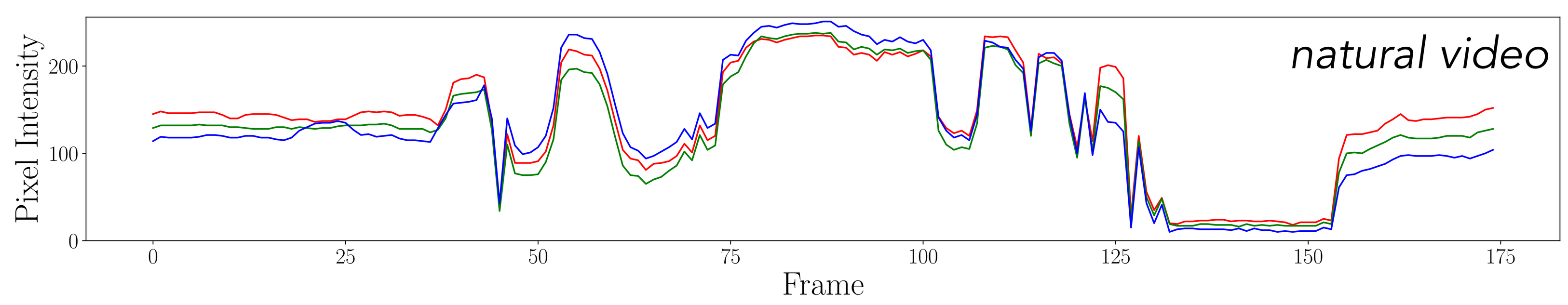
Improving Sequential Latent Variable Models with Autoregressive Flows

Joseph Marino[1], Lei Chen[2], Jiawei He[2], Stephan Mandt[3]

[1]California Institute of Technology, [2]Simon Fraser University, [3]UC Irvine

introduction

Sequences, e.g. video, often contain ***temporally redundant structure***. We would prefer to focus model capacity on learning the most complex temporal dependencies by first removing 'lower-level' correlations.



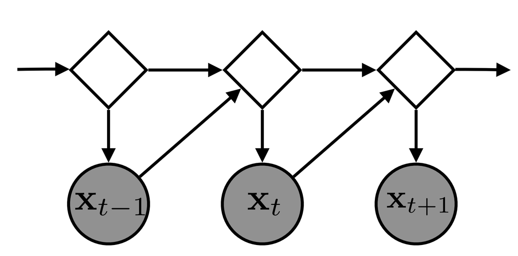
We formulate this procedure by applying affine autoregressive flows across time [Kingma et al., 2016; Papamakarios et al., 2017]. These flows act as a ***moving reference frame***, providing less correlated sequences to downstream modeling of more complex dependencies.

background

Autoregressive / Sequential Models

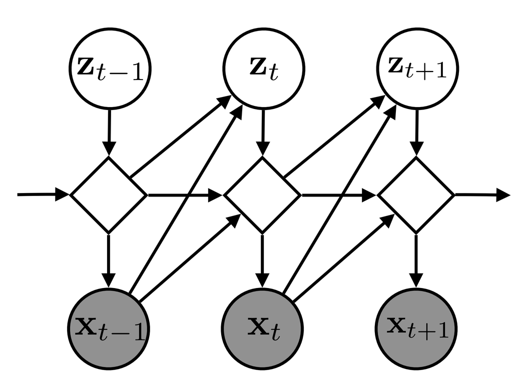
fully-visible model

$$p_\theta(\mathbf{x}_{1:T}) = \prod_{t=1}^{T} p_\theta(\mathbf{x}_t|\mathbf{x}_{<t})$$



latent variable model

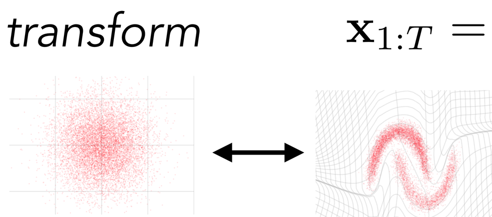
$$p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^{T} p_\theta(\mathbf{x}_t|\mathbf{x}_{<t}, \mathbf{z}_{\leq t}) p_\theta(\mathbf{z}_t|\mathbf{x}_{<t}, \mathbf{z}_{<t})$$



Normalizing Flows

base distribution $p_\theta(\mathbf{y}_{1:T})$

transform $\mathbf{x}_{1:T} = f_\theta(\mathbf{y}_{1:T})$



Dinh et al. 2017

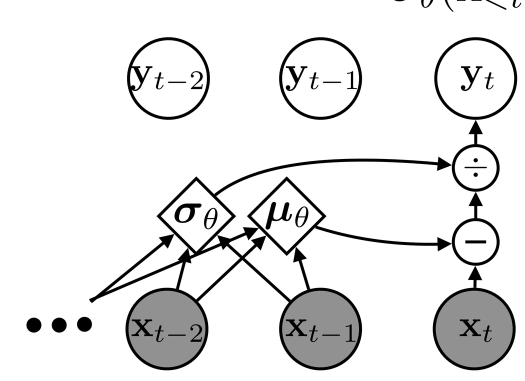
change of variables formula

$$p_\theta(\mathbf{x}_{1:T}) = p_\theta(\mathbf{y}_{1:T}) \left| \det \left(\frac{\partial \mathbf{x}_{1:T}}{\partial \mathbf{y}_{1:T}} \right) \right|^{-1}$$

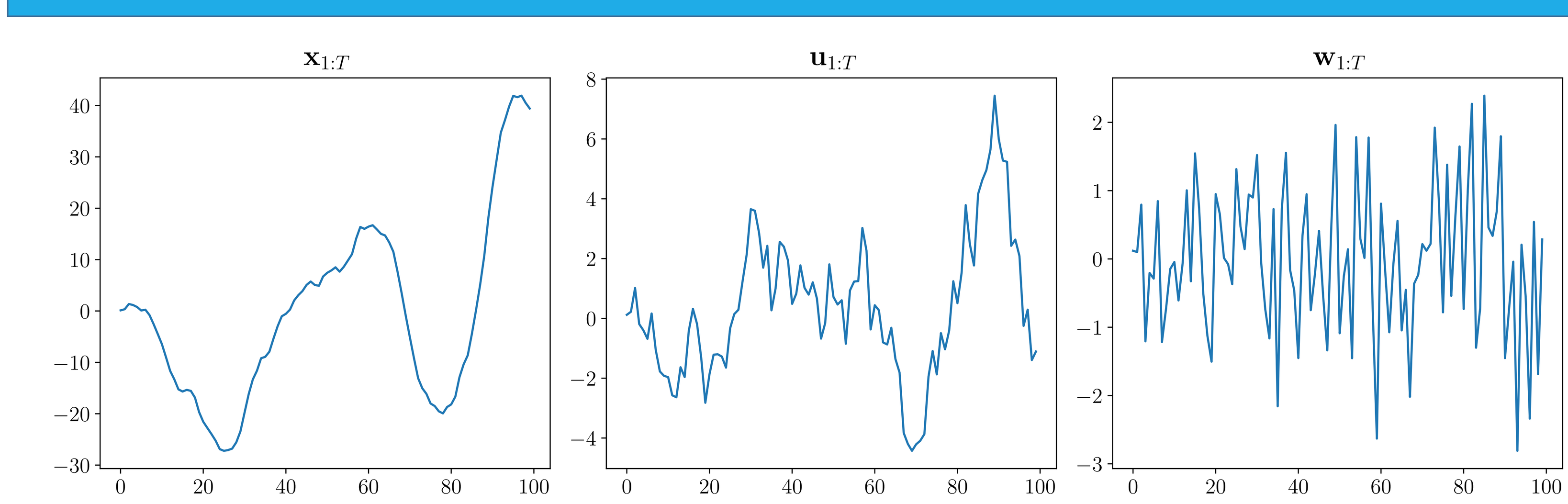
Autoregressive Normalizing Flows

forward $\mathbf{x}_t = \boldsymbol{\mu}_\theta(\mathbf{x}_{<t}) + \boldsymbol{\sigma}_\theta(\mathbf{x}_{<t}) \odot \mathbf{y}_t$

inverse $\mathbf{y}_t = \dfrac{\mathbf{x}_t - \boldsymbol{\mu}_\theta(\mathbf{x}_{<t})}{\boldsymbol{\sigma}_\theta(\mathbf{x}_{<t})}$



motivating example



high temporal dependence   *temporally independent*

linear dynamical system

position $\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{u}_t$

velocity $\mathbf{u}_t = \mathbf{u}_{t-1} + \mathbf{w}_t$

noise $\mathbf{w}_t \sim \mathcal{N}(\mathbf{w}_t; \mathbf{0}, \boldsymbol{\Sigma})$

$\longrightarrow$

$$\mathbf{u}_t \sim \mathcal{N}(\mathbf{u}_t; \mathbf{u}_{t-1}, \boldsymbol{\Sigma})$$

$$\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_t; \mathbf{x}_{t-1} + \mathbf{u}_{t-1}, \boldsymbol{\Sigma}) \sim \mathcal{N}(\mathbf{x}_t; \mathbf{x}_{t-1} + \mathbf{x}_{t-1} - \mathbf{x}_{t-2}, \boldsymbol{\Sigma})$$

$\mathbf{u}_t = \mathbf{x}_t - \mathbf{x}_{t-1}$ and $\mathbf{w}_t = \mathbf{u}_t - \mathbf{u}_{t-1}$ are special cases of $\dfrac{\mathbf{x}_t - \boldsymbol{\mu}_\theta(\mathbf{x}_{<t})}{\boldsymbol{\sigma}_\theta(\mathbf{x}_{<t})}$

$\partial \mathbf{x}_t / \partial \mathbf{u}_t = \mathbf{I}$ and $\partial \mathbf{u}_t / \partial \mathbf{w}_t = \mathbf{I}$, so from the change of variables formula,

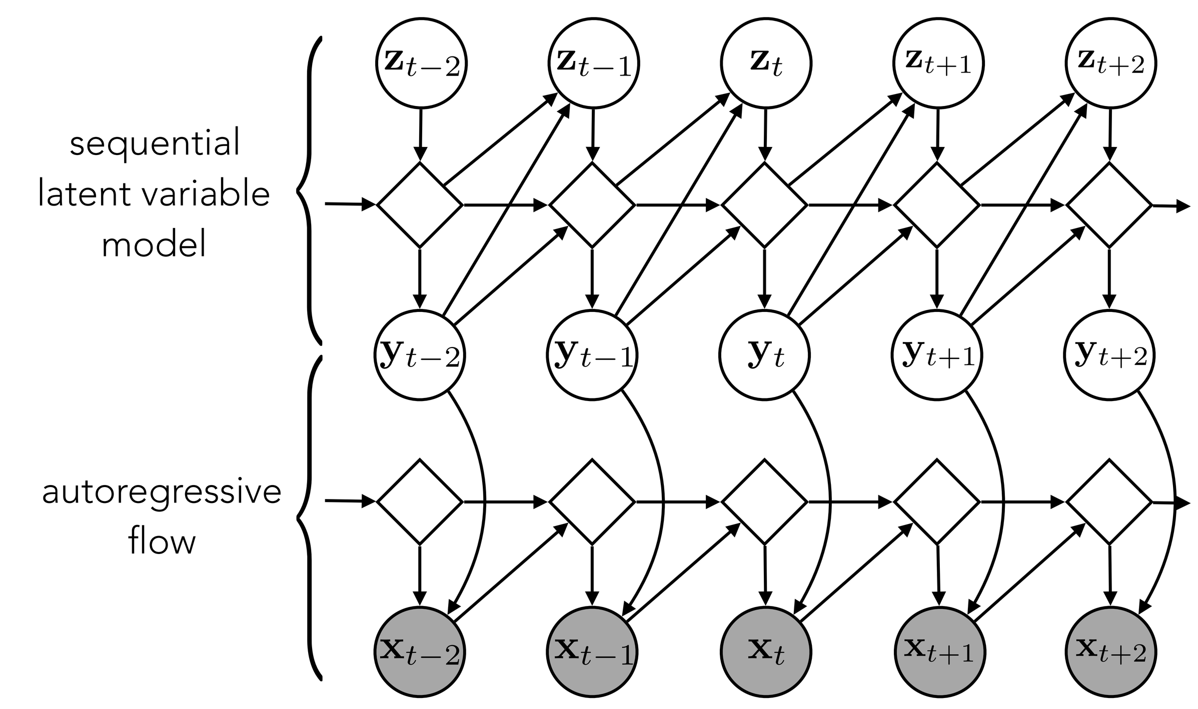
$$\underbrace{p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_{t-2})}_{\text{2nd order}} = \underbrace{p(\mathbf{u}_t|\mathbf{u}_{t-1})}_{\text{1st order}} = \underbrace{p(\mathbf{w}_t)}_{\text{independent}}$$

autoregressive flows on sequences

Seq. Latent Variable Model + Autoregressive Flow Conditional Likelihood

joint distribution $$p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = p_\theta(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}) \left| \det \left(\frac{\partial \mathbf{x}_{1:T}}{\partial \mathbf{y}_{1:T}} \right) \right|^{-1}$$

with $$p_\theta(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^{T} p_\theta(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{z}_{\leq t}) p_\theta(\mathbf{z}_t|\mathbf{y}_{<t}, \mathbf{z}_{<t})$$



Variational Inference

filtering approx. posterior $$q(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}) = \prod_{t=1}^{T} q(\mathbf{z}_t|\mathbf{x}_{\leq t}, \mathbf{z}_{<t})$$

evidence lower bound (ELBO)

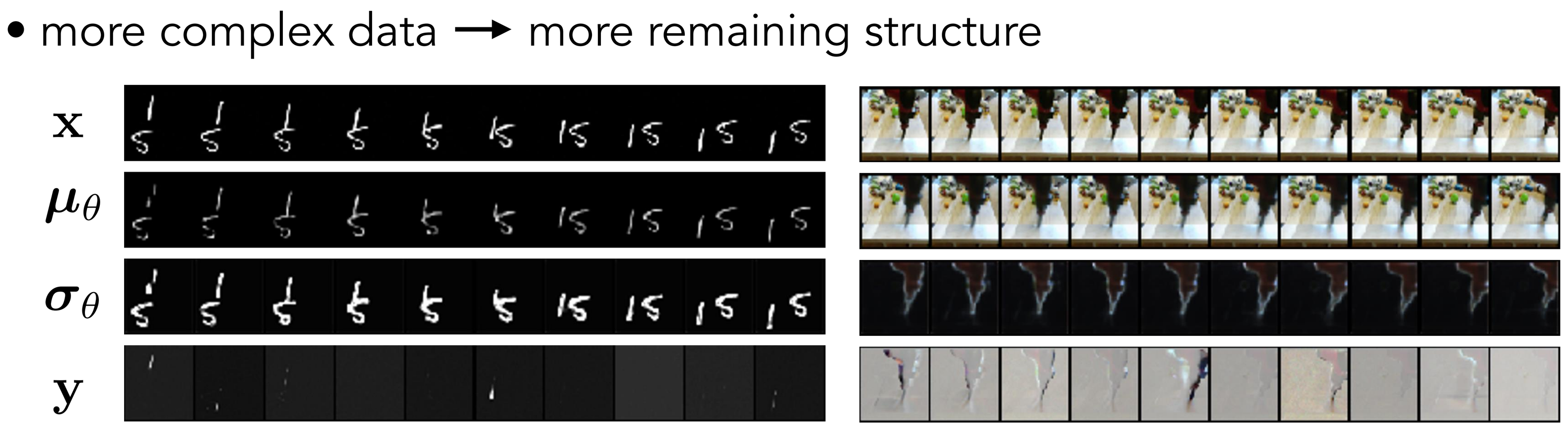
$$\log p_\theta(\mathbf{x}_{1:T}) \geq \mathcal{L}(\mathbf{x}_{1:T}; q, \theta) \equiv \mathbb{E}_{q(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})} \left[\log p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) - \log q(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}) \right]$$

$$\mathcal{L} = \sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{z}_{\leq t}|\mathbf{y}_{\leq t})} \left[\log p_\theta(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{z}_{\leq t}) - \log \frac{q(\mathbf{z}_t|\mathbf{y}_{\leq t}, \mathbf{z}_{<t})}{p_\theta(\mathbf{z}_t|\mathbf{y}_{<t}, \mathbf{z}_{<t})} - \log \left| \det \left(\frac{\partial \mathbf{x}_t}{\partial \mathbf{y}_t} \right) \right| \right]$$

results

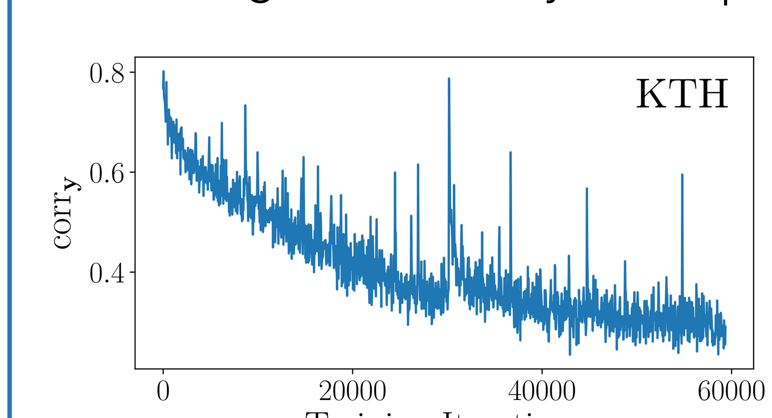
Visualizing Autoregressive Flows

- $\boldsymbol{\mu}_\theta$ is the prediction of the low-level model, $\boldsymbol{\sigma}_\theta$ expresses areas of uncertainty
- $\mathbf{y}$ captures any remaining spatiotemporal structure
- more complex data $\longrightarrow$ more remaining structure



Quantifying Temporal Decorrelation

- autoregressive flows yield sequences with reduced temporal correlation



temporal correlation in adjacent frames

	M-MNIST	BAIR	KTH
$\text{corr}_\mathbf{x}$	0.24	0.87	0.96
$\text{corr}_\mathbf{y}$	0.02	0.43	0.31

Improved Model Performance

- autoregressive flows (AF) perform well as standalone models and
- improve modeling performance of a sequential latent variable model (SLVM)

test negative log-likelihood in *nats / dim*

	M-MNIST	BAIR	KTH
1-AF	2.15	3.05	3.34
2-AF	2.13	2.90	3.35
SLVM	≤ 1.92	≤ 3.57	≤ 4.63
SLVM w/ 1-AF	$\leq \mathbf{1.86}$	$\leq \mathbf{2.35}$	$\leq \mathbf{2.39}$