10417-617
Deep Learning: Fall 2020

Andrej Risteski

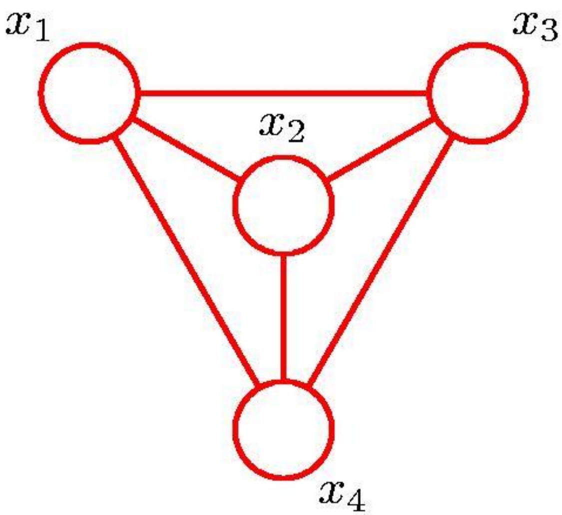
Machine Learning Department

Lecture 13:

Variational methods, applications to learning latent-variable directed models

Graphical Models

Recall: graph contains a set of nodes connected by edges.



In a probabilistic graphical model, each node represents a random variable, links represent "probabilistic dependencies" between random variables.

Graph specifies how joint distribution over all random variables decomposes into a **product** of factors, each factor depending on a subset of the variables.

Two types of graphical models:

- **Bayesian networks**, also known as Directed Graphical Models (the links have a particular directionality indicated by the arrows)
- **Markov Random Fields**, also known as Undirected Graphical Models (the links do not carry arrows and have no directional significance).

Algorithmic pros/cons of latent-variable models (so far)

RBM's

- Hard to draw samples ✗

(In fact, #P-hard provably, even in Ising models)

- Easy to sample posterior distribution over latents ✓

Directed models

- Easy to draw samples ✓

- Hard to sample posterior distribution over latents ✗

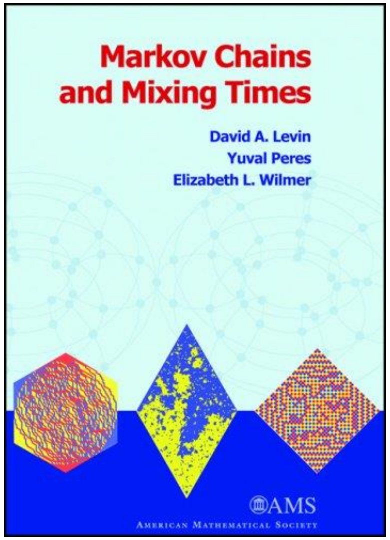
(In fact, #P-hard even in mixtures)

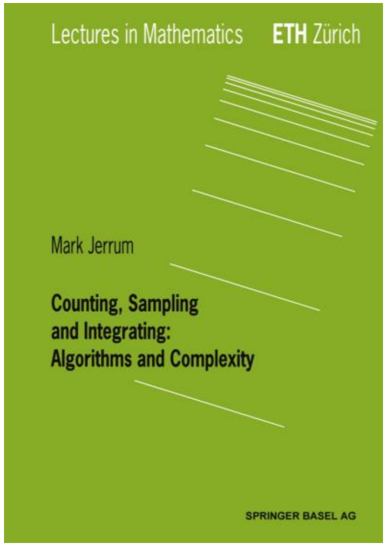
Algorithmic approaches

When faced with a difficult to calculate probabilistic quantity (partition function, difficult posterior), there are two families of approaches:

MARKOV CHAIN MONTE CARLO

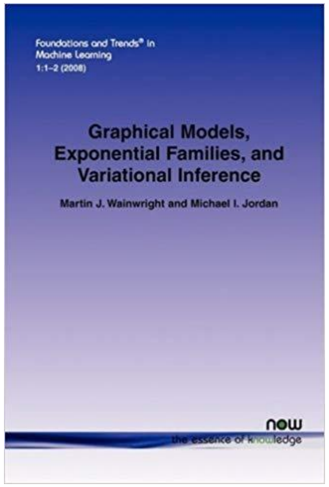
- Random walk w/ equilibrium distribution the one we are trying to sample from.





VARIATIONAL METHODS

- Based on solving an optimization problem.



Part I: approximating posteriors via variational methods

Sampling posteriors in latent-variable directed models

Recall, sampling from the posterior distribution P(z|x) is **hard**:

$$P(\text{Diseases}, \text{Symptoms}) = P(\text{Diseases}) \;\; P(\text{Symptoms}|\text{Diseases})$$

Latent

Data

Simple, explicit

By Bayes rule, $P(\text{Diseases}|\text{Symptoms}) \propto P(\text{Diseases}, \text{Symptoms})$

Up to normalizing const, simple...

Complicated partition function:

$$\sum_{\text{Diseases}} P(\text{Diseases}, \text{Symptoms})$$

Again, can be #P-hard to sample from!!

Variational methods for approximating posteriors

Gibbs variational principle: Let $p(z, x)$ be a joint distribution over latent variables and observables. Then,

$$p(z|x) = \underset{q(z|x)\text{:distribution over } Z}{\operatorname{argmax}} \mathbb{E}_{q\ (z|x)} \log q\ (z|x) - \mathbb{E}_{q\ (z|x)} \log p\ (z, x)\}$$

$$-H\big(q\ (z|x)\big) \quad - \quad \mathbb{E}_{z\sim q}[\log\ p(z, x)]$$

In fact, for every $q(z|x)$, we have

$$\log p(x) \ = -\big(-H\big(q\ (z|x)\big) - \mathbb{E}_{z\sim q}[\log\ p(z, x)]\big) + KL(q\ (z|x)||\ p(z|x))$$

Variational methods for partition functions

Gibbs variational principle: Let $p(z,x)$ be a joint distribution over latent variables and observables. Then,

$$p(z|x) = \underset{q(z|x)\text{:distribution over } Z}{\operatorname{argmax}} \mathbb{E}_{q\,(z|x)} \log q\,(z|x) - \mathbb{E}_{q\,(z|x)} \log p\,(z,x)\}$$

In fact, for every $q(z|x)$, we have

$$\log p(x) = KL(q\,(z|x)||\,p(z|x)) - (-H(q\,(z|x)) - \mathbb{E}_{z\sim q}[\log\ p(z,x)])$$

Why:

$$0 \leq KL(q\,(z|x)||\,p(z|x)) = \mathbb{E}_{q\,(z|x)} \log q\,(z|x) - \mathbb{E}_{q\,(z|x)} p(z|x)$$

$$= -H(q\,(z|x)) - \mathbb{E}_{q\,(z|x)} \log \frac{p(z,x)}{p(x)}$$

$$= -H(q\,(z|x)) - \mathbb{E}_{q\,(z|x)} \log p\,(z,x) + \log p(x)$$

Equality is attained if and only if $KL(q\,(z|x)||\,p(z|x))$=0 i. e. $q\,(z|x) = p(z|x)$

Variational methods for approximating posteriors

Gibbs variational principle: Let $p(z,x)$ be a joint distribution over latent variables and observables. Then,

$$p(z|x) = \underset{q(z|x):\text{distribution over } Z}{\operatorname{argmax}} \mathbb{E}_{q\,(z|x)} \log q\,(z|x) - \mathbb{E}_{q\,(z|x)} \log p\,(z,x)\}$$

$$\log p(x) \;= -\big(-H\big(q\,(z|x)\big) - \mathbb{E}_{z\sim q}[\log\; p(z,x)]\big) + KL\big(q\,(z|x)\big|\big|\, p(z|x)\big)$$

Why is this useful?

(1) Instead of finding the argmax over **all** distributions over Z, we can maximize over some **simpler** parametric family Q, i.e. we can solve

$$\max_{q(z|x)\in Q} \mathbb{E}_{q\,(z|x)} \log q\,(z|x) - \mathbb{E}_{q\,(z|x)} \log p\,(z,x)$$

The argmax of the above distribution solves $\min_{q(Z|X)\in Q} KL\big(q\,(z|x)\big|\big|\, p(z|x)\big)$.

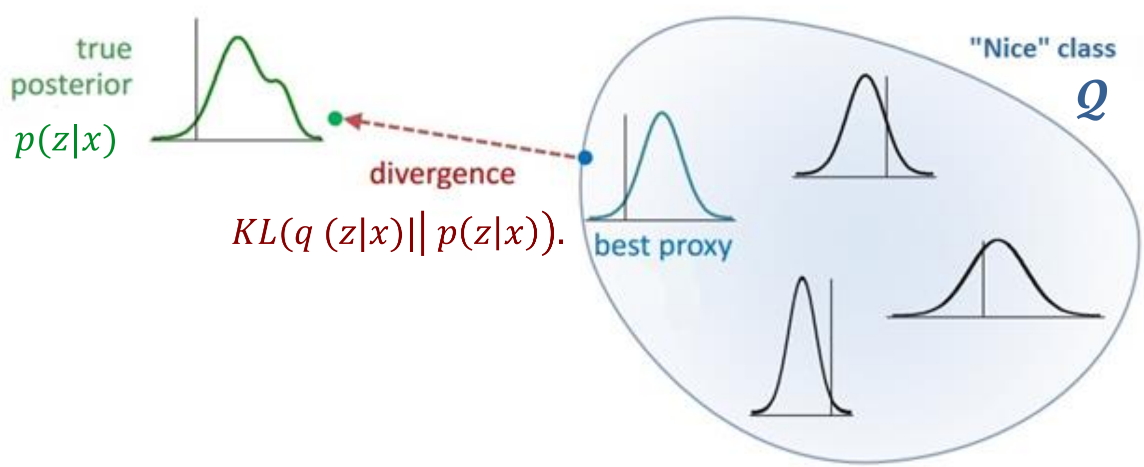
In other words, we are finding the **projection** of $p(z|x)$ onto Q.

Variational methods for approximating posteriors

Gibbs variational principle: Let $p(z,x)$ be a joint distribution over latent variables and observables. Then,

$$p(z|x) = \underset{q(z|x):\text{distribution over } Z}{\operatorname{argmax}} \mathbb{E}_{q\,(z|x)} \log q\,(z|x) - \mathbb{E}_{q\,(z|x)} \log p\,(z,x)\}$$

$$\log p(x) \quad = -\big(-H\big(q\,(z|x)\big) - \mathbb{E}_{z\sim q}[\log\ p(z,x)]\big) + KL\big(q\,(z|x)\big\|\,p(z|x)\big)$$



Variational methods for approximating posteriors

Gibbs variational principle: Let $p(z, x)$ be a joint distribution over latent variables and observables. Then,

$$p(z|x) = \underset{q(z|x):\text{distribution over } Z}{\text{argmax}} \; \mathbb{E}_{q\,(z|x)} \log q\,(z|x) - \mathbb{E}_{q\,(z|x)} \log p\,(z, x)\}$$

$$\log p(x) \;\; = -\big(-H\big(q\,(z|x)\big) - \mathbb{E}_{z\sim q}[p(z,x)]\big) + KL\big(q\,(z|x)\big|\big|\,p(z|x)\big)$$

Why is this useful?

(1) Instead of finding the argmax over *all* distributions over Z, we can maximize over some **simpler** parametric family $\mathcal{Q}$, i.e. we can solve

$$\max_{q(z|x)\in\mathcal{Q}} \mathbb{E}_{q\,(z|x)} \log q\,(z|x) - \mathbb{E}_{q\,(z|x)} \log p\,(z, x)\}$$

There are several common families $\mathcal{Q}$ that are used for which the above optimization is solveable – we will see **mean-field** family today, **neural-net** parametrized families when we study variational autoencoders.

Variational methods for approximating posteriors

Gibbs variational principle: Let $p(z, x)$ be a joint distribution over latent variables and observables. Then,

$$p(z|x) = \underset{q(z|x):\text{distribution over } Z}{\operatorname{argmax}} \mathbb{E}_{q\,(z|x)} \log q\,(z|x) - \mathbb{E}_{q\,(z|x)} \log p\,(z, x)\}$$

$$\log p(x) \;\; = -\big(-H\big(q\,(z|x)\big) - \mathbb{E}_{z\sim q}[p(z, x)]\big) + KL\big(q\,(z|x)\big|\big|\, p(z|x)\big)$$

Why is this useful?

(2) Provides a lower bound on $\log p(x)$ -- sometimes called the **ELBO (evidence lower bound)**, since

$$\log p(x) \geq \max_{q(z|x)\in Q} \mathbb{E}_{q\,(z|x)} \log q\,(z|x) - \mathbb{E}_{q\,(z|x)} \log p\,(z, x)$$

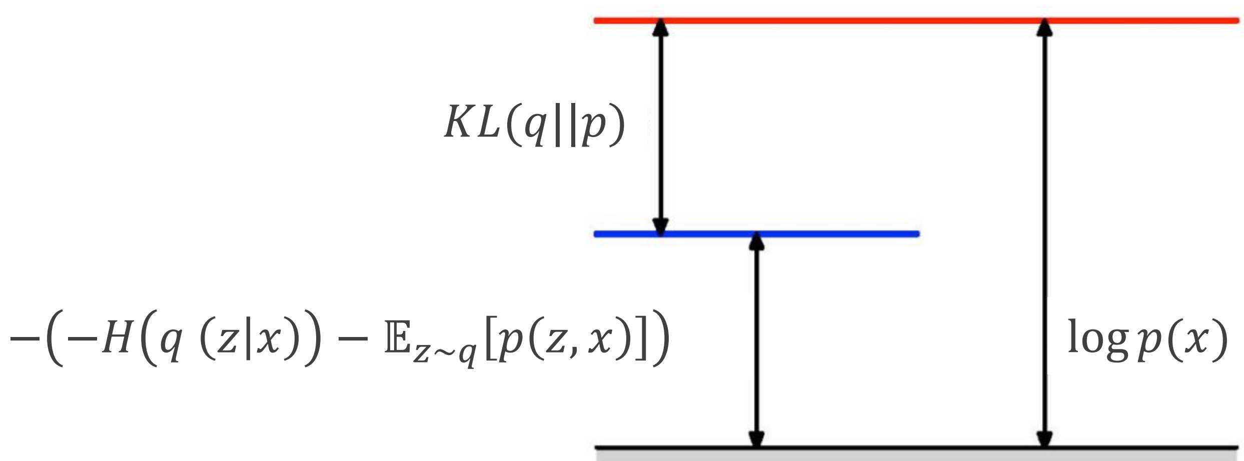
This will be useful when learning latent-variable directed models (stay tuned !).

Variational methods for approximating posteriors

Gibbs variational principle: Let $p(z,x)$ be a joint distribution over latent variables and observables. Then,

$$p(z|x) = \underset{q(z|x):\text{distribution over } Z}{\text{argmax}} \; \mathbb{E}_{q\,(z|x)} \log q\,(z|x) - \mathbb{E}_{q\,(z|x)} \log p\,(z,x)\}$$

$$\log p(x) \;\; = -\big(-H\big(q\,(z|x)\big) - \mathbb{E}_{z\sim q}[p(z,x)]\big) + KL\big(q\,(z|x)\big|\big|\, p(z|x)\big)$$



Solving the mean-field relaxation: coordinate ascent

Inspiration from physics: consider the case where Q contains product distributions, that is, for every $q(\cdot\,|x) \in Q$:

$$q(z|x) = \Pi_{i=1}^{d} q_i(z_i|x).$$

Consider updating a **single** coordinate of the mean-field distribution, that is keep $q_{-i}\,(z_i|x)$ fixed, and optimize for $q_i\,(z_i|x)$. We have:

$$KL(q\,(z|x)||\,p(z|x)) \;= \mathbb{E}_{q\,(z|x)} \log q\,(z|x) - \mathbb{E}_{q\,(z|x)} \log p\,(z,x)$$

$$= \sum_i \mathbb{E}_{q_i\,(z_i|x)} \log q_i\,(z_i|x) - \mathbb{E}_{q_i\,(z_i|x)} \left[\mathbb{E}_{q_{-i}(z_{-i}|x)} \log p\,(z_i, z_{-i}, x)\right]$$

$$= \mathbb{E}_{q_i\,(z_i|x)} \log q_i\,(z_i|x) - \mathbb{E}_{q_i\,(z_i|x)}[\log \tilde{p}\,(z_i, x)] + C$$

Renormalize to make it a distribution

Solving the mean-field relaxation: coordinate ascent

Inspiration from physics: consider the case where $\mathcal{Q}$ contains product distributions, that is, for every $q(\cdot\,|x) \in \mathcal{Q}$:

$$q(z|x) = \Pi_{i=1}^{d} q_i(z_i|x).$$

Consider updating a **single** coordinate of the mean-field distribution, that is keep $q_{-i}\,(z_i|x)$ fixed, and optimize for $q_i\,(z_i|x)$. We have:

$$KL(q\,(z|x)||\,p(z|x)) \;= \mathbb{E}_{q_i\,(z_i|x)} \log q_i\,(z_i|x) - \;\mathbb{E}_{q_i\,(z_i|x)}[\log \tilde{p}\,(z_i, x)] + C$$

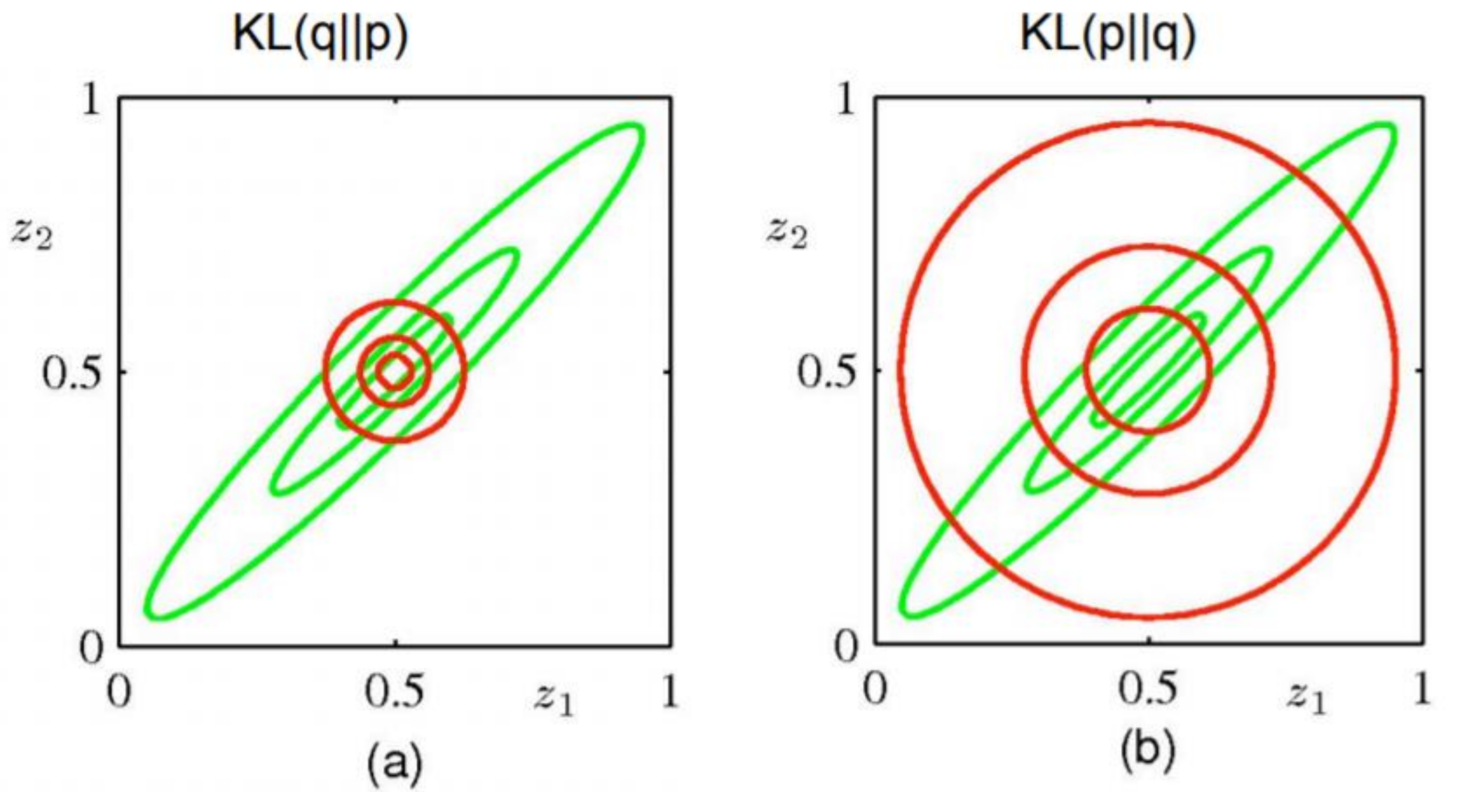
$$= KL(q_i\,(z_i|x)||\,\tilde{p}(z_i, x)) + C$$

Optimum is $q_i\,(z_i|x) = \tilde{p}(z_i, x)$

$$= \frac{\mathbb{E}_{q_{-i}(z_{-i}|x)} \log p\,(z_i, z_{-i}, x)}{\int_{z_i} \mathbb{E}_{q_{-i}(z_{-i}|x)} \log p\,(z_i, z_{-i}, x)}$$

Coordinate ascent: iterate above updates!

What if we changed the order of p, q in KL divergence?



Approximation is too compact.

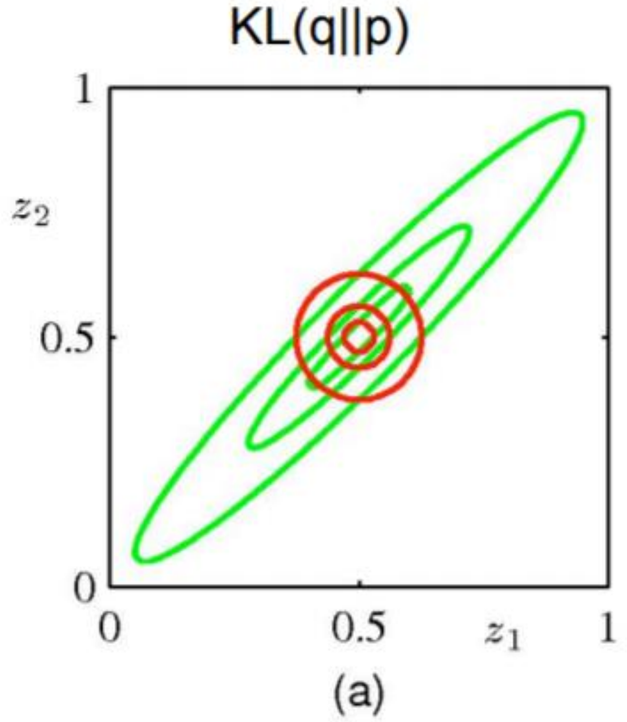
Approximation is too spread.

What if we changed the order of p, q in KL divergence?

$$\mathrm{KL}(q||p) = -\int q(\mathbf{Z}) \ln \frac{p(\mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z}.$$

There is a large positive contribution to the KL divergence from regions of Z space in which:

- p(Z) is near zero
- unless q(Z) is also close to zero.



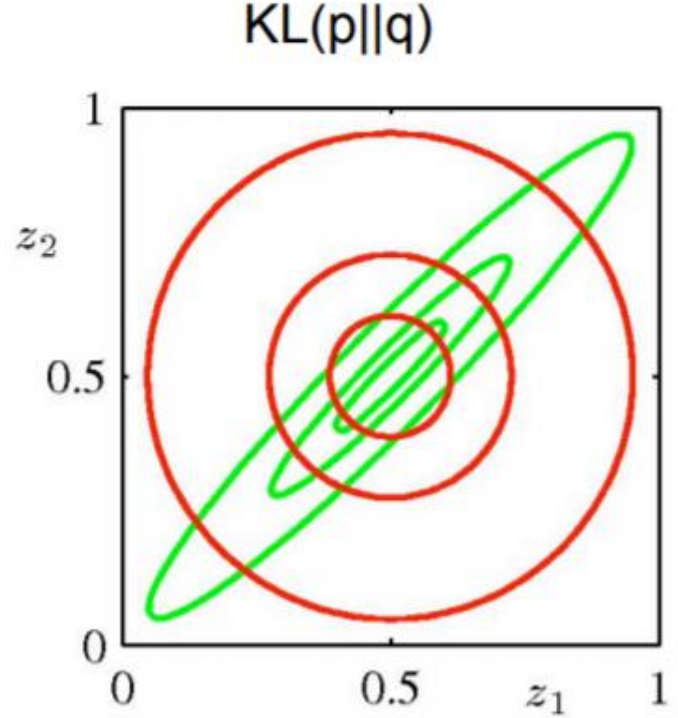
Minimizing KL(q||p) leads to distributions q(Z) that avoid regions in which p(Z) is small.

What if we changed the order of p, q in KL divergence?

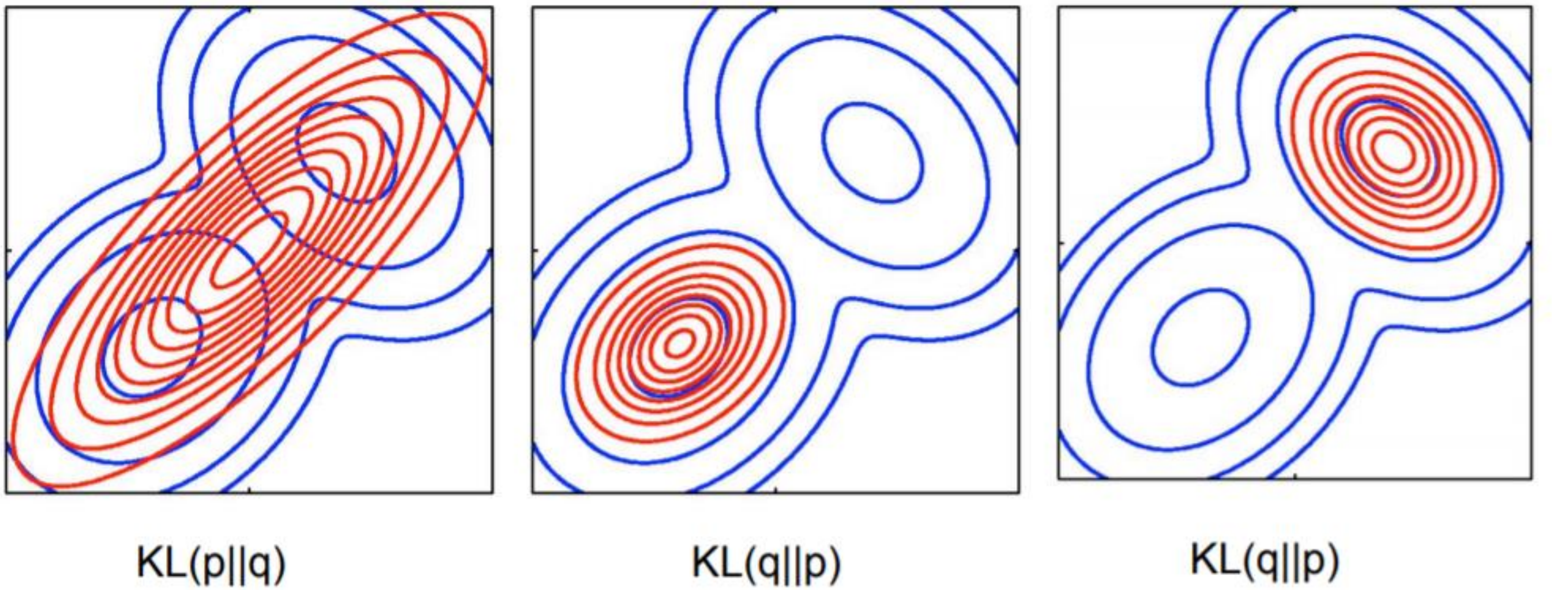
$$\mathrm{KL}(p||q) = -\int p(\mathbf{Z}) \ln \frac{q(\mathbf{Z})}{p(\mathbf{Z})} \mathrm{d}\mathbf{Z}.$$

There is a large positive contribution to the KL divergence from regions of Z space in which:
- q(Z) is near zero,
- unless p(Z) is also close to zero.

Minimizing KL(p||q) leads to distributions q(Z) that are nonzero in regions where p(Z) is nonzero.



What happens when posterior class is not rich enough?



Blue contours show bimodal distribution, red contours single Gaussian distribution that best approximates it.

KL(q||p) will tend to find a single mode, whereas KL(p||q) will average across all of the modes.

Part II: Learning latent-variable directed models

Learning latent-variable directed graphical models

How should we try to learn the parameters of a graphical model?

The most obvious strategy: maximum likelihood estimation

Given data $x_1, x_2, \ldots, x_n$, solve the optimization problem

$$\max_{\theta \in \Theta} \sum_{i=1}^{n} \log p(x_i)$$

Latent variables: we will use the Gibbs variational principle again!

$$\log_\theta \; p(x) = \max_{q(z|x)\text{: distribution over } z} H(q(z|x)) + \mathbb{E}_{q(z|x)}[\log p_\theta(x, z)]$$

Hence, MLE objective can be written as double maximization:

$$\max_{\theta \in \Theta} \max_{\{q_i(z|x_i)\}} \sum_{i=1}^{n} H(q_i(z|x_i)) + \mathbb{E}_{q_i(z|x_i)}[\log p_\theta(x_i, z)]$$

Expectation-maximization/ variational inference

The canonical algorithm for learning a single-layer latent-variable Bayesian network is an iterative algorithm as follows.

Consider the max-likelihood objective, rewritten as in the previous slide:

$$\max_{\theta\in\Theta} \max_{\{q_i(z|x_i)\in \mathcal{Q}\}} \sum_{i=1}^{n} H(q_i(z|x_i)) + \mathbb{E}_{q_i(z|x_i)}[\log p_\theta(x_i, z)]$$

Algorithm maintains iterates $\theta^t, \{q_i^t(z|x_i)\}$, and updates them iteratively

(1) **Expectation (E)-step**:

Keep θ^t fixed, set $\{q_i^{t+1}(z|x_i) \in \mathcal{Q}\}$, s.t. they maximize the objective above.

(2) **Maximization (M)-step**:

Keep $\{q_i^t(z|x_i)\}$ fixed, set θ^{t+1} s.t. it maximizes the objective above.

Clearly, every step cannot make the objective worse!

Does *not* mean it converges to global optimum – could, e.g. get stuck in a local minimum.

Expectation-maximization/ variational inference

The canonical algorithm for learning a single-layer latent-variable Bayesian network is an iterative algorithm as follows.

Consider the max-likelihood objective, rewritten as in the previous slide: $\max_{\theta \in \Theta} \max_{q_i(z|x_i)} \sum_{i=1}^{n} H(q_i(z|x_i)) + \mathbb{E}_{q_i(z|x_i)}[\log p_\theta(x_i, z)]$

Algorithm maintains iterates $\theta^t, q_i^t(z|x_i)$, and updates them iteratively

(1) **Expectation step**:

Keep θ^t and set $q_i^{t+1}(z|x_i)$, s.t. they maximize the objective above.

If the class is infinitely rich, the optimum is $q_i^{t+1}(z|x_i) = p_{\theta^t}(z|x_i)$

This is called **expectation-maximization (EM)**.
If class is not infinitely rich, it's called **variational inference**.

Examples

1: Mixtures of spherical Gaussians

Consider a mixture of K Gaussians with unknown means $p = \sum_{i=1}^{K} \frac{1}{K} \mathcal{N}(\mu_i, I_d)$

Let's try to calculate the E and M steps.

E-step: the optimal $q_i^{t+1}(z|x_i)$ is $p_{\theta^t}(z|x_i)$. Can we calculate this?

By Bayes rule, $p_{\theta^t}(z = k|x_i) \propto p(x_i|z = k) \propto e^{-\left|\left|x_i - \mu_k^t\right|\right|^2}$

Writing out the normalizing constant, we have

$$p_{\theta^t}(z = k|x_i) = \frac{e^{-\left|\left|x_i - \mu_k^t\right|\right|^2}}{\sum_{k'} e^{-\left|\left|x_i - \mu_{k'}^t\right|\right|^2}}$$

"Soft" version of assigning point to nearest cluster

Examples

<u>1: Mixtures of spherical Gaussians</u>

Consider a mixture of K Gaussians with unknown means $p = \sum_{i=1}^{K} \frac{1}{K} \mathcal{N}(\mu_i, I_d)$

Let's try to calculate the E and M steps.

M-step: given a quess $q_i^t(z|x_i)$, we can rewrite the maximization for θ as:

$$\max_{\theta \in \Theta} \sum_{i=1}^{n} H\left(q_i^t(z|x_i)\right) + \mathbb{E}_{q_i^t(z|x_i)}[\log p_\theta(x_i, z)]$$

$$= \mathbb{E}_{q_i^t(z|x_i)}\,[\cancel{\log p(z)} + \log p_\theta(x|z)]$$

$$\mathbb{E}_{q_i^t(z|x_i)}[\ \log p_\theta(x|z)]$$

Doesn't depend on θ

Examples

<u>1: Mixtures of spherical Gaussians</u>

Consider a mixture of K Gaussians with unknown means $p = \sum_{i=1}^{K} \frac{1}{K} \mathcal{N}(\mu_i, I_d)$

Let's try to calculate the E and M steps.

M-step: given a quess $q_i^t(z|x_i)$, we can rewrite the maximization for θ as:

$$\max_{\theta} \ \mathbb{E}_{q_i^t(z|x_i)}[\ \log p_\theta(x|z)] = \max_{\theta} \ -\sum_{i=1}^{n} \sum_{k=1}^{K} q_i^t(z=k|x_i)||x_i - \mu_k||^2$$

Setting the derivative wrt to μ_k to 0, we have:

$$\mu_k^{t+1} = \sum_{i=1}^{n} \frac{e^{-\left\|x_i - \mu_k^t\right\|^2}}{\sum_{k'} e^{-\left\|x_i - \mu_{k'}^t\right\|^2}} x_i$$

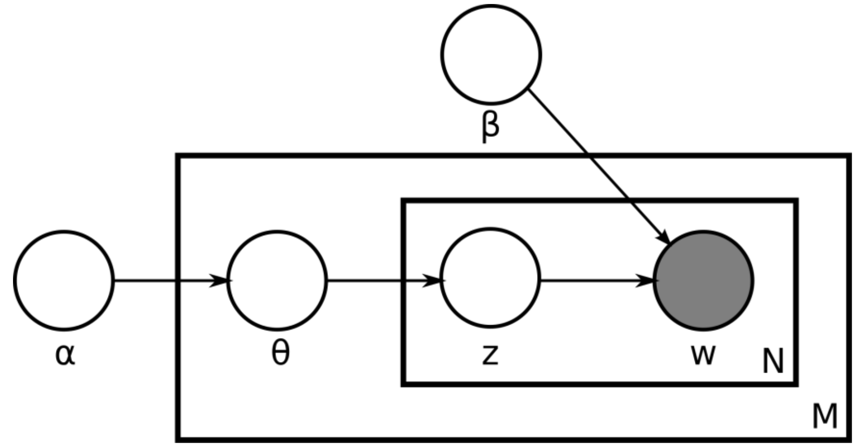
Average points, weighing nearby points more

Examples

2: Latent Dirichlet Allocation

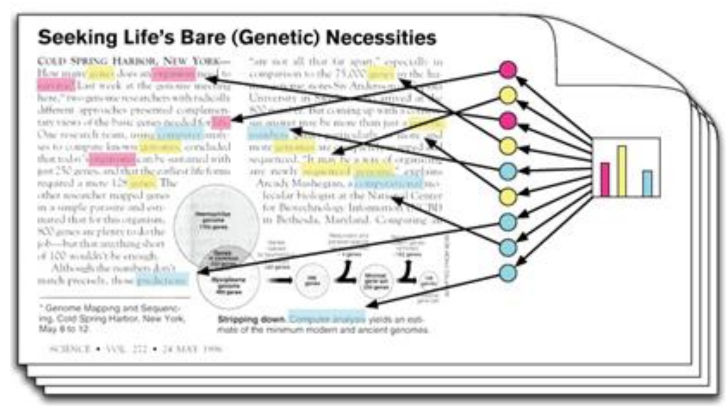
The **parameters** are: $\{\alpha_i\}_{i=1}^{K}$ (Dirichlet parameters) and matrix $\beta \in \mathbb{R}_{+}^{N\times K}$, where N is the size of the vocabulary.

The columns of β satisfy $\sum_{j=1}^{N} \beta_{ij} = 1$ (the distribution of words in a topic i)



To produce document:

- First, sample $\theta \sim \mathrm{Dir}(\cdot\,|\alpha)$: this will be the topic proportion vector for the document.
- Each word in the document is generated in order, independently.
- To generate word i:
  - Sample topic z_i with categorical distribution with parameters θ
  - Sample word w_i with categorical distribution with parameters β_{z_i}



Examples

The E-step cannot be done in closed form:

$$p(\vec{\theta}_{1:D}, z_{1:D,1:N}, \vec{\beta}_{1:K} \mid w_{1:D,1:N}, \alpha, \eta) = \frac{p(\vec{\theta}_{1:D}, \vec{z}_{1:D}, \vec{\beta}_{1:K} \mid \vec{w}_{1:D}, \alpha, \eta)}{\int_{\vec{\beta}_{1:K}} \int_{\vec{\theta}_{1:D}} \sum_{\vec{z}} p(\vec{\theta}_{1:D}, \vec{z}_{1:D}, \vec{\beta}_{1:K} \mid \vec{w}_{1:D}, \alpha, \eta)}$$

(In fact, can be shown to be #P-hard to perform in the worst case.)

The variational family to approximate the posterior is commonly chosen to be a mean-field family:

$$q(\vec{\theta}_{1:D}, z_{1:D,1:N}, \vec{\beta}_{1:K}) = \prod_{k=1}^{K} q(\vec{\beta}_k \mid \vec{\lambda}_k) \prod_{d=1}^{D} \left(q(\vec{\theta}_{d\,d} \mid \vec{\gamma}_d) \prod_{n=1}^{N} q(z_{d,n} \mid \vec{\phi}_{d,n}) \right)$$

- **Probability of topic** z **given document** d: $q(\theta_d \mid \gamma_d)$
  Each document has its own Dirichlet prior γ_d
- **Probability of word** w **given topic** z: $q(\beta_z \mid \lambda_z)$
  Each topic has its own Dirichlet prior λ_z
- **Probability of topic assignment to word** $w_{d,n}$: $q(z_{d,n} \mid \varphi_{d,n})$
  Each word position *word[d][n]* has its own prior $\varphi_{d,n}$

Examples

$$q(\vec{\theta}_{1:D}, z_{1:D,1:N}, \vec{\beta}_{1:K}) = \prod_{k=1}^{K} q(\vec{\beta}_k \mid \vec{\lambda}_k) \prod_{d=1}^{D} \left(q(\vec{\theta}_{d} \mid \vec{\gamma}_d) \prod_{n=1}^{N} q(z_{d,n} \mid \vec{\phi}_{d,n}) \right)$$

- **Probability of topic** z **given document** d: $q(\theta_d \mid \gamma_d)$
  Each document has its own Dirichlet prior γ_d
- **Probability of word** w **given topic** z: $q(\beta_z \mid \lambda_z)$
  Each topic has its own Dirichlet prior λ_z
- **Probability of topic assignment to word** $w_{d,n}$: $q(z_{d,n} \mid \varphi_{d,n})$
  Each word position *word[d][n]* has its own prior $\varphi_{d,n}$

One iteration of mean field variational inference for LDA

(1) For each topic k and term v:

(8) $$\lambda_{k,v}^{(t+1)} = \eta + \sum_{d=1}^{D} \sum_{n=1}^{N} 1(w_{d,n} = v)\phi_{n,k}^{(t)}.$$

(2) For each document d:

(a) Update γ_d:

(9) $$\gamma_{d,k}^{(t+1)} = \alpha_k + \sum_{n=1}^{N} \phi_{d,n,k}^{(t)}.$$

(b) For each word n, update $\vec{\phi}_{d,n}$:

(10) $$\phi_{d,n,k}^{(t+1)} \propto \exp\left\{ \Psi(\gamma_{d,k}^{(t+1)}) + \Psi(\lambda_{k,w_n}^{(t+1)}) - \Psi(\textstyle\sum_{v=1}^{V} \lambda_{k,v}^{(t+1)}) \right\},$$

where Ψ is the digamma function, the first derivative of the $\log \Gamma$ function.

Parameter updates:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j}.$$