Stat. Learning

(1) Basic limit Thms:

The i.i.d model is good because of the existence of SLLN.

Next, we consider DTMC model:

Let $S = \{c_1, \cdots c_d\}$. finite set

Lem. $e = (e_s)$ is initial dist. $P = (P_{ss'})$ is transfer matrix. $\Rightarrow$ dist. of X_t is

$$e_s^{(t)} := \mathbb{P}(X_t = s) = (e^T P^t)_s .$$

Pf: $e_s^{(t+1)} \overset{\text{induct}}{=} \sum_{s'} P_{s's} e_{s'}^{(t)} = (e^T P^t)_s .$

Def: i) discrete density $\bar{e}$ on S is called stationary state if $\bar{e}^T P = \bar{e}^T$.

ii) A DTMC is strongly mixing if $\exists$ stationary state $\bar{e}$. st. for any initial dist. e. : $\lim_{t\to\infty} e^T P^t = \bar{e}^T$.

Rmk: Stationary state for the strongly mixing DTMC is unique.

Next, we let $g : S \to \mathbb{R}^1$. And we want to estimate $\mathbb{E}_{\bar{e}}(g) := \sum_{s} g(s)\,\bar{e}(s)$.

Thm. (X_t) is strongly mixing MC on S with initial dist e. Then:

$$\frac{1}{T}\sum_{0}^{T-1} g(X_t) \xrightarrow{pr} \mathbb{E}_{\bar{e}}(g)$$

where $\bar{e}$ is its stationary state.

Pf:

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{0}^{T-1} g(X_t) - \mathbb{E}_{\bar{e}}(g)\right| \geq \epsilon\right) \leq \frac{1}{\epsilon^2}\frac{1}{T^2}\sum_{t,t'}^{T-1}\mathbb{E}\left((g(X_t)-\mathbb{E}_{\bar{e}}(g))(g(X_{t'})-\mathbb{E}_{\bar{e}}(g))\right)$$

$$\overset{A}{=} \frac{1}{T^2\epsilon^2}\sum_{t,t'\geq 0}^{T-1} A_{t,t'}.$$

$$A_{t,t'} = \sum_{s,s'}[e^{\top}P^{t}]_s\,(g(s)-\mathbb{E}_{\bar{e}}(g))\,P^{t'-t}_{s,s'}\cdot(g(s')-\mathbb{E}_{\bar{e}}(g))$$

$$= \sum \square\,(P^{t'-t}_{s,s'} - \bar{e}_{s'})\,\square + \sum \square\,\bar{e}_{s'}\underbrace{\square}_{=0} \quad \text{for } t \leq t'.$$

So: $|A_{t,t'}| \le 4\|g\|_\infty^2 |S| \max_{s,s'} |P_{s,s'}^{t'-t} - \bar{e}_{s'}| \xrightarrow{t'-t\to\infty} 0$

$\Rightarrow$ RHS $= \left(\sum_{|t'-t|\ge \log T}^{T-1} + \sum_{|t'-t|\le \log T}^{T-1} \square \right) / T^2\varepsilon^2$

$\le C \left(\sum_{|t'-t|\ge \log T}^{T-1} A_{t,t'} + T\log T \right) / T^2 \varepsilon^2$

$\le C \left(T^2 \cdot o(1) + T(\log T) \right) / T^2 \varepsilon^2 \xrightarrow{T\to\infty} 0$

__Cor.__ It can converge $\mathbb{P}$-a.s. if $A_{t,t'}$ satisfy $|A_{t,t'}| \le c^{|t-t'|}$. for $c<1$.

__Pf__: Set $Y_T = \frac{1}{T}\sum_0^{T-1} g(X_t) - E_{\bar{e}}(g)$ and

$S_T = T Y_T$.

So: $\mathbb{P}(|Y_{T^2}| \ge \varepsilon) \lesssim T^{-2} \to 0$ which follows from prop. of $A_{t,t'}$.

For $m \in [T^2, (T+1)^2)$. We have:

$\mathbb{P}(|S_m - S_{T^2}| \ge T^2 \varepsilon)$

$\lesssim T^{-4} \left(\sum_{T^2+1}^{(T+1)^2-1} m - T^2 \right)$

So: $\sum_T \mathbb{P}(|S_m - S_{T^2}| \ge T^2\varepsilon) \le \sum \frac{1}{T^2} < \infty$

__Thm.__ If $P = (P_{ss'})_{S\times S}$ satisfies $P_{ss'} > 0$. $\forall s,s'$

Then: the DTMC has a unique stationary

State $\bar{e}$ and satisfies:

$$\sup_{s,s'} |P^t_{s,s'} - \bar{e}_{s'}| \xrightarrow{t\to\infty} 0 \text{ exponentially.}$$

Pf: $V := \{ e = (e_s)_s \mid e_s \in [0,1],\ \sum_s e_s = 1 \}$.

With metric $d(e, e') = \sum_s |e_s - e'_s|$

is complete (CLS of $(\mathbb{R}^{|S|}, \|\cdot\|_1)$)

Set $T : V \to V$. $e \mapsto T(e) = (e^T P)^T$

$\varepsilon := \inf \{ P_{s,s'} \mid s, s' \in S \} > 0$.

Next, we apply Banach fixed pt:

$$d(T(e), T(e')) = \sum_{s'} \Big| \sum_s (e_s - e'_s) P_{s,s'} \Big|$$

$$= \sum_{s'} \Big| \sum_s (e_s - e'_s)(P_{s,s'} - \varepsilon) \Big|$$

$$\le \sum_{s'} \sum_s |e_s - e'_s| (P_{s,s'} - \varepsilon)$$

$$= (1 - |S|\varepsilon)\, d(e, e').$$

If $\varepsilon = 1/|S|$. Set $\bar{e} = (\frac{1}{|S|}, \cdots, \frac{1}{|S|})$.

(2) Stat. Learning Algorithm:

Next, we consider set $S \subseteq \mathbb{R}^n$. $|S| < \infty$.

Pf: 1) $d(\cdot \| \cdot)$ is divergence on $\mathcal{M}^+_1(\mathbb{R}^n)$. the

space of p.m.'s on $\mathbb{R}^n$. if $d : \mathcal{M}^+_1(\mathbb{R}^n)$

$\times \mathcal{M}_1^+(\mathbb{R}^d) \to \overline{\mathbb{R}}_{\geq 0}$. s.t. $d(\mu \| \nu) = 0 \iff \mu = \nu$

If d is true metric. We write:

$\mu(\cdot \| \cdot) = d(\cdot, \cdot)$.

<u>Rmk</u>: d takes role of measuring how close our estimate $\hat{\mu}_n$ to μ.

ii) Unsupervised Stat. learning algorithm is collection of func. $\{\hat{\mu}_n\}_n$ defined by $\hat{\mu}_n : \mathbb{R}^{d \times n} \to \mathcal{M}_1^+(\mathbb{R}^d)$. s.t. $\forall \mu \in \mathcal{M}_1^+(\mathbb{R}^d)$. $\mathbb{R}^{d \times n} \ni (X_1, \cdots X_n) =: \mathcal{X} \mapsto d(\mu \| \hat{\mu}_n(\mathcal{X}))$ is measurable from $(\mathbb{R}^{dn}, \mathcal{B}_{\mathbb{R}^{dn}})$ to $(\overline{\mathbb{R}}_{\geq 0}, \mathcal{B}_{\overline{\mathbb{R}}_{\geq 0}})$.

<u>Rmk</u>: Supervised Stat. learning is a input-output model. But there's no input (i.e. labeled training data) here.

iii) $X_j : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d})$ is data model $\mathcal{X}_n = (X_1, \cdots X_n)$ is sample of size n.

Let $\hat{\mu}_n = \hat{\mu}_n \circ \mathcal{X}_n$ random SLA. $X_j \overset{i.i.d}{\sim} \mu$.

We call μ is d-learnable for the divergence d if $d(\mu \| \hat{\mu}_n) \xrightarrow{pr} 0$.

Rmk: It's kind of weak learnable.

$\mathcal{J} \subseteq \mathcal{M}_1^+(\mathbb{R}^d)$ is called L-learnable if $\forall \mu \in \mathcal{J}$ is L-learnable.

$\mathcal{J}$ is called L-PAC-learnable (probably approxi. correct) if $\exists n(\cdot,\cdot) : (0,1)\times(0,1) \to \mathbb{N}$. s.t. $\forall \mu \in \mathcal{J}$. $X_j \sim \mu$. $\forall \varepsilon < 1, \delta > 0$ $\exists n(\varepsilon, \delta)$. We have:

$P(L(\mu \| \hat{\mu}_n) > \varepsilon) \leq \delta$. $\forall n \geq n(\varepsilon, \delta)$.

Rmk: i) PAC-learnable is stronger than L-learnable. Since $n(\varepsilon,\delta)$ won't depend on $\mu \in \mathcal{J}$.

ii) Restriction of data generating dist. on $\mathcal{J}$ represent the prior knowledge of data.

iii) Note that $\mathcal{J} \subseteq \overline{\varliminf_n \mathrm{Im}(\hat{\mu}_n)}^d$. But if $\mathcal{J} \nsubseteq F$, set $\varepsilon_{\max}(\mu) \overset{\Delta}{=} \inf_n \inf_{\nu \in \mathrm{Im}(\hat{\mu}_n)} L(\mu \| \nu)$ then, $\exists \mu \in \mathcal{J}$:

s.t. $\Sigma_{n.d}(\mu) > 0$. We can replace

$d(\mu \| \hat{\mu}_n) \to 0$ by $d(\mu \| \hat{\mu}_n) \to \Sigma_{n.d}(\mu)$

① Learning on DTMC:

Denote: i) $k_s(\mu, \nu) = \max_s |\mu(\{s\}) - \nu(\{s\})|$.

$$I_{\{\mu(s) = \nu(s) = 1\}} + I_{\{\mu \neq \nu\} \cap \{\mu(s) = \nu(s) = 1\}^c}$$

ii) Empirical measure $\hat{\mu}_{T-1} = \frac{1}{T} \sum_{0}^{T-1} \delta_{X_t}$

Lemma. $\{\hat{\mu}_T\}$ learns $\mathcal{J} = \{\mu \in \mathcal{M}_1^+(\mathbb{R}^d) \mid \mu(s) = 1\}$ w.r.t. k_s for $\{X_t\}$ of a strongly mixing DTMC with invar. measure μ

Pf: Set $\bar{e}_s = \mu(\{s\})$. $\hat{e}_{s,T} = \hat{\mu}_T(\{s\})$

$$\mathbb{P}(k_s(\bar{e}, \hat{e}_T) > \varepsilon) \leq$$

$$\sum_s \mathbb{P}(|\bar{e}_s - \hat{e}_{s,T}| > \varepsilon) \xrightarrow[\text{mixing}]{T \to \infty} 0 \quad (|S| < \infty)$$

② Learning on i.i.d model:

Pf: $\mathcal{G} \subseteq \{g : \mathbb{R}^d \to \mathbb{R}^1 \mid g \text{ is bdd. measurable}\}$

$\mathcal{G}$ - weak topo is generated by seminorms

$k_g(\mu, \nu) = |\int_{\mathbb{R}^d} g \, d\mu - g \, d\nu|$. $g \in \mathcal{G}$.

i.e. $\mu_n \xrightarrow{\mathcal{G}\text{-}w} \mu$ if $d_g(\mu, \mu_n) \to 0$. $\forall g \in \mathcal{G}$.

$\mathcal{G}$ - strong topo is generated by $\mathcal{G}$ - divergence

$d_{\mathcal{G}}(\mu \| \mu_n) := \sup_{\mathcal{G}} d_g(\mu, \mu_n) \to 0.$

<u>Rmk</u>: i) We say $\mathcal{G}$ is separating if:

$d_{\mathcal{G}}$ is a norm, i.e. $d_{\mathcal{G}}(\mu, \nu) = 0$

$\Leftrightarrow \mu = \nu$

ii) $M_1^+(\mathbb{R}^k)$ isn't LS. But it can be embedded into sign measure space which's LS. So the "norm" will make sense.

iii) The Lem. in ① satisfies Def by chosing $\mathcal{G} = \{ I_{(s)} \mid s \in S \}$.

Next, we want to investigate the empirical measure $\hat{\mu}_n := \frac{1}{n} \sum_1^n \delta_{x_j}$.

<u>Lem.</u> $\mathcal{G} := \{ g$ is measurable. $|g| \le 1 \}$. For $\mu \in \mathcal{T}$. s.t. $\mu = f\,dx$. $f \ge 0$. $f \in L^1$. Then:

$d_{\mathcal{G}}(\mu \| \hat{\mu}_n) \equiv 2$. $\forall n \in \mathbb{N}$.

Pf: Set $g_n^w(x) = \sum_1^n I_{\{X_j(w)\}}(x) - I_{\{X_j(w),\, j\le n\}^c}$

We have $\sup_{\{g_n^w\}_{w\in\Omega}} k_g(\mu, \hat{\mu}_n) = 2$

Def: $\mathcal{G}_{GC} := \{ g = I_{(-\infty, \vec{a}]},\ \vec{a} \in \mathbb{R}^k \}$. Glivenko-Cantelli divergence $k_{GC}(\mu, \nu) = k_{\mathcal{G}_{GC}}(\mu, \nu)$

$= \sup_a | F_\mu(a) - F_\nu(a) |$. F_μ is d.f. of μ.

Rmk: i) By Lem. above, we restrict $\mathcal{G}$ on a small family.

ii) k_{GC} is truly metric. Since:

$F_\mu(a) = F_\nu(a),\ \forall a \Rightarrow \mu \sim \nu$.

Thm. (Glivenko-Cantelli)

$\mathcal{J} = \mathcal{M}_1^+(\mathbb{R}^1)$ is learnable w.r.t. k_{GC} by SLA $\hat{\mu}_n := \frac{1}{n}\sum_1^n \delta_{X_j}$

Rmk: It holds for $\mathbb{R}^k$, $\forall k \ge 1$.

Pf: Set $F_X(a-) = X_* \mathbb{P}(-\infty, a)$

$F_X(a) = X_* \mathbb{P}(-\infty, a]$.

and similarly for $\hat{F}_n(a)$, $\hat{F}_n(a-)$.

Set $q_{\frac{j}{N}} = \inf\{x \mid F_X(x) \geq \frac{j}{N}\}$. $\frac{j}{N}$-quantile of $X \sim \mu$. $q_0 = 0$. $q_1 = \infty$.

Let $\Gamma_n^{(N)} := \max_{1 \leq j \leq N-1} |F_X(q_{\frac{j}{N}}) - \hat{F}_n(q_{\frac{j}{N}})| \vee$

$$|F_X(q_{\frac{j}{N}}-) - \hat{F}_n(q_{\frac{j}{N}}-)|.$$

By SLLN. $\Gamma_n^{(N)} \to 0$ $(n \to \infty)$. $\mathbb{P}$-a.s.

For $\forall x \in \mathbb{R}^1$. $\exists j$. s.t. $x \in (q_{\frac{j-1}{N}}, q_{\frac{j}{N}})$.

So: $\hat{F}_n(x) \leq \hat{F}_n(q_{\frac{j}{N}}-) \leq F_X(q_{\frac{j}{N}}-) + \Gamma_n^{(N)}$

$$\leq F_X(x) + \frac{1}{N} + \Gamma_n^{(N)}.$$

(Note: $F_X(q_{\frac{j-1}{N}}-) - F_X(q_{\frac{j}{N}}-) \leq \frac{1}{N}$).

Similar for $\hat{F}_n(x) \geq F_X(x) - \frac{1}{N} - \Gamma_n^{(N)}$.

So $d_{KS}(\hat{\mu}_n, \mu_X) \leq \frac{1}{N} + \Gamma_n^{(N)} \xrightarrow[n\to\infty]{N\to\infty} 0$. $\mathbb{P}$-a.s.

Rmk: For $\mathcal{J} = \mathcal{M}_+^1(\mathbb{R}^d)$, it's only learnable for some weak distance like d_{KS}.