Some Recent Insights on Transfer Learning

$$P \to Q?$$

Samory Kpotufe
Columbia University

Based on work with Guillaume Martinet, and (ongoing) Steve Hanneke

Transfer Learning:

Given data $\{X_i, Y_i\} \sim_{\text{i.i.d.}} P$, produce a classifier for $(X, Y) \sim Q$.

Case study: Apple Siri's voice assistant

- Initially trained on data from American English speakers ...
- Could not understand 30M+ nonnative speakers in the US!

Costly Solution $\equiv$ **5+ years acquiring more data and retraining!**

A Main Practical Goal:

Cheaply **transfer** ML software between related populations.

Transfer Learning:

Given data $\{X_i, Y_i\} \sim_{\text{i.i.d.}} P$, produce a classifier for $(X, Y) \sim Q$.

Case study: Apple Siri's voice assistant

- Initially trained on data from American English speakers ...
- Could not understand 30M+ nonnative speakers in the US!

Costly Solution $\equiv$ **5+ years acquiring more data and retraining!**

A Main Practical Goal:

Cheaply **transfer** ML software between related populations.

Transfer Learning:

Given data $\{X_i, Y_i\} \sim_{\text{i.i.d.}} P$, produce a classifier for $(X, Y) \sim Q$.

<u>Case study: Apple Siri's voice assistant</u>

- Initially trained on data from American English speakers ...
- Could not understand 30M+ nonnative speakers in the US!

Costly Solution $\equiv$ **5+ years acquiring more data and retraining!**

A Main Practical Goal:

Cheaply **transfer** ML software between related populations.

Transfer is of general relevance:

AI for Judicial Systems

- **Source Population:** prison inmates
- **Target Population:** everyone arrested

Over 60% inaccurate risk assessments on minorities
(2016 Pro-Publica study)

Main Issue: Good Target data is hard or expensive to acquire

AI in medicine, Genomics, Insurance Industry, Smart cities, ...

Transfer is of general relevance:

AI for Judicial Systems

- **Source Population:** prison inmates
- **Target Population:** everyone arrested

Over 60% inaccurate risk assessments on minorities
(2016 Pro-Publica study)

Main Issue: Good Target data is hard or expensive to acquire

AI in medicine, Genomics, Insurance Industry, Smart cities, ...

Transfer is of general relevance:

AI for Judicial Systems

- **Source Population:** prison inmates
- **Target Population:** everyone arrested

Over 60% inaccurate risk assessments on minorities
(2016 Pro-Publica study)

Main Issue: Good Target data is hard or expensive to acquire

AI in medicine, Genomics, Insurance Industry, Smart cities, ...

Transfer is of general relevance:

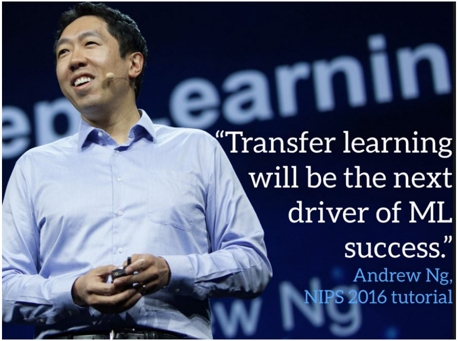
AI for Judicial Systems

- **Source Population:** prison inmates
- **Target Population:** everyone arrested

Over 60% inaccurate risk assessments on minorities
(2016 Pro-Publica study)

Main Issue: Good Target data is hard or expensive to acquire

AI in medicine, Genomics, Insurance Industry, Smart cities, ...



Many heuristics ... but no mature theory or principles

Basic questions remain largely unanswered:

Suppose: $\hat{h}$ is trained on source data $\sim P$, to be transferred to target Q.

- Is there enough information in source P about target Q?
- If not, how much new data should we collect, and how?
- Would unlabeled target data suffice? Or help at least?

Basic questions remain largely unanswered:

Suppose: $\hat{h}$ is trained on source data $\sim P$, to be transferred to target Q.

- Is there enough information in source P about target Q?
- If not, how much new data should we collect, and how?
- Would unlabeled target data suffice? Or help at least?

Basic questions remain largely unanswered:

Suppose: $\hat{h}$ is trained on source data $\sim P$, to be transferred to target Q.

- Is there enough information in source P about target Q?
- If not, how much new data should we collect, and how?
- Would unlabeled target data suffice? Or help at least?

Basic questions remain largely unanswered:

Suppose: $\hat{h}$ is trained on source data $\sim P$, to be transferred to target Q.

- Is there enough information in source P about target Q?
- If not, how much new data should we collect, and how?
- Would unlabeled target data suffice? Or help at least?

Formal Setup:

Covariate-Shift: $P_X \neq Q_X$ but $P_{Y|X} = Q_{Y|X}$.

Given: source data $\{X_i, Y_i\} \sim P^{n_P}$, target data $\{X_i, Y_i\} \sim Q^{n_Q}$.

Goal: $\hat{h}$ with small target error $\text{err}_Q(\hat{h}) = \mathbb{E}_Q\, \mathbb{1}(\hat{h}(X) \neq Y)$.

What is the best $\text{err}_Q(\hat{h})$ achievable in terms of n_P, n_Q?

Depends on distance$(P_X \to Q_X)$

Formal Setup:

Covariate-Shift: $P_X \neq Q_X$ but $P_{Y|X} = Q_{Y|X}$.

For $P_{Y|X} \neq Q_{Y|X}$: [Scott 19], [Blanchard et al. 19], [Cai & Wei 19].

Given: source data $\{X_i, Y_i\} \sim P^{n_P}$, target data $\{X_i, Y_i\} \sim Q^{n_Q}$.

Goal: $\hat{h}$ with small target error $\text{err}_Q(\hat{h}) = \mathbb{E}_Q \, \mathbb{1}(\hat{h}(X) \neq Y)$.

What is the best $\text{err}_Q(\hat{h})$ achievable in terms of n_P, n_Q?

Depends on distance$(P_X \to Q_X)$

Formal Setup:

Covariate-Shift: $P_X \neq Q_X$ but $P_{Y|X} = Q_{Y|X}$.

Given: source data $\{X_i, Y_i\} \sim P^{n_P}$, target data $\{X_i, Y_i\} \sim Q^{n_Q}$.

Goal: $\hat{h}$ with small target error $\text{err}_Q(\hat{h}) = \mathbb{E}_Q \, \mathbb{1}(\hat{h}(X) \neq Y)$.

What is the best $\text{err}_Q(\hat{h})$ achievable in terms of n_P, n_Q?

Depends on distance$(P_X \to Q_X)$

Formal Setup:

Covariate-Shift: $P_X \neq Q_X$ but $P_{Y|X} = Q_{Y|X}$.

Given: source data $\{X_i, Y_i\} \sim P^{n_P}$, target data $\{X_i, Y_i\} \sim Q^{n_Q}$.

Goal: $\hat{h}$ with small target error $\text{err}_Q(\hat{h}) = \mathbb{E}_Q \, \mathbb{1}(\hat{h}(X) \neq Y)$.

What is the best $\text{err}_Q(\hat{h})$ achievable in terms of n_P, n_Q?

Depends on distance$(P_X \to Q_X)$

Formal Setup:

Covariate-Shift: $P_X \neq Q_X$ but $P_{Y|X} = Q_{Y|X}$.

Given: source data $\{X_i, Y_i\} \sim P^{n_P}$, target data $\{X_i, Y_i\} \sim Q^{n_Q}$.

Goal: $\hat{h}$ with small target error $\text{err}_Q(\hat{h}) = \mathbb{E}_Q\, \mathbb{1}(\hat{h}(X) \neq Y)$.

What is the best $\text{err}_Q(\hat{h})$ achievable in terms of n_P, n_Q?

Depends on distance$(P_X \to Q_X)$

Formal Setup:

Covariate-Shift: $P_X \neq Q_X$ but $P_{Y|X} = Q_{Y|X}$.

Given: source data $\{X_i, Y_i\} \sim P^{n_P}$, target data $\{X_i, Y_i\} \sim Q^{n_Q}$.

Goal: $\hat{h}$ with small target error $\text{err}_Q(\hat{h}) = \mathbb{E}_Q \, \mathbb{1}(\hat{h}(X) \neq Y)$.

What is the best $\text{err}_Q(\hat{h})$ achievable in terms of n_P, n_Q?

Depends on distance$(P_X \to Q_X)$

Formal Setup:

Covariate-Shift: $P_X \neq Q_X$ but $P_{Y|X} = Q_{Y|X}$.

Given: source data $\{X_i, Y_i\} \sim P^{n_P}$, target data $\{X_i, Y_i\} \sim Q^{n_Q}$.

Goal: $\hat{h}$ with small target error $\text{err}_Q(\hat{h}) = \mathbb{E}_Q \mathbb{1}(\hat{h}(X) \neq Y)$.

What is the best $\text{err}_Q(\hat{h})$ achievable in terms of n_P, n_Q?

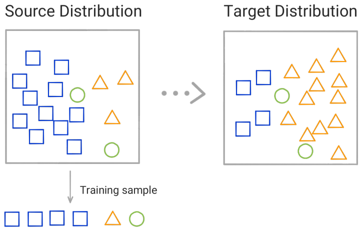
Depends on distance$(P_X \to Q_X)$

However, the right notion of distance$(P_X \to Q_X)$ remains unclear

...

However, the right notion of distance$(P_X \to Q_X)$ remains unclear ...

Basic intuition: Transfer is easiest if P has sufficient mass in regions of large Q-mass.



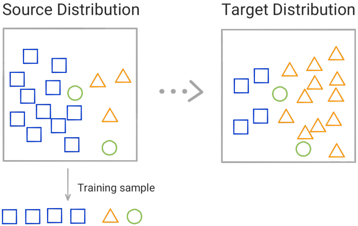
$\neq$ proportions of □ and △

(e.g. $\neq$ proportions of English accents)

However, the right notion of distance$(P_X \to Q_X)$ remains unclear

...

Basic intuition: Transfer is easiest if P has sufficient mass in regions of large Q-mass.

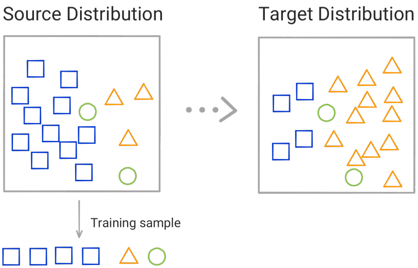


$\neq$ proportions of $\square$ and $\triangle$

(e.g. $\neq$ proportions of English accents)

Many foundational results quantify this intuition ...

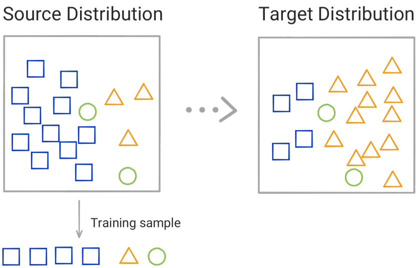
Common notions of distance$(P_X \to Q_X)$:



- **Extensions of TV:** differences $|P_X(A) - Q_X(A)|$, suitable A
  (e.g. $d_{\mathcal{A}}$ divergence/$\mathcal{Y}$-discrepancy of S. Ben David, M. Mohri, ...)
- **Density Ratios:** ratio dQ_X/dP_X over data space
  (e.g., Sugiyama, Belkin, Jordan, Wainwright, ...)

How well do these notions measure transferability?

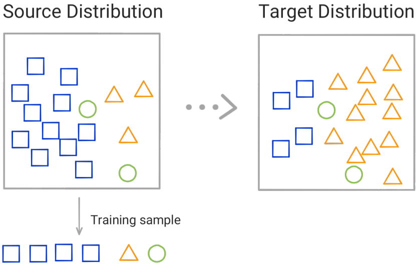
Common notions of distance$(P_X \to Q_X)$:



- **Extensions of TV:** differences $|P_X(A) - Q_X(A)|$, suitable A
  (e.g. $d_{\mathcal{A}}$ divergence/$\mathcal{Y}$-discrepancy of S. Ben David, M. Mohri, ...)
- **Density Ratios:** ratio dQ_X/dP_X over data space
  (e.g., Sugiyama, Belkin, Jordan, Wainwright, ...)

How well do these notions measure transferability?

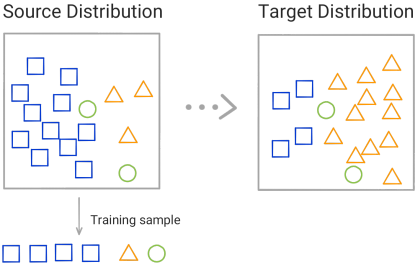
Common notions of distance$(P_X \to Q_X)$:



- **Extensions of TV:** differences $|P_X(A) - Q_X(A)|$, suitable A
  (e.g. $d_{\mathcal{A}}$ divergence/$\mathcal{Y}$-discrepancy of S. Ben David, M. Mohri, ...)
- **Density Ratios:** ratio dQ_X/dP_X over data space
  (e.g., Sugiyama, Belkin, Jordan, Wainwright, ...)

How well do these notions measure transferability?

Common notions of distance($P_X \to Q_X$):



- **Extensions of TV:** differences $|P_X(A) - Q_X(A)|$, suitable A
  (e.g. $d_{\mathcal{A}}$ divergence/$\mathcal{Y}$-discrepancy of S. Ben David, M. Mohri, ...)
- **Density Ratios:** ratio dQ_X/dP_X over data space
  (e.g., Sugiyama, Belkin, Jordan, Wainwright, ...)

How well do these notions measure transferability?

Seminal results: (TV, $d_{\mathcal{A}}$, $\mathcal{Y}$-disc, dQ/dP, KL, Rényi, Wasserstein)

The closer P_X is to Q_X $\implies$ the easier Transfer is ...

However: P_X <u>far</u> from Q_X $\not\Longrightarrow$ Transfer is Hard

These notions can be pessimistic in measuring transferability ...

Seminal results: (TV, $d_{\mathcal{A}}$, $\mathcal{Y}$-disc, dQ/dP, KL, Rényi, Wasserstein)

The closer P_X is to Q_X $\implies$ the easier Transfer is ...

However: P_X **<u>far</u> from** Q_X $\not\Longrightarrow$ **Transfer is Hard**

These notions can be pessimistic in measuring transferability ...

Seminal results: (TV, $d_{\mathcal{A}}$, $\mathcal{Y}$-disc, dQ/dP, KL, Rényi, Wasserstein)

The closer P_X is to Q_X $\Longrightarrow$ the easier Transfer is ...

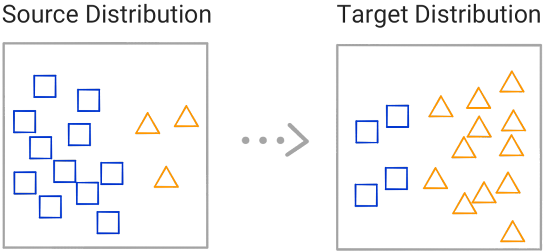
However: P_X **<u>far</u> from** Q_X $\not\Longrightarrow$ **Transfer is Hard**

These notions can be pessimistic in measuring transferability ...

Seminal results: (TV, $d_{\mathcal{A}}$, $\mathcal{Y}$-disc, dQ/dP, KL, Rényi, Wasserstein)

The closer P_X is to Q_X $\Longrightarrow$ the easier Transfer is ...

However: P_X **far from** Q_X $\not\Longrightarrow$ **Transfer is Hard**



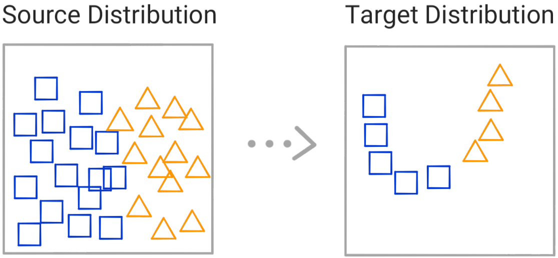
Large TV, $d_{\mathcal{A}}$, $\mathcal{Y}$-disc $\approx 1/2$

These notions can be pessimistic in measuring transferability ...

Seminal results: (TV, $d_{\mathcal{A}}$, $\mathcal{Y}$-disc, dQ/dP, KL, Rényi, Wasserstein)

The closer P_X is to Q_X $\implies$ the easier Transfer is ...

However: P_X **far** **from** Q_X $\cancel{\implies}$ **Transfer is Hard**



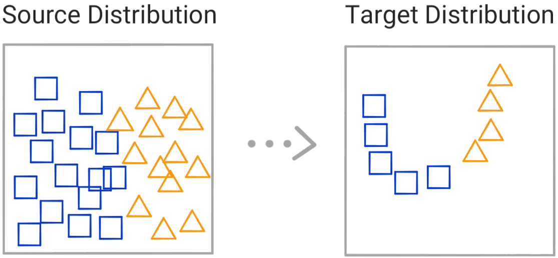
Large dQ/dP, KL-div $\approx \infty$

These notions can be pessimistic in measuring transferability ...

Seminal results: (TV, $d_{\mathcal{A}}$, $\mathcal{Y}$-disc, dQ/dP, KL, Rényi, Wasserstein)

The closer P_X is to Q_X $\Longrightarrow$ the easier Transfer is ...

However: P_X **far** **from** Q_X $\not\Longrightarrow$ **Transfer is Hard**



Large dQ/dP, KL-div $\approx \infty$

These notions can be pessimistic in measuring transferability ...

We propose a new distance$(P \to Q)$ shown to control transfer ...

Relating source P to target Q [Kpo., Martinet, COLT 18]

Main intuition: P_X needs mass in regions of significant Q_X mass.

Transfer exponent $\gamma \geq 0$:

$$\forall^Q x, \forall r \in (0,1], \quad P(B(x,r)) \geq C \cdot r^\gamma \cdot Q_X(B(x,r))$$

γ captures a continuum of easy to hard transfers.

Relating source P to target Q [Kpo., Martinet, COLT 18]

Main intuition: P_X needs mass in regions of significant Q_X mass.

Transfer exponent $\gamma \geq 0$:

$$\forall^Q x, \forall r \in (0,1], \quad P(B(x,r)) \geq C \cdot r^\gamma \cdot Q_X(B(x,r))$$

γ captures a continuum of easy to hard transfer ...

Relating source P to target Q [Kpo., Martinet, COLT 18]

Main intuition: P_X needs mass in regions of significant Q_X mass.

Transfer exponent $\gamma \geq 0$:

$$\forall^Q x, \forall r \in (0,1], \quad P(B(x,r)) \geq C \cdot r^{\gamma} \cdot Q_X(B(x,r))$$

γ captures a continuum of easy to hard transfer ...

Relating source P to target Q [Kpo., Martinet, COLT 18]

Main intuition: P_X needs mass in regions of significant Q_X mass.

Transfer exponent $\gamma \geq 0$:

$$\forall^Q x, \forall r \in (0,1], \quad P(B(x,r)) \geq C \cdot r^{\gamma} \cdot Q_X(B(x,r))$$

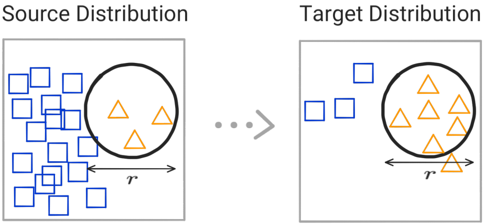
γ **captures a continuum of easy to hard transfer ...**

Relating source P to target Q [Kpo., Martinet, COLT 18]

Main intuition: P_X needs mass in regions of significant Q_X mass.

Transfer exponent $\gamma \geq 0$:

$$\forall^Q x, \forall r \in (0,1], \quad P(B(x,r)) \geq C \cdot r^\gamma \cdot Q_X(B(x,r))$$



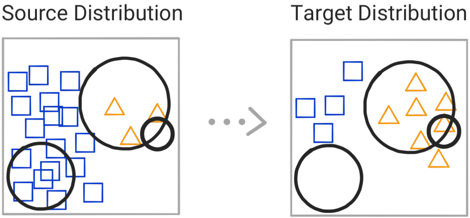
γ captures a continuum of easy to hard transfer ...

Relating source P to target Q [Kpo., Martinet, COLT 18]

Main intuition: P_X needs mass in regions of significant Q_X mass.

Transfer exponent $\gamma \geq 0$:

$$\forall^Q x, \forall r \in (0,1], \quad P(B(x,r)) \geq C \cdot r^\gamma \cdot Q_X(B(x,r))$$



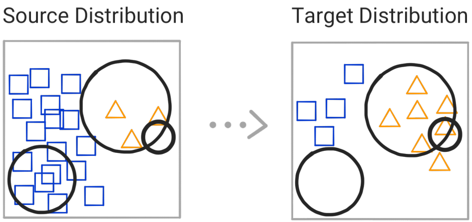
γ captures a continuum of easy to hard transfer ...

Relating source P to target Q [Kpo., Martinet, COLT 18]

Main intuition: P_X needs mass in regions of significant Q_X mass.

Transfer exponent $\gamma \geq 0$:

$$\forall^Q x, \forall r \in (0,1], \quad P(B(x,r)) \geq C \cdot r^{\gamma} \cdot Q_X(B(x,r))$$



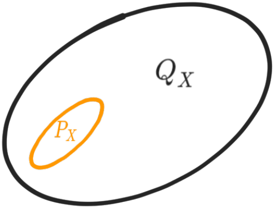
γ captures a continuum of easy to hard transfer ...

First let's look at extremes $\gamma = \infty$ or 0

Notice that $\gamma \doteq \gamma(P \to Q)$ is asymmetric

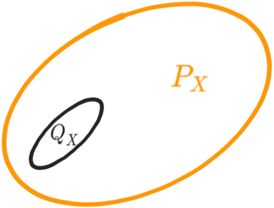
(unlike TV, $d_{\mathcal{A}}$, $\mathcal{Y}$-discrepancy, Wasserstein, ...)

First let's look at extremes $\gamma = \infty$ or 0



$\gamma = \infty$

Bad Coverage



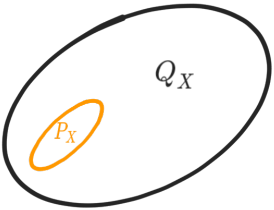
$\gamma = 0$

Good Coverage

Notice that $\gamma \doteq \gamma(P \to Q)$ is asymmetric

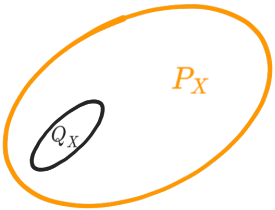
(unlike TV, $d_{\mathcal{A}}$, $\mathcal{Y}$-discrepancy, Wasserstein, ...)

First let's look at extremes $\gamma = \infty$ or 0



$\gamma = \infty$

Bad Coverage



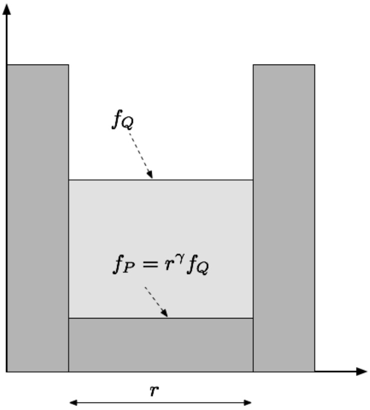
$\gamma = 0$

Good Coverage

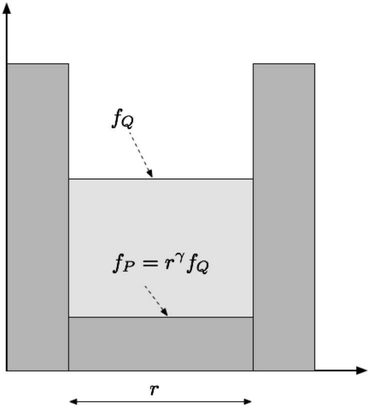
Notice that $\gamma \doteq \gamma(P \to Q)$ is asymmetric

(unlike TV, $d_{\mathcal{A}}$, $\mathcal{Y}$-discrepancy, Wasserstein, ...)

The continuum $0 < \gamma < \infty$:

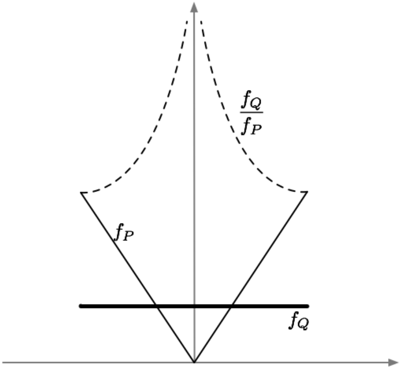


The continuum $0 < \gamma < \infty$:



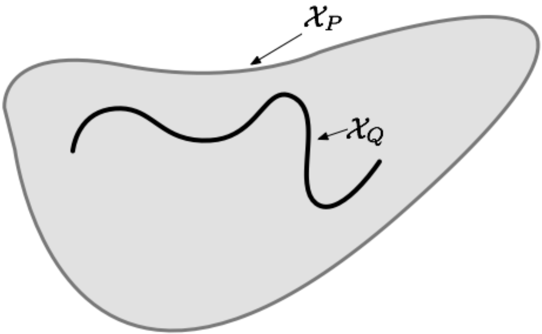
$\gamma \equiv$ How fast P_X shifts mass away from Q_X-dense regions.

The continuum $0 < \gamma < \infty$:



$\gamma \equiv$ How fast f_Q/f_P goes to ∞.

The continuum $0 < \gamma < \infty$:



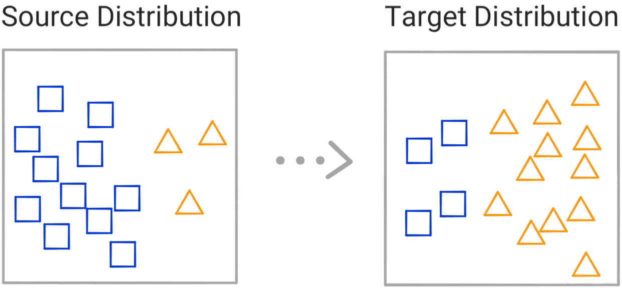
$\gamma \equiv$ Difference in support dimension.

Optimistic:

γ is often small when other measures are not

Optimistic:

γ is often small when other measures are not

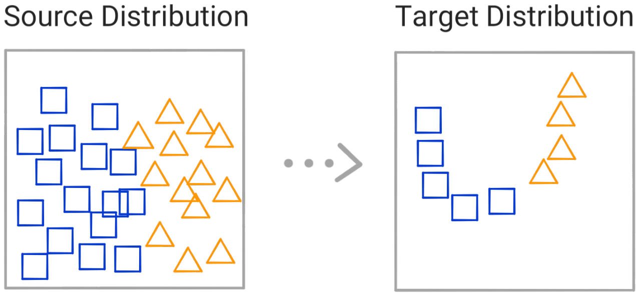


Large TV, $d_{\mathcal{A}}$, $\mathcal{Y}$-disc $\approx 1/2$

but here typically $\gamma \approx 0$

Optimistic:

γ is often small when other measures are not



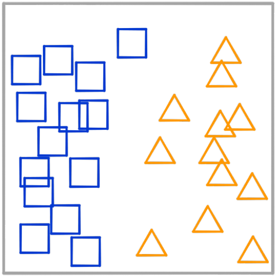
Large dQ_X/dP_X, KL-div $\approx \infty$

but $\gamma = 1 \equiv \dim(P_X) - \dim(Q_X)$

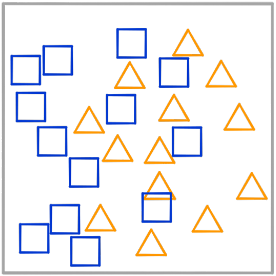
γ captures performance limits (minimax rates) under transfer ...

Performance depends on γ + hardness of Q:

Easy to hard Target $Q_{X,Y}$ classification



Easy $Q_{X,Y}$

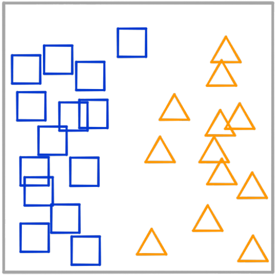


Hard $Q_{X,Y}$

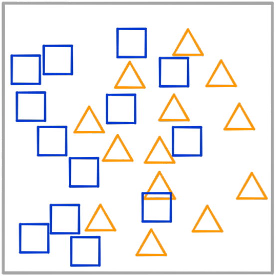
Essential: Noise in $Q_{Y|X}$, and Q_X-mass near decision boundary

Performance depends on γ + hardness of Q:

Easy to hard Target $Q_{X,Y}$ classification



Easy $Q_{X,Y}$

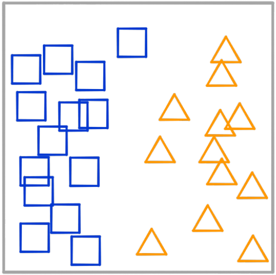


Hard $Q_{X,Y}$

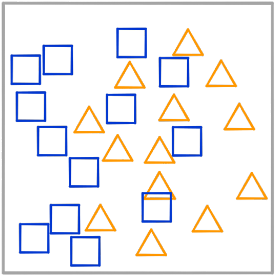
Essential: Noise in $Q_{Y|X}$, and Q_X-mass near decision boundary

Performance depends on γ + hardness of Q:

Easy to hard Target $Q_{X,Y}$ classification



Easy $Q_{X,Y}$



Hard $Q_{X,Y}$

Essential: Noise in $Q_{Y|X}$, and Q_X-mass near decision boundary

Parametrizing easy to hard $Q_{X,Y}$:

Setup: $X \in$ compact $\mathcal{X}$, $Y \in \{0,1\}$

Noise conditions on $\eta(x) = \mathbb{E}[Y|x]$:

- Smoothness: $|\eta(x) - \eta(x')| \le \lambda \rho(x, x')^\alpha$.
- Noise Margin: $Q_X\left(x : |\eta(x) - 1/2| < t\right) \le C t^\beta$.

2 types of regularity on Q_X:

- Near-uniform mass: for any ball B_r, $Q_X(B_r) \ge C r^d$.
- Support regularity: $\mathcal{X}_Q$ has r-cover size $\le C r^{-d}$.

d above acts like the intrinsic dimension of Q_X, for $X \in \mathbb{R}^D$.

Classification is easiest with large α, β, small d .... so is transfer

Parametrizing easy to hard $Q_{X,Y}$:

Setup: $X \in$ compact $\mathcal{X}$, $Y \in \{0,1\}$

Noise conditions on $\eta(x) = \mathbb{E}[Y|x]$:

- Smoothness: $|\eta(x) - \eta(x')| \le \lambda \rho(x,x')^\alpha$.
- Noise Margin: $Q_X\,(x : |\eta(x) - 1/2| < t) \le Ct^\beta$.

2 types of regularity on Q_X:

- Near-uniform mass: for any ball B_r, $Q_X(B_r) \ge Cr^d$.
- Support regularity: $\mathcal{X}_Q$ has r-cover size $\le Cr^{-d}$.

d above acts like the intrinsic dimension of Q_X, for $X \in \mathbb{R}^D$.

Classification is easiest with large α, β, small d .... so is transfer

Parametrizing easy to hard $Q_{X,Y}$:

Setup: $X \in$ compact $\mathcal{X}$, $Y \in \{0, 1\}$

Noise conditions on $\eta(x) = \mathbb{E}[Y|x]$:

- Smoothness: $|\eta(x) - \eta(x')| \leq \lambda\rho(x, x')^\alpha$.
- Noise Margin: $Q_X(x : |\eta(x) - 1/2| < t) \leq Ct^\beta$.

2 types of regularity on Q_X:

- Near-uniform mass: for any ball B_r, $Q_X(B_r) \geq Cr^d$.
- Support regularity: $\mathcal{X}_Q$ has r-cover size $\leq Cr^{-d}$.

d above acts like the intrinsic dimension of Q_X, for $X \in \mathbb{R}^D$.

Classification is easiest with large α, β, small d .... so is transfer

Parametrizing easy to hard $Q_{X,Y}$:

Setup: $X \in$ compact $\mathcal{X}$, $Y \in \{0, 1\}$

Noise conditions on $\eta(x) = \mathbb{E}[Y|x]$*:*

- Smoothness: $|\eta(x) - \eta(x')| \leq \lambda\rho(x, x')^{\alpha}$.
- Noise Margin: $Q_X\left(x : |\eta(x) - 1/2| < t\right) \leq Ct^{\beta}$.

2 types of regularity on Q_X*:*

- Near-uniform mass: for any ball B_r, $Q_X(B_r) \geq Cr^d$.
- Support regularity: $\mathcal{X}_Q$ has r-cover size $\leq Cr^{-d}$.

d above acts like the intrinsic dimension of Q_X, for $X \in \mathbb{R}^D$.

Classification is easiest with large α, β, small d .... so is transfer

Parametrizing easy to hard $Q_{X,Y}$:

Setup: $X \in$ compact $\mathcal{X}$, $Y \in \{0, 1\}$

Noise conditions on $\eta(x) = \mathbb{E}[Y|x]$:

- Smoothness: $|\eta(x) - \eta(x')| \leq \lambda\rho(x, x')^\alpha$.
- Noise Margin: $Q_X\left(x : |\eta(x) - 1/2| < t\right) \leq Ct^\beta$.

2 types of regularity on Q_X:

- Near-uniform mass: for any ball B_r, $Q_X(B_r) \geq Cr^d$.
- Support regularity: $\mathcal{X}_Q$ has r-cover size $\leq Cr^{-d}$.

d above acts like the intrinsic dimension of Q_X, for $X \in \mathbb{R}^D$.

Classification is easiest with large α, β, small d .... so is transfer

Parametrizing easy to hard $Q_{X,Y}$:

Setup: $X \in$ compact $\mathcal{X}$, $Y \in \{0, 1\}$

Noise conditions on $\eta(x) = \mathbb{E}[Y|x]$:

- Smoothness: $|\eta(x) - \eta(x')| \leq \lambda \rho(x, x')^{\alpha}$.
- Noise Margin: $Q_X\,(x : |\eta(x) - 1/2| < t) \leq C t^{\beta}$.

2 types of regularity on Q_X:

- Near-uniform mass: for any ball B_r, $Q_X(B_r) \geq C r^d$.
- Support regularity: $\mathcal{X}_Q$ has r-cover size $\leq C r^{-d}$.

d above acts like the intrinsic dimension of Q_X, for $X \in \mathbb{R}^D$.

Classification is easiest with large α, β, small d .... so is transfer

Parametrizing easy to hard $Q_{X,Y}$:

Setup: $X \in$ compact $\mathcal{X}$, $Y \in \{0, 1\}$

Noise conditions on $\eta(x) = \mathbb{E}[Y|x]$:

- Smoothness: $|\eta(x) - \eta(x')| \leq \lambda\rho(x, x')^{\alpha}$.
- Noise Margin: $Q_X\,(x : |\eta(x) - 1/2| < t) \leq Ct^{\beta}$.

2 types of regularity on Q_X:

- Near-uniform mass: for any ball B_r, $Q_X(B_r) \geq Cr^{d}$.
- Support regularity: $\mathcal{X}_Q$ has r-cover size $\leq Cr^{-d}$.

d above acts like the intrinsic dimension of Q_X, for $X \in \mathbb{R}^D$.

Classification is easiest with large α, β, small d .... so is transfer

Parametrizing easy to hard $Q_{X,Y}$:

Setup: $X \in$ compact $\mathcal{X}$, $Y \in \{0, 1\}$

Noise conditions on $\eta(x) = \mathbb{E}[Y|x]$:

- Smoothness: $|\eta(x) - \eta(x')| \leq \lambda\rho(x, x')^{\alpha}$.
- Noise Margin: $Q_X(x : |\eta(x) - 1/2| < t) \leq Ct^{\beta}$.

2 types of regularity on Q_X:

- Near-uniform mass: for any ball B_r, $Q_X(B_r) \geq Cr^d$.
- Support regularity: $\mathcal{X}_{\mathbf{Q}}$ has r-cover size $\leq Cr^{-d}$.

d above acts like the intrinsic dimension of Q_X, for $X \in \mathbb{R}^D$.

Classification is easiest with large α, β, small d .... so is transfer

Parametrizing easy to hard $Q_{X,Y}$:

Setup: $X \in$ compact $\mathcal{X}$, $Y \in \{0,1\}$

Noise conditions on $\eta(x) = \mathbb{E}[Y|x]$:

- Smoothness: $|\eta(x) - \eta(x')| \leq \lambda \rho(x, x')^{\alpha}$.
- Noise Margin: $Q_X\left(x : |\eta(x) - 1/2| < t\right) \leq C t^{\beta}$.

2 types of regularity on Q_X:

- Near-uniform mass: for any ball B_r, $Q_X(B_r) \geq C r^{d}$.
- Support regularity: $\mathcal{X}_{\mathbf{Q}}$ has r-cover size $\leq C r^{-d}$.

d above acts like the intrinsic dimension of Q_X, for $X \in \mathbb{R}^D$.

Classification is easiest with large α, β, small d .... so is transfer

Parametrizing easy to hard $Q_{X,Y}$:

Setup: $X \in$ compact $\mathcal{X}$, $Y \in \{0, 1\}$

Noise conditions on $\eta(x) = \mathbb{E}[Y|x]$:

- Smoothness: $|\eta(x) - \eta(x')| \leq \lambda\rho(x, x')^{\alpha}$.
- Noise Margin: $Q_X\left(x : |\eta(x) - 1/2| < t\right) \leq Ct^{\beta}$.

2 types of regularity on Q_X:

- Near-uniform mass: for any ball B_r, $Q_X(B_r) \geq Cr^{d}$.
- Support regularity: $\mathcal{X}_{\mathbf{Q}}$ has r-cover size $\leq Cr^{-d}$.

d above acts like the intrinsic dimension of Q_X, for $X \in \mathbb{R}^D$.

Classification is easiest with large α, β, small d .... so is transfer

Minimax rates of Transfer:

Given: labeled source and target data $\{X_i, Y_i\} \sim P^{n_P} \times Q^{n_Q}$.
Excess error: $\mathcal{E}_Q(\hat{h}) \equiv \text{err}_Q(\hat{h}) - \inf_h \text{err}_Q(h)$.

Theorem. Define $\hat{h}$ on $\{X_i, Y_i\}$, even with knowledge of P_X, Q_X:

$$\inf_{\hat{h}} \sup_{\text{dist}(P,Q)=\gamma} \mathbb{E}\, \mathcal{E}_Q(\hat{h}) \approx \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q\right)^{-(\beta+1)/d_0},$$

$d_0 = 2 + d/\alpha$ for near-uniform Q_X, and $d_0 = 2 + \beta + d/\alpha$ otherwise.

Immediate message:
Transfer is easiest as $\gamma \to 0$, hardest as $\gamma \to \infty$ ...

Minimax rates of Transfer:

Given: labeled source and target data $\{X_i, Y_i\} \sim P^{n_P} \times Q^{n_Q}$.
Excess error: $\mathcal{E}_Q(\hat{h}) \equiv \text{err}_Q(\hat{h}) - \inf_h \text{err}_Q(h)$.

Theorem. Define $\hat{h}$ on $\{X_i, Y_i\}$, even with knowledge of P_X, Q_X:

$$\inf_{\hat{h}} \sup_{\text{dist}(P,Q)=\gamma} \mathbb{E}\, \mathcal{E}_Q(\hat{h}) \approx \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q\right)^{-(\beta+1)/d_0},$$

$d_0 = 2 + d/\alpha$ for near-uniform Q_X, and $d_0 = 2 + \beta + d/\alpha$ otherwise.

Immediate message:
Transfer is easiest as $\gamma \to 0$, hardest as $\gamma \to \infty$ ...

Minimax rates of Transfer:

Given: labeled source and target data $\{X_i, Y_i\} \sim P^{n_P} \times Q^{n_Q}$.
Excess error: $\mathcal{E}_Q(\hat{h}) \equiv \text{err}_Q(\hat{h}) - \inf_h \text{err}_Q(h)$.

Theorem. Define $\hat{h}$ on $\{X_i, Y_i\}$, even with knowledge of P_X, Q_X:

$$\inf_{\hat{h}} \sup_{\text{dist}(P,Q)=\gamma} \mathbb{E}\, \mathcal{E}_Q(\hat{h}) \approx \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q\right)^{-(\beta+1)/d_0},$$

$d_0 = 2 + d/\alpha$ for near-uniform Q_X, and $d_0 = 2 + \beta + d/\alpha$ otherwise.

Immediate message:
Transfer is easiest as $\gamma \to 0$, hardest as $\gamma \to \infty$ ...

Minimax rates of Transfer:

Given: labeled source and target data $\{X_i, Y_i\} \sim P^{n_P} \times Q^{n_Q}$.
Excess error: $\mathcal{E}_Q(\hat{h}) \equiv \text{err}_Q(\hat{h}) - \inf_h \text{err}_Q(h)$.

Theorem. Define $\hat{h}$ on $\{X_i, Y_i\}$, even with knowledge of P_X, Q_X:

$$\inf_{\hat{h}} \sup_{\text{dist}(P,Q)=\gamma} \mathbb{E}\, \mathcal{E}_Q(\hat{h}) \approx \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q\right)^{-(\beta+1)/d_0},$$

$d_0 = 2 + d/\alpha$ for near-uniform Q_X, and $d_0 = 2 + \beta + d/\alpha$ otherwise.

Immediate message:

Transfer is easiest as $\gamma \to 0$, hardest as $\gamma \to \infty$ ...

Minimax rates of Transfer:

Given: labeled source and target data $\{X_i, Y_i\} \sim P^{n_P} \times Q^{n_Q}$.
Excess error: $\mathcal{E}_Q(\hat{h}) \equiv \text{err}_Q(\hat{h}) - \inf_h \text{err}_Q(h)$.

Theorem. Define $\hat{h}$ on $\{X_i, Y_i\}$, even with knowledge of P_X, Q_X:

$$\inf_{\hat{h}} \sup_{\text{dist}(P,Q)=\gamma} \mathbb{E}\, \mathcal{E}_Q(\hat{h}) \approx \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q\right)^{-(\beta+1)/d_0},$$

$d_0 = 2 + d/\alpha$ for near-uniform Q_X, and $d_0 = 2 + \beta + d/\alpha$ otherwise.

Immediate message:

Transfer is easiest as $\gamma \to 0$, hardest as $\gamma \to \infty$ ...

Minimax rates of Transfer:

Given: labeled source and target data $\{X_i, Y_i\} \sim P^{n_P} \times Q^{n_Q}$.
Excess error: $\mathcal{E}_Q(\hat{h}) \equiv \text{err}_Q(\hat{h}) - \inf_h \text{err}_Q(h)$.

Theorem. Define $\hat{h}$ on $\{X_i, Y_i\}$, even with knowledge of P_X, Q_X:

$$\inf_{\hat{h}} \sup_{\text{dist}(P,Q)=\gamma} \mathbb{E}\, \mathcal{E}_Q(\hat{h}) \approx \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q\right)^{-(\beta+1)/d_0},$$

$d_0 = 2 + d/\alpha$ for near-uniform Q_X, and $d_0 = 2 + \beta + d/\alpha$ otherwise.

Immediate message:

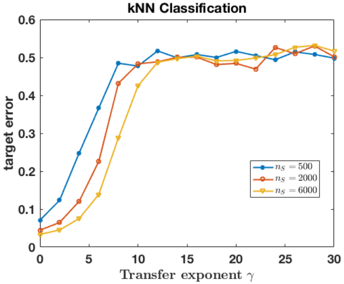
Transfer is easiest as $\gamma \to 0$, hardest as $\gamma \to \infty$ ...

Simulations with increasing γ:

$n_S \equiv$ Source sample size (no target sample used)

Transfer requires more source data for larger γ values.

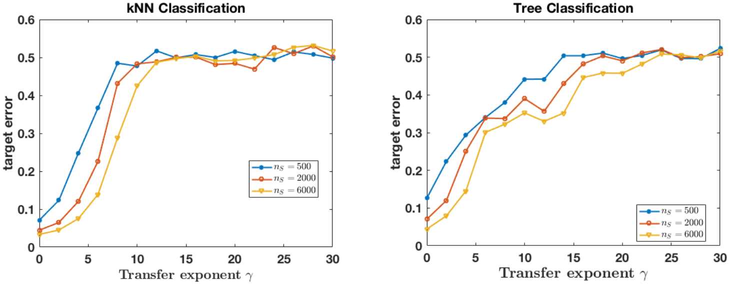
Simulations with increasing γ:



$n_{\mathbf{S}} \equiv$ Source sample size (no target sample used)

Transfer requires more source data for larger γ values.

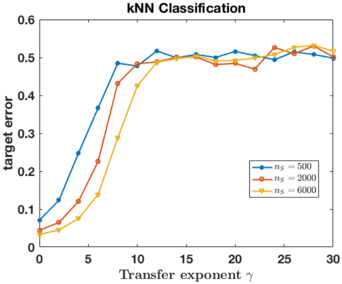
Simulations with increasing γ:

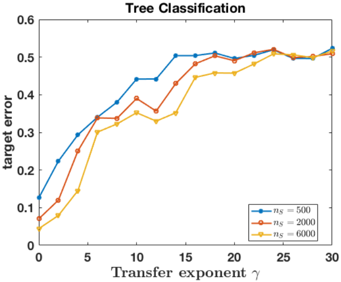


$n_S \equiv$ Source sample size (no target sample used)

Transfer requires more source data for larger γ values.

Simulations with increasing γ:

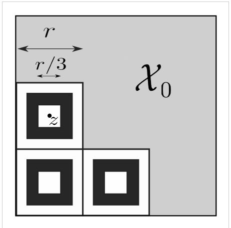




$n_S \equiv$ Source sample size (no target sample used)

Transfer requires more source data for larger γ values.

Lower-bound analysis:

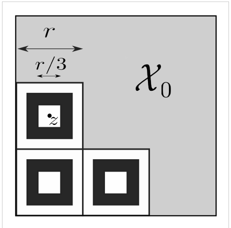


Main Ingredients:

- Established techniques based on Fano's inequality.

- $\hat{h}$ has access to (P, Q) samples, but has to do well on just Q ...

Lower-bound analysis:

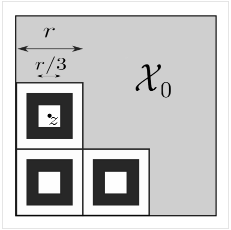


Main Ingredients:

- Established techniques based on Fano's inequality.

- $\hat{h}$ has access to (P, Q) samples, but has to do well on just Q ...

Lower-bound analysis:



Main Ingredients:

- Established techniques based on Fano's inequality.

- $\hat{h}$ has access to (P, Q) samples, but has to do well on just Q ...

Upper-bound analysis:

Rate achieved by k-NN on combined source and target data.

Main ingredient: show that NN distances depend on ρ.

One open problem we had to solve:

Rates for Vanilla k-NN without uniform density assumption

Upper-bound analysis:

Rate achieved by k-NN on combined source and target data.

Main ingredient: show that NN distances depend on ρ.

One open problem we had to solve:

Rates for Vanilla k-NN without uniform density assumption

Upper-bound analysis:

Rate achieved by k-NN on combined source and target data.

Main ingredient: show that NN distances depend on ρ.

One open problem we had to solve:

Rates for Vanilla k-NN without uniform density assumption

Upper-bound analysis:

Rate achieved by k-NN on combined source and target data.

Main ingredient: show that NN distances depend on ρ.

One open problem we had to solve:

Rates for Vanilla k-NN without uniform density assumption

Many new messages:

Recall Minimax Rates $\approx \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q\right)^{-(\beta+1)/d_0}$

- **Fast rates $O(1/n)$ are possible even with large γ.**

- **Target data most beneficial when $n_Q \gg n_P^{d_0/(d_0+\gamma/\alpha)}$.**

- **Unlabeled data does not improve rates beyond constants ...**

Many new messages:

Recall Minimax Rates $\approx \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q \right)^{-(\beta+1)/d_0}$

- **Fast rates $O(1/n)$ are possible even with large γ.**

- **Target data most beneficial when $n_Q \gg n_P^{d_0/(d_0+\gamma/\alpha)}$.**

- **Unlabeled data does not improve rates beyond constants ...**

Many new messages:

Recall Minimax Rates $\approx \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q\right)^{-(\beta+1)/d_0}$

- **Fast rates $O(1/n)$ are possible even with large γ.**

- **Target data most beneficial when $n_Q \gg n_P^{d_0/(d_0+\gamma/\alpha)}$.**

- **Unlabeled data does not improve rates beyond constants ...**

Many new messages:

Recall Minimax Rates $\approx \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q\right)^{-(\beta+1)/d_0}$

- **Fast rates $O(1/n)$ are possible even with large γ.**

- **Target data most beneficial when $n_Q \gg n_P^{d_0/(d_0+\gamma/\alpha)}$.**

- **Unlabeled data does not improve rates beyond constants ...**

Many new messages:

Recall Minimax Rates $\approx \left(n_P^{d_0/(d_0+\gamma/\alpha)} + n_Q\right)^{-(\beta+1)/d_0}$

- **Fast rates $O(1/n)$ are possible even with large γ.**
- **Target data most beneficial when $n_Q \gg n_P^{d_0/(d_0+\gamma/\alpha)}$.**
- **Unlabeled data does not improve rates beyond constants ...**

All good but ...

Can we automatically decide how much target data to sample?

(Ongoing work...)

All good but ...

Can we automatically decide how much target data to sample?

(Ongoing work...)

All good but ...

Can we automatically decide how much target data to sample?

(Ongoing work...)

γ yields insights into adaptive sampling: (ongoing work)

Setup:

n_P labeled samples from P, n_Q unlabeled samples from Q.

Adaptive Sampling:

Sample in low-confidence regions $A \subset \mathcal{X}$ with large $\gamma(A)$.

$(\gamma(A) \leftarrow$ compares P_X and Q_X in region $A)$

Essentially label in Q-massive regions with few samples from P ...

Above refines a procedure of [Berlind, Urner, 15]

γ yields insights into adaptive sampling: (ongoing work)

Setup:

n_P labeled samples from P, n_Q unlabeled samples from Q.

Adaptive Sampling:

Sample in low-confidence regions $A \subset \mathcal{X}$ with large $\gamma(A)$.

$(\gamma(A) \leftarrow$ compares P_X and Q_X in region $A)$

Essentially label in Q-massive regions with few samples from P ...

Above refines a procedure of [Berlind, Urner, 15]

γ yields insights into adaptive sampling: (ongoing work)

Setup:

n_P labeled samples from P, n_Q unlabeled samples from Q.

Adaptive Sampling:

Sample in low-confidence regions $A \subset \mathcal{X}$ with large $\gamma(A)$.

($\gamma(A) \leftarrow$ compares P_X and Q_X in region A)

Essentially label in Q-massive regions with few samples from P ...

Above refines a procedure of [Berlind, Urner, 15]

γ yields insights into adaptive sampling: (ongoing work)

Setup:

n_P labeled samples from P, n_Q unlabeled samples from Q.

Adaptive Sampling:

Sample in low-confidence regions $A \subset \mathcal{X}$ with large $\gamma(A)$.

$(\gamma(A) \leftarrow$ compares P_X and Q_X in region $A)$

Essentially label in Q-massive regions with few samples from P ...

Above refines a procedure of [Berlind, Urner, 15]

γ yields insights into adaptive sampling: (ongoing work)

Setup:

n_P labeled samples from P, n_Q unlabeled samples from Q.

Adaptive Sampling:

Sample in low-confidence regions $A \subset \mathcal{X}$ with large $\gamma(A)$.

$(\gamma(A) \leftarrow$ compares P_X and Q_X in region $A)$

Essentially label in Q-massive regions with few samples from P ...

Above refines a procedure of [Berlind, Urner, 15]

Initial experiments with CIFAR-10:

Transfer Setup: Target Q has 50% images of cats and dogs.

$n_P = 20K$, $n_Q = 6K$ unlabeled Q data

$\hat{h}$: Deep NN with 10 layers, and k-NN on top.

Initial experiments with CIFAR-10:

Transfer Setup: Target Q has 50% images of cats and dogs.

$n_P = 20K$, $n_Q = 6K$ unlabeled Q data

$\hat{h}$: Deep NN with 10 layers, and k-NN on top.

Initial experiments with CIFAR-10:

Transfer Setup: Target Q has 50% images of cats and dogs.

$n_P = 20K$, $n_Q = 6K$ unlabeled Q data

$\hat{h}$: Deep NN with 10 layers, and k-NN on top.

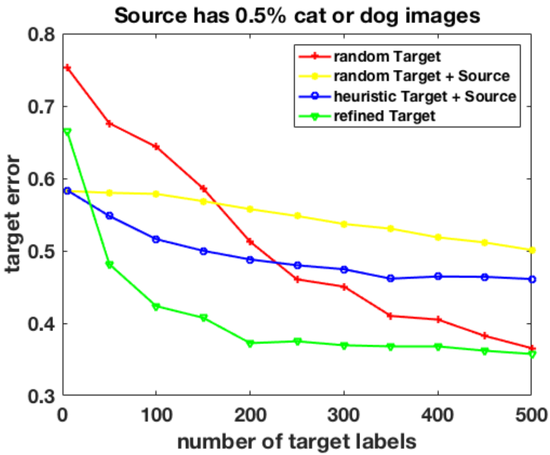
Initial experiments with CIFAR-10:

Transfer Setup: Target Q has 50% images of cats and dogs.

$n_P = 20K$, $n_Q = 6K$ unlabeled Q data

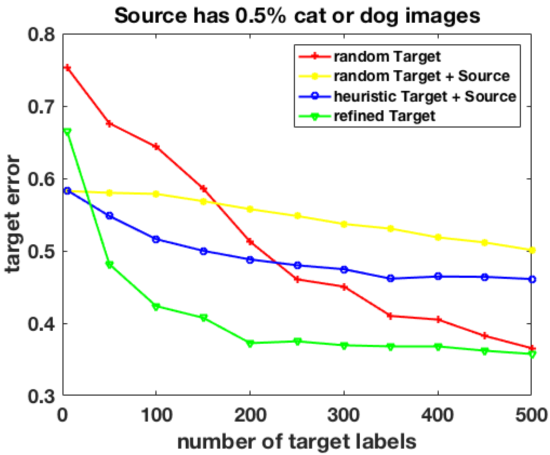
$\hat{h}$: Deep NN with 10 layers, and k-NN on top.

Results



Best performance achieved after relatively few label requests ...

Results



Best performance achieved after relatively few label requests ...

Quick Summary and some New Directions ...

Quick Summary:

- γ captures a more optimistic view of transferability $P \to Q$.
- Unlabeled data can only improve constants in the rates.
- Adaptive sampling is possible with no knowledge of γ.

Quick Summary:

- γ captures a more optimistic view of transferability $P \to Q$.
- Unlabeled data can only improve constants in the rates.
- Adaptive sampling is possible with no knowledge of γ.

Quick Summary:

- γ captures a more optimistic view of transferability $P \to Q$.
- Unlabeled data can only improve constants in the rates.
- Adaptive sampling is possible with no knowledge of γ.

Quick Summary:

- γ captures a more optimistic view of transferability $P \to Q$.
- Unlabeled data can only improve constants in the rates.
- Adaptive sampling is possible with no knowledge of γ.

Quick Summary:

- γ captures a more optimistic view of transferability $P \to Q$.
- Unlabeled data can only improve constants in the rates.
- Adaptive sampling is possible with no knowledge of γ.

New direction: refining γ ...

Often in practice, a family $\mathcal{H}$ of predictors is fixed
(NN, trees, SVMs, Neural Nets ...)

Intuition:
Consider regions of $\mathcal{X}$ most relevant to $\mathcal{H}$ (with S. Hanneke)

This yields $\mathcal{H}$ specific performance limits ...

New direction: refining γ ...

Often in practice, a family $\mathcal{H}$ of predictors is fixed
(NN, trees, SVMs, Neural Nets ...)

Intuition:
Consider regions of $\mathcal{X}$ most relevant to $\mathcal{H}$ (with S. Hanneke)

This yields $\mathcal{H}$ specific performance limits ...

New direction: refining γ ...

Often in practice, a family $\mathcal{H}$ of predictors is fixed
(NN, trees, SVMs, Neural Nets ...)

Intuition:
Consider regions of $\mathcal{X}$ most relevant to $\mathcal{H}$ (with S. Hanneke)

This yields $\mathcal{H}$ specific performance limits ...

New direction: refining γ ...

Often in practice, a family $\mathcal{H}$ of predictors is fixed
(NN, trees, SVMs, Neural Nets ...)

Intuition:
Consider regions of $\mathcal{X}$ most relevant to $\mathcal{H}$ (with S. Hanneke)

This yields $\mathcal{H}$ specific performance limits ...

New direction: refining γ ...

Often in practice, a family $\mathcal{H}$ of predictors is fixed
(NN, trees, SVMs, Neural Nets ...)

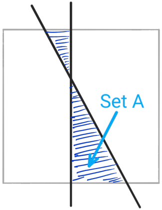
Intuition:

Consider regions of $\mathcal{X}$ most relevant to $\mathcal{H}$ (with S. Hanneke)

In particular:

Consider disagreements between classifiers

$$\gamma : \quad P_X(h \neq h') \gtrsim Q_X(h \neq h')^{1+\gamma}$$



Set A where 2 half-spaces disagree

This yields $\mathcal{H}$ specific performance limits ...

New direction: refining γ ...

Often in practice, a family $\mathcal{H}$ of predictors is fixed
(NN, trees, SVMs, Neural Nets ...)

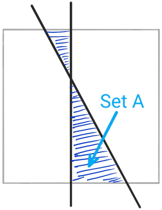
Intuition:

Consider regions of $\mathcal{X}$ most relevant to $\mathcal{H}$ (with S. Hanneke)

In particular:

Consider disagreements between classifiers

$$\gamma : \quad P_X(h \neq h') \gtrsim Q_X(h \neq h')^{1+\gamma}$$



Set A where 2 half-spaces disagree

This yields $\mathcal{H}$ specific performance limits ...

New Messages: [Kpo. & Hanneke, 19]

Near optimal heuristics for bounded VC classes:
(no need to estimate γ)

No Classification noise:
ERM on combined source and target data is minimax-optimal.

Any Level of Noise:
Minimize $\hat{R}_Q(h)$ subject to $\hat{R}_P(h) \leq \min_{h'} \hat{R}_P(h') + \Delta_{n_P}(h)$

☹ Hard to implement in general...

New Messages: [Kpo. & Hanneke, 19]

Near optimal heuristics for bounded VC classes:
(no need to estimate γ)

No Classification noise:
ERM on combined source and target data is minimax-optimal.

Any Level of Noise:
Minimize $\hat{R}_Q(h)$ subject to $\hat{R}_P(h) \leq \min_{h'} \hat{R}_P(h') + \Delta_{n_P}(h)$

☹ Hard to implement in general...

New Messages: [Kpo. & Hanneke, 19]

Near optimal heuristics for bounded VC classes:

(no need to estimate γ)

No Classification noise:

ERM on combined source and target data is minimax-optimal.

Any Level of Noise:

Minimize $\hat{R}_Q(h)$ subject to $\hat{R}_P(h) \leq \min_{h'} \hat{R}_P(h') + \Delta_{n_P}(h)$

☹ Hard to implement in general...

New Messages: [Kpo. & Hanneke, 19]

Near optimal heuristics for bounded VC classes:
(no need to estimate γ)

No Classification noise:

ERM on combined source and target data is minimax-optimal.

Any Level of Noise:

Minimize $\hat{R}_Q(h)$ subject to $\hat{R}_P(h) \leq \min_{h'} \hat{R}_P(h') + \Delta_{n_P}(h)$

☹ Hard to implement in general...

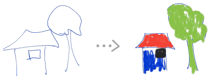
Somehow we are still just scratching the surface of what's possible

...

Results extend beyond covariate-shift to $P_{Y|X} \neq Q_{Y|X}$

Mostly Open:

- More complex transfer regimes?
- Multitask, Curriculum, Lifelong, Fairness, Robustness ?



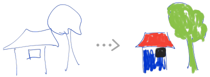
Thanks!

Somehow we are still just scratching the surface of what's possible ...

Results extend beyond covariate-shift to $P_{Y|X} \neq Q_{Y|X}$

Mostly Open:

- More complex transfer regimes?
- Multitask, Curriculum, Lifelong, Fairness, Robustness ?



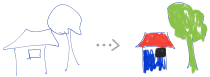
Thanks!

Somehow we are still just scratching the surface of what's possible ...

Results extend beyond covariate-shift to $P_{Y|X} \neq Q_{Y|X}$

Mostly Open:

- More complex transfer regimes?
- Multitask, Curriculum, Lifelong, Fairness, Robustness ?



Thanks!

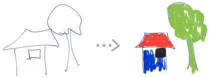
Somehow we are still just scratching the surface of what's possible

...

Results extend beyond covariate-shift to $P_{Y|X} \neq Q_{Y|X}$

Mostly Open:

- More complex transfer regimes?
- Multitask, Curriculum, Lifelong, Fairness, Robustness ?



Thanks!

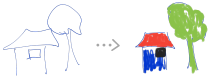
Somehow we are still just scratching the surface of what's possible

...

Results extend beyond covariate-shift to $P_{Y|X} \neq Q_{Y|X}$

Mostly Open:

- More complex transfer regimes?
- Multitask, Curriculum, Lifelong, Fairness, Robustness ?



Thanks!