

Variational Inference

CS 185/285

Instructor: Sergey Levine
UC Berkeley

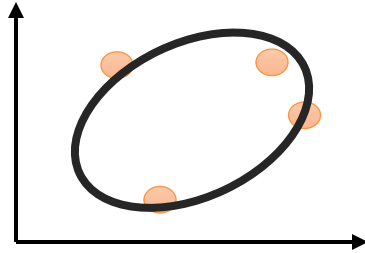


Part 1:

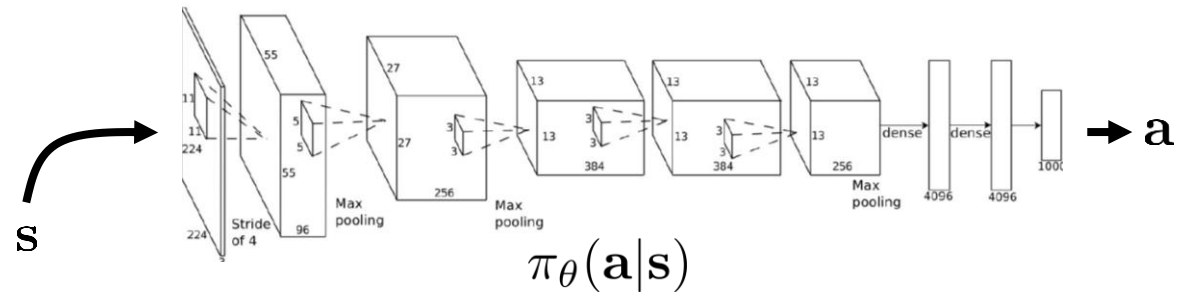
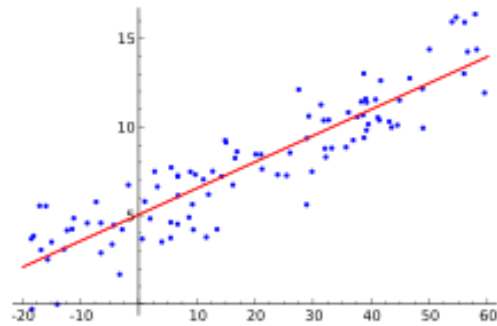
Latent variable models

Probabilistic models

$p(x)$



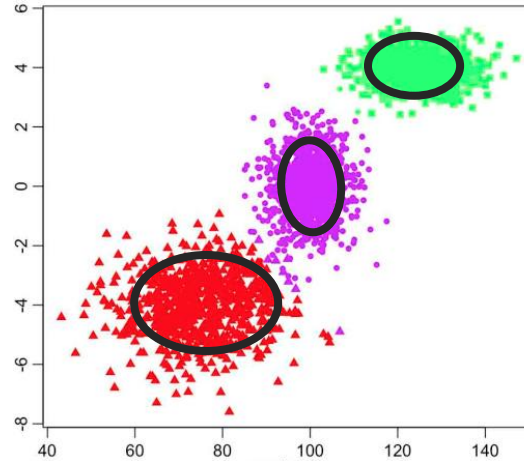
$p(y|x)$



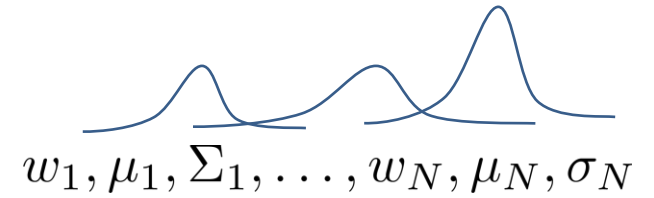
Latent variable models

$$p(x) = \sum_z p(x|z)p(z)$$

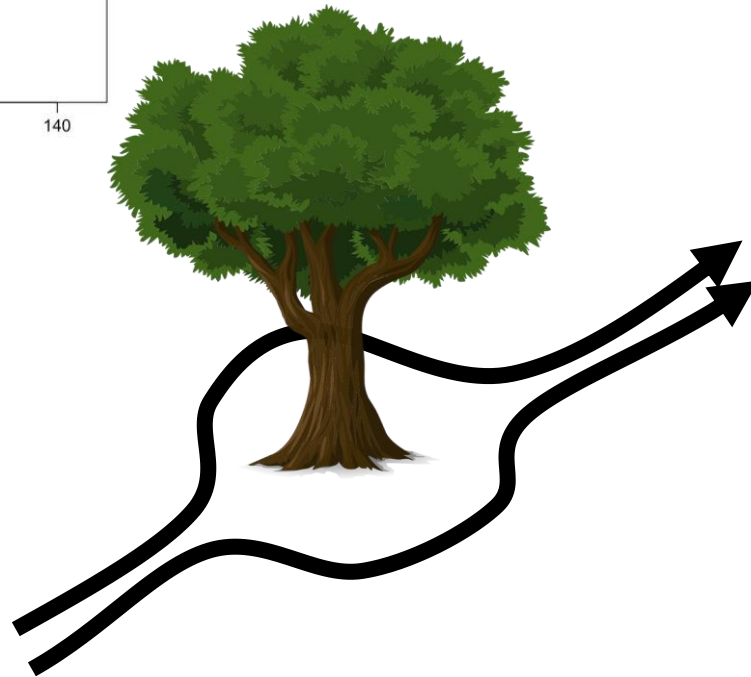
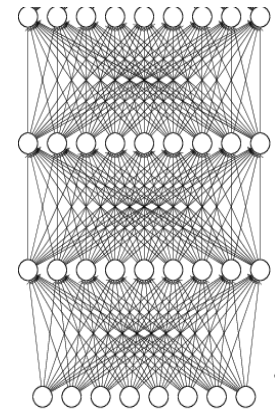
↑
mixture
element



$$p(y|x) = \sum_z p(y|x, z)p(z)$$



$w_1, \mu_1, \Sigma_1, \dots, w_N, \mu_N, \sigma_N$

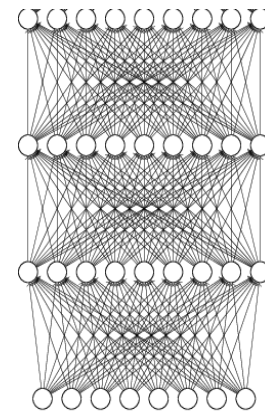
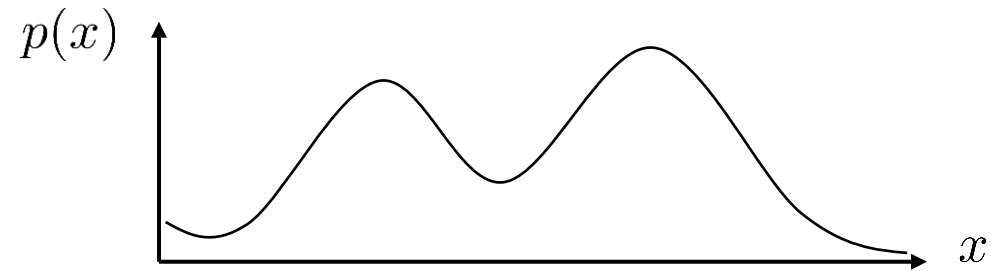


Latent variable models in general

$$p(x) = \int p(x|z)p(z)dz$$

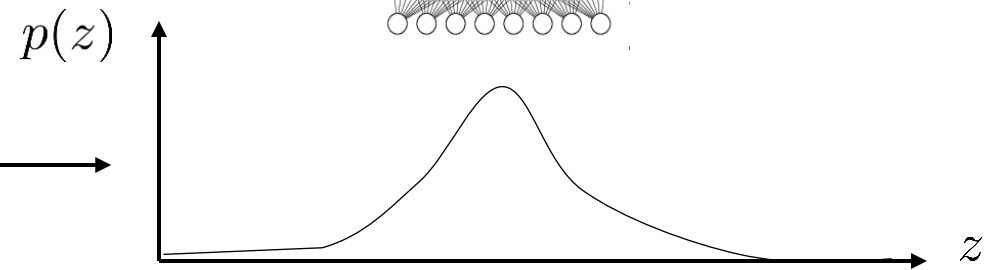
“easy” distribution
(e.g., conditional Gaussian)

“easy” distribution
(e.g., Gaussian)



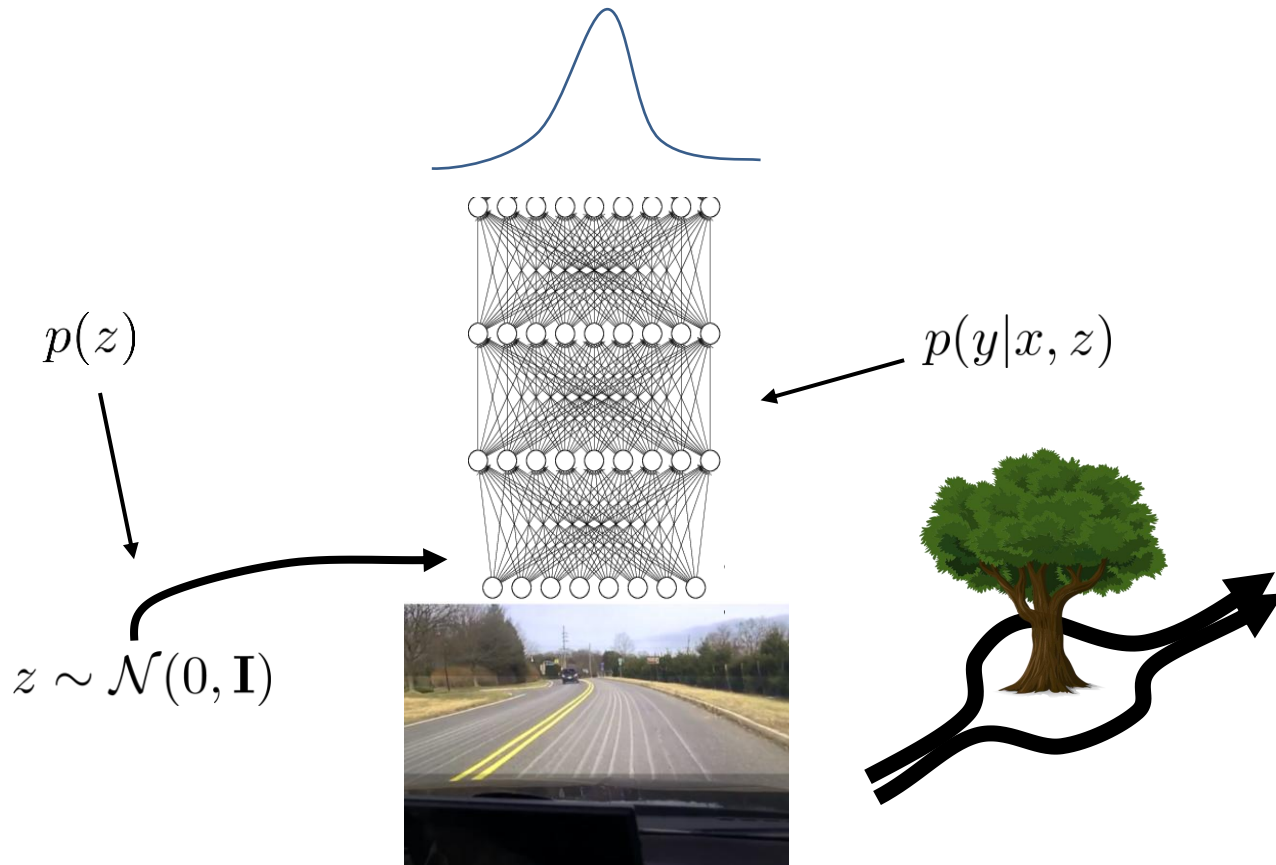
$$p(x|z) = \mathcal{N}(\mu_{\text{nn}}(z), \sigma_{\text{nn}}(z))$$

“easy” distribution
(e.g., Gaussian)

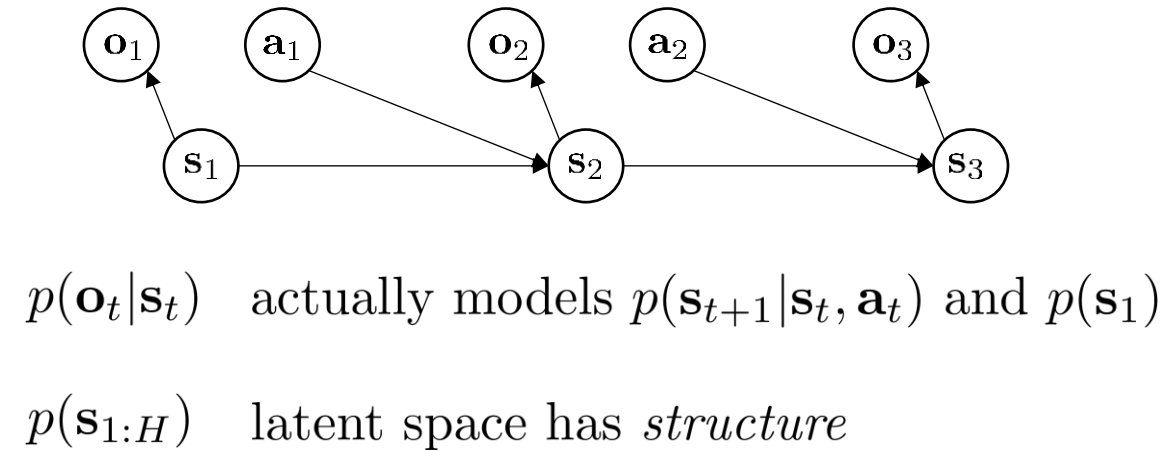


Latent variable models in RL

conditional latent variable models for multi-modal policies



latent variable models for model-based RL



How do we train latent variable models?

the model: $p_{\theta}(x)$

the data: $\mathcal{D} = \{x_1, x_2, x_3, \dots, x_N\}$

maximum likelihood fit:

$$\theta \leftarrow \arg \max_{\theta} \frac{1}{N} \sum_i \log p_{\theta}(x_i)$$

$$p(x) = \int p(x|z)p(z)dz$$

$$\theta \leftarrow \arg \max_{\theta} \frac{1}{N} \sum_i \log \left(\int p_{\theta}(x_i|z)p(z)dz \right)$$



completely intractable

Estimating the log-likelihood

alternative: *expected* log-likelihood:

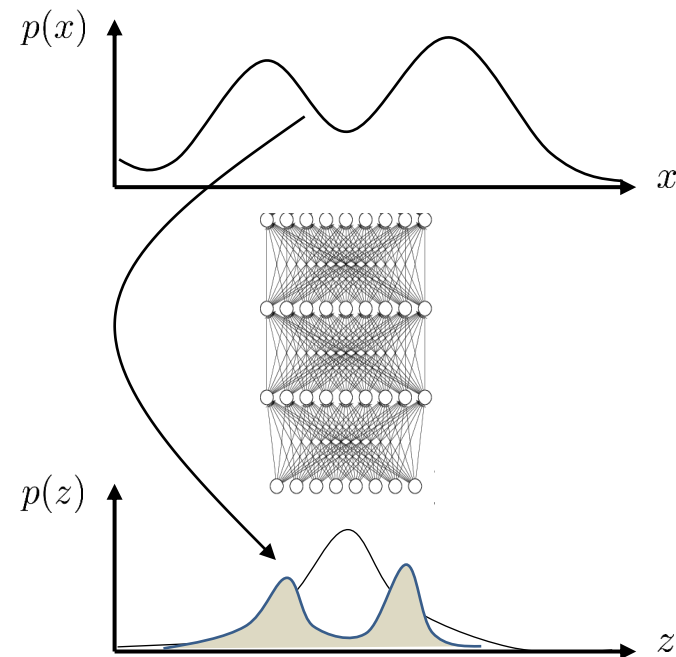
$$\theta \leftarrow \arg \max_{\theta} \frac{1}{N} \sum_i E_{z \sim p(z|x_i)} [\log p_{\theta}(x_i, z)]$$

but... how do we calculate $p(z|x_i)$?

this is called *probabilistic inference*

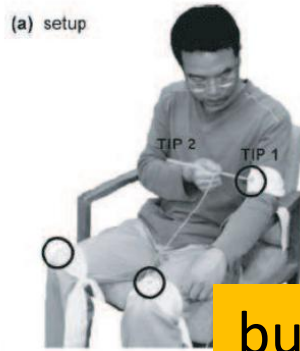
intuition: “guess” most likely z given x_i ,
and pretend it’s the right one

...but there are many possible values of z
so use the distribution $p(z|x_i)$



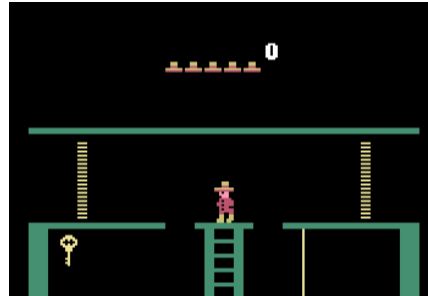
Other places we'll see probabilistic inference

using RL/control + variational inference to model human behavior

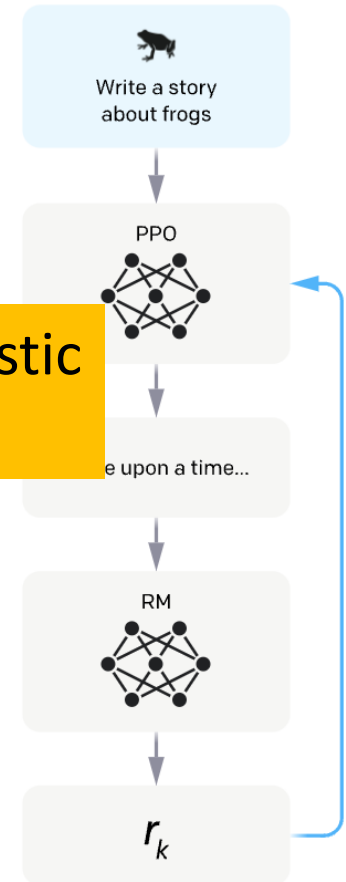


Li & Todorov

using generative models and variational inference for exploration



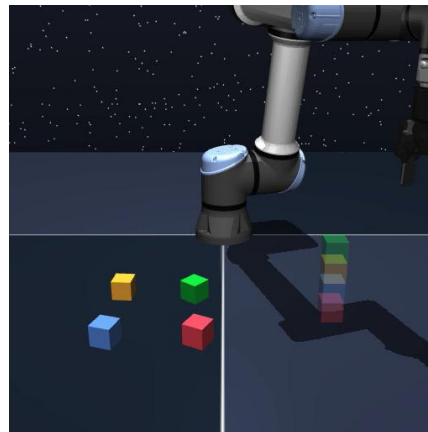
learning from human feedback



but today we'll focus on the fundamentals of probabilistic (variational) inference and latent variable models



Ziebart '08



Part 2: Variational inference

The variational approximation

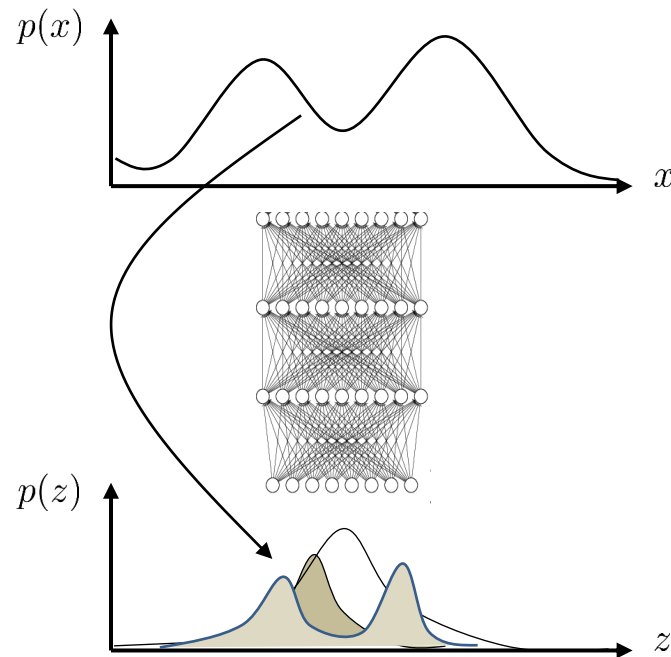
but... how do we calculate $p(z|x_i)$?

what if we approximate with $q_i(z) = \mathcal{N}(\mu_i, \sigma_i)$

can bound $\log p(x_i)$!

$$\begin{aligned}\log p(x_i) &= \log \int_z p(x_i|z)p(z) \\ &= \log \int_z p(x_i|z)p(z) \frac{q_i(z)}{q_i(z)} \\ &= \log E_{z \sim q_i(z)} \left[\frac{p(x_i|z)p(z)}{q_i(z)} \right]\end{aligned}$$

this is *incorrect* but *very convenient*



The variational approximation

but... how do we calculate $p(z|x_i)$?

can bound $\log p(x_i)$!

$$\log p(x_i) = \log \int_z p(x_i|z)p(z)$$

$$= \log \int_z p(x_i|z)p(z) \frac{q_i(z)}{q_i(z)}$$

$$= \log E_{z \sim q_i(z)} \left[\frac{p(x_i|z)p(z)}{q_i(z)} \right]$$

$$\geq E_{z \sim q_i(z)} \left[\log \frac{p(x_i|z)p(z)}{q_i(z)} \right] = E_{z \sim q_i(z)} [\log p(x_i|z) + \log p(z)] + \mathbf{H}_{q_i(z)}$$

maximizing this maximizes $\log p(x_i)$



Jensen's inequality

$$\log E[y] \geq E[\log y]$$

A brief aside...

Entropy:

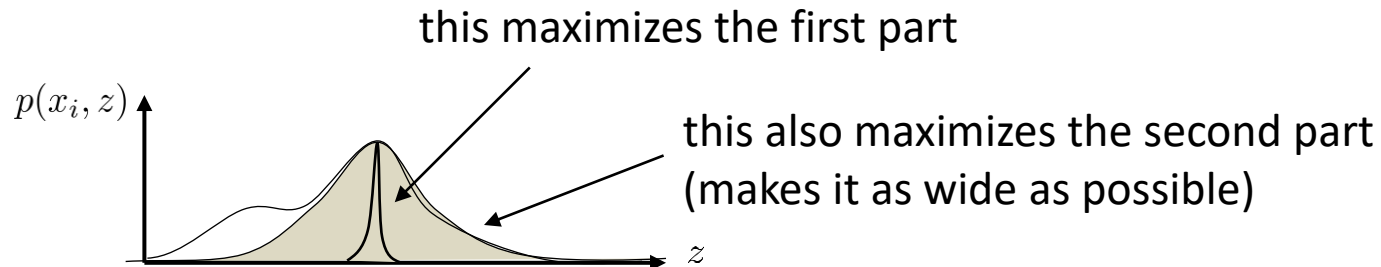
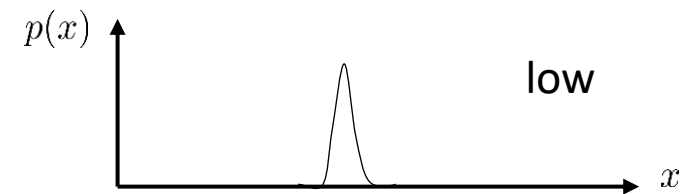
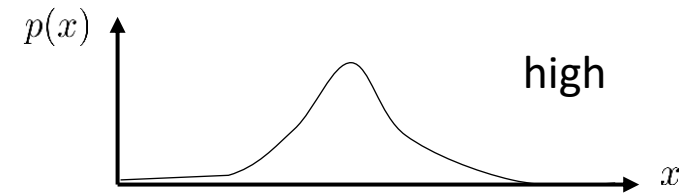
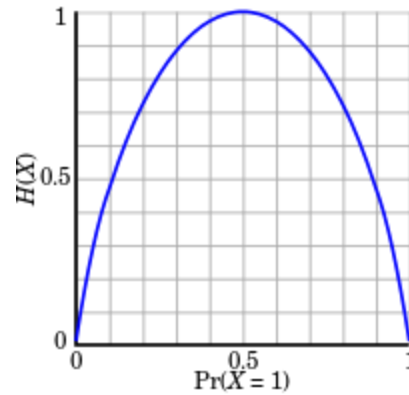
$$\mathcal{H}(p) = -E_{x \sim p(x)}[\log p(x)] = - \int_x p(x) \log p(x) dx$$

Intuition 1: how *random* is the random variable?

Intuition 2: how large is the log probability in expectation *under itself*

what do we expect this to do?

$$E_{z \sim q_i(z)}[\log p(x_i|z) + \log p(z)] + \mathcal{H}(q_i)$$



A brief aside...

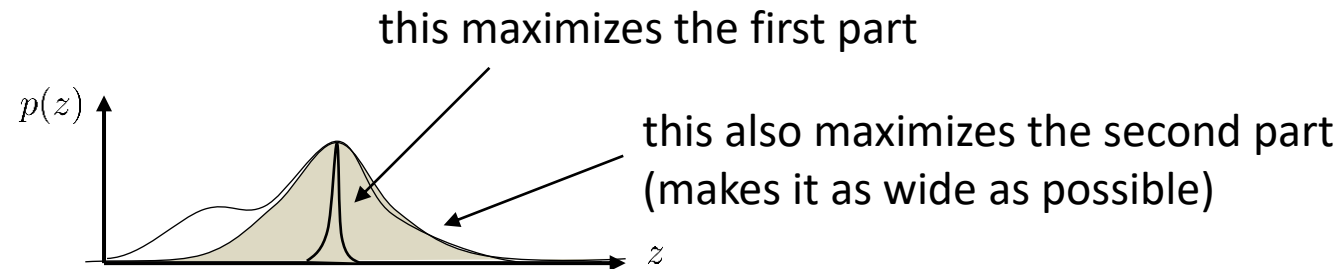
KL-Divergence:

$$D_{\text{KL}}(q||p) = E_{x \sim q(x)} \left[\log \frac{q(x)}{p(x)} \right] = E_{x \sim q(x)} [\log q(x)] - E_{x \sim q(x)} [\log p(x)] = -E_{x \sim q(x)} [\log p(x)] - \mathcal{H}(q)$$

Intuition 1: how *different* are two distributions?

Intuition 2: how small is the expected log probability of one distribution under another, minus entropy?

why entropy?



The variational approximation

$$\mathcal{L}_i(p, q_i)$$

$$\log p(x_i) \geq \overbrace{E_{z \sim q_i(z)} [\log p(x_i|z) + \log p(z)]} + \mathcal{H}(q_i)$$

what makes a good $q_i(z)$?

intuition: $q_i(z)$ should approximate $p(z|x_i)$

approximate in what sense?

compare in terms of KL-divergence: $D_{\text{KL}}(q_i(z) \| p(z|x))$

why?

$$\begin{aligned} D_{\text{KL}}(q_i(z) \| p(z|x_i)) &= E_{z \sim q_i(z)} \left[\log \frac{q_i(z)}{p(z|x_i)} \right] = E_{z \sim q_i(z)} \left[\log \frac{q_i(z)p(x_i)}{p(x_i, z)} \right] \\ &= -E_{z \sim q_i(z)} [\log p(x_i|z) + \log p(z)] + E_{z \sim q_i(z)} [\log q_i(z)] + E_{z \sim q_i(z)} [\log p(x_i)] \\ &= -E_{z \sim q_i(z)} [\log p(x_i|z) + \log p(z)] - \mathcal{H}(q_i) + \log p(x_i) \\ &= -\mathcal{L}_i(p, q_i) + \log p(x_i) \end{aligned}$$

$$\log p(x_i) = D_{\text{KL}}(q_i(z) \| p(z|x_i)) + \mathcal{L}_i(p, q_i)$$

$$\log p(x_i) \geq \mathcal{L}_i(p, q_i)$$

The variational approximation

$$\mathcal{L}_i(p, q_i)$$

$$\log p(x_i) \geq \overbrace{E_{z \sim q_i(z)} [\log p(x_i|z) + \log p(z)]}^{\mathcal{L}_i(p, q_i)} + \mathcal{H}(q_i)$$

$$\log p(x_i) = D_{\text{KL}}(q_i(z) \| p(z|x_i)) + \mathcal{L}_i(p, q_i)$$

$$\log p(x_i) \geq \mathcal{L}_i(p, q_i)$$

$$D_{\text{KL}}(q_i(z) \| p(z|x_i)) = E_{z \sim q_i(z)} \left[\log \frac{q_i(z)}{p(z|x_i)} \right] = E_{z \sim q_i(z)} \left[\log \frac{q_i(z)p(x_i)}{p(x_i, z)} \right]$$

$$= \underbrace{-E_{z \sim q_i(z)} [\log p(x_i|z) + \log p(z)]}_{-\mathcal{L}_i(p, q_i)} + \log p(x_i)$$

independent of q_i !

\Rightarrow maximizing $\mathcal{L}_i(p, q_i)$ w.r.t. q_i minimizes KL-divergence!

How do we use this?

$$\log p(x_i) \geq \overbrace{E_{z \sim q_i(z)} [\log p_\theta(x_i|z) + \log p(z)]}^{\mathcal{L}_i(p, q_i)} + \mathcal{H}(q_i)$$

~~$$\theta \leftarrow \arg \max_{\theta} \frac{1}{N} \sum_i \log p_\theta(x_i)$$~~

$$\theta \leftarrow \arg \max_{\theta} \frac{1}{N} \sum_i \mathcal{L}_i(p, q_i)$$

for each x_i (or mini-batch):

calculate $\nabla_{\theta} \mathcal{L}_i(p, q_i)$:

sample $z \sim q_i(z)$

$$\nabla_{\theta} \mathcal{L}_i(p, q_i) \approx \nabla_{\theta} \log p_\theta(x_i|z)$$

$$\theta \leftarrow \theta + \alpha \nabla_{\theta} \mathcal{L}_i(p, q_i)$$

update q_i to maximize $\mathcal{L}_i(p, q_i)$

how?

let's say $q_i(z) = \mathcal{N}(\mu_i, \sigma_i)$

use gradient $\nabla_{\mu_i} \mathcal{L}_i(p, q_i)$ and $\nabla_{\sigma_i} \mathcal{L}_i(p, q_i)$

gradient ascent on μ_i, σ_i

What's the problem?

for each x_i (or mini-batch):

calculate $\nabla_{\theta} \mathcal{L}_i(p, q_i)$:

sample $z \sim q_i(z)$

$\nabla_{\theta} \mathcal{L}_i(p, q_i) \approx \nabla_{\theta} \log p_{\theta}(x_i|z)$

$\theta \leftarrow \theta + \alpha \nabla_{\theta} \mathcal{L}_i(p, q_i)$

update q_i to maximize $\mathcal{L}_i(p, q_i)$

let's say $q_i(z) = \mathcal{N}(\mu_i, \sigma_i)$

use gradient $\nabla_{\mu_i} \mathcal{L}_i(p, q_i)$ and $\nabla_{\sigma_i} \mathcal{L}_i(p, q_i)$

gradient ascent on μ_i, σ_i

How many parameters are there?

$$|\theta| + (|\mu_i| + |\sigma_i|) \times N$$

next time we'll learn about how deep learning makes this tractable!