Inference & Representation

DS-GA 1005, Fall 2023

Lecture 1, Course Overview

Yoav Wald, 09/06/2023

Logistics

- Office Hours, Wed 12:00-13:00, 60 5th Ave Room 600
- TA: Aiqing Li, Office Hours TBA
- Web resources
  - Notion: Lecture notes, schedule, syllabus
  - Brightspace: HW upload, grades
  - Ed Discussion: Q&A
- Grading: 30% Homework (4 assignments), 30% mid-term, 40% final project

Probabilistic Models: Toy Example

- Consider a medical diagnosis system
- Each patient fills out a questionnaire with symptoms:

  coughing 😮‍💨 , chest pains 🪎 , muscle aches 💪 , fever 🤒, …

  X_1 X_2 X_3 X_4 $X_N \in \{0,1\}^N$

- Then there are possible conditions:

  asthma, seasonal flu, covid-19, …

  X_{N+1} X_{N+2} X_{N+3} X_{N+M} $\in \{0,1\}^M$

- Denote $d = M + N$ and define $\mathcal{F} = \left\{ f : \mathbb{H}_d \to [0,1] \mid \Sigma_{\mathbf{x} \in \mathbb{H}_d} f(\mathbf{x}) = 1 \right\}$

Probabilistic Inference: Toy Example

- $\mathcal{F} = \left\{ f : \mathbb{H}_d \rightarrow [0,1] \mid \Sigma_{\mathbf{x} \in \mathbb{H}_d} f(\mathbf{x}) = 1 \right\}$ is the set of ***all*** possible probability distributions over $\mathbb{H}_d = \{0,1\}^d$

- *Diagnosis example*: $f(\mathbf{x})$ is the probability of observing a patient with certain symptoms and conditions

- We will explore several questions about learning and using such models

Statistical Questions

- How to define useful models in $\mathscr{F}$?
- How to estimate (learn) models from data?
  - Given a sample $\{\mathbf{x}_i\}_{i=1}^m$, how do we estimate $P(X_1, \ldots, X_d)$?
- What can we say about estimation error of a model as a function of the sample size? (sample complexity)
- What can we do with these models once they are trained? (inference)
  - E.g. can we retrieve $P(\text{Asthma} = 1 \mid X_1 = x_1, \ldots, X_N = x_n)$?

Computational Questions

- Algorithms for estimating model parameters (learning)
  - Given a sample $\{\mathbf{x}_i\}_{i=1}^{m}$, what are computationally efficient algorithms for estimating $P(X_1, \ldots, X_d)$?
- Algorithms for querying probabilistic models (inference)
  - How to compute $P(\text{Asthma} = 1 \mid X_1 = x_1, \ldots, X_N = x_n)$?

The Curse of Dimensionality

- The space $\mathscr{F}$ of all possible distributions has $2^d - 1$ free parameters
- *Statistical question*: How can we estimate the correct distribution $P \in \mathscr{F}$ from a dataset $\{\mathbf{x}_i\}_{i=1}^m$ sampled i.i.d from P?
  - **Example:** consider a “brute-force” approach of estimating all parameters of a multinomial model over $\mathbb{H}_d$
  - ***Multinomial model:*** denoting $K = 2^d$, we have to estimate the parameters $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_K]$, which must satisfy $\theta_i \in [0,1] \ \forall j \in [K]$ and $\Sigma_{i=1}^K \theta_i = 1$

Estimating Multinomial Model with Maximum Likelihood

- *Input:* dataset $D = \{\mathbf{x}_i\}_{i=1}^m$ sampled i.i.d from P
- *Output:* model $f_{\hat{\boldsymbol{\theta}}} \in \mathscr{F}$ where $f_{\hat{\boldsymbol{\theta}}}(\mathbf{y}_j) := \hat{\theta}_j \approx \theta_j^* := P(\mathbf{x} = \mathbf{y}_j)$
  - We enumerate $\mathbb{H}_d = \{\mathbf{y}_1, \ldots, \mathbf{y}_K\}$ (where $K = 2^d$)
- Estimate using *MLE*: $\hat{\theta}_j = m^{-1}\Sigma_{i=1}^m \ \mathbf{1}[\mathbf{x}_i = \mathbf{y}_j]$
- *Question:* How large does m need to be?

To (roughly) quantify this we need a target error $\mathbb{E}_{D\sim P^m}\dfrac{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2}{\|\boldsymbol{\theta}^*\|^2} \leq \varepsilon$

The Curse of Dimensionality

$$\hat{\theta}_j = m^{-1} \sum_{i=1}^{m} \mathbf{1}[\mathbf{x}_i = \mathbf{y}_j]$$

Define RVs $z_j = \mathbf{1}[\mathbf{x} = \mathbf{y}_j] \quad \forall \boldsymbol{y}_j \in \mathbb{H}_d$

$z_j \sim \mathrm{Ber}(\theta_j^*) \Rightarrow \mathbb{E}_P[z_j] = \theta_j^*, \mathrm{Var}(z_j) = \theta_j^*(1 - \theta_j^*)$

Let us calculate the expected distance:

$$\mathbb{E}_{D \sim P^m}[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2] = \mathbb{E}_{D \sim P^m} \sum_j (\hat{\theta}_j - \theta_j^*)^2 = \sum_j \mathbb{E}(\hat{\theta}_j - \theta_j^*)^2 = \sum_j \mathrm{Var}(\hat{\theta}_j) = \sum_j m^{-1} \theta_j^* (1 - \theta_j^*)$$

$\mathbb{E}[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}^*$

Then expected error is:

$$\mathbb{E}_{D \sim P^m}\left[\frac{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2}{\|\boldsymbol{\theta}^*\|^2}\right] = \frac{\sum_j \theta_j^* - \|\boldsymbol{\theta}^*\|^2}{m\|\boldsymbol{\theta}^*\|^2} = m^{-1}\left(\|\boldsymbol{\theta}^*\|^{-2} - 1\right) \approx m^{-1}K = m^{-1}2^d$$

😱

$\sum_j \theta_j^* = 1$

$\|\boldsymbol{\theta}^*\| \approx K^{-1/2}$ for “generic” $\boldsymbol{\theta}^*$

The Curse of Dimensionality

$$\mathbb{E}_{D \sim P^m}\left[\frac{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2}{\|\boldsymbol{\theta}^*\|^2}\right] \approx m^{-1}2^d$$ 😱

Conclusion:

- The multinomial model is “cursed by dimension”
- To work in high dimensions, simplifying assumptions must be made
- How can we incorporate structure into probabilistic models?

Structure in Probabilistic Models

Example: we wish to estimate $P(\text{Illness}, \text{Cough}, \text{Fatigue}) := P(I, C, F)$

- $\text{Illness} \in \{\text{Asthma}, \text{Flu}, \ldots\}, \text{Cough} \in \{0,1\}, \text{Fatigue} \in \{0,1\}$
- Chain rule: $P(I, C, F) = P(I)P(C \mid I)P(F \mid C, I)$
- ***Assumption***: if we know the type of illness, then coughing and fever are independent

$$P(I, C, F) = P(I)P(C \mid I)\,P(F \mid I)$$

- Consequence: fewer parameters to estimate

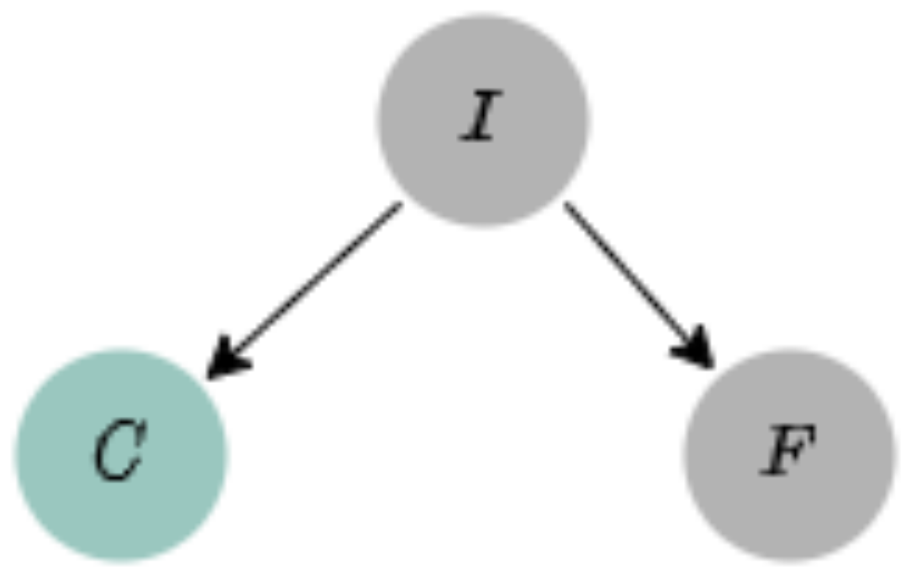
Structure in Probabilistic Models

More generally, if we have d variables $X_1, \ldots, X_d$

- Chain rule:
  $P(X_1, \ldots, X_d) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1, X_2)\ldots P(X_d \mid X_1, \ldots X_{d-1})$
- Still suffers from the curse of dimensionality
- But what if we can drop some of the conditioning variables?
  - Corresponds to conditional independence relationships $X_i \perp\!\!\!\perp X_j \mid X_k$
- ***Example***: Markov process $X_1, \ldots, X_T$, such that $X_{t+1} \perp\!\!\!\perp X_1, \ldots X_{t-1} \mid X_t$.
  Future is independent of past, given the present

Probabilistic Graphical Models

- Conditional independence relations can be encoded by graphs
- Random Variable ~ Vertex in graph
- Edges ~ factorization of joint distribution



- Probabilistic graphical models formalize this and leverage graph-theoretic algorithms for the purpose on inference and learning

Birdseye View of Course

- We will study 3 major families of distributions
  - Mixture models / Latent variable models
  - Gibbs distributions
  - Implicit models

Mixture Models / Latent Variables Models

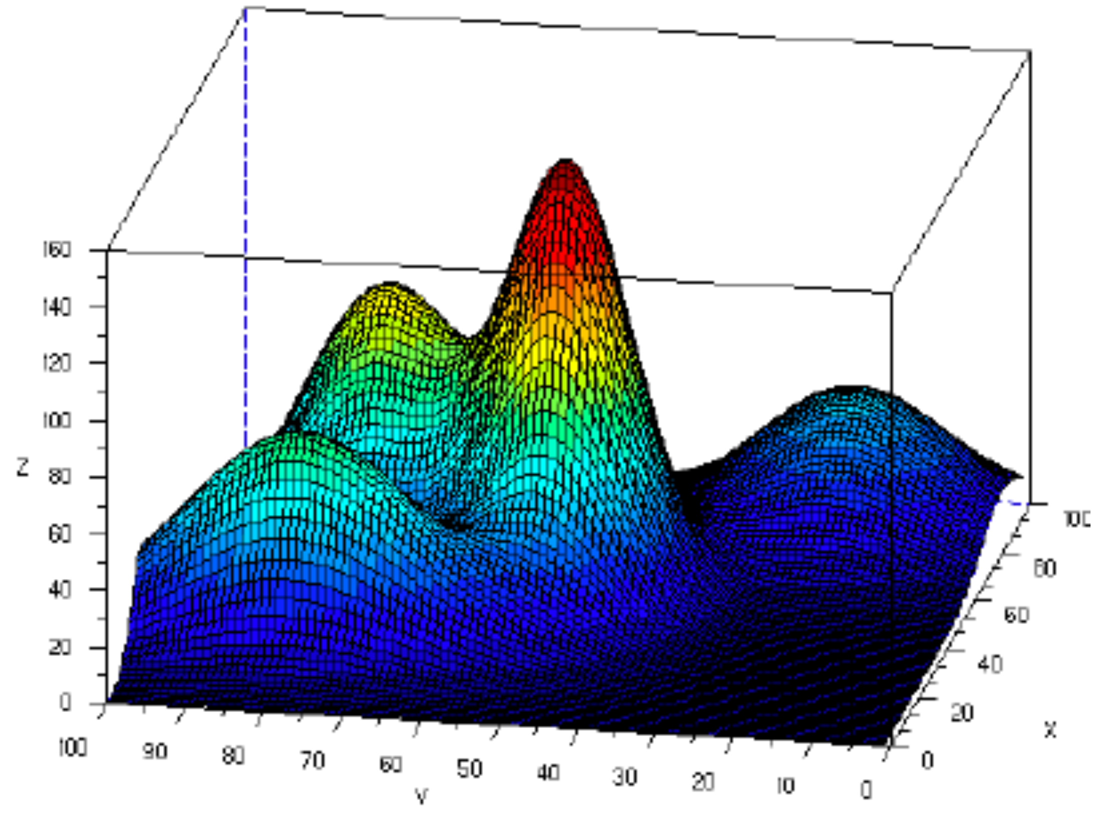
- Mixture models / Latent variable models

$$\mathbf{x} \in \mathbb{R}^d \qquad\qquad p(\mathbf{x}) = \int_{\Omega} p_{\boldsymbol{\theta}}(\mathbf{x} \mid z) p_{\boldsymbol{\beta}}(z) dz$$

- <u>*Linear*</u> combination of templates $p(\mathbf{x} \mid z)$

- *Example*: Gaussian Mixture Model,

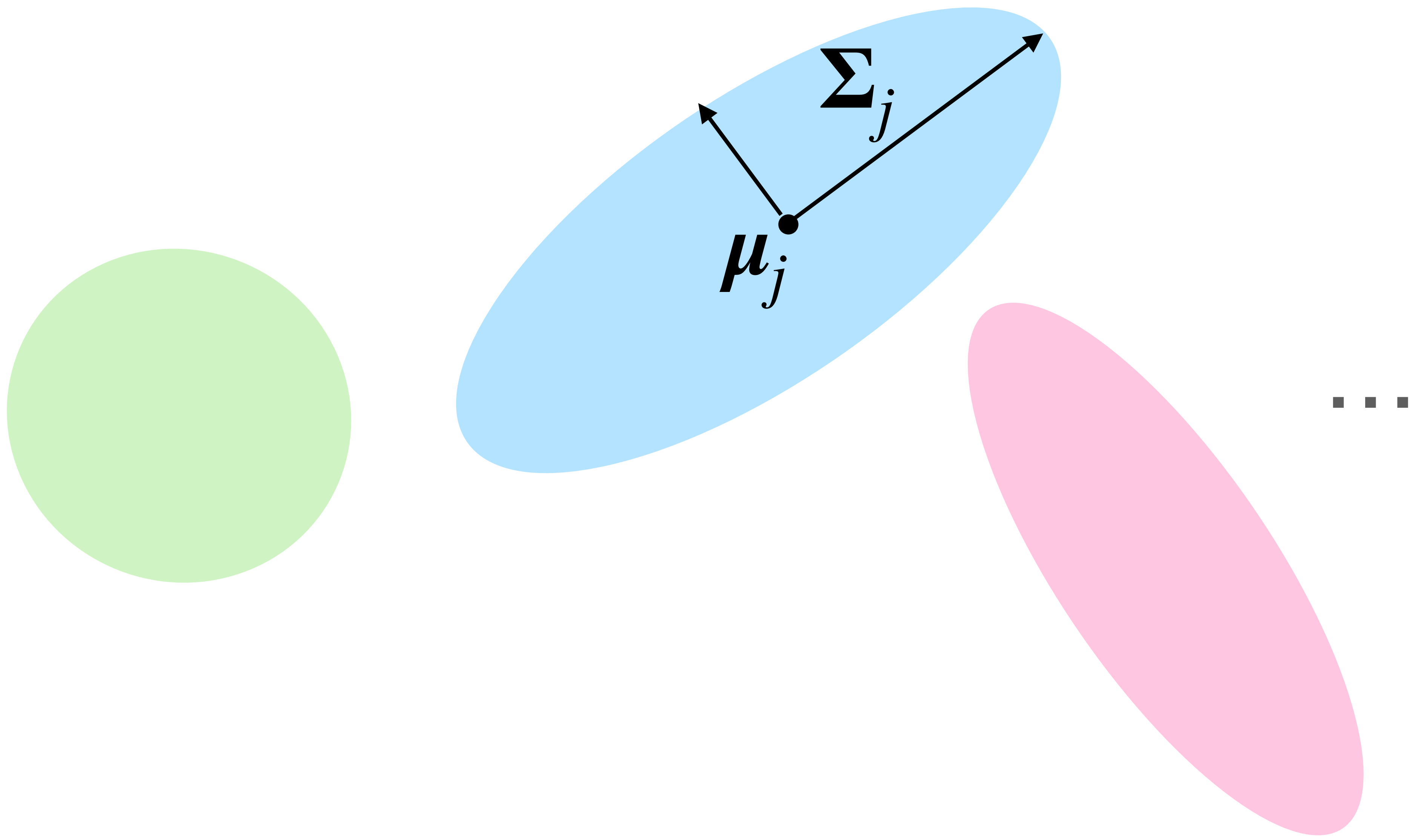
$$p(\mathbf{x}) = \sum_{j=1}^{k} \alpha_j p(\mathbf{x} \mid z = j) \text{ and } p(\mathbf{x} \mid z = j) = \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$



Mixture Models / Latent Variables Models

- *Example*: Gaussian Mixture Model,

$$p(\mathbf{x}) = \sum_{j=1}^{k} \alpha_j p(\mathbf{x} \mid z = j) \text{ and } p(\mathbf{x} \mid z = j) = \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$



- Parameters of the model: $\left\{\alpha_1, \ldots, \alpha_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_{\mathbf{1}}, \ldots, \boldsymbol{\Sigma}_k\right\}$

Gibbs Distributions

- Gibbs or "Energy Based" models are distributions of the form

$$p(\mathbf{x}) = \frac{1}{Z}\exp\{-H(\mathbf{x})\}$$

- $H : \mathbb{R}^d \to \mathbb{R}$ is the *energy function*,

  Z is the *partition function* of p, $Z = \displaystyle\int \exp\{H(\mathbf{x})\}d\mathbf{x}$

- Gibbs/Boltzmann distributions originate in statistical physics

  - H is modeling interactions between variables (cores. to undirected graphs)

  - Attractive statistical properties (exponential families), but computationally challenging to use (e.g. computing Z).

Implicit Models

- Implicit, or transport-based models specify the ***transformation*** from some variable $\mathbf{z} \in \Omega$, of known distribution, to the variable of interest $\mathbf{x} \in \mathscr{X}$
  - e.g. $\mathbf{z} \sim \mathscr{N}(0, \mathbf{I}_{d_z})$ and $\mathbf{x} \in \mathbb{R}^d$ are images
  - $\mathbf{z} \sim \mu_0; \mathbf{x} = \phi(\mathbf{z}), \phi : \Omega \to \mathscr{X}$
  - Examples: in normalizing flows ϕ is a diffeomorphism, but ϕ can also be unstructured (e.g. GANs use neural nets for ϕ)

Diffusion Models

- Mix of Gibbs dist. and implicit models where ϕ is a “diffusion process”

[Signh 2023]

Birdseye View of Course

Main questions we will study

- Computational aspects of learning and inference in these families of models
    - Emphasis: *high-dimensional regime*
- Computational and statistical tradeoffs
    - Focus on foundations and theory!
- Two basic problems: *learning* and *inference*

Learning Probabilistic Models

- Learning means fitting the parameters of a probabilistic model to data
- *Example*: When learning a Gaussian Mixture Model, a model $p_{\boldsymbol{\theta}}$ is specified by
  - Means $\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_k} \in \mathbb{R}^d$
  - Covariance matrices $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K \in \mathbb{R}^{d \times d}$ (symmetric, positive semi-definite)
  - Mixture weights $\boldsymbol{\alpha} \in \Delta^K$ $(\alpha_j \geq 0, \Sigma_{j=1}^K \alpha_j = 1)$

$$\boldsymbol{\theta} = \left(\boldsymbol{\alpha}, \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=1}^K\right) \text{ and } p_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{j=1}^K \alpha_j q(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

- Given a sample $\{\boldsymbol{x}_i\}_{i=1}^m$, how do we find $\boldsymbol{\theta}$ that best explains the data?

Learning Probabilistic Models

- Learning means fitting the parameters of a probabilistic model to data

- *Example*: When learning a Gaussian Mixture Model, a model $p_{\boldsymbol{\theta}}$ is specified by

$$\boldsymbol{\theta} = \left(\boldsymbol{\alpha}, \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=1}^{K}\right) \text{ and } p(\mathbf{x}; \boldsymbol{\theta}) = \sum_{j=1}^{K} \alpha_j q(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

- Given a sample $\{\boldsymbol{x}_i\}_{i=1}^{m}$, how do we find $\boldsymbol{\theta}$ that best explains the data?

    - “Classic” solution is to use Maximum Likelihood Estimation (MLE):

$$\max_{\boldsymbol{\theta}} \frac{1}{m} \sum_{i=1}^{m} \log p(\boldsymbol{x}_i; \boldsymbol{\theta})$$

Probabilistic Inference

- Given a model of the joint distribution $P(X_1, \ldots, X_d)$, answer a query about a subset of the variables
  - $P(X_1, X_2 \mid X_3 = x_3, \ldots, X_d = x_d)$
  - $\max_{x_1, x_2} P(X_1 = x_1, X_2 = x_2 \mid X_3 = x_3, \ldots, X_d = x_d)$
- *Example*: Recall our $P(\text{Illness}, \text{Cough}, \text{Fever}) := P(I, C, F)$ model. Assume **we know** the values of all joint probabilities $P(I = k, C = c, F = f)$, for $k \in [K], c \in \{0,1\}, f \in \{0,1\}$.

Probabilistic Inference

- Given a model of the joint distribution $P(X_1, \ldots, X_d)$, answer a query about a subset of the variables

- *Example*: Recall our $P(\text{Illness}, \text{Cough}, \text{Fatigue}) := P(I, C, F)$s model. Assume **we know** the values of all joint probabilities $P(I = k, C = c, F = f)$, for $k \in [K], c \in \{0,1\}, f \in \{0,1\}$.

summing 2 numbers

  - How do we calculate $P(I \mid C = 1)$?

Easy: $P(I \mid C = 1) = \dfrac{P(I, C = 1)}{P(C = 1)} = \dfrac{\sum_{f \in \{0,1\}} P(I, C = 1, f = f)}{\sum_{f,c \in \{0,1\}^2} P(I, C = c, f = f)}$

summing 4 numbers

  - But what if we have d other symptoms and not only F?

Probabilistic Inference

- *Example*: Assume **we know** the values of all joint probabilities $P(I = k, C = c, S = \mathbf{s})$, for $k \in [K], c \in \{0,1\}, \mathbf{s} \in \{0,1\}^d$.
  - How do we calculate $P(I \mid C = 1)$?
  - $S = \{$Fever, Muscle Pain, $\ldots\}$

summing 2^d numbers

$$P(I \mid C = 1) = \frac{P(I, C = 1)}{P(C = 1)} = \frac{\sum_{\mathbf{s} \in \{0,1\}^d} P(I, C = 1, S = \mathbf{s})}{\sum_{\mathbf{s} \in \{0,1\}^d, c \in \{0,1\}} P(I, C = c, S = \mathbf{s})}$$ 😱

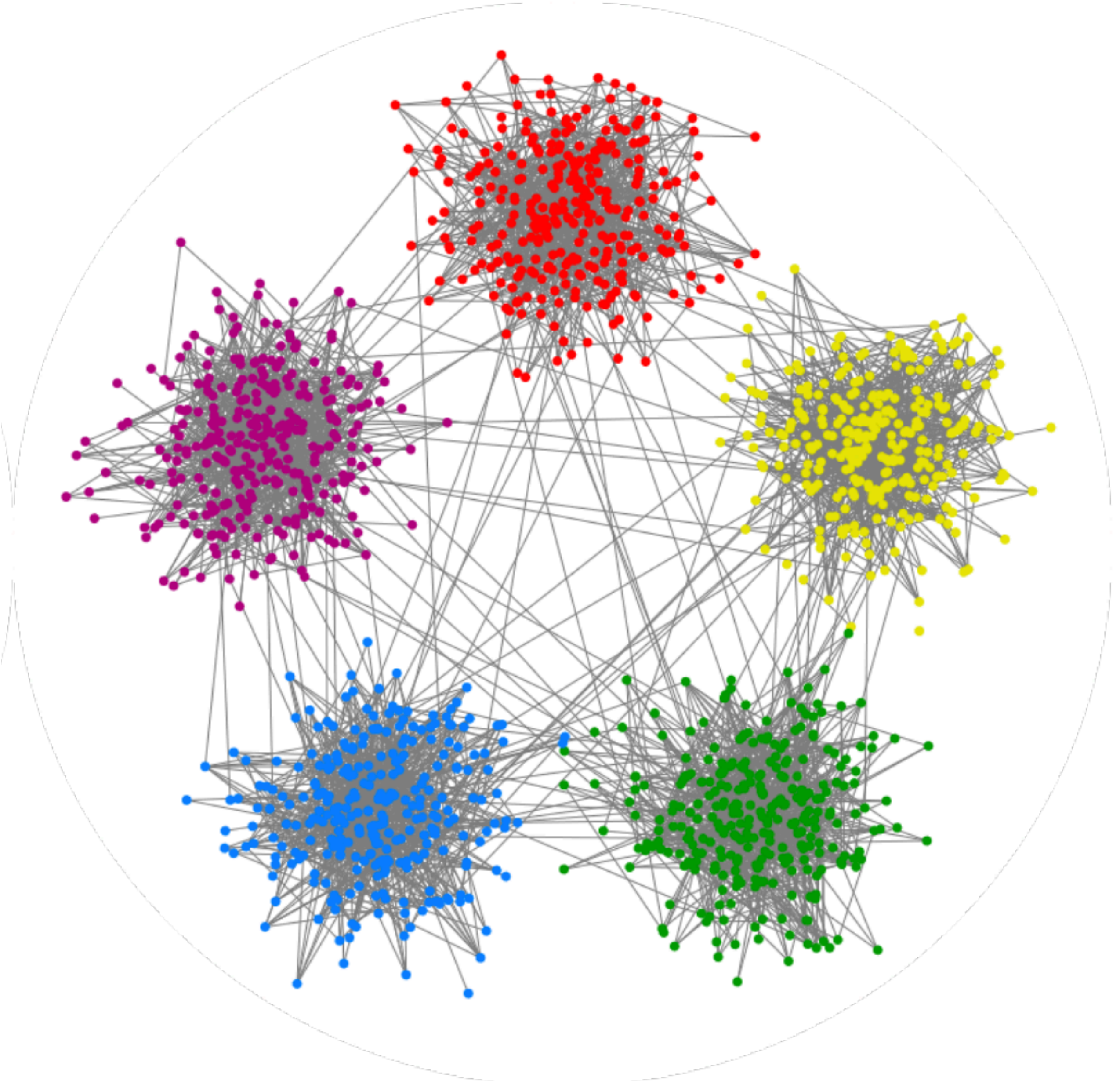
  - When and how can we avoid the exponential blowup?

summing 2^{d+1} numbers

Probabilistic Inference

- *Example 2*: Stochastic block models
  - Probabilistic models that generate random graphs
- $Z_1, Z_2, \ldots, Z_d$ variables for each node in the graph, stands for membership in a community (a.k.a "block")
- Edges, $\boldsymbol{E} = \{E_{ij}\}_{(i,j)\in[d]\times[d]}$, also binary variables such that

$$P(E_{ij} = 1) = \begin{cases} p & Z_i = Z_j \\ q & Z_i \neq Z_j \end{cases}$$



[Abbe 2018]

Probabilistic Inference

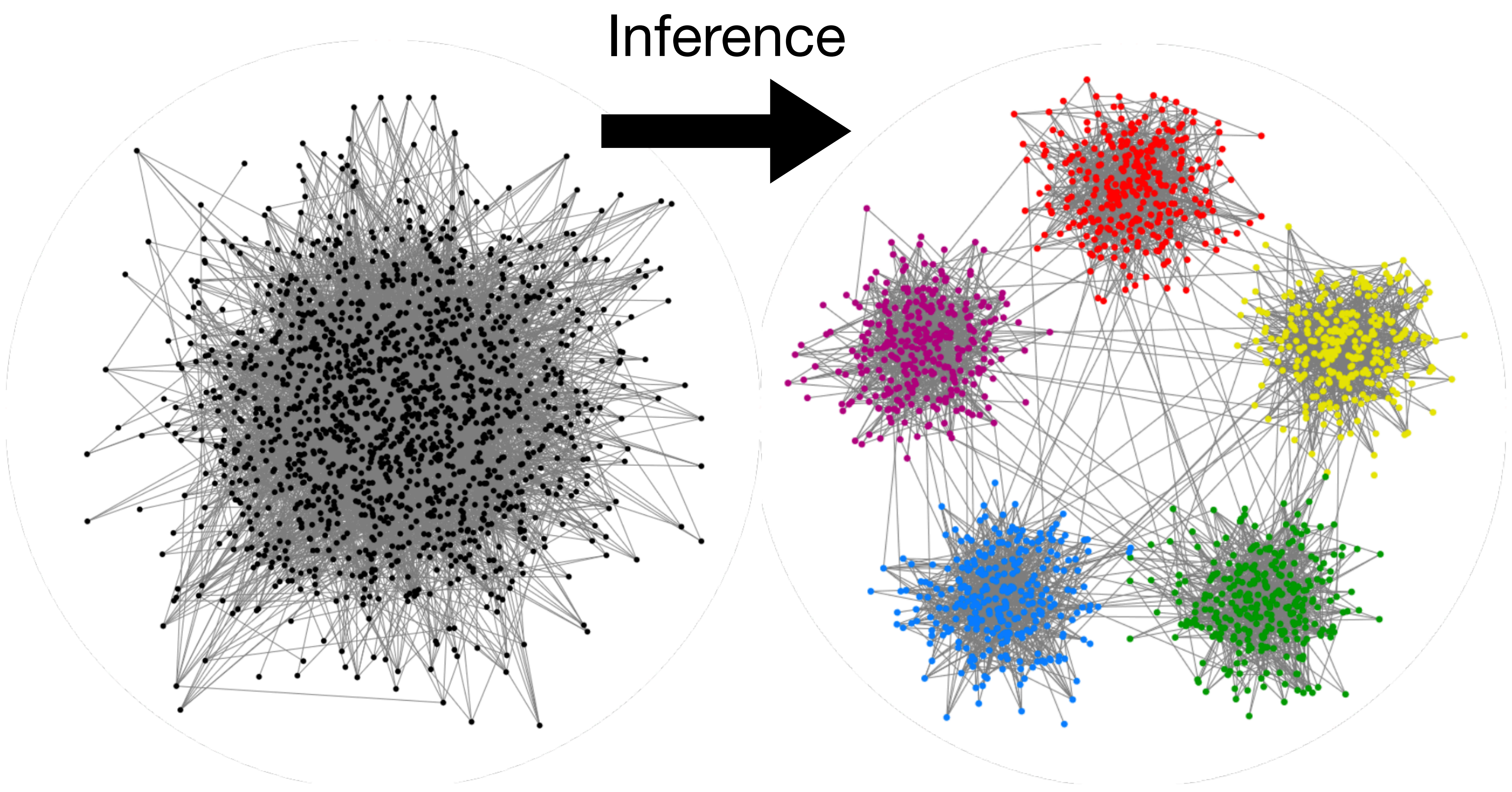
- $Z_1, Z_2, \ldots, Z_d$ variables for each node in the graph, stands for membership in a community (a.k.a "block")

- Edges, $\boldsymbol{E} = \{E_{ij}\}_{(i,j)\in[d]\times[d]}$, also binary variables such that

$$P(E_{ij} = 1) = \begin{cases} p & Z_i = Z_j \\ q & Z_i \neq Z_j \end{cases}$$

- Inference problem: calculate

$$P(Z_i \mid \boldsymbol{E} = \mathbf{e}) \quad \forall i \in [d]$$



[Abbe 2018]

Probabilistic Inference

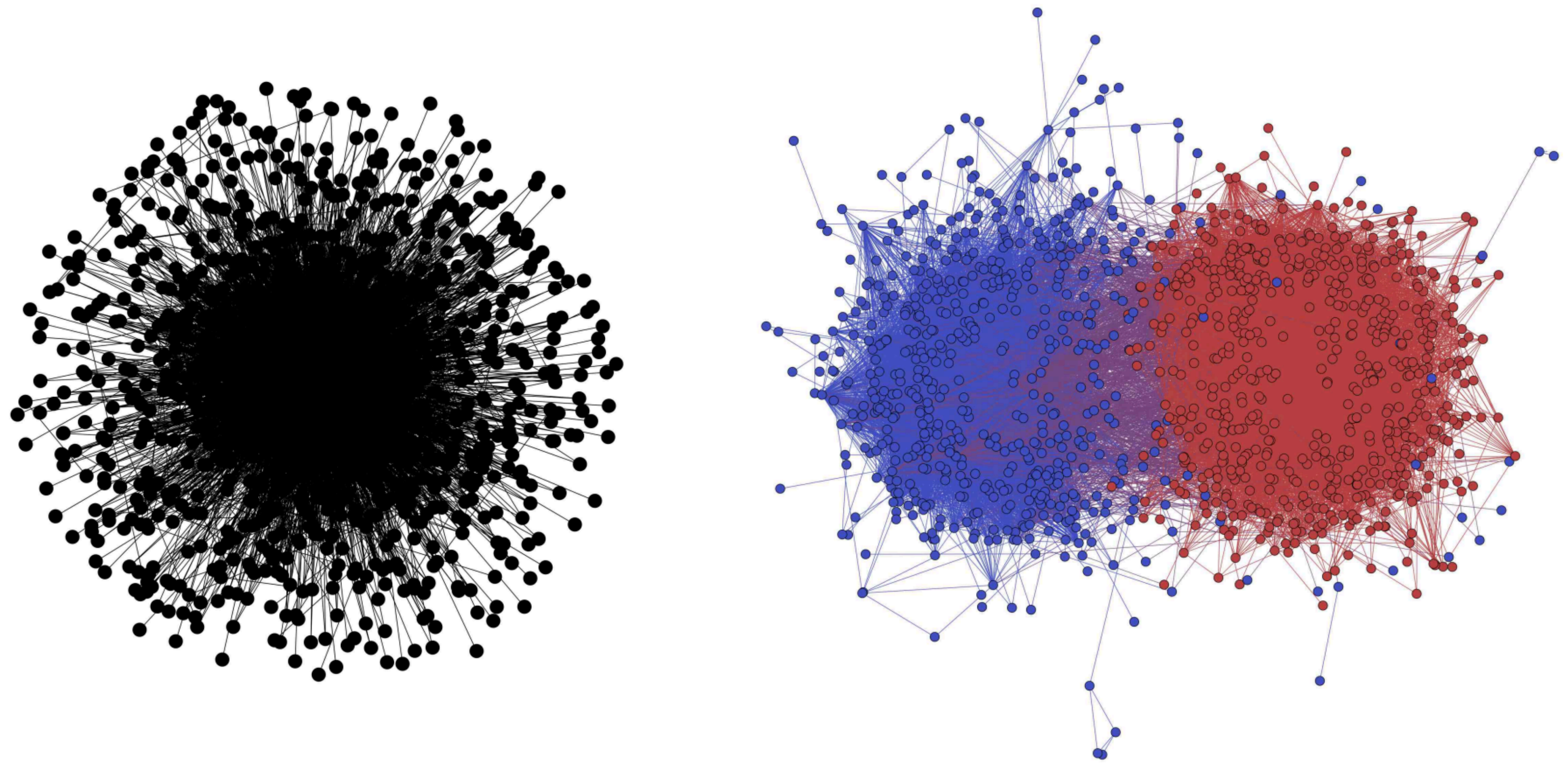
- $Z_1, Z_2, \ldots, Z_d$ variables for each node in the graph, stands for membership in a community (a.k.a "block")

- Edges, $\boldsymbol{E} = \{E_{ij}\}_{(i,j)\in[d]\times[d]}$, also binary variables such that

$$P(E_{ij} = 1) = \begin{cases} p & Z_i = Z_j \\ q & Z_i \neq Z_j \end{cases}$$

- Inference problem: calculate

$$P(Z_i \mid \boldsymbol{E} = \mathbf{e}) \quad \forall i \in [d]$$



[Abbe 2018]

Solution Approaches for Learning and Inference

- **Variational Inference:** Relies on connection between learning/inference and optimization
- **Sampling-Based Methods:** Relies on the *concentration* of empirical estimates (Law of Large Numbers, Central Limit Theorem etc.)
  - Elementary example: assume we would like to compute $\boldsymbol{\theta} = \mathbb{E}_{x \sim q}\left[f(\mathbf{x})\right]$
  - *Monte-Carlo estimator*: given data $\left\{\mathbf{x}_i\right\}_{i=1}^{L}$ sampled i.i.d from q, estimate

$$\hat{\boldsymbol{\theta}} = \frac{1}{L}\sum_{l=1}^{L} f(\boldsymbol{x}_l)$$

Course Plan

- Gaussian Estimation (lecture 2)
- Probabilistic Graphical Models (lectures 3-4)
- Variational Inference + Intro to Causality (lectures 5-7)
- Midterm
- Sampling-based Methods (lectures 8-10)
- Optimal Transport and Implicit Models (lecture 11)
- Score-based Models, Diffusion Models (lecture 12)