Homework 0

Q1.(d) $F(w) = \sum_{i=1}^{n} (x_i^T w - y_i)^2 + \lambda \sum_{j=1}^{d} w_j^2$

$x_i \in \mathbb{R}^d$, $w \in \mathbb{R}^d$

$x_i^T w = \sum_{j=1}^{d} x_i[j] \cdot w[j]$

$(1 \times d)(d \times 1)$

$\nabla F(w) = \left[\frac{\partial F(w)}{\partial w_1}, \dots, \frac{\partial F(w)}{\partial w_d} \right]^T$

$\frac{\partial F(w)}{\partial w_j} = \sum_{i=1}^{n} 2(x_i^T w - y_i) \cdot (x_{i,j} \cdot w_j)'$

$x_{i,j}$

$\nabla F(w) = \begin{pmatrix} \sum_{i=1}^{n} 2(x_i^T w - y_i) \cdot x_{i,1} + \lambda 2 w_1 \\ \sum_{i=1}^{n} 2(x_i^T w - y_i) \cdot x_{i,d} + \lambda 2 w_d \end{pmatrix} + \lambda 2 \cdot w_j$

$= \sum_{i=1}^{n} 2(x_i^T w - y_i) \vec{x}_i + 2\lambda \vec{w}$

$\nabla_w (x^T w) = x$

(e) $f(x \dots x_n) = \log \sum_{i=1}^{n} \exp(x_i)$

$x \in \mathbb{R}^n$ $\log \exp(a) = a$

(Soft max) $\approx \max(x_1 \dots x_n)$

why? check out Q1(f)

$\nabla f(x)$

$\frac{\partial f(x)}{\partial x_j} = \frac{1}{\sum_{i=1}^{n} \exp(x_i)} \left(\frac{\partial \sum_{i=1}^{n} \exp(x_i)}{\partial x_j} \right)$

$$= \frac{1}{\sum_{i=1}^{n} \exp(x_i)} \frac{\partial \exp(x_j)}{\partial x_j}$$

$$= \left[\frac{\exp(x_j)}{\sum_{i=1}^{n} \exp(x_i)} \right]$$

note that $\frac{\partial e^a}{\partial a} = e^a$

$$\nabla f(x) = \frac{1}{\sum_{i=1}^{n} \exp(x_i)} \left[\exp(x_1), \exp(x_2) \dots \exp(x_n) \right]$$

Note that $\nabla f(x)$ converts any $x \in \mathbb{R}^n$ into a probability distribution
a.k.a. "Softmax transform"

We call it "soft-argmax" function

Example:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}$$

$\boxed{\underset{i}{\text{argmax}}\ x_i = 2}$

we can also represent it as $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$

Soft-argmax($\vec{x}$)

i.e. $\nabla f(\vec{x}) = \begin{pmatrix} 3.35e-4 \\ 0.99954 \\ 1.23e-4 \end{pmatrix}$

Approximation

Why "soft"? this function is differentiable.