Graph Tracking in Dynamic Probabilistic Programs via Source Transformations

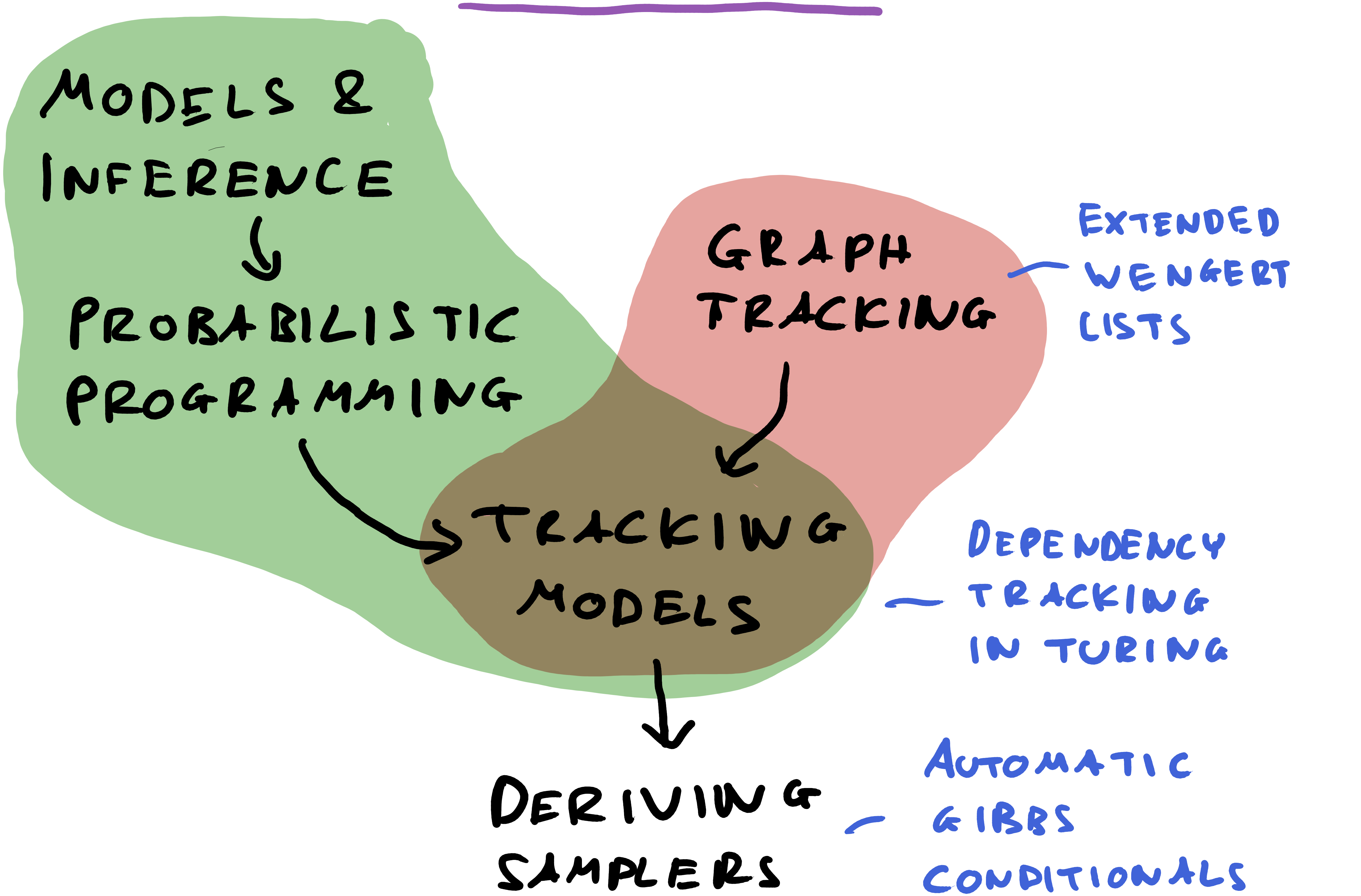
Philipp Gabler

2021-03-11

ABSTRACT

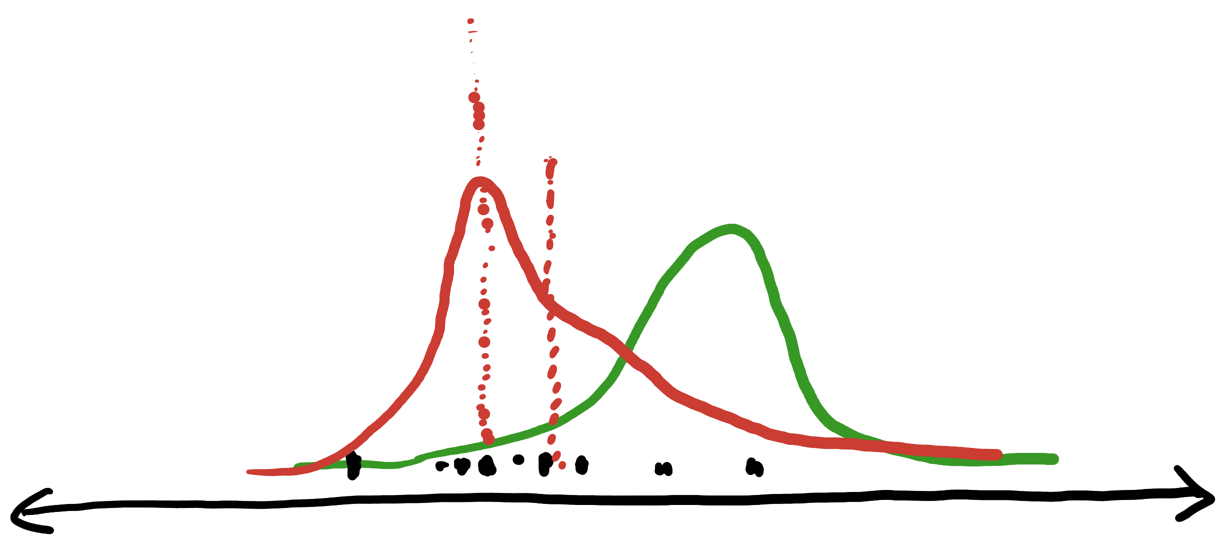
- Q: HOW TO ADD STRUCTURE TO TURING MODELS?
- OUTCOMES:
  - "EXTENDED WENGERT LIST" TRACKING TECHNIQUE
  - DEPENDENCY GRAPH EXTRACTION (SLICING) FROM TURING MODELS
  - AUTOMATIC GIBBS SAMPLER

ROADMAP


MODELS & INFERENCE
PROBABILISTIC PROGRAMMING
GRAPH TRACKING
EXTENDED WENGERT LISTS
TRACKING MODELS
DEPENDENCY TRACKING IN TURING
DERIVING SAMPLERS
AUTOMATIC GIBBS CONDITIONALS

7

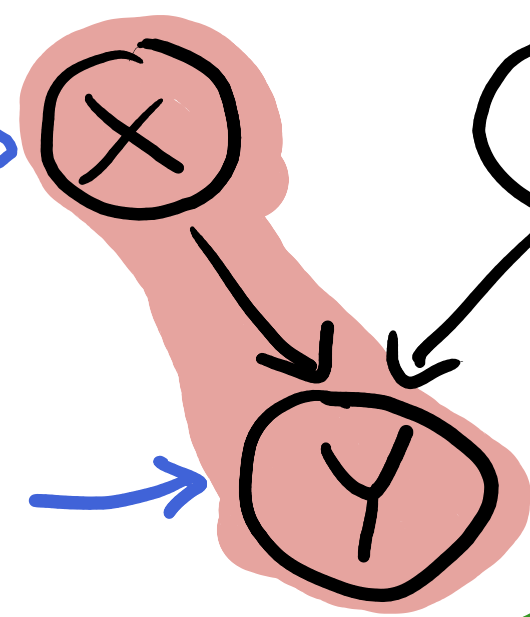
BAYES-ICS



GENERATIVE MODELS

INPUT

PARAMETERS, LATENT VARS, ...



OUTPUT

PRIOR

LIKELIHOOD

GIVEN: GENERATIVE PROCESS

$$P(Y,\Theta|X) = \underbrace{P(\Theta)}_{\text{PRIOR}}\,\underbrace{P(Y|X,\Theta)}_{\text{LIKELIHOOD}}$$

WANTED: POSTERIOR

$$P(\Theta|D) = \frac{P(Y,\Theta|X)}{P(Y|X)}$$

OBSERVED DATA

$D = \{(x_1, y_1), \ldots\}$

EVIDENCE

POSTERIOR EXPECTATION

DATA

- $\mathbb{E}[f(\theta) \mid D]$

POSTERIOR

$$= \int f(\vartheta)\, p(\vartheta \mid D)\, d\nu(\vartheta)$$

- INTERESTING QUANTITIES (OFTEN) ARE INTRACTABLE INTEGRALS!

E.G. PPD, CREDIBLE INTERVALS, MAP, MODEL COMPARISON

APPROXIMATE INFERENCE

- VI: MINIMIZE

$$D\left(q_\lambda(D)\,\|\,p(\vartheta \mid D)\right)$$

- SAMPLING:

USE LLN!

$$I^{(N)}(f) = \frac{1}{N}\sum_n f(Y^{(n)})$$

$$\xrightarrow{a.s.} \mathbb{E}[f(\Theta) \mid D]$$

THE MARKOV CHAIN

- CONSTRUCT TRANSITION KERNEL:

$$\mathbb{P}[Y^{(k+1)} \in dy \mid Y^{(k)} = y^{(k)}, \dots, Y^{(1)} = y^{(1)}]$$
$$= K(y^{(k)}, dy)$$
$$\text{s.t. } \underbrace{p(\vartheta | D)\, K(\vartheta, S) = \mathbb{P}[\theta \in S \mid D] \;\; \forall S}_{\widehat{=}\; \pi K = \pi}$$

- SCHEMA:
  - PROPOSE NEW $Y^{(i)}$ FROM $Y^{(i-1)}$
  - ACCEPT/REJECT

LIVE DEMO

KERNELS ≈ SAMPLERS

- METROPOLIS-HASTINGS, HMC, ...
- OFTEN EASY: GIBBS CONDITIONALS!

$$p(\vartheta_1 \mid \vartheta_2, D) \qquad p(\vartheta_2 \mid \vartheta_1, D)$$

NO ACCEPTANCE STEP!

- "WITHIN-GIBBS": MIX SAMPLERS PER VARIABLE

2

PROBABILISTIC PROGRAMMING

$$X \sim \mathcal{N}(\mu, \sigma_1)$$

$$Y \sim \mathcal{N}(X, \sigma_2)$$

PROBABILISTIC PROGRAMS I

HOW TO IMPLEMENT MODELS?

OBSERVED DATA

PARAMETERS

```
@model function hierarchical_gaussian(x)
    λ ~ Gamma(2.0, inv(3.0))
    m ~ Normal(0, sqrt(1 / λ))
    x ~ Normal(m, sqrt(1 / λ))
end
```

JULIA FUNCTIONS

Probabilistic Programs II

PARAMETERS

```
@model function normal_mixture(x, K, m, s, σ)
    N = length(x)

    μ = Vector{Float64}(undef, K)
    for k = 1:K
        μ[k] ~ Normal(m, s)
    end

    z = Vector{Int}(undef, N)
    for n = 1:N
        z[n] ~ Categorical(K)
    end

    for n = 1:N
        x[n] ~ Normal(μ[z[n]], σ)
    end

    return x
end
```

DATA STRUCTURES

MUTATION

LOOPS / CONTROL FLOW

PPL CHARACTERISTICS

- DENSITY EVALUATION
- MODEL STRUCTURE
- AUTOMATIC DIFFERENTIATION
- SAMPLING, DATA GENERATION
- DIAGNOSTICS, EVALUATION
- PROGRAMMING & EXTERNAL LIBS
- COMPOSITION

PPL IMPLEMENTATIONS

METAPROGRAMMING

SOURCE $\longrightarrow$ DENSITY PROGRAM $\longrightarrow$ SAMPLER

RESTRICTED DYNAMICS

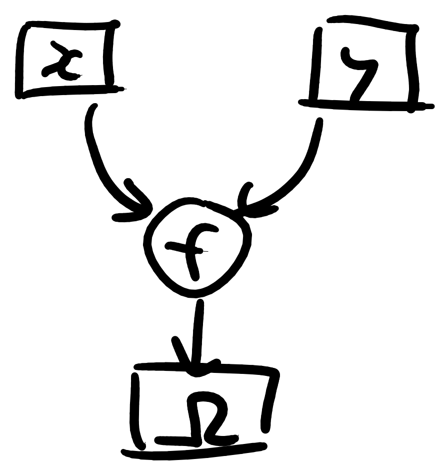
NO STRUCTURE

TRACED PROGRAM $\rightleftarrows$ SAMPLER

CALLBACKS

3

GRAPH TRACKING



Computation Graphs

$$g(\sin(x), y)$$

EXPRESSION / PROGRAM

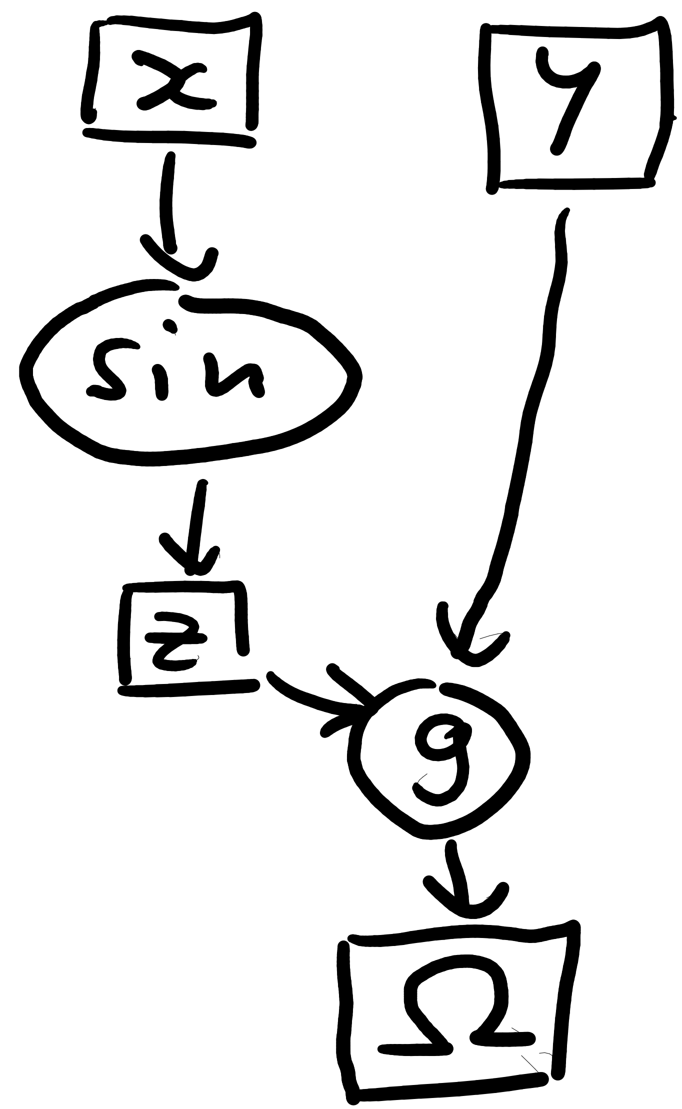
$$x = ?$$

$$y = ?$$

$$z = \sin(x)$$

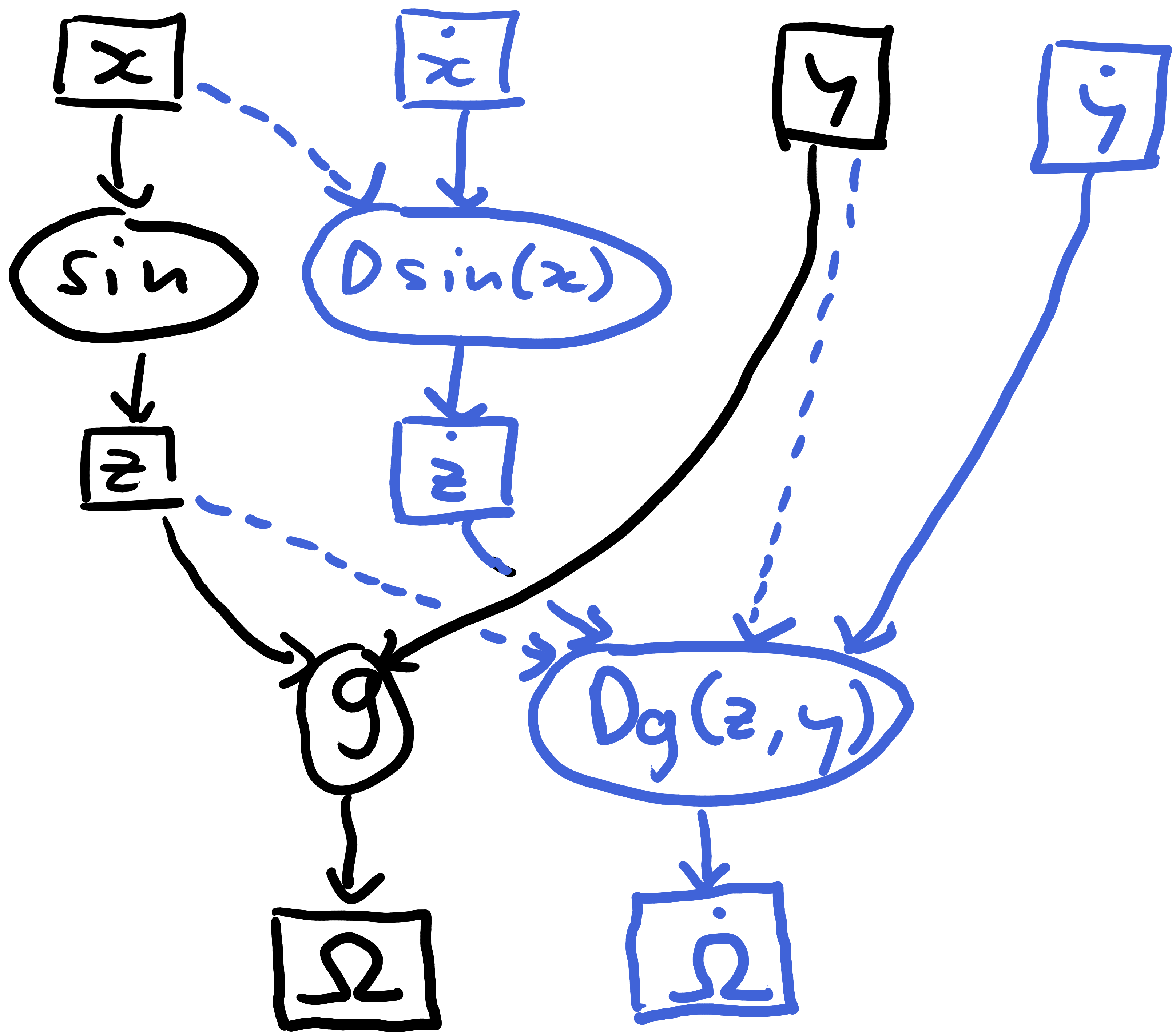
$$\Omega = g(z, y)$$

SSA FORM



GRAPH

GRAPHS IN AD



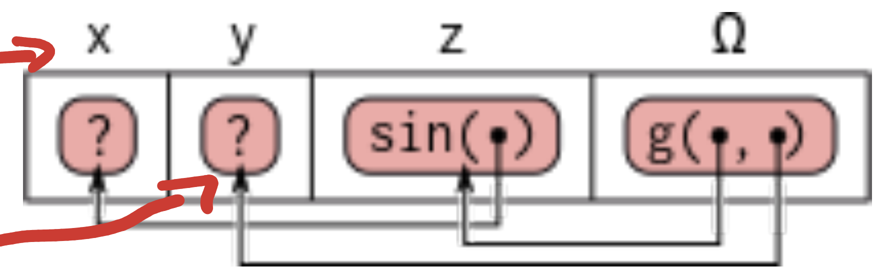
IMPLEMENTATON TECHNIQUES

- OPERATOR OVERLOADING:

x = TRACKED(?)

y = TRACKED(?)

$\vdots$



WENGERT LIST, A.K.A. TAPE

- SOURCE TRANSFORMATION:

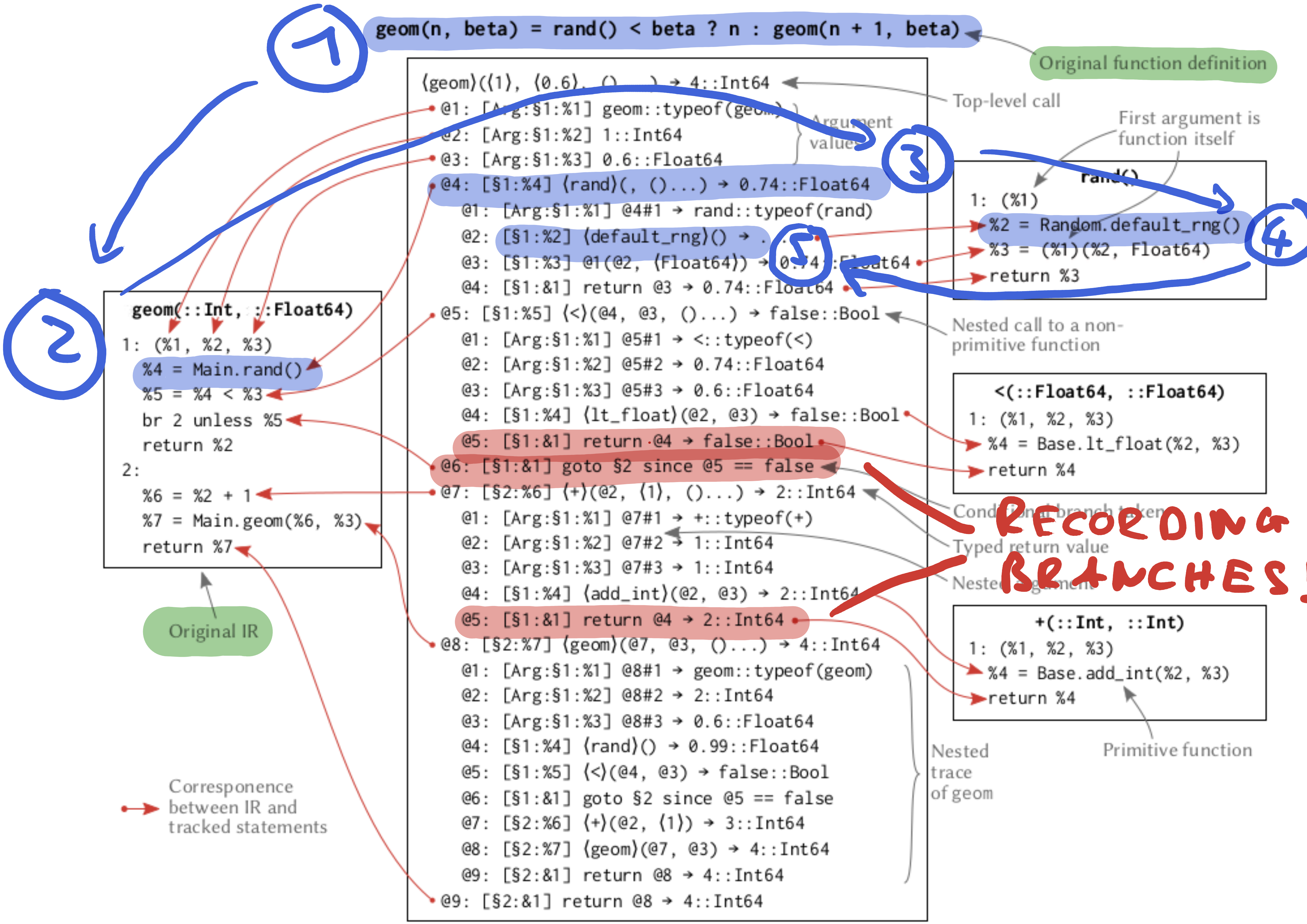
$$x = ? \qquad z = \sin(x)$$
$$\dot{x} = \Delta_1 \qquad \dot{z} = D\sin(x)(\dot{x})$$
$$y = ? \qquad \Omega = g(z, y)$$
$$\dot{y} = \Delta_2 \qquad \dot{\Omega} = Dg(z,y)(\dot{z}, \dot{y})$$

JULIA IR

f(x,y) = y > 0 ? g(sin(x),y) : y

```
    f    x    y
1: (%1, %2, %3)
  %4 = %3 > 0
  br 3 unless(%4)
2:
  %5 = sin(%2)        z = sin(x)
  %6 = g(%5, %3)      Ω = g(z,y)
  return %6
3:
  return %3
```

EXTENDED WENGERT LIST



IR TRANSFORMATION

TRACKED IR

```
%15 = record!(%5, %14)
%16 = TapeConstant(Main.rand)
%17 = Base.tuple()
%18 = trackedcall(%5, %16, %17, $(QuoteNode(§1:%4)))
%19 = record!(%5, %18)
%20 = TapeConstant(Main.:<)
%21 = trackedvariable(%5, $(QuoteNode(%4)), %19)
%22 = trackedvariable(%5, $(QuoteNode(%3)), %3)
%23 = Base.tuple(%21, %22)
%24 = trackedcall(%5, %20, %23, $(QuoteNode(§1:%5)))
%25 = record!(%5, %24)
%26 = Base.tuple()
%27 = trackedvariable(%5, $(QuoteNode(%5)), %25)
%28 = trackedjump(%5, 2, %26, %27, $(QuoteNode(§1:&1)))
%29 = trackedvariable(%5, $(QuoteNode(%2)), %2)
%30 = trackedreturn(%5, %29, $(QuoteNode(§1:&2)))
br 2 (%28) unless %25
br 3 (%2, %30)
```

Actual jump is recorded

```
1: (%1, %2, %3)
  %4 = Main.rand()
  %5 = %4 < %3
  br 2 unless %5
  return %2
```

ORIGINAL IR

Jumps and returns are passed down to the next block

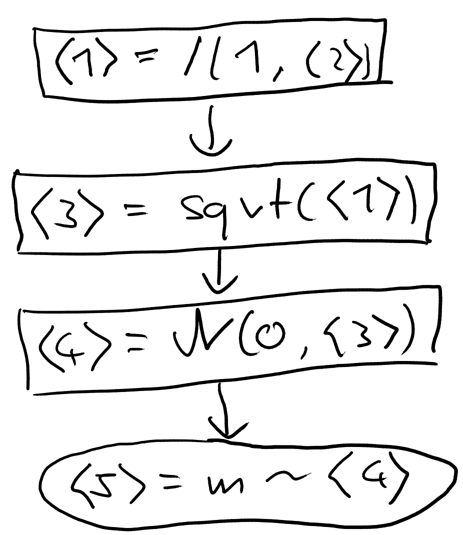
⋮

```
3: (%46, %47)
  %48 = record!(%5, %47)
  return %46
```

Special extra block for return values

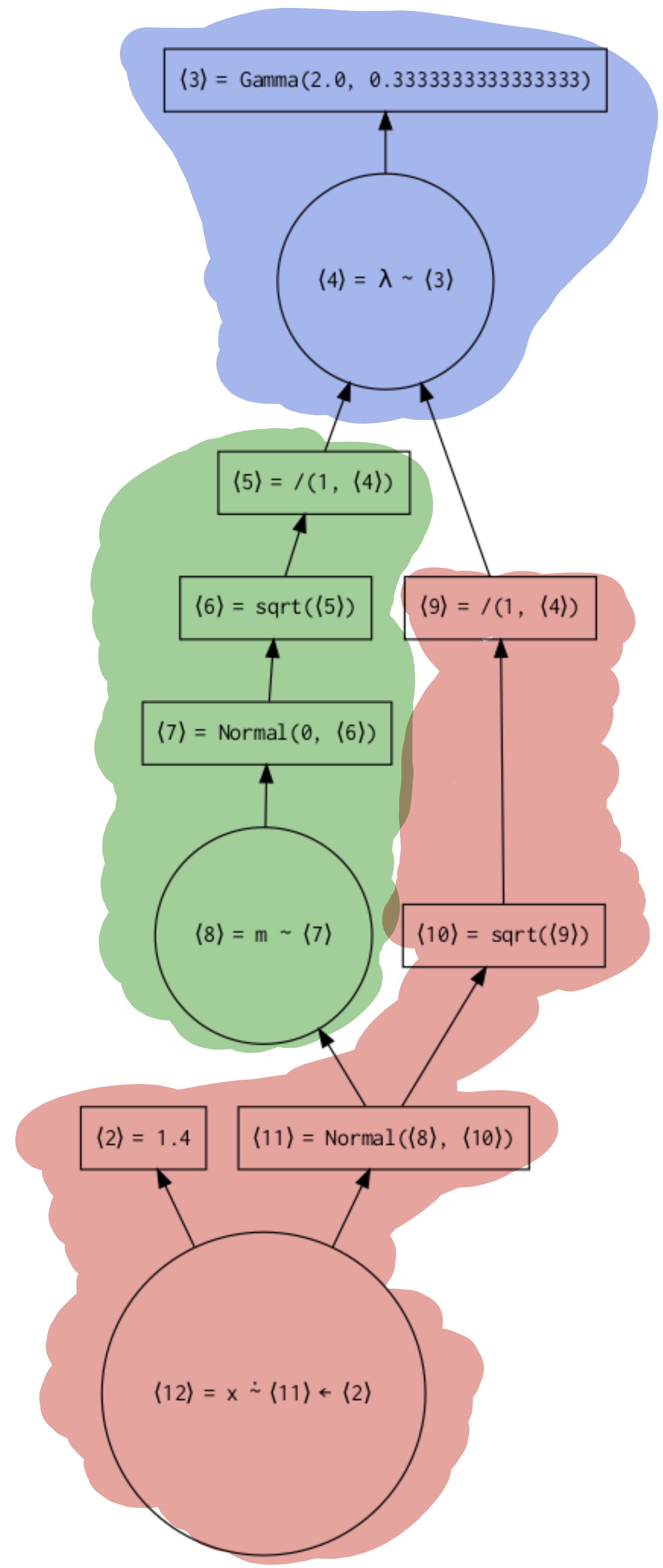
4

Tracking Turing Models

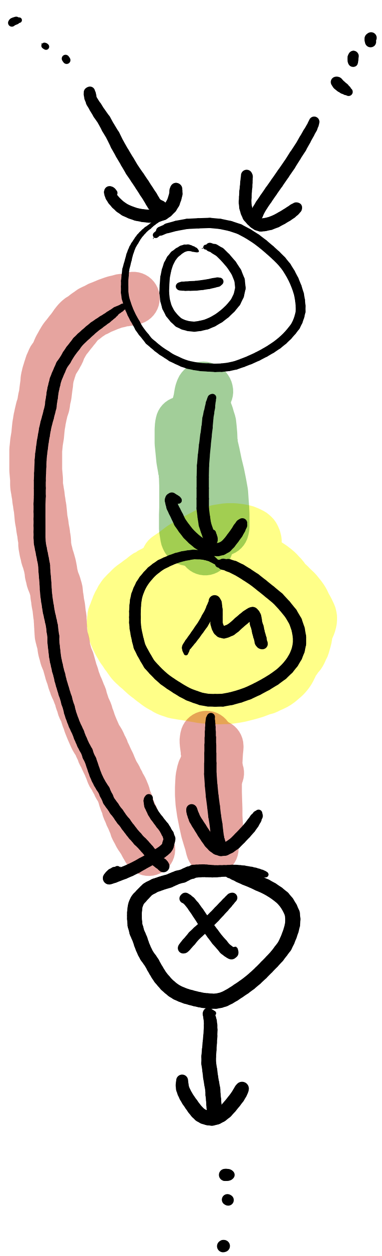


Dependency Extraction

```
@model function hierarchical_gaussian(x)
    λ ~ Gamma(2.0, inv(3.0))
    m ~ Normal(0, sqrt(1 / λ))
    x ~ Normal(m, sqrt(1 / λ))
end
```



(DISCRETE) GIBBS CONDITIONALS



$$p(m, \vartheta, x)$$
$$= p(\vartheta)\, p(m \mid \vartheta)\, p(x \mid m, \vartheta)$$

MARKOV BLANKET

$$p(m \mid \vartheta, x) = \frac{p(m \mid \vartheta)\, p(x \mid m, \vartheta)}{Z}$$

$$Z = \sum_m p(m \mid \vartheta)\, p(x \mid m, \vartheta)$$

A Test Model

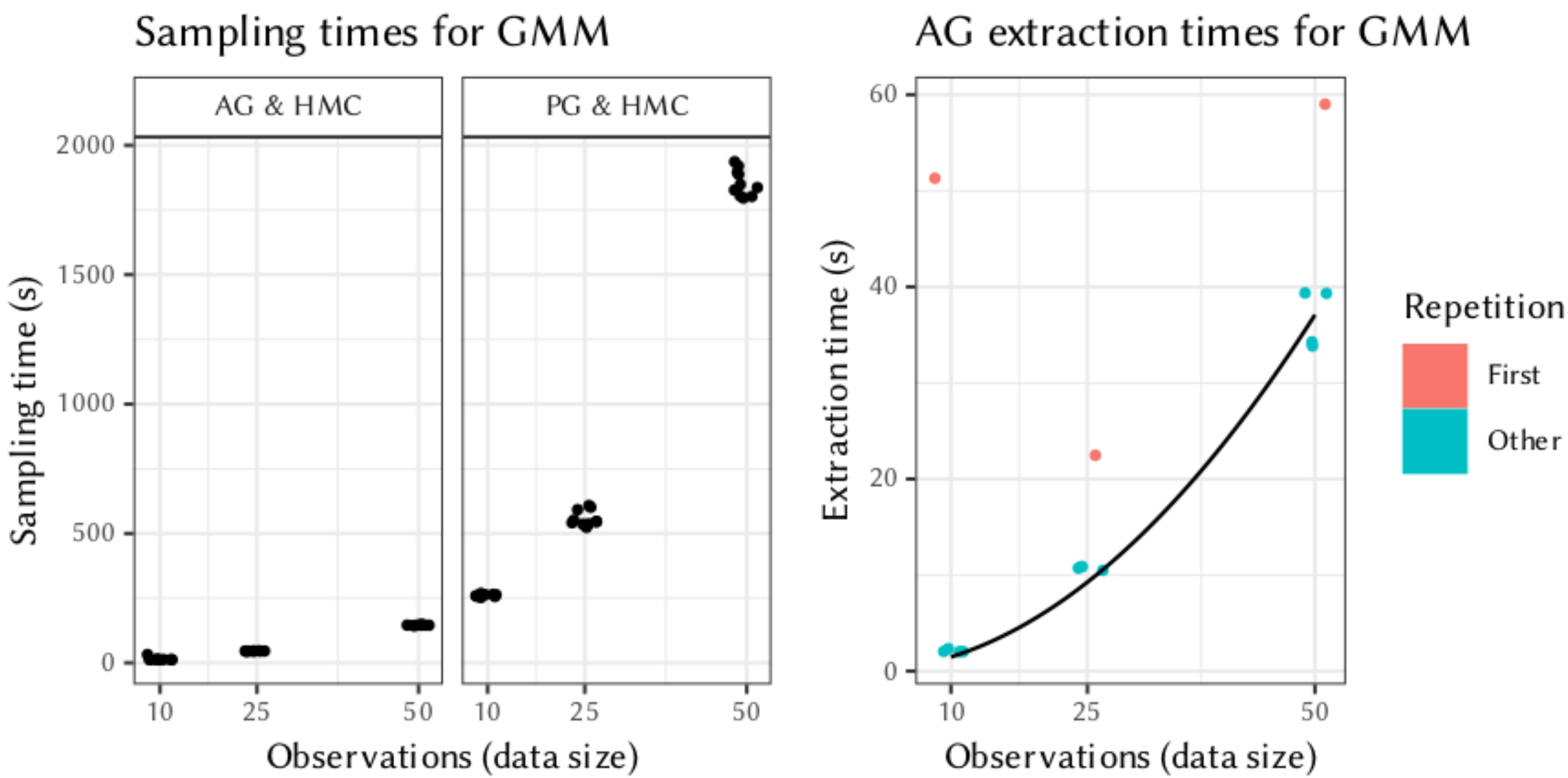
```
@model function gmm(x, K)
    N = length(x)
    w ~ Dirichlet(K, 1/K)  # Cluster association prior
    z ~ filldist(Categorical(w), N)  # Cluster assignments
    μ ~ filldist(Normal(0.0, s1_gmm), K)  # Cluster centers

    for n = 1:N
        x[n] ~ Normal(μ[z[n]], s2_gmm)  # Observations
    end
end
```

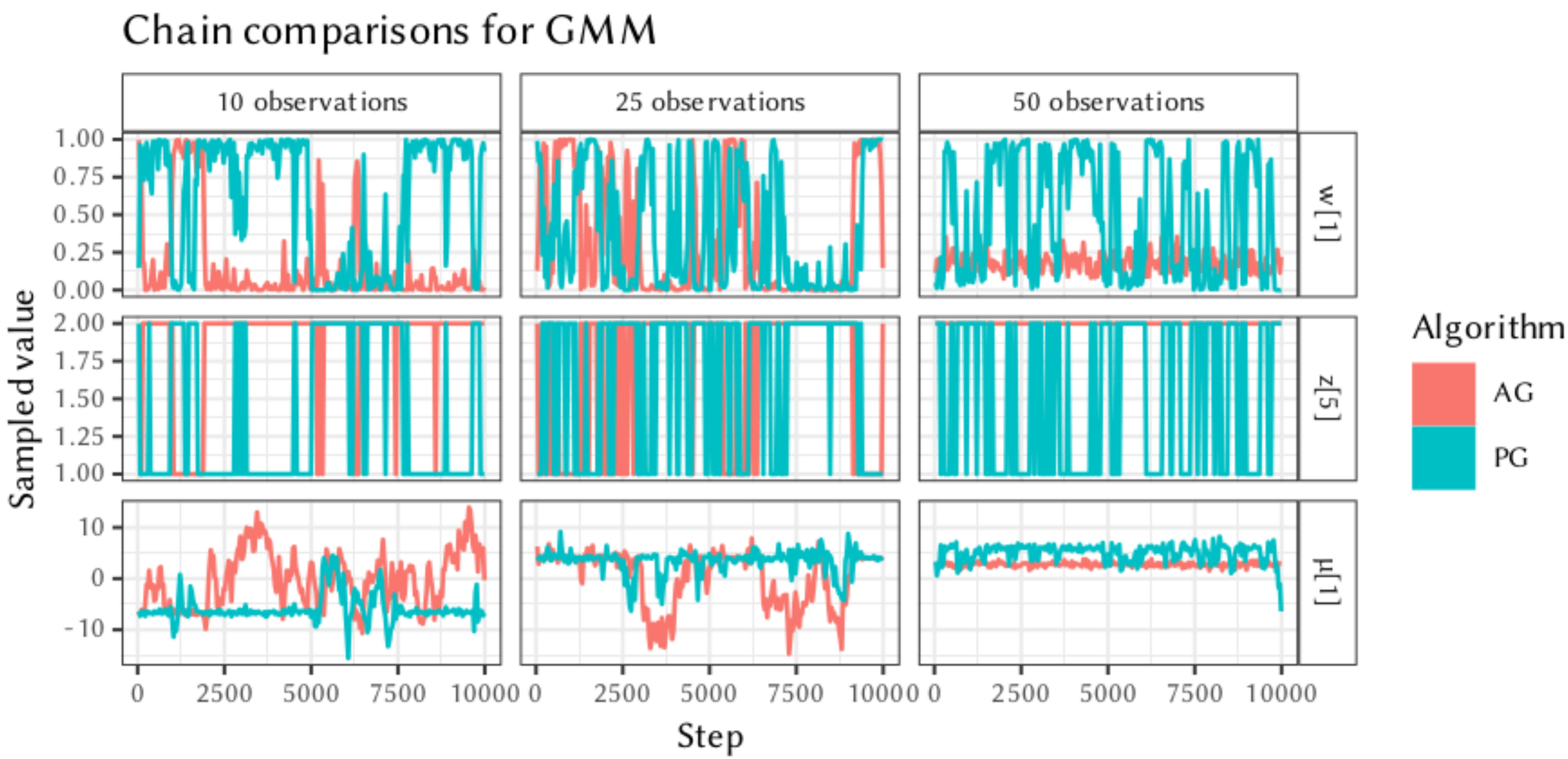
z: AutoGibbs vs. Particle Gibbs

w, μ: HMC

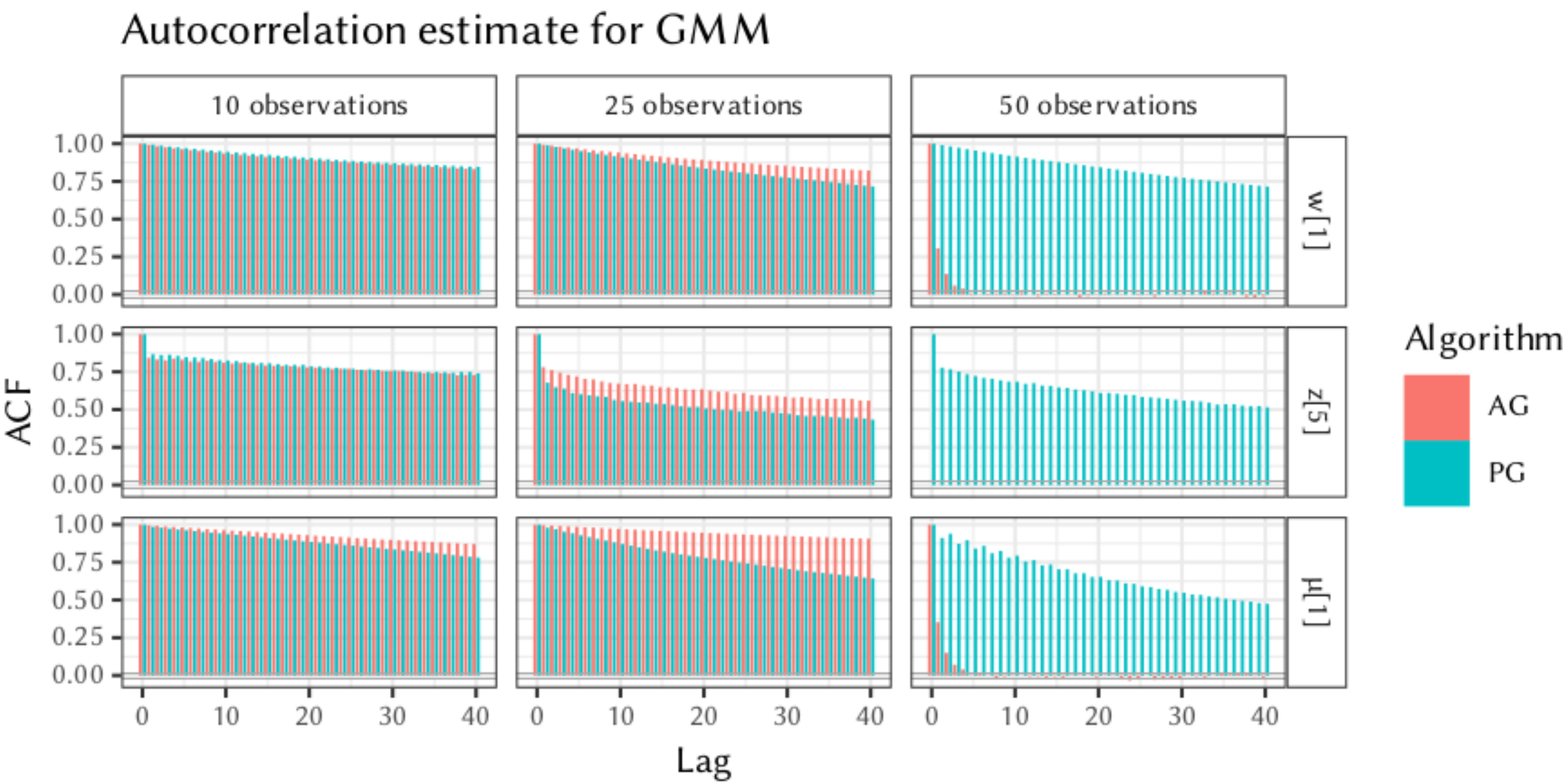
EVALUATION: TIME



Evaluation: Chains



EVALUATION: CONVERGENCE

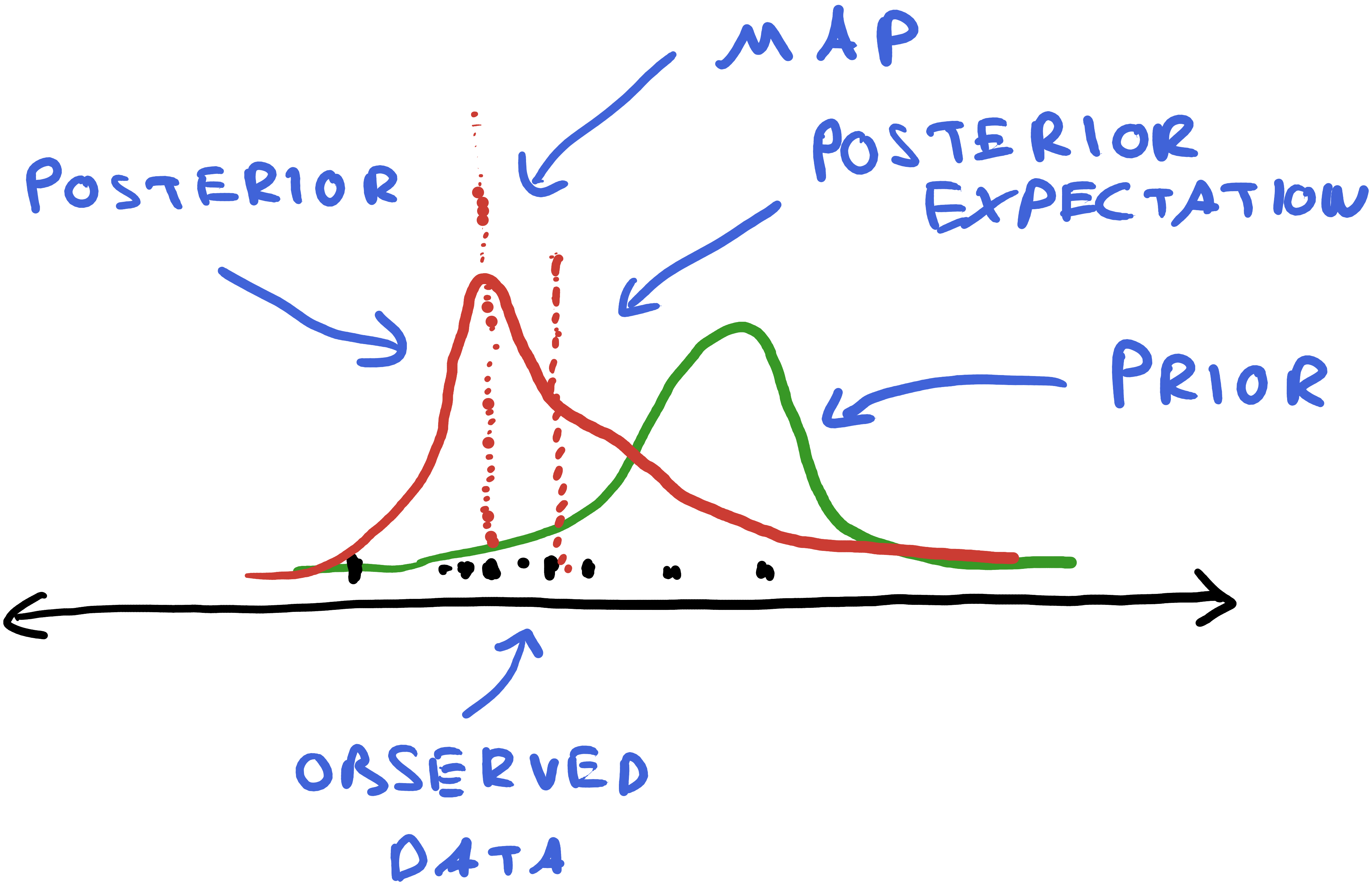


CONCLUSIONS

- STATIC DEPENDENCIES, FINITE CONDITIONALS: 
- SLICING DYNAMIC MODELS: 
- RECOVERING DYNAMIC STRUCTURE: 
- FUTURE PROOF:  ?

scratch

"Bayes-ics"



PREDICTION

- PLUGIN ESTIMATOR:

arg max!

$$\hat{p}(y|x,D) = p(y|x,\vartheta^*(D))$$

- POSTERIOR PREDICTIVE:

$$p(y|x,D) = \int p(y|x,\vartheta)\,p(\vartheta|D)\,d\nu(\vartheta)$$

POSTERIOR EXPECTATION!