

A Sufficient Condition to Assumption 3.2

Proposition 3.3 requires an upper bound on $\|\nabla\varphi_0(x)\|_2$. The following result shows that such a bound can be derived from appropriate concentration properties of μ and ν . We begin with a definition describing the right-tail concentration behavior of the target measure ν :

Definition A.1. Let ν be a probability measure on \mathbb{R}^d . For function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, we denote $\nu f := \int f(x) \nu(dx) = \mathbb{E}_{X \sim \nu}[f(X)]$.

1. $X \sim \nu$ satisfies a Gaussian concentration inequality with constant $\beta > 0$, if for any 1-Lipschitz function f and any $r \geq 0$,

$$\mathbb{P}(f(X) - \nu f \geq r) \leq \exp(-\beta \cdot r^2/2). \quad (24)$$

2. $X \sim \nu$ satisfies an exponential concentration inequality with constant $\alpha > 0$, if for any 1-Lipschitz function f , and any $r \geq 0$,

$$\mathbb{P}(f(X) - \nu f \geq r) \leq \exp(-\alpha r). \quad (25)$$

3. $X \sim \nu$ satisfies a Polynomial concentration inequality with constant $\gamma, C_\gamma, r_0 > 0$, if for any 1-Lipschitz function f and any $r \geq r_0 \geq 0$,

$$\mathbb{P}(f(X) - \nu f \geq r) \leq C_\gamma \cdot r^{-\gamma}. \quad (26)$$

We now present a sufficient condition for Assumption 3.2.

Proposition A.2 (Sufficient condition to Assumption 3.2). Suppose μ has density $p(x) = \exp(-V(x))$, and suppose there exists a function $M : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$\|\nabla V(x)\|_2 \leq M(r), \quad \text{for all } x \in B(0, r)$$

Let $\nabla\varphi_0$ be the OT map from μ to ν . Then,

1. If ν satisfies Equation (24),

$$\|\nabla\varphi_0(x)\|_2 \leq \sqrt{\frac{128}{\beta} \left((\|x\|_2 + 6\sqrt{d}) \cdot M(\|x\|_2 + 4\sqrt{d}) + V(0) - \log C_d \right)}$$

2. If ν satisfies Equation (25),

$$\|\nabla\varphi_0(x)\|_2 \leq \frac{8}{\alpha} \left((\|x\|_2 + 6\sqrt{d}) \cdot M(\|x\|_2 + 4\sqrt{d}) + V(0) - \log C_d \right).$$

3. If ν satisfies Equation (26),

$$\log\left(\frac{\|\nabla\varphi_0(x)\|_2}{8}\right) \leq \frac{1}{\gamma} \left((\|x\|_2 + 6\sqrt{d}) \cdot M(\|x\|_2 + 4\sqrt{d}) + V(0) - \log(C_d/C_\gamma) \right).$$

where $C_d := \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} (2\sqrt{d})^d$.

Proof of Proposition A.2. The proposition is indeed a direct extension of Theorem 1.1 in [2]. We only focus on the case where ν satisfies the Gaussian concentration inequality (24). The two remaining cases can be handled by exactly the same argument.

Without loss of generality, assume

$$\|\nabla\varphi_0(x)\|_2 \geq 8(1 + \|x\|_2^2), \quad \|\nabla\varphi_0(x)\|_2 \geq 6\sqrt{d}.$$

Otherwise the bound is trivial.

For an arbitrary x , define $u := \frac{\nabla\varphi_0(x) - x}{\|\nabla\varphi_0(x) - x\|_2}$. Meanwhile, define $f : \mathbb{R}^d \rightarrow \mathbb{R}$ as

$$f(z) := \langle z - \nabla\varphi_0(x), u \rangle + \frac{1}{2} \|z - \nabla\varphi_0(x)\|_2.$$

From the proof of Theorem 1.1 of [2], f is $3/2$ -lipschitz, and $\nabla\varphi_0(B) \subset \{f \geq 0\}$, where $B := B(x + 4\sqrt{d} \cdot u, 2\sqrt{d})$. Moreover, it holds that

$$\int f d\mu \leq -\|\nabla\varphi_0(x)\|_2/8, \quad \text{for all } x.$$

Since ν satisfies Equation (24), we have: $\mathbb{P}(f(X) - \nu f \geq r) \leq \exp(-\beta \cdot r^2/2)$.

Meanwhile, since ν is standard Gaussian, for some $C > 0$,

$$\nu(\nabla\varphi_0(B)) \leq \nu(\{f \geq 0\}) \leq \nu\left(\left\{f \geq \int f d\nu + \frac{\|\nabla\varphi_0(x)\|_2}{8}\right\}\right) \leq \exp\left\{-\frac{\beta}{128}\|\nabla\varphi_0(x)\|_2^2\right\}.$$

On the other hand, we can lower bound $\mu(B)$ as

$$\begin{aligned} \mu(B) &= \int_{B(x+4\sqrt{d}u, 2\sqrt{d})} p(v) dv = \int_{B(0, 2\sqrt{d})} p(x + 4\sqrt{d}u + v) dv \\ &\geq \int_{B(0, 2\sqrt{d})} p(x + 4\sqrt{d}u) \cdot \exp\{-2\sqrt{d} \cdot M(\|x\|_2 + 4\sqrt{d})\} dv \\ &= \gamma(B(0, 2\sqrt{d})) \cdot \exp\{-2\sqrt{d} \cdot M(\|x\|_2 + 4\sqrt{d}) - V(x + 4\sqrt{d}u)\} \\ &= C_d \cdot \exp\{-2\sqrt{d} \cdot M(\|x\|_2 + 4\sqrt{d}) - V(x + 4\sqrt{d}u)\} \end{aligned} \quad (27)$$

where $C_d := \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} (2\sqrt{d})^d$ is the volume of the ball $B(0, 2\sqrt{d})$. Now combining $\mu(B) \leq \nu(\nabla\varphi_0(B))$, we have

$$C_d \cdot \exp\{-2\sqrt{d} \cdot M(\|x\|_2 + 4\sqrt{d}) - V(x + 4\sqrt{d}u)\} \leq \exp\left\{-\frac{\beta}{128}\|\nabla\varphi_0(x)\|_2^2\right\}$$

Thus

$$\begin{aligned} \frac{\beta}{128}\|\nabla\varphi_0(x)\|_2^2 &\leq 2\sqrt{d} \cdot M(\|x\|_2 + 4\sqrt{d}) + V(x + 4\sqrt{d}u) - \log C_d \\ &\leq 2\sqrt{d} \cdot M(\|x\|_2 + 4\sqrt{d}) + M(\|x\|_2 + 4\sqrt{d})(\|x\|_2 + 4\sqrt{d}) + V(0) - \log C_d \\ &\leq (\|x\|_2 + 6\sqrt{d}) \cdot M(\|x\|_2 + 4\sqrt{d}) + V(0) - \log C_d \end{aligned} \quad (28)$$

Re-organizing terms in Equation (28) gives an upper bound of $\|\nabla\varphi_0(x)\|_2$:

$$\|\nabla\varphi_0(x)\|_2 \leq \sqrt{\frac{128}{\beta} \left((\|x\|_2 + 6\sqrt{d}) \cdot M(\|x\|_2 + 4\sqrt{d}) + V(0) - \log C_d \right)}$$

□

Example. Let μ has density $\mu \propto \exp(-\|x\|_2^{2p}/2)$ for some $p > 0$, and $Y \sim \nu$ has polynomial tails, i.e., $\nu(\|Y\|_2 > t) \lesssim t^{-q}$ for some $q > 0$. Then, it holds that

$$\|\nabla\varphi_0(x)\|_2 \lesssim \exp\left(c \cdot \frac{p}{q}\|x\|_2^{2p}\right) =: l(x),$$

where $c > 0$ is only related to p . Furthermore, when $q > 8cp$, we have $l(X) \in L^4(\mu)$.

Proof. We first show that ν satisfies Equation (26).

For any 1-Lipschitz function f , let $m := |\nu f - f(0)|$, it holds that

$$f(Y) - \nu f \leq f(Y) - f(0) + m \leq \|Y\|_2 + m$$

Then for any $r > 2m$,

$$\nu(\{f(Y) - \nu f \geq r\}) \leq \nu(\{\|Y\|_2 \geq r - m\}) \leq \nu(\{\|Y\|_2 \geq \frac{r}{2}\}) \lesssim \left(\frac{r}{2}\right)^{-q}$$

Hence, polynomial concentration inequality (26) holds with constant $C_\gamma = 2^q$ and $\gamma = q$.

Now we apply Proposition A.2 with

$$V(x) = \frac{1}{2} \|x\|_2^{2p}, \quad M(r) = p r^{2p-1}.$$

The Proposition gives

$$\log\left(\frac{\|\nabla\varphi_0(x)\|_2}{8}\right) \lesssim \frac{1}{q} \left((\|x\|_2 + 6\sqrt{d}) p (\|x\|_2 + 4\sqrt{d})^{2p-1} - \log \frac{C_d}{C_\gamma} \right).$$

Exponentiating and absorbing constants into a single $c > 0$ yields

$$\|\nabla\varphi_0(x)\|_2 \lesssim 8 \cdot C_{p,q,d} \exp \left\{ c \frac{p}{q} \|x\|_2^{2p} \right\} =: l(x),$$

where the suppressed constants come from the definitions of μ, ν . Constant $C_{p,q,d}$ is only related to p, q, d , and constant c is only related to p .

In the last step, we check the integrability of $l(X)$. Recall μ has density $p(x) \propto \exp(-\|x\|_2^{2p}/2)$, we have

$$\int l(x)^4 \mu(dx) \lesssim \int \exp \left\{ \left(4c \frac{p}{q} - \frac{1}{2} \right) \cdot \|x\|_2^{2p} \right\} dx$$

RHS is integrable when $4c \frac{p}{q} - \frac{1}{2} < 0$. Hence, when $q > 8cp$, $l(X) \in L^4(\mu)$.

□

B Proof of Proposition 3.3

Proof of Proposition 3.3. Let $y = \nabla\varphi_0(x)$, and $z = x + \frac{1}{D_x}(y - \nabla\varphi(x))$. Note that z is a point totally determined by x , and $D_x > 0$ to be specified.

Note that z lies in a bigger ball $B(0, r(x))$, i.e.

$$\|z\|_2 \leq \|x\|_2 + \frac{1}{D_x} u(x) =: r(x)$$

In particular, if we want to restrict $x, z \in B(0, R)$, it is sufficient to restrict x within the area $\{x : r(x) \leq R\}$.

We begin by proving that the sieved convex conjugate $\varphi_R^*(y)$ satisfies a quadratic lower bound: The second order Taylor expansion of φ gives

$$\varphi(z) \leq \varphi(x) + \langle \nabla\varphi(x), z - x \rangle + \frac{1}{2} U_2(r) \cdot \|z - x\|_2^2 \quad (29)$$

Then on $\{x : r(x) \leq R\}$, the corresponding z lies in $B(0, R)$. Replacing $\varphi(z)$ with its quadratic upper bound derived in Equation (29), the sieved convex conjugate restricted on $B(0, R)$ satisfies

$$\begin{aligned} \varphi_R^*(y) &\geq \langle z, y \rangle - \varphi(z) \geq \langle z, y \rangle - \varphi(x) - \langle \nabla\varphi(x), z - x \rangle - \frac{1}{2} U_2(r) \cdot \|z - x\|_2^2 \\ &\geq \langle x, y \rangle - \varphi(x) + \frac{1}{D_x} \|y - \nabla\varphi(x)\|_2^2 - \frac{1}{2} U_2(r) \cdot \frac{1}{D_x^2} \|y - \nabla\varphi(x)\|_2^2 \\ &= \varphi_0(x) + \varphi_0^*(y) - \varphi(x) + \frac{1}{D_x} \|y - \nabla\varphi(x)\|_2^2 \cdot \left(1 - \frac{U_2(r)}{2D_x}\right) \\ &\geq \varphi_0(x) + \varphi_0^*(y) - \varphi(x) + \frac{1}{2D_x} \|y - \nabla\varphi(x)\|_2^2, \end{aligned} \quad (30)$$

where the last inequality holds if $D_x \geq U_2(r(x))$. Re-organizing Equation (30) gives

$$\frac{1}{2D_x} \|y - \nabla\varphi(x)\|_2^2 \leq \varphi(x) + \varphi_R^*(y) - \varphi_0(x) - \varphi_0^*(y) \quad (31)$$

Note that

$$D_x := U_2(\|x\|_2 + u(x)) \geq U_2\left(\|x\|_2 + \frac{u(x)}{D_x}\right) = U_2(r(x)),$$

Thus, we set $D_x := U_2(\|x\|_2 + u(x)) \geq 1$. Consequently,

$$r(x) = \|x\|_2 + \frac{u(x)}{U_2(\|x\|_2 + u(x))}$$

Now we define $R_\varepsilon, \tilde{R}_\varepsilon$ as

$$\mu(\|X\|_2 > R_\varepsilon) \leq \varepsilon, \quad \tilde{R}_\varepsilon \geq \sup_{\|x\|_2 \leq R_\varepsilon} \left\{ \|x\|_2 + \frac{u(x)}{U_2(\|x\|_2 + u(x))} \right\} \quad (32)$$

It then holds that, for any $x \in B(0, R_\varepsilon), z \in B(0, \tilde{R}_\varepsilon)$. As a result, Equation (31) holds for any $x \in B(0, R_\varepsilon)$.

We integrate both sides of (31) on $B(0, R_\varepsilon)$ to get

$$\int_{B(0, R_\varepsilon)} \frac{1}{2D_x} \|y - \nabla \varphi(x)\|_2^2 \mu(\mathrm{d}x) \leq \int_{B(0, R_\varepsilon)} \varphi(x) + \varphi_R^*(y) - \varphi_0(x) - \varphi_0^*(y) \mu(\mathrm{d}x) \quad (33)$$

Note that $1 \leq \frac{2U_2(R_\varepsilon + U_1(R_\varepsilon))}{2D_x}$ on $B(0, R_\varepsilon)$, so that

$$\int_{B(0, R_\varepsilon)} \|y - \nabla \varphi(x)\|_2^2 \mu(\mathrm{d}x) \leq 2U_2(R_\varepsilon + U_1(R_\varepsilon)) \cdot \int_{B(0, R_\varepsilon)} \frac{1}{2D_x} \|y - \nabla \varphi(x)\|_2^2 \mu(\mathrm{d}x) \quad (34)$$

Finally, on $\{\|x\|_2 > R_\varepsilon\}$ we bound $\|y - \nabla \varphi(x)\|_2 \leq u(x)$ and apply Cauchy–Schwarz Inequality:

$$\int_{\|x\|_2 > R_\varepsilon} \|y - \nabla \varphi(x)\|_2^2 \mu(\mathrm{d}x) \leq \int_{\|x\|_2 > R_\varepsilon} u(x)^2 \mu(\mathrm{d}x) \leq \sqrt{\mu(u^4(X)) \cdot \mu(\|X\|_2 > R_\varepsilon)} \quad (35)$$

Combining (33), (34), and (35), we conclude that

$$\|\nabla \varphi - \nabla \varphi_0\|_{L^2(\mu)}^2 \leq 2U_2(R_\varepsilon + U_1(R_\varepsilon)) \cdot r_\varepsilon(\varphi) + \|u\|_{L^4(\mu)}^2 \cdot \varepsilon^{\frac{1}{2}}, \quad (36)$$

where

$$r_\varepsilon(\varphi) := \int_{B(0, R_\varepsilon)} \varphi(x) + \varphi_{\tilde{R}_\varepsilon}^*(\nabla \varphi_0(x)) \mu(\mathrm{d}x) - \int_{B(0, R_\varepsilon)} \varphi_0(x) + \varphi_0^*(\nabla \varphi_0(x)) \mu(\mathrm{d}x). \quad (37)$$

□

C Proof of Proposition 3.4

Proof of Proposition 3.4. Let $\varepsilon > 0$ and R_ε satisfies $\mu(\|X\|_2 > R_\varepsilon) \leq \varepsilon$.

For any estimated distribution μ_n, ν_m , we estimate Brenier potential within \mathcal{F} , i.e.

$$\tilde{\varphi}_{n,m} := \arg \min_{\varphi \in \mathcal{F}} \mu_n \varphi + \nu_m \varphi_{\tilde{R}_\varepsilon}^* \quad (38)$$

Since $\tilde{\varphi}_{n,m}$ minimizes the empirical loss, it holds for any $\bar{\varphi} \in \mathcal{F}$ that

$$\mu_n \tilde{\varphi}_{n,m} + \nu_m \tilde{\varphi}_{n,m, \tilde{R}_\varepsilon}^* \leq \mu_n \bar{\varphi} + \nu_m \bar{\varphi}_{\tilde{R}_\varepsilon}^* \quad (39)$$

Recall the truncated excess risk define in Equation (11):

$$\tilde{r}_{n,m,\varepsilon} := \int_{B(0, R_\varepsilon)} \tilde{\varphi}_{n,m}(x) + \tilde{\varphi}_{n,m, \tilde{R}_\varepsilon}^*(y) - \varphi_0(x) - \varphi_0^*(y) \mu(\mathrm{d}x),$$

where $y := \nabla \varphi_0(x)$. We write for brevity

$$\tilde{r}_{n,m} := \tilde{r}_{n,m,\varepsilon}, \quad \mu_\varepsilon f := \int_{\|x\| \leq R_\varepsilon} f(x) \mu(\mathrm{d}x), \quad \nu_\varepsilon g := \int_{y = \nabla \varphi_0(x), \|x\| \leq R_\varepsilon} g(y) \nu(\mathrm{d}y).$$

Then it holds for $\tilde{r}_{n,m}$ that

$$\begin{aligned}
\tilde{r}_{n,m} &= \mu_\varepsilon(\tilde{\varphi}_{n,m} - \varphi_0) + \nu_\varepsilon(\tilde{\varphi}_{n,m,\tilde{R}_\varepsilon}^* - \varphi_0^*) \\
&= \mu_\varepsilon(\tilde{\varphi}_{n,m} - \bar{\varphi}) + \mu_\varepsilon(\bar{\varphi} - \varphi_0) + \nu_\varepsilon(\tilde{\varphi}_{n,m,\tilde{R}_\varepsilon}^* - \bar{\varphi}_{\tilde{R}_\varepsilon}^*) + \nu_\varepsilon(\bar{\varphi}_{\tilde{R}_\varepsilon}^* - \varphi_0^*) \\
&= \mu_\varepsilon(\tilde{\varphi}_{n,m} - \bar{\varphi}) + \nu_\varepsilon(\tilde{\varphi}_{n,m,\tilde{R}_\varepsilon}^* - \bar{\varphi}_{\tilde{R}_\varepsilon}^*) + \mu_\varepsilon(\bar{\varphi} - \varphi_0) + \nu_\varepsilon(\bar{\varphi}_{\tilde{R}_\varepsilon}^* - \varphi_0^*) \quad (40)
\end{aligned}$$

The last two terms are approximation errors for φ_0 and φ_0^* . It's easy to see $|\mu_\varepsilon(\bar{\varphi} - \varphi_0)| \leq \|\bar{\varphi} - \varphi_0\|_{L^\infty(B(0,\tilde{R}_\varepsilon))}$. Now we consider the upper bound of $\nu_\varepsilon(\bar{\varphi}_{\tilde{R}_\varepsilon}^* - \varphi_0^*)$. Take $u^* \in B(0,\tilde{R}_\varepsilon)$ which attains the supremum in $\bar{\varphi}_{\tilde{R}_\varepsilon}^*(y)$, it holds that

$$\begin{aligned}
\bar{\varphi}_{\tilde{R}_\varepsilon}^*(y) - \varphi_0^*(y) &= \sup_{\|u\|_2 \leq \tilde{R}_\varepsilon} \{\langle u, y \rangle - \bar{\varphi}(u)\} - \sup_{u \in \mathbb{R}} \{\langle u, y \rangle - \varphi_0(u)\} \\
&\leq \langle u^*, y \rangle - \bar{\varphi}(u^*) - \langle u^*, y \rangle + \varphi_0(u^*) \leq \|\bar{\varphi} - \varphi_0\|_{L^\infty(B(0,\tilde{R}_\varepsilon))} \quad (41)
\end{aligned}$$

While Equation (41) provides an upper bound for $\bar{\varphi}_{\tilde{R}_\varepsilon}^* - \varphi_0^*$, by $\varphi_0(x) + \varphi_0^*(\nabla \varphi_0(x)) = \langle x, \nabla \varphi_0(x) \rangle$, the lower bound follows that

$$\begin{aligned}
\nu_\varepsilon(\bar{\varphi}_{\tilde{R}_\varepsilon}^* - \varphi_0^*) &= \int_{B(0,R_\varepsilon)} \sup_{\|u\|_2 \leq \tilde{R}_\varepsilon} \{\langle u, \nabla \varphi_0(x) \rangle - \bar{\varphi}(u)\} - \langle x, \varphi_0(x) \rangle + \varphi_0(x) \mu(dx) \\
&\geq \int_{B(0,R_\varepsilon)} \langle x, \nabla \varphi_0(x) \rangle - \bar{\varphi}(x) - \langle x, \varphi_0(x) \rangle + \varphi_0(x) \mu(dx) \\
&= \int_{B(0,R_\varepsilon)} \varphi_0(x) - \bar{\varphi}(x) \mu(dx), \quad (42)
\end{aligned}$$

where the first inequality holds because $R_\varepsilon \leq \tilde{R}_\varepsilon$. Thus, combining Equations (41) and (42), we have

$$|\nu_\varepsilon(\bar{\varphi}_{\tilde{R}_\varepsilon}^* - \varphi_0^*)| \leq \|\bar{\varphi} - \varphi_0\|_{L^\infty(B(0,\tilde{R}_\varepsilon))},$$

Now Equation (40) can be bounded as

$$\tilde{r}_{n,m} \leq \mu_\varepsilon(\tilde{\varphi}_{n,m} - \bar{\varphi}) + \nu_\varepsilon(\tilde{\varphi}_{n,m,\tilde{R}_\varepsilon}^* - \bar{\varphi}_{\tilde{R}_\varepsilon}^*) + 2\|\bar{\varphi} - \varphi_0\|_{L^\infty(B(0,\tilde{R}_\varepsilon))}$$

Combine with Equation (39),

$$\tilde{r}_{n,m} \leq (\mu_\varepsilon - \mu_n)(\tilde{\varphi}_{n,m} - \bar{\varphi}) + (\nu_\varepsilon - \nu_m)(\tilde{\varphi}_{n,m,\tilde{R}_\varepsilon}^* - \bar{\varphi}_{\tilde{R}_\varepsilon}^*) + 2\|\bar{\varphi} - \varphi_0\|_{L^\infty(B(0,\tilde{R}_\varepsilon))} \quad (43)$$

Next, since $\tilde{\varphi}_{n,m}, \bar{\varphi} \in \mathcal{F}$, their difference lies in $\bar{\mathcal{F}}$, where $\bar{\mathcal{F}} := \{\varphi_1 - \varphi_2 : \varphi_1, \varphi_2 \in \mathcal{F}\}$. The first term on RHS can be bounded as

$$(\mu_n - \mu_\varepsilon)(\bar{\varphi} - \tilde{\varphi}_{n,m}) \leq \sup_{f \in \bar{\mathcal{F}}} \int_{B(0,R_\varepsilon)} f d(\mu - \mu_n) + 2\mathbb{E}_{X \sim \mu_n}[\Phi(X) \cdot \mathbb{I}(\|X\|_2 > R_\varepsilon)] \quad (44)$$

where $\bar{\mathcal{F}} := \{\varphi_1 - \varphi_2 : \varphi_1, \varphi_2 \in \mathcal{F}\}$. The first term on the RHS represents the statistical error, and the second error comes from sieving.

Similarly, for conjugate function class $\bar{\mathcal{G}} := \{g := \varphi_{1,\tilde{R}_\varepsilon}^* - \varphi_{2,\tilde{R}_\varepsilon}^* : \varphi_1, \varphi_2 \in \mathcal{F}\}$ with envelope function $G(\cdot)$,

$$\varphi_{\tilde{R}_\varepsilon}^*(y) = \sup_{x \in B(0,\tilde{R}_\varepsilon)} \langle x, y \rangle - \varphi(x) \leq \tilde{R}_\varepsilon \|y\|_2 + \sup_{x \in B(0,\tilde{R}_\varepsilon)} \Phi(x) =: G(y)$$

The second term on the RHS of Equation (43) can be bounded and decomposed in the same way,

$$\begin{aligned}
(\nu_m - \nu_\varepsilon)(\bar{\varphi}^* - \tilde{\varphi}_{n,m}^*) &\leq \sup_{g \in \bar{\mathcal{G}}} \int_{\nabla \varphi_0 B(0,R_\varepsilon)} g d(\nu - \nu_m) + 2\mathbb{E}_{Y \sim \nu_m}[G(Y) \cdot \mathbb{I}(Y \in (\nabla \varphi_0 B(0,R_\varepsilon))^c)] \\
&= \sup_{g \in \bar{\mathcal{G}}} \int_{\nabla \varphi_0 B(0,R_\varepsilon)} g d(\nu - \nu_n) + 2\mathbb{E}_{Y \sim \nu_m}[G(Y) \cdot \mathbb{I}(\|(\nabla \varphi_0)^{-1}(Y)\|_2 > R_\varepsilon)] \quad (45)
\end{aligned}$$

Therefore, combine (43),(44) and (45), we have

$$\begin{aligned}\tilde{r}_{n,m} \leq & 2\|\bar{\varphi} - \varphi_0\|_{L^\infty(B(0,\tilde{R}_\varepsilon))} + \sup_{f \in \bar{\mathcal{F}}} \int_{B(0,R_\varepsilon)} f d(\mu - \mu_n) + 2\mathbb{E}_{X \sim \mu_n}[\Phi(X) \cdot \mathbb{I}(\|X\|_2 > R_\varepsilon)] \\ & + \sup_{g \in \bar{\mathcal{G}}} \int_{\nabla \varphi_0 B(0,R_\varepsilon)} g d(\nu - \nu_m) + 2\mathbb{E}_{Y \sim \nu_m}[G(Y) \cdot \mathbb{I}(\|(\nabla \varphi_0)^{-1}(Y)\|_2 > R_\varepsilon)]\end{aligned}$$

The stated oracle inequality holds with the following definitions of the three error components.

$$\begin{aligned}\mathcal{E}_{stat} &:= \sup_{f \in \bar{\mathcal{F}}} \int_{B(0,R_\varepsilon)} f d(\mu_n - \mu) + \sup_{g \in \bar{\mathcal{G}}} \int_{B(0,R_\varepsilon)} g d(\nu_m - \nu), \\ \mathcal{E}_{sieve} &:= 2\mathbb{E}_{X \sim \mu_n}[\Phi(X) \cdot \mathbb{I}(\|X\|_2 > R_\varepsilon)] + 2\mathbb{E}_{Y \sim \nu_m}[G(Y) \cdot \mathbb{I}(\|(\nabla \varphi_0)^{-1}(Y)\|_2 > R_\varepsilon)], \\ \mathcal{E}_{app} &:= 2 \inf_{\varphi \in \bar{\mathcal{F}}} \|\varphi - \varphi_0\|_{L^\infty(B(0,\tilde{R}_\varepsilon))}.\end{aligned}$$

□

D Proof of Theorem 3.6

In this appendix, we state a complete version of Theorem 3.6 and provide its proof.

Theorem D.1 (Statistical Error with Empirical Distribution). *Under assumptions of Proposition 3.4, and Assumption 3.5, the sieved estimator $\tilde{\varphi}_{n,m}$ from Equation (8), computed with empirical measures μ_n and ν_m , satisfies:*

$$\mathbb{E}[\mathcal{E}_{stat}] \lesssim \sqrt{\frac{D_{\mathcal{F}}}{n}} \left(R_\varepsilon^{\frac{\eta}{2}} + \sup_{x \in B(0,R_\varepsilon)} \Phi(x)^{\frac{1}{2}} \right) \log(n)^{\frac{\gamma'}{2}} + \sqrt{\frac{D_{\mathcal{F}}}{m}} \left(\tilde{R}_\varepsilon^{\frac{\eta}{2}} + M^{\frac{1}{2}} \right) \log(m)^{\frac{\gamma'}{2}}, \quad (46)$$

and

$$\mathbb{E}[\mathcal{E}_{sieve}] \leq 2(\|\Phi\|_{L^2(\mu)} + \|G\|_{L^2(\nu)}) \cdot \varepsilon^{\frac{1}{2}}, \quad (47)$$

where $M := \tilde{R}_\varepsilon \sup_{x \in B(0,R_\varepsilon)} \|\nabla \varphi_0(x)\|_2 + \sup_{x \in B(0,\tilde{R}_\varepsilon)} \Phi(x)$, and the suppressed constant depends only on γ, γ' . Moreover, the bounds also hold in high probability, i.e. with probability at least $1 - \delta_1 - \delta_2$:

$$\begin{aligned}\mathcal{E}_{stat} &\lesssim \sqrt{\frac{D_{\mathcal{F}}}{n}} \left(R_\varepsilon^{\frac{\eta}{2}} + \sup_{x \in B(0,R_\varepsilon)} \Phi(x)^{\frac{1}{2}} \right) \log(n)^{\frac{\gamma'}{2}} + \sqrt{\frac{D_{\mathcal{F}}}{m}} \left(\tilde{R}_\varepsilon^{\frac{\eta}{2}} + M^{\frac{1}{2}} \right) \log(m)^{\frac{\gamma'}{2}} \\ &\quad + n^{-\frac{1}{2}} F \cdot \log^{\frac{1}{2}}\left(\frac{2}{\delta_1}\right) + m^{-\frac{1}{2}} M \cdot \log^{\frac{1}{2}}\left(\frac{2}{\delta_2}\right),\end{aligned}$$

where $F := \sup_{x \in B(0,R_\varepsilon)} 2\Phi(x)$. And with probability at least $1 - \delta_3 - \delta_4$:

$$\mathcal{E}_{sieve} \leq 2(\|\Phi\|_{L^2(\mu)} + \|G\|_{L^2(\nu)}) \cdot \varepsilon^{\frac{1}{2}} + n^{-\frac{1}{2}} 2\sqrt{\mu(\Phi^2(X))} \cdot \delta_3^{-\frac{1}{2}} + m^{-\frac{1}{2}} 2\sqrt{\nu(G^2(Y))} \cdot \delta_4^{-\frac{1}{2}}$$

Remark D.2. In the \mathcal{E}_{sieve} we only derive polynomial-type high-probability bounds, as they follow directly from the finite-moment assumptions on $\Phi(X)$. Under stronger tail conditions (e.g. sub-Gaussian or sub-Weibull concentration), one can replace these moment-based estimates with exponential-type concentration inequalities to obtain substantially sharper high-probability guarantees.

Proof of Theorem D.1. Step 1. Bounding the statistical error \mathcal{E}_{stat} : We first consider

$$\sup_{f \in \bar{\mathcal{F}}} \int_{B(0,R_\varepsilon)} f d(\mu - \mu_n)$$

We can take $\bar{\mathcal{F}}_\varepsilon := \{f \cdot \mathbb{I}(B(0,R_\varepsilon)) : f \in \bar{\mathcal{F}}\}$ as a new function class, which is bounded w.r.t. $\|\cdot\|_{L^\infty}$ norm, i.e.

$$\sup_{f \in \bar{\mathcal{F}}} \left\{ \|f \cdot \mathbb{I}(B(0,R_\varepsilon))\|_{L^\infty} \right\} \leq \sup_{x \in B(0,R_\varepsilon)} 2\Phi(x) =: F.$$

Since $B(0, R_\varepsilon) \subset [-R_\varepsilon, R_\varepsilon]^d$, it holds that

$$\log \mathcal{N}(h, \bar{\mathcal{F}}_\varepsilon, \|\cdot\|_{L^\infty}) \leq \log \mathcal{N}(h, \mathcal{F}, L^\infty([-R_\varepsilon, R_\varepsilon]^d))$$

We also observe that $\bar{\mathcal{F}}_\varepsilon$ can be recovered by expanding the restricted class \mathcal{F}_ε , i.e.

$$\bar{\mathcal{F}}_\varepsilon = \{\varphi_1 \cdot \mathbb{I}(B(0, R_\varepsilon)) - \varphi_2 \cdot \mathbb{I}(B(0, R_\varepsilon))\}.$$

So by Lemma F.2, we get the covering number of $\bar{\mathcal{F}}_\varepsilon$:

$$\begin{aligned} \log \mathcal{N}(h, \bar{\mathcal{F}}_\varepsilon, \|\cdot\|_{L^\infty}) &\leq \log \mathcal{N}(h/3, \mathcal{F}, \|\cdot\|_{L^\infty}) + 6F \log_+(1/h) \\ &\lesssim D_{\mathcal{F}} \cdot h^{-\gamma} \cdot \log_+(1/h)^{\gamma'} \cdot R_\varepsilon^\eta + F \cdot \log_+(1/h) \end{aligned} \quad (48)$$

Lemma F.1 provides the expectation and high probability upper bounds for $\sup_{f \in \bar{\mathcal{F}}} \int_{B(0, R_\varepsilon)} f d(\mu - \mu_n)$: For any $\xi > 0$,

$$\mathbb{P}\left(\sup_{f \in \bar{\mathcal{F}}} \int f d(\mu_n - \mu) \geq \mathbb{E}\left[\sup_{f \in \bar{\mathcal{F}}} \int f d(\mu_n - \mu)\right] + \xi\right) \leq 2 \exp\left\{-\frac{n\xi^2}{2F^2}\right\} \quad (49)$$

and

$$\mathbb{E}\left[\sup_{f \in \bar{\mathcal{F}}} \int f d(\mu_n - \mu)\right] \leq 2 \inf_{0 < \delta < F} \left(2\delta + \frac{12}{\sqrt{n}} \int_\delta^F \sqrt{\log \mathcal{N}(h, \bar{\mathcal{F}}_\varepsilon, \|\cdot\|_{L^\infty})} dh\right). \quad (50)$$

A straightforward calculation then yields

$$\begin{aligned} \int_\delta^F h^{-\frac{\gamma}{2}} \cdot \log_+(1/h)^{\frac{\gamma'}{2}} dh &= \int_\delta^e h^{-\frac{\gamma}{2}} \cdot \log_+(1/h)^{\frac{\gamma'}{2}} dh \\ &\leq \int_\delta^{1/e} h^{-\frac{\gamma}{2}} \cdot \log_+(1/\delta)^{\frac{\gamma'}{2}} dh \leq \log_+(1/\delta)^{\frac{\gamma'}{2}} \frac{1}{1-\gamma/2} (e^{\frac{\gamma}{2}-1} - \delta^{1-\frac{\gamma}{2}}) \\ &\lesssim \log_+(1/\delta)^{\frac{\gamma'}{2}}, \end{aligned} \quad (51)$$

and

$$\int_\delta^{1/e} \log_+(1/h)^{\frac{1}{2}} dh \leq \log_+(1/\delta)^{\frac{1}{2}} (e^{-1} - \delta) \lesssim \log_+(1/\delta)^{\frac{1}{2}} \quad (52)$$

Since $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, Combining Equations (50), (51) and (52) yields that

$$\begin{aligned} \mathbb{E}\left[\sup_{f \in \bar{\mathcal{F}}} \int f d(\mu_n - \mu)\right] &\leq 2 \inf_{0 < \delta < F} \left(2\delta + \frac{12}{\sqrt{n}} \int_\delta^F \sqrt{\log \mathcal{N}(h, \bar{\mathcal{F}}_\varepsilon, \|\cdot\|_{L^\infty})} dh\right), \\ &\lesssim \inf_{0 < \delta < F} \left(\delta + n^{-\frac{1}{2}} D_{\mathcal{F}}^{\frac{1}{2}} R_\varepsilon^{\frac{\eta}{2}} \log_+(1/\delta)^{\frac{\gamma'}{2}} + n^{-\frac{1}{2}} F^{\frac{1}{2}} \cdot \log_+(1/\delta)^{\frac{1}{2}}\right) \\ &\leq n^{-\frac{1}{2}} D_{\mathcal{F}}^{\frac{1}{2}} R_\varepsilon^{\frac{\eta}{2}} \log_+(n)^{\frac{\gamma'}{2}} + n^{-\frac{1}{2}} F^{\frac{1}{2}} \cdot \log_+(n)^{\frac{1}{2}}, \end{aligned} \quad (53)$$

where we suppressed constants related to γ, γ' , and the last inequality holds by taking $\delta = n^{-\frac{1}{2}}$. Taking $\xi = n^{-\frac{1}{2}} F \cdot \log^{\frac{1}{2}}(\frac{2}{\delta_1})$, combining with Equation (49), it holds that with probability at least $1 - \delta_1$

$$\sup_{f \in \bar{\mathcal{F}}} \int_{B(0, R_\varepsilon)} f d(\mu - \mu_n) \leq n^{-\frac{1}{2}} D_{\mathcal{F}}^{\frac{1}{2}} R_\varepsilon^{\frac{\eta}{2}} \log_+(n)^{\frac{\gamma'}{2}} + n^{-\frac{1}{2}} F^{\frac{1}{2}} \cdot \log_+(n)^{\frac{1}{2}} + n^{-\frac{1}{2}} F \cdot \log^{\frac{1}{2}}\left(\frac{2}{\delta_1}\right) \quad (54)$$

Now we consider

$$\sup_{g \in \bar{\mathcal{G}}} \int_{\nabla \varphi_0 B(0, R_\varepsilon)} g d(\nu - \nu_m)$$

Recall $\bar{\mathcal{G}} := \{\varphi_{\bar{R}_\varepsilon}^* : \varphi \in \mathcal{F}\}$ with

$$\|\varphi_{\bar{R}_\varepsilon}^* - \varphi_{\bar{R}_\varepsilon}^*\|_{L^\infty} \leq \|\varphi_1 - \varphi_2\|_{L^\infty(B(0, \bar{R}_\varepsilon))} \leq \|\varphi_1 - \varphi_2\|_{L^\infty[-\bar{R}_\varepsilon, \bar{R}_\varepsilon]^d}$$

It implies that

$$\log \mathcal{N}(h, \bar{\mathcal{G}}, \|\cdot\|_{L^\infty}) \leq \log \mathcal{N}(h, \mathcal{F}, L^\infty([-\bar{R}_\varepsilon, \bar{R}_\varepsilon]^d))$$

We also note that $G(y) = \tilde{R}_\varepsilon \|y\|_2 + \sup_{x \in B(0, \tilde{R}_\varepsilon)} \Phi(x)$ is the envelope function for \mathcal{G} . We can first restrict it to \mathcal{G}_ε :

$$\mathcal{G}_\varepsilon := \{g \cdot \mathbb{I}(\nabla \varphi_0 B(0, R_\varepsilon)) \mid g \in \mathcal{G}\},$$

which is bounded in $\|\cdot\|_{L^\infty}$ norm, i.e.

$$\sup_{g \in \mathcal{G}} \left\{ \left\| g \cdot \mathbb{I}(\nabla \varphi_0 B(0, R_\varepsilon)) \right\|_{L^\infty} \right\} \leq \tilde{R}_\varepsilon \sup_{x \in B(0, R_\varepsilon)} \|\nabla \varphi_0(x)\| + \sup_{x \in B(0, \tilde{R}_\varepsilon)} \Phi(x) =: M,$$

and then expand \mathcal{G}_ε to $\bar{\mathcal{G}}_\varepsilon$. Similar to the tricks in step 1, Lemma F.2 implies the covering number for $\bar{\mathcal{G}}_\varepsilon$, i.e.

$$\begin{aligned} \log \mathcal{N}(h, \bar{\mathcal{G}}_\varepsilon, \|\cdot\|_{L^\infty}) &\leq \log \mathcal{N}(h/3, \mathcal{G}_\varepsilon, \|\cdot\|_{L^\infty}) + 6M \log_+(1/h) \\ &\lesssim D_{\mathcal{F}} \cdot h^{-\gamma} \cdot \log_+(1/h)^{\gamma'} \cdot \tilde{R}_\varepsilon^\eta + M \cdot \log_+(1/h) \end{aligned} \quad (55)$$

By applying Lemma F.1, one can repeat the procedures in for $\bar{\mathcal{F}}$ to have

$$\mathbb{P}\left(\sup_{g \in \bar{\mathcal{G}}_\varepsilon} \int g d(\nu_m - \nu) \geq \mathbb{E}\left[\sup_{g \in \bar{\mathcal{G}}_\varepsilon} \int g d(\nu_m - \nu)\right] + \varepsilon\right) \leq 2 \exp\left\{-\frac{m\varepsilon^2}{2M^2}\right\} \quad (56)$$

and

$$\mathbb{E}\left[\sup_{g \in \bar{\mathcal{G}}_\varepsilon} \int g d(\nu_m - \nu)\right] \leq m^{-\frac{1}{2}} D_{\mathcal{F}}^{\frac{1}{2}} \tilde{R}_\varepsilon^\eta \log_+(m)^{\frac{\gamma'}{2}} + m^{-\frac{1}{2}} M^{\frac{1}{2}} \cdot \log_+(m)^{\frac{1}{2}}, \quad (57)$$

Combining Equations (53) and (57), we have

$$\begin{aligned} \mathbb{E}[\mathcal{E}_{stat}] &\lesssim n^{-\frac{1}{2}} D_{\mathcal{F}}^{\frac{1}{2}} R_\varepsilon^{\frac{\eta}{2}} \log_+(n)^{\frac{\gamma'}{2}} + n^{-\frac{1}{2}} \sup_{x \in B(0, R_\varepsilon)} \Phi^{\frac{1}{2}}(x) \cdot \log_+(n)^{\frac{1}{2}} \\ &\quad + m^{-\frac{1}{2}} D_{\mathcal{F}}^{\frac{1}{2}} \tilde{R}_\varepsilon^{\frac{\eta}{2}} \log_+(m)^{\frac{\gamma'}{2}} + m^{-\frac{1}{2}} M^{\frac{1}{2}} \cdot \log_+(m)^{\frac{1}{2}}. \end{aligned}$$

And Equations (54) and (56) provide that with high probability $1 - \delta_1 - \delta_2$:

$$\begin{aligned} \mathcal{E}_{stat} &\lesssim n^{-\frac{1}{2}} D_{\mathcal{F}}^{\frac{1}{2}} R_\varepsilon^{\frac{\eta}{2}} \log_+(n)^{\frac{\gamma'}{2}} + n^{-\frac{1}{2}} F^{\frac{1}{2}} \cdot \log_+(n)^{\frac{1}{2}} + n^{-\frac{1}{2}} F \cdot \log_+^{\frac{1}{2}}\left(\frac{2}{\delta_1}\right) \\ &\quad + m^{-\frac{1}{2}} D_{\mathcal{F}}^{\frac{1}{2}} \tilde{R}_\varepsilon^{\frac{\eta}{2}} \log_+(m)^{\frac{\gamma'}{2}} + m^{-\frac{1}{2}} M^{\frac{1}{2}} \cdot \log_+(m)^{\frac{1}{2}} + m^{-\frac{1}{2}} M \cdot \log_+^{\frac{1}{2}}\left(\frac{2}{\delta_2}\right) \end{aligned}$$

Since $\gamma' \geq 1$ and $D_{\mathcal{F}} > 1$, re-organizing terms yields the stated bounds for \mathcal{E}_{stat} .

Step 2. Bounding the sieve error \mathcal{E}_{sieve} : We first consider the source part $2\mathbb{E}_{X \sim \mu_n}[\Phi(X) \cdot \mathbb{I}(\|X\|_2 > R_\varepsilon)]$, where $\mu_n := \frac{1}{n} \sum_{i=1}^n \mathbb{I}(x = X_i)$ is the (random) empirical distribution with $X_i \sim \mu$ i.i.d. We apply Cauchy inequality to get

$$\mathbb{E}\left[2\mathbb{E}_{X \sim \mu_n}[\Phi(X) \cdot \mathbb{I}(\|X\|_2 > R_\varepsilon)]\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[(2\Phi(X) \cdot \mathbb{I}(\|X\|_2 > R_\varepsilon))\right] \leq 2\sqrt{\mu(\Phi^2(X))} \varepsilon^{\frac{1}{2}} \quad (58)$$

By Chebyshev inequality and $\text{Var}(2\Phi(X) \mathbb{I}(\|X\|_2 > R_\varepsilon)) \leq E(4\Phi^2(X))$, we have

$$\mathbb{P}\left(\left|\mu_n(2\Phi(X) \cdot \mathbb{I}(\|X\|_2 > R_\varepsilon)) - \mathbb{E}[\mu_n(2\Phi(X) \cdot \mathbb{I}(\|X\|_2 > R_\varepsilon))]\right| > \xi\right) \leq \frac{4\mu(\Phi^2(X))}{n\xi^2}$$

Taking $\xi = n^{-\frac{1}{2}} 2\sqrt{\mu(\Phi^2(X))} \cdot \delta_3^{-\frac{1}{2}}$, it holds that with probability at least $1 - \delta_3$,

$$\mu_n(2\Phi(X) \cdot \mathbb{I}(\|X\|_2 > R_\varepsilon)) \leq 2\sqrt{\mu(\Phi^2(X))} \varepsilon^{\frac{1}{2}} + n^{-\frac{1}{2}} 2\sqrt{\mu(\Phi^2(X))} \cdot \delta_3^{-\frac{1}{2}} \quad (59)$$

For the target part, still letting $G(y) := \tilde{R}_\varepsilon \|y\|_2 + \sup_{x \in B(0, \tilde{R}_\varepsilon)} \Phi(x)$, then similarly,

$$\begin{aligned} \mathbb{E}\left[2\mathbb{E}_{Y \sim \nu_m}[G(Y) \cdot \mathbb{I}(\|(\nabla \varphi_0)^{-1}(Y)\|_2 > R_\varepsilon)]\right] &\leq 2\sqrt{\nu(G^2(Y))} \varepsilon^{\frac{1}{2}} \\ &\leq 2\sqrt{\tilde{R}_\varepsilon^2 \cdot \nu(Y^2) + \sup_{x \in B(0, \tilde{R}_\varepsilon)} \Phi(x)^2} \cdot \varepsilon^{\frac{1}{2}} \end{aligned} \quad (60)$$

And with probability at least $1 - \delta_4$

$$2\mathbb{E}_{Y \sim \nu_m} [G(Y) \cdot \mathbb{I}(\|(\nabla \varphi_0)^{-1}(Y)\|_2 > R_\varepsilon)] \leq 2\sqrt{\nu(G^2(Y))\varepsilon^{\frac{1}{2}}} + m^{-\frac{1}{2}} 2\sqrt{\nu(G^2(Y))} \cdot \delta_4^{-\frac{1}{2}} \quad (61)$$

Combine Equations (58) and (60), we have

$$\mathbb{E}[\mathcal{E}_{sieve}] \leq 2\sqrt{\mu(\Phi^2(X))\varepsilon^{\frac{1}{2}}} + 2\sqrt{\nu(G^2(Y))\varepsilon^{\frac{1}{2}}}$$

Meanwhile, by (59) and (61), with probability at least $1 - \delta_3 - \delta_4$,

$$\begin{aligned} \mathcal{E}_{sieve} &\leq 2\sqrt{\mu(\Phi^2(X))\varepsilon^{\frac{1}{2}}} + n^{-\frac{1}{2}} 2\sqrt{\mu(\Phi^2(X))} \cdot \delta_3^{-\frac{1}{2}} \\ &\quad + 2\sqrt{\nu(G^2(Y))\varepsilon^{\frac{1}{2}}} + m^{-\frac{1}{2}} 2\sqrt{\nu(G^2(Y))} \cdot \delta_4^{-\frac{1}{2}} \end{aligned}$$

□

E Proof of Theorem 3.7

In this appendix, we state a complete version of Theorem 3.7 and provide its proof.

Theorem E.1 (OT Estimation via Tanh Neural Network). *Let $n = m$, and suppose μ is (λ, K) -sub-Weibull, and the true Brenier potential $\varphi_0 \in C^\alpha(\mathbb{R}^d)$ with $\alpha \geq 2$ is (β, b) -smooth. By setting $R_\varepsilon, \tilde{R}_\varepsilon$ according to Equation (10), there exists a $(\beta, k + b + 1)$ -smooth deep TNN function class $\mathcal{F} := \mathcal{F}(L, W, \kappa)$, with*

$$3 \leq L = \mathcal{O}(1), \quad W = \mathcal{O}(N^d), \quad \kappa = \mathcal{O}(N^{\frac{d(d+(\lfloor \alpha \rfloor + 2)^2 + 4) + 2}{2}}),$$

where N is the network size parameter, such that the sieved estimator $\nabla \tilde{\varphi}_n$ with sieve radius

$$\tilde{R}_n = C_{K,d,\alpha} \cdot (\log n)^{\frac{1}{\lambda}}, \quad \text{for some constant } C_{K,d,\alpha} \geq 4K \frac{2\alpha}{d + 2\alpha} + 2 \quad (62)$$

satisfies

$$\mathbb{E} \|\nabla \tilde{\varphi}_n - \nabla \varphi_0\|_{L^2(\mu)}^2 \lesssim_{\log n} n^{-\frac{\alpha}{d+2\alpha}}. \quad (63)$$

Proof of Theorem 3.7. We first build the candidate TNN class \mathcal{F} .

Since μ is (λ, K) -sub-Weibull, for any $\varepsilon > 0$ we can choose $R_\varepsilon = K \left(\log \frac{2}{\varepsilon} \right)^{1/\lambda}$, so that $\mu(\|X\|_2 > R_\varepsilon) \leq \varepsilon$. By the assumption that φ_0 is (β, b) -smooth, it holds that

$$\|\nabla^2 \varphi_0(x)\|_{op} \leq \beta \langle x \rangle^b, \quad \|\nabla \varphi_0(x)\|_2 \leq \beta \langle x \rangle^{b+1} + \beta, \quad |\varphi_0(x)| \leq \beta \langle x \rangle^{b+2} + \beta$$

Fix any $\tilde{R}_\varepsilon \geq R_\varepsilon$ to be determined, it follows that $\|\varphi_0\|_{W^{2,\infty}([- \tilde{R}_\varepsilon, \tilde{R}_\varepsilon]^d)} \leq 3\beta \langle \tilde{R}_\varepsilon \rangle^{b+2}$. Then we apply lemma F.4 to have: there exists an two hidden layers Tanh Neural Network φ_N in \mathcal{F}_N with width $\mathcal{O}(N^d)$, and parameters bounded by $\mathcal{O}(N^{\frac{d(d+(\lfloor \alpha \rfloor + 2)^2 + 4) + 2}{2}})$, such that

$$\begin{aligned} \|\varphi_N - \varphi_0\|_{L^\infty([- \tilde{R}_\varepsilon, \tilde{R}_\varepsilon]^d)} &\lesssim \log(\tilde{R}_\varepsilon) \langle \tilde{R}_\varepsilon \rangle^{k+b+2} N^{-\alpha} \\ \|\nabla \varphi_N - \nabla \varphi_0\|_{L^\infty([- \tilde{R}_\varepsilon, \tilde{R}_\varepsilon]^d)} &\lesssim \log(\tilde{R}_\varepsilon) \langle \tilde{R}_\varepsilon \rangle^{k+b+1} \\ \|\nabla^2 \varphi_N\|_{op,([- \tilde{R}_\varepsilon, \tilde{R}_\varepsilon]^d)} &\lesssim \log(\tilde{R}_\varepsilon) \langle \tilde{R}_\varepsilon \rangle^{k+b}, \end{aligned} \quad (64)$$

Equivalently, if we choose a subclass $\mathcal{F} \subset \mathcal{F}_N$ consisting of all φ that is $(\beta, k + b + 1)$ -smooth:

$$\|\nabla^2 \varphi(x)\|_{op} \leq \beta \langle x \rangle^{k+b+1} := \beta \langle x \rangle^{b'}$$

then the network approximation φ_N produced by Lemma F.4 automatically lies in \mathcal{F} .

Given the growing rate of \mathcal{F} , now we determine \tilde{R}_ε .

Recall in Proposition 3.3, with $\|\nabla \varphi(x) - \nabla \varphi_0(x)\|_2 \leq u(x) := \beta \langle x \rangle^{b'+1}$, $U_2(R) := \beta \langle x \rangle^{b'}$, it holds that

$$\tilde{R}_\varepsilon \geq 2\langle R_\varepsilon \rangle \geq R_\varepsilon + \sup_{\|x\|_2 \leq R_\varepsilon} \frac{u(x)}{U_2(\|x\|_2 + u(x))}$$

So we set $\tilde{R}_\varepsilon := 2\langle R_\varepsilon \rangle$.

With Propositions 3.3 (stability) and 3.4 (oracle inequality) in hand, we are now ready to derive the desired convergence rate.

Denote the truncated excess risk as

$$\tilde{r}_n := \int_{B(0, R_\varepsilon)} (\tilde{\varphi}_n(x) + \tilde{\varphi}_{n, \tilde{R}_\varepsilon}^*(y) - \varphi_0(x) - \varphi_0^*(y)) \mu(dx),$$

Since $\tilde{\varphi}_n \in \mathcal{F}$, Proposition 3.3 gives that

$$\begin{aligned} \|\nabla \tilde{\varphi}_n - \nabla \varphi_0\|_{L^2(\mu)}^2 &\leq 2U_2(R_\varepsilon + U_1(R_\varepsilon)) \cdot r_\varepsilon(\varphi) + \|u\|_{L^4(\mu)}^2 \cdot \varepsilon^{\frac{1}{2}} \\ &\lesssim \langle R_\varepsilon \rangle^{b'(b'+1)} \cdot \tilde{r}_n + \varepsilon^{\frac{1}{2}}. \end{aligned} \quad (65)$$

And Proposition 3.4 further gives that

$$\tilde{r}_n \leq \mathcal{E}_{app} + \mathcal{E}_{sieve} + \mathcal{E}_{stat},$$

with

$$\begin{aligned} \mathcal{E}_{app} &:= 2 \inf_{\varphi \in \mathcal{F}} \|\varphi - \varphi_0\|_{L^\infty(B(0, \tilde{R}_\varepsilon))} \lesssim \log(\tilde{R}_\varepsilon) \langle \tilde{R}_\varepsilon \rangle^{k+b+2} N^{-\alpha} \quad (\text{By Equation (64)}) \\ \mathcal{E}_{sieve} &:= 2\mathbb{E}_{X \sim \mu_n} [\Phi(X) \cdot \mathbb{I}(\|X\|_2 > R_\varepsilon)] + 2\mathbb{E}_{Y \sim \nu_n} [G(Y) \cdot \mathbb{I}(\|(\nabla \varphi_0)^{-1}(Y)\|_2 > R_\varepsilon)] \\ \mathcal{E}_{stat} &:= \sup_{f \in \mathcal{F}} \int_{B(0, R_\varepsilon)} f d(\mu - \mu_n) + \sup_{g \in \tilde{\mathcal{G}}} \int_{B(0, R_\varepsilon)} g d(\nu_n - \nu), \end{aligned} \quad (66)$$

where $\Phi(x) := \beta \langle x \rangle^{b'+2}$, $G(y) := \tilde{R}_\varepsilon \|y\|_2 + \beta \langle \tilde{R}_\varepsilon \rangle^{b'+2}$. Thus, it remains to control the statistical error.

Given the architecture of Tanh NN, the covering number has an upper bound through Proposition 1 of [9], i.e.

$$\log \mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty) \lesssim N^d \log N \log_+(\frac{1}{\delta}) \quad (67)$$

Note that Equation (67) indicates that $\gamma = 0, \gamma' = 1, \eta = 0$ and $D_{\mathcal{F}} = N^d \log N$ as defined in Assumption 3.5. Plugging this bound in Theorem D.1, we have

$$\begin{aligned} \mathbb{E} \mathcal{E}_{stat} &\lesssim n^{-\frac{1}{2}} (N^{\frac{d}{2}} \log^{\frac{1}{2}} N + R_\varepsilon^{\frac{b'+2}{2}}) \cdot \log(n)^{\frac{\gamma' \vee 1}{2}} \\ \mathbb{E} \mathcal{E}_{sieve} &\lesssim R_\varepsilon^{b'+2} \varepsilon^{\frac{1}{2}} \end{aligned} \quad (68)$$

Consequently, it holds for \tilde{r}_n that

$$\begin{aligned} \tilde{r}_n &\lesssim \log(\tilde{R}_\varepsilon) \langle \tilde{R}_\varepsilon \rangle^{b'+1} N^{-\alpha} + n^{-\frac{1}{2}} (N^{\frac{d}{2}} \log^{\frac{1}{2}} N + R_\varepsilon^{\frac{b'+2}{2}}) \cdot \log(n)^{\frac{\gamma' \vee 1}{2}} + R_\varepsilon^{b'+2} \varepsilon^{\frac{1}{2}} \\ &\lesssim_{\log n} n^{-\frac{\alpha}{d+2\alpha}} \log(\tilde{R}_\varepsilon) R_\varepsilon^{b'+1} + R_\varepsilon^{b'+2} \varepsilon^{\frac{1}{2}} \end{aligned}$$

where the last equality holds by utilizing $\gamma' = 1$ and taking $N = O(n^{\frac{1}{d+2\alpha}})$. Combine with (65), we have

$$\mathbb{E} \|\nabla \varphi_0 - \nabla \tilde{\varphi}_n\|_{L^2(\mu)}^2 \lesssim_{\log n} \langle R_\varepsilon \rangle^{b'(b'+1)} \cdot \left(n^{-\frac{\alpha}{d+2\alpha}} \log(\tilde{R}_\varepsilon) R_\varepsilon^{b'+1} + R_\varepsilon^{b'+2} \varepsilon^{\frac{1}{2}} \right) + \varepsilon^{\frac{1}{2}} \quad (69)$$

$$\lesssim_{\log n} n^{-\frac{\alpha}{d+2\alpha}} \quad (70)$$

where the last equality holds by taking $\varepsilon = O(n^{-\frac{2\alpha}{d+2\alpha}})$. \square

F Auxillary Lemmas

F.1 Supporting Lemmas for Theorem 3.6

Lemma F.1. *For a symmetric function class \mathcal{F} with $\sup_{f \in \mathcal{F}} \|f\|_{L^\infty} \leq M$ for a constant M , define $d_{\mathcal{F}}(\mu, \nu) := \sup_{f \in \mathcal{F}} \int f d(\mu - \nu)$ and let $\hat{\mu}_n$ to be i.i.d empirical distribution, then we have*

$$\mathbb{P}\left(\left|d_{\mathcal{F}}(\hat{\mu}_n, \mu) - \mathbb{E}[d_{\mathcal{F}}(\hat{\mu}_n, \mu)]\right| \geq \varepsilon\right) \leq 2 \exp\left\{-\frac{n\varepsilon^2}{2M^2}\right\}$$

and

$$\mathbb{E}[d_{\mathcal{F}}(\mu_n, \mu)] \leq 2 \inf_{0 < \delta < M} \left(2\delta + \frac{12}{\sqrt{n}} \int_{\delta}^M \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_{\infty})} d\epsilon \right),$$

where $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_{L^\infty})$ denotes the ϵ -covering number of \mathcal{F} with respect to the $\|\cdot\|_{L^\infty}$ norm.

Proof of Lemma F.1. Let $X_1, X_2, \dots, X_n \sim \mu$ to be i.i.d samples,

$$d_{\mathcal{F}}(\hat{\mu}_n, \mu) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \frac{1}{n} \sum_{i=1}^n \mathbb{E}f(X_i) \right|$$

To apply McDiarmid's inequality on the function

$$G(x_1, x_2, \dots, x_n) := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \frac{1}{n} \sum_{i=1}^n \mathbb{E}f(X_i) \right|,$$

we need to verify the self-bounding property when varying each single coordinate. Denote $\mu_n^{\setminus i} f = \frac{1}{n} \sum_{j \neq i} \delta_{x_j} + \delta_{x'_i}$, then

$$\begin{aligned} & \sup_{x_1, \dots, x_n, x_i, x'_i} |G(x_1, \dots, x_i, \dots, x_n) - G(x_1, \dots, x'_i, \dots, x_n)| \\ &= \sup_{x_1, \dots, x_n, x_i, x'_i} \left| \sup_{f \in \mathcal{F}} |\mu_n f - \mu f| - \sup_{f \in \mathcal{F}} |\mu_n^{\setminus i} f - \mu f| \right| \\ &\leq \sup_{x_1, \dots, x_n, x_i, x'_i} \sup_{f \in \mathcal{F}} |\mu_n f - \mu f - (\mu_n^{\setminus i} f - \mu f)| \\ &= \sup_{x_1, \dots, x_n, x_i, x'_i} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} f(x_i) - \frac{1}{n} f(x'_i) \right| \\ &\leq \frac{1}{n} \sup_{f \in \mathcal{F}} \sup_{x, x'} |f(x) - f(x')| \leq \frac{1}{n} \sup_{f \in \mathcal{F}} (\sup_x |f(x)| + \sup_{x'} |f(x')|) \\ &= \frac{2}{n} \sup_{f \in \mathcal{F}} \|f\|_{\infty} \leq \frac{2}{n} M, \quad i = 1, \dots, n. \end{aligned}$$

Thus McDiarmid's inequality gives the first inequality:

$$\begin{aligned} & \mathbb{P} \left(\left| G(X_1, X_2, \dots, X_n) - \mathbb{E}[G(X_1, X_2, \dots, X_n)] \right| \geq \varepsilon \right) \\ &= \mathbb{P} \left(\left| d_{\mathcal{F}}(\hat{\mu}_n, \mu) - \mathbb{E}[d_{\mathcal{F}}(\hat{\mu}_n, \mu)] \right| \geq \varepsilon \right) \leq 2 \exp \left\{ -\frac{n\varepsilon^2}{2M^2} \right\} \end{aligned}$$

The second inequality is indeed Lemma 7 of [1]. The proof utilizes the symmetrization technique and Dudley's entropy integral, which can be found in empirical process theory [11]. We omit the proof here. □

Lemma F.2 (Covering Numbers for Expanded Class). *Suppose class \mathcal{F} is bounded in norm $\|\cdot\|$, i.e. $\sup_{f \in \mathcal{F}} \|f\| \leq F$. Let $\bar{\mathcal{F}} := \{(1-t)\varphi_1 + t\varphi_2 : \varphi_1, \varphi_2 \in \mathcal{F}, t \in [0, 1]\}$, we have*

$$\log \mathcal{N}(h, \bar{\mathcal{F}}, \|\cdot\|) \leq \log \mathcal{N}(h/3, \mathcal{F}, \|\cdot\|) + 6F \log_+(1/h)$$

Proof of Lemma F.2. For any $f_1 := (1-t)\varphi_1 + t\varphi_2, f_2 := (1-\tilde{t})\tilde{\varphi}_1 + \tilde{t}\tilde{\varphi}_2 \in \bar{\mathcal{F}}$, it holds that

$$f_1 - f_2 = (1-t)(\varphi_1 - \tilde{\varphi}_1) + t(\varphi_2 - \tilde{\varphi}_2) - (t - \tilde{t})(\tilde{\varphi}_1 - \tilde{\varphi}_2)$$

Then

$$\|f_1 - f_2\| \leq \|\varphi_1 - \tilde{\varphi}_1\| + \|\varphi_2 - \tilde{\varphi}_2\| + 2F|t - \tilde{t}|$$

Therefore, any $\frac{h}{3}$ -covering of \mathcal{F} and $\frac{h}{6F}$ -covering of $[0, 1]$ can form a h -covering of $\bar{\mathcal{F}}$.

$$\log \mathcal{N}(h, \bar{\mathcal{F}}, \|\cdot\|) \leq \log \mathcal{N}(h/3, \mathcal{F}, \|\cdot\|) + 6F \log_+(1/h) \quad \square$$

□

F.2 Properties of Tanh Neural Networks

Here we use the standard multi-index notation. A *multi-index* β is a d -tuple

$$\beta = (\beta_1, \dots, \beta_d) \in \mathbb{N}_0^d, \quad |\beta| := \sum_{i=1}^d \beta_i, \quad D^\beta := \frac{\partial^{\beta_1}}{\partial x_1^{\beta_1}} \cdots \frac{\partial^{\beta_d}}{\partial x_d^{\beta_d}}.$$

Lemma F.3 (Extension of Theorem 5.1 in [3]). *Let $d, s \in \mathbb{N}$, $B > 0$, and suppose $f \in W^{s,\infty}([-B, B]^d)$, then there exists a two layers tanh neural network f_N , with widths of order $O(N^d)$, parameters bounded by $O(B^{-\frac{s^2}{2}} \|f\|_{W^{s,\infty}([-B, B]^d)} N^{\frac{d(d+s^2+4)}{2}})$, such that*

$$\|f - f_N\|_{W^{k,\infty}([-B, B]^d)} \lesssim C_{d,s} \log(B \|f\|_{W^{s,\infty}([-B, B]^d)} N) \frac{B^{s-k} \|f\|_{W^{s,\infty}([-B, B]^d)}}{N^{s-k}},$$

for any $0 \leq k \leq s-1$, where N is the network size parameter.

Proof of Lemma F.3. We will construct a two-layer tanh network f_N on $[-B, B]^d$ by rescaling f to $[0, 1]^d$ and applying Theorem 5.1 of [3], then rescaling back.

Let $g(y) := f(2By - B\mathbf{1}_d)$, so $g(\cdot)$ has domain on $[0, 1]^d$. Here, we write $\mathbf{1}_d = (1, 1, \dots, 1)^\top \in \mathbb{R}^d$ to denote the d -dimensional column vector of all ones.

Then g is s -times differentiable on $[0, 1]^d$, and by the chain rule, for any multi-index β with $|\beta| \leq s$,

$$D^\beta g(y) = (2B)^{|\beta|} D^\beta f(2By - B\mathbf{1}_d).$$

Hence

$$\|g\|_{W^{s,\infty}([0,1]^d)} = \max_{|\alpha| \leq s} \sup_{y \in [0,1]^d} |D^\alpha g(y)| \leq (2B)^s \|f\|_{W^{s,\infty}([-B, B]^d)}.$$

Now we approximate g using TNN. By applying Theorem 5.1 of [3], we immediately have a two layers tanh neural network g_N , for any $0 \leq k \leq s-1$, such that

$$\|g - g_N\|_{W^{k,\infty}([0,1]^d)} \lesssim C_{d,s} \log(B \|f\|_{W^{s,\infty}([-B, B]^d)} N) \frac{B^s \|f\|_{W^{s,\infty}([-B, B]^d)}}{N^{s-k}},$$

where $C_{d,s}$ is constant only related to d, s . According to the same theorem, g_N has widths of order N^d , parameters bounded by $O(B^{-\frac{s^2}{2}} \|f\|_{W^{s,\infty}([-B, B]^d)} N^{\frac{d(d+s^2+4)}{2}})$.

In the last step, we rescaling g_N back to $[-B, B]^d$. Let $f_N(x) := g_N(\frac{x}{2B} + \frac{1}{2}\mathbf{1}_d)$, so that

$$\|f - f_N\|_{W^{k,\infty}([-B, B]^d)} \lesssim C_{d,s} \log(B \|f\|_{W^{s,\infty}([-B, B]^d)} N) \frac{B^{s-k} \|f\|_{W^{s,\infty}([-B, B]^d)}}{N^{s-k}},$$

If we write $g_N(x) = W_2 \cdot \sigma(W_1 x + b_1) + b_2$, then f_N can be written as

$$f_N(x) = W_2 \cdot \sigma\left(\frac{1}{2B} W_1 + \frac{1}{2} W_1 \mathbf{1}_d + b_1\right) + b_2$$

One only need to modify weights and biases in the first-layer of g_N , i.e. $W'_1 := \frac{1}{2B} W_1$ and $b'_1 := \frac{1}{2} W_1 \mathbf{1}_d + b_1$. The overall parameters if f_N are still bounded by $O(B^{-\frac{s^2}{2}} \|f\|_{W^{s,\infty}([-B, B]^d)} N^{\frac{d(d+s^2+4)}{2}})$. \square

Lemma F.4 (Approximation with Controlled Hessian). *Suppose $\varphi_0 \in C^\alpha(\mathbb{R}^d)$, with $\alpha \geq 2$. Then for any $B > 0$ and $R := \sqrt{d}B$, there exists an two layers Tanh Neural Network φ_N with width $O(N^d)$, and parameters bounded by $O(c_4(R) N^{\frac{d(d+(\lfloor \alpha \rfloor + 2)^2 + 4)}{2}})$, such that*

$$\begin{aligned} \|\varphi_0 - \varphi_N\|_{L^\infty([-B, B]^d)} &\lesssim c_7(R) N^{-\alpha} \\ \|\nabla^2 \varphi_N\|_{op,([-B, B]^d)} &\lesssim c_8(R), \end{aligned}$$

where $c_i(R)$ are constants depend only on R, φ_0 , smoothness parameters α, k , and dimension d .

Proof. We first construct a kernel smoothed $\varphi_h(\cdot)$ to approximate $\varphi_0(\cdot)$ on $B(0, R)$, where $R := \sqrt{d}B$, so that $[-B, B]^d \subset B(0, R)$.

Let $k = \lfloor \alpha \rfloor \geq 1$, and $K(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$ be a C^∞ kernel of order k with compact support, i.e.

$$\int_{\mathbb{R}^d} x^\beta K(x) dx = 0, \quad 0 \leq |\beta| \leq k, \quad \int_{\mathbb{R}^d} \|x\|_2^{k+1} K(x) dx < \infty,$$

where $\beta = (\beta_1, \dots, \beta_d)$, $|\beta| := \beta_1 + \dots + \beta_d$, and $x^\beta := x_1^{\beta_1} \dots x_d^{\beta_d}$.

For bandwidth $0 < h < 1$, we define $K_h(x) := h^{-d}K(x/h)$ and the smoothed potential

$$\varphi_h(x) = \int_{\mathbb{R}^d} K_h(y) \varphi_0(x - y) dy = \int_{\mathbb{R}^d} \varphi_0(x - ht) K(t) dt.$$

Denote $\beta! = (\beta_1! \dots \beta_d!)$, $(-ht)^\beta = \prod_{i=1}^d (-ht_i)^{\beta_i}$ and restrict $x \in B(0, R)$. Since K has vanishing moments up to order k , a Taylor expansion of $\varphi_0(x - ht)$ around x to order $k - 1$ gives

$$\begin{aligned} \varphi_h(x) - \varphi_0(x) &= \int_{\mathbb{R}^d} [\varphi_0(x - y) - \varphi_0(x)] K_h(y) dy \\ &= \int_{\mathbb{R}^d} \left[\sum_{|\beta| \leq k-1} \frac{(-ht)^\beta}{\beta!} D^\beta \varphi_0(x) + \sum_{|\beta|=k} \frac{(-ht)^\beta}{\beta!} D^\beta \varphi_0(x - \eta ht) \right] K(t) dt \\ &= \sum_{|\beta|=k} \frac{(-h)^\beta}{\beta!} \int_{\mathbb{R}^d} t^\beta D^\beta \varphi_0(x - \eta ht) K(t) dt \\ &= \sum_{|\beta|=k} \frac{(-h)^\beta}{\beta!} \int_{\mathbb{R}^d} t^\beta \left(D^\beta \varphi_0(x - \eta ht) - D^\beta \varphi_0(x) \right) K(t) dt, \end{aligned}$$

where $\eta \in (0, 1)$, and the last implication is consequence of the fact that $K(\cdot)$ is a kernel of order k .

Let $L(R)$ be the $(\alpha - k)$ -Hölder continuity constant of each $D^\beta \varphi_0$ on $B(0, R)$, i.e. for any $x, y \in B(0, R)$

$$\sup_{|\beta|=\lfloor \alpha \rfloor} \left| D^\beta \varphi_0(x) - D^\beta \varphi_0(y) \right| \leq L(R) \|x - y\|_2^{\alpha - \lfloor \alpha \rfloor}$$

It follows that

$$\left| D^\beta \varphi_0(x - \eta ht) - D^\beta \varphi_0(x) \right| \leq L(R + \|t\|_2) \|\eta ht\|_2^{\alpha - k} \leq L(R + \|t\|_2) h^{\alpha - k} \|t\|_2^{\alpha - k}.$$

Then we obtain the bound

$$\begin{aligned} \left| \varphi_h(x) - \varphi_0(x) \right| &\leq \frac{h^k}{k!} \binom{k+d-1}{k} \int_{\mathbb{R}^d} \|t\|_2^k L(R + \|t\|_2) \|th\|_2^{\alpha - k} K(t) dt \\ &= \frac{h^\alpha}{k!} \binom{k+d-1}{k} \int_{\mathbb{R}^d} L(R + \|t\|_2) \|t\|_2^\alpha K(t) dt \\ &\lesssim W_\alpha(R) \cdot h^\alpha, \end{aligned}$$

where

$$W_\alpha(R) := \int_{\mathbb{R}^d} L(R + \|t\|_2) \|t\|_2^\alpha K(t) dt < \infty$$

The finite is guaranteed by the compact support of $K(\cdot)$.

Now we control the Hessian of φ_h . Since convolution commutes with differentiation, it holds that

$$\begin{aligned}
\|\nabla^2 \varphi_h(x)\|_{op} &= \|\nabla^2(\varphi_0 * K_h)(x)\|_{op} = \left\| \int_{\mathbb{R}^d} K_h(y) \nabla^2 \varphi_0(x-y) dy \right\|_{op} \\
&\leq \int_{\mathbb{R}^d} K_h(y) \|\nabla^2 \varphi_0(x-y)\|_{op} dy \\
&= \int_{\mathbb{R}^d} \frac{1}{h^d} K\left(\frac{y}{h}\right) \|\nabla^2 \varphi_0(x-y)\|_{op} dy \\
&= \int_{\mathbb{R}^d} K(t) \|\nabla^2 \varphi_0(x-h t)\|_{op} dt \\
&\leq \int_{\mathbb{R}^d} \sup_{\|u\|_2 \leq R+\|t\|_2} \|\nabla^2 \varphi_0(u)\|_{op} K(t) dt := c_1(R),
\end{aligned}$$

Here and in what follows, $c_i(R)$ ($i = 1, 2, \dots$) denote positive constants that may change from line to line and depend only on R , φ_0 , smoothness parameters α, k , and dimension d .

Meanwhile, it holds that

$$\|\nabla \varphi_h(0)\|_2 \leq \int_{\mathbb{R}^d} K(t) \|\nabla \varphi_0(-h t)\|_2 dt \leq \int_{\mathbb{R}^d} K(t) \sup_{\|u\|_2 \leq R+\|t\|_2} \|\nabla \varphi_0(u)\|_2 dt$$

Then

$$\|\nabla \varphi_h(x)\|_2 \leq c_1(R)R + \|\nabla \varphi_h(0)\|_2 =: c_2(R)$$

Now it remains construct an TNN φ_N to approximate $\varphi_h \in C^s \subset C^\infty$, for any $s \in \mathbb{N}$. The approximation property of φ_h gives that

$$\begin{aligned}
\|\varphi_h\|_{W^{2,\infty}([-B,B]^d)} &\leq \|\varphi_h\|_{W^{2,\infty}(B(0,R))} \\
&\leq \|\varphi_0\|_{L^\infty(B(0,R))} + W_\alpha h^\alpha + c_2(R) + c_1(R) \lesssim c_3(R)
\end{aligned}$$

By lemma F.3, there exist a two layers tanh neural network φ_N with width $O(N^d)$, and parameters bounded by $O(c_4(R)N^{\frac{d(d+s^2+4)}{2}})$, such that

$$\|\varphi_h - \varphi_N\|_{\infty,([-B,B]^d)} \lesssim c_5(R) \frac{1}{N^s}$$

and

$$\|\nabla^2 \varphi_h - \nabla^2 \varphi_N\|_{op,([-B,B]^d)} \lesssim c_6(R) \frac{1}{N^{s-2}},$$

where we only consider $s \geq 3$. Since φ_h approximates φ_0 at $O(h^\alpha)$, we then set $N^{-s} = h^{\lfloor \alpha \rfloor + 2}$, $h^{-1} = N$, $s = \lfloor \alpha \rfloor + 2$, and conclude that

$$\begin{aligned}
\|\varphi_0 - \varphi_N\|_{\infty,([-B,B]^d)} &\lesssim (c_5(R) + W_\alpha(R)) N^{-\alpha} := c_7(R) \cdot N^{-\alpha} \\
\|\nabla^2 \varphi_N\|_{op,([-B,B]^d)} &\lesssim c_3(R) + c_6(R) := c_8(R)
\end{aligned}$$

□

Remark F.5 (Comment to Lemma F.4). For any $f \in W^{s,\infty}([-B,B]^d)$ with $s \in \mathbb{N}_0$, Theorem 5.1 of [3] guarantees a Tanh neural network approximation f_N of width $O(N^d)$ such that

$$\|f - f_N\|_{W^{k,\infty}([-B,B]^d)} = O(N^{-(s-k)}), \quad 0 \leq k \leq s-1.$$

In particular, when $f \in C^\alpha([-B,B]^d)$ with $2 < \alpha < 3$, this only yields

$$\|f - f_N\|_{W^{1,\infty}([-B,B]^d)} = O(N^{-1}),$$

and provides no control on the second derivatives $\|\nabla^2 f_N\|_\infty$. By contrast, Lemma F.4 slightly extend their results, by showing that for non-integer $\alpha > 2$, the same network architecture satisfies

$$\|f - f_N\|_{L^\infty([-B,B]^d)} = O(N^{-\alpha}), \quad \text{and} \quad \|f - f_N\|_{W^{2,\infty}([-B,B]^d)} \leq O(1),$$

i.e. all second derivatives remain uniformly bounded as $N \rightarrow \infty$, even $2 \leq \alpha < 3$.

Remark F.6. We list $c_i(R)$'s here:

$$\begin{aligned}
W_\alpha(R) &:= \int_{\mathbb{R}^d} L(R + \|t\|_2) \|t\|_2^\alpha K(t) dt \\
c_1(R) &:= \int_{\mathbb{R}^d} \sup_{\|u\|_2 \leq R + \|t\|_2} \|\nabla^2 \varphi_0(u)\|_{op} K(t) dt \\
c_2(R) &:= c_1(R)R + \int_{\mathbb{R}^d} \sup_{\|u\|_2 \leq R + \|t\|_2} \|\nabla \varphi_0(u)\|_2 K(t) dt \\
c_3(R) &:= \|\varphi_0\|_{L^\infty(B(0,R))} + c_2(R) + c_1(R) \\
c_4(R) &:= R^{-\frac{k^2}{2}} \left(\|\varphi_0\|_{W^{k,\infty}([-B,B]^d)} + \int_{\mathbb{R}^d} \sup_{\|u\|_2 \leq R + \|t\|_2} |\varphi_0(u)| \cdot K(t) dt \right) \\
c_5(R) &:= \log(R \|f\|_{W^{k,\infty}([-B,B]^d)}) R^k \|f\|_{W^{s,\infty}([-B,B]^d)} \\
c_6(R) &:= \log(R \|f\|_{W^{k,\infty}([-B,B]^d)}) R^{k-2} \|f\|_{W^{s,\infty}([-B,B]^d)} \\
c_7(R) &:= c_5(R) + W_\alpha(R) \\
c_8(R) &:= c_3(R) + c_6(R)
\end{aligned}$$

G Additional Numerical Experiments

In this section, we present numerical simulations to evaluate the performance of our sieved-TNN estimator and compare it against the original dual-type estimators from [7, 25, 21]. We consider four synthetic OT problems that pose challenges for existing theoretical frameworks:

- *Non- (β, b) -smooth maps:* $\mathcal{N}(0, 1) \rightarrow t_6$, $\text{Uniform}(0, 1) \rightarrow \mathcal{N}(0, 1)$.
- *Non- (α, a) -convex maps:* $t_6 \rightarrow \mathcal{N}(0, 1)$, $\mathcal{N}(0, 1) \rightarrow \text{Uniform}(0, 1)$.

We evaluate performance in dimensions $d = 5, 10, 20$. For each fixed setting (d, μ, ν) , we draw $n = m \in \{64, 128, 256, 512, 1024, 2048\}$ i.i.d. samples $X_i \sim \mu$ and $Y_j \sim \nu$, and repeat each experiment over 50 independent trials. For reproducibility, in the r th trial ($r = 1, \dots, 50$), we set the random seed as $\text{seed} = 12345 + r$. This ensures that each of the 50 repetitions uses a distinct, but deterministic, initialization.

All experiments were run on a computational cluster powered by Intel Xeon Gold 6342 processors. Specifically, each experimental setting (i.e. a fixed triple (d, μ, ν)) was executed on a single 20-core node, allocating 2 GB of RAM per core. In total, our numerical study consumed approximately 4,000 CPU-core hours.

Data Generation and true OT map $\nabla \varphi_0$ To generate multivariate data with dimension d , we set source and target measures μ and ν on \mathbb{R}^d whose coordinates are i.i.d. according to one of three 1-D marginals: $\mathcal{N}(0, 1)$, $\text{Uniform}(0, 1)$, or t_6 . Because both μ and ν have identical, independent marginals, the true OT map $\nabla \varphi_0$ is a composition of simple 1-D quantile transform applied to each coordinate. Concretely, if F and G denote the cumulative distribution functions of the source and target marginals respectively, then the true OT map is

$$\nabla \varphi_0(x) = (G^{-1}(F(x_1)), G^{-1}(F(x_2)), \dots, G^{-1}(F(x_d)))^\top.$$

That is, each component is pushed forward by the 1-D OT map $x \mapsto G^{-1}(F(x))$.

Sieve radius As discussed in Equation (18), we choose the sieve radius \tilde{R} as

$$\tilde{R} = \max_{1 \leq i \leq n} \|X_i\|_2 + C \cdot \max_{1 \leq j \leq n} \|Y_j\|_2, \quad C \in \{0, 1, 2, 3, \infty\}.$$

Here, $C = 0$ corresponds to the setting in [4], while $C = \infty$ recovers the original dual-type estimator studied in [7, 25, 21], enabling direct comparison.

Implementation setup We implement our dual-type sieved TNN estimator $\tilde{\varphi}_{n,m}$ in PyTorch [10] with objective function in Equation (8). The candidate class \mathcal{F}_Θ consists of TNNs with two hidden layers ($L = 3$) of width 10 (for $d = 5$), 20 (for $d = 10$), and 30 (for $d = 20$).

Our implementation follows the algorithms in [4]. In the outer loop (Algorithm 1), we estimate the Brenier potential by solving the sieved empirical semi-dual (Equation 8):

$$\min_{\varphi \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varphi(X_i) + \frac{1}{m} \sum_{j=1}^m \varphi_R^*(Y_j)$$

In the inner loop (Algorithm 2), we compute (sieved) convex conjugates φ_R^* with sieve radius \tilde{R} . For completeness, we restate both algorithms with TNNs below.

Training proceeds in two phases: a warm-up phase of $\max\{\lfloor 10000/n \rfloor, 50\}$ epochs at learning rate 5×10^{-3} , followed by 300 epochs at 10^{-3} . In both phases we use mini-batches of size $n' = m' = 64$ drawn independently from the source and target samples. Within each outer-loop update (Algorithm 1), the inner subproblem for the sieved convex conjugates (Algorithm 2) is solved using 300 iterations.

In each experiment, 10% of the samples are held out for validation. We report the Maximum Mean Discrepancy (MMD) between the transported source samples and target samples, using the Gaussian RBF kernel $k(x, y) = \exp\left\{-\frac{1}{d} \|x - y\|_2^2\right\}$.

Algorithm 1 Dual-type OT Map Estimators $\nabla \tilde{\varphi}_{n,m}$ with TNN

Require: Samples $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$; TNN model \mathcal{F}_Θ ; number of epochs T ; batch sizes n', m' .

- 1: Initialize $\varphi_\theta \in \mathcal{F}_\Theta$
- 2: **for** $t = 1, \dots, T$ **do**
- 3: **for** mini-batch $(X_{i_k})_{k=1}^{n'}$ in $(X_i)_{i=1}^n$, and $(Y_{j_k})_{k=1}^{m'}$ in $(Y_j)_{j=1}^m$ **do**
- 4: Compute $(\varphi_\theta^*(Y_{j_k}))_{k=1}^{m'}$ with Algorithm 2
- 5: Compute loss: $L \leftarrow \frac{1}{n'} \sum_{k=1}^{n'} \varphi_\theta(X_{i_k}) + \frac{1}{m'} \sum_{k=1}^{m'} \varphi_\theta^*(Y_{j_k})$
- 6: Update $\theta \in \Theta$ by minimizing L with Adam
- 7: **end for**
- 8: **end for**
- 9: **Return** $\nabla \varphi_\theta$

Algorithm 2 Computing the original or sieve convex conjugates

Require: Function φ ; value $y \in \mathbb{R}^d$; number of iterations T ; projection radius M_n .

- 1: Initialize $x \leftarrow 0$
- 2: **def** closure(φ):
- 3: Compute loss: $l \leftarrow \varphi(x) - \langle x, y \rangle$
- 4: **return** l
- 5: **for** $t = 1, \dots, T$ **do**
- 6: Update x with $x \leftarrow \text{GradientDescent}(\text{closure}, x)$
- 7: Project x onto $B(0, \tilde{R})$
- 8: **end for**
- 9: Calculate convex conjugate: $\varphi^*(y) \leftarrow \langle x, y \rangle - \varphi(x)$
- 10: **Return** $\varphi^*(y)$

Evaluation We assess the estimation error using unexplained variance proportion (UVP) [8]: $L^2\text{-UVP}(\nabla \hat{\varphi}_n) := \|\nabla \hat{\varphi}_n - \nabla \varphi_0\|_{L^2(\mu)}^2 / \text{Var}_\nu(\|Y\|_2)$. Lower values indicate better performance. For each experiment, we approximate the L^2 -UVP using an independent set of 10^6 samples.

We report the L^2 -UVP error versus sample size n for $d = 5, 10, 20$ in Figure 2 - 4 here.

From here are revised version

Superior Performance of the Sieved-TNN Estimator Across all experimental configurations, our sieved TNN estimator exhibits reliable convergence as the sample size n grows, thereby empirically validating the non-asymptotic convergence derived in Section 3.4.

Moreover, the sieved estimator consistently outperforms the classical dual-type estimator (i.e. with sieve constant $C = \infty$), especially in the small-sample regime. For example, in Figure 4a, the sieve-based estimators achieve lower L^2 -UVP errors for $n \leq 512$ and smaller variance, whereas the classical estimator only matches their performance once n becomes large $n \geq 1024$. This phenomenon is most evident in the Gaussian $\rightarrow t_6$ and Uniform \rightarrow Gaussian transports, across all dimensions considered.

We also observe that sieved estimators with small sieve constants (e.g. $C = 0$ or 1) not only attain lower training loss but also stabilize earlier on the held-out validation set, compared to those with larger C . This indicates that overly large sieve radii can complicate the optimization landscape, while a modest radius effectively mitigates these issues without introducing too much sieved bias.

These observations corroborate our theoretical findings: large sieve radii may degrade convergence. These results also offer practical guidance: a modest sieve constant (e.g. $C = 0$ or 1) is sufficient to yield better and robust performance across a wide range of settings, without incurring too much sieved bias.

Comparable Training Time We also compare the computational speed between sieved estimator and the original estimator, since they have different optimization problem of projection. The mean training time between the sieved estimator ($C = 2$) and the original estimator ($C = \infty$) for $d = 10$ from normal to uniform is summarized in Table 2:

Table 2: Mean training time (in seconds) for the sieved estimator and the original estimator.

C	64	128	256	512	1024	2048
2	97.8096	159.7494	293.5467	587.9228	1107.5363	2159.0094
∞	85.7766	140.1795	257.9078	513.5465	968.8996	1890.1076

Table 2 shows that the sieved estimator incurs an additional computational cost of approximately 15% due to the projection step. However, we believe this increase in computational time is moderate and acceptable, given the practical benefits of the sieved estimator.

Parameters in TNNs are Well Bounded In all our experiment settings, no explicit truncation of the TNN parameters is required in practice, although a theoretical bound $\kappa = \mathcal{O}(n^{\frac{d(d+(\lfloor \alpha \rfloor + 2)^2 + 4) + 2}{2(d+2\alpha)}})$ is established in Theorem 3.7. This can be understood theoretically and verified experimentally.

Across all settings, the underlying Brenier potentials φ_0 are infinitely smooth. Taking the limit $\alpha \rightarrow \infty$ leads to $\kappa \rightarrow \infty$, implying that truncation of TNN parameters becomes unnecessary.

Furthermore, Table 3 reports the mean of maximum TNN parameter over 50 repetitions for $d = 10$. The result suggests that parameter values remain small and increase slowly with sample size.

Table 3: Mean of Max Parameter over 50 repetition ($d = 10$)

C	Normal \rightarrow t						Normal \rightarrow Uniform					
	64	128	256	512	1024	2048	64	128	256	512	1024	2048
0	1.188	1.317	1.737	2.387	3.103	4.021	0.917	0.873	1.029	1.271	1.473	1.719
1	1.307	1.551	2.039	2.814	3.617	4.529	0.890	0.840	0.987	1.194	1.352	1.581
2	1.330	1.583	2.062	2.859	3.713	4.550	0.876	0.829	0.968	1.146	1.316	1.521
3	1.340	1.594	2.087	2.860	3.726	4.654	0.891	0.826	0.965	1.123	1.287	1.497
∞	1.357	1.620	2.108	2.849	3.682	4.563	0.893	0.815	0.955	1.102	1.280	1.439
C	t \rightarrow Normal						Uniform \rightarrow Normal					
	64	128	256	512	1024	2048	64	128	256	512	1024	2048
0	1.195	1.368	1.732	2.216	2.751	3.551	1.140	1.037	1.134	1.272	1.389	1.620
1	1.207	1.413	1.776	2.309	2.888	3.968	1.110	1.070	1.164	1.292	1.436	1.626
2	1.224	1.434	1.785	2.362	2.970	3.969	1.116	1.125	1.220	1.337	1.468	1.627
3	1.236	1.456	1.782	2.373	2.983	3.991	1.113	1.155	1.267	1.366	1.483	1.641
∞	1.262	1.467	1.757	2.354	2.966	3.977	1.135	1.236	1.362	1.428	1.533	1.669

To further investigate the effect of truncation, we conducted supplementary experiments in which all network parameters were clamped within the range ± 10 for $d = 10$ and $C = 2, \text{inf}$. The corresponding $L^2\text{-UVP}$ results are summarized in Table 4. Interestingly, we observed that truncation has a negligible impact on the performance of TNN.

Table 4: $L^2\text{-UVP}$, Truncation vs No Truncation ($d = 10$)

C	Normal \rightarrow t						Normal \rightarrow Uniform					
	64	128	256	512	1024	2048	64	128	256	512	1024	2048
2 (No)	0.4074	0.2591	0.1419	0.0916	0.0648	0.0593	0.9058	0.2717	0.1371	0.0720	0.0369	0.0267
2 (10)	0.4280	0.2589	0.1449	0.0929	0.0654	0.0595	0.9045	0.2775	0.1354	0.0703	0.0370	0.0268
∞ (No)	0.5823	0.4262	0.2340	0.0997	0.0646	0.0588	0.8815	0.2745	0.1369	0.0703	0.0364	0.0266
∞ (10)	0.5808	0.4413	0.2450	0.1016	0.0656	0.0586	0.8862	0.2768	0.1357	0.0697	0.0364	0.0266

C	t \rightarrow Normal						Uniform \rightarrow Normal					
	64	128	256	512	1024	2048	64	128	256	512	1024	2048
2 (No)	0.5526	0.3390	0.2002	0.1094	0.0689	0.0531	0.2765	0.1556	0.1055	0.0821	0.0674	0.0616
2 (10)	0.5701	0.3357	0.2029	0.1104	0.0692	0.0535	0.2779	0.1600	0.1077	0.0821	0.0680	0.0617
∞ (No)	0.5679	0.3678	0.2017	0.1064	0.0664	0.0517	0.2758	0.1811	0.1019	0.0788	0.0651	0.0609
∞ (10)	0.5686	0.3699	0.2065	0.1075	0.0670	0.0526	0.2999	0.1863	0.1021	0.0787	0.0652	0.0609

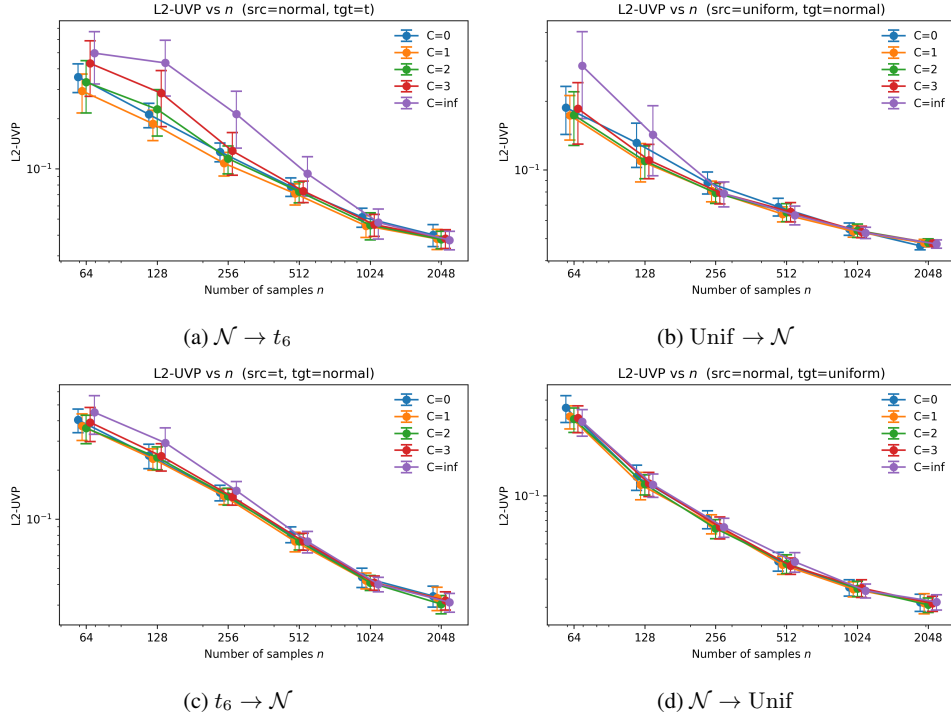


Figure 2: $L^2\text{-UVP}$ when $d = 5$. Each curve shows the mean $L^2\text{-UVP}$ over 50 random trials with one standard deviation.

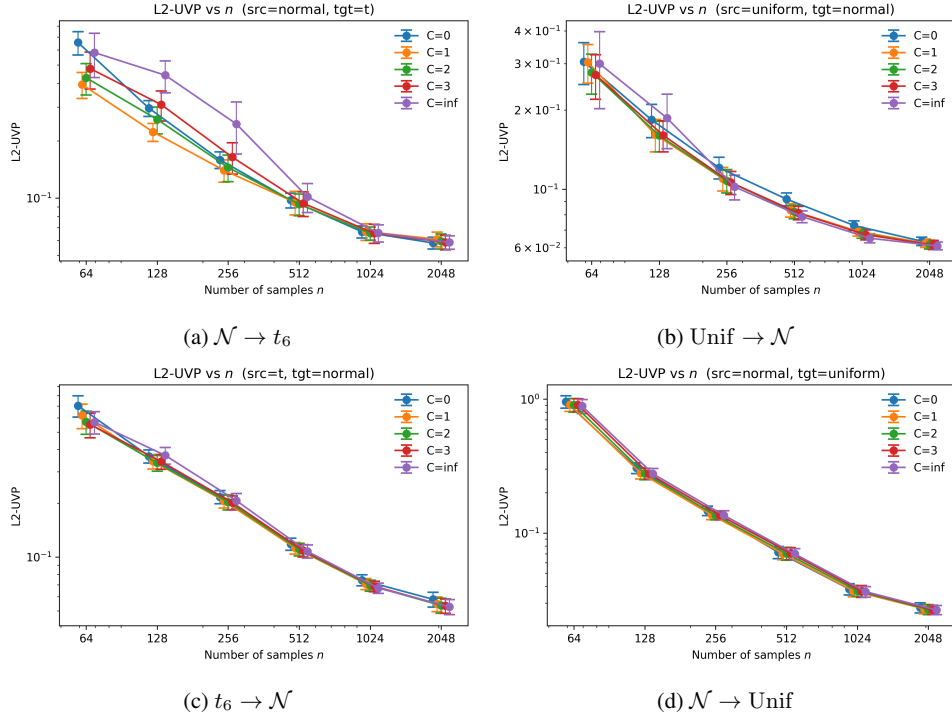


Figure 3: L^2 -UVP when $d = 10$. Each curve shows the mean L^2 -UVP over 50 random trials with one standard deviation.

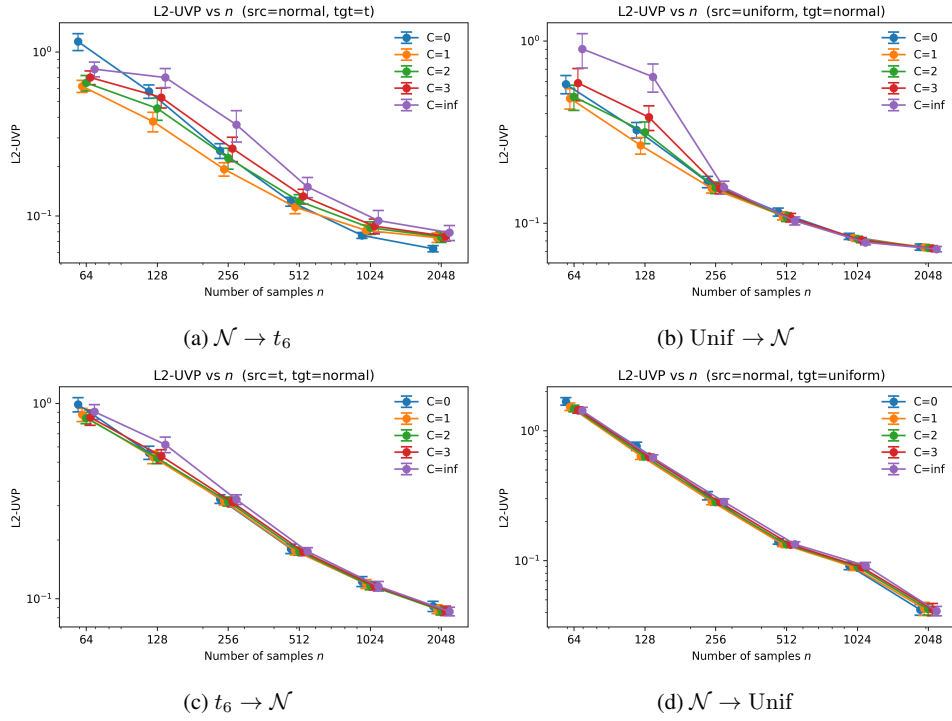


Figure 4: L^2 -UVP when $d = 20$. Each curve shows the mean L^2 -UVP over 50 random trials with one standard deviation.

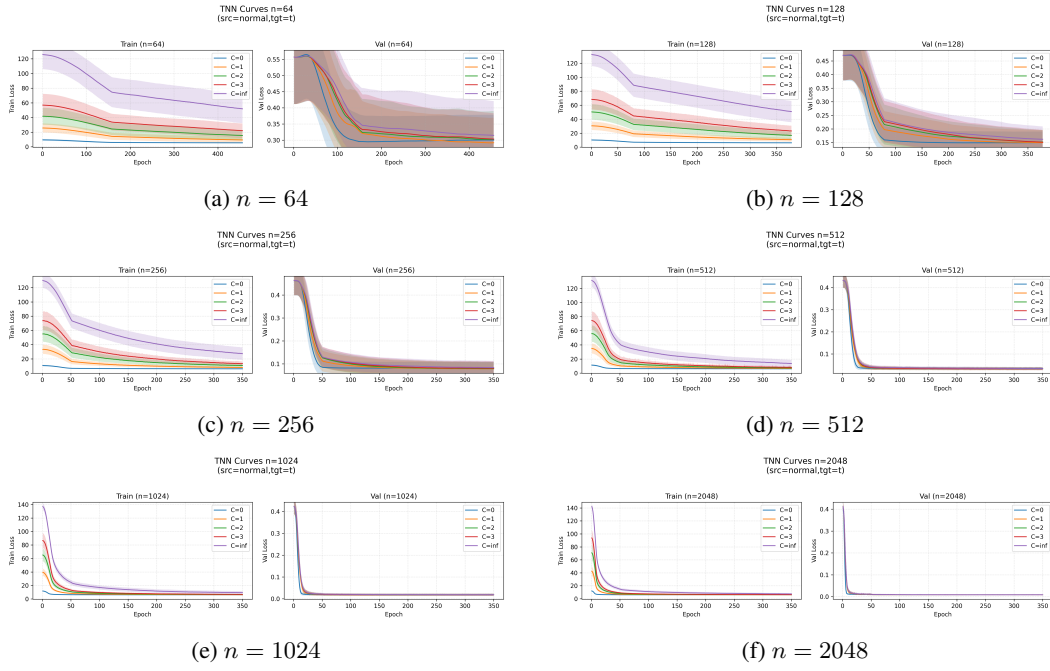


Figure 5: TNN training and validation loss curves for $d = 5$, source = Normal, target = Student- t , across different sample sizes n .

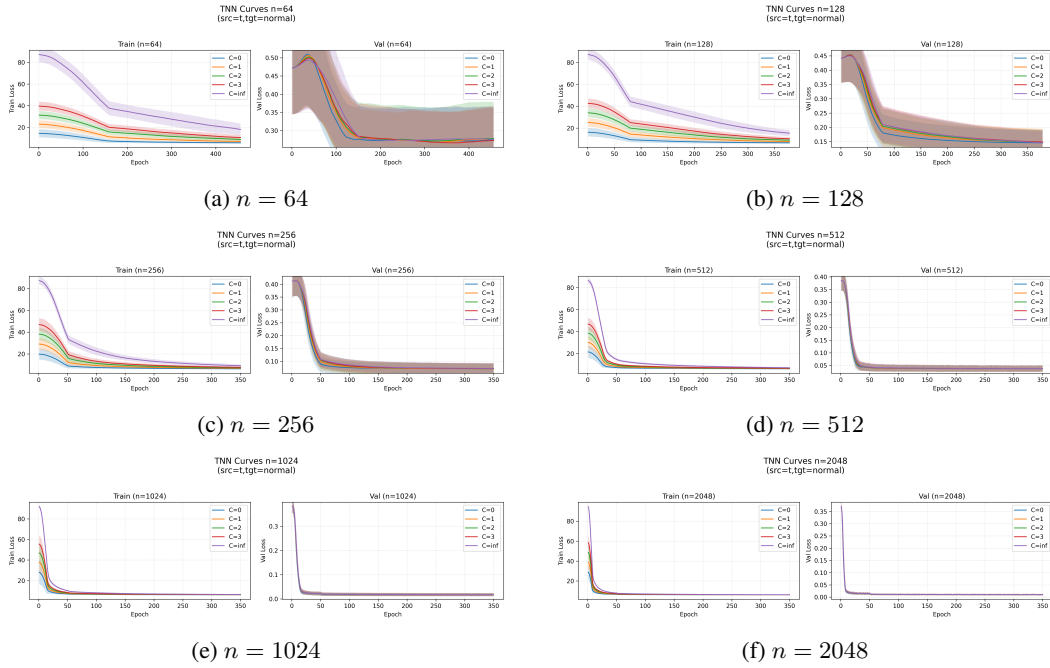


Figure 6: TNN training and validation loss curves for $d = 5$, source = Student- t , target = Normal, across different sample sizes n .

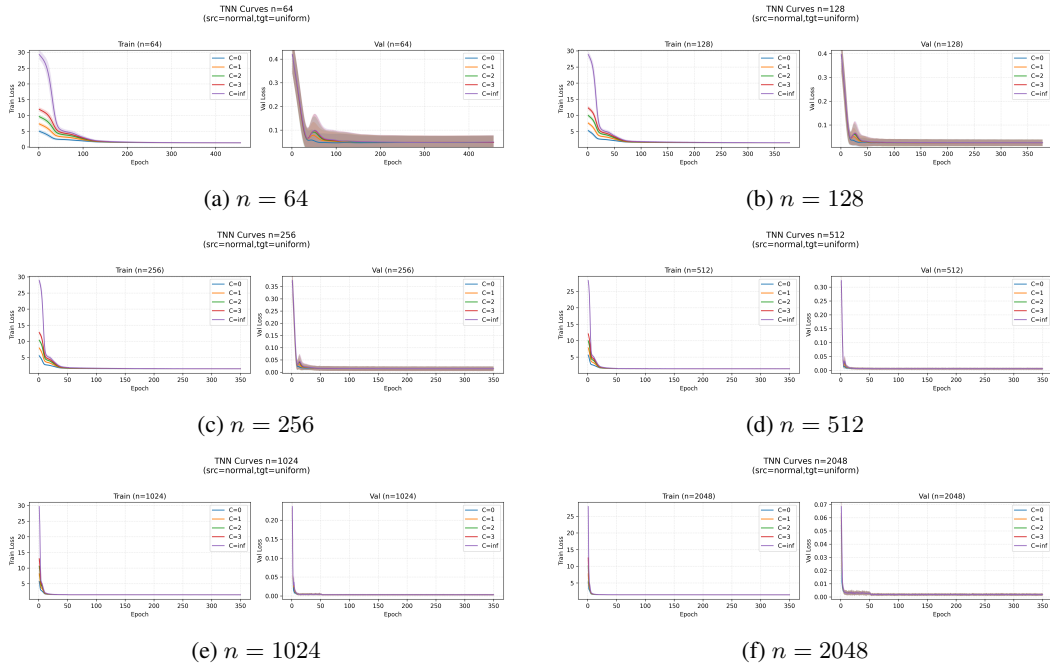


Figure 7: TNN training and validation loss curves for $d = 5$, source = Normal, target = Uniform, across different sample sizes n .

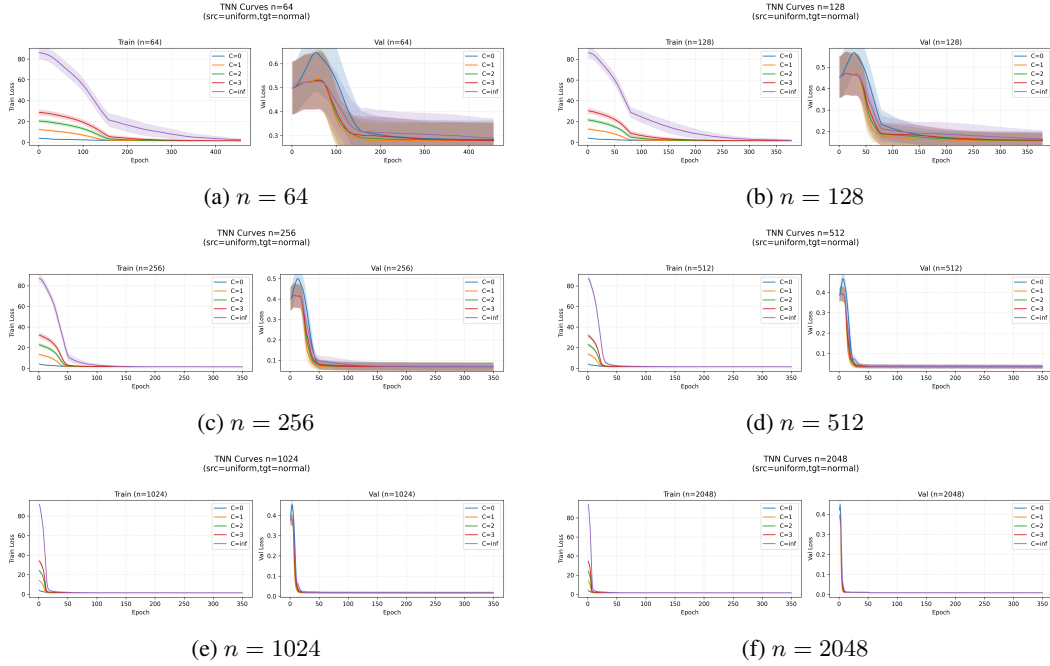


Figure 8: TNN training and validation loss curves for $d = 5$, source = Uniform, target = Normal, across different sample sizes n .

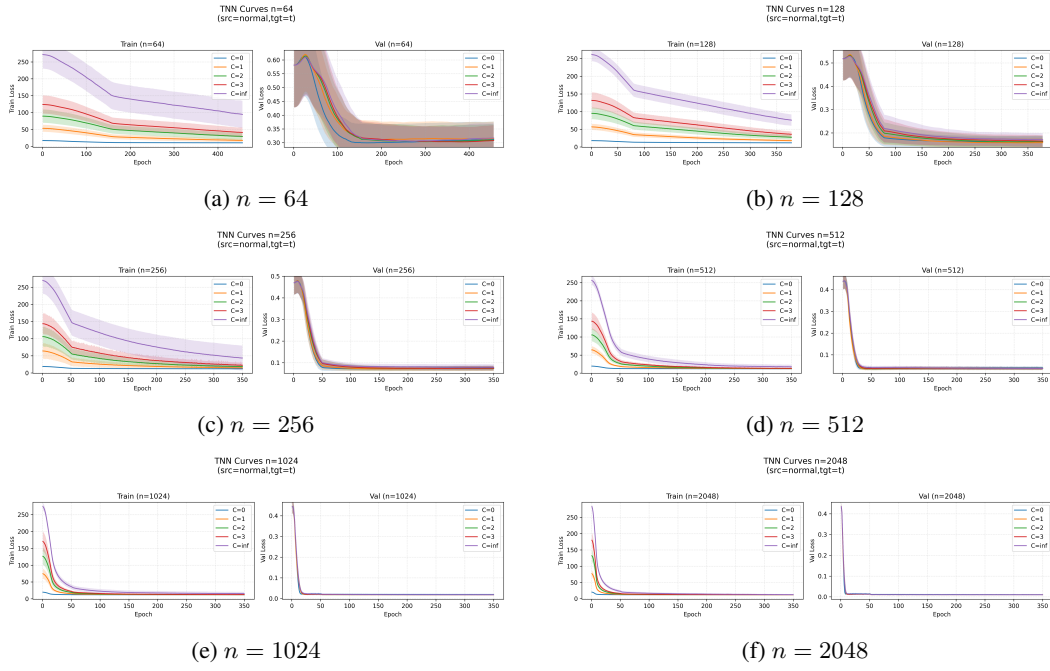


Figure 9: TNN training and validation loss curves for $d = 10$, source = Normal, target = Student- t , across different sample sizes n .

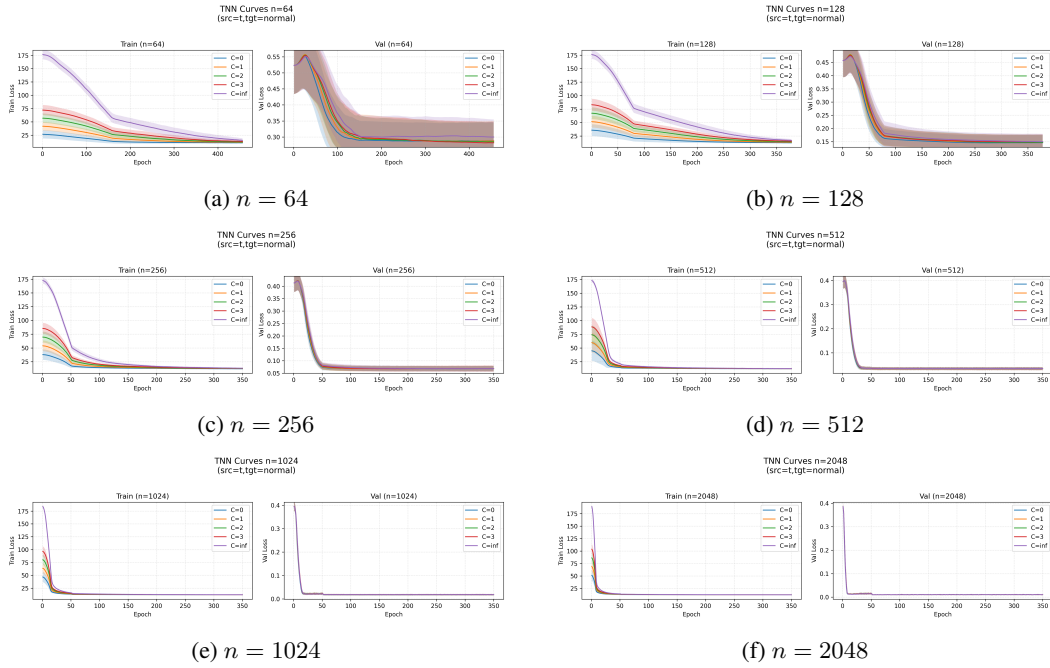


Figure 10: TNN training and validation loss curves for $d = 10$, source = Student- t , target = Normal, across different sample sizes n .

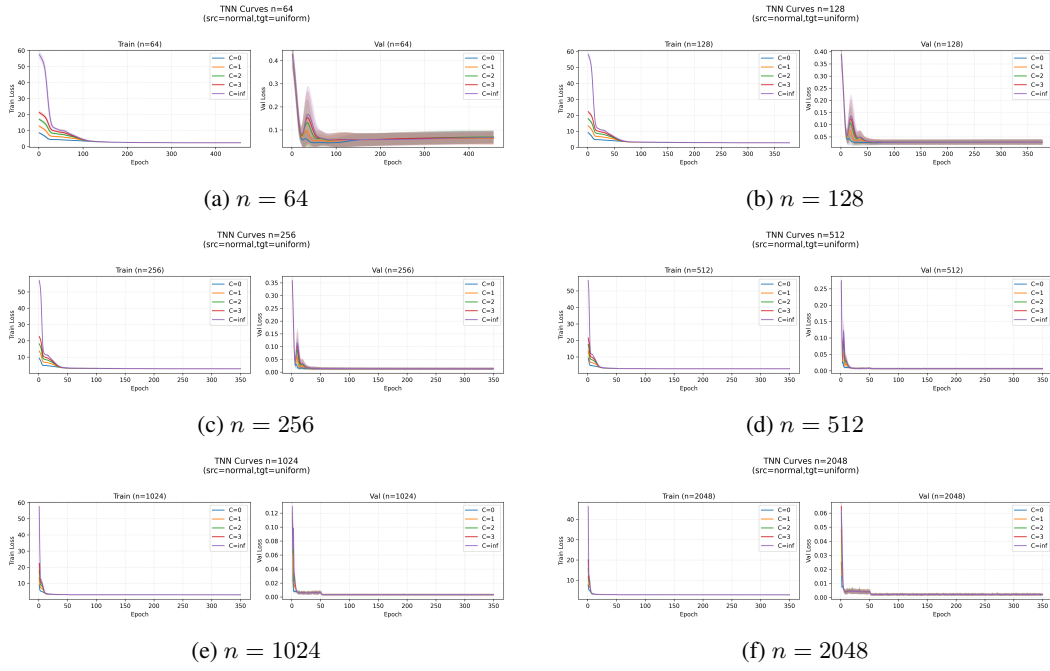


Figure 11: TNN training and validation loss curves for $d = 10$, source = Normal, target = Uniform, across different sample sizes n .

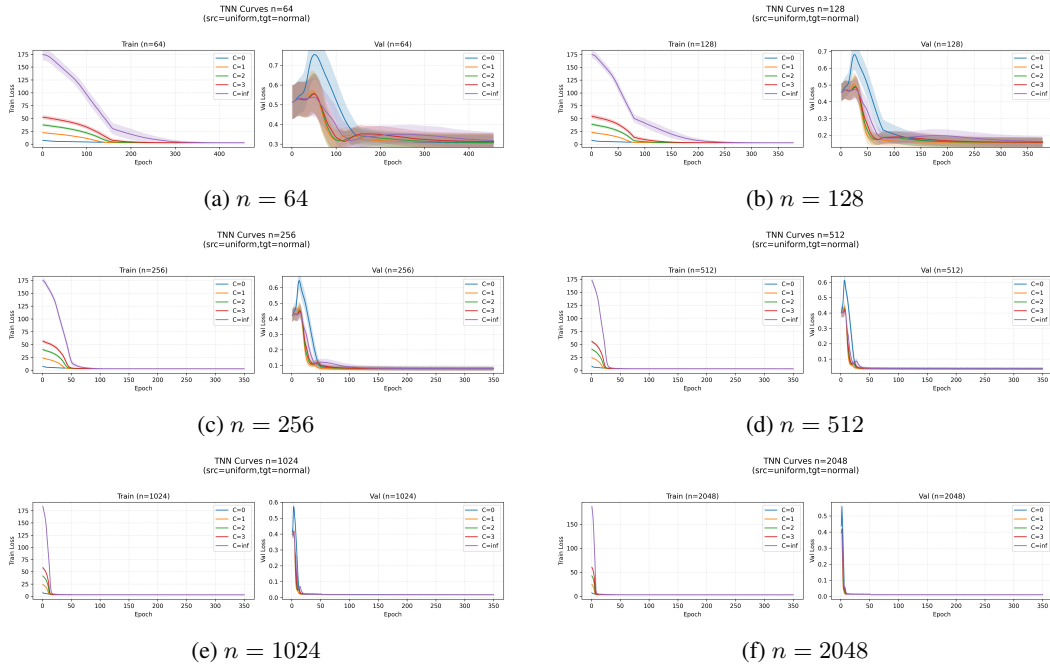


Figure 12: TNN training and validation loss curves for $d = 10$, source = Uniform, target = Normal, across different sample sizes n .

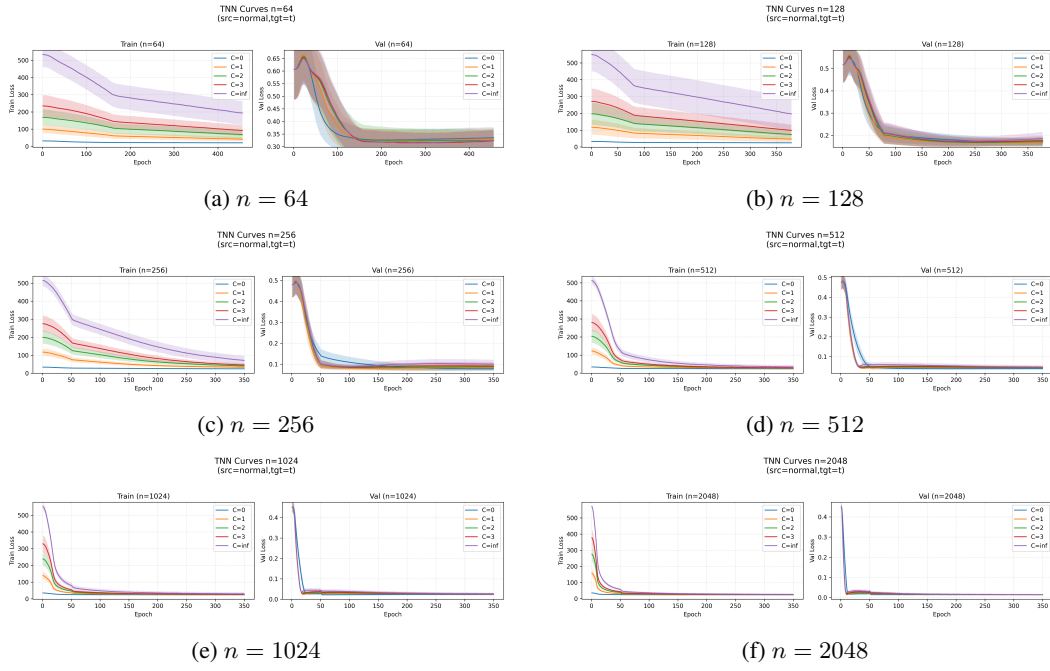


Figure 13: TNN training and validation loss curves for $d = 20$, source = Normal, target = Student- t , across different sample sizes n .

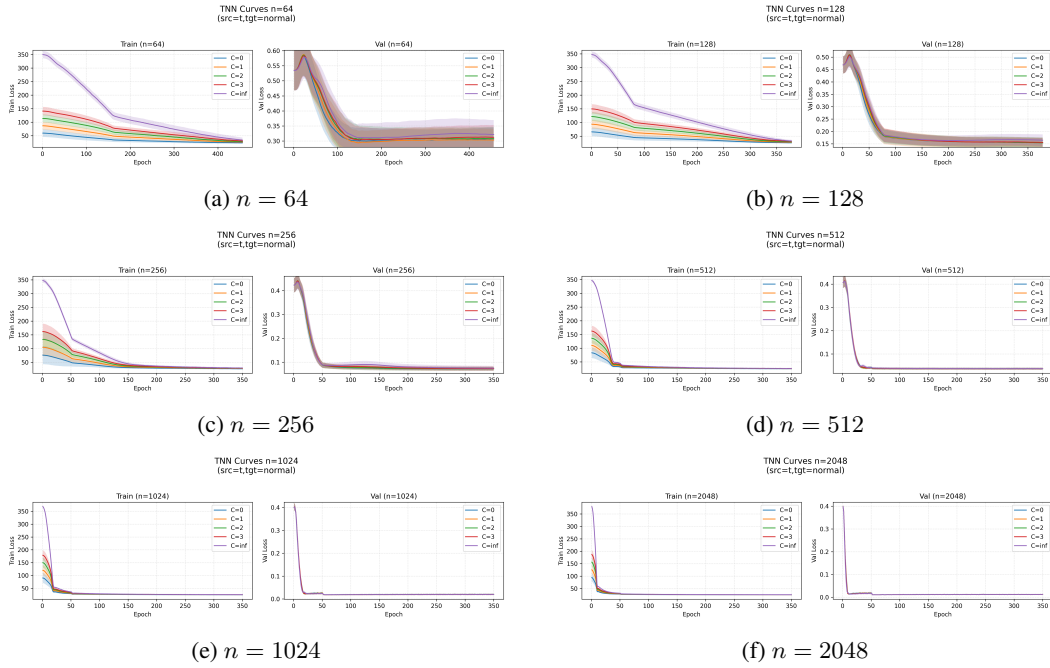


Figure 14: TNN training and validation loss curves for $d = 20$, source = Student- t , target = Normal, across different sample sizes n .

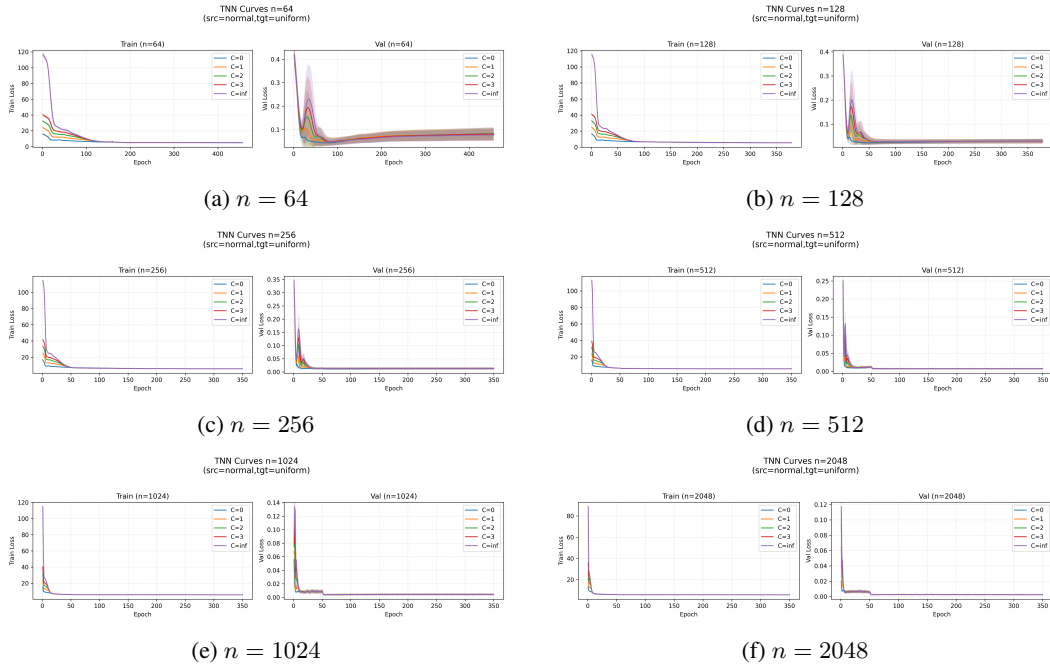


Figure 15: TNN training and validation loss curves for $d = 20$, source = Normal, target = Uniform, across different sample sizes n .

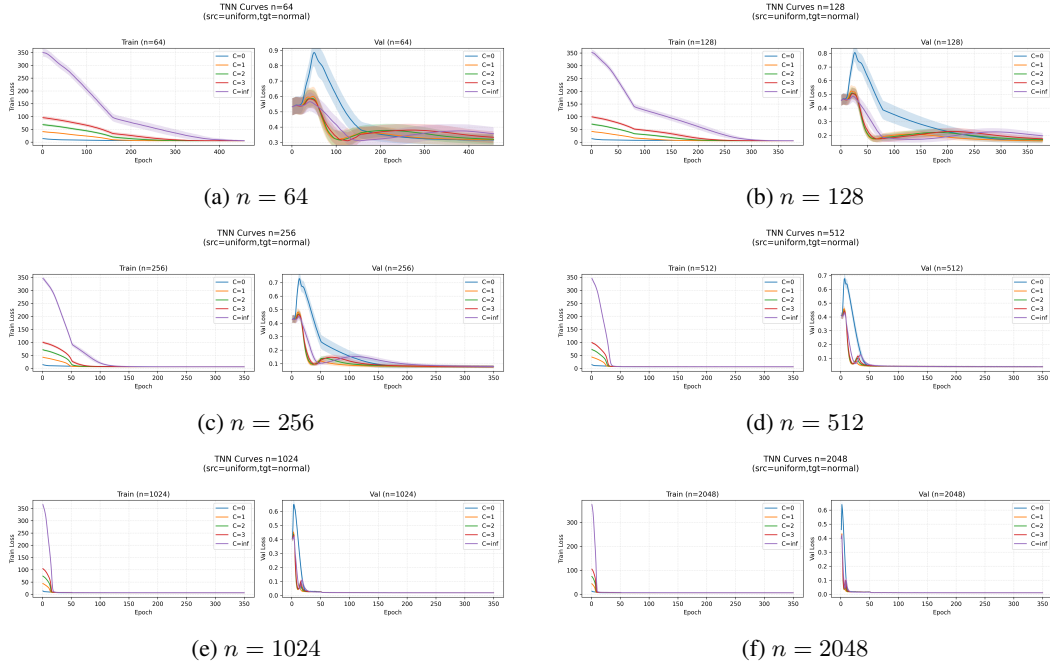


Figure 16: TNN training and validation loss curves for $d = 20$, source = Uniform, target = Normal, across different sample sizes n .

Appendix References

- [1] Minshuo Chen, Wenjing Liao, Hongyuan Zha, and Tuo Zhao. Distribution approximation and statistical estimation guarantees of generative adversarial networks. *arXiv preprint arXiv:2002.03938*, 2020.
- [2] Maria Colombo and Max Fathi. Bounds on optimal transport maps onto log-concave measures. *Journal of Differential Equations*, 271:1007–1022, 2021.
- [3] Tim De Ryck, Samuel Lanthaler, and Siddhartha Mishra. On the approximation of functions by tanh neural networks. *Neural Networks*, 143:732–750, 2021.
- [4] Yizhe Ding, Runze Li, and Lingzhou Xue. Statistical convergence rates of optimal transport map estimation between general distributions. *arXiv preprint arXiv:2412.08064*, 2024.
- [5] Vincent Divol, Jonathan Niles-Weed, and Aram-Alexandre Pooladian. Optimal transport map estimation in general function spaces. *The Annals of Statistics*, 53(3):963–988, 2025.
- [6] Florian F Gunsilius. On the convergence rate of potentials of brenier maps. *Econometric Theory*, 38(2):381–417, 2022.
- [7] Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 2021.
- [8] Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. In *International Conference on Learning Representations*, 2021.
- [9] Ilsang Ohn and Yongdai Kim. Smooth function approximation by deep neural networks with general activation functions. *Entropy*, 21(7):627, 2019.
- [10] A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.
- [11] Jon Wellner. *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer Science & Business Media, 2013.