Introduction to Variational Inference and its Applications

Valentina Staneva
Senior Data Scientist, eScience Institute

UNIVERSITY *of* WASHINGTON

Estimating Posteriors: Sampling Techniques

MCMC Methods

- iterative dependent sampling
- sampling in high dimensions is hard
- slow convergence, hard diagnostics

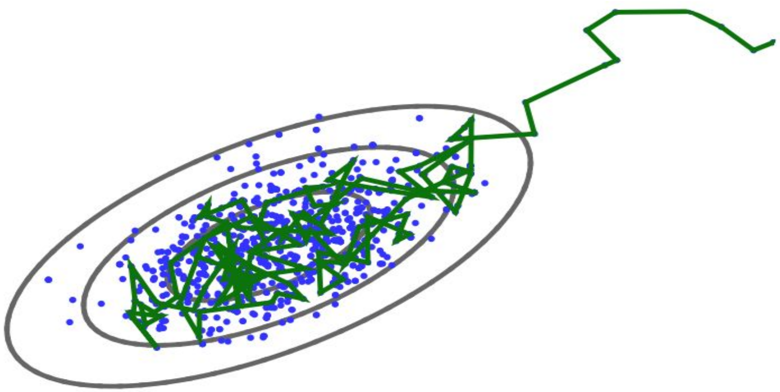


Image Source

Importance Sampling

- independent samples from approximate distribution
- weight degenaracy
- requires even larger samples

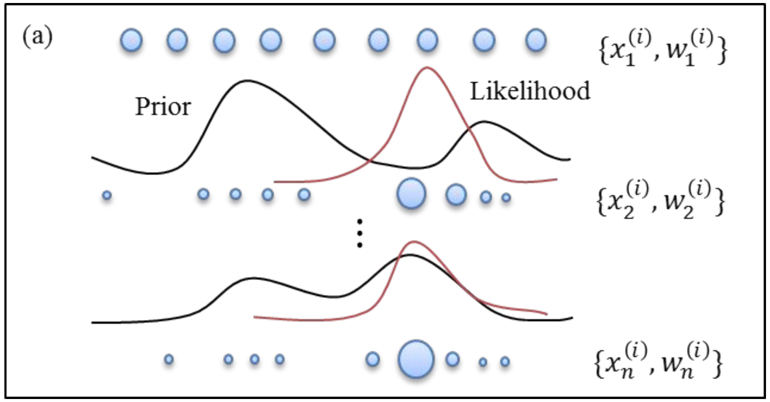
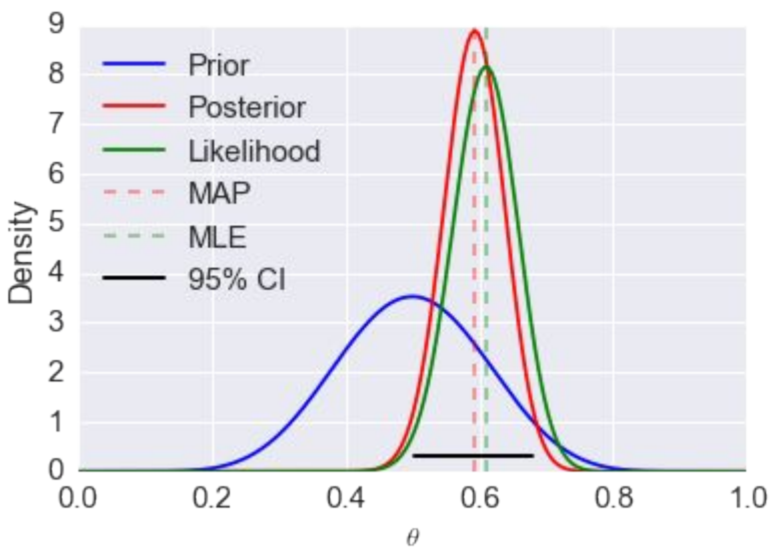


Image Source

Estimating Posteriors: Optimization Techniques

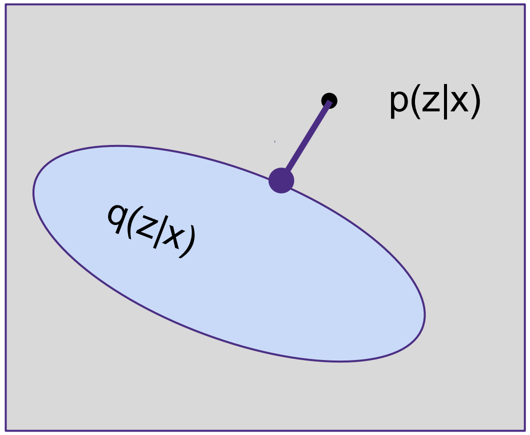
Expectation-Maximization (EM)

- finds only the mode, not the complete distribution



Variational Inference

- select a distribution family q(z|x)
- minimize D(q(z|x), p(z|x))



Minimization Loss

- Minimize KL divergence

$$D_{KL}(Q(Z|X)||P(Z|X)) = \int_Z Q(Z|X) \log \frac{Q(Z|X)}{P(Z|X)} dZ$$

- Maximize Evidence Lower Bound (ELBO)

$$\mathcal{L}(Q) = -D_{KL}(Q(Z|X)||P(Z)) + \mathbb{E}_{Q(Z|X)} \log(P(X|Z))$$

Notes

- Sometimes the first term can be computed analytically.
- The second term is obtained by sampling from Q(Z|X)

Stochastic Variational Inference

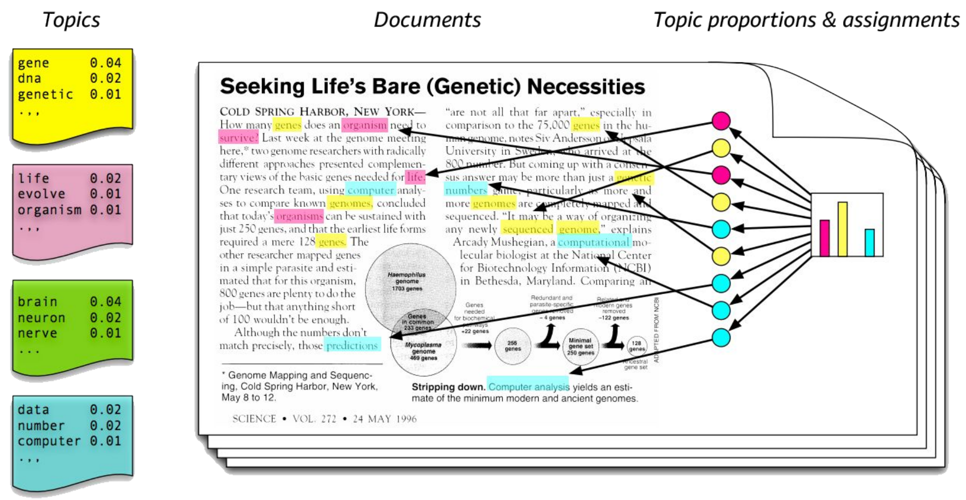
Stochastic Gradient Descent:

- evaluate gradient at individual (or batch of observations)
- reduce time step to optimize function

Stochastic Variational Inference:

- evaluate gradients at individual (or batch of observations)
- keep time step fixed to achieve stationary distribution

Example: Topic Models



Generate j'th word in document i:

- Choose topic z_i,j ~ Multinomial(θ_i)
- Choose word w_i,j ~ Multinomial(φ(z_ij))

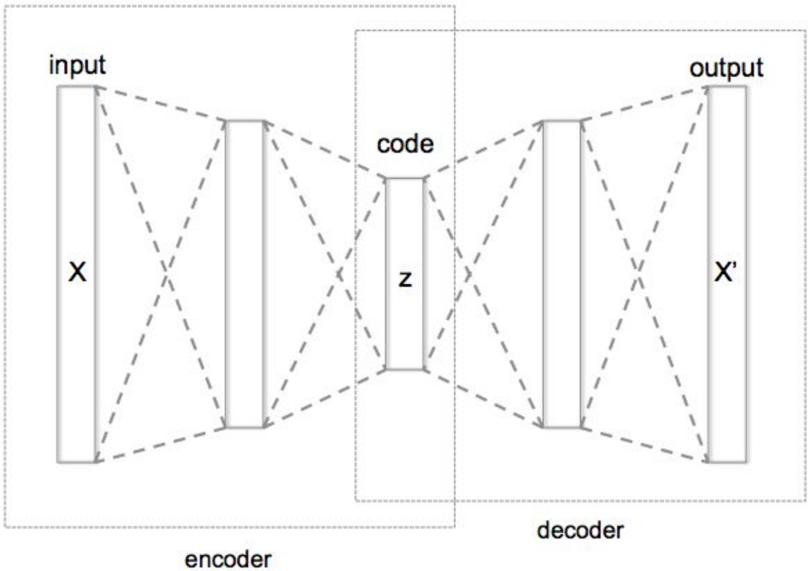
θ ~ Dirichlet(α) (topic distribution per document)

φ ~ Dirichlet(β) (word distribution of topic)

Sample from a distribution for which the latent variables are decoupled.

[Blei'03]

Autoencoders



https://en.wikipedia.org/wiki/Autoencoder#/media/File:Autoencoder_structure.png

- Minimize reconstruction cost
- Mappings represented by a neural network

Note, when only one hidden layer: X = WZ (matrix decomposition)

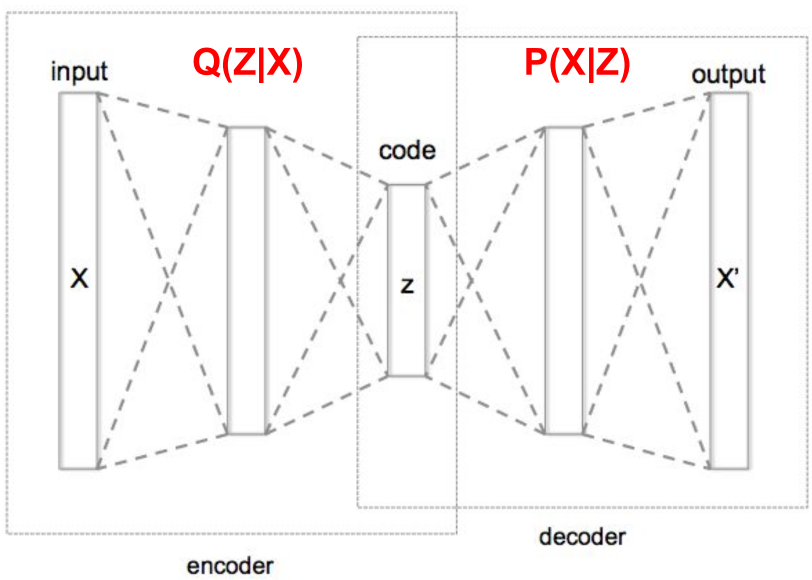
Variational Autoencoders

decoder = generative network $p(z)$, $p(x|z)$

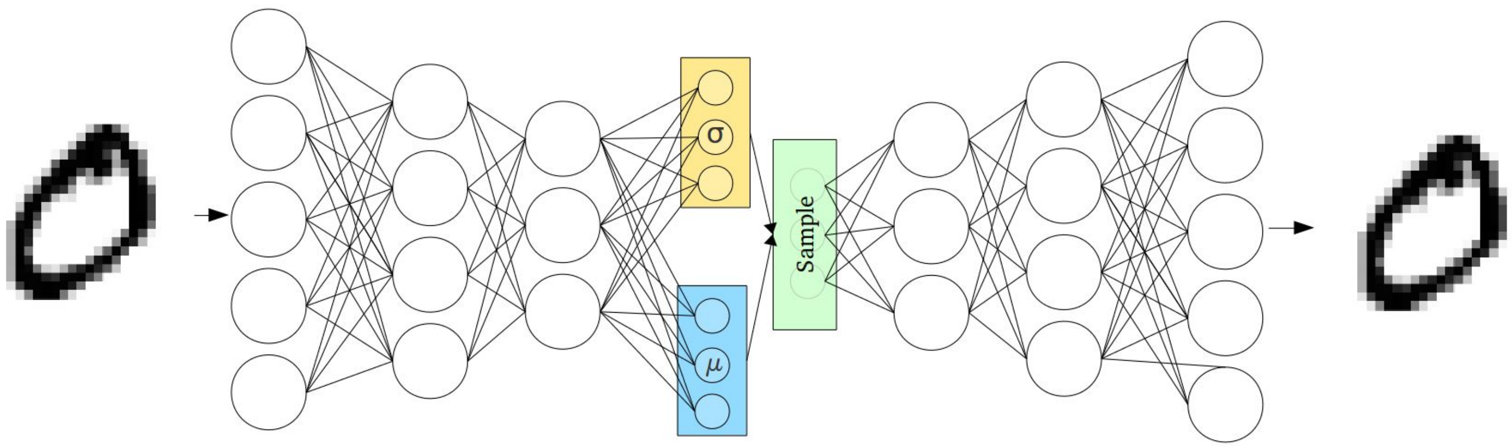
encoder = inference network $q(z|x)$

reconstruction cost = -E_Q(Z|X)[log(P(Z|X))]

regularization = D_{KL}(Q(Z|X)||P(Z))



Example: MNIST Digits



Find low dimensional latent variables z from which we can generate the digits.

$$Q : z|x \sim \mathcal{N}(\mu, \sigma I)$$ where ($\boldsymbol{\mu}$, $\boldsymbol{\sigma}$) = inference_network(x)

$$P : z \sim \mathcal{N}(0, I)$$

$$P : x_i|z \sim Bernoulli(logit)$$ where logit = generative_network(z)

Computations

Why Tensorflow & Keras?

- GPU support
- automatic differentiation
- stochastic gradient descent methods
- built-in tools for batch processing and evaluation
- can add deep models

Tensorflow Probability

Edward Library

References

- Auto-Encoding Variational Bayes
- Stochastic Variational Inference
- Stochastic Gradient Descent for Variational Inference
- Keras MNIST Example (dense layers)
- Colab MNIST Example (conv layers)
- https://github.com/valentina-s/Variational_Inference
- Tensorflow Probability Examples