Supplementary material for paper: Margin-aware Adversarial Domain Adaptation with Optimal Transport

In this supplementary material, we provide proofs for the theoretical claims of the main paper, and we detail some other points.

Problem setup and notations

We recall the problem setup of our study with the notations used throughout the paper. We consider a binary classification setting, in which source and target data are respectively drawn from $\mathcal{S}$ and $\mathcal{T}$, the joint distributions over the product space of instances and labels $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{-1, 1\}$. We denote their corresponding marginal distributions as $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{T}$ and use bold upper-case letters for matrices (e.g., $\mathbf{D}$) and bold lower-case letters for vectors (e.g., $\mathbf{x}$). Although both domains are assumed to be labeled, only the labels of the source instances are observable during the learning stage. This settings is often referred to as unsupervised domain adaptation.

Furthermore, let $\mathcal{H}$ and $\mathcal{H}'$ denote two compact classes of hypotheses acting on $\mathcal{X}$ and taking values in $[-1, 1]$. For further developments, we define several quantities to assess classifiers' performances on different domains. Let $l^{\rho,\beta}$ be the loss function defined by

$$l^{\rho,\beta}(t) := \begin{cases} 1 - \frac{(t-\rho)}{\beta}, & \text{if } \rho \leq t \leq \beta + \rho \\ [t < \rho], & \text{otherwise} \end{cases}$$

where $1 > \rho, \beta > 0$, and $[\cdot]$ denotes the Iverson bracket for indicator functions. From its definition, we note that $l^{\rho,0}(t) = [t < \rho]$ and that it verifies the following inequality for all $\rho, \beta > 0$ and $t \in \mathbb{R}$

$$l^{\rho,0}(t) = [t < \rho] < l^{\rho,\beta}(t) < l^{\rho+\beta,0}(t) = [t < \rho + \beta] \quad (1)$$

illustrated in Figure 1.

For any domain $\mathcal{P}$ with marginal feature distribution $\mathcal{D}_\mathcal{P}$ and any hypotheses h, f, we define their disagreement associated to the loss $l^{\rho,\beta}$ as

$$\epsilon_\mathcal{P}^{\rho,\beta}(h, f) := \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}_\mathcal{P}} \left[l^{\rho,\beta}(f(\mathbf{x})h(\mathbf{x})) \right], \quad (2)$$

This quantity can be further generalized to non deterministic hypotheses that define the labeling of domain $\mathcal{P}$, in which case the expectation is taken over $\mathcal{P}$:

$$\epsilon_\mathcal{P}^{\rho,\beta}(h) := \mathop{\mathbb{P}}_{\mathbf{x}, y \sim \mathcal{P}} \left[l^{\rho,\beta}(yh(\mathbf{x})) \right]. \quad (3)$$

This definition stands for the classification risk on $\mathcal{P}$: for $\beta = 0$, it is the $\rho-$margin violation rate measuring the probability of the event $\{yh(\mathbf{x}) < \rho\}$, while for $\rho = \beta = 0$, it is the 0-1 or misclassification rate.

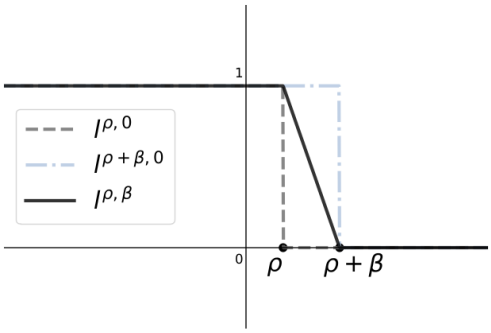

Figure 1. Loss function $l^{\rho,\beta}$ with its characteristic points and an illustration of the property from Equation (1).

Section 2

Details on the bound of (Zhang et al., 2019) in the binary case: In their work, (Zhang et al., 2019) use vector-valued scoring functions that take their values in $\mathbb{R}^2$ in the case of binary classification. To this end, let us denote by $\underline{h} = (h_1, h_2)$ and $\underline{h}' = (h'_1, h'_2)$ the respective $\mathbb{R}^2$-valued counterparts of our real valued h and h' scoring functions, and we define the labels in $\{0,1\}$ as $\underline{y}$ analogously. Then, according to the notations used by the authors, we have:

$$\rho_{\underline{h}}(\mathbf{x}, 1) = \frac{1}{2}(h_1(\mathbf{x}) - h_2(\mathbf{x})),$$
$$\rho_{\underline{h}}(\mathbf{x}, 0) = \frac{1}{2}(h_2(\mathbf{x}) - h_1(\mathbf{x})).$$

By linking their notations and ours via the relation $h = \frac{1}{2}(h_1 - h_2)$, the last two equations can be written:

$$\rho_{\underline{h}}(\mathbf{x}, \underline{y}) = yh(\mathbf{x}).$$

In particular, when we consider the class associated to a scoring function, i.e., $\underline{y}_{\underline{h}} := [h_1 > h_2]$ corresponding in our

case to $y_h = \operatorname{sgn}(h)$, we have:

$$\rho_{\underline{h}'}(\mathbf{x}, \underline{y}_{\underline{h}}) = \operatorname{sgn}(h(\mathbf{x}))\, h'(\mathbf{x}).$$

Hence, we have for all $\beta > 0$,

$$l^{0,\beta} \circ \rho_{\underline{h}'}(\mathbf{x}, \underline{y}_{\underline{h}}) = l^{0,\beta}(\operatorname{sgn}(h(\mathbf{x}))\, h'(\mathbf{x})).$$

where $l^{0,\beta}$ is denoted Φ_β in (Zhang et al., 2019).

Remark For any $\kappa > 0$, the parametrization

$$h = \kappa(h_1 - h_2)$$

also works as long as $|h| \leq 1$. In this case, we have:

$$\begin{aligned} l^{0,\beta} \circ \rho_{\underline{h}'}(\mathbf{x}, \underline{y}_{\underline{h}}) &= l^{0,\beta}\left(\frac{1}{\kappa}\operatorname{sgn}(h(\mathbf{x}))\, h'(\mathbf{x})\right) \\ &= l^{0,\kappa\beta}(\operatorname{sgn}(h(\mathbf{x}))\, h'(\mathbf{x})) \end{aligned}$$

and $\kappa\beta$ spans all of the positive numbers when $\beta > 0$.

Section 3

Bound with non convex divergence between distributions

Theorem 1. *Assume that for any $h' \in \mathcal{H}'$, we have $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathcal{S}}[h'(\mathbf{x}) = 0] = \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathcal{T}}[h'(\mathbf{x}) = 0] = 0$. Let $\rho, \beta, \alpha > 0$ be such that $\rho + \beta < \alpha < 1$. Then, for any $h \in \mathcal{H}$, the following bound holds:*

$$\epsilon_\mathcal{T}^{\rho,0}(h) \leq \epsilon_\mathcal{S}^{\overset{+}{\mathcal{H}},0}(h) + d_{h,\mathcal{H}'}^{\rho,\beta}(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T}) + \lambda_\alpha$$

where

$$d_{h,\mathcal{H}'}^{\rho,\beta}(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T}) := \sup_{h'\in\mathcal{H}'} \left|\epsilon_\mathcal{S}^{\rho,\beta}(h,h') - \epsilon_\mathcal{T}^{\rho,\beta}(h,h')\right|$$

and

$$\lambda_\alpha := \inf_{f\in\mathcal{H}'} \epsilon_\mathcal{T}^{0,0}(f) + \epsilon_\mathcal{S}^{0,0}(f) + \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathcal{S}}[|f| < \alpha].$$

Proof. Let $h \in \mathcal{H}$, $f \in \mathcal{H}'$ and $0 \leq \theta < 1$.

$$\begin{aligned} &\mathbb{P}_{\mathbf{x},y\sim\mathcal{T}}[yh < \theta\rho] && (4)\\ &= \mathbb{P}_{\mathbf{x},y\sim\mathcal{T}}\left[yh.f^2 < \theta\rho.f^2\right] && (5)\\ &\leq \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathcal{T}}[hf < \rho|f|] + \mathbb{P}_{\mathbf{x},y\sim\mathcal{T}}[yf < \theta|f|] && (6)\\ &\leq \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathcal{T}}[hf < \rho] + \mathbb{P}_{\mathbf{x},y\sim\mathcal{T}}[y\operatorname{sgn}(f) < \theta] \\ &= \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathcal{T}}[hf < \rho] + \mathbb{P}_{\mathbf{x},y\sim\mathcal{T}}[y\operatorname{sgn}(f) < 0] && (7)\\ &= \epsilon_\mathcal{T}^{\rho,0}(h,f) + \epsilon_\mathcal{T}^{0,0}(f). && (8) \end{aligned}$$

In the previous developments, we first use the fact that $yh(\mathbf{x})ff^2(\mathbf{x}) < \theta\rho f^2(\mathbf{x})$ implies that either $h(\mathbf{x})f(\mathbf{x}) < \rho|f|(\mathbf{x})$ or $yf(\mathbf{x}) < \theta|f|(\mathbf{x})$ to obtain (6). The next inequality comes from $|f| \leq 1$ as $f \in \mathcal{H}'$. Then, we use the fact that $\frac{f}{|f|} = \operatorname{sgn}(f)$. Finally, as $0 < \theta < 1$ and $y\operatorname{sgn}(f) \in \{-1, 1\}$, the inequality $y\operatorname{sgn}(f) < \theta$ is equivalent to $y\operatorname{sgn}(f) = -1 < 0$ and to $yf(\mathbf{x}) < 0$, implying (7) and (8).
Taking the limit as $\theta \to 1$ in the previous inequality, we get:

$$\epsilon_\mathcal{T}^{\rho,0}(h) \leq \epsilon_\mathcal{T}^{\rho,0}(h,f) + \epsilon_\mathcal{T}^{0,0}(f). \tag{9}$$

Now, let us concentrate on bounding $\epsilon_\mathcal{T}^{\rho,0}(h,f)$.

$$\begin{aligned} &\epsilon_\mathcal{T}^{\rho,0}(h,f) \\ &= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathcal{T}}[[hf < \rho]] \\ &\leq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathcal{T}}\left[l^{\rho,\beta}(hf)\right] \\ &= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathcal{T}}\left[l^{\rho,\beta}(hf)\right] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathcal{S}}\left[l^{\rho,\beta}(hf)\right] + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathcal{S}}\left[l^{\rho,\beta}(hf)\right] \\ &\leq \sup_{h'\in\mathcal{H}'}\left|\epsilon_{\mathcal{D}_\mathcal{T}}^{\rho,\beta}(h,h') - \epsilon_{\mathcal{D}_\mathcal{S}}^{\rho,\beta}(h,h')\right| + \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathcal{S}}[hf < \rho + \beta], \end{aligned} \tag{10}$$

where we used the lower bound on the ramp function $l^{\rho,\beta}$ from Equation (1). Finally, from the fact that $f \in \mathcal{H}'$, we take the supremum over $\mathcal{H}'$ and we use the upper bound on $l^{\rho,\beta}$, again from Equation (1), to obtain (10).
What is left to bound is $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathcal{S}}[hf < \rho + \beta]$. For any $0 < \theta < 1$, we have:

$$\begin{aligned} &\mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathcal{S}}[hf < \rho + \beta] \\ &= \mathbb{P}_{\mathbf{x},y\sim\mathcal{S}}\left[yh.yf < \frac{\rho+\beta}{\theta|f|}\theta|f|\right] \\ &\leq \mathbb{P}_{\mathbf{x},y\sim\mathcal{S}}\left[yh < \frac{\rho+\beta}{\theta|f|}\right] + \mathbb{P}_{\mathbf{x},y\sim\mathcal{S}}[yf < \theta|f|] \\ &= \mathbb{P}_{\mathbf{x},y\sim\mathcal{S}}\left[yh < \frac{\rho+\beta}{\theta|f|}\right] + \mathbb{P}_{\mathbf{x},y\sim\mathcal{S}}[y\operatorname{sgn}(f) < \theta] \\ &= \mathbb{P}_{\mathbf{x},y\sim\mathcal{S}}\left[yh < \frac{\rho+\beta}{\theta|f|}\right] + \mathbb{P}_{\mathbf{x},y\sim\mathcal{S}}[y\operatorname{sgn}(f) < 0] && (11)\\ &\xrightarrow{\theta\to 1} \mathbb{P}_{\mathbf{x},y\sim\mathcal{S}}\left[yh < \frac{\rho+\beta}{|f|}\right] + \epsilon_\mathcal{S}^{0,0}(f) \\ &\leq \mathbb{P}_{\mathbf{x},y\sim\mathcal{S}}\left[yh < \frac{\rho+\beta}{\alpha}\right] + \mathbb{P}_{\mathbf{x}\sim\mathcal{D}_\mathcal{S}}[|f| < \alpha] + \epsilon_\mathcal{S}^{0,0}(f) && (12) \end{aligned}$$

where $\alpha > 0$ is arbitrarily chosen. To obtain (11), we applied the same technique as the one used to prove (9). Then, taking the limit as $\theta \to 1$ is justified by the monotonous convergence theorem. Finally, we use conditioning on the event $\{|f(\mathbf{x})| \geq \alpha\}$, and the fact that the event $\{yh(\mathbf{x}) < \frac{\rho+\beta}{|f|(\mathbf{x})} \wedge |f|(\mathbf{x}) \geq \alpha\}$ implies event $\{yh(\mathbf{x}) < \frac{\rho+\beta}{\alpha}\}$, to obtain (12).

To sum up the different developments established up to now, i.e., (9), (10) and (12), we have:

$$\epsilon_{\mathcal{T}}^{\rho,0}(h) \leq \epsilon_{\mathcal{T}}^{0,0}(f) + \epsilon_{\mathcal{S}}^{0,0}(f) + \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{S}}}[|f| < \alpha] + d_{h,\mathcal{H}}^{\rho,\beta}(\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}}) + \epsilon_{\mathcal{S}}^{\overset{+}{-},0}(h).$$

Given that previous inequality holds for any choice of $f \in \mathcal{H}'$, minimizing it over this choice yields the result and introduces the ideal joint error. □

Corollary 1. *If $\mathcal{H} = \mathcal{H}'$ is a class of binary hypotheses taking values in $\{-1, 1\}$, the bound from Theorem 1 is equivalent to the one in (Ben-David et al., 2010).*

Proof. Let $h \in \mathcal{H}$, $h' \in \mathcal{H}'$ and $y \in \{-1, 1\}$. For $0 < \rho, \beta < 1$ and $1 > \alpha > \rho + \beta$, we have:

$$l^{\rho,\beta}(h(\mathbf{x})h'(\mathbf{x})) = [h(\mathbf{x})h(\mathbf{x}') < 0] \tag{13}$$

$$[y.h(\mathbf{x}) < \frac{\rho+\beta}{\alpha}] = l^{\overset{+}{-},0}(y.h(\mathbf{x})) = [y.h(\mathbf{x}) < 0]. \tag{14}$$

This holds because for any $1 > \rho, \beta > 0$ verifying $\rho + \beta < 1$, $l^{\rho,\beta}$ and $[. < 0]$ take the same values when restricted to the set $\{-1, 1\}$. Consequently, all the margins in the estimable part of our bound can be omitted, and it becomes equal to $\epsilon_{\mathcal{S}}^{0,0}(h) + d_{h,\mathcal{H}'}^{0,0}(\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}})$.
Omitting all of the margins in the estimable part of our bound, we have:

$$\epsilon_{\mathcal{T}}^{0,0}(h) \leq \epsilon_{\mathcal{S}}^{0,0}(h) + d_{h,\mathcal{H}}^{0,0}(\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}}) + \lambda_\alpha \tag{15}$$

$$= \epsilon_{\mathcal{S}}^{0,0}(h) + \sup_{h'\in\mathcal{H}} \left|\epsilon_{\mathcal{S}}^{0,0}(h,h') - \epsilon_{\mathcal{T}}^{0,0}(h,h')\right| + \lambda_\alpha \tag{16}$$

$$\leq \epsilon_{\mathcal{S}}^{0,0}(h) + \sup_{h,h'\in\mathcal{H}} \left|\epsilon_{\mathcal{D}_{\mathcal{S}}}^{0,0}(h,h') - \epsilon_{\mathcal{D}_{\mathcal{T}}}^{0,0}(h,h')\right| + \lambda_\alpha$$

$$= \epsilon_{\mathcal{S}}^{0,0}(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}}) + \lambda_\alpha \tag{17}$$

where (15) is from Theorem 1 and (16) comes from properties (13) and (14). Then, taking the supremum over $h \in \mathcal{H}$ and the definition of $d_{\mathcal{H}\Delta\mathcal{H}}$ yield the rest of the developments. Finally, for any $f \in \mathcal{H}$, $\mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{S}}}[|f(\mathbf{x}) < \alpha|] = 0$ since $|f(\mathbf{x})| = 1$ for all $\mathbf{x} \in \mathcal{X}$ as f is a binary hypothesis. Hence,

$$\lambda_\alpha = \inf_{f\in\mathcal{H}} \epsilon_{\mathcal{T}}^{0,0}(f) + \epsilon_{\mathcal{S}}^{0,0}(f) + \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{S}}}[|f| < \alpha] = \inf_{f\in\mathcal{H}} \epsilon_{\mathcal{T}}^{0,0}(f) + \epsilon_{\mathcal{S}}^{0,0}(f) = \lambda$$

Combining this result with Equation (17) yields the final result. □

PROOF OF THE BOUND ON λ_α

When we analyzed the previous bound in the main paper, we gave the following bound on λ_α.

$$\lambda_\alpha \leq \min_{f\in\mathcal{H}'} \epsilon_{\mathcal{S}}^{\alpha,0}(f) + \epsilon_{\mathcal{T}}^{\alpha,0}(f).$$

We hereby provide its proof.

Proof. We have for $f \in \mathcal{H}'$ and for $\alpha > 0$:

$$\epsilon_{\mathcal{T}}^{0,0}(f) = \mathop{\mathbb{P}}_{\mathbf{x},y\sim\mathcal{T}}[y.f(\mathbf{x}) < 0] \leq \mathop{\mathbb{P}}_{\mathbf{x},y\sim\mathcal{T}}[y.f(\mathbf{x}) < \alpha] \tag{18}$$

Also

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\mathbf{x},y\sim\mathcal{S}}[y.f(\mathbf{x}) < \alpha] \\
&= \mathop{\mathbb{P}}_{\mathbf{x},y\sim\mathcal{S}}[y.f(\mathbf{x}) < 0] + \mathop{\mathbb{P}}_{\mathbf{x},y\sim\mathcal{S}}[0 \leq y.f(\mathbf{x}) < \alpha] \\
&= \mathop{\mathbb{P}}_{\mathbf{x},y\sim\mathcal{S}}[y.f(\mathbf{x}) < 0] \\
&+ \mathop{\mathbb{P}}_{\mathbf{x},y\sim\mathcal{S}}[0 \leq y.f(\mathbf{x}) < \alpha \wedge y = \operatorname{sgn}(f(\mathbf{x}))] \\
&+ \mathop{\mathbb{P}}_{\mathbf{x},y\sim\mathcal{S}}[0 \leq y.f(\mathbf{x}) < \alpha \wedge y = -\operatorname{sgn}(f(\mathbf{x}))] \\
&\geq \mathop{\mathbb{P}}_{\mathbf{x},y\sim\mathcal{S}}[y.f(\mathbf{x}) < 0] \\
&+ \mathop{\mathbb{P}}_{\mathbf{x},y\sim\mathcal{S}}[0 \leq y.f(\mathbf{x}) < \alpha \wedge y = \operatorname{sgn}(f(\mathbf{x}))] \\
&= \mathop{\mathbb{P}}_{\mathbf{x},y\sim\mathcal{S}}[y.f(\mathbf{x}) < 0] + \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{S}}}[|f(\mathbf{x})| < \alpha]
\end{aligned}$$

Summing the last inequality and inequality (18) yields for any $f \in \mathcal{H}'$:

$$\epsilon_{\mathcal{T}}^{0,0}(f) + \epsilon_{\mathcal{S}}^{0,0}(f) + \mathop{\mathbb{P}}_{\mathbf{x}\sim\mathcal{D}_{\mathcal{S}}}[|f| < \alpha] \leq \epsilon_{\mathcal{T}}^{\alpha,0}(f) + \epsilon_{\mathcal{S}}^{\alpha,0}(f)$$

We take the infimum over $f \in \mathcal{H}'$ on both sides of this inequality to complete the proof. □

A convex domain divergence based on optimal transport

Before proceeding to the rest of the proofs, we recall some notions related to the optimal transport theory. Let us introduce two projection operators $\pi_1 : (\mathbf{x}_1, \mathbf{x}_2) \mapsto \mathbf{x}_1$ and $\pi_2 : (\mathbf{x}_1, \mathbf{x}_2) \mapsto \mathbf{x}_2$ defined over $\mathcal{X} \times \mathcal{X}$. The set Π of transport plans between $\mathcal{D}_{\mathcal{S}}$ and $\mathcal{D}_{\mathcal{T}}$ is the set of probability distributions $\mathcal{D}$ over $\mathcal{X} \times \mathcal{X}$ that verify the following two properties:

$$\pi_1 \# \mathcal{D} = \mathcal{D}_{\mathcal{S}}, \qquad \pi_2 \# \mathcal{D} = \mathcal{D}_{\mathcal{T}},$$

where $\#$ denotes the pushforward measure.

Proposition 1 (Convex bound for alignment term)**.** *For any $\rho, \beta > 0$, we have*

$$d_{h,\mathcal{H}'}^{\rho,\beta}(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T}) \leq \frac{1}{\beta} \inf_{\mathcal{D}\in\Pi} \Delta_{\mathcal{H}'}(h, \mathcal{D}) \tag{19}$$

where

$$\Delta_{\mathcal{H}'}(h, \mathcal{D}) := \sup_{h'\in\mathcal{H}'} \mathop{\mathbb{E}}_{\mathbf{x}_s,\mathbf{x}_t\sim\mathcal{D}} \left[|hh'(\mathbf{x}_s) - hh'(\mathbf{x}_t)|\right].$$

Proof.

$$\begin{aligned}
&d_{h,\mathcal{H}}^{\rho,\beta}(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T}) \\
&= \sup_{h'\in\mathcal{H}'} \left| \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_\mathcal{T}} \left[l^{\rho,\beta}(hh')\right] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_\mathcal{S}} \left[l^{\rho,\beta}(hh')\right] \right| \\
&\leq \sup_{h'\in\mathcal{H}'} \sup_{|\varphi|_{\text{Lip}}\leq \frac{1}{\beta}} \left| \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_\mathcal{T}} [\varphi(hh')] - \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}_\mathcal{S}} [\varphi(hh')] \right| \qquad (20) \\
&= \frac{1}{\beta} \sup_{h'\in\mathcal{H}'} W_1(hh'\#\mathcal{D}_\mathcal{S}, hh'\#\mathcal{D}_\mathcal{T}) \qquad (21) \\
&= \frac{1}{\beta} \sup_{h'\in\mathcal{H}'} \inf_{\mathcal{D}\in\Pi} \mathop{\mathbb{E}}_{\mathbf{x}_s,\mathbf{x}_t\sim\mathcal{D}} [|hh'(\mathbf{x}_s) - hh'(\mathbf{x}_t)|] \qquad (22) \\
&\leq \frac{1}{\beta} \inf_{\mathcal{D}\in\Pi} \sup_{h'\in\mathcal{H}'} \mathop{\mathbb{E}}_{\mathbf{x}_s,\mathbf{x}_t\sim\mathcal{D}} [|hh'(\mathbf{x}_s) - hh'(\mathbf{x}_t)|] \qquad (23)
\end{aligned}$$

where we used the $\frac{1}{\beta}-$Lipchtiz continuity of $l^{\rho,\beta}$ to obtain (20), in which the term $\sup_{|\varphi|_{\text{Lip}}\leq \frac{1}{\beta}}$ denotes a supremum over all $\frac{1}{\beta}-$Lipschitz functions. Then using the dual form of the Wasserstein distance W_1 between 1-dimensional distributions $hh'\#\mathcal{D}_\mathcal{S}$ and $hh'\#\mathcal{D}_\mathcal{T}$, we obtain (21). In the next line (22), we express the Wasserstein distance in its primal form. Finally, we use the inf-sup inequality to obtain (23). □

Proposition 2 (Optimal transport bound on the target risk)**.** *With the assumptions and notations of Theorem 1 and Proposition 1, we have for any $h \in \mathcal{H}$:*

$$\epsilon_\mathcal{T}^{\rho,0}(h) \leq \epsilon_{\overline{\mathcal{S}}}^{\rho,0}(h) + \frac{1}{\beta} \inf_{\mathcal{D}\in\Pi} \Delta_{\mathcal{H}'}(h, \mathcal{D}) + \lambda_\alpha \tag{24}$$

Proof. Follows directly by combining results of Theorem 1 and Proposition 1. □

Proposition 3 (Bounding by the Wasserstein distance)**.** *Let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a metric, and assume that all of the hypotheses from $\mathcal{H}$ and $\mathcal{H}'$ verify the $L-$Lipschitz continuity with respect to metric c for some $L > 0$. Then, the following holds*

$$\sup_{h\in\mathcal{H}} \inf_{\mathcal{D}\in\Pi} \Delta_{\mathcal{H}'}(h, \mathcal{D}) \leq 2LW_1(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T})$$

where

$$W_1(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T}) := \inf_{\mathcal{D}\in\Pi} \mathop{\mathbb{E}}_{\mathbf{x}_s,\mathbf{x}_t\sim\mathcal{D}} [c(\mathbf{x}_s, \mathbf{x}_t)] \tag{25}$$

is the Wasserstein distance associated to metric c.

Proof. For any $\mathbf{x}_s, \mathbf{x}_t \in \mathcal{X}$,

$$\begin{aligned}
&|hh'(\mathbf{x}_s) - hh'(\mathbf{x}_t)| \\
&= |h(\mathbf{x}_s)(h'(\mathbf{x}_s) - h'(\mathbf{x}_t)) + h'(\mathbf{x}_t)(h(\mathbf{x}_s) - h(\mathbf{x}_t))| \\
&\leq |h(\mathbf{x}_s)||h'(\mathbf{x}_s) - h'(\mathbf{x}_t)| + |h'(\mathbf{x}_t)||h(\mathbf{x}_s) - h(\mathbf{x}_t)| \qquad (26) \\
&\leq |h'(\mathbf{x}_s) - h'(\mathbf{x}_t)| + |h(\mathbf{x}_s) - h(\mathbf{x}_t)| \qquad (27) \\
&\leq 2L \cdot c(\mathbf{x}_s, \mathbf{x}_t)
\end{aligned}$$

where we apply the triangle inequality to obtain (26) and we use the fact that the hypotheses from $\mathcal{H}$ and $\mathcal{H}'$ have values in $[-1, 1]$ to obtain (27). Then we use the $L-$Lipschitz property of h and h'. Taking the expectation with respect to an arbitrary distribution $\mathcal{D} \in \Pi$, then the supremum over $h' \in \mathcal{H}'$ yields:

$$\Delta_{\mathcal{H}'}(h, \mathcal{D}) \leq 2LW_1(\mathcal{D}_\mathcal{S}, \mathcal{D}_\mathcal{T})$$

Then taking the infimum over $\mathcal{D} \in \Pi$ followed by the supremum over $h \in \mathcal{H}$ yields the result. □

Section 4

We recall that we try to solve the following problem:

$$\min_{\substack{h\in\mathcal{H}\\ \mathcal{D}\in\Pi}} \mathop{\mathbb{E}}_{\mathbf{x},y\sim\mathcal{S}} [(\rho' - y.h(\mathbf{x}))_+] + \frac{1}{\beta}\Delta_{\mathcal{H}'}(h, \mathcal{D}). \tag{28}$$

Proposition 4. *Let $\mathcal{H}$ be the space of linear classifier with bounded ℓ^2 norm, and $\mathcal{H}'$ be the space of linear classifiers with bounded ℓ^1 norm. Let l denote the hinge loss, $\mathbf{D}_{st} := \mathbf{x}_s\mathbf{x}_s^T - \mathbf{x}_t\mathbf{x}_t^T$ and $(|\mathbf{D}_{st}\mathbf{w}|)_i := |(\mathbf{D}_{st}\mathbf{w})_i|$ for $1 \leq i \leq d$, where $\mathbf{w} \in \mathbb{R}^d$. Then, Problem (28) can be equivalently expressed as the following convex program:*

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^n\\ \mathcal{D}\in\Pi}} \mathop{\mathbb{E}}_{\mathbf{x},y\sim\mathcal{S}} \left[l(y.\mathbf{w}^T\mathbf{x})\right] + \delta \left\| \mathop{\mathbb{E}}_{\mathbf{x}_s,\mathbf{x}_t\sim\mathcal{D}} \left[|\mathbf{D}_{st}\mathbf{w}|\right] \right\|_\infty + \zeta\|\mathbf{w}\|_2^2 \tag{29}$$

where $\delta, \zeta > 0$ are two hyper-parameters related to the bounds on $\mathcal{H}$ and $\mathcal{H}'$.

Proof. Let $\nu > 0$ and $\eta > 0$ be the respective radii of $\mathcal{H}$ and $\mathcal{H}'$, i.e:

$$\mathcal{H} \simeq \{\mathbf{w} \in \mathbb{R}^d; \|\mathbf{w}\|_2 \leq \nu\} \tag{30}$$

$$\mathcal{H}' \simeq \{\mathbf{v} \in \mathbb{R}^d; \|\mathbf{v}\|_1 \leq \eta\} \tag{31}$$

were $\simeq$ denotes an equality up to an isomorphism of vector spaces. Also, let $\{\mathbf{u}_1, ..., \mathbf{u}_d\}$ be the canonical basis of $\mathbb{R}^d$. For $h \in \mathcal{H}$, i.e., for $\mathbf{w} \in \mathbb{R}^d$ with $\|\mathbf{w}\| \leq \nu$, the alignment term becomes:

$$\begin{aligned}
&\Delta_{\mathcal{H}'}(h, \mathcal{D}) \\
&= \sup_{h'\in\mathcal{H}'} \mathop{\mathbb{E}}_{\mathbf{x}_s,\mathbf{x}_t\sim\mathcal{D}} [|hh'(\mathbf{x}_s) - hh'(\mathbf{x}_t)|]
\end{aligned}$$

$$
\begin{aligned}
&= \sup_{\|v\|_1 \leq \eta} \mathop{\mathbb{E}}_{\mathbf{x}_s, \mathbf{x}_t \sim \mathcal{D}} \left[\left| \mathbf{v}^T \mathbf{D}_{st} \mathbf{w} \right| \right] \\
&= \eta \sup_{1 \leq k \leq d} \mathop{\mathbb{E}}_{\mathbf{x}_s, \mathbf{x}_t \sim \mathcal{D}} \left[\left| \mathbf{u}_k^T \mathbf{D}_{st} \mathbf{w} \right| \right] \quad (32) \\
&= \eta \sup_{1 \leq k \leq d} \mathop{\mathbb{E}}_{\mathbf{x}_s, \mathbf{x}_t \sim \mathcal{D}} \left[\mathbf{u}_k^T |\mathbf{D}_{st} \mathbf{w}| \right] \quad (33) \\
&= \eta \sup_{1 \leq k \leq d} \mathbf{u}_k^T \mathop{\mathbb{E}}_{\mathbf{x}_s, \mathbf{x}_t \sim \mathcal{D}} \left[|\mathbf{D}_{st} \mathbf{w}| \right] \\
&= \eta \left\| \mathop{\mathbb{E}}_{\mathbf{x}_s, \mathbf{x}_t \sim \mathcal{D}} \left[|\mathbf{D}_{st} \mathbf{w}| \right] \right\|_\infty ,
\end{aligned}
$$

where after replacing h and h' by linear classifiers $\mathbf{w}$ and $\mathbf{v}$, we use the convexity of $\mathbf{v} \mapsto \mathop{\mathbb{E}}_{\mathbf{x}_s, \mathbf{x}_t \sim \mathcal{D}} \left[\left| \mathbf{v}^T \mathbf{D}_{st} \mathbf{w} \right| \right]$ for a fixed $\mathbf{w}$, and the fact that the ℓ_1 unit ball is a polytope with vertices $\{\pm \eta \mathbf{u}_k; 1 \leq k \leq d\}$ to obtain (32). Then, we use the identity $\left| \mathbf{u}_k^T \mathbf{u} \right| = \mathbf{u}_k^T |\mathbf{u}|$, $\forall \mathbf{u} \in \mathbb{R}^d$ to obtain (33). The last steps are deduced by the linearity of the expectation and the definition of the ℓ_∞ norm.
Satisfying constraint $\|\mathbf{w}\|_2 \leq \nu$ is equivalent to adding $\zeta \|\mathbf{w}\|^2$, where ζ is a Lagrange multiplier who has a one-to-one correspondance with ν. Finally, as $\frac{1}{\beta}$ multiplies $\Delta_{\mathcal{H}'}(h, \mathcal{D})$ in Problem (28), setting $\delta = \frac{\eta}{\beta}$ yields the result. □

Empirical case and optimization problem

We recall that in the empirical case, one has access to finite data sets $\mathbf{S}_m = \{((\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\} \sim \mathcal{S}^m$, and $\mathbf{T}_n = \{(\mathbf{x}'_1, y'_1), ..., (\mathbf{x}'_n, y'_n)\} \sim \mathcal{T}^n$, where the labels of $\mathbf{T}_n$ are not used for learning classifier $\mathbf{w}$. We define $\hat{\Pi}$, the empirical counterpart of Π, as the set:

$$
\hat{\Pi} = \{\mathbf{\Gamma} \in \mathbb{R}_+^{m \times n}; \mathbf{\Gamma} \mathbf{1}_n = \frac{1}{m} \mathbf{1}_m; \mathbf{\Gamma}^T \mathbf{1}_m = \frac{1}{n} \mathbf{1}_n\} \quad (34)
$$

where $\mathbf{1}_p = (1, 1, ..., 1) \in \mathbb{R}^p$. Denoting $\mathbf{D}_{ij} = \mathbf{x}_i \mathbf{x}_i^T - \mathbf{x}'_j {\mathbf{x}'_j}^T \in \mathbb{R}^{d \times d}$, the empirical cost function of Problem (29) is:

$$
\frac{1}{m} \sum_{1 \leq i \leq m} l(y_{s,i}. \mathbf{w}^T \mathbf{x}_{s,i}) + \delta \left\| \sum_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} \gamma_{ij} |\mathbf{D}_{ij} \mathbf{w}| \right\|_\infty + \zeta \|\mathbf{w}\|_2^2 \quad (35)
$$

is a function of $\mathbf{w} \in \mathbb{R}^d$ and $\mathbf{\Gamma} \in \hat{\Pi}$ having elements γ_{ij}.

For a fixed $\mathbf{w} \in \mathbb{R}^d$, we would like to find $\mathbf{\Gamma} \in \hat{\Pi}$ minimizing:

$$
\left\| \sum_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} \gamma_{ij} |\mathbf{D}_{ij} \mathbf{w}| \right\|_\infty \quad (36)
$$

However, we have:

$$
\begin{aligned}
&\min_{\mathbf{\Gamma} \in \hat{\Pi}} \left\| \sum_{ij} \gamma_{ij} |\mathbf{D}_{ij} \mathbf{w}| \right\|_\infty \\
&= \min_{\mathbf{\Gamma} \in \hat{\Pi}} \max_{1 \leq k \leq d} \sum_{ij} \gamma_{ij} \mathbf{u}_k^T |\mathbf{D}_{ij} \mathbf{w}| \quad (37) \\
&= \min_{\mathbf{\Gamma} \in \hat{\Pi}} \max_{\mathbf{q} \in \mathbf{\Delta}_d} \sum_{ij} \gamma_{ij} \mathbf{q}^T |\mathbf{D}_{ij} \mathbf{w}| \quad (38) \\
&= \max_{\mathbf{q} \in \mathbf{\Delta}_d} \min_{\mathbf{\Gamma} \in \hat{\Pi}} \sum_{ij} \gamma_{ij} \mathbf{q}^T |\mathbf{D}_{ij} \mathbf{w}| \quad (39) \\
&= - \min_{\mathbf{q} \in \mathbf{\Delta}_d} \max_{\mathbf{\Gamma} \in \hat{\Pi}} \left(- \sum_{ij} \gamma_{ij} \mathbf{q}^T |\mathbf{D}_{ij} \mathbf{w}| \right) \quad (40)
\end{aligned}
$$

where $\mathbf{u}_k$ are the vectors of the canonical basis of $\mathbb{R}^d$. Due to the positivity of coordinates of $\sum_{ij} \gamma_{ij} \mathbf{u}_k^T |\mathbf{D}_{ij} \mathbf{w}|$, we obtain (37). Then, since the function $\mathbf{q} \mapsto \sum_{ij} \gamma_{ij} \mathbf{q}^T |\mathbf{D}_{ij} \mathbf{w}|$ is linear, its maximum is achieved over the vertices of the probability simplex, hence the equality between (37) and (38). Furthermore, due to the linearity of both $\mathbf{q} \mapsto \sum_{ij} \gamma_{ij} \mathbf{q}^T |\mathbf{D}_{ij} \mathbf{w}|$ and $\mathbf{\Gamma} \mapsto \sum_{ij} \gamma_{ij} \mathbf{q}^T |\mathbf{D}_{ij} \mathbf{w}|$, and to the convexity and compactness of Δ_d and $\hat{\Pi}$, Von-Neumann's minimax theorem allows us to permute the maximum and the minimum to obtain (39). Finally, introducing minus sign yields (40).

Smooth proxies used for optimization

We use the smooth proxies to provide smooth functions for `scipy`'s L-BFGS optimizer. They all depend on a parameter $\kappa > 0$. For all our experiments, we set $\kappa = 0.1$.

SMOOTH PROXY OF THE POSITIVE PART

We define the smooth proxy of the positive part function as:

$$
\mathrm{pos}_\kappa : t \mapsto \begin{cases} \frac{1}{2\kappa} \left(t + \frac{\kappa}{2} \right)^2 & \text{if } -\frac{\kappa}{2} \leq t \leq \frac{\kappa}{2} \\ (t)_+ & \text{otherwise} \end{cases}
$$

plotted in Figure 2 (left) and verifying $\mathrm{pos}_\kappa(t) \xrightarrow{\kappa \to 0} (t)_+$ for all $t \in \mathbb{R}$.

SMOOTH PROXY OF THE ABSOLUTE VALUE

Since for any $t \in \mathbb{R}$, one has $|t| = (t)_+ + (-t)_+$, we define a smooth proxy of the absolute value in a similar manner:

$$
\mathrm{abs}_\kappa(t) = \mathrm{pos}_\kappa(t) + \mathrm{pos}_\kappa(-t)
$$

plotted in Figure 2 (right) and verifying $\mathrm{abs}_\kappa(t) \xrightarrow{\kappa \to 0} |t|$ for all $t \in \mathbb{R}$.

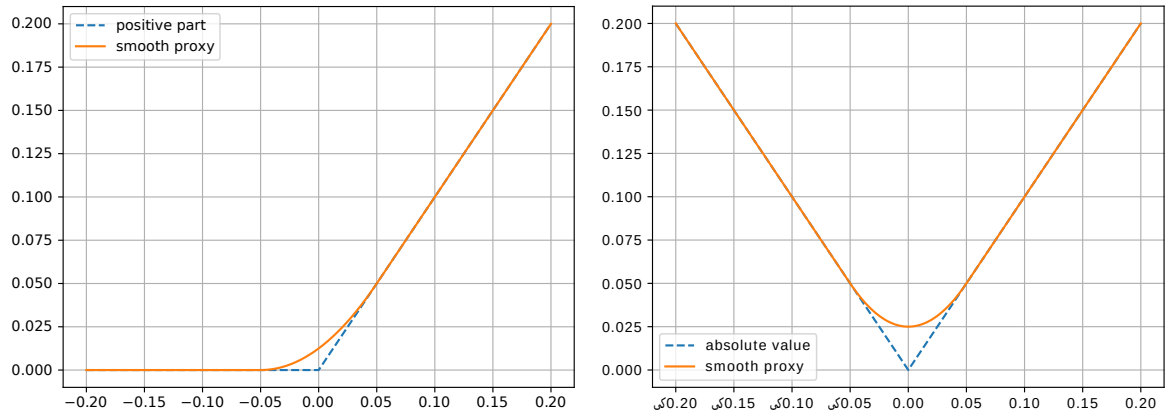

Figure 2. (**left**) Smooth proxy of the positive part function; (**right**) Smooth proxy of the absolute value function.

SMOOTH PROXY OF THE INFINITE NORM

As in Problem (29), the coordinates of vector $\mathop{\mathbb{E}}_{\mathbf{x}_s,\mathbf{x}_t\sim\mathcal{D}}[|\mathbf{D}_{st}\mathbf{w}|]$ are positive, computing its infinite norm is simply computing its largest coordinate. Hence, we use the $\mathrm{logsumexp}_\kappa$ function (Nesterov, 2005, Lemma 4) defined for $\mathbf{t} = (t_1, ..., t_d)$ in $\mathbb{R}^d$ as:

$$\mathrm{logsumexp}_\kappa(\mathbf{t}) = \kappa \log\left(\sum_{k=1}^{d} e^{\frac{t_k}{\kappa}}\right)$$

verifying $\mathrm{logsumexp}_\kappa(\mathbf{t}) \xrightarrow{\kappa\to 0} \max(t_1, ..., t_d)$ for all $\mathbf{t} \in \mathbb{R}^d$.

Proposition 5. *The margin violation loss does not respect the triangle inequality.*

Proof. Let $\rho, \epsilon_1, \epsilon_2 > 0$ such that $0 < \sqrt{\rho} \leq \epsilon_1 < \epsilon_2 < 1$. Also, let

$$x = y = \epsilon_2\sqrt{\rho} \quad \text{and} \quad z = \frac{\sqrt{\rho}}{\epsilon_1}.$$

We have $0 \leq x, y, z \leq 1$ and :

$$xz = yz = \epsilon_2\sqrt{\rho}\frac{\sqrt{\rho}}{\epsilon_1} = \frac{\epsilon_2}{\epsilon_1}\rho > \rho.$$

Hence,

$$[xz < \rho] = [yz < \rho] = 0.$$

However,

$$xy = \epsilon_2^2\rho < \rho,$$

hence $[xy < \rho] = 1$ and we have:

$$1 = [xy < 1] \not\leq [xz < 1] + [yz < 1] = 0.$$

□

References

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

Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

Zhang, Y., Liu, T., Long, M., and Jordan, M. Bridging Theory and Algorithm for Domain Adaptation. In *International Conference on Machine Learning*, pp. 7404–7413, 2019.