# CSC 412/2506:
# Probabilistic Learning and Reasoning

## Week 6 - 2/2: Variational Inference I

Michal Malyska

University of Toronto

# Overview

- Variational Inference
- M-projection
- I-projection
- Naive mean-field approach

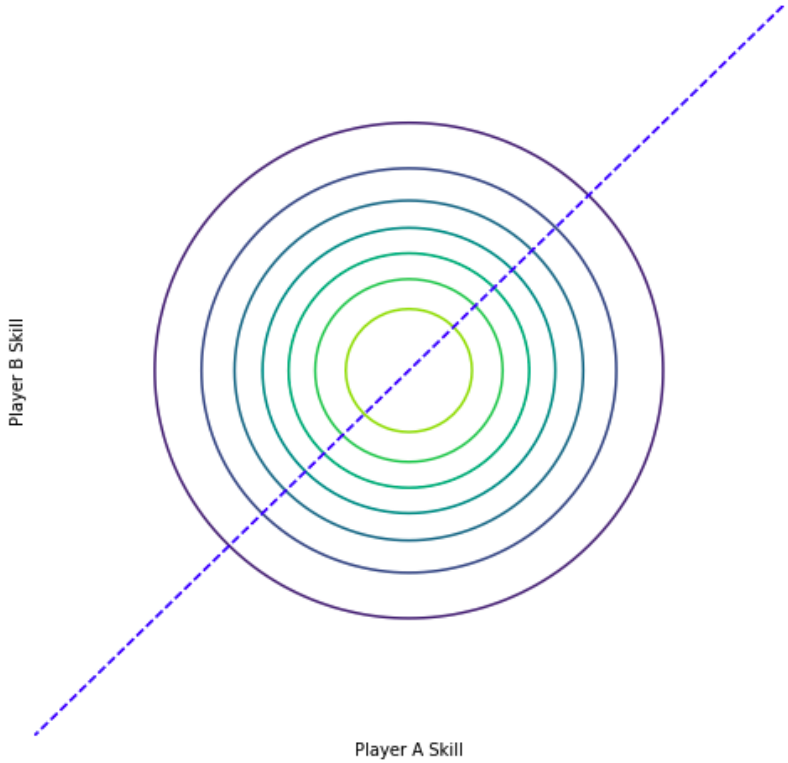
# Posterior Inference for Latent Variable Models

We've worked with a few latent variable models, such as the generative image model and the trueskill model.
These models have a factorization $p(x, z) = p(z)p(x|z)$ where

- $x$ are the observations or data,
- $z$ are the unobserved (latent) variables
- $p(z)$ is usually called the **prior**
- $p(x|z)$ is usually called the **likelihood**
- The conditional distribution of the unobserved variables given the observed variables (aka the **posterior**) is

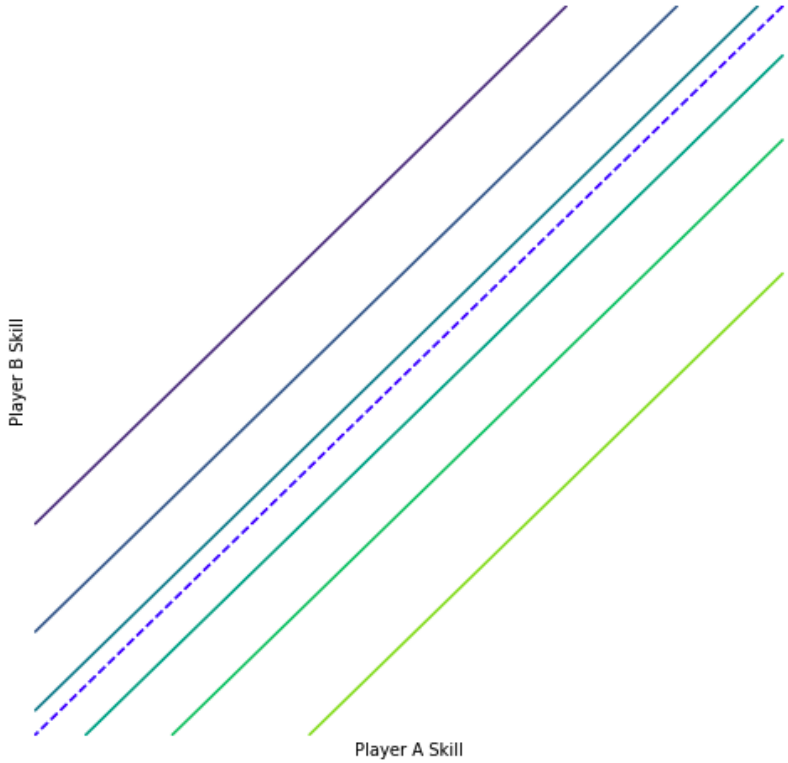
$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} = \frac{p(x|z)p(z)}{\int p(x, z)dz}$$

Prior:



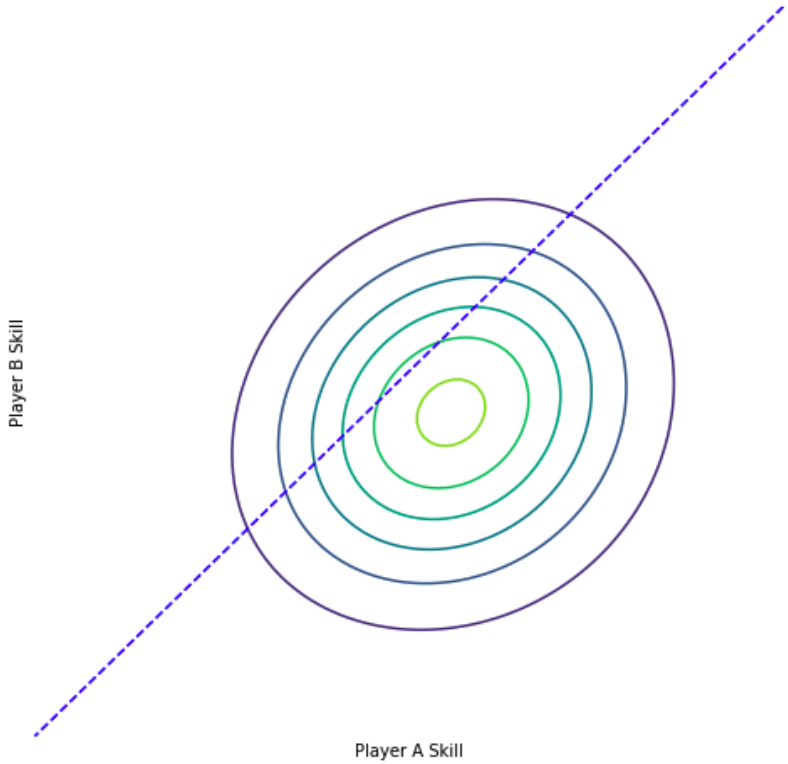

Says we're very uncertain about both player's skill.

Likelihood:



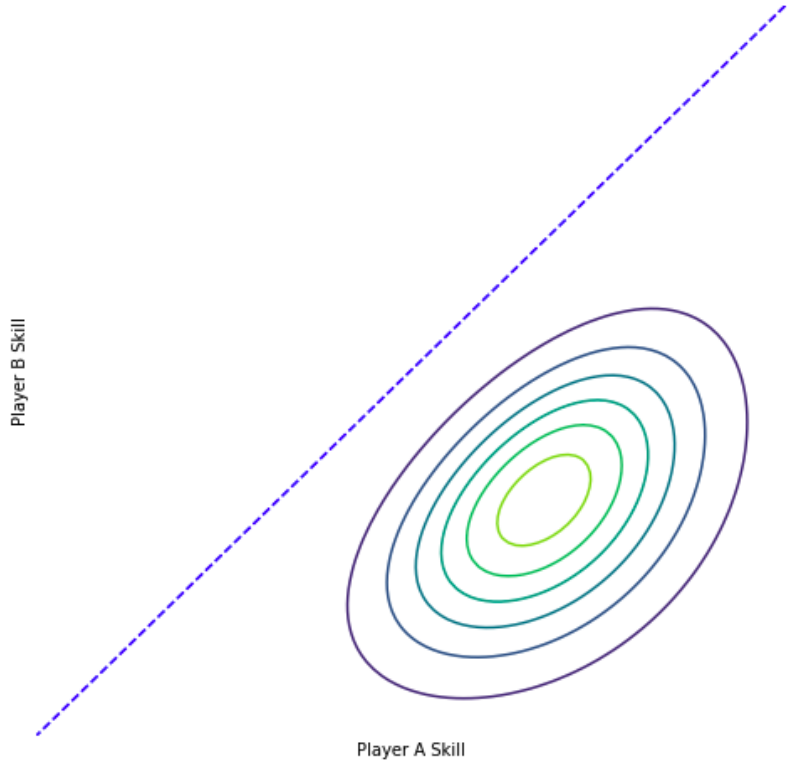

This is the part of the model that gives meaning to the latent variables.

Posterior:



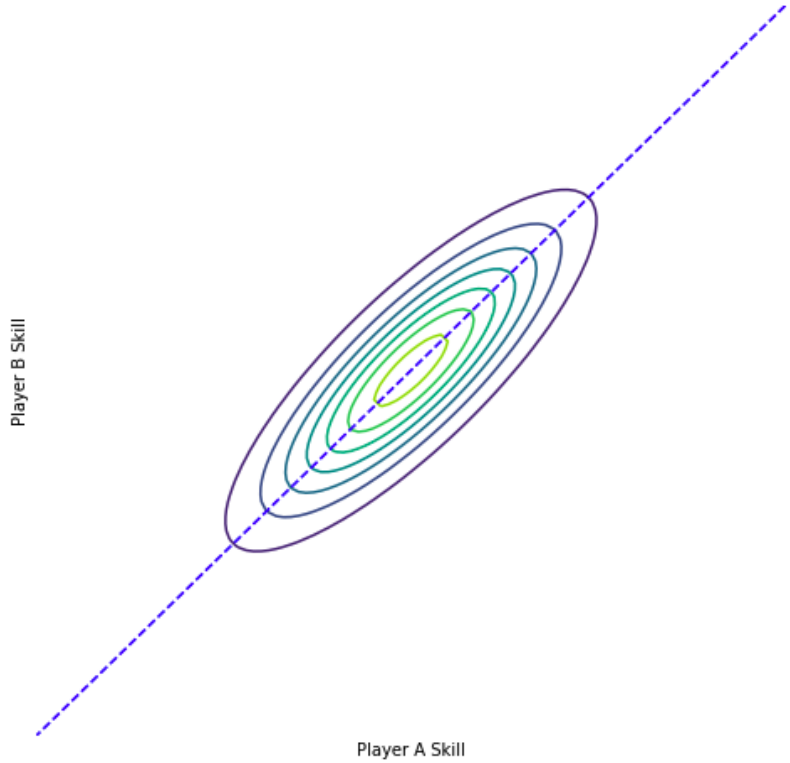

The posterior isn't Gaussian anymore.

Posterior after A beats B 10 times:



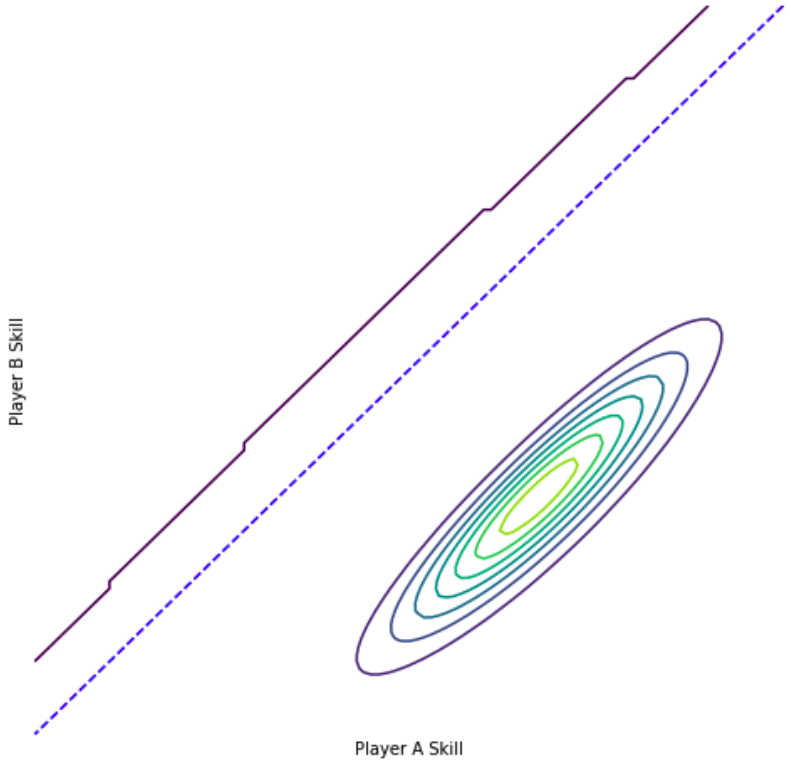

Now the posterior is certain that A is better than B.

Posterior after both beat each other 10 times:




Now the posterior is certain that neither player is much better than the other, but is uncertain how good they both are in an absolute sense.

Posterior after 90 wins vs 10:




In general, the more evidence we have, the more the posterior will shrink.

# What is hard to compute about the posterior?

- The integral $p(x) = \int p(x, z)dz$ is intractable whenever $z$ is high dimensional. This makes evaluating the normalized posterior $p(z|x)$ for a given $x$ and $z$ also intractable and sampling difficult.
- Here is a list of operations that are expensive:
  - Computing the evidence / marginal likelihood $p(x) = \int p(z, x)dz$
    - Useful for choosing between models, or fitting model parameters.
  - Computing a posterior probability: $p(z|x) = \frac{p(z)p(x|z)}{p(x)}$
  - Computing marginals of $p(z_1|x) = \int p(z_1, z_2, \ldots z_D|x)dz_2, dz_3, \ldots dz_D$
    - E.g. finding the posterior over a single tennis player's skill given all games.
  - Sampling $z \sim p(z|x)$
    - Useful for summarizing which hypotheses are likely given the data, making predictions, and decisions.

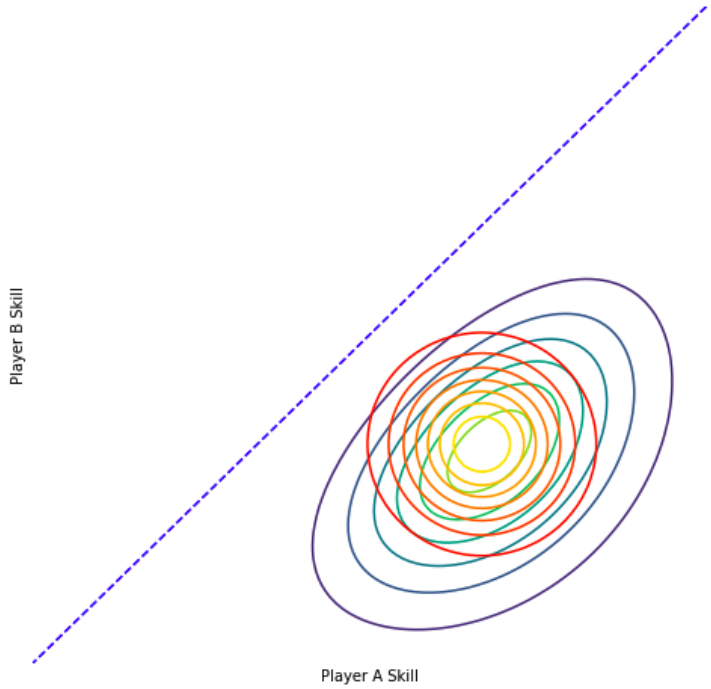
# Variational methods

- Variational inference is closely related to the calculus of variations, developed in the 1700s by Euler, Lagrange.
- Calculus of variations is the calculus of functionals (which take functions as arguments).
- Variational inference is an approximate inference method where we seek a tractable (e.g., factorized) approximation to the target intractable distribution.

# Variational methods

To be more formal, variational inference works as follows:

- Choose a tractable distribution $q(z) \in Q$ from a feasible set $Q$. This distribution will be used to approximate $p(z|x)$.
  - For example, $q(z) = \mathcal{N}(z|\mu, \Sigma)$. The idea is that we'll try choose a $Q$ that makes $q(z)$ a good approximation of the true posterior $p(z|x)$.
- Encode some notion of "difference" between $p(z|x)$ and $q$ that can be effciently estimated. Usually we will use the KL divergence.
- Minimize this difference. Usually we will use an iterative optimization method.




- Whatever feasible set we choose for $Q$, it's usually not the case that there is any $q \in Q$ that exactly matches the true posterior.
- But computing the true posterior is intractable, so we have to take a shortcut somewhere.

# How to measure closeness: KL divergence

We will measure the difference between $q$ and $p$ using the **Kullback-Leibler divergence**

$$KL(q(z)||p(z|x)) = \int q(z) \log \frac{q(z)}{p(z|x)} dz$$
$$= \mathop{\mathbb{E}}_{z\sim q} \log \frac{q(z)}{p(z|x)}$$

Properties of the KL Divergence

- $KL(q||p) \geq 0$
- $KL(q||p) = 0 \Leftrightarrow q = p$
- $KL(q||p) \neq KL(p||q)$
- KL divergence is not a metric, since it's not symmetric

# Which direction of KL to use? $KL(q||p)$ vs $KL(p||q)$

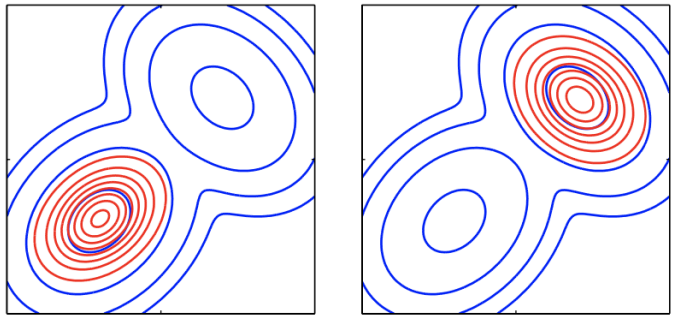
- We could minimize $KL(q||p)$ or $KL(p||q)$
- Which one to choose?
- As always, we will go with the tractable one.

# Information (I-)Projection:

I-projection: $q^* = \arg\min_{q \in Q} KL(q||p) = \mathbb{E}_{x \sim q(x)} \log \frac{q(x)}{p(x)}$:

- $p \approx q \implies KL(q||p)$ small
- I-projection underestimates support, and does not yield the correct moments.
- $KL(q||p)$ penalizes $q$ having mass where $p$ has none.

$p(x)$ is mixture of two 2D Gaussians and $Q$ is the set of all 2D Gaussian distributions (with arbitrary covariance matrices)



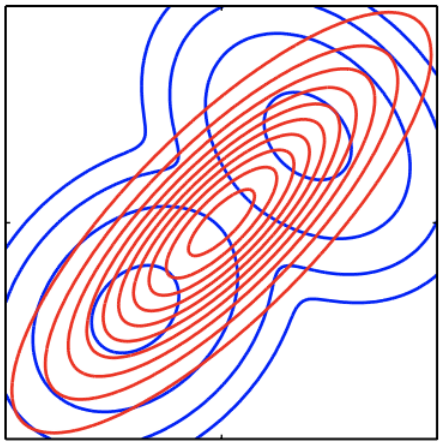
$p$=Blue, $q^*$=Red (two equivalently good solutions!)

# Moment (M-)projection

M-projection: $q^* = \arg\min_{q \in Q} KL(p||q) = \mathbb{E}_{x \sim p(x)} \log \frac{p(x)}{q(x)}$:

- $p \approx q \implies KL(p||q)$ small
- $KL(p||q)$ penalizes $q$ missing mass where $p$ has some.
- M-projection yields a distribution $q(x)$ with the correct mean and covariance.

$p(x)$ is mixture of two 2D Gaussians and $Q$ is the set of all 2D Gaussian distributions (with arbitrary covariance matrices)



$p$=Blue, $q^*$=Red

# Maximum entropy interpretation

- A related quantity is the **entropy**:

$$H(p) = -\mathbb{E}_{x\sim p(x)} \log p(x)$$

  measuring the uncertainty in the distribution $p$.
- Consider the optimization problem

$$\begin{aligned} &\text{maximize } H(p) \\ &\text{subject to } \mathbb{E}_{x\sim p(x)}[f_i(x)] = t_i \quad \text{for } i = 1, .., k. \end{aligned}$$

- **Theorem**: Exponential family of distributions maximize the entropy $H(p)$ over all distributions satisfying

$$\mathbb{E}_{x\sim p(x)}[f_i(x)] = t_i \quad \text{for } i = 1, .., k.$$

- In M-projection, if $Q$ is set of exponential families, then the expected sufficient statistics wrt $q^*(x)$ is the same as that wrt $p(x)$.
- M-projection require expectation wrt $p$, hence intractable.
- Most variational inference algorithms make use of the I-projection.

# Mean-field approach

- Say we have an arbitrary MRF:

$$p(x|\theta) = \exp\left\{\sum_{c\in\mathcal{C}} \phi_c(x_c) - \log Z(\theta)\right\}$$

- We find an approximate distribution $q(x) \in Q$ by performing I-projection to $p(x)$.

$$q^* = \arg\min_{q\in Q} KL(q||p) = \mathbb{E}_{x\sim q(x)} \log \frac{q(x)}{p(x|\theta)}$$

$$\arg\min_{q\in Q} KL(q||p) = \mathbb{E}_{x\sim q(x)}\left[\log q(x) - \sum_{c\in\mathcal{C}} \phi_c(x_c) + \log Z(\theta)\right]$$

$$= \arg\max_{q\in Q} \sum_{c\in\mathcal{C}} \mathbb{E}_q[\phi_c(x_c)] + H(q)$$

- For tractability, we need a nice set $Q$. If $p \in Q$, then $q^* = p$. But this almost never happens.

# Naive Mean-Field

- One way to proceed is the mean-field approach where we assume:

$$q(x) = \prod_{i \in V} q_i(x_i)$$

  the set $Q$ is composed of those distributions that factor out.
- Using this in the maximization problem, we can simplify things

$$q^* = \arg\max_{q \in Q} \sum_{c \in \mathcal{C}} \sum_{x_c} q(x_c)\phi_c(x_c) + H(q)$$

- We notice $q(x_c) = \prod_{i \in c} q_i(x_i)$ and also

$$
\begin{aligned}
H(q) =& \mathbb{E}_q[-\log q(x)] = -\sum_x q(x) \log q(x) \\
&= -\sum_x q(x) \Big[ \sum_i \log q_i(x_i) \Big] \\
&= -\sum_i \sum_x \Big[ q_i(x_i) \log q_i(x_i) \Big] \frac{q(x)}{q_i(x_i)} \\
&= -\sum_i \sum_{x_i} \Big[ q_i(x_i) \log q_i(x_i) \Big] \sum_{x \setminus x_i} \frac{q(x)}{q_i(x_i)} \\
&= -\sum_i \sum_{x_i} \Big[ q_i(x_i) \log q_i(x_i) \Big] \\
&= \sum_i H(q_i)
\end{aligned}
$$

# Example: Pairwise MRF

- Thus the final optimization problem reduces to

$$q^* = \arg\max_q \sum_{c \in \mathcal{C}} \sum_{x_c} \phi_c(x_c) \prod_{i \in c} q_i(x_i) + \sum_i H(q_i)$$

$$\text{subject to: } q_i(x_i) \geq 0 \text{ and } \sum_{x_i} q_i(x_i) = 1.$$

- Let's further simplify the setting and assume that we have a pairwise MRF. Then the optimization problem becomes

$$q^* = \arg\max_q \sum_{(i,j) \in \mathcal{E}} \sum_{x_i, x_j} \phi_{ij}(x_i, x_j) q_i(x_i) q_j(x_j) - \sum_i \sum_{x_i} q_i(x_i) \log(q_i(x_i))$$

$$\text{subject to: } q_i(x_i) \geq 0 \text{ and } \sum_{x_i} q_i(x_i) = 1.$$

# Coordinate maximization

This problem is hard as it has many local maxima! But we can still try to optimize using block coordinate ascent.

- Initialize $\{q_i(x_i)\}_{i \in V}$ uniformly
- Iterate over $i \in V$
  - Greedily maximize the objective over $q_i(x_i)$
  - This is equivalent to: $q_i(x_i) \propto \exp\left\{ \sum_{j \in N(i)} \sum_{x_j} q_j(x_j) \phi_{ij}(x_i, x_j) \right\}$
    - which follows from: write the Lagrangian, take the derivative, set to zero, and solve
  - Repeat until convergence.

This is guaranteed to converge but can converge to local optima.

# Summary

- Approximate the complex (intractable) distribution with a simpler (tractable) distribution
- I-projection & M-projection measure the distance to true posterior
- Mean field approximation is a way to simplify the set of distributions
- More variational inference after midterm.