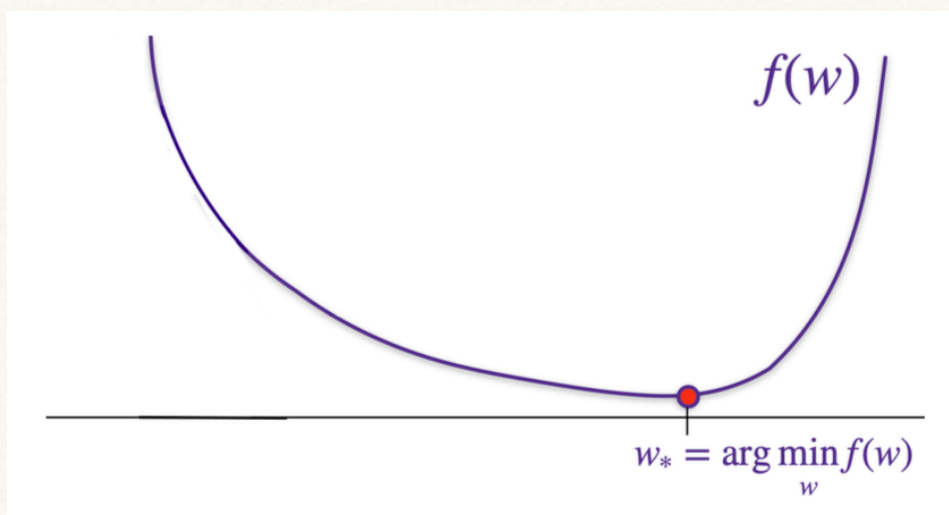
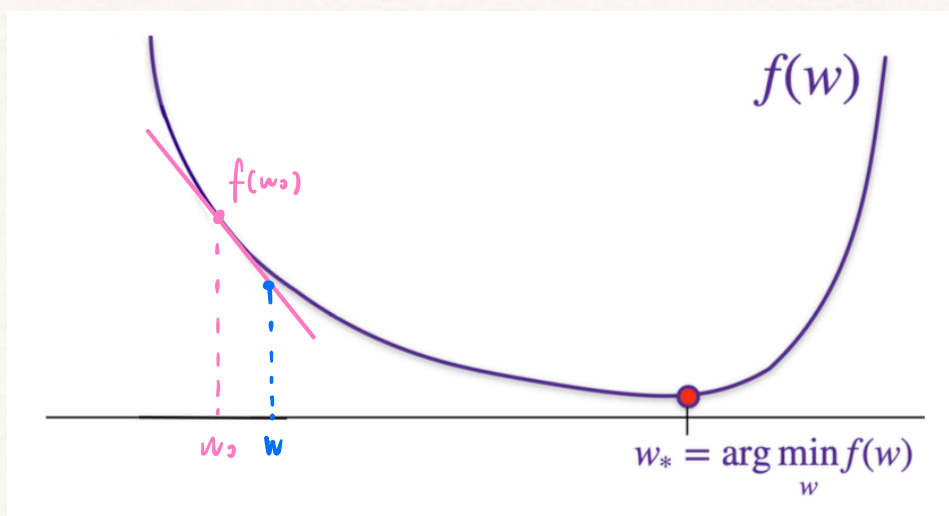


1. Gradient Descent



(a) For some w close to w_0 , give Taylor approx of $f(w)$ at $f(w_0)$.



$$f(w) \approx f(w_0) + (w - w_0) f'(w_0)$$

(b) Let $\eta > 0$ be small. Assume $w_1 = w_0 - \eta f'(w_0)$.

Use part (a), express $f(w_1)$. *nudge in gradient direction*

$$\begin{aligned} f(w_1) &\approx f(w_0 - \eta f'(w_0)) \\ &= f(w_0) + (w_0 - \eta f'(w_0) - w_0) f'(w_0) \\ &= f(w_0) - \eta f'(w_0)^2 \end{aligned}$$

(c) If η is small enough and Taylor approx is good,
we are guaranteed to move in direction that decreases $f(w)$.

$$\text{Because } \eta > 0, f'(w_0)^2 \geq 0$$

$$\Rightarrow \eta f'(w_0)^2 \geq 0$$

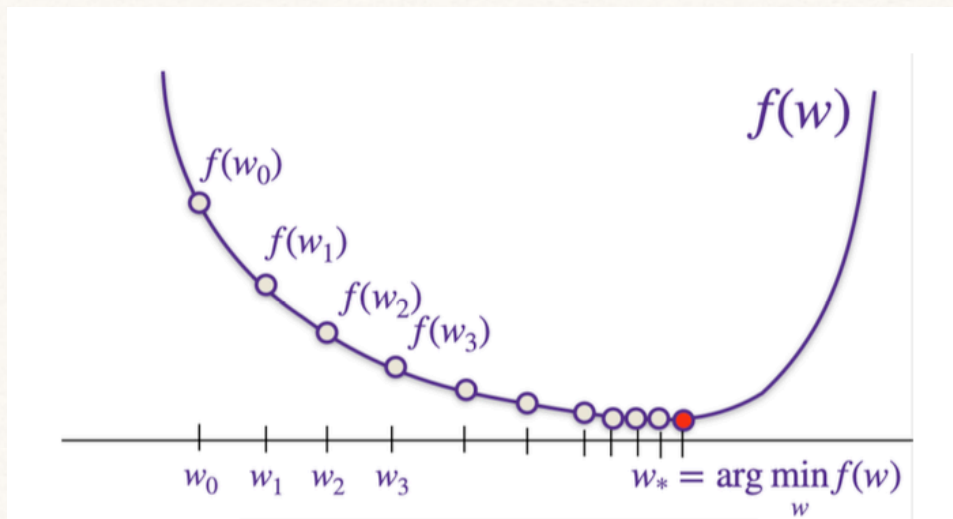
$$\Rightarrow f(w_1) \leq f(w_0)$$

(d) Write general form of gradient descent.

$$\text{For } k = 1, 2, 3, \dots$$

$$w(k+1) = w_k - \eta f'(w_k)$$

$$\text{As } k \rightarrow \infty, f'(w_k) \rightarrow 0.$$



2. Generalized Least Squares

$$\hat{w} = \underset{w}{\operatorname{argmin}} \sum_{i=1}^n (y_i - x_i^T w)^2 + \lambda \sum_{i=1}^d D_{ii} w_i^2, \quad D_{ii} > 0$$

different weights
↓

$$= \underset{w}{\operatorname{argmin}} \|Xw - Y\|_2^2 + \lambda w^T D w$$

2.1 Find closed form for $\hat{w}$.

$$\begin{aligned}\hat{w} &= \underset{w}{\operatorname{argmin}} \|Xw - y\|_2^2 + \lambda w^T D w \\ &= \underset{w}{\operatorname{argmin}} (Xw - y)^T (Xw - y) + \lambda w^T D w\end{aligned}$$

Take gradient and set to 0:

$$2 X^T (Xw - y) + \lambda \underbrace{(D^T + D)}_{= 2D} w = 0$$

$$2 X^T (Xw - y) + 2\lambda D w = 0$$

$$X^T (Xw - y) + \lambda D w = 0$$

$$X^T X w - X^T y + \lambda D w = 0$$

$$(X^T X + \lambda D) w = X^T y$$

$$\hat{w} = \underline{(X^T X + \lambda D)^{-1} X^T y}$$

invertible if $X^T X + \lambda D$ is full rank.

$$X^T X = V D_X V^{-1} \quad D_X \succeq 0, V \text{ orthogonal}$$

$$\lambda D = \lambda V D V^{-1}$$

$$X^T X + \lambda D = V \underbrace{(D_X + \lambda D)}_{\succeq 0 \text{ if } \lambda > 0} V^{-1} \rightarrow \text{full rank}$$

2.2 (a) what happens if we double all values of y_i ?

This doubles $\hat{w}_i$ too.

(b) what happens if we double x_{ij} ?

This halves $\hat{w}_i$.

(c) Suppose $D = I \rightarrow$ ridge regression. Why always invertible?

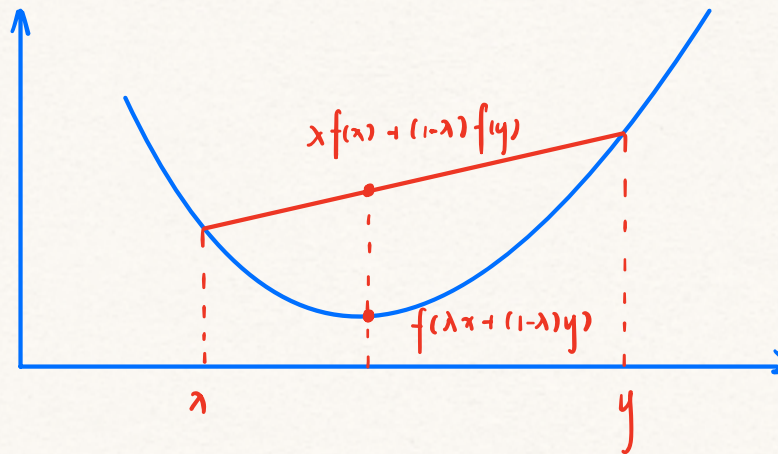
we've show for any $\lambda > 0$, $D_{ii} > 0$.

$(X^T X + \lambda D)$ is invertible.

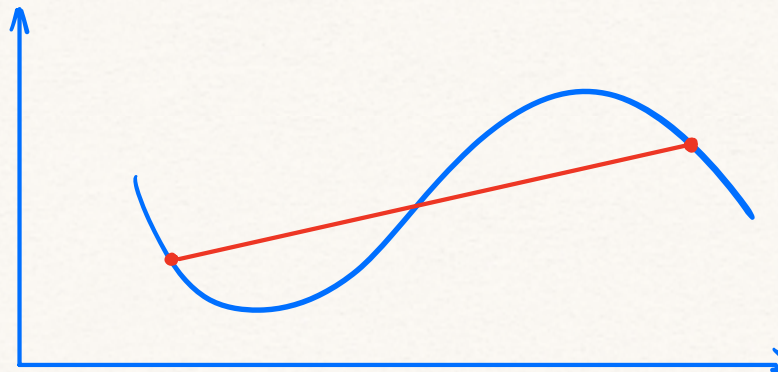
3. Convexity

Def: $f: \mathbb{R}^d \rightarrow \mathbb{R}$ is convex on a set A if
for all $x, y \in A$ and $\lambda \in [0, 1]$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$



convex



not convex

(a) Why do we care if f is convex?

Because local minima = global minima

Proof:

Let x^* be a local minimum

Suppose we have $f(x_0) < f(x^*)$

Since x^* is a local minimum, we can find $\lambda \in [0, 1]$ s.t

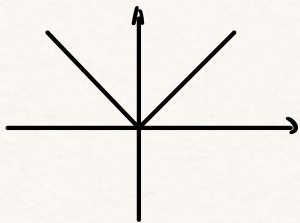
$$y = \lambda x^* + (1-\lambda)x_0, \quad f(y) > f(x^*)$$

But by convexity,

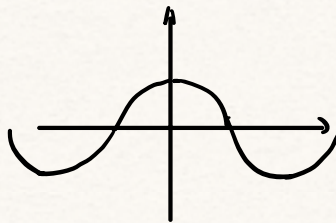
$$\underbrace{f(\lambda x_0 + (1-\lambda)x^*)}_{f(y)} \leq \lambda f(x_0) + (1-\lambda)f(x^*) < f(x^*)$$

Therefore, we have $f(y) < f(x^*) < f(y)$ $\nexists$

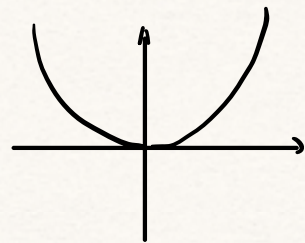
(b) Which functions are convex?



(1) ✓ $|x|$



(2) $\cos(x)$



(3) ✓ x^7

Proof of $|x|$:

$$\begin{aligned} f(\lambda|x| + (1-\lambda)|y|) &= |\lambda x + (1-\lambda)y| \\ &\leq \lambda|x| + (1-\lambda)|y| \end{aligned}$$

(c) Can a function be both convex and concave?

Yes, linear functions.

4. Checking convexity

- Differentiable functions : check $f(y) \geq f(x) + \nabla f(x)^T (y - x)$
- Twice differentiable functions : check $\nabla^2 f(x) \succeq 0$

(a) Prove Jensen's inequality: $E f(x) \geq f(E x)$ when f convex.

Let $\mu = E[x]$. Since f is convex,

$$f(x) \geq f(\mu) + \nabla f(\mu)^T (x - \mu)$$

with prob 1. Taking expectation of both sides

$$\begin{aligned} E[f(x)] &\geq E[f(\mu)] + \nabla f(\mu)^T \underbrace{E[x - \mu]}_{= 0} \\ &= E[f(\mu)] \end{aligned}$$

(b) Show that objective for linear regression is convex.

$$\begin{aligned} L(w) &= \|Xw - y\|_2^2 \\ &= (Xw - y)^T (Xw - y) \\ \nabla L(w) &= 2X^T (Xw - y) \\ \nabla^2 L(w) &= 2X^T X \succeq 0 \end{aligned}$$

(c) - (f) : exercise

5. MAP as regularization

Maximum Likelihood Estimation:

$$\max_{\theta} P(\text{data} | \theta)$$

Maximum A Posteriori Estimation:

$$\max_{\theta} P(\underbrace{\theta}_{\text{posterior}} | \text{data}) = \frac{P(\text{data} | \theta) \underbrace{P(\theta)}_{\text{prior}}}{P(\text{data})}$$

$$\hat{w}_{\text{MAP}} = \underset{w}{\operatorname{argmax}} P(w | X, Y)$$

$$= \underset{w}{\operatorname{argmax}} \frac{P(X, Y | w) P(w)}{P(X, Y)}$$

$$= \underset{w}{\operatorname{argmax}} \log P(X, Y | w) + \log P(w) - \log P(X, Y) \quad \text{const}$$

(a) Suppose we have Laplacian prior

$$P(w_i) = \frac{\lambda}{4b^2} \exp\left(-|w_i| \frac{\lambda}{2b^2}\right)$$

What's the MAP objective?

$$\underset{w}{\operatorname{argmax}} P(w | X, Y) = \underset{w}{\operatorname{argmax}} \log P(X, Y | w) + \log P(w)$$

$$= \underset{w}{\operatorname{argmax}} \log \left\{ (2\pi b^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2} (Y - Xw)^T (b^2 I)^{-1} (Y - Xw)\right) \right\} \\ + \sum_{i=1}^n \log \frac{\lambda}{4b^2} \exp\left(-|w_i| \frac{\lambda}{2b^2}\right)$$

$$= \underset{w}{\operatorname{argmax}} -\frac{n}{2} \log(2\pi b^2) - \frac{1}{2b^2} (Y - Xw)^T (Y - Xw) \\ + n \log \frac{\lambda}{4b^2} - \frac{\lambda}{2b^2} \sum_{i=1}^n |w_i|$$

$$= \underset{w}{\operatorname{argmax}} -\frac{1}{2b^2} \left((Y - Xw)^T (Y - Xw) + \lambda \sum_{i=1}^n |w_i| \right)$$

$$= \underset{w}{\operatorname{argmax}} -\|Y - Xw\|_2^2 - \lambda \|w\|_1$$

$$= \underset{w}{\operatorname{argmin}} \|Y - Xw\|_2^2 + \lambda \|w\|_1 \rightarrow \text{LASSO}$$

(b) Show RIDGE regression = Gaussian prior. (exercise)