Statistical modeling with stochastic processes

Alexandre Bouchard-Côté

Lecture 5, Monday March 14

Program for today

- Wrapping up Variational methods
  - Examples: Mean field and Belief propagation
  - Theoretical framework
- Introduction to Bayesian non-parametrics
  - The Dirichlet Process: Theoretical foundations

Review

Precisions from last time

Importance sampling vs. Independence chain

Theoretical work: ‘Comparing Importance Sampling and the Metropolis Algorithm’ Federico Bassetti and Persi Diaconis

‘It follows that importance sampling and the Metropolis algorithm are roughly comparable for this example.’

Empirical comparison: ??

Find the potential bugs

for $t = 1$ .. T

Pick a kernel $q = q_\alpha$, $\alpha \sim M(x_{t-1})$

Loop....

1. Propose a new state x_{prop} according to $q(\ x \mid x_{t-1})$
2. Compute:

$$A(x_{t-1} \to x_{\text{prop}}) = \min\left\{1, \frac{\text{target}(x_{t-1})q(x_{\text{prop}}|x_{t-1})}{\text{target}(}\right\}$$

3. Generate a Unif[0,1] number u

.... while $u > A(x_{t-1} \to x_{prop})$

Set x_t to x_{prop}

- Ratio is upside down !
- Mixing of kernels distribution should not depend on x
- No while loop !
(c.f. rejection sampling)
- Inequality is reversed!

Variational inference

Quick review of exponential family

Sufficient statistic

Parameter

$$\mathbb{P}(\boldsymbol{X}_{\boldsymbol{\theta}} \in B) = \sum_{x \in B} \exp\{\langle \boldsymbol{\phi}(x), \boldsymbol{\theta} \rangle - A(\boldsymbol{\theta})\}\nu(x),$$

$$A(\boldsymbol{\theta}) = \log \sum_{x \in \boldsymbol{\mathcal{X}}} \exp\{\langle \boldsymbol{\phi}(x), \boldsymbol{\theta} \rangle\}\nu(x),$$

Log partition function

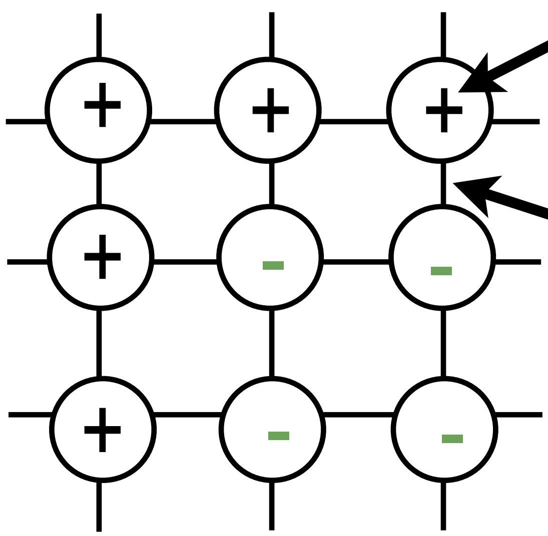
Large discrete set
(e.g. all configs of an Ising model)

A counting measure

Example of sufficient statistics

Ising model

One node



Pairs of nodes

$$\phi(x) = \begin{bmatrix} \mathbf{1}[x_{1,1} = +] \\ \mathbf{1}[x_{1,1} = -] \\ \mathbf{1}[x_{1,2} = +] \\ \vdots \\ \mathbf{1}[x_{1,1} = +, x_{1,2} = +] \\ \mathbf{1}[x_{1,1} = +, x_{1,2} = -] \\ \vdots \end{bmatrix} \begin{bmatrix} \theta_{1,1,+} \\ \theta_{1,1,-} \\ \theta_{1,2,+} \\ \vdots \\ \theta_{1,1,+;1,2,+} \\ \theta_{1,1,+;1,2,-} \\ \vdots \end{bmatrix} D$$

'Over-complete' sufficient statistic

What we are trying to compute

Moments:

$$\mu = \mathrm{E}[\phi(X)] = \begin{bmatrix} \mu_{1,1,+} \\ \mu_{1,1,-} \\ \mu_{1,2,+} \\ \vdots \\ \mu_{1,1,+;1,2,+} \\ \mu_{1,1,+;1,2,-} \\ \vdots \end{bmatrix} \begin{bmatrix} \mathbf{1}[x_{1,1} = +] \\ \mathbf{1}[x_{1,1} = -] \\ \mathbf{1}[x_{1,2} = +] \\ \vdots \\ \mathbf{1}[x_{1,1} = +, x_{1,2} = +] \\ \mathbf{1}[x_{1,1} = +, x_{1,2} = -] \\ \vdots \end{bmatrix} \begin{bmatrix} \theta_{1,1,+} \\ \theta_{1,1,-} \\ \theta_{1,2,+} \\ \vdots \\ \theta_{1,1,+;1,2,+} \\ \theta_{1,1,+;1,2,-} \\ \vdots \end{bmatrix} \quad D$$

and *log partition function*: $A(\boldsymbol{\theta}) = \log \sum_{x \in \boldsymbol{\mathcal{X}}} \exp\{\langle \boldsymbol{\phi}(x), \boldsymbol{\theta} \rangle\} \nu(x)$

Important properties

The gradient of the log partition function is equal to the moments:

$$\nabla A(\boldsymbol{\theta}) = \mathbb{E}[\phi(\boldsymbol{X}_{\boldsymbol{\theta}})]$$

The hessian of the log partition function is equal to the covariance matrix:

$$H(A(\boldsymbol{\theta})) = \mathbf{V}ar[\phi(\boldsymbol{X}_{\boldsymbol{\theta}})].$$

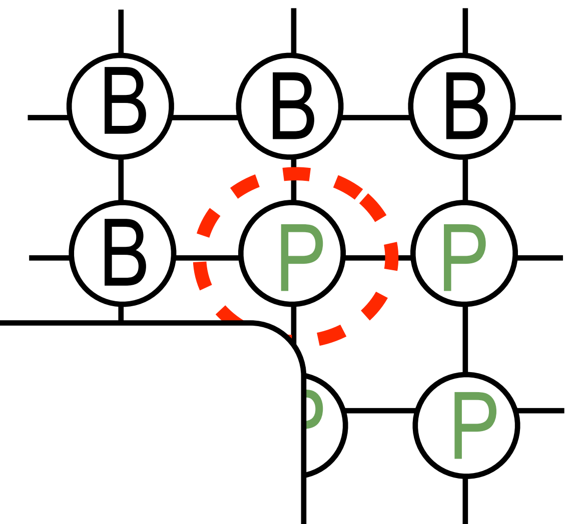
Consequence: A is a convex function

Road map

Hard probabilistic inference problems

1

Deterministic algorithms



$\mathrm{target}(x) = \mathbb{P}(X = x | \mathrm{obs}, \ \mathrm{params})$

2

Two examples:
- Mean field
- Loopy Belief Propagation

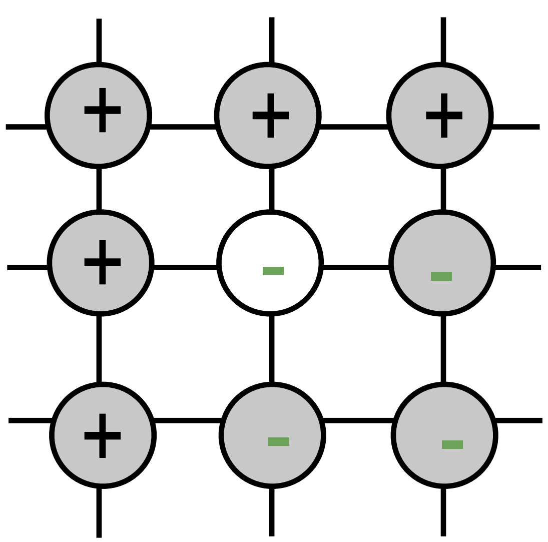
$\lambda_{s_1}(A) =$

$\lambda_{s_1,s_2}(A_1, A_2) = \lambda_{s_2,s_1}(A_2, A_1)$

Probabilistic inference as an optimization problem

Naive mean-field and connection with Gibbs

First: let rewrite the Gibbs sampler with the exponential family notation



$$\phi(x) = \begin{bmatrix} \mathbf{1}[x_{1,1} = +] \\ \mathbf{1}[x_{1,1} = \text{-}] \\ \mathbf{1}[x_{1,2} = +] \\ \vdots \\ \mathbf{1}[x_{1,1} = +, x_{1,2} = +] \\ \mathbf{1}[x_{1,1} = +, x_{1,2} = \text{-}] \\ \vdots \end{bmatrix} \begin{bmatrix} \theta_{1,1,+} \\ \theta_{1,1,\text{-}} \\ \theta_{1,2,+} \\ \vdots \\ \theta_{1,1,+;1,2,+} \\ \theta_{1,1,+;1,2,\text{-}} \\ \vdots \end{bmatrix} \updownarrow D$$

Mean field

Note: Gibbs can be seen in this case as keeping around one vector $s_t = \phi(x_t)$ at each iteration (where each component of s is in the set {0, 1}

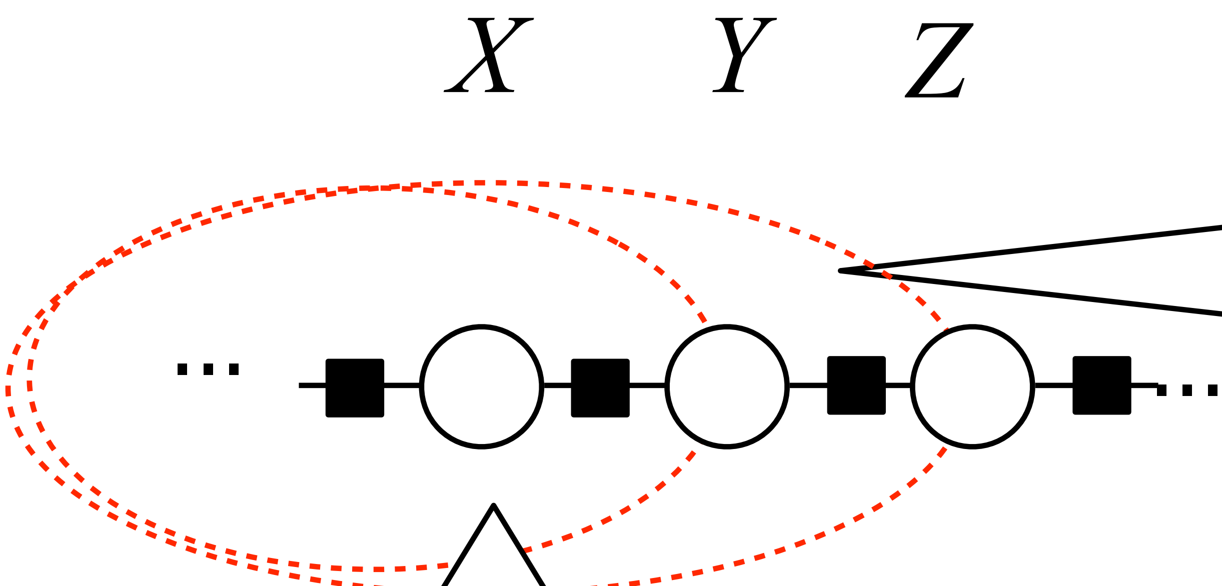
Idea: we are going to keep around one vector μ_t where each component of μ_t is in the set [0, 1], and hope that this gives a good approximation to

$$\mu = \mathrm{E}[\phi(X)]$$

$$\phi(x) = \begin{bmatrix} \mathbf{1}[x_{1,1} = +] \\ \mathbf{1}[x_{1,1} = \text{-}] \\ \mathbf{1}[x_{1,2} = +] \\ \vdots \\ \mathbf{1}[x_{1,1} = +, x_{1,2} = +] \\ \mathbf{1}[x_{1,1} = +, x_{1,2} = \text{-}] \\ \vdots \end{bmatrix}$$

Loopy Belief Propagation (BP)

Looking back at exact inference on a chain/tree:



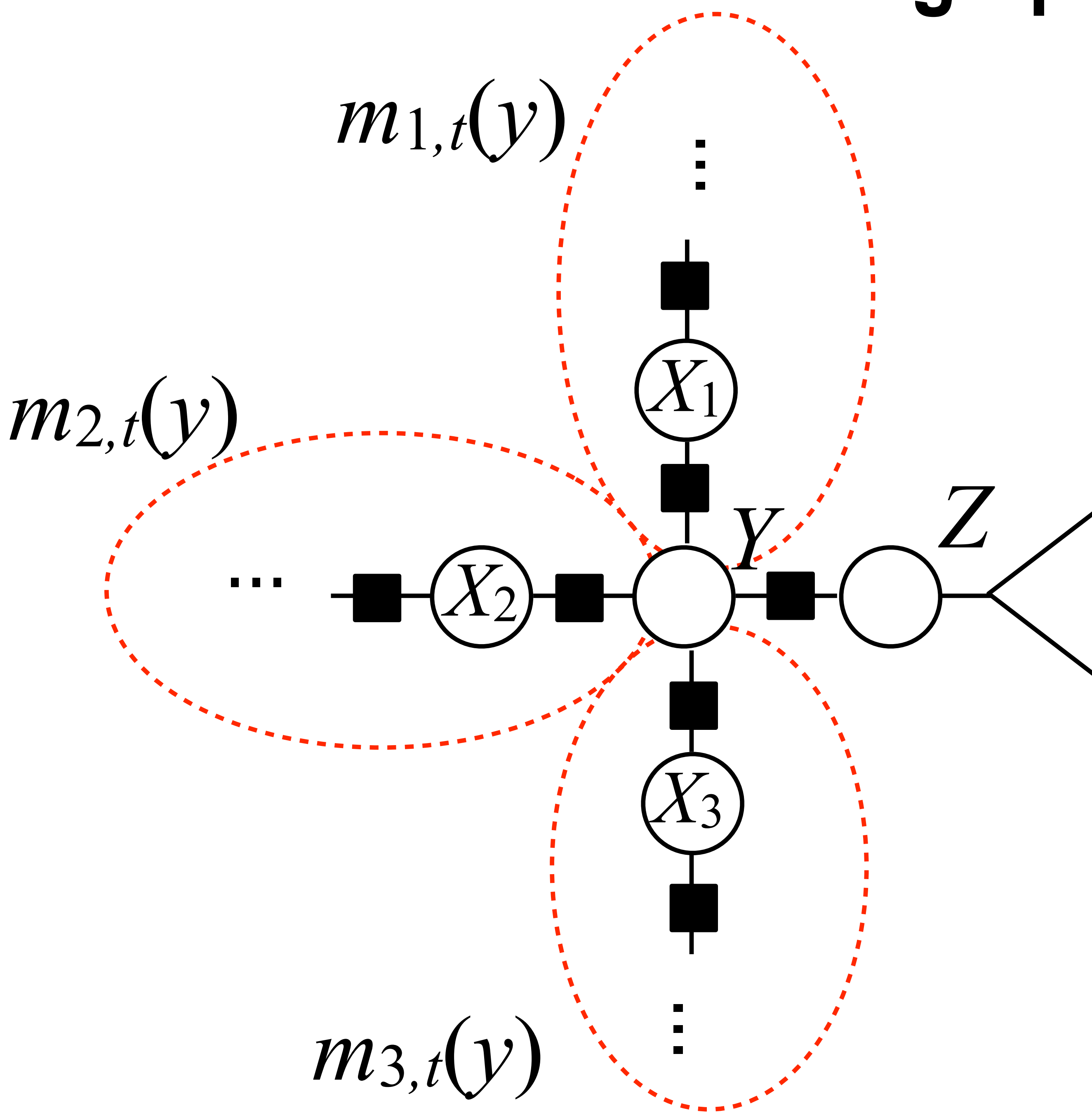
What would be the next message, $m_{t+1}(z)$ that Y would send to the node Z at the right of it?

Use the notation $f(y,z)$ for the factor between Y and Z

View the process of eliminating all the variable at the left of X as a message sent from X to Y: $m_t(y)$

Loopy Belief Propagation (BP)

Idea: do this even if the graph is not a chain/tree



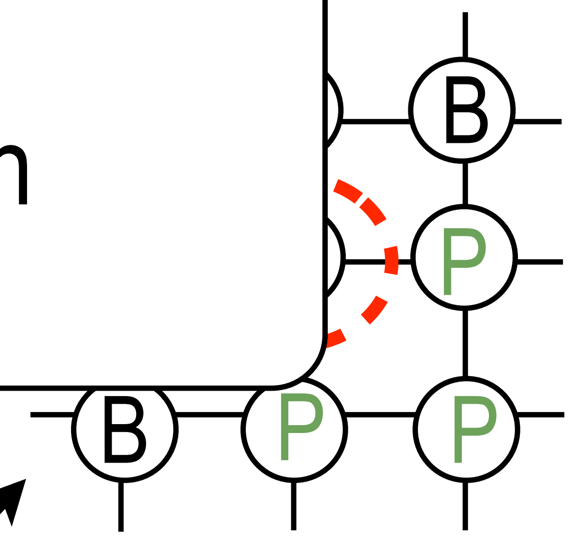
What would be the next message, $m_{t+1}(z)$ that Y would send to the node Z at the right of it?

Using the notation $f(y,z)$ for the factors

Road map

Deterministic

thms



Hard probabilistic inference problems

$\text{target}(x) = \mathbb{P}(X = x|\text{obs, para}$

Next step: expressing the inference tasks as a constrained optimization problem

2

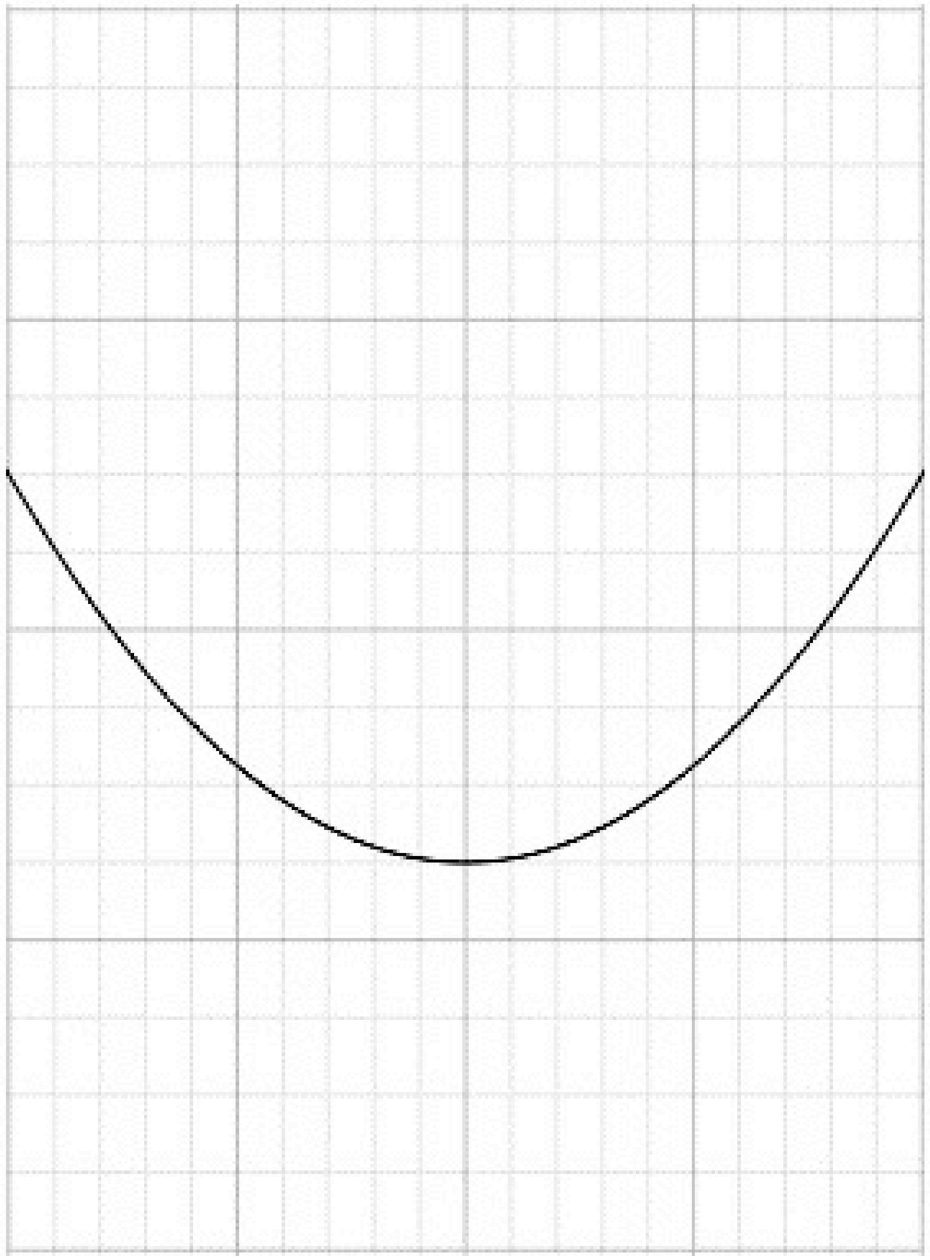
3

$$\lambda_{s_1}(A) = \lambda_{s_1,s_2}(A, \mathbf{R})$$ [marginalization]

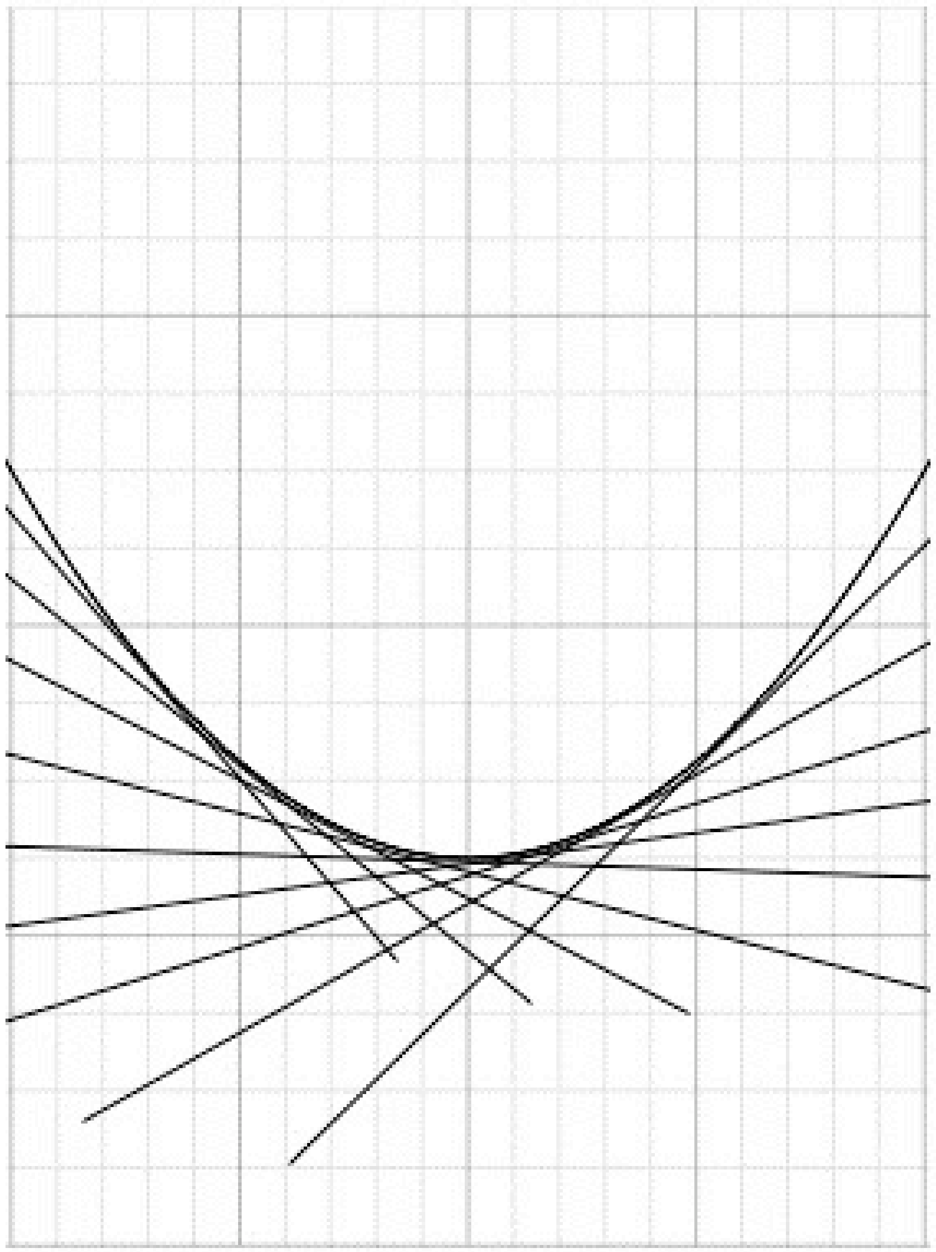
$$\lambda_{s_1,s_2}(A_1, A_2) = \lambda_{s_2,s_1}(A_2, A_1)$$

Probabilistic inference as an optimization problem

Representation of convex functions



Standard / pointwise encoding



Encoded by intercepts of the supporting tangents

Connexion: Legendre-Fenchel transformation

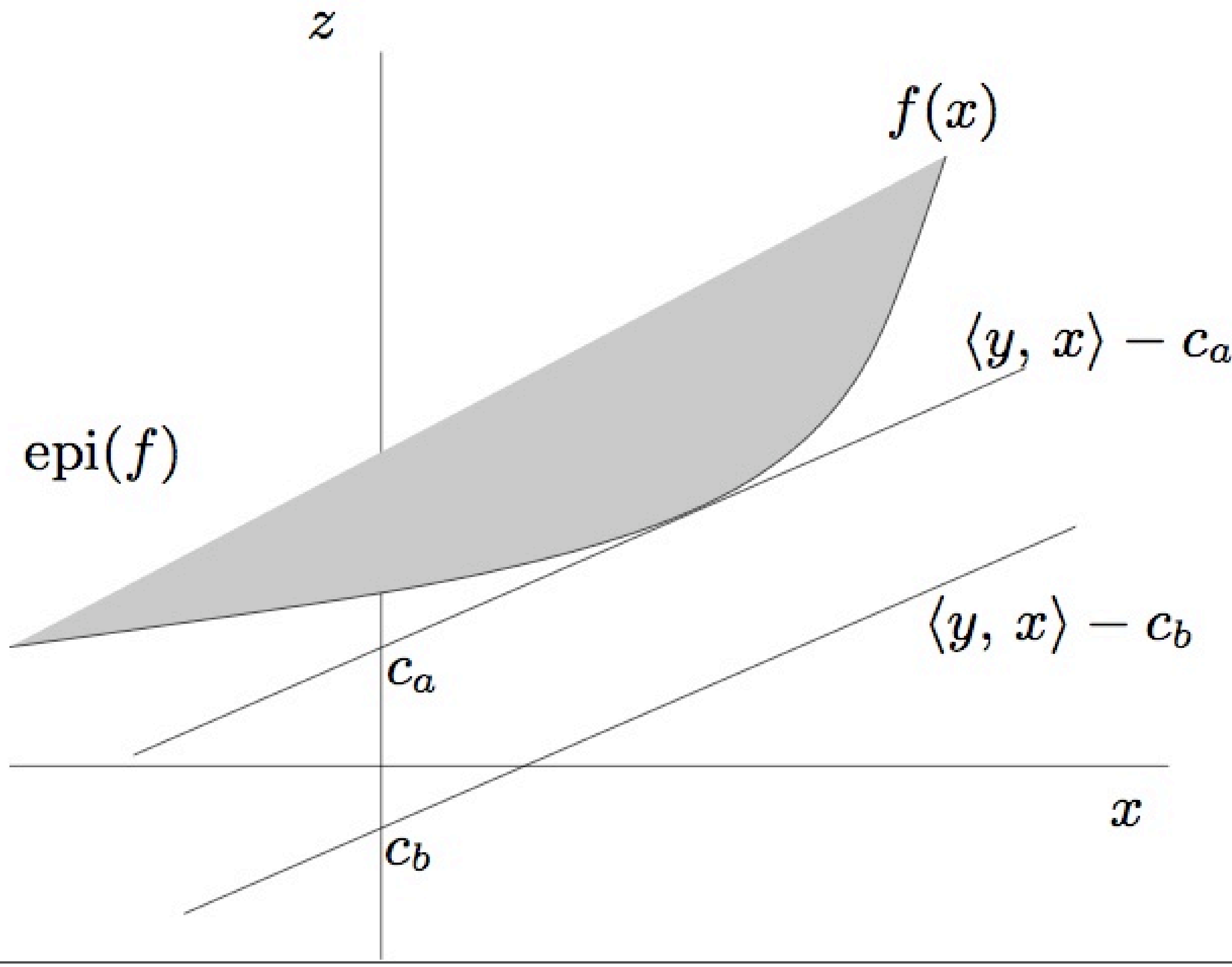
An operator (a function that takes a function and transforms it into another function) denoted by *

$$f^*(y) := \sup_{x \in \mathrm{dom}(f)} \left\{ \langle y, x \rangle - f(x) \right\},$$

Warning: for pedagogical reasons, assume for now that f is univariate, twice differentiable and strictly convex (can be made more general!!)

Intuition

"f acts on points, f^* acts on tangents": Suppose I give you a tangent/supporting plane. Encoding a convex function can be done by giving the intercept c_a



Why this particular ‘encoding’?

Theorem:
When f is convex (and lower semi-continuous): $f^{**} = f$

Consequence: the log partition function satisfies $A^{**} = A$

What we will do with this: First, apply the definition of Fenchel dual to the function A^*, get:

$$A^{**}(\boldsymbol{\theta}) = \sup\{\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - A^*(\boldsymbol{\mu}) : \boldsymbol{\mu} \in \mathscr{M}\},$$

$$||$$

$$A(\theta)$$

This is just the domain of A^*

Done?

Convex function are easy to optimize, right?

$$A(\boldsymbol{\theta}) = \sup\{\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - A^*(\boldsymbol{\mu}) : \boldsymbol{\mu} \in \mathscr{M}\},$$

Problems: there are exponentially many constraints

Constraints: realizable moments

Suppose I give you a D-dimensional vector μ and I claim it is the moment of a distribution for some parameters θ (which I don't give you θ, but the sufficient statistics are known)

I.e. claim there is a θ such that:

$$\mu = \mathrm{E}[\phi(X_\theta)]$$

What could you check?

$$\begin{bmatrix} \mu_{1,1,+} \\ \mu_{1,1,-} \\ \mu_{1,2,+} \\ \vdots \\ \mu_{1,1,+;1,2,+} \\ \mu_{1,1,+;1,2,-} \\ \vdots \end{bmatrix}$$

Constraints: realizable moments

Suppose I give you a D-dimensional vector μ and I claim it is the moment of a distribution for some parameters θ (which I don't give you θ, but the sufficient statistics are known)

I.e. claim there is a θ such that:

$$\mu = \mathrm{E}[\phi(X_\theta)]$$

What could you check?

$$\mu_{1,1,+} = \sum_{x \in \{+,-\}} \mu_{1,1,+;1,2,x}$$

Looks familiar?

$$\begin{bmatrix} \mu_{1,1,+} \\ \mu_{1,1,-} \\ \mu_{1,2,+} \\ \vdots \\ \mu_{1,1,+;1,2,+} \\ \mu_{1,1,+;1,2,-} \\ \vdots \end{bmatrix}$$

Constraints: realizable moments

Theorem: for *trees*, μ is a realizable moment if and only if *pairwise* marginalization conditions are met

In *cyclic* graphs, higher order marginalization constraints needed!

Belief propagation

Main idea: even if there are cycles, use only pairwise marginalization constraints (a relaxation of the optimization problem)

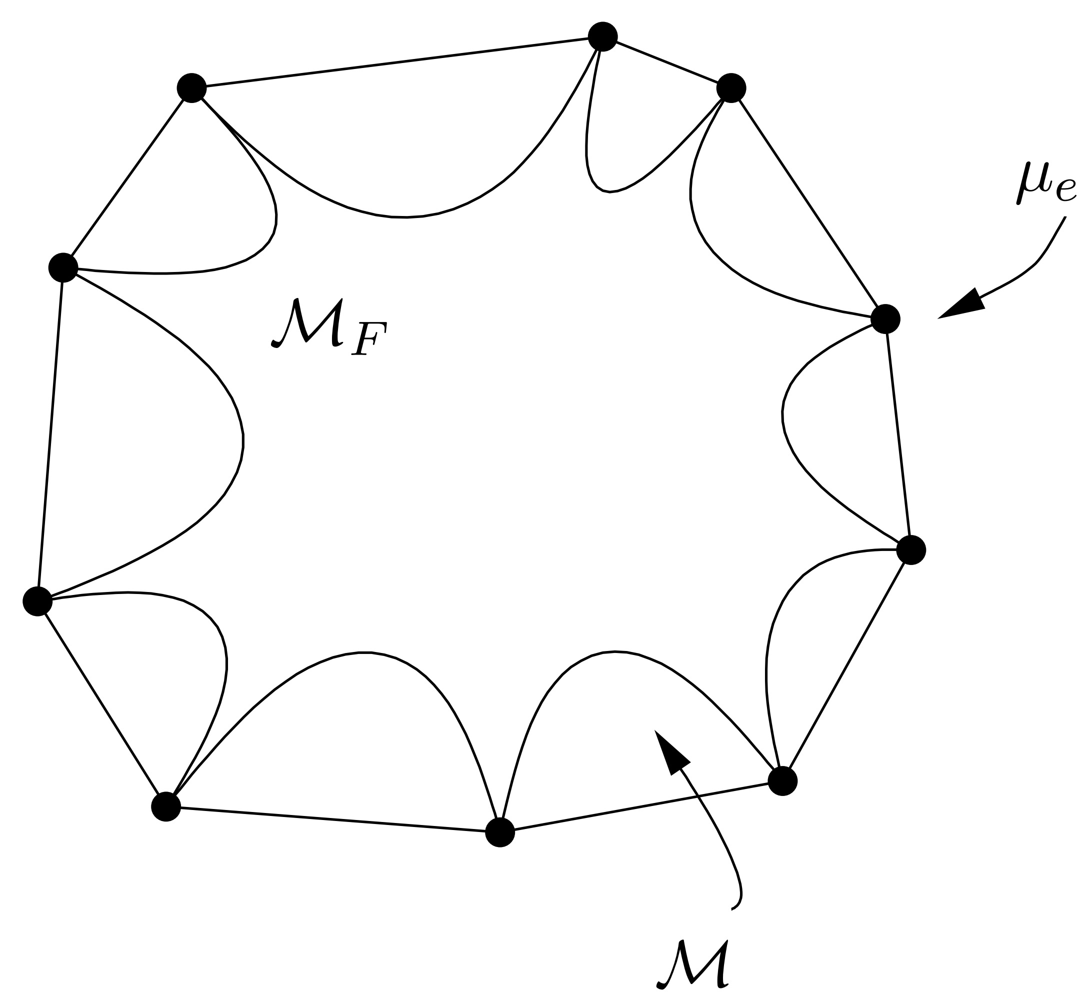
It can be shown that optimizing this relaxed problem yields the familiar BP algorithm
(the objective also needs to be simplified a little bit)

Mean-field: inference by making the set of realizable moments simpler

$$A(\boldsymbol{\theta}) = \sup\{\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - A^*(\boldsymbol{\mu}) : \boldsymbol{\mu} \in \mathscr{M}\},$$

AND: $\boldsymbol{\mu} \in \mathscr{M}_{\mathrm{MF}}$

Cartoon:



Recommended readings

MCMC:

- **Overview of theory and practice: '**Markov chains for exploring posterior distributions.' (1994) L. Tierney.
- **Tricks of the trade:** Part IV of 'Information Theory, Inference, and Learning Algorithms.' (2003) D. MacKay.
- **Fast sampler for Ising model I haven't covered: '**Nonuniversal critical dynamics in Monte Carlo simulations. ' (1987) R.H. Swendsen and J.-S. Wang.
- **Computing partition function from samples:** 'Marginal likelihood from the Gibbs output' (1995) S. Chib.; Also: 'Simulating ratios of normalizing constants via a simple identity: a theoretical exploration' (1996) X.-L. Meng and W.H. Wong.

Recommended readings

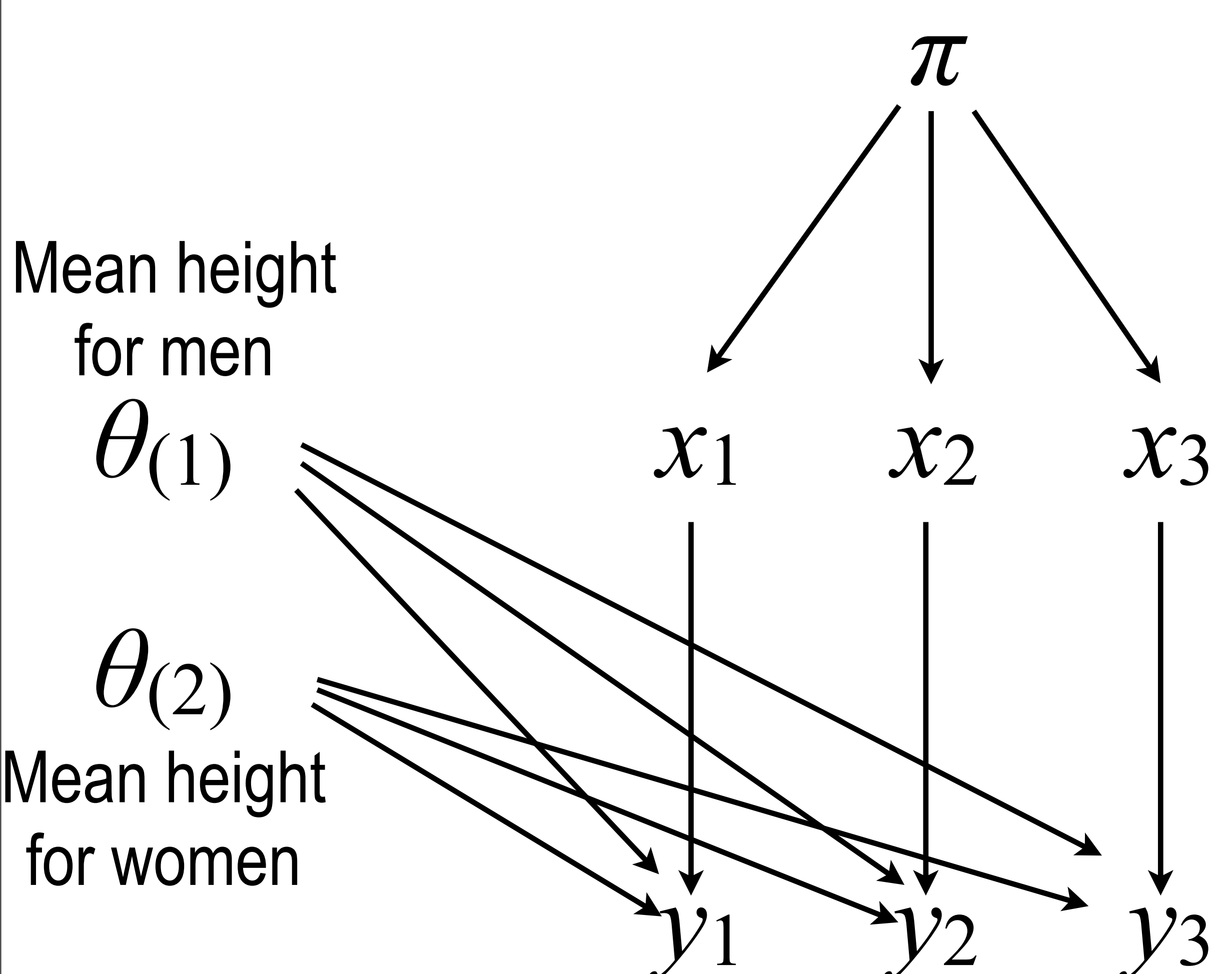
Variational:

- **Overview of theory:** Chapters 1-5 of 'Graphical models, exponential families, and variational inference.' (2008) M. J. Wainwright and M. I. Jordan.
- **More on the Mean Field:** Background section of 'Optimization of Structured Mean Field Objectives'. (2009) A. Bouchard and M.I. Jordan.

Dirichlet Processes

Recall: motivation in density estimation

Mixture model: (UBC student height with 2 components)
say we have only 3 observations



1- Generate a male/female relative frequence

$$\pi \sim \mathrm{Beta}(\text{male prior pseudo counts, female P.C})$$

2- Generate the sex of each student for each i

$$x_i \mid \pi \sim \mathrm{Mult}(\pi)$$

3- Generate the mean height of each cluster c

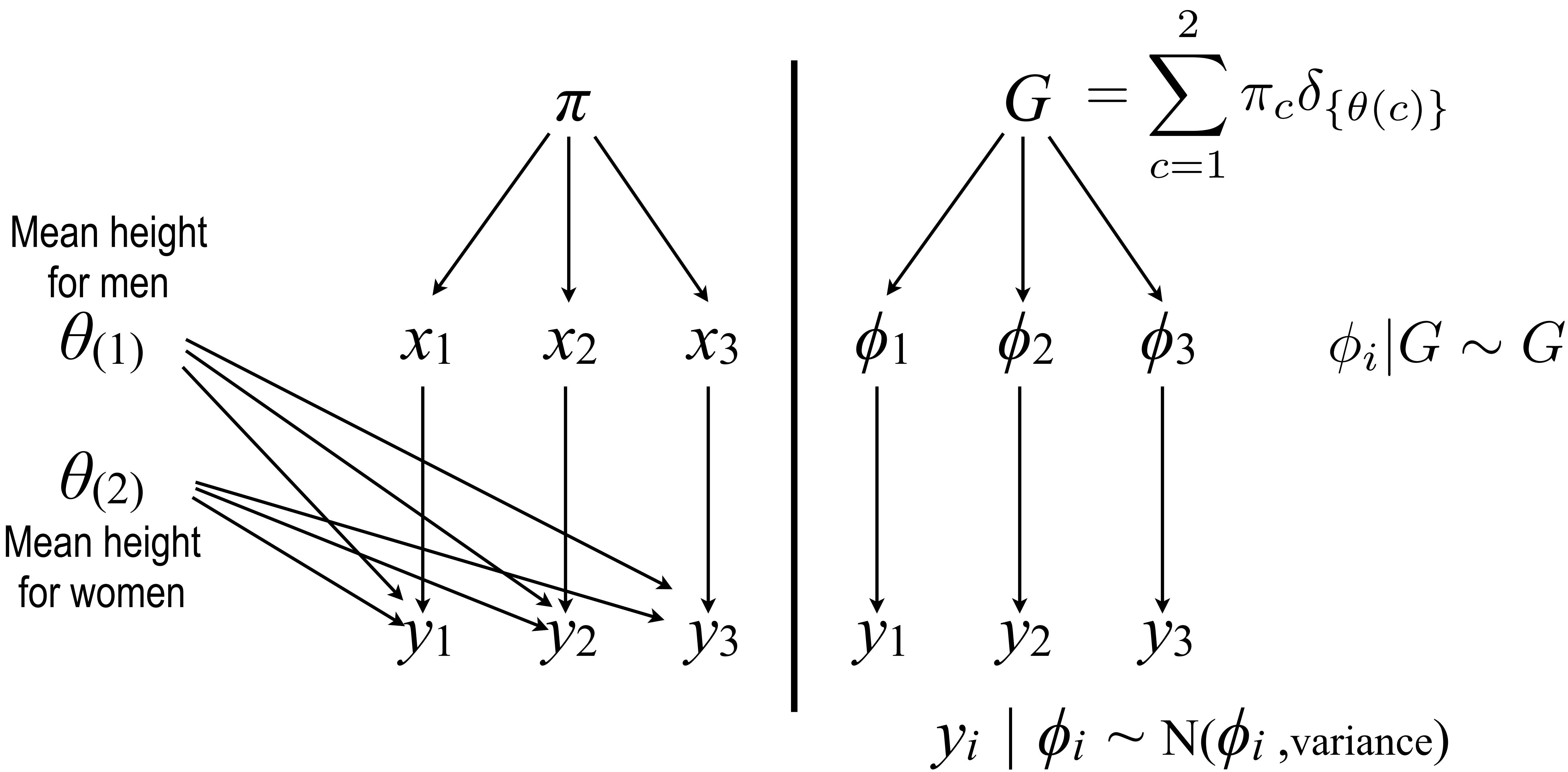
$$\theta_{(c)} \sim \mathrm{N}(\text{prior height, how confident prior})$$

4- Generate student heights for each i

$$y_i \mid x_i,\ \theta_{(1)},\ \theta_{(2)} \sim \mathrm{N}(\theta_{(x_i)}\ \text{,variance})$$

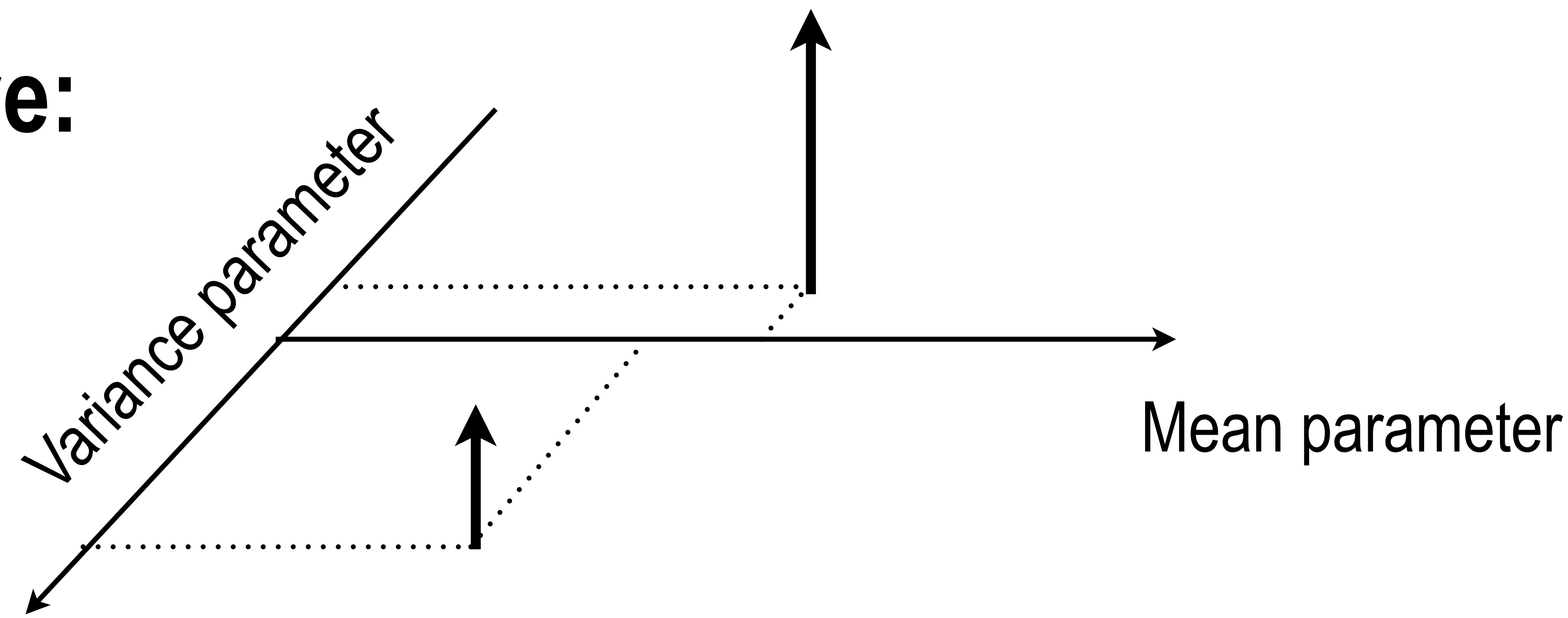
Equivalent notation

Mixture model: (UBC student height with 2 components)
say we have only 3 observations

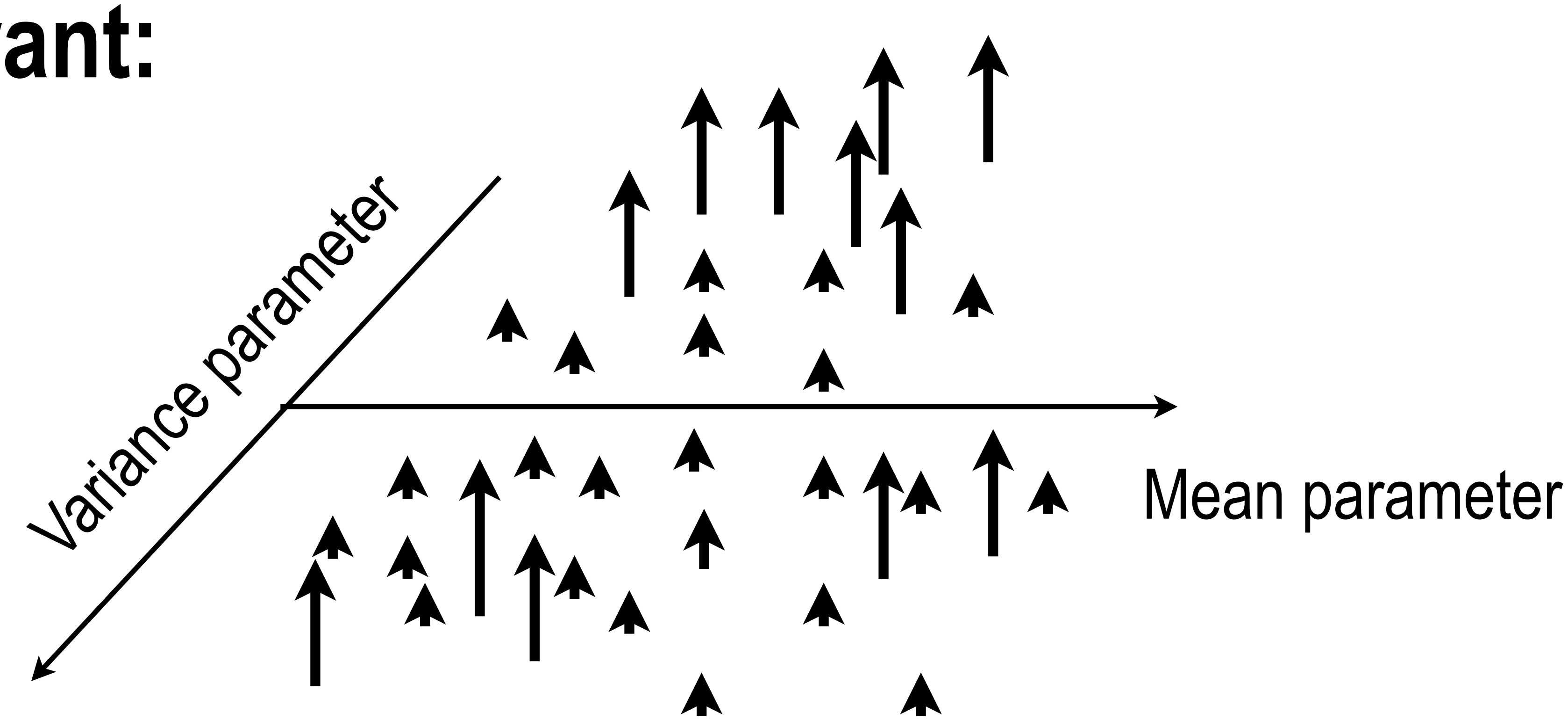


Samples from G

What we have:



What we want:



Definition: Dirichlet Process

Let G_0 be a distribution on a sample space Ω (the base distribution) α_0 be a positive real number (the concentration), and $(A_1, ..., A_\mathrm{k})$ be a partition of Ω. We say

$$G \sim \mathrm{DP}(\alpha_0, G_0)$$

i.e., G is a Dirichlet Process, if

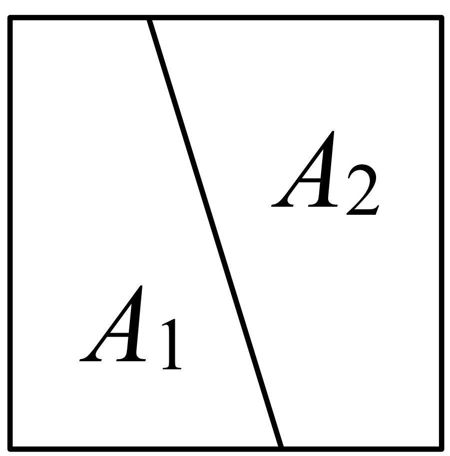
$$(G(A_1), \ldots, G(A_k)) \sim \mathrm{Dir}(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_k))$$

for all partitions and all k.

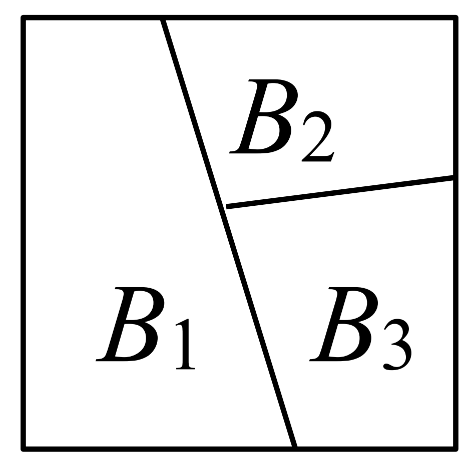
Does this make sense/exists?

Kolmogorov consistency: check the marginals are consistent under marginalization

In this case: check that the marginals are consistent when refining partitions



$$(G(A_1), G(A_2))$$
$$(U_1, U_2)$$



$$(G(B_1), G(B_2), G(B_3))$$
$$(V_1, V_2, V_3)$$

Constructive argument

Claim: the random probability distribution constructed below is the Dirichlet process with base distribution G_0 and concentration α_0

$$\beta_j \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha_0)$$

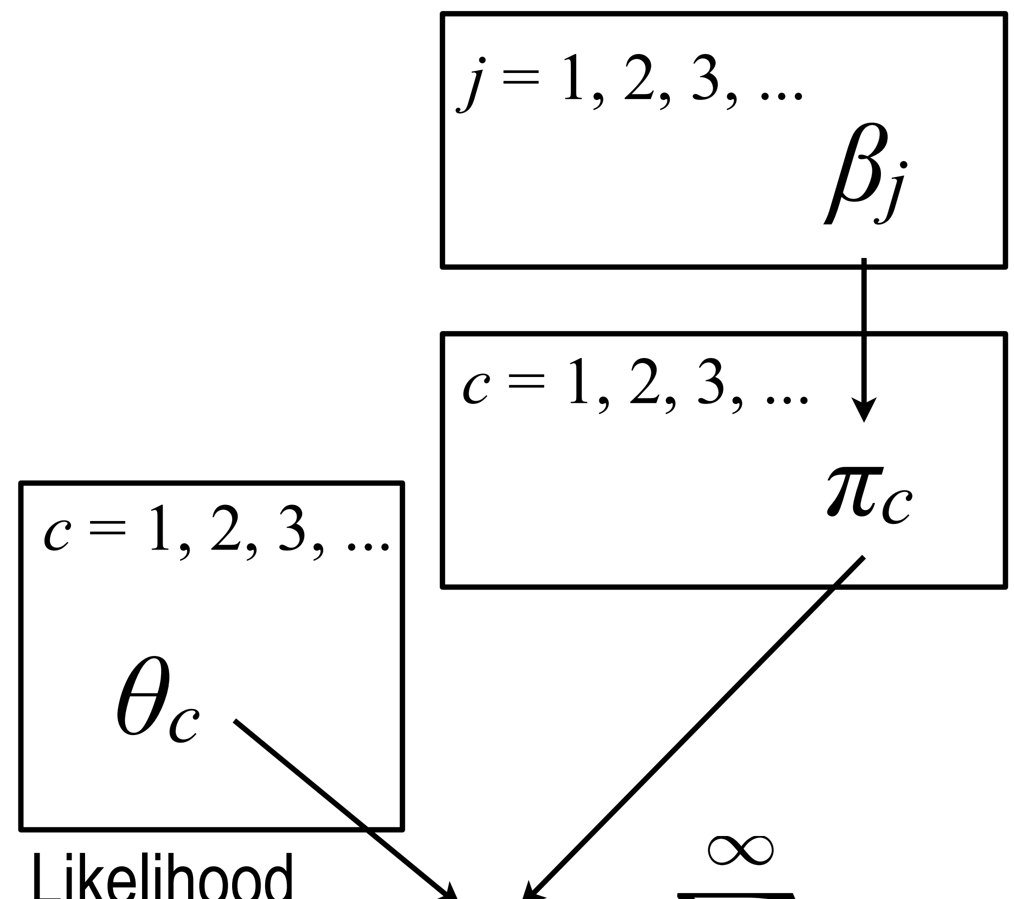
$$\theta_c \overset{\text{iid}}{\sim} G_0$$

Start with a stick of length 1, and break a segment of length β_1 for π_1, keep the rest

$$\pi_1 = \beta_1$$

At step c, if the length of the stick remaining is L, set:

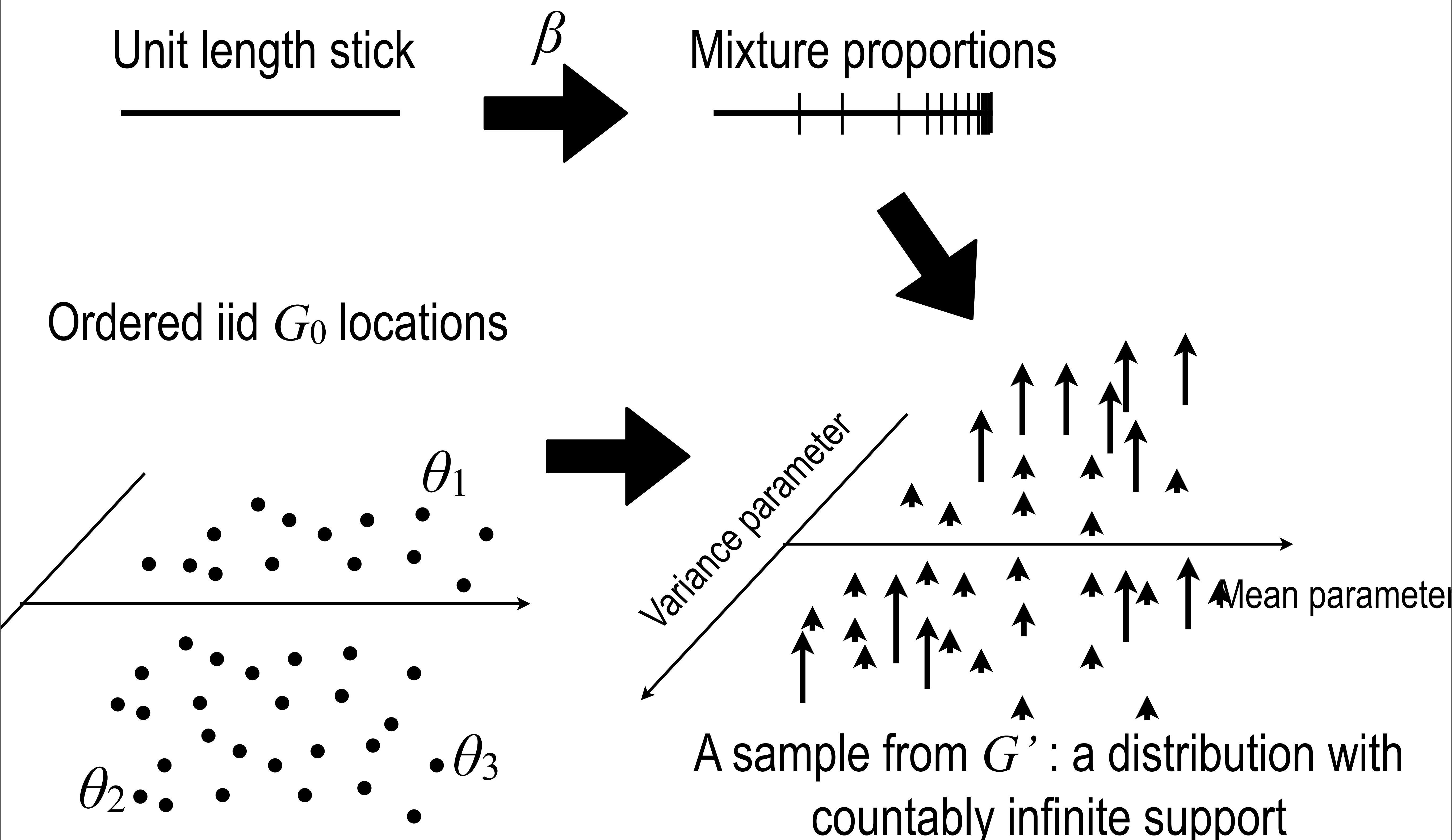
$$\pi_c = \beta_c L = \beta_c \prod_{j:j<c} (1 - \beta_j)$$



Likelihood mixture component parameters

$$G' = \sum_{c=1}^{\infty} \pi_c \delta_{\{\theta(c)\}}$$

Samples from G



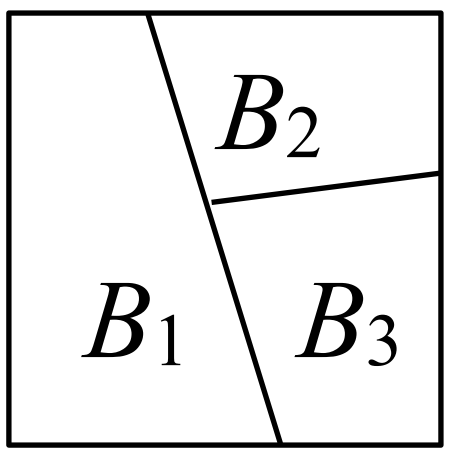
Are the samples indeed probability distributions?

Need to check: $$\sum_{c=1}^{n} \pi_c \xrightarrow{a.s.} 1$$

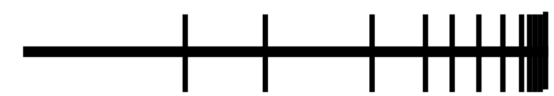
Recall: $$\pi_c = \beta_c L = \beta_c \prod_{j:j<c} (1 - \beta_j)$$

Goal: showing two definitions are equivalent

Kolmogorov consistency



Stick-breaking construction



Strategy: showing that for all partitions $(A_1, \ldots, A_k)$, the constructed process G' has finite Dirichlet marginals

$$(G'(A_1), \ldots, G'(A_k)) \sim \mathrm{Dir}(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_k))$$

Key observation: ‘self-similarity’

Definitions: $$G' = f(\beta,\theta) = \sum_{c=1}^{\infty} \pi_c \delta_{\{\theta(c)\}}$$

$$\beta^* = (\beta_1, \beta_2, \ldots)^* = (\beta_2, \beta_3, \ldots)$$

Observation: $$G' = \pi_1 \delta_{\{\theta(1)\}} + (1-\pi_1) f(\beta^*, \theta^*)$$

$$= \pi_1 \delta_{\{\theta(1)\}} + (1-\pi_1) G''$$ for $G' \stackrel{d}{=} G''$

Notation: $$G' \stackrel{st}{=} \pi_1 \delta_{\{\theta(1)\}} + (1-\pi_1) G'$$ *

How we’ll use it: we will show that if there is a distribution that satisfies this equation, it is unique; and that the finite Dirichlet distribution satisfies it

Detailed plan

Finite Dirichlet distributions satisfy equation (*)

Equation (*) has a unique solution

G' satisfies equation (*) $\Longrightarrow$ The marginals of G' satisfy equation (*) $\Longrightarrow$ G' has Dirichlet marginals

$\Downarrow$

$$G \overset{d}{=} G'$$

To be continued...