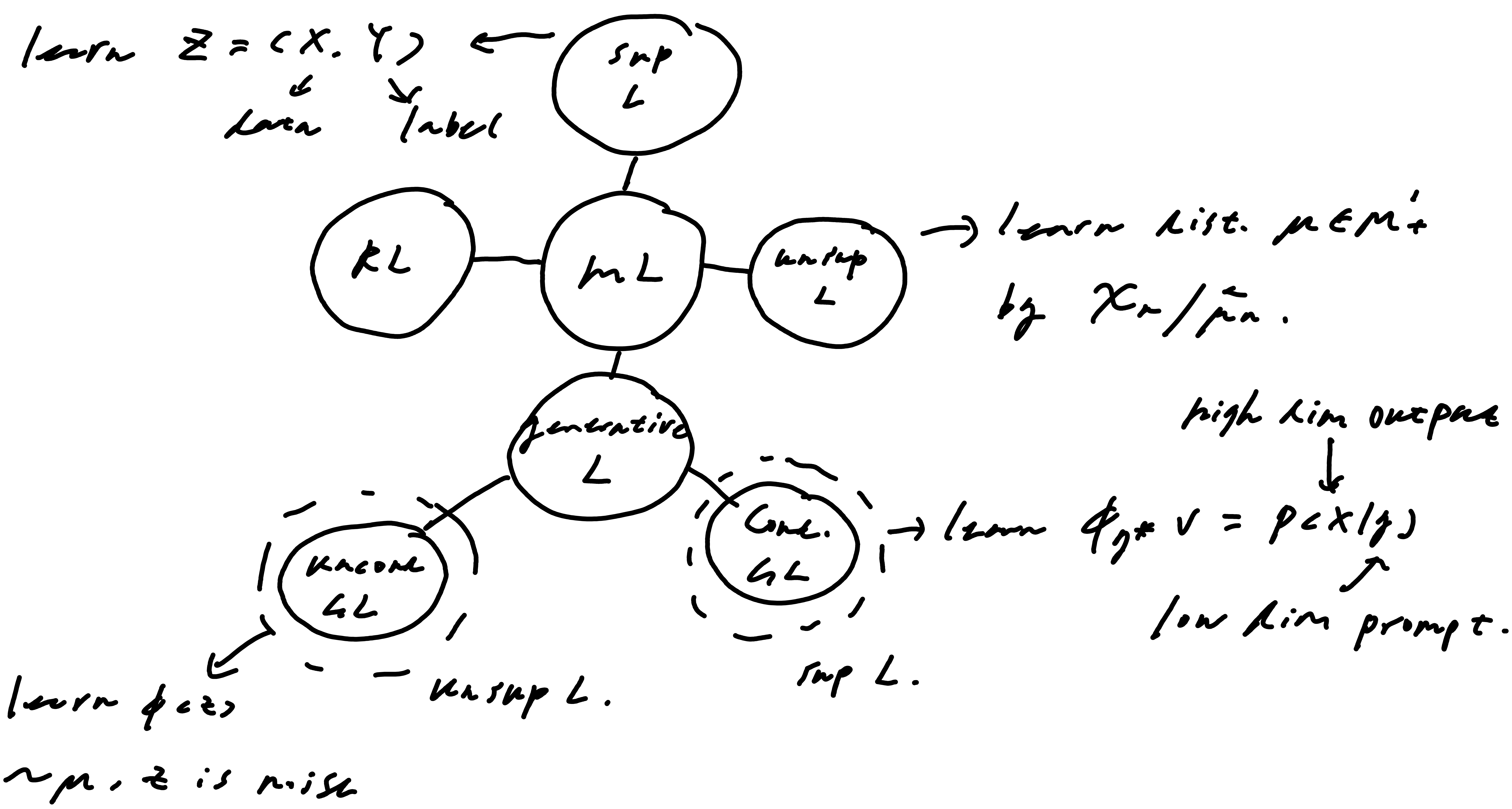
Supervised Learning



Rmk: Supervised learning is easier than unsupervised one since the dim. of label is low in common but data of unsupervised has high dim.

In supervised learning, we observe $Z_j = (Y_j, X_j) : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}^s \times \mathbb{R}^d, \mathcal{B}_{\mathbb{R}^s \times \mathbb{R}^d})$ in sample $\mathcal{X}_n = (Z_1, \cdots Z_n)$. Z_k i.i.d. where Y_j is label/output and X_j is covariate/input.

Our goal is not to learn distribution of Z_j

rather to learn the dist. of $Y|X$. i.e. propose a dist. $\hat{\mu}_{|X, \mathcal{X}_n}$ from $\mathcal{X}_n = (Z_1 \cdots Z_n)$ to get close to $\mu_{|X} = P(Y|X)$.

e.g. Y is r.v. of products customer will buy

X is shopping history ... info. of customer

$\Rightarrow$ interested in $\mathbb{P}(Y = \text{"soap"} | X) = ?$

(1) Regular Conditional proba.:

Lem: For $\tilde{A} \subset A$. $\Rightarrow \mathbb{E}(\cdot | \tilde{A}) = \text{Proj.} : L^2(\Omega, A, \mathbb{P}; \mathbb{R}') \to L^2(\Omega, \tilde{A}, \mathbb{P}; \mathbb{R}')$. ortho. proj.

This fixes $\mathbb{E}(\cdot | \tilde{A})$. $\mathbb{P}$-a.s.

Rmk: Recall i) $\text{Proj}\, X = \underset{Y \in L^2(\tilde{A})}{\arg\min}\, \mathbb{E}((X - Y)^2)$.

ii) $\text{Proj} \circ \text{Proj} = \text{Proj}$. iii) $\text{Proj} = id_{L^2(\tilde{A})}$

iv) $\mathbb{E}(X\, \text{Proj}\, Y) = \mathbb{E}(\text{Proj}\, X \cdot Y)$

(By ortho. decompose of X, Y)

Lem. If $Y \in \sigma(X)$. $Y : \Omega \to \mathbb{R}'$. r.v. $X : \Omega \to \mathbb{R}^d$. r.v. Then: $\exists g : (\mathbb{R}^d, B_{\mathbb{R}^d}) \to (\mathbb{R}', B_{\mathbb{R}'})$ measurable. st. $Y = g(X)$. g is unique

$\mathbb{P}_X$ - a.s. i.e. $g' = g$. $\mathbb{P}_X$ - a.s. $\Rightarrow g'(X) = Y$. $\mathbb{P}$-a.s.

__Def__: i) $\{P_{Y|X}(A)\}_{A \in \mathcal{A}}$ is regular conditional probability (r.c.p.) if it's induced by a Markov kernel $K(x, B) = P_{Y|X=x}(B)$. $\mathbb{P}_X$.a.s. x &

a) $B \mapsto K(x, B)$ is p.m. for $\forall x \in \mathbb{R}^d$.

b) $x \mapsto K(x, B)$ is $\mathcal{B}_{\mathbb{R}^d}$-measurable. $\forall B \in \mathcal{A}$.

ii) $P_{Y|X=x}(B)$ admits a regular version if $\bar{P}_{Y|X}(\cdot)$ is r.c.p. and $P_{Y|X} = \bar{P}_{Y|X}$ a.s.

__Rmk__: Note if let $P_{Y|X}(A) = \mathbb{E}(I_{\{Y \in A\}} | \sigma(X))$. then: $P_{Y|X}(\cdot)$ also only satisfies property of p.m. $\mathbb{P}$-a.s. But the null set N will depend on $\{A_n\}$ we choose if it's not r.c.p.

$\Rightarrow$ it may not be p.m. for $\mathbb{P}_X$-a.s. x.

Supervised learning is to learn Markov kernel.

__e.g.__ $f(y|x) \stackrel{\Delta}{=} (\sqrt{2\pi\sigma^2})^{-\frac{1}{2}} e^{-\frac{1}{2}(\frac{y-x}{\sigma})^2} dy \Rightarrow$

$K(x, B) = \int_B f(y|x) dy$ is a Markov kernel.

for linear regression

Thm. $Z = (Y, X) : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}^s \times \mathbb{R}^d, \mathcal{B}_{\mathbb{R}^s \times \mathbb{R}^d})$ s.t. $Y \in L^2$-r.v. $\Rightarrow$ There exists a regular version $P_{Y|X}(A)$.

Lem. The dist. of $Z = (Y, X)$: P_Z is uniquely determined by its marginal $\mathbb{P}_X$ and r.c.p. $\mathbb{P}_{Y|X=x}(\cdot)$. i.e. we have: $\forall B \in \mathcal{B}_{\mathbb{R}^d \times \mathbb{R}^s}$.

$$\mathbb{P}_Z(B) = \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^s} I_B(y, x) \, d\mathbb{P}_{Y|X=x}(y) \right) d\mathbb{P}_X(x).$$

Lem. If $Z = (X, Y)$ has density $f_Z(x, y)$. let $f_X(x) = \int_{\mathbb{R}^s} f_Z(x, y) \, dy$ and set

$$f(y|x) = \begin{cases} f_Z(x, y) / f_X(x). & \text{if } f_X(x) > 0 \\ I_{[0,1]^s}(y) . & \text{otherwise.} \end{cases}$$

Then : $P_{Y|X=x}(B) = \int_{\mathbb{R}^s} I_B(y) f(y|x) \, dy$ def a r.c.p. of Y over $X = x$.

(2) **Divergence:**

Next we want to introduce a divergence to measure success of learning from $\hat{P}_{Y|X, x_n}$ to true Markov kernel $P_{Y|X}$.

Def: For two Markov kernel $\mu|X=x$, $\nu|X=x$ and divergence $d(\cdot\|\cdot)$ on $M_1^+(\mathbb{R}^l)$. s.t. $x\in\mathbb{R}^d \mapsto d(\mu|X=x \| \nu|X=x) \in (\mathbb{R}^1, \mathcal{B}_{\mathbb{R}^1})$ is measurable for $\forall \mu|X, \nu|X \in \mathcal{K}_1^+(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^l})$ space of Markov kernels. Set:

$D_p(\mu|X \| \nu|X) = D_{p,X,d}(\mu|X \| \nu|X)$

$= \mathbb{E}_X(d(\mu|X \| \nu|X)^p)^{\frac{1}{p}}$

$= (\int_{\mathbb{R}^d} d(\mu|X=x \| \nu|X=x)^p \, dP_X(x))^{\frac{1}{p}}$. $1\le p<\infty$.

$D_\infty(\mu|X \| \nu|X) = \operatorname*{ess\,sup}_{x\in\mathbb{R}^d} d(\mu|X=x \| \nu|X=x)$

Rmk: i) $p\uparrow$. then sensitivity of $d(\mu|X=x \| \nu|X=x)$ at each pt. $x \uparrow$.

ii) $D_{1,TV}$ is weaker than $D_{2,KL}$.

(D will mostly inherit the strength of d. But it's not from Pinsker's)

iii) If $X\sim P_X$ but we let $X\sim \tilde{P}_X$ during inference. Suppose that

$D_{1,X\sim P_X,d}(\mu|X \| \hat{\mu}|X,n) \to 0$. a.s.

Note $D_{1,X\sim \tilde{P}_X,d}(\mu|X \| \hat{\mu}|X,n) =$

$\mathbb{E}_{x\sim p_x}\left(d(\mu_{|x}\|\hat{\mu}_{|x,n})\frac{d\bar{P_x}}{dP_x}\right) \le$

$\left\|\frac{d\bar{P_x}}{dP_x}\right\|_\infty D_{1,x\sim p_x,d}(\mu_{|x}\|\hat{\mu}_{|x,n}) \xrightarrow{a.s.} 0$ if $\left\|\frac{d\bar{P_x}}{dP_x}\right\|_\infty < \infty$

So the real task can still be covered in training

Lem. Denote $V_{|x}P_x(B) \overset{\Delta}{=} \int\int I_B(y,x)\,dV_{|X=x}(y)\,dP_x(x)$

For $V_{|x}, \mu_{|x} \in \mathcal{K}_1^+(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d})$. We have:

$D_{1,x,KL}(\mu_{|x}\|V_{|x}) = d_{KL}(\mu\|V)$. where $\mu_{|x}$ $= \mu_{|x}P_x$. $V = V_{|x}P_x$.

Rmk: In supervised learning we always know μ and V.

Pf: i) We first prove: $\frac{dV}{d\mu}(y,x) \overset{\mu\text{-a.s.}}{=} \frac{dV_{|X=x}}{d\mu_{|X=x}}(y)$.

a) $\mu \ll V \Rightarrow \mu_{|X=x} \ll V_{|X=x}$. P_x-a.s. x.

otherwise. $\exists B_1$ s.t. $P_x(B_1)>0$ satisfy:

$\exists B_2(x)$. s.t. $\mu_{|X=x}(B_2(x)) > 0 = V_{|X=x}(B_2(x))$

Set $B = \{(y,x): x\in B_1, y\in B_2(x)\}$

$\Rightarrow \mu(B) = \int \mu_{|X=x}(B_2(x))\,I_{B_1}(x)\,dP_x > 0$

But $V(B) = \int V_{|X=x}(B_2(x))\,I_{B_1}(x)\,dP_x = 0$.

$\Rightarrow$ contradiction.

b) And we set $\int I_B(y,x) \frac{dV_{|X=x}}{d\mu_{|X=x}}(y) d\mu(y,x)$

$= \int \left(\int I_B(y,x) \frac{dV_{|X=x}}{d\mu_{|X=x}}(y) \, d\mu_{|X=x}(y) \right) dP_X(x)$

$= \int \left(\int I_B \, dV_{|X=x}(y) \right) dP_X(x) = V(B)$

So: $\mu_{|X=x} \ll V_{|X=x}$, μ-a.s. $x \Rightarrow \mu \ll V$.

and $\frac{dV}{d\mu}(y,x) = \frac{dV_{|X=x}}{d\mu_{|X=x}}(y)$. μ.a.s.

2) $d_{KL}(\mu \| V) \overset{i)}{=} -\int \log\left(\frac{dV_{|X=x}}{d\mu_{|X=x}}(y) \right) d\mu(y,x)$

$= \int \left(-\int \log\left(\frac{dV_{|X=x}}{d\mu_{|X=x}}(y) \right) d\mu_{|X=x}(y) \right) dP_X(x)$

$= D_{1.X.KL}(\mu_{|X} \| V_{|X})$.

(3) <u>Framework:</u>

Denote $\mathcal{J} \subseteq \mathcal{K}_1^+(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^s})$ is set of target Markov kernels. $\mathcal{X}_n = (Z_1, \cdots Z_n)$ is samples and $\hat{\mu}_{n|X} \overset{\Delta}{=} \hat{\mu}_{n|X} \circ \mathcal{X}_n$. For D div. We'll require: $\mathcal{X} \in \mathbb{R}^{(s+d)n} \mapsto D(\mu_{|X} \| \hat{\mu}_{n,|X}(\mathcal{X})) \in \mathbb{R}^+$ is $\mathcal{B}_{\mathbb{R}^{(d+s)n}}$ - measurable.

<u>Def</u>: $\mathcal{H}_n \subseteq \mathcal{K}_1^+(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^s})$ hypothesis space of Markov kernel. $Z_j = (Y_j, X_j) \overset{i.i.d}{\sim} \mu = \mu_{|X} P_X$

Let $D = D_{p,X,d}$ div. on $\mathcal{K}_1^+$.

i) Empirical risk function $\hat{\mathcal{L}}_n = \hat{\mathcal{L}}_n(\nu_{|X}, \mathcal{X}_n)$ satisfies:

a) $\mathbb{R}^{(d+1)n} \ni \mathcal{X}_n \mapsto \hat{\mathcal{L}}_n(\nu_{|X}, \mathcal{X}_n)$ is measurable for $\forall \nu_{|X} \in \mathcal{H}_n$

b) $\exists h(\cdot)$ on $\mathcal{M}_1^+$, $(c_n) \subset \mathbb{R}^+$ s.t.

$$c_{n_k} \hat{\mathcal{L}}_{n_k}(\nu_{|X}, \mathcal{X}_{n_k}) + h_{n_k}(\mu) \xrightarrow{pr} D_{p,X,d}(\mu_{|X} \| \nu_{|X})$$

$\forall (n_k)$ s.t. $\nu_{|X} \in \mathcal{H}_{n_k}$.

ii) $\hat{\mathcal{L}}_n$ is unbiased if $E_z(c_{n_k} \hat{\mathcal{L}}_{n_k}(\nu_{|X}, \mathcal{X}_{n_k}) + h_{n_k}(\mu)) = D_{p,X,d}(\mu_{|X} \| \nu_{|X})$.

iii) $\hat{\mu}_{n|X}$ is supervised ERM-learning for $\hat{\mathcal{L}}_n$ if $\hat{\mu}_{n|X} \in \underset{\nu_{|X} \in \mathcal{H}_n}{\arg\min}\, \hat{\mathcal{L}}_n(\nu_{|X}, \mathcal{X}_n)$.

Note we still have error decomposition:

$$0 \leq D_{p,X,d}(\mu_{|X} \| \hat{\mu}_{n,|X}) \leq \varepsilon_{n,\mathrm{mod}}(\mu_{|X}) + \varepsilon_{n,\mathrm{learn}} + 2\varepsilon_{n,\mathrm{sampl}}$$

$$\varepsilon_{n,\mathrm{mod.}}(\mu) = \inf_{\nu_{|X} \in \mathcal{H}_n} D_{p,X,d}(\mu_{|X} \| \nu_{|X});$$

$$\varepsilon_{n,\mathrm{learn}} = c_n \left(\hat{\mathcal{L}}_n(\hat{\mu}_{n|X}, \mathcal{X}_n) - \inf_{\nu_{|X} \in \mathcal{H}_n} \hat{\mathcal{L}}_n(\nu_{|X}, \mathcal{X}_n) \right);$$

$$\varepsilon_{n,\mathrm{sampl.}} = \sup_{\nu_{|X} \in \mathcal{H}_n} \left| D(\mu_{|X} \| \nu_{|X}) - \left(c_n \hat{\mathcal{L}}_n(\nu_{|X}, \mathcal{X}_n) + h_n(\mu) \right) \right|.$$

__Cor.__ For $\mathcal{J} \subset \mathcal{H}$, $|\mathcal{H}| < \infty$. Then: $\mathcal{J}$ is PAC-learnable.

__Cor.__ $\mathcal{H}$ is cpt. $\mathcal{J} \subset \mathcal{H}$. For $\hat{\mathcal{L}}_n(V_{1X}, \mathcal{X}_n) = \sum_1^n \ell(z_j | V_{1X})$ with $|\ell(z|V_{1X})| \le k(z) \in L^1(\mathbb{R}^{s+h}, \mu)$, $\forall \mu = \mu_{1X} P_X$, $\mu_{1X} \in \mathcal{H}$. If $V_{1X} \in \mathcal{H} \mapsto \ell(z|V_{1X})$ is $D_{P,X}(\cdot \| \cdot)$-conti.

Then $\mathcal{J}$ is learned by ERM learner $\hat{\mu}_n$.

__Thm.__ $\mathcal{J}, \mathcal{H} \subseteq \mathcal{K}_1^T(\mathbb{R}^h, \mathcal{B}_{\mathbb{R}^h}) \cap \{ dV_{1X}(y) = f_{V_{1X}}(y|x) dy \}$

If $\ell(z | V_{1X}) \stackrel{\Delta}{=} -\log f_{V_{1X}}(y|x) \in L^1(\mathbb{R}^{s+h}, \mu)$ $\forall \mu = \mu_{1X} P_X$. $\forall \mu_{1X} \in \mathcal{J}$. $\forall V_{1X} \in \mathcal{H}$. And let:

$\hat{\mathcal{L}}_n(V_{1X}, \mathcal{X}_n) = \sum_1^n \ell(z_j | V_{1X})$. Then:

i) $\hat{\mathcal{L}}_n$ is unbiased EMF with $c_n = n^{-1}$ and $h_n(\mu) = h(\mu) = \int_{\mathbb{R}^{s+h}} \log f_{\mu_{1X=x}}(y|x) \, d\mu_{1X=x}(y) \, dP_X(x)$

ii) If $X \sim f_X(x) dx$. $f_V(z) \stackrel{\Delta}{=} f_{V|X}(y|x) f_X(x)$

$\bar{\mathcal{J}} \stackrel{\Delta}{=} \{ \mu = \mu_{1X} P_X \mid \mu_{1X} \in \mathcal{H} \}$.

$\bar{\mathcal{H}} \stackrel{\Delta}{=} \{ V = V_{1X} P_X \mid V_{1X} \in \mathcal{J} \}$. Then:

$\hat{\mu}_n$ is ERM learner w.r.t. d_{KL} (i.e. $\hat{\mu}_n \in \arg\min_{V \in \bar{\mathcal{H}}} - \sum_1^n \log f_V(z_j)$) $\Longleftrightarrow$ $\hat{\mu}_{n|X}$ r.c.p. of

$\hat{\mu}_n$ w.r.t $\sigma(X)$ is ERM learner for $\hat{I}_n$

<u>Pf:</u> i) $E_Z((L_n \hat{I}_n(V_{1X}, X_n) + h_n(\mu)) \overset{\text{Lem.}}{=}$

$- \int \log(f_{V|X}(y|x)/f_{\mu|X}(y|x))\, d\mu_{X=x}(y)\, dP_X(x)$

$= \int d_{KL}(\mu_{1X=x} \| V_{1X=x})\, dP_X(x) = D_{1,X,KL}(\mu_{1X} \| V_{1X})$

ii) Note $f_V(z) = f_{V|X}(y|X) f_X(x)$

$- \tilde{\sum_i} (\log f_V(z_j)) = \hat{I}_n(V_{1X}, X_n) - \sum_i \log f_X(x_j)$

$\Rightarrow$ The 2^{nd} term doesn't depend on V_{1X}

with last Lem. of (2). and def of $\bar{\pi}$. We obtain the corresponding.

<u>Rmk:</u> We can see supervised learning is as unsupervised learning with part of data.

(4) <u>Common Principles for models:</u>

For $(P(y|x))$ Markov kernel. Next. we want to know what data type of y:

a) continuous b) discrete: $\mathbb{N}$. $\{1,2,\cdots Q\}$. $\{0,1\}$ ...

$p(y|X,\theta)$ should be constructed to give all cond. proba. and label space of y.

e.g., i) (Binary labels)

$B(p)$ = Bernoulli dist. on $\{0,1\}$. i.e.

$B(p)\{1\} = p = 1 - B(p)\{0\}$.

Note all dist. on $\{0,1\}$ can be repre in $\{B(p(x)) \mid p(x)$ is measurable$\}$ all Markov kernels over $\mathbb{R}^d$ and label space $= \{0,1\}$.

ii) (Conti. dist.: classical regression)

$y \in \mathbb{R}^1$. $p(y|x) = (\sqrt{2\pi}\,\sigma)^{-1} \exp(-\frac{1}{2}(\frac{y - m(x)}{\sigma})^2)$

RMK: If we use it to model all conti. dist. $\Rightarrow$ error will be from:

a) Model error (may not Gaussian)

b) Deviation of $m(x)$

Steps:

1) identify the data structure of labels: i.e. conti. or discrete ?

2) choose a parametric family of dist. $\{M_\xi : \xi \in \Xi,\ \xi \mapsto M_\xi(B)$ is measurable. $\forall B \in A\}$.

3) choose a parametric class of functions $\{m_\theta :$

$\theta \in \mathbb{R}^d \mapsto \Xi \ni m_\theta$ is measurable$\}$. Derive the hypothesis space $\mathcal{H} = \{\mu_{m_\theta} : \theta \in \mathbb{R}\}$. $\mu_{|X=x}(B) := \mu_{m_\theta}(B)$ is Markov kernel$\}$.

4) Choose a divergence $D(\cdot\|\cdot)$ (may D_{KL}) and corresponding ERF $\hat{\mathcal{L}}_n$ (may neg. log likelihood)

5) Train with ERM.

① <u>Binary classification:</u> (logistic regression)

label space $= \{0,1\}$. $\sigma(z) = e^z/(1+e^z)$.

$m_\theta(x) = \bar{\theta}^T x + \theta_0$. $x \in \mathbb{R}^d$. $\theta = (\theta_0, \bar{\theta}) \in \mathbb{R}^{d+1}$.

(θ_0 is bias and $\bar{\theta}$ is weight.)

Set $B(p(x))\{1\} = p(x) = \sigma(m_\theta(x))$ and use Neg. log likelihood:

$$\hat{\mathcal{L}}_n((y_j, x_j)_1^n) = -\sum_1^n y_j \log p(x_j) + (1-y_j)\log(1-p(x_j))$$

$\Rightarrow$ Choose $y_j = \arg\max_{y \in \{0,1\}} B(p(x))\{y\}$

Or we can use other functions:

$m_\theta(\lambda) = \phi_{\theta_L}^{(L)} \circ \cdots \phi_{\theta_1}^{(1)}$. $\{\phi_{\theta_k}^{(k)}\}$ is layers. $\theta = (\theta_1, \cdots \theta_L)$. L is depth. $\theta_j = (\bar{\theta}_j, \theta_j^0)$. s.t. $\bar{\theta}_j \in Mat_{d_{j-1} \times d_j}(\mathbb{R})$. $\theta_j^0 \in \mathbb{R}^{d_j}$.

We set $\phi^{(j)}_{\theta_j}(x^{(j-1)}) = \chi(\theta_j^T x^{(j-1)} + \theta_j^0) \in \mathbb{R}^{d_j}$

where $\chi : \mathbb{R}^{d_j} \to \mathbb{R}^{d_j}$. $\chi(z) \triangleq (\chi(z_1) \cdots \chi(z_{d_j}))$

for $z \in \mathbb{R}^{d_j}$. $\forall j$. activation function

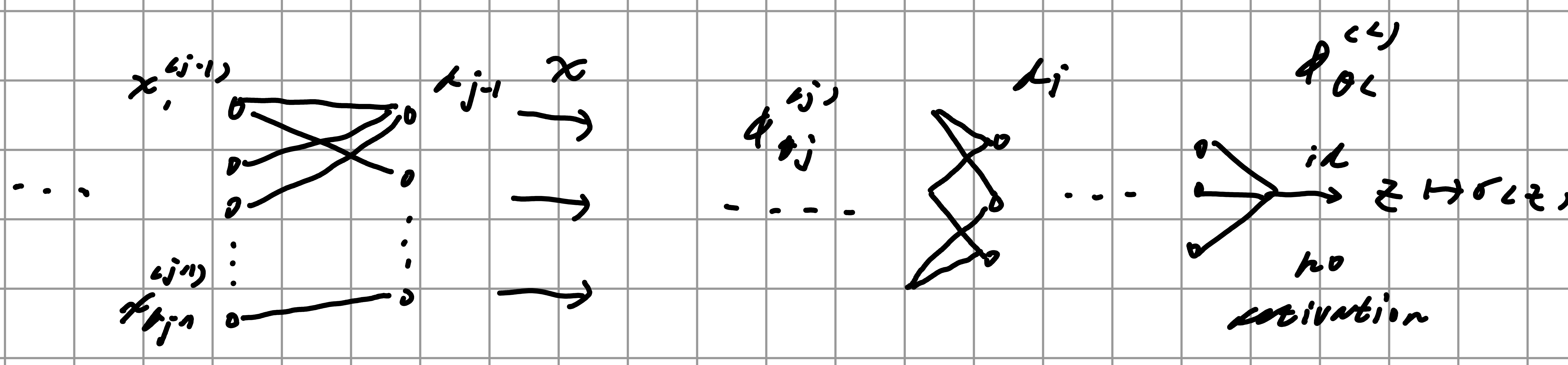
e.g. $\chi(z_j) = e^{z_j}/(1+e^{z_j})$ sigmoid function

and $\mathrm{ReLU}(\xi) = \xi \vee 0$.

Rmk: i) We don't use polynomials generally because of compu. ineff. and hard to optimize. ($|\nabla(p'(x))|$ is large)

ii) $\max\{\ ,\ \}$ isn't differentiable. So: optimal can't be trained to get

Fully Connected Neural Network (FCNN):



② Multiclass Classification:

$y \in \mathcal{C} = \{1, 2, \cdots, q\}$. Set $\mathrm{softmax}(\xi)_y =$

$$e^{\xi_y} / \sum_{y' \in \mathcal{C}} e^{\xi_{y'}}. \quad \xi \in \mathbb{R}^q.$$

Rmk: Note $\xi + c\mathbb{1}$ still correspd the same $c\, p(y|\xi)$. So it's not identificable. Hence we generally force $\xi_1 = 0$.

Next, we do a logistic regression: $\theta = (\bar{\theta}, \theta_0)$

$\bar{\theta} \in \mathrm{Mat}_{d \times 2-1}(\mathbb{R}^1)$. $\theta_0 \in \mathbb{R}^{2-1}$. $m_\theta(x) = \bar{\theta}^T x + \theta_0$.

$\mu^{\theta}_{|X=x}(\{y\}) = P_\theta(y|x) := \mathrm{softmax}(0, m_\theta(x))_y$

So: $\sum_y P_\theta(y|x) = 1 \Rightarrow$ it's dist. on $\mathcal{L}$.

And we get Markov kernel $\mu_{|x}(\cdot)$.

Rmk: All dist. on $\mathcal{L}$ can be represented by one ξ.

$$\tilde{L}_n((y_j, x_j)) = -\sum_1^n \log(\mathrm{softmax}(0, \bar{\theta} x_j + \theta_0)_{y_j})$$

$$= -\sum_{j=1}^{n} \sum_{i=1}^{2} \delta_{i, y_j} \log \square.$$

③ <u>Count regression</u>:

$\mathrm{Poi}(\xi)\{y\} := e^{-\xi} \xi^y / y!$. $y \in \mathcal{L} = \mathbb{N}$. Set:

$$m_\theta(x) = \begin{cases} \theta^T x + \theta_0 & \text{or} \\ \varphi_\theta(x) & (\mathrm{FCNN}) \end{cases}$$

and Markov kernel $\mu^{\theta}_{|X=x}\{y\} = p_\theta(y|x) = e^{-m_\theta(x)} m_\theta(x)^y / y!$

$\Rightarrow \hat{\mathcal{L}}((y_j, x_j), \theta) = \sum_i -m_\theta(\{x_j\}) + y_j \log m_\theta(x_j) - \log y_j!$.

④ <u>Regression:</u>

Markov kernel is $(\sqrt{2\pi}\sigma)^{-1} e^{-\frac{1}{2}\left(\frac{y - m_\theta(x)}{\sigma}\right)^2} dy$

$$\hat{\mathcal{L}}_n((y_j, x_j), \theta, \sigma) = -\sum -\frac{1}{2}\left(\frac{y_j - m_\theta(x_j)}{\sigma}\right)^2 - \log\sqrt{2\pi}\sigma$$

$$= (2\sigma^2)^{-1} \sum (y_j - m_\theta(x_j))^2 + n\log\sqrt{2\pi}\sigma$$

$$= \frac{n}{2\sigma^2} \mathcal{L}_n^{MSE}(\theta) + n\log\sqrt{2\pi}\sigma.$$

<u>Rmk:</u> For $m_\theta(X) = \theta X + \theta_0$. We call it ordinary least square (OLS) linear regression: if $\theta_0 = 0$, then $\theta^* = (X^T X)^{-1} X^T Y$ is the MLE if $|X^T X| \neq 0$, where $X = (X_1 \cdots X_n)$, $Y = (y_1 \cdots y_n)$. Also: $\sigma_*^2 = \mathcal{L}_n^{MSE}(\theta^*)$ is optimal.