Learning Rational Probability Distributions with Bisequential Variational Autoencoders

Georgi Shopov

Probabilistic Models

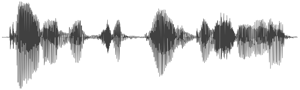
L'essentiel est invisible pour les yeux

French sentence $\mathcal{S}$

$\xrightarrow{\mathbb{P}_{\mathcal{S}}}$

The essential is invisible to the eyes

English translation



Utterance $\mathcal{U}$

$\xrightarrow{\mathbb{P}_{\mathcal{U}}}$

There's no smoke without fire

Transcription

A delicate and precise watercolor illustration ...

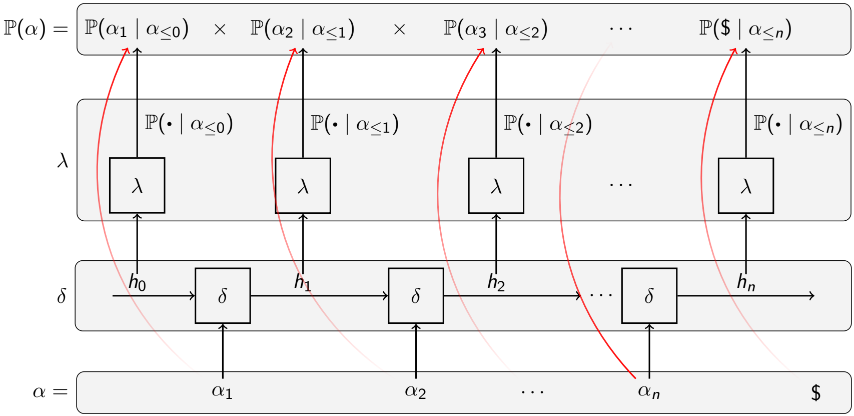
Text description $\mathcal{D}$

$\xrightarrow{\mathbb{P}_{\mathcal{D}}}$

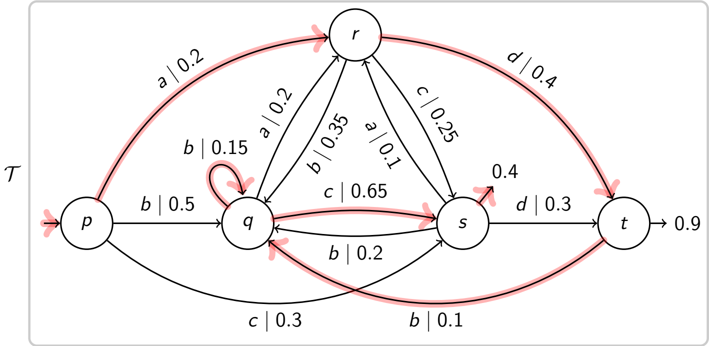
Image

Recurrent Probabilistic Models

$h_0 \in \mathbb{R}^d$ $\delta \colon \mathbb{R}^d \times \Sigma \to \mathbb{R}^d$ $\lambda \colon \mathbb{R}^d \to \Delta^{|\Sigma|}$



Sequential Transducers



$$[\![\mathcal{T}]\!](adbbc) = 0.2 \times 0.4 \times 0.1 \times 0.15 \times 0.65 \times 0.4$$

Sequential transducers $\longrightarrow$ sequential functions

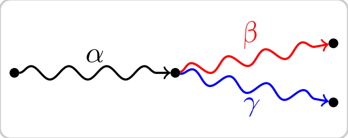
Arbitrary transducers $\longrightarrow$ rational relations

Expressive Power of Sequential Transducers

Definition

For $\alpha, \beta, \gamma \in \Sigma^*$ such that $\beta \wedge \gamma = \epsilon$, the ***prefix distance*** between $\alpha\beta$ and $\alpha\gamma$ is defined as

$$d_p(\alpha\beta, \alpha\gamma) = |\beta| + |\gamma|.$$



Theorem (Mohri [3])

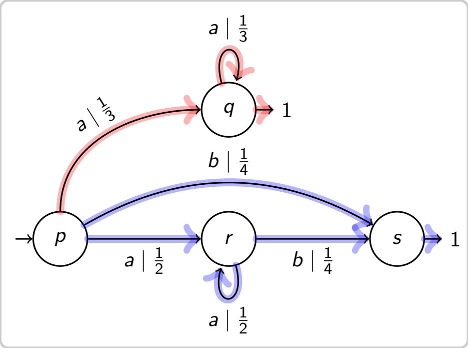
A rational probability distribution $\mathbb{P}$ over Σ^* is sequential if and only if

$$\left\{ \frac{\mathbb{P}(\alpha)}{\mathbb{P}(\beta)} \;\middle|\; \alpha, \beta \in \mathrm{Supp}(\mathbb{P}) \;\&\; d_p(\alpha, \beta) \leq n \right\}$$

is finite for all $n \in \mathbb{N}$.

Limitations of Sequential Transducers

$$\mathbb{P}(\alpha) = \begin{cases} \left(\frac{1}{3}\right)^{n+1} & \text{if } \alpha = a^{n+1} \\ \left(\frac{1}{2}\right)^{n+2} & \text{if } \alpha = a^n b \\ 0 & \text{otherwise} \end{cases}$$

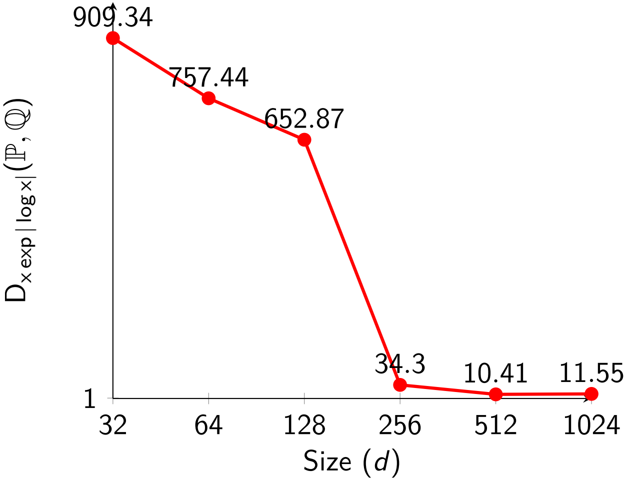


$\mathbb{P}$ is rational but not sequential because for every $n \in \mathbb{N}$,

$$d_p(a^{n+1}, a^n b) = 2 \quad \text{and} \quad \frac{\mathbb{P}(a^n b)}{\mathbb{P}(a^{n+1})} = \frac{\left(\frac{1}{2}\right)^{n+2}}{\left(\frac{1}{3}\right)^{n+1}} = \frac{1}{2}\left(\frac{3}{2}\right)^{n+1}.$$

Limitations of Recurrent Probabilistic Models

$$\mathsf{D}_{\mathsf{x}\exp|\log \mathsf{x}|}(\mathbb{P},\mathbb{Q}) = \mathbb{E}_{\alpha\sim\mathbb{P}}\left[\exp\left|\log\frac{\mathbb{P}(\alpha)}{\mathbb{Q}(\alpha)}\right|\right] = \mathbb{E}_{\alpha\sim\mathbb{P}}\left[\frac{\max\left\{\mathbb{P}(\alpha),\mathbb{Q}(\alpha)\right\}}{\min\left\{\mathbb{P}(\alpha),\mathbb{Q}(\alpha)\right\}}\right]$$

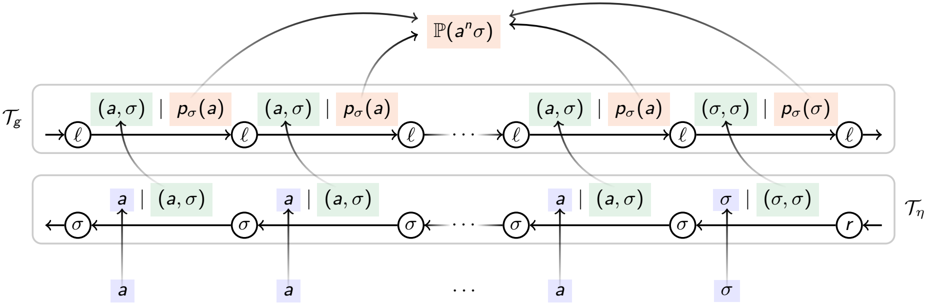


Bisequential Decompositions

Definition

A ***bisequential decomposition*** of a probability distribution $\mathbb{P}$ over Σ^* is a tuple (Γ, η, g), where

- Γ is a ***latent alphabet***;
- $\eta\colon \Sigma^* \to (\Sigma \times \Gamma)^*$ is a co-sequential function s.t. $\eta \circ \pi_{\Sigma^*} = \mathrm{id}_{\Sigma^*}$;
- $g\colon (\Sigma \times \Gamma)^* \to [0, 1]$ is a sequential probability distribution;
- $\mathbb{P} = \eta \circ g$.



Expressive Power of Bisequential Decompositions

Theorem (Elgot and Mezei [1])

A probability distribution is rational if and only if it admits a bisequential decomposition.

Theorem (Shopov and Gerdjikov [5])

A probability distribution $\mathbb{P}$ *over* Σ^* *is rational if and only if there exists a finite partition* $\{L_i\}_{i=1}^n$ *of* Σ^* *such that, for* $1 \leq i \leq n$, L_i *is regular and* $\left\{\mathbb{P}(\,\cdot\,\alpha \mid L_i\alpha)\right\}_{\alpha\in\Sigma^*}$ *is finite. In this case,*

$$\alpha \sim_i \beta \iff \mathbb{P}(\,\cdot\,\alpha \mid L_i\alpha) = \mathbb{P}(\,\cdot\,\beta \mid L_i\beta)$$

is a left congruence and the latent alphabet can be chosen to be

$$\Gamma = \left\{ \left(\mathbb{P}(\,\cdot\,\alpha \mid L_i\alpha)\right)_{i=1}^n \,\middle|\, \alpha \in \Sigma^* \right\}.$$

Latent Variable Models

- A latent variable model specifies a joint distribution

$$p_\theta(x, z)$$

over observed variables (data) $x \in \mathcal{X}$ and latent variables $z \in \mathcal{Z}$.

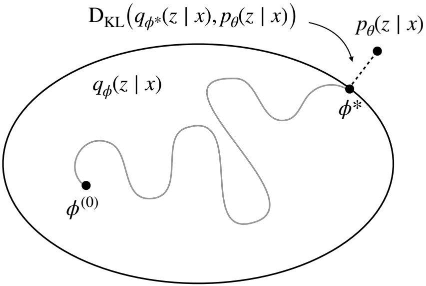
- The latent variables explain the ***hidden structure used to generate the data***.
- However, the marginal distribution over the data

$$p_\theta(x) = \int_{z \in \mathcal{Z}} p_\theta(x, z) dz$$

is often intractable.

Evidence Lower Bound

$$
\begin{aligned}
\log p_\theta(x) &= \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x)\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)\, q_\phi(z \mid x)}{p_\theta(z \mid x)\, q_\phi(z \mid x)}\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z \mid x)}\right] + \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{q_\phi(z \mid x)}{p_\theta(z \mid x)}\right] \\
&= \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z \mid x)}\right]}_{\text{Evidence Lower Bound}} + \underbrace{\mathrm{D}_{\mathrm{KL}}\big(q_\phi(z \mid x), p_\theta(z \mid x)\big)}_{\geq 0}
\end{aligned}
$$



Variational Autoencoders

Definition (Kingma and Welling [2], Rezende et al. [4])

A ***variational autoencoder*** is a pair (ϕ, θ) of parameters defining

- a variational encoder $q_\phi(z \mid x)$, and
- a generative model $p_\theta(x, z) = p_\theta(x \mid z)p_\theta(z)$

that are optimised by maximising the evidence lower bound

$$\begin{aligned}
\mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z \mid x)}\right] &= \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x \mid z)p_\theta(z)}{q_\phi(z \mid x)}\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] + \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(z)}{q_\phi(z \mid x)}\right] \\
&= \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right]}_{\text{Reconstruction term}} - \underbrace{\mathrm{D}_{\mathrm{KL}}\left(q_\phi(z \mid x), p_\theta(z)\right)}_{\text{Regularisation term}}.
\end{aligned}$$

Bisequential Variational Autoencoders

Definition

A variational autoencoder is called ***bisequential*** if

- $q_\phi(z \mid x) = \prod_{i=1}^{n} q_\phi(z_i \mid x_{\geq i})$, and
- $p_\theta(x, z) = \prod_{i=1}^{n} p_\theta(z_i \mid x_{<i}, z_{<i}) p_\theta(x_i \mid x_{<i}, z_{\leq i})$.

Theorem

The evidence lower bound of a bisequential variational autoencoder (ϕ, θ) *can be expressed as*

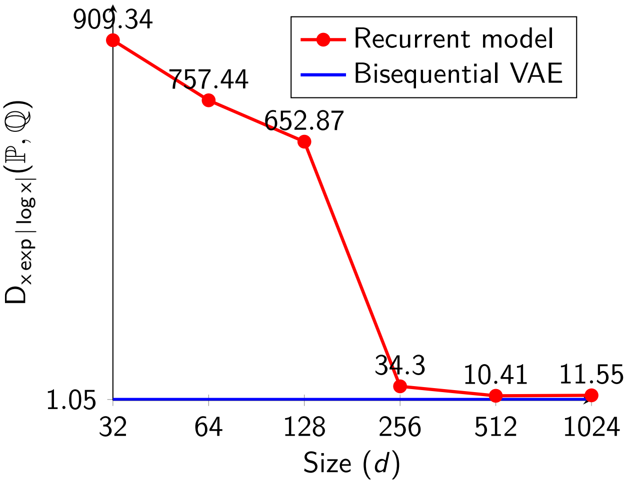
$$\mathbb{E}_{q_\phi(z|x)}\left[\sum_{i=1}^{n} \underbrace{\log p_\theta(x_i \mid x_{<i}, z_{\leq i})}_{\textit{Reconstruction term}} - \underbrace{\mathrm{D}_{\mathrm{KL}}\big(q_\phi(z_i \mid x_{\geq i}), p_\theta(z_i \mid x_{<i}, z_{<i})\big)}_{\textit{Regularisation term}}\right].$$

Furthermore, when approximating a rational probability distribution, this lower bound will be tight.

Expressive Power of Bisequential VAEs

Theorem

Every rational probability distribution can be represented by a bisequential variational autoencoder.



Conclusion

We demonstrated how automata theory can be used to:

- ***Characterise the expressive power of recurrent models.***
- ***Identify key limitations*** – in particular, their inability to represent ***non-sequential rational probability distributions***.
- ***Motivate more expressive architectures***, such as ***Bisequential VAEs***, which overcome these limitations and can model ***the full class of rational probability distributions***.

References I

[1] C. C. Elgot and J. E. Mezei. On relations defined by generalized finite automata. *IBM Journal of Research and Development*, 9(1): 47–68, 1965.

[2] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *The Second International Conference on Learning Representations*, 2014.

[3] Mehryar Mohri. Finite-state transducers in language and speech processing. *Computational Linguistics*, 23(2):269–311, 1997.

[4] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the Thirty-First International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286. PMLR, June 2014.

References II

[5] Georgi Shopov and Stefan Gerdjikov. Consistent bidirectional language modelling: Expressive power and representational conciseness. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5724–5768. Association for Computational Linguistics, November 2024.