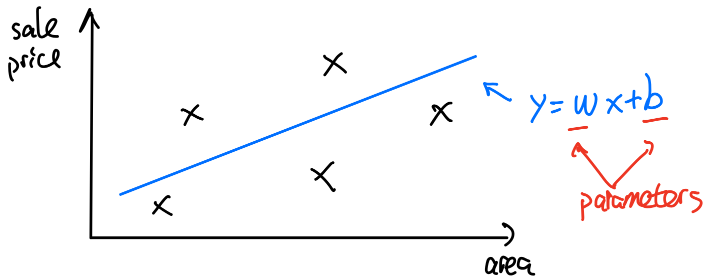
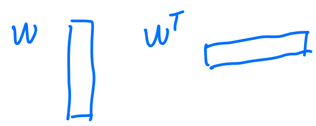
Regression: predicting a real number

$D = \{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\}$, $x^{(i)} \in \mathbb{R}^d$, $y^{(i)} \in \mathbb{R}$

Each dimension in x is called a “feature”, d features in total



General linear predictor: $\hat{y} = \sum_{i=1}^{d} w_i x_i + b = \underbrace{w^T x}_{\text{dot product}} + b$



How to choose w and b?

Loss function

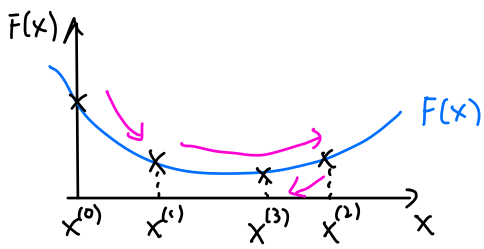
$$L(w,b) = \frac{1}{n} \sum_{i=1}^{n} (\underbrace{w^T x^{(i)} + b}_{\text{prediction}} - \underbrace{y^{(i)}}_{\text{true output “response”}})^2$$

Find (w,b) that minimize L(w,b) (optimization)

Gradient Descent

Function f from $\mathbb{R}^d \to \mathbb{R}$, differentiable

Goal: minimize F



$F(x) \approx F(x^{(0)}) + (x - x^{(0)}) \cdot F'(x^{(0)}) \leftarrow$ tangent line

(best approximation of F(x) around $x^{(0)}$)

$$\nabla_x F(x) = \left[\frac{\partial F}{\partial x_1}, \dots, \frac{\partial F}{\partial x_d} \right] \in \mathbb{R}^d$$

$$F(x) \approx F(x^{(0)}) + (x_1 - x_1^{(0)}) \nabla_x F(x^{(0)})_1$$
$$+ \dots$$
$$+ (x_d - x_d^{(0)}) \nabla_x F(x^{(0)})_d$$
$$= F(x^{(0)}) + (x - x^{(0)})^T \nabla_x F(x^{(0)})$$

Fact: Gradient is the direction of steepest ascent; negative gradient of steepest descent.

Algorithm

$x^{(0)} \leftarrow \vec{0} \in \mathbb{R}^d$

For $t = 1, \cdots, T$

learning rate

$x^{(t)} \leftarrow x^{(t-1)} - \eta \nabla_x F(x^{(t-1)})$

return $x^{(T)}$

Let's start at $u \in \mathbb{R}^d$, and take a step $v \in \mathbb{R}^d$, $\|v\| < \varepsilon$
How to maximally increase F(u+v)?

$$F(u+v) \approx F(u) + v^T \nabla F(u)$$

$$= \underbrace{\|v\|}_{\text{const}} \cdot \underbrace{\|\nabla F(u)\|}_{\text{const}} \cdot \cos(\underbrace{\alpha}_{\substack{\text{largest} = 1 \\ \text{smallest} = -1}}) + F(u)$$

$$L(w) = \frac{1}{n} \sum_{i=1}^{n} (w^T x^{(i)} - y^{(i)})^2$$

$$\nabla L(w) = \frac{1}{n} \sum_{i=1}^{n} 2(w^T x^{(i)} - y^{(i)}) \cdot x^{(i)}$$

Algorithm for linear regression

$w^{(0)} \leftarrow \vec{0} \in \mathbb{R}^d$

for $t = 1, \cdots, T$

$w^{(t)} \leftarrow w^{(t-1)} - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} 2(w^{(t-1)T} x^{(i)} - y^{(i)}) \cdot x^{(i)}$