Variational Inference

Bryan Pardo

Northwestern University

(updated fall 2022)

Some basic definitions

Some definitions

- **Sample Space:** The set, S,of possible values a random variable can take. These are mutually exclusive (e.g. coin flips: heads or tails).

- **Random Variable**: A mapping from a sample space to actual measured outputs. We denote the whole mapping with a capital letter (e.g. X) and a particular sample output with a lower case (e.g. x)

- **Support:** For a random variable X, is the portion of the sample space that has non-zero probability. (if a flipped coin always turns up “heads”, then the support is “heads”)

Some definitions

- A **Discrete random variable** has a countable sample space. For us, we'll use FINITE sample spaces, like heads/tails or words in a dictionary.

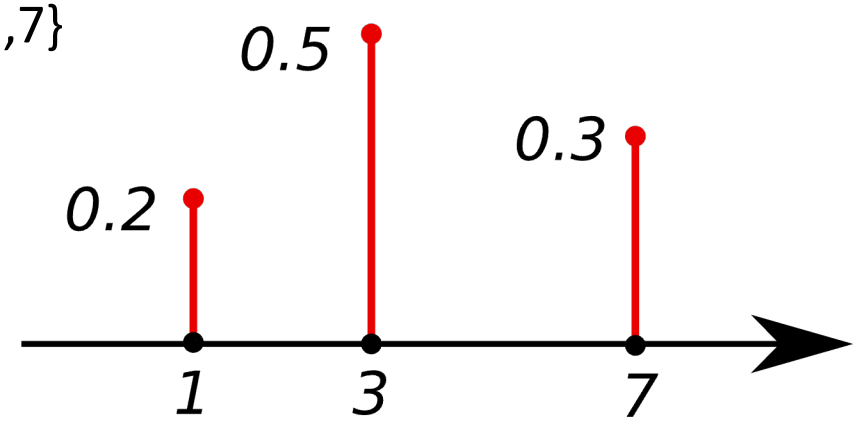
- A **Continuous random variable** has an uncountably infinite sample space, like real numbers on the interval (0,1).

Probability Mass Function

… specifies the probability of a **discrete random variable** taking each of the values in the sample space.

The PMF is nonnegative, and the sum of its probabilities = 1.0

In this example, the sample space is {1,3,7}



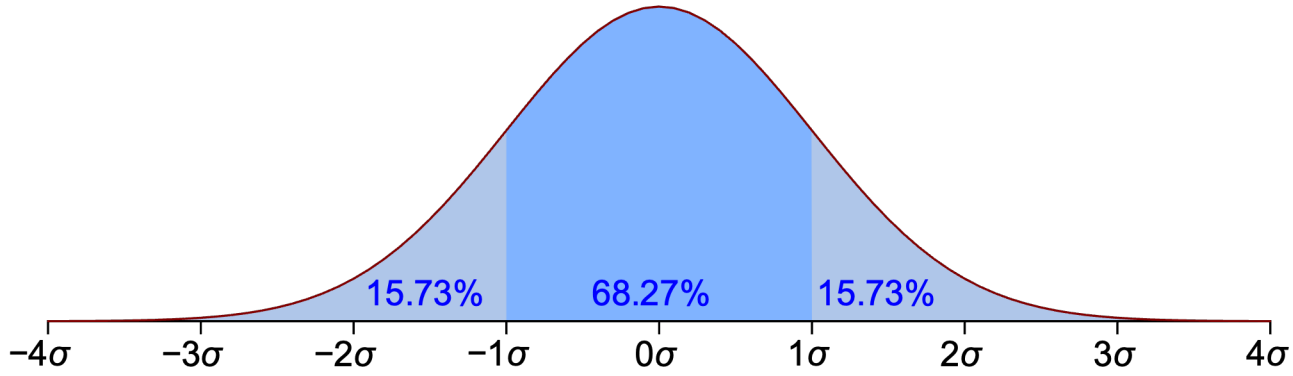
By Oleg Alexandrov - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=2073424

Probability density function (PDF)

… specifies the probability of a **continuous random variable** falling within a particular range of values.

This probability is given by the integral of of the PDF over that range (i.e. the area under the curve).

The PDF is nonnegative, and its integral over the entire space = 1.0



Expected value E[X] of a random variable X

- For a finite, discrete random variable X, E[X] is defined as

$$E[X] = \sum_{x \in S} xp(x)$$

- For a continuous random variable on real numbers, E[X] is defined as

We often drop the -∞, ∞ and assume they're implied

$$E[x] = \int_{-\infty}^{\infty} xp(x)dx$$

x is a particular element in the sample space S
p(x) is the probability mass (or density) function for X applied to the sample x.

An integrable random variable.

- For a finite, discrete random variable X, E[X] is integrable iff:

$$\infty > \sum_{x \in S} |x| p(x)$$

- For a continuous random variable on real numbers, E[X] is integrable iff:

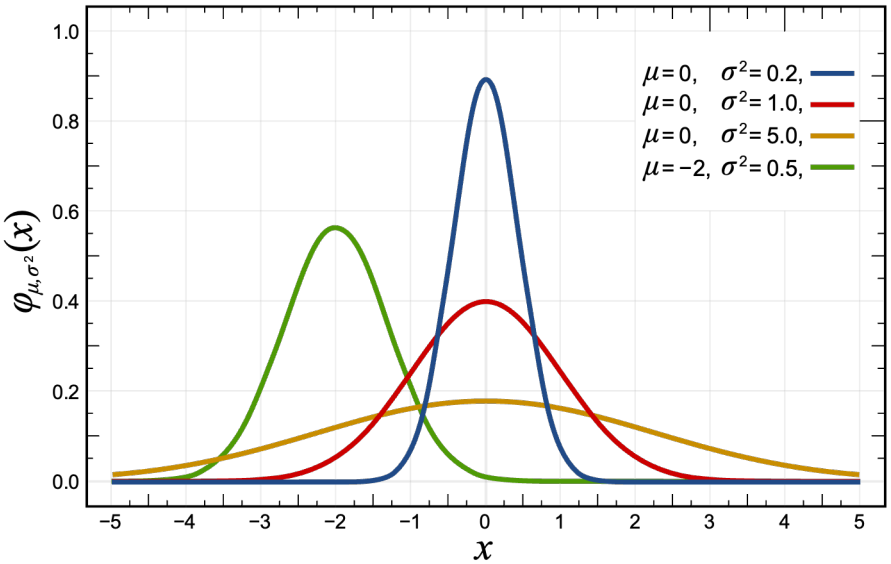
$$\infty > \int_{-\infty}^{\infty} x p(x) dx$$

x is a particular element in the sample space S
p(x) is the probability mass (or density) function for X applied to the sample x.

The Normal (Gaussian) Distribution

The Gaussian a.k.a. Normal Distribution

$$x \sim N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



By Inductiveload - self-made, Mathematica, Inkscape, Public Domain, https://commons.wikimedia.org/w/index.php?curid=3817954

The multivariate normal distribution of a *k*-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_k)^{\mathrm{T}}$ following notation:

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

or to make it explicitly known that *X* is *k*-dimensional,

$$\mathbf{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

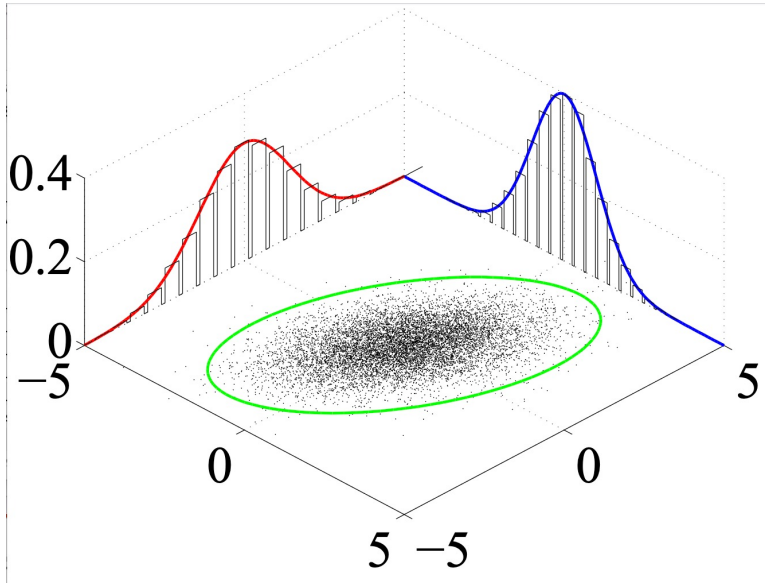
with *k*-dimensional mean vector

$$\boldsymbol{\mu} = \mathrm{E}[\mathbf{X}] = (\mathrm{E}[X_1], \mathrm{E}[X_2], \ldots, \mathrm{E}[X_k])^{\mathrm{T}},$$

and $k \times k$ covariance matrix

$$\Sigma_{i,j} = \mathrm{E}[(X_i - \mu_i)(X_j - \mu_j)] = \mathrm{Cov}[X_i, X_j]$$

such that $1 \leq i \leq k$ and $1 \leq j \leq k$. The inverse of the covariance matrix is called the preci



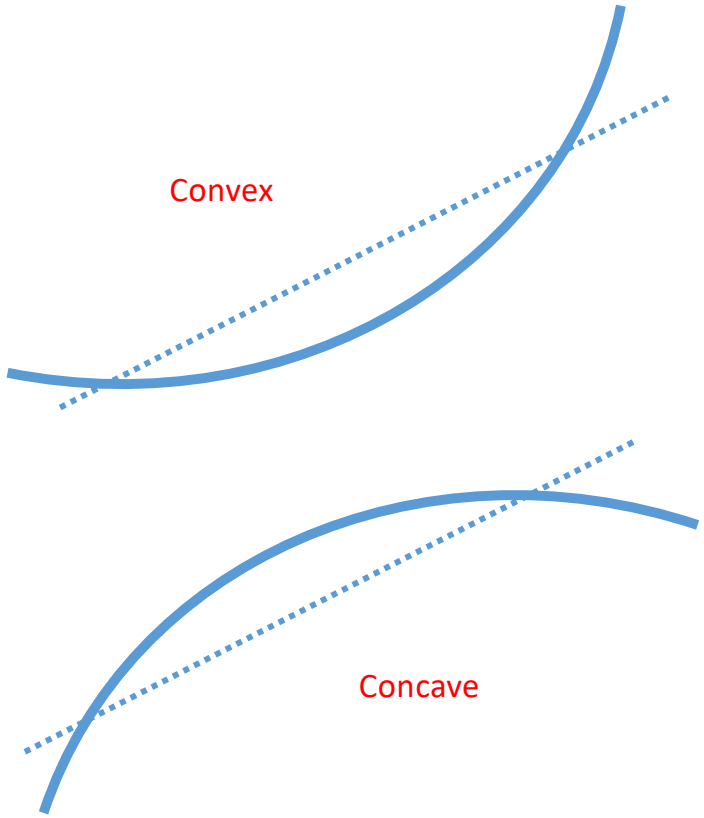
https://en.wikipedia.org/wiki/Multivariate_normal_distribution#Notation_and_parameterization

Jensen’s Inequality

Some definitions

- A real-valued function is **convex** if the line segment between any two points on the graph of the function lies **above** the graph between the two points.

- A real-valued function is **concave** if the line segment between any two points on the graph of the function lies **below** the graph between the two points.



Is my function convex/concave?

- A differentiable function is (strictly) convex if its second derivative is (strictly) positive;
- That function is (strictly) concave if its second derivative is (strictly) negative.
- Example The natural logarithm is strictly concave because…

$$\frac{d^2}{dx^2}\log(x) = -x^{-2}$$

Now for the inequality…

- let X be an integrable random variable a random variable and ϕ be a concave function, then Jensen's inequality is defined as:

$$E[\phi(X)] \leq \phi(E[X])$$

- We know log is a concave function. This means:

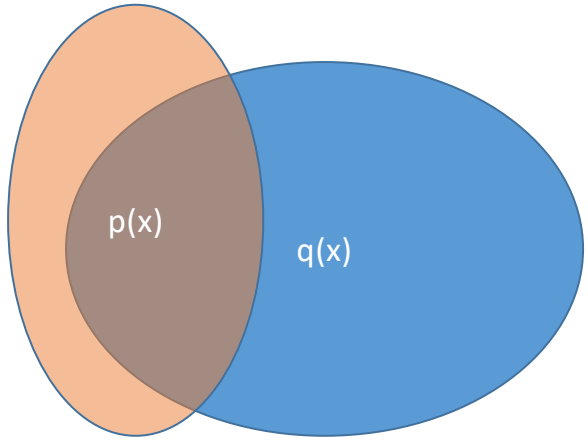
$$E[log(X)] \leq log(E[X])$$

- …we'll be using this later.

Kullback-Leibler divergence

Comparing two PDFs

- Given two PDFs, p(x) and q(x), how do I tell how similar they are?
- Maybe we could measure the overlap in some way?
- People use Kullback-Leibler divergence (KL divergence) for this.



KL-divergence

- P and Q are both probability functions on the same sample space.
- For finite sample spaces:

$$D_{KL}(P||Q) = \sum_{x} P(x) log\left(\frac{P(x)}{Q(x)}\right)$$

- For real-valued sample spaces :

$$D_{KL}(P||Q) = \int p(x) log\left(\frac{p(x)}{q(x)}\right) dx$$

Relation to entropy

$$D_{KL}(P||Q) = \sum_x P(x) log\left(\frac{P(x)}{Q(x)}\right)$$

$$= \sum_x P(x)\log(P(x)) - \sum_x P(x)\log\big(Q(x)\big)$$

This is the entropy of P(x)

This is the cross-entropy of P(x) compared to Q(x)

Jensen-Shannon Divergence

- Symmetric measure
- People often use this, because it is symmetric and avoids divide by 0

$$D_{JS}(P||Q) = D_{KL}(P||M) + D_{KL}(Q||M)$$

…where $M = \frac{P+Q}{2}$

Joint distributions

Joint Distribution

- Given two random variables, X and Z, defined on the same probability space, the joint probability distribution is the probability distribution on all possible pairs of outputs.

$$\mathrm{P}(X, Z)$$

The Chain Rule & Conditional Probability

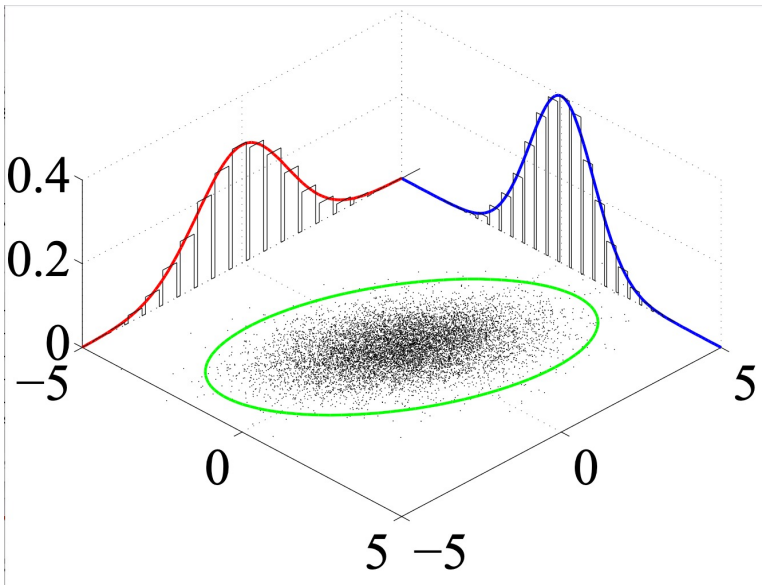
Definition of conditional probability:

$$P(X|Z) = P(X,Z)/P(Z)$$

$$P(Z|X) = P(X,Z)/P(X)$$

We can factor the joint distribution using conditional probability:

$$\begin{aligned} P(X,Z) &= P(X|Z)P(Z) \\ &= P(Z|X)P(X) \end{aligned}$$



By IkamusumeFan - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=30432580

Bayes Rule

With conditional probability:

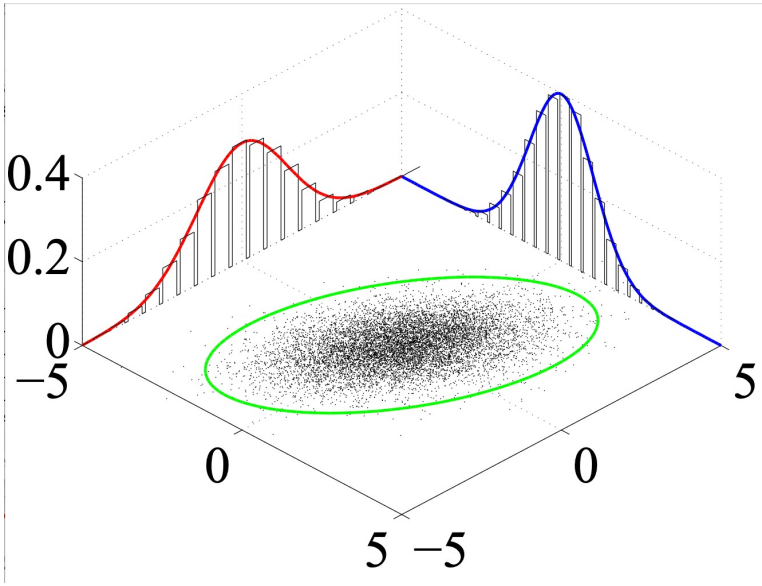
$$P(X|Z) = P(X,Z)/P(Z)$$

…and the chain rule:

$$P(X,Z) = P(Z|X)P(X)$$

We get to Bayes rule:

$$\begin{aligned} P(Z|X) &= P(X,Z)/P(X) \\ &= P(X|Z)P(Z)/P(X) \end{aligned}$$



A finite Joint Distribution of Booleans

- A truth table listing all combinations of variable values
- This embodies P(A,B,C)
- Each combination has a probability that must be estimated
- How big is this table for 100 Boolean variables?

A	B	C	Prob
0	0	0	0.1
0	0	1	0.2
0	1	0	0.1
0	1	1	0.05
1	0	0	0.05
1	0	1	0.2
1	1	0	0.25
1	1	1	0.05

Finding a marginal distribution

- To find P(A) we must “marginalize”
- Sum the probabilities of all rows where A=1

$$P(A) = \sum_b \sum_c P(1, b, c)$$
$$= 0.05 + 0.2 + 0.25 + 0.05$$
$$= 0.55$$

What happens if there are 100 variables?

A	B	C	Prob
0	0	0	0.1
0	0	1	0.2
0	1	0	0.1
0	1	1	0.05
1	0	0	0.05
1	0	1	0.2
1	1	0	0.25
1	1	1	0.05

Inferring latent variables

More definitions!

- If you can directly sample a random variable X and see the outcome x, we call x an **observation**.

- If you can't directly sample a random variable Z and see the outcome z, we call Z a **latent variable**.

- We'll typically use those letters with those implications: X is observable, Z is latent.

A grounded example with names

- What is the chance I have Covid, if I have a positive Covid test?
- Let Z be the latent variable ”Covid: yes/no”
- Let X be the observable variable “test: positive/negative”

- P(X) is the unconditioned (aka **prior**) probability of a positive test.
- P(Z) is the **prior** probability of having Covid.
- P(Z|X) is the **posterior** probability of Covid, given the observed outcome.
- P(X|Z) is the probability of a test outcome, given the truth of whether you have Covid. (aka the **likelihood**)

The problem

Let $\mathbf{x} = \mathrm{x}_1 \ldots \mathrm{x}_{\boldsymbol{n}}$ be a set of observed variables.

Let $\mathbf{z} = \mathrm{z}_1 \ldots \mathrm{z}_{\boldsymbol{m}}$ be a set of latent variables of interest.

We want to infer these latent variables from the evidence. We want to know $p(\mathbf{z}|\mathbf{x})$.

$$p(\mathbf{z}|\mathbf{x}) = \frac{\boldsymbol{p}(\mathbf{z}, \mathbf{x})}{\boldsymbol{p}(\mathbf{x})}$$

Note that this simple expression hides a lot of complexity.

Marginalization is not your friend.

$$p(\mathbf{z}|\mathbf{x}) = \frac{\boldsymbol{p}(\mathbf{z}, \mathbf{x})}{\boldsymbol{p}(\mathbf{x})}$$

To calculate $\boldsymbol{p}(\mathbf{x})$ we often have to marginalize all the variables in $\mathbf{x}$ over all the variables in $\mathbf{z}$.

This is often intractable. Remember the joint probability table.

Variational Inference

Based on (but modified from):

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: A review for statisticians." Journal of the American statistical Association 112.518 (2017): 859-877.

https://arxiv.org/pdf/1601.00670.pdf

Set this up as an estimation problem.

- We want to learn $p(\mathbf{z}|\mathbf{x})$
- Make a family of density functions θ
- Search through θ to find $q^*(\mathbf{z})$, which optimizes this equation:

$$q^*(\mathbf{z}) = \underset{q(\mathbf{z})\in\theta}{\operatorname{argmin}} D_{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}))$$

This is Blei et al's formulation. These folks came up with Variational Inference.

In the end we'll want to do this…

$$q^*(\mathbf{z}|\mathbf{x}) = \underset{q(\mathbf{z})\in\theta}{\operatorname{argmin}} D_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x}))$$

This is Kingma et al's formulation. These folks invented the Variational Autoencoder (VAE).

The big difference is making z depend on x.

…but for now, we'll go back to Blei et al's formulation.

What are we doing, again?

- We're to skip that marginalization process on the evidence.
- We're using KL divergence to find the best fit from a family of distributions θ .
- We must pick a θ flexible enough to have a member that models p(z|x) well, while still being tractable to search.
- Often, people use the Gaussians as the family of distributions.

So does this make things any easier?

$$q^*(z) = \underset{q(\mathbf{z})\in\theta}{\operatorname{argmin}} D_{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}))$$

Let's pick apart that KL divergence a bit, taking expectation with respect to z

$$\begin{aligned} D_{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) &= \mathrm{E}[\log q(\mathbf{z})] - \mathrm{E}[\log p(\mathbf{z}|\mathbf{x})] \\ &= \mathrm{E}[\log q(\mathbf{z})] - \mathrm{E}[\log p(\mathbf{z},\mathbf{x})] + \log p(\mathbf{x}) \end{aligned}$$

But wait….that term $p(\mathbf{x})$ is back. That's the one we said was intractable!

No worries. This is going to the other side of the equation.

Getting to the ELBO

From previous slide:

$$D_{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) = \mathrm{E}[\log q(\mathbf{z})] - \mathrm{E}[\log p(\mathbf{z},\mathbf{x})] + \log p(\mathbf{x})$$

Negate both side:

$$-D_{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) = -\mathrm{E}[\log q(\mathbf{z})] + \mathrm{E}[\log p(\mathbf{z},\mathbf{x})] - \log p(\mathbf{x})$$

Add $\log p(\mathbf{x})$ to both sides:

This is called the Evidence Lower Bound (ELBO)

$$-D_{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) + \log p(\mathbf{x}) = -\mathrm{E}[\log q(\mathbf{z})] + \mathrm{E}[\log p(\mathbf{z},\mathbf{x})]$$

Put our evidence on one side:

$$\log p(\mathbf{x}) = D_{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) + \mathrm{ELBO}$$

Why it is called the “Evidence Lower Bound”

From previous slide:

$$\log p(\mathbf{x}) = D_{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) + \text{ELBO}$$

By definition $D_{KL}() \geq 0$, so the ELBO is a lower bound on the log evidence. The more we reduce this KL divergence, the closer the ELBO gets.

Rewriting the ELBO

$$\text{ELBO} = \mathrm{E}[\log p(\mathbf{z}, \mathbf{x})] - \mathrm{E}[\log q(\mathbf{z})]$$

by the chain rule $$= \mathrm{E}[\log p(\mathbf{x}|\mathbf{z})] + \mathrm{E}[\log p(\mathbf{z})] - \mathrm{E}[\log q(\mathbf{z})]$$

By def of KL divergence $$= \mathrm{E}[\log p(\mathbf{x}|\mathbf{z})] - D_{KL}(q(\mathbf{z})||p(\mathbf{z}))$$

This is starting to look like a good function to optimize.

We have the expected log likelihood of the data and the KL divergence between the true distribution for the hidden variables and our guestimate.