

A Organization of the supplementary

The organization of this supplementary document is as follows: In Section B, we lists definitions of the notations used in the proofs; Section C presents the theorems and propositions referenced in this study and their proofs; In Section D, the numerical experimental results related to the discussion in this study are provided; In Section E, the backpropagation algorithm referred to in Section 6 is presented. In addition to this material, the code used in the numerical experiments is also submitted as supplementary material.

B Notations

Table 1 lists the notations and definitions used in the proofs of Section C.

C Proofs

C.1 Proofs for Section 5

In this Section, we provide propositions and their proofs, as referred to in Section 5.

Lemma C.1. *A variational representation of α -divergence is given as*

$$D_\alpha(Q||P) = \sup_{\phi \geq 0} \left\{ \frac{1}{\alpha(1-\alpha)} - \frac{1}{\alpha} E_Q[\phi^{-\alpha}] - \frac{1}{1-\alpha} E_P[\phi^{1-\alpha}] \right\}, \quad (27)$$

where supremum is considered over all measurable functions with $E_P[\phi^{1-\alpha}] < \infty$ and $E_Q[\phi^{-\alpha}] < \infty$. The maximum value is achieved at $\phi = dQ/dP$.

proof of Lemma C.1. Let $f_\alpha(t) = \{t^{1-\alpha} - (1-\alpha) \cdot t - \alpha\} / \{\alpha(\alpha-1)\}$ for $\alpha \neq 0, 1$, then

$$\begin{aligned} E_P \left[f_\alpha \left(\frac{dQ}{dP} \right) \right] &= E_P \left[\frac{1}{\alpha(\alpha-1)} \left(\frac{dQ}{dP} \right)^{1-\alpha} + \frac{1}{\alpha} \left(\frac{dQ}{dP} \right) + \frac{1}{1-\alpha} \right] \\ &= \frac{1}{\alpha(\alpha-1)} E_P \left[\left(\frac{dQ}{dP} \right)^{1-\alpha} \right] + \frac{1}{\alpha} + \frac{1}{1-\alpha} \\ &= D_\alpha(Q||P). \end{aligned} \quad (28)$$

Note that, the Legendre transform for $g_\alpha(x) = x^{1-\alpha}/(1-\alpha)$ is obtained as

$$g_\alpha^*(x) = \frac{\alpha}{\alpha-1} x^{1-\frac{1}{\alpha}}. \quad (29)$$

In addition, note that, for the Legendre transforms of any fuction $h(x)$, it hold that

$$\{C \cdot h(x)\}^* = C \cdot h^* \left(\frac{x}{C} \right) \quad \text{and} \quad \{h(x) + C \cdot t + D\}^* = h^*(x - C) - D. \quad (30)$$

Here, A^* denotes the the Legendre transform of A .

From (29) and (30), we have

$$\begin{aligned} f_\alpha^*(t) &= \left\{ \frac{1}{(-\alpha)} g_\alpha(t) + \frac{1}{\alpha} t + \frac{1}{1-\alpha} \right\}^* \\ &= \frac{1}{(-\alpha)} g_\alpha^* \left(-\alpha \cdot \left\{ t - \frac{1}{\alpha} \right\} \right) - \frac{1}{1-\alpha} \\ &= -\frac{1}{\alpha} g_\alpha^*(1-\alpha t) + \frac{1}{\alpha-1} \\ &= -\frac{1}{\alpha} \left\{ \frac{\alpha}{\alpha-1} (1-\alpha t)^{1-\frac{1}{\alpha}} \right\} + \frac{1}{\alpha-1} \\ &= \frac{1}{1-\alpha} (1-\alpha t)^{1-\frac{1}{\alpha}} + \frac{1}{\alpha-1}. \end{aligned} \quad (31)$$

Table 1: Notations and definitions used in the proofs

Notations	Definitions, Meanings
$\mathbf{1}(\cdot)$	A propositional function: $\mathbf{1}(cond) = 1$ if $cond$ is true, and $\mathbf{1}(cond) = 0$ otherwise.
id_A	The identity function of a set A : $id_A(x) = 1$ if $x \in A$, and $id_A(x) = 0$ otherwise.
$\ \cdot\ $	the Euclidean norm.
$O_a(x)$	A term such that $\lim_{x \rightarrow 0} O(x)/x = C_a < \infty$, where C_a is a scalar value determined by a .
$O(x)$	A term such that $\lim_{x \rightarrow 0} O(x)/x < C$, where C is constant.
$f \lesssim g$	A relationship between two functions f and g such that $\limsup_{n \rightarrow \infty} f(n)/g(n) < \infty$.
$P \ll Q$	P is absolutely continuous with respect to Q .
P, Q	A pair of probability measures with $P \ll Q$ and $Q \ll P$.
μ	A probability measure with $P \ll \mu$ and $Q \ll \mu$.
$\frac{dP}{dQ}$	The Radon–Nikodým derivative of P with respect to Q . When $\frac{dQ}{d\mu}(\mathbf{x}) = 0$, this is defined as $\frac{dP}{dQ}(\mathbf{x}) = 0$.
\mathbf{X}	A random variable with a probability distribution μ .
$\mathbf{X}_{\sim P}$	A random variable obtained from \mathbf{X} by changing the probability distributions from μ to P : $P(\mathbf{X}_{\sim P} \leq \mathbf{x}) = \mu(\mathbf{X} \leq \mathbf{x})$, $\forall \mathbf{x} \in \mathbb{R}^d$. Intuitively, an observed value of \mathbf{X} in P .
$\mathbf{X}_{\sim Q}$	A random variable obtained from \mathbf{X} by changing the probability distributions from μ to Q : $P(\mathbf{X}_{\sim Q} \leq \mathbf{x}) = \mu(\mathbf{X} \leq \mathbf{x})$, $\forall \mathbf{x} \in \mathbb{R}^d$. Intuitively, an observed value of \mathbf{X} in Q .
$\mathbf{X}^{(N)}$	N i.i.d. random variables from μ : $\mathbf{X}^{(N)} = \{\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^N\}$, $\mathbf{X}^i \stackrel{\text{iid}}{\sim} \mu$.
$\mathbf{X}_P^{(N)}$	Random variables obtained from $\mathbf{X}^{(N)}$ by changing the probability distributions from μ to P : $\mathbf{X}_P^{(N)} = \{\mathbf{X}_{\sim P}^1, \mathbf{X}_{\sim P}^2, \dots, \mathbf{X}_{\sim P}^N\}$, $\mu(\mathbf{X}^i \leq \mathbf{x})$, $\forall \mathbf{x} \in \mathbb{R}^d$, $(1 \leq i \leq N)$. Intuitively, observed values of $\mathbf{X}^{(N)}$ in P .
$\mathbf{X}_Q^{(N)}$	Random variables obtained from $\mathbf{X}^{(N)}$ by changing the probability distributions from μ to Q : $\mathbf{X}_Q^{(N)} = \{\mathbf{X}_{\sim Q}^1, \mathbf{X}_{\sim Q}^2, \dots, \mathbf{X}_{\sim Q}^N\}$, $\mu(\mathbf{X}^i \leq \mathbf{x})$, $\forall \mathbf{x} \in \mathbb{R}^d$, $(1 \leq i \leq N)$. Intuitively, observed values of $\mathbf{X}^{(N)}$ in Q .
$\hat{P}^{(N)}$	The (empirical) distributions of $\mathbf{X}_P^{(N)}$: $\hat{P}^{(N)}(x) = \frac{1}{N} \sum_i \mathbf{1}(\mathbf{X}_{\sim P}^i = x)$.
$\hat{Q}^{(N)}$	The (empirical) distributions of $\mathbf{X}_Q^{(N)}$: $\hat{Q}^{(N)}(x) = \frac{1}{N} \sum_i \mathbf{1}(\mathbf{X}_{\sim Q}^i = x)$.
\mathcal{T}^α	The set of functions defined in Theorem 6.1.
$\mathbf{U} = \{U_1, \dots, U_m\}$	Unobserved random variables.
$\mathbf{V} = \{V_1, \dots, V_n\}$	Observed random variables.
$\mathcal{X}_\mathbf{A}$	The domain of variables \mathbf{A} .
$G = G_{\mathbf{V}\mathbf{U}}$	The causal graph for $\mathbf{V} \cup \mathbf{U}$.
$Pa(\mathbf{A})_G$	All the parents of the observed variables in G for for $\mathbf{A} \subset \mathbf{V}$.
$Ch(\mathbf{A})_G$	All the children of the observed variables in G for for $\mathbf{A} \subset \mathbf{V}$.
$An(\mathbf{A})_G$	All the ancestors of the observed variables in G for for $\mathbf{A} \subset \mathbf{V}$.
$De(\mathbf{A})_G$	All the descendants of the observed variables in G for for $\mathbf{A} \subset \mathbf{V}$.
W_p	The Wasserstein distance of order p .

433 By differentiating $f_\alpha(t)$, we obtain

$$f'_\alpha(t) = -\frac{1}{\alpha}t^{-\alpha} + \frac{1}{\alpha}. \quad (32)$$

434 Thus, we have

$$E_Q[f'_\alpha(\phi)] = E_Q\left[-\frac{1}{\alpha}\phi^{-\alpha} + \frac{1}{\alpha}\right]. \quad (33)$$

435 From (31) and (32), we obtain

$$\begin{aligned} E_P[f'_\alpha(f'_\alpha(\phi))] &= E_P\left[\frac{1}{1-\alpha}\left\{1-\alpha\cdot\left(-\frac{1}{\alpha}\phi^{-\alpha} + \frac{1}{\alpha}\right)\right\}^{1-\frac{1}{\alpha}} + \frac{1}{\alpha-1}\right] \\ &= E_P\left[\frac{1}{1-\alpha}\phi^{1-\alpha} + \frac{1}{\alpha-1}\right]. \end{aligned} \quad (34)$$

436 In additionm, from (33) and (34), we see for both $E_P[\phi^{1-\alpha}] < \infty$ and $E_Q[\phi^{-\alpha}] < \infty$ to hold is
437 equivalent for both $E_P[f'_\alpha(f'_\alpha(\phi))] < \infty$ and $E_Q[f'_\alpha(\phi)] < \infty$ to hold.

438 Finally, substituting (33) and (34) for (10), we have

$$\begin{aligned} D_\alpha(Q||P) &= \sup_{\phi \geq 0} \{E_Q[f'_\alpha(\phi)] - E_P[f'_\alpha(f'_\alpha(\phi))]\} \\ &= \sup_{\phi \geq 0} \left\{E_Q\left[-\frac{1}{\alpha}\phi^{-\alpha} + \frac{1}{\alpha}\right] - E_P\left[\frac{1}{1-\alpha}\phi^{1-\alpha} + \frac{1}{\alpha-1}\right]\right\} \\ &= \sup_{\phi \geq 0} \left\{\frac{1}{\alpha(1-\alpha)} - \frac{1}{\alpha}E_Q[\phi^{-\alpha}] - \frac{1}{1-\alpha}E_P[\phi^{1-\alpha}]\right\}. \end{aligned}$$

439 This completes the proof. \square

440 **Proposition C.2.** α -divergence can be written as follows:

$$D_\alpha(Q||P) = \sup_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \left\{ \frac{1}{\alpha(1-\alpha)} - \frac{1}{\alpha}E_Q[e^{\alpha \cdot T}] - \frac{1}{1-\alpha}E_P[e^{(\alpha-1) \cdot T}] \right\}, \quad (35)$$

441 where supremum is considered over all measurable functions $T: \mathbb{R}^d \rightarrow \mathbb{R}$ with $E_P[e^{(\alpha-1) \cdot T}] < \infty$
442 and $E_Q[e^{\alpha \cdot T}] < \infty$. The equality holds for T^* satisfying

$$\frac{dQ}{dP} = e^{-T^*}. \quad (36)$$

443 *proof of Proposition C.2.* Substituting e^{-T} for ϕ in (27), we have

$$\begin{aligned} D_\alpha(Q||P) &= \sup_{\phi \geq 0} \left\{ \frac{1}{\alpha(1-\alpha)} - \frac{1}{\alpha}E_Q[\phi^{-\alpha}] - \frac{1}{1-\alpha}E_P[\phi^{1-\alpha}] \right\} \\ &= \sup_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \left\{ \frac{1}{\alpha(1-\alpha)} - \frac{1}{\alpha}E_Q[\{e^{-T}\}^{-\alpha}] - \frac{1}{1-\alpha}E_P[\{e^{-T}\}^{1-\alpha}] \right\} \\ &= \sup_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \left\{ \frac{1}{\alpha(1-\alpha)} - \frac{1}{\alpha}E_Q[e^{\alpha \cdot T}] - \frac{1}{1-\alpha}E_P[e^{(\alpha-1) \cdot T}] \right\}. \end{aligned} \quad (37)$$

444 Finally, from Lemma C.1, the equality for (37) holds if and only if

$$\frac{dQ}{dP} = e^{-T^*}. \quad (38)$$

445 This completes the proof. \square

446 **Proposition C.3.** For $T \in \mathcal{T}^\alpha$, let

$$l_\alpha(\mathbf{X}_{\sim Q}, \mathbf{X}_{\sim P}; T) = \frac{1}{\alpha}e^{\alpha \cdot T(\mathbf{X}_{\sim Q})} + \frac{1}{1-\alpha}e^{(\alpha-1) \cdot T(\mathbf{X}_{\sim P})}. \quad (39)$$

447 Then the optimal function T^* for $\inf_{T: \mathbb{R}^d \rightarrow \mathbb{R}} l_\alpha(\mathbf{X}_{\sim Q}, \mathbf{X}_{\sim P}; T)$ is obtained as $T^* = -\log dQ/dP$,
448 μ -almost everywhere.

449 *proof of Proposition C.3.* From the definition for $\mathbf{X}_{\sim Q}$ and $\mathbf{X}_{\sim P}$, we see

$$e^{\alpha \cdot T(\mathbf{X}_{\sim Q}=\mathbf{x})} = e^{\alpha \cdot T(\mathbf{X}=\mathbf{x})} \cdot \frac{dQ}{d\mu}(\mathbf{x}),$$

450 and

$$e^{(\alpha-1) \cdot T(\mathbf{X}_{\sim P}=\mathbf{x})} = e^{(\alpha-1) \cdot T(\mathbf{X}=\mathbf{x})} \cdot \frac{dP}{d\mu}(\mathbf{x}).$$

451 Subsequently, we obtain

$$l_\alpha(\mathbf{X}_{\sim Q} = \mathbf{x}, \mathbf{X}_{\sim P} = \mathbf{x}; T) = \frac{1}{\alpha} \cdot e^{\alpha \cdot T(\mathbf{X}=\mathbf{x})} \cdot \frac{dQ}{d\mu}(\mathbf{x}) + \frac{1}{1-\alpha} \cdot e^{(\alpha-1) \cdot T(\mathbf{X}=\mathbf{x})} \cdot \frac{dP}{d\mu}(\mathbf{x}).$$

452 Note that, from Jensen's inequality, it holds that

$$\log(p \cdot X + q \cdot Y) \geq p \cdot \log(X) + q \cdot \log(Y), \quad (40)$$

453 for $X, Y > 0$ and $p, q > 0$ with $p + q = 1$, and the equality holds when $X = Y$.

454 From this equation, by letting $X = e^{\alpha \cdot T(\mathbf{X}=\mathbf{x})} \cdot \frac{dQ}{d\mu}(\mathbf{x})$, $Y = e^{(\alpha-1) \cdot T(\mathbf{X}=\mathbf{x})} \cdot \frac{dP}{d\mu}(\mathbf{x})$, $p = 1 - \alpha$ and
455 $q = \alpha$, we observe that

$$\log(p \cdot X + q \cdot Y) = \log\left(\frac{1}{\alpha \cdot (1-\alpha)} \cdot l_\alpha(\mathbf{X}_{\sim Q} = \mathbf{x}, \mathbf{X}_{\sim P} = \mathbf{x}; T)\right),$$

456 and $\log\left(\frac{1}{\alpha \cdot (1-\alpha)} \cdot l_\alpha(\mathbf{X}_{\sim Q} = \mathbf{x}, \mathbf{X}_{\sim P} = \mathbf{x}; T)\right)$ is minimized when $e^{\alpha \cdot T(\mathbf{X}=\mathbf{x})} \cdot \frac{dQ}{d\mu}(\mathbf{x}) =$
457 $e^{(\alpha-1) \cdot T(\mathbf{X}=\mathbf{x})} \cdot \frac{dP}{d\mu}(\mathbf{x})$, μ -almost everywhere.

458 Then, we see that $\inf_{T: \mathbb{R}^d \rightarrow \mathbb{R}} l_\alpha(\mathbf{X}_{\sim Q} = \mathbf{x}, \mathbf{X}_{\sim P} = \mathbf{x}; T)$ is achieved at $e^{-T^*}(\mathbf{x}) = \frac{dQ}{dP}$, μ -almost
459 everywhere. Hence, we have $T^* = -\log dQ/dP$, μ -almost everywhere.

460 □

461 **Proposition C.4.** For $T \in \mathcal{T}^\alpha$, let $T^{+k} = T + k$. Then the optimal function T^* for
462 $\inf_{k \in \mathbb{R}} l_\alpha(\mathbf{X}_{\sim Q}, \mathbf{X}_{\sim P}; T^{+k})$ is satisfying that $E_P[e^{-T^*}] = 1$, where $l_\alpha(\mathbf{X}_{\sim Q}, \mathbf{X}_{\sim P}; T)$ is de-
463 fined as (39).

464 *proof of Proposition C.4.* From the definition for $\mathbf{X}_{\sim Q}$ and $\mathbf{X}_{\sim P}$, we see

$$e^{\alpha \cdot T^{+k}(\mathbf{X}_{\sim Q}=\mathbf{x})} = e^{\alpha \cdot T^{+k}(\mathbf{X}=\mathbf{x})} \cdot \frac{dQ}{d\mu}(\mathbf{x}) = e^{\alpha \cdot k} \cdot e^{\alpha \cdot T(\mathbf{X}=\mathbf{x})} \cdot \frac{dQ}{d\mu}(\mathbf{x}),$$

465 and

$$e^{(\alpha-1) \cdot T(\mathbf{X}_{\sim P}=\mathbf{x})} = e^{(\alpha-1) \cdot T^{+k}(\mathbf{X}=\mathbf{x})} \cdot \frac{dP}{d\mu}(\mathbf{x}) = e^{(\alpha-1) \cdot k} \cdot e^{(\alpha-1) \cdot T(\mathbf{X}=\mathbf{x})} \cdot \frac{dP}{d\mu}(\mathbf{x}).$$

466 Subsequently, we obtain

$$l_\alpha(\mathbf{X}_{\sim Q} = \mathbf{x}, \mathbf{X}_{\sim P} = \mathbf{x}; T) = e^{\alpha \cdot k} \cdot e^{\alpha \cdot T(\mathbf{X}=\mathbf{x})} \cdot \frac{dQ}{d\mu}(\mathbf{x}) + e^{\alpha \cdot k} \cdot e^{\alpha \cdot T(\mathbf{X}=\mathbf{x})} \cdot \frac{dQ}{d\mu}(\mathbf{x}).$$

467 For Jensen's inequality (40), let $X = e^{\alpha \cdot k} \cdot e^{\alpha \cdot T(\mathbf{X}=\mathbf{x})} \cdot \frac{dQ}{d\mu}(\mathbf{x})$, $Y = e^{(\alpha-1) \cdot k} \cdot e^{(\alpha-1) \cdot T(\mathbf{X}=\mathbf{x})} \cdot \frac{dP}{d\mu}(\mathbf{x})$,
468 $p = 1 - \alpha$ and $q = \alpha$. Then, $l_\alpha(\mathbf{X}_{\sim Q}, \mathbf{X}_{\sim P}; T^{+k})$ is minimized when $e^{\alpha \cdot k_*} \cdot e^{\alpha \cdot T(\mathbf{X}=\mathbf{x})} \cdot \frac{dQ}{d\mu}(\mathbf{x}) =$
469 $e^{(\alpha-1) \cdot k_*} \cdot e^{(\alpha-1) \cdot T(\mathbf{X}=\mathbf{x})} \cdot \frac{dP}{d\mu}(\mathbf{x})$, μ -almost everywhere.

470 Then, we see

$$e^{-k_*} \cdot \frac{dQ}{d\mu}(\mathbf{x}) = e^{-T(\mathbf{x})} \cdot \frac{dP}{d\mu}(\mathbf{x}).$$

471 By integrating both sides of the above equality over \mathbb{R}^d with μ , we obtain

$$e^{-k_*} = E_P[e^{-T}].$$

472 From this, we have

$$e^{k_*} \cdot E_P [e^{-T}] = 1.$$

473 Now, since $T_* = T + k_*$, we see

$$E_P [e^{-T_*}] = E_P [e^{-T+k_*}] = e^{k_*} \cdot E_P [e^{-T}] = 1.$$

474 This completes the proof. \square

475 **Proposition C.5.** For a fixed point $\mathbf{x}_0 \in \mathbb{R}^d$, suppose that $\frac{dQ}{d\mu}(\mathbf{x}_0) > 0$ and $\frac{dP}{d\mu}(\mathbf{x}_0) > 0$. For a
 476 constant $L > 0$, let I_L denote an interval $[-L - \log \frac{dQ}{dP}(\mathbf{x}_0), L - \log \frac{dQ}{dP}(\mathbf{x}_0)]$. Subsequently, let
 477 $f : I_L \rightarrow \mathbb{R}$ be a function as follows:

$$f(t) = \frac{1}{\alpha} \cdot e^{\alpha \cdot t} \cdot \frac{dQ}{d\mu}(\mathbf{x}_0) + \frac{1}{1-\alpha} \cdot e^{(\alpha-1) \cdot t} \cdot \frac{dP}{d\mu}(\mathbf{x}_0). \quad (41)$$

478 Let $\frac{\lambda}{2} = e^{-L} \cdot \left\{ \frac{dQ}{d\mu}(\mathbf{x}_0) \right\}^{1-\alpha} \left\{ \frac{dP}{d\mu}(\mathbf{x}_0) \right\}^{\alpha}$. Then, $f(t)$ satisfies $f''(t) \cdot t^2 \geq \frac{\lambda}{2} \cdot t^2$ for all $t \in I_L$,
 479 that is, $f(t)$ is λ -strongly convex. In addition, $|f'(t)| \leq 2 \cdot e^L \cdot \left\{ \frac{dQ}{d\mu}(\mathbf{x}_0) \right\}^{1-\alpha} \left\{ \frac{dP}{d\mu}(\mathbf{x}_0) \right\}^{\alpha}$ holds
 480 for all $t \in I_L$, and $f(t)$ is minimized at $t_* = -\log \frac{dQ}{dP}(\mathbf{x}_0)$.

481 *proof of Proposition C.5.* By repeating the derivative of $f(t)$, we obtain

$$f'(t) = e^{\alpha \cdot t} \cdot \frac{dQ}{d\mu}(\mathbf{x}) - e^{(\alpha-1) \cdot t} \cdot \frac{dP}{d\mu}(\mathbf{x}), \quad (42)$$

482 and

$$f''(t) = \alpha \cdot e^{\alpha \cdot t} \cdot \frac{dQ}{d\mu}(\mathbf{x}) + (1-\alpha) \cdot e^{(\alpha-1) \cdot t} \cdot \frac{dP}{d\mu}(\mathbf{x}). \quad (43)$$

483 First, we see that $f''(t) \geq \frac{\lambda}{2}$ holds for all $t \in I_L$. From (43), we have

$$\begin{aligned} f''(t) &\geq \alpha \cdot e^{\alpha \cdot (t_* - L)} \cdot \frac{dQ}{d\mu}(\mathbf{x}) + (1-\alpha) \cdot e^{(\alpha-1) \cdot (t_* + L)} \cdot \frac{dP}{d\mu}(\mathbf{x}) \\ &\geq \alpha \cdot e^{-\alpha \cdot L} \cdot e^{\alpha \cdot t_*} \cdot \frac{dQ}{d\mu}(\mathbf{x}) + (1-\alpha) \cdot e^{(\alpha-1) \cdot L} \cdot e^{(\alpha-1) \cdot t_*} \cdot \frac{dP}{d\mu}(\mathbf{x}) \\ &\geq \alpha \cdot e^{-L} \cdot e^{\alpha \cdot t_*} \cdot \frac{dQ}{d\mu}(\mathbf{x}) + (1-\alpha) \cdot e^{-L} \cdot e^{(\alpha-1) \cdot t_*} \cdot \frac{dP}{d\mu}(\mathbf{x}) \\ &= e^{-L} \cdot \left\{ \alpha \cdot e^{\alpha \cdot t_*} \cdot \frac{dQ}{d\mu}(\mathbf{x}) + (1-\alpha) \cdot e^{(\alpha-1) \cdot t_*} \cdot \frac{dP}{d\mu}(\mathbf{x}) \right\}. \end{aligned} \quad (44)$$

484 Note that, from (40), we see

$$p \cdot X + q \cdot Y \geq p \cdot X^p \cdot Y^q, \quad (45)$$

485 for $X, Y > 0$ and $p, q > 0$ with $p + q = 1$, and the equality holds when $X = Y$. By letting
 486 $X = e^{\alpha \cdot t_*} \cdot \frac{dQ}{d\mu}(\mathbf{x})$, $Y = e^{(\alpha-1) \cdot t_*} \cdot \frac{dP}{d\mu}(\mathbf{x})$, $p = \alpha$, and $q = 1 - \alpha$ in the above equality, we obtain

$$\begin{aligned} \alpha \cdot e^{\alpha \cdot t_*} \cdot \frac{dQ}{d\mu}(\mathbf{x}) + (1-\alpha) \cdot e^{(\alpha-1) \cdot t_*} \cdot \frac{dP}{d\mu}(\mathbf{x}) &\geq \left\{ e^{\alpha \cdot t_*} \cdot \frac{dQ}{d\mu}(\mathbf{x}) \right\}^{\alpha} \left\{ e^{(\alpha-1) \cdot t_*} \cdot \frac{dP}{d\mu}(\mathbf{x}) \right\}^{1-\alpha} \\ &= e^{\{\alpha^2 - (1-\alpha)^2\} \cdot t_*} \left\{ \frac{dQ}{d\mu}(\mathbf{x}) \right\}^{\alpha} \left\{ \frac{dP}{d\mu}(\mathbf{x}) \right\}^{1-\alpha} \\ &= \left\{ \frac{dQ}{dP}(\mathbf{x}) \right\}^{1-2\alpha} \left\{ \frac{dQ}{d\mu}(\mathbf{x}) \right\}^{\alpha} \left\{ \frac{dP}{d\mu}(\mathbf{x}) \right\}^{1-\alpha} \\ &= \left\{ \frac{dQ}{d\mu}(\mathbf{x}) \right\}^{1-\alpha} \left\{ \frac{dP}{d\mu}(\mathbf{x}) \right\}^{\alpha}. \end{aligned} \quad (46)$$

487 Thus, from (44) and (46), we see $f''(t) \geq \frac{\lambda}{2}$ holds for all $t \in I_L$.

488 Next, we obtain $|f'(t)| \leq 2e^L \left\{ \frac{dQ}{d\mu}(\mathbf{x}_0) \right\}^{1-\alpha} \left\{ \frac{dP}{d\mu}(\mathbf{x}_0) \right\}^\alpha$. From (42), we have

$$\begin{aligned}
|f'(t)| &\leq e^{\alpha \cdot (t_* + L)} \cdot \frac{dQ}{d\mu}(\mathbf{x}) + e^{(\alpha-1) \cdot (t_* - L)} \cdot \frac{dP}{d\mu}(\mathbf{x}) \\
&= e^{\alpha \cdot t_*} \cdot e^{\alpha \cdot L} \cdot \frac{dQ}{d\mu}(\mathbf{x}) + e^{(\alpha-1) \cdot t_*} \cdot e^{(1-\alpha) \cdot L} \cdot \frac{dP}{d\mu}(\mathbf{x}) \\
&\leq e^L \left\{ e^{\alpha \cdot t_*} \cdot \frac{dQ}{d\mu}(\mathbf{x}) + e^{(\alpha-1) \cdot t_*} \cdot \frac{dP}{d\mu}(\mathbf{x}) \right\} \\
&= e^L \left[\left\{ \frac{dQ}{dP}(\mathbf{x}) \right\}^{-\alpha} \frac{dQ}{d\mu}(\mathbf{x}) + \left\{ \frac{dQ}{dP}(\mathbf{x}) \right\}^{1-\alpha} \frac{dP}{d\mu}(\mathbf{x}) \right] \\
&= e^L \left[\left\{ \frac{dQ}{d\mu}(\mathbf{x}) \right\}^{1-\alpha} \left\{ \frac{dP}{d\mu}(\mathbf{x}) \right\}^\alpha + \left\{ \frac{dQ}{d\mu}(\mathbf{x}) \right\}^{1-\alpha} \left\{ \frac{dP}{d\mu}(\mathbf{x}) \right\}^\alpha \right]. \quad (47)
\end{aligned}$$

489 Here, we see $f'(t) \leq 2 \cdot e^L \left\{ \frac{dQ}{d\mu}(\mathbf{x}_0) \right\}^{1-\alpha} \left\{ \frac{dP}{d\mu}(\mathbf{x}_0) \right\}^\alpha$.

490 The rest of the proposition statement follows from Lemma C.3.

491 This completes the proof. \square

492 **Corollary C.6.** For N fixed points $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^d$, suppose that $\frac{dQ}{d\mu}(\mathbf{x}_i) > 0$ and $\frac{dP}{d\mu}(\mathbf{x}_i) > 0$
493 ($1 \leq i \leq N$). For a constant $L > 0$, let I_L^i denote an interval $[-L - \log \frac{dQ}{dP}(\mathbf{x}_i), L - \log \frac{dQ}{dP}(\mathbf{x}_i)]$.
494 Subsequently, let $f^{(N)} : I_L^1 \times I_L^2 \times \dots \times I_L^N \rightarrow \mathbb{R}$ be a function as follows:

$$f^{(N)}(\mathbf{t}) = f^{(N)}(t_1, t_2, \dots, t_N) = \frac{1}{\alpha} \frac{1}{N} \sum_{i=1}^N e^{\alpha \cdot t_i} \cdot \frac{dQ}{d\mu}(\mathbf{x}_i) + \frac{1}{1-\alpha} \frac{1}{N} \sum_{i=1}^N e^{(\alpha-1) \cdot t_i} \cdot \frac{dP}{d\mu}(\mathbf{x}_i), \quad (48)$$

495 and let

$$\frac{\lambda}{2} = \frac{1}{N} \cdot \min_{1 \leq i \leq N} \left\{ e^{-L} \cdot \left\{ \frac{dQ}{d\mu}(\mathbf{x}_i) \right\}^{1-\alpha} \left\{ \frac{dP}{d\mu}(\mathbf{x}_i) \right\}^\alpha \right\}. \quad (49)$$

496 Then, $f^{(N)}(\mathbf{t})$ satisfies

$$\mathbf{t}^T \cdot \nabla^2 f^{(N)}(\mathbf{t}) \cdot \mathbf{t} = \sum_{1 \leq i, j \leq N} t_i \cdot t_j \cdot \frac{\partial^2 f^{(N)}}{\partial t_i \partial t_j} \geq \frac{\lambda}{2} \cdot \|\mathbf{t}\|^2, \quad (50)$$

497 that is, $f^{(N)}(\mathbf{t})$ is λ -strongly convex.

498 In addition, let $D_i = 2 \cdot e^L \cdot \left\{ \frac{dQ}{d\mu}(\mathbf{x}_i) \right\}^{1-\alpha} \left\{ \frac{dP}{d\mu}(\mathbf{x}_i) \right\}^\alpha$, and let $D = \max \{D_1, D_2, \dots, D_N\}^2$.

499 Then,

$$\left\| \nabla f^{(N)}(\mathbf{t}) \right\|^2 = \left\| \left(\frac{\partial}{\partial t_1} f^{(N)}(\mathbf{t}), \frac{\partial}{\partial t_2} f^{(N)}(\mathbf{t}), \dots, \frac{\partial}{\partial t_N} f^{(N)}(\mathbf{t}) \right) \right\|^2 \leq D^2, \quad (51)$$

500 for all $\mathbf{t} \in I_L^1 \times I_L^2 \times \dots \times I_L^N$, and $f^{(N)}(\mathbf{t})$ is minimized at $\mathbf{t}_* = (t_*^1, t_*^2, \dots, t_*^N) = (-\log \frac{dQ}{dP}(\mathbf{x}_1),$
501 $-\log \frac{dQ}{dP}(\mathbf{x}_2), \dots, -\log \frac{dQ}{dP}(\mathbf{x}_N))$.

502 *proof of Corollary C.6.* Let

$$f_i(t) = \frac{1}{\alpha} e^{\alpha \cdot t} \cdot \frac{dQ}{d\mu}(\mathbf{x}_i) + \frac{1}{1-\alpha} e^{(\alpha-1) \cdot t} \cdot \frac{dP}{d\mu}(\mathbf{x}_i), \quad (52)$$

503 and let $\frac{\lambda_i}{2} = e^{-L} \cdot \left\{ \frac{dQ}{d\mu}(\mathbf{x}_i) \right\}^{1-\alpha} \left\{ \frac{dP}{d\mu}(\mathbf{x}_i) \right\}^\alpha$. From Proposition C.5, for each $1 \leq i \leq N$,
504 $f_i(t)$ satisfies that $f_i''(t) \geq \frac{\lambda_i}{2} \cdot t$ and $f_i'(t) \leq D_i$ holds for all $t \in I_L^i$, and $f_i(t)$ is minimized at
505 $t_*^i = -\log \frac{dQ}{dP}(\mathbf{x}_i)$.

506 Note that,

$$\begin{aligned} \nabla^2 f^{(N)}(\mathbf{t}) &= \begin{pmatrix} \frac{\partial^2}{\partial t_1^2} f^{(N)}(\mathbf{t}) & \frac{\partial^2}{\partial t_1 \partial t_2} f^{(N)}(\mathbf{t}) & \cdots & \frac{\partial^2}{\partial t_1 \partial t_N} f^{(N)}(\mathbf{t}) \\ \frac{\partial^2}{\partial t_2 \partial t_1} f^{(N)}(\mathbf{t}) & \frac{\partial^2}{\partial t_2^2} f^{(N)}(\mathbf{t}) & & \\ \vdots & & \ddots & \\ \frac{\partial^2}{\partial t_N \partial t_j} f^{(N)}(\mathbf{t}) & \frac{\partial^2}{\partial t_i \partial t_j} f^{(N)}(\mathbf{t}) & \cdots & \frac{\partial^2}{\partial t_N \partial t_N} f^{(N)}(\mathbf{t}) \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{N} \cdot f_1''(t_1) & & & \\ & \frac{1}{N} \cdot f_2''(t_2) & & \mathbf{0} \\ & & \ddots & \\ & \mathbf{0} & & \ddots \\ & & & & \frac{1}{N} \cdot f_N''(t_N) \end{pmatrix}. \end{aligned} \quad (53)$$

507 From this, we see

$$\begin{aligned} \mathbf{t}^T \cdot \nabla^2 f^{(N)}(\mathbf{t}) \cdot \mathbf{t} &= \frac{1}{N} \sum_{i=1}^N f_i''(t_i) \cdot t_i^2 \\ &\geq \frac{1}{N} \sum_{i=1}^N \frac{\lambda_i}{2} \cdot t_i^2 \\ &= \sum_{i=1}^N \frac{1}{N} \cdot \frac{\lambda_i}{2} \cdot t_i^2 \\ &\geq \sum_{i=1}^N \frac{\lambda}{2} \cdot t_i^2 \\ &= \frac{\lambda}{2} \cdot \|\mathbf{t}\|^2. \end{aligned} \quad (54)$$

508 In addition, since $f^{(N)}(t) = \frac{1}{N} \sum_{i=1}^N f_i(t_i)$, we have

$$\begin{aligned} \|\nabla f^{(N)}(\mathbf{t})\|^2 &= \left\| \left(\frac{\partial}{\partial t_1} f^{(N)}(\mathbf{t}), \frac{\partial}{\partial t_2} f^{(N)}(\mathbf{t}), \dots, \frac{\partial}{\partial t_N} f^{(N)}(\mathbf{t}) \right) \right\|^2, \\ &= \left\| \left(\frac{1}{N} \cdot f_1'(t_1), \frac{1}{N} \cdot f_2'(t_2), \dots, \frac{1}{N} \cdot f_N'(t_N) \right) \right\|^2 \\ &\leq \left\| \left(\frac{1}{N} \cdot D_1, \frac{1}{N} \cdot D_2, \dots, \frac{1}{N} \cdot D_N \right) \right\|^2 \\ &\leq \left\| \left(\frac{1}{N} \cdot D, \frac{1}{N} \cdot D, \dots, \frac{1}{N} \cdot D \right) \right\|^2 \\ &= D^2, \end{aligned} \quad (55)$$

509 and $f^{(N)}(\mathbf{t})$ is minimized at $\mathbf{t}_* = (t_*^1, t_*^2, \dots, t_*^N) = (-\log \frac{dQ}{dP}(\mathbf{x}_1), -\log \frac{dQ}{dP}(\mathbf{x}_2), \dots,$
 510 $-\log \frac{dQ}{dP}(\mathbf{x}_N))$.

511 This completes the proof. □

512 **Lemma C.7.** For $T \in \mathcal{T}^\alpha$, let

$$l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T) = \frac{1}{\alpha} \cdot e^{\alpha \cdot T(\mathbf{X}_{\sim Q}^i)} + \frac{1}{1-\alpha} \cdot e^{(\alpha-1) \cdot T(\mathbf{X}_{\sim P}^i)}, \quad (56)$$

$$L_\alpha(Q, P; T) = \frac{1}{\alpha} \cdot E_Q \left[e^{\alpha \cdot T(\mathbf{X})} \right] + \frac{1}{1-\alpha} \cdot E_P \left[e^{(\alpha-1) \cdot T(\mathbf{X})} \right], \quad (57)$$

513 and let

$$l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i) = \inf_{T: \mathbb{R}^d \rightarrow \mathbb{R}} l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T), \quad (58)$$

$$L_\alpha(Q, P) = \inf_{T: \mathbb{R}^d \rightarrow \mathbb{R}} L_\alpha(Q, P; T), \quad (59)$$

514 where the infimums of (58) and (59) are considered over all measurable functions with $T : \mathbb{R}^d \rightarrow \mathbb{R}$
 515 with $E_P[e^{(\alpha-1) \cdot T}] < \infty$ and $E_Q[e^{\alpha \cdot T}] < \infty$.

516 In addition, let

$$\hat{L}_\alpha^{(N)}(Q, P; T) = \frac{1}{\alpha} \frac{1}{N} \sum_{i=1}^N e^{\alpha \cdot T(\mathbf{X}_{\sim Q}^i)} + \frac{1}{1-\alpha} \frac{1}{N} \sum_{i=1}^N e^{(\alpha-1) \cdot T(\mathbf{X}_{\sim P}^i)}, \quad (60)$$

$$\hat{L}_\alpha^{(N)}(Q, P) = \inf_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \hat{L}_\alpha^{(N)}(Q, P; T). \quad (61)$$

517 Then, it holds that

$$E_\mu \left[\hat{L}_\alpha^{(N)}(Q, P; T) \right] = E_\mu \left[l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T) \right] = L_\alpha(Q, P; T), \quad (62)$$

518

$$E_\mu \left[\hat{L}_\alpha^{(N)}(Q, P) \right] = E_\mu \left[l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i) \right] = L_\alpha(Q, P). \quad (63)$$

519 *proof of Lemma C.7.* We first show the last equality in (63) holds. Now, we consider the following
 520 integral:

$$\begin{aligned} E \left[\frac{1}{\alpha(1-\alpha)} - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i) \right] &= \int \left\{ \frac{1}{\alpha(1-\alpha)} - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i) \right\} d\mu \\ &= \int \left\{ \frac{1}{\alpha(1-\alpha)} - \inf_{T: \mathbb{R}^d \rightarrow \mathbb{R}} l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T) \right\} d\mu \\ &= \int \sup_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \left\{ \frac{1}{\alpha(1-\alpha)} - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T) \right\} d\mu. \end{aligned} \quad (64)$$

521 Let T^* be the optimal function for (35) in Lemma C.2. Let $T_k = -\log dQ/dP + 1/k$, then from
 522 Proposition C.3, we have

$$\lim_{k \rightarrow \infty} l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_k) = \inf_{T: \mathbb{R}^d \rightarrow \mathbb{R}} l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T). \quad (65)$$

523 From this, we obtain

$$\begin{aligned} \lim_{k \rightarrow \infty} E \left[\frac{1}{\alpha(1-\alpha)} - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_k) \right] &= \frac{1}{\alpha(1-\alpha)} - \lim_{k \rightarrow \infty} E \left[l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_k) \right] \\ &= \frac{1}{\alpha(1-\alpha)} - \inf_{T: \mathbb{R}^d \rightarrow \mathbb{R}} l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T) \\ &= \sup_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \left\{ E \left[\frac{1}{\alpha(1-\alpha)} - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T) \right] \right\}. \end{aligned} \quad (66)$$

524 Now, from Lemma C.2, we see

$$\begin{aligned} \left| \frac{1}{\alpha(1-\alpha)} - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i) \right| &= \left| \sup_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \left\{ \frac{1}{\alpha(1-\alpha)} - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T) \right\} \right| \\ &= \frac{1}{\alpha(1-\alpha)} - \frac{1}{\alpha} e^{\alpha \cdot T^*(\mathbf{X}_{\sim Q}^i)} - \frac{1}{1-\alpha} e^{(\alpha-1) \cdot T^*(\mathbf{X}_{\sim P}^i)} \\ &= \frac{1}{\alpha(1-\alpha)} - \frac{1}{\alpha} \left(\frac{dQ}{dP}(\mathbf{X}_{\sim Q}^i) \right)^\alpha - \frac{1}{1-\alpha} \left(\frac{dQ}{dP}(\mathbf{X}_{\sim P}^i) \right)^{\alpha-1}. \end{aligned} \quad (67)$$

525 Let $\phi(\mathbf{X})$ denote the term on the right hand side of (67). Then, we observe that

$$\left| \frac{1}{\alpha(1-\alpha)} - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T) \right| \leq \phi(\mathbf{X}) \quad \text{and} \quad E[\phi(\mathbf{X})] < \infty.$$

526 That is, we see that the following sequence is uniformly integrable for μ :

$$\left\{ \frac{1}{\alpha(1-\alpha)} - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_k) \right\}_{k=1}^N.$$

527 Thus, from the property of the Lebesgue integral (25, P188, Theorem 4), we obtain

$$E \left[\lim_{k \rightarrow \infty} \left\{ \frac{1}{\alpha(1-\alpha)} - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_k) \right\} \right] = \lim_{k \rightarrow \infty} E \left[\frac{1}{\alpha(1-\alpha)} - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_k) \right]. \quad (68)$$

528 Finally, from (66) and (68), we have

$$\begin{aligned} \frac{1}{\alpha(1-\alpha)} - E[l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i)] &= E \left[\frac{1}{\alpha(1-\alpha)} - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i) \right] \\ &= E \left[\sup_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \left\{ \frac{1}{\alpha(1-\alpha)} - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T) \right\} \right] \\ &= E \left[\lim_{k \rightarrow \infty} \left\{ \frac{1}{\alpha(1-\alpha)} - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_k) \right\} \right] \\ &= \lim_{k \rightarrow \infty} E \left[\frac{1}{\alpha(1-\alpha)} - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_k) \right] \quad \therefore (68) \\ &= \sup_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \left\{ E \left[\frac{1}{\alpha(1-\alpha)} - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T) \right] \right\} \quad \therefore (66) \\ &= \frac{1}{\alpha(1-\alpha)} - \inf_{T: \mathbb{R}^d \rightarrow \mathbb{R}} l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T) \\ &= \frac{1}{\alpha(1-\alpha)} - L_\alpha(Q, P). \end{aligned} \quad (69)$$

529 Here, we see

$$E[l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i)] = L_\alpha(Q, P). \quad (70)$$

530 Next, we show the first equality in (63) holds. Note that, it holds that

$$\frac{1}{N} \sum_{i=1}^N \inf_{T_i: \mathbb{R}^d \rightarrow \mathbb{R}} l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_i) \leq \inf_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \frac{1}{N} \sum_{i=1}^N l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T) \leq \frac{1}{N} \sum_{i=1}^N l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T^*). \quad (71)$$

531 Since $\inf_{T_i: \mathbb{R}^d \rightarrow \mathbb{R}} l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_i) = l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_*)$ from Proposition C.3, we have

$$\frac{1}{N} \sum_{i=1}^N l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T^*) \leq \inf_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \frac{1}{N} \sum_{i=1}^N l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T) \leq \frac{1}{N} \sum_{i=1}^N l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T^*). \quad (72)$$

532 Therefore,

$$\inf_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \frac{1}{N} \sum_{i=1}^N l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T) = \frac{1}{N} \sum_{i=1}^N l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T^*). \quad (73)$$

533 From this, we see

$$\begin{aligned} \hat{L}_\alpha^{(N)}(Q, P) &= \inf_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \frac{1}{N} \sum_{i=1}^N l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T) \\ &= \frac{1}{N} \sum_{i=1}^N l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T^*) \\ &= \frac{1}{N} \sum_{i=1}^N l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i) \\ &= l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i). \end{aligned} \quad (74)$$

Subsequently, by integrating both sides of the above equation, we have

$$E\left[\hat{L}_\alpha^{(N)}(Q, P)\right] = E\left[l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i)\right]. \quad (75)$$

Here, we have (63) from (70) and (75).

To see (62), note that, it holds that

$$\hat{L}_\alpha^{(N)}(Q, P; T) = \frac{1}{N} \sum_{i=1}^N l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T). \quad (76)$$

By integrating both sides of the above equation, we have

$$\begin{aligned} E\left[\hat{L}_\alpha^{(N)}(Q, P; T)\right] &= E\left[\frac{1}{N} \sum_{i=1}^N l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T)\right] \\ &= \frac{1}{N} \sum_{i=1}^N E\left[l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T)\right] \\ &= E\left[l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T)\right] \\ &= L_\alpha(Q, P; T). \end{aligned} \quad (77)$$

Here, we see that (62) holds.

This completes the proof. \square

Proposition C.8. Let $T_\theta(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}$ be a function such that the map $\theta = (\theta_1, \theta_2, \dots, \theta_p) \in \Theta \mapsto T_\theta(\mathbf{x})$ is differentiable for all θ and μ -almost every $\mathbf{x} \in \mathbb{R}^d$. Assume that, for a point $\bar{\theta} \in \Theta$, it holds that $E_P[e^{(\alpha-1) \cdot T_{\bar{\theta}}(\mathbf{x})}] < \infty$ and $E_Q[e^{\alpha \cdot T_{\bar{\theta}}(\mathbf{x})}] < \infty$, and there exist a compact neighborhood of the $\bar{\theta}$, which is denoted by $B_{\bar{\theta}}$, and a constant value L , such that $|T_\psi(\mathbf{x}) - T_{\bar{\theta}}(\mathbf{x})| < L\|\psi - \bar{\theta}\|$ holds. Then, for $l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T)$ and $\hat{L}_\alpha^{(N)}(Q, P; T)$, $L_\alpha(Q, P; T)$ in Proposition C.12, it holds that

$$E\left[\nabla_\theta L_\alpha(\hat{Q}^{(N)}, \hat{P}^{(N)}; T_\theta)|_{\theta=\bar{\theta}}\right] = E\left[\nabla_\theta l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_\theta)|_{\theta=\bar{\theta}}\right] = \nabla_\theta L_\alpha(Q, P; T_\theta)|_{\theta=\bar{\theta}}. \quad (78)$$

Here, $E[\cdot]$ denotes $E_P[E_Q[\cdot]]$.

proof of Proposition C.8. We now consider the values, as $\psi \rightarrow \bar{\theta}$, of the following two integrals:

$$\int \frac{1}{\|\psi - \bar{\theta}\|} \left\{ \frac{1}{\alpha} e^{\alpha \cdot T_\psi} - \frac{1}{\alpha} e^{\alpha \cdot T_{\bar{\theta}}} \right\} dQ, \quad (79)$$

and

$$\int \frac{1}{\|\psi - \bar{\theta}\|} \left\{ \frac{1}{1-\alpha} e^{(\alpha-1) \cdot T_\psi} - \frac{1}{1-\alpha} e^{(\alpha-1) \cdot T_{\bar{\theta}}} \right\} dP. \quad (80)$$

Note that, it follows from the intermediate value theorem that

$$\left| \frac{1}{\alpha} e^{\alpha \cdot x} - \frac{1}{\alpha} e^{\alpha \cdot y} \right| = |x - y| \cdot e^{\alpha \cdot \{\gamma \cdot y + (1-\gamma) \cdot x\}} \quad (\exists \gamma \in [0, 1]). \quad (81)$$

By using the above equation as $x = T_\psi(\mathbf{x})$ and $y = T_{\bar{\theta}}(\mathbf{x})$ for the integrand of (79), we see

$$\begin{aligned} &\left| \frac{1}{\|\psi - \bar{\theta}\|} \left\{ \frac{1}{\alpha} e^{\alpha \cdot T_\psi(\mathbf{x})} - \frac{1}{\alpha} e^{\alpha \cdot T_{\bar{\theta}}(\mathbf{x})} \right\} \right| \\ &= \frac{1}{\|\psi - \bar{\theta}\|} |T_\psi(\mathbf{x}) - T_{\bar{\theta}}(\mathbf{x})| \cdot e^{\alpha \cdot \{\gamma_{\mathbf{x}} \cdot (T_{\bar{\theta}}(\mathbf{x}) + (1-\gamma_{\mathbf{x}}) \cdot (T_\psi(\mathbf{x}) - T_{\bar{\theta}}(\mathbf{x})))\}} \quad (\gamma_{\mathbf{x}} \in [0, 1]) \\ &= \frac{1}{\|\psi - \bar{\theta}\|} |T_\psi(\mathbf{x}) - T_{\bar{\theta}}(\mathbf{x})| \cdot e^{\alpha \cdot \gamma_{\mathbf{x}} \cdot (T_\psi(\mathbf{x}) - T_{\bar{\theta}}(\mathbf{x}))} \cdot e^{\alpha \cdot T_{\bar{\theta}}(\mathbf{x})} \\ &\leq \frac{1}{\|\psi - \bar{\theta}\|} |T_\psi(\mathbf{x}) - T_{\bar{\theta}}(\mathbf{x})| \cdot e^{\alpha \gamma_{\mathbf{x}} |T_\psi(\mathbf{x}) - T_{\bar{\theta}}(\mathbf{x})|} \cdot e^{\alpha \cdot T_{\bar{\theta}}(\mathbf{x})} \\ &\leq L \cdot e^{\alpha L \cdot \|\psi - \bar{\theta}\|} \cdot e^{\alpha \cdot T_{\bar{\theta}}(\mathbf{x})}, \end{aligned} \quad (82)$$

550 for all $\psi \in B_{\bar{\theta}}$.

551 Thus, integrating the term on the left hand side of (82) by Q , we see

$$\begin{aligned} & \int \left| \frac{1}{\|\psi - \bar{\theta}\|} \left\{ \frac{1}{\alpha} e^{\alpha \cdot T_\psi(\mathbf{X})} - \frac{1}{\alpha} e^{\alpha \cdot T_{\bar{\theta}}(\mathbf{X})} \right\} \right| dQ \\ & \leq \int L \cdot e^{\alpha L \cdot \|\psi - \bar{\theta}\|} \cdot e^{\alpha \cdot T_{\bar{\theta}}(\mathbf{X})} dQ \\ & = L \cdot e^{\alpha L \cdot \|\psi - \bar{\theta}\|} E_Q [e^{\alpha \cdot T_{\bar{\theta}}}] . \end{aligned} \quad (83)$$

552 Considering the supremum for $\psi \in B_{\bar{\theta}}$ of (83), it holds that

$$\begin{aligned} & \sup_{\psi \in B_{\bar{\theta}}} \left\{ \int \left| \frac{1}{\|\psi - \bar{\theta}\|} \left\{ \frac{1}{\alpha} e^{\alpha \cdot T_\psi} - \frac{1}{\alpha} e^{\alpha \cdot T_{\bar{\theta}}} \right\} \right| dQ \right\} \\ & \leq \sup_{\psi \in B_{\bar{\theta}}} \left\{ L \cdot e^{\alpha L \cdot \|\psi - \bar{\theta}\|} E_Q [e^{\alpha \cdot T_{\bar{\theta}}}] \right\} \\ & = E_Q [e^{\alpha \cdot T_{\bar{\theta}}}] \cdot \sup_{\psi \in B_{\bar{\theta}}} L \cdot e^{\alpha L \cdot \|\psi - \bar{\theta}\|} < \infty, \end{aligned} \quad (84)$$

553 since $B_{\bar{\theta}}$ is compact.

554 Similarly, as for (80), we see

$$\begin{aligned} & \sup_{\psi \in B_{\bar{\theta}}} \int \left| \frac{1}{\|\psi - \bar{\theta}\|} \left\{ \frac{1}{1-\alpha} e^{(\alpha-1) \cdot T_\psi(\mathbf{X})} - \frac{1}{1-\alpha} e^{(\alpha-1) \cdot T_{\bar{\theta}}(\mathbf{X})} \right\} \right| dP \\ & \leq \sup_{\psi \in B_{\bar{\theta}}} \left\{ L \cdot e^{(1-\alpha)L \cdot \|\psi - \bar{\theta}\|} E_P [e^{(1-\alpha) \cdot T_{\bar{\theta}}}] \right\} \\ & = E_P [e^{(1-\alpha) \cdot T_{\bar{\theta}}}] \cdot \sup_{\psi \in B_{\bar{\theta}}} L \cdot e^{(1-\alpha)L \cdot \|\psi - \bar{\theta}\|} < \infty. \end{aligned} \quad (85)$$

555 From (84) and (85), we obtain

$$\begin{aligned} & \sup_{\psi \in B_{\bar{\theta}}} \int \int \left| \frac{1}{\|\psi - \bar{\theta}\|} \left\{ l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_\psi) - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_{\bar{\theta}}) \right\} \right| dP dQ \\ & = \sup_{\psi \in B_{\bar{\theta}}} \int \int \left| \frac{1}{\|\psi - \bar{\theta}\|} \left\{ \frac{1}{\alpha} e^{\alpha \cdot T_\psi(\mathbf{X}_{\sim Q}^i)} - \frac{1}{\alpha} e^{\alpha \cdot T_{\bar{\theta}}(\mathbf{X}_{\sim Q}^i)} \right\} \right. \\ & \quad \left. + \frac{1}{\|\psi - \bar{\theta}\|} \left\{ \frac{1}{1-\alpha} e^{(\alpha-1) \cdot T_\psi(\mathbf{X}_{\sim P}^i)} - \frac{1}{1-\alpha} e^{(\alpha-1) \cdot T_{\bar{\theta}}(\mathbf{X}_{\sim P}^i)} \right\} \right| dP dQ \\ & \leq \sup_{\psi \in B_{\bar{\theta}}} \int \int \left| \frac{1}{\|\psi - \bar{\theta}\|} \left\{ \frac{1}{\alpha} e^{\alpha \cdot T_\psi(\mathbf{X}_{\sim Q}^i)} - \frac{1}{\alpha} e^{\alpha \cdot T_{\bar{\theta}}(\mathbf{X}_{\sim Q}^i)} \right\} \right| \\ & \quad + \left| \frac{1}{\|\psi - \bar{\theta}\|} \left\{ \frac{1}{1-\alpha} e^{(\alpha-1) \cdot T_\psi(\mathbf{X}_{\sim P}^i)} - \frac{1}{1-\alpha} e^{(\alpha-1) \cdot T_{\bar{\theta}}(\mathbf{X}_{\sim P}^i)} \right\} \right| dP dQ \\ & = \sup_{\psi \in B_{\bar{\theta}}} \left\{ \int \left| \frac{1}{\|\psi - \bar{\theta}\|} \left\{ \frac{1}{\alpha} e^{\alpha \cdot T_\psi(\mathbf{X})} - \frac{1}{\alpha} e^{\alpha \cdot T_{\bar{\theta}}(\mathbf{X})} \right\} \right| dQ \right. \\ & \quad \left. + \int \left| \frac{1}{\|\psi - \bar{\theta}\|} \left\{ \frac{1}{1-\alpha} e^{(\alpha-1) \cdot T_\psi(\mathbf{X})} - \frac{1}{1-\alpha} e^{(\alpha-1) \cdot T_{\bar{\theta}}(\mathbf{X})} \right\} \right| dP \right\} \\ & \leq \sup_{\psi \in B_{\bar{\theta}}} \left\{ \int \left| \frac{1}{\|\psi - \bar{\theta}\|} \left\{ \frac{1}{\alpha} e^{\alpha \cdot T_\psi(\mathbf{X})} - \frac{1}{\alpha} e^{\alpha \cdot T_{\bar{\theta}}(\mathbf{X})} \right\} \right| dQ \right\} \\ & \quad + \sup_{\psi \in B_{\bar{\theta}}} \left\{ \int \left| \frac{1}{\|\psi - \bar{\theta}\|} \left\{ \frac{1}{1-\alpha} e^{(\alpha-1) \cdot T_\psi(\mathbf{X})} - \frac{1}{1-\alpha} e^{(\alpha-1) \cdot T_{\bar{\theta}}(\mathbf{X})} \right\} \right| dP \right\} \\ & < \infty. \end{aligned} \quad (86)$$

556 Therefore, the following set is uniformly integrable for μ :

$$\left\{ \frac{1}{\|\psi - \bar{\theta}\|} \left\{ l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_\psi) - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_{\bar{\theta}}) \right\} : \psi \in B_{\bar{\theta}} \right\}. \quad (87)$$

557 Then, from the property of the Lebesgue integral (25, P188, Theorem 4), the integral $\int \int (\cdot) dP dQ$
 558 and the limitation $\lim_{\psi \rightarrow \bar{\theta}}$ for the above term are exchangeable.

559 Hence, we have

$$\begin{aligned}
 & \lim_{\psi \rightarrow \bar{\theta}} \int \int \frac{1}{\|\psi - \bar{\theta}\|} \left\{ l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_\psi) - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_{\bar{\theta}}) \right\} dP dQ \\
 &= \int \int \lim_{\psi \rightarrow \bar{\theta}} \left[\frac{1}{\|\psi - \bar{\theta}\|} \left\{ l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_\psi) - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_{\bar{\theta}}) \right\} \right] dP dQ \\
 &= \int \int \nabla_\theta l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_\theta)|_{\theta=\bar{\theta}} dP dQ \\
 &= E \left[\nabla_\theta l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_\theta)|_{\theta=\bar{\theta}} \right]. \tag{88}
 \end{aligned}$$

560 On the other hand, for the term on the left hand side of (88), we obtain

$$\begin{aligned}
 & \lim_{\psi \rightarrow \bar{\theta}} \int \int \frac{1}{\|\psi - \bar{\theta}\|} \left\{ l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_\psi) - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_{\bar{\theta}}) \right\} dP dQ \\
 &= \lim_{\psi \rightarrow \bar{\theta}} \frac{1}{\|\psi - \bar{\theta}\|} \int \int \left\{ l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_\psi) - l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_{\bar{\theta}}) \right\} dP dQ \\
 &= \lim_{\psi \rightarrow \bar{\theta}} \frac{1}{\|\psi - \bar{\theta}\|} \left\{ L_\alpha(Q, P; T_\psi) - L_\alpha(Q, P; T_{\bar{\theta}}) \right\} \\
 &= \nabla_\theta L_\alpha(Q, P; T_\theta)|_{\theta=\bar{\theta}}. \tag{89}
 \end{aligned}$$

561 From (88) and (89), we obtain

$$E \left[\nabla_\theta l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_\theta)|_{\theta=\bar{\theta}} \right] = \nabla_\theta L_\alpha(Q, P; T_\theta)|_{\theta=\bar{\theta}}. \tag{90}$$

562 From this, we also have

$$\begin{aligned}
 E \left[\nabla_\theta l_\alpha(\hat{Q}^{(N)}, \hat{P}^{(N)}; T_\theta)|_{\theta=\bar{\theta}} \right] &= E \left[\nabla_\theta|_{\theta=\bar{\theta}} \frac{1}{N} \sum_{i=1}^N l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_\theta) \right] \\
 &= E \left[\frac{1}{N} \sum_{i=1}^N \nabla_\theta l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_\theta)|_{\theta=\bar{\theta}} \right] \\
 &= \frac{1}{N} \sum_{i=1}^N E \left[\nabla_\theta l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T_\theta)|_{\theta=\bar{\theta}} \right] \\
 &= \frac{1}{N} \sum_{i=1}^N \nabla_\theta L_\alpha(Q, P; T_\theta)|_{\theta=\bar{\theta}} \\
 &= \nabla_\theta L_\alpha(Q, P; T_\theta)|_{\theta=\bar{\theta}}. \tag{91}
 \end{aligned}$$

563 Here, we see (78) from (90) and (91).

564 This completes the proof. \square

565 **Proposition C.9.** *Let*

$$\hat{D}_\alpha^{(N)}(Q||P) = \sup_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \left[\frac{1}{\alpha(1-\alpha)} - \frac{1}{\alpha} \left\{ \frac{1}{N} \sum_{i=1}^N e^{\alpha \cdot T(\mathbf{X}_{\sim Q}^i)} \right\} - \frac{1}{1-\alpha} \left\{ \frac{1}{N} \sum_{i=1}^N e^{(\alpha-1) \cdot T(\mathbf{X}_{\sim P}^i)} \right\} \right], \tag{92}$$

566 where supremum is considered over all measurable functions $T: \mathbb{R}^d \rightarrow \mathbb{R}$ with $E_P[e^{(\alpha-1) \cdot T}] < \infty$
 567 and $E_Q[e^{\alpha \cdot T}] < \infty$.

568 Then, it holds that if $\alpha \neq 1/2$,

$$\begin{aligned}
 & \sqrt{N} \left\{ \hat{D}_\alpha^{(N)}(Q||P) - D_\alpha(Q||P) \right\} \\
 & \xrightarrow{d} \mathcal{N} \left(0, C_\alpha^1 \cdot D_{2\alpha-1}(Q||P) + C_\alpha^2 \cdot D_\alpha(Q||P) + C_\alpha^3 \cdot D_\alpha(Q||P)^2 \right), \tag{93}
 \end{aligned}$$

569 where

$$C_\alpha^1 = \left(\frac{1}{\alpha^2} + \frac{1}{(1-\alpha)^2} \right) \cdot (2\alpha - 1) \cdot (2\alpha - 2), \quad (94)$$

$$C_\alpha^2 = \frac{2\{\alpha^2 + (1-\alpha)^2\}}{\alpha \cdot (1-\alpha)} \quad \text{and} \quad C_\alpha^3 = -\alpha^2 - (1-\alpha)^2, \quad (95)$$

570 and if $\alpha = 1/2$,

$$\begin{aligned} & \sqrt{N} \left\{ \hat{D}_\alpha^{(N)}(Q||P) - D_\alpha(Q||P) \right\} \\ & \xrightarrow{d} \mathcal{N} \left(0, 4 D_\alpha(Q||P) - \frac{1}{2} D_\alpha(Q||P)^2 \right). \end{aligned} \quad (96)$$

571 *proof of Proposition C.9.* First, we note that

$$\begin{aligned} \hat{D}_\alpha^{(N)}(Q||P) &= \sup_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \left[\frac{1}{\alpha(1-\alpha)} - \frac{1}{\alpha} \left\{ \frac{1}{N} \sum_{i=1}^N e^{\alpha \cdot T(\mathbf{X}_{\sim Q}^i)} \right\} - \frac{1}{1-\alpha} \left\{ \frac{1}{N} \sum_{i=1}^N e^{(\alpha-1) \cdot T(\mathbf{X}_{\sim P}^i)} \right\} \right] \\ &= \frac{1}{\alpha(1-\alpha)} - \inf_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \left[\frac{1}{\alpha} \left\{ \frac{1}{N} \sum_{i=1}^N e^{\alpha \cdot T(\mathbf{X}_{\sim Q}^i)} \right\} + \frac{1}{1-\alpha} \left\{ \frac{1}{N} \sum_{i=1}^N e^{(\alpha-1) \cdot T(\mathbf{X}_{\sim P}^i)} \right\} \right] \\ &= \frac{1}{\alpha(1-\alpha)} - \inf_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \frac{1}{N} \sum_{i=1}^N \left[\frac{1}{\alpha} e^{\alpha \cdot T(\mathbf{X}_{\sim Q}^i)} + \frac{1}{1-\alpha} e^{(\alpha-1) \cdot T(\mathbf{X}_{\sim P}^i)} \right] \\ &= \frac{1}{\alpha(1-\alpha)} - \inf_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \frac{1}{N} \sum_{i=1}^N [l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i; T)] \\ &= \frac{1}{\alpha(1-\alpha)} - \frac{1}{N} \sum_{i=1}^N l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i). \end{aligned} \quad (97)$$

572 On the other hand, from Lemma C.7, it holds that

$$\begin{aligned} D_\alpha(Q||P) &= \sup_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \left\{ \frac{1}{\alpha(1-\alpha)} - \frac{1}{\alpha} E_Q [e^{\alpha \cdot T}] - \frac{1}{1-\alpha} E_P [e^{(\alpha-1) \cdot T}] \right\} \\ &= \frac{1}{\alpha(1-\alpha)} - \inf_{T: \mathbb{R}^d \rightarrow \mathbb{R}} \left\{ \frac{1}{\alpha} E_Q [e^{\alpha \cdot T}] - \frac{1}{1-\alpha} E_P [e^{(\alpha-1) \cdot T}] \right\} \\ &= \frac{1}{\alpha(1-\alpha)} - \frac{1}{N} \sum_{i=1}^N L_\alpha(Q, P) \\ &= \frac{1}{\alpha(1-\alpha)} - \frac{1}{N} \sum_{i=1}^N E [l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i)]. \end{aligned} \quad (98)$$

573 Subtracting (98) from (97), we have

$$\hat{D}_\alpha^{(N)}(Q||P) - D_\alpha(Q||P) = \frac{1}{N} \sum_{i=1}^N \{ l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i) - E [l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i)] \}. \quad (99)$$

574 Let $L_i = l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i) - E[l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i)]$. Then $\{L_i\}_{i=1}^N$ are independently identically
 575 distributed variables whose means and variances are as follows:

$$\begin{aligned}
 E[L_i] &= 0, \\
 \text{Var}[L_i] &= E\left[\left\{l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i) - E[l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i)]\right\}^2\right] \\
 &= E_P\left[E_Q\left[\frac{1}{\alpha}\left\{\left(\frac{dQ}{dP}\right)^{-\alpha}(\mathbf{X}_{\sim Q}^i) - E_Q\left[\left(\frac{dQ}{dP}\right)^{-\alpha}\right]\right\}\right.\right. \\
 &\quad \left.\left. + \frac{1}{1-\alpha}\left\{\left(\frac{dQ}{dP}\right)^{1-\alpha}(\mathbf{X}_{\sim P}^i) - E_P\left[\left(\frac{dQ}{dP}\right)^{1-\alpha}\right]\right\}\right]^2\right] \\
 &= \frac{1}{\alpha^2} \cdot E_Q\left\{\left(\frac{dQ}{dP}\right)^{-\alpha}(\mathbf{X}^i) - E_Q\left[\left(\frac{dQ}{dP}\right)^{-\alpha}\right]\right\}^2 \\
 &\quad + \frac{1}{(1-\alpha)^2} \cdot E_P\left\{\left(\frac{dQ}{dP}\right)^{1-\alpha}(\mathbf{X}^i) - E_P\left[\left(\frac{dQ}{dP}\right)^{1-\alpha}\right]\right\}^2 \\
 &= \frac{1}{\alpha^2} \cdot E_P\left\{\frac{dQ}{dP} \cdot \left(\frac{dQ}{dP}\right)^{-\alpha}(\mathbf{X}^i) - E_P\left[\frac{dQ}{dP} \cdot \left(\frac{dQ}{dP}\right)^{-\alpha}\right]\right\}^2 \\
 &\quad + \frac{1}{(1-\alpha)^2} \cdot E_P\left\{\left(\frac{dQ}{dP}\right)^{1-\alpha}(\mathbf{X}^i) - E_P\left[\left(\frac{dQ}{dP}\right)^{1-\alpha}\right]\right\}^2 \\
 &= \left\{\frac{1}{\alpha^2} + \frac{1}{(1-\alpha)^2}\right\} \cdot E_P\left\{\left(\frac{dQ}{dP}\right)^{1-\alpha} - E_P\left[\left(\frac{dQ}{dP}\right)^{1-\alpha}\right]\right\}^2 \\
 &= \left\{\frac{1}{\alpha^2} + \frac{1}{(1-\alpha)^2}\right\} \left\{E_P\left[\left(\frac{dQ}{dP}\right)^{2 \cdot (1-\alpha)}\right] - \left\{E_P\left[\left(\frac{dQ}{dP}\right)^{1-\alpha}\right]\right\}^2\right\} \\
 &= \left\{\frac{1}{\alpha^2} + \frac{1}{(1-\alpha)^2}\right\} \\
 &\quad \times \left[(2\alpha-1)(2\alpha-2) \left\{\frac{1}{(2\alpha-1)(2\alpha-2)} E_P\left[\left(\frac{dQ}{dP}\right)^{1-(2\alpha-1)} - 1\right]\right\} + 1\right. \\
 &\quad \left. - \alpha^2(1-\alpha)^2 \left\{\frac{1}{\alpha \cdot (\alpha-1)} E_P\left[\left(\frac{dQ}{dP}\right)^{1-\alpha} - 1\right]\right\}^2\right. \\
 &\quad \left. + \alpha^2(1-\alpha)^2 \left\{\frac{2}{\alpha \cdot (\alpha-1)} E_P\left[\left(\frac{dQ}{dP}\right)^{1-\alpha} - 1\right]\right\} - 1\right]. \tag{101}
 \end{aligned}$$

576 From this, if $\alpha \neq 1/2$, we have

$$\text{Var}[L_i] = C_\alpha^1 \cdot D_{2\alpha-1}(Q||P) + C_\alpha^2 \cdot D_\alpha(Q||P) + C_\alpha^3 \cdot D_\alpha(Q||P)^2, \tag{102}$$

577 where

$$\begin{aligned}
 C_\alpha^1 &= \left(\frac{1}{\alpha^2} + \frac{1}{(1-\alpha)^2}\right) \cdot (2\alpha-1) \cdot (2\alpha-2), \\
 C_\alpha^2 &= \frac{2\{\alpha^2 + (1-\alpha)^2\}}{\alpha \cdot (1-\alpha)} \quad \text{and} \quad C_\alpha^3 = -\alpha^2 - (1-\alpha)^2,
 \end{aligned}$$

578 and if $\alpha = 1/2$, we obtain

$$\text{Var}[L_i] = 4D_\alpha(Q||P) - \frac{1}{2}D_\alpha(Q||P)^2. \tag{103}$$

Therefore, by the central limit theorem, we see

$$\frac{1}{N} \sum_{i=1}^N \{l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i) - E[l_\alpha(\mathbf{X}_{\sim Q}^i, \mathbf{X}_{\sim P}^i)]\} \xrightarrow{d} \mathcal{N}(0, *). \quad (104)$$

Here the “*” is (102) or (103), which corresponds to the cases that $\alpha \neq 1/2$ or $\alpha = 1/2$, respectively.

This completes the proof. \square

We mention that the statement of the following corollary is the same as Corollary 1 in Birrell et al.(2022).

Corollary C.10 (Birrell et al.(2022), P19, Corollary 1). *For $\alpha = 1/2$, it holds that*

$$\lim_{N \rightarrow \infty} \frac{N \cdot \text{Var}[\hat{D}_{1/2}^{(N)}(Q||P)]}{D_{1/2}(Q||P)^2} = \frac{8 D_{1/2}(Q||P) - D_{1/2}(Q||P)^2}{2 D_{1/2}(Q||P)^2}. \quad (105)$$

Thus, the sample complexity of D_α for $\alpha = 1/2$ is $O(1)$.

proof of Corollary C.10. The statement of the corollary follows from (96) in Proposition C.9. \square

Proposition C.11. *Let $\{T_k\}_{k=1}^\infty$ be a sequence of functions in \mathcal{T}^α with $E_P[e^{-T_k(\mathbf{X})}] = 1$ such that $\lim_{k \rightarrow \infty} T_k = -\log dQ/dP$, P -almost everywhere. Subsequently, let $\{\mathbf{X}_k^Q\}_{k=1}^\infty$ be a sequence of measures on \mathbb{R}^d defined as follows:*

$$\mathbf{X}_k^Q = e^{-T_k(\mathbf{X}^P)} \cdot \mathbf{X}^P.$$

Then, it holds that

$$\mathbf{X}_k^Q \xrightarrow{d} \mathbf{X}^Q, \quad \text{as } k \rightarrow \infty. \quad (106)$$

proof of Proposition C.11. Let Q_k denote the probability distribution of \mathbf{X}_k^Q : $Q_k(A) = P(\mathbf{X}_k^Q \in A)$ for all $A \in \mathcal{F}$. Then, since $\frac{dQ_k}{dP} = e^{-T_k(\mathbf{X})}$, we see

$$\frac{dQ_k}{dQ} = e^{-T_k(\mathbf{X})} \cdot \frac{dP}{dQ}. \quad (107)$$

Now, from Corollary 6 in [7], for probability measures A and B with $A \ll \mu$ and $B \ll \mu$, it holds that

$$\frac{1}{2} \left\{ E_\mu \left| \frac{dA}{d\mu} - \frac{dB}{d\mu} \right| \right\}^2 \leq D_\alpha(A||B). \quad (108)$$

By substituting $A = Q_k$ and $B = Q$ into (108), we have

$$\begin{aligned} \frac{1}{2} \left\{ E_\mu \left| \frac{dQ_k}{d\mu} - \frac{dQ}{d\mu} \right| \right\}^2 &\leq D_\alpha(Q_k||Q) \\ &= \int \frac{1}{\alpha(\alpha-1)} \left\{ \left(\frac{dQ_k}{dQ} \right)^{1-\alpha} - 1 \right\} dQ \\ &= \int \frac{1}{\alpha(\alpha-1)} \int \left(\frac{dQ_k}{dQ} \right)^{1-\alpha} dQ - \frac{1}{\alpha(\alpha-1)} \\ &= \frac{1}{\alpha(\alpha-1)} \int \left(\frac{dP}{dQ} \right)^{1-\alpha} \cdot e^{(\alpha-1) \cdot T_k} \frac{dQ}{dP} dP - \frac{1}{\alpha(\alpha-1)} \\ &= \frac{1}{\alpha(\alpha-1)} \int \left(\frac{dQ}{dP} \right)^\alpha \cdot e^{(\alpha-1) \cdot T_k} dP - \frac{1}{\alpha(\alpha-1)}. \end{aligned} \quad (109)$$

596 Now, from Hölder's inequality, we have

$$\begin{aligned}
\int \left| \left(\frac{dQ}{dP} \right)^\alpha \cdot e^{(\alpha-1) \cdot T_k} \right| dP &\leq \left\{ \int \left(\frac{dP}{dQ} \right)^{\frac{-\alpha}{\alpha-1}} dP \right\}^\alpha \\
&\quad \times \left\{ \int \left(e^{(\alpha-1) \cdot T_k} \right)^{\frac{1}{1-\alpha}} dP \right\}^{1-\alpha} \\
&= \left(E_P [e^{-T_k}] \right)^{1-\alpha} \\
&= 1 < \infty.
\end{aligned} \tag{110}$$

597 Hence, we see the following sequence is uniformly integrable for P :

$$\left\{ \left(\frac{dQ}{dP} \right)^\alpha \cdot e^{(\alpha-1) \cdot T_k} \right\}_{k=1}^N. \tag{111}$$

598 Then, for (109) as $k \rightarrow \infty$, we see

$$\begin{aligned}
\lim_{k \rightarrow \infty} \frac{1}{2} \left\{ E_\mu \left| \frac{dQ_k}{d\mu} - \frac{dQ}{d\mu} \right| \right\}^2 &\leq \lim_{k \rightarrow \infty} \frac{1}{\alpha(\alpha-1)} \int \left(\frac{dQ}{dP} \right)^\alpha \cdot e^{(\alpha-1) \cdot T_k} dP - \frac{1}{\alpha(\alpha-1)} \\
&= \frac{1}{\alpha(\alpha-1)} \int \left(\frac{dQ}{dP} \right)^\alpha \cdot \left(\frac{dQ}{dP} \right)^{1-\alpha} dP - \frac{1}{\alpha(\alpha-1)} \\
&= \frac{1}{\alpha(\alpha-1)} \int \frac{dQ}{dP} dP - \frac{1}{\alpha(\alpha-1)} \\
&= 0.
\end{aligned}$$

599 Thus, Q_k converges to Q in total variation.

600 The statement (106), the convergence of $\hat{\mathbf{X}}_k^Q$ to \mathbf{X}^Q in distribution, is derived from the convergence
601 of Q_k to Q in total variation.

602 This completes the proof. \square

603 **Corollary C.12.** Let $\{T_k\}_{k=1}^\infty$ be a sequence of functions in \mathcal{T}^α such that

$$\begin{aligned}
D_\alpha(\hat{Q}^{(N)} || \hat{P}^{(N)}) \\
= \lim_{k \rightarrow \infty} \left\{ \frac{1}{\alpha(1-\alpha)} - \frac{1}{\alpha} \frac{1}{N} \sum_{i=1}^N e^{\alpha \cdot T_k(\mathbf{X}_{\sim Q}^i)} + \frac{1}{1-\alpha} \frac{1}{N} \sum_{i=1}^N e^{(\alpha-1) \cdot T_k(\mathbf{X}_{\sim P}^i)} \right\}. \tag{112}
\end{aligned}$$

604 Subsequently, let $\{\hat{\mathbf{X}}_{\mathbb{Q}}^{(N)}(k)\}_{k=1}^\infty$ be a sequence of measures on \mathbb{R}^d defined as follows:

$$\hat{\mathbf{X}}_{\mathbb{Q}}^{(N)}(k) = e^{-T_k} \cdot \mathbf{X}_{\mathbb{P}}^{(N)} \tag{113}$$

605 Then, it holds that, as $k \rightarrow \infty$,

$$\hat{\mathbf{X}}_{\mathbb{Q}}^{(N)}(k) \xrightarrow{d} \hat{\mathbf{X}}_{\mathbb{Q}}^{(N)}. \tag{114}$$

606 *proof of Corollary C.12.* Let ν be the countable measure on $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N\}$:

$$\nu(\mathbf{x}) = \begin{cases} 1 & \text{if } 1 \leq \exists i \leq N \text{ s.t. } \mathbf{X}_i = \mathbf{x}, \\ 0 & \text{otherwise.} \end{cases} \tag{115}$$

607 Then, $\hat{P}^{(N)} \ll \nu$ and $\hat{Q}^{(N)} \ll \nu$.

608 For Proposition C.11 and its proof, substituting $\hat{P}^{(N)}$ for P , $\hat{Q}^{(N)}$ for Q , and ν for μ , we see that
609 the statement of the corollary holds.

610 This completes the proof. \square

611 **Proposition C.13.** For $\hat{T} \in \mathcal{T}^\alpha$, let \hat{Q} and \hat{P} be two probabilities defined as

$$d\hat{Q} = e^{-\hat{T}} \cdot dP \quad \text{and} \quad d\hat{P} = e^{\hat{T}} \cdot dQ,$$

612 and let $T_* = -\log dQ/dP$.

613 Then, it holds that

$$\begin{aligned} & \frac{\alpha}{2} \left\{ E_\mu \left| \hat{Q} - Q \right| \right\}^2 + \frac{1-\alpha}{2} \left\{ E_\mu \left| \hat{P} - P \right| \right\}^2 \\ & \leq \frac{1}{\alpha(1-\alpha)} - \frac{1}{\alpha} E_Q \left[e^{\alpha \cdot (T_* - \hat{T})} \right] - \frac{1}{1-\alpha} E_P \left[e^{(\alpha-1) \cdot (T_* - \hat{T})} \right] \end{aligned} \quad (116)$$

$$= L_\alpha(Q, P; \hat{T}) - L_\alpha(Q, P; T_*). \quad (117)$$

614 Here, $L_\alpha(Q, P; \cdot)$ in (117) is defined as (57) in Lemma C.7

615 *proof of Proposition C.13.* First, we see (116). Note that, it holds that

$$\frac{d\hat{Q}}{dQ} = e^{-\hat{T}} \cdot \frac{dP}{dQ} \quad \text{and} \quad \frac{d\hat{P}}{dP} = e^{\hat{T}} \cdot \frac{dQ}{dP}. \quad (118)$$

616 By using (108), we have

$$\frac{\alpha}{2} \left\{ E_\mu \left| \hat{Q} - Q \right| \right\}^2 + \frac{1-\alpha}{2} \left\{ E_\mu \left| \hat{P} - P \right| \right\}^2 \leq (1-\alpha) \cdot D_{1-\alpha}(\hat{Q}||Q) + \alpha \cdot D_\alpha(\hat{P}||P).$$

617 Thus, we obtain

$$\begin{aligned} & \frac{\alpha}{2} \left\{ E_\mu \left| \hat{Q} - Q \right| \right\}^2 + \frac{1-\alpha}{2} \left\{ E_\mu \left| \hat{P} - P \right| \right\}^2 \\ & \leq (1-\alpha) \cdot D_{1-\alpha}(\hat{Q}||Q) + \alpha \cdot D_\alpha(\hat{P}||P) \\ & = \frac{1-\alpha}{\alpha(\alpha-1)} \left\{ \int \left(\frac{d\hat{Q}}{dQ} \right)^\alpha dQ - 1 \right\} \\ & \quad + \frac{\alpha}{\alpha(\alpha-1)} \left\{ \int \left(\frac{dQ}{dP} \right)^{1-\alpha} dP - 1 \right\} \\ & = -\frac{1}{\alpha} \left\{ \int e^{-\alpha \cdot \hat{T}} \left(\frac{dP}{dQ} \right)^\alpha dQ - 1 \right\} \\ & \quad - \frac{1}{1-\alpha} \left\{ \int e^{(1-\alpha) \cdot \hat{T}} \left(\frac{dQ}{dP} \right)^{1-\alpha} dP - 1 \right\} \\ & = -\frac{1}{\alpha} \left\{ \int e^{-\alpha \cdot \hat{T}} e^{\alpha \cdot T_*} dQ - 1 \right\} \\ & \quad - \frac{1}{1-\alpha} \left\{ \int e^{-(\alpha-1) \cdot \hat{T}} e^{(\alpha-1) \cdot T_*} dP - 1 \right\} \\ & = \frac{1}{\alpha(1-\alpha)} - \frac{1}{\alpha} \int e^{\alpha \cdot (T_* - \hat{T})} dQ - \frac{1}{1-\alpha} \int e^{(\alpha-1) \cdot (T_* - \hat{T})} dP. \end{aligned} \quad (119)$$

618 Here, we see (116).

619 To obtain (117), we have

$$\begin{aligned} & \frac{1}{\alpha(1-\alpha)} - \frac{1}{\alpha} \int e^{\alpha \cdot (T_* - \hat{T})} dQ - \frac{1}{1-\alpha} \int e^{(\alpha-1) \cdot (T_* - \hat{T})} dP \\ & = \frac{1}{\alpha} + \frac{1}{1-\alpha} - \frac{1}{\alpha} \int e^{\alpha \cdot (T_* - \hat{T})} dQ - \frac{1}{1-\alpha} \int e^{(\alpha-1) \cdot (T_* - \hat{T})} dP \\ & = \int \left\{ \frac{1}{\alpha} \cdot \frac{dQ}{d\mu} - \frac{1}{\alpha} \cdot e^{\alpha \cdot (T_* - \hat{T})} \frac{dQ}{d\mu} \right\} d\mu \\ & \quad + \int \left\{ \frac{1}{1-\alpha} \cdot \frac{dP}{d\mu} - \frac{1}{1-\alpha} \cdot e^{(\alpha-1) \cdot (T_* - \hat{T})} \frac{dP}{d\mu} \right\} d\mu. \end{aligned} \quad (120)$$

By replacing the measures μ of the two integrals in (120) with

$$\nu = e^{-\alpha \cdot \hat{T}} d\mu \quad \text{and} \quad \tau = e^{-(\alpha-1) \cdot \hat{T}} d\mu, \quad (121)$$

respectively, we obtain

$$\begin{aligned} & \int \left\{ \frac{1}{\alpha} \cdot \frac{dQ}{d\nu} - \frac{1}{\alpha} \cdot e^{\alpha \cdot (T_* - \hat{T})} \frac{dQ}{d\nu} \right\} e^{\alpha \cdot \hat{T}} \cdot d\nu \\ & \quad + \int \left\{ \frac{1}{1-\alpha} \cdot \frac{dP}{d\tau} - \frac{1}{1-\alpha} \cdot e^{(\alpha-1) \cdot (T_* - \hat{T})} \frac{dP}{d\tau} \right\} e^{(\alpha-1) \cdot \hat{T}} \cdot d\tau \\ = & \int \left\{ \frac{1}{\alpha} \cdot e^{\alpha \cdot \hat{T}} \frac{dQ}{d\nu} - \frac{1}{\alpha} \cdot e^{\alpha \cdot T_*} \frac{dQ}{d\nu} \right\} d\nu \\ & \quad + \int \left\{ \frac{1}{1-\alpha} \cdot e^{(\alpha-1) \cdot \hat{T}} \frac{dP}{d\tau} - \frac{1}{1-\alpha} \cdot e^{(\alpha-1) \cdot T_*} \frac{dP}{d\tau} \right\} d\tau \\ = & \left\{ \frac{1}{\alpha} \int e^{\alpha \cdot \hat{T}} dQ + \frac{1}{1-\alpha} \int e^{(\alpha-1) \cdot \hat{T}} dP \right\} \\ & \quad - \left\{ \frac{1}{\alpha} \int e^{\alpha \cdot T_*} dQ + \frac{1}{1-\alpha} \int e^{(\alpha-1) \cdot T_*} dP \right\} \\ = & L_\alpha(Q, P; \hat{T}) - L_\alpha(Q, P; T_*). \end{aligned} \quad (122)$$

This completes the proof. \square

C.2 Proofs for Section 6

In this Section, we present two theorems for the proposed method in Section 6. Before presenting the first theorem, we briefly review Pearl's *do*-calculus (Pearl(1995)) used in the proof of the first theorem.

Theorem C.14 (*do*-calculus, Pearl(1995)). *Causal effects can be transformed by following rules R1-R3:*

- R1. $P(\mathbf{Y}|\text{do}(\mathbf{X}), \mathbf{Z}, \mathbf{W}) = P(\mathbf{Y}|\text{do}(\mathbf{X}), \mathbf{W})$, if $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}, \mathbf{W})_{\overline{G}(\mathbf{X})}$.
- R2. $P(\mathbf{Y}|\text{do}(\mathbf{X}), \text{do}(\mathbf{Z}), \mathbf{W}) = P(\mathbf{Y}|\text{do}(\mathbf{X}), \mathbf{Z}, \mathbf{W})$, if $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}, \mathbf{W})_{\overline{G}(\mathbf{X}, \mathbf{Z})}$.
- R3. $P(\mathbf{Y}|\text{do}(\mathbf{X}), \text{do}(\mathbf{Z}), \mathbf{W}) = P(\mathbf{Y}|\text{do}(\mathbf{X}), \mathbf{W})$, if $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}, \mathbf{W})_{\overline{G}(\mathbf{X}, \mathbf{Z}^*)}$, where $\mathbf{Z}^* = \mathbf{Z} \setminus \text{An}(\mathbf{W})_{\overline{G}(\mathbf{X})}$.

Here, $\overline{G}(\mathbf{A})$ denotes a graph obtained from G by deleting all arrows emerging from variables to \mathbf{A} , and $\overline{G}(\mathbf{A}, \mathbf{B})$ denotes a graph obtained from G by deleting both of all arrows emerging from any variables to \mathbf{A} and all arrows emerging from \mathbf{B} to any variables, and $(\mathbf{A} \perp\!\!\!\perp \mathbf{B})_G$ represents that there is no path between \mathbf{A} and \mathbf{B} in G .

We now provide the first theorem, which presents a sufficient condition for explanatory variables to be available for estimating causal effects.

Theorem C.15. *Let G be a DAG for \mathbf{V} and \mathbf{U} . For disjoint sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subset \mathbf{V}$, suppose that $P(\mathbf{Y}|\text{do}(\mathbf{X}), \mathbf{Z})$ is identifiable in G , and $\mathbf{X} \subset \text{An}(\mathbf{Y})_G$. Let $\mathbf{Z}_{De} = \mathbf{Z} \cap \text{De}(\mathbf{Y})_G$. Then,*

$$\begin{aligned} & P(\mathbf{Y}|\text{do}(\mathbf{X}), \mathbf{Z}) \\ = & \begin{cases} P(\mathbf{Y}|\mathbf{X}, \mathbf{Z}) & \text{if } \mathbf{Z}_{De} = \emptyset, \\ \frac{P(\mathbf{Y}|\mathbf{X}, \mathbf{Z} \setminus \mathbf{Z}_{De}) P(\mathbf{Z}_{De}|\mathbf{Y}, \mathbf{X}, \mathbf{Z} \setminus \mathbf{Z}_{De})}{P(\mathbf{Z}_{De}|\mathbf{X}, \mathbf{Z} \setminus \mathbf{Z}_{De})} & \text{if } \mathbf{Z}_{De} \neq \emptyset. \end{cases} \end{aligned} \quad (123)$$

proof of Theorem C.15. We note that each $Z_i \in \mathbf{Z}$ can be assumed to be that either $Z_i \in \text{An}(\mathbf{Y})_G$ or $Z_i \in \text{De}(\mathbf{Y})_G$. To see this, suppose that there exist some $Z_i \in \mathbf{Z}$ such that $Z_i \notin \text{An}(\mathbf{Y})_G$ and $Z_i \notin \text{De}(\mathbf{Y})_G$. Let $\mathbf{V}' = \mathbf{V} \setminus (\mathbf{Y} \cup \mathbf{X} \cup \mathbf{Z})$. Since $(\mathbf{Y} \perp\!\!\!\perp Z_i|\mathbf{X}, \mathbf{V}', \mathbf{U})_{\overline{G}(\mathbf{X})}$ holds for the Z_i , by applying *do*-calculus R1 in Theorem C.14, we have

$$P(\mathbf{Y}|\text{do}(\mathbf{X}), \mathbf{Z} \setminus \{Z_i\}, Z_i, \mathbf{V}', \mathbf{U}) = P(\mathbf{Y}|\text{do}(\mathbf{X}), \mathbf{Z} \setminus \{Z_i\}, \mathbf{V}', \mathbf{U}). \quad (124)$$

645 By marginalizing both sides of (124) for $\mathcal{X}_{\mathbf{V}' \cup \mathbf{U}}$, we obtain

$$P(\mathbf{Y}|do(\mathbf{X}), \mathbf{Z} \setminus \{Z_i\}, Z_i) = P(\mathbf{Y}|do(\mathbf{X}), \mathbf{Z} \setminus \{Z_i\}). \quad (125)$$

646 Thus, after repeating the above calculation, $P(\mathbf{Y}|do(\mathbf{X}), \mathbf{Z})$ finally includes only $Z_i \in \mathbf{Z}$ such that
 647 $Z_i \in An(\mathbf{Y})_G$ and $Z_i \in De(\mathbf{Y})_G$.

648 Therefore, in this proof, we assume that

$$\mathbf{Z} = An(\mathbf{Y})_G \cup De(\mathbf{Y})_G. \quad (126)$$

649 Next, we note that $\mathbf{Z} \cap An(\mathbf{X})_G \cap De(\mathbf{Y})_G = \phi$. To see this, suppose $\mathbf{Z} \cap An(\mathbf{X})_G \cap De(\mathbf{Y})_G \neq \phi$.

650 Let $\overset{\mathbf{V}'}{\dashrightarrow}$ denote a path through only variables of \mathbf{V}' . Then there exists a directed path such that
 651 $\mathbf{Y} \overset{\mathbf{V}'}{\dashrightarrow} \mathbf{Z} \overset{\mathbf{V}'}{\dashrightarrow} \mathbf{X}$, which contradicts the assumption $\mathbf{X} \subset An(\mathbf{Y})_G$.

652 From the above discussion, \mathbf{Z} can be divided into the three disjoint sets as follows:

$$\begin{aligned} \mathbf{Z} &= \mathbf{Z}_1 \cup \mathbf{Z}_2 \cup \mathbf{Z}_3 \\ \mathbf{Z}_1 &= (\mathbf{Z} \setminus De(\mathbf{X})_G) \cap An(\mathbf{Y})_G, \\ \mathbf{Z}_2 &= \mathbf{Z} \cap De(\mathbf{X})_G \cap An(\mathbf{Y})_G, \\ \mathbf{Z}_3 &= (\mathbf{Z} \setminus An(\mathbf{X})_G) \cap De(\mathbf{Y})_G. \end{aligned}$$

653 Then, each of the paths between \mathbf{Z}_1 , \mathbf{Z}_2 and \mathbf{Z}_3 is one of the following P1, P2 and P3:

654 P1. $\mathbf{Z}_1 \overset{\mathbf{V}'}{\dashrightarrow} \mathbf{Z}_2$,

655 P2. $\mathbf{Z}_2 \overset{\mathbf{V}'}{\dashrightarrow} \mathbf{Z}_3$,

656 P3. $\mathbf{Z}_1 \overset{\mathbf{V}'}{\dashrightarrow} \mathbf{Z}_3$.

657 In fact, if there exists a directed path in the opposite direction of P1, that is $\mathbf{Z}_2 \overset{\mathbf{V}'}{\dashrightarrow} \mathbf{Z}_1$, then there
 658 exists a path such that $X_i \overset{\mathbf{V}'}{\dashrightarrow} \mathbf{Z}_2 \overset{\mathbf{V}'}{\dashrightarrow} \mathbf{Z}_1$. This contradicts the assumption $\mathbf{Z}_1 \subset \mathbf{Z} \setminus De(\mathbf{X})_G$.

659 Similarly, if there exists a directed path in the opposite direction of P2, that is $\mathbf{Z}_3 \overset{\mathbf{V}'}{\dashrightarrow} \mathbf{Z}_2$, then
 660 there exists a path such that $Y_i \overset{\mathbf{V}'}{\dashrightarrow} \mathbf{Z}_3 \overset{\mathbf{V}'}{\dashrightarrow} \mathbf{Z}_2$, which contradicts the assumption $\mathbf{Z}_2 \subset An(\mathbf{Y})_G$.

661 In addition, if there exists a directed path in the opposite direction of P3, that is $\mathbf{Z}_3 \overset{\mathbf{V}'}{\dashrightarrow} \mathbf{Z}_1$, then
 662 there exists a path such that $Y_i \overset{\mathbf{V}'}{\dashrightarrow} \mathbf{Z}_3 \overset{\mathbf{V}'}{\dashrightarrow} \mathbf{Z}_1$, which contradicts the assumption $\mathbf{Z}_1 \subset An(\mathbf{Y})_G$.
 663 Therefore, all paths except P1, P2 and P3 are denied.

664 Hence, by marginalizing $P(\mathbf{V})$ for $\mathcal{X}_{\mathbf{V}'}$, we obtain

$$\begin{aligned} P(\mathbf{Y}, \mathbf{X}, \mathbf{Z}) &= \sum_{\mathcal{X}_{\mathbf{V}'}} P(\mathbf{V}) \\ &= P(\mathbf{Y}|\mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2) \cdot P(\mathbf{X}|\mathbf{Z}_1) \cdot P(\mathbf{Z}_1) \\ &\quad \times P(\mathbf{Z}_2|\mathbf{X}, \mathbf{Z}_1) \cdot P(\mathbf{Z}_3|\mathbf{Y}, \mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2). \end{aligned}$$

665 In addition, from (1), we have

$$\begin{aligned} P(\mathbf{Y}, \mathbf{Z}|do(\mathbf{X})) &= P(\mathbf{Y}|\mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2) \cdot P(\mathbf{Z}_1) \cdot P(\mathbf{Z}_2|\mathbf{X}, \mathbf{Z}_1) \\ &\quad \times P(\mathbf{Z}_3|\mathbf{Y}, \mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2). \end{aligned} \quad (127)$$

666 In the case that $\mathbf{Z}_3 = \phi$, by marginalizing out \mathbf{Y} of (127), we have

$$\begin{aligned} P(\mathbf{Z}|do(\mathbf{X})) &= \sum_{\mathbf{y} \in \mathcal{X}_{\mathbf{Y}}} P(\mathbf{Y} = \mathbf{y}, \mathbf{Z}|do(\mathbf{X})) \\ &= \sum_{\mathbf{y} \in \mathcal{X}_{\mathbf{Y}}} P(\mathbf{Y} = \mathbf{y}|\mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2) \cdot P(\mathbf{Z}_1) \cdot P(\mathbf{Z}_2|\mathbf{X}, \mathbf{Z}_1) \\ &= P(\mathbf{Z}_1) \cdot P(\mathbf{Z}_2|\mathbf{X}, \mathbf{Z}_1) \sum_{\mathbf{y} \in \mathcal{X}_{\mathbf{Y}}} P(\mathbf{Y} = \mathbf{y}|\mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2) \\ &= P(\mathbf{Z}_1) \cdot P(\mathbf{Z}_2|\mathbf{X}, \mathbf{Z}_1). \end{aligned}$$

667 On the other hand, in the case that $\mathbf{Z}_3 \neq \phi$, we obtain

$$\begin{aligned}
P(\mathbf{Z}|do(\mathbf{X})) &= \sum_{\mathbf{y} \in \mathcal{X}_{\mathbf{Y}}} P(\mathbf{Y} = \mathbf{y}, \mathbf{Z}|do(\mathbf{X})) \\
&= \sum_{\mathbf{y} \in \mathcal{X}_{\mathbf{Y}}} P(\mathbf{Y} = \mathbf{y}|\mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2) \cdot P(\mathbf{Z}_1) \\
&\quad \times P(\mathbf{Z}_2|\mathbf{X}, \mathbf{Z}_1) \cdot P(\mathbf{Z}_3|\mathbf{Y} = \mathbf{y}, \mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2) \\
&= P(\mathbf{Z}_1) \cdot P(\mathbf{Z}_2|\mathbf{X}, \mathbf{Z}_1) \\
&\quad \times \sum_{\mathbf{y} \in \mathcal{X}_{\mathbf{Y}}} P(\mathbf{Z}_3|\mathbf{Y} = \mathbf{y}, \mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2) \cdot P(\mathbf{Y} = \mathbf{y}|\mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2) \\
&= P(\mathbf{Z}_1) \cdot P(\mathbf{Z}_2|\mathbf{X}, \mathbf{Z}_1) \cdot P(\mathbf{Z}_3|\mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2).
\end{aligned}$$

668 Summarizing the above results, we have

$$P(\mathbf{Z}|do(\mathbf{X})) = \begin{cases} P(\mathbf{Z}_1) \cdot P(\mathbf{Z}_2|\mathbf{X}, \mathbf{Z}_1), & \text{if } \mathbf{Z}_3 = \phi, \\ P(\mathbf{Z}_1) \cdot P(\mathbf{Z}_2|\mathbf{X}, \mathbf{Z}_1) \cdot P(\mathbf{Z}_3|\mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2) & \text{if } \mathbf{Z}_3 \neq \phi. \end{cases} \quad (128)$$

669 Inserting (127) and (128) into (2), we see

$$\begin{aligned}
P(\mathbf{Y}|do(\mathbf{X}), \mathbf{Z}) &= \frac{P(\mathbf{Y}, \mathbf{Z}|do(\mathbf{X}))}{P(\mathbf{Z}|do(\mathbf{X}))} \\
&= \begin{cases} P(\mathbf{Y}|\mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2) & \text{if } \mathbf{Z}_3 = \phi, \\ \frac{P(\mathbf{Y}|\mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2)P(\mathbf{Z}_3|\mathbf{Y}, \mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2)}{P(\mathbf{Z}_3|\mathbf{X}, \mathbf{Z}_1, \mathbf{Z}_2)} & \text{if } \mathbf{Z}_3 \neq \phi. \end{cases} \quad (129)
\end{aligned}$$

670 Note that, $\mathbf{Z}_3 = \mathbf{Z} \cap De(\mathbf{Y})_G$, since $\mathbf{Z} \cap An(\mathbf{X})_G \cap De(\mathbf{Y})_G = \phi$.

671 Therefore, by rewriting \mathbf{Z}_3 as \mathbf{Z}_{De} and $\mathbf{Z}_1 \cup \mathbf{Z}_2$ as $\mathbf{Z} \setminus \mathbf{Z}_{De}$ for (129), we obtain (123).

672 This completes the proof. \square

673 Next, we provide the main theorem presented in Section 6.

674 **Theorem C.16** (Theorem 6.1 restated). *Given disjoint sets of $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$, $\mathbf{Y}, \mathbf{Z} \subset \mathbf{V}$*
675 *satisfying*

$$\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\} \subset An(\mathbf{Y})_G, \quad (130)$$

676 and

$$\mathbf{Z} \cap De(\mathbf{Y})_G = \phi. \quad (131)$$

677 Let $\mathbb{P} = P(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n, \mathbf{Z})$ and $\mathbb{Q} = P(\mathbf{X}_1) \times P(\mathbf{X}_2) \times \dots \times P(\mathbf{X}_n) \times P(\mathbf{Z})$, and $\tilde{P} =$
678 $P(\mathbf{Y}|do(\mathbf{X}), \mathbf{Z}) \times P(\mathbf{X}_1) \times P(\mathbf{X}_2) \times \dots \times P(\mathbf{X}_n) \times P(\mathbf{Z})$.

679 Suppose P satisfies Assumptions 1 and 2 in the above setting, and it holds that $E_{\mathbb{P}} \left[(d\mathbb{Q}/d\mathbb{P})^{1-\alpha} \right] <$
680 ∞ for some $0 < \alpha < 1$, then, for the optimal function T^* , such that

$$\begin{aligned}
T^*(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n, \mathbf{Z}) &= \arg \inf_{T \in \mathcal{T}^\alpha} \left\{ \frac{1}{\alpha} E_{\mathbb{Q}} [e^{\alpha \cdot T}] \right. \\
&\quad \left. + \frac{1}{1-\alpha} E_{\mathbb{P}} [e^{(\alpha-1) \cdot T}] \right\}, \quad (132)
\end{aligned}$$

681 it holds that

$$\frac{d\tilde{P}}{dP} = e^{-T^*(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n, \mathbf{Z})}. \quad (133)$$

682 Here, \mathcal{T}^α denotes the set of all non-constant functions $T(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}$ with $E_{\mathbb{P}}[e^{(\alpha-1) \cdot T(\mathbf{X})}] < \infty$.

683 *proof of Theorem C.16.* From Theorem C.15 and the assumption (131), we have

$$\begin{aligned}
\tilde{P} &= P(\mathbf{Y}|do(\mathbf{X}), \mathbf{Z}) \times P(\mathbf{X}_1) \times P(\mathbf{X}_2) \times \dots \times P(\mathbf{X}_n) \times P(\mathbf{Z}) \\
&= P(\mathbf{Y}|\mathbf{X}, \mathbf{Z}) \times P(\mathbf{X}_1) \times P(\mathbf{X}_2) \times \dots \times P(\mathbf{X}_n) \times P(\mathbf{Z}).
\end{aligned}$$

Thus, from Lemma C.2, we obtain

$$e^{-T^*(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n, \mathbf{Z})} = \frac{d\mathbb{Q}}{d\mathbb{P}} = \frac{d\tilde{P}}{dP}. \quad (134)$$

This completes the proof. \square

C.3 Proofs for Section 7

In this section, we first present a proposition for obtaining the density ratio between empirical distributions of the source and target distributions. Next, we present a proposition and lemmas for the early stopping method proposed in this study.

Proposition C.17. *It holds that*

$$\frac{d\hat{Q}^{(N)}}{d\hat{P}^{(N)}}(\mathbf{x}) = \begin{cases} dQ/dP(\mathbf{x}) & \text{if } 1 \leq \exists i \leq N \text{ s.t. } \mathbf{X}_i = \mathbf{x}, \\ 0 & \text{otherwise.} \end{cases} \quad (135)$$

proof of Proposition C.24. Let ν be the countable measure on $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N\}$:

$$\nu(\mathbf{x}) = \begin{cases} 1 & \text{if } 1 \leq \exists i \leq N \text{ s.t. } \mathbf{X}_i = \mathbf{x}, \\ 0 & \text{otherwise.} \end{cases} \quad (136)$$

Then, $\hat{P}^{(N)} \ll \nu$ and $\hat{Q}^{(N)} \ll \nu$.

Note that, from the definitions of $\hat{P}^{(N)}(\mathbf{x})$ and $\hat{Q}^{(N)}(\mathbf{x})$, we have

$$\hat{P}^{(N)}(\mathbf{x}) = \frac{1}{N} \sum_i \mathbf{1}(\mathbf{X}_{\sim P}^i = \mathbf{x}) = \frac{1}{N} \sum_i \mathbf{1}(\mathbf{X}^i = \mathbf{x}) \cdot \frac{dP}{d\mu}(\mathbf{x}), \quad (137)$$

and

$$\hat{Q}^{(N)}(\mathbf{x}) = \frac{1}{N} \sum_i \mathbf{1}(\mathbf{X}_{\sim Q}^i = \mathbf{x}) = \frac{1}{N} \sum_i \mathbf{1}(\mathbf{X}^i = \mathbf{x}) \cdot \frac{dQ}{d\mu}(\mathbf{x}), \quad (138)$$

where $\mathbf{1}(\cdot)$ equals one if the statement in parentheses is true and zero otherwise.

From (137) and (138), if $\mathbf{X}_i = \mathbf{x}$, we see

$$\frac{d\hat{P}^{(N)}}{d\nu}(\mathbf{x}) = \hat{P}^{(N)}(\mathbf{x}) = \frac{1}{N} \frac{dP}{d\mu}(\mathbf{x}), \quad (139)$$

and

$$\frac{d\hat{Q}^{(N)}}{d\nu}(\mathbf{x}) = \hat{Q}^{(N)}(\mathbf{x}) = \frac{1}{N} \frac{dQ}{d\mu}(\mathbf{x}). \quad (140)$$

Then, we have

$$\frac{d\hat{Q}^{(N)}}{d\hat{P}^{(N)}}(\mathbf{x}) = \frac{\frac{d\hat{P}^{(N)}}{d\nu}(\mathbf{x})}{\frac{d\hat{Q}^{(N)}}{d\nu}(\mathbf{x})} = \frac{dQ}{dP}(\mathbf{x}). \quad (141)$$

For $\mathbf{x} \notin \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N\}$, we observe $d\hat{Q}^{(N)}/d\nu(\mathbf{x}) = 0$. Note that, $d\hat{Q}^{(N)}/d\hat{P}^{(N)}(\mathbf{x})$ is defined as zero for $\mathbf{x} \in \Omega$ such that $d\hat{Q}^{(N)}/d\nu(\mathbf{x}) = 0$. Subsequently, we see $d\hat{Q}^{(N)}/d\hat{P}^{(N)}(\mathbf{x}) = 0$. \square

Next, we present a proposition for the early stopping method proposed in Section 7.1. We obtain an early stopping step as the step that minimizes the W_1 distance of the balanced distribution and target distribution, $\hat{Q}_k^{(N)}$ and Q in (22). To obtain the early stopping step, we assume that the two distributions differ the worst outside the neighborhood of the observations because we cannot know the closeness of the two distributions, $\hat{Q}_k^{(N)}$ and Q in (22) except in the neighborhood of the observations.

709 We now provide a note on the convergence rate for optimizing the loss function (16). Let

$$f^{(N)}(\mathbf{t}) = f^{(N)}(t_1, t_2, \dots, t_N) = \frac{1}{\alpha} \frac{1}{N} \sum_{i=1}^N e^{\alpha \cdot t_i} \cdot \frac{dQ}{d\mu}(\mathbf{x}_i) + \frac{1}{1-\alpha} \frac{1}{N} \sum_{i=1}^N e^{(\alpha-1) \cdot t_i} \cdot \frac{dP}{d\mu}(\mathbf{x}_i). \quad (142)$$

710 Subsequently, let \mathbf{t}_K denote a model at step K when optimizing (142) with a Stochastic Gradient Descent (SGD) algorithm. Because, from Corollary C.6, $f^{(N)}(\mathbf{t})$ is strongly convex with
 711 $\|\nabla f^{(N)}(\mathbf{t})\|^2 \leq D^2$ ($\exists D \in \mathbb{R}$) around the optimal point $\mathbf{t}_* = (t_*^1, t_*^2, \dots, t_*^N) = (-\log \frac{dQ}{dP}(\mathbf{x}_1),$
 712 $-\log \frac{dQ}{dP}(\mathbf{x}_2), \dots, -\log \frac{dQ}{dP}(\mathbf{x}_N))$, an $O(1/K)$ convergence rate can be achieved at step K when
 713 optimizing (142) with SGD algorithms under regular conditions for \mathbf{t} :
 714

$$E[f_N(\bar{\mathbf{t}}_K)] - f_N(\mathbf{t}_*) \leq \frac{C}{K+1}, \quad (143)$$

715 where $\bar{\mathbf{t}}_K$ is a weighted averaging such that $\bar{\mathbf{t}}_K = \frac{1}{(K+1)(K+2)} \sum_k (k+1) \cdot \mathbf{t}_k$ and $C > 0$ is
 716 constant. Here $E[\cdot]$ denotes the expectation for the randomness of batch sampling of SGD.⁶ As
 717 assumptions close to (143), we briefly assume (144) in Assumption E1 and (145) in Assumption E2
 718 to obtain an early stopping step, which are simpler and more relaxed than (143).

719 Herein, we make the following assumptions for the early stopping method presented in Section 7.

- 720 • Assumption E1. Let $\{T_k^{(N)}\}_{k=1}^\infty$ a sequence of functions in \mathcal{T}^α such that
 721 $\lim_{k \rightarrow \infty} T_k^{(N)}(\mathbf{X}_i) = -\log dQ/dP(\mathbf{X}_i)$, for $1 \leq \forall i \leq N$. Suppose that

$$\hat{L}_\alpha^{(N)}(Q, P; T_k^{(N)}) - \hat{L}_\alpha^{(N)}(Q, P; T_*) \leq \frac{C_0}{K}, \quad (144)$$

722 where $\hat{L}_\alpha^{(N)}(Q, P; \cdot)$ is defined as (60) in Lemma C.7 and $C_0 > 0$ is constant.

- 723 • Assumption E2. Let $\{T_k\}_{k=1}^\infty$ be a sequence of functions in \mathcal{T}^α with $E_P[e^{-T_k(\mathbf{X})}] = 1$
 724 such that $\lim_{k \rightarrow \infty} T_k = -\log dQ/dP$, P -almost everywhere. Suppose that

$$L_\alpha(Q, P; T_k) - L_\alpha(Q, P; T_*) \leq \frac{C_1}{K}, \quad (145)$$

725 where $L_\alpha(Q, P; \cdot)$ is defined as (57) in Lemma C.7 and $C_1 > 0$ is constant.

726 In addition, we make the following assumptions to simplify the discussion in the proofs.

- 727 • Assumption E3. Let Ω be a compact set in \mathbb{R}^d . Then λ denotes the Lebesgue measure on
 728 \mathbb{R}^d .
- 729 • Assumption E4. Let Q and P be two probabilities on Ω with continuous probability densi-
 730 ties $p(\mathbf{x})$ $q(\mathbf{x})$, respectively. Assume $0 < p_{\min} \leq p(\mathbf{x}) \leq p_{\max}$ and $0 < q_{\min} \leq q(\mathbf{x}) \leq$
 731 q_{\max} for all $\mathbf{x} \in \Omega$.
- 732 • Assumption E5. For $\{T_k^{(N)}\}_{k=1}^\infty$ in Assumption E1, assume that each function of $T_k^{(N)}(\mathbf{X})$
 733 is Lipschitz continuous: for $1 \leq k \leq \infty$,

$$|T_k^{(N)}(\mathbf{x}) - T_k^{(N)}(\mathbf{y})| \leq \rho_k \cdot \|\mathbf{x} - \mathbf{y}\|. \quad (146)$$

- 734 • Assumption E6. For $\{T_k^{(N)}\}_{k=1}^\infty$ in Assumption E2, assume that each function of $T_k(\mathbf{X})$ is
 735 Lipschitz continuous: for $1 \leq k \leq \infty$,

$$|T_k(\mathbf{x}) - T_k(\mathbf{y})| \leq \tilde{\rho}_k \cdot \|\mathbf{x} - \mathbf{y}\|. \quad (147)$$

736 Note that, the Lipschitz coefficient in Assumption E5 does not depend on the sample size N .

737

738

739 **Lemma C.18.** For $\{T_k^{(N)}\}_{k=1}^\infty$ in Assumption E5, it holds that for $\mathbf{x} \in \Omega$ and $\|\mathbf{y} - \mathbf{x}\| < D$,

$$e^{-T_k^{(N)}(\mathbf{y})} = e^{-T_k^{(N)}(\mathbf{x})} + e^{-T_k^{(N)}(\mathbf{x})} \cdot \{O(D) + O(D^2)\} + O_{\mathbf{x}}(D). \quad (148)$$

⁶For the convergence rate of SGD algorithms, for example, readers can refer to [12].

740 *proof of Lemma C.18.* From the intermediate value theorem for the second derivative of e^{-x} , we
 741 have

$$e^{-y} = e^{-x} - e^{-x} \cdot (y - x) + \frac{e^{-x+\theta \cdot (x-y)}}{2} \cdot (y - x)^2, \quad (149)$$

742 where $0 < \theta < 1$.

743 By substituting $y = T_k^{(N)}(\mathbf{y})$ and $x = T_k^{(N)}(\mathbf{x})$ into the above formula, we obtain

$$\begin{aligned} e^{-T_k^{(N)}(\mathbf{y})} &= e^{-T_k^{(N)}(\mathbf{x})} - e^{-T_k^{(N)}(\mathbf{x})} \left(T_k^{(N)}(\mathbf{y}) - T_k^{(N)}(\mathbf{x}) \right) \\ &\quad + \frac{e^{-T_k^{(N)}(\mathbf{x})+\theta(\mathbf{x},\mathbf{y}) \cdot (T_k^{(N)}(\mathbf{y})-T_k^{(N)}(\mathbf{x}))}}{2} \left(T_k^{(N)}(\mathbf{y}) - T_k^{(N)}(\mathbf{x}) \right)^2, \end{aligned} \quad (150)$$

744 where $0 < \theta(\mathbf{x}, \mathbf{y}) < 1$.

745 Now, note that, from Assumption E5,

$$T_k^{(N)}(\mathbf{y}) - T_k^{(N)}(\mathbf{x}) \leq \rho_k \cdot \|\mathbf{x} - \mathbf{y}\| \leq \rho_k \cdot D. \quad (151)$$

746 From (150) and (151), we see

$$\begin{aligned} \left| e^{-T_k^{(N)}(\mathbf{y})} - e^{-T_k^{(N)}(\mathbf{x})} \right| &= \left| e^{-T_k^{(N)}(\mathbf{x})} \cdot \left(T_k^{(N)}(\mathbf{y}) - T_k^{(N)}(\mathbf{x}) \right) \right. \\ &\quad \left. + \frac{e^{-T_k^{(N)}(\mathbf{x})+\theta(\mathbf{x},\mathbf{y}) \cdot (T_k^{(N)}(\mathbf{y})-T_k^{(N)}(\mathbf{x}))}}{2} \cdot \left(T_k^{(N)}(\mathbf{y}) - T_k^{(N)}(\mathbf{x}) \right)^2 \right| \\ &\leq e^{-T_k^{(N)}(\mathbf{x})} \cdot \rho_k \cdot D + \frac{e^{-T_k^{(N)}(\mathbf{x})+D}}{2} \cdot D^2 \\ &= e^{-T_k^{(N)}(\mathbf{x})} \cdot \{O(D) + O(D^2)\} + O_{\mathbf{x}}(D). \end{aligned} \quad (152)$$

747 Therefore, we have

$$e^{-T_k^{(N)}(\mathbf{y})} = e^{-T_k^{(N)}(\mathbf{x})} + e^{-T_k^{(N)}(\mathbf{x})} \cdot \{O(D) + O(D^2)\} + O_{\mathbf{x}}(D). \quad (153)$$

748 This completes the proof. \square

749 **Lemma C.19.** For $\{T_k\}_{k=1}^{\infty}$ in Assumption E6, it holds that for $\mathbf{x} \in \Omega$ and $\|\mathbf{y} - \mathbf{x}\| < D$,

$$e^{-T_k(\mathbf{y})} = e^{-T_k(\mathbf{x})} + e^{-T_k(\mathbf{x})} \cdot \{O(D) + O(D^2)\} + O_{\mathbf{x}}(D). \quad (154)$$

750 *proof of Lemma C.19.* Note that, we use only the Lipschitz continuity of $T_k^{(N)}(\mathbf{x})$ to prove Lemma
 751 C.18. From Assumption E6, $T_k(\mathbf{x})$ is Lipschitz continuous. Then, (154) can be proven in a manner
 752 similar to Lemma C.18.

753 This completes the proof. \square

754 **Lemma C.20.** Let $T_* = -\log dQ/dP$. Under Assumption E3 and E4, it holds that for $\mathbf{x} \in \Omega$ and
 755 $\|\mathbf{y} - \mathbf{x}\| < D$,

$$e^{-T_*}(\mathbf{y}) = e^{-T_*}(\mathbf{x}) - e^{-T_*}(\mathbf{x}) \cdot \{O(D) + O(D^2)\} + O_{\mathbf{x}}(D). \quad (155)$$

756 *proof of Lemma C.20.* Note that, we use only the Lipschitz continuity of $T_k^{(N)}(\mathbf{x})$ to prove Lemma
 757 C.18. From Assumption E3 and Assumption E4, $T_*(\mathbf{x})$ is a bounded continuous function on Ω .
 758 Since bounded continuous functions are Lipschitz continuous, $T_*(\mathbf{x})$ is Lipschitz continuous. Thus,
 759 (155) can be proven in a manner similar to Lemma C.18.

760 This completes the proof. \square

761 **Lemma C.21.** Let $B(\mathbf{x}_0, D) = \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}_0\| < D\}$. Then,

$$p_{\min} \cdot D^d \leq P(B(\mathbf{x}_0, D)) \leq p_{\max} \cdot D^d. \quad (156)$$

762 *proof of Lemma C.21.* From Assumption E4, $p_{\min} \leq p(\mathbf{x}) \leq p_{\max}$ holds, and by integrating over
 763 $B(\mathbf{x}_0, D)$ with λ , we obtain

$$\begin{aligned} \int_{B(\mathbf{x}_0, D)} p_{\min} d\lambda &\leq P(B(\mathbf{x}_0, D)) \leq \int_{B(\mathbf{x}_0, D)} p_{\max} d\lambda \\ \therefore p_{\min} \cdot D^d &\leq P(B(\mathbf{x}_0, D)) \leq p_{\max} \cdot D^d. \end{aligned} \quad (157)$$

764 This completes the proof. \square

765 **Lemma C.22.** Let $B(\mathbf{x}_0, D) = \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}_0\| < D\}$. Then,

$$P(B(\mathbf{x}_0, D)) = C \cdot p(\mathbf{x}_0) \cdot D^d, \quad (158)$$

766 where C is constant.

767 *proof of Lemma C.22.* From Assumption E4, p is a bounded continuous function on Ω . Since
 768 bounded continuous functions are Lipschitz continuous, $p(\mathbf{x})$ is Lipschitz continuous.

769 Then, there exist a constant C such that

$$p(\mathbf{x}) \leq p(\mathbf{x}_0) + C \cdot \|\mathbf{x} - \mathbf{x}_0\|, \quad (159)$$

770 and by integrating over $B(\mathbf{x}_0, D)$ with λ , we obtain

$$P(B(\mathbf{x}_0, D)) \leq C \cdot p(\mathbf{x}_0) \cdot D^d. \quad (160)$$

771 This completes the proof. \square

772 **Proposition C.23.** For $\{T_k^{(N)}\}_{k=1}^\infty$ in Assumption E1, let $\hat{Q}_k^{(N)}$ be a probability defined as

$$d\hat{Q}_k^{(N)} = e^{-T_k^{(N)}} \cdot dP.$$

773 Then, under Assumption E1-E6, for a sufficiently large $K > 0$, it holds that

$$E_{\mathbf{X}_P^{(N)}}[W_1(Q, \hat{Q}_K^{(N)})] \leq 2 - N \cdot K^{-\frac{d}{2}} + K^{-\frac{1}{2}}. \quad (161)$$

774 **Corollary C.24.** Let $K_0 = N^{\frac{2}{d+\delta}}$ with $\delta > 0$. Then, under Assumption E1-E6, for a sufficiently
 775 large N , it holds that

$$E_{\mathbf{X}_P^{(N)}}[W_1(Q, \hat{Q}_{K_0}^{(N)})] \leq 2 - K_0^{\frac{\delta}{2}} + K_0^{-\frac{1}{2}}. \quad (162)$$

776 **Corollary C.25.** In Corollary C.24, let $\delta' > 0$ such that $N^{\frac{\delta'}{d+\delta'}} = 2$, and let $K_0 = N^{\frac{2}{d+\delta'}}$. Then,
 777 under Assumption E1-E6, for a sufficiently large N , it holds that

$$E_{\mathbf{X}_P^{(N)}}[W_1(Q, \hat{Q}_{K_0}^{(N)})] \leq K_0^{-\frac{1}{2}}. \quad (163)$$

778 Thus, if $N > \left(\frac{1}{\varepsilon}\right)^{d+\delta'}$ then $W_1(Q, \hat{Q}_{K_0}^{(N)}) < \varepsilon$.

779 *proof of Proposition C.23.* Let Q_K be a probability defined as

$$dQ_K = e^{-T_K} \cdot dP. \quad (164)$$

780 Intuitively, Q_K is the true balanced probability distribution at a step K .

781 First, from the triangle inequality for the L_1 norm, we have

$$E_\mu \left| \hat{Q}_K^{(N)} - Q \right| \leq E_\mu \left| \hat{Q}_K^{(N)} - Q_K \right| + E_\mu \left| Q_K - Q \right|. \quad (165)$$

782 Considering the expectation $E_{\mathbf{X}_P^{(N)}}[\cdot]$ for the both sides of the above equation, we see

$$E_{\mathbf{X}_P^{(N)}} \left[E_\mu \left| \hat{Q}_K^{(N)} - Q \right| \right] \leq E_{\mathbf{X}_P^{(N)}} \left[E_\mu \left| \hat{Q}_K^{(N)} - Q_K \right| \right] + E_{\mathbf{X}_P^{(N)}} \left[E_\mu \left| Q_K - Q \right| \right]. \quad (166)$$

783 Next, we obtain the upper bound of the first term in (166).

784 Let $\Delta_i = B(\mathbf{X}_i, 1/\sqrt{K})$. Subsequently, let $\Delta = \bigcup_{i=1}^N \Delta_i$. Then, we have

$$\begin{aligned}
& E_\mu \left| \hat{Q}_K^{(N)} - Q_K \right| \\
&= \int \left| e^{-T_k^{(N)}} \cdot \frac{dP}{d\mu} - e^{-T_k} \cdot \frac{dP}{d\mu} \right| d\mu \\
&= \int \left| e^{-T_k^{(N)}} - e^{-T_k} \right| \frac{dP}{d\mu} d\mu \\
&= E_P \left| e^{-T_k^{(N)}} - e^{-T_k} \right| \\
&= E_P \left[id_\Delta \left| e^{-T_k^{(N)}} - e^{-T_k} \right| \right] + E_P \left[id_{\Omega \setminus \Delta} \left| e^{-T_k^{(N)}} - e^{-T_k} \right| \right] \\
&\leq E_P \left[id_\Delta \left| e^{-T_k^{(N)}} - e^{-T_k} \right| \right] + E_P \left[id_{\Omega \setminus \Delta} \left| e^{-T_k^{(N)}} \right| \right] + E_P \left[id_{\Omega \setminus \Delta} \left| e^{-T_k} \right| \right].
\end{aligned}$$

785 Considering the expectation $E_{\mathbf{X}_P^{(N)}}[\cdot]$ for the both sides of the above equation, we see

$$\begin{aligned}
& E_{\mathbf{X}_P^{(N)}} \left[E_\mu \left| \hat{Q}_K^{(N)} - Q_K \right| \right] \\
&\leq E_{\mathbf{X}_P^{(N)}} \left[E_P \left[id_\Delta \left| e^{-T_k^{(N)}} - e^{-T_k} \right| \right] \right] \\
&\quad + E_{\mathbf{X}_P^{(N)}} \left[E_P \left[id_{\Omega \setminus \Delta} \left| e^{-T_k^{(N)}} \right| \right] \right] + E_{\mathbf{X}_P^{(N)}} \left[E_P \left[id_{\Omega \setminus \Delta} \left| e^{-T_k} \right| \right] \right]. \quad (167)
\end{aligned}$$

786 To obtain the upper bound of the first term in (167), we see

$$\begin{aligned}
& E_P \left[id_\Delta \left| e^{-T_k^{(N)}} - e^{-T_k} \right| \right] \\
&= E_P \left[\sum_{i=1}^N id_{\Delta_i} \left| e^{-T_k^{(N)}} - e^{-T_k} \right| \right] \\
&= \sum_{i=1}^N E_P \left[id_{\Delta_i} \left| e^{-T_k^{(N)}} - e^{-T_k} \right| \right] \\
&= \sum_{i=1}^N E_P \left[id_{B(\mathbf{X}_i, 1/\sqrt{K})} \left| e^{-T_k^{(N)}} - e^{-T_k} \right| \right] \\
&= E_P \left[\left| \sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})} \cdot e^{-T_k^{(N)}} - \sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})} \cdot e^{-T_k(\mathbf{X}_i)} \right| \right]. \quad (168)
\end{aligned}$$

787 Subsequently, we have

$$\begin{aligned}
& E_P \left[\left| \sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})} \cdot e^{-T_k^{(N)}} - \sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})} \cdot e^{-T_k(\mathbf{X}_i)} \right| \right] \\
= & E_P \left[\left| \sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \cdot e^{-T_k^{(N)}(\mathbf{x})} - \sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \cdot e^{-T_k^{(N)}(\mathbf{X}_i)} \right. \right. \\
& + \sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \cdot e^{-T_k^{(N)}(\mathbf{X}_i)} - \sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \cdot e^{-T_*(\mathbf{X}_i)} \\
& + \sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \cdot e^{-T_*(\mathbf{X}_i)} - \sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \cdot e^{-T_*(\mathbf{x})} \\
& \left. \left. + \sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \cdot e^{-T_*(\mathbf{x})} - \sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \cdot e^{-T_k(\mathbf{x})} \right| \right] \\
= & E_P \left[\left| \sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \left\{ e^{-T_k^{(N)}(\mathbf{x})} - e^{-T_k^{(N)}(\mathbf{X}_i)} \right\} \right. \right. \\
& + \sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})} \left\{ e^{-T_k^{(N)}(\mathbf{X}_i)} - e^{-T_*(\mathbf{X}_i)} \right\} \\
& + \sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})} \left\{ e^{-T_*(\mathbf{X}_i)} - e^{-T_*(\mathbf{x})} \right\} \\
& \left. \left. + \sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \left\{ e^{-T_*(\mathbf{x})} - e^{-T_k(\mathbf{x})} \right\} \right| \right] \\
\leq & E_P \left[\sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_k^{(N)}(\mathbf{x})} - e^{-T_k^{(N)}(\mathbf{X}_i)} \right| \right] \\
& + E_P \left[\sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_k^{(N)}(\mathbf{X}_i)} - e^{-T_*(\mathbf{X}_i)} \right| \right] \\
& + E_P \left[\sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_*(\mathbf{X}_i)} - e^{-T_*(\mathbf{x})} \right| \right] \\
& + E_P \left[\sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_*(\mathbf{x})} - e^{-T_k(\mathbf{x})} \right| \right].
\end{aligned}$$

788 Considering the expectation $E_{\mathbf{X}_P^{(N)}}[\cdot]$ for the both sides of the above equation, we obtain

$$\begin{aligned}
& E_{\mathbf{X}_P^{(N)}} \left[E_P \left[id_{\Delta} \left| e^{-T_k^{(N)}} - e^{-T_k} \right| \right] \right] \\
\leq & E_{\mathbf{X}_P^{(N)}} \left[E_P \left[\sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_k^{(N)}(\mathbf{x})} - e^{-T_k^{(N)}(\mathbf{X}_i)} \right| \right] \right] \\
& + E_{\mathbf{X}_P^{(N)}} \left[E_P \left[\sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_k^{(N)}(\mathbf{X}_i)} - e^{-T_*(\mathbf{X}_i)} \right| \right] \right] \\
& + E_{\mathbf{X}_P^{(N)}} \left[E_P \left[\sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_*(\mathbf{X}_i)} - e^{-T_*(\mathbf{x})} \right| \right] \right] \\
& + E_{\mathbf{X}_P^{(N)}} \left[E_P \left[\sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_*(\mathbf{x})} - e^{-T_k(\mathbf{x})} \right| \right] \right]. \quad (169)
\end{aligned}$$

789 Now, from Lemma C.19, we have, for $\mathbf{x} \in id_{B(\mathbf{X}_i, 1/\sqrt{K})}$

$$\left| e^{-T_k^{(N)}(\mathbf{x})} - e^{-T_k^{(N)}(\mathbf{X}_i)} \right| = e^{-T_k^{(N)}(\mathbf{X}_i)} \left\{ O\left(\frac{1}{\sqrt{K}}\right) + O\left(\frac{1}{K}\right) \right\} + O_{\mathbf{X}_i}\left(\frac{1}{K}\right).$$

790 Then, we see

$$\begin{aligned} & E_P \left[\sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_k^{(N)}(\mathbf{x})} - e^{-T_k^{(N)}(\mathbf{X}_i)} \right| \right] \\ &= \sum_{i=1}^N E_P \left[id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \left\{ e^{-T_k^{(N)}(\mathbf{X}_i)} \left\{ O\left(\frac{1}{\sqrt{K}}\right) + O\left(\frac{1}{K}\right) \right\} + O_{\mathbf{X}_i}\left(\frac{1}{\sqrt{K}}\right) \right\} \right] \\ &= \sum_{i=1}^N E_P \left[id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \left\{ e^{-T_k^{(N)}(\mathbf{X}_i)} \left\{ O\left(\frac{1}{\sqrt{K}}\right) + O\left(\frac{1}{K}\right) \right\} + O_{\mathbf{X}_i}\left(\frac{1}{\sqrt{K}}\right) \right\} \right] \\ &= \sum_{i=1}^N P\left(B(\mathbf{X}_i, 1/\sqrt{K})\right) \left\{ e^{-T_k^{(N)}(\mathbf{X}_i)} \left\{ O\left(\frac{1}{\sqrt{K}}\right) + O\left(\frac{1}{K}\right) \right\} + O_{\mathbf{X}_i}\left(\frac{1}{\sqrt{K}}\right) \right\} \\ &= \sum_{i=1}^N P\left(B(\mathbf{X}_i, 1/\sqrt{K})\right) \cdot O_{\mathbf{X}_i}\left(\frac{1}{\sqrt{K}}\right) \\ &\leq \sum_{i=1}^N p_{max} \cdot O_{\mathbf{X}_i}\left(\frac{1}{(\sqrt{K})^d}\right) \cdot O_{\mathbf{X}_i}\left(\frac{1}{\sqrt{K}}\right) \\ &= \sum_{i=1}^N O_{\mathbf{X}_i}\left(K^{-\frac{d+1}{2}}\right). \end{aligned} \tag{170}$$

791 Here, we obtain (170) by using Lemma C.21.

792 Considering the expectation $E_{\mathbf{X}_P^{(N)}}[\cdot]$ for the both sides of the above equation, we have

$$\begin{aligned} & E_{\mathbf{X}_P^{(N)}} \left[E_P \left[\sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_k^{(N)}(\mathbf{x})} - e^{-T_k^{(N)}(\mathbf{X}_i)} \right| \right] \right] \\ &= E_{\mathbf{X}_P^{(N)}} \left[\sum_{i=1}^N O_{\mathbf{X}_i}\left(K^{-\frac{d+1}{2}}\right) \right] \\ &= \sum_{i=1}^N O\left(K^{-\frac{d+1}{2}}\right) \\ &= N \cdot O\left(K^{-\frac{d+1}{2}}\right). \end{aligned} \tag{171}$$

793 In addition, from Lemma C.19 and C.20, it holds that, for $\mathbf{x} \in id_{B(\mathbf{X}_i, 1/\sqrt{K})}$,

$$\left| e^{-T_k(\mathbf{x})} - e^{-T_k(\mathbf{X}_i)} \right| = e^{-T_k(\mathbf{X}_i)} \left\{ O\left(\frac{1}{\sqrt{K}}\right) + O\left(\frac{1}{K}\right) \right\} + O_{\mathbf{X}_i}\left(\frac{1}{K}\right),$$

794 and

$$\left| e^{-T_*(\mathbf{x})} - e^{-T_*(\mathbf{X}_i)} \right| = e^{-T_*(\mathbf{X}_i)} \left\{ O\left(\frac{1}{\sqrt{K}}\right) + O\left(\frac{1}{K}\right) \right\} + O_{\mathbf{X}_i}\left(\frac{1}{K}\right).$$

795 In the similar manner to obtain (C.23), it holds that, for $\mathbf{x} \in id_{B(\mathbf{X}_i, 1/\sqrt{K})}$, we obtain

$$E_{\mathbf{X}_P^{(N)}} \left[E_P \left[\sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_k(\mathbf{x})} - e^{-T_k(\mathbf{X}_i)} \right| \right] \right] = N \cdot O\left(K^{-\frac{d+1}{2}}\right), \tag{172}$$

796 and

$$E_{\mathbf{X}_P^{(N)}} \left[E_P \left[\sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_*(\mathbf{x})} - e^{-T_*(\mathbf{X}_i)} \right| \right] \right] = N \cdot O \left(K^{-\frac{d+1}{2}} \right). \quad (173)$$

797 Here, we obtain the upper bounds of the first, third, and fourth terms in (169).

798 We now have the upper bounds of the second term in (169).

799 First, we obtain

$$\begin{aligned} & E_P \left[\sum_{i=1}^N id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_k^{(N)}(\mathbf{X}_i)} - e^{-T_*(\mathbf{X}_i)} \right| \right] \\ &= \sum_{i=1}^N E_P \left[id_{B(\mathbf{X}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_k^{(N)}(\mathbf{X}_i)} - e^{-T_*(\mathbf{X}_i)} \right| \right] \\ &= \sum_{i=1}^N P \left(B(\mathbf{X}_i, 1/\sqrt{K}) \left| e^{-T_k^{(N)}(\mathbf{X}_i)} - e^{-T_*(\mathbf{X}_i)} \right| \right). \end{aligned} \quad (174)$$

800 Then, from Lemma C.22, we have

$$\begin{aligned} & \sum_{i=1}^N P \left(B(\mathbf{X}_i, 1/\sqrt{K}) \left| e^{-T_k^{(N)}(\mathbf{X}_i)} - e^{-T_*(\mathbf{X}_i)} \right| \right) \\ &= C \sum_{i=1}^N p(\mathbf{X}_i) \cdot O \left(\frac{1}{(\sqrt{K})^d} \right) \left| e^{-T_k^{(N)}(\mathbf{X}_i)} - e^{-T_*(\mathbf{X}_i)} \right|. \end{aligned} \quad (175)$$

801 Next, note that,

$$\sum_{i=1}^N p(\mathbf{X}_i) \cdot \left| e^{-T_k^{(N)}(\mathbf{X}_i)} - e^{-T_*(\mathbf{X}_i)} \right| = N \cdot E_\nu \left| \hat{Q}_k^{(N)} - Q^{(N)} \right|. \quad (176)$$

802 Here, ν is the countable measure on $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N\}$ defined as (136).

803 In addition, since

$$\hat{L}_\alpha^{(N)}(Q, P; T) = L_\alpha(\hat{Q}^{(N)}, \hat{P}^{(N)}; T),$$

804 holds, we obtain, from Proposition C.13 and Assumption E1,

$$\begin{aligned} \sqrt{\frac{\alpha}{2}} \cdot E_\nu \left| \hat{Q}_K^{(N)} - Q^{(N)} \right| &\leq \sqrt{L_\alpha(Q^{(N)}, P^{(N)}; T_K^{(N)}) - L_\alpha(Q^{(N)}, P^{(N)}; T_*)} \\ &= \sqrt{\hat{L}_\alpha^{(N)}(Q, P; T_K^{(N)}) - \hat{L}_\alpha^{(N)}(Q, P; T_*)} \\ &= O \left(\frac{1}{\sqrt{K}} \right). \end{aligned} \quad (177)$$

805 Finally, from (174), (175), (176), and (177), we have

$$\begin{aligned}
& E_P \left[\sum_{i=1}^N id_{B(\mathbf{x}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_k^{(N)}(\mathbf{x}_i)} - e^{-T_*(\mathbf{x}_i)} \right| \right] \\
& \leq O \left(\frac{1}{(\sqrt{K})^d} \right) \sum_{i=1}^N p(\mathbf{x}_i) \left| e^{-T_k^{(N)}(\mathbf{x}_i)} - e^{-T_*(\mathbf{x}_i)} \right| \\
& \leq O \left(\frac{1}{(\sqrt{K})^d} \right) \cdot N \cdot E_\nu \left| \hat{Q}_k^{(N)} - Q \right| \\
& \leq O \left(\frac{1}{(\sqrt{K})^d} \right) \cdot N \cdot O \left(\frac{1}{\sqrt{K}} \right) \\
& = N \cdot O \left(K^{-\frac{d+1}{2}} \right). \tag{178}
\end{aligned}$$

806 Considering the expectation $E_{\mathbf{X}_P^{(N)}}[\cdot]$ for the both sides of the above equation, we have

$$\begin{aligned}
& E_{\mathbf{X}_P^{(N)}} \left[E_P \left[\sum_{i=1}^N id_{B(\mathbf{x}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_k^{(N)}(\mathbf{x}_i)} - e^{-T_*(\mathbf{x}_i)} \right| \right] \right] \\
& = E_{\mathbf{X}_P^{(N)}} \left[N \cdot O \left(K^{-\frac{d+1}{2}} \right) \right] \\
& = N \cdot O \left(K^{-\frac{d+1}{2}} \right). \tag{179}
\end{aligned}$$

807 Summarizing (169), (171), (172), (173), and (179),

$$\begin{aligned}
& E_{\mathbf{X}_P^{(N)}} \left[E_P \left[id_\Delta \left| e^{-T_k^{(N)}} - e^{-T_k} \right| \right] \right] \\
& \leq E_{\mathbf{X}_P^{(N)}} \left[E_P \left[\sum_{i=1}^N id_{B(\mathbf{x}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_k^{(N)}(\mathbf{x})} - e^{-T_k^{(N)}(\mathbf{x}_i)} \right| \right] \right] \\
& \quad + E_{\mathbf{X}_P^{(N)}} \left[E_P \left[\sum_{i=1}^N id_{B(\mathbf{x}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_k^{(N)}(\mathbf{x}_i)} - e^{-T_*(\mathbf{x}_i)} \right| \right] \right] \\
& \quad + E_{\mathbf{X}_P^{(N)}} \left[E_P \left[\sum_{i=1}^N id_{B(\mathbf{x}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_*(\mathbf{x}_i)} - e^{-T_*(\mathbf{x})} \right| \right] \right] \\
& \quad + E_{\mathbf{X}_P^{(N)}} \left[E_P \left[\sum_{i=1}^N id_{B(\mathbf{x}_i, 1/\sqrt{K})}(\mathbf{x}) \left| e^{-T_*(\mathbf{x})} - e^{-T_k(\mathbf{x})} \right| \right] \right] \\
& = N \cdot O \left(K^{-\frac{d+1}{2}} \right) + N \cdot O \left(K^{-\frac{d+1}{2}} \right) + N \cdot O \left(K^{-\frac{d+1}{2}} \right) + N \cdot O \left(K^{-\frac{d+1}{2}} \right) \\
& = N \cdot O \left(K^{-\frac{d+1}{2}} \right). \tag{180}
\end{aligned}$$

808 Here, we see the upper bound of the first term in (167).

809 Next, we obtain the upper bound of the second and third term in (167).

810 First, we obtain the upper bound of the second term in (167). Now, we have

$$\begin{aligned}
& E_{\mathbf{X}_P^{(N)}} \left[E_P \left[id_{\Omega \setminus \Delta} \left| e^{-T_k} \right| \right] \right] \\
&= E_{\mathbf{X}_P^{(N)}} \left[E_P \left[id_{\Omega \setminus \bigcup_{i=1}^N \Delta_i}(\mathbf{x}) \cdot \left| e^{-T_k^{(N)}} \right| \right] \right] \\
&= E_{\mathbf{X}_P^{(N)}} \left[E_P \left[\left\{ 1 - id_{\bigcup_{i=1}^N \Delta_i}(\mathbf{x}) \right\} \cdot e^{-T_k^{(N)}} \right] \right] \\
&= E_{\mathbf{X}_P^{(N)}} \left[1 - E_P \left[\sum_{i=1}^N id_{B(\mathbf{x}_i, 1/\sqrt{K})}(\mathbf{x}) \cdot e^{-T_k^{(N)}} \right] \right] \\
&= E_{\mathbf{X}_P^{(N)}} \left[1 - \sum_{i=1}^N E_P \left[id_{B(\mathbf{x}_i, 1/\sqrt{K})}(\mathbf{x}) \cdot e^{-T_k^{(N)}} \right] \right]. \tag{181}
\end{aligned}$$

811 Then, from Lemma C.18 and Lemma C.21, we have

$$\begin{aligned}
& E_P \left[id_{B(\mathbf{x}_i, 1/\sqrt{K})}(\mathbf{x}) \cdot e^{-T_k^{(N)}(\mathbf{x})} \right] \\
&= E_P \left[id_{B(\mathbf{x}_i, 1/\sqrt{K})}(\mathbf{x}) \cdot \left\{ e^{-T_k^{(N)}(\mathbf{x}_i)} + e^{-T_k^{(N)}(\mathbf{x}_i)} \cdot \left\{ O\left(\frac{1}{\sqrt{K}}\right) + O\left(\frac{1}{K}\right) \right\} + O_{\mathbf{x}}\left(\frac{1}{\sqrt{K}}\right) \right\} \right] \\
&= E_P \left[id_{B(\mathbf{x}_i, 1/\sqrt{K})}(\mathbf{x}) \cdot \left\{ e^{-T_k^{(N)}(\mathbf{x}_i)} + e^{-T_k^{(N)}(\mathbf{x}_i)} \cdot \left\{ O\left(\frac{1}{\sqrt{K}}\right) + O\left(\frac{1}{K}\right) \right\} + O_{\mathbf{x}}\left(\frac{1}{\sqrt{K}}\right) \right\} \right] \\
&= P\left(B(\mathbf{x}_i, 1/\sqrt{K})\right) \cdot \left\{ e^{-T_k^{(N)}(\mathbf{x}_i)} + e^{-T_k^{(N)}(\mathbf{x}_i)} \cdot \left\{ O\left(\frac{1}{\sqrt{K}}\right) + O\left(\frac{1}{K}\right) \right\} + O_{\mathbf{x}}\left(\frac{1}{\sqrt{K}}\right) \right\} \\
&\geq p_{min} \cdot O_{\mathbf{x}_i} \left(\frac{1}{(\sqrt{K})^d} \right) \cdot \left\{ e^{-T_k^{(N)}(\mathbf{x}_i)} + e^{-T_k^{(N)}(\mathbf{x}_i)} \cdot \left\{ O\left(\frac{1}{\sqrt{K}}\right) + O\left(\frac{1}{K}\right) \right\} + O_{\mathbf{x}}\left(\frac{1}{\sqrt{K}}\right) \right\} \\
&= O_{\mathbf{x}_i} \left(K^{-\frac{d}{2}} \right) + O_{\mathbf{x}_i} \left(K^{-\frac{d+1}{2}} \right). \tag{182}
\end{aligned}$$

812 From (181) and (182), we see

$$\begin{aligned}
& E_{\mathbf{X}_P^{(N)}} \left[E_P \left[id_{\Omega \setminus \Delta} \left| e^{-T_k^{(N)}} \right| \right] \right] \\
&= E_{\mathbf{X}_P^{(N)}} \left[1 - \sum_{i=1}^N E_P \left[id_{B(\mathbf{x}_i, 1/\sqrt{K})}(\mathbf{x}) \cdot e^{-T_k^{(N)}} \right] \right] \\
&\leq E_{\mathbf{X}_P^{(N)}} \left[1 - \sum_{i=1}^N O_{\mathbf{x}_i} \left(K^{-\frac{d}{2}} \right) + O_{\mathbf{x}_i} \left(K^{-\frac{d+1}{2}} \right) \right] \\
&= 1 - N \cdot \left\{ O\left(K^{-\frac{d}{2}}\right) + O\left(K^{-\frac{d+1}{2}}\right) \right\} \\
&= 1 - N \cdot O\left(K^{-\frac{d}{2}}\right) \tag{183}
\end{aligned}$$

813 In the similar manner to obtain (183), we obtain the upper bound of the third term in (167):

$$E_{\mathbf{X}_P^{(N)}} \left[E_P \left[id_{\Omega \setminus \Delta} \left| e^{-T_k} \right| \right] \right] = 1 - N \cdot O\left(K^{-\frac{d}{2}}\right). \tag{184}$$

814 Summarizing (167), (180), (183) and (184), we have

$$\begin{aligned}
& E_{\mathbf{X}_P^{(N)}} \left[E_{\mu} \left| \hat{Q}_K^{(N)} - Q_K \right| \right] \\
&\leq N \cdot O\left(K^{-\frac{d+1}{2}}\right) + \left\{ 1 - N \cdot O\left(K^{-\frac{d}{2}}\right) \right\} + \left\{ 1 - N \cdot O\left(K^{-\frac{d}{2}}\right) \right\} \\
&= 2 - N \cdot O\left(K^{-\frac{d}{2}}\right). \tag{185}
\end{aligned}$$

815 For the upper bound of the second term in (166), from Propsition C.13 we obtain

$$\sqrt{\frac{\alpha}{2}} E_\mu |Q_K - Q| \leq \sqrt{L_\alpha(Q, P; T_K) - L_\alpha(Q, P; T_*)}. \quad (186)$$

816 Thus, under Assumption E2, we see

$$E_\mu |Q_K - Q| \leq \frac{C'_0}{\sqrt{K}}, \quad (187)$$

817 where $C'_0 = \sqrt{(2 \cdot C_0)/\alpha}$.

818 Considering the expectation $E_{\mathbf{X}_P^{(N)}}[\cdot]$ for the both sides of the above equation, we have

$$E_{\mathbf{X}_P^{(N)}} [E_\mu |Q_K - Q|] = O(K^{-\frac{1}{2}}). \quad (188)$$

819 Finally, (165), (185), and (188), we have

$$\begin{aligned} E_{\mathbf{X}_P^{(N)}} [W_1(Q, \hat{Q}_{K_0}^{(N)})] &\leq E_{\mathbf{X}_P^{(N)}} [E_\mu |\hat{Q}_K^{(N)} - Q|] \\ &\leq E_{\mathbf{X}_P^{(N)}} [E_\mu |\hat{Q}_K^{(N)} - Q_K|] + E_{\mathbf{X}_P^{(N)}} [E_\mu |Q_K - Q|] \\ &= 2 - N \cdot O(K^{-\frac{d}{2}}) + O(K^{-\frac{1}{2}}). \end{aligned} \quad (189)$$

820 From this, for sufficiently large $K > 0$, we see

$$E_{\mathbf{X}_P^{(N)}} [W_1(Q, \hat{Q}_{K_0}^{(N)})] \leq 2 - N \cdot K^{-\frac{d}{2}} + K^{-\frac{1}{2}}.$$

821 Here, we show (161).

822 This completes the proof. \square

823 *proof of Corollary C.24.* For 161, substituting K_0 for K , and $N = K_0^{\frac{d+\delta}{2}}$, we have

$$\begin{aligned} E_{\mathbf{X}_P^{(N)}} [W_1(Q, \hat{Q}_{K_0}^{(N)})] &\leq 2 - K_0^{\frac{d+\delta}{2}} \cdot K_0^{-\frac{d}{2}} + K_0^{-\frac{1}{2}} \\ &= 2 - K_0^{\frac{\delta}{2}} + K_0^{-\frac{1}{2}}. \end{aligned} \quad (190)$$

824 This completes the proof. \square

825 *proof of Corollary C.25.* For the setting of the proposition, we have $K_0^{\frac{\delta'}{2}} = N^{\frac{\delta'}{d+\delta'}} = 2$. Thus, for
826 190, we see

$$\begin{aligned} E_{\mathbf{X}_P^{(N)}} [W_1(Q, \hat{Q}_{K_0}^{(N)})] &\leq 2 - K_0^{\frac{d+\delta}{2}} \cdot K_0^{-\frac{d}{2}} + K_0^{-\frac{1}{2}} \\ &= 2 - 2 + K_0^{-\frac{1}{2}} \\ &= K_0^{-\frac{1}{2}}. \end{aligned}$$

827 This completes the proof. \square

D Numerical Experiments

In this section, we report the results of numerical experiments conducted in this study.

D.1 Experiments on convergence for different values of α

In this section, we report the results of the numerical experiments related to the discussion in Section 5: the results of the numerical experiments on the convergence of learning for different values of α are presented.

Experimental Setup. For $\alpha = -3, -2, -1, 0.2, 0.5, 0.8, 2.0, 3.0$, and 4.0 , we generated training and test dataset, and then trained an NGB model with the training dataset while estimating the α divergence at each learning step with the test dataset. One hundred numerical simulations were performed for each α . As a result of the experiment, the median of the estimated value and ranges between the 45th and 55th percentile quartiles and between the 5th and 95th percentile quartiles at each learning step are reported.

Synthetic Data. We generated synthetic data of size 5000 from 5-dimensional normal distribution $\{X_1, X_2, \dots, X_5\}$ such that $E[X_i] = 0$, $\text{Var}[X_i] = 1$ and $E[X_i \cdot X_j] = 0.8$ ($i \neq j$), for each of the training and test datasets.

Estimating the α divergence. The α divergence was estimated in the following way

$$\hat{D}_\alpha(Q||P)(t) = \frac{1}{\alpha \cdot (1 - \alpha)} - \frac{1}{\alpha} \hat{E}_Q \left[e^{\alpha \cdot T_{\theta_t}(\mathbf{x}^{te})} \right] - \frac{1}{1 - \alpha} \hat{E}_P \left[e^{(\alpha-1) \cdot T_{\theta_t}(\mathbf{x}^{te})} \right] \quad (191)$$

$$= \frac{1}{\alpha \cdot (1 - \alpha)} - \mathcal{L}_\alpha(\theta_t) \quad (192)$$

where T_{θ_t} is a model at learning step t in Algorithm 1 and \mathbf{x}^{te} denotes the test dataset. Note that, decreasing of the estimated divergence $\hat{D}_\alpha(Q||P)(t)$ in (191) implies increasing of the loss $\mathcal{L}_\alpha(\theta_t)$ in (192).

Implementation and Training Details. We used a neural network which has 3 hidden layers of 100 units in each layer. The Adam algorithm in PyTorch was used. For the hyperparameters in the training, the learning rate was 0.001, BatchSize was 2500, and the number of epochs was 500. A NVIDIA Tesla K80 GPU was used. It took approximately four hours to conduct all simulations for each value of α .

Results. Figure 1 and 2 show the results of estimating the α divergence over the number of learning steps during the optimization. Figure 1 is for $\alpha = -3, -2, -1, 2, 3$ and 4 , and Figure 2 is for $\alpha = 0.2, 0.5$, and 0.8 . The y -axis of each graph represents the estimated value of the α divergence, and the x -axis of each graph represents the learning step. The solid blue line shows the median of the estimates of the α divergence. The dark blue area shows the ranges of the estimates between the 45th and 55th percentiles, and the light blue area shows the range of the estimates between the 5th and 95th percentile quartiles.

Discussion. As shown in Figure 1, the estimates of the α divergence diverged. This corresponds to a negative divergence of the loss function $\mathcal{L}_\alpha(\theta_t)$ in (192), and then implies that $E_Q[e^{T_{\theta_t}}] \rightarrow 0$ for $\alpha > 1$, and $E_Q[e^{T_{\theta_t}}] \rightarrow \infty$ for $\alpha < 0$ in (191). The discussion in Section 5 suggests that $E[\nabla_\theta \mathcal{L}_\alpha(\theta)] \rightarrow \vec{0}$. That is, the gradients of the neural networks in this case vanished for $\alpha = -3, -2, -1, 2, 3$ and 4 . However, as shown in Figure 1, the estimates of the α divergence converge stably for $\alpha = 0.2, 0.5$, and 0.8 .

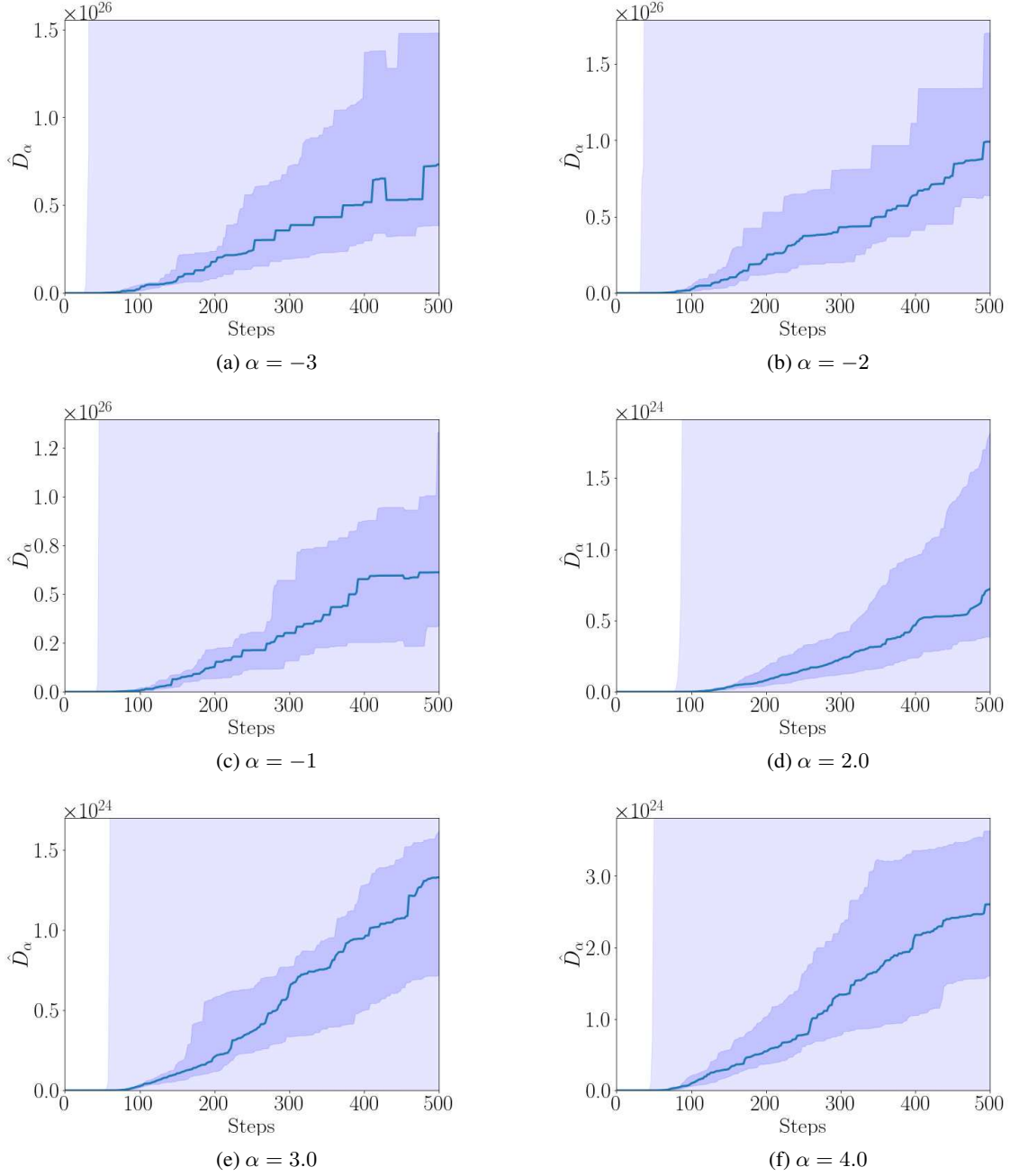


Figure 1: Results of estimating the α divergence for $\alpha = -3, -2, -1, 2, 3$ and 4 , over the number of learning steps during the optimization. The y -axis of each graph represents the estimated value of the α divergence, and the x -axis of each graph represents the learning step. The solid blue line shows the median of the estimates of the α divergence. The dark blue area shows the ranges of the estimates between the 45th and 55th percentiles, and the light blue area shows the range of the estimates between the 5th and 95th percentile quartiles.

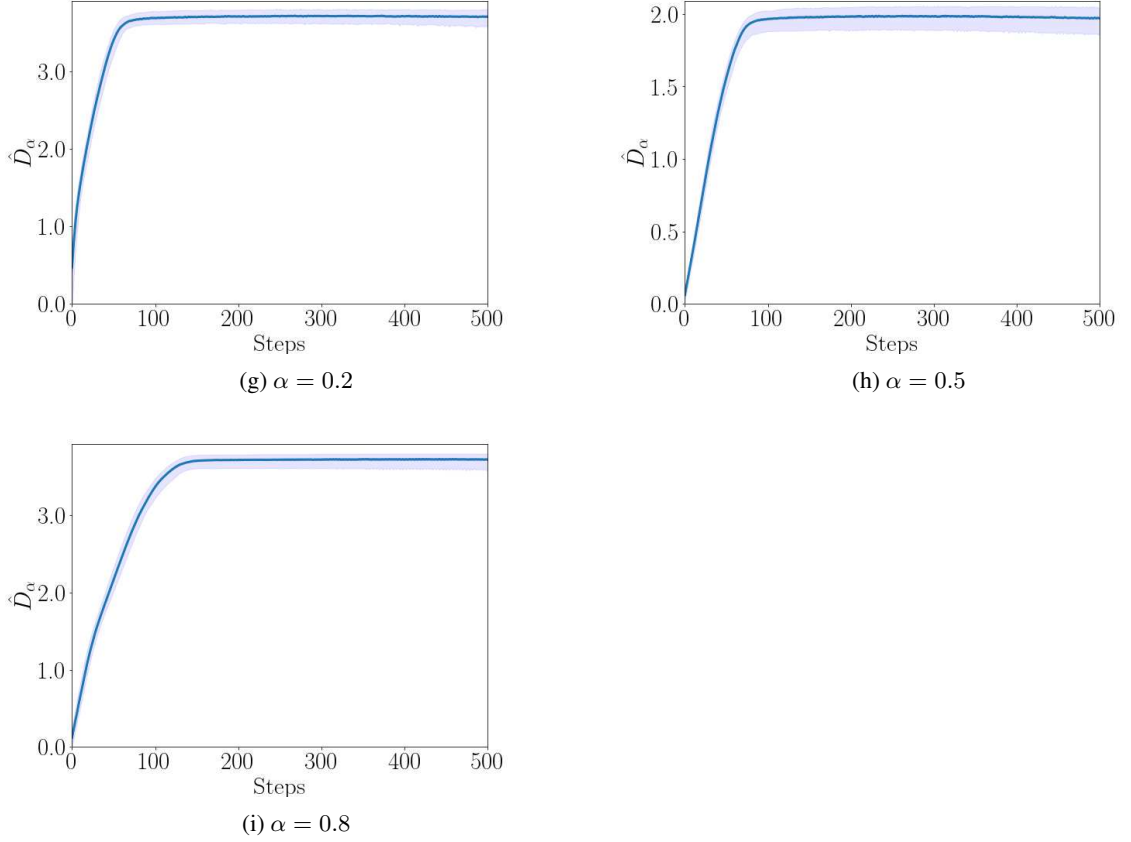


Figure 2: Results of estimating the α divergence for $\alpha = 0.2, 0.5$ and 0.8 , over the number of learning steps during the optimization. The y -axis of each graph represents the estimated value of the α divergence, and the x -axis of each graph represents the learning step. The solid blue line shows the median of the estimates of the α divergence. The dark blue area shows the ranges of the estimates between the 45th and 55th percentiles, and the light blue area shows the range of the estimates between the 5th and 95th percentile quartiles.

D.2 Experiments to confirm the relationship between dimensions of dataset and steps in training

In this section, we report the results of numerical experiments related to the discussion in Section 7: the results of numerical experiments to confirm the relationship between dimensions of dataset and steps in training are presented.

Experimental Setup. We generated training and test datasets of dimensions $d = 2, 3, 4, 5, 6$, and 7 , and then trained an NGB model with the training dataset while estimating the α divergence at each learning step with the test dataset. One hundred numerical simulations were performed for each dimension d . As a result of the experiment, the median of the estimated value and ranges between the 5th and 95th percentile quartiles at each learning step are reported.

Synthetic Data. For each $d = 2, 3, 4, 5, 6$, and 7 , we generated the training and test datasets of size 5000 from d -dimensional normal distribution $\{X_1, X_2, \dots, X_d\}$, such that $E[X_i] = 0$, $\text{Var}[X_i] = 1$ and $E[X_i \cdot X_j] = 0.8$ ($i \neq j$).

Table 2: The early stop step ($N^{2/d}$) and the median of the steps at which the estimated divergence reaches its maximum ($\text{median}(K_{\max})$), for each dimension $d = 2, 3, 4, 5, 6$, and 7.

	$d = 2$	$d = 3$	$d = 4$	$d = 5$	$d = 6$	$d = 7$
$N^{2/d}$	5000	292	71	30	17	11
$\text{median}(K_{\max})$	130	112	130	136	50	50

878 **Estimating the α divergence.** The α divergence was estimated in the following way:

$$\hat{D}_\alpha(Q||P)(t) = \frac{1}{\alpha \cdot (1 - \alpha)} - \frac{1}{\alpha} \hat{E}_Q \left[e^{\alpha \cdot T_{\theta_t}(\mathbf{x}^{te})} \right] - \frac{1}{1 - \alpha} \hat{E}_P \left[e^{(\alpha - 1) \cdot T_{\theta_t}(\mathbf{x}^{te})} \right] \quad (193)$$

$$= \frac{1}{\alpha \cdot (1 - \alpha)} - \mathcal{L}_\alpha(\theta_t) \quad (194)$$

879 where T_{θ_t} is a model at learning step t in Algorithm 1 and \mathbf{x}^{te} denotes the test dataset. Note that,
880 decreasing of the estimated divergence $\hat{D}_\alpha(Q||P)(t)$ in (193) implies increasing of the loss $\mathcal{L}_\alpha(\theta_t)$
881 in (194).

882 **Implementation and Training Details.** We used a neural network which has 3 hidden layers of
883 100 units in each layer. The Adam algorithm in PyTorch was used. For the hyperparameters in the
884 training, the learning rate was 0.001, BathSize was 2500, and the number of epochs was 500. A
885 NVIDIA Tesla K80 GPU was used. It took approximately four hours to conduct all simulations for
886 each d .

887 **Results.** Let K_{\max} denote the step at which the estimated divergence reaches its maximum:

$$K_{\max} = \underset{t}{\operatorname{argmax}} \hat{D}_\alpha(Q||P)(t). \quad (195)$$

888 Table 3 lists $N^{2/d}$, the early stop step obtained from (23), and the median of K_{\max} , for each dimen-
889 sion $d = 2, 3, 4, 5, 6$, and 7. In Figure 3, we show the results of estimating the α divergence over
890 the number of learning steps during the optimization. Since the value of the α divergence changes
891 as the dimension of the dataset changes, we divided by the the estimated value of the divergence
892 by the true value of the divergence to normalize the results of each dimension. The y -axis of each
893 graph represents the estimated value of the α divergence divided by the true value of the divergence,
894 and the x -axis of each graph represents the learning step. The solid blue line shows the median of
895 the estimates of the α divergence. The light blue area shows the range of the estimates between
896 the 5th and 95th percentile quartiles. The dashed red line indicates $Y=1$, which corresponds to the
897 theoretical value of the estimate for each d .

898 divided

899 **Discussion.** As shown in Table 2, the steps from the early stop method and those at which the
900 estimates decreased were approximately consistent, except in the case of $d = 2$. However, the
901 estimates of the data of the low dimensions, particularly $d = 2$, decreased earlier than the early
902 stop method suggests. This may be because C in (23) for the data of low dimensions can be small
903 because the neural network learns quickly when the dimensions of the data are low. However, Figure
904 3 shows that the estimates of the divergence decreased slowly when the dimensions of the data are
905 low, and they decreased more quickly when the dimensions of the data were higher. These results
906 suggest that the curse of dimensionality of balancing is easier to observe when dimensions of data
907 are higher.

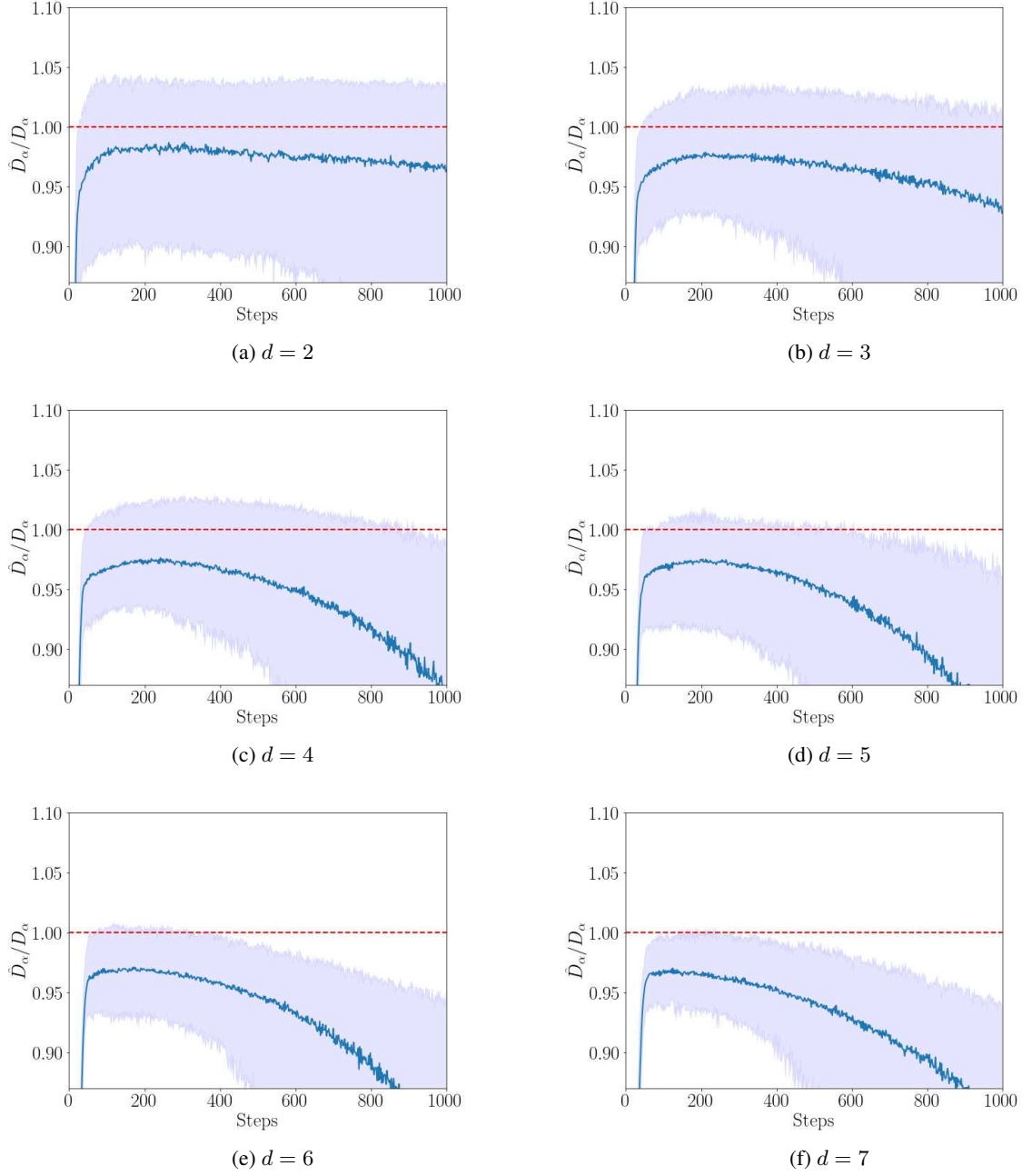


Figure 3: Results of estimating the α divergence for $\alpha = -3, -2, -1, 2, 3$ and 4 , over the number of learning steps during the optimization. The y -axis of each graph represents the estimated value of the α divergence divided by the true value of the divergence, and the x -axis of each graph represents the learning step. The dashed red line indicates $Y=1$, which corresponds to the theoretical value of the estimate for each d . The solid blue line shows the median of the estimates of the α divergence. The dark blue area shows the ranges of the estimates between the 45th and 55th percentiles, and the light blue area shows the range of the estimates between the 5th and 95th percentile quartiles.

D.3 Experiments for estimating causal effects of joint and multidimensional interventions with different sample sizes

In this section, we report the results of numerical experiments related to the discussion in Section 8: the results of numerical experiments for estimating causal effects of joint and multidimensional interventions with different sample sizes are presented.

Experimental Setup. The following two experiments were conducted, in which synthetic data of size $N = 1000, 10000$, and 100000 were generated using the method developed by Vegetabile et al.(2021).

- Experiment 1. An experiment on estimating the causal effect of a single intervention, especially for continuous intervention, $E[Y|do(A), \mathbf{X}]$.
- Experiment 2. An experiment to estimate the causal effect of a mixture of both arbitrary discrete and continuous interventions, $E[Y|do(A), \overline{do}(\mathbf{X}_1), do(X_2), \overline{do}(\mathbf{X}_3)]$.

Experimental Details. Experiments 1 and 2 were conducted using the following steps.

Step 1: We created training dataset of size $N = 1000, 10000$, and 100000 , and test dataset with size $N = 1000$. The training dataset were generated using the method developed by Vegetabile et al.(2021). The test dataset were generated from the following distribution:

- Experiment 1. $P(Y|do(A), \mathbf{X}) \times P(A) \times P(\mathbf{X})$,
- Experiment 2. $P(Y|do(A), do(\mathbf{X}_1), do(X_2), do(\mathbf{X}_3)) \times P(A) \times P(\mathbf{X}_1) \times P(X_2) \times P(\mathbf{X}_3)$,

where P denotes the distribution of the training dataset. To create the test dataset, we shuffled the dataset generated from the same distribution as the training dataset.

Step 2: The balancing weights were estimated for each experiment. We estimated $BW(A, \mathbf{X} : T_{\theta_*})$ for Experiment 1, and $BW(A, \mathbf{X}_1, X_2, \mathbf{X}_3 : T_{\theta_*})$ for Experiment 2.

Step 3: We created models for each experiment using the linear regression (LR) or the gradient boosting tree (GBT) algorithm with our weights from the previous step. The hyperparameters were tuned to create models of GBT.

Step 4: We estimate the average causal effects $E[Y|do(A), \mathbf{X}]$ and $E[Y|do(A), \overline{do}(\mathbf{X}_1), do(X_2), \overline{do}(\mathbf{X}_3)]$ using the predictions of the models from Step 2 with the test dataset. Finally, we report the mean squared error (RMSE) between the true and estimated values.

Baseline Method. The main baseline method used in our experiments is entropy balancing [30]. We compared our method with the method for balancing \mathbf{X} with A for each of the moments from 1 to 4. For Experiment 1, both our method and the baseline method estimated the same target: $E[Y|do(A), \mathbf{X}]$. However, no existing method can fully deal with the target of Experiment 2: $E[Y|do(A), \overline{do}(\mathbf{X}_1), do(X_2), \overline{do}(\mathbf{X}_3)]$. Therefore, the same entropy balancing as in Experiment 1 was used in Experiment 2. This may be an unfair comparison to the baseline method. In addition, we included a “naïve” estimation, using algorithms with no sample weights, as a baseline. For the calculation of entropy balancing weights, WeightIt library in R was used.⁷

Training Data Set. Specifically, we used the following steps to generate the dataset. First, $\mathbf{W} = (W_1, W_2, W_3, W_4, W_5)$ were generated independently, such that $W_1 \sim \mathcal{N}(-0.5, 1)$, $W_2 \sim \mathcal{N}(1, 1)$, $W_3 \sim \mathcal{N}(0, 1)$, $W_4 \sim \mathcal{N}(1, 1)$, and $X_{W_5} = \{0, 1, 2\}$ with $P(W_5 = 0) = 0.70$ and $P(W_5 = 1) = P(W_5 = 2) = 0.15$. Second, A and Y were generated as follows:

$$\begin{aligned} A &\sim \mathcal{X}^2(df = 3, \mu_A(W_1, W_2, W_4, W_5)), \\ Y &= \frac{1}{50} [(-0.15A^2 + A(W_1^2 + W_2^2) - 15) \\ &\quad + ((W_1 + 3)^2 + 2(W_2 - 25)^2 + W_3) \\ &\quad - C + \varepsilon], \end{aligned} \tag{196}$$

⁷<https://cran.r-project.org/web/packages/WeightIt/index.html>

948 where $\mu_A(W_1, W_2, W_4, W_5) = 5|W_1| + 6|W_2| + |W_4| + a$, and $a = 0$ if $W_5 = 0$, and $a = 1$
 949 if $W_5 = 1$, and $a = 5$ if $W_5 = 2$, and $C = E[(W_1 + 3)^2] + 2E[(W_2 - 25)^2] + E[W_3]$, and
 950 $\varepsilon \sim \mathcal{N}(0, 1)$. Here, $\mathcal{X}^2(df = n, \mu)$ is the noncentral χ^2 distribution with n degrees of freedom and
 951 a noncentral parameter μ . Finally, we create new variables $\mathbf{X} = (\mathbf{X}_1, X_2, \mathbf{X}_3)$, as observed values
 952 of \mathbf{W} using the following transformation:

$$\mathbf{X}_1 = (X_{(1,1)}, X_{(1,2)}, X_{(1,3)}), \quad (197)$$

$$\begin{aligned} \text{where } X_{(1,1)} &= \exp(W_1/2), \\ X_{(1,2)} &= W_2/(1 + \exp(W_1)) + 10, \\ X_{(1,3)} &= W_1 W_3/25 + 0.6, \end{aligned}$$

$$X_2 = (W_4 - 1)^2, \quad (198)$$

$$\mathbf{X}_3 = \begin{cases} (1, 0) & \text{if } W_5 = 0, \\ (0, 1) & \text{if } W_5 = 1, \\ (0, 0) & \text{if } W_5 = 2. \end{cases} \quad (199)$$

953 **Test Data Set.** We first generated dataset from the same distribution as the training dataset.
 954 Second, the dataset were shuffled by the index, with the following divided parts treated as a single
 955 piece of data: for Experiment 1, A and \mathbf{X} were shuffled by the index, and for Experiment 2, each
 956 of A , \mathbf{X}_1 , X_2 and \mathbf{X}_3 were shuffled by the index. Third, using the inverse transformation of Eq.
 957 (197)-(199), we calculated $(W_1, W_2, W_3, W_4, W_5)$ from $\mathbf{X}_1 = (X_{(1,1)}, X_{(1,2)}, X_{(1,3)})$, X_2 , and
 958 \mathbf{X}_3 of the shuffled dataset:

$$\begin{aligned} W_1 &= 2 \log X_{(1,1)}, & W_2 &= X_{(1,2)} \cdot (1 + X_{(1,1)}^2), \\ W_3 &= \frac{25(X_{(1,3)} - 0.6)}{2 \log X_{(1,1)}}, & W_4 &= \sqrt{X_2} + 1, \\ W_5 &= \begin{cases} 0 & \text{if } \mathbf{X}_3 = (1, 0), \\ 1 & \text{if } \mathbf{X}_3 = (0, 1), \\ 2 & \text{if } \mathbf{X}_3 = (0, 0). \end{cases} \end{aligned}$$

959 Finally, the true values of Y for causal effects were calculated using the terms in Eq. (196) without
 960 the term ε .

961 **Implementation and Training Details.** $N = 1000$: For experiments with the dataset of size $N =$
 962 1000, we used a neural network which has 10 hidden layers of 100 units in each layer. $\alpha = 0.5$ was
 963 used to estimate the divergence. The Adam algorithm in PyTorch was used. For the hyperparameters
 964 in the training, the learning rate was 0.0001, BathSize was 1000, and the number of epochs was 70.
 965 A NVIDIA Tesla K80 GPU was used. It took approximately 40 min to conduct all the simulations
 966 for each experiment.

967 $N = 10000$: For experiments with the dataset of size $N = 10000$, We used a neural network which
 968 has 10 hidden layers of 100 units in each layer. $\alpha = 0.5$ was used to estimate the divergence. The
 969 Adam algorithm in PyTorch was used. For the hyperparameters in the training, the learning rate was
 970 0.0001, BathSize was 2500, and the number of epochs was 200. A NVIDIA Tesla K80 GPU was
 971 used. It took approximately 7 h to conduct all the simulations for each experiment.

972 $N = 100000$: For experiments with the dataset of size $N = 100000$, We used a neural network
 973 which has 10 hidden layers of 100 units in each layer. $\alpha = 0.5$ was used to estimate the divergence.
 974 The Adam algorithm in PyTorch was used. For the hyperparameters in the training, the learning rate
 975 was 0.0001, BathSize was 2500, and the number of epochs was 200. A NVIDIA Tesla K80 GPU was
 976 used. It took approximately 78 h to conduct all the simulations for each experiment.

Table 3: Average RMSE for estimation in Experiments 1 and 2 for dataset of size $N = 1000, 10000$, and 100000 . For entropy balancing, the number to the right side of the method name, “ (m) ,” denotes the number of moments that are balanced. The results from 100 simulations are in the form of “mean (std. err.)”.

(a) $N = 1000$				
Method	Experiment 1		Experiment 2	
	LR	GBT	LR	GBT
Unweighted	1.347(0.039)	0.739(0.066)	1.347(0.033)	0.741(0.068)
Entropy Balancing(1)	1.303(0.056)	0.724(0.058)	1.303(0.052)	0.726(0.060)
Entropy Balancing(2)	1.206(0.029)	0.693(0.056)	1.206(0.026)	0.698(0.055)
Entropy Balancing(3)	1.201(0.026)	0.690(0.054)	1.201(0.024)	0.698(0.061)
Entropy Balancing(4)	1.203(0.027)	0.699(0.057)	1.203(0.025)	0.699(0.061)
NBW	1.347(0.039)	0.745(0.065)	1.347(0.034)	0.738(0.063)
(b) $N = 10000$				
Method	Experiment 1		Experiment 2	
	LR	GBT	LR	GBT
Unweighted	1.342(0.030)	0.489(0.035)	1.342(0.026)	0.489(0.039)
Entropy Balancing(1)	1.295(0.033)	0.486(0.026)	1.295(0.030)	0.487(0.035)
Entropy Balancing(2)	1.194(0.025)	0.466(0.036)	1.194(0.025)	0.468(0.041)
Entropy Balancing(3)	1.187(0.025)	0.459(0.032)	1.187(0.024)	0.457(0.036)
Entropy Balancing(4)	1.189(0.024)	0.457(0.035)	1.189(0.023)	0.452(0.034)
NBW	1.274(0.038)	0.488(0.035)	1.273(0.031)	0.485(0.032)
(c) $N = 100000$				
Method	Experiment 1		Experiment 2	
	LR	GBT	LR	GBT
Unweighted	1.342(0.027)	0.457(0.048)	1.342(0.023)	0.459(0.044)
Entropy Balancing(1)	1.299(0.029)	0.453(0.037)	1.298(0.027)	0.455(0.036)
Entropy Balancing(2)	1.195(0.025)	0.391(0.034)	1.194(0.023)	0.386(0.039)
Entropy Balancing(3)	1.186(0.024)	0.361(0.025)	1.186(0.023)	0.360(0.023)
Entropy Balancing(4)	1.188(0.024)	0.353(0.022)	1.187(0.023)	0.356(0.020)
NBW	1.239(0.095)	0.376(0.033)	1.252(0.080)	0.388(0.030)

Results. We report the average and standard errors of the root mean squared error (RMSE) between the estimated and true values of the average causal effects for synthetic data of size $N = 1000, 10000$, and 100000 . Table 3 lists the results of Experiments 1 and 2 for each N . Each result is in the form of “mean (std. err.)” from 100 simulations.

Discussion. As shown in all the results, the results of NBW were less accurate than those of the entropy-balancing method. Moreover, the results for $N = 1000$ shows that NBW were less accurate than the unweighted estimation. However, as seen in all results for $N = 100000$, the accuracy of NBW was superior to that of the unweighted estimation, which was close to the accuracy of the entropy-balancing method. These results imply that the sample size requirements of the proposed method are larger than those of the entropy balancing method.

Algorithm 3 Back-Propagation Algorithm using Neural Balancing Weights

Input: Data $(y, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \mathbf{z}) = \{(y^i, \mathbf{x}_1^i, \mathbf{x}_2^i, \dots, \mathbf{x}_n^i, \mathbf{z}^i) | i = 1, 2, \dots, N\}$
1: A Neural Balancing Weight Model T
Output: A Neural Network Model f_ϕ for Estimating $E_{\bar{P}}[Y|\mathbf{X}, \mathbf{Z}]$
2: **repeat**
3: $\hat{y} \leftarrow f_\phi(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \mathbf{z})$ // Forward Propagation
4: $BW(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \mathbf{z}) \leftarrow \frac{e^{-T(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \mathbf{z})}}{MEAN(e^{-T(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \mathbf{z})})}$
5: $Err_\phi \leftarrow y - \hat{y}$ // Obtaining Errors for f_ϕ
6: $\mathcal{L}(\phi) \leftarrow MEAN(Err_\phi \otimes Err_\phi \otimes BW(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \mathbf{z}))$ // Calculating Loss $\mathcal{L}(\phi)$
7: $\phi \leftarrow \phi - \nabla \mathcal{L}(\phi)$
8: **until** convergence

E Back-Propagation Algorithm using Neural Balancing Weights

We show a back-propagation algorithm using NBW for MSE loss in Algorithm 3. The MSE loss here is calculated by the mean of the element wise product of both the original squared errors and the balancing weights.