



Flexible Domain Adaptation for Multimedia Indexing

Emilie Morvant, Amaury Habrard, Stéphane Ayache

► To cite this version:

Emilie Morvant, Amaury Habrard, Stéphane Ayache. Flexible Domain Adaptation for Multimedia Indexing. 2011. hal-00634881

HAL Id: hal-00634881

<https://hal.science/hal-00634881>

Preprint submitted on 24 Oct 2011

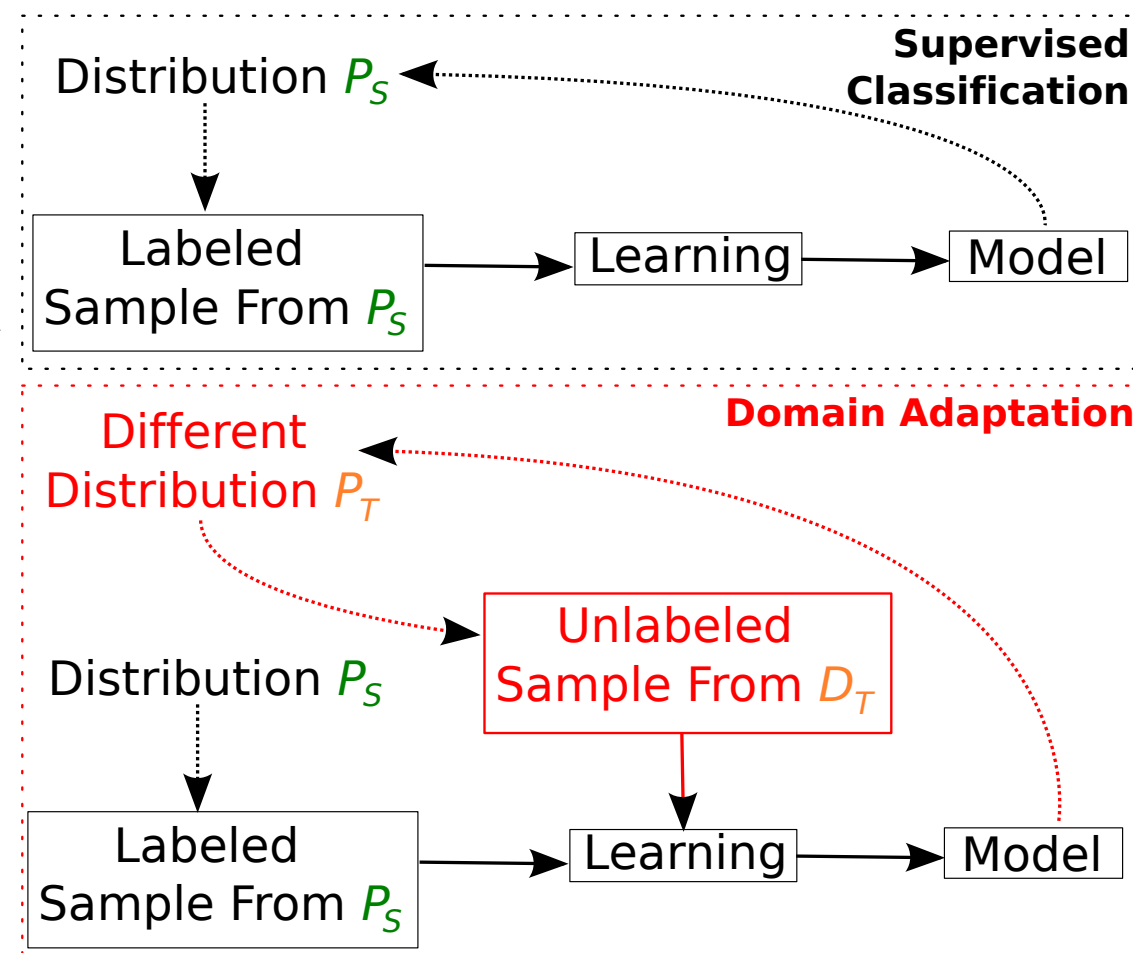
HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

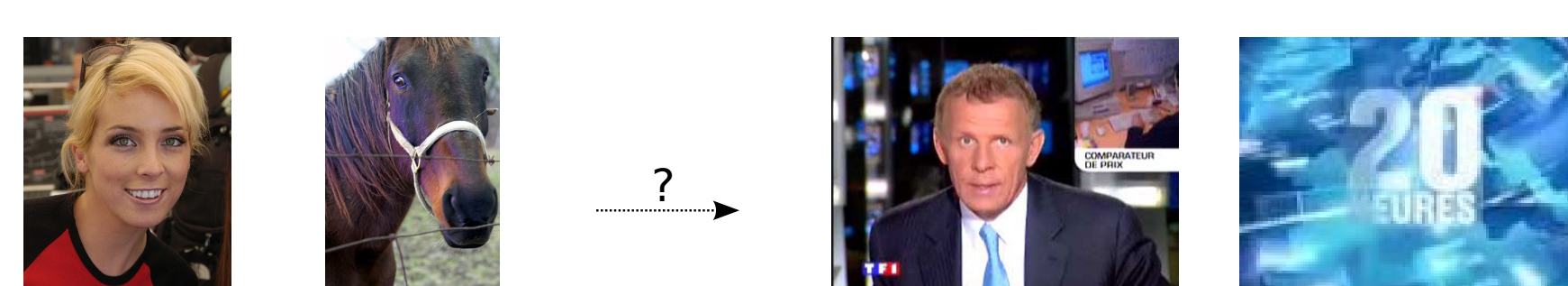
INTRODUCTION, NOTATIONS AND MOTIVATION

We consider binary classification task:

- X input space, $Y = \{-1, 1\}$ label set
- P_S **source** domain: distribution over $X \times Y$
 D_S marginal distribution over X
- P_T **target** domain: different distribution over $X \times Y$
 D_T marginal distribution over X
- errors of a hypothesis $h : X \rightarrow Y$
 - $err_S(h)$, $\bar{err}_S(h)$ **source** domain errors
 - $err_T(h)$, $\bar{err}_T(h)$ **target** domain errors
- **Supervised Classification objective:**
 - $h \in \mathcal{H}$ with a **low** $err_S(h)$
- **Domain Adaptation objective:**
 - $h \in \mathcal{H}$ with a **low** $err_T(h)$



For example:

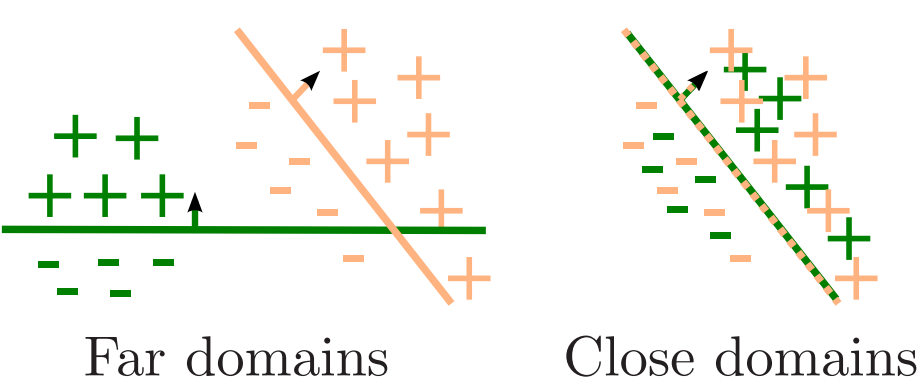
- We have **labeled** images from a **Web image corpus**, i.e. $\sim P_S$
 - Is there a **Person** in **unlabeled** images from a **Video corpus**, i.e. $\sim D_T$?
- 
- Person no Person $P_S \neq P_T$
- ⇒ The **Learning** distribution is **different** from the **Testing** distribution
 ⇒ How can we learn, from the **source domain**, a low-error classifier on the **target domain** ?

DOMAIN ADAPTATION

Theorem 1 ([2]). Let $\mathcal{H}$ an hypothesis space. If D_S and D_T are respectively the marginal distributions of source and target instances, then for all $\delta \in]0, 1]$, with probability at least $1 - \delta$, for every $h \in \mathcal{H}$:

$$err_T(h) \leq err_S(h) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu,$$

where $d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T)$ is the $\mathcal{H}\Delta\mathcal{H}$ -distance between D_S and D_T
 and $\nu = err_S(h^*) + err_T(h^*)$, with $h^* = \operatorname{argmin}_{h \in \mathcal{H}} (err_S(h) + err_T(h))$.



Idea
 Build a new projection space to move closer the domains.

LEARNING WITH GOOD SIMILARITY FUNCTIONS

Definition 1 ([1]). $K : X \times X \rightarrow [-1, 1]$ is an (ϵ, γ, τ) -good similarity function for a binary classification problem P if

(i) A $1 - \epsilon$ probability mass of examples $(\mathbf{x}, y)$ satisfy

$$\mathbb{E}_{(\mathbf{x}', y') \sim P} [yy' K(\mathbf{x}, \mathbf{x}') | R(\mathbf{x}')] \geq \gamma,$$

(ii) $\Pr_{\mathbf{x}'} [R(\mathbf{x}')] \geq \tau$ (Notation: R set of reasonable points).

Properties

- Generalization of kernels: K may be **not** symmetric and **not** PSD
- A low-error linear classifier can be learned by minimizing the Pb. (SF) in the explicit projection space defined by $R = \{\mathbf{x}'_j\}_{j=1}^{d_u}$:

$$\phi^R(\cdot) = \langle K(\cdot, \mathbf{x}'_1), \dots, K(\cdot, \mathbf{x}'_{d_u}) \rangle$$

(Notation: $\mathcal{H}_{SF}$ the hypothesis space of such classifiers)

DOMAIN ADAPTATION OF LINEAR CLASSIFIERS BASED ON GOOD SIMILARITY FUNCTIONS

Building a new projection $\phi_{new}^{R'}$ to move closer D_S and D_T with $\mathcal{C}_{ST}$ a pair set $(\mathbf{x}_s, \mathbf{x}_t) \in U_S \sim D_S \times U_T \sim D_T$ such that the deviation of losses of $\mathbf{x}_s$ and $\mathbf{x}_t$ is small

$$\left| \left[1 - y \sum_{j=1}^{d'_u} \alpha_j K(\mathbf{x}_s, \mathbf{x}'_j) \right]_+ - \left[1 - y \sum_{j=1}^{d'_u} \alpha_j K(\mathbf{x}_t, \mathbf{x}'_j) \right]_+ \right| \leq \frac{\left\| ({}^t\phi^{R'}(\mathbf{x}_s) - {}^t\phi^{R'}(\mathbf{x}_t)) \operatorname{diag}(\boldsymbol{\alpha}) \right\|_1}{\left\| {}^t\phi_{new}^{R'}(\mathbf{x}_s) - {}^t\phi_{new}^{R'}(\mathbf{x}_t) \right\|_1} \implies \phi_{new}^{R'}(\cdot) = \left\langle \underbrace{\alpha_1 K(\cdot, x'_1)}_{K_{new}(\cdot, x'_1)}, \dots, \underbrace{\alpha_{d_u} K(\cdot, x'_{d_u})}_{K_{new}(\cdot, x'_{d_u})} \right\rangle$$

Our global optimization problem

With $\{(\mathbf{x}_i, y_i)\}_{i=1}^{d_l}$ i.i.d. from P_S , $R' = \{\mathbf{x}'_j\}_{j=1}^{d'_u}$ and $\mathcal{C}_{ST}$ a set of pairs $(\mathbf{x}_s, \mathbf{x}_t) \sim D_S \times D_T$.

At iteration l , we build the $\phi_{l+1}^{R'}$ space with the help of $\boldsymbol{\alpha}^l$ inferred by (SF)

$$\min_{\boldsymbol{\alpha}^l} \frac{1}{d_l} \sum_{i=1}^{d_l} \left[1 - y_i \sum_{j=1}^{d'_u} \alpha_j^l K_l(\mathbf{x}_i, \mathbf{x}'_j) \right]_+ + \lambda \|\boldsymbol{\alpha}^l\|_1 + \beta \sum_{(\mathbf{x}_s, \mathbf{x}_t) \in \mathcal{C}_{ST}} \left\| ({}^t\phi_l^{R'}(\mathbf{x}_s) - {}^t\phi_l^{R'}(\mathbf{x}_t)) \operatorname{diag}(\boldsymbol{\alpha}^l) \right\|_1 \quad (\text{DASF})$$

Some little results

- **Sparsity Analysis.** With $B_R = \min_{\mathbf{x}'_j \in R} \left\{ \max_{(\mathbf{x}_s, \mathbf{x}_t) \in \mathcal{C}_{ST}} |K(\mathbf{x}_s, \mathbf{x}'_j) - K(\mathbf{x}_t, \mathbf{x}'_j)| \right\}$,

$$\|\boldsymbol{\alpha}^*\|_1 \leq \frac{1}{\beta B_R + \lambda}$$

- **Generalization bound.** Following the robustness notion of Xu&Mannor [3], the Problem (DASF) is $(2M_\eta, \frac{N_\eta}{\beta B_R + \lambda})$ **robust on P_S** , with $\eta > 0$, M_η is the η -covering number of X and $N_\eta = \max_{\substack{\mathbf{x}_a, \mathbf{x}_b \sim D_S \\ \rho(\mathbf{x}_a, \mathbf{x}_b) \leq \eta}} \|{}^t\phi^R(\mathbf{x}_a) - {}^t\phi^R(\mathbf{x}_b)\|_\infty$. Thus for any $h \in \mathcal{H}_{SF}$, for any $\delta > 0$, with probability at least $1 - \delta$,

$$err_T(h) \leq \widehat{err}_S(h) + \frac{N_\eta}{\beta B_R + \lambda} + \sqrt{\frac{4M_\eta \ln 2 + 2 \ln \frac{1}{\delta}}{d_l}} + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(D_S, D_T) + \nu.$$

EXPERIMENTS ON MULTIMEDIA INDEXING

Experimental Setup

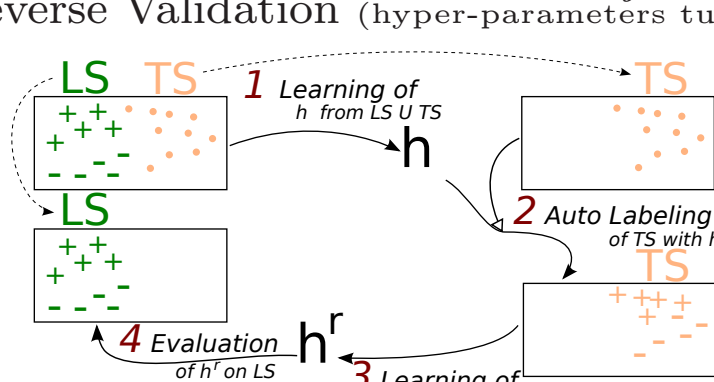
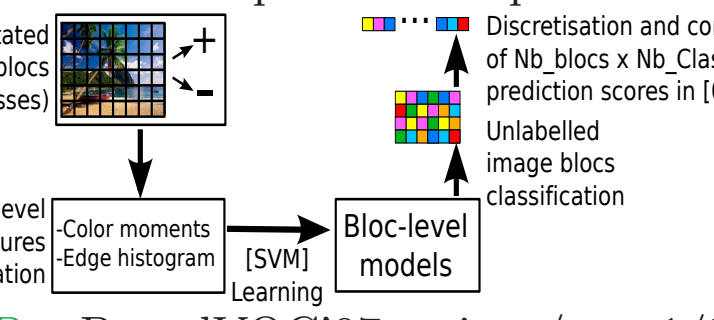
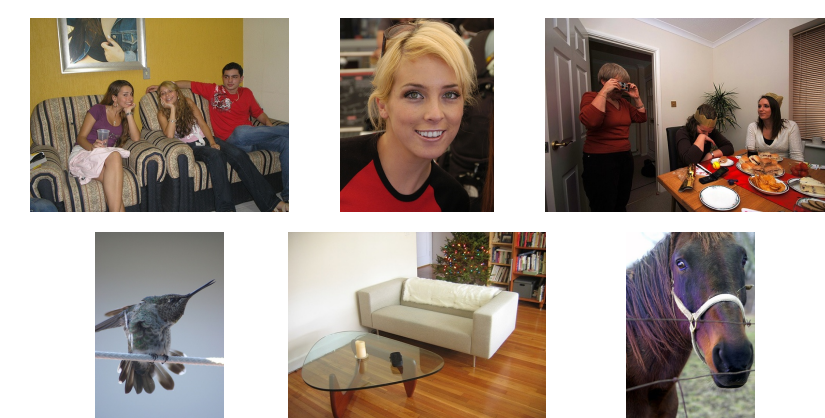
- Similarity function:
 κ Gaussian kernel, $K^*(\cdot, \mathbf{x}'_j) = \frac{K(\cdot, \mathbf{x}'_j) - \mu_{\mathbf{x}'_j}}{\sigma_{\mathbf{x}'_j}} \in [-1, 1]$
 - Reverse Validation (hyper-parameters tuning)
- 
- Toy problem “inter-twinning moons”
 - 1 P_S , 8 P_T according to 8 rotations
 - Image Indexing (according to F-measure)
 - Visual descriptors: Percepts
- 
- P_S : PascalVOC'07 ratio $\pm 1/3$
 - P_T : $\pm 1/3$: PascalVOC'07 Test $\pm 1/3$: TrecVid'07

Image Indexing: PascalVOC'07 Vs PascalVOC'07

Conc.	bird	boat	bottle	bus	car	cat	chair	cycle	cow	dining table	dog	horse	monitor	motorbike	person	plane	plants	sheep	sofa	train
SVM	0.18	0.29	0.01	0.16	0.28	0.23	0.24	0.10	0.15	0.15	0.24	0.31	0.16	0.17	0.56	0.34	0.12	0.16	0.16	0.36
SV	867	351	587	476	1096	882	1195	392	681	534	436	761	698	670	951	428	428	261	631	510
SF	0.18	0.27	0.11	0.12	0.34	0.20	0.21	0.10	0.11	0.10	0.18	0.24	0.12	0.17	0.46	0.34	0.13	0.12	0.13	0.20
Reas.	237	203	233	212	185	178	241	139	239	253	200	247	203	243	226	178	236	128	224	202
TSVM	0.14	0.14	0.11	0.16	0.37	0.14	0.22	0.13	0.12	0.13	0.22	0.17	0.12	0.12	0.44	0.18	0.10	0.12	0.15	0.19
SV	814	704	718	445	631	779	864	390	888	515	704	828	861	861	1111	585	406	474	866	652
DASVM	0.16	0.22	0.11	0.14	0.37	0.20	0.23	0.14	0.11	0.15	0.22	0.23	0.12	0.14	0.55	0.30	0.12	0.13	0.17	0.28
SV	922	223	295	421	866	1011	1418	706	335	536	180	802	668	841	303	356	1434	246	486	407
DASF	0.20	0.32	0.12	0.17	0.38	0.23	0.26	0.16	0.16	0.16	0.25	0.32	0.16	0.18	0.58	0.35	0.15	0.20	0.18	0.42
Reas.	50	184	78	94	51	378	229	192	203	372	391	384	287	239	6	181	293	153	167	75

The reasonable points for **Person**:



Toy Problem

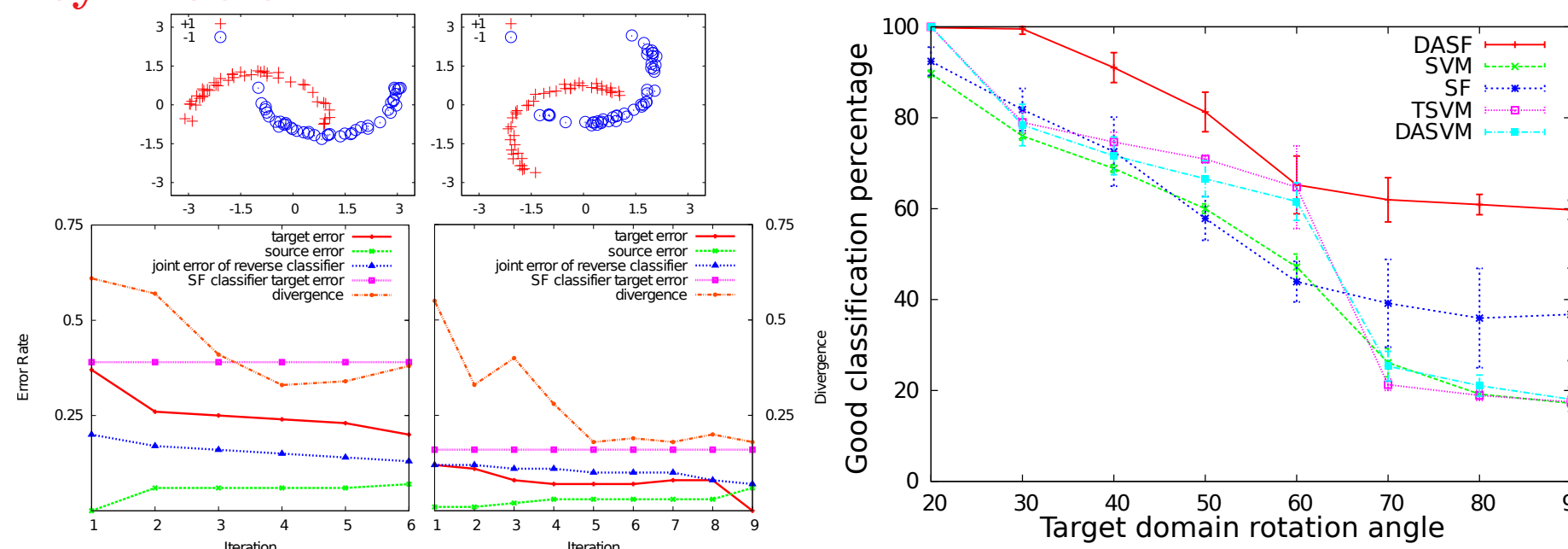


Image Indexing: PascalVOC'07 Vs TrecVid'07

Conc.	boat	bus	car	monitor	person	plane	Conc.	Average
SVM	0.56	0.25	0.43	0.19	0.52	0.32	SVM	0.38
SV	351	476	1096	698	951	428	SV	667
SF	0.49	0.46	0.50	0.34	0.45	0.54	SF	0.46
Reas.	214	224	176	246	226	178	Reas.	211
TSVM	0.56	0.48	0.52	0.37	0.46	0.61	TSVM	0.50
SV	498	535	631	741	1024	259	SV	615
DASVM	0.52	0.46	0.55	0.30	0.54	0.52	DASVM	0.48
SV	202	222	627	523	274	450	SV	383
DASF	0.57	0.49	0.55	0.42	0.57	0.66	DASF	0.54
Reas.	120	130	254	151	19	7	Reas.	113

REFERENCES

- [1] M.-F. Balcan, A. Blum, and N. Srebro. Improved guarantees for learning via similarity functions. In *Proceedings of COLT*, pages 287–298, 2008.
- [2] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J.W. Vaughan. A theory of learning from different domains. *Machine Learning Journal*, 79(1-2):151–175, 2010.
- [3] H. Xu and S. Mannor. Robustness and generalization. In *Proceedings of COLT*, pages 503–515, 2010.