# Deep learning
episode 12
# Bayesian methods in DL

Yandex
Data Factory

LAMBDA

British Hedgehog
Preservation Society

What is probability?

# What is probability?

What does it mean if P(event) = 0.25?

How evaluate P(coin lands heads)?

A coin was flipped 2 times

Heads both times

P(heads) = ?

# Frequentist Vs Bayesian

- Probability is objective nondeterminism
- There is no prior, there's data

- Hypothesis testing
- Quantum Physics

- Trust Regions
- Maximum Likelihode Estimate

- Probability is subjective ignorance
- Through prior I gain strength

- Regularization
- Structured learning

- Posterior distributions
- Maximum a-posteriori

# Bayes theorem

## Conditional probability

*Probability of **a** given **b***

$$P(a|b)=\frac{P(ab)}{P(b)}$$

## Bayes theorem

$$P(a|b)=\frac{P(ab)}{P(b)}=\frac{P(b|a)\cdot P(a)}{P(b)}$$

# Bayes theorem

## Conditional probability

*Probability of **a** given **b***

$$P(a|b)=\frac{P(ab)}{P(b)}$$

## Bayes theorem

$$P(ab)=P(b|a)\cdot P(a)$$

$$P(a|b)=\frac{P(ab)}{P(b)}=\frac{P(b|a)\cdot P(a)}{P(b)}$$

# Marginalization

$$P(b) = \int_a P(b|a) \cdot P(a)\, da = E_{a \sim P(a)} P(b|a)$$

$$P(a|b) = \frac{P(b|a) \cdot P(a)}{P(b)} = \frac{P(b|a) \cdot P(a)}{\int_a P(b|a) \cdot P(a)\, da}$$

# Logistic regression

- Binary classification
- Data: X (objects) Y(answers)
- Model:

What parameters does logreg have?

# Logistic regression

- Binary classification
- Data: X (objects) Y(answers)
- Model:

$$\theta=[w_\theta, b_\theta]$$

How do we estimate $P(y|x,\theta)$?

# Logistic regression

- Binary classification
- Data: X (objects) Y(answers)
- Model:

$$\theta = [w_\theta, b_\theta]$$

$$P(y|x,\theta) = \sigma(w_\theta \cdot x + b_\theta)$$

$$P(\bar{y}|x,\theta) = 1 - \sigma(w_\theta \cdot x + b_\theta)$$

# Logistic regression

- Binary classification
- Data: X (objects) Y(answers)
- Model:

$$\theta = [w_\theta, b_\theta]$$

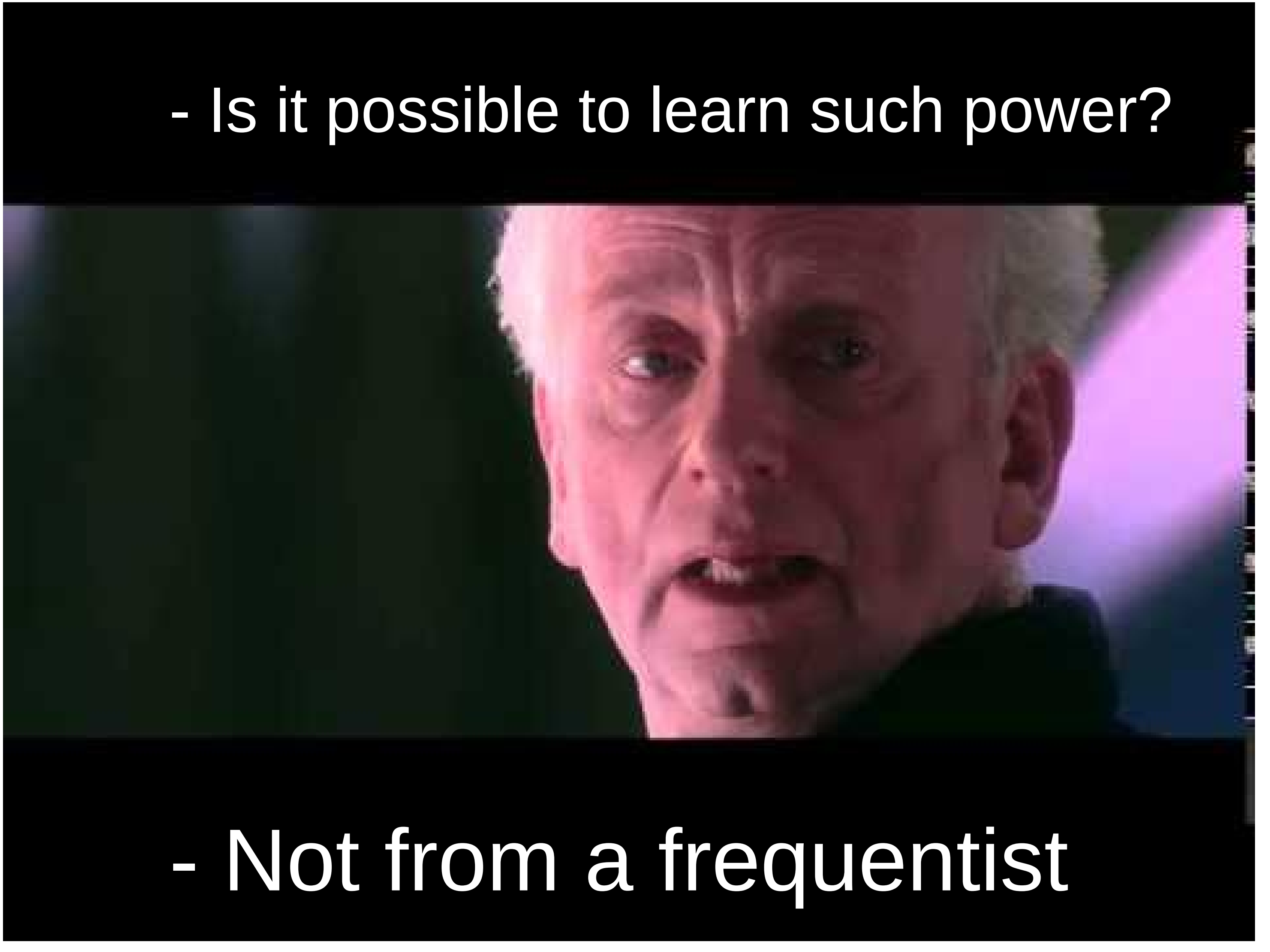
$$P(y|x,\theta) = \sigma(w_\theta \cdot x + b_\theta)$$

$$P(\bar{y}|x,\theta) = 1 - \sigma(w_\theta \cdot x + b_\theta)$$

Objective:

$$\theta' = \underset{\theta}{argmax}\, P(\theta|X, Y)$$

# Frequentist Vs Bayesian



# Maximum a-posteriori

Objective:

$$\theta' = \underset{\theta}{argmax}\, P(\theta | X, Y)$$

$$P(\theta | X, Y) = \frac{P(X, Y | \theta) \cdot P(\theta)}{P(X, Y)}$$

# Maximum a-posteriori

Objective:

$$\theta' = \underset{\theta}{argmax}\, P(\theta | X, Y)$$

**const(θ)**

$$P(Y|X,\theta)\cdot P(X|\theta)$$

$$P(\theta|X,Y) = \frac{P(X,Y|\theta)\cdot P(\theta)}{P(X,Y)} \sim P(Y|X,\theta)\cdot P(\theta)$$

**const(θ)**

# Maximum a-posteriori

Objective:

$$\theta' = \underset{\theta}{argmax}\, P(\theta | X, Y) = \underset{\theta}{argmax}\, P(Y | X, \theta) \cdot P(\theta)$$

Likelihood:

$$P(Y | X, \theta) = \prod_i P(y_i | x_i, \theta)$$

Result:

$$\underset{\theta}{argmax} \prod_i P(y_i | x_i, \theta) \cdot P(\theta)$$

Product of many <1 terms,
Computationally unstable

**Quiz**: can we optimize something more stable?

# Maximum a-posteriori

Objective:

$$\theta' = \underset{\theta}{argmax}\, P(\theta|X, Y) = \underset{\theta}{argmax}\, P(Y|X, \theta) \cdot P(\theta)$$

$$\underset{\theta}{argmax} \log \left[ \prod_i P(y_i|x_i, \theta) \cdot P(\theta) \right]$$

$$\underset{\theta}{argmax} \sum_i \log P(y_i|x_i, \theta) + \log P(\theta)$$

# Logistic regression

- Model:

$$P(y_i | x_i, \theta) = \begin{matrix} if\ y_i = 1, \sigma(w_\theta \cdot x + b_\theta) \\ if\ y_i = 0, 1 - \sigma(w_\theta \cdot x + b_\theta) \end{matrix}$$

# Logistic regression

- Model:

$$P(y_i|x_i,\theta)=\begin{matrix} if\ y_i=1, \sigma(w_\theta \cdot x+b_\theta) \\ if\ y_i=0, 1-\sigma(w_\theta \cdot x+b_\theta) \end{matrix}$$

$$\log P(y_i|x_i,\theta)=\begin{matrix} if\ y_i=1, \log\sigma(w_\theta \cdot x+b_\theta) \\ if\ y_i=0, \log(1-\sigma(w_\theta \cdot x+b_\theta)) \end{matrix}$$

- Replace if with multiplication

$$y_i \cdot \log\sigma(w_\theta \cdot x+b_\theta)+(1-y_i)\cdot\log(1-\sigma(w_\theta \cdot x+b_\theta))$$

# Logistic regression

- Model:

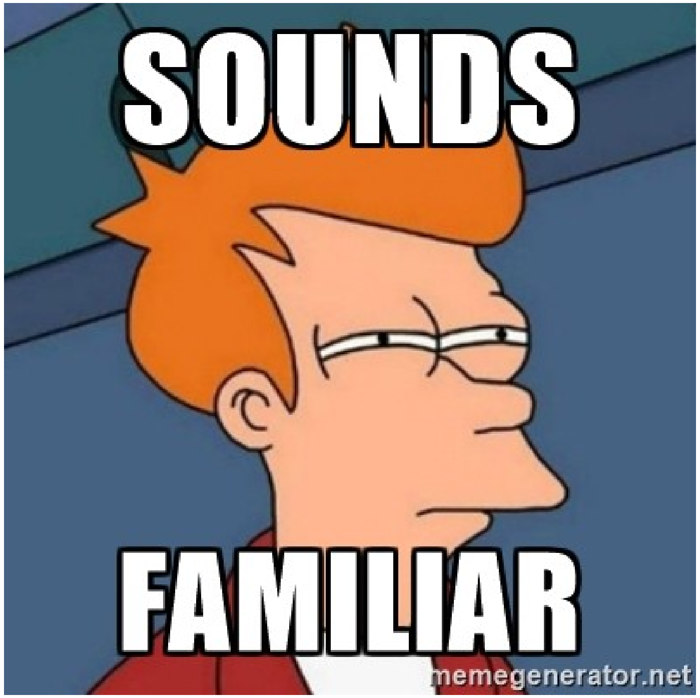
$$P(y_i|x_i,\theta) = \begin{matrix} if\ y_i = 1, \sigma(w_\theta \cdot x + b_\theta) \\ if\ y_i = 0, 1 - \sigma(w_\theta \cdot x + b_\theta) \end{matrix}$$

$$\underset{\theta}{argmax} \sum_i \log P(y_i|x_i,\theta) + \log P(\theta)$$

# Logistic Regression

- Assume uniform prior (const)
- Replace max(a) by min(-a)

$$-\sum_i P(y_i|x_i,\theta) = y_i \cdot \sigma(w_\theta \cdot x + b_\theta) + (1 - y_i) \cdot (1 - \sigma(w_\theta \cdot x + b_\theta))$$



# Prior

- Information about weights before observation
- Which of these weights you'd prefer?

First set of weights:

-0.554, 2.726, 0.999, 2.573, -0.694, 0.323, -1.903, -0.070

Second set of weights:

154016218671.074, 133023030621.400, 72847832520.938, 130909237163.079, -134435263422.709, 72550546946.769, 121468470400.514, -55724178429.301

# Prior

- Weights should be small,

$$P(0.1) > P(10^5)$$

- Which distributions support that?

# Prior

- Weights should be small,

$$P(0.1) > P(10^5)$$

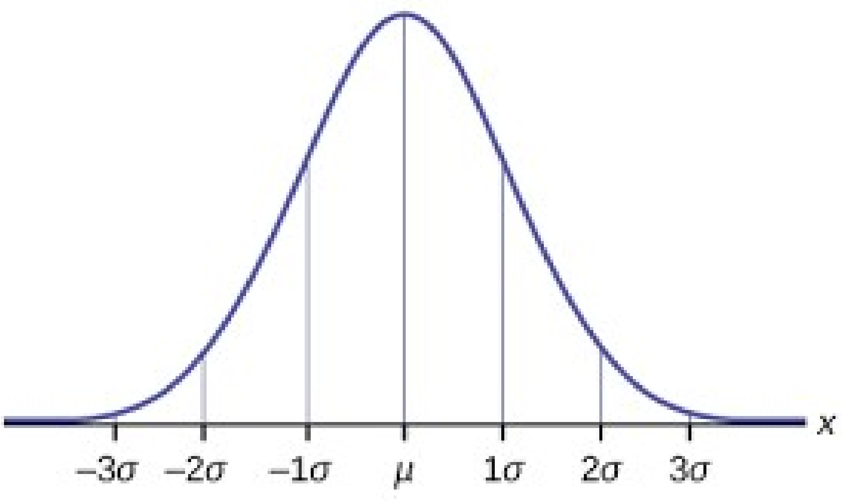
- Which distributions support that?
- Actually, all kinds of distributions, but we'll name a few...

# Prior

- Weights should be small,

P(0.1) > P(10^5)

- Which distributions support that?
- Normal




$$P(\theta)=\frac{1}{\sigma\cdot\sqrt{2\pi}}\cdot e^{\frac{-(\theta-\mu)^2}{\sigma^2}}$$

- Assumuing mean = 0

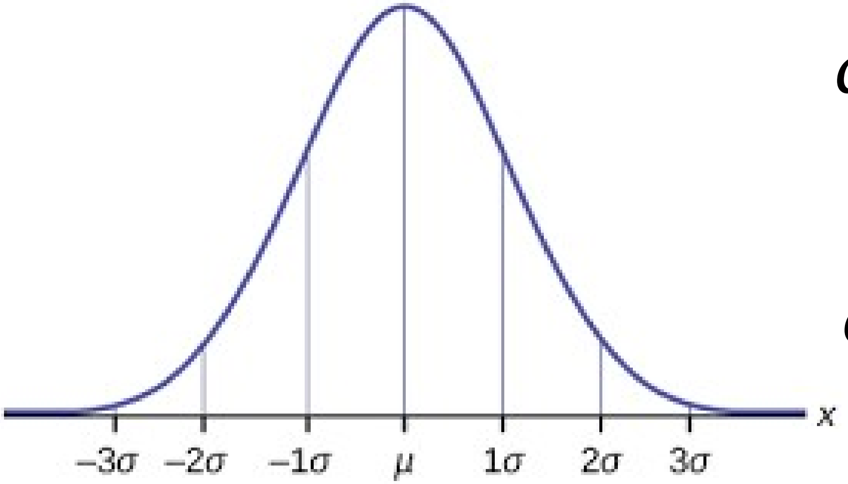
$$\log P(\theta)\sim\frac{-\theta^2}{\sigma^2}$$

# Prior

- Weights should be small,

P(0.1) > P(10^5)

- Which distributions support that?
- Normal

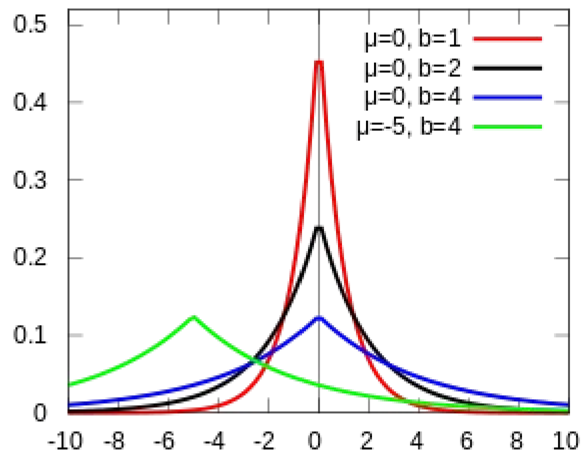



$$\underset{\theta}{argmax}\left[P\left(Y|X,\theta\right)+\frac{-1}{\sigma^2}\sum_j \theta_j^2\right]$$

$$\underset{\theta}{argmin}\left[-P\left(Y|X,\theta\right)+\frac{1}{\sigma^2}\sum_j \theta_j^2\right]$$

# Moar

- Laplacian prior




$$P(\theta)=\frac{1}{2b}\cdot e^{\frac{-|\theta-\mu|}{b}}$$

- Guess the final term for loss... :)

- Other: exponential for ensembling, mixtures, etc.

# Linear regression

- Regression
- Predict P(y|x) with a gaussian
  - mean = wx+b
  - variance = 1

$$P(y_i|x,\theta)=\frac{1}{\sqrt{2\pi}}\cdot e^{-(y_i-[w_\theta\cdot x_i+b_\theta])^2}$$

$$\log P(y_i|x,\theta)=what\,?$$

# Linear regression

- Regression
- Predict P(y|x) with a gaussian
  - mean = wx+b
  - variance = 1

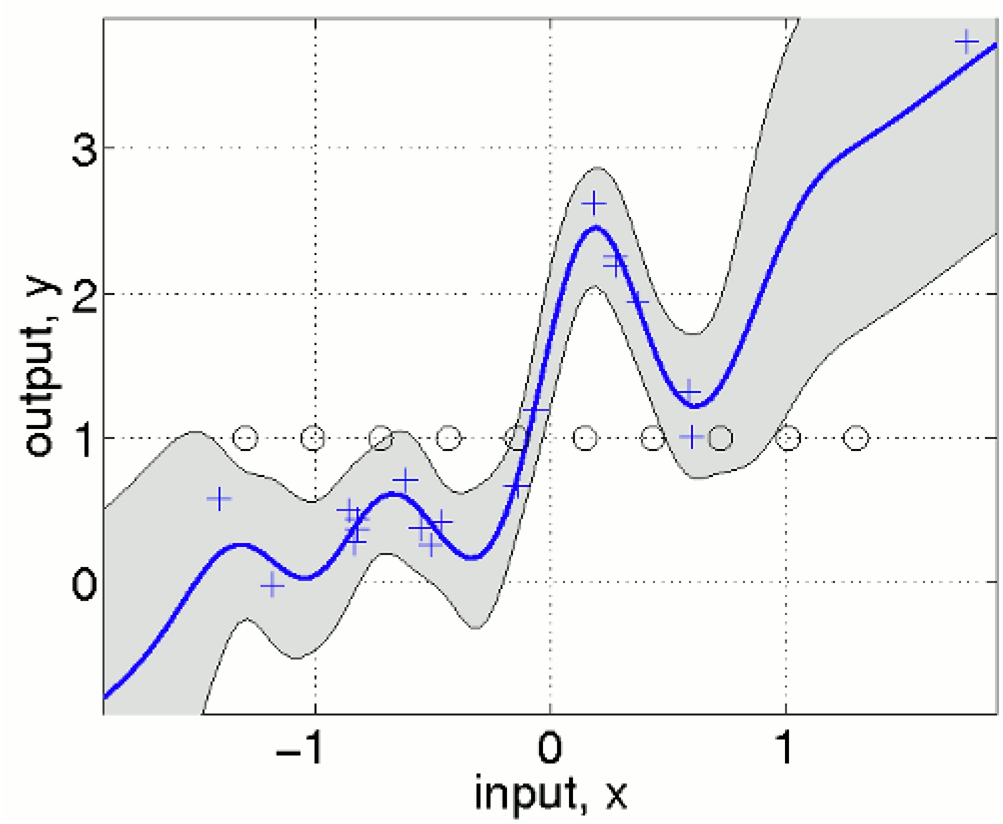
$$P(y_i|x,\theta)=\frac{1}{\sqrt{2\pi}}\cdot e^{-(y_i-[w_\theta\cdot x_i+b_\theta])^2}$$

$$\log P(Y|X,\theta)\sim-\sum_i (y_i-[w_\theta\cdot x_i+b_\theta])^2$$

**Hence, MSE**

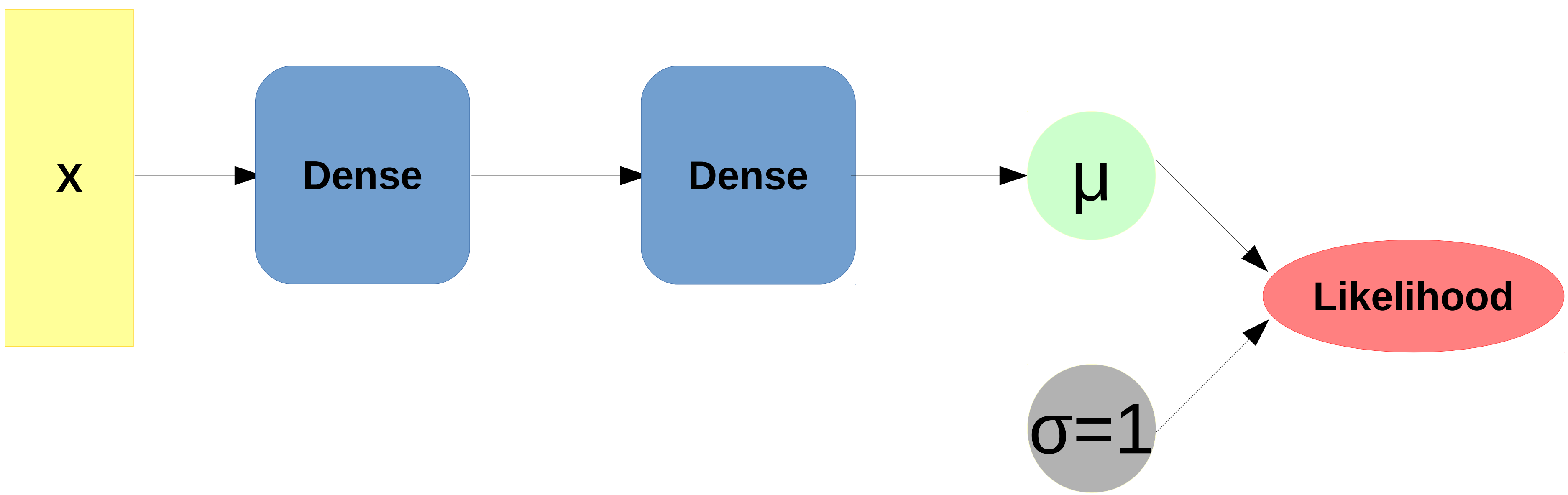
# Why do we need certainty?

- To which extent do we trust our model?
- Sampling from generative models



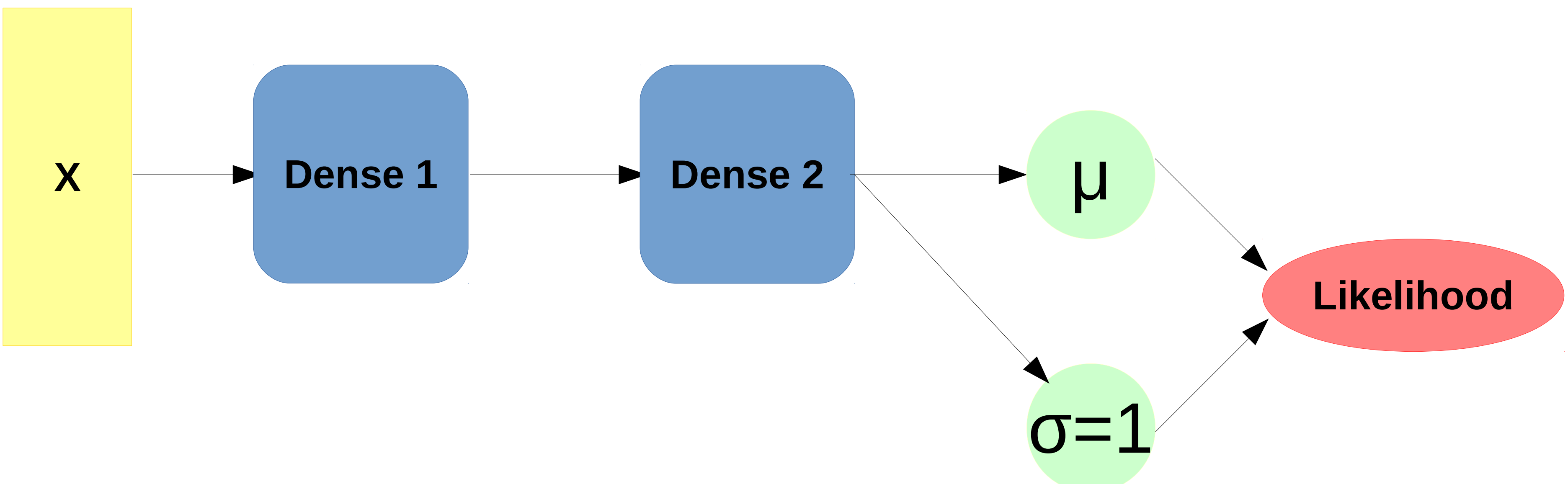
# Predicting distributions

- Before:



# Predicting distributions

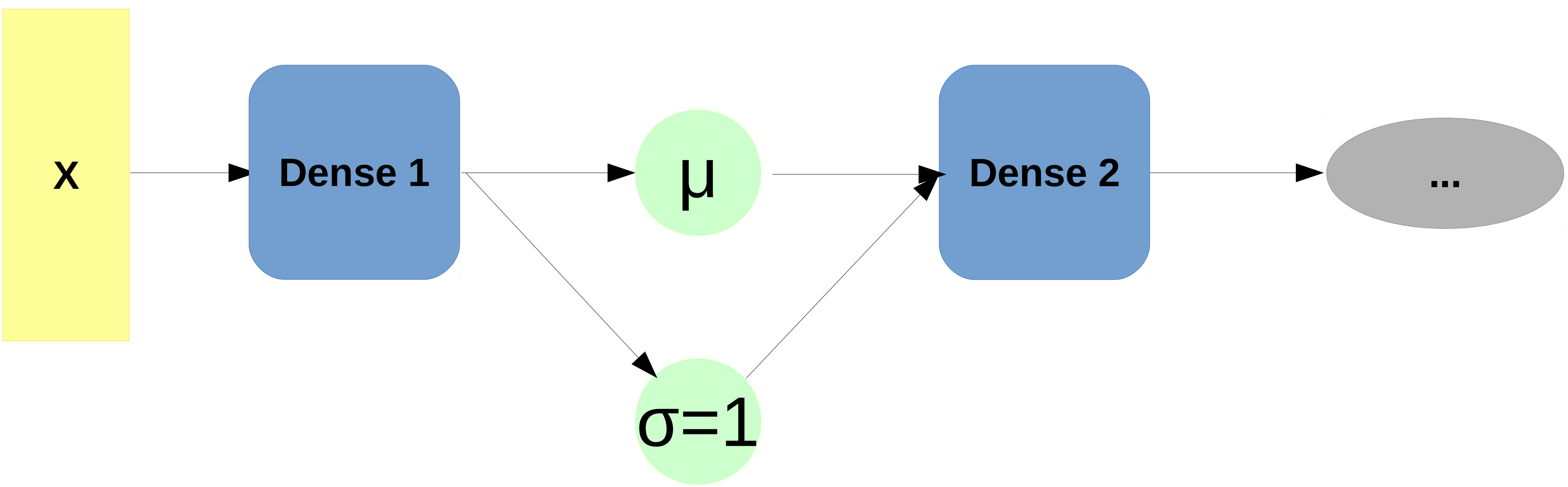
- After:




- Sigma = how certain are you?

# Predicting distributions

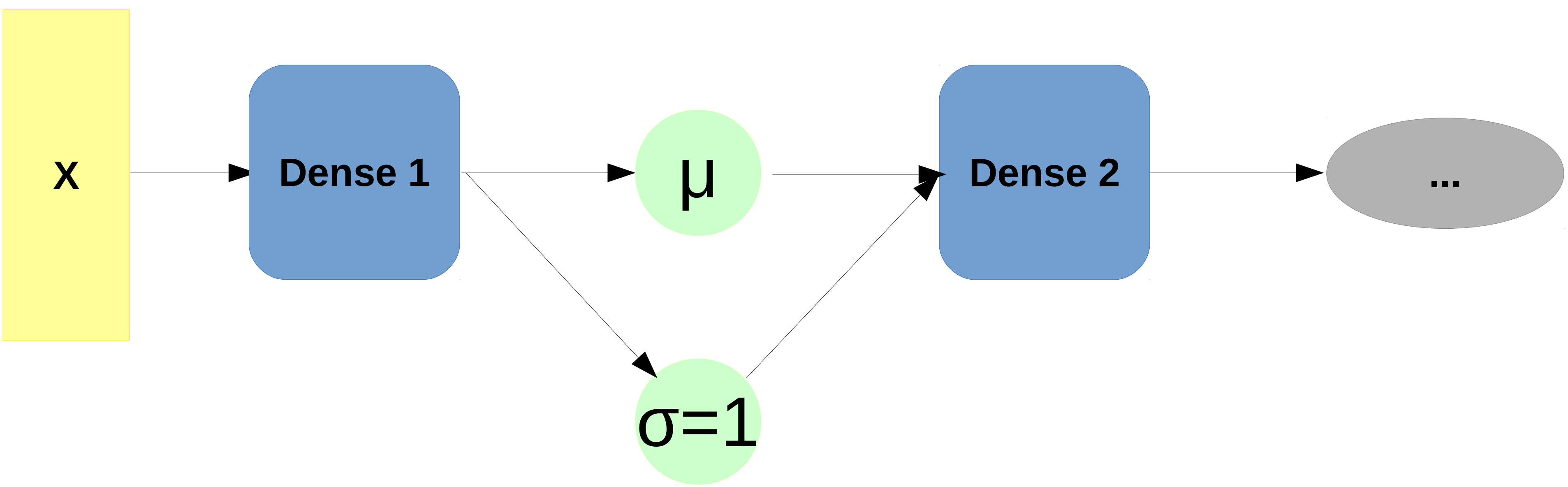
- Distribution on activations:




- How do we train that thing?

# Predicting distributions

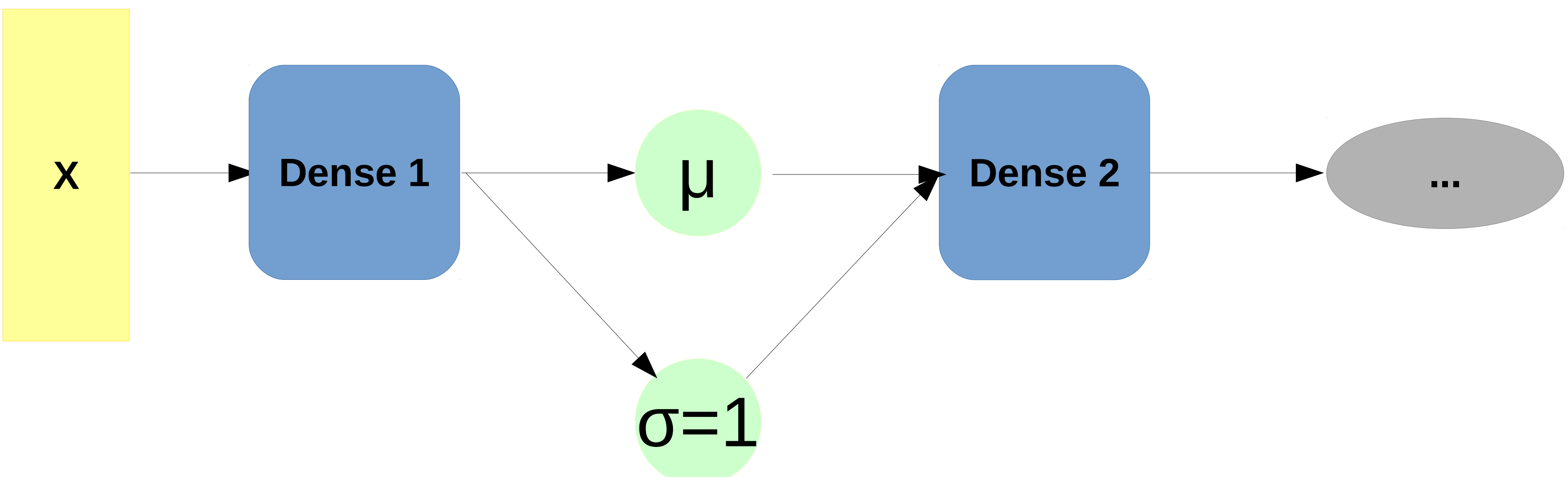
- Distribution on activations:




$$\log P(y|x,\theta) = E_{a\sim\mu(x),\sigma(x)} dense2(a)$$

# Predicting distributions

- Distribution on activations:




$$\log P(y|x,\theta) = E_{a\sim\mu(x),\sigma(x)}\, dense2(a)$$

weights

# Reparameterization trick

- Idea:

Replace parameterized distribution

with some expression over noise

$$N(\mu,\sigma)=\mu+\sigma\cdot N(0,1)$$

$$\frac{\delta[\mu+\sigma\cdot N(0,1)]}{\delta\mu,\delta\sigma} \textit{ is okay}$$

Works with many(but not all) distributions

# How to train?

- Reparameterization trick:

Was:

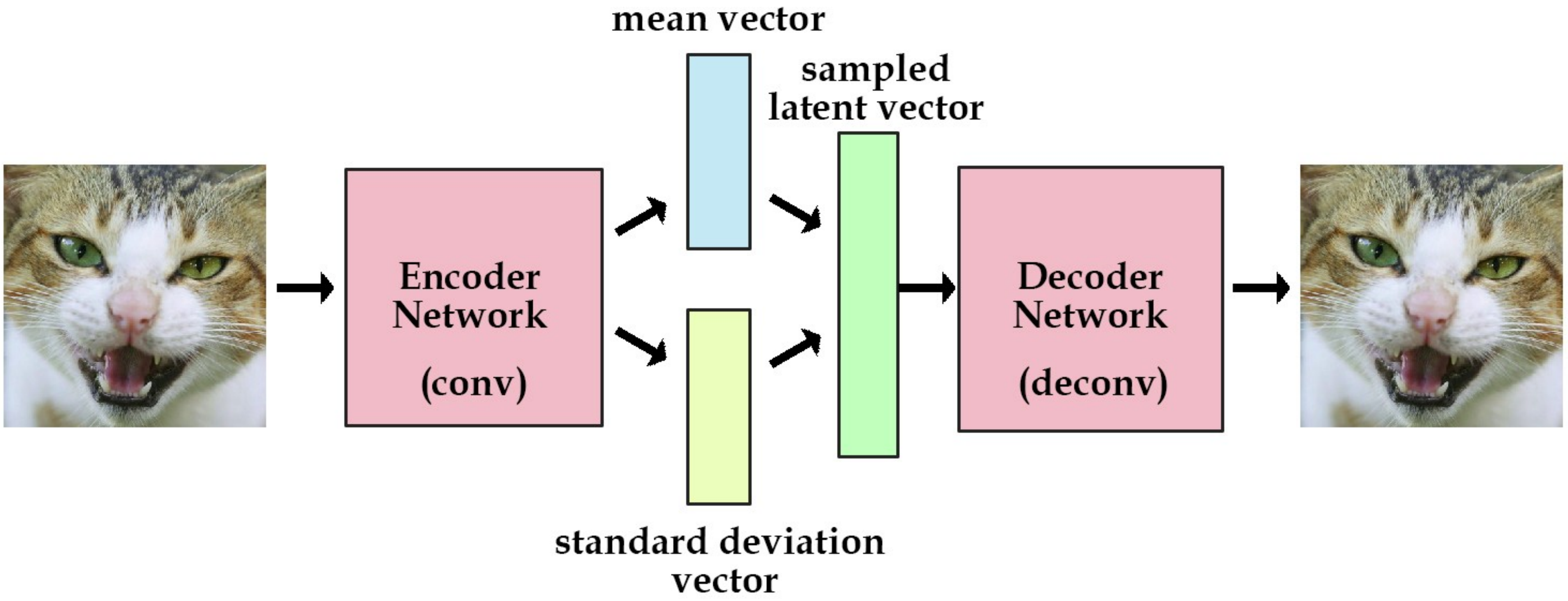
$$\log P(y|x,\theta) = E_{a\sim\mu(x),\sigma(x)}\, dense2(a)$$

Now:

$$\log P(y|x,\theta) = E_{\xi\sim N(0,1)}\, dense2(\mu(x)+\sigma(x)\cdot\xi)$$

**noise**

# Seminar announcement



# Seminar announcement