Mean Field Variational Inference: Computational and Statistical Guarantees

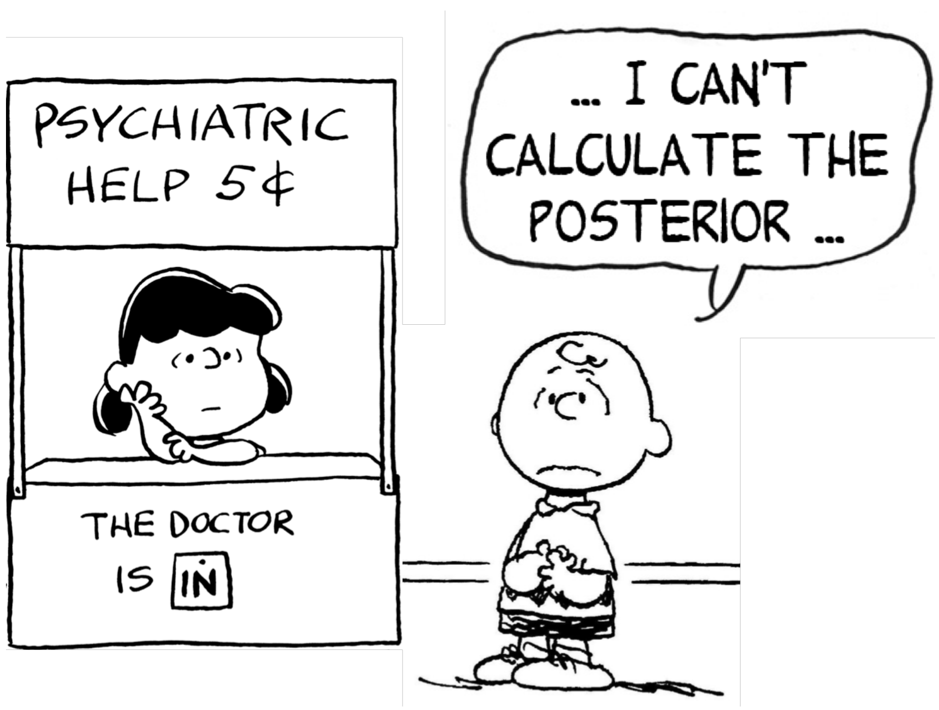
Anderson Ye Zhang

Department of Statistics
University of Pennsylvania

Bayesian Inference

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}$$

Challenge: Often computationally intractable



Bayesian Inference

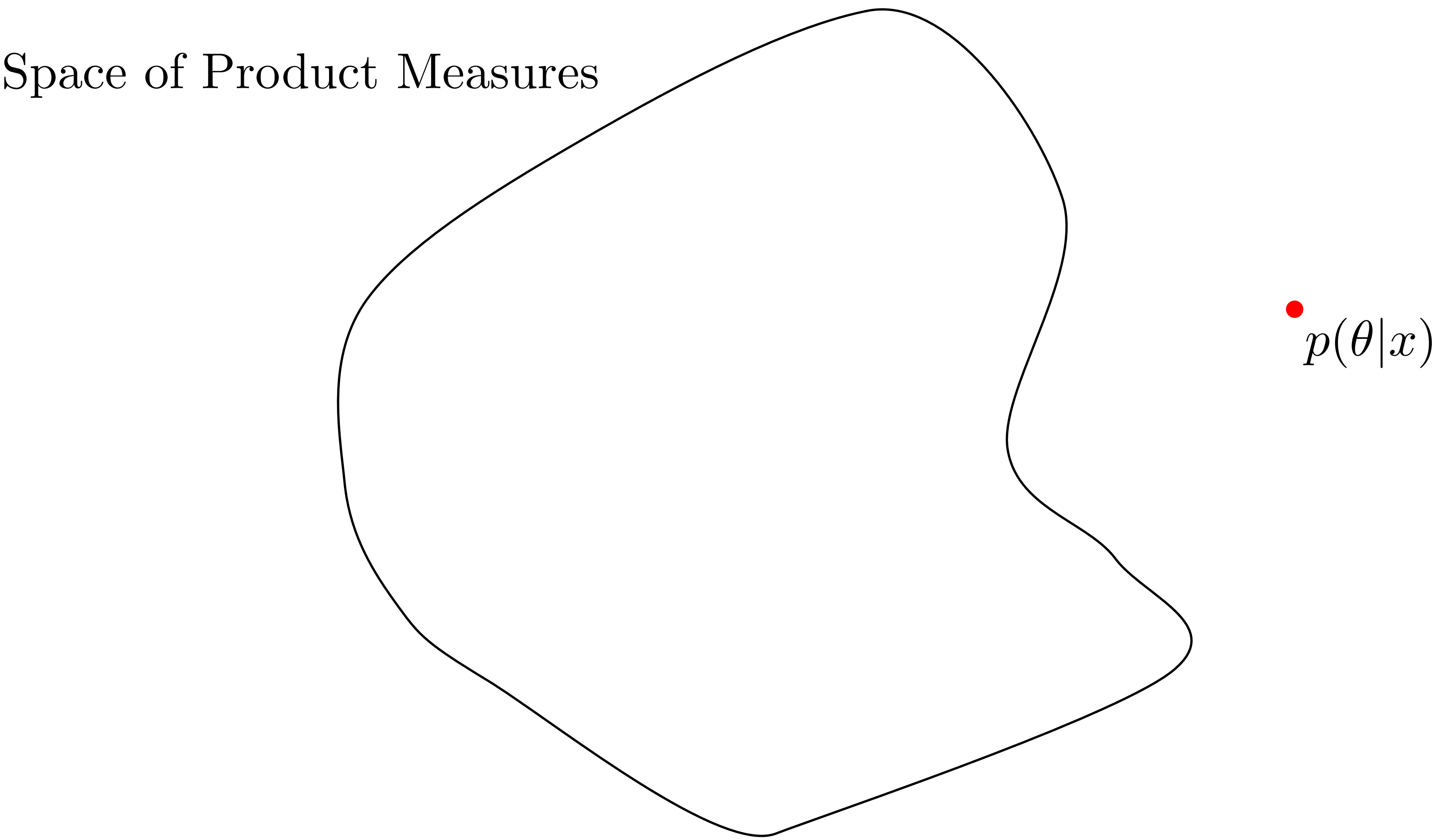
$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}$$

Challenge: Often computationally intractable

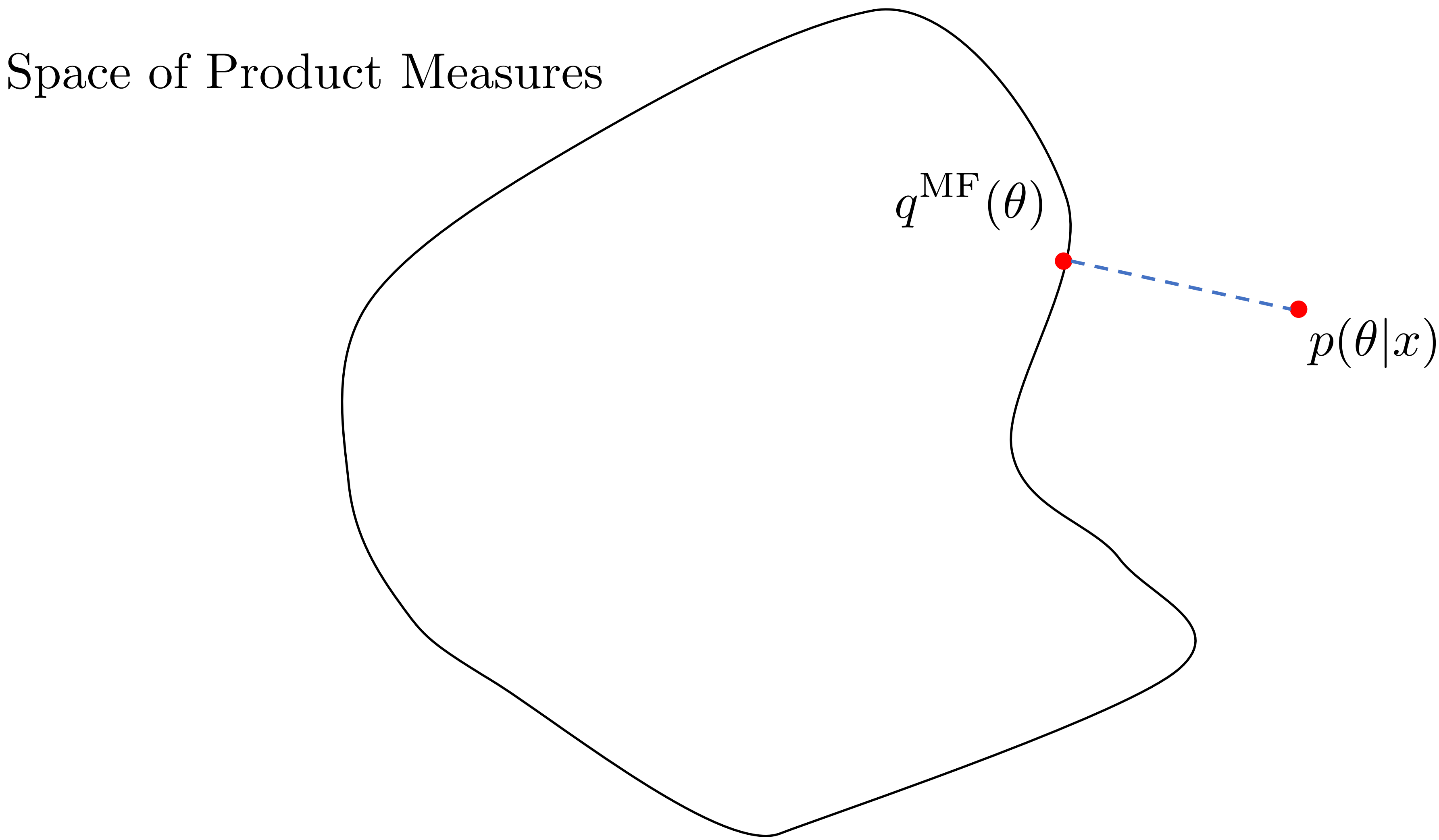
Remedy:

- MCMC (e.g., Gibbs Sampler …
- Mean Field Variational Inference
- …

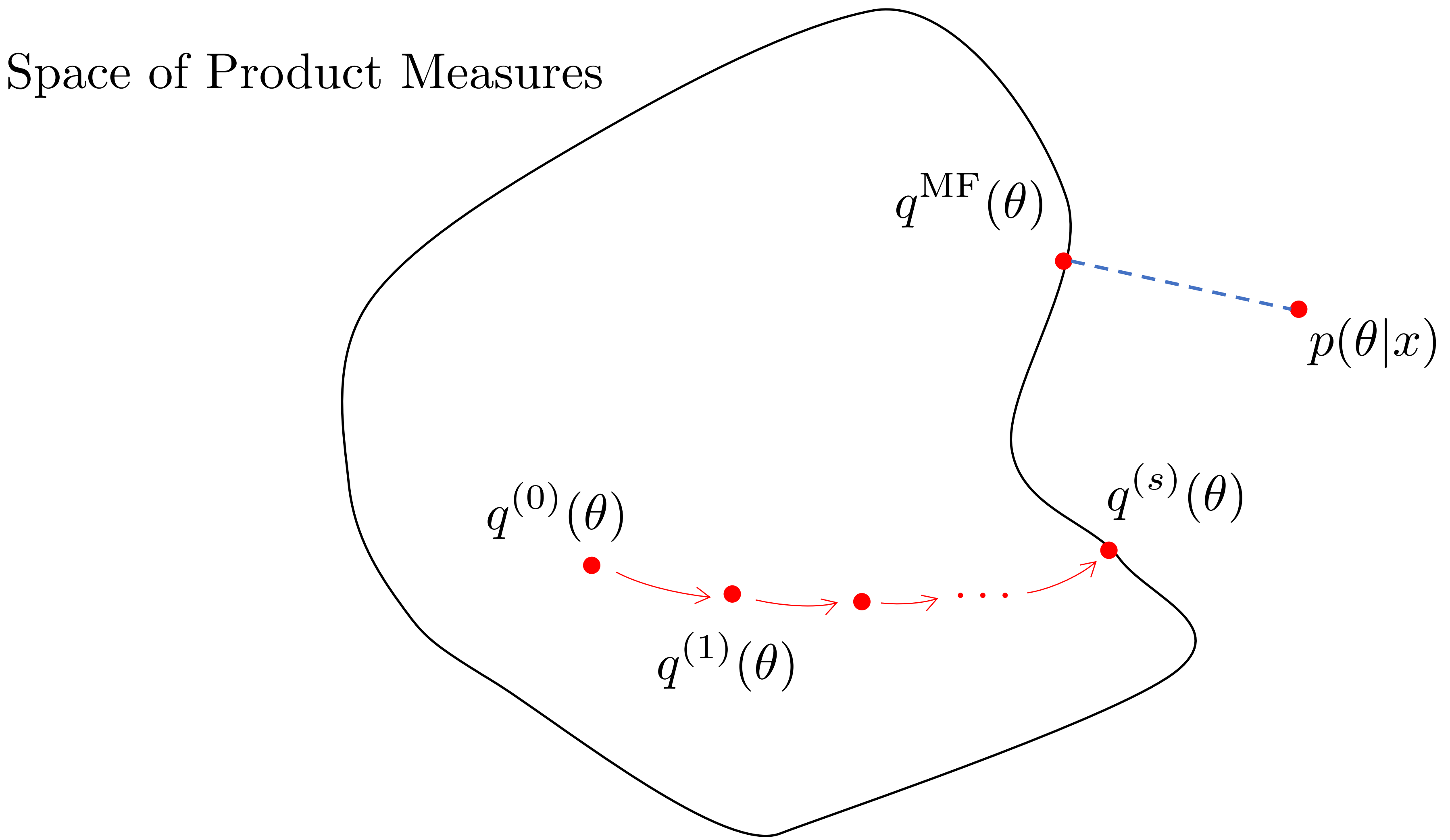
Big Picture



Big Picture



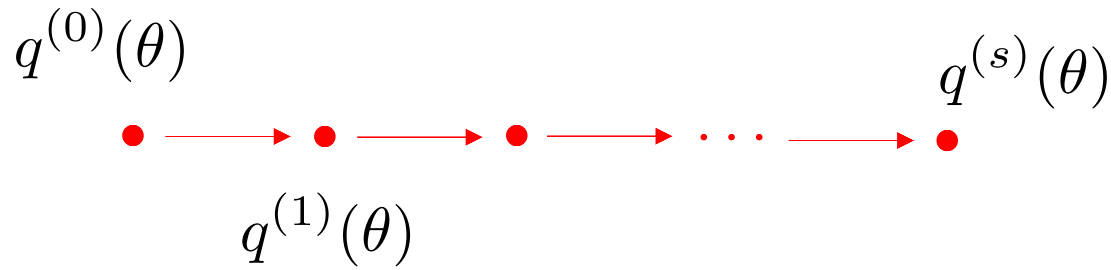
Big Picture



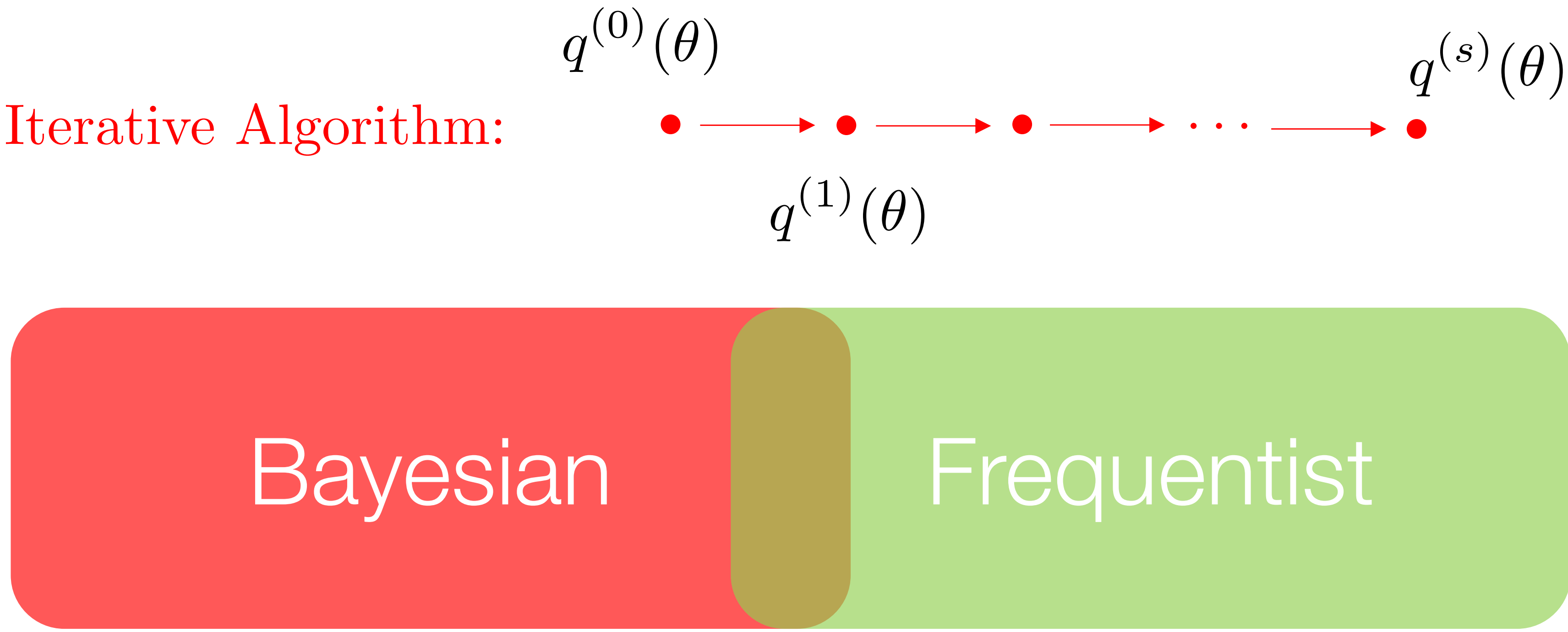
Space of Product Measures
$q^{\mathrm{MF}}(\theta)$
$p(\theta|x)$
$q^{(0)}(\theta)$
$q^{(1)}(\theta)$
$q^{(s)}(\theta)$

Big Picture

Iterative Algorithm:

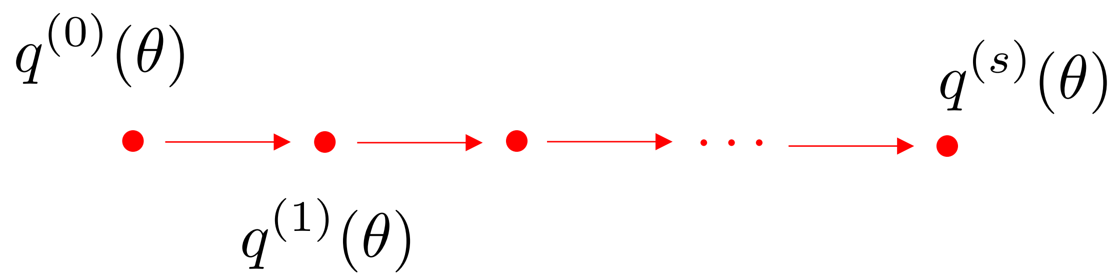


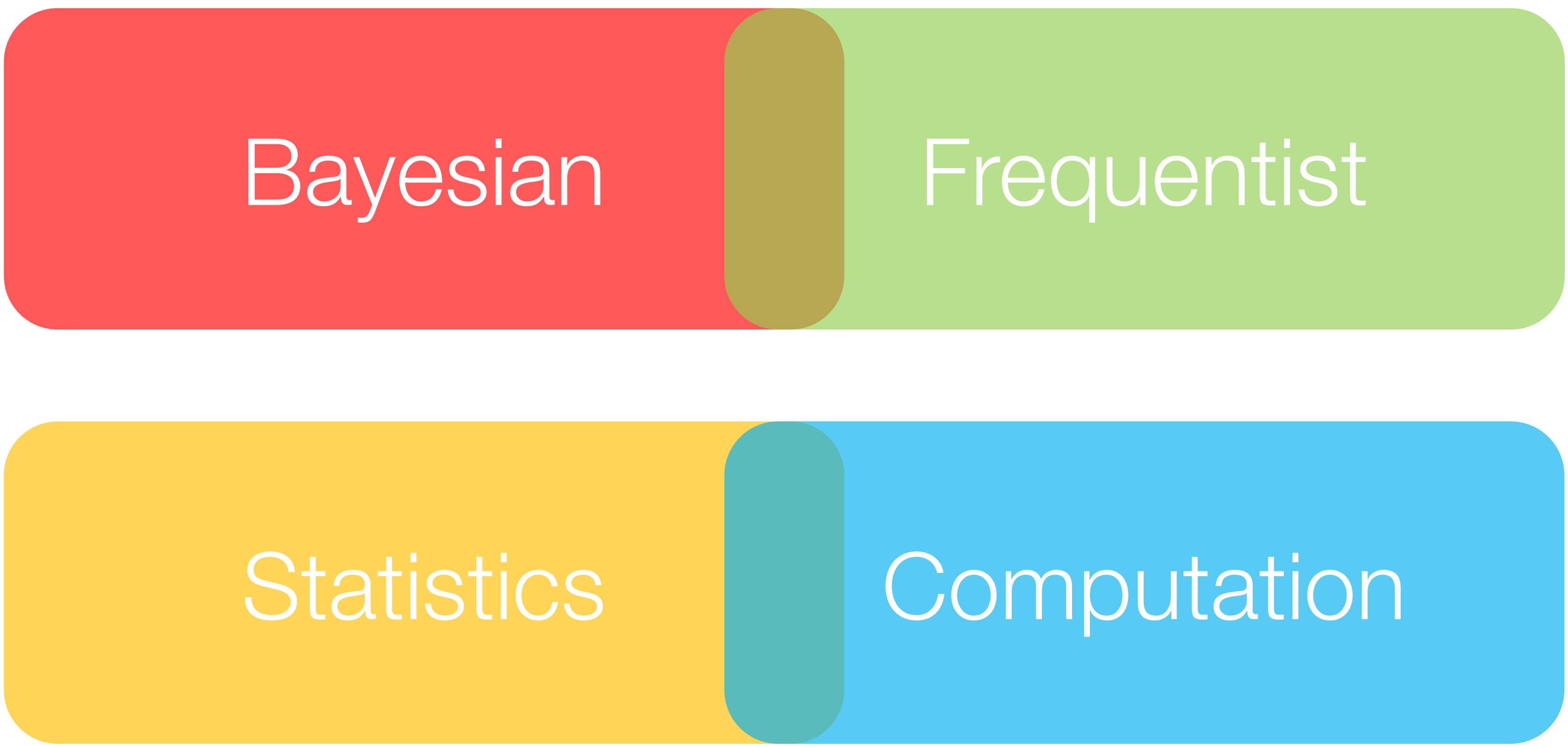
Big Picture


Iterative Algorithm:
$q^{(0)}(\theta)$
$q^{(1)}(\theta)$
$\cdots$
$q^{(s)}(\theta)$
Bayesian
Frequentist

Big Picture

Iterative Algorithm:





Outline:

- Motivating Examples
- Mean Field Variational Inference
- Guarantees of Mean Field Variational Inference on Community Detection
- Three Siblings: Mean Field, Gibbs Sampler, EM

Outline:

- Motivating Examples
- Mean Field Variational Inference
- Guarantees of Mean Field Variational Inference on *Community Detection*

  A Test Case
- Three Siblings: Mean Field, Gibbs Sampler, EM

- Motivating Examples
- Mean Field Variational Inference
- Guarantees of Mean Field Variational Inference on Community Detection
- Three Siblings: Mean Field, Gibbs Sampler, EM

Example I: Topic Models

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Ref: David Blei, Andrew Ng, and Michael Jordan. "Latent Dirichlet Allocation." (2003)

Example I: Latent Dirichlet Allocation

$$Z_{\text{William}},\ Z_{\text{foundation}},\ Z_{\text{performing}},\ \cdots$$
$$\theta_{\text{Arts}},\ \theta_{\text{Education}}, \cdots$$
$$\cdots$$

Example I: Latent Dirichlet Allocation

$$p\left(\begin{matrix} Z_{\text{William}},\ Z_{\text{foundation}},\ Z_{\text{performing}},\ \cdots \\ \theta_{\text{Arts}},\ \theta_{\text{Education}},\cdots \\ \cdots \end{matrix}\middle|\text{Text Corpus}\right)$$

Example I: Latent Dirichlet Allocation

$$p\left(\begin{array}{c} Z_{\text{William}},\ Z_{\text{foundation}},\ Z_{\text{performing}},\ \cdots \\ \theta_{\text{Arts}},\ \theta_{\text{Education}}, \cdots \\ \cdots \end{array} \middle| \text{Text Corpus}\right)$$

approx.

$$\begin{array}{c} \mathbf{q} = q\left(Z_{\text{William}}\right) q\left(Z_{\text{foundation}}\right) q\left(Z_{\text{performing}}\right) \cdots \\ q\left(\theta_{\text{Arts}}\right) q\left(\theta_{\text{Education}}\right) \\ \cdots \end{array}$$

Ref: David Blei, Andrew Ng, and Michael Jordan. "Latent Dirichlet Allocation." (2003)

Example I: Latent Dirichlet Allocation

$$p\left(\begin{array}{c} Z_{\text{William}},\ Z_{\text{foundation}},\ Z_{\text{performing}},\ \cdots \\ \theta_{\text{Arts}},\ \theta_{\text{Education}}, \cdots \\ \cdots \end{array} \middle| \text{Text Corpus} \right)$$

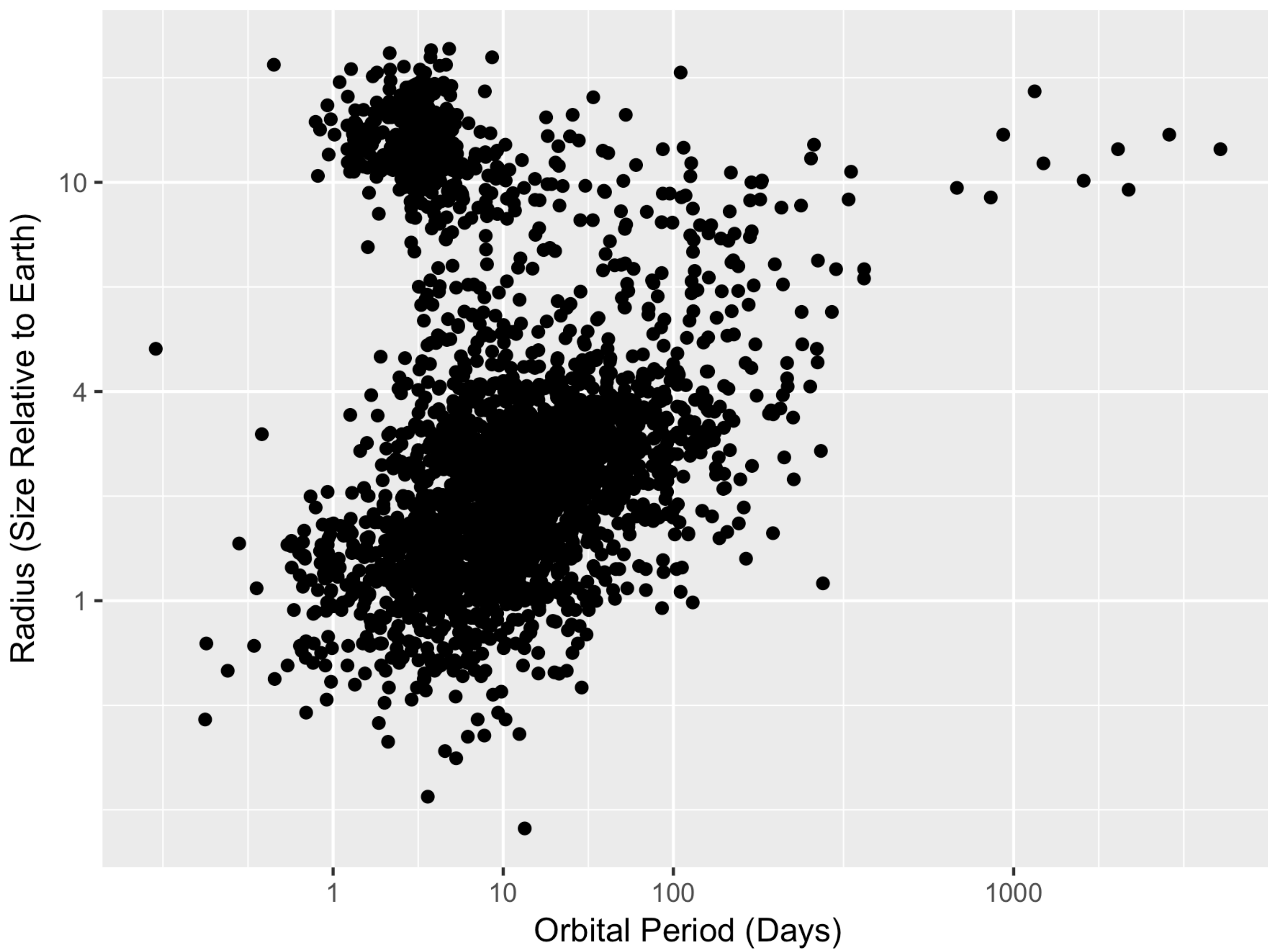
⇧ *approx.*

$$\begin{array}{c} \mathbf{q} = q\left(Z_{\text{William}}\right) q\left(Z_{\text{foundation}}\right) q\left(Z_{\text{performing}}\right) \cdots \\ q\left(\theta_{\text{Arts}}\right) q\left(\theta_{\text{Education}}\right) \\ \cdots \end{array}$$

Iterative Algorithm: $\mathbf{q}^{(0)} \to \mathbf{q}^{(1)} \to \ldots \to \mathbf{q}^{(s)}$

Ref: David Blei, Andrew Ng, and Michael Jordan. "Latent Dirichlet Allocation." (2003)

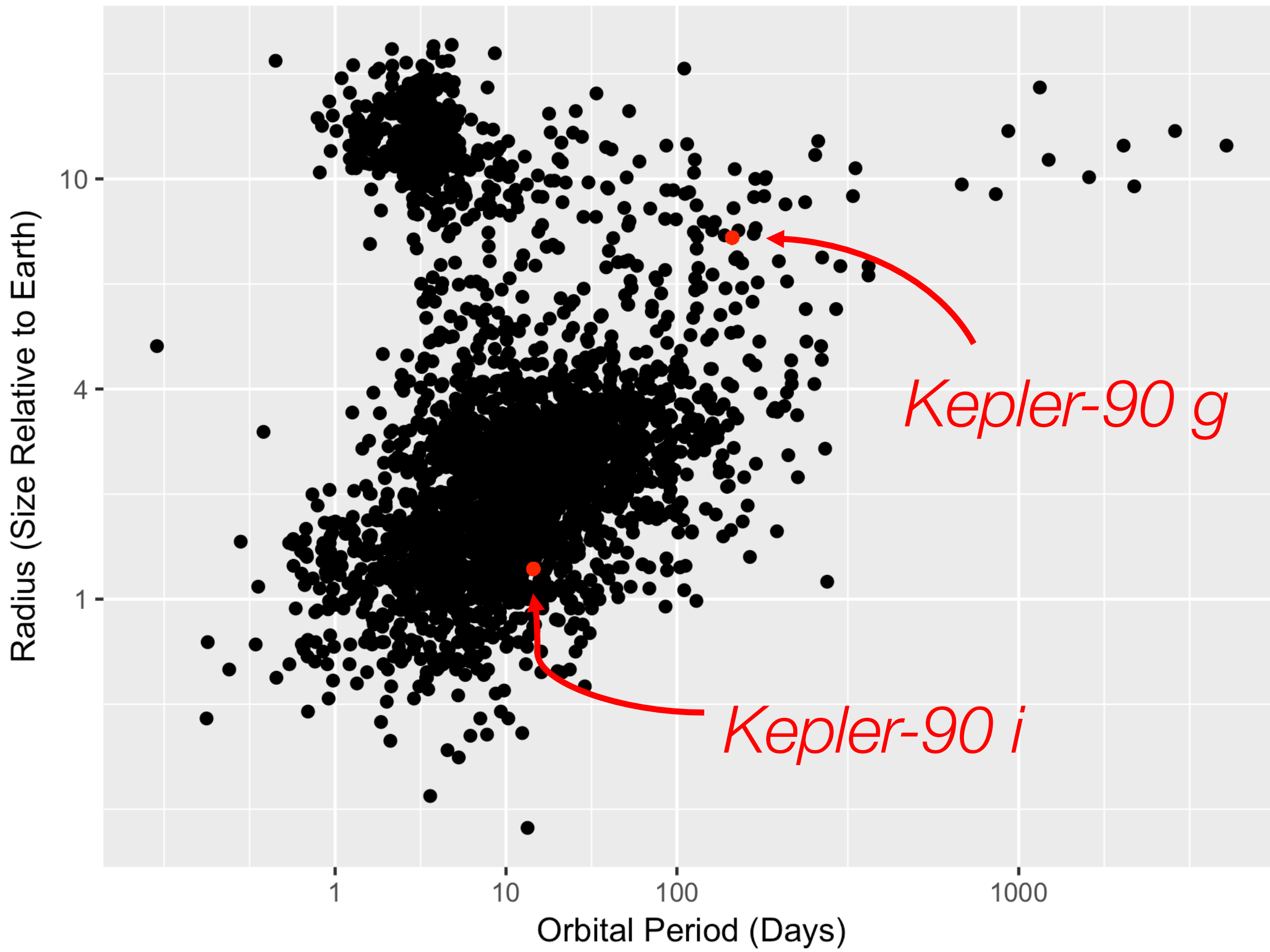
Example II: Clustering



Exoplanets: Orbital Period vs. Radius

Data Source: NASA Exoplanet Archive (https://exoplanetarchive.ipac.caltech.edu)

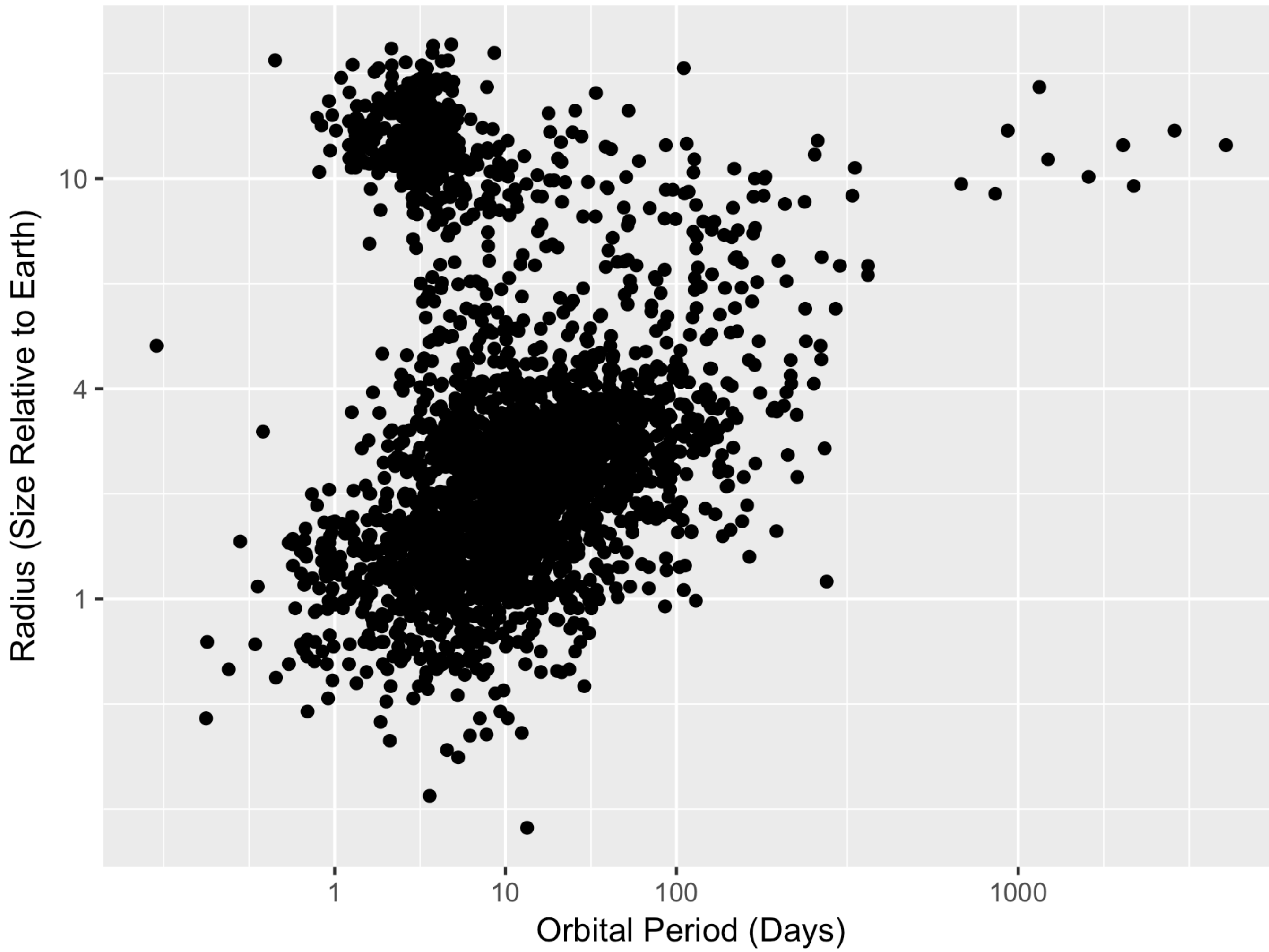
Example II: Clustering



Exoplanets: Orbital Period vs. Radius

Data Source: NASA Exoplanet Archive (https://exoplanetarchive.ipac.caltech.edu)

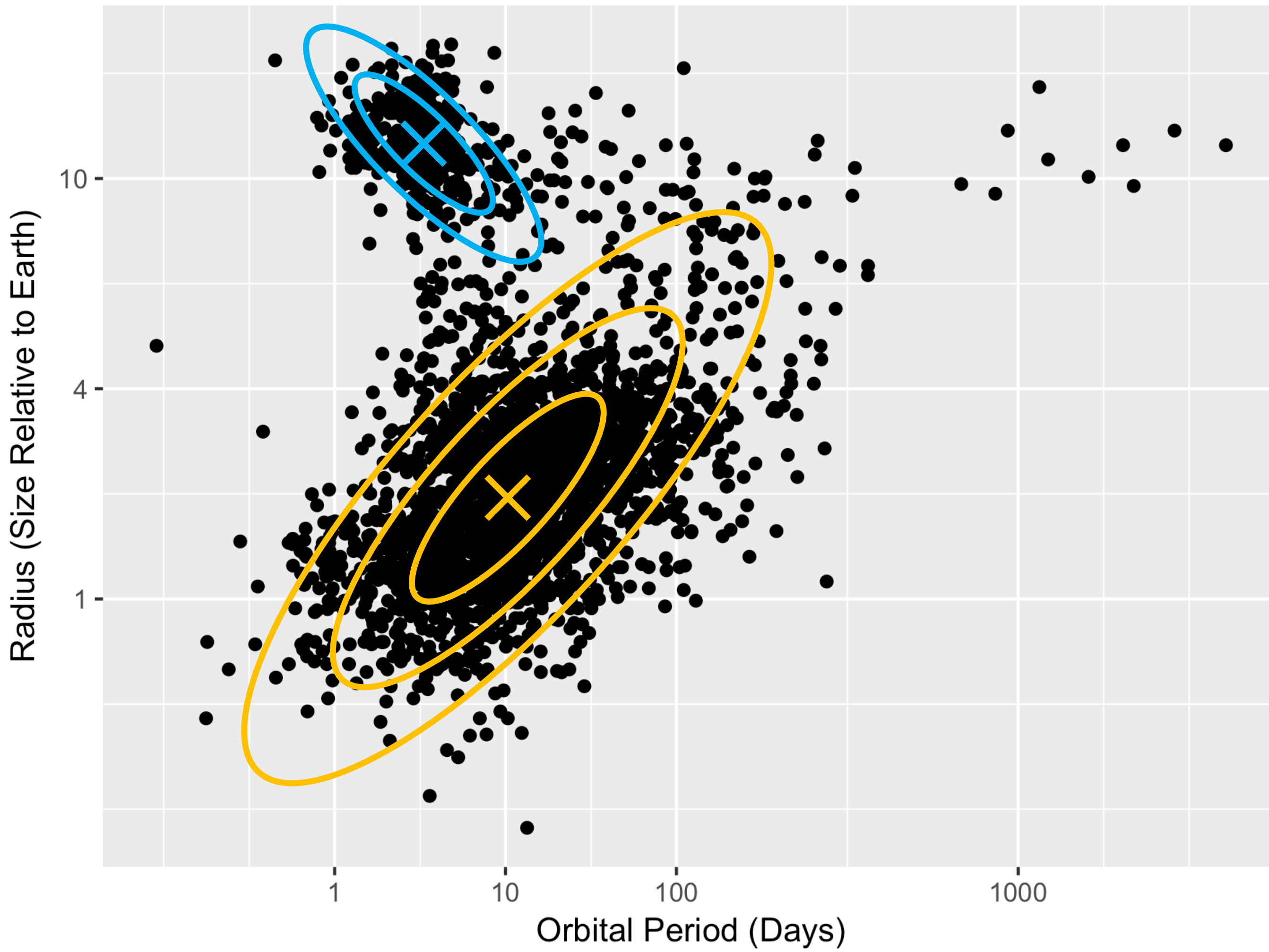
Example II: Clustering



Exoplanets: Orbital Period vs. Radius

Data Source: NASA Exoplanet Archive (https://exoplanetarchive.ipac.caltech.edu)

Example II: Clustering



Exoplanets: Orbital Period vs. Radius

Data Source: NASA Exoplanet Archive (https://exoplanetarchive.ipac.caltech.edu)

Example II: Gaussian Mixture Model

$$Z_{\text{Kepler-90 i}},\ Z_{\text{Kepler-90 g}},\ Z_{\text{HD 114762 b}},\ \cdots$$
$$\mu_1,\ \mu_2,$$
$$\Sigma_1,\ \Sigma_2$$

Example II: Gaussian Mixture Model

$$p\left(\begin{matrix} Z_{\text{Kepler-90 i}},\ Z_{\text{Kepler-90 g}},\ Z_{\text{HD 114762 b}},\ \cdots \\ \mu_1,\ \mu_2, \\ \Sigma_1,\ \Sigma_2 \end{matrix}\ \middle|\ \text{data}\right)$$

Example II: Gaussian Mixture Model

$$p\left(\begin{array}{c} Z_{\text{Kepler-90 i}},\ Z_{\text{Kepler-90 g}},\ Z_{\text{HD 114762 b}},\ \cdots \\ \mu_1,\ \mu_2, \\ \Sigma_1,\ \Sigma_2 \end{array}\middle|\,\text{data}\right)$$

approx.

$$\begin{gathered}\mathbf{q} = q\left(Z_{\text{Kepler-90 i}}\right) q\left(Z_{\text{Kepler-90 g}}\right) q\left(Z_{\text{HD 114762 b}}\right)\cdots \\ q\left(\mu_1\right) q\left(\mu_2\right) \\ q\left(\Sigma_1\right) q\left(\Sigma_2\right)\end{gathered}$$

Ref: Bo Wang and DM Titterington (2006)

Example II: Gaussian Mixture Model

$$p\left(\begin{array}{c} Z_{\text{Kepler-90 i}},\ Z_{\text{Kepler-90 g}},\ Z_{\text{HD 114762 b}},\ \cdots \\ \mu_1,\ \mu_2, \\ \Sigma_1,\ \Sigma_2 \end{array}\middle|\,\text{data}\right)$$

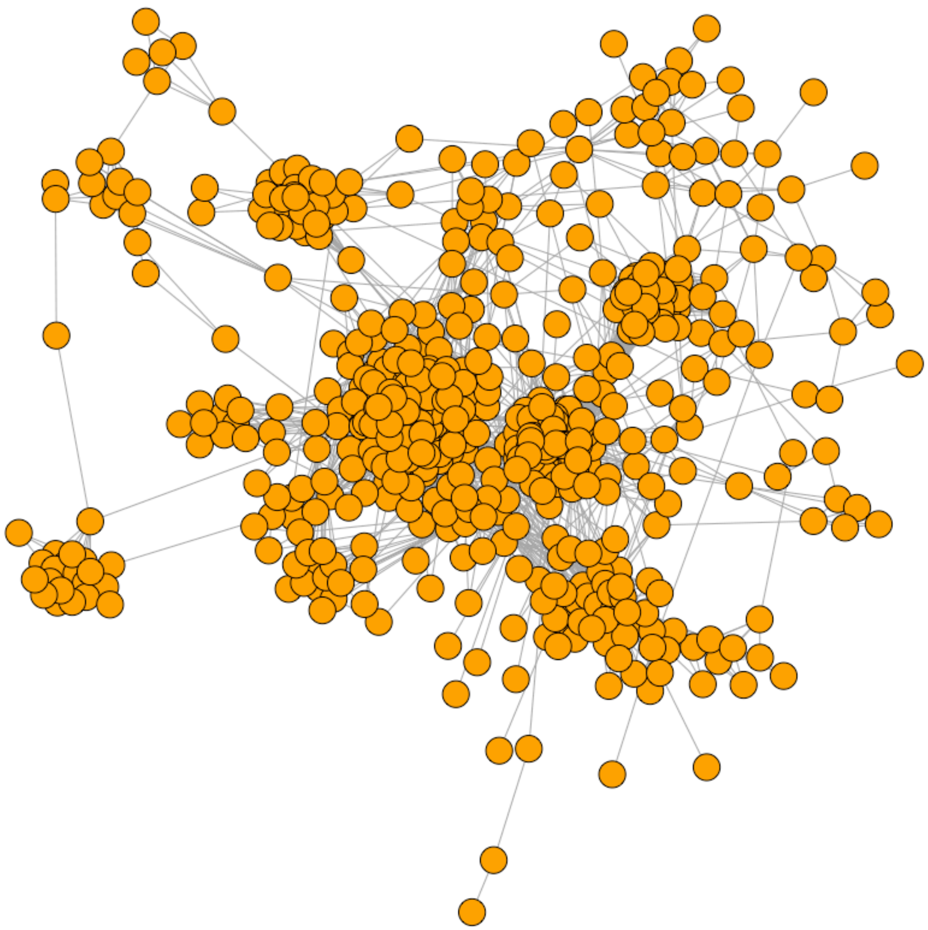
⇧ *approx.*

$$\begin{gathered}\mathbf{q} = q\left(Z_{\text{Kepler-90 i}}\right) q\left(Z_{\text{Kepler-90 g}}\right) q\left(Z_{\text{HD 114762 b}}\right)\cdots \\ q\left(\mu_1\right) q\left(\mu_2\right) \\ q\left(\Sigma_1\right) q\left(\Sigma_2\right)\end{gathered}$$

Iterative Algorithm: $\mathbf{q}^{(0)} \rightarrow \mathbf{q}^{(1)} \rightarrow \ldots \rightarrow \mathbf{q}^{(s)}$

Ref: Bo Wang and DM Titterington (2006)

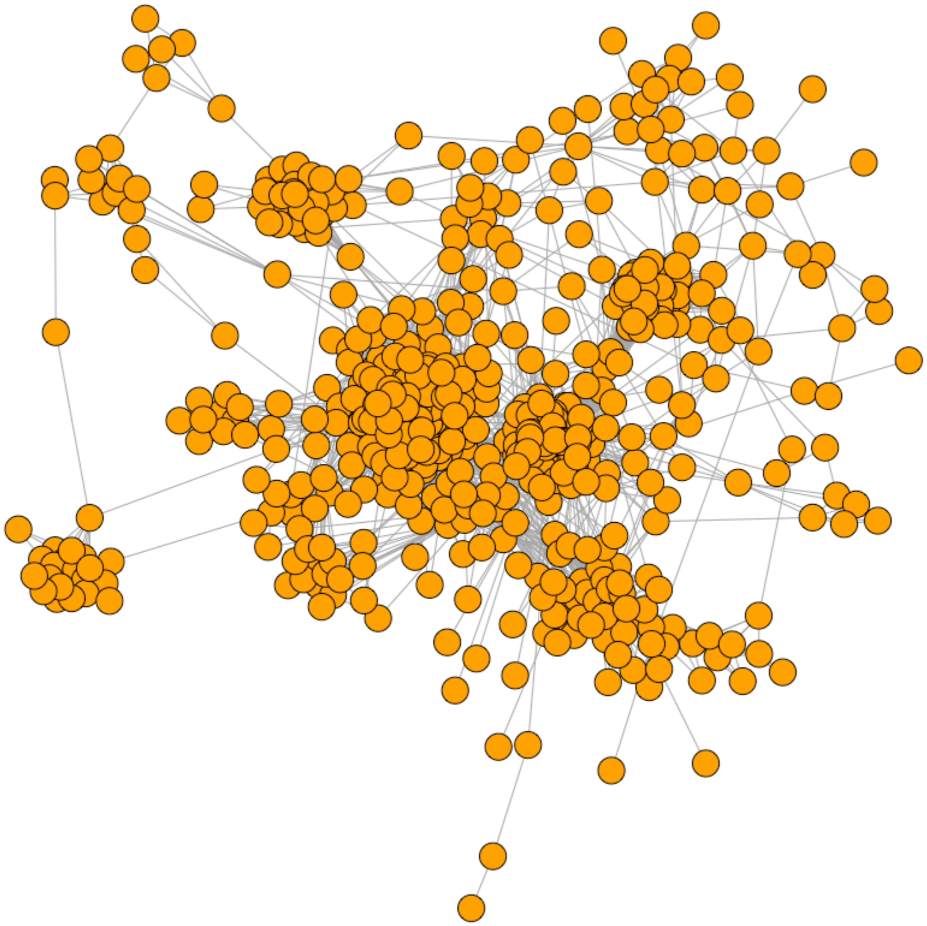
Example III: Community Detection

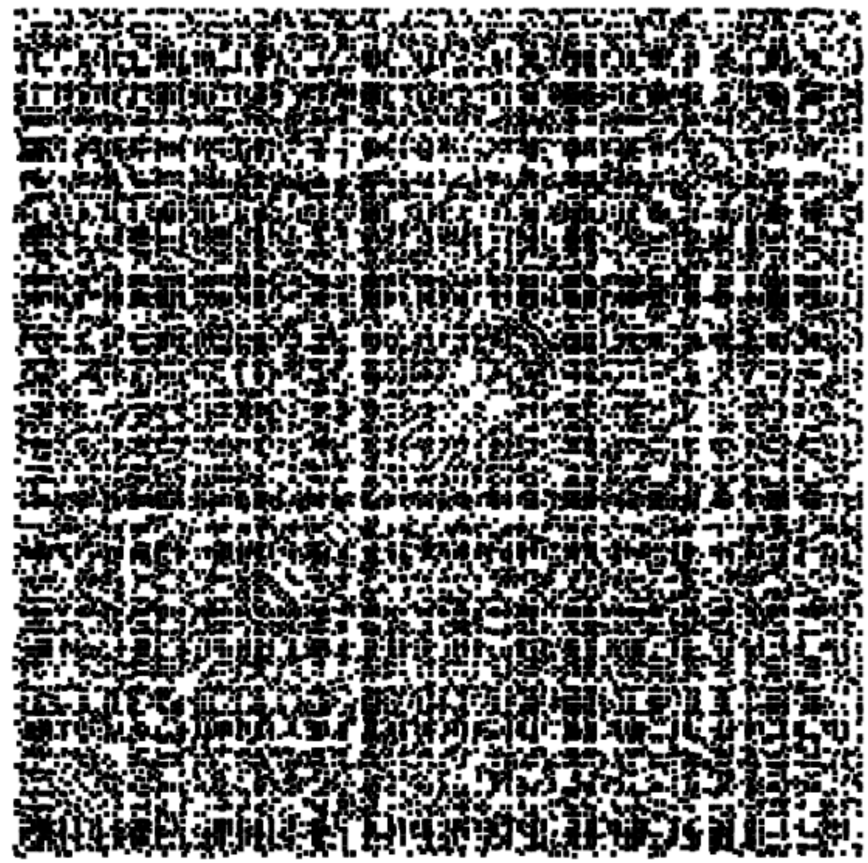


Human Gene–gene Co-association Network

Ref: Mark B Gerstein et al. Nature (2014)

Example III: Community Detection

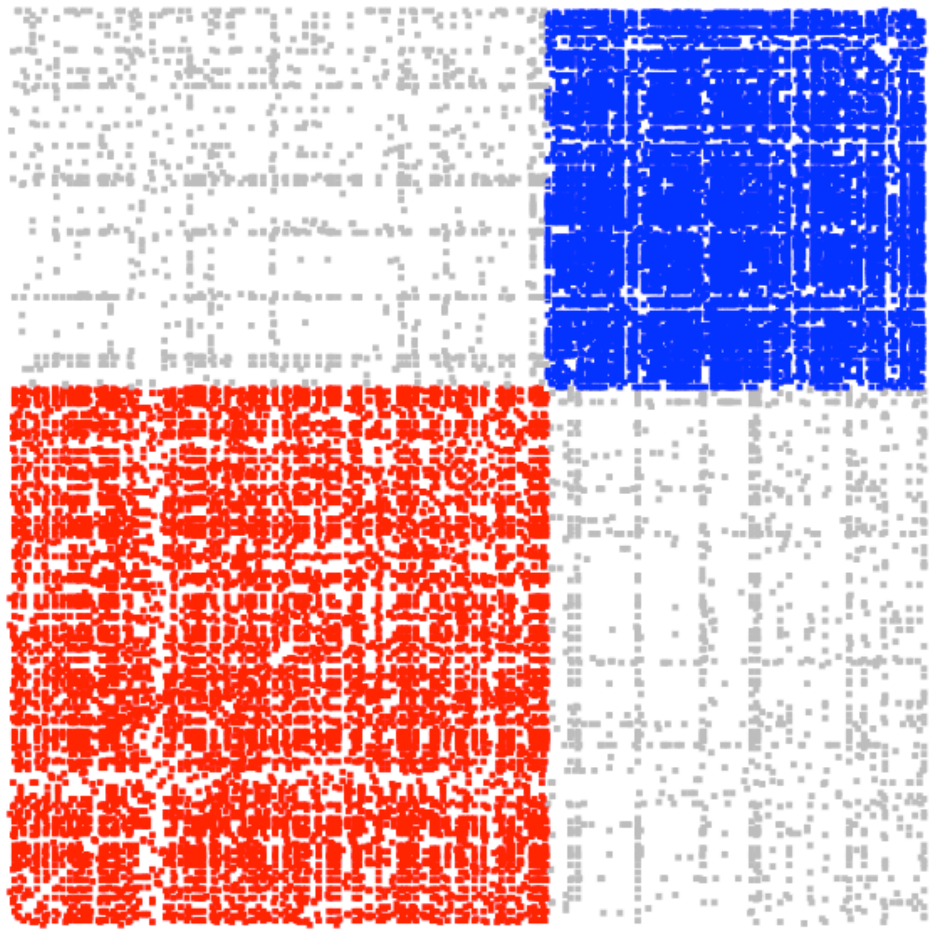




Human Gene–gene Co-association Network

Ref: Mark B Gerstein et al. Nature (2014)

Example III: Community Detection



Human Gene–gene Co-association Network

Ref: Mark B Gerstein et al. Nature (2014)

Example III: Stochastic Block Model

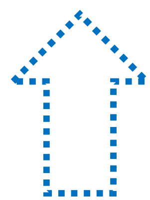
$$Z_{\text{gene }1},\ Z_{\text{gene }2},\ Z_{\text{gene }3},\ \ldots$$
$$p_{\text{within}},\ p_{\text{cross}}$$

Example III: Stochastic Block Model

$$p\left(\begin{array}{c} Z_{\text{gene 1}},\ Z_{\text{gene 2}},\ Z_{\text{gene 3}},\ \ldots \\ p_{\text{within}},\ p_{\text{cross}} \end{array} \middle| \text{network} \right)$$

Example III: Stochastic Block Model

$$p\left(\begin{matrix} Z_{\text{gene 1}},\ Z_{\text{gene 2}},\ Z_{\text{gene 3}},\ \ldots \\ p_{\text{within}},\ p_{\text{cross}} \end{matrix} \middle| \text{network} \right)$$



approx.

$$\mathbf{q} = q\left(Z_{\text{gene 1}}\right) q\left(Z_{\text{gene 2}}\right) q\left(Z_{\text{gene 3}}\right) \ldots \\ q\left(p_{\text{within}}\right) q\left(p_{\text{cross}}\right)$$

Ref: Peter Bickel et al, Annals of Statistics (2013)

Example III: Stochastic Block Model

$$p\left(\begin{matrix} Z_{\text{gene 1}},\ Z_{\text{gene 2}},\ Z_{\text{gene 3}},\ \ldots \\ p_{\text{within}},\ p_{\text{cross}} \end{matrix} \middle| \text{network} \right)$$

approx.

$$\mathbf{q} = q\left(Z_{\text{gene 1}}\right) q\left(Z_{\text{gene 2}}\right) q\left(Z_{\text{gene 3}}\right) \ldots$$
$$q\left(p_{\text{within}}\right) q\left(p_{\text{cross}}\right)$$

Iterative Algorithm: $\mathbf{q}^{(0)} \rightarrow \mathbf{q}^{(1)} \rightarrow \ldots \rightarrow \mathbf{q}^{(s)}$

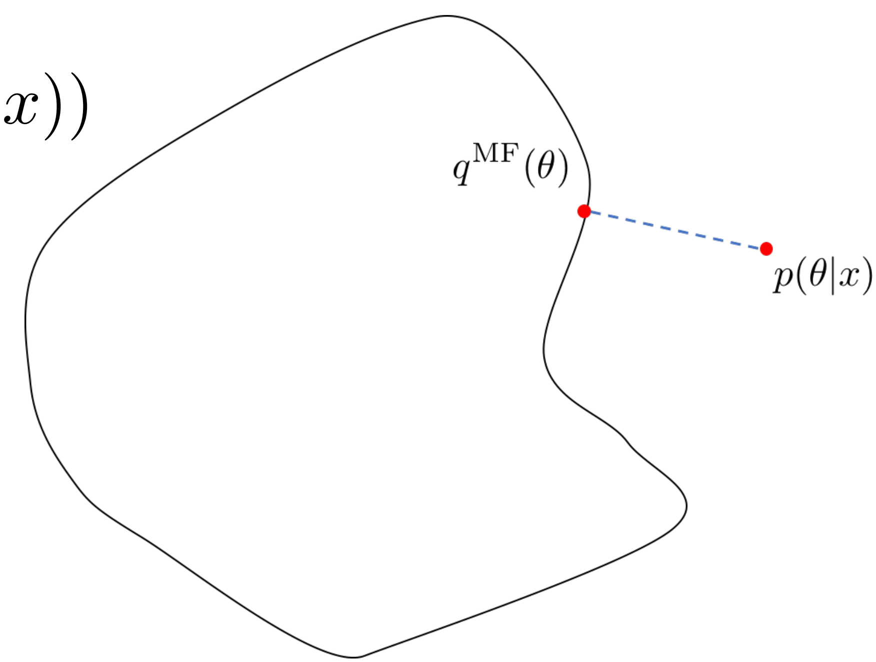
Ref: Peter Bickel et al, Annals of Statistics (2013)

- Motivating Examples
- Mean Field Variational Inference
- Guarantees of Mean Field Variational Inference on Community Detection
- Three Siblings: Mean Field, Gibbs Sampler, EM

Mean Field Variational Inference

- Approximate $p(\theta|x)$ by some $q(\theta)$
- **Product Measure:** $q(\theta) = \prod q_i(\theta_i)$ where $\theta = (\theta_1, \theta_2, \ldots)$
- Minimize the Kullback-Leibler divergence

$$\hat{q}^{\mathrm{MF}} = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q(\theta) \| p(\theta|x))$$



Coordinate Ascent Variational Inference (CAVI)

Iterative Algorithm: Coordinate-wise update on

$$\theta = (\theta_1, \theta_2, \ldots, \theta_i, \ldots)$$

$$\arg\min_q \mathrm{KL}\left(q(\theta) \middle\| p(\theta|x) \right)$$

Ref: C. Bishop, Pattern Recognition and Machine Learning, (2006)

Coordinate Ascent Variational Inference (CAVI)

Iterative Algorithm: Coordinate-wise update on

$$\theta = (\theta_1, \theta_2, \ldots, \theta_i, \ldots)$$

$$\underset{q_1 q_2 \ldots q_i \ldots}{\arg\min} \, \mathrm{KL}\left(q_1(\theta_1) \times q_2(\theta_2) \times \ldots \times q_i(\theta_i) \times \ldots \middle\| p(\theta|x) \right)$$

Ref: C. Bishop, Pattern Recognition and Machine Learning, (2006)

Coordinate Ascent Variational Inference (CAVI)

Iterative Algorithm: Coordinate-wise update on

$$\theta = (\theta_1, \theta_2, \ldots, \theta_i, \ldots)$$

$$\underset{q_1 q_2 \ldots q_i \ldots}{\arg\min} \ \mathrm{KL}\left(q_1(\theta_1) \times q_2(\theta_2) \times \ldots \times q_i(\theta_i) \times \ldots \middle\| p(\theta | x) \right)$$

Ref: C. Bishop, Pattern Recognition and Machine Learning, (2006)

Coordinate Ascent Variational Inference (CAVI)

Iterative Algorithm: Coordinate-wise update on

$$\theta = (\theta_1, \theta_2, \ldots, \theta_i, \ldots)$$

$$\arg\min_{q_1 q_2 \ldots q_i \ldots} \mathrm{KL}\left(q_1(\theta_1) \times q_2(\theta_2) \times \ldots \times q_i(\theta_i) \times \ldots \middle\| p(\theta|x) \right)$$

Ref: C. Bishop, Pattern Recognition and Machine Learning, (2006)

Coordinate Ascent Variational Inference (CAVI)

Iterative Algorithm: Coordinate-wise update on

$$\theta = (\theta_1, \theta_2, \ldots, \theta_i, \ldots)$$

$$\underset{q_1 q_2 \ldots q_i \ldots}{\arg\min} \; \mathrm{KL}\left(q_1(\theta_1) \times q_2(\theta_2) \times \ldots \times q_i(\theta_i) \times \ldots \middle\| p(\theta|x) \right)$$

Ref: C. Bishop, Pattern Recognition and Machine Learning, (2006)

Coordinate Ascent Variational Inference (CAVI)

Iterative Algorithm: Coordinate-wise update on

$$\theta = (\theta_1, \theta_2, \ldots, \theta_i, \ldots)$$

$$\underset{q_1 q_2 \ldots q_i \ldots}{\arg\min} \; \mathrm{KL}\left(q_1(\theta_1) \times q_2(\theta_2) \times \ldots \times q_i(\theta_i) \times \ldots \middle\| p(\theta|x) \right)$$

⇕

Explicit Formula: $q_i(\theta_i) \propto \exp\{\mathbb{E}_{q_{-i}}[\log p(\theta_i|\theta_{-i}, x)]\}$

Ref: C. Bishop, Pattern Recognition and Machine Learning, (2006)

Coordinate Ascent Variational Inference (CAVI)

Iterative Algorithm: Coordinate-wise update on

$$\theta = (\theta_1, \theta_2, \dots, \theta_i, \dots)$$

$$\underset{q_1 q_2 \dots q_i \dots}{\arg\min} \, \mathrm{KL}\left(q_1(\theta_1) \times q_2(\theta_2) \times \dots \times q_i(\theta_i) \times \dots \middle\| p(\theta|x) \right)$$

Explicit Formula: $q_i(\theta_i) \propto \exp\{\cancel{\mathbb{E}}_{q_{-i}}[\log p(\theta_i|\theta_{-i}, x)]\}$

Ref: C. Bishop, Pattern Recognition and Machine Learning, (2006)

Coordinate Ascent Variational Inference (CAVI)

Iterative Algorithm: Coordinate-wise update on

$$\theta = (\theta_1, \theta_2, \dots, \theta_i, \dots)$$

$$\underset{q_1 q_2 \dots q_i \dots}{\arg\min} \ \mathrm{KL}\left(q_1(\theta_1) \times q_2(\theta_2) \times \dots \times q_i(\theta_i) \times \dots \middle\| p(\theta|x) \right)$$

Explicit Formula: $q_i(\theta_i) \propto \exp\{\mathbb{E}_{q_{-i}}[\log p(\theta_i|\theta_{-i}, x)]\}$

Ref: C. Bishop, Pattern Recognition and Machine Learning, (2006)

Coordinate Ascent Variational Inference (CAVI)

Iterative Algorithm: Coordinate-wise update on

$$\theta = (\theta_1, \theta_2, \ldots, \theta_i, \ldots)$$

$$\underset{q_1 q_2 \ldots q_i \ldots}{\arg\min}\ \mathrm{KL}\left(q_1(\theta_1) \times q_2(\theta_2) \times \ldots \times q_i(\theta_i) \times \ldots \middle\| p(\theta|x) \right)$$

Explicit Formula: $q_i(\theta_i) \propto \exp\{\mathbb{E}_{q_{-i}}[\log p(\theta_i|\theta_{-i}, x)]\}$

Ref: C. Bishop, Pattern Recognition and Machine Learning, (2006)

Coordinate Ascent Variational Inference (CAVI)

Iterative Algorithm: Coordinate-wise update on

$$\theta = (\theta_1, \theta_2, \ldots, \theta_i, \ldots)$$

$$\arg\min_{q_1 q_2 \ldots q_i \ldots} \mathrm{KL}\left(q_1(\theta_1) \times q_2(\theta_2) \times \ldots \times q_i(\theta_i) \times \ldots \middle\| p(\theta|x) \right)$$

⇕

Explicit Formula: $q_i(\theta_i) \propto \exp\{\mathbb{E}_{q_{-i}}[\log p(\theta_i|\theta_{-i}, x)]\}$

a "deterministic" version of Gibbs Sampler

Ref: C. Bishop, Pattern Recognition and Machine Learning, (2006)

Coordinate Ascent Variational Inference (CAVI)

Iterative Algorithm: Coordinate-wise update on

$$\theta = (\theta_1, \theta_2, \ldots, \theta_i, \ldots)$$

$$\arg\min_{q_1 q_2 \ldots q_i \ldots} \mathrm{KL}\left(q_1(\theta_1) \times q_2(\theta_2) \times \ldots \times q_i(\theta_i) \times \ldots \middle\| p(\theta|x) \right)$$

⇕

Explicit Formula: $q_i(\theta_i) \propto \exp\{\mathbb{E}_{q_{-i}}[\log p(\theta_i|\theta_{-i}, x)]\}$

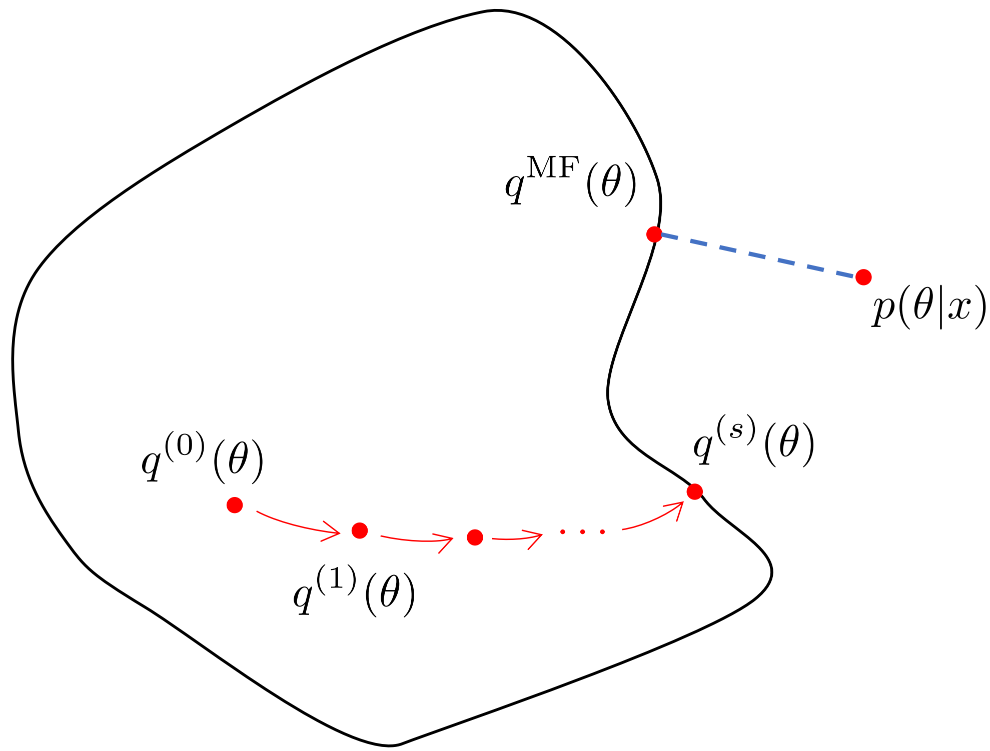
"mean-field"

a "deterministic" version of Gibbs Sampler

Ref: C. Bishop, Pattern Recognition and Machine Learning, (2006)

Goal

Two Approximations:



Questions:

- Provably Statistical Guarantee
- Computation Cost (i.e., # Iterations)

- Motivating Examples
- Mean Field Variational Inference
- Guarantees of Mean Field Variational Inference on Community Detection
- Three Siblings: Mean Field, Gibbs Sampler, EM

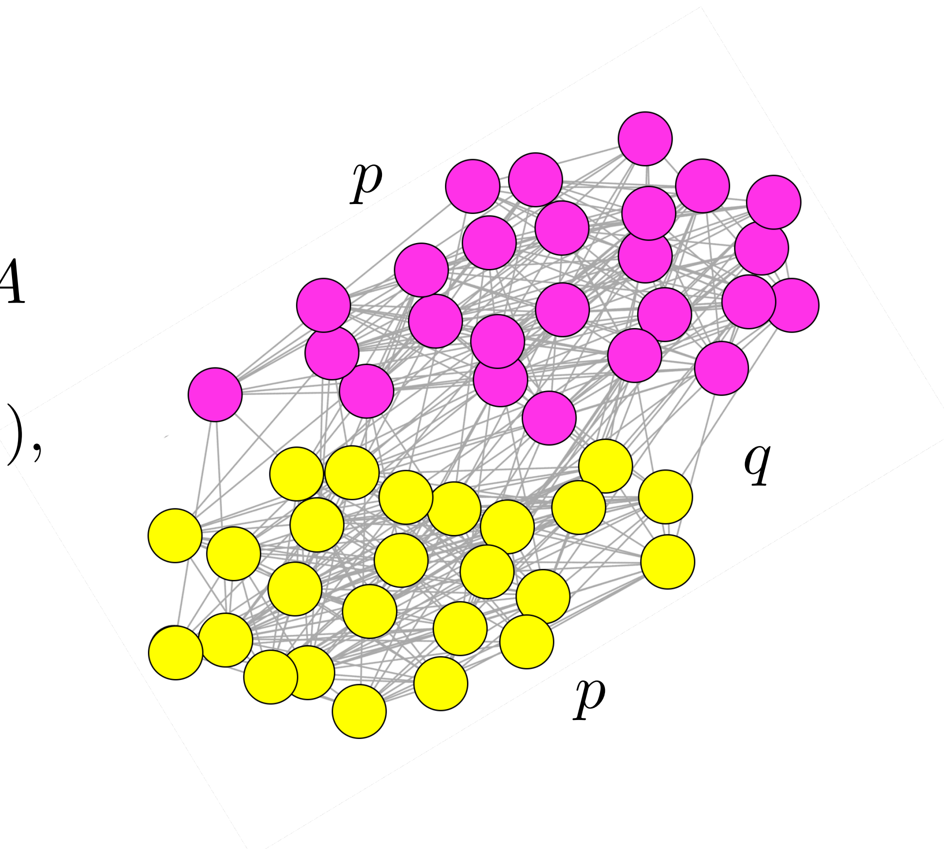
Stochastic Block Model (Two Communities)

Partition: $z \in \{0,1\}^n$

Observation: Adjacency matrix A

$$A_{ij} \sim \begin{cases} \text{Bernoulli}(p), & \text{if } z(i) = z(j), \\ \text{Bernoulli}(q), & \text{o.w.} \end{cases}$$

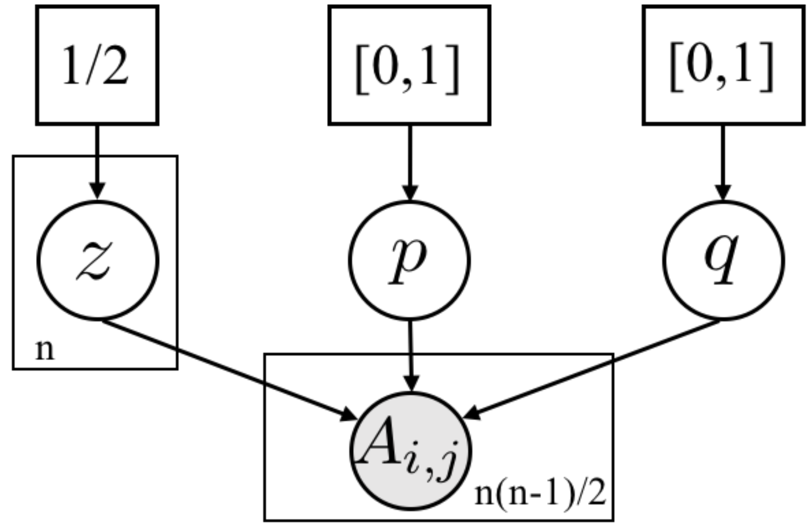
Goal: Recover z from A.



Bayesian Inference for SBM

Prior:

$$\left.\begin{aligned} z_i &\sim \text{Bernoulli}(1/2), \forall i \\ p, q &\sim \text{Uniform}[0,1] \end{aligned}\right\} \text{independently}$$



Posterior:

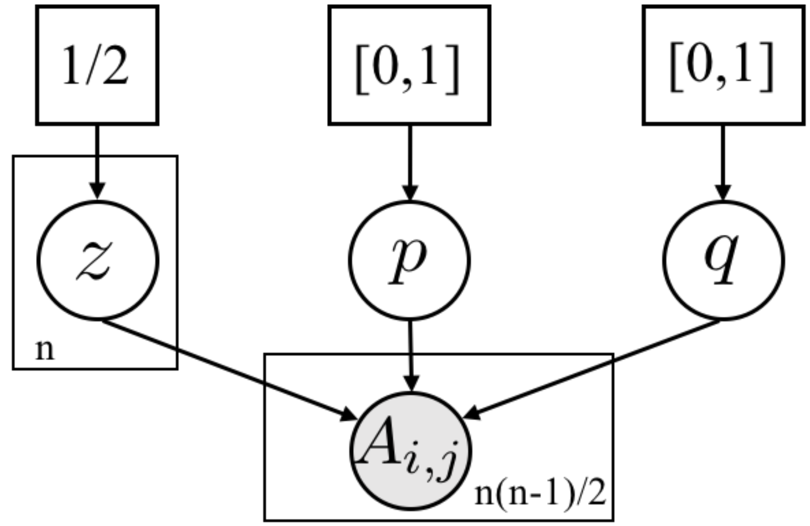
$$\mathbf{p}(z, p, q | A) = \frac{\mathbf{p}(z, p, q, A)}{\sum_{z \in \{0,1\}^n} \int_{p,q} \mathbf{p}(z, p, q, A)}$$

computationally intractable

Bayesian Inference for SBM

Prior:

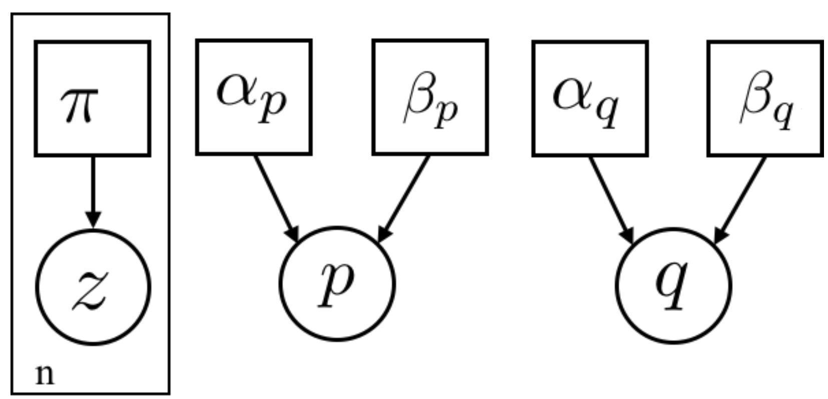
$$\left.\begin{array}{l} z_i \sim \text{Bernoulli}(1/2), \forall i \\ p, q \sim \text{Uniform}[0,1] \end{array}\right\} \text{independently}$$



Posterior: $\mathbf{p}(z, p, q|A)$

Product Measure:

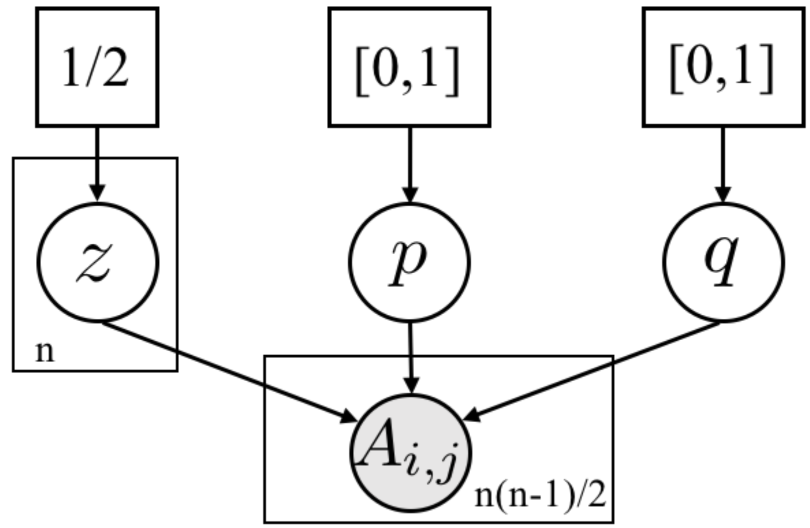
$$\left.\begin{array}{l} z_i \sim \text{Bernoulli}(\pi_i), \forall i \\ p \sim \text{Beta}(\alpha_p, \beta_p) \\ q \sim \text{Beta}(\alpha_q, \beta_q) \end{array}\right\} \text{independently}$$



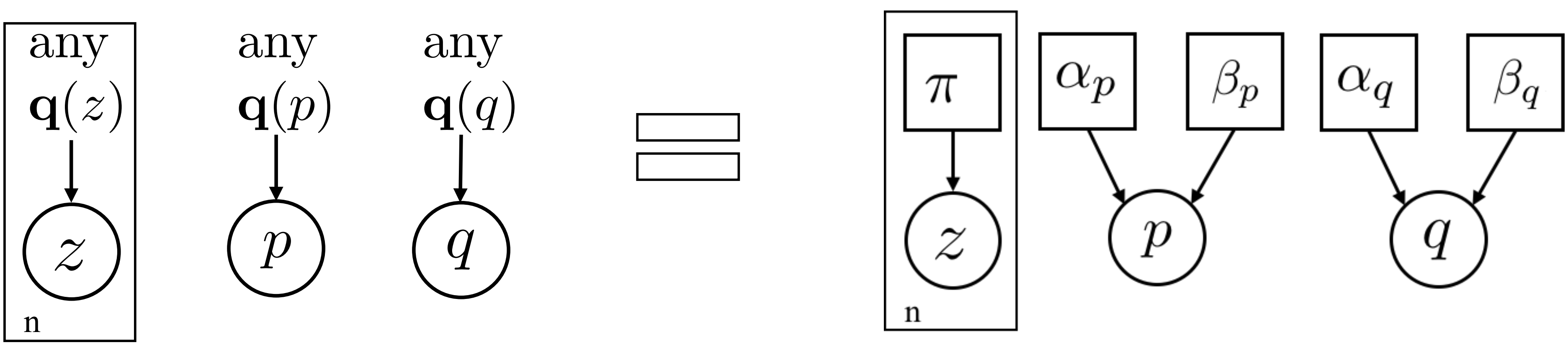
Bayesian Inference for SBM

Prior:

$$\left.\begin{array}{l} z_i \sim \text{Bernoulli}(1/2), \forall i \\ p, q \sim \text{Uniform}[0,1] \end{array}\right\} \text{independently}$$



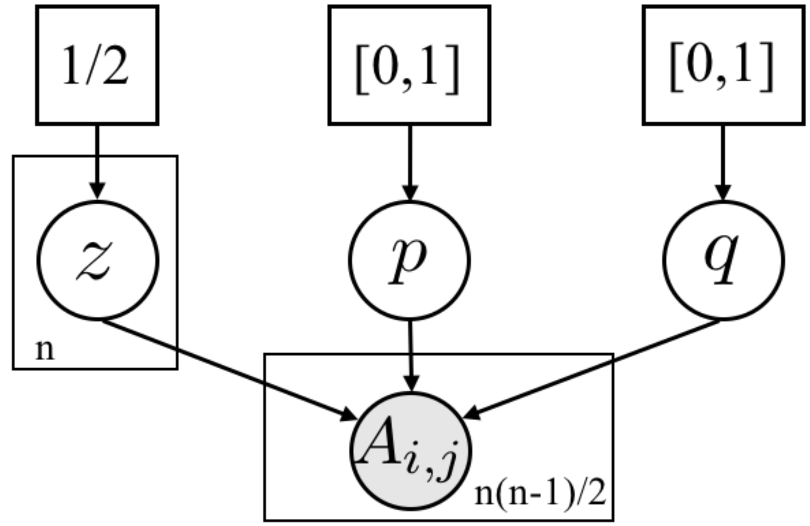
Posterior: $\mathbf{p}(z, p, q | A)$



Bayesian Inference for SBM

Prior:

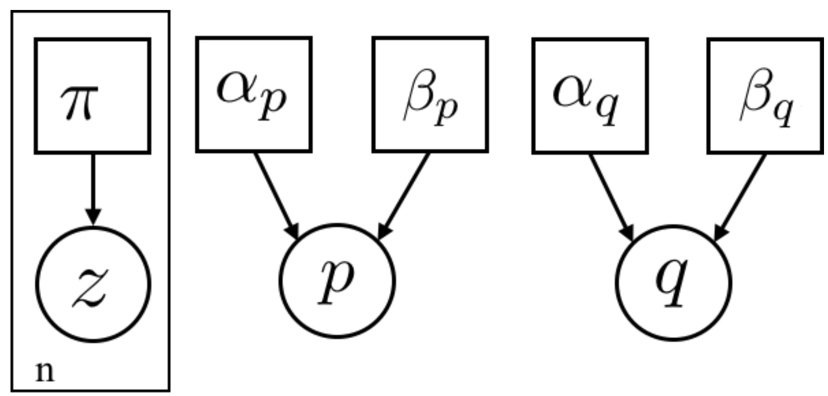
$$\left.\begin{aligned} z_i &\sim \text{Bernoulli}(1/2), \forall i \\ p, q &\sim \text{Uniform}[0,1] \end{aligned}\right\} \text{independently}$$



Posterior: $\mathbf{p}(z, p, q | A)$

Product Measure:

$$\left.\begin{aligned} z_i &\sim \text{Bernoulli}(\pi_i), \forall i \\ p &\sim \text{Beta}(\alpha_p, \beta_p) \\ q &\sim \text{Beta}(\alpha_q, \beta_q) \end{aligned}\right\} \text{independently}$$



Iterative Algorithm

Initializer: $\pi^{(0)}$

Updates on p, q: $\mathbf{q}^{(s)}(p) \sim \mathrm{Beta}(\alpha_p^{(s)}, \beta_p^{(s)})$, $\mathbf{q}^{(s)}(q) \sim \mathrm{Beta}(\alpha_q^{(s)}, \beta_q^{(s)})$

Updates on z: $\mathbf{q}^{(s)}(z_i) \sim \mathrm{Bernoulli}(\pi_i^{(s)}), \forall i \in [n]$

Iterative Algorithm

$$q_i(\theta_i) \propto \exp\{\mathbb{E}_{q_{-i}}[\log p(\theta_i|\theta_{-i}, x)]\}$$

Initializer: $\pi^{(0)}$

Updates on p, q: $\mathbf{q}^{(s)}(p) \sim \text{Beta}(\alpha_p^{(s)}, \beta_p^{(s)}),\ \mathbf{q}^{(s)}(q) \sim \text{Beta}(\alpha_q^{(s)}, \beta_q^{(s)})$

Updates on z: $\mathbf{q}^{(s)}(z_i) \sim \text{Bernoulli}(\pi_i^{(s)}), \forall i \in [n]$

Iterative Algorithm

$$q_i(\theta_i) \propto \exp\{\mathbb{E}_{q_{-i}}[\log p(\theta_i|\theta_{-i}, x)]\}$$

Initializer: $\pi^{(0)}$

Updates on p, q: $\mathbf{q}^{(s)}(p) \sim \text{Beta}(\alpha_p^{(s)}, \beta_p^{(s)})$, $\mathbf{q}^{(s)}(q) \sim \text{Beta}(\alpha_q^{(s)}, \beta_q^{(s)})$

$$\alpha_p^{(s)} \leftarrow 1 + \sum_{i<j} A_{i,j} \left(\pi_i^{(s-1)}\pi_j^{(s-1)} + \left(1 - \pi_i^{(s-1)}\right)\left(1 - \pi_j^{(s-1)}\right)\right),$$

$$\beta_p^{(s)} \leftarrow 1 + \sum_{i<j} (1 - A_{i,j}) \left(\pi_i^{(s-1)}\pi_j^{(s-1)} + \left(1 - \pi_i^{(s-1)}\right)\left(1 - \pi_j^{(s-1)}\right)\right),$$

$$\alpha_q^{(s)} \leftarrow 1 + \sum_{i<j} A_{i,j}\pi_i^{(s-1)} \left(1 - \pi_j^{(s-1)}\right), \quad \beta_q^{(s)} \leftarrow 1 + \sum_{i<j} (1 - A_{i,j})\pi_i^{(s-1)} \left(1 - \pi_j^{(s-1)}\right).$$

Updates on z: $\mathbf{q}^{(s)}(z_i) \sim \text{Bernoulli}(\pi_i^{(s)}), \forall i \in [n]$

$$\pi_i^{(s)} \leftarrow \frac{\exp\left(2t^{(s)} \sum_{j \neq i} \pi_j^{(s-1)}(A_{i,j} - \lambda^{(s)})\right)}{\exp\left(2t^{(s)} \sum_{j \neq i} \pi_j^{(s-1)}(A_{i,j} - \lambda^{(s)})\right) + \exp\left(2t^{(s)} \sum_{j \neq i}(1 - \pi_j^{(s-1)})(A_{i,j} - \lambda^{(s)})\right)},$$

where $t^{(s)}, \lambda^{(s)}$ are functions of $\alpha_p^{(s)}, \beta_p^{(s)}, \alpha_q^{(s)}, \beta_q^{(s)}$.

Iterative Algorithm

$$q_i(\theta_i) \propto \exp\{\mathbb{E}_{q_{-i}}[\log p(\theta_i|\theta_{-i}, x)]\}$$

Initializer: $\pi^{(0)}$

Updates on p, q: $\mathbf{q}^{(s)}(p) \sim \text{Beta}(\alpha_p^{(s)}, \beta_p^{(s)}),\ \mathbf{q}^{(s)}(q) \sim \text{Beta}(\alpha_q^{(s)}, \beta_q^{(s)})$

$$\alpha_p^{(s)} \leftarrow 1 + \sum_{i<j} A_{i,j}\left(\pi_i^{(s-1)}\pi_j^{(s-1)} + \left(1-\pi_i^{(s-1)}\right)\left(1-\pi_j^{(s-1)}\right)\right),$$

$$\beta_p^{(s)} \leftarrow 1 + \sum_{i<j}(1-A_{i,j})\left(\pi_i^{(s-1)}\pi_j^{(s-1)} + \left(1-\pi_i^{(s-1)}\right)\left(1-\pi_j^{(s-1)}\right)\right),$$

Computationally Easy

$$\alpha_q^{(s)} \leftarrow 1 + \sum_{i<j} A_{i,j}\pi_i^{(s-1)}\left(1-\pi_j^{(s-1)}\right), \quad \beta_q^{(s)} \leftarrow 1 + \sum_{i<j}(1-A_{i,j})\pi_i^{(s-1)}\left(1-\pi_j^{(s-1)}\right).$$

Updates on z: $\mathbf{q}^{(s)}(z_i) \sim \text{Bernoulli}(\pi_i^{(s)}), \forall i \in [n]$

$$\pi_i^{(s)} \leftarrow \frac{\exp\left(2t^{(s)}\sum_{j\neq i}\pi_j^{(s-1)}(A_{i,j}-\lambda^{(s)})\right)}{\exp\left(2t^{(s)}\sum_{j\neq i}\pi_j^{(s-1)}(A_{i,j}-\lambda^{(s)})\right) + \exp\left(2t^{(s)}\sum_{j\neq i}(1-\pi_j^{(s-1)})(A_{i,j}-\lambda^{(s)})\right)},$$

Computationally Easy

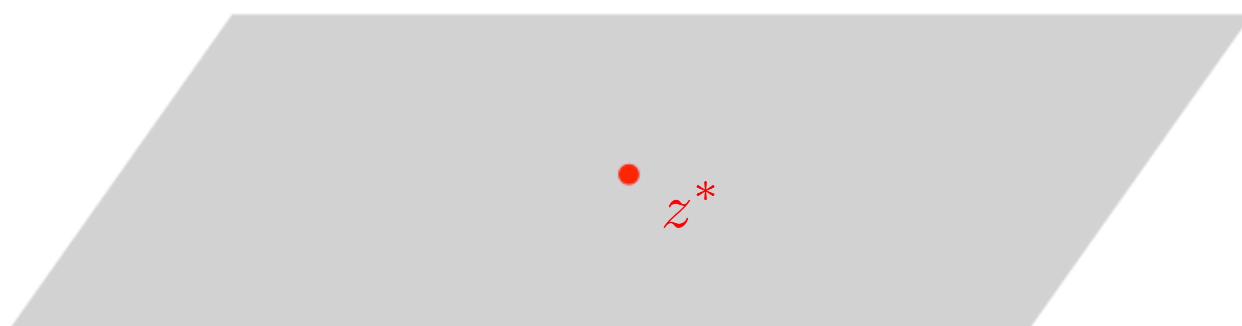
where $t^{(s)}, \lambda^{(s)}$ are functions of $\alpha_p^{(s)}, \beta_p^{(s)}, \alpha_q^{(s)}, \beta_q^{(s)}$.

Evaluation: Posterior Contraction

$$\textbf{Output:}\ \ z \sim \mathbf{q}^{(s)}(z) = \prod_i \mathrm{Bernoulli}(\pi_i^{(s)})$$

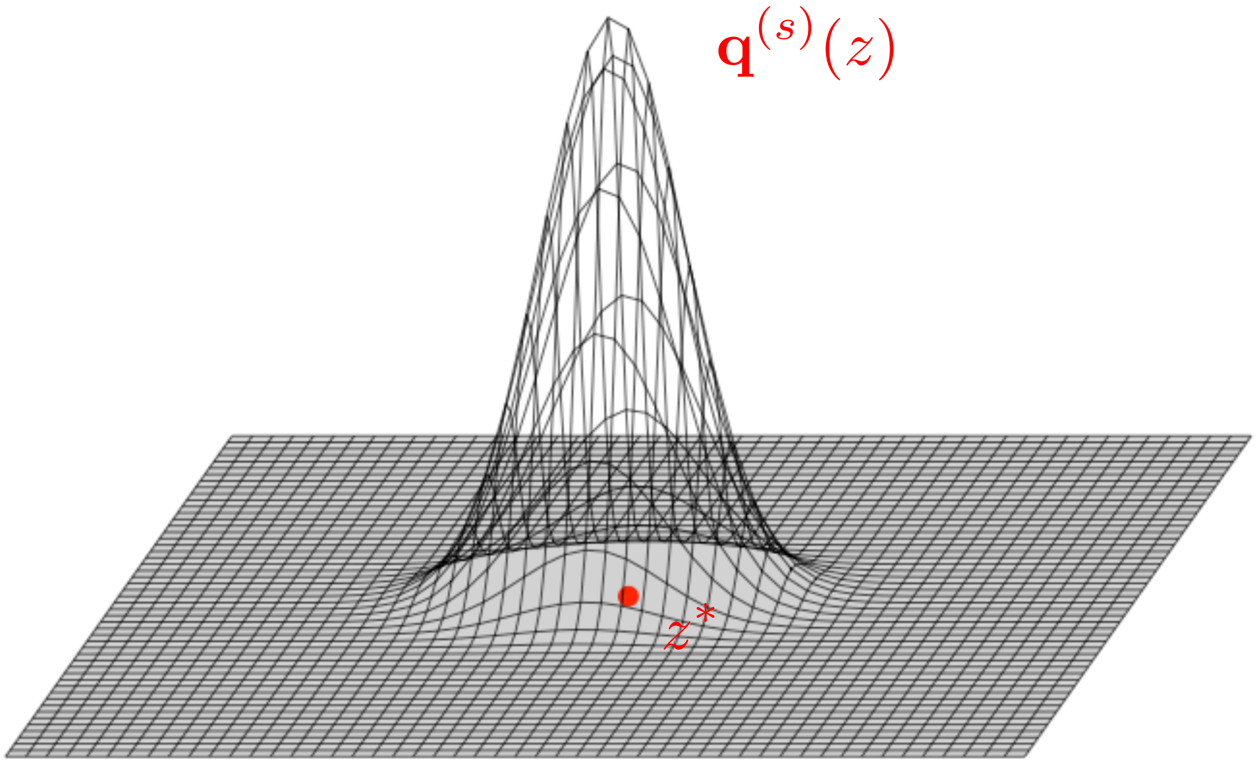
Evaluation: Posterior Contraction

$$\textbf{Output:}\ z \sim \mathbf{q}^{(s)}(z) = \prod_i \mathrm{Bernoulli}(\pi_i^{(s)})$$



Evaluation: Posterior Contraction

Output: $z \sim \mathbf{q}^{(s)}(z) = \prod_i \text{Bernoulli}(\pi_i^{(s)})$



Evaluation: Posterior Contraction

Output: $z \sim \mathbf{q}^{(s)}(z) = \prod_i \text{Bernoulli}(\pi_i^{(s)})$

Ground Truth: z^*, p^*, q^*

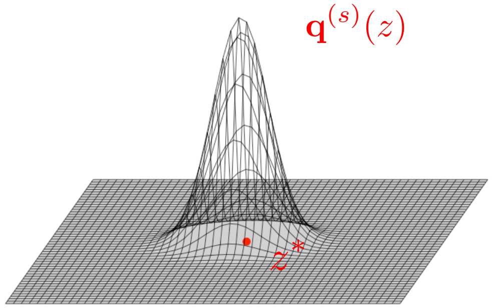
Evaluation: Posterior Contraction

Output: $z \sim \mathbf{q}^{(s)}(z) = \prod_i \text{Bernoulli}(\pi_i^{(s)})$

Ground Truth: z^*, p^*, q^*

Concentration of $\mathbf{q}^{(s)}(z)$ around z^*:

$$\mathbb{E}_{z \sim \mathbf{q}^{(s)}(z)}[\ell(z, z^*)]$$



Evaluation: Posterior Contraction

Output: $z \sim \mathbf{q}^{(s)}(z) = \prod_i \text{Bernoulli}(\pi_i^{(s)})$

Ground Truth: z^*, p^*, q^*

Concentration of $\mathbf{q}^{(s)}(z)$ around z^*:

$$\mathbb{E}_{z \sim \mathbf{q}^{(s)}(z)}[\ell(z, z^*)]$$

Loss Function: $\ell(z, z^*) = \frac{1}{n}\left\{ \|z - z^*\|_0 \qquad\qquad\qquad\qquad \right\}$

Evaluation: Posterior Contraction

Output: $z \sim \mathbf{q}^{(s)}(z) = \prod_i \text{Bernoulli}(\pi_i^{(s)})$

Ground Truth: z^*, p^*, q^*

Concentration of $\mathbf{q}^{(s)}(z)$ around z^*:

$$\mathbb{E}_{z \sim \mathbf{q}^{(s)}(z)}[\ell(z, z^*)]$$

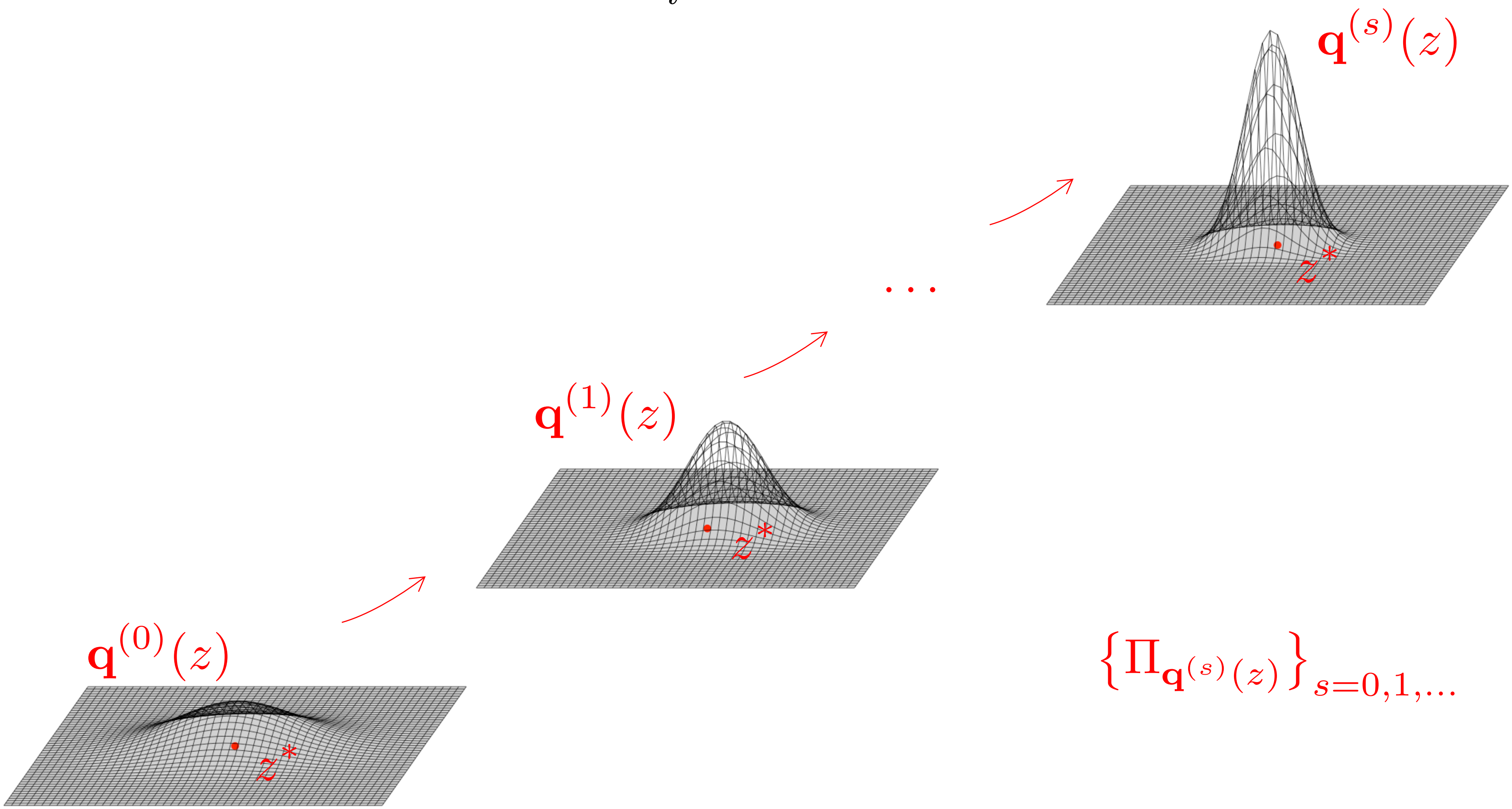
Loss Function: $\ell(z, z^*) = \frac{1}{n} \min \left\{ \|z - z^*\|_0 , \|z - (1 - z^*)\|_0 \right\}$

(1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0)

(0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1)

Evaluation: Posterior Contraction

$$\textbf{Output:}\ \ z \sim \mathbf{q}^{(s)}(z) = \prod_i \text{Bernoulli}(\pi_i^{(s)})$$



Statistical Guarantee

Theorem

- Signal-to-noise Ratio: $I = \text{Rényi}_{\frac{1}{2}}(\text{Bernoulli}(p^*), \text{Bernoulli}(q^*))$

Statistical Guarantee

Theorem

- Signal-to-noise Ratio: $I \asymp (p^* - q^*)^2/(p^* + q^*)$

Statistical Guarantee

Theorem

- Signal-to-noise Ratio: $I \asymp (p^* - q^*)^2/(p^* + q^*)$
- SNR Strength: $nI \to \infty$

Statistical Guarantee

Theorem

- Signal-to-noise Ratio: $I \asymp (p^* - q^*)^2/(p^* + q^*)$
- SNR Strength: $nI \to \infty$
- Initializer: $\mathbb{E}_{z \sim \mathbf{q}^{(0)}(z)}\left[\ell(z, z^*)\right] < c$ for some constant $c > 0$

Statistical Guarantee

Theorem

- Signal-to-noise Ratio: $I \asymp (p^* - q^*)^2/(p^* + q^*)$
- SNR Strength: $nI \to \infty$
- Initializer: $\mathbb{E}_{z \sim \mathbf{q}^{(0)}(z)}[\ell(z, z^*)] < c$ for some constant $c > 0$

With high probability, for ALL iterations $s \geq 1$:

$$\mathbb{E}_{z \sim \mathbf{q}^{(s)}(z)}[\ell(z, z^*)|A] \leq \exp\left(-(1 - o(1))\frac{nI}{2}\right) + \left[\frac{1}{\sqrt{nI}}\right]^s E_{z \sim \mathbf{q}^{(0)}(z)}[\ell(z, z^*)|A].$$

optimal rate

linear convergence

Running Time Upper Bound

Corollary

For $s \geq \log n$:

$$\mathbb{E}_A \left[\mathbb{E}_{z \sim \mathbf{q}^{(s)}(z)} \left(\ell(z, z^*) \geq \exp\left(-(1 - o(1)) \frac{nI}{2} \right) \middle| A \right) \right] \to 0.$$

- Remark 1: Rate-optimal. *[Z. & Zhou, Annals of Statistics, 2016]*
- Remark 2: Practical. Spectral Clustering, SDP

Running Time Upper Bound

Corollary

For $s \geq \log n$:

$$\mathbb{E}_A \left[\mathbb{E}_{z \sim \mathbf{q}^{(s)}(z)} \left(\ell(z, z^*) \geq \exp\left(-(1 - o(1)) \frac{nI}{2} \right) \middle| A \right) \right] \to 0.$$

- Remark 3: When p^*, q^* known, $\mathbb{E}_{z \sim \mathbf{q}^{(0)}(z)} [\ell(z, z^*)] < \frac{1}{2} - \frac{1}{(nI)^{\frac{1}{2} - \delta}}$
- Remark 4: $nI \to \infty$ sufficent and necessary condition for consistency

- Motivating Examples
- Mean Field Variational Inference
- Guarantees of Mean Field Variational Inference on Community Detection
- Three Siblings: Mean Field, Gibbs Sampler, EM

Three Siblings

Similarity: Coordinate updates with $\mathbf{p}(\theta_i|\theta_{-i}, x)$.

Three Siblings

Similarity: Coordinate updates with $\mathbf{p}(\theta_i|\theta_{-i}, x)$.

Mean Field Variational Inference	Gibbs Sampler	Expectation Maximization
For all $i \in [n]$, $\exp\{\mathbb{E}_{\mathbf{q}_{-i}}[\log \mathbf{p}(\theta_i|\theta_{-i}, \mathbf{x})]\}$	$\forall i \in [n]$, sample from $\mathbf{p}(\theta_i|\theta_{-i}, \mathbf{x})$	E-Step: *local* variables (e.g., z) $\exp\{\mathbb{E}_{\mathbf{q}_{-i}}[\log \mathbf{p}(\theta_i|\theta_{-i}, \mathbf{x})]\}$ M-Step: *global* variables (e.g., p, q) $\arg\max_{\theta_i} \mathbb{E}_{\mathbf{q}_{-i}}[\log \mathbf{p}(\theta_i|\theta_{-i}, \mathbf{x})]$

Guarantees for Gibbs Sampler

Theorem

- Signal-to-noise Ratio: $I \asymp (p^* - q^*)^2/(p^* + q^*)$
- SNR Strength: $nI \to \infty$
- Initializer: $\ell(z^{(0)}, z^*) < c$ for some constant $c > 0$

With high probability, for ALL iterations $1 \leq s \leq e^n$:

$$\mathbb{E}\left[\ell(z^{(s)}, z^*)|A\right] \leq \exp\left(-(1-o(1))\frac{nI}{2}\right) + \left[\frac{1}{\sqrt{nI}}\right]^s \ell(z^{(0)}, z^*).$$

Guarantees for (a variant of) EM

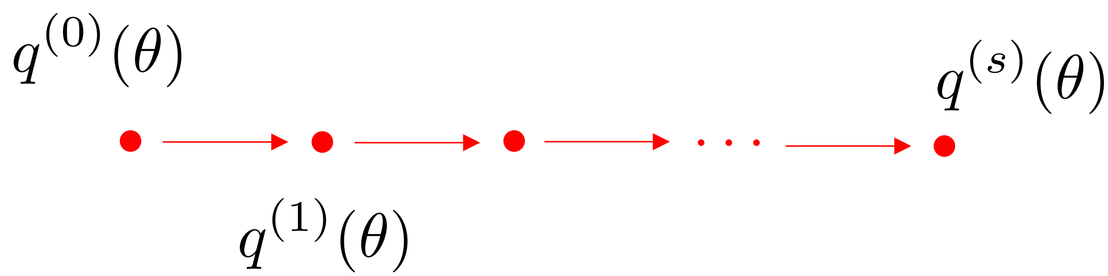
Theorem

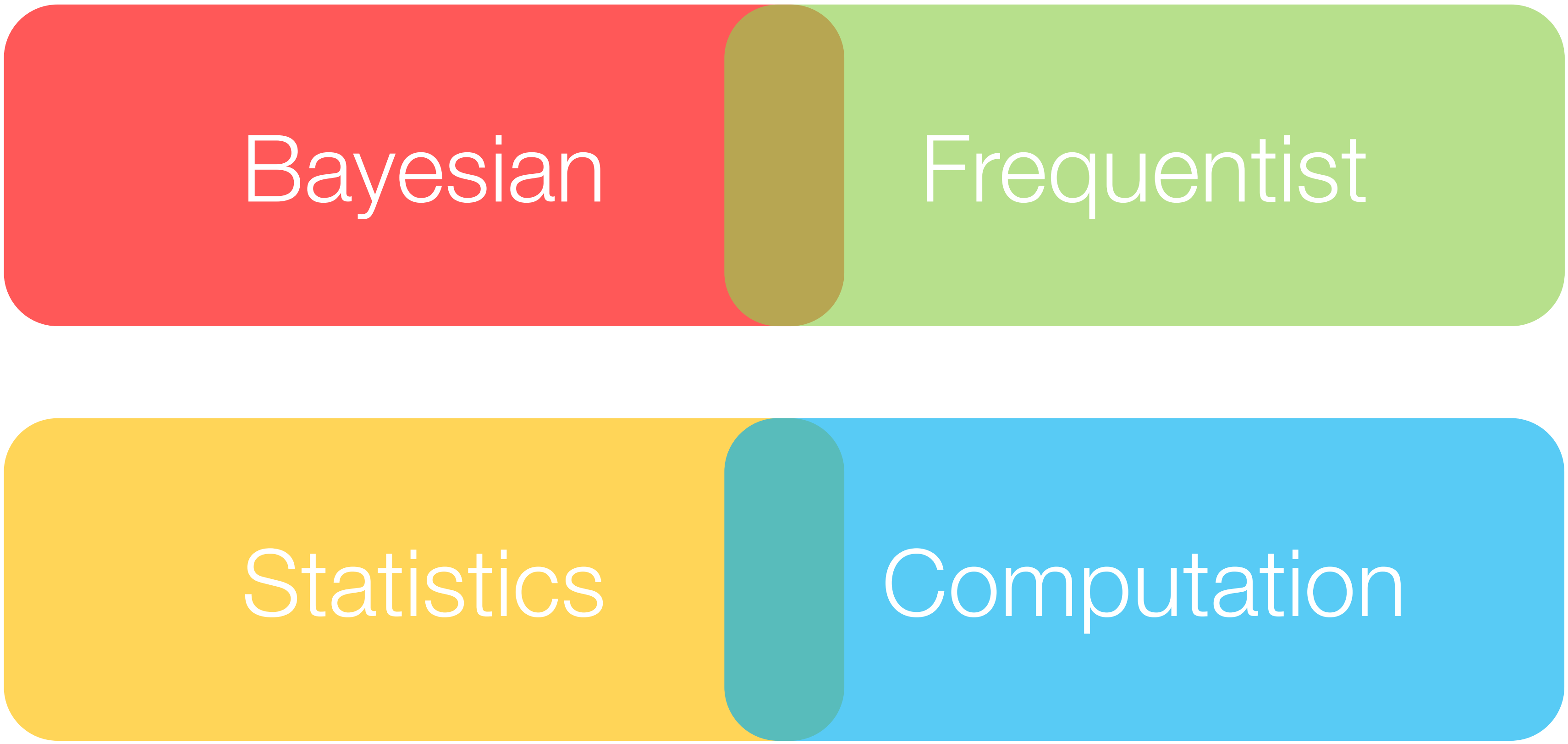
- Signal-to-noise Ratio: $I \asymp (p^* - q^*)^2/(p^* + q^*)$
- SNR Strength: $nI \to \infty$
- Initializer: $\ell_1(\pi^{(0)}, z^*) < c$ for some constant $c > 0$

With high probability, for ALL iterations $s \geq 1$:

$$\ell_1(\pi^{(s)}, z^*) \leq \exp\left(-(1 - o(1))\frac{nI}{2}\right) + \left[\frac{1}{\sqrt{nI}}\right]^s \ell_1(\pi^{(0)}, z^*).$$

Summary


$q^{(0)}(\theta)$
$q^{(1)}(\theta)$
$q^{(s)}(\theta)$


Bayesian
Frequentist
Statistics
Computation

Thank You

☞ Anderson Zhang & Harrison Zhou. "Theoretical and computational guarantees of mean field variational inference for community detection". *Annals of Statistics.* 48.5 (2020): 2575-2598

<u>Related Reference:</u>

☞ Zhang, Anderson Y., and Harrison H. Zhou. "Minimax rates of community detection in stochastic block models". *The Annals of Statistics* 44.5 (2016): 2252-228

☞ Gao, C., Ma, Z., Zhang, A., and Zhou, H. "Achieving Optimal Misclassification Proportion in Stochastic Block Model". *Journal of Machine Learning Research*. 18.60 (2017): 1-45

Proof (High-level Idea)

$$\mathbb{E}_{z\sim \mathbf{q}^{(s)}(z)}\left[\ell(z,z^*)|A\right] \leq \exp\left(-(1-o(1))\frac{nI}{2}\right) + \left[\frac{1}{\sqrt{nI}}\right]^s E_{z\sim \mathbf{q}^{(0)}(z)}\left[\ell(z,z^*)|A\right].$$

Proof (High-level Idea)

$$\mathbb{E}_{z \sim \mathbf{q}^{(s)}(z)}[\ell(z, z^*)|A] \leq \exp\left(-(1 - o(1))\frac{nI}{2}\right) + \left[\frac{1}{\sqrt{nI}}\right]^s E_{z \sim \mathbf{q}^{(0)}(z)}[\ell(z, z^*)|A].$$

- Assume p^*, q^* known

Proof (High-level Idea)

$$\mathbb{E}_{z\sim\mathbf{q}^{(s)}(z)}\left[\ell(z,z^*)|A\right] \leq \exp\left(-(1-o(1))\frac{nI}{2}\right) + \left[\frac{1}{\sqrt{nI}}\right]^s E_{z\sim\mathbf{q}^{(0)}(z)}\left[\ell(z,z^*)|A\right].$$

- Assume p^*, q^* known
- $z \sim \mathbf{q}^{(s)}(z) = \prod_i \text{Bernoulli}(\pi_i^{(s)})$ $\quad \pi^{(0)} \to \pi^{(1)} \to \ldots \to \pi^{(s)}$

Proof (High-level Idea)

$$\mathbb{E}_{z\sim\mathbf{q}^{(s)}(z)}\left[\ell(z,z^*)|A\right] \leq \exp\left(-(1-o(1))\frac{nI}{2}\right) + \left[\frac{1}{\sqrt{nI}}\right]^s E_{z\sim\mathbf{q}^{(0)}(z)}\left[\ell(z,z^*)|A\right].$$

- Assume p^*, q^* known
- $z \sim \mathbf{q}^{(s)}(z) = \prod_i \text{Bernoulli}(\pi_i^{(s)})$ $\quad \pi^{(0)} \to \pi^{(1)} \to \ldots \to \pi^{(s)}$
- $\pi \to \pi^{\text{new}}$

Proof (High-level Idea)

$$\mathbb{E}_{z\sim\mathbf{q}^{(s)}(z)}\left[\ell(z,z^*)|A\right] \le \exp\left(-(1-o(1))\frac{nI}{2}\right) + \left[\frac{1}{\sqrt{nI}}\right]^s E_{z\sim\mathbf{q}^{(0)}(z)}\left[\ell(z,z^*)|A\right].$$

- Assume p^*, q^* known
- $z \sim \mathbf{q}^{(s)}(z) = \prod_i \text{Bernoulli}(\pi_i^{(s)})$ $\quad \pi^{(0)} \to \pi^{(1)} \to \ldots \to \pi^{(s)}$
- $\pi \to \pi^{\text{new}}$

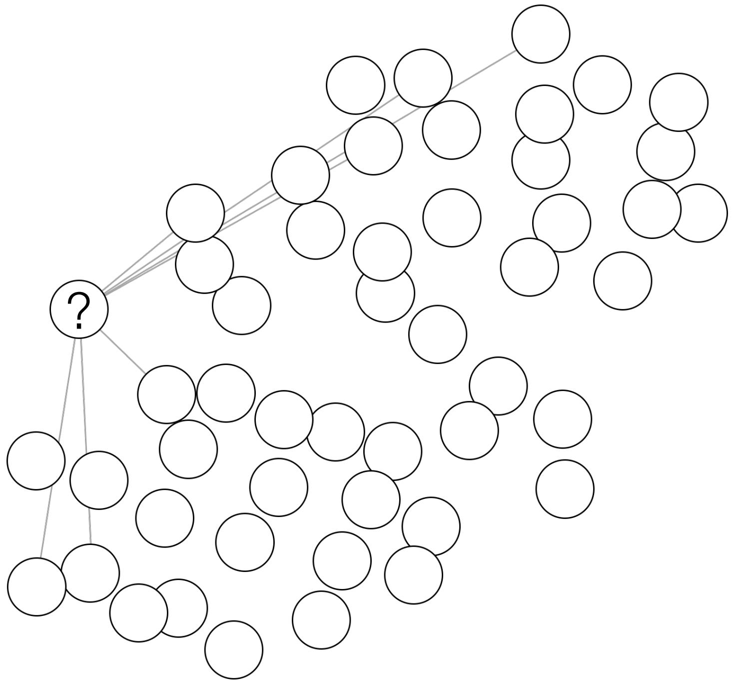
$$\begin{aligned}
\pi_1^{\text{new}} &\propto \exp\{\mathbb{E}_{z_2,\ldots,z_n}[\log p(z_1|z_2,\ldots,z_n,A)]\} \\
&= \exp\{\mathbb{E}_{z_2,\ldots,z_n}[\log p(z_1|z_2,\ldots,z_n,A_{1,\cdot})]\} \\
&= f(\pi_2,\ldots,\pi_n,A_{1,\cdot})
\end{aligned}$$

Proof (High-level Idea)

$$\mathbb{E}_{z\sim\mathbf{q}^{(s)}(z)}\left[\ell(z,z^*)|A\right] \leq \exp\left(-(1-o(1))\frac{nI}{2}\right) + \left[\frac{1}{\sqrt{nI}}\right]^s E_{z\sim\mathbf{q}^{(0)}(z)}\left[\ell(z,z^*)|A\right].$$

- Assume p^*, q^* known
- $z \sim \mathbf{q}^{(s)}(z) = \prod_i \text{Bernoulli}(\pi_i^{(s)})$ $\quad \pi^{(0)} \to \pi^{(1)} \to \ldots \to \pi^{(s)}$
- $\pi \to \pi^{\text{new}}$

$$\begin{aligned}\pi_1^{\text{new}} &\propto \exp\{\mathbb{E}_{z_2,\ldots,z_n}[\log p(z_1|z_2,\ldots,z_n,A)]\} \\ &= \exp\{\mathbb{E}_{z_2,\ldots,z_n}[\log p(z_1|z_2,\ldots,z_n,A_{1,\cdot})]\} \\ &= f(\pi_2,\ldots,\pi_n,A_{1,\cdot})\end{aligned}$$

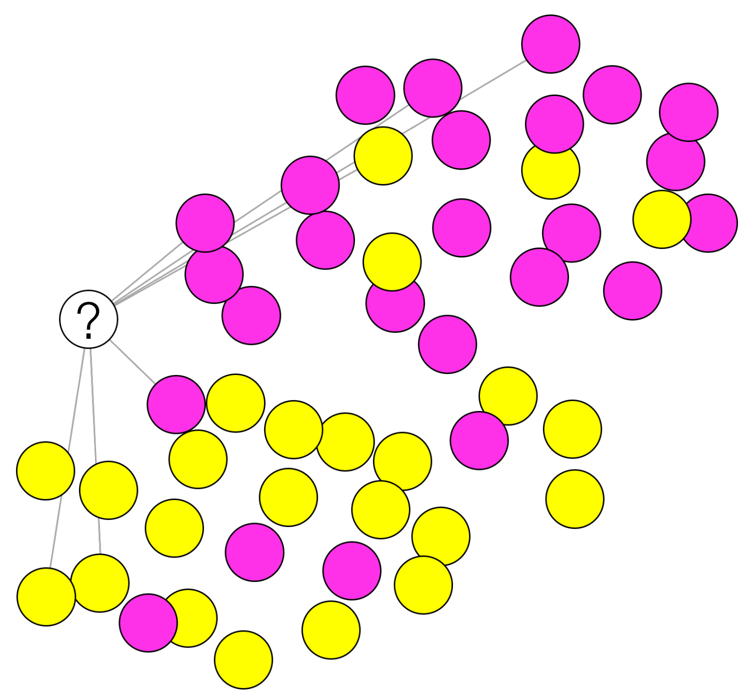


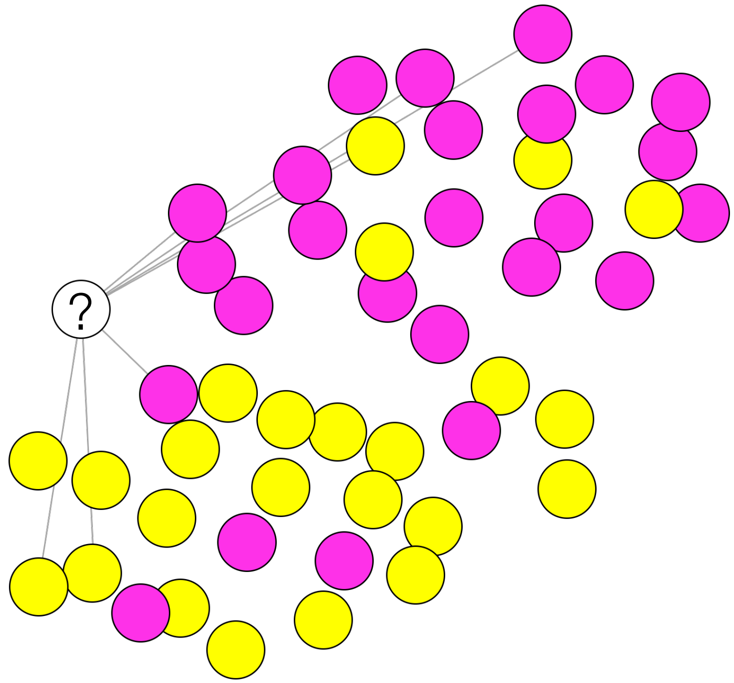
Proof (High-level Idea)

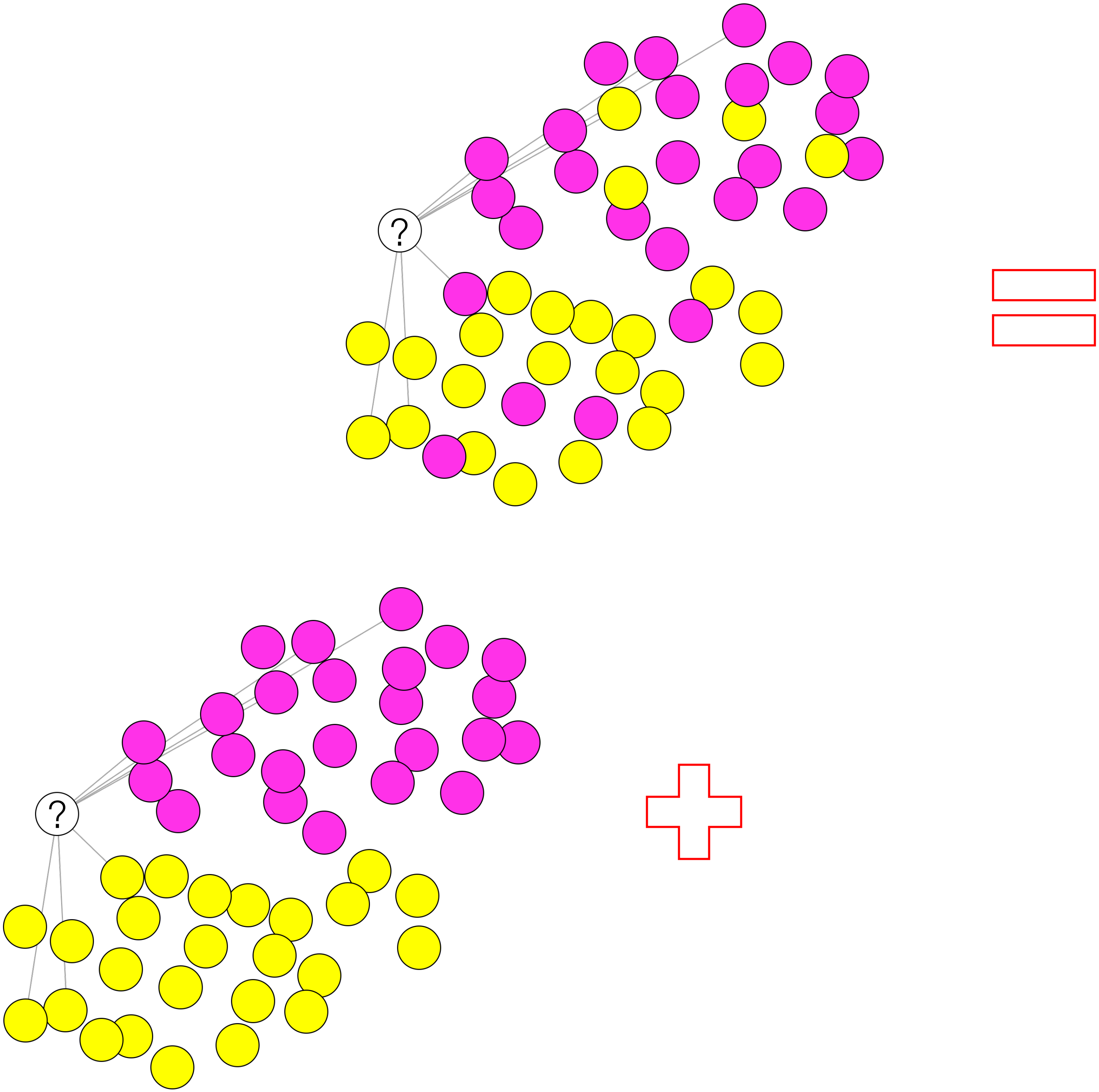
$$\mathbb{E}_{z\sim\mathbf{q}^{(s)}(z)}\left[\ell(z,z^*)|A\right] \le \exp\left(-(1-o(1))\frac{nI}{2}\right) + \left[\frac{1}{\sqrt{nI}}\right]^s E_{z\sim\mathbf{q}^{(0)}(z)}\left[\ell(z,z^*)|A\right].$$

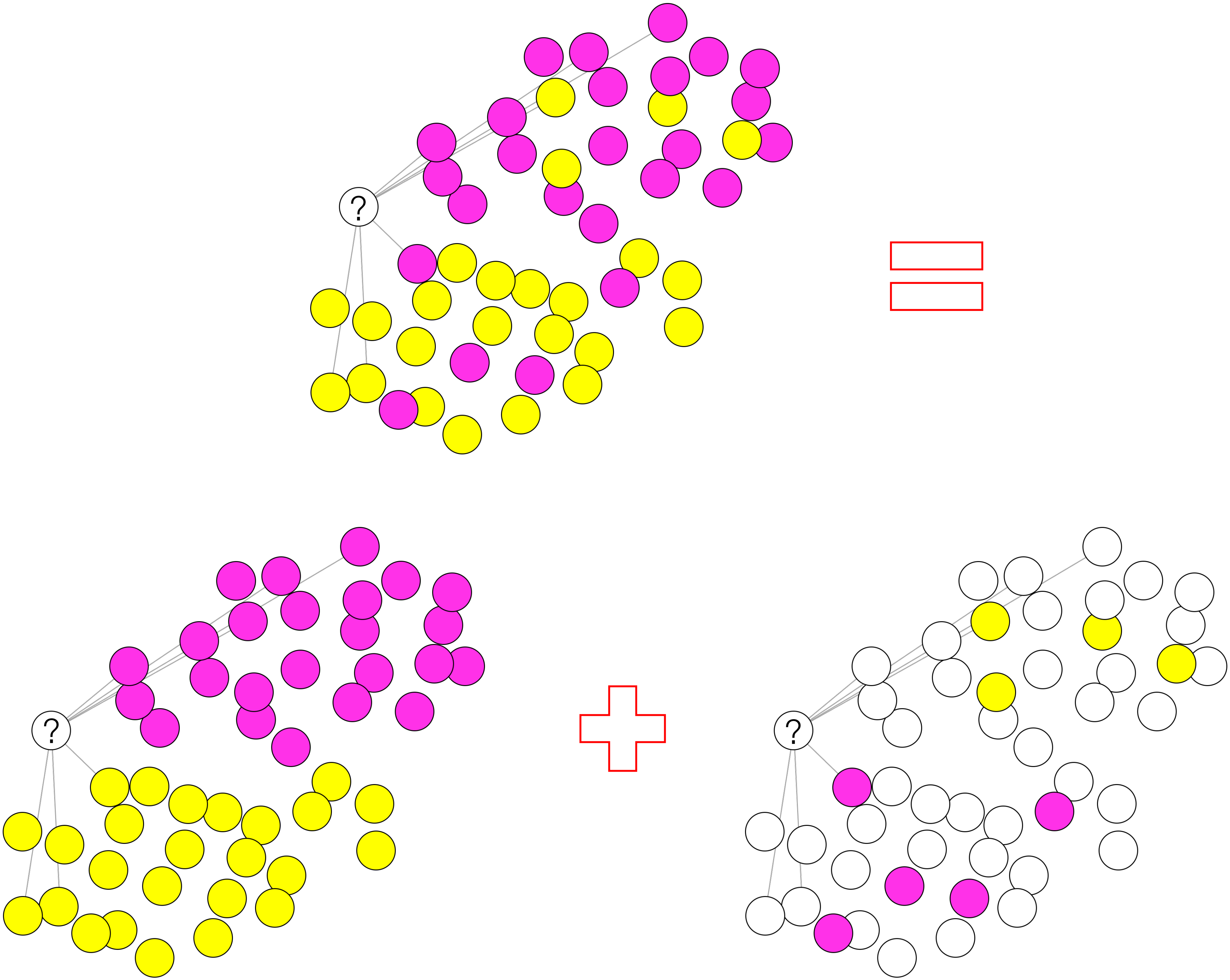
- Assume p^*, q^* known
- $z \sim \mathbf{q}^{(s)}(z) = \prod_i \text{Bernoulli}(\pi_i^{(s)})$ $\quad \pi^{(0)} \to \pi^{(1)} \to \ldots \to \pi^{(s)}$
- $\pi \to \pi^{\text{new}}$

$$\begin{aligned}\pi_1^{\text{new}} &\propto \exp\{\mathbb{E}_{z_2,\ldots,z_n}[\log p(z_1|z_2,\ldots,z_n,A)]\} \\ &= \exp\{\mathbb{E}_{z_2,\ldots,z_n}[\log p(z_1|z_2,\ldots,z_n,A_{1,\cdot})]\} \\ &= f(\pi_2,\ldots,\pi_n,A_{1,\cdot})\end{aligned}$$









Proof (High-level Idea)

$$
\begin{aligned}
|\pi_1^{\text{new}} - z_1^*| &= |f(\pi_2, \ldots, \pi_n, A_{1,\cdot}) - z_1^*| \\
&\leq |f(z_2^*, \ldots, z_n^*, A_{1,\cdot}) - z_1^*| + |f(\pi_2, \ldots, \pi_n, A_{1,\cdot}) - f(z_2^*, \ldots, z_n^*, A_{1,\cdot})|
\end{aligned}
$$

Proof (High-level Idea)

$$|\pi_1^{\text{new}} - z_1^*| = |f(\pi_2, \ldots, \pi_n, A_{1,\cdot}) - z_1^*|$$
$$\leq |f(z_2^*, \ldots, z_n^*, A_{1,\cdot}) - z_1^*| + |f(\pi_2, \ldots, \pi_n, A_{1,\cdot}) - f(z_2^*, \ldots, z_n^*, A_{1,\cdot})|$$

optimal rate — *deviation between* π & z^*

Proof (High-level Idea)

$$
\begin{aligned}
|\pi_1^{\text{new}} - z_1^*| &= |f(\pi_2, \ldots, \pi_n, A_{1,\cdot}) - z_1^*| \\
&\leq |f(z_2^*, \ldots, z_n^*, A_{1,\cdot}) - z_1^*| + |f(\pi_2, \ldots, \pi_n, A_{1,\cdot}) - f(z_2^*, \ldots, z_n^*, A_{1,\cdot})|
\end{aligned}
$$

optimal rate   *deviation between* π & z^*

$$
\mathbb{E}_{z\sim \mathbf{q}_{\pi^{\text{new}}}}\left[\ell(z, z^*)|A\right] \leq \exp\left(-(1-o(1))\frac{nI}{2}\right) + \frac{1}{\sqrt{nI}} E_{z\sim \mathbf{q}_{\pi}}\left[\ell(z, z^*)|A\right].
$$

Proof (High-level Idea)

$$
\begin{aligned}
|\pi_1^{\text{new}} - z_1^*| &= |f(\pi_2, \ldots, \pi_n, A_{1,\cdot}) - z_1^*| \\
&\leq |f(z_2^*, \ldots, z_n^*, A_{1,\cdot}) - z_1^*| + |f(\pi_2, \ldots, \pi_n, A_{1,\cdot}) - f(z_2^*, \ldots, z_n^*, A_{1,\cdot})|
\end{aligned}
$$

optimal rate *deviation between* π & z^*

$$
\mathbb{E}_{z\sim\mathbf{q}_{\pi^{\text{new}}}}\left[\ell(z, z^*)|A\right] \leq \exp\left(-(1-o(1))\frac{nI}{2}\right) + \frac{1}{\sqrt{nI}} E_{z\sim\mathbf{q}_{\pi}}\left[\ell(z, z^*)|A\right].
$$

⇓

$$
\boxed{\mathbb{E}_{z\sim\mathbf{q}^{(s)}(z)}\left[\ell(z, z^*)|A\right] \leq \exp\left(-(1-o(1))\frac{nI}{2}\right) + \left[\frac{1}{\sqrt{nI}}\right]^s E_{z\sim\mathbf{q}^{(0)}(z)}\left[\ell(z, z^*)|A\right].}
$$