

7. Gradient methods with generalized distances

- Bregman distances
- variant of Nesterov's method
- example

7-1

Gradient method and extension

basic gradient method for minimizing f (lecture 1)

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

extension for minimizing $f + g$ over C (lectures 4-5)

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

- g a simple nondifferentiable function; C a simple convex set
- interesting if projection/thresholding operation S_t is inexpensive

Generalization

replace $(1/2)\|z - x\|_2^2$ with 'generalized distance function' $d(z, x)$

- basic gradient update

$$\operatorname{argmin}_z \left(f(x) + \nabla f(x)^T(z - x) + \frac{1}{t}d(z, x) \right)$$

- extension with projection/thresholding

$$\operatorname{argmin}_{z \in C} \left(f(x) + \nabla f(x)^T(z - x) + \frac{1}{t}d(z, x) + g(z) \right)$$

potential benefits

- select $d(z, x)$ to fit the curvature of f , or geometry of C
- simplify the thresholding/projection

Bregman distance functions

Bregman distance associated with strictly convex, differentiable h :

$$d(x, y) = h(x) - h(y) - \nabla h(y)^T(x - y)$$

h is called the *kernel* function of d

properties

- convex in x for fixed y
- $d(x, y) \geq 0$ for all x, y ; $d(x, y) = 0$ if and only if $x = y$
- not a real distance (not symmetric)
- $d(x, y) \geq (\mu/2)\|x - y\|_2^2$ if h is strongly convex with constant μ

first two properties follow from (strict) convexity of h

Examples

quadratic function: $h(x) = \|x\|_2^2/2$

$$d(x, y) = \frac{1}{2}\|x - y\|_2^2$$

negative entropy: $h(x) = \sum_{i=1}^n x_i \log x_i$ with $\text{dom } h = \mathbf{R}_{++}^n$

$$d(x, y) = \sum_{i=1}^n (x_i \log(x_i/y_i) - x_i + y_i)$$

the relative entropy or Kullback-Leibler divergence

logarithm barrier: $h(x) = -\sum_{i=1}^n \log x_i$ with $\text{dom } h = \mathbf{R}_{++}^n$

$$d(x, y) = \sum_{i=1}^n (x_i/y_i - \log(x_i/y_i)) - n$$

inverse barrier: $h(x) = \sum_{i=1}^n 1/x_i$ with $\text{dom } h = \mathbf{R}_{++}^n$

$$d(x, y) = \sum_{i=1}^n \frac{1}{y_i} \left(\sqrt{\frac{x_i}{y_i}} - \sqrt{\frac{y_i}{x_i}} \right)^2$$

log-det barrier: $h(X) = -\log \det X$ with $\text{dom } h = \mathbf{S}_{++}^n$

$$d(X, Y) = \text{tr}(XY^{-1}) - \log \det(XY^{-1}) - n$$

(follows from $\nabla h(X) = -X^{-1}$)

matrix entropy: $h(X) = \text{tr}(X \log X) = \sum_i \lambda_i(X) \log(\lambda_i(X))$ on \mathbf{S}_{++}^n

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

(follows from $\nabla h(X) = I + \log X$)

Triangle identity

the triangle identity

$$\|x - z\|_2^2 = \|x - y\|_2^2 + \|y - z\|_2^2 + 2(y - z)^T(x - y)$$

was used in the convergence proofs of the gradient methods

triangle identity for Bregman distances

$$d(x, z) = d(x, y) + d(y, z) + (\nabla h(y) - \nabla h(z))^T(x - y)$$

follows directly from the definition of d

Optimization with Bregman regularization

$$\begin{aligned} & \text{minimize} && f(x) + d(x, y) \\ & \text{subject to} && x \in C \end{aligned}$$

- f convex and subdifferentiable on C
- C a convex set
- $y \in \text{dom } h$ (h is the kernel function associated with d)

property: if \hat{x} is optimal, then

$$f(x) + d(x, y) \geq f(\hat{x}) + d(\hat{x}, y) + d(x, \hat{x}) \quad \forall x \in C$$

proof: if \hat{x} is optimal then there is a subgradient $g \in \partial f(\hat{x})$ with

$$(g + \nabla h(\hat{x}) - \nabla h(y))^T (x - \hat{x}) \geq 0 \quad \forall x \in C$$

from the triangle identity

$$\begin{aligned} & f(x) + d(x, y) \\ &= f(x) + (\nabla h(\hat{x}) - \nabla h(y))^T (x - \hat{x}) + d(x, \hat{x}) + d(\hat{x}, y) \\ &\geq f(\hat{x}) + (g + \nabla h(\hat{x}) - \nabla h(y))^T (x - \hat{x}) + d(x, \hat{x}) + d(\hat{x}, y) \\ &\geq f(\hat{x}) + d(\hat{x}, y) + d(x, \hat{x}) \end{aligned}$$

for all $x \in C$

Variant of Nesterov's method

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in C \end{array}$$

- f convex, differentiable; ∇f Lipschitz continuous on C with constant L

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y \in C$$

- $C \subseteq \text{dom } f$ is a closed convex set
- we assume the problem is solvable

$d(x, y)$ is a Bregman distance on C , and

$$d(x, y) \geq \frac{1}{2}\|x - y\|_2^2 \quad \forall x, y \in C$$

(i.e., kernel h is strongly convex, with strong convexity constant 1)

algorithm

choose $x^{(0)} \in C$ and set $v^{(0)} = x^{(0)}$

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

$$\begin{aligned} y^{(k)} &= x^{(k-1)} + \frac{2}{k+1}(v^{(k-1)} - x^{(k-1)}) \\ v^{(k)} &= \operatorname{argmin}_{z \in C} \left(\nabla f(y^{(k)})^T z + \frac{2L}{k+1} d(z, v^{(k-1)}) \right) \\ x^{(k)} &= x^{(k-1)} + \frac{2}{k+1}(v^{(k)} - x^{(k-1)}) \end{aligned}$$

note

- $v^{(k)}, x^{(k)}, y^{(k)}$ are feasible for all k
- not a descent algorithm: $f(x^{(k)})$ can be greater than $f(x^{(k-1)})$

Interpretation

for $C \in \mathbf{R}^n$ and $d(z, v) = (1/2)\|z - v\|_2^2$, step 2 reduces to

$$v^{(k)} = v^{(k-1)} - \frac{k+1}{2L} \nabla f(y^{(k)})$$

eliminating $y^{(k)}$, $v^{(k)}$ gives $x^{(1)} = x^{(0)} - (1/L)\nabla f(x^{(0)})$ and for $k \geq 2$,

$$\begin{aligned} x^{(k)} &= x^{(k-1)} + \frac{k-2}{k+1}(x^{(k-1)} - x^{(k-2)}) \\ &\quad - \frac{1}{L} \nabla f \left(x^{(k-1)} + \frac{k-2}{k+1}(x^{(k-1)} - x^{(k-2)}) \right) \end{aligned}$$

a gradient method with two-step 'momentum' term

Analysis of one iteration

with $x = x^{(i-1)}$, $x^+ = x^{(i)}$, $y = y^{(i)}$, $v = v^{(i-1)}$, $v^+ = v^{(i)}$, $\theta = 2/(i+1)$

1. from Lipschitz continuity of ∇f

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

2. plug in $x^+ = x + \theta(v^+ - x)$ (algorithm step 3)

$$f(x^+) \leq f(y) + \nabla f(y)^T((1-\theta)x + \theta v^+ - y) + \frac{L}{2}\|(1-\theta)x + \theta v^+ - y\|_2^2$$

3. substitute $y = x + \theta(v - x)$ (algorithm step 1) in last term

$$f(x^+) \leq f(y) + \nabla f(y)^T((1-\theta)x + \theta v^+ - y) + \frac{L\theta^2}{2}\|v^+ - v\|_2^2$$

4. from convexity of f and $(1/2)\|v^+ - v\|_2^2 \leq d(v^+, v)$

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

5. v^+ minimizes the right hand side over C ; from page 7–9 this means

$$\begin{aligned} f(x^+) + L\theta^2 d(x^*, v^+) \\ \leq (1 - \theta)f(x) + \theta (f(y) + \nabla f(y)^T(x^* - y) + L\theta d(x^*, v)) \end{aligned}$$

6. from convexity of f

$$f(x^+) + L\theta^2 d(x^*, v^+) \leq (1 - \theta)f(x) + \theta f^* + L\theta^2 d(x^*, v)$$

conclusion

$$\frac{1}{\theta^2}(f(x^+) - f^*) + Ld(x^*, v^+) \leq \frac{1 - \theta}{\theta^2}(f(x) - f^*) + Ld(x^*, v)$$

Iteration complexity

we showed that for $i \geq 1$,

$$\begin{aligned} \frac{(i + 1)^2}{4}(f(x^{(i)}) - f^*) + Ld(x^*, v^{(i)}) \\ \leq \frac{(i - 1)(i + 1)}{4}(f(x^{(i-1)}) - f^*) + Ld(x^*, v^{(i-1)}) \end{aligned}$$

now, iterate from $i = 1$ to $i = k$ and use $(i - 1)(i + 1) \leq i^2$ to get

$$\frac{(k + 1)^2}{4}(f(x^{(k)}) - f^*) + Ld(x^*, v^{(k)}) \leq Ld(x^*, v^{(0)})$$

since the distance function d is nonnegative and $v^{(0)} = x^{(0)}$, we obtain

$$f(x^{(k)}) - f^* \leq \frac{4Ld(x^*, x^{(0)})}{(k + 1)^2}$$

Variations

- step 3 of the algorithm can be replaced by

$$x^{(k)} = \operatorname{argmin}_{z \in C} \left(\nabla f(y^{(k)})^T z + \frac{L}{2} \|z - y^{(k)}\|_2^2 \right)$$

in the analysis of this variant we start at the inequality in 2 (page 7–14)

- if L is unknown, we start with an initial guess and increase L if the inequality in 4 (page 7–15) is not satisfied;
this can be interpreted as backtracking on the ‘step size’ $1/L$

Projection operator

$$Q_t(u, v) := \operatorname{argmin}_{z \in C} \left(u^T z + \frac{1}{t} d(z, v) \right)$$

- allows us to write step 2 of the algorithm as

$$v^{(k)} = Q_t \left(\nabla f(y^{(k)}), v^{(k-1)} \right), \quad t = \frac{k+1}{2L}$$

- the method is well suited for problems where Q_t is inexpensive

Extension

$$\begin{aligned} & \text{minimize} && f(x) + g(x) \\ & \text{subject to} && x \in C \end{aligned}$$

- f convex, ∇f Lipschitz continuous with constant L
- g a 'simple' nondifferentiable convex function

replace step 2 of the algorithm with

$$v^{(k)} = Q_t \left(\nabla f(y^{(k)}), v^{(k-1)} \right), \quad t = \frac{k+1}{2L}$$

where

$$Q_t(u, v) = \operatorname{argmin}_{z \in C} \left(u^T z + g(z) + \frac{1}{t} d(z, v) \right)$$

Example

projection on probability simplex, with relative entropy distance

$$C = \{x \mid x \succeq 0, \mathbf{1}^T x = 1\}, \quad d(x, y) = \sum_{i=1}^n x_i \log(x_i/y_i)$$

projection operator $S_t(u, v)$ (for $v \succ 0$, $\mathbf{1}^T v = 1$) is the solution of

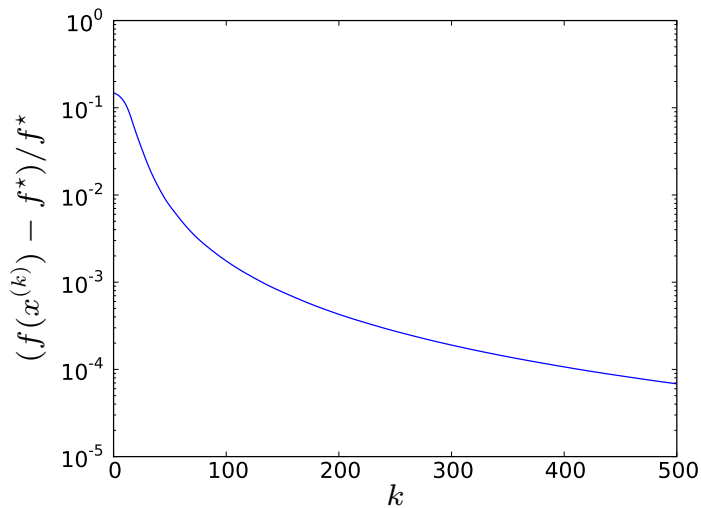
$$\begin{aligned} & \text{minimize} && tu^T z + \sum_{i=1}^n z_i \log(z_i/v_i) \\ & \text{subject to} && \mathbf{1}^T z = 1 \end{aligned}$$

closed form solution:

$$Q_t(u, v)_i = \frac{v_i e^{-tu_i}}{\sum_{j=1}^n v_j e^{-tu_j}}, \quad i = 1, \dots, n$$

example

$$\begin{aligned} & \text{minimize} && (1/2)\|Ax - b\|_2^2 \\ & \text{subject to} && x \succeq 0, \quad \mathbf{1}^T x = 1 \end{aligned}$$



randomly generated A of size 5000×2000

References

- P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization* (2008)
the variant of Nesterov's algorithm is essentially algorithm 1 in the paper; the analysis on page 7–14 also follows this paper
- G. Lan, Z. Lu, R.D.C. Monteiro, *Primal-dual first-order methods with $O(1/\epsilon)$ iteration-complexity for cone programming*, Mathematical programming (2009)
the variant of Nesterov's algorithm with the alternative step 3 (page 7–17) is the second algorithm of section 3 of the paper
- A. Auslender and M. Teboulle, *Interior gradient and proximal methods for convex and cone optimization*, SIAM J. Optim. (2006)
- Yu. Nesterov, *Smooth minimization of non-smooth functions*, Mathematical Programming (2005)
a variant of the algorithm that uses gradients of all past iterations