REDWOOD CENTER for Theoretical Neuroscience
Berkeley Herbert Wertheim School of Optometry & Vision Science

Poisson Variational Autoencoder

Hadi Vafaii[1,2], Dekel Galor[3], Jacob L. Yates[2,4]
@hadivafaii @dekelgalor @jcbyts

(1) Redwood Center for Theoretical Neuroscience, (2) Herbert Wertheim School of Optometry & Vision Science, (3) Electrical Engineering and Computer Sciences (4) Helen Wills Neuroscience Institute; UC Berkeley, CA, 94720, USA

Code (GitHub) Read the Paper

Conjecture

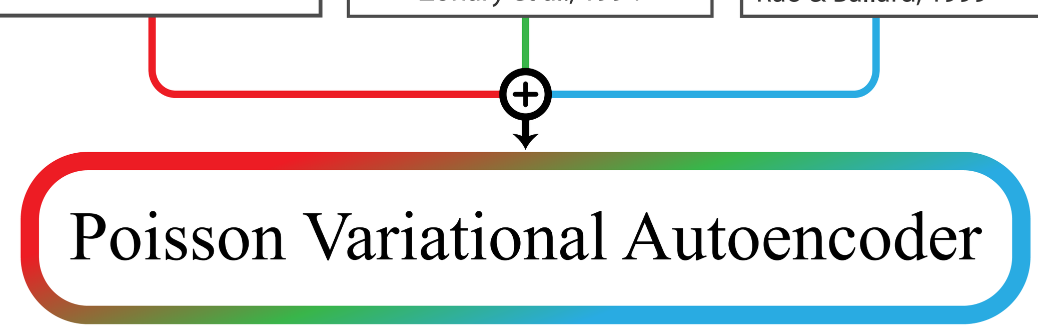
We are going to "understand" the brain through the study of brain-like artificial neural networks (ANN)

→ ***Challenge*** : design ANNs that exhibit brain-like structure and function

Contributions

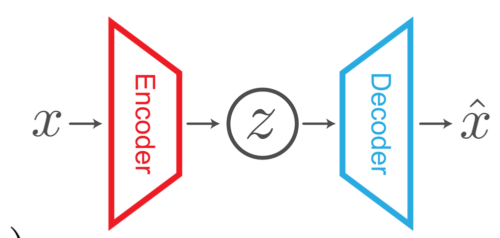
We introduce the "*Poisson Variational Autoencoder*" (P-VAE), a novel architecture that draws inspiration from centuries of neuroscience research and links them with modern machine learning

Perception as Inference	Rate Coding	Predictive Coding
Alhazen, ~1000 AD; Helmholtz, 1860s; Lee & Mumford, 2003; Barlow, 2003; Knill & Pouget, 2004; Yuille & Kersten, 2006; Friston, 2010	Adrian & Zotterman, 1926; Perkel & Bullock, 1968; Barlow, 1972; Zohary et al., 1994	Neisser, 1967; Gregory, 1980; Srinivasan et al., 1982; Friston, 2005; Rao & Ballard, 1999; Clark, 2013



Background: Variational Autoencoders (VAE)

Posterior inference: which set of latents z are likely given a data sample x?



True (unknown) posterior: $p(z|x) \propto p(x|z)p(z)$

Approx. posterior q that minimizes: $\mathcal{D}_{\mathrm{KL}}\big(q(z|x)\,\|\,p(z|x)\big) \Rightarrow \cdots \Rightarrow$

$$\Rightarrow \mathcal{L}_{\mathrm{VAE}}(q) = -\mathbb{E}_{z\sim q(z|x)}\big[\log p(x|z)\big] + \mathcal{D}_{\mathrm{KL}}\big(q(z|x)\,\|\,p(z)\big)$$ VAE Loss

Most **common** parameterization of VAEs: **Gaussian** latent space $\begin{cases} p(z) = \mathcal{N}(z; \mathbf{0}, \mathbf{1}) \\ q(z|x) = \mathcal{N}(z; \mu(x), \sigma^2(x)) \end{cases}$ $\{\mu(x), \sigma^2(x)\}$ are produced by the encoder network

Gaussian latents: Continuous ✗ Unconstrained (both +/- values) ✗

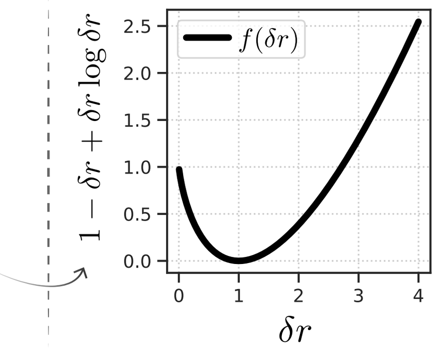
The brain: Uses **rate** coding, encodes inputs into **discrete** spike counts ✓

P-VAE: a VAE with Poisson-distributed latents

Poisson-distributed latent space $\begin{cases} p(z) = \mathcal{P}\mathrm{ois}(z; r_{\mathrm{prior}}) \\ q(z|x) = \mathcal{P}\mathrm{ois}(z; r_{\mathrm{post.}}(x)) \end{cases}$ + Predictive coding assumption $\begin{cases} r_{\mathrm{prior}} \to r \text{ (interpretation: representation units)} \\ r_{\mathrm{post.}} \to r \odot \delta r(x) \text{ residual parameterization} \end{cases}$

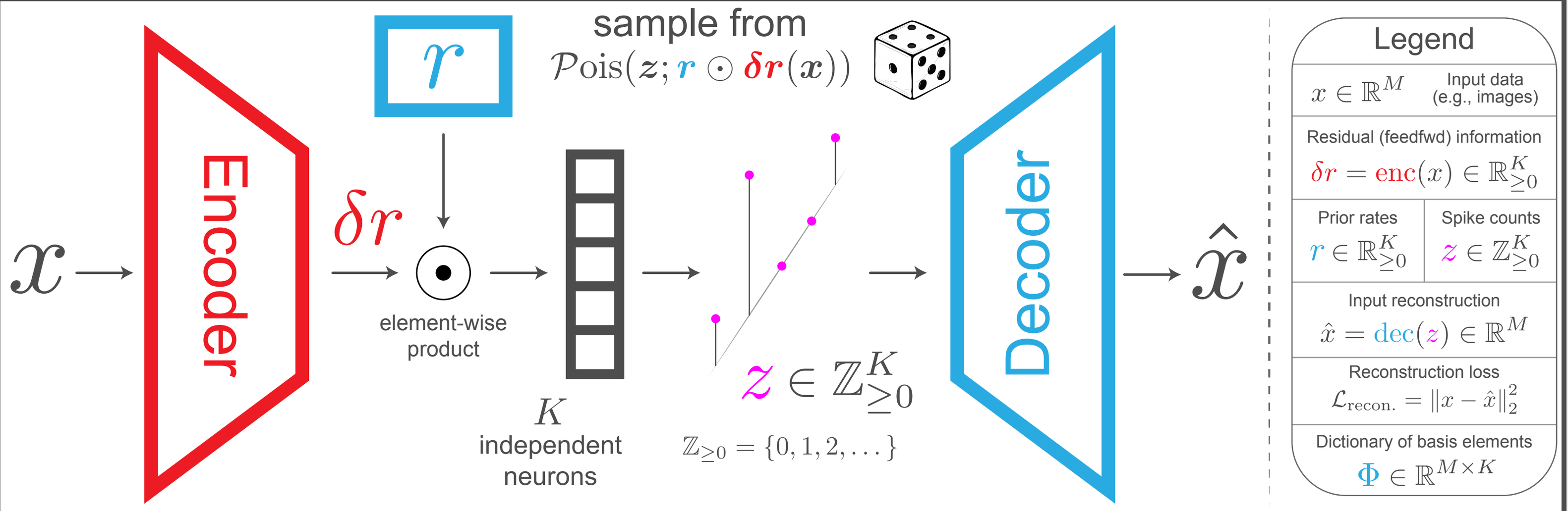
Poisson KL-term has closed-form solution!

$$\mathcal{D}_{\mathrm{KL}}\big(\mathcal{P}\mathrm{ois}(z; r\odot\delta r)\,\|\,\mathcal{P}\mathrm{ois}(z; r)\big) = \sum_{i=1}^{K} r_i f(\delta r_i)$$



(see paper appendix for the full derivation)

The P-VAE: Architecture & Objective



$$\Rightarrow \mathcal{L}_{\mathcal{P}\mathrm{VAE}} = \mathbb{E}_{z\sim\mathcal{P}\mathrm{ois}(z; r\delta r)}\Big[\|x - \mathrm{dec}(z)\|_2^2\Big] + \sum_{i=1}^{K} r_i f(\delta r_i)$$ P-VAE Loss

Reconstruction term "***Metabolic cost***" term

where we have $f(y) = 1 - y + y\log y$

Linear P-VAE = Amortized Sparse Coding

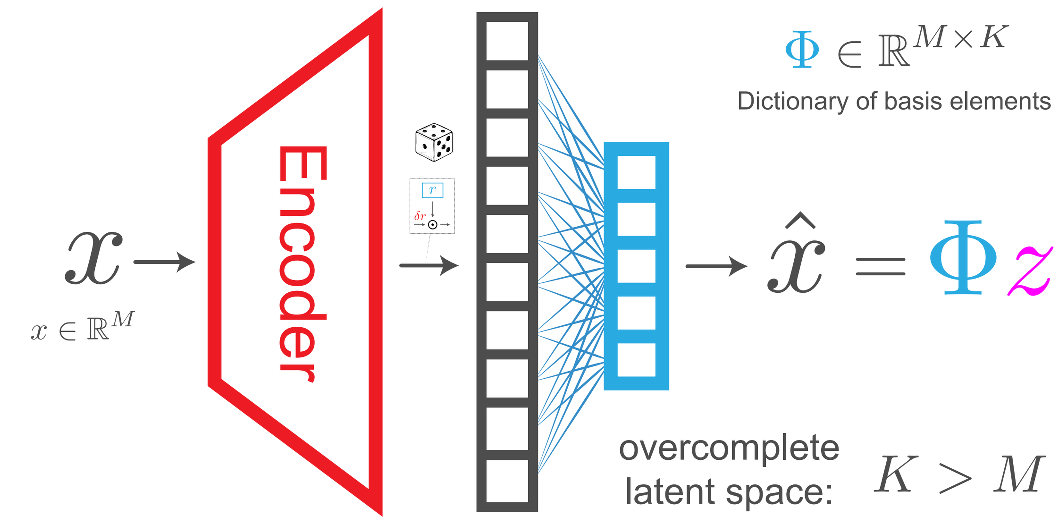
The **metabolic cost term** in the objective is reminiscent of **sparse coding**. To make this connection more explicit, we adopt additional assumptions commonly made in sparse coding (SC):

P-VAE with:

(i) **linear** decoder

(ii) **overcomplete** latent space

(typical sparse coding (SC) assumptions [6])



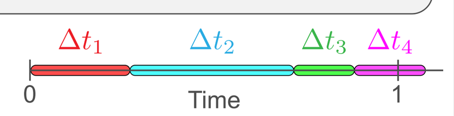
$$\mathcal{L}_{\text{SC-PVAE}}(x; \delta r, r, \Phi) = \|x - \Phi\lambda\|_2^2 + \lambda^T \mathrm{diag}(\Phi^T\Phi) + \beta\sum_{i=1}^{K} r_i f(\delta r_i)$$

posterior rates $\lambda := r \odot \delta r(x)$

Poisson Reparameterization algorithm

Technical challenge: gradient descent with discrete, stochastic latents

Intuition: think about the *Poisson process* generating the spike count samples



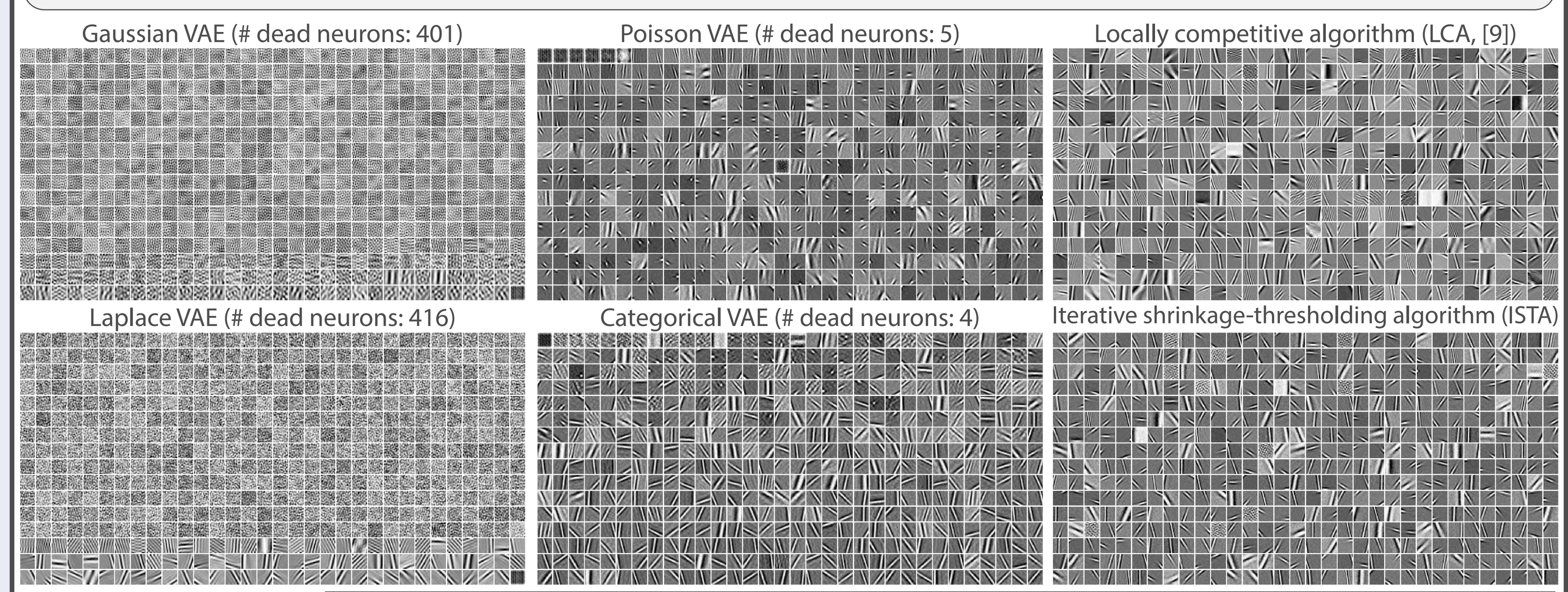
Algorithm 1 Reparameterized sampling (rsample) for Poisson distribution.

Input:
$\boldsymbol{\lambda} \in \mathbb{R}^{B\times K}_{>0}$ # rate parameter; B, batch size; K, latent dimensionality
n_exp # number of exponential samples to generate
temperature # controls the sharpness of the thresholding

procedure RSAMPLE($\boldsymbol{\lambda}$, n_exp, temperature)
  Exp ← Exponential($\boldsymbol{\lambda}$) ▷ create exponential distribution
  Δt ← Exp.rsample((n_exp,)) ▷ sample inter-event times, $\Delta t : [n\_\mathrm{exp} \times B \times K]$
  times ← cumsum(Δt, dim=0) ▷ compute arrival times, same shape as Δt
  indicator ← sigmoid$\left(\frac{1-\mathrm{times}}{\mathrm{temperature}}\right)$ ▷ soft indicator for events within unit time
  z ← sum(indicator, dim=0) ▷ event counts, or number of spikes, $z : [B \times K]$
  return z
end procedure

The P-VAE develops Gabor-like feature selectivity

Trained on 16 x 16 natural image patches [8], 512 dictionary elements $\Rightarrow \Phi \in \mathbb{R}^{256\times 512}$

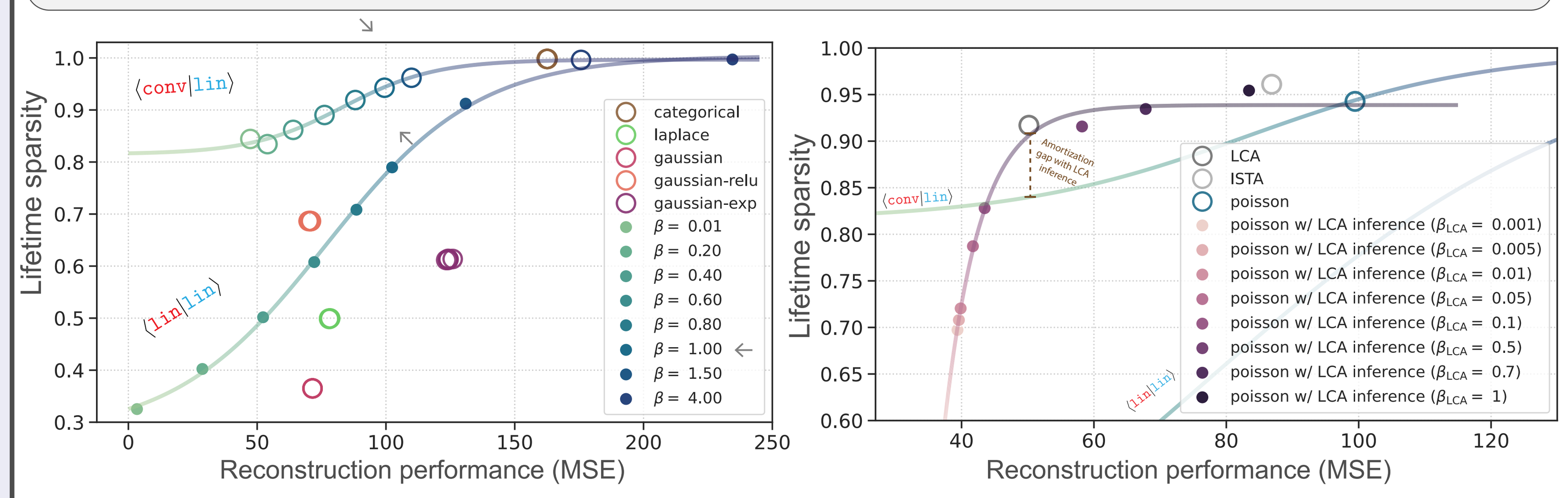


The P-VAE largely avoids the posterior collapse issue

Model	van Hateren		$\text{CIFAR}_{16\times16}$		MNIST	
	linear	conv	linear	conv	linear	conv
$\mathcal{P}$-VAE	**0.984**±.011	**0.819**±.041	**0.999**±.002	**0.928**±.045	**0.537**±.008	**0.426**±.011
$\mathcal{L}$-VAE	0.188±.000	0.222±.003	0.193±.003	0.230±.000	0.027±.000	0.034±.002
$\mathcal{G}$-VAE	0.218±.003	0.246±.000	0.105±.008	0.246±.000	0.027±.000	0.031±.000

The P-VAE learns sparse representations

Rate-distortion curves [10]: rate ↔ sparsity, distortion ↔ recon. (MSE)



Locally Competetive Algorithm (LCA, [7]), a "shallow" model from 2008, outperforms heavily-parameterized P-VAE with a convlutional encoder

inference suboptimality: **iterative > amortized**

Conclusions and future work

- We introduced the P-VAE, an architecture that draws inspiration from well-established concepts in neurosceice, and integrates them with modern machine learning.
- The P-VAE encodes its inputs in **discrete spike counts**, thus, it is one step closer to the brain (but still work in progress).
- A **metabolic cost term** emerges in the P-VAE loss "for free," suggesting a connection to **sparse coding**, which we verify.
- What we're *really* excited about, a **hierarchical P-VAE** with linear decoders all the way down:
  - *A sweet spot between interpretable/expressive?*

References

[1] Higgins et al., Nature Communications (2021)
[2] Vafaii, Yates, & Butts, NeurIPS (2023)
[3] Helmholtz, Treatise on Physio. Optics (1860s)
[4] Adrian & Zotterman, J Physiol. (1926)
[5] Rao & Ballard, Nature Neuroscience (1999)
[6] Olshausen & Field, Nature (1996)
[7] Rozell et al., Neural Computation (2008)
[8] Kingma & Welling, ICLR (2014)
[9] van Hateren & van der Schaaf, Proc. R. Soc. B (1998)
[10] Alemi et al., ICML (2018)

Contact info
@hadivafaii
vafaii@berkeley.edu

Acknowledgements
NIH R00EY032179