Adaptivity in Domain Adaptation and Friends

$$P + Q \to Q?$$

Samory Kpotufe
Columbia University, Statistics

Based on various works with **G. Martinet**, **S. Hanneke** and **J. Suk**

Domain Adaptation (or 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.

Domain Adaptation (or 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.

Domain Adaptation (or 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:

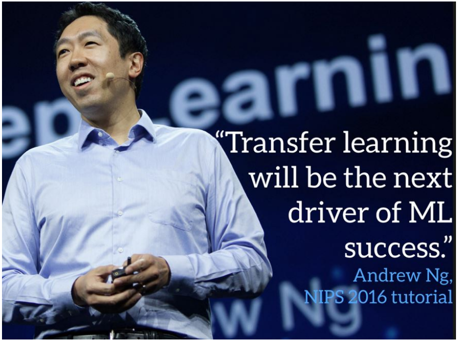
AI for Judicial Systems

- **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 theory and principles are still evolving

Basic questions remain largely unanswered:

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

- Is there sufficient information in source P about target Q?
- If not, how much new data should be collected?
- Would unlabeled data help?
- What's the right mix of P and Q data w.r.t. $$ sampling costs?

What's the relative statistical value of P and Q data?

Depends on how *far* P is from Q ...

Basic questions remain largely unanswered:

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

- Is there sufficient information in source P about target Q?
- If not, how much new data should be collected?
- Would unlabeled data help?
- What's the right mix of P and Q data w.r.t. $$ sampling costs?

What's the relative statistical value of P and Q data?

Depends on how *far* P is from Q ...

Basic questions remain largely unanswered:

Suspect:

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

- Is there sufficient information in source P about target Q?
- If not, how much new data should be collected?
- Would unlabeled data help?
- What's the right mix of P and Q data w.r.t. $$ sampling costs?

What's the relative statistical value of P and Q data?

Depends on how *far* P is from Q ...

Basic questions remain largely unanswered:

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

- Is there sufficient information in source P about target Q?
- If not, how much new data should be collected?
- Would unlabeled data help?
- What's the right mix of P and Q data w.r.t. $$ sampling costs?

What's the relative statistical value of P and Q data?

Depends on how *far* P is from Q ...

Basic questions remain largely unanswered:

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

- Is there sufficient information in source P about target Q?
- If not, how much new data should be collected?
- Would unlabeled data help?
- What's the right mix of P and Q data w.r.t. $$ sampling costs?

What's the relative statistical value of P and Q data?

Depends on how *far* P is from Q ...

Basic questions remain largely unanswered:

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

- Is there sufficient information in source P about target Q?
- If not, how much new data should be collected?
- Would unlabeled data help?
- What's the right mix of P and Q data w.r.t. $$ sampling costs?

What's the relative statistical value of P and Q data?

Depends on how *far* P is from Q ...

Basic questions remain largely unanswered:

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

- Is there sufficient information in source P about target Q?
- If not, how much new data should be collected?
- Would unlabeled data help?
- What's the right mix of P and Q data w.r.t. \$\$ sampling costs?

What's the relative statistical value of P and Q data?

Depends on how *far* P is from Q ...

How do we proceed?

Formal Setup:
Classification $X \mapsto Y$, fixed VC class $\mathcal{H}$

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

Goal: $\hat{h} \in \mathcal{H}$ with small *excess* target error

$$\mathcal{E}_Q(\hat{h}) = \mathbb{E}_Q\left[\hat{h}(X) \neq Y\right] - \inf_{h \in \mathcal{H}} \mathbb{E}_Q\left[h(X) \neq Y\right]$$

Basic Information-theoretic Question:

Which $\mathcal{E}_Q(\hat{h})$ is achievable in terms of sample sizes n_P and n_Q?

Which notion of dist$(P \to Q)$ captures this error?

How do we proceed?

Nonparametric work

- (Covariate Shift) [Kpo. and Martinet, AoS 21]
- (Posterior Drift) [Scott 19] [Cai and Wei, AoS 19]
- (Covariate Shift, Posterior Drift) [Reeve, Cannings, Samworth, AoS 21]
- (Covariate Shift) [Pathak, Ma, Wainwright, ICML 22]

Formal Setup:

Classification $X \mapsto Y$, fixed VC class $\mathcal{H}$

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

Goal: $\hat{h} \in \mathcal{H}$ with small *excess* target error

$$\mathcal{E}_Q(\hat{h}) = \mathbb{E}_Q \left[\hat{h}(X) \neq Y\right] - \inf_{h \in \mathcal{H}} \mathbb{E}_Q \left[h(X) \neq Y\right]$$

Basic Information-theoretic Question:

Which $\mathcal{E}_Q(\hat{h})$ is achievable in terms of sample sizes n_P and n_Q?

How do we proceed?

Formal Setup:
Classification $X \mapsto Y$, fixed VC class $\mathcal{H}$

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

Goal: $\hat{h} \in \mathcal{H}$ with small *excess* target error

$$\mathcal{E}_Q(\hat{h}) = \mathbb{E}_Q\left[\hat{h}(X) \neq Y\right] - \inf_{h \in \mathcal{H}} \mathbb{E}_Q\left[h(X) \neq Y\right]$$

Basic Information-theoretic Question:
Which $\mathcal{E}_Q(\hat{h})$ is achievable in terms of sample sizes n_P and n_Q?

Which notion of dist$(P \to Q)$ captures this error?

How do we proceed?

Formal Setup:

Classification $X \mapsto Y$, fixed VC class $\mathcal{H}$

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

Goal: $\hat{h} \in \mathcal{H}$ with small *excess* target error

$$\mathcal{E}_Q(\hat{h}) = \mathbb{E}_Q\left[\hat{h}(X) \neq Y\right] - \inf_{h \in \mathcal{H}} \mathbb{E}_Q\left[h(X) \neq Y\right]$$

Basic Information-theoretic Question:

Which $\mathcal{E}_Q(\hat{h})$ is achievable in terms of sample sizes n_P and n_Q?

Which notion of dist$(P \to Q)$ captures this error?

How do we proceed?

Formal Setup:

Classification $X \mapsto Y$, fixed VC class $\mathcal{H}$

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

Goal: $\hat{h} \in \mathcal{H}$ with small *excess* target error

$$\mathcal{E}_Q(\hat{h}) = \mathbb{E}_Q\left[\hat{h}(X) \neq Y\right] - \inf_{h \in \mathcal{H}} \mathbb{E}_Q\left[h(X) \neq Y\right]$$

Basic Information-theoretic Question:

Which $\mathcal{E}_Q(\hat{h})$ is achievable in terms of sample sizes n_P and n_Q?

Which notion of dist$(P \to Q)$ captures this error?

How do we proceed?

Formal Setup:

Classification $X \mapsto Y$, fixed VC class $\mathcal{H}$

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

Goal: $\hat{h} \in \mathcal{H}$ with small *excess* target error

$$\mathcal{E}_Q(\hat{h}) = \mathbb{E}_Q\left[\hat{h}(X) \neq Y\right] - \inf_{h \in \mathcal{H}} \mathbb{E}_Q\left[h(X) \neq Y\right]$$

Basic Information-theoretic Question:

Which $\mathcal{E}_Q(\hat{h})$ is achievable in terms of sample sizes n_P and n_Q?

Which notion of dist$(P \to Q)$ captures this error?

Similar Questions in Regression, RL & Bandits (even harder) ...

(Classification) Many competing notions of dist$(P \to Q)$ ...

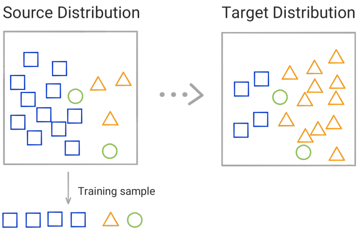
- **Extensions of TV:** consider $|P(A) - Q(A)|$ over suitable A
  (e.g. $d_{\mathcal{A}}$ divergence/$\mathcal{Y}$-discrepancy of S. Ben David, M. Mohri, ...)

$$\mathcal{E}_Q(\hat{h}) \lesssim o_P(1) + \text{dist}(P \to Q)$$

- **Density Ratios:** consider ratio dQ/dP over data space
  (e.g., Sugiyama, Belkin, Jordan, Wainwright, ...)

$$\mathcal{E}_Q(\hat{h}) \lesssim o_P(1) + \text{estimation error}(d_Q/d_P)$$

(Classification) Many competing notions of dist$(P \to Q)$ ...



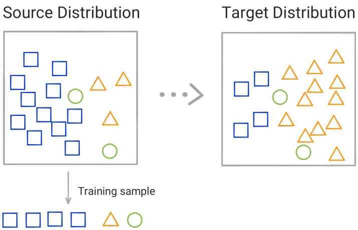
- **Extensions of TV:** consider $|P(A) - Q(A)|$ over suitable A
  (e.g. $d_{\mathcal{A}}$ divergence/$\mathcal{Y}$-discrepancy of S. Ben David, M. Mohri, ...)

$$\mathcal{E}_Q(\hat{h}) \lesssim o_P(1) + \text{dist}(P \to Q)$$

- **Density Ratios:** consider ratio dQ/dP over data space
  (e.g., Sugiyama, Belkin, Jordan, Wainwright, ...)

$$\mathcal{E}_Q(\hat{h}) \lesssim o_P(1) + \text{estimation error}(d_Q/d_P)$$

(Classification) Many competing notions of dist$(P \to Q)$ ...



- **Extensions of TV:** consider $|P(A) - Q(A)|$ over suitable A
  (e.g. $d_{\mathcal{A}}$ divergence/$\mathcal{Y}$-discrepancy of S. Ben David, M. Mohri, ...)

$$\mathcal{E}_Q(\hat{h}) \lesssim o_P(1) + \text{dist}(P \to Q)$$

- **Density Ratios:** consider ratio dQ/dP over data space
  (e.g., Sugiyama, Belkin, Jordan, Wainwright, ...)

$$\mathcal{E}_Q(\hat{h}) \lesssim o_P(1) + \text{estimation error}(d_Q/d_P)$$

Many notions: (TV, $d_{\mathcal{A}}$, $\mathcal{Y}$-disc, KL, Renyi, MMD, Wasserstein ...)

They all tend to be over-pessimistic about transfer ☹

Namely: P **far from** Q $\not\Longrightarrow$ **Transfer is Hard**

Many notions: (TV, $d_{\mathcal{A}}$, $\mathcal{Y}$-disc, KL, Renyi, MMD, Wasserstein ...)

They all tend to be over-pessimistic about transfer ☹

Namely: P **far from** Q $\not\Longrightarrow$ **Transfer is Hard**

Many notions: (TV, $d_{\mathcal{A}}$, $\mathcal{Y}$-disc, KL, Renyi, MMD, Wasserstein ...)

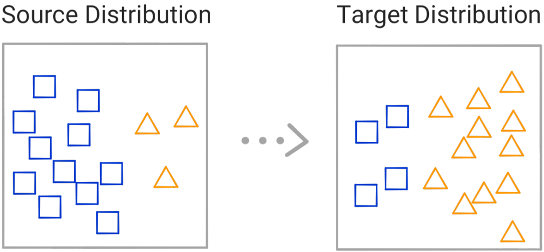
They all tend to be over-pessimistic about transfer ☹

Namely: P **far from** Q $\not\Longrightarrow$ **Transfer is Hard**

Many notions: (TV, $d_{\mathcal{A}}$, $\mathcal{Y}$-disc, KL, Renyi, MMD, Wasserstein ...)

They all tend to be over-pessimistic about transfer ☹

Namely: P **far from** Q $\not\Longrightarrow$ **Transfer is Hard**

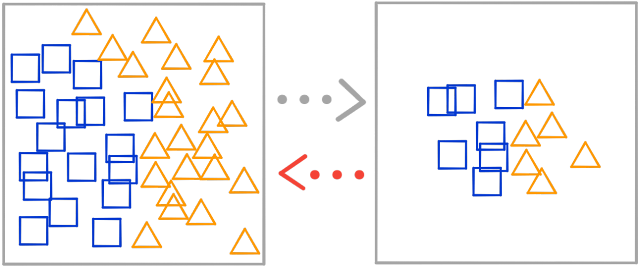


Large TV, $d_{\mathcal{A}}$, $\mathcal{Y}$-disc $\approx 1/2$

Many notions: (TV, $d_{\mathcal{A}}$, $\mathcal{Y}$-disc, KL, Renyi, MMD, Wasserstein ...)

They all tend to be over-pessimistic about transfer ☹

Namely: P **far from** Q $\not\Longrightarrow$ **Transfer is Hard**

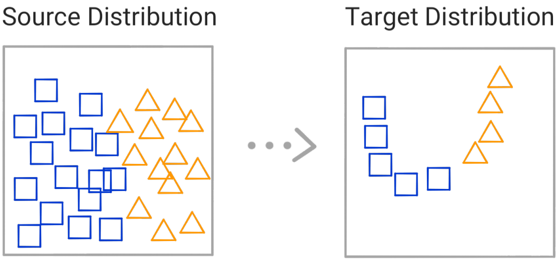


Asymmetry in transfer $\Longrightarrow$ Metrics are inappropriate

Many notions: (TV, $d_{\mathcal{A}}$, $\mathcal{Y}$-disc, KL, Renyi, MMD, Wasserstein ...)

They all tend to be over-pessimistic about transfer ☹

Namely: P **far from** Q $\not\Longrightarrow$ **Transfer is Hard**



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

Relating source P to target Q [Hanneke, Kpo. NeurIPS 19]

Intuition: $h \in \mathcal{H}$ has low error under $P \implies$ low error under Q

For now assume $h_P^* = h_Q^*$ ...

Transfer exponent $\rho > 0$:

$$\forall h \in \mathcal{H}, \quad \mathcal{E}_Q(h) \leq c \cdot \mathcal{E}_P^{1/\rho}(h)$$

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

Relating source P to target Q [Hanneke, Kpo. NeurIPS 19]

Intuition: $h \in \mathcal{H}$ has low error under $P \implies$ low error under Q

For now assume $h_P^* = h_Q^*$ ...

Transfer exponent $\rho > 0$:

$$\forall h \in \mathcal{H}, \quad \mathcal{E}_Q(h) \leq c \cdot \mathcal{E}_P^{1/\rho}(h)$$

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

Relating source P to target Q [Hanneke, Kpo. NeurIPS 19]

Intuition: $h \in \mathcal{H}$ has low error under $P \implies$ low error under Q

For now assume $h_P^* = h_Q^*$ ...

Transfer exponent $\rho > 0$:

$$\forall h \in \mathcal{H}, \quad \mathcal{E}_Q(h) \leq c \cdot \mathcal{E}_P^{1/\rho}(h)$$

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

Relating source P to target Q [Hanneke, Kpo. NeurIPS 19]

Intuition: $h \in \mathcal{H}$ has low error under $P \implies$ low error under Q

For now assume $h_P^* = h_Q^*$ ...

Transfer exponent $\rho > 0$:

$$\forall h \in \mathcal{H}, \quad \mathcal{E}_Q(h) \leq c \cdot \mathcal{E}_P^{1/\rho}(h)$$

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

Relating source P to target Q [Hanneke, Kpo. NeurIPS 19]

Intuition: $h \in \mathcal{H}$ has low error under $P \implies$ low error under Q

For now assume $h^*_P = h^*_Q$ ...

Transfer exponent $\rho > 0$:

$$\forall h \in \mathcal{H}, \quad \mathcal{E}_Q(h) \leq c \cdot \mathcal{E}_P^{1/\rho}(h)$$

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

Examples:

Transfer exponent $\rho > 0$:

$$\forall h \in \mathcal{H}, \quad \mathcal{E}_Q(h, h^*) \le c \cdot \mathcal{E}_P^{1/\rho}(h, h^*)$$

Examples:

Transfer exponent $\rho > 0$:

$$\forall h \in \mathcal{H}, \quad \mathcal{E}_Q(h, h^*) \leq c \cdot \mathcal{E}_P^{1/\rho}(h, h^*)$$

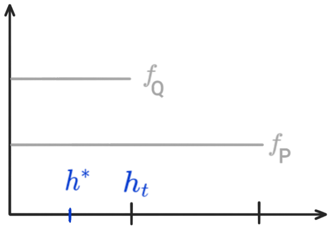
For deterministic $Y = h^*(X)$ this reduces to:

$$Q_X(h \neq h^*) \leq c \cdot {P_X}^{1/\rho}(h \neq h^*)$$

Examples:

Transfer exponent $\rho > 0$:

$$\forall h \in \mathcal{H}, \quad \mathcal{E}_Q(h, h^*) \le c \cdot \mathcal{E}_P^{1/\rho}(h, h^*)$$

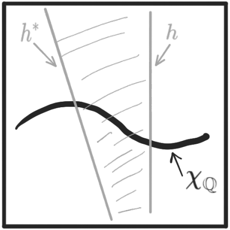


$\rho = 1$ but $d_{\mathcal{A}}(P, Q) = \mathcal{Y}$-disc$(P, Q) = 1/4$

Examples:

Transfer exponent $\rho > 0$:

$$\forall h \in \mathcal{H}, \quad \mathcal{E}_Q(h, h^*) \le c \cdot \mathcal{E}_P^{1/\rho}(h, h^*)$$

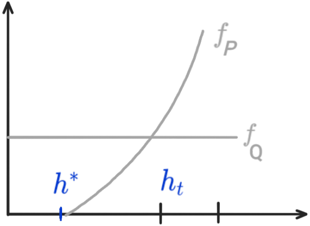


$\rho = 1$ but KL, Renyi, blow up ...

Examples:

Transfer exponent $\rho > 0$:

$$\forall h \in \mathcal{H}, \quad \mathcal{E}_Q(h, h^*) \leq c \cdot \mathcal{E}_P^{1/\rho}(h, h^*)$$

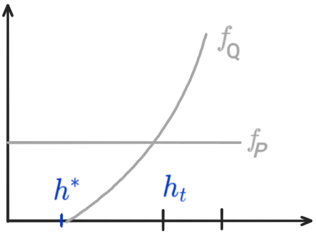


$\rho > 1 \equiv$ how much P covers decision boundary

Examples:

Transfer exponent $\rho > 0$:

$$\forall h \in \mathcal{H}, \quad \mathcal{E}_Q(h, h^*) \leq c \cdot \mathcal{E}_P^{1/\rho}(h, h^*)$$

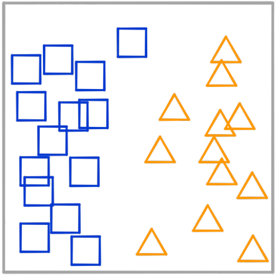


$0 < \rho < 1 \equiv$ Super Transfer (P has better coverage of decision boundary)

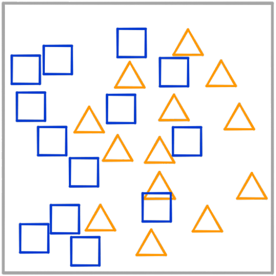
ρ captures performance limits (minimax rates) under transfer ...

Performance depends on ρ + hardness of classification:

Easy to hard classification



Easy Classification



Hard Classification

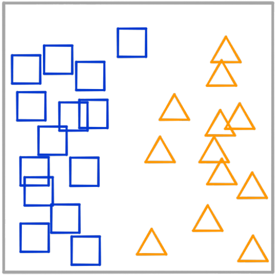
Essential: Noise in $Y|X$, and X-mass near decision boundary

Bernstein condition: $\quad Q_X(h \neq h^*) \lesssim \mathcal{E}_Q{}^{\beta}(h; h^*), \quad \beta \in [0, 1]$

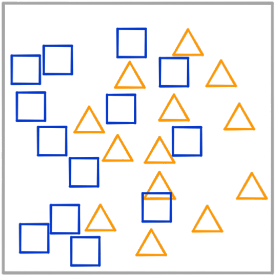
Similar noise condition on P.

Performance depends on ρ + hardness of classification:

Easy to hard classification



Easy Classification



Hard Classification

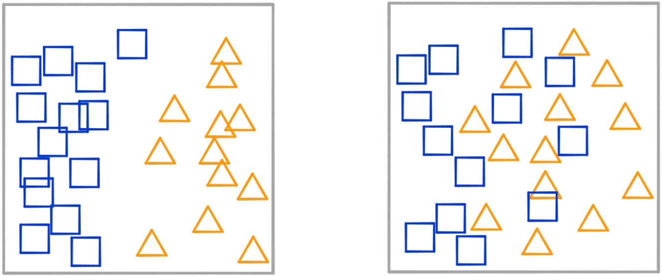
Essential: Noise in $Y|X$, and X-mass near decision boundary

Bernstein condition: $Q_X(h \neq h^*) \lesssim \mathcal{E}_Q{}^{\beta}(h; h^*), \quad \beta \in [0, 1]$

Similar noise condition on P.

Performance depends on ρ + hardness of classification:

Easy to hard classification



Easy Classification   Hard Classification

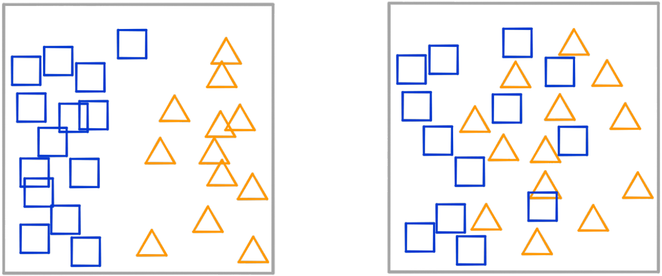
Essential: Noise in $Y|X$, and X-mass near decision boundary

Bernstein condition: $Q_X(h \neq h^*) \lesssim {\mathcal{E}_Q}^{\beta}(h; h^*), \quad \beta \in [0, 1]$

Similar noise condition on P.

Performance depends on ρ + hardness of classification:

Easy to hard classification



Easy Classification   Hard Classification

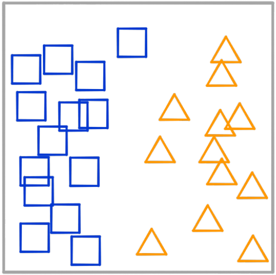
Essential: Noise in $Y|X$, and X-mass near decision boundary

Bernstein condition: $Q_X(h \neq h^*) \lesssim \mathcal{E}_Q{}^{\beta}(h; h^*), \quad \beta \in [0,1]$

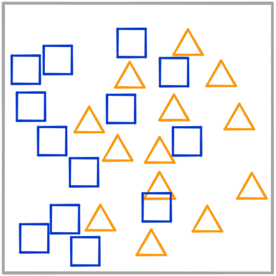
Similar noise condition on P.

Performance depends on ρ + hardness of classification:

Easy to hard classification



Easy Classification



Hard Classification

Essential: Noise in $Y|X$, and X-mass near decision boundary

Bernstein condition: $Q_X(h \neq h^*) \lesssim \mathcal{E}_Q{}^{\beta}(h; h^*), \quad \beta \in [0, 1]$

Similar noise condition on P.

Minimax rates of Transfer: [Hanneke, Kpo. NeurIPS 19]

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

Theorem. Let $\hat{h}$ trained on samples from $P + Q$:

$$\inf_{\hat{h}} \sup_{(P,Q)} \mathcal{E}_Q(\hat{h}) \propto \left(n_P^{1/\rho} + n_Q\right)^{-1/(2-\beta)}$$

Tight for any $\mathcal{H}, \rho \geq 1, \beta, n_P, n_Q$ ...

- **Benefits of Unlabeled data:** cannot improve the rates ...
- **Benefits of Labeled Q data:** transition at $n_Q > n_P^{1/\rho}$
- **Adaptive sampling at optimal $$ costs:** possible in some regimes

Minimax rates of Transfer: [Hanneke, Kpo. NeurIPS 19]

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

Theorem. Let $\hat{h}$ trained on samples from $P + Q$:

$$\inf_{\hat{h}} \sup_{(P,Q)} \mathcal{E}_Q(\hat{h}) \propto \left(n_P^{1/\rho} + n_Q\right)^{-1/(2-\beta)}$$

Tight for any $\mathcal{H}, \rho \geq 1, \beta, n_P, n_Q$ ...

- **Benefits of Unlabeled data:** cannot improve the rates ...
- **Benefits of Labeled Q data:** transition at $n_Q > n_P^{1/\rho}$
- **Adaptive sampling at optimal $$ costs:** possible in some regimes

Minimax rates of Transfer: [Hanneke, Kpo. NeurIPS 19]

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

Theorem. Let $\hat{h}$ trained on samples from $P + Q$:

$$\inf_{\hat{h}} \sup_{(P,Q)} \mathcal{E}_Q(\hat{h}) \propto \left(n_P^{1/\rho} + n_Q\right)^{-1/(2-\beta)}$$

Tight for any $\mathcal{H}, \rho \geq 1, \beta, n_P, n_Q$ ...

- **Benefits of Unlabeled data:** cannot improve the rates ...
- **Benefits of Labeled Q data:** transition at $n_Q > n_P^{1/\rho}$
- **Adaptive sampling at optimal** $\$\$$ **costs:** possible in some regimes

Minimax rates of Transfer: [Hanneke, Kpo. NeurIPS 19]

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

Theorem. Let $\hat{h}$ trained on samples from $P + Q$:

$$\inf_{\hat{h}} \sup_{(P,Q)} \mathcal{E}_Q(\hat{h}) \propto \left(n_P^{1/\rho} + n_Q\right)^{-1/(2-\beta)}$$

Tight for any $\mathcal{H}, \rho \geq 1, \beta, n_P, n_Q$ ...

- **Benefits of Unlabeled data:** cannot improve the rates ...
- **Benefits of Labeled Q data:** transition at $n_Q > n_P^{1/\rho}$
- **Adaptive sampling at optimal $$ costs:** possible in some regimes

Minimax rates of Transfer: [Hanneke, Kpo. NeurIPS 19]

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

Theorem. Let $\hat{h}$ trained on samples from $P + Q$:

$$\inf_{\hat{h}} \sup_{(P,Q)} \mathcal{E}_Q(\hat{h}) \propto \left(n_P^{1/\rho} + n_Q\right)^{-1/(2-\beta)}$$

Tight for any $\mathcal{H}, \rho \geq 1, \beta, n_P, n_Q$ ...

- **Benefits of Unlabeled data:** cannot improve the rates ...
- **Benefits of Labeled Q data:** transition at $n_Q > n_P^{1/\rho}$
- **Adaptive sampling at optimal $\$\$$ costs:** possible in some regimes

Lower-Bound Analysis

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

Construction: family $\{(P, Q)_h\}$, any $\mathcal{H}$, $\rho \geq 1$, β:

- $(P^{n_P} \times Q^{n_Q})_h$ are close in KL-divergence
- But far under distance $Q_h(h' \neq h)$

The rest is extensions of Fano (see e.g. Tsybakov, or Barron and Li) ...

Lower-Bound Analysis

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

Construction: family $\{(P, Q)_h\}$, any $\mathcal{H}$, $\rho \geq 1$, β:

- $(P^{n_P} \times Q^{n_Q})_h$ are close in KL-divergence
- But far under distance $Q_h(h' \neq h)$

The rest is extensions of Fano (see e.g. Tsybakov, or Barron and Li) ...

Lower-Bound Analysis

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

Construction: family $\{(P, Q)_h\}$, any $\mathcal{H}$, $\rho \geq 1$, β:

- $(P^{n_P} \times Q^{n_Q})_h$ are close in KL-divergence
- But far under distance $Q_h(h' \neq h)$

The rest is extensions of Fano (see e.g. Tsybakov, or Barron and Li) ...

Upper-bound Analysis:

Performance limits: $\mathcal{E}_Q(\hat{h}) \propto \left(n_P^{1/\rho} + n_Q\right)^{-1/(2-\beta)}$

(Optimal Heuristics for unknown ρ)

Low Classification noise ($\beta = 1$):

ERM on combined source and target data.

Non i.i.d. Bernstein + usual fixed point argument

Unknown Noise Level ($\beta \in [0, 1]$):

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

Lepski-type argument

Upper-bound Analysis:

Performance limits: $\mathcal{E}_Q(\hat{h}) \propto \left(n_P^{1/\rho} + n_Q\right)^{-1/(2-\beta)}$

We are interested in *adaptivity* to ρ ...

(Optimal Heuristics for unknown ρ)

Low Classification noise ($\beta = 1$):

ERM on combined source and target data.

Non i.i.d. Bernstein + usual fixed point argument

Unknown Noise Level ($\beta \in [0, 1]$):

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

Lepski-type argument

Upper-bound Analysis:

Performance limits: $\mathcal{E}_Q(\hat{h}) \propto \left(n_P^{1/\rho} + n_Q\right)^{-1/(2-\beta)}$

(Optimal Heuristics for unknown ρ)

Low Classification noise ($\beta = 1$):

ERM on combined source and target data.

Non i.i.d. Bernstein + usual fixed point argument

Unknown Noise Level ($\beta \in [0,1]$):

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

Lepski-type argument

Upper-bound Analysis:

Performance limits: $\mathcal{E}_Q(\hat{h}) \propto \left(n_P^{1/\rho} + n_Q\right)^{-1/(2-\beta)}$

(Optimal Heuristics for unknown ρ)

Low Classification noise ($\beta = 1$):

ERM on combined source and target data.

Non i.i.d. Bernstein + usual fixed point argument

Unknown Noise Level ($\beta \in [0, 1]$):

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

Lepski-type argument

Upper-bound Analysis:

Performance limits: $\mathcal{E}_Q(\hat{h}) \propto \left(n_P^{1/\rho} + n_Q\right)^{-1/(2-\beta)}$

(Optimal Heuristics for unknown ρ)

Low Classification noise ($\beta = 1$):

ERM on combined source and target data.

Non i.i.d. Bernstein + usual fixed point argument

Unknown Noise Level ($\beta \in [0, 1]$):

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

Lepski-type argument

Quick Summary and some New Directions ...

Quick Summary:

- ρ captures a more optimistic view of transferability $P \to Q$.
- Reveals general form of optimal heuristics:

Minimize $\hat{R}_P(h)$ subject to $\hat{R}_Q(h)$ not too large ...

- Cost-sensitive sampling is possible with no knowledge of ρ.
- Results extend to $h_P^* \neq h_Q^*$: $\exists \hat{h}$ s.t.

$$\mathcal{E}_Q(\hat{h}) \lesssim \min\left\{ n_P^{-1/(2-\beta)\rho} + \mathcal{E}_Q(h_P^*), n_Q^{-1/(2-\beta)} \right\}$$

Quick Summary:

- ρ captures a more optimistic view of transferability $P \to Q$.
- Reveals general form of optimal heuristics:

  Minimize $\hat{R}_P(h)$ subject to $\hat{R}_Q(h)$ not too large ...

- Cost-sensitive sampling is possible with no knowledge of ρ.
- Results extend to $h_P^* \neq h_Q^*$: $\exists \hat{h}$ s.t.

$$\mathcal{E}_Q(\hat{h}) \lesssim \min\left\{ n_P^{-1/(2-\beta)\rho} + \mathcal{E}_Q(h_P^*), n_Q^{-1/(2-\beta)} \right\}$$

Quick Summary:

- ρ captures a more optimistic view of transferability $P \to Q$.
- Reveals general form of optimal heuristics:

Minimize $\hat{R}_P(h)$ subject to $\hat{R}_Q(h)$ not too large ...

- Cost-sensitive sampling is possible with no knowledge of ρ.
- Results extend to $h_P^* \neq h_Q^*$: $\exists \hat{h}$ s.t.

$$\mathcal{E}_Q(\hat{h}) \lesssim \min\left\{ n_P^{-1/(2-\beta)\rho} + \mathcal{E}_Q(h_P^*), n_Q^{-1/(2-\beta)} \right\}$$

Quick Summary:

- ρ captures a more optimistic view of transferability $P \to Q$.
- Reveals general form of optimal heuristics:

  Minimize $\hat{R}_P(h)$ subject to $\hat{R}_Q(h)$ not too large ...

- Cost-sensitive sampling is possible with no knowledge of ρ.
- Results extend to $h_P^* \neq h_Q^*$: $\exists \hat{h}$ s.t.

$$\mathcal{E}_Q(\hat{h}) \lesssim \min\left\{ n_P^{-1/(2-\beta)\rho} + \mathcal{E}_Q(h_P^*), n_Q^{-1/(2-\beta)} \right\}$$

Quick Summary:

- ρ captures a more optimistic view of transferability $P \to Q$.
- Reveals general form of optimal heuristics:

  Minimize $\hat{R}_P(h)$ subject to $\hat{R}_Q(h)$ not too large ...

- Cost-sensitive sampling is possible with no knowledge of ρ.
- Results extend to $h_P^* \neq h_Q^*$: $\exists \hat{h}$ s.t.

$$\mathcal{E}_Q(\hat{h}) \lesssim \min\left\{ n_P^{-1/(2-\beta)\rho} + \mathcal{E}_Q(h_P^*), n_Q^{-1/(2-\beta)} \right\}$$

Recent work:

Limits of Adaptivity in Multi-Task (AoS 2022 with S. Hanneke)

$$P_1 + P_2 + \cdots + P_N + Q \to Q?$$

Prior theory only yields single source rates ...

Recent work:

Limits of Adaptivity in Multi-Task (AoS 2022 with S. Hanneke)

$$P_1 + P_2 + \cdots + P_N + Q \to Q?$$

Prior theory only yields single source rates ...

Setup:

N sources $\{P_t\}_{t=1}^N \mapsto Q$ with $\mathcal{E}_t(h) \gtrsim \mathcal{E}_Q^{\rho_t}(h)$

$$\text{Minimax Rate on } \mathcal{E}_Q(\hat{h}) : \quad \min_{t\in[N+1]} \left(\sum_{s=1}^{t} n_{(s)}\right)^{-1/(2-\beta)\bar{\rho}_t}$$

Adaptive Strategies (as $N \to \infty$):

Low noise ($\beta = 1$): ERM on combined data

Information on ranking $\rho_{(1)} \leq ... \leq \rho_{(N)}$: Greedy ICI strategy ...

No adaptive strategy outside above regimes !!!

$$\inf_{\hat{h}} \sup_{\text{rankings}} \sup_{\{P_t\}\times Q} \mathcal{E}_Q(\hat{h}) \gtrsim n_Q^{-1/(2-\beta)}.$$

Setup:

N sources $\{P_t\}_{t=1}^N \mapsto Q$ with $\mathcal{E}_t(h) \gtrsim \mathcal{E}_Q^{\rho_t}(h)$

Minimax Rate on $\mathcal{E}_Q(\hat{h})$: $$\min_{t \in [N+1]} \left(\sum_{s=1}^{t} n_{(s)} \right)^{-1/(2-\beta)\bar{\rho}_t}$$

Adaptive Strategies (as $N \to \infty$):

Low noise ($\beta = 1$): ERM on combined data

Information on ranking $\rho_{(1)} \leq \ldots \leq \rho_{(N)}$: Greedy ICI strategy ...

No adaptive strategy outside above regimes !!!

$$\inf_{\hat{h}} \sup_{\text{rankings}} \sup_{\{P_t\} \times Q} \mathcal{E}_Q(\hat{h}) \gtrsim n_Q^{-1/(2-\beta)}.$$

Setup:

N sources $\{P_t\}_{t=1}^N \mapsto Q$ with $\mathcal{E}_t(h) \gtrsim \mathcal{E}_Q^{\rho_t}(h)$

Minimax Rate on $\mathcal{E}_Q(\hat{h})$: $$\min_{t \in [N+1]} \left(\sum_{s=1}^{t} n_{(s)} \right)^{-1/(2-\beta)\bar{\rho}_t}$$

Adaptive Strategies (as $N \to \infty$):

Low noise ($\beta = 1$): ERM on combined data

Information on ranking $\rho_{(1)} \le \ldots \le \rho_{(N)}$: Greedy ICI strategy ...

No adaptive strategy outside above regimes !!!

$$\inf_{\hat{h}} \sup_{\text{rankings}} \sup_{\{P_t\} \times Q} \mathcal{E}_Q(\hat{h}) \gtrsim n_Q^{-1/(2-\beta)}.$$

Setup:

N sources $\{P_t\}_{t=1}^N \mapsto Q$ with $\mathcal{E}_t(h) \gtrsim \mathcal{E}_Q^{\rho_t}(h)$

Minimax Rate on $\mathcal{E}_Q(\hat{h})$:
$$\min_{t \in [N+1]} \left(\sum_{s=1}^{t} n_{(s)} \right)^{-1/(2-\beta)\bar{\rho}_t}$$

Adaptive Strategies (as $N \to \infty$):

Low noise ($\beta = 1$): ERM on combined data

Information on ranking $\rho_{(1)} \leq ... \leq \rho_{(N)}$: Greedy ICI strategy ...

No adaptive strategy outside above regimes !!!

$$\inf_{\hat{h}} \sup_{\text{rankings}} \sup_{\{P_t\} \times Q} \mathcal{E}_Q(\hat{h}) \gtrsim n_Q^{-1/(2-\beta)}.$$

Setup:

N sources $\{P_t\}_{t=1}^{N} \mapsto Q$ with $\mathcal{E}_t(h) \gtrsim \mathcal{E}_Q^{\rho_t}(h)$

$$\text{Minimax Rate on } \mathcal{E}_Q(\hat{h}) : \quad \min_{t \in [N+1]} \left(\sum_{s=1}^{t} n_{(s)} \right)^{-1/(2-\beta)\bar{\rho}_t}$$

Adaptive Strategies (as $N \to \infty$):

Low noise ($\beta = 1$): ERM on combined data

Information on ranking $\rho_{(1)} \leq \ldots \leq \rho_{(N)}$: Greedy ICI strategy ...

No adaptive strategy outside above regimes !!!

$$\inf_{\hat{h}} \sup_{\text{rankings}} \sup_{\{P_t\} \times Q} \mathcal{E}_Q(\hat{h}) \gtrsim n_Q^{-1/(2-\beta)}.$$

Setup:

N sources $\{P_t\}_{t=1}^N \mapsto Q$ with $\mathcal{E}_t(h) \gtrsim \mathcal{E}_Q^{\rho_t}(h)$

Minimax Rate on $\mathcal{E}_Q(\hat{h})$: $$\min_{t \in [N+1]} \left(\sum_{s=1}^{t} n_{(s)} \right)^{-1/(2-\beta)\bar{\rho}_t}$$

Adaptive Strategies (as $N \to \infty$):

Low noise $(\beta = 1)$: ERM on combined data

Information on ranking $\rho_{(1)} \leq ... \leq \rho_{(N)}$: Greedy ICI strategy ...

No adaptive strategy outside above regimes !!!

$$\inf_{\hat{h}} \sup_{\text{rankings}} \sup_{\{P_t\} \times Q} \mathcal{E}_Q(\hat{h}) \gtrsim n_Q^{-1/(2-\beta)}.$$

Driving Philosophy: which aspects of the change affect learning?

Unknown Distribution Changes in Bandits (with Joe Suk)

- *No Change-Point Detection* under Covariate-Shifts. ALT 21 .
- Detecting *(In-)Significant Changes* in Best-Arms. COLT 22.

Model Selection and Transfer (with S. Hanneke, to be written :))

··· (sample sizes, model complexity, model transferability)

Somehow we are still just scratching the surface ...

Thanks!

Driving Philosophy: which aspects of the change affect learning?

Unknown Distribution Changes in Bandits (with Joe Suk)

- *No Change-Point Detection* under Covariate-Shifts. ALT 21 .
- Detecting *(In-)Significant Changes* in Best-Arms. COLT 22.

Model Selection and Transfer (with S. Hanneke, to be written :))

⋯ (sample sizes, model complexity, model transferability)

Somehow we are still just scratching the surface ...

Thanks!

Driving Philosophy: which aspects of the change affect learning?

Unknown Distribution Changes in Bandits (with Joe Suk)

- *No Change-Point Detection* under Covariate-Shifts. ALT 21 .
- Detecting *(In-)Significant Changes* in Best-Arms. COLT 22.

Model Selection and Transfer (with S. Hanneke, to be written :))

··· (sample sizes, model complexity, model transferability)

Somehow we are still just scratching the surface ...

Thanks!