Lecture 5: Latent Variable Models 2 and Bits-back Coding

Scribed by: Terence Neo, Eloise Lim, Xueqi, Rishav Chourasia
Lecture presented by: David Yam, Hitoshi, Song Kai

Lecture 4a: Latent Variable Models 2

Variational Inference

- Evaluate marginal likelihood to train the latent variable model

$$p(x) = \sum_z p(z, x)$$

VI as Importance Sampling

- If z is high dimension and probability mass concentrated over one z (eg. car in an image, only the car is important)
  - Use Importance Sampling to sample high density regions
  - The lower bound of log (p(x)) computed via importance sampling is

$$\mathbb{E}_{z \sim q(z|x)} \left[\log \frac{p(z, x)}{q(z|x)} \right]$$

- Importance ratio for some sample of z, zi:

$$w_i = \frac{p(z_i, x)}{q(z_i|x)}$$

- The lower bound of log p(x) can be tightened by taking k samples:

$$\mathcal{L}_k = \mathbb{E}\left[\log \frac{1}{k} \sum_{i=1}^{k} w_i\right]$$

- **Theorem**: For all k, the lower bounds satisfy

$$\log p(\mathbf{x}) \geq \mathcal{L}_{k+1} \geq \mathcal{L}_k.$$

Moreover, if $p(\mathbf{h}, \mathbf{x})/q(\mathbf{h}|\mathbf{x})$ is bounded, then $\mathcal{L}_k$ approaches $\log p(\mathbf{x})$ as k goes to infinity.

- Implications:
  - 1st inequality:

- 2nd inequality: If you have more samples, you will not be worse off (lower bound increases as k increases)
-

Improving VAEs

- Reducing Variational Gap:
  - Result of mismatch between approx posterior and true posterior
  - Use importance sampling: IWAE (Importance Weighted Autoencoder)

Lecture 4b: Bits-Back Coding

Challenges from previous encodings: **Continuous** data and **high dimension** data

Use multiple Gaussian to encode many distributions that the original distribution cannot decode

Scheme 1: "Max-mode" Coding
To code x:
Find i that maximises p(i|x)
Send i --- cost: log1/p(i)
Send x --- cost:log 1/p(x|i)
Limitation: may not be the max a gaussian can encode, cost = H(x) + KL(P||Q)

Scheme 2: Posterior-Mode Sampling Coding
Optimal?
Yes if we like to send (i,x) b/c we use log 1/p(x,i)
BUT: we are looking to send just x, so the overhead of log 1/(pi|x)

Scheme 3: Bits-Back Coding (the best)
Recipient decodes i,x + knows p(i|x)
-> can reconstruct the random bits used to sample p(i|x)
-> those random bits were also sent -> these are log1/p(i|x) random bits, which we now don't have to count
-> the cost is the lowest -> Optimal!!!

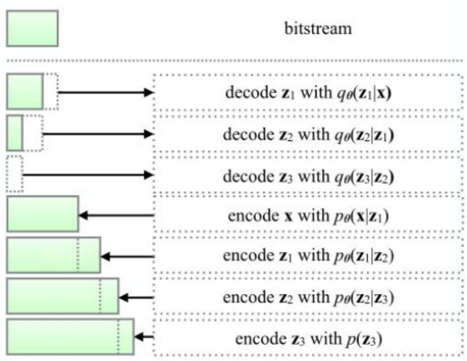
BB-ANS
How well does BB-ANS work?
Assumptions to investigate:
- Finite precision approximation of log 1/p
- Inefficiency in encoding the first data point
- VAE has **continuous** latent variables

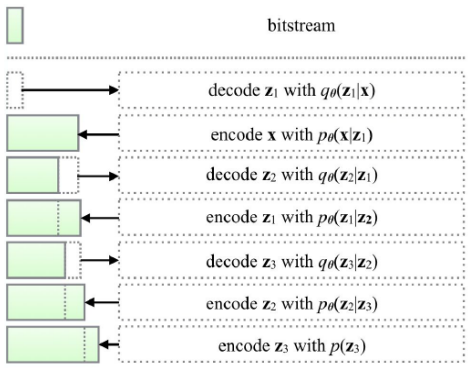
    - Expected encoding length is given by KL(continuous distribution || discrete distribution)
- Are the bits **clean**

Can we do even better?
Quality of encoding depends on quality ELBO
Latent variable models with multiple latent layers tend to achieve better ELBO than with a single latent layer

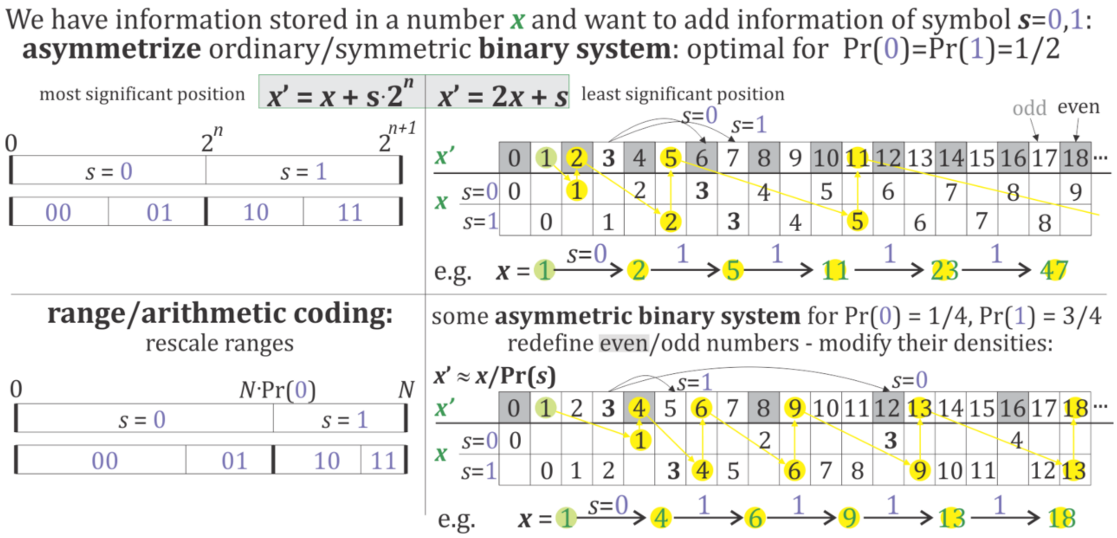
Bit-Swap

Bit-Swap Encoding
-treat information as whole
-By doing all decoding at first → larger information bits
- require fewer initial bits than ANS





Asymmetric Numeral System (ANS)
Assign natural numbers to a and b, a and b are unique sets, a union b is the universal set
Eg. if p(a) = ¼, a contains every 4th number, eg a = {0,4,8,12,16,...}, b = {1,2,3,5,6,7,...}



Some helpful links:

1. Importance weighted autoencoders tutorial http://dustintran.com/blog/importance-weighted-autoencoders
2. Asymmetric numeral systems: entropy coding combining speed of Huffman coding with compression rate of arithmetic coding: https://arxiv.org/pdf/1311.2540.pdf
3. Importance sampling tutorial https://www.roe.ac.uk/ifa/postgrad/pedagogy/2010_grocutt.pdf
4. Importance weighted autoencoder paper https://arxiv.org/abs/1509.00519
5. Bits-back coding https://www.cs.helsinki.fi/u/ahonkela/papers/infview.pdf
6. 2019 paper, exceptional performance with just one VAE bits-back coding, BB-ANS https://arxiv.org/pdf/1901.04866.pdf
7. Asymmeric numeral system https://en.wikipedia.org/wiki/Asymmetric_numeral_systems