

Contents

1 Preliminaries:	2
1.0.1 Ex 15.18 Wainright	2
1.1 M-ary Hypotheses testing	2
1.2 Fano’s Inequality	2
1.2.1 Measure Theory background	2
1.2.2 Information Theory background	2
1.2.3 Fano and M-ary Testing Intuition	3
1.2.4 Fano’s bound for M-ary testing problem	3
1.3 Minimax Lower Bounds for Sparse Linear Regression Recovery Problem	3
1.3.1 Goal	3
1.3.2 Setting	3
1.3.3 Minimax to M-ary testing	3
1.3.4 Fano’s method	5
1.4 Results	7
2 Lower Bound for Sparse Causal Estimator	7
2.1 Goal and Problem Description	7
2.2 Formalism	7
2.2.1 Model	7
2.2.2 Data	8
2.3 Minimax to M-ary Hypotheses Testing	8
2.3.1 Discretization of the Parameter Space	8
2.3.2 Minimax after Discretization	9
2.4 Fano’s Method	9
2.5 Fano’s continued with Lemma 2	10
2.5.1 Optimization problem	11
3 Fano: Continued with upper bounding using convexity argument	13
3.1 Some Tools	13
3.2 New upper bound for mutual information $I(\underline{\mathbf{y}}_i, \underline{\mathbf{J}} \underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i)$	13
3.3 Interpretation	15

1 Preliminaries:

1.0.1 Ex 15.18 Wainright

The content in this section is strongly related to the example 15.18 in the book "High Dimensional Statistics"

1.1 M-ary Hypotheses testing

Note: $\mathbb{P}\{\underline{\mathbf{x}}\}$ denote the prob of the event whereas $\mathbb{P}(\underline{\mathbf{x}}) =: \mathbb{P}_{\underline{\mathbf{x}}}(x)$ is the distribution
Assuming the following family of distribution

$$\mathcal{P} := \{\mathbb{P}_{\theta_j} \mid \theta_j \in \Theta := \{\theta_1, \dots, \theta_M\}\} \quad (1)$$

An M-ary testing problem defined by \mathcal{P} and the following

$$\underline{\mathbf{J}} \sim \text{Uni}\{[M]\} \quad (2)$$

$$\underline{\mathbf{Z}} \sim \mathbb{P}_{\theta_{\underline{\mathbf{J}}}} \quad (3)$$

Defining a testing function $\psi : \mathcal{Z} \rightarrow [M]$ we have to following error event with probability taken jointly over $(\underline{\mathbf{Z}}, \underline{\mathbf{J}})$

$$\mathbb{Q}\{\psi(\underline{\mathbf{Z}}) \neq \underline{\mathbf{J}}\} \quad (4)$$

where \mathbb{Q} denotes the joint distribution of $(\underline{\mathbf{Z}}, \underline{\mathbf{J}})$

Note: the marginal distribution of $\underline{\mathbf{Z}}$ is

$$\bar{\mathbb{Q}} = \sum_{j \in [M]} \mathbb{Q}(\underline{\mathbf{Z}}, \underline{\mathbf{J}} = j) = \sum_{j \in [M]} \mathbb{Q}(\underline{\mathbf{Z}} | \underline{\mathbf{J}} = j) \mathbb{Q}(\underline{\mathbf{J}} = j) \quad (5)$$

$$= \sum_{j \in [M]} \mathbb{P}_{\theta_{\underline{\mathbf{J}}=j}} \frac{1}{M} \quad (6)$$

Note that $\bar{\mathbb{Q}}[\text{Event}(\underline{\mathbf{Z}})]$ takes event from $\sigma(\mathcal{Z})$ i.e does not depends on $\underline{\mathbf{J}}$.

Meaning that the data generated by the rv $\underline{\mathbf{Z}}$ have a mixture distribution

1.2 Fano's Inequality

1.2.1 Measure Theory background

Abslute Continuity of Measures: Let ν, μ be two measures on the same measurable space (S, \mathcal{A}) . The measure ν is absolutely continous wrt the measure μ written

$$\nu \ll \mu \iff \mu(A) = 0 \implies \nu(A) = 0 \quad \forall A \in \mathcal{A}$$

Radon-Nikodym Thm:

Let μ and ν be measures on (S, \mathcal{A}) s.t $\nu \ll \mu$, then there exists a \mathcal{A} -measurable function $f : S \rightarrow [0, \infty)$ such that for any measurable set $A \subseteq S$ we have

$$\nu(A) = \int_A f d\mu \quad \text{or the radon derivative:} \quad f := \frac{d\nu}{d\mu}$$

Chain Rule: Suppose the measures ν, μ, η on the measurable space (S, \mathcal{A}) s.t $\nu \ll \mu \ll \eta$ then

$$\frac{d\nu}{d\mu}(s) = \frac{d\nu}{d\eta} \frac{d\eta}{d\mu}(s) \quad a.e[\eta]$$

1.2.2 Information Theory background

KL divergence:

$$D(\mathbb{P} || \mathbb{Q}) = \begin{cases} \mathbb{E}_{\mathbb{P}} \log \frac{d\mathbb{P}}{d\mathbb{Q}} & \text{when } \mathbb{P} \ll \mathbb{Q} \\ 0 & \text{else} \end{cases}$$

KL with densities: $\mathbb{P}, \mathbb{Q} \ll \nu$ with densities: $p = \frac{d\mathbb{P}}{d\nu}, q = \frac{d\mathbb{Q}}{d\nu}$

$$D(\mathbb{P} || \mathbb{Q}) = \int_{\mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right) \nu(dx)$$

KL Tensorization for $\mathbb{P}_j \ll \mathbb{Q}_j, \forall j \in [n]$

$$D(\otimes_{j \in [n]} \mathbb{P}_j || \otimes_{j \in [n]} \mathbb{Q}_j) = \sum_{j \in [n]} D(\mathbb{P}_j || \mathbb{Q}_j)$$

Mutual Information:

$$I(\underline{\mathbf{Z}}, \underline{\mathbf{J}}) := D(\mathbb{Q}_{\underline{\mathbf{Z}}, \underline{\mathbf{J}}} || \mathbb{Q}_{\underline{\mathbf{Z}}} \otimes \mathbb{Q}_{\underline{\mathbf{J}}})$$

Shannon entropy:

Given $\underline{\mathbf{X}} \sim \mathbb{Q} \ll \mu \quad \frac{d\mathbb{Q}}{d\mu} =: q$

$$H(\mathbb{Q}) = H(\underline{\mathbf{X}}) = -\mathbb{E}_{\underline{\mathbf{X}}} \log q(\underline{\mathbf{X}}) = -\int_{\mathcal{X}} q(x) \log q(x) \mu(dx)$$

Conditional Entropy:

$$H(\underline{\mathbf{X}} | \underline{\mathbf{Y}}) = \mathbb{E}_{\underline{\mathbf{Y}}} H(\mathbb{Q}_{\underline{\mathbf{X}} | \underline{\mathbf{Y}}}) = \mathbb{E}_{\underline{\mathbf{Y}}} \int_{\mathcal{X}} q(x | \underline{\mathbf{Y}}) \log q(x | \underline{\mathbf{Y}}) \mu(dx)$$

Chain Rule:

$$H(\underline{\mathbf{X}}_1, \dots, \underline{\mathbf{X}}_n) = \sum_{i \in [n]} H(\underline{\mathbf{X}}_i | \underline{\mathbf{X}}_1, \dots, \underline{\mathbf{X}}_{i-1})$$

Mutual Information and Entropy

$$I(\underline{\mathbf{X}}, \underline{\mathbf{Y}}) = H(\underline{\mathbf{X}}) + H(\underline{\mathbf{Y}}) - H(\underline{\mathbf{Y}} | \underline{\mathbf{X}})$$

Notice for the mutual information

$$\begin{aligned} I(\underline{\mathbf{Z}}, \underline{\mathbf{J}}) &:= D(\mathbb{Q}_{\underline{\mathbf{Z}}, \underline{\mathbf{J}}} || \mathbb{Q}_{\underline{\mathbf{Z}}} \otimes \mathbb{Q}_{\underline{\mathbf{J}}}) = \mathbb{E}_{\underline{\mathbf{Z}}, \underline{\mathbf{J}}} \log \frac{d\mathbb{Q}_{\underline{\mathbf{Z}}, \underline{\mathbf{J}}}}{d\mathbb{Q}_{\underline{\mathbf{Z}}} d\mathbb{Q}_{\underline{\mathbf{J}}}} \\ &= \mathbb{E}_{\underline{\mathbf{J}}} \mathbb{E}_{\underline{\mathbf{Z}} | \underline{\mathbf{J}}} \log \frac{d\mathbb{Q}_{\underline{\mathbf{Z}} | \underline{\mathbf{J}}} d\mathbb{Q}_{\underline{\mathbf{J}}}}{d\mathbb{Q}_{\underline{\mathbf{Z}}} d\mathbb{Q}_{\underline{\mathbf{J}}}} \\ &= \mathbb{E}_{\underline{\mathbf{J}}} D(\mathbb{Q}_{\underline{\mathbf{Z}} | \underline{\mathbf{J}}} || \mathbb{Q}_{\underline{\mathbf{Z}}}) \end{aligned}$$

1.2.3 Fano and M-ary Testing Intuition

Recall in M ary testing problem our observation $\mathbf{Z} \sim \frac{1}{M} \sum_{j \in [M]} \mathbb{P}_{\theta^j}$ i.e follows a mixture distribution and the goal is to recover from which index $\mathbf{J} = j$ a given observation $\mathbf{Z} = Z$ has been drawn. In the extreme case where $\mathbf{Z} \perp \mathbf{J}$ observing Z has no value to our problem. A way of measuring the "amount" of depedence between rv is mutual information $I(\mathbf{Z}, \mathbf{J}) \geq 0$. Given our setting we can write the mutual information in the following way

$$I(\mathbf{Z}, \mathbf{J}) = \mathbb{E}_{\mathbf{J}} D(\mathbb{Q}_{\mathbf{Z}|\mathbf{J}} \| \mathbb{Q}_{\mathbf{Z}}) \quad (7)$$

$$= \frac{1}{M} \sum_{j \in [M]} D(\mathbb{P}_{\theta^j} \| \bar{\mathbb{Q}}) \quad (8)$$


Meaning the mutual information is small if the distributions \mathbb{P}_{θ^j} are hard to distinguish from the mixture distribution $\bar{\mathbb{Q}}$ on average.

1.2.4 Fano's bound for M-ary testing problem

Lemma 1.1 Fanno


Given the M-ary testing problem setting from section 1.1 Fanno's Lemma says

$$\mathbb{Q}\{\psi(\mathbf{Z}) \neq \mathbf{J}\} \geq 1 - \frac{I(\mathbf{Z}, \mathbf{J}) + \log(2)}{\log(M)} \quad (9)$$

Proof. The Proof of Fanno's lemma is skipped and can be found in Wainright book's [2]. A proof that I personally liked can be found in Giraud's book [1]  \square

Lemma 1.2 Mutual Information Bound for $\mathbf{Z}|\mathbf{J} \sim \mathbb{Q}_{\mathcal{N}(\mu, \sigma^2)}$

$$I(\mathbf{Z}, \mathbf{J}) \leq \frac{1}{2} \left\{ \log \det \text{cov}(\mathbf{Z}) - \frac{1}{M} \sum_{j \in [M]} \log \det \Sigma_j \right\} \quad (10)$$

Proof. This lemma comes from Wainright [2]  \square

1.3 Minimax Lower Bounds for Sparse Linear Regression Recovery Problem

1.3.1 Goal

"Give a lower bound on any procedure (Algorithms) for recovering the support of θ depending on $(n, \theta, |Supp(\theta)|)$ ".

1.3.2 Setting

Consider the following setting

$$\mathbf{x}_i \sim \mathcal{N}(0, \mathbb{I}_{d \times d}) \quad \forall i \in [n] \quad (11)$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \quad \forall i \in [n] \quad (12)$$

$$\mathbf{y}_i = \langle \mathbf{x}_i, \theta^* \rangle + \epsilon_i \quad (13)$$

We assume the data to be iid distributed i.e

$$(\mathbf{x}_i, \mathbf{y}_i) \sim iid \quad P_{\theta^*} \in \mathcal{P} := \{P_{\theta}; \theta \in \Theta\} \quad (14)$$

$$\mathbf{D}_n := \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in [n]} \sim \otimes_{i \in [n]} P_{\theta^*} =: \mathbb{P}_{\theta^*} \quad (15)$$

We also define the matrix form

$$\mathbf{Y}_n = (\mathbf{y}_1, \dots, \mathbf{y}_n) \quad \mathbf{X}_n = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T \quad (16)$$

$$\mathbf{Y}_n = \mathbf{X}_n \theta^* + \epsilon \quad (17)$$

where $\theta^* \in \mathbb{S}(k, d) =: \Theta$

and $\mathbb{S}(s, d) := \{\theta \in \mathbb{R}^d; \quad |\theta|_0 = s < d; \quad \theta_j \geq \theta_{min} \quad \forall j \in supp(\theta)\}$

1.3.3 Minimax to M-ary testing

Assume $(\mathbb{P}_{\theta})_{\theta \in \Theta}$ a set of Prob distribution on a measurable space $(\mathcal{D}, \mathcal{A})$. We have access to an observation $(\mathbf{Y}_n, \mathbf{X}_n) = \mathbf{D}_n \in \mathcal{D}$ with $\mathbf{D}_n \sim \mathbb{P}_{\theta}$. The goal is to recover the support of θ from \mathbf{D}_n with measurable map $\hat{\mathbf{S}} : \mathcal{D} \rightarrow supp(\Theta)$. We define the following metric on Θ

$$\mathbb{I}\{\hat{\mathbf{S}}(\mathbf{D}_n) \neq supp(\theta)\} \quad (18)$$

Recalling that minimax risk correspond to the best possible error uniformly over the class Θ we have

$$\inf_{\hat{\mathbf{S}}: \mathcal{D} \rightarrow supp(\Theta)} \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_{\theta}} \mathbb{I}\{\hat{\mathbf{S}}(\mathbf{D}_n) \neq supp(\theta)\} = \quad (19)$$

$$\inf_{\hat{\mathbf{S}}: \mathcal{D} \rightarrow supp(\Theta)} \sup_{\theta \in \Theta} \mathbb{P}_{\theta} \{\hat{\mathbf{S}}(\mathbf{D}_n) \neq supp(\theta)\} \quad (20)$$

We can define the following finite subset $\Theta^M \subseteq \Theta$ as follow

$$\Theta^M := \left\{ \theta^l \in \mathbb{S}(k, d); \quad \theta_j^l = \theta_{min} \quad \forall j \in supp(\theta^l) \quad \forall l \in [M] := \left[\binom{d}{k} \right]; \quad supp(\theta^l) \in \Gamma_k \right\} \quad (21)$$

where

$$\Gamma_k := \left\{ T; \quad T \subset \{1, \dots, d\}; \quad |T| = k \right\} \quad (22)$$

Notice that

$$|\Gamma_k| = \binom{d}{k} = M = |\Theta^M| \quad (23)$$

and recall that

$$\mathbb{S}(k, d) := \left\{ \theta \in \mathbb{R}^d; \quad |\theta|_0 = k \ll d; \quad \theta_j \geq \theta_{\min} \quad \forall j \in \text{supp}(\theta) \right\} \quad (24)$$

We have the following lemma:

Lemma 1.3 Minimax to M-ary Testing Problem for Sparse Linear Regression Support Recovery

Given the above setup the minimax risk can be lower bounded by a quantity corresponding to an M-ary testing problem:

$$\inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\Theta)} \sup_{\theta \in \Theta} \mathbb{P}_{\theta} \{ \hat{\mathbf{S}}(\mathbf{D}_n) \neq \text{supp}(\theta) \} \geq \inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\Theta)} \mathbb{Q}_{\mathbf{J}, \mathbf{D}_n} \{ \hat{\mathbf{S}}(\mathbf{D}_n) \neq \mathbf{J} \} \quad (25)$$

where

$$\mathbf{J} \sim \text{Unif}[\Gamma_k] \quad (26)$$

Proof. we can lower the minimax setting as follow:

$$\inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\Theta)} \sup_{\theta \in \Theta} \mathbb{P}_{\theta} \{ \hat{\mathbf{S}}(\mathbf{D}_n) \neq \text{supp}(\theta) \} \geq \inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\Theta)} \max_{\theta \in \Theta^M} \mathbb{P}_{\theta} \{ \hat{\mathbf{S}}(\mathbf{D}_n) \neq \text{supp}(\theta) \} \quad (27)$$

$$= \inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\Theta)} \max_{\text{supp}(\theta^l) \in \Gamma_k} \mathbb{P}_{\theta^l} \{ \hat{\mathbf{S}}(\mathbf{D}_n) \neq \text{supp}(\theta^l) \} \quad (28)$$

$$\geq \inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\Theta)} \frac{1}{|\Gamma_k|} \sum_{S^l \in \Gamma_k} \mathbb{P}_{\theta^l} \{ \hat{\mathbf{S}}(\mathbf{D}_n) \neq S^l \} \quad | \quad S^l := \text{supp}(\theta^l) \quad (29)$$

$$(30)$$

Now we notice that the last equation has the same form as an M-ary testing problem

$$\mathbf{J} := \mathbf{S}^1 \sim \text{Unif}[\Gamma_k] \quad (31)$$

$$\mathbf{Z} := \mathbf{D}_n \sim \mathbb{P}_{\theta^1 \triangleq \mathbf{S}^1 = \mathbf{J}} \quad (32)$$

$$\hat{S} := \psi : \mathcal{Z} := \mathcal{D} \rightarrow \Gamma_k \quad (33)$$

$$\mathbb{Q}(\mathbf{J}, \mathbf{Z}) = \mathbb{Q}(\mathbf{S}^1, \mathbf{D}_n) \quad (34)$$

hence we can write

$$\inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\Theta)} \frac{1}{|\Gamma_k|} \sum_{S^l \in \Gamma_k} \mathbb{P}_{\theta^l} \{ \hat{\mathbf{S}}(\mathbf{D}_n) \neq S^l \} \quad (35)$$

$$= \inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\Theta)} \mathbb{Q}_{\mathbf{J}, \mathbf{D}_n} \{ \hat{\mathbf{S}}(\mathbf{D}_n) \neq \mathbf{J} \} \quad (36)$$

where \mathbb{Q} denotes the joint distribution over \mathbf{D}_n and \mathbf{J} .

Note: for a fixed $S^l \in \Gamma_k$ corresponding to the l^{th} subset we have a **unique** $\theta^l \in \Theta^M$ which allow 35 to be true i.e $\theta^l \triangleq \mathbf{S}^1$ where both follow the same distribution uniformly over a discrete set of the same size. This uniqueness motivate the choice of $|\Theta^M| = \binom{d}{k}$. if we choose a discretization of Θ with more vector than $\binom{d}{k}$ uniqueness of θ^l would be broken.

Note: Eq 36 is true because of the following

$$\frac{1}{|\Gamma_k|} \sum_{S^l \in \Gamma_k} \mathbb{P}_{\theta^l} \{ \hat{\mathbf{S}}(\mathbf{D}_n) \neq S^l \} = \frac{1}{M} \sum_{j \in \mathcal{J}} \mathbb{Q}_{\mathbf{Z}|\mathbf{J}} \{ \mathbf{Z} \neq \mathbf{J} | \mathbf{J} = j \} \quad (37)$$

$$= \int_{\mathcal{J}} \mathbb{Q}_{\mathbf{Z}|\mathbf{J}} \{ \mathbf{Z} \neq \mathbf{J} | \mathbf{J} = j \} d\mathbb{Q}_{\mathbf{J}} \{ \mathbf{J} = j \} \quad (38)$$

$$= \mathbb{E}_{\mathbf{J}} \left[\mathbb{Q}_{\mathbf{Z}|\mathbf{J}} \{ \mathbf{Z} \neq \mathbf{J} \} \right] \quad (39)$$

$$= \mathbb{E}_{\mathbf{J}} \left[\mathbb{E}_{\mathbf{Z}|\mathbf{J}} \left[\mathbb{I}(\mathbf{Z} \neq \mathbf{J}) \right] \right] \quad (40)$$

$$= \mathbb{E}_{\mathbf{Z}, \mathbf{J}} \left[\mathbb{I}(\mathbf{Z} \neq \mathbf{J}) \right] \quad (41)$$

$$= \mathbb{Q}_{\mathbf{Z}, \mathbf{J}} \{ \mathbf{Z} \neq \mathbf{J} \} \quad (42)$$

☺ □

1.3.4 Fano's method

In order to use fano's method we will condition on a particular instance of the Design matrix $\mathbf{X}_n = \{x_i\}_{i \in [n]} =: \mathbb{X}_n$ and $\mathbf{J} \sim \text{Uni}[M]$.

Claim 1.1 Applying Fanno to Lemma 1.3

A direct application of Fanno's Lemma 1.1 to lemma 1.3 yield the following lower bound for the minimax risk

$$\inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\Theta)} \sup_{\theta \in \Theta} \mathbb{P}_{\theta} \{ \hat{\mathbf{S}}(\mathbf{D}_n) \neq \text{supp}(\theta) \} \geq 1 - \frac{\mathbb{E}_{\mathbf{X}_n} I(\mathbf{Y}_n, \mathbf{J} | \mathbf{X}_n) + \log 2}{\log M} \quad (43)$$

Proof.

$$\inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\Theta)} \mathbb{Q}_{\mathbf{D}_n, \mathbf{J}} \{ \hat{\mathbf{S}}(\mathbf{D}_n) \neq \mathbf{J} \} := \inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\Theta)} \mathbb{Q}_{\mathbf{Y}_n, \mathbf{X}_n, \mathbf{J}} \{ \hat{\mathbf{S}}(\mathbf{Y}_n, \mathbf{X}_n) \neq \mathbf{J} \} \quad (44)$$

$$= \inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\Theta)} \mathbb{E}_{\mathbf{X}_n} \mathbb{Q}_{\mathbf{Y}_n, \mathbf{J} | \mathbf{X}_n} \{ \hat{\mathbf{S}}(\mathbf{Y}_n, \mathbf{X}_n) \neq \mathbf{J} | \mathbf{X}_n = \{x_i\}_{i \in [n]} \} \quad (45)$$

$$\geq 1 - \frac{\mathbb{E}_{\mathbf{X}_n} I(\mathbf{Y}_n, \mathbf{J} | \mathbf{X}_n) + \log 2}{\log M} \quad (46)$$

Notice the considerable advantage of fanno's bound; it is procedure independent i.e independent of $\hat{\mathbf{S}}(\mathbf{D}_n)$

😊 □

Hence we are interested in upper bounding the term

$$\mathbb{E}_{\mathbf{X}_n} I(\mathbf{Y}_n, \mathbf{J} | \mathbf{X}_n)$$

Claim 1.2 A Upper Bound on $I(\mathbf{Y}_n, \mathbf{J} | \mathbf{X}_n)$

Given the setup in section 1.3.2, the mutual information $I(\mathbf{Y}_n, \mathbf{J} | \mathbf{X}_n)$ can be upper bounded as follow

$$I(\mathbf{Y}_n, \mathbf{J} | \mathbf{X}_n) \leq \sum_{i \in [n]} I(\mathbf{y}_i, \mathbf{J} | \mathbf{x}_i) \quad (47)$$

Proof. Given the setting in section 1.3.2 we have for a given observation $\mathbf{x}_i = x_i$ the following model for the output variable \mathbf{y}_i

$$(\mathbf{y}_i | \mathbf{x}_i = x_i) =: \mathbf{y}_i^{x_i} = \langle x_i, \theta^{\mathbf{J}} \rangle + \epsilon_i \quad \forall x_i \in \mathcal{X} \quad (48)$$

hence the distribution over $\mathbf{y}_i^{x_i}$ is

$$\mathbb{Q}_{\mathbf{y}_i^{x_i}}(\cdot) = \int_{j \in \mathcal{J}} \mathbb{Q}_{\mathbf{y}_i^{x_i} | \mathbf{J}}(\cdot | \mathbf{J} = j) \mathbb{Q}_{\mathbf{J}}(\mathbf{J} = j) \quad (49)$$

$$= \frac{1}{M} \sum_{j \in [M]} \mathbb{Q}_{\mathbf{y}_i^{x_i} | \mathbf{J}}(\cdot | \mathbf{J} = j) \quad (50)$$

$$= \frac{1}{M} \sum_{j \in [M]} \mathbb{Q}_{\mathcal{N}(\langle x_i, \theta^{\mathbf{J}=j} \rangle, \sigma^2)}(\cdot) \quad (51)$$

also notice that $\mathbf{y}_i^{x_i} \not\perp \mathbf{y}_j^{x_j} \quad \forall i \neq j$ but we have $(\mathbf{y}_i^{x_i} | \mathbf{J}) \perp (\mathbf{y}_j^{x_j} | \mathbf{J}) \quad \forall i \neq j$ because of the setting assumptions.

We can upper bound the mutual information $I(\mathbf{Y}_n, \mathbf{J} | \mathbf{X}_n)$ as follow

$$I(\mathbf{Y}_n, \mathbf{J} | \mathbf{X}_n) = H(\mathbf{y}_1, \dots, \mathbf{y}_n | \mathbf{X}_n) - H(\mathbf{y}_1, \dots, \mathbf{y}_n | \mathbf{J}, \mathbf{X}_n) \quad (52)$$

$$=: H_{\mathbf{X}_n}(\mathbf{y}_1, \dots, \mathbf{y}_n) - H_{\mathbf{X}_n}(\mathbf{y}_1, \dots, \mathbf{y}_n | \mathbf{J}) \quad (53)$$

$$= \sum_{i \in [n]} H_{\mathbf{X}_n}(\mathbf{y}_i | \mathbf{y}_1, \dots, \mathbf{y}_{i-1}) - H_{\mathbf{X}_n}(\mathbf{y}_1, \dots, \mathbf{y}_n | \mathbf{J}) \quad \text{Chain rule} \quad (54)$$

$$\leq \sum_{i \in [n]} H_{\mathbf{X}_n}(\mathbf{y}_i) - H_{\mathbf{X}_n}(\mathbf{y}_1, \dots, \mathbf{y}_n | \mathbf{J}) \quad \text{Conditioning reduces Entropy} \quad (55)$$

$$= \sum_{i \in [n]} H_{\mathbf{X}_n}(\mathbf{y}_i) - \sum_{i \in [n]} H_{\mathbf{X}_n}(\mathbf{y}_i | \mathbf{J}) \quad \text{Chain rule and Independence} \quad (56)$$

$$= \sum_{i \in [n]} H_{\mathbf{X}_n}(\mathbf{y}_i) - H_{\mathbf{X}_n}(\mathbf{y}_i | \mathbf{J}) \quad (57)$$

$$= \sum_{i \in [n]} I(\mathbf{y}_i, \mathbf{J} | \mathbf{X}_n) \quad (58)$$

$$= \sum_{i \in [n]} I(\mathbf{y}_i, \mathbf{J} | \mathbf{x}_i) \quad (59)$$

😊 □

Claim 1.3 An Upper Bound on $I(\mathbf{y}_i, \mathbf{J} | \mathbf{x}_i)$

Using lemma 1.2 we obtain the following upper bound on $I(\mathbf{y}_i, \mathbf{J} | \mathbf{x}_i)$

$$I(\mathbf{y}_i, \mathbf{J} | \mathbf{x}_i) \leq \frac{1}{2} \left\{ \log \frac{\text{var}(\mathbf{y}_i | \mathbf{x}_i)}{\sigma^2} \right\} \quad (60)$$

Proof. We notice that $\mathbf{J} \sim \text{Uni}[M]$ and $(\mathbf{y}_i | \mathbf{J} = j, \mathbf{x}_i = x_i) \sim \mathcal{N}(\langle x_i, \theta^{\mathbf{J}=j} \rangle, \sigma^2)$ hence due to lemma 1.2 we have

$$I(\mathbf{y}_i, \mathbf{J} | \mathbf{x}_i = x_i) =: I_{x_i}(\mathbf{y}_i, \mathbf{J}) \leq \frac{1}{2} \left\{ \log \text{var}(\mathbf{y}_i^{x_i}) - \frac{1}{M} \sum_{j \in [M]} \log \text{var}(\mathbf{y}_i^{x_i} | \mathbf{J} = j) \right\} \quad \forall x_i \in \mathcal{X} \quad (61)$$

$$= \frac{1}{2} \left\{ \log \frac{\text{var}(\mathbf{y}_i^{x_i})}{\sigma^2} \right\} \quad \forall x_i \in \mathcal{X} \quad (62)$$

$$(63)$$

😊 □

Lemma 1.4 Final Minimax Lower Bound for Sparse lin. Reg. Support Recovery

Using claims 1.1, 1.2, 1.3 as well as lemma 1.3 we have the following minimax lower bound in term of the parameters $\theta_{\min}, k, d, \sigma^2, n$

$$\inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\Theta)} \sup_{\theta \in \Theta} \mathbb{P}_{\theta} \{ \hat{\mathbf{S}}(\mathbf{D}_n) \geq 1 - \frac{\frac{n}{2} \log(1 + \frac{k\theta_{\min}^2}{\sigma^2}) + \log 2}{\log \binom{d}{k}} \} \quad (64)$$

Proof. Recalling $\mathbf{J} \sim \text{Uni}[M]$ and $(\mathbf{y}_i | \mathbf{J} = j, \mathbf{x}_i = x_i) \sim \mathcal{N}(\langle x_i, \theta^{\mathbf{J}=j} \rangle, \sigma^2)$ hence due to lemma ?? we have

$$I(\mathbf{y}_i, \mathbf{J} | \mathbf{x}_i = x_i) =: I_{x_i}(\mathbf{y}_i, \mathbf{J}) \leq \frac{1}{2} \left\{ \log \text{var}(\mathbf{y}_i^{x_i}) - \frac{1}{M} \sum_{j \in [M]} \log \text{var}(\mathbf{y}_i^{x_i} | \mathbf{J} = j) \right\} \quad \forall x_i \in \mathcal{X} \quad (65)$$

$$= \frac{1}{2} \left\{ \log \frac{\text{var}(\mathbf{y}_i^{x_i})}{\sigma^2} \right\} \quad \forall x_i \in \mathcal{X} \quad (66)$$

$$\iff \quad (67)$$

$$\sum_{i \in [n]} I_{x_i}(\mathbf{y}_i, \mathbf{J}) \leq \sum_{i \in [n]} \frac{1}{2} \left\{ \log \frac{\text{var}(\mathbf{y}_i^{x_i})}{\sigma^2} \right\} \quad \forall x_i \in \mathcal{X} \quad (68)$$

As the data $(\mathbf{x}_i, \mathbf{y}_i)$ are IID we have

$$I(\mathbf{Y}_n, \mathbf{J} | \mathbf{X}_n) \leq \frac{n}{2} \log \frac{\text{var}(\mathbf{y}_1 | \mathbf{x}_1)}{\sigma^2} \quad (69)$$

Recalling that we aim at lower bounding the average over \mathbf{X}_n

$$\mathbb{E}_{\mathbf{X}_n} I(\mathbf{Y}_n, \mathbf{J} | \mathbf{X}_n) \leq \frac{n}{2} \log \frac{\mathbb{E}_{\mathbf{x}_1} \text{var}(\mathbf{y}_1 | \mathbf{x}_1)}{\sigma^2} \quad (70)$$

Using concavity of log and Jensen. Writing more explicit we need an upper bound on

$$\mathbb{E}_{\mathbf{x}_1} \text{var}(\mathbf{y}_1(\mathbf{J}) | \mathbf{x}_1) \leq \mathbb{E}_{\mathbf{x}_1} \mathbb{E}_{\mathbf{y}_1, \mathbf{J} | \mathbf{x}_1} \left[\mathbf{y}_1(\mathbf{J})^2 | \mathbf{x}_1 \right] \quad \text{Def. of Variance} \quad (71)$$

$$\text{Recall} \quad (\mathbf{y}_1(\mathbf{J}) | \mathbf{x}_1 = x_1) = \langle x_1, \theta^{\mathbf{J}} \rangle + \epsilon_1 \quad (72)$$

$$\mathbb{E}_{\mathbf{y}_1, \mathbf{J} | \mathbf{x}_1} \left[\mathbf{y}_1(\mathbf{J})^2 | \mathbf{x}_1 \right] = \mathbb{E}_{\mathbf{J}} \mathbb{E}_{\mathbf{y}_1 | \mathbf{x}_1, \mathbf{J}} \left[\mathbf{y}_1^2 | \mathbf{x}_1, \mathbf{J} \right] \quad \text{Iterated Expectation} \quad (73)$$

$$\text{Recall} \quad (\mathbf{y}_1 | \mathbf{x}_1 = x_1, \mathbf{J} = j) = \langle x_1, \theta^j \rangle + \epsilon_1 \quad (74)$$

$$\iff \quad (75)$$

$$\mathbb{E}_{\mathbf{x}_1} \text{var}(\mathbf{y}_1(\mathbf{J}) | \mathbf{x}_1) \leq \mathbb{E}_{\mathbf{x}_1} \mathbb{E}_{\mathbf{J}} \mathbb{E}_{\mathbf{y}_1 | \mathbf{x}_1, \mathbf{J}} \left[(\langle x_1, \theta^j \rangle + \epsilon_1)^2 \right] \quad (76)$$

$$= \mathbb{E}_{\mathbf{x}_1} \mathbb{E}_{\mathbf{J}} \mathbb{E}_{\epsilon} \left[(\langle x_1, \theta^j \rangle + \epsilon_1)^2 \right] \quad (77)$$

$$= \mathbb{E}_{\mathbf{x}_1} \mathbb{E}_{\mathbf{J}} \mathbb{E}_{\epsilon} \left[\text{Tr} \{ x_1^T \theta^j \otimes \theta^j x_1 \} + 2\epsilon_1 x_1^T \theta^j + \epsilon_1^2 \right] \quad (78)$$

$$= \mathbb{E}_{\mathbf{x}_1} \mathbb{E}_{\mathbf{J}} \left[\text{Tr} \{ x_1^T \theta^{\mathbf{J}} \otimes \theta^{\mathbf{J}} x_1 \} \right] + \sigma^2 \quad \text{See noise dist.} \quad (79)$$

$$= \mathbb{E}_{\mathbf{x}_1} \text{Tr} \{ \mathbf{x}_1^T \mathbb{E}_{\mathbf{J}} [\theta^{\mathbf{J}} \otimes \theta^{\mathbf{J}}] \mathbf{x}_1 \} + \sigma^2 \quad (80)$$

Due to the distribution of \mathbf{J} we have

$$\mathbb{E}_{\mathbf{J}} \theta^{\mathbf{J}} \otimes \theta^{\mathbf{J}} = \frac{1}{M} \sum_{j \in [M]} \theta^j \otimes \theta^j \quad (81)$$

$$\implies \quad (82)$$

$$\frac{1}{M} \sum_{j \in [M]} \mathbb{E}_{\mathbf{x}_1} \text{Tr} \{ \mathbf{x}_1^T \theta^j \otimes \theta^j \mathbf{x}_1 \} + \sigma^2 = \frac{1}{M} \sum_{j \in [M]} \text{Tr} \{ \theta^j \otimes \theta^j \mathbb{E}_{\mathbf{x}_1} \mathbf{x}_1 \otimes \mathbf{x}_1 \} + \sigma^2 \quad (83)$$

$$= \frac{1}{M} \sum_{j \in [M]} \text{Tr} \{ \theta^j \otimes \theta^j \} + \sigma^2 \quad \text{Isotropic Gaussian Covariates} \quad (84)$$

Recalling eq. 21 we have

$$\theta^j \in \Theta^M \implies \text{Tr} \{ \theta^j \otimes \theta^j \} = k\theta_{\min}^2 \quad \forall j \in [M] \quad (85)$$

Hence we have

$$\mathbb{E}_{\mathbf{x}_1} \text{var}(\mathbf{y}_1(\mathbf{J})|\mathbf{x}_1) \leq k\theta_{min}^2 + \sigma^2 \quad (86)$$

$$\Longleftrightarrow \quad (87)$$

$$\mathbb{E}_{\mathbf{x}_n} I(\mathbf{Y}_n, \mathbf{J}|\mathbf{x}_n) \leq \frac{n}{2} \log \frac{\mathbb{E}_{\mathbf{x}_1} \text{var}(\mathbf{y}_1|\mathbf{x}_1)}{\sigma^2} \quad (88)$$

$$\leq \frac{n}{2} \log \frac{k\theta_{min}^2 + \sigma^2}{\sigma^2} \quad (89)$$

$$\Longleftrightarrow \quad (90)$$

$$\inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\Theta)} \mathbb{Q}_{\mathbf{D}_n, \mathbf{J}}\{\hat{\mathbf{S}}(\mathbf{D}_n) \neq \mathbf{J}\} \geq 1 - \frac{\frac{n}{2} \log(1 + \frac{k\theta_{min}^2}{\sigma^2}) + \log 2}{\log M} \quad (91)$$

Moreover

$$\inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\Theta)} \mathbb{Q}_{\mathbf{D}_n, \mathbf{J}}\{\hat{\mathbf{S}}(\mathbf{D}_n) \neq \mathbf{J}\} \geq 1 - \frac{\frac{n}{2} \log(1 + \frac{k\theta_{min}^2}{\sigma^2}) + \log 2}{\log M} \quad (92)$$

$$\Longleftrightarrow \quad (93)$$

$$\inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\Theta)} \sup_{\theta \in \Theta} \mathbb{P}_{\theta}\{\hat{\mathbf{S}}(\mathbf{D}_n) \neq \text{supp}(\theta)\} \geq 1 - \frac{\frac{n}{2} \log(1 + \frac{k\theta_{min}^2}{\sigma^2}) + \log 2}{\log M} \quad (94)$$

😊 □

1.4 Results

Result 1.1 A Lower on the number of data needed to be better than fully random decision process

In order for any procedure or algorithm to achieve a probability of error in the recovery process **below** 1/2 we need at least

$$n > \frac{\log \binom{d}{k} + \log(2)}{\log(1 + \frac{k\theta_{min}^2}{\sigma^2})} \quad (95)$$

data.

Proof. From lemma 1.4

$$\frac{1}{2} > \inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\Theta)} \mathbb{Q}_{\mathbf{D}_n, \mathbf{J}}\{\hat{\mathbf{S}}(\mathbf{D}_n) \neq \mathbf{J}\} \quad (96)$$

$$\geq 1 - \frac{\frac{n}{2} \log(1 + \frac{k\theta_{min}^2}{\sigma^2}) + \log 2}{\log M} \quad (97)$$

$$\Longleftrightarrow \quad (98)$$

$$n > \frac{\log \binom{d}{k} + \log(2)}{\log(1 + \frac{k\theta_{min}^2}{\sigma^2})} \quad (99)$$

😊 □

2 Lower Bound for Sparse Causal Estimator

2.1 Goal and Problem Description

Assume a data set of patient's covariates (think: blood analysis, symptomatic, etc...) drawn from some population. We give treatment with a certain probability to this population set and we observe the outcome variable on both patient with treatment and without. Further assume an estimate of the difference of effects between the two groups (treatment vs no treatment) which is assumed to be sparse. The goal is to show a lower bound on any procedure (algorithm) for recovering the support of the difference in treatment effect.

2.2 Formalism

Sparse Vector space: $\mathbb{S}(s, d) := \{\theta \in \mathbb{R}^d; \quad |\theta|_0 = s \ll d; \quad \theta_j \geq \theta_{min} \quad \forall j \in \text{supp}(\theta)\}$

2.2.1 Model

$$\mathbf{T}_i \sim \text{Ber}(p) \quad \text{i.e } \mathbf{T}_i \in \{0, 1\} \quad (100)$$

$$\epsilon_i^{\mathbf{T}=0} \sim \mathcal{N}(0, \sigma_{\mathbf{T}=0}^2) \quad (101)$$

$$\epsilon_i^{\mathbf{T}=1} \sim \mathcal{N}(0, \sigma_{\mathbf{T}=1}^2) \quad (102)$$

$$\mathbf{x}_i \sim \mathcal{N}(0, \mathbb{I}_{d \times d}) \quad (\text{Isotropic gaussian}) \quad (103)$$

$$\mathbf{y}_i^{\mathbf{T}} = \langle \mathbf{x}_i, \vartheta^{\mathbf{T}} \rangle + \epsilon_i^{\mathbf{T}} \quad (104)$$

where we assume

$$\vartheta^{\mathbf{T}=0}, \vartheta^{\mathbf{T}=1} \in \mathbb{R}^d \quad (105)$$

$$\vartheta^{\mathbf{T}=0} - \vartheta^{\mathbf{T}=1} \in \mathbb{S}(s, d) \quad (106)$$

Hence we can derive the following equivalence for the linear parameter:

$$\vartheta^{\mathbf{T}=0} := \beta, \beta \in \mathbb{R}^d \quad (107)$$

$$\vartheta^{\mathbf{T}=1} := \beta + \theta, \theta \in \mathbb{S}(s, d) \quad (108)$$

we can rewrite the model as

$$\underline{\mathbf{y}}_i^{\mathbf{T}=0} = \langle \underline{\mathbf{x}}_i, \beta \rangle + \underline{\epsilon}_i^{\mathbf{T}=0} \quad (109)$$

$$\underline{\mathbf{y}}_i^{\mathbf{T}=1} = \langle \underline{\mathbf{x}}_i, \beta + \theta \rangle + \underline{\epsilon}_i^{\mathbf{T}=1} \quad (110)$$

We can even write the model in a more compact form

$$\underline{\mathbf{y}}_i = \langle \underline{\mathbf{x}}_i, \beta + \underline{\mathbf{T}}_i \theta \rangle + \underline{\epsilon}_i \quad (111)$$

if we assume $\sigma_{\mathbf{T}=0} = \sigma_{\mathbf{T}=1}$

Defining the following

$$\underline{\mathbf{Y}}_n := (\underline{\mathbf{y}}_1, \dots, \underline{\mathbf{y}}_n)^T \in \mathbb{R}^n \quad | \quad \underline{\mathbf{X}}_n = (\underline{\mathbf{x}}_1, \dots, \underline{\mathbf{x}}_n)^T \in \mathbb{R}^{n \times d} \quad | \quad \underline{\mathbf{T}}_n = \text{diag}(\underline{\mathbf{T}}_1, \dots, \underline{\mathbf{T}}_n) \in \mathbb{R}^{n \times d} \quad (112)$$

$$\underline{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T \in \mathbb{R}^n \quad (113)$$

we can write the model in matrix form:

$$\underline{\mathbf{Y}}_n = \underline{\mathbf{X}}_n \beta + \underline{\mathbf{T}}_n \underline{\mathbf{X}}_n \theta + \underline{\epsilon} \quad (114)$$

2.2.2 Data

Assuming:

$$\{(\underline{\mathbf{y}}_i, \underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i)\}_{i=1}^n \sim iid \quad P_{(\beta, \theta), p} \in \mathcal{P} := \{P_{(\beta, \theta), p} \mid (\beta, \theta) \in \Theta \mid p \in [0, 1]\} \quad (115)$$

with parameter space

$$\Theta := \mathbb{S}(d, s) \quad \otimes \quad \{\beta \in \mathbb{R}^d; \quad \|\beta\|_2 \leq 1\} \quad (116)$$

we define the random sample as:

$$\underline{\mathbb{D}}_n := ((\underline{\mathbf{y}}_1, \underline{\mathbf{x}}_1, \underline{\mathbf{T}}_1), \dots, (\underline{\mathbf{y}}_n, \underline{\mathbf{x}}_n, \underline{\mathbf{T}}_n)) \sim \mathbb{P}_{(\beta, \theta), p} := P_{(\beta, \theta), p}^{\otimes n} \quad (117)$$

We assume the parameter p to be given and we write the probability mass function of $\underline{\mathbf{T}}_i$ as follow

$$f_p(T_i) = p^{T_i} \underbrace{(1-p)^{1-T_i}}_{=: q} \quad (118)$$

2.3 Minimax to M-ary Hypotheses Testing

Assuming the following paramter space Θ

$$\mathbb{B}_2^d := \mathbb{B}(\mathbb{R}^d, \|\cdot\|_2) = \{\beta \in \mathbb{R}^d; \quad \|\beta\|_2 \leq 1\} \quad (119)$$

$$\Theta := \mathbb{S}(d, k) \quad \otimes \quad \mathbb{B}_2^d \quad (120)$$

and considering the following family of distribution

$$(\mathbb{P}_{(\beta, \theta)})_{(\beta, \theta) \in \Theta} \quad (121)$$

on a measurable space $(\mathcal{D}, \mathcal{A})$.

We observe the data $\underline{\mathbb{D}}_n \in \mathcal{D}$ distributed as $\underline{\mathbb{D}}_n \sim \mathbb{P}_{(\beta^*, \theta^*) \in \Theta}$.

Given the data the goal is to recover $\text{supp}(\theta^*)$ which correspond to recover

$$\text{supp}(\vartheta^{\mathbf{T}=0} - \vartheta^{\mathbf{T}=1}) \quad (\text{see 106}) \quad (122)$$

Hence we take the 0-1 loss as error metric and we define any procedure or algorithm recovering the problem as the following measurable map $\hat{\underline{\mathbf{S}}} : \mathcal{D} \rightarrow \text{supp}(\mathbb{S}(k, d))$

Hence in a minimax risk framework we want to lower bound:

$$\inf_{\hat{\underline{\mathbf{S}}} : \mathcal{D} \rightarrow \text{supp}(\mathbb{S}(k, d))} \sup_{(\beta, \theta) \in \Theta} \mathbb{P}_{(\beta, \theta)} \{ \hat{\underline{\mathbf{S}}}(\underline{\mathbb{D}}_n) \neq \text{supp}(\theta) \} \quad (123)$$

We notice similarities to the problem describe in 1.3.3. The parameter space is different, but as describe in the next section, the discretization is strongly inspired from the discretization in eq. 21.

Note: Remember that we assume the parameter p to be fixed (or given e.g $p = 1/2$)

2.3.1 Discretization of the Parameter Space

we define the following finite dimensional space

$$\Theta^M := \left\{ \omega^j = (\theta^j, \beta^j); \quad j \in [M]; \quad \theta_i^j = \theta_{\min} \forall i \in [d]; \quad \beta^j \in \mathbb{B}_2^d \right\} \quad (124)$$

As for the discretization in eq 21 we define the size of the discretized space as $M := \binom{d}{k}$ in order to have a one to one correspondence between all possible support set of a k-sparse vector and each element in the discretized space. In other world defining the set of all possible support for a given k-sparse θ as

$$\Gamma_k := \left\{ T; \quad T \subset \{1, \dots, d\}; \quad |T| = k \right\} \quad (125)$$

we have the following correspondence

$$\Gamma_k \triangleq \Theta^M \quad (126)$$

Notice that in Θ^M we specified the discretization of $\mathbb{S}(d, k)$ and it remains to define the discretization of the euclidean unit ball \mathbb{B}_2^d i.e what are β_j 's are.

2.3.2 Minimax after Discretization

Lemma 2.1 Minimax to M-ary Hypotheses Testing For Our Setup

Given the discretized space Θ^M in the same spirit as lemma 1.3 we have

$$\inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\mathbb{S}(k,d))} \sup_{(\beta, \theta) \in \Theta} \mathbb{P}_{(\beta, \theta)} \{ \hat{\mathbf{S}}(\underline{\mathbb{D}}_{\mathbf{n}}) \neq \text{supp}(\theta) \} \geq \inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\mathbb{S}(k,d))} \mathbb{Q}_{\underline{\mathbb{D}}_{\mathbf{n}}, \underline{\mathbf{J}}} \{ \hat{\mathbf{S}}(\underline{\mathbb{D}}_{\mathbf{n}}) \neq \underline{\mathbf{J}} \} \quad | \quad \underline{\mathbf{J}} \sim \text{Uni}[M] \quad (127)$$

Proof. Given the discretized space Θ^M we have

$$\sup_{(\beta, \theta) \in \Theta} \mathbb{P}_{(\beta, \theta)} \{ \hat{\mathbf{S}}(\underline{\mathbb{D}}_{\mathbf{n}}) \neq \text{supp}(\theta) \} \geq \max_{w^j \in \Theta^M} \mathbb{P}_{w^j} \{ \hat{\mathbf{S}}(\underline{\mathbb{D}}_{\mathbf{n}}) \neq \text{supp}(\theta^j) \} \quad (128)$$

$$= \max_{S^l \in \Gamma_k} \mathbb{P}_{w^l} \{ \hat{\mathbf{S}}(\underline{\mathbb{D}}_{\mathbf{n}}) \neq S^l \} \quad S^l := \text{supp}(\theta^l) \quad (129)$$

$$\geq \frac{1}{M} \sum_{S^l \in \Gamma_k} \mathbb{P}_{w^l} \{ \hat{\mathbf{S}}(\underline{\mathbb{D}}_{\mathbf{n}}) \neq S^l \} \quad (130)$$

Using the same argumentation as in the proof of lemma 1.3 we have

$$\inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\mathbb{S}(k,d))} \sup_{(\beta, \theta) \in \Theta} \mathbb{P}_{(\beta, \theta)} \{ \hat{\mathbf{S}}(\underline{\mathbb{D}}_{\mathbf{n}}) \neq \text{supp}(\theta) \} \geq \inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\mathbb{S}(k,d))} \mathbb{Q}_{\underline{\mathbb{D}}_{\mathbf{n}}, \underline{\mathbf{J}}} \{ \hat{\mathbf{S}}(\underline{\mathbb{D}}_{\mathbf{n}}) \neq \underline{\mathbf{J}} \} \quad | \quad \underline{\mathbf{J}} \sim \text{Uni}[M] \quad (131)$$

😊 □

2.4 Fano's Method

Claim 2.1 Fanno's Lemma on Lemma 2.1

$$\inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\mathbb{S}(k,d))} \sup_{(\beta, \theta) \in \Theta} \mathbb{P}_{(\beta, \theta)} \{ \hat{\mathbf{S}}(\underline{\mathbb{D}}_{\mathbf{n}}) \neq \text{supp}(\theta) \} \geq 1 - \frac{\mathbb{E}_{\underline{\mathbf{X}}_{\mathbf{n}}, \underline{\mathbb{T}}_{\mathbf{n}}} I(\underline{\mathbf{Y}}_{\mathbf{n}}, \underline{\mathbf{J}} | \underline{\mathbf{X}}_{\mathbf{n}}, \underline{\mathbb{T}}_{\mathbf{n}}) + \log 2}{\log \binom{d}{k}} \quad (132)$$

Proof. Using Fanno's lemma 1.1 we have

$$\inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\mathbb{S}(k,d))} \mathbb{Q}_{\underline{\mathbb{D}}_{\mathbf{n}}, \underline{\mathbf{J}}} \{ \hat{\mathbf{S}}(\underline{\mathbb{D}}_{\mathbf{n}}) \neq \underline{\mathbf{J}} \} = \quad (133)$$

$$\inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\mathbb{S}(k,d))} \mathbb{E}_{\underline{\mathbf{X}}_{\mathbf{n}}, \underline{\mathbb{T}}_{\mathbf{n}}} \mathbb{Q}_{\underline{\mathbf{Y}}_{\mathbf{n}}, \underline{\mathbf{J}} | \underline{\mathbf{X}}_{\mathbf{n}}, \underline{\mathbb{T}}_{\mathbf{n}}} \{ \hat{\mathbf{S}}(\underline{\mathbf{Y}}_{\mathbf{n}}, \underline{\mathbf{X}}_{\mathbf{n}}, \underline{\mathbb{T}}_{\mathbf{n}}) \neq \underline{\mathbf{J}} | \underline{\mathbf{X}}_{\mathbf{n}}, \underline{\mathbb{T}}_{\mathbf{n}} \} \quad (134)$$

$$\geq 1 - \frac{\mathbb{E}_{\underline{\mathbf{X}}_{\mathbf{n}}, \underline{\mathbb{T}}_{\mathbf{n}}} I(\underline{\mathbf{Y}}_{\mathbf{n}}, \underline{\mathbf{J}} | \underline{\mathbf{X}}_{\mathbf{n}}, \underline{\mathbb{T}}_{\mathbf{n}}) + \log 2}{\log M} \quad (135)$$

😊 □

Hence we are interested in upper bounding the quantity

$$\mathbb{E}_{\underline{\mathbf{X}}_{\mathbf{n}}, \underline{\mathbb{T}}_{\mathbf{n}}} I(\underline{\mathbf{Y}}_{\mathbf{n}}, \underline{\mathbf{J}} | \underline{\mathbf{X}}_{\mathbf{n}}, \underline{\mathbb{T}}_{\mathbf{n}}) \quad (136)$$

To upper this quantity we first claim an upper bound on $I(\underline{\mathbf{y}}_i, \underline{\mathbf{J}} | \underline{\mathbf{x}}_i = x_i, \underline{\mathbb{T}}_i = T_i)$ using simple information theory (**Important: tight bound** because independent of discretization space).

Claim 2.2 An Upper Bound for $I(\underline{\mathbf{y}}_i, \underline{\mathbf{J}} | \underline{\mathbf{x}}_i = x_i, \underline{\mathbb{T}}_i = T_i)$

Given the settings of the problem we have

$$I(\underline{\mathbf{y}}_i, \underline{\mathbf{J}} | \underline{\mathbf{x}}_i = x_i, \underline{\mathbb{T}}_i = T_i) \leq \sum_{i \in [n]} I(\underline{\mathbf{y}}_i, \underline{\mathbf{J}} | \underline{\mathbf{x}}_i, \underline{\mathbb{T}}_i) \quad (137)$$

Proof. Notice that from the section 2.2 we have the following model for the output variable

$$\underline{\mathbf{y}}_i = \left\langle \underbrace{\begin{bmatrix} \underline{\mathbf{x}}_i \\ \underline{\mathbb{T}}_i \underline{\mathbf{x}}_i \end{bmatrix}}_{=: \underline{\mathbf{v}}_i(\underline{\mathbf{x}}_i, \underline{\mathbb{T}}_i) \in \mathbb{R}^{2d}}, \omega^{\underline{\mathbf{J}}} \right\rangle + \epsilon_i \quad | \quad \omega^{\underline{\mathbf{J}}} = \begin{bmatrix} \beta^{\underline{\mathbf{J}}} \\ \theta^{\underline{\mathbf{J}}} \end{bmatrix} \in \Theta^M \quad (138)$$

Now conditioned on $\underline{\mathbf{x}}_i = x_i$ and $\underline{\mathbb{T}}_i = T_i$ we have

$$(\underline{\mathbf{y}}_i | \underline{\mathbf{x}}_i = x_i, \underline{\mathbb{T}}_i = T_i) =: \underline{\mathbf{y}}_i^{x_i T_i} = \langle v_i(x_i, T_i), \omega^{\underline{\mathbf{J}}} \rangle + \epsilon_i \quad | \quad \forall (x_i, T_i) \in \mathcal{X} \times \{0, 1\} \quad (139)$$

Hence we notice that

$$\underline{\mathbf{y}}_i^{x_i T_i} \not\perp \underline{\mathbf{y}}_j^{x_j T_j} \quad \forall i \neq j \quad | \quad \underline{\mathbf{y}}_i^{x_i T_i} | \underline{\mathbf{J}} \perp \underline{\mathbf{y}}_j^{x_j T_j} | \underline{\mathbf{J}} \quad \forall i \neq j \quad | \quad \forall (x_i, T_i) \in \mathcal{X} \times \{0, 1\} \quad (140)$$

We can upper bound the mutual information $I(\underline{\mathbf{Y}}_{\mathbf{n}}, \underline{\mathbf{J}} | \underline{\mathbf{X}}_{\mathbf{n}}, \underline{\mathbb{T}}_{\mathbf{n}})$ as follow

$$I(\underline{\mathbf{Y}}_{\mathbf{n}}, \underline{\mathbf{J}} | \underline{\mathbf{X}}_{\mathbf{n}}, \underline{\mathbb{T}}_{\mathbf{n}}) = H(\underline{\mathbf{y}}_1, \dots, \underline{\mathbf{y}}_n | \underline{\mathbf{X}}_{\mathbf{n}}, \underline{\mathbb{T}}_{\mathbf{n}}) - H(\underline{\mathbf{y}}_1, \dots, \underline{\mathbf{y}}_n | \underline{\mathbf{J}}, \underline{\mathbf{X}}_{\mathbf{n}}, \underline{\mathbb{T}}_{\mathbf{n}}) \quad (141)$$

$$= \sum_{i \in [n]} H(\underline{\mathbf{y}}_i | \underline{\mathbf{y}}_1, \dots, \underline{\mathbf{y}}_{(i-1)}, \underline{\mathbf{x}}_i, \underline{\mathbb{T}}_i) - H(\underline{\mathbf{y}}_1, \dots, \underline{\mathbf{y}}_n | \underline{\mathbf{J}}, \underline{\mathbf{X}}_{\mathbf{n}}, \underline{\mathbb{T}}_{\mathbf{n}}) \quad | \quad \text{Chain Rule} \quad (142)$$

$$\leq \sum_{i \in [n]} H(\underline{\mathbf{y}}_i | \underline{\mathbf{x}}_i, \underline{\mathbb{T}}_i) - H(\underline{\mathbf{y}}_1, \dots, \underline{\mathbf{y}}_n | \underline{\mathbf{J}}, \underline{\mathbf{X}}_{\mathbf{n}}, \underline{\mathbb{T}}_{\mathbf{n}}) \quad | \quad \text{Conditioning reduces entropy} \quad (143)$$

$$= \sum_{i \in [n]} H(\underline{\mathbf{y}}_i | \underline{\mathbf{x}}_i, \underline{\mathbb{T}}_i) - \sum_{i \in [n]} H(\underline{\mathbf{y}}_i | \underline{\mathbf{J}}, \underline{\mathbf{x}}_i, \underline{\mathbb{T}}_i) \quad | \quad \text{Chain Rule and Ind.} \quad (144)$$

$$= \sum_{i \in [n]} I(\underline{\mathbf{y}}_i, \underline{\mathbf{J}} | \underline{\mathbf{x}}_i, \underline{\mathbb{T}}_i) \quad (145)$$

Note: the content up to this section constitute the foundation of this work. The methods used up to this point will be fixed for all the resulting upcoming.

2.5 Fano's continued with Lemma 2

Claim 2.3 An upper bound on $I(\underline{\mathbf{y}}_i, \underline{\mathbf{J}}|\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i)$ using lemma 1.2

$$I(\underline{\mathbf{y}}_i, \underline{\mathbf{J}}|\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i) \leq \frac{1}{2} \left\{ \log \frac{\text{var}(\underline{\mathbf{y}}_i^{x_i T_i})}{\sigma^2} \right\} \quad (146)$$

Proof. We recall that $\underline{\mathbf{J}} \sim \text{Uni}[M]$ and from eq. 2.1. $\underline{\mathbf{y}}_i^{x_i T_i}|\underline{\mathbf{J}} = j \sim \mathcal{N}(\langle v(x_i, T_i), \omega^j \rangle, \sigma^2)$ hence applying lemma 1.2

$$I(\underline{\mathbf{y}}_i, \underline{\mathbf{J}}|\underline{\mathbf{x}}_i = x_i, \underline{\mathbf{T}}_i = T_i) \leq \frac{1}{2} \left\{ \log \text{var}(\underline{\mathbf{y}}_i^{x_i T_i}) - \frac{1}{M} \sum_{j \in [M]} \log \text{var}(\underline{\mathbf{y}}_i^{x_i T_i}|\underline{\mathbf{J}} = j) \right\} \quad (147)$$

$$= \frac{1}{2} \left\{ \log \frac{\text{var}(\underline{\mathbf{y}}_i^{x_i T_i})}{\sigma^2} \right\} \quad (148)$$

$$(149)$$

Claim 2.4 An upper on $\mathbb{E}_{\underline{\mathbf{X}}_n, \underline{\mathbf{T}}_n} I(\underline{\mathbf{Y}}_n, \underline{\mathbf{J}}|\underline{\mathbf{X}}_n, \underline{\mathbf{T}}_n)$ to be optimized wrt $\{\beta_j\}_{j \in [M]}$

Using claim 2.3 and 2.3 we have the following bound on $\mathbb{E}_{\underline{\mathbf{X}}_n, \underline{\mathbf{T}}_n} I(\underline{\mathbf{Y}}_n, \underline{\mathbf{J}}|\underline{\mathbf{X}}_n, \underline{\mathbf{T}}_n)$

$$\mathbb{E}_{\underline{\mathbf{X}}_n, \underline{\mathbf{T}}_n} I(\underline{\mathbf{Y}}_n, \underline{\mathbf{J}}|\underline{\mathbf{X}}_n, \underline{\mathbf{T}}_n) \leq \sum_{i \in [n]} \log \frac{\mathbb{E}_{\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i} \text{var}(\underline{\mathbf{y}}_i|\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i)}{\sigma^2} \quad (150)$$

and

$$\mathbb{E}_{\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i} \text{var}(\underline{\mathbf{y}}_i(\underline{\mathbf{J}})|\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i) \leq \sigma^2 + \frac{1}{M} \sum_{j \in [M]} \underbrace{\left[(1-p) \text{Tr}(\beta^j \otimes \beta^j) + p \text{Tr}((\beta^j + \theta^j) \otimes (\beta^j + \theta^j)) \right]}_{\text{to be optimised wrt } \{\beta^j\}_{j \in [M]}} \quad (151)$$

Proof. Using claim 2.3

$$I(\underline{\mathbf{Y}}_n, \underline{\mathbf{J}}|\underline{\mathbf{X}}_n, \underline{\mathbf{T}}_n) \leq \sum_{i \in [n]} \log \frac{\text{var}(\underline{\mathbf{y}}_i|\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i)}{\sigma^2} \quad (152)$$

$$\iff \quad (153)$$

$$\mathbb{E}_{\underline{\mathbf{X}}_n, \underline{\mathbf{T}}_n} I(\underline{\mathbf{Y}}_n, \underline{\mathbf{J}}|\underline{\mathbf{X}}_n, \underline{\mathbf{T}}_n) \leq \mathbb{E}_{\underline{\mathbf{X}}_n, \underline{\mathbf{T}}_n} \sum_{i \in [n]} \log \frac{\text{var}(\underline{\mathbf{y}}_i|\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i)}{\sigma^2} \quad (154)$$

$$= \sum_{i \in [n]} \mathbb{E}_{\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i} \log \frac{\text{var}(\underline{\mathbf{y}}_i|\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i)}{\sigma^2} \quad (155)$$

$$\leq \sum_{i \in [n]} \log \frac{\mathbb{E}_{\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i} \text{var}(\underline{\mathbf{y}}_i|\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i)}{\sigma^2} \quad | \quad \text{Jensen and Concavity} \quad (156)$$

$$(157)$$

to further upper bound we have

$$\mathbb{E}_{\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i} \text{var}(\underline{\mathbf{y}}_i(\underline{\mathbf{J}})|\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i) \leq \mathbb{E}_{\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i} \mathbb{E}(\underline{\mathbf{y}}_i(\underline{\mathbf{J}})^2|\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i) \quad (158)$$

$$= \mathbb{E}_{\underline{\mathbf{J}}} \mathbb{E}_{\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i} \mathbb{E}(\underline{\mathbf{y}}_i(\underline{\mathbf{J}})^2|\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i, \underline{\mathbf{J}}) \quad (159)$$

$$= \frac{1}{M} \sum_{j \in [M]} \mathbb{E}_{\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i} \mathbb{E}(\underline{\mathbf{y}}_i(\underline{\mathbf{J}})^2|\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i, \underline{\mathbf{J}} = j) \quad (160)$$

$$= \frac{1}{M} \sum_{j \in [M]} \mathbb{E}_{\underline{\mathbf{T}}_i} \mathbb{E}_{\underline{\mathbf{x}}_i|\underline{\mathbf{T}}_i} \mathbb{E}(\underline{\mathbf{y}}_i(\underline{\mathbf{J}})^2|\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i, \underline{\mathbf{J}} = j) \quad (161)$$

$$= \frac{1}{M} \sum_{j \in [M]} \left[f_p(\underline{\mathbf{T}}_i = 0) \mathbb{E}_{\underline{\mathbf{x}}_i|\underline{\mathbf{T}}_i} \mathbb{E}(\underline{\mathbf{y}}_i(\underline{\mathbf{J}})^2|\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i = 0, \underline{\mathbf{J}} = j) \right. \quad (162)$$

$$\left. + f_p(\underline{\mathbf{T}}_i = 1) \mathbb{E}_{\underline{\mathbf{x}}_i|\underline{\mathbf{T}}_i} \mathbb{E}(\underline{\mathbf{y}}_i(\underline{\mathbf{J}})^2|\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i = 1, \underline{\mathbf{J}} = j) \right] \quad (163)$$

$$= \frac{1}{M} \sum_{j \in [M]} \left[q \mathbb{E}_{\underline{\mathbf{x}}_i} \mathbb{E}(\underline{\mathbf{y}}_i(\underline{\mathbf{J}})^2|\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i = 0, \underline{\mathbf{J}} = j) + p \mathbb{E}_{\underline{\mathbf{x}}_i} \mathbb{E}(\underline{\mathbf{y}}_i(\underline{\mathbf{J}})^2|\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i = 0, \underline{\mathbf{J}} = j) \right] \quad (164)$$

Notice that

$$(\underline{\mathbf{y}}_i^2|\underline{\mathbf{x}}_i = x_i, \underline{\mathbf{T}}_i = 0, \underline{\mathbf{J}} = j) = (\langle x_i, \beta^j \rangle + \epsilon_i)^2 \quad \forall x_i \in \mathcal{X} \quad (165)$$

$$= \text{Tr}(\beta^j \otimes \beta^j x_i \otimes x_i) + 2\epsilon_i \langle x_i, \beta^j \rangle + \epsilon_i^2 \quad \forall x_i \in \mathcal{X} \quad (166)$$

$$(\underline{\mathbf{y}}_i^2|\underline{\mathbf{x}}_i = x_i, \underline{\mathbf{T}}_i = 1, \underline{\mathbf{J}} = j) = (\langle x_i, \beta^j + \theta^j \rangle + \epsilon_i)^2 \quad \forall x_i \in \mathcal{X} \quad (167)$$

$$= \text{Tr}((\beta^j + \theta^j) \otimes (\beta^j + \theta^j) x_i \otimes x_i) + 2\epsilon_i \langle x_i, (\beta^j + \theta^j) \rangle + \epsilon_i^2 \quad \forall x_i \in \mathcal{X} \quad (168)$$

Hence summarising the last 2 succession of equation and using the fact that we modelled $\underline{\mathbf{x}}_i$ as isotropic gaussian we have

$$\mathbb{E}_{\underline{\mathbf{x}}_i} \mathbb{E}(\underline{\mathbf{y}}_i(\underline{\mathbf{J}})^2 | \underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i = 0, \underline{\mathbf{J}} = j) = \text{Tr}(\beta^j \otimes \beta^j) + \sigma^2 \quad (169)$$

$$\mathbb{E}_{\underline{\mathbf{x}}_i} \mathbb{E}(\underline{\mathbf{y}}_i(\underline{\mathbf{J}})^2 | \underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i = 1, \underline{\mathbf{J}} = j) = \text{Tr}((\beta^j + \theta^j) \otimes (\beta^j + \theta^j)) + \sigma^2 \quad (170)$$

putting things together

$$\mathbb{E}_{\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i} \text{var}(\underline{\mathbf{y}}_i(\underline{\mathbf{J}}) | \underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i) \leq \sigma^2 + \frac{1}{M} \sum_{j \in [M]} \left[(1-p) \text{Tr}(\beta^j \otimes \beta^j) + p \text{Tr}((\beta^j + \theta^j) \otimes (\beta^j + \theta^j)) \right] \quad (171)$$

😊 □

2.5.1 Optimization problem

We aim at tight bounds i.e

$$(\cdot) \geq 1 - \frac{\mathbb{E}I + (\cdot)}{(\cdot)} \quad (172)$$

$$\iff \quad (173)$$

$$\mathbb{E}I \leq \underbrace{(\cdot)}_{\text{to minimize in order to have tight bounds}} \quad (174)$$

Claim 2.5 Convex Programm

The optimization problem from claim 2.4

$$\{\hat{\beta}^j\}_{j \in [M]} = \arg \min_{\{\beta^j\}_{j \in [M]} \in \mathbb{B}_2^d} \frac{1}{M} \sum_{j \in [M]} \langle \beta^j, \beta^j \rangle + \frac{2p}{M} \sum_{j \in [M]} \langle \beta^j, \theta^j \rangle + \frac{p}{M} \sum_{j \in [M]} \langle \theta^j, \theta^j \rangle \quad (175)$$

$$(176)$$

can be rewritten as the mathematical programm $\mathcal{P}(\mathcal{M}, f(\mathcal{B}))$

$$\hat{\mathcal{B}} = \arg \min_{\mathcal{B} \in \mathcal{M}} f(\mathcal{B}) = \arg \min_{\mathcal{B} \in \mathcal{M}} \frac{1}{M} \text{Tr}\{\mathcal{B}^T \mathcal{B}\} + \frac{2p}{M} \text{Tr}\{\mathcal{B}^T \mathcal{O}\} + pk\theta_{min}^2 \quad (177)$$

$$\mathcal{M} := \left\{ \mathcal{B} \in \mathbb{R}^{d \times M}; \|\text{col}(\mathcal{B})_i\|_2 \leq 1 \quad \forall i \in [M] \right\} \quad (178)$$

where

$$\mathcal{B} := [\beta_1, \dots, \beta_M] \in \mathbb{R}^{d \times M} \quad (179)$$

$$\mathcal{O} := [\theta_1, \dots, \theta_M] \in \mathbb{R}^{d \times M} \quad (180)$$

Moreover

$$\mathcal{P}(\mathcal{M}, f(\mathcal{B})) \quad \text{is a convex programm} \quad (181)$$

Proof.

$$\{\hat{\beta}^j\}_{j \in [M]} = \arg \min_{\{\beta^j\}_{j \in [M]} \in \mathbb{B}_2^d} \frac{1}{M} \sum_{j \in [M]} \left[(1-p) \text{Tr}(\beta^j \otimes \beta^j) + p \text{Tr}((\beta^j + \theta^j) \otimes (\beta^j + \theta^j)) \right] \quad (182)$$

$$= \arg \min_{\{\beta^j\}_{j \in [M]} \in \mathbb{B}_2^d} \frac{1}{M} \sum_{j \in [M]} \left[(1-p) \langle \beta^j, \beta^j \rangle + p \langle \beta^j + \theta^j, \beta^j + \theta^j \rangle \right] \quad (183)$$

$$= \arg \min_{\{\beta^j\}_{j \in [M]} \in \mathbb{B}_2^d} \frac{1}{M} \sum_{j \in [M]} \langle \beta^j, \beta^j \rangle + \frac{2p}{M} \sum_{j \in [M]} \langle \beta^j, \theta^j \rangle + \frac{p}{M} \sum_{j \in [M]} \langle \theta^j, \theta^j \rangle \quad (184)$$

$$(185)$$

We can rewrite the above mathematical program by defining the following matrix

$$\mathcal{B} := [\beta_1, \dots, \beta_M] \in \mathbb{R}^{d \times M} \quad (186)$$

$$\mathcal{O} := [\theta_1, \dots, \theta_M] \in \mathbb{R}^{d \times M} \quad (187)$$

which yield

$$\hat{\mathcal{B}} = \arg \min_{\mathcal{B} \in \mathcal{M}} \frac{1}{M} \text{Tr}\{\mathcal{B}^T \mathcal{B}\} + \frac{2p}{M} \text{Tr}\{\mathcal{B}^T \mathcal{O}\} + pk\theta_{min}^2 \quad (188)$$

$$\mathcal{M} := \left\{ \mathcal{B} \in \mathbb{R}^{d \times M}; \|\text{col}(\mathcal{B})_i\|_2 \leq 1 \quad \forall i \in [M] \right\} \quad (189)$$

we notice the following assuming $A, B \in \mathcal{M}$ and $\lambda \in [0, 1]$

$$\|\lambda \text{col}(A)_i + (1-\lambda) \text{col}(B)_i\|_2^2 = \langle \lambda \text{col}(A)_i + (1-\lambda) \text{col}(B)_i, \lambda \text{col}(A)_i + (1-\lambda) \text{col}(B)_i \rangle \quad (190)$$

$$= \lambda^2 \|\text{col}(A)_i\|_2^2 + 2\lambda(1-\lambda) \|\text{col}(A)_i\|_2 \|\text{col}(B)_i\|_2 + (1-\lambda)^2 \|\text{col}(B)_i\|_2^2 \quad (191)$$

$$\leq \lambda^2 + 2\lambda(1-\lambda) + (1-\lambda)^2 \quad | \text{C.S and } A, B \in \mathcal{M} \quad (192)$$

$$= 1 \quad \forall i \in [M] \quad (193)$$

Hence the constraint set \mathcal{M} is a convex set.

We can also rewrite the optimization problem by vectorizing the matrices \mathcal{B}, \mathcal{O} as follow

$$\text{vec}(\mathcal{B}) =: B = \begin{bmatrix} \beta^1 \\ \vdots \\ \beta^M \end{bmatrix} \in \mathbb{R}^{dM} \quad (194)$$

$$\text{vec}(\mathcal{O}) =: O = \begin{bmatrix} \theta^1 \\ \vdots \\ \theta^M \end{bmatrix} \in \mathbb{R}^{dM} \quad (195)$$

which yield

$$\hat{B} = \arg \min_{B \in \mathcal{M}} f(B) = \arg \min_{B \in \mathcal{M}} \frac{1}{M} \langle B, B \rangle + \frac{2p}{M} \langle B, O \rangle + pk\theta_{min}^2 \quad (196)$$

We notice that $\nabla^2 f(B) \succeq 0 \quad \forall B \in \mathbb{R}^{dM} \iff f$ is convex over \mathbb{R}^{dM}

😊 □

Claim 2.6 A Solution to Optimization Problem of claim 2.5 when Global Optima $\in \mathcal{M}$

Assuming

$$p\sqrt{k}\theta_{min} \leq 1 \quad (197)$$

the solution of $\mathcal{P}(\mathcal{M}, f(\mathcal{B}))$ yield

$$\hat{\beta}^j = -p\theta^j \quad \forall j \in [M] \quad (198)$$

Proof. If the global minimum belongs $\hat{B} = \arg \min f(B)$ of the to the constraint region \mathcal{M} it yield a solution to the convex constraint program. By first order characterization of convexity we have

$$\nabla_{|B=\hat{B}} f(B) = 0 \quad (199)$$

$$\Leftrightarrow \quad (200)$$

$$B = -pO \quad (201)$$

$$\Leftrightarrow \quad (202)$$

$$\hat{\beta}^j = -p\theta^j \quad \forall j \in [M] \quad (203)$$

furthermore the optimal belongs to the constraint set in the following setting:

$$\|\hat{\beta}^j\|_2 \leq 1 \Leftrightarrow p\|\theta^j\|_2 \leq 1 \quad (204)$$

$$\Leftrightarrow p\sqrt{k}\theta_{min} \leq 1 \quad (205)$$

😊 □

under the last constraint, we can now choose the discretization of \mathbb{B}_2^d to be \hat{B} and plug it into fano's bound yielding the following lemma

Lemma 2.2 Minimax Lower Bound Lower Bound for our Sparse Causal Estimator

Using lemma 2.1 and claims 2.1,2.2,2.3,2.4,2.5 we have the following minimax lower bound

$$\inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\mathbb{S}(k,d))} \sup_{(\beta, \theta) \in \Theta} \mathbb{P}_{(\beta, \theta)} \{ \hat{\mathbf{S}}(\mathbb{D}_{\mathbf{n}}) \neq \text{supp}(\theta) \} \geq 1 - \frac{n \log \left[1 + \frac{k\theta_{min}^2 p(1-p)}{\sigma^2} \right] + \log 2}{\log \binom{d}{k}} \quad (206)$$

Proof.

$$\mathbb{E}_{\mathbf{x}_i, \mathbf{T}_i} \text{var}(\mathbf{y}_i(\mathbf{J}) | \mathbf{x}_i, \mathbf{T}_i) \leq \sigma^2 + \frac{1}{M} \sum_{j \in [M]} \left[(p-1) \text{Tr}(\beta^j \otimes \beta^j) + p \text{Tr}((\beta^j + \theta^j) \otimes (\beta^j + \theta^j)) \right] \quad (207)$$

$$= \sigma^2 + \frac{1}{M} \sum_{j \in [M]} \left[p^2(1-p) \text{Tr}(\theta^j \otimes \theta^j) + p(1-p)^2 \text{Tr}(\theta^j \otimes \theta^j) \right] \quad (208)$$

$$= \sigma^2 + k\theta_{min}^2 (p^2(1-p) + p(1-p)^2) \quad (209)$$

$$= \sigma^2 + k\theta_{min}^2 p(1-p) \quad (210)$$

$$\Leftrightarrow \quad (211)$$

$$\mathbb{E}_{\mathbf{x}_n, \mathbf{T}_n} I(\mathbf{Y}_{\mathbf{n}}, \mathbf{J} | \mathbf{x}_n, \mathbf{T}_n) \leq \sum_{i \in [n]} \log \left[1 + \frac{k\theta_{min}^2 p(1-p)}{\sigma^2} \right] \quad (212)$$

$$= n \log \left[1 + \frac{k\theta_{min}^2 p(1-p)}{\sigma^2} \right] \quad (213)$$

$$\Leftrightarrow \quad (214)$$

$$\inf_{\hat{S}: \mathcal{D} \rightarrow \text{supp}(\mathbb{S}(k,d))} \sup_{(\beta, \theta) \in \Theta} \mathbb{P}_{(\beta, \theta)} \{ \hat{\mathbf{S}}(\mathbb{D}_{\mathbf{n}}) \neq \text{supp}(\theta) \} \geq 1 - \frac{n \log \left[1 + \frac{k\theta_{min}^2 p(1-p)}{\sigma^2} \right] + \log 2}{\log M} \quad (215)$$

😊 □

Result 2.1 A lower bound on the data needed to be better than randomness

Using lemma 2.2, in order for any procedure or algorithm to achieve probability of error in the support recovery process of $\vartheta^{\mathbf{T}=0} - \vartheta^{\mathbf{T}=1}$ below $1/2$ we need at least

$$n > \left\lceil \frac{\log \binom{d}{k}}{2} + \log 2 \right\rceil \frac{1}{\log \left[1 + \frac{k\theta_{\min}^2 p(1-p)}{\sigma^2} \right]} \quad (216)$$

3 Fano: Continued with upper bounding using convexity argument

3.1 Some Tools

Claim 3.1 Controlling the KL Divergence with Convexity

Given the distributions $\mathbb{Q}, \{\mathbb{P}_j\}_{j \in [M]}$ and $\sum_{j \in [M]} \lambda_j = 1 \quad \lambda_j \in \mathbb{R}, \forall j \in [M]$ we have

$$D(\mathbb{Q} \parallel \sum_{j \in [M]} \lambda_j \mathbb{P}_j) \leq \sum_{j \in [M]} \lambda_j D(\mathbb{Q} \parallel \mathbb{P}_j)$$

Proof. Starting with the definition of the KL divergence and defining $\mathbb{P} := \sum_{j \in [M]} \lambda_j \mathbb{P}_j$ we have

$$D(\mathbb{Q} \parallel \mathbb{P}) = \mathbb{E}_{\mathbb{Q}} \log \frac{d\mathbb{Q}}{d\mathbb{P}} \quad (217)$$

$$= -\mathbb{E}_{\mathbb{Q}} \log \frac{d\mathbb{P}}{d\mathbb{Q}} \quad (218)$$

In the quantity above we would like to pour out the sum i.e $\frac{d(\sum_j \lambda_j \mathbb{P}_j)}{d\mathbb{Q}}$. By defining $\mathbb{K} := \frac{1}{2}(\mathbb{P} + \mathbb{Q})$ we have $\mathbb{Q} \ll \mathbb{P} \ll \mathbb{K}$ which implies by radon Theorem and chain rule that the following is true

$$\frac{d\mathbb{P}}{d\mathbb{Q}} = \frac{d\mathbb{P}}{d\mathbb{K}} \frac{d\mathbb{K}}{d\mathbb{Q}} \quad | \quad \text{Chain Rule} \quad (219)$$

$$= \sum_{j \in [M]} \lambda_j \frac{d\mathbb{P}_j}{d\mathbb{K}} \frac{d\mathbb{K}}{d\mathbb{Q}} \quad | \quad \mathbb{P} \ll \mathbb{K} \implies \mathbb{P}_j \ll \mathbb{K} \forall j \in [M] \quad + \text{Existence of Radon derivative} \quad (220)$$

Using Jensen inequality we have

$$\mathbb{E}_{\mathbb{Q}} - \log \frac{d\mathbb{P}}{d\mathbb{Q}} = \mathbb{E}_{\mathbb{Q}} - \log \sum_{j \in [M]} \lambda_j \frac{d\mathbb{P}_j}{d\mathbb{Q}} \quad (221)$$

$$\leq \sum_{j \in [M]} \lambda_j \mathbb{E}_{\mathbb{Q}} - \log \frac{d\mathbb{P}_j}{d\mathbb{Q}} \quad (222)$$

$$= \sum_{j \in [M]} \lambda_j D(\mathbb{Q} \parallel \mathbb{P}_j) \quad (223)$$

😊 □

3.2 New upper bound for mutual information $I(\underline{\mathbf{y}}_i, \underline{\mathbf{J}} | \underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i)$

Lemma 3.1 Upper bound for $I(\underline{\mathbf{y}}_i, \underline{\mathbf{J}} | \underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i)$ using convexity argument

Given the setup described in 2.2 and the claim 3.1 the following upper bound on the mutual information $I(\underline{\mathbf{y}}_i, \underline{\mathbf{J}} | \underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i)$ holds true

$$I(\underline{\mathbf{y}}_i, \underline{\mathbf{J}} | \underline{\mathbf{x}}_i = x_i, \underline{\mathbf{T}}_i = T_i) \leq \frac{1}{2\sigma^2 M^2} \sum_{\substack{l, j \in [M] \\ l \neq j}} Tr \left\{ \left[(\beta^l - \beta^j)^{\otimes 2} + T_i [(\theta^l - \theta^j) \otimes (\beta^l - \beta^j) + (\beta^l - \beta^j) \otimes (\theta^l - \theta^j)] + T_i^2 (\theta^l - \theta^j)^{\otimes 2} \right] x_i^{\otimes 2} \right\} \quad (224)$$

Proof. We can write the mutual information in term of Kl divergence as follow

$$I(\underline{\mathbf{y}}_i, \underline{\mathbf{J}} | \underline{\mathbf{x}}_i = x_i, \underline{\mathbf{T}}_i = T_i) = \mathbb{E}_{\underline{\mathbf{Q}}_{\underline{\mathbf{J}}}} D(\mathbb{Q}_{\underline{\mathbf{y}}_i | \underline{\mathbf{J}}, \underline{\mathbf{x}}_i = x_i, \underline{\mathbf{T}}_i = T_i} \parallel \mathbb{Q}_{\underline{\mathbf{y}} | \underline{\mathbf{x}}_i = x_i, \underline{\mathbf{T}}_i = T_i}) \quad (225)$$

$$= \frac{1}{M} \sum_{j \in [M]} D(\mathbb{Q}_{\underline{\mathbf{y}}_i | \underline{\mathbf{J}} = j, \underline{\mathbf{x}}_i = x_i, \underline{\mathbf{T}}_i = T_i} \parallel \mathbb{Q}_{\underline{\mathbf{y}} | \underline{\mathbf{x}}_i = x_i, \underline{\mathbf{T}}_i = T_i}) \quad (226)$$

where

$$\mathbb{Q}_{\underline{\mathbf{y}}_i | \underline{\mathbf{J}} = j, \underline{\mathbf{x}}_i = x_i, \underline{\mathbf{T}}_i = T_i} = \mathbb{Q}_{\mathcal{N}(\langle x_i, \beta^j + T_i \theta^j \rangle, \sigma^2)} \ll \nu \quad (227)$$

$$\mathbb{Q}_{\underline{\mathbf{y}} | \underline{\mathbf{x}}_i = x_i, \underline{\mathbf{T}}_i = T_i} = \frac{1}{M} \sum_{j \in [M]} \mathbb{Q}_{\mathcal{N}(\langle x_i, \beta^j + T_i \theta^j \rangle, \sigma^2)} \ll \nu \quad (228)$$

$$(229)$$

where ν represents the lebesgue measure. Hence by absolute continuity we can replace the gaussian distribution by their respective Radon derivative i.e their probability density functions

$$\frac{d\mathbb{Q}_{\mathcal{N}(\langle x_i, \beta^l + T_i \theta^l \rangle, \sigma^2)}}{d\nu} =: g_l \quad (230)$$

$$\frac{1}{M} \sum_{j \in [M]} \frac{d\mathbb{Q}_{\mathcal{N}(\langle x_i, \beta^j + T_i \theta^j \rangle, \sigma^2)}}{d\nu} =: \frac{1}{M} \sum_{j \in [M]} g_j =: g \quad (231)$$

from claim 3.1 we have

$$D(g_l || g) \leq \frac{1}{M} \sum_{j \in [M]} D(g_l || g_j) \quad (232)$$

Now we use the closed form for KL divergence of gaussian distributions

$$D(g_l || g_j) = \frac{1}{2} \left[\log \frac{|\Sigma_{g_j}|}{|\Sigma_{g_l}|} + \text{Tr}[\Sigma_{g_j}^{-1} \Sigma_{g_l}] - 1 + (\mu_{g_l} - \mu_{g_j})^T \Sigma_{g_j}^{-1} (\mu_{g_l} - \mu_{g_j}) \right] \quad (233)$$

$$= \frac{1}{2\sigma^2} (\mu_{g_l} - \mu_{g_j})^2 \quad (234)$$

where

$$\mu_{g_l} = \langle x_i, \beta^l + T_i \theta^l \rangle \quad (235)$$

$$\mu_{g_j} = \langle x_i, \beta^j + T_i \theta^j \rangle \quad (236)$$

$$(237)$$

implying

$$(\mu_{g_l} - \mu_{g_j}) = x_i^T \underbrace{(\beta^l - \beta^j + T_i(\theta^l - \theta^j))}_{=:c} \quad (238)$$

$$\Leftrightarrow \quad (239)$$

$$(\mu_{g_l} - \mu_{g_j})^2 = x_i^T c c^T x_i = \text{Tr}(x_i^T c c^T x_i) = \text{Tr}(c c^T x_i x_i^T) \quad (240)$$

$$= \text{Tr} \left\{ \left((\beta^l - \beta^j)^{\otimes 2} + T(\theta^l - \theta^j) \otimes (\beta^l - \beta^j) + T(\beta^l - \beta^j) \otimes (\theta^l - \theta^j) + T^2(\theta^l - \theta^j)^{\otimes 2} \right) x_i^{\otimes 2} \right\} \quad (241)$$

Putting things together we have

$$I(\underline{\mathbf{y}}_i, \underline{\mathbf{J}} | \underline{\mathbf{x}}_i = x_i, \underline{\mathbf{T}}_i = T_i) = \frac{1}{M} \sum_{j \in [M]} D(\mathbb{Q}_{\underline{\mathbf{y}}_i | \underline{\mathbf{J}}=j, \underline{\mathbf{x}}_i=x_i, \underline{\mathbf{T}}_i=T_i} || \mathbb{Q}_{\underline{\mathbf{y}} | \underline{\mathbf{x}}_i=x_i, \underline{\mathbf{T}}_i=T_i}) \quad (242)$$

$$\leq \frac{1}{M^2} \sum_{\substack{l, j \in [M] \\ l \neq j}} D(g_l || g_j) \quad (243)$$

$$= \frac{1}{2\sigma^2 M^2} \sum_{\substack{l, j \in [M] \\ l \neq j}} \text{Tr} \left\{ \left((\beta^l - \beta^j)^{\otimes 2} + T_i(\theta^l - \theta^j) \otimes (\beta^l - \beta^j) + T_i(\beta^l - \beta^j) \otimes (\theta^l - \theta^j) \right. \right. \quad (244)$$

$$\left. + T_i^2(\theta^l - \theta^j)^{\otimes 2} \right) x_i^{\otimes 2} \Big\} \quad (245)$$

😊 □

Result 3.1 Semi - Result: New Bound for $\mathbb{E}_{\underline{\mathbf{x}}_n, \underline{\mathbf{T}}_n} I(\underline{\mathbf{Y}}_n, \underline{\mathbf{J}} | \underline{\mathbf{X}}_n, \underline{\mathbf{T}}_n)$ not Optimized

$$\mathbb{E}_{\underline{\mathbf{x}}_n, \underline{\mathbf{T}}_n} I(\underline{\mathbf{Y}}_n, \underline{\mathbf{J}} | \underline{\mathbf{X}}_n, \underline{\mathbf{T}}_n) \leq \quad (246)$$

$$\frac{n}{2\sigma^2 M^2} \sum_{\substack{l, j \in [M] \\ l \neq j}} \text{Tr} \left\{ \underbrace{(1-p)[(\beta^l - \beta^j)^{\otimes 2}] + p[(\theta^l - \theta^j) \otimes (\beta^l - \beta^j) + (\beta^l - \beta^j) \otimes (\theta^l - \theta^j) + (\theta^l - \theta^j)^{\otimes 2}]}_{\text{To optimize wrt } \{\beta^j\}_{j \in [M]}} \right\} \quad (247)$$

Proof. Recalling BLABLA and the lemma 3.1

$$\mathbb{E}_{\underline{\mathbf{x}}_n, \underline{\mathbf{T}}_n} I(\underline{\mathbf{Y}}_n, \underline{\mathbf{J}} | \underline{\mathbf{X}}_n, \underline{\mathbf{T}}_n) \leq \sum_{i \in [n]} \mathbb{E}_{\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i} I(\underline{\mathbf{y}}_i, \underline{\mathbf{J}} | \underline{\mathbf{x}}_i = x_i, \underline{\mathbf{T}}_i = T_i) \quad (248)$$

$$\leq \frac{1}{2\sigma^2 M^2} \sum_{i \in [n]} \sum_{\substack{l, j \in [M] \\ l \neq j}} \mathbb{E}_{\underline{\mathbf{x}}_i, \underline{\mathbf{T}}_i} \text{Tr} \{ c c^T \underline{\mathbf{x}}_i \underline{\mathbf{x}}_i^T \} \quad (249)$$

where

$$c c^T = \left[(\beta^l - \beta^j)^{\otimes 2} + \underline{\mathbf{T}}_i [(\theta^l - \theta^j) \otimes (\beta^l - \beta^j) + (\beta^l - \beta^j) \otimes (\theta^l - \theta^j)] + \underline{\mathbf{T}}_i^2 (\theta^l - \theta^j)^{\otimes 2} \right] \quad (250)$$

Decomposing the expectation yield

$$\mathbb{E}_{\mathbf{X}_i, \mathbb{T}_i} Tr\{cc^T \mathbf{X}_i \mathbf{X}_i^T\} = \mathbb{E}_{\mathbb{T}_i} \mathbb{E}_{\mathbf{X}_i | \mathbb{T}_i} Tr\{cc^T \mathbf{X}_i \mathbf{X}_i^T\} \quad (251)$$

$$= \mathbb{E}_{\mathbb{T}_i} Tr\{cc^T \mathbb{E}_{\mathbf{X}_i} \mathbf{X}_i^{\otimes 2}\} \quad (252)$$

$$= \mathbb{E}_{\mathbb{T}_i} Tr\{cc^T\} \quad (253)$$

$$= Tr\left\{(1-p)[(\beta^l - \beta^j)^{\otimes 2}] + p[(\theta^l - \theta^j) \otimes (\beta^l - \beta^j) + (\beta^l - \beta^j) \otimes (\theta^l - \theta^j) + (\theta^l - \theta^j)^{\otimes 2}]\right\} \quad (254)$$

$$\iff \quad (255)$$

$$\mathbb{E}_{\mathbf{X}_n, \mathbb{T}_n} I(\mathbf{Y}_n, \mathbf{J} | \mathbf{X}_n, \mathbb{T}_n) \leq \frac{1}{2\sigma^2 M^2} \sum_{i \in [n]} \sum_{\substack{l, j \in [M] \\ l \neq j}} Tr\left\{(1-p)[(\beta^l - \beta^j)^{\otimes 2}] + \right. \quad (256)$$

$$\left. p[(\theta^l - \theta^j) \otimes (\beta^l - \beta^j) + (\beta^l - \beta^j) \otimes (\theta^l - \theta^j) + (\theta^l - \theta^j)^{\otimes 2}]\right\} \quad (257)$$

$$= \frac{n}{2\sigma^2 M^2} \sum_{\substack{l, j \in [M] \\ l \neq j}} Tr\left\{(1-p)[(\beta^l - \beta^j)^{\otimes 2}] + p[(\theta^l - \theta^j) \otimes (\beta^l - \beta^j) + (\beta^l - \beta^j) \otimes (\theta^l - \theta^j) + (\theta^l - \theta^j)^{\otimes 2}]\right\} \quad (258)$$

To optimize wrt $\{\beta^j\}_{j \in [M]}$

😊 □

3.3 Interpretation

Result 3.2 Simple Case: Growth in Term of $\theta_{min}, k, d, n, \sigma^2$

Given the Result 3.1 with $p = 1$ and $\beta_j = 0 \quad \forall j \in [M]$ which corresponds to the setting in chapter 1 we obtain the following order

$$\mathbb{E}_{\mathbf{X}_n, \mathbb{T}_n} I(\mathbf{Y}_n, \mathbf{J} | \mathbf{X}_n, \mathbb{T}_n) \in \mathcal{O}\left(\frac{nk\theta_{min}^2}{\sigma^2} \left(1 - \frac{1}{\binom{d}{k}}\right)\right) \quad (259)$$

Proof. Using result 3.1 and $p = 1$ and $\beta_j = 0 \quad \forall j \in [M]$ and recalling $M =: \binom{d}{k}$ we have

$$\mathbb{E}_{\mathbf{X}_n, \mathbb{T}_n} I(\mathbf{Y}_n, \mathbf{J} | \mathbf{X}_n, \mathbb{T}_n) \leq \frac{n}{2\sigma^2 M^2} \sum_{\substack{l, j \in [M] \\ l \neq j}} Tr\{(\theta^l - \theta^j)^{\otimes 2}\} \quad (260)$$

$$= \frac{nk\theta_{min}^2}{\sigma^2} \frac{M(M-1)}{M^2} \quad (261)$$

😊 □

References

- [1] Christophe Giraud. *Introduction to High-Dimensional Statistics*. 2nd. Chapman and Hall/CRC, 2021. DOI: 10.1201/9781003158745.
- [2] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. DOI: 10.1017/9781108627771.