Agenda

Gradient method with non-Euclidean distances

1. Bregman distance
2. Examples
3. Accelerated non-Euclidean gradient methods
4. Entropic descent algorithm (EDA)

Proximal distance-like function

Basic gradient method

$$x_+ = \mathsf{argmin}_{x \in C} \left\{ f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2t} \|x - x_0\|^2 \right\}$$

with extension to composite functions

Generalization: replace $\| \cdot \|^2$ with some distance-like function

$$x_+ = \mathsf{argmin}_{x \in C} \left\{ f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2t} d(x, x_0) \right\}$$

Extension to composite function $f = g + h$

$$x_+ = \mathsf{argmin}_{x \in C} \left\{ g(x_0) + \nabla g(x_0)^T (x - x_0) + h(x) + \frac{1}{2t} d(x, x_0) \right\}$$

Minimal required properties

- $d(\cdot, x_0)$ cvx for any x_0
- $d(\cdot, \cdot) \geq 0$ and $d(x, x_0) = 0$ iff $x = x_0$

d is not a distance: no symmetry or triangle inequality

Bregman distance functions

- Kernel h is strongly convex
- Bregman distance

$$d(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle$$

- Interpretation: distance above tangent line
- Obeys minimal requirements
- Lack of symmetry is evident

How to choose h?

- Select h to fit geometry of C
- Select h to fit curvature of f, i.e. can add curvature when needed (h strongly convex on feasible set)
- Simplify the projection-like computation

Examples

(1) Negative entropy over simplex $\Delta_n = \{x \in \mathbb{R}^n : x \geq 0,\ 1^T x = 1\}$

$$h(x) = \sum_i x_i \log x_i$$

h is strongly convex wrt to ℓ_1 norm: $d(x,y) \geq \frac{1}{2}\|x-y\|_1^2$ for all x, y in Δ_n

$$\begin{aligned} d(x,y) &= \sum_i (x_i \log x_i - y_i \log y_i) - \sum_i (\log y_i + 1)(x_i - y_i) \\ &= \sum_i (x_i \log(x_i/y_i) - x_i + y_i) \\ &= \sum_i x_i \log(x_i/y_i) \end{aligned}$$

(2) Negative entropy over positive orthant

$$d(x,y) = \sum_i (x_i \log(x_i/y_i) - x_i + y_i)$$

(3) Negative entropy over PSD cone

$$h(X) = \sum_i \lambda_i(X) \log \lambda_i(X) = \mathsf{tr}(X \log X)$$

and

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

(4) Negative entropy over $\{X : X \succeq 0 \text{ and } \mathsf{tr}(X) = 1\}$

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

(5) logarithmic barrier $(x) = -\sum_i \log x_i$ over $\mathbb{R}^n_+$

$$d(x, y) = \sum_i [(x_i/y_i - \log(x_i/y_i) - 1]$$

Logarithmic barrier $h(X) = -\log\det(X)$ over PSD cone

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

Accelerated non-Euclidean gradient method

$$\begin{array}{ll} \text{min} & f(x) \\ \text{s.t.} & x \in C \end{array}$$

f cvx with Lipschitz gradient

Auslender and Teboulle (2006) (h strongly cvx with $\mu \geq 1$)

- Choose x_0 , set $v_0 = x_0, \theta_0 = 1$
- Loop: for $k = 0, 1, 2, \dots$

$$\begin{array}{ll} \text{(a)} & y_k = (1-\theta_k)x_k + \theta_k v_k \\ \text{(b)} & v_{k+1} = \text{argmin}_{x \in C}\{\nabla f(y_k)^T x + L\theta_k d(x, v_k)\} \\ \text{(c)} & x_{k+1} = (1-\theta_k)x_k + \theta_k v_{k+1} \\ \text{(d)} & \theta_{k+1} = \frac{2}{1+\sqrt{1+4/\theta_k^2}} \end{array}$$

x_k, y_k, v_k feasible for all h

Can be extended to composite functions

Interesting if

$$\text{argmin}_{z \in C} \quad u^T z + t^{-1} d(z, v)$$

is computationally cheap

Interpretation: Vandenberghe

$C = \mathbb{R}^n$ and $d(x, y) = \frac{1}{2}\|x - y\|^2$

$$v_{k+1} = v_k - L/\theta_k \nabla f(y_k)$$

Eliminating y_k and v_k and with $\beta_k = \theta_k(1 - \theta_{k-1})/\theta_{k-1}$

$$\begin{aligned} x_{k+1} &= x_k + \theta_k(v_{k+1} - x_k) \\ &= x_k + \beta_k(x_k - x_{k-1}) - (L/\theta_k)\nabla f(x_{k-1} + \beta_k(x_k - x_{k-1})) \end{aligned}$$

Gradient method with two-step momentum term

Extensions

- Can be used with backtracking if L is not known

  Idea: satisfy key inequality in convergence proof (Nesterov ('04), Beck and Teboulle ('09))

- Extension to composite functions $f = g + h$: replace (b) with

$$v_{k+1} = \mathsf{argmin}_{x \in C} \{\nabla g(y_k)^T x + h(x) + L\theta_k d(x, v_k)\}$$

Complexity analysis

Theorem (Auslender Teboulle, 2006)

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

Variations and other schemes

- Nesterov (2005), see 'smoothing lecture': gradient history + 2 prox (one quadratic and one h based)
- Tseng (2008): gradient history + 2 prox h based

Key relationship

Three-point identity

$$\forall x, y, z : d(x,z) = d(x,y) + d(y,z) + \langle \nabla h(y) - \nabla h(z), x - y \rangle$$

Plays a crucial role in the analysis of any optimization method based on Bregman distances

With $h = \frac{1}{2}\|\cdot\|^2$, this is

$$\|x - z\|^2 = \|x - y\|^2 + \|y - z\|^2 + 2\langle y - z, x - y \rangle$$

which played a crucial role in convergence proofs (see proximal and fast proximal lectures)

Entropic descent algorithm (EDA)

$$\begin{array}{ll} \text{min} & f(x) \\ \text{s.t.} & x \in \Delta_n \end{array}$$

- $d(x, y) = \sum_i x_i \log(x_i/y_i)$
- Projection step

$$\text{argmin}_{z \in \Delta_n} \{u^T z + t^{-1} d(z, v)\}$$

  is solution to

$$\begin{array}{ll} \text{min} & t \sum_i u_i z_i + \sum_i z_i \log(z_i/x_i) \\ \text{s.t.} & z_i \geq 0 \\ & \sum_i z_i = 1 \end{array}$$

  and given by

$$z_i = \frac{v_i e^{-tu_i}}{\sum_j v_j e^{-tu_j}}$$

- Convergence: since $d(x^\star, x_0) \leq \log n$

$$f(x_k) - f^\star \leq \frac{4L \cdot \log n}{(k+1)^2}$$

References

- Y. Nesterov. Smooth minimization of non-smooth functions, *Math. Program., Serie A*, **103** (2005)
- A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, **16**(3) (2006)
- L. Vandenberghe. Lecture Notes for EE 236C, UCLA