15-859(B) Machine Learning Theory

Transfer Learning and Prior Estimation for VC Classes

Liu Yang

04/25/2012

Notation

- Instance space $X = R^n$
- Concept space C of classifiers h: X -> {0,1}
  - Assume C has VC dimension vc < ∞
- Data Distribution D on X
- Unknown target function h*: the true labeling function (Realizable case: h* in C)
- Assume ρ(h, g)=$P_{x\sim D}$[h(x) ≠ g(x)] for any classifiers h, g, is a **metric** on C
- Err (h) = $P_{x\sim D}$ [h(x) ≠ h*(x)]

Transfer Learning

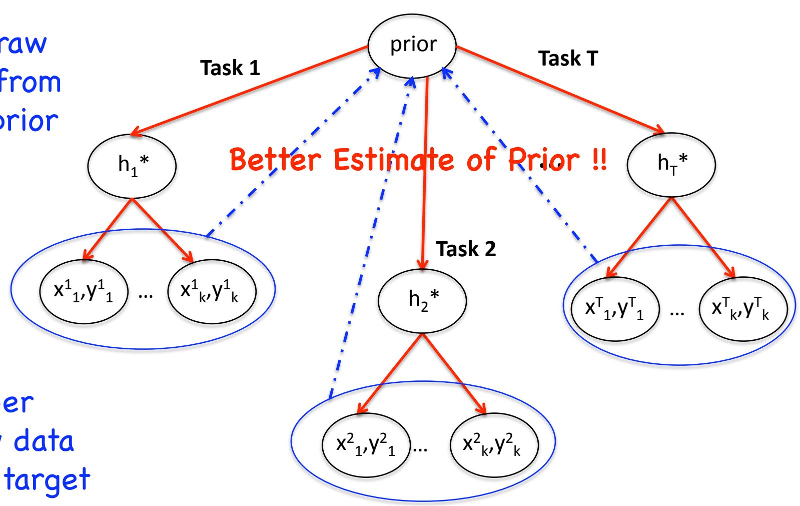
- Principle: solving a new learning problem is easier given that we've solved several already !
- How does it help?
  - New task directly ``related" to previous task [e.g., Ben-David & Schuller 03; Evgeniou, Micchelli, & Pontil 2005]
  - Previous tasks give us useful sub-concepts [e.g., Thrun 96]
  - Can gather statistical info on the variety of concepts [e.g., Baxter 97; Ando & Zhang 04]
- Example: Speech Recognition
  - After training a few times, figured out the dialects.
  - Next time, just identify the dialect.
  - Much easier than training a recognizer from scratch

 3

Model of Transfer Learning

Motivation: Learners often Not Too Altruistic

Layer 1: draw task i.i.d. from unknown prior



Layer 2: per task, draw data i.i.d. from target

 4

Identifiability of priors from joint distribs

- Let prior π be any distribution on C
  - example: (w, b) ~ multivariate normal
- Target $h^*_\pi \sim \pi$
- Data $X = (X_1, X_2, \ldots)$ i.i.d. D indep h^*_π
- $Z(\pi) = ((X_1, h^*_\pi(X_1)), (X_2, h^*_\pi(X_2)), \ldots)$.
- Let $[m] = \{1, \ldots, m\}$.
- Denote $X_I = \{X_i\}_{i \text{ in } I}$ (I : subset of natural numbers)
- $Z_I(\pi) = \{(X_i, h^*_\pi(X_i))\}_{I \text{ in } I}$

Theorem: $Z_{[VC]}(\pi_1) =_d Z_{[VC]}(\pi_2)$ iff $\pi_1 = \pi_2$.

Identifiability of priors by VC-dim joint distri.

- Threshold:

– for two points x_1, x_2, if $x_1 < x_2$, then
Pr(+,+)=Pr(+.), Pr(-,-)=Pr(.-), Pr(+,-)=0,
So Pr(-,+)=Pr(.+)-Pr(++) = Pr(.+)-Pr(+.)

– for any k > 1 points, can directly to reduce number of labels in the joint prob from k to 1

P(-----------(-+)++++++++++++++++)
= P((-+))
= P((.+)) - P((++))
= P((.+)) - P((+.))
+ P((+-)) (unrealized labeling !!)
= P((.+)) - P((+.))

- **Theorem: $Z_{[VC]}(\pi_1) =_d Z_{[VC]}(\pi_2)$ iff $\pi_1 = \pi_2$.**

  Proof Sketch
- Let $\rho_m(h,g) = 1/m \sum_{i=1}^{m} \mathbb{I}(h(X_m) \neq g(X_m))$
  Then vc < ∞ implies w.p.1 forall h, g in C with h ≠ g
  $\lim_{m \to \infty} \rho_m(h,g) = \rho(h,g) > 0$
- ρ is a metric on C by assumption,
  so w.p.1 each h in C labels ∞-seq $(X_1, X_2 \ldots)$ distinctly $(h(X_1), h(X_2), \ldots)$
- => w.p.1 conditional distribution of the label seq $Z(\pi)|X$ identifies π
  => distrib of $Z(\pi)$ identifies π
  i.e. $Z_\infty(\pi_1) =_d Z_\infty(\pi_2)$ implies $\pi_1 = \pi_2$

Identifiability of Priors from Joint Distributions

Theorem: $Z_{[\mathrm{vc}]}(\pi_1) \stackrel{d}{=} Z_{[\mathrm{vc}]}(\pi_2) \Leftrightarrow \pi_1 = \pi_2$.

Proof Sketch:

Fix any $m > \mathrm{vc}$, $x_1, \ldots, x_m \in \mathcal{X}$, $y_1, \ldots, y_m \in \{0,1\}$.

Note $\mathbb{C}$ cannot shatter $(x_1, \ldots, x_m)$.

Let $\tilde{y}_1, \ldots, \tilde{y}_m \in \{0,1\}$ be s.t. $\nexists h \in \mathbb{C}$ with $\forall i, h(x_i) = \tilde{y}_i$.

Clearly $\mathbb{P}\left(Z_{[m]}(\pi) = \{(x_i, \tilde{y}_i)\}_{i \in [m]} \middle| \mathbb{X}_{[m]} = \{x_i\}_{i \in [m]}\right) = 0$.

If $\exists k$ s.t. $y_k \neq \tilde{y}_k$, then letting $y'_i = y_i$ for $i \neq k$, and $y'_k = \tilde{y}_k$,

$$\mathbb{P}\left(Z_{[m]}(\pi) = \{(x_i, y_i)\}_{i \in [m]} \middle| \mathbb{X}_{[m]} = \{x_i\}_{i \in [m]}\right)$$

$$= \mathbb{P}\left(Z_{[m]\setminus\{k\}}(\pi) = \{(x_i, y_i)\}_{i \in [m]\setminus\{k\}} \middle| \mathbb{X}_{[m]\setminus\{k\}} = \{x_i\}_{i \in [m]\setminus\{k\}}\right)$$

$$- \mathbb{P}\left(Z_{[m]}(\pi) = \{(x_i, y'_i)\}_{i \in [m]} \middle| \mathbb{X}_{[m]} = \{x_i\}_{i \in [m]}\right).$$

lower–dim cond distrib

y' closer to ỹ

Induction: $\mathbb{P}\left(Z_{[m]}(\pi) = \cdot \middle| \mathbb{X}_{[m]}\right)$ function of $\mathbb{P}\left(Z_{[\mathrm{vc}]}(\pi) = \cdot \middle| \mathbb{X}_{[\mathrm{vc}]}\right)$.

Identifiability of Priors from Joint Distributions

Theorem: $Z_{[\mathrm{vc}]}(\pi_1)\overset{d}{=}Z_{[\mathrm{vc}]}(\pi_2) \Leftrightarrow \pi_1 = \pi_2$.

Proof Sketch:

By the above,
$Z_{[\mathrm{vc}]}(\pi_1)\overset{d}{=}Z_{[\mathrm{vc}]}(\pi_2) \Rightarrow \forall m \in \mathbb{N}, Z_{[m]}(\pi_1)\overset{d}{=}Z_{[m]}(\pi_2)$.

Classic result:
set of distribs of $Z_{[m]}(\pi) : m \in \mathbb{N}$ identify distrib of $Z(\pi)$, so
$Z_{[m]}(\pi_1)\overset{d}{=}Z_{[m]}(\pi_2), \forall m \in \mathbb{N} \Rightarrow Z(\pi_1)\overset{d}{=}Z(\pi_2)$.

Showed above that
$Z(\pi_1)\overset{d}{=}Z(\pi_2) \Rightarrow \pi_1 = \pi_2$. □

Identifiability of Priors from Joint Distributions

Theorem: $Z_{[\mathrm{vc}]}(\pi_1)\overset{d}{=}Z_{[\mathrm{vc}]}(\pi_2) \Leftrightarrow \pi_1 = \pi_2$.

Theorem: $\exists \mathcal{D}$, $\pi_1 \neq \pi_2$ s.t. $\forall m < \mathrm{vc}$, $Z_{[m]}(\pi_1)\overset{d}{=}Z_{[m]}(\pi_2)$.

Proof Sketch:
Let $(x_1, \ldots, x_{\mathrm{vc}})$ be shattered by $\mathcal{H} = \{h_1, \ldots, h_{2^{\mathrm{vc}}}\} \subseteq \mathbb{C}$.
Let $\mathcal{D}$ be uniform on $\{x_1, \ldots, x_{\mathrm{vc}}\}$,
let π_1 be uniform on $\mathcal{H}$.
Let $\mathcal{H}' = \{h'_1, \ldots, h'_{2^{\mathrm{vc}-1}}\} \subset \mathcal{H}$ shatter $(x_1, \ldots, x_{\mathrm{vc}-1})$
s.t. $h'_i(x_{\mathrm{vc}}) = \mathrm{Parity}(\{h'_i(x_1), \ldots, h'_i(x_{\mathrm{vc}-1})\})$.
Let π_2 be uniform on $\mathcal{H}'$.
Clearly $\pi_1 \neq \pi_2$.
But for $m < \mathrm{vc}$, $Z_{[m]}(\pi_1)\overset{d}{=}Z_{[m]}(\pi_2)$:
unif cond on labels given distinct $X_1, \ldots, X_m$. □

Transfer Learning Setting

- Collection Π of distribs on C. (known)
- Target distrib π^* in Π. (unknown)
- Indep target fns $h_1^*, \ldots, h_T^* \sim \pi^*$ (unknown)
- Indep i.i.d. D data sets $X^{(t)} = (X_1^{(t)}, X_2^{(t)}, \ldots)$, t in [T].
- Define $Z^{(t)} = ((X_1^{(t)}, h_t^*(X_1^{(t)})), (X_2^{(t)}, h_t^*(X_2^{(t)})), \ldots)$.
- Learning alg. "gets" $Z^{(1)}$, then produces $\hat{h}_1$, then "gets" $Z^{(2)}$, then produces $\hat{h}_2$, etc. in sequence.
- Interested in: values of $\rho(\hat{h}_t, h^*(t))$, and the number of $h^*_t(X_j^{(t)})$ value alg. needs to access.

Estimating the prior

- Principle: learning would be easier if know π^*
- Fact: π^* is identifiable by distrib of $Z_{[VC]}^{(t)}$
- Strategy: Take samples $Z_{[VC]}^{(i)}$ from past tasks 1, …, t-1, use them to estimate distrib of $Z_{[VC]}^{(i)}$, convert that into an estimate π'_{t-1} of π^*,
- Use π'_{t-1} in a prior-dependent learning alg for new task h_t^*
- Assume Π is totally bounded in total variation
- Can estimate π^* at a bounded rate:

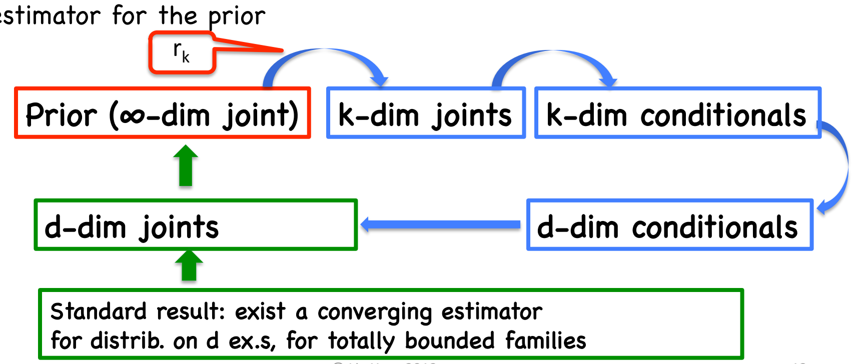
  $\|\pi^* - \pi'_t\| < \delta_t$ converges to 0 (holds whp)

Main Theorem

Theorem 1 *There exists an estimator* $\hat{\theta}_{T\theta_\star} = \hat{\theta}_T(\mathcal{Z}_{1d}(\theta_\star), \ldots, \mathcal{Z}_{Td}(\theta_\star))$, *and functions* $R : \mathbb{N}_0 \times (0,1] \to [0,\infty)$ *and* $\delta : \mathbb{N}_0 \times (0,1] \to [0,1]$, *such that for any* $\alpha > 0$, $\lim_{T\to\infty} R(T,\alpha) = \lim_{T\to\infty} \delta(T,\alpha) = 0$ *and for any* $T \in \mathbb{N}_0$ *and* $\theta_\star \in \Theta$,

$$\mathbb{P}\left(\|\pi_{\hat{\theta}_{T\theta_\star}} - \pi_{\theta_\star}\| > R(T,\alpha)\right) \le \delta(T,\alpha) \le \alpha.$$

Pf Idea: relate convergence of estimator for d-dim joint to convergence of estimator for the prior



 13

Transfer Learning

- Given a prior-dependent learning A(ε , π), with E[# labels accessed] = Λ(ε , π) and producing $\hat{h}$ with E[$\rho(\hat{h}, h^*)$] ≤ ε

For t = 1,…, T
  If $\delta_{t-1} > \varepsilon/4$,
    run prior-indep learning on $Z_{[VC/\varepsilon]}^{(t)}$ to get $\hat{h}_t$
  Else let $\pi''_t = \mathrm{argmin}_{\pi \text{ in } B(\pi'_{t-1}, \delta_{t-1})} \Lambda(\varepsilon/2, \pi)$
  and run A(ε /2, π''_t) on $Z^{(t)}$ to get $\hat{h}_t$

Theorem: Forall t, E[$\rho(\hat{h}_t, h_t^*)$] ≤ ε , and $\limsup_{T\to\infty}$ E[#labels accessed]/T ≤ Λ(ε /2, π*) + vc.

 14

Relate Prior to k-dim joint

Lemma: *There exists a sequence* $r_k = o(1)$ *such that* $\forall t, k \in \mathbb{N}, \forall \theta, \theta' \in \Theta$,

$$\|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| \le \|\pi_\theta - \pi_{\theta'}\| \le \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + r_k.$$

Proof: - The left inequality follows from, for any θ,θ' in Θ and t (natural num), $\| P_{Z_{tk}(\theta)} - P_{Z_{tk}(\theta')} \| \le \| P_{Z_t(\theta)} - P_{Z_t(\theta')} \| = \| \pi_\theta - \pi_{\theta'} \|$

- To show the right inequality: Fix θ,θ' in Θ, let γ>0,let B subseteq $(X\times\{-1, +1\})^\infty$ be a measurable set s.t.

$$\|\pi_\theta - \pi_{\theta'}\| = \|\mathbb{P}_{\mathcal{Z}_t(\theta)} - \mathbb{P}_{\mathcal{Z}_t(\theta')}\| < \mathbb{P}_{\mathcal{Z}_t(\theta)}(B) - \mathbb{P}_{\mathcal{Z}_t(\theta')}(B) + \gamma.$$

- Carathéodory's extention theorem implies there exist disjoint sets $\{A_i\}_{i \text{ in } \mathbb{N}}$ where A_i is an event for finite number of data pts, s.t. $B \subseteq \bigcup_{i\in\mathbb{N}} A_i$

$$P_{Z_t(\theta)}(B) - P_{Z_t(\theta')}(B) < \Sigma_{i \text{ in } \mathbb{N}} P_{Z_t(\theta)}(A_i) - \Sigma_{i \text{ in } \mathbb{N}} P_{Z_t(\theta')}(A_i) + \gamma$$

- Since these sums are bounded, there must exist n in $\mathbb{N}$ s.t.

$$\Sigma_{i \text{ in } \mathbb{N}} P_{Z_t(\theta)}(A_i) < \gamma + \Sigma_{i=1}^{n} P_{Z_t(\theta)}(A_i)$$

Relate Prior to k-dim joint

So that
$$\sum_{i\in\mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i) - \sum_{i\in\mathbb{N}} \mathbb{P}_{\mathcal{Z}_t(\theta')}(A_i) < \gamma + \sum_{i=1}^{n} \mathbb{P}_{\mathcal{Z}_t(\theta)}(A_i) - \sum_{i=1}^{n} \mathbb{P}_{\mathcal{Z}_t(\theta')}(A_i)$$
$$= \gamma + \mathbb{P}_{\mathcal{Z}_t(\theta)}\left(\bigcup_{i=1}^{n} A_i\right) - \mathbb{P}_{\mathcal{Z}_t(\theta')}\left(\bigcup_{i=1}^{n} A_i\right).$$

-As $\bigcup_{i=1}^{n} A_i \in \mathcal{A}$, there exists k' (natural num) & measurable A' subset of $(X \times \{-1, +1\})^k$ s.t. $\bigcup_{i=1}^{n} A_i = A' \times (\mathcal{X} \times \{-1,+1\})^\infty$

Thus
$$\mathbb{P}_{\mathcal{Z}_t(\theta)}\left(\bigcup_{i=1}^{n} A_i\right) - \mathbb{P}_{\mathcal{Z}_t(\theta')}\left(\bigcup_{i=1}^{n} A_i\right) = \mathbb{P}_{\mathcal{Z}_{tk'}(\theta)}(A') - \mathbb{P}_{\mathcal{Z}_{tk'}(\theta')}(A')$$
$$\le \|\mathbb{P}_{\mathcal{Z}_{tk'}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk'}(\theta')}\| \le \lim_{k\to\infty} \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\|.$$

In sum, $\|\pi_\theta - \pi_{\theta'}\| \le \lim_{k\to\infty} \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + 3\gamma.$

- Taking the limit as γ-> 0 implies $\|\pi_\theta - \pi_{\theta'}\| \le \lim_{k\to\infty} \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\|$

- Particularly, it implies there exists a sequence $r_k(\theta, \theta') = o(1)$ s.t.

$$\forall k \in \mathbb{N}, \|\pi_\theta - \pi_{\theta'}\| \le \|\mathbb{P}_{\mathcal{Z}_{tk}(\theta)} - \mathbb{P}_{\mathcal{Z}_{tk}(\theta')}\| + r_k(\theta, \theta').$$

QED

Relate k-dim Joint to k-dim Cond.

- Want to bound between tvd of k-dim joints
- Easier to bound diff between tvd of k-dim cond. Distri.s
- Use Jensen's ineqn to relate tvd of k-dim joint distri. to k-dim cond. distri. :

  $\| P_{Z_{tk}(\theta)} - P_{Z_{tk}(\theta')} \| \leq E\, [\| P_{Y_{tk}(\theta)\, | X_{tk}} - P_{Y_{tk}(\theta')\, | X_{tk}} \|]$

Relate k-dim Cond.to d-dim Cond.

- By def of total variation dist.

$$\|\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k} - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}\| = (1/2) \sum_{\bar{y}^k \in \{-1,+1\}^k} |\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k}(\bar{y}^k) - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}(\bar{y}^k)|,$$

- By Sauer's Lemma this is ≤ $(ek)^d \max_{\bar{y}^k \in \{-1,+1\}^k} |\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k}(\bar{y}^k) - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}(\bar{y}^k)|,$

- Notations:

  $I \subseteq \{1, \ldots, k\}$, fix $\bar{x}_I \in \mathcal{X}^{|I|}$ and $\bar{y}_I \in \{-1,+1\}^{|I|}$. Then the $\tilde{y}_I \in \{-1,+1\}^{|I|}$ for which no $h \in \mathbb{C}$ has $h(\bar{x}_I) = \tilde{y}_I$ for which $\|\bar{y}_I - \tilde{y}_I\|_1$ is minimal, has $\|\bar{y}_I - \tilde{y}_I\|_1 \leq d+1$, and for any $i \in I$ with $\bar{y}_i \neq \tilde{y}_i$, letting $\bar{y}'_j = \bar{y}_j$ for $j \in I \setminus \{i\}$ and $\bar{y}'_i = \tilde{y}_i$, we have

$$\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I) = \mathbb{P}_{\mathbb{Y}_{I\setminus\{i\}}(\theta)|\mathbb{X}_{I\setminus\{i\}}}(\bar{y}_{I\setminus\{i\}}|\bar{x}_{I\setminus\{i\}}) - \mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}'_I|\bar{x}_I),$$

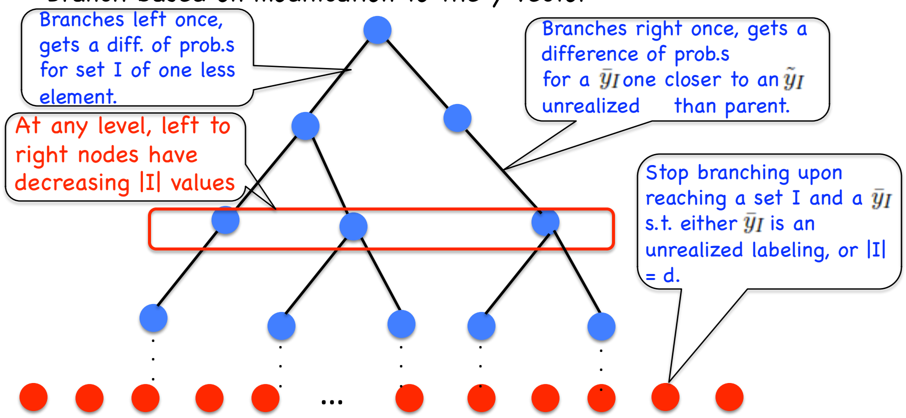
  (By P(A and B) = P(A) - P(A and not B). Two terms, one reduce dim by 1, the other brought y vector closer to the unrealizable labeling by one bit)

- Apply this to theta and theta', interested in the tvd between the cond. Prob.

$$\begin{aligned} &|\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I) - \mathbb{P}_{\mathbb{Y}_I(\theta')|\mathbb{X}_I}(\bar{y}_I|\bar{x}_I)| \\ &\leq |\mathbb{P}_{\mathbb{Y}_{I\setminus\{i\}}(\theta)|\mathbb{X}_{I\setminus\{i\}}}(\bar{y}_{I\setminus\{i\}}|\bar{x}_{I\setminus\{i\}}) - \mathbb{P}_{\mathbb{Y}_{I\setminus\{i\}}(\theta')|\mathbb{X}_{I\setminus\{i\}}}(\bar{y}_{I\setminus\{i\}}|\bar{x}_{I\setminus\{i\}})| \\ &+ |\mathbb{P}_{\mathbb{Y}_I(\theta)|\mathbb{X}_I}(\bar{y}'_I|\bar{x}_I) - \mathbb{P}_{\mathbb{Y}_I(\theta')|\mathbb{X}_I}(\bar{y}'_I|\bar{x}_I)|. \end{aligned}$$

Tree Argument: Combinatorics

- Consider these two terms inductively define a binary tree
- Branch based on modification to the y vector



- Any path can branch left ≤ k – d times (total) before reaching a set I w/ only d elements; can branch right ≤ d + 1 times in a row before reaching a $\bar{y}_I$ s.t. both prob.s zero, so the diff is zero.

 19

Tree Argument: Conclusions

- Bound original (root node) diff of prob.s by sum of the diff of prob.s for leaf nodes with |I| = d.
- Depth of any leaf node with |I| = d is at most (k – d)d.
- Maximum width of the tree is at most k – d.
- So total #leaf nodes with |I| = d is at most d (k – d)².
  - For any $\bar{y}_I \in$ {-1, +1}^{k}, $\bar{x} \in \mathcal{X}^k$

$$|\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k}(\bar{y}|\bar{x}) - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}(\bar{y}|\bar{x})|$$
$$\leq (k-d)^2 d \cdot \max_{\bar{y}^d \in \{-1,+1\}^d} \max_{D \in \{1,\ldots,k\}^d} |\mathbb{P}_{\mathbb{Y}_d(\theta)|\mathbb{X}_d}(\bar{y}^d|\bar{x}_D) - \mathbb{P}_{\mathbb{Y}_d(\theta')|\mathbb{X}_d}(\bar{y}^d|\bar{x}_D)|.$$

 20

Relate k-dim Joint to d-dim Joint

- Note $\mathbb{E}\left[\max_{\tilde{y}^d \in \{-1,+1\}^d} \max_{D \in \{1,\dots,k\}^d} |\mathbb{P}_{Y_d(\theta)|X_D}(\tilde{y}^d) - \mathbb{P}_{Y_d(\theta')|X_D}(\tilde{y}^d)|\right]$

$$\leq \sum_{\tilde{y}^d \in \{-1,+1\}^d} \sum_{D \in \{1,\dots,k\}^d} \mathbb{E}\left[|\mathbb{P}_{Y_d(\theta)|X_D}(\tilde{y}^d) - \mathbb{P}_{Y_d(\theta')|X_D}(\tilde{y}^d)|\right]$$

$$\leq (2k)^d \max_{\tilde{y}^d \in \{-1,+1\}^d} \max_{D \in \{1,\dots,k\}^d} \mathbb{E}\left[|\mathbb{P}_{Y_d(\theta)|X_D}(\tilde{y}^d) - \mathbb{P}_{Y_d(\theta')|X_D}(\tilde{y}^d)|\right]$$

- By exchangeability, the last line equals

$$(2k)^d \max_{\tilde{y}^d \in \{-1,+1\}^d} \mathbb{E}\left[|\mathbb{P}_{Y_d(\theta)|X_d}(\tilde{y}^d) - \mathbb{P}_{Y_d(\theta')|X_d}(\tilde{y}^d)|\right].$$

- Want d-dim joint instead of d-dim cond.

Claim: $\mathbb{E}\left[\left|\mathbb{P}_{Y_d(\theta)|X_d}(\tilde{y}^d|X_d) - \mathbb{P}_{Y_d(\theta')|X_d}(\tilde{y}^d|X_d)\right|\right] \leq 4\sqrt{\|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\|},$

Proof of the Claim

Proof:

Suppose $$\mathbb{E}\left[\left|\mathbb{P}_{Y_d(\theta)|X_d}(\tilde{y}^d|X_d) - \mathbb{P}_{Y_d(\theta')|X_d}(\tilde{y}^d|X_d)\right|\right] \geq \varepsilon,$$

for some $\tilde{y}^d$. Then either

$$\mathbb{P}\left(\mathbb{P}_{Y_d(\theta)|X_d}(\tilde{y}^d|X_d) - \mathbb{P}_{Y_d(\theta')|X_d}(\tilde{y}^d|X_d) \geq \varepsilon/4\right) \geq \varepsilon/4,$$

or

$$\mathbb{P}\left(\mathbb{P}_{Y_d(\theta')|X_d}(\tilde{y}^d|X_d) - \mathbb{P}_{Y_d(\theta)|X_d}(\tilde{y}^d|X_d) \geq \varepsilon/4\right) \geq \varepsilon/4.$$

For which ever is the case, let A_ε denote the corresponding measurable subset of $\mathcal{X}^d$, of probability at least $\varepsilon/4$. Then

$$\|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\| \geq \left|\mathbb{P}_{\mathcal{Z}_{td}(\theta)}(A_\varepsilon \times \{\tilde{y}^d\}) - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}(A_\varepsilon \times \{\tilde{y}^d\})\right|$$
$$\geq (\varepsilon/4)\mathbb{P}_{X_d}(A_\varepsilon) \geq \varepsilon^2/16.$$

Therefore,

$$\mathbb{E}\left[\left|\mathbb{P}_{Y_d(\theta)|X_d}(\tilde{y}^d|X_d) - \mathbb{P}_{Y_d(\theta')|X_d}(\tilde{y}^d|X_d)\right|\right] \leq 4\sqrt{\|\mathbb{P}_{\mathcal{Z}_{td}(\theta)} - \mathbb{P}_{\mathcal{Z}_{td}(\theta')}\|},$$

Reflect the Path of Proof

Earlier, $||\pi_\theta - \pi_{\theta'}|| \le ||P_{Z_{tk}(\theta)} - P_{Z_{tk}(\theta')}|| + r_k$

Just showed

$||P_{Z_t(\theta)} - P_{Z_t(\theta')}|| \le ||P_{Z_{tk}(\theta)} - P_{Z_{tk}(\theta')}|| + r_k$

Carathéodory's extention Thm in general form

(tree) $\le g_k(||P_{Z_{td}}(\theta) - P_{Z_{td}}(\theta')||) + r_k$

where $||P_{Z_{tk}(\theta)} - P_{Z_{tk}(\theta')}|| \le 4\ (2ek)^{2d+2}\sqrt{||P_{Z_{td}(\theta)} - P_{Z_{td}(\theta')}||}$

So in total

For any k in $\mathbb{N}$, $||\pi_\theta - \pi_{\theta'}|| \le 4\ (2ek)^{2d+2}\sqrt{||P_{Z_{td}(\theta)} - P_{Z_{td}(\theta')}||} + r_k$

In particular, $r_k \to 0$ as $k \to \infty$. Let $g(\varepsilon) = \min_k(4\ (2ek)^{2d+2}\sqrt{\varepsilon} + r_k)$.

Claim: $g(\varepsilon) \to 0$ as $\varepsilon \to 0$.

(Why? Let $\varepsilon_k = (r_k/(4\ (2ek)^{2d+2}))^2$. $\varepsilon_k = o(1)$. $g(\varepsilon_k) \le 4\ (2ek)^{2d+2}\sqrt{\varepsilon_k} + r_k = 2r_k$

g is monotonic in ε => $\lim_{\varepsilon \to \infty} g(\varepsilon) = \lim_{k \to \infty} g(\varepsilon_k) = \lim_{k \to \infty} 2r_k = 0$.)

 23

Distri. Estimation Rate

- The last component: rate of conv. of our estimate of $\mathbb{P}_{\mathcal{Z}_d(\theta_\star)}$
  - N(ε) is the ε -covering number $\{\mathbb{P}_{\mathcal{Z}_d(\theta)} : \theta \in \Theta\}$
  - Taking $\hat{\theta}_{T\theta_\star}$ as the minimum distance skeleton estimate of Yatracos (1985) achieves expected tvd ε from $\pi_{\theta_\star}$, for some $T = O((1/\varepsilon^2)\log N(\varepsilon/4))$.

Solving for eps in terms of T implies E[tvd of d-dim] -> 0 as T-> ∞

- Conclusion for prior estimation:
  - Pick the sequence of R_t s.t. R_t -> 0, but with $E[w_t]/R_t$ -> 0
  - Let w_t be E[tvd of d-dim]. For any t, apply Markov ineq. => $P(w_t > R_t) < E[w_t]/R_t$
  - Since E[tvd of d-dim] -> 0, Markov's ineq. => there is a bound on tvd -> 0 which holds with prob. that -> 1, as T-> ∞
  - If tvd of d-dim joints -> 0, plugging into g() (just proved), tvd of priors -> 0.
- Together we just proved the theorem

Theorem 1 *There exists an estimator* $\hat{\theta}_{T\theta_\star} = \hat{\theta}_T(\mathcal{Z}_{1d}(\theta_\star), \ldots, \mathcal{Z}_{Td}(\theta_\star))$, *and functions* $R : \mathbb{N}_0 \times (0,1] \to [0,\infty)$ *and* $\delta : \mathbb{N}_0 \times (0,1] \to [0,1]$, *such that for any* $\alpha > 0$, $\lim_{T\to\infty} R(T,\alpha) = \lim_{T\to\infty} \delta(T,\alpha) = 0$ *and for any* $T \in \mathbb{N}_0$ *and* $\theta_\star \in \Theta$,

$$\mathbb{P}\left(\|\pi_{\hat{\theta}_{T\theta_\star}} - \pi_{\theta_\star}\| > R(T,\alpha)\right) \le \delta(T,\alpha) \le \alpha.$$

 24

Rate of Conv. under Hölder-Smooth

Definition: For $L \in (0, \infty)$ and $\alpha \in (0, 1]$, a function $f : \mathbb{C} \to \mathbb{R}$ is (L, α)-Hölder smooth if
$$\forall h, g \in \mathbb{C}, |f(h) - f(g)| \le L\rho(h, g)^{\alpha}.$$

Theorem. *For Π_Θ any class of priors on $\mathbb{C}$ having (L, α)-Hölder smooth densities $\{f_\theta : \theta \in \Theta\}$, for any $T \in \mathbb{N}$, there exists an estimator $\hat{\theta}_{T\theta} = \hat{\theta}_T(\mathcal{Z}_{1d}(\theta), \ldots, \mathcal{Z}_{Td}(\theta))$ such that*
$$\sup_{\theta_\star \in \Theta} \mathbb{E}\|\pi_{\hat{\theta}_T} - \pi_{\theta_\star}\| = \tilde{O}\left(LT^{-\frac{\alpha^2}{2(d+2\alpha)(\alpha+2(d+1))}}\right).$$

$$\|\pi_\theta - \pi_{\theta'}\| \le \min_k 4\,(2ek)^{2d+2} \| P_{Z_{td}(\theta)} - P_{Z_{td}(\theta')}\|^{1/2} + r_k$$

Under Hölder-smooth
$r_k = O(L(d/k \log(k/d))^{\alpha})$

Rate of Conv. under Hölder-Smooth

Definition: For $L \in (0, \infty)$ and $\alpha \in (0, 1]$, a function $f : \mathbb{C} \to \mathbb{R}$ is (L, α)-Hölder smooth if
$$\forall h, g \in \mathbb{C}, |f(h) - f(g)| \le L\rho(h, g)^{\alpha}.$$

Theorem. *For Π_Θ any class of priors on $\mathbb{C}$ having (L, α)-Hölder smooth densities $\{f_\theta : \theta \in \Theta\}$, for any $T \in \mathbb{N}$, there exists an estimator $\hat{\theta}_{T\theta} = \hat{\theta}_T(\mathcal{Z}_{1d}(\theta), \ldots, \mathcal{Z}_{Td}(\theta))$ such that*
$$\sup_{\theta_\star \in \Theta} \mathbb{E}\|\pi_{\hat{\theta}_T} - \pi_{\theta_\star}\| = \tilde{O}\left(LT^{-\frac{\alpha^2}{2(d+2\alpha)(\alpha+2(d+1))}}\right).$$

Proof:

- By PAC bound, for any $\gamma>0$, w.p.$>1-\gamma$, a sample of $k = O((d/\gamma)\log(1/\gamma))$ partition C into regions of width $< \gamma$.
- For any $\theta \in \Theta$, π'_θ denote a (conditional on $X_1, \ldots, X_k$) distribution
  f'_θ denote the (conditional on $X_1, \ldots, X_k$) density function of π'_θ with respect to π_0
- For any $g \in \mathbb{C}$, $$f'_\theta(g) = \frac{\pi_\theta(\{h \in \mathbb{C} : \forall i \le k, h(X_i) = g(X_i)\})}{\pi_0(\{h \in \mathbb{C} : \forall i \le k, h(X_i) = g(X_i)\})}$$
  (or 0 if $\pi_0(\{h \in \mathbb{C} : \forall i \le k, h(X_i) = g(X_i)\}) = 0$).
- By smoothness, w. p. $>1-\gamma$, we have everywhere $|f_\theta(h) - f'_\theta(h)| < L\gamma^{\alpha}$.

Rate of Conv. under Hölder-Smooth

- Thus for any $\theta, \theta' \in \Theta$, w.p.>1-γ,

$$\|\pi_\theta - \pi_{\theta'}\| = (1/2)\int |f_\theta - f_{\theta'}| d\pi_0 < L\gamma^\alpha + (1/2)\int |f'_\theta - f'_{\theta'}| d\pi_0.$$

- Since the regions that define f'_θ and $f'_{\theta'}$ are the same,

$$(1/2)\int |f'_\theta - f'_{\theta'}| d\pi_0$$
$$= (1/2) \sum_{y_1,\ldots,y_k \in \{-1,+1\}} |\pi_\theta(\{h \in \mathbb{C} : \forall i \le k, h(X_i) = y_i\}) - \pi_{\theta'}(\{h \in \mathbb{C} : \forall i \le k, h(X_i) = y_i\})|$$
$$= \|\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k} - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}\|.$$

r_k

- Thus, w.p. ≥ 1-γ, $\|\pi_\theta - \pi_{\theta'}\| < L\gamma^\alpha + \|\mathbb{P}_{\mathbb{Y}_k(\theta)|\mathbb{X}_k} - \mathbb{P}_{\mathbb{Y}_k(\theta')|\mathbb{X}_k}\|.$
- Proceed as before, we get

$$\|\pi_\theta - \pi_{\theta'}\| < (L+1)\gamma^\alpha + 4(2ek)^{2d+2}\sqrt{\|\mathbb{P}_{\mathcal{Z}_d(\theta)} - \mathbb{P}_{\mathcal{Z}_d(\theta')}\|}.$$

- Plug in $k = c(d/\gamma)\log(1/\gamma)$, get $(L+1)\gamma^\alpha + 4\left(2ec\frac{d}{\gamma}\log\left(\frac{1}{\gamma}\right)\right)^{2d+2}\sqrt{\|\mathbb{P}_{\mathcal{Z}_d(\theta)} - \mathbb{P}_{\mathcal{Z}_d(\theta')}\|}.$ (*

Rate of Conv. under Hölder-Smooth

- Rate of conv. of estimate of $\mathbb{P}_{\mathcal{Z}_d(\theta_\star)}$
- ε -cover size bounded by grid-argument under hölder-smooth, plug that into the SC of Yachocos (1985), get T = O(ε^{-2} (L/ε)$^{d/\alpha}$ log(1/ε)) for ε. Solving for ε, we get ε = O(L (log(TL)/T)$^{\alpha/(d+2\alpha)}$).
- Plug this into (*), get the follow (hold for any γ)

$$\mathbb{E}\|\pi_{\hat\theta_T} - \pi_{\theta_\star}\| < (L+1)\gamma^\alpha + 4\left(2ec\frac{d}{\gamma}\log\left(\frac{1}{\gamma}\right)\right)^{2d+2} O\left(L\left(\frac{\log(TL)}{T}\right)^{\frac{\alpha}{2d+4\alpha}}\right)$$

- With $\gamma = \tilde{O}\left(T^{-\frac{\alpha}{2(d+2\alpha)(\alpha+2(d+1))}}\right)$ $\mathbb{E}\|\pi_{\hat\theta_T} - \pi_{\theta_\star}\| = \tilde{O}\left(LT^{-\frac{\alpha^2}{2(d+2\alpha)(\alpha+2(d+1))}}\right)$

QED

Is this Better than without Transfer ?

- The question becomes:
  - How much does knowledge of target distrib π^* help?
- There are some (constant factor) gains for passive learning [e.g. HKS1992]
- It really helps in Active learning:
  - Earlier, we showed can get $o(1/\varepsilon)$ for all π
- For many C (e.g. linear separators), no prior-indep alg has this guarantee.
- Plugging in that method, transfer method accesses $o(1/\varepsilon)$ labels on avg.

An Example of Prior-Dependent Learning

Self-verifying Bayesian Active Learning

(a special type of stopping criterion)

- Given ε, adaptively decides # of query, then halts
- has the property that $E[err] < \varepsilon$ when halts

Question: Can you do with $E[\#query] = o(1/\varepsilon)$? (passive learning need $1/\varepsilon$ labels)

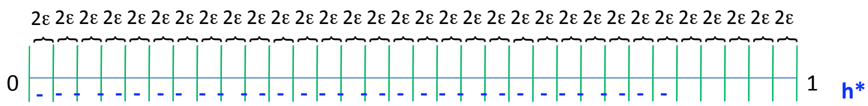
Example: Intervals

Verification Lower Bound

In non-Bayesian setting, supposing h* is empty interval.

Given any classifier h,
just to verify err(h) < ε,
Need to verify h* is not an interval of width 2ε.

Need an example in Ω(1/ε) regions to verify this fact.



Suppose h* is empty interval, D is uniform on [0,1]

 31

Interval Example with prior

- - - - - |++++++++|- - - - - - -

- Algorithm: Query random pts till find first +, do binary search to find end-pts. Halt when reach a pre-specified prior-based query budget. Output posterior's Bayes classifier.
- Let budget N be high enough so E[err] < ε
  - N = o(1/ε) sufficient for E[err|w*>0] < ε : if w* > 0, even prior-independent analysis needs only E[#queries|w*] = O(1/w* + log(1/ε)) = o(1/ε).
  - N = o(1/ε) sufficient for E[err|w*=0] < ε : if P(w*=0)>0, then after some L = O(log(1/ε)) queries, w.p.> 1-ε, most prob. mass on empty interval, so posterior's Bayes classifier has 0 error rate

 32

Can do o(1/eps) for any VC-class

Theorem: With the prior, can get $o(1/\varepsilon)$ QC

- There are methods that find a good classifier in o(1/eps) queries (though they aren't self-verifying) [BHW08]
- Need set a stopping criterion for those alg
- The stop criterion we use : budget
- Set the budget to be just large enough so $E[err] < \varepsilon$.