Probabilistic Graphical Models

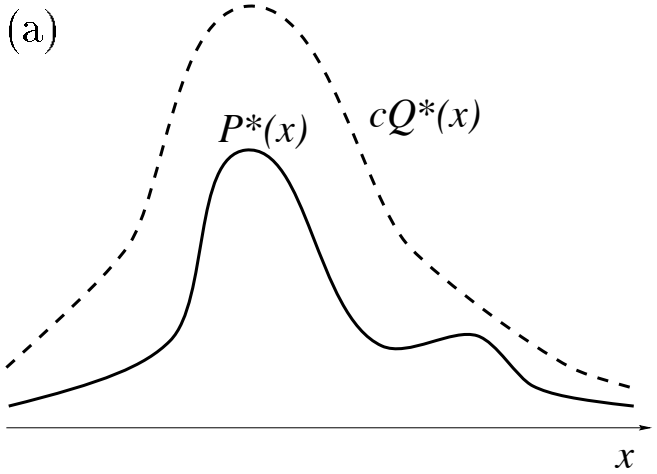
Brown University CSCI 2950-P, Spring 2013
Prof. Erik Sudderth

Lecture 15:
Importance Sampling, Particle Filters,
Sequential Monte Carlo Methods

Basic Monte Carlo Methods

Monte Carlo: $x^{(\ell)} \sim p(x)$

$$\mathbb{E}[f] = \int f(x)p(x)\,dx \approx \frac{1}{L}\sum_{\ell=1}^{L} f(x^{(\ell)})$$



Target Distribution:

$$p(x) = \frac{1}{Z}p^*(x)$$

Proposal Distribution:

$$q(x) = \frac{1}{Z'}q^*(x)$$

Rejection Sampling: Exact sampling with random computation

- Auxiliary variable method: $p(x,u) = p(x)p(u \mid x) = p(x)\mathrm{Unif}(u \mid 0, p^*(x))$
- Scaled proposal must bound target: $cq^*(x) > p^*(x) \text{ for all } x$

  we must know the envelope constant c

Importance Sampling: Approximate expectations, not samples

- Larger class of permissible proposals: $q(x) > 0 \text{ where } p(x) > 0$

$$\hat{f}_L = \sum_{\ell=1}^{L} w_\ell f(x^{(\ell)}) \qquad w_\ell = \frac{w^*(x^{(\ell)})}{\sum_{m=1}^{L} w^*(x^{(m)})}$$

$$w^*(x) = p^*(x)/q^*(x)$$

$$x^{(\ell)} \sim q(x)$$

Failures for High-Dimensional Posteriors

Monte Carlo: Computationally intractable

- In most cases, must be able to tractably manipulate inverse CDF
- Discrete variables: Sums exponential in number of variables
- Continuous variables: Intractable integrals, quadrature exponential

Rejection Sampling: Computationally intractable

- Small errors in matching individual marginal distributions compound to produce a very small overall rejection rate
- Problem worse if we can only find a conservative envelope bound:

$$cq^*(x) > p^*(x) \text{ for all } x$$

we must know the envelope constant c

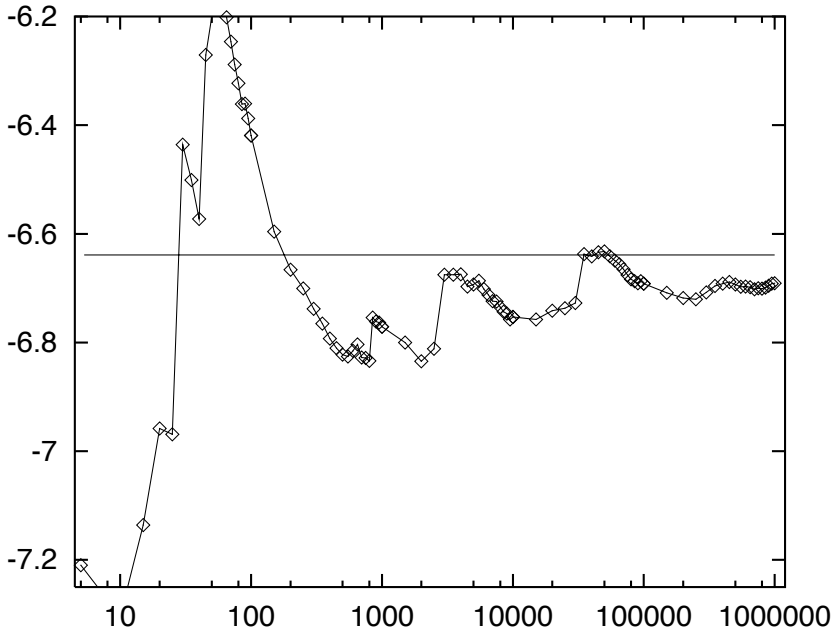
Importance Sampling: Efficiently gives inaccurate estimates

$$\mathbb{E}_q[\hat{f}_L] = \mathbb{E}_p[f] \triangleq \mu \qquad \mathrm{Var}_q[f(x)w(x)] = \mathbb{E}_q[f^2(x)w^2(x)] - \mu^2$$

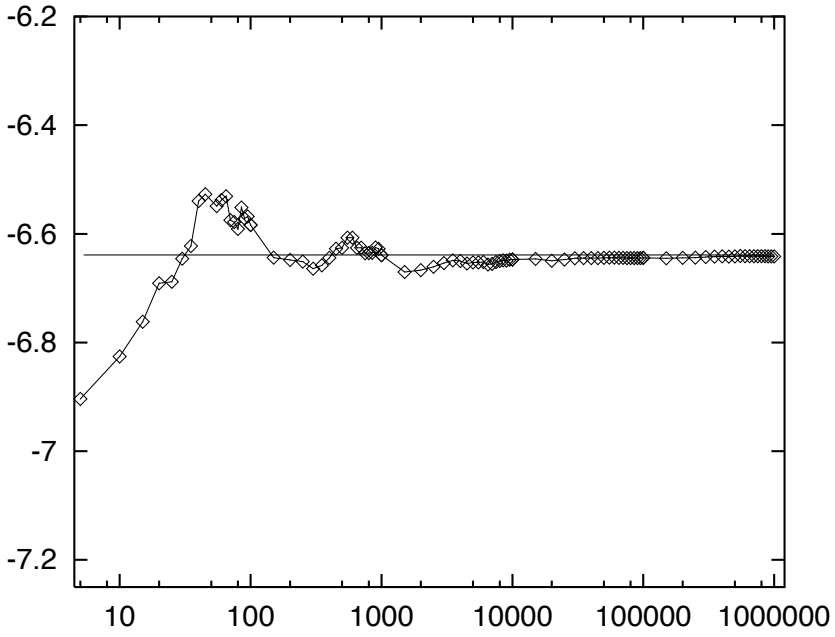
- If we don't satisfy the (difficult) envelope bound condition from rejection sampling, the weights for some samples will be very large
- For some proposals, the variance of the estimator may be infinite
- If we actually want samples, doesn't directly provide them

Selecting Proposal Distributions

- For a toy one-dimensional, heavy-tailed target distribution:



Gaussian Proposal



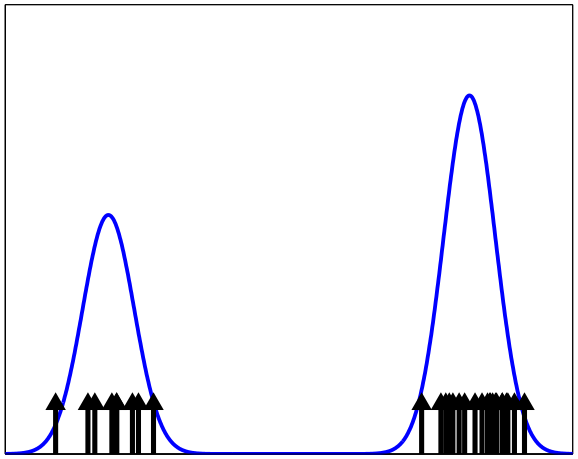
Cauchy (Student's-t) Proposal

Empirical variance of weights may not predict estimator variance

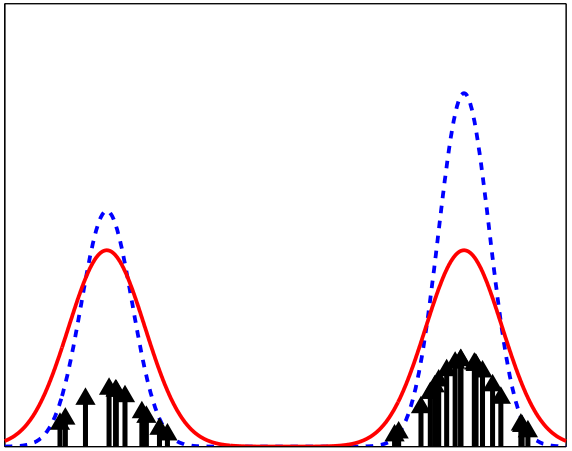
- Always (asymptotically) unbiased, but variance of estimator can be enormous unless weight function bounded above:

$$\mathbb{E}_q[\hat{f}_L] = \mathbb{E}_p[f] \qquad \mathrm{Var}_q[\hat{f}_L] = \frac{1}{L}\mathrm{Var}_q[f(x)w(x)] \qquad w(x) = \frac{p(x)}{q(x)}$$

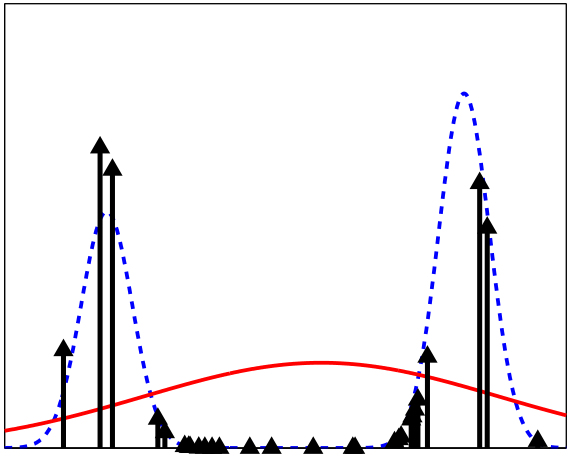
Selecting Proposal Distributions



Target Distribution



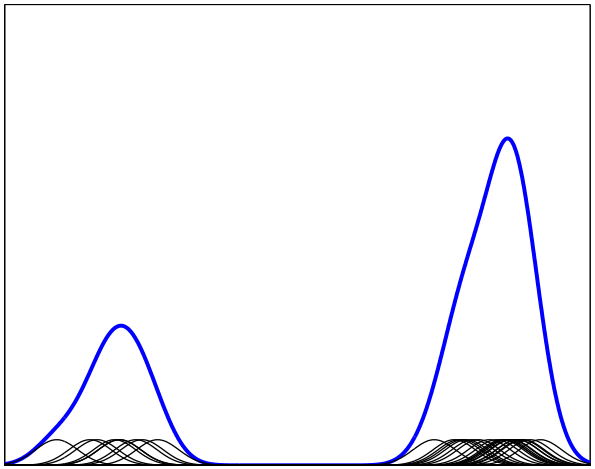
Good Proposal

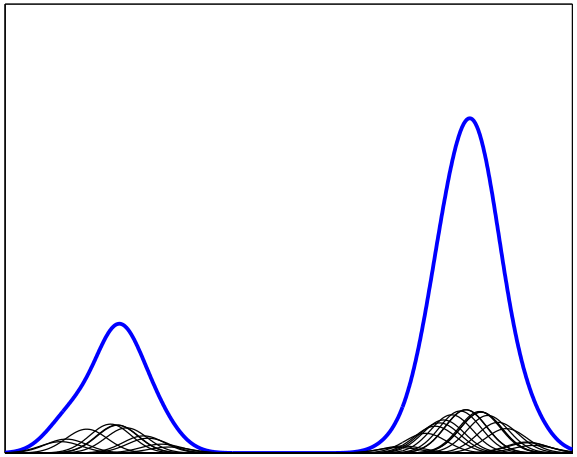


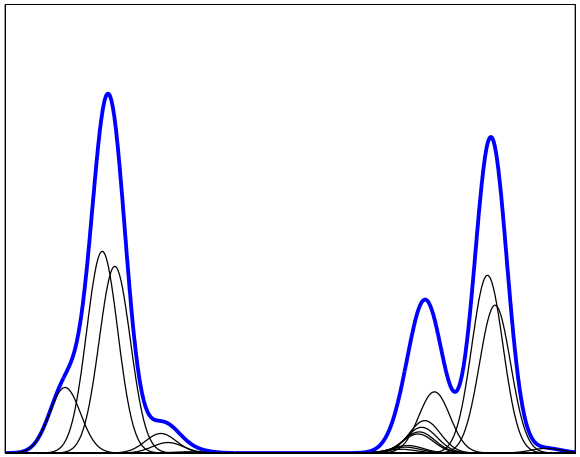
Poor Proposal

Kernel or Parzen window estimators interpolate for nonparametric density prediction

$$\hat{p}(x) = \sum_{\ell=1}^{L} w^{(\ell)} \mathcal{N}(x; x^{(\ell)}, \Lambda)$$







High-Dimensional Importance Sampling

- Consider an N-dimensional importance sampling problem:
- By Gaussian central limit theorem, norm of samples from proposal is nearly

$$\rho^2 \sim N\sigma^2 \pm \sqrt{2N}\sigma^2$$

- After 100 samples, ratio of largest weight to median weight will then be approximately

$$\frac{w_r^{\max}}{w_r^{\text{med}}} = \exp\left(\sqrt{2N}\right)$$

Uniform Target Distribution:

$$P^*(\mathbf{x}) = \begin{cases} 1 & 0 \le \rho(\mathbf{x}) \le R_P \\ 0 & \rho(\mathbf{x}) > R_P \end{cases}$$

$$\rho(\mathbf{x}) \equiv \left(\sum_i x_i^2\right)^{1/2}$$

Gaussian Proposal Distribution:

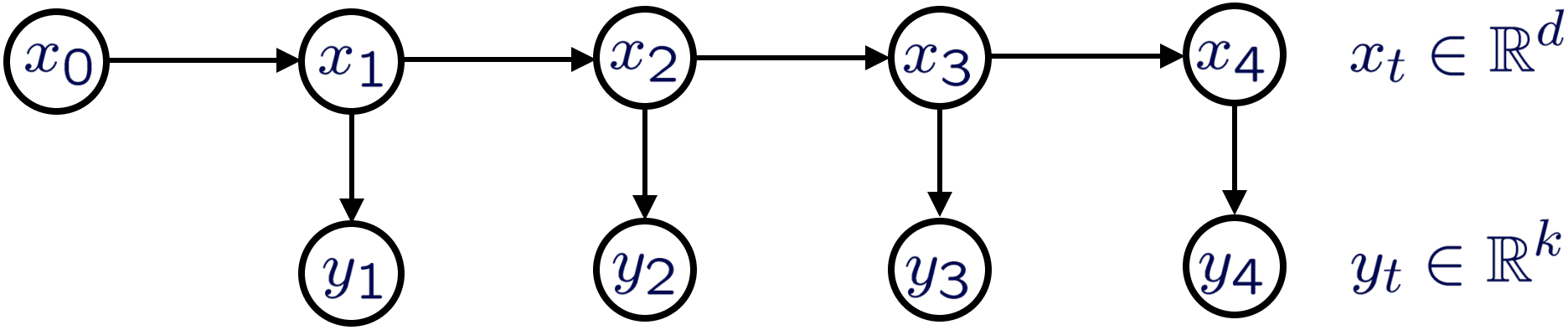
$$q(x) = \mathcal{N}(x \mid 0, \sigma^2 I_N)$$

Empirical variance of weights may not predict estimator variance

- Always (asymptotically) unbiased, but variance of estimator can be enormous unless weight function bounded above:

$$\mathbb{E}_q[\hat{f}_L] = \mathbb{E}_p[f] \qquad \mathrm{Var}_q[\hat{f}_L] = \frac{1}{L}\mathrm{Var}_q[f(x)w(x)] \qquad w(x) = \frac{p(x)}{q(x)}$$

Nonlinear State Space Models

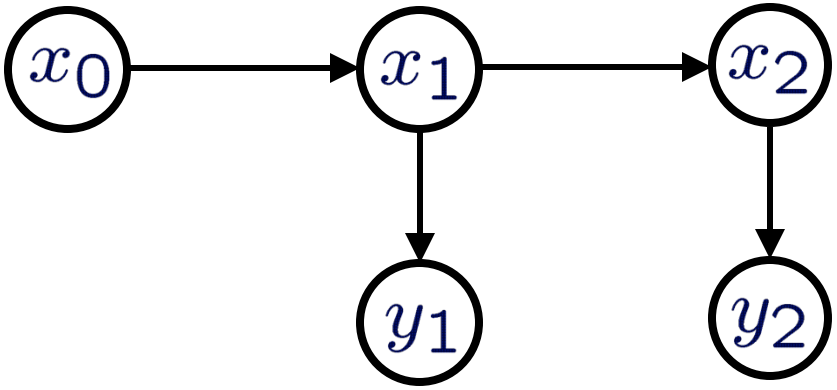


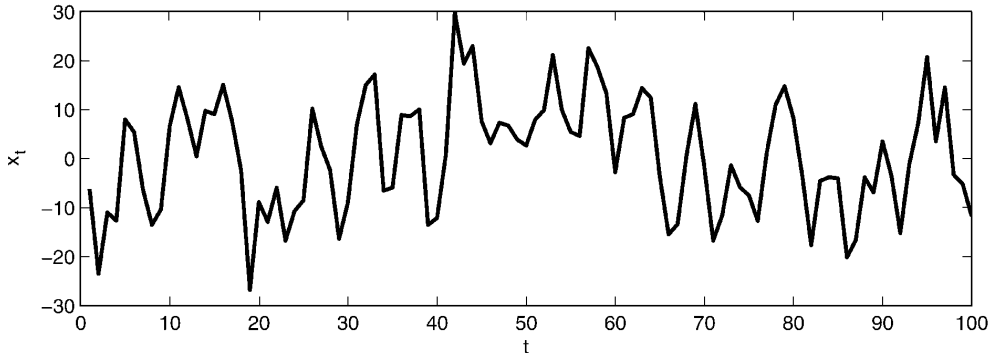
$$x_{t+1} = f(x_t, w_t) \qquad w_t \sim \mathcal{F}$$

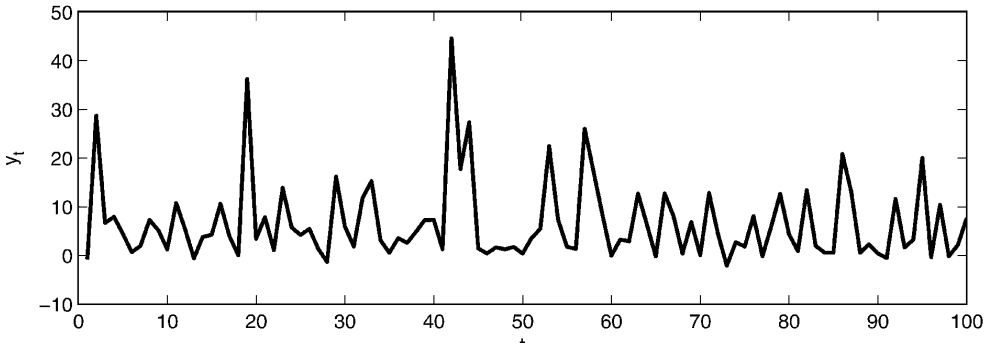
$$y_t = g(x_t, v_t) \qquad v_t \sim \mathcal{G}$$

- State dynamics and measurements given by potentially complex *nonlinear functions*
- Noise sampled from *non-Gaussian* distributions

A Toy Nonlinear Model







$$x_t = \frac{x_{t-1}}{2} + 25\frac{x_{t-1}}{1+x_{t-1}^2} + 8\cos(1.2t) + u_t$$

$$y_t = \frac{x_t^2}{20} + v_t$$

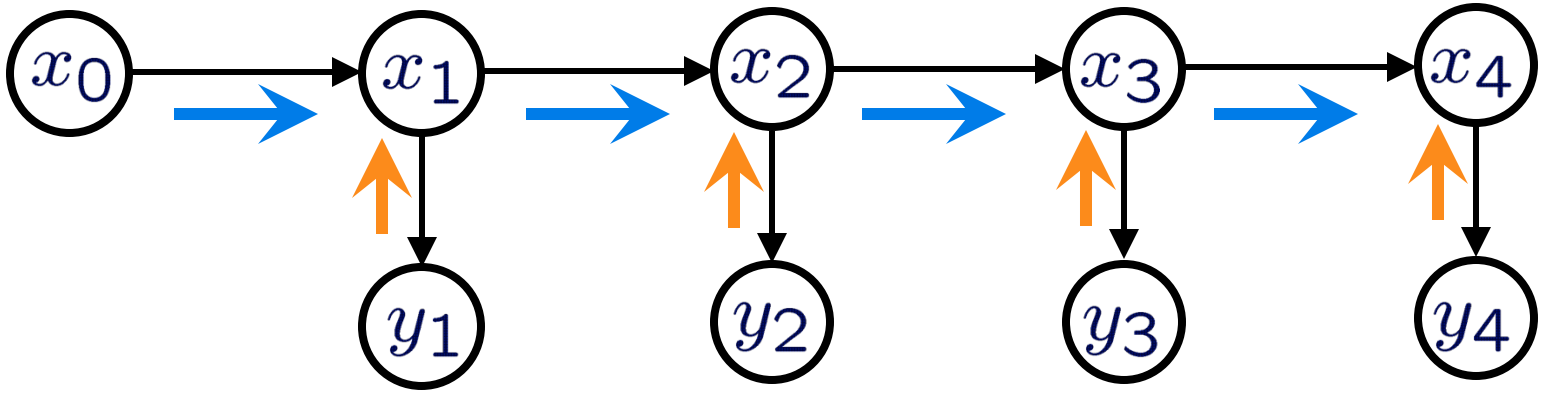
$$u_t \sim \mathcal{N}(0, \sigma_u^2)$$

$$v_t \sim \mathcal{N}(0, \sigma_v^2)$$

$$f(x_t|x_{t-1}) = \mathcal{N}\left(x_t \middle| \frac{x_{t-1}}{2} + 25\frac{x_{t-1}}{1+x_{t-1}^2} + 8\cos(1.2t), \sigma_u^2\right)$$

$$g(y_t|x_t) = \mathcal{N}\left(y_t \middle| \frac{x_t^2}{20}, \sigma_v^2\right).$$

Nonlinear Filtering



$$p(x_t \mid y_1, \ldots, y_{t-1}) = \tilde{q}_t(x_t)$$

$$p(x_t \mid y_1, \ldots, y_t) = q_t(x_t)$$

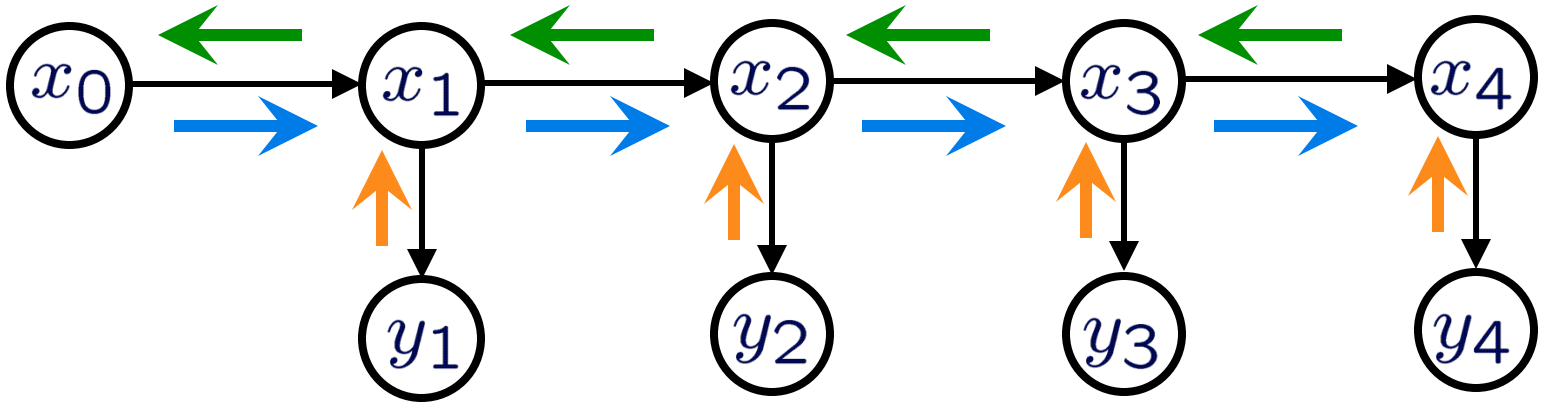
Prediction:

$$\tilde{q}_t(x_t) = \int p(x_t \mid x_{t-1}) q_{t-1}(x_{t-1}) \, dx_{t-1}$$

Update:

$$q_t(x_t) = \frac{1}{Z_t} \tilde{q}_t(x_t) p(y_t \mid x_t)$$

Monte Carlo Estimation



- Suppose interested in some complex, global function of state:

$$\mathbb{E}[f] = \int f(x)p(x \mid y)\, dx \approx \frac{1}{L}\sum_{\ell=1}^{L} f(x^{(\ell)}) \qquad x^{(\ell)} \sim p(x \mid y)$$

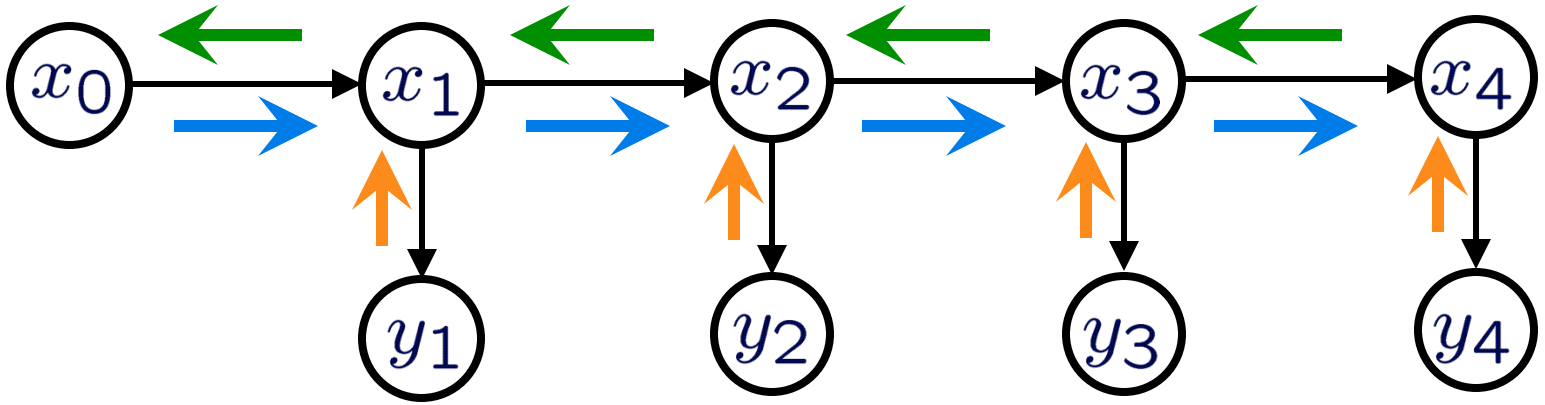
- Can efficiently draw joint samples from posterior marginals:
  - Forward Message Passing:
  - Backwards Sampling:

$$p(x_t \mid y), p(x_t, x_{t+1} \mid y)$$

$$x_T^{(\ell)} \sim p(x_T \mid y)$$
$$x_{T-1}^{(\ell)} \sim p(x_{T-1} \mid x_T^{(\ell)}, y)$$
$$x_{T-2}^{(\ell)} \sim p(x_{T-2} \mid x_{T-1}^{(\ell)}, y)$$

$$(x_1^{(\ell)}, x_2^{(\ell)}, \ldots, x_T^{(\ell)}) \sim p(x \mid y)$$

Monte Carlo Estimation



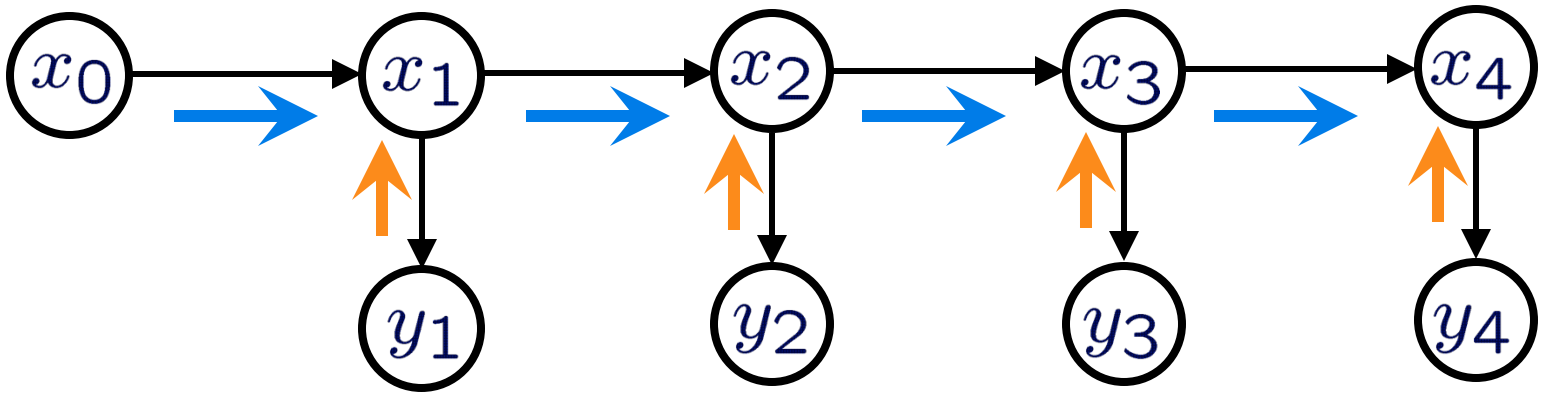
- Procedure only tractable for a limited class of models:
  - Discrete states: Sum-product belief propagation algorithm
  - Gaussian continuous states: Kalman smoothing algorithm

- Can efficiently draw joint samples from posterior marginals:
  - Forward Message Passing: $p(x_t \mid y), p(x_t, x_{t+1} \mid y)$
  - Backwards Sampling:

$$x_T^{(\ell)} \sim p(x_T \mid y)$$
$$x_{T-1}^{(\ell)} \sim p(x_{T-1} \mid x_T^{(\ell)}, y)$$
$$x_{T-2}^{(\ell)} \sim p(x_{T-2} \mid x_{T-1}^{(\ell)}, y)$$

$$(x_1^{(\ell)}, x_2^{(\ell)}, \ldots, x_T^{(\ell)}) \sim p(x \mid y)$$

Standard Importance Sampling



- Suppose interested in some complex, global function of state:

$$\mathbb{E}[f] = \int f(x)p(x \mid y)\, dx \approx \frac{1}{L}\sum_{\ell=1}^{L} f(x^{(\ell)}) \qquad x^{(\ell)} \sim p(x \mid y)$$

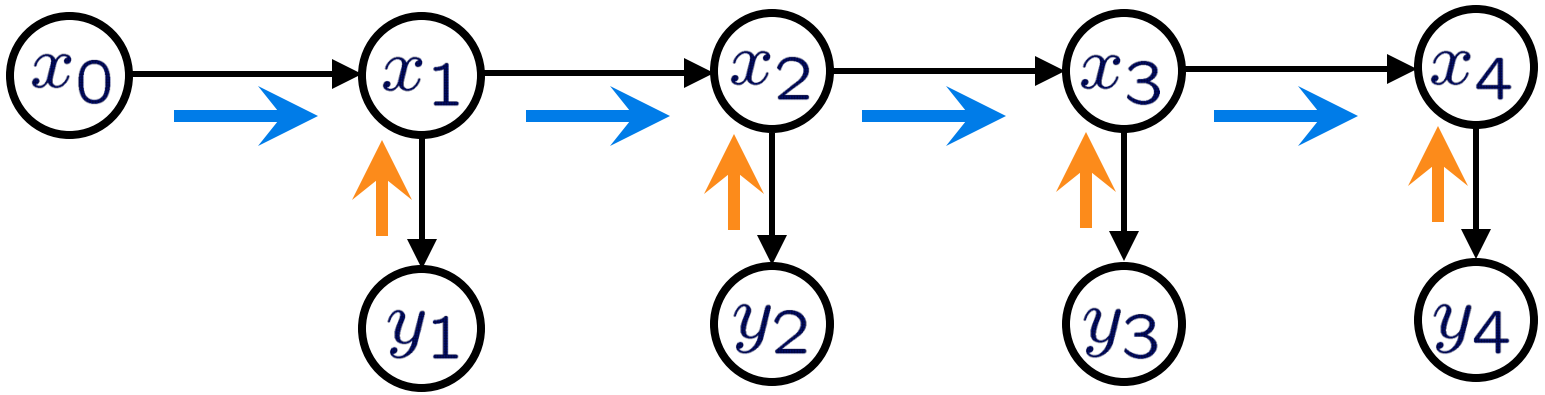
- Could use Markov structure to construct efficient proposal:

$$q(x \mid y) = q(x_0)\prod_{t=1}^{T} q(x_t \mid x_{t-1}, y_t)$$

$$q(x_t \mid x_{t-1}, y_t) \approx p(x_t \mid x_{t-1}, y)$$

Small local errors give large global estimator variance

Sequential Monte Carlo



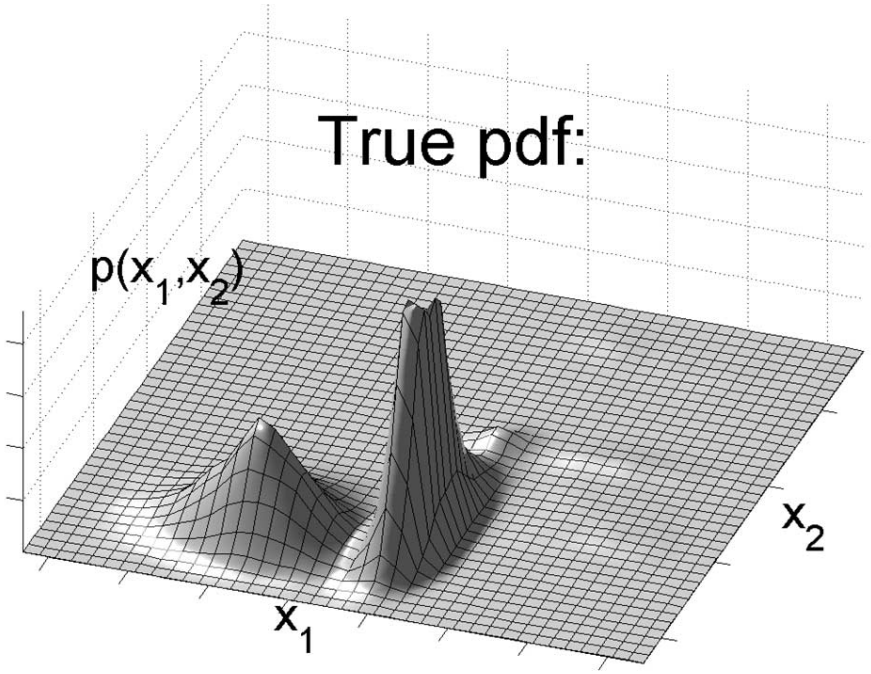
$$m_{t,t+1}(x_{t+1}) \propto \int_{\mathcal{X}_t} p(x_{t+1} \mid x_t)\, p(y_t \mid x_t)\, m_{t-1,t}(x_t)\; dx_t$$

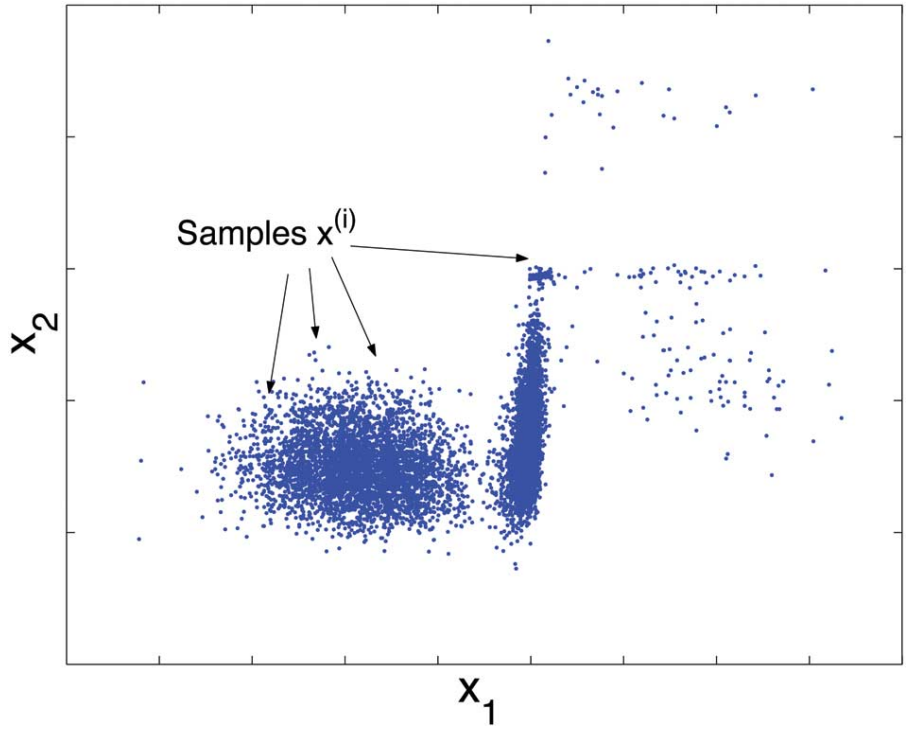
$$m_{t-1,t}(x_t) \propto p\left(x_t \mid y_{\overline{t-1}}\right)$$

$$m_{t-1,t}(x_t)\, p(y_t \mid x_t) \propto p(x_t \mid y_{\overline{t}})$$

$$y_{\overline{t}} = \{y_0, y_1, \ldots, y_t\}$$

Particle-Based Density Estimates





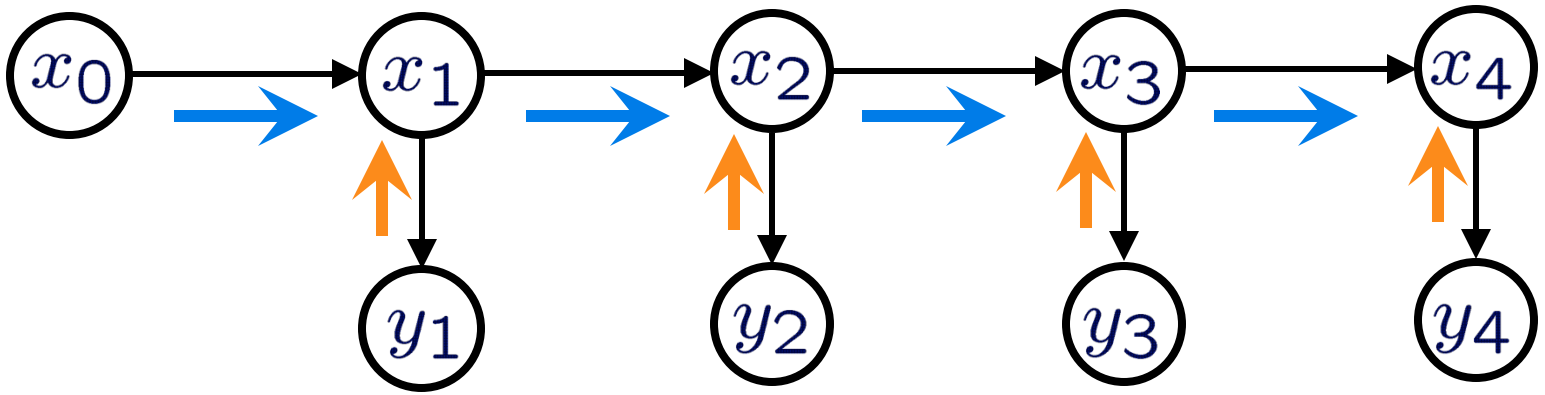
Particle-Based Posterior State Estimates:

- Approximate density by set of (possibly weighted) samples
- Dynamically move samples to the most probable parts of space
- Do this in a way which minimizes bias

$$m_{t-1,t}(x_t) \approx \sum_{\ell=1}^{L} w_{t-1,t}^{(\ell)} \delta(x_t, x_t^{(\ell)})$$

$$\sum_{\ell=1}^{L} w_{t-1,t}^{(\ell)} = 1$$

Sequential Monte Carlo



Measurement Update:

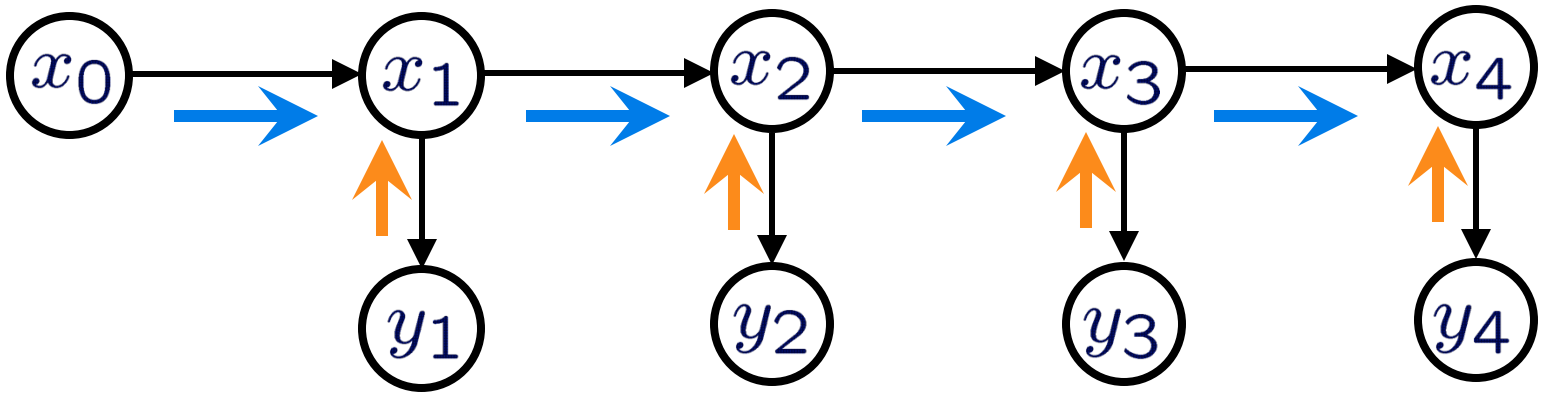
$$m_{t-1,t}(x_t) \approx \sum_{\ell=1}^{L} w_{t-1,t}^{(\ell)} \delta(x_t, x_t^{(\ell)}) \qquad \sum_{\ell=1}^{L} w_{t-1,t}^{(\ell)} = 1$$

$$q_{\bar{t}}(x_t) \propto m_{t-1,t}(x_t)\, p(y_t \mid x_t) \propto \sum_{\ell=1}^{L} w_{t-1,t}^{(\ell)}\, p(y_t \mid x_t^{(\ell)})\, \delta(x_t, x_t^{(\ell)})$$

$$q_{\bar{t}}(x_t) = \sum_{\ell=1}^{L} w_t^{(\ell)} \delta(x_t, x_t^{(\ell)}) \qquad w_t^{(\ell)} \triangleq \frac{w_{t-1,t}^{(\ell)}\, p(y_t \mid x_t^{(\ell)})}{\sum_{m=1}^{L} w_{t-1,t}^{(m)}\, p(y_t \mid x_t^{(m)})}$$

*Variance of importance weights **increases** with each update*

Sequential Monte Carlo



Sample Propagation:

$$q_{\overline{t}}(x_t) = \sum_{\ell=1}^{L} w_t^{(\ell)} \delta(x_t, x_t^{(\ell)})$$

$$m_{t,t+1}(x_{t+1}) = \sum_{\ell=1}^{L} w_{t,t+1}^{(\ell)} \delta(x_{t+1}, x_{t+1}^{(\ell)})$$

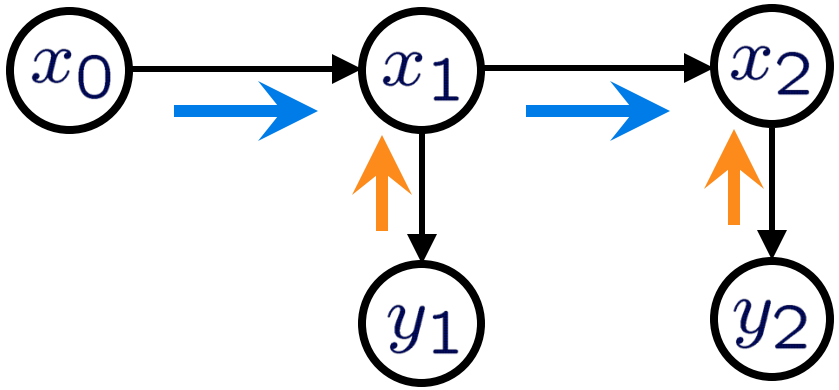
$$x_{t+1}^{(\ell)} \sim p(x_{t+1} \mid x_t^{(\ell)})$$

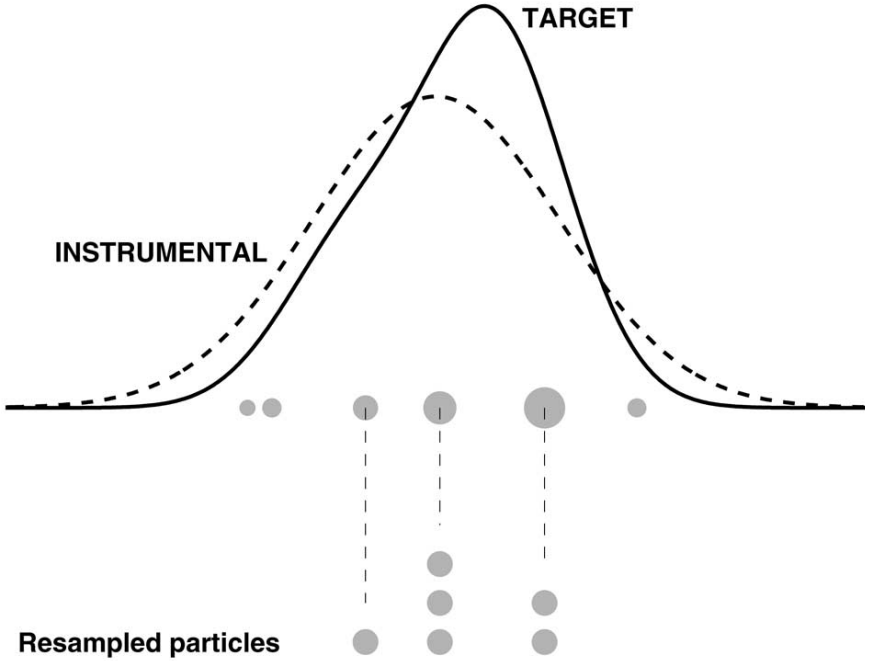
$$w_{t,t+1}^{(\ell)} = w_t^{(\ell)}$$

Justify as importance estimate of joint distribution $p(x_t, x_{t+1} \mid y_{\overline{t}})$

Assumption for now: Can simulate temporal dynamics

Sequential Monte Carlo





Resampling to Avoid Depletion:

$$q_{\overline{t}}(x_t) = \sum_{\ell=1}^{L} w_t^{(\ell)} \delta(x_t, x_t^{(\ell)})$$

$$m_{t,t+1}(x_{t+1}) = \sum_{\ell=1}^{L} w_{t,t+1}^{(\ell)} \delta(x_{t+1}, x_{t+1}^{(\ell)})$$

$$\tilde{x}_t^{(\ell)} \sim q_{\overline{t}}(x_t)$$

$$x_{t+1}^{(\ell)} \sim p(x_{t+1} \mid \tilde{x}_t^{(\ell)})$$

$$w_{t,t+1}^{(\ell)} = 1/L$$

Resampling with replacement preserves expectations, but increases the variance of subsequent estimators

Sequential Monte Carlo

Effective Sample Size:

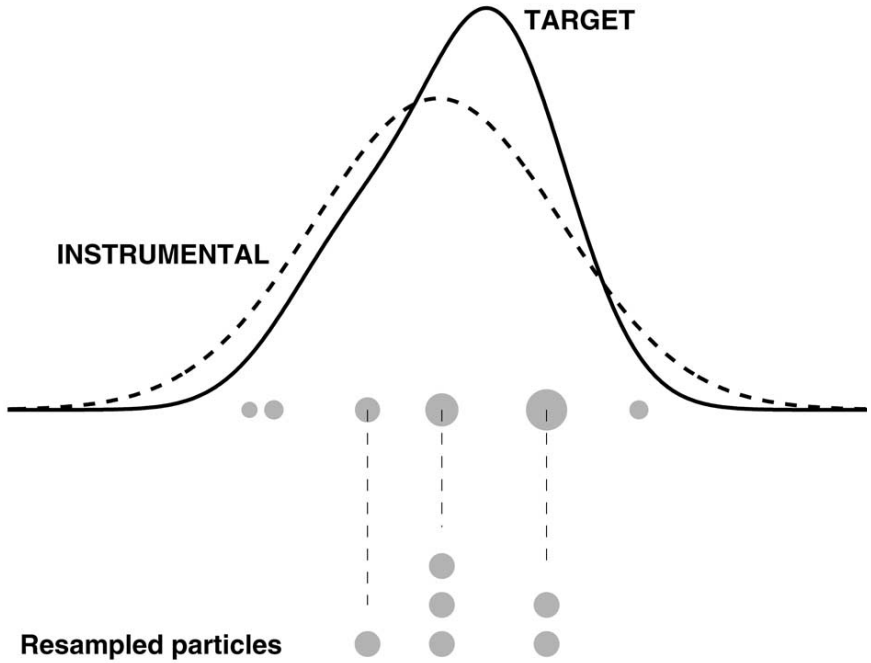
$$L_{\text{eff}} = \left(\sum_{\ell=1}^{L} \left(w^{(\ell)} \right)^2 \right)^{-1}$$

$$1 \leq L_{\text{eff}} \leq L$$

Resampling to Avoid Depletion:

$$q_{\overline{t}}(x_t) = \sum_{\ell=1}^{L} w_t^{(\ell)} \delta(x_t, x_t^{(\ell)})$$

$$m_{t,t+1}(x_{t+1}) = \sum_{\ell=1}^{L} w_{t,t+1}^{(\ell)} \delta(x_{t+1}, x_{t+1}^{(\ell)})$$



$$\tilde{x}_t^{(\ell)} \sim q_{\overline{t}}(x_t)$$

$$x_{t+1}^{(\ell)} \sim p(x_{t+1} \mid \tilde{x}_t^{(\ell)})$$

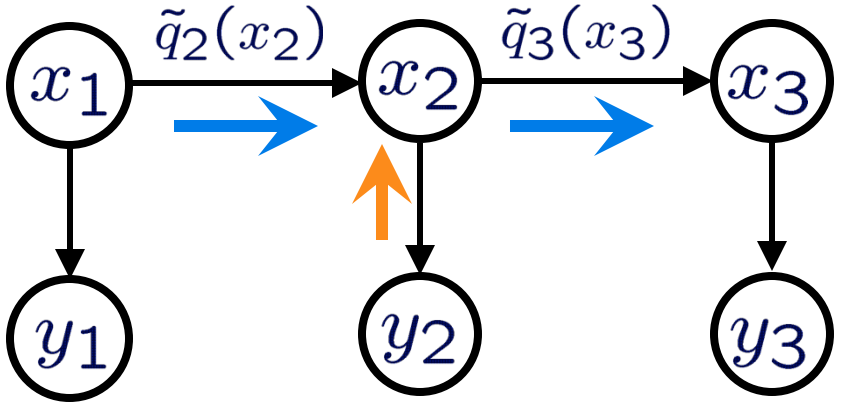
$$w_{t,t+1}^{(\ell)} = 1/L$$

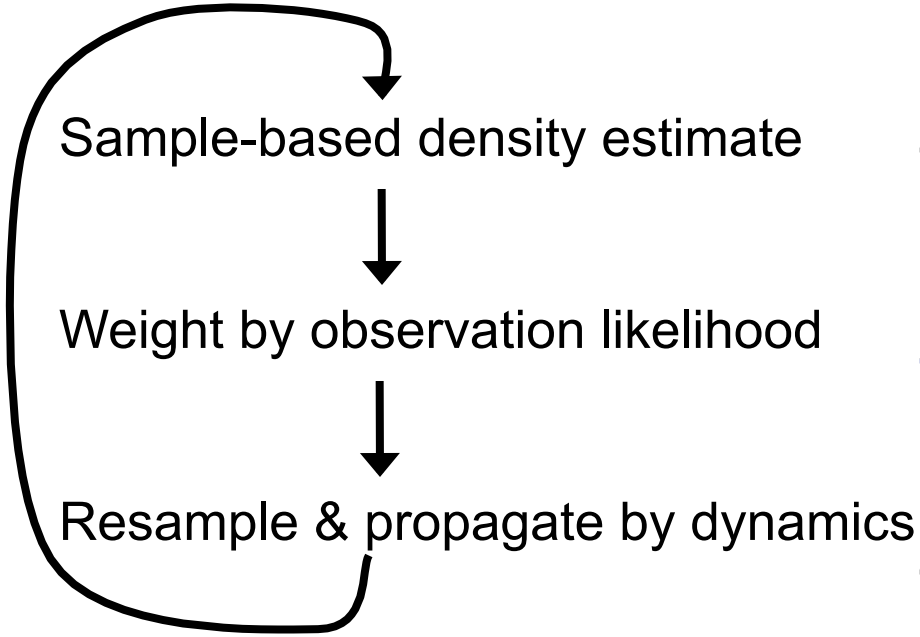
Resampling with replacement preserves expectations, but increases the variance of subsequent estimators

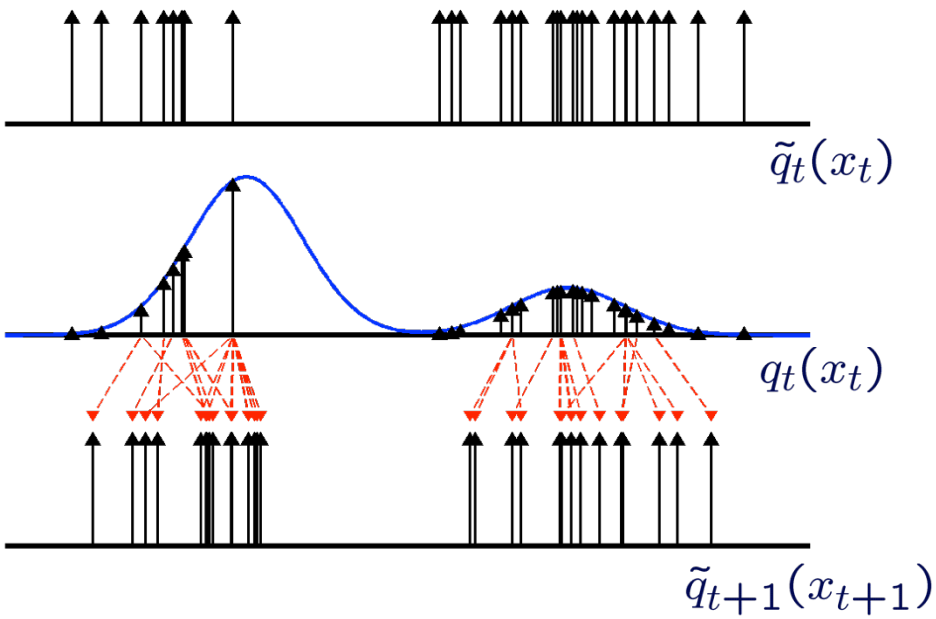
Particle Filters

Condensation, Sequential Monte Carlo, Survival of the Fittest,…

- Represent state estimates using a set of samples
- Propagate over time using *importance sampling*





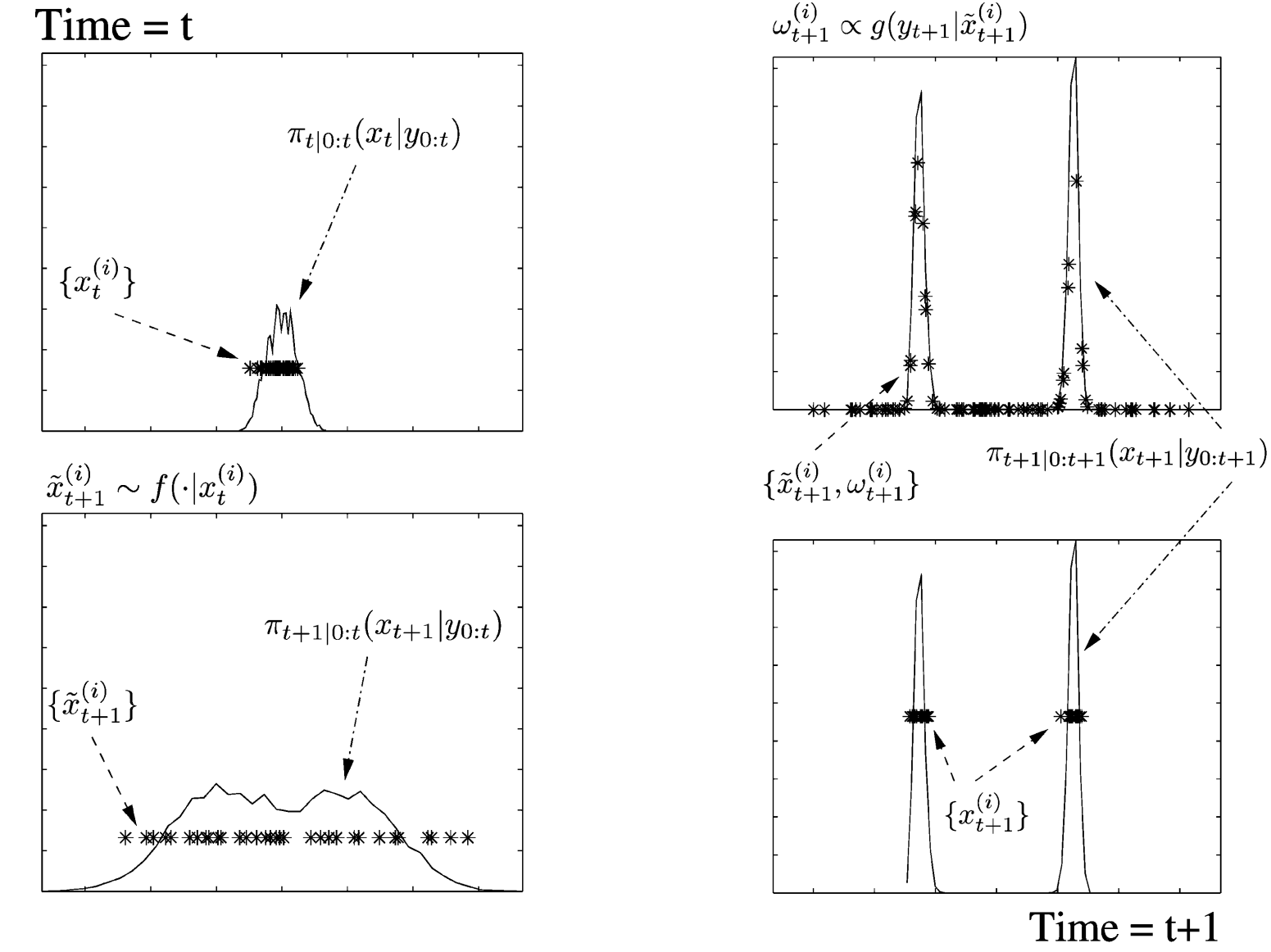


Particle Filtering Movie

(M. Isard, 1996)

Bootstrap Filters

Particle filters where temporal dynamics used to propagate samples.



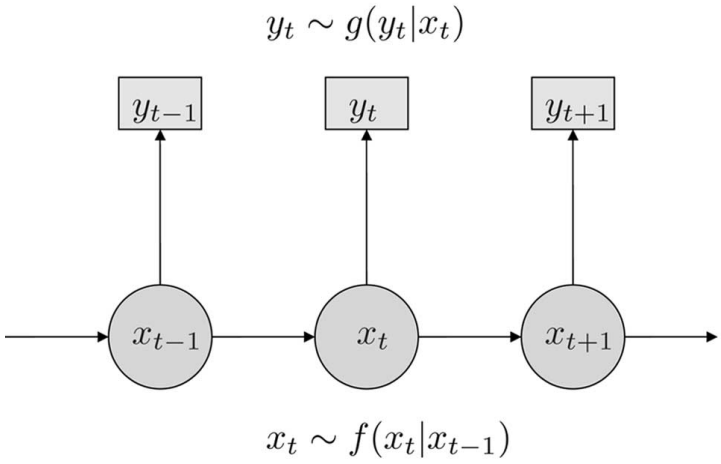
General Sequential Monte Carlo

$$\bar{h} = \int h(x_{0:t})\pi_{0:t|0:t}(x_{0:t}|y_{0:t})dx_{0:t}$$
$$\approx \sum_{i=1}^{N} \frac{\tilde{\omega}_t^{(i)}}{\sum_{j=1}^{N}\tilde{\omega}_t^{(j)}} h\left(\tilde{x}_{0:t}^{(i)}\right).$$

$$\tilde{\omega}_t^{(i)} = \frac{\pi_{0:t|0:t}\left(\tilde{x}_{0:t}^{(i)}|y_{0:t}\right)}{q_{0:t}\left(\tilde{x}_{0:t}^{(i)}|y_{0:t}\right)}$$

Target Posterior

Proposal Density



- Exploit temporal structure to propose sequences recursively:

$$q_{0:t}(x_{0:t}|y_{0:t}) = \overbrace{q_{0:t-1}(x_{0:t-1}|y_{0:t-1})}^{\text{Keep existing path}}\overbrace{q_t(x_t|x_{t-1},y_t)}^{\text{extend path}}$$

$$\tilde{\omega}_t^{(i)} = \frac{\pi_{0:t|0:t}\left(\tilde{x}_{0:t}^{(i)}|y_{0:t}\right)}{q_{0:t}\left(\tilde{x}_{0:t}^{(i)}|y_{0:t}\right)} \propto \omega_{t-1}^{(i)} \times \frac{f\left(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)}\right)g\left(y_t|\tilde{x}_t^{(i)}\right)}{q_t\left(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)},y_t\right)\ell_{t|0:t-1}(y_t|y_{0:t-1})}$$

- Sequential importance sampling without resampling:

$$\tilde{x}_t^{(i)} \sim q_t\left(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)},y_t\right) \qquad \tilde{\omega}_t^{(i)} = \omega_{t-1}^{(i)}\frac{g\left(y_t|\tilde{x}_t^{(i)}\right)f\left(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)}\right)}{q_t\left(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)},y_t\right)} \qquad \omega_t^{(i)} = \tilde{\omega}_t^{(i)}/\sum_{j=1}^{N}\tilde{\omega}_t^{(j)}$$

General Sequential Monte Carlo

- Resampling happens at the end of one of the updates below, after importance reweighting but before propagation:
  - Mean is unchanged, does not introduce extra bias
  - Variance for estimates up to that time *strictly increases*
  - But, equalizing weights improves subsequent steps
- Fancier proposals try to approximate $q_t(x_t|x_{t-1},y_t) = \frac{f(x_t|x_{t-1})g(y_t|x_t)}{\int f(x|x_{t-1})g(y_t|x)dx}$
- Exploit temporal structure to propose sequences recursively:

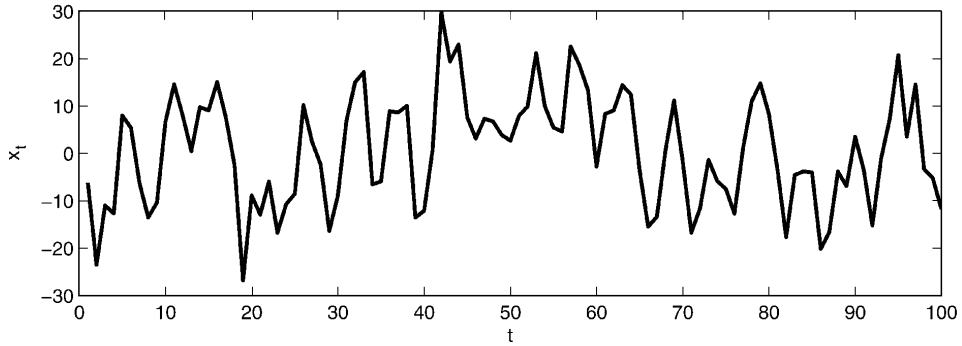
$$q_{0:t}(x_{0:t}|y_{0:t}) = \overbrace{q_{0:t-1}(x_{0:t-1}|y_{0:t-1})}^{\text{Keep existing path}}\overbrace{q_t(x_t|x_{t-1},y_t)}^{\text{extend path}}$$

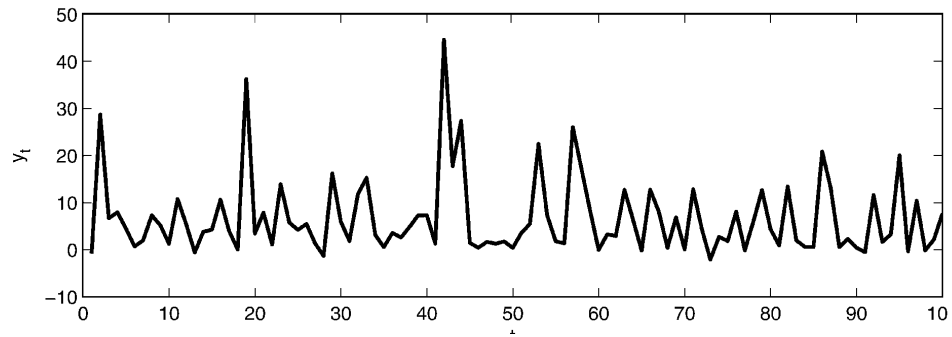
$$\tilde{\omega}_t^{(i)} = \frac{\pi_{0:t|0:t}\left(\tilde{x}_{0:t}^{(i)}|y_{0:t}\right)}{q_{0:t}\left(\tilde{x}_{0:t}^{(i)}|y_{0:t}\right)} \propto \omega_{t-1}^{(i)} \times \frac{f\left(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)}\right)g\left(y_t|\tilde{x}_t^{(i)}\right)}{q_t\left(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)},y_t\right)\ell_{t|0:t-1}(y_t|y_{0:t-1})}$$

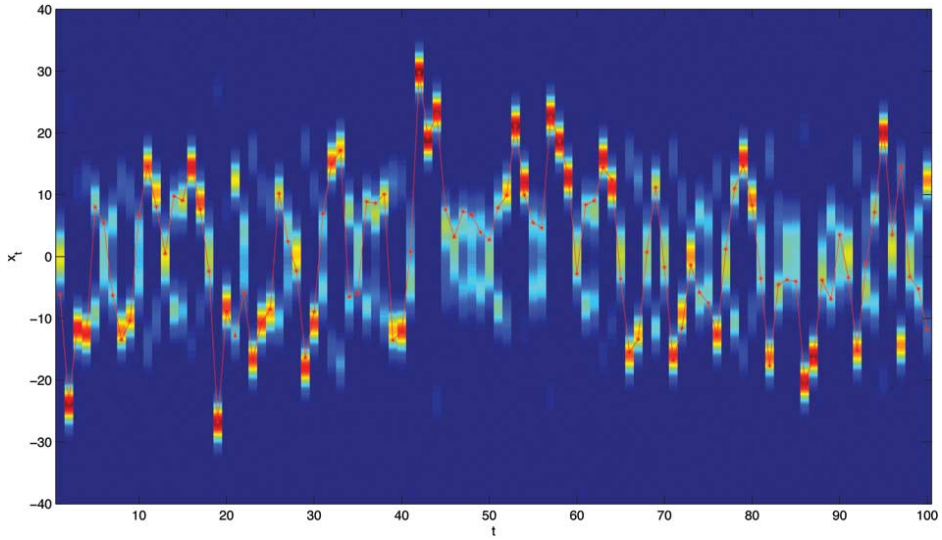
- Sequential importance sampling without resampling:

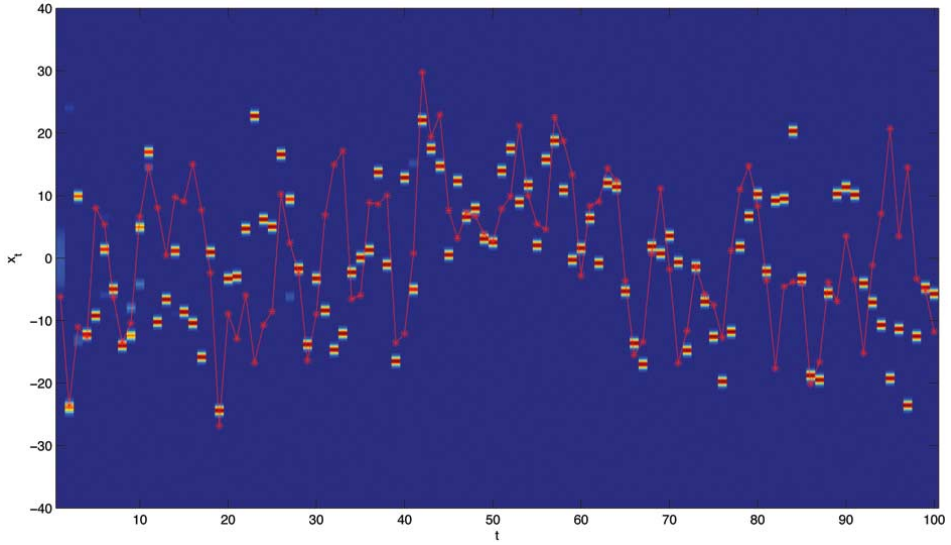
$$\tilde{x}_t^{(i)} \sim q_t\left(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)},y_t\right) \qquad \tilde{\omega}_t^{(i)} = \omega_{t-1}^{(i)}\frac{g\left(y_t|\tilde{x}_t^{(i)}\right)f\left(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)}\right)}{q_t\left(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)},y_t\right)} \qquad \omega_t^{(i)} = \tilde{\omega}_t^{(i)} / \sum_{j=1}^{N}\tilde{\omega}_t^{(j)}$$

Resampling: The Good







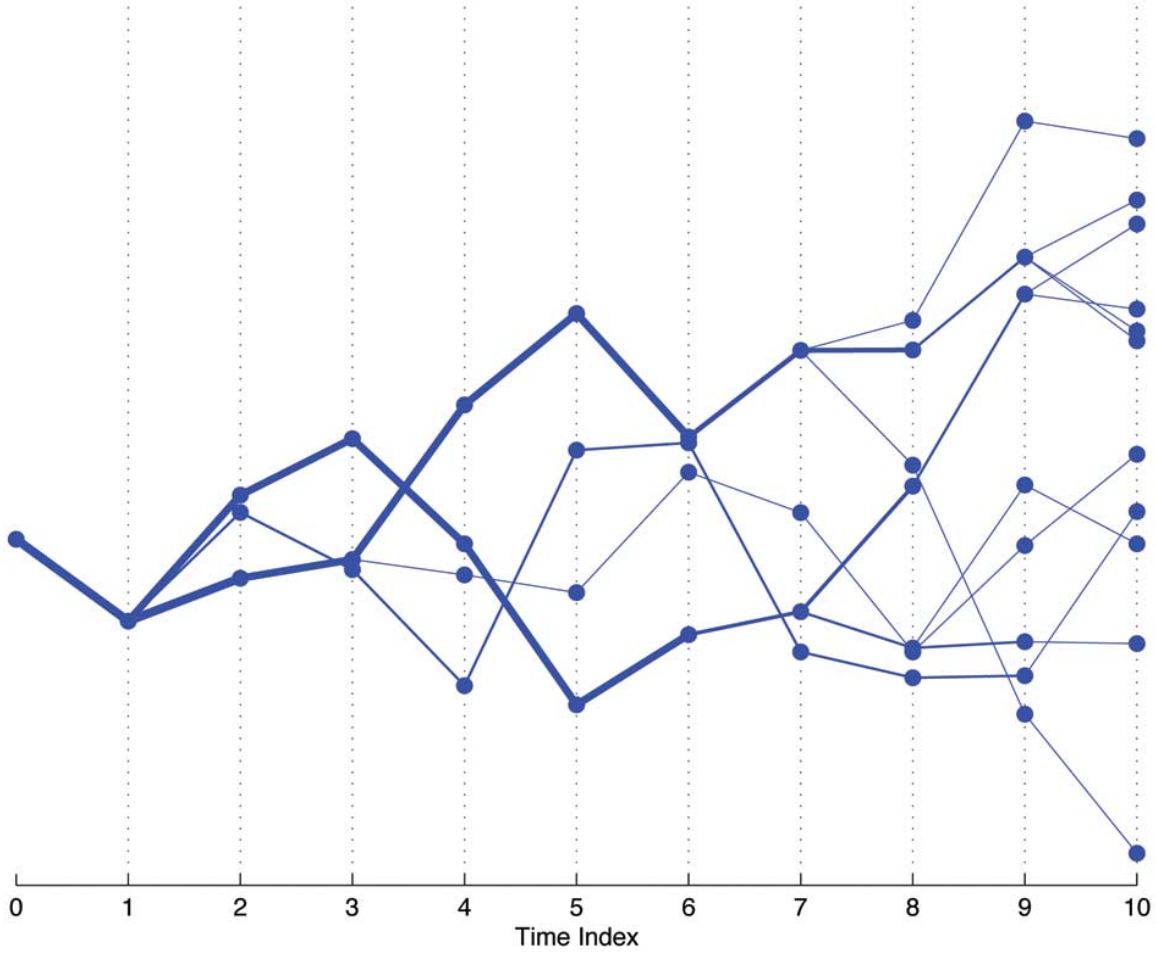


Full Particle Filter KDEs

Full Sequence Importance Sampling

- What is the probability that a state sequence, sampled from the prior model, is consistent with all observations?
- Marginal estimates degenerate on single, mediocre sample

Resampling: The Bad



- After each resampling step, some particles are discarded, and can never be restored in subsequent stages.
- Estimates of *smoothed* marginals/sequences typically poor