# Cookbook

## Recipe

# Open problem

**Tail of hypergeometric distribution**

If we estimate the maximum from the sampled subset of the population, what is the tail probability of the cardinality of a set whose value is greater than the estimated maximum?

2025-06-17

# 1. Random Variable Control

## 1.1 Maximum of i.i.d gaussian

Let $\xi_1, \ldots, \xi_k$ be $k$ independent samples from $\mathcal{N}(0,1)$. Then

$$\mathbb{E}\left[\max\left\{\xi_1^2, \ldots, \xi_k^2\right\}\right] \leq 2\log(2k) \tag{1.1}$$

## 1.2 Union bound for partial sums

**Etemadi's inequality.** Let $X_1, \ldots, X_n$ be independent random variables. For $i \in [n]$, let $Y_i = \sum_{j=1}^{i} X_j$ denote the partial sum up to $i$. Then for all $\alpha \geq 0$,

$$\Pr[\max_{i=1}^{n} |Y_i| > 3 \cdot \alpha] \leq 3 \cdot \max_{i=1}^{n} \Pr[|Y_i| > \alpha]. \tag{1.2}$$

*Proof Sketch.* $\mathbb{P}[|Y_i| > \alpha]$ term을 얻기 위해서 $|Y_i - Y_n|$과 $|Y_i|$ 사이의 independence를 사용함. 그리고 partial sum의 maximum과 각 partial sum을 연결하기 위해서 $i$번 째 partial sum이 처음으로 $3\alpha$ 보다 큰 event로 분해함. (Detail)

## 1.3 Random Singed Summation Bound

**Khintchine ineqaulity.** Let $\{\varepsilon_n\}_{n=1}^{N}$ be i.i.d. Rademacher random variables. Let $0 < p < \infty$ and let $x_1, \ldots, x_N \in \mathbb{C}$. Then

$$A_p \left(\sum_{n=1}^{N} |x_n|^2\right)^{1/2} \leq \left(\mathbb{E}\left|\sum_{n=1}^{N} \varepsilon_n x_n\right|^p\right)^{1/p} \leq B_p \left(\sum_{n=1}^{N} |x_n|^2\right)^{1/2} \tag{1.3}$$

## 1.4 Summation Bound

**Marcinkiewicz-Zygmund ineqaulity.** If $X_i$, $i = 1, \ldots, n$ are independent random variables with $\mathbb{E}[X_i] = 0$ and $\mathbb{E}[|X_i|^p], 1 < p < +\infty$, then

$$A_p\, \mathbb{E}\left[\left(\sum_{i=1}^{n} |X_i|^2\right)^{p/2}\right] \leq \mathbb{E}\left[\left|\sum_{i=1}^{n} X_i\right|^p\right] \leq B_p\, \mathbb{E}\left[\left(\sum_{i=1}^{n} |X_i|^2\right)^{p/2}\right] \tag{1.4}$$

where $A_p$ and $B_p$ are positive constants, which depend only on $p$. for some constants $A_p, B_p$ depending only on $p$.

**Latala's inequality.** If $p \geq 2$ and $X, X_1, \ldots, X_n$ are i.i.d. mean 0 random variables, then we have

$$\left\|\sum_{i=1}^{n} X_i\right\|_{L^p} \sim \sup\left\{\frac{p}{s}\left(\frac{n}{p}\right)^{1/s} \|X\|_{L^s} \;\middle|\; \max\left\{2, \frac{p}{n}\right\} \leq s \leq p\right\} \tag{1.5}$$

## 1.5 Concentraion ineqaulity

### 1.5.1 Bernstein's ineqaulity

Let $X_1, \dots X_n$ be independent random variables. Assume $\mathbb{E}[X_i] = 0, \mathbb{E}[X_i^2] = \sigma_i^2$, and $\Pr[|X_i| \leq 1] = 1$ for each $i \in [n]$. Let $\sigma^2 := \sum_{i=1}^n \sigma_i^2$. Then for all $t \geq 0$,

$$\Pr\left[\sum_{i=1}^n X_i \geq t\right] \leq \exp\left(\frac{-3t^2}{6\sigma^2 + 2t}\right) \tag{1.6}$$

*Proof Sketch.* First bound the MGF of each $X_i$ using taylor expansion. Then use Markov inequality for $\Pr\left[\sum_{i=1}^n X_i \geq t\right]$ with $\exp(\lambda \cdot)$ and minimize the upper bound with $\lambda$. Then use the following lemma to finish the proof.

**Lemma.** Let $v > -1$. Then $(1+v)\log(1+v) \geq v + \frac{3v^2}{2v+6}$

### 1.5.2 Matrix version Bernstein's inequality

Let $\boldsymbol{B}$ a fixed $q \times d$ matrix. Construct $q \times d$ matrix $\boldsymbol{R}$ such that

$$\mathbb{E}[\boldsymbol{R}] = \boldsymbol{B}, \quad \|\boldsymbol{R}\|_{\text{op}} \leq L \tag{1.7}$$

Form the matrix sampling estimator

$$\bar{\boldsymbol{R}}_m = \frac{1}{m}\sum_{i=1}^m \boldsymbol{R}_i, \tag{1.8}$$

where each $\boldsymbol{R}_i$ is an independent copy of $\boldsymbol{R}$. Then for every $t > 0$, the estimator satisfies

$$\mathbb{P}\left[\left\|\bar{\boldsymbol{R}}_m - \boldsymbol{B}\right\|_{\text{op}} \geq t\right] \leq (q+d) \cdot \exp\left(\frac{-mt^2}{m_2(\boldsymbol{R}) + 2Lt/3}\right), \tag{1.9}$$

where $m_2(R)$ is the second moment $m_2(\boldsymbol{R}) = \max\left\{\left\|\mathbb{E}[\boldsymbol{R}^*\boldsymbol{R}]\right\|_{\text{op}}, \left\|\mathbb{E}[\boldsymbol{R}\boldsymbol{R}^*]\right\|_{\text{op}}\right\}$.

### 1.5.3 Hoeffding's inequality

Let $X_1, \dots, X_n$ be independent random variables such that $a_i \leq X_i \leq b_i$ a.s. Then for all $t > 0$,

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \tag{1.10}$$

where $S_n = X_1 + \dots X_n$. Also, consider a set of $r$ i.i.d. random variables $X_1 \dots X_r$ such that $-\Delta \leq X_i \leq \Delta$ and $\mathbb{E}[X_i] = 0$ for each $i \in [r]$. Let $\sum_{i=1}^r X_i$. Then for any $\alpha \in (0, 1/2)$

$$\mathbb{P}[|M| > \alpha] \leq 2\exp(-\frac{\alpha^2}{2r\Delta^2}) \tag{1.11}$$

The proof uses the following:

**Lemma.** Let $X$ be any real-valued random variable such that $a \leq X \leq b$ a.s. Then. for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right) \tag{1.12}$$

COOL FACT: conditional expectation also works for the lemma.

**Variance-only form.** Consider a set of $r$ independent random variables $X_1, \dots X_r$. Let $M = \sum_{i=1}^r X_i$. Then for $\alpha \in (0, 2\text{Var}[M]/(\max_i |X_i - \mathbb{E}[X_i]|))$

$$\mathbb{P}[|M - \mathbb{E}[M]| > \alpha] \leq 2\exp\left(\frac{-\alpha^2}{4\sum_{i=1}^r \text{Var}[X_i]}\right). \tag{1.13}$$

### 1.5.4 Paley-Zygmund inequality

If $Z \geq 0$ is a random variable with finite variance, and if $0 \leq \theta \leq 1$, then

$$\mathbb{P}(Z > \theta\,\mathbb{E}[Z]) \geq (1-\theta)^2 \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]} \tag{1.14}$$

### 1.5.5 Max of independent Gaussians

Let $X_1, X_2, \ldots, X_n$ i.i.d. $\mathcal{N}(0,1)$, then

$$\mathbb{E}[\max(X_1, \ldots, X_n)] = \sqrt{2\log(n)} + o(\sqrt{\log(n)}) \tag{1.15}$$

### 1.5.6 DKW inequality

DKW inequality provides a bound on the worst-case distance of empirical CDF and the true CDF:

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| > \varepsilon\right) \leq Ce^{-2n\varepsilon^2} \quad \text{for every } \varepsilon > 0. \tag{1.16}$$

For multivariate case, let $X_1, X_2, \ldots, x_n$ be an i.i.d. sequence of $k$-dimensional vectors,

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}^k} |F_n(t) - F(t)| > \varepsilon\right) \leq (n+1)ke^{-2n\varepsilon^2} \tag{1.17}$$

for every $\varepsilon, n, k > 0$.

Also see local DKW inequality

Steinke version Let $X_1, \ldots, X_n$ be independent random variables with CDF $f(v) := \mathbb{P}[X_i \leq v]$ for all $i \in [n]$ and $v \in \mathbb{R}$. Let the empirical CDF be $F_x(v) := \frac{1}{n}\sum_{i=1}^n 1[X_i \leq v]$ for all $v \in \mathbb{R}$. Then, for all $\beta > 0$,

$$\mathbb{P}_X\left[\sup_{v\in\mathbb{R}} F_X(v) - f(v) \leq \sqrt{\frac{2\log(1/\beta)}{n}} + \frac{\log(1/\beta)}{2n}\right] \geq 1-\beta. \tag{1.18}$$

**Lemma.** For all $t, \lambda > 0$,

$$\mathbb{P}\left[\sup_{v\in\mathbb{R}} F_x(v)\log\left(1+\frac{t}{f(v)}\right) > \frac{\lambda}{n}\right] \leq (1+t)^n e^{-\lambda} \leq e^{tn-\lambda}. \tag{1.19}$$

Note: maximum이 bound되는 event 확률 구할 때는 martingale construction해서 optional stopping theorem적용하는 것도 좋음 ㅋ Lemma에서는 bionamial exponent에 놓아서 martingale 만듬.

### 1.5.7 Quadratic form

**Definition (Subgaussian random variable).** A centered random variable $X$ is said to be $v$-subgaussian if its cumulant generating function is subquadratic:

$$\xi_X(t) \leq \frac{1}{2}vt^2 \quad \forall t \in \mathbb{R} \tag{1.20}$$

**Hanson-Wright tail bound.** Let $\boldsymbol{x}$ be a random vector with independent centered $v$-subgaussian entries and let $\boldsymbol{A}$ be a square matrix. Then

$$\mathbb{P}\left(\left|\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}]\right| \geq t\right) \leq 2\exp\left(-\frac{c\cdot t^2}{v^2\left\|\boldsymbol{A}\right\|_F^2 + v\left\|A\right\| t}\right), \tag{1.21}$$

where $c > 0$ is a constant independent of $v, \boldsymbol{x}, t$ or $\boldsymbol{A}$.

### 1.5.8 Gaussian CCDF bound

$$1 - \Phi(w) \leq \min\left\{\frac{1}{2}, \frac{1}{w\sqrt{2\pi}}\right\} e^{-w^2/2}, \quad w > 0 \tag{1.22}$$

### 1.5.9 McDiarmid's inequality

A function $f : \mathcal{X} \times \mathcal{X} \times \ldots \mathcal{X}_n \to \mathbb{R}$ satisfies the bounded differences property if substituting the value of the $i$th coordinate $x_i$ changes the value of $f$ by at most $c_i$. More formally, if there are constants $c_1, c_2, \ldots, c_n$ such that for all $i \in [n]$, and all $x_i \in \mathcal{X}_1, x_2 \in \mathcal{X}_2, \ldots, x_n \in \mathcal{X}_3$,

$$\sup_{x_i' \in \mathcal{X}_i} |f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i \tag{1.23}$$

Let $f : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ satisfy the bounded differences property with bounds $c_1, c_2, \ldots, c_n$.
Consider independent random variables $X_1, X_2, \ldots, X_n$ where $X_i \in C_i$ for all $i$. Then, for any $\varepsilon > 0$,

$$\mathbb{P}(f(X_1, X_2, \ldots, X_n) - \mathbb{E}[f(X_1, X_2, \ldots, X_n)] \geq \varepsilon) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right), \tag{1.24}$$

$$\mathbb{P}(f(X_1, X_2, \ldots, X_n) - \mathbb{E}[f(X_1, X_2, \ldots, X_n)] \leq -\varepsilon) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right) \tag{1.25}$$

and as an immediate consequence,

$$\mathbb{P}(|f(X_1, X_2, \ldots, X_n) - \mathbb{E}[f(X_1, X_2, \ldots X_n)]| \geq \varepsilon) \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right) \tag{1.26}$$

## 1.6 Decoupling lemma

### 1.6.1 Quadratic form

Let $f : \mathbb{R} \to \mathbb{R}$ be convex. Let $X_1, \ldots X_n \in \mathbb{R}$ be independent mean-zero random variables. For $i, j \in [n]$, let $a_{i,j} \in \mathbb{R}$ be a constant. Then

$$\mathbb{E}\left[f\left(\sum_{i \neq j} a_{ij} X_i X_j\right)\right] \leq \mathbb{E}\left[f\left(4\sum_{i \neq j} a_{ij} X_i X_j'\right)\right], \tag{1.27}$$

where $X_1', \ldots, X_n'$ are independent copies of $X_1, \ldots, X_n$.

Note: We can analyze $\boldsymbol{x}^* \boldsymbol{A} \boldsymbol{x}$ by $\boldsymbol{x}^* \boldsymbol{A} \boldsymbol{x}'$ with independent $\boldsymbol{x}'$.

## 1.7 Global variance control

### 1.7.1 Efron-Stein inequality

For $i \in [n]$ and tuple $Z = (Z_1, \ldots, Z_n)$, let $Z^{(i)}$ denote the tuple $(Z_1, \ldots, Z_{i-1}, \tilde{Z}_i, Z_{i+1}, \ldots, Z_n)$, where $\tilde{Z}_i$ is an independent copy of $Z_i$. For a scalar function $f(Z)$, the Efron-Stein inequality states that

$$\mathrm{Var}[f(Z)] = \mathbb{E}[(f(Z) - \mathbb{E}[f(Z)])^2] \leq \frac{1}{2} \cdot \sum_{i \in [n]} \mathbb{E}\left[\left(f(Z) - f(Z^{(i)})\right)^2\right] \tag{1.28}$$

$$\stackrel{\dagger}{=} \underbrace{\sum_{i \in [n]} \mathbb{E}\left[\left(f(Z) - E_i[f(Z^{(i)}])\right)^2\right]}_{\text{sum of conditional variance}} \tag{1.29}$$

† : Note that $\mathbb{E} = \mathbb{E}_{-i}\,\mathbb{E}_i, \mathbb{E}[f(Z^{(i)}) \mid Z] = \mathbb{E}_i[f(Z^{(i)})]$ and

$$\mathop{\mathbb{E}}_{i}[(Z - \mathop{\mathbb{E}}_{i}[Z])^2] = \frac{1}{2}\mathop{\mathbb{E}}_{i}[(Z - Z^{(i)})^2] \tag{1.30}$$

## 1.8 Information

### 1.8.1 Fano's inequality

Let $X \in \{0,1\}^d$ be uniformly random and let $Y \in \mathbb{R}^d$ be a random variable that depends on $X$.
If $\mathbb{E}[\|X - Y\|_1] \leq \alpha \cdot d$ for $\alpha \leq \frac{1}{2}$, then

$$I(X;Y) \geq d \cdot D_{\mathrm{KL}}\left(\mathrm{Ber}(\alpha) || \mathrm{Ber}\left(\frac{1}{2}\right)\right). \tag{1.31}$$

## 1.9 Do you like martingale?

### 1.9.1 Tail Distribution

Let $X$ be a nonnegative cadlag submartingale. Then, for each $K, t > 0$,

$$K\mathbb{P}(X_t^* \geq K) \leq \mathbb{E}[1_{\{X_t^* \geq K\}} X_t] \tag{1.32}$$

## 1.10 Stochastic Dominance

**Definition.** Let $X, Y \in \mathbb{R}$ be random variables. We say $X$ is *stochastically dominated* by $Y$ if $\mathbb{P}[X > t] \leq \mathbb{P}[Y > t]$ for all $t \in \mathbb{R}$. Equivalently, $X$ is stochastically dominated by $Y$ if there exists a coupling such that $\mathbb{P}[X \leq Y] = 1$.

### 1.10.1 SD is preserved under sums/convolutions

**Lemma.** Suppose $X_1$ is stochastically dominate by $Y_1$. Suppose that, for all $x \in \mathbb{R}$, the conditional distribution $X_2|X_1 = x$ is stochastically dominated by $Y_2$. Assume that $Y_1$ and $Y_2$ are independent. Then $X_1 + X_2$ is stochastically dominated by $Y_1 + Y_2$.

# 2. Privacy

## 2.1 Privacy Loss

**Definition.** Let $Y$ and $Z$ be two random variables. The privacy loss random variables $\mathcal{L}_{Y||Z}$ is distributed by drawing $t \sim \mathrm{Law}(Y)$, and outputting $\log\left(\frac{\mathbb{P}[Y=t]}{\mathbb{P}[Z=t]}\right)$. If the support of $Y$ and $Z$ are not equal, then the privacy loss random variable is undefined.

## 2.2 $\varepsilon - \delta$ DP

### 2.2.1 4 ways to see $\delta$

**Proposition.** Let $P$ and $Q$ be two probability distributions on $\mathcal{Y}$ such that the privacy loss distribution $\mathrm{PrivLoss}(P||Q)$ is well-defined. Fix $\varepsilon \geq 0$ and define

$$\delta := \sup_{S \subset \mathcal{Y}} P(S) - e^{\varepsilon} Q(S). \tag{2.1}$$

Then

$$\begin{aligned}\delta &= \mathop{\mathbb{P}}_{Z\sim\text{PrivLoss}(P||Q)}[Z > \varepsilon] - e^{\varepsilon} \cdot \mathop{\mathbb{P}}_{Z'\sim\text{PrivLoss}(Q||P)}[-Z' > \varepsilon] \\ &= \mathop{\mathbb{E}}_{Z\sim\text{PrivLoss}(P||Q)}[\max\{0, 1 - \exp(\varepsilon - Z\}] \\ &= \int_{\varepsilon}^{\infty} e^{\varepsilon - z} \mathop{\mathbb{P}}_{Z\sim\text{PrivLoss}(P||Q)}[Z > z]\,\mathrm{d}z \\ &\leq \mathop{\mathbb{P}}_{Z\sim\text{PrivLoss}(P||Q)}[Z > \varepsilon].\end{aligned}$$

### 2.2.2 Moment difference bound

Let $X$ and $Y$ be a random variable supported on $[-\Delta, \Delta]$ satisfying $\mathbb{P}[X \in S] \leq e^{\varepsilon}\,\mathbb{P}[Y \in S] + \delta$ for all measurable $S$ and vice versa. Then

$$\mathbb{E}[X] - \mathbb{E}[Y] \leq (e^{\varepsilon} - 1)\,\mathbb{E}[|Y|] + 2\delta\Delta \tag{2.2}$$

## 2.3 zCDP

**Definition.** A randomised mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ is $(\xi, \rho)$-zero-concentrated differentially private if, for all $x, x' \in \mathcal{X}^n$ differing on a single entry and all $\alpha \in (1, \infty)$,

$$\mathbb{E}[e^{(\alpha-1)Z}] \leq e^{(\alpha-1)(\xi+\rho\alpha)}, \tag{2.3}$$

where $Z = \text{PrivLoss}(M(x)||M(x'))$ is the privacy loss random variable.

### 2.3.1 Key properties

1. Pure $\varepsilon$-DP imples $\frac{1}{2}\varepsilon^2$-zCDP
2. The composition of $k$ independent $\frac{1}{2}\varepsilon^2$-zCDP algorithms satisfies $\frac{1}{2}\varepsilon^2 k$-zCDP.
3. $\frac{1}{2}\varepsilon^2 k$-zCDP implies approximate $(\varepsilon', \delta)$-DP with $\delta \in (0, 1)$ arbitrary and $\varepsilon' = \varepsilon \cdot \sqrt{2k\log(1/\delta)} + \frac{1}{2}\varepsilon^2 k$.

## 2.4 Approximate Rényi Differential Privacy

> "Rényi differential privacy was introduced by Minorov and was motivated by analyzing privacy amplification by subsampling interleaved with composition, which arises in differentially private deep learning"

— Thomas Steinke **Defintion (RDP).** An algorithm $M$ is said to be $(\lambda, \varepsilon)$-RDP with $\lambda \geq 1$ and $\varepsilon \geq 0$, if for any adjacent inputs $x, x'$

$$D_\lambda(M(x)||M(x')) := \frac{1}{\lambda - 1}\log \mathop{\mathbb{E}}_{Y\leftarrow M(x)}\left[\left(\frac{\mathbb{P}[M(x) = Y]}{\mathbb{P}[M(x') = Y]}\right)^{\lambda-1}\right] \leq \varepsilon \tag{2.4}$$

**Tip:** The $\varepsilon$ should be thought of as a function $\varepsilon(\lambda)$, rather than a single number.

**Properties**

Let $P, Q$ be probability distributions over $\mathcal{Y}$ with a common sigma-algebra such that $P$ is absolutely continuous with respect to $Q$.

1. **Postprocessing (a.k.a. data processing inequality) & non-negativity:** Let $f : \mathcal{Y} \to \mathcal{Z}$ be a measurable function. Let $f(P)$ denote the distribution on $\mathcal{Z}$

obtained by applying $f$ to a sample from $P$; define $f(Q)$ similarly. Then

$$0 \leq D_\alpha(f(P)\|f(Q)) \leq D_\alpha(P\|Q) \quad \text{for all } \alpha \in [1,\infty].$$

2. **Composition:** If $P = P' \times P''$ and $Q = Q' \times Q''$ are product distributions, then

$$D_\alpha(P\|Q) = D_\alpha(P'\|Q') + D_\alpha(P''\|Q'') \quad \text{for all } \alpha \in [1,\infty].$$

More generally, suppose $P$ and $Q$ are distributions on $\mathcal{Y} = \mathcal{Y}' \times \mathcal{Y}''$. Let $P'$ and $Q'$ be the marginal distributions on $\mathcal{Y}'$ induced by $P$ and $Q$ respectively. For $y' \in \mathcal{Y}'$, let $P''_{y'}$ and $Q''_{y'}$ be the conditional distributions on $\mathcal{Y}''$ induced by $P$ and $Q$ respectively. That is, we can generate a sample $Y = (Y', Y'') \leftarrow P$ by first sampling $Y' \leftarrow P'$ and then sampling $Y'' \leftarrow P''_{Y'}$, and similarly for $Q$. Then

$$D_\alpha(P\|Q) \leq D_\alpha(P'\|Q') + \sup_{y' \in \mathcal{Y}'} D_\alpha(P''_{y'}\|Q''_{y'}) \quad \text{for all } \alpha \in [1,\infty].$$

3. **Monotonicity:** For all $1 \leq \alpha \leq \alpha' \leq \infty$,

$$D_\alpha(P\|Q) \leq D_{\alpha'}(P\|Q).$$

4. **Gaussian divergence:** For all $\mu, \mu' \in \mathbb{R}$ with $\sigma > 0$ and all $\alpha \in [1,\infty)$,

$$D_\alpha(\mathcal{N}(\mu,\sigma^2)\|\mathcal{N}(\mu',\sigma^2)) = \alpha \cdot \frac{(\mu-\mu')^2}{2\sigma^2}.$$

5. **Pure DP to Concentrated DP:** For all $\alpha \in [1,\infty)$,

$$D_\alpha(P\|Q) \leq \frac{\alpha}{8} \cdot \left(D_\infty(P\|Q) + D_\infty(Q\|P)\right)^2.$$

6. **Quasi-convexity:** Let $P'$ and $Q'$ be probability distributions over $\mathcal{Y}$ such that $P'$ is absolutely continuous with respect to $Q'$. For $s \in [0,1]$, let $(1-s)\cdot P + s \cdot P'$ denote the convex combination of the distributions $P$ and $P'$ with weighting $s$. For all $\alpha \in (1,\infty)$ and all $s \in [0,1]$,

$$\begin{aligned}
&D_\alpha\left((1-s)\cdot P + s\cdot P' \,\middle\|\, (1-s)\cdot Q + s\cdot Q'\right) \\
&\leq \frac{1}{\alpha-1}\log\left((1-s)\cdot\exp\left((\alpha-1)D_\alpha(P\|Q)\right) + s\cdot\exp\left((\alpha-1)D_\alpha(P'\|Q')\right)\right) \\
&\leq \max\left\{D_\alpha(P\|Q), D_\alpha(P'\|Q')\right\},
\end{aligned}$$

and

$$D_1\left((1-s)\cdot P + s\cdot P' \,\middle\|\, (1-s)\cdot Q + s\cdot Q'\right) \leq (1-s)\cdot D_1(P\|Q) + s\cdot D_1(P'\|Q').$$

7. **Triangle-like inequality (a.k.a. group privacy):** Let $R$ be a distribution on $\mathcal{Y}$ and assume that $Q$ is absolutely continuous with respect to $R$. For all $1 < \alpha < \alpha' < \infty$,

$$D_\alpha(P\|R) \leq \frac{\alpha'}{\alpha'-1} \cdot D_{\alpha' \cdot \frac{\alpha'-1}{\alpha'-\alpha}}(P\|Q) + D_{\alpha'}(Q\|R).$$

In particular, if $D_\alpha(P\|Q) \leq \rho_1 \cdot \alpha$ and $D_\alpha(Q\|R) \leq \rho_2 \cdot \alpha$ for all $\alpha \in (1,\infty)$, then

$$D_\alpha(P\|R) \leq \left(\sqrt{\rho_1} + \sqrt{\rho_2}\right)^2 \cdot \alpha \quad \text{for all } \alpha \in (1,\infty).$$

8. **Conversion to approximate DP:** For all measurable $S \subset \mathcal{Y}$, all $\alpha \in (1,\infty)$, and all $\tilde{\varepsilon} \geq D_\alpha(P\|Q)$,

$$P(S) \leq e^{\tilde{\varepsilon}}\cdot Q(S) + e^{-(\alpha-1)(\tilde{\varepsilon}-D_\alpha(P\|Q))}\cdot\frac{1}{\alpha}\left(1-\frac{1}{\alpha}\right)^{\alpha-1} \leq e^{\tilde{\varepsilon}}\cdot Q(S) + e^{-(\alpha-1)(\tilde{\varepsilon}-D_\alpha(P\|Q))}.$$

**Definition (Approximate RDP).** A randomized algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ is $\delta$-approximately $(\lambda, \varepsilon)$-Rényi differentially private if, for all neighboring pairs of inputs $x, x' \in \mathcal{X}^n$, $D_\lambda^\delta(M(x)||M(x')) \leq \varepsilon$.

### 2.4.1 Properties

1. $(\varepsilon, \delta)$-DP is equivalent to $\delta$-approximate $(\infty, \delta)$-RDP.
2. $(\varepsilon, \delta)$-DP implies $\delta$-approximate $(\lambda, \frac{1}{2}\varepsilon^2\delta)$-RDP for all $\lambda \in (1, \infty)$.
3. $\delta$-approximate $(\lambda, \varepsilon)$-RDP implies $(\hat{\varepsilon}, \hat{\delta})$-DP for

$$\hat{\delta} = \delta + \frac{\exp((\lambda-1)(\hat{\varepsilon}-\varepsilon))}{\lambda} \cdot \left(1 - \frac{1}{\lambda}\right)^{\lambda-1}. \tag{2.5}$$

4. $\delta$-approximate $(\lambda, \varepsilon)$-RDP is closed under postprocessing.
5. If $M_1$ is $\delta_1$-approximately $(\lambda, \varepsilon_1)$-RDP and $M_2$ is $\delta_2$-approximately $(\lambda, \varepsilon_2)$-RDP, then their composition is $(\delta_1 + \delta_2)$-approximately $(\lambda, \varepsilon_1 + \varepsilon_2)$-RDP.

## 2.5 Composition

### 2.5.1 Advanced composition $(\varepsilon, \delta)$

**Theorem.** If each mechanism $m_i$ is in a $k$-fold adaptive composition $m_1, \ldots, m_k$ satisfies $\varepsilon$-differential privacy, then for any $\delta' \geq 0$, the entire $k$-fold adaptive composition satisfies $(\varepsilon', \delta')$-differential privacy, where

$$\varepsilon' = \varepsilon\sqrt{2k\log(1/\delta')} + k\varepsilon(e^\varepsilon - 1) \tag{2.6}$$

**Theorem.** For $j \in [k]$, let $M_j \in \mathcal{X}^n \times \mathcal{Y}_{i-1} \to \mathcal{Y}_i$ be randomized algorithms. Suppose $M_j$ is $(\varepsilon_j, \delta_j)$-DP for each $j \in [k]$. For $j \in [k]$, inductively define $M_{1\ldots j} : \mathcal{X}^n \to \mathcal{Y}_j$ by $M_{1\ldots j}(x) = M_j(x, M_{1\ldots(j-1)}(x))$, where each algorithm is run independently and $M_{1\ldots 0} = y$ for some fixed $y_0 \in \mathcal{Y}_0$. Then $M_{1\ldots k}$ is $(\varepsilon, \delta)$-DP for any $\delta > \sum_{j=1}^k \delta_j$ with

$$\varepsilon = \min\left\{\sum_{j=1}^k \varepsilon_j, \frac{1}{2}\sum_{j=1}^k \varepsilon_j^2 + \sqrt{2\log(1/\delta')\sum_{k=1}^k \varepsilon_j^2}\right\} \tag{2.7}$$

## 2.6 Joint Differential Privacy

**Definition.** For $\varepsilon, \delta \geq 0$, a randomized algorithm $\mathcal{M} : \mathbb{N}^{\mathcal{X}} \to \mathcal{Y}^N$ is $(\varepsilon, \delta)$-joint differentially private if for every possible pair of $z, z' \in \mathcal{X}$, for every $i \in [N]$, and for every subset of possible outputs $E \subseteq \mathcal{Y}^{N-1}$, we have

$$\mathbb{P}_{\mathcal{M}}[\mathcal{M}(z \cup D_{-z})_{-i} \in E] \leq e^\varepsilon \mathbb{P}_{\mathcal{M}}[\mathcal{M}(z' \cup D_{-z})_{-i} \in E] + \delta \tag{2.8}$$

where $\mathcal{M}_{-i}$ denotes the output of $\mathcal{M}$ that excludes the $i$th dimension.

## 2.7 Lower bound tools

### 2.7.1 Query moment w.r.t binary data

**Correlation–Variance Dichotomy.** Let $f : \{0,1\}^d \to [0,1]$ be an arbitrary function. Let $P \in [0,1]$ be uniformly random and, conditioned on $P$, let $X_1, X_2, \ldots X_n$ be independent with $\mathbb{E}[X_i] = P$ for each $i \in [n]$. Then

$$\underbrace{\mathop{\mathbb{E}}_{X,P}\left[(f(X) - P) \cdot \sum_{i=1}^n (X_i - P)\right]}_{\text{일종의 total correlation}} + \mathop{\mathbb{E}}_{P}\left[\mathop{\mathbb{E}}_{X}\left[f(X) - \overline{X}\right]^2\right] \geq \frac{1}{12} \tag{2.9}$$

## 2.8 Decomposition

### 2.8.1 Basic decomposition

Let $P$ and $Q$ be probability distributions over $\mathcal{Y}$. Fix $\varepsilon, \delta \geq 0$. Suppose that, for all measurable $S \subset \mathcal{Y}$, we have $P(S) \leq e^{\varepsilon} Q(S) + \delta$ and vice versa.
Then there exist $\delta' \in [0, \delta]$ and distributions $P', Q', P''$ and $Q''$ over $\mathcal{Y}$ such that the following three properties are all satisfied.

1. We can express $P$ and $Q$ as convex combinations:
$$P = (1-\delta')P' + \delta' P''$$
$$Q = (1-\delta')Q' + \delta' Q''$$
2. Second, for all measurable $S \subset \mathcal{Y}$, we have $e^{-\varepsilon} P'(S) \leq Q'(S) \leq e^{\varepsilon} P'(S)$
3. There exists measurable $S, T \subset \mathcal{Y}$ such that $P''(S) = 1, Q''(T) = 1, \forall S' \subset S\ P(S') \geq Q(S')$, and $\forall T' \subset T\ Q(T') \geq P(T')$

**Corollary.** Let $P$ and $Q$ be probability distribution over $\mathcal{Y}$. Fix $\varepsilon, \delta$. Suppose that for all measurable $S \subset \mathcal{Y}$, we have $P(S) \leq e^{\varepsilon} Q(S) + \delta$ and $Q(S) \leq e^{\varepsilon} P(S) + \delta$. Then there exist distributions $A, B, P''$, and $Q''$ over $\mathcal{Y}$ such that

$$P = (1-\delta)\frac{e^{\varepsilon}}{e^{\varepsilon}+1}A + (1-\delta)\frac{1}{e^{\varepsilon}+1}B + \delta P'',$$
$$Q = (1-\delta)\frac{e^{\varepsilon}}{e^{\varepsilon}+1}B + (1-\delta)\frac{1}{e^{\varepsilon}+1}A + \delta Q''$$

**Interpretation:** All $(\varepsilon, \delta)$ DP distributions can be represented as a postprocessing of the $(\varepsilon, \delta)$ randomized response with the postprocessing F such that $F(0, \perp) = A, F(1, \perp) = B, F(0, \top) = P''$ and $F(1, \top) = Q''$

### 2.8.2 Bayesian version [2, Lemma 5.6]

**Question**

Suppose we observe a sample from either $P$ or $Q$ and we have a prior on these two possibilities, what is the posterior distribution of possibilities? We need to account for the event with $\delta$ where things "fail" arbitrarily.

Let $P$ and $Q$ be probability distributions over $\mathcal{Y}$. Fix $\varepsilon, \delta \geq 0$. Suppose that, for all measurable $S \subset \mathcal{Y}$, we have $P(S) \leq e^{\varepsilon} Q(S) + \delta$ and vice versa.
Then there exists a randomized function $E_{P,Q} : \mathcal{Y} \to \{0, 1\}$ with the following properties:

1. Fix $p \in [0, 1]$ and suppose $X \sim \text{Bernoulli}(p)$. If $X = 1$, sample $Y \sim P$ else $Y \sim Q$. Then for all $Y \in \mathcal{Y}$, we have
$$\mathop{\mathbb{P}}_{\substack{X \sim \text{Bernoulli(p)} \\ Y \sim XP + (1-X)Q}} [X = 1 \wedge E_{P,Q}(Y) = 1 | Y = y] \leq \frac{p}{p + (1-p)e^{-\varepsilon}}$$
2. Under each hypothesis $Y \sim P$ and $Y \sim Q$, the expected value $E_{P,Q}(Y)$ is equal or greater than $1 - \delta$.

# 3. Probability and Statistics

## 3.1 Moment

### 3.1.1 Moment and CDF

For random variable $X$,

$$\mathbb{E}[X] = \int_0^{\infty} \mathbb{P}(X \geq t)\, \mathrm{d}t - \int_{-\infty}^{0} \mathbb{P}(X \leq t)\, \mathrm{d}t \tag{3.1}$$

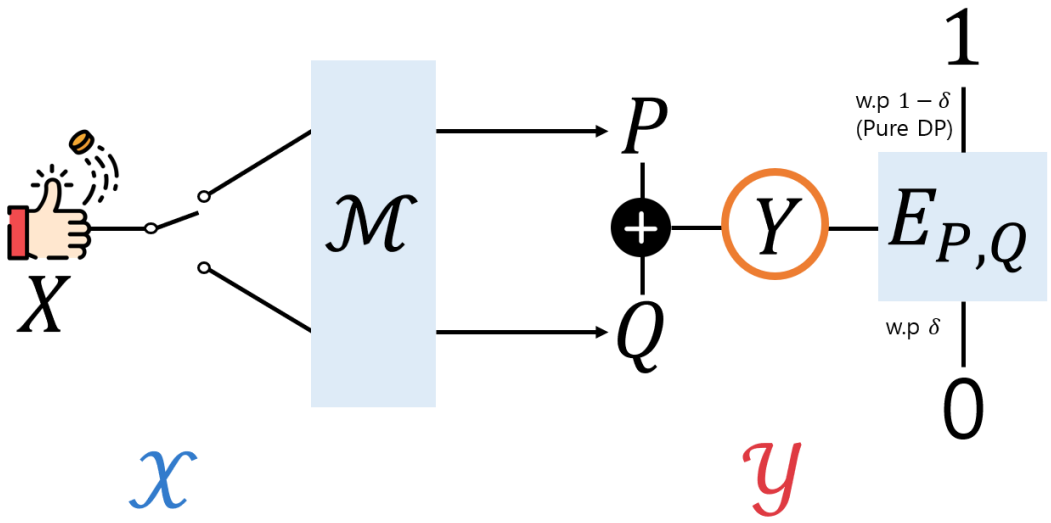


Figure 1: Visualization of the Bayesian version decomposition

In general, $X \geq 0$ and a smooth function $g$ with $g(0) = 0$

$$\mathbb{E}[g(X)] = \int_0^\infty g'(t)P(X > t)\,\mathrm{d}t \tag{3.2}$$

#### 3.1.2 $k$-th moment in the lens of CDF

$$\frac{1}{k}\,\mathbb{E}[X^k] = -\int_{-\infty}^0 x^{k-1}F(x)\,\mathrm{d}x + \int_0^\infty x^{k-1}(1 - F(x))\,\mathrm{d}x \tag{3.3}$$

#### 3.1.3 Clipped random variable

For complementary CDF $\bar{F}$ of random variable $X \geq 0$,

$$\mathbb{E}[\min(X, k)] = \int_0^k xf(x)\,\mathrm{d}x + k\bar{F}(k), \tag{3.4}$$

## 3.2 Conditional Distribution

**Tip:** ML/Statistics 분야에서 흔히 쓰는 notation $p(x|y), p(x, y)$ 같은 것은 $X, Y$ 들의 pdf/pmf라고 생각하자. 거의 measure는 안나옴.

**Conditioning on event.** Let $f_{X,Y}$ be the joint density of $X$ and $Y$, and $f_X(x)$ is the marginal density of $X$.

1. Single point conditioning $X = 1$

$$f_{Y|X=1}(y) = \frac{f_{X,Y}(1, y)}{f_X(1)}$$

2. Set conditioning $X \in S$

$$f_{Y|X\in S}(y) = \frac{\int_S f_{X,Y}(x, y)\,\mathrm{d}x}{\int_S f_X(x)\,\mathrm{d}x} = \frac{\int_S f_{Y|X=x}(y)f_X(x)\,\mathrm{d}x}{\int_S f_X(x)\,\mathrm{d}x}$$

3. Event conditioning

$$f_{Y|A}(y) = \frac{f_Y(y)1[y \in A]}{\int_A f_Y(y)\,\mathrm{d}y}$$

## 3.3 Weak convergence

### 3.3.1 Delta method

If there is a sequence of random variables $X_n$ satisfying

$$\sqrt{n}[X_n - \theta] \xrightarrow{w} \mathcal{N}(0, \sigma^2) \tag{3.5}$$

then

$$\sqrt{n}[g(X_n) - g(\theta)] \xrightarrow{w} \mathcal{N}(0, \sigma^2 \cdot [g'(\theta)]^2) \tag{3.6}$$

given that $g'(\theta)$ exists and is non-zero value.

### 3.3.2 Slutsky's theorem

If $X_n$ converges in distribution to a random element $X$ and $Y_n$ converges in probability to a constant $c$, then

1. $X_n + Y_n \xrightarrow{w} X + c$
2. $X_n Y_n \xrightarrow{w} Xc$
3. $X_n / Y_n \xrightarrow{w} X/c$

where $\xrightarrow{w}$ denotes convergence in distribution.

## 3.4 Central limit theorem

### 3.4.1 Salem-Zygmund

**Theorem.** Let $U$ be a uniform random variable with support $(0, 2\pi)$, and let $X_k = r_k \cos(n_k U + a_k)(0 \le a_k < 2\pi)$, where

1. $n_k$ satisfy the *lacunarity condition*: there exists $q > 1$ such that $n_{k+1} \ge q n_k$ for all $k$
2. $\sum_{i=1}^{\infty} r_i^2 = \infty$ and $\frac{r_k^2}{r_1^2 + \cdots + r_k^2} \to 0$

Then,

$$\frac{X_1 + \cdots + X_k}{\sqrt{r_1^2 + \cdots + r_k^2}} \tag{3.7}$$

converges in distribution to $\mathcal{N}(0, 1/2)$.

## 3.5 Conditional independence

**Theorem.** Let $p_{XYZ}$ be the joint PDF/PMF of $X, Y$ and $Z$. Then the following are equivalent with up to almost-everywhere equivalence:

1. $X \perp Y \mid Z$
2. $p_{XYZ}(x, y, |z) = p_{X|Z}(x|z) p_{Y|Z}(y|z)$
3. $p_{X|YZ}(x|y, z) = p_{X|Z}(x|z)$
4. $p_{XYZ}(x, y, z) = \frac{p_{XZ}(x,z) p_{YZ}(y,z)}{p_Z(z)}$
5. $p_{XYZ}(x, y, z) = g(x, z) h(y, z)$ for some measurable functions $g$ and $h$
6. $p_{X|YZ}(x|y, z) = w(x, z)$ for some measurable function $w$

**Properties.** Let $X, Y, Z, W$ be RVs

1. (symmetry) $X \perp Y \mid Z \iff Y \perp X \mid Z$
2. (decomposition) $X \perp Y \mid Z \Rightarrow h(X) \perp Y \mid Z$ for any measurable function $h$
3. (weak union) $X \perp Y \mid Z \Rightarrow X \perp Y \mid Z, h(X)$ for any measurable function $h$
4. (contraction)
$$X \perp Y \mid Z \text{ and } X \perp W \mid (Y, Z) \iff X \perp (W, Y) \mid Z \tag{3.8}$$
5. If the joint PDF $P_{XYZW}(x, y, z, w)$ satisfies $f_{YZW}(y, z, w) > 0$ almost everywhere. Then
$$X \perp Y \mid (W, Z) \text{ and } X \perp W \mid (Y, Z) \iff X \perp (W, Y) \mid Z \tag{3.9}$$

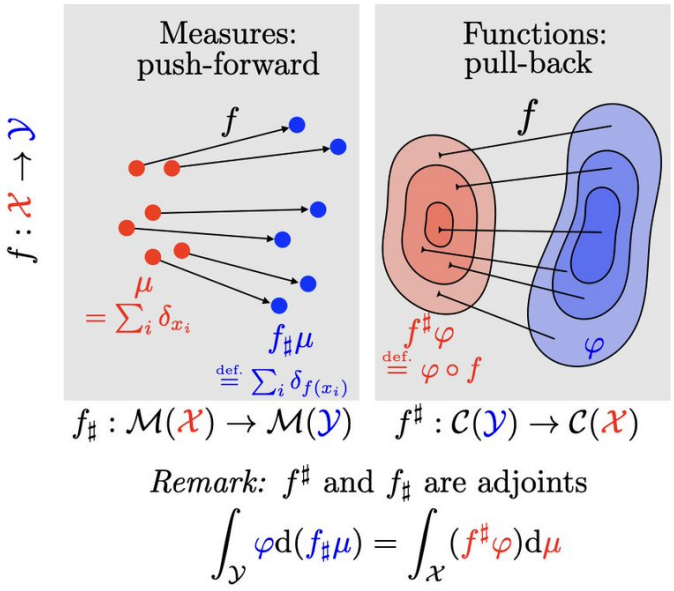


Figure 2: Pullback of functions and pushforward of measures are dual one with each other!

### 3.5.1 Bayes' Theorem

Assume that $X$ is a random variable on $(\Omega, \mathcal{F}, P)$,and let $Q$ be another probability measure on $(\Omega, \mathcal{F})$ with Radon-Nikodym derivative

$$L = \frac{\mathrm{d}Q}{\mathrm{d}P} \text{ on } \mathcal{F} \tag{3.10}$$

Assume that $X \in L^1(\Omega, \mathcal{F}, Q)$ and that $\mathcal{G}$ is a sigma-algebra with $\mathcal{G} \subseteq \mathcal{F}$. Then

$$\overset{Q}{\mathbb{E}}[X \mid \mathcal{G}] = \frac{\mathbb{E}^P[L \cdot X \mid \mathcal{G}]}{\mathbb{E}^P[L \mid \mathcal{G}]}, \quad Q - a.s. \tag{3.11}$$

### 3.5.2 Conditional expectation under independence

**Proposition.** Let $(\Omega, \mathcal{B}, \mathbb{P})$ be a probability space, $(\mathbb{X}, \mathcal{M})$, $(\mathbb{Y}, \mathcal{N})$ be measurable spaces, $X : \Omega \to X$ and $Y : \Omega \to \mathbb{Y}$ be measurable functions. If $X$ and $Y$ are *independent* and $f \in (\mathcal{M} \otimes \mathcal{N})_b$ then

$$\mathbb{E}[f(X, Y) \mid X] = \mathbb{E}[f(x, Y)]|_{x=X} \text{ a.s.} \tag{3.12}$$

## 3.6 Regular conditional distribution

**Theorem.** If $X$ is a real random variable defined on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ then for every $\sigma$-algebra $\mathcal{G} \subset \mathcal{F}$ there is a regular conditional distribution for $X$ given $\mathcal{G}$.

Regular conditional distributions are useful in part because they allow one to reduce many problems concerning conditional expectations to problems concerning only ordinary expectations. For such applications the following disintegration formula for conditional expectations is essential.

### 3.6.1 Disintegration formula

**Theorem.** Let $\mu_w(\mathrm{d}x)$ be a regular conditional distribution for $X$ given $\mathcal{G}$, let $Y$ be $\mathcal{G}$-measurable, and let $f(x, y)$ be a jointly measurable real-valued function such that $\mathbb{E}[|f(X, Y)|] < \infty$. Then,

$$\mathbb{E}[f(X, Y) \mid \mathcal{G}] = \int f(x, Y(w)) \mu_w(\mathrm{d}x) \quad \text{a.s.} \tag{3.13}$$

**Theorem2.** Let $Y$ and $X$ be two Radon spaces. Let $\mu \in P(Y)$, let $\pi : Y \to X$ be a Borel-measurable function, and let $\nu \in P(X)$ be the pushforward measure from $Y$ to $X$ by $\pi$. Then there exists a $\nu$-almost everywhere uniquely determined family of probability measures $\{\mu_x\}_{x \in X} \subseteq P(Y)$ such that

1. the function $x \mapsto \mu_x$ is Borel measurable
2. $\mu_x$ lives on the fiber $\pi^{-1}(x)$
3. for every Borel-measurable function $f : Y \to [0, +\infty]$,

$$\int_Y f(y)\,\mathrm{d}\mu(y) = \int_X \int_{\pi^{-1}(x)} f(y)\,\mathrm{d}\mu_x(y)\,\mathrm{d}\nu(x) \tag{3.14}$$

## 3.7 Markov kernel

A Markov kernel (also called transition kernel, stochastic kernel, or probability kernel) is a mathematical formalization of a "function with random outcomes".

## 3.8 Mutual Information

### 3.8.1 Concavity of Mutual information

Let $\alpha$ be the law of $X$ and $\pi$ be the conditions law of $Y|X$. Let $I_1$ be $I(X;Y)$ where $(X,Y) \sim (\alpha_1, \pi)$, let $I_2$ be $I(X;Y)$ where $(X,Y) \sim (\alpha_2, \pi)$, let $I$ be $I(X;Y)$ where $(X,Y) \sim (\lambda\alpha_1 + (1-\lambda)\alpha_2, \pi)$, for some $0 \leq \lambda \leq 1$, then

$$I \geq \lambda I_1 + (1-\lambda) I_2.$$

## 3.9 Fisher Information

Given the score function $\log p(\theta; X)$, the Fisher Information is defined as

$$I(\theta) := \mathop{\mathbb{E}}_{\theta}\left[-\frac{\partial^2}{\partial\theta^2}\log p(\theta; X)\right] \tag{3.15}$$

It gives you uncertainty about the estimation since

$$\underbrace{\mathrm{Var}\left[\frac{\partial \ell(\theta; X)}{\partial\theta}\right]}_{\text{variance of score}} = -\,\mathbb{E}\left[\frac{\partial^2 \ell(\theta; X)}{\partial\theta^2}\right] \tag{3.16}$$

holds.

### 3.9.1 Cramér–Rao bound

Unbiased means there

$$\mathbb{E}[\hat{\theta}(X) - \theta \mid \theta] = \int \Big(\hat{\theta}(x) - \theta\Big) f(x;\theta)\,\mathrm{d}x = 0 \text{ regardless of the value of } \theta \tag{3.17}$$

Then, the following holds

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)} \tag{3.18}$$

The precision to which we can estimate $\theta$ is fundamentally limited by the Fisher information of the likelihood function.

## 3.10 MLE estimation

### 3.10.1 Asymptotic normality

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} \mathcal{N}(0, I^{-1}(\theta)) \tag{3.19}$$

The mean square error (MSE) of $\hat{\theta}_n$ is

$$\mathrm{MSE}(\hat{\theta}_n, \theta_0) = \mathrm{bias}^2(\hat{\theta}) + \mathrm{Var}(\hat{\theta}_n) \approx \frac{1}{nI(\theta_0)} \tag{3.20}$$

Moreover, if we know about $I(\theta_0)$, we can construct a $1-\alpha$ confidence interval using

$$\left[\hat{\theta}_n - \frac{z_{1-\alpha/2}}{\sqrt{n\hat{I}(\theta_0)}}, \hat{\theta}_n + \frac{z_{1-\alpha/2}}{\sqrt{n\hat{I}(\theta_0)}}\right] \tag{3.21}$$

[1]

## 3.11 Famous Family

### 3.11.1 Poission

**Binomial of Poisson trials is Poisson.** Let $\lambda \geq 0, p \in [0,1]$. Suppose $(X_i)_{i=1}^{\infty}$ are i.i.d Bernoulli random variables with parameter $p$, and $N$ is a Poisson($\lambda$) random variable independent of the $X_i$'s. Then $\sum_{i=1}^{N} X_i \sim \mathrm{Poi}(\lambda p)$.

**Tail Distribution.** Let $X \sim \mathrm{Poi}(\lambda)$, for some parameter $\lambda > 0$. Then for any $x > 0$, we have

$$\mathbb{P}[X \geq \lambda + x] \leq e^{-\frac{x^2}{2\lambda} h\left(\frac{x}{\lambda}\right)} \tag{3.22}$$

, and, for any $0 < x < \lambda$,

$$\mathbb{P}[X \leq \lambda - x] \leq e^{-\frac{x^2}{2\lambda} h\left(-\frac{x}{\lambda}\right)}. \tag{3.23}$$

where $h(u) := 2\frac{(1+u)\log(1+u)-u}{u^2}$. In particular, this implies that for every $x > 0$,

$$\mathbb{P}[|X - \lambda| \geq x] \leq 2e^{-\frac{x^2}{2(\lambda+x)}}. \tag{3.24}$$

### 3.11.2 Binomial

**Fact.** If $X$ is Binomial$(n,p)$, then $\mathbb{E}[1/(X+1)] \leq 1/((n+1)\cdot p)$

### 3.11.3 Negative Binomial

**Negative binomial as a Poisson + logarithmic.**

1. Draw $T$ from a Poisson distribution and draw $K_1, K_2, \ldots, K_T$ independently from a logarithmic distribution.
2. Then $K = \sum_{t=1}^{T} K_t$ follows a negative binomial distribution.

## 3.12 M-estimator

### 3.12.1 Consistency

**Theorem.** Suppose that

1. $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \to 0$ in probability (ULLN)
2. For all $\varepsilon > 0$, $\sup\{M(\theta) : d(\theta,\theta_0) \geq \varepsilon\} < M(\theta_0)$ (identifiability)
3. $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_p(1)$

Then $\hat{\theta}_n \to \theta_0$ in probability.

[1] CI: estimator ± z-value * (SD of estimator)

# 4. Convexity

## 4.1 Characterization

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable over an open domain. Then, the following are equivalent

1. $f$ is convex
2. $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$, for all $x, y, \in \text{dom}(f)$
3. $\nabla^2 f(x) \succeq 0$, for all $x \in \text{dom}(f)$

여기서 Condition 3은 모든 점에서 non-negative curvature를 가지고 있다는 의미.

### 4.1.1 Characterization of Strict Convexity

1. $\nabla^2 f(x) \succ 0, \forall \in \Omega$ ( The converse is not true)
2. A function $f$ is strictly convex on $\Omega \subseteq \mathbb{R}^n$ if and only if

$$f(y) > f(x) + \nabla^\top f(x)(y - x), \ \forall x, y, \in \Omega, x \neq y$$

3. $f$ is strongly convex if and only if there exists $m > 0$ such that

$$\begin{aligned} & f(y) \geq f(x) + \nabla^\top f(x)(y - x) + m \left\| y - x \right\|^2, \forall x, y, \in \text{dom}(f) \\ \iff & \nabla^2 f(x) \succeq m\boldsymbol{I}, \forall x \in \text{dom}(f) \end{aligned}$$

### 4.1.2 When can we assume equal variables?

If the **constraints and the function to be optimized are both symmetric with respect to a group of permutations of the variables**, then the solution set will also be symmetric with respect to this group.

## 4.2 Dual Problem

$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu)$ is concave and lower bound of the optimal value.

**Example.**

$$\begin{aligned} \text{maxmize} \quad & \sum_{i=0}^{m} \mathop{\mathbb{P}}_{F}[F(t) = i] \cdot \mathop{\mathbb{P}}_{\check{W}}[\check{W}(t) \geq v - i] \\ \text{subject to} \quad & \sum_{i=0}^{m} \mathbb{P}[F(t) = i] \cdot i \leq 2m \cdot \delta \\ & \sum_{i=0}^{m} \mathop{\mathbb{P}}_{F}[F(t) = i] = 1, \text{and} \mathop{\mathbb{P}}_{F}[F(t) = i] \geq 0 \ \forall i \in \{0, 1, \cdots, m\} \end{aligned}$$

**Step 1. Find Lagrangian**

$$L(F, \alpha, \beta, \lambda) = \sum_{i=0}^{m} \mathop{\mathbb{P}}_{F}[F(t) = i](\mathop{\mathbb{P}}_{\check{W}}[\check{W}(t) \geq v - i] - \alpha i - \beta - \lambda_i) + 2m\delta\alpha + \beta \tag{4.1}$$

with $\alpha \geq 0, \lambda \geq 0$. Then, $g(\alpha, \beta, \lambda) := \sup_F L(F, \alpha, \beta, \lambda) \geq \sup_{F \in C} L(F, \alpha, \beta, \lambda) \geq f^*$. To drop the first term, add the constraint: $\alpha \cdot i + \beta \geq \mathbb{P}_{\check{W}}[\check{W}(t) \geq v - i] \ \forall i \in \{0, 1, \cdots, m\}$.

**Step 2. Optimize the Dual function**

$$\begin{aligned} \text{minimize} \quad & 2m\delta\alpha + \beta \\ \text{subject to} \quad & \alpha \cdot i + \beta \geq \mathop{\mathbb{P}}_{\check{W}}[\check{W}(t) \geq v - i] \quad \forall i \in \{0, 1, \ldots, m\} \\ & \alpha \geq 0 \end{aligned}$$

Make $\alpha, \beta$ as small as possible!

$$\beta = \mathop{\mathbb{P}}_{\check{W}^*}[\check{W} \geq v],$$
$$\alpha = \max\left(\{0\} \cup \left\{\frac{1}{i}\left(\mathop{\mathbb{P}}_{\check{w}^*}[\check{W} \geq v - i] - \beta\right) : i \in \{1, 2, \cdots, m\}\right\}\right)$$

where $\hat{W}^*$ is a distribution on $\mathbb{R}$ such that $\mathbb{P}[\check{W}^* \geq v - i] \geq \mathbb{P}[\check{W}(t) \geq v - i]$ for all $i \in \{0, 1, \cdots, m\}$ and all $t$ in the support of $T$.

**Remark.** Dual problem을 구하는 과정에서 constraint 추가해도 괜찮음.

# 5. Integral Technique

## 5.1 Gaussian Integral

### 5.1.1 Gauss-Hermite quadrature

$$\int_{-\infty}^{\infty} e^{-x^2/2} f(x)\,\mathrm{d}x \approx \sum_{i=1}^{n} w_i f(x_i) \tag{5.1}$$

where $n$ is the number of sampled points used. The $x_i$ are the roots of the physicists' version of the Hermite polynomial $H_n(x)(i = 1, 2, \ldots, n)$ and the associated weights $w_i$ are given by

$$w_i = \frac{2^{n-1} n! \sqrt{\pi}}{n^2 [H_{n-1}(x_i)]^2} \tag{5.2}$$

### 5.1.2 Stein's lemma

For a differentiable function $\phi : \mathbb{R} \to \mathbb{R}$ with $\mathbb{E}_{x\sim\mathcal{N}(0,1)}[|\phi'(x)|] < \infty$

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,1)}[\phi(x)x] = \mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,1)}[\phi'(x)] \tag{5.3}$$

### 5.1.3 Change of measure

$$\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)\,\mathrm{d}\boldsymbol{w} = \mathcal{N}(\boldsymbol{x}^\top \boldsymbol{w}|\boldsymbol{x}^\top \mu_n, \boldsymbol{x}^\top \boldsymbol{\Sigma}_n \boldsymbol{x})\,\mathrm{d}(\boldsymbol{x}^\top \boldsymbol{w}) \tag{5.4}$$

# 6. Stochastic process

## 6.1 Predictable process

### 6.1.1 How to understand the predictable process?

**Fact 1.** The predictable sigma-algebra is generated by continuous and adapted processes.

**Fact 2.** The predictable sigma-algebra is generated by the sets of the form

$$\{(s, t] \times A : t > x \geq 0, A \in \mathcal{F}_s\} \cup \{\{0\} \times A : A \in \mathcal{F}_0\} \tag{6.1}$$

## 6.2 Local martingale

### 6.2.1 Quadratic variation

**Fact.** If $X$ is a continuous local martingale, then $[X]_t < \infty$ a.s. for every $t \geq 0$, where $[X]$ denote the quadratic variation of the process $X$.

### 6.2.2 Stochastic integral

**Proposition.** For any continuous $L^2$-martingale $M$ where $M_0 = 0$, and any predictable step process $V$ where $|V| \leq 1$, the process $(V \cdot M)$ is an $L^2$-martingale with $\mathbb{E}(V \cdot M)_t^2 \leq \mathbb{E}\, M_t^2$.

## 6.3 Doob's h transform

Set

$$h(x) = \underset{x}{\mathbb{P}}(\tau_A < \tau_B) \tag{6.2}$$

Then, $h(x)$ is the probability, starting from $x$ to hit $A$ before hitting $B$. Then $h$ is positive on $\mathcal{X} \setminus (A \cup B)$. Furthermore, for $x \notin A \cup B$

$$\hat{P}(x, y) = \underset{x}{\mathbb{P}}[X_1 = y | \tau_A < \tau_B] \tag{6.3}$$

Finally, $h(x) = \mathbb{P}_x(\tau_A < \tau_B)$ satisfies both

1. $h(x) = 1$ for $x \in A$ and $h(z) = 0$ for $z \in B$
2. $h$ is harmonic at $x$ for every $x \notin A \cup B$

and $h(\cdot)$ is the unique solution of linear system given by 1 and 2 above. (Good source)

# 7. Fundamental Algebra

## 7.1 Series

### 7.1.1 Geometric

$$\sum_{n=a}^{b} r^n = r^a \frac{1 - r^{b-a+1}}{1 - r} = \frac{r^a - r^{b+1}}{1 - r} \tag{7.1}$$

$$\sum_{j=1}^{\infty} \frac{j^2}{\rho^j} = \frac{\rho(\rho + 1)}{(\rho - 1)^3} \tag{7.2}$$

### 7.1.2 Quotient Stack

(source)

$$\sum_{k=1}^{n} \lfloor \frac{n}{k} \rfloor = 2 \sum_{k=1}^{\lfloor \sqrt{n} \rfloor} \lfloor \frac{n}{k} \rfloor - \lfloor \sqrt{n} \rfloor^2 \tag{7.3}$$

## 7.2 Inequalities

### 7.2.1 Ratio of Summation

$$\frac{\sum_i a_i}{\sum_i b_i} \leq \max_i \frac{a_i}{b_i} \tag{7.4}$$

### 7.2.2 Weighted AM-GM Inequality

Let the nonnegative numbers $x_1, x_2, \ldots, x_n$ and the nonnegative weights $w_1, w_2, \ldots w_n$ be given. Set $w = w_1 + w_2 + \ldots w_n$. If $w > 0$, then the inequality

$$\frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w} \geq \sqrt[w]{x_1^{w_1} x_2^{w_2} \cdots x_n^{w_n}} \tag{7.5}$$

holds.

### 7.2.3 Titu's Lemma

**Summation Form.** For any real numbers $a_1, a_2, \ldots a_n$ and positive reals $b_1, b_2, b_3, \ldots b_n$, we have

$$\frac{a_1^2}{b_1} + \frac{a_2^2}{b_2} + \cdots + \frac{a_n^2}{b_n} \geq \frac{(a_1 + a_2 + \cdots + a_n)^2}{b_1 + b_2 + \ldots b_n} \tag{7.6}$$

**Probabilistic Form.** Let $X$ be a real random variable and $Y$ be a positive random variable such that $\mathbb{E}[|X|]$ and $\mathbb{E}[Y]$ are well defined. Then

$$\mathbb{E}[X^2/Y] \geq \mathbb{E}[|X|^2]/\mathbb{E}[Y] \geq \mathbb{E}[X]^2/\mathbb{E}[Y] \tag{7.7}$$

#### 7.2.4 Cauchy-Schwarz Inequality

For any non-zero vector $\boldsymbol{x}$,

$$\|\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{x}\|_1^2 \leq \|\boldsymbol{x}\|_0 \|\boldsymbol{x}\|_2^2 \tag{7.8}$$

(Note: Useful in binary matrix multiplication.)

#### 7.2.5 Chebyshev's Sum Inequality

If $a_1 \geq a_2 \geq \cdots \geq a_n$ and $b_1 \geq b_2 \geq \cdots \geq b_n$, then

$$\frac{1}{n}\sum_{k=1}^{n} a_k b_k \geq \left(\frac{1}{n}\sum_{k=1}^{n} a_k\right)\left(\frac{1}{n}\sum_{k=1}^{n} b_k\right) \tag{7.9}$$

#### 7.2.6 Symmetric Parametric Inequality

$$\left(1-p+\frac{p}{x}\right)^{\alpha-2} \cdot (1-p+px)^{\alpha-2} \geq 1 \tag{7.10}$$

#### 7.2.7 We love Jensen

By convexity of $(u, v) \mapsto u^\lambda v^{1-\lambda}$

$$\mathbb{E}[U]^\lambda \, \mathbb{E}[V]^{1-\lambda} \leq \mathbb{E}[U^\lambda V^{1-\lambda}] \tag{7.11}$$

## 7.3 Bounds and Approximations

#### 7.3.1 Exponential Bound on Hyperbolic Ratio

For $0 \leq y < x \leq 2$,

$$\frac{\sinh(x)-\sinh(y)}{\sinh(x-y)} \leq e^{\frac{1}{2}xy} \tag{7.12}$$

#### 7.3.2 Exponential Inequalities

For all $t \in \mathbb{R}$ and $0 \leq p \leq 1$,

$$1-p+p\cdot x \leq e^{p(x-1)} \tag{7.13}$$

Additionally:

$$\forall x, y \geq 0 \quad \frac{1+e^{x+y}}{e^x+e^y} \leq e^{xy/2} \tag{7.14}$$

$$\frac{1}{2}(e^x+e^{-x}) \leq e^{x^2/2} \tag{7.15}$$

#### 7.3.3 Softplus Quadratic Bound

For all $a, x \in \mathbb{R}$ with $a \neq 0$, we have

$$\log(1+e^x) \leq \log(1+e^a) + \frac{x-a}{1+e^{-a}} + \frac{(e^a-1)\cdot(x-a)^2}{4\cdot a\cdot(e^a+1)} \tag{7.16}$$

#### 7.3.4 Miscellaneous upper Bounds

**Linear-Rational function.** For all $x > 0$, $t \in [0, 1]$:

$$\frac{1}{1-t+tx} \leq 1 - t(1-t)(1+3t)(x-1) + t^2\left((1-t)x^2 + \frac{t}{x} - 1\right) \tag{7.17}$$

**Inverse logarithmic.** For all $u > 0$:

$$\frac{1}{\log(1+1/u)} \leq u + \frac{1}{2} \tag{7.18}$$

$x \log_+ y$ **decoupling.** For non negative reals $x$ and $y$,

$$x \log_+ y + 1 \wedge +1 \leq x \log x + e^{-1} y + 1 \tag{7.19}$$

**Weighted reciprocal.** For all $p \in [0,1]$ and $x \in (0,\infty)$,

$$\frac{1}{1-p+p/x} \leq 1 - p + p \cdot x \tag{7.20}$$

### 7.3.5 Order of Rademacher Sums

Let $\sigma \in \{-1,1\}^n$ be a random Rademacher sequence and let $a \in \mathbb{R}^n$ be an arbitrary real vector with sorted entries $|a_1| \geq |a_2| \geq \cdots \geq |a_n|$. Then

$$\|\langle a, \sigma \rangle\|_{L^p} \sim \sum_{i \leq p} a_i + \sqrt{p} (\sum_{i>p} a_i^2)^{1/2} \tag{7.21}$$

### 7.3.6 Sum Approximation

$$\max(a,b) \leq a + b \leq 2 \max(a,b), \quad a, b \geq 0 \tag{7.22}$$

## 7.4 Combinatorics

### 7.4.1 Expansion

$$\binom{2n}{m} = \sum_{j=0}^{\lfloor \frac{m}{2} \rfloor} \binom{n}{j} \binom{n-j}{m-2j} 2^{m-2j} \tag{7.23}$$

### 7.4.2 Vandermonde's Identity

$$\sum_{i=0}^{r} \binom{m}{i} \binom{n}{r-i} = \binom{n+m}{r} \tag{7.24}$$

$$\sum_{m=0}^{n} \binom{m}{j} \binom{n-m}{k-j} = \binom{n+1}{k+1} \tag{7.25}$$

Special form (**hockey-stick identity**):

$$\sum_{m=k}^{n} \binom{m}{k} = \binom{n+1}{k+1} \tag{7.26}$$

(c.f. $\prod_{\ell=0}^{k-1} \left(\frac{\ell+\eta}{\ell+1}\right) = \binom{k+\eta-1}{k}$)

# 8. Useful real analysis results

## 8.1 Leibniz's Integral Rule

Let $\mu$ be a probability distribution with support $\Omega$, let $I \subset \mathbb{R}$ be a nontrivial open interval, also let $f : \Omega \times I \to \mathbb{R}$ be a map with the following properties:

1. For any $x \in I$, $\mathbb{E}_{w \sim \mu}[|f(w,x)|] < \infty$ (**Uniformly finite moment**)
2. For almost all $w \in \infty$, the map $x \mapsto f(w,x)$ is differentiable with derivative $\frac{\partial}{\partial x} f(w,x)$ (**differentiability**)
3. There is a map $h : \Omega \to \mathbb{R}$ with the property that $\mathbb{E}_{w \sim \mu}[|h(w)|] < \infty$, such that $\left|\frac{\partial}{\partial x} f(\cdot, x)\right| \leq h$ (**derivative bound**).

Then, for any $x \in I$, $E_{w \sim \mu}\left[\left|\frac{\partial}{\partial x} f(w,x)\right|\right] < \infty$ and the function $F : x \to \mathbb{E}_{w \sim \mu}[f(w,x)]$ is differentiable with derivative

$$F'(x) = \mathop{\mathbb{E}}_{w \sim \mu}\left[\frac{\partial}{\partial x} f(w,x)\right] \tag{8.1}$$

# 9. Linaer Algebra

## 9.1 Determinant

### 9.1.1 Expansion formula

For any $A \subseteq \mathcal{Y}$,

$$\sum_{A \subseteq Y \subseteq \mathcal{Y}} \det(\boldsymbol{L}_Y) = \det(\boldsymbol{L} + \boldsymbol{I}_{\bar{A}}), \tag{9.1}$$

### 9.1.2 Rearrangement

$$\sum_{(I',J') \in \mathcal{S}(I,J)} \det(\boldsymbol{Z}_{Y,I'}) \det(\boldsymbol{Z}_{Y,J'}) \leq \sum_{(I',*) \in \mathcal{S}(I,j)} \det(\boldsymbol{Z}_{Y,I'})^2 \tag{9.2}$$

where $\boldsymbol{I}_{\bar{A}}$ is the diagonal matrix with ones in the diagonal positions with indices in $\bar{A}$ and zeros elsewhere.

### 9.1.3 Weinstein-Aronszajin identity

If $\boldsymbol{A}$ and $\boldsymbol{B}$ are matrices of size $m \times n$ and $n \times m$ respectively, given that $\boldsymbol{AB}$ is of trass class, then

$$\det(\boldsymbol{I} + \boldsymbol{AB}) = \det(\boldsymbol{I} + \boldsymbol{BA}) \tag{9.3}$$

### 9.1.4 DPP related

**Propsosal matrix for NDPP.** Given $\boldsymbol{V}, \boldsymbol{B}, \boldsymbol{D}$ such that $\boldsymbol{L} = \boldsymbol{V}\boldsymbol{V}^\top + \boldsymbol{B}(\boldsymbol{D} - \boldsymbol{D}^\top)\boldsymbol{B}^\top$, let $\{\rho_i, \boldsymbol{v}_i\}_{i=1}^K$ be the eigendecomposition of $\boldsymbol{V}\boldsymbol{V}^\top$ and $\{(\sigma_j, \boldsymbol{y}_{2j-1}, \boldsymbol{y}_{2j})\}$ be the Youla decomposition of $\boldsymbol{B}(\boldsymbol{D} - \boldsymbol{D}^\top)\boldsymbol{B}^\top$. Denote $\boldsymbol{Z} := [\boldsymbol{v}_1, \dots, \boldsymbol{v}_K, \boldsymbol{y}_1, \dots, \boldsymbol{y}_K] \in \mathbb{R}^{M \times 2K}$ and

$$\boldsymbol{X} := \operatorname{diag}\left(\rho, \dots, \rho_K, \begin{bmatrix} 0 & \sigma_1 \\ -\sigma_1 & 0 \end{bmatrix}, \dots \begin{bmatrix} 0 & \sigma_{K/2} \\ -\sigma_{K/2} & 0 \end{bmatrix}\right),$$
$$\hat{\boldsymbol{X}} := \operatorname{diag}\left(\rho, \dots, \rho_K, \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_1 \end{bmatrix}, \dots \begin{bmatrix} \sigma_{K/2} & 0 \\ 0 & \sigma_{K/2} \end{bmatrix}\right),$$

so that $\boldsymbol{L} = \boldsymbol{Z}\boldsymbol{X}\boldsymbol{Z}^\top$ and $\hat{\boldsymbol{L}} = \boldsymbol{Z}\hat{\boldsymbol{X}}\boldsymbol{Z}^\top$. Then, for every subset $\boldsymbol{Y} \subseteq [M]$, it holds that

$$\det(\boldsymbol{L}_Y) \leq \det(\hat{\boldsymbol{L}}_Y) \tag{9.4}$$

and the equality holds when the size of $\boldsymbol{Y}$ is equal to the rank of $\boldsymbol{L}$.

**Propsosal matrix for NDPP II.** Given $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{W}^A \in \mathbb{R}^{d \times d}$ Then,

$$\det([\boldsymbol{X}\boldsymbol{W}^A\boldsymbol{X}^\top]_S) \leq \det([\boldsymbol{X}\hat{\boldsymbol{W}}^A\boldsymbol{X}^\top]_S) \tag{9.5}$$

for every $S \subseteq [n]$. In addition, equality holds when $|S| \geq d$.

**DPP probability expansion [1, Lemma 2.6].**

$$\mathbb{P}_{\hat{\boldsymbol{L}}}(Y) = \frac{\det(\hat{\boldsymbol{L}}_Y)}{\det(\hat{\boldsymbol{L}} + \boldsymbol{I})} = \sum_{E \subseteq [2K], |E| = |Y|} \det(\underbrace{\boldsymbol{Z}_{Y,E}\boldsymbol{Z}_{Y,E}^\top}_{\text{elementary DPP}}) \prod_{i \in E} \frac{\lambda_i}{\lambda_i + 1} \prod_{i \notin E} \frac{1}{\lambda_i + 1} \tag{9.6}$$

1. Choose an elementary DPP according to its mixture weight
2. Sample a subset from the selected elementary DPP

**DPP probability expansion II.** The probability of sampling $S \in \binom{n}{k}$ from the $k$-DPP with $\hat{\boldsymbol{L}}$ can be decomposed into the following

$$\frac{\det(\hat{\boldsymbol{L}}_S)}{e_k(\{\lambda_i\}_{i=1}^d)} = \sum_{E \in \binom{[d]}{k}} \frac{\prod_{i \in E} \lambda_i}{e_k(\{\lambda_i\}_{i=1}^d)} \cdot \det(\boldsymbol{K}_S^E) \tag{9.7}$$

where $\boldsymbol{K}^E$ is a rank-$k$ projection matrix consisting of eigenvalues of $\hat{\boldsymbol{L}}$.

### 9.1.5 Ratio

Given that $\det(\boldsymbol{QSQ}^\top) \neq 0$

$$\frac{\det(\boldsymbol{Q}(\boldsymbol{S}+\boldsymbol{R})\boldsymbol{Q}^\top)}{\det(\boldsymbol{QSQ}^\top)} \leq \det(\boldsymbol{I}_2 + (\boldsymbol{QSQ}^\top)^{-1/2}\boldsymbol{QRQ}^\top(\boldsymbol{QSQ}^\top)^{-1/2}) \tag{9.8}$$

### 9.1.6 Inverse of trace

For an invertible matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$,

$$\operatorname{tr}(\boldsymbol{A}^{-1}) = \sum_{i=1}^{n} \det(\boldsymbol{A}_{-i})/\det(\boldsymbol{A}), \tag{9.9}$$

where $\boldsymbol{A}_{-i} \in \mathbb{R}^{n-1\times n-1}$ is the submatrix of $\boldsymbol{A}$ where the $i$th row and column of $\boldsymbol{A}$ are removed.

## 9.2 Vectorziation

$$\operatorname{vec}(\boldsymbol{AXB}) = (\boldsymbol{B}^\top \otimes \boldsymbol{A})\operatorname{vec}(\boldsymbol{X}) \tag{9.10}$$

**Lyapunov Equation.**

$$\boldsymbol{AX} + \boldsymbol{XB} = \boldsymbol{C} \tag{9.11}$$
$$\boldsymbol{AXI} + \boldsymbol{IXB} = \boldsymbol{C} \tag{9.12}$$
$$(\boldsymbol{I}\otimes\boldsymbol{A})\operatorname{vec}(\boldsymbol{X}) + (\boldsymbol{B}^\top\otimes\boldsymbol{I})\operatorname{vec}(\boldsymbol{X}) = \operatorname{vec}(\boldsymbol{C}) \tag{9.13}$$
$$\operatorname{vec}(\boldsymbol{X}) = (\boldsymbol{I}\otimes\boldsymbol{A} + \boldsymbol{B}^\top\otimes\boldsymbol{I})^{-1}\operatorname{vec}(\boldsymbol{C}) \tag{9.14}$$

## 9.3 Trace

$$\operatorname{tr}(\boldsymbol{ABC}) = \operatorname{vec}(\boldsymbol{A}^\top)^\top(\boldsymbol{I}\otimes\boldsymbol{B})\operatorname{vec}(\boldsymbol{C}) \tag{9.15}$$
$$\operatorname{tr}(\boldsymbol{A}^\top\boldsymbol{BCD}^\top) = \operatorname{vec}(\boldsymbol{A})^\top(\boldsymbol{D}\otimes\boldsymbol{B})\operatorname{vec}(\boldsymbol{C}) \tag{9.16}$$

### 9.3.1 Von Neumann's trace inequqality

**Theorem.** If $\boldsymbol{A}, \boldsymbol{B}$ are complex $n\times n$ matrices with singular values

$$\alpha_1 \geq \cdots \geq \alpha_n, \quad \beta_1 \geq \ldots \beta_n, \tag{9.17}$$

respectively, then

$$|\operatorname{tr}(\boldsymbol{AB})| \leq \sum_{i=1}^{n} \alpha_i\beta_i \tag{9.18}$$

## 9.4 Inversion

### 9.4.1 Woodbury identity

$$[\boldsymbol{A}+\boldsymbol{BCD}]^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}[\boldsymbol{C}^{-1}+\boldsymbol{DA}^{-1}\boldsymbol{B}]^{-1}\boldsymbol{DA}^{-1} \tag{9.19}$$

given that $\boldsymbol{A}^{-1}$ and $\boldsymbol{C}^{-1}$ exist. If $\boldsymbol{B}=\boldsymbol{x}, \boldsymbol{C}=\boldsymbol{I}, \boldsymbol{D}=\boldsymbol{y}^\top$

$$(\boldsymbol{A}+\boldsymbol{xy}^\top)^{-1} = \boldsymbol{A}^{-1} - \frac{(\boldsymbol{A}^{-1}\boldsymbol{x})(\boldsymbol{yA}^{-1})}{1+\boldsymbol{y}^\top\boldsymbol{A}^{-1}\boldsymbol{x}} \tag{9.20}$$

### 9.4.2 Schur Complement

Schur Complement essentially is a block Cholesky factorization of a matrix

$$\begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{bmatrix} = \begin{bmatrix}\boldsymbol{I} & \boldsymbol{BD}^{-1}\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix}\begin{bmatrix}(\boldsymbol{A}-\boldsymbol{BD}^{-1}\boldsymbol{C}) & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{D}\end{bmatrix}\begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{D}^{-1}\boldsymbol{C} & \boldsymbol{I}\end{bmatrix} \tag{9.21}$$

$\boldsymbol{A}-\boldsymbol{BD}^{-1}\boldsymbol{A}$ is called the *Schur complement* of $\boldsymbol{D}$.

## 9.5 Hadamard Product

### 9.5.1 Quadratic Relation

$$\boldsymbol{x}^\top(\boldsymbol{A}\odot\boldsymbol{B})\boldsymbol{y}=\operatorname{tr}\left(\operatorname{Diag}(\boldsymbol{x})\boldsymbol{A}\operatorname{Diag}(\boldsymbol{y})\boldsymbol{B}^\top\right) \tag{9.22}$$

By setting $\boldsymbol{x}=\boldsymbol{y}$, it shows that the Hadamard product of two PSD matrices is PSD.

### 9.5.2 Rank Relation

$$rank(\boldsymbol{A}\odot\boldsymbol{B})\leq rank(\boldsymbol{A})rank(\boldsymbol{B}) \tag{9.23}$$

### 9.5.3 Spectrum Relation

$$\prod_{i=k}^{n}\lambda_i(\boldsymbol{A}\odot\boldsymbol{B})\geq\prod_{i=k}^{n}\lambda_i(\boldsymbol{A}\boldsymbol{B}),\ \forall k=1,\dots,n \tag{9.24}$$

with $\lambda_i(\cdot)$ denotes PD matrix.

### 9.5.4 Determinant

$$|\boldsymbol{A}\odot\boldsymbol{B}|\geq|\boldsymbol{A}|\,|\boldsymbol{B}| \tag{9.25}$$

## 9.6 Matrix Calculus

### 9.6.1 Matrix Chain rule

$$[\nabla_{\boldsymbol{X}}f(g(\boldsymbol{X}))]_{ij}=\sum_{k=1}^{p}\sum_{\ell=1}^{q}\frac{\partial f(G)}{\partial g_{k\ell}}\frac{\partial g_{k\ell}}{x_{ij}} \tag{9.26}$$

### 9.6.2 Differentials

$$\mathrm{d}(\operatorname{tr}\boldsymbol{X})=\operatorname{tr}\,\mathrm{d}\boldsymbol{X} \tag{9.27}$$
$$\mathrm{d}(\boldsymbol{X}\otimes\boldsymbol{Y})=(\mathrm{d}\boldsymbol{X})\otimes\boldsymbol{Y}+\boldsymbol{X}\otimes(\mathrm{d}\boldsymbol{Y}) \tag{9.28}$$
$$\mathrm{d}\boldsymbol{X}^{-1}=-\boldsymbol{X}^{-1}\cdot\mathrm{d}\boldsymbol{X}\cdot\boldsymbol{X}^{-1} \tag{9.29}$$
$$\mathrm{d}(\det(\boldsymbol{X}))=\operatorname{tr}(\operatorname{adj}(\boldsymbol{X})\,\mathrm{d}\boldsymbol{X}) \tag{9.30}$$
$$\mathrm{d}\det(\boldsymbol{X})=\det(\boldsymbol{X})\operatorname{tr}(\boldsymbol{X}^{-1}\,\mathrm{d}\boldsymbol{X}) \tag{9.31}$$
$$\mathrm{d}\log(\det(\boldsymbol{X}))=\operatorname{tr}(\boldsymbol{X}^{-1}\,\mathrm{d}\boldsymbol{X}) \tag{9.32}$$
$$\mathrm{d}\sigma(a)=(\operatorname{Diag}(\sigma)-\operatorname{Diag}(\sigma)^2)\,\mathrm{d}a \tag{9.33}$$
$$\mathrm{d}(\operatorname{softmax}(\theta))=(\operatorname{Diag}(\boldsymbol{y})-\boldsymbol{y}\boldsymbol{y}^\top)\,\mathrm{d}\theta \tag{9.34}$$

Note: Elementwise function은 일단 Diagonal 형태로 바꿔서 생각해 보삼 ㅋ

### 9.6.3 Useful first derivatives

$$\frac{\partial \operatorname{tr} \boldsymbol{X}}{\partial \boldsymbol{X}} = \boldsymbol{I} \tag{9.35}$$

$$\frac{\partial \operatorname{tr} \boldsymbol{X}^{-1}}{\partial \boldsymbol{X}} = -\boldsymbol{X}^{-2} \tag{9.36}$$

$$\frac{\partial \operatorname{tr} (\boldsymbol{A}\boldsymbol{X})}{\partial \boldsymbol{X}} = \boldsymbol{A}^\top \tag{9.37}$$

$$\frac{\partial \operatorname{tr} (\boldsymbol{X}^k)}{\partial \boldsymbol{X}} = k \cdot (\boldsymbol{X}^\top)^{k-1} \tag{9.38}$$

$$\frac{\partial \operatorname{tr} (\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}\boldsymbol{B})}{\partial \boldsymbol{X}} = \boldsymbol{B}^\top \boldsymbol{X}^\top \boldsymbol{A}^\top + \boldsymbol{A}^\top \boldsymbol{X}^\top \boldsymbol{B}^\top \tag{9.39}$$

$$\frac{\partial \boldsymbol{A}\boldsymbol{X}^{-1}\boldsymbol{B}}{\partial \boldsymbol{X}} = -\boldsymbol{X}^{-\top} \boldsymbol{A}^\top \boldsymbol{B}^\top \boldsymbol{X}^{-\top} \tag{9.40}$$

$$\frac{\partial \log \det(\boldsymbol{X})}{\partial \boldsymbol{X}} = \boldsymbol{X}^{-\top} \tag{9.41}$$

$$\frac{\partial \det(\boldsymbol{X}^{-1})}{\partial \boldsymbol{X}} = \frac{\boldsymbol{X}^{-\top}}{\det \boldsymbol{X}} \tag{9.42}$$

$$\frac{\partial \det(\boldsymbol{X}^k)}{\partial \boldsymbol{X}} = k \det(\boldsymbol{X}^k) \boldsymbol{X}^{-\top} \tag{9.43}$$

$$\frac{\partial \log \det(\boldsymbol{X}\boldsymbol{X}^\top)}{\partial \boldsymbol{X}} = 2\boldsymbol{X}[\boldsymbol{X}^\top \boldsymbol{X}]^{-1} \cdot \det(\boldsymbol{X}\boldsymbol{X}^\top) \tag{9.44}$$

$$\frac{\det(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B})}{\partial \boldsymbol{X}} = \det(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) \boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{X}\boldsymbol{B})^{-\top} \boldsymbol{B}^\top \tag{9.45}$$

### 9.6.4 Quadratic form

$$\frac{\partial (\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^\top \boldsymbol{W} (\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})}{\partial \boldsymbol{s}} = -2\boldsymbol{A}^\top \boldsymbol{W} (\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s}) \tag{9.46}$$

$$\frac{\partial (\boldsymbol{x} - \boldsymbol{s})^\top \boldsymbol{W} (\boldsymbol{x} - \boldsymbol{s})}{\partial \boldsymbol{x}} = 2\boldsymbol{W} (\boldsymbol{x} - \boldsymbol{s}) \tag{9.47}$$

$$\frac{\partial (\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^\top \boldsymbol{W} (\boldsymbol{s} - \boldsymbol{A}\boldsymbol{s})}{\partial \boldsymbol{x}} = 2\boldsymbol{W} (\boldsymbol{s} - \boldsymbol{A}\boldsymbol{s}) \tag{9.48}$$

$$\frac{\partial (\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^\top \boldsymbol{W} (\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})}{\partial \boldsymbol{A}} = -2\boldsymbol{W} (\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s}) \boldsymbol{s}^\top \tag{9.49}$$

### 9.6.5 Hessian product rule

Given two functions $f, g : \mathbb{R}^n \to \mathbb{R}$,

$$H_c(fg) = (H_c f) g(c) + \nabla_c f^\top \nabla_c g + \nabla_c g^\top \nabla_c f + f(c) H_c g \tag{9.50}$$

### 9.6.6 Integration by parts

Given vector valued function $\varphi$ and scalar function $f$ with vanishing condition,

$$\int_{\mathbb{R}^d} \varphi(\boldsymbol{x}) \cdot \nabla f(\boldsymbol{x}) \, \mathrm{d}x = -\int_{\mathbb{R}^d} (\nabla \cdot \varphi(\boldsymbol{x})) f(\boldsymbol{x}) \, \mathrm{d}x \tag{9.51}$$

## 9.7 Eigenvalues and Eigenvectors

### 9.7.1 General Properties

Assume that $\boldsymbol{A} \in \mathbb{R}^{n \times m}$ and $\boldsymbol{B} \in \mathbb{R}^{m \times n}$,

$$\operatorname{eig}(\boldsymbol{A}\boldsymbol{B}) = \operatorname{eig}(\boldsymbol{B}\boldsymbol{A}) \tag{9.52}$$

$$\operatorname{rank}(\boldsymbol{A}) = r \Rightarrow \text{ At most } r \text{ non-zero } \lambda_i \tag{9.53}$$

### 9.7.2 Symmetric

Assume $\boldsymbol{A}$ is symmetric, then

$$\boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{I} \tag{9.54}$$
$$\lambda_i \in \mathbb{R} \tag{9.55}$$
$$\operatorname{tr}(\boldsymbol{A}^p) = \sum_i \lambda_i^p \tag{9.56}$$
$$\operatorname{eig}(\boldsymbol{I} + c\boldsymbol{A}) = 1 + c\lambda_i \tag{9.57}$$
$$\operatorname{eig}(\boldsymbol{A} - c\boldsymbol{I}) = \lambda_i - c \tag{9.58}$$
$$\operatorname{eig}(\boldsymbol{A}^{-1}) = \lambda_i^{-1} \tag{9.59}$$

For a symmetric, positive matrix $\boldsymbol{A}$

$$\operatorname{eig}(\boldsymbol{A}^\top\boldsymbol{A}) = \operatorname{eig}(\boldsymbol{A}\boldsymbol{A}^\top) = \operatorname{eig}(\boldsymbol{A}) \circ \operatorname{eig}(\boldsymbol{A}) \tag{9.60}$$

### 9.7.3 Singular Value Decomposition

Any $n \times m$ matrix $\boldsymbol{A}$ can be written as

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \tag{9.61}$$

where

$$\begin{aligned}\boldsymbol{U} &= \text{ eigenvectors of } \boldsymbol{A}\boldsymbol{A}^\top \; n \times n\\ \boldsymbol{D} &= \sqrt{\operatorname{diag}(\operatorname{eig}(\boldsymbol{A}\boldsymbol{A}^\top))} \quad n \times m\\ \boldsymbol{V} &= \text{ eigenvectors of } \boldsymbol{A}^\top\boldsymbol{A} \; m \times m\end{aligned}$$

**Square decomposed into rectangular.** Assume $\boldsymbol{V}_*\boldsymbol{D}_*\boldsymbol{U}_*^\top = 0$ then we can expand the SVD of $\boldsymbol{A}$ into

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{V} & \boldsymbol{V}_* \end{bmatrix} \left[\begin{array}{c|c} \boldsymbol{D} & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{D}_* \end{array}\right] \begin{bmatrix} \boldsymbol{U}^\top \\ \hline \boldsymbol{U}_*^\top \end{bmatrix} \tag{9.62}$$

where the SVD of $\boldsymbol{A}$ is $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top$

### 9.7.4 LU decomposition

Assume $\boldsymbol{A}$ is a square matrix with non-zero leading prcinipal minors, then

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U} \tag{9.63}$$

where $\boldsymbol{L}$ is a unique unit lower triangular matrix and $\boldsymbol{U}$ is a unique upper triangular matrix.

### 9.7.5 Cholesky decomposition

Assume $\boldsymbol{A}$ is a symmetric positive definite square matrix, then

$$\boldsymbol{A} = \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{L}\boldsymbol{L}^\top \tag{9.64}$$

where $\boldsymbol{U}$ is an unique upper triangular matrix and $\boldsymbol{L}$ is a lower triangular matrix.

### 9.7.6 Eigenvalues of its reverse

**Proposition.** Given $M \times K$ matrix $\boldsymbol{A}, \boldsymbol{B}$, the nonzero eigenvalues of $\boldsymbol{A}\boldsymbol{B}^\top \in \mathbb{C}^{M\times M}$ and $\boldsymbol{B}^\top\boldsymbol{A} \in \mathbb{C}^{K\times K}$ are identical. If addition, if $(\lambda, \boldsymbol{v})$ is an eigenpair of $\boldsymbol{B}^\top\boldsymbol{A}$ with $\lambda \neq 0$, then $(\lambda, \boldsymbol{A}\boldsymbol{v}/\|\boldsymbol{A}\boldsymbol{v}\|_2)$ is an eigenpair of $\boldsymbol{A}\boldsymbol{B}^\top$.

### 9.7.7 Row stochastic matrix

**Fact.** The operator norm of a row-stochastic matrix is 1.

## 9.8 Inverses

### 9.8.1 Rank-1 update of the inverse of inner product

Denote $\boldsymbol{A} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ and that $\boldsymbol{X}$ is extended to include a new column vector in the end $\tilde{\boldsymbol{X}} = [\boldsymbol{X}, \boldsymbol{v}]$, let $N = \boldsymbol{v}^\top (\boldsymbol{I} - \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top)\boldsymbol{v}$ then

$$(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{-1} = N^{-1} \begin{bmatrix} N\boldsymbol{A} + \boldsymbol{A}\boldsymbol{X}^\top \boldsymbol{v}(\boldsymbol{A}\boldsymbol{X}^\top \boldsymbol{v})^\top & -\boldsymbol{A}\boldsymbol{X}^\top \boldsymbol{v} \\ -\boldsymbol{v}^\top \boldsymbol{X}\boldsymbol{A}^\top & 1 \end{bmatrix} \tag{9.65}$$

### 9.8.2 Approximations

The following identity is known as the *Neuman series* of a matrix, which holds when $|\lambda_i| < 1$ for all eigenvalues $\lambda_i$

$$(\boldsymbol{I} - \boldsymbol{A})^{-1} = \sum_{n=0}^{\infty} \boldsymbol{A}^n \tag{9.66}$$

$$(\boldsymbol{I} + \boldsymbol{A})^{-1} = \sum_{n=0}^{\infty} (-1)^n \boldsymbol{A}^n \tag{9.67}$$

$$\begin{aligned} \boldsymbol{A} - \boldsymbol{A}(\boldsymbol{I} + \boldsymbol{A})^{-1}\boldsymbol{A} &= \boldsymbol{A} - \boldsymbol{A}(\boldsymbol{I} + \boldsymbol{A}^{-1})^{-1} && (9.68)\\ &= \boldsymbol{A}(\boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{A}^{-1})^{-1}) && (9.69)\\ &\approx \boldsymbol{A}(\boldsymbol{I} - \boldsymbol{I} + \boldsymbol{A}^{-1} - \boldsymbol{A}^{-2}) && (9.70)\\ &= \boldsymbol{I} - \boldsymbol{A}^{-1} && (9.71) \end{aligned}$$

### 9.8.3 Block matrix

Using Schur complements

$$\boldsymbol{C}_1 = \boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} \tag{9.72}$$

$$\boldsymbol{C}_2 = \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} \tag{9.73}$$

as

$$\left[\begin{array}{c|c} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \hline \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{array}\right]^{-1} = \left[\begin{array}{c|c} \boldsymbol{C}_1^{-1} & -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{C}_2^{-1} \\ \hline -\boldsymbol{C}_2^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & \boldsymbol{C}_2^{-1} \end{array}\right] \tag{9.74}$$

## 9.9 PSD matrix

### 9.9.1 Decomposition

1. The matrix is PSD with rank $r \iff$ there exists a matrix $\boldsymbol{B}$ of rank $r$ such that $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{B}^\top$
2. The matrix is PD $\iff$ there exists an invertible matrix $\boldsymbol{B}$ such that $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{B}^\top$
3. Given $\boldsymbol{A}$ is an $n \times n$ PSD matrix, there exists an $n \times r$ matrix $\boldsymbol{B}$ of rank $r$ such that $\boldsymbol{B}^\top \boldsymbol{A}\boldsymbol{B} = \boldsymbol{I}$.

### 9.9.2 Sylvester's characterization

$$\boldsymbol{A} \succeq 0 \iff \text{All } 2^n - 1 \text{ principal minors are nonnegative.} \tag{9.75}$$

$$\boldsymbol{A} \succ 0 \iff \text{All } n \text{ leading principal minors are positive.} \tag{9.76}$$

### 9.9.3 Eqution with zeros

Assume $\boldsymbol{A}$ is PSD, then $\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} = \boldsymbol{0} \Rightarrow \boldsymbol{A}\boldsymbol{X} = 0$

### 9.9.4 Rank of product

Assume $\boldsymbol{A}$ is positive definite, then $\text{rank}(\boldsymbol{B}\boldsymbol{A}\boldsymbol{B}^\top) = \text{rank}(\boldsymbol{B})$

### 9.9.5 Outer product

If $\boldsymbol{X} \in n \times r$, where $n \leq r$ and $\text{rank}(\boldsymbol{X}) = n$, then $\boldsymbol{X}\boldsymbol{X}^\top$ is positive definite.

### 9.9.6 Small pertubations

If $\boldsymbol{A}$ is positive definite, and $\boldsymbol{B}$ is symmetric, then $\boldsymbol{A} - t\boldsymbol{B}$ is positive definite for sufficiently small $t$.

### 9.9.7 Hadamard inequality

If $\boldsymbol{A}$ is a positive definite or semi-definite matrix, then

$$\det(\boldsymbol{A}) \leq \prod_i A_{ii} \tag{9.77}$$

### 9.9.8 Loewner order

**Fact.** Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be hermitian positive definite. Then

$$\boldsymbol{A} \succeq \boldsymbol{B} \iff \boldsymbol{I} \succeq \boldsymbol{A}^{-1/2}\boldsymbol{B}\boldsymbol{A}^{-1/2} \tag{9.78}$$

### 9.9.9 Inverse of PSD

**Fact.** Suppose that $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{A} - \boldsymbol{B}$ are all positive definite, then $\boldsymbol{B}^{-1} - \boldsymbol{A}^{-1}$ is also positive definite.

## 9.10 Symmetric and skew-symmetric matrix

### 9.10.1 Properties of symmetric matrix

1. Every real symmetric matrix can be orthogonally diagonalizable. [2]
2. The rank of a symmetric matrix $\boldsymbol{A}$ is equal to the number of non-zero eigenvalues of $\boldsymbol{A}$.
3. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are $n \times n$ real symmetric matrices that commute, then they can be simultaneously diagonalized by an orthogonal matrix.

### 9.10.2 Youla decomposition

Given $\boldsymbol{B} \in \mathbb{R}^{M \times K}$ and $D \in \mathbb{R}^{K \times K}$, consider a rank-$K$ skew-symmetric matrix $\boldsymbol{B}^\top(\boldsymbol{D} - \boldsymbol{D}^\top)\boldsymbol{B}^\top$. Then, we can write

$$\boldsymbol{B}(\boldsymbol{D} - \boldsymbol{D}^\top)\boldsymbol{B} = \sum_{j=1}^{K/2} i\sigma_j(\boldsymbol{a}_j + i\boldsymbol{b}_j)(\boldsymbol{a}_j + i\boldsymbol{b}_j)^H - i\sigma_j(\boldsymbol{a} - J - i\boldsymbol{b}_j)(\boldsymbol{a}_j - i\boldsymbol{b}_j)^H \tag{9.79}$$

$$= \sum_{j=1}^{K/2} 2\sigma_j(\boldsymbol{a}_j\boldsymbol{b}_j^\top - \boldsymbol{b}_j\boldsymbol{a}_j^\top) \tag{9.80}$$

$$= \sum_{j=1}^{K/2} \begin{bmatrix}\boldsymbol{a}_j - \boldsymbol{b}_j & \boldsymbol{a}_j + \boldsymbol{b}_j\end{bmatrix} \begin{bmatrix}0 & \sigma_j \\ -\sigma_j & 0\end{bmatrix} \begin{bmatrix}\boldsymbol{a}_j^\top - \boldsymbol{b}_j^\top \\ \boldsymbol{a}_j^\top + \boldsymbol{b}_j^\top\end{bmatrix} \tag{9.81}$$

Note that $\boldsymbol{a}_1 \pm \boldsymbol{b}_1, \ldots \boldsymbol{a}_{K/2} \pm \boldsymbol{b}_{K/2}$ are real-valued orthonormal vectors. The pair $\{(\sigma_j, \boldsymbol{a}_j - \boldsymbol{b}_j, \boldsymbol{a}_j + \boldsymbol{b}_j)\}_{j=1}^{K/2}$ is often called the **Youla decomposition** of $\boldsymbol{B}(\boldsymbol{D} - \boldsymbol{D}^\top)\boldsymbol{B}^\top$.

[2] Think of it this way: every symmetric matrix can be triangulated and normality is preserved under a similar transform. When is the triangular matrix normal? Of course, it is the diagonal matrix.

## 9.11 Some techniques

### 9.11.1 Binary analysis

어떤 matrix의 operator norm을 분석하기 위해 matrix를 binary matrix로 decomposition하는 것은 유용할 수 있다.

**Example**

Let $\boldsymbol{v}$ be the unit-normed vector that realizes the operator norm of $\boldsymbol{D}^{-1}\boldsymbol{A}$. We define the sequence of binary matrices $\boldsymbol{B}^0, \boldsymbol{B}^1, \boldsymbol{B}^2$ as follows:

$$B^t_{i,j} := 1_{\left\{2^{-t-1}\sqrt{\alpha/n} < [\boldsymbol{D}^{-1}\boldsymbol{A}]_{i,j} \le 2^{-t}\sqrt{\alpha/n}\right\}} \text{ for every integers } t \ge 0, \tag{9.82}$$

where $\sqrt{\alpha/n}$ is the upper bound for entries of $\boldsymbol{D}^{-1}\boldsymbol{A}$. Then we have the following inequalities for the entries and the $l_2$-norm:

$$[\boldsymbol{D}^{-1}\boldsymbol{A}]_{i,j} \le \sqrt{\alpha/n}\sum_{t=0}^{\infty} 2^{-t} \cdot [\boldsymbol{B}^t]_{i,j} \tag{9.83}$$

$$\left\|\boldsymbol{D}^{-1}\boldsymbol{A}\cdot\boldsymbol{v}\right\|_2 \le \sqrt{\alpha/n}\sum_{t=0}^{\infty} 2^{-t} \cdot \left\|\boldsymbol{B}^t\boldsymbol{v}\right\|_2 \tag{9.84}$$

만약 $\boldsymbol{B}^t$ matrix의 row, column들의 non-zero elements를 estimate 하면 $\left\|\boldsymbol{B}^t\boldsymbol{v}\right\|_2^2$도 estimate 할 수 있고 $\boldsymbol{D}^{-1}\boldsymbol{A}$의 operator norm bound도 estimate 할 수 있다.

# Deferred proof

Section 9.7.7 Suppose $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ for some $\lambda > 1$. Since the rows of $\boldsymbol{A}$ are nonnegative and sum to 1, each element of vector $\boldsymbol{A}\boldsymbol{x}$ is a convex combination of the components of $\boldsymbol{x}$, which can be no greater than $\boldsymbol{x}_{max}$. On the other hand, at least one element of $\lambda\boldsymbol{x}$ is greater than $\boldsymbol{x}_{max}$, which proves that $\boldsymbol{x}_{max}$, which shows that $\lambda > 1$ is impossible.

# References

[1] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *Found. Trends Mach. Learn.*, 5:123–286, 2012.

[2] T. Steinke, M. Nasr, and M. Jagielski. Privacy auditing with one (1) training run. *ArXiv*, abs/2305.08846, 2023.