普通の勾配法

$$\min \ f(x)$$

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

更新式 $$x^{k+1} \leftarrow x^k - \underbrace{\alpha}_{\text{学習率}} \nabla f(x^k)$$

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

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

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

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

.

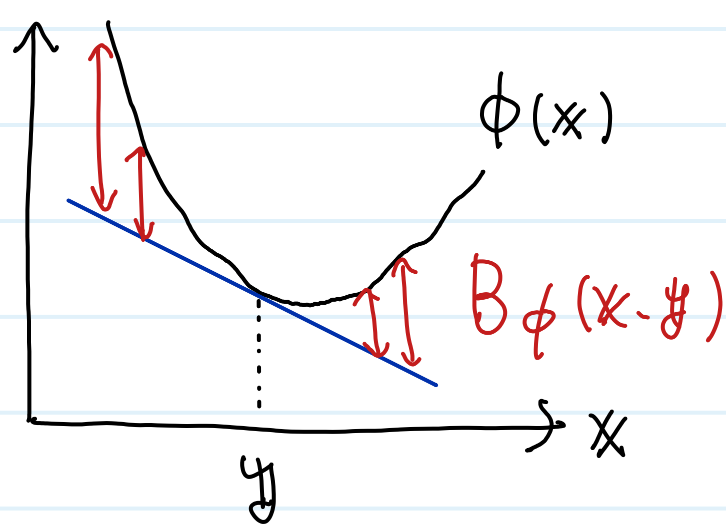
鏡像降下法は、$\|x - x^k\|_2^2$ を
Bregman divergence に置き換えたもの。

1. Bregman divergence とは。

任意の凸関数 $\phi(x)$ は、定義から

$$\phi(x) \geq \phi(x) + \underbrace{\langle \nabla\phi(y), x - y \rangle}_{y\text{での接線の傾き}}$$

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。

<u>例</u>. $\phi(x) = \sum_i x_i \log x_i$ 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。

$$x^{k+1}$$

$$= \underset{x}{\mathrm{argmin}} \left\{ f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{\alpha} B_\phi(x, x^k) \right\}$$

$$= \underset{x}{\mathrm{argmin}} \left\{ f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{\alpha} \left(\phi(x) - \phi(x^k) - \langle \nabla \phi(x^k), x \rangle \right) \right\}$$

$$= \underset{x}{\mathrm{argmin}} \left\{ \langle \nabla f(x^k), x \rangle + \frac{1}{\alpha} \left(\phi(x) - \langle \nabla \phi(x^k), x \rangle \right) \right\}$$

$$= \underset{x}{\mathrm{argmin}} \left\{ \langle \nabla f(x^k) - \frac{1}{\alpha} \phi(x^k), x \rangle + \frac{1}{\alpha} \phi(x) \right\}$$

ここで関数 $P(x, x^k)$ を次のように定義する。

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

$\nabla P(x, x^k) = 0$ をみたすように x を更新。

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

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

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

これは

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

のときに0になる。ここで、$\nabla\phi$の逆関数を用いれば、x^{k+1}について解くことができる

xをθとおいて、$\eta = \nabla\phi(\theta)$と定義すると、
更新式はまさしく

$$\eta^{k+1} \leftarrow \eta^{k+1} - \alpha\nabla f(\theta)$$

となって、とても情報幾何である

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

$\nabla \phi(\mathbb{x}) = \mathbb{1} + \log \mathbb{x}$ であるから

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

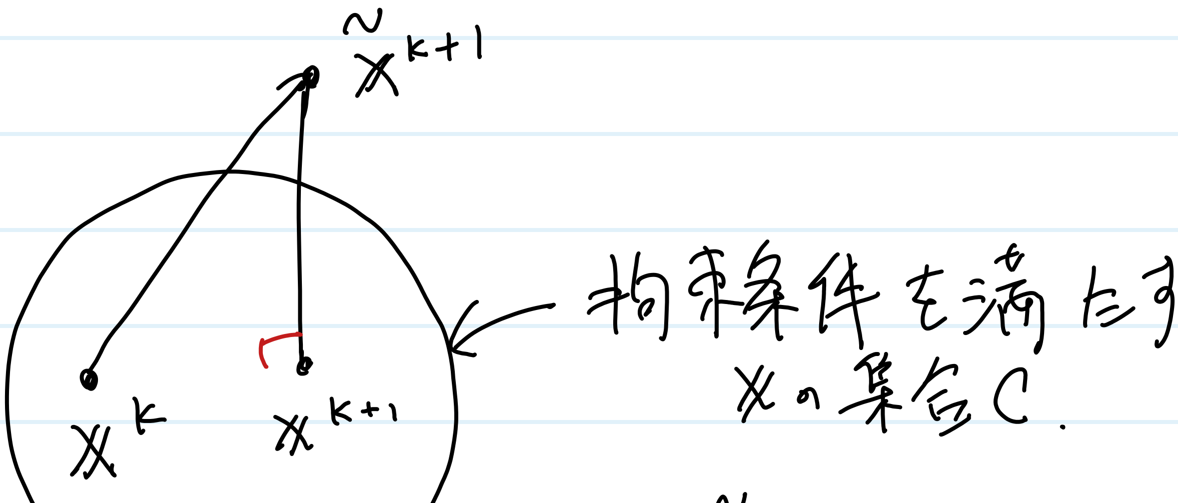
$$\mathbb{x}^{k+1} = \mathbb{x}^k \exp\left(-\alpha \nabla f(\mathbb{x}^k)\right)$$

である.

補足：

ここまで、xには拘束条件はなかった。
でも、xに拘束条件がある場合はどうする？

イテレーションを進めて、xが拘束条件をやぶってしまったら、拘束条件を満たすxの集合の中から、最も近い点に更新する。



$$x^{k+1} = P_C \tilde{x}^{k+1}$$

つまり、$\min_{x \in C} f(x)$ では

$$x^{k+1} \leftarrow P_C \left(x^k - \alpha \nabla f(x^k) \right)$$

普通の勾配法なら P_C はユークリッド射影.

鏡像降下法では, Bregman divergence の意味での射影を考える. (Bregman projection)

例えば, $C = \Delta^d = \{ \mathbf{x} \mid x_d \geq 0, \sum_d x_d = 1 \}$

として, $\phi(\mathbf{x}) = \sum_i x_i \log x_i$ とすると, 制約条件を

満たさない $\tilde{\mathbf{x}}^{t+1}$ から, Δ^d 中の最も近い点は

規格化されていない.

$$
\begin{aligned}
\mathbf{x}^{t+1} &= \underset{\mathbf{x} \in \Delta^d}{\operatorname{argmin}} \; B_\phi(\tilde{\mathbf{x}}^{t+1}, \mathbf{x}) \\
&= \underset{\mathbf{x} \in \Delta^d}{\operatorname{argmin}} \sum_i \tilde{x}_i^{t+1} \log \frac{\tilde{x}_i^{t+1}}{x_i} \\
&\qquad - \sum_i x_i - \sum_i \tilde{x}_i^{t+1}
\end{aligned}
$$

を解くことで求まる.

ラグランジュ関数を用いて $\sum_i x_i = 1$ を課しながら最適な $\boldsymbol{x}^{t+1}$ を求めると、次を得る。

$$x_i^{t+1} = \frac{\tilde{x}_i^{t+1}}{\sum_i \tilde{x}_i^{t+1}}$$

以上から、$\min_{\boldsymbol{x} \in \Delta^n} f(\boldsymbol{x})$ を $\phi(\boldsymbol{x}) = \sum_i x_i \log x_i$ の鏡像降下法で更新するとアルゴリズムは…

repeat

$$\tilde{\boldsymbol{x}}^{t+1} \leftarrow \boldsymbol{x}^t \exp\left(-\alpha \nabla f(\boldsymbol{x}^t)\right)$$

$$\boldsymbol{x}^{t+1} \leftarrow \frac{\tilde{x}_i^{t+1}}{\sum_i \tilde{x}_i^{t+1}}$$