Appendix

A. *Proof of Proposition 1*

$1^\circ(\Rightarrow)$ For all $w=(z,y), s=(u,v)\in\mathbb{R}^D\times\mathbb{R}^C$, and $t>0$, we have

$$\begin{aligned}\left|\left\langle\int_0^t \nabla f(w+\tau s)d\tau, s\right\rangle\right| &= |f(w+ts)-f(w)|\\ &= |f((z,y)+t(u,v))-f(z,y)|\\ &= |f(z+tu,y+tv)-f(z,y)|\\ &\le \|tu\|+\alpha\|tv\|\\ &= t\|s\|_\alpha,\end{aligned}$$

where the inequality follows from (5). Subsequently,

$$\left|\left\langle\frac{1}{t}\int_0^t \nabla f(w+\tau s)d\tau, s\right\rangle\right| \le \|s\|_\alpha, \forall w,s,t.$$

Taking limit $t\to 0$ for both sides, by L'Hôpital's rule we have,

$$\lim_{t\to 0}\langle\nabla f(w+ts),s\rangle \le \|s\|_\alpha,\ \forall w,s.$$

Thus,

$$\frac{\langle\nabla f(w),s\rangle}{\|s\|_\alpha}\le 1,\forall w,s.$$

Taking supremum over s for both sides yields,

$$\sup_{s\neq\mathbf{0}}\frac{\langle\nabla f(w),s\rangle}{\|s\|_\alpha}\le 1,\forall w.$$

By the definition of dual norm, we obtain $\|\nabla f(w)\|_{\alpha*}\le 1$ for all w. Equivalently, $\|\nabla f(z,y)\|_{\alpha*}\le 1$ for all z and y.

$2^\circ(\Leftarrow)$ For all $w=(z,y)$ and $w'=(z',y')$, we have

$$\begin{aligned}|f(w)-f(w')| &= \left|\left\langle\int_0^1 \nabla f\left(w'+\tau(w-w')\right)d\tau, w-w'\right\rangle\right|\\ &\le \left\|\int_0^1 \nabla f(w'+\tau(w-w'))d\tau\right\|_{\alpha*}\cdot\|w-w'\|_\alpha\\ &\le \int_0^1 \|\nabla f(w'+\tau(w-w'))\|_{\alpha*}d\tau\cdot\|w-w'\|_\alpha\\ &\le 1\cdot\|w-w'\|_\alpha\\ &= \|z-z'\|+\alpha\|y-y'\|,\end{aligned}$$

where the first inequality comes from the definition of dual norm, the second inequality is due to the property of integral, and the last inequality follows from $\|\nabla f(z,y)\|_{\alpha*}\le 1$.

B. *Proof of Proposition 2*

We first provide the following lemma as a preparation of proving Proposition 2.

Lemma 1. *Consider the optimization problem given $x\in\mathbb{R}^n$*

$$\begin{aligned}&\min_{r\in\mathbb{R}^n}\ r_1x_1+\cdots+r_nx_n\\ &\textit{s.t.}\quad |r_1|+\cdots+|r_n|=1.\end{aligned}\tag{12}$$

Its optimal solution is

$$r_i^*=\begin{cases}\mathrm{sign}(x_i), & \textit{if } i=\arg\max_j\{|x_j|\},\\ 0, & \textit{otherwise},\end{cases}$$

and the corresponding optimal objective value is $\max\{|x_1|, \ldots, |x_n|\}$.

Proof. First we know that $\sum_{i=1}^{n} r_i x_i \leq \sum_{i=1}^{n} |r_i||x_i|$, and the equality can be attained by if $r_i x_i \geq 0$ for all i. Denote $\bar{r}_i = |r_i|$ and $\bar{x}_i = |x_i|$, we have the following equivalent reformulation of (12):

$$\begin{aligned} \min_{\bar{r}\in\mathbb{R}^n} \ & \bar{r}_1\bar{x}_1 + \cdots + \bar{r}_n\bar{x}_n \\ \text{s.t.} \ & \bar{r}_1 + \cdots + \bar{r}_n = 1, \\ & \bar{r}_i \geq 0, \end{aligned} \tag{13}$$

where $\bar{x}_i \geq 0$. Suppose $k = \arg\max_j\{\bar{x}_j\}$, then

$$\bar{r}_1\bar{x}_1 + \cdots + \bar{r}_n\bar{x}_n \leq \bar{r}_1\bar{x}_k + \cdots + \bar{r}_n\bar{x}_k = (\bar{r}_1 + \cdots + \bar{r}_n)\bar{x}_k = \bar{x}_k.$$

where the equality can be attained by

$$\bar{r}_i^* = \begin{cases} 1, & \text{if } i = \arg\max_j\{\bar{x}_j\}, \\ 0, & \text{otherwise.} \end{cases}$$

Since $\bar{r}_i = |r_i|$ and $\bar{x}_i = |x_i|$, we have

$$|r_i^*| = \begin{cases} 1, & \text{if } i = \arg\max_j\{|x_j|\}, \\ 0, & \text{otherwise.} \end{cases}$$

Further by $r_i x_i \geq 0$ for all i, we obtain the optimal solution to (12):

$$r_i^* = \begin{cases} \text{sign}(x_i), & \text{if } i = \arg\max_j\{|x_j|\}, \\ 0, & \text{otherwise.} \end{cases}$$

Then, the corresponding optimal objective value is

$$r_k x_k = \text{sign}(x_k)x_k = |x_k| = \max\{|x_1|, \ldots, |x_n|\}.$$

□

We are now ready to prove Proposition 2 formally. By definition,

$$\|g\|_{\alpha*} = \sup_{s\neq\mathbf{0}} \frac{|\langle g, s\rangle|}{\|s\|_\alpha} = \sup_{(u,v)\neq\mathbf{0}} \frac{|\langle a, u\rangle + \langle b, v\rangle|}{\|u\| + \alpha\|v\|}.$$

1° If $\|\cdot\|_\alpha = \|\cdot\|_1 + \alpha\|\cdot\|_1$, then

$$\begin{aligned} \|g\|_{\alpha*} &= \sup_{(u,v)\neq\mathbf{0}} \frac{|\langle a, u\rangle + \langle b, v\rangle|}{\|u\|_1 + \alpha\|v\|_1} \\ &= \sup_{u,v} \{|\langle a, u\rangle + \langle b, v\rangle| : \|u\|_1 + \alpha\|v\|_1 \leq 1\} \\ &= \sup_{u,v} \left\{ \sum_{i=1}^{K} a_i u_i + \sum_{j=1}^{C} b_j v_j : \sum_{i=1}^{K} |u_i| + \sum_{j=1}^{C} |\alpha v_j| \leq 1 \right\} \\ &= \sup_{u,v} \left\{ \sum_{i=1}^{K} a_i u_i + \sum_{j=1}^{C} \alpha v_j \frac{b_j}{\alpha} : \sum_{i=1}^{K} |u_i| + \sum_{j=1}^{C} |\alpha v_j| \leq 1 \right\} \\ &= \max\left\{ |a_1|, \ldots |a_K|, \frac{|b_1|}{\alpha}, \ldots, \frac{|b_C|}{\alpha} \right\} \qquad (14) \\ &= \max\left\{ \|a\|_\infty, \frac{\|b\|_\infty}{\alpha} \right\}, \end{aligned}$$

where equality (14) directly follows from Lemma 1 by simple variable substitution.

2° If $\|\cdot\|_\alpha = \|\cdot\|_2 + \alpha\|\cdot\|_2$, then

$$\|g\|_{\alpha*} = \sup_{(u,v)\neq\mathbf{0}} \frac{|\langle a, u\rangle + \langle b, v\rangle|}{\|u\|_2 + \alpha\|v\|_2}$$

By Cauchy-Schwartz inequality, we have

$$\frac{|\langle a, u\rangle + \langle b, v\rangle|}{\|u\|_2 + \alpha\|v\|_2} \leq \frac{\|a\|_2\|u\|_2 + \|b/\alpha\|_2\|\alpha v\|_2}{\|u\|_2 + \|\alpha v\|_2},$$

where the equality can be attained when $u = \mu a$ and $\alpha v = \nu b/\alpha$ for some $\mu, \nu \in \mathbb{R}$. Hence,

$$\begin{aligned}\|g\|_{\alpha*} &= \max_{(\mu,\nu)\neq(0,0)} \frac{\mu\|a\|_2^2 + \nu\|b/\alpha\|_2^2}{\mu\|a\|_2 + \nu\|b/\alpha\|_2} \\ &= \max_{(\mu,\nu)\neq(0,0)} \|a\|_2 + \frac{\nu\|b/\alpha\|_2^2 - \nu\|a\|_2\|b/\alpha\|_2}{\mu\|a\|_2 + \nu\|b/\alpha\|_2}.\end{aligned}$$

The optimal solution (μ^*, ν^*) depends on the sign of $\nu\|b/\alpha\|_2^2 - \nu\|a\|_2\|b/\alpha\|_2$. Therefore, if $\|b/\alpha\|_2 > \|a\|_2$, let $\mu^* = 0$ and $\nu^* > 0$, we have $\|g\|_{\alpha*} = \|b/\alpha\|_2$; if $\|b/\alpha\|_2 \leq \|a\|_2$, let $\mu^* > 0$ and $\nu^* = 0$, we have $\|g\|_{\alpha*} = \|a\|_2$. To summary, we obtain $\|g\|_{\alpha*} = \max\{\|a\|_2, \|b\|_2/\alpha\}$.

C. Corollary of Proposition 2

Based on Proposition 2, we can readily differentiate α-norm w.r.t. each dimension.

Corollary 1. *Let $g = (a, b)$, where $a \in \mathbb{R}^K$ and $b \in \mathbb{R}^C$. If $\|\cdot\|_\alpha$ is induced by ℓ_1 norm, then*

$$\frac{\partial}{\partial a_i}\|\cdot\|_{\alpha*} = \begin{cases} 1, & \textit{if } |a_i| > |a_k| \ \forall k \in [K]\backslash\{i\}, \textit{and } |a_i| > |b_j|/\alpha \ \forall j \in [C], \\ 0, & \textit{otherwise}, \end{cases}$$

$$\frac{\partial}{\partial b_j}\|\cdot\|_{\alpha*} = \begin{cases} \frac{1}{\alpha}, & \textit{if } |b_j| > |b_l| \ \forall l \in [C]\backslash\{j\}, \textit{and } |b_j| > |a_i|/\alpha \ \forall i \in [K], \\ 0, & \textit{if otherwise}. \end{cases}$$

If $\|\cdot\|_\alpha$ is induced by ℓ_2 norm, then

$$\frac{d}{dg}\|\cdot\|_{\alpha*} = \begin{cases} (a/\|a\|_2, \mathbf{0}_C), & \textit{if } \|a\|_2 > \|b\|_2/\alpha, \\ (\mathbf{0}_K, b/(\alpha\|b\|_2)), & \textit{if } \|a\|_2 < \|b\|_2/\alpha. \end{cases}$$

D. Proof of Theorem 1

Proof. Denote P'_{t,h_p} be the joint distribution whose density function is defined as $p'_{t,h_p}(z,y) \coloneqq \nu_t(z)\delta(y - h_p(z))$, where ν_t is the marginal density of the target feature z_t, and $\delta(\cdot)$ is the Dirac delta function. Correspondingly, let $\widehat{P}'_s$ be the source empirical joint distribution whose density function is $\hat{p}'_s(z,y) \coloneqq \frac{1}{n_s}\sum_{i=1}^{n_s}\delta((z,y)-(z_s^i, y_s^i))$, and $\widehat{P}'_{t,h_p}$ be the target empirical joint distribution whose density function is $\hat{p}'_{t,h_p}(z,y) \coloneqq \frac{1}{n_t}\sum_{i=1}^{n_t}\delta(z - z_t^i)\delta(y - h_p(z_t^i))$. We consider the expected error of any function $h_p : \mathcal{Z} \to \mathbb{R}^C$ on the target domain:

$$\begin{aligned}
\mathrm{err}_T(h_p) &= \mathbb{E}_{(z,y)\sim P'_t}[\|y - h_p(z)\|] \\
&\leq \mathbb{E}_{(z,y)\sim P'_t}[\|y - h^*(z)\| + \|h^*(z) - h_p(z)\|] \\
&= \mathrm{err}_T(h^*) + \mathbb{E}_{(z,y)\sim P'_t}[\|h^*(z) - h_p(z)\|] \\
&= \mathrm{err}_T(h^*) + \mathbb{E}_{z\sim\Xi_T}[\|h^*(z) - h_p(z)\|] \\
&= \mathrm{err}_T(h^*) + \int_{\mathcal{Z}} \|h^*(z) - h_p(z)\|\nu_t(z)dz \\
&= \mathrm{err}_T(h^*) + \int_{\mathcal{Z}}\int_{\mathcal{C}} \|h^*(z) - y\|\nu_t(z)\delta(y - h_p(z))\,dzdy \\
&= \mathrm{err}_T(h^*) + \int_{\mathcal{Z}\times\mathcal{C}} \|h^*(z) - y\|p'_{t,h_p}(z,y)dzdy \\
&= \mathrm{err}_T(h^*) + \mathbb{E}_{(z,y)\sim P'_{t,h_p}}[\|h^*(z) - y\|] \\
&= \mathrm{err}_T(h^*) + \mathrm{err}_S(h^*) + \mathrm{err}_{T,h_p}(h^*) - \mathrm{err}_S(h^*) \\
&\leq \mathrm{err}_T(h^*) + \mathrm{err}_S(h^*) + \left|\mathrm{err}_{T,h_p}(h^*) - \mathrm{err}_S(h^*)\right| \qquad (15)
\end{aligned}$$

Then for given $\kappa > 0$, we bound the last term in (15) as

$$\begin{aligned}
&\left|\text{err}_{T,h_p}(h^*) - \text{err}_S(h^*)\right| \\
=&\left|\int_{\mathcal{Z}\times\mathcal{C}} \|y_t - h^*(z_t)\| dP'_{t,h_p}(z_t, y_t) - \int_{\mathcal{Z}\times\mathcal{C}} \|y_s - h^*(z_s)\| dP'_s(z_s, y_s)\right| \\
=&\left|\int_{(\mathcal{Z}\times\mathcal{C})^2} \|y_s - h^*(z_s)\| - \|\hat{y}_t - h^*(z_t)\| d\Gamma^*((z_s, y_s), (z_t, \hat{y}_t))\right| && (16) \\
\leq&\int_{(\mathcal{Z}\times\mathcal{C})^2} |\|y_s - h^*(z_s)\| - \|\hat{y}_t - h^*(z_t)\|| \, d\Gamma^*((z_s, y_s), (z_t, \hat{y}_t)) \\
\leq&\int_{(\mathcal{Z}\times\mathcal{C})^2} |\|y_s - h^*(z_s)\| - \|\hat{y}_t - h^*(z_s)\|| + |\|\hat{y}_t - h^*(z_s)\| - \|\hat{y}_t - h^*(z_t)\|| \, d\Gamma^*((z_s, y_s), (z_t, \hat{y}_t)) \\
\leq&\int_{(\mathcal{Z}\times\mathcal{C})^2} \|y_s - \hat{y}_t\| + |h^*(z_s) - h^*(z_t)| \, d\Gamma^*((z_s, y_s), (z_t, \hat{y}_t)) && (17) \\
\leq&\int_{(\mathcal{Z}\times\mathcal{C})^2} \|y_s - \hat{y}_t\| + \kappa\|z_s - z_t\| d\Gamma^*((z_s, y_s), (z_t, \hat{y}_t)) + M\phi(\kappa) && (18) \\
=&\frac{1}{\alpha}\int_{(\mathcal{Z}\times\mathcal{C})^2} \|z_s - z_t\| + \alpha\|y_s - \hat{y}_t\| d\Gamma^*((z_s, y_s), (z_t, \hat{y}_t)) + M\phi\left(\frac{1}{\alpha}\right) && (19) \\
=&\frac{1}{\alpha} W_1\left(P'_s, P'_{t,h_p}\right) + M\phi\left(\frac{1}{\alpha}\right). && (20)
\end{aligned}$$

We provide some explanations for (16) to (20). (16) results from the fact that Γ^* is a joint distribution with marginals P_s and P'_{t,h_p}. (17) is due to the triangle inequality of $\|\cdot\|$. Further, we denote

$$\begin{aligned}
\Omega &\coloneqq \left\{(z_s, z_t) \in \mathcal{Z}\times\mathcal{Z} : \|h(z_s) - h(z_t)\| \leq \kappa\|z_s - z_t\|\right\}, \\
\bar{\Omega} &\coloneqq \left\{(z_s, z_t) \in \mathcal{Z}\times\mathcal{Z} : \|h(z_s) - h(z_t)\| > \kappa\|z_s - z_t\|\right\}.
\end{aligned}$$

Note that $\Omega \cup \bar{\Omega} = \mathcal{Z}\times\mathcal{Z}$ and $P(\Omega) \geq 1 - \phi(\kappa)$, then (18) can be obtained by the M-boundedness of h^*, i.e.,

$$\begin{aligned}
&\int_{(\mathcal{Z}\times\mathcal{C})^2} |h^*(z_s) - h^*(z_t)| \, d\Gamma^*((z_s, y_s), (z_t, \hat{y}_t)) \\
=&\int_{\mathcal{Z}\times\mathcal{Z}} |h^*(z_s) - h^*(z_t)| \, d\Gamma^*_z(z_s, z_t) \\
=&\int_{\Omega} |h^*(z_s) - h^*(z_t)| \, d\Gamma^*_z(z_s, z_t) + \int_{\bar{\Omega}} |h^*(z_s) - h^*(z_t)| \, d\Gamma^*_z(z_s, z_t) \\
\leq&\int_{\Omega} \kappa\|z_s - z_t\| d\Gamma^*_z(z_s, z_t) + \int_{\bar{\Omega}} M d\Gamma^*_z(z_s, z_t) \\
\leq&\int_{\mathcal{Z}\times\mathcal{Z}} \kappa\|z_s - z_t\| d\Gamma^*_z(z_s, z_t) + MP(\Omega) \\
\leq&\int_{\mathcal{Z}\times\mathcal{Z}} \kappa\|z_s - z_t\| d\Gamma^*_z(z_s, z_t) + M\phi(\kappa) \\
=&\int_{(\mathcal{Z}\times\mathcal{C})^2} \kappa\|z_s - z_t\| d\Gamma^*((z_s, y_s), (z_t, \hat{y}_t)) + M\phi(\kappa).
\end{aligned}$$

where Γ^*_z is the marginal distribution of (z_s, z_t) by marginalizing out y_s and $\hat{y}_t$ in Γ^*. Letting $\alpha = 1/\kappa$ gives (19). Besides, (20) results from the definition of Wasserstein distance.

Plugging (20) into (15), together with the triangle inequality of Wasserstein distance, we have

$$\begin{aligned}
\text{err}_T(h_p) \leq&\text{err}_T(h^*) + \text{err}_S(h^*) + \frac{1}{\alpha} W_1\left(P'_s, P'_{t,h_p}\right) + M\phi\left(\frac{1}{\alpha}\right) \\
\leq&\text{err}_T(h^*) + \text{err}_S(h^*) + \frac{1}{\alpha} W_1\left(\widehat{P}'_s, \widehat{P}'_{t,h_p}\right) + \frac{1}{\alpha} W_1\left(P'_s, \widehat{P}'_s\right) + \frac{1}{\alpha} W_1\left(P'_{t,h_p}, \widehat{P}'_{t,h_p}\right) + M\phi\left(\frac{1}{\alpha}\right) && (21)
\end{aligned}$$

Based on the Theorem 2.1 in [36], there exist β and n, if $\min\{n_s, n_t\} > n$, we have

$$\mathbb{P}\left\{W_1\left(P'_s, \widehat{P}'_s\right) \leq \sqrt{\frac{2\log(2/\varepsilon)}{\beta n_s}}\right\} \geq 1 - \frac{\varepsilon}{2}, \tag{22}$$

$$\mathbb{P}\left\{W_1\left(P'_{t,h_p}, \widehat{P}'_{t,h_p}\right) \leq \sqrt{\frac{2\log(2/\varepsilon)}{\beta n_t}}\right\} \geq 1 - \frac{\varepsilon}{2}. \tag{23}$$

Combining (22) and (23) with (21), we conclude that with probability at least $1-\varepsilon$, the following inequality holds

$$\mathrm{err}_T(h_p) \leq \frac{1}{\alpha} W_1\left(\widehat{P}'_s, \widehat{P}'_{t,h_p}\right) + \frac{1}{\alpha}\sqrt{\frac{2}{\beta}\log\left(\frac{2}{\varepsilon}\right)}\left(\frac{1}{\sqrt{n_s}} + \frac{1}{\sqrt{n_t}}\right) + \mathrm{err}_S(h^*) + \mathrm{err}_T(h^*) + M\phi\left(\frac{1}{\alpha}\right).$$

$\square$