Unlabeled data: Now it helps, now it doesn't

Aarti Singh, Robert Nowak, Xiaojin Zhu

University of Wisconsin–Madison

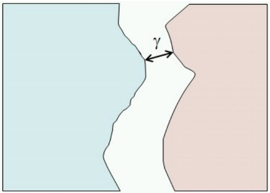
NIPS 2008

Semi-Supervised Learning under Cluster Assumption

- $f(X)$ is the optimal predictor of Y given P_{XY}
- Data: n labeled points $\stackrel{iid}{\sim} P_{XY}$, m unlabeled points $\stackrel{iid}{\sim} P_X$, $m \gg n$
- Goal: learn $f(X)$ from data

Semi-Supervised Learning under Cluster Assumption

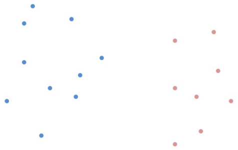
- $f(X)$ is the optimal predictor of Y given P_{XY}
- Data: n labeled points $\stackrel{iid}{\sim} P_{XY}$, m unlabeled points $\stackrel{iid}{\sim} P_X$, $m \gg n$
- Goal: learn $f(X)$ from data
- The cluster assumption:
  - P_X is a mixture of components in d-dim
  - $f(X)$ smooth on each component
  - γ is the margin (> 0 separation, < 0 overlap), characterizes difficulty of learning problem



Does Unlabeled Data Help?

[BB05,BDLP08,BL07,CC95,LW08,Ni08,Ri07]

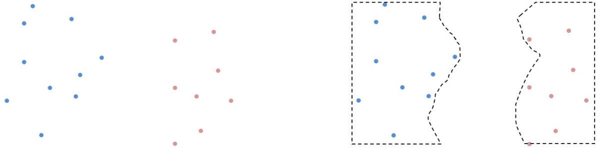
- **Unlabeled data doesn't help**



Does Unlabeled Data Help?

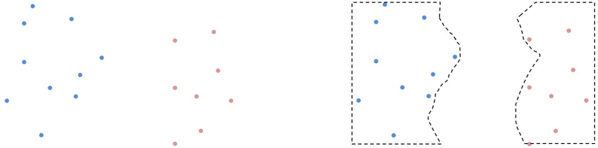
[BB05,BDLP08,BL07,CC95,LW08,Ni08,Ri07]

- **Unlabeled data doesn't help**



Does Unlabeled Data Help?
[BB05,BDLP08,BL07,CC95,LW08,Ni08,Ri07]

- **Unlabeled data doesn't help**

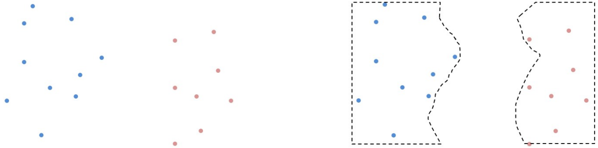


For any $\gamma > 0$, given enough labeled data, unlabeled data is superfluous (SSL does not result in faster rates of convergence).

Does Unlabeled Data Help?

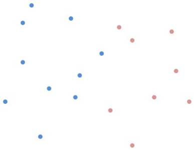
[BB05,BDLP08,BL07,CC95,LW08,Ni08,Ri07]

- **Unlabeled data doesn't help**



For any $\gamma > 0$, given enough labeled data, unlabeled data is superfluous (SSL does not result in faster rates of convergence).

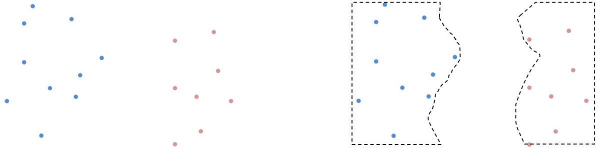
- **Unlabeled data helps**



Does Unlabeled Data Help?

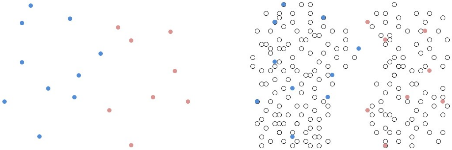
[BB05,BDLP08,BL07,CC95,LW08,Ni08,Ri07]

- **Unlabeled data doesn't help**



For any $\gamma > 0$, given enough labeled data, unlabeled data is superfluous (SSL does not result in faster rates of convergence).

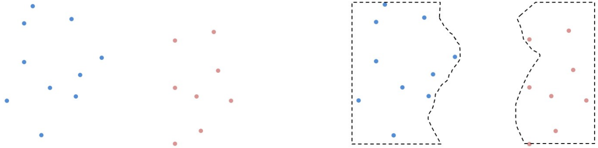
- **Unlabeled data helps**



Does Unlabeled Data Help?

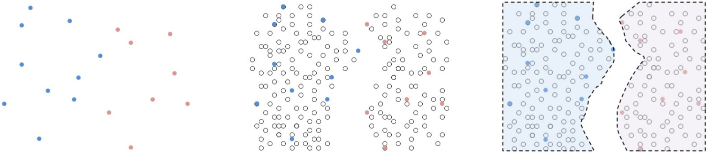
[BB05,BDLP08,BL07,CC95,LW08,Ni08,Ri07]

- **Unlabeled data doesn't help**



For any $\gamma > 0$, given enough labeled data, unlabeled data is superfluous (SSL does not result in faster rates of convergence).

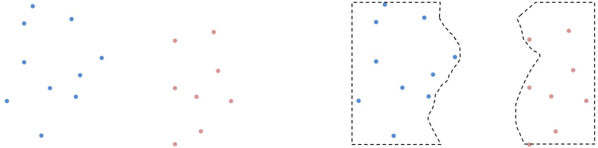
- **Unlabeled data helps**



Does Unlabeled Data Help?

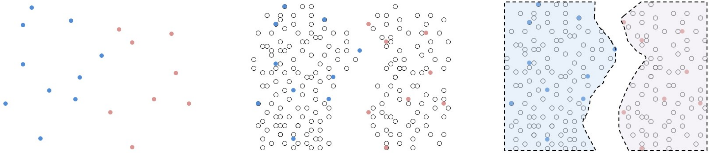
[BB05,BDLP08,BL07,CC95,LW08,Ni08,Ri07]

- **Unlabeled data doesn't help**



For any $\gamma > 0$, given enough labeled data, unlabeled data is superfluous (SSL does not result in faster rates of convergence).

- **Unlabeled data helps**



Given a finite labeled data, there are learning problems with small enough γ that SL fails, whereas perfect knowledge of components would yield small error.

Our Contributions

1. Benefits of SSL not always revealed through asymptotic analysis and rates
2. Instead, we quantify them with finite sample analysis
3. We show SSL sometimes helps, sometimes not
4. There are cases in which SSL has faster rates than SL

Finite Sample Bounds

- $f_{m,n}$: predictor learned from m unlabeled and n labeled points
  - $m = 0$: supervised
  - $m > 0$: semi-supervised
  - $m = \infty$: oracle (full knowledge of P_X, but not f)

Finite Sample Bounds

- $f_{m,n}$: predictor learned from m unlabeled and n labeled points
  - $m = 0$: supervised
  - $m > 0$: semi-supervised
  - $m = \infty$: oracle (full knowledge of P_X, but not f)
- $R(f_{m,n})$: Risk under loss function ℓ, e.g., $\ell = (f_{m,n}(X) - Y)^2$

$$R(f_{m,n}) = \mathbb{E}_{(X,Y)\sim P_{XY}} \left[\ell(f_{m,n}(X), Y)\right]$$

Finite Sample Bounds

- $f_{m,n}$: predictor learned from m unlabeled and n labeled points
  - $m = 0$: supervised
  - $m > 0$: semi-supervised
  - $m = \infty$: oracle (full knowledge of P_X, but not f)
- $R(f_{m,n})$: Risk under loss function ℓ, e.g., $\ell = (f_{m,n}(X) - Y)^2$

$$R(f_{m,n}) = \mathbb{E}_{(X,Y)\sim P_{XY}}\left[\ell(f_{m,n}(X), Y)\right]$$

- $\mathcal{E}(f_{m,n})$: Excess Risk, the difference between expected Risk (over random draws of training set) and Bayes Risk

$$\mathcal{E}(f_{m,n}) = \mathbb{E}_{\text{training}}\left[R(f_{m,n})\right] - \inf_{\tilde{f}} R(\tilde{f})$$

Finite Sample Bounds

- $f_{m,n}$: predictor learned from m unlabeled and n labeled points
  - $m = 0$: supervised
  - $m > 0$: semi-supervised
  - $m = \infty$: oracle (full knowledge of P_X, but not f)
- $R(f_{m,n})$: Risk under loss function ℓ, e.g., $\ell = (f_{m,n}(X) - Y)^2$

$$R(f_{m,n}) = \mathbb{E}_{(X,Y)\sim P_{XY}}\left[\ell(f_{m,n}(X), Y)\right]$$

- $\mathcal{E}(f_{m,n})$: Excess Risk, the difference between expected Risk (over random draws of training set) and Bayes Risk

$$\mathcal{E}(f_{m,n}) = \mathbb{E}_{\text{training}}\left[R(f_{m,n})\right] - \inf_{\tilde{f}} R(\tilde{f})$$

- Minimax error

$$\epsilon_{m,n,\gamma} \overset{\text{polylog}}{\sim} \inf_{f_{m,n}} \sup_{P(\gamma)} \mathcal{E}(f_{m,n})$$

Finite Sample Bounds

- $f_{m,n}$: predictor learned from m unlabeled and n labeled points
  - $m = 0$: supervised
  - $m > 0$: semi-supervised
  - $m = \infty$: oracle (full knowledge of P_X, but not f)
- $R(f_{m,n})$: Risk under loss function ℓ, e.g., $\ell = (f_{m,n}(X) - Y)^2$

$$R(f_{m,n}) = \mathbb{E}_{(X,Y)\sim P_{XY}}\left[\ell(f_{m,n}(X), Y)\right]$$

- $\mathcal{E}(f_{m,n})$: Excess Risk, the difference between expected Risk (over random draws of training set) and Bayes Risk

$$\mathcal{E}(f_{m,n}) = \mathbb{E}_{\text{training}}\left[R(f_{m,n})\right] - \inf_{\tilde{f}} R(\tilde{f})$$

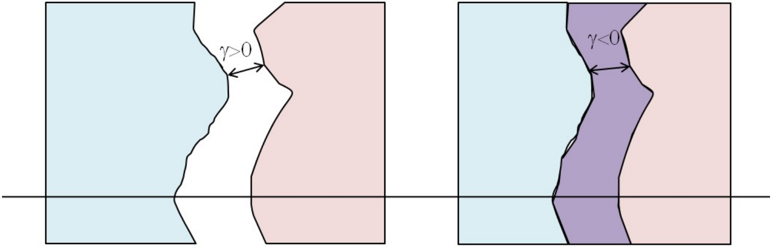
- Minimax error

$$\epsilon_{m,n,\gamma} \overset{\text{polylog}}{\sim} \inf_{f_{m,n}} \sup_{P(\gamma)} \mathcal{E}(f_{m,n})$$

- $\epsilon_{\infty,n,\gamma} \leq \epsilon_{m,n,\gamma} \leq \epsilon_{0,n,\gamma}$

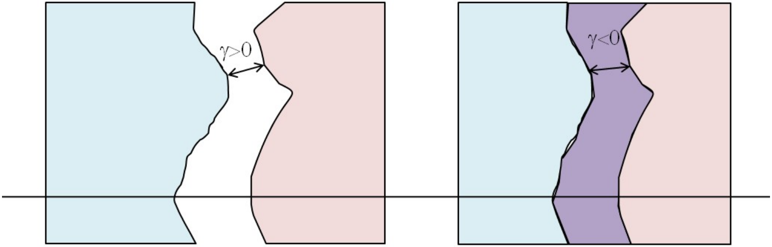
Mathematical Formalization of Cluster Assumption

- Components (compact support, Lipschitz boundary)
- Density bounded from below and above, Hölder-α smooth



Mathematical Formalization of Cluster Assumption

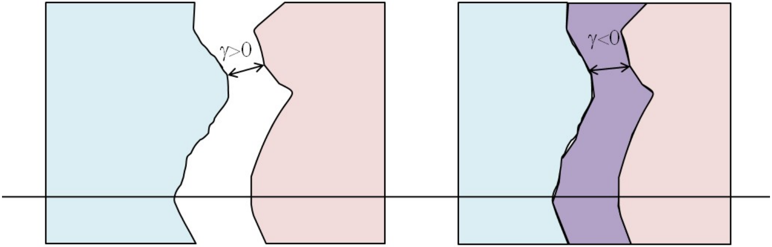
- Components (compact support, Lipschitz boundary)
- Density bounded from below and above, Hölder-α smooth



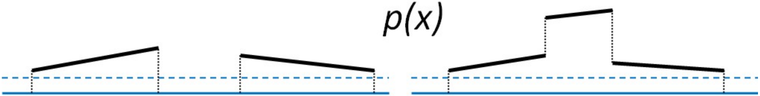
- Decision sets $\mathcal{D}$: all intersections of components 

Mathematical Formalization of Cluster Assumption

- Components (compact support, Lipschitz boundary)
- Density bounded from below and above, Hölder-α smooth



- Decision sets $\mathcal{D}$: all intersections of components 
- Overall density *jumps* at decision set boundaries



SSL Approach

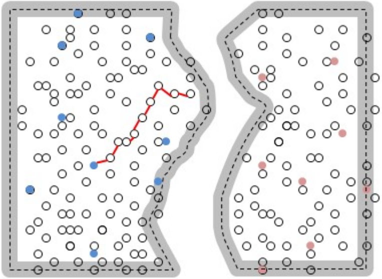
- Oracle knows the shape of decision sets, learns within a decision set.

SSL Approach

- Oracle knows the shape of decision sets, learns within a decision set.
- SSL mimics Oracle, learns only from *connected* labeled points

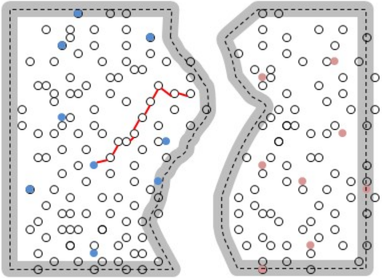
SSL Approach

- Oracle knows the shape of decision sets, learns within a decision set.
- SSL mimics Oracle, learns only from *connected* labeled points
- Connected: $x_1 \leftrightarrow x_2$ if there is a sequence of unlabeled steppingstones: (1) close together, (2) similar local density



SSL Approach

- Oracle knows the shape of decision sets, learns within a decision set.
- SSL mimics Oracle, learns only from *connected* labeled points
- Connected: $x_1 \leftrightarrow x_2$ if there is a sequence of unlabeled steppingstones: (1) close together, (2) similar local density



- Connectedness is almost as good as knowing the decision sets: **Lemma**: if $|\gamma| > Cm^{-1/d}$, then for all pairs x_1, x_2 not in a small tube around decision set boundaries, with large probability

x_1, x_2 **in same decision set if and only if** $x_1 \leftrightarrow x_2$

SSL Error

Corollary: if $|\gamma| > Cm^{-1/d}$, then SSL is only "a bit worse" than oracle:

$$\epsilon_{m,n,\gamma} \leq \epsilon_{\infty,n,\gamma} + O\left(nm^{-1/d}\right)$$

SSL Error

Corollary: if $|\gamma| > Cm^{-1/d}$, then SSL is only "a bit worse" than oracle:

$$\epsilon_{m,n,\gamma} \leq \epsilon_{\infty,n,\gamma} + O\left(nm^{-1/d}\right)$$

- The value of unlabeled data: if $m \gg n$ s.t. $nm^{-1/d} \leq \epsilon_{\infty,n,\gamma}$, then SSL is as good as Oracle.
  - if $\epsilon_{\infty,n,\gamma}$ decays polynomially, m must grow polynomially with n
  - if $\epsilon_{\infty,n,\gamma}$ decays exponentially, m must grow exponentially with n

SSL Error

Corollary: if $|\gamma| > Cm^{-1/d}$, then SSL is only "a bit worse" than oracle:

$$\epsilon_{m,n,\gamma} \leq \epsilon_{\infty,n,\gamma} + O\left(nm^{-1/d}\right)$$

- The value of unlabeled data: if $m \gg n$ s.t. $nm^{-1/d} \leq \epsilon_{\infty,n,\gamma}$, then SSL is as good as Oracle.
  - if $\epsilon_{\infty,n,\gamma}$ decays polynomially, m must grow polynomially with n
  - if $\epsilon_{\infty,n,\gamma}$ decays exponentially, m must grow exponentially with n
- If, in addition, Oracle is better than any ordinary SL

$$\epsilon_{\infty,n,\gamma} < \epsilon_{0,n,\gamma}$$

then SSL helps.

Application to SSL Regression

- Assumption: target function Hölder-α smooth within a decision set, but may be discontinuous across decision sets.

Application to SSL Regression

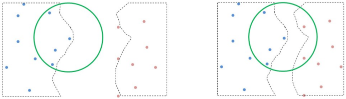
- Assumption: target function Hölder-α smooth within a decision set, but may be discontinuous across decision sets.
- Two possible sources of error:
  1. regression error within decision sets $n^{-2\alpha/(2\alpha+d)}$
  2. error in estimating boundaries of decision sets $n^{-1/d}$

Application to SSL Regression

- Assumption: target function Hölder-α smooth within a decision set, but may be discontinuous across decision sets.
- Two possible sources of error:
  1. regression error within decision sets $n^{-2\alpha/(2\alpha+d)}$
  2. error in estimating boundaries of decision sets $n^{-1/d}$
- Oracle: learn f on each decision set separately, $\epsilon_{\infty,n,\gamma} = n^{-2\alpha/(2\alpha+d)}$

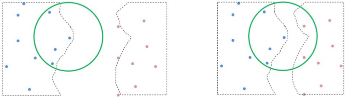
Application to SSL Regression

- Assumption: target function Hölder-α smooth within a decision set, but may be discontinuous across decision sets.
- Two possible sources of error:
  1. regression error within decision sets $n^{-2\alpha/(2\alpha+d)}$
  2. error in estimating boundaries of decision sets $n^{-1/d}$
- Oracle: learn f on each decision set separately, $\epsilon_{\infty,n,\gamma} = n^{-2\alpha/(2\alpha+d)}$
- SL: if $\gamma > cn^{-1/d}$ then $\epsilon_{0,n,\gamma} = n^{-2\alpha/(2\alpha+d)}$, otherwise $\epsilon_{0,n,\gamma} = n^{-1/d}$ (worse: blur across decision sets).

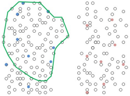


Application to SSL Regression

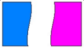
- Assumption: target function Hölder-α smooth within a decision set, but may be discontinuous across decision sets.
- Two possible sources of error:
  1. regression error within decision sets $n^{-2\alpha/(2\alpha+d)}$
  2. error in estimating boundaries of decision sets $n^{-1/d}$
- Oracle: learn f on each decision set separately, $\epsilon_{\infty,n,\gamma} = n^{-2\alpha/(2\alpha+d)}$
- SL: if $\gamma > cn^{-1/d}$ then $\epsilon_{0,n,\gamma} = n^{-2\alpha/(2\alpha+d)}$, otherwise $\epsilon_{0,n,\gamma} = n^{-1/d}$ (worse: blur across decision sets).



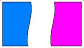
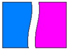
- SSL: if $|\gamma| > Cm^{-1/d}$ and $m \gg n^{2d}$, then the same as Oracle.



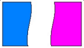
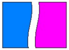
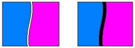
Unlabeled data: now it helps, now it doesn't

	margin	Oracle $\epsilon_{\infty,n,\gamma}$	SL $\epsilon_{0,n,\gamma}$	SSL $\epsilon_{m,n,\gamma}$	SSL helps?
	$n^{-\frac{1}{d}} \leq \gamma$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	no

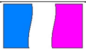
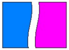
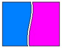
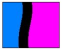
Unlabeled data: now it helps, now it doesn't

	margin	Oracle $\epsilon_{\infty,n,\gamma}$	SL $\epsilon_{0,n,\gamma}$	SSL $\epsilon_{m,n,\gamma}$	SSL helps?
	$n^{-\frac{1}{d}} \leq \gamma$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	no
	$m^{-\frac{1}{d}} \leq \gamma < n^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	yes

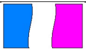
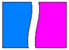
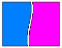
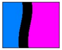
Unlabeled data: now it helps, now it doesn't

	margin	Oracle $\epsilon_{\infty,n,\gamma}$	SL $\epsilon_{0,n,\gamma}$	SSL $\epsilon_{m,n,\gamma}$	SSL helps?
	$n^{-\frac{1}{d}} \leq \gamma$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	no
	$m^{-\frac{1}{d}} \leq \gamma < n^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	yes
	$\lvert\gamma\rvert < m^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{1}{d}}$	$n^{-\frac{1}{d}}$	no

Unlabeled data: now it helps, now it doesn't

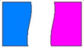
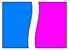
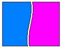
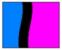
	margin	Oracle $\epsilon_{\infty,n,\gamma}$	SL $\epsilon_{0,n,\gamma}$	SSL $\epsilon_{m,n,\gamma}$	SSL helps?
	$n^{-\frac{1}{d}} \le \gamma$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	no
	$m^{-\frac{1}{d}} \le \gamma < n^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	yes
 	$\lvert\gamma\rvert < m^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{1}{d}}$	$n^{-\frac{1}{d}}$	no
	$\gamma < -m^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	yes

Unlabeled data: now it helps, now it doesn't

	margin	Oracle $\epsilon_{\infty,n,\gamma}$	SL $\epsilon_{0,n,\gamma}$	SSL $\epsilon_{m,n,\gamma}$	SSL helps?
	$n^{-\frac{1}{d}} \leq \gamma$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	no
	$m^{-\frac{1}{d}} \leq \gamma < n^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	yes
 	$\lvert\gamma\rvert < m^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{1}{d}}$	$n^{-\frac{1}{d}}$	no
	$\gamma < -m^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	yes

In particular, with $\gamma < -\gamma_0$, SSL has a faster rate of error convergence than SL, provided $m \gg n^{2d}$.

Unlabeled data: now it helps, now it doesn't

	margin	Oracle $\epsilon_{\infty,n,\gamma}$	SL $\epsilon_{0,n,\gamma}$	SSL $\epsilon_{m,n,\gamma}$	SSL helps?
	$n^{-\frac{1}{d}} \leq \gamma$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	no
	$m^{-\frac{1}{d}} \leq \gamma < n^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	yes
 	$\lvert\gamma\rvert < m^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{1}{d}}$	$n^{-\frac{1}{d}}$	no
	$\gamma < -m^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	$n^{-\frac{1}{d}}$	$n^{-\frac{2\alpha}{2\alpha+d}}$	yes

In particular, with $\gamma < -\gamma_0$, SSL has a faster rate of error convergence than SL, provided $m \gg n^{2d}$.

Thank you

Backup Slides

Hölder Smoothness

If f is Hölder-α, then the $k = \lfloor \alpha \rfloor$ Taylor polynomial at x_0, p_{k,f,x_0}, yields the approximation error bound:

$$|p_{k,f,x_0}(x) - f(x)| \leq C|x - x_0|^\alpha$$

The Corollary

Even when $|\gamma| > Cm^{-1/d}$, the Lemma may fail for two reasons:

- One of the n labeled points or the test point falls in the small uncertain tube.
  - Volume of the tube $O(m^{-1/d})$
  - This is the probability that one point falls in the tube
  - Union bound gives $O(nm^{-1/d})$
  - Risk is bounded
  - The contribution to excess error is $O(nm^{-1/d})$
- With probability $1/m$ connectedness does not imply same decision set
  - The contribution to excess error is $O(1/m)$
- Overall, $O(1/m + nm^{-1/d}) \sim O(nm^{-1/d})$.

The lemma does not apply when $|\gamma| \leq Cm^{-1/d}$.