Approximation of the Inverse CDF using Transport Map

Dawen Wu[1,2] Ludovic Chamoin[3]

[1]CNRS@CREATE, Singapore [2]National University of Singapore [3]Université Paris-Saclay, France

In memory of Prof. Stéphane Bressan

Problem Statement

Many important probability distributions, such as the normal distribution, lack closed-form analytical solutions for their inverse cumulative distribution functions (inverse CDFs or quantile functions).

Traditional non-parametric methods rely on numerical integration and interpolation, which can be computationally expensive and limit accuracy.

Goal: Develop novel, more accurate **parametric methods** to approximate the inverse CDF.

Background: Transport Maps

A transport map T creates a coupling between a simple reference distribution (e.g., standard uniform) and a complex target distribution.

Key Insight: For a 1D problem, if the reference distribution is Uniform(0,1), the optimal transport map T is exactly the inverse CDF (Φ^{-1}) of the target distribution.

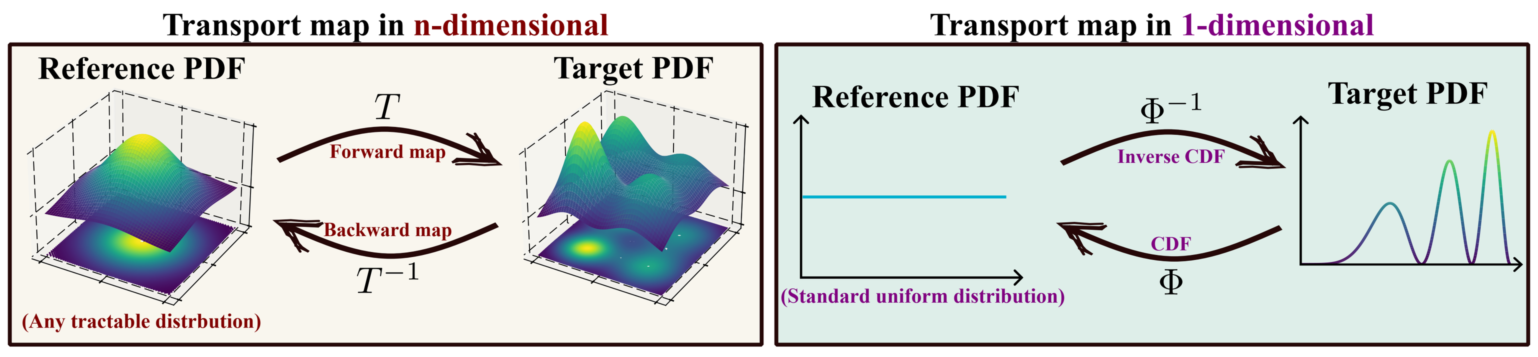

Figure 1. Left: General n-dimensional transport map. Right: In 1D, the map from a uniform distribution to a target distribution is the inverse CDF.

Proposed Methods

We propose a composite approximation function combining a **logit function** with a **neural network (NN)**:

$$\hat{\Phi}_{\text{inv}}(u; w) = N\left(\log\left(\frac{u}{1-u}\right); w\right) \approx \Phi^{-1}(u)$$

We introduce two distinct training strategies for the NN weights w.

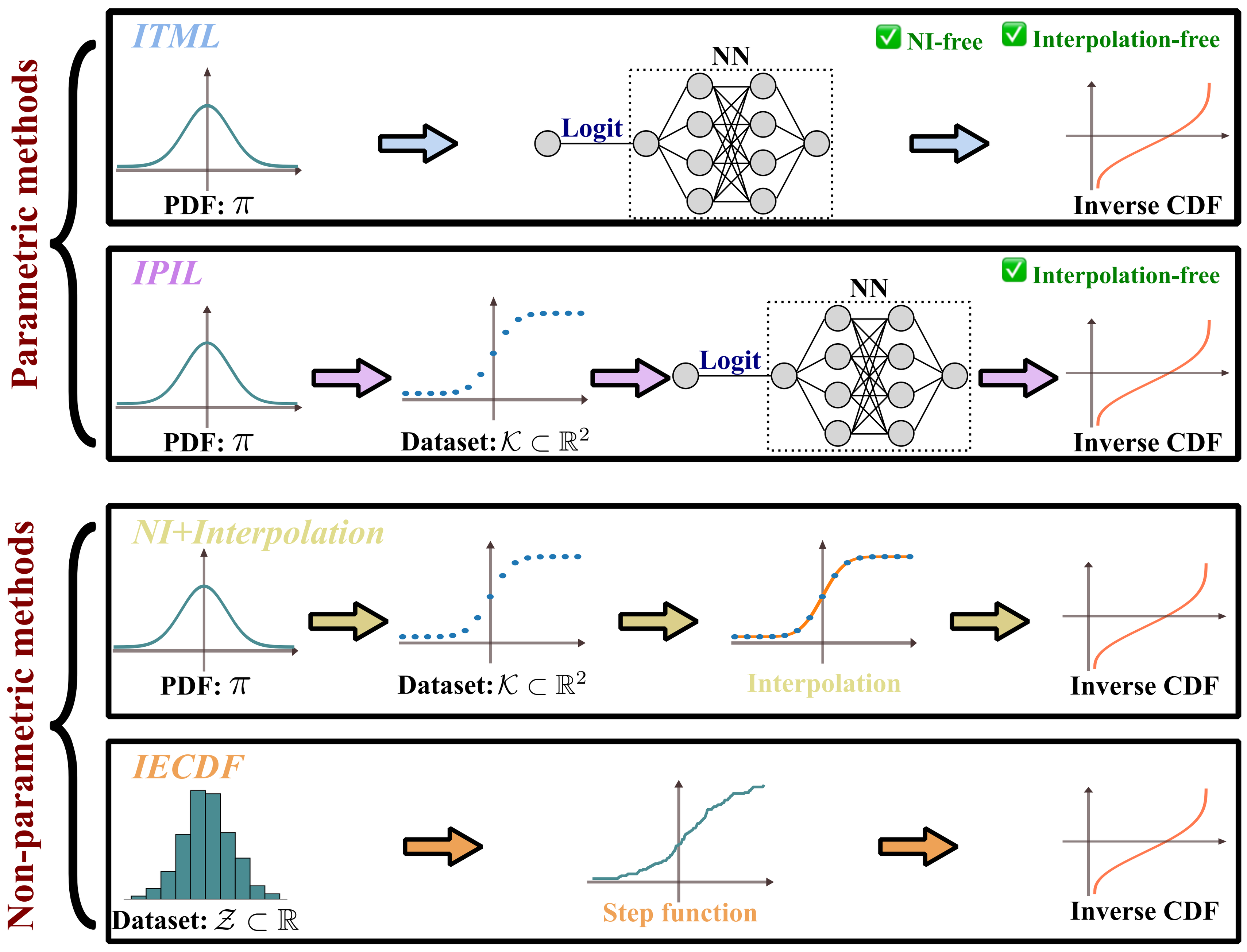

Figure 2. Overview of our parametric methods (ITML, IPIL) versus existing non-parametric approaches. Our methods avoid interpolation and ITML is also free from numerical integration.

Method 1: Inverse Transport Map Learning (ITML)

This approach is derived from the transport map theory by minimizing the Kullback-Leibler (KL) divergence.

- **Loss Function:** Based on minimizing the KL divergence between the reference and the transformed distribution.
- **Key Advantage:** It directly uses the Probability Density Function (PDF) $\pi(z)$ and is completely **free from numerical integration (NI-free)** and interpolation.
- **Constraint:** The derivative of the approximation must be positive to ensure invertibility.

Method 2: Inverse Physics-informed Learning (IPIL)

This approach formulates the problem as solving a differential equation using a Physics-Informed Neural Network (PINN).

- **Governing Equation:** The derivative of our approximation should match the derivative of the true inverse CDF, which is $1/\pi(\Phi^{-1}(u))$.
- **Training Data:** Requires a dataset of (u, z) pairs, where $u = \Phi(z)$ is computed via numerical integration.
- **Key Advantage:** Leverages the underlying physics (the PDF) to achieve high accuracy and avoids explicit interpolation (**Interpolation-free**).

Experiments and Results

We validated our methods on standard normal, Beta, Gamma, and an abstract distribution. The NN used has 3 hidden layers with 10 neurons each.

Case Study: Standard Normal Distribution

- ITML uses only the PDF, $\pi(z) \propto e^{-z^2/2}$.
- IPIL and other baseline methods use a training set of 10,000 (u, z) points.
- Performance is evaluated on a test set of 1 million points.

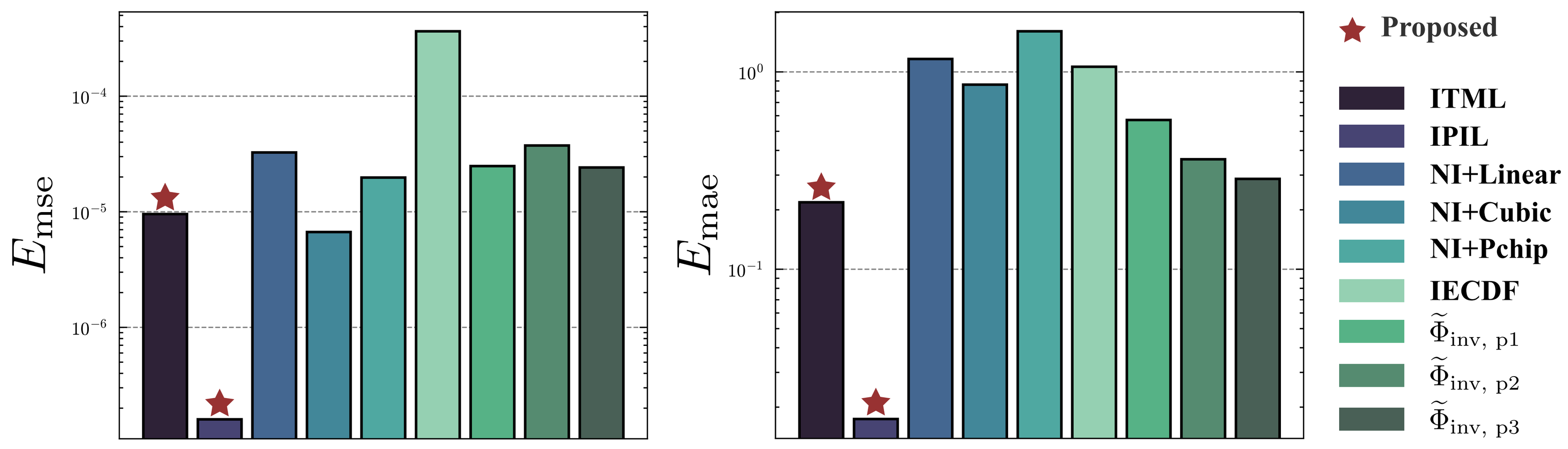

Figure 3. Performance comparison on the standard normal distribution. Our proposed methods (ITML, IPIL, marked with ⋆) are compared against various existing methods, including numerical integration with interpolation (NI+...) and explicit approximations ($\Phi_{\text{inv, p}^*}$).

Key Findings:

- As shown in the figure, **IPIL** achieves the lowest Mean Squared Error (E_{mse}), outperforming all other methods.
- **IPIL** and **ITML** show the best performance in Maximum Absolute Error (E_{mae}), indicating a robust approximation across the entire domain.
- Both proposed methods significantly outperform traditional non-parametric approaches like NI+Interpolation and the Inverse Empirical CDF (IECDF).
- Similar superior performance was observed for Beta and Gamma distributions.

Conclusions

- We introduced two novel parametric methods, **ITML** and **IPIL**, for approximating inverse CDFs using neural networks.
- Both methods achieve state-of-the-art accuracy, surpassing traditional non-parametric techniques.
- **ITML** is a powerful tool when only the PDF is known, as it requires no numerical integration to generate training data.
- **IPIL** acts as an advanced, physics-informed interpolation method, delivering extremely high accuracy when CDF observations can be generated.

Future Work

- Exploring approximation functions (e.g., polynomials) whose inverses are directly computable, to also obtain a CDF approximation.
- Extending the methodology to high-dimensional distributions, which poses significant conceptual challenges regarding inter-variable correlations.

Reference

- Michael Giles and Oliver Sheridan-Methven. Analysis of nested multilevel monte carlo using approximate normal random variables. SIAM/ASA Journal on Uncertainty Quantification, 2022.
- Tarek A El Moselhy and Youssef M Marzouk. Bayesian inference with optimal maps. Journal of Computational Physics, 2012.
- George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. Journal of Machine Learning Research, 2021.