# Statistical Learning Theory summary

Michael Van Huffel

October 11, 2023

Statistical Learning Theory summary created by *michavan@student.ethz.ch*

This summary has been written based on the Lecture 252-0526-00 S Statistical Learning Theory by Prof. J. M. Buhmann (Spring 2023). This serves as complementary summary to the lecture script which is also allowed to take to the exam. There is no guarantee for completeness and/or correctness regarding the content of this summary. Use it at your own discretion

## Information Theory

**Surprise:** $S(\mathbb{P}(A)) : [0,1] \to [0,\infty)$. $S(1)=0$, $S(u)>S(v) \to u<v$, continuous, $S(uv)=S(u)+S(v)$ if indip, $S(u) = -c\log u$, $c$ const **Entropy:** $H(P(x)) = H(X) = \mathbb{E}_{x\sim P}[S] = -\sum_x P(x)\log P(x)$ Cont: H can $\le 0$ unif $(0,1/2)$, Discr: $0 \le H \le \log N$, $H=0$ iff $\exists x: p(x)=1$, $N = |\mathcal{X}|$, $H_N(p(x_1)..p(x_N)) = H_{N-1}(p(x_1)+p(x_2),..)$ $+(p(x_1)+p(x_2))H_2(\frac{p(x_1)}{p(x_1)+p(x_2)}, \frac{p(x_2)}{p(x_1)+p(x_2)})$

**BinH:** $H_{bin}(\delta) = -\delta\log\delta - (1-\delta)\log(1-\delta)$

**ConH** $H(Y|X) = \int P(X=x)H(Y|X=x)dx = \int p(x)\int p_{Y|X}(y,x)\log p_{Y|X}(y,x)dydx$ $H(X,Y)=H(X)+H(Y|X)$ (CR)

**MI** $I(X;Y)=I(Y;X)=H(X)-H(X|Y) = \mathbb{E}_{X,Y}[\log\frac{p(X,Y)}{p(X)p(Y)}]$, $I(X;X)=H(X)$

**ConI** $I(X;Y|Z) = H(X|Z)-H(X|Y,Z)$ $I(X,Y;Z) = H(X,Y)-H(X,Y|Z) = H(X)+H(Y|X)-H(X|Z)-H(Y|X,Z) = I(X;Z)+I(Y;Z|X)$ Given $X \to Y \to Z$: $I(X;Z)\le I(X;Y)$ **Chain:** $I(X;Y,Z) = I(X;Y)+I(X;Z|Y) = I(X;Z)+I(X;Y|Z)$. $I(X_1,...X_n;Z) = \sum_{i=1}^n I(X_i;Z|X_1,...X_{i-1})$

**DPi:** $I(X;Y|Z)\ge 0$ and $I(X;Z|Y)=0$ $X_1\to..\to X_n$, $I(X_1;X_2..X_n) = I(X_1;X_2)$

**KL** $D_{KL}(p\|q) = \sum_x p(x)\log(\frac{p(x)}{q(x)})$

**Msc** $I(X;Y) = KL(p_{X,Y}(x,y)\|p_X(x)p_Y(y))$, $H(X|Y)\le H(X)$, $H(X,Y|Z)\ge H(X|Z)$, $H(5X)=H(X)$ if $X$ discrete, $>H(X)$ cont. $H(X,Y,Z)-H(X,Y)\le H(X,Z)-H(X)$, $H(X|Y)\ge H(X|Y,Z)$, $H(X,Y)\le H(X)+H(Y)$, $H(g(X))\le H(X)$, $X$ discr. $I(X,Y;Z) \ge I(X;Z), I(X;Z|Y) = -I(Z;Y)+I(Z;Y|X)+I(X;Z), I(g(X),Y) \le I(X;Y)$ **Codes:** source code $C$ for RV $X$ is map $\mathcal{X}\to D^\star$, domain $X$, set of finite length strings from D-ary alphabet.

**Exp length:** $L(C)=\sum_{x\in\mathcal{X}}P(x)l(x)$

**Prefix code:** no codeword is prefix of other cw. **Kraft ineq:** prefix code over |alphabet| $= D$: $\sum_{i=1}^m D^{-l_i}\le 1$ ($l_i$ = length code word)

**Opt. codes:** minimize $L=\sum_{i=1}^m p_i l_i$ s.t. $\sum_{i=1}^m D^{-l_i}\leqslant 1 \Rightarrow L^\star = -\sum_1^m p_i\log p_i$

**Inform. Bottle:** efficient code $X\mapsto C$ Preserve relevant info on context variable $Y$ $\mathcal{R}^{IB} = I(X;C)-\lambda I(C;Y), \lambda>0$

**Scheme:** Obj space → Data groups (by minze mutual info) → Minze cost → Get $c^{opt}$ → Constr on context MI → Feat space

**Rate dist theory:** $d:\mathcal{X}\times\hat{\mathcal{X}}\to\mathbb{R}^+$, w\ $\mathbb{E}_{x,\hat x}[d(x,\hat x)] = \sum_{x\in\mathcal{X}}\sum_{\hat x\in\hat{\mathcal X}} p(\hat x|x)p(x)d(x,\hat x)$

**Shannon/Kolmogorov:** opt repr satisfies $R(D) = \min_{\{p(\hat x|x):\mathbb{E}_{x,\hat x}[d(x,\hat x)]<D\}} I(x,\hat x)$ $\mathrm{L}(p(\hat x|x)) = I(x,\hat x)+\beta(\mathbb{E}_{x,\hat x}[d(x,\hat x)]-D)$ $\frac{\delta L}{\delta p(\hat x|x)} = 0 \Rightarrow p(\hat x|x) = \frac{p(\hat x)}{Z(x,\beta)}\exp(-\beta d(x,\hat x))$

## Max Entropy: $p(c|X)\propto\exp(-\beta R(c,X))$

**Requirements for** $p(\cdot|X)$: $R(c_1,X)\le R(c_2,X) \to p(c_1|X)\ge p(c_2|X)$, Jaynes's (see below) $p_\beta(c) = \exp[-\beta\cdot(R(c)-F(\beta))]$,

**Free energy** $F(\beta) = \mathbb{E}[R(c,X)] - T\cdot H = -\frac{1}{\beta}\log Z(\beta)$. **Gibbs free energy:** $G(p) = \sum_{c\in C}p(c)R(c)+\frac{1}{\beta}\sum_c p(c)\log p(c)$, $G(p) = \frac1\beta KL(p\|p_\beta)+F(\beta)$

Measurements: $r_j(x), 1\le j\le m$, yield constraints $\mathbb{E}\{r_j(X)\} = \mu_j$. **Jaynes's principle** $\sup_{p(X)}\{-\int_{\mathcal X}p(x)\log p(x)dx\}$, $\int_{\mathcal X}p(x)dx=1$ $p(x)\ge0$, $\int_{\mathcal X}p(x)r_j(x)dx = \mu_j, 1\le j\le m$ $J(p+\delta p) = \int_x(p+\delta p)\cdot (-\log(p(1+\delta p/p)) + \lambda_0 - \sum_{j=1}^m\lambda_j r_j(x))dx$. *Variation equal zero:* $p(x) = \frac1Z\exp(-\sum_1^m\lambda_j r_j(x))$, $Z=\int_x\ldots$ **TTL '84:** $0 = \int_x\frac{\partial p(x)}{\partial\mu_i}dx$, $1 = \int_x\frac{\partial p(x)}{\partial\mu_i}r_i(x)dx$ $1\le\sqrt{\mathbb{E}[(\frac{\partial\log p(x)}{\partial\mu_i})^2]}\sqrt{\mathrm{Var}[r_i(x)]}$, estimated error $\mu_i\sim\sqrt{\mathrm{Var}\{r_i\}}\Rightarrow\delta\mu_i = \beta_i\sqrt{\mathrm{Var}\{r_i\}}$ $\beta_i\le\sqrt{\mathbb{E}[(\frac{\delta_{\mu_i}p(x)}{p})^2]}$, eq. if $\frac{\partial\ln p(x)}{\partial\mu_i}\sim r_i(x)-\mu_i$

Distrib with min sensitivity: $\frac{\partial\ln p(x)}{\partial\mu_i} = \alpha_i(r_i(x)-\mu_i)$, $\alpha_i$ constants. Cov matrix $C_{i,j} = \mathbb{E}(r_i(x)-\mu_i)(r_j(x)-\mu_j)$ is diag, $\alpha_i(\mu_1,\ldots,\mu_m)$ dep only on $\mu_i$ → Gibbs distr → min sensitiv to change moments $\mu_i$.

**Kmeans:** $R^{\mathrm{km}}(c,\mathbf{Y},\mathbf{X}) = \sum_{i\le n}\|x_i-y_{c(i)}\|^2$ $\frac{\partial}{\partial y_\alpha}R^{\mathrm{km}}(\ldots) = -2\sum_{i:c(i)=\alpha}(x_i-y_\alpha) = 0$ $y_\alpha = \frac{1}{n_\alpha}\sum_{i:c(i)=\alpha}x_i$ w\ $n_\alpha = \#\{i:c(i)=\alpha\}$

**Centroid equation:** posterior $p(c|X,\theta)$ depends on $\theta$. ME for $\theta$, $S$ = entropy: $\frac{\partial}{\partial\theta}S = -\sum_c(\frac{\partial}{\partial\theta}p(c\mid x,\theta))\log p(c\mid x,\theta) + \sum_c p(c|x,\theta)\frac{\frac{\partial}{\partial\theta}p(c|x,\theta)}{p(c|x,\theta)} = \beta\sum_c R(\frac{\partial}{\partial\theta}p(c\mid x,\theta))$ $\frac{\partial}{\partial\theta}\beta\sum_c Rp(c\mid x,\theta) = \frac{\partial}{\partial\theta}\mu = 0$, $\mu$ is guaranteed by Lagr variabl $\beta$ (not dep on $\theta$) $\frac{\partial}{\partial\theta}\sum_c Rp(c|\ldots) = 0 \to \sum_c p(c|\ldots)(\frac{\partial}{\partial\theta}R) = 0$

**DA:** stop splitting via CV, MDL, PA

## Markov Chains $\sum_{c'}P(c,c')=1$

$\mathrm{P}(X_{t+1}=c'|X_t=c) = P(c,c')$, **Irreducible:** can go from any stato to any state (finite steps), **Periodic:** if state $i$ is visited after number of steps multiple of integer $d>1$ ($d=1$ aperiodic), **Stationarity:** $\sum_{c\in\mathcal C}\pi(c)P(c,c') = \pi(c')$. If all above satisfied, $\lim_{t\to\infty}\mathbb{P}[X_t=c] = \pi(c)$, $\lim_{t\to\infty}\frac1t\sum_1^t f(X_s) = \sum_c\pi(c)f(c)$, **Detailed balnc** $\pi(c)P(c,c') = \pi(c')P(c',c)$

**Mixing time** of MC $\propto\frac{1}{\lambda_1-\lambda_2}$ where $\lambda_1=1$ is first eig and $\lambda_2\le\lambda_1$ is second eig of $P$.

## Least Angle Clustering $e_i := X_i/|X_i|$

$S(x_i,x_j) = w_{ij}\cos\phi_{ij} = w_{ij}e_ie_j$, $w_{ij} = v_i\cdot v_j$ $v_i = 1/\sqrt{p_k n}, p_k = \frac1n\sum_i M_{ik}, v_iv_j = \frac{1}{p_\alpha n} = \bullet$ $R(M,X) = -\frac12\sum_\nu^k\sum_i^n\sum_j^n M_{i\nu}M_{j\nu}v_iv_j\cos\phi_{ij}$ $+\frac12\sum_\nu^k\sum_i^n M_{i\nu}^2v_i^2 = k/2$

$Z = \sum_{\{M\}}\prod_\nu^k\exp(\frac{\beta p_\nu n}{2}\|\frac{1}{p_\nu n}\sum_{i=1}^n M_{i\nu}e_i\|^2)$ $y_\alpha = \frac{1}{p_\alpha n}\sum_{i=1}^n e_ip_{i\alpha}$, $p_\alpha = \frac1n\sum_{i=1}^n p_{i\alpha}$

**Laplace:** $f(x^*)=0, I = \int d\mathbf{x}\,h(\mathbf{x})e^{-nf(\mathbf{x})}$ $f(\mathbf{x})\approx f(x^*)+\frac12(x-x^*)^\top H_f(x^*)(x-x^*)$ $I\approx e^{-nf(x^\star)}\int h(x^\star)e^{-\frac n2(x-x^*)^\top H_f(x^*)(x-x^*)}$ $H = Q^\top\Lambda Q, H^{1/2} = Q\Lambda^{1/2}, y = \sqrt n H^{1/2}(x-x^\star), x = y\frac{H^{-1/2}}{\sqrt n}+x^\star$ Jacobian, $n^{-d/2}\det(\Lambda)^{1/2}$ $I\approx e^{-nf(x^*)}h(x^*)\int\mathrm{d}y\,e^{-\frac{\|y\|^2}{2}}\cdot|\det|J_y|| = \frac{e^{-nf(x^*)}h(x^*)\cdot(2\pi)^{d/2}}{n^{d/2}\sqrt{\det(H_f(x^*))}}$ **Gibbs distribution** $p(M|X) = \frac1z\exp(\frac{\beta}{2n}\sum_{\nu=1}^k\frac{1}{p_\nu}(\sum_{i=1}^n M_{i\nu}e_i)^2)$

**BW:** $e^{b^2/2a^2} = \int_{\mathbb R}\frac{a}{\sqrt{2\pi}}\exp(-\frac{a^2x^2}{2}+bx)dx$

Nominator: $\sqrt{\beta n/\pi}^{dk}\prod_\nu^k p_\nu^{\frac d2}\int_{\mathbb{R}^d}dy_1\cdots dy_k$ $\exp(-\beta n(\frac12\sum_\nu p_\nu y_\nu^2 - \frac1n\sum_\nu y_\nu\sum_i M_{i\nu}e_i))$ $p_\alpha^*y_\alpha^* = \frac1n\sum_{i=1}^n e_ip_{i\alpha}^*$, $p_{i\alpha}^* = \frac{\exp(\beta e_iy_\alpha^*)}{\sum_{\nu=1}^k\exp(\beta e_iy_\nu^*)}$ $Z = \int dy_1..\int dy_k\exp(-\beta nf_X(y)), p(M|X)\approx \frac{\exp(-\beta n(\frac12\sum_i p_\nu^*y_\nu^{*2} - \frac1n\sum_\nu y_\nu^*\sum_i M_{i\nu}e_i))}{\exp(-\beta n(\frac12\sum_\nu^* p_\nu^* y_\nu^* - \frac{1}{\beta n}\sum_{i=1}^n\log\sum_{\nu=1}^n\exp(\beta e_iy_\nu^*)))} = \exp(\beta\sum_\nu y_\nu^*\sum_i M_{i\nu}e_i - \sum_i^n\log\sum_\nu^n\exp(\beta e_iy_\nu^*))$

**MAP:** $\frac1n\sum_i M_{i\alpha}e_i = \frac1n\sum_{i=1}^n e_ip_{i\alpha}^*$

## Histo Cluster $\Delta_{i,j} = 1(0)$ if $i=j (i\ne j)$

$n$ obj, $m$ feat, $l$ dyads $\{(x_{i(r)},y_{j(r)})\}_{r=1}^l$

**Generative model:** 1. select $x_i\in\mathcal X$ with unif $1/n$; 2. choose clust using clust membership $c(i)$ of $x_i$; select $y_j\in\mathcal Y$ from $q(y_j|c(i))$. $\hat p(x_i,y_j) = \frac1l\sum_{r=1}^l\Delta_{x_i,x_{i(r)}}\Delta_{y_j,y_{j(r)}}$, $\hat p(y_j|x_i) = \frac{\hat p(x_i,y_j)}{\hat p(x_i)} = \frac{\hat p(x_i,y_j)}{\frac1l\sum_{r=1}^l\sum_{j=1}^m\Delta_{x_i,x_{i(r)}}\Delta_{y_j,y_{j(r)}}}$

$\mathcal L = \prod_{i=1}^n\prod_{j=1}^m\mathbf{P}(x_i,y_j|c(i),q)^{l\hat p(x_i,y_j)}$ $\mathbf P(x_i,y_j|c,q) = q(y_j|c(i))p(c(i))$, $p(\alpha) = 1/k$ $R^{\mathrm{hc}}(c;\{q(\cdot|\alpha)\}) = -\log(\mathcal L)$, $x_i\to i, y_j\to j$ $R^{\mathrm{hc}}\to R^{\mathrm{hc}}+l\sum_i\sum_j \hat p(j|i)\hat p(i)\log\hat p(j|i) = l\sum_i\hat p(i)\sum_m\hat p(j|i)(\log\hat p(j|i)-\log q(j|c(i))) - l\sum_i\sum_j\hat p(j|i)\hat p(i)\log(p(c(i))) = l\sum_i\hat p(i)\sum_j \hat p(j|i)\log(\frac{\hat p(j|i)}{q(j|c(i))}) - l\sum_i\hat p(i)\log p(c(i))$

$\hat p(i)\approx 1/n$ since obj dawn from unif distrib $R^{hc}(c,q,\hat p)\propto -l\sum_i^n\sum_j^m\hat p(i,j)\log q(j|c(i))$

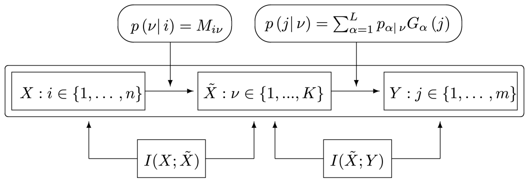
**Centr. cond:** $\mathbb{E}_{c|\hat p}\frac{\partial}{\partial q(j|\alpha)}R^{hc}(c,q,\hat p) = 0 = -\sum_i\sum_j l\hat p(i,j)\mathbb{E}_{c|\hat p}\mathbb{I}_{\{c(i)=\alpha\}}\frac{\partial}{\partial q_{j\alpha}}\log q_{j\alpha} + \frac{\partial}{\partial q_{ja}}\sum_\nu\lambda_\nu\sum_j(q_{j\nu}-1)$, $p_i(\alpha|Q,\mathcal X)$

**Limitations:** lack topology (permutations bin index do not change KL, neglect info due to noise-induced errors during the histogramming process ca), Histograms repr cat info while most feat spaces equipped w\ natural topo (feature similarity), cant impose structure (idk nothing about pixels by themselves).

**PDC:** replace empirical histogram centroids by prototypical distribution parametrized by Gaussian mixture (more robust centroids), assume given set of obj $\mathbf o_i, i\in\{1..n\}$, $M_{i\nu}=1$, if obj $\mathbf o_i$ assigned to cluster $\nu = 1..k$, $\sum_{\nu<k}M_{i\nu}=1$. Each obj $\mathbf o_i$ equipped with set of $n_i$ obs $\mathcal X_i = \{x_{i1},\ldots,x_{in_i}\}, x_{ij}\in\mathbb R^d$, $p(x|\nu) = \sum_{\alpha=1}^l p_{\alpha|\nu}g(x|\mu_\alpha,\Sigma_\alpha)$, prob of various groups $p_\nu$, feat val restr to specific domains → rectified Gauss. Domain $I = \cup_{j=1}^m I_j$, $I_j\cap I_s = \emptyset$ for $j\ne s$, region weight $G_a(j) = \int_{I_j}g_a(y)dy$,

$\Theta = \{p_\nu, p_{\alpha|\nu},\mu_\alpha|\alpha=1,\ldots,l;\nu=1,\ldots k\}$ $\mathcal L = \prod_{i\le n}\sum_{\nu\le k}M_{i\nu}p_\nu p(\mathcal X_i|\nu,\Theta) = \prod_{i\le n}\prod_{\nu\le k}[p_\nu p(\mathcal X_i|\nu,\Theta)]^{M_{i\nu}}$ $n_{ij} = \#$ observation at site $i$ is in $I_j$, $\mathcal L =$

$\prod_{i\le n}\prod_{\nu\le k}[p_\nu\prod_{j\le m}(\sum_{\alpha\le l}p_{\alpha|\nu}G_\alpha(j))^{n_{ij}}]^{M_{i\nu}}$ $\mathcal{R}^{phc}(c,\{p(a|\nu)\},\theta) = -\log\mathcal{L}$, two-part coding scheme: expected codelength when encoding the cluster memberships and encoding the individual feat values. $\mathrm{E}[\mathcal{R}] = -\sum_i\sum_\nu q_{i\nu}[\log p_\nu + \sum_j n_{ij}\log(\sum_\alpha p_{\alpha|\nu}G_\alpha(j))]$
**EM:** *E-step:* $h_{i\nu} = -\log p_\nu - \sum_j n_{ij}\log(\sum_\alpha p_{\alpha|\nu}G_\alpha(j))$, $q_{i\nu} = \mathbb{E}[M_{i\nu}] = p(M_{i\nu} = 1) \propto \exp(-\frac{h_{i\nu}}{T})$, *M-step:* $p_\nu = \frac{1}{n}\sum_i q_{i\nu}$
$X$ prob of choose obj $i$, assume $p_i = 1/n$




$H(\tilde{X}|X) = -\sum_i p_i\sum_\nu M_{i\nu}\log M_{i\nu} = 0$
$I(X;\tilde{X}) = H(\tilde{X}) = -\sum_{\nu<k}p_\nu\log p_\nu$ with $p_\nu = \sum_{i=1}^n p(C=\nu|X=i)p(X=i)$, $M_{i\nu}$
$H(Y|\tilde{X}) = -\sum_{\tilde{X},Y}p(Y,\tilde{X})\log p(Y|\tilde{X}) = -\sum_{\tilde{X},Y}\sum_X p(Y,\tilde{X},X)\log p(Y|\tilde{X})$,
$\mathcal{R}^{IB} = -\sum_i^n\sum_\nu^k M_{i\nu}[\log p_\nu + \frac{\lambda}{n_i}\sum_{j=1}^m n_{ij}\log(\sum_{\alpha=1}^L p_{\alpha|\nu}G_\alpha(j))]$ , $\lambda H(Y)$ const in $\Theta$

## Graph Clustering $S_{ij} = \exp(-D_{ij}/\Delta)$

OR: $S_{ij} = \max_{ij}D_{ij} - D_{ij}$, $W$ weight matrix of $(\mathcal{V},\mathcal{E})(A,B\subset\mathcal{V}, A\cap B=\emptyset)$
$\mathrm{cut}(A,B) = \sum_{i\in A,j\in B}W_{i,j}$, $\mathrm{assoc}(A,\mathcal{V}) = \sum_{i\in A,j\in\mathcal{V}}W_{i,j}$ (measures total connection strength from nodes A to all nodes)
**SCC:** modify CC by choosing continuous similarities ($S_{ij}\in[-1,+1]$) and sum them up relative to threshold $u$, $\frac{1}{2}(|X|\pm X) = \max\{0,\pm X\}$, $\mathcal{R}^{cc}(c;\mathcal{D}) = -\frac{1}{2}\sum_{\nu\le k}\sum_{(i,j)\in\mathcal{E}_{\nu\nu}}(|S_{ij}-u| + S_{ij}-u) - \frac{1}{2}\sum_{\nu\le k}\sum_{\substack{\mu\le k\\ \mu\ne\nu}}\sum_{(i,j)\in\mathcal{E}_{\nu\mu}}(|S_{ij}+u| - S_{ij}-u)$.
**PC & DA:** given $S = (S_{ij})_{i,j\le N}\in\mathbb{R}^{N\times N}$ specifies how much $i$ likes $j$. Want to cluster in $K$ groups based on personality, estimate hm person like another based on personality type. Cluster assignment matrix $M = (M_{ik})_{i<N.k<K}$ , $r = \{r_{kk'}\}_{k,k'<N}$ (hm person type $k$ likes type $k'$, $\sum_{k\le K}M_{ik} = 1$
$R(M,r) = \sum_{ij\le N,kk'\le K}M_{ik}M_{jk'}(S_{ij}\text{-}r_{kk'})^2$
With $r$ fix, Lagrangian $\mathcal{L}(\mathbf{P}(\cdot\mid r),\lambda_0,\lambda_1) = H[\mathbf{P}(\cdot|r)] + \lambda_0(\mathbb{E}[R]\text{-cst}) + \lambda_1(\sum_M\mathbf{P}(M|r)\text{-}1)$

$\mathbf{P}^*(S,M|r)\propto\exp(-R(M,r)/T)$, compute now values for $r$ that maximize $H[\mathbf{P}^*(\cdot|r)]$
$\arg\max_r H[\mathbf{P}^*(\cdot|r)] = \arg\max_r F[R(\cdot,r)]$
w/ $F[R(\cdot,r)] = \log(\sum_M\exp(-\frac{R(M,r)}{T}))$
EM to compute ($\star$) $= \arg\max\log\tilde{\mathbf{P}}(S|r)$
w/ $\tilde{\mathbf{P}}(S\mid r)$ is the marginal pdf of ($\star$)
$\sum_M\int_S\tilde{\mathbf{P}}(S,M|r)dS = \sum_M\prod_{i,j}\int_{S_{ij}}\prod_{k,k'}\exp(-M_{ik}M_{jk'}(S_{ij} - r_{kk'})^2/T)dS_{ij}$ **EM:** *E-step* compute $\tilde{\mathbf{P}}(\cdot\mid S,r_0) = \mathbf{P}^*(\cdot|r)$, *M-step* compute $r^* = \arg\max_r Q(r,r_0)$, w/ $Q(r,r_0) = \mathbb{E}_{M\sim\mathbf{P}^*(\cdot|r_0)}[\log\tilde{\mathbf{P}}(S,M\mid r')] = -\frac{1}{T}\sum_{ij}\sum_{kk'}\mathbb{E}_{M\sim\mathbf{P}^*(\cdot|r_0)}[M_{ik}M_{jk'}](S_{ij}\text{-}r_{kk'})^2$
$r_{kk'}^{(\ell+1)} = \frac{\sum_{i,j\le N}\mathbf{P}^*(M_{ik}M_{jk'}=1|r^{(\ell)})S_{ij}}{\sum_{i,j\le N}\mathbf{P}^*(M_{ik}M_{jk'}=1|r^{(\ell)})}$ cant use, norm const of ($\star$) is $\Theta(2^{N\times K})$ **MFA:** ($\star$) $\approx\mathbf{q}(\cdot|r) = \arg\min_q D_{KL}(q\|\mathbf{P}^*(\cdot\mid r))$ with $\mathbf{q}(M|r) = \prod_{i\le N,k\le K}\mathbf{q}_{ik}(M_{ik}\mid r)$, $\mathbf{q}_{ik}(M_{ik}|r)$ Ber, $\mathbf{P}^*(M_{ik}M_{jk'}=1|r) = \mathbf{q}(M_{ik}=1|r)\mathbf{q}(M_{jk'}=1|r)$

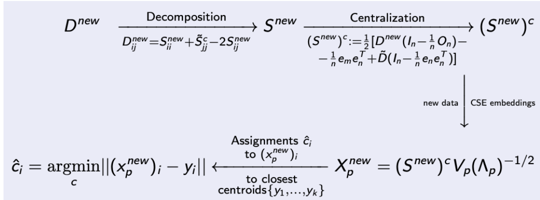
**Algorithm 2** DA

**procedure** SOLVE($\epsilon\in\mathbb{R}$, $S\in\mathbb{R}^{N\times N}$)
  $T\leftarrow\infty$
  **while** $T>\epsilon$ **do**
    For $i,k$, initialize $\mathbf{q}_{ik}(1\mid r)$ with a random value in $[0,1]$. ▷ MFA starts
    **repeat**
      $h_{ik,b}\leftarrow\mathbb{E}_{ik\mapsto b}[L(M,r)]$, for $i\le N, k\le K, b\in\{0,1\}$.
      $\mathbf{q}_{ik}(1\mid r)\leftarrow\frac{\exp(-h_{ik,1}/T)}{\sum_{b\in\{0,1\}}\exp(-h_{ik,b}/T)}$
    **until** Convergence of all $\mathbf{q}_{ik}(1\mid r)$ ▷ MFA ends
    Initialize $r^{\text{old}}$ with random values. ▷ EM starts
    **repeat**
      $r_{kk'} = \frac{\sum_{i,j\le N}\mathbf{q}_{ik}(1|r)\mathbf{q}_{jk'}(1|r^{\text{old}})S_{ij}}{\sum_{i,j\le N}\mathbf{q}_{ik}(1|r)\mathbf{q}_{jk'}(1|r^{\text{old}})}$,   for $k,k'\le K$.
      $r^{\text{old}} = r$.
    **until** Convergence of $r$. ▷ EM ends
    Reduce $T$.
  **end while**
  **Compute** $M$ **from** $\mathbf{q}(\cdot\mid r)$.
  **return** $M,r$.
**end procedure**

**Graph partitioning**, $D_{ij}\in\mathbb{R}$: $\mathcal{R}^{\text{gp}}(c;\mathcal{D}) = \sum_{\nu\le k}\sum_{(i,j)\in\mathcal{E}_{\nu\nu}}D_{ij} = \sum_{(i,j)\in\mathcal{E}}D_{ij} - \sum_{\nu,\mu:\nu<\mu}\sum_{(i,j)\in\mathcal{E}_{\nu\mu}}D_{ij} = \text{const} - \sum_{\nu\le k}\mathrm{cut}(\mathcal{G}_\nu(\mathcal{D}),\mathcal{V}\backslash\mathcal{G}_\nu(\mathcal{D})) = \text{const}_2 + \sum_{\nu\le k}\mathrm{cut}(\mathcal{G}_\nu(\mathcal{S}),\mathcal{V}\backslash\mathcal{G}_\nu(\mathcal{S}))$. *Problem:* strong bias for unbalances (very small or large) clusters due to lack normalization.
**Alternative costs:** Norm cut $\mathcal{R}^{\text{nc}}(c;\mathcal{S}) = \sum_{\nu\le k}(\frac{\mathrm{cut}(\mathcal{G}_\nu,\mathcal{V}\backslash\mathcal{G}_\nu)}{\mathrm{assoc}(\mathcal{G}_\nu,\mathcal{V})}) = k\text{-}\sum_{\nu\le k}(\frac{\mathrm{assoc}(\mathcal{G}_\nu,\mathcal{G}_\nu)}{\mathrm{assoc}(\mathcal{G}_\nu,\mathcal{V})})$, Average cut $\mathcal{R}^{\text{ac}}(c;\mathcal{S}) = \sum_{\nu\le k}(\frac{\mathrm{cut}(\mathcal{G}_\nu,\mathcal{V}\backslash\mathcal{G}_\nu)}{|\mathcal{G}_\nu|})$, Min-Max $\mathcal{R}^{\text{mmc}}(c;\mathcal{S}) = \sum_{\nu\le k}(\frac{\mathrm{cut}(\mathcal{G}_\nu,\mathcal{V}\backslash\mathcal{G}_\nu)}{\mathrm{assoc}(\mathcal{G}_\nu,\mathcal{G}_\nu)})$ (biased over equipatitions), ARC: $\mathcal{R}^{\text{rc}}(c;\mathcal{S}) = \sum\sum_{\mu=\nu+1}^k\mathrm{cut}(\mathcal{G}_\nu,\mathcal{G}_\mu)(|\mathcal{G}_\nu|^{-\frac{1}{p-1}} + |\mathcal{G}_\mu|^{-\frac{1}{p-1}})^{p-1}$ ($p=2$ standard Ratio Cut, $p=1$ multiway Cheeger Cut), Chg. Cut $\mathcal{R}^{\text{Cheeger}}(c;\mathcal{S}) = \sum_{\nu=1}^k\sum_{\mu=\nu+1}^k(\frac{\mathrm{cut}(\mathcal{G}_\nu,\mathcal{G}_\mu)}{\min\{|\mathcal{G}_\nu|,|\mathcal{G}_\mu|\}})$. **Validation** of graph based methods poor understood and no general accepted principle available.
**Data space**: pow(2, $\binom{n}{2}$), **Sol spc**: $2^n$.
Data space is exponentially larger than solution space. Estimating probability distribution over solution space feasible when data distributions cannot be estimated.
**PC & Norm Cut:** $k=2$, $G=(V,E)$, $Ncut(A,B) = \frac{\mathrm{cut}(A,B)}{\mathrm{assoc}(A,V)} + \frac{\mathrm{cut}(A,B)}{\mathrm{assoc}(B,V)} = \frac{\mathrm{assoc}(A,V)-\mathrm{assoc}(A,A)}{\mathrm{assoc}(A,V)} + \frac{\mathrm{assoc}(B,V)-\mathrm{assoc}(B,B)}{\mathrm{assoc}(B,V)} = 2 - (\frac{\mathrm{assoc}(A,A)}{\mathrm{assoc}(A,V)} + \frac{\mathrm{assoc}(B,B)}{\mathrm{assoc}(B,V)})$

## Constant Shift Embedding

$\tilde{D}$ decomposition same $D$ but tilde, $\tilde{S}^c = -\frac{1}{2}\tilde{D}^c$, $X_t = V_t(\Lambda_t)^{1/2}$, *Prob:* cluster new obj given $M\times N$ dissimilarities $D_{ij}^{\text{new}}$ betw new obj and all $n$ original
**Note:** $\tilde{S}^cV_p = V_p\Lambda_p\to X_p = \tilde{S}^cV_p(\Lambda_p)^{-1/2}$



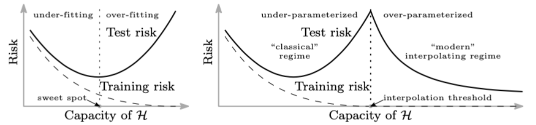

## MFA: $c(i), c(j), i\ne j$ not dependent

$\frac{\partial^2\mathcal{B}}{\partial q_u(\alpha)\partial q_v(\gamma)} = \frac{\partial h_u(\alpha)}{\partial q_v(\gamma)} = \sum_c\prod_{\substack{i=1\\ i\ne u,v}}^n q_i(c(i))\,\mathbb{I}_{\{c(u)=\alpha,\,c(v)=\gamma\}}R(c,X)$
**MDS:** given $\mathbf{D}$, $\mathcal{Z} = \{\mathbf{z}_1,\dots,\mathbf{z}_n\}\subset\mathbb{R}^d$ find $\mathbf{X} = \{\mathbf{x}_1,\dots,\mathbf{x}_n\}\subset\mathbb{R}^m$, $m=1,2,3$
$J^{\text{Sam}SStr}(\mathbf{X},\mathbf{D}) = \sum_{i\le n}\sum_{k\le n}w_{ik}(\|\mathbf{x}_i - \mathbf{x}_k\|^2 - D_{ik}^2)^2 = \sum_{i,k\le n}w_{ik}(\|\mathbf{x}_i\|^2 + \|\mathbf{x}_k\|^2 - (2\mathbf{x}_i^\top\mathbf{x}_k + D_{ik}^2))^2 = \sum_{i,k\le n}w_{ik}(2\|\mathbf{x}_i\|^4 + 2\|\mathbf{x}_i\|^2\|\mathbf{x}_k\|^2 - 8\|\mathbf{x}_i\|^2\mathbf{x}_i^\top\mathbf{x}_k - 4\|\mathbf{x}_i\|^2D_{ik}^2 + 4\,\mathrm{Tr}(\mathbf{x}_i\mathbf{x}_i^\top)(\mathbf{x}_k\mathbf{x}_k^\top) + 4\mathbf{x}_i^\top\mathbf{x}_kD_{ik}^2 + D_{ik}^4)$,
SSTRESS (quartic cost func) → analysed w\ linear algebra, SAMMONs (irrational cost func) mapping smoother embeddings than SSTRESS. **Metric MDS:** $D(\cdot)$ yields numerical values meaningful/ interpretable.
**Non-Metric MDS:** $D(\cdot)$ respects rank dissimilarities, but not necessarily their numerical values, $\delta_{ij}<\delta_{kl}\to D(\delta_{ij})<D(\delta_{kl})$.
$0 = \frac{\partial\mathcal{D}^{\text{KL}}(\mathbf{P}^0\|\mathbf{P}^G)}{\beta\partial\theta_{ip}} = \alpha_i^0\frac{\partial\mathbb{E}\{\|\mathbf{x}_i\|^4\}}{\partial\theta_{ip}} + \hat{\mathbf{h}}_i^T\frac{\partial\mathbb{E}\{\|\mathbf{x}_i\|^2\mathbf{x}_i\}}{\partial\theta_{ip}} + \frac{\partial}{\partial\theta_{ip}}\mathrm{Tr}[\mathbf{H}_i\mathbb{E}\{\mathbf{x}_i\mathbf{x}_i^T\}] + \mathbf{h}_i^T\frac{\partial\mathbb{E}\{\mathbf{x}_i\}^T}{\partial\theta_{ip}} + T\frac{\partial}{\partial\theta_{ip}}\mathbb{E}\{\log q_i\}$, $1\le i\le n$
$\hat{\mathbf{h}}_i = -8\sum_{k=1}^N w_{ik}\mathbb{E}\{\mathbf{x}_k\}$, $\alpha_i^0 = 2\sum_{k=1}^N w_{ik}$
$\mathbf{h}_i = 8\sum_{k=1}^N w_{ik}(D_{ik}^2\mathbb{E}\{\mathbf{x}_k\} - \mathbb{E}\{\|\mathbf{x}_k\|^2\mathbf{x}_k\})$,
$\mathbf{H}_i = \sum_{k=1}^N w_{ik}[8\mathbb{E}\{\mathbf{x}_k\mathbf{x}_k^T\} + 4(\mathbb{E}\{\|\mathbf{x}_k\|^2\} - D_{ik}^2)\mathbf{I}]$, $\Theta_i = (\alpha_i^0,\mathbf{h}_i,\mathbf{H}_i,\hat{\mathbf{h}}_i)$

initialize the parameters $\Theta$ of $\mathbf{P}^0(\mathbf{X}|\Theta)$ randomly
$T\leftarrow T_{max}$
**while** $T>T_{min}$ **do**
  **while** change of KL divergence $\mathcal{D}^{\text{KL}}>\varepsilon$ **do**
    $\Theta^0\leftarrow\Theta$
    **for all** $1\le i\le N$ in random order **do**
      Compute the statistics $\Phi_i$ from the expectations $\{\mathbb{E}\{\mathbf{x}_k\},\mathbb{E}\{\mathbf{x}_k\mathbf{x}_k^T\},\mathbb{E}\{\|\mathbf{x}_k\|^2\mathbf{x}_k\} : 1\le k\le n,\ k\ne i\}$, taken w.r.t. $\mathbf{P}^0(\mathbf{X}|\Theta)$
      Minimize $\mathcal{D}^{\text{KL}}(\mathbf{P}^0(\mathbf{X}|\Theta)\|\mathbf{P}^{G(\mathbf{X})})$ w.r.t. $\Theta_i$
    **end for**
  **end while**
  $T\leftarrow\eta T$, $0<\eta<1$
**end while**

## Model Selection for Clustering

**Validation Methods:** 1) procedures and concepts for quantitative and objective assessment of clustering solutions. 2) evaluate specific quality measure. 3) external (compare with ground-truth/teacher information) or internal. 4) used for model selection, i.e. use validity measures to select a model.




Second descent in the interpolation regime when the training error vanishes and the generalization error decays with a capacity increase. **Complexity MS:** assume model-based clustering, problem → log-like diverges for vanishing variance. Strategy → add regularization to neg log-like. **Occam's razor:** choose model with shortest description of the data → MDL, BIC
**MDL:** minimize description length $-\log p(\mathbf{X}|\Theta_k) - \log p(\Theta_k)$ (-data-model).
Asymptot approx (-neg log-like + penalty)
$\hat{k}\in\underset{1\le k\le K_{\max}}{\operatorname{argmin}}(-\log(\mathrm{p}(\mathbf{X}\mid\hat{\Theta}_k)) + \frac{1}{2}k'\log n)$
$k'\to\#$ param in model enc. by MLE $\hat{\Theta}_k$
**BIC:** $\dim\theta = p$, $f'(x)_{x=x_0} = 0$, flat prior
$p(M|X) = \frac{p(M)}{p(X)}\int\exp(\log p(X|M,\theta))p(\theta|M)d\theta$
$\int_{\mathbb{R}}\exp(cf(x))dx\propto\sqrt{2\pi}C|f''(x_0)|\exp(cf(x_0))$
$\bar{l}(\theta) = \frac{1}{n}\log p(X|\theta,M) = \frac{1}{n}\sum_i\log p(x_i|\theta,M)$
$\bar{l}(\theta) = \bar{l}(\hat{\theta}) - \frac{1}{2}(\hat{\theta}-\theta)^\top(-\frac{\partial^2\bar{l}(\theta)}{\partial\theta\partial\theta^\top}{}_{\theta=\hat{\theta}})(\theta-\hat{\theta})$
$p(X|M)\propto C\cdot\exp(l(\hat{\theta}))(\frac{2\pi}{n})^{\frac{k}{2}}|\mathcal{I}(\hat{\theta})|^{-\frac{1}{2}}$

$p(X|M) = \exp(\log l(\hat\theta) - \frac{k}{2}\log(n) + O(1)) = \exp(-\frac{\mathrm{BIC}}{2} + O(1))$, $p(M|X) \propto p(X|M)p(M) \approx \exp(-\frac{\mathrm{BIC}}{2})p(M)$. MDL and BIC are consistent as a model selection criterion for $n \to \infty$.
⊕ Well-motivated model selection schemes. ⊖ MDL/BIC methods to model selection rely on likelihood optimization. (not gen applicable) ⊕ Many variants, ⊖$\lambda$ tuning.
**Stability-based Validation** if no prior knowledge avaible → solutions on two data sets from same source should be similar. Stable solut. transfered to second data (same distribution) at minimal model misfit *Two sample scenario:* 1) Draw two data sets from same source 2) Cluster both data sets 3) Compute disagreement. *In Practice:* only one dataset avaible, 1) Estimate expected agreement by resampling 2) Cluster entire dataset with optimal $k$. *How to measure disagreement of two clustering $c$ and $c'$ on disjoint data?* Extend solution from $X'$ to $X$ with classifiers. 1) Train a predictor $\phi_{\mathbf{Z}'}(\cdot)$ on data $\mathbf{Z}' := (\mathbf{X}', \hat c(\mathbf{X}'))$ 2) Predict labels on $\mathbf{X}$ using $\phi_{\mathbf{Z}'}$ 3) Compare clustering solutions $(\phi_{\mathbf{Z}'}(X_i))_{i\in[n]}$ and $\hat c(\mathbf{X})$ on $\mathbf{X}$. *Symmetry Breaking:* problem → labeling unique only up to $\pi \in \mathfrak{S}_k$. Solution: Stability index $S :=$ expected minimal disagreement over all $\pi \in \mathfrak{S}_k$: Hungarian method $O(k^3)(O(n)$ preprocessing). *Final Measure for Stability:*

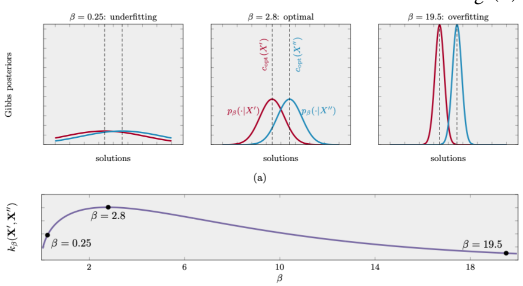
$$S_k(\hat c) = \mathbb{E}_{\mathbf{X},\mathbf{X}'}(\underbrace{\min_{\pi\in\mathfrak{S}_k}\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{\{\hat c_k(\mathbf{X})_i \neq \pi\circ\phi_{\mathbf{z}'}(X_i)\}}}_{d_{\tilde E_k}(\hat c_k(\mathbf{X}),\phi'_{\mathbf{Z}}(\mathbf{X}))})$$

*In Practice:* Estimate $\mathbb{E}_{\mathbf{X},\mathbf{X}'}$ by resampling. $t(\mathbf{X})$ any target labelling. Then $S_k(\hat c) \le \mathbb{E}_{\mathbf{X}} d_{\mathfrak{S}_k}(\hat c_k(\mathbf{X}), t(\mathbf{X})) + \mathbb{E}_{\mathbf{X},\mathbf{X}'} d_{\mathfrak{S}_k}(t(\mathbf{X}), \phi'_{\mathbf{Z}}(\mathbf{X}))$ Large $S_k$ → not close to any target labelling on average! ⊕ Performance on experimental datasets. Stability method overlooks solution complexity. Tradeoff betw informative and stability needs quantitative eval, requiring algo validation theory.

## Nature of DSA: train $X'$, validation $X''$

Substitute data-hypothesis relation with posterior distribution. *How select scoring for posterior:* conciseness (minimize MDL), generalization (high score validation)

**Design constraints hyp $\Theta$**: a priori, all hypotheses $\theta \in \Theta$ necessary for inference (If can rule out hypotheses before measured data → exclude such hypothesis a priori) **Generalization DL:** for Gibbs distrib $\mathbb{E}_{\theta|X'}$-log $\mathbf{P}(\theta|X'') = \beta(\mathbb{E}_{\theta|X'}R(\theta,X'')\text{-}\mathcal{F}(X''))$ $X'_{\mathcal{E}}, X''_{\mathcal{E}}$ generated by same experiment $\mathcal{E}$. **Object:** MDL post given test data when hyp. sampled from post given train data **M1** Train/val data cond. independent $\mathbf{P}^{\mathcal{A}}(X'_{\mathcal{E}}, X''_{\mathcal{E}}|\varphi_{\mathcal{E}}) = \mathbf{P}^{\mathcal{A}}(X'_{\mathcal{E}}|\varphi_{\mathcal{E}})\mathbf{P}^{\mathcal{A}}(X''_{\mathcal{E}}|\varphi_{\mathcal{E}})$, **M2** valdata perturbation of train data (iid) $\mathbf{P}^{\mathcal{A}}(X'_{\mathcal{E}}, X''_{\mathcal{E}}|\varphi_{\mathcal{E}}) = \mathbf{P}^{\mathcal{A}}(X''_{\mathcal{E}}|X'_{\mathcal{E}})\mathbf{P}^{\mathcal{A}}(X'_{\mathcal{E}}|\varphi_{\mathcal{E}})$ M2 corresponds to noisy channel transmission where random codebook vector $X'_{\mathcal{E}}$ corrupted by channel noise $\mathbf{P}^{A}(X''_{\mathcal{E}}|X'_{\mathcal{E}})$. **Scale oo sample DL** M1, M2 $k_{\mathcal{A}}(X'_{\mathcal{E}}, X''_{\mathcal{E}})$: $-\log\mathbf{P}^{\mathcal{A}}_{\mathcal{E}}(\theta|\varphi_{\mathcal{E}}) = -\log\mathbb{E}_{X''_{\mathcal{E}}|\varphi_{\mathcal{E}}}\mathbf{P}^{\mathcal{A}}(\theta|X''_{\mathcal{E}},\varphi_{\mathcal{E}})$ $= -\log\mathbb{E}_{X'_{\mathcal{E}}|\varphi_{\mathcal{E}}}\mathbb{E}_{X''_{\mathcal{E}}|X'_{\mathcal{E}}}\mathbf{P}^{\mathcal{A}}(\theta|X''_{\mathcal{E}},\varphi_{\mathcal{E}})$,
**Score** $\min_{\mathcal{A}}\mathbb{E}_{X'_{\mathcal{E}},X''_{\mathcal{E}}}\mathbb{E}_{\theta|X'_{\mathcal{E}}}(-\log\frac{\mathbf{P}^{\mathcal{A}}(\theta|X''_{\mathcal{E}})}{\mathbf{P}^{\mathcal{A}}_{\mathcal{E}}(\theta)}) \ge \min_{\mathcal{A}}\mathbb{E}_{X'_{\mathcal{E}},X''_{\mathcal{E}}}(-\log\mathbb{E}_{\theta|X'_{\mathcal{E}}}\frac{\mathbf{P}^{\mathcal{A}}(\theta|X''_{\mathcal{E}})}{\mathbf{P}^{\mathcal{A}}_{\mathcal{E}}(\theta)}) \ge 0$, with $\mathbb{E}_{X'_{\mathcal{E}},X''_{\mathcal{E}}}\mathbb{E}_{\theta|X'_{\mathcal{E}}}(-\log\frac{\mathbf{P}^{\mathcal{A}}(\theta|X''_{\mathcal{E}})}{\mathbf{P}^{\mathcal{A}}_{\mathcal{E}}(\theta)}) = \mathbb{E}_{X'_{\varepsilon},X''_{\varepsilon}}$ $\mathcal{D}(\mathbf{P}^{\mathcal{A}}(\theta|X'_{\mathcal{E}})\|\mathbf{P}^{\mathcal{A}}(\theta|X''_{\mathcal{E}})) - \mathcal{I}(\theta;X_{\mathcal{E}}) \ge 0$ **Prob Richness:** sampling hyp uniform from all experiments yield uniform prior $\pi(\theta) := \mathbb{E}_{\mathcal{E}}\mathbf{P}^{\mathcal{A}}_{\mathcal{E}}(\theta) \approx |\Theta|^{-1}$. $\min_{\mathcal{A}}\mathcal{D}(\pi(\theta)\||\Theta|^{-1}) = \log|\Theta| - \max_{\mathcal{A}}H_\pi(\theta)$ **Post. selection:** $\min_{\mathcal{A}}(\mathbb{E}_{X'_{\mathcal{E}},X''_{\mathcal{E}}}\mathbb{E}_{\theta|X'_{\mathcal{E}}}(-\log\frac{\mathbf{P}^{\mathcal{A}}(\theta|X''_{\mathcal{E}})}{\mathbf{P}^{\mathcal{A}}_{\mathcal{E}}(\theta)}) + \lambda\mathcal{D}(\pi(\theta)\||\Theta|^{-1})) \ge -\max_{\mathcal{A}}(\lambda H_\pi(\theta) - \lambda\log|\Theta| + \mathbb{E}_{X',X''}\log k_{\mathcal{A}}(X'_{\mathcal{E}},X''_{\mathcal{E}})) \ge$- $\max_{\mathcal{A}}\mathbb{E}_{X',X''}\log(\frac{\exp(\lambda H_\pi(\theta))}{|\Theta|^\lambda}k_{\mathcal{A}}(X'_{\mathcal{E}},X''_{\mathcal{E}})) \le 1$ $\ge \max\mathbb{E}_{X'_{\mathcal{E}},X''_{\mathcal{E}}}\log(k_{\mathcal{A}}(X'_{\mathcal{E}},X''_{\mathcal{E}}))$, $H_\pi(\theta) = \log|\Theta|$, Kernel topo concept, part ordering normalized desc length $\mathbb{E}_{\theta|X'_{\mathcal{E}}}(-\log\frac{\mathbf{P}^{\mathcal{A}}(\theta|X''_{\mathcal{E}})}{\mathbf{P}^{\mathcal{A}}_{\mathcal{E}}(\theta)})$




**Algorithm 2** Posterior Selection Algorithm
1: Derive empirical lower bound on PA kernel score $\mathbb{E}_{X',X''}\log k_{\mathcal{A}}(X',X'') \ge \frac{1}{L}\sum_{l\le L}\log k_{\mathcal{A}}(X'_l,X''_l) -$ penalty
2: Estimate the optimal empirical posterior distribution:
3: $\mathbf{P}^{\mathcal{A}}_{\mathrm{opt}}(.\,|\,.) \in \arg\max_{\mathcal{A}}(\frac{1}{L}\sum_{l\le L}\log k_{\mathcal{A}}(X'_l,X''_l) -$ penalty)
4: Sample hypotheses $\theta \sim \mathbf{P}^{\mathcal{A}}_{\mathrm{opt}}(\theta\,|\,X''')$ from optimized posterior $\mathbf{P}^{\mathcal{A}}_{\mathrm{opt}}$ given future data $X'''$ or from "posterior agreement" $\theta \sim \mathbf{P}^{\mathcal{A}}_{\mathrm{opt}}(\theta\,|\,X')\mathbf{P}^{\mathcal{A}}_{\mathrm{opt}}(\theta\,|\,X'')$

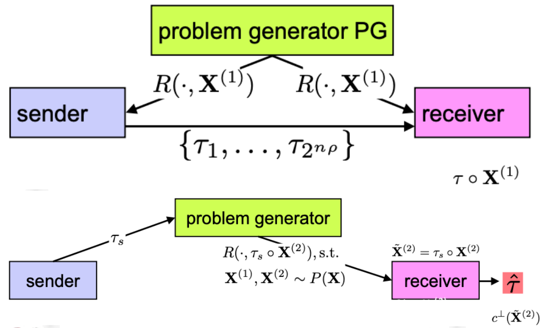
$\mathbf{P}(X'_{\mathcal{E}},X''_{\mathcal{E}}) = \sum_\theta\mathbf{P}(X'_{\mathcal{E}},X''_{\mathcal{E}}|\theta)\mathbf{P}(\theta) = \sum_\theta\frac{\mathbf{P}(X'_{\mathcal{E}},X''_{\mathcal{E}}|\theta)}{\mathbf{P}(X'_{\mathcal{E}}|\theta)\mathbf{P}(X''_{\mathcal{E}}|\theta)}\mathbf{P}(X'_{\mathcal{E}}|\theta)\mathbf{P}(X''_{\mathcal{E}}|\theta)\mathbf{P}(\theta) = \sum_\theta\frac{\mathbf{P}(X'_{\mathcal{E}},X''_{\mathcal{E}}|\theta)}{\mathbf{P}(X'_{\mathcal{E}}|\theta)\mathbf{P}(X''_{\mathcal{E}}|\theta)}\frac{\mathbf{P}(\theta|X'_{\mathcal{E}})}{\mathbf{P}(\theta)}\mathbf{P}(X'_{\mathcal{E}})\frac{\mathbf{P}(\theta|X''_{\mathcal{E}})}{\mathbf{P}(\theta)}\cdot\mathbf{P}(X''_{\mathcal{E}})\mathbf{P}(\theta)$, $\mathcal{W}(X'_{\mathcal{E}},X''_{\mathcal{E}},\theta) \le \mathbf{P}(X'_{\mathcal{E}})\mathbf{P}(X''_{\mathcal{E}})\cdot\sum_{\theta\in\Theta}\frac{\mathbf{P}(\theta|X'_{\mathcal{E}})\mathbf{P}(\theta|X''_{\mathcal{E}})}{\mathbf{P}(\theta)}\max_\theta\{\mathcal{W}(X'_{\mathcal{E}},X''_{\mathcal{E}},\theta)\}$ Insert in Mutual information $\mathcal{I}(X'_{\mathcal{E}},X''_{\mathcal{E}}) = \mathbb{E}_{X'_{\mathcal{E}},X''_{\mathcal{E}}}[\log\frac{\mathbf{P}(X'_{\mathcal{E}},X''_{\mathcal{E}})}{\mathbf{P}(X'_{\mathcal{E}})\mathbf{P}(X''_{\mathcal{E}})}] = \mathbb{E}_{X'_{\mathcal{E}},X''_{\mathcal{E}}}[\log(\sum_{\theta\in\Theta}\frac{\mathbf{P}(\theta|X'_{\mathcal{E}})\mathbf{P}(\theta|X''_{\mathcal{E}})}{\mathbf{P}(\theta)}\mathcal{W}(X'_{\mathcal{E}},X''_{\mathcal{E}},\theta))] \le \mathbb{E}_{X'_{\mathcal{E}},X''_{\mathcal{E}}}[\log\sum_{\theta\in\Theta}\frac{\mathbf{P}(\theta|X'_{\mathcal{E}})\mathbf{P}(\theta|X''_{\mathcal{E}})}{\mathbf{P}(\theta)}] + \mathbb{E}_{X'_{\mathcal{E}},X''_{\mathcal{E}}}[\log\max_\theta\{\mathcal{W}(X'_{\mathcal{E}},X''_{\mathcal{E}},\theta)\}]$ MI → weighted version of PA bounded above by PA and constant (vanishes if solutions $\theta$ are suff. stat of $X'_{\mathcal{E}},X''_{\mathcal{E}}$. **Binary sym channel:** minimize Hamming dist $\in[0,n]$, $x,c\in\{-1,1\}^n$ $d(c,x) = \sum_{i=1}^n\frac{1}{2}(1-x_ic_i) = \frac{n}{2} - \frac{1}{2}\sum_{i=1}^n x_ic_i$, $R^{\mathrm{Ham}}(c,x) = -\sum_{i=1}^n x_ic_i$ $p(c|X) = \prod_1^n\frac{\exp(\beta x_ic_i)}{\sum_c\exp(\beta\sum_1^n x_ic_i)} = \prod_1^n\frac{\exp(\beta x_ic_i)}{2\cosh(\beta x_i)}$, $Z^2 = 2^{2n}(\cosh\beta)^{2n}$ Expected alignment $\mathbb{E}_{c|x}\frac{1}{n}\sum_{i=1}^n x_ic_i = \tanh\beta$, Channel noise $\delta := \frac{1}{n}\sum_{i\le n}\mathbb{I}_{\{x'_i\neq x''_i\}}$ $\frac{1}{2}|\mathbf{x}'-\mathbf{x}''| = n\delta$, $k(x',x'') = \sum_c\frac{e^{-\beta x'c}}{Z'}\frac{e^{-\beta x''c}}{Z''} = \frac{1}{Z^2}\sum_c\prod e^{-\beta(x'_i+x''_i)c_i} = \frac{1}{Z^2}\prod e^{-\beta(x'_i+x''_i)} + e^{\beta(x'_i+x''_i)c_i} = \frac{1}{Z^2}2^n(\cosh2\beta)^{n(1-\delta)}$ $I = \mathbb{E}_{x',x''}\log|c|k(x',x'') \propto n(1-\delta)\cdot\log\cosh2\beta - 2n\log\cosh\beta \to \frac{\partial I}{\partial\beta} = 0$ $\cosh\beta = \sqrt{\frac{1}{2}(\cosh2\beta+1)} = (2\delta)^{-1/2}$ $\frac{1}{n}I(\beta^{opt}) = (1-\delta)\log\cosh2\beta - 2\log\cosh\beta = \log2 - H_{bin}(\delta)$, for generalized ofSDL: $2\delta\tanh\beta - (1-2\delta)\frac{\beta}{(\cosh\beta)^2} = 0$

## Info Theoretical Algorithm Validation

Data noise reduces solution resolution; simple solution space leads to underfitting. Data analysis focuses on concentrated data distributions with signals and noise. Worst case analysis neglects the signal. *Objective:* rank algorithms based on signal sensitivity and noise robustness.

**Algorithm 3** Structure Validation IT
1: Sample a hypothesis $\tilde c \sim p^{\mathcal{A}}(c\,|\,\mathbf{X})$
2: **for** $j = 1,\dots,M$ **do**
3: Select random transformation $\tau_j \in \mathbb{T}$
4: Define code vector $\tilde c_i \sim p^{\mathcal{A}}(c\,|\,\tau_i\circ\mathbf{X})$
5: **end for**
6: **return** Codebook $\mathcal{T} := \{\tilde c_1,\dots,\tilde c_M\}$

**Approximation sets (AS):** data in high dim space, measure sensitivity sol sets to noise. **Coding** with sets of approximate rankings: define set code problems. **Communication** by AS: estimate coding error




**Algorithm 4** Communication by approximation sets
1: Sender sends transformation $\tau_s$ to problem generator
2: Problem generator sends a new problem with transformed indices to receiver without revealing $\tau_s$.
3: Receiver identifies transformation $\hat\tau$ by comparing approximation sets.

**Approximate Sorting:** given $N$ objects $o_1..o_i..o_N$, rank them based on pairwise noisy comparisons represented by dataset $X$, with $X_{ij} = 1$ if $o_i < o_j$, 0 else, $X_{ii} = 0$ and $X_{ij} = 1 - X_{ji}\ \forall i, j \neq i$, feasible solutions = all permutations $c$ over items, where $c_i = \ell$ = position (rank) of obj $o_i$ $\mathcal{R}^{sort}(c|X) := |\{(i,j)|X_{ij} = 1 \text{ but } c_i > c_j\}|$ Let $M$ assignment matrix with $M_{il} = 1$ if $o_i$ gets rank $l$, 0 else, $\sum_{\ell i}M_{i\ell} = 1$, item/rank get unique rank/item, $M(i) = \ell \Leftrightarrow M_{i\ell} = 1$ $\mathcal{R}^{\mathrm{sort}}(M|X) = \sum_i\sum_j\sum_\ell\sum_{\bar\ell>\ell}X_{ij}M_{i\bar\ell}M_{j\ell}$ **MFA:** $q_{i\ell} := \frac{e^{-\beta\mathcal{E}_{i\ell}}}{\sum_h e^{-\beta\mathcal{E}_{ih}}}$, $Q(M|\mathcal{E}) = q_{1M(1)}\cdot q_{2M(2)}\cdots q_{NM(N)}$, $\mathcal{E}_{i\ell} = \mathbb{E}_{M\sim Q_{i\to\ell}}[\mathcal{R}^{\mathrm{sort}}]$, in before equation notation means $o_i$ to rank $l$, Q satisfies this constraint, not both C1) $X_{ij} = 1 \Rightarrow i\to\ell$ and $j\to\underline{\ell}$ with $\underline{\ell}<\ell$

C2) $X_{ji} = 1 \Rightarrow i \to \ell$ and $j \to \underline{\ell}$ with $\overline{l} > \ell$
C3) $X_{ab} = 1 \Rightarrow a \to \overline{\ell}$ and $b \to \underline{\ell}$ w/ $\underline{\ell} < \overline{\ell}$
$\mathcal{E}_{i\ell} = \sum_j \sum_{\underline{\ell}<\ell} X_{ij} q_{j\underline{\ell}} + \sum_j \sum_{\overline{l}<\ell} X_{ji} q_{j\overline{\ell}}$ $+ \sum_{a \neq i} \sum_{b \neq i} X_{ab} \sum_{\underline{\ell}} \sum_{\overline{l} > \underline{\ell}} q_{a\overline{\ell}} q_{b\underline{\ell}}$, const
$\mathcal{R}^{MF}_{fix}(M|\mathcal{E}) = \sum_{i,l} M_{il}\mathcal{E}_{il} + \sum_\ell \lambda_\ell (\sum_j M_{j\ell} - 1)$ $= \sum_{il} M_{il}(\mathcal{E}_{il} + \lambda_l) + \text{cst}$, $\lambda_l$ from $\sum_i q_{i\ell} = 1$
**sMBP:** given $G = (V, E, W)$, select subset of nodes $U$, find min bisection for $U$. If $|U| 0\Theta(n^{2/7}) \to$ NP-hard. **REM:** given $N = \Theta(\exp(n))$ states $J \in \mathcal{H}$, i.i.d. score $(R(c, \mathbf{X}))$ values score$(J) \sim \mathcal{N}(0, \sqrt{n})$. Find $\widehat{J}_{\max} = \arg\max_{J \in \mathcal{H}} \text{score}(J)$, model max disordered and without structure for searching $n^s = \Theta(n^{2/7})$. By construction, different solutions $J', J''$ do not share common parameters (statistical independent) *sMBP is lower/upper bounded by REM:*
$\lim_{n\to\infty} \frac{\mathbb{E}[\log Z(\beta,\mathbf{X})] + \hat{\beta}\mu\sqrt{N \log m}}{\log m}$
$= 1 + \frac{\hat{\beta}^2\sigma^2}{2}$ if $\hat{\beta}\sigma < \sqrt{2}$ or $= \hat{\beta}\sigma\sqrt{2}$ (else)
**Generalization capacity** of sMBP for noise perturbed random graph: $\mathbf{X}' = \mathbf{X} + \delta\mathbf{X}', \mathbf{X}'' = \mathbf{X} + \delta\mathbf{X}'', \gamma = \sigma/\tilde{\sigma}$, 2 transitions
$\lim_{n\to\infty} \frac{\mathbb{E}_{X,\delta X',\delta X''} \log(|\mathcal{C}|\hat{k}_\beta(X',X''))}{\log m} = \eta(\hat{\beta})$

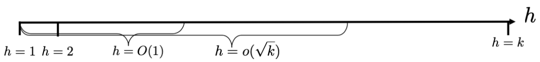
$$\eta(\hat{\beta}) = \begin{cases} (\hat{\beta}\tilde{\sigma})^2, & \hat{\beta}\tilde{\sigma} < \frac{\sqrt{2}}{\sqrt{4+2\gamma^2}} \\ \hat{\beta}\tilde{\sigma}\sqrt{2}\sqrt{4+2\gamma^2} - (\hat{\beta}\tilde{\sigma})^2(1+\gamma^2) - 1, & \frac{\sqrt{2}}{\sqrt{4+2\gamma^2}} \le \hat{\beta}\tilde{\sigma} < \frac{\sqrt{2}}{\sqrt{1+\gamma^2}} \\ \hat{\beta}\tilde{\sigma}\sqrt{2}\left(\sqrt{4+2\gamma^2} - 2\sqrt{1+\gamma^2}\right) + 1, & \frac{\sqrt{2}}{\sqrt{1+\gamma^2}} \le \hat{\beta}\tilde{\sigma} \end{cases}$$

REM has no information for searching $\to$ evaluate $M = 2^n$ cost values. Random sMBP, CDP $\approx$ REM (free E gen. func.)

## Learning and algorithmic complexity

**Strategy for combinatorial search:** 1) Define concept- and structurally simple comb. problem P1 2) Define reference prob. P2 where exhaustive search is optimal 3) Study relation betw. P1 and P2.
**PREM (P2):** given $N = O(\exp(n))$ states $J \in \mathcal{H}$ with scores $X_J \sim \mathcal{N}(X \mid 0, \sigma^2)$, define planted state $I^\star \sim \mathcal{N}(X|\mu, 1)$, find state $\widehat{J}_{\max} = \arg\max_{J \in \mathcal{H} \cup I^\star} \text{score}(J)$. *Signal:* Bias $\mu$, unbiased scores act as observation noise. Minimal bias $\mu_0$ for recovery in probability? (Unlimited positive bias can't improve recovery beyond exhaustive search due to lack of shared information with neighboring states) **Recover prob:**
$\mathbf{P}_{\mathcal{X}}(\max_{J \in \mathcal{H}\setminus\{I^\star\}} X_J < X_{I^\star}) \xrightarrow{n\uparrow\infty} 1$

**Success prob.** to recover Planted REM:
$\mathbf{P}(\widehat{J} = I^\star) \geq \exp(-\exp(\ln|\mathcal{H}| - \frac{\mu^2}{4}))$
0/1 behaviour, threshold $\mu_0 = 2\sqrt{\ln|\mathcal{H}|}$:
$\lim_{n\uparrow\infty} \mathbf{P}(\widehat{J} = I^\star) = 1$ if $\mu > \mu_0$, $= 0$ else
$\mathbf{P}(\hat{J} = I^\star) = \mathbb{E}_{X_{I^\star}} \mathbf{P}(\hat{J} = I^\star | X_{I^\star}) = \mathbb{E}_{X_{I^\star}}$
$\mathbf{P}(\bigwedge_{J \neq I^\star} X_J < X_{I^\star} | X_{I^\star}) = \mathbb{E}_{X_{I^\star}} \prod_{J \neq I^\star}$
$\mathbf{P}(X_J < X_{I^\star}|X_{I^\star}) = \mathbb{E}_{X_{I^\star}} (\int_{-\infty}^{X_{I^\star}} d\mathcal{N}(z)$
$)^{|\mathcal{H}|-1} = \mathbb{E}_{X_{I^\star}}(1 - \int_{X_{I^\star}}^\infty d\mathcal{N}(z))^{|\mathcal{H}|-1} \geq (1 -$
$\mathbb{E}_{X_{I^\star}} \int_{X_{I^\star}}^\infty d\mathcal{N}(z))^{|\mathcal{H}|} \geq (1 - \mathbb{E}_{X_{I^\star}} \frac{\mathcal{N}(X_{I^\star})}{X_{I^\star}})^{|\mathcal{H}|}$
$\approx \exp(-|\mathcal{H}| \exp(-\mu^2/4))$
**Note:** $\frac{1}{u}e^{-u^2/2}(1 + u^{-2}) \leq \int_u^\infty e^{-z^2/2} dz \leq \frac{1}{u}e^{-u^2/2}$, $d\mathcal{N}(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2})dz$
**Planted sub-hypergraph (P1):** Hypothesis $J$ is set of $k$ indices $J = \{j_1, \ldots, j_k\} \subset \{1, \ldots, n\} =: [n]$, $\mathcal{H} = \{J \subset [n] : |J| = k\}$
$|\mathcal{H}| = \binom{n}{k} < (\frac{ne}{k})^k$, sol $I^\star = \{i_1^\star, \ldots, i_k^\star\}$
Subsets $H$ of $h$ items assigned score $X_H, H = \{j_1, \ldots, j_h\}$ Scores $X_H$ are data of planted sub-hypergraph recovery. score$(J) = \sum_{H \subset J:|H|=h} X_H$. All proper subsets $H$ of planted sol $I^\star$ are biased.
$X_H = \begin{cases} \mu + \xi_H & H \subset I^\star \\ \xi_H & H \not\subset I^\star \end{cases}$, $\xi_H \sim \mathcal{N}(0, \sigma^2)$
**Note:** $J$ defined by Bin$(k, h)$ parameters $X_H$. When $h$ increases, localize planted solution $I^\star$ by using Bin$(n, h)$ RV as parameters (localization of solutions).
**Interaction degree** $h$: $h = 1$ sort the $X_i$ values, select $k$ largest score$(J) = \sum_{i \in J} X_i$, $h = 2$ find $k$-clique with only biased edges, score$(J) = \sum_{(i,j) \in J \times J} X_{i,j}$, $h \geq 3$ sub-hypergraph recovery, $h = k$ identify set $J = \arg\max_{J \subset [n]} X_J$ employing exhaustive search of Bin$(n, k)$ states, score$(J) = X_J$.




Increasing $h \to k$ localizes influence planted solution $I^\star$ up to single state at expense of increasing amount of (random) data.
**REM recovery:** $k$ sub-hypergraph recovery with $h = k$ interaction suppresses parameter sharing and eliminates statistical dependence between solutions. This independence implies REM behavior since all of the Bin$(n, k)$ subsets of size $k$ of $n$ items are characterized by individual score value.

**Algorithm 5** Hypergraph recovery problem
1: Define spectrum of hypergraph recovery prob $(1 \leq h \leq k)$ with increased localization of solutions.
2: Characterize the statistical dependencies of solutions.
3: Compare the community detection case $h = \mathcal{O}(n^0)$ with REM as limit case $h = k$ for $k = n^\alpha, 0 < \alpha < \frac{1}{2}$.
4: Estimate number of effectively independent solutions to measure the computational complexity of the search problem.
5: Prove absence of effective statistical tests to amplify asymptotically vanishing search information.

*How many states share param with planted solution $I^\star$?* Assume $|J \cap I^\star| = l \geq h \Rightarrow$ Bin$(k, l) \cdot$ Bin$(n - k, n - l)$. $\mathcal{D}(I^\star)$ set of solu $J$ statistically dependent on $I^\star$ (share at least 1 param). Let $|\mathcal{D}(I^\star)| = \sum_{l=h}^k \text{Bin}(k, l)\,\text{Bin}(n - k, n - l)$, $k = n^\alpha, 0 < \alpha < \frac{1}{2}$, Bin$(k, h) \cdot$ Bin$(n - k, n - h)$, then $(*) \leq |\mathcal{D}(I^\star)| \leq (*) \cdot (k - h + 1)$,
*Proof:* upper bound holds since $(*)$ monotonically decreases in $l$ for $k = o(\sqrt{n})$.
**Effective independence:** if sol $J$ shares $l > h$ items with $I^\star$, then Bin$(l, h)$ parameters of $J$ are biased and Bin$(k, h) -$ Bin$(l, h)$ are unbiased. If total bias much small than standard dev, (Bin$(l, h)\mu \ll \sqrt{\text{Bin}(k, h)\sigma}$), then cant distinguish betw sol with $l$ shared items and independent solution due to fluctuations. Assume that overlap $h \leq l \leq l_\perp$ ($l_\perp$ cutoff parameter) too weak to significant dependence, # effectively dependent states: $|\mathcal{D}^{\text{eft}}(I^\star)| = \sum_{l=l_\perp}^k \text{Bin}(k, l) \cdot \text{Bin}(n - k, k - l) \approx \text{Bin}(k, l_\perp)\,\text{Bin}(n - k, k - l_\perp)(k - l_\perp + 1)$
**Sub hyp det rate:** $\mathfrak{R} = \ln \frac{|\mathcal{H}|}{|\mathcal{D}^{\text{eft}}(I^\star)||\mathcal{X}|}$
$\geq \ln \text{Bin}(n, k) - \ln[\text{Bin}(k, l_\perp) \cdot \text{Bin}(n - k, k - l_\perp)(k - l_\perp + 1)] - \ln \text{Bin}(n, h)$
If $\mathfrak{R} = \mathcal{O}(n^\beta)$, exponentially many independent states that define REM and this REM is exponentially larger than the number of parameters! Since a REM requires exhaustive search, conclude that searching for planted sub-hypergraph is exponentially more expensive than testing. **Stirling:**
$n! = \sqrt{2\pi n}(\frac{n}{e})^n \cdot (1 + \frac{1}{12n} + \mathcal{O}(n^{-2}))$
$\ln \text{Bin}(a, b) \cong b \ln \frac{a}{b} + b(1 - r) \leq b(\ln \frac{a}{b} + 1)$
with $r = \sum_{\nu=1}^\infty \frac{1}{\nu(\nu+1)} \left(\frac{b}{a}\right)^\nu +$
$\frac{1}{2b} \ln 2\pi b \left(1 - \frac{b}{a}\right) = \mathcal{O}\left(\max\left\{\frac{b}{a}, \frac{1}{b}\right\}\right)$
**Recovery Theo:** if $k = n^\alpha, l_\perp = n^\beta, h = \mathcal{O}(n^0), 2\alpha < 1$, then $\mathfrak{R} \geq l_\perp \ln \frac{l_\perp n}{k^2} - l_\perp - h(\ln \frac{n}{h} + 1) - \ln(k - l_\perp + 1) = \mathcal{O}(n^\beta \ln n)$
For case $\mathcal{D}(I^\star) = \mathcal{D}^{\text{eft}}(I^\star)$ with $l_\perp = h$, rate scale polynomial in $n$ with $\mathfrak{R} = \mathcal{O}(h \ln n)$, $\oplus$ refined if $h = \mathcal{O}(n^\gamma)$. *Proof:* $\geq k(\ln \frac{n}{k} + 1) - kr(n, k) - l_\perp(\ln \frac{k}{l_\perp} + 1) - (k - l_\perp)(\ln \frac{n-k}{k-l_\perp} + 1) - \ln(k - l_\perp + 1) - h(\ln \frac{n}{h} + 1)$

The terms linear in $l_\perp$ define the leading contributions in $n$ (since the terms linear in $k$ vanish asymptotically), i.e.,

$$-l_\perp(\ln \frac{k}{l_\perp} + 1) + l_\perp(\ln \frac{n-k}{k-l_\perp} + 1) = l_\perp \ln \frac{(n-k)l_\perp}{(k-l_\perp)k} \geq l_\perp \ln \frac{nl_\perp}{k^2} = (1 + \beta - 2\alpha)\, n^\beta \ln n = \mathcal{O}(n^\beta \ln n)$$

The sum of terms proportional to $k$ is negative; therefore, we bound its absolute value from above, i.e.,

$$-k(\ln \frac{n}{k} + 1) + k(\ln \frac{n-k}{k-l_\perp} + 1) + kr(n,k) = k \ln \frac{k(n-k)}{n(k-l_\perp)} + kr(n,k) = k \ln \frac{1-k/n}{1-l_\perp/k} + kr(n,k)$$
$$= k \sum_{\nu=1}^\infty \frac{1}{\nu}\left(\frac{l_\perp^\nu}{k^\nu} - \frac{k^\nu}{n^\nu}\right) + kr(n,k) = l_\perp - \frac{k^2}{n} + \sum_{\nu=2}^\infty \frac{k}{\nu}\left(\frac{l_\perp^\nu}{k^\nu} - \frac{k^\nu}{n^\nu}\right) + kr(n,k) \leq l_\perp + kr(n,k) = \mathcal{O}(n^\beta),$$

since $kr(n, k) \sim \frac{k^2}{n} = O(n^{2\alpha-1}) \to 0$ for $2\alpha < 1$. Collecting all the terms yields the final rate

$$\mathfrak{R} \geq l_\perp \ln \frac{nl_\perp}{k^2} - l_\perp - h(\ln \frac{n}{h} + 1) - \ln(k - l_\perp + 1) = \mathcal{O}(n^\beta \ln n)$$

since the positive term $l_\perp \ln \frac{nl_\perp}{k^2} = \mathcal{O}(n^\beta \ln n)$ dominates the negative terms that scale as $l_\perp = \mathcal{O}(n^\beta)$ and it also dominates $h(\ln \frac{n}{h} + 1) + \ln(k - l_\perp + 1) = \mathcal{O}(\ln n)$ in this scaling limit $h = \mathcal{O}(1)$.

**Exponential scaling:** $h = O(n^\gamma), 0 < \gamma \leq \beta < \alpha, |\mathcal{X}| = \text{Bin}(n, h) = O(n^{(n^\gamma)})$, $l_\perp = \omega h$ with $\omega = O(1)$, $\mathfrak{R} \geq l_\perp \ln \frac{nl_\perp}{k^2} - l_\perp - h(\ln \frac{n}{h} + 1) - \ln(k - l_\perp + 1) \cong \omega h \ln \frac{n\omega h}{k^2} - \omega h - h(\ln \frac{n}{h} + 1)$, $\frac{\mathfrak{R}}{h \ln n} \geq \omega(1 + \gamma - 2\alpha) - 1 + \gamma + \frac{\omega(\ln \omega - 1) - 1}{\ln n} > 0$, $\alpha - \gamma = \frac{\psi}{\ln n}, \omega - 1 = \frac{\tau}{\ln n}$
Rate is positive in asymptotic limit ($\mathcal{R} \geq n^\gamma$) for $\tau > 2\frac{\psi+1}{1-\alpha} \Rightarrow \frac{\mathfrak{R}}{h \ln n} \geq 0$, so $\omega > 1 + \frac{2}{1-\alpha}(\alpha - \gamma) + \frac{2}{(1-\alpha)\ln n}$
**Small overlap:** sol $J'$ weakly stat dependent on $I^\star$ ($|J' \cap I^\star| = l > h$); sol$J''$ stat independent on $I^\star$,($|J'' \cap I^\star| < h$); $J'$ and $J''$ are stat independent ($|J' \cap J''| < h$)

$$\text{score}(J') = \sum_{H \subset J' \cap I^\star} X_H + \sum_{\tilde{H} \subset J' \wedge \tilde{H} \not\subset I^\star} X_{\tilde{H}} \sim \mathcal{N}\left(\binom{l}{h}\mu, \binom{k}{h}\sigma^2\right)$$

$$\text{score}(J'') = \sum_{H \subset J''} X_H \sim \mathcal{N}\left(0, \binom{k}{h}\sigma^2\right)$$

score$(J', J'') \in \text{Bin}(l, h)\mu \pm \sqrt{\text{Bin}(k, h)\sigma}$
select maximal overlap $l_\perp$ as Bin$(l_\perp, h) = \epsilon\sqrt{\text{Bin}(k, h)}\frac{\sigma}{\mu} \to l_\perp = \sqrt{k}(\epsilon\frac{\sigma}{\mu})^{\frac{1}{h}}$, Induce subsampled hyp class $\mathcal{H}^\Delta$, with $|\mathcal{H}^\Delta| = \frac{|\mathcal{H}|}{|\mathcal{D}^{\text{eft}}(I^\star)|} = \mathcal{O}(n \exp(n^\beta))$, REM recovery condition: $\frac{\mu}{\sigma} > 2\sqrt{\ln|\mathcal{H}^\Delta|}$, $\beta, \alpha = \frac{1}{7}, \frac{1}{3}$
$l_\perp = n^\beta < \sqrt{k}(\frac{\epsilon}{2\sqrt{\ln|\mathcal{H}^\Delta|}})^{\frac{1}{h}} = \mathcal{O}(n^{\frac{\alpha}{2}} n^{-\frac{\beta}{2h}})$
**Ising Model,** $N = \#$ **pixel,** $\lambda \,\&\, J_{ij} \geq 0$
$E(\boldsymbol{\sigma}|\mathbf{h}) = -\lambda \sum_{i=1}^N h_i \sigma_i - \sum_{\langle i,j \rangle} J_{ij} \sigma_i \sigma_j$,
$\mathbf{h} = \{h_i\}$ noisy img, $\boldsymbol{\sigma} = \{\sigma_i, h_i\}(\sigma_i \in \{-1, 1\})$ denoised img (sol), like, prior
**MCMC:** $p_i(\boldsymbol{\sigma}) = \exp(-\beta \max[0, \Delta E_i(\boldsymbol{\sigma})])$
$P(\boldsymbol{\sigma}, \boldsymbol{\sigma}') = \frac{1}{N} p_i(\boldsymbol{\sigma})$ if $\boldsymbol{\sigma}' = \boldsymbol{\sigma}^{(i)}$ for some $i$, $= 1 - \frac{1}{N} \sum_{i=1}^N p_i(\boldsymbol{\sigma})$ if $\boldsymbol{\sigma}' = \boldsymbol{\sigma}$ , else 0.

Irreducible and stationary, detailed balance hold, periodic (if $E(\boldsymbol{\sigma}) = E$ for all state, period 2, only case → no loop = E const)

*Periodic p-prior*: $-\sum_{i=1}^{N-p} \sigma_i \sigma_{i+p}$ (E is sum of $p$ 1-periodic Ising energies)

**Algorithm 1** Metropolis-Hastings

Define $\{q(\cdot \mid c)\}_{c \in C}$ s.t $\mathcal{G}_q$ is connected, $q(c \mid c) > 0$
$c_0 \leftarrow \$$
**for** $t = 1, 2, \ldots$ **do**
  $\tilde{c} \xleftarrow{\$} q(\cdot \mid c_{t-1})$
  $b \xleftarrow{\$} \mathrm{Ber}(\min\{1, \frac{q(c_{t-1}|\tilde{c})p(\tilde{c})}{q(\tilde{c}|c_{t-1})p(c_{t-1})}\})$
  **if** $b = 1$ **then**
    Set $c_t = \tilde{c}$ (Accept the proposal)
  **else**
    Set $c_t = c_{t-1}$ (Reject the proposal)
  **end if**
  $t \leftarrow t + 1$
**end for**

**Algorithm 2** Posterior Selection Algorithm

Derive empirical lower bound on PA kernel score $\mathbb{E}_{X',X''} \log k_{\mathcal{A}}(X', X'') \geq \frac{1}{L}\sum_{l \leq L} \log k_{\mathcal{A}}(X'_l, X''_l) -$ penalty
Estimate the optimal empirical posterior distribution:
$\mathbf{P}^{\mathcal{A}}_{\text{opt}}(. \mid .) \in \arg\max_{\mathcal{A}} \left(\frac{1}{L}\sum_{l \leq L} \log k_{\mathcal{A}}(X'_l, X''_l) - \text{ penalty}\right)$
Sample hypotheses $\theta \sim \mathbf{P}^{\mathcal{A}}_{\text{opt}}(\theta \mid X''')$ from optimized posterior $\mathbf{P}^{\mathcal{A}}_{\text{opt}}$ given future data $X'''$ or from "posterior agreement" $\theta \sim \mathbf{P}^{\mathcal{A}}_{\text{opt}}(\theta \mid X')\, \mathbf{P}^{\mathcal{A}}_{\text{opt}}(\theta \mid X'')$

**Algorithm 3** Structure Validation IT

Sample a hypothesis $\tilde{c} \sim p^{\mathcal{A}}(c \mid \mathbf{X})$
**for** $j = 1, \ldots, M$ **do**
  Select random transformation $\tau_j \in \mathbb{T}$
  Define code vector $\tilde{c}_i \sim p^{\mathcal{A}}(c \mid \tau_i \circ \mathbf{X})$
**end for**
**return** Codebook $\mathcal{T} := \{\tilde{c}_1, \ldots, \tilde{c}_M\}$

**Algorithm 4** Communication by approximation sets

Sender sends transformation $\tau_s$ to problem generator
Problem generator sends a new problem with transformed indices to receiver without revealing $\tau_s$.
Receiver identifies transformation $\hat{\tau}$ by comparing approximation sets.

$$\eta(\hat{\beta}) = \begin{cases} (\hat{\beta}\tilde{\sigma})^2, & \hat{\beta}\tilde{\sigma} < \frac{\sqrt{2}}{\sqrt{4+2\gamma^2}} \\ \hat{\beta}\tilde{\sigma}\sqrt{2}\sqrt{4+2\gamma^2} - (\hat{\beta}\tilde{\sigma})^2\left(1+\gamma^2\right) - 1, & \frac{\sqrt{2}}{\sqrt{4+2\gamma^2}} \leq \hat{\beta}\tilde{\sigma} < \frac{\sqrt{2}}{\sqrt{1+\gamma^2}} \\ \hat{\beta}\tilde{\sigma}\sqrt{2}\left(\sqrt{4+2\gamma^2} - 2\sqrt{1+\gamma^2}\right) + 1, & \frac{\sqrt{2}}{\sqrt{1+\gamma^2}} \leq \hat{\beta}\tilde{\sigma} \end{cases}$$

**Algorithm 5** Hypergraph recovery problem

1: Define spectrum of hypergraph recovery prob
$(1 \leq h \leq k)$ with increased localization of solutions.
2: Characterize the statistical dependencies of solutions.
3: Compare the community detection case $h = \mathcal{O}\left(n^0\right)$ with
REM as limit case $h = k$ for $k = n^\alpha, 0 < \alpha < \frac{1}{2}$.
4: Estimate number of effectively independent solutions to
measure the computational complexity of the search problem.
5: Prove absense of effective statistical tests to
amplify asymptotically vanishing search information.

$$\text{score}\left(J'\right) = \sum_{H \subset J' \cap I^\star} X_H + \sum_{\tilde{H} \subset J' \wedge \tilde{H} \not\subset I^\star} X_{\tilde{H}} \sim \mathcal{N}\left(\left(\begin{array}{c} l \\ h \end{array}\right)\mu, \left(\begin{array}{c} k \\ h \end{array}\right)\sigma^2\right)$$

$$\text{score}\left(J''\right) = \sum_{H \subset J''} X_H \sim \mathcal{N}\left(0, \left(\begin{array}{c} k \\ h \end{array}\right)\sigma^2\right)$$