Bregman Divergence and Convex Analysis

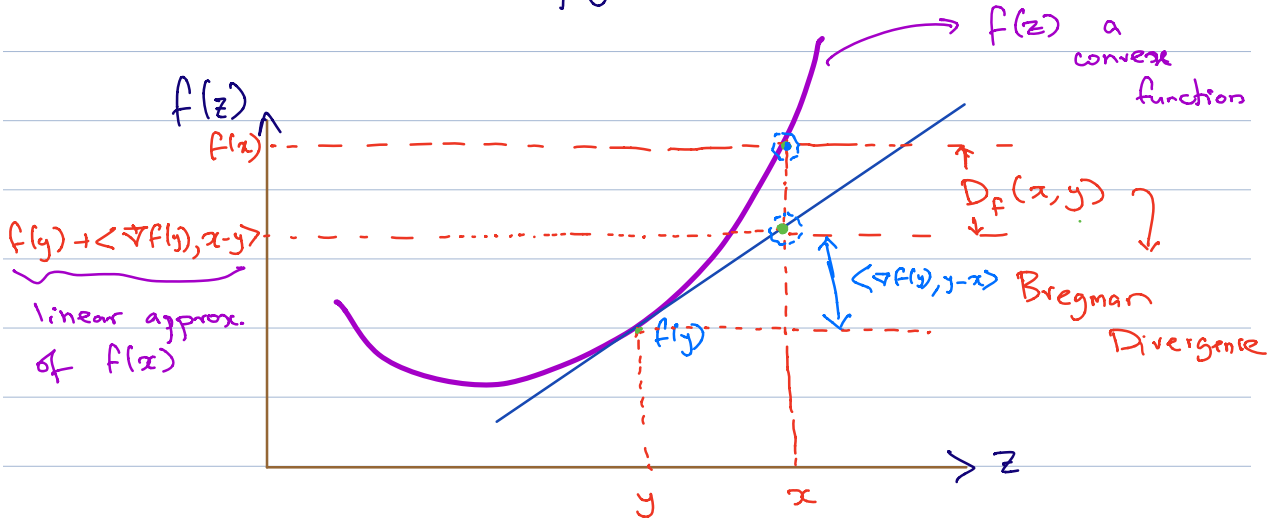
Sources:

1. Bandit Algorithms, L. & S, Chap. 26
2. Bregman Divergence and Mirror Descent, Notes by Prof. Yinhua Zhang, UI Chicago.

(Online: https://www2.cs.uic.edu/%7Ezhangx/teaching/bregman.pdf).

Setting: $f: \mathbb{R}^d \to \mathbb{R}$, a convex function

C a non-empty convex set.



Taylor's Thm: $f(x) = f(y) + \langle \nabla f(y), x-y \rangle$

$+ \frac{1}{2}(x-y)^T \nabla^2 f(y)(x-y) + \text{H.O.T.}$

($\nabla^2 f(y)$: psd (or pd if strictly convex))

Jensen's Inequality: X a $\mathbb{R}^d$ valued r.v, and f a convex fn. Suppose X is proper, and $E[X]$ exists. Then,

$$E[f(X)] \geq f(E[X]).$$

---

ex: $f(x) = x^2$.

$$E[X^2] \geq (E[X])^2, \text{ i.e., } \sigma^2 \geq 0.$$

Strong Convexity: $f : \mathbb{R}^d \to \mathbb{R}$ σ- strongly convex if

$$f(x) \geq f(y) + \langle \nabla f(y), x-y \rangle + \frac{\sigma^2}{2} \|x-y\|^2$$

i.e., in figure above, $D_f(x,y) \geq \frac{\sigma^2}{2} \|x-y\|^2$.

Legendre Function f a strictly convex function is Legendre if it's domain: (a) has a non-empty interior, (b) it is strictly convex and differentiable, and (c) is a steep function.

Steep: $\|\nabla f(x_n)\| \to \infty$ as $x_n \to$ Boundary (Domain f)

Legendre-Fenchel Transform $f^*: \mathbb{R}^d \to \mathbb{R}$ is called the L-F Transform / Fenchel Transform / Convex Dual / Convex Transform if

$$f^*(u) = \sup_x \left(\langle u, x \rangle - f(x) \right)$$

Further, when f is a Legendre function,

(a) $$f(x) = \sup_u \langle u, x \rangle - f^*(x)$$

(b) $$f^{**}(x) = f(x).$$

(c) f^* is a Legendre function

(d) $$(\nabla f)^{-1} = \nabla f^*, \text{ i.e., } \nabla f (\nabla f^*(z)) = z.$$

Bregman Divergence $f: \mathbb{R}^d \to \mathbb{R}$ a convex function. Then,

$$D_f(x,y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle$$

* We will restrict below to Legendre functions (though some of the discussions below hold somewhat more generally).

Examples:

① Suppose f is quadratic, i.e., $f(x) = \frac{1}{2}\|x\|^2$.

Then,

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

More generally, $f(x) = \frac{1}{2} x^T Q x$ → real, symm., p.d. $= U^T \Lambda U$.

From Taylor's Thm,

$$f(x) = f(y) + \langle \nabla f, x-y \rangle + \frac{1}{2}(x-y)^T \nabla^2 f (x-y) + \text{H.O.T.}$$

$\nabla f(y) = Qy$, $\nabla^2 f(y) = Q$. — higher order derivatives $= 0$.

$$\Rightarrow f(x) = f(y) + (x-y)^T Q y + \frac{1}{2}(x-y)^T Q (x-y)$$

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

② Suppose $f(y) = \sum_{i=1}^{d} (y_i \ln(y_i) - y_i)$,

$y \in \rho^{d-1}$, d-dim. simplex, i.e. $\sum_{i=1}^{d} y_i = 1$, $y_i \geq 0$.

Then, $\nabla f(y) = \begin{pmatrix} \ln(y_1) \\ \vdots \\ \ln(y_d) \end{pmatrix} \triangleq \ln(y)$.

$x \in \rho^{d-1}$. Then,

$$D_f(x,y) = \sum_{i=1}^{d} x_i \ln\left(\frac{x_i}{y_i}\right) = KL(x,y)$$

Properties:

ⓐ $D_f(x,y)$ strictly convex in x, nonnegative, with $D_f(x,y) = 0 \iff x = y$.

ⓑ $D_{pf+qg}(x,y) = p D_f(x,y) + q D_g(x,y)$

↑ ↑ Legendre fns. ↳ scalars

ⓒ $\nabla D_f(x,y) = \nabla f(x) - \nabla f(y)$.

ⓓ $D_f(x,y) = D_{f^*}(\nabla f(y), \nabla f(x))$.

(e) Projections: C a closed convex set, $C \subseteq \mathbb{R}^d$, f a Legendre function.

$$\Pi_{C,f}(x) = \text{Projection of } f(x) \text{ onto } C,$$

i.e., $$\Pi_{C,f}(x) = \underset{y \in C}{\operatorname{argmin}} \; D_f(y,x).$$

e.g.: $f = \frac{1}{2}\|x\|^2 \implies \Pi_{C,f}(x) = \underset{y \in C}{\operatorname{argmin}} \frac{1}{2}\|x-y\|^2$

$\Pi_{C,f}(x)$: least squares projection

Theorem (20.5 in textbook): f is Legendre, C a closed convex set. Further suppose:

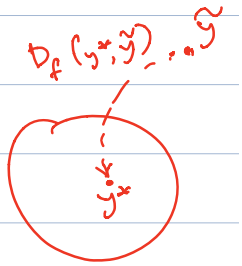
$$\tilde{y} = \underset{y \in \mathbb{R}^d}{\operatorname{argmin}} \; f(y) \quad \text{exists.}$$

Then,

$$y^* = \underset{y \in C}{\operatorname{argmin}} \; f(y) \quad \text{exists and is unique.}$$

Further,

$$y^* = \underset{z \in C}{\operatorname{argmin}} \; D_f(z, \tilde{y})$$



i.e., $y^* = \underset{z \in C}{\operatorname{argmin}} \; D_f\left(z, \underset{y \in \mathbb{R}^d}{\operatorname{argmin}} f(y)\right)$

{ (a) Solve an unconstrained minimization
(b) Project the solution onto C using the Bregman Divergence

$\Longrightarrow$ Solving constrained minimization over C.

Algorithms for Optimization (see ② by X. Zhang for proofs).

(Summary from Ref ②; Also recall class by Constantine Caramanis)

$\min_{x \in C} f(x)$, C closed convex, f is Legendre.

Projected gradient descent: At iteration $k \geq 1$:

$$\tilde{x}_{k+1} = x_k - \eta_k \nabla f(x_k) \quad \longrightarrow \text{unconstrained gradient step}$$

$$x_{k+1} = \underset{x \in C}{\operatorname{argmin}} \; \frac{1}{2} \| x - \tilde{x}_{k+1} \|^2 \quad \longrightarrow \text{projection using } \ell_2\text{-regularizer}$$

penalty / regularization

This is equivalent to:

linear approx. of f at x_k

$$x_{k+1} = \underset{x \in C}{\operatorname{argmin}} \left(f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\eta_k} \|x - x_k\|^2 \right)$$

$\frac{1}{2\eta_k}$ ~ $\psi(\cdot)$

Why: Expand both and check that they are the same.

Motivates **Mirror Descent**

Legendre function: ψ

$= \frac{1}{2}\|\cdot\|^2$ above

$$x_{k+1} = \underset{x \in C}{\operatorname{argmin}} \left(f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\eta_k} D_\psi(x, x_k) \right)$$

Equivalent to:

$$\tilde{x}_{k+1} = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \left(f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\eta_k} D_\psi(x, x_k) \right) \longrightarrow (A)$$

unconstrained

$$x_{k+1} = \underset{x \in C}{\operatorname{argmin}} \; D_\psi(x, \tilde{x}_{k+1}) \longrightarrow (B)$$

Projection using Bregman Divergence

(A): Set gradient $= 0$ to evaluate $\tilde{x}_{k+1}$ (unconstrained problem)

Then, $0 = \nabla f(x_k) + \frac{1}{\eta_k}\left(\nabla\psi(\tilde{x}_{k+1}) - \nabla\psi(x_k)\right)$

Also, $(\nabla\psi)^{-1} = \nabla\psi^*$

$\nabla D_\psi(x, x_k) = \nabla\psi(x) - \nabla\psi(x_k)$, with $x = \tilde{x}_{k+1}$

$$\Rightarrow \quad \tilde{x}_{k+1} = \nabla\psi^*\left(\nabla\psi(x_k) - \eta_k \nabla f(x_k)\right)$$

i.e.,

$x_k \xrightarrow[\nabla\psi(\cdot)]{\text{mirror map}} \nabla\psi(x_k)$

$\downarrow$ gradient step in mirror space

$\tilde{x}_{k+1} \xleftarrow[\nabla\psi^*(\cdot)]{\text{inverse mirror map.}} \nabla\psi(x_k) - \eta_k \nabla f(x_k)$

$\downarrow$ projection onto C using $D_\psi(\cdot,\cdot)$

(B): $x_{k+1} = \underset{x \in C}{\operatorname{argmin}}\ D_\psi(x, \tilde{x}_{k+1})$.

Reason for Studying Mirror Descent

Let suppose $\min_{x \in \rho^{d-1}} f(x)$, i.e., optimizing over the simplex of prob. vectors.

Use $\psi(y) = \sum_{i=1}^{d} (y_i \ln y_i - y_i)$, $y \in \rho^{d-1}$.

Then, (A) becomes:

$$\tilde{x}_{k+1}(i) = x_k(i) e^{-\eta_k \nabla f(x_k)_i}$$

(i^{th} coordinate of $\begin{pmatrix} \nabla f(x_k)_1 \\ \nabla f(x_k)_2 \\ \vdots \\ \nabla f(x_k)_d \end{pmatrix}$)

(B): $$x_{k+1}(i) = \frac{\tilde{x}_{k+1}(i)}{\sum_j \tilde{x}_{k+1}(j)} = \left(\frac{x_k(i) e^{-\eta_k \nabla f(x_k)_i}}{\sum_j x_k(j) e^{-\eta_k \nabla f(x_k)_j}} \right)$$

Further,

$$\|x_k - x^*\| \simeq \left(\frac{\log d}{\sqrt{k}} \right)$$ ← dimension.

whereas standard projected gradient descent in this setting results in:

$$\|x_k - x^*\| \simeq \frac{\sqrt{d}}{\sqrt{k}}$$

} major benefit in terms of scaling with dimension d for mirror descent.

See section 2.1, 2.2 in Notes by X. Zhang for full details.