Bayesian Optimisation for Likelihood Free Inference

Make model parameterisation go brrr

Jacob Cumming

University of Melbourne, Walter and Eliza Hall Institute

May 2024

Notation

- Model a (random) function $f : \boldsymbol{\Theta} \to \boldsymbol{\mathcal{X}}$.
  - $\boldsymbol{\Theta}$: parameter space
  - $\boldsymbol{\mathcal{X}}$: model output space
  - $\mathbf{X}_{\boldsymbol{\theta}} := f(\boldsymbol{\theta})$ (assumed same form as $\mathbf{X}_{\text{obs}}$).

Notation

- Model a (random) function $f : \boldsymbol{\Theta} \to \boldsymbol{\mathcal{X}}$.
  - $\boldsymbol{\Theta}$: parameter space
  - $\boldsymbol{\mathcal{X}}$: model output space
  - $\mathbf{X}_{\boldsymbol{\theta}} := f(\boldsymbol{\theta})$ (assumed same form as $\mathbf{X}_{\text{obs}}$).
- $\mathbf{X}_{\text{obs}}$: a vector of observed data (incidence, prevalence, hospitalisations etc.)

Notation

- Model a (random) function $f : \Theta \to \mathcal{X}$.
  - Θ : parameter space
  - $\mathcal{X}$: model output space
  - $\mathbf{X}_\theta := f(\theta)$ (assumed same form as $\mathbf{X}_{\text{obs}}$).
- $\mathbf{X}_{\text{obs}}$: a vector of observed data (incidence, prevalence, hospitalisations etc.)
- $S(\mathbf{X}_{\text{obs}})$: summary statistic (vector) of observed data (average weekly incidence etc.)

Champagne Model Parameters

- α : proportion of those infected but cleared of blood stage infections (through treatment)
- β : a further proportion that are also cleared of liver stage parasites, given that they were also cleared of blood stage infection (radical cure)
- λ : the rate of infection
- γ_L : rate of clearance of liver stage disease
- f : rate of relapse
- r : rate of blood stage clearance
- δ : importation rate (which we assume is 0)

Parameter inference would become easy if we had...

- An explicit form for the likelihood:
  - $\mathcal{L}(\theta|\mathbf{X}_{\text{obs}}) := \Pr(\mathbf{X}_{\text{obs}}|\theta)$

Parameter inference would become easy if we had...

- An explicit form for the likelihood:
  - $\mathcal{L}(\theta|\mathbf{X}_{\text{obs}}) := \Pr(\mathbf{X}_{\text{obs}}|\theta)$
  - Or $\mathcal{L}(\boldsymbol{\theta}|S(\mathbf{X}_{\text{obs}})) := \Pr(S(\mathbf{X}_{\text{obs}})|\boldsymbol{\theta})$

Parameter inference would become easy if we had...

- An explicit form for the likelihood:
  - $\mathcal{L}(\theta|\mathbf{X}_{\text{obs}}) := \Pr(\mathbf{X}_{\text{obs}}|\theta)$
  - Or $\mathcal{L}(\boldsymbol{\theta}|S(\mathbf{X}_{\text{obs}})) := \Pr(S(\mathbf{X}_{\text{obs}})|\boldsymbol{\theta})$
- $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|S(\mathbf{X}_{\text{obs}}))$

Parameter inference would become easy if we had...

- An explicit form for the likelihood:
  - $\mathcal{L}(\theta|\mathbf{X}_{\text{obs}}) := \Pr(\mathbf{X}_{\text{obs}}|\theta)$
  - Or $\mathcal{L}(\boldsymbol{\theta}|S(\mathbf{X}_{\text{obs}})) := \Pr(S(\mathbf{X}_{\text{obs}})|\boldsymbol{\theta})$
- $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|S(\mathbf{X}_{\text{obs}}))$
- $\Pr(\boldsymbol{\theta}|S(\mathbf{X}_{\text{obs}})) \propto \Pr(S(\mathbf{X}_{\text{obs}})|\boldsymbol{\theta}) \Pr(\boldsymbol{\theta})$

Reality Sets In

- Explicit likelihoods often don't exist/are intractible
  - eg. agent based models

A Standard Bayesian Solution

- Approximate Bayesian Computation (ABC)
  1. Sample $\boldsymbol{\theta}$ from prior
  2. Run model
  3. Accept or reject parameters run based on how well $\mathbf{X}_{\boldsymbol{\theta}}$ 'matches' $\mathbf{X}_{\text{obs}}$.

What is 'matches'

- Match if $\mathbf{X}_{\boldsymbol{\theta}} = \mathbf{X}_{\text{obs}}$? (no)

What is 'matches'

- Match if $\mathbf{X}_{\boldsymbol{\theta}} = \mathbf{X}_{\text{obs}}$? (no)
- Discrepency function $D : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$
  - Can be a norm such as
    $||S(\mathbf{X}_{\boldsymbol{\theta}}) - S(\mathbf{X}_{\text{obs}})||_p := (\sum_{i=1}^{d} |S(\mathbf{X}_{\boldsymbol{\theta}}) - S(\mathbf{X}_{\text{obs}})|^p)^{1/p}$

What is 'matches'

- Match if $\mathbf{X}_{\boldsymbol{\theta}} = \mathbf{X}_{\text{obs}}$? (no)
- Discrepency function $D : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$
  - Can be a norm such as
    $||S(\mathbf{X}_{\boldsymbol{\theta}}) - S(\mathbf{X}_{\text{obs}})||_p := (\sum_{i=1}^{d} |S(\mathbf{X}_{\boldsymbol{\theta}}) - S(\mathbf{X}_{\text{obs}})|^p)^{1/p}$
  - Rescale $S(\cdot)$ appropriately (ie via a covariance matrix).

What is 'matches'

- Match if $\mathbf{X}_{\boldsymbol{\theta}} = \mathbf{X}_{\text{obs}}$? (no)
- Discrepency function $D : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$
  - Can be a norm such as
    $||S(\mathbf{X}_{\boldsymbol{\theta}}) - S(\mathbf{X}_{\text{obs}})||_p := (\sum_{i=1}^{d} |S(\mathbf{X}_{\boldsymbol{\theta}}) - S(\mathbf{X}_{\text{obs}})|^p)^{1/p}$
  - Rescale $S(\cdot)$ appropriately (ie via a covariance matrix).
- We consider $\mathcal{D}(\boldsymbol{\theta}) := D(S(\mathbf{X}_{\boldsymbol{\theta}}), S(\mathbf{X}_{\text{obs}}))$

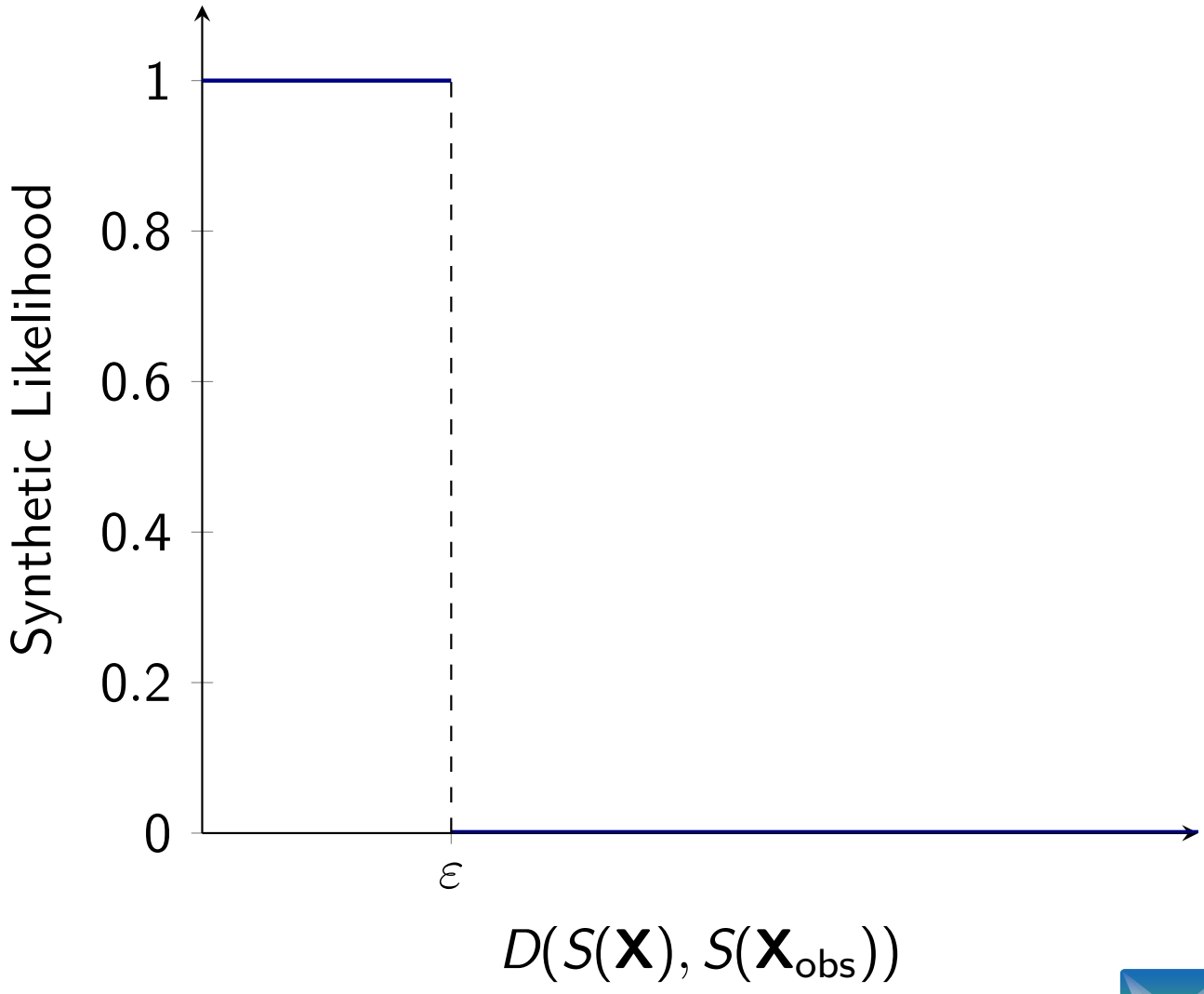
THE UNIVERSITY OF MELBOURNE

WEHI brighter together

What is 'matches'

- Match if $\mathbf{X}_{\boldsymbol{\theta}} = \mathbf{X}_{\text{obs}}$? (no)
- Discrepency function $D : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$
  - Can be a norm such as $||S(\mathbf{X}_{\boldsymbol{\theta}}) - S(\mathbf{X}_{\text{obs}})||_p := (\sum_{i=1}^{d} |S(\mathbf{X}_{\boldsymbol{\theta}}) - S(\mathbf{X}_{\text{obs}})|^p)^{1/p}$
  - Rescale $S(\cdot)$ appropriately (ie via a covariance matrix).
- We consider $\mathcal{D}(\boldsymbol{\theta}) := D(S(\mathbf{X}_{\boldsymbol{\theta}}), S(\mathbf{X}_{\text{obs}}))$
- $\mathcal{D}(\boldsymbol{\theta})$ is 'how close' we were using parameters $\boldsymbol{\theta}$.

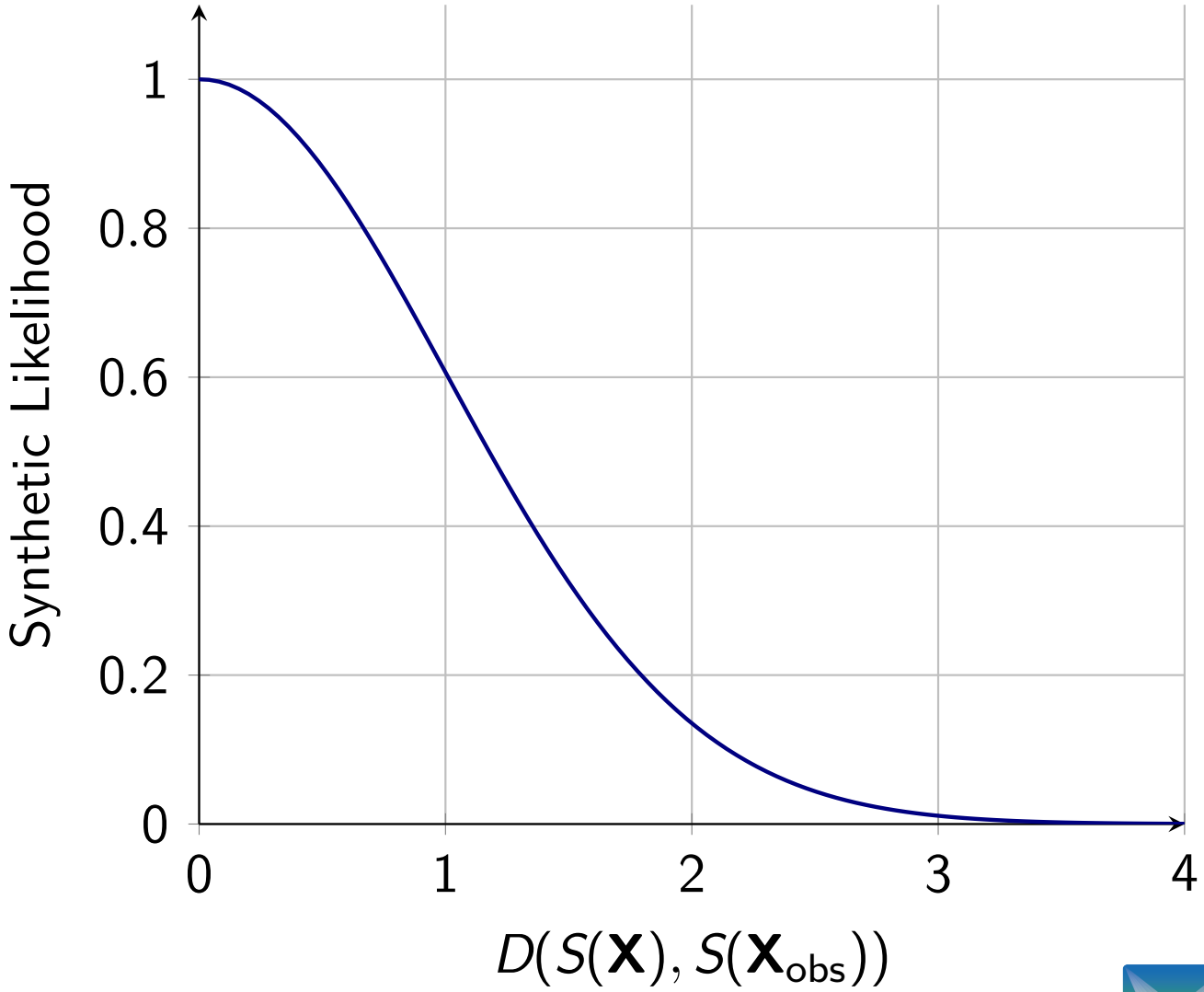
Uniform Acceptance Probability



THE UNIVERSITY OF MELBOURNE

WEHI
brighter together

Acceptance Probability



THE UNIVERSITY OF MELBOURNE

WEHI
brighter together

Overall Idea of my Research

- Can we predict $\mathcal{D}(\boldsymbol{\theta})$ for simulated $\boldsymbol{\theta}$?

Overall Idea of my Research

- Can we predict $\mathcal{D}(\boldsymbol{\theta})$ for simulated $\boldsymbol{\theta}$?
- Locally, hopefully yes.

THE UNIVERSITY OF MELBOURNE

WEHI brighter together

Gaussian Processes

- Random functions
- Common examples - Brownian motion, Ornstein Uhlenbeck process

Gaussian Processes

- Random functions
- Common examples - Brownian motion, Ornstein Uhlenbeck process
- Nap time for wet-lab people

Gaussian Processes on $\mathbb{R}^d$

Definition (Gaussian Process)

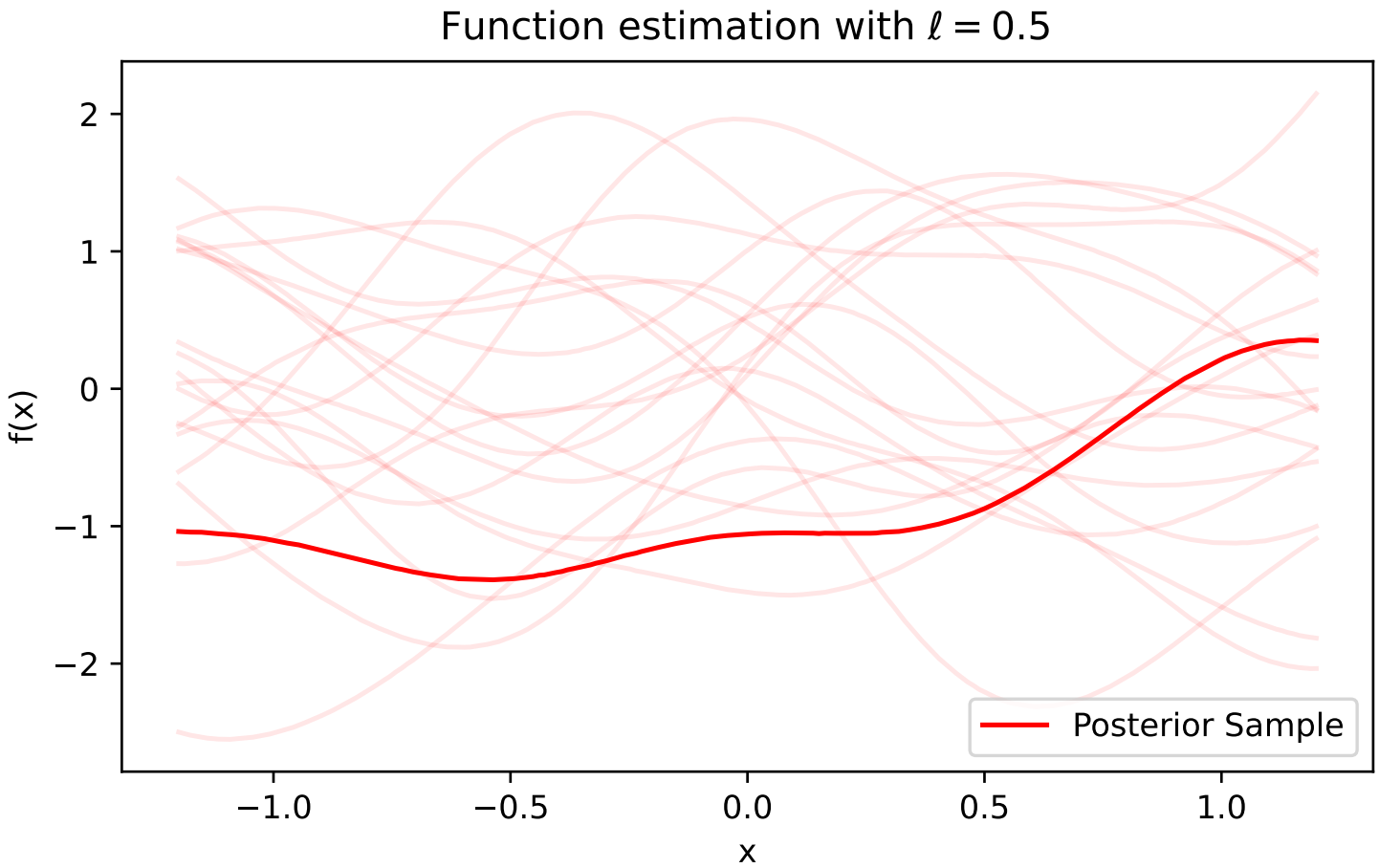
A collection of random variables $\{f(\mathbf{x})\}_{\mathbf{x}\in\mathbb{R}^d}$ is a Gaussian process if all finite dimensional distributions are multivariate normal distributed. That is, there is a function $m : \mathbf{x} \to \mathbb{R}$ and kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for all finite sets $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$,

$$\begin{bmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_n) \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} m(\mathbf{x}_1) \\ m(\mathbf{x}_2) \\ \vdots \\ m(\mathbf{x}_n) \end{bmatrix}, \mathbf{K} \right)$$

where

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \ldots & k(\mathbf{x}_1, \mathbf{x}_n) \\ k(\mathbf{x}_2, \mathbf{x}_1) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \ldots & \ldots & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}$$

Gaussian Process Example Realisations



THE UNIVERSITY OF MELBOURNE

WEHI
brighter together

Covariance Kernel Motivation

- Kernel determines the amount of covariance between sets of indices.
- When the distance between indices is small, covariance needs to be large

Common Covariance Kernels

- Matern Kernel

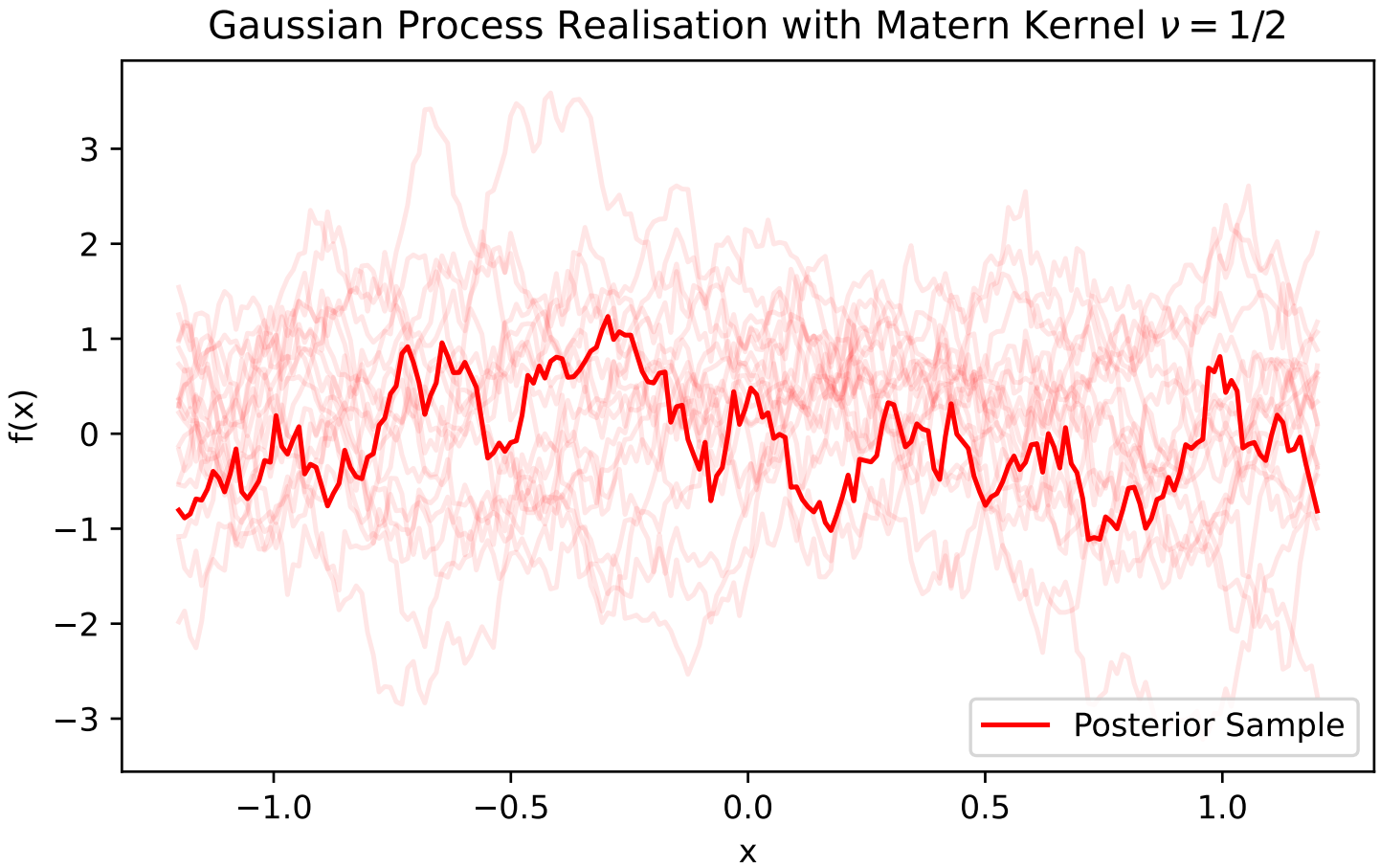
$$k_\nu(x, x') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}||x - x'||}{\ell} \right)^\nu K_\nu \left(-\frac{\sqrt{2\nu}||x - x'||}{\ell} \right)$$

  where K_ν is a modified Bessel function ($||\cdot||$ is the euclidean distance)
- $\lfloor \nu \rfloor$ times mean square differentiable.
- As $\nu \to \infty$ you get squared exponential covariance function, which results in realisations that are infinitely mean square differentiable:

$$k(x, x') = \sigma^2 \exp(-\frac{||x - x'||^2}{\ell})$$

Time to wake up

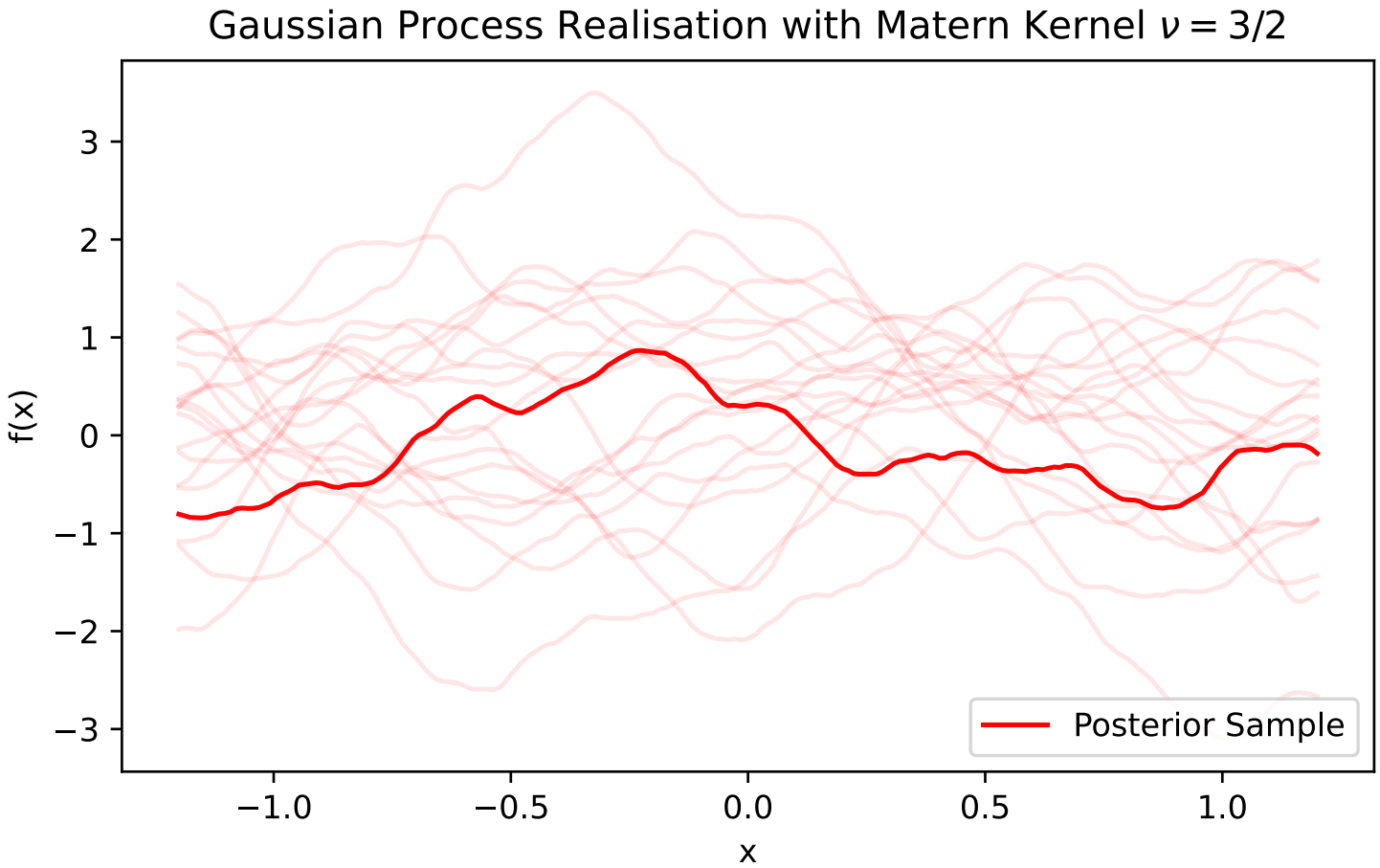
Kernel Choices - Kernel Type



THE UNIVERSITY OF MELBOURNE

WEHI brighter together

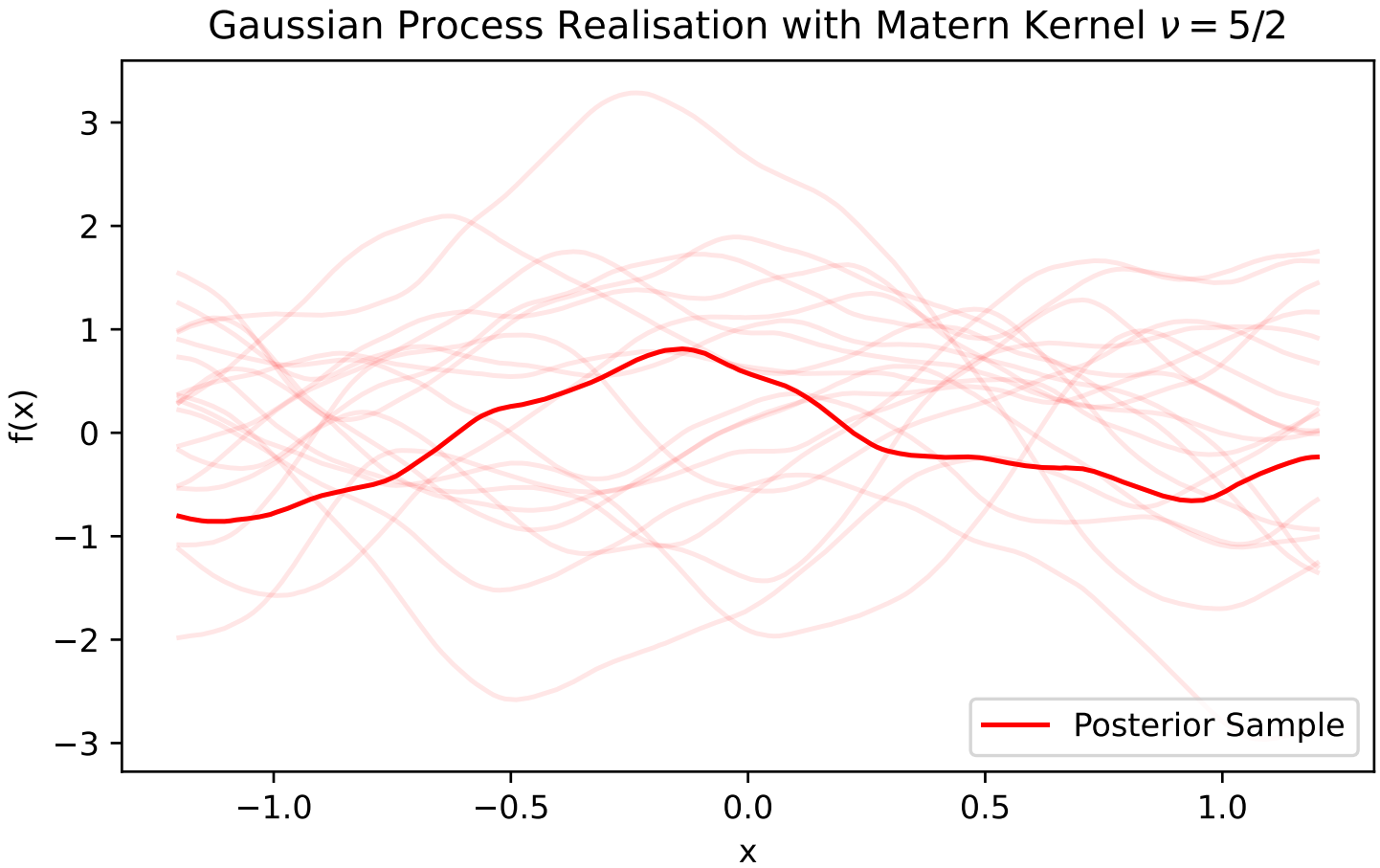
Kernel Choices - Kernel Type



THE UNIVERSITY OF MELBOURNE

WEHI brighter together

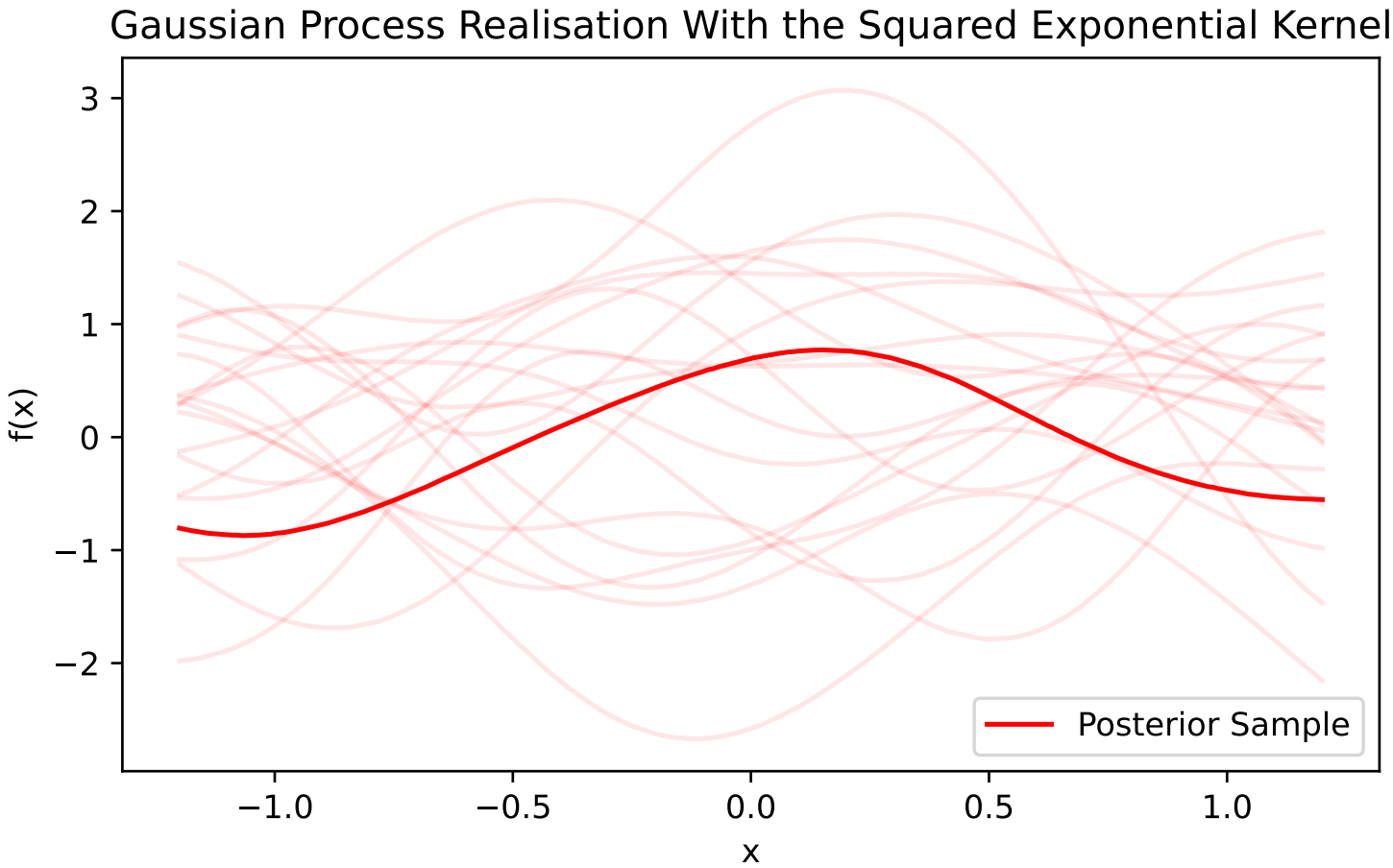
Kernel Choices - Kernel Type



THE UNIVERSITY OF MELBOURNE

WEHI brighter together

Kernel Choices - Kernel Type



THE UNIVERSITY OF MELBOURNE

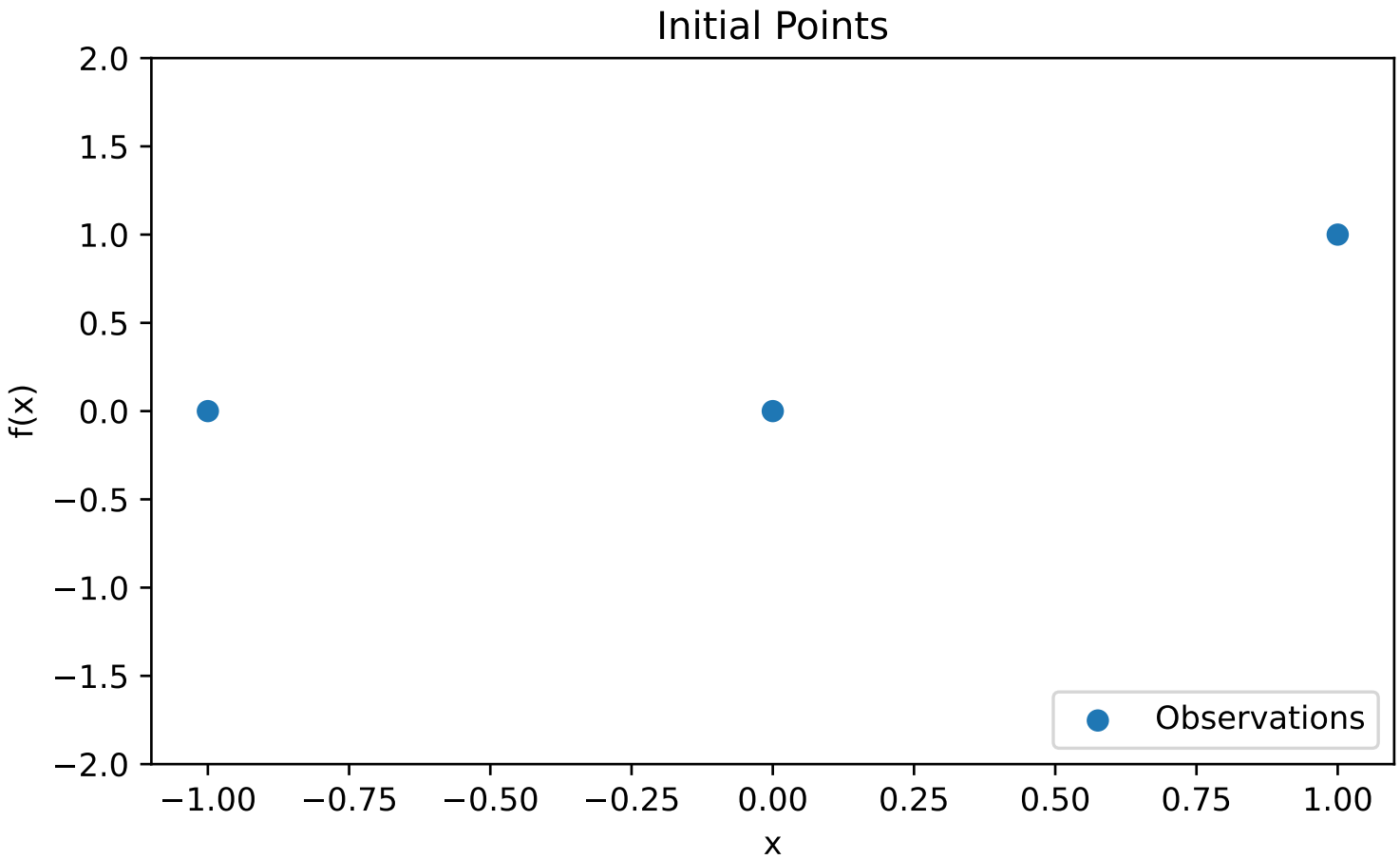
WEHI
brighter together

Fitting our GP to data

GPs are 'priors'

Fitting our GP to data

GPs are 'priors'

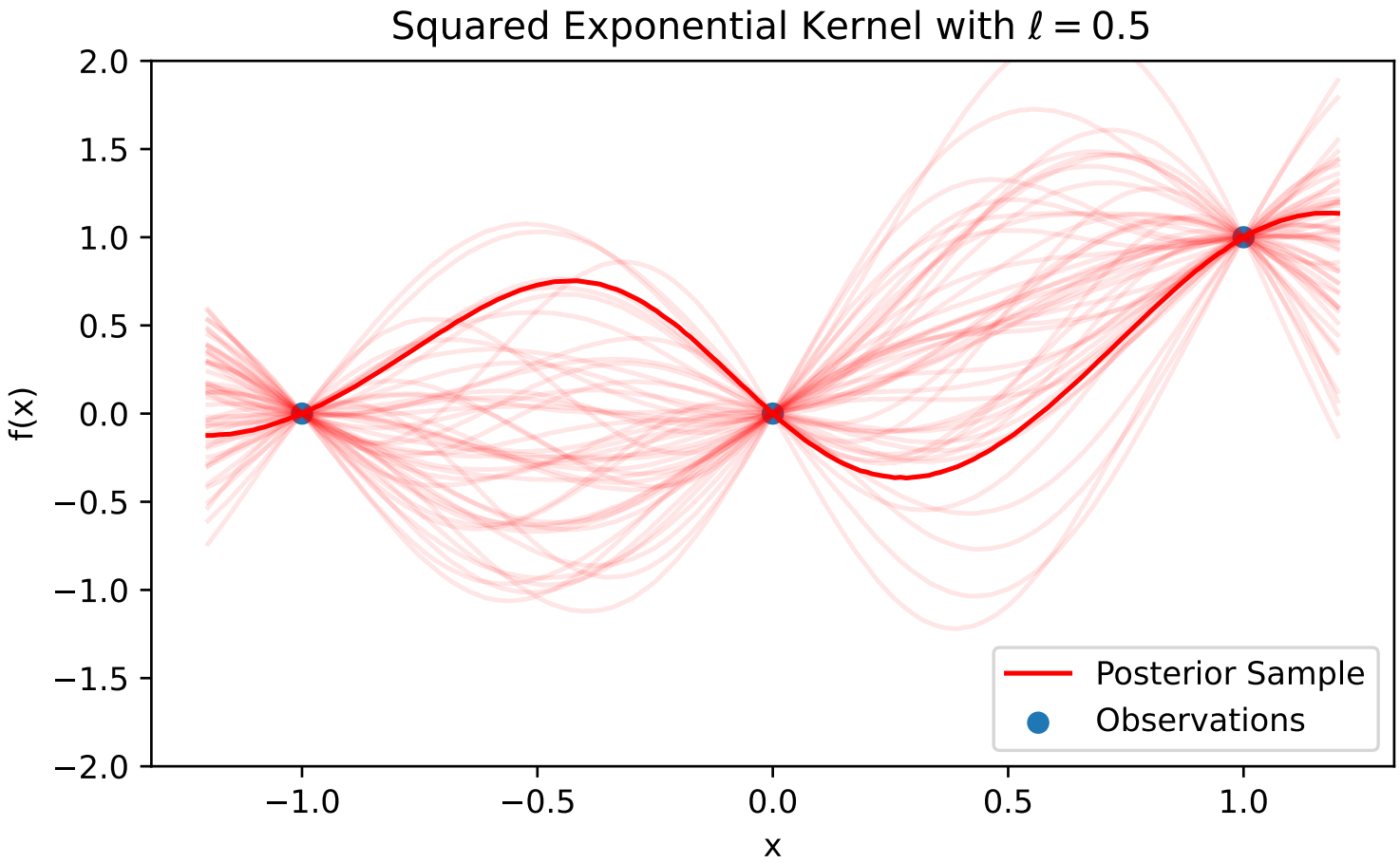


THE UNIVERSITY OF MELBOURNE

WEHI
brighter together

Fitting our GP to data

GPs are 'priors'



THE UNIVERSITY OF MELBOURNE

WEHI brighter together

Fitting our GP to data

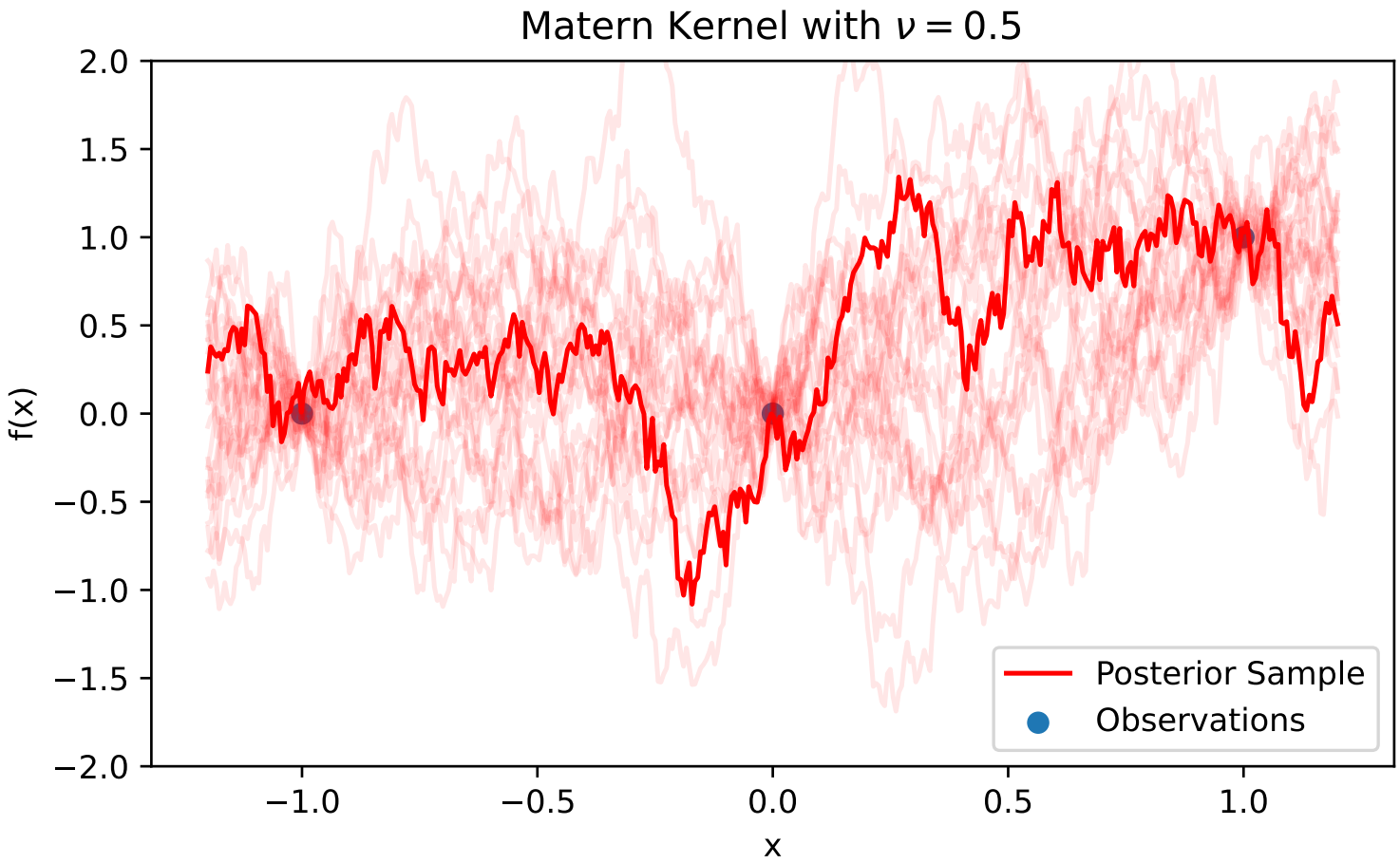
GPs are 'priors'

Fitting our GP to data

GPs are 'priors'

Fitting our GP to data

GPs are 'priors'



THE UNIVERSITY OF MELBOURNE

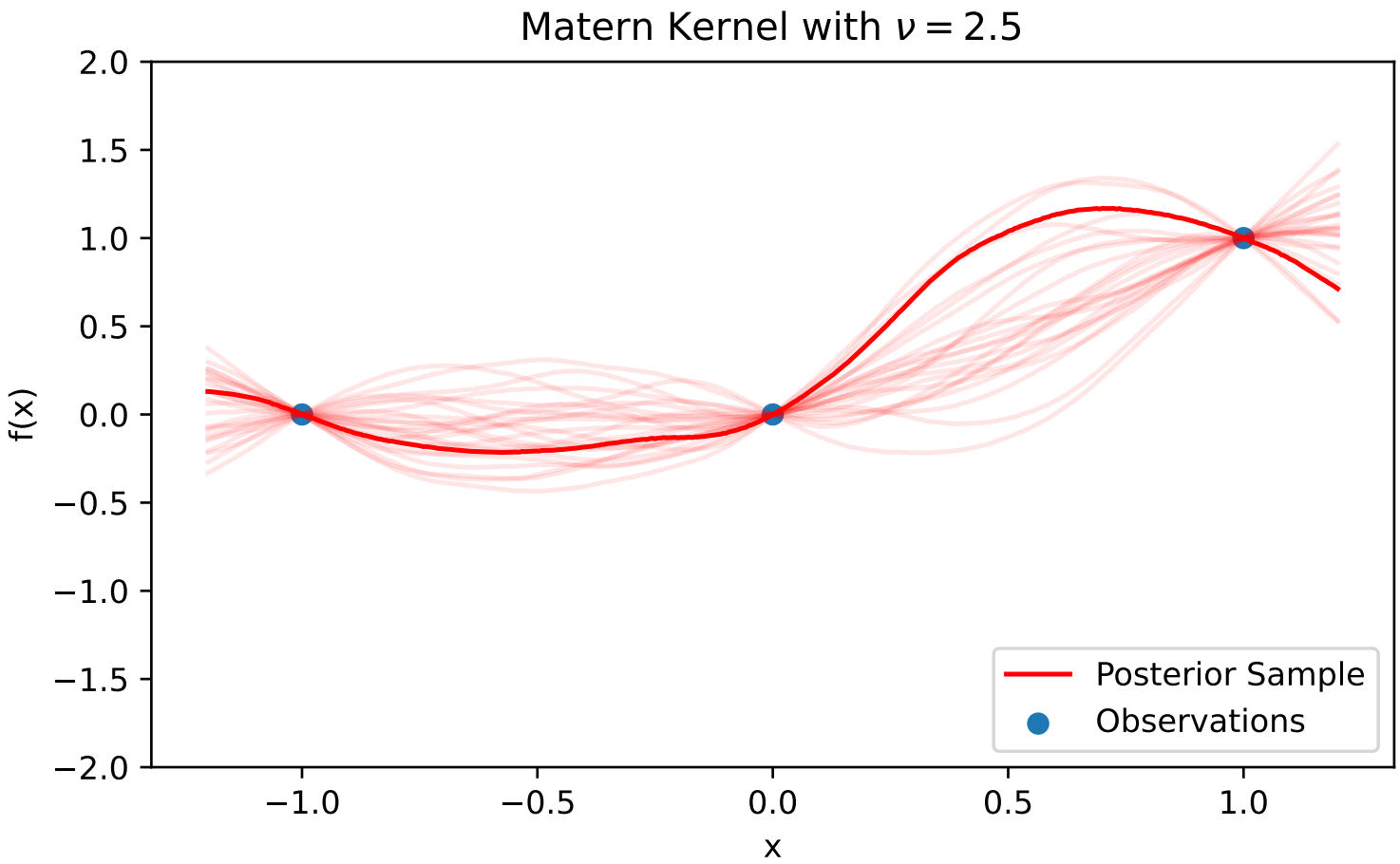
WEHI brighter together

Fitting our GP to data

GPs are 'priors'

Fitting our GP to data

GPs are 'priors'

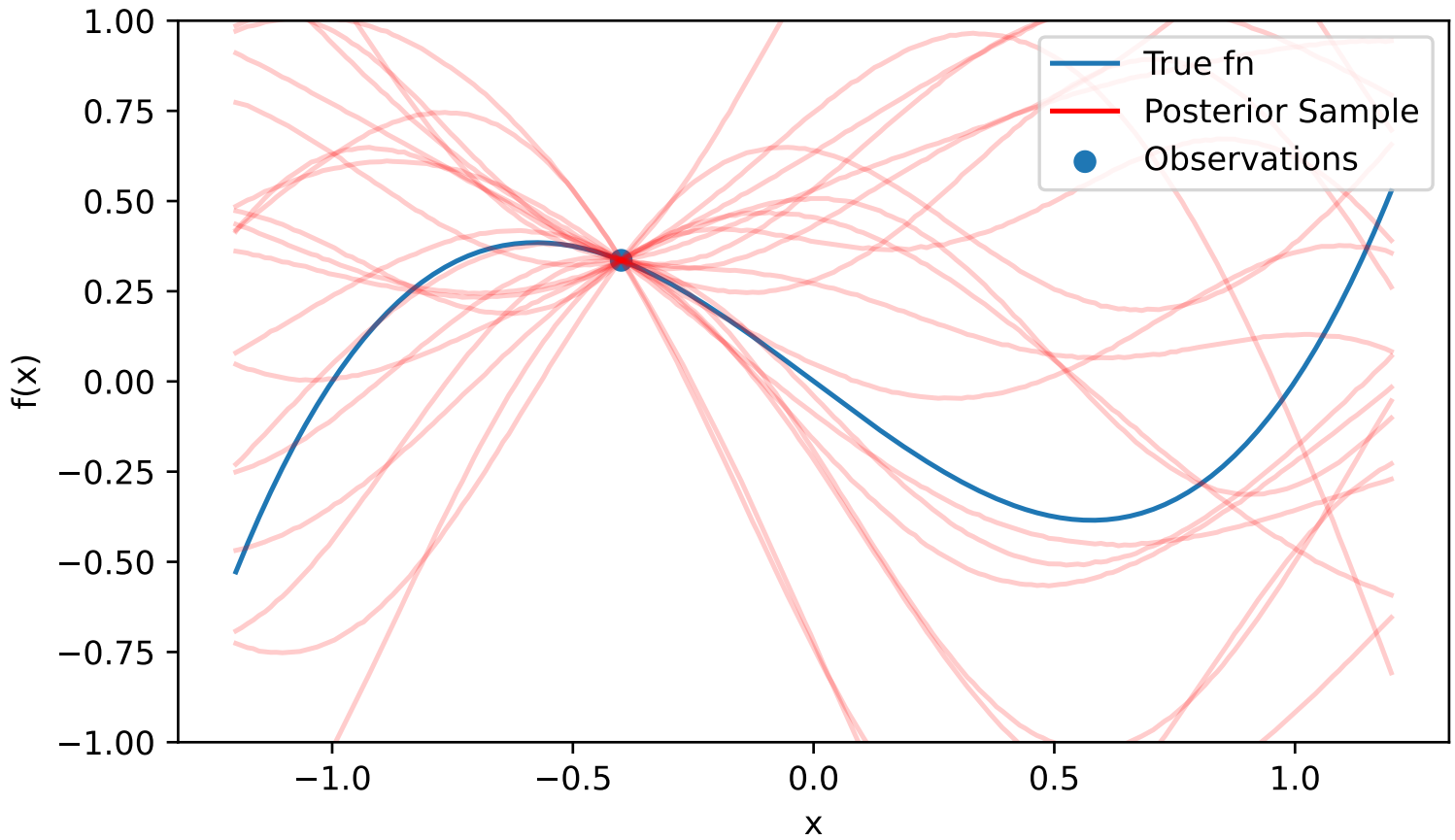


THE UNIVERSITY OF MELBOURNE

WEHI brighter together

GP regression on $x(x-1)(x+1)$

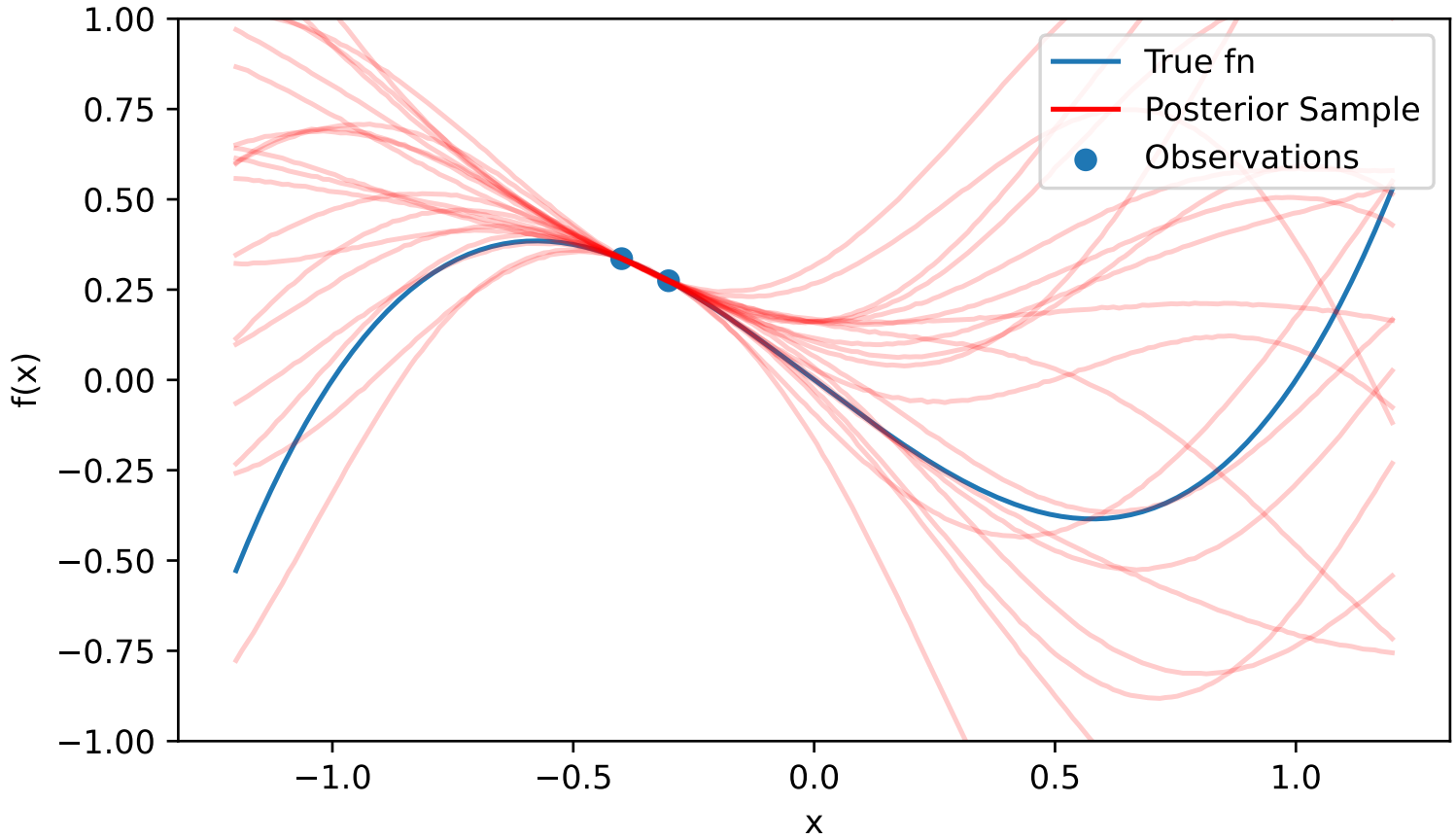
GP regression on $x(x-1)(x+1)$



THE UNIVERSITY OF MELBOURNE

WEHI
brighter together

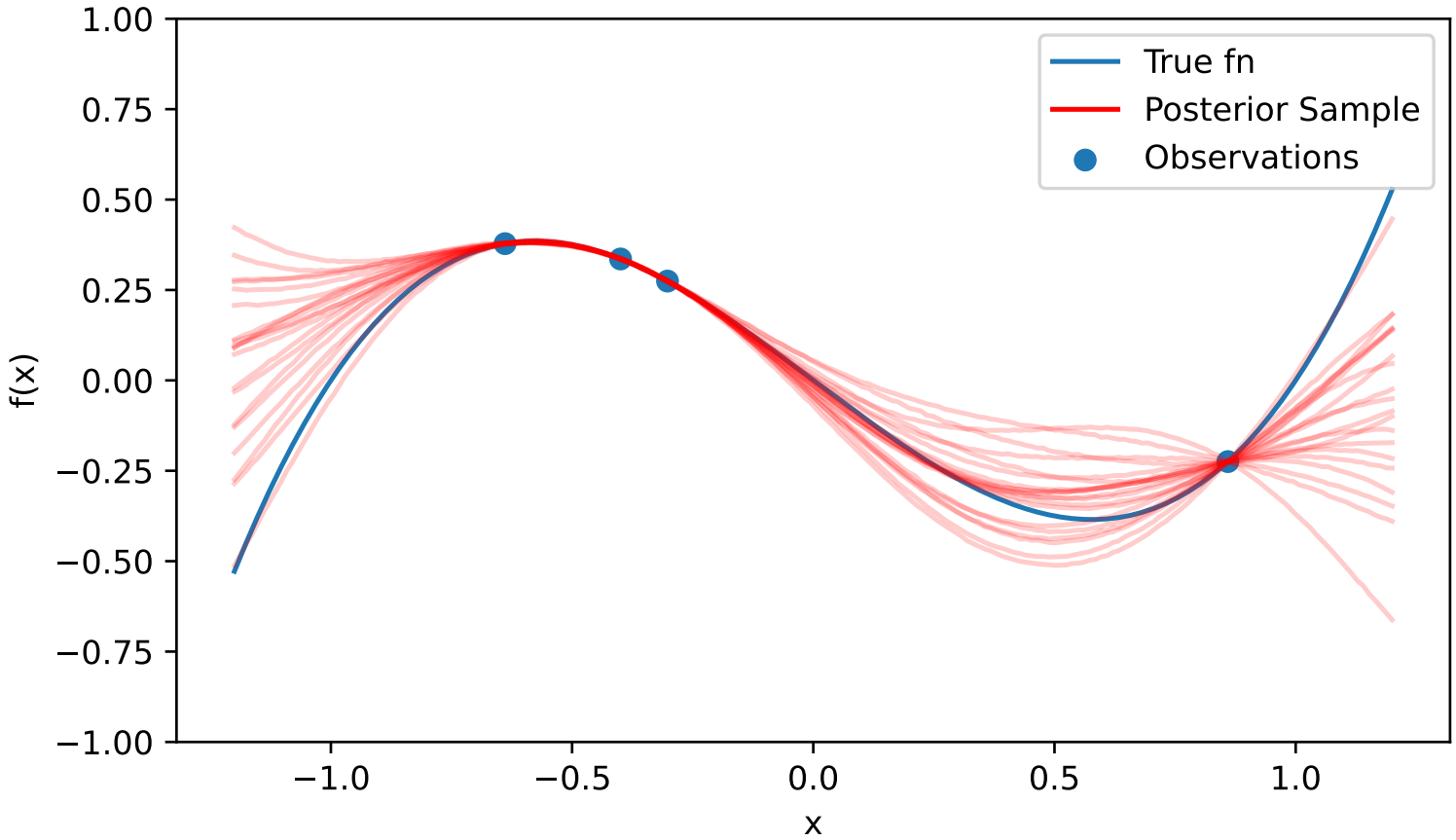
GP regression on $x(x-1)(x+1)$



THE UNIVERSITY OF MELBOURNE

WEHI
brighter together

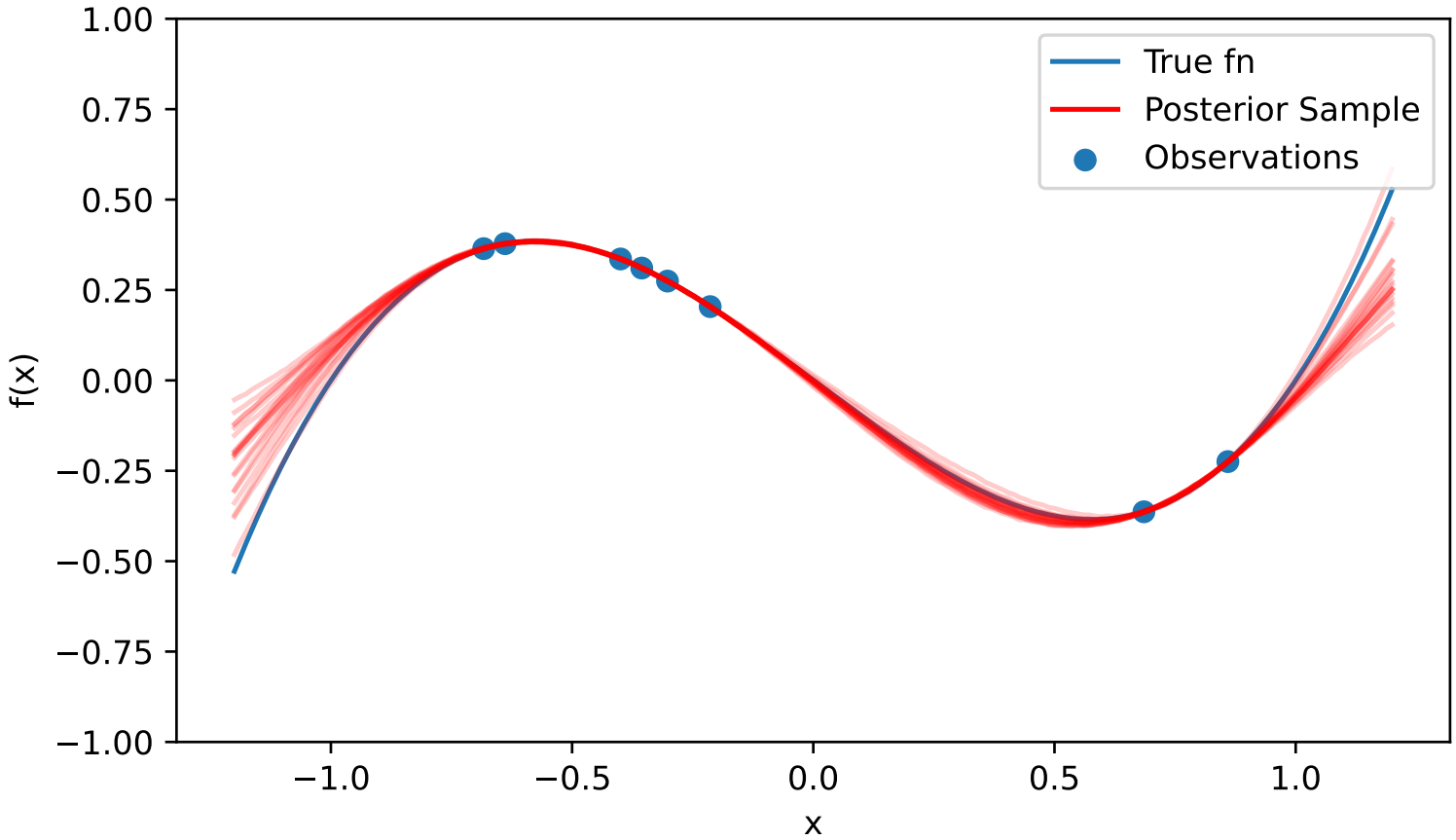
GP regression on $x(x-1)(x+1)$



THE UNIVERSITY OF MELBOURNE

WEHI brighter together

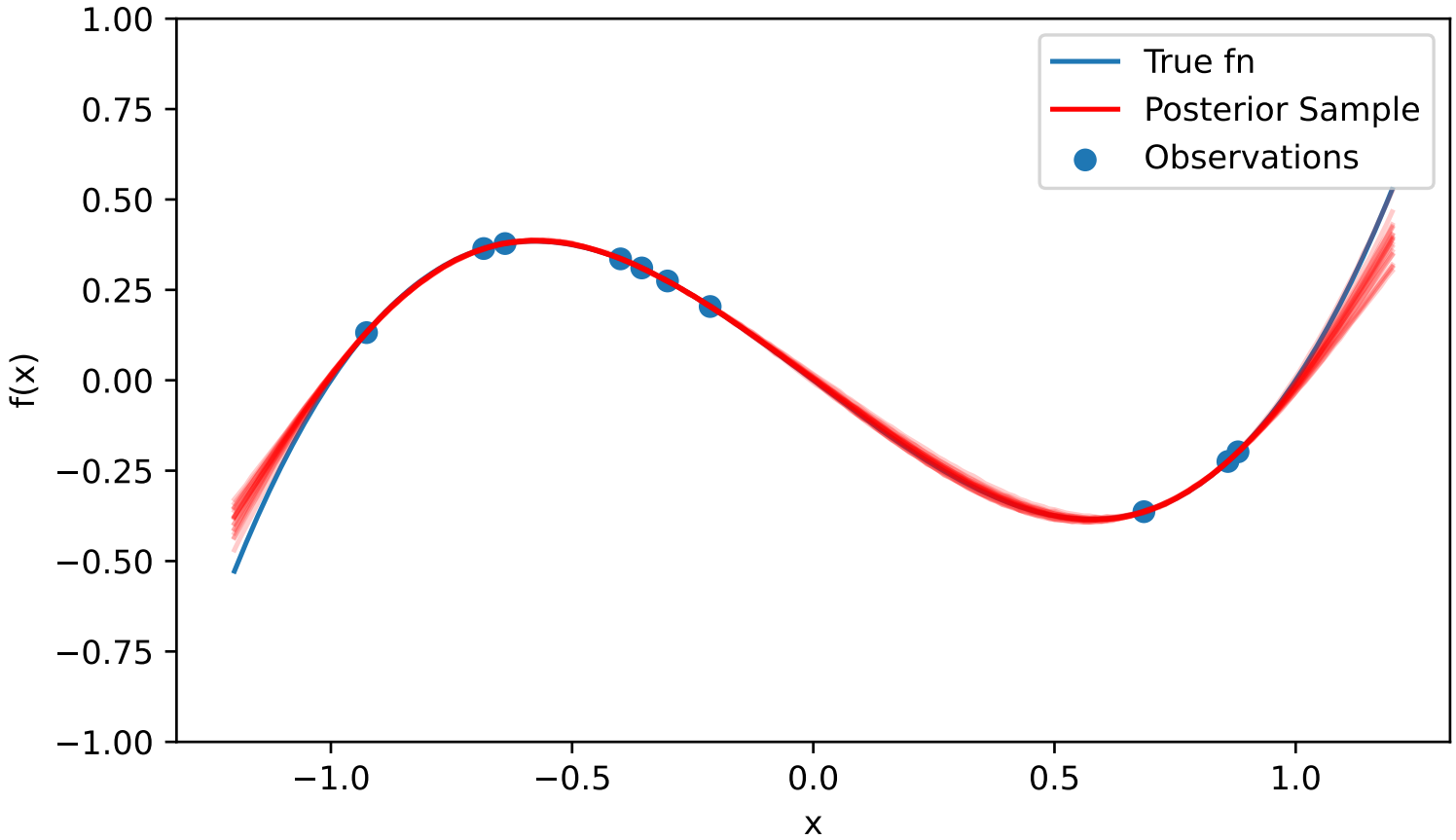
GP regression on $x(x-1)(x+1)$



THE UNIVERSITY OF MELBOURNE

WEHI brighter together

GP regression on $x(x-1)(x+1)$



THE UNIVERSITY OF MELBOURNE

WEHI brighter together

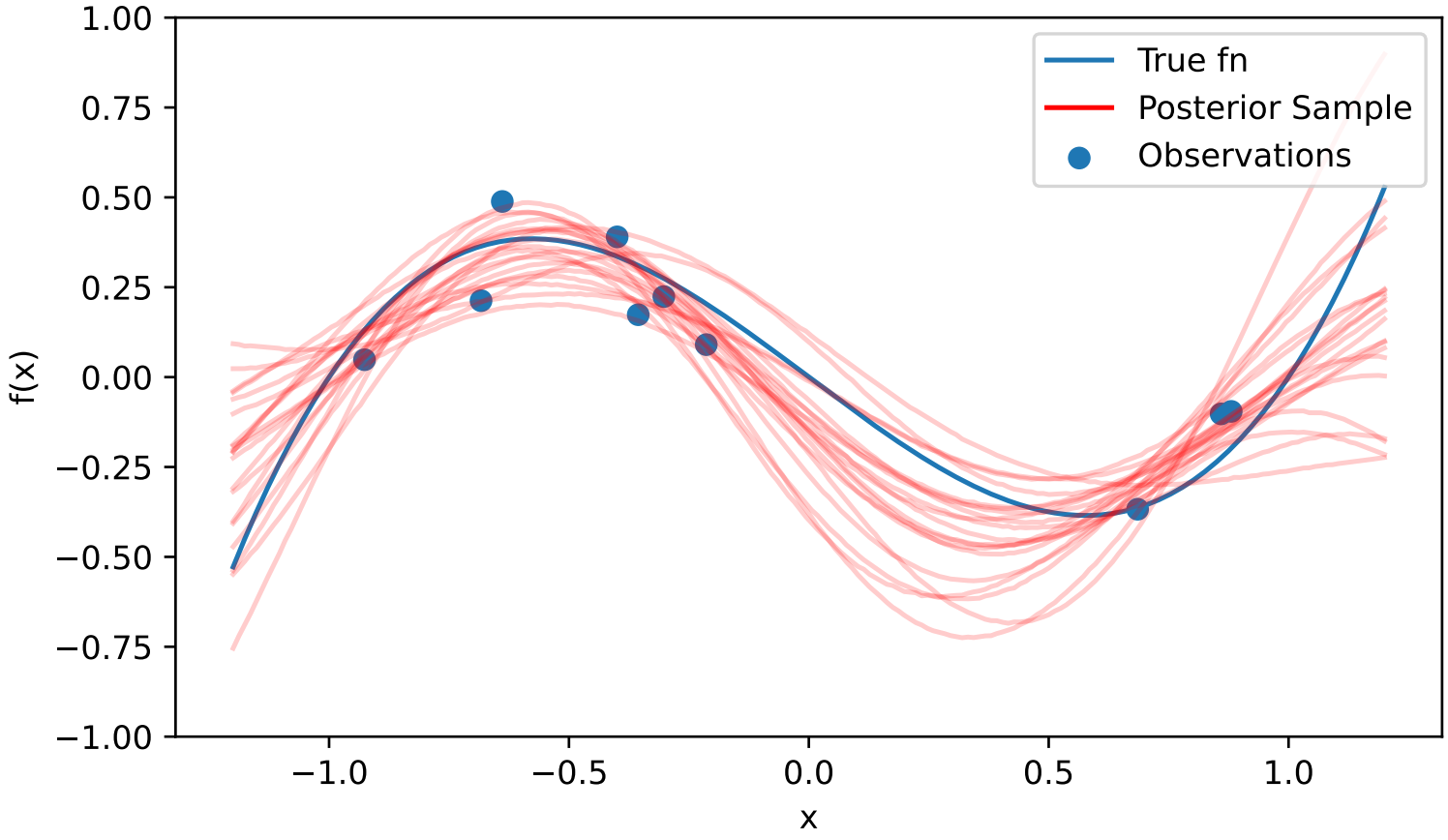
What if we have noise?

Add observation variance σ^2,

$$\begin{bmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_n) \end{bmatrix} \sim \text{MVN} \left(\begin{bmatrix} m(\mathbf{x}_1) \\ m(\mathbf{x}_2) \\ \vdots \\ m(\mathbf{x}_n) \end{bmatrix}, \mathbf{K} + \sigma^2 \mathbf{I}_n \right)$$

GP regression on $x(x-1)(x+1)$

GP regression on $x(x-1)(x+1)$



THE UNIVERSITY OF MELBOURNE

WEHI
brighter together

Overall Idea again

- Can we predict $\mathcal{D}(\boldsymbol{\theta})$ for simulated $\boldsymbol{\theta}$

Overall Idea again

- Can we predict $\mathcal{D}(\boldsymbol{\theta})$ for simulated $\boldsymbol{\theta}$
- $\mathcal{D}(\boldsymbol{\theta}) \approx \mathcal{D}(\boldsymbol{\theta}')$ for $\boldsymbol{\theta}$, $\boldsymbol{\theta}$ close.

Overall Idea again

- Can we predict $\mathcal{D}(\boldsymbol{\theta})$ for simulated $\boldsymbol{\theta}$
- Use Gaussian process to predict discrepency function

About Vivax Malaria

- Has dormant liver stage on top of blood stage infection

Champagne Model Parameters

- α : proportion of those infected but cleared of blood stage infections (through treatment)
- β : a further proportion that are also cleared of liver stage parasites, given that they were also cleared of blood stage infection (radical cure)
- λ : the rate of infection
- γ_L : rate of clearance of liver stage disease
- f : rate of relapse
- r : rate of blood stage clearance
- δ : importation rate (which we assume is 0)

Champagne Model Transition Rates

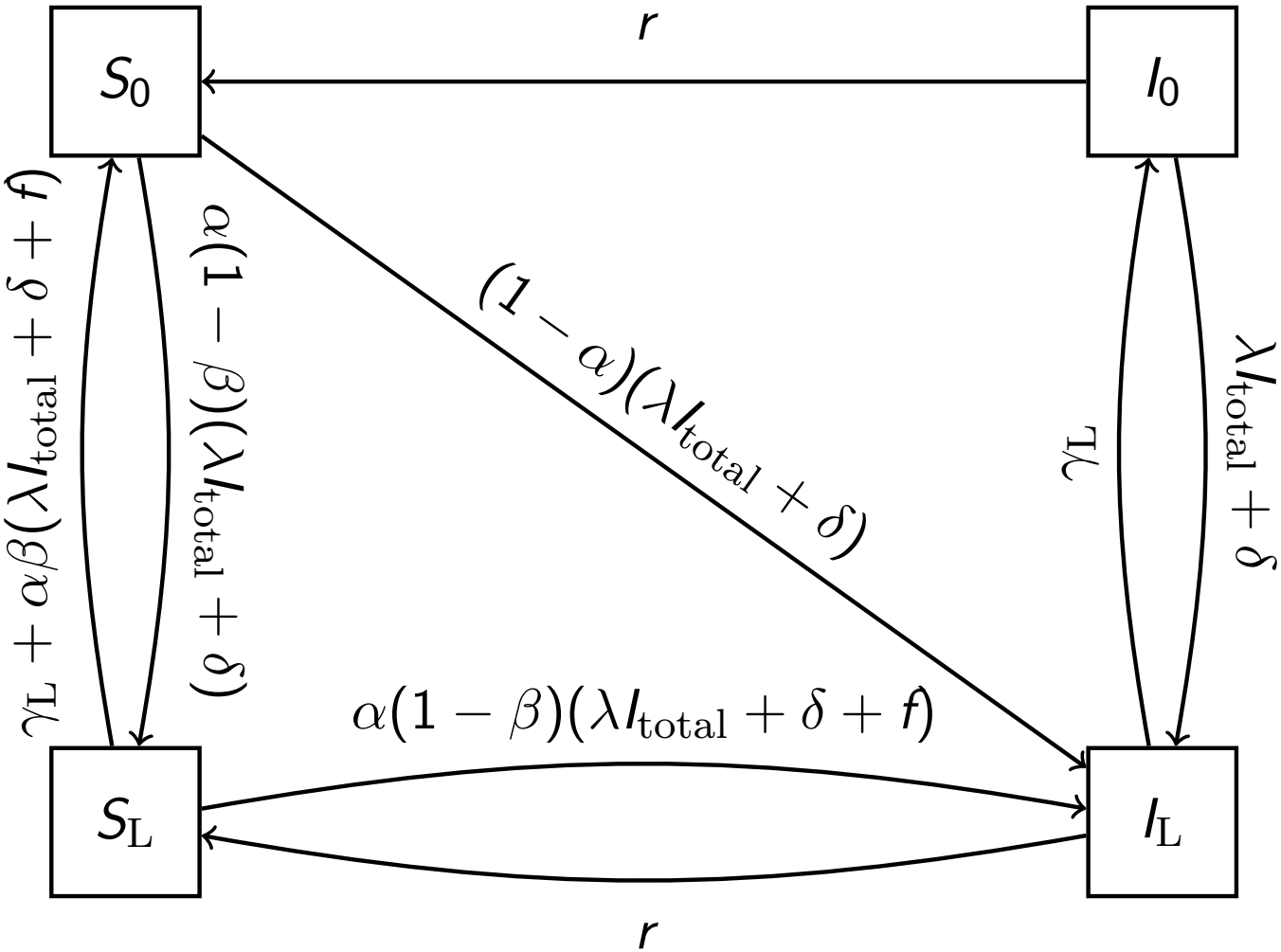


Figure: *P. vivax* model described by Champagne et al. 2022

Champagne ODEs

$$
\begin{aligned}
\frac{\mathrm{d}I_{\mathrm{L}}}{\mathrm{d}t} =& (1-\alpha)(\lambda I_{\mathrm{total}}+\delta)(S_0+S_{\mathrm{L}})+(\lambda I_{\mathrm{total}}+\delta)I_0 \\
&+(1-\alpha)fS_{\mathrm{L}}-\gamma_{\mathrm{L}}I_{\mathrm{L}}-rI_{\mathrm{L}} \\
\frac{\mathrm{d}I_0}{\mathrm{d}t} =& -(\lambda I_{\mathrm{total}}+\delta)I_0+\gamma_{\mathrm{L}}I_{\mathrm{L}}-rI_0 \\
\frac{\mathrm{d}S_{\mathrm{L}}}{\mathrm{d}t} =& -(1-\alpha(1-\beta))(\lambda I_{\mathrm{total}}+\delta+f)S_{\mathrm{L}}+\alpha(1-\beta)(\lambda I_{\mathrm{total}} \\
&+\delta)S_0-\gamma_{\mathrm{L}}S_{\mathrm{L}}+rI_{\mathrm{L}} \\
\frac{\mathrm{d}S_0}{\mathrm{d}t} =& -(1-\alpha\beta)(\lambda I_{\mathrm{total}}+\delta)S_0+(\lambda I_{\mathrm{total}}+\delta)\alpha\beta S_{\mathrm{L}}+\alpha\beta fS_{\mathrm{L}} \\
&+\gamma_{\mathrm{L}}S_{\mathrm{L}}+rI_0
\end{aligned}
$$

THE UNIVERSITY OF MELBOURNE

WEHI brighter together

Model Calibration Data

- Doob-Gillespie algorithm with paper parameters reported in the original paper for 'observed data', 10 initial infections, 1000 people.

Model Calibration Data

- Doob-Gillespie algorithm with paper parameters reported in the original paper for 'observed data', 10 initial infections, 1000 people.
- $S(\mathbf{X}_{\text{obs}}) := \{w_{\text{obs}}, p_{\text{obs}}, m_{\text{obs}}\}$
  - w_{obs} : weekly incidence around (stochastic) equilibrium
  - p_{obs} : prevalence around (stochastic) equilibrium
  - m_{obs} : incidence in the first month of the epidemic

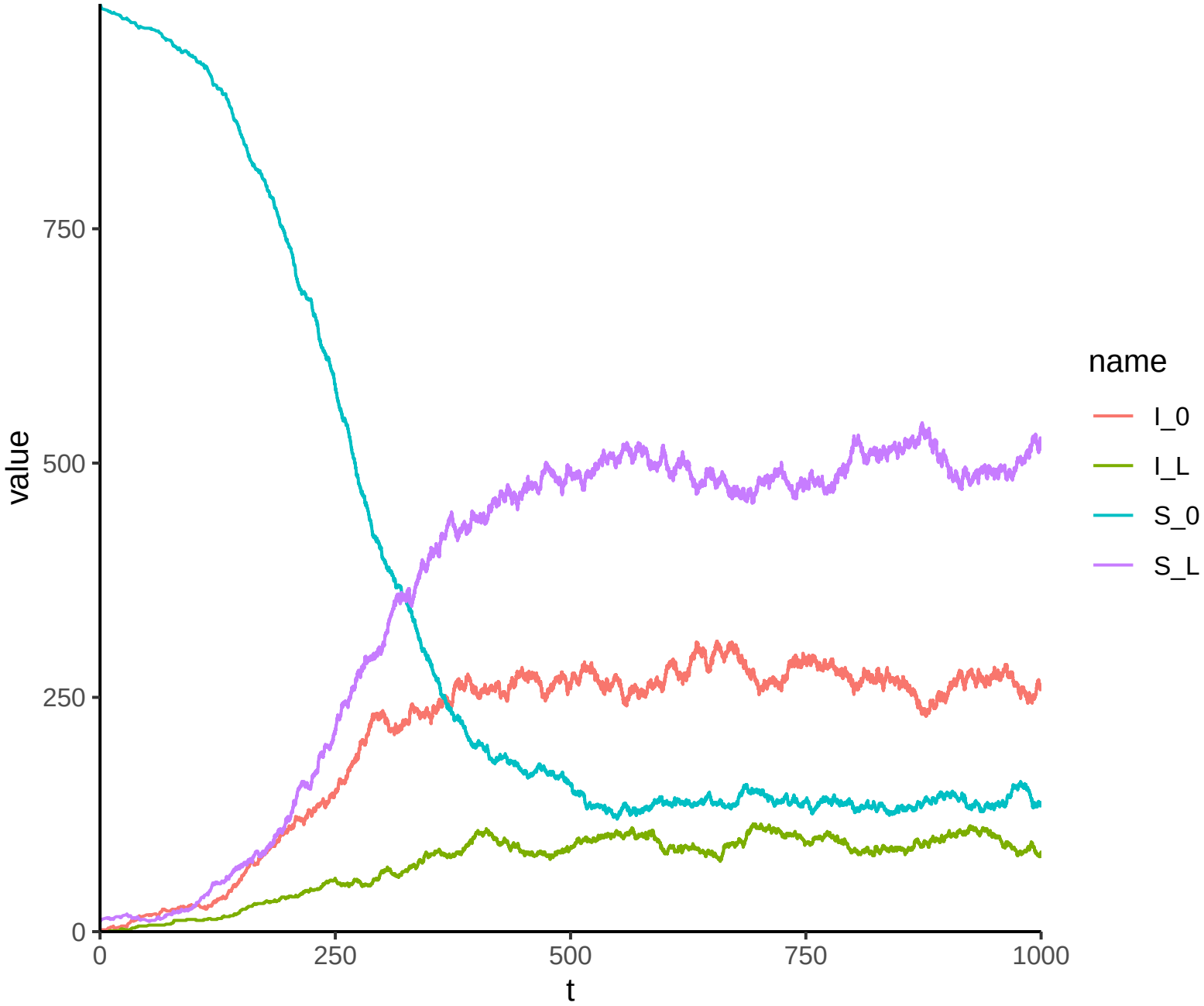
Model Calibration Data

- Doob-Gillespie algorithm with paper parameters reported in the original paper for 'observed data', 10 initial infections, 1000 people.
- $S(\mathbf{X}_{\text{obs}}) := \{w_{\text{obs}}, p_{\text{obs}}, m_{\text{obs}}\}$
  - w_{obs} : weekly incidence around (stochastic) equilibrium
  - p_{obs} : prevalence around (stochastic) equilibrium
  - m_{obs} : incidence in the first month of the epidemic
- $\mathcal{D}(\boldsymbol{\theta}) := \left|\frac{w_{\text{obs}}-w}{w_{\text{obs}}}\right| + \left|\frac{p_{\text{obs}}-p}{p_{\text{obs}}}\right| + \left|\frac{m_{\text{obs}}-m}{m_{\text{obs}}}\right|$
  - (L_1 norm on the relative differences)

THE UNIVERSITY OF MELBOURNE

WEHI brighter together

Example Simulation



What's the Bayesian part?

- Bayesian acquisition function $\arg\min_{\boldsymbol{\theta}} A(\boldsymbol{\theta})$

What's the Bayesian part?

- Bayesian acquisition function $\arg\min_{\boldsymbol{\theta}} A(\boldsymbol{\theta})$
- Gutmann and Cor 2016 uses

$$\mu(\boldsymbol{\theta}) - \eta_t \sqrt{\mathrm{v}(\boldsymbol{\theta})}$$

  - $\eta_t := \sqrt{c + 2\ln(t^{d/2+2})}$, and c can be chosen
  - $\mu(\boldsymbol{\theta})$ and $\mathrm{v}(\boldsymbol{\theta})$ are the posterior mean and variance

What's the Bayesian part?

- Bayesian acquisition function $\arg\min_{\boldsymbol{\theta}} A(\boldsymbol{\theta})$
- Gutmann and Cor 2016 uses

$$\mu(\boldsymbol{\theta}) - \eta_t \sqrt{\mathrm{v}(\boldsymbol{\theta})}$$

  - $\eta_t := \sqrt{c + 2\ln(t^{d/2+2})}$, and c can be chosen
  - $\mu(\boldsymbol{\theta})$ and $\mathrm{v}(\boldsymbol{\theta})$ are the posterior mean and variance
- Could use expected information

$$(\mu_{\min} - \mu(\boldsymbol{\theta}))\Phi\left(\frac{\mu_{\min} - \mu(\boldsymbol{\theta})}{\sqrt{\mathrm{v}(\boldsymbol{\theta})}}\right) + \sqrt{\mathrm{v}(\boldsymbol{\theta})}\phi\left(\frac{\mu_{\min} - \mu(\boldsymbol{\theta})}{\sqrt{\mathrm{v}(\boldsymbol{\theta})}}\right)$$

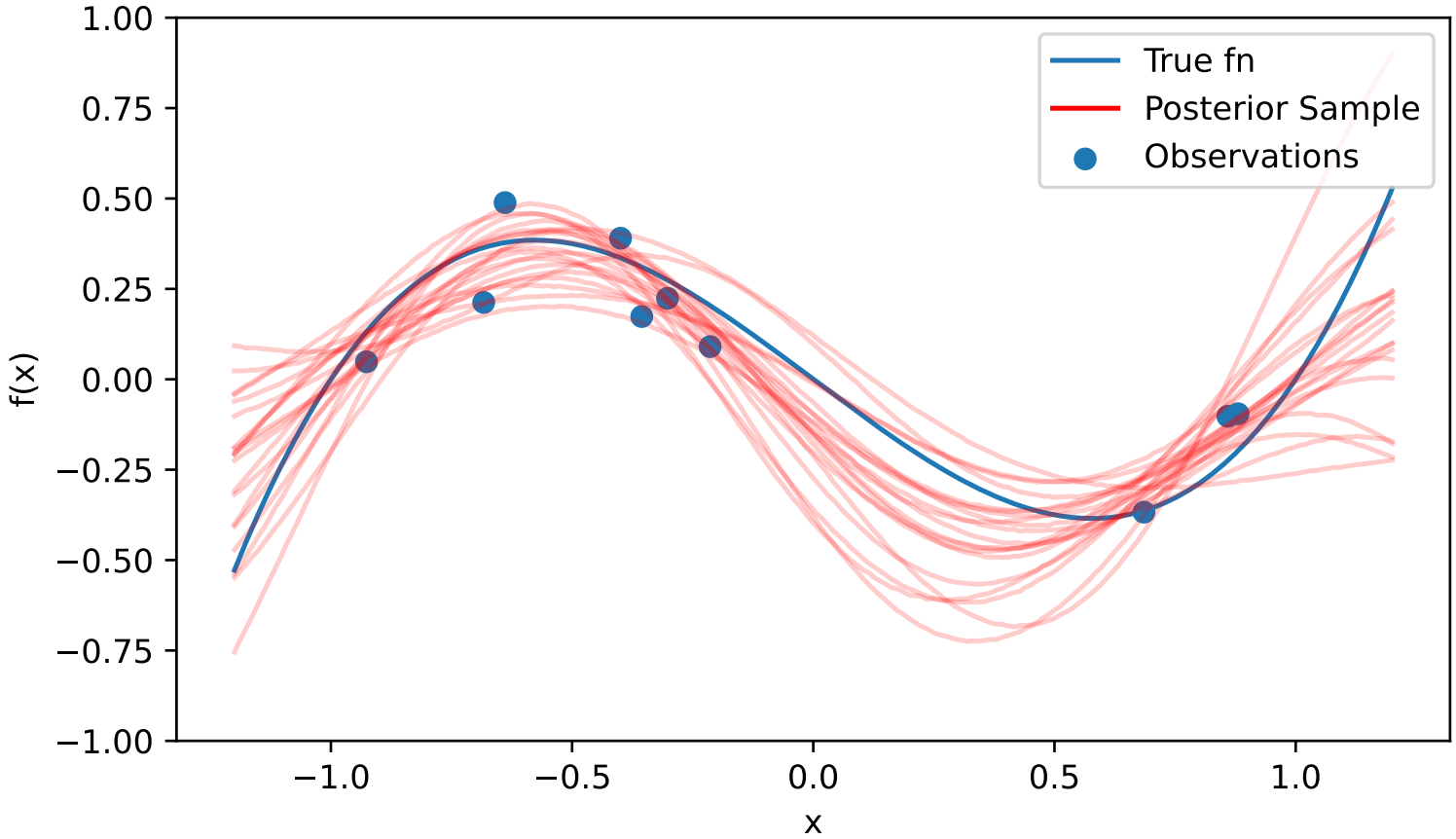
  - $\mu_{\min} := \min_{\boldsymbol{\theta}} \mu(\boldsymbol{\theta})$
  - Φ, ϕ CDF and PDF of standard normal

THE UNIVERSITY OF MELBOURNE

WEHI brighter together

Acquisition ‘example’

Acquisition ‘example’



THE UNIVERSITY OF MELBOURNE

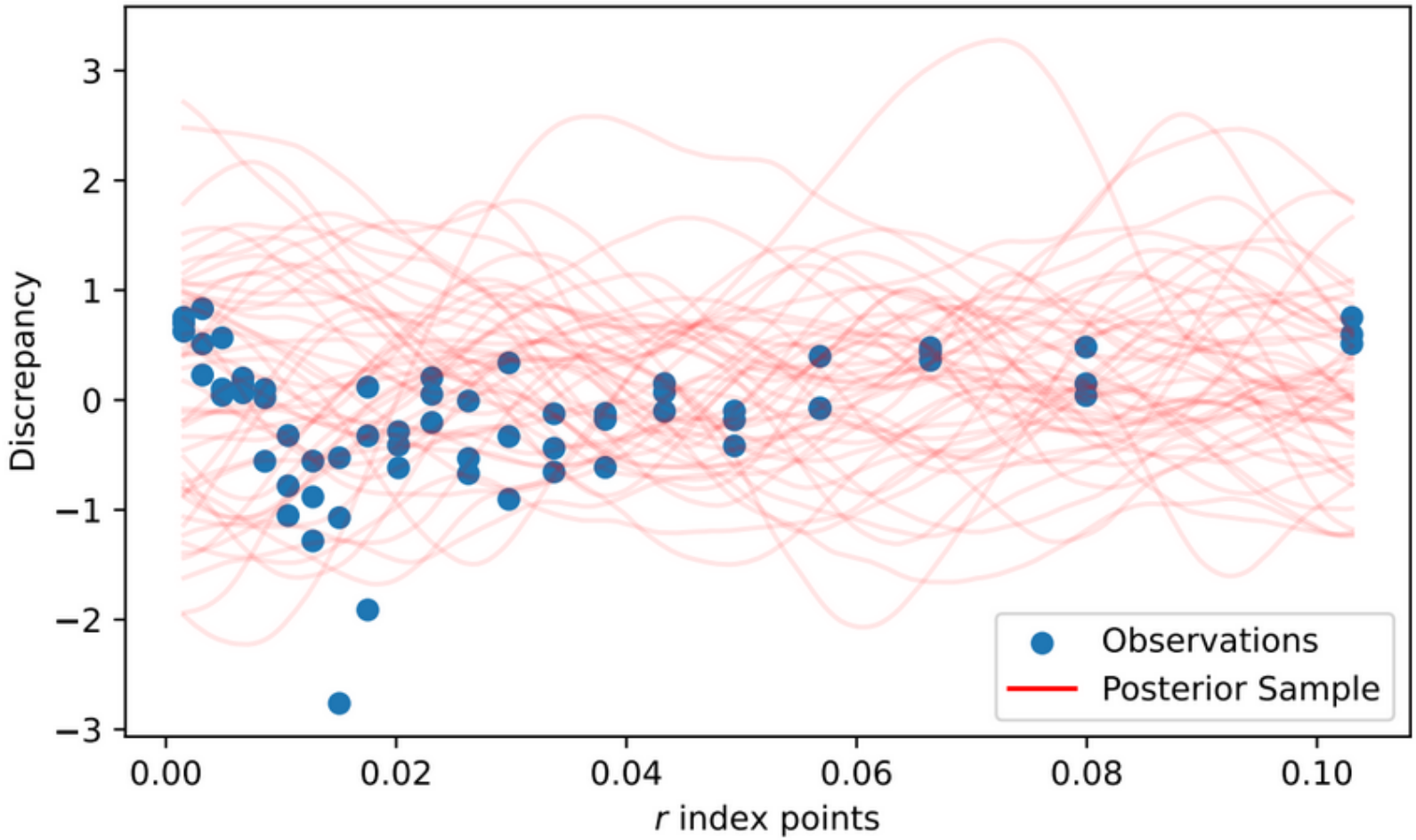
WEHI
brighter together

Synthetic Likelihood

- $L(\boldsymbol{\theta}|\mathbf{X}_{\text{obs}}) \approx P(\mathcal{D}_{\mathcal{GP}}(\theta) < \varepsilon)$ (up to a proportion) where $\mathcal{D}_{\mathcal{GP}}$ is the discrepency modelled the Gaussian process

How did it go?

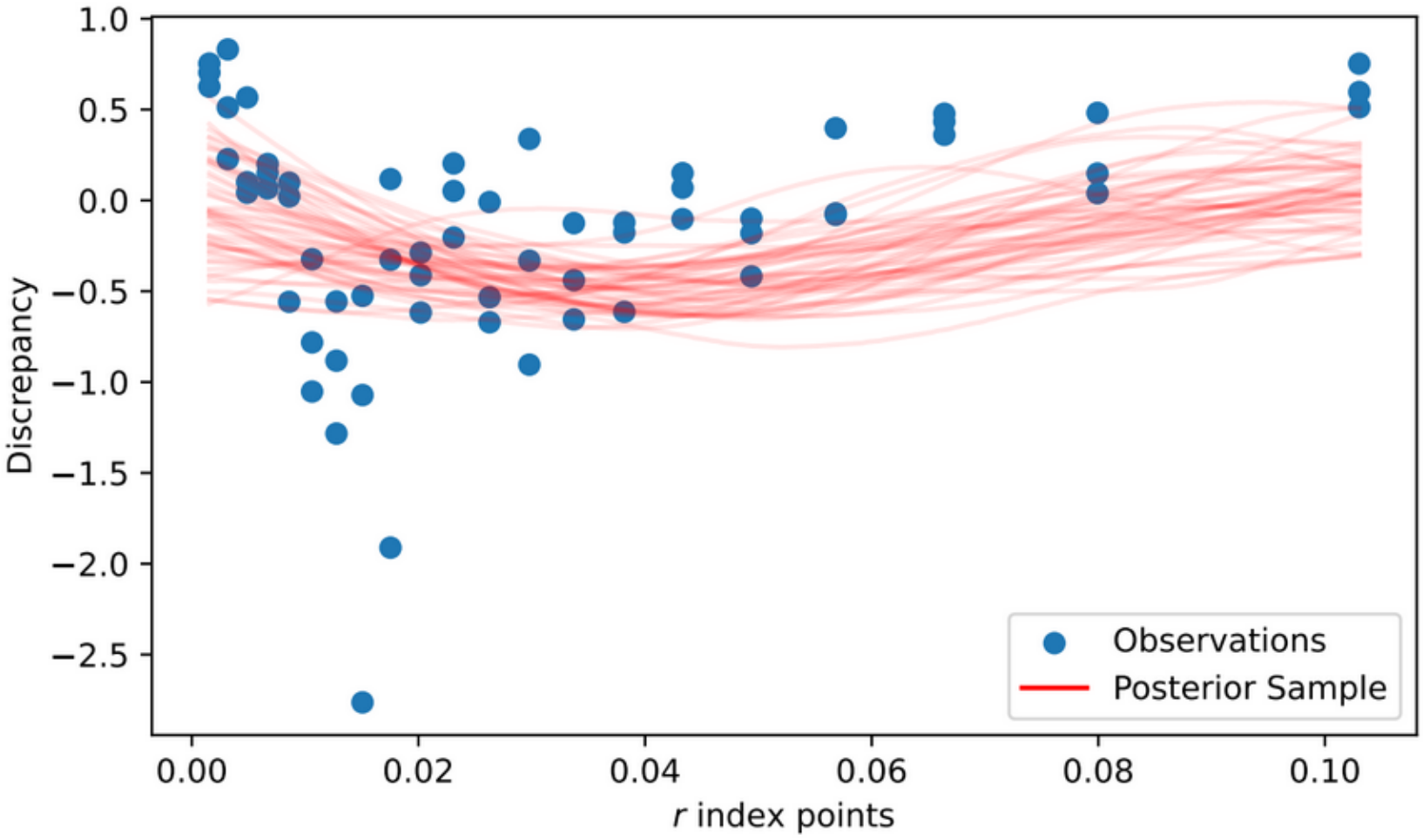
How did it go?



THE UNIVERSITY OF MELBOURNE

WEHI
brighter together

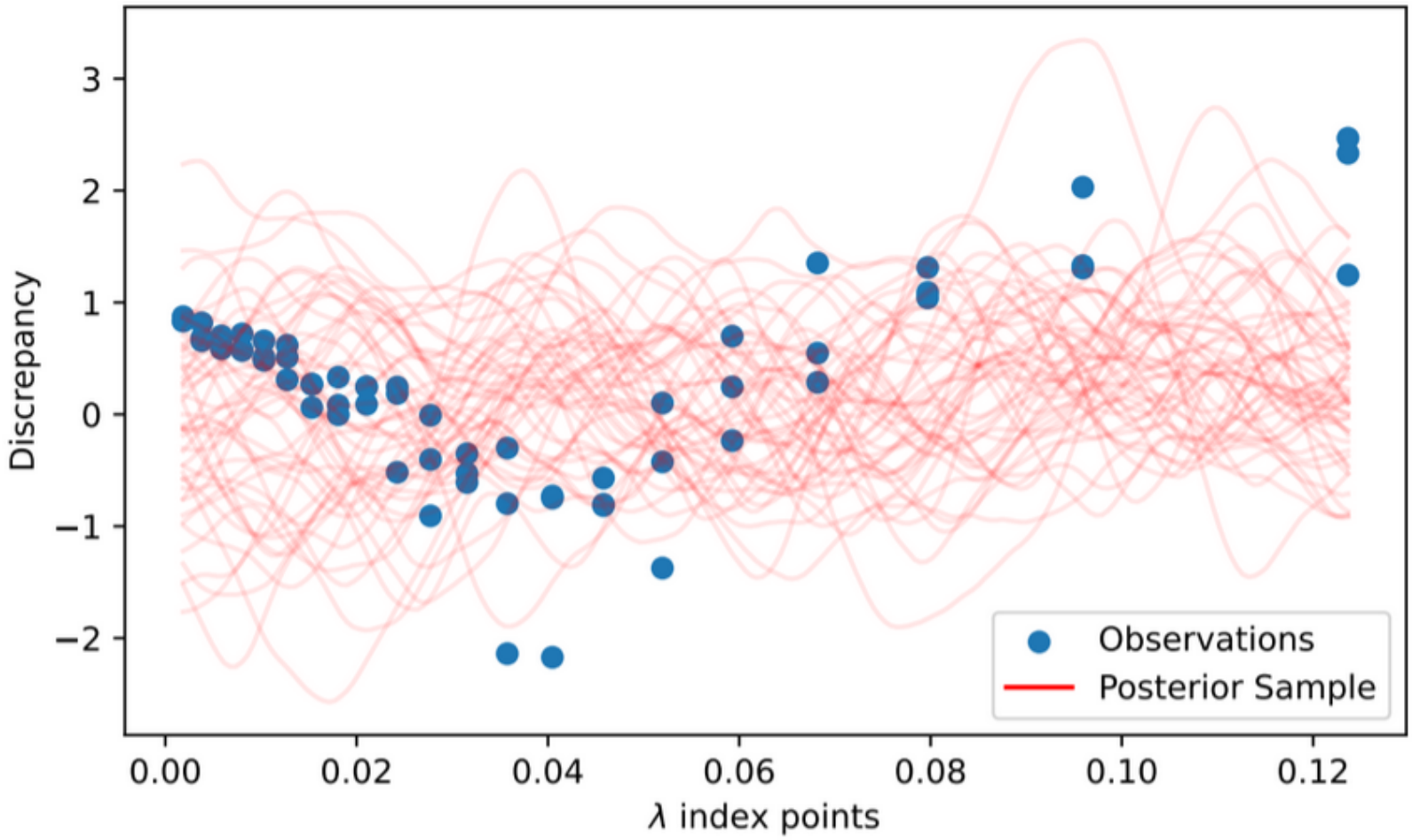
How did it go?



THE UNIVERSITY OF MELBOURNE

WEHI
brighter together

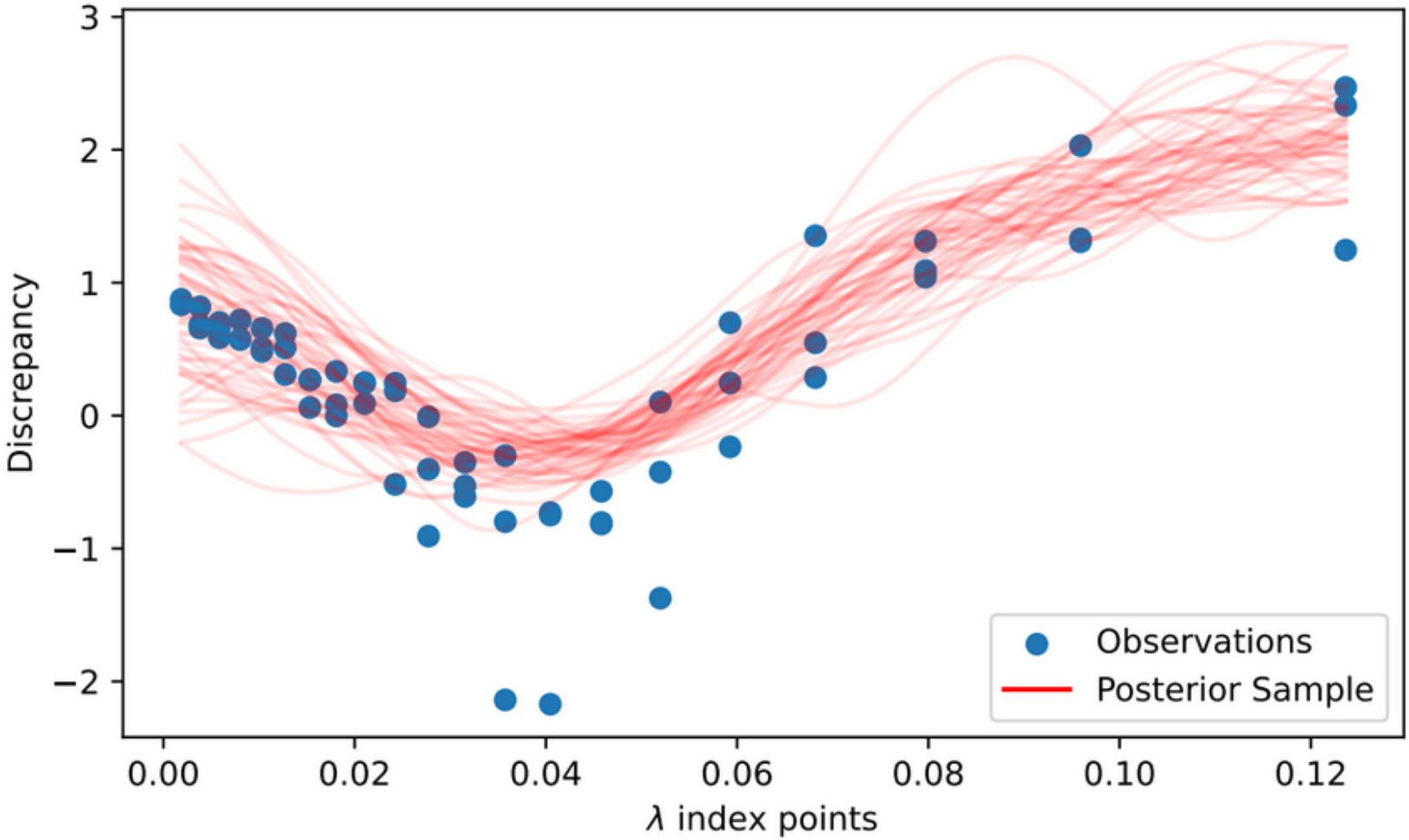
How did it go?



THE UNIVERSITY OF MELBOURNE

WEHI
brighter together

How did it go?



THE UNIVERSITY OF MELBOURNE

WEHI
brighter together

Big Problem (Big Solutions?)

- Observation variance is considered constant across the GP
  - Particularly a problem at the threshold
- Assumes that normal distribution approximates $\mathcal{D}(\boldsymbol{\theta})$
  - Log-normal might be more realistic?
- Jumps where there is threshold/bifurcation behaviour
  - Use Student $t-$Process to allow for jumps?

Big Problem (Big Solutions?)

- Observation variance is considered constant across the GP
  - Particularly a problem at the threshold
  - Fix by modelling observation variance as another GP
- Assumes that normal distribution approximates $\mathcal{D}(\boldsymbol{\theta})$
  - Log-normal might be more realistic?
- Jumps where there is threshold/bifurcation behaviour
  - Use Student $t-$Process to allow for jumps?

Thanks to

- Eamon Conway
- Jennifer Flegg
- August for explaining GPs