Supplementary Material for Support and Invertibility in Unsupervised Domain Adaptation

Fredrik D. Johansson MIT
David Sontag MIT
Rajesh Ranganath NYU

A Proofs

A.1 Proof of bounds for support sufficiency divergence

Lemma 1. *The support sufficiency divergence is bounded with* $0 \leq d^{\epsilon}_{\text{supp}}(p \parallel q) \leq 1$*, and the bounds are tight.*

Proof. The lower bound holds and is tight because

$$\begin{aligned} d^{\epsilon}_{\text{supp}}(p \parallel q) &= \int_x (q(x) - p(x))\delta_{p,q}(x) \\ &= \int_x \max(q(x) - p(x), 0)\mathbb{1}[p(x) \leq \epsilon]\mathbb{1}[p(x) \leq \epsilon] \, . \end{aligned}$$

which is clearly non-negative. Moreover, for $\epsilon \leq \inf_x p(x)$, $d^{\epsilon}_{\text{supp}}(p \parallel q) = 0$. The upper bound holds trivially as $\delta_{p,q}(x) \leq 1$. For tightness, let q, p be discrete densities over two states, $q = [1., 0.]$ and $p = [0., 1.]$. Then with $\epsilon > 0$, $d^{\epsilon}_{\text{supp}}(p \parallel q) = 1$. □

Recall that

$$w^{\epsilon}_{p,q}(x) = \begin{cases} q(x)/p(x) & \text{if } p(x) \geq \epsilon \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

A.2 Proof of Lemma 1

Lemma 2. *Let* p, q *be densities over* $\mathcal{X}$*. Further, define* $\delta_{p,q}(x) = \mathbb{1}[p(x) \leq \epsilon \text{ and } p(x) \leq q(x)]$*. Then,*

$$\mathbb{E}_q[f(x)] \leq \mathbb{E}_p\left[w^{\epsilon}_{p,q}(x)f(x)\right] + M \cdot \left(\mathbb{E}_q[\delta^{\epsilon}_{p,q}(x)] - \mathbb{E}_p[\delta^{\epsilon}_{p,q}(x)]\right)$$

Proof. We have,

$$\begin{aligned} \mathop{\mathbb{E}}_{q}[f(x)] &= \int_x q(x)f(x)dx && \text{(By def.)} \\ &= \int_{x:p(x)>\epsilon} q(x)f(x)dx + \int_{x:p(x)\leq\epsilon} q(x)f(x)dx && \text{(Dividing domain)} \\ &= \int_{x:p(x)>\epsilon} \frac{q(x)}{p(x)}p(x)f(x)dx + \int_{x:p(x)\leq\epsilon} q(x)f(x)dx && \text{(Multiply and divide)} \\ &\leq \mathbb{E}_p\left[w^{\epsilon}_{p,q}(x)f(x)\right] + \int_{x:p(x)\leq\epsilon} (q(x) - p(x))f(x)dx && \text{(Add and subtract p(x))} \\ &\leq \mathbb{E}_p\left[w^{\epsilon}_{p,q}(x)f(x)\right] + M \int_{\substack{x:p(x)\leq\epsilon \\ p(x)\leq q(x)}} (q(x) - p(x))dx && \text{(By assmp. \& } f, p, q \geq 0) \end{aligned}$$

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

$$
\begin{aligned}
&= \mathbb{E}_p\left[w^{\epsilon}_{p,q}(x) f(x)\right] \\
&+ M \int_x (q(x) - p(x)) \underbrace{\mathbb{1}[p(x) \le \epsilon \wedge p(x) \le q(x)]}_{\delta^{\epsilon}_{p,q}(x)} dx && \text{(By domain of int.)} \\
&= \mathbb{E}_p\left[w^{\epsilon}_{p,q}(x) f(x)\right] + M \cdot \left(\mathop{\mathbb{E}}_{q}[\delta_{p,q}(x)] - \mathop{\mathbb{E}}_{p}[\delta^{\epsilon}_{p,q}(x)]\right) && \text{(By def.)}
\end{aligned}
$$

Further, $p = q$ implies equality when $\epsilon \ge \sup_x q(x)$.

□

A.3 Proof of Theorem 2

Lemma 3. *Assume that $p_t(Y \mid X) = p_s(Y \mid X)$. Define $Z = \phi(X)$ and let $h(x) = f(\phi(x))$. Then,*

$$
\mathop{\mathbb{E}}_{x,y\sim q(x,y)}[\ell(h(x), y)] = \mathop{\mathbb{E}}_{z,y\sim q(z,y)}[\ell(f(z), y)]
$$

Proof.

$$
\begin{aligned}
\mathop{\mathbb{E}}_{z,y\sim q(z,y)}[\ell(f(z), y)] &= \int_{z,y} q(z,y)\ell(f(z), y) dz dy \\
&= \int_{z,y} \ell(f(z), y) \int_{x\in\phi^{-1}(z)} q(x,y) dx dz dy \\
&= \int_{x,y} q(x,y) \int_z \mathbb{1}[z = \phi(x)]\ell(f(z), y) dz dx dy \\
&= \int_{x,y} q(x,y)\ell(h(x), y) dx dy \\
&= \mathop{\mathbb{E}}_{x,y\sim q(x,y)}[\ell(h(x), y)]
\end{aligned}
$$

□

Lemma 4.

$$
R_t(h) = E_{q(z)p(y|z)}[\ell(f(z), y)] + \eta^{\ell}_{\phi}(f, y)
$$

Proof. By Lemma A.3

$$
R_t(h) = E_{q(x,y)}[\ell(h(x), y)] = E_{q(z,y)}[\ell(f(z), y)]
$$

We have that

$$
\begin{aligned}
E_{q(z)q(y|z)}[\ell(f(z), y)] &= \int_z \int_y q(z) q(y \mid z)\ell(f(z), y) dy dz \\
&= \int_z \int_y \left(\int_{x:\phi(x)=z} q(x) dx\right) q(y \mid z)\ell(f(z), y) dy dz \\
&= \int_z \int_y \int_{x\in\phi^{-1}(z)} q(x) q(y \mid \phi(x))\ell(h(\phi(x)), y) dx dy dz \\
&= \int_x \int_y q(x) q(y \mid \phi(x))\ell(h(\phi(x)), y) dx dy \\
&= \int_x \int_y q(x)\ell(h(\phi(x)), y) \left(\underbrace{q(y \mid x)}_{=p(y|x) \text{ (assmp.)}} + q(y \mid \phi(x)) - q(y \mid x)\right) dx dy
\end{aligned}
$$

and by the same argument,

$$E_{q(z)p(y|z)}[\ell(f(z),y)] = \int_x \int_y q(x)\ell(h(\phi(x)),y)\left(p(y \mid x) + p(y \mid \phi(x)) - p(y \mid x)\right) dxdy$$

and as a result,

$$\begin{aligned}
&E_{q(z)q(y|z)}[\ell(f(z),y)] - E_{q(z)p(y|z)}[\ell(f(z),y)] \\
&\quad = \mathop{\mathbb{E}}_{q}(x)[\mathop{\mathbb{E}}_{q(y|\phi(x))}\ell(f(\phi(x)),y)] - \mathop{\mathbb{E}}_{q(y|x)}\ell(f(\phi(x)),y)] \\
&\quad + \mathop{\mathbb{E}}_{q}(x)[\mathop{\mathbb{E}}_{p(y|\phi(x))}\ell(f(\phi(x)),y)] - \mathop{\mathbb{E}}_{p(y|x)}\ell(f(\phi(x)),y)] \\
&\quad = \eta^{\ell}_{\phi}(f,y)
\end{aligned}$$

□

Theorem 2 (Restated). Consider any feature representation $z = \phi(x)$ with $\phi : \mathcal{X} \to \mathcal{Z}$ and prediction function $f : \mathcal{Z} \to \mathcal{Y}$, and define $h = f \circ \phi$. Further, let $p(Z)$ and $q(Z)$ be the two distributions induced by the representation ϕ applied to X distributed according to $p(X), q(X)$. Further, assume that for any hypothesis h and a loss function ℓ, $\sup_{x\in\mathcal{X},h\in\mathcal{H}} \ell(h(x),y) \leq M$. Now, with $\epsilon > 0$, we have the following result.

$$R_q(h) \leq \mathbb{E}_p\left[w^{\epsilon}_{p_s,p_t}(z)\ell(f(z),y)\right] + M d^{\epsilon}_{\text{supp}}(p(z) \parallel q(z)) + \eta^{\ell}_{\phi}(f,y)\ .$$

Proof. By Lemma 4, we have that

$$R_q(h) \leq E_{q(z)p(y|z)}[\ell(f(z),y)] + \eta^{\ell}_{\phi}(f,y)\ .$$

Further,

$$\begin{aligned}
E_{q(z)p(y|z)}[\ell(f(z),y)] &= \iint_{z\in\mathcal{Z},y\in\mathcal{Y}} q(z)p(y \mid z)\ell(f(z),y)dydz \\
&= \iint_{z:p(z)\geq\epsilon,y\in\mathcal{Y}} q(z)p(y \mid z)\ell(f(z),y)dydz \\
&+ \iint_{z:p(z)<\epsilon,y\in\mathcal{Y}} q(z)p(y \mid z)\ell(f(z),y)dydz \\
&= \iint_{z\in\mathcal{Z},y\in\mathcal{Y}} w^{\epsilon}_{p_s,p_t}(z)p(z)p(y \mid z)\ell(f(z),y)dydz \\
&+ \int_{z:\substack{p(z)<\epsilon \\ p(z)\leq q(z)}} \underbrace{(q(z)-p(z))}_{\geq 0} \underbrace{\int_{y\in\mathcal{Y}} p(y \mid z)\ell(f(z),y)dy}_{\in[0,M]}\, dz \\
&+ \underbrace{\int_{z:\substack{p(z)<\epsilon \\ p(z)>q(z)}} (q(z)-p(z)) \int_{y\in\mathcal{Y}} p(y \mid z)\ell(f(z),y)dydz}_{\leq 0}\ .
\end{aligned}$$

□

A.4 Kernel support divergence

Theorem 3 may be viewed as a measuring differences in density only where supports differ significantly. In the case where $\mathcal{L}$ is a Hilbert space, similar to the maximum mean discrepancy (Gretton *et al.*, 2012), we may decompose $d^{\mathcal{L},\epsilon}_{\text{supp}}(p \parallel q)$ using reproducing kernels.

Lemma 5. *Let $\mathcal{H}$ be the reproducing-kernel Hilbert space with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then,*

$$\begin{aligned}
d^{\epsilon}_{\text{supp}}(p \parallel q)_{\mathcal{G}} &= \mathop{\mathbb{E}}_{x,x'\sim p}[\delta^{\epsilon}_p(x,x')k(x,x') \\
&- 2\mathop{\mathbb{E}}_{x\sim p,x'\sim q}[\delta^{\epsilon}_p(x,x')k(x,x')] + \mathop{\mathbb{E}}_{x,x'\sim q}[\delta^{\epsilon}_p(x,x')k(x,x')]
\end{aligned} \tag{2}$$

Proof. The proof follows from Gretton *et al.* (2012) and can be found in Appendix A.4. □

Let $\delta_p^\epsilon(x) = \mathbb{1}[p(x) \leq \epsilon]$ and $\delta_p^\epsilon(x, x') = \delta_p^\epsilon(x)\delta_p^\epsilon(x')$

$$\begin{aligned} d^\epsilon_{\text{supp}}(p \parallel q)_\mathcal{G} &\coloneqq \sup_{g \in \mathcal{G}} \left| \mathop{\mathbb{E}}_{q}[\delta_p^\epsilon(x)g(x)] - \mathop{\mathbb{E}}_{p}[\delta_p^\epsilon(x)g(x)] \right| \\ &= \mathop{\mathbb{E}}_{x,x' \sim p}[\delta_p^\epsilon(x, x')k(x, x')] \\ &\quad - 2 \mathop{\mathbb{E}}_{x \sim p, x' \sim q}[\delta_p^\epsilon(x, x')k(x, x')] \\ &\quad + \mathop{\mathbb{E}}_{x,x' \sim q}[\delta_p^\epsilon(x, x')k(x, x')] \end{aligned} \tag{3}$$

Proof. Follow `http://alex.smola.org/teaching/iconip2006/iconip_3.pdf` page 18–20. □

B Model

We may bound $d^\epsilon_{\text{supp}}(p \parallel q)$ using the hinge loss as follows,

$$\begin{aligned} d^\epsilon_{\text{supp}}(p \parallel q) &\leq \mathop{\mathbb{E}}_{x \sim q}\left[\max\left(0, 2 - \frac{p(x)}{\epsilon}\right) \max\left(0, 2 - \frac{p(x)}{q(x)}\right)\right] \\ &\quad - \mathop{\mathbb{E}}_{x \sim p}\left[\max\left(0, 1 - \frac{p(x)}{\epsilon}\right) \max\left(0, 1 - \frac{p(x)}{q(x)}\right)\right] \\ &=: \tilde{d}^\epsilon_{\text{supp}}(p \parallel q) \, . \end{aligned}$$

C Experiments

In Figure 1, we see that the embeddings learned using DANN models under label marginal shift show worse separation between classes, than the embeddings learned under equal label marginal distributions.

D Consistent domain-invariant variable selection

Consider a matrix $A \in \{0, 1\}^{k \times d}$ with $k < d$ such that $\forall j : \sum_{i=1} a_{ij} \leq 1$ and $\forall i : \sum_{j=1} a_{ij} \leq 1$. In other words, A is a variable selection operator on X. Now, assume that $Z \coloneqq \Phi(X) \coloneqq AX$ is sufficient for Y on p_s and p_t and that $p_s(AX) = p_t(AX)$. Further, assume labeled data is observed under p and unlabeled data observed under p_t. Then, is Y identifiable based on domain-invariance and source predictive loss?

Condition 1 (Smoothness)**.** *With $\Sigma_L = \{f : \sum_{k \in \mathbb{Z}^d} k_2^2 \langle f, \varphi_k \rangle^2 \leq L; \forall j \in \{1, \ldots, d\}$ for $L > 0$, f is L-smooth if $f \in \Sigma_L$, with φ_k the trigonometric fourier basis.*

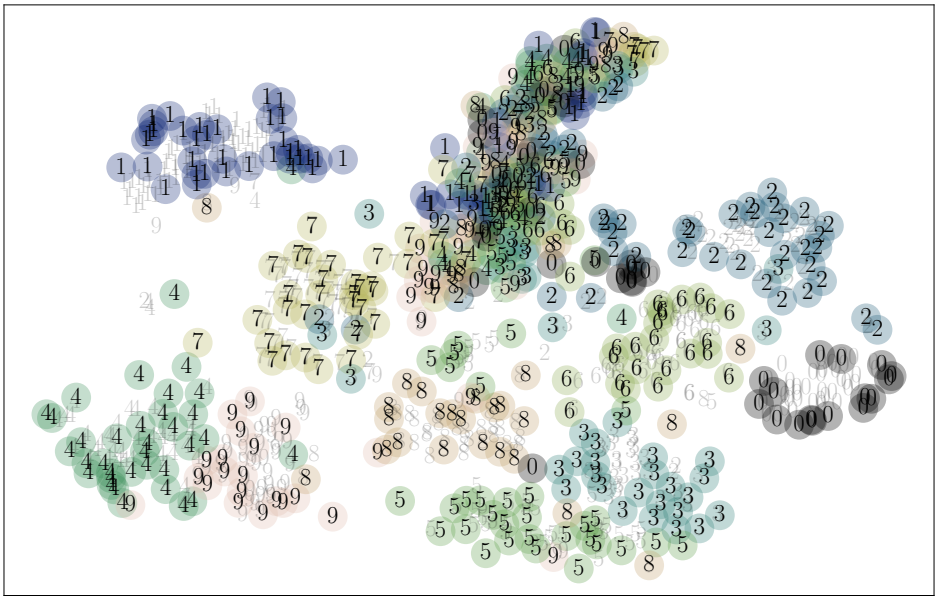
Condition 2 (Identifiability)**.** *A sufficient set of variables J, such that $\exists \bar{f} : f(x) = \bar{f}(x_J)$ for all $x \in \mathbb{R}^d$, is κ-identifiable if for all $j \in J$,*

$$\int_{[0,1]^d} (f(x) - \int_0^1 f(x) dx_j)^2 dx \geq \kappa \, .$$

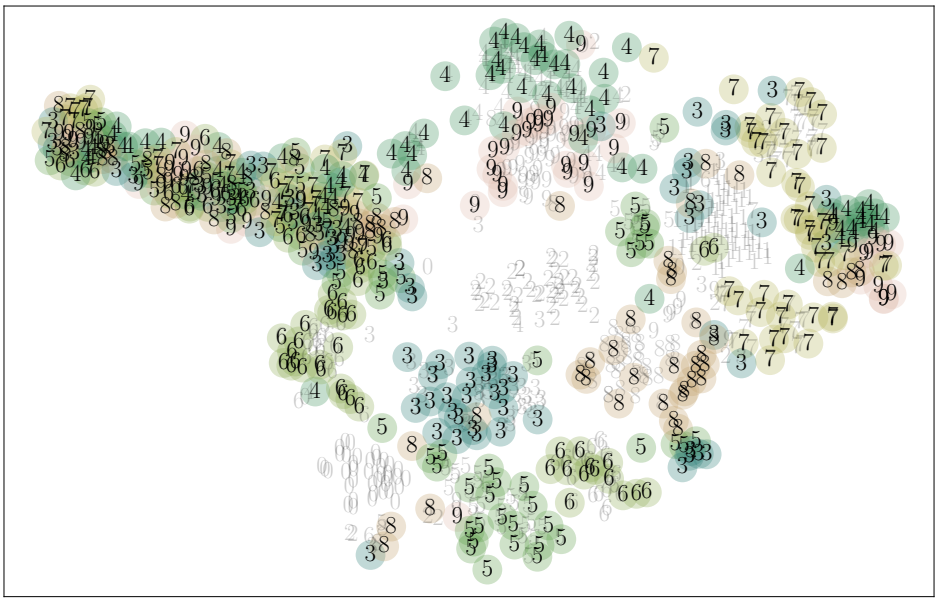
Condition 3 (Positive bounded support)**.** *The density $p_s(x)$ has positive bounded support over $[0, 1]^d$ if with $p_{s\min} > 0$, $\forall x \in [0, 1]^d : p_s(x) \geq p_{s\min}$ and $\forall x \notin [0, 1]^d : p_s(x) = 0$.*

Condition 4 (Bounded ∞-norm and 2-norm)**.** *A function f has bounded ∞-norm and 2-norm with respect to p if $\Pr_{X \sim p_s}(|f(X)| \leq L_\infty) = 1$ and $\mathbb{E}_{X \sim p_s}[f(X)^2] \leq L_2^2$.*

Condition 5 (Sub-gaussian additive noise)**.** *The observed outcome may be written as $Y_i = f(x_i) + \sigma\epsilon_i$ with $\mathbb{E}[e^{t\epsilon_i} \mid X_i] \leq e^{t^2/2}$ for all $t > 0$.*



(a) MNIST→MNIST-M



(b) MNIST→MNIST-M$\backslash\{0, 1, 2\}$

Figure 1: Embeddings learned by DANN with equal (top) and unequal (bottom) label marginal distributions. In MNIST-M$\backslash\{0, 1, 2\}$, all images of digits 0,1,2 have been removed. Grey digits are embeddings and labels of the source domain. Black digits against colored background are from the target domian.

Theorem 1 (Variable selection in non-parametric regression (Comminges *et al.*, 2012))**.** *Assume that Conditions 1–5 hold, with known parameters* $p_{s\min}$, $\theta = 2L/\kappa$ *and* L_2. *Then, there is an estimator* $\hat{J}$ *that satisfies* $\Pr(\hat{J} \neq J) \leq (8d/d^*)^{-d^*}$.

Comminges *et al.* (2012) give a constructive proof of Theorem 1 in which the chosen estimator is allowed to depend on the density $p_s(x)$.

References

Comminges, L., Dalalyan, A.S. *et al.* (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistics*, **40**, 2667–2696.

Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B. & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, **13**, 723–773.