Novel Data Assimilation Methods for Highly Nonlinear Problems

Michael Turmon

Machine Learning Group
Caltech / Jet Propulsion Laboratory
Pasadena, CA, USA

September 2004

- Introduction and Motivation
- Monte Carlo Sampling
- Monte Carlo Inference
- Particle Filter Inference

turmon@aig.jpl.nasa.gov
http://www-aig.jpl.nasa.gov/home/turmon/

Jet Propulsion Laboratory
California Institute of Technology

Nonlinear Time Series

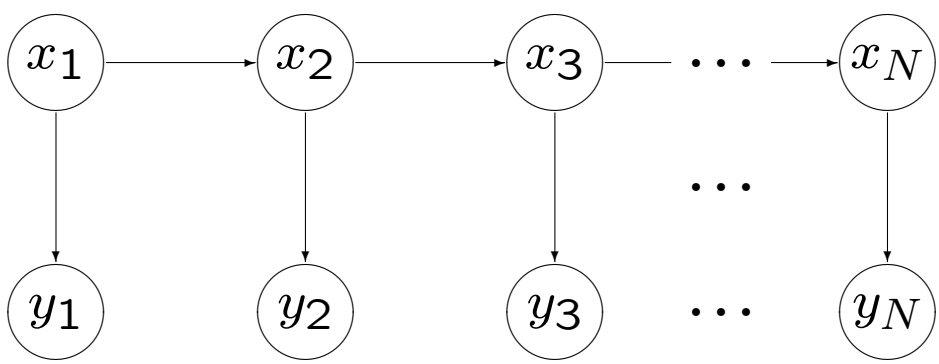
Familiar nonlinear state-space model

$$
\begin{aligned}
x_{n+1} &= g(x_n) + r_n \qquad & r_n &\sim N(0, \kappa^2) \\
y_n &= C\, x_n \;\; + s_n \qquad & s_n &\sim N(0, \sigma^2)
\end{aligned}
$$

for vector x and y

There is a corresponding continuous-time diffusion

The Bayes net shows dependences among this collection of variables



- **Fundamental Inference Problem**

Given y_n, for various $n \in \{1, ..., N\}$ (with gaps), estimate x_n, $1 \leq n \leq N$.

Jet Propulsion Laboratory
California Institute of Technology

Learning Problems

Inference: Given y, estimate x

Large-scale (e.g., weather) applications solve tractable approximations of this problem

We focus on the other end of the spectrum, computationally expensive procedures that:

- Account for nonlinear dynamics $g(\cdot)$
- Provide error bars, up to and including estimates of the probability density of x_n
- Explain how to properly weight the members of an ensemble according to cumulative credibility

- **Related Problems**

Smoothing, filtering are specific versions of inference as we have defined it

Sampling (e.g. of x given y) can be a building block for inference

Parameter learning: estimate $g(\cdot)$, κ, σ, or other parameters

Structure learning: determine network structure (e.g., identify covariates influencing x)

Jet Propulsion Laboratory
California Institute of Technology

Motivating Applications

- **General nonlinear systems**

  These systems exhibit threshold effects, shocks, fronts

  Nongaussian posterior distributions
  Skew distributions or heavy tails
  Multimodal posterior: > 1 distinct outcome

- **Dynamics of Lagrangian features**

  The internal state of fronts can be viewed as satisfying a state equation for a (learned) dynamic $g(\cdot)$

  These Lagrangian features are low-dimensional and allow computationally expensive modeling

  These two broad areas overlap

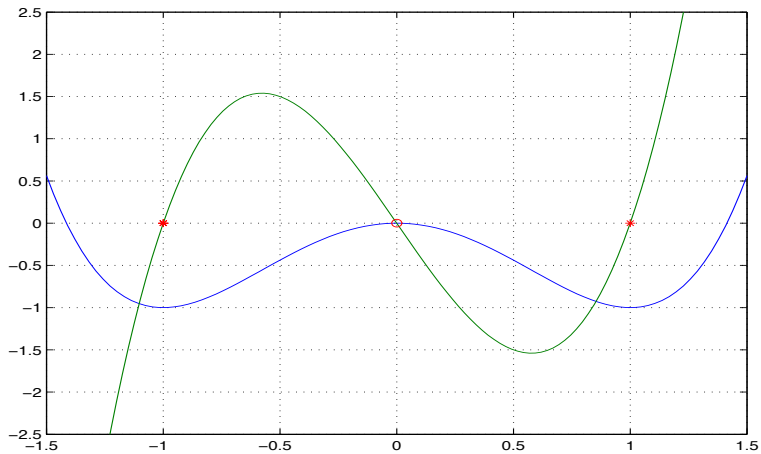
  E.g., nonlinearities couple fourier modes, causing coherent fronts to build rather than dissipate

Jet Propulsion Laboratory
California Institute of Technology

Double-Well Model

Particle in a potential well given by
$f(x) = -2x^2 + x^4$ and $g = f'$

Two minima at ± 1 and one stationary point at 0



Trivial climate
cf. Miller et al., Eyinck & Restrepo, etc.

Corresponds to the continuous-time diffusion

$$dx_t = -g(x_t) + \kappa\, dB_t \qquad \kappa \text{ constant}$$
$$y_t = x_t + s_t \qquad s_t \sim N(0, \sigma^2)$$

Approximate by δ-width finite differences

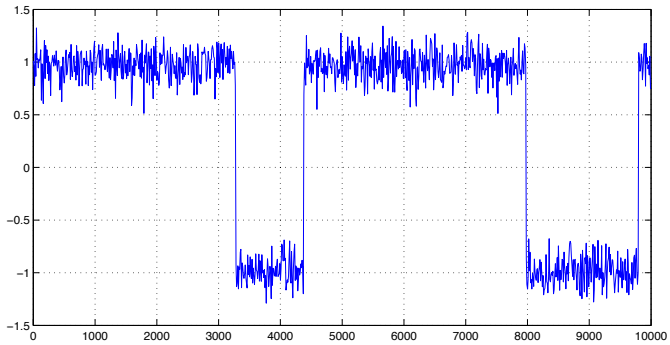
$$x_{n+1} = x_n - \delta g(x_n) + r_n \qquad r_n \sim N(0, \delta\kappa^2)$$
$$y_n = x_n + s_n \qquad s_n \sim N(0, \sigma^2)$$

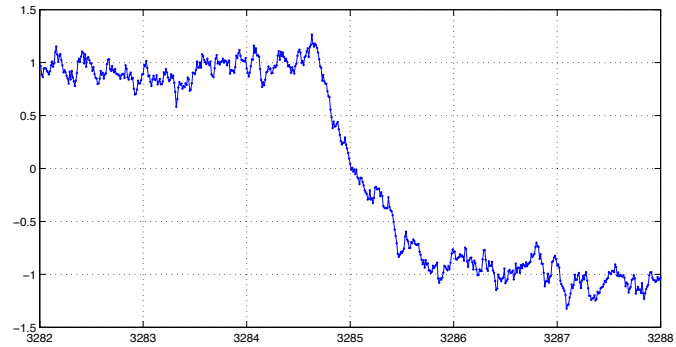
Need $\delta|g(\cdot)|, \sqrt{\delta}\kappa \ll 1$

Jet Propulsion Laboratory
California Institute of Technology

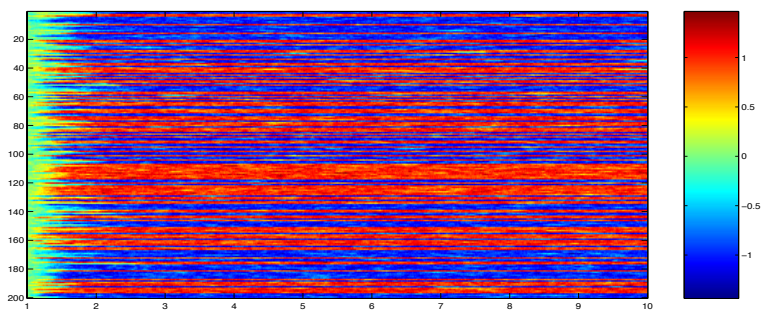
Sample Paths

$\kappa^2 = 0.24$, $ET_{\pm 1} \simeq 4600$ $(\delta = .01,\ x_1 = 1,\ T = 10^4)$





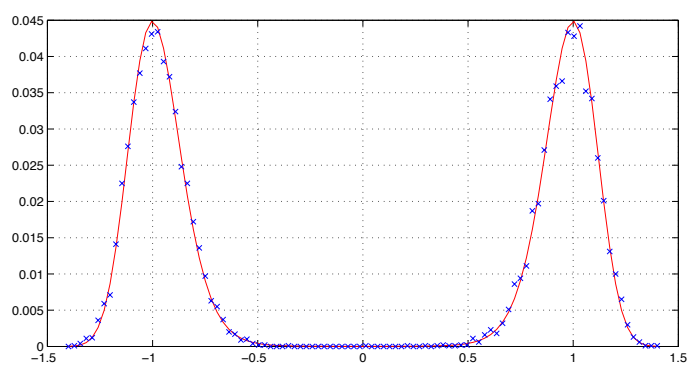
Stacked ensemble of 200, $x_1 = 0$ $(\delta = .01,\ T = 10)$



Limiting density $p_\infty(x) \propto \exp(-2f(x)/\kappa^2)$

Histogram:

Ensemble of 10000, $x_1 = 0$, $T = 10^3$



Can we compute more than just limit densities using Monte Carlo?

Jet Propulsion Laboratory
California Institute of Technology

MCMC: Essence

- **Monte Carlo integration...**

$$Ef(X) = \int f(x)\,\pi(x)\,dx \simeq \frac{1}{N}\sum_{t=1}^{N} f(x_t) \qquad (x_t \text{ from } \pi)$$

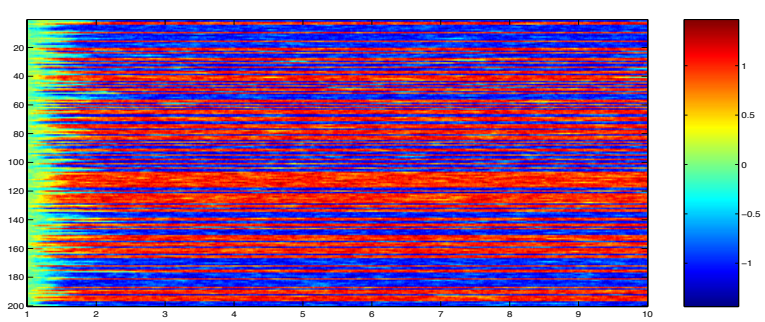
Typically x_t are sampled iid: but this can be hard

- **...driven by Markov chain samples**

Craft a MC having $\pi(\cdot)$ as its stationary dist'n
  Draw the x_t's above from this MC
  They are not independent, but they are from π

The long-term sample averages will converge,
  same as the ensemble averages do

We used this trick to find histograms:
  Ensemble averaging goes vertically
  MCMC sampling goes horizontally



MCMC: Metropolis-Hastings

One particular kind of MC uses the...

- **Metropolis-Hastings recipe**

  Say the MC is now at x (time t)
  Draw a candidate x' from $q(x' \,|\, x)$
  Replace x by the new x' with probability

$$\rho = \min\Big(1, \frac{\pi(x')\, q(x \,|\, x')}{\pi(x)\, q(x' \,|\, x)}\Big)$$

  Under broad conditions, sample mean $\to Ef(X)$

  But: rate of convergence is critical, and hard to assess

  Can use un-normalized versions of $\pi(\cdot)$

- **Acceptance probability ρ**

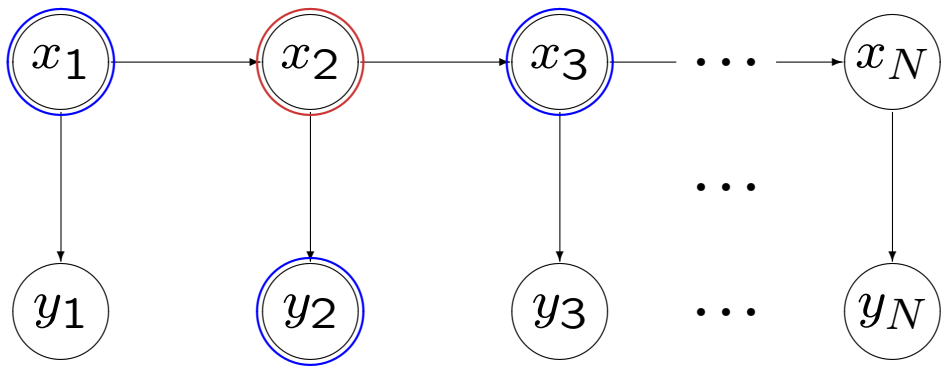
  Favors increasing the posterior, but allows $\pi(x)$ to decrease occasionally

  Adjusts for any bias introduced by our choice of q

Jet Propulsion Laboratory
California Institute of Technology

MCMC: Time Series

Recall the Bayes network showing dependences



For now, we only want to estimate x_2 and other values are known

Apply Metropolis-Hastings recipe:

Propose a new value x_2', e.g. from $N(x_2, \delta\kappa^2)$

($q(x_2' \,|\, x_2)$ is symmetric)

Compute the ratio

$$\frac{\pi(x')\, q(x \,|\, x')}{\pi(x)\, q(x' \,|\, x)} = \frac{p(x_2' \,|\, x_1)\, p(x_3 \,|\, x_2')}{p(x_2 \,|\, x_1)\, p(x_3 \,|\, x_2)} \times \frac{p(y_2 \,|\, x_2')}{p(y_2 \,|\, x_2)}$$

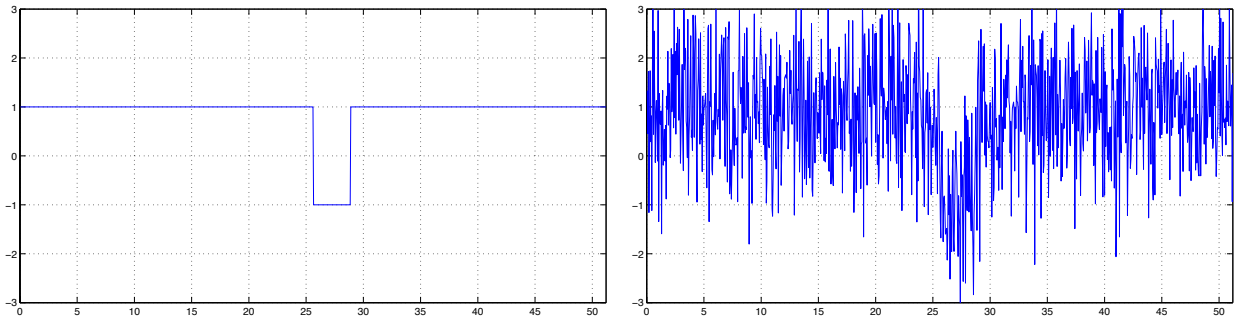
Combines a smoothness term and a data term

The full MCMC scheme sweeps over all unknown variables $x_1, ..., x_N$, proposing changes to each

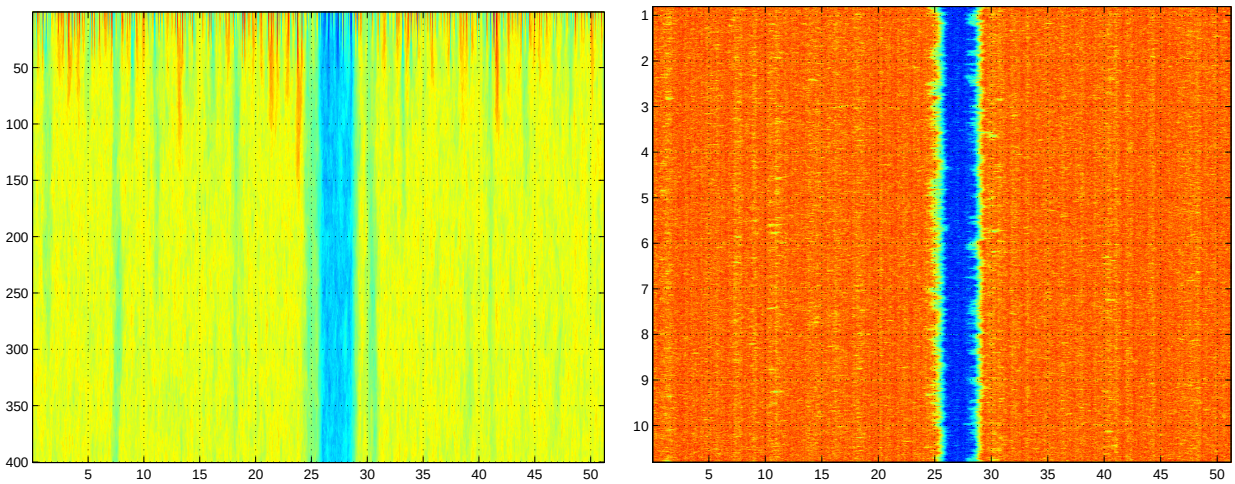
Jet Propulsion Laboratory
California Institute of Technology

MCMC Results

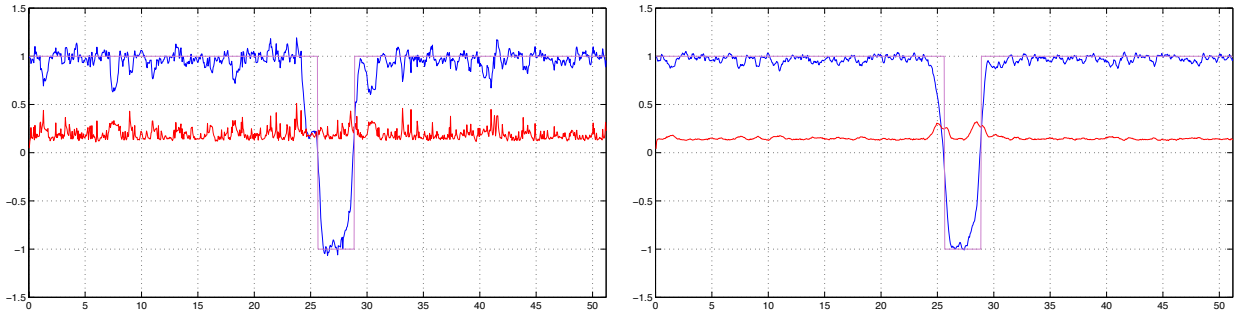
True x_n with data ($\kappa = 0.5$, $\sigma = 1$, $\delta = .05$, $T = 1024$)



Initial (1:400) and tail (10^4:10^5 downsampled)



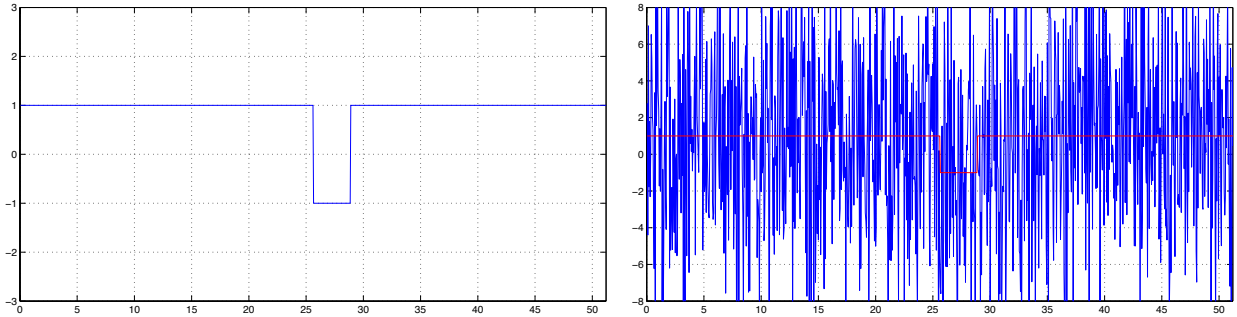
Statistics, initial and tail: Mean, RMS, and truth



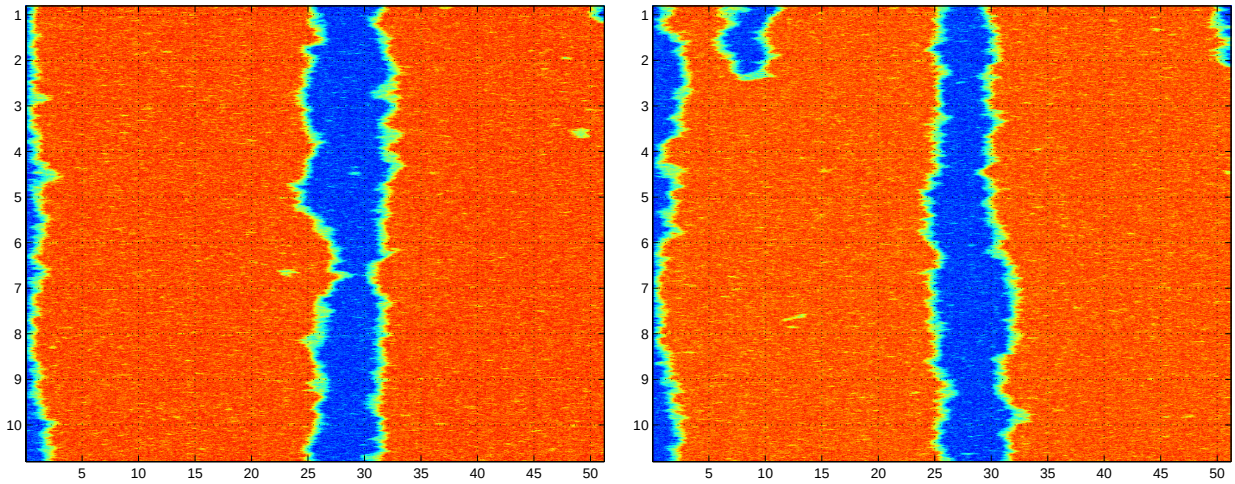
Jet Propulsion Laboratory
California Institute of Technology

MCMC Results: High Noise

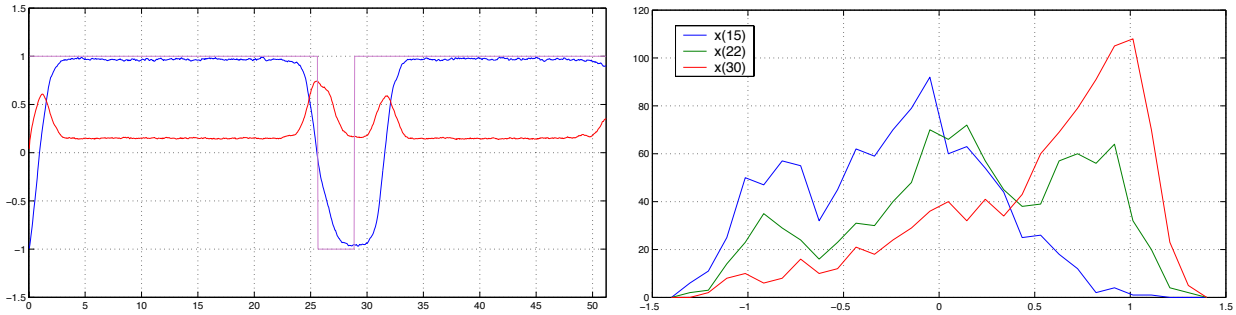
True x_n with data ($\kappa = 0.5$, $\sigma = 5$, $\delta = .05$, $T = 1024$)



Two MCMC series (10^4:10^5 downsampled 100:1)



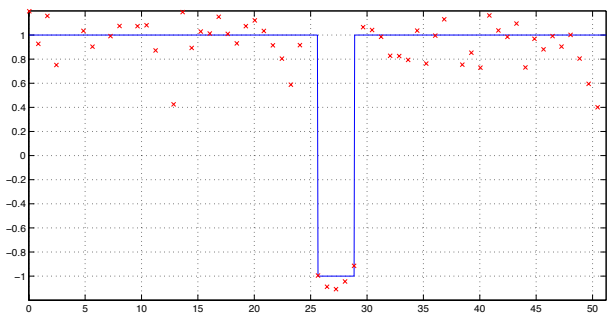
$E\,x_n$, $\sigma(x_n)$, truth; histograms of x_{15}, x_{22}, x_{30}

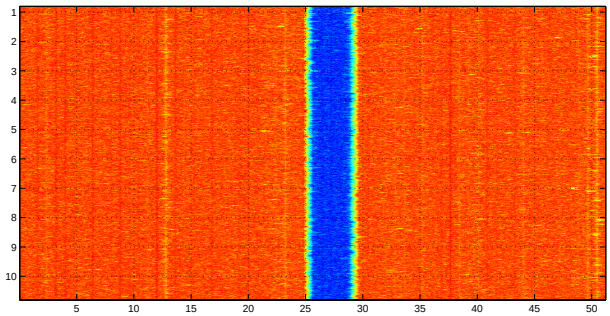


Jet Propulsion Laboratory
California Institute of Technology

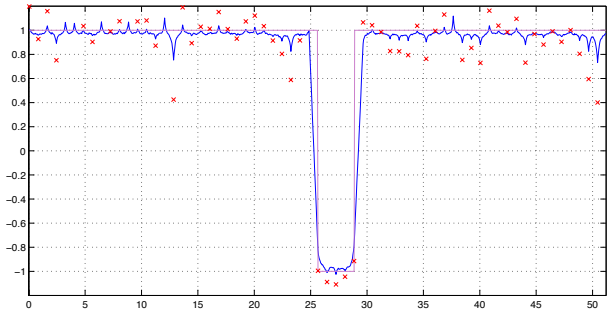
MCMC Results: Sparse Data

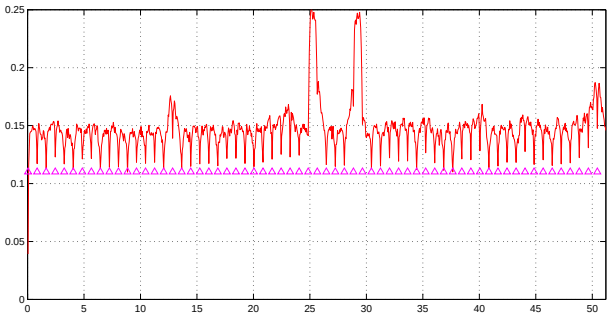
True x_n with data ($\kappa = 0.5$, $\sigma = 0.2$, $\delta = .05$, $T = 1024$)
But: downsample 16$\times$ (data rate $\ll$ simulation)



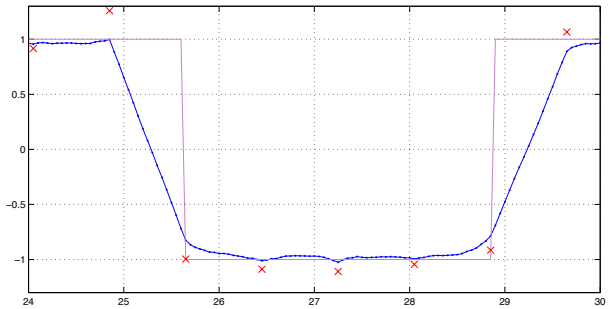


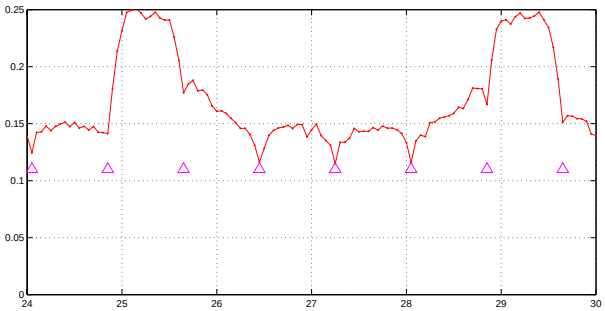
$E\,x_n$ and $\sigma(x_n)$: sampling interval indicated





Data assimilation versus path smoothness





Near-gaussian statistics disallow fast jumps

Jet Propulsion Laboratory
California Institute of Technology

Midpoint

- **MCMC**

  Represents global density with a single, random sample which is evolved across simulation epochs

- **Particle Filter**

  Represents density at a particular time with a set of random samples which are pushed forward in physical time

  Bootstrap sample: represent distribution with samples, not parameters like means and covariances

  Potentially much more expressive than Gaussian modeling
  Handles heavy tails
  Handles multimodal distributions

  Particle representation has a natural affinity with ensembles which lends itself to exploitation by practical forward modelers

Jet Propulsion Laboratory
California Institute of Technology

Bootstrap Representations

Approximate p with point masses drawn from p

$$p(x) \simeq R^{-1} \sum_{k=1}^{K} \delta(x - x^{(k)})$$

Fully analogous to the familiar

$$E\, f(x) = \int f(x)\, p(x)\, dx \simeq K^{-1} \sum_{k=1}^{K} f(x^{(k)})$$

- **Conditioning**

  Observe a certain y that is related to x above

$$\begin{aligned} p(x \,|\, y) &= p(y \,|\, x)\, p(x) \,/\, p(y) \\ &\simeq \frac{1}{K\, p(y)} \sum_{k=1}^{K} p(y \,|\, x^{(k)})\, \delta(x - x^{(K)}) \\ &= \sum_{k=1}^{K} w_k\, \delta(x - x^{(k)}) \qquad \Big(\sum w_k \equiv 1\Big) \end{aligned}$$

  Further conditioning multiplies new factors into w_k

- **Resampling**

  Some weights will decrease to near zero

  Regeneration: sample K' new points *with replacement* from the K current point masses

Jet Propulsion Laboratory
California Institute of Technology

Importance Sampling

We can develop a bootstrap approximation to p even when we can not sample directly from it

Use importance sampling

Suppose: can sample from $\pi(x)$

If $\pi(x) > 0$ where-ever $p(x) > 0$, represent:

$$p(x) \simeq \sum_{k=1}^{K} w_k \, \delta(x - x^{(k)})$$

$$w_k := p(x)/\pi(x), \text{ and } \sum w_k \equiv 1$$

This works because

$$E\, f(x) = \int f(x)\, p(x)\, dx = \int f(x) \frac{p(x)}{\pi(x)} \pi(x)\, dx$$

and we absorb the fraction into the modified weights

$\pi(\cdot)$ is the importance distribution

Can show that the quality of the approximation decreases as var(w_k) increases

(Ideally, all samples are counted equally)

Jet Propulsion Laboratory
California Institute of Technology

Particle Filter Algorithm

Begin the cycle with weighted representation

$$p(x_n \,|\, y_{1:n}) \sim \{\, (w_k, x^{(k)}) \,\}_{k=1}^{K}$$

Push sampled $x^{(k)}$'s thru nonlinearity, plus noise:

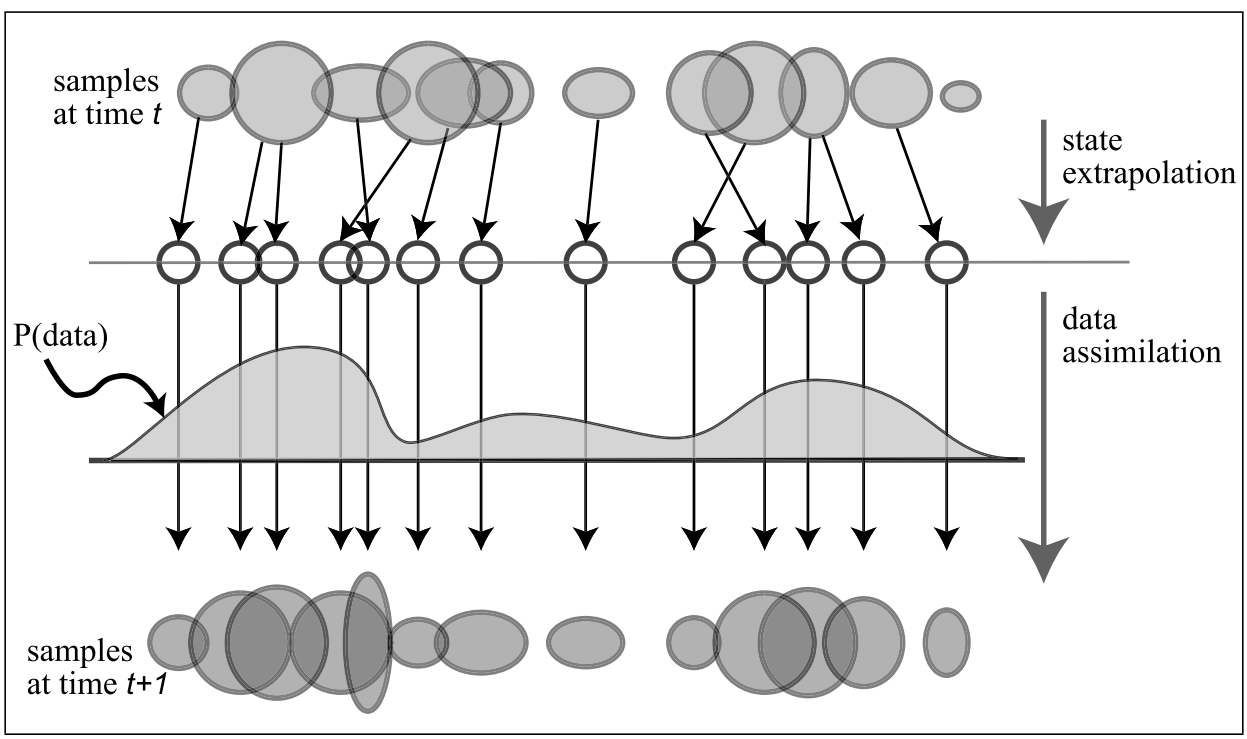
$$x_{\text{new}}^{(k)} = g(x^{(k)}) + r^{(k)} \qquad r^{(k)} \sim N(0, \kappa^2)$$

$$p(x_{n+1} \,|\, y_{1:n}) \sim \{\, x_{\text{new}}^{(k)} \,\}_{k=1}^{K}$$

Condition on new data y_{n+1} by reweighting

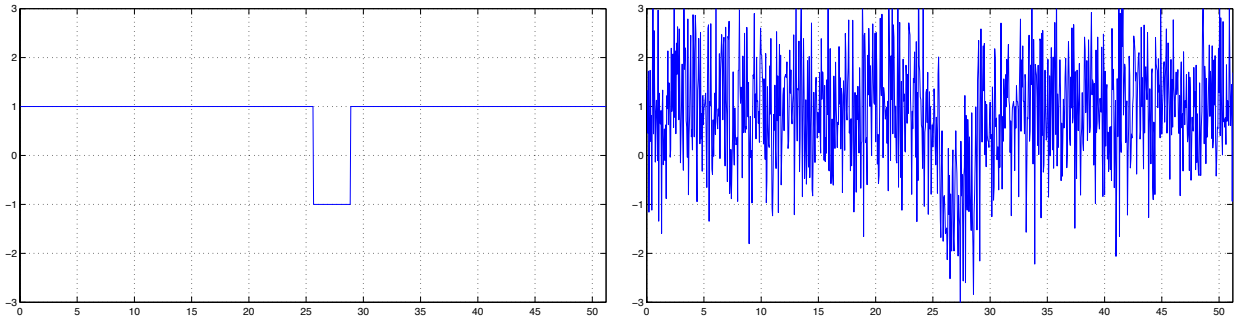
$$p(x_{n+1} \,|\, y_{1:n+1}) \sim \{\, (w_k, x_{\text{new}}^{(k)}) \,\}_{k=1}^{K}$$

$$w_k \propto p(y_{n+1} \,|\, x_{\text{new}}^{(k)})$$

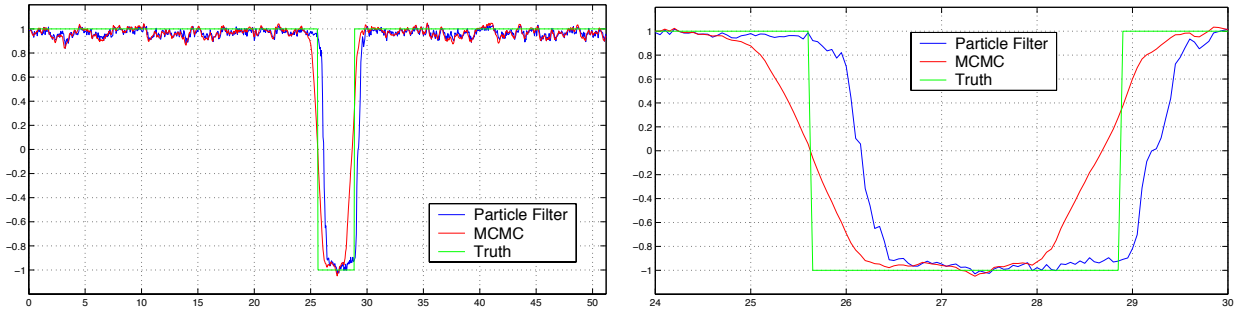


Sample Paths

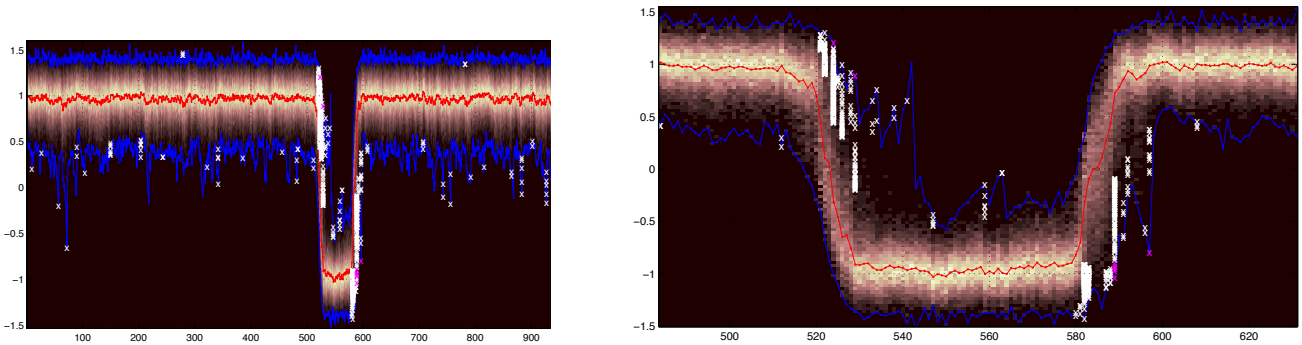
True x_n with data ($\kappa = 0.5$, $\sigma = 1$, $\delta = .05$, $T = 1024$)



Particle filter is 1-sided; a 2-sided version exists



Note occasional loss of diversity of particle set



Use a different importance density to patch this

Jet Propulsion Laboratory
California Institute of Technology

Future Work

- **MCMC**

  Proposals changing more than one value at a time

  Proposals using physical criteria

- **Particle Filter**

  Loss of particle diversity

  Measure this (e.g. via entropy) and adapt the importance density

- **Futures**

  More systematic comparisons, e.g. using an L1 error between target and estimate
  Kalman filter, MCMC, and particle filter

  Precise comparisons with the literature, e.g. Eyink and Restrepo

  Reasonable proposals for higher-dimensional systems

http://www-aig.jpl.nasa.gov/home/turmon/