A Prior of a Googol Gaussians: a Tensor Ring Induced Prior for Generative Models

Maksim Kuznetsov[1,*], Daniil Polykovskiy[1,*], Dmitry Vetrov[2], Alexander Zhebrak[1]
*Equal contribution [1]Insilico Medicine [2]National Research University Higher School of Economics

Abstract

We propose a new family of prior distributions—*Tensor Ring Induced Prior*—that packs an exponential number of Gaussians into a high-dimensional lattice with a relatively small number of parameters. In this paper we:

- Investigate applications of proposed prior in GANs and VAE
- Study generative models with TRIP in the conditional generation setup with missing conditions

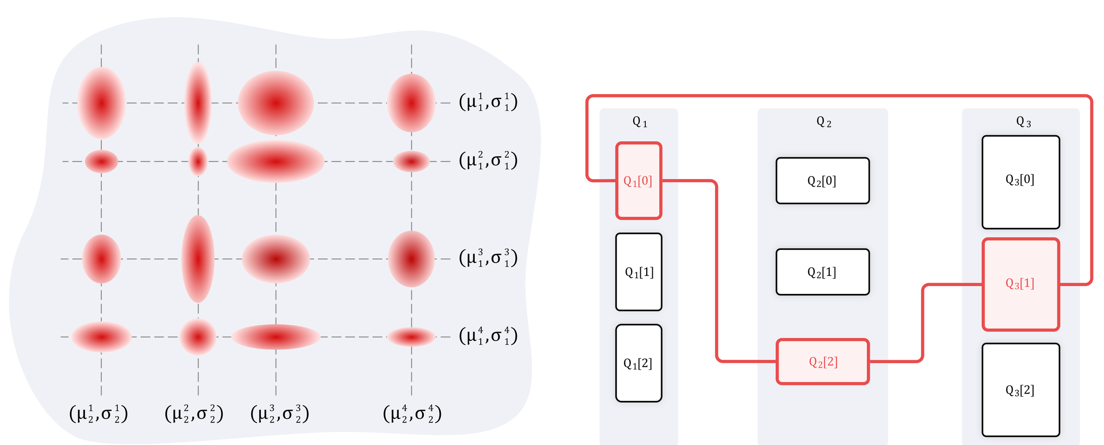


Fig. 1: **Left:** The TRIP distribution. **Right:** How to compute an element of a distribution defined with a Tensor Ring format.

Discrete distribution in the Tensor Ring format

Consider a joint distribution $p(r_1, r_2, \dots r_d)$ of d discrete random variables r_k taking values from $\{0, 1, \dots N_k{-}1\}$. We can write these probabilities as elements of a d-dimensional tensor $P[r_1, r_2, \dots r_d] = p(r_1, r_2, \dots r_d)$, represented in Tensor Ring decomposition [3]:

$$p(r_{1:d}) \propto \widehat{P}[r_{1:d}] = \mathrm{Tr}\left(\prod_{j=1}^{d} Q_j[r_j]\right) \tag{1}$$

Properties:

1. To marginalize out the random variable r_k, we replace cores Q_k in Eq 1 with a matrix $\widetilde{Q}_k = \sum_{r_k=0}^{N_k-1} Q_k[r_k]$:

$$p(r_{1:k-1}, r_{k+1:d}) \propto \mathrm{Tr}\left(\prod_{j=1}^{k-1} Q_j[r_j] \cdot \widetilde{Q}_k \cdot \prod_{j=k+1}^{d} Q_j[r_j]\right) \tag{2}$$

2. We can store a nonnormalized distribution and compute a normalizing constant Z:

$$Z = \mathrm{Tr}\left(\prod_{j=1}^{d} \widetilde{Q}_k\right) \tag{3}$$

3. Conditional probabilities can be computed as a ratio between joint and marginal probabilities.

$$p(r_A | r_B) = \frac{\mathrm{Tr}\left(\prod_{j\in(A\cup B)} Q_j[r_j] \prod_{k\notin(A\cup B)} \widetilde{Q}_k\right)}{\mathrm{Tr}\left(\prod_{j\in B} Q_j[r_j] \prod_{k\notin B} \widetilde{Q}_k\right)} \tag{4}$$

4. Sample from this distribution using chain rule:

$$r_1 \sim p(r_1) \Rightarrow r_2 \sim p(r_2|r_1) \Rightarrow \dots \Rightarrow r_d \sim p(r_d|r_1, \cdots, r_{d-1}) \tag{5}$$

Tensor Ring Induced Prior

To generalize this approach for continuous random variables $p(z)$, we use a multidimensional Gaussian Mixture Model with modes placed in the nodes of a multidimensional lattice. A discrete distribution of mixture components $p(s)$ is stored as a tensor $\widehat{P}[s]$ in a Tensor Ring decomposition:

$$p(z_{1:d}) = \sum_{s_{1:d}} p(s_{1:d}) p(z_{1:d} \mid s_{1:d}) \propto \sum_{s_{1:d}} \widehat{P}[s_{1:d}] \prod_{j=1}^{d} \mathcal{N}(z_j \mid \mu_j^{s_j}, \sigma_j^{s_j}) \tag{6}$$

We call this family of distributions a Tensor Ring Induced Prior (TRIP). TRIP provides a fast and differentiable way to compute log-probabilities.

Algorithm 1 Calculation of marginal probabilities in TRIP

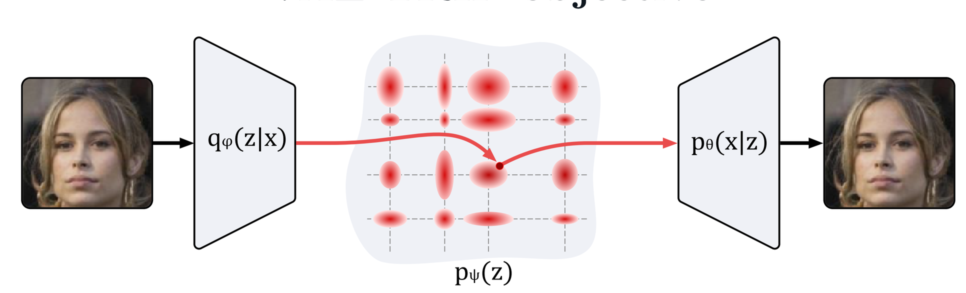
Input: A set M of variable indices for which we compute the probability, values of z_i for $i \in M$
Output: Joint probability $\log p(z_M)$, where $z_M = \{z_i \ \forall i \in M\}$
Initialize $Q_{\text{buff}} = I \in \mathbb{R}^{m_1 \times m_1}$, $Q_{\text{norm}} = I \in \mathbb{R}^{m_1 \times m_1}$
for $j = 1$ **to** d **do**
  if j is marginalized out ($j \notin M$) **then**
    $Q_{\text{buff}} = Q_{\text{buff}} \cdot \left(\sum_{k=0}^{N_j-1} Q_j[k]\right)$
  else
    $Q_{\text{buff}} = Q_{\text{buff}} \cdot \left(\sum_{k=0}^{N_j-1} Q_j[k] \cdot \mathcal{N}\left(z_k \mid \mu_j^{s_j}, \sigma_j^{s_j}\right)\right)$
  end if
  $Q_{\text{norm}} = Q_{\text{norm}} \cdot \left(\sum_{k=0}^{N_j-1} Q_j[k]\right)$
end for
$\log p(z_M) = \log \mathrm{Tr}\left(Q_{\text{buff}}\right) - \log \mathrm{Tr}\left(Q_{\text{norm}}\right)$

An important property of the proposed TRIP family is that we can derive its one-dimensional conditional distributions in a closed form which is useful for sampling:

$$p_\psi(z_k \mid z_{1:k-1}) = \sum_{s_k=0}^{N_k-1} p_\psi(s_k \mid z_{1:k-1}) \mathcal{N}(z_k \mid \mu_k^{s_k}, \sigma_k^{s_k}) \tag{7}$$

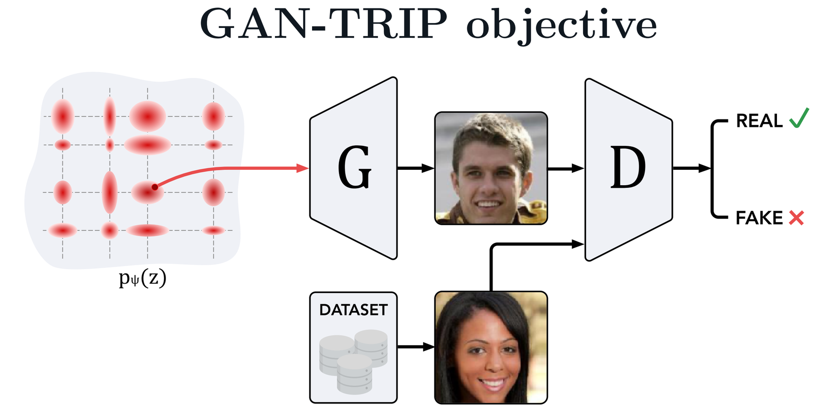
VAE-TRIP and GAN-TRIP

VAE-TRIP objective



$$\mathcal{L}(\theta, \phi, \psi) \approx \frac{1}{l} \sum_{i=1}^{l} \log\left(\frac{p_\theta(x \mid z_i) p_\psi(z_i)}{q_\phi(z_i \mid x)}\right), \quad z_i = \epsilon_i \cdot \sigma_\phi(x) + \mu_\phi(x) \tag{8}$$

GAN-TRIP objective



$$\min_{G,\psi} \max_{D} \mathcal{L}_{GAN} = \mathbb{E}_{x\sim p(x)} \log D(x) + \mathbb{E}_{z\sim p_\psi(z)} \log\left(1 - D(G(z))\right) \tag{9}$$

$$\nabla_\psi \mathcal{L}_{GAN} \approx \frac{1}{l} \sum_{i=1}^{l} \nabla_\psi \log p_\psi(z_i) \left[d_i - \frac{1}{l} \sum_{j=1}^{l} d_j\right] \tag{10}$$

Experiments

Metric	Model	$\mathcal{N}(0, I)$	GMM F	GMM L	TRIP F	TRIP L
FID	VAE	86.72	85.64	84.48	85.31	**83.54**
	WGAN	63.46	67.10	61.82	62.48	**57.6**
	WGAN-GP	54.71	57.82	62.10	63.06	**52.86**
ELBO	VAE	-194.16	-201.60	-193.88	-202.04	**-193.32**
IWAE ELBO ($k = 100$)	VAE	-185.09	-191.99	-184.73	-190.09	**-184.43**

Tab. 1: FID for WGAN, WGAN-GP, VAE trained on CelebA dataset, and ELBO for VAE. F = Fixed, L = Learnable

	$\mathcal{N}(0, 1)$	GMM	TRIP	+ RealNVP $\mathcal{N}(0, I)$	+ RealNVP GMM	+ RealNVP TRIP
Parameters (model)	11.4M	11.1M	10.7M	11.3M	10.7M	10.4M
Parameters (prior)	0	0.2M	0.6M	0.3M	0.5M	0.7M
Parameters (total)	11.4M	11.3M	11.1M	11.5M	11.2M	11.1M
ELBO	-192.6	-190.05	-189.1	-185.3	-186.0	**-184.7**

Tab. 2: ELBO on CELEBA for different learnable priors. Number of parameters are the same.

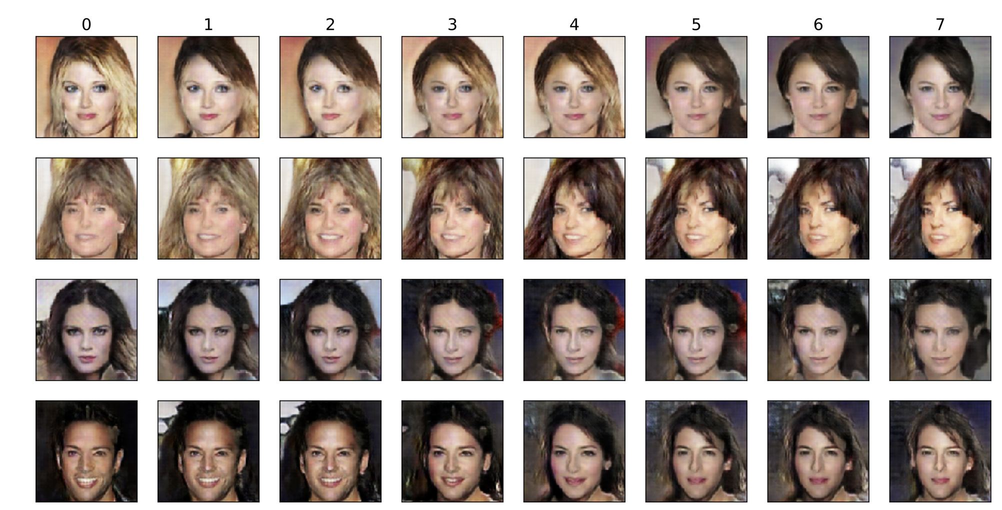


Fig. 4: Mode hopping in WGAN-GP-TRIP.

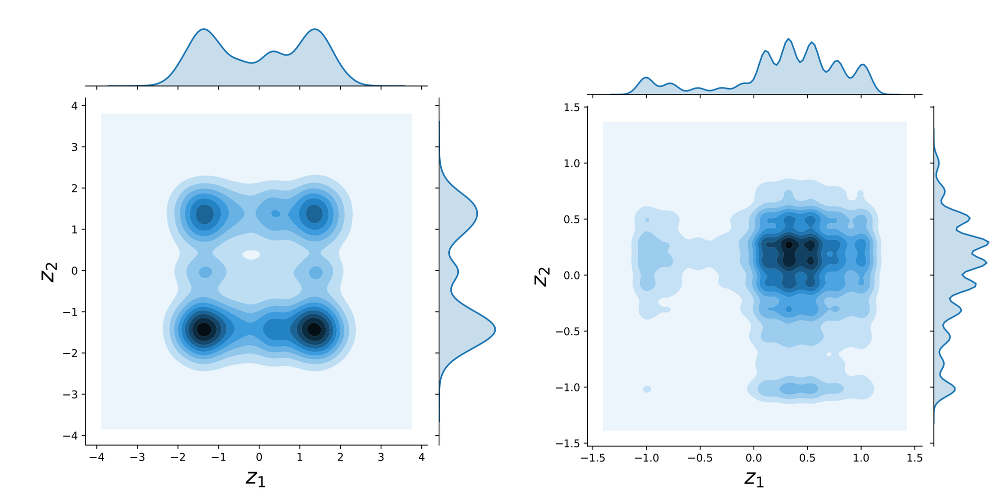


Fig. 5: Visualization of the first two dimensions of the learned prior $p_\psi(z_1, z_2)$ in VAE-TRIP and WGAN-GP-TRIP.

References

[1] Ian Goodfellow et al. "Generative Adversarial Nets". In: *Advances in neural information processing systems* (2014), pp. 2672–2680.

[2] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *International Conference on Learning Representations* (2013).

[3] Qibin Zhao et al. "Tensor Ring Decomposition". In: *arXiv preprint arXiv:1606.05535* (2016).

[4] Alex Zhavoronkov et al. "Deep learning enables rapid identification of potent DDR1 kinase inhibitors". In: *Nature biotechnology* (2019), pp. 1–4.