

#### Lecture 9:

# Variational Autoencoders & **Lossy Neural Compression**

Robert Bamler • Summer Term of 2023

These slides are part of the course "Data Compression With and Without Deep Probabilistic Models" taught at University of Tübingen. More course materials—including video recordings, lecture notes, and problem sets with solutions—are publicly available at https://robamler.github.io/teaching/compress23/.

### Recall: Variational Inference





- Idea:
  - $\blacktriangleright$  approximate the (inaccessible) true posterior  $P(\mathbf{Z} | \mathbf{X} = \mathbf{x})$  with a variational distribution  $Q_{\phi}(\mathbf{Z})$ .
  - Find the best approximation  $\phi^* := \arg \max_{\phi} \mathsf{ELBO}(\phi, \mathbf{x})$ .
- ► Evidence Lower Bound:  $|ELBO(\phi, \mathbf{x}) = \mathbb{E}_{Q_{\phi}(\mathbf{Z})}[\log P(\mathbf{Z}, \mathbf{X} = \mathbf{x}) \log Q_{\phi}(\mathbf{Z})]$ 
  - negative expected net bit rate of bits-back coding:  $ELBO(\phi, \mathbf{x}) = -\mathbb{E}_{\mathbf{s}}[R_{\phi}^{net}(\mathbf{x} \mid \mathbf{s})]$
  - **b** bound on the evidence:  $\mathsf{ELBO}(\phi, \mathbf{x}) = \log P(\mathbf{X} = \mathbf{x}) D_{\mathsf{KL}}(Q_{\phi}(\mathbf{Z}) \parallel P(\mathbf{Z} \mid \mathbf{X} = \mathbf{x})) \leq \log P(\mathbf{X} = \mathbf{x})$
  - regularized maximum likelihood: ELBO $(\phi, \mathbf{x}) = \mathbb{E}_{Q_{\phi}(\mathbf{Z})} \left[ \log P(\mathbf{X} = \mathbf{x} \mid \mathbf{Z}) \right] D_{\mathsf{KL}} \left( Q_{\phi}(\mathbf{Z}) \parallel P(\mathbf{Z}) \right)$
  - **today:** rate/distortion-tradeoff:  $|\mathsf{ELBO}_{\beta}(\phi, \mathbf{x}) = \mathbb{E}_{Q_{\phi}(\mathbf{Z})}[\log P(\mathbf{X} = \mathbf{x} \mid \mathbf{Z})] \beta D_{\mathsf{KL}}(Q_{\phi}(\mathbf{Z}) \parallel P(\mathbf{Z}))$ (actually, next neet (i))

#### Problems:

- $\blacktriangleright$  What's the generative model  $P(\mathbf{Z}, \mathbf{X})$ ?  $\longrightarrow$  variational expectation maximization
- $\blacktriangleright$  Expensive "arg max<sub> $\phi$ </sub>" for each message **x** in both encoder & decoder.  $\longrightarrow$  amortized inference

Robert Bamler - Lecture 9 of the course "Data Compression With and Without Deep Probabilistic Models" - Summer Term of 2023 - more course materials at https://robamler.github.io/teaching/compress23/

## Part 1: Learning the Generative Model





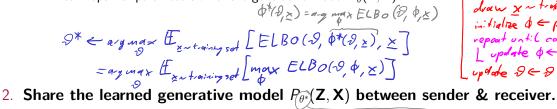
- ▶ **Goal:** learn optimal parameters  $\theta^*$  of the *generative* model  $P_{\theta}(\mathbf{Z}, \mathbf{X}) = P_{\theta}(\mathbf{Z}) P_{\theta}(\mathbf{X} \mid \mathbf{Z})$ .
  - ► Thus, the ELBO now depends on  $\theta$ , i.e., ELBO $(\theta, \phi, \mathbf{x}) = Q_{\phi}(\mathbf{Z}) \left[ \log P_{\theta}(\mathbf{Z}, \mathbf{X} = \mathbf{x}) \log Q_{\phi}(\mathbf{Z}) \right]$
  - **Example:** data  $\mathbf{X} = (X_i)_i$  are binarized images, i.e., each  $X_i$  is a pixel value  $\in \{0,1\}$ .
    - $\rightarrow$  Prior is fixed:  $P(\mathbf{Z} = \mathbf{z}) = \mathcal{N}(\mathbf{z}; 0, I)$  (standard normal distribution)
    - ightarrow Likelihood is parameterized by a (deconvolutional) neural network  $g_{\theta}$ :  $ho_{\theta}(\mathbf{X} \mid \mathbf{Z}) = \prod_{i} P_{\theta}(X_{i} \mid \mathbf{Z})$  with  $P_{\theta}(X_{i} = \mathbf{I} \mid \mathbf{Z} = \mathbf{z}) = \sigma(g_{\theta,i}(\mathbf{z}))$  in jury ternote book.

#### Distinguish:

- ▶ global parameters  $\theta^*$  ("model parameters"):
  - $\rightarrow$  specify the *generative model*  $P_{\theta^*}(\mathbf{Z}, \mathbf{X})$
  - $\rightarrow$  same for all data points  $\mathbf{x} \Longrightarrow$  known to both sender & receiver
- *local* parameters  $\phi^*$  ("variational parameters"):
  - $\rightarrow$  specify an approximation  $Q_{\phi^*}(\mathbf{Z})$  to the posterior  $P_{\theta^*}(\mathbf{Z} \mid \mathbf{X} = \mathbf{x})$  for a specific data point  $\mathbf{x}$
  - ightarrow different for each data point  ${f x}\Longrightarrow$  not available to the receiver until it has decoded  ${f x}$

### **Variational Expectation Maximization**

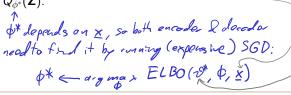
- 1. In order to develop a new compression method:
  - learn optimal parameters  $\theta^*$  of the generative model  $P_{\theta}(\mathbf{Z}, \mathbf{X})$ :



in: Halize & = roundown repeat until compagn

- 3. In deployment: encode / decode a given data point x

▶ Use entropy model  $Q_{\phi^*}(\mathbf{Z})$ .



# Part 2: Learning How to Do Inference (Fast)





- **Problems:** 
  - 1. Learning the generative model requires an expensive inner loop for every training step.
  - 2. Expensive optimization over  $\phi$  for each message **x** we want compress / decompress.
- **Solution:** *amortized* variational inference
  - $\blacktriangleright$  learn a mapping f from x to variational parameters such that setting  $\phi \leftarrow f(\mathbf{x})$  approximately maximizes ELBO $(\theta^*, \phi, \mathbf{x})$  for a given  $\mathbf{x}$ .
  - ▶ Notation: inference network  $f_{\phi}(\mathbf{x})$ ; variational distribution  $Q_{\phi}(\mathbf{Z} \mid \mathbf{X} = \mathbf{x})$  ← in the notation we've used so far, this would be Qfi(x) (Z)
  - **Example:** Gaussian mean field variational distribution:
    - o inference network  $f_\phi({f x})=(m{\mu}_\phi({f x}),\logm{\sigma}_\phi^2({f x}))$  outputs means and (log) variances
    - o these parameterize a variational distribution  $Q_{\phi}(\mathbf{Z}\,|\,\mathbf{x}) = \mathcal{N}ig(\mu_{\phi}(\mathbf{x}), \mathrm{diag}ig(\sigma_{\phi,1}^2(\mathbf{x}), \ldots, \sigma_{\phi,k}^2(\mathbf{x})ig)ig)$

training algorithm now:  $\begin{cases} \text{in:Halize } \mathcal{D}, \emptyset \leftarrow \text{random} \\ \text{in:Halize } \mathcal{D}, \emptyset \leftarrow \text{random} \\ \text{training algorithm now:} \end{cases} \text{ ELBO for amorbited in ference:} \\ \text{repeat ontil convergence:} \\ \text{ELBO(D, <math>\phi, \times ) = \mathbb{E}_{Q_{\phi}(Z|X=X)} \Big[ \log P(Z,X=X) - \log Q_{\phi}(Z|X=X) \Big]} \\ \text{In order } \{0, \emptyset \leftarrow \{0, \phi\} + S \setminus \nabla_{Q, \phi} \in \text{ELBO}(D, \phi, X) \} \Rightarrow \text{no expensive inner loop: } \mathcal{D} \text{ and } \phi \end{cases}$  $(LBO(-9, \phi, x))$   $\Rightarrow$  no expensive inner loop; -9 and  $-\phi$  terials at https://robanler.github.io/teaching/cospress23/ ave (earned concurrently

Variational Autoencoders (VAEs)



Combine variational expectation maximization with amortized variational inference. That's all.

- Lossless compression with variational autoencoders:
  - use bits-back trick  $\rightarrow$  Problem 9.1
- Lossy compression with variational autoencoders:
  - **Example:** data  $\mathbf{X} = (X_i)_i$  are color images, i.e., each  $X_i$  is a continuous RGB value  $\in [0,1]$ .
    - ightarrow Prior may be learned, e.g.:  $P_{\theta}(\mathbf{Z} = \mathbf{z}) = \mathcal{N}(\mathbf{z}; 0, \mathsf{diag}(\sigma_1^2, \dots, \sigma_{\mathsf{num channels}}^2)^{\otimes \mathsf{spatial\_dim}})$
    - $\rightarrow$  Likelihood is parameterized by a (deconvolutional) neural network  $g_{\theta}$ :  $P_{\theta}(\mathbf{X} \mid \mathbf{Z}) = \prod_{i} P_{\theta}(X_{i} \mid \mathbf{Z})$  with density function  $P_{\theta}(x_{i} \mid \mathbf{Z} = \mathbf{z}) = \mathcal{N}(x_{i}; g_{\theta,i}(\mathbf{z}), \frac{1}{\beta}I)$
  - ▶ **Idea:** just use  $g_{\theta,i}(\mathbf{z})$  as the reconstruction of an image. (Don't bother using the likelihood  $P_{\theta}(\mathbf{X} | \mathbf{Z} = \mathbf{z})$  to encode the true image.)
  - Likelihood no longer has a probabilistic meaning. But  $-\log P_{\theta}(\mathbf{X} \mid \mathbf{Z} = \mathbf{z})$  is a distortion metric. ⇒ ELBO becomes a *rate-distortion* trade-off
  - next week