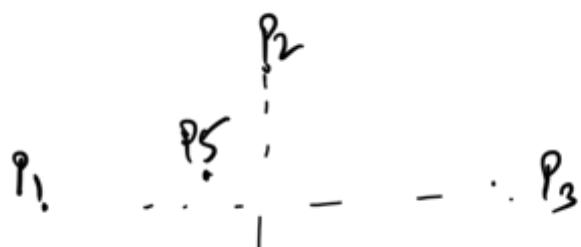


COH774
Major Solutions

Q.1. First we will show that $Vc\text{-Dim}(\mathcal{H}) \geq 5$

Now $\mathcal{H} = \mathcal{H}^+ \cup \mathcal{H}^-$
 consider the following set of 5 points
 (in 2-D)



!
P4

There are 32 possible different labelings & we need to show that all of those are achievable using $h \in \mathcal{H}$.

① Clearly all +ve & all -ve can be achieved :- we can draw a rectangle $\mathcal{H} \in \mathcal{H}^+$ enclosing all points. Similarly, all -ve is achievable since we can draw a rectangle $\mathcal{H} \in \mathcal{H}^-$ enclosing all points.

② Consider where 4 of the points are +ve & 1 -ve.

↳ ③ If the point P_5 is -ve then we can draw a rectangle

2 possibilities

$\leftarrow P^+$

Let's consider the case when 4 points are $-ve$.
 If any other point is $-ve$, we can change them by symmetry. We can easily draw a rectangle like shown to the left for each such combination.

Possibilities

Consider the case when 3 points are $+ve$ & 1 is $-ve$.

↳ 10 such possibilities.
 First let us consider when P_5 is one of the points.
 \Rightarrow 2 of P_1, P_2, P_3, P_4 are $+ve$ & 2 are $-ve$.

See figure to left. We can find a H_2T^+ to achieve this labeling. Similarly, see the labeling to left. \Rightarrow $3H_2T^+$ to achieve the labeling of 1st of 6.

Ans

P_1
 +ve
 L --- -

By similar argument,
 such possibilities can be
 achieved [4 like fig (a) & 2
 like fig (b)]

Consider:- 3+ve & 2-ve where
 P_5 is one of the points

P_2

⇒ 4 such possibilities

P_1 | $\begin{matrix} -ve \\ \text{is} \\ | \end{matrix}$ | $\begin{matrix} P_3 \\ -ve \end{matrix}$
 | | | |
 P_4

⇒ we can easily draw a
 rectangular hypothesis
 set enclosing P_5 &
 the other two points
 like shown on left.

Similar argument holds for
 any chosen negative point
 among P_1, P_2, P_3, P_4 .

⇒ All 16 possible labelings
 when 3+ve & 2-ve can
 be achieved by PCH

Finally:-

case when 3-ve & 2+ve
 is identical to 3+ve & 2-ve
 case with roles of x^+ & x^-
 reversed

⇒ All 32 labelings can
 be done by PCH

be achievable by \mathcal{H}^n

→ 3 set of points in \mathcal{D}

$P_1, P_2, P_3, P_4 \& P_5$ which can

be shattered by \mathcal{H} .

→ $VC-D(\mathcal{H}) \geq 5$.

→ A

Next, we will show / argue that

$$VC-D(\mathcal{H}) \leq 6$$

We will show that $\mathcal{H}\{P_1 \dots P_6\}$ which

$\subseteq \mathcal{E}R^2$

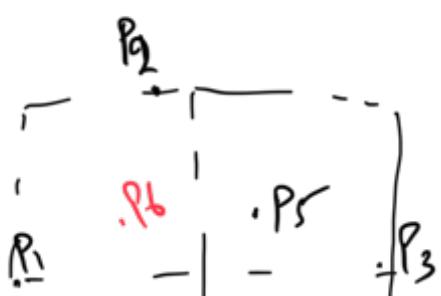
can be shattered by \mathcal{H} .

Argument :-

Consider some set of 6 points. We can safely assume that points are distinct. If not, :- clearly we can assign opposite labels to 2 points which are not distinct & the points can not be shattered.

We will explain our argument with the help of following example

↳ choose P_1, P_3, P_4, P_5 such that they correspond min-max values of x-coordinate & y-coordinate respectively. If the same $\rightarrow P_2$ the max/min



$$\left. \begin{array}{l} P_1.x \leq P_i.x \quad \forall i \\ P_3.x \geq P_i.x \quad \forall i \\ P_4.y \leq P_i.y \quad \forall i \quad [1, 2, 3] \\ P_2.y \geq P_i.y \quad \forall i \quad [1, 2, 3] \end{array} \right\} \text{By construction:-}$$

P₁, P₂, P₃, P₄ are distinct.

P₅ & P₆ lie inside the rectangle enclosing P₁, P₂, P₃, P₄.

Next we will construct a labeling to as follows - our labeling with form 3 +ve points, 3 -ve points.

Wlog:- Assume.

$$\left. \begin{array}{l} P_5.x \geq P_i.x \\ P_5.y \geq P_i.y \end{array} \right\} \text{case 1: if } P_5.x \geq P_i.x \quad \begin{array}{l} P_5.y \leq P_i.y \\ \text{then in the following step} \\ \text{we will choose point with min } x \text{ value} \\ (\text{R}) \& \text{min } y \text{ value} \\ \text{i.e. } P_5, P_1, P_2, \text{ +ve} \\ P_6, P_3, P_4, \text{ -ve} \end{array}$$

then Assign:-

P₅, P₁, P₄ : +ve

P₆, P₂, P₃ : -ve

Clearly:-

Rectangle enclosing P₁, P₄, P₅ also contains P₆. :- Not part of R₁ but part of R₂ which achieves derived labeling.

Similarly, rectangle enclosing L₁, D₁

$P_6, P_2, \text{ and } P_3$ also covers 15.
 \rightarrow $\nexists h \in \mathcal{H}$ which
 achieves $\sum_{i=1}^6 x_i$ which
 covered labeling

$\Rightarrow \nexists h \in \mathcal{H}$ which can
 achieve labeling constraint
 in the above manner.

∴ Since, $P_1, P_2, P_3, P_4, P_5, P_6$ were
 chosen generically, \nexists set of six
 points which can be shattered by
 $\mathcal{H} \Rightarrow \text{VC-Dim } (\mathcal{H}) \leq 6$ - \textcircled{A}

Combining \textcircled{A} & \textcircled{B} we get

$$\boxed{\text{VC-Dim } (\mathcal{H}) = 5}$$

Q.2: This question is similar to Question 1
 in Minor examination.

(Q) Logistic Regression loss:-

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m [y^{(i)} - h_{\theta}(x^{(i)})]^2$$

$$h_{\theta}(x^{(i)}) = \frac{1}{1 + e^{-\theta^T \phi(x^{(i)})}}$$

sigmoid
function
applied to
 $\theta^T \phi(x^{(i)})$

(b) First let us look at the "gradient"

$$\nabla_{\theta} L(\theta) = \frac{1}{2m} \sum_{l=1}^m [y^{(l)} - h_{\theta}(\phi(x^{(l)})]^2$$

$$= \frac{1}{2m} \sum_{l=1}^m (2) [y^{(l)} - h_{\theta}(\phi(x^{(l)})] \nabla_{\theta} h_{\theta}(\phi(x^{(l)})$$

Note: $\frac{\partial}{\partial \theta}$

$$\nabla_{\theta} g(\theta) = -\frac{1}{m} \sum_{l=1}^m [y^{(l)} - h_{\theta}(\phi(x^{(l)})] g'(\theta \phi(x^{(l)}))$$

$$= g(\theta) (1 - g(\theta))$$

g: Sigmoid where $g(\theta \phi(x^{(l)})) = \frac{1}{1 + e^{-\theta \phi(x^{(l)})}}$

Sigmoid function

$$\nabla_{\theta} \theta \phi(x^{(l)}) = \phi(x^{(l)})$$

$$\Rightarrow \nabla_{\theta} L(\theta) = \frac{-1}{m} \sum_{l=1}^m \frac{[y^{(l)} - h_{\theta}(\phi(x^{(l)})]}{[\phi(x^{(l)})] (1 - \phi(x^{(l)}))}$$

$$\text{Now, } h_{\theta}(\phi(x^{(l)})) = \frac{1}{1 + e^{-\theta \cdot \phi(x^{(l)})}}$$

Let us assume that θ can be represented as coefficients $\alpha_1, \dots, \alpha_m$ of data points $x^{(1)}, x^{(2)}, \dots, x^{(m)}$

$$\text{then } \theta = \sum_{i=1}^m \alpha_i \phi(x^{(i)}) \quad \left\{ \begin{array}{l} \text{And we will} \\ \text{show that this} \\ \text{assumption is} \end{array} \right.$$

At iteration t of learning

satisfied during the entire learning process] → see below.

$$\theta^{(t)} = \sum_{i=1}^m d_i^{(t)} \phi(x^{(i)})$$

Clearly At $f=0$, $\alpha_1 - \alpha_m = 0$

$$\text{Since } \theta^{(0)} = \vec{0}$$

Next, we will show that this holds wrong an induction argument.

Learning update,-

$$t=0 \\ \theta(t) \leftarrow \text{init}();$$

do {

$$\theta(t+1) \leftarrow \theta(t) - \eta \cdot \nabla_{\theta} L(\theta)|_{\theta(t)}$$

$$t \leftarrow t+1;$$

} (until ! converged)

$$\begin{aligned} \text{if } (\theta) \\ = & \frac{1}{m} \sum_{i=1}^m \left[y_i - \text{hol}(\phi(x^{(i)})) \right. \\ & \cdot \text{hol}(\phi(x^{(i)})) \left. \right] \end{aligned}$$

Assume our assumption holds at iteration t : Then, at $t+1$:
(it clearly holds at $t=0$)

$$\theta^{(t+1)} = \theta^{(t)} + \eta \cdot \sum_{i=1}^m \left[y^{(i)} - h_{\theta}(x^{(i)}) \right] h_{\theta}(x^{(i)}) \begin{cases} 1 - h_{\theta}(x^{(i)}) \\ h_{\theta}(x^{(i)}) \end{cases}$$

By inductive assumption:-

$$\theta^{(t)} = \sum_{i=1}^m \alpha_i^{(t)} \phi(x^{(i)})$$

using a different variable

$$\theta^{(t)T} \cdot \phi(x^{(i)}) = \left[\sum_{e=1}^m \alpha_e^{(t)} \phi(x^{(e)}) \right]^T \cdot \phi(x^{(i)})$$

$$= \sum_{e=1}^m \alpha_e^{(t)} [\phi(x^{(e)})^T \phi(x^{(i)})]$$

can be computed using only dot products

$$\left\{ \begin{aligned} &= \sum_{e=1}^m \alpha_e^{(t)} K(x^{(e)}, x^{(i)}) \\ &\equiv \beta_i^{(t)} \quad \text{[Another term for this expansion]} \end{aligned} \right.$$

$$\Rightarrow \theta^{(t+1)} \leftarrow \theta^{(t)} + \eta \cdot \frac{1}{m} \sum_{i=1}^m \left[y^{(i)} - \frac{1}{1+e^{-\beta_i^{(t)}}} \right] \begin{bmatrix} 1 \\ \frac{1}{1+e^{-\beta_i^{(t)}}} \end{bmatrix} \cdot \phi(x^{(i)})$$

$$\Rightarrow \theta^{(t+1)} \leftarrow \theta^{(t)} + \eta \cdot \frac{1}{m} \sum_{i=1}^m \gamma_i^{(t)} \phi(x^{(i)})$$

using inductive assumption

$$\phi(t) = \sum_{i=1}^m d_i^{(t)} \phi(x^{(i)})$$

$$\Rightarrow \phi(t+1) \leftarrow \sum_{i=1}^m (d_i^{(t)} + \eta \cdot y_i t) \phi(x^{(i)})$$

Defmc it this way.

$$= \sum_{i=1}^m d_i^{(t+1)} \phi(x^{(i)})$$

Hence, $\phi(t+1)$ can be expressed as
linear combination

$$\sum_{i=1}^m d_i^{(t+1)} \phi(x^{(i)})$$

where $d_i^{(t+1)} = d_i^{(t)} + \frac{\eta}{m} y_i t$

where $y_i t$ is as defined
earlier & can be computed easily
in terms of dot products

~~⇒ Entire learning algorithm~~

Since we only needed to compute

$k(\cdot, \cdot)$ during the entire process
without explicitly involving $\phi(x^{(i)})$'s

Proof
of
inductive
step

computation \rightarrow entire learning algorithm
can be implemented in terms of kernel
operations.

- (c) practically:- Assuming $k(x_1, x_2)$ can be
computed efficiently ($O(n)$ time)
 \rightarrow entire learning update can be
done in $O(n)$ time as
opposed to $O(N^2)$ if we were
to represent $\phi(x_i)$ s explicitly
to perform computations over
them directly.
This can be a big saving
if $N \gg n$.

Q.3. (a) f is convex if

$$f(x^1, x^2) \in \mathbb{R}^m$$

$$f(\underbrace{\alpha x^1 + (1-\alpha)x^2}_{\text{where } \alpha \in [0, 1]}) \leq \alpha f(x^1) + (1-\alpha)f(x^2)$$

convex combination of $x^1 + x^2$.

- (b) We will show the required inequality
holds using induction over k .
If $K=2$, this holds (using part (a))

[Also it is true for $k=1$
since $f(1 \cdot x^1) = 1 \cdot f(x^1)$]

Assume it holds for $(k-1)$ Consider (k)

LHS:-

$$\sum_{i=1}^k d_i f(x^{(i)}) = \sum_{i=1}^{(k-1)} d_i f(x^{(i)}) + \alpha_k f(x^{(k)})$$

$$= \left[\sum_{i=1}^{(k-1)} \alpha_i f(x^{(i)}) \right] \cdot \frac{1 \cdot \left[\sum_{i=1}^{k-1} d_i \right]}{\sum_{i=1}^{k-1} d_i} + \alpha_k f(x^{(k)})$$

Multiply
& divide
by this
quantity

$$= \underbrace{\sum_{i=1}^{k-1} d_i}_{\text{Term 2}} \left[\sum_{i=1}^{k-1} \frac{\alpha_i}{\sum_{j=1}^{k-1} \alpha_j} f(x^{(i)}) \right] + \alpha_k f(x^{(k)})$$

$$\text{Let } \frac{d_i}{\sum_{i=1}^{k-1} d_i} = \beta_i \Rightarrow \sum_{i=1}^{k-1} \beta_i = 1$$

→ Applying Jensen's inequality in Term 1

$$\left[\sum_{i=1}^{k-1} \alpha_i \right] \sum_{i=1}^{k-1} \beta_i f(x^{(i)}) + \alpha_k f(x^{(k)})$$

$$\geq \left[\sum_{i=1}^{k-1} d_i \right] f \left[\sum_{i=1}^{k-1} \beta_i x^{(i)} \right] + 1 \cdot f(x^{(k)})$$

$$\text{RHS} = \left[\sum_{i=1}^{k-1} \alpha_i \right] f \left\{ \frac{\sum_{i=1}^{k-1} [\alpha_i f(x^{(i)})]}{\sum_{i=1}^{k-1} \alpha_i} \right\} + \alpha_k f(x^{(k)})$$

$$\text{Let } \sum_{i=1}^{k-1} \beta_i x^{(i)} = \underline{x'} \text{ where } \beta_i = \frac{\alpha_i}{\sum_{i=1}^{k-1} \alpha_i}$$

$$\text{And } \sum_{i=1}^{k-1} \alpha_i = \alpha'$$

Note that $\boxed{\alpha' + \alpha_k = 1}$ $\frac{0 < \alpha' \leq 1}{0 \leq \alpha_k \leq 1}$

$$\text{RHS} = \alpha' f(\underline{x'}) + \alpha_k f(x^{(k)})$$

Applying Jensen's inequality again

$$\begin{aligned} \text{RHS} &= f(\underline{x'} \alpha' + \alpha_k x^{(k)}) \\ &= f\left(\sum_{i=1}^{k-1} \frac{x^{(i)} \alpha_i}{\alpha'} + \frac{(k-1) \alpha'}{\alpha'} + \alpha_k x^{(k)}\right) \\ &= f\left(\sum_{i=1}^{k-1} x^{(i)} \alpha_i + \alpha_k x^{(k)}\right) \\ &= f\left(\sum_{i=1}^k \alpha_i x^{(i)}\right) \end{aligned}$$

$\Rightarrow *$

$$\text{LHS} = \sum_{i=1}^k \alpha_i f(x^{(i)}) \geq f\left(\sum_{i=1}^k \alpha_i x^{(i)}\right) = \text{RHS}$$

in the original

(equation)

Hence, proved

(C) Jensen's inequality - f :- convex

$$f(E_p[x]) \leq E_p[f(x)]$$

under a given distribution $P(X=x)$

We will first prove this when $x \in \mathbb{R}^+$.

Let $P(X=x) \geq 0$ when $x \in (a, b)$

we will eventually take the limit Given:

$$a \rightarrow -\infty, b \rightarrow \infty$$

$$\underbrace{x^{(1)} - x^{(k)}}_{\text{equally spaced}}$$

Then, let

$$\Delta x = \frac{b-a}{k}$$

$$\text{Let } \underline{x}^{(0)} = a \quad x^{(i)} = a + i \Delta x$$

$$d_i = P(x^{(i-1)} \leq X \leq x^{(i)})$$

Then using Jensen's inequality

$$\sum_{i=1}^k d_i f(x^{(i)}) \geq \sum_{i=1}^k f(d_i x^{(i)})$$

where d_i as given above

$$x^{(i-1)} + \Delta x = x^{(i)} \quad \underline{x^{(i)}}$$

$$\sum_{i=1}^k P(x^{(i-1)} \leq X \leq x^{(i)} + \Delta x) f(x^{(i)})$$

$$\geq f(\sum_{i=1}^k P(x^{(i-1)} \leq X \leq x^{(i)}) \cdot x^{(i)})$$

Take $\lim K \rightarrow \infty$,

$$\Delta x = \frac{b-a}{K} \rightarrow 0$$

Sum becomes integral

$$\int_a^b P_d(x) f(x) dx = \int_a^b P_d(x) \Delta x$$

$$\Rightarrow \left[E_{P_d(x)} [f(x)] \right] \geq \int_a^b [E_{P_d(x)} [x]]$$

This holds for any given value of a &
in particular we can keep

$$a \rightarrow -\infty, b \rightarrow \infty.$$

Finally when $x \in \mathbb{R}^n$

Then intervals have to be defined along
each dimension:-

Δx corresponds to the interval between

$$x_s^{(i+1)} \text{ & } x_s^{(i)} \text{ in each dimension}$$

Jensen's inequality becomes:-

We assume K^n such
intervals (one along
each dimension)

$$\prod_{i=1}^K \prod_{j=1}^n \left[\int_{x_j^{(i)}}^{x_j^{(i+1)}} P(x_j^{(i)}) \leq x_j \leq x_j^{(i+1)} + \Delta x_j \right] f(x_1^{(1)}, \dots, x_n^{(n)})$$

$$\geq f \left(\prod_{i=1}^k \prod_{j=1}^{n_i} P(x_j^{(i-1)} \leq x_j \leq x_j^{(i-1)} + \Delta x_j) \right) \\ (x_1^{(i)}, \dots, x_n^{(i)})$$

$$\Rightarrow \prod_{i=1}^k \prod_{j=1}^{n_i} \prod_{\text{intervals}} P(x_j^{(i-1)} \leq x_j \leq x_j^{(i-1)} + \Delta x_j) \\ f(x^{(i)})$$

Becomes integral
in the limit $\geq f \left(\prod_{i=1}^k \prod_{j=1}^{n_i} P(x_j^{(i-1)} \leq x_j \leq x_j^{(i-1)} + \Delta x_j) x^{(i)} \right)$

$$\Rightarrow \iint \cdots \int_{\bar{x}_1=a_1}^{b_1} \cdots \int_{\bar{x}_n=a_n}^{b_n} P_d(x_1, \dots, x_n) \cdot dx_1 \cdots dx_n \int f(x^{(i)})$$

$$\geq f \left(\int_{x_1=a_1}^{b_1} \int_{x_2=a_2}^b \cdots \int_{x_n=a_n}^{b_n} P_d(x_1, \dots, x_n) \right) dx_1 \cdots dx_n$$

$$\Rightarrow E_{P(x)}[f(x)] \geq f_{P(x)}[E[x]]$$

We can take $\lim a_1, \dots, a_n \rightarrow -\infty$
& $b_1, \dots, b_n \rightarrow \infty$

To argue for the general case.

Q.4. Unconstrained SVMs:-

$$\min_{w, b, \epsilon} \frac{1}{2} w^T w + C \sum_{i=1}^m \epsilon_i$$

$$y^{(i)} (w^T x^{(i)} + b) \geq 1 - \epsilon_i \quad \text{--- (1)}$$

$$\epsilon_i \geq 0 \quad \text{--- (2)}$$

- not all equations in (1) we get

From 5 to 8 years

$$g_{e_i} \geq 1 - y^{(i)} / (\omega^T x^{(i)} + b)$$

& ② tells us that

$$k_{ij} \geq 0$$

Together these equations tell us that

$$g_{\text{ei}} = \max(0, 1 - y^{(1)}(w^T x^{(1)} + b))$$

Also, since we are optimizing w w.r.t. ϵ_i , we have

$$\min_{w, b, \epsilon_i} \quad y_i w^T w + c \sum_{i=1}^n \epsilon_i$$

it must be the case that

$$L_{ij} = \max(0, 1 - y^{(i)}(\omega^T x^{(i)} + b))$$

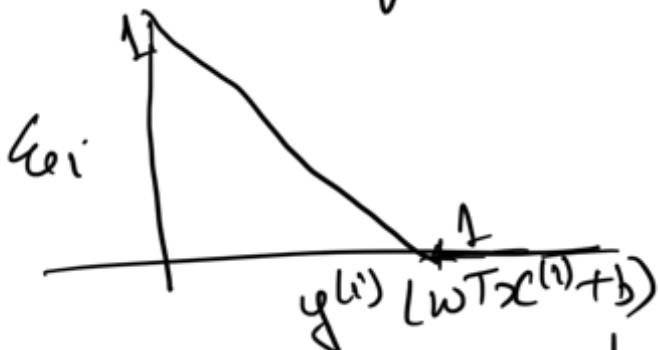
(since we need to minimize wrt w_i 's)

$$\Rightarrow \text{hei} = 1 - y^{(i)}(\bar{w}^T x^{(i)} + b) \quad \text{if } y^{(i)}(\bar{w}^T x^{(i)} + b) \leq 1$$

O. W.

→ Hinge loss wrt $y^w(w^T x^w + b)$

$$L_H(y^{w_1(\bar{w}^T x^w + b)})$$



\Rightarrow unconstrained objective can be written as
 obtained by substituting the value for x_1, x_2, \dots, x_m

$$\text{Loss function} = \min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^m h_H(z_i)$$

unconstrained loss \mathcal{L}

$$z_i = y^{(i)}(w^T x^{(i)} + b)$$

$$h_H(z_i) = 1 - z_i \quad \text{if } z_i \leq 1$$

$h_H(w, b) \downarrow$

$$\Rightarrow \min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^m \max(0, 1 - y^{(i)}(w^T x^{(i)} + b)) \quad \text{--- (3)}$$

Intuitively, the loss term is non zero if point is inside the margin boundary

(b) Let us differentiate wrt w

$$\nabla_w \frac{1}{2} w^T w + C \sum_{i=1}^m \max(0, 1 - y^{(i)}(w^T x^{(i)} + b))$$

$$= w + C \sum_{i=1}^m \nabla_w \max(0, 1 - y^{(i)}(w^T x^{(i)} + b))$$

$$\text{if } y^{(i)}(w^T x^{(i)} + b) > 1$$

$$\underbrace{\nabla_w y^{(i)}(w^T x^{(i)} + b)}_{\text{if } y^{(i)}(w^T x^{(i)} + b) = 1} = 0$$

0
(wrong right)

at the point
of discontinuity)

$$C[-y^{(i)}x^{(i)}]$$

Gradient descent update -

$$\left. \begin{array}{l} w^{(t+1)} \leftarrow w^{(t)} - \eta \cdot \nabla_w L_u(w, b) \\ b^{(t+1)} \leftarrow b^{(t)} - \eta \cdot \nabla_b L_u(w, b) \end{array} \right\}$$

where gradients are as defined above.

Note :- loss is -ve of $L(\theta)$. We will show $L(\theta)$ is concave \Rightarrow loss is Convex.

Q.S.

L logistic Regression!

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$L_2(0) = \sum_{i=1}^m y^{(i)} \log h_0(x^{(i)}) + (1-y^{(i)}) \log (1-h_0(x^{(i)}))$$

$$\nabla_{\theta} \text{LL}(\theta) = \sum_{l=1}^m \frac{y^{(l)}}{h_{\theta}(x^{(l)})} \cdot \nabla_{\theta} h_{\theta}(x^{(l)}) + \frac{(1-y^{(l)})}{1-h_{\theta}(x^{(l)})} \nabla_{\theta} (-h_{\theta}(x^{(l)}))$$

$$= \sum_{l=1}^m \nabla_{\theta} h_{\theta}(x^{(l)}) \left[\frac{y^{(l)}}{h_{\theta}(x^{(l)})} - \frac{(-g^{(l)})}{1-h_{\theta}(x^{(l)})} \right]$$

$$= \sum_{l=1}^m \nabla_{\theta} h_{\theta}(x^{(l)}) \left[\frac{y^{(l)}}{h_{\theta}(x^{(l)})} - \frac{(-y^{(l)})}{1-h_{\theta}(x^{(l)})} \right]$$

$$= \sum_{l=1}^m \cancel{h_{\theta}(x^{(l)})} \cancel{(1-h_{\theta}(x^{(l)}))} x^{(l)} \left[\frac{y^{(l)} (h_{\theta}(x^{(l)})) - (1-y^{(l)}) h_{\theta}(x^{(l)})}{\cancel{1-h_{\theta}(x^{(l)})}} \right]$$

$$= \frac{\sum_{i=1}^m x^{(i)} (y^{(i)} - h_\theta(x^{(i)}))}{m}$$

$$\begin{aligned} \nabla_{\theta}^2 L(\theta) &= \sum_{i=1}^m h_{\theta}(x^{(i)}) (1 - h_{\theta}(x^{(i)})) [x^{(i)} x^{(i)\top}] \\ &= - \sum_{i=1}^m [h_{\theta}(x^{(i)}) (1 - h_{\theta}(x^{(i)}))] x^{(i)} x^{(i)\top} \\ &\equiv \text{Hessian Matrix} \end{aligned}$$

For Convexity, we need to show that H is
-ve semi-definite \Rightarrow

$$\begin{aligned} Z^T H Z &\leq 0 \quad Z \in \mathbb{R}^{(n+1) \times (n+1)} \\ \Rightarrow Z^T \left[- \sum_{i=1}^m [h_{\theta}(x^{(i)}) (1 - h_{\theta}(x^{(i)}))] x^{(i)} x^{(i)\top} \right] Z &\leq 0 \\ \Rightarrow - \sum_{i=1}^m h_{\theta}(x^{(i)}) (1 - h_{\theta}(x^{(i)})) Z^T x^{(i)} x^{(i)\top} Z &\\ = - \sum_{i=1}^m h_{\theta}(x^{(i)}) (1 - h_{\theta}(x^{(i)})) \frac{[(x^{(i)\top} Z)^2]}{\leq 0} & \\ \Rightarrow - \sum_{i=1}^m h_{\theta}(x^{(i)}) (1 - h_{\theta}(x^{(i)})) \underbrace{(x^{(i)\top} Z)^2}_{\geq 0} &\leq 0 \end{aligned}$$

\Rightarrow Hessian matrix is -ve semi-definite
 $\Rightarrow L(\theta)$ is convex \Rightarrow Loss function $= -L(\theta)$

(a) by $P(X) = \frac{1}{Z} \prod_k \phi_k(x_{ck})$

$x_j \in \{0, 1\}$ for (Nodes represent Boolean valued variables in the given setting)

Since probabilities should add up to 1

$$Z = \sum_{\substack{x \in \{0, 1\}^n \\ x_1 \in \{0, 1\}, x_2 \in \{0, 1\} \dots x_m \in \{0, 1\}}} \frac{1}{Z} \prod_k \phi_k(x_{ck})$$

Here x_{ck} denote the value of x restricted to variables in X_{ck}

(b) Recall that a distribution defined by a Bayesian network is given as:-

$$P(X) = \prod_{j=1}^n P(X_j | P_a(X_j))$$

Now, given a Bayesian network \mathcal{B} , let us define a ~~prob~~ corresponding network where π_i factors are present.

Markov
involves a variable and lots of
factors
and there are as many factors
as there are variables then

Markov equivalent network

$$P(X) = \frac{1}{Z} \prod_{j=1}^n \phi_j(x_j, P(x_j))$$

where $Z = 1$

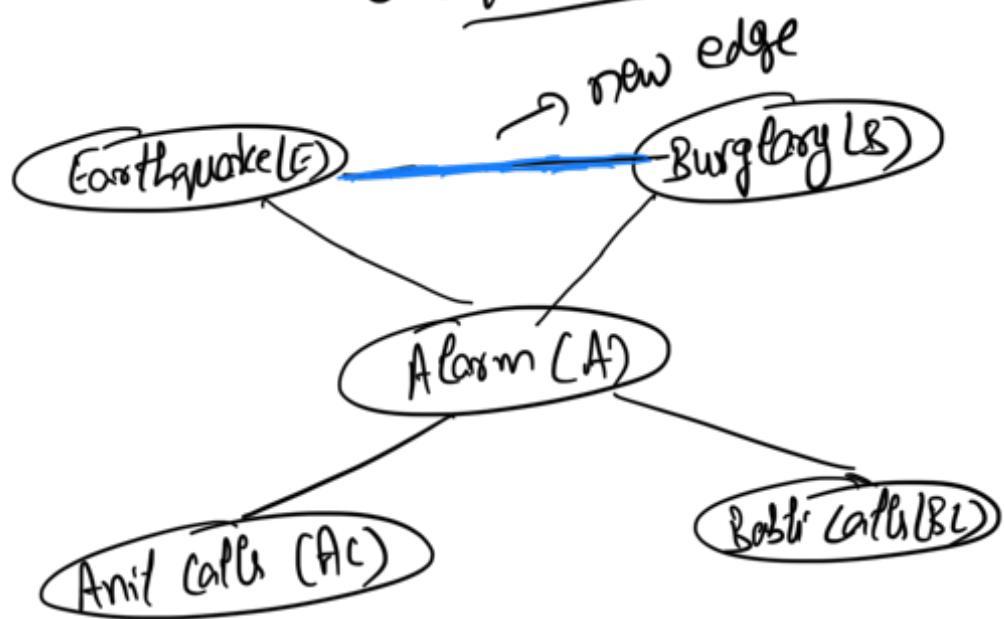
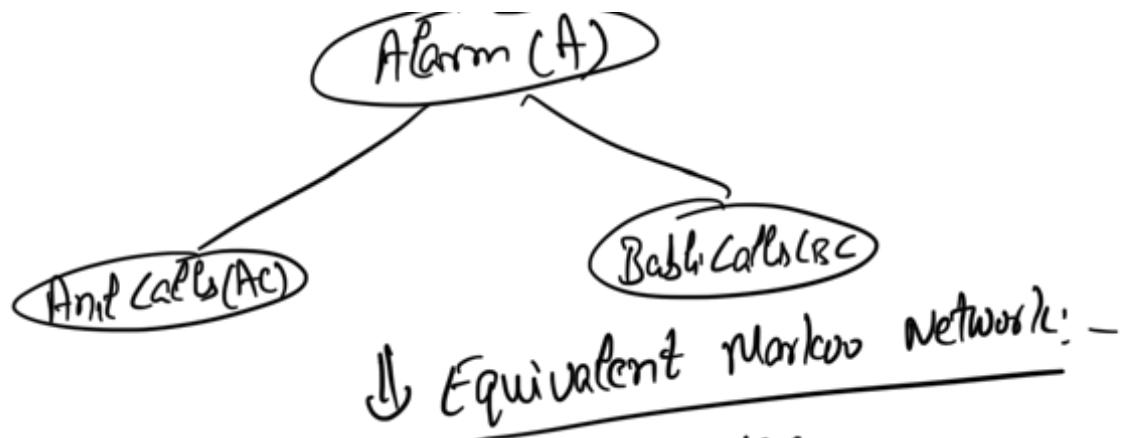
$$\phi_j(x_j, P(x_j)) = \frac{P(x_j | P(x_j))}{\prod_i}$$

comes from the
parameters of
CTs in the
Bayesian
network

In particular, the
equivalent Markov network
graph is obtained by connecting
all the parents of each node
(directionally) with each other & dropping the
arrows in Bayesian network arcs.
Since factors are defined over cliques
& above procedure results in dense set
of cliques.

Example (for illustration only): -





Q.T: (a) $\min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^m \epsilon_i$

$$y^{(i)} (w^T x^{(i)} + b) \geq 1 - \epsilon_i \quad \forall i$$

$$\epsilon_i \geq 0 \quad \forall i$$

In the limit $C \rightarrow \infty$, model strictly adheres to data, disallowing any form of noise. Any point outside the margin boundary is given a very high penalty (which goes to infinity). Therefore, in the limit $C \rightarrow \infty$

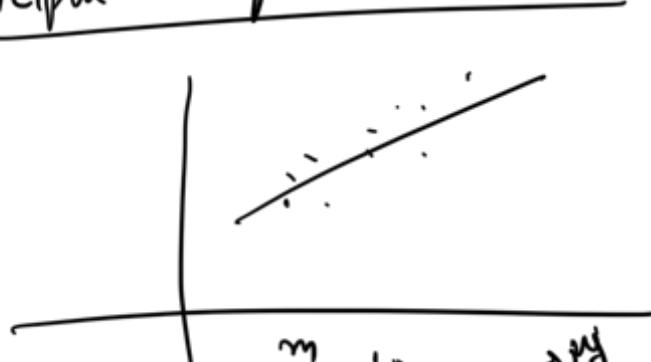
as $C \rightarrow \infty$, it's more likely to overfit
the model is since it really tries hard
to fit every point on the "right" side of
margin, even if it happened to be an
outlier. For smaller values of C , this
is not the case, & model can softly
deal with noise (i.e., by passing a
finite ρ).

- (b) Since both training & testing errors are low,
this is the cause of underfitting [the model
is not able to fit even the patterns
observable in training data as evident
from high training error].
- As a remedy to underfitting as
described above, we would try a more
sophisticated model first. Giving
more training data is not going to
help since the model may not simply
be able to fit this additional amount
of data. A more sophisticated
model may reduce the training
errors by being able to fit better
fit patterns seen in the training data.

(C) Decision trees are unstable learners since minor changes in the training distribution can lead to substantial changes in the learned tree by altering the choice of a node/attribute to split on at a particular level in the tree growing phase.

On the other hand, Naive Bayes is a relatively stable learner since minor changes in the training distribution will only result in minor change in the learned probabilities (parameters) of the underlying model.

Q.8. Principal Component Analysis (PCA)



PCA tries to

Maximize the variance of the projected data

$$\left\{ \sum_{i=1}^m x_i^{(w)} \right\}_{i=1}^m \text{ s.t. } \frac{1}{m} \sum_{i=1}^m x_i^{(w)} = 0 \quad \begin{matrix} \rightarrow \mu_w \\ \text{(mean is zero)} \end{matrix} \right\}$$

$$s.t. - \frac{1}{m} \sum_{i=1}^m (x_i^{(w)} - 0)^2 = 1 \quad \begin{matrix} \rightarrow \sigma_w^2 \\ \text{(variance is 1)} \end{matrix} \quad \text{given}$$

Let the projected dimensions be $w_1 - w_k$.

Then projection of x^w along the direction u_e is given as -

$$x^w T u_e$$

Note since the original data is 0 mean,
projected data is also zero mean

$$\frac{1}{m} \sum_{l=1}^m [x^{(l)T} u_e] = \frac{1}{m} \sum_{l=1}^m x^{(l)T} u_e = 0^T u_e = 0$$

Total variance of the projected data is -

$$\sum_{u_1-u_k} \left[\frac{1}{m} \sum_{l=1}^m [x^{(l)T} u_e]^2 \right] \quad \begin{array}{l} \text{i.e. sum of variances} \\ \text{along each of the} \\ \text{projected dimension}. \end{array}$$

$$= \frac{1}{m} \sum_{u_1-u_k} \sum_{l=1}^m [u_e^T x^{(l)}] [x^{(l)T} u_e]$$

$$= \frac{1}{m} \sum_{u_1-u_k} \sum_{l=1}^m u_e^T [x^{(l)} x^{(l)T}] u_e$$

$$= \sum_{u_1-u_k} u_e^T \underbrace{\frac{1}{m} \sum_{l=1}^m x^{(l)} x^{(l)T}}_{\text{Empirical co-variance matrix}} u_e$$

\sum :- Empirical co-variance matrix

$$= \sum_{u_1-u_k} [u_e^T \sum u_e] : - \text{Projected variance.}$$

Therefore the goal in PCA is to find
 $u_1 - u_k$ such that

$$\sum_{u_1-u_k} u_e^T \sum u_e \quad \begin{array}{l} \text{is maximized} \\ \text{Empirical co-variance} \\ \text{matrix} \end{array}$$

$$\text{arg max } (u_e^T \sum u_e)$$

:- $\max_{\mathbf{u}_1, \dots, \mathbf{u}_k} \mathbf{u}_1^T \mathbf{u}_2 + \dots + \mathbf{u}_{k-1}^T \mathbf{u}_k$
 Subject to:- $\|\mathbf{u}_i\|^2 = 1$ i.e. $\mathbf{u}_i^T \mathbf{u}_i = 1$ &
 $\mathbf{u}_i^T \mathbf{u}_j = 0$ for $i \neq j$
 (orthonormality of $\mathbf{u}_1, \dots, \mathbf{u}_k$)

(b)

First let us solve for $k=1$

$$\max_{\mathbf{u}_1} \mathbf{u}_1^T \sum_i \mathbf{u}_i$$

$$\mathbf{u}_1^T \mathbf{u}_1 = 1$$

Lagrangian:-

$$L(\mathbf{u}, \alpha_1) = \mathbf{u}_1^T \sum_i \mathbf{u}_i + \alpha_1 (1 - \mathbf{u}_1^T \mathbf{u}_1)$$

using KKT (necessary condition for optimality)

$$\nabla_{\mathbf{u}} L(\mathbf{u}, \alpha_1) = 0$$

$$\Rightarrow \frac{\partial}{\partial \mathbf{u}_1} L(\mathbf{u}, \alpha_1) = \sum_i \mathbf{u}_i - \alpha_1 \mathbf{u}_1 = 0$$

$$\Rightarrow \frac{\partial}{\partial \mathbf{u}_1} \left[\sum_i \mathbf{u}_i - \alpha_1 \mathbf{u}_1 \right] = \alpha_1 \mathbf{u}_1$$

$\Rightarrow \mathbf{u}_1$ is eigenvector of $\sum_i \mathbf{u}_i$.

Also, since we want to maximize

$$\mathbf{u}_1^T \left[\sum_i \mathbf{u}_i \right] = \mathbf{u}_1^T \alpha_1 \mathbf{u}_1 = \alpha_1$$

α_1 must correspond to highest eigenvalue

:- \mathbf{u}_1 is eigenvector corresponding to highest eigenvalue

② $k=2$ we want to maximize $\mathbf{u}_1^T \sum_i \mathbf{u}_i \mathbf{u}_1$

$$\max_{u_1, u_2} \left[u_1^T \sum u_i + u_2^T \sum u_i \right]$$

Aside

Empirical covariance between $u_1^T x$ and $u_2^T x$ is

$$\begin{aligned} & \rightarrow \sum_{i=1}^m u_1^T x^{(i)} x^{(i) T} u_1 \\ & = \sum_{i=1}^m \sum_{j=1}^m u_1^T (x^{(i)} x^{(j) T}) u_1 \\ & = \frac{1}{m} \sum_{i=1}^m u_1^T u_1 \\ & = \frac{1}{m} \sum_{i=1}^m u_2^T u_2 \\ & = \frac{1}{m} \sum_{i=1}^m u_2^T u_2 = 0 \\ & = \sum_{i=1}^m u_2^T u_2 \text{ will be zero} \end{aligned}$$

subject to constraint $u_1^T u_1 = 1$
 $u_2^T u_2 = 1$

$$u_1^T u_2 = 0$$

$$L(u_1, u_2, \alpha_1, \alpha_2, \beta)$$

$$\begin{aligned} &= u_1^T \sum u_i + u_2^T \sum u_i \\ &\quad + \alpha_1 (1 - u_1^T u_1) \\ &\quad + \alpha_2 (1 - u_2^T u_2) \\ &\quad + \beta u_1^T u_2 \end{aligned}$$

$$\nabla_{u_2} L(u_1, u_2, \alpha_1, \alpha_2, \beta) = 0$$

[using KKT]

$$= 2 \sum u_2 + 2\alpha_2 (-u_2)$$

$$+ \beta u_1 = 0$$

\Rightarrow Pre-multiplying by u_1^T

$$\underbrace{2 u_1^T \sum u_2}_{0} + \underbrace{\alpha_2 u_1^T (-u_2)}_0 + \underbrace{\beta (-u_1^T u_1)}_{-1} = 0$$

Empirical covariance between projections along u_1 & u_2

$$\Rightarrow \underline{\beta = 0}$$

$$\Rightarrow 2 \sum u_2 = \alpha_2 u_2 \Rightarrow \boxed{\sum u_2 = \alpha_2 u_2}$$

... corresponds to estimator of α_2

$\rightarrow u_2$ ~~and~~

further we want to maximize

$$\max_{u_1, u_2} [u_1^T \Sigma u_1 + u_2^T \Sigma u_2]$$

$$= \max_{u_1, u_2} [d_1 + d_2]$$

must correspond to highest
eigenvalues of Σ .
Empirical co-variance matrix.

D.g. Decision Tree Pruning:-

Let

T_p :- Tree obtained by greedy
pruning strategy

T_{best} :- best tree obtained any given
pruning strategy. Let T_{best} be
minimal ES such tree i.e.
no more nodes in T_{best} can
be pruned.

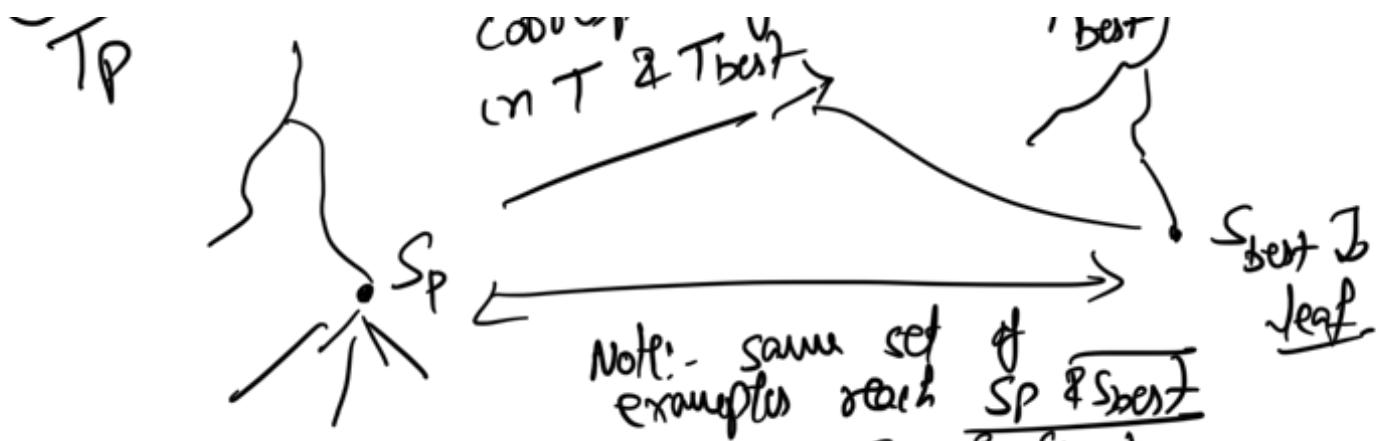
Let $\rightarrow \text{Aval}(T_{best}) > \text{Aval}(T_p)$
We will prove by contradiction that
above is not possible

Consider some sub-tree in T_p which was

(A)

pruned in T_{best} :-
corresponding sub-trees

$T_{...}$



Then let us consider S_P & S_{best}
if $Aval(S_{best}) > Aval(S_P)$

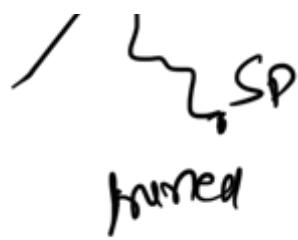
Then, clearly T_P is not final
tree after following greedy strategy
because we can prune tree below
 S_P & get more accuracy on
Validation set.

$$\Rightarrow Aval(S_{best}) \leq Aval(S_P)$$

Since S_{best} is best part of
best possible in tree T_{best}
it must be the case that
 $Aval(S_{best}) = Aval(S_P)$

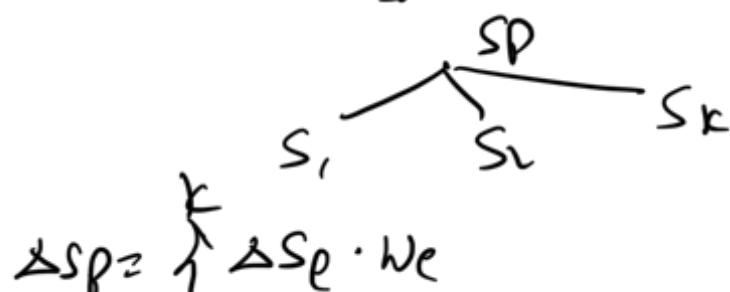
Next, let us consider subtree such
that it is pruned in T_P but
not in T_{best}





Then let $Aval(Sbest) > Aval(SP)$
But this is not possible since

claim:- if in greedy strategy if any subtree SP
is pruned, then it must be the
case that pruning any of its children
would have also resulted in no loss in
accuracy. If not :- *



$$\Delta SP = \sum \Delta Sp \cdot we$$

$$l=1$$

$$\text{If } \Delta Sp < 0$$

for some l :

3 l' s.t

$$\boxed{\Delta Sp > \Delta Sp'}$$

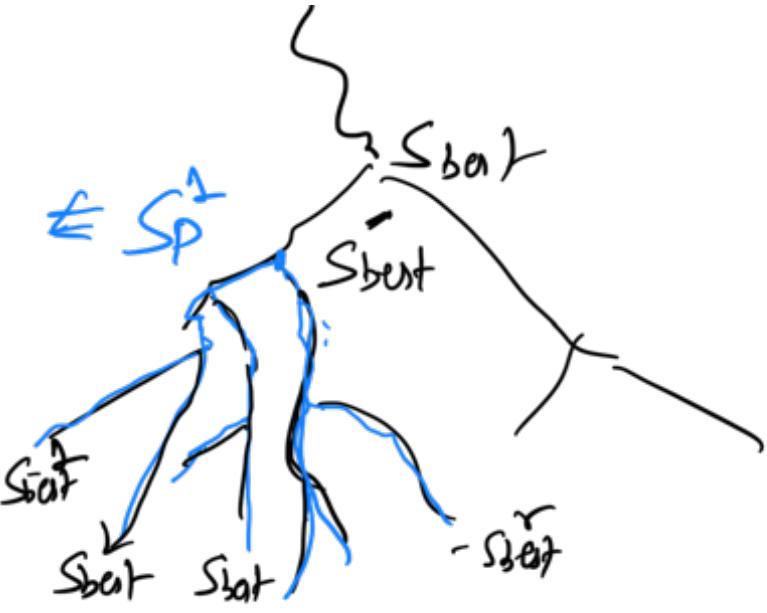
$0 \leq we \leq 1$
weight
of that
branch

if any of its
children result
in loss in accuracy
by pruning them, then
clearly one of its
other children must give
more increase in accuracy
by pruning it

Thus is true for every
intermediate tree obtained
during pruning

Let us consider

Frost sub-tree
pruned such
that parts
of S_p^2 are
present in
 S_{best} .



Then:- clearly recursively speaking

$\Delta \text{Avul}(S_p^2)$ is at least as
good as $\Delta \text{Avul}(S_{best})$ \Leftarrow F.C.D

If not:- we would have pruned that
sub-tree first \Rightarrow contradiction.

$$\Rightarrow \text{Avul}(S_p^2) \geq \text{Avul}(S_{best})$$

but since S_{sat} is best possible

$$\rightarrow \text{Avul}(S_p^2) = \text{Avul}(S_{best})$$

\Rightarrow We can get T'_{best} s.t.

tree below S'_{best} is pruned

$$\& \text{Avul}(T'_{best}) = \text{Avul}(T_{best})$$

$\Rightarrow S_{best}$ is not minimal

\rightarrow contradiction.

- \nexists T_p & S_{best} s.t. T_p is
granted by S_{best} is not
~~(differing)~~
- \Rightarrow All ~~sub-trees~~ of T_p are
either identical to T_{best}
or have same ~~if wrong~~. contradiction
- $T_A \sqsubset T_B = \text{Avail}(T_{best})$