Mirror descent

Yanjie Ze, July 2021

Outline

1. Mirror descent
2. Bregman divergence
3. Alternative forms of mirror descent
4. Convergence analysis

A proximal viewpoint of projected GD

projected GD:

$$x^{t+1} = \arg\min_{x \in C} \left\{ f(x^t) + \langle \nabla f(x^t), x - x^t \rangle + \frac{1}{2\eta_t} \|x - x^t\|_2^2 \right\}$$

The quadratic proximal term is used to monitor the discrepancy between $f(\cdot)$ and its first-order approximation

The quadratic term is based on "certain belief", homogenous penalty $(2\eta_t)^{-1} \|x - x^t\|_2^2$ well approximate

Issue: the local geometry might be highly inhomogeneous

or even non-Euclidean

Example: quadratic minimization $f(x) = \frac{1}{2}(x - x^*)^T Q (x - x^*)$

accelerate:

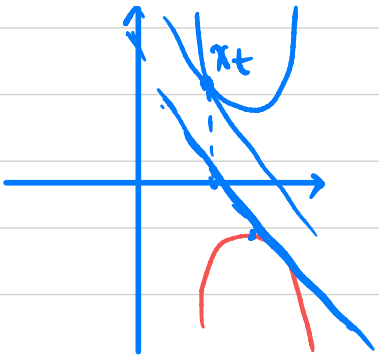
$$x^{t+1} = x^t - \eta_t Q^{-1} \nabla f(x^t)$$

$$= x^t - \eta_t (x^t - x^*)$$

This is the same as:

$$x^{t+1} = \arg\min_{x \in R^n} \left\{ \langle \nabla f(x^t), x - x^t \rangle + \frac{1}{2\eta_t} (x - x^t)(2(x - x^t)) \right\}$$

fits geometry better



why ?

Example: probability simplex

$$\min_{x \in \Delta} f(x)$$

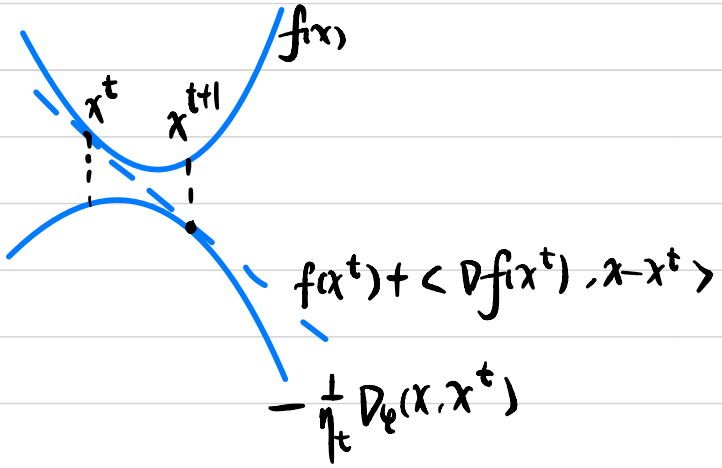
where $\Delta := \{ x \in R_+^n \mid 1^T x = 1 \}$ is probability simplex

Euclidean distance is not recommended here.

Probability divergence is better, like KL divergence

Mirror descent (MD)

We replace the quadratic proximity $\|x - x^t\|_2^2$ with distance-like metric D_φ



$$x^{t+1} = \operatorname*{argmin}_{x \in C} \left\{ f(x^t) + \langle \nabla f(x^t), x - x^t \rangle + \frac{1}{\eta_t} D_\varphi(x, x^t) \right\}$$

where $D_\varphi(x, z) := \varphi(x) - \varphi(z) - \langle \nabla \varphi(z), x - z \rangle$,

Bregman Divergence

for convex and differentiable φ

More generally,

$$x^{t+1} = \operatorname*{argmin}_{x \in C} \left\{ \langle g^t, x - x^t \rangle + \frac{1}{\eta_t} R_\varphi(x, x^t) \right\}$$

with $g^t \in \partial f(x^t)$

Principles in choosing Bregman divergence:

1. fits the local curvature of $f(\cdot)$
2. fits the geometry of the constraint set C
3. makes sure the Bregman projection is inexpensive

Bregman divergence

Let $\varphi : C \to R$ be strictly convex and differentiable on C, then

$$D_\varphi(x, z) = \varphi(x) - \varphi(z) - \langle \nabla\varphi(z), x - z \rangle$$

This is a locally quadratic measure. think of it as

$$D_\varphi(x, z) = (x-z)^T \nabla^2\varphi(\xi)\,(x-z)$$

for some ξ depending on x and z

Example: squared Mahalanobis distance

Let $\varphi(x) = \frac{1}{2}x^T Q x \quad (Q \succ 0)$

Then $D_\varphi(x, z) = \frac{1}{2}x^T Q x - \frac{1}{2}z^T Q z - z^T Q(x-z)$

$$= \frac{1}{2}x^T Q x + \frac{1}{2}z^T Q z - z^T Q x$$

$$= \frac{1}{2}(x-z)^T Q (x-z)$$

When $D_\varphi(x, z) = \frac{1}{2}(x-z)^T Q (x-z)$, $C = R^n$, and f is differentiable MD has a closed-form expression:

$$x^{t+1} = x^t - \eta_t Q^{-1} \nabla f(x^t)$$

In general,

$$x^{t+1} = \arg\min_{x \in C} \{ \eta_t \langle g^t, x \rangle + \frac{1}{2}(x - x^t) Q (x - x^t) \}$$

$$= \arg\min_{x \in C} \{ \frac{1}{2} x^T Q x + \frac{1}{2} x^{tT} Q x^t - \langle Q x^t, x \rangle + \eta_t \langle g^t, x \rangle \}$$

$$= \arg\min_{x \in C} \{ \frac{1}{2} x^T Q x - \langle Q(x^t - \eta_t Q^{-1} g^t), x \rangle + \underbrace{\frac{1}{2} x^{tT} Q x^t}_{\text{can be removed}} \}$$

$$= \arg\min_{x \in C} \{ \frac{1}{2} (x - (x^t - \eta_t Q^{-1} g^t))^T Q (x - (x^t - \eta_t Q^{-1} g^t)) \}$$

This is projection of $x^t - \eta_t Q^{-1} g^t$ based on the weighted l_2 distance $\|z\|_Q^2 = z^T Q z$

Example: KL divergence

Let $\varphi(x) = \sum_i x_i \log x_i$ (negative entropy)

$C = \Delta := \{ x \in R_+^n \mid \sum_i x_i = 1 \}$ is the probability simplex

Then we can generate $D_\varphi(x, z)$:

$$D_\varphi(x, z) = \varphi(x) - \varphi(z) - \langle \nabla \varphi(z), x - z \rangle$$

$$= \sum_i x_i \log x_i - \sum_i z_i \log z_i - \sum_i (\log z_i + 1)(x_i - z_i)$$

$$= \sum_i x_i \log x_i - \cancel{\sum_i z_i \log z_i} - \sum_i x_i \log z_i + \cancel{\sum_i z_i \log z_i} - \underbrace{\sum_i x_i}_{=1} + \underbrace{\sum_i z_i}_{=1}$$

$$= \sum_i x_i \log \frac{x_i}{z_i} = KL(x \| z)$$

When $R_\varphi(x, z) = KL(x \| z)$. ($= \Delta$) , f is differentiable.

MD has closed-form:

$$x^{t+1} = \frac{x_i^t \exp(-\eta_t [\nabla f(x^t)]_i)}{\sum_{j=1}^n x_j^t \exp(-\eta_t [\nabla f(x^t)]_j)} \qquad 1 \le i \le n$$

This is often called exponentiated gradient descent or entropy descent

Example: generalized KL divergence

Example: von Neumann divergence

Common families of Bregman divergence

Function Name	$\varphi(x)$	dom φ	$D_\varphi(x; y)$
Squared norm	$\frac{1}{2}x^2$	$(-\infty, +\infty)$	$\frac{1}{2}(x-y)^2$
Shannon entropy	$x \log x - x$	$[0, +\infty)$	$x \log \frac{x}{y} - x + y$
Bit entropy	$x \log x + (1-x) \log(1-x)$	$[0, 1]$	$x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y}$
Burg entropy	$-\log x$	$(0, +\infty)$	$\frac{x}{y} - \log \frac{x}{y} - 1$
Hellinger	$-\sqrt{1-x^2}$	$[-1, 1]$	$(1-xy)(1-y^2)^{-1/2} - (1-x^2)^{1/2}$
ℓ_p quasi-norm	$-x^p$ $(0 < p < 1)$	$[0, +\infty)$	$-x^p + p\,xy^{p-1} - (p-1)\,y^p$
ℓ_p norm	$\lvert x\rvert^p$ $(1 < p < \infty)$	$(-\infty, +\infty)$	$\lvert x\rvert^p - p\,x \operatorname{sgn} y \lvert y\rvert^{p-1} + (p-1)\lvert y\rvert^p$
Exponential	$\exp x$	$(-\infty, +\infty)$	$\exp x - (x - y + 1) \exp y$
Inverse	$1/x$	$(0, +\infty)$	$1/x + x/y^2 - 2/y$

Basic properties of Bregman divergence:

Let $\varphi: C \to \mathbb{R}$ be u-strongly and differentiable on C

1. non-negativity: $D_\varphi(x, z) \geq 0$ and $D_\varphi(x, z) = 0$ iff $x = z$

2. convexity: $D_\varphi(x, z)$ is convex in x,
   but not necessarily convex in z.

3. lack of symmetry: in general, $D_\varphi(x, z) \neq D_\varphi(z, x)$

4. linearity: for φ_1, φ_2 strictly convex and $\lambda \geq 0$

$$D_{\varphi_1 + \lambda\varphi_2}(x, z) = D_{\varphi_1}(x, z) + \lambda D_{\varphi_2}(x, z)$$

5. unaffected by linear terms:

   let $\varphi_2(x) = \varphi_1(x) + a^T x + b$
   then $D_{\varphi_2} = D_{\varphi_1}$

6. gradient: $\nabla_x D_\varphi(x, z) = \nabla\varphi(x) - \nabla\varphi(z)$

7. Three-point Lemma:

   For every three points x, y, z,

$$D_\varphi(x, z) = D_\varphi(x, y) + D_\varphi(y, z) - \langle \nabla\varphi(z) - \nabla\varphi(y), x - y \rangle$$