On Learning Invariant Representations for Domain Adaptation

Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J. Gordon
Presented by: Han Zhao
han.zhao@cs.cmu.edu
Machine Learning Department, Carnegie Mellon University

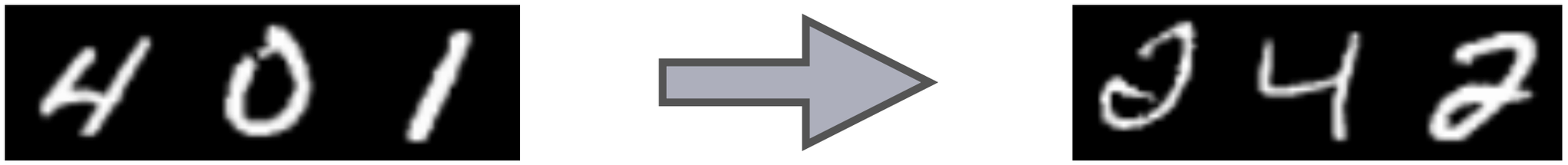
June 11th, 2019

Microsoft Research

Carnegie Mellon University

Background

Generalization: Source (Train) = Target (Test)



Source

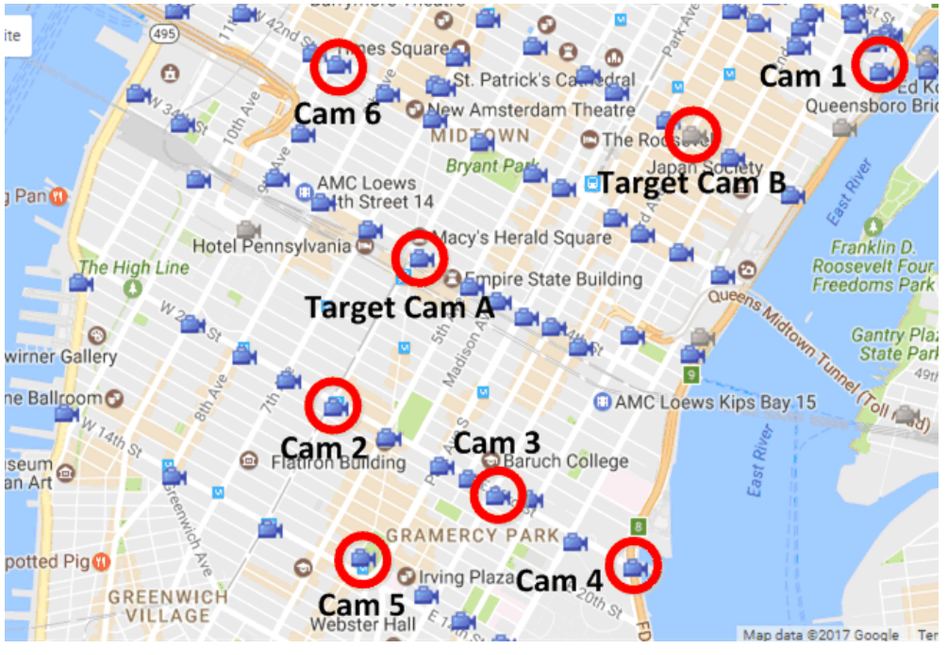
Target

Background

Domain adaptation: Source $\neq$ Target

Source (with Labels)

Target (No Labels)



Background

Many applications…

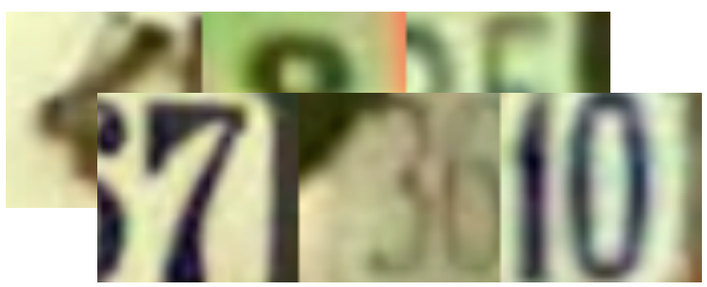
Background

Many applications…

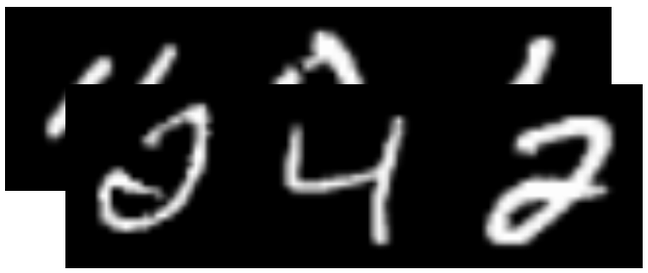
- Face recognition (Sohn et al.' 17)
- Object recognition (Ghifary et al.' 15)
- Sentiment analysis (Glorot et al.' 11)
- Accented speech recognition (Sun et al.' 17)
- Indoor WiFi localization (Pan et al.' 08)
- Relational reasoning (Davis et al.' 09)
- ……

Background

Learning domain-invariant representations:



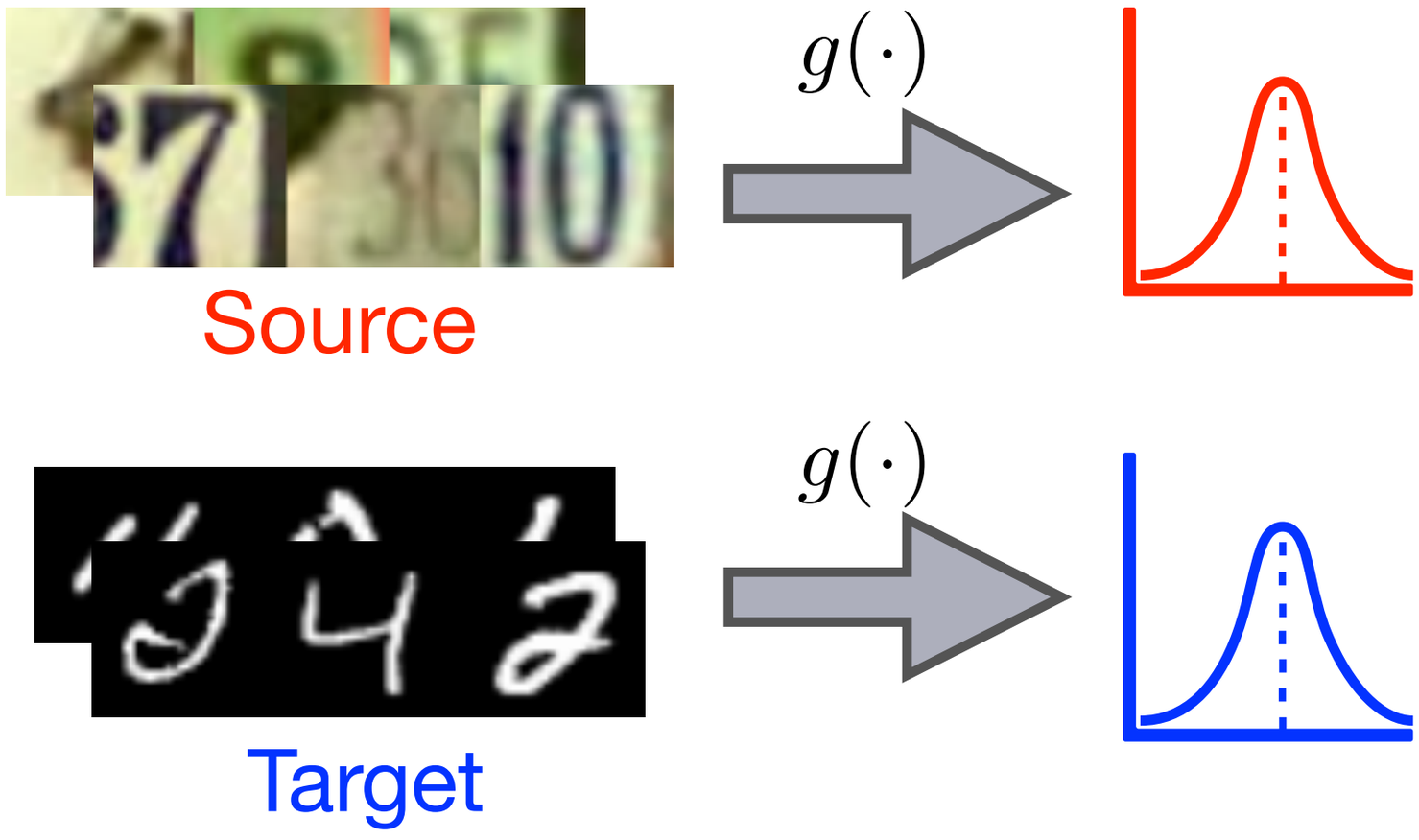
Source



Target

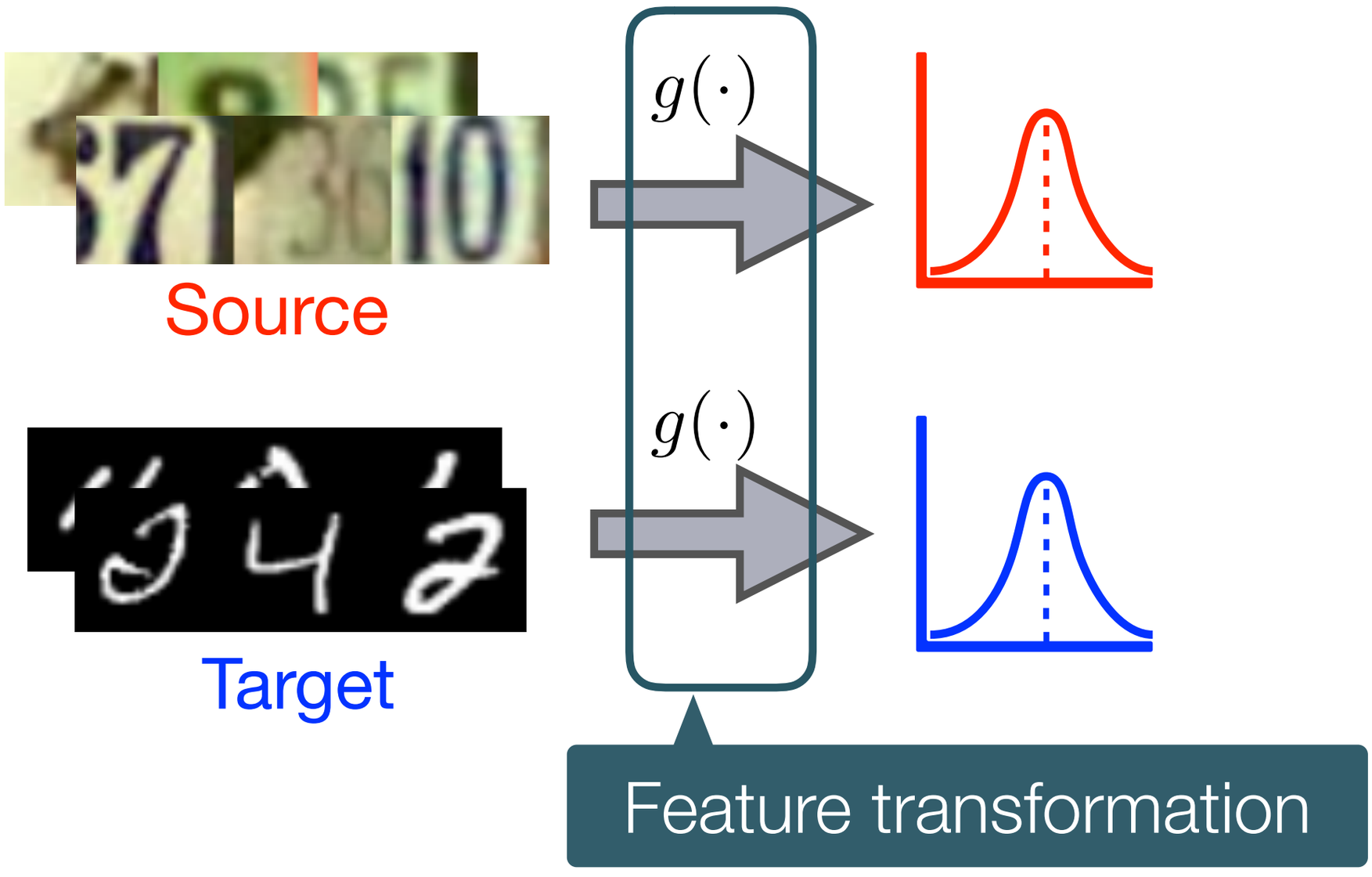
Background

Learning domain-invariant representations:



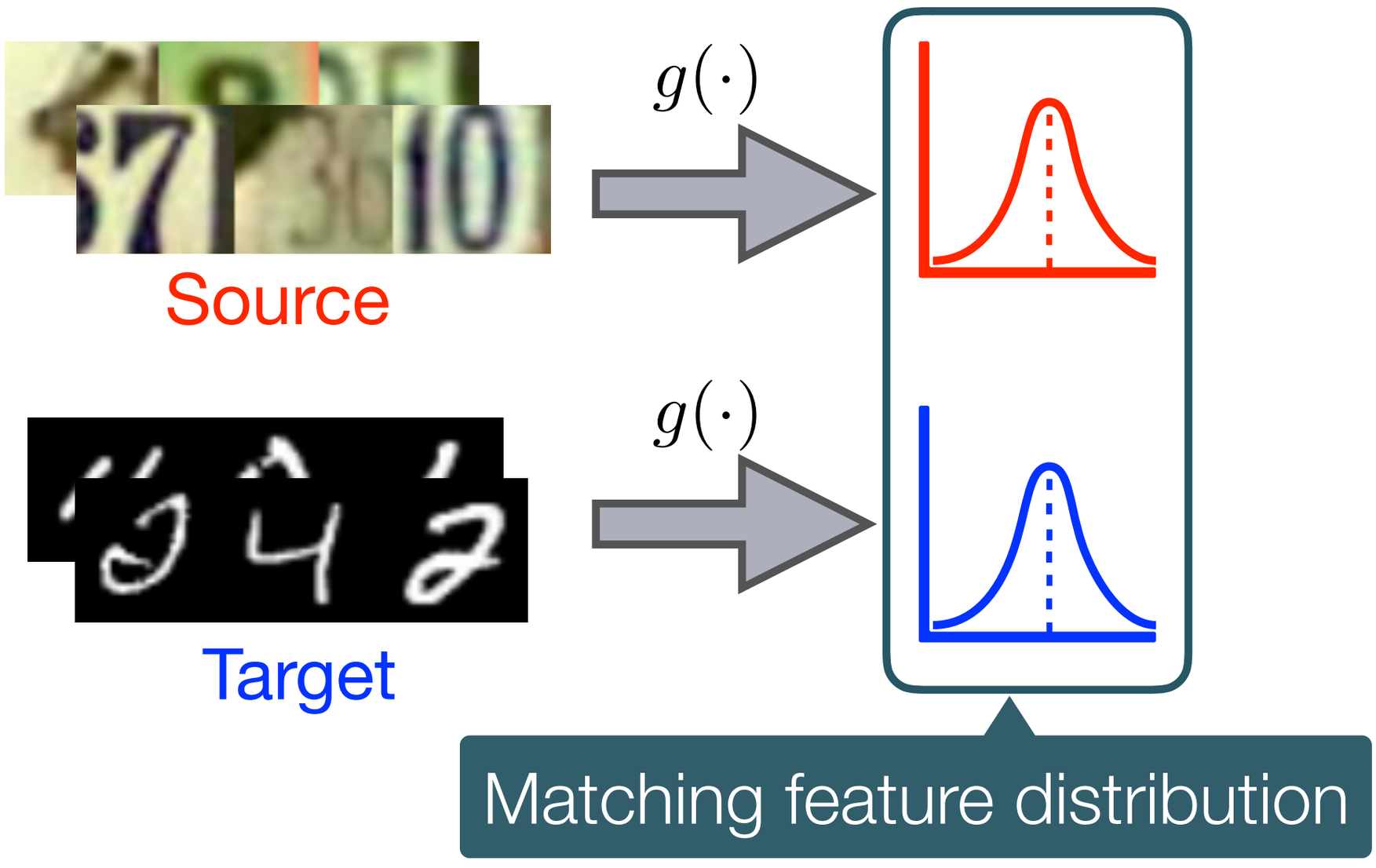
Background

Learning domain-invariant representations:



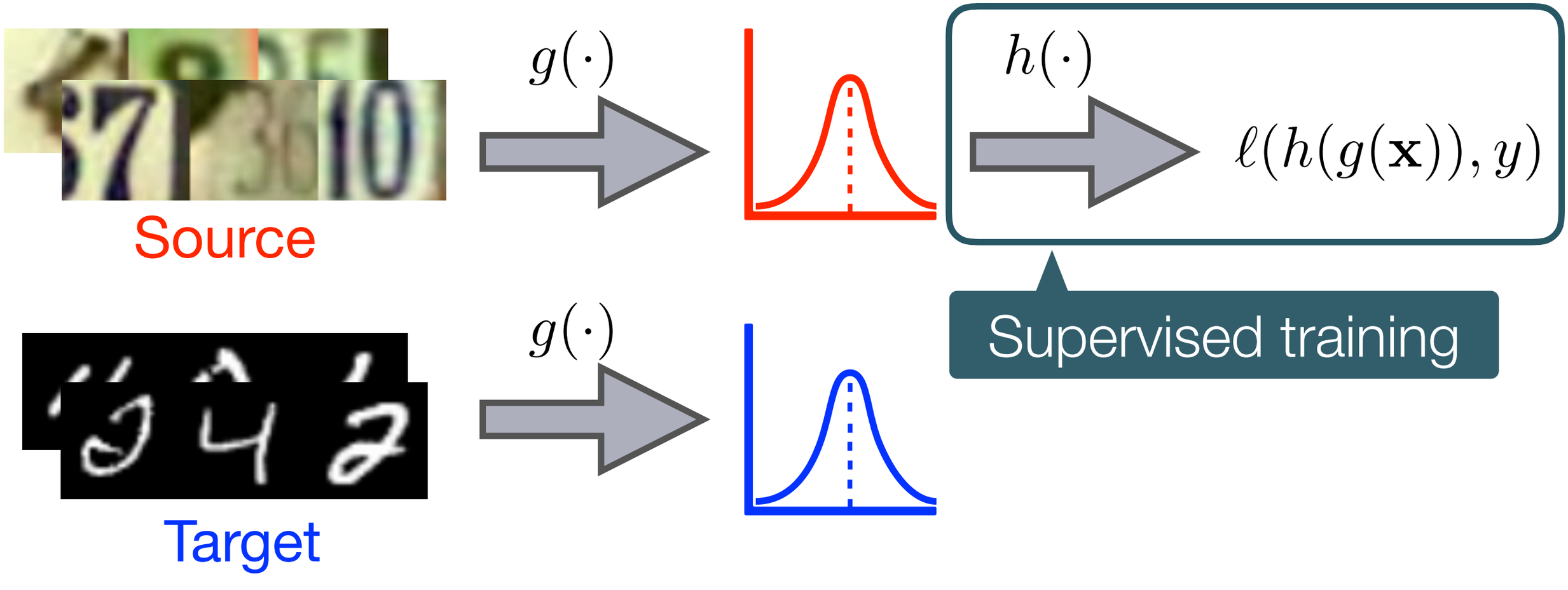
Background

Learning domain-invariant representations:



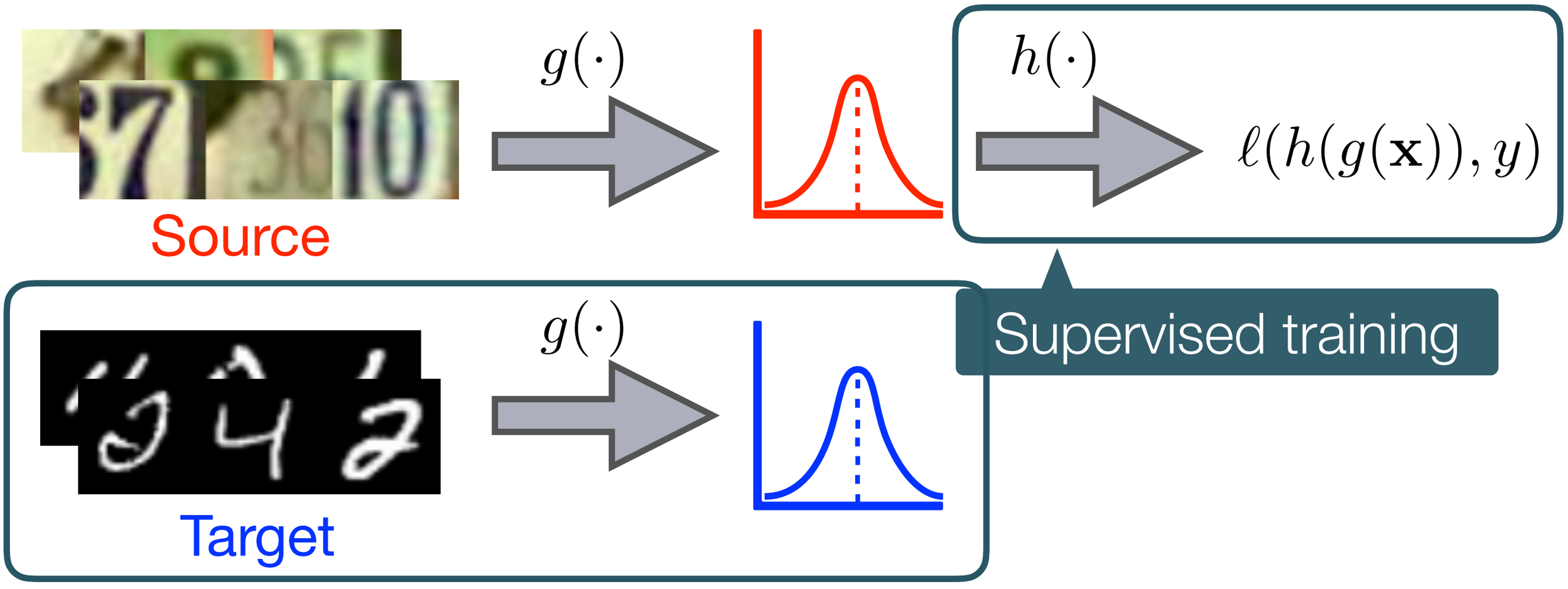
Background

Learning domain-invariant representations:



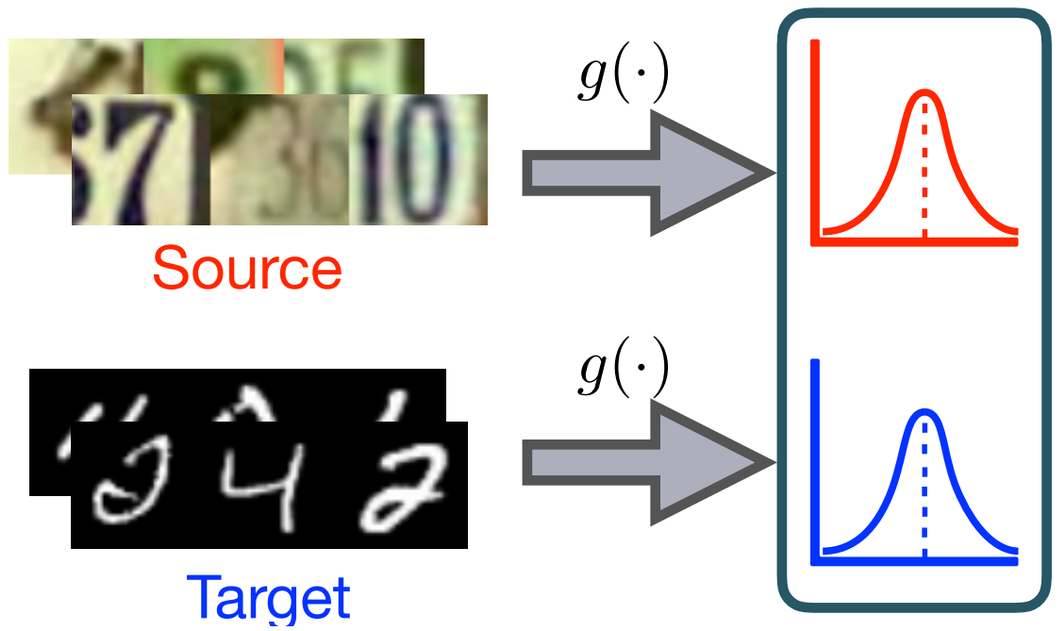
Background

Learning domain-invariant representations:



Background

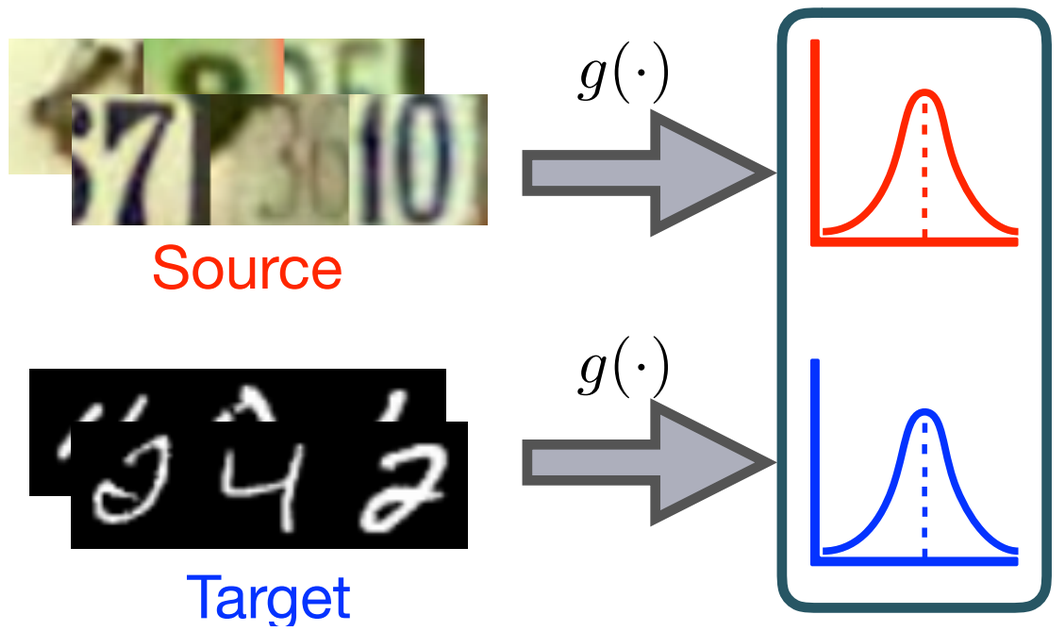
Learning domain-invariant representations:



Background

Learning domain-invariant representations:

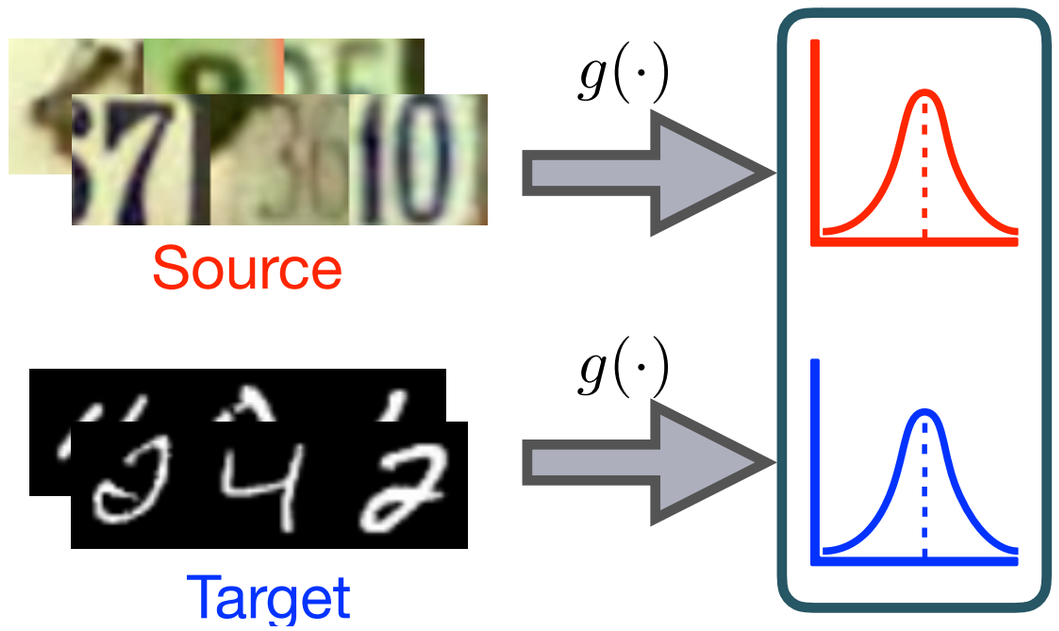
- Adversarial discriminator (DANN, Ganin et al.' 15)



Background

Learning domain-invariant representations:

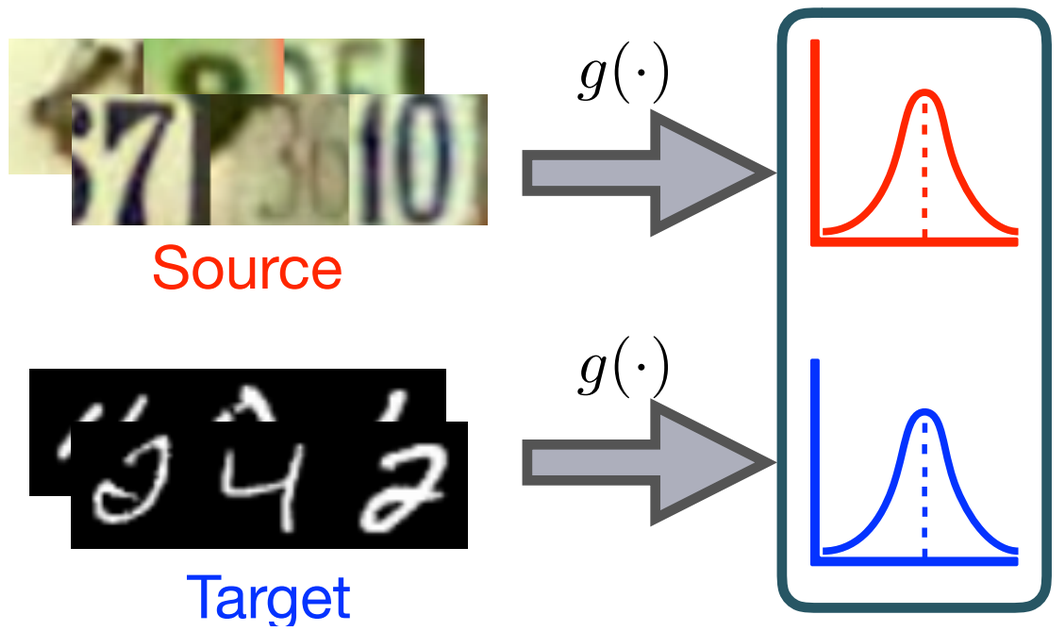
- Adversarial discriminator (DANN, Ganin et al.' 15)
- Maximum mean discrepancy (DAN, Long et al.' 15)



Background

Learning domain-invariant representations:

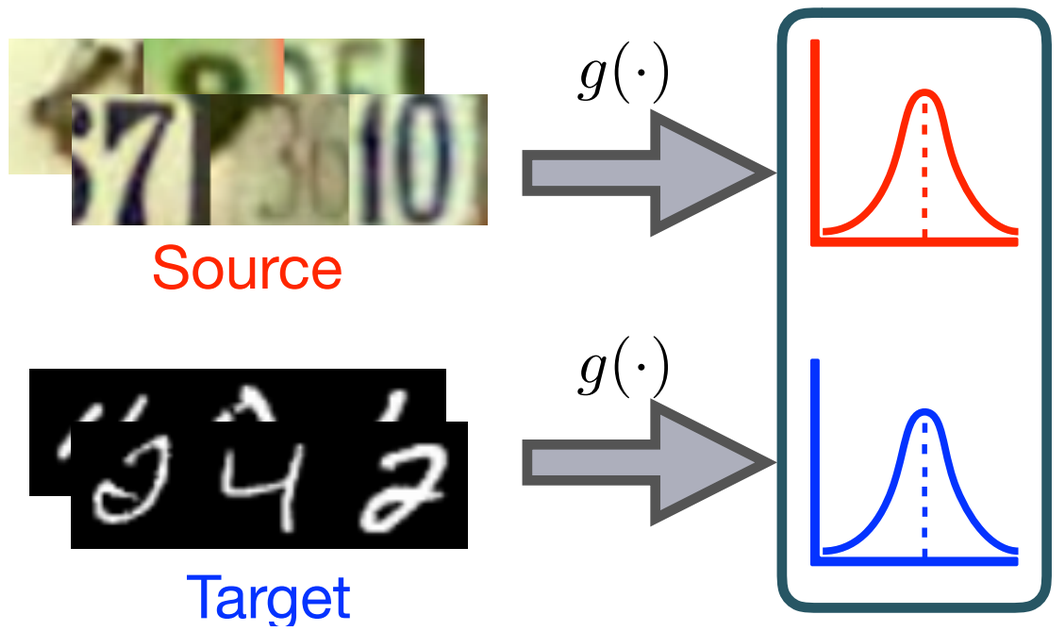
- Adversarial discriminator (DANN, Ganin et al.' 15)
- Maximum mean discrepancy (DAN, Long et al.' 15)
- Wasserstein distance (Shen et al.' 18)



Background

Learning domain-invariant representations:

- Adversarial discriminator (DANN, Ganin et al.' 15)
- Maximum mean discrepancy (DAN, Long et al.' 15)
- Wasserstein distance (Shen et al.' 18)
- .....



Background

Justification of learning domain-invariant representations:

Theorem (Ben-David et al.' 10):

$$\varepsilon_T(h) \leq \varepsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(T; S) + \lambda^*$$

- $\varepsilon_T(h)/\varepsilon_S(h)$: true target/source errors
- $d_{\mathcal{H}\Delta\mathcal{H}}(T; S)$: divergence between target/source distributions
- $\lambda^* := \min_{h' \in \mathcal{H}} \varepsilon_S(h') + \varepsilon_T(h')$: optimal joint error

Background

Justification of learning domain-invariant representations:

Theorem (Ben-David et al.' 10):

$$\varepsilon_T(h) \leq \varepsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(T; S) + \lambda^*$$

- $\varepsilon_T(h)/\varepsilon_S(h)$: true target/source errors
- $d_{\mathcal{H}\Delta\mathcal{H}}(T; S)$: divergence between target/source distributions
- $\lambda^* := \min_{h' \in \mathcal{H}} \varepsilon_S(h') + \varepsilon_T(h')$: optimal joint error

Bound-minimizing algorithm:

$$\min \quad \varepsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(T; S)$$

Motivation

Question:

Is finding invariant representations while at the same time achieving a small source error sufficient to guarantee a small target error?

Motivation

Question:

Is finding invariant representations while at the same time achieving a small source error sufficient to guarantee a small target error?

- If not, under what conditions is it?

- Is there any necessary condition for this family of methods?

Overview

Question:

Is finding invariant representations while at the same time achieving a small source error sufficient to guarantee a small target error?

- If not, under what conditions is it?

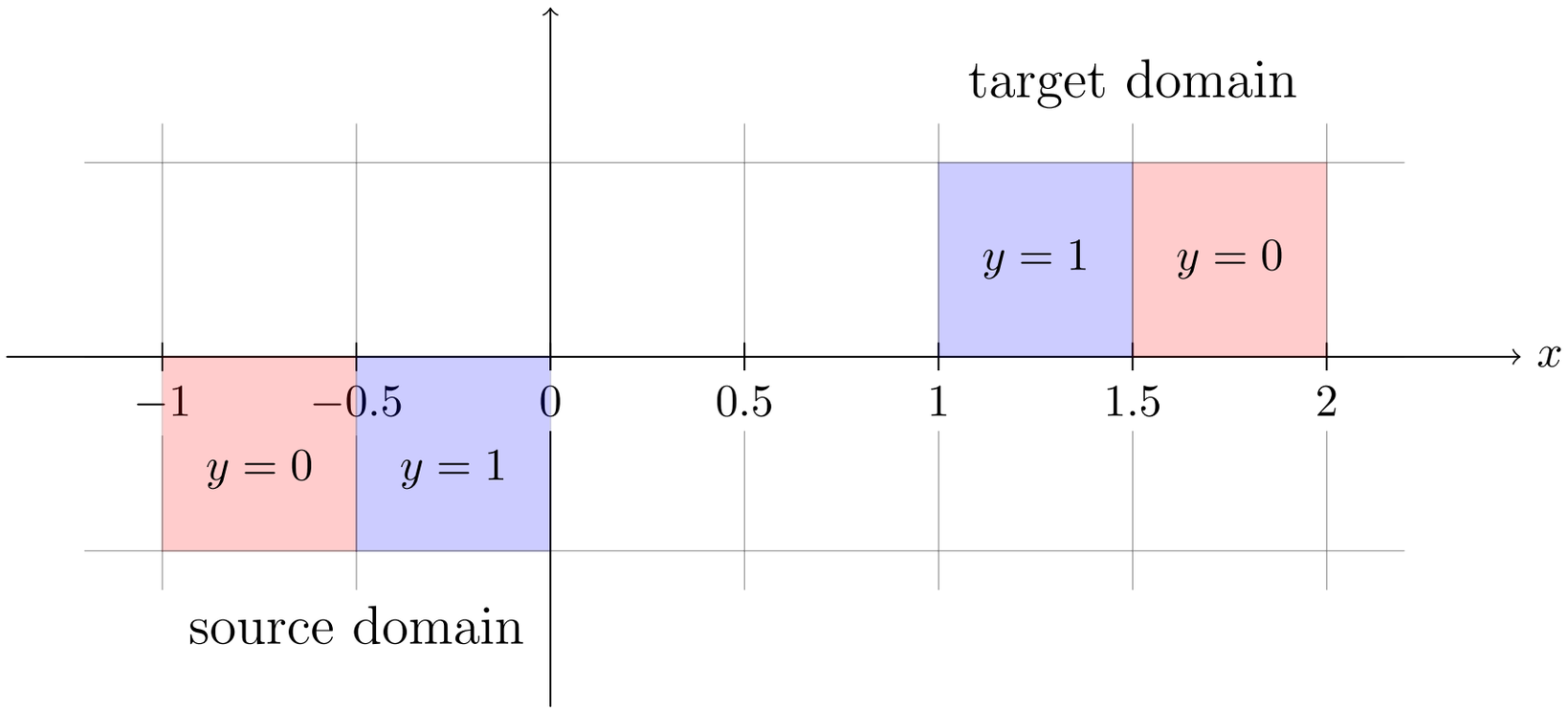
Sufficient condition: only when the conditional distributions match $\mathcal{D}_S^{Y|X} \approx \mathcal{D}_T^{Y|X}$

- Is there any necessary condition for this family of methods?

Necessary condition: only when marginal label distributions are close $\mathcal{D}_S^{Y} \approx \mathcal{D}_T^{Y}$

A Simple Example

Consider a 1D adaptation problem:

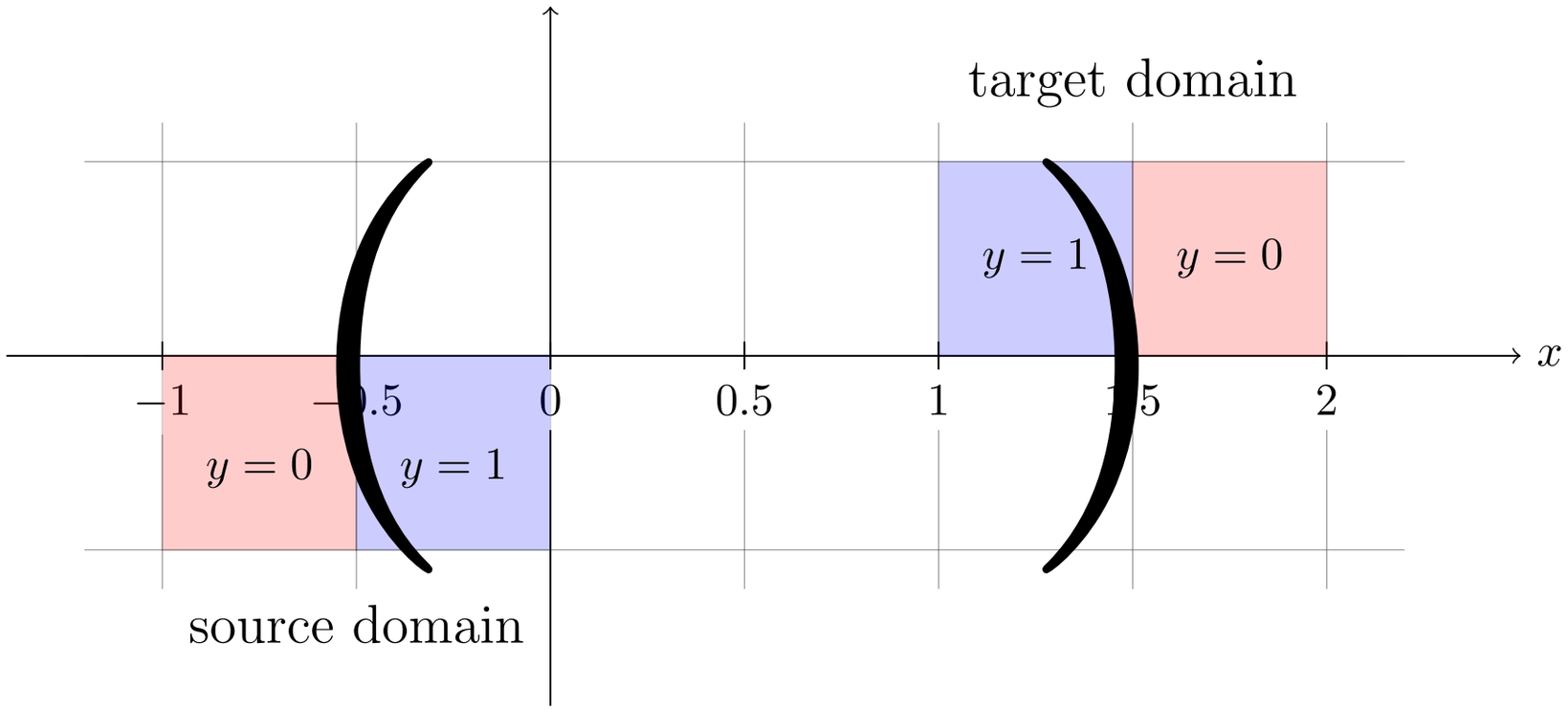


Source: $\mathcal{D}_S = U(-1, 0)$

Target: $\mathcal{D}_T = U(1, 2)$

A Simple Example

Consider a 1D adaptation problem:



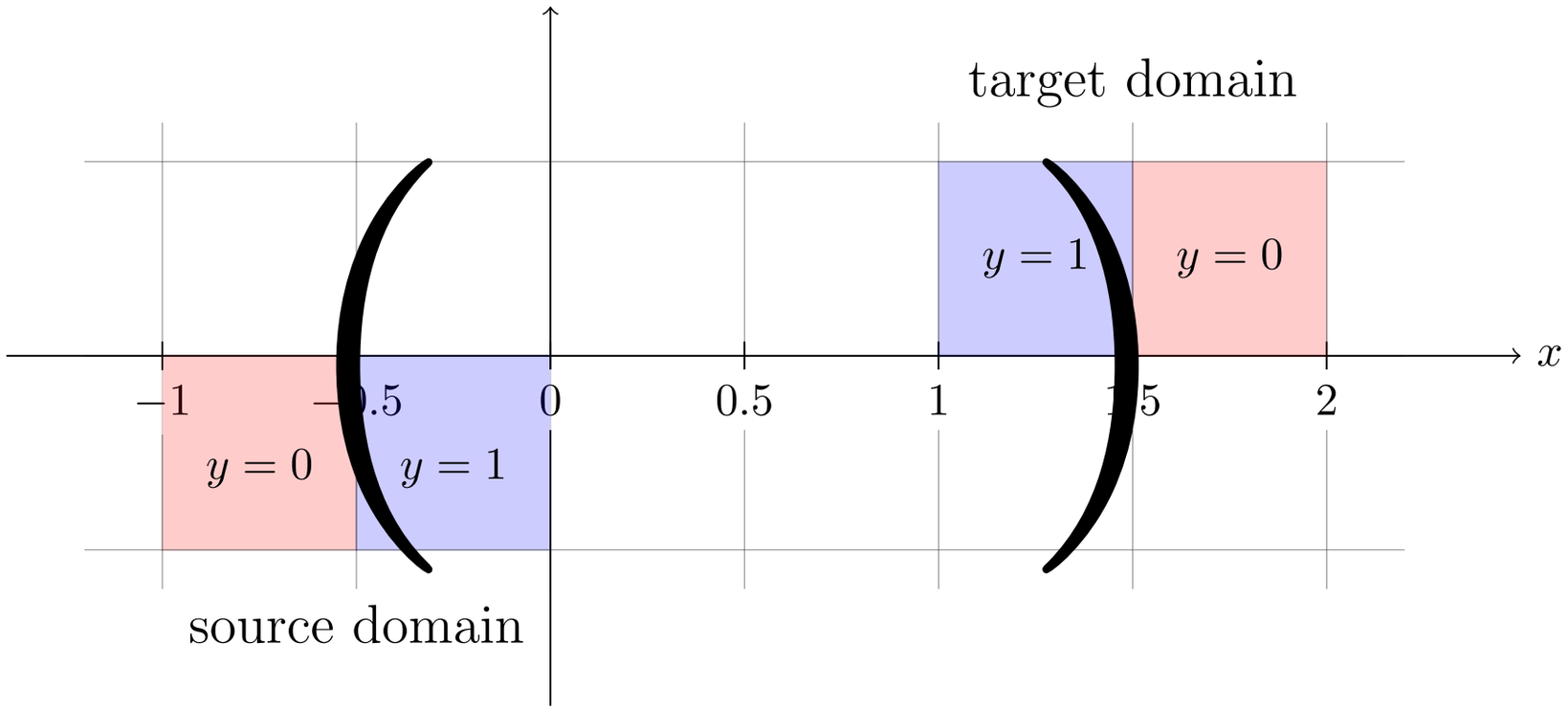
Source: $\mathcal{D}_S = U(-1, 0)$

Target: $\mathcal{D}_T = U(1, 2)$

$h^*(x) = 1$ iff $x \in (-1/2, 3/2)$

A Simple Example

Consider a 1D adaptation problem:



Source: $\mathcal{D}_S = U(-1, 0)$

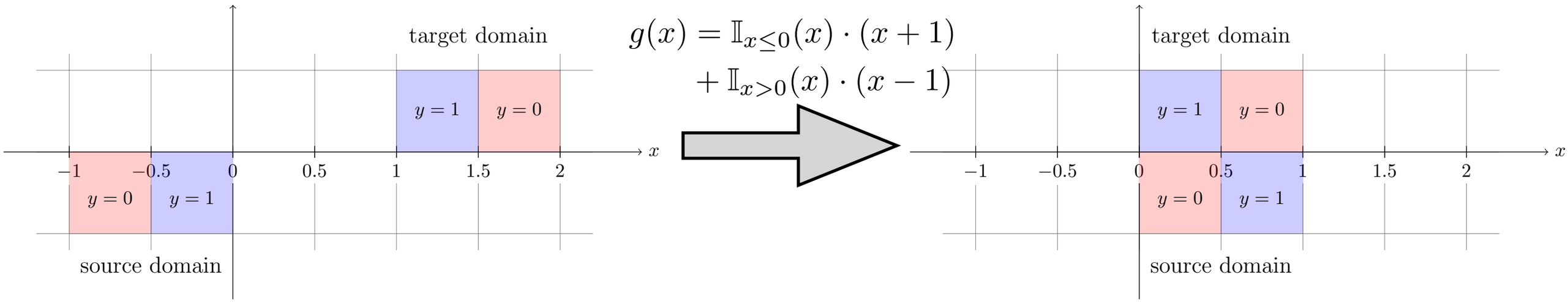
Target: $\mathcal{D}_T = U(1, 2)$

$h^*(x) = 1$ iff $x \in (-1/2, 3/2)$

$\lambda^* = \min_{h'} \varepsilon_S(h') + \varepsilon_T(h') = 0$

A Simple Example

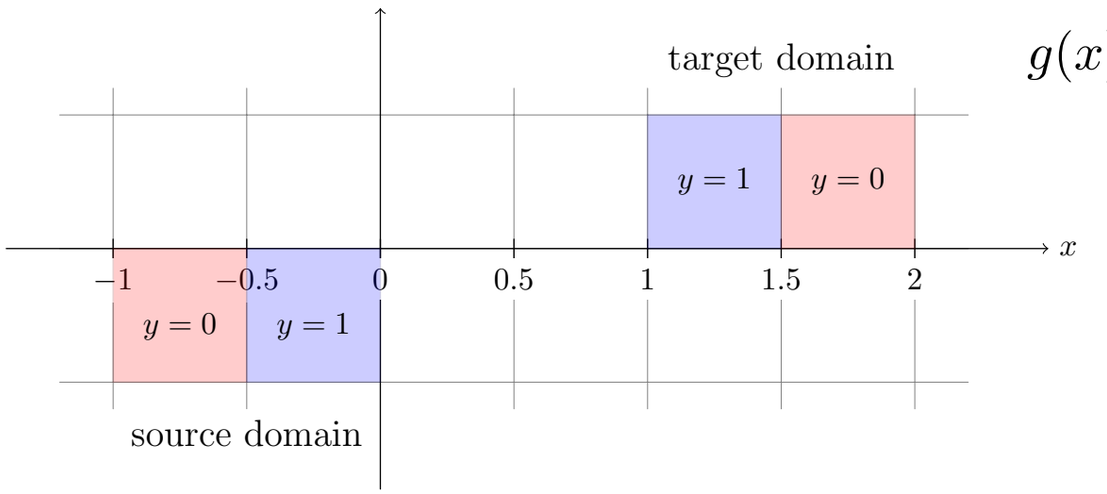
Consider a 1D adaptation problem: Adaptation



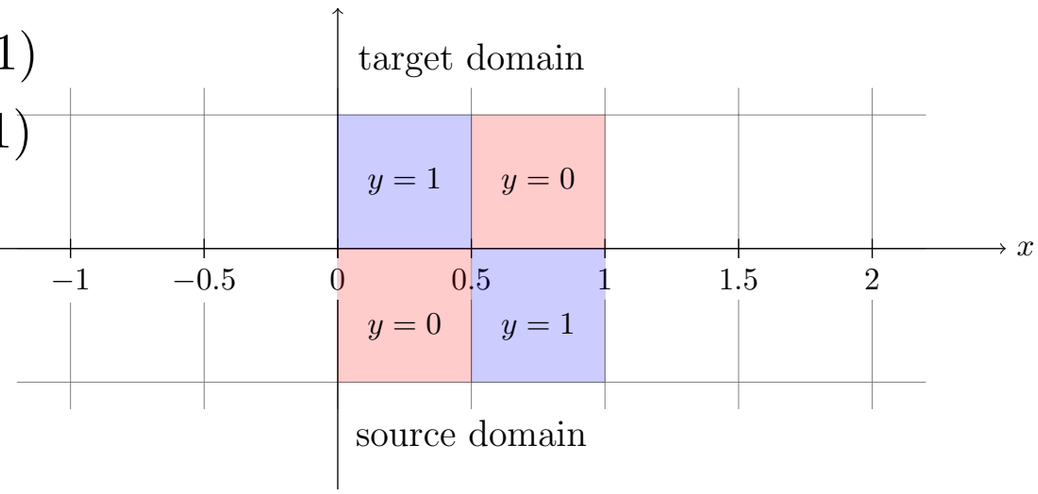
Bound-minimizing algorithm: $\min \quad \varepsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(T; S)$

A Simple Example

Consider a 1D adaptation problem: Adaptation



$$g(x) = \mathbb{I}_{x \leq 0}(x) \cdot (x + 1) + \mathbb{I}_{x > 0}(x) \cdot (x - 1)$$



Bound-minimizing algorithm: $\min \quad \varepsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(T; S)$

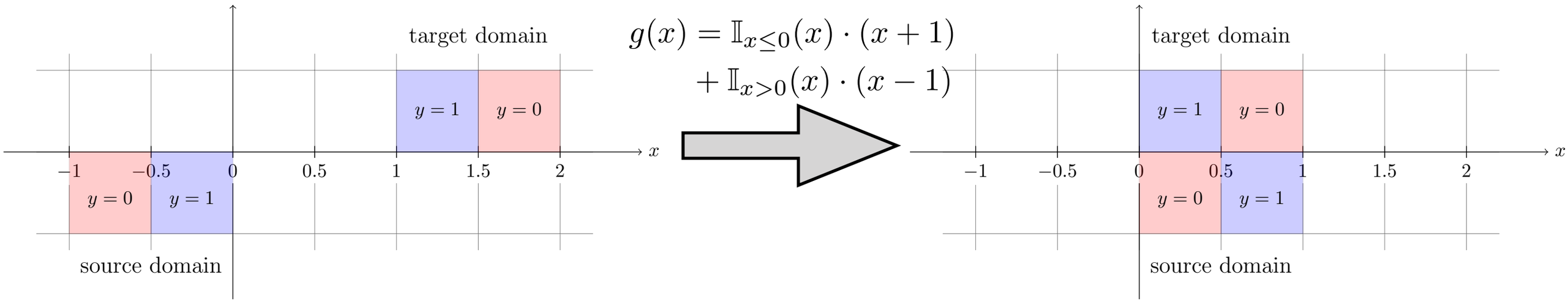
Source: $\mathcal{D}'_S = U(0, 1)$

Target: $\mathcal{D}'_T = U(0, 1)$

$$d_{\mathcal{H}\Delta\mathcal{H}}(S, T) = 0$$

A Simple Example

Consider a 1D adaptation problem: Adaptation



Bound-minimizing algorithm: $\min \quad \varepsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(T; S)$

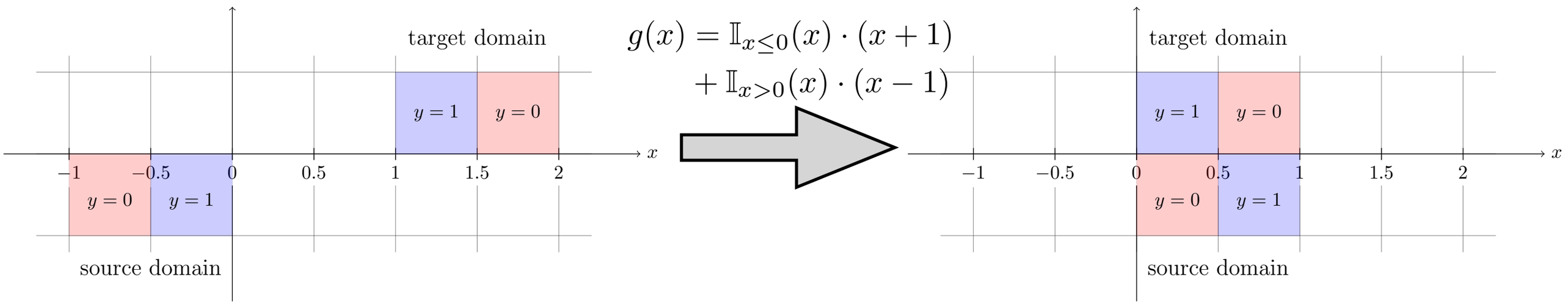
Source: $\mathcal{D}'_S = U(0, 1)$

Target: $\mathcal{D}'_T = U(0, 1)$

$$d_{\mathcal{H}\Delta\mathcal{H}}(S, T) = 0$$

If $\varepsilon_S(h) = 0$, then $\varepsilon_T(h) = 1$!

A Simple Example

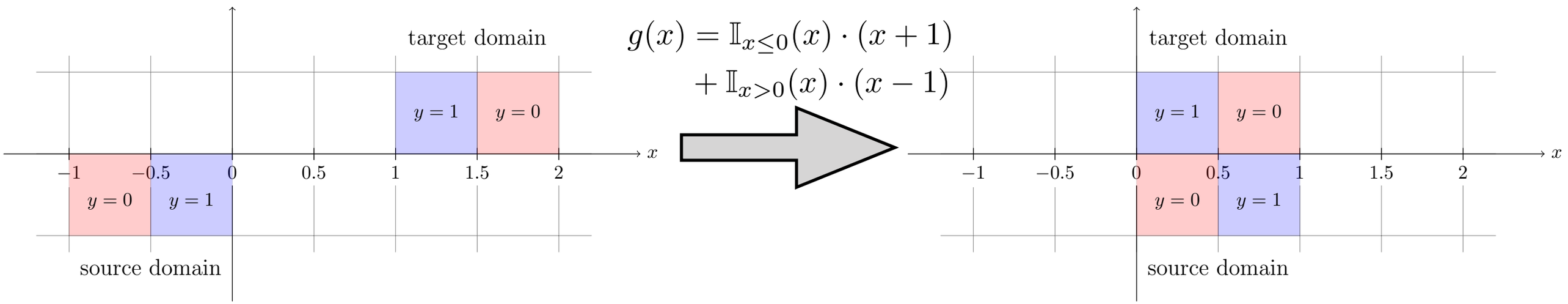


Bound-minimizing algorithm: $\min \quad \varepsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(T; S)$

What's wrong with this example?

$$\varepsilon_T(h) \leq \varepsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(T; S) + \lambda^*$$
$$\| \qquad\qquad \|$$
$$0 \qquad\qquad 0$$

A Simple Example



Bound-minimizing algorithm: $\min \quad \varepsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(T; S)$

What's wrong with this example?

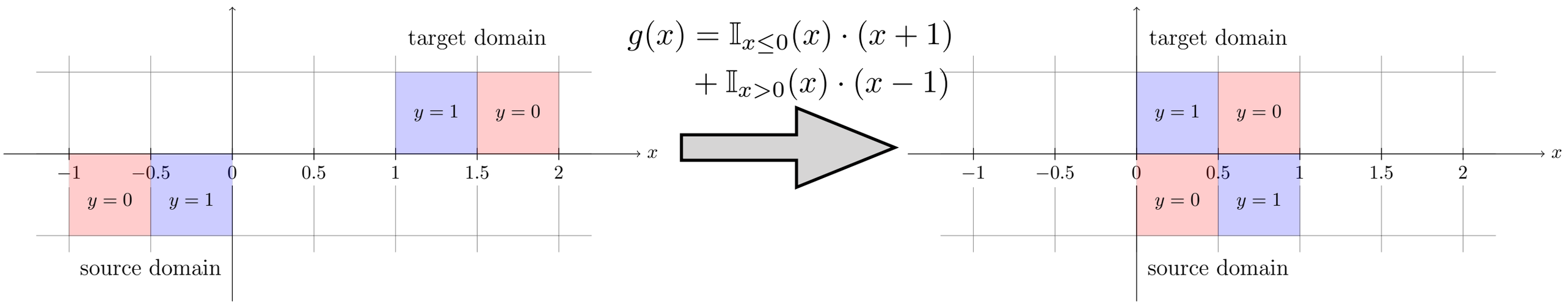
$$\varepsilon_T(h) \leq \varepsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(T; S) + \lambda^*$$

$$\| \qquad\qquad \|$$

$$0 \qquad\qquad 0$$

In fact, $\forall h \in 2^{[0,1]}, \quad \varepsilon_S(h) + \varepsilon_T(h) = 1 \Rightarrow \lambda^* = 1$

A Simple Example



Bound-minimizing algorithm: $\min \quad \varepsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(T; S)$

What's wrong with this example?

$$\varepsilon_T(h) \leq \varepsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(T; S) + \lambda^*$$

$$\| \qquad\qquad \|$$

$$0 \qquad\qquad 0$$

In fact, $\forall h \in 2^{[0,1]}, \quad \varepsilon_S(h) + \varepsilon_T(h) = 1 \Rightarrow \lambda^* = 1$

Conditional distributions are maximally different after adaptation:

$$\mathbb{E}_{x \sim U(0,1)} \left[\left| \mathcal{D}_S^{Y|X} - \mathcal{D}_T^{Y|X} \right| \right] = 1$$

A Generalization Upper Bound

Failure mode: conditional shift increases during adaptation

Could we also take into account the conditional shift?

A Generalization Upper Bound

Failure mode: conditional shift increases during adaptation

Could we also take into account the conditional shift?

Theorem (deterministic setting): Let $\langle \mathcal{D}_S, f_S \rangle$ and $\langle \mathcal{D}_T, f_T \rangle$ be the source and target domains. For any function class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, and $\forall h \in \mathcal{H}$, the following inequality holds:

$$\left|\varepsilon_S(h) - \varepsilon_T(h)\right| \leq d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) + \min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\}$$

- $\tilde{\mathcal{H}} := \{\text{sgn}(|h(x) - h'(x)| - t) \mid h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$, a hypothesis class induced by $\mathcal{H}$
- f_S / f_T labeling function in source/target domains

A Generalization Upper Bound

Failure mode: conditional shift increases during adaptation

Could we also take into account the conditional shift?

Theorem (deterministic setting): Let $\langle \mathcal{D}_S, f_S \rangle$ and $\langle \mathcal{D}_T, f_T \rangle$ be the source and target domains. For any function class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, and $\forall h \in \mathcal{H}$, the following inequality holds:

$$\left|\varepsilon_S(h) - \varepsilon_T(h)\right| \leq d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) + \min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\}$$

Shift between data distributions

- $\tilde{\mathcal{H}} := \{\text{sgn}(|h(x) - h'(x)| - t) \mid h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$, a hypothesis class induced by $\mathcal{H}$
- f_S / f_T labeling function in source/target domains

A Generalization Upper Bound

Failure mode: conditional shift increases during adaptation

Could we also take into account the conditional shift?

Theorem (deterministic setting): Let $\langle \mathcal{D}_S, f_S \rangle$ and $\langle \mathcal{D}_T, f_T \rangle$ be the source and target domains. For any function class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, and $\forall h \in \mathcal{H}$, the following inequality holds:

$$\left|\varepsilon_S(h) - \varepsilon_T(h)\right| \leq d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) + \min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\}$$

Shift between data distributions

Shift between conditional distributions

- $\tilde{\mathcal{H}} := \{\text{sgn}(|h(x) - h'(x)| - t) \mid h, h' \in \mathcal{H}, 0 \leq t \leq 1\}$, a hypothesis class induced by $\mathcal{H}$
- f_S / f_T labeling function in source/target domains

A Generalization Upper Bound

Theorem (deterministic setting): Let $\langle \mathcal{D}_S, f_S \rangle$ and $\langle \mathcal{D}_T, f_T \rangle$ be the source and target domains. For any function class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, and $\forall h \in \mathcal{H}$, the following inequality holds:

$$\left|\varepsilon_S(h) - \varepsilon_T(h)\right| \leq d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) + \min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\}$$

Shift between data distributions

Shift between conditional distributions

A Generalization Upper Bound

Theorem (deterministic setting): Let $\langle \mathcal{D}_S, f_S \rangle$ and $\langle \mathcal{D}_T, f_T \rangle$ be the source and target domains. For any function class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, and $\forall h \in \mathcal{H}$, the following inequality holds:

$$\left|\varepsilon_S(h) - \varepsilon_T(h)\right| \leq d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) + \min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\}$$

Shift between data distributions

Shift between conditional distributions

- Free of λ^*, the optimal joint error

A Generalization Upper Bound

Theorem (deterministic setting): Let $\langle \mathcal{D}_S, f_S \rangle$ and $\langle \mathcal{D}_T, f_T \rangle$ be the source and target domains. For any function class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, and $\forall h \in \mathcal{H}$, the following inequality holds:

$$\left|\varepsilon_S(h) - \varepsilon_T(h)\right| \leq d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) + \min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\}$$

Shift between data distributions

Shift between conditional distributions

- Free of λ^*, the optimal joint error
- Two more noise terms if we extend the bound to stochastic setting

A Generalization Upper Bound

Theorem (deterministic setting): Let $\langle \mathcal{D}_S, f_S \rangle$ and $\langle \mathcal{D}_T, f_T \rangle$ be the source and target domains. For any function class $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$, and $\forall h \in \mathcal{H}$, the following inequality holds:

$$\left|\varepsilon_S(h) - \varepsilon_T(h)\right| \leq d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) + \min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\}$$

Shift between data distributions

Shift between conditional distributions

- Free of λ^*, the optimal joint error
- Two more noise terms if we extend the bound to stochastic setting
- Can use standard concentration argument to get high probability bound of $\varepsilon_S(h)$ and $d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T)$

An Information-Theoretic Lower Bound

Can we extend the counter-example to more general cases?
Consider the general adaptation scenario:

$$X \xrightarrow{g} Z \xrightarrow{h} \hat{Y}$$

- $g(\cdot)$ (nonlinear) feature transformation
- $h(\cdot) \in \{0, 1\}$ (randomized) classification function

An Information-Theoretic Lower Bound

Can we extend the counter-example to more general cases?
Consider the general adaptation scenario:

$$X \xrightarrow{g} Z \xrightarrow{h} \hat{Y}$$

- $g(\cdot)$ (nonlinear) feature transformation
- $h(\cdot) \in \{0, 1\}$ (randomized) classification function

Theorem: suppose the Markov chain holds and $d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, then

$$\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) \geq \frac{1}{2}\left(d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)\right)^2$$

- $\mathcal{D}_S^Z / \mathcal{D}_T^Z$: induced distributions by $g(\cdot)$ from $\mathcal{D}_S / \mathcal{D}_T$
- $d_{\mathrm{JS}}(\cdot, \cdot)$: Jensen-Shannon distance

$$d_{\mathrm{JS}}^2(\mathcal{D}, \mathcal{D}') = \frac{1}{2} D_{\mathrm{KL}}(\mathcal{D} \mid\mid \mathcal{D}_M) + \frac{1}{2} D_{\mathrm{KL}}(\mathcal{D}' \mid\mid \mathcal{D}_M), \quad \mathcal{D}_M := \frac{1}{2}(\mathcal{D} + \mathcal{D}')$$

An Information-Theoretic Lower Bound

Theorem: suppose the Markov chain holds and $d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, then

$$\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) \geq \frac{1}{2}\left(d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)\right)^2$$

- $\mathcal{D}_S^Z / \mathcal{D}_T^Z$: induced distributions by $g(\cdot)$ from $\mathcal{D}_S / \mathcal{D}_T$
- $d_{\mathrm{JS}}(\cdot, \cdot)$: Jensen-Shannon distance

An Information-Theoretic Lower Bound

Theorem: suppose the Markov chain holds and $d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, then

$$\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) \geq \frac{1}{2}\left(d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)\right)^2$$

- $\mathcal{D}_S^Z / \mathcal{D}_T^Z$: induced distributions by $g(\cdot)$ from $\mathcal{D}_S / \mathcal{D}_T$
- $d_{\mathrm{JS}}(\cdot, \cdot)$: Jensen-Shannon distance

- For domain-invariant representation, $d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z) = 0$, lower bound achieves maximum value

An Information-Theoretic Lower Bound

Theorem: suppose the Markov chain holds and $d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, then

$$\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) \geq \frac{1}{2}\left(d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)\right)^2$$

- $\mathcal{D}_S^Z / \mathcal{D}_T^Z$: induced distributions by $g(\cdot)$ from $\mathcal{D}_S / \mathcal{D}_T$
- $d_{\mathrm{JS}}(\cdot, \cdot)$: Jensen-Shannon distance

- For domain-invariant representation, $d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z) = 0$, lower bound achieves maximum value
- After a given threshold, minimizing source error and domain discrepancy leads to larger target error

An Information-Theoretic Lower Bound

Theorem: suppose the Markov chain holds and $d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, then

$$\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) \geq \frac{1}{2}\left(d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)\right)^2$$

Proof Sketch:

An Information-Theoretic Lower Bound

Theorem: suppose the Markov chain holds and $d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, then

$$\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) \geq \frac{1}{2}\left(d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)\right)^2$$

Proof Sketch:

JS distance is a metric: $d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \leq d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_S^{\hat{Y}}) + d_{\mathrm{JS}}(\mathcal{D}_S^{\hat{Y}}, \mathcal{D}_T^{\hat{Y}}) + d_{\mathrm{JS}}(\mathcal{D}_T^{\hat{Y}}, \mathcal{D}_T^Y)$

An Information-Theoretic Lower Bound

Theorem: suppose the Markov chain holds and $d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, then

$$\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) \geq \frac{1}{2}\left(d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)\right)^2$$

Proof Sketch:

JS distance is a metric: $d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \leq d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_S^{\hat{Y}}) + d_{\mathrm{JS}}(\mathcal{D}_S^{\hat{Y}}, \mathcal{D}_T^{\hat{Y}}) + d_{\mathrm{JS}}(\mathcal{D}_T^{\hat{Y}}, \mathcal{D}_T^Y)$

Data-processing Principle: $d_{\mathrm{JS}}(\mathcal{D}_S^{\hat{Y}}, \mathcal{D}_T^{\hat{Y}}) \leq d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$

An Information-Theoretic Lower Bound

Theorem: suppose the Markov chain holds and $d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, then

$$\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) \geq \frac{1}{2}\left(d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)\right)^2$$

Proof Sketch:

JS distance is a metric: $d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \leq d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_S^{\hat{Y}}) + d_{\mathrm{JS}}(\mathcal{D}_S^{\hat{Y}}, \mathcal{D}_T^{\hat{Y}}) + d_{\mathrm{JS}}(\mathcal{D}_T^{\hat{Y}}, \mathcal{D}_T^Y)$

Data-processing Principle: $d_{\mathrm{JS}}(\mathcal{D}_S^{\hat{Y}}, \mathcal{D}_T^{\hat{Y}}) \leq d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$

Lin's Lemma (1991): $d_{\mathrm{JS}}(\mathcal{D}^Y, \mathcal{D}^{\hat{Y}}) \leq \sqrt{\varepsilon(h \circ g)}$

An Information-Theoretic Lower Bound

Theorem: suppose the Markov chain holds and $d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, then

$$\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) \geq \frac{1}{2}\left(d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)\right)^2$$

Proof Sketch:

JS distance is a metric: $d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \leq d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_S^{\hat{Y}}) + d_{\mathrm{JS}}(\mathcal{D}_S^{\hat{Y}}, \mathcal{D}_T^{\hat{Y}}) + d_{\mathrm{JS}}(\mathcal{D}_T^{\hat{Y}}, \mathcal{D}_T^Y)$

Data-processing Principle: $d_{\mathrm{JS}}(\mathcal{D}_S^{\hat{Y}}, \mathcal{D}_T^{\hat{Y}}) \leq d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$

Lin's Lemma (1991): $d_{\mathrm{JS}}(\mathcal{D}^Y, \mathcal{D}^{\hat{Y}}) \leq \sqrt{\varepsilon(h \circ g)}$

Combining all three inequalities with AM-GM inequality finishes the proof.

An Information-Theoretic Lower Bound

Theorem: suppose the Markov chain holds and $d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, then

$$\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) \geq \frac{1}{2}\left(d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)\right)^2$$

An Information-Theoretic Lower Bound

Theorem: suppose the Markov chain holds and $d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, then

$$\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) \geq \frac{1}{2}\left(d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)\right)^2$$

- Lower bound still holds when separate transformations are used in source/target domains

An Information-Theoretic Lower Bound

Theorem: suppose the Markov chain holds and $d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) \geq d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)$, then

$$\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) \geq \frac{1}{2}\left(d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)\right)^2$$

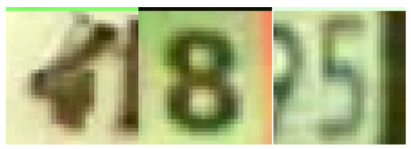
- Lower bound still holds when separate transformations are used in source/target domains
- The proof can be extended using other distance metrics as well

Experiments

Digit classification:

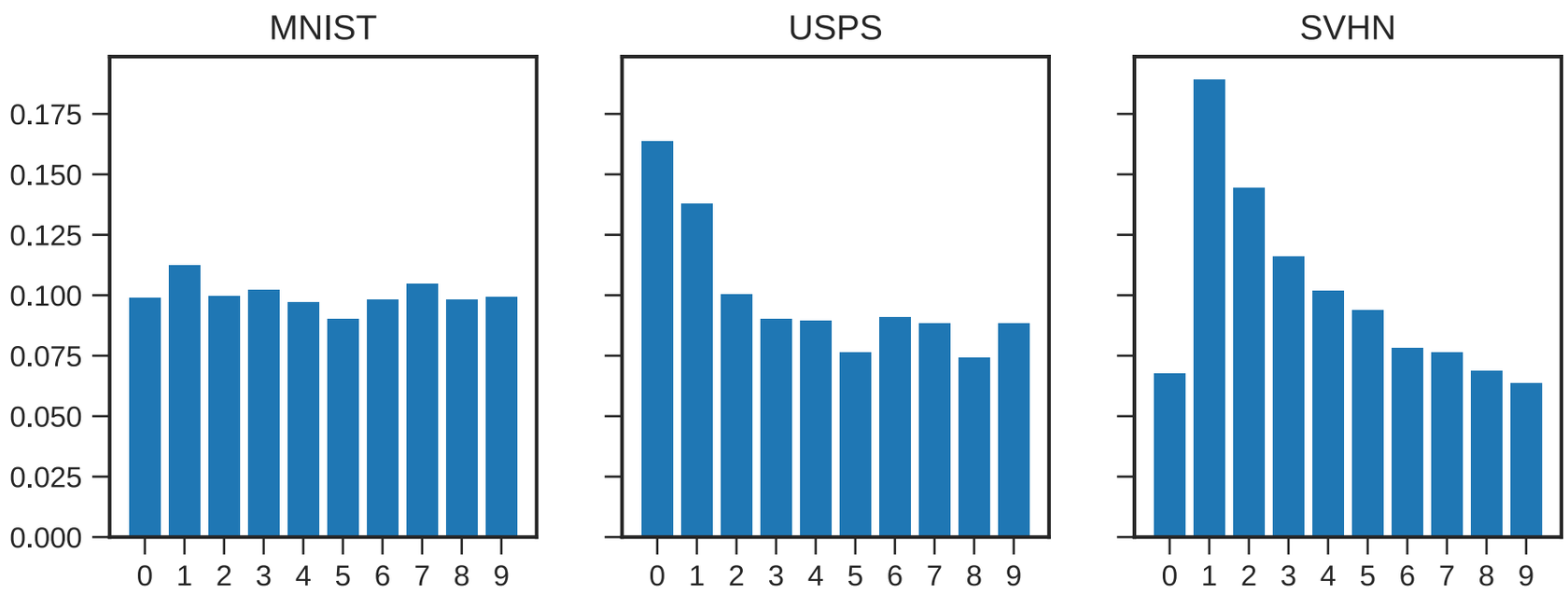
MNIST

SVHN



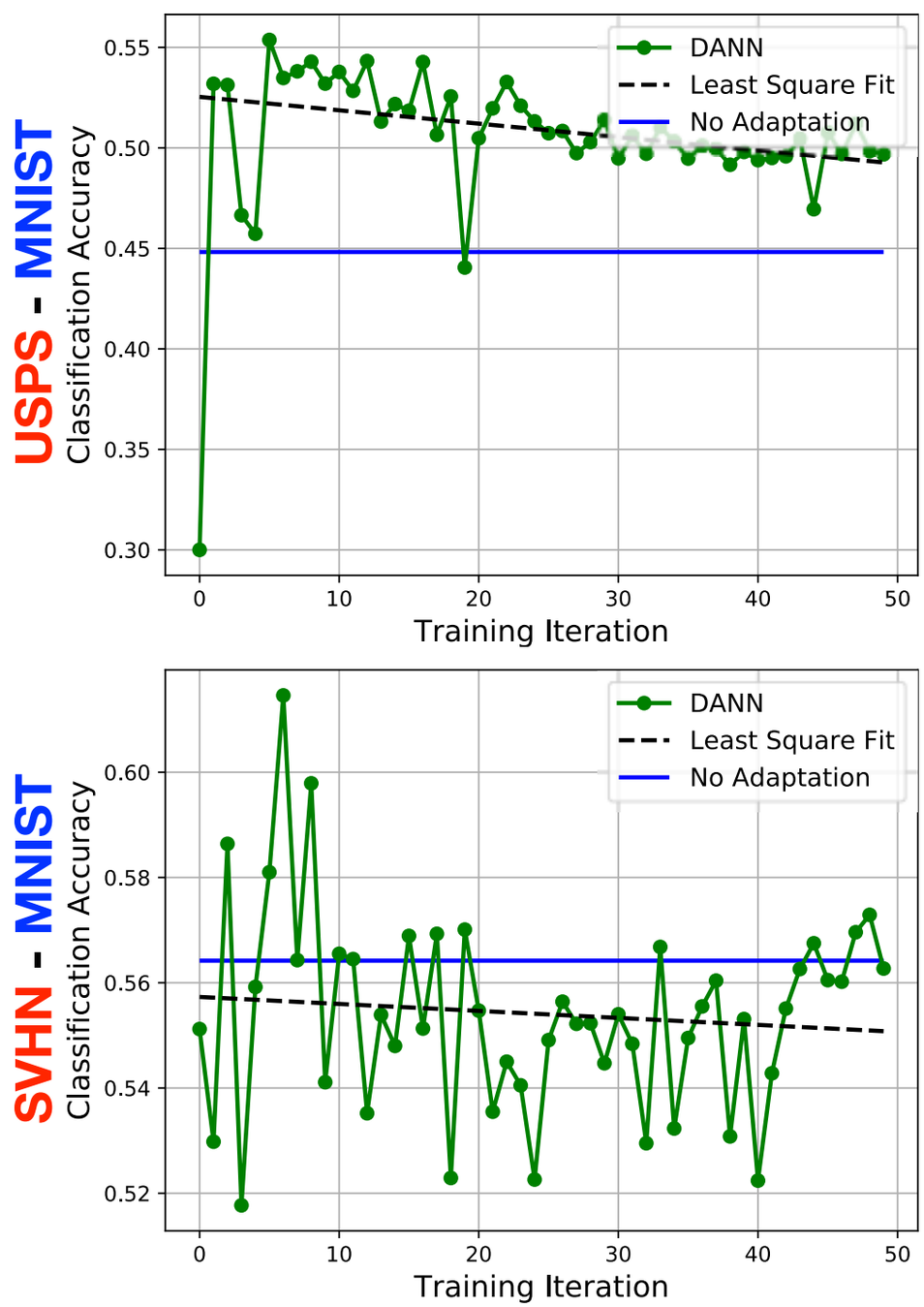
USPS

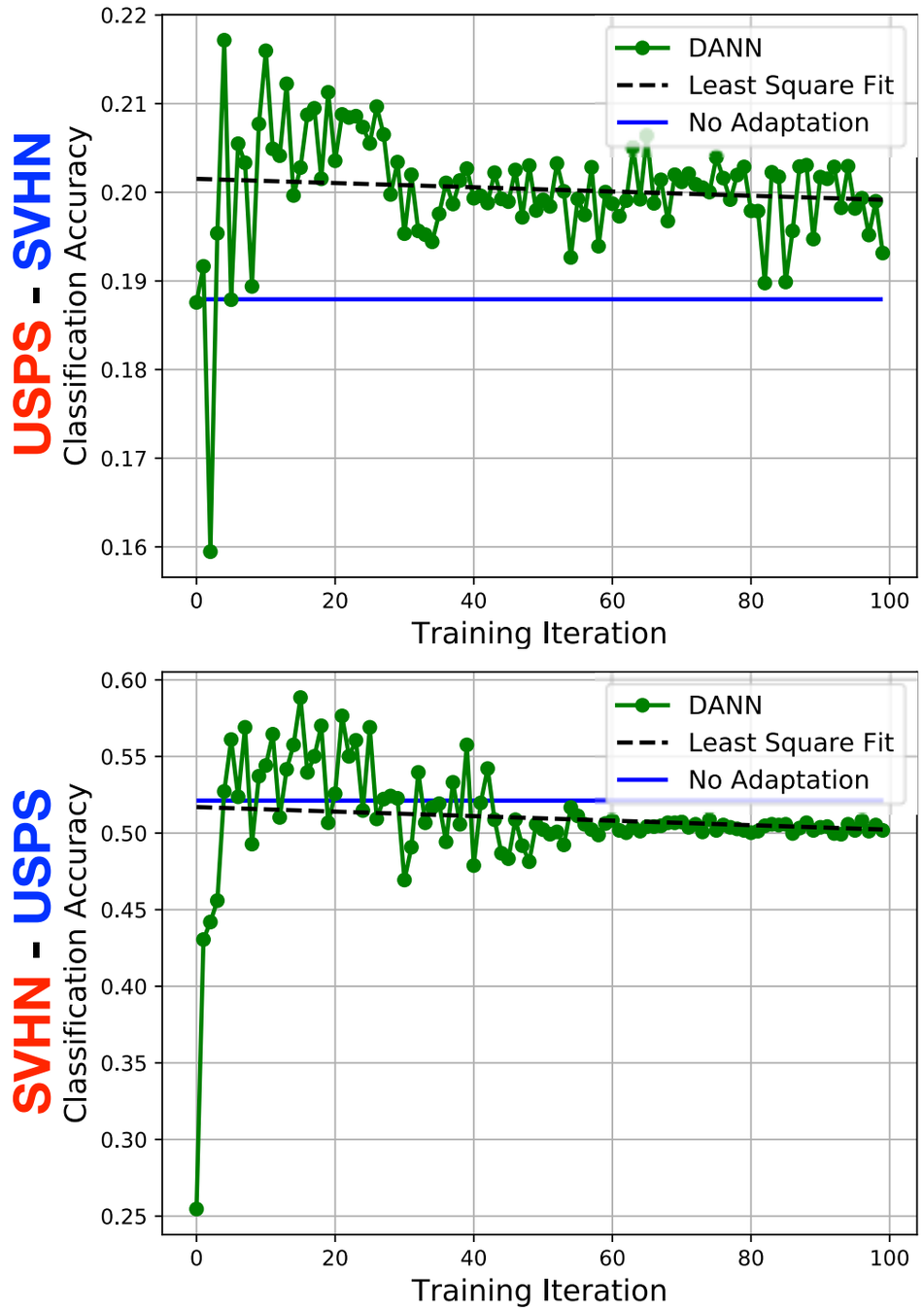
Marginal label distribution:



Experiments

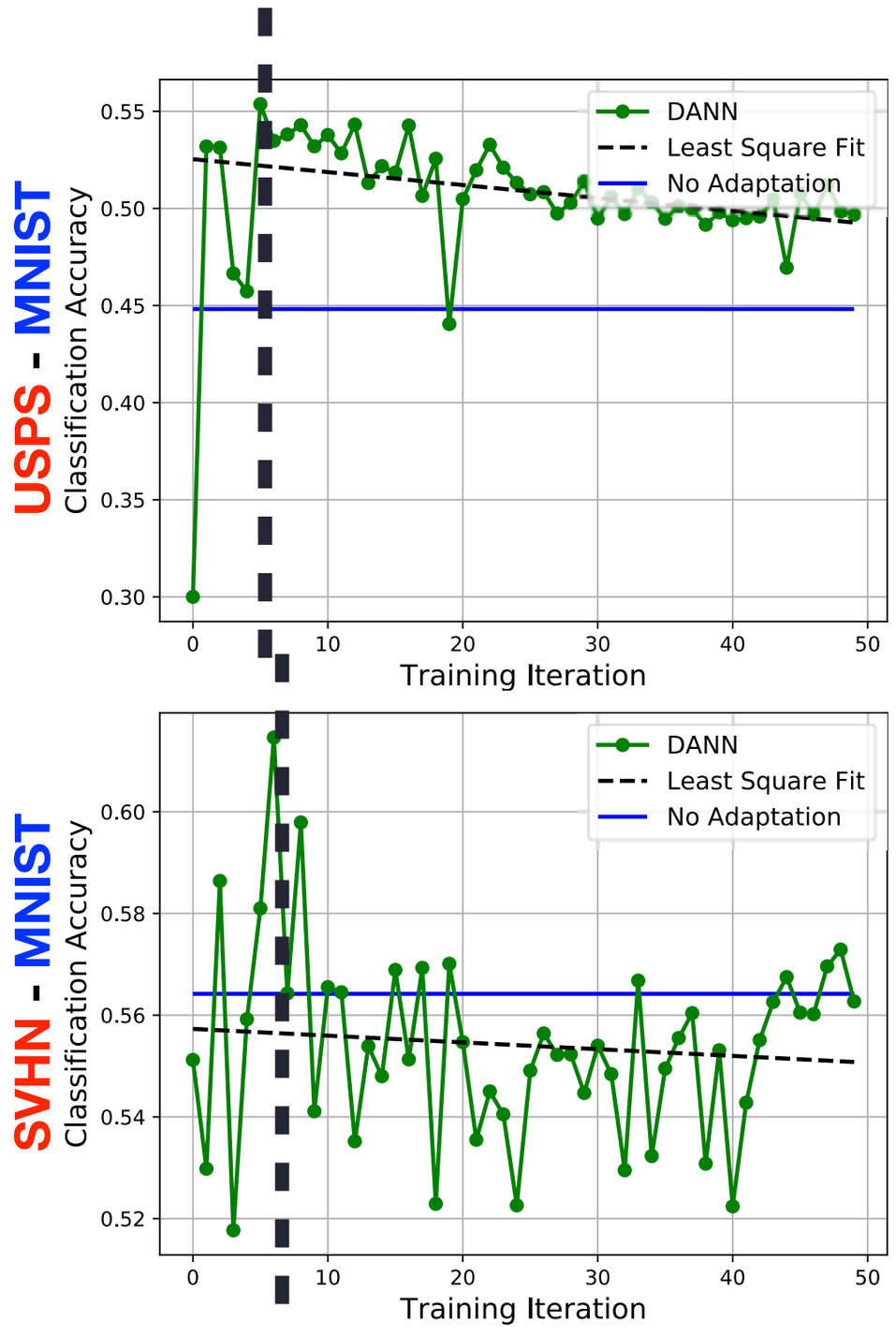
Source
Target

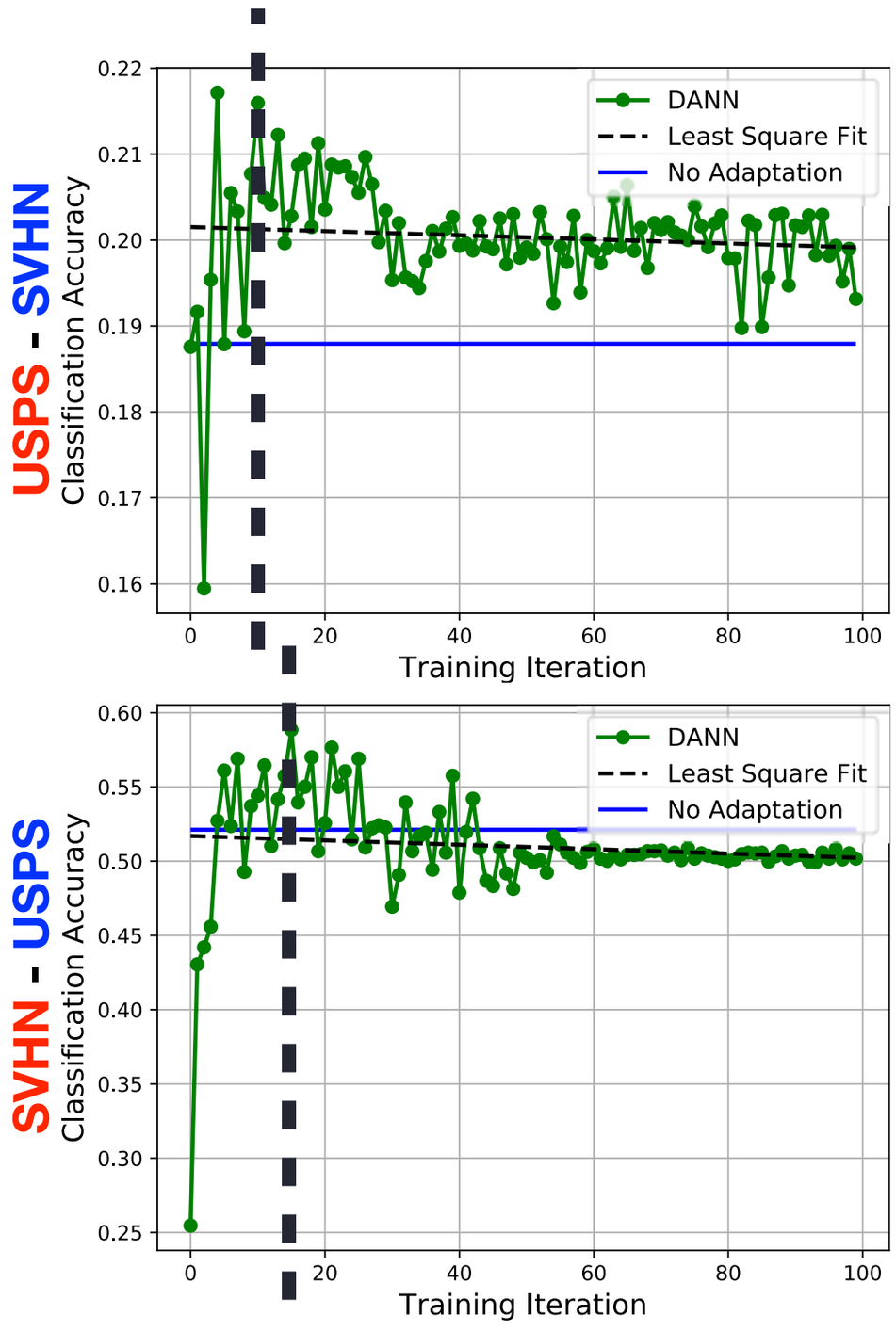




Experiments

Source
Target





Summary

- Sufficient condition: matching conditional distribution

$$\left|\varepsilon_S(h) - \varepsilon_T(h)\right| \leq d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) + \min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\}$$

- Necessary condition: matching marginal label distribution

$$\varepsilon_S(h \circ g) + \varepsilon_T(h \circ g) \geq \frac{1}{2}\left(d_{\mathrm{JS}}(\mathcal{D}_S^Y, \mathcal{D}_T^Y) - d_{\mathrm{JS}}(\mathcal{D}_S^Z, \mathcal{D}_T^Z)\right)^2$$

Collaborators:

Remi Tachet des Combes

Kun Zhang

Geoffrey J. Gordon

Poster: #71, Pacific Ballroom, Today