Empirical Risk Min.

$\mathcal{L} : \mathcal{H} \ni \nu \mapsto d(\mu \| \nu)$ is loss func. and $\mathcal{H}$ is set of all prim. candidates for learning

<u>Def</u>: i) $\mathcal{H}$ is called hypothesis space.

<u>Rmk</u>: Choice of $\mathcal{H}$ is related to data

If we want to learn the dist. μ by choosing ν from $\mathcal{H}$. The best choice is $\nu \in \underset{\nu \in \mathcal{H}}{\arg\min}\, \mathcal{L}(\nu)$. But often we don't know real dist. μ.

So we need $\hat{\mathcal{L}}_n(\cdot)$ empirical risk func. computed from data $\mathcal{X} = (X_1 \cdots X_n)$. And it should be good approxi. for $\mathcal{L}(\cdot)$.

<u>Rmk</u>: i) $\mathcal{L}(\nu)$ measure the degree of failure from choice $\nu \in \mathcal{H}$.

ii) For $\mathcal{H}$ cpt. $\nu \in \mathcal{H} \mapsto d(\mu \| \nu)$ is conti. $\Rightarrow \underset{\mathcal{H}}{\arg\min}\, \mathcal{L}(\nu) \neq \emptyset$.

For d is div. $\Rightarrow \underset{\mathcal{H}}{\arg\min}\, \mathcal{L}(\nu)$

is singleton: $d(\mu \| \nu) = d(\tilde{\mu} \| \nu) + d(\mu \| \tilde{\mu}) \Rightarrow d(\mu \| \tilde{\mu}) = 0. \Rightarrow \mu = \tilde{\mu}$

ii) Empirical risk func. $\hat{\mathcal{L}} := \{\hat{\mathcal{L}}_n\}$ for given div. d is family of map $\hat{\mathcal{L}}_n : \mathcal{H}_n \times \mathbb{R}^{ln} \to \mathbb{R}'$. st.

a) $\mathbb{R}^{ln} \ni X_n \mapsto \hat{\mathcal{L}}_n(\nu, X_n)$ is measurable for $\forall \nu \in \operatorname{argmin} \hat{\mathcal{L}}_n(\nu, X_n)$

b) $\exists c_n > 0$ and $h_n : \mathcal{J} \to \mathbb{R}'$. st. $\exists (n_k)$ for $\forall \nu \in \mathcal{H}_{n_k}$, $\forall k$. $\forall \mu \in \mathcal{J}$. we have $c_{n_k} \hat{\mathcal{L}}_{n_k}(\nu, X_{n_k}) + h_{n_k}(\mu) \xrightarrow{pr} d(\mu \| \nu)$ provided $X_k \overset{i.i.d.}{\sim} \mu$.

Rmk: If $\mathcal{H}_n \uparrow$. Then: we can replace (n_k) by (n) and consider $n \to \infty$.

iii) ERF is unbiased if $\forall \mu \in \mathcal{J}$. $\nu \in \mathcal{H}_n$ $X_j \overset{i.i.d.}{\sim} \mu$. $\mathbb{E}_\mu(c_n \hat{\mathcal{L}}_n(\nu, X_n) + h_n(\mu)) = d(\mu \| \nu)$

iv) SLA $\hat{\mu}_n(X_n)$ is called empirical risk minimizer (ERM)-learner if:

$\hat{\mu}_n(X_n) \in \underset{\mathcal{H}_n}{\operatorname{argmin}} \hat{\mathcal{L}}_n(\nu, X_n)$.

and the process of minimizing $\hat{\mathcal{L}}_n(\nu, X_n)$

is called training of ERM-learner.

Next we omit n and focus on $\mathcal{H} = \mathcal{H}_n$.

We want to define ERM-learner w.r.t. the i.i.d data model (similar for DTMC)

(1) <u>max. Likelihood estimate as ERM:</u>

<u>Def</u>: $\nu \in \mathcal{M}_1^+(\mathbb{R}^d)$ p.m. $X_n = (X_1 \cdots X_n)$ is n-sample of i.i.d r.v. X_k with value in $\mathbb{R}^d$.

i) $\mathcal{H} \subset \mathcal{M}_1^+(\mathbb{R}^d)$ hypothesis space st. its elements have discrete or conti. density.

ii) $$\mathcal{L}_n(X_n \mid \nu) = \begin{cases} \prod_i^n \nu(\{X_i\}), & \nu \text{ is discrete} \\ \prod_i^n f_\nu(X_i), & \nu(dx) = f_\nu(x)dx. \end{cases}$$

for $\nu \in \mathcal{H}$.

Recall MLE is SLA st. it fulfills:

$\hat{\mu}_n(X_n) \in \arg\max \{\mathcal{L}_n(X_n \mid \nu) \mid \nu \in \mathcal{H}\}$.

<u>Rmk</u>: We want to minimize $-\log \mathcal{L}_n(X_n \mid \nu)$

Next $\log f_\nu(x_i)$ & $\log \nu\{x_i\} \stackrel{\Delta}{=} \ell(x_i \mid \nu)$.

Lem. $X_j \overset{i.i.d}{\sim} \mu \in \mathcal{T} \subset \mathcal{M}_1^+(\mathcal{X}^\wedge)$. $\mathcal{X}_n$ is n-observation

Assume $\widetilde{\mathcal{H}} \ll \mathcal{T} \subseteq$ p.m. with conti. density

i.e. $\nu(x) = f_\nu(x)dx$ & $\ell(x, \nu) = \log f_\nu(x) \in L^1(\mu)$

for $\forall \nu \in \widetilde{\mathcal{H}}$ and $\mu \in \mathcal{T}$.

Then: $\hat{\mathcal{L}}_n(\nu, \mathcal{X}_n) = -\log \mathcal{L}(\mathcal{X}_n, \nu) = -\sum_1^n \ell(x_j|\nu)$

is unbiased ERF w.r.t. d_{KL} with $C_n = \frac{1}{n}$

and $h_n(\mu) = \mathbb{E}_\mu(\ell(X|\mu))$.

<u>RMK</u>: It's similar to prove for discrete case.

<u>Pf</u>: Unbiased is from i.i.d. data $\{X_j\}_1^n$

And by SLLN: $C_n \hat{\mathcal{L}}_n(\nu, \mathcal{X}_n) =$

$-\frac{1}{n}\sum_1^n \ell(X_j|\nu) \longrightarrow -\mathbb{E}_\mu(\ell(X|\nu))$

$\Rightarrow C_n \hat{\mathcal{L}}_n(\nu, \mathcal{X}_n) + h_n(\mu) \longrightarrow -\mathbb{E}_\mu(\log \frac{d\nu}{d\mu})$

(2) <u>Error Decomp.</u>:

<u>Thm.</u> For i.i.d model $X_k \sim \mu$ and d divergence

$\mathcal{L}(\mu\|\nu) = k(\mu\|\nu)$. $\mathcal{H}_n$ is hypo space and

$\{\hat{\mathcal{L}}_n\}$ is unbiased ERF w.r.t. $\mathcal{L}$ with C_n, h_n

If $\hat{\mu}_n$ is a SLA. Then we have:

i) $0 \le L(\mu \| \hat{\mu}_n) \le \Sigma_{n,\mathrm{mod}}(\mu) + \Sigma_{n,\mathrm{learn}} + 2\,\Sigma_{n,\mathrm{sample}}(\mu)$

Where $\Sigma_{n,\mathrm{mod}}(\mu) = \inf_{\mathcal{H}_n} L(\mu \| \nu)$.

$\Sigma_{n,\mathrm{learn}} = C_n\left(\hat{\mathcal{L}}_n(\hat{\mu}_n, X_n) - \inf_{\mathcal{H}_n} \hat{\mathcal{L}}_n(\nu, X_n)\right)$

$\Sigma_{n,\mathrm{sample}}(\mu) = \sup_{\mathcal{H}_n} \left| L(\mu \| \nu) - \left(C_n \hat{\mathcal{L}}_n(\nu, X_n) + h_n(\mu)\right)\right|$.

ii) For ERM-learner $\hat{\mu}_n$, we have:

$L(\mu \| \hat{\mu}_n) \le \Sigma_{n,\mathrm{mod}}(\mu) + 2\,\Sigma_{n,\mathrm{sample}}(\mu)$.

iii) In addition with ii), for $\mu \in \mathcal{T} \subset \mathcal{H}_n$,

we have: $L(\mu \| \hat{\mu}_n) \le 2\,\Sigma_{n,\mathrm{sample}}(\mu)$.

($\mathcal{T}$ is set of "true" measures to be learned)

<u>Pf</u>: For $\nu \in \mathcal{H}_n$. Note that

$$L(\mu \| \hat{\mu}_n) \overset{(*)}{=} L(\mu \| \nu) + \left[C_n \hat{\mathcal{L}}_n(\hat{\mu}_n) + h_n(\mu) - \left(C_n \hat{\mathcal{L}}_n(\nu) + h_n(\mu)\right)\right] + \left\{ \left[L(\mu \| \hat{\mu}_n) - \left(C_n \hat{\mathcal{L}}_n(\hat{\mu}_n) + h_n(\mu)\right)\right] + \left[\left(C_n \hat{\mathcal{L}}_n(\nu) + h_n(\mu) - L(\mu \| \nu)\right)\right]\right\}.$$

$$\le L(\mu \| \nu) + C_n\left(\hat{\mathcal{L}}_n(\hat{\mu}_n) - \underbrace{\hat{\mathcal{L}}_n(\nu)}_{\ge \inf_{\mathcal{H}_n} \hat{\mathcal{L}}_n(\nu)}\right) + 2 \cdot \sup_{\nu' \in \mathcal{H}_n} \left| L(\mu \| \nu') - \left(C_n \hat{\mathcal{L}}_n(\nu') + h_n(\mu)\right)\right|$$

Take $\inf\{\cdots \mid \nu \in \mathcal{H}_n\}$ on RHS.

Rmk: i) If $\exists V_n^* \in \mathcal{H}_n$. s.t. $L(\mu \| V_n^*) = \inf_{\mathcal{H}_n} L(\mu \| v)$

$\Rightarrow$ We can modify above:

$$0 \le L(\mu \| \hat{\mu}_n)$$

$$\le L(\mu \| v) + C_n(\hat{\mathcal{I}}_n(\hat{\mu}_n) - \hat{\mathcal{I}}_n(v))$$

$$+ [(C_n \hat{\mathcal{I}}_n(v) + h_n(\mu)) - L(\mu \| v)]$$

$$+ [L(\mu \| \hat{\mu}_n) - C_n \hat{\mathcal{I}}_n(\hat{\mu}_n) - h_n(\mu)]$$

$$\le \square_{(v)} + \sup_{\mathcal{H}_n} | L(\mu \| v) - (C_n \hat{\mathcal{I}}_n(v) + h_n(\mu)) |$$

And we take $v = V_n^*$ on RHS. $\Rightarrow$ We have a better estimate. Note that one term $\xrightarrow{pr} 0$ by def of ERFs.

We only need to control $\sup \square$

ii) For $\{\hat{\mathcal{I}}_n\}$ unbiased and $\hat{\mu}_n$ ERM. in case i). We take $\mathbb{E}(\cdot)$ on the result i): Since $\hat{\mathcal{I}}_n(\hat{\mu}_n) \le \hat{\mathcal{I}}_n(V_n^*)$

So $\mathbb{E}(L(\mu \| \hat{\mu}_n)) \overset{\text{unbiased}}{\le} L(\mu \| V_n^*) +$

$\mathbb{E}(\sup_{v \in \mathcal{H}_n} | L(\mu \| v) - (C_n \hat{\mathcal{I}}_n(v) + h_n(\mu)) |)$

(By Chebyshev inequi. We can estimate $\mathbb{P}(L(\mu \| \hat{\mu}_n) > \varepsilon)$ or $\mathbb{P}(L(\mu \| \hat{\mu}_n) - L(\mu \| V_n^*) > \varepsilon)$)

$\Rightarrow$ If $\mathcal{H}_n \uparrow$. To get learnability of μ:

a) $\mu \in \overline{\mathcal{H}} := \overline{\bigcup_n \mathcal{H}_n}^{L}$

b) $\sup_{\mathcal{H}_n} | L(\mu \| \nu) - (C_n \hat{\mathcal{L}}_n(\nu) + h_n(\mu)) | \to 0$

(Note for pointwise ν. b) holds by the def. of EMF $\{\hat{\mathcal{L}}_n\}$)

if $\mu \in \mathcal{T} \subset \mathcal{H}$. then $L(\mu \| \hat{\nu}_n) \to 0$. and we only need cond. b).

iii) Note $\hat{\mu}_n = \hat{\mu}_n(X_n)$ depend on X_n. So:

$\mathbb{E}(C_n \hat{\mathcal{L}}_n(\hat{\mu}_n) + h(\mu)) \neq L(\mu \| \hat{\mu}_n)$ in general. $\Rightarrow$ Take expectation on (*) in proof may not give better estimate.

iv) When capacity $\mathcal{H} \uparrow$ $\Rightarrow$ $\Sigma_{n,\text{model}}(\mu) \downarrow$ but $\Sigma_{n,\text{sample}}(\mu)$ (and $\Sigma_{n,\text{learn}}$) $\uparrow$.

v) $L(\mu \| \hat{\mu}_n) - (C_n \hat{\mathcal{L}}_n(\hat{\mu}_n) + h_n(\mu))$ can be large when $\hat{\mathcal{L}}_n(\hat{\mu}_n) \downarrow$ by increasing the data. It's known as overfitting

$\Rightarrow$ To decouple it. we take suprema so let $\hat{\mu}_n$ disappear.

Cor. Under cond. of Thm iii) above with $\mathcal{H}$ finite. $\Rightarrow \forall\, \mathcal{J} \subseteq \mathcal{H}$ is PAC-learnable.

Pf: $\mathbb{P}(L(\mu \| \hat{\mu}_n) > \varepsilon) \overset{\text{decomp}}{\leq} \mathbb{P}(2 \sup_{\mathcal{H}} |\square| > \varepsilon)$

$= \mathbb{P}(\bigcup_{\mathcal{H}} \{ |L(\mu \| \nu) - (C_n \hat{L}_n(\nu) + f_n(\mu))| > \frac{\varepsilon}{2} \})$

$\leq \sum_{\mathcal{H}} \mathbb{P}(|\cdots| > \frac{\varepsilon}{2}) \to 0$

So we can define $N(\varepsilon, \delta) = \max$

$\{ N_{\mu,\nu}(\frac{\varepsilon}{2}, \frac{\delta}{|\mathcal{H}|}) \mid \mu \in \mathcal{J},\ \nu \in \mathcal{H} \}$.

Rmk: Define $\mathcal{H}$ is agnostically learnable if $L(\mu \| \hat{\mu}_n) - \Sigma_{n,\min}(\mu) \to 0$ in pr.

(When $\Sigma_{n,\min}(\mu) \not\to 0$ then we put it on LHS. And PAC case is similar)

So if $\mathcal{J} \not\subseteq \mathcal{H}$. $\mathcal{J}$ is still agnostically PAC-learnable.

(3) Cpt hypo space:

$|\mathcal{H}| < \infty$ is too strong and restrictive in last cor. So we want to consider infinite hypo. space.

eg. Consider parametric space $\Theta = [\mu_-, \mu_+] \times [\sigma_0^2, \sigma_1^2]$. $\mu_\theta = N(\mu, \sigma^2)$ for $\theta = (\mu, \sigma^2)$

$d(\mu_\theta, \mu_{\theta'}) = \|\theta - \theta'\|_2$. It's infinitely uncount.

Next, we assume ERM-learner (so $\mathcal{E}_{n,\text{learn}} = 0$) and $\mathcal{T} \subseteq \mathcal{H}$ ($\mathcal{E}_{n,\text{model}} = 0$). Thus, we only need

$\mathcal{E}_{n,\text{sample}} = \sup_{\mathcal{H}} | C_n \hat{L}_n(v) + h(\mu) - L(\mu \| v) | \to 0$ in pr.

for $\mu \in \mathcal{T}$ and $X_k \overset{\text{i.i.d.}}{\sim} \mu$.

Consider unbiased EMF has form: $C_n \hat{L}_n(v) = \frac{1}{n} \sum_1^n \ell(X_i | v)$ if $v \in \mathcal{H}$ is conti. distributed

$dv(x) = f(x|v)dx$

Note that there're two questions:

1°) Whether supremum is measurable. i.e. Is $\mathcal{E}_{n,\text{sample}}$ a measurable r.v. ?

It's true if $\mathcal{H}$ is d-separable, then $\exists \mathcal{H}_0$ dense countable & $v \mapsto \ell(X|v)$ conti. $\Rightarrow$

$\mathcal{E}_{n,\text{sampl.}} = \sup_{\mathcal{H}_0} | \square |$.

2°) Show $\mathcal{E}_{n,\text{sample}} \xrightarrow{pr} 0$

We require some uniform LLN. It can

be achieved by assuming $\mathcal{H}$ is cpt.

$\Rightarrow$ We can consider $|c_n \hat{I}_n(\mu) + h(\mu) - L(\mu\|\nu)| \xrightarrow{pr} 0$ in every small ball by conti.: $\nu \mapsto l(x|\nu)$

<u>Thm.</u> (Uniform LLN)

$\mathcal{J} \subset \mathcal{H} \subset M_1^+(\mathbb{R}^1)$. $\mathcal{H}$ is λ-cpt. For $X_k \overset{i.i.d}{\sim} \mu \in \mathcal{J}$. i.i.d data model. If $\nu \mapsto l(x|\nu)$ is conti. and $k(x) = \sup_{\mathcal{H}} |l(x|\nu)| \in L^1(\mu)$. Then: $\Sigma_{n,\text{sample}}(\mu) \to 0$. a.s. (So the ERM learner $\hat{\mu}_n$ learns $\mu \in \mathcal{J}$)

<u>Pf:</u> We first prove:

$$\mathbb{P}\Big(\overline{\lim_{n\to\infty}} \sup_{\mathcal{H}} \frac{1}{n}\sum_1^n l(X_j|\nu) \le \sup_{\mathcal{H}} \mathbb{E}(l(X|\nu))\Big) = 1$$

Set $\varphi(x|\nu,e) = \sup_{\nu' \in \mathcal{H}:\, L(\nu'\|\nu) < e} l(x|\nu') \le k(x)$.

$\varphi(x|\nu,e) \downarrow_{e\to 0} l(x|\nu)$ by conti. of l.

With DCT: $\mathbb{E}(\varphi(x|\nu,e)) \downarrow_{e\to 0} \mathbb{E}(l(x|\nu))$

So for $\varepsilon > 0$. $\exists e_\nu$. s.t.

$\mathbb{E}(\varphi(x|\nu,e_\nu)) \le \mathbb{E}(l(x|\nu)) + \varepsilon$.

$\Rightarrow \exists \{L(\nu_i\|\nu) < e_{\nu_i}\}_{i \le r}$ covers $\mathcal{H}$.

By def, $\exists k$. s.t. $\frac{1}{n}\sum_1^n \ell(X_j|v) \le \frac{1}{n}\sum_j^n \varphi(X_j|v_k, \varepsilon v_k)$

take $\sup_k$ and then $\sup_{v\in\mathcal{V}}$:

$$\frac{1}{n}\sum_1^n \varphi(X_j|v_k, \varepsilon_k) \le \frac{1}{n}\sum_1^n \mathbb{E}(\ell(X_j|v_k)) + \varepsilon$$

$$\to \mathbb{E}(\ell(X|v_k)) + \varepsilon. \ \mathbb{P}\text{-a.s.}$$

So: $\varlimsup_{n\to\infty} \sup_{1\le k\le r} \frac{1}{n}\sum_1^n \varphi(X_j|v_k, \varepsilon_k) \le \sup_{k\le r} \mathbb{E}(\ell(X|v_k)) + \varepsilon.$

Let $\varepsilon\to 0$. We have:

$$\varlimsup_{n\to\infty} \sup_{v\in\mathcal{V}} \frac{1}{n}\sum_1^n \ell(X_j|v) \le \sup_{v\in\mathcal{V}} \mathbb{E}(\ell(X|v)). \ \mathbb{P}\text{-a.s.}$$

Next, we see $\ell(X|v)$ is conti. so that $\mathbb{E}(\ell(X|v))$ is conti. Set $\bar{\ell}(X|v) = \ell(X|v) - \mathbb{E}(\ell(X|v))$. Repeat above on $\{\bar{\ell}(X|v)\}$.

$\Rightarrow \varlimsup \sup_{v\in\mathcal{V}} \frac{1}{n}\sum_1^n \bar{\ell}(X_k|v) \le 0. \ \mathbb{P}\text{-a.s.}$

Apply on $-\bar{\ell}(X|v)$. We have $\varliminf \inf_k \square \le 0$.

Besides, we note that $\frac{1}{n}\sum_1^n \bar{\ell}(X_k|v) =$

$\frac{1}{n}\sum_1^n \ell(X_j|v) + h_n(\mu) - h_n(\mu) - \mathbb{E}(C_n\hat{L}_n(v))$

$= C_n\hat{L}_n(v) + h_n - h(\mu\|v)$.

<u>Rmk</u>: i) It's not PAC-learnable because

the regularity condition.

ii) $\mathcal{M}$ can be $\tilde{\lambda}$ - cpt. $\tilde{\lambda}$ is other metric. And assume $\nu \mapsto L(X|\nu)$ is $\tilde{\lambda}$ - conti. The THM still holds. (So we can choose weaker topo. to let $\mathcal{M}$ be cpt.)

(4) Consistent Para. MLE:

For model of parametric statistics $\{\mu_\theta\}_{\theta \in \Theta}$

Recall that $\mathcal{M} = \mathrm{Im}(\mu_n(\cdot))$ is λ_Θ - cpt $\Leftrightarrow$ $\Theta \subset \mathbb{R}^d$ is $|\cdot|$ - cpt. (if $\mu_n(\cdot)$ is injective)

Rmk: $\mathcal{M}$ will also be λ - cpt if Θ is $|\cdot|$ - cpt and $\Theta \ni \theta \mapsto \mu_\theta$ is λ - conti.

Set $\hat{\theta}_n = \arg\min\limits_{\theta \in \Theta} \hat{\mathcal{L}}_n(\mu_\theta)$. We say consistency of para. estimate $\{\hat{\theta}_n\}$ for θ_0 if $|\hat{\theta}_n - \theta_0| \xrightarrow{pr} 0$

Rmk: We have $\mu_{\hat{\theta}_n}$ is ERM w.r.t. $\hat{\mathcal{L}}_n(\cdot)$.

Thm. If para. model $\{\mu_\theta\}_\Theta \subseteq \mathcal{M}_1^+(\mathbb{R}^k)$ s.t. $\Theta \subseteq \mathbb{R}^k$ cpt. $\Theta \ni \theta \mapsto \mu_\theta$ is injective & λ - conti. Then:

$\hat{\mu}_n = \mu_{\hat{\theta}_n}$ learns $J = \mathcal{H}$ w.r.t. d $\Longleftrightarrow$

$\hat{\theta}_n$ is consistent for Θ.

__Cor.__ So if cond. of uniform LLN holds.

then : the maximal likelihood para.

estimator $\hat{\theta}_n$ is consistent

__Pf:__ By Lem. before, we have :

$d(\hat{\mu}_n, \mu) \sim d_\Theta(\mu_{\hat{\theta}_n}, \mu_{\theta_0}) = |\hat{\theta}_n - \theta_0|$.

__Rmk:__ It works for normal / exponential / Poi. / binomial / geo. dist. But uniform dist. is special since it isn't even conti. w.r.t d_{KL}.

__e.g.__ Choose $d = d_{KL}$. $\Theta = [\mu_-, \mu_+] \times [\sigma_-^2, \sigma_+^2]$.

$\theta = (\mu, \sigma) \in \Theta \mapsto \mu_\theta = e^{-\frac{1}{2}(x-\mu)^2/\sigma^2} / \sqrt{2\pi\sigma}\, dx$

is injective. And its d_{KL} - conti. is also easy to check. Also Θ is cpt.