Latent Variable Models

Stefano Ermon, Yang Song

Stanford University

Lecture 5

Recap of last lecture

1. Autoregressive models:
   - Chain rule based factorization is fully general
   - Compact representation via *conditional independence* and/or *neural parameterizations*
2. Autoregressive models Pros:
   - Easy to evaluate likelihoods
   - Easy to train
3. Autoregressive models Cons:
   - Requires an ordering
   - Generation is sequential
   - Cannot learn features in an unsupervised way

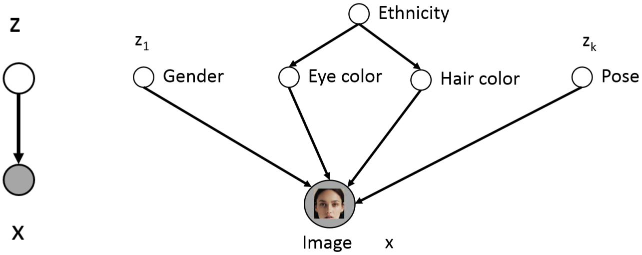
Plan for today

1. Latent Variable Models
    - Mixture models
    - Variational autoencoder
    - Variational inference and learning

Latent Variable Models: Motivation

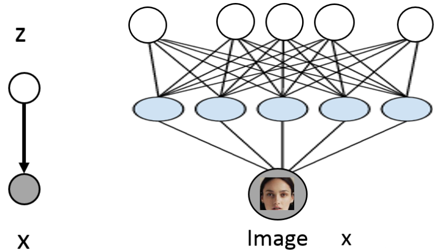
1. Lots of variability in images x due to gender, eye color, hair color, pose, etc. However, unless images are annotated, these factors of variation are not explicitly available (latent).
2. **Idea**: explicitly model these factors using latent variables z

Latent Variable Models: Motivation



1. Only shaded variables x are observed in the data (pixel values)
2. Latent variables z correspond to high level features
   - If z chosen properly, $p(\mathrm{x}|\mathrm{z})$ could be much simpler than $p(\mathrm{x})$
   - If we had trained this model, then we could identify features via $p(\mathrm{z} \mid \mathrm{x})$, e.g., $p(EyeColor = Blue|\mathrm{x})$
3. **Challenge:** Very difficult to specify these conditionals by hand

Deep Latent Variable Models

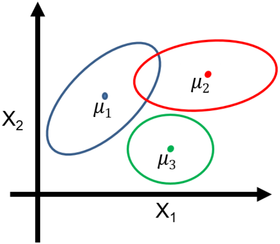


- Use neural networks to model the conditionals (deep latent variable models):
  1. $\mathrm{z} \sim \mathcal{N}(0, I)$
  2. $p(\mathrm{x} \mid \mathrm{z}) = \mathcal{N}\left(\mu_\theta(\mathrm{z}), \Sigma_\theta(\mathrm{z})\right)$ where μ_θ,Σ_θ are neural networks
- *Hope* that after training, z will correspond to meaningful latent factors of variation (*features*). Unsupervised representation learning.
- As before, features can be computed via $p(\mathrm{z} \mid \mathrm{x})$

Mixture of Gaussians: a Shallow Latent Variable Model

Mixture of Gaussians. Bayes net: $z \rightarrow x$.

1. $z \sim \text{Categorical}(1, \cdots, K)$
2. $p(x \mid z = k) = \mathcal{N}(\mu_k, \Sigma_k)$



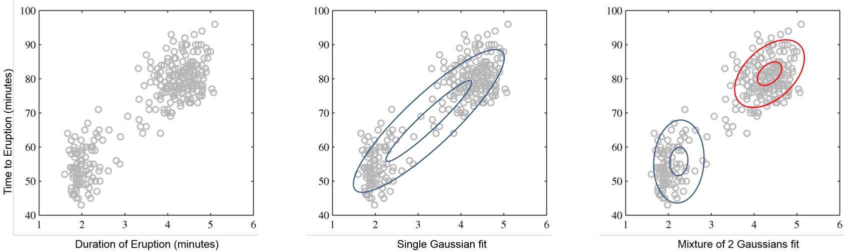
Generative process

1. Pick a mixture component k by sampling z
2. Generate a data point by sampling from that Gaussian

Mixture of Gaussians: a Shallow Latent Variable Model

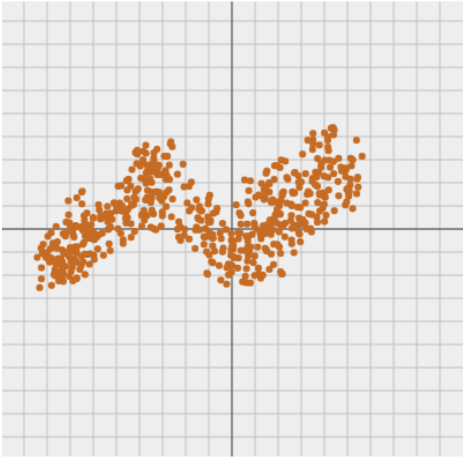
Mixture of Gaussians:

1. $z \sim \text{Categorical}(1, \cdots, K)$
2. $p(x \mid z = k) = \mathcal{N}(\mu_k, \Sigma_k)$

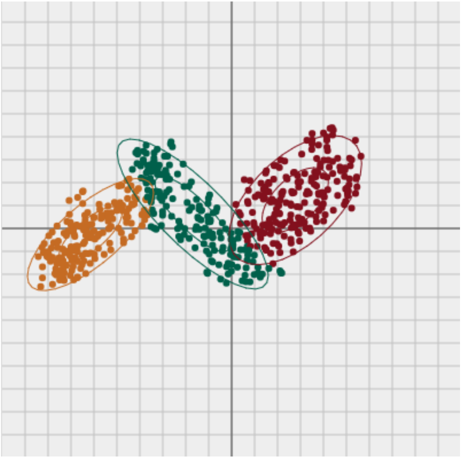


- **Clustering:** The posterior $p(z \mid x)$ identifies the mixture component
- **Unsupervised learning:** We are hoping to learn from unlabeled data (ill-posed problem)

Unsupervised learning

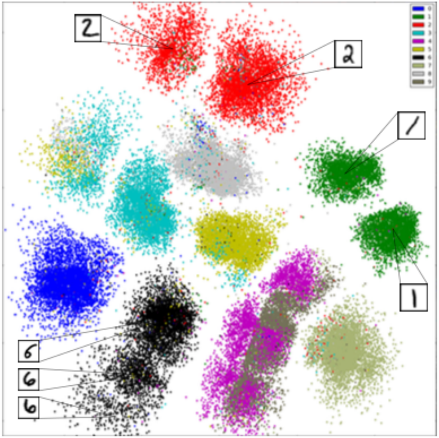


Unsupervised learning



Shown is the posterior probability that a data point was generated by the i-th mixture component, $P(z = i|x)$

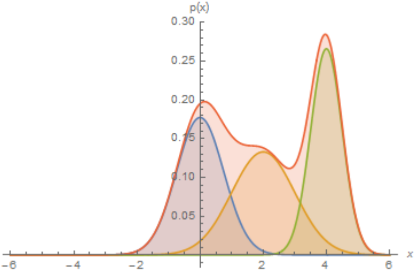
Unsupervised learning



Unsupervised clustering of handwritten digits.

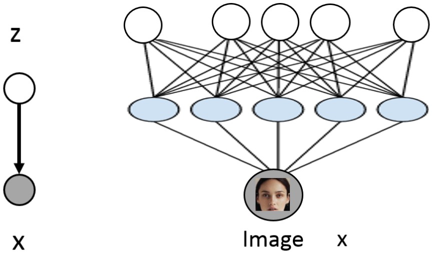
Mixture models

Alternative motivation: Combine simple models into a more complex and expressive one



$$p(\mathsf{x}) = \sum_{\mathsf{z}} p(\mathsf{x}, \mathsf{z}) = \sum_{\mathsf{z}} p(\mathsf{z}) p(\mathsf{x} \mid \mathsf{z}) = \sum_{k=1}^{K} p(\mathsf{z} = k) \underbrace{\mathcal{N}(\mathsf{x}; \mu_k, \Sigma_k)}_{\text{component}}$$

Variational Autoencoder



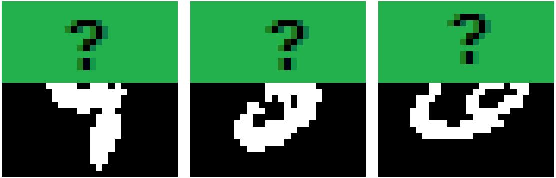
A mixture of an infinite number of Gaussians:

1. $z \sim \mathcal{N}(0, I)$
2. $p(x \mid z) = \mathcal{N}(\mu_\theta(z), \Sigma_\theta(z))$ where μ_θ,Σ_θ are neural networks
   - $\mu_\theta(z) = \sigma(Az + c) = (\sigma(a_1 z + c_1), \sigma(a_2 z + c_2)) = (\mu_1(z), \mu_2(z))$
   - $\Sigma_\theta(z) = diag(\exp(\sigma(Bz + d))) = \begin{pmatrix} \exp(\sigma(b_1 z + d_1)) & 0 \\ 0 & \exp(\sigma(b_2 z + d_2)) \end{pmatrix}$
   - $\theta = (A, B, c, d)$
3. Even though $p(x \mid z)$ is simple, the marginal $p(x)$ is very complex/flexible

Recap

- Latent Variable Models
  - Allow us to define complex models $p(\mathrm{x})$ in terms of simpler building blocks $p(\mathrm{x} \mid \mathrm{z})$
  - Natural for unsupervised learning tasks (clustering, unsupervised representation learning, etc.)
  - No free lunch: much more difficult to learn compared to fully observed, autoregressive models

Marginal Likelihood



- Suppose some pixel values are missing at train time (e.g., top half)
- Let X denote observed random variables, and Z the unobserved ones (also called hidden or latent)
- Suppose we have a model for the joint distribution (e.g., PixelCNN)

$$p(\mathrm{X}, \mathrm{Z}; \theta)$$

What is the probability $p(\mathrm{X} = \bar{\mathrm{x}}; \theta)$ of observing a training data point $\bar{\mathrm{x}}$?

$$\sum_{\mathrm{z}} p(\mathrm{X} = \bar{\mathrm{x}}, \mathrm{Z} = \mathrm{z}; \theta) = \sum_{\mathrm{z}} p(\bar{\mathrm{x}}, \mathrm{z}; \theta)$$

- Need to consider all possible ways to complete the image (fill green part)

Variational Autoencoder Marginal Likelihood

A mixture of an infinite number of Gaussians:

1. $z \sim \mathcal{N}(0, I)$
2. $p(x \mid z) = \mathcal{N}(\mu_\theta(z), \Sigma_\theta(z))$ where μ_θ,Σ_θ are neural networks
3. Z are unobserved at train time (also called hidden or latent)
4. Suppose we have a model for the joint distribution. What is the probability $p(X = \bar{x}; \theta)$ of observing a training data point $\bar{x}$?

$$\int_z p(X = \bar{x}, Z = z; \theta) dz = \int_z p(\bar{x}, z; \theta) dz$$

Partially observed data

- Suppose that our joint distribution is

$$p(\mathrm{X}, \mathrm{Z}; \theta)$$

- We have a dataset $\mathcal{D}$, where for each datapoint the X variables are observed (e.g., pixel values) and the variables Z are never observed (e.g., cluster or class id.). $\mathcal{D} = \{\mathrm{x}^{(1)}, \cdots, \mathrm{x}^{(M)}\}$.
- Maximum likelihood learning:

$$\log \prod_{\mathrm{x} \in \mathcal{D}} p(\mathrm{x}; \theta) = \sum_{\mathrm{x} \in \mathcal{D}} \log p(\mathrm{x}; \theta) = \sum_{\mathrm{x} \in \mathcal{D}} \log \sum_{\mathrm{z}} p(\mathrm{x}, \mathrm{z}; \theta)$$

- Evaluating $\log \sum_{\mathrm{z}} p(\mathrm{x}, \mathrm{z}; \theta)$ can be intractable. Suppose we have 30 binary latent features, $\mathrm{z} \in \{0, 1\}^{30}$. Evaluating $\sum_{\mathrm{z}} p(\mathrm{x}, \mathrm{z}; \theta)$ involves a sum with 2^{30} terms. For continuous variables, $\log \int_{\mathrm{z}} p(\mathrm{x}, \mathrm{z}; \theta) dz$ is often intractable. Gradients ∇_θ also hard to compute.
- Need **approximations**. One gradient evaluation per training data point $\mathrm{x} \in \mathcal{D}$, so approximation needs to be cheap.

First attempt: Naive Monte Carlo

Likelihood function $p_\theta(\mathbf{x})$ for Partially Observed Data is hard to compute:

$$p_\theta(\mathbf{x}) = \sum_{\text{All values of } \mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z}) = |\mathcal{Z}| \sum_{\mathbf{z} \in \mathcal{Z}} \frac{1}{|\mathcal{Z}|} p_\theta(\mathbf{x}, \mathbf{z}) = |\mathcal{Z}| \mathbb{E}_{\mathbf{z} \sim Uniform(\mathcal{Z})} \left[p_\theta(\mathbf{x}, \mathbf{z}) \right]$$

We can think of it as an (intractable) expectation. Monte Carlo to the rescue:

1. Sample $\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(k)}$ uniformly at random
2. Approximate expectation with sample average

$$\sum_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z}) \approx |\mathcal{Z}| \frac{1}{k} \sum_{j=1}^{k} p_\theta(\mathbf{x}, \mathbf{z}^{(j)})$$

Works in theory but not in practice. For most z, $p_\theta(\mathbf{x}, \mathrm{z})$ is very low (most completions don't make sense). Some are very large but will never "hit" likely completions by uniform random sampling. Need a clever way to select $\mathbf{z}^{(j)}$ to reduce variance of the estimator.

Second attempt: Importance Sampling

Likelihood function $p_\theta(\mathbf{x})$ for Partially Observed Data is hard to compute:

$$p_\theta(\mathbf{x}) = \sum_{\text{All possible values of } \mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z} \in \mathcal{Z}} \frac{q(\mathrm{z})}{q(\mathrm{z})} p_\theta(\mathbf{x}, \mathbf{z}) = \mathbb{E}_{\mathrm{z} \sim q(\mathrm{z})} \left[\frac{p_\theta(\mathbf{x}, \mathbf{z})}{q(\mathrm{z})} \right]$$

Monte Carlo to the rescue:

1. Sample $\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(k)}$ from $q(\mathrm{z})$
2. Approximate expectation with sample average

$$p_\theta(\mathbf{x}) \approx \frac{1}{k} \sum_{j=1}^{k} \frac{p_\theta(\mathbf{x}, \mathbf{z}^{(j)})}{q(\mathbf{z}^{(j)})}$$

What is a good choice for $q(\mathrm{z})$? Intuitively, choose likely completions. It would then be tempting to estimate the *log*-likelihood as:

$$\log\left(p_\theta(\mathbf{x})\right) \approx \log\left(\frac{1}{k} \sum_{j=1}^{k} \frac{p_\theta(\mathbf{x}, \mathbf{z}^{(j)})}{q(\mathbf{z}^{(j)})}\right) \overset{k=1}{\approx} \log\left(\frac{p_\theta(\mathbf{x}, \mathbf{z}^{(1)})}{q(\mathbf{z}^{(1)})}\right)$$

However, it's clear that $\mathbb{E}_{\mathbf{z}^{(1)} \sim q(\mathrm{z})} \left[\log\left(\frac{p_\theta(\mathbf{x}, \mathbf{z}^{(1)})}{q(\mathbf{z}^{(1)})}\right)\right] \neq \log\left(\mathbb{E}_{\mathbf{z}^{(1)} \sim q(\mathrm{z})} \left[\frac{p_\theta(\mathbf{x}, \mathbf{z}^{(1)})}{q(\mathbf{z}^{(1)})}\right]\right)$

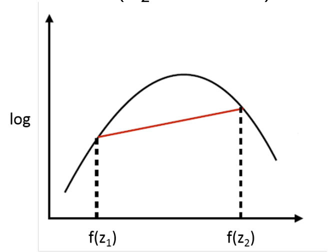
Evidence Lower Bound

Log-Likelihood function for Partially Observed Data is hard to compute:

$$\log\left(\sum_{\mathbf{z}\in\mathcal{Z}} p_\theta(\mathbf{x},\mathbf{z})\right) = \log\left(\sum_{\mathbf{z}\in\mathcal{Z}} \frac{q(\mathrm{z})}{q(\mathrm{z})} p_\theta(\mathbf{x},\mathbf{z})\right) = \log\left(\mathbb{E}_{\mathrm{z}\sim q(\mathrm{z})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathrm{z})}\right]\right)$$

- $\log()$ is a concave function. $\log(px + (1-p)x') \geq p\log(x) + (1-p)\log(x')$.
- Idea: use Jensen Inequality (for concave functions)

$$\log\left(\mathbb{E}_{\mathrm{z}\sim q(\mathrm{z})}[f(\mathrm{z})]\right) = \log\left(\sum_{\mathrm{z}} q(\mathrm{z})f(\mathrm{z})\right) \geq \sum_{\mathrm{z}} q(\mathrm{z})\log f(\mathrm{z})$$



Evidence Lower Bound

Log-Likelihood function for Partially Observed Data is hard to compute:

$$\log\left(\sum_{\mathbf{z}\in\mathcal{Z}} p_\theta(\mathbf{x},\mathbf{z})\right) = \log\left(\sum_{\mathbf{z}\in\mathcal{Z}} \frac{q(\mathrm{z})}{q(\mathrm{z})} p_\theta(\mathbf{x},\mathbf{z})\right) = \log\left(\mathbb{E}_{\mathrm{z}\sim q(\mathrm{z})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathrm{z})}\right]\right)$$

- $\log()$ is a concave function. $\log(px + (1-p)x') \geq p\log(x) + (1-p)\log(x')$.
- Idea: use Jensen Inequality (for concave functions)

$$\log\left(\mathbb{E}_{\mathrm{z}\sim q(\mathrm{z})}\left[f(\mathrm{z})\right]\right) = \log\left(\sum_{\mathrm{z}} q(\mathrm{z}) f(\mathrm{z})\right) \geq \sum_{\mathrm{z}} q(\mathrm{z}) \log f(\mathrm{z})$$

Choosing $f(\mathrm{z}) = \frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathrm{z})}$

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

Called Evidence Lower Bound (**ELBO**).

Variational inference

- Suppose $q(\mathrm{z})$ is **any** probability distribution over the hidden variables
- **Evidence lower bound** (ELBO) holds for any q

$$
\begin{aligned}
\log p(\mathrm{x};\theta) &\geq \sum_{\mathrm{z}} q(\mathrm{z}) \log \left(\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathrm{z})} \right) \\
&= \sum_{\mathrm{z}} q(\mathrm{z}) \log p_\theta(\mathbf{x},\mathbf{z}) \underbrace{- \sum_{\mathrm{z}} q(\mathrm{z}) \log q(\mathrm{z})}_{\text{Entropy } H(q) \text{ of } q} \\
&= \sum_{\mathrm{z}} q(\mathrm{z}) \log p_\theta(\mathbf{x},\mathbf{z}) + H(q)
\end{aligned}
$$

- Equality holds if $q = p(\mathrm{z}|\mathrm{x};\theta)$

$$
\log p(\mathrm{x};\theta) {\color{red}=} \sum_{\mathrm{z}} q(\mathrm{z}) \log p(\mathrm{z},\mathrm{x};\theta) + H(q)
$$

- (Aside: This is what we compute in the E-step of the EM algorithm)

Why is the bound tight

- We derived this lower bound that holds holds for any choice of $q(\mathrm{z})$:

$$\log p(\mathrm{x};\theta) \geq \sum_{\mathrm{z}} q(\mathrm{z}) \log \frac{p(\mathrm{x},\mathrm{z};\theta)}{q(\mathrm{z})}$$

- If $q(\mathrm{z}) = p(\mathrm{z}|\mathrm{x};\theta)$ the bound becomes:

$$\begin{aligned}
\sum_{\mathrm{z}} p(\mathrm{z}|\mathrm{x};\theta) \log \frac{p(\mathrm{x},\mathrm{z};\theta)}{p(\mathrm{z}|\mathrm{x};\theta)} &= \sum_{\mathrm{z}} p(\mathrm{z}|\mathrm{x};\theta) \log \frac{p(\mathrm{z}|\mathrm{x};\theta)p(\mathrm{x};\theta)}{p(\mathrm{z}|\mathrm{x};\theta)} \\
&= \sum_{\mathrm{z}} p(\mathrm{z}|\mathrm{x};\theta) \log p(\mathrm{x};\theta) \\
&= \log p(\mathrm{x};\theta) \underbrace{\sum_{\mathrm{z}} p(\mathrm{z}|\mathrm{x};\theta)}_{=1} \\
&= \log p(\mathrm{x};\theta)
\end{aligned}$$

- Confirms our previous importance sampling intuition: we should choose likely completions.
- What if the posterior $p(\mathrm{z}|\mathrm{x};\theta)$ is intractable to compute? How loose is the bound?

Variational inference continued

- Suppose $q(\mathrm{z})$ is **any** probability distribution over the hidden variables. A little bit of algebra reveals

$$D_{KL}(q(\mathrm{z}) \| p(\mathrm{z}|\mathrm{x};\theta)) = -\sum_{\mathrm{z}} q(\mathrm{z}) \log p(\mathrm{z},\mathrm{x};\theta) + \log p(\mathrm{x};\theta) - H(q) \geq 0$$

- Rearranging, we re-derived the **Evidence lower bound** (ELBO)

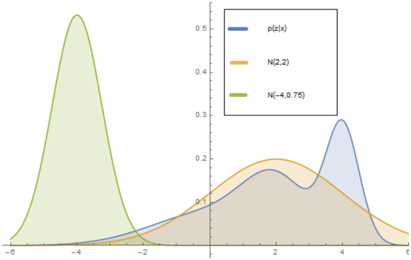
$$\log p(\mathrm{x};\theta) \geq \sum_{\mathrm{z}} q(\mathrm{z}) \log p(\mathrm{z},\mathrm{x};\theta) + H(q)$$

- Equality holds if $q = p(\mathrm{z}|\mathrm{x};\theta)$ because $D_{KL}(q(\mathrm{z}) \| p(\mathrm{z}|\mathrm{x};\theta)) = 0$

$$\log p(\mathrm{x};\theta) {\color{red}=} \sum_{\mathrm{z}} q(\mathrm{z}) \log p(\mathrm{z},\mathrm{x};\theta) + H(q)$$

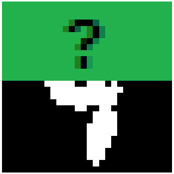
- In general, $\log p(\mathrm{x};\theta) = \mathrm{ELBO} + D_{KL}(q(\mathrm{z}) \| p(\mathrm{z}|\mathrm{x};\theta))$. The closer $q(\mathrm{z})$ is to $p(\mathrm{z}|\mathrm{x};\theta)$, the closer the ELBO is to the true log-likelihood

The Evidence Lower bound



- What if the posterior $p(\mathrm{z}|\mathrm{x};\theta)$ is intractable to compute?
- Suppose $q(\mathrm{z};\phi)$ is a (tractable) probability distribution over the hidden variables parameterized by ϕ (variational parameters)
  - For example, a Gaussian with mean and covariance specified by ϕ
$$q(\mathrm{z};\phi) = \mathcal{N}(\phi_1, \phi_2)$$
- **Variational inference**: pick ϕ so that $q(\mathrm{z};\phi)$ is as close as possible to $p(\mathrm{z}|\mathrm{x};\theta)$. In the figure, the posterior $p(\mathrm{z}|\mathrm{x};\theta)$ (blue) is better approximated by $\mathcal{N}(2,2)$ (orange) than $\mathcal{N}(-4,0.75)$ (green)

A variational approximation to the posterior

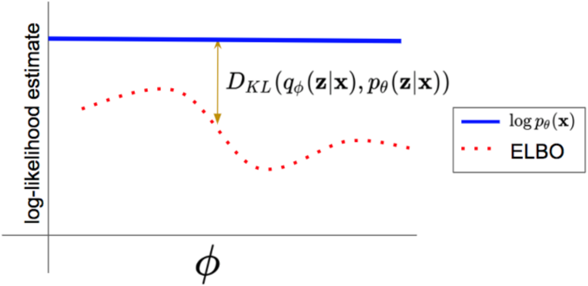


- Assume $p(\mathrm{x}^{top}, \mathrm{x}^{bottom}; \theta)$ assigns high probability to images that look like digits. In this example, we assume $\mathrm{z} = \mathrm{x}^{top}$ are unobserved (latent)
- Suppose $q(\mathrm{x}^{top}; \phi)$ is a (tractable) probability distribution over the hidden variables (missing pixels in this example) x^{top} parameterized by ϕ (variational parameters)

$$q(\mathrm{x}^{top}; \phi) = \prod_{\text{unobserved variables } \mathrm{x}_i^{top}} (\phi_i)^{\mathrm{x}_i^{top}} (1 - \phi_i)^{(1 - \mathrm{x}_i^{top})}$$

- Is $\phi_i = 0.5 \ \forall i$ a good approximation to the posterior $p(\mathrm{x}^{top} | \mathrm{x}^{bottom}; \theta)$? No
- Is $\phi_i = 1 \ \forall i$ a good approximation to the posterior $p(\mathrm{x}^{top} | \mathrm{x}^{bottom}; \theta)$? No
- Is $\phi_i \approx 1$ for pixels i corresponding to the top part of digit **9** a good approximation? Yes

The Evidence Lower bound



$$\begin{aligned}\log p(\mathrm{x};\theta) &\geq \sum_{\mathrm{z}} q(\mathrm{z};\phi)\log p(\mathrm{z},\mathrm{x};\theta) + H(q(\mathrm{z};\phi)) = \underbrace{\mathcal{L}(\mathrm{x};\theta,\phi)}_{\text{ELBO}} \\ &= \mathcal{L}(\mathrm{x};\theta,\phi) + D_{KL}(q(\mathrm{z};\phi)\|p(\mathrm{z}|\mathrm{x};\theta))\end{aligned}$$

The better $q(\mathrm{z};\phi)$ can approximate the posterior $p(\mathrm{z}|\mathrm{x};\theta)$, the smaller $D_{KL}(q(\mathrm{z};\phi)\|p(\mathrm{z}|\mathrm{x};\theta))$ we can achieve, the closer ELBO will be to $\log p(\mathrm{x};\theta)$. Next: jointly optimize over θ and ϕ to maximize the ELBO over a dataset

Summary

- Latent Variable Models Pros:
  - Easy to build flexible models
  - Suitable for unsupervised learning
- Latent Variable Models Cons:
  - Hard to evaluate likelihoods
  - Hard to train via maximum-likelihood
  - Fundamentally, the challenge is that posterior inference $p(\mathrm{z} \mid \mathrm{x})$ is hard. Typically requires variational approximations
- Alternative: give up on KL-divergence and likelihood (GANs)