Generative Conditional Distributions by Neural (Entropic) Optimal Transport

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

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.

1. Introduction

The conditional distribution of a response variable given the covariate information is a key quantity for numerous tasks in data science and artificial intelligence. 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 et al., 2015). Having access to the conditional probability distribution of a response also leads to an estimate of the treatment effect without explicit, additional experiments.

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

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

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 also 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 covariates 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 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.

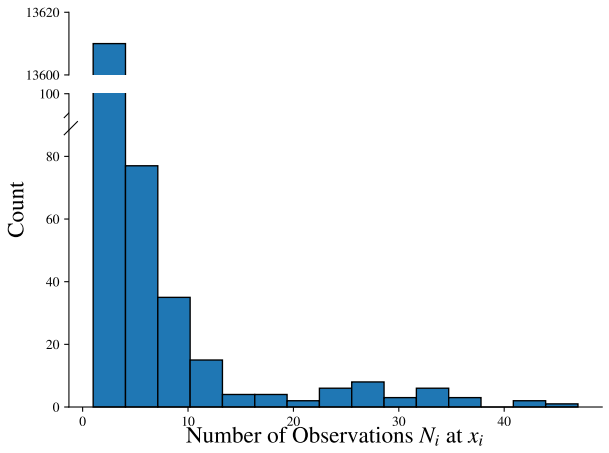


Figure 1. Histogram count of Number of observed responses for distinct covariate values in the LDW dataset (LaLonde, 1986). This dataset has over 13,000 covariates with only one response.

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 $\mathcal{D} = \{(x_i, \{y_{i,k}\}_{k=1,\dots,N_i}) \in \mathbb{R}^d \times \mathbb{R}$. Without any loss of generality, we suppose that the covariates x_i are distinct, and for each x_i, $\{y_{ik}\}_{k=1,\dots,N_i}$ 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 expectations or quantiles, especially under a small training sample size.

Contributions. We introduce GENTLE, a generative neural transport learning 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.

Notations. Throughout, 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 (Makkuva et al., 2020; Bunne et al., 2022; 2023; Jacob et al., 2018); 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.

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 et al. 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

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

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

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

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:

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

Cuturi (2013) showed that (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, (3) can also be interpreted as a specific instance of the Schrödinger bridge problem, which has a rich history in physics. For an in-depth 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)

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

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

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

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*

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

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:

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

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

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

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 $F^{-1}_{\mu_X}$ can be unbounded, and the magnitude of $F^{-1}_{\mu_X}$ 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

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

4.2. Regularization Measure

Minimizing the $\text{Fit}(\theta)$ 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 the covariate distance and the Wasserstein distance of the conditional distributions

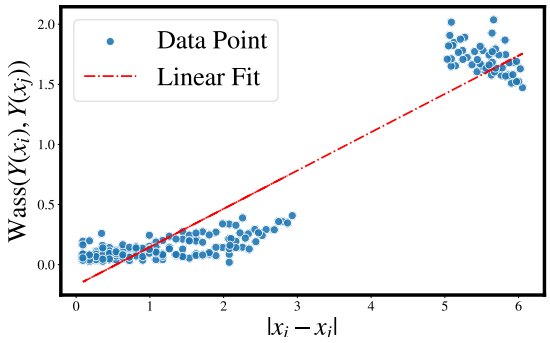


Figure 2. Positive correlation between the Wasserstein distance between $Y(x_i)$ and $Y(x_j)$ and the covariate distance $\|x_i - x_j\|$. Only covariates x with more than 18 observations are selected.

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 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:

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

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

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

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

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

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 $v_\phi^\varepsilon(x_i, y)$

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

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:

$$\mathrm{Reg}(\theta) \stackrel{\text{def.}}{=} \max_{\phi} \ \mathcal{R}(\theta, \phi), \tag{10a}$$

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

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

In this way, $\mathrm{Reg}(\theta)$ 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 $\mathcal{E}$.

Remark 4.2. In (8), we use the unregularized OT distance as a discrepancy function for the fit term $\mathrm{Fit}(\theta)$ 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 $\mathrm{Reg}(\theta)$, where we need to compare two distributions that are both outputs of the generative network T_θ. Using the unregularized OT distance for the $\mathrm{Reg}(\theta)$ 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:

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

Naively optimizing (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

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

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 cross-validation 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 grid of parameters are given in the Appendix.

Algorithm 1 Empirical loss evaluation

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 $\varepsilon, \lambda, r_1, r_2$
Output: loss estimate $\widehat{\mathcal{L}}(\theta, \phi, p, q)$
Initialize: take a batch of B training nodes $\{x_i^{(b)}\}_{b\in[B]}$
for each $(x_i^{(b)}, x_j^{(b)}) \in \mathcal{E}$ **do**
  Sample $U^{(b)} \sim \mathcal{U}[0,1]$
  for $m = 1, \cdots, M$ **do**
    Sample $U^{(b,m)} \sim \mathcal{U}[0,1]$
  end for
end for
$\widehat{\text{Fit}}(\theta) \leftarrow \frac{1}{B}\sum_b \left[U^{(b)} - \hat{F}_{x_i^{(b)}}(T_\theta(x_i^{(b)}, U^{(b)}))\right]_2^2$ (following (8))
Compute following (10):

$$\hat{\mathcal{R}}(\theta,\phi) \leftarrow \frac{1}{M\times B}\sum_{(x_i^{(b)}, x_j^{(b)}, U^{(b,m)})} \left[v_\phi^\varepsilon(x_i^{(b)}, T_\theta(x_j^{(b)}, U^{(b,m)})) + v_\phi(x_i^{(b)}, T_\theta(x_i^{(b)}, U^{(b,m)}))\right]$$

Compute following (12):

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

Algorithm 2 Training of the Minimax loss (12)

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
Initialize: $\theta^0, \phi^0, p^0, q^0$
for each number of training iteration t **do**
  $\theta^{t+1} \leftarrow \theta^t - \alpha\nabla_\theta\widehat{\mathcal{L}}(\theta^t, \phi^t, p^t, q^t)$
  $\phi^{t+1} \leftarrow \phi^t + \beta\nabla_\phi\widehat{\mathcal{L}}(\theta^{t+1}, \phi^t, p^t, q^t)$
  $p^{t+1} \leftarrow p^t + \gamma(\theta^{t+1} - p^t)$
  $q^{t+1} \leftarrow q^t + \delta(\phi^{t+1} - q^t)$
end for

5. Experiments

To benchmark our proposed method GENTLE, we use the following two datasets along with the same preprocessing in Athey et al. (2021):

- 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 sample consists of nine attributes, including two earnings measures, two indicators for ethnicity and marital status, and two education measures, age, and a binary variable for whether this individual receives treatment. Each covariate x_i includes eight attributes, respectively, `is_treated`, `age`, `education`, `black`, `hispanic`, `married`, `nodegree`, `earning 74`, `earning 75`. The response y_i is `earning 78`, this individual's yearly income in 1978.

- 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).

We compare our method with CWGAN (Athey et al., 2021). We keep the architecture of $T_\theta(x, U)$ and v_ϕ simple with seven fully connected layers with ReLU activation.

We choose the top 0.1% covariates with the highest frequency as the test dataset D_{test}, while the others are in the training dataset. For each distinct $x \in D_{\text{test}}$, the set of actual respective responses is denoted as $Y_{GT}(x)$, while we use CWGAN and GENTLE to generate $K = 10^5$ approximating points for $Y_{\text{GT}}(x)$, denoted as $Y_{\text{GENTLE}}(x)$, and $Y_{\text{CWGAN}}(x)$ respectively. We then compute the average *Wasserstein distance* and *Kolmogorov-Smirnov statistic* (Kolmogorov, 1933) over D_{test} between $Y_{\text{GENTLE}}(x)$, $Y_{\text{CWGAN}}(x)$ and $Y_{\text{GT}}(x)$ 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(↓)	KS(↓)
LDW-CPS	GENTLE	**3767.62 (± 1337.56)**	**0.48 ± 0.05**
	CWGAN	13371.99 (± 3637.66)	0.88 ± 0.09
ECM	GENTLE	**0.97 ± 0.14**	**0.21 ± 0.02**
	CWGAN	4.42 ± 2.011	0.94 ± 0.05

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

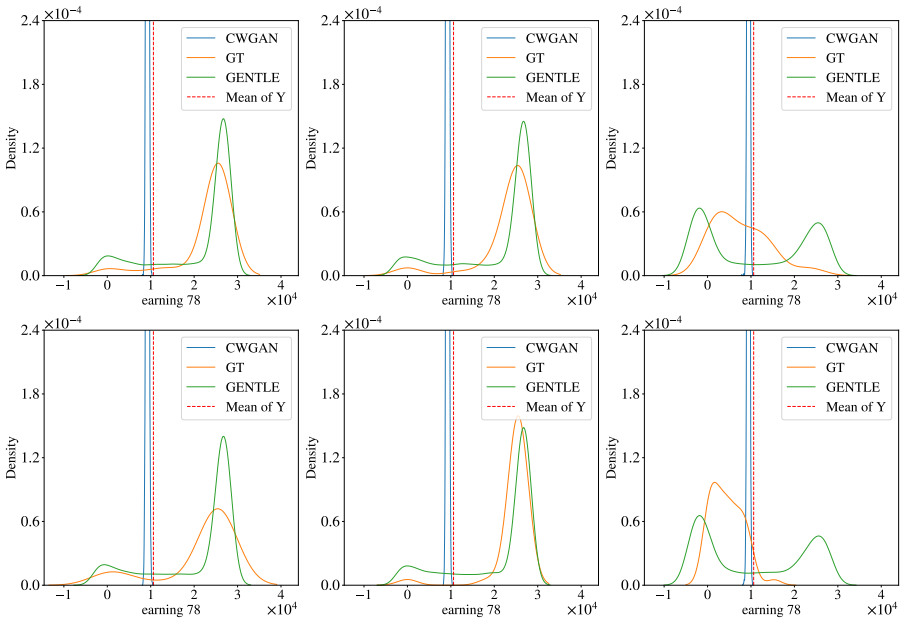


Figure 3. The qualitative results of methods on the LDW-CPS dataset. The density graph 'GT', **'GENTLE' (ours)**, and 'CWGAN' is constructed by applying kernel density estimate (KDE) on $Y_{\text{GT}}(x)$, $Y_{\text{GENTLE}}(x)$, and $Y_{\text{CWGAN}}(x)$, respectively. The vertical line 'Mean of Y' denotes the average value of Y across the training dataset. We observe that GENTLE can produce the true distribution much more effectively than CWGAN, which seems to suffer from the mode collapse into the mean value of Y over the training dataset.

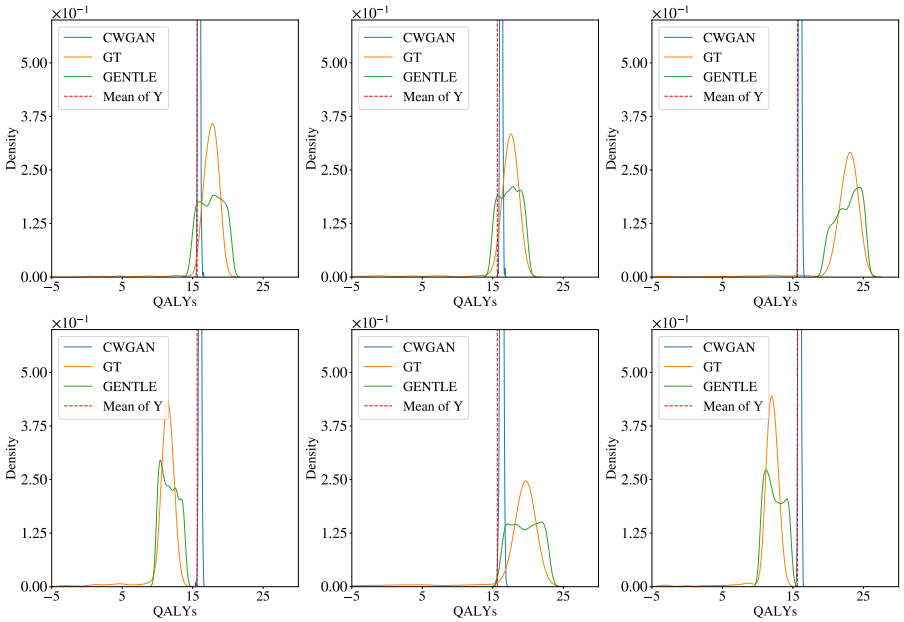


Figure 4. The qualitative results of methods on the ECM dataset. The density graph 'GT', **'GENTLE' (ours)**, and 'CWGAN' is constructed by applying kernel density estimate (KDE) on $Y_{\text{GT}}(x)$, $Y_{\text{GENTLE}}(x)$, and $Y_{\text{CWGAN}}(x)$, respectively. The vertical line 'Mean of Y' denotes the average value of Y across the training dataset. We observe that GENTLE can produce the true distribution much more effectively than CWGAN, which seems to suffer from the mode collapse into the mean value of Y over the training dataset.

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 CWGAN deviate significantly from the data-generating distributions. This is in agreement with quantitative results from Table 1 that not only the data generated by CWGAN have smaller WD and KS values, but they also have smaller standard deviations. This suggests GENTLE's superior performance compared to CWGAN's.

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 $\mathrm{Reg}(\theta)$ 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 phenomenon 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 phenomenon 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.

Table 2. Wasserstein distance and K-S statistic on generated $\hat{Y}(x)$ using different ablation versions of GENTLE. The symbol $\pm$ represents the standard deviation.

Dataset	Method	WD(↓)	KS(↓)
LDW-CPS	GENTLE	**3767.62** (± **1337.56**)	0.48 ± 0.05
	GENTLE without regularization	13625.83 (± 562.10)	0.57 ± 0.04
	GENTLE without smoothing	6075.02 (± 1340.75)	**0.41 ± 0.04**
ECM	GENTLE	**0.97 ± 0.14**	**0.21 ± 0.02**
	GENTLE without regularization	1.13 ± 0.47	0.29 ± 0.04
	GENTLE without smoothing	1.80 ± 1.28	0.30 ± 0.05

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 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 Statements

Our paper on Conditional Distribution Learning via Neural Entropic Optimal Transport presents an advancement in the field of machine learning and also 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.

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., Imbens, G. W., et al. 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.

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 pre-processing 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.

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.

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*, NIPS'16, pp. 3440–3448, 2016.

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. 2018.

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 lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91, 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.

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.

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.

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

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. et al. *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.

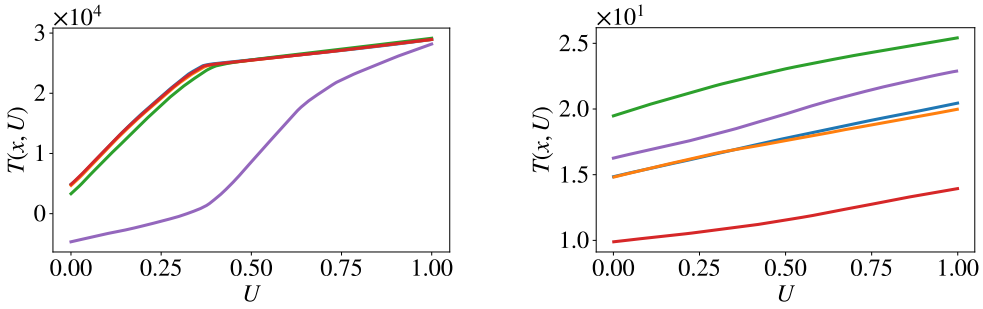
A. Additional Experimental Details

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 procedure and found that the best combination for our parameters is KDE bandwidth of 0.1, $\varepsilon = 1.0$, $\lambda = 0.3$, $\gamma = 1.0$, $\delta = 2.0$. We fix these parameter values for all our experiments in the main paper.

A.2. Additional Experimental Results

In Figure 5, we plot the value of $T_\theta(x, U)$ for an increasing grid of $U \sim \mathcal{U}(0, 1)$. It is clearly observed 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 increasingly monotonic attribute

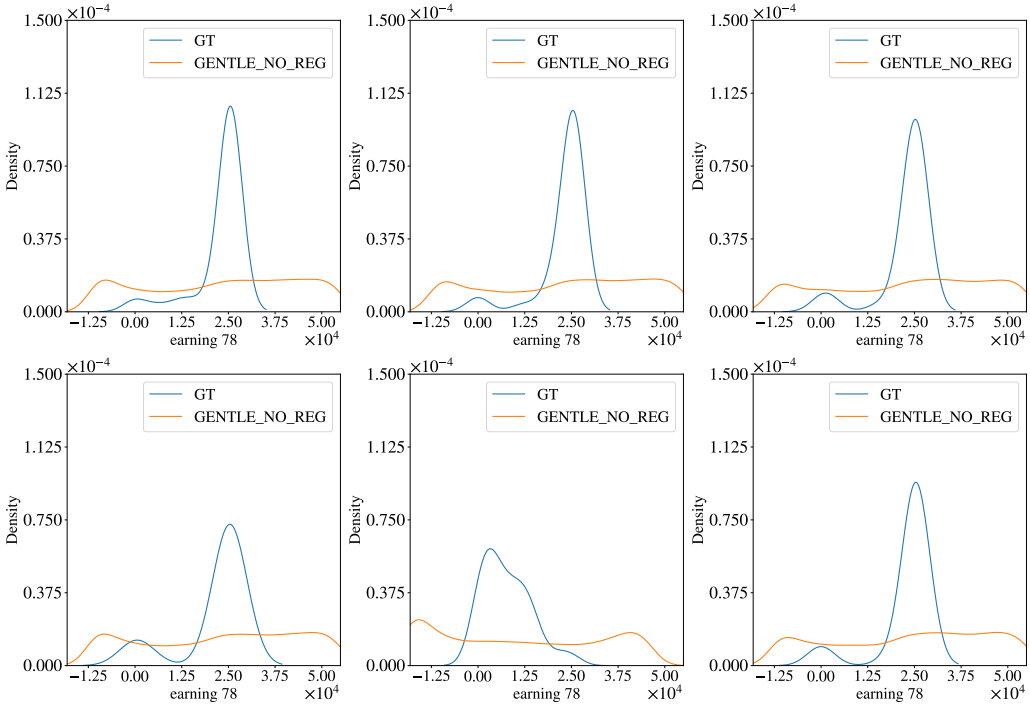


Figure 6. Qualitative result of the GENTLE version without regularization term on the LDW-CPS dataset. The learned distribution of this version tends to uniformly span over the range of the response term.

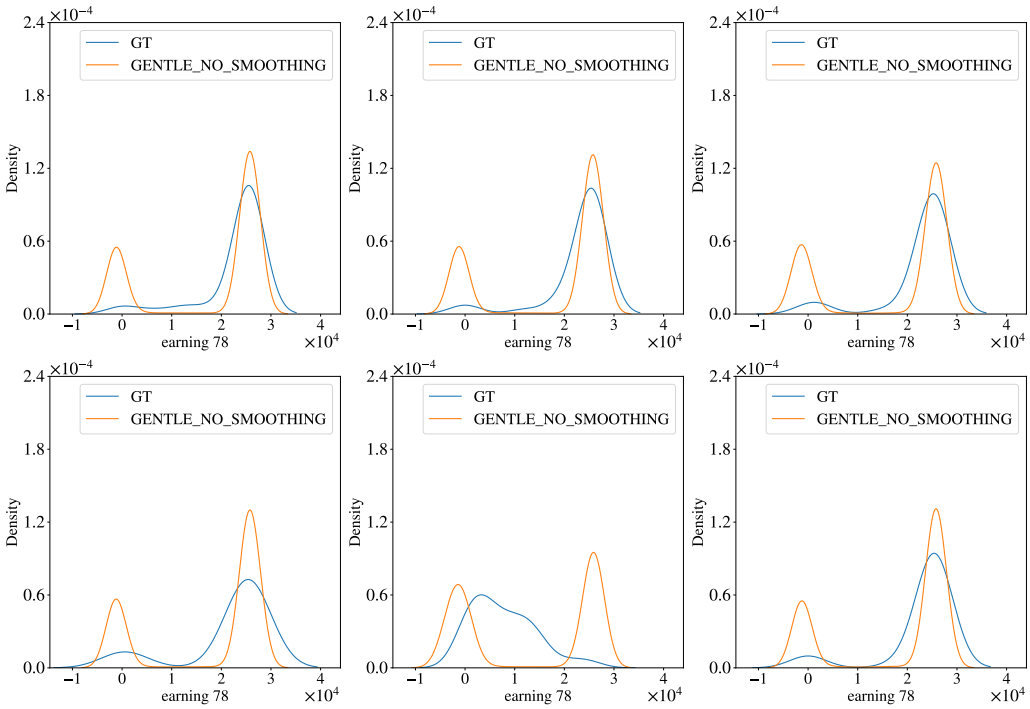


Figure 7. Qualitative result of the GENTLE version without smoothing term on the LDW-CPS dataset.

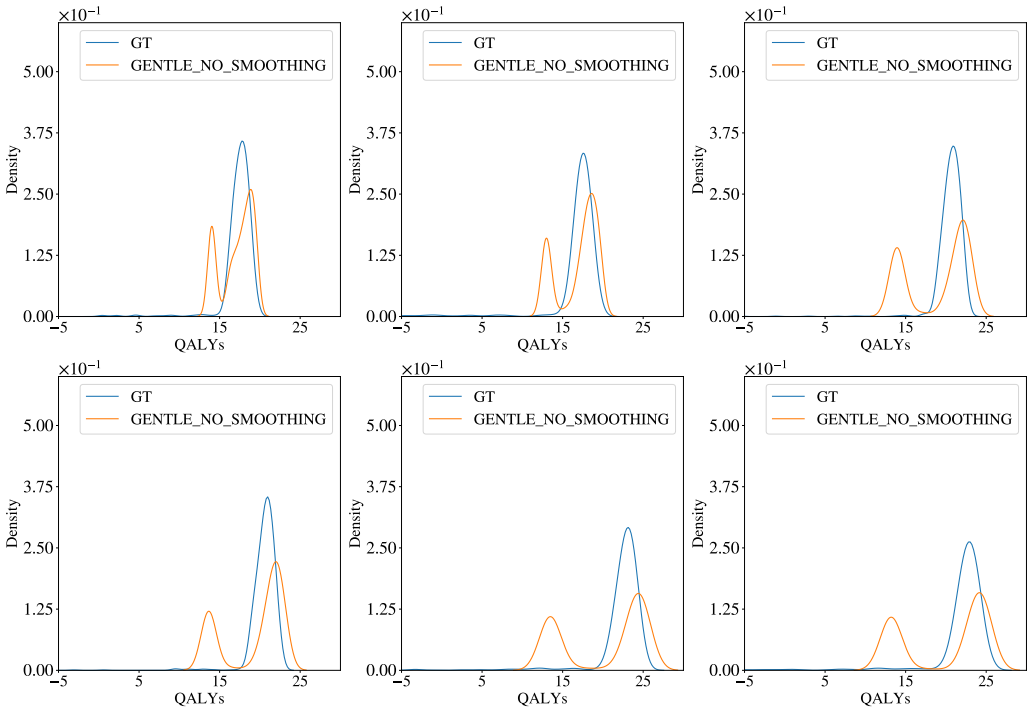


Figure 8. Qualitative result of the GENTLE version without smoothing term on the ECM dataset.