

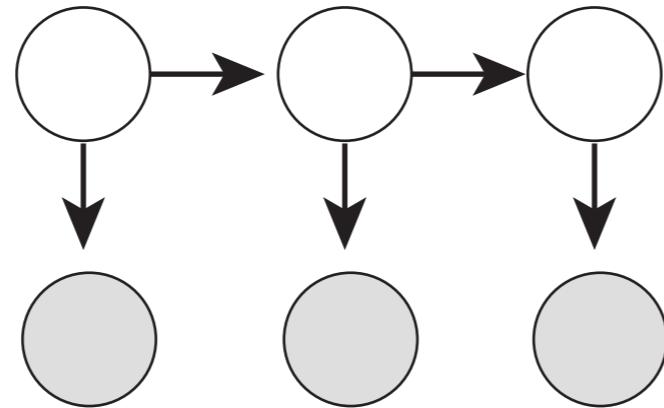
DS-GA 3001.001
Probabilistic time series analysis
Lecture 3
Latent state model: Kalman filtering

Instructor: Cristina Savin
NYU, CNS & CDS

Latent state models

In the latent space the temporal dependencies are simple:

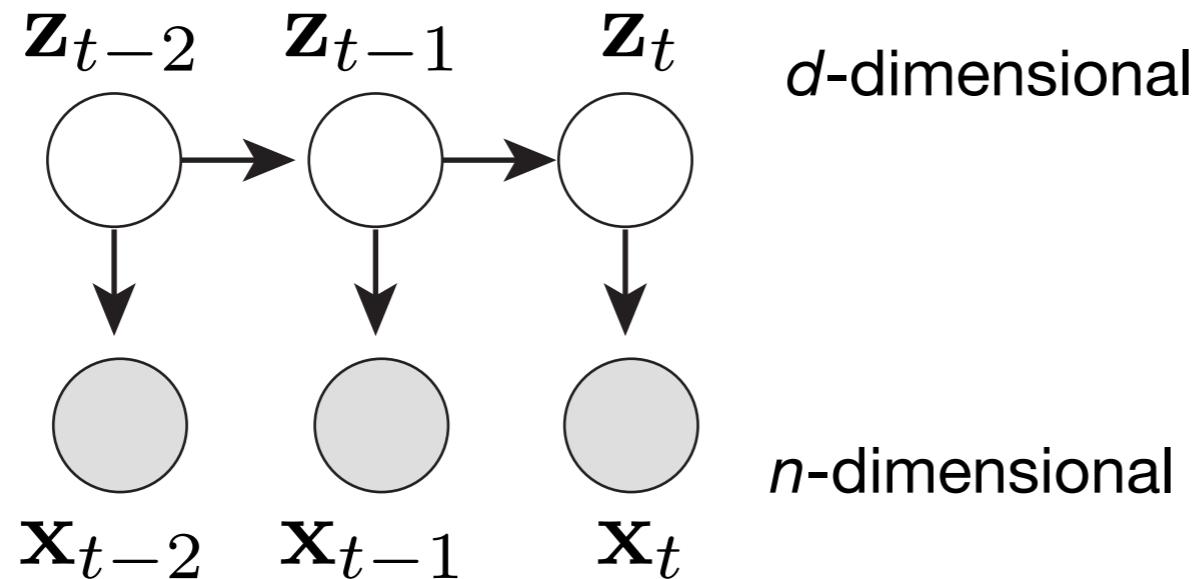
1st order Markov



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



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

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

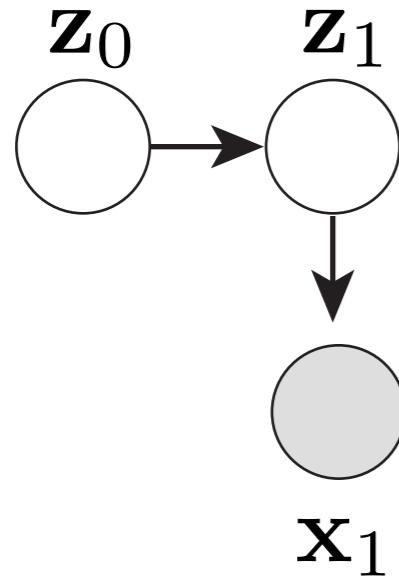
Where the noise terms
are iid Gaussian:

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

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

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

Intuitively, we can construct the forward pass by induction



2 sub-steps:

$$z_1 | z_0$$

Forecasting

$$\mu_{1|0} = A\mu_0$$

$$\Sigma_{1|0} = A\Sigma_0 A^T + Q$$

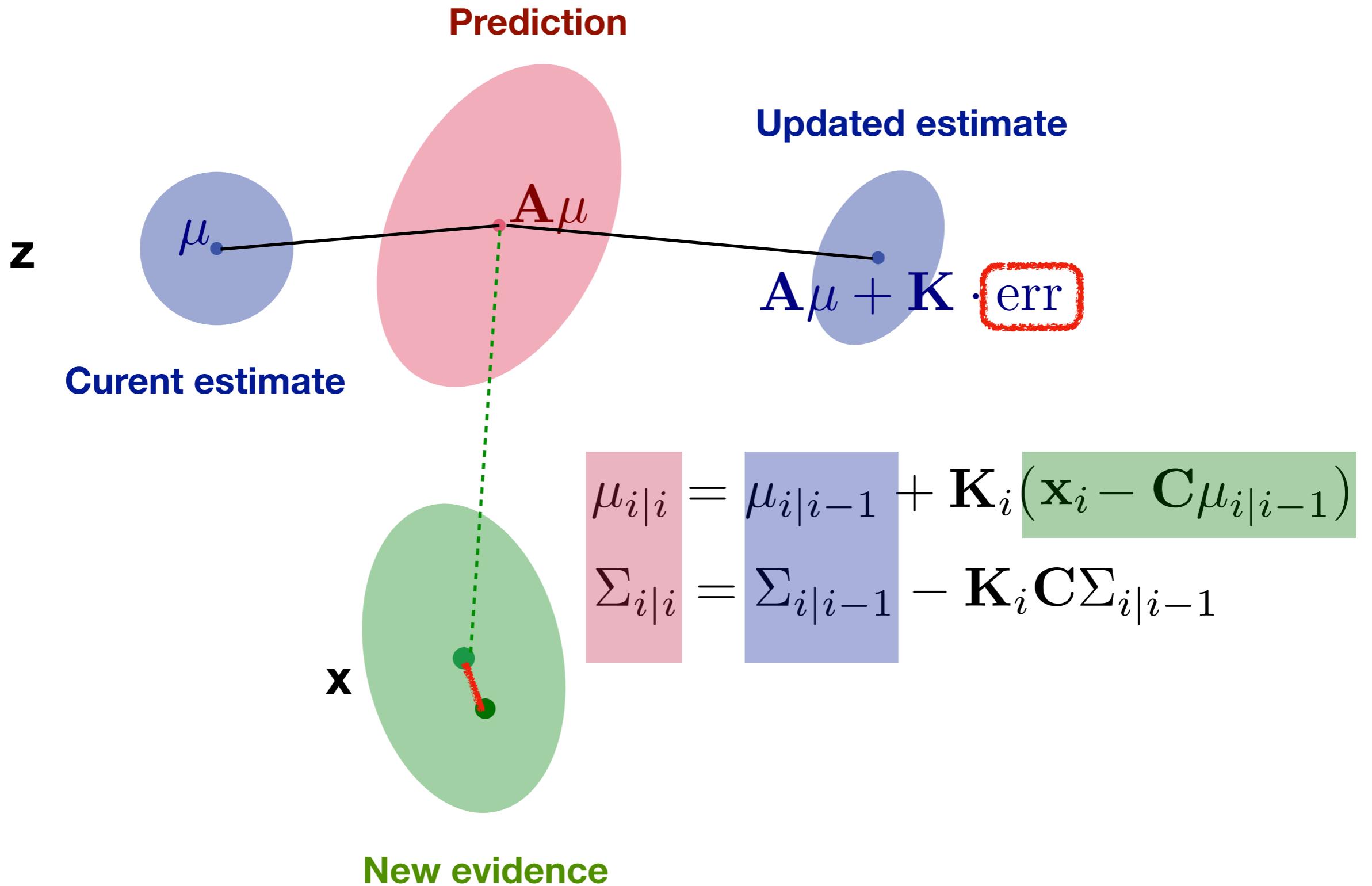
$$z_1 | x_1, z_0$$

Combine with new evidence

$$\mu_{1|1} = \mu_{1|0} + K(x_1 - C\mu_{1|0})$$

$$\Sigma_{1|1} = \Sigma_{1|0} - K C \Sigma_{1|0}$$

$$K = \Sigma_{1|0} C^T (C \Sigma_{1|0} C^T + R)^{-1}$$



$$K_i = \Sigma_{i|i-1} C^T (C \Sigma_{i|i-1} C^T + R)^{-1}$$

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)

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

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

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

How do we use learn the parameters?

Estimating parameters via expectation maximization

Find parameters that are most consistent with the observed data

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

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

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

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

Bounding the log-likelihood

$$\begin{aligned}\log P(\mathbf{x}|\theta) &= \log P(\mathbf{x}|\theta) \int_Z q(\mathbf{z}) d\mathbf{z} \\ &= \int_Z q(\mathbf{z}) \log P(\mathbf{x}|\theta) d\mathbf{z} \\ &= \int_Z q(\mathbf{z}) (\log P(\mathbf{x}, \mathbf{z}|\theta) - \log P(\mathbf{z}|\mathbf{x}, \theta)) d\mathbf{z} \\ &= \int_Z q(\mathbf{z}) (\log P(\mathbf{x}, \mathbf{z}|\theta) - \log q(\mathbf{z}) + \log q(\mathbf{z}) - \log P(\mathbf{z}|\mathbf{x}, \theta)) d\mathbf{z} \\ &= \int_Z q(\mathbf{z}) \log \frac{P(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z})} d\mathbf{z} - \int_Z q(\mathbf{z}) \log \frac{P(\mathbf{z}|\mathbf{x}, \theta)}{q(\mathbf{z})} d\mathbf{z} \\ &= \mathcal{L}(q, \theta) + \text{KL}(q(\mathbf{z}) || P(\mathbf{z}|\mathbf{x}, \theta)) \geq 0\end{aligned}$$

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.

E step:

$$Q_{k+1} \leftarrow \arg \max_Q \mathcal{F}(Q, \theta_k)$$

M step:

$$\theta_{k+1} \leftarrow \arg \max_\theta \mathcal{F}(Q_{k+1}, \theta)$$

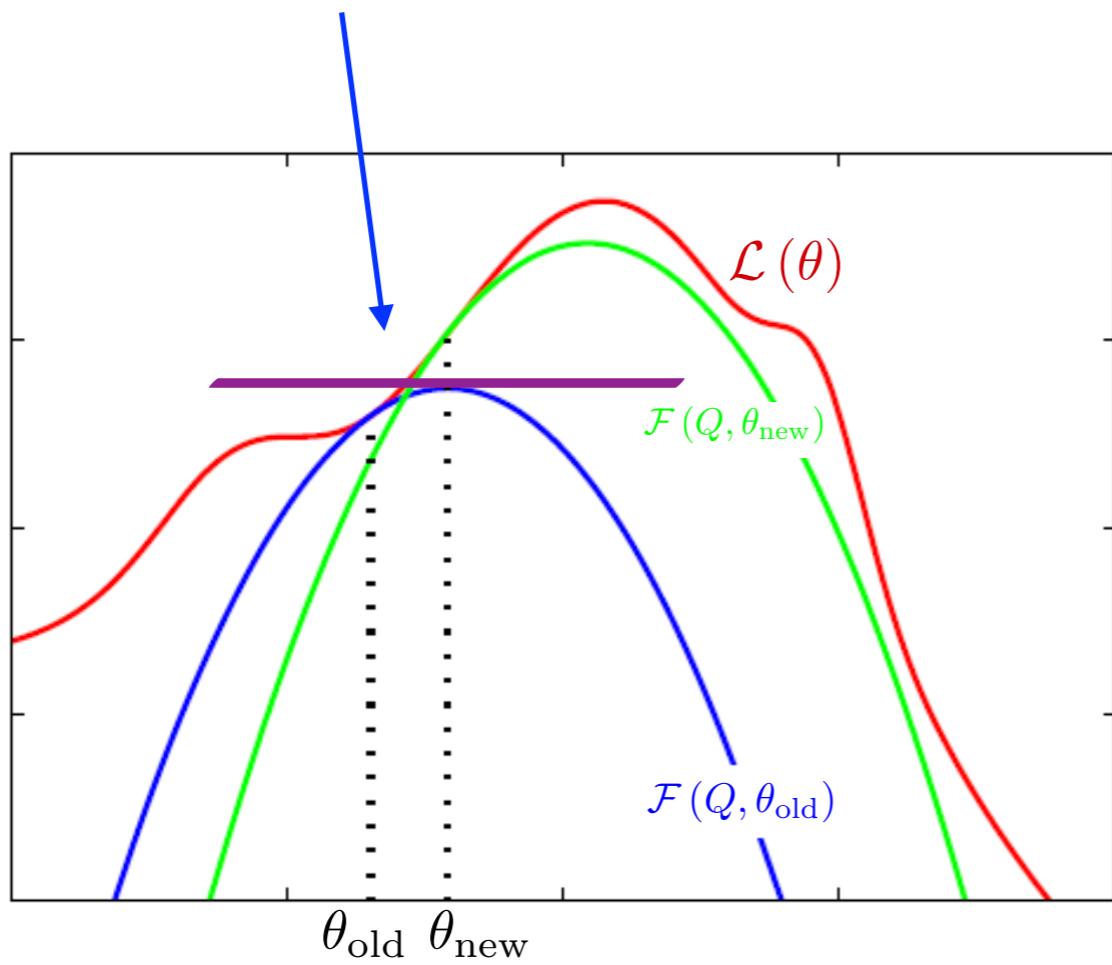
It is guaranteed that the likelihood will **never decrease** during this procedure.

General philosophy:

use simple parametric form 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: KL=0

M-step



\mathbf{q}^*

$$\begin{aligned}
 \mathcal{L}(q, \theta) &= \int_Z P(\mathbf{z}|\mathbf{x}, \theta^{\text{old}}) \log \frac{P(\mathbf{x}, \mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{x}, \theta^{\text{old}})} d\mathbf{z} \\
 &= \int_Z P(\mathbf{z}|\mathbf{x}, \theta^{\text{old}}) \log P(\mathbf{x}, \mathbf{z}|\theta) - \int_Z P(\mathbf{z}|\mathbf{x}, \theta^{\text{old}}) \log P(\mathbf{z}|\mathbf{x}, \theta^{\text{old}}) \\
 &= \mathcal{Q}(\theta, \theta^{\text{old}}) + \text{const}
 \end{aligned}$$

Let's apply this idea to our LDS model

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

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

We use the output of the Kaman smoother as an approximation of the joint distribution.

*shorthand: $\theta = \{\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\}$

E-step approximation:

instead of full joint dependencies,
we only use posterior (joint) marginals

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

$$\mathbb{E}[\mathbf{z}_i \mathbf{z}_i^T | \mathbf{x}_*] = \boldsymbol{\Sigma}_{i|t} + \boldsymbol{\mu}_{i|t} \boldsymbol{\mu}_{i|t}^T$$

$$\mathbb{E}[\mathbf{z}_{i+1} \mathbf{z}_i | \mathbf{x}_*] = \boldsymbol{\Sigma}_{i+1|t} \mathbf{F}_i + \boldsymbol{\mu}_{i+1|t} \boldsymbol{\mu}_{i|t}^T$$

The complete data (log)likelihood is :

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

$$= (\mathbf{z}_0 - \boldsymbol{\mu}_o)^t \boldsymbol{\Sigma}^{-1} (\mathbf{z}_0 - \boldsymbol{\mu}_o) - \frac{1}{2} \log |\boldsymbol{\Sigma}| \quad \text{Initial condition}$$

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

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

+const

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

M step: initial condition

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

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:

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

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

Detour: simple linear regression reminder

Linear model $\mathbf{y} = \mathbf{W}\mathbf{x} + \epsilon$ data $(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})$

ML parameter estimates:

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

Note: check at home, derivation in handout.

M step: latent dynamics

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

This has the maximum likelihood estimates:

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

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

M step: observations

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

This has the maximum likelihood parameters estimates:

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

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

Dealing with non-gaussian observations: Particle filtering

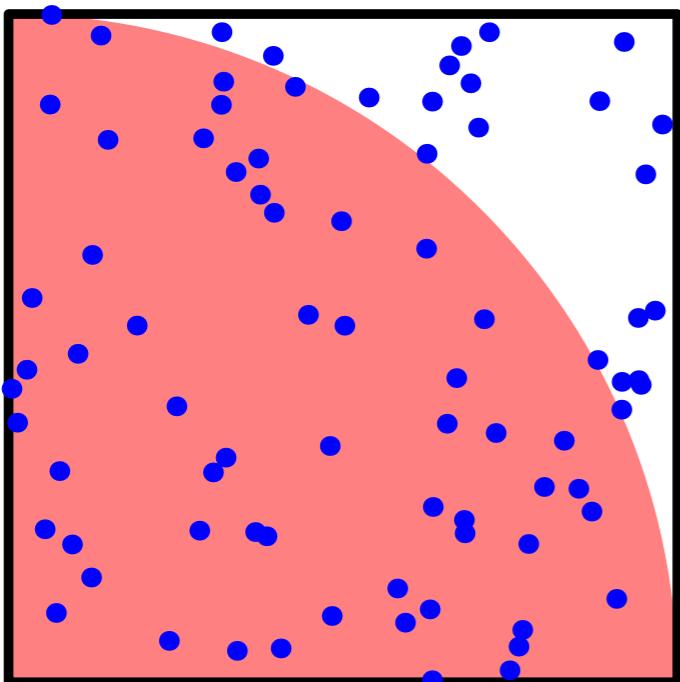
Basic idea: sampling

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.

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

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

Example: computing pi



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

$$\pi = 4 \iint \mathbb{I}((x^2 + y^2) < 1) P(x, y) \, dx \, dy$$

In general, most bayesian computations are ultimately about computing expectations under some distribution

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

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

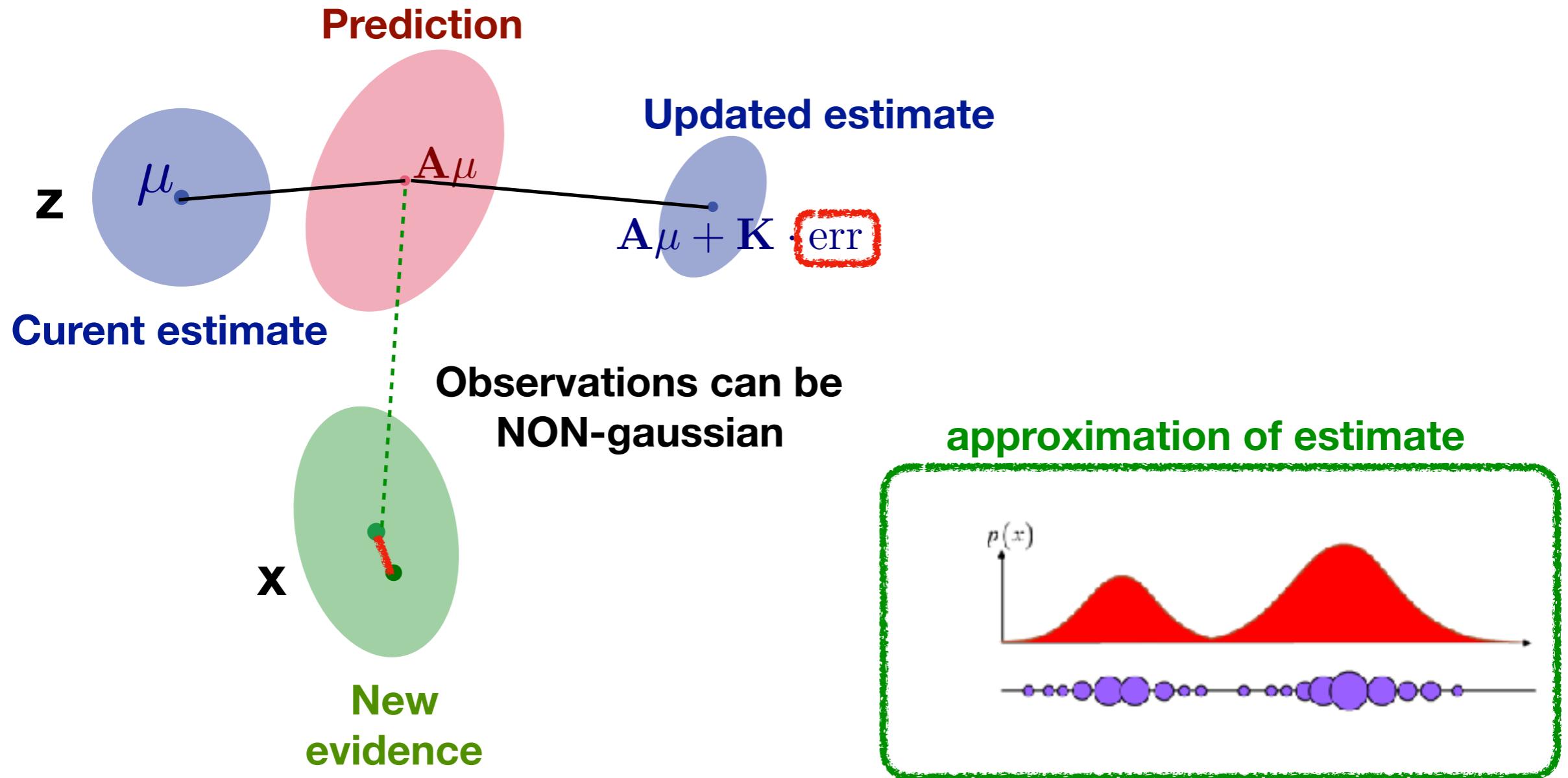
$$\begin{aligned} p(x|\mathcal{D}) &= \int P(x|\theta, \mathcal{D}) P(\theta|\mathcal{D}) d\theta \\ &\approx \frac{1}{S} \sum_{s=1}^S P(x|\theta^{(s)}, \mathcal{D}), \quad \theta^{(s)} \sim P(\theta|\mathcal{D}) \end{aligned}$$

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**.

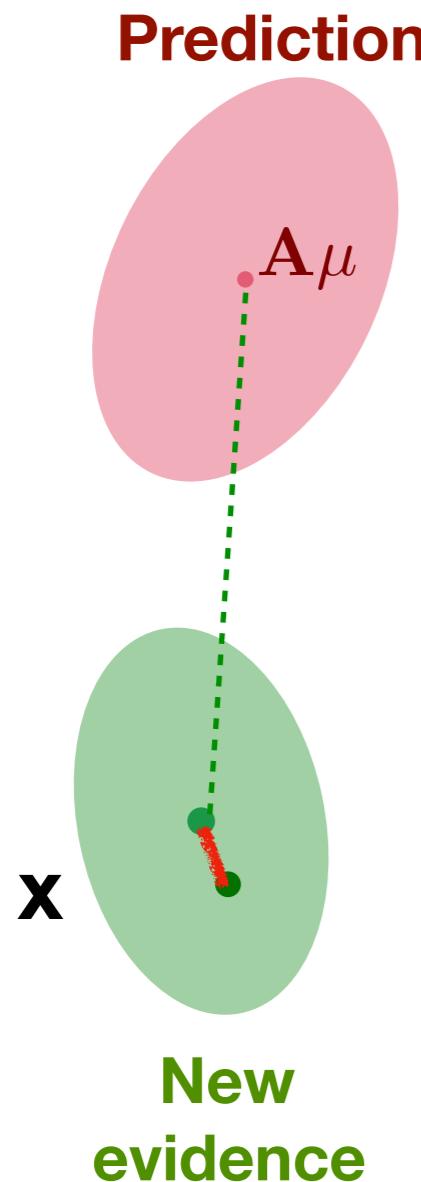
$$\begin{aligned} \int f(x)P(x) dx &= \int f(x) \frac{P(x)}{Q(x)} Q(x) dx, \quad (Q(x) > 0 \text{ if } P(x) > 0) \\ &\approx \frac{1}{S} \sum_{s=1}^S f(x^{(s)}) \frac{P(x^{(s)})}{Q(x^{(s)})}, \quad x^{(s)} \sim Q(x) \end{aligned}$$

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

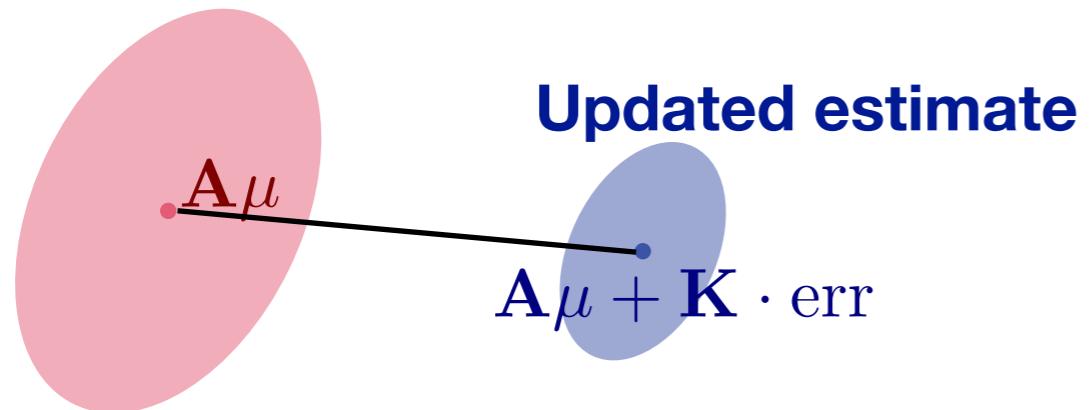
$$\mathbf{z}_i^{(k)} \sim P(\mathbf{z}_i | x_{1:i-1})$$

**The posterior can be approximated
Using importance sampling:**

$$w_i^{(k)} = \frac{P(x_i | z_i^{(k)})}{\sum_l P(x_i | z_i^{(l)})}$$

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

Prediction



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

mixture distribution

2-step sampling procedure:

Sample **class** according to w

Sample **observations** conditioning on class

