ECE-GY 6143
Lecture 3

- [x] Recap: Linear regression
- [x] Gradient descent
- [ ] Application: linear regression (again).

Recap: Data (x_i, y_i) $i = 1, \dots, n$.

- fit a linear model

$$y = w_0 + w_1 x.$$

- Loss function

$$MSE(w) = \frac{1}{n} \sum_{i=1}^{n} \left[y_i - (w_0 + w_1 x_i) \right]^2$$

-

$$w_0^* = \bar{y} - w_1^* \bar{x}$$

$$w_1^* = \frac{\sum_i x_i y_i - \bar{x}^2}{\sum x_i^2 - \bar{x}^2}$$

← Univariate case

Multivariate

$$x = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \end{bmatrix}, \quad y$$

$$\begin{bmatrix} \vdots \\ x^{(d)} \end{bmatrix}$$

Linear model

$$y = w_0 + w_1 x^{(1)} + w_2 x^{(2)} + \cdots + w_d x^{(d)}$$

w_0: intercept

$w_1, \dots, w_d$: coefficients.

Assumption 1: Data & labels are zero mean.

$\Rightarrow$ Ignore w_0.

$$y = \sum_{i=1}^{d} w_i x_i \qquad w \cdot x$$

Step 1

$$y = \langle w, x \rangle$$

Step 2

$$MSE = \frac{1}{2} \sum_{i=1}^{n} (y_i - \langle w, x_i \rangle)^2$$

Step 3 Minimize MSE with respect to w.

$$\begin{bmatrix} \langle w, x_1 \rangle \\ \langle w, x_2 \rangle \\ \vdots \end{bmatrix} = Xw = \begin{bmatrix} - x_1 - \\ - x_2 - \\ \vdots \\ - x_n - \end{bmatrix} w \qquad X = \begin{bmatrix} - x_1 - \\ - x_2 - \\ \vdots \\ - x_n - \end{bmatrix}$$

$$\langle w, x_n \rangle$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$MSE = \frac{1}{2} \| y - Xw \|_2^2$$

Minimize with respect to w.

$$w^* = \left(\sum_{i=1}^{n} x_i x_i^T \right)^{-1} \left[\sum_{i=1}^{n} y_i x_i \right]$$

$$\boxed{w^* = \left(X^T X \right)^{-1} X^T y}$$

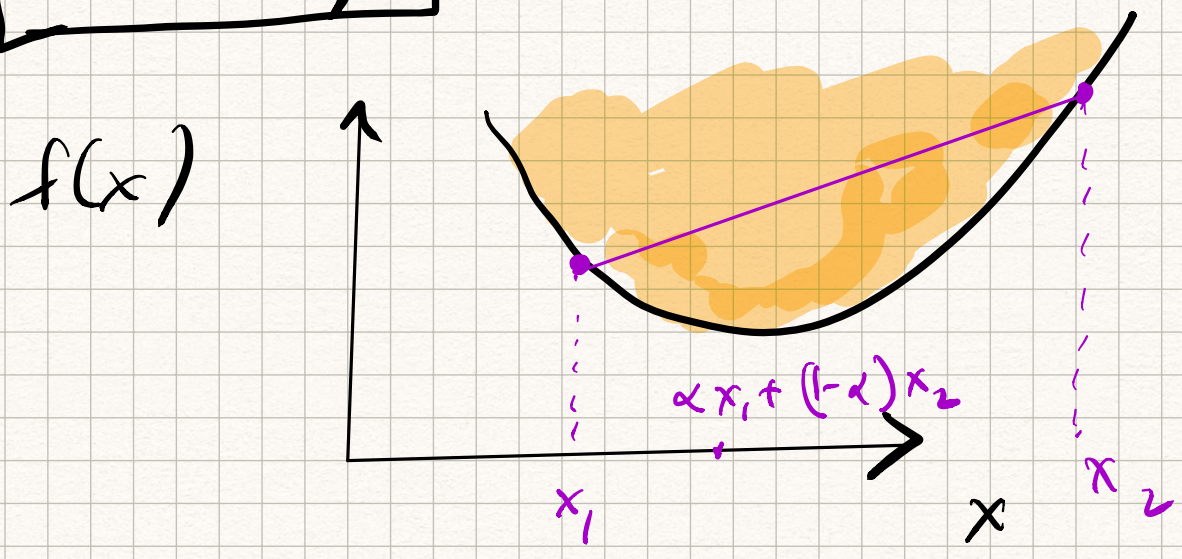
Multivariate case.

$d \times 1$ $\quad$ $d \times d$ $\quad$ $d \times n$ $\quad$ $n \times 1$

Bad! Inversion takes $O(d^3)$ time.

Goal: Improve this running time sufficiently. Using gradient descent.

Convexity



Words: Straight line b/w $(x_1, f(x_1))$ & $(x_2, f(x_2))$ should lie "above" $f(x)$

Math: $f: \mathbb{R}^d \to \mathbb{R}$. For any $x, y \in \mathbb{R}^d$

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

for all α in $[0,1]$.

then f is **convex**.

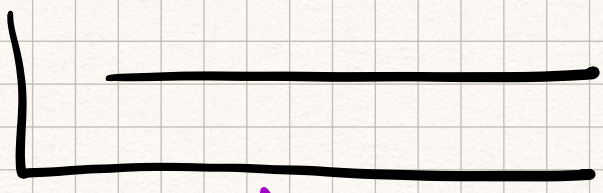
Opposite of convex: "concave"

↑

Sinusoid:

Neither concave nor convex.

Constant:

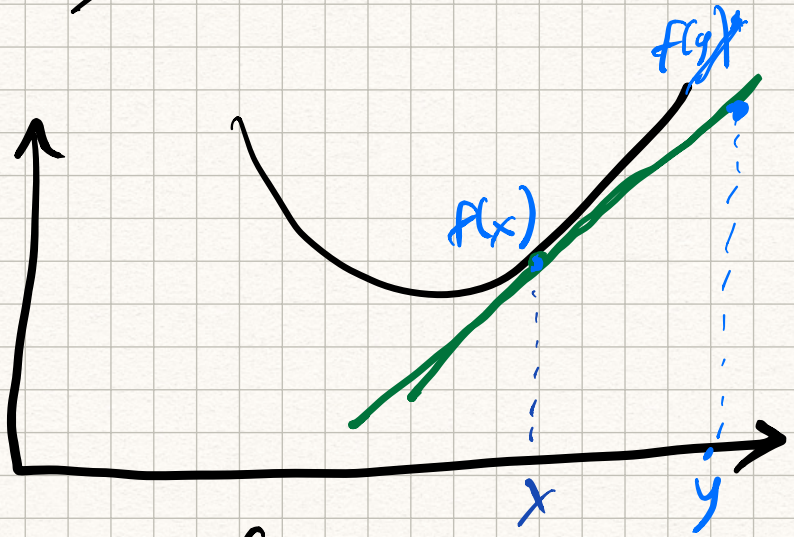


Both concave and convex

Straight

Both concave/convex.

Property



Words

Tangent of f at any point x lies "below" the curve.

Math

For any x, y,

Scalar $f(y) \geq f(x) + \frac{df(x)}{dx}(y-x)$.

Vectors

$$\begin{bmatrix} \frac{\partial f}{\partial x^{(1)}} \\ \frac{\partial f}{\partial x^{(2)}} \\ \vdots \\ \frac{\partial f}{\partial x^{(d)}} \end{bmatrix} = \nabla f(x).$$

For all x, y.

$$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle.$$

Observation

If $\nabla f(x) = 0$, then

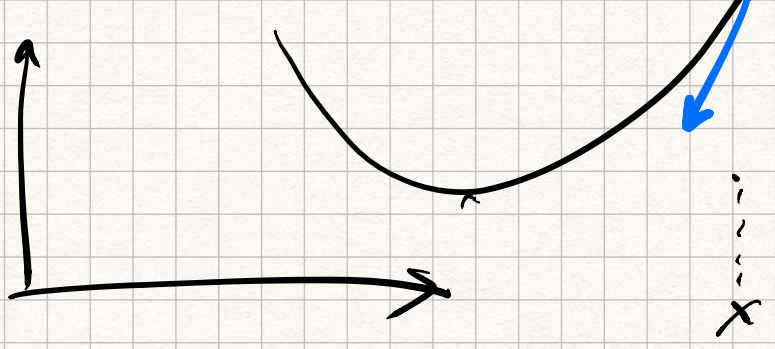
$f(y) \geq f(x)$ for all y

i.e. x is a global minimum of f.

Caveat: Above discussion assumes differentiability / smoothness

Gradient descent [GD]

$\nabla f(x)$



Goal: find x^* that minimizes $f(x)$.

Approach
Find Δ $\quad f(w+\Delta) < f(w)$.

- $w \leftarrow w + \Delta$
- Repeat

Principle of steepest descent:

$\nabla f(x)$ points in the direction of maximum change in f.

Hence

$$w_{k+1} = w_k - \alpha_k \nabla f(w_k)$$

for $k = 0, 1, 2, \ldots, T$

"Step size" / "Learning rate"

"Number of epochs"

Terminate after T epochs

or when ∇f becomes small.

Applications of GD: Linear regression

$$L(w) = \frac{1}{2} \|y - Xw\|_2^2$$

GD:
- Initialize w_0.
- Iterate.

$\nabla L(w) = -X^T(y - Xw)$

$$w_{k+1} \leftarrow w_k - \alpha_k X^T(y - Xw_k)$$

- Until T or ∇ becomes small.

Running time: $O(nd) \cdot T$

Need to bound T, set step size.

- Set error parameter ε, terminate when $\|\nabla F(x)\|_2 \leq \varepsilon$
- Can prove that this happens after $T \geq \log\left(\frac{\|w_0\|_2}{\varepsilon}\right)$

$$\log \frac{1}{p}\left[\frac{n_0 n_2}{\Sigma}\right]$$

where $p = \dfrac{1 - \frac{\ell}{L}}{1 + \frac{\ell}{L}}$

ℓ, L smallest & largest eigenvalues of $X^T X$.

Eigenvalue: $A v = \lambda v$

v ↑ eigenvector

λ ↑ eigenvalue