Variational Inference

From basics to modern applications

Adam Arany

17 October 2022
3rd AIDD School, Leuven, Belgium

#IStandWithTheWomenOfIran

Overview

- Bayesian models and inference recap (from Lugano)
  - Bayesian probability
  - Graphical model notation
  - Predictive inference
- Variational Inference Basics
  - Calculus of Variation
  - Derivation of the Evidence Lower Bound (ELBO)
  - Mean Field Approximation
  - Gradient based VI
- Variational Autoencoders
  - Amortized inference
  - Reparametrization trick
- Bayesian NNs

Bayesian Probability

Recap from Lugano

Probability of the hypothesis *after observing the evidence*

"a posteriori"

Probability of the hypothesis *before observing the evidence*

"a priori"

$$\overbrace{P(H|D)}^{\text{posterior}} = \frac{\overbrace{P(D|H)}^{\text{likelihood}}\overbrace{P(H)}^{\text{prior}}}{\underbrace{P(D)}_{\text{marginal likelihood}}}$$

$$P(D) = \int P(D|H)P(H)dH$$

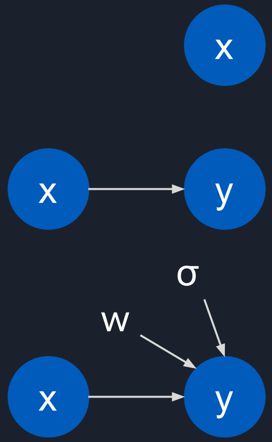
This integral is most often intractable!

- have an epistemic / subjectivist interpretation
- true independent of interpretations

Pierre-Simon Laplace

Elements of probabilistic models

Recap from Lugano



Random variables $X \sim \mathcal{N}(0,1)$

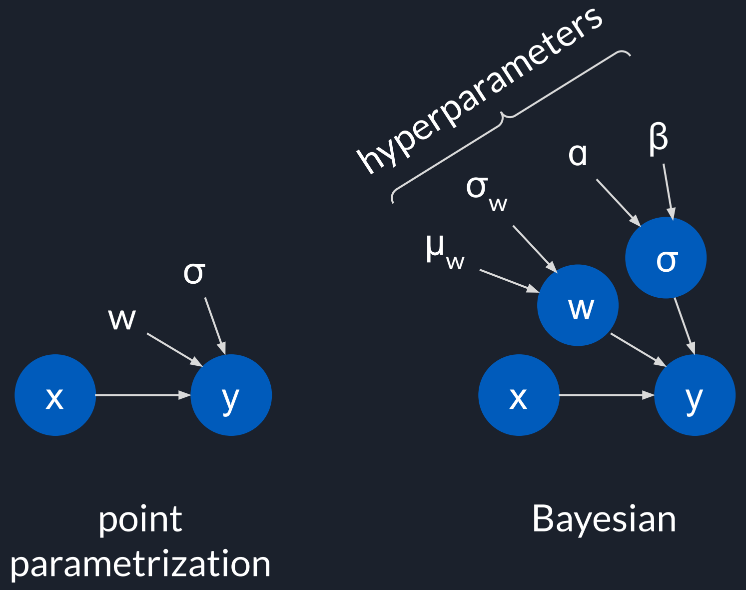
Dependencies $Y \sim \mathcal{N}(2X, 0.5)$

Parameters $Y \sim \mathcal{N}(wX, \sigma)$

Probabilistic graphical models (PGMs), Bayesian networks WARNING!

Bayesian models

Recap from Lugano

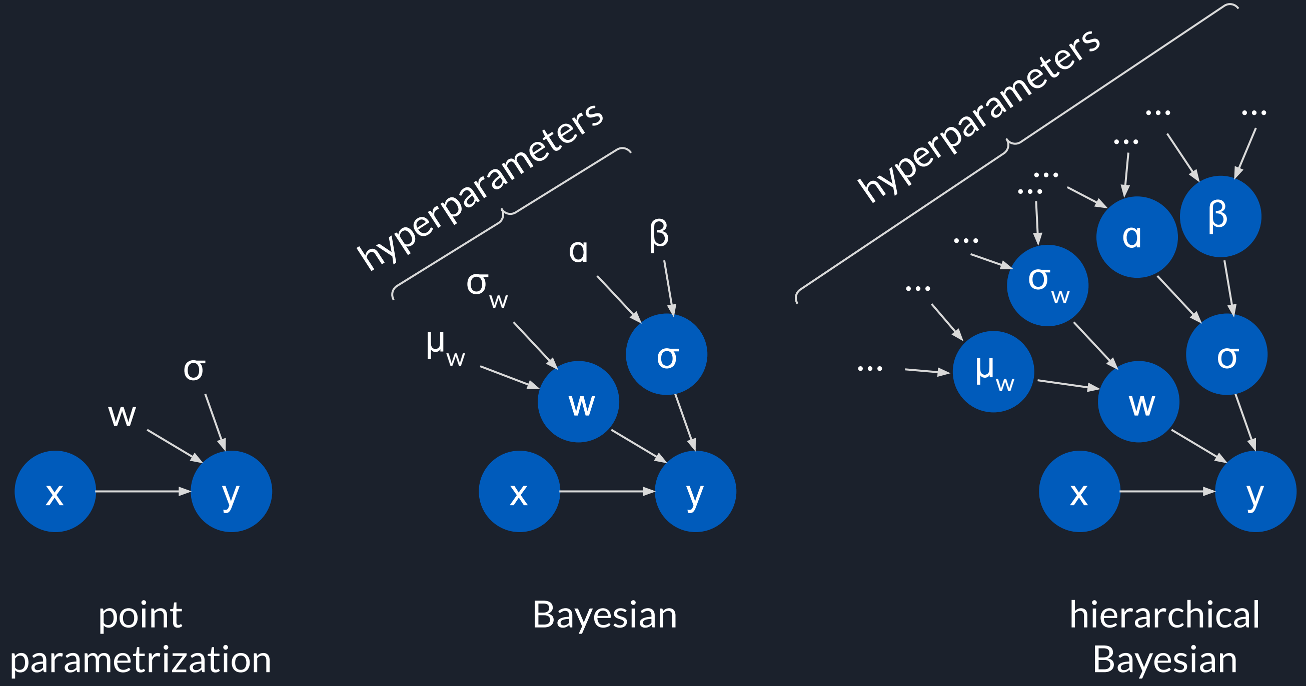


Bayesian treatment:

- Parameter is just a random variable
- We do not expect to find 'the real' parameter value exactly
- We search for the distribution of the parameters supported by the data.

Bayesian models

Recap from Lugano



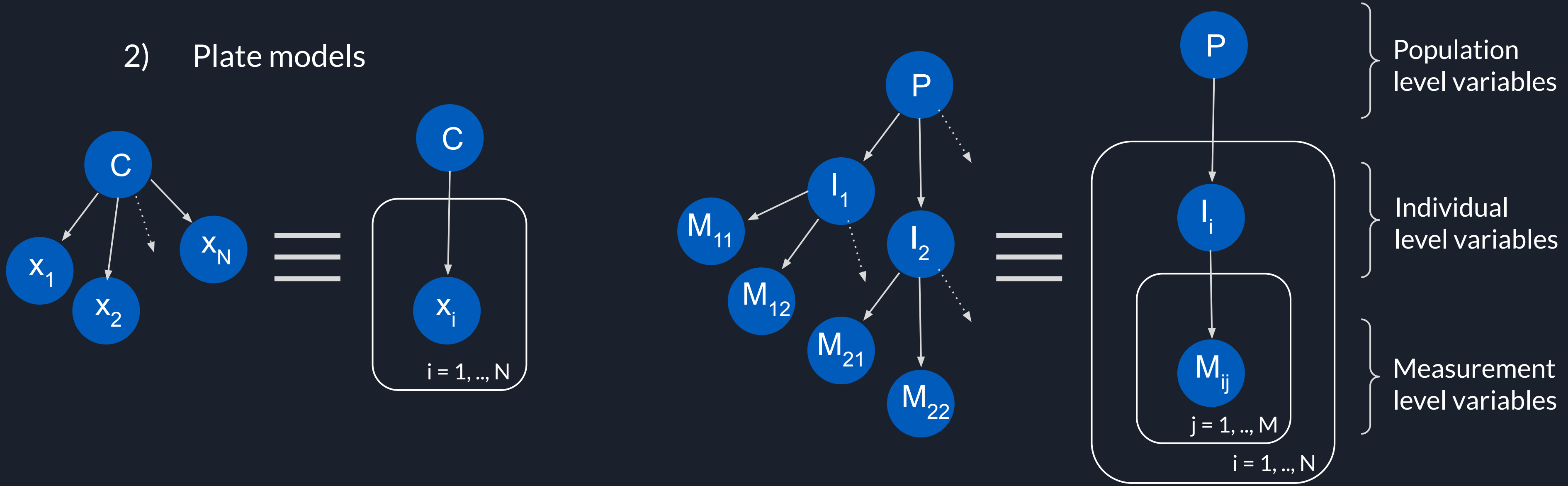
Frequently used shorthand notations

Recap from Lugano

1) Vector, matrix, tensor valued random variables

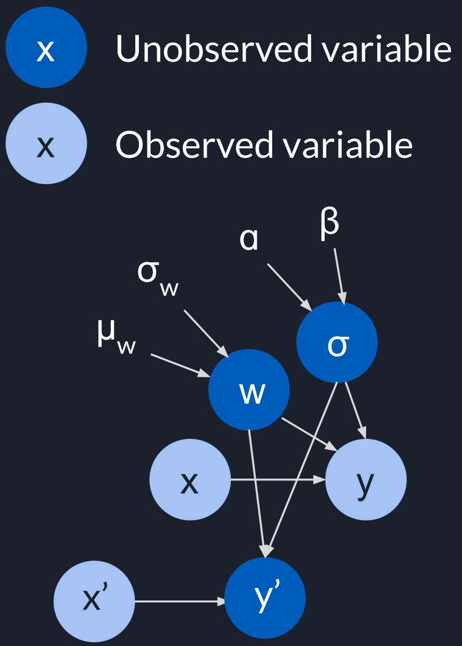
$$y \sim \mathcal{N}(W^\top x, \Sigma)$$

2) Plate models



Predictive inference

Recap from Lugano



Predicted outcome? Just another random variable

$$\underbrace{P(y'|\underbrace{x',x,y}_{\mathcal{D}})}_{\text{Posterior predictive distribution}} = P(y'|x',w,\sigma)\underbrace{P(w,\sigma|\underbrace{x,y}_{\mathcal{D}})}_{\text{(model) posterior}}$$

What is the mean?

$$\mathbb{E}[y'|x',\mathcal{D}] = \mathbb{E}_{P(w,\sigma|\mathcal{D})}[f_{w,\sigma}(x')]$$

"Bayesian model averaging"

Note the similarity with ensembles!

Calculus of variation

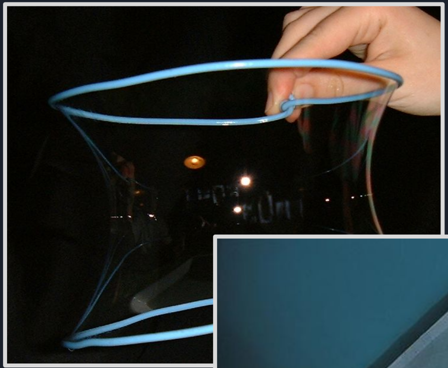
Calculus: concerned with functions $f(x)$
mapping "numbers" to "numbers"

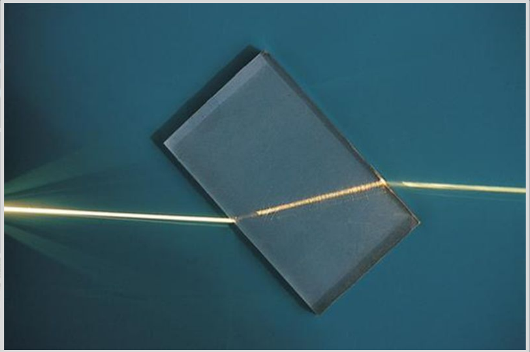
Variational calculus: concerned with $F[f]$ functionals
mapping functions to "numbers"

Searching for extremal points of F functional

Several physics problem:

- Brachistochrone problem
- Shape of hanging chain
- Soap bubbles
- Fermat principle
- Principle of least action / principle of stationary action





Calculus of variation

$$H[P]=\int_{-\infty}^{\infty} P(x)\log P(x)dx$$

Entropy

$$S[L]=\int_{t_1}^{t_2} L(q(t),\dot{q}(t))dt$$

Action

$$D[Q]=D_{KL}(Q\|P)=\int_{-\infty}^{\infty} Q(x)\log\frac{Q(x)}{P(x)}dx$$

Kullback-Leibler divergence

Calculus of variation

$$H[P] = \int_{-\infty}^{\infty} P(x) \log P(x) dx$$

Entropy

$$S[L] = \int_{t_1}^{t_2} L(q(t), \dot{q}(t)) dt$$

Action

$$D[Q] = D_{KL}(Q \| P) = \int_{-\infty}^{\infty} Q(x) \log \frac{Q(x)}{P(x)} dx$$

Kullback-Leibler divergence

Calculus of variation

Searching for a distribution (a function) Q, that minimize a KL-divergence (a functional) is a problem in the calculus of variation.

Therefore the inference method using this approximation is named Variational.

$$D[Q] = D_{KL}(Q \| P) = \int_{-\infty}^{\infty} Q(x) \log \frac{Q(x)}{P(x)} dx$$

Kullback-Leibler divergence

General inference problem

Given a model (here as a graphical model).
Define a set of observed variables E

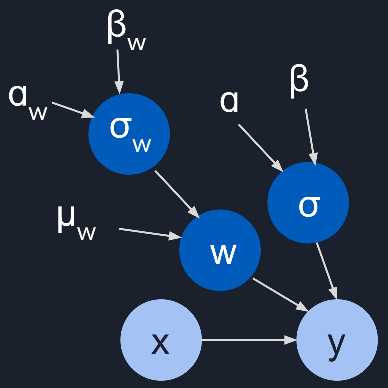
$$E = \{x, y\}$$

And the variables of interest H

$$H = \{w, \sigma, \sigma_w\}$$

We want to estimate the posterior of H conditioned on E, P(H|E)

$$p(w, \sigma, \sigma_w | x, y)$$



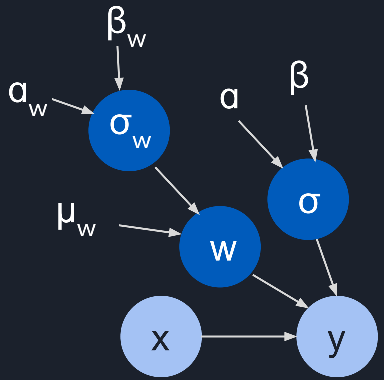
Variational inference

If we would have $p(w, \sigma, \sigma_w | x, y)$
in analytical form, our job would be done.

- Often there is no such analytical form
- Search for a function $q(w, \sigma, \sigma_w) \approx p(w, \sigma, \sigma_w | x, y)$

$$\min_{\phi} D_{KL}(q_\phi(w, \sigma, \sigma_w) || p(w, \sigma, \sigma_w | x, y))$$

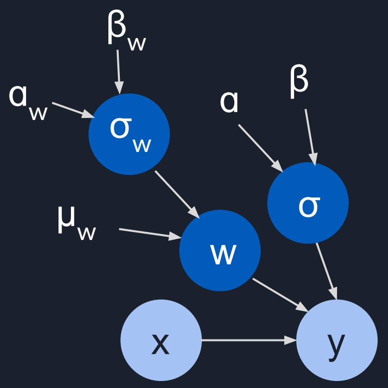
ϕ: variational parameter



Variational inference

If we would have $p(w,\sigma,\sigma_w|x,y)$
in analytical form, our job would be done.

- Often there is no such analytical form
- Search for a function $q(w,\sigma,\sigma_w) \approx p(w,\sigma,\sigma_w|x,y)$



$$\min_{\phi} D_{KL}(q_{\phi}(w,\sigma,\sigma_w) || p(w,\sigma,\sigma_w|x,y))$$

ϕ: variational parameter

Every evidence result in a different q!

Every inference query require optimization. This is a price we pay.

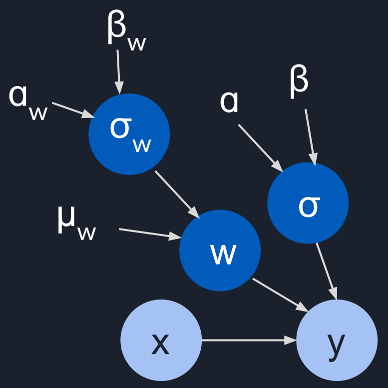
Variational inference

If we would have $p(w, \sigma, \sigma_w | x, y)$
in analytical form, our job would be done.

- Often there is no such analytical form
- Search for a function $q(w, \sigma, \sigma_w) \approx p(w, \sigma, \sigma_w | x, y)$

$$\min_{\phi} D_{KL}(q_{\phi}(w, \sigma, \sigma_w) || p(w, \sigma, \sigma_w | x, y))$$

ϕ: variational parameter



It can be shown that equivalently we can take the following objective:

$$\min_{\phi} \underbrace{D_{KL}(q_{\phi}(w, \sigma, \sigma_w) || p(w, \sigma, \sigma_w)) - \mathbb{E}_{q_{\phi}(w, \sigma, \sigma_w)}[\log p(x, y | w, \sigma, \sigma_w)]}_{\text{Evidence lower bound (ELBO)}}$$

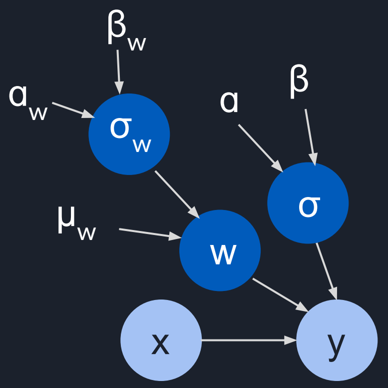
Variational inference

If we would have $p(w,\sigma,\sigma_w|x,y)$
in analytical form, our job would be done.

- Often there is no such analytical form
- Search for a function $q(w,\sigma,\sigma_w) \approx p(w,\sigma,\sigma_w|x,y)$

$$\min_\phi D_{KL}(q_\phi(w,\sigma,\sigma_w)||p(w,\sigma,\sigma_w|x,y))$$

ϕ: variational parameter



It can be shown that equivalently we can take the following objective:

$$\min_\phi \underbrace{D_{KL}(q_\phi(w,\sigma,\sigma_w)||p(w,\sigma,\sigma_w))}_{\text{Regularization}} - \underbrace{\mathbb{E}_{q_\phi(w,\sigma,\sigma_w)}[\log p(x,y|w,\sigma,\sigma_w)]}_{\text{Expected likelihood}}$$

Derivation of the ELBO

$$\min_{\phi} D_{KL}(q_{\phi}(H) \| p(H|E))$$

Derivation of the ELBO

$$\min_{\phi} D_{KL}(q_{\phi}(H) \| p(H|E))$$

$$\min_{\phi} \int q_{\phi}(H) \log \frac{q_{\phi}(H)}{p(H|E)} dH = \mathbb{E}_{q_{\phi}(H)}\left[\log q_{\phi}(H) - \log p(H|E)\right]$$

Derivation of the ELBO

$$\frac{p(E|H)p(H)}{P(E)}$$

$$\min_{\phi} D_{KL}(q_{\phi}(H) \| p(H|E))$$

$$\min_{\phi} \int q_{\phi}(H) \log \frac{q_{\phi}(H)}{p(H|E)} dH = \mathbb{E}_{q_{\phi}(H)}\left[\log q_{\phi}(H) - \log p(H|E)\right]$$

Derivation of the ELBO

$$\frac{p(E|H)p(H)}{P(E)}$$

$$\min_{\phi} D_{KL}(q_{\phi}(H) \| p(H|E))$$

$$\min_{\phi} \int q_{\phi}(H) \log \frac{q_{\phi}(H)}{p(H|E)} dH = \mathbb{E}_{q_{\phi}(H)}\left[\log q_{\phi}(H) - \log p(H|E)\right]$$

$$\min_{\phi} \mathbb{E}_{q_{\phi}(H)}\left[\log q_{\phi}(H) - \log p(H) - \log p(E|H) + \log p(E)\right]$$

Derivation of the ELBO

$$\frac{p(E|H)p(H)}{P(E)}$$

$$\min_{\phi} D_{KL}(q_{\phi}(H)||p(H|E))$$

$$\min_{\phi} \int q_{\phi}(H) \log \frac{q_{\phi}(H)}{p(H|E)} dH = \mathbb{E}_{q_{\phi}(H)}[\log q_{\phi}(H) - \log p(H|E)]$$

$$\min_{\phi} \mathbb{E}_{q_{\phi}(H)}[\log q_{\phi}(H) - \log p(H) - \log p(E|H) + \log p(E)]$$

Derivation of the ELBO

$$\frac{p(E|H)p(H)}{P(E)}$$

$$\min_{\phi} D_{KL}(q_\phi(H) \| p(H|E))$$

$$\min_{\phi} \int q_\phi(H) \log \frac{q_\phi(H)}{p(H|E)} dH = \mathbb{E}_{q_\phi(H)}\left[\log q_\phi(H) - \log p(H|E)\right]$$

$$\min_{\phi} \mathbb{E}_{q_\phi(H)}\left[\log q_\phi(H) - \log p(H) - \log p(E|H) + \log p(E)\right]$$

Derivation of the ELBO

$$\frac{p(E|H)p(H)}{P(E)}$$

$$\min_{\phi} D_{KL}(q_{\phi}(H) \| p(H|E))$$

$$\min_{\phi} \int q_{\phi}(H) \log \frac{q_{\phi}(H)}{p(H|E)} dH = \mathbb{E}_{q_{\phi}(H)}[\log q_{\phi}(H) - \log p(H|E)]$$

$$\min_{\phi} \mathbb{E}_{q_{\phi}(H)}[\log q_{\phi}(H) - \log p(H) - \log p(E|H) + \log p(E)]$$

Derivation of the ELBO

$$\frac{p(E|H)p(H)}{P(E)}$$

$$\min_{\phi} D_{KL}(q_{\phi}(H)||p(H|E))$$

$$\min_{\phi} \int q_{\phi}(H) \log \frac{q_{\phi}(H)}{p(H|E)} dH = \mathbb{E}_{q_{\phi}(H)}[\log q_{\phi}(H) - \log p(H|E)]$$

$$\min_{\phi} \mathbb{E}_{q_{\phi}(H)}[\log q_{\phi}(H) - \log p(H) - \log p(E|H) + \log p(E)]$$

Derivation of the ELBO

$$\frac{p(E|H)p(H)}{P(E)}$$

$$\min_{\phi} D_{KL}(q_{\phi}(H)||p(H|E))$$

$$\min_{\phi} \int q_{\phi}(H) \log \frac{q_{\phi}(H)}{p(H|E)} dH = \mathbb{E}_{q_{\phi}(H)}[\log q_{\phi}(H) - \log p(H|E)]$$

$$\min_{\phi} \mathbb{E}_{q_{\phi}(H)}[\underbrace{\log q_{\phi}(H) - \log p(H)}_{D_{KL}(q_{\phi}(H)||p(H))} - \log p(E|H) + \log p(E)]$$

Derivation of the ELBO

$$\frac{p(E|H)p(H)}{P(E)}$$

$$\min_{\phi} D_{KL}(q_{\phi}(H)\|p(H|E))$$

$$\min_{\phi} \int q_{\phi}(H) \log \frac{q_{\phi}(H)}{p(H|E)} dH = \mathbb{E}_{q_{\phi}(H)}\left[\log q_{\phi}(H) - \log p(H|E)\right]$$

$$\min_{\phi} \mathbb{E}_{q_{\phi}(H)}\left[\underbrace{\log q_{\phi}(H) - \log p(H)}_{D_{KL}(q_{\phi}(H)\|p(H))} - \log p(E|H) + \log p(E)\right]$$

Derivation of the ELBO

$$\frac{p(E|H)p(H)}{P(E)}$$

$$\min_{\phi} D_{KL}(q_{\phi}(H)||p(H|E))$$

$$\min_{\phi} \int q_{\phi}(H) \log \frac{q_{\phi}(H)}{p(H|E)} dH = \mathbb{E}_{q_{\phi}(H)}[\log q_{\phi}(H) - \log p(H|E)]$$

$$\min_{\phi} \mathbb{E}_{q_{\phi}(H)}[\underbrace{\log q_{\phi}(H) - \log p(H)}_{D_{KL}(q_{\phi}(H)||p(H))} - \log p(E|H) + \log p(E)]$$

$$\min_{\phi} D_{KL}(q_{\phi}(H)||p(H)) - \mathbb{E}_{q_{\phi}(H)}[\log p(E|H)] + \text{const.}$$

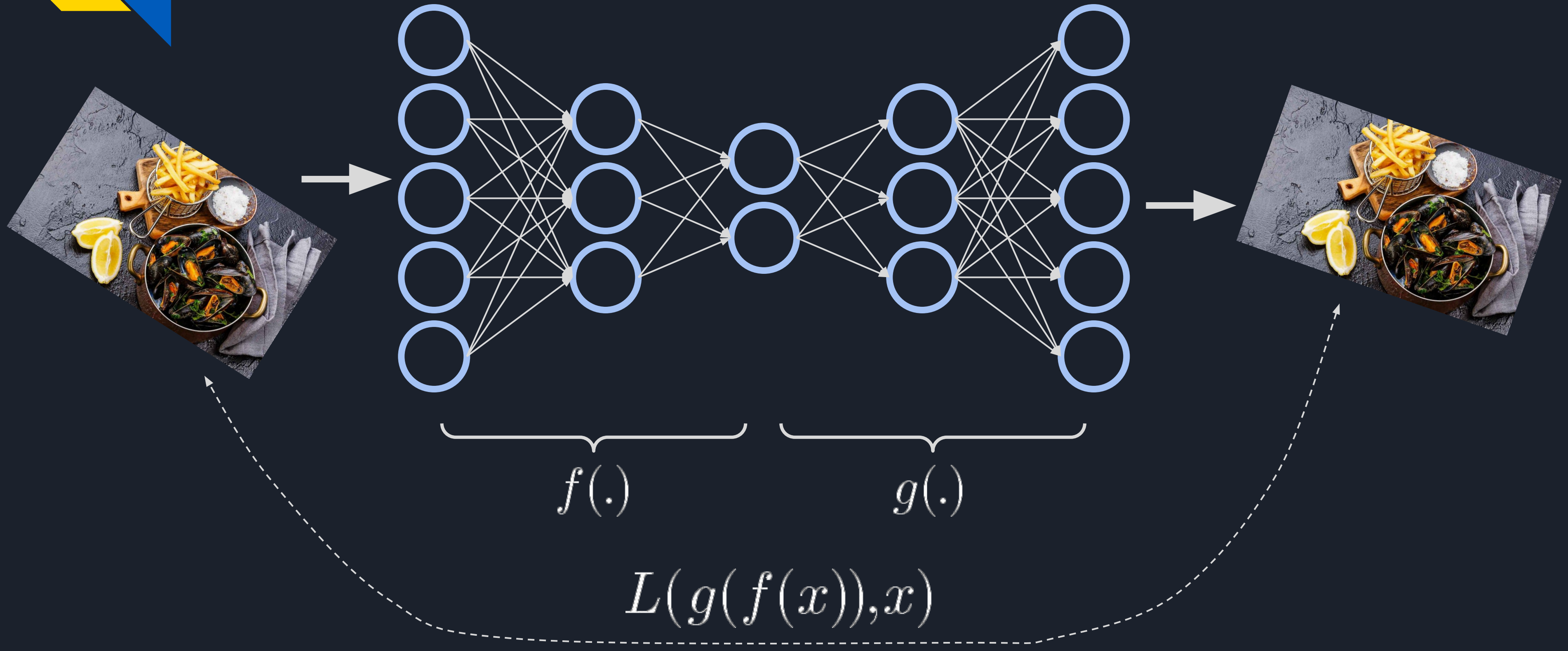
Variational Autoencoders

Diederik P. Kingma and Max Welling. "Auto-encoding variational bayes”

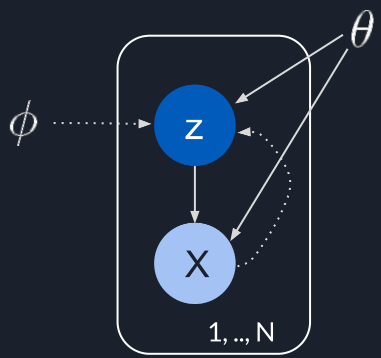
Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. “Stochastic backpropagation and variational inference in deep latent gaussian models”

KU LEUVEN

(Classical) Autoencoders



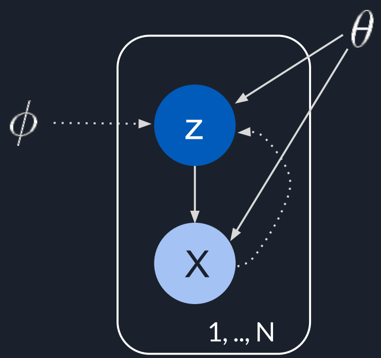
Variational Autoencoder



$p_\theta(z)$ prior e.g.: $\mathcal{N}(0,\boldsymbol{I})$

$p_\theta(x|z)$ likelihood e.g: $\mathcal{N}(g_\mu(z),g_\sigma(z))$

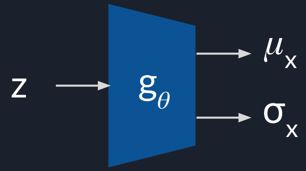
Variational Autoencoder



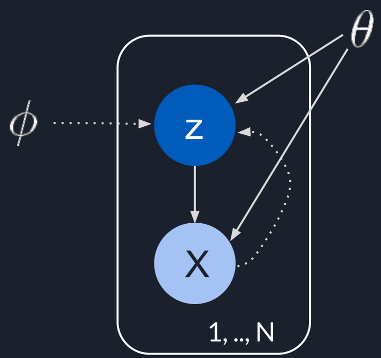
$p_\theta(z)$ prior e.g.: $\mathcal{N}(0, \boldsymbol{I})$

$p_\theta(x|z)$ likelihood e.g: $\mathcal{N}(g_\mu(z), g_\sigma(z))$

In real life applications often the mean is regarded as output



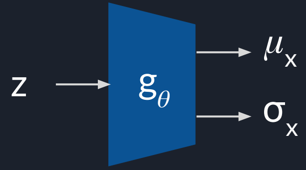
Variational Autoencoder

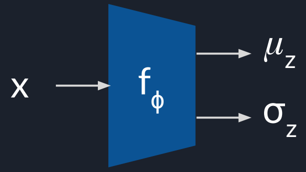


$p_\theta(z)$ prior e.g.: $\mathcal{N}(0, \boldsymbol{I})$

$p_\theta(x|z)$ likelihood e.g: $\mathcal{N}(g_\mu(z), g_\sigma(z))$

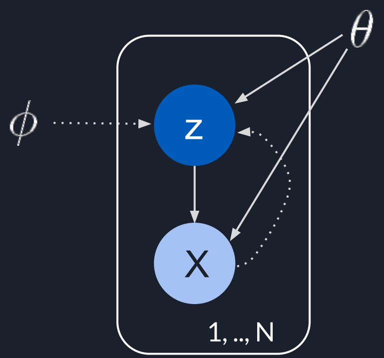
In real life applications often the mean is regarded as output





Search for a variational distribution in the form $\mathcal{N}(f_\mu(x), f_\sigma(x))$
Note: an entire family of variational distributions is learned!

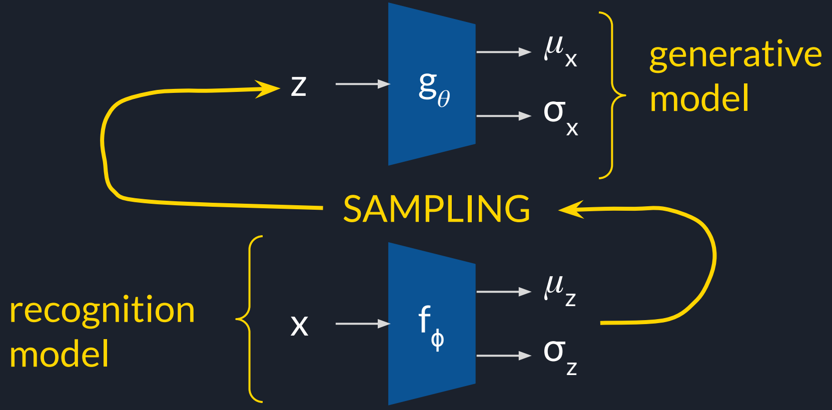
Variational Autoencoder



$p_\theta(z)$ prior e.g.: $\mathcal{N}(0, \boldsymbol{I})$

$p_\theta(x|z)$ likelihood e.g: $\mathcal{N}(g_\mu(z), g_\sigma(z))$

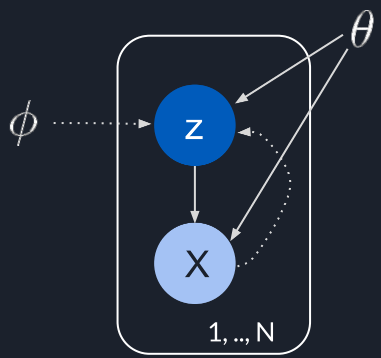
In real life applications often the mean is regarded as output



Search for a variational distribution in the form $\mathcal{N}(f_\mu(x), f_\sigma(x))$

Note: an entire family of variational distributions is learned!

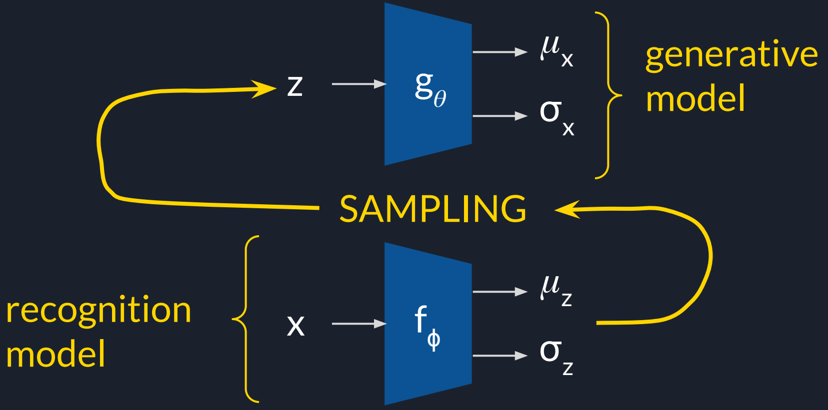
Variational Autoencoder



$p_\theta(z)$ prior e.g.: $\mathcal{N}(0, \boldsymbol{I})$

$p_\theta(x|z)$ likelihood e.g: $\mathcal{N}(g_\mu(z), g_\sigma(z))$

In real life applications often the mean is regarded as output



Search for a variational distribution in the form $\mathcal{N}(f_\mu(x), f_\sigma(x))$

Note: an entire family of variational distributions is learned!

Learning θ and ϕ jointly.

Reparameterization trick

$$\min_{\phi} D_{KL}(q_{\phi}(H) \| p(H)) - \mathbb{E}_{q_{\phi}(H)}\big[\log p(E|H)\big] + \text{const.}$$

We need to evaluate gradients of the following form: $\nabla_{\phi}\, \mathbb{E}_{q_{\phi}(H)}\big[f(H)\big]$

Reparameterization trick

$$\min_{\phi} D_{KL}(q_\phi(H) \| p(H)) - \mathbb{E}_{q_\phi(H)}[\log p(E|H)] + \text{const.}$$

We need to evaluate gradients of the following form: $\nabla_\phi \mathbb{E}_{q_\phi(H)}[f(H)]$

Let us assume the following form: $q_\phi(H) = \mathcal{N}(\mu(X), \sigma(X))$

Reparameterization trick

$$\min_{\phi} D_{KL}(q_{\phi}(H) \| p(H)) - \mathbb{E}_{q_{\phi}(H)}[\log p(E|H)] + \text{const.}$$

We need to evaluate gradients of the following form: $\nabla_{\phi} \mathbb{E}_{q_{\phi}(H)}[f(H)]$

Let us assume the following form: $q_{\phi}(H) = \mathcal{N}(\mu(X), \sigma(X))$

Samples from the distribution can be generated as:

$$\mu(X) + \epsilon\sigma(X);\ \epsilon \sim \mathcal{N}(0, I)$$

Reparameterization trick

$$\min_{\phi} D_{KL}(q_{\phi}(H) \| p(H)) - \mathbb{E}_{q_{\phi}(H)}[\log p(E|H)] + \text{const.}$$

We need to evaluate gradients of the following form: $\nabla_{\phi} \mathbb{E}_{q_{\phi}(H)}[f(H)]$

Let us assume the following form: $q_{\phi}(H) = \mathcal{N}(\mu(X), \sigma(X))$

Samples from the distribution can be generated as:

$$\mu(X) + \epsilon\sigma(X);\ \epsilon \sim \mathcal{N}(0, I)$$

Substituting to the gradient expression, we get the reparameterized form:

$$\nabla_{\phi} \mathbb{E}_{\mathcal{N}(\epsilon|0,I)}[f(\mu(X) + \epsilon\sigma(X))]$$

Reparameterization trick

$$\min_{\phi} D_{KL}(q_\phi(H)||p(H)) - \mathbb{E}_{q_\phi(H)}[\log p(E|H)] + \text{const.}$$

We need to evaluate gradients of the following form: $\nabla_\phi \, \mathbb{E}_{q_\phi(H)}[f(H)]$

Let us assume the following form: $q_\phi(H) = \mathcal{N}(\mu(X), \sigma(X))$

Samples from the distribution can be generated as:

$$\mu(X) + \epsilon\sigma(X);\ \epsilon \sim \mathcal{N}(0, I)$$

Substituting to the gradient expression, we get the reparameterized form:

$$\nabla_\phi \mathbb{E}_{\underbrace{\mathcal{N}(\epsilon|0,I)}_{\text{parameter free distribution}}}[f(\underbrace{\mu(X) + \epsilon\sigma(X)}_{\text{function parameterized by }\phi})]$$

parameter free distribution

function parameterized by ϕ

Reparameterization trick

$$\min_{\phi} D_{KL}(q_{\phi}(H)||p(H)) - \mathbb{E}_{q_{\phi}(H)}[\log p(E|H)] + \text{const.}$$

We need to evaluate gradients of the following form: $\nabla_{\phi} \mathbb{E}_{q_{\phi}(H)}[f(H)]$

Let us assume the following form: $q_{\phi}(H) = \mathcal{N}(\mu(X), \sigma(X))$

Samples from the distribution can be generated as:

$$\mu(X) + \epsilon\sigma(X);\ \epsilon \sim \mathcal{N}(0, I)$$

Substituting to the gradient expression, we get the reparameterized form:

$$\nabla_{\phi} \mathbb{E}_{\underbrace{\mathcal{N}(\epsilon|0,I)}}[f(\underbrace{\mu(X)+\epsilon\sigma(X)})] \approx \frac{1}{L}\sum_{l=1}^{L} \dots$$

parameter free distribution

function parameterized by ϕ

Approximated by a finite sum
Kingma and Welling used L=1 ;)

Reparameterization trick

$I)] + \text{const.}$

We need to

Let us assu

Samples fr

Substitutin

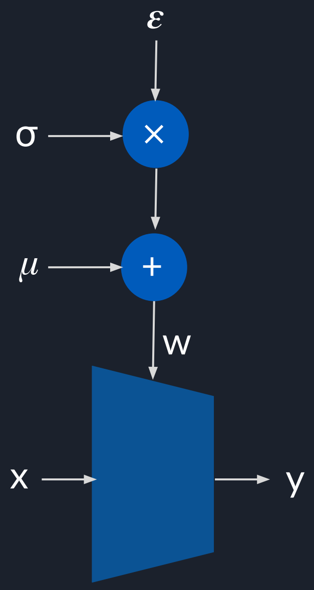
Target	Basis	Differentiable transformation
$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{N}(0,1)$	$\mu + \epsilon\sigma$
$\mathcal{N}(\mu, RR^\top)$	$\mathcal{N}(0,I)$	$\mu + R\epsilon$
any with tractable inverse CDF	$\mathcal{U}(0,1)$	the inverse CDF
$\mathcal{Cat}(\pi_1, \ldots \pi_k)$	$\mathcal{Gumbel}(0,1)$	$\operatorname{argmax}_i \epsilon_i + \log \pi_i$
$\ldots$	$\ldots$	$\ldots$

parameter free distribution

function parameterized by ϕ

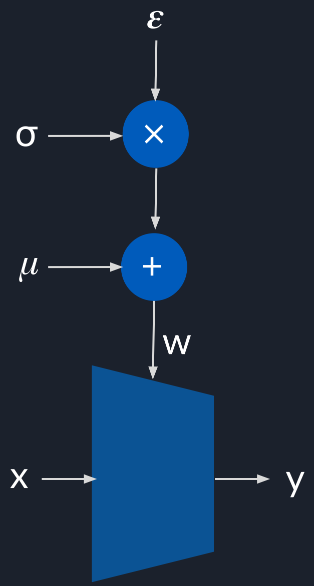
Approximated by a finite sum
Kingma and Welling used L=1 ;)

Bayesian treatment of NN weights

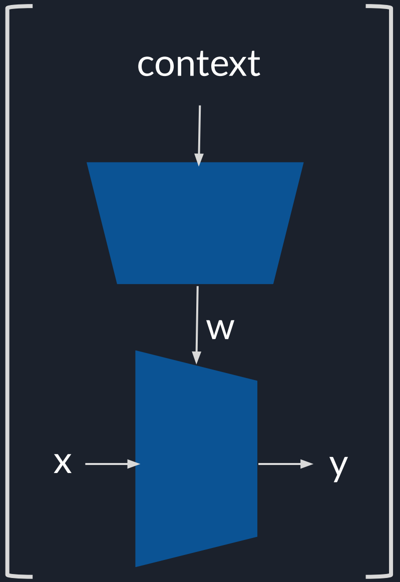


reparameterization trick

Bayesian treatment of NN weights

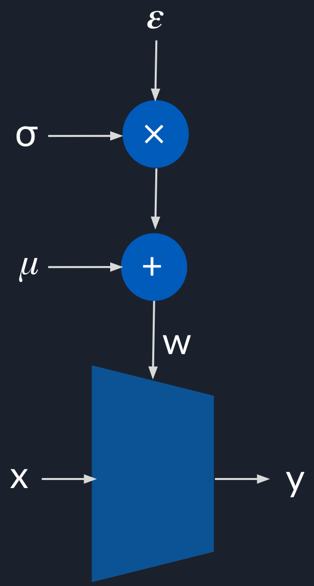


reparameterization trick

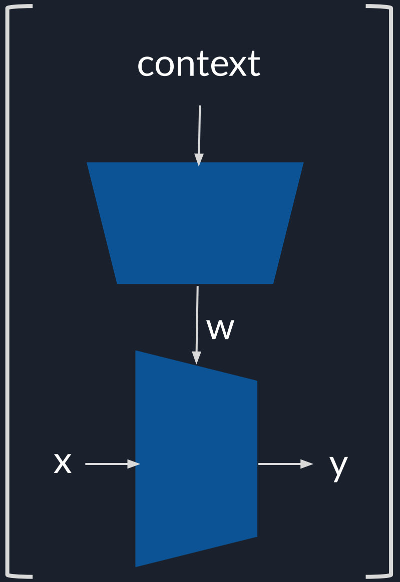


meta networks

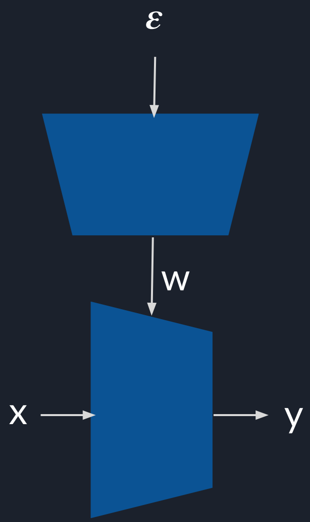
Bayesian treatment of NN weights



reparameterization trick

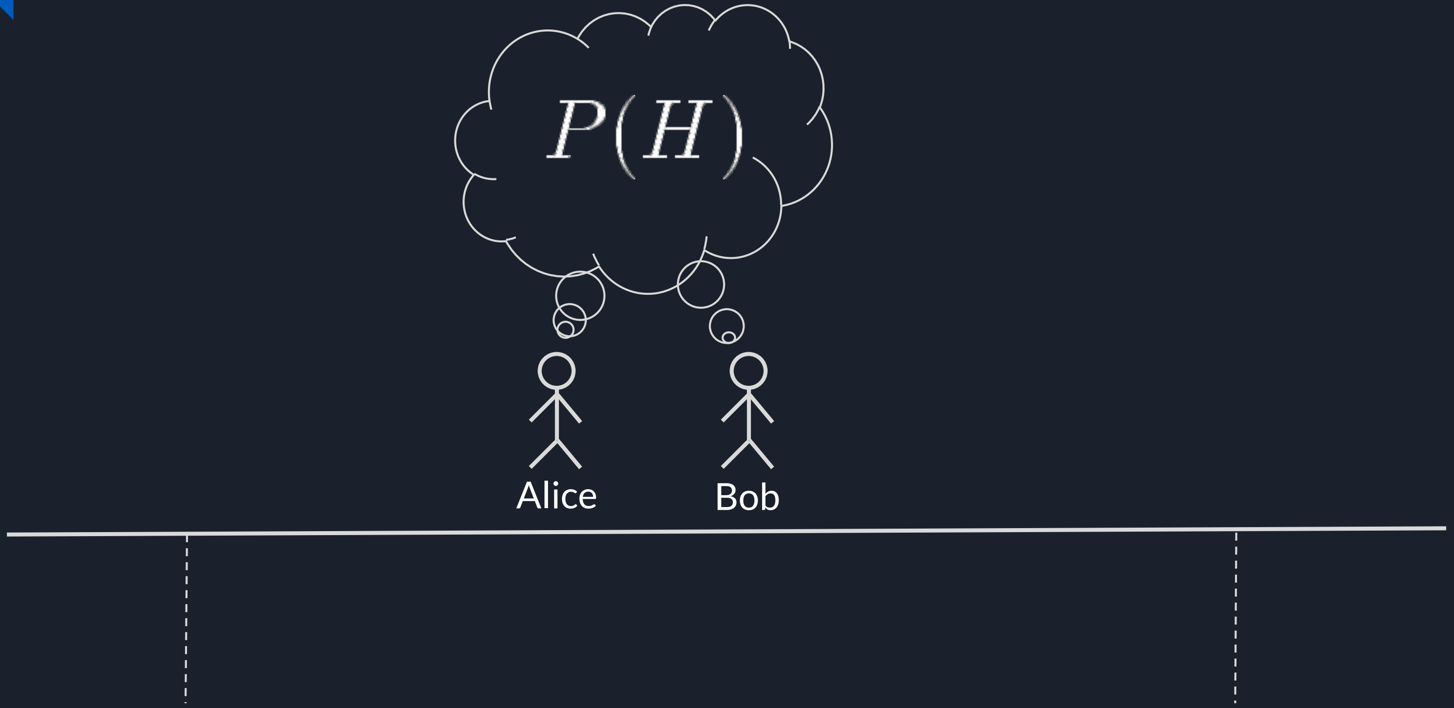


meta networks

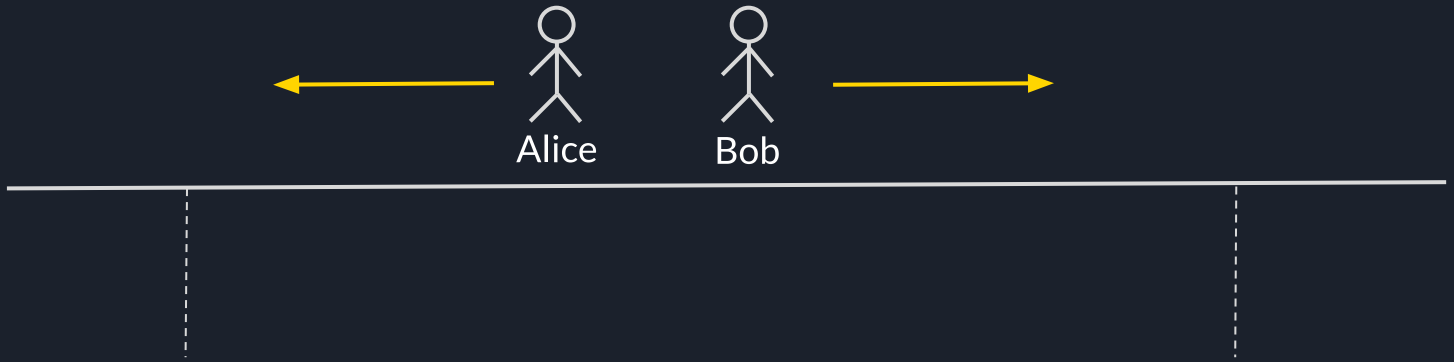


Bayesian meta networks

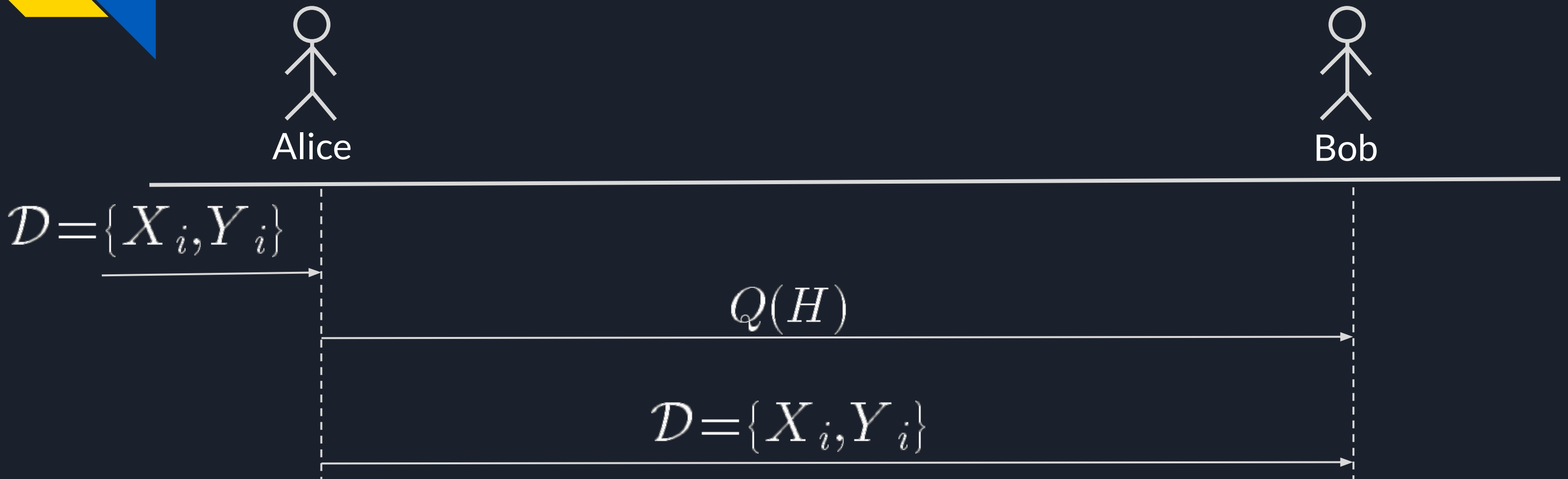
Information theoretical interpretation



Information theoretical interpretation



Information theoretical interpretation



Can this be done in practice?

Yes, see: Geoffrey E. Hinton and Richard S. Zemel. Autoencoders, minimum description length, and helmholtz free energy. In Advances in Neural Information Processing Systems 6, 1994.

Thank you for your attention!