APPENDIX A
PROOFS

We first introduce a technical result from approximation theory that will be used in the subsequent derivations. The following result, which is a restatement of Proposition 10 from [17], states that a sufficiently smooth function over a compact domain can be approximated to within $O(k^{-1/2})$ error by a shallow NN.

Proposition 1 (Approximation of smooth functions; Proposition 10 from [17])**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be compact and $g : \mathcal{X} \to \mathbb{R}$. Suppose that there exists an open set $\mathcal{U} \supset \mathcal{X}$, $b \geq 0$, and $\tilde{g} \in \mathcal{C}_b^{s_{\mathsf{KB}}}(\mathcal{U})$, $s_{\mathsf{KB}} \coloneqq \lfloor d/2 \rfloor + 3$, such that $g = \tilde{g}\big|_{\mathcal{X}}$. Then, there exists $f \in \mathcal{F}_{k,d}\left(\bar{c}_{b,d,\|\mathcal{X}\|}\right)$, where $\bar{c}_{b,d,\|\mathcal{X}\|}$ is given in Equation (A.15) of [17], such that*

$$\|f - g\|_\infty \lesssim \bar{c}_{b,d,\|\mathcal{X}\|} d^{\frac{1}{2}} k^{-\frac{1}{2}}.$$

This proposition will allow us to control the approximation error of the EOT NE. To invoke it, we will establish smoothness of the semi-dual EOT potentials (see Lemma 1 ahead). The smoothness of potentials stems from the presence of the entropic penalty and the smoothness of the quadratic cost function.

A. Proof of Theorem 1

For $(\mu, \nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$, define the population-level neural EOT cost as

$$\mathsf{OT}^{\varepsilon}_{k,a}(\mu, \nu) \coloneqq \sup_{f \in \mathcal{F}_{k,d}(a)} \int f d\mu + \int f^{c,\varepsilon} d\nu. \tag{15}$$

We decompose the neural estimation error into the approximation and empirical estimation errors:

$$\begin{aligned}\mathbb{E}\left[\left|\widehat{\mathsf{OT}}^{\varepsilon}_{k,a}(X^n, Y^n) - \mathsf{OT}^{\varepsilon}(\mu, \nu)\right|\right] &\leq \underbrace{\left|\mathsf{OT}^{\varepsilon}_{k,a}(\mu, \nu) - \mathsf{OT}^{\varepsilon}(\mu, \nu)\right|}_{\text{Approximation error}} \\ &\quad + \underbrace{\mathbb{E}\left[\left|\mathsf{OT}^{\varepsilon}_{k,a}(\mu, \nu) - \widehat{\mathsf{OT}}^{\varepsilon}_{k,a}(X^n, Y^n)\right|\right]}_{\text{Estimation error}}, \end{aligned} \tag{16}$$

and analyze each term separately.

Approximation error. Proposition 1 provides a sup-norm approximation error bound of a smooth function by a NN. To invoke it, we first study the semi-dual EOT potentials and show that they are indeed smooth functions, i.e., admit an extension to an open set with sufficiently many bounded derivatives. The following lemma establishes regularity of semi-dual potentials for $\mathsf{OT}^{\varepsilon}(\mu, \nu)$; after stating it we shall account for the extension.

Lemma 1 (Uniform regularity of EOT potentials)**.** *There exist semi-dual EOT potentials $(\varphi, \varphi^{c,\varepsilon})$ for $\mathsf{OT}^{\varepsilon}(\mu, \nu)$, such that*

$$\begin{aligned}&\|\varphi\|_{\infty,\mathcal{X}} \leq 2d \\ &\|D^\alpha \varphi\|_{\infty,\mathcal{X}} \leq C_s\left(1 + \varepsilon^{1-s}\right)\left(1 + 2\sqrt{d}\right)^s \text{ with } 1 \leq |\alpha| \leq s,\end{aligned} \tag{17}$$

for any $s \geq 2$ and some constant C_s that depends only on s. Analogous bounds hold for $\varphi^{c,\varepsilon}$.

The lemma is proven in Appendix B-A. The derivation is similar to that of Lemma 4 in [49], but the bounds are adapted to the compactly supported case and present an explicit dependence on ε (as opposed to the $\varepsilon = 1$ assumption that was imposed in that work).

Let $(\varphi, \varphi^{c,\varepsilon})$ be semi-dual potentials as in Lemma 1 (i.e., satisfying (17)) with the normalization $\int \varphi d\mu = \int \varphi^{c,\varepsilon} d\nu = \frac{1}{2}\mathrm{OT}^{\varepsilon}(\mu, \nu)$. Define the natural extension of φ to the open ball of radius $\sqrt{2d}$:

$$\tilde{\varphi}(x) \coloneqq -\varepsilon \log \int \exp\left(\frac{\psi(y) - c(x,y)}{\varepsilon}\right) d\nu(y), x \in B_d(\sqrt{2d}),$$

and notice that $\tilde{\varphi}\big|_{\mathcal{X}} = \varphi$, pointwise on $\mathcal{X}$. Similarly, consider its (c,ε)-transform $\tilde{\varphi}^{c,\varepsilon}$ extended to $B_d(\sqrt{2d})$, and again observe that $\tilde{\varphi}^{c,\varepsilon}\big|_{\mathcal{Y}} = \varphi^{c,\varepsilon}$. Following the proof of Lemma 1, one readily verifies that for any $s \geq 2$, we have

$$\begin{aligned}&\|\tilde{\varphi}\|_{\infty, B_d(\sqrt{2d})} \leq 4d \\ &\|D^\alpha \tilde{\varphi}\|_{\infty, B_d(\sqrt{2d})} \leq C_s\left(1 + \varepsilon^{1-s}\right)(1 + 2\sqrt{2d})^s, 1 \leq |\alpha| \leq s.\end{aligned} \tag{18}$$

Recall that $s_{\mathsf{KB}} = \lfloor d/2 \rfloor + 3$ and set

$$C_d \coloneqq (1 + 2\sqrt{2d})^{s_{\mathsf{KB}}}. \tag{19}$$

By (18) and (19), we now have

$$\max_{\alpha:\, |\alpha| \leq s_{\mathsf{KB}}} \|D^\alpha \tilde{\varphi}\|_{\infty, B_d(\sqrt{2d})} \leq C_{s_{\mathsf{KB}}} C_d (1 + \varepsilon^{1 - s_{\mathsf{KB}}}) \coloneqq b, \tag{20}$$

and so $\tilde{\varphi} \in \mathcal{C}_b^{s_{\mathsf{KB}}}\left(B_d(\sqrt{2d})\right)$.

Noting that $\mathcal{X} \subset B_d(\sqrt{2d})$, by Proposition 1, there exists $f \in \mathcal{F}_{k,d}\left(\bar{c}_{b,d}\right)$ such that

$$\|\varphi - f\|_{\infty,\mathcal{X}} \lesssim \bar{c}_{b,d} d^{\frac{1}{2}} k^{-\frac{1}{2}}, \tag{21}$$

where $\bar{c}_{b,d} = b\,\bar{c}_d$ and $\bar{c}_d$ is defined as (see [17, Equation (A.15)])

$$\begin{aligned}\bar{c}_d \coloneqq &\left(\kappa_d d^{\frac{3}{2}} \vee 1\right) \pi^{\frac{d}{2}} \Gamma\left(\frac{d}{2} + 1\right)^{-1} (\mathrm{rad}(\mathcal{X}) + 1)^d \\ &\times 2^{s_{\mathsf{KB}}} d \left(\frac{1 - d^{\frac{s_{\mathsf{KB}}}{2}}}{1 - \sqrt{d}} + d^{\frac{s_{\mathsf{KB}}}{2}}\right) \max_{\|\alpha\|_1 \leq s_{\mathsf{KB}}} \|D^\alpha \Psi\|_{\infty, B_d(0.5)},\end{aligned} \tag{22}$$

with $\kappa_d^2 \coloneqq \left(d + d^{(s_{\mathsf{KB}}-1)}\right) \int_{\mathbb{R}^d} \left(1 + \|\omega\|^{2(s_{\mathsf{KB}}-2)}\right)^{-1} d\omega$, $\mathrm{rad}(\mathcal{X}) = 0.5 \sup_{x,x' \in \mathcal{X}} \|x - x'\|$ and $\Psi(x) \propto \exp\left(-\frac{1}{0.5 - \|x\|^2}\right) \mathbb{1}_{\{\|x\| < 0.5\}}$ as the canonical mollifier normalized to have unit mass.

Our last step is to lift the sup-norm neural approximation bound on the semi-dual potential from (21) to a bound on the approximation error of the corresponding EOT cost. The following lemma is proven in Appendix B-B.

Lemma 2 (Neural approximation error reduction)**.** *Fix $(\mu, \nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ and let φ be the semi-dual EOT*

potential for $\mathsf{OT}^{\varepsilon}(\mu,\nu)$ *from Lemma 1. For any* $f \in \mathcal{F}_{k,d}(a)$*, we have*

$$\left|\mathsf{OT}^{\varepsilon}(\mu,\nu) - \mathsf{OT}^{\varepsilon}_{k,a}(\mu,\nu)\right| \leq 2\left\|\varphi - f\right\|_{\infty,\mathcal{X}}.$$

Setting $a = \bar{c}_{b,d}$ and combining Lemma 2 with (21), we obtain

$$\left|\mathsf{OT}^{\varepsilon}_{k,a}(\mu,\nu) - \mathsf{OT}^{\varepsilon}(\mu,\nu)\right| \lesssim_d \left(1 + \frac{1}{\varepsilon^{\left\lfloor \frac{d}{2} \right\rfloor + 2}}\right) k^{-\frac{1}{2}}. \quad (23)$$

Estimation error. Set $\mathcal{F}^{c,\varepsilon}(a) := \{f^{c,\varepsilon} : f \in \mathcal{F}_{k,d}(a)\}$, and first bound

$$\begin{aligned}
&\mathbb{E}\left[\left|\mathsf{OT}^{\varepsilon}_{k,a}(\mu,\nu) - \widehat{\mathsf{OT}}^{\varepsilon}_{k,a}(X^n,Y^n)\right|\right] \\
&\leq \underbrace{n^{-\frac{1}{2}}\mathbb{E}\left[\sup_{f\in\mathcal{F}_{k,d}(a)} n^{-\frac{1}{2}}\left|\sum_{i=1}^{n}\left(f(X_i) - \mathbb{E}_{\mu}[f]\right)\right|\right]}_{\text{(I)}} \quad (24) \\
&\quad + \underbrace{n^{-\frac{1}{2}}\mathbb{E}\left[\sup_{f\in\mathcal{F}^{c,\varepsilon}(a)} n^{-\frac{1}{2}}\left|\sum_{j=1}^{n}\left(f(Y_j) - \mathbb{E}_{\nu}[f]\right)\right|\right]}_{\text{(II)}}. \quad (25)
\end{aligned}$$

To control these expected suprema, we again require regularity of the involved function, as stated in the next lemma.

Lemma 3. *Fix* $c \in \mathcal{C}^{\infty}$*, the* (c,ε)*-transform of NNs class* $\mathcal{F}_{k,d}(a)$ *satisfies the following uniform smoothness properties:*

$$\max_{\alpha:|\alpha|_1\leq s} \left\|D^{\alpha} f^{c,\varepsilon}\right\|_{\infty,\mathcal{Y}} \leq R_s(1+a)(1+\varepsilon^{1-s})$$

for any $s \geq 2$*,* $f \in \mathcal{F}_{k,d}(a)$ *and some constant* R_s *that depends only on* s, d*.*

The only difference between Lemmas 1 and 3 is that here we consider the (c,ε)-transform of NNs, rather than of dual EOT potentials. As our NNs are also compactly supported and bounded, the derivation of this result is all but identical to the proof of Lemma 1, and is therefore omitted to avoid repetition.

We proceed to bound Terms (I) and (II) from (24). For the first, consider

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}_{k,d}(a)} n^{-\frac{1}{2}}\left|\sum_{i=1}^{n}\left(f(X_i) - \mathbb{E}_{\mu}[f]\right)\right|\right] \quad (26) \\
&\overset{(a)}{\lesssim} \mathbb{E}\left[\int_0^{\infty}\sqrt{\log N\left(\delta, \mathcal{F}_{k,d}(a), \|\cdot\|_{2,\mu_n}\right)}d\delta\right] \\
&\leq \int_0^{\infty}\sqrt{\sup_{\gamma\in\mathcal{P}(\mathcal{X})}\log N\left(\delta, \mathcal{F}_{k,d}(a), \|\cdot\|_{2,\gamma}\right)}d\delta \\
&\overset{(b)}{=} \int_0^{12a}\sqrt{\sup_{\gamma\in\mathcal{P}(\mathcal{X})}\log N\left(\delta, \mathcal{F}_{k,d}(a), \|\cdot\|_{2,\gamma}\right)}d\delta \\
&\lesssim a\int_0^{1}\sqrt{\sup_{\gamma\in\mathcal{P}(\mathcal{X})}\log N\left(6a\delta, \mathcal{F}_{k,d}(a), \|\cdot\|_{2,\gamma}\right)}d\delta \\
&\overset{(c)}{\lesssim} ad^{\frac{3}{2}}, \quad (27)
\end{aligned}$$

where:

(a) follows by [50, Corollary 2.2.8] since $n^{-\frac{1}{2}}\sum_{i=1}^{n}\sigma_i f(X_i)$, where $\{\sigma_i\}_{i=1}^{n}$ are i.i.d Rademacher random variables, is sub-Gaussian w.r.t. pseudo-metric $\|\cdot\|_{2,\mu_n}$ (by Hoeffding's inequality) ;

(b) is since $\bar{C}\left(\left|\mathcal{F}_{k,d}(a)\right|, \mathcal{X}\right) \leq 3a(\|\mathcal{X}\|+1) = 6a$ and $N\left(\delta, \mathcal{F}_{k,d}(a), \|\cdot\|_{2,\gamma}\right) = 1$, whenever $\delta > 12a$;

(c) uses the bound

$$\int_0^{1}\sqrt{\sup_{\gamma\in\mathcal{P}(\mathcal{X})}\log N\left(6a\delta, \mathcal{F}_{k,d}(a), \|\cdot\|_{2,\gamma}\right)}d\delta \lesssim d^{\frac{3}{2}},$$

which follows from step (A.33) in [17].

For Term (II), let $s = \lceil d/2 \rceil + 1$, and consider

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}^{c,\varepsilon}(a)} n^{-\frac{1}{2}}\left|\sum_{i=1}^{n}\left(f(Y_i) - \mathbb{E}_{\mu}[f]\right)\right|\right] \quad (28) \\
&\overset{(a)}{\lesssim} \int_0^{12a}\sqrt{\sup_{\gamma\in\mathcal{P}(\mathcal{Y})}\log N\left(\delta, \mathcal{F}^{c,\varepsilon}(a), \|\cdot\|_{2,\gamma}\right)}d\delta \\
&\lesssim \int_0^{12a}\sqrt{\sup_{\gamma\in\mathcal{P}(\mathcal{Y})}\log N_{[]}\left(2\delta, \mathcal{F}^{c,\varepsilon}(a), \|\cdot\|_{2,\gamma}\right)}d\delta \\
&\overset{(b)}{\lesssim} K_s\int_0^{12a}\left(\frac{R_s(1+a)(1+\varepsilon^{1-s})}{2\delta}\right)^{\frac{d}{2s}}d\delta \\
&\lesssim_d a(1+a)(1+\varepsilon^{1-s}), \quad (29)
\end{aligned}$$

where R_s is the constant from Lemma 3 (which depends only s, d), (a) follows by a similar argument to that from the bound on Term (I), along with equation (38), which specifies the upper limit for entropy integral, while (b) follows by Lemma 3 and [50, Corollary 2.7.2], which upper bounds the bracketing entropy number of smooth functions on a bounded convex support.

To arrive at the effective error bound from Theorem 1, we provide a second bound on Term (II). This second bound yields a better dependence on dimension (namely, only the smaller dimension d appears in the exponent) at the price of another $\sqrt{k}$ factor. Neither bound is uniformly superior over the other, and hence our final result will simply take the minimum of the two. By (38) from the proof of Lemma 2, we have

$$N\left(\delta, \mathcal{F}^{c,\varepsilon}(a), \|\cdot\|_{2,\gamma}\right) \leq N\left(\delta, \mathcal{F}_{k,d}(a), \|\cdot\|_{\infty,\mathcal{X}}\right).$$

Invoking Lemma 2 from [17], which upper bounds the metric entropy of ReLU NNs class on the RHS above, we further obtain

$$\begin{aligned}
&\log N\left(\delta, \mathcal{F}_{k,d}(a), \|\cdot\|_{\infty,\mathcal{X}}\right) \\
&\leq \left((d+2)k + d + 1\right)\log\left(1 + 20a(\|\mathcal{X}\|+1)\delta^{-1}\right), \quad (30)
\end{aligned}$$

and proceed to bound Term (II) as follows:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}^{c,\varepsilon}} n^{-\frac{1}{2}}\left|\sum_{j=1}^{n}\big(f(Y_j)-\mathbb{E}_\nu[f]\big)\right|\right] \tag{31}$$
$$\lesssim a\int_0^1\sqrt{\sup_{\gamma\in\mathcal{P}(\mathcal{X})}\log N\left(6a\delta,\mathcal{F}^{c,\varepsilon}(a),\|\cdot\|_{2,\gamma}\right)}d\delta$$
$$\lesssim ad^{\frac{1}{2}}\sqrt{k}\int_0^1\sqrt{\log(1+7\delta^{-1})}d\delta$$
$$\lesssim ad^{\frac{1}{2}}\sqrt{k}. \tag{32}$$

Inserting (27), (29), and (32) back into (24), we obtain the desired bound on the empirical estimation error by setting $a=\bar{c}_{b,d}$, as was defined in approximation error analysis:

$$\mathbb{E}\left[\left|\mathsf{OT}^{\varepsilon}_{k,a}(\mu,\nu)-\widehat{\mathsf{OT}}^{\varepsilon}_{k,a}\left(X^n,Y^n\right)\right|\right] \lesssim_d \min\left\{1+\frac{1}{\varepsilon^{\left\lceil\frac{3d}{2}\right\rceil+4}},\left(1+\frac{1}{\varepsilon^{\left\lfloor\frac{d}{2}\right\rfloor+2}}\right)\sqrt{k}\right\}n^{-\frac{1}{2}}. \tag{33}$$

The proof is concluded by plugging the approximation error bound from (23) and the estimation error bounds from (33) into (16), and supremizing over $(\mu,\nu)\in\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y})$, while noting that all the above bounds holds uniformly in the two distributions. □

B. Proof of Theorem 2

This proof is similar to that of Theorem 1, up to minor modifications. For brevity, we only highlight the required changes. Note that for k with $m_k\geq\bar{c}_{b,d}$, where the latter is defined in proof of Theorem 1 (see (20) and (22)), we have $\mathcal{F}_{k,d}(\bar{c}_{b,d})\subset\mathcal{F}_{k,d}(m_k)$. Hence, by Lemma 2 and (21), there exists a NN $f\in\mathcal{F}_{k,d}(\bar{c}_{b,d})$, such that

$$\left|\mathsf{OT}^{\varepsilon}(\mu,\nu)-\mathsf{OT}^{\varepsilon}_{k,m_k}(\mu,\nu)\right|\leq 2\left\|\varphi-f\right\|_{\infty,\mathcal{X}} \lesssim_d\left(1+\frac{1}{\varepsilon^{\left\lfloor\frac{d}{2}\right\rfloor+2}}\right)k^{-\frac{1}{2}}.$$

Next, for estimation error, by setting $a=m_k:=\log k\vee 1$ in (27), (29), and (32) (instead of $a=\bar{c}_{b,d}$ as in the proof of Theorem 1), we arrive at

$$\mathbb{E}\left[\left|\mathsf{OT}^{\varepsilon}_{k,m_k}(\mu,\nu)-\widehat{\mathsf{OT}}^{\varepsilon}_{k,m_k}\left(X^n,Y^n\right)\right|\right] \lesssim_d\min\left\{\left(1+\frac{1}{\varepsilon^{\left\lfloor\frac{d}{2}\right\rfloor}}\right)(\log k)^2,\sqrt{k}\log k\right\}n^{-\frac{1}{2}}.$$

Combining both bounds completes the proof. □

C. Proof of Theorem 3

Define $\Gamma(f):=\int_{\mathcal{X}}fd\mu+\int_{\mathcal{Y}}f^{c,\varepsilon}d\nu$, and let $\varphi_\star$ be optimal potential of $\mathsf{OT}^{\varepsilon}(\mu,\nu)$, solving semi-dual formulation. Denote the corresponding optimal coupling by $\pi^{\varepsilon}_\star$. We first show that for any continuous $f:\mathcal{X}\to\mathbb{R}$, the following holds:

$$\Gamma(\varphi_\star)-\Gamma(f)=\varepsilon\mathsf{D}_{\mathsf{KL}}\big(\pi^{\varepsilon}_\star\big\|\pi^{\varepsilon}_f\big), \tag{34}$$

where (see (11))

$$d\pi^{\varepsilon}_f(x,y)=\frac{\exp\left(\frac{f(x)-c(x,y)}{\varepsilon}\right)}{\int_{\mathcal{X}}\exp\left(\frac{f-c(\cdot,y)}{\varepsilon}\right)d\mu}d\mu\otimes\nu(x,y).$$

The derivation is inspired by the proof of [37, Theorem 2], with several technical modifications. Since $\mu\in\mathcal{P}_{\mathsf{ac}}(\mathcal{X})$ with Lebesgue density $\frac{d\mu}{dx}$, define its energy function $E_\mu:\mathcal{X}\to\mathbb{R}$ by $\frac{d\mu(x)}{dx}\propto\exp\left(-E_\mu(x)\right)$. Also define conditional distribution $d\pi^{\varepsilon}_f(\cdot|y):=\frac{d\pi_f(\cdot,y)}{d\nu(y)}$, and set $\tilde{f}:=f-\varepsilon E_\mu(x)$. We have

$$\frac{d\pi^{\varepsilon}_f(x|y)}{dx}=\frac{d\pi^{\varepsilon}_f(x,y)}{d\nu(y)dx}=\frac{\exp\left(\frac{f(x)-c(x,y)}{\varepsilon}\right)\frac{d\mu(x)}{dx}}{\int_{\mathcal{X}}\exp\left(\frac{f(x')-c(x',y)}{\varepsilon}\right)\frac{d\mu}{dx'}(x')dx'}=\frac{\exp\left(\frac{\tilde{f}(x)-c(x,y)}{\varepsilon}\right)}{\int_{\mathcal{X}}\exp\left(\frac{\tilde{f}(x')-c(x',y)}{\varepsilon}\right)dx'}.$$

Define the shorthands $F_f(y):=\int_{\mathcal{X}}\exp\left(\frac{f(x)-c(x,y)}{\varepsilon}\right)dx$ and $Z:=\int_{\mathcal{X}}\exp\big(-E_\mu(x)\big)dx$, and note that the (c,ε)-transform of f can be expressed as

$$\begin{aligned}f^{c,\varepsilon}(y)&=-\varepsilon\log\left(\int_{\mathcal{X}}\exp\left(\frac{f(x)-c(x,y)}{\varepsilon}\right)\frac{d\mu(x)}{dx}dx\right)\\&=-\varepsilon\log\left(\int_{\mathcal{X}}\exp\left(\frac{\tilde{f}(x)-c(x,y)}{\varepsilon}\right)dx\right)+\varepsilon\log(Z)\\&=-\varepsilon\log\big(F_{\tilde{f}}(y)\big)+\varepsilon\log(Z).\end{aligned}$$

We are now ready to prove (34). For $\rho\in\mathcal{P}_{\mathsf{ac}}(\mathcal{X})$ with Lebesgue density $\frac{d\rho}{dx}$, denote the differential entropy of ρ by $\mathsf{H}(\rho):=-\int_{\mathcal{X}}\log\left(\frac{d\rho}{dx}\right)d\rho$. Consider:

$$\begin{aligned}&\Gamma(\varphi_\star)-\Gamma(f)\\&=\int_{\mathcal{X}\times\mathcal{Y}}cd\pi^{\varepsilon}_\star-\varepsilon\int_{\mathcal{Y}}\mathsf{H}\big(\pi^{\varepsilon}_\star(\cdot|y)\big)d\nu(y)+\varepsilon\mathsf{H}(\mu)\\&\qquad-\int_{\mathcal{X}}fd\mu-\int_{\mathcal{Y}}f^{c,\varepsilon}d\nu\\&=\int_{\mathcal{X}\times\mathcal{Y}}\big(c(x,y)-\tilde{f}(x)\big)d\pi^{\varepsilon}_\star(x,y)-\varepsilon\int_{\mathcal{Y}}\mathsf{H}\big(\pi^{\varepsilon}_\star(\cdot|y)\big)d\nu(y)\\&\qquad+\varepsilon\int_{\mathcal{Y}}\log\big(F_{\tilde{f}}\big)d\nu\\&=-\varepsilon\int_{\mathcal{X}\times\mathcal{Y}}\frac{\tilde{f}(x)-c(x,y)}{\varepsilon}d\pi^{\varepsilon}_\star(x,y)+\varepsilon\int_{\mathcal{X}\times\mathcal{Y}}\log\big(F_{\tilde{f}}(y)\big)d\pi^{\varepsilon}_\star(x,y)\\&\qquad-\varepsilon\int_{\mathcal{Y}}\mathsf{H}\big(\pi^{\varepsilon}_\star(\cdot|y)\big)d\nu(y)\\&=-\varepsilon\int_{\mathcal{X}\times\mathcal{Y}}\log\left(\frac{1}{F_{\tilde{f}}(y)}\exp\left(\frac{\tilde{f}(x)-c(x,y)}{\varepsilon}\right)\right)d\pi^{\varepsilon}_\star(x,y)\\&\qquad-\varepsilon\int_{\mathcal{Y}}\mathsf{H}\big(\pi^{\varepsilon}_\star(\cdot|y)\big)d\nu(y)\\&=-\varepsilon\int_{\mathcal{X}\times\mathcal{Y}}\log\left(\frac{d\pi^{\varepsilon}_f(x\mid y)}{dx}\right)d\pi^{\varepsilon}_\star(x,y)-\varepsilon\int_{\mathcal{Y}}\mathsf{H}\big(\pi^{\varepsilon}_\star(\cdot|y)\big)d\nu(y)\end{aligned}$$

$$
\begin{aligned}
&= -\varepsilon \int_{\mathcal{Y}} \int_{\mathcal{X}} \log\left(\frac{d\pi_f^\varepsilon(x|y)}{dx}\right) d\pi_\star^\varepsilon(x|y) d\nu(y) \\
&\qquad + \varepsilon \int_{\mathcal{Y}} \int_{\mathcal{X}} \log\left(\frac{d\pi^\varepsilon(x|y)}{dx}\right) d\pi_\star^\varepsilon(x|y) d\nu(y) \\
&= \varepsilon \int_{\mathcal{Y}} \int_{\mathcal{X}} \log\left(\frac{d\pi_\star^\varepsilon(x|y)}{d\pi_f^\varepsilon(x|y)}\right) d\pi_\star^\varepsilon(x|y) d\nu(y) \\
&= \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi_\star^\varepsilon(x,y)}{d\pi_f^\varepsilon(x,y)}\right) d\pi_\star^\varepsilon(x,y) \\
&= \varepsilon \mathsf{D}_{\mathsf{KL}}\left(\pi_\star^\varepsilon \middle\| \pi_f^\varepsilon\right).
\end{aligned}
$$

Recalling that $\hat{f}_\star$ is a NN that optimizes the NE $\widehat{\mathsf{OT}}_{k,a}^\varepsilon(X^n, Y^n)$ from (10), and plugging it into (34) yields $\mathsf{D}_{\mathsf{KL}}\left(\pi_\star^\varepsilon \middle\| \pi_{\hat{f}_\star}^\varepsilon\right) = \varepsilon^{-1}\left[\Gamma\left(\varphi_\star\right) - \Gamma(\hat{f}_\star)\right]$. Thus, to prove the KL divergence bound from Theorem 3, it suffices to control the gap between the Γ functionals on the RHS above.

Write $f_\star$ for a NN that maximizes the population-level neural EOT cost $\mathsf{OT}_{k,a}^\varepsilon(\mu,\nu)$ (see (15)). Define $\widehat{\Gamma}(f) := \frac{1}{n}\sum_{i=1}^n \left[f(X_i) + f^{c,\varepsilon}(Y_i)\right]$ for the optimization objective in the problem $\widehat{\mathsf{OT}}_{k,a}^\varepsilon(X^n, Y^n)$ (see (10)), and note that $\hat{f}_\star$ is a maximizer of $\widehat{\Gamma}$. We now have

$$
\begin{aligned}
&\varepsilon \mathsf{D}_{\mathsf{KL}}\left(\pi_\star^\varepsilon \middle\| \pi_{\hat{f}_\star}^\varepsilon\right) \\
&= \Gamma\left(\varphi_\star\right) - \Gamma\left(f_\star\right) + \Gamma\left(f_\star\right) - \widehat{\Gamma}\left(\hat{f}_\star\right) + \widehat{\Gamma}\left(\hat{f}_\star\right) - \Gamma\left(\hat{f}_\star\right) \\
&= \underbrace{\mathsf{OT}^\varepsilon(\mu,\nu) - \mathsf{OT}_{k,a}^\varepsilon(\mu,\nu)}_{\text{(I)}} + \underbrace{\mathsf{OT}_{k,a}^\varepsilon(\mu,\nu) - \widehat{\mathsf{OT}}_{k,a}^\varepsilon(X^n, Y^n)}_{\text{(II)}} \\
&+ \underbrace{\widehat{\Gamma}\left(\hat{f}_\star\right) - \Gamma\left(\hat{f}_\star\right)}_{\text{(III)}}.
\end{aligned}
\tag{35}
$$

Setting $a = \bar{c}_{b,d}$ as in the proof of Theorem 1 (see (20) and (22)) and taking an expectation (over the data) on both sides, Terms (I) and (II) are controlled, respectively, by the approximation error and empirical estimation error from (23) and (33). For Term (III), consider

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{\Gamma}\left(\hat{f}_\star\right) - \Gamma\left(\hat{f}_\star\right)\right] \\
&\le \mathbb{E}\left[\sup_{f\in\mathcal{F}_{k,d}(\bar{c}_{b,d})} \left|\Gamma(f) - \widehat{\Gamma}(f)\right|\right] \\
&\le n^{-\frac{1}{2}}\mathbb{E}\left[\sup_{f\in\mathcal{F}_{k,d}(\bar{c}_{b,d})} n^{-\frac{1}{2}}\left|\sum_{i=1}^n \left(f\left(X_i\right) - \mathbb{E}_\mu[f]\right)\right|\right] \\
&\qquad + n^{-\frac{1}{2}}\mathbb{E}\left[\sup_{f\in\mathcal{F}^{c,\varepsilon}} n^{-\frac{1}{2}}\left|\sum_{j=1}^n \left(f\left(Y_j\right) - \mathbb{E}_\nu[f]\right)\right|\right].
\end{aligned}
$$

where the last step follows similarly to (24). Notably, the RHS above is also bounded by the estimation error bound from (33). Combining the above completes the proof. □

APPENDIX B
PROOFS OF TECHNICAL LEMMAS

A. Proof of Lemma 1

The existence of optimal potentials follows by standard EOT arguments [38, Lemma 1]. Recall that EOT potentials are unique up to additive constants. Thus, let $(\varphi_0, \psi_0) \in L^1(\mu) \times L^1(\nu)$ be optimal EOT potentials for the cost c, solving dual formulation (6), and we can assume without loss of generality that $\int \varphi_0 d\mu = \int \psi_0 d\nu = \frac{1}{2}\mathsf{OT}^\varepsilon(\mu,\nu)$.

Recall that the optimal potentials satisfies the Schrödinger system from (8). Define new functions φ and ψ as

$$
\begin{aligned}
&\varphi(x) := -\varepsilon \log \textstyle\int_{\mathcal{Y}} \exp\left(\frac{\psi_0(y) - c(x,y)}{\varepsilon}\right) d\nu(y), && x \in \mathcal{X} \\
&\psi(y) := \varphi^{c,\varepsilon}(y), && y \in \mathcal{Y}.
\end{aligned}
$$

These integrals are clearly well-defined as the integrands are everywhere positive on $\mathcal{X}$ and $\mathcal{Y}$, and φ_0, ψ_0 are defined on the supports of μ, ν respectively. Now We show that φ, ψ are pointwise finite. For the upper bound, by Jensen's inequality, we have

$$
\varphi(x) \le \int_{\mathcal{Y}} \frac{1}{2}\|x - y\|^2 - \psi_0(y) d\nu(y) \le 2d,
$$

the second inequality follows from $\int \psi_0 d\nu = \frac{1}{2}\mathsf{OT}^\varepsilon(\mu,\nu) \ge 0$. The upper bound holds similarly for ψ on $\mathcal{Y}$ and ψ_0 on the support of ν. For lower bound, we have

$$
-\varphi(x) \le \varepsilon \log \int_{\mathcal{Y}} \exp\left(\frac{2d}{\varepsilon}\right) d\nu(y) = 2d.
$$

Note that φ is defined on $\mathcal{X}$, with pointwise bounds proven above. By Jensen's inquality,

$$
\begin{aligned}
&\int_{\mathcal{X}} \left(\varphi_0 - \varphi\right) d\mu + \int_{\mathcal{Y}} \left(\psi_0 - \psi\right) d\nu \\
&\le \varepsilon \log \int_{\mathcal{X}} \exp\left(\frac{\varphi_0 - \varphi}{\varepsilon}\right) d\mu + \log \int_{\mathcal{Y}} \exp\left(\frac{\psi_0 - \psi}{\varepsilon}\right) d\nu \\
&= \varepsilon \log \int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{\varphi_0(x) + \psi_0(y) - c(x,y)}{\varepsilon}\right) d\mu \otimes \nu \\
&\qquad + \varepsilon \log \int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{\varphi(x) + \psi_0(y) - c(x,y)}{\varepsilon}\right) d\mu \otimes \nu \\
&= 0.
\end{aligned}
$$

Since (φ_0, ψ_0) maximizes (6), so does (φ, ψ) and thus they are also optimal potentials. Therefore, φ solves semi-dual formulation (7). By the strict concavity of the logarithm function we further conclude that $\varphi = \varphi_0$ μ-a.s and $\psi = \psi_0$ ν-a.s.

The differentiability of (φ, ψ) is clear from their definition. For any multi-index α, the multivariate Faa di Bruno formula (see [51, Corollary 2.10]) implies

$$
\begin{aligned}
&- D^\alpha \varphi(x) = \\
&\varepsilon \sum_{r=1}^{|\alpha|} \sum_{p(\alpha,r)} l(\alpha, r, \boldsymbol{k}, \boldsymbol{\beta}) \prod_{j=1}^{|\alpha|} \left(\frac{D^{\beta_j} \int \exp(\frac{\psi_0(y) - c(x,y)}{\varepsilon}) d\nu(y)}{\int \exp(\frac{\psi_0(y) - c(x,y)}{\varepsilon}) d\nu(y)}\right)^{k_j},
\end{aligned}
\tag{36}
$$

where $l(\alpha, r, \boldsymbol{k}, \boldsymbol{\beta}) = \frac{\alpha!(r-1)!(-1)^{r-1}}{\prod_{j=1}^{|\alpha|}(k_j!)(\beta_j!)^{k_j}}$, $p(\alpha, r)$ is the collection of all tuples $\left(k_1, \cdots, k_{|\alpha|}; \beta_1, \cdots, \beta_{|\alpha|}\right) \in \mathbb{N}^{|\alpha|} \times \mathbb{N}^{d\times|\alpha|}$ satisfying $\sum_{i=1}^{|\alpha|} k_i = r, \sum_{i=1}^{|\alpha|} k_i\beta_i = \alpha$, and for which there exists $s \in \{1, \ldots, |\alpha|\}$ such that $k_i = 0$ and $\beta_i = 0$ for all $i = 1, \ldots |\alpha| - s, k_i > 0$ for all $i = |\alpha| - s + 1, \ldots, |\alpha|$, and $0 \prec \beta_{|\alpha|-s+1} \prec \cdots \prec \beta_{|\alpha|}$. For a detailed discussion of this set including the linear order $\prec$, please refer to [51]. For the current proof we only use the fact that the number of elements in this set solely depends on $|\alpha|$ and r. Given the above, it clearly suffices to bound $\left|D^{\beta_j}\int \exp\left(\psi_0(y) - c(x,y)\right) d\nu(y)\right|$. First, we apply the same formula to $D^{\beta_j} e^{-c(x,y)/\varepsilon}$ and obtain

$$D^{\beta_j} e^{\frac{-c(x,y)}{\varepsilon}} = \sum_{r'=1}^{|\beta_j|} (\frac{-1}{2\varepsilon})^{r'} \sum_{p(\beta_j, r')} l(\beta_j, r', \boldsymbol{k'}, \boldsymbol{\eta}) e^{\frac{-c(x,y)}{\varepsilon}} \prod_{i=1}^{|\beta_j|} \left(D^{\eta_i}\|x-y\|^2\right)^{k'_i}, \tag{37}$$

where $p\left(\beta_j, r'\right)$ defined similarly to the above. Observe that

$$\left|D^{\eta_i}\left(\|x-y\|^2\right)\right| \le 4\left(1 + \|x-y\|\right) \le 4(1 + 2\sqrt{d}),$$

where the first inequality follows from proof of Lemma 3 in [49]. Consequently, for $0 < \varepsilon < 1$, we have

$$\left|\frac{D^{\beta_j}\int \exp(\frac{\psi_0(y)-c(x,y)}{\varepsilon}) d\nu(y)}{\int \exp(\frac{\psi_0(y)-c(x,y)}{\varepsilon}) d\nu(y)}\right| \le C_{\beta_j}\varepsilon^{-|\beta_j|}(1 + 2\sqrt{d})^{|\beta_j|},$$

and for $\varepsilon \ge 1$,

$$\left|\frac{D^{\beta_j}\int \exp(\frac{\psi_0(y)-c(x,y)}{\varepsilon}) d\nu(y)}{\int \exp(\frac{\psi_0(y)-c(x,y)}{\varepsilon}) d\nu(y)}\right| \le C_{\beta_j}\varepsilon^{-1}(1 + 2\sqrt{d})^{|\beta_j|},$$

Plugging back, we obtain

$$|D^\alpha\varphi(x)| \le C_{|\alpha|}(1 + \varepsilon^{1-|\alpha|})(1 + 2\sqrt{d})^{|\alpha|}.$$

Analogous bound holds for ψ. □

B. Proof of Lemma 2

For any $f \in \mathcal{F}_{k,a}(a)$, we know that $\|f\|_{\infty,\mathcal{X}} \le 3a(\|\mathcal{X}\| + 1) = 6a$, so NNs are uniformly bounded. This implies that $\mathsf{OT}^\varepsilon(\mu,\nu) \ge \mathsf{OT}^\varepsilon_{k,a}(\mu,\nu)$. Since φ satisfies (17), it's uniformly bounded on $\mathcal{X}$. Then, the following holds:

$$\begin{aligned}\left|\mathsf{OT}^\varepsilon(\mu,\nu) - \mathsf{OT}^\varepsilon_{k,a}(\mu,\nu)\right| =& \mathsf{OT}^\varepsilon(\mu,\nu) - \mathsf{OT}^\varepsilon_{k,a}(\mu,\nu)\\ \le& \mathbb{E}_\mu|\varphi - f| + \mathbb{E}_\nu|\varphi^{c,\varepsilon} - f^{c,\varepsilon}|\\ \le& 2\|\varphi - f\|_{\infty,\mathcal{X}}.\end{aligned}$$

The last inequality holds by an observation that,

$$|\varphi^{c,\varepsilon}(y) - f^{c,\varepsilon}(y)| \le \|\varphi - f\|_{\infty,\mathcal{X}}, \quad \forall y \in \mathcal{Y}. \tag{38}$$

Indeed, note that for any $y \in \mathcal{Y}$,

$$\begin{aligned}&\varphi^{c,\varepsilon}(y) - f^{c,\varepsilon}(y)\\ &= -\varepsilon\log\left(\frac{\int \exp\left(\frac{\varphi(x)-c(x,y)}{\varepsilon}\right) d\mu(x)}{\int \exp\left(\frac{f(x)-c(x,y)}{\varepsilon}\right) d\mu(x)}\right)\\ &= -\varepsilon\log\left(\frac{\int \exp\left(\frac{\varphi(x)-f(x)}{\varepsilon}\right)\exp\left(\frac{f(x)-c(x,y)}{\varepsilon}\right) d\mu(x)}{\int \exp\left(\frac{f(x)-c(x,y)}{\varepsilon}\right) d\mu(x)}\right)\\ &\ge -\varepsilon\log\left(\frac{\int \exp\left(\frac{\|\varphi-f\|_{\infty,\mathcal{X}}}{\varepsilon}\right)\exp\left(\frac{f(x)-c(x,y)}{\varepsilon}\right) d\mu(x)}{\int \exp\left(\frac{f(x)-c(x,y)}{\varepsilon}\right) d\mu(x)}\right)\\ &= -\|\varphi - f\|_{\infty,\mathcal{X}}.\end{aligned}$$

Similarly, we can have that

$$\begin{aligned}&\varphi^{c,\varepsilon}(y) - f^{c,\varepsilon}(y)\\ &\le -\varepsilon\log\left(\frac{\int \exp\left(\frac{-\|\varphi-f\|_{\infty,\mathcal{X}}}{\varepsilon}\right)\exp\left(\frac{f(x)-c(x,y)}{\varepsilon}\right) d\mu(x)}{\int \exp\left(\frac{f(x)-c(x,y)}{\varepsilon}\right) d\mu(x)}\right)\\ &= \|\varphi - f\|_{\infty,\mathcal{X}}.\end{aligned}$$

□