A PROOFS

For completeness, we include the proofs of the main results in Carlier et al. (2016) with slight modifications in the statements and proofs to fit our variation of the problem.

Theorem 2 (Carlier et al. 2017)**.** *Assume $U \in \mathbb{R}^d$ is random vector with distribution μ, $(X, Y) \in \mathbb{R}^m \times \mathbb{R}^d$ is random vector with joint distribution ν. Furthermore, assume $\mathbb{E}(X|U) = \mathbb{E}(X) = 0$. If there exists smooth function $\varphi : \mathbb{R}^d \to \mathbb{R}$ and smooth function $b : \mathbb{R}^d \to \mathbb{R}^m$ such that $\varphi_x(u) = \varphi(u) + b(u)^\top x$ is convex for $\mathrm{Law}(X)$-almost every value of x such that $Y = \nabla\varphi(U) + \nabla b(U)^\top X$, then U solves the correlation maximization problem (7).*

Proof. Let $\Phi_X(U) = \varphi(U) + b(U)^\top X$ and define V with $\mathrm{Law}(V) = \mu$ and $\mathbb{E}(X|V) = 0$. From the Fenchel-Young inequality,

$$V^\top Y \leq \Phi_X(V) + \Phi_X^*(Y). \tag{15}$$

By assumption, $Y = \nabla\Phi_X(U)$ implying $Y \in \partial\varphi_X(U)$. The Fenchel-Young inequality satisfies equality if and only if $Y \in \partial\varphi_X(U)$, thus $U^\top Y = \Phi_X(U) + \Phi_X^*(Y)$. Taking the expectation of both sides of the equality, U is optimal by the Strong Duality Theorem. ☐

Theorem 3 (Carlier et al. 2017)**.** *Let ν be an absolutely continuous probability measure over $\mathbb{R}^m \times \mathbb{R}^d$ with density g. Assume the support of ν is $\bar{\Omega}$ where Ω is an open bounded convex subset of $\mathbb{R}^m \times \mathbb{R}^d$, and g is bounded on Ω and bounded away from zero on compact subsets of Ω. Then, the dual problem admits at least one solution.*

Proof. Let us denote the barycenter of ν as $(\bar{x}, \bar{y}) \equiv (0, \bar{y})$. Clearly we have $(0, \bar{y}) \in \Omega$. Otherwise, by convexity of Ω, ν would be supported on $\partial\Omega$ contradicting the assumption that $\nu \in L^\infty(\Omega)$.

Recall that ψ satisfies,

$$\psi(x, y) = \sup_{u \in \mathbb{R}^d} \{u^\top y - \varphi(u) - b(u)^\top x\}. \tag{16}$$

We can choose ψ to be 1-Lipschitz convex w.r.t. y such that,

$$|\psi(x, y) - \psi(x, \bar{y})| \leq \|y - \bar{y}\|. \tag{17}$$

The minimizer to our problem does not change up to additive constant C and so we choose C such that $\psi(0, \bar{y}) = 0$. Combining this into our constraint,

$$\varphi(t) \geq t\bar{y} - \psi(0, \bar{y}) \geq -|\bar{y}|. \tag{18}$$

Combining $\psi(x, \bar{y}) \geq 0$ with (17),

$$\psi(x, y) \geq -\|y - \hat{y}\| \geq -C, \tag{19}$$

where the last inequality comes from the boundedness of Ω. Let us take a minimizing sequence $(\psi_n, \varphi_n, b_n) \in C(\bar{\Omega}, \mathbb{R}) \times C(\mathbb{R}^d, \mathbb{R}) \times (\mathbb{R}^d, \mathbb{R}^N)$ s.t. all the aforementioned properties are satisfied for all n. Note that ψ_n, φ_n are bounded sequences in L^1 as $\varphi_n \geq -|y|$ and $\psi_n \geq -C$ and the sequence is minimizing. Set $z := (x, y) \in \Omega$ and $r > 0$ such that $B_r(z)$ is the set of all points at least distance r away from $\partial\Omega$. There exists $\alpha > 0$ s.t. $g \geq \alpha$ on $B_r(z)$ as we assumed g is bounded away from zero on compact subsets of Ω. We chose ψ_n to be convex, thus

$$-C \leq \psi_n(z) \leq \frac{1}{|B_r(z)|} \int_{B_r(z)} \psi_n \leq \frac{1}{\alpha|B_r(z)|} \int_{B_r(z)} |\psi_n| g \leq \frac{1}{\alpha|B_r(z)|} \|\psi_n\|_{L^1(\nu)} \tag{20}$$

$\implies \psi_n$ is locally bounded. Furthermore, the following inequality also holds due to convexity,

$$\|\nabla\psi_n\|_{L^\infty(B_r(z))} \leq \frac{2}{R - r} \|\psi_n\|_{L^\infty(B_R(z))}, \tag{21}$$

for $R > r$, $B_R(z) \subset \Omega$. This suggests ψ_n is locally uniformly Lipschitz as well. By the Arzelà-Ascoli theorem, there exists a subsequence of ψ_n that converges uniformly to ψ.

Now, take $r > 0$ s.t. $B_{2r}(0, \bar{y}) \in \Omega$. $\forall\, x \in B_r(0)$ and any $t \in \mathbb{R}^d$,

$$-b(t)^\top x \leq \varphi_n(t) - t^\top \bar{y} + \|\psi_n\|_{L^\infty(B_r(0,\bar{y}))} \leq C + \varphi_n(t) \tag{22}$$

for some $C > 0$. By maximizing $x \in B_r(0)$, we get $|b_n(t)|^\top r \leq C + \varphi_n(t)$. Since φ_n is bounded on L^1, then so is b_n. By Komlo's theorem, there exists a sub-sequence such that $\frac{1}{n}\sum_{k=1}^n \varphi_k \to \varphi$ and $\sum_{k=1}^n b_k \to b$ almost-everywhere. Indeed, our constraint inequality is still satisfied. Since the sequence $(\bar{\psi_n}, \bar{\varphi_n}, \bar{b_n}) = \sum_{k=1}^n \frac{(\psi_k, \varphi_k, b_k)}{n}$ is minimizing as well, we can use Fatou's lemma to arrive at,

$$\int_\Omega \varphi(x,y)\mathrm{d}\nu(x,y) + \int_{\mathbb{R}^d} \varphi(u)\mathrm{d}\mu(u) \leq \liminf_n \int_\Omega \bar{\varphi}(x,y)\mathrm{d}\nu(x,y) + \int_{\mathbb{R}^d} \bar{\varphi(u)}\mathrm{d}\mu(u) \tag{23}$$

$$= \inf_{\psi,\varphi,b} \int_\Omega \varphi(x,y)\mathrm{d}\nu(x,y) + \int_{\mathbb{R}^d} \varphi(u)\mathrm{d}\mu(u).$$

□

Theorem 4 (Carlier et al. (2017))**.** *Let $U \in \mathbb{R}^d$ be a solution to (7) and let $\Psi : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}, \varphi : \mathbb{R}^d \to \mathbb{R}, b : \mathbb{R}^d \to \mathbb{R}^m$ be solutions to the corresponding dual problem (13). Let $\varphi_x(t) = \varphi(t) + b(t)^\top x \;\forall (t,x) \in [0,1]^d \times \mathrm{support}(\mathrm{Law}(X))$. Then, $\varphi_X(U) = \varphi_X^{**}(U)$ and $U \in \partial\varphi_X^*(Y)$ almost surely.*

Proof. Define $\Phi_X(U) = \varphi(U) + b(U)^\top X$ and $\psi_X(Y) = \psi(X,Y)$. By the constraint $\psi_X(Y) + \Phi_X(U) \geq U^\top Y$ and by definition of $\psi(y) := \sup_{U \in R^d}\{U^\top Y - \varphi(U) - b(U)^\top X\}$ and order-reversibility of the Legendre transformation,

$$\psi_X \geq \Phi_X^*. \tag{24}$$

By strong duality, we have $\psi_X(Y) + \Phi_X(U) = UY$ almost surely. Then, $U^\top Y = \psi_X(Y) + \Phi_X(U) \geq \Phi_X^*(Y) + \Phi_X(U)$ thus $U^\top Y = \Phi_X^*(Y) + \Phi_X(U)$ and $\Phi_X^{**} \geq U^\top Y - \Phi_X^*(Y) = \Phi_X(U)$. Combining the above, we arrive at $\Phi_X(U) = \Phi_X^{**}(U)$ and $U^\top Y = \Phi^*(Y) + \Phi_X^{**}(U)$ which also suggests $U \in \partial\Phi_X^*(Y)$ and $Y \in \partial\varphi_X^{**}(U)$ almost surely. □ □

Proposition 1. *If $\varphi_x(U) : \mathbb{R}^d \to \mathbb{R}$ is convex w.r.t. $U \in \mathbb{R}^d$ and $U_k \leq V_k$, then $[\nabla\varphi_x(U)]_k \leq [\nabla\varphi_x(V)]_k$ for all $k = 1, \ldots, d$.*

Proof. Let H_{φ_x} denote the Hessian of φ_x. By convexity of φ_x, $H_{\varphi_x} \succcurlyeq 0$ so the diagonal entries of H_{φ_x} are non-negative. Otherwise if the i^{th} diagonal entry is negative, let e_i be the standard basis at the i^{th} coordinate, then $e_i^\top H_{\varphi_x} e_i < 0$, arriving at a contradiction. Consider the k^{th} component of $\nabla\varphi_x$, $(\nabla\varphi_x)_k = \frac{\partial\varphi_x}{\partial U_k}$, since $H_{\varphi_x} \succcurlyeq 0$ then $\frac{\partial^2\varphi_x}{\partial U_k^2} = diag(H_{\varphi_x})_k \geq 0$. Thus, $(\nabla\varphi_x)_k$ is comonotonic w.r.t. U. □

Proposition 2. *For any continuous conditional quantile function $Q_x(u)$ we can find large n and functions $f(x) : \mathbb{R}^m \to \mathbb{R}^n$, $\varphi(u) : \mathbb{R}^d \to \mathbb{R}$ and $b(u) : \mathbb{R}^d \to \mathbb{R}^n$ such that the gradient of $\varphi_x(u) := \varphi(u) + b(u)^\top f(x)$ approximates $Q_x(u)$ uniformly over any compact region of (u, x).*

Proof. Fixing x, the function $u \mapsto Q_x(u)$ is the gradient of some convex function $q_x(u) =: q(x,u)$. Thus, using results in Chen et al. (2019), we may approximate $q_x(u)$ with a Relu network $g(x,u)$. Using the standard compactness argument, we may approximate $q(x,u)$ over any compact convex region K by $\sum_{i=1}^n q_i(x_i, u)$. Now we use the results of Chen et al. (2019) to approximate each $q_i(x_i, u)$ with a convex Relu network $g_i(u)$. Define $f(x) = \mathbf{1}$ and $b(u) = [g_1(u), \ldots, g_n(u)]$ we then have $b(u)^\top f(x) = \sum_i g_i(u)$ which approximates $q(x,u)$ uniformly over the compact region K. Finally, we note that uniform approximation of a convex function also leads to approximation of its gradient (Rockafellar, 1970).

We note that we can also modify the arguments of Chen et al. (2019) to provide a more direct and possibly tighter proof. □

B EXPERIMENTAL DETAILS

B.1 PSEUDO-CODE OF DUAL OBJECTIVE

We present the pseudo-code for the correlation maximization dual objective below. Each tensor is assumed to be two dimensional; first dimension is the batch axis and the second dimension is the feature axis. Flatten all dimensions except the batch dimension to allow for vector dot product otherwise.

Algorithm 1: PyTorch-style code for computing the correlation maximization dual objective.

```
function loss $(a, b)$;
Input: $U$, $\hat{Y}$, $Y$, $X$
    $\varphi, b = \hat{Y}$
    $Y = Y$.permute(1, 0)
    $X = X$.permute(1, 0)
    $BX =$ torch.mm($b$, $X$)
    loss = torch.mean($\alpha$)
    $UY$ = torch.mm($U$, $Y$)
    $\psi = UY - \alpha - BX$
    sup, _ = torch.max($\psi$, dim=0)
    loss += torch.mean(sup)
    return loss
```

B.2 MODEL ARCHITECTURE

Table 3: Summary of quantile network architecture used for each experiment. The hidden dimension of the LSTMs are 4 times the input dimensions.

Experiment	**Input Dims**	φ **Layers**	φ **Units**	b **Layers**	b **Units**	f
`Toy`	2	3	128	1	128	Identity
`MNIST (No VAE)`	784	3	512	1	512	Identity
`MNIST+VAE`	5	3	128	1	128	Identity
`CelebA+VAE`	2048	3	4096	1	4096	Identity
`Energy`	28	3	128	1	128	2 Layer LSTM
`Stocks`	6	3	128	1	128	2 Layer LSTM

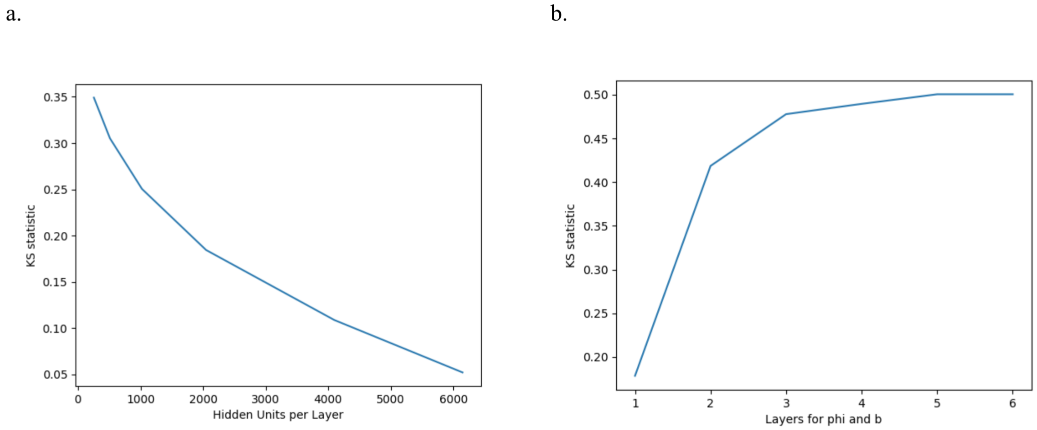

Figure 4: **Kolmogorov-Smirnov Test Ablation on High Dimensional Data.** The target distribution is a 2048-dimensional isotropic standard Gaussian. (a.) The KS statistic w.r.t. number of hidden units per layer. (b.) The KS statistic w.r.t. the number of layers. Both φ and b are scaled evenly. Lower KS statistic value is better.

C ADDITIONAL EXPERIMENTS AND RESULTS

C.1 HIGH DIMENSIONAL OPTIMAL TRANSPORT

Optimal transport approaches can suffer in high dimensional settings. Here, we study the effect of high dimensions on our dual objective, and suggest a way we empirically found to mitigate the curse of dimensionality. We set the target distribution as a $2048-$dimensional isotropic standard Gaussian. The baseline network is 3-layer $\varphi : \mathbb{R}^{2048} \rightarrow \mathbb{R}$ with a 2048-dimensional hidden layer.

We use the Kolmogorov-Smirnov test to compare the closeness of the generated distribution against the target distribution. In particular, the statistic is given by,

$$D_n = \sup_x |F_n(x) - F(x)| \tag{25}$$

where F is the cumulative distribution function of the distribution of X (where X_i assumed iid). In essence, this statistic measures the largest variation of the empirical against the target distribution.

In particular, we find that scaling up depth of the network leads to no improvements on learning the high dimensional target distribution, but scaling up the width leads to improvements. Figure 4 shows the effect of scaling width and scaling depth independently on fitting the high dimensional Gaussian. A theoretical justification is left for future work.

C.2 INPUT CONVEX VS. SMOOTH NEURAL NETWORKS

Previous works (Makkuva et al., 2020; Huang et al., 2021) that construct a Brenier map between two distributions parameterize their function as an ICNN. Doing so guarantees that the trained model is convex w.r.t. the input data. Contrary to these works, our work demonstrates an ICNN parameterization is not necessary and that the trained model is still convex w.r.t. the input as expected from optimal transport theory. Here, we study the effect of restricting the model to be input convex a priori against relaxing this assumption by using smooth neural networks to approximate convex functions. Table 4 displays the results of on the two time-series datasets `Energy` and `Stocks` averaged over 5 runs 10 epochs each.

We notice slight improvements of a smooth parameterization compared to the convex parameterization. We suspect this is due to the greater flexibility of smooth neural networks as the weights are not constrained to be non-negative, thus enabling better fitting the optimal convex potential.

Table 4: Performance evaluation on the multivariate time-series datasets `Energy` and `Stocks`. Results are averaged over 5 runs. Lower score is better. We use boldface for the lowest score.

Dataset	**Model**	**MaxAE**	**MeanAE**	**QL50**	**QL90**	**RMSE**	**sMAPE**
`Energy`	ICNN	0.863	0.063	0.030	0.018	0.100	0.246
	Smooth	0.624	0.051	0.026	0.038	0.091	0.210
`Stocks`	ICNN	0.739	0.024	0.014	0.009	0.040	0.205
	Smooth	0.660	0.024	0.012	0.014	0.036	0.220

C.3 More CelebA Samples

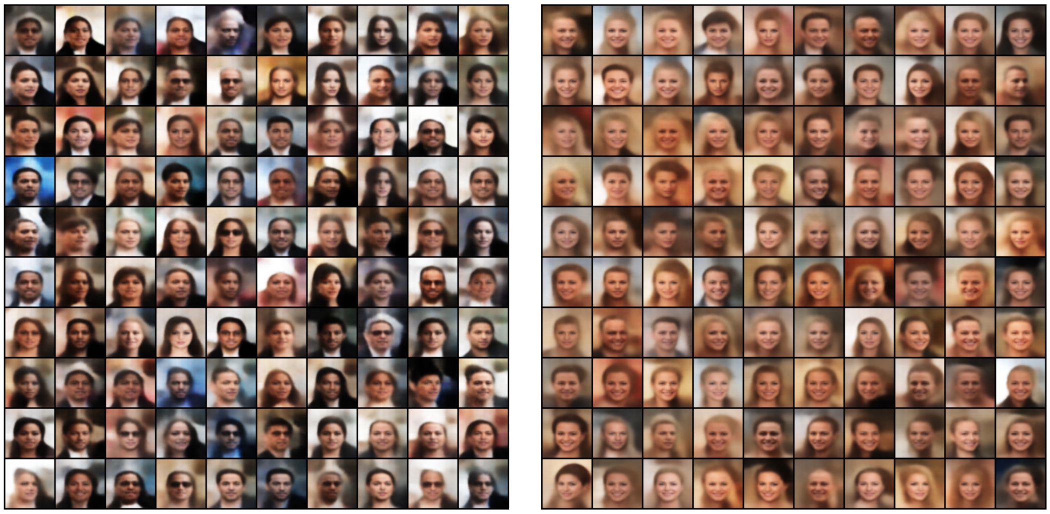

Figure 5: **Left**. Unconditioned. **Right**. Conditioned on Blond, Young, Smiling, Female, Mouth_Slightly_Open.

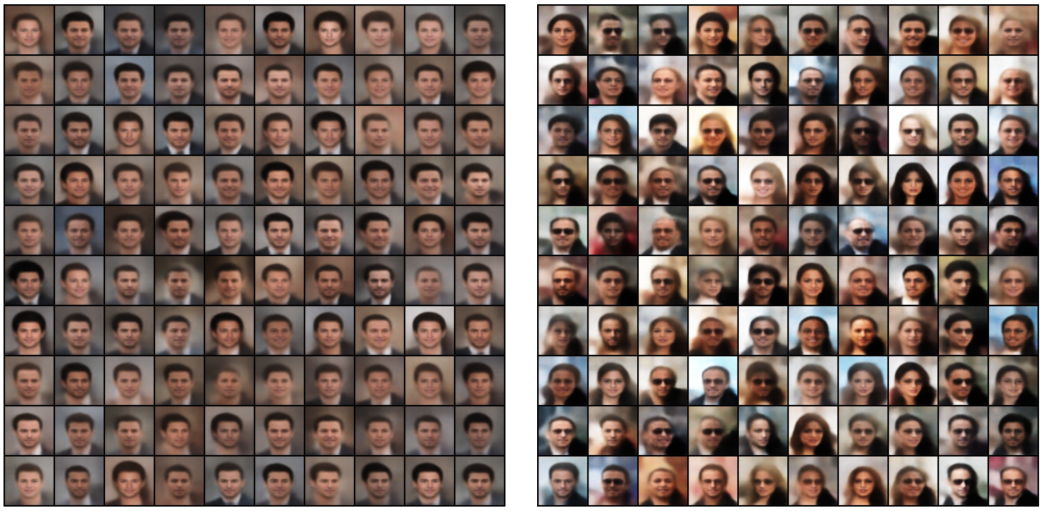

Figure 6: **Left**. Conditioned on 5_O_Clock_Shadow, Male, No_Beard, Straight_Hair. **Right**. Conditioned on Black Hair, Eyeglasses, Male, Smiling, Young.