普通の勾配法

$$\min \quad f(x)$$

$$\nabla = \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \cdots, \frac{\partial}{\partial x_d}\right)^T$$

更新式 $x^{k+1} \leftarrow x^k - \alpha \nabla f(x^k)$

α：学習率

右辺 $x^k - \alpha \nabla f(x^k)$ は次のように書き直せる．

$$x^k - \nabla f(x^k) = \operatorname{argmin}_x \quad Q(x, x^k)$$

ここで関数 $Q(x, x^k)$ は次で定義される．

$$Q(x, x^k) = f(x^k) + \underbrace{\langle \nabla f(x^k), x - x^k \rangle}_{f(x)\text{の}x = x^k\text{まわりの1次展開}} + \frac{1}{2\alpha} \underbrace{\|x - x^k\|_2^2}_{\text{正則化（遠くに行きすぎない）}}$$

つまり、$f(x^k)$ の最適化のために、各イテレーションで関数 $Q(x, x^k)$ を最適化している。

このことは、$\nabla Q(x, x^k) = 0$ を x について解くと更新式 $x^{k+1} \leftarrow x^k - \nabla f(x^k)$ が得られることから確認できる。

$$\nabla Q(x, x^k)$$

$$= \nabla \left\{ f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2\alpha} \| x - x^k \|_2^2 \right\}$$

$$= \nabla \langle f(x^k), x \rangle + \frac{1}{\alpha} (x - x^k)$$

$$= \nabla f(x^k) + \frac{1}{\alpha} (x - x^k)$$

これを x について解いて

$$x = x^k - \alpha \nabla f(x^k)$$

鏡像降下法は、$\|x - x^k\|_2^2$ を Bregman divergence に置き換えたもの.

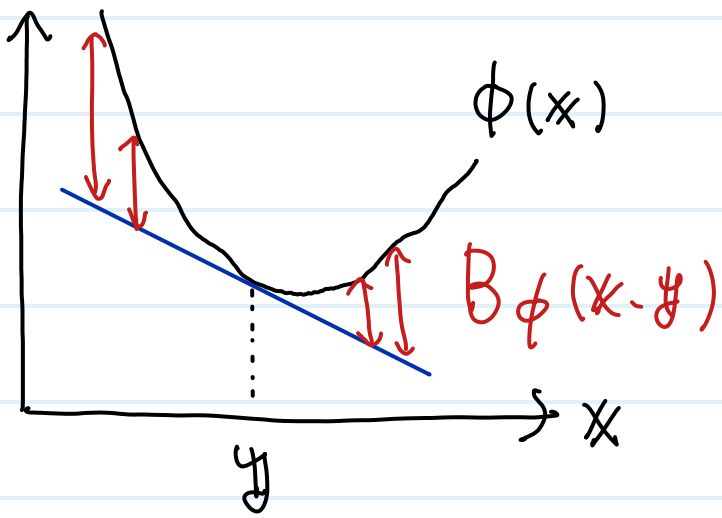
1. Bregman divergence とは.

任意の凸関数 $\phi(x)$ は、定義から

$$\phi(x) \geq \phi(x) + \langle \nabla\phi(y), x - y \rangle$$

$\nabla\phi(y)$：y での接線の傾き

を満たす. つまり、任意の点 y での接線よりも常に上側に値がある.



凸関数 ϕ の Bregman divergence B_ϕ は、

$$B_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y \rangle$$

で定義する。

例. $\phi(x) = \sum_i x_i \log x_i$ とすると

$B_\phi(x, y)$ は、KL 情報量となる。

2. 鏡像降下法

普通の勾配法の $\|x - x^k\|_2^2$ を

$B_\phi(x, x^k)$ で置き換えて、更新式を得る。