Continuous Vector Quantile Regression

Sanketh Vedula [*] [1 2] **Irene Tallini** [*] [1 3] **Aviv A. Rosenberg** [1 2] **Marco Pegoraro** [1 3]
Emanuele Rodolá [3] **Yaniv Romano** [1] **Alex M. Bronstein** [1 2]

Abstract

Vector quantile regression (VQR) estimates the conditional vector quantile function (CVQF), a fundamental quantity which fully represents the conditional distribution of $\mathbf{Y}|\mathbf{X}$. VQR is formulated as an optimal transport (OT) problem between a uniform $\mathbf{U} \sim \mu$ and the target $(\mathbf{X}, \mathbf{Y}) \sim \nu$, the solution of which is a unique transport map, co-monotonic with $\mathbf{U}$. Recently non linear VQR (NL-VQR) has been proposed to estimate support non-linear CVQFs, together with fast solvers which enabled the use of this tool in practical applications. Despite its utility, the scalability and estimation quality of NL-VQR is limited due to a discretization of the OT problem onto a grid of quantile levels. We propose a novel *continuous* formulation and parametrization of VQR using partial input-convex neural networks (PICNNs). Our approach allows for accurate, scalable, differentiable and invertible estimation of non-linear CVQFs. We further demonstrate, theoretically and experimentally, how continuous CVQFs can be used for general statistical inference tasks: estimation of likelihoods, CDFs, confidence sets, coverage, sampling, and more. This work is an important step towards unlocking the full potential of VQR.

1. Introduction

Quantile regression (QR) (Koenker & Bassett, 1978) is a widely-known approach for modeling the conditional quantiles of a target variable Y given covariates $\mathbf{X}$. Despite its power and usefulness, QR is inherently limited in that it deals only with scalar-valued random variables Y. This limitation stems from the fact that QR minimizes the *pinball loss*, which is not defined for multivariate inputs. Moreover, even the notion of quantiles is not trivial to define for high dimensional variables.

Seminal works from Carlier et al. (2016) and Chernozhukov et al. (2017) introduced the notion of *vector quantiles functions*, defining them through the extension of two properties of scalar quantile functions, namely co-monotonicity (Eq. 1) and strong representation (Eq. 2), to the vector case:

$$\left(Q_{\mathbf{Y}}(\boldsymbol{u}) - Q_{\mathbf{Y}}(\boldsymbol{u'})\right)^{\top} (\boldsymbol{u} - \boldsymbol{u'}) \geq 0, \; \forall \, \boldsymbol{u}, \boldsymbol{u'} \in [0,1]^d \quad (1)$$

$$\mathbf{Y} = Q_{\mathbf{Y}}(\mathbf{U}), \; \mathbf{U} \sim \mathbb{U}[0,1]^d \quad (2)$$

where $\mathbf{Y}$ is a d-dimensional variable, and $Q_{\mathbf{Y}} : [0,1]^d \mapsto \mathbb{R}^d$ is its *vector quantile function* (VQF). Carlier et al. (2016) also proposed *vector quantile regression* (VQR), an extension of QR to vector-valued targets which estimates the *conditional vector quantile function* (CVQF) $Q_{\mathbf{Y}|\mathbf{X}}$ from samples drawn from $P_{(\mathbf{X},\mathbf{Y})}$. VQR is formulated as an optimal transport (OT) problem between the uniform base distribution of $\mathbf{U}$ and the target conditional distribution of $\mathbf{Y}|\mathbf{X}$, for which the resulting transport map is the CVQF, $Q_{\mathbf{Y}|\mathbf{X}}$. In order to account for the conditioning on $\mathbf{X}$, the additional mean independence constraint $\mathbb{E}\left[\mathbf{U}|\mathbf{X}\right] = \mathbb{E}\left[\mathbf{X}\right]$ is added to the OT problem, making it challenging to solve. By modelling the CVQF as linear in $\mathbf{X}$, i.e. $Q_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{u}; \boldsymbol{x}) = \boldsymbol{B}(\boldsymbol{u})^{\top}\boldsymbol{x} + \boldsymbol{a}(\boldsymbol{u})$, the primal of the VQR OT problem can be naïvely solved as a linear program. In a recent work, Rosenberg et al. (2023) demonstrated that this approach is intractable for real-world datasets, and proposed fast solvers for VQR by solving an entropic-regularized dual formulation of the OT problem which is amenable to gradient-based optimization. They further proposed *nonlinear VQR* (NL-VQR), an extension which overcomes the linearity assumption about the CVQF.

Despite the advantages of the fast nonlinear approach of Rosenberg et al. (2023), they nevertheless solve a *discrete* OT problem, based on the original formulation of Carlier et al. (2016). We argue that using a discrete formulation has significant drawbacks which hinder the full potential of VQR. First, the discrete solvers estimate the CVQF on a grid of T quantile levels per dimension. The number of optimization variables is thus proportional to T^d which hinders scaling to data beyond a few dimensions. Moreover, the CVQF is obtained through the gradient of a convex po-

[*] Equal contribution [1] Department of Computer Science, Technion [2] Sibylla, UK [3] Sapienza, University of Rome. Correspondence to: Sanketh Vedula <sanketh@campus.technion.ac.il>, Irene Tallini <tallini@di.uniroma1.it>.

Workshop on New Frontiers in Learning, Control, and Dynamical Systems at the International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA, 2023.

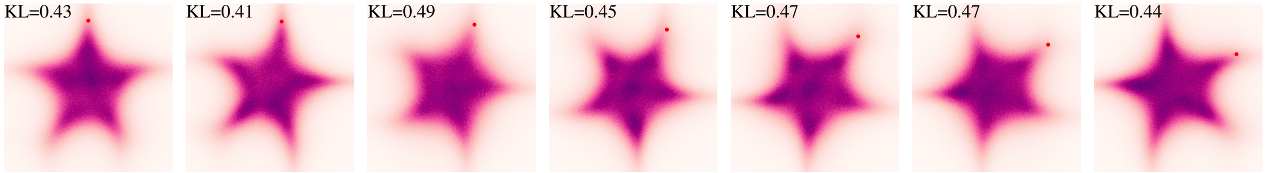

Figure 1. **Continuous VQR accurately estimates the conditional likelihoods on the rotating stars dataset.** Depicted from left to right are likelihoods of a star shaped distribution $f_{\mathbf{Y}|\mathrm{X}}$, where X defines the rotation angle. Overlaid on each plot is the KL divergence of the estimated likelihood with respect to the corresponding ground truth likelihood. Red marker on the plots is added to aid visualization.

tential function, which in this case is also discretized on the grid of quantile levels. In the discrete case, this gradient is obtained by first-order differences, making it prone to numerical errors which degrade performance, especially when T is small. Second, as discussed in Section 3, the CVQF, the inverse CVQF, and their derivatives are meaningful statistical quantities. Estimating the convex potentials discretely therefore hinders the accurate calculation of their higher-order derivatives and therefore limits the utility of VQR for general statistical inference. Third, NL-VQR cannot exploit the underlying structure of the target variable $\mathbf{Y}$. This becomes especially important as d increases, since high dimensional data, such as images, often lie on lower-dimensional manifolds embedded in $\mathbb{R}^d$. Lastly, another drawback of the NL-VQR approach is that by solving the entropic-regularized relaxed dual VQR OT problem co-monotonicity of the CVQF is promoted but not enforced, and since the discrete formulation parametrizes the convex potentials as simple parameter vectors, it is not possible to make them convex by construction.

Contributions. To address the aforementioned limitations, our first contribution is a novel continuous formulation of VQR, together with a convex parametrization based on partial input convex neural networks (PICNNs). Instead of *explicitly* evaluating the convex potentials over a discrete grid, we propose to learn an *implicit* representation of the convex potentials by modeling them via PICNNs (Amos et al., 2017). By constraining our model space to input-convex functions, we guarantee the co-monotonicity of the estimated CVQF. Furthermore, one can encode any inductive bias into the PICNN architecture to leverage the structure present in $\mathbf{Y}$. Our approach is builds on recent progress in neural optimal transport, where ICNNs are employed to model transport maps, estimating e.g. Wasserstein distances between high-dimensional distributions (Makkuva et al., 2020; Korotin et al., 2021b; 2022). To the best of our knowledge, ours is the first approach to estimate continuous CVQFs. Our second contribution is to leverage continuous CVQFs for general statistical inference on arbitrarily data distributions. We derive numerous statistical quantities as a function of their CVQF, namely: *exact* conditional and unconditional likelihoods, cumulative distribution functions (CDFs), confidence sets with their areas, and statistical coverage in terms of the VQF and CVQF respectively. Estimating statistical coverage of a high-dimentional variable is notably difficult, as it requires testing the insideness of a point in a high-dimensional set; however, we show that using the inverse CVQF the estimation becomes trivial. As far as we know, ours is the first work to estimate the aforementioned statistical quantities directly from the (conditional) vector quantile function. Finally, we employ challenging synthetic data experiments to evaluate the performance of the estimated quantile functions in terms of sampling quality, likelihood estimation, and the statistical validity of the confidence sets both in modeling conditional and unconditional distributions. We demonstrate that continuous VQR is more accurate and scales more effectively to higher dimensions than its discrete counterparts.

Notation. Throughout, Y, $\mathbf{X}$ denote random variables and vectors, respectively; deterministic scalars, vectors and matrices are denoted as y, $\boldsymbol{x}$, and $\boldsymbol{A}$. $|\boldsymbol{A}|$ denotes the matrix determinant. $P_{(\mathbf{X},\mathrm{Y})}$ denotes the joint distribution of the $\mathbf{X}$ and Y. $\mathbf{1}_N$ denotes an N-dimensional vector of ones, $\odot$ denotes the elementwise product, and $\mathbb{I}\left\{\cdot\right\}$ is an indicator. We denote by N the number of samples, d the dimension of the target variable, k the dimension of the covariates, and T the number of vector quantile levels per target dimension. $Q_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{u};\boldsymbol{x})$ is the CVQF of the variable $\mathbf{Y}|\mathbf{X}$ evaluated at the vector quantile level $\boldsymbol{u}$ for $\mathbf{X}=\boldsymbol{x}$. The Jacobian of a function f is denoted with $\boldsymbol{J}_f$.

2. Continuous VQR

We refer the reader to Appendix A which provides a gentle introduction to quantile regression and its optimal transport formulation, and presents the discrete formulations of VQR and NL-VQR on which we build our approach.

The semi-discrete dual formulation of NL-VQR (Rosenberg et al. (2023), appendix eq. 10) suffers from three drawbacks: (i) the number of dual variables in $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ grows exponentially with the dimension of the target variable d; (ii) approximating the the convex potential's gradient via finite-differences is leads to inaccuracies which get worse when coarsening the quantile level grid, as required in higher dimensions; (iii) it is unclear how to compute the inverse CVQF, $Q^{-1}_{\mathbf{Y}|\mathbf{X}}$, since the required derivative is w.r.t. $\boldsymbol{y}$, whose discretization onto a fixed grid is infeasible.

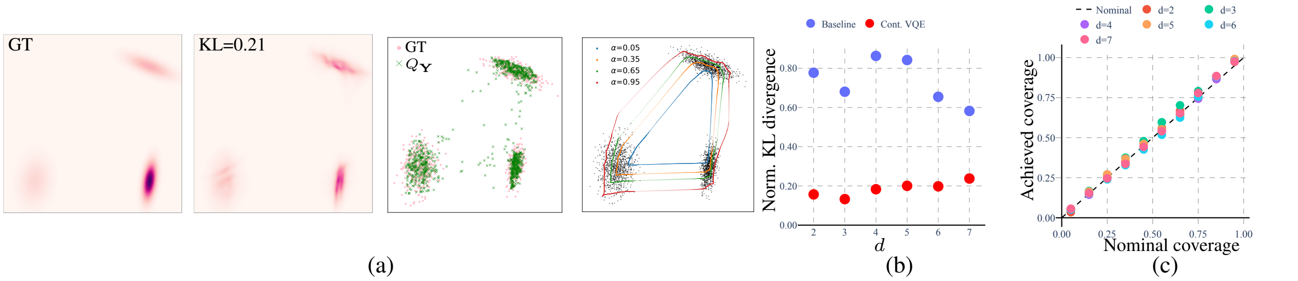

Figure 2. **Continuous VQE accurately captures the sampling process, likelihoods, and confidence sets.** **(a)** The first and second panels depict the ground truth and estimated likelihoods, respectively, for $d = 2$. The third panel presents samples drawn from the quantile function and the ground truth distribution. The fourth panel presents the α-confidence sets overlaid on the groundtruth distribution. **(b)** The x-axis depicts the dimension d of the target variable. The y-axis presents the normalized KL divergence between the estimated and the ground truth likelihoods. **(c)** Plots nominal vs achieved coverage for different α-confidence sets.

To overcome these limitations, we implicitly represent the CVQF as a gradient of a partial input-convex neural network $f(\boldsymbol{u};\boldsymbol{x}) : [0,1]^d \times \mathbb{R}^k \to \mathbb{R}$, that is convex in $\boldsymbol{u}$ and nonlinear in $\boldsymbol{x}$ (see Appendix A for further details). We propose solving an alternative formulation of the semi-discrete dual formulation (Eq. 10), given as follows

$$\begin{aligned}\min_f \;& \sum_{i=1}^{T} \mu_i \sum_{j=1}^{N} \nu_j f(\boldsymbol{u}_i;\boldsymbol{x}_j) \\ &+ \sum_{j=1}^{N} \nu_j \max_{\boldsymbol{u}\in[0,1]^d} \left\{ \left(\boldsymbol{u}^\top \boldsymbol{y}_j - f(\boldsymbol{u};\boldsymbol{x}_j)\right)\right\}.\end{aligned} \tag{3}$$

Note that the inner optimization problem calculates the convex conjugate of the potential $f(\boldsymbol{u};\boldsymbol{x})$ for a given $(\boldsymbol{x}_j, \boldsymbol{y}_j)$. We solve the inner problem by evaluating the maximum of over samples $\{\boldsymbol{u}_i\}_{i=1}^{T} \sim \mathbb{U}[0,1]^d$ and approximate the maximum with a soft-maximum. In our experiments we use an affine formulation for the partial ICNN, given by $f(\boldsymbol{u};\boldsymbol{x}) = \beta(\boldsymbol{u})^\top g(\boldsymbol{x}) + \varphi(\boldsymbol{u})$, where $\beta : \mathbb{R}^d \to \mathbb{R}^k$, $\varphi : \mathbb{R}^d \to \mathbb{R}$ are ICNNs, and $g : \mathbb{R}^{k'} \to \mathbb{R}^k_+$ is an MLP with a ReLU activation over the outputs to ensure convexity of $f(\boldsymbol{u};\boldsymbol{x})$ with respect to $\boldsymbol{u}$. The conditional vector quantile function is then obtained by

$$Q_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{u};\boldsymbol{x}) = \nabla_{\boldsymbol{u}} \left[f(\boldsymbol{u};\boldsymbol{x}) \right]. \tag{4}$$

Note that in the case of vector quantile estimation (VQE), $f : [0,1]^d \to \mathbb{R}$ can simply be an ICNN, then the VQF becomes $\nabla_{\boldsymbol{u}}[f(\boldsymbol{u})]$.

3. Deriving statistical quantities from CVQFs

In what follows, we derive different statistical quantities as a function of the VQF $Q_{\mathbf{Y}}$. The conditional analogues can be simply obtained by substituting $Q_{\mathbf{Y}}$ with $Q_{\mathbf{Y}|\mathbf{X}}$.

Vector rank function. The vector rank function is defined as the inverse of the VQF. It is derived through computing the convex conjugate:

$$\psi(\boldsymbol{y}) = \max_{\boldsymbol{u}\in[0,1]^d} \left\{ \boldsymbol{u}^\top \boldsymbol{y} - f(\boldsymbol{u}) \right\},$$

and taking its gradient with respect to $\boldsymbol{y}$:

$$R_{\mathbf{Y}}(\boldsymbol{y}) := Q_{\mathbf{Y}}^{-1}(\boldsymbol{y}) = \nabla_{\boldsymbol{y}} \left[\psi(\boldsymbol{y}) \right].$$

Cumulative distribution function. The cumulative distribution function for multivalued functions is defined as:

$$F_{\mathbf{Y}}(\boldsymbol{y}) = \mathbb{P}\left[\mathrm{Y}_1 \le y_1, \ldots, \mathrm{Y}_d \le y_d \right].$$

From the strong representation property (2), we have that, $\forall i \in \{1,\ldots,d\}$, $\mathrm{Y_i} = Q^i_{\mathbf{Y}}(\mathbf{U})$. Applying the change of variables formula for multivariable integration yields:

$$F_{\mathbf{Y}}(\boldsymbol{y}) = \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_d} \left| \boldsymbol{J}_{Q_{\mathbf{Y}}^{-1}}(\boldsymbol{y}') \right| dy'_1 \ldots dy'_d, \tag{5}$$

Likelihood. The likelihood of a multivariate continuous random variable $\mathbf{Y}$, $f_{\mathbf{Y}}(\boldsymbol{y})$ is defined as:

$$F_{\mathbf{Y}}(\boldsymbol{y}) = \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_d} f_{\mathbf{Y}}(\boldsymbol{y}') dy'_1 \ldots dy'_d$$

From Equation (5), we have the explicit expression for the likelihood:

$$f_{\mathbf{Y}}(\boldsymbol{y}) = \left| \boldsymbol{J}_{Q_{\mathbf{Y}}^{-1}}(\boldsymbol{y}) \right|.$$

Confidence sets. A valid α-confidence set for a random vector $\mathbf{Y}$ is defined as a set $\mathcal{C}^{\mathbf{Y}}_\alpha$ for which the property $\mathbb{P}\left[\mathbf{Y} \in \mathcal{C}^{\mathbf{Y}}_\alpha\right] = \alpha$ holds.

In order to construct valid confidence sets for the target distribution, we first build valid confidence intervals $\mathcal{C}^{\mathbf{U}}_\alpha$ in the base distribution, namely hypercubes centered in $(0.5,\ldots,0.5)$ and with side length equal to $\sqrt[d]{\alpha}$. Then we can construct a target α-confidence interval as:

$$\mathcal{C}^{\mathbf{Y}}_\alpha = Q_{\mathbf{Y}}\left(\mathcal{C}^{\mathbf{U}}_\alpha\right).$$

Validity is given by the following equalities:

$$\mathbb{P}\left[\mathbf{Y} \in \mathcal{C}^{\mathbf{Y}}_\alpha\right] = \mathbb{P}\left[Q_{\mathbf{Y}}\left(\mathbf{U}\right) \in Q_{\mathbf{Y}}\left(\mathcal{C}^{\mathbf{U}}_\alpha\right)\right] = \mathbb{P}\left[\mathbf{U} \in \mathcal{C}^{\mathbf{U}}_\alpha\right] = \alpha,$$

where the second equality is a consequence of the monotonicity property of quantile functions.

Area of confidence sets. The area of the α-confidence set

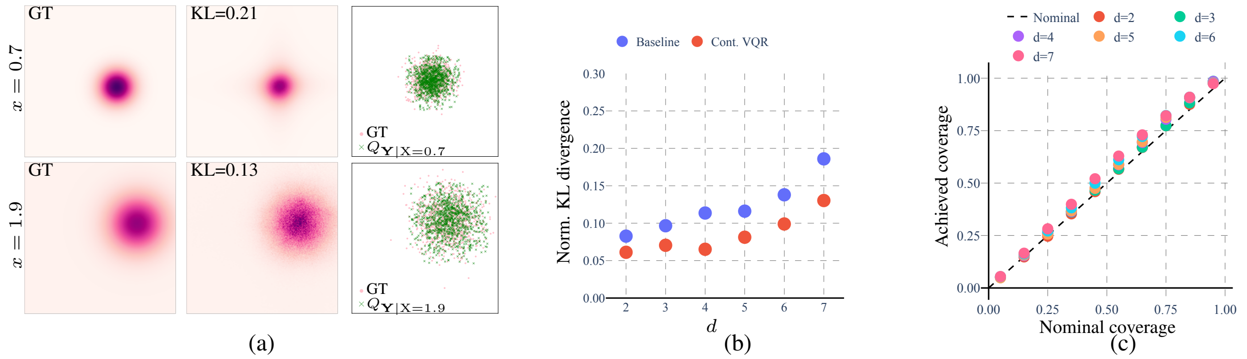

Figure 3. **Continuous VQR accurately captures the conditional sampling process, likelihoods, and confidence sets.** **(a)** The left and center panels depict the ground truth and estimated likelihoods, respectively, for $d = 2$. The right panel presents samples drawn from the conditional quantile function and the ground truth distribution. Top and bottom rows correspond to different conditioning values, namely $x = 0.7$ and $x = 1.9$. **(b)** The x-axis depicts the dimension d of the target variable. The y-axis presents the KL divergence between the estimated likelihoods and the ground truth likelihoods, normalized by the entropy of the respective ground truth distributions. **(c)** Plots expected vs achieved marginal coverage values for α-confidence sets for different values of α, averaged over the conditioning value x.

is obtained as:

$$\mathcal{A}(\mathcal{C}_\alpha^{\mathbf{Y}}) = \int_{\mathcal{C}_\alpha^{\mathbf{Y}}} d\boldsymbol{y} = \int_{\mathcal{C}_\alpha^{\mathbf{U}}} |\boldsymbol{J}_{Q_{\mathbf{Y}}}(\boldsymbol{u})|\, d\boldsymbol{u}$$

Inclusion in a confidence set. Testing if a point is contained in a given α-confidence set is a challenging problem in high-dimensional domains, as it requires testing if a point is contained into a general polygon. However, using the continuous VQF, we can test if a sample $\boldsymbol{y}$ is contained in $\mathcal{C}_\alpha^{\mathbf{Y}}$ by checking if $Q_{\mathbf{Y}}^{-1}(\boldsymbol{y})$ is contained in $\mathcal{C}_\alpha^{\mathbf{U}}$, which is a much simpler problem, i.e.,

$$\mathbb{I}\left\{\boldsymbol{y} \in \mathcal{C}_\alpha^{\mathbf{Y}}\right\} = \mathbb{I}\left\{Q_{\mathbf{Y}}^{-1}(\boldsymbol{y}) \in \mathcal{C}_\alpha^{\mathbf{U}}\right\}. \tag{6}$$

This procedure also makes it easier to measure *statistical coverage*, the rate of inclusion in a α-confidence set, a standard metric to evaluate the quality of confidence sets.

4. Experimental Results

To evaluate the proposed continuous VQR formulation, we conduct carefully designed synthetic data experiments where groundtruth conditional likelihoods are known in closed form. We evaluate the quality of CVQF based on three criteria that capture different aspects of the estimated conditional distribution: (i) statistical validity of the confidence sets, (ii) likelihood estimation quality, and (iii) sampling quality. Appendix B presents the details of the datasets and evaluation metrics used in the experiments. We evaluate continuous VQE and VQR by varying $d \in \{2, \ldots, 7\}$. We note that the discrete VQR formulation (Rosenberg et al., 2023) is computationally infeasible for $d > 3$.

The qualitative and quantitative results for continuous VQE on multimodal Gaussian data across different ds is presented in Figure 2. Fig. 2b demonstrates that the normalized KL divergence of the estimated likelihood w.r.t. the ground truth stays small as the dimension increases indicating that the continuous approach scales reasonably well with the target dimension. Fig. 2c presents that the α-confidence sets constructed over the distribution are statistically valid, i.e., the achieved coverage matches the nominal rate. Fig. 2a further shows that the quality of estimated likelihoods and sampling is visually consistent with the ground truth.

The results of continuous VQR that evaluate the quality of estimated CVQFs are presented in Fig. 1 for the rotating stars data, and in Fig. 3 for the conditional Gaussians data. Fig 1 and Fig 3a demonstrate that the conditional likelihood estimation via the CVQF is visually accurate. This is further evidenced by the small normalized KL divergence values presented in Fig. 3b. Finally, Fig 3c presents that the marginal coverage of the conditional α-confidence sets matches the nominal rate indicating their validity.

5. Conclusions

We present the first continuous formulation of VQR by modeling the CVQF as the gradient of a partial ICNN. This allows for accurate, scalable, differentiable and invertible estimation of non-linear CVQFs. We further derived a variety of statistical quantities (CDF, likelihood, confidence sets and areas) as a function of the CVQF, and were able to estimate them by leveraging our continuous approach. Through synthetic experiments, we verified the validity of the derived quantities and demonstrated that the continuous formulation offers better estimation quality and greater scalability than its discrete counterpart.

Limitations. Our approach explicitly calculates the convex-conjugate of the potential, thus leading to a min-max optimization problem. We solved the inner problem by approximating the maximum with a finite sample approximation, the accuracy of which deteriorates as the dimensionality increases, thus posing a scalability challenge. In future works, we aim to adapt our solvers to prevent this limitation, e.g. by incorporating recent advances in neural OT (Korotin et al., 2021a;b). Notwithstanding these limitations, we believe that our continuous approach will help unlock the full potential of VQR in a wide array of applications.

Acknowledgement

S.V., A.A.R, and A.M.B. were partially supported by the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No. 863839), by the Council For Higher Education - Planning & Budgeting Committee, and by the Israeli Smart Transportation Research Center (ISTRC). I.T, M.P. and E.R were partially supported by ERC Grant no. 802554 "SPECGEO" and PRIN 2020 project no. 2020TA3K9N "LEGO.AI".

References

Amos, B., Xu, L., and Kolter, J. Z. Input convex neural networks. In *International Conference on Machine Learning*, pp. 146–155. PMLR, 2017.

Carlier, G., Chernozhukov, V., and Galichon, A. Vector quantile regression: An optimal transport approach. *Annals of Statistics*, 44(3):1165–1192, 2016. ISSN 00905364. doi: 10.1214/15-AOS1401.

Chen, Y., Shi, Y., and Zhang, B. Optimal control via neural networks: A convex approach. *arXiv preprint arXiv:1805.11835*, 2018.

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. MongeKantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223 – 256, 2017. doi: 10.1214/16-AOS1450. URL https://doi.org/10.1214/16-AOS1450.

Koenker, R. and Bassett, G. Regression Quantiles. *Econometrica*, 46(1):33, 1978. ISSN 00129682. doi: 10.2307/1913643.

Korotin, A., Egiazarian, V., Asadulaev, A., Safin, A., and Burnaev, E. Wasserstein-2 generative networks. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=bEoxzW_EXsa.

Korotin, A., Li, L., Genevay, A., Solomon, J. M., Filippov, A., and Burnaev, E. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. *Advances in Neural Information Processing Systems*, 34: 14593–14605, 2021b.

Korotin, A., Selikhanovych, D., and Burnaev, E. Neural optimal transport. *arXiv preprint arXiv:2201.12220*, 2022.

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.

Rosenberg, A. A., Vedula, S., Romano, Y., and Bronstein, A. M. Fast nonlinear vector quantile regression. *International Conference on Learning Representations (ICLR)*, 2023.

A. Background

Quantiles. The u-th quantile $Q_{\mathrm{Y}}(u)$ of a r.v. Y is defined as the smallest scalar y such that $\mathbb{P}\left[\mathrm{Y} \leq y\right] = u$.

A well known property of the quantile function (*strong representation*) is that of being the only *monotonic* function such that $\mathrm{Y} = Q_{\mathrm{Y}}(\mathrm{U})$ with probability 1. In other words, any density can be obtained by transforming a uniform random variable by the quantile function.

Quantile Regression. The goal of quantile regression (QR) is to estimate the quantiles of a variable Y *conditioned* on a vector $\mathbf{X}$, i.e., of $\mathrm{Y}|\mathbf{X}$. Assuming a linear model $Q_{\mathrm{Y}}(u) = \boldsymbol{b}_u{}^\top \boldsymbol{x} + a_u$ for the quantiles, QR amounts to solving the following optimization problem (Koenker & Bassett, 1978):

$$\min_{\boldsymbol{b},a} \mathbb{E}_{(\mathbf{X},\mathrm{Y})}\left[\rho_u(\mathrm{Y} - \boldsymbol{b}^\top \mathbf{X} - a)\right],$$

where $\rho_u(z)$, known as the *pinball loss*, is given by $\rho_u(z) = \max\{0, z\} + (u-1)z$.

Solving this problem produces an estimate of $Q_{\mathrm{Y}|\mathbf{X}}$ for a single quantile level u. In order to estimate the full *conditional quantile function* (CQF) $Q_{\mathrm{Y}|\mathbf{X}}(u)$, the problem must be solved at all levels of u with additional monotonicity constraints, the quantile function being non-decreasing in u. The CQF discretized at T quantile levels can be estimated from N samples $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N} \sim P_{(\mathbf{X},\mathrm{Y})}$ by solving:

$$\begin{aligned} &\min_{\boldsymbol{B},\boldsymbol{a}} \sum_{u} \sum_{i=1}^{N} \rho_u(y_i - \boldsymbol{b}_u{}^\top \boldsymbol{x}_i - a_u) \\ &\text{s.t. } \forall i,\ u' \geq u \implies \boldsymbol{b}_{u'}{}^\top \boldsymbol{x}_i + a_{u'} \geq \boldsymbol{b}_u{}^\top \boldsymbol{x}_i + a_u, \end{aligned} \tag{7}$$

where $\boldsymbol{B}$ and $\boldsymbol{a}$ aggregate all the $\boldsymbol{b}_u$ and a_u, respectively. We refer to Equation (7) as *simultaneous linear quantile regression* (SLQR).

This problem is undefined for a vector-valued Y, due to the inherently 1D formulation of the monotonicity constraints and of the pinball loss.

Optimal Transport Formulation. Carlier et al. (2016) showed that SLQR (7) can be equivalently formulated as an *optimal transport* (OT) problem between the target variable and the quantile levels, with an additional constraint of mean independence. Given N data samples arranged as $\boldsymbol{y} \in \mathbb{R}^N$, $\boldsymbol{X} \in \mathbb{R}^{N\times k}$, and T quantile levels denoted by $\boldsymbol{u} = \left[\frac{1}{T}, \frac{2}{T}, ..., 1\right]^\top$ we can write:

$$\begin{aligned} &\max_{\boldsymbol{\Pi}\geq 0} \boldsymbol{u}^\top \boldsymbol{\Pi} \boldsymbol{y} \\ &\text{s.t. } \boldsymbol{\Pi}^\top \mathbf{1}_T = \boldsymbol{\nu} \\ &\qquad \boldsymbol{\Pi}\mathbf{1}_N = \boldsymbol{\mu} \qquad [\boldsymbol{\varphi}] \\ &\qquad \boldsymbol{\Pi}\boldsymbol{X} = \bar{\boldsymbol{X}} \qquad [\boldsymbol{\beta}] \end{aligned} \tag{8}$$

where $\boldsymbol{\Pi}$ is the transport plan between quantile levels $\boldsymbol{u}$ and samples $(\boldsymbol{x}, \boldsymbol{y})$, with marginal constraints $\boldsymbol{\nu} = \frac{1}{N}\mathbf{1}_N$, $\boldsymbol{\mu} = \frac{1}{T}\mathbf{1}_T$ and mean-independence constraint $\bar{\boldsymbol{X}} = \frac{1}{T}\mathbf{1}_T \frac{1}{N}\mathbf{1}_N{}^\top \boldsymbol{X}$. The dual variables are $\boldsymbol{\varphi} = \boldsymbol{D}^{-\top}\boldsymbol{a}$ and $\boldsymbol{\beta} = \boldsymbol{D}^{-\top}\boldsymbol{B}$, where $\boldsymbol{D}^\top$ is a first-order finite differences matrix, and $\boldsymbol{a} \in \mathbb{R}^T$, $\boldsymbol{B} \in \mathbb{R}^{T\times k}$ contain the regression coefficients for all quantile levels.

Vector quantile regression. The optimal transport formulation has the advantage of being amenable to extension to the vector-valued case. Denote covariates and targets by $\{(\boldsymbol{x}_j, \boldsymbol{y}_j)\}_{i=1}^{N} \sim \mathbb{P}_{(\mathbf{X},\mathbf{Y})}$, and vector quantile levels $\{\boldsymbol{u}_i\}_{i=1}^{T^d}$ sampled on a uniform grid on $[0,1]^d$ with T evenly spaced points in each dimension. Assume a linear specification for the CVQF, $Q_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{u};\boldsymbol{x}) = \boldsymbol{B}(\boldsymbol{u})^\top \boldsymbol{x} + \boldsymbol{a}(\boldsymbol{u})$. Performing vector quantile regression amounts to solving the following optimal transport problem:

$$\begin{aligned} &\max_{\boldsymbol{\Pi}\geq 0} \sum_{i=1}^{T^d} \sum_{j=1}^{N} \boldsymbol{u}_i{}^\top \boldsymbol{y}_j \Pi_{i,j} \\ &\text{s.t. } \boldsymbol{\Pi}^\top \mathbf{1}_{T^d} = \boldsymbol{\nu} \qquad [\boldsymbol{\psi}] \\ &\qquad \boldsymbol{\Pi}\mathbf{1}_N = \boldsymbol{\mu} \qquad [\boldsymbol{\varphi}] \\ &\qquad \boldsymbol{\Pi}\boldsymbol{X} = \bar{\boldsymbol{X}} \qquad [\boldsymbol{\beta}] \end{aligned} \tag{9}$$

where $\mathbf{\Pi} \in \mathbb{R}^{T^d \times N}$ represents the optimal transport *plan*, and the dual variables $\boldsymbol{\varphi} \in \mathbb{R}^{T^d}$, $\boldsymbol{\beta} \in \mathbb{R}^{T^d \times k}$, $\boldsymbol{\psi} \in \mathbb{R}^N$ are the discretized convex potentials. The discrete CVQF is then computed as follows $\hat{Q}_{\mathbf{Y}|\mathbf{X}}\left(\boldsymbol{u}; \boldsymbol{x}\right) = \boldsymbol{D}^\top \left(\boldsymbol{\beta}^\top \boldsymbol{x} + \boldsymbol{\varphi}\right)$, where $\boldsymbol{D}$ is the first-order finite differences matrix; thus, the estimated CVQF satisfies the co-monotonicity (Eq. 1).

Nonlinear VQR. (Rosenberg et al., 2023) propose solving the semi-discrete dual formulation of the aforementioned primal OT problem which is amenable to gradient-based optimization. Furthermore, they proposed performing VQR in the embedding domain of $\mathbf{X}$, where the embedding function is learned jointly with the discretized potentials, the optimization problem reads as follows:

$$\min_{\boldsymbol{\psi}, \boldsymbol{\beta}, \theta} \ \boldsymbol{\psi}^\top \boldsymbol{\nu} + \operatorname{tr}\left(\boldsymbol{\beta}^\top g_\theta(\bar{\boldsymbol{X}})\right) + \sum_{i=1}^{T^d} \mu_i \max_j \left\{ \left(\boldsymbol{u}_i{}^\top \boldsymbol{y}_j - \boldsymbol{\beta}_i{}^\top g_\theta(\boldsymbol{x}_j) - \psi_j \right) \right\}. \tag{10}$$

In practice, the inner maximum is approximated by a soft-maximum. The resulting nonlinear CVQF is given as $\hat{Q}_{\mathbf{Y}|\mathbf{X}}\left(\boldsymbol{u}; \boldsymbol{x}\right) = \boldsymbol{D}^\top \left(\boldsymbol{\beta}^\top g_\theta(\boldsymbol{x}) + \boldsymbol{\varphi}\right)$.

Partial ICNNs. We model the CVQF as the gradient of a convex function $f(\boldsymbol{u}; \boldsymbol{x})$. We implement f as a partial input convex neural network (PICNN) (Amos et al., 2017). Given $(\boldsymbol{u}; \boldsymbol{x}) \in \mathbb{R}^d \times \mathbb{R}^k$, the mapping $(\boldsymbol{u}; \boldsymbol{x}) \mapsto \mathbb{R}$ is given by L-layer of feed-forward neural network following the equations for $l = 0, \ldots, L-1$:

$$\begin{aligned} \boldsymbol{x}_{l+1} =& \tilde{\sigma}_l(\tilde{W}_l \boldsymbol{x}_l + \tilde{b}_l) \\ \boldsymbol{z}_{l+1} =& \sigma_l \left(W_l^{(\boldsymbol{z})} \left(\boldsymbol{z}_l \circ [W_l^{(\boldsymbol{zx})} \boldsymbol{x}_l + b_l^{(\boldsymbol{z})}]_+ \right) + \right. \\ & \left. W_l^{(\boldsymbol{u})} \left(\boldsymbol{u} \circ (W_l^{(\boldsymbol{ux})} u_l + b_l^{(\boldsymbol{u})}) \right) + W_l^{(\boldsymbol{x})} \boldsymbol{x}_l + b_l \right) \\ f(\boldsymbol{u}; \boldsymbol{x}) =& \boldsymbol{z}_k, \ \boldsymbol{x}_0 = \boldsymbol{x} \end{aligned} \tag{11}$$

where $\boldsymbol{z}_i \in \mathbb{R}^{m_i}$ and $\boldsymbol{x}_i \in \mathbb{R}^{n_i}$ denote the hidden units for the "$\boldsymbol{u}$-path" and "$\boldsymbol{x}$-path", and where $\circ$ denotes the Hadamard product, the elementwise product between two vectors. The total set of parameters is composed by the weight matrices $(\{W_l\}, \{W_l^{(\boldsymbol{z})}\}, \{W_l^{(\boldsymbol{u})}\}, \{W_l^{(\boldsymbol{ux})}\}, \{W_l^{(\boldsymbol{x})}\})$ and the bias terms $(\{\tilde{b}_l\}, \{b_l^{(\boldsymbol{z})}\}, \{b_l^{(\boldsymbol{u})}\}, \{b_l\})$. To ensure the convexity in $\boldsymbol{u}$, Equation (11) needs to satisfy the following constrains: *(i)* all $W_{1:L}^{(\boldsymbol{z})}$ must be non-negative, *(ii)* $\tilde{\sigma}_l$ and σ_l must be convex and non-decreasing entry-wise activation functions.

The class of PICNN can represent any ICNN (see Amos et al., 2017, Proposition 2) and therefore approximate any convex function over a compact domain (see Chen et al., 2018, Theorem 1), making it a suitable parametrization of convex potential in the OT problem.

B. Datasets, Metrics, and Baselines

Datasets. In order to compare our samples and likelihoods to the ground truth, we perform evaluations on synthetic datasets for which sampling and likelihood evaluation are straightforward. For the experiments presented in the paper, we use the following two synthetic datasets, and the rotating stars dataset introduced in (Rosenberg et al., 2023).

- **Multi-modal Gaussians.** To evaluate the performance of VQE, we use multi-modal high-dimensional Gaussian data, whose density is given as follows:

  $$f_{\mathbf{Y}}(\boldsymbol{y}) = \sum_{i=1}^{3} \alpha_i \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i),$$

  with coefficients $\alpha_1 = \alpha_2 = 0.25$ and $\alpha_3 = 0.5$. We choose $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ to lie on the hypersphere of unit radius, and the covariances are generated as random positive semidefinite matrices.

- **Conditional Gaussians.** To evaluate the performance of VQR, we generate data from the following data-generating process

  $$\mathrm{X} \sim [0.5, 2.5], \quad f_{\mathbf{Y}}(\boldsymbol{y} \,|\, \mathrm{X} = x) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_x, x\boldsymbol{I}), \quad \boldsymbol{\mu}_x = (x, \ldots, x).$$

Metrics. We measure the quality of CVQFs by measuring the normalized KL divergence and statistical coverage defined as given below.

- **Normalized KL divergence.** To evaluate the quality of likelihood estimation, we use the normalized KL divergence defined as: KL divergence of the estimated likelihood w.r.t the ground truth likelihood, normalized by entropy of the ground truth distribution, it can be written as
$$\text{NormKL}(f_{\mathbf{Y}^*}||f_{\hat{\mathbf{Y}}}) = \frac{\text{KL}(f_{\mathbf{Y}^*}||f_{\hat{\mathbf{Y}}})}{\text{Entropy}(f_{\mathbf{Y}^*})}.$$
KL divergence between continuous distributions $f_{\hat{\mathbf{Y}}}$ and $f_{\mathbf{Y}^*}$ measures the minimum amount of bits wasted by representing a K-bin discretization of $f_{\hat{\mathbf{Y}}}$ with a code tailored for the K-bin discretization of $f_{\mathbf{Y}^*}$. The normalized KL divergence thus indicates the *proportion* of bits wasted.

- **Statistical coverage.** The statistical validity of the α-confidence is measured by calculating the coverage. Namely, given the α confidence level, we estimate its validity by sampling $\{\boldsymbol{y}_i\}_{i=1}^N$ from the ground truth distribution $f_{\mathbf{Y}}(\boldsymbol{y})$ and measuring:
$$\frac{1}{N}\sum_{i}^{N}\mathbb{I}\left\{\boldsymbol{y}_i \in \mathcal{C}_\alpha^{\mathbf{Y}}\right\} = \frac{1}{N}\sum_{i}^{N}\mathbb{I}\left\{Q_{\mathbf{Y}}^{-1}(\boldsymbol{y}_i) \in \mathcal{C}_\alpha^{\mathbf{U}}\right\}.$$
This computation trivially extends to the conditional case.

Baselines. In order to have a baseline for our estimated likelihoods, we fit simple parametric distributions both in the conditional and unconditional settings.

- **Unconditional.** As a simple baseline for the multi-modal Gaussian data, we define a parametric multivariate Gaussian distribution with the parameters given by the sample mean and sample covariance estimated from samples drawn from the true distribution.

- **Conditional.** As a baseline for the conditional case, we first sample $\{x_i\}_{i=1}^N \sim \mathcal{U}([0.5, 2.5])$ and, for each i, we sample $\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_{x_1}, x_i \boldsymbol{I})$. We then define a parametric multivariate Gaussian distribution with the parameters given by the sample mean and sample covariance computed from the samples $\{\boldsymbol{y}_i\}_{i=1}^N$.

Machine configuration and training time. All experiments were run on a machine with an Intel Xeon E5 CPU, 256GB of RAM and an Nvidia Titan 2080Ti GPU with 11GB dedicated graphics memory. Training time for PICNNs used in VQR and VQE experiments was approximately 1 hour.