$$w_{k+1,j} = w_{k,j} - \alpha \boxed{\frac{\partial \ell(w_k)}{\partial w_{k,j}}}$$

$$\ell(w_k) = \frac{1}{2}\sum_{i=0}^{n}\left(y_i - f_{w_k}(x_i)\right)^2$$

$$f_w(x_i) = \sum_{j=1}^{m} w_j x_{ij}$$

$$\frac{\partial}{\partial x} x^2 \to 2x$$

$$\frac{\partial}{\partial x} f(x)^2 \to 2f(x)\frac{\partial f(x)}{\partial x}$$

$$\frac{\partial}{\partial w_{k,j}} \ell(w_k) = \frac{\partial}{\partial w_{k,j}} \frac{1}{2}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2$$

$$= \sum_{i=0}^{n} \frac{\partial}{\partial w_{k,j}} \frac{1}{2}(y_i - \hat{y}_i)^2$$

$$= \sum_{i=0}^{n} \cancel{2}\cdot\frac{1}{\cancel{2}}(y_i - \hat{y}_i)\frac{\partial}{\partial w_{k,j}}(y_i - \hat{y}_i)$$

$$= \sum_{i=0}^{n}(y_i - \hat{y}_i)\frac{\partial}{\partial w_{k,j}}\sum_{\beta=1}^{m} w_{k,\beta} x_{i,\beta}$$

skipping steps!

$$= -\sum_{i=0}^{n}(y_i - \hat{y}_i)\frac{\partial}{\partial w_{k,j}} w_{k,j} x_{i,j}$$

$\frac{\partial}{\partial z} z \cdot \text{const} = \text{const}$

$$= -\sum_{i=0}^{n}(y_i - \hat{y}_i) x_{i,j}$$

What if $x_{i,j}$ is magnitude $\sim 10^{10}$?

$f(x) = x^2$

$\nabla f(x) = 2x$

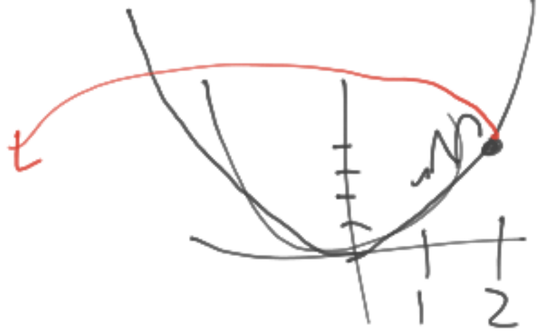
$x_0 = 2$

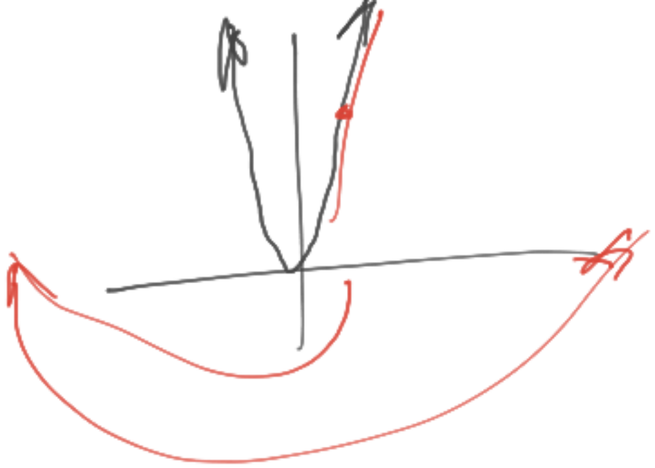
$\alpha = .0001$

$x_1 = 2 - .0001\ 2(2)$

$= 2 - .0004$

$= 1.9996$





$= 2 - 20^7(4)$

massive negative number.

$f(x) = 10^{10} x^2$

$\nabla f(x) = 20^{10} x$

$x_0 = 2$

$\alpha = .0001$

$x_1 = 2 - .0001 \cdot 20^{10}(4)$

$x_{i,1} \sim 10^{10}$ — $\alpha \simeq \frac{1}{10^{10}}$

too big step. $\alpha = \frac{1}{10}$

$x_{i,2} \sim 1$ step size is too small for $w_{i,2}$

$$\frac{\partial l(w_k)}{\partial w_{kj}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)\, x_{ij}$$

Input normalization

$x'_{i,j} \in [-1, 1]$

Before $x_{ij} \in [a, b]$

$$x'_{ij} = 2\frac{x_{ij} - a}{(b-a)} - 1$$

$x'_{ij} \in [0, 1]$

$$x'_{ij} = \frac{x_{ij} - a}{(b-a)}$$

Normalize:

- Mean zero
- Standard deviation 1 (variance 1)

$$X'_{i,j} = X_{i,j} - \frac{1}{n}\sum_{\beta=1}^{n} X_{\beta,j}$$

	X	1	2	3	m=3
1	X_1:	7	14	31	
2	X_2:	-2	-1	16	
3	X_3:	1	2	3	
4	X_4:	7	6	5	
5	X_5:	12	14	3	
	mean				

mean zero.

$X_{5,2}$

$$X'_{i,j} = \frac{X_{i,j}}{\sqrt{\mathrm{Var}(X_{\bullet,j})}}$$

$\sigma^2 = \mathrm{Var}(X_{\bullet,j})$

$$\mathrm{Var}(X'_{\bullet,j}) = \mathrm{Var}\left(\frac{X_{\bullet,j}}{\sqrt{\mathrm{Var}(X_{\bullet,j})}}\right)$$

$$= \frac{1}{\sigma^2}\mathrm{Var}(X_{\bullet,j})$$

$$= \frac{1}{\sigma^2}\sigma^2 = 1.$$

$\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$

$\frac{1}{c^2}\mathrm{Var}\left(\frac{1}{c}\right)$

$\mathrm{Var}\left(\frac{\cdot}{c}\right)$

$$\mathrm{Var}(X_{1,j}, X_{2,j}, \ldots, X_{n,j}) =$$

!

sample $$\frac{1}{n-1} \sum_{i=1}^{n} (X_{i,j} - \bar{X}_j)^2$$

$$\bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{i,j}$$

$$\forall i,j \quad X'_{i,j} = X_{i,j} - \underbrace{\frac{1}{n}\sum_{\beta=1}^{n} X_{\beta,j}}_{\bar{X}_j}$$

For each j

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{i,j}$$

σ_j^2 = variance

$$\sigma_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_{i,j} - \bar{X}\right)^2$$

$$\forall i,j \rightarrow \quad X'_{i,j} = X_{i,j} / \sqrt{\sigma^2}$$

For $j = 1$ to m

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{i,j}$$

$$\sigma_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_{i,j} - \bar{X}_j\right)^2$$

$n-1$ ← Bessel's correction

For $i = 1$ to n

For $j = 1$ to m

$$X_{i,j} \leftarrow \frac{X_{i,j} - \bar{X}_j}{\sqrt{\sigma_j^2}}$$

$$l(w_k) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$\frac{1}{n}$ scale (1)

$\sum$ scale (100,000)

$y_i \in [0, 4]$

$l(w_k) = 10,672$

X $n = 1$

$n = 100,000$

1) Check scaling of y_i.

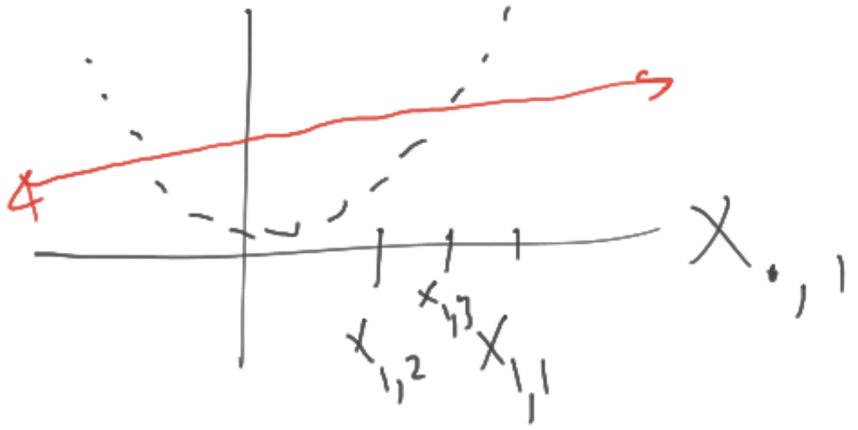
2) Divide by n.

1) Miscaled objectives hard to interpret.

2) Miscaled objectives hard to set α.

Basis functions.

$n = 1$



"linear"

$$f_w(x_i) = \sum_{j=1}^{m} w_j x_{i,j}$$

linear model is linear w.r.t. w.

"Affine"

Linear or Affine means

$\frac{\partial}{\partial w} f_w(x_i) \times$ constant.

Const function of w.

$$\frac{\partial}{\partial w_\beta} f_w(x_i) = \sum_{i=1}^{n} x_{i,\beta}$$

$X, Y \longrightarrow X', Y$

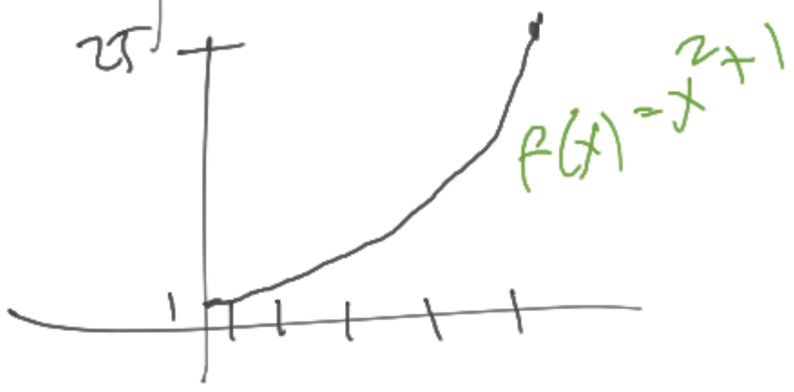
$m = 1$

1	2
2	5
3	10
4	17
5	26

$m' = 2$

$x^2, 1$

1	1	2
4	1	5
9	1	10
16	1	17
25	1	26



$w_1 = 1$

$w_2 = 1$

$f_w(x_i) = w_1 x'_{i,1} + w_2 x'_{i,2}$

$= w_1 x^2 + w_2 1$

$= x^2 + 1$

ϕ "phi" phi-tures features.

$$x_i' = \phi(x_i)$$

↓ vector of features

$$\phi : \mathbb{R}^m \to \mathbb{R}^{m'}$$

↓ Hyperparameter.

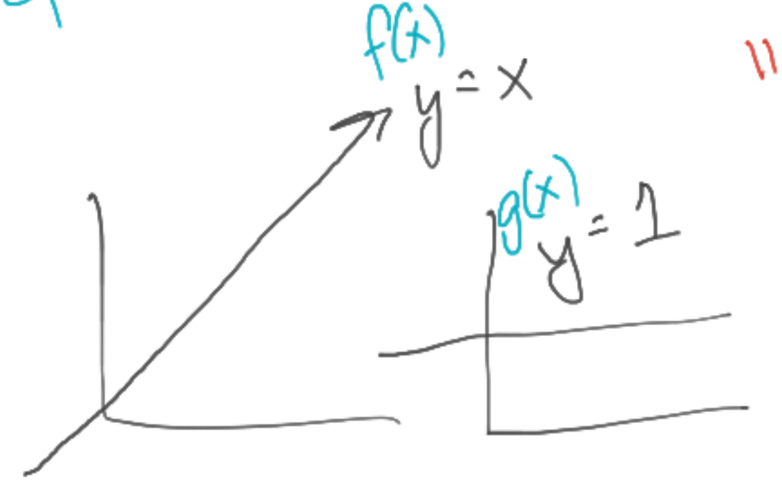
"basis function"

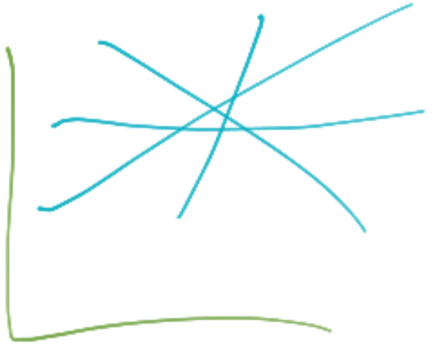
$$f_w(x_i) = \sum_{j=1}^{m} x_{ij} w_j$$

$$f_w(x_i) = \sum_{j=1}^{m'} \phi(x_{ij}) w_j$$

How use features from $\mathbb{R}^2$

$c_1 f(x) + c_2 g(x)$





c_1 [0,1]

+ c_2 [1,0]

↓ Basis for $\mathbb{R}^2$

Polynomial basis. (monomial basis)

$m=3$ $\quad m'$

$x_{i,1} \quad x_{i,2} \quad x_{i,3}$

"order" → Hyperparameter.

↳ $x_{i,1} \quad x_{i,1}^2 \quad x_{i,1}^3 \cdots x_{i,1}^0, x_{i,2}, x_{i,2}^2, \cdots, x_{i,2}^0, \cdots$

(No term with both $x_{i,j}$ and $x_{i,j'}$)

$x_{i,1}^4 \quad x_{i,2}^3$ → Fourier Basis.

→ cos sin.

- Normalize before applying basis.