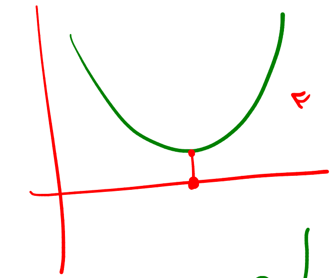
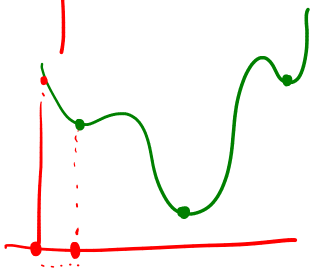
COMPUTATIONAL CONSIDERATIONS

$$w^* = (XX^T)^{+} X y$$

inverse computation is expensive if d is large $O(d^3)$



- We know w^* is the solution of an unconstrained optimization



- We can apply GRADIENT DESCENT.

$$w^{t+1} = w^t - \underbrace{\eta^t}_{\text{Scalar stepsize.}} \underbrace{\nabla f(w^t)}_{\text{Gradient at } w^t}$$

$$f(w) = \|Xw - y\|^2 = \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

$$\nabla f(w) = 2(XX^T)w - 2Xy \quad \text{[verify this]}$$

Gradient descent update for Linear regression

$$w^{t+1} = w^t - \eta^t \left[\underbrace{2(XX^T)w - 2(Xy)} \right]$$

- What if n is large [hundreds of millions], might want to avoid XX^T

- How to adapt gradient descent ?

STOCHASTIC GRADIENT DESCENT

for $t = 1, \dots, T$

- At each step, sample a bunch of datapoints (k) uniformly at random from the set of all points

- PRETEND this sample is the entire dataset and take a gradient step w.r.t it

$$\underbrace{2\left(\tilde{X}\tilde{X}^T w^t - \tilde{X}\tilde{y}\right)}$$

↳ Manageable because $\tilde{X} \in \mathbb{R}^{d \times k}$ ↳ # Sampled points.

end

After T rounds, use

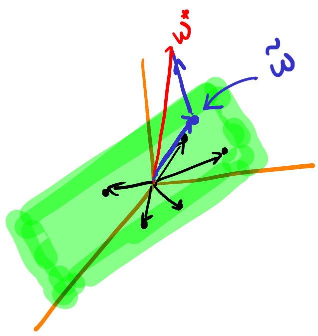
$$\boxed{\underline{w^T_{SGD}} = \frac{1}{T}\sum_{i=1}^{T} w^t}$$

as opposed to w^T as in Standard Gradient descent

Guaranteed to converge to optima with high probability

NON-LINEAR REGRESSION

$$w^* = (XX^T)^{\dagger} Xy$$



- $\underline{w^*}$ must lie in the span of data points.

$$\sum_{i=1}^{n}(\underline{w^{*T} x_i} - y_i)^2 = \sum_{i=1}^{n}(\underline{\tilde{w}^T x_i} - y_i)$$

$$w^* = \tilde{w} + \tilde{w}_{\perp}$$

$$\forall i \quad w^{*T} x_i = (\tilde{w} + \tilde{w}_{\perp})^T x_i = \tilde{w}^T x_i + \underbrace{\tilde{w}_{\perp}^T x_i}_{0}$$

$w^* = \boxed{X\alpha^*}$ for some $\alpha^* \in \mathbb{R}^n$

$= (XX^T)^\dagger X y$

$$w^* = \underset{\min}{\arg} \sum_{i=1}^{n} \left(\phi(x_i) - y_i\right)^2$$

$X\alpha^* = (XX^T)^\dagger X y$

$(XX^T) X\alpha^* = \cancel{(XX^T)}\cancel{(XX^T)}^\dagger X y$

$(XX^T) X\alpha^* = Xy$

$X^T(XX^T) X\alpha^* = X^T X y$

$\underbrace{(X^T X)^2}_{:= K}\alpha^* = (X^T X) y$

$\Rightarrow \quad K^2 \alpha^* = K y$

$\boxed{\alpha^* = K^{-1} y}$

$K^{-1} \in \mathbb{R}^{n \times n}$

KERNEL - REGRESSION

PREDICTION

for some $x_{test} \in \mathbb{R}^d$

$w^{*T} \phi(x_{test})$

$= \left(\sum_{i=1}^{n} \alpha_i^* \phi(x_i)\right)^T \phi(x_{test})$

$= \boxed{\sum_{i=1}^{n} \alpha_i^* K(x_i, x_{test})}$

How important is i^{th} point towards w^*

How similar is x_{test} to x_i

PROBABILISTIC VIEW OF LINEAR REGRESSION

$x \in \mathbb{R}^d \qquad y \in \mathbb{R}$

Dataset

$\{(x_1, y_1) \ldots, (x_n, y_n)\}$

$y/x \sim w^T x + \epsilon$

w: unknown but fixed $\in \mathbb{R}^d$

ϵ: Noise $N(0, \sigma^2)$ — Gaussian.

- Can view this as an "ESTIMATION" problem
- Solution approach - Maximum Likelihood.

Likelihood

$$\mathcal{L}\left(w;\ \begin{matrix} x_1, \dots, x_n \\ y_1, \dots, y_n \end{matrix}\right) = \prod_{i=1}^{n} e^{\frac{-(w^T x_i - y_i)^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi}\,\sigma}$$

$$\log \mathcal{L}\left(w;\ \begin{matrix} x_1, \dots, x_n \\ y_1, \dots, y_n \end{matrix}\right) = \sum_{i=1}^{n} -\frac{(w^T x_i - y_i)^2}{2\sigma^2} \cdot \frac{1}{\sqrt{2\pi}\,\sigma}$$

equivalently

$$\max_{w} \sum_{i=1}^{n} -(w^T x_i - y_i)^2$$

$$\equiv \min_{w} \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

$$\hat{w}_{ML} = w^* = (XX^T)^{\dagger} X y$$

CONCLUSION: Maximum Likelihood estimator assuming ZERO MEAN GAUSSIAN NOISE is same as LINEAR Regression with SQUARED ERROR!

What else have we gained?

- Can study properties of estimators $\hat{w}_{ML}$. !