Welcome back to Computational Learning Theory (Trinity Term).

Learning Real-valued Functions

Let's start with a simple problem.

$\underline{x}_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$ $\quad (\underline{x}_i, y_i) \sim$ dist over $\mathbb{R}^n \times \mathbb{R}$.

There is a class of functions $\mathcal{G}$, of $f^n : \mathbb{R}^n \to \mathbb{R}$

There exists a $g \in \mathcal{G}$, s.t. $\mathbb{E}[y \mid \underline{x}] = g(\underline{x})$.

$$y = g(\underline{x}) + \xi \quad \text{(noise)}$$

GOAL : Find some $\hat{g} : \mathbb{R}^n \to \mathbb{R}$ s.t. $\mathbb{E}_{x \sim D}\left[(\hat{g}(\underline{x}) - g(\underline{x}))^2 \right] \le \varepsilon$.

$$\mathbb{E}_{(x,y)}\left[(\hat{g}(\underline{x}) - y)^2 \right] = \underbrace{\mathbb{E}\left[(\hat{g}(x) - g(x))^2 \right]}_{\text{want to control.}} + \underbrace{2\,\mathbb{E}\left[(\hat{g}(x) - g(x))(g(x) - y) \right]}_{\text{0 expectation}} + \underbrace{\mathbb{E}\left[(g(x) - y)^2 \right]}_{\substack{\text{doesn't depend on } \hat{g} \\ \text{(noise-level)}}}$$

Linear Regression:

$g(\underline{x}) = \underline{w}^* \cdot \underline{x} \qquad \underline{w}^*, \underline{x} \in \mathbb{R}^n$

$\hat{g}(\underline{x}) = \hat{\underline{w}} \cdot \underline{x} \qquad$ (estimate $\hat{w}$).

Training data : $\{(\underline{x}_i, y_i)\}_{i=1}^m$

$$R(w) = \frac{1}{m} \sum_i (w \cdot x_i - y_i)^2$$

} Least-squares method.

Alg: Find $\hat{w} \in \operatorname{argmin} R(w)$.

Ex : Find a "closed"-form solution for the above problem.

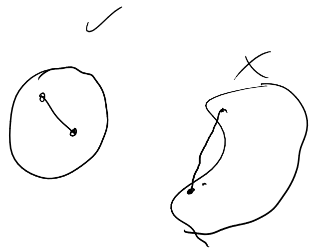
GOAL : $\mathbb{E}\left[(\hat{w} \cdot x - w^* \cdot x)^2 \right] \le \varepsilon$;

How large does m need to be to guarantee that the Least-Squares algorithm works?

Convex Optimization:

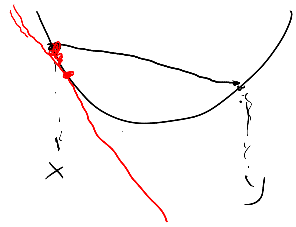
$$R(w) = \frac{1}{m} \sum_i (w \cdot x_i - y_i)^2 \qquad (\text{convex } f^n).$$

Let $K \subseteq \mathbb{R}^n$ which is a convex set,
i.e. $\forall \underline{x}, \underline{y} \in K$, $\forall \lambda \in [0,1]$, $\lambda \underline{x} + (1-\lambda)\underline{y} \in K$.



A function $f : K \to \mathbb{R}$ is convex if
$\forall \underline{x}, \underline{y} \in K$, $\forall \lambda \in [0,1]$, $\quad f(\lambda \underline{x} + (1-\lambda)\underline{y}) \leq \lambda f(\underline{x}) + (1-\lambda) f(\underline{y})$

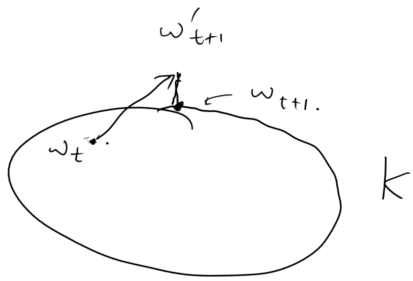
If $f : K \to \mathbb{R}$ convex is differentiable,
then $\forall \underline{x}, \underline{y} \in K$, $\quad f(\underline{y}) \geq f(\underline{x}) + \nabla f(\underline{x}) \cdot (\underline{y} - \underline{x})$



Goal: Find $\hat{w} \in K$, $\quad f : K \to \mathbb{R}$ is convex.

$$\text{s.t. } f(\hat{w}) \leq \min_{w \in K} f(w) + \varepsilon$$

Projected Gradient Descent:

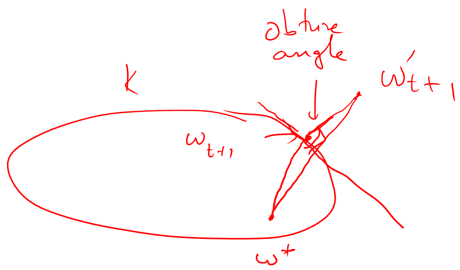


Let $w_1 \in K$,
For $t = 1, \dots, T$

$$w'_{t+1} = w_t - \eta \nabla f(w_t).$$

$$w_{t+1} = \underbrace{\Pi_K (w'_{t+1})}.$$

(i) For any $\underline{w}^* \in K$, $\quad \|w^* - w'_{t+1}\| \geq \|w^* - w_{t+1}\|$ (Projection decreases distance to points in K)



(ii) $\quad f(w_t) + \nabla f(w_t) \cdot (w^* - w_t) \leq f(w^*)$. $\qquad$ (defn of convexity).

$$f(w_t) - f(w^*) \leq \nabla f(w_t) \cdot (w_t - w^*)$$

(Potential-based Proof)

Let $w^* \in K$, s.t. $w^* \in \arg\min f(w)$

$$\boxed{w'_{t+1} = w_t - \eta \nabla f(w_t).}$$

$$\|w'_{t+1} - w^*\|^2 = \|w_t - \eta \nabla f(w_t) - w^*\|^2$$

$$= \|w_t - w^*\|^2 - 2\eta\, \nabla f(w_t)\cdot(w_t - w^*) + \eta^2 \|\nabla f(w_t)\|^2.$$

$$\|w_{t+1} - w^*\|^2 \le \|w_t - w^*\|^2 - 2\eta\, \nabla f(w_t)\cdot(w_t - w^*) + \eta^2 \|\nabla f(w_t)\|^2$$

$$\nabla f(w_t)\cdot(w_t - w^*) \le \frac{\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2}{2\eta} + \frac{\eta}{2}\|\nabla f(w_t)\|^2 \qquad (*)$$

$$f(w_t) - f(w^*) \le \frac{\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2}{2\eta} + \frac{\eta}{2}\|\nabla f(w_t)\|^2$$

$$\frac{1}{T}\sum_{t=1}^{T}\left(f(w_t) - f(w^*)\right) \le \frac{1}{2\eta T}\underbrace{\|w_1 - w^*\|^2}_{\text{how far you start from optimal}} + \frac{1}{T}\cdot\frac{\eta}{2}\underbrace{\sum_{t=1}^{T}\|\nabla f(w_t)\|^2}_{\text{behaviour of } f}.$$

Suppose $\|\nabla f(w)\| \le L \quad \forall w \in K$. ($f$ is L-Lipschitz).

Suppose $\max_{w, w' \in K} \|w - w'\| \le R$ (diameter of K is at most R)

$$\frac{1}{T}\sum_{t=1}^{T}\left(f(w_t) - f(w^*)\right) \le \frac{1}{2\eta T}R^2 + \frac{\eta}{2}L^2 \le \frac{RL}{\sqrt{T}} \quad \text{if } \eta = \frac{R}{L\sqrt{T}}.$$

$$f\left(\frac{1}{T}\sum_t w_t\right) \le \frac{1}{T}\sum_t f(w_t) \quad \text{(because of convexity)}$$

(Polyak-averaging).

Suppose that $\|w^*\|_2 \le W$,

$x, y \in \mathbb{R}^n \times \mathbb{R}$, Let D be the joint distribution, s.t. $\forall (x, y) \in \text{support}(D)$

$\|x\|_2 \le B$, $y \in [-M, M]$.

$$f(w) = \frac{1}{m}\sum_i (w\cdot x_i - y_i)^2,$$

$$\nabla f(w) = \frac{1}{m}\sum_i 2(w\cdot x_i - y_i)\, x_i \qquad |w\cdot x| \le B\cdot W \text{ (Cauchy Schwartz)}$$

$$\|\nabla f(w)\| \le 2\cdot(BW + M)\cdot B.$$

K = unit ball of radius W in $\mathbb{R}^n$.

Generalized Linear Models:

$g : \mathbb{R}^n \to \mathbb{R}$

$g(\underline{x}) = u(\underline{w} \cdot \underline{x})$.

$\underline{w} \in \mathbb{R}^n$, $u : \mathbb{R} \to \mathbb{R}$.

u is monotonically increasing

& u is 1-Lipschitz

$(|u(x) - u(y)| \leq |x - y|)$

$\underline{x}, y$

s.t. $\mathbb{E}[y \mid \underline{x}] = u(\underline{w} \cdot \underline{x})$.

$\left[u \text{ is the sigmoid function. } u(z) = \dfrac{1}{1+e^{-z}} \right]$ (Logistic Regression).

$$\min_{w} \frac{1}{m} \left(\sum_i (u(\underline{w} \cdot \underline{x}_i) - y_i)^2 \right) \quad \text{(Objective function).}$$

(This is not a convex function in $\underline{w}$).

(Surrogate loss functions.)

$$\ell(\underline{w}; \underline{x}, y) = \int_0^{\underline{w} \cdot x} (u(z) - y)\, dz.$$

$$\left[\ell_2(\underline{w}; \underline{x}, y) = (u(\underline{w} \cdot x) - y)^2 \right.$$

$$\nabla_{w} \ell = (u(\underline{w} \cdot x) - y) \cdot \underline{x}$$

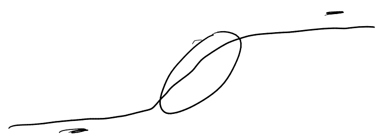
$$\nabla_{w} \ell_2 (w; x, y) = 2(u(\underline{w} \cdot x - y) \boxed{u'(\underline{w} \cdot x)} \cdot \underline{x}$$

"Vanishing gradient"

$$\frac{\partial \ell}{\partial w_i \partial w_j} = u'(\underline{w} \cdot \underline{x}) \cdot x_i x_j$$

$$H(\ell) = \underbrace{u'(\underline{w} \cdot x)}_{\geq 0}\, \underline{x}\, \underline{x}^T \succeq 0 \quad \text{(so convex)}$$

u



$$\mathbb{E}_y\, \ell(w; \underline{x}, y) = \mathbb{E}_y \int_0^{w \cdot x} (u(z) - y)\, dz = \int_0^{w \cdot x} (u(z) - u(\underline{w}^* \cdot \underline{x}))\, dz \qquad \mathbb{E}[y \mid \underline{x}] = u(\underline{w}^* \cdot \underline{x}).$$

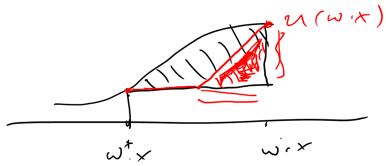
$$\mathbb{E}_y\, \ell(w^*; x, y) = \mathbb{E}_y \int_0^{w^* \cdot x} (u(z) - y)\, dz = \int_0^{w^* \cdot x} (u(z) - u(w^* \cdot x))\, dz$$

$$\mathbb{E}_y (\ell(w; x, y) - \ell(w^*; x, y)) = \int_{w^* \cdot x}^{w \cdot x} (u(z) - u(w^* \cdot x))\, dz.$$

$\underline{w}^* \cdot x < \underline{w} \cdot x$

$$\geq \frac{1}{2} (u(\underline{w} \cdot x) - u(w^* \cdot x))^2.$$

$$\geq 0$$



If you find $\underline{w}$ s.t. $\mathbb{E}_{x,y}[\ell(w; x, y)] \leq \mathbb{E}[\ell(w^*; x, y)] + \varepsilon/2$

then it must be that

$$\mathbb{E}_{x}\left[(u(w\cdot x)-u(w^{*}\cdot x))^{2}\right] \leq \varepsilon .$$