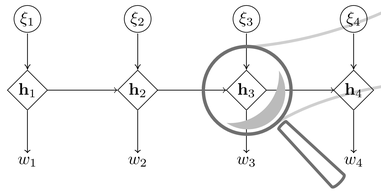
Deep State Space Models for Unconditional Word Generation

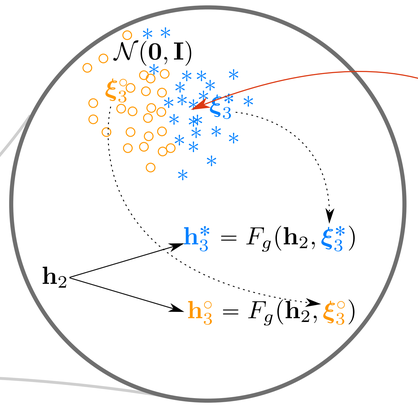
Florian Schmidt Thomas Hofmann
florian.schmidt@inf.ethz.ch

ETH zürich
Data Analytics Lab

Generative Model

$$P(\mathbf{w}) = \int \prod_{t=1}^{T} p(\mathbf{h}_t|\mathbf{h}_{t-1}) P(w_t|\mathbf{h}_t)\, d\mathbf{h}$$





Uncertainty about continuation

Express $p(\mathbf{h}_t|\mathbf{h}_{t-1})$ via transition function

$$\mathbf{h}_t = F_g(\mathbf{h}_{t-1}, \boldsymbol{\xi}_t), \quad \boldsymbol{\xi}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Simple generative model with a single source of noise and a deterministic transition function.

Objective

ELBO

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{h}|\mathbf{w})}\left[\sum_{t=1}^{T} \overbrace{\log P(w_t|\mathbf{h}_t)}^{\text{Reconstruction}} + \overbrace{\log \frac{p(\mathbf{h}_t|\mathbf{h}_{t-1})}{q(\mathbf{h}_t|\mathbf{h}_{t-1}, \mathbf{w}_{t:T})}}^{\text{Model missmatch}}\right]$$

+ Inference Flow

$$= \mathbb{E}_{q(\boldsymbol{\xi}|\mathbf{w})}\left[\sum_{t=1}^{T} \log P(w_t|F_q(\boldsymbol{\xi}_t)) + \log \frac{p(F_q(\boldsymbol{\xi}_t))}{q(\boldsymbol{\xi}|\mathbf{h}_{t-1}, \mathbf{w}_{t:T})} + \log \left|\det \mathbf{J}_{F_q}(\boldsymbol{\xi}_t)\right|\right]$$

+ Generative Flow

$$= \mathbb{E}_{q(\boldsymbol{\xi}|\mathbf{w})}\left[\sum_{t=1}^{T} \log P(w_t|F_q(\boldsymbol{\xi}_t)) + \log \frac{p(F_g{}^{-1}F_q(\boldsymbol{\xi}_t))}{q(\boldsymbol{\xi}|\mathbf{h}_{t-1}, \mathbf{w}_{t:T})} + \log \left|\det \mathbf{J}_{F_g{}^{-1}\circ F_q}(\boldsymbol{\xi}_t)\right|\right]$$

Variational lower-bound that exposes how the generative and the inference model both learn to relate noise to states via transition functions.

Inference Model

Factorizes like true posterior

$$q(\boldsymbol{\xi}|w) = \prod_{t=1}^{T} q(\boldsymbol{\xi}_t|\mathbf{h}_{t-1}, w_{t:T}) \qquad \text{with transition function} \qquad F_q : H \times \Xi \to H$$

where the future data $w_{t:T}$ is encoded by a standard RNN.

Extension: Importance weighted ELBO

- Draw K samples $\mathcal{Q} = \{\mathbf{h}_t^{(k)}\}_{k=1}^K \sim q(\mathbf{h}_t|\mathbf{h}_{t-1}, \mathbf{w}_{t:T})$. Effectively re-weights gradients:

$$\nabla\mathcal{L} = \mathbb{E}_{\mathcal{Q}}\left[\sum_{k=1}^{K} \frac{\omega^{(k)}}{\sum_{k'}\omega^{(k')}} \nabla \log \omega^{(k)}\right] \qquad \text{with} \qquad \omega^{(k)} = \frac{P(\mathbf{w}, \mathbf{h})}{q(\mathbf{h})}$$

- Here: $\omega_t^{(k)} = \underbrace{\frac{P(w_t, \mathbf{h}_t^{(k)}|\mathbf{h}_{t-1})}{q(\mathbf{h}_t^{(k)}|\mathbf{h}_{t-1}, \mathbf{w}_{t:T})}}_{\substack{\text{immediate} \\ \text{ELBO-cost}}} \underbrace{\mathbb{E}_{q(\mathbf{h}_{t+1:T}|\mathbf{h}_t^{(k)})}\left[\frac{P(w_{t+1:T}, \mathbf{h}_{t+1:T}|\mathbf{h}_t^{(k)})}{q(\mathbf{h}_{t+1:T}|\mathbf{h}_t^{(k)}, w_{t+1:T})}\right]}_{\text{future ELBO-cost}}$
- Condition q on F_g

  $q(\mathbf{h}_t|\mathbf{h}_{t-1}, \mathbf{w}_{t:T}) \quad \to \quad q(\mathbf{h}_t|\mathbf{h}_{t-1}, \mathbf{h}_g^{(k)}, \mathbf{w}_{t:T}) \quad \text{where} \quad \mathbf{h}_g^{(k)} = F_G(\mathbf{h}_{t-1}, \boldsymbol{\xi}^{(k)}), \boldsymbol{\xi}^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

  Stabilizes training by informing q about the multimodality/variance of F_g.

The inference model proposes a state trajectory by providing data-informed noise samples given the future observations.

A more advanced version proposes states iteratively, by "simulating" the generative model in each step.

Results

Flows

1. Triangular TRIL with special case DIAG
   $$F(\mathbf{h}_{t-1}, \boldsymbol{\xi}_t) = g(\mathbf{h}_{t-1}) + \mathbf{G}(\mathbf{h}_{t-1})\boldsymbol{\xi}_t$$
   $g : \mathbb{R}^d \to \mathbb{R}^d$ MLP
   $\mathbf{G} : \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$ MLP s.t. $\mathbf{G}(\mathbf{h})$ is lower triangular
2. RealNVP
3. Complex flows via composition $F = f_1 \circ \cdots \circ f_n$

Sequence Cross Entropy

Model	$H[P_{\text{train}}, \hat{P}]$	$H[P_{\text{test}}, \hat{P}]$	$\mathbf{w} \in V$ unique	$\mathbf{w} \in V$	$\bar{I}$
TRIL	12.13±.11	11.99±.11	0.18±.00	0.43±.03	0.95±.04
TRIL, K=2	11.76±.12	11.82±.12	0.16±.01	0.46±.02	1.06±.16
TRIL, K=5	11.46±.05	11.51±.05	0.16±.01	0.48±.02	1.08±.13
TRIL, K=10	11.43±.05	11.47±.05	0.16±.01	0.49±.02	1.12±.12
2×TRIL	11.91±.08	11.86±.13	0.17±.01	0.45±.02	0.89±.07
2×TRIL, K=2	11.55±.09	11.61±.09	0.16±.00	0.47±.01	1.00±.13
2×TRIL, K=5	11.42±.07	11.46±.06	0.16±.00	0.49±.01	1.20±.12
2×TRIL, K=10	**11.33±.05**	**11.38±.06**	0.16±.00	0.49±.01	**1.28±.13**
2×TRIL, K=10, BIDI	**11.33±.09**	11.39±.10	0.16±.01	0.48±.00	1.25±.16
$d = 16$ 2×TRIL, K=10	11.21	11.43	0.15	0.48	1.43
$d = 32$ 2×TRIL, K=10	11.27	11.13	0.15	0.50	1.31
REAL-NVP-[2,3,4,5,6,7]	11.77	11.81	0.12	0.53	0.94
BASELINE-5D	12.92	12.97	0.13	0.53	–
BASELINE-16D	12.55	12.60	0.14	0.62	–
ORACLE-TRAIN	7.0	7.02[1]	0.27	1.0	–

Generative vs. Inference Flow

Flow F_g \ Flow F_q	ID	DIAG	TRIL	2×TRIL
ID	14.23±.00	14.23±.00	14.23±.00	–
DIAG	12.82±.37	12.35±.37	12.20±.25	–
TRIL	13.55±.01	11.99±.11	–	–
2×TRIL	–	11.86±.13	–	–

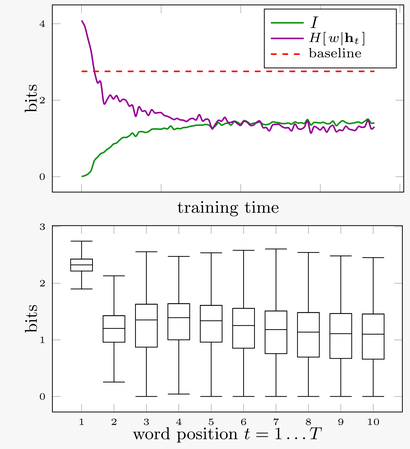
(a) Test cross entropy $H[P_{\text{test}}, \hat{P}]$

Flow F_g \ Flow F_q	ID	DIAG	TRIL	2×TRIL
ID	0±.00	0±.00	0±.00	–
DIAG	0.93±.15	0.85±.16	0.92±.13	–
TRIL	0.65±.01	0.95±.04	–	–
2×TRIL	–	0.89±.07	–	–

(b) Average mutual information $\bar{I}$

Noise Mutual Information

$$I(t) = I(w_t; \boldsymbol{\xi}_t|\mathbf{h}_{t-1}) = \mathbb{E}_{\mathbf{h}_{t-1}}\left[H[w_t|\mathbf{h}_{t-1}] - H[w_t|\boldsymbol{\xi}_t, \mathbf{h}_{t-1}]\right] \geqslant 0$$



BGS15 Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov *Importance Weighted Autoencoders*, CoRR 2015

DSB16 Laurent Dinh and Jascha Sohl-Dickstein and Samy Bengio *Density estimation using Real NVP*, ICLR 2016

RM15 Danilo Jimenez Rezende and Shakir Mohamed *Variational Inference with Normalizing Flows*, ICML 2015