aligned} f(x_{k+1}) &\leq g(x_k) + \nabla g(x_k)^T (x_{k+1} - x_k) + h(x_{k+1}) + \frac{1}{t_k} d(x_{k+1}, x_k) \\ &\leq g(x_k) + \nabla g(x_k)^T (x - x_k) + h(x) + \frac{1}{t_k} (d(x, x_k) - d(x, x_{k+1})) \\ &\leq f(x) + \frac{1}{t_k} (d(x, x_k) - d(x, x_{k+1})) \end{aligned} \quad (3)$$

2nd line is optimality condition for $\text{prox}_{t_k h}^d$ on p.13.21; 3rd line is convexity of g

Descent properties

- substituting $x = x_k$ in (3) shows that

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \frac{1}{t_k} d(x_k, x_{k+1}) \\ &\leq f(x_k) \end{aligned}$$

strict inequality holds if $x_k \neq x_{k+1}$ and the kernel ϕ is strictly convex

- substituting $x = x^\star$ in (3) shows that

$$\begin{aligned} d(x^\star, x_{k+1}) - d(x^\star, x_k) &\leq t_k(f^\star - f(x_{k+1})) \\ &\leq 0 \end{aligned} \tag{4}$$

Convergence of function values

suppose (2) holds at every iteration

$$\begin{aligned} \left(\sum_{i=0}^{k-1} t_i\right)(f(x_k) - f^\star) &\leq \sum_{i=1}^k t_{i-1}(f(x_i) - f^\star) \\ &\leq \sum_{i=1}^k (d(x^\star, x_{i-1}) - d(x^\star, x_i)) \\ &= d(x^\star, x_0) - d(x^\star, x_k) \\ &\leq d(x^\star, x_0) \end{aligned}$$

- first inequality holds because function values $f(x_i)$ are non-increasing
- second inequality is (4)

this shows that

$$f(x_k) - f^\star \leq \frac{d(x^\star, x_0)}{\sum_{i=0}^{k-1} t_i}$$

Step size selection

Fixed step size: for $t_i = 1/L$, the upper bound on the previous page is

$$f(x_k) - f^\star \leq \frac{Ld(x^\star, x_0)}{k}$$

Line search: start at $t_k = \hat{t}$ and backtrack ($t_k := \beta t_k$, with $\beta \in (0, 1)$) until (2) holds

- since (2) holds for $t_k \leq 1/L$, the selected step size satisfies

$$t_k \geq t_{\min} = \min\{\hat{t}, \beta/L\}$$

- the upper bound on the previous page implies that

$$f(x_k) - f^\star \leq \frac{d(x^\star, x_0)}{kt_{\min}}$$

Outline

- proximal gradient method with Bregman distance
- **accelerated proximal gradient method**

Accelerated proximal gradient method

we discuss a Bregman distance variant of FISTA (p. 7.8) for the problem on p. 14.2

Algorithm: start at $x_0 = v_0 \in \text{dom } f \cap \text{int}(\text{dom } \phi)$, and repeat for $k = 0, 1, \dots$:

$$y_{k+1} = x_k + \theta_k(v_k - x_k)$$

$$v_{k+1} = \underset{v}{\operatorname{argmin}} (h(v) + \nabla g(y_{k+1})^T v + \frac{1}{\tau_k} d(v, v_k))$$

$$x_{k+1} = x_k + \theta_k(v_{k+1} - x_k)$$

- step 2 can be written as $v_{k+1} = \operatorname{prox}_{\tau_k h}^d(v_k, \tau_k \nabla g(y_{k+1}))$
- choice of parameters $\theta_k \in (0, 1]$, $\tau_k > 0$ will be discussed on page 14.16
- known as the *improved interior gradient algorithm* (Ausslender & Teboulle, 2006)
- Bregman extension of a gradient projection method by Nesterov (1988)

Feasibility of the iterates

step 2 requires that $\nabla g(y_{k+1})$ exists and that $v_k \in \text{int}(\text{dom } \phi)$

$$y_{k+1} = \theta_k v_k + (1 - \theta_k) x_k$$

$$v_{k+1} = \underset{v}{\text{argmin}} (h(v) + \nabla g(y_{k+1})^T v + \frac{1}{\tau_k} d(v, v_k))$$

$$x_{k+1} = \theta_k v_{k+1} + (1 - \theta_k) x_k$$

suppose $x_0 = v_0 \in \text{dom } f \cap \text{int}(\text{dom } \phi)$ and $\text{dom } \phi \subseteq \text{dom } g$

- step 1: y_{k+1} is a convex combination of v_k and x_k
- step 2: $v_{k+1} \in \text{dom } h \cap \text{int}(\text{dom } \phi)$, by assumption that $\text{prox}_{\tau_k h}^d$ is well defined
- step 3: x_{k+1} is a convex combination of v_{k+1} and x_k

hence, the sequences y_k, v_k, x_k remain in $\text{dom } f \cap \text{int}(\text{dom } \phi)$

Quadratic kernel

for the quadratic distance $d(x, y) = \frac{1}{2}\|x - y\|_2^2$ the algorithm can be written as

$$\begin{aligned}y_{k+1} &= x_k + \theta_k(v_k - x_k) \\v_{k+1} &= \text{prox}_{\tau_k h}(v_k - \tau_k \nabla g(y_{k+1})) \\x_{k+1} &= x_k + \theta_k(v_{k+1} - x_k)\end{aligned}$$

- compare with FISTA (page 7.8): same y -update, different x -, v -updates

$$\begin{aligned}y_{k+1} &= x_k + \theta_k(v_k - x_k) \\x_{k+1} &= \text{prox}_{t_k h}(y_{k+1} - t_k \nabla g(y_{k+1})) \\v_{k+1} &= x_k + \frac{1}{\theta_k}(x_{k+1} - x_k)\end{aligned}$$

- if $h = 0$ and $t_k = \theta_k \tau_k$, the two methods are equivalent
- if $h \neq 0$, points v_k, y_k in FISTA may be outside $\text{dom } h$ (in contrast to 1st method)

Assumptions

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

we make the same assumptions as on page 14.3 with one difference

- ∇g is L -Lipschitz continuous for some norm $\|\cdot\|$:

$$g(x) \leq g(y) + \nabla g(y)^T(x - y) + \frac{L}{2}\|x - y\|^2 \quad \text{for all } x, y \in \text{dom } g$$

- the Bregman kernel ϕ is 1-strongly convex with respect to the same norm:

$$d(x, y) \geq \frac{1}{2}\|x - y\|^2 \quad \text{for all } (x, y) \in \text{dom } d$$

these two assumptions replace the relative smoothness assumption on page 14.3:

$$g(x) \leq g(y) + \nabla g(y)^T(x - y) + Ld(x, y)$$

Consequence of Lipschitz continuity of gradient

- the following inequality holds if $0 < \tau_k \leq 1/(L\theta_k)$:

$$g(x_{k+1}) \leq (1 - \theta_k)g(x_k) + \theta_k \left(g(y_{k+1}) + \nabla g(y_{k+1})^T (v_{k+1} - y_{k+1}) + \frac{1}{\tau_k} d(v_{k+1}, v_k) \right) \quad (5)$$

- if this inequality holds, then for all $x \in \text{dom } f \cap \text{dom } \phi$,

$$\frac{\tau_k}{\theta_k} (f(x_{k+1}) - f(x)) + d(x, v_{k+1}) \leq \frac{\tau_k(1 - \theta_k)}{\theta_k} (f(x_k) - f(x)) + d(x, v_k) \quad (6)$$

(proofs on next pages)

Proof: we show that the inequality (5) holds for $\tau_k = 1/(L\theta_k)$

- we use notation $x^+ = x_{k+1}$, $x = x_k$, $v^+ = v_{k+1}$, $v = v_k$, $y = y_{k+1}$, $\theta = \theta_k$
- from the Lipschitz continuity of ∇g :

$$g(x^+) \leq g(y) + \nabla g(y)^T(x^+ - y) + \frac{L}{2}\|x^+ - y\|^2$$

- from steps 1 and 2 in the algorithm, $\theta(v^+ - v) = x^+ - y$:

$$g(x^+) \leq g(y) + \nabla g(y)^T(x^+ - y) + \frac{L\theta^2}{2}\|v^+ - v\|^2$$

- from strong convexity of the Bregman kernel:

$$g(x^+) \leq g(y) + \nabla g(y)^T(x^+ - y) + L\theta^2 d(v^+, v)$$

- from step 3 in the algorithm, $x^+ = (1 - \theta)x + \theta v^+$:

$$g(x^+) \leq g(y) + (1 - \theta)\nabla g(y)^T(x - y) + \theta\nabla g(y)^T(v^+ - y) + L\theta^2 d(v^+, v)$$

- inequality (5) now follows from $g(y) + \nabla g(y)^T(x - y) \leq g(x)$ (convexity of g)

Proof: we show that (5) implies that (6) holds for all $x \in \text{dom } f \cap \text{dom } \phi$

- the optimality condition for the prox evaluation in step 2 of the algorithm is

$$h(v_{k+1}) \leq h(x) + \nabla g(y_{k+1})^T (x - v_{k+1}) + \frac{1}{\tau_k} (d(x, v_k) - d(x, v_{k+1}) - d(v_{k+1}, v_k))$$

- from Jensen's inequality and $x_{k+1} = (1 - \theta_k)x_k + \theta_k v_{k+1}$:

$$h(x_{k+1}) \leq (1 - \theta_k)h(x_k) + \theta_k \left(h(x) + \nabla g(y_{k+1})^T (x - v_{k+1}) + \frac{1}{\tau_k} (d(x, v_k) - d(x, v_{k+1}) - d(v_{k+1}, v_k)) \right)$$

- combine this with (5):

$$f(x_{k+1}) \leq (1 - \theta_k)f(x_k) + \theta_k \left(h(x) + g(y_{k+1}) + \nabla g(y_{k+1})^T (x - y_{k+1}) + \frac{1}{\tau_k} (d(x, v_k) - d(x, v_{k+1})) \right)$$

- from convexity of g :

$$f(x_{k+1}) \leq (1 - \theta_k)f(x_k) + \theta_k \left(f(x) + \frac{1}{\tau_k} (d(x, v_k) - d(x, v_{k+1})) \right)$$

Parameter selection

- the parameters $\theta_k \in (0, 1]$, $\tau_k > 0$ will be chosen to satisfy (5) and

$$\theta_0 = 1, \quad \frac{\tau_k(1 - \theta_k)}{\theta_k} \leq \frac{\tau_{k-1}}{\theta_{k-1}} \quad \text{for } k \geq 1 \quad (7)$$

- this allows us to combine the inequalities (6) at $x = x^\star$ recursively to obtain

$$\begin{aligned} \frac{\tau_{k-1}}{\theta_{k-1}}(f(x_k) - f(x^\star)) + d(x^\star, v_k) &\leq \frac{\tau_0}{\theta_0}(f(x_1) - f(x^\star)) + d(x^\star, v_1) \\ &\leq \frac{\tau_0(1 - \theta_0)}{\theta_0}(f(x_0) - f(x^\star)) + d(x^\star, v_0) \\ &= d(x^\star, x_0) \end{aligned}$$

hence,

$$f(x_k) - f^\star \leq \frac{\theta_{k-1}}{\tau_{k-1}} d(x^\star, x_0) \quad (8)$$

Fixed step size

if L is known, we choose $\tau_k = 1/(L\theta_k)$ and θ_k that satisfies

$$\theta_0 = 1, \quad \frac{\theta_k^2}{1 - \theta_k} \geq \theta_{k-1}^2 \quad \text{for } k \geq 1$$

- a simple choice is $\theta_k = 2/(k + 2)$
- alternatively, find the smallest allowable θ_k by solving $\theta_k^2/(1 - \theta_k) = \theta_{k-1}^2$:

$$\theta_0 = 1, \quad \theta_k = \frac{-\theta_{k-1}^2 + \sqrt{\theta_{k-1}^4 + 4\theta_{k-1}^2}}{2}, \quad k \geq 1$$

with these choices the bound (8) implies $1/k^2$ convergence:

$$f(x_k) - f^\star \leq \frac{4L}{(k + 1)^2} d(x^\star, x_0)$$

Variable step size

if L is unknown, we take $\tau_k = t_k/\theta_k$, where t_k is estimate of $1/L$, and solve θ_k from

$$\theta_0 = 1, \quad \frac{t_k(1 - \theta_k)}{\theta_k^2} = \frac{t_{k-1}}{\theta_{k-1}^2} \quad \text{for } k \geq 1$$

- to find t_k , we start at $t_k = \hat{t}_k$ and backtrack ($t_k := \beta t_k$) until (5) holds
- for each tentative t_k , we need to recompute y_{k+1} , v_{k+1} , x_{k+1} to evaluate (5)
- since (5) holds for $\tau_k \leq 1/(L\theta_k)$, the selected t_k satisfies $t_k \geq \min \{\hat{t}_k, \beta/L\}$
- it was shown in lecture 7, equation (3), that

$$\frac{\theta_{k-1}^2}{t_{k-1}} = \frac{1}{t_0} \prod_{i=1}^{k-1} (1 - \theta_i) \leq \frac{4}{(2\sqrt{t_0} + \sum_{i=1}^{k-1} \sqrt{t_i})^2}$$

- if $t_{\min} = \min \{\min_i \hat{t}_i, \beta/L\} > 0$, the bound (8) shows $1/k^2$ convergence:

$$f(x_k) - f^* \leq \frac{4/t_{\min}}{(k+1)^2} d(x^*, x_0)$$

Example

Primal problem (variable $x \in \mathbf{R}^n$)

$$\text{minimize } f(x) + \lambda_{\max}(\mathcal{A}(x) + B)$$

- f is strongly convex
- \mathcal{A} maps n -vector x to $m \times m$ symmetric matrix $\mathcal{A}(x) = x_1 A_1 + \cdots + x_n A_n$
- coefficient matrices A_1, \dots, A_n, B are symmetric $m \times m$ matrices

Dual problem (variable $X \in \mathbf{S}^m$)

$$\begin{aligned} &\text{maximize } \text{tr}(BX) - f^*(-\mathcal{A}^{\text{adj}}(X)) \\ &\text{subject to } \text{tr}(X) = 1 \\ & \quad X \geq 0 \end{aligned}$$

\mathcal{A}^{adj} maps symmetric matrix X to n -vector $\mathcal{A}^{\text{adj}}(X) = (\text{tr}(A_1 X), \dots, \text{tr}(A_n X))$

Bregman proximal mapping

we'll apply the generalized proximal gradient method to the dual problem

- kernel is matrix entropy (p.13.11): $\phi(X) = \text{tr}(X \log X)$ with $\text{dom } \phi = \mathbf{S}_+^m$,

$$d(X, Y) = \text{tr}(X \log X - X \log Y - X + Y)$$

- proximal mapping of indicator δ_C of the set $C = \{X \geq 0 \mid \text{tr}(X) = 1\}$ is

$$\underset{\text{tr}(X)=1, X \geq 0}{\text{argmin}} (\text{tr}(AX) + d(X, Y)) = \frac{\exp(-A + \log Y)}{\text{tr}(\exp(-A + \log Y))}$$

exponential and logarithm of symmetric matrix are defined as

$$\log U = \sum_i (\log \lambda_i) q_i q_i^T, \quad \exp U = \sum_i (\exp \lambda_i) q_i q_i^T$$

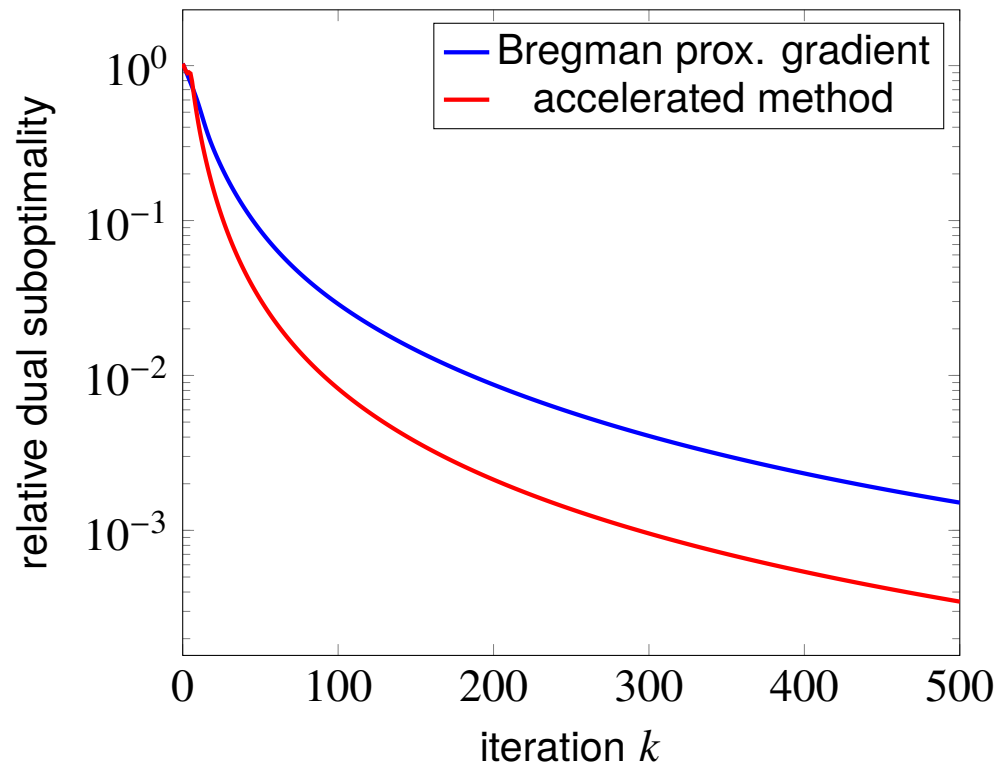
if U has eigenvalue decomposition $U = \sum_i \lambda_i q_i q_i^T$

Example

$$\text{minimize} \quad \frac{1}{2}\|x\|_2^2 + \lambda_{\max}(\mathcal{A}(x) + B)$$

$$\begin{aligned} &\text{maximize} \quad \text{tr}(BX) - \frac{1}{2}\|\mathcal{A}^{\text{adj}}(X)\|_2^2 \\ &\text{subject to} \quad \text{tr}(X) = 1, \quad X \geq 0 \end{aligned}$$

- randomly generated data with $m = 200, n = 100$
- basic and accelerated method, with the same, fixed step size



References

- A. Auslender and M. Teboulle, *Interior gradient and proximal methods for convex and cone optimization*, SIAM J. Optim. (2006).
- P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization* (2008).
The algorithm on page 14.8 is Algorithm 1 in Tseng's paper.