CS5275 The Algorithm Designer's Toolkit

Final Examination Helpsheet

AY2024/25 Semester 2 · Prepared by Tian Xiao *@snoidetx*

1 Concentration Inequalities

Markov's inequality: Z non-negative: $\Pr[Z \geq t] \leq \frac{\mathbb{E}[Z]}{t}$.

- Z any, ϕ non-decreasing & non-negative:
$$\Pr[Z \geq t] \leq \Pr[\phi(Z) \geq \phi(t)] \leq \frac{\mathbb{E}[\phi(Z)]}{\phi(t)}.$$

Chebyshev's inequality: $\Pr[|Z - \mathbb{E}[Z]| \geq t] \leq \frac{\mathrm{Var}[Z]}{t^2}$.

Chernoff bound: $\Pr[Z \geq t] \leq e^{-\lambda t}\mathbb{E}[e^{\lambda Z}]$.

- Cramér-Chernoff inequality: $\Pr[Z \geq t] \leq e^{-\psi_Z^*(t)}$, where
  - ▷ $\psi_Z^*(t) = \sup_{\lambda \geq 0}(\lambda t - \psi_Z(\lambda))$;
  - ▷ $\psi_Z(\lambda) = \log \mathbb{E}[e^{\lambda Z}]$ for $\lambda \geq 0$.
- Sum of independent rv: $Z = X_1 + \cdots + X_n$:
$$\Pr\left[\frac{1}{n}|Z - \mathbb{E}[Z]| \geq \epsilon\right] \leq \frac{\mathrm{Var}[X]}{n\epsilon^2};$$
$$\Pr[Z \geq n\epsilon] \leq e^{-n\psi_X^*(\epsilon)}.$$
  - ▷ Gaussian $X \sim \mathcal{N}(0, \sigma^2)$: $\Pr[|Z| \geq n\epsilon] \leq 2e^{-\frac{n\epsilon^2}{2\sigma^2}}$.

Sub-Gaussian: A zero-mean rv X is *sub-Gaussian with parameter* σ^2 (i.e., $\in \mathcal{G}(\sigma^2)$) if $\psi_X(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$ for all $\lambda > 0$.

- Equivalent definitions:
  - ▷ $\exists K_0 > 0$ s.t. $\psi_X(\lambda) \leq K_0^2\lambda^2$ for all $\lambda > 0$;
  - ▷ $\exists K_1 > 0$ s.t. $\Pr[|X| \geq t] \leq 2\exp\left(-\frac{t^2}{K_1^2}\right)$ for all $t \geq 0$;
  - ▷ $\exists K_2 > 0$ s.t. $\mathbb{E}[|X|^p]^{\frac{1}{p}} \leq K_2\sqrt{p}$ for all $p \geq 1$.
- Concentration:
  - ▷ $\Pr[|X| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2}}$;
  - ▷ Independent $\sum a_i X_i \in \mathcal{G}(\sum a_i \sigma_i^2)$;
  - ▷ Sum of independent $\Pr[|Z| \geq n\epsilon] \leq 2e^{-\frac{n\epsilon^2}{2\sigma^2}}$.

Bounded: A zero-mean bounded rv $X \in [a, b]$ satisfies $X \in \mathcal{G}(\frac{(b-a)^2}{4})$.

- $\Pr[|X - \mathbb{E}[X]| \geq t] \leq 2e^{-\frac{2t^2}{(b-a)^2}}$;
- Hoeffding's inequality: $Z = X_1 + \cdots + X_n$ (independent & bounded):
$$\Pr\left[\frac{1}{n}|Z - \mathbb{E}[Z]| \geq \epsilon\right] \leq 2\exp\left(-\frac{2n\epsilon^2}{\frac{1}{n}\sum_{i=1}^n (b_i - a_i)^2)}\right);$$
$$\Pr\left[\frac{1}{n}|Z - \mathbb{E}[Z]| \geq \epsilon\right] \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right) \text{ if all } a_i/b_i\text{'s equal.}$$
  - ▷ Setting RHS $= \delta$, we have $n \geq \frac{(b-a)^2}{2\epsilon^2}\log\frac{2}{\delta}$.

2 Probability Method

Aim: Prove the existence of certain objects with certain properties.

- Counting: Construct a set of m bad events each with probability at most p. The object of interest must exist when none of the events happens (w.p. at least $1 - mp$).
- Expectation: The probability that a rv is larger/smaller than its expectation is positive.
- Second moment: Use Chebyshev's inequality.
- Sample and modify
- Lovász local lemma: Let $\mathcal{B}_1, \cdots, \mathcal{B}_m$ be bad events such that for each $i \in [m]$, $\Pr[\mathcal{B}_i] \leq p$ and $\mathcal{B}_i$ is mutually independent to all but $\leq d$ events. If $4pd \leq 1$, then $\Pr\left[\bigcap_{i=1}^m \bar{\mathcal{B}}_i\right] \geq (1 - 2p)^m > 0$.

3 Convex Optimization

Convexity:

- Convex set: If $\mathbf{x}, \mathbf{x}' \in D$, then $\lambda\mathbf{x} + (1-\lambda)\mathbf{x}' \in D$ for all $\lambda \in [0, 1]$.
- Convex function: $f(\lambda\mathbf{x} + (1-\lambda)\mathbf{x}') \leq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{x}')$.
  - ▷ Any local minimum is also a global minimum.
  - ▷ f differentiable $\Rightarrow$ convex iff $f(\mathbf{x}') \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{x}' - \mathbf{x})$ for all $\mathbf{x}, \mathbf{x}'$.
  - ▷ f twice differentiable $\Rightarrow$ convex iff $\nabla^2(\mathbf{x}) \succeq \mathbf{0}$ for all $\mathbf{x}$.
  - ▷ If f_1, f_2 convex, $\alpha_1, \alpha_2 > 0$, then $\alpha_1 f_1 + \alpha_2 f_2$ convex.
  - ▷ If $f_1, \cdots, f_L$ convex, then $\max_{\ell \in [L]} f_\ell$ convex.
  - ▷ If h linear/affine and g convex, then $g \circ h$ convex.
  - ▷ Jensen's inequality: For any random vector $\mathbf{X}$ and convex function f, $f(\mathbb{E}[\mathbf{X}]) \leq \mathbb{E}[f(\mathbf{X})]$.

Convex optimization: (1) f_0 and all f_i are convex; (2) all h_i affine.
$$\begin{aligned} \min_{\mathbf{x}} \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) \leq 0, \quad \forall i = 1, \cdots, m_{\text{ineq}} \\ & h_i(\mathbf{x}) = 0, \quad \forall i = 1, \cdots, m_{\text{eq}}. \end{aligned}$$

Lagrangian: $L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_{i \in [m_{\text{ineq}}]} \lambda_i f_i(\mathbf{x}) + \sum_{i \in [m_{\text{eq}}]} \nu_i h_i(\mathbf{x})$.

- $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$ are *Lagrangian multipliers*.
- Lagrangian dual: $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$.
- Lagrangian dual problem: $\max_{\boldsymbol{\lambda}, \boldsymbol{\mu}}\ g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ s.t. $\boldsymbol{\lambda} \geq \mathbf{0}$.
- Weak duality: $g(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \leq f_0(\mathbf{x}^*)$.
- Strong duality: If original problem is convex and a mild regularity condition holds, then $g(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) = f_0(\mathbf{x}^*)$.
  - ▷ Slater's condition: There exists at least one feasible $\mathbf{x}$ s.t. all $f_i(\mathbf{x}) < 0$ and all ${}_i(\mathbf{x}) = 0$.
  - ▷ Another sufficient condition: All f_i are linear.

Lagrangian of LP:
$$\begin{aligned} (\mathbf{P}) \min_{\mathbf{x}} \quad & \mathbf{c}^\top\mathbf{x} \\ \text{s.t.} \quad & \mathbf{A}\mathbf{x} = \mathbf{b}; \mathbf{x} \geq \mathbf{0}. \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} (\mathbf{D}) \max_{\boldsymbol{\nu}} \quad & \mathbf{b}^\top\boldsymbol{\nu} \\ \text{s.t.} \quad & \mathbf{A}^\top\boldsymbol{\nu} \leq \mathbf{c}. \end{aligned}$$

- If we replace by $\mathbf{A}\mathbf{x} \geq \mathbf{b}$, then add constraint $\boldsymbol{\nu} \geq \mathbf{0}$.
- Strong duality: $\min(\mathbf{P}) = \max(\mathbf{D})$.

Examples of convex optimization formulation:

Directed graph $G = (V, E)$; source s, sink t;
$$\begin{aligned} \max_{\{f_{uv}\}} \quad & \sum_{v:(s,v)\in E} f_{sv} \\ \text{s.t.} \quad & 0 \leq f_{uv} \leq c_{uv} \\ & \sum_{u:(u,v)\in E} f_{uv} = \sum_{w:(v,w)\in E} f_{vw}, \\ & \forall v \in V \setminus \{s, t\}. \end{aligned}$$

MaxFlow

$$\begin{aligned} \min_{\boldsymbol{\theta}, \theta_0} \quad & \frac{1}{2}\|\boldsymbol{\theta}\|^2 \\ \text{s.t.} \quad & y_i(\boldsymbol{\theta}^\top\mathbf{x}_i + \theta_0) \geq 1, \\ & \forall i = 1, 2, \cdots, n. \end{aligned}$$

MaxMarginClassifier

Noisy channel $R_i = \frac{1}{2}\log\left(1 + \frac{P_i}{\sigma_i^2}\right)$;
$$\begin{aligned} \max_{P_1, \cdots, P_K} \quad & \sum_{i=1}^K \frac{1}{2}\log\left(1 + \frac{P_i}{\sigma_i^2}\right) \\ \text{s.t.} \quad & \sum_{i=1}^K P_i \leq P_{\text{total}} \\ & P_i \geq 0, \forall i = 1, \cdots, K. \end{aligned}$$

PowerAllocation

Ensure that expected return is at least $r_{\min}$ while minimizing the risk;
$$\begin{aligned} \min \quad & \mathbf{x}^\top\Sigma_{\mathbf{p}}\mathbf{x} \\ \text{s.t.} \quad & \mu_{\mathbf{p}}^\top\mathbf{x} \geq r_{\min} \\ & \sum_{i=1}^n x_i = 1. \end{aligned}$$

PortfolioOptimization

- Dual of MaxFlow:
$$\begin{aligned} \min_{\boldsymbol{\lambda}, \boldsymbol{\mu}} \quad & \sum_{(u,v)\in E} c_{uv}\lambda_{uv} \\ \text{s.t.} \quad & \mu_s = 1, \mu_t = 0 \\ & \lambda_{uv} \geq \mu_u - \mu_v, \lambda_{uv} \geq 0, \ \forall (u, v) \in E. \end{aligned}$$
  - ▷ Max flow = min cut.

4 Submodular

Submodularity: $\forall S \subseteq T \subseteq V,\ e \in V \setminus T,\ \Delta(e|S) \geq \Delta(e|T)$.

- Related notions:
  - ▷ Monotonicity: $S \subseteq T \subseteq V \Rightarrow f(S) \leq f(T)$.
  - ▷ Modularity: $\Delta(e|S) = \Delta(e|T)$.
  - ▷ Supermodularity: $\Delta(e|S) \leq \Delta(e|T)$.
- Equivalent definitions:
  - ▷ $\forall S, T,\ f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$.
  - ▷ $\forall S, e, e',\ \Delta(e|S) \geq \Delta(e|S \cup \{e'\})$.
  - ▷ (If f monotone) $\forall S, T,\ f(T) \leq f(S) + \sum_{e \in T \setminus S} \Delta(e|S)$.
- Relation to concavity:
  - ▷ Diminishing returns;
  - ▷ (Non-monotone case) Any local maximum is within 1/2 of global maximum.
  - ▷ Maximization (unconstrained or constrained) can be done approximately efficiently.
  - ▷ $f(S) = g(|S|)$ is submodular if g is concave.
- Relation to convexity:
  - ▷ Unconstrained minimization can be done exactly efficiently.
  - ▷ An extension from sets to continuous values called *Lovász extension* is a convex function.
- Properties: Suppose f_1, f_2 are submodular:
  - ▷ Linear combinations: $c_1, c_2 > 0 \Rightarrow c_1 f_1 + c_2 f_2$ submodular.
  - ▷ Concave of modular: g modular, h concave $\Rightarrow h \circ g$ submodular.
  - ▷ Residual: $f(S) = f_1(S \cup B) - f_1(B)$ submodular for any B.
  - ▷ Conditioning: $f(S) = f_1(S \cap A)$ submodular for any A.
  - ▷ Reflection: $f(S) = f_1(V \setminus S)$ submodular.
  - ▷ Truncation: If f_1 also monotone, then $f(S) = \min\{c, f_1(S)\}$ is submodular for any c.
  - ▷ Minimum: $\min\{f_1, f_2\}$ is submodular if either $f_1 - f_2$ or $f_2 - f_1$ is monotone.

- Examples:
  - ▷ $f(S)$ = area covered by activating all sensors in S.
  - ▷ Let $\mathbf{X}$ be a matrix, V be the set of column indices, $\mathbf{X}_S$ is the submatrix indexed by $S \subseteq V$. Then $r_S = \text{rank}(\mathbf{X}_S)$ is monotone submodular.
  - ▷ $f(S) =$ total number of users influenced by advertising to S (in a graph).
  - ▷ $f(S) =$ representativeness of images in S.
  - ▷ $f(S) =$ number of edges between S and S^c is submodular but non-monotone.
  - ▷ $f(S) = H(\mathbf{X}_S)$ where *entropy* $H_X = \sum_x P_X(x) \log \frac{1}{P_X(x)}$ is monotone submodular.

Cardinality-constrained submodular maximization:

$$\max_{S \in \mathcal{S}} \quad f(S) \qquad \text{s.t.} \quad \mathcal{S} = \{S : |S| \le k\}.$$

- **Greedy algorithm**: For k times, add $e = \arg\max_{e \in V \setminus S_{i-1}} \Delta(e|S_{i-1})$.
- Useful fact: $1 - x \le e^{-x}, \forall x \in \mathbb{R}$.
- Approximation: If f monotone submodular with $f(\emptyset) = 0$, then $f(S_k) \ge (1 - 1/e) f(S_k^*)$.
- Generalization: If we perform ℓ instead of k iterations, then $f(S_\ell) \ge (1 - e^{-\ell/k}) f(S_k^*)$.

Proof.

$$\begin{aligned} f(S^*) &\le f(S^* \cup S_i) \quad (\text{monotonicity}) \\ &= f(S_i) + \sum_{j=1}^{k} \Delta(e_j^* | S_i \cup \{e_1^*, \cdots, e_{j-1}^*\}) \\ &\le f(S_i) + \sum_{j=1}^{k} \Delta(e_j^* | S_i) \quad (\text{submodularity}) \\ &\le f(S_i) + \sum_{j=1}^{k} \Delta(e_{i+1}^* | S_i) \quad (\text{greedy}) \\ &\le f(S_i) + k(f(S_{i+1}) - f(S_i)). \end{aligned}$$

So $f(S^*) - f(S_{i+1}) \le (1 - 1/k)(f(S^*) - f(S_i))$. Since $(1-1/k)^\ell \le e^{-\ell/k}$, we have proven the theorem.

5 Multiplicative Weight Update

Simple majority: Binary prediction and a perfect expert exists:

① Let $S_t \subseteq [n]$ be the set of experts that make no mistake at the first $t-1$ iterations;
② At iteration t, predict the majority vote from S_t.

- The simple majority algorithm makes at most $\log n$ mistakes.

Proof. Each mistakes eliminate $\ge \frac{1}{2}$ of remaining experts.

Weighted majority: Binary prediction:

① Fix $\eta \in (0, \frac{1}{2}]$; initialize each expert's weight to 1;
② At each iteration $t \in [T]$:
   (a) Predict the weighted majority vote;
   (b) For those who predict wrongly, decay their weight to $1 - \eta$.

- The weighted majority algorithm makes at most $2(1+\eta)M_i + \frac{2 \log n}{\eta}$ mistakes for any expert i, where i makes M_i mistakes.

Proof. Each mistake decreases $\ge \frac{\eta}{2}$ of total weight. Hence, final weight of i, $(1-\eta)^{M_i} \le$ final total weight $\le n \cdot \left(1 - \frac{\eta}{2}\right)^M$.

Randomized weighted majority: At each iteration, predict 0 or 1 with probability proportional to its total weight.

- The randomized weighted majority algorithm, in expectation, makes at most $(1+\eta)M_i + \frac{\log n}{\eta}$ mistakes for any expert i.

Proof. Each mistake decreases $\ge \eta f^{(t)}$ of total weight, where $f^{(t)}$ is the weighted fraction of mistakes at t. Hence, final weight of i, $(1-\eta)^{M_i} \le$ final total weight $\le n \cdot \prod (1 - \eta f^{(t)}) \le n \cdot \exp\left(-\eta \sum f^{(t)}\right) = n \cdot \exp(-\eta \mathbb{E}[M])$.

Multiplicative weight update: Real-valued bounded loss $\in [-1, 1]$:

① Fix $\eta \in (0, \frac{1}{2}]$; initialize each expert's weight to 1;
② At each iteration $t \in [T]$:
   (a) Follow expert i's advice w.p. its normalized weight;
   (b) Decay each expert i's weight to $1 - \eta \cdot$ its loss.

- The MWU algorithm, in expectation, has loss at most $\sum_{t \in [T]} m_i^{(t)} + \sum_{t \in [T]} \left| m_i^{(t)} \right| + \frac{\log n}{\eta}$ (proof ≈ randomized weighted majority).

6 Fourier Transform

Fourier series: Let $f : [-\pi, \pi] \to \mathbb{R}$ be piecewise continuous:

- General bases: $f(x) = \sum_{n=1}^{\infty} \langle f, e_n \rangle e_n(x)$.
  - ▷ Inner product: $\langle f, g \rangle = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \overline{g(x)} \, \mathrm{d}x$.
- Trigonometric bases: $f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx)$,
  - ▷ $a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \, \mathrm{d}x$;
  - ▷ $a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx \, \mathrm{d}x$; 0 for odd functions.
  - ▷ $b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx \, \mathrm{d}x$; 0 for even functions.
- Complex exponential bases: $f(x) = \sum_{n=-\infty}^{\infty} c_n e^{inx}$,
  - ▷ $c_n = \hat{f}_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx} \, \mathrm{d}x$.
    - ∗ $c_k = c_{-k}$ for even functions; $-c_{-k}$ for odd functions.

Parseval's theorem: $\|f\|^2 = \sum_{n=-\infty}^{\infty} |\hat{f}_n|^2 = \sum_{n=-\infty}^{\infty} |\langle f, e^{inx} \rangle|^2$.

- Energy of a signal = energy of its Fourier transform.

Fourier transform: Let $f : \mathbb{R} \to \mathbb{C}$ be piecewise continuous on every finite interval & absolutely integrable ($\int_{-\infty}^{\infty} |f(x)| \, \mathrm{d}x < \infty$):

$$\hat{f}(\omega) = \mathcal{F}(f(x)) = \int_{-\infty}^{\infty} f(x) \, e^{-i\omega x} \, \mathrm{d}x;$$

$$f(x) = \mathcal{F}^{-1}(\hat{f}(\omega)) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega) \, e^{i\omega x} \, \mathrm{d}\omega.$$

- $f(x) = \begin{cases} 1 & |x| \le \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \Rightarrow \hat{f}(\omega) = \frac{\sin(\omega/2)}{\omega/2} = \text{sinc}(\omega/2\pi)$.
- Linearity: $\mathcal{F}(af + bg) = a\mathcal{F}(f) + b\mathcal{F}(g)$.
- Shifting: $f(ax) \Rightarrow \frac{1}{a} \hat{f}(\omega/a)$; $f(x - c) \Rightarrow \hat{f}(\omega) e^{-ic\omega}$.
- Convolution: $(f * g)(x) = \int_{-\infty}^{\infty} f(x - y) g(y) \, \mathrm{d}y \Rightarrow \hat{f}(\omega) \hat{g}(\omega)$.
- Derivative: $\mathcal{F}(f'(x)) = i\omega \mathcal{F}(f(x))$.
- Parseval's theorem: $\int_{-\infty}^{\infty} |f(x)|^2 \, \mathrm{d}x = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\hat{f}(\omega)|^2 \, \mathrm{d}\omega$.

7 Information Theory

Information of an event: If event A occurs with probability p,

$$\text{Info}(A) = \psi(p) = \log_b \frac{1}{p}.$$

- When $b = 2$, information is measured in bits.
- Axiomatization of $\psi(p)$:
  - ▷ Non-negativity: $\psi(p) > 0$;
  - ▷ Zero for definite events: $\psi(1) = 1$;
  - ▷ Monotonicity: $p \le p' \Rightarrow \psi(p) \ge \psi(p')$;
  - ▷ Continuity: $\psi(p)$ is continuous in p;
  - ▷ Additivity under independence: $\psi(p_1 p_2) = \psi(p_1) + \psi(p_2)$.

Shannon entropy: Let X be a discrete random variable with probability mass function P_X. The *Shannon entropy* of X is the average information we learn from observing $X = x$ (note: $0 \log_2 \frac{1}{0} = 0$):

$$H(X) = \mathbb{E}_{X \sim P_X} [\psi(X = x)] = \sum_x P_X(x) \log_2 \frac{1}{P_X(x)}.$$

- Joint entropy:
$$\begin{aligned} H(X, Y) &= \mathbb{E}_{(X,Y) \sim P(X,Y)} [\psi(X = x, Y = y)] \\ &= \sum_{x,y} P_{XY}(x, y) \log_2 \frac{1}{P_{XY}(x, y)}. \end{aligned}$$
- Conditional entropy:
$$\begin{aligned} H(Y|X) &= \mathbb{E}_{(X,Y) \sim P(X,Y)} [\psi(Y = y | X = x)] \\ &= \sum_{x,y} P_{XY}(x, y) \log_2 \frac{1}{P_{Y|X}(y|x)} \\ &= \sum_x P_X(x) H(Y | X = x). \end{aligned}$$
- Entropy measures information or uncertainty in X.
  - ▷ Binary source: $H(X) = H_2(p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}$;
  - ▷ Uniform source: $H(X) = \log_2 |\mathcal{X}|$.
- Properties of entropy:
  - ▷ Non-negativity: $H(X) \ge 0$;
  - ▷ Upper bound: $H(X) \le \log_2 |X|$;
  - ▷ Chain rule (2 var): $H(X, Y) = H(X) + H(Y|X)$;
  - ▷ Chain rule (n var): $H(X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i | X_1, \cdots, X_{i-1})$;
  - ▷ Conditioning reduces entropy: $H(X|Y) \le H(X)$ with equality if and only if X and Y are independent;
  - ▷ Sub-additivity: $H(X_1, \cdots, X_n) \le \sum_{i=1}^{n} H(X_i)$.

KL divergence:

$$D(P||Q) = \mathbb{E}_{X \sim P} \left[\log_2 \frac{P(x)}{Q(x)} \right] = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)}.$$

- $D(P||Q) \ge 0$ with equality if and only if $P = Q$.

Mutual information: Information between random variables:

$$I(X; Y) = H(Y) - H(Y|X).$$

- Terminologies:

- ▷ $H(Y)$: Prior uncertainty in Y;
- ▷ $H(Y|X)$: Remaining uncertainty in Y after observing X;
- ▷ $I(X;Y)$: Information we learn about Y after observing X.
- Joint mutual information:
$$I(X_1, X_2; Y_1, Y_2) = H(Y_1.Y_2) - H(Y_1, Y_2|X_1, X_2).$$
- Conditional mutual information:
$$I(X;Y|Z) = H(Y|Z) - H(Y|X,Z).$$
- Properties of mutual information:
  - ▷ Alternative Forms:
$$I(X;Y) = D(P_{XY}||P(X) \times P(Y)) = \sum_{x,y} P_{XY}(x,y) \log_2 \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} = \sum_{x,y} P_{XY}(x,y) \log_2 \frac{P_{Y|X}(y|x)}{P_Y(y)};$$
  - ▷ Symmetry: $I(X;Y) = I(Y;X) = H(X)+H(Y)-H(X,Y)$;
  - ▷ Non-negativity: $I(X;Y) \geq 0$ with equality if and only if X and Y are independent;
  - ▷ Upper bounds: $I(X;Y) \leq H(X)$; $I(X;Y) \leq H(Y)$.
  - ▷ Chain rule: $I(X_1, \cdots, X_n|Y) = \sum_{i=1}^{n} I(X_i; Y|X_1, \cdots, X_{i-1})$;
  - ▷ Data processing inequality: If X and Z are conditionally independent given Y, then $I(X;Z) \leq I(X;Y)$;
  - ▷ Partial sub-additivity: If $Y_1, \cdots, Y_n$ are conditionally independent given $X_1, \cdots, X_n$, and Y_i depends on $X_1, \cdots, X_n$ only through X_i, then
$$I(X_1, \cdots, X_n; Y_1, \cdots, Y_n) \leq \sum_{i=1}^n I(X_i; Y_i).$$

8 Error-Correcting Codes

Linear code: Any code with parity checks is a *linear code*.

- Types of linear code $\mathbf{u} \to \mathbf{x}$:
  - ▷ Systematic parity-check code: The first k out of n bits of $\mathbf{x}$ are always precisely the original k bits, and the remaining $n-k$ bits are parity checks.
  - ▷ parity-check code: All n codeword bits may be arbitrarily parity checks.
- Generator matrix: $\mathbf{x} = \mathbf{uG}$, $\mathbf{G}$ is the generator matrix.
- Linearity: $\mathbf{x} \oplus \mathbf{x}' = (\mathbf{u} + \mathbf{u}')\mathbf{G}$.
- Parity-check matrix: $\mathbf{xH} = \mathbf{0} \Leftrightarrow \mathbf{x}$ is valid.
  - ▷ For systematic codes, $\mathbf{G} = \begin{bmatrix}\mathbf{I}_k & \mathbf{P}\end{bmatrix} \Rightarrow \mathbf{H} = \begin{bmatrix}\mathbf{P} \\ \mathbf{I}_{n-k}\end{bmatrix}$.

Distance properties:

- Hamming distance: The *Hamming distance* between two vectors $\mathbf{x}$ and $\mathbf{x}'$ is the number of positions in which they differ:
$$d_H(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{n} \mathbb{I}[x_i \neq x_i'].$$
- Minimum distance: The *minimum distance* of a codebook $\mathcal{C}$ of length-n codewords is
$$d_{\min} = \min_{\mathbf{x} \neq \mathbf{x}' \in \mathcal{C}} d_H(\mathbf{x}, \mathbf{x}').$$
  - ▷ If minimum distance is $d_{\min}$, then it is possible to correct up to $d_{\min} - 1$ erasures and $\frac{d_{\min}-1}{2}$ bit flips.
- Weight: $w(\mathbf{x}) = \sum_{i=1}^n \mathbb{I}[x_i = 1]$.
  - ▷ For linear codes, minimum distances equal minimum weights.

Minimum distance decoding:

- Maximum-likelihood decoder: For any channel $P_{\mathbf{Y}|\mathbf{X}}$ and any codebook $\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(M)}\}$, the decoding rule that minimizes the error probabiltiy P_e is the maximum-likelihood decoder:
$$\hat{m} = \arg\max_{j=1,\cdots,M} P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}^{(j)}).$$
  - ▷ For a linear code, if the syndrome is $\mathbf{S} = \mathbf{yH} = \mathbf{zH}$, then the minimum-distance codeword to $\mathbf{y}$ can be obtained by
$$\hat{\mathbf{z}} = \arg\min_{\tilde{\mathbf{z}}:\tilde{\mathbf{z}}\mathbf{H}=\mathbf{S}} w(\tilde{\mathbf{z}}),$$
then computing $\hat{\mathbf{x}} = \mathbf{y} \oplus \hat{\mathbf{z}}$.

9 Expander Graphs

***d*-regular**: Every vertex has degree d.

Edge expansion: For a graph $G = (V, E)$ with n edges, let
$$\phi[S] = \frac{\text{no. of edges } (u,v) \text{ with } u \in S, v \notin S}{|S|}.$$
Cheeger's constant is defined as $\phi_G = \min_{0<|S|\leq n/2} \phi[S]$.

Vertex expansion: For a graph $G = (V, E)$ with n edges, let
$$\phi'[S] = \frac{\text{no. of vertices in } V \setminus S \text{ connected to } S}{|S|}.$$
Vertex expansion number is defined as $\phi_G = \min_{0<|S|\leq n/2} \phi'[S]$.

Bipartite expander: A bipartite graph with $|L| = n$, $|R| = m$, $\deg(u) = d$ for all $u \in L$ is called a *$(n, m, d, \gamma, \epsilon)$-expander* if for all $S \subseteq L$ with $0 \leq |S| \leq \gamma n$ we have $|N(S)| \geq \epsilon d|S|$, where $N(S)$ is the neighbors of S in R.

- Theorem: Suppose the edges in a bipartite graph with $|L| = n$, $|R| = m$ are constructed by: for each $u \in L$, select d vertices in R uniformly at random without replacement and connect them. Then for $d \geq 32, m \geq 3n/4$ and large enough n, w.p. $\geq \frac{18}{19}$ that the graph is an $(n, m, d, \frac{5}{8}, \frac{1}{10d})$-expander.

> *Proof.* Union bound the bad event $|N(S)| < \frac{5}{8}d|S|$ for all S.

- Regular expanders can be converted to bipartite expanders by *double covering* (i.e., maintaining two copies of each vertex).

Explicitness: A deterministic algorithm outputs the expander graph's entire adjacency matrix in poly(n) time.

- Strong explicitness: Given any $u \in [n], i \in [d]$, a deterministic algorithm outputs the i-th neighbor of u in poly($\log n$) time.

10 Communication Complexity

Problem setting:

- Alice has access to $x \in \mathcal{X} = \{0,1\}^n$;
- Bob has access to $y \in \mathcal{Y} = \{0,1\}^n$;
- Goal: Compute $f(x,y)$ where $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ and $\mathcal{Z} = \{0,1\}$.
- Deterministic protocol π: Determines which player sends the next message and what to send.
  - ▷ *π computes a function f* if the value $f(x,y)$ can be deterministically computed following π.
- Communication cost: Total maximum number of bits exchanged.
- Communication complexity: Smallest communication cost.

Protocol tree: A binary tree branched based on a bit is 0 or 1.

- Communication complexity = smallest depth among all trees that compute f.

Rectangle: $A \subseteq \mathcal{X}, B \subseteq \mathcal{Y}$, then $A \times B$ is a *rectangle*.

- Lemma: For every node v in the protocol tree, define
$$S_v = \{\text{all input pairs } (x,y) \text{ that leads to } v\};$$
$$X_v = \{x \in \mathcal{X} : \exists y \in \mathcal{Y} \text{ s.t. } (x,y) \in S_v\};$$
$$Y_v = \{y \in \mathcal{Y} : \exists x \in \mathcal{X} \text{ s.t. } (x,y) \in S_v\}.$$
Then, $S_v = X_v \times Y_v$. Also, the rectangles correspond to all leaves form a partition of $\mathcal{X} \times \mathcal{Y}$.

> *Proof.* Induction: A node v has left child u and right child w. WLOG suppose v is Alice sending a bit, then X_v is split into X_u and X_w based on whether it is 0 or 1.

- Monochromatic rectangle: Given $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ and $z \in \mathcal{Z}$, rectangle R is *z-monochromatic* if $f(x,y) = z$ for all $(x,y) \in R$.
  - ▷ Fact: If v is a leaf, then R_v is monochromatic.
  - ▷ Theorem: If communication complexity of f is c, then $\mathcal{X} \times \mathcal{Y}$ can be partitioned to at most 2^c monochromatic rectangles with respect to f.
    - ∗ Corollary: If cannot partition, then must exceed c.
  - ▷ Theorem: If there exists a partition to at most 2^c monochromatic rectangles, then exists a protocol of length $O(c^2)$ that computes f.

Lower bounds from rectangles:

- `EQUALS`: $\geq n+1$ as each 1-monochromatic rectangle has size 1×1.
- `DISJ`: There are 3^n 1's and each 1-monochromatic rectangle has size at most 2^n. So $\geq \log(3^n/2^n + 1)$.

Rank bound:

- Communication complexity of $f \leq \text{rank}(M_f) + 1$.

> *Proof.* Factorize $M_f = AB$. Alice sending the r-bit row of A to Bob suffices, Here $r \geq \text{rank}(M_f)$.

- Communication complexity of $f \geq \log(\text{rank}(M_f) + 1)$ if M_f is not the all-1 matrix.

> *Proof.* Rank $c \Rightarrow$ at most 2^c monochromatic rectangles R with value z_R. Let M_R be the matrix indicating if $(x,y) \in R$ (z_R if so otherwise 0). If $z_R = 0$, M_R has rank 0; 1 otherwise. M is the sum of all M_R with at least 1 being all-0. Hence $\text{rank}(M) \leq 2^c - 1$ and $c \geq \log(\text{rank}(M_f) + 1)$.

Fooling set: Every monochromatic rectangle with respect to g can share at most one element with S.

- Theorem: Exists fooling set size $s \Rightarrow$ complexity $\geq \log s$.
- The set $S = \{(X, N \setminus X) : X \subseteq [n]\}$ is a fooling set for `DISJ`.
  - ▷ So communication complexity $\geq n + 1$.

Most functions require high communication: Only a vanishingly small (as $n \to \infty$) fraction of such functions can be computed with $n-2$ bits (or fewer) of communication using deterministic protocols.