Appendix

A Empirical Estimates

Lemma 1. *As $|\mathcal{D}| \to \infty$, if $\mathcal{W}_1(p_S, p_{S_{\boldsymbol{a}}}) < \infty$ for all $\boldsymbol{a}$, the empirical barycenter satisfies $\lim \sum_{\boldsymbol{a}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{\bar{S}}, \hat{p}_{S_{\boldsymbol{a}}}) \to \sum_{\boldsymbol{a}} p_{\boldsymbol{a}} \mathcal{W}_1(p_{\bar{S}}, p_{S_{\boldsymbol{a}}})$ almost surely*[7].

Proof. By triangle inequality:

$$\sum_{\boldsymbol{a}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{\bar{S}}, p_{S_{\boldsymbol{a}}}) \leq \sum_{\boldsymbol{a}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{\bar{S}}, \hat{p}_{S_{\boldsymbol{a}}}) + \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(p_{S_{\boldsymbol{a}}}, \hat{p}_{S_{\boldsymbol{a}}}), \tag{4}$$

$$\sum_{\boldsymbol{a}} p_{\boldsymbol{a}} \mathcal{W}_1(p_{\bar{S}}, \hat{p}_{S_{\boldsymbol{a}}}) \leq \sum_{\boldsymbol{a}} p_{\boldsymbol{a}} \mathcal{W}_1(p_{\bar{S}}, p_{S_{\boldsymbol{a}}}) + p_{\boldsymbol{a}} \mathcal{W}_1(p_{S_{\boldsymbol{a}}}, \hat{p}_{S_{\boldsymbol{a}}}). \tag{5}$$

Since $p_{\bar{S}}$ and $\hat{p}_{\bar{S}}$ are the weighted barycenters of $\{p_{S_{\boldsymbol{a}}}\}$ and $\{\hat{p}_{S_{\boldsymbol{a}}}\}$ respectively:

$$\sum_{\boldsymbol{a}} p_{\boldsymbol{a}} \mathcal{W}_1(p_{\bar{S}}, p_{S_{\boldsymbol{a}}}) \leq \sum_{\boldsymbol{a}} p_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{\bar{S}}, p_{S_{\boldsymbol{a}}}), \tag{6}$$

$$\sum_{\boldsymbol{a}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{\bar{S}}, \hat{p}_{S_{\boldsymbol{a}}}) \leq \sum_{\boldsymbol{a}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(p_{\bar{S}}, \hat{p}_{S_{\boldsymbol{a}}}). \tag{7}$$

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

$$\begin{aligned}
\sum_{\boldsymbol{a}} p_{\boldsymbol{a}} \mathcal{W}_1(p_{\bar{S}}, p_{S_{\boldsymbol{a}}}) &\leq \sum_{\boldsymbol{a}} p_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{\bar{S}}, \hat{p}_{S_{\boldsymbol{a}}}) + p_{\boldsymbol{a}} \mathcal{W}_1(p_{S_{\boldsymbol{a}}}, \hat{p}_{S_{\boldsymbol{a}}}) \\
&\leq \sum_{\boldsymbol{a}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{\bar{S}}, \hat{p}_{S_{\boldsymbol{a}}}) + |\hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{\bar{S}}, \hat{p}_{S_{\boldsymbol{a}}}) - p_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{\bar{S}}, \hat{p}_{S_{\boldsymbol{a}}})| + p_{\boldsymbol{a}} \mathcal{W}_1(p_{S_{\boldsymbol{a}}}, \hat{p}_{S_{\boldsymbol{a}}}) \\
&\leq \sum_{\boldsymbol{a}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{\bar{S}}, \hat{p}_{S_{\boldsymbol{a}}}) + |\hat{p}_{\boldsymbol{a}} - p_{\boldsymbol{a}}| \cdot |\mathcal{W}_1(\hat{p}_{\bar{S}}, \hat{p}_{S_{\boldsymbol{a}}})| + p_{\boldsymbol{a}} \mathcal{W}_1(p_{S_{\boldsymbol{a}}}, \hat{p}_{S_{\boldsymbol{a}}}) \\
\sum_{\boldsymbol{a}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{\bar{S}}, \hat{p}_{S_{\boldsymbol{a}}}) &\leq \sum_{\boldsymbol{a}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(p_{\bar{S}}, p_{S_{\boldsymbol{a}}}) + \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(p_{S_{\boldsymbol{a}}}, \hat{p}_{S_{\boldsymbol{a}}}) \\
&\leq \sum_{\boldsymbol{a}} p_{\boldsymbol{a}} \mathcal{W}_1(p_{\bar{S}}, p_{S_{\boldsymbol{a}}}) + |p_{\boldsymbol{a}} \mathcal{W}_1(p_{\bar{S}}, p_{S_{\boldsymbol{a}}}) - \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(p_{\bar{S}}, p_{S_{\boldsymbol{a}}})| + \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(p_{S_{\boldsymbol{a}}}, \hat{p}_{S_{\boldsymbol{a}}}) \\
&\leq \sum_{\boldsymbol{a}} p_{\boldsymbol{a}} \mathcal{W}_1(p_{\bar{S}}, p_{S_{\boldsymbol{a}}}) + |p_{\boldsymbol{a}} - \hat{p}_{\boldsymbol{a}}| \cdot |\mathcal{W}_1(p_{\bar{S}}, p_{S_{\boldsymbol{a}}})| + \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(p_{S_{\boldsymbol{a}}}, \hat{p}_{S_{\boldsymbol{a}}}).
\end{aligned}$$

Therefore the following inequality holds almost surely:

$$\begin{aligned}
\Big| \sum_{\boldsymbol{a}} p_{\boldsymbol{a}} \mathcal{W}_1(p_{\bar{S}}, p_{S_{\boldsymbol{a}}}) - \sum_{\boldsymbol{a}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{\bar{S}}, \hat{p}_{S_{\boldsymbol{a}}}) \Big| &\leq \sum_{\boldsymbol{a}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(p_{S_{\boldsymbol{a}}}, \hat{p}_{S_{\boldsymbol{a}}}) + |p_{\boldsymbol{a}} - \hat{p}_{\boldsymbol{a}}| \cdot \mathcal{W}_1(p_{\bar{S}}, p_{S_{\boldsymbol{a}}}) \\
&\leq \sum_{\boldsymbol{a}} \mathcal{W}_1(p_{S_{\boldsymbol{a}}}, \hat{p}_{S_{\boldsymbol{a}}}) + |p_{\boldsymbol{a}} - \hat{p}_{\boldsymbol{a}}| \cdot \mathcal{W}_1(p_{\bar{S}}, p_{S_{\boldsymbol{a}}}) \\
&\leq \sum_{\boldsymbol{a}} \mathcal{W}_1(p_{S_{\boldsymbol{a}}}, \hat{p}_{S_{\boldsymbol{a}}}) + |p_{\boldsymbol{a}} - \hat{p}_{\boldsymbol{a}}| \cdot \mathcal{W}_1(p_S, p_{S_{\boldsymbol{a}}}).
\end{aligned}$$

Since $\mathcal{W}_1(p_{S_{\boldsymbol{a}}}, \hat{p}_{S_{\boldsymbol{a}}}) \to 0$ almost surely for all $\boldsymbol{a}$ (see Weed and Bach (2017)), and $\hat{p}_{\boldsymbol{a}} \to p_{\boldsymbol{a}}$ almost surely (by the strong law of large numbers) and $\mathcal{W}_1(p_S, p_{S_{\boldsymbol{a}}}) < \infty$ for all $\boldsymbol{a}$, the result follows:

$$\lim \sum_{\boldsymbol{a}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{\bar{S}}, \hat{p}_{S_{\boldsymbol{a}}}) \to \sum_{\boldsymbol{a}} p_{\boldsymbol{a}} \mathcal{W}_1(p_{\bar{S}}, p_{S_{\boldsymbol{a}}}),$$

almost surely. □

[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 $\mathcal{W}_1(p_{S_{\boldsymbol{a}}}, p_{\bar{S}}) \leq L$ for all $\boldsymbol{a} \in \mathcal{A}$. Let $P_S, P_{S_{\boldsymbol{a}}}$ and $P_{\bar{S}}$ be the cumulative density functions of S, $S_{\boldsymbol{a}}$ and $\bar{S}$. Assume these random variables all have domain $\Omega = [0, 1]$ and that all $P \in \{P_S, P_{\bar{S}}\} \cup \{P_{S_{\boldsymbol{a}}}\}_{\boldsymbol{a} \in \mathcal{A}}$ are continuous, then:

Lemma 5. *For any* $\epsilon, \delta > 0$, *if* $\min\left[\bar{N}, \min_{\boldsymbol{a}}\left[N_{\boldsymbol{a}}\right]\right] \geq \frac{16 \log(2|\mathcal{A}|/\delta)|\mathcal{A}|^2 \max[1,L]^2}{\epsilon^2}$, *with probability* $1 - \delta$*:*

$$\sum_{\boldsymbol{a} \in \mathcal{A}} p_{\boldsymbol{a}} \mathcal{W}_1(p_{S_{\boldsymbol{a}}}, p_{\bar{S}}) \leq \sum_{\boldsymbol{a} \in \mathcal{A}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{S_{\boldsymbol{a}}}, \hat{p}_{\bar{S}}) + \epsilon \,.$$

In other words, provided access to sufficient samples, a low value of $\sum_{\boldsymbol{a}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{S_{\boldsymbol{a}}}, \hat{p}_{\bar{S}})$ *implies a low value for* $\sum_{\boldsymbol{a}} p_{\boldsymbol{a}} \mathcal{W}_1(p_{S_{\boldsymbol{a}}}, p_{\bar{S}})$ *with high probability and therefore good performance at test time.*

Proof. We start with the case when $p_{\bar{S}} = p_S$. By the triangle inequality for Wasserstein-1 distances, for all $\boldsymbol{a} \in \mathcal{A}$:

$$\hat{p}_{\boldsymbol{a}} \mathcal{W}_1(p_{S_{\boldsymbol{a}}}, p_{\bar{S}}) \leq \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{S_{\boldsymbol{a}}}, \hat{p}_{\bar{S}}) + \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{\bar{S}}, p_{\bar{S}}) + \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{S_{\boldsymbol{a}}}, p_{S_{\boldsymbol{a}}}) \,. \tag{8}$$

Let $\hat{P}$ for $P \in \{P_S, P_{\bar{S}}\} \cup \{P_{S_{\boldsymbol{a}}}\}_{\boldsymbol{a} \in \mathcal{A}}$ denote the empirical CDF of P. Since their domain is restricted to $[0, 1]$ and are one dimensional random variables:

$$\mathcal{W}_1(\hat{p}_{S_*}, p_{S_*}) = \int_0^1 |\hat{P}(x) - P(x)| dx \tag{9}$$

For $S_* \in \{S, \bar{S}\} \cup \{S_{\boldsymbol{a}}\}_{\boldsymbol{a} \in \mathcal{A}}$. Since $P \in \{P_S, P_{\bar{S}}\} \cup \{P_{S_{\boldsymbol{a}}}\}_{\boldsymbol{a} \in \mathcal{A}}$ are all continuous, the Dvorestky-Kiefer-Wolfowitz theorem (see main theorem in Massart (1990)) and the condition $\min\left[\bar{N}, \min_{\boldsymbol{a}}\left[N_{\boldsymbol{a}}\right]\right] \geq \frac{16 \log(2|\mathcal{A}|/\delta)|\mathcal{A}|^2 \max[1,L]^2}{\epsilon^2}$ implies that:

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

Since all the random variables have domain $[0, 1]$ this in turn implies that for all $S_* \in \{S, \bar{S}\} \cup \{S_{\boldsymbol{a}}\}_{\boldsymbol{a} \in \mathcal{A}}$:

$$\mathbb{P}\left(\mathcal{W}_1(\hat{p}_{S_*}, p_{S_*}) \geq \frac{\epsilon}{4}\right) \leq \frac{\delta}{2|\mathcal{A}|}$$

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

$$\hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{\bar{S}}, p_{\bar{S}}) \leq \frac{\hat{p}_{\boldsymbol{a}} \epsilon}{4}, \quad \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{S_{\boldsymbol{a}}}, p_{S_{\boldsymbol{a}}}) \leq \frac{\hat{p}_{\boldsymbol{a}} \epsilon}{4} \,. \tag{10}$$

Summing Eq. (8) over $\boldsymbol{a}$ and applying the last observation yields

$$\sum_{\boldsymbol{a} \in \mathcal{A}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(p_{S_{\boldsymbol{a}}}, p_{\bar{S}}) \leq \sum_{\boldsymbol{a} \in \mathcal{A}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{S_{\boldsymbol{a}}}, \hat{p}_{\bar{S}}) + \frac{\epsilon}{2} \,.$$

Recall that we assume $\forall \boldsymbol{a} \in \mathcal{A}$,

$$\mathcal{W}_1(p_{S_{\boldsymbol{a}}}, p_{\bar{S}}) \leq L \,.$$

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

$$|p_{\boldsymbol{a}} - \hat{p}_{\boldsymbol{a}}| \leq \frac{\epsilon}{4|\mathcal{A}| \max[L, 1]} \,. \tag{11}$$

Consequently the desired result holds:

$$\sum_{\boldsymbol{a} \in \mathcal{A}} p_{\boldsymbol{a}} \mathcal{W}_1(p_{S_{\boldsymbol{a}}}, p_{\bar{S}}) \leq \sum_{\boldsymbol{a} \in \mathcal{A}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{S_{\boldsymbol{a}}}, \hat{p}_{\bar{S}}) + \epsilon \,.$$

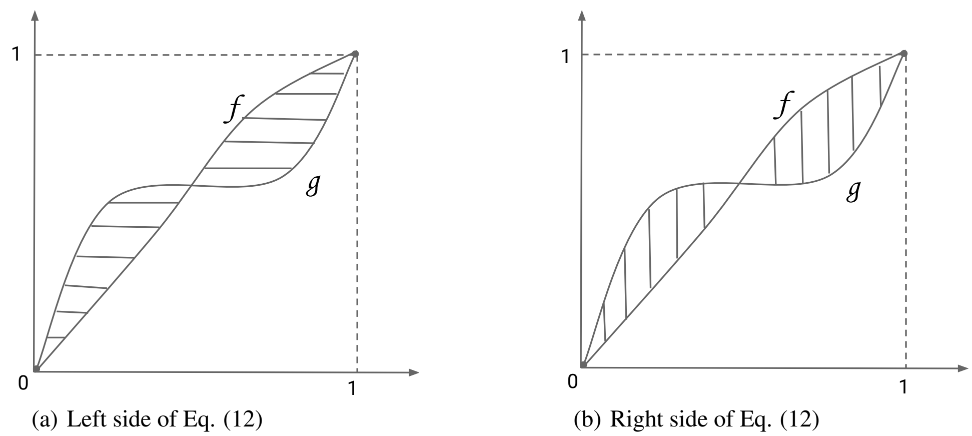

Figure 3: Integrating $|f^{-1} - g^{-1}|$ along the x axis (left) and integrating $|f - g|$ along the y axis (right) both compute the area of the same shaded region, thus the equality in Eq. (12).

If $p_{\bar{S}}$ equals the weighted barycenter of the population level distributions $\{p_{S_a}\}$, then

$$\sum_{\boldsymbol{a}\in\mathcal{A}} p_{\boldsymbol{a}} \mathcal{W}_1(p_{S_a}, p_{\bar{S}}) \leq \sum_{\boldsymbol{a}\in\mathcal{A}} p_{\boldsymbol{a}} \mathcal{W}_1(p_{S_a}, \hat{p}_{\bar{S}}) \,.$$

Since $\hat{p}_{\boldsymbol{a}}\mathcal{W}_1(p_{S_{\boldsymbol{a}}}, \hat{p}_{\bar{S}}) \leq \hat{p}_{\boldsymbol{a}}\mathcal{W}_1(\hat{p}_{S_{\boldsymbol{a}}}, \hat{p}_{\bar{S}}) + \hat{p}_{\boldsymbol{a}}\mathcal{W}_1(\hat{p}_{S_{\boldsymbol{a}}}, p_{S_a})$, with probability $1-\delta$:

$$\begin{aligned}\sum_{\boldsymbol{a}\in\mathcal{A}} p_{\boldsymbol{a}} \mathcal{W}_1(p_{S_a}, p_{\bar{S}}) &\leq \sum_{\boldsymbol{a}\in\mathcal{A}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(p_{S_a}, p_{\bar{S}}) + \frac{\epsilon}{2} \\ &\leq \sum_{\boldsymbol{a}\in\mathcal{A}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{S_{\boldsymbol{a}}}, \hat{p}_{\bar{S}}) + \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{S_{\boldsymbol{a}}}, p_{S_a}) + \frac{\epsilon}{2} \\ &\leq \sum_{\boldsymbol{a}\in\mathcal{A}} \hat{p}_{\boldsymbol{a}} \mathcal{W}_1(\hat{p}_{S_{\boldsymbol{a}}}, \hat{p}_{\bar{S}}) + \epsilon\end{aligned}$$

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 $f, g : [0, 1] \to [0, 1]$, we have*

$$\int_{s=0}^{1} |f^{-1}(s) - g^{-1}(s)| ds = \int_{\tau=0}^{1} |f(\tau) - g(\tau)| d\tau. \tag{12}$$

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

$$\int_{x=x_1}^{x_2} f(x) dx + \int_{y=y_1}^{y_2} f^{-1}(y) dy = x_2 y_2 - x_1 y_1. \tag{13}$$

The proof of Eq. (13) is the following (see also Laisant (1905)).

$$
\begin{aligned}
& f^{-1}(f(x)) = x \\
\Longrightarrow & f'(x)f^{-1}(f(x)) = f'(x)x & \text{(multiply both sides by } f'(x)\text{)} \\
\Longrightarrow & \int_{x=x_1}^{x_2} f'(x)f^{-1}(f(x))dx = \int_{x=x_1}^{x_2} f'(x)xdx & \text{(integrate both sides)} \\
\Longrightarrow & \int_{y=y_1}^{y_2} f^{-1}(y)dy = \int_{x=x_1}^{x_2} f'(x)xdx & \text{(apply change of variable } y = f(x) \text{ on the left side)} \\
\Longrightarrow & \int_{y=y_1}^{y_2} f^{-1}(y)dy = xf(x)\Big|_{x=x_1}^{x_2} - \int_{x=x_1}^{x_2} f(x)dx & \text{(integrate by parts on the right side)} \\
\Longrightarrow & \int_{y=y_1}^{y_2} f^{-1}(y)dy + \int_{x=x_1}^{x_2} f(x)dx = x_2y_2 - x_1y_1.
\end{aligned}
$$

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 $\forall i > 0$.

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

$$
\begin{aligned}
&\int_{\tau=a_i}^{b_i} f(\tau)d\tau + \int_{s=f(a_i)}^{f(b_i)} f^{-1}(s)ds = b_if(b_i) - a_if(a_i) \\
&= b_ig(b_i) - a_ig(a_i) = \int_{\tau=a_i}^{b_i} g(\tau)d\tau + \int_{s=g(a_i)}^{g(b_i)} g^{-1}(s)ds.
\end{aligned}
$$

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 $y \in [f(a_i), f(b_i)] = [g(a_i), g(b_i)]$, we have $g^{-1}(y) \in [a_i, b_i]$ and $f(g^{-1}(y)) > g(g^{-1}(y)) = y = f(f^{-1}(y))$. 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,

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

□