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

\[ x_+ = \arg\min_{x \in C} \left\{ f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2t} \|x - x_0\|^2 \right\} \]

with extension to composite functions

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

\[ x_+ = \arg\min_{x \in C} \left\{ f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2t} d(x, x_0) \right\} \]

Extension to composite function \( f = g + h \)

\[ x_+ = \arg\min_{x \in C} \left\{ g(x_0) + \nabla g(x_0)^T (x - x_0) + h(x) + \frac{1}{2t} d(x, x_0) \right\} \]
Minimal required properties

- $d(\cdot, x_0)$ convex 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
- Obey 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 $\ell_1$ norm: $d(x, y) \geq \frac{1}{2} \|x - y\|_1^2$ for all $x, y$ in $\Delta_n$

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

\[
= \sum_i (x_i \log(x_i/y_i) - x_i + y_i)
\]

\[
= \sum_i x_i \log(x_i/y_i)
\]

(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

\[ h(X) = \sum_i \lambda_i(X) \log \lambda_i(X) = \text{tr}(X \log X) \]

and

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

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

\[ d(X, Y) = \text{tr}(X (\log X - \log Y)) \]

(5) logarithmic barrier \( (x) = -\sum_i \log x_i \text{ over } \mathbb{R}_+^n \)

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

Logarithmic barrier \( h(X) = -\log \text{det}(X) \text{ over PSD cone} \)

\[ d(X, Y) = \text{tr}(XY^{-1}) - \log \text{det}(XY^{-1}) - n \]
Accelerated non-Euclidean gradient method

\[
\begin{align*}
\min & \quad f(x) \\
\text{s.t.} & \quad x \in C
\end{align*}
\]

\(f\) 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 \(k = 0, 1, 2, \ldots\)
  
  \[\begin{align*}
  (a) & \quad y_k = (1 - \theta_k)x_k + \theta_k v_k \\
  (b) & \quad v_{k+1} = \arg\min_{x \in C} \{\nabla f(y_k)^T x + L\theta_k d(x, v_k)\} \\
  (c) & \quad x_{k+1} = (1 - \theta_k)x_k + \theta_k v_{k+1} \\
  (d) & \quad \theta_{k+1} = \frac{2}{1 + \sqrt{1 + 4/\theta_k^2}}
  \end{align*}\]

\(x_k, y_k, v_k\) 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

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

\[ v_{k+1} = v_k - \frac{L}{\theta_k} \nabla f(y_k) \]

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

\[ x_{k+1} = x_k + \theta_k (v_{k+1} - x_k) \]
\[ = x_k + \beta_k (x_k - x_{k-1}) - \left( \frac{L}{\theta_k} \right) \nabla f(x_{k-1} + \beta_k (x_k - x_{k-1})) \]

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

  $$v_{k+1} = \arg\min_{x \in C} \{\nabla g(y_k)^T x + h(x) + L\theta_k d(x, v_k)\}$$
Complexity analysis

Theorem (Auslender Teboulle, 2006)

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

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)

\[
\begin{align*}
\min & \quad f(x) \\
\text{s.t.} & \quad x \in \Delta_n \\
\end{align*}
\]

- \(d(x, y) = \sum_i x_i \log(x_i/y_i)\)
- **Projection step**
  \[
  \arg\min_{z \in \Delta_n} \{u^T z + t^{-1} d(z, v)\}
  \]
  is solution to
  \[
  \begin{align*}
  \min & \quad t \sum_i u_i z_i + \sum_i z_i \log(z_i/x_i) \\
  \text{s.t.} & \quad z_i \geq 0 \\
  & \quad \sum_i z_i = 1
  \end{align*}
  \]
  and given by
  \[
  z_i = \frac{v_i e^{-tu_i}}{\sum_j v_j e^{-tu_j}}
  \]
- **Convergence:** since \(d(x^*, x_0) \leq \log n\)

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


L. Vandenberghe. Lecture Notes for EE 236C, UCLA