NF 4: Discrete, Mixed Flows

CS 598: Deep Generative and Dynamical Models

Instructor: Arindam Banerjee

October 12, 2021

Motivation and Background

- Data is on discrete domain, models are continuous
- Need for de-quantization to move back and forth
- *Continuous* change of variables

$$p_X(\mathrm{x}) = p_Z(\mathrm{z}) \left| \det \frac{d\mathrm{z}}{d\mathrm{x}} \right| \ , \quad \mathrm{z} = f(\mathrm{x})$$

- Variety of flows have been developed
  - Coupling flows: Input $\mathrm{x} = [\mathrm{x}_a, \mathrm{x}_b]$, deep nets $s(\cdot), t(\cdot)$

$$\mathrm{z} = [\mathrm{z}_a, \mathrm{z}_b] = f(\mathrm{x}) = [\mathrm{x}_a, \mathrm{x}_b \odot s(\mathrm{x}_a) + t(\mathrm{x}_a)]$$

  - Factor-out layers:

$$[\mathrm{z}_1, \mathrm{y}_1] = f_1(\mathrm{x}) \ , \quad \mathrm{z}_2 = f_2(\mathrm{y}_1) \ , \quad \mathrm{z} = [\mathrm{z}_1, \mathrm{z}_2]$$

$$p(\mathrm{x}) = p(\mathrm{z}_2) \left| \det \frac{\partial d_2(\mathrm{y}_1}{\partial \mathrm{y}_1} \right| p(\mathrm{z}_1 | \mathrm{y}_1) \left| \det \frac{\partial f_1(\mathrm{x})}{\partial \mathrm{x}} \right|$$

Lossless Compression, Entropy Encoding

- True data distribution $x \sim \mathcal{D}$
  - Encode x with $-\log \mathcal{D}(x)$ bits
  - Expected code length is entropy $H(\mathcal{D}) = \mathbb{E}_{x \sim \mathcal{D}}[-\log \mathcal{D}(x)$
- Model using distribution $p_X(x)$
  - x encoded as $c(x)$, with $|c(x)| \approx -\log p_X(x)$ bits
- Expected code length

$$\mathbb{E}_{x \sim \mathcal{D}}[|c(x)|] \approx \mathbb{E}_{x \sim \mathcal{D}}[-\log p_X(x)] \geq \mathcal{H}(\mathcal{D})$$

- Map symbols to bits, use entropy encoders
- Compression: Map x to z, encode using $p_Z(z)$

Integer Discrete Flows (IDFs)

- IDFs learn probability mass functions on $\mathbb{Z}^d$, over integer vectors
- Prior p_Z on $\mathbb{Z}^d$, bijective map $f : \mathbb{Z}^d \mapsto \mathbb{Z}^d$

$$p_X(\mathrm{x}) = p_Z(\mathrm{z}) \ , \quad z = f(\mathrm{x})$$

- Stack multiple IDF layers $\{f_l\}_{l=1}^{L}$, ensure composition is closed
- Integer discrete coupling: Additive coupling with rounding

$$\mathrm{z}_a = \mathrm{x}_a \ , \quad \mathrm{z}_b = \mathrm{x}_b + \lfloor t(\mathrm{x}_a) \rceil$$

- Backprop through rounding: Ignore rounding, i.e., $\nabla_x \lfloor x \rceil = 1$
  - Incurs bias in gradient computation

Tractable Discrete Distributions

- Logistic$(z|\mu, s)$ has CDF sigmoid on $(x-\mu)/s$, i.e., $\sigma((x-\mu)/s)$
- Prior p_Z is factored discreteized logistic distribution

$$\begin{aligned}\text{DLogistic}(z|\mu, s) &= \int_{z-1/2}^{z+1/2} \text{Logistic}(z'|\mu, s)dz' \\ &= \sigma\left(\frac{z+1/2-\mu}{s}\right) - \sigma\left(\frac{z-1/2-\mu}{s}\right)\end{aligned}$$

- For stacked layers, μ, s are deep nets
- Factor out layers $[\mathrm{z}_l, y_l]$, z_l uses $\mu(y_l), s(\mathrm{y}_l)$
- Beyond unimodality: Use a mixture model

$$p(z|\mu, s, \pi) = \sum_k \pi_k \cdot p(z|\mu_k, s_k)$$

Architecture, Gradient Bias

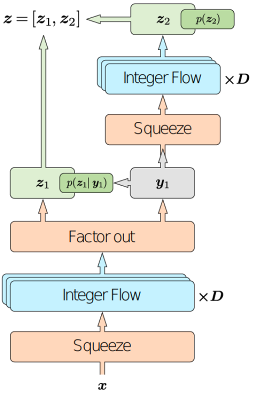


Figure 4: Example of a 2-level flow architecture. The squeeze layer reduces the spatial dimensions by two, and increases the number of channels by four. A single integer flow layer consists of a channel permutation and an integer discrete coupling layer. Each level consists of D flow layers.

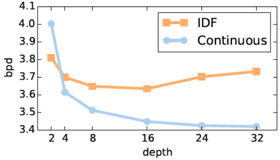


Figure 5: Performance of flow models for different depths (i.e. coupling layers per level). The networks in the coupling layers contain 3 convolution layers. Although performance increases with depth for continuous flows, this is not the case for discrete flows.

Results: Image Compression, Vision benchmarks

Table 1: Compression performance of IDFs on CIFAR10, ImageNet32 and ImageNet64 in bits per dimension, and compression rate (shown in parentheses). The Bit-Swap results are retrieved from [23]. The column marked IDF† denotes an IDF trained on ImageNet32 and evaluated on the other datasets.

Dataset	IDF	IDF†	Bit-Swap	FLIF [35]	PNG	JPEG2000
CIFAR10	**3.34 (2.40**$\times$**)**	3.60 (2.22$\times$)	3.82 (2.09$\times$)	4.37 (1.83$\times$)	5.89 (1.36$\times$)	5.20 (1.54$\times$)
ImageNet32	**4.18 (1.91**$\times$**)**	**4.18 (1.91**$\times$**)**	4.50 (1.78$\times$)	5.09 (1.57$\times$)	6.42 (1.25$\times$)	6.48 (1.23$\times$)
ImageNet64	**3.90 (2.05**$\times$**)**	3.94 (2.03 $\times$)	–	4.55 (1.76$\times$)	5.74 (1.39$\times$)	5.10 (1.56$\times$)

Results: Image Compression, Histology

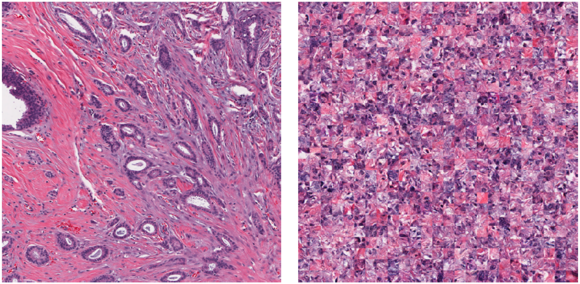


Figure 6: Left: An example from the ER + BCa histology dataset. Right: 625 IDF samples of size 80×80px.

Table 2: Compression performance on the ER + BCa histology dataset in bits per dimension and compression rate. JP2-WSI is a specialized format optimized for virtual microscopy.

Dataset	IDF	JP2-WSI	FLIF [35]	JPEG2000
Histology	**2.42 (3.19×)**	3.04 (2.63×)	4.00 (2.00×)	4.26 (1.88×)

Results: Progressive Image Rendering

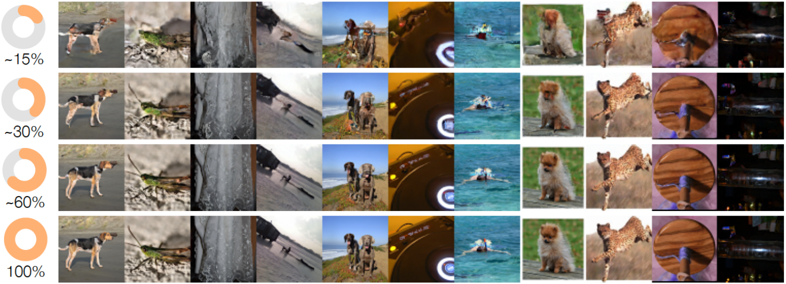


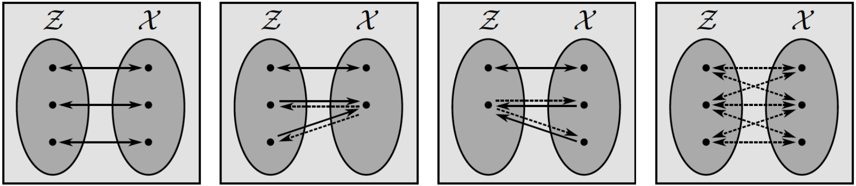
Figure 8: Progressive display of the data stream for images taken from the test set of ImageNet64. From top to bottom row, each image uses approximately 15%, 30%, 60% and 100% of the stream, where the remaining dimensions are sampled. Best viewed electronically.

Results: Probability Mass Estimation

Table 3: Generative modeling performance of IDFs and comparable flow-based methods in bits per dimension (negative $\log_2$-likelihood).

Dataset	IDF	Continuous	RealNVP	Glow	Flow++
CIFAR10	3.32	3.31	3.49	3.35	3.08
ImageNet32	4.15	4.13	4.28	4.09	3.86
ImageNet64	3.90	3.85	3.98	3.81	3.69

Surjections for Flows



(a) Bijective (b) Surjective (Gen.) (c) Surjective (Inf.) (d) Stochastic

Figure 1: Classes of SurVAE transformations $\mathcal{Z} \to \mathcal{X}$ and their inverses $\mathcal{X} \to \mathcal{Z}$. Solid lines indicate deterministic transformations, while dashed lines indicate stochastic transformations.

Unifying Framework: VAEs, Flows, etc.

- Surjective maps:
  - Generative: $\forall \mathrm{x} \in \mathcal{X}$, $\exists \mathrm{z} \in \mathcal{Z}$, $\mathrm{x} = f(\mathrm{z})$
  - Inference: $\forall \mathrm{z} \in \mathcal{Z}$, $\exists \mathrm{x} \in \mathcal{X}$, $\mathrm{z} = g(\mathrm{x})$
  - Inverse map is stochastic with support on preimage
- For bijections, e.g., flows

$$\log p(\mathrm{x}) = \log p(\mathrm{z}) + \log |\det J| \ , \quad \mathrm{z} = f^{-1}(\mathrm{x})$$

- For stochastic transformations, e.g., VAEs

$$\log p(\mathrm{x}) = \mathbb{E}_{q(\mathrm{z}|\mathrm{x})}[\log p(\mathrm{x}|\mathrm{z})] - KL(q(\mathrm{z}|\mathrm{x}) \| p(\mathrm{z})) + KL(q(\mathrm{z}|\mathrm{x}) \| p(\mathrm{z}|\mathrm{x}))$$

- General perspective

$$\log p(\mathrm{x}) \approx \log p(\mathrm{z}) + \mathcal{V}(\mathrm{x},\mathrm{z}) + \mathcal{E}(\mathrm{x},\mathrm{z}) \ , \quad \mathrm{z} \sim q(\mathrm{z}|\mathrm{x})$$

  - $\mathcal{V}(\mathrm{x},\mathrm{z})$: Likelihood contribution
  - $\mathcal{E}(\mathrm{x},\mathrm{z})$: Bound looseness

Log-Likelihood Computation

Algorithm 1: $\mathsf{log - likelihood}(\boldsymbol{x})$

Data: $\boldsymbol{x}$, $p(\boldsymbol{z})$ & $\{f_t\}_{t=1}^{T}$
Result: $\mathcal{L}(\boldsymbol{x})$
for t *in* $\mathsf{range}(T)$, **do**
  if f_t *is bijective* **then**
    $\boldsymbol{z} = f_t^{-1}(\boldsymbol{x})$;
    $\mathcal{V}_t = \log \left|\det \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}\right|$;
  else if f_t *is stochastic* **then**
    $\boldsymbol{z} \sim q_t(\boldsymbol{z}|\boldsymbol{x})$;
    $\mathcal{V}_t = \log \frac{p_t(\boldsymbol{x}|\boldsymbol{z})}{q_t(\boldsymbol{z}|\boldsymbol{x})}$;
  $\boldsymbol{x} = \boldsymbol{z}$;
end
return $\log p(\boldsymbol{z}) + \sum_{t=1}^{T} \mathcal{V}_t$

Composable Blocks of SurVAE Flows

Table 1: Composable building blocks of SurVAE Flows.

Transformation	**Forward** $x \leftarrow z$	**Inverse** $z \leftarrow x$	**Likelihood Contribution** $\mathcal{V}(x, z)$	**Bound Gap** $\mathcal{E}(x, z)$
Bijective	$x = f(z)$	$z = f^{-1}(x)$	$\log \lvert \det \nabla_x z \rvert$	0
Stochastic	$x \sim p(x\|z)$	$z \sim q(z\|x)$	$\log \frac{p(x\|z)}{q(z\|x)}$	$\log \frac{q(z\|x)}{p(z\|x)}$
Surjective (Gen.)	$x = f(z)$	$z \sim q(z\|x)$	$\log \frac{p(x\|z)}{q(z\|x)}$ as $p(x\|z) \rightarrow \delta(x - f(z))$	$\log \frac{q(z\|x)}{p(z\|x)}$
Surjective (Inf.)	$x \sim p(x\|z)$	$z = f^{-1}(x)$	$\log \frac{p(x\|z)}{q(z\|x)}$ as $q(z\|x) \rightarrow \delta(z - f^{-1}(x))$	0

Examples: Inference Surjection Layers

Table 2: Summary of selected inference surjection layers. See App. C for more SurVAE layers.

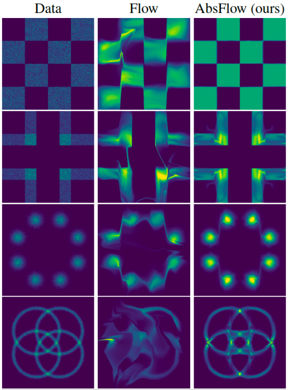
Surjection	**Forward**	**Inverse**	$\mathcal{V}(\boldsymbol{x}, \boldsymbol{z})$
Abs	$s \sim \text{Bern}(\pi(z))$ $x = s \cdot z, \quad s \in \{-1, 1\}$	$s = \text{sign}\, x$ $z = \lvert x \rvert$	$\log p(s \mid z)$
Max	$k \sim \text{Cat}(\boldsymbol{\pi}(z))$ $x_k = z, \boldsymbol{x}_{-k} \sim p(\boldsymbol{x}_{-k} \mid z, k)$	$k = \arg\max \boldsymbol{x}$ $z = \max \boldsymbol{x}$	$\log p(k \mid z) + \log p(\boldsymbol{x}_{-k} \mid z, k)$
Sort	$\mathcal{I} \sim \text{Cat}(\boldsymbol{\pi}(\boldsymbol{z}))$ $\boldsymbol{x} = \boldsymbol{z}_{\mathcal{I}}$	$\mathcal{I} = \text{argsort}\, \boldsymbol{x}$ $\boldsymbol{z} = \text{sort}\, \boldsymbol{x}$	$\log p(\mathcal{I} \mid \boldsymbol{z})$

Unifying Framework, Redux

Table 3: SurVAE Flows as a unifying framework.

Model	**SurVAE Flow architecture**
Probabilistic PCA (Tipping and Bishop, 1999) VAE (Kingma and Welling, 2014; Rezende et al., 2014) Diffusion Models (Sohl-Dickstein et al., 2015; Ho et al., 2020)	$\mathcal{Z} \xrightarrow{\text{stochastic}} \mathcal{X}$
Dequantization (Uria et al., 2013; Ho et al., 2019)	$\mathcal{Z} \xrightarrow{\text{round}} \mathcal{X}$
ANFs, VFlow (Huang et al., 2020; Chen et al., 2020)	$\mathcal{X} \xrightarrow{\text{augment}} \mathcal{X} \times \mathcal{E} \xrightarrow{\text{bijection}} \mathcal{Z}$
Multi-scale Architectures (Dinh et al., 2017)	$\mathcal{X} \xrightarrow{\text{bijection}} \mathcal{Y} \times \mathcal{E} \xrightarrow{\text{slice}} \mathcal{Y} \xrightarrow{\text{bijection}} \mathcal{Z}$
CIFs, Discretely Indexed Flows, DeepGMMs (Cornish et al., 2019; Duan, 2019; Oord and Dambre, 2015)	$\mathcal{X} \xrightarrow{\text{augment}} \mathcal{X} \times \mathcal{E} \xrightarrow{\text{bijection}} \mathcal{Z} \times \mathcal{E} \xrightarrow{\text{slice}} \mathcal{Z}$
RAD Flows (Dinh et al., 2019)	$\mathcal{X} \xrightarrow{\text{partition}} \mathcal{X}_{\mathcal{E}} \times \mathcal{E} \xrightarrow{\text{bijection}} \mathcal{Z} \times \mathcal{E} \xrightarrow{\text{slice}} \mathcal{Z}$

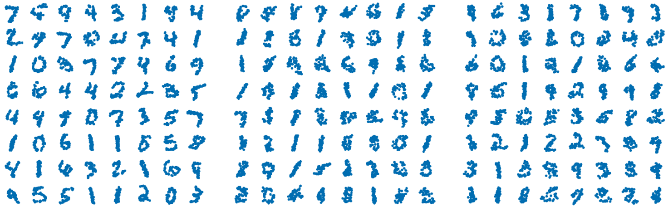
Results: Synthetic Data



Dataset	Flow	AbsFlow (ours)
Checkerboard	3.65	**3.49**
Corners	3.19	**3.03**
Gaussians	3.01	**2.86**
Circles	3.44	**2.99**

Figure 4: Comparison of flows with and without absolute value surjections modelling anti-symmetric (top row) and symmetric (3 bottom rows) 2-dimensional distributions.

Results: Point Clouds



(a) Data (b) SortFlow (c) PermuteFlow

Figure 5: Point cloud samples from permutation-invariant SurVAE flows trained on SpatialMNIST.

Results: Bits/dim in benchmarks

Table 4: Unconditional image modeling results in bits/dim.

Model	**CIFAR-10**	**ImageNet32**	**ImageNet64**
RealNVP (Dinh et al., 2017)	3.49	4.28	-
Glow (Kingma and Dhariwal, 2018)	3.35	4.09	3.81
Flow++ (Ho et al., 2019)	3.08	3.86	3.69
Baseline (Ours)	**3.08**	**4.00**	**3.70**
MaxPoolFlow (Ours)	3.09	4.01	3.74

Results: Comparison with GANs

Model	Inception ↑	FID ↓
DCGAN*	6.4	37.1
WGAN-GP*	6.5	36.4
PixelCNN*	4.60	65.93
PixelIQN*	5.29	49.46
Baseline (Ours)	5.08	49.56
MaxPoolFlow (Ours)	**5.18**	**49.03**

Table 5: Inception score and FID for CIFAR-10.
*Results taken from Ostrovski et al. (2018).

References

- E. Hoogeboom, J. Peters, R. van den Berg, M. Welling. Integer discrete flows and lossless compression. NeurIPS, 2019.
- D. Nielsen, P. Jaini, E. Hoogeboom, O. Winther, M. Welling, SurVAE flows: Surjections to bridge the gap between VAEs and flows, NeurIPS, 2020.