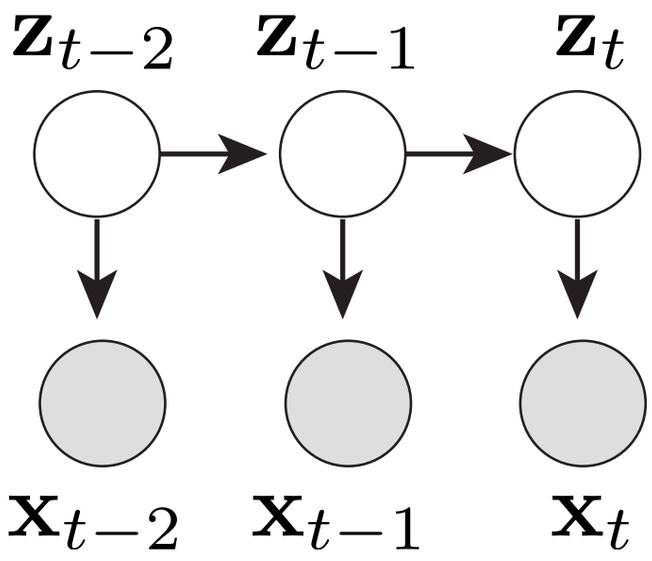
DS-GA 3001.008 Modelling time series data

L4. Latent state models 2: EM. Particle filtering

Instructor: Cristina Savin

NYU, CNS & CDS

LDS recap



$$\mathbf{z}_t = \mathbf{A}\mathbf{z}_{t-1} + \mathbf{w}_t$$
$$\mathbf{x}_t = \mathbf{C}\mathbf{z}_t + \mathbf{v}_t$$

$$\mathbf{w}_t \sim \mathcal{N}(0, \mathbf{Q})$$
$$\mathbf{v}_t \sim \mathcal{N}(0, \mathbf{R})$$
$$\mathbf{z}_0 \sim \mathcal{N}(\mu_0, \Sigma)$$

Inference: observe data, want to know the corresponding latents

Forward pass (Kalman filtering)

$$\mu_{i|i} = \mu_{i|i-1} + \mathbf{K}_i(\mathbf{x}_i - \mathbf{C}\mu_{i|i-1})$$
$$\Sigma_{i|i} = \Sigma_{i|i-1} - \mathbf{K}_i\mathbf{C}\Sigma_{i|i-1}$$
$$\mathbf{K}_i = \Sigma_{i|i-1}\mathbf{C}^{\mathrm{T}}(\mathbf{C}\Sigma_{i|i-1}\mathbf{C}^{\mathrm{T}} + R)^{-1}$$

Backward pass (Kalman smoothing)

$$\mu_{i|t} = \mu_{i|i} + \mathbf{F}_i(\mu_{i+1|t} - \mu_{i+1|i})$$
$$\Sigma_{i|t} = \mathbf{F}_i(\Sigma_{i+1|t} - \Sigma_{i+1|i})\mathbf{F}_i^{\mathrm{T}} + \Sigma_{i|$$
$$\mathbf{F}_i = \Sigma_{i|i}\mathbf{A}^{\mathrm{T}}\Sigma_{i+1|i}^{-1}$$

Note: check out kalmanderivation.pdf for complete derivation

How do we do learning?

General idea: find parameters that are most consistent with the observed data

The goal is to find the parameters that maximize the (log) likelihood

$$\mathcal{L}(\theta) = \log \mathrm{P}(\mathbf{x}_*|\theta)$$

Which we get after marginalizing out the latents

$$\mathrm{P}(\mathbf{x}_*|\theta) = \int \mathrm{P}(\mathbf{x}_*, \mathbf{z}_*|\theta)\mathbf{dz}$$

Learning with latent variables: expectation maximization

$$\log \int_z \mathrm{P}(\mathbf{x}, \mathbf{z}|\theta)\mathbf{dz} = \log \int_z Q(\mathbf{z}) \frac{\mathrm{P}(\mathbf{x},\mathbf{z}|\theta)}{Q(\mathbf{z})}\mathbf{dz} \qquad = \log\left(\mathbb{E}_Q\left[\frac{\mathrm{P}(\mathbf{x},\mathbf{z}|\theta)}{Q(\mathbf{z})}\right]\right)$$

via **Jensen's Inequality** $\geq \int_z Q(\mathbf{z}) \log \frac{\mathrm{P}(\mathbf{x},\mathbf{z}|\theta)}{Q(\mathbf{z})}\mathbf{dz}$ $= \mathbb{E}_Q\left[\log \frac{\mathrm{P}(\mathbf{x},\mathbf{z}|\theta)}{Q(\mathbf{z})}\right]$

$$= \int_Q Q(\mathbf{z}) \log \mathrm{P}(\mathbf{x}, \mathbf{z}|\theta))\mathbf{dz} - \int_Q Q(\mathbf{z}) \log Q(\mathbf{z})\mathbf{dz}$$

$$= \mathcal{F}(Q, \theta)$$ **Free energy**

EM alternates between finding a good approximation Q,
and then changing the parameters to improve the approx likelihood

This is done coordinate-wise: improve one keeping the other fixed, then swap.

$$\textbf{E step:} \qquad Q_{k+1} \leftarrow \underset{Q}{\arg\max} \;\; \mathcal{F}(Q, \theta_k)$$

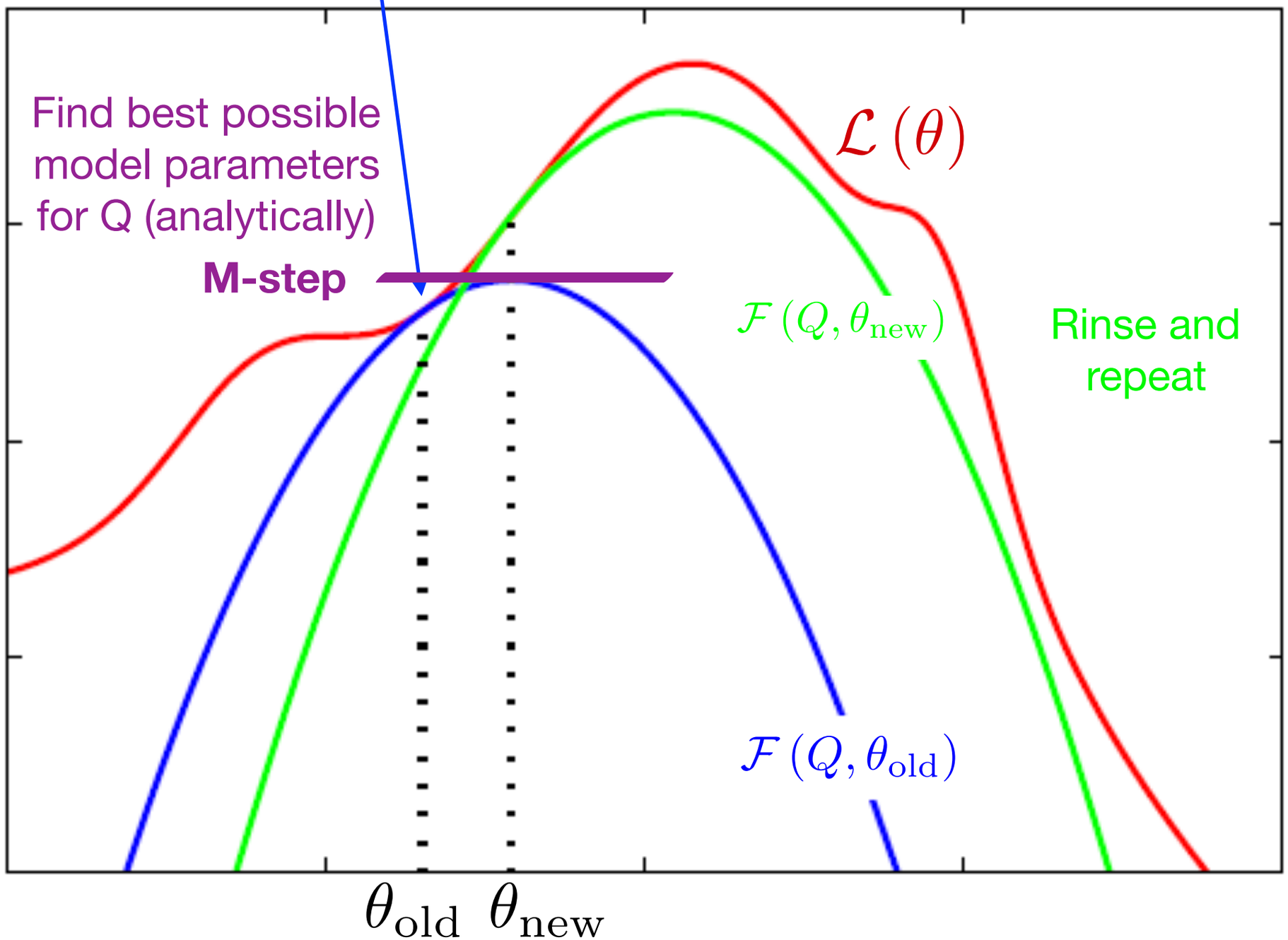
$$\textbf{M step:} \qquad \theta_{k+1} \leftarrow \underset{\theta}{\arg\max} \;\; \mathcal{F}(Q_{k+1}, \theta)$$

It is guaranteed that the likelihood will **never decrease** during this procedure.

General philosophy:
use smoothing estimates to guesstimate the state of the
latent variables, given current model parameters. Then use this fictitious
complete data to find new model parameters.

E-step: $Q_{k+1}(\mathbf{z}) = \mathrm{P}(\mathbf{z}|\mathbf{x}, \theta_k)$

The approximation is locally exact



Let's apply this idea to our LDS model

Log likelihood: $\mathcal{L}(\theta) = \log \mathrm{P}(\mathbf{x}_*|\theta)$

$$\mathrm{P}(\mathbf{x}_*, \mathbf{z}_*|\theta) = \mathrm{P}_\theta(\mathbf{z}_0) \prod_i \mathrm{P}_\theta(\mathbf{z}_{i+1}|\mathbf{z}_i) \prod_i \mathrm{P}_\theta(\mathbf{x}_i|\mathbf{z}_i)$$

To simplify notation we use shorthand $\theta = \{\mathbf{A},\, \mathbf{C},\, \mathbf{Q},\, \mathbf{R}, \mu_0, \Sigma_0\}$

For the lower bound we use the output of the Kaman smoother as an approximation of the joint distribution

What this means is that we don't consider the full joint dependencies, but only look at the posterior (joint) marginals for the latent variables **z**

$$\begin{aligned}
\mathbb{E}\left[\mathbf{z}_i \middle| \mathbf{x}_*\right] &= \mu_{i|t} \\
\mathbb{E}\left[\mathbf{z}_i \mathbf{z}_i \middle| \mathbf{x}_*\right] &= \Sigma_{i|t} + \mu_{i|t} \mu_{i|t}^{\mathrm{t}} \\
\mathbb{E}\left[\mathbf{z}_{i+1} \mathbf{z}_i \middle| \mathbf{x}_*\right] &= \Sigma_{i+1|t} \mathbf{F}_i + \mu_{i+1|t} \mu_{i|t}^{\mathrm{t}}
\end{aligned}$$

Note: I'll add the proof for the last bit to the kalmanderivation.

The complete data (log)likelihood is :

$$\log \mathrm{P}(\mathbf{x}_*, \mathbf{z}_*|\theta) = \log \mathrm{P}_\theta(\mathbf{z}_0) + \sum_i \log \mathrm{P}_\theta(\mathbf{z}_{i+1}|\mathbf{z}_i) + \sum_i \log \mathrm{P}_\theta(\mathbf{x}_i|\mathbf{z}_i)$$

$$= (\mathbf{z}_0 - \mu_o)^{\mathrm{t}} \Sigma^{-1} (\mathbf{z}_0 - \mu_o) - \tfrac{1}{2} \log |\Sigma|$$

$$+ \sum_i (\mathbf{z}_{i+1} - \mathbf{A}\mathbf{z}_i)^{\mathrm{t}} \mathbf{Q}^{-1} (\mathbf{z}_{i+1} - \mathbf{A}\mathbf{z}_i) - \tfrac{t}{2} |Q|$$

$$+ \sum_i (\mathbf{x}_i - \mathbf{C}\mathbf{z}_i)^{\mathrm{t}} \mathbf{R}^{-1} (\mathbf{x}_i - \mathbf{C}\mathbf{z}_i) - \tfrac{t}{2} |R|$$

$$+ \mathrm{const}$$

It’s just a sum of quadratic forms, and individual parameters show up in separate terms so they can be optimized individually

Expectation part

We define the auxiliary distribution *Q* by taking the expectation of the complete log with respect to the posterior distribution

$$\mathrm{P}(\mathbf{z}_*|\mathbf{x}_*, \theta_{\mathrm{old}})$$

$$\mathcal{F}(\theta, \theta_{\mathrm{old}}) = \mathbb{E}_{z|\theta_{old}} \log \mathrm{P}(\mathbf{x}_*, \mathbf{z}_*|\theta)$$

First we guestimate the latent values given current setting of parameters, then pretend we do actually know the latents and optimize for parameters

M step: optimize this function with respect to the components of \theta

1. Initial state z_0

$$Q(\theta, \theta_{\text{old}}) = -\tfrac{1}{2}\log|\Sigma_0| - \tfrac{1}{2}\mathbb{E}_{z|\theta_{old}}\left[(\mathbf{z}_1 - \mu_0)^{\text{t}}\Sigma_0^{-1}(\mathbf{z}_1 - \mu_0)\right] + \text{const}$$

const - absorbs all terms that do not depend on μ_0, Σ_0

This is just a gaussian so maximum likelihood solution is given by the empirical moments:

$$\mu_0^{\text{new}} = \mathbb{E}[\mathbf{z}_1]$$

$$\Sigma_0^{\text{new}} = \mathbb{E}[\mathbf{z}_1\mathbf{z}_1^{\text{t}}] + \mathbb{E}[\mathbf{z}_1]\mathbb{E}[\mathbf{z}_1]^t$$

2. latent dynamics

$$Q(\theta, \theta_{\text{old}}) = -\tfrac{t}{2}\log|\mathbf{Q}| - \mathbb{E}_{Z|\theta_{\text{old}}}\left[\tfrac{1}{2}\sum_i (\mathbf{z}_{i+1} - \mathbf{A}\mathbf{z}_i)^{\text{t}}\mathbf{Q}^{-1}(\mathbf{z}_{i+1} - \mathbf{A}\mathbf{z}_i)\right] + \text{const}$$

This has the maximum likelihood estimates:

$$\mathbf{A}^{\text{new}} = \left(\sum_i \mathbb{E}[\mathbf{z}_{i+1}\mathbf{z}_i^{\text{t}}]\right)\left(\sum_i \mathbb{E}[\mathbf{z}_i\mathbf{z}_i^{\text{t}}]\right)^{-1}$$

$$\begin{aligned}\mathbf{Q}^{\text{new}} = \tfrac{1}{t}\sum_i \big(& \mathbb{E}[\mathbf{z}_{i+1}\mathbf{z}_{i+1}^{\text{t}}] - \mathbf{A}^{\text{new}}\mathbb{E}[\mathbf{z}_i\mathbf{z}_{i+1}^{\text{t}}] \\ & - \mathbb{E}[\mathbf{z}_{i+1}\mathbf{z}_i^{\text{t}}]\mathbf{A}^{\text{new t}} + \mathbf{A}^{\text{new}}\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^{\text{t}}]\mathbf{A}^{\text{new t}}\big)\end{aligned}$$

*Using **multivariate linear regression**

model $\mathbf{y} = \mathbf{W}\mathbf{x} + \epsilon$

data $\left(\mathbf{x}^{(n)}, \mathbf{y}^{(n)}\right)$

ML estimates:

$$\hat{\mathbf{W}} = \left(\sum_n \mathbf{y}^{(n)}\mathbf{x}^{(n)\,\text{t}}\right)\left(\sum_n \mathbf{x}^{(n)}\mathbf{x}^{(n)\,\text{t}}\right)^{-1}$$

$$\hat{\Sigma} = \frac{1}{N}\sum_n \left(\mathbf{y}^{(n)} - \mathbf{W}\mathbf{x}^{(n)}\right)\left(\mathbf{y}^{(n)} - \mathbf{W}\mathbf{x}^{(n)}\right)^{\text{t}}$$

Note: check this at home!

3. observation model

$$Q(\theta, \theta_{\text{old}}) = -\tfrac{t}{2} \log |\mathbf{R}| - \mathbb{E}_{z|\theta_{\text{old}}} \left[\tfrac{1}{2} \sum_i (\mathbf{x}_i - \mathbf{C}\mathbf{z}_i)^{\text{t}} \mathbf{R}^{-1} (\mathbf{x}_i - \mathbf{C}\mathbf{z}_i) \right] + \text{const}$$

This has the maximum likelihood parameters estimates:

$$\mathbf{C}^{\text{new}} = \left(\sum_i \mathbf{x}_i \mathbb{E}[\mathbf{z}_i^{\text{t}}] \right) \left(\sum_i \mathbb{E}[\mathbf{z}_i \mathbf{z}_i^{\text{t}}] \right)^{-1}$$

$$\mathbf{R}^{\text{new}} = \tfrac{1}{t} \sum_i \left(\mathbf{x}_i \mathbf{x}_i^{\text{t}} - \mathbf{C}^{\text{new t}} \mathbb{E}[\mathbf{z}_i] \mathbf{x}_i^{\text{t}} - \mathbf{x}_i \mathbb{E}[\mathbf{z}_i^{\text{t}}] \mathbf{C}^{\text{new}} + \mathbf{C}^{\text{new t}} \mathbb{E}[\mathbf{z}_i \mathbf{z}_i^{\text{t}}] \mathbf{C}^{\text{new}} \right)$$

Note: we'll see some practical use of this in the lab.

Particle filtering

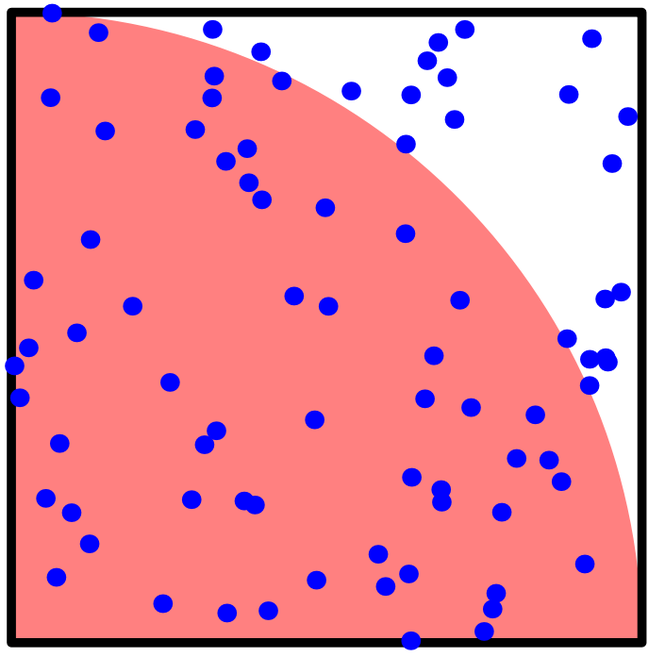
What if we can't compute the distributions exactly?

When the posterior distribution is too complicated to write out in closed form,
or to compute marginals and expectations, we can use sampling based approximate techniques

$$\mathbf{z}^{(k)} \sim \mathrm{P}(\mathbf{z}|\mathbf{x}_*, \theta)$$

$$\mu_{z_i} = \frac{1}{K} \sum z_i^{(k)}$$

Example: approximating pi



$$P(x,y) = \begin{cases} 1 & 0<x<1 \quad \text{and} \quad 0<y<1 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi = 4 \iint \mathbb{I}\left((x^2+y^2)<1\right) P(x,y)\ \mathrm{d}x\ \mathrm{d}y$$

In general:

$$\int f(x)P(x)\,\mathrm{d}x \;\approx\; \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)}), \quad x^{(s)} \sim P(x)$$

Example: making predictions

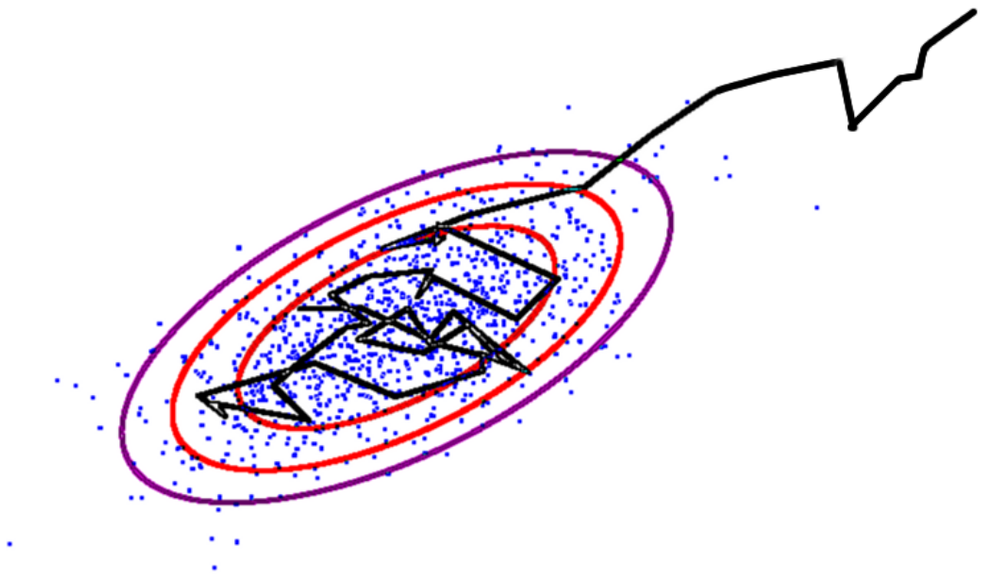
$$\begin{aligned} p(x|\mathcal{D}) &= \int P(x|\theta,\mathcal{D})P(\theta|\mathcal{D})\,\mathrm{d}\theta \\ &\approx \frac{1}{S}\sum_{s=1}^{S} P(x|\theta^{(s)},\mathcal{D}), \quad \theta^{(s)} \sim P(\theta|\mathcal{D}) \end{aligned}$$

Also during learning: E-step in EM

How do we construct stochastic dynamics that do the 'right' thing (Metropolis Hastings)?

Construct a biased random walk that explores target dist $P^\star(x)$

Markov steps, $x_t \sim T(x_t \leftarrow x_{t-1})$



MCMC gives approximate, correlated samples from $P^\star(x)$

- Propose a move from the current state $Q(x'; x)$, e.g. $\mathcal{N}(x, \sigma^2)$
- Accept with probability $\min\left(1, \frac{P(x')Q(x;x')}{P(x)Q(x';x)}\right)$
- Otherwise next state in chain is a copy of current state

http://videolectures.net/mlss09uk_murray_mcmc/

Making things more efficient: importance sampling

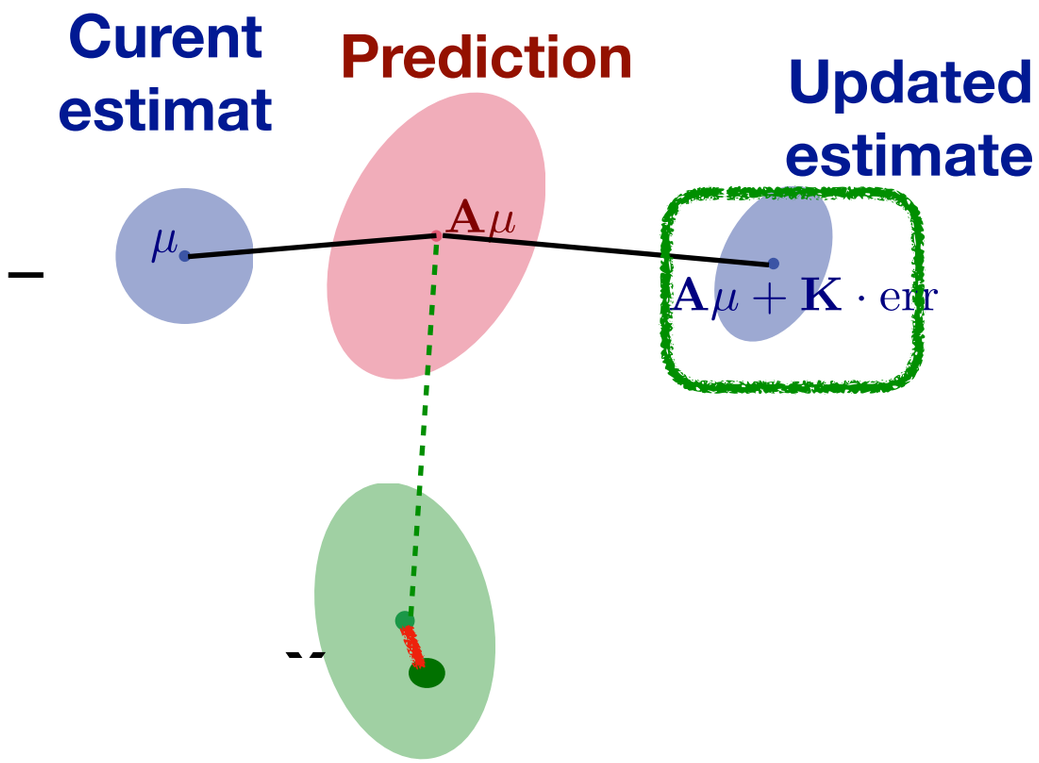
Sometimes we can use samples even
if they don't come from the right distribution
(but are cheap to generate)

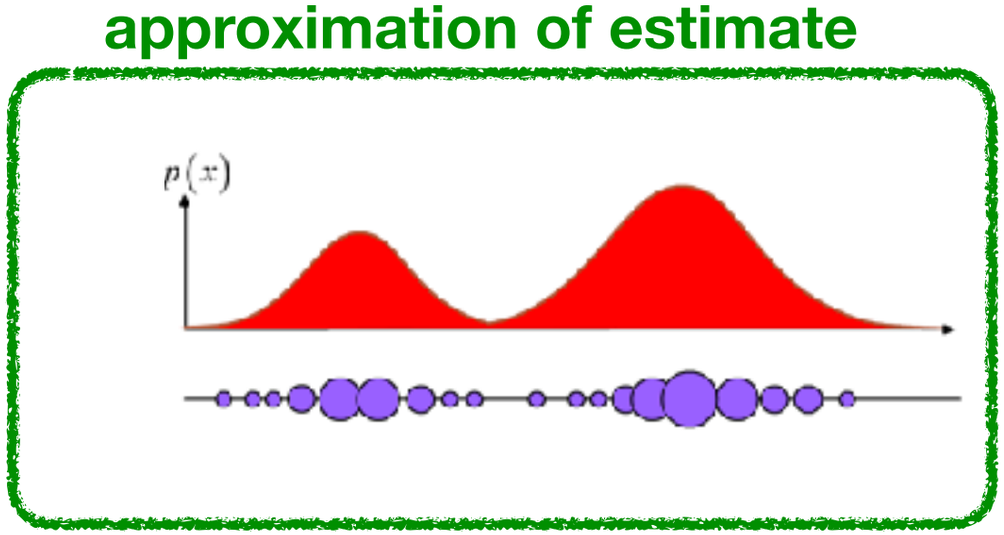
$$\int f(x)P(x)\,\mathrm{d}x = \int f(x)\frac{P(x)}{Q(x)}Q(x)\,\mathrm{d}x, \qquad (Q(x) > 0 \text{ if } P(x) > 0)$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)})\frac{P(x^{(s)})}{Q(x^{(s)})}, \quad x^{(s)} \sim Q(x)$$

Can be quite efficient if Q a decent approximation of P

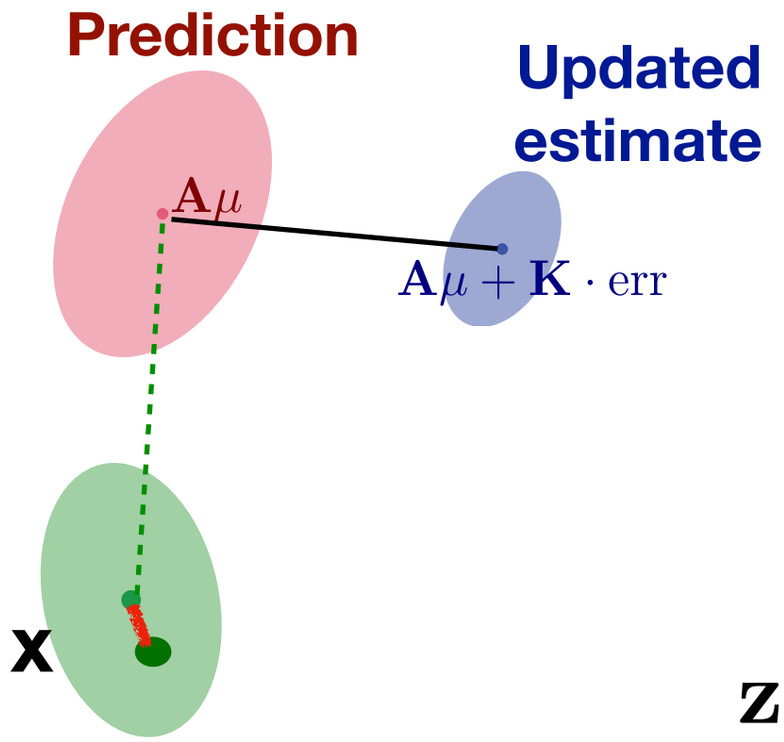
Particle filtering





Main idea: represent distributions as a collection of weighted samples manipulate this samples to change corresponding posterior as we run through the data

Revisiting Kalman filtering steps



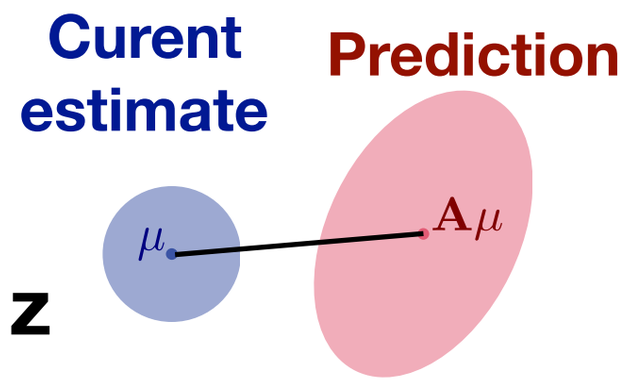
$$z_i^{(k)} \sim \mathrm{P}(\mathbf{z}_i | \mathbf{x}_{1:i})$$

$$\mathbf{z}_i^{(k)} \sim \mathrm{P}(\mathbf{z}_i | x_{1:i-1})$$

$$w_i^{(k)} = \frac{\mathrm{P}(\mathbf{x}_i | z_i^{(k)})}{\sum_l \mathrm{P}(\mathbf{x}_i | z_i^{(l)})}$$

$$0 \leq w_i^{(k)} \leq 1 \quad \sum_k w_i^{(k)} = 1$$

Samples in, samples out



$$\mathrm{P}(\mathbf{z}_{i+1}|\mathbf{x}_{1:i}) \approx \sum_k w_i^{(k)} \mathrm{P}(\mathbf{z}_{i+1}|\mathbf{z}_i^{(k)})$$

this is a mixture distribution

To sample from this:

draw mixture component according to w,
then draw from corresponding component.

Note: Bishop figure is nicer…