1. Linear regression

① Statistical linear regression model

$\boxed{Y = X\beta^* + W}$, $W \sim N(0, \sigma^2 \cdot I_n)$

- $Y \in \mathbb{R}^n$ the response
- $X \in \mathbb{R}^{n\times d}$ the design matrix
- $\beta^* \in \mathbb{R}^d$ the features
- $W \in \mathbb{R}^n$ the Gaussian noise
- Assumption: $\text{rank}(X) = d$

Columns of X are orthonormal $\frac{1}{n}X^TX = I_d$

Remark orthonormal matrix $Q \in \mathbb{R}^{n\times n}$, columns & rows are mutually orthogonal $Q^{-1} = Q^T$, $\det(Q) = \pm 1$, $Q^TQ = QQ^T = I$, $\text{col}(Q)\cdot\text{col}(Q) = 0$, $\|Qx\| = \|x\|$.

② Least squares

$\boxed{\hat\beta = \arg\min_\beta \|Y - X\beta\|^2}$

Theorem The LS estimator $\hat\beta$ satisfies $\boxed{E[\|\beta^* - \hat\beta\|^2] = \sigma^2 \cdot d/n}$ (Proof note 1 sec 5) Statistical guarantee

Remark

- $-\frac{1}{2n}\nabla_\beta f(\beta) = \underbrace{\frac{1}{n}X^TX}_{I_d}(\beta - \beta^*) - \frac{1}{n}X^TW$
- $E[(X^TX)^{-1}X^TY] = \beta^*$
- $E[W] = 0$
- $\frac{1}{n}X^TX = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$
- $E[\hat\beta] = \beta^*$

③ Minimax optimality of LS

With $X \in \mathbb{R}^{n\times d}$, $\frac{1}{n}X^TX = I_d$, $W \sim N(0, I_n)$, we have an arbitrary estimator $\hat\beta_{LS}: \mathbb{R}^n \to \mathbb{R}^d$, $Y \to \hat\beta(Y)$, then the risk is defined as

$\boxed{R(\hat\beta; \beta^*) = E[\|\hat\beta(X\beta^* + W) - \beta^*\|^2] = \frac{d}{n}}$

Theorem For all estimator $\hat\beta: \mathbb{R}^n \to \mathbb{R}^d$, $\boxed{\sup_{\beta^* \in \mathbb{R}^d} R(\hat\beta, \beta^*) \ge \frac{d}{n}}$ — the upper bound. (Proof note 2 sec 8)

Proof strategy: $\forall t \ge 0$, $\forall \hat\beta: \mathbb{R}^n \to \mathbb{R}^d$, $E[R(\hat\beta; \beta^*)] \ge \frac{d}{n + 1/t} \xrightarrow{t\to\infty} \frac{d}{n}$. $\beta \sim N(0, t \cdot I_d)$

④ Joint Gaussian variables (interlude)

Random variables $A_1, \dots, A_m$ on same prob space are jointly Gaussian. If $\exists \mu, \Sigma$ $(A_1, \dots, A_m) \sim N(\mu, \Sigma)$.

Fact 1: linear transformation $A_1 + A_2$, $0.1A_1 + 10A_2 - 5A_3$.

Fact 2: Joint Gaussian A, B indep. $\to$ uncorrelated. $E[A\cdot B] = E[A]\cdot E[B]$

⑤ Bayes-optimal estimator

Theorem $\forall$ dist. P, $\forall$ est. $\hat\beta: \mathbb{R}^n \to \mathbb{R}^d$, $\boxed{E[R(\hat\beta;\beta)] \ge E[R(\hat\beta_P;\beta)]}$

Posterior-mean estimator

Dist. P over $\mathbb{R}^d$, $\beta \sim P$, then

$\boxed{\hat\beta_P(Y) = E[\beta \mid X\beta + W]}$ $W \sim N(0, I_d)$ (Proof note 2 sec 3)

Gaussian prior $X \in \mathbb{R}^{n\times d}$, $\frac{1}{n}X^TX = I_d$, $W \sim N(0, I_n)$, $\beta \sim N(0, t\cdot I_n)$, $\beta \perp W$, $Y = X\beta + W$, $\hat\beta_{mean} = E[\beta \mid Y]$.

To show $E[\|\hat\beta_{mean} - \beta\|^2] = d/(n + \frac{1}{t})$.

Lemma $\boxed{\{\beta \mid Y\} = N(\frac{1}{n+1/t}X^TY, \frac{1}{n+1/t}I_d)}$ imply (Proof note 2 sec 6)

Corollary $Y = X\beta + W$, $\beta \sim N(0, I_d)$, $W \sim N(0, I_n)$, then $\{X\beta \mid Y\} = N(BY, B)$, where $B = X(X^TX + I_d)^{-1}X^T$.

2. Sparse linear regression

① Definition

Similar to linear regression, but having a sparsity parameter $k \in [d]$ $d \ge 2$. $\beta^* \in \mathbb{R}^d$ is k-sparse.

$\boxed{Y = X\beta^* + W}$, $W \sim N(0, I_n)$.

Remark a vector is k-sparse if it has at most k non-zero entries.

② Best-subset-selection estimator (Brute force, NP-hard in worst case)

Known: $X \in \mathbb{R}^{n\times d}$, $k \in [d]$ $k \ll n \ll d$

Hidden: $\beta^* \in \mathbb{R}^d$, k-sparse $\le k$ non-zero

Observe: $Y = X\beta^* + W$, $W \sim N(0, I_n)$ $\binom{d}{k} \le (\frac{ed}{k})^k$

Estimator: $\boxed{\hat\beta = \arg\min\{\|X\beta - Y\|^2 \mid \beta \in \mathbb{R}^d \text{ } k\text{-sparse}\}}$

Theorem $\forall \beta^* \in \mathbb{R}^d$ k-sparse, w.h.p., $\boxed{\frac{1}{n}\|X(\hat\beta - \beta^*)\|^2 \le \frac{k}{n}\cdot O(\log\frac{2d}{k})}$

error if relevant features known; Cost of not knowing relevant features

(Proof note 3 sec 3)

Remark (on proof) 1. $f(\beta) = \frac{1}{2n}\|X\beta - Y\|^2$; 2. $\nabla f(\beta^*) = -\frac{1}{n}X^TW$; 3. $f(\beta^* + u) = f(\beta^*) + \langle\nabla f(\beta^*), u\rangle + \frac{1}{2n}\|Xu\|^2$ (taylor series); 4. $f(\beta^*) \ge f(\hat\beta)$, $u = \hat\beta - \beta^*$; 5. $\|Xu\|^2 \le 4\frac{\langle W, Xu\rangle^2}{\|Xu\|^2} \le k\cdot O(\log\frac{2d}{k})$

③ LASSO

Replace sparsity constraint with tractable constraint.

$\boxed{\hat\beta = \arg\min\{\|X\beta - Y\|^2 \mid \|\beta\|_1 \le R\}}$ $R = \|\beta^*\|_1$ canonical choice.

Remark normalization columns $X_1, \dots, X_d$ of X satisfy $\|X_i\|^2 = n$.

④ Slow rate LASSO

Theorem If $R = \|\beta^*\|_1$, then $\boxed{E[\frac{1}{n}\|X(\hat\beta - \beta^*)\|^2] \le \frac{\|\beta^*\|_1}{\sqrt{n}}\cdot O(\log 2d)^{1/2}}$ (Proof note 3 sec 6)

Remark (on proof) 1. $u = \hat\beta - \beta^*$; 2. $\frac{1}{2n}\|Xu\|^2 \le \frac{1}{n}\langle u, X^TW\rangle$; 3. Hölder $\langle u, X^TW\rangle \le \|u\|_1 \cdot \|X^TW\|_\infty$; 4. $\|u\|_1 \le 2\|\beta^*\|_1$; 5. $E[\|X^TW\|_\infty] \le O(\log d)^{1/2}\cdot\sqrt{n}$; 6. $E[\frac{1}{n}\|Xu\|^2] \le \frac{1}{\sqrt{n}}\|\beta^*\|_1 \cdot O(\log d)^{1/2}$.

⑤ Fast rate LASSO

Assume: $X \in \mathbb{R}^{n\times d}$, $S \subseteq [d]$, $|S| \le k$. $\gamma > 0$, so that (γ, S)-RE: $\forall u \in C(S)$, $\frac{1}{n}\|Xu\|^2 \ge \gamma \cdot \|u\|^2$.

$C(S) = \{u \in \mathbb{R}^d \mid \|u_S\|_1 \ge \|u_{\bar S}\|_1\}$. $\|u_S\|_1 = \sum_{i \in S}|u_i|$, $\|u_{\bar S}\|_1 = \sum_{i \notin S}|u_i|$.

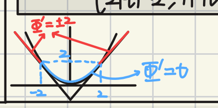
Theorem $\forall \beta^*$, $\text{support}(\beta^*) \subseteq S$, lasso $\hat\beta$ satisfies for $R = \|\beta^*\|_1$, $\forall X$, (γ, S)-RE, $\boxed{E[\frac{1}{n}\|X(\hat\beta - \beta^*)\|^2] \le \frac{k}{\gamma\cdot n}\ O(\log d)}$. (Proof note 3 sec 6)

Remark (on proof) 1. $u = \hat\beta - \beta^*$; 2. $\frac{1}{n}\|Xu\|^2 \le \|u\|_1 \cdot \|X^TW\|_\infty$; 3. $E[\|Xu\|^2] \le n\cdot O(\log d)$; 4. $E[\|Xu\|^2] \le O(\frac{k}{\gamma}\log d)$.

Combine with LS loss

$$\Phi[t] = \begin{cases} \frac{1}{2}t^2, & \text{if } |t| \le 2 \\ 2|t| - 2, & \text{if } |t| > 2 \end{cases}$$

Theorem Let $\alpha = \min_i P(|\eta_i| \le 1)$, $C = \max_i \frac{\|X_i\|}{\sqrt{d}}$ rows $X_1, \dots, X_n \in \mathbb{R}^d$ of X, suppose $n \ge O(\frac{d^2}{C^2\alpha^2})$, then $\boxed{\|\hat\beta - \beta^*\|^2 \le O(\frac{d}{\alpha^2 n})}$.



(Proof note 4 sec 4)

3. Huber loss and regression with outliers

① Linear regression with outliers

LS is sensitive to outliers. Single outlier can change outcome completely.

- $X \in \mathbb{R}^{n\times d}$, $\frac{1}{n}X^TX = I_d$
- Observation: $\boxed{Y = X\beta^* + \eta}$, $\beta^* \in \mathbb{R}^d$
- $\{\eta\} = \{-\eta\}$ symmetry $\{\eta\} = \{\eta_1\} \otimes \dots \otimes \{\eta_n\}$ independence
- $W \sim N(0, I_n)$, arbitrary $Q \in \mathbb{R}^n$
- $\eta_i = \begin{cases} W_i, & \text{with prob. } \alpha \\ +Q_i, & \text{with prob. } \frac{1}{2}(1-\alpha) \\ -Q_i, & \text{with prob. } \frac{1}{2}(1-\alpha) \end{cases}$ outliers
- $E[\eta_i] = 0$
- $E[\eta_i^2] = \alpha\cdot 1 + (1-\alpha)\cdot Q_i^2$

② Huber loss

For some $\Phi: \mathbb{R} \to \mathbb{R}^+$, we have

$\boxed{f(\beta) = \sum_i \Phi[\langle X, \beta_i\rangle - Y_i] = \Phi[X\beta - Y]}$ In LS, $\Phi[t] = t^2$

Remark (on proof) 1. $u = \hat\beta - \beta^*$; 2. $f(\beta^* + u) = \frac{1}{n}\sum_{i=1}^n \Phi[\langle X_i, u\rangle - \eta_i]$; 3. $\nabla f(\beta^* + u) = \frac{1}{n}\sum_{i=1}^n \Phi'[\langle X_i, u\rangle - \eta_i]$, where $\Phi'[t] = \text{sign}(t)\cdot\min\{|t|, 2\}$; 4. $\forall u$ with $\|u\| \le R$, $f(\beta^* + u) \ge f(\beta^*) - \|\nabla f(\beta^*)\|\cdot\|u\| + \frac{\alpha}{10}\|u\|^2$. 5. $f(\beta^*) \ge f(\beta^*) + \langle\nabla f(\beta^*), u\rangle + \frac{\alpha}{10}\|u\|^2 \to \|u\|^2 \le \frac{100}{\alpha^2}\|\nabla f(\beta^*)\|^2$ 6. $E[\|\nabla f(\beta^*)\|^2] = \underbrace{\frac{1}{n^2}\sum_{i=1}^n \|X_i\|^2 \cdot E[\Phi'[\eta_i]^2]}_{|t| \le 2} + \underbrace{\frac{1}{n^2}\langle X_i, X_j\rangle E[\Phi'[\eta_i]\cdot\Phi'[\eta_j]]}_{|t| > 2}$

$X^TX = dn$ orthogonal $\le \frac{4}{n^2}\boxed{\sum_{i=1}^n \|X_i\|^2} = \frac{4d}{n}$

③ Local quadratic lower bound

With high prob., $\forall u$ with $\|u\| \in R$, $f(\beta^*, u) \ge f(\beta^*) - \underbrace{\sqrt{\frac{400d}{n}}}_{C_1}\|u\| + \underbrace{\frac{\alpha}{10}}_{C_2}\|u\|^2$, if $C_1 \cdot R = C_2 \cdot R^2$, then $\|\hat\beta - \beta^*\|^2 \le \frac{C_1^2}{C_2} = \frac{400d}{\alpha^2\cdot n}\cdot 100 \le O(\frac{d}{\alpha^2 n})$.

$f(\beta^* + u) = f(\beta^*) + \langle\nabla f(\beta^*), u\rangle + \langle u, (\int_0^1 H[f(\beta^* + t\cdot u)]\cdot(1-t)dt)\cdot u\rangle$

where $H[f(\beta^* + t\cdot u)] = \frac{1}{n}\sum_{i=1}^n \Phi''[\langle X_i, u\rangle - \eta_i]\cdot X_iX_i^T$. $R = \frac{1}{C\cdot\sqrt{d}}$

$M(u) = \frac{1}{n}\sum_{i=1}^n (\int_0^1 [\![|t\cdot\langle X_i, u\rangle - \eta_i| \le 2]\!]\cdot(1-t)dt)\cdot\langle X_i, u\rangle^2$

$\ge \frac{1}{n}\sum_{i=1}^n [\![|\eta_i| \le 1]\!]\cdot\langle X_i, u\rangle^2 \cdot \int_0^1 (1-t)dt \; 1/2 = M^{low}(u)$

$E[M^{low}(u)] = \frac{1}{2n}\sum_{i=1}^n P(|\eta_i| \le 1)\cdot\langle X_i, u\rangle^2 \ge \frac{\alpha}{2n}\sum_{i=1}^n \langle X_i, u\rangle^2 = \frac{\alpha}{2}\|u\|^2$

With high prob., $\forall u$, $M^{low}(u) \ge \frac{\alpha}{4}\|u\|^2$.

4. Low-rank factorization models and PCA

① Single-spike model

$\boxed{Y = X^* + W = (a^*)(b^*)^T + W}$, $W_{ij} \sim N(0, 1)$

- $X^* \in \mathbb{R}^{m\times n}$, $a^* \in \mathbb{R}^m$, $b^* \in \mathbb{R}^n$ unknown (truth)
- Goal is to estimate X^* given Y
- Error measure: $\frac{1}{n\cdot m}\|\hat X - X^*\|_F^2 = \frac{1}{n\cdot m}\sum_{i,j}(\hat X_{ij} - X^*_{ij})^2$
- Error bound: $\boxed{E[\frac{1}{n\cdot m}\|a^*(b^*)^T - \hat a(\hat b)^T\|_F^2] \le \max\{\frac{1}{m}, \frac{1}{n}\}}$

② Maximum likelihood estimator

Best rank-1 approximation

$\boxed{\hat X = \arg\min\{\|X - Y\|_F^2 \mid \text{rank}(X) \le 1\}}$

- $\min \|ab^T - Y\|_F^2$ Poly-time estimator exists!
- $a \in \mathbb{R}^m$, $b \in \mathbb{R}^n$

Theorem $\boxed{E[\frac{1}{nm}\|\hat X - X^*\|_F^2] \le O(\frac{1}{n} + \frac{1}{m})}$

③ Singular vector decomposition – SVD (interlude)

$\boxed{Y = \sum_{i=1}^r \sigma_i \cdot u_i \cdot v_i}$, $\text{rank}(Y) = r$

Fact: SVD always exists and poly-time computable. Not always unique!

- Singular values: $\sigma_1 \ge \dots \ge \sigma_r > 0$
- Left singular value vectors: $u_1, \dots, u_r$ orthonormal $Y \in \mathbb{R}^{m\times n}$, $u_i \in \mathbb{R}^m$
- Right singular value vectors: $v_1, \dots, v_r$ orthonormal $v_i \in \mathbb{R}^n$

Fact: $\sigma_1 \cdot u_1 \cdot v_1^T$ is the best rank-1 approximation to Y.

Matrix norms: - Spectral norm: $\|X\| = \sigma_1$ - Frobenius norm (ℓ-2 norm) $\|X\|_F = \sqrt{\sigma_1^2 + \dots + \sigma_r^2}$

- Nuclear norm: $\|X\|_* = \sigma_1 + \dots + \sigma_r$

Inequalities: $\|Y\| \le \|Y\|_F \le \|Y\|_*$, $\|Y\|_* \le \sqrt{r}\cdot\|Y\|_F$, $\langle X, Y\rangle \le \|X\|\cdot\|Y\|_*$

Theorem: $W_{ij} \sim N(0, 1)$, $\boxed{E[\|W\|^2] \le O(n + m)}$ * Exp. of standard Gaussian $m\times n$ matrix is about $\sqrt{m} + \sqrt{n}$

Remark (on proof): 1. $Y = X^* + W$, $\text{rank}(Y) = 1$, $W_{ij} \sim N(0, 1)$; 2. $\hat X = \arg\min\{\|X - Y\|_F^2 \mid \text{rank}(X) \le 1\}$; 3. $U = \hat X - X^* \to \text{rank}(U) \le 2$; 4. $\|U\|_F^2 \le 2\langle U, W\rangle$; 5. Hölder: $\langle U, W\rangle \le \|U\|_* \cdot \|W\|$; 6. $\|U\|_* \le \sqrt{2}\|U\|_F$; 7. $\|U\|_F^2 \le 2\langle U, W\rangle \le 2\|U\|_* \cdot \|W\| \le 2\sqrt{2}\cdot\|U\|_F \cdot \|W\|$; 8. $\|U\|_F^2 \le 8\|W\|^2$; 9. $E[\|U\|_F^2] \le 8E[\|W\|^2] \le O(n + m)$ (Proof note 5 sec 5)

④ Multi-spike matrix model

$Y = X^* + W$, $\boxed{Y_{ij} = (\sum_k a^*_{ik}\cdot b^*_{jk}) + W_{ij}}$, $W_{ij} \sim N(0, 1)$

- $X^* = A^*(B^*)^T$, $\text{rank}(X^*) \le r$, unknown
- r: archetypes of user.
- a^*_{ik}: quality of movie i for user archetype k.
- user$_j$: combination of all r archetypes with weights $b^*_{j1}, \dots, b^*_{jk}$.

MLE: $\boxed{\hat X = \arg\min\{\|X - Y\|_F^2 \mid \text{rank}(X) \le r\}} = \sigma_1 \cdot u_1 \cdot v_1^T + \dots + \sigma_r \cdot u_r \cdot v_r^T$

Theorem: $\boxed{E[\frac{1}{nm}\|\hat X - X^*\|_F^2] \le r\cdot O(\frac{1}{n} + \frac{1}{m})}$

⑤ Single spike v.s. multiple spike

Single spike $Y = a^*(b^*)^T + W$ rank-1

$Y_{ij} = a^*_i(b^*_j)^T + W_{ij}$, $a^* \in \mathbb{R}^m$, $b^* \in \mathbb{R}^n$

$$\begin{bmatrix} Y_{11} & \cdots & Y_{1n} \\ \vdots & \ddots & \vdots \\ Y_{m1} & \cdots & Y_{mn} \end{bmatrix} = \begin{bmatrix} a^*_1 \\ \vdots \\ a^*_m \end{bmatrix}\begin{bmatrix} b^*_1 & \cdots & b^*_n \end{bmatrix} + W$$

Multiple spike $Y = A^*(B^*)^T + W$

$Y_{ij} = \sum_{k=1}^r A^*_{ik}B^*_{jk} + W_{ij}$, $A^* \in \mathbb{R}^{m\times r}$, $B^* \in \mathbb{R}^{r\times n}$

$$\begin{bmatrix} Y_{11} & \cdots & Y_{1n} \\ \vdots & \ddots & \vdots \\ Y_{m1} & \cdots & Y_{mn} \end{bmatrix} = \begin{bmatrix} A^*_{11} & \cdots & A^*_{1r} \\ \vdots & \ddots & \vdots \\ A^*_{m1} & \cdots & A^*_{mr} \end{bmatrix}\begin{bmatrix} B^*_{11} & \cdots & B^*_{n1} \\ \vdots & \ddots & \vdots \\ B^*_{1r} & \cdots & B^*_{nr} \end{bmatrix} + W$$

5. Matrix completion

① Formalization

Multi-spike model with some missing data.

$\boxed{Y_i = \langle X_i, \theta^*\rangle + W_i}$, rank $r \ge 2$, dimension $d \ge 2$, $n \ge 2$

- $\theta^* \in \mathbb{R}^{d\times d}$, $\text{rank}(\theta^*) \le r$, unknown
- Few missing entries of θ^*: $(X_1, Y_1), \dots, (X_n, Y_n)$
- $X_i = e_{a(i)}e^T_{b(i)}$, $a(1), \dots, a(n), b(1), \dots, b(n) \sim \text{Unif}\{1, \dots, d\}$
- $W_1, \dots, W_n \overset{iid}{\sim} N(0, \sigma^2)$

Remark $\langle X_i, \theta^*\rangle = \theta^*_{a(i), b(i)}$ Make sure every column has at least one observation

- Error measure: $\frac{1}{d^2}\|\hat\theta - \theta^*\|_F^2$
- Necessary assumptions: $n \gg r\cdot d$, $n \ge d\cdot\log d$
- MLE: $\boxed{\hat\theta = \arg\min\{\sum_i(Y_i - \langle X_i, \theta\rangle)^2 \mid \text{rank}(\theta) \le r\}}$ NP hard!

② Bernstein tail bound

$Z \in \mathbb{R}^{d\times d}$, $d \ge 2$, $E[Z] = 0$, $\{Z\} = \{Z^T\}$. **Definition**: Z satisfies (σ_Z, b_Z)-Bernstein condition if $\forall j \in \mathbb{N}_{\ge 2}$, $\boxed{\|E[Z^{j-2}\cdot ZZ^T]\| \le j!\cdot b_Z^{j-2}\cdot\sigma_Z^2}$.

Theorem: If Z satisfies (σ_Z, b_Z)-Bernstein condition, then with probability $1 - d^{-100}$, for $Z_1, \dots, Z_n \overset{iid}{\sim} \{Z\}$

$\boxed{\|\frac{1}{n}\sum_{i=1}^n Z_i\| \lesssim \sigma_Z\sqrt{\frac{\log d}{n}} + b_Z\cdot\frac{\log d}{n}}$.

③ Polynomial-time estimator

$\boxed{Y = \frac{1}{n}\sum_{i=1}^n Y_i \cdot d^2 \cdot X_i}$

- Y is a function of $\{(Y_i, X_i)\}_{i=1}^n$
- $E[Y] = \frac{1}{d^2}\sum_{a,b}^d \theta^*_{ab}\cdot d^2 \cdot e_a e_b^T = \theta^*$

Estimator: $\boxed{\hat\theta = \arg\min_{\theta \in \mathbb{R}^{d\times d}}\{\|\theta - Y\|_F^2 \mid \text{rank}(\theta) \le r\}}$ (Proof note 6 sec 6)

Theorem If $n \ge r\cdot d\log d$, w.h.p. $\boxed{\frac{1}{d^2}\|\hat\theta - \theta^*\|_F^2 \le O(\frac{(\sigma^2 + \|\theta^*\|^2_{max})\cdot r d\log d}{n})}$

$\|\theta^*\|_{max} = \max_{a,b}|\theta^*_{ab}|$. Use Bernstein tail bound to prove.

Remark (on proof): 1. Let $U = \hat\theta - \theta^*$, $W = Y - \theta^*$; 2. $\|U\|_F^2 \le 8\cdot r\cdot\|W\|^2$; 3. $W = Y - \theta^* = \frac{1}{n}\sum(Y_i\cdot d^2\cdot X_i - \theta^*) = \frac{1}{n}\sum(\boxed{\langle X_i, \theta^*\rangle\cdot d^2 X_i - \theta^*} + \boxed{W_i\cdot d^2\cdot X_i}) = \frac{1}{n}\sum(Z_i^{(1)} + Z_i^{(2)})$; 4. $W^{(1)} = \frac{1}{n}\sum Z_i^{(1)}$, $W^{(2)} = \frac{1}{n}\sum Z_i^{(2)}$; 5.* $\|W^{(1)}\|^2 \le O(\|\theta^*\|^2_{max}\cdot\frac{d^3\log d}{n})$, ** $\|W^{(2)}\| \le O(\sigma^2\cdot\frac{d^3\log d}{n})$; 6. $\|U\|_F^2 \le 8r\|W\|^2 \le 8r(\|W^{(1)}\| + \|W^{(2)}\|)^2 \le 16r\cdot(\|W^{(1)}\|^2 + \|W^{(2)}\|^2) \le O(\frac{(\sigma^2 + \|\theta^*\|^2_{max})\cdot r\cdot d^3\log d}{n})$.

* and ** at Step 5 have detailed proof.

6. Community detection

① Stochastic block model

graph G

$$\mathbb{P}\{ij \in E(G)\} = \begin{cases} (1+\varepsilon)\frac{d}{n}, & x_i^* = x_j^* \quad \text{same community} \\ (1-\varepsilon)\frac{d}{n}, & x_i^* \neq x_j^* \end{cases} \quad \forall i \neq j \in [n]$$

- $X^* \in \{\pm 1\}^n$ labels unknown
- $\varepsilon \in [0,1]$ bias, degree (#links) $d \in \mathbb{N}_{\geq 1}$

Goal: Compute $\hat{X}$ in poly-time such that w.h.p., $\boxed{\|\hat{X} - X^*\|_F^2 \leq 0.001 \cdot n^2}$, $X^* = x^*(x^*)^T$

② Matrix Bernstein inequality variant

Theorem: Let $Z_1, \cdots, Z_m$ be independent n-by-n random matrices. Suppose $\mathbb{E}[Z_i] = 0$, $Z_i = Z_i^T$. Choosing $\sigma, b > 0$ such that $\|Z_i\| \leq b$ for all $i \in [m]$ and $\boxed{\left\|\sum_{i=1}^m \mathbb{E}[Z_i Z_i^T]\right\| \leq \sigma^2}$. Then, with probability at least $1 - n^{-100}$, it holds that $\boxed{\left\|\sum_{i=1}^m Z_i\right\| \lesssim \sigma \cdot \sqrt{\log n} + b \cdot \log n}$.

③ Recovery guarantees (spectral norm)

Theorem. $\boxed{\|\hat{X} - X^*\|_F^2 \leq 0.001 \cdot n^2}$ is possible if $\boxed{d \geq O\left(\frac{\log n}{\varepsilon^2}\right)}$.

Remark (on proof): 1. $\mathbb{E}[Y] = X^*$, output a best rank-1 approximation $\hat{X}$ of Y. 2. Let A be adjacent matrix of G. $\mathbb{E}[A_{ij}] = (1 + \varepsilon \cdot X_i^* \cdot X_j^*)\frac{d}{n}$ for $i \neq j$, $Y_{ij} = \frac{n}{\varepsilon d}(A_{ij} - \frac{d}{n})$, $Y_{ii} = 1$; 3. have $Y_{ij} = \begin{cases} \frac{n}{\varepsilon d}(1 - \frac{d}{n}), & ij \in E(G) \quad A_{ij} = 1 \\ -\frac{1}{\varepsilon}, & ij \notin E(G) \quad A_{ij} = 0 \end{cases}$; 4. $\|\hat{X} - X\|_F^2 \leq 8 \cdot \|Y - X^*\|^2$; 5. Let $Z_{ij} = (Y_{ij} - X_{ij}^*)(e_i e_j^T + e_j e_i^T)$, $Y - X^* = \sum_{i<j} Z_{ij}$; 6. $\|Z_{ij}\| = |Y_{ij} - X_{ij}^*| \leq 1 + \frac{1}{\varepsilon}(\frac{n}{d} - 1) = \frac{n}{\varepsilon d} \leq b$; 7. $\mathbb{E}[Z_{ij} Z_{ij}^T] = \mathbb{E}[(Y_{ij} - X_{ij}^*)^2] \cdot (e_i e_i^T + e_j e_j^T) \leq \frac{2n}{\varepsilon^2 d} \cdot (e_i e_i^T + e_j e_j^T)$; 8. $\mathbb{E}[(Y_{ij} - X_{ij}^*)^2] \leq \mathbb{E}[Y_{ij}^2] \leq (1+\varepsilon)\frac{d}{n} \cdot \alpha_1^2 + \alpha_0^2 \leq (1+\varepsilon)\frac{n}{\varepsilon^2 d} \leq \frac{2n}{\varepsilon^2 d}$; 9. $\sum_{i<j} \mathbb{E}[Z_{ij} Z_{ij}^T] \leq \frac{2n}{\varepsilon^2 d} \cdot \sum_{i<j}(e_i e_i^T + e_j e_j^T) \leq \frac{2n^2}{\varepsilon^2 d} \cdot I_n = \sigma^2 \cdot I_n$; 10. $\|Y - X^*\| \lesssim \sigma \cdot \sqrt{\log n} + b \cdot \log n = \sqrt{\frac{2n^2 \log n}{\varepsilon^2 d}} + \frac{n \log n}{\varepsilon d} \lesssim \sqrt{\frac{n^2 \log n}{\varepsilon^2 d}}$; 11. $\|\hat{X} - X^*\|_F^2 \lesssim \frac{n^2 \log n}{\varepsilon^2 d}$. (Proof note 7 sec 4)

④ Bernstein inequality scalars

Theorem: $Z_1, \cdots, Z_N$ independent real-valued random variables. $\mathbb{E}[Z_1] = \cdots = \mathbb{E}[Z_N] = 0$. Let $\sigma, b > 0$ satisfy $\mathbb{E}[Z_1^2] + \cdots + \mathbb{E}[Z_N^2] < \sigma^2$ and $|Z_1|, \cdots, |Z_N| \leq b$. For all $\delta > 0$, with probability $1 - \delta$,

$$\boxed{|Z_1 + \cdots + Z_N| \lesssim \sigma \cdot \sqrt{\log(1/\delta)} + b \cdot \log(1/\delta)}.$$

⑤ Recovery guarantees (cut norm)

$$\boxed{\hat{X} = \operatorname{argmin}\{\|X - Y\|_F^2 \mid X \in C\}}$$

$C \subseteq \mathbb{R}^{n \times n}$, set of all rank-1 matrices

$C = \{xx^T \mid x \in \{\pm 1\}^n\}$

$C = \{uv^T \mid u, v \in \{\pm 1\}^n\}$ set of all possible choices of X^*

	$v_j=1$, $x_j=1$	$v_j=-1$, $x_j=-1$
$u_i=1$, $x_i=1$	1	-1
$u_i=-1$, $x_i=-1$	-1	1

Theorem $\boxed{\|\hat{X} - X^*\|_F^2 \lesssim \frac{n^2}{\varepsilon\sqrt{d}}}$

Remark (on proof)

$\|\hat{X} - X^*\|_F^2 \leq 2\langle \hat{X} - X^*, Y - X^* \rangle \leq 4\|Y - X^*\|_{cut}$

$\|M\|_{cut} = \max_{W \in C}\langle W, M\rangle = \max_{u,v \in \{\pm 1\}^n}\langle u, Mv\rangle$

Let $Z_{ij} = (Y_{ij} - X_{ij}^*) \cdot W_{ij}$, then $\langle W, Y - X^*\rangle = 2\sum_{i<j} Z_{ij}$, $\mathbb{E}[Z_{ij}] = 0$

$|Z_{ij}| \leq \frac{1}{\varepsilon}(\frac{n}{d} - 1) + 1 \leq \frac{n}{\varepsilon d}$

$\mathbb{E}[Z_{ij}^2] \leq \mathbb{E}[Y_{ij}^2] = (1 + \varepsilon X_{ij}^*)\frac{d}{n} \cdot \left(\frac{1}{\varepsilon^2} \cdot (\frac{n}{d} - 1)^2 - \frac{1}{\varepsilon^2}\right) + \frac{1}{\varepsilon^2} \leq (1+\varepsilon)\frac{n}{\varepsilon^2 d} + \frac{1}{\varepsilon^2} \leq \frac{3n}{\varepsilon^2 d}$.

Choose $\delta = 2^{-3n}$, $b = \frac{n}{\varepsilon d}$, $\sigma > 0$, then

$\langle W, Y - X^*\rangle \lesssim \sigma \cdot \sqrt{\log_2(1/\delta)} + b \cdot \log_2(1/\delta) = \frac{\sqrt{3}\, n^{3/2}}{\varepsilon\sqrt{d}} \sqrt{3n} + \frac{n}{\varepsilon d} \cdot 3n \lesssim \frac{n^2}{\varepsilon\sqrt{n}}$

Choose $t \lesssim \frac{n^2}{\varepsilon^2 d}$,

$\mathbb{P}\{\|Y - X^*\|_{cut} > t\} \leq \sum_{W \in C}\mathbb{P}\{\langle W, Y - X^*\rangle > t\} \leq 2^{2n} \cdot 2^{-3n} = 2^{-n}$.

Union bound

(Proof note 7 sec 6)

⑥ Grothendieck's inequality

$\boxed{\mathcal{E} = \{Z \mid Z \succeq 0, \operatorname{diag}(Z) = I_{2n}\}}$ is a set of symmetric matrices convex and poly-time separable.

Grothendieck norm:

$$\boxed{\|M\|_G = \max\left\{\left\langle Z, \begin{pmatrix} 0 & M \\ 0 & 0 \end{pmatrix}\right\rangle \,\middle|\, Z \in \mathcal{E}\right\}}$$

equal to $\langle Z_{1,2}, M\rangle$, $Z_{1,2}$ is the upper right n-by-n block of Z.

Theorem: For all matrices M, $\boxed{\|M\|_{cut} \leq \|M\|_G \leq k_G \cdot \|M\|_{cut}}$ where $k_G = [1, 2C]$ absolute const

Choose $C = \{Z_{1,2} \mid Z \in \mathcal{E}\}$ obtain a poly-time computable estimator in the family, which implies $\|M\|_{cut} \leq \|M\|_G$. We still want to have the bound $\|\hat{X} - X^*\|_F^2 \lesssim \frac{n^2}{\varepsilon\sqrt{d}}$. **Remark**: $\|\hat{X} - X^*\|_F^2 \leq 2|\langle \hat{X} - X^*, Y - X^*\rangle| \leq 4\|Y - X^*\|_G \leq 4k_G \cdot \|Y - X^*\|_{cut}$. Continue as part 5.

Remark (on theorem proof): 1. $\|M\|_{cut} \leq \|M\|_G$ simply because Grothendieck norm optimizes the same objective but larger set; 2. Show $\|M\|_G \leq k_G \|M\|_{cut}$; 3. $\|M\|_{cut} = \max_{x,y \in \{\pm 1\}^n}\langle x, My\rangle$; 4. $\max_{\|x\|_\infty \leq 1, \|y\|_\infty \leq 1}\langle x, My\rangle = \max_{x,y \in \mathbb{R}^n \setminus \{0\}}\frac{\langle x, My\rangle}{\|x\|_\infty\|y\|_\infty}$; 5. $\|M\|_G = \max_{u_1 \cdots u_n v_1 \cdots v_n \in B^{2n}} M_{ij}\langle u_i, v_j\rangle$, $B^{2n} = \{w \in \mathbb{R}^{2n} \mid \|w\|_2 \leq 1\}$; 6. $\|M\|_G = \sum_{i,j} M_{i,j}\langle u_i, v_j\rangle$, $x_i = \langle g, u_i\rangle$, $y_i = \langle g, v_j\rangle$, $\mathbb{E}[\langle x, My\rangle] = \mathbb{E}[\sum_{i,j} M_{i,j} x_i y_j] = \sum_{i,j} M_{ij}\langle u_i, \mathbb{E}[gg^T] v_j\rangle = \|M\|_G \leq a\|M\|_G + c\|M\|_{cut}$; 7. $x_i^{\leq\tau} = \operatorname{clamp}_\tau(x_i)$, $x_i^{>\tau} = x_i - x_i^{\leq\tau}$, $y_j^{\leq\tau} = \operatorname{clamp}_\tau(y_j)$, $y_j^{>\tau} = y_j - y_j^{\leq\tau}$, $\operatorname{clamp}_\tau(z) = \begin{cases} z, & |z| < \tau \\ \tau \cdot \operatorname{sign}(z), & \text{otherwise} \end{cases}$ $\|M\|_G = \mathbb{E}[\langle x, My\rangle] = \mathbb{E}[\langle x^{\leq\tau}, My^{\leq\tau}\rangle + \langle x^{>\tau}, My^{\leq\tau}\rangle + \langle x^{\leq\tau}, My^{>\tau}\rangle + \langle x^{>\tau}, My^{>\tau}\rangle] = \mathbb{E}[R]$; (Proof note 7 sec 8) 8. $\mathbb{E}[R] = \langle W, \begin{pmatrix} 0 & M \\ 0 & 0 \end{pmatrix}\rangle \leq W_{11} \cdot \|M\|_G$; 9. $w_1 = \begin{pmatrix} \lambda x^{>\tau} \\ \lambda^{-1} y^{\leq\tau} \end{pmatrix}$, $w_2 = \begin{pmatrix} \lambda^{-1} x^{\leq\tau} \\ \lambda y^{>\tau} \end{pmatrix}$, $w_3 = \begin{pmatrix} x^{>\tau} \\ y^{>\tau} \end{pmatrix}$, $W = \mathbb{E}[w_1 w_1^T + w_2 w_2^T + w_3 w_3^T]$, $W_{ii} = \lambda^{-2}\mathbb{E}[\operatorname{clamp}_\tau(x_i^2)] + (\lambda^2 + 1)\mathbb{E}[(|x_i| - \tau)_+^2]$; 10. $\mathbb{E}[(|x_i| - \tau)_+^2] = 2\int_\tau^\infty (x - \tau)^2 \gamma(x)\,dx \leq \max_{x > \tau}\frac{\gamma(x)}{\gamma(x - \tau)} \leq e^{-\tau^2/2}$.

7. Non-negative matrix factorization & topic models

① Topic model

- Document $D_1 \cdots D_n$
- Topics $A_1 \cdots A_k$ unknown
- Topic proportion $W_1 \cdots W_n \in \mathbb{R}_{\geq 0}^k$ sum to 1 unknown

Bag of words:

- $D_l(i) = \frac{\#\text{ occurence of word } i \text{ in doc } l}{\#\text{ word in doc } l} \in \mathbb{R}_{\geq 0}^d$
- $A_j(i) =$ significance of word i for topic $j \in \mathbb{R}_{\geq 0}^d$

Statistical model

- $\boxed{M_l = \sum_{j=1}^k W_l(j) \cdot A_j}$
- D_l is avg. of L-size sample of M_l
- $\boxed{\mathbb{E}[D_l] = M_l}$

Matrix view $\boxed{M = AW}$

- $M = [M_1 \cdots M_n] \in \mathbb{R}^{d \times n}$
- $A = [A_1 \cdots A_k] \in \mathbb{R}^{d \times k}$
- $W = [W_1 \cdots W_n] \in \mathbb{R}^{k \times n}$
- $D = [D_1 \cdots D_n] \to \boxed{\mathbb{E}[D] = M}$

Best rank-k approximation

Does not work! For rotation matrix $R \in \mathbb{R}^{k \times k}$, we have $M = (AR)(R^{-1}W)$, giving negative number, mixed topics.

② Non-negative matrix factorization

"Gorbachev" and "virus" are anchor word.

$$M = \underbrace{\begin{array}{c|ccc} & \text{Cold war} & \text{Finance} & \text{Pandemic} \\ \hline \text{"Gorbachev"} & 0.2 & 0 & 0 \\ \text{"Stocks"} & 0 & 0.3 & 0 \\ \text{"Virus"} & 0 & 0 & 0.4 \\ \text{"Government"} & 0.8 & 0.7 & 0.6 \end{array}}_{A} \times \underbrace{\begin{array}{c|cc} & \text{doc1} & \text{doc2} \\ \hline \text{Cold war} & 0.6 & 0 \\ \text{Finance} & 0.4 & 0.3 \\ \text{Pandemic} & 0 & 0.7 \end{array}}_{W} = \begin{array}{c|cc} & \text{doc1} & \text{doc2} \\ \hline \text{"Gorbachev"} & \boxed{0.12} & 0 \\ \text{"Stocks"} & 0.12 & 0.09 \\ \text{"Virus"} & 0 & \boxed{0.28} \\ \text{"Government"} & 0.76 & 0.63 \\ & \text{sum to 1} & \text{sum to 1} \end{array}$$

Definition: Any pair (A, W) with $A \in \mathbb{R}_{\geq 0}^{d \times k}$, $W \in \mathbb{R}_{\geq 0}^{k \times n}$, and $M = AW$.

Column stochastic matrix has all columns sum up to 1. Row stochastic matrix has all rows sum up to 1. If M has rank-k NMF and M is column stochastic, then exists rank-k NMF for M with both factors col. stochastic.

③ Separability

- Word i is an anchor for topic j, row i of A is multiple of e_j.
- Separable A: every topic has anchor word, i.e., the rows of A contain $e_1, \cdots, e_k$ up scaling

Theorem: Suppose $M = AW$ for a separable A and full-rank W. Then, given M, there exists an algorithm that computes the columns of A in polynomial time. Assume M, A, W are column/row stochastic w.l.o.g.

Algorithm: Step 1: rescale the rows of M to sum to 1. Step 2: Find the corners of the convex hull of the rows of M and set them to be the rows of W. Step 3: solve $M = AW$ to get A.

④ Error bound analysis

$D = [D_1 \cdots D_n] \in \mathbb{R}^{d \times n}$, draw L-size sample M_i, set $D_i(j) = \frac{1}{L}$ (count of word j in doc i). We estimate $\boxed{\frac{1}{n}MM^T = A(\frac{1}{n}MM^T)A^T}$.

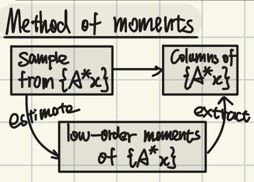
Gram matrix: $G = \frac{1}{n}DD^T = \frac{1}{n}\sum_{k=1}^n D_k D_k^T$, $Q = \mathbb{E}[G] = \frac{1}{n}\sum_{k=1}^n \mathbb{E}[D_k D_k^T]$, $D_k(i) = \frac{1}{L}\sum_{l=1}^L D_{k,i,l}$. $\mathbb{E}[D_k(i) D_k(i')] = \frac{1}{L}M_k(i) \cdot [\![i = i']\!] + (1 - \frac{1}{L}) \cdot M_k(i) \cdot M_k(i') \longrightarrow \mathbb{E}[D_k D_k^T] = \frac{1}{L} \cdot \operatorname{diag}(M_k) + \frac{L-1}{L} \cdot M_k M_k^T \longrightarrow$ unbiased estimator $Y_k = \frac{L}{L-1}(D_k D_k^T - \frac{1}{L} \cdot \operatorname{diag}(D_k))$, such that $\mathbb{E}[Y_k] = M_k M_k^T \longrightarrow Y_k = M_k M_k^T + Z_k$, $\mathbb{E}[Z_k] = 0$ and $Z_1 \cdots Z_k$ independent. For $n \gg d$, using concentration bound, $\|\frac{1}{n}\sum_{k=1}^n Z_k\|$ is small w.h.p., $\|\frac{1}{n}\sum_{k=1}^n Y_k - \frac{1}{n}M_k M_k^T\| = \|\frac{1}{n}\sum_{k=1}^n Z_k\|$. Hence, $\frac{1}{n}\sum_{k=1}^n Y_k$ is an estimator of $\frac{1}{n}MM^T$.

③ Independent component analysis

Example: In a party, we want to take the d messy, mixed conversation audio to separate them back to d original, clean, individual voice.

- Independent source $x \in \mathbb{R}^d$
- Mixing matrix $A^* \in \mathbb{R}^{d \times d}$, $\operatorname{rank}(A^*) = r$
- $\boxed{y_1 \cdots y_n \overset{iid}{\sim} \{A^* x\} \text{ or } Y = A^* X}$

Method of moments



$u \in \mathbb{R}^d$, $v \in (A^*)^T u \in \mathbb{R}^r$, $\mathbb{E}[x] = 0$, $\mathbb{E}[xx^T] = I_r$, $d = r$

First moment: $\mathbb{E}[\langle u, A^* x\rangle] = \langle u, A^* \mathbb{E}[x]\rangle = 0$

Second moment: $\mathbb{E}[\langle u, A^* x\rangle^2] = v^T \mathbb{E}[XX^T] v = \|v\|^2 = \|u\|^2$

Third moment: $\mathbb{E}[\langle u, A^* x\rangle^3] = \sum_{i,j,k=1}^d v_i v_j v_k \mathbb{E}[X_i X_j X_k] = \sum_{i=1}^d v_i^3 \mathbb{E}[X_i^3]$

Can be used to recover the columns of mixing matrix A^*.

Fourth moment: $\mathbb{E}[\langle u, A^* x\rangle^4] = \sum_{i,j,k,l=1}^d v_i v_j v_k v_l \mathbb{E}[X_i X_j X_k X_l] = \sum_{i=1}^d v_i^4 \cdot \mathbb{E}[X_i^4] + 6\sum_{i,j=1, i<j}^d v_i^2 v_j^2 \mathbb{E}[X_i^2 X_j^2] = \sum_{i=1}^d v_i^4(\mathbb{E}[X_i^4] - 3) + 3\|u\|^4 = F(u)$

$M_3 = \sum_{i=1}^d \mathbb{E}[x_i^3] \cdot (A_i^*)^{\otimes 3}$, $M_4 = \mathbb{E}[(A^* x)^{\otimes 4}]$ (Proof note 9 sec 6)

8. Tensor decomposition

order	0	1	2	3	...
	scalar	col/row	matrix	...	...

① Tensors (interlude)

A tensor is an array of numbers with multiple indices. The order of a tensor is the number of its indices.

Tensor product: $(a \otimes a')_{i,i'} = a_i \cdot a'_{i'}$; $a^{\otimes k} = \overbrace{a \otimes \cdots \otimes a}^{k \text{ times}}$; $\langle a \otimes a', b \otimes b'\rangle = \langle a, b\rangle \cdot \langle a', b'\rangle$

Contraction: $a^T \cdot b = \langle a, b\rangle$; $(a^T \otimes I) \cdot (b \otimes b') = \langle a, b\rangle \cdot b'$; $(a^T \otimes I) \cdot (b \otimes b') = \langle a, b\rangle \cdot (b \otimes b')$; for $i \in [r]$, let $M_i \in \mathbb{R}^{p_i \times q_i}$, rank-1 tensors $v_1 \otimes \cdots \otimes v_r \in \mathbb{R}^{p_1} \otimes \cdots \otimes \mathbb{R}^{p_r}$, then $(M_1 \otimes \cdots \otimes M_r) \cdot (v_1 \otimes \cdots \otimes v_r) = M_1 v_1 \otimes \cdots \otimes M_r v_r \in \mathbb{R}^{q_1} \otimes \cdots \otimes \mathbb{R}^{q_r}$,

② Random contraction algorithm

Theorem: Given an order-3 tensor $X \in (\mathbb{R}^d)^{\otimes 3}$ that admits a decomposition of the form $\boxed{X = a_1^{\otimes 3} + \cdots + a_r^{\otimes 3}}$ for orthonormal vectors $a_1 \cdots a_r \in \mathbb{R}^d$, we can efficiently compute the set of vectors $\{a_1, \cdots, a_r\}$. (Proof note 9 sec 5)

Theorem (asymmetric): Let $X \in \mathbb{R}^{d_1} \otimes \mathbb{R}^{d_2} \otimes \mathbb{R}^{d_3}$ be an order-3 tensor that admits a decomposition of the form $\boxed{X = a_1 \otimes b_1 \otimes c_1 + \cdots + a_r \otimes b_r \otimes c_r}$, where $a_1 \cdots a_r \in \mathbb{R}^{d_1}$, $b_1 \cdots b_r \in \mathbb{R}^{d_2}$ are orthonormal, $c_1 \cdots c_r \in \mathbb{R}^{d_3}$ are pairwise distinct. Can efficiently compute up to sign.

Remark (on uniqueness of SVD): SVD is not always unique. But if a a matrix has singular values that are all strictly positive and pairwise distinct, its left and right singular values are guaranteed to be unique.

Remark (on asymmetric proof): 1. Choose $g \sim N(0, I_{d_3})$. 2. Compute contraction $M = (I_{d_1} \otimes I_{d_2} \otimes g^T) \cdot X$. 3. $M = \sum_{i=1}^r \underbrace{\langle g, c_i\rangle}_{\text{singular}} \cdot \underbrace{a_i \otimes b_i}_{\text{orthonormal}}$. 4. $\{|\langle g, c_i\rangle| : i \in [r]\}$, c_i pairwise distinct.

④ Gaussian mixture model

Idea: K-means clustering has limitations, using moments with Gaussian mixture models to recover the true cluster means.

- Centers $v_1^*, \cdots, v_k^*$ linearly independent, $v_j^* \in \mathbb{R}^d$ (unknown)
- Cluster proportion $w_1, \cdots, w_k$ the weight for cluster. $\sum_{j=1}^k w_j = 1$.
- Observation $y_t = v_{I^*(t)}^* + w_t$, $w_t \sim N(0, I_d)$.

First moment: $M_1 = \mathbb{E}[\frac{1}{N}\sum_{t=1}^N y_t] = \sum_{j=1}^k w_j v_j^*$

Second moment: $M_2 = \mathbb{E}[\frac{1}{N}\sum_{t=1}^N y_t^{\otimes 2}] = \sum_{j=1}^k w_j (v_j^*)^{\otimes 2} + I_d$

$\longrightarrow P = M_2 - I_d = \sum_{j=1}^k w_j (v_j^*)^{\otimes 2}$

Third moment: $M_3 = \mathbb{E}[\frac{1}{N}\sum_{t=1}^N y_t^{\otimes 3}] = \sum_{j=1}^k w_j (v_j^*)^{\otimes 3} + C(M_1)$

$\to S = M_3 - C(M_1) = \sum_{j=1}^k w_j (v_j^*)^{\otimes 3}$ (Proof note 9 sec 7)

For othogonality **Whitening**: $U = (P^{1/2})^+$, $v_j' = U v_j^*$

Transformation & contraction,

$v_j' = c_j \hat{v}_j'$ $\begin{cases} S' = (U \otimes U \otimes U) \cdot S = \sum_{j=1}^k w_j (v_j')^{\otimes 3} \\ M = (g^T \otimes I_k \otimes I_k) \cdot S' = \sum_{j=1}^k \langle g, v_j'\rangle \cdot w_j \cdot v_j'(v_j')^T \end{cases}$

Scale recovery via linear system

$S_1' = UM_1 = \sum_{j=1}^k w_j v_j' = \sum_{j=1}^k w_j c_j \hat{v}_j'$

$\left.\begin{array}{l} a_j = (\hat{v}_j')^T P' \hat{v}_j' = w_j c_j^2 \|\hat{v}_j'\|_2^4 \\ b_j = \langle S_1', \hat{v}_j'\rangle = w_j c_j \|\hat{v}_j'\|_2^4 \end{array}\right\} c_j = \frac{a_j}{b_j \|\hat{v}_j'\|_2^2}$

$\longrightarrow \hat{v}_j^* = P^{1/2} v_j'$

9. Sum-of-squares

Purpose of SOS: Convert NP-hard estimation problem into efficient, poly-time algorithms.

① Sum-of-square proof system

Definition: – X vector/matrix of N indeterminates over the reals.

– $P_1(X), \cdots, P_m(X),\ q(X) \in \mathbb{R}[X]$ polynomials

A degree-l SOS proof of the statement "$q(X) \geq 0$ over $\{P_1(x) \geq 0, \cdots, P_m(X)\}$" is any finite decomposition $\boxed{q(X) = \sum_{r=1}^{R} S_r(X)^2 \cdot \bar{P}_r(X)}$, where

– $\bar{P}_r$ is product of subset of $\{P_1(X), \cdots, P_m(X)\}$.

– degree$(S_r(X)^2 \cdot \bar{P}_r(X)) \leq l$. Notation: $\boxed{\vec{P}(X) \geq 0 \ \vdash_l^X \ q(X) \geq 0}$

② Useful SOS list (Proof note 11 sec 8)

– $\vdash_2^{a,b}\ a \cdot b \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$

– $\vdash_2^{a,b} \langle a, b\rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$

– $\vdash_4^{a,b} \langle a, b\rangle^2 \leq \|a\|^2 + \|b\|^2$

– $\vdash_2^{x} \langle x, Mx\rangle \leq \|M\| \cdot \|x\|^2$

– $\vdash_4^{M,x} \langle x, Mx\rangle \leq \frac{1}{2}(1 + \|M\|_F^2) \cdot \|x\|^2$

– $\vdash_t^{z} z_1 \cdots z_t \leq \frac{1}{t}(z_1^t + \cdots + z_t^t)$

– $\vdash_t^{A,B} (A+B)^t \leq 2^{t-1} \cdot (A^t + B^t)$

– $X^2 \leq C \cdot X \vdash_2^X X < C$, for $C > 0$

③ SOS meta theorem

There exists an algorithm that given a vector of N-variate polynomials $\vec{P}(X) \in \mathbb{R}[X]^m$ and numbers $l \in \mathbb{N}$, $\varepsilon > 0$ outputs in time $\boxed{(N + m + \log(1/\varepsilon))^{O(l)}}$ a vector $\hat{X} \in \mathbb{R}^N$ with the following property: whenever there exists a solution $X^* \in \mathbb{R}^N$ satisfying $\{\vec{P}(X^*) > 0\}$ and for some $C \in \mathbb{R}$, $\boxed{\{\vec{P}(X) \geq 0\} \vdash_l^X \{\|X - X^*\|^2 \leq C\}}$, then $\hat{X}$ satisfies $\boxed{\|\hat{X} - X^*\|^2 \leq C + \varepsilon}$

Additional assumptions: 1. The bit-complexity of $\vec{P}$, X^*, and polynomials in the decomposition of SOS are bounded by $\boxed{(N+m)^{O(l)}}$. 2. Observations $Y = [y_1 \cdots y_n]$ do not appear explicitly. We use $\{\vec{P}(X, Y) \geq 0\}$ then we substitute Y to Y after we observe Y. Finally get a new poly system $\boxed{\{\vec{P}(x) \geq 0\} = \{\vec{P}(X, Y) \geq 0\}}$.

Corollary: Add auxiliary variables for the ploynomial system to increase the odds to have SOS. There exists an algorithm that given a vector of $N + N'$-variate polynomials $\vec{P}(X, Z) \in \mathbb{R}[X, Z]^m$ and numbers $l \in \mathbb{N}$, $\varepsilon > 0$ output in time $\boxed{(N + N' + m + \log(1/\varepsilon))^{O(l)}}$ a vector $\hat{X} \in \mathbb{R}^N$ with property: whenever there exists a solution $X^* \in \mathbb{R}^N$, $Z^* \in \mathbb{R}^{N'}$ satisfying $\{\vec{P}(X^*, Z^*) \geq 0\}$ and for some $C \in \mathbb{R}$, $\boxed{\{\vec{P}(X, Z) > 0\} \vdash_l^{X,Z} \{\|X - X^*\|^2 \leq C\}}$, then $\hat{X}$ satisfies $\boxed{\|\hat{X} - X^*\|^2 \leq C + \varepsilon}$. (Proof note 10 sec 3)

10. Clustering, GMM, and SOS

① GMM for clustering $\boxed{Y = X^* + W}$, $W_{ij} \overset{iid}{\sim} N(0, I_d)$

– Unknown: centers $\mu_1^*, \cdots, \mu_k^* \in \mathbb{R}^d$; assignment $X_1^*, \cdots, X_n^* \in \{\mu_1^*, \cdots, \mu_k^*\}$. rank$(X^*) \leq k$

– Observe: $Y_1, \cdots, Y_n$ with $Y_i = X_i^* + W_i$, $W_1, \cdots, W_n \overset{iid}{\sim} N(0, 1)$

– MLE, $\hat{X} = \text{argmin}\{\|X - Y\|_F \mid X$ has $\leq k$ dist. columns$\}$. NP-hard

Theorem. If sample size is large enough so that $n \geq (100d)^t$ for some even $t \in \mathbb{N}$ with $t \geq 4$, then the MLE $\hat{X}$ satisfies w.h.p. $\boxed{\frac{1}{n}\|\hat{X} - X^*\|_F^2 \leq 4t \cdot k^{4/t}}$ $\Leftarrow \Theta(\log k)$

(Proof note 11 sec 5)

Remark: Exponential improvement over PCA. PCA error is $\Theta(k \cdot \frac{d+n}{n})$. Allows us to cluster it means have min. separation $\gg \log k$. Make MLE guarantees efficient.

② Hölders inequality (interlude) *Cauchy-Schwarz generalized

$\forall a_1, \cdots, a_n, b_1, \cdots, b_n \in \mathbb{R}$, $t > 1$, $\boxed{\sum_{i=1}^n a_i \cdot b_i \leq \left(\sum_{i=1}^n |a_i|^t\right)^{1/t} \cdot \left(\sum_{i=1}^n |b_i|^{\frac{t}{t-1}}\right)^{1 - 1/t}}$

Cauchy-Schwarz is a special case where $t = 2$.

Jensens inequality $\mathbb{E}[X] \leq (\mathbb{E}[X^t])^{1/t}$, $t \geq 2$, $\mathbb{E}[X] = \sum_{i=1}^n P_i \cdot x_i \leq \Big(\underbrace{\sum_{i=1}^n (P_i^{1/t} \cdot x_i)^t}_{\mathbb{E}[X^t]}\Big)^{1/t} \cdot \Big(\underbrace{\sum_{i=1}^n (P_i^{1 - 1/t})^{\frac{t}{t-1}}}_{1}\Big)^{1 - 1/t}$.

③ Gaussian (sample) moments

Theorem. If $n \geq (100d)^t$ for even t in the range $2 \leq t \leq d/2$, then w.h.p., $\forall v \in \mathbb{R}^d$, $\boxed{\left(\frac{1}{n}\sum_{i=1}^n \langle W_i, v\rangle^t\right)^{1/t} \leq \sqrt{t} \cdot \|v\|}$, $W_i \overset{iid}{\sim} N(0, I_d)$.

11. Robust mean estimation

① Model of corruption

naive estimator $\hat{\mu}(z_1 \cdots z_n) = \frac{1}{n}\sum_{i=1}^n z_i$

– Unknown: $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$, $\Sigma \leq I$

– Distribution: $D \overset{iid}{\sim} N(\mu, \Sigma)$

– Sample: $y_1^*, \cdots, y_n^* \overset{iid}{\sim} D$, $Y^* = (y_1^*, \cdots, y_n^*) \in \mathbb{R}^{d \times n}$

– ε-corruption: $C_\varepsilon(Y^*) = \{Z \in \mathbb{R}^{d \times n} \mid Y^* - Z$ has non-zero cols$\}$

– Adversial error: For estimator $\hat{\mu}(Z)$, we have $\sup\{\|\hat{\mu}(Z) - \mu\| \mid Z \in C_\varepsilon(Y^*)\}$. unbounded.

② An inefficient estimator (Proof note 12 sec 2)

Lemma. Let Y and Y' be two random vectors suppose that $\mathbb{P}(Y \neq Y') \leq \varepsilon$ with $\varepsilon \leq 1/4$, and $\text{Cov}(Y) - I_d$ and $\text{Cov}(Y') \leq I_d$. Then $\boxed{\|\mathbb{E}[Y] - \mathbb{E}[Y']\| < O(\sqrt{\varepsilon})}$.

Theorem. There exists an estimator $\hat{\mu}$ such that w.h.p. (over Y^*), the adversarial error is at most $\boxed{O(\sqrt{\varepsilon})}$.

Remark (on proof): 1. $u = Y - Y'$, $\bar{u} = \mathbb{E}[Y] - \mathbb{E}[Y']$, $\|u\|^2 \leq (\mathbb{E}[\mathbb{1}]_{u \neq 0}^2]^{1/2} (\mathbb{E}[\langle \bar{u}, u\rangle^2])^{1/2}$. 2. $(\mathbb{E}\langle \bar{u}, u\rangle^2)^{1/2} \leq 2\|\bar{u}\| + \|\bar{u}\|^2$. 3. $\|\bar{u}\|^2 \leq \sqrt{\varepsilon}(2\|\bar{u}\| + \|\bar{u}\|^2)$, $\|u\| \leq 4\sqrt{\varepsilon}$. 4. $\|\frac{1}{n}\sum_{i=1}^n y_i^* - \mu^*\| \leq O(\sqrt{\varepsilon})$, $\text{Cov}(\text{Unif}\{y_1^*, \cdots, y_n^*\}) \leq 2 \cdot I_d$, $\mathbb{P}(Y \neq Y') \leq 2\varepsilon$. 5. $\|\frac{1}{n}\sum_{i=1}^n y_i^* - \hat{\mu}(Z)\| \leq O(\sqrt{\varepsilon})$. 6. $\|\mu^* - \hat{\mu}(Z)\| \leq \|\mu^* - \frac{1}{n}\sum_{i=1}^n y_i^*\| + \|\frac{1}{n}\sum_{i=1}^n y_i^* - \hat{\mu}(Z)\| \leq O(\sqrt{\varepsilon})$. (Proof note 12 sec 2)

③ An efficient estimator

Consider polynomial system A in variables $Y_1, \cdots, Y_n \in \mathbb{R}^d$, $\mu \in \mathbb{R}^d$, $W_1, \cdots, W_n \in \mathbb{R}$, and $R \in \mathbb{R}^{d \times d}$, defined as following. R is used to enforce the constraint $\frac{1}{n}\sum_{i=1}^n (Y_i - \mu)(Y_i - \mu)^T \leq 2 I_d$.

$$\boxed{A = \begin{cases} W_i^2 = W_i,\ i \in [n] \\ W_i(Z_i - Y_i) = 0,\ i \in [n] \\ \Sigma_{i=1}^n W_i \geq (1 - \varepsilon)n \\ \mu = \frac{1}{n}\Sigma_{i=1}^n Y_i \\ \frac{1}{n}\Sigma_{i=1}^n (Y_i - \mu)(Y_i - \mu)^T + R^T R = 2 I_d \end{cases}}$$

Theorem. $\boxed{A \vdash_4 \|\mu - \mu^*\|^2 < O(\sqrt{\varepsilon})}$

Remark (SOS): $\vdash_4^{a,b} \langle a, b\rangle^2 = \langle a \otimes b, b \otimes a\rangle \leq \frac{1}{2}\|a\|^2\|b\|^2 + \frac{1}{2}\|a\|^2\|b\|^2 = \|a\|^2 \cdot \|b\|^2$

$\{X^2 \leq CX\} \vdash X = (\frac{1}{\sqrt{C}}X) \cdot \sqrt{C} \leq \frac{X^2}{2C} + \frac{C}{2} \leq \frac{X}{2} + \frac{C}{2}$.

(Proof note 12 sec 3)

Appendix: Preliminaries

① Linear algebra & calculus

Inner product, for $\forall u, v \in \mathbb{R}^n$, $\boxed{\langle u, v\rangle = u^T v = \sum_{i=1}^n u_i v_i}$

*Inner product $= 0 \to$ orthogonal

Norms: l_p-norm $\boxed{\|v\|_p = \left(\sum_{i=1}^n v_i^p\right)^{1/p}}$

* l_1 norm $\|v\|_1 = \sum_{i=1}^n |v_i|$ (lasso)

l_2 norm is Euclidean norm (ridge)

l_∞ norm $\|v\|_\infty = \max_{i \in [n]} |v_i|$

Cauchy-Schwarz: $\langle u, v\rangle \leq \|u\| \cdot \|v\|$

Holder: $\langle u, v\rangle \leq \|u\|_p \cdot \|v\|_q$, $\frac{1}{p} + \frac{1}{q} = 1$

Jensen: $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$ convex

$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$ concave

Matrix: $A \in \mathbb{R}^{m \times n}$, m rows, n columns

Trace: $\text{Tr}(A) = \sum_{i=1}^n A_{ii}$, $A = \mathbb{R}^{n \times n}$

Cyclicity: $\text{Tr}(ABC) = \text{Tr}(CAB) = \text{Tr}(BCA)$

Linearity: $\text{Tr}(aA + bB) = a\text{Tr}(A) + b\text{Tr}(B)$

Trace & inner product, $\langle A, B\rangle = \text{Tr}(AB^T) = \text{Tr}(A^T B)$

Example, $\langle x, Mx\rangle = x^T M x$ ($\in \mathbb{R}^{1 \times n}$)

$= \text{Tr}(x^T M x) = \text{Tr}(M x x^T)$

$= \text{Tr}(M^T x x^T) = \langle M, x x^T\rangle$

Orthogonal matrices: $U \in \mathbb{R}^{n \times n}$ is orthogonal if its rows & columns are orthogonal unit vectors – also called orthonormal. * $U^T U = U U^T = I_n$

$U^{-1} = U^T$, for $v \in \mathbb{R}^n$, we have $\|Uv\| = \|v\|$.

Positive semidefinite (PSD): $A \in \mathbb{R}^{n \times n}$ is PSD, i.e., $A \succeq 0$, iff $x^T A x \geq 0$ for all $x \in \mathbb{R}^n$. A satisfies all eigenvalues of A are non-negative, $\langle A, M\rangle \geq 0$ for all PSD matrices M, unique PSD $C \in \mathbb{R}^{n \times n}$, $A = CC$.

Positive definite (PD): $A \in \mathbb{R}^{n \times n}$ is PD, i.e., $A \succ 0$ iff $x^T A x > 0$ for all non-zero $x \in \mathbb{R}^n$. A satisfies all eigenvalues of A are positive, $\langle A, M\rangle > 0$ for all PD matrices M, unique PD $C \in \mathbb{R}^{n \times n}$, $A = CC$.

Eigen decomposition: $A \in \mathbb{R}^{n \times n}$, a vector-scalar pair $(w, \lambda) \in \mathbb{R}^n \times \mathbb{R}$ is an eigenvalue-eigenvector-pair of the matrix A if $\boxed{Aw = \lambda w}$.

Taylor series: $f(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \frac{f'''(a)}{3!}(x - a)^3 + \cdots$

Exponential series: $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!}$

Logarithmic expansion: $\ln(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots$, $|x| < 1$

Geometric series: $\sum_{k=0}^{\infty} ar^k = a + ar + ar^2 + \cdots = \frac{a}{1 - r}$, $|r| < 1$.

Bound for binomial term: $\binom{n}{k} \leq \frac{n^k}{k!} < \frac{n^k}{(k/e)^k} = \left(\frac{en}{k}\right)^k$

② Concentration bounds

Markov inequalities: For X a non-negative RV, for $t > 0$, $\boxed{\mathbb{P}(X > t) \leq \frac{\mathbb{E}[X]}{t}}$. * Extension: For X a RV with mean μ, for $t > 0$, $\boxed{\mathbb{P}(|X - \mu| \geq t) \leq \frac{\mathbb{E}[|X - \mu|^k]}{t^k}}$.

Sub-Gaussian RVs: A RV X with mean μ is σ-sub-Gaussian if $\forall \lambda \in \mathbb{R}$, $\boxed{\mathbb{E}[e^{\lambda(X - \mu)}] \leq e^{\sigma^2 \lambda^2 / 2}}$. Sub-Gaussian tail bound: For X a sub-Gaussian RV with mean μ, for $t > 0$ $\boxed{\mathbb{P}(X - \mu \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)}$

Hoeffding inequality: For X_i, $i = 1, \cdots, n$ independent sub-Gaussian RV with μ_i and σ_i, for $t \geq 0$, $\boxed{\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mu_i) \geq t\right) \leq \exp\left(-\frac{nt^2}{2\frac{1}{n}\sum_{i=1}^n \sigma_i^2}\right)}$. For $X_i \cdots X_n$ be independent RVs with μ_i such that for each $i = 1, \cdots, n$, it holds that $a_i \leq X_i \leq b_i$ almost surely, for $a_i, b_i \in \mathbb{R}$, $\boxed{\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mu_i) \geq t\right) \leq \exp\left(-\frac{2nt^2}{\frac{1}{n}\sum_{i=1}^n (b_i - a_i)^2}\right)}$.

Sub-exponential RVs: A RV X with mean μ is (ν, α)-sub-exponential if $\forall |\lambda| < \frac{1}{\alpha}$, $\boxed{\mathbb{E}[e^{\lambda(X - \mu)}] \leq e^{\nu^2 \lambda^2 / 2}}$. Sub-exponential tail bound: $\boxed{\mathbb{P}(X - \mu \geq t) \leq \begin{cases} \exp\left(-\frac{t^2}{2\nu^2}\right), & 0 \leq t \leq \frac{\nu^2}{\alpha} \\ \exp\left(-\frac{t}{2\alpha}\right), & t > \frac{\nu^2}{\alpha} \end{cases}}$

Sum tail bound: For X_i, $i = 1, \cdots, n$, independent sub-exponential RVs with μ and (ν_i, α_i) for $t \geq 0$, $\boxed{\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mu_i) \geq t\right) \leq \begin{cases} \exp\left(-\frac{nt^2}{2\nu_*^2/n}\right), & 0 \leq t \leq \frac{\nu_*^2}{n\alpha_*} \\ \exp\left(-\frac{nt}{2\alpha_*}\right), & t > \frac{\nu_*^2}{n\alpha_*} \end{cases}}$ $\nu_* = \sqrt{\sum_{i=1}^n \nu_i^2}$, $\alpha_* = \max_{i \in [n]} \alpha_i$

Bernstein condition: A RV X with μ, σ^2 satisfies Bernstein condition with parameter b if, for $k = 2, 3, 4 \cdots$ $\boxed{|\mathbb{E}[(X - \mu)^k]| \leq \frac{1}{2}k! \sigma^2 \cdot b^{k-2}}$. *Also X is a sub-exponential with parameters $(\sqrt{2}\sigma, 2b)$.

Bernstein inequality: For a X that satisfies Bernstein condition, we have for all $|\lambda| < \frac{1}{b}$, for $t \geq 0$, $\boxed{\mathbb{E}[e^{\lambda(X - \mu)}] \leq \exp\left(\frac{\lambda^2\sigma^2/2}{1 - b|\lambda|}\right)}$. Furthermore, $\boxed{\mathbb{P}(X - \mu \geq t) \leq \exp\left(-\frac{t^2}{2(\sigma^2 + bt)}\right)}$.