Supplement to: ‘On Translation and Reconstruction Guarantees of the Cycle-Consistent Generative Adversarial Networks’

Appendix

Proof of Lemma (2). Let us begin by specifying the class of discriminators \( \mathcal{L}_X \equiv \mathcal{L}^1_\epsilon \). Now, given \( \alpha, \beta \in \mathcal{P}(Y) \)
\[
d_{\mathcal{L}_X}(\phi_{\#\alpha}, \phi_{\#\beta}) = \sup_{l \in \mathcal{L}_X} \left[ \mathbb{E}_{\phi_{\#\alpha}}l - \mathbb{E}_{\phi_{\#\beta}}l \right] = \sup_{l \in \mathcal{L}_X} \left[ \mathbb{E}_{\alpha}(l \circ \phi) - \mathbb{E}_{\beta}(l \circ \phi) \right].
\]
Due to the definition of supremum, for any \( \epsilon > 0 \exists \ l_\epsilon \in \mathcal{L}_X \) for which
\[
d_{\mathcal{L}_X}(\phi_{\#\alpha}, \phi_{\#\beta}) \leq \mathbb{E}_{\alpha}(l_\epsilon \circ \phi) - \mathbb{E}_{\beta}(l_\epsilon \circ \phi) + \epsilon
\]
Here, \( l_\epsilon \circ G_{Lip} := \{ l_\epsilon \circ f : f \in G_{Lip} \} \). Now,
\[
\sup_{l \in \mathcal{L}_X} \left[ \mathbb{E}_{\alpha}(l \circ g^*) - \mathbb{E}_{\beta}(l \circ g^*) \right] = \inf_{g \in L_{G_{Lip}}} \left\{ \mathbb{E}_{\alpha}|(l_\epsilon \circ \phi) - g| - \mathbb{E}_{\beta}|(l_\epsilon \circ \phi) - g| + \mathbb{E}_{\alpha}(g) - \mathbb{E}_{\beta}(g) \right\} + \epsilon,
\]
where \( [4] \) is due to the fact that \( g^* \in G_{Lip} \). As such,
\[
d_{\mathcal{L}_X}(\phi_{\#\alpha}, \phi_{\#\beta}) \leq 2 \inf_{g \in G_{Lip}} \| \phi - g' \|_\infty + L_G \sup_{l \in \mathcal{L}_X} \left[ \mathbb{E}_{\alpha}(l \circ g^*) - \mathbb{E}_{\beta}(l \circ g^*) \right] \]
(1)

Proof of Corollary (1). We have already noticed \( \mathbb{E}_\nu[d_{\mathcal{L}_X}(\nu, \hat{\nu}_{n_2})] \leq \mathcal{O}(k^2 n_2^{-\frac{1}{2}}), k \geq 2 \). Since the distance \( d_{\mathcal{L}_X}(\nu, \hat{\nu}_{n_2}) \) satisfies the bounded difference inequality, the application of McDiarmid’s inequality leads to
\[
\mathbb{P}\left(d_{\mathcal{L}_X}(\nu, \hat{\nu}_{n_2}) \leq \mathcal{O}(k^2 n_2^{-\frac{1}{2}}) + t \right) \geq 1 - \exp\left\{ -\frac{2 n_2 t^2}{B_y^2} \right\},
\]
where \( B_y = \text{diam}(\Omega_2) \) with respect to the metric \( c' \). We point out that \( [5] \) is a generalized version of Proposition 20 in [1]. Now, Theorem (1) tells us,
\[
d_{\mathcal{L}_X}(\mu_{n_1}, \phi_{\#} \hat{\nu}_{n_2}) \leq \epsilon + L_G d_{\mathcal{L}_X}(\nu, \hat{\nu}_{n_2}) + \mathcal{O}(C_1 W^{-\frac{1}{2}} L^{-\frac{1}{2}}),
\]
given $\epsilon > 0$ and $n_1 \leq \frac{W-d-1}{2} \lfloor \frac{|W-d-1|}{2} \rfloor + 2$. Combining these two results, we get

$$P\left(d_{\mathcal{L}^1}(\hat{\mu}_{n_1}, \phi_{#} \hat{\nu}_{n_2}) \leq \mathcal{O}(k^2 n_2)^{-\frac{1}{2}} + \frac{1}{2} \sqrt{\ln \left(\frac{1}{\delta}\right) + \mathcal{O}(C_1 W^{-\frac{1}{2}} L^{-\frac{1}{2}})} \right) \geq 1 - \delta,$$

by taking $\delta = \exp \left\{-\frac{2n_1^2}{L^2}\right\}$. The statement also holds if we replace the two sample sizes $n_1, n_2$ with $\min(n_1, n_2)$. In such a case, the Borel-Cantelli lemma implies that $d_{\mathcal{L}^1}(\hat{\mu}_{n_1}, \phi_{#} \hat{\nu}_{n_2}) \to 0$ almost surely (under $P$), provided $d, k$ remain fixed.

**Remark.** We draw the attention of the reader to a particular consequence of this result. Observe that the width ($W$) and depth ($L$) of the translator network are intrinsically related to the sample size ($n_1$) from the target law. In case $\min(n_1, n_2) \to \infty$, $W$ also follows suit, given that $L$ remains constant. As such, our ideal backward translator, achieving generation consistency, is a finite sample approximation of an infinitely wide ReLU network. Maps induced by such an infinitely wide network converge in distribution to a Gaussian process [2]. This determines the large sample property of $\phi$. Finding out the exact statistical properties of such a process in a parametric setup might be taken up as future work.

**Remark.** For any $n_1 \in \mathbb{N}^+, d_{\mathcal{L}^1}(\mu, \phi_{#} \hat{\nu}_{n_2}) \leq d_{\mathcal{L}^1}(\mu, \hat{\mu}_{n_1}) + d_{\mathcal{L}^1}(\hat{\mu}_{n_1}, \phi_{#} \hat{\nu}_{n_2})$. We have already seen that the second term on the right-hand side of the inequality vanishes eventually [Corollary 1]. Moreover, similar to [2]

$$P\left(d_{\mathcal{L}^1}(\mu, \hat{\mu}_{n_1}) \leq \mathcal{O}(d^2 n_1)^{-\frac{1}{2}} + t\right) \geq 1 - \exp \left\{-\frac{2n_1 t^2}{B_2^2}\right\}.$$

As a result, $d_{\mathcal{L}^1}(\mu, \hat{\mu}_{n_1}) \overset{a.s.}{\longrightarrow} 0$ (using Borel-Cantelli lemma). Hence, it can be concluded that $\phi_{#} \hat{\nu}_{n_2}$ converges weakly to $\mu$ in $\mathcal{P}(\mathcal{X})$ [Theorem 6.9 in [3]].

**Proof of Theorem (2).** Let us carry out the decomposition of the realized backward translation error, similar to that in Theorem (1).

$$d_{\mathcal{W}^1_{m, \infty}}(\hat{\mu}_{n_1}, \phi_{#} \hat{\nu}_{n_2}) \leq d_{\mathcal{W}^1_{m, \infty}}(\hat{\mu}_{n_1}, \phi_{#} \nu) + d_{\mathcal{W}^1_{m, \infty}}(\phi_{#} \nu, \phi_{#} \hat{\nu}_{n_2}).$$

Observe that $\mathcal{W}^1_{m, \infty} \subset \mathcal{W}^1_{1, \infty}$, for any positive integer $m$. Also, the class $\mathcal{W}^1_{1, \infty}$ is a dense subset of 1-Lipschitz functions on $\mathcal{X}$. As such, $d_{\mathcal{W}^1_{m, \infty}}(\hat{\mu}_{n_1}, \phi_{#} \nu) \leq d_{\mathcal{L}^1}(\hat{\mu}_{n_1}, \phi_{#} \nu) \leq \epsilon$, where $\epsilon > 0$ (as in the proof of Theorem (1)).

The remaining approximation error can similarly be upper bound using the same technique. However, it would be far from tight. Let us define a class of functions that help in the pursuit of sharper bounds.

**Definition (Hölder Space).** For $s \in \mathbb{R}_{>0}$, with $|s|$ indicating the largest integer strictly smaller than $s$, the Hölder space of order $s$ is defined as

$$C^s_L(\mathbb{R}^d) = \left\{ f \in C_u(\mathbb{R}^d) : \|f\|_{C^s} = \|f\|_{\mathcal{W}^s} + \sum_{|\alpha| = |s|} \sup_{x,y \in \mathbb{R}^d} \frac{|D^\alpha f(x) - D^\alpha f(y)|}{|x-y|^{s-|\alpha|}} < L \right\}.$$

Now, similar to the proof of Lemma (2), for any $\epsilon' > 0 \exists f_\epsilon \in \mathcal{W}^1_{1, \infty}$ such that

$$d_{\mathcal{W}^1_{m, \infty}}(\phi_{#} \alpha, \phi_{#} \beta) \leq d_{\mathcal{W}^1_{m, \infty}}(l_\epsilon \circ \phi) - d_{\mathcal{W}^1_{m, \infty}}(l_\epsilon \circ \phi) + \epsilon', \text{ where } \alpha, \beta \in \mathcal{P}(\mathcal{Y})$$

$$= \inf_{g \in G_{Lip}} \left\{ D^\alpha\{l_\epsilon \circ \phi\} - D^\beta\{l_\epsilon \circ \phi\} - g + \mathcal{E}_\alpha\{g\} - \mathcal{E}_\beta\{g\} \right\} + \epsilon'$$

$$\leq 2 \inf_{g \in G_{Lip}} \|\phi - g\|_{\infty} + \left\{ \sup_{l \in \mathcal{W}^1_{m, \infty}} \left[ D^\alpha\{l \circ g\} - D^\beta\{l \circ g\} \right] \right\} + \epsilon', \forall g \in G_{Lip}. \tag{3}$$

The first term in (3) is obtained due to the Lipschitz property of $l_\epsilon$. Here,

$$\sup_{l \in \mathcal{W}^1_{m, \infty}} \left[ D^\alpha\{l \circ g\} - D^\beta\{l \circ g\} \right] = d_{\mathcal{W}^1_{m, \infty}}(g^*_\alpha \alpha, g^*_\beta \beta) \leq d_{\mathcal{C}^s_L}(g^*_\alpha \alpha, g^*_\beta \beta) \tag{4}$$

$$= \sup_{l \in \mathcal{C}^s_L} \left\{ \mathcal{E}_{x \sim \alpha}\{l(x)\} - \mathcal{E}_{x \sim \beta}\{l(x)\} \right\}. \tag{5}$$
Inequality (4) is based on the observation that there exists \( r > 0 \) for which \( \mathcal{W}_r^{m, \infty} \subset C_r^m \). Given any \( f \in C_r^m \) and \( g^* \in G_{Lip} \),

\[
|f \circ g^*|_\infty = \left\{ \sup \{ f(g^*(y)) : y \in \mathbb{R}^k \} = \left\{ \sup \{ f(x) : x = g^*(y) \in \mathbb{R}^d, y \in \mathbb{R}^k \} \right. \right. \leq \left. \left. \sup \{ f(x) : x \in \mathbb{R}^d \} = \|f\|_\infty. \right. \right. \]

Moreover, for \( x, y \in \mathbb{R}^k, x \neq y \)

\[
\frac{|D^\alpha f(g^*(x)) - D^\alpha f(g^*(y))|}{|x - y|^{\alpha - |\alpha|}} = \frac{|D^\alpha f(g^*(x)) - D^\alpha f(g^*(y))|}{|g^*(x) - g^*(y)|^{\alpha - |\alpha|}} \leq \frac{|D^\alpha f(g^*(x)) - D^\alpha f(g^*(y))|}{|x^* - y^*|^{\alpha - |\alpha|}} (L_G)^{\alpha - |\alpha|},
\]

assuming \( x^* \neq y^* \in \mathbb{R}^d \). Here, we choose both the metrics \( c, c' \) to be \( L^1 \) in their respective spaces. This convention conforms to the rest of the discussion as well.

Also, for \( 1 \leq |\alpha| \leq m \) we have

\[
D^\alpha (f \circ g^*)(x) = s! \sum_{1 \leq |i| \leq |\alpha|} \frac{(D^i f)(g^*(x))}{i!} P_{s,i}(g^*; x),
\]

where \( P_{s,i}(g^*; x) \) is a homogeneous polynomial of degree \( |i| \). Schreuder et al. [Lemma 7.2 in [5]] show that \( |D^\alpha (f \circ g^*)(x)| < C \), where \( C > 0 \) is a constant. This implies that there exists \( r^* > 0 \) for which \( f \circ g^* \in C_r^{m} (\mathbb{R}^k) \). As such, we may upper bound [5] by replacing the supremum over \( C_r^{m} (\mathbb{R}^d) \circ g^* \) by the same over \( C_r^{m} (\mathbb{R}^k) \).

Hence, for \( \epsilon > 0 \)

\[
d_{\mathcal{W}_r^{m, \infty}}(\bar{\mu}_{n_1}, \phi \# \hat{\nu}_{n_2}) \leq 2 \inf_{g^* \in G_{Lip}} \|\phi - g^*\|_\infty + d_{C_r^{m}} (\nu, \hat{\nu}_{n_2}) + \epsilon.
\]

The expected approximation error in the base domain can be put under a deterministic upper bound given by \( \mathbb{E}_\nu \left[ d_{C_r^{2}} (\nu, \hat{\nu}_{n_2}) \right] \leq n_2^{-\frac{m}{2}} + \frac{\log n_2}{\sqrt{n_2}} \) [Lemma 2.8 in [6]]. As such, we get

\[
\mathbb{E}\left[ d_{\mathcal{W}_r^{m, \infty}}(\bar{\mu}_{n_1}, \phi \# \hat{\nu}_{n_2}) \right] \leq C(n_2^{-\frac{m}{2}} + \frac{\log n_2}{\sqrt{n_2}}) + C(\sqrt{L_G} B_{n} W^{-\frac{1}{2}} L^{-\frac{1}{2}}).
\]

**Proof of Proposition (1).** Let us denote the VC dimension of \( \mathcal{Y}(\mathcal{P}(\chi)) \) by \( v_x < \infty \). This criteria ensures that the target class of distributions are ‘learnable’. For example, VC-dim[\( \mathcal{Y}(G_d) \)] = \( O(d^2) \), where \( G_d \) = the class of \( d \)-dimensional Gaussian distributions [7].

Now, given \( g \in G_{Lip} \), for any \( n \in \mathbb{N}^+ \)

\[
d_{\mathcal{L}_2^2}(g \# \hat{\nu}_n, (g \# \nu)_n) \leq d_{\mathcal{L}_2^2}(g \# \hat{\nu}_n, g \# \nu) + d_{\mathcal{L}_2^2}(g \# \nu, (g \# \nu)_n) \leq L_G d_{\mathcal{L}_2^2}(\hat{\nu}_n, \nu) + B_x \|g \# \nu - (g \# \nu)_n\|_{TV}.
\]

Inequality (6) exploits the relation between Wasserstein and TV metrics [Theorem 4 in [8]]. We know there exist constants \( \bar{C}_1, \bar{C}_2 > 0 \) such that

\[
\mathbb{P}\left( \|g \# \nu - (g \# \nu)_n\|_{TV} \geq \bar{C}_1 \sqrt{\frac{v_x}{n} + t} \right) \leq \exp (-\bar{C}_2 n t^2),
\]

[Lemma 2 in [9]]. Using this argument along with (4) we obtain

\[
\mathbb{P}\left( d_{\mathcal{L}_2^2}(g \# \hat{\nu}_n, (g \# \nu)_n) \leq t + O(n^{-\frac{1}{2}}) + O(\sqrt{v_x} n^{-\frac{1}{2}}) \right) \geq 1 - 2 \exp (-C_2 n t^2),
\]

where \( C_2 = \frac{1}{2} \min \left\{ \frac{2}{L_G B_{\nu}}, \frac{\bar{C}_1}{B_x} \right\} > 0 \). As such, the function \( g \) is an information preserving map of degree 1, under the 1-Wasserstein metric, with a decaying error of order \( O(n^{-\frac{1}{2}}) \).
Proof of Lemma (4). Our characterization of the critics allow \( \mathcal{L}_X \) to be \( \mathcal{L}^1_c \) or \( W^{1,\infty} \). Under this setup, for any backward translator \( G \)

\[
d_{\mathcal{L}_X}(\hat{\mu}_{n_1}, G_\# \hat{\nu}_{n_2}) \leq d_{\mathcal{L}_X}(\hat{\mu}_{n_1}, (G_\# \nu)_{n_2}) + d_{\mathcal{L}_X}((G_\# \nu)_{n_2}, G_\# \hat{\nu}_{n_2}) \leq B_x \| \hat{\mu}_{n_1} - (G_\# \nu)_{n_2} \|_{TV} + \mathcal{E}_3 \leq B_x \| \hat{\mu}_{n_1} - \Gamma_{n_1} \|_{TV} + \Lambda(n_{n_1}, n_2) + \mathcal{E}_3,
\]

where \( \Gamma_{n_1} = \arg\min_{\tau \in \mathcal{P}(\mathcal{X})} \| \tau - \hat{\mu}_{n_1} \|_{TV} \). It is often called the Empirical Yatracos Minimizer \cite{10}. Observe that, \( \| \hat{\mu}_{n_1} - \Gamma_{n_1} \|_{TV} \leq \| \hat{\mu}_{n_1} - \mu \|_{TV} \). Now, in case the OT map \( T \) exists such that \( T_\# \nu = \mu \), we get \( \| \hat{\mu}_{n_1} - \Gamma_{n_1} \|_{TV} \leq \mathcal{E}_3 \)

Remark. The information loss (in the right-hand side of (7)) can be taken care of by deploying an IPT as the translator. As such, it is the term \( d_{\mathcal{L}_X}(\hat{\mu}_{n_1}, (G_\# \nu)_{n_2}) \) that mainly contributes to the upper bound. We have built the empirical distribution \( \hat{\mu}_{n_1} \) based on \( \{X_i\}_{i=1}^{n_1} \sim \mu \). Similarly, let \( (G_\# \nu)_{n_2} \)

be based on \( \{Y_i\}_{i=1}^{n_2} \sim G_\# \nu \). We may write

\[
d_{\mathcal{L}_X}(\hat{\mu}_{n_1}, (G_\# \nu)_{n_2}) = \sup_{f \in \mathcal{L}_X} \left| \sum_{i=1}^{N} W_i f(Z_i) \right|,
\]

where \( N = n_1 + n_2 \); \( W_i = \frac{1}{n_1} \) when \( Z_i = X_i \), \( i = 1, \ldots, n_1 \) and \( W_{n_1+j} = -\frac{1}{n_2} \) when \( Z_{n_1+j} = Y_j \), \( j = 1, \ldots, n_2 \). Under this framework, the solution to (3) can be achieved by solving a linear program, given that \( \mathcal{L}_X \equiv \mathcal{L}^1 \) [Theorem 2.1 in \cite{11}]. This provides a pathway to get hold of the realized approximation error, making the upper bound deterministic.

Proof of Lemma (5). Given translator maps \( G \in \mathcal{F}(\mathcal{Y}, \mathcal{P}(\mathcal{X})) \) and \( F \in \mathcal{F}(\mathcal{X}, \mathcal{P}(\mathcal{Y})) \), the cyclic loss in the space \( \mathcal{X} \) can be broken down as the following:

\[
\| \mu - (G \circ F)_\# \mu \|_1 \leq \| \mu - G_\# \nu \|_1 + \| G_\# \nu - (G \circ F)_\# \mu \|_1,
\]

where

\[
\| G_\# \nu - (G \circ F)_\# \mu \|_1 = \| G_\# \nu - G_\# (F_\# \mu) \|_1 = 2 \sup_{\omega \in \mathcal{S}(\mathcal{X})} | G_\# \nu(\omega) - G_\# (F_\# \mu)(\omega) |
\]

\begin{align*}
&= 2 \sup_{\omega \in \mathcal{S}(\mathcal{X})} | \nu(G^{-1}(\omega)) - F_\# \mu(G^{-1}(\omega)) | \\
&\leq 2 \sup_{\omega \in \mathcal{S}(\mathcal{Y})} | \nu(\omega') - F_\# \mu(\omega') | = \| \nu - F_\# \mu \|_1.
\end{align*}

The inequality holds by taking supremum over all measurable sets belonging to the Borel \( \sigma \)-algebra on \( \mathcal{Y} \) instead of the particular path directed by \( G^{-1} \). As such

\[
\| \mu - (G \circ F)_\# \mu \|_1 \leq \| \mu - G_\# \nu \|_1 + \| \nu - F_\# \mu \|_1.
\]

Similarly, \( \| \nu - (F \circ G)_\# \nu \|_1 \leq \| \nu - F_\# \mu \|_1 + \| \mu - G_\# \nu \|_1 \). Hence the proof.

Proof of Theorem (3). Given a measurable function \( f : \mathbb{R}^d \to \mathbb{R} \), let us define its convolution with the kernel \( K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \) as the following:

\[
K_h(f) = \int_{\mathbb{R}^d} K_h(\cdot, y)f(y)dy = \frac{1}{h^d} \int_{\mathbb{R}^d} K(\frac{\cdot - y}{h})f(y)dy,
\]

where \( \frac{y}{h} = (\frac{y_1}{h}, \ldots, \frac{y_d}{h}) \), \( h > 0 \). We begin by taking \( K \) to be regularly invariant. Now,

\[
\| p_\mu - p_{\phi_\# \nu} \|_1 \leq \| p_\mu - K_h(p_\mu) \|_1 + \| K_h(p_\mu) - K_h(p_{\phi_\# \nu}) \|_1 + \| K_h(p_{\phi_\# \nu}) - p_{\phi_\# \nu} \|_1
\]

\begin{align*}
&\leq J \| p_\mu - K_h(p_\mu) \|_p + \| K_h(p_\mu) - K_h(p_{\phi_\# \nu}) \|_1 + J \| K_h(p_{\phi_\# \nu}) - p_{\phi_\# \nu} \|_{p'},
\end{align*}

(9)
where $J > 0$. The existence of such a constant, and hence the inequality (9), is ensured by the fact $\|f\|_1 \leq J \|f\|_p$, $p \geq 1$ since we have $\lambda(\Omega_x) < \infty$. Also, there exists a constant $l$ depending upon $m_s$ and $K$, such that $\|K_\mu(p, \mu) - \mu\|_p \leq l \|D^{m_s} p, \mu\|_p$ [Proposition 4.3.33 in [12]]. As such, we get hold of a constant $J^* = Jl$ for which

$$
\left\| p - p_{\phi_{\#} \nu} \right\|_1 \leq J^* \left\{ \left\| D^{m_s} p, \mu \right\|_p + \left\| D^{m_s} p_{\phi_{\#} \nu} \right\|_p \right\}\left( \kappa + J^* \right),
$$

(by Assumption 2). Observe that,

$$
K_\mu(p, \mu) - K_\mu(p_{\phi_{\#} \nu}, \nu) = \frac{1}{h^d} \int \left\{ K(\frac{x}{h}, \frac{y}{h}) - K(\frac{x}{h}, \frac{z}{h}) \right\} d\kappa(y, z),
$$

where $\kappa$ is a coupling between $\mu$ and $\phi_{\#} \nu$. Hence,

$$
\left\| K_\mu(p, \mu) - K_\mu(p_{\phi_{\#} \nu}, \nu) \right\|_1 \leq \int \left\{ \frac{1}{h^d} \int \left| K(\frac{x}{h}, \frac{y}{h}) - K(\frac{x}{h}, \frac{z}{h}) \right| dx \right\} d\kappa(y, z) \leq \frac{M^*}{h} \int |y - z| d\kappa(y, z),
$$

where $M^*$ is a positive constant. The step (10) is due to Jensen’s inequality, whereas (11) exploits the invariance of $K$. Since the inequality holds for all possible measure couples $\kappa$, we conclude

$$
\left\| K_\mu(p, \mu) - K_\mu(p_{\phi_{\#} \nu}, \nu) \right\|_1 \leq \frac{M^*}{h} W_1^1(\mu, \phi_{\#} \nu),
$$

where $c \equiv L^1$. A similar inference can be drawn for a general class of metrics $c$ by altering the specification of the same in the definition of invariance. Now, choose

$$
h = \left\{ \frac{W_1^1(\mu, \phi_{\#} \nu)}{\left\| D^{m_s} p, \mu \right\|_p + \left\| D^{m_s} p_{\phi_{\#} \nu} \right\|_p} \right\}^{\frac{1}{m_s + \tau}}.
$$

Finally, we obtain

$$
\left\| p - p_{\phi_{\#} \nu} \right\|_1 \leq M \left\{ \left\| D^{m_s} p, \mu \right\|_p + \left\| D^{m_s} p_{\phi_{\#} \nu} \right\|_p \right\} \left( \frac{1}{m_s + \tau} \right),
$$

where $M = 2(J^* \lor M^*)$.

\[\square\]

**Proof of Proposition (2).** Using Lemma (5),

$$
\mathcal{L}_{cyc}(\hat{\mu}_{\# n_1}, \hat{\nu}_{\# n_2}, F, G) = \|\hat{\mu}_{n_1} - (G \circ F)_{\#} \hat{\mu}_{n_1}\|_1 + \|\hat{\nu}_{n_2} - (F \circ G)_{\#} \hat{\nu}_{n_2}\|_1 \leq 4 \left\{ \|\hat{\mu}_{n_1} - G_{\#} \hat{\nu}_{n_2}\|_{TV} + \|\hat{\nu}_{n_2} - F_{\#} \hat{\mu}_{n_1}\|_{TV} \right\}.
$$

Now, a similar decomposition of the translation errors under the TV metric, as in the proof of Lemma (4), results in the following:

$$
\|\hat{\mu}_{n_1} - G_{\#} \hat{\nu}_{n_2}\|_{TV} \leq \|\hat{\mu}_{n_1} - \Gamma_{n_1}\|_{TV} + \|\hat{\Gamma}_{n_1} - (G_{\#} \hat{\nu})_{n_2}\|_{TV} + \|\hat{\Gamma}_{n_1} - G_{\#} \hat{\nu}_{n_2}\|_{TV} \leq \|\hat{\mu}_{n_1} - \mu\|_{TV} + \frac{\Lambda_{\mu n_1 n_2}}{B_x} + \|\hat{\Gamma}_{n_1} - G_{\#} \hat{\nu}_{n_2}\|_{TV}.
$$

Similarly, given that $\hat{\Gamma}_{n_2} = \text{argmin}_{\tau \in \mathcal{P}(Y)} \|\tau - \hat{\nu}_{n_2}\|_{TV}$

$$
\|\hat{\nu}_{n_2} - F_{\#} \hat{\mu}_{n_1}\|_{TV} \leq \|\hat{\nu}_{n_2} - \nu\|_{TV} + \frac{\Lambda_{\mu n_1 n_2}}{B_y} + \|\hat{\Gamma}_{n_1} - (F_{\#} \hat{\mu})_{n_1} - F_{\#} \hat{\mu}_{n_1}\|_{TV}.
$$

\[\square\]
Proof of Theorem (4). Let \( \phi \in \Phi(W, L) \), as specified in Theorem (1). Also, let \( \psi \in \Phi(W', L') \) be a forward translator that achieves consistency. Observe that
\[
\hat{L}_{cyc}(\hat{\mu}, \hat{\nu}, \psi, \phi) \leq \|\hat{\mu} - \mu\|_1 + \|\hat{\nu} - \nu\|_1 + \hat{L}_{cyc}(\mu, \nu, \psi, \phi)
\]
\[
\leq \|\hat{\mu} - \mu\|_1 + \|\hat{\nu} - \nu\|_1 + 2\{\|\mu - \phi\#\nu\|_1 + \|\nu - \psi\#\mu\|_1\}. \tag{12}
\]
For \( 1 \leq p, q < \infty \), we know that
\[
\mathbb{E}\left[\|\hat{\phi}(\mu) - \phi\|_p\right] \lesssim n_1 - \frac{m_{x+y}}{n_1^q},
\]
[Theorem 6.1 in \( [13] \).] Similarly, for the estimation error in \( \mathcal{Y} \), \( \mathbb{E}\left[\|\hat{\phi}(\mu) - \phi\|_q\right] \lesssim n_2 - \frac{m_{x+y}}{n_2^q} \).
Moreover, Theorem (3) implies that
\[
\left\{\|\hat{\mu} - \mu\phi\|_1 \right\}^{\frac{m_{x+y}}{m_{x+y}} - 1} \leq R d(\mu, \phi\#\nu) \leq R \left\{\left\|d(\mu, \hat{\mu}, 1) + d(\phi\#\nu, \hat{\mu}, 1)\right\|_1\right\}, \tag{13}
\]
where \( R = \frac{m_{x+y}}{m_{x+y}} \left\|D^{mx}p_{\mu}\right\|_p + \|D^{mx}p_{\phi\#\nu}\|_p \right\|^\frac{1}{r} \), and \( \hat{\mu}_1 \) is an usual empirical measure corresponding to \( \mu \). The term \( d(\mu, \hat{\mu}, 1) \) can be made arbitrarily small due to the construction of \( \phi \) [Lemma (1)]. Also, we have already seen that \( \mathbb{E}\left[\left\|d(\mu, \hat{\mu}, 1)\right\|_1\right] \lesssim n_1 - \frac{1}{n_1^q} \).
As such,
\[
\mathbb{E}\left[\|\hat{\mu} - \mu\|_1 + 2\|\mu - \phi\#\nu\|_1\right] \leq O \left(n_1 - \frac{m_{x+y}}{n_1^q}\right),
\]
by applying Jensen’s inequality to \( [13] \). This bound, together with a similar result corresponding to its forward counterpart, will imply
\[
\mathbb{E}\left[\hat{L}_{cyc}(\hat{\mu}, \hat{\nu}, \psi, \phi)\right] \lesssim \max\left\{n_1 - \frac{m_{x+y}}{n_1^q}, n_2 - \frac{m_{x+y}}{n_2^q}\right\}.
\]

Proof of Corollary (2). We point out that, \( K(x, y) \) can be taken in particular as \( \tilde{K}(|x - y|) \), where \( \tilde{K} : \mathbb{R}^d \to \mathbb{R} \) identically follows the traits of \( K \). Under such a kernel function,
\[
\left\|\mathbb{E}[\hat{\rho}_{\mu, n}^{1, 1}] - \rho_{\mu}\right\|_1 \leq l^* h^m_x,
\]
for some constant \( l^* > 0 \). Now, given an \( \epsilon \leq \frac{1}{2^d} \), concentration inequalities on kernel density estimates tell us: there exists constants \( E_2 > 0 \) such that
\[
\mathbb{P}\left(\left\|\hat{\rho}_{\mu, n} - \mathbb{E}[\hat{\rho}_{\mu, n}]\right\|_\infty > \epsilon\right) \leq E_1 \left(\frac{\sqrt{d} B_{\epsilon}}{\nu^{d+1} \epsilon^{d}}\right)^d \exp \left(- E_2 n_1 \epsilon^2 h^d\right).
\]
The exact value of \( E_2 = \frac{\epsilon}{28 K(0)} \) can be obtained based on the convention that \( \tilde{K}(\cdot) \) achieves its modal value at 0. Such a centering can always be done. Hence,
\[
\mathbb{P}\left(\left\|\hat{\rho}_{\mu, n} - \rho_{\mu}\right\|_1 > \epsilon + l^* h^m_x\right) \leq E_1 \left(\frac{\sqrt{d} B_{\epsilon}}{\nu^{d+1} \epsilon^{d}}\right)^d \exp \left(- E_2 n_1 \epsilon^2 h^d\right). \tag{14}
\]
By applying Borel-Cantelli lemma one can show that \( \left\|\hat{\rho}_{\mu, n} - \rho_{\mu}\right\|_1 \to 0 \) almost surely, under suitable choice of \( h \equiv h(n_1, m_x, d) \). This inspires a similar concentration for the estimate \( \hat{\rho}_{\nu, n} \) around \( \rho_{\nu} \), under \( L^1 \). As such, by taking the corresponding bandwidth \( \hat{h} \equiv \hat{h}(n_2, m_{\nu}, k) \), it can also be said that \( \left\|\hat{\rho}_{\nu, n} - \rho_{\nu}\right\|_1 \to 0 \) almost surely. To unify the two processes, one may assess the convergence based on \( n = \min\{n_1, n_2\} \). Putting these results back in \( [12] \), along with \( [13] \), we conclude
\[
\hat{L}_{cyc}(\hat{\mu}, \hat{\nu}, \psi, \phi) \to 0, \text{ almost surely.}
\]
In other words, \( (\phi \circ \psi)\#\hat{\mu}_{n_1} \to \mu \) and \( (\psi \circ \phi)\#\hat{\mu}_{n_2} \to \nu \), both in total variation. \( \square \)
Identity loss

Let us first rewrite the identity loss in terms of the underlying measures. Based on the notations in our framework,

$$
\mathcal{L}_{id}(\mu, \nu, F, G) = \|\mu - F_{\#}\mu\|_1 + \|\nu - G_{\#}\nu\|_1.
$$

Observe that the distributions must be equivariate to conform to this loss. Moreover,

$$
\|\mu - \nu\|_1 - \|F_{\#}\mu - \nu\|_1 \leq \|\mu - F_{\#}\mu\|_1. \tag{15}
$$

If the forward translated law $F_{\#}\mu$ is Sobolev-smooth of order $m_y$ (Assumption 2), Theorem (3) asserts the existence of a constant $R' > 0$ such that

$$
\left\| p_{\nu} - p_{F_{\#}\mu} \right\|_1 \leq R' \left[ d_{W_2}(\nu, F_{\#}\mu) \right]^{\frac{m_y}{m_y + 1}}.
$$

In case $F$ is also translation consistent, the second term on the left-hand side of (15) vanishes. A similar conclusion can be drawn for the quantity $\|\nu - G_{\#}\nu\|_1$ as well. As such, the cumulative identity loss from both domains cannot be minimized beyond the intrinsic discrepancy between the input distributions.

References


