Fast Functional Integrals with Applications to Stochastic Dynamical Systems

John Tillinghast, Ph.D.
Adjunct Professor
Department of Statistics
George Washington University

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

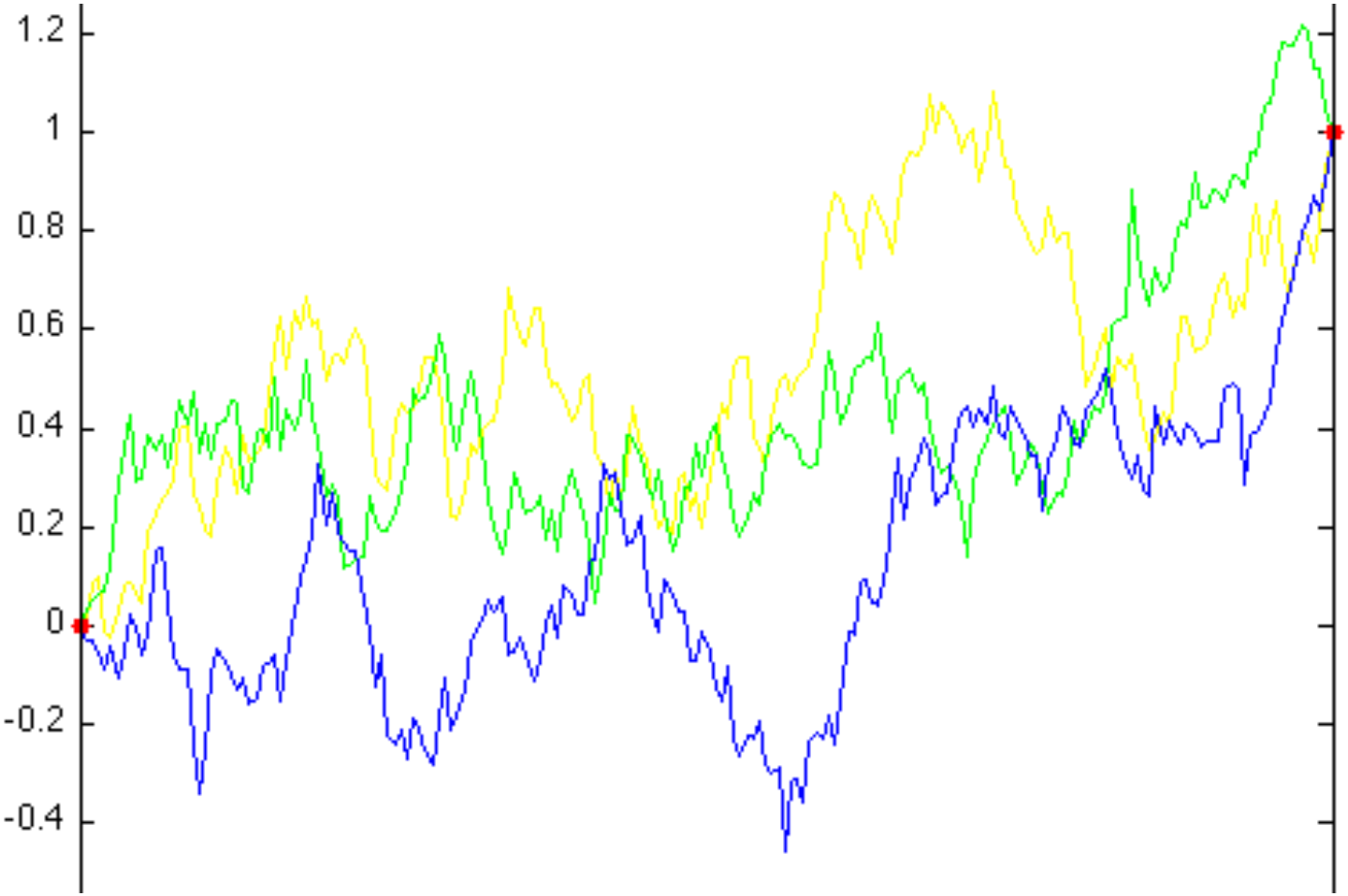
Overview

- Summary of functional integrals
- New application: parameter estimation for dynamical systems
- New method of calculating functional integrals
- Results for an infectious disease model (SIR) – equivalent results as Monte Carlo but much faster

Functional ("path") integrals

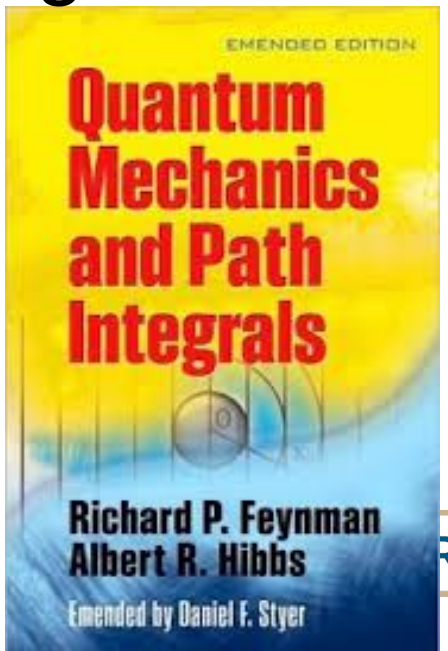
- Also called "sum over histories"
- Early application: Brownian motion (limit of random walks with infinitesimal steps)
- Gives likelihood of going from point A to point B in time t
  - Likelihood for individual paths is simple
  - Sum over all possible paths

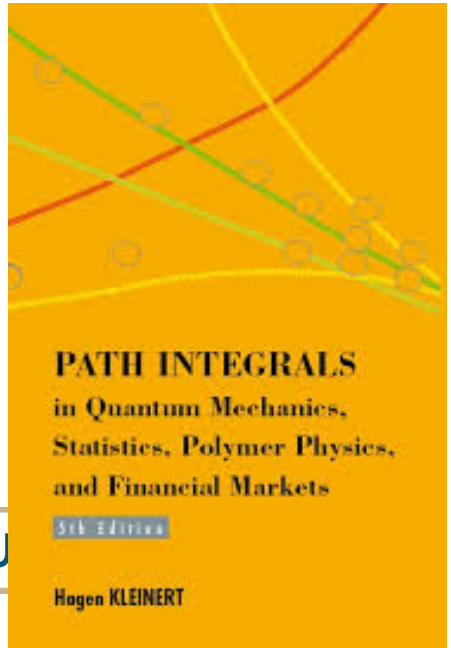
A few of the infinite number of histories to integrate



Many other applications

- Quantum and statistical mechanics
- Probability (SDEs, finance)
- Computationally intensive, can require long simulations





The new method: Sparse Laplace Approximation (SLA)

- Laplace approximation computes high-dimensional integrals by fitting a multivariate Gaussian to the integrand
- Basic version just integrates the Gaussian
- Higher-order version uses the Gaussian combined with higher-order derivatives
- **For functional integrals, the second- and higher-order derivative tensors are sparse (block-tridiagonal etc**.)

What's new here?

- Variance-stabilizing transform to get critical path
- Efficient way to get the higher-order terms: $O(Nd^4)$ instead of $O(N^4d^4)$

Application: Estimating Parameters in Dynamical Systems

- Dynamical systems are everywhere
- We want to estimate their parameters
- Traditional and recent methods
- How to use functional integrals for this
- Results and conclusions

Dynamical systems (differential equations models) are ubiquitous

- Life science, chemistry, economics
- These models incorporate:
  - *Variables*, which evolve over the time frame of the data
  - *Parameters*, which should be constant for the system – we want to estimate these.

Systems modeled by differential equations are ubiquitous

- Predator-prey models: numbers next year depend on numbers this year
  - Variables: numbers of predators and prey
  - Parameters: fertility and predation rate
- Chemical reactions
  - Variables are concentrations
  - Parameters are reaction rates

Systems modeled by differential equations are ubiquitous

- Infectious disease models
  - Variables: numbers of people who are susceptible, infected, recovered, etc.
  - Parameters: infectiousness, lethality, recovery rate, etc.

A simple example illustrates some of the challenges

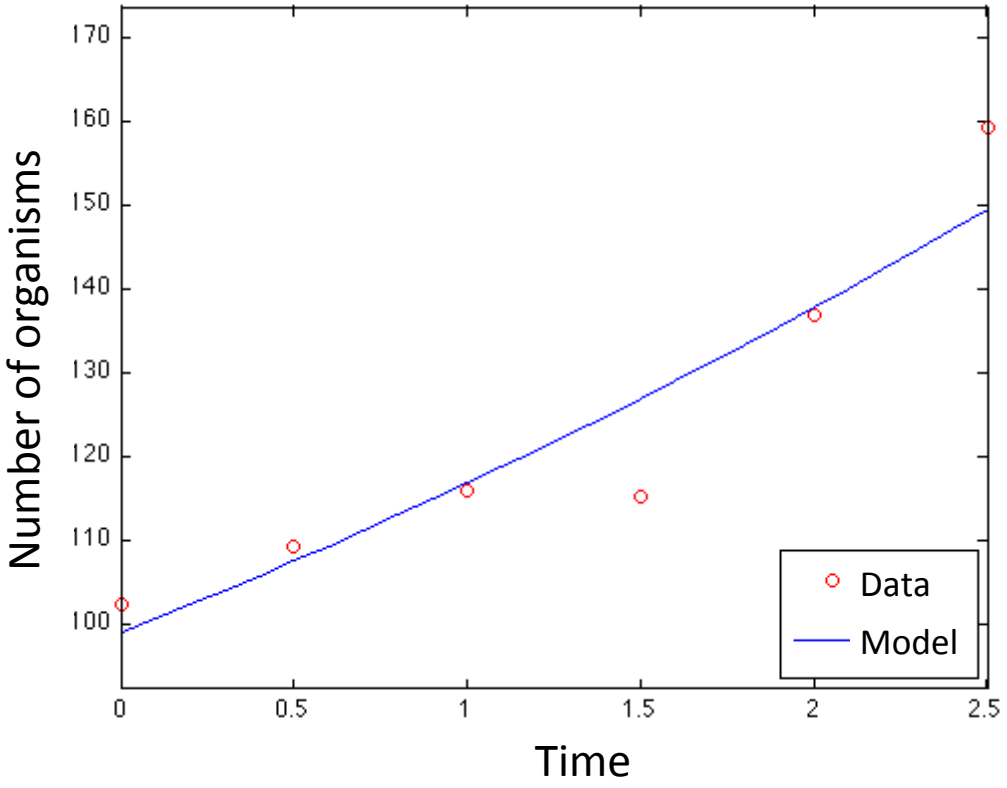
- In a bacterial colony a fixed fraction of cells divide every hour:

$$\frac{dy}{dt} = \theta y$$

where y is the number of bacteria, θ is the fraction that divide each hour, and t is time.

The traditional method solves the deterministic equations for different parameter values

- Parameter values are chosen to optimize fit to data



Traditional method has significant limitations

- It is not stochastic
- It does not allow for imperfectly followed differential equations
- In some systems, there are multiple optima of the parameters

More recent work

- Ramsay, Hooker, et al (JRSSB 2007) use spline functions and cross-validation
- Campbell (2010) uses splines and MCMC (similar to a functional integral but in a much smoother function space)

Spline-based approach

- Find a good compromise between data fit and ODE fit
- For the bacteria example, this means minimize

$$\sum_i \left(y^i_{data} - \hat{y}(t_i) \right)^2 + \lambda \int \left(\frac{d\hat{y}}{dt} - \theta \hat{y} \right)^2 dt$$

- Tradeoff (λ) is chosen by cross-validation or MCMC

Anatomy of the methods

- What all the approaches have in common:
- (1) for given params, find a "best guess" path $\hat{y}$ by minimizing "inner" objective function
- (2) given $\hat{y}$, calculate derivatives of "outer" objective function wrt parameters
- (3) repeat until outer function has been minimized

Functional Integral Approach

- Discretize time

$$N = \#\, time\ points,\ d = \#\, variables$$

- Treat true values as a latent, underlying Markov process
- This allows for rougher paths, more like Brownian motion

Overall likelihood as a functional integral

$$L(y_{data}|\theta,\sigma) = \int L_{data}(y_{data}|y_{true},\sigma)\, L_{Markov}(y_{true}|\theta)\, d^{Nd}y_{true}$$

where

y_{data} are the data values at whatever time points we have them

y_{true} is a possible history of the system

σ are parameters of measurement error

θ are dynamical parameters

Rewrite in terms of log-likelihood

- If we say

$$\ell_{data} = -\log(L_{data})$$
$$\ell_{Markov} = -\log(L_{Markov})$$
$$\ell = \ell_{data} + \ell_{Markov}$$

Then the second-order derivative matrix of ℓ is very sparse, and the higher-order derivative tensors are also sparse!

Example: bacteria colony

- L_{data} is the likelihood of the data given the true value

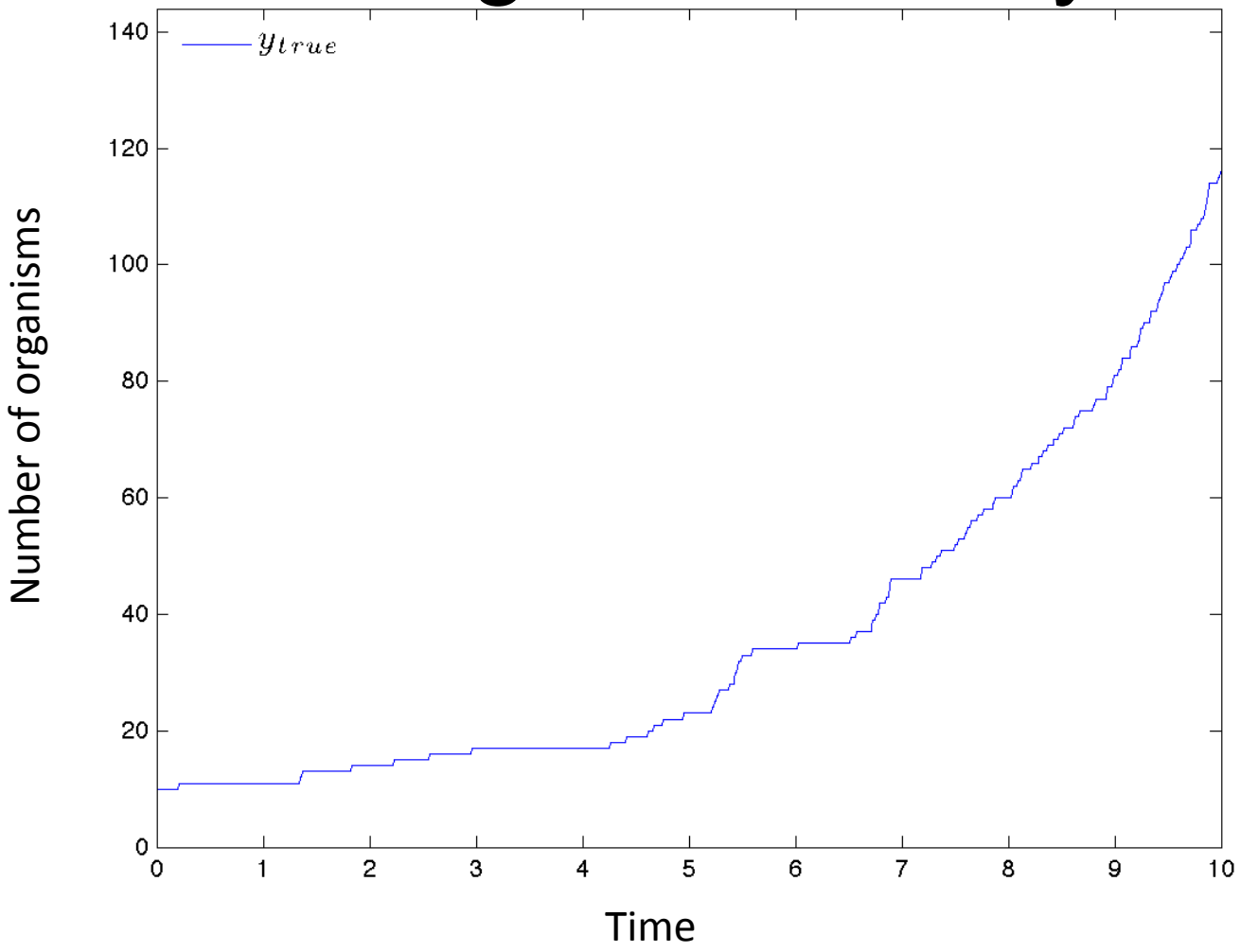
$$\log(y_{data}) \sim Normal\left(\log(y_{true}), \sigma^2\right)$$

- L_{Markov} is the likelihood of the sequence y^i_{true}

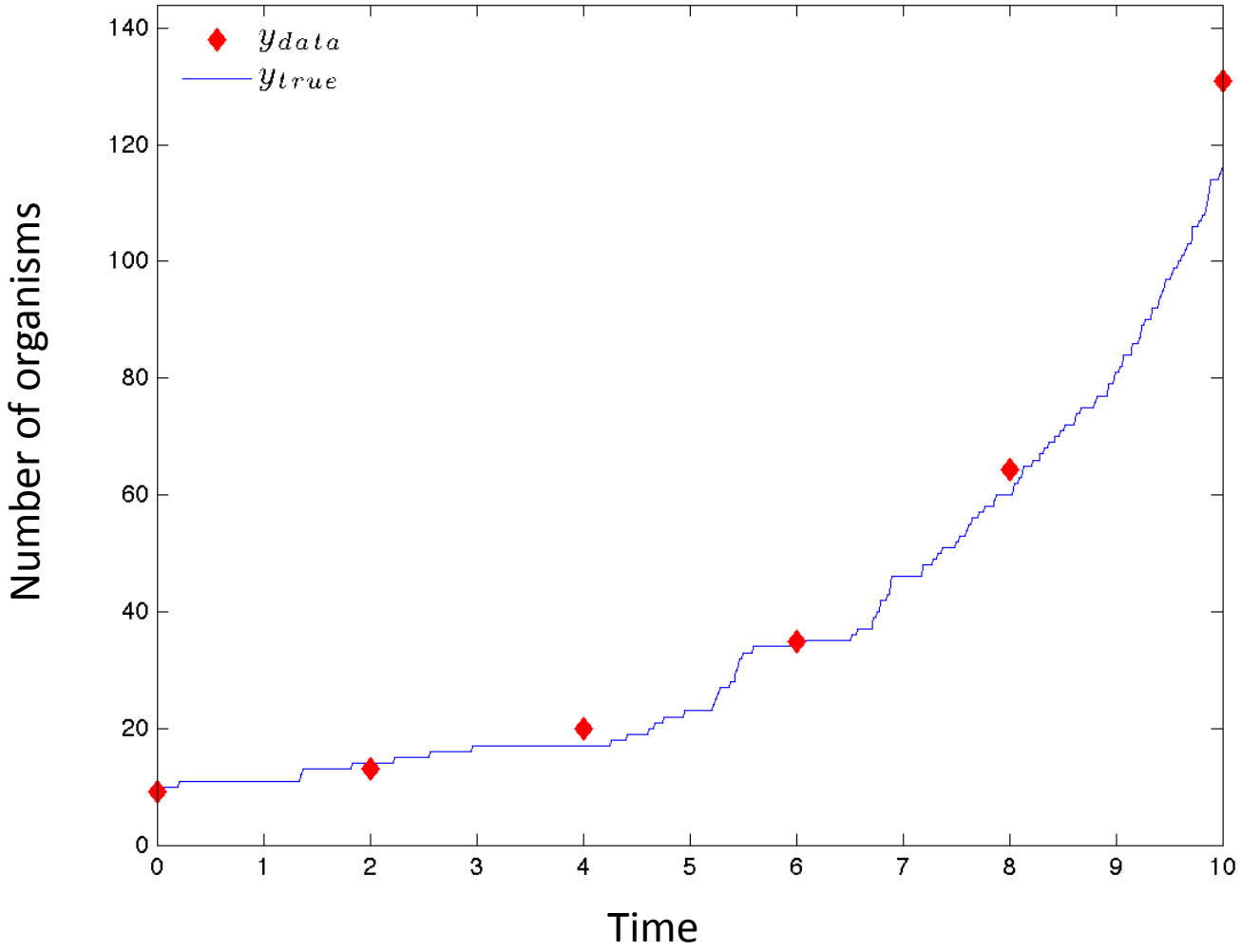
assuming that the real system follows the Markov process

$$\Delta_i y_{true} \sim Poiss(\theta \cdot y^i_{true} \cdot \Delta_i t)$$

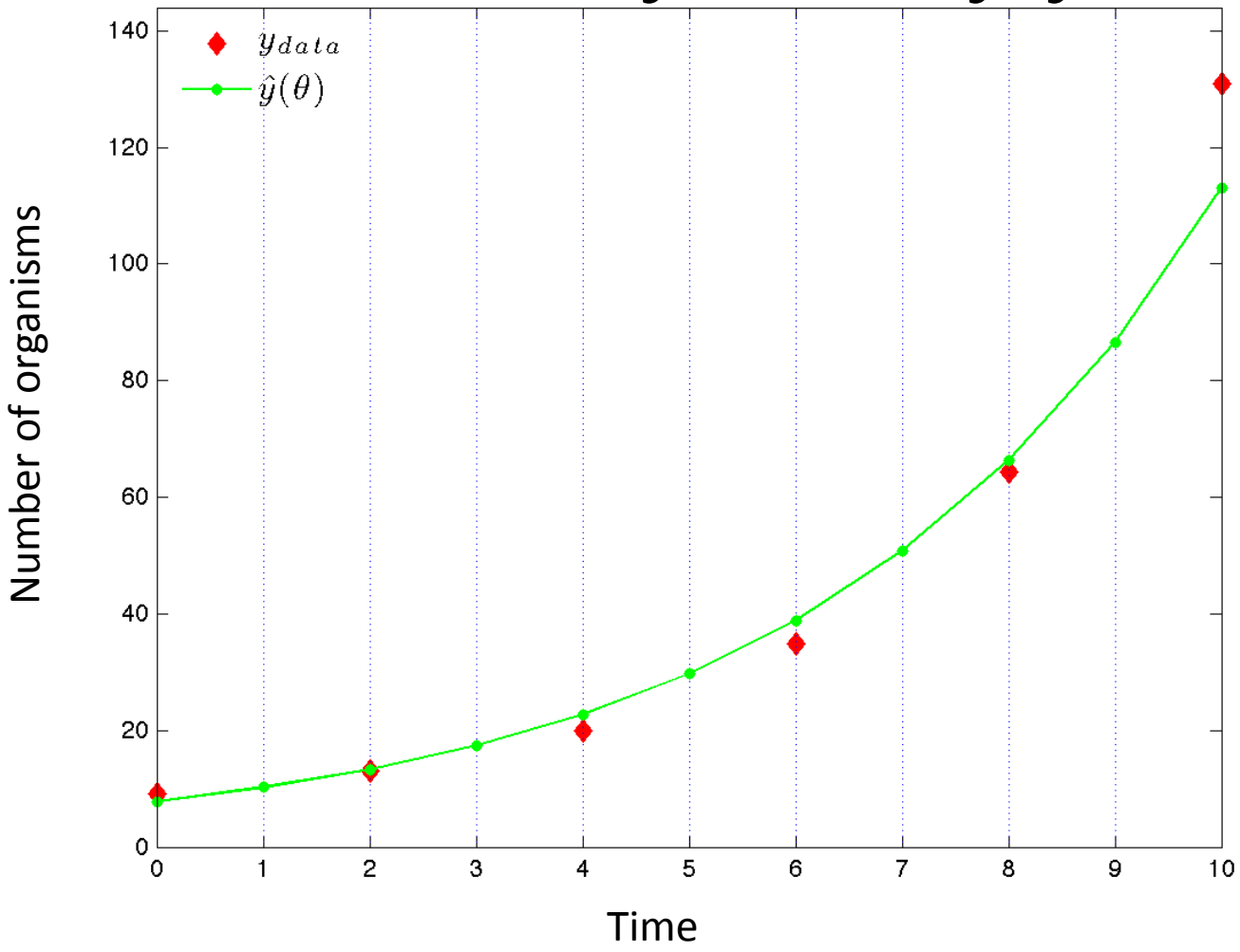
Bacterial colony example: "true" growth history



Data only at some time points, with measurement error



For given parameters, there is a most likely history ŷ



Expand around critical path

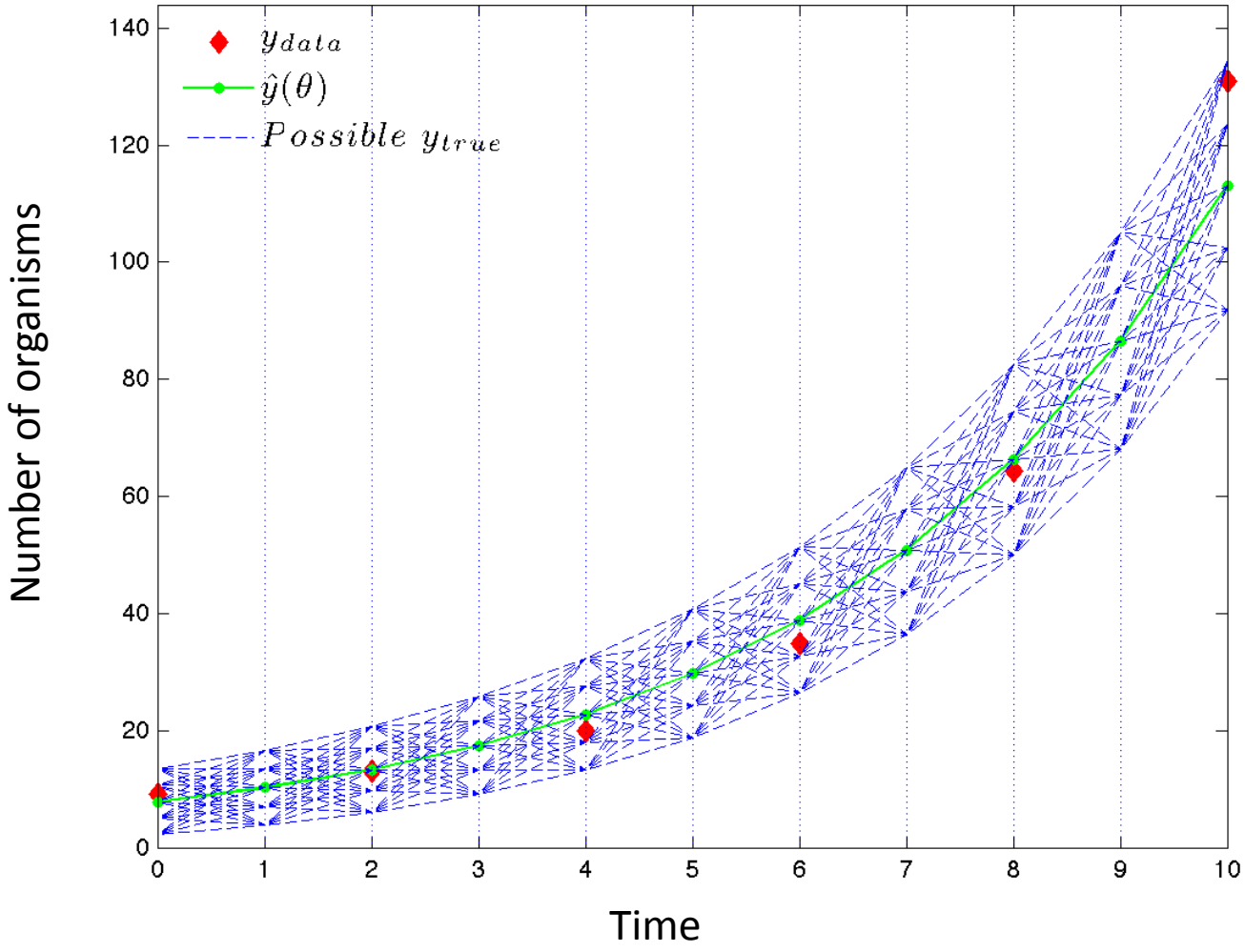
$$y = \hat{y} + \epsilon$$

$$\int_{\mathbb{R}^n} e^{-\ell(y)} d^n y = \int_{\mathbb{R}^n} e^{-\ell(\hat{y}+\epsilon)} d^n \epsilon$$

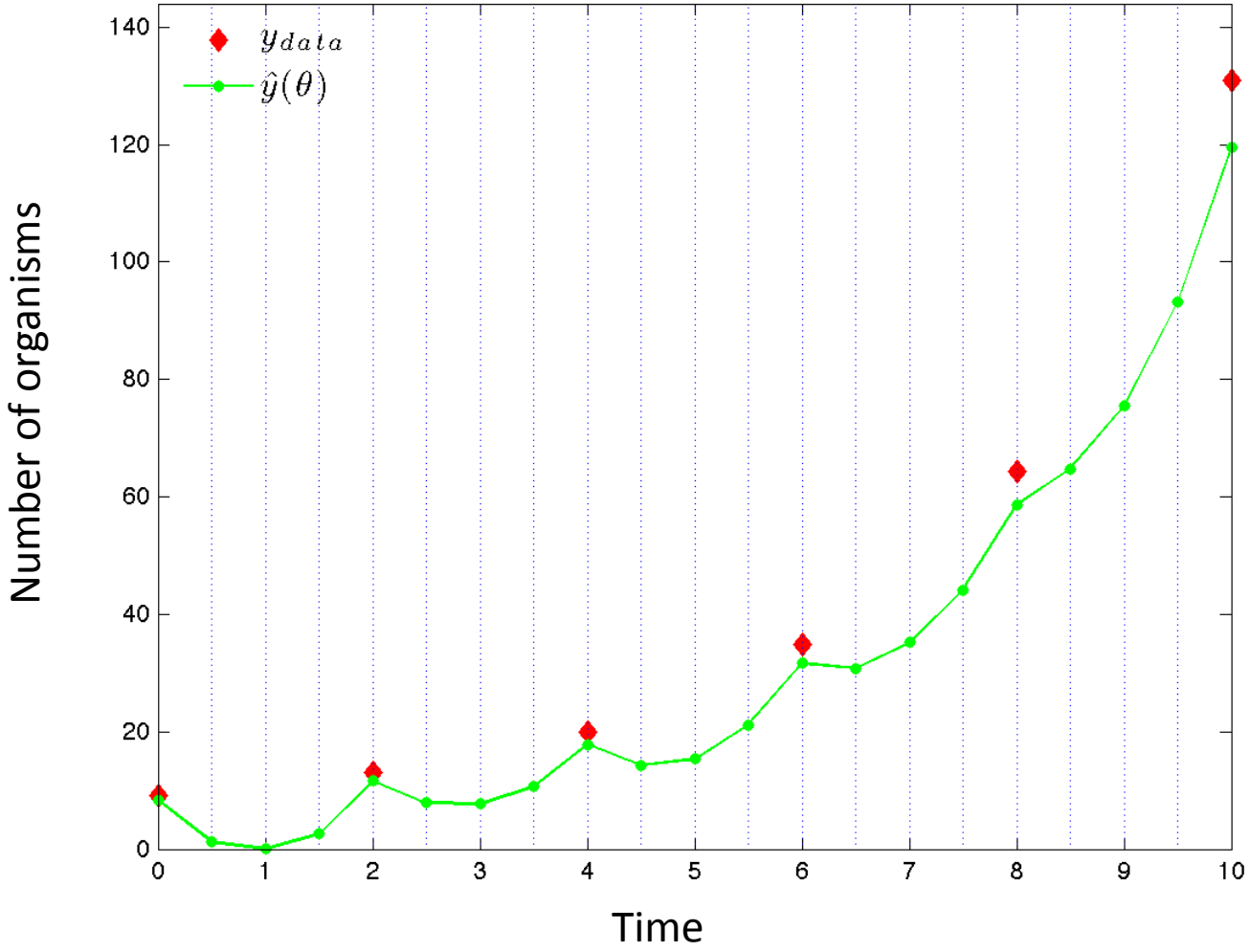
$$= \int_{\mathbb{R}^n} exp\left(-\ell\left(\hat{y}\right) - \frac{1}{2}\ell^{(2)}_{ij}\left(\hat{y}\right)\epsilon_i\epsilon_j \right.$$

$$\left. - \frac{1}{3!}\ell^{(3)}_{ijk}\left(\hat{y}\right)\epsilon_i\epsilon_j\epsilon_k - \ldots \right) d^n\epsilon$$

We use it to compute the integral over all paths



All very good, unless this happens…



What's going on?

In the normal approximation, the Markov variance is proportional to y:

$$L_{Markov} = \prod_{i<N} \frac{1}{\sqrt{2\pi\theta y_i \Delta_i t}} \exp\left(-\frac{1}{2} \frac{(\Delta_i y - \theta y_i \Delta_i t)^2}{\theta y_i \Delta_i t} \right)$$

The $\frac{1}{\sqrt{y_i}}$ means that moving y to zero increases the likelihood.

Why this is not okay

- It moves the critical path away from the realistic paths
- Likelihood *density* is maximized in a region whose *total likelihood* is small
- With the right change of variable, this problem disappears

Variance-stabilizing transform

- The solution is to make a change of variable that stabilizes the variance.

$$v_y = \sqrt{y}$$

$$L_{Markov} \prod_i dy_i \propto \prod_i \frac{e^{-q_i}}{\sqrt{y_i}} dy_i = 2 \prod_i e^{-q_i} \, dv_y = 2 \exp\left(-\sum_i q_i \right) dv_y$$

Reducing $\sum_i q_i$ does not bias y toward zero.

Cumulant form of Laplace approximation (Shun & McCullagh, 1995)

$$\log \int e^{-\ell(y)} d^N y \sim -\ell(\hat{y}) + \frac{1}{2}\log\frac{(2\pi)^N}{|H|}$$
$$- \frac{3}{24} F_{ijkl} H^{-1}_{ij} H^{-1}_{kl}$$
$$+ \frac{9}{72} T_{ijk} T_{lmn} H^{-1}_{ij} H^{-1}_{kl} H^{-1}_{mn}$$
$$+ \frac{6}{72} T_{ijk} T_{lmn} H^{-1}_{il} H^{-1}_{jm} H^{-1}_{kn}$$
$$+ \ldots$$

Basic approximation is the first line

Other terms are higher-order

Watch your accounting: higher-order terms

- There are three distinct higher-order terms
- For the first two, it is “obvious” how to use sparsity and compute efficiently
- The last one involves a complicated analytical trick
- Total time is $O(N\,d^4)$ (N = time points, d = variables)
- Without sparsity, the total time would have been $O(N^4\,d^4)$

Testing on a simple infectious disease model

- SIR Model (Kermack and McKendrick, 1927)
  - **S**usceptible
  - **I**nfected (and contagious)
  - **R**ecovered (and immune)

Usually this is solved deterministically

$$\frac{dS}{dt} = -\beta SI$$

$$\frac{dI}{dt} = \beta SI - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

SLA permits a stochastic approach

$$\text{New Infections } (t_i \text{ to } t_{i+1}) \sim Poiss(\beta S^i_{true} I^i_{true} \Delta_i t)$$

$$\text{Recoveries } (t_i \text{ to } t_{i+1}) \sim Poiss(\gamma I^i_{true} \Delta_i t)$$

Normal approximation to the Poisson distribution: write as SDE

$$dS = -\beta SIdt + \sqrt{\beta SI}\,dW_1$$

$$dI = -dS - \gamma Idt + \sqrt{\gamma I}\,dW_2$$

Variance-Stabilizing SIR

Bias is removed by this transformation:

$$v_S = \sqrt{S}$$

$$v_I = \log(I)$$

$$L_{Markov} dS_i dI_i \propto \frac{dS_i dI_i}{\sqrt{S_i \cdot I_i}} = 2 dv_S dv_I$$

Comparison with STAN on a real data set

- STAN is a new (2012) general-purpose system for Hamiltonian Monte Carlo.
- You specify the model; then it generates C++ code and compiles with optimization.
- Developed by Andrew Gelman's group at Columbia; replaces BUGS
- Try it! http://mc-stan.org

Data set

- British boarding school data set: 14-day flu epidemic
- Data is only on number infected (number susceptible is latent!)
- Still possible to implement both in STAN and SLA

Results (parameter estimates)

		β	γ	σ	Time (sec)	Opt steps
Initial guess		2.18E-03	0.440	(0.1)*	N/A	N/A
STAN	5%	1.31E-03	0.667	0.122	267	
	Mean	1.53E-03	0.760	0.152	325	N/A
	95%	1.77E-03	0.882	0.156	366	
SLA basic		1.45E-03	0.777	0.156	6	23
SLA higher		1.47E-03	0.770	0.157	42	4

*Paper doesn't specify initial sigma, so we chose 0.1. STAN runs use 4 chains each with 2000 steps. There were 100 such runs.

Results (marginal likelihood)

	Log(L)	% diff from IS estimate
SLA basic	5.353	5.7%
SLA higher	5.300	0.4%
Importance Sampling (N=6.6e8)	LB: 5.295 Est: 5.296 UB: 5.298 (99+% CI)	-0.1% 0.0% 0.2%

Day 0 removed to avoid a pathology that happens when I=1.
Importance sampling uses multivariate T, df=2.
Sample values constrained to sqrt(S) and log(I) positive.

Conclusion

- For this problem, SLA gives near-identical results to STAN, but is vastly faster
  - Basic is > 50 times faster
  - Basic + Higher is ~6 times faster
  - Real difference should be bigger, because STAN is compiled with high optimization

More to be done: methods

- Further efficiencies
- More independent variables (not just time) requires different algorithm
- Analytical error bounds
  - Notoriously hard for Laplace approximations, but methods exist
  - The usual, hand-waving explanation is not the real reason why they are so accurate (Olver, 1968)

More to be done: other applications

- Larger problems
- Bayesian model selection
- Graphical models
- Physics
  - "Likelihood" becomes complex (not that bad)
  - In some problems the path would be in a group, not $\mathbf{R}^n$

Acknowledgements

- Giles Hooker (Cornell) for piquing my interest and helpful discussion
- Karen Bandeen-Roche and Ravi Varadhan (Hopkins) for encouragement
- Tammy Kolda (Sandia) for help with the MATLB Tensor Toolbox
- Neil Abernethy (UW SPH) for suggesting the data set
- Robert Rudnitsky (NIST) and Kelley Burridge (FDA) for editing ideas

Thank you!

Work was begun at the Hopkins School of Public Health under Training Grant T32 AG000247 (epidemiology and biostatistics of aging)

References

- Spline methods: Ramsay J., Hooker G., Campbell, D. and Cao, J. (2007) Parameter estimation in differential equations: A generalized smoothing approach. J. R. Statist. Soc. B 69, Part 5, pp. 741-96.
- Spline MCMC: Campbell, D. and Steele, R. (2012) Smooth Functional Tempering for Nonlinear Differential Equation Models. Statistics and Computing, 22(2), pp 429-43.
- STAN (and its papers) are available at http://mc-stan.org

More references

- Laplace approximation in cumulant form: Shun, Z. and McCullagh, P. (1995) Laplace approximation of high-dimensional integrals, J. Royal Stat. Soc. B, 53, pp. 749-60.
- SIR Model: Kermack, W. O. and McKendrick, A. G. (1927), “A Contribution to the Mathematical Theory of Epidemics”. Proc. Roy. Soc. Lond. A 115, pp. 700-721.
- British boarding school data: J.D. Murray (2002), *Mathematical Biology I, An Introduction*, pp. 325-326.
- Error bounds: Olver (1968). “Error Bounds for the Laplace Approximation for Definite Integrals”, J. Approximation Theory v.1, pp. 293-313.