COMP5212: Machine Learning

Review

Minhao CHENG

Final Project

Presentation

- 6-7 mins per group (background, methodology, initial result)
  - By group number
- 4 pages long report (addition content in appendix)
  - Due on Dec 15

Final exam

- 90 mins
- Close-book

Review

Matrix derivate

- Chain rule: f is a function of Y, let Y=AXB, to get $\frac{\partial f}{\partial X}$
  - $df = tr(\frac{\partial f}{\partial Y}^T dY) = tr(\frac{\partial f}{\partial Y}^T AdXB) = tr(B\frac{\partial f}{\partial Y}^T AdX) = tr((A^T\frac{\partial f}{\partial Y}B^T)^T dX)$
    - Since $dY = d(A)XB + AdXB + AXdB = AdXB$ as $dA = 0, dB = 0$
  - So we get $\frac{\partial f}{\partial X} = A^T\frac{\partial f}{\partial Y}B^T$

Review

Matrix derivate

- Ex 1: $f = a^T X b$, solve $\dfrac{\partial f}{\partial X}$, where a is $m \times 1$ vector, X is $m \times n$ matrix, b is $n \times 1$ vector
- Ex 2: $f = a^T exp(Xb)$, solve $\dfrac{\partial f}{\partial X}$, where a is $m \times 1$ vector, X is $m \times n$ matrix, b is $n \times 1$ vector
- Ex 3: $f = \|Xw - y\|^2$, solve $\dfrac{\partial f}{\partial w}$, where y is $m \times 1$ vector, X is $m \times n$ matrix, w is $n \times 1$ vector
  -

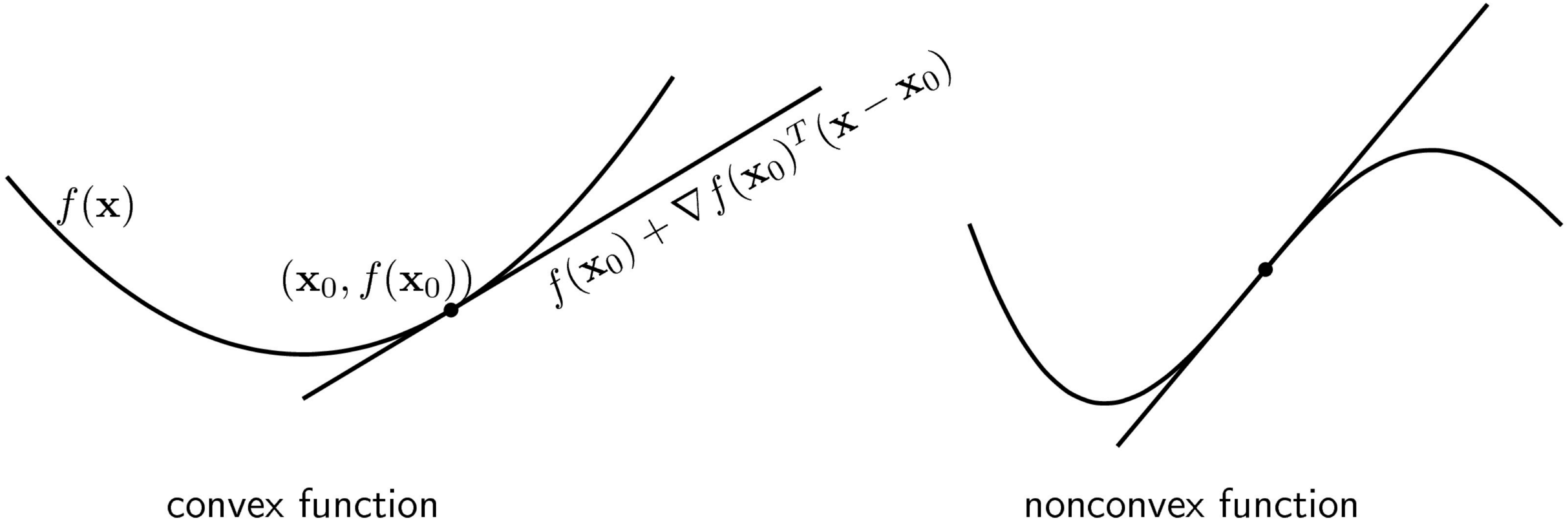
Review
Convexity

- A function $f: \mathbb{R}^n \to \mathbb{R}$ is a convex function
- $\Leftrightarrow$ the function f is below any line segment between two points on f:
    - $\forall x_1, x_2, \forall t \in [0,1]$,
    - $f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$

Review

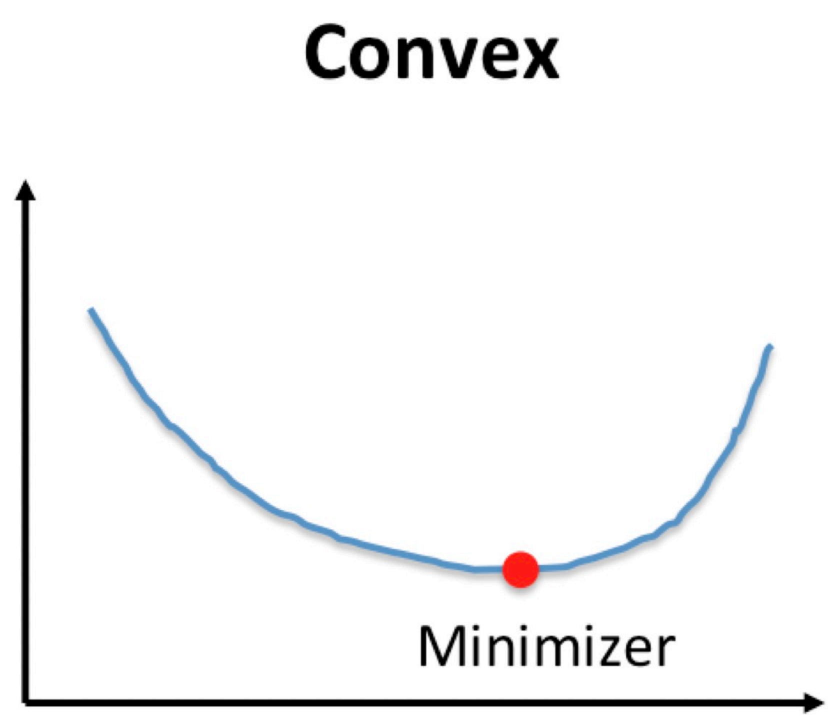
Convexity

- Another equivalent definition for differentiable function:
  - f is convex if and only if $f(x) \geq f(x_0) + \nabla f(x_0)^T(x - x_0), \forall x, x_0$



Review

Convexity



- Convex function:
  - (For differentiable function) $\nabla f(w^*) = 0 \Leftrightarrow w^*$ is a global minimum
  - If f is twice differentiable $\Rightarrow$
    - F is convex if and only if $\nabla^2 f(w)$ is **positive semi-definite**
    - Example: linear regression, logistic regression, …

Review

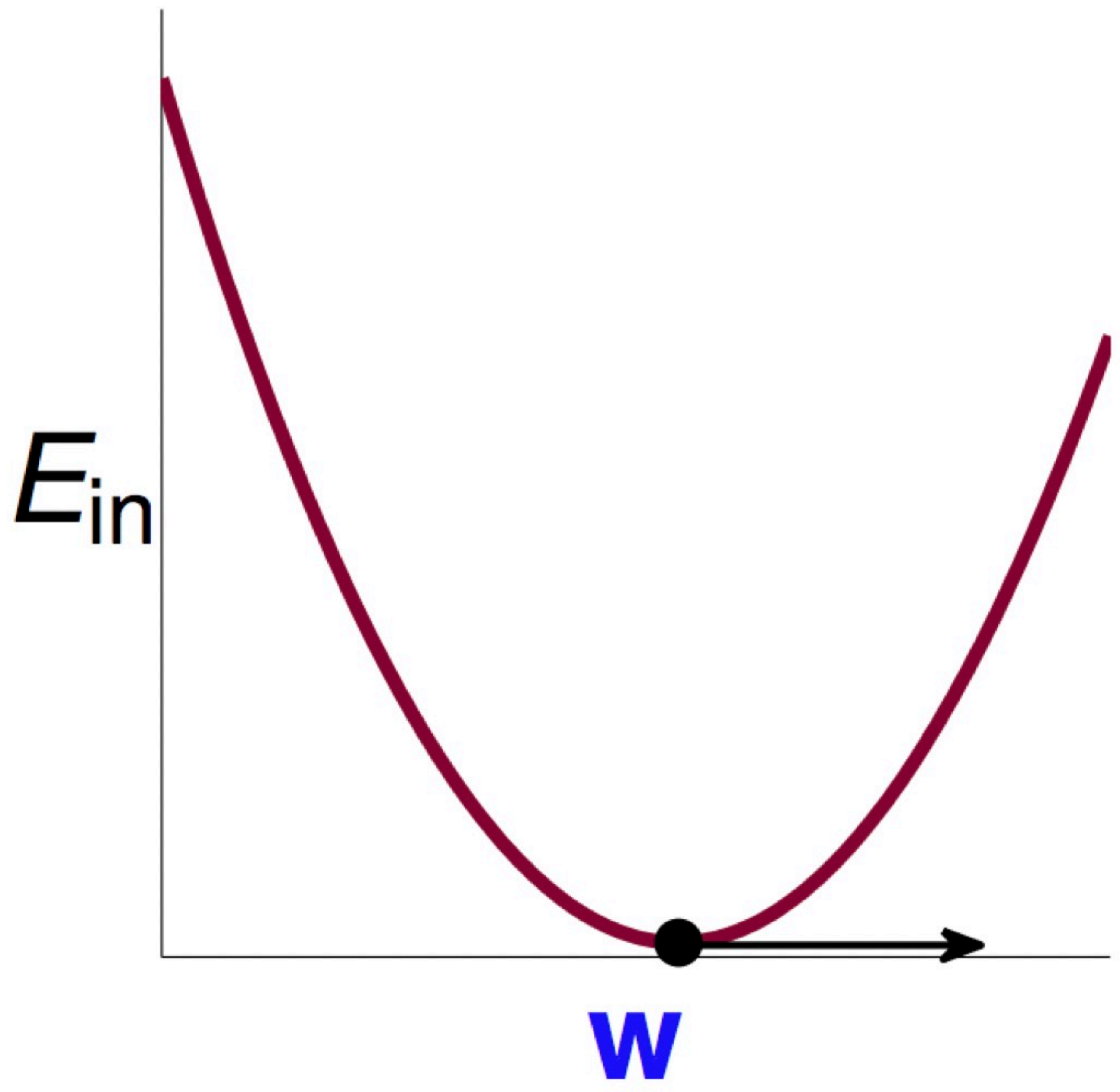
Lipchitz continuous/smooth

- A differential function f is said to be L-Lipschitz continuous:
  - $\|f(x_1) - f(x_2)\|_2 \leq L\|x_1 - x_2\|_2$
- A differential function f is said to be L-smooth: its gradient are Lipschitz continuous:
  - $\|\nabla f(x_1) - \nabla f(x_2)\|_2 \leq L\|x_1 - x_2\|_2$
  - And we could get
    - $\nabla^2 f(x) \preceq LI$
    - $f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{1}{2}L\|y - x\|^2$

Review

Linear regression

- $min_w f(w) = \|Xw - y\|^2$
  - E_{train}: continuous, differentiable, convex
  - Necessary condition of optimal w:
  - $$\nabla f(w^*) = \begin{bmatrix} \frac{\partial f}{\partial w_0}(w^*) \\ \vdots \\ \frac{\partial f}{\partial w_d}(w^*) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$



Review

Linear regression

$$f(w) = \|Xw - y\|^2 = w^T X^T X w - 2 w^T X^T y + y^T y$$

$$\nabla f(w) = 2(X^T X w - X^T y)$$

- $$\nabla f(w^*) = 0 \Rightarrow \underbrace{X^T X w^* = X^T y}_{\text{normal equation}}$$

- $\Rightarrow w^* = (X^T X)^{-1} X^T y$

Review

Optimization

- Gradient descent
- Stochastic gradient descent
- Adagrad
- Momentum
- Adam

Review

Nonlinear mapping

- Can now freely do quadratic classification, quadratic regression
- Can easily extend to any degree of polynomial mappings
    - E.g.,
      $\phi(x) = (x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1x_2^2, x_1x_3^2, x_1x_2^2, x_2^2x_3, x_2^2x_3, x_1^3, x_2^3, x_3^3)$

Review

Generalization bound

- $$P[\neg\exists h \in \mathscr{H} \,|E_{tr}(h) - E(h)| > \epsilon] = P[\forall h \in \mathscr{H} \,|E_{tr}(h) - E(h)| \leq \epsilon]$$
  $$\geq 1 - 2|\mathscr{H}|e^{-2\epsilon^2 N}$$
- Given N and some δ, we have
  - $$|E_{tr}(h) - E(h)| \leq \sqrt{\frac{1}{2N}\log\frac{2|\mathscr{H}|}{\delta}}$$
  - i.e $|E_{tr}(h) - E(h)| \leq \gamma$ for all $h \in \mathscr{H}$

Review

VC Dimension

- Given a set $S = \{x^{(i)}, \ldots, x^{(d)}\}$ (no relation to the training set) of points $x^{(i)} \in \mathscr{X}$, we say that $\mathscr{H}$ shatters S if $\mathscr{H}$ can realize any labeling on S. I.e, if for any set of labels $\{y^{(i)}, \ldots, y^{(d)}\}$, there exist some $h \in \mathscr{H}$ so that $h(x^{(i)}) = y^{(i)}$ for all $i = 1,\ldots, d$
- If no data set of size k can be shattered by $\mathscr{H}$, then k is a break point for $\mathscr{H}$
  - $m_{\mathscr{H}}(k) < 2^k$
- VC dimension for linear model

Review

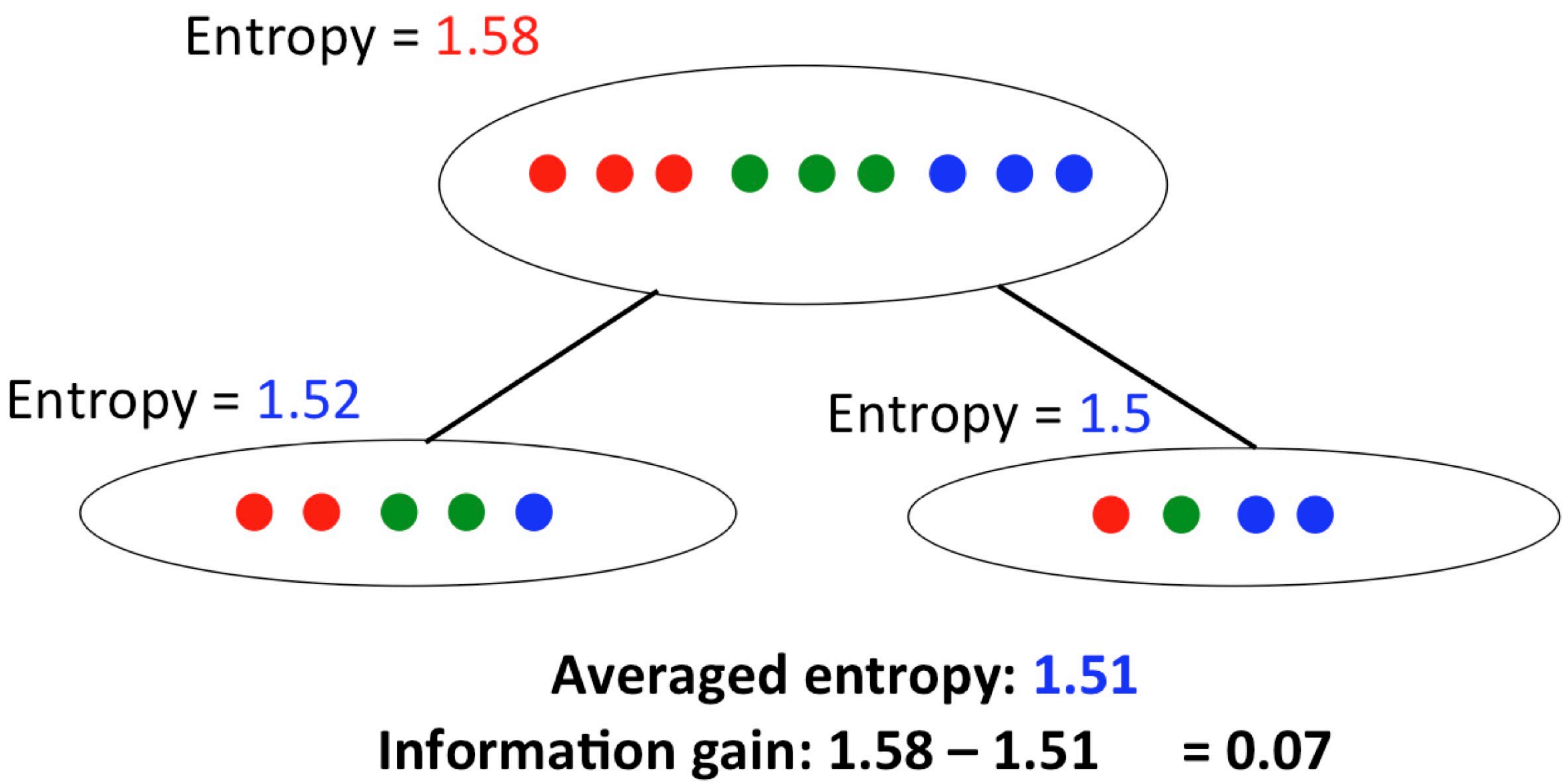
Regularization

- Calling the regularizer $\Omega = \Omega(h)$, we minimize
  - $E_{\text{reg}}(h) = E_{\text{tr}}(h) + \frac{\lambda}{N}\Omega(h)$
- In general, $\Omega(h)$ can be any measurement for the "size" of h

Review

Decision Tree

- The averaged entropy of a split $S \rightarrow S_1, S_2$
  - $\frac{|S_1|}{|S|}H(S_1) + \frac{|S_2|}{|S|}H(S_2)$
- Information gain: measure how good is the split
  - $H(S) - ((|S_1|/|S|)H(S_1) + (|S_2|/|S|)H(S_2))$



Review

Model ensemble

- Bagging
  - Random Forest (Bootstrap ensemble for decision trees):
    - Create T trees
    - Learn each tree using a subsampled dataset S_i and subsampled feature set D_i
    - Prediction: Average the results from all the T trees
- Boosting
  - Direct loss minimization: at each stage m, find the best function to minimize loss
    - Solve $f_m = \arg\min_{f_m} \sum_{i=1}^{N} \ell(y_i, F_{m-1}(x_i) + f_m(x_i))$
    - Update $F_m \leftarrow F_{m-1} + f_m$

Neural networks

- Forward/ backward propagation
- Activation function
- Convolution neural networks: kernel, stride, padding, pooling
- Overfitting
- Gradient vanish/exploding

Exam

- 12-Dec-2023 12:30PM - 02:30PM
  - Lecture Theater D
- SFQ before Nov 30