Feature-Based Approaches for Unsupervised Domain Adaptation

A. de Mathelin, M. Atiq

Sloan Kettering Institute - CEA

ECAS 2025

Notations

Let's introduce the following notations :

- **loss function** : $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
- **Hypothesis space** : $\mathcal{H}$, a functional space such that $h : \mathcal{X} \to \mathcal{Y}$, $\forall h \in \mathcal{H}$.

Notations

Let's introduce the following notations :

- **loss function** : $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
- **Hypothesis space** : $\mathcal{H}$, a functional space such that $h : \mathcal{X} \to \mathcal{Y}$, $\forall h \in \mathcal{H}$.

For any $h \in \mathcal{H}$, the **average loss** or **risk** over the joint distribution P_{XY} defined as follows :

$$\mathcal{L}_P(h) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(h(x), y)\, dP(x, y) \tag{1}$$

Notations

Let's introduce the following notations :

- **loss function** : $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
- **Hypothesis space** : $\mathcal{H}$, a functional space such that $h : \mathcal{X} \to \mathcal{Y}$, $\forall h \in \mathcal{H}$.

For any $h \in \mathcal{H}$, the **average loss** or **risk** over the joint distribution P_{XY} defined as follows :

$$\mathcal{L}_P(h) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(h(x), y)\, dP(x, y) \tag{1}$$

We also define the **empirical risk**, computed on the set of observations $\mathcal{S}$ drawn according to P_{XY} :

$$\mathcal{L}_{\mathcal{S}}(h) = \frac{1}{n} \sum_{(x,y)\in\mathcal{S}} \ell(h(x), y). \tag{2}$$

Notations

Let's introduce the following notations :

- **loss function** : $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
- **Hypothesis space** : $\mathcal{H}$, a functional space such that $h : \mathcal{X} \to \mathcal{Y}$, $\forall h \in \mathcal{H}$.

For any $h \in \mathcal{H}$, the **average loss** or **risk** over the joint distribution P_{XY} defined as follows :

$$\mathcal{L}_P(h) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(h(x), y)\, dP(x, y) \tag{1}$$

We also define the **empirical risk**, computed on the set of observations $\mathcal{S}$ drawn according to P_{XY} :

$$\mathcal{L}_{\mathcal{S}}(h) = \frac{1}{n} \sum_{(x,y)\in\mathcal{S}} \ell(h(x), y). \tag{2}$$

Additionally, we define the risk and its empirical estimation between two hypotheses $h, h' \in \mathcal{H}$ over the input distribution P_X as follows :

$$\mathcal{L}_{P_X}(h, h') = \int_{x\in\mathcal{X}} \ell(h(x), h'(x))\, dP_X(x) \tag{3}$$

$$\mathcal{L}_{\mathcal{S}_{\mathcal{X}}}(h, h') = \frac{1}{n} \sum_{x\in\mathcal{S}_{\mathcal{X}}} \ell(h(x), h'(x)). \tag{4}$$

Rademacher complexity

Definition (Rademacher complexity)

Let $\mathcal{G}$ be a set of functions mapping $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$. Given a sample $\mathcal{S} = \{(x_1, y_1)..., (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$, the empirical Rademacher complexity of $\mathcal{G}$ computed over $\mathcal{S}$ is defined as follows :

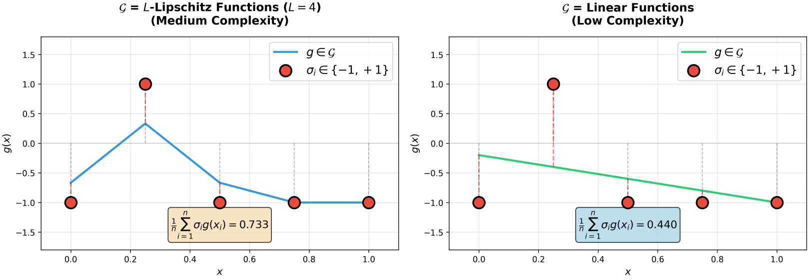
$$\mathfrak{R}_{\mathcal{S}}(\mathcal{G}) = \mathop{\mathrm{E}}_{\sigma \sim \mathcal{U}(\{-1,1\}^n)} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(x_i, y_i) \right] .$$

Rademacher complexity

Definition (Rademacher complexity)

Let $\mathcal{G}$ be a set of functions mapping $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$. Given a sample $\mathcal{S} = \{(x_1, y_1)..., (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$, the empirical Rademacher complexity of $\mathcal{G}$ computed over $\mathcal{S}$ is defined as follows :

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{G}) = \mathop{\mathrm{E}}_{\sigma \sim \mathcal{U}(\{-1,1\}^n)} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(x_i, y_i) \right] .$$

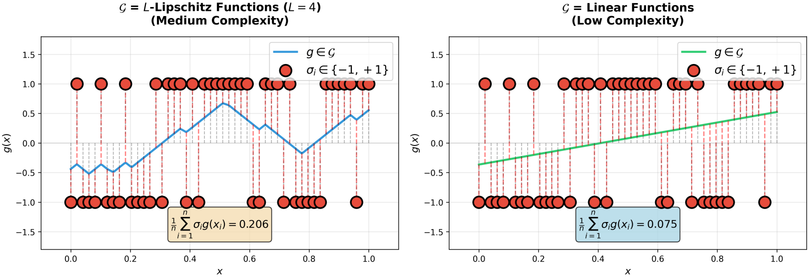


Rademacher complexity

Definition (Rademacher complexity)

Let $\mathcal{G}$ be a set of functions mapping $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$. Given a sample $\mathcal{S} = \{(x_1, y_1)..., (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$, the empirical Rademacher complexity of $\mathcal{G}$ computed over $\mathcal{S}$ is defined as follows :

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{G}) = \mathop{\mathbb{E}}_{\sigma \sim \mathcal{U}(\{-1,1\}^n)} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(x_i, y_i) \right] .$$



$\Longrightarrow$ If $\mathcal{H}$ is finite, we have $\mathfrak{R}_{\mathcal{S}}(\mathcal{G}) = \mathcal{O}(\sqrt{\log(|\mathcal{H}|)/n})$ [Mohri et al., 2018]

Traditional Learning Guarantees

Proposition 1 (cf. [Mohri et al., 2018])

Let P be a distribution over $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{S}$ a sample of size $n \in \mathbb{N}$ drawn iid according to P. Assuming $\mathcal{G} = \{(x, y) \to \ell(h(x), y);\ h \in \mathcal{H}\}$ bounded by $M > 0$, then, for any $h \in \mathcal{H}$ and for any $\delta > 0$, the following bound holds with probability at least $1 - \delta$:

$$\mathcal{L}_P(h) \leq \mathcal{L}_{\mathcal{S}}(h) + 2\,\mathfrak{R}_{\mathcal{S}}(\mathcal{G}) + 3M\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}\,. \tag{5}$$

Traditional Learning Guarantees

Proposition 1 (cf. [Mohri et al., 2018])

Let P be a distribution over $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{S}$ a sample of size $n \in \mathbb{N}$ drawn iid according to P. Assuming $\mathcal{G} = \{(x, y) \to \ell(h(x), y);\ h \in \mathcal{H}\}$ bounded by $M > 0$, then, for any $h \in \mathcal{H}$ and for any $\delta > 0$, the following bound holds with probability at least $1 - \delta$:

$$\mathcal{L}_P(h) \leq \mathcal{L}_\mathcal{S}(h) + 2\,\mathfrak{R}_\mathcal{S}(\mathcal{G}) + 3M\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}\,. \tag{5}$$

Remark : If $\mathcal{S}$ is indepent of h, then the Hoeffding's inequality implies that :

$$\mathcal{L}_P(h) \overset{1-\delta}{\leq} \mathcal{L}_\mathcal{S}(h) + M\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}\,. \tag{6}$$

Proposition 2

*Let P^s, P^t be a **source** and **target** distribution over $\mathcal{X} \times \mathcal{Y}$, we have :*

$$\mathcal{L}_{P^t}(h) \leq \mathcal{L}_{P^s}(h) + \underline{\mathrm{div}\left(P^s, P^t\right)}.$$

Consequently, under the same assumptions as in Proposition 2, for any $\delta > 0$, the following bound holds with probability at least $1 - \delta$:

$$\mathcal{L}_{P^t}(h) \leq \mathcal{L}_{\mathcal{S}}(h) + \underline{\mathrm{div}\left(P^s, P^t\right)} + 2\,\mathfrak{R}_{\mathcal{S}}(\mathcal{G}) + 3M\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}\,,$$

with $\mathrm{div}(P^s, P^t)$ a positive real number that measures the divergence between P^s and P^t.

Learning Guarantees Under Domain Shift

Proposition 2

Let P^s, P^t be a ***source*** *and* ***target*** *distribution over $\mathcal{X} \times \mathcal{Y}$, we have :*

$$\mathcal{L}_{P^t}(h) \leq \mathcal{L}_{P^s}(h) + \underline{\text{div}\left(P^s, P^t\right)}.$$

Consequently, under the same assumptions as in Proposition 2, for any $\delta > 0$, the following bound holds with probability at least $1 - \delta$:

$$\mathcal{L}_{P^t}(h) \leq \mathcal{L}_{\mathcal{S}}(h) + \underline{\text{div}\left(P^s, P^t\right)} + 2\,\mathfrak{R}_{\mathcal{S}}(\mathcal{G}) + 3M\sqrt{\frac{\log(\frac{2}{\delta})}{2n}},$$

with $\text{div}(P^s, P^t)$ a positive real number that measures the divergence between P^s and P^t.

Remark : In practice, $\text{div}(P^s, P^t)$ may depend on $\mathcal{H}$, ℓ, or h, in which case it is more precise to write $\text{div}(P^s, P^t, \mathcal{H}, \ell, h)$.

Learning Guarantees Under Domain Shift

Examples of $\text{div}(P^s, P^t)$:

Examples of $\text{div}(P^s, P^t)$:

Definition 1 ($\mathcal{Y}$-discrepancy [Mohri and Muñoz Medina, 2012])

Given two distributions, P^t, P^s defined over $\mathcal{X} \times \mathcal{Y}$, and an hypothesis set $\mathcal{H}$ of functions mapping $\mathcal{X}$ to $\mathcal{Y}$, the $\mathcal{Y}$-discrepancy between P^t and P^s is defined as follows :

$$\mathcal{Y}\text{-disc}_{\mathcal{H}}\left(P^t, P^s\right) = \sup_{h \in \mathcal{H}} |\mathcal{L}_{P^t}(h) - \mathcal{L}_{P^s}(h)| \tag{7}$$

Learning Guarantees Under Domain Shift

Examples of $\text{div}(P^s, P^t)$:

Definition 1 ($\mathcal{Y}$-discrepancy [Mohri and Muñoz Medina, 2012])

Given two distributions, P^t, P^s defined over $\mathcal{X} \times \mathcal{Y}$, and an hypothesis set $\mathcal{H}$ of functions mapping $\mathcal{X}$ to $\mathcal{Y}$, the $\mathcal{Y}$-discrepancy between P^t and P^s is defined as follows :

$$\mathcal{Y}\text{-disc}_{\mathcal{H}}\left(P^t, P^s\right) = \sup_{h \in \mathcal{H}} \left|\mathcal{L}_{P^t}(h) - \mathcal{L}_{P^s}(h)\right| \tag{7}$$

Definition 2 (discrepancy [Ben-David et al., 2007, Ben-David et al., 2010])

Let P^s_X, P^t_X be two marginal distributions over $\mathcal{X}$, the discrepancy between P^s_X and P^t_X is defined as follows :

$$\text{disc}_{\mathcal{H}}\left(P^s_X, P^t_X\right) = \sup_{h', h'' \in \mathcal{H}} \left|\mathcal{L}_{P^t_X}(h', h'') - \mathcal{L}_{P^s_X}(h', h'')\right| \tag{8}$$

Learning Guarantees Under Domain Shift

Proposition 3 (cf. [Ben-David et al., 2007, Ben-David et al., 2010])

Let P^t, P^s, two distributions defined over $\mathcal{X} \times \mathcal{Y}$, and an hypothesis set $\mathcal{H}$ of functions mapping $\mathcal{X}$ to $\mathcal{Y}$. If the loss function ℓ verifies the triangular inequality, the following bound holds for any $h \in \mathcal{H}$:

$$\mathcal{L}_{P^t}(h) \leq \mathcal{L}_{P^s}(h) + \mathrm{disc}_{\mathcal{H}}\left(P^s_X, P^t_X\right) + \epsilon \tag{9}$$

with,

$$\mathrm{disc}_{\mathcal{H}}\left(P^s_X, P^t_X\right) = \sup_{h', h'' \in \mathcal{H}} \left| \mathcal{L}_{P^t_X}(h', h'') - \mathcal{L}_{P^s_X}(h', h'') \right| \tag{10}$$

and,

$$\epsilon = \epsilon\left(P^s, P^t\right) = \inf_{h \in \mathcal{H}} \left(\mathcal{L}_{P^s}(h) + \mathcal{L}_{P^t}(h)\right) \tag{11}$$

Learning Guarantees Under Domain Shift

Proposition 3 (cf. [Ben-David et al., 2007, Ben-David et al., 2010])

Let P^t, P^s, two distributions defined over $\mathcal{X} \times \mathcal{Y}$, and an hypothesis set $\mathcal{H}$ of functions mapping $\mathcal{X}$ to $\mathcal{Y}$. If the loss function ℓ verifies the triangular inequality, the following bound holds for any $h \in \mathcal{H}$:

$$\mathcal{L}_{P^t}(h) \leq \mathcal{L}_{P^s}(h) + \mathrm{disc}_{\mathcal{H}}\left(P^s_X, P^t_X\right) + \epsilon \tag{9}$$

with,

$$\mathrm{disc}_{\mathcal{H}}\left(P^s_X, P^t_X\right) = \sup_{h', h'' \in \mathcal{H}} \left| \mathcal{L}_{P^t_X}(h', h'') - \mathcal{L}_{P^s_X}(h', h'') \right| \tag{10}$$

and,

$$\epsilon = \epsilon\left(P^s, P^t\right) = \inf_{h \in \mathcal{H}} \left(\mathcal{L}_{P^s}(h) + \mathcal{L}_{P^t}(h)\right) \tag{11}$$

Remark : If there exists a hypothesis h that fits both domains well, then ϵ is small.

Divergences Between Distributions

$$\mathrm{div}\left(P_X^s, P_X^t\right) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}_{P_X^t}\left[g(X)\right] - \mathbb{E}_{P_X^s}\left[g(X)\right] \right|.$$

Divergences Between Distributions

$$\text{div}\left(P_X^s, P_X^t\right) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}_{P_X^t}[g(X)] - \mathbb{E}_{P_X^s}[g(X)] \right| .$$

Divergence	Δ ; $\text{div}\left(P_X^s, P_X^t\right) = \sup_{g \in \mathcal{G}} (\Delta)$	$\mathcal{G}$
Discrepancy	$\left| \mathbb{E}_{P_X^t}[g(X)] - \mathbb{E}_{P_X^s}[g(X)] \right|$	$\{g : x \to \ell(h(x), h'(x)); (h, h') \in \mathcal{H}^2\}$
MMD	idem	$\{g : \mathcal{X} \to \mathbb{R}; \ \|\|g\|\|_{\mathscr{H}} \leq 1\}$
Wasserstein-1	idem	$\{g : \mathcal{X} \to \mathbb{R}; \ \|\|g\|\|_{\text{Lip}} \leq 1\}$
$\mathcal{H}$-divergence	idem	$\mathcal{H}$ (with $\mathcal{Y} = \{0, 1\}$)
TV	idem	$\{g : \mathcal{X} \to [0, 1]\}$
KL	$\mathbb{E}_{P_X^t}[g(X)] - \mathbb{E}_{P_X^s}\left[e^{g(X)-1}\right]$	$\{g : \mathcal{X} \to \mathbb{R}\}$

Divergences Between Distributions

$$\operatorname{div}\left(P_X^s, P_X^t\right) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}_{P_X^t}[g(X)] - \mathbb{E}_{P_X^s}[g(X)] \right|.$$

Divergence	Δ ; $\operatorname{div}\left(P_X^s, P_X^t\right) = \sup_{g \in \mathcal{G}} (\Delta)$	$\mathcal{G}$
Discrepancy	$\left\| \mathbb{E}_{P_X^t}[g(X)] - \mathbb{E}_{P_X^s}[g(X)] \right\|$	$\{g : x \to \ell(h(x), h'(x)); (h, h') \in \mathcal{H}^2\}$
MMD	idem	$\{g : \mathcal{X} \to \mathbb{R};\ \|\|g\|\|_{\mathscr{H}} \leq 1\}$
Wasserstein-1	idem	$\{g : \mathcal{X} \to \mathbb{R};\ \|\|g\|\|_{\text{Lip}} \leq 1\}$
$\mathcal{H}$-divergence	idem	$\mathcal{H}$ (with $\mathcal{Y} = \{0, 1\}$)
TV	idem	$\{g : \mathcal{X} \to [0, 1]\}$
KL	$\mathbb{E}_{P_X^t}[g(X)] - \mathbb{E}_{P_X^s}\left[e^{g(X)-1}\right]$	$\{g : \mathcal{X} \to \mathbb{R}\}$

$\Longrightarrow$ All divergences verify Proposition 3 ($\mathcal{L}_{P^t}(h) \leq \mathcal{L}_{P^s}(h) + \operatorname{div}\left(P_X^s, P_X^t\right) + \epsilon$) under appropriate assumptions on $\mathcal{H}$ and ℓ.

Divergences Between Distributions

Divergence	Primal Formulation
Discrepancy	NA
$\mathcal{H}$-divergence	NA
MMD	$\left\|\left\| \mathbb{E}_{P_X^s}\left[\psi(X)\right] - \mathbb{E}_{P_X^t}\left[\psi(X)\right] \right\|\right\|_{\mathscr{H}}$
Wasserstein-1	$\inf_{\gamma \in \Gamma\left(P_X^s, P_X^t\right)} \mathbb{E}_{(X,X') \sim \gamma}\left[\|\|X - X'\|\|_1\right]$
TV	$\frac{1}{2}\int_{x \in \mathcal{X}} \|dP_X^s(x) - dP_X^t(x)\| dx$
KL	$\mathbb{E}_{P_X^t}\left[\log\left(dP_X^t(X) \big/ dP_X^s(X)\right)\right]$

Table – $\mathscr{H}$ is a reproducing kernel Hilbert space with $\psi : \mathcal{X} \to \mathscr{H}$ the corresponding feature map. $\Gamma(P_X^s, P_X^t)$ is the set of all joint probability measure on $\mathcal{X} \times \mathcal{X}$ whose marginals are P_X^s and P_X^t on the first and second factors.

Why considering different divergences ?

Why considering different divergences ?

- Divergences are computed on empirical distributions. The emprirical Wasserstein converges as $\mathcal{O}(1/n^{1/p})$ (dimension p), while MMD as $\mathcal{O}(1/\sqrt{n})$. For high-dimensional $\mathcal{X}$, MMD may be preferred.

Divergences Between Distributions

Why considering different divergences ?

- Divergences are computed on empirical distributions. The emprirical Wasserstein converges as $\mathcal{O}(1/n^{1/p})$ (dimension p), while MMD as $\mathcal{O}(1/\sqrt{n})$. For high-dimensional $\mathcal{X}$, MMD may be preferred.
- Computational complexity varies : $\mathcal{H}$-divergence reduces to fitting a binary classifier, MMD requires $\mathcal{O}(n^2)$ kernel computations, which can be costly for large datasets.

Why considering different divergences ?

- Divergences are computed on empirical distributions. The emprirical Wasserstein converges as $\mathcal{O}(1/n^{1/p})$ (dimension p), while MMD as $\mathcal{O}(1/\sqrt{n})$. For high-dimensional $\mathcal{X}$, MMD may be preferred.
- Computational complexity varies : $\mathcal{H}$-divergence reduces to fitting a binary classifier, MMD requires $\mathcal{O}(n^2)$ kernel computations, which can be costly for large datasets.
- Discrepancy is task-specific and gives tighter generalization bounds, but is hard to compute for large or unbounded hypothesis spaces.

Why considering different divergences ?

- Divergences are computed on empirical distributions. The emprirical Wasserstein converges as $\mathcal{O}(1/n^{1/p})$ (dimension p), while MMD as $\mathcal{O}(1/\sqrt{n})$. For high-dimensional $\mathcal{X}$, MMD may be preferred.
- Computational complexity varies : $\mathcal{H}$-divergence reduces to fitting a binary classifier, MMD requires $\mathcal{O}(n^2)$ kernel computations, which can be costly for large datasets.
- Discrepancy is task-specific and gives tighter generalization bounds, but is hard to compute for large or unbounded hypothesis spaces.
- Choice depends on problem : KL can be infinite for disjoint supports, while Wasserstein and discrepancy remain finite. KL is suitable for importance weighting when target support is included in the source.

Feature-based Approach

For a divergence defined previously, we have :

$$\mathcal{L}_{P^t}(h) \leq \mathcal{L}_{P^s}(h) + \underline{\mathrm{div}\left(P^s, P^t\right)} + \epsilon\,,$$

For a divergence defined previously, we have :

$$\mathcal{L}_{P^t}(h) \leq \mathcal{L}_{P^s}(h) + \underline{\text{div}\left(P^s, P^t\right)} + \epsilon\,,$$

Then, for any feature maps $\phi^s, \phi^t : \mathcal{X} \to \mathcal{Z}$ and any $h : \mathcal{Z} \to \mathcal{Y}$, we have :

$$\mathcal{L}_{P^t}(h \circ \phi^t) \leq \mathcal{L}_{P^s}(h \circ \phi^s) + \underline{\text{div}\left(\phi^s\left(P^s_X\right), \phi^t\left(P^t_X\right)\right)} + \epsilon\,,$$

Feature-based Approach

For a divergence defined previously, we have :

$$\mathcal{L}_{P^t}(h) \leq \mathcal{L}_{P^s}(h) + \underline{\text{div}\left(P^s, P^t\right)} + \epsilon\,,$$

Then, for any feature maps $\phi^s, \phi^t : \mathcal{X} \to \mathcal{Z}$ and any $h : \mathcal{Z} \to \mathcal{Y}$, we have :

$$\mathcal{L}_{P^t}(h \circ \phi^t) \leq \mathcal{L}_{P^s}(h \circ \phi^s) + \underline{\text{div}\left(\phi^s\left(P^s_X\right), \phi^t\left(P^t_X\right)\right)} + \epsilon\,,$$

$\Longrightarrow$ Find feature map $\hat{\phi}^s, \hat{\phi}^t : \mathcal{X} \to \mathcal{Z}$ and hypothesis $\hat{h}$ such that :

$$\hat{\phi}^s, \hat{\phi}^t, \hat{h} = \underset{\phi^s, \phi^t, h}{\text{argmin}}\, \mathcal{L}_{\hat{P}^s}(h \circ \phi^s) + \text{div}\left(\phi^s\left(\hat{P}^s_X\right), \phi^t\left(\hat{P}^t_X\right)\right)$$

Feature-based Domain Adaptation

Two-stage Approach (Asymetric)

$$\hat{\phi}^s, \hat{\phi}^t = \underset{\phi^s, \phi^t : \mathcal{X} \to \mathcal{Z}}{\operatorname{argmin}} \operatorname{div}\left(\phi^s\left(\hat{P}_X^s\right), \phi^t\left(\hat{P}_X^t\right)\right)$$

$$\hat{h} = \underset{h : \mathcal{Z} \to \mathcal{Y}}{\operatorname{argmin}} \mathcal{L}_{\hat{P}_X^s}(h \circ \hat{\phi}^s)$$

Feature-based Domain Adaptation

Two-stage Approach (Asymetric)

$$\hat{\phi}^s, \hat{\phi}^t = \underset{\phi^s, \phi^t : \mathcal{X} \to \mathcal{Z}}{\operatorname{argmin}} \ \operatorname{div}\left(\phi^s\left(\hat{P}^s_X\right), \phi^t\left(\hat{P}^t_X\right)\right)$$

$$\hat{h} = \underset{h : \mathcal{Z} \to \mathcal{Y}}{\operatorname{argmin}} \ \mathcal{L}_{\hat{P}^s_X}(h \circ \hat{\phi}^s)$$

Two-stage Approach (Symetric)

$$\hat{\phi} = \underset{\phi : \mathcal{X} \to \mathcal{Z}}{\operatorname{argmin}} \ \operatorname{div}\left(\phi\left(\hat{P}^s_X\right), \phi\left(\hat{P}^t_X\right)\right)$$

$$\hat{h} = \underset{h : \mathcal{Z} \to \mathcal{Y}}{\operatorname{argmin}} \ \mathcal{L}_{\hat{P}^s_X}(h \circ \hat{\phi}^s)$$

Feature-based Domain Adaptation

Two-stage Approach (Asymetric)

$$\hat{\phi}^s, \hat{\phi}^t = \underset{\phi^s, \phi^t : \mathcal{X} \to \mathcal{Z}}{\text{argmin}} \ \text{div}\left(\phi^s\left(\hat{P}_X^s\right), \phi^t\left(\hat{P}_X^t\right)\right)$$

$$\hat{h} = \underset{h: \mathcal{Z} \to \mathcal{Y}}{\text{argmin}} \ \mathcal{L}_{\hat{P}_X^s}(h \circ \hat{\phi}^s)$$

Two-stage Approach (Symetric)

$$\hat{\phi} = \underset{\phi : \mathcal{X} \to \mathcal{Z}}{\text{argmin}} \ \text{div}\left(\phi\left(\hat{P}_X^s\right), \phi\left(\hat{P}_X^t\right)\right)$$

$$\hat{h} = \underset{h: \mathcal{Z} \to \mathcal{Y}}{\text{argmin}} \ \mathcal{L}_{\hat{P}_X^s}(h \circ \hat{\phi}^s)$$

One-stage Approach (Symetric)

$$\hat{h}, \hat{\phi} = \underset{h, \phi}{\text{argmin}} \ \mathcal{L}_{\hat{P}^s}(h \circ \phi) + \text{div}\left(\phi\left(\hat{P}_X^s\right), \phi\left(\hat{P}_X^t\right), \underline{h}\right)$$

Feature-based Assumptions

Perfect Matching Exists

Feature-based approaches assume that there exist ϕ^* such that :

$$P^s(\phi^*(X), Y) = P^t(\phi^*(X), Y) \ \text{(symetric)} \tag{12}$$

or ϕ^s, ϕ^t such that :

$$P^s(\phi^s(X), Y) = P^t(\phi^t(X), Y) \ \text{(asymetric)} \tag{13}$$

It is also implicitely assumed that ϕ^* is the $\mathrm{div}\left(P^s(\phi(X)), P^s(\phi(X))\right)$ minimizer.

Feature-based Assumptions

Perfect Matching Exists

Feature-based approaches assume that there exist ϕ^* such that :

$$P^s(\phi^*(X), Y) = P^t(\phi^*(X), Y) \ \ \text{(symetric)} \tag{12}$$

or ϕ^s, ϕ^t such that :

$$P^s(\phi^s(X), Y) = P^t(\phi^t(X), Y) \ \ \text{(asymetric)} \tag{13}$$

It is also implicitely assumed that ϕ^* is the $\text{div}\,(P^s(\phi(X)), P^s(\phi(X)))$ minimizer.

No Label Shift

Feature-based approaches implicitely assume that there is no label shift :

$$P^s(Y) = P^t(Y)$$

Feature-based Assumptions

Perfect Matching Exists

Feature-based approaches assume that there exist ϕ^* such that :

$$P^s(\phi^*(X), Y) = P^t(\phi^*(X), Y) \ \ \text{(symetric)} \tag{12}$$

or ϕ^s, ϕ^t such that :

$$P^s(\phi^s(X), Y) = P^t(\phi^t(X), Y) \ \ \text{(asymetric)} \tag{13}$$

It is also implicitely assumed that ϕ^* is the $\text{div}\left(P^s(\phi(X)), P^s(\phi(X))\right)$ minimizer.

No Label Shift

Feature-based approaches implicitely assume that there is no label shift :

$$P^s(Y) = P^t(Y)$$

Indeed :

$$P^s(\phi(X), Y) = P^t(\phi(X), Y) \implies \int_{\mathcal{X}} P^s(\phi(x), Y)dx = \int_{\mathcal{X}} P^t(\phi(x), Y)dx \implies P^s(Y) = P^t(Y)$$

Feature-based Mapping Examples

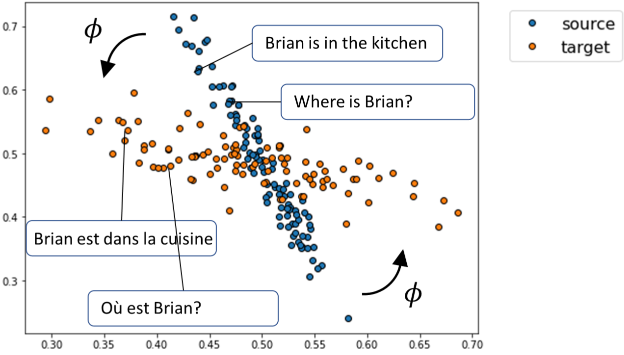


Figure – **Rotation** In this example, the space of feature transformation is $\Phi = \{x \to Mx^T; M \in \mathbb{R}^{p \times p}, M^T M = \mathrm{Id}_p\}$ and is used to match a dataset of english sentences to a dataset of french sentences.

Feature-based Mapping Examples

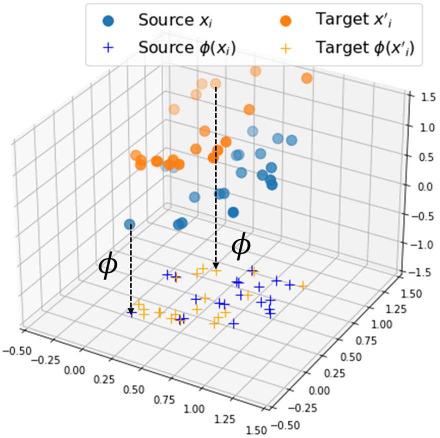


Figure – **Projection** In this example, the space of feature transformation is $\Phi = \{x \to Px^T; P \in \mathrm{Proj}(\mathbb{R}^{p\times p})\}$

Feature-based Mapping Examples

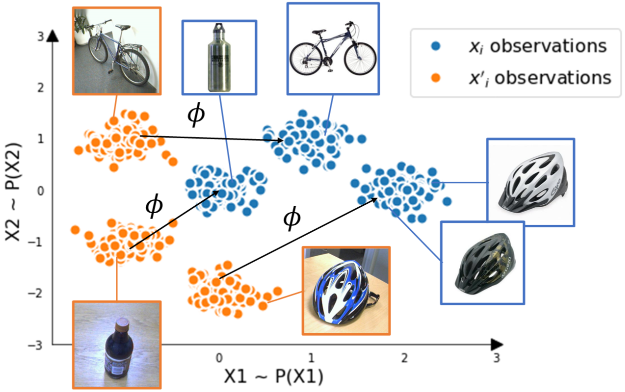


Figure – **Continuous Transformation** In this example, the space of feature transformation is $\Phi = \mathcal{C}_0(\mathcal{X})$ and is used to match a dataset of pictures from Amazon to a dataset of webcam pictures.

Feature-based Mapping Examples

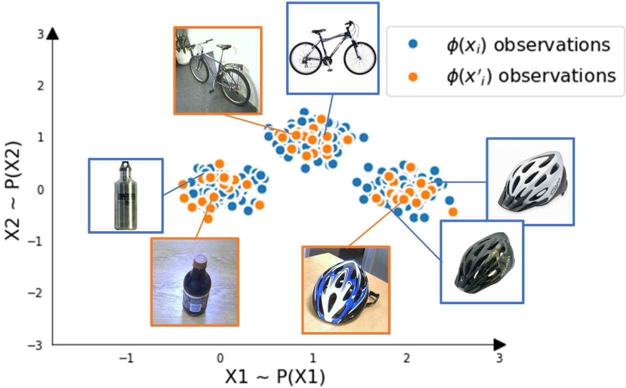


Figure – **Continuous Transformation** In this example, the space of feature transformation is $\Phi = \mathcal{C}_0(\mathcal{X})$ and is used to match a dataset of pictures from Amazon to a dataset of webcam pictures.

Subspace Alignment (SA)

The purpose of Subspace Alignment [Fernando et al., 2013] is to find a linear transformation of the source PCA eigenvectors which match as close as possible the target PCA eigenvectors :

$$\widehat{\phi_T^*} = x \to xW_T^d$$
$$\widehat{\phi_S^*} = x \to xW_S^d M^*$$
$$M^* = \underset{M \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} ||W_S^d M - W_T^d||_2^2$$

Where W_T^d and W_S^d are respectively the matrixes of the d first eigenvectors of the target and source PCA.

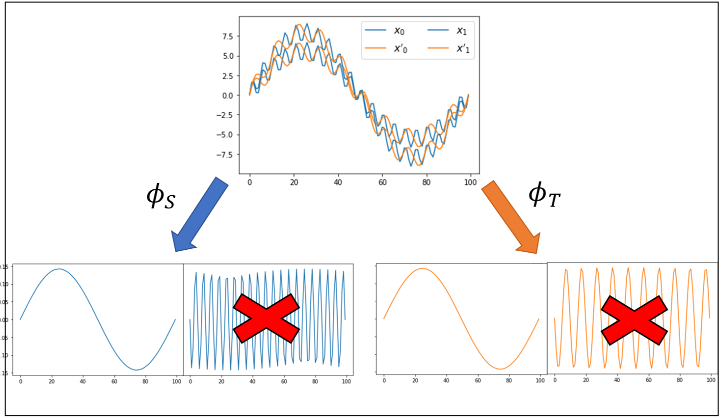


Figure – **Subspace Alignment** In this example, the source dataset is composed of signals : $t \rightarrow 5X_1 \sin(2\pi t) + X_2 \sin(40\pi t)$ and the target dataset of signals : $t \rightarrow 5X_1 \sin(2\pi t) + X_2 \sin(20\pi t)$ $(X_1, X_2 \sim \mathcal{U}([0, 2]))$

Correlation Alignment (CORAL)

The purpose of Correlation Alignment [Sun et al., 2016] is to find a linear transformation of the source data which covariance matrix match as cloas as possible the target data covariance matrix :

$$\widehat{\phi_S^*} = x \rightarrow xA^*$$

$$A^* = \underset{A \in \mathbb{R}^{p \times p}}{\operatorname{argmin}} ||A^T C_S A - C_T||_2^2$$

Where C_T and C_S are respectively the covariance matrixes of the target and source dataset.

Correlation Alignment Example

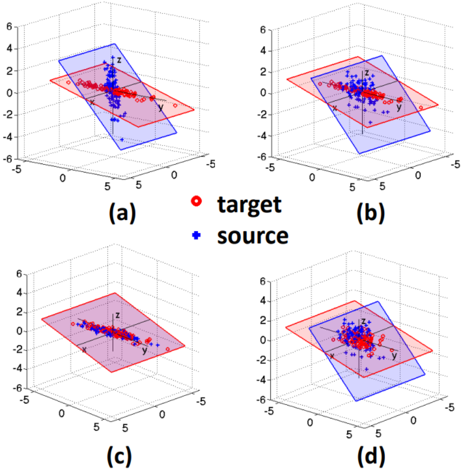


Figure – **Correlation Alignment** (source [Sun et al., 2016]) (a) : initial situation, (b) : source data are "whitened", (c) : source data are "re-colored" with the target covariance.

Feature-based approach : Optimal Transport

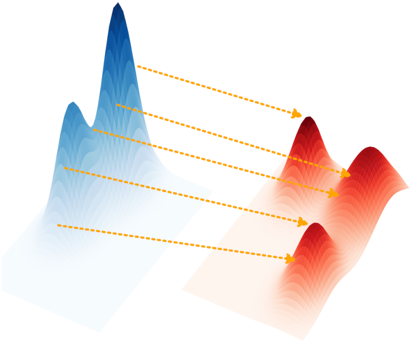


Figure – Optimal Transport consists in finding the minimal-cost mapping to move one distribution onto another distribution. Each arrow indicates the transport plan between the two distributions.

(Image Source : Laboratoire Hubert Curien's Data Intelligence Team)

Feature-based approach : Optimal Transport

Optimal Transport for Domain Adaptation (OTDA)

The purpose of OTDA [Courty et al., 2016] is to find the optimal transportation from the source data to the target data :

$$\widehat{\phi_S^*} = x \to \sum_{x' \in \mathcal{T}} \gamma^*(x, x')x'$$

$$\gamma^* = \operatorname*{argmin}_{\gamma \in \Gamma} \sum_{x \in \mathcal{S}} \sum_{x' \in \mathcal{T}} \gamma(x, x')||x - x'||_2^2$$

Where, for any $\gamma \in \Gamma$, $\gamma : \mathcal{S} \times \mathcal{T} \to [0, 1]$ and :

$$\sum_{x \in \mathcal{S}} \gamma(x, x') = 1 \ \forall x' \in \mathcal{T}$$

$$\sum_{x' \in \mathcal{T}} \gamma(x, x') = 1 \ \forall x \in \mathcal{S}$$

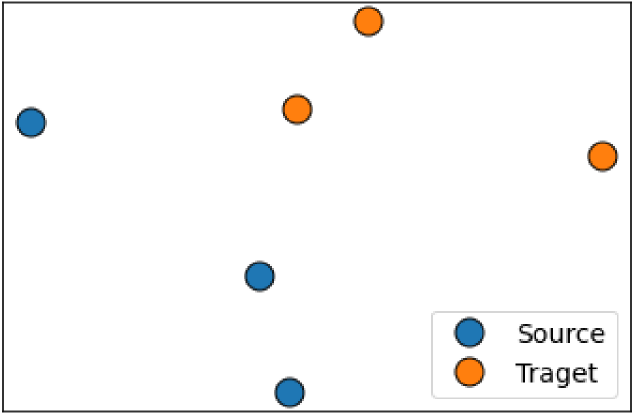


Figure – **OTDA** : In this example, we are looking at the optimal pairing between three source and target data points. The optimal pairing minimizes the sum of distances between paired data points.

OTDA

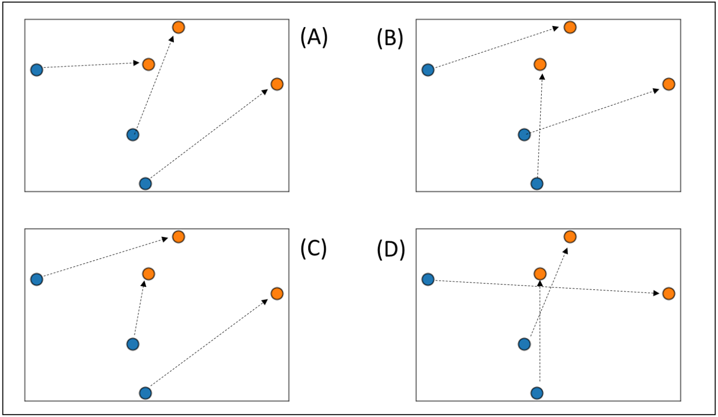


Figure – **OTDA** : In this example, we are looking at the optimal pairing between three source and target data points. The optimal pairing minimizes the sum of distances between paired data points.

OTDA

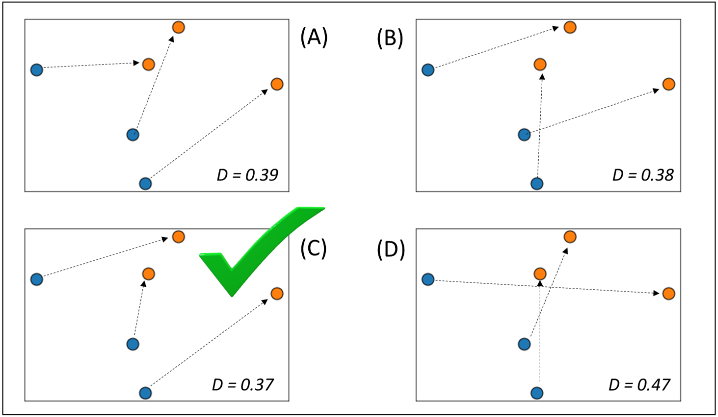


Figure – **OTDA** : The optimal pairing is the pairing (C). Finding the optimal pairing for large datasets is often intractable, we are then looking at approximated solution.

Feature-Based Approach : Adversarial Neural Networks

Let $\Phi = \{\phi \in \mathrm{NN}; \phi : \mathcal{X} \to \mathcal{Z}\}$ and $\mathcal{H} = \{h \in \mathrm{NN}; h : \mathcal{Z} \to \mathcal{Y}\}$

Feature-Based Approach : Adversarial Neural Networks

Let $\Phi = \{\phi \in \mathrm{NN}; \phi : \mathcal{X} \to \mathcal{Z}\}$ and $\mathcal{H} = \{h \in \mathrm{NN}; h : \mathcal{Z} \to \mathcal{Y}\}$

One-stage Approach (Symetric)

$$\hat{h}, \hat{\phi} = \underset{h,\,\phi}{\mathrm{argmin}}\ \mathcal{L}_{\hat{P}^s}(h \circ \phi) + \mathrm{div}\Big(\phi\left(\hat{P}^s_X\right), \phi\left(\hat{P}^t_X\right)\Big)$$

Let $\Phi = \{\phi \in \text{NN}; \phi : \mathcal{X} \to \mathcal{Z}\}$ and $\mathcal{H} = \{h \in \text{NN}; h : \mathcal{Z} \to \mathcal{Y}\}$

One-stage Approach (Symetric)

$$\hat{h}, \hat{\phi} = \underset{h,\, \phi}{\text{argmin}} \;\; \mathcal{L}_{\hat{P}^s}(h \circ \phi) + \text{div}\Big(\phi\left(\hat{P}_X^s\right), \phi\left(\hat{P}_X^t\right)\Big)$$

Let's consider the $\mathcal{H}$-divergence, with $\mathcal{Y} = \{0, 1\}$ and $\ell = L_{01}$:

$$\mathcal{H}\text{-div}\Big(\phi\left(\hat{P}_X^s\right), \phi\left(\hat{P}_X^t\right)\Big) = \max_{h' \in \mathcal{H}} \left| \mathbb{E}_{\hat{P}_X^s}[h' \circ \phi(x)] - \mathbb{E}_{\hat{P}_X^t}[h' \circ \phi(x)] \right|$$

Feature-Based Approach : Adversarial Neural Networks

Let $\Phi = \{\phi \in \text{NN}; \phi : \mathcal{X} \rightarrow \mathcal{Z}\}$ and $\mathcal{H} = \{h \in \text{NN}; h : \mathcal{Z} \rightarrow \mathcal{Y}\}$

One-stage Approach (Symetric)

$$\hat{h}, \hat{\phi} = \underset{h,\,\phi}{\text{argmin}} \;\; \mathcal{L}_{\hat{P}^s}(h \circ \phi) + \text{div}\Big(\phi\left(\hat{P}_X^s\right), \phi\left(\hat{P}_X^t\right)\Big)$$

Let's consider the $\mathcal{H}$-divergence, with $\mathcal{Y} = \{0, 1\}$ and $\ell = L_{01}$:

$$\mathcal{H}\text{-div}\Big(\phi\left(\hat{P}_X^s\right), \phi\left(\hat{P}_X^t\right)\Big) = \max_{h' \in \mathcal{H}} \left| \mathbb{E}_{\hat{P}_X^s}[h' \circ \phi(x)] - \mathbb{E}_{\hat{P}_X^t}[h' \circ \phi(x)] \right|$$

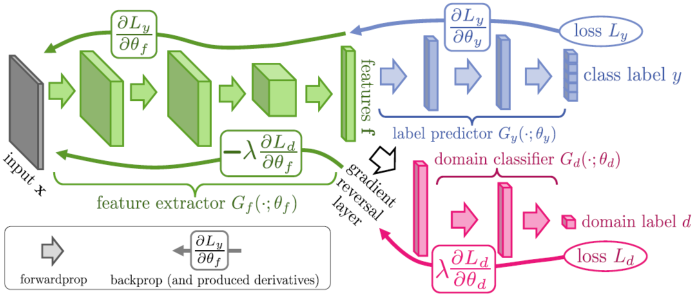
Then, the optimization become,

$$\min_{h,\,\phi} \max_{h'} \;\; \mathcal{L}_{\hat{P}^s}(h \circ \phi) + \lambda \left| \mathbb{E}_{\hat{P}_X^s}[h' \circ \phi(x)] - \mathbb{E}_{\hat{P}_X^t}[h' \circ \phi(x)] \right|$$

Adversarial Neural Networks : DANN

Discriminative Adversarial Neural Network (DANN) [Ganin et al., 2016]

$$\min_{h,\phi} \max_{h'} \; \mathcal{L}_{\hat{P}^s}(h \circ \phi) + \lambda \left(\mathbb{E}_{\hat{P}^s_X}[\log\left(1 - h' \circ \phi(x)\right)] - \mathbb{E}_{\hat{P}^t_X}[\log\left(h' \circ \phi(x)\right)] \right)$$

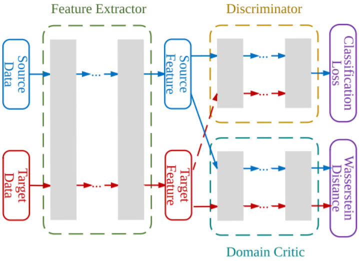


Adversarial Neural Networks : WDGRL

Wasserstein Distance Guided Representation Learning (WDGRL) [Shen et al., 2018]

$$\min_{h,\,\phi} \mathcal{L}_{\hat{P}^s}(h \circ \phi) + \lambda \left(\mathbb{E}_{\hat{P}^s_X}[h' \circ \phi(x)] - \mathbb{E}_{\hat{P}^t_X}[h' \circ \phi(x)] \right)$$

$$\max_{h'} \left(\mathbb{E}_{\hat{P}^s_X}[h' \circ \phi(x)] - \mathbb{E}_{\hat{P}^t_X}[h' \circ \phi(x)] \right) + \mathbb{E}_{\alpha \sim \mathcal{U}([0,1])} \mathbb{E}_{x \sim \hat{P}^s_X} \mathbb{E}_{x' \sim \hat{P}^t_X} \left[|| h' \left(\alpha\phi(x) + (1-\alpha)\phi(x') \right) - 1 ||_2^2 \right]$$



Bibliography

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010).
A theory of learning from different domains.
Mach. Learn., 79(1-2) :151–175.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2007).
Analysis of representations for domain adaptation.
In Schölkopf, B., Platt, J. C., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 137–144. MIT Press.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2016).
Optimal transport for domain adaptation.
IEEE transactions on pattern analysis and machine intelligence, 39(9) :1853–1865.

Fernando, B., Habrard, A., Sebban, M., and Tuytelaars, T. (2013).
Unsupervised visual domain adaptation using subspace alignment.
In *Proceedings of the IEEE international conference on computer vision*, pages 2960–2967.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016).
Domain-adversarial training of neural networks.
J. Mach. Learn. Res., 17(1) :2096–2030.

Mohri, M. and Muñoz Medina, A. (2012).
New analysis and algorithm for learning with drifting distributions.
In Bshouty, N. H., Stoltz, G., Vayatis, N., and Zeugmann, T., editors, *Algorithmic Learning Theory*, pages 124–138, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018).
Foundations of machine learning.
MIT press.

Shen, J., Qu, Y., Zhang, W., and Yu, Y. (2018).
Wasserstein distance guided representation learning for domain adaptation.