Probabilistic Graphical Models & Probabilistic AI

Ben Lengerich

Lecture 13: Variational Inference

March 11, 2025

Reading: See course homepage

Logistics

- Next week:
  - HW5 due **March 18.**
  - Midterm exam March 20 **in-class**.
    - Study guide released.
- Looking ahead:
  - Project midway report due **April 11**.

Ben Lengerich © University of Wisconsin-Madison 2025

Project Proposals

- Dynamically Pruning PGMs Inferred from an LLM
- Modeling Mental and Physical Health with PGMs
- Use PGMs to Intelligently Route LLM Queries
- Predicting Wildfire Occurrence using Bayesian Networks
- Contextualized PGMs for Cancer Genomic Analysis
- Optimizing Variable Elimination in PGMs
- Discovering Macroeconomic Factors of Stock Market with BNs
- HMM-Based Offline Handwriting Recognition
- PGMs for Social Network Analysis
- Time-varying Co-occurrence networks with Graph Networks
- Integrating Multi-source Medical Data
- Simulating Stock Market Trajectories with Diffusion-based PGMs
- Hybrid PGMs and DL for Robust Deepfake Detection
- Reducing Dimensionality of Cancer Genomic PGMs

Ben Lengerich © University of Wisconsin-Madison 2025

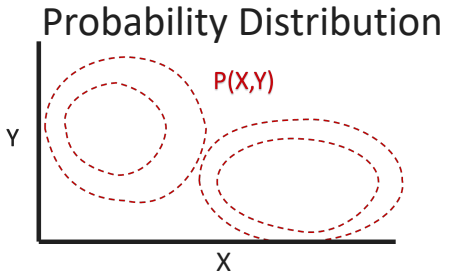
Today

- Variational Inference
  - Mean-field variational inference

Ben Lengerich © University of Wisconsin-Madison 2025

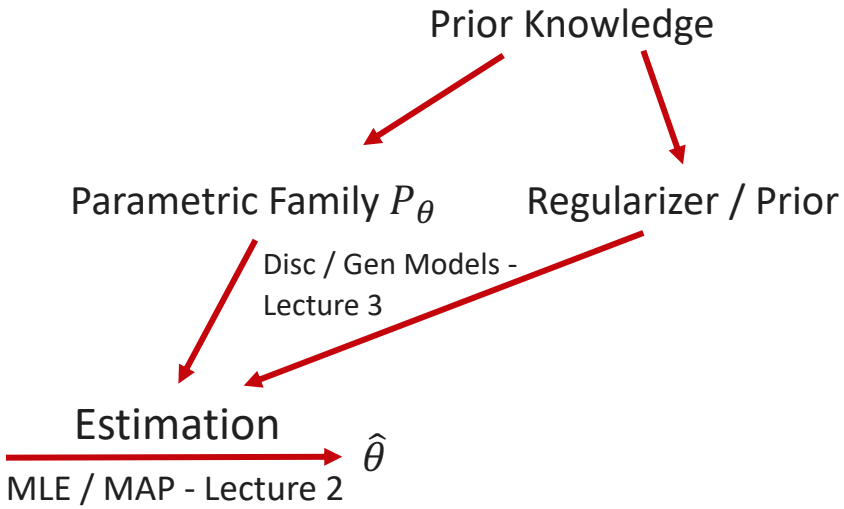
Variational Inference

A Brief Recap of our Roadmap



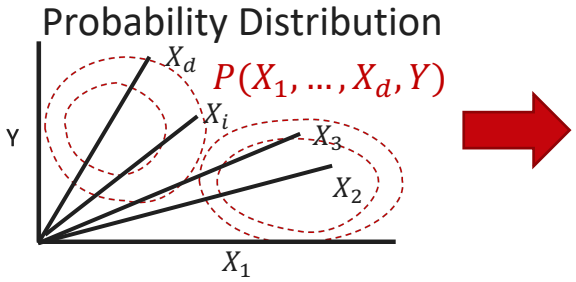
Observations

X	Y
0	0
0.1	1
0.2	1



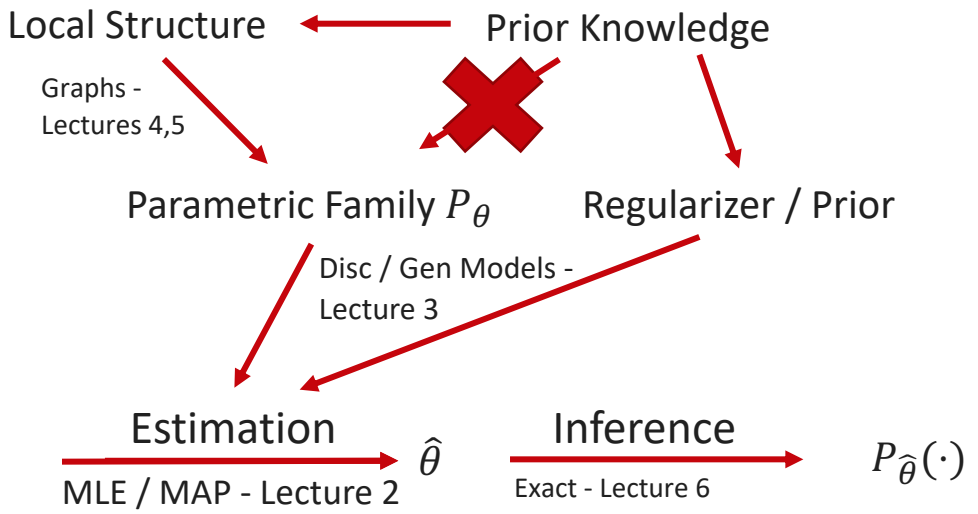
Ben Lengerich © University of Wisconsin-Madison 2025

A Brief Recap of our Roadmap



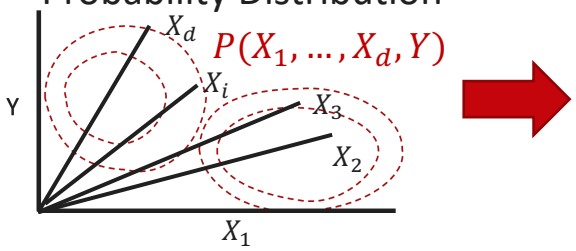
Observations

X	Y
0	0
0.1	1
0.2	1



Ben Lengerich © University of Wisconsin-Madison 2025

A Brief Recap of our Roadmap



Observations

X	Y
0	0
0.1	1
0.2	1

Structure Learning - Lectures 10, 11

Architecture Design, Lectures 16-23

Local Structure

Prior Knowledge

Graphs - Lectures 4,5

Parametric Family P_θ

Regularizer / Prior

Disc / Gen Models - Lecture 3

Estimation → $\hat{\theta}$

MLE / MAP - Lecture 2

Learning on Graphs - Lectures 7 - 9, 12-13

Inference → $P_{\hat{\theta}}(\cdot)$

Exact - Lecture 6

Approximate – Lectures 13-14

Ben Lengerich © University of Wisconsin-Madison 2025

Motivation for Variational Inference (VI)

- We can't compute the posterior for many interesting models.
- Consider the Bayesian mixture of Gaussians

1. Draw $\mu_k \sim \mathcal{N}(0, \tau^2)$ for $k = 1 \dots K$.
2. For $i = 1 \dots n$:
   (a) Draw $z_i \sim \text{Mult}(\pi)$;
   (b) Draw $x_i \sim \mathcal{N}(\mu_{z_i}, \sigma^2)$.

Ben Lengerich © University of Wisconsin-Madison 2025

Motivation for Variational Inference (VI)

- For the Bayesian mixture of Gaussians, the posterior distribution is

$$p(\mu_{1:K}, z_{1:n} \mid x_{1:n}) = \frac{\prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \mid z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \mid z_i, \mu_{1:K})}.$$

Hard to compute!

- Let's try to compute it. First, we can take advantage of the conditional independence of the z_i 's given the cluster centers,

$$p(x_{1:n}) = \int_{\mu_{1:K}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} \sum_{z_i} p(z_i) p(x_i \mid z_i, \mu_{1:K}).$$

K^n terms!

Ben Lengerich © University of Wisconsin-Madison 2025

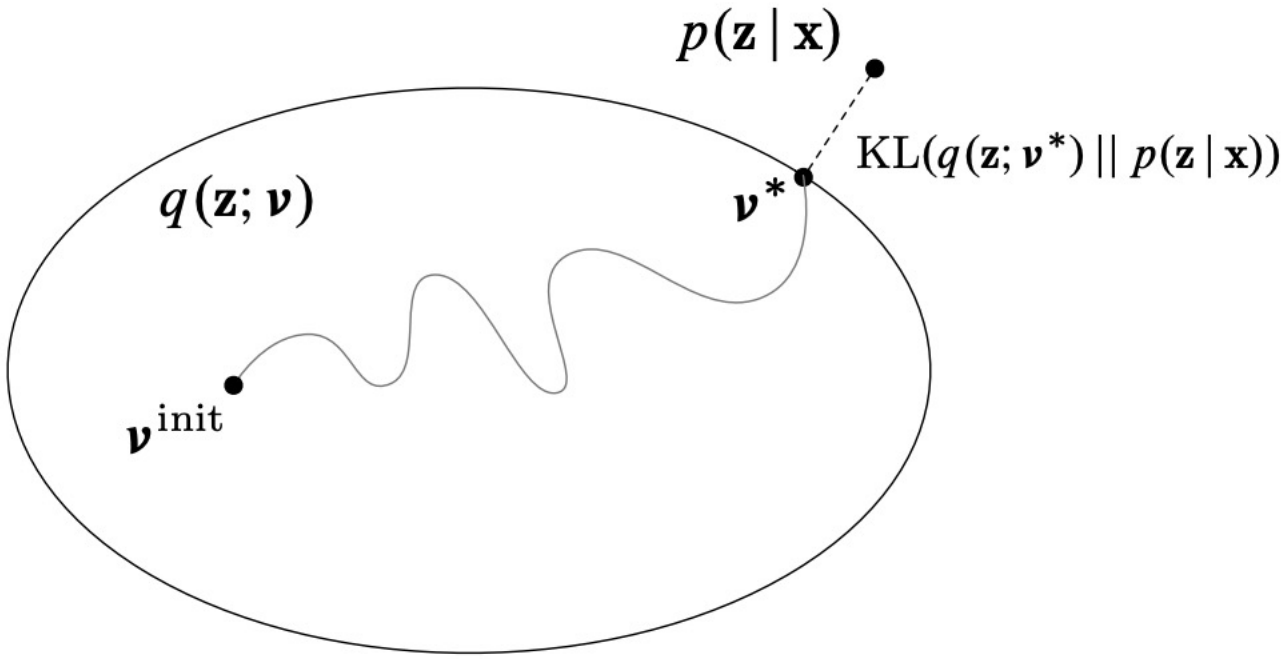
Variational Inference

- The main idea behind variational methods is to pick a family of distributions over the latent variables with its own **variational parameters**,

$$q(z_{1:m} \mid \nu).$$

- Then, find the setting of the parameters that makes q close to the posterior of interest.
- Use q with the fitted parameters as a proxy for the posterior, e.g., to form predictions about future data or to investigate the posterior distribution of the hidden variables

Variational Inference



VI solves **inference** with **optimization**.

Recall EM and the ELBO

$$\log p(x \mid \theta) = E_{z\sim q}[\log p(x, z \mid \theta)] + H(q) + KL(q(z \mid x) \,||\, p(z \mid x, \theta))$$

EM: Let $q_t(z \mid x) = p(z \mid x, \theta_t)$.
Max $p(x \mid \theta)$ by iterating:

$$Q(\theta', \theta_t) = E_{z\sim p(z|\theta_t)}[\log p(x, z \mid \theta')]$$
$$\theta_{t+1} = \text{argmax}_{\theta'} Q(\theta', \theta_t)$$

Variational Inference: Let $q(z \mid x)$ be some family that's easier to optimize.

$$\log p(x \mid \theta) \geq \underbrace{E_{z\sim q}[\log p(x, z \mid \theta)] + H(q)}$$

"ELBO": Evidence Lower Bound

equivalently,

$$\text{ELBO} = \log p(x \mid \theta) - KL(q(z \mid x)||p(z, x\theta))$$

Can we optimize q to maximize the ELBO?

Variational Inference

$$\log p(x \mid \theta) \geq \underbrace{E_{z\sim q}[\log p(x, z \mid \theta)] + H(q)}$$

"ELBO": Evidence Lower Bound

- We choose a family of variational distributions (i.e., a parameterization of a distribution of the latent variables) such that the expectations are computable.
- Then, we maximize the ELBO to find the parameters that gives as tight a bound as possible on the marginal probability of x.

Is ELBO convex? How to pick q?

Mean-Field VI

Mean-field VI

- In mean field variational inference, we assume that the variational family factorizes

$$q(z_1, \ldots, z_m) = \prod_{j=1}^{m} q(z_j).$$

- Each variable is independent. (We are suppressing the parameters νj .)
- This is more general that it initially appears—the hidden variables can be grouped and the distribution of each group factorizes.

Does this family include the true posterior for the Guassian mixture model?

Optimizing the ELBO for a factorized distribution

- We will use coordinate ascent inference, interatively optimizing each variational distribution holding the others fixed.
- First, decompose the joint

$$p(z_{1:m}, x_{1:n}) = p(x_{1:n}) \prod_{j=1}^{m} p(z_j \mid z_{1:(j-1)}, x_{1:n})$$

- Notice that the z variables can occur in any order in this chain. The indexing from 1 to m is arbitrary. (This will be important later.)
- Second, decompose the entropy of the variational distribution,

$$\mathrm{E}[\log q(z_{1:m})] = \sum_{j=1}^{m} \mathrm{E}_j[\log q(z_j)],$$

- where Ej denotes an expectation with respect to q(z_j).

Optimizing the ELBO for a factorized distribution

- This makes the ELBO:

$$\mathcal{L} = \log p(x_{1:n}) + \sum_{j=1}^{m} \mathrm{E}[\log p(z_j \mid z_{1:(j-1)}, x_{1:n})] - \mathrm{E}_j[\log q(z_j)].$$

- Consider the ELBO as a function of q(z).
- This leads to the objective function

$$\mathcal{L} = \mathrm{E}[\log p(z_k \mid z_{-k}, x)] - \mathrm{E}_j[\log q(z_k)] + \text{const.}$$

Optimizing the ELBO for a factorized distribution

$$\mathcal{L} = \mathrm{E}[\log p(z_k \mid z_{-k}, x)] - \mathrm{E}_j[\log q(z_k)] + \text{const.}$$

- As a function of q_k:

$$\mathcal{L}_k = \int q(z_k)\mathrm{E}_{-k}[\log p(z_k \mid z_{-k}, x)]dz_k - \int q(z_k)\log q(z_k)dz_k.$$

- Optimize:

$$\frac{d\mathcal{L}_j}{dq(z_k)} = \mathrm{E}_{-k}[\log p(z_k \mid z_{-k}, x)] - \log q(z_k) - 1 = 0$$

$$q^*(z_k) \propto \exp\{\mathrm{E}_{-k}[\log p(z_k, Z_{-k}, x)]\}$$

Optimizing the ELBO for a factorized distribution

- Bottom line:
  - The coordinate ascent algorithm is to iteratively update each q(zk).
  - The ELBO converges to a local maximum.
  - Use the resulting q is as a proxy for the true posterior.

Example: Multinomial conditionals

$$q^*(z_k) \propto \exp\{\mathrm{E}_{-k}[\log p(z_k, Z_{-k}, x)]\}$$

- Suppose the conditional is multinomial

$$p(z_j \mid z_{-j}, x_{1:n}) := \pi(z_{-j}, x_{1:n})$$

- Then the optimal $q(z_j)$ is also a multinomial,

$$q^*(z_j) \propto \exp\{\mathrm{E}[\log \pi(z_{-j}, x)]\}$$

Example: Exponential Family Conditionals

$$q^*(z_k) \propto \exp\{\mathrm{E}_{-k}[\log p(z_k, Z_{-k}, x)]\}$$

Suppose each conditional is in the exponential family

$$p(z_j \mid z_{-j}, x) = h(z_j) \exp\{\eta(z_{-j}, x)^\top t(z_j) - a(\eta(z_{-j}, x))\}$$

Then

$$q^*(z_j) \propto h(z_j) \exp\{\mathrm{E}[\eta(z_{-j}, x)]^\top t(z_j)\}$$

and the normalizing constant is $a(\mathrm{E}[\eta(z_{-j}, x)])$.

Optimal q is in the same family as the conditional.

Example: Exponential Family Conditionals

Coordinate ascent algorithm

- Give each hidden variable a variational parameter ν_j, and put each one in the same exponential family as its model conditional,

$$q(z_{1:m} \mid \nu) = \prod_{j=1}^{m} q(z_j \mid \nu_j)$$

The coordinate ascent algorithm iteratively sets each natural variational parameter ν_j equal to the expectation of the natural conditional parameter for variable z_j given all the other variables and the observations,

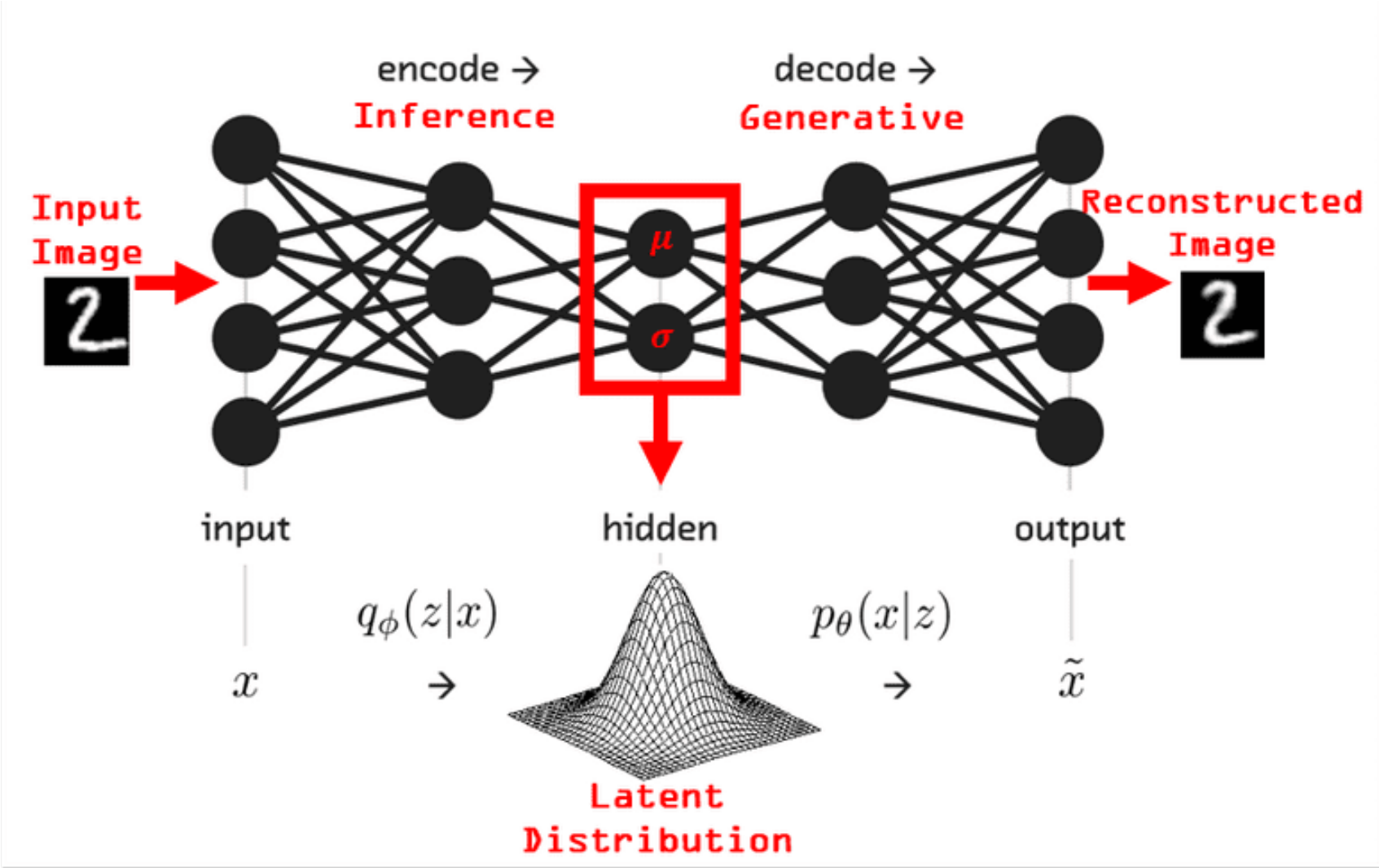
$$\nu_j^* = \mathrm{E}[\eta(z_{-j}, x)].$$

Example: Bayesian Mixture of Gaussians

$$q^*(z_i = k) \propto \exp\{\log \pi_k + x_i \mathrm{E}[\mu_k] - \mathrm{E}[\mu_k^2]/2\}.$$

$$\begin{aligned} \mathrm{E}[\mu_k] &= \frac{\mu_0/\sigma_0^2 + \sum_{i=1}^n \mathrm{E}[z_i^k] x_i}{1/\sigma_0^2 + \sum_{i=1}^n \mathrm{E}[z_i^k]} \\ \mathrm{Var}(\mu_k) &= 1/(1/\sigma_0^2 + \textstyle\sum_{i=1}^n \mathrm{E}[z_i^k]). \end{aligned}$$

Example: Variational Autoencoder



https://theaisummer.com/Autoencoder/

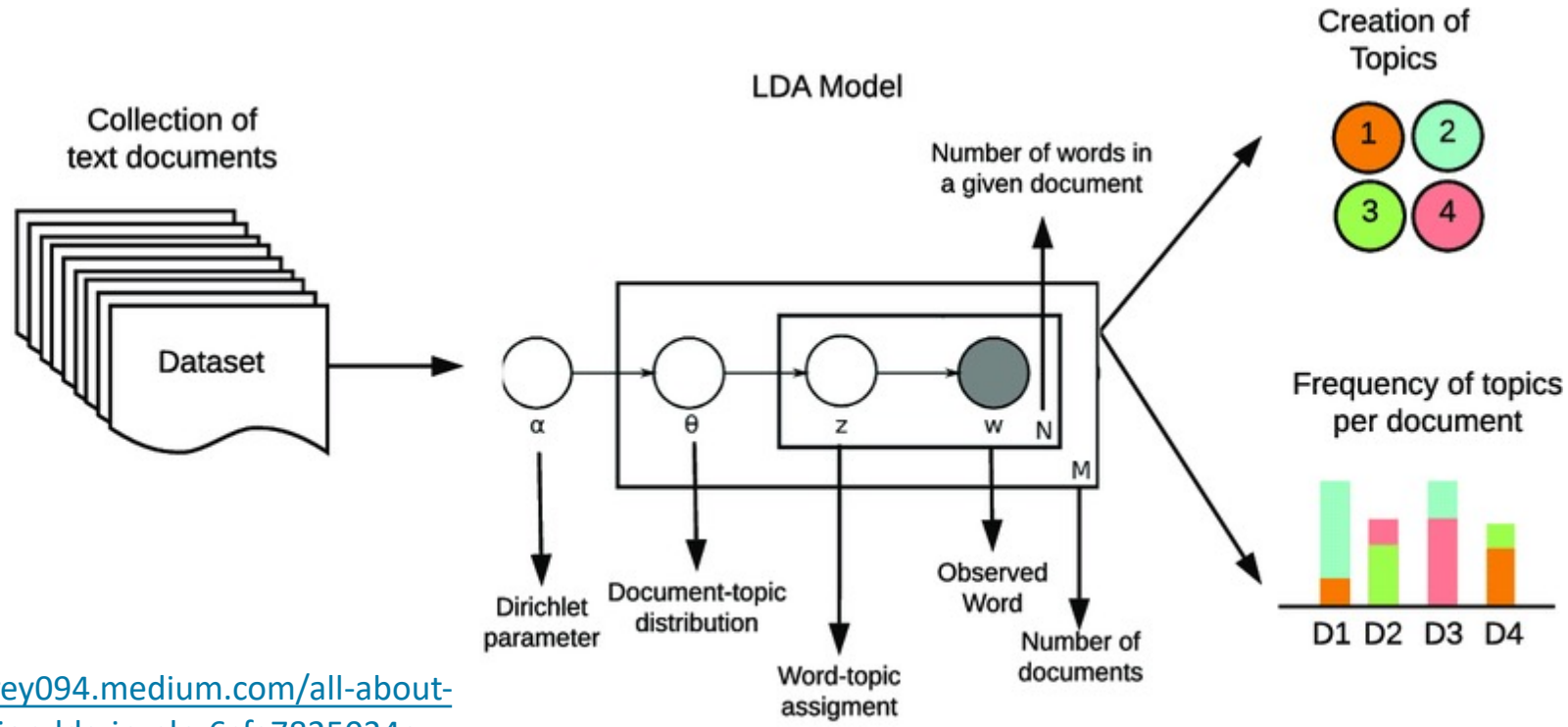
Example: Latent Dirichlet Allocation

[PDF] Latent dirichlet allocation

DM Blei, AY Ng, MI Jordan - Journal of machine Learning research, 2003 - jmlr.org

We describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in …

☆ Save Cite Cited by 56217 Related articles »



https://mohamedbakrey094.medium.com/all-about-latent-dirichlet-allocation-lda-in-nlp-6cfa7825034e

Questions?