Appendix

A  Empirical Estimates

Lemma 1. As $|D| \to \infty$, if $W_1(p_S, p_{S_a}) < \infty$ for all $a$, the empirical barycenter satisfies $\lim \sum_a \hat{p}_a W_1(\hat{p}_S, \hat{p}_{S_a}) \to \sum_a p_a W_1(p_S, p_{S_a})$ almost surely\(^7\).

Proof. By triangle inequality:

\[ \sum_a \hat{p}_a W_1(\hat{p}_S, p_{S_a}) \leq \sum_a \hat{p}_a W_1(\hat{p}_S, \hat{p}_{S_a}) + \hat{p}_a W_1(p_{S_a}, \hat{p}_{S_a}) ; \tag{4} \]

\[ \sum_a p_a W_1(p_S, \hat{p}_{S_a}) \leq \sum_a p_a W_1(p_S, p_{S_a}) + p_a W_1(p_{S_a}, \hat{p}_{S_a}) . \tag{5} \]

Since $p_S$ and $\hat{p}_S$ are the weighted barycenters of $\{p_{S_a}\}$ and $\{\hat{p}_{S_a}\}$ respectively:

\[ \sum_a p_a W_1(p_S, p_{S_a}) \leq \sum_a p_a W_1(p_S, p_{S_a}) ; \tag{6} \]

\[ \sum_a \hat{p}_a W_1(\hat{p}_S, \hat{p}_{S_a}) \leq \sum_a \hat{p}_a W_1(\hat{p}_S, \hat{p}_{S_a}) . \tag{7} \]

Combining Eqs. (4) and (6), and (5) and (7):

\[ \sum_a p_a W_1(p_S, p_{S_a}) \leq \sum_a p_a W_1(p_S, p_{S_a}) + p_a W_1(p_{S_a}, \hat{p}_{S_a}) \]
\[ \leq \sum_a \hat{p}_a W_1(\hat{p}_S, \hat{p}_{S_a}) + |\hat{p}_a W_1(\hat{p}_S, \hat{p}_{S_a}) - p_a W_1(p_{S_a}, \hat{p}_{S_a})| + p_a W_1(p_{S_a}, \hat{p}_{S_a}) \]
\[ \leq \sum_a \hat{p}_a W_1(\hat{p}_S, \hat{p}_{S_a}) + |\hat{p}_a - p_a| \cdot |W_1(p_{S_a}, \hat{p}_{S_a})| + p_a W_1(p_{S_a}, \hat{p}_{S_a}) \]

\[ \sum_a \hat{p}_a W_1(\hat{p}_S, \hat{p}_{S_a}) \leq \sum_a \hat{p}_a W_1(p_S, p_{S_a}) + \hat{p}_a W_1(p_{S_a}, \hat{p}_{S_a}) \]
\[ \leq \sum_a p_a W_1(p_S, p_{S_a}) + |p_a W_1(p_S, p_{S_a}) - \hat{p}_a W_1(p_{S_a}, p_{S_a})| + \hat{p}_a W_1(p_{S_a}, \hat{p}_{S_a}) \]
\[ \leq \sum_a p_a W_1(p_S, p_{S_a}) + |p_a - \hat{p}_a| \cdot |W_1(p_{S_a}, p_{S_a})| + \hat{p}_a W_1(p_{S_a}, \hat{p}_{S_a}) . \]

Therefore the following inequality holds almost surely:

\[ \left| \sum_a p_a W_1(p_S, p_{S_a}) - \sum_a \hat{p}_a W_1(\hat{p}_S, \hat{p}_{S_a}) \right| \leq \sum_a \hat{p}_a W_1(p_{S_a}, \hat{p}_{S_a}) + |p_a - \hat{p}_a| \cdot W_1(p_S, p_{S_a}) \]
\[ \leq \sum_a W_1(p_{S_a}, \hat{p}_{S_a}) + |p_a - \hat{p}_a| \cdot W_1(p_S, p_{S_a}) \]
\[ \leq \sum_a W_1(p_{S_a}, \hat{p}_{S_a}) + |p_a - \hat{p}_a| \cdot W_1(p_S, p_{S_a}) . \]

Since $W_1(p_{S_a}, \hat{p}_{S_a}) \to 0$ almost surely for all $a$ (see Weed and Bach (2017)), and $\hat{p}_a \to p_a$ almost surely (by the strong law of large numbers) and $W_1(p_S, p_{S_a}) < \infty$ for all $a$, the result follows:

\[ \lim \sum_a \hat{p}_a W_1(\hat{p}_S, \hat{p}_{S_a}) \to \sum_a p_a W_1(p_S, p_{S_a}) , \]

almost surely. \hfill \Box

\(^7\)See Klenke (2013) for a formal definition of almost sure convergence of random variables.
B Generalization

The following lemma addresses generalization of the Wasserstein-1 objective. Assume \( W_1(p_{S_a}, p_S) \leq L \) for all \( a \in A \). Let \( P_S, P_{S_a} \) and \( P_S \) be the cumulative density functions of \( S, S_a \) and \( S \). Assume these random variables all have domain \( \Omega = [0, 1] \) and that all \( P \in \{ P_S, P_S \}_a \in A \) are continuous, then:

**Lemma 5.** For any \( \epsilon, \delta > 0 \), if \( \min \left[ \tilde{N}, \min_a \left[N_a \right] \right] \geq \frac{16 \log(2|A|/\delta)|A|^2 \max[1,L]^2}{\epsilon^2} \), with probability \( 1 - \delta \):

\[
\sum_{a \in A} \hat{p}_a W_1(p_{S_a}, p_S) \leq \sum_{a \in A} \hat{p}_a W_1(\hat{p}_{S_a}, \hat{p}_S) + \epsilon.
\]

In other words, provided access to sufficient samples, a low value of \( \sum_{a} \hat{p}_a W_1(\hat{p}_{S_a}, \hat{p}_S) \) implies a low value for \( \sum_{a} p_a W_1(p_{S_a}, p_S) \) with high probability and therefore good performance at test time.

**Proof.** We start with the case when \( p_S = p_{S_a} \). By the triangle inequality for Wasserstein-1 distances, for all \( a \in A \):

\[
\hat{p}_a W_1(p_{S_a}, p_S) \leq \hat{p}_a W_1(p_{S_a}, \hat{p}_S) + \hat{p}_a W_1(\hat{p}_S, p_S) + \hat{p}_a W_1(\hat{p}_{S_a}, p_{S_a}).
\]

Let \( \hat{P} \) for \( P \in \{ P_S, P_S \}_a \in A \) denote the empirical CDF of \( P \). Since their domain is restricted to \([0, 1]\) and are one dimensional random variables:

\[
W_1(\hat{p}_{S_a}, p_{S_a}) = \int_0^1 |\hat{P}(x) - P(x)| dx
\]

For \( S_a \in \{ S, \tilde{S} \}_a \in A \). Since \( P \in \{ P_S, P_S \}_a \in A \) are all continuous, the Dvorestky-Kiefer-Wolfowitz theorem (see main theorem in Massart (1990)) and the condition \( \min \left[ \tilde{N}, \min_a \left[N_a \right] \right] \geq \frac{16 \log(2|A|/\delta)|A|^2 \max[1,L]^2}{\epsilon^2} \) implies that:

\[
\mathbb{P} \left( \sup_{x \in [0,1]} |\hat{P}(x) - P(x)| \geq \frac{\epsilon}{4} \right) \leq \frac{\delta}{2|A|}
\]

Since all the random variables have domain \([0, 1]\) this in turn implies that for all \( S_a \in \{ S, \tilde{S} \}_a \in A \):

\[
\mathbb{P} \left( W_1(\hat{p}_{S_a}, p_{S_a}) \geq \epsilon \right) \leq \frac{\delta}{2|A|}
\]

And therefore that with probability \( 1 - \frac{\delta}{2} \) the following inequalities hold simultaneously for all \( a \in A \):

\[
\hat{p}_a W_1(\hat{p}_S, p_S) \leq \frac{\hat{p}_a \epsilon}{4} \quad \text{and} \quad \hat{p}_a W_1(\hat{p}_{S_a}, p_{S_a}) \leq \frac{\hat{p}_a \epsilon}{4}.
\]

Summing Eq. (8) over \( a \) and applying the last observation yields

\[
\sum_{a \in A} \hat{p}_a W_1(p_{S_a}, p_S) \leq \sum_{a \in A} \hat{p}_a W_1(\hat{p}_{S_a}, \hat{p}_S) + \frac{\epsilon}{2}.
\]

Recall that we assume \( \forall a \in A \)

\[
W_1(p_{S_a}, p_S) \leq L.
\]

By concentration of measure of Bernoulli random variables, with probability \( 1 - \frac{\delta}{2} \) the following inequality holds simultaneously for all \( a \in A \):

\[
|p_a - \hat{p}_a| \leq \frac{\epsilon}{4|A| \max[1,L]}. \quad (11)
\]

Consequently the desired result holds:

\[
\sum_{a \in A} p_a W_1(p_{S_a}, p_S) \leq \sum_{a \in A} \hat{p}_a W_1(\hat{p}_{S_a}, \hat{p}_S) + \epsilon.
\]
If $p_S$ equals the weighted barycenter of the population level distributions \{p_{S_a}\}, then
\[
\sum_{a \in A} p_a W_1(p_{S_a}, p_S) \leq \sum_{a \in A} p_a W_1(p_{S_a}, \hat{p}_S).
\]
Since $\hat{p}_a W_1(p_{S_a}, \hat{p}_S) \leq \hat{p}_a W_1(\hat{p}_{S_a}, \hat{p}_S) + \hat{p}_a W_1(\hat{p}_{S_a}, p_{S_a})$, with probability $1 - \delta$:
\[
\sum_{a \in A} p_a W_1(p_{S_a}, p_S) \leq \sum_{a \in A} \hat{p}_a W_1(p_{S_a}, p_S) + \epsilon \frac{\epsilon}{2}
\]
\[
\leq \sum_{a \in A} \hat{p}_a W_1(\hat{p}_{S_a}, \hat{p}_S) + \hat{p}_a W_1(\hat{p}_{S_a}, p_{S_a}) + \epsilon \frac{\epsilon}{2}
\]
\[
\leq \sum_{a \in A} \hat{p}_a W_1(\hat{p}_{S_a}, \hat{p}_S) + \epsilon.
\]

The first inequality follows from Eq. (11), and the third one by Eq. (10). The result follows.

C  Inverse CDFs

Lemma 6. Given two differentiable and invertible cumulative distribution functions $f, g$ over the probability space $\Omega = [0, 1]$, thus $f, g : [0, 1] \rightarrow [0, 1]$, we have
\[
\int_{s=0}^{1} |f^{-1}(s) - g^{-1}(s)|ds = \int_{\tau=0}^{1} |f(\tau) - g(\tau)|d\tau.
\]  
(12)

Intuitively, we see that the left and right side of Eq. (12) correspond to two ways of computing the same shaded area in Figure 3. Here is a complete proof.

Proof. Invertible CDFs $f, g$ are strictly increasing functions due to being bijective and non-decreasing. Furthermore, we have $f(0) = 0, f(1) = 1$ by definition of CDFs and $\Omega = [0, 1]$, since $P(X \leq 0) = 0, P(X \leq 1) = 1$ where $X$ is the corresponding random variable. The same holds for the function $g$. Given an interval $(x_1, x_2) \subset [0, 1]$, let $y_1 = f(x_1), y_2 = f(x_2)$. Since $f$ is differentiable, we have
\[
\int_{x=x_1}^{x_2} f(x)dx + \int_{y=y_1}^{y_2} f^{-1}(y)dy = x_2y_2 - x_1y_1.
\]  
(13)
The proof of Eq. (13) is the following (see also Laisant (1905)).

\[
f^{-1}(f(x)) = x
\]

\[
\implies f'(x)f^{-1}(f(x)) = f'(x)x
\]

(multiply both sides by \( f'(x) \))

\[
\implies \int_{x=x_1}^{x_2} f'(x)f^{-1}(f(x))\,dx = \int_{x=x_1}^{x_2} f'(x)x\,dx
\]

(integrate both sides)

\[
\implies \int_{y=y_1}^{y_2} f^{-1}(y)\,dy = \int_{x=x_1}^{x_2} f'(x)x\,dx
\]

(apply change of variable \( y = f(x) \) on the left side)

\[
\implies \int_{y=y_1}^{y_2} f^{-1}(y)\,dy = x f(x)\bigg|_{x=x_1}^{x_2} - \int_{x=x_1}^{x_2} f(x)\,dx
\]

(integrate by parts on the right side)

\[
\implies \int_{y=y_1}^{y_2} f^{-1}(y)\,dy + \int_{x=x_1}^{x_2} f(x)\,dx = x_2 y_2 - x_1 y_1.
\]

Define a function \( h := f - g \) on \([0, 1]\). Then \( h \) is differentiable and thus continuous. Define the set of roots \( A := \{ x \in [0, 1] \mid h(x) = 0 \} \). Define the set of open intervals on which either \( h > 0 \) or \( h < 0 \) by \( B := \{(a, b) \mid b = \inf\{s \in A \mid a < s\}, 0 \leq a < b \leq 1, a \in A\} \). By continuity of \( h \), for any \((a, b) \in B\), we have \( b \in A \), i.e. \( b \) is also a root of \( h \). Since there are no other roots of \( h \) in \((a, b)\), by continuity of \( h \), we must have either \( h > 0 \) or \( h < 0 \) on \((a, b)\). For any two elements \((a, b), (c, d) \in B\), we argue that they must be disjoint intervals. Without loss of generality, we assume \( a < c \). Since \( b = \inf\{s \in A \mid a < s\} \leq c, i.e. b \leq c \), then \((a, b) \cap (c, d) = \emptyset \). For any open interval \((a, b) \in B\), there exists a rational number \( q \in \mathbb{Q} \) such that \( a < q < b \). We pick such a rational number and call it \( q(a, b) \). Since all elements of \( B \) are disjoint, for any two intervals \((a_0, b_0), (a_1, b_1) \) containing \( q(a_0, b_0), q(a_1, b_1) \in \mathbb{Q} \) respectively, we must have \( q(a_0, b_0) \neq q(a_1, b_1) \). We define the set \( Q_B := \{ q(a, b) \in \mathbb{Q} \mid (a, b) \in B \} \). Then \( Q_B \subset \mathbb{Q} \) and \(|Q_B| = |B| \). Since the set of rational numbers \( \mathbb{Q} \) is countable, the set \( B \) must also be countable. Let \( B = \{(a_i, b_i)\}_{i=0}^{N} \) where \( N \in \mathbb{N} \) or \( N = \infty \). Recall that \( h = f - g \) on \([0, 1]\), \( h(a_i) = 0 \), \( h(b_i) = 0 \) and either \( h < 0 \) or \( h > 0 \) on \((a_i, b_i) \) for all \( i > 0 \).

Consider the interval \((a_i, b_i)\) for some \( i > 0 \), by Eq.13 we have

\[
\int_{\tau=a_i}^{b_i} f(\tau)\,d\tau + \int_{s=f(a_i)}^{f(b_i)} f^{-1}(s)\,ds = b_i f(b_i) - a_i f(a_i)
\]

\[
= b_i g(b_i) - a_i g(a_i) = \int_{\tau=a_i}^{b_i} g(\tau)\,d\tau + \int_{s=g(a_i)}^{g(b_i)} g^{-1}(s)\,ds.
\]

Thus

\[
\int_{\tau=a_i}^{b_i} f(\tau) - g(\tau)\,d\tau = \int_{s=f(a_i)}^{f(b_i)} g^{-1}(s) - f^{-1}(s)\,ds.
\]

Notice that if \( f > g \) on \([a_i, b_i]\), then \( f^{-1} < g^{-1} \) on \([f(a_i), f(b_i)]\). This is due to the following. Given any \( y \in [f(a_i), f(b_i)] \), we have \( g^{-1}(y) \in [a_i, b_i] \) and \( f(g^{-1}(y)) = f(y) \). Thus \( g^{-1} > f^{-1} \) since \( f \) is strictly increasing. The contrary holds by the same reasoning, i.e. if \( f < g \) on \([a_i, b_i]\), then \( f^{-1} > g^{-1} \) on \([f(a_i), f(b_i)]\). Therefore,

\[
\int_{\tau=a_i}^{b_i} |f(\tau) - g(\tau)|\,d\tau = \int_{s=f(a_i)}^{f(b_i)} |g^{-1}(s) - f^{-1}(s)|\,ds,
\]

which holds for all intervals \((a_i, b_i)\). Summing over \( i \) on both sides, we have

\[
\sum_{i=0}^{N} \int_{\tau=a_i}^{b_i} |f(\tau) - g(\tau)|\,d\tau = \sum_{i=0}^{N} \int_{s=f(a_i)}^{f(b_i)} |g^{-1}(s) - f^{-1}(s)|\,ds,
\]

or equivalently,

\[
\int_{s=0}^{1} |f^{-1}(s) - g^{-1}(s)|\,ds = \int_{\tau=0}^{1} |f(\tau) - g(\tau)|\,d\tau.
\]