Generative Conditional Distributions by Neural (Entropic) Optimal Transport

Bao Nguyen [1 2] **Binh Nguyen** [3] **Hieu Trung Nguyen** [1 4] **Viet Anh Nguyen** [1]

Abstract

Learning conditional distributions is challenging because the desired outcome is not a single distribution but multiple distributions that correspond to multiple instances of the covariates. We introduce a novel neural entropic optimal transport method designed to effectively learn generative models of conditional distributions, particularly in scenarios characterized by limited sample sizes. Our method relies on the minimax training of two neural networks: a generative network parametrizing the inverse cumulative distribution functions of the conditional distributions and another network parametrizing the conditional Kantorovich potential. To prevent overfitting, we regularize the objective function by penalizing the Lipschitz constant of the network output. Our experiments on real-world datasets show the effectiveness of our algorithm compared to state-of-the-art conditional distribution learning techniques. Our implementation can be found at https://github.com/nguyenngocbaocmt02/GENTLE.

1. Introduction

The conditional distribution of a response variable given the covariate information is a key quantity for numerous tasks in machine learning and data science. For example, knowing the conditional distribution of a patient's health outcome (the response) given the patient's characteristics (the covariate) improves risk assessment and prediction accuracy and allows personalized interventions wherein healthcare providers can tailor treatments based on individual patient profiles. It is also a useful tool to evaluate the effects of a policy in economics (Díaz, 2020; Athey & Imbens, 2015). Access to a response's conditional probability distribution also leads to an estimate of the treatment effect without explicit, additional experiments.

Unfortunately, learning the relationship between the response and the covariate is challenging, especially in high-dimensional covariate space or complex dependency settings. Many simple and interpretable methods, such as kernel density estimation, may struggle to capture the complexity of the data-generating conditional distribution accurately (Athey et al., 2021). On the other hand, there are overly complex (deep neural network) models that perform well on the training data but generalize poorly to new data; this overfitting behavior triggers unreliable predictions and interpretations. Additionally, performing inferences on these overly complex models can be time-consuming and environmentally unfriendly. Alternatively, using additional expert inputs, one can build physically informed simulation models to represent the real physical world. However, running simulations can be time-consuming, making them ill-suited for real-time decision-making.

Another critical obstacle underlying conditional distribution learning is that we are not simply learning one distribution; instead, we need to learn a collection of distributions, one conditional distribution for each possible value of the covariate information. This challenge intensifies because, in reality, the data is scarce. This scarcity is particularly prevalent in healthcare, where privacy concerns or budget constraints prohibit the collection and dissemination of samples. As an illustrative example, we will consider the Lalonde-Dehjia-Wahba (LDW) (LaLonde, 1986; Dehejia & Wahba, 1999) dataset, which is a widely used dataset for analyzing the average treatment effects (Athey et al., 2021). The LDW dataset comprises 16,177 samples aggregated from actual experiments and the Current Population Survey. Each observed sample includes nine covariate attributes of a person and a scalar response reporting the individual's earnings in 1978. Since two persons may have the exact attributes, it happens that for a specific covariate value, there are multiple observations of the response. The histogram in Figure 1 shows the count numbers per distinct covariates for the LDW dataset. We observe that more than 13,000 covariates have only one response. This is an extreme case where we must learn the conditional distribution with a single data point observed at that covariate. On the other hand, we also observe several popular covari-

---

[1]The Chinese University of Hong Kong [2]VinUniversity [3]Department of Mathematics, National University of Singapore [4]VinAI Research. Correspondence to: Bao Nguyen <bao.nn2@vinuni.edu.vn>.

ates with many responses: the most popular covariate has 47 responses. This again induces unbalanced sampling in which the number of responses is not balanced across different covariate values. A naive solution to fix this problem could be under-sampling (remove data) the more frequent covariate and over-sampling (create synthetic data) for the rare covariate to balance the dataset using popular methods such as SMOTE (Chawla et al., 2002), SMOTER (Torgo et al., 2013) and SMOGN (Branco et al., 2017). Unfortunately, these methods do not scale well for large datasets with more than thousands of covariates.
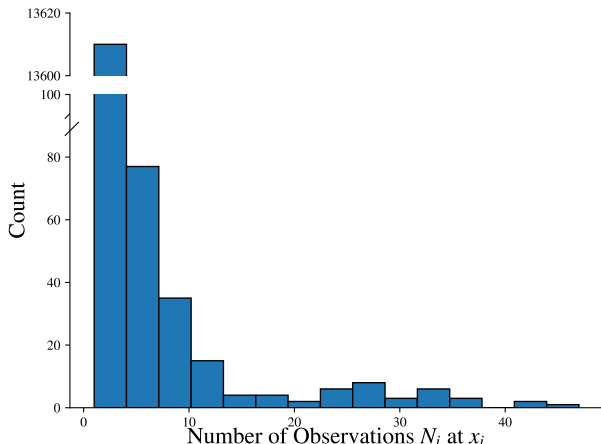

Figure 1. Histogram count of Number of observed responses for distinct covariate values in the LDW dataset (LaLonde, 1986). This dataset has over 13,000 covariates with only one response.

The current literature on conditional distribution learning reveals a notable gap in methodologies that can effectively address the aforementioned difficulties. Specifically, our experimental section shows that CWGAN (Athey et al., 2021) suffers from mode collapse and fails to recover the data-generating conditional distributions.

1.1. Problem Statement and Contributions

We consider X a multi-dimensional covariate and Y a one-dimensional response. For any specific value x, the data-generating conditional distribution of $Y|X = x$ is denoted by μ_x. Our goal is to learn these conditional distributions μ_x uniformly over all x from a finite dataset $\mathcal{D} = \{(x_i, \{y_{i,k}\}_{k=1,...,N_i}) \in \mathbb{R}^d \times \mathbb{R}$. Without any loss of generality, we suppose that the covariates x_i are distinct, and for each x_i, $\{y_{ik}\}_{k=1,...,N_i}$ are N_i independent samples drawn from μ_{x_i}.

More importantly, we aim to learn a *generative* model of the conditional distributions, which can allow us to draw new, unseen samples. This task is more challenging than learning specific finite-dimensional statistics of the conditional distributions, such as learning conditional expecta-

tions or quantiles, especially under a small training sample size. We reemphasize that conditional distribution learning is challenging: our goal is not simply learning one distribution, but instead, a collection of distributions – one conditional distribution for each possible value of the covariate information. This challenge intensifies because of data scarcity. This difficulty is often encountered in real-world deployment, for instance, in health care and finance.

Contributions. We introduce GENTLE, a **ge**nerative **n**eural **t**ransport **le**arning model for conditional distributions. GENTLE is represented by a neural network T_θ that takes the covariate x and a uniform $(0, 1)$ random variable U as input, and it outputs $T_\theta(x, U)$ with a distribution that is close to μ_x, the data-generating conditional distribution of $Y|X = x$. Due to the low sample size, T_θ can overfit, and we propose learning θ by minimizing the sum of the fitness and the overfitting regularization trade-off. The overfitting regularizer ensures that T_θ is sufficiently smooth in the covariate x, and we use an entropic optimal transport distance to measure the discrepancy of the generative distribution outputs. To this end, we create a second network v_ϕ to approximate the conditional potential associated with the semi-dual formulation of the entropic optimal transport distance. This leads to a min-max optimization problem over the pair of generative-conditional potential network parameters (θ, ϕ). This min-max problem is then solved using a state-of-the-art gradient descent-ascent algorithm.

Using real-life datasets, we empirically validate that GENTLE performs better than state-of-the-art methods such as conditional Wasserstein GANs across multiple performance metrics (Wasserstein metric and Kolmogorov-Smirnov metric). In particular, even for the case of LDW-CPS data that exhibits severe imbalances in observed responses, GENTLE still demonstrates a remarkable ability to generate observations that follow ground truth distributions closely, quantitatively, and qualitatively.

This paper unfolds as follows: Section 2 provides an overview of the existing literature. Section 3 reviews the technical background on optimal transports and the tools we will utilize. Section 4 presents our approach to generative conditional distributions. The numerical results are provided in Section 5.

Notations. We write the capital letter X for the random variable and the case letter x for a specific realization of X. For simplicity, we use $Y(x)$ to denote the conditional random variable $Y|X = x$. For a measurable map $f : \mathbb{R} \to \mathbb{R}$ and any probability measure $\mathbb{P}$ on $\mathbb{R}$, we use $f_\#\mathbb{P}$ to denote the pushforward measure of $\mathbb{P}$ under f. The (standard) uniform distribution on $(0, 1)$ is denoted $\mathcal{U}(0, 1)$, and we use $\mathbb{U}$ to represent the probability measure associated with this uniform distribution. We denote $N = \sum_i N_i$ as the total number of observations in the training dataset.

2. Related Works

Metamodeling is a computational approach that efficiently captures the relationship between input variables and key output statistics like means or quantiles. This is crucial for real-time simulation applications such as personalized decision making (Shen et al., 2021) and online risk monitoring (Jiang et al., 2020). Traditional metamodeling techniques, while effective, often require pre-specification of relevant statistics, limiting their flexibility. To address this, generative metamodeling has emerged as a promising technique that estimates conditional distributions quickly and dynamically. Generative metamodeling differs from traditional approaches by simulating a wider range of possible outcomes, providing a more comprehensive analysis of potential scenarios. Athey et al. (2021) propose using a Wasserstein Generative Adversarial Model (WGAN, Arjovsky et al. 2017) to simulate the distribution of outcomes conditioned on covariates, offering a novel approach to outcome simulation. Similarly, Hong et al. (2023) propose quantile-regression-based generative metamodeling (QRGMM), another method of conditional distribution learning. QRGMM leverages the conditional quantile regression technique to provide a more detailed understanding of the distribution tails, which is crucial in risk assessment and decision-making processes. These advancements in generative metamodeling can significantly enhance the accuracy and applicability of real-time simulations, paving the way for more personalized and precise decision-making tools in various fields.

Synthetic evaluation data. Previous works (van Breugel et al., 2024; Parikh et al., 2022) leverage the deep generative model to approximate conditional distribution for synthetic data generation. (Shrivastava & Tuzel, 2019) parameterize a conditionally Gaussian model with a Deep Neural Network to model error distribution for each state of the object.

Neural Optimal Transport is a recent emergent framework that proposes to learn the optimal transport plan by deep neural networks. Various existing works in this direction include parameterization of the OT map as the gradient of an input convex neural network (Jacob et al., 2019; Makkuva et al., 2020; Bunne et al., 2022; 2023); or solutions of convex-concave problems (Korotin et al., 2022; Gushchin et al., 2022; Choi et al., 2023). However, these formulations are primarily suited to computer vision or computational biology applications. To the best of our knowledge, our paper is the first work attempting to formulate the neural OT framework into a conditional distribution learning task.

Learning and decision-making with conditional distributions also attracted tremendous attention in artificial intelligence and operations research communities. Nguyen et al. (2020b) studied the robust parameter estimation of conditional distributions under the parametric exponential family of distributions, while Nguyen et al. (2020a) studied the non-parametric Wasserstein robust estimation of conditional statistics. Making conditional decisions based on observed covariates is a well-known problem in finance (Nguyen et al., 2024) and operations research (Esteban-Pérez & Morales, 2022); interested readers are invited to the survey by Sadana et al. (2024) for recent developments.

3. Background

We focus on a one-dimensional response space for Y. The optimal transport distance between two distributions supported on $\mathbb{R}$ is formally defined as follows.

Definition 3.1 (Optimal Transport Distance, Villani 2009). Given two probability measures μ and ν supported on $\mathbb{R}$ with finite second moments, the optimal transport (2-Wasserstein) distance between them is defined as

$$\mathrm{W}_2^2(\mu,\nu) \stackrel{\text{def.}}{=} \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} |y-y'|^2 \, \mathrm{d}\pi(y,y'), \quad (1)$$

where $\Pi(\mu,\nu)$ denotes the set of probability distributions on $\mathbb{R}^2$ with marginals μ and ν, respectively.

Optimal transport in 1d. In one dimension, the 2-Wasserstein distance admits a closed form expression (Bobkov & Ledoux, 2019, Theorem 2.10): for two measures μ and ν with cumulative density functions (CDF) F_μ and F_ν, respectively, we have

$$\begin{aligned} \mathrm{W}_2^2(\mu,\nu) &= \int_0^1 |F_\mu^{-1}(t) - F_\nu^{-1}(t)|^2 \mathrm{d}t \\ &= \mathbb{E}_{U\sim\mathcal{U}(0,1)}\left[|F_\mu^{-1}(U) - F_\nu^{-1}(U)|^2\right]. \end{aligned} \quad (2)$$

Entropic-regularized optimal transport. The recent popularity in applications of OT to machine learning is largely due to its efficient computation by adding negative entropy of the transport plan as a form of regularization to the linear OT program (1). This problem is called entropic optimal transport (EOT), which is smooth and strongly convex, therefore has a unique minimizer:

$$\mathrm{W}_{2,\varepsilon}^2(\mu,\nu) \stackrel{\text{def.}}{=} \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^2} |y-y'|^2 \mathrm{d}\pi - \varepsilon H(\pi). \quad (3)$$

Cuturi (2013) showed that the minimization problem (3) can be solved by Sinkhorn's algorithm (a type of matrix scaling algorithm, also called iterative proportional fitting procedure, Kullback 1968) that can be implemented at large scale and is analytically tractable. In addition to its computational advantages, problem (3) can also be interpreted as a specific instance of the Schrödinger bridge

problem, which has a rich history in physics. For an indepth theoretical introduction to EOT, readers can refer to Nutz (2021) or Léonard (2014).

A nice property of the EOT problem (3) is that it allows an equivalent smooth unbounded semi-dual formulation (Genevay et al., 2016, Proposition 2.1)

$$
\min_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^2} |y - y'|^2 \mathrm{d}\pi - \varepsilon H(\pi)
$$
$$
= \sup_{v(\cdot) \in \mathcal{C}} \int_{\mathbb{R}} v^\varepsilon(y) \mathrm{d}\mu + \int_{\mathbb{R}} v(y') \mathrm{d}\nu - \varepsilon, \tag{4}
$$

where $\mathcal{C}$ is the space of all continuous functions from $\mathbb{R}$ to $\mathbb{R}$, and v^ε is defined as

$$
v^\varepsilon(y) = -\varepsilon \log \left(\int_{\mathbb{R}} \exp \left(\frac{v(y') - |y - y'|^2)}{\varepsilon} \right) \mathrm{d}\nu \right). \tag{5}
$$

The function v is called Kantorovich potential, along with its smoothed c-transform v^ε, with c being the absolute difference bi-function. Using the *semi*-dual form is that we need to optimize over only one function v associating with ν, and we can obtain v^ε as the potential associating with μ.

4. Methodologies

We propose to learn a mapping $T : \mathbb{R}^d \times (0, 1) \to \mathbb{R}$ such that uniformly over all x, $T(x, U) = Y(x)$ in distribution when $U \sim \mathcal{U}(0, 1)$. The existence of this mapping $T(x, U)$ is guaranteed by the following celebrated result.

Lemma 4.1 (Noise Outsourcing, Austin 2015). *Suppose that $\mathcal{X}$ and $\mathcal{Y}$ are standard Borel spaces and that (X, Y) is an $(\mathcal{X} \times \mathcal{Y})$-valued random variable. Then, there are random variables $U \sim \mathcal{U}(0, 1)$ coupled with $\mathcal{X}$ and $\mathcal{Y}$ and a Borel function $T : \mathcal{X} \times (0, 1) \to \mathcal{Y}$ such that U is independent from X and*

$$
(X, Y) = (X, T(X, U)) \quad \text{almost surely.}
$$

In general, the mapping T is not unique. Following the neural OT framework (Korotin et al., 2022; Makkuva et al., 2020; Bunne et al., 2022), we use a neural network to parametrize a possible instance of $T(x, U)$, and we denote this network as $T_\theta(x, U)$. For any specific value of the parameter θ, the network outputs a probability measure $T_\theta(x, \cdot)_\# \mathbb{U}$, which represents the distribution of $T_\theta(x, U)$ with $U \sim \mathcal{U}(0, 1)$. We propose the following regularized loss to find the optimal map T_θ:

$$
\min_\theta \mathrm{Fit}(\theta) + \lambda \mathrm{Reg}(\theta), \tag{6}
$$

where the first term is a measure of the fitness of the map, the second is a regularization to prevent overfitting, and $\lambda \geq 0$ is the regularization weight. We explain the detailed formulation of each term and its rationale in the next sections.

4.1. Fitness Measure

Motivated by the OT distance in (2), a natural choice to measure the discrepancy between the network outputs and the conditional data-generating distributions is by integrating the discrepancy of the respective inverse CDFs over all possible values x, *i.e.*, by evaluating

$$
\mathbb{E}_{X \sim \mathcal{D}, U \sim \mathcal{U}(0,1)}[|F_{\mu_X}^{-1}(U) - T_\theta(X, U)|^2]
$$
$$
= \mathbb{E}_{X \sim \mathcal{D}} \left[\mathbb{E}_{U \sim \mathcal{U}(0,1)}[|F_{\mu_X}^{-1}(U) - T_\theta(X, U)|^2] \right]. \tag{7}
$$

Above, F_{μ_x} is the cumulative distribution function of $Y(x)$, and the outer expectation is taken over the (uniform) sampling of all distinct values of X in the training dataset $\mathcal{D}$. Unfortunately, the inverse CDF $F_{\mu_X}^{-1}$ can be unbounded, and the magnitude of $F_{\mu_X}^{-1}$ explodes to infinity for tail events whenever U is to close to the boundary 0 or 1. Hence, learning T_θ with the above metric is not stable. To improve stability, we use the following fitness function

$$
\mathrm{Fit}(\theta) \stackrel{\text{def.}}{=} \mathbb{E}_{X \sim \mathcal{D}} \left[\mathbb{E}_{U \sim \mathcal{U}(0,1)}[|U - F_{\mu_X}(T_\theta(X, U))|^2] \right]. \tag{8}
$$

In practice, we do not have access to the real cumulative distribution of $Y(x)$. However, we have a set of responses sampled from $Y(x)$, so we propose to use a simple method such as Kernel Density Estimator (KDE) with the kernel bandwidth of h to have a smooth conditional density from responses of the covariate x in the training dataset. The kernel bandwidth plays a vital role in our objective function, and tuning this parameter is important. In theory, our KDE smoothness using the Gaussian kernel is strongly related to the smoothed Wasserstein distance. From Nietert et al. (2021, Lemma 1 and Corollary 1), the smooth-2-Wasserstein distance is continuous and monotonically non-increasing in the Gaussian kernel bandwidth. This means that the larger the bandwidth (the more smoothing applied), the smaller the fitness term will become. In practice, we could implement an automatic kernel bandwidth selection such as the Scott (Scott, 1979) or the Silverman (Silverman, 2018) methods. In our implementation, the value of the bandwidth is selected by a grid search to minimize the Wasserstein distance on the validation set.

4.2. Regularization Measure

Minimizing the $\mathrm{Fit}(\theta)$ term can lead to overfitting due to the low sample sizes for many values of x_i. To combat this problem, we add a regularization term to promote the transfer learning across different values of x_i. Consider the CPS dataset, we examine the relationship between the covariate distance $\|x_i - x_j\|$ versus the empirical Wasserstein distance between the conditional random variables $Y(x_i)$ and $Y(x_j)$. For better accuracy, we select only pairs (x_i, x_j) such that both N_i and N_j are greater than 18. The scatter plot in Figure 2 indicates a positive correlation between
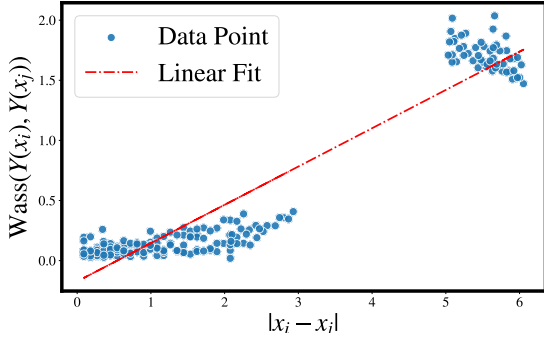
Figure 2. Positive correlation between the Wasserstein distance between $Y(x_i)$ and $Y(x_j)$ and the covariate distance $\|x_i - x_j\|$. Only covariates x with more than 18 observations are selected.

the covariate distance and the Wasserstein distance of the conditional distributions. We remark that we observed the same trend in other datasets.

From this empirical evidence, we can assume that if two covariates x_i and x_j are similar, their respective conditional distributions of $Y(x_i)$ and $Y(x_j)$ should also be similar. We emphasize that relying on this assumption to design the regularization term does not restrict the generalization of our framework. This assumption, in general, holds for the case where the data-generating process follows the form $Y = f(X) + \text{noise}$, where f is a continuous function. This is a classic setting in distribution learning, a more general form of linear assumption. Moreover, classic algorithms for supervised and unsupervised learning, such as kNN or k-clustering, also require this assumption to work. We can, therefore, assume a Lipschitz condition that there exists a finite positive number L such that

$$W_2^2(\mu_{x_i}, \mu_{x_j}) \leq L\|x_i - x_j\| \qquad \forall i, j,$$

where μ_x is the data-generating conditional probability measure for $Y(x)$. If the generative network T_θ matches sufficiently well the conditional distributions, then it is reasonable to translate the above Lipschitz assumption into a regularization term: let $\mathcal{E}$ be a set of pairs (x_i, x_j) for which we would like to impose Lipschitz conditions, we can regularize using the sum of the pairwise *entropic* OT distance:

$$\sum_{(x_i, x_j) \in \mathcal{E}} W_{2,\varepsilon}^2 \left(T_\theta(x_i, \cdot)_{\#} \mathbb{U}, T_\theta(x_j, \cdot)_{\#} \mathbb{U} \right). \qquad (9)$$

The semi-dual form of the EOT (4) leads to

$$\sum_{(x_i, x_j) \in \mathcal{E}} \left(\max_{v_{ij}(\cdot) \in \mathcal{C}} \int_{\mathbb{R}} v_{ij}^\varepsilon(y) d\mathbb{P}_{\theta,j} + \int_{\mathbb{R}} v_{ij}(y) d\mathbb{P}_{\theta,i} - \varepsilon \right).$$

Above, we have used the shorthand $\mathbb{P}_{\theta,i} = T_\theta(x_i, \cdot)_{\#} \mathbb{U}$, the pushforward measure of the uniform measure under the

map $T_\theta(x_i, \cdot)$. Each maximization problem inside the sum is an infinite-dimensional optimization problem, in which we need to search over all potentials v_{ij} continuous. To alleviate this computational burden, a naive approach is to use a neural network $v_\phi(x_i, x_j, \cdot) \approx v_{ij}(\cdot)$, where ϕ is a learnable parameter. By omitting the constant ε and switching the order of the summation and the maximization operator, we can regularize using

$$\max_\phi$$
$$\sum_{(x_i, x_j) \in \mathcal{E}} \left(\int_{\mathbb{R}} v_\phi^\varepsilon(x_i, x_j, y) d\mathbb{P}_{\theta,j} + \int_{\mathbb{R}} v_\phi(x_i, x_j, y) d\mathbb{P}_{\theta,i} \right).$$

A downside of this parametrization is that it requires both x_i and x_j as input to v_ϕ to output the potential. This leads to redundancy in the parametrization. To combat this problem, we propose a *single* input network $v_\phi(x_i, y)$ to approximate $v_{ij}(y)$, the potential associating with $\mathbb{P}_{\theta,i}$. Further, we compute the smoothed transform $v_\phi^\varepsilon(x_i, y)$

$$v_\phi^\varepsilon(x_i, y) \overset{\text{def.}}{=}$$
$$- \varepsilon \log \left(\int_{\mathbb{R}} \exp \left(\frac{v_\phi(x_i, y') - |y - y'|^2)}{\varepsilon} \right) d\mathbb{P}_{\theta,i} \right),$$

and use $v_\phi^\varepsilon(x_i, y)$ to approximate $v_{ij}^\varepsilon(y)$, the potential associating with $\mathbb{P}_{\theta,j}$. The above single input parametrization no longer takes x_j as input; thus, to make the network well-defined, we need disambiguation by imposing the following condition: for each distinct value x_i, there is *at most one* pair (x_i, x_j) that belongs to $\mathcal{E}$. If this condition on $\mathcal{E}$ holds, then we can uniquely identify the x_j such that $v_\phi^\varepsilon(x_i, y)$ is serving as the potential.

Thus, for a valid set $\mathcal{E}$, we propose the following regularization term:

$$\text{Reg}(\theta) \overset{\text{def.}}{=} \max_\phi \ \mathcal{R}(\theta, \phi), \qquad (10a)$$

where $\mathcal{R}(\theta, \phi)$ is

$$\mathcal{R}(\theta, \phi) \overset{\text{def.}}{=} \qquad\qquad\qquad\qquad\qquad (10b)$$
$$\sum_{(x_i, x_j) \in \mathcal{E}} \left(\int_{\mathbb{R}} v_\phi^\varepsilon(x_i, y) d\mathbb{P}_{\theta,j} + \int_{\mathbb{R}} v_\phi(x_i, y) d\mathbb{P}_{\theta,i} \right).$$

In this way, $\text{Reg}(\theta)$ is the sum of the semi-dual neural entropic OT distances of all pairwise distributions specified by $\mathcal{E}$.

To conclude this section, we elaborate on a reasonable approach to building $\mathcal{E}$. Ideally, $\mathcal{E}$ should capture neighborhood information about which pair (x_i, x_j) is close to each other; further, it should satisfy our disambiguation criterion that for each distinct value x_i, there is *at most one* pair (x_i, x_j) that belongs to $\mathcal{E}$. A reasonable choice of $\mathcal{E}$ that is

also computationally efficient is to build $\mathcal{E}$ as the *directed* edge set of a minimum spanning tree of all x. Here, the spanning tree minimizes the sum of distance $\|x_i - x_j\|$ between two connected nodes, and we can employ Kruskal's algorithm (Kruskal, 1956). For the direction of the edge: at each iteration, when we add x_i to the incumbent minimum spanning tree, then x_i is chosen as the tail of the edge, and the other endpoint of the adding edge (which already belongs to the incumbent tree) is chosen as the head of the edge. This construction will satisfy the disambiguation criterion for the set $\mathcal{E}$. We emphasize that the graph-based approach to identify which x_j is associated with x_i in the regularization term is not restricted only to a discrete set of covariates. This is because the graph-based approach is only employed during the training when the number of covariates is finite. In inference phases, the model can give the distributional prediction to any covariates.

Remark 4.2. In (8), we use the unregularized OT distance as a discrepancy function for the fit term $\text{Fit}(\theta)$ because therein, we are measuring the discrepancy between a fixed target (a uniform distribution U) and a quantity that depends on the generative network output $T_\theta(X, U)$ and the data-generating CDFs F_{μ_X}. This contrasts with the regularization term $\text{Reg}(\theta)$, where we need to compare two distributions that are both outputs of the generative network T_θ. Using the unregularized OT distance for the $\text{Reg}(\theta)$ will require evaluating the CDFs of T_θ. To alleviate this difficulty, we use the semi-dual form of EOT, where we can parametrize a single conditional potential network to evaluate all the pairwise EOT distances between covariate pairs in $\mathcal{E}$.

4.3. Min-Max Optimization with Smooth Gradient Descent-Ascent

By substituting the form of the regularization term (10) into (6), we obtain the min-max optimization problem:

$$\min_\theta \max_\phi \ \text{Fit}(\theta) + \lambda \mathcal{R}(\theta, \phi). \tag{11}$$

Naively optimizing problem (11) will result in a common pitfall in optimization of training GANs: gradient descent-ascent (GDA) on nonconvex-nonconcave objective is, in general, hard to converge (Heusel et al., 2017; Lin et al., 2020; Zheng et al., 2023). To alleviate this problem, we employ the state-of-the-art first-order algorithm for GDA (Zheng et al., 2023) with a convergence guarantee. First, for parameters $r_1 \neq r_2$, $r_1, r_2 > 0$, we add two extra regularization (smoothing) terms with auxiliary variables p and q into the loss function. This leads to the augmented objective

$$\mathcal{L}(\theta, \phi, p, q)$$
$$\stackrel{\text{def.}}{=} \text{Fit}(\theta) + \lambda \mathcal{R}(\theta, \phi) + \frac{r_1}{2}\|\theta - p\|_2^2 - \frac{r_2}{2}\|\phi - q\|_2^2. \tag{12}$$

Including the quadratic terms enhances the smoothness of the primal and dual update, facilitating a more balanced trade-off between the gradient updates in our nonconvex-nonconcave setting. In practice, we must evaluate this loss with training samples, detailed in Algorithm 1. Our training procedure is summarized in Algorithm 2. This doubly-smoothed GDA update is guaranteed to converge to a stationary point (Zheng et al., 2023, Theorem 2) as long as the max term $\mathcal{R}(\theta, \phi)$ satisfies Kurdyka-Łojasiewicz condition (a nonsmooth extension of the more popular Polyak-Łojasiewicz condition).

In general, we resort to a validation procedure to choose the appropriate parameters $h, \lambda, \varepsilon, r_1, r_2$, and the optimization algorithm's learning rates $(\alpha, \beta, \gamma, \delta)$. Details on the parameters grid are given in the Appendix A.

Algorithm 1 Empirical loss evaluation

1: **Input:** data $(x_i, \hat{F}_{x_i})$, network T_θ and v_ϕ, auxiliary variables p, q, batch size B, set of pair covariates $\mathcal{E}$, regularization parameter $\varepsilon, \lambda, r_1, r_2$
2: **Output:** loss estimate $\widehat{\mathcal{L}}(\theta, \phi, p, q)$
3: **Initialize:** take a batch of B training nodes $\{x_i^{(b)}\}_{b \in [B]}$
4: **for** each $(x_i^{(b)}, x_j^{(b)}) \in \mathcal{E}$ **do**
5:     Sample $U^{(b)} \sim \mathcal{U}[0, 1]$
6:     **for** $m = 1, \cdots, M$ **do**
7:         Sample $U^{(b,m)} \sim \mathcal{U}[0, 1]$
8:     **end for**
9: **end for**
10: Compute following (8):

$$\widehat{\text{Fit}}(\theta) \leftarrow \frac{1}{B} \sum_b \left[U^{(b)} - \hat{F}_{x_i^{(b)}}(T_\theta(x_i^{(b)}, U^{(b)})) \right]_2^2$$

11: Compute following (10):

$$\sum_{(x_i^{(b)}, x_j^{(b)}, U^{(b,m)})} \left[v_\phi^\varepsilon(x_i^{(b)}, T_\theta(x_j^{(b)}, U^{(b,m)})) \right.$$
$$\left. + v_\phi(x_i^{(b)}, T_\theta(x_i^{(b)}, U^{(b,m)})) \right] / (M \times B)$$

12: Compute following (12):

$$\widehat{\mathcal{L}}(\theta, \phi, p, q)$$
$$\leftarrow \widehat{\text{Fit}}(\theta) + \lambda \hat{\mathcal{R}}(\theta, \phi) + \frac{r_1}{2}\|\theta - p\|_2^2 - \frac{r_2}{2}\|\phi - q\|_2^2$$

---

5. Experiments

Datasets. To benchmark our proposed method GENTLE, we use the following two datasets and the same preprocessing steps as in Athey et al. (2021).

Algorithm 2 Training of the Minimax loss (12)

1: **Input:** data $\mathcal{D} = (x_i, \{y_{ik}\}_{k=1,\ldots,N_i})$, network T_θ and v_ϕ, learning rate α, β, extrapolation parameters γ, δ
2: **Initialize:** $\theta^0, \phi^0, p^0, q^0$
3: **for** each number of training iteration t **do**
4: $\quad \theta^{t+1} \leftarrow \theta^t - \alpha \nabla_\theta \widehat{\mathcal{L}}(\theta^t, \phi^t, p^t, q^t)$
5: $\quad \phi^{t+1} \leftarrow \phi^t + \beta \nabla_\phi \widehat{\mathcal{L}}(\theta^{t+1}, \phi^t, p^t, q^t)$
6: $\quad p^{t+1} \leftarrow p^t + \gamma(\theta^{t+1} - p^t)$
7: $\quad q^{t+1} \leftarrow q^t + \delta(\phi^{t+1} - q^t)$
8: **end for**

The LDW-CPS dataset, constructed by LaLonde (1986); Dehejia & Wahba (1999), is widely used in studies of Average Treatment Effects. This dataset contains 16177 samples from the real-world experiment and the Current Population Survey. Each covariate x_i includes the individual's eight attributes: two earnings measure (`earning 74`, `earning 75`), two indicators for ethnicity (`black`, `hispanic`), marital status (`married`), age (`age`), two education measures (`education`, `nodegree`), and a binary variable for whether this individual receives treatment (`is_treated`). The response y_i is `earning 78`, this individual's yearly income in 1978. This dataset has limited samples for each covariate, shown in Figure 1. To achieve better testing evaluation and model selection, we choose covariates based on their frequency for the test and validation sets. Specifically, covariates with a frequency higher than 30 are included in the test set, while those with a frequency higher than 20 but less than or equal to 30 are included in the validation set. According to this split strategy, the training set contains 15,255 samples corresponding to 13,745 covariates, the validation set contains 488 samples corresponding to 19 covariates, and the test set contains 434 samples corresponding to 20 covariates.

We also apply frameworks to a practical simulator called the Esophageal Cancer Markov Chain Simulation Model (ECM).[1] This simulator enables us to simulate patients' quality-adjusted life years (QALYs) from their current treatment until death. The approach to generate data for training and testing follows the authors in Hong et al. (2023). Each covariate x_i consists of five attributes which are `Barrett`, `aspirinEffect`, `statinEffect`, `drugIndex`, and `initialAge`. The response feature for this dataset is `QALY` (quality-adjusted life years). Because we have access to the simulator, we can sample infinite responses for each covariate. We follow the dataset generation in Hong et al. (2023) in which they use the simulator to uniformly generate 2×10^4 different covariates for the train set, and each covariate has one response. The validation set has 200 covariates that are different from those in

---

[1] Code is publicly available at https://simopt.github.io/ECSim

the train set, while the test set also has 200 covariates that are different from both the train and validation sets. Each covariate has 10^4 responses.

Baselines. We compare our method against state-of-the-art baselines, including CWGAN (Athey et al., 2021), WGAN-GP (Gulrajani et al., 2017), MGAN (Baptista et al., 2020), and CDSB (Shi et al., 2022). We keep the architecture of $T_\theta(x, U)$ and v_ϕ simple with seven fully connected layers with ReLU activation.

We denote D_{train}, D_{val}, and D_{test} as the set of covariates in the training, validation, and test sets, respectively. While the covariates in the training set are used for training the model parameters, the ones in the validation test are utilized to select hyperparameters. For each distinct $x \in D_{\text{test}}$, the set of actual respective responses is denoted as $Y_{GT}(x)$, while we use GENTLE, CWGAN, WGAN-GP, MGAN, and CDSB to generate $K = 10^4$ approximating points for $Y_{GT}(x)$, denoted as $Y_{\text{GENTLE}}(x)$, $Y_{\text{CWGAN}}(x)$, $Y_{\text{WGAN-GP}}(x)$, $Y_{\text{MGAN}}(x)$, $Y_{\text{CDSB}}(x)$ respectively. We then compute the average *Wasserstein distance* and *Kolmogorov-Smirnov statistic* (Kolmogorov, 1933) over D_{test} between $Y_{GT}(x)$ and the outputs $Y_{\text{GENTLE}}(x)$, $Y_{\text{CWGAN}}(x)$, $Y_{\text{WGAN-GP}}(x)$, $Y_{\text{MGAN}}(x)$, and $Y_{\text{CDSB}}(x)$, as shown in Table 1.

Table 1. Wasserstein distance and K-S statistic on generated $\hat{Y}(x)$ using different methods vs. corresponding ground truth. The symbol $\pm$ represents the standard deviation.

DATASET	METHOD	WD($\downarrow$)	KS($\downarrow$)
LDW-CPS	GENTLE	**3767.62** (**$\pm$ 1337.56**)	**0.48 $\pm$ 0.05**
	CWGAN	13371.99 ($\pm$ 3637.66)	0.88 $\pm$ 0.09
	MGAN	4096.75 ($\pm$ 1004.5)	0.91 $\pm$ 0.19
	WGAN-GP	7828.66 ($\pm$ 2306.00)	0.99 $\pm$ 0.02
	CDSB	4970.66 ($\pm$ 1094.74)	0.56 $\pm$ 0.14
ECM	GENTLE	**0.97 $\pm$ 0.14**	**0.21 $\pm$ 0.02**
	CWGAN	4.42 $\pm$ 2.01	0.94 $\pm$ 0.05
	MGAN	1.09 $\pm$ 0.59	0.55 $\pm$ 0.13
	WGAN-GP	1.59 $\pm$ 0.77	0.66 $\pm$ 0.27
	CDSB	2.04 $\pm$ 0.43	0.47 $\pm$ 0.12

Results. We plot the qualitative results of estimated densities for different covariates X in Figure 3 and 4. Overall, synthetic observations generated by GENTLE exhibit a strong resemblance to the ground truth distributions in both datasets, while the ones generated by baselines deviate significantly from the data-generating distributions. For each covariate, the distributions generated by CWGAN seem to collapse into the mean of the training data distribution, while those of MGAN and WGAN-GP tend to shrink

towards the mean of the conditional data distribution of the covariate. Although CDSB does not suffer from this collapse tendency, it still fails to match the data-generating distributions accurately. This is consistent with the quantitative results in Table 1, which show that the data generated by GENTLE have not only smaller Wasserstein distance (WD) and Kolmogorov-Smirnov (KS) values but also smaller standard deviations. This suggests GENTLE's superior performance compared to the other baselines.

5.1. Ablation Study on Training Loss Terms

We explore the impact of removing specific components from the loss function (12) by considering two ablation versions. The first version omits the smoothing term in the loss function $\mathcal{L}$ by setting $r_1 = r_2 = 0$, while the second version omits the regularization term by setting $\lambda = 0$. When we exclude the regularization term, the model generates uniformly distributed responses in the LDW-CPS dataset, as illustrated in Figure 6. We hypothesize that the absence of $\text{Reg}(\theta)$ compromises the model's ability to enforce the Lipschitz condition between the proximity in the covariates and responses space. This absence significantly hampers the method's performance in the dataset with an unbalanced observation of covariates like LDW-CPS. This problem does not exist in the ECM dataset which covariates are uniformly drawn from the simulation, so the performance of this version in ECM just slightly decreases. On the other hand, removing the smoothing term leads the model to get trapped in bad local optima, learning only simple two-peak patterns in both two datasets, as depicted in Figure 7 and Figure 8. This becomes more apparent when examining Table 2, where the KS statistic of this version surpasses that of the main method. This improvement may be attributed to the model's inclination to match high-density regions exclusively, lacking overall generalization. In summary, the results in Table 2 highlight that both ablation versions exhibit significantly worse performance than the official method, particularly regarding Wasserstein distance. This underscores the importance of the regularization and smoothing terms in maintaining the model's robustness and enhancing its overall performance.

6. Conclusions

This paper introduces a neural generative model for conditional distributions, whose parameters can be learned from limited data. Our method demonstrates remarkable efficacy in generating distributions over response spaces from given inputs. This is exemplified by successfully applying conditional dataset synthesis tasks on LDW-CPS and ECM datasets, where our framework exhibits superior performances.

An exciting extension of the current work could involve

Table 2. Wasserstein distance and K-S statistic on generated $\hat{Y}(x)$ using different ablation versions of GENTLE. The version GENTLE w/o SMO does not use the smoothing term while GENTLE w/o REG does not use the regularization term. The symbol $\pm$ represents the standard deviation.

DATASET	METHOD	WD($\downarrow$)	KS($\downarrow$)
LDW-CPS	GENTLE	**3767.62** (**$\pm$ 1337.56**)	0.48 ± 0.05
	GENTLE W/O REG	13625.83 (± 562.10)	0.57 ± 0.04
	GENTLE W/O SMO	6075.02 (± 1340.75)	**0.41 ± 0.04**
ECM	GENTLE	**0.97 ± 0.14**	**0.21 ± 0.02**
	GENTLE W/O REG	1.13 ± 0.47	0.29 ± 0.04
	GENTLE W/O SMO	1.80 ± 1.28	0.30 ± 0.05

exploring the application of our proposed method to dynamic or time-series data. Incorporating temporal dependencies could enhance the model's ability to capture evolving conditional distributions over time, enabling more accurate predictions in dynamic environments. Additionally, extending the setting of our approach beyond one-dimensional responses could further broaden its scope and impact. Finally, exploring interpretability techniques to elucidate the underlying mechanisms driving the learned conditional distributions could enhance understanding of the model's predictions in real-world decision-making scenarios.

Impact Statement

Our paper presents an advancement in the field of machine learning and also has potential societal implications. By improving the understanding and capabilities of conditional distribution modeling, our framework can contribute to various real-world applications. For instance, it can aid in personalized treatment planning in healthcare by modeling conditional distributions of patient responses to different therapies. In finance, it can enhance risk assessment by predicting conditional distributions of asset returns under varying market conditions. By enabling more accurate and nuanced predictions in diverse domains, our research can positively impact decision-making processes, resource allocation, and, ultimately, the well-being of individuals and society.
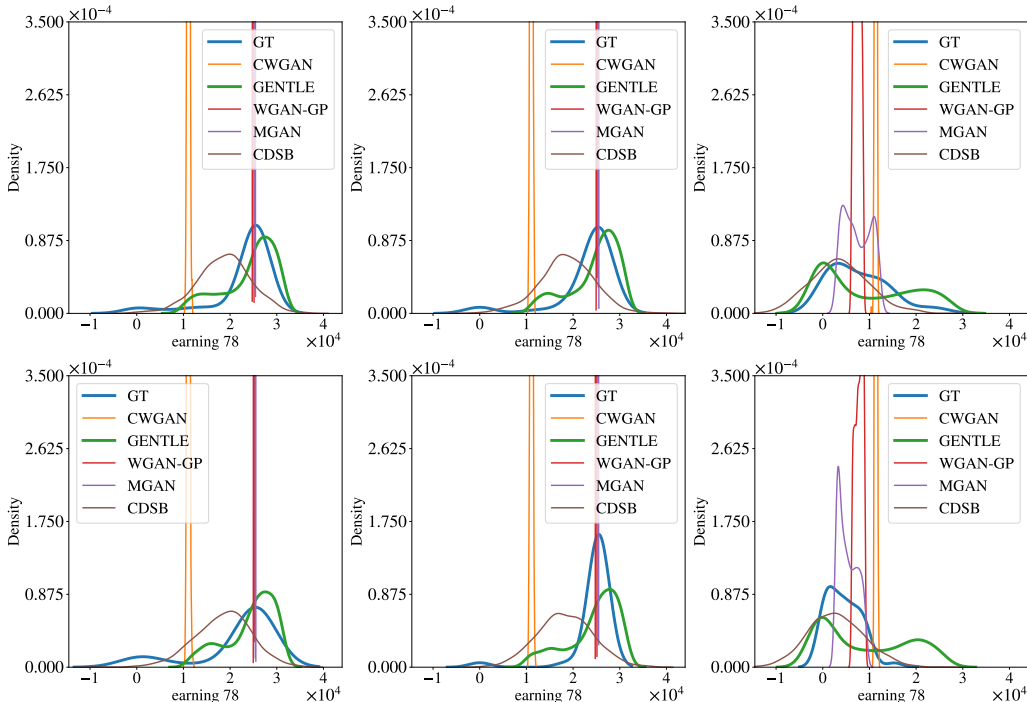
8

Figure 3. The qualitative results of methods on the LDW-CPS dataset. The density graph 'GT', **'GENTLE' (ours)**, 'CWGAN', 'WGAN-GP', 'MGAN', 'CDSB' is constructed by applying kernel density estimate (KDE) on $Y_{GT}(x)$, $Y_{GENTLE}(x)$, $Y_{CWGAN}(x)$, $Y_{WGAN\text{-}GP}(x)$, $Y_{MGAN}(x)$, $Y_{CDSB}(x)$, respectively. Each subfigure is for a different value x on the test set. We observe that GENTLE can produce the true distribution much more effectively than baselines.
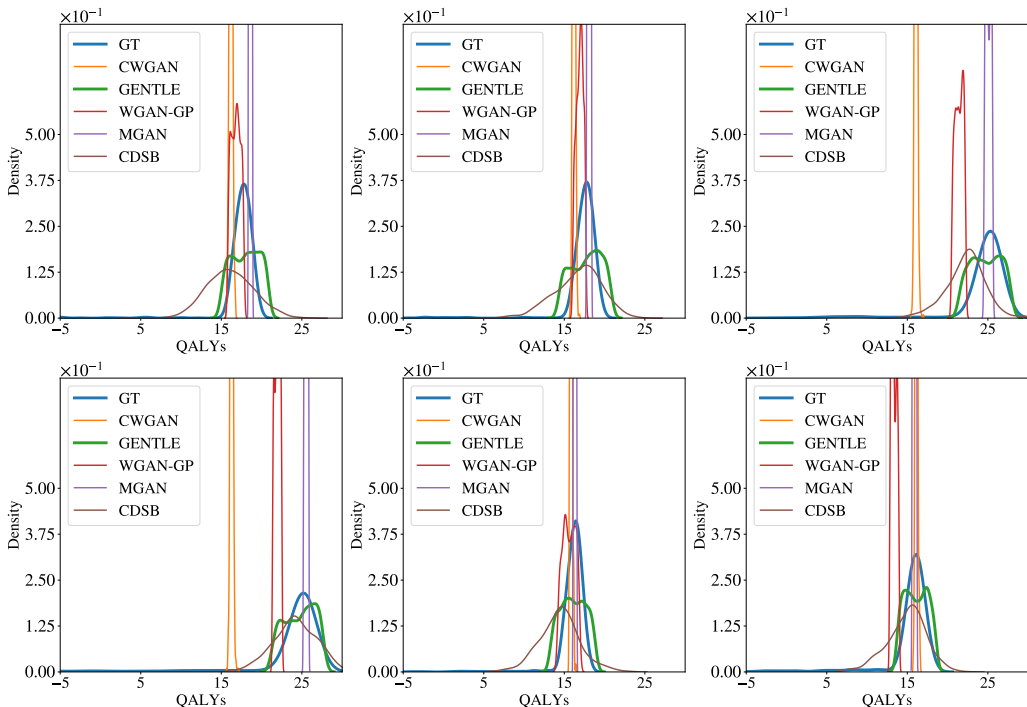

Figure 4. The qualitative results of methods on the ECM dataset. The density graph 'GT', **'GENTLE' (ours)**, 'CWGAN', 'WGAN-GP', 'MGAN', 'CDSB' is constructed by applying kernel density estimate (KDE) on $Y_{GT}(x)$, $Y_{GENTLE}(x)$, $Y_{CWGAN}(x)$, $Y_{WGAN\text{-}GP}(x)$, $Y_{MGAN}(x)$, $Y_{CDSB}(x)$, respectively. Each subfigure is for a different value x on the test set. We observe that GENTLE can produce the true distribution much more effectively than baselines.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223. PMLR, 2017.

Athey, S. and Imbens, G. W. Machine learning for estimating heterogeneous causal effects. Technical report, 2015.

Athey, S., Imbens, G. W., Metzger, J., and Munro, E. Using Wasserstein generative adversarial networks for the design of Monte Carlo simulations. *Journal of Econometrics*, 2021.

Austin, T. Exchangeable random measures. In *Annales de l'IHP Probabilités et Statistiques*, volume 51, pp. 842–861, 2015.

Baptista, R., Hosseini, B., Kovachki, N. B., and Marzouk, Y. Conditional sampling with monotone GANs: From generative models to likelihood-free inference. *arXiv preprint arXiv:2006.06755*, 2020.

Bobkov, S. and Ledoux, M. *One-Dimensional Empirical Measures, Order Statistics, and Kantorovich Transport distances*, volume 261. American Mathematical Society, 2019.

Branco, P., Torgo, L., and Ribeiro, R. P. SMOGN: A preprocessing approach for imbalanced regression. In *First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, pp. 36–50. PMLR, 2017.

Bunne, C., Krause, A., and Cuturi, M. Supervised training of conditional Monge maps. *Advances in Neural Information Processing Systems*, 35:6859–6872, 2022.

Bunne, C., Stark, S. G., Gut, G., Del Castillo, J. S., Levesque, M., Lehmann, K.-V., Pelkmans, L., Krause, A., and Rätsch, G. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Choi, J., Choi, J., and Kang, M. Generative modeling through the semi-dual formulation of unbalanced optimal transport. *arXiv preprint arXiv:2305.14777*, 2023.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.

Dehejia, R. H. and Wahba, S. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

Díaz, I. Machine learning in the estimation of causal effects: Targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*, 21(2):353–358, 2020.

Esteban-Pérez, A. and Morales, J. M. Distributionally robust stochastic programs with side information based on trimmings. *Mathematical Programming*, 195(1):1069–1105, 2022.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 3440–3448, 2016.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of Wasserstein GANs. *Advances in Neural Information Processing Systems*, 30, 2017.

Gushchin, N., Kolesov, A., Korotin, A., Vetrov, D., and Burnaev, E. Entropic neural optimal transport via diffusion processes. *arXiv preprint arXiv:2211.01156*, 2022.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

Hong, L. J., Hou, Y., Zhang, Q., and Zhang, X. Learning to simulate: Generative metamodeling via quantile regression. *arXiv preprint arXiv:2311.17797*, 2023.

Jacob, L., She, J., Almahairi, A., Rajeswar, S., and Courville, A. W2GAN: Recovering an optimal transport map with a GAN, 2019. URL https://openreview.net/forum?id=BJx9f305t7.

Jiang, G., Hong, L. J., and Nelson, B. L. Online risk monitoring using offline simulation. *INFORMS Journal on Computing*, 32(2):356–375, 2020.

Kolmogorov, A. Sulla determinazione empirica di una legge di distribuzione. *Giorn Dell'inst Ital Degli Att*, 4:89–91, 1933.

Korotin, A., Selikhanovych, D., and Burnaev, E. Neural optimal transport. In *The Eleventh International Conference on Learning Representations*, 2022.

Kruskal, J. B. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1):48–50, 1956.

Kullback, S. Probability densities with given marginals. *The Annals of Mathematical Statistics*, 39(4):1236–1243, 1968.

LaLonde, R. J. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pp. 604–620, 1986.

Léonard, C. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems-Series A*, 34 (4):1533–1574, 2014.

Lin, T., Jin, C., and Jordan, M. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pp. 6083–6093. PMLR, 2020.

Makkuva, A., Taghvaei, A., Oh, S., and Lee, J. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pp. 6672–6681. PMLR, 2020.

Nguyen, V. A., Zhang, F., Blanchet, J., Delage, E., and Ye, Y. Distributionally robust local non-parametric conditional estimation. *Advances in Neural Information Processing Systems*, 33:15232–15242, 2020a.

Nguyen, V. A., Zhang, X., Blanchet, J., and Georghiou, A. Distributionally robust parametric maximum likelihood estimation. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7922–7932, 2020b.

Nguyen, V. A., Zhang, F., Wang, S., Blanchet, J., Delage, E., and Ye, Y. Robustifying conditional portfolio decisions via optimal transport. *Operations Research*, 2024.

Nietert, S., Goldfeld, Z., and Kato, K. Smooth p-Wasserstein distance: Structure, empirical approximation, and statistical applications. In *International Conference on Machine Learning*, pp. 8172–8183. PMLR, 2021.

Nutz, M. Introduction to Entropic Optimal Transport. *Lecture notes, Columbia University*, 2021.

Parikh, H., Varjao, C., Xu, L., and Tchetgen, E. T. Validating causal inference methods. In *International Conference on Machine Learning*, pp. 17346–17358. PMLR, 2022.

Sadana, U., Chenreddy, A., Delage, E., Forel, A., Frejinger, E., and Vidal, T. A survey of contextual optimization methods for decision-making under uncertainty. *European Journal of Operational Research*, 2024.

Scott, D. W. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

Shen, H., Hong, L. J., and Zhang, X. Ranking and selection with covariates for personalized decision making. *INFORMS Journal on Computing*, 33(4):1500–1519, 2021.

Shi, Y., De Bortoli, V., Deligiannidis, G., and Doucet, A. Conditional simulation using diffusion Schrödinger bridges. In *Uncertainty in Artificial Intelligence*, pp. 1792–1802. PMLR, 2022.

Shrivastava, A. and Tuzel, O. Learning conditional error model for simulated time-series data. In *CVPR Workshops*, pp. 91–94, 2019.

Silverman, B. W. *Density Estimation for Statistics and Data Analysis*. Routledge, 2018.

Torgo, L., Ribeiro, R. P., Pfahringer, B., and Branco, P. SMOTE for regression. In *Portuguese Conference on Artificial Intelligence*, pp. 378–389. Springer, 2013.

van Breugel, B., Seedat, N., Imrie, F., and van der Schaar, M. Can you rely on your model evaluation? Improving model evaluation with synthetic test data. *Advances in Neural Information Processing Systems*, 36, 2024.

Villani, C. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

Zheng, T., Zhu, L., So, A. M.-C., Blanchet, J., and Li, J. Universal gradient descent ascent method for nonconvex-nonconcave minimax optimization. In *Advances in Neural Information Processing Systems*, 2023.

11

A. Additional Experimental Details

A.1. Parameter Tuning

We heuristically fix the learning rates $\alpha = \beta = 0.001$ as a reasonable value for almost every simple multilayer perception model. For other parameters, we conducted a grid search on minimizing the WD metric of the validation set. We found that the best combination for our parameters in the LDW-CPS dataset is KDE bandwidth of $h = 0.3$, $\varepsilon = 1.0$, $\lambda = 0.4$, $r_1 = 3.0$, $r_2 = 2.0$, $\gamma = 0.5$, $\delta = 0.7$. Regarding the ECM dataset, the best combination is $h = 0.2$, $\varepsilon = 1.0$, $\lambda = 0.4$, $r_1 = 3.0$, $r_2 = 2.0$, $\gamma = 0.5$, $\delta = 0.7$. We fix these parameter values for all our experiments in the main paper.

A.2. Comparison between GENTLE and Baselines on Other Metrics

In this section, we report two additional metrics, including Mean Squared Error (MSE) and R^2 score, to evaluate the performance of GENTLE and baselines. Tables 3 and 4 represent the average results of algorithms across ten different seed numbers in terms of these two metrics. We observe that GENTLE outperforms baselines on both MSE and R^2 criteria. Although MGAN has comparable results to GENTLE in terms of MSE and R^2, the qualitative results shown in Figures 3 and 4 and the statistical metrics of WD and KS in Tables 1 reveals its trend of collapsing to the mean of the ground truth distribution.

Table 3. MSE and R^2 on generated $\hat{Y}(x)$ using different methods vs. corresponding ground truth on the LDW-CPS dataset. The symbol $\pm$ represents the standard deviation.

| Method	MSE ($\times 10^6$)($\downarrow$)	R^2($\uparrow$)
GENTLE	**3.70 $\pm$ 0.73**	**0.92 $\pm$ 0.05**
CWGAN	108.43 $\pm$ 42.93	-1.83 $\pm$ 0.13
MGAN	3.82 $\pm$ 0.71	**0.92 $\pm$ 0.09**
WGAN-GP	59.64 $\pm$ 9.65	-0.59 $\pm$ 0.07
CDSB	8.54 $\pm$ 1.21	0.77 $\pm$ 0.08

Table 4. MSE and R^2 on generated $\hat{Y}(x)$ using different methods vs. corresponding ground truth on the ECM dataset. The symbol $\pm$ represents the standard deviation.

| Method	MSE ($\downarrow$)	R^2($\uparrow$)
GENTLE	**0.62 $\pm$ 0.32**	**0.96 $\pm$ 0.05**
CWGAN	24.30 $\pm$ 4.36	0.00 $\pm$ 0.13
MGAN	1.24 $\pm$ 0.22	0.91 $\pm$ 0.10
WGAN-GP	2.11 $\pm$ 0.33	0.86 $\pm$ 0.08
CDSB	2.01 $\pm$ 0.39	0.89 $\pm$ 0.04

A.3. Additional Experimental Results

In Figure 5, we plot the value of $T_\theta(x, U)$ for an increasing grid of $U \sim \mathcal{U}(0, 1)$. We observe that the output of T_θ also increased with U, empirically confirming that T_θ is an increasing function, which conforms with the fact that $T_\theta(x, U)$ is an approximation of an increasing function of U.
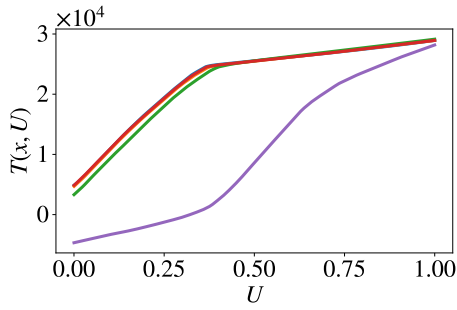
(a) GENTLE trained on LDW-CPS
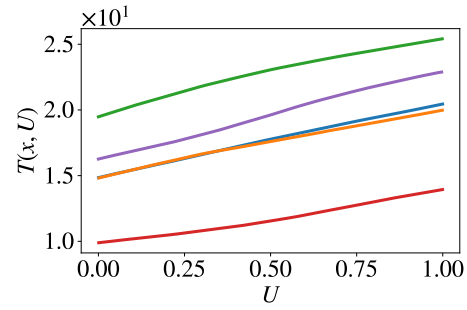
(b) GENTLE trained on ECM

Figure 5. Empirical evidence for the monotonically increasing property of the learned network $T_\theta(x, U)$ in the variable U.
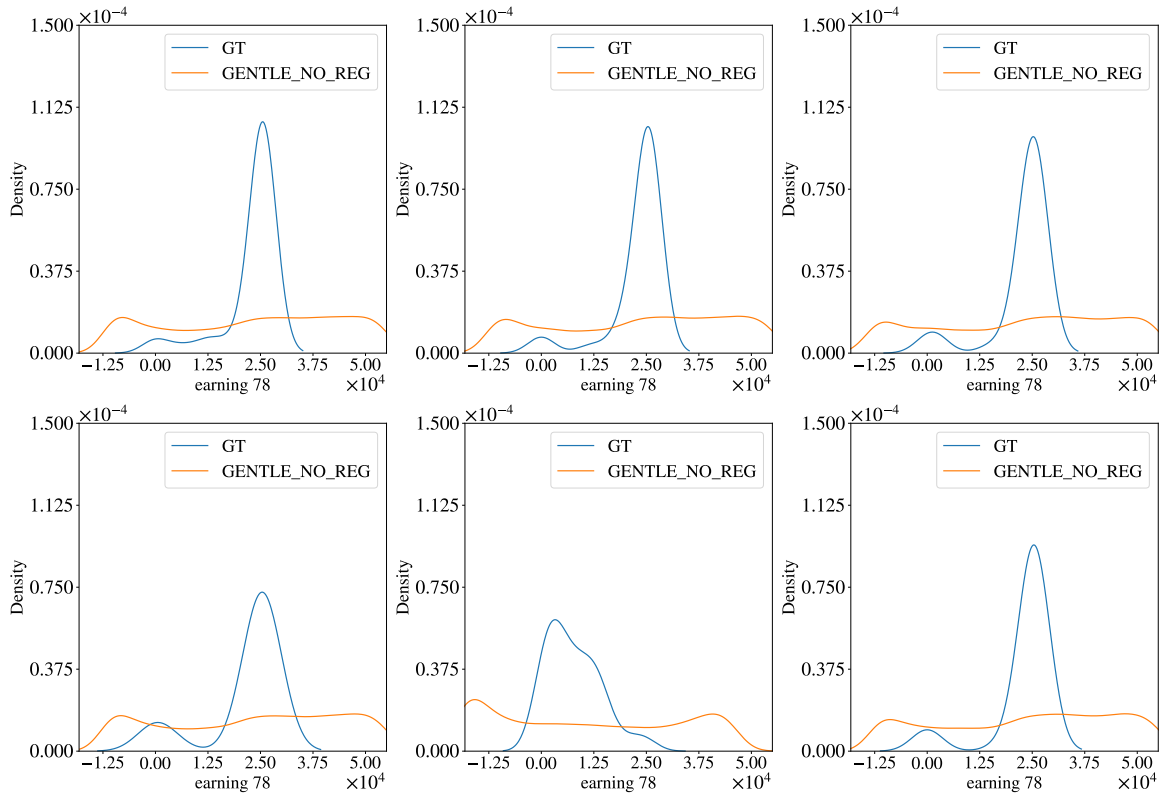

Figure 6. Qualitative result of the GENTLE version without regularization term on the LDW-CPS dataset. Each subfigure represents the results of a different covariate value in the test set. The learned distribution of this version tends to span over the range of the response term uniformly.
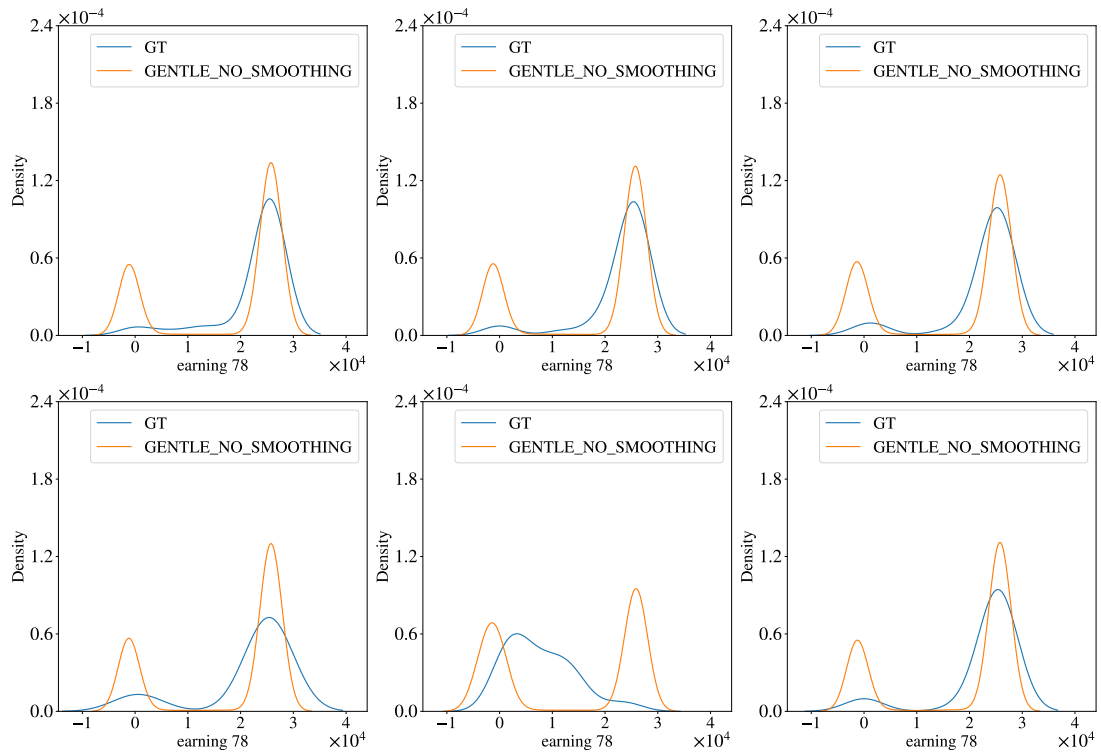
13

Figure 7. Qualitative result of the GENTLE version without smoothing term on the LDW-CPS dataset. Each subfigure represents the results of a different covariate value in the test set.
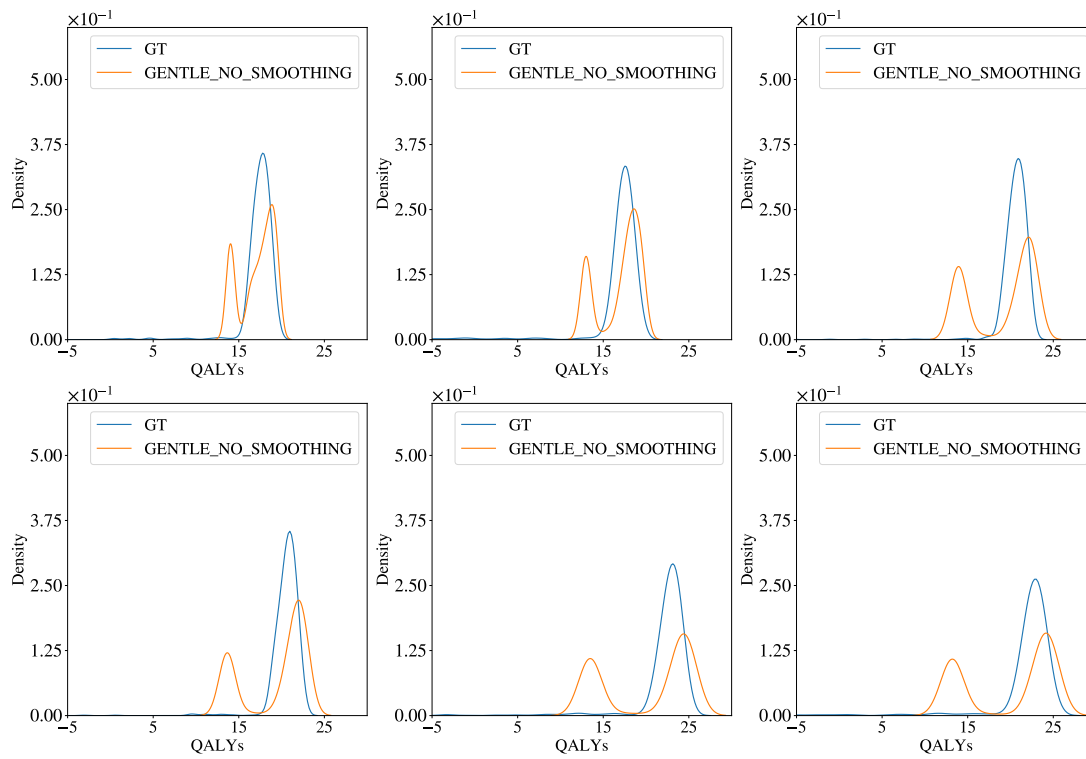
Figure 8. Qualitative result of the GENTLE version without smoothing term on the ECM dataset. Each subfigure represents the results of a different covariate value in the test set.