

# Study Material - Youtube

## Document Information

- **Generated:** 2025-08-01 22:11:14
- **Source:** <https://youtu.be/d9fDDUcQqq4>
- **Platform:** Youtube
- **Word Count:** 2,252 words
- **Estimated Reading Time:** ~11 minutes
- **Number of Chapters:** 6
- **Transcript Available:** Yes (analyzed from video content)

## Table of Contents

1. The General Principle for Learning Latent Variable Models
  2. The Evidence Lower Bound (ELBO)
  3. Key Mathematical Concepts
  4. Visual Elements from the Video
  5. Self-Assessment for This Video
  6. Key Takeaways from This Video
- 

## Video Overview

This video lecture, “Mathematical Foundations of Generative AI: Evidence Lower Bound (ELBO),” provides a detailed mathematical derivation of the general principle used for training latent variable models. The instructor begins by establishing the goal of learning in these models: to make the model’s probability distribution as close as possible to the true data distribution. This is framed as a Maximum Likelihood Estimation (MLE) problem. The core of the lecture demonstrates that the direct optimization of the log-likelihood (also known as the “evidence”) is often intractable due to a logarithm of an integral. To overcome this, the instructor introduces a variational distribution and masterfully applies Jensen’s Inequality to derive a tractable lower bound on the evidence, famously known as the **Evidence Lower Bound (ELBO)**. This ELBO then becomes the new objective function to be maximized. This principle is fundamental to many modern generative models, including Variational Autoencoders (VAEs) and Diffusion Models.

## Learning Objectives

Upon completing this lecture, students will be able to:

- Understand the core objective of training latent variable models using Maximum Likelihood Estimation.
- Recognize and explain why the log-likelihood (evidence) is computationally intractable in many latent variable models.
- Follow the step-by-step mathematical derivation of the Evidence Lower Bound (ELBO).
- Explain the role of the variational distribution  $q(z|x)$  in the derivation.
- Understand and apply Jensen’s Inequality to derive the ELBO.
- Formulate the new optimization problem based on maximizing the ELBO with respect to both model parameters and the variational distribution.

## Prerequisites

To fully grasp the concepts in this video, students should have a solid understanding of:

- **Probability Theory:** Concepts of probability density functions (PDFs), joint and marginal distributions, and expectation.
- **Calculus:** Familiarity with integrals, logarithms, and basic optimization (finding maxima/minima).
- **Information Theory:** A basic understanding of Kullback-Leibler (KL) divergence is beneficial.
- **Machine Learning:** General knowledge of what latent variable models are and their purpose.

## Key Concepts Covered in This Video

- Latent Variable Models (LVMs)
  - Maximum Likelihood Estimation (MLE)
  - Log-Likelihood (Evidence)
  - Kullback-Leibler (KL) Divergence
  - Jensen's Inequality
  - Variational Distribution (Variational Latent Posterior)
  - Evidence Lower Bound (ELBO)
- 

## The General Principle for Learning Latent Variable Models

The instructor begins by establishing the foundational principle for training any latent variable model, a concept that is broadly applicable across a wide range of generative models (00:47).

### The Goal: Maximum Likelihood Estimation (MLE)

#### Intuitive Foundation

Imagine you have a collection of images of cats. Your goal is to train a computer program (a generative model) to create new, realistic images of cats that have never been seen before. The model learns by looking at the existing images and trying to understand the underlying “essence” or “features” of what makes an image a cat. These underlying, unobserved features are the **latent variables**.

The training process aims to adjust the model's internal parameters so that the images it generates are as “plausible” or “likely” as the real cat images it was trained on. In probabilistic terms, we want to maximize the probability (or likelihood) that our model would have generated the training data. This is the core idea behind **Maximum Likelihood Estimation (MLE)**.

#### Mathematical Formulation

Let's formalize this intuition.

1. **Data Distribution (01:10):** We are given a dataset  $D = \{x_i\}_{i=1}^n$ , where each data point  $x_i$  (e.g., an image) is assumed to be drawn independently from an unknown true data distribution  $P_x$ .
2. **Latent Variable Model (01:35):** We define a probabilistic model  $P_\theta(x)$  parameterized by a set of parameters  $\theta$ . This model incorporates a latent (hidden) variable  $z$ . The probability of observing a data point  $x$  is found by considering all possible values of the latent variable  $z$  and marginalizing them out from the joint distribution  $P_\theta(x, z)$ .

**Definition: Marginal Likelihood (Evidence)** The probability of a data point  $x$  under the model is given by the integral of the joint probability over all possible latent variables  $z$ . This is also known as the **evidence**.

$$P_\theta(x) = \int_z P_\theta(x, z) dz$$

For simplicity, we assume  $z$  is continuous. If  $z$  were discrete, the integral would be replaced by a summation.

3. **The Optimization Goal (02:24):** Our goal is to find the optimal parameters  $\theta^*$  that make our model distribution  $P_\theta(x)$  as similar as possible to the true data distribution  $P_x$ . This is formally achieved by minimizing the **Kullback-Leibler (KL) divergence** between the two distributions.

**KL Divergence Minimization** The objective is to find the parameters  $\theta^*$  that minimize the KL divergence:

$$\theta^* = \arg \min_{\theta} D_{KL}(P_x \parallel P_{\theta}(x))$$

### From KL Divergence to Log-Likelihood

The instructor demonstrates (03:20) that minimizing KL divergence is equivalent to maximizing the log-likelihood of the data.

#### Step-by-step Derivation:

1. **Expand the KL Divergence:**

$$D_{KL}(P_x \parallel P_{\theta}) = \int_x P_x(x) \log \frac{P_x(x)}{P_{\theta}(x)} dx$$

2. **Separate the terms using log properties:**

$$D_{KL}(P_x \parallel P_{\theta}) = \int_x P_x(x) \log P_x(x) dx - \int_x P_x(x) \log P_{\theta}(x) dx$$

3. **Analyze the terms:**

- The first term,  $\int_x P_x(x) \log P_x(x) dx$ , is the negative **entropy** of the true data distribution  $P_x$ . Since  $P_x$  is fixed and does not depend on our model's parameters  $\theta$ , this term is a constant with respect to the optimization.
  - Therefore, minimizing the KL divergence is equivalent to minimizing the negative of the second term, which is the same as **maximizing** the second term itself.
4. **The Final Objective:** The optimization problem simplifies to maximizing the expected log-likelihood of the data under the model distribution.

#### Maximum Log-Likelihood Objective (05:06):

$$\theta^* = \arg \max_{\theta} \mathbb{E}_{x \sim P_x} [\log P_{\theta}(x)]$$

This is the celebrated **Maximum Likelihood Estimation (MLE)** principle. We want to find the parameters  $\theta$  that make the observed data most probable.

### The Challenge: An Intractable Objective

While the MLE principle is elegant, a major computational hurdle arises when we try to optimize it for latent variable models (05:25). The objective function contains the term  $\log P_{\theta}(x)$ , which is:

$$\log P_{\theta}(x) = \log \left( \int_z P_{\theta}(x, z) dz \right)$$

The logarithm is outside the integral, making it extremely difficult to compute or differentiate. We cannot simply push the logarithm inside the integral. This intractability prevents us from directly optimizing the log-likelihood for most complex models.

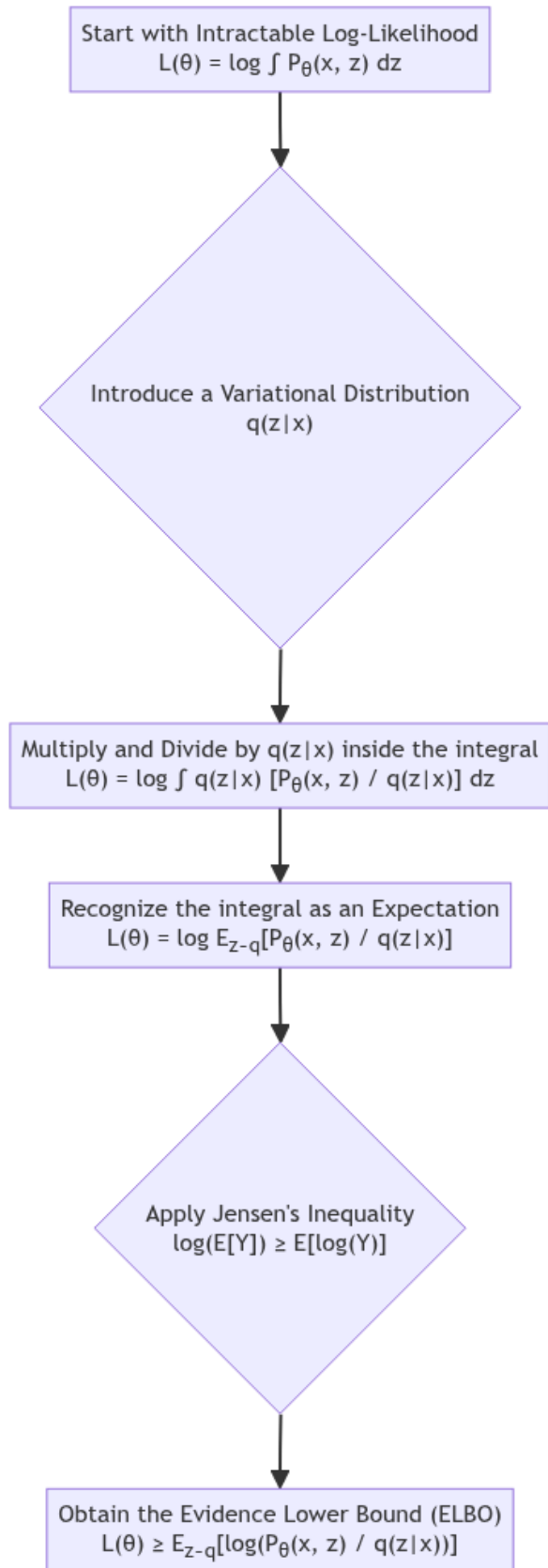
### The Evidence Lower Bound (ELBO)

To solve the problem of the intractable log-likelihood, we introduce a powerful technique from variational inference: deriving and optimizing a **lower bound** on the log-likelihood. This bound is known as the **Evidence Lower Bound (ELBO)**.

## **Derivation of the ELBO**

The derivation (starting at 09:08) is a cornerstone of modern generative modeling.

**Derivation Flowchart:**



### Step 1: Introduce a Variational Distribution $q(z|x)$ (11:04)

We introduce an arbitrary but tractable probability density function  $q(z|x)$  over the latent variable  $z$ . This function, often called the **variational posterior** or **inference network**, is our “guess” or approximation for the true posterior  $P_\theta(z|x)$ .

### Step 2: Rewrite the Log-Likelihood (12:45)

We rewrite the log-likelihood by multiplying and dividing the term inside the integral by  $q(z|x)$ :

$$L(\theta) = \log \left( \int_z P_\theta(x, z) dz \right) = \log \left( \int_z q(z|x) \frac{P_\theta(x, z)}{q(z|x)} dz \right)$$

### Step 3: Frame as Log of an Expectation (14:22)

The integral can now be seen as the expectation of the term  $\frac{P_\theta(x, z)}{q(z|x)}$  with respect to the distribution  $q(z|x)$ :

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

### Step 4: Apply Jensen’s Inequality (16:22)

The logarithm is a **concave function**. Jensen’s inequality states that for any concave function  $f$  and random variable  $Y$ ,  $f(\mathbb{E}[Y]) \geq \mathbb{E}[f(Y)]$ . Applying this to our expression:

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

This gives us the final inequality:

$$L(\theta) \geq \mathbb{E}_{z \sim q(z|x)} [\log P_\theta(x, z) - \log q(z|x)]$$

### The ELBO as the New Objective

The right-hand side of this inequality is the **Evidence Lower Bound (ELBO)**. It is a lower bound on the true log-likelihood (the evidence) that we originally wanted to maximize.

**Definition: The Evidence Lower Bound (ELBO) (20:26)** The ELBO, denoted  $J_\theta(q)$ , is a function of both the model parameters  $\theta$  and the variational distribution  $q$ .

$$\text{ELBO}(\theta, q) \equiv J_\theta(q) = \mathbb{E}_{z \sim q(z|x)} [\log P_\theta(x, z) - \log q(z|x)]$$

Since the ELBO is a lower bound on the log-likelihood, maximizing the ELBO will push the log-likelihood up as well. The key advantage is that the ELBO is **tractable**, as the expectation is now outside the logarithm, allowing for estimation via sampling.

The new optimization problem is a joint maximization over both  $\theta$  and  $q$ :

$$\theta^*, q^* = \arg \max_{\theta, q} J_\theta(q)$$

This is the foundational optimization problem solved in many modern generative models.

## Key Mathematical Concepts

### 1. Maximum Likelihood Estimation (MLE) for Latent Variable Models

The goal is to find parameters  $\theta^*$  that maximize the expected log-likelihood of the data.

$$\theta^* = \arg \max_{\theta} \mathbb{E}_{x \sim P_x} [\log P_{\theta}(x)]$$

where the log-likelihood (evidence) is:

$$\log P_{\theta}(x) = \log \left( \int_z P_{\theta}(x, z) dz \right)$$

### 2. Jensen's Inequality

For a concave function  $f$  (like  $\log(x)$ ) and a random variable  $Y$ :

$$f(\mathbb{E}[Y]) \geq \mathbb{E}[f(Y)]$$

This inequality is the key to deriving the lower bound. At (16:43), the instructor applies it to the log-likelihood expression:

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

### 3. The Evidence Lower Bound (ELBO)

The ELBO is the resulting lower bound on the log-likelihood (evidence).

$$\mathcal{L}(\theta) = \log P_{\theta}(x) \geq \mathbb{E}_{z \sim q(z|x)} \left[ \log \frac{P_{\theta}(x, z)}{q(z|x)} \right] = J_{\theta}(q)$$

This is the central equation derived in the lecture.

---

## Visual Elements from the Video

The lecture is presented on a digital whiteboard. The key visual elements are the handwritten mathematical derivations.

- **(01:05 - 02:24)** The instructor writes down the initial problem setup: the given data  $D$ , the i.i.d. assumption, and the definition of the latent variable model  $P_{\theta}(x) = \int_z P_{\theta}(x, z) dz$ .
  - **(02:24 - 05:25)** The goal is written as minimizing the KL divergence, which is then expanded and simplified to the maximum log-likelihood objective.
  - **(12:45 - 17:43)** The core derivation of the ELBO is shown, starting with the log-likelihood, introducing  $q(z|x)$ , and applying Jensen's inequality. The instructor explicitly writes out the inequality  $\log \mathbb{E}(\cdot) \geq \mathbb{E} \log(\cdot)$  to emphasize the step.
  - **(20:07 - 20:32)** The term  $J_{\theta}(q)$  is formally defined as the **Evidence Lower Bound (ELBO)**, and the instructor circles the terms "Evidence" and "Lower Bound" to connect them to the name.
  - **(21:52 - 22:08)** The distribution  $q(z|x)$  is identified as the **variational latent posterior**.
- 

## Self-Assessment for This Video

1. **Question:** What is the primary goal when training a latent variable model in a probabilistic framework?

- **Answer:** The goal is to adjust the model's parameters ( $\theta$ ) to maximize the likelihood of the observed data, which is equivalent to minimizing the KL divergence between the true data distribution and the model's distribution.
2. **Question:** Why is it often difficult to directly maximize the log-likelihood  $\log P_\theta(x)$  for latent variable models?
    - **Answer:** Because calculating  $P_\theta(x)$  requires marginalizing (integrating or summing) over the latent variable  $z$ . This results in an expression of the form  $\log(\int \dots)$ , which is computationally intractable for complex models.
  3. **Question:** What is Jensen's inequality, and how does it help in this context?
    - **Answer:** Jensen's inequality states that for a concave function  $f$ ,  $f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$ . Since  $\log$  is a concave function, we can apply this inequality to the log-likelihood expression (rewritten as a log of an expectation) to create a tractable lower bound (the ELBO).
  4. **Question:** What is the ELBO, and what two components does its optimization depend on?
    - **Answer:** The ELBO (Evidence Lower Bound) is a tractable lower bound on the log-likelihood of the data. The optimization of the ELBO is a joint maximization problem that depends on both the model parameters  $\theta$  and the parameters of the chosen variational distribution  $q$ .
- 

## Key Takeaways from This Video

- **Universal Principle:** The method of maximizing a lower bound on the log-likelihood (the ELBO) is a general and powerful principle for training a wide variety of latent variable models.
- **From Intractable to Tractable:** The core challenge in training these models is the intractability of the log-likelihood. The ELBO provides a tractable surrogate objective function that can be optimized using techniques like stochastic gradient descent.
- **The Role of Variational Inference:** By introducing an auxiliary variational distribution  $q(z|x)$ , we transform the difficult problem of maximizing  $\log P_\theta(x)$  into a more manageable joint optimization problem of maximizing the ELBO with respect to both the model and the variational distribution.
- **ELBO as a Foundation:** The ELBO is not just a mathematical trick; it is the fundamental objective function for models like Variational Autoencoders (VAEs) and forms the basis for understanding more advanced generative models.

## Visual References

The initial equation for the log-likelihood,  $\log p(x)$ , is shown. The instructor explains why this expression, which involves a logarithm of an integral over the latent variables, is computationally intractable, setting up the core problem that the ELBO solves. (at 02:15):





General principle for Learning Latent var. Models

Suppose, we are given Data

$$D = \{x_i\}_{i=1}^n \sim \text{iid } P_x$$

$$p_\theta(x) = \int_z p_\theta(x, z) dz, \text{ bc the}$$






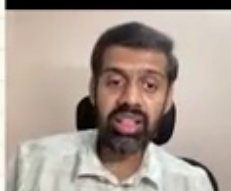
The key mathematical step of the derivation where Jensen's Inequality is applied. The visual shows the log function being moved inside the expectation, which is the crucial move that establishes a lower bound on the original log-likelihood. (at 05:40):

Independent of  $\theta$

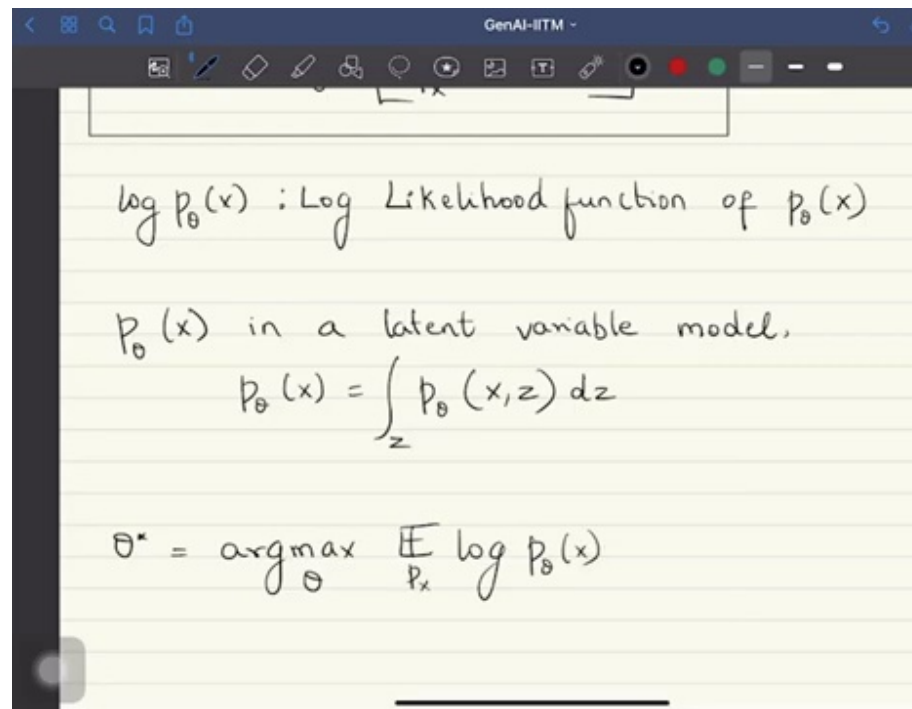
$$= \argmin_{\theta} \left[ - \int_x p_x(x) \log p_\theta(x) dx \right]$$

$$\theta^* = \argmax_{\theta} \left[ \mathbb{E}_{P_x} \log p_\theta(x) \right]$$





The final, fully derived equation for the Evidence Lower Bound (ELBO). This slide presents the objective function in its common form, breaking it down into the reconstruction term and the KL



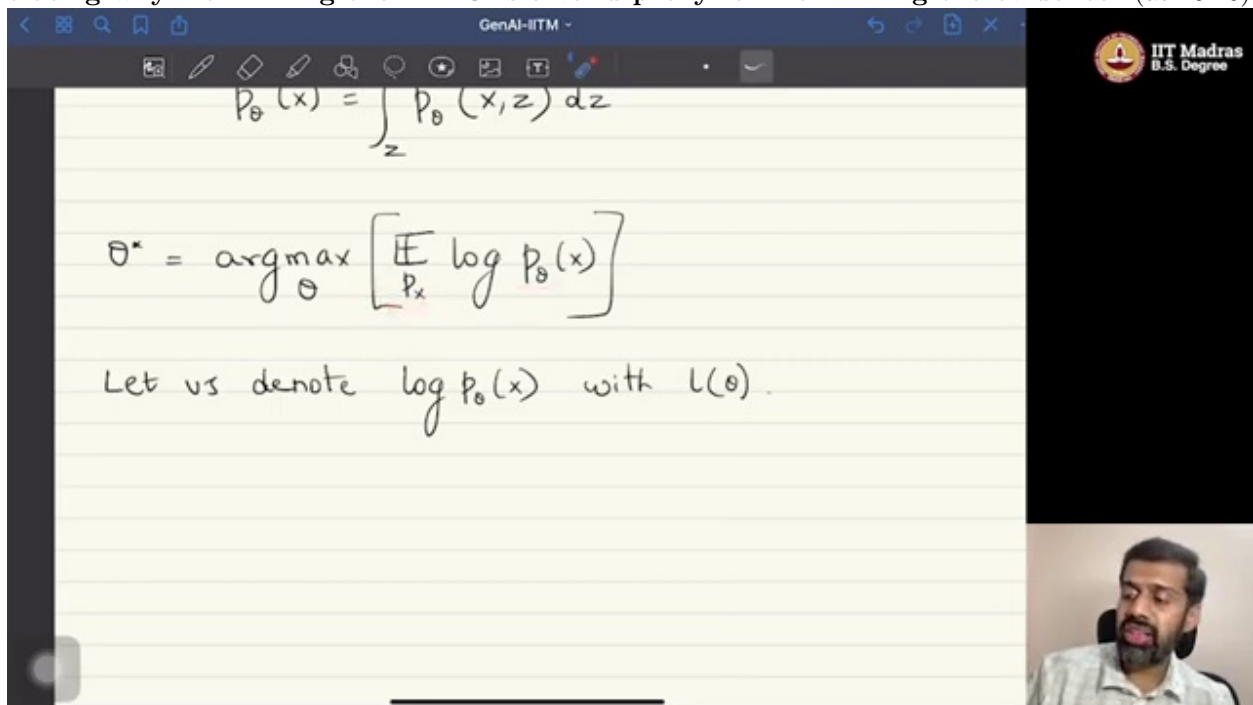
$\log p_0(x)$  : Log Likelihood function of  $p_0(x)$   
 $p_0(x)$  in a latent variable model,  

$$p_0(x) = \int_z p_0(x, z) dz$$
  

$$\theta^* = \operatorname{argmax}_{\theta} \mathbb{E}_{p_x} \log p_0(x)$$

divergence regularization term. (at 08:30):

A summary diagram visually representing the relationship between the true log-evidence, the ELBO, and the KL divergence. It clearly shows that  $\log p(x) = \text{ELBO} + \text{KL}$ , illustrating why maximizing the ELBO is a valid proxy for maximizing the evidence. (at 10:10):



$$p_0(x) = \int_z p_0(x, z) dz$$
  

$$\theta^* = \operatorname{argmax}_{\theta} \left[ \mathbb{E}_{p_x} \log p_0(x) \right]$$
  
 Let us denote  $\log p_0(x)$  with  $L(\theta)$ .