

# ODE parameter estimation in drug research: a deep dive into a Bayesian inference engine

Yi Zhang

<2023-05-26 Fri>

# Outline

The problem

You're Gonna Need a Bigger Boat

Up In the Air

Small moves, Ellie. Small moves

# Thanks



Statistics & Pharmacometrics

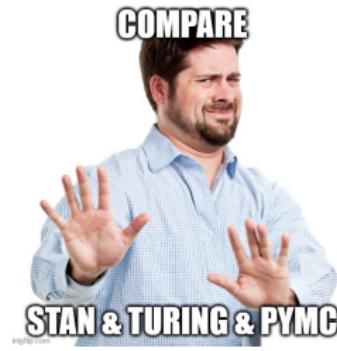
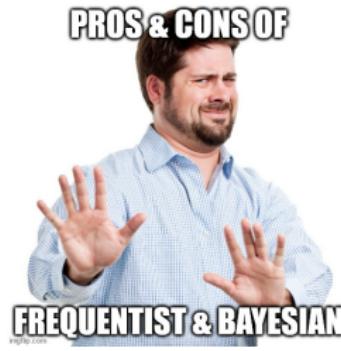


SPECIAL INTEREST GROUP



- ▶ Stan development team  
(Bob Carpenter, Daniel Lee, Sebastian Weber, ...)
- ▶ Bill Gillespie (Metrum Research Group)

# Disclaimer



Inference  
engine/Probabilistic  
programing  
language(PPL)

- ▶ Language design
- ▶ Inference algorithms
- ▶ Model checking & diagnostics
- ▶ Model assessment and comparison

# Table of contents

The problem

You're Gonna Need a Bigger Boat

Up In the Air

Small moves, Ellie. Small moves

# Pharmacokinetics (PK) data

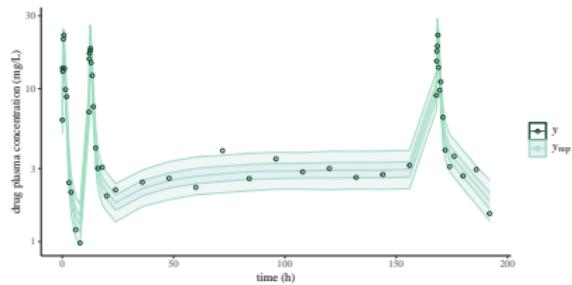
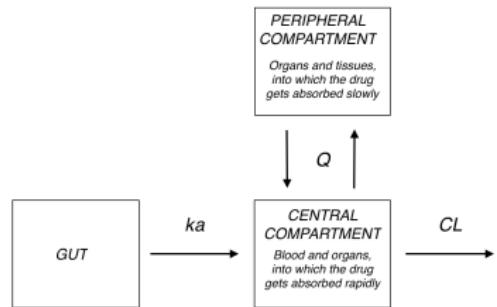


Figure: Subject plasma concentration history (q12hx14).



# A Pharmacokinetics (PK) model

$$\frac{d\hat{z}_{\text{gut}}}{dt} = -k_a \hat{z}_{\text{gut}}$$

$$\frac{d\hat{z}_{\text{cent}}}{dt} = k_a \hat{z}_{\text{gut}} - \left( \frac{\text{CL}}{V_{\text{cent}}} + \frac{Q}{V_{\text{cent}}} \right) \hat{z}_{\text{cent}} + \frac{Q}{V_{\text{peri}}} \hat{z}_{\text{peri}}$$

$$\frac{d\hat{z}_{\text{peri}}}{dt} = \frac{Q}{V_{\text{cent}}} \hat{z}_{\text{cent}} - \frac{Q}{V_{\text{peri}}} \hat{z}_{\text{peri}},$$

With  $\hat{y} = \frac{\hat{z}}{V_{\text{cent}}}$ :

$$\theta \equiv \{k_a, \text{CL}, Q, V_{\text{cent}}, V_{\text{peri}}, \dots\},$$

$$\hat{y}(t) = \hat{y}(t; \theta),$$

$$y(t) \sim \text{Normal}(\hat{y}(t; \theta), \sigma).$$

# A statistical model

- ▶ play mix & match with two models
- ▶ Likelihood  $p(y|\theta)$

$$p(y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(\hat{y} - y)^2}{2\sigma^2},$$

$$\log p(y|\theta) = -\frac{(\hat{y} - y)^2}{2\sigma^2} - \log 2\pi\sigma^2$$

- ▶ Usually  $\sigma$  is also unknown so likelihood  $p(y|\theta, \sigma)$

# Estimate $\theta$

- ▶ Maximum likelihood estimation  $\theta_{\text{argmax}}(p(y|\theta))$

## Estimate $\theta$

- ▶ Maximum likelihood estimation  $\theta_{\text{argmax}}(p(y|\theta))$
- ▶ Posterior distribution

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\bigcirc}$$

$$\log p(\theta|y) = C + \log p(y|\theta) + \log p(\theta)$$

## Estimate $\theta$

- ▶ Maximum likelihood estimation  $\theta_{\text{argmax}}(p(y|\theta))$
- ▶ Posterior distribution

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\bigcirc}$$

$$\log p(\theta|y) = C + \log p(y|\theta) + \log p(\theta)$$

- ▶ Connection to regularization:  $\log p(y|\theta) = -\frac{(\hat{y}-y)^2}{2\sigma^2}$

$$\min \|\hat{y}(\theta) - y\|^2 + \lambda \|\theta\|^2,$$

$$\log p(\theta|y) = -\frac{(\hat{y} - y)^2}{2\sigma^2} + \log p(\theta)$$

"Tikhonov regularization", "ridge regression", "shrinkage"

# Prior & likelihood

$$\log p(\theta|y) = C + \log p(y|\theta) + \log p(\theta)$$



## Sample the posterior

target (distribution) =  $\log p(\theta|y) = \log p(y|\theta) + \log p(\theta)$

- ▶ Posterior from an oracle:  $\theta \rightarrow \log p(\theta|y)$

## Sample the posterior

target (distribution) =  $\log p(\theta|y) = \log p(y|\theta) + \log p(\theta)$

- ▶ Posterior from an oracle:  $\theta \rightarrow \log p(\theta|y)$
- ▶ Every query costs one ODE solution

## Sample the posterior

target (distribution) =  $\log p(\theta|y) = \log p(y|\theta) + \log p(\theta)$

- ▶ Posterior from an oracle:  $\theta \rightarrow \log p(\theta|y)$
- ▶ Every query costs one ODE solution
- ▶ Sampling procedure for a distribution:  $\theta^{(1)}, \theta^{(2)}, \dots$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum f(\theta^{(i)}) = \mathbb{E}(f)$$

## Sample the posterior

target (distribution) =  $\log p(\theta|y) = \log p(y|\theta) + \log p(\theta)$

- ▶ Posterior from an oracle:  $\theta \rightarrow \log p(\theta|y)$
- ▶ Every query costs one ODE solution
- ▶ Sampling procedure for a distribution:  $\theta^{(1)}, \theta^{(2)}, \dots$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum f(\theta^{(i)}) = \mathbb{E}(f)$$

- ▶ MCMC:  $p(\theta^{(i)}|y, \theta^{(i-1)}, \theta^{(i-2)}, \dots, \theta^{(1)}) = p(\theta^{(i)}|y, \theta^{(i-1)})$

# Posterior samples: trace

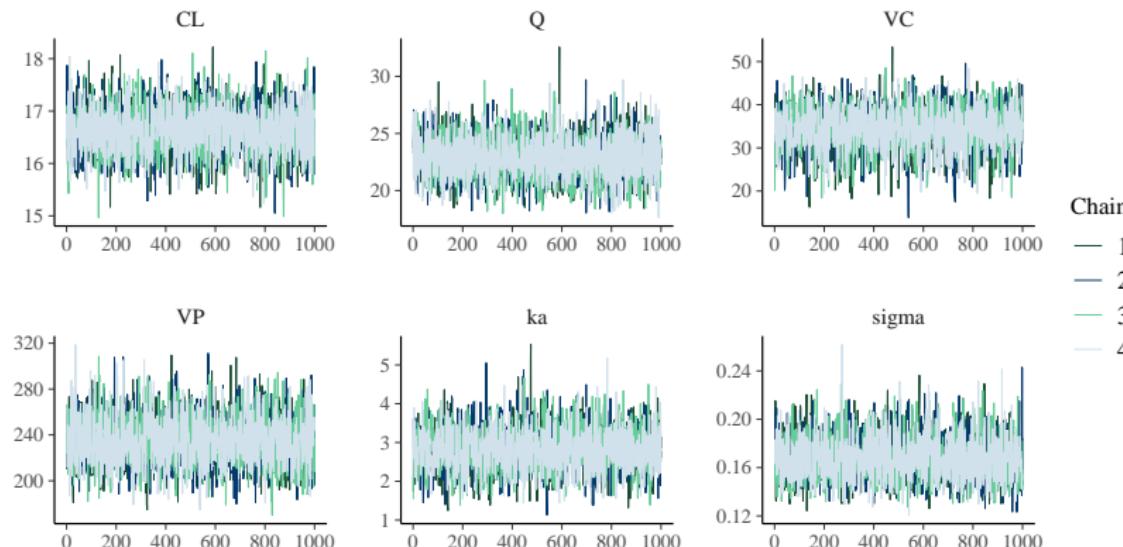


Figure: Trace of the MCMC samples (Margossian, Zhang, Gillespie 2022)

# Posterior samples: density

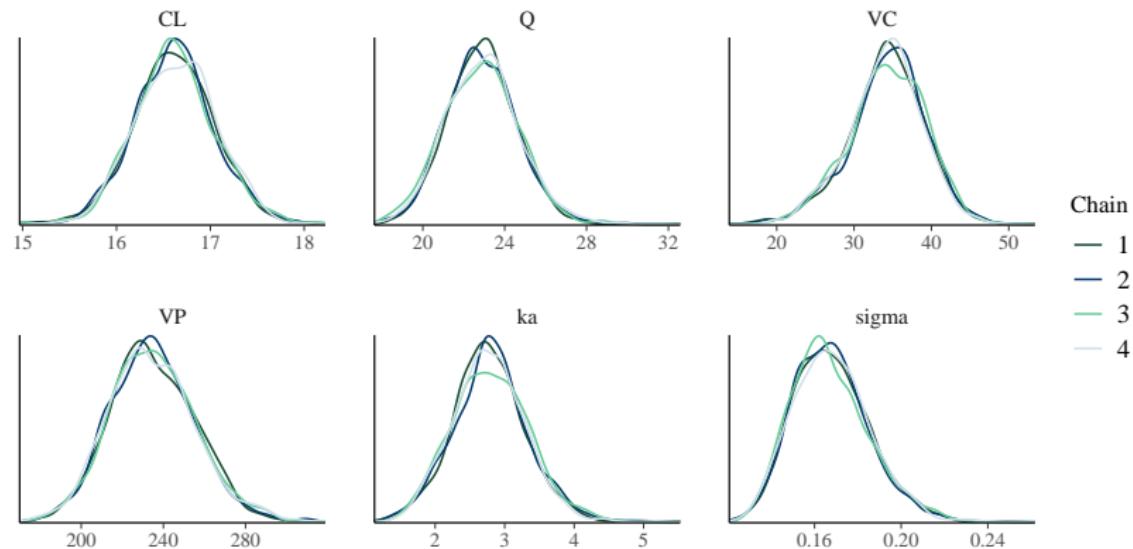


Figure: Estimated density (Margossian, Zhang, Gillespie 2022)

# The Metropolis-Hastings algorithm

Proposal-rejection from  $\theta^{(i)}$  to  $\theta^{(i+1)}$ .

1. Propose  $\theta^{(i+1)}$  according to transition density  $q(\theta^{(i+1)}, \theta^{(i)})$ .
2. Accept  $\theta^{(i+1)}$  with probability

$$\alpha(\theta^{(i)}, \theta^{(i+1)}) = \min \left[ 1, \frac{p(\theta^{(i+1)}|y)q(\theta^{(i)}, \theta^{(i+1)})}{p(\theta^{(i)}|y)q(\theta^{(i+1)}, \theta^{(i)})} \right] \quad (1)$$

otherwise reject.

- ▶ Chain generated by M-H has detailed balance with  $p(\theta|y)$  as its stationary distribution.
- ▶ Convergence in TVD with proper proposal that guarantees irreducibility and aperiodicity.

# Why MCMC?

*Monte Carlo is an extremely bad method; it should be used only when all alternative methods are worse.*

*– Sokal, A. D. (1989). "Monte carlo methods in statistical mechanics: foundations and new algorithms."*



# Table of contents

The problem

You're Gonna Need a Bigger Boat

Up In the Air

Small moves, Ellie. Small moves

# The problem

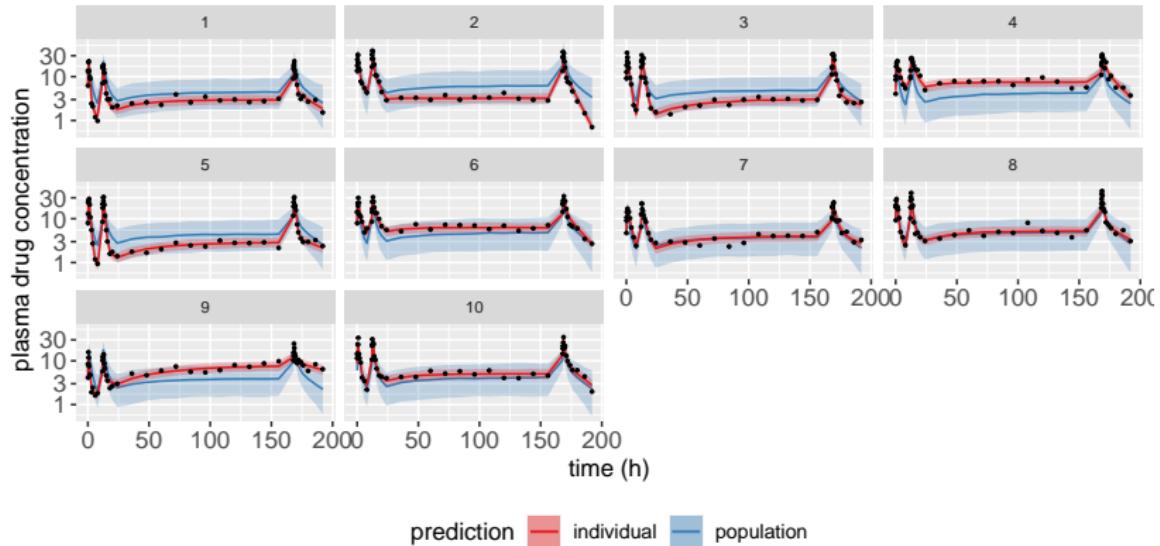


Figure: PK observation of ten subjects (Margossian, Zhang, Gillespie 2022)

# Hierarchical model

- ▶ Increased premium for posterior oracle

$$\theta_0 \sim \text{Prior}(\cdot),$$

$$\theta_i \sim \text{MultiNormal}(\theta_0, \Sigma),$$

$$y_i \sim \text{Normal}(\hat{y}_i(\theta_i), \sigma),$$

$$\theta = \{\theta_0, \theta_1, \dots, \theta_n, \Sigma\}$$

# Hierarchical model

$$\theta_0 \sim \text{Prior}(\cdot),$$

$$\theta_i \sim \text{MultiNormal}(\theta_0, \Sigma),$$

$$y_i \sim \text{Normal}(\hat{y}_i(\theta_i), \sigma),$$

$$\theta = \{\theta_0, \theta_1, \dots, \theta_n, \Sigma\}$$

- ▶ Increased premium for posterior oracle
- ▶ Curse of dimensionality
  - ▶ Computational tractability

# Hierarchical model

$$\theta_0 \sim \text{Prior}(\cdot),$$

$$\theta_i \sim \text{MultiNormal}(\theta_0, \Sigma),$$

$$y_i \sim \text{Normal}(\hat{y}_i(\theta_i), \sigma),$$

$$\theta = \{\theta_0, \theta_1, \dots, \theta_n, \Sigma\}$$

- ▶ Increased premium for posterior oracle
- ▶ Curse of dimensionality
  - ▶ Computational tractability
  - ▶ Concentration of measure  
(e.g. high dimensional gaussian distribution is like uniform distribution)

# Hierarchical model

$$\theta_0 \sim \text{Prior}(\cdot),$$

$$\theta_i \sim \text{MultiNormal}(\theta_0, \Sigma),$$

$$y_i \sim \text{Normal}(\hat{y}_i(\theta_i), \sigma),$$

$$\theta = \{\theta_0, \theta_1, \dots, \theta_n, \Sigma\}$$

- ▶ Increased premium for posterior oracle
- ▶ Curse of dimensionality
  - ▶ Computational tractability
  - ▶ Concentration of measure  
(e.g. high dimensional gaussian distribution is like uniform distribution)
- ▶ Geometry of posterior

# Challenges: high dimensional gaussian distribution

$$p(|\|y_d\|_2 - \sqrt{d}| \geq t) \leq 2 \exp(-ct^2), \forall t \geq 0 \text{ (Vershynin, 2018).}$$

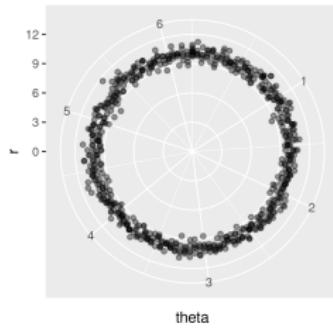
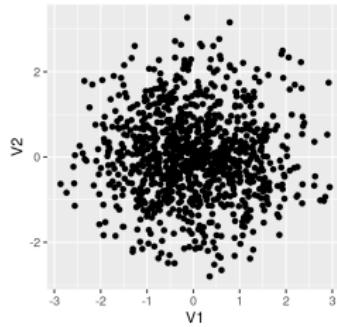
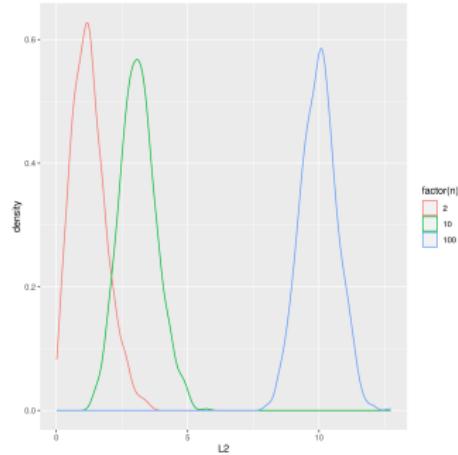


Figure:  $\|y_d\|_2, y_d \sim \text{Normal}(0, \mathbb{I}_d)$

Average is not representative.  
Random Walk sampler is not  
efficient.

# Challenges: Geometry of posterior

$\theta_0 = 0,$   
 $\kappa \sim \text{Normal}(0, 3),$   
 $\theta_i(k) \sim \text{Normal}(0, \exp(\kappa/2)),$   
 $k = 1, 2, \dots$

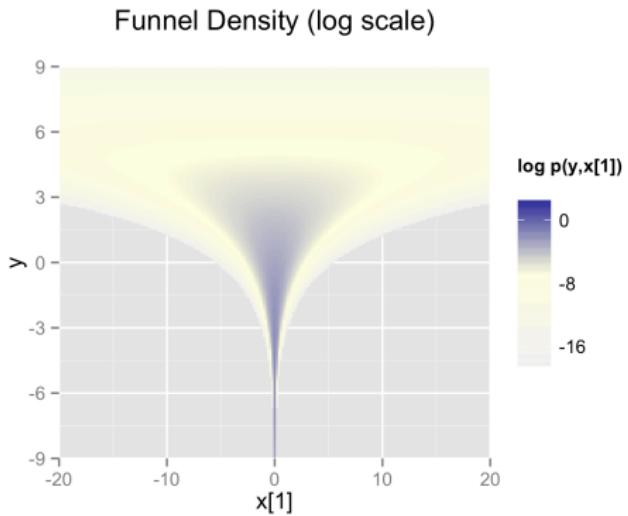


Figure: Neal's funnel (Stan dev team, 2023)

Mode is not representative. Optimizer is not efficient.

# Table of contents

The problem

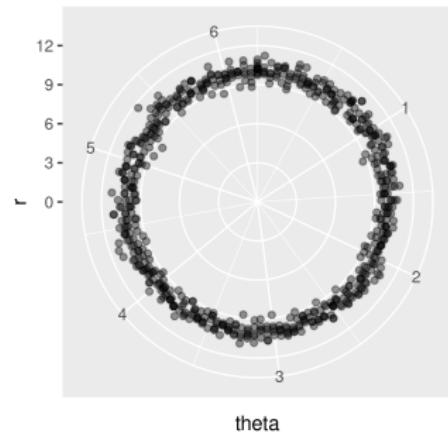
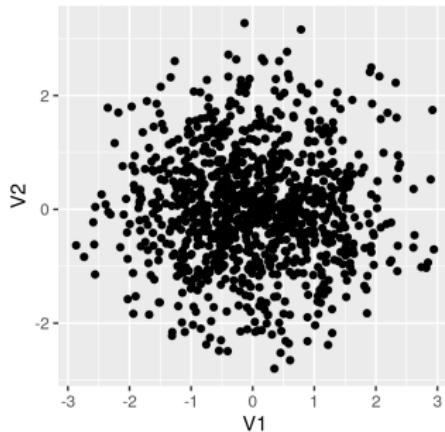
You're Gonna Need a Bigger Boat

Up In the Air

Small moves, Ellie. Small moves

# Hamiltonian Monte Carlo

$$H(\theta, r) = T(r) + V(\theta) = \frac{1}{2} r^T M^{-1} r - \log p(\theta|y),$$
$$\frac{d\theta}{dt} = \frac{\partial H}{\partial r}, \quad \frac{dr}{dt} = -\frac{\partial H}{\partial \theta},$$



# Hamiltonian Monte Carlo

$$H(\theta, r) = T(r) + V(\theta) = \frac{1}{2} r^T M^{-1} r - \log p(\theta|y),$$
$$\frac{d\theta}{dt} = \frac{\partial H}{\partial r}, \quad \frac{dr}{dt} = -\frac{\partial H}{\partial \theta},$$

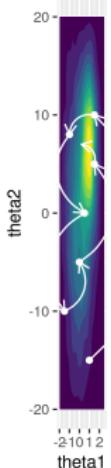
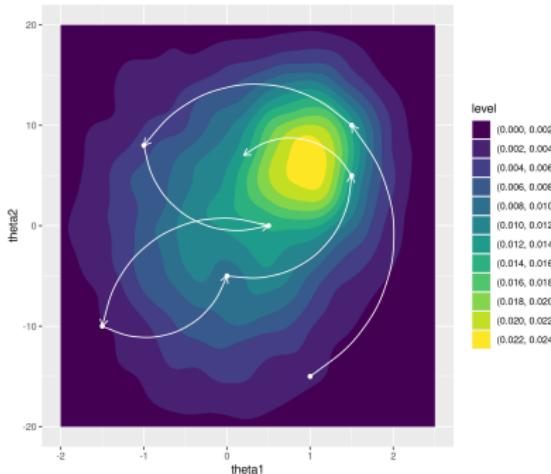
Apply M-H to  $p(r, \theta) \triangleq \frac{\exp[-H(\theta, r)]}{\bigcirc}$

$$\alpha((r^{(i)}, \theta^{(i)}), (r^{(i+1)}, \theta^{(i+1)})) = \min \left[ 1, \frac{p(r^{(i+1)}, \theta^{(i+1)}|y)q()}{p(r^{(i)}, \theta^{(i)}|y)q()} \right]$$
$$= \min \left[ 1, \frac{p(r^{(i+1)}, \theta^{(i+1)}|y)}{p(r^{(i)}, \theta^{(i)}|y)} \right]$$

$$\alpha(\cdot, \cdot) = \min [1, \exp(H(\theta^{(i)}, r^{(i)}) - H(\theta^{(i+1)}, r^{(i+1)}))]$$

# A principled sampler for $p(\theta|y)$

$$p(\theta^{(i)} \rightarrow \theta^{(i+1)}) = \min [1, \exp(H(\theta^{(i)}, r^{(i)}) - H(\theta^{(i+1)}, r^{(i+1)}))]$$
$$r \sim \text{MultiNormal}(0, M^{-1}) \implies (r^{(i)}, \theta^{(i)}) \rightarrow (r^{(i+1)}, \theta^{(i+1)})$$



# Table of contents

The problem

You're Gonna Need a Bigger Boat

Up In the Air

Small moves, Ellie. Small moves

# A tale of two ODEs

$$\theta^{(i)} = \{\theta^{(i)}_j\}, j=1,2,\dots,n: (r, \theta)(\tau^{(i)}) \rightarrow (r, \theta)(\tau^{(i+1)})$$

$$\begin{cases} \theta_j^{(i)} \rightarrow \hat{y}_j(t; \theta_j^{(i)}), \\ y_{jk} \sim \text{Normal}(\hat{y}_{jk}(\theta_j), \sigma), \quad p(y_{jk} | \theta_j) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(\hat{y}_{jk}(\theta_j) - y_{jk})^2}{2\sigma^2}\right] \end{cases}$$

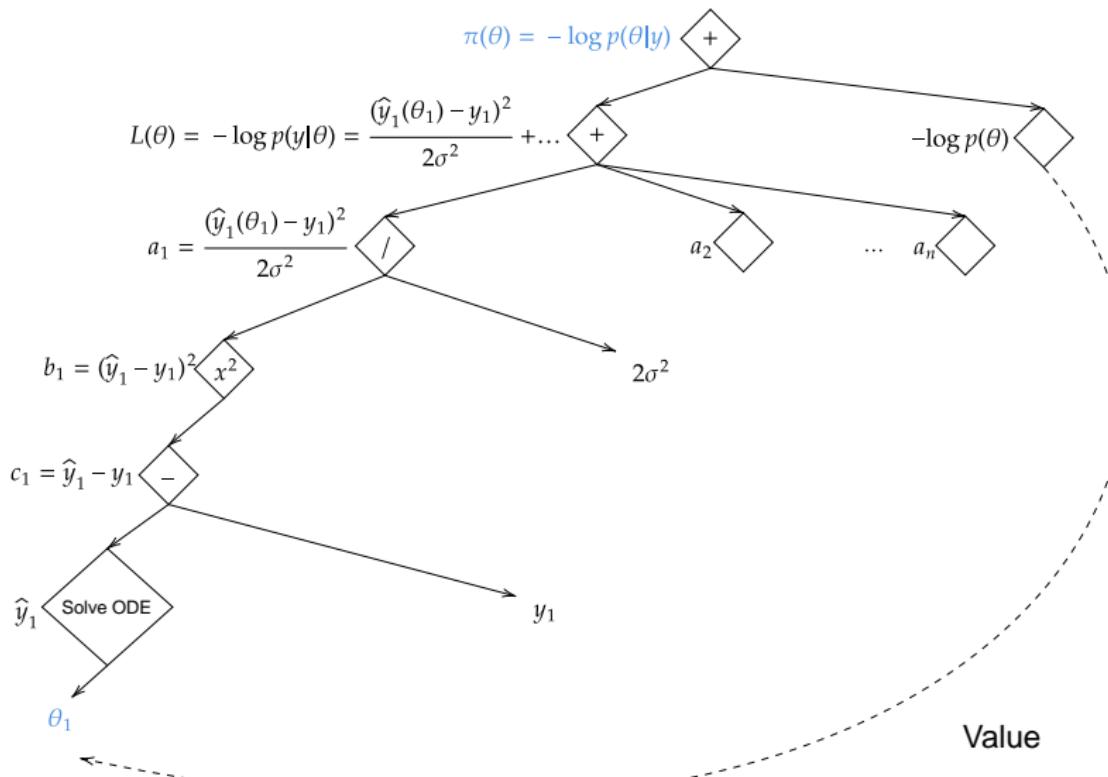
Nested ODE solvers:

$$r^{(i+1/2)} = r^{(i)} - \frac{h}{2} \nabla_\theta \log p(\theta^{(i)} | y), \text{ a step in leapfrog}$$

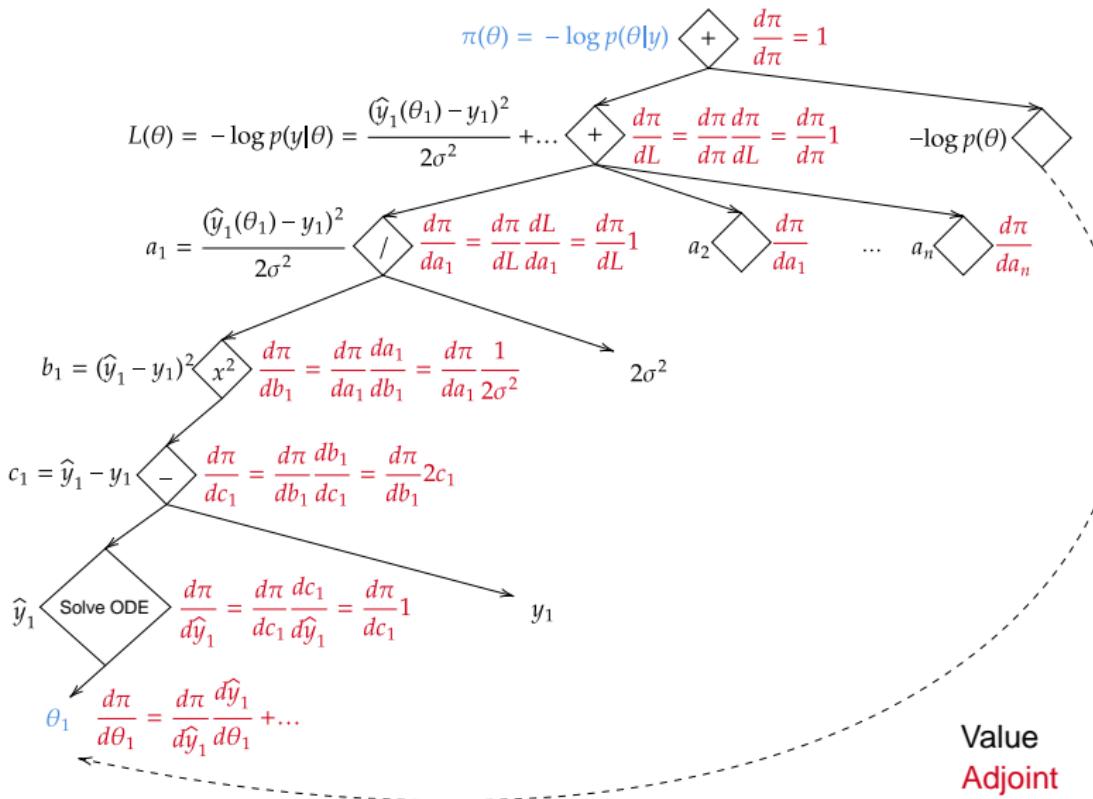
$$\nabla_\theta \log p(\theta^{(i)} | y) = \nabla_\theta \log p(\theta^{(i)}) + \nabla_\theta \log p(y | \theta^{(i)}),$$

$$\nabla_\theta \log p(y | \theta^{(i)}) = -(\dots) \sum_{j,k} \nabla_\theta \frac{\hat{y}_{jk}(\theta_j^{(i)}) - y_{jk}}{\sigma^{(i)}} + \dots$$

# Automatic differentiation (1 obsv/subject: k=1)



# Automatic differentiation (1 obsv/subject: k=1)



## Automatic differentiation

<i>function</i>	<i>arguments</i>	<i>return</i>
abs	(real x)	real
acos	(real x)	real
acosh	(real x)	real
asin	(real x)	real
asinh	(real x)	real
atan2	(real x, real y)	real
atan	(real x)	real
atanh	(real x)	real
cos	(real x)	real
cosh	(real x)	real
cbrt	(real x)	real
ceil	(real x)	real
erf	(real x)	real
erfc	(real x)	real
exp2	(real x)	real
exp	(real x)	real
fdim	(real x, real y)	real
floor	(real x)	real
fma	(real x, real y, real z)	real
fmax	(real x, real y)	real
fmin	(real x, real y)	real
fmod	(real x, real y)	real
hypot	(real x, real y)	real
log	(real x)	real
log10	(real x)	real
log1p	(real x)	real
log2	(real x)	real
pow	(real x, real y)	real
round	(real x)	real
sin	(real x)	real
sinh	(real x)	real
sqrt	(real x)	real
tan	(real x)	real
tanh	(real x)	real
trunc	(real x)	real



## **EXP GETS ADJOINT**



## **ERF GETS ADJOINT**



# SINH GETS ADJOINT



# **EVERYONE GETS AN ADJOINT**

# Sensitivity solution

$$\frac{d\hat{y}}{dt} = f(t, \hat{y}; \theta) \rightarrow \nabla_{\theta} \frac{d\hat{y}}{dt} = \nabla_{\theta} f(t, \hat{y}; \theta)$$
$$\frac{d\nabla_{\theta}\hat{y}}{dt} = f_{\theta} + f_{\hat{y}} \nabla_{\theta}\hat{y}$$

Use the autodiff calculate  $f_{\theta}$  and  $f_{\hat{y}}$ .

# Thank you