DS-GA 1018.001

Probabilistic time series analysis

Lecture 5

Latent state model: EM; Particle filtering

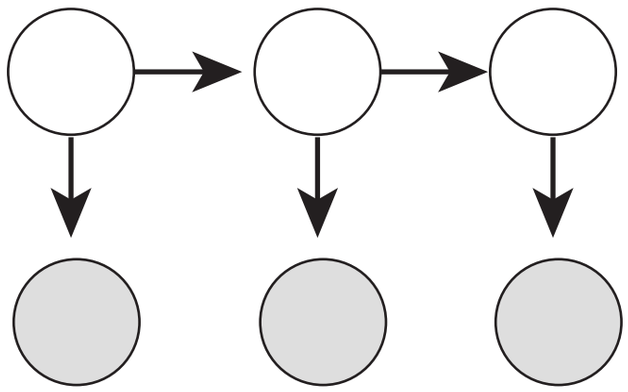
Instructor: Cristina Savin

NYU, CNS & CDS

Latent state models

In the latent space the temporal dependencies are simple:

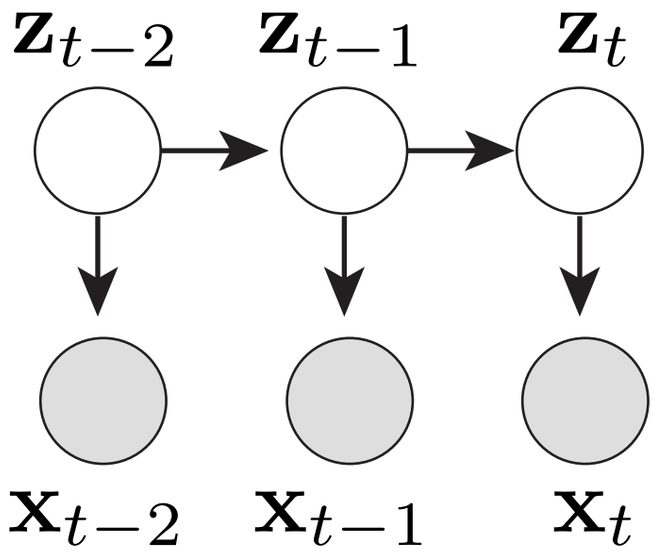
1st order Markov



Goal: find a **latent** variable
that summarizes
the relevant **history**
while keeping math simple

The latent variable is either discrete for **hidden Markov models** (HMMs) or continuous for **latent dynamical systems** (LDS)

Latent state models



d-dimensional

n-dimensional

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

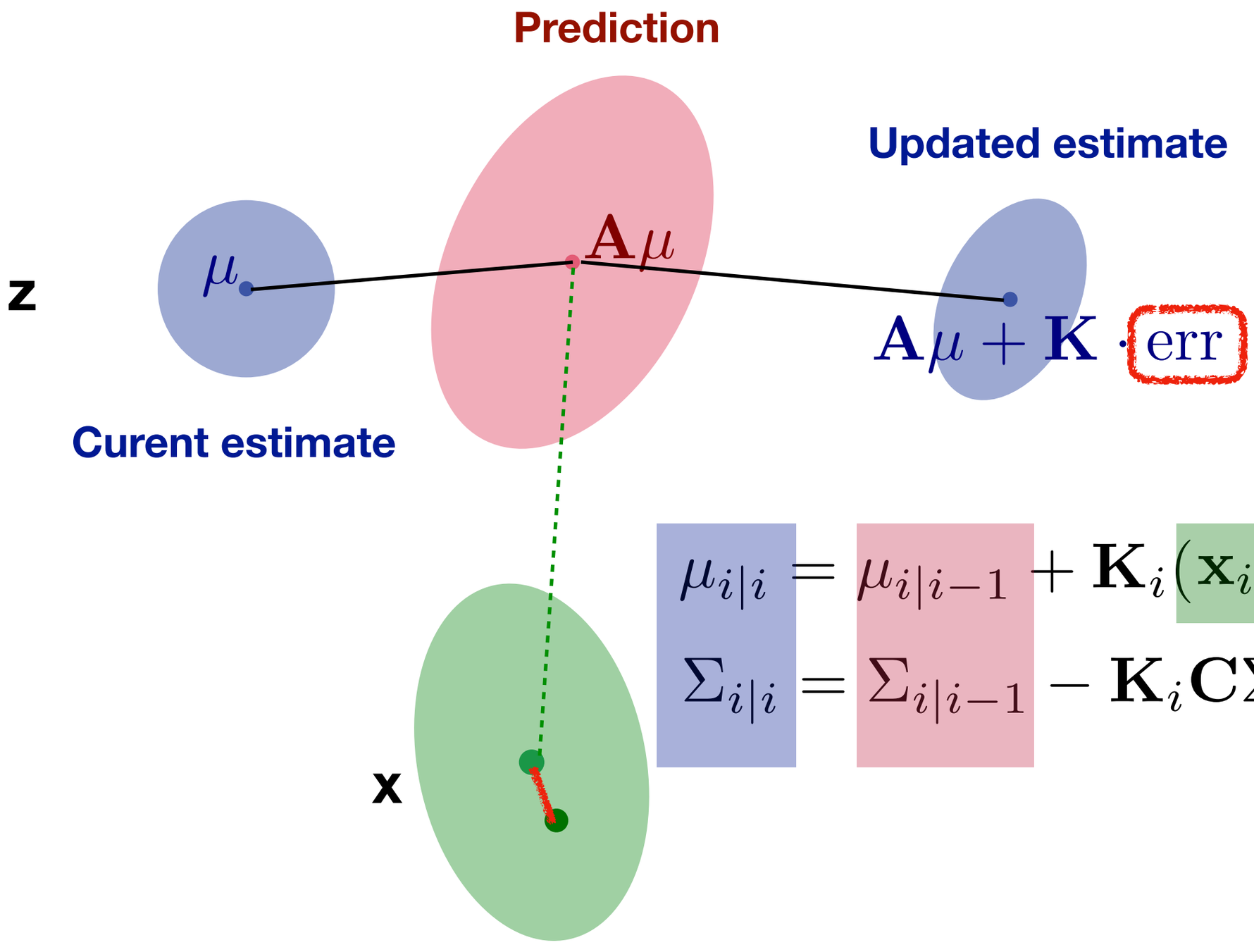
$$\mathbf{x}_t = \mathbf{C}\mathbf{z}_t + \mathbf{v}_t$$

Where the noise terms are iid Gaussian:

$$\mathbf{w}_t \sim \mathcal{N}(0, \mathbf{Q})$$

$$\mathbf{v}_t \sim \mathcal{N}(0, \mathbf{R})$$

$$\mathbf{z}_0 \sim \mathcal{N}(\mu_0, \Sigma)$$



$$\mu_{i|i} = \mu_{i|i-1} + \mathbf{K}_i(\mathbf{x}_i - \mathbf{C}\mu_{i|i-1})$$

$$\Sigma_{i|i} = \Sigma_{i|i-1} - \mathbf{K}_i\mathbf{C}\Sigma_{i|i-1}$$

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

Smoothing - backward sweep

Same principle, but now we propagate information backwards
in time, using the estimates we have already
(we no longer need the data)

$$\mu_{i|t} = \mu_{i|i} + \mathbf{F}_i(\mu_{i+1|t} - \mu_{i+1|i})$$

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

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

How do we learn the parameters?

Estimating parameters via expectation maximization (EM)

Find parameters that are most consistent with the observed data

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

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

For that we need to **marginalize out the latents**

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

EM provides a general *framework* to do this,
we'll apply it for multiple models in the context of the class

General idea of EM

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

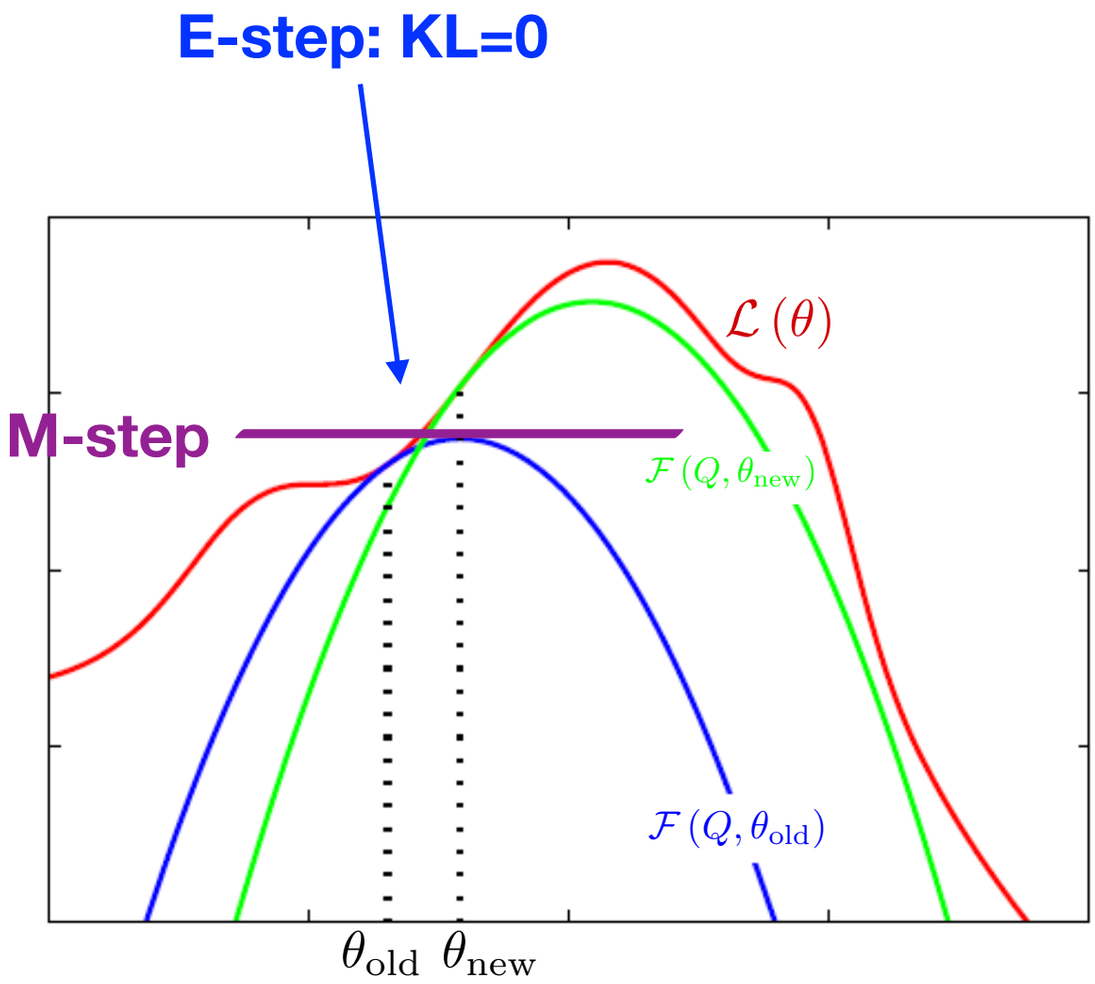
via **Jensen's Inequality** $$\geq \int_z Q(\mathbf{z}) \log \frac{\mathrm{P}(\mathbf{x},\mathbf{z}|\theta)}{Q(\mathbf{z})}\mathbf{dz} \qquad = \mathbb{E}_Q\left[\log \frac{\mathrm{P}(\mathbf{x},\mathbf{z}|\theta)}{Q(\mathbf{z})}\right]$$

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

Big picture: log of integral is not nice, so we replace it with bound on integral of log

A slightly different way of bounding the log-likelihood

$$
\begin{aligned}
\log \mathrm{P}\left(\mathbf{x}|\theta\right) &= \log \mathrm{P}\left(\mathbf{x}|\theta\right) \int_Z q(\mathbf{z}) \mathbf{dz} \\
&= \int_Z q(\mathbf{z}) \log \mathrm{P}\left(\mathbf{x}|\theta\right) \mathbf{dz} \\
&= \int_Z q(\mathbf{z}) \left(\log \mathrm{P}\left(\mathbf{x}, \mathbf{z}|\theta\right) - \log \mathrm{P}\left(\mathbf{z}|\mathbf{x}, \theta\right) \right) \mathbf{dz} \\
&= \int_Z q(\mathbf{z}) \left(\log \mathrm{P}\left(\mathbf{x}, \mathbf{z}|\theta\right) - \log q(\mathbf{z}) + \log q(\mathbf{z}) - \log \mathrm{P}\left(\mathbf{z}|\mathbf{x}, \theta\right) \right) \mathbf{dz} \\
&= \int_Z q(\mathbf{z}) \log \frac{\mathrm{P}\left(\mathbf{x}, \mathbf{z}|\theta\right)}{q(\mathbf{z})} \mathbf{dz} - \int_Z q(\mathbf{z}) \log \frac{\mathrm{P}\left(\mathbf{z}|\mathbf{x}, \theta\right)}{q(\mathbf{z})} \mathbf{dz} \\
&= \mathcal{L}(q, \theta) + \underbrace{\mathrm{KL}\left(q(\mathbf{z}) || \mathrm{P}\left(\mathbf{z}|\mathbf{x}, \theta\right)\right)}_{\geq 0}
\end{aligned}
$$



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.

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

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

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

Let's apply this idea to our LDS model

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

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

We use the output of the Kaman smoother for q.

*shorthand: $\theta = \{\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}, \mu_0, \Sigma_0\}$

E-step approximation:

instead of full joint dependencies,

we only use posterior (joint) marginals

$$\mathbb{E}[\mathbf{z}_i|\mathbf{x}_*] = \mu_{i|t}$$

$$\mathbb{E}[\mathbf{z}_i\mathbf{z}_i^{\mathrm{t}}|\mathbf{x}_*] = \Sigma_{i|t} + \mu_{i|t}\mu_{i|t}^{\mathrm{t}}$$

$$\mathbb{E}[\mathbf{z}_{i+1}\mathbf{z}_i^{\mathrm{t}}|\mathbf{x}_*] = \Sigma_{i+1|t}\mathbf{F}_i + \mu_{i+1|t}\mu_{i|t}^{\mathrm{t}}$$

The complete data (log)likelihood is :

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

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

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

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

$$+ \mathrm{const}$$

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

Trick:

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

M step: initial condition

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

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:

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

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

Detour: simple linear regression reminder

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

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

ML parameter estimates:

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

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

Note: check at home, derivation in handout.

M step: 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:

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

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

M step: observations

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

This has the maximum likelihood parameters estimates:

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

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

Dealing with non-gaussian observations:

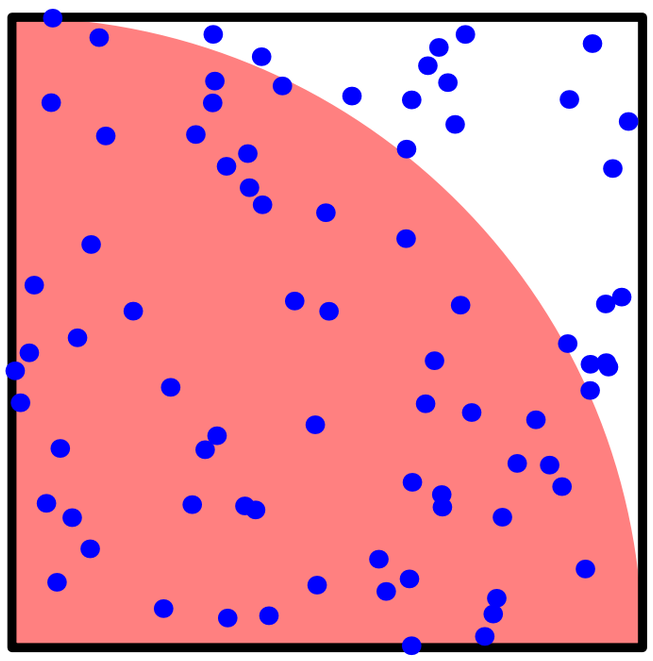
Particle filtering

Basic idea: sampling

If the true distribution is too complicated to deal with analytically
we can do (approximate) inference as long as we can generate **samples** from it.

$$\mathbf{z}^{(k)} \sim \mathrm{P}(\mathbf{z}|\mathbf{x}_*, \theta) \qquad \mu_{z_i} = \tfrac{1}{K} \sum z_i^{(k)}$$

Example: computing 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, most bayesian computations are ultimately about computing expectations under some distribution

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

Example: taking into account parameter uncertainty for **predictions**

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

Importance sampling:

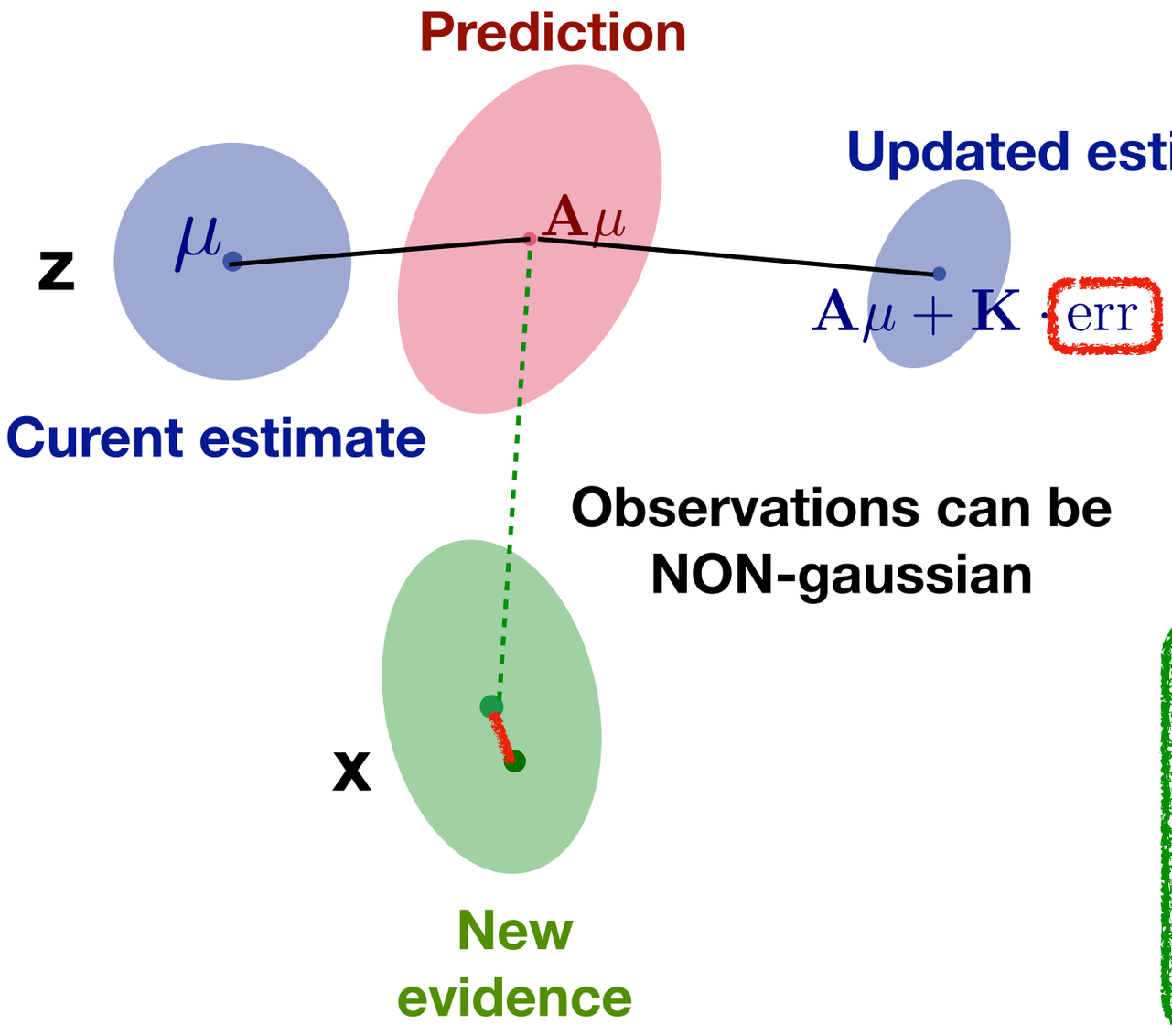
Basic idea: if sampling from the right distribution is not convenient

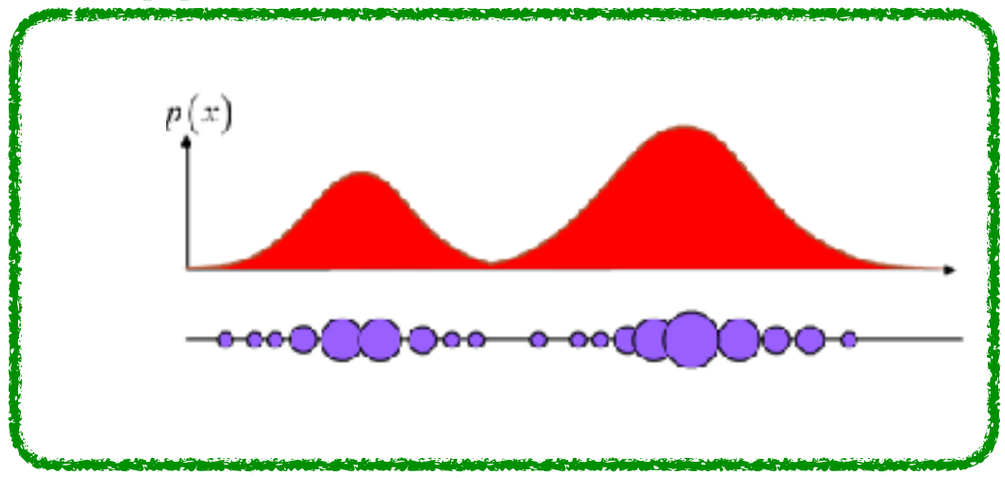
we can **use samples from the wrong distribution**
(but that are cheap to generate) and **correct for the difference**.

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

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

Importance weights

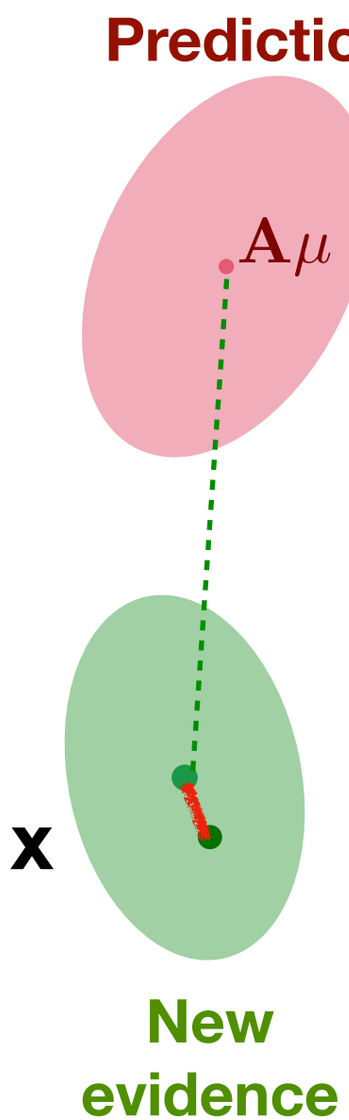




Main idea: represent marginals as a collection of weighted samples

manipulate this samples to approximate kalman-filtering

Incorporate evidence



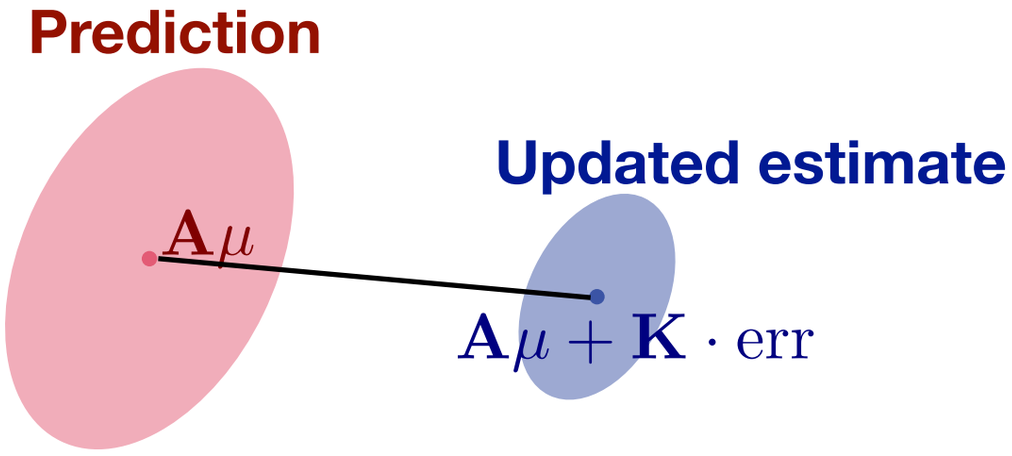
Given samples from the prior

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

The posterior can be approximated Using importance sampling:

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

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



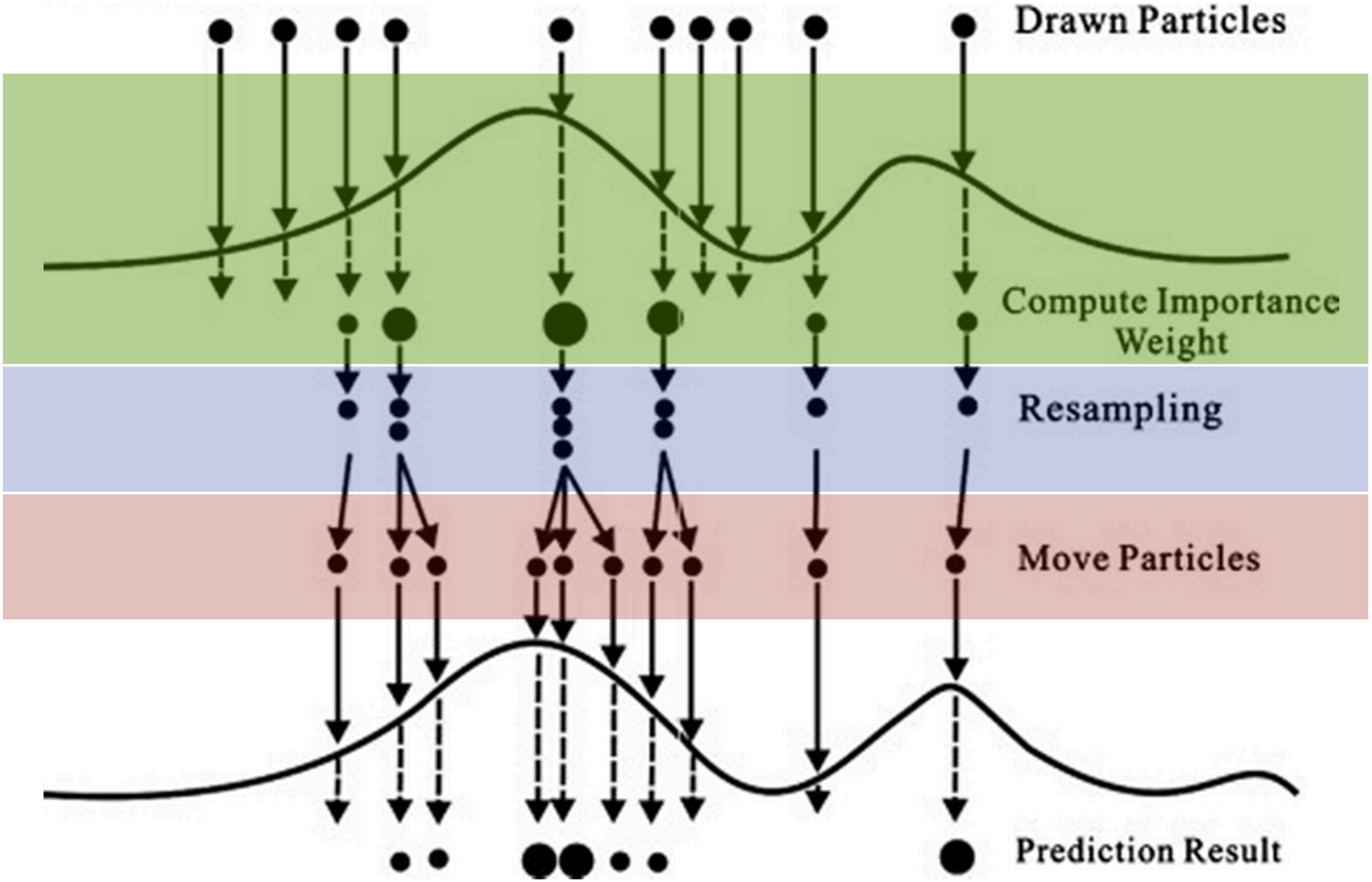
$$\mathrm{P}(\mathbf{z}_{i+1}|\mathbf{x}_{1:i}) \approx \sum_k w_i^{(k)} \mathrm{P}(\mathbf{z}_{i+1}|\mathbf{z}_i^{(k)})$$

mixture distribution

2-step sampling procedure:

Sample **class** according to w

Sample **observations** conditioning on class


Drawn Particles
Compute Importance Weight
Resampling
Move Particles
Prediction Result