Transfer Learning and Prior Estimation for VC Classes

Liu Yang
04/25/2012

Notation

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

---

**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

Better Estimate of Prior!!
Identifiability of priors from joint distribs

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

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

Identifiability of priors by VC-dim joint distri.

- Threshold:

```
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>+---------------</td>
</tr>
</tbody>
</table>
```

- 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: $Z_{[VC]}(\pi_1) \equiv_d Z_{[VC]}(\pi_2)$ 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 $\text{vc} < \infty$ implies w.p.1 for all $h, g$ in $C$ with $h \neq g$
  $\lim_{m \to \infty} \rho_m(h,g) = \rho(h,g) > 0$
- $\rho$ is a metric on $C$ by assumption,
  so w.p.1 each $h$ in $C$ labels $\infty$-seq $(X_1, X_2, \ldots)$
  distinctly $(h(X_1), h(X_2), \ldots)$
- $\Rightarrow$ w.p.1 conditional distribution of the label seq
  $Z(\pi|X)$ identifies $\pi$
  $\Rightarrow$ distrib of $Z(\pi)$ identifies $\pi$
  i.e. $Z_{\infty}(\pi_1) \equiv_d Z_{\infty}(\pi_2)$ implies $\pi_1 = \pi_2$

Identifiability of Priors from Joint Distributions

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

Proof Sketch:
Fix any $m > \text{vc}$, $x_1, \ldots, x_m \in \mathcal{X}$, $y_1, \ldots, y_m \in \{0, 1\}$.
Note $C$ cannot shatter $(x_1, \ldots, x_m)$.
Let $\tilde{y}_1, \ldots, \tilde{y}_m \in \{0, 1\}$ be s.t. $\exists h \in C$ with $\forall i, h(x_i) = \tilde{y}_i$.
Clearly $\mathbb{P}(Z_m(\pi) = \{(x_i, y_i)\}_{i \in [m]} | X_m = \{x_i\}_{i \in [m]}) = 0$.
If $\exists k$ s.t. $y_k \neq \tilde{y}_k$, then letting $y'_i = y_i$ for $i \neq k$, and $y'_k = \tilde{y}_k$,

\[
\begin{align*}
\mathbb{P}(Z_m(\pi) & = \{(x_i, y'_i)\}_{i \in [m]} | X_m = \{x_i\}_{i \in [m]}) \\
& = \mathbb{P}(Z_m(\pi _{\{k\}}) = \{(x_i, y'_i)\}_{i \in [m]\{k\}} | X_m \{k\} = \{x_i\}_{i \in [m]\{k\}}) \\
& \quad - \mathbb{P}(Z_m(\pi) = \{(x_i, y'_i)\}_{i \in [m]} | X_m = \{x_i\}_{i \in [m]})
\end{align*}
\]

Induction: $\mathbb{P}(Z_m(\pi) = \cdot | X_m)$ function of $\mathbb{P}(Z_{[VC]}(\pi) = \cdot | X_{[VC]})$.
Identifiability of Priors from Joint Distributions

**Theorem:** \( Z_{[\text{vc}]}(\pi_1) \overset{d}{=} Z_{[\text{vc}]}(\pi_2) \iff \pi_1 = \pi_2. \)

**Proof Sketch:**

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

Classic result:
set of distibs 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
\( Z(\pi_1) \overset{d}{=} Z(\pi_2) \Rightarrow \pi_1 = \pi_2. \)

---

**Theorem:** \( Z_{[\text{vc}]}(\pi_1) \overset{d}{=} Z_{[\text{vc}]}(\pi_2) \iff \pi_1 = \pi_2. \)

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

**Proof Sketch:**

Let \( (x_1, \ldots, x_{\text{vc}}) \) be shattered by \( \mathcal{H} = \{h_1, \ldots, h_{2\text{vc}}\} \subseteq \mathbb{C}. \)
Let \( \mathcal{D} \) be uniform on \( \{x_1, \ldots, x_{\text{vc}}\} \),
let \( \pi_1 \) be uniform on \( \mathcal{H}. \)
Let \( \mathcal{H}' = \{h'_1, \ldots, h'_{2\text{vc}-1}\} \subset \mathcal{H} \) shatter \( (x_1, \ldots, x_{\text{vc}-1}) \)
s.t. \( h'_i(x_{\text{vc}}) = \text{Parity}({h'_1(x_1), \ldots, h'_i(x_{\text{vc}-1})}). \)
Let \( \pi_2 \) be uniform on \( \mathcal{H}'. \)
Clearly \( \pi_1 \neq \pi_2. \)

But for \( m < \text{vc}, Z_{[m]}(\pi_1) \overset{d}{=} Z_{[m]}(\pi_2): \)
unif cond on labels given distinct \( X_1, \ldots, X_m. \)
Transfer Learning Setting

- Collection $\Pi$ of distributions on $C$. (known)
- Target distribution $\pi^*$ in $\Pi$. (unknown)
- Independent target functions $h_1^*, \ldots, h_T^* \sim \pi^*$ (unknown)
- Independent i.i.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 algorithm “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 $h_t^*$ at $X_j^{(t)}$ value that alg. needs to access.

Estimating the prior

- **Principle**: learning would be easier if know $\pi^*$
- **Fact**: $\pi^*$ is identifiable by distribution of $Z_{[VC]}^{(t)}$
- **Strategy**: Take samples $Z_{[VC]}^{(i)}$ from past tasks 1, $\ldots, t-1$, use them to estimate distribution of $Z_{[VC]}^{(i)}$, convert that into an estimate $\pi'_t$ of $\pi^*$.
- Use $\pi'_t$ in a prior-dependent learning alg for new task $h_t^*$
- Assume $\Pi$ is totally bounded in total variation
- Can estimate $\pi^*$ at a bounded rate:
  $\| \pi^* - \pi'_t \| < \delta_t$ converges to 0 (holds whp)
Main Theorem

**Theorem 1** There exists an estimator \( \theta_{j,k} = \theta_T(Z_{1}(\theta_k), \ldots, Z_{T}(\theta_k)) \) 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 \( \alpha > 0 \), \( \lim_{T \to \infty} R(T, \alpha) = \infty \), and for any \( T \in \mathbb{N}_0 \) and \( \theta \in \Theta \),

\[
P \left( \| \theta_{j,k} - \theta_k \| > R(T, \alpha) \right) \leq \delta(T, \alpha) \leq \alpha.
\]

**Pf Idea:** relate convergence of estimator for \( d \)-dim joint to convergence of estimator for the prior

Transfer Learning

- Given a prior-dependent learning \( A(\varepsilon, \pi) \), with \( E[\# \text{ labels accessed}] = \Lambda(\varepsilon, \pi) \) and producing \( \hat{h} \) with \( E[\rho(\hat{h}, h^*)] \leq \varepsilon \)

For \( t = 1, \ldots, T \)

- **If** \( \delta_{t-1} > \varepsilon / 4 \),
  - run prior-indep learning on \( Z_{[\text{VC} / \varepsilon]}(t) \) to get \( \hat{h}_t \)
- **Else** let \( \pi''_t = \arg\min_{\pi} \Delta_{t+1, \delta_{t+1}}(\varepsilon / 2, \pi) \) and run \( A(\varepsilon / 2, \pi''_t) \) on \( Z(t) \) to get \( \hat{h}_t \)

**Theorem:** For all \( t \), \( E[\rho(\hat{h}_t, h_t^*)] \leq \varepsilon \), and \( \limsup_{T \to \infty} E[\# \text{labels accessed}] / T \leq \Lambda(\varepsilon / 2, \pi^*) + \text{vc} \).
Relate Prior to k-dim joint

Lemma: There exists a sequence \( r_k = o(1) \) such that \( \forall k \in \mathbb{N}, \forall \theta, \theta' \in \Theta, \)
\[ \| P_{Z_t(\theta)} - P_{Z_t(\theta')} \| \leq \| \pi_\theta - \pi_{\theta'} \| \leq \| P_{Z_{t_k}(\theta)} - P_{Z_{t_k}(\theta')} \| + r_k. \]

Proof:

- The left inequality follows from, for any \( \theta, \theta' \) in \( \Theta \) and \( t \) (natural num), \( \| P_{Z_{t_k}(\theta)} - P_{Z_{t_k}(\theta')} \| \leq \| \pi_\theta - \pi_{\theta'} \| = \| \nabla_{\theta} \| \).
- To show the right inequality: Fix \( \theta, \theta' \) in \( \Theta \), let \( \gamma > 0 \), let \( B \) subset eq \((X \times \{-1, +1}\))^\infty be a measurable set s.t.
\[ \| \pi_\theta - \pi_{\theta'} \| = \| P_{Z_{t_k}(\theta)}(B) - P_{Z_{t_k}(\theta')}(B) \| > 0. \]
- Carathéodory’s extension theorem implies there exist disjoint sets \( \{A_i\}_{i \in \mathbb{N}} \) in \( \mathbb{N} \) where \( A_i \) is an event for finite number of data pts, s.t. \( B \subseteq \bigcup_{i \in \mathbb{N}} A_i \).
\[ P_{Z_{t_k}(\theta)}(B) - P_{Z_{t_k}(\theta')}(B) < \sum_{i \in \mathbb{N}} P_{Z_{t_k}(\theta)}(A_i) - \sum_{i \in \mathbb{N}} P_{Z_{t_k}(\theta')}(A_i) + \gamma. \]
- Since these sums are bounded, there must exist \( n \) in \( \mathbb{N} \) s.t.
\[ \sum_{i \in \mathbb{N}} P_{Z_{t_k}(\theta)}(A_i) \leq \gamma + \sum_{i \in \mathbb{N}} P_{Z_{t_k}(\theta')}(A_i) + \gamma. \]

In sum, \( \| \pi_\theta - \pi_{\theta'} \| \leq \lim_{k \to \infty} \| P_{Z_{t_k}(\theta)} - P_{Z_{t_k}(\theta')} \| + 3\gamma. \)

- Taking the limit as \( \gamma \to 0 \) implies \( \| \pi_\theta - \pi_{\theta'} \| \leq \lim_{k \to \infty} \| P_{Z_{t_k}(\theta)} - P_{Z_{t_k}(\theta')} \| \)
- Particularly, it implies there exists a sequence \( r_k(\theta, \theta') = o(1) \) s.t.
\[ \forall k \in \mathbb{N}, \| \pi_\theta - \pi_{\theta'} \| \leq \| P_{Z_{t_k}(\theta)} - P_{Z_{t_k}(\theta')} \| + r_k(\theta, \theta'). \]

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.
  \[ ||P_{Y_{tk}(x)} - P_{Y_{tk}(y')}|| = (1/2) \sum_{y^d \in \{-1, +1\}^d} |P_{Y_{tk}(x)}(y^d) - P_{Y_{tk}(y')}(y^d)|. \]
- By Sauer’s Lemma this is \[ \leq \]
- By def of total variation dist.
- Notations:
  \[ I \subseteq \{1, \ldots, k\}, \text{fix } \bar{x}_I \in A^{|I|} \text{and } \bar{y}_I \in \{-1, +1\}^{|I|}. \text{Then the } \bar{y}_I \in \{-1, +1\}^{|I|} \text{for which no } h \in C \text{has } h(\bar{x}_I) = \bar{y}_I \text{for which } ||\bar{y}_I - \bar{y}_I||, \text{is minimal, has } ||\bar{y}_I - \bar{y}_I|| \leq d + 1, \text{and for any } i \in I \text{ with } \bar{y}_I \neq \bar{y}_I, \text{letting } \bar{y}_I = \bar{y}_I \text{for } j \in I \setminus \{i\} \text{and } \bar{y}_I = \bar{y}_I, \text{we have} \]
  \[ P_{Y_{tk}(x)}(\bar{y}_I|\bar{x}_I) = P_{Y_{tk}(x)}(\bar{y}_I|\bar{x}_I), \]
  (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.
  \[ |P_{Y_{tk}(y|x)}(y|x) - P_{Y_{tk}(x|x)}(y|x)| \]
  \[ \leq |P_{Y_{tk}(y|x)}(y|x) - P_{Y_{tk}(x|x)}(y|x)| \]
  \[ + |P_{Y_{tk}(y|x)}(y|x) - P_{Y_{tk}(x|x)}(y|x)|. \]

© Liu Yang 2012
Tree Argument: Combinatorics

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

At any level, left to right nodes have decreasing \(|I|\) values.

Branches left once, gets a diff. of prob.s for set \( I \) of one less element.

Branches right once, gets a difference of prob.s for a \( y_I \) one closer to an \( y_I \) unrealized than parent.

Stop branching upon reaching a set \( I \) and a \( y_I \) s.t. either \( y_I \) is an unrealized labeling, or \(|I| = d\).

Any path can branch left \( \leq k - d \) times (total) before reaching a set \( I \) w/ only \( d \) elements; can branch right \( \leq d + 1 \) times in a row before reaching a \( y_I \) s.t. both prob.s zero, so the diff is zero.

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 \( y_I \in \{-1, +1\}^k\), \( x \in \mathbb{X}^k\)
    
    \[
    \left| P_{Y_A(y)|X_A}(y|x) - P_{Y_A(y')}|X_A}(y'|x) \right| \\
    \leq (k - d)^2d \cdot \max_{y \in \{-1, +1\}^d} \max_{D \in \{1, \ldots, k\}^d} \left[ |P_{Y_A(y)|X_A}(D|x_D) - P_{Y_A(y')}|X_A}(D'|x_{D'})| \right].
    \]
Relate k-dim Joint to d-dim Joint

- Note

\[
\max_{g^d \in \{-1,+1\}^d} \max_{D \in \{1,\ldots,k\}^d} \mathbb{E} \left[ \max_{g^d \in \{-1,+1\}^d} \mathbb{E} \left[ P_{Y_d|X_d}(g^d^k) - P_{Y_d|X_d}(g^d) \right] \right]
\]

\[
\leq \sum_{g^d \in \{-1,+1\}^d} \sum_{D \in \{1,\ldots,k\}^d} \mathbb{E} \left[ P_{Y_d|X_d}(g^d) - P_{Y_d|X_d}(g^d) \right]
\]

\[
\leq (2k)^d \max_{g^d \in \{-1,+1\}^d} \mathbb{E} \left[ P_{Y_d|X_d}(g^d) - P_{Y_d|X_d}(g^d) \right]
\]

- By exchangeability, the last line equals

\[
(2k)^d \max_{g^d \in \{-1,+1\}^d} \mathbb{E} \left[ P_{Y_d|X_d}(g^d) - P_{Y_d|X_d}(g^d) \right].
\]

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

Claim:

\[
\mathbb{E} \left[ P_{Y_d|X_d}(g^d|X_d) - P_{Y_d|X_d}(g^d|X_d) \right] \leq 4 \|P_{Z_d}(\theta) - P_{Z_d}(\theta')\|
\]

Proof of the Claim

Proof:
Suppose

\[
\mathbb{E} \left[ P_{Y_d}(g^d|X_d) - P_{Y_d}(g^d|X_d) \right] \geq \epsilon,
\]

for some \( g^d \). Then either

\[
P \left( P_{Y_d}(g^d|X_d) - P_{X_d}(g^d|X_d) \geq \epsilon/4 \right) \geq \epsilon/4,
\]

or

\[
P \left( P_{Y_d}(g^d|X_d) - P_{X_d}(g^d|X_d) \geq \epsilon/4 \right) \geq \epsilon/4.
\]

For whichever is the case, let \( A_\epsilon \) denote the corresponding measurable subset of \( X^d \), of probability at least \( \epsilon/4 \). Then

\[
\|P_{Z_d}(\theta) - P_{Z_d}(\theta')\| \geq \|P_{Z_d}(\theta)(A_\epsilon \times \{g^d\}) - P_{Z_d}(\theta')(A_\epsilon \times \{g^d\})\|
\]

\[
\geq \left( \epsilon/4 \right) P_{X_d}(A_\epsilon) \geq \epsilon^2/16.
\]

Therefore,

\[
\mathbb{E} \left[ P_{Y_d}(g^d|X_d) - P_{Y_d}(g^d|X_d) \right] \leq 4 \|P_{Z_d}(\theta) - P_{Z_d}(\theta')\|.
\]
Reflect the Path of Proof

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

Just showed

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

So in total

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

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

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

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

$g$ is monotonic in $\epsilon \Rightarrow \lim_{\epsilon \to 0} g(\epsilon) = \lim_{k \to \infty} g(\epsilon_k) = \lim_{k \to \infty} 2r_k = 0$.)

Distri. Estimation Rate

• The last component: rate of conv. of our estimate of $P_{Z_{d}(\theta)}$.
  - $N(\epsilon)$ is the $\epsilon$-covering number $\{P_{\delta,\epsilon}(\theta) : \theta \in \Theta\}$
  - Taking $\theta_{T_{d}}$ as the minimum distance skeleton estimate of Yatracos (1985) achieves expected tvd $\epsilon$ from $\theta_{T_{d}}$, for some $T = O((1/\epsilon^2)\log N(\epsilon/4))$

Solving for $\epsilon$ in terms of $T$ implies $E[\text{tvd of d-dim}] \to 0$ as $T \to \infty$

• Conclusion for prior estimation:
  - Pick the sequence of $R_t$ s.t. $R_t \to 0$, but with $E[w_t]/R_t \to 0$
  - Let $w_t$ be $E[\text{tvd of d-dim}]$ for any $t$, apply Markov ineq. $\Rightarrow P(w_t > R_t) < E[w_t]/R_t$
  - Since $E[\text{tvd of d-dim}] \to 0$, Markov’s ineq. $\Rightarrow$ there is a bound on tvd $\to 0$ which holds with prob. that $\to 1$, as $T \to \infty$
  - If tvd of d-dim joints $\to 0$, plugging into $g()$ (just proved), tvd of priors $\to 0$.

• Together we just proved the theorem

**Theorem 1** There exists an estimator $\hat{\Theta}_{T_{d}(\theta)} = \hat{\theta}(Z_{td}(\theta), \ldots, Z_{td}(\theta))$, 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 $\alpha > 0$, $\lim_{T \to \infty} \delta(T, \alpha) = 0$ and for any $T \in \mathbb{N}$ and $\theta \in \Theta$,

$$P\left(\|\pi_{\hat{\theta}_T} - \pi_{\theta}\| > R(T, \alpha)\right) \leq \delta(T, \alpha) \leq \alpha.$$
Rate of Conv. under Hölder–Smooth

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

Theorem. For $\Theta$, any class of priors on $C$ having $(L, \alpha)$-Hölder smooth densities $\{f_\theta : \theta \in \Theta\}$, for any $T \in \mathbb{N}$, there exists an estimator $\hat{\theta}_T = \hat{\theta}_T(Z_{1d}(\theta), \ldots, Z_{Td}(\theta))$ such that 
\[ \sup_{\theta, \theta'} \mathbb{E} \|\pi_{\theta'} - \pi_{\theta}\| = O \left( LT^{-\frac{\rho^2}{(d+2)(d+2d+1)}} \right). \]

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

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 $f: C \to \mathbb{R}$ is $(L, \alpha)$-Hölder smooth if 
\[ \forall h, g \in C, |f(h) - f(g)| \leq L \rho(h, g)^{\alpha}. \]

Theorem. For $\Theta$, any class of priors on $C$ having $(L, \alpha)$-Hölder smooth densities $\{f_\theta : \theta \in \Theta\}$, for any $T \in \mathbb{N}$, there exists an estimator $\hat{\theta}_T = \hat{\theta}_T(Z_{1d}(\theta), \ldots, Z_{Td}(\theta))$ such that 
\[ \sup_{\theta, \theta'} \mathbb{E} \|\pi_{\theta'} - \pi_{\theta}\| = O \left( LT^{-\frac{\rho^2}{(d+2)(d+2d+1)}} \right). \]

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 $\theta \in \Theta$, $\pi'$ denote a (conditional on $X_1, \ldots, X_d$) distribution 
  \[ f_\theta' \] denote the (conditional on $X_1, \ldots, X_d$) density function of $\pi'$ with respect to $\pi_0$.
- For any $g \in C$, 
  \[ f_\theta'(g) = \frac{\pi'(\{h \in C : h(X_i) = g(X_i)\})}{\pi'(\{h \in C : h(X_i) = g(X_i)\})}. \]
  (or 0 if $\pi_0(\{h \in C : h(X_i) = g(X_i)\}) = 0$).
- By smoothness, w. p. $\gamma > 1 - \gamma$, we have everywhere $|f_\theta(h) - f_\theta'(h)| < L_\gamma^\alpha$.

© Liu Yang 2012
Rate of Conv. under Hölder-Smooth

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

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

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

$$\begin{align*}
(1/2) \int |f'_\theta - f'_{\theta'}|d\pi_0 &= (1/2) \sum_{y_{k-1} = -1}^{y_{k+1}} \pi_\theta(\{h \in C : \forall i \leq k, h(X_i) = y_i\}) - \pi_{\theta'}(\{h \in C : \forall i \leq k, h(X_i) = y_i\}) \\
&= \|P_{\{h \in C : \forall i \leq k, h(X_i) = y_i\}} - P_{\{h \in C : \forall i \leq k, h(X_i) = y_i\}}\|.
\end{align*}$$

- Thus, w.p. $\geq 1-\gamma$,

$$\|\pi_\theta - \pi_{\theta'}\| < L \gamma^n + \|P_{\{h \in C : \forall i \leq k, h(X_i) = y_i\}} - P_{\{h \in C : \forall i \leq k, h(X_i) = y_i\}}\|.$$ 

- Proceed as before, we get

$$\|\pi_\theta - \pi_{\theta'}\| < (L + 1) \gamma^n + 4(2c)^{d+2} \frac{\sqrt{\|P_{\{h \in C : \forall i \leq k, h(X_i) = y_i\}} - P_{\{h \in C : \forall i \leq k, h(X_i) = y_i\}}\|}}{\gamma}.$$ 

- Plug in $k = c(d/\gamma) \log(1/\gamma)$, get $(L + 1) \gamma^n + 4(2c)^{d+2} \frac{\sqrt{\|P_{\{h \in C : \forall i \leq k, h(X_i) = y_i\}} - P_{\{h \in C : \forall i \leq k, h(X_i) = y_i\}}\|}}{\gamma}$.

Rate of Conv. under Hölder-Smooth

- Rate of conv. of estimate of $P_{\{h \in C : \forall i \leq k, h(X_i) = y_i\}}$

- $\varepsilon$-cover size bounded by grid-argument under hölder-smooth, plug that into the SC of Yachocos (1985), get $T = O(\varepsilon^{-2} (L/\varepsilon)^{d/\alpha} \log(1/\varepsilon))$ for $\varepsilon$. Solving for $\varepsilon$, we get $\varepsilon = O(L (\log(TL)/T)^{\alpha/(\alpha+2d)}).$

- Plug this into (*)*, get the follow (hold for any $\gamma$)

$$\mathbb{E}\|\pi_{\theta'} - \pi_{\theta}\| < (L + 1) \gamma^n + 4 \left(2c \frac{\varepsilon}{\gamma} \log \left(\frac{1}{\gamma}\right)\right)^{d+2} \frac{\sqrt{\|P_{\{h \in C : \forall i \leq k, h(X_i) = y_i\}} - P_{\{h \in C : \forall i \leq k, h(X_i) = y_i\}}\|}}{\gamma}.$$ 

- With $\gamma = \tilde{O}(T^{-\frac{\alpha}{2d+2\alpha+2\alpha(d+2)}})$, $\mathbb{E}\|\pi_{\theta'} - \pi_{\theta}\| = \tilde{O}(LT^{-\frac{\alpha}{2d+2\alpha+2\alpha(d+2)}}).$

QED
Is this Better than without Transfer?

• The question becomes:
  - How much does knowledge of target distrib $\pi^*$ 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 $\pi$
• 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 $\varepsilon$, adaptively decides # of query, then halts
- has the property that $E[\text{err}] < \varepsilon$ when halts
Question: Can you do with $E[\text{#query}] = o(1/\varepsilon)$? (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 $\text{err}(h) < \varepsilon$.

Need to verify $h^*$ is not an interval of width $2\varepsilon$.

Need an example in $\Omega(1/\varepsilon)$ regions to verify this fact.

![Diagram showing an example with intervals](image)

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 $E[\text{err}] < \varepsilon$**
  - $N = o(1/\varepsilon)$ sufficient for $E[\text{err}|w^*>0] < \varepsilon$: if $w^* > 0$, even prior-independent analysis needs only $E[\#\text{queries}|w^*] = O(1/w^* + \log(1/\varepsilon)) = o(1/\varepsilon)$.
  - $N = o(1/\varepsilon)$ sufficient for $E[\text{err}|w^*=0] < \varepsilon$: if $P(w^*=0)>0$, then after some $L = O(\log(1/\varepsilon))$ queries, w.p.$>1-\varepsilon$, most prob. mass on empty interval, so posterior's Bayes classifier has 0 error rate.
Can do o(1/eps) for any VC-class

Theorem: With the prior, can get o(1/ε) 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] < ε.