PAC-Learning for Histogram

Note when target measures $\in \mathcal{F} \neq \mathcal{N}$. There will be a nonzero ε_{mod}. We want to increase $\mathcal{N}$ to diminish ε_{mod} but recall it will also increase ε_{sample}. So we will do this sufficiently slowly.

Next, we consider p.m. $d\mu(x) = f_\mu(x)dx$ with conti. density $f_\mu \in L^2$. And $d_2(\mu,\nu)^2 = \|f_\mu - f_\nu\|_{L^2}^2$.

(1) <u>EMF for Histogram:</u>

<u>Lem</u>: $\mathbb{R}^d$-valued r.v. $X_j \overset{i.i.d}{\sim} \mu \in \mathcal{M}_1^+(\mathbb{R}^d)$. $\mathcal{N} := \{ d\mu(x) = f_\mu(x)dx$ with conti. density $f_\mu \in L^2 \} \subset \mathcal{M}_1^+(\mathbb{R}^d)$. If $\mu, \nu \in \mathcal{N}$. Then:

$$\hat{I}_n(\nu, X_n) = \widehat{\sum_j}\left(\int f(x|\nu)^2 dx - 2f(X_j|\nu)\right)$$

is unbiased EMF w.r.t. $d_2(\cdot,\cdot)^2$ with $c_n = n^{-1}$. $h_n(\mu) = \int f^2(x|\mu)dx$.

Pf: i) $f(x|v) \in L^1(\mu)$. Since

$$\mathbb{E}_X(f(X|v)) = \int f(x|v) f(x|\mu)$$
$$\leq \frac{1}{2}\int f^2(x|v) + f^2(x|\mu)$$

2) $\mathbb{E}(c_n \hat{\mathcal{L}}_n(v) + h_n(\mu)) = \int (f(x|v) - f(x|\mu))^2$
$= L_2(\mu, v)^2$.

3) $c_n \hat{\mathcal{L}}_n(v) + h_n(\mu) \xrightarrow{a.s.} L_2(\mu, v)^2$. by LLN.

Prop. For $\mathbb{R}^d = \Sigma D_j$. decomposition. $\mathcal{H} = \{ f(x)dx \mid f(x) = \Sigma f_j I_{D_j}.\ f_j \geq 0.\ \Sigma f_j \int_{D_j} dx = 1\}$ space of histogram func. over $\{D_j\}$.

For $X_j \sim \mu \in M_1^+(\mathbb{R}^d)$. Then: $\hat{\mu}_n = \hat{f}(x)dx$ with $\hat{f}(x) = \Sigma \hat{f}_j I_{D_j}$. $\hat{f}_j(X_1, \cdots X_n) = |\{ l = 1, \cdots n \mid X_l \in D_j \}| / n \int_{D_j} dx$ is an ERM learner for $\hat{\mathcal{L}}_n$ define above.

Pf: $\hat{\mathcal{L}}_n(v) = \sum_i^n \{ \int f(x|v)^2 - 2 f(X_i|v) \}$

$$= n \int (\sum_j f_j I_{D_j})^2 - 2 \sum_j \sum_l f_j I_{D_j}(X_l)$$

$$= n \sum_j (f_j^2 \int_{D_j} dx - 2 f_j |\{\cdots\}| / n)$$

So we can choose $\hat{f_j} = n^{-1}|\{i=1,\cdots n \mid X_i \in D_j\}|$

to minimize $\hat{\mathcal{L}}_n(\nu)$.

Note $\hat{f_j} \geq 0$. $\int \hat{f_j} I_{D_j} = \sum_1^{\Delta} |\{\cdots\}|/n = 1$.

So we set $d\hat{\mu} = \sum_1^n \hat{f_j} I_{D_j} dx \in \mathcal{U}$.

$\Rightarrow \hat{\mu}$ is an ERM-learner.

(2) Quantitative Results:

Denote: $\mathcal{J}_L = \{\nu \in \mathcal{M}_1^+([0,1]^d) \mid d\nu = f_\nu dx$. where $|f_\nu(x) - f_\nu(y)| \leq L|x-y|\}$.

Rmk: i) L is to help control $\mathcal{E}_{\text{approx}}$.

ii) $\mathcal{J}_L$ is uniformly bdd. i.e. $\exists f_{max}$ s.t. $0 \leq f_\nu(x) \leq f_{max}$. $\forall \nu \in \mathcal{J}_L$.

Lem. For $\Delta^{-1} \in \mathbb{N}$. $(D_j)_1^{\Delta^{-d}}$ is subdivision of $[0,1]^d$ with length Δ. $\mathcal{U} \triangleq \{d\nu = f_\nu dx \mid f_\nu(x) = \sum_1^{\Delta^{-d}} f_j I_{D_j}$. s.t. $0 \leq f_j \leq f_{max}$. $\sum_1^{\Delta^{-d}} f_j \Delta^d = 1\}$.

Then: $\sup\limits_{\mu \in \mathcal{J}_L} \mathcal{E}_{\text{approx}} = \sup\limits_{\mu \in \mathcal{J}_L} \inf\limits_{\nu \in \mathcal{U}} d_2(\mu,\nu)^2 \leq (L\sqrt{d}\Delta)^2$.

Pf: For $\mu \in \mathcal{J}_L$. $d\mu = f_\mu dx$.

Set $f_j = \Delta^{-d} \int_{D_j} f_\mu \, dx \le f_{max}$.

And $\sum_1^{\Delta^{-d}} f_j \Delta^d = \sum_j \int_{D_j} f_\mu = \int f_\mu = 1$.

Let $d\nu = f_\nu \, dx$. $f_\nu = \sum f_j I_{D_j}$

So: $d_2(\mu, \nu)^2 = \sum \int_{D_j} (f_\mu - \Delta^{-d} \int_{D_j} f_\mu)^2$

$= \sum \int_{D_j} (\Delta^{-d} \int_{D_j} f_\mu(x) - f_\mu(x') \, dx')^2$

$\overset{Lip}{\le} \sum \int_{D_j} (L\sqrt{d}\Delta)^2 = (L\sqrt{d}\Delta)^2$.

($\sqrt{d}\Delta$ is diameter of D_j. $\forall j$).

Rmk: It controls model error.

Lem^3. (Sample error.)

Under cond. of Lem^2. Use $(\hat{\mu}_n, \mu_n(\alpha))$

$d_2(\cdot, \cdot)^2$ in Lem^1. We have:

$\mathbb{P}(\varepsilon_{n,sample} \ge \varepsilon) \le 2\Delta^{-d} \exp(-\frac{n}{2}(\frac{\varepsilon}{f_{max}})^2)$