THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

Tensorizing-Flow: Generative Modeling with Matrix Product State

The University of Chicago AI + Science Summer School *August 8th – 12th*

Yuehaw Khoo
The University of Chicago, *Department of Statistics*

Michael Lindsey
University of California, Berkeley, *Department of Mathematics*

Hongli Zhao
The University of Chicago, *Committee on Computational and Applied Mathematics*

Motivation

- Variational Inference (VI) is widely applied in approximating high-dimensional probability distributions from data.
- Normalizing Flow (NF) approximates a distribution as pushforward of a base distribution (e.g., Gaussian) via an invertible map
- Performance of NF relies on base distribution being absolutely continuous with respect to the target. Difficult to learn when target has complex structure (e.g., multi-modality).
- Tensor-train (TT) can represent a high-dimensional distribution in a low-parametric form achievable by linear algebra routines.
- We propose to combine TT initialization with the expressiveness of NF to achieve more accurate approximations while supporting efficient **sampling** and **density evaluation**.

Inference Model

- Given un-normalized target density $h(\boldsymbol{x})$ with $\boldsymbol{x} \in \mathbb{R}^d$ with true density defined as:

$$p(\boldsymbol{x}) = \frac{1}{Z} h(\boldsymbol{x})$$

where Z is normalizing constant.

- Optimize the variational lower bound over a family of distributions:

$$\min_{\theta} [D_{KL}(p_{\boldsymbol{\theta}} || p) - \log Z]$$

- We parameterize $p_{\boldsymbol{\theta}}$ using a flow-based model:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) \coloneqq p_0\left(f_{\boldsymbol{\theta}}^{-1}(\boldsymbol{x})\right) \cdot |\mathbf{D}\mathbf{f}_{\boldsymbol{\theta}}^{-1}|$$

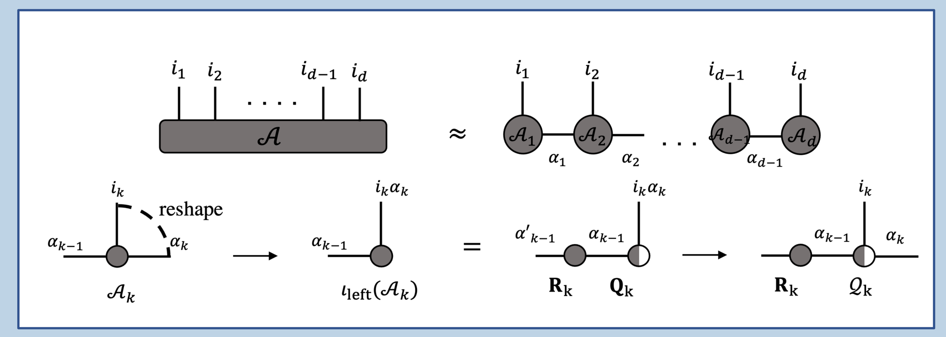
where $f_{\boldsymbol{\theta}}$ can be taken as a deep neural network (DNN). p_0 is a simple, known base distribution

- Empirical loss function estimated from N samples:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} [\log p_0(f_{\boldsymbol{\theta}}^{-1}(\boldsymbol{x}^{(i)})) - \log |\mathbf{D}\mathbf{f}_{\boldsymbol{\theta}}| - \log h(\boldsymbol{x}^{(i)})]$$

Tensor-Train Decomposition

- Tensor-Train (TT) decomposition is a compression method that generalizes truncated singular value decomposition to d-dimensional arrays.
- A tensor $\mathcal{A}$ of size n^d can be decomposed as $\mathcal{A}[i_1, \dots, i_d] \approx \mathcal{A}_1[:, i_1, :] \cdots \mathcal{A}_d[:, i_d, :]$, where each $\mathcal{A}_k$ is a 3-dimensional tensor that has size nr^2, with $r < n$
- Furthermore, each core can be put in right-left canonical form through QR decomposition.



Base Distribution using Squared-TT Ansatz

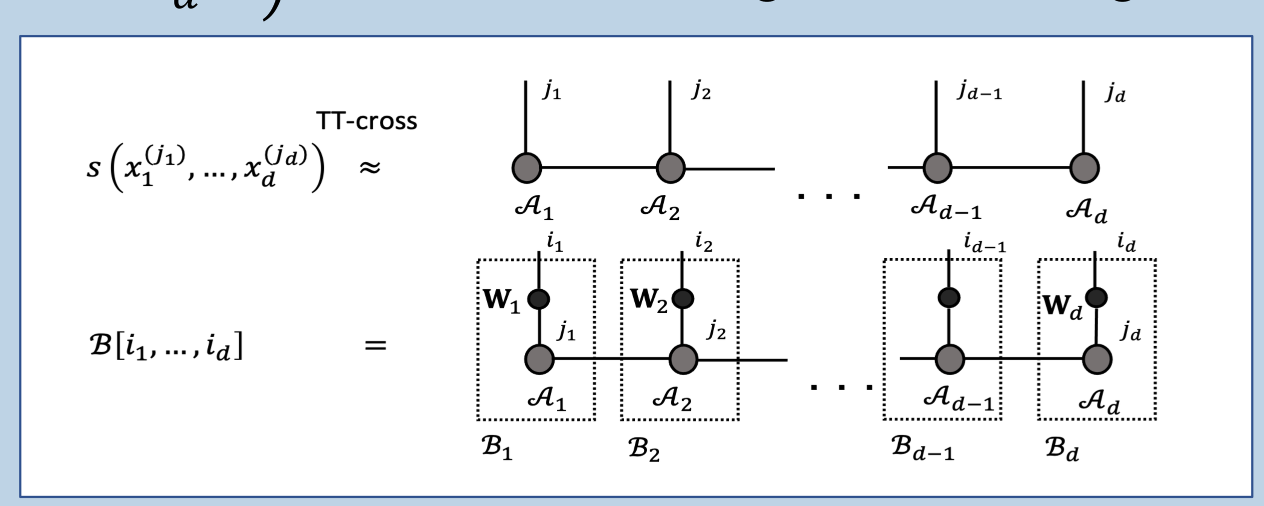
- Given a black-box to evaluate the un-normalized h, we propose to obtain an initial approximation to the target using multivariate basis expansion $s_{TT} \approx s = \sqrt{h}$

$$s_{TT}(x_1, \dots, x_d) = \sum_{i_1, \dots, i_d} \mathcal{B}[i_1, \dots, i_d] \, \phi_{i_1}(x_1) \cdots \phi_{i_d}(x_d)$$

- Coefficient tensor $\mathcal{B}$ is computed by projecting onto the orthogonal polynomial product basis $\prod_{j=1}^{d} \phi_{i_j}$ and evaluating on a quadrature given by quadrature $\left\{x_i^{(1)}, \dots, x_i^{(N_i)}\right\}$ and weights $\left\{w_i^{(1)}, \dots, w_i^{(N_i)}\right\}$ $(1 \le i \le d)$:.

$$\mathcal{B}[i_1, \dots, i_d] = \sum_{j_1, \dots, j_d} w_1^{(j_1)} \cdots w_d^{(j_d)} s\left(x_1^{(j_1)}, \dots, x_d^{(j_d)}\right) \phi_{i_1}\left(x_1^{(j_1)}\right) \cdots \phi_{i_d}\left(x_d^{(j_d)}\right)$$

- $s[j_1, \dots, j_d] \coloneqq s\left(x_1^{(j_1)}, \dots, x_d^{(j_d)}\right)$ is evaluated using the cross algorithm [1] as a TT.

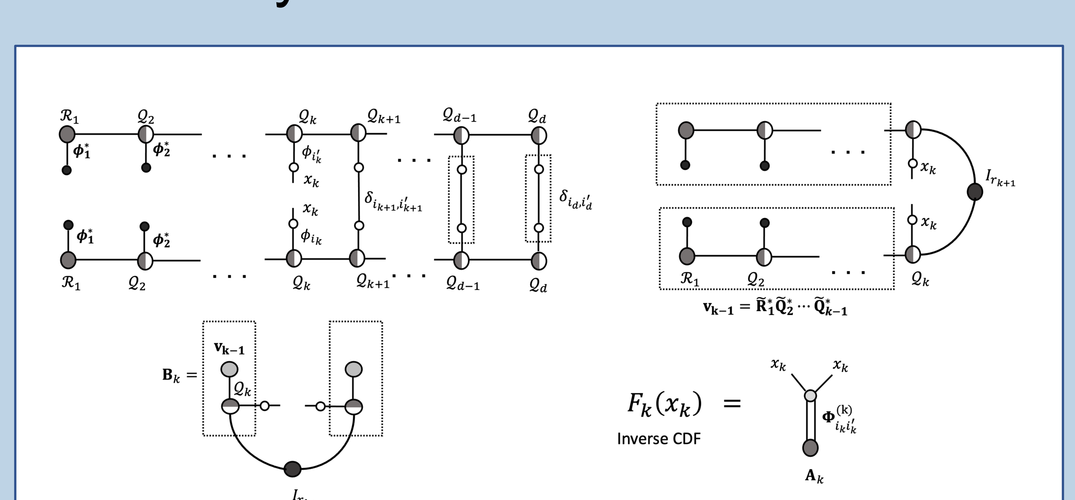


Computing coefficient tensor

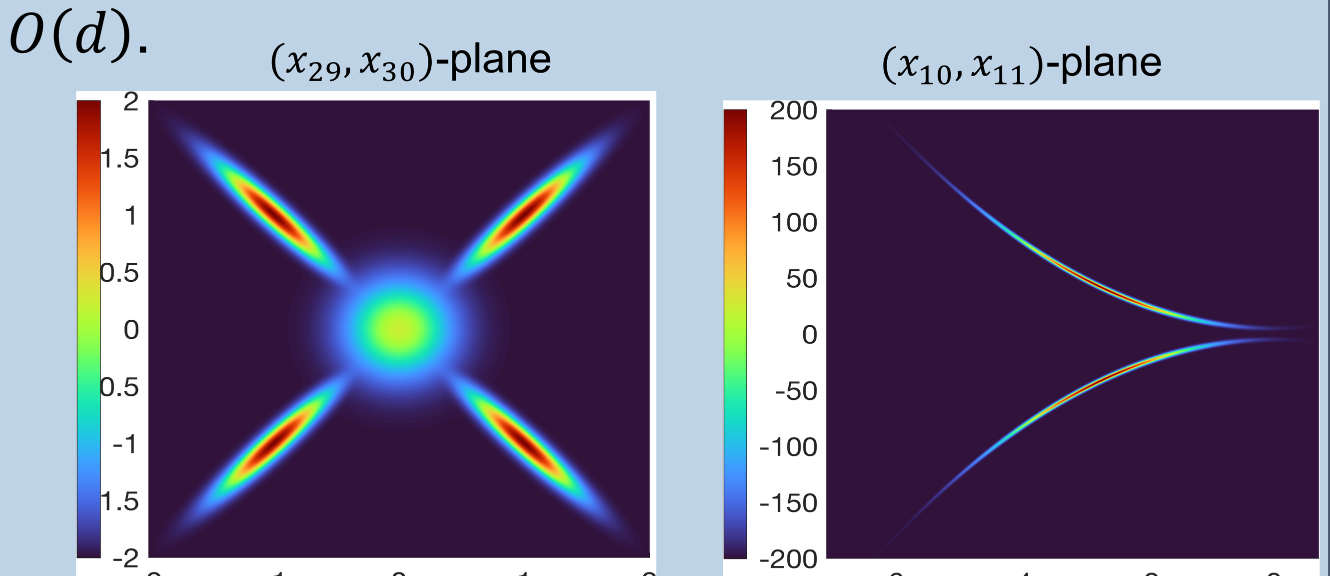
- Define base $p_{TT} \coloneqq s_{TT}^2 / ||s_{TT}||_{L^2}^2$, which always has unit mass and is non-negative.
- Putting $\mathcal{B}$ in right-left orthogonal form, conditional distribution sampling of the form:

$$p(x_1, x_2, \dots, x_d) = p(x_1) p(x_2|x_1) \cdots p(x_d | x_1, \dots x_{d-1})$$

and density evaluations can be achieved in $O(d)$.



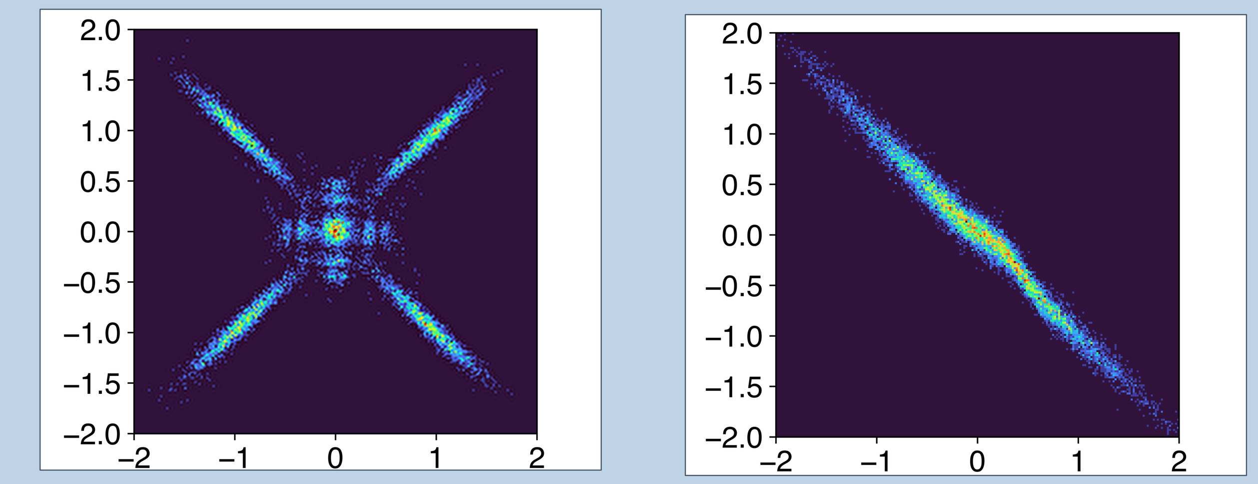
Schema for conditional distribution sampling



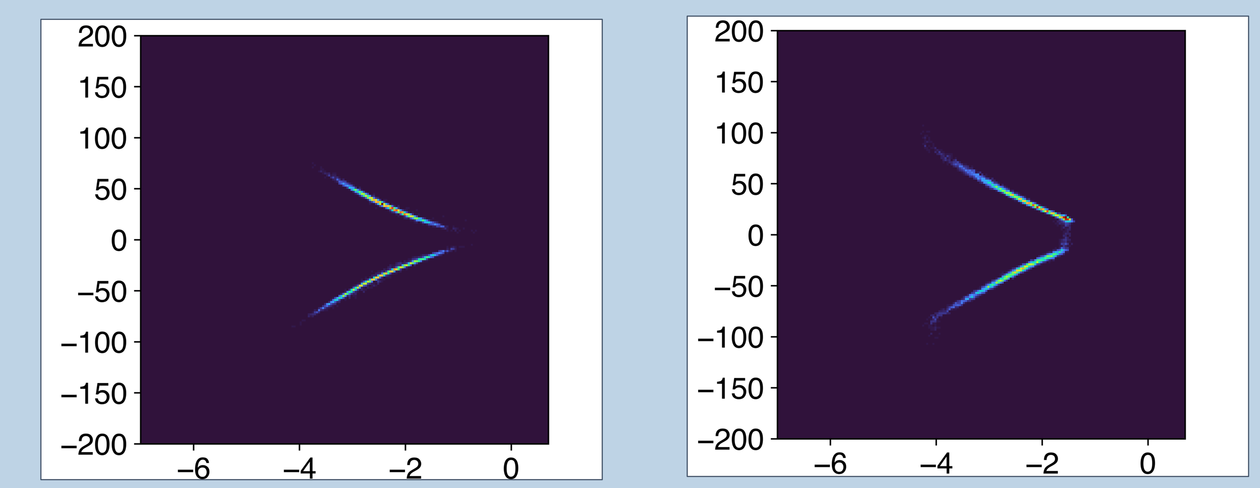
Squared TT reconstruction of high-dimensional distributions.
Left: mixture Gaussian (d=30), Right: double Rosenbrock function (d=11)

Accuracy Comparison

- Visualization of samples from learned distributions

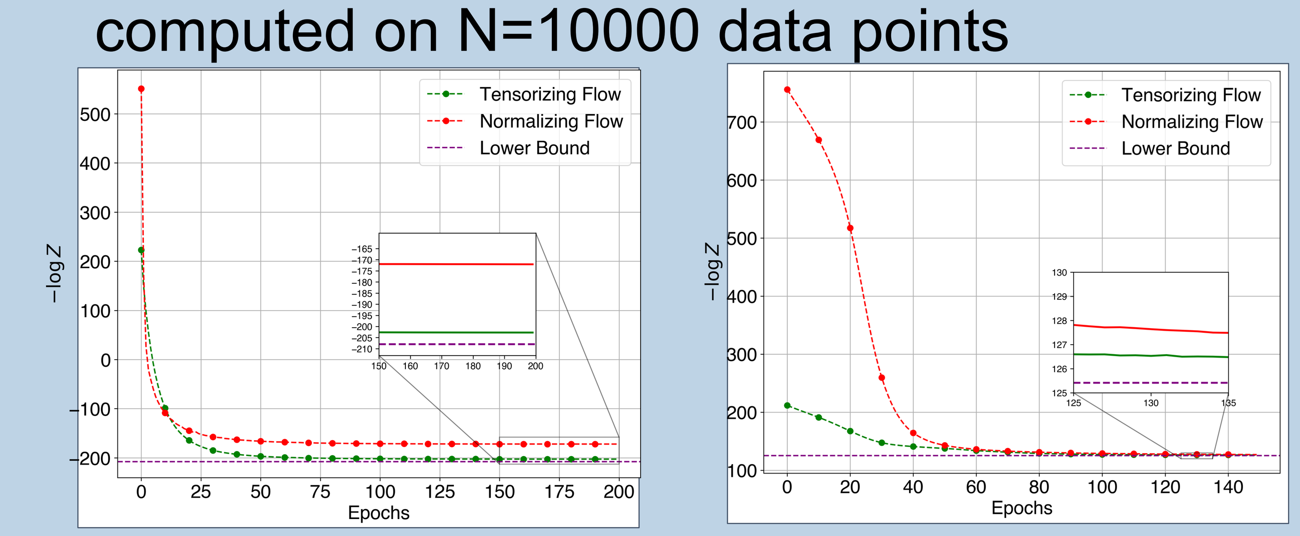


N=10000 samples from Tensorizing Flow   N=10000 samples from Normalizing Flow



N=10000 samples from Tensorizing Flow   N=10000 samples from Normalizing Flow

- Comparison of estimated lower bound, $-\log Z$, computed on N=10000 data points



Mixture Gaussian distribution   Double Rosenbrock function

Tensorizing-Flow

- Combining low-parametric initialization of base distribution in flow-based models as the squared TT format and a flexible push-forward map $f_{\boldsymbol{\theta}}$ gives an efficient and expressive generative model.

Future Work

- General learning tasks where model response is required to be non-negative.
- More complex kernel functions.

References

[1] *Oseledets et al.*, TT-cross approximation for multidimensional arrays