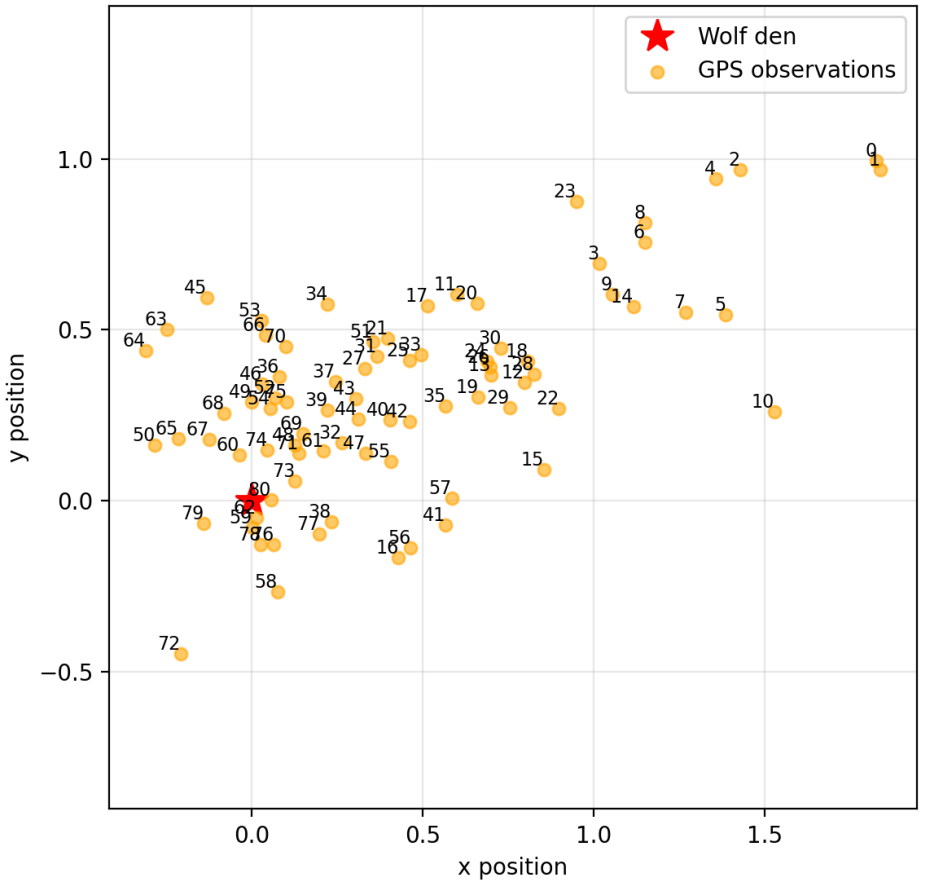
Foundations of Data Science

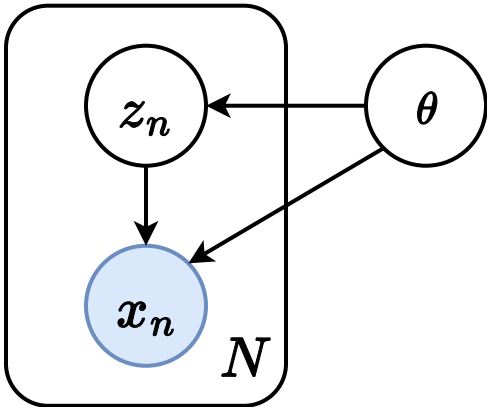
Lecture 5: State-space models

Prof. Gilles Louppe
g.louppe@uliege.be



Today's case study: **tracking** the location of a wolf over time using noisy GPS observations.

Discrete-time models



Static latent variable models

We previously defined latent variable models as probabilistic models that explain observed data x in terms of unobserved (latent) variables z and parameters θ,

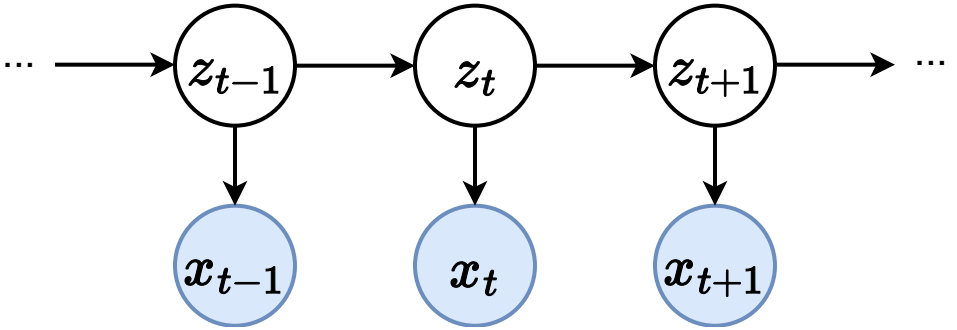
$$p(x, z, \theta) = p(x|z, \theta)p(z|\theta)p(\theta).$$

What if the system evolves over time and we have a sequence of observations $x_{1:T} = (x_1, \ldots, x_T)$ collected at discrete time steps $t = 1, \ldots, T$?

State-space models

To model a dynamical system, we can first assume a discretization of time and introduce a sequence of latent variables z_t that represent the state of the system at each time step t.

In this context, a **state-space model** is a latent variable model that explains a sequence $x_{1:T} = (x_1, \ldots, x_T)$ of observations in terms of a sequence $z_{1:T} = (z_1, \ldots, z_T)$ of latent variables.



In a Markovian state-space model, the latent variables form a Markov chain

$$p(z_t|z_{1:t-1}) = p(z_t|z_{t-1}),$$

where the conditional distribution $p(z_t|z_{t-1})$ is called the **transition model**.

The observations are conditionally independent given the latent variables,

$$p(x_t|x_{1:t-1}, z_{1:t}) = p(x_t|z_t),$$

where the conditional distribution $p(x_t|z_t)$ is called the **observation model**.

Example

We want to track the location of a wild animal (e.g., a wolf) over time using noisy GPS observations.

Assumptions:

- The animal has a home location (den, nest) at $\mu \in \mathbb{R}^2$.
- The animal moves according to a random walk with drift towards the home location.
- The GPS observations are noisy measurements of the animal's true location.
- Time is discretized regularly every Δt time units.

State-space model:

- States $z_t \in \mathbb{R}^2$ represent the true location of the animal at time t. They evolve as a random walk with drift towards the home location μ.

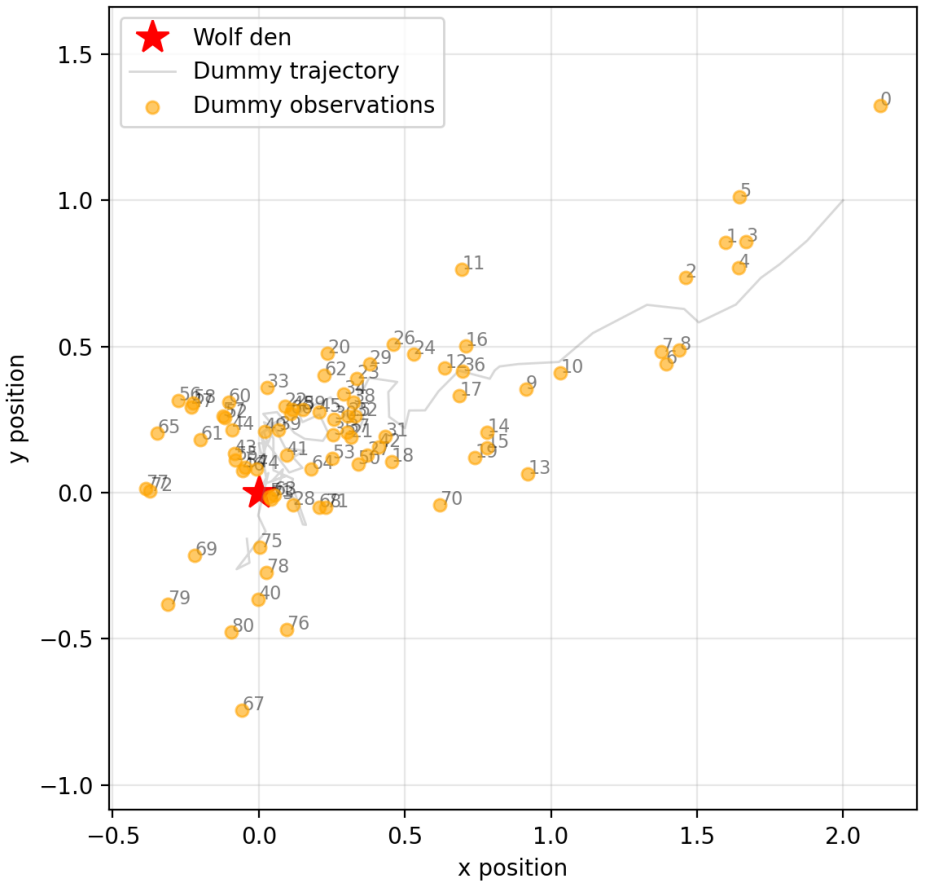
$$p(z_t|z_{t-1}) = \mathcal{N}(z_t|z_{t-1} - \kappa(z_{t-1} - \mu)\Delta t, \sigma^2 \Delta t I),$$

  where $\kappa > 0$ is the strength of attraction to the home location and σ^2 is the variance of the random walk.

- Observations $x_t \in \mathbb{R}^2$ represent noisy GPS measurements of the animal's location at time t. They relate to the states via

$$p(x_t|z_t) = \mathcal{N}(x_t|z_t, R),$$

  where R is the observation noise covariance.



Example of trajectory and observations
generated from the discrete-time state-space model ($\Delta t = 0.25$).

Inference in state-space models

Given a state-space model and a sequence of observations $x_{1:T}$, we are typically interested in solving one or more of the following inference problems:

- Prediction: $p(z_{t+k}|x_{1:t})$ for $k \geq 1$.
- Filtering: $p(z_t|x_{1:t})$.
- Smoothing: $p(z_t|x_{1:T})$.

Bayes filter

The Bayes filter is a recursive algorithm for estimating the filtering distributions $p(z_t|x_{1:t})$ as

$$p(z_t|x_{1:t}) = \frac{p(x_t|z_t)p(z_t|x_{1:t-1})}{p(x_t|x_{1:t-1})},$$

for $t = 1, 2, \ldots, T$, with the base case $p(z_1|x_1) = \frac{p(x_1|z_1)p(z_1)}{p(x_1)}$.

Proof. The filtering distribution can be derived in two steps:

1. Prediction: Push the filtering distribution from the previous time step through the transition model to obtain a prediction of the current state. That is,

$$p(z_t|x_{1:t-1}) = \int p(z_t|z_{t-1})p(z_{t-1}|x_{1:t-1})dz_{t-1}.$$

2. Update: Condition on the new observation to obtain the filtering distribution,

$$p(z_t|x_{1:t}) = \frac{p(x_t|z_t)p(z_t|x_{1:t-1})}{p(x_t|x_{1:t-1})},$$

   where the marginal likelihood $p(x_t|x_{1:t-1})$ is given by

$$p(x_t|x_{1:t-1}) = \int p(x_t|z_t)p(z_t|x_{1:t-1})dz_t.$$

Once we have computed the filtering distributions $p(z_t|x_{1:t})$ for $t = 1, \ldots, T$, we can compute the **prediction distributions** $p(z_{t+k}|x_{1:t})$ for $k \geq 1$ using the prediction step of the Bayes filter iteratively,

$$p(z_{t+k}|x_{1:t}) = \int p(z_{t+k}|z_{t+k-1})p(z_{t+k-1}|x_{1:t})dz_{t+k-1},$$

for $k = 1, 2, \ldots$.

Bayes smoother

The Bayes smoother computes the smoothing distributions $p(z_t|x_{1:T})$ for $t = 1, \ldots, T$ using the filtering distributions $p(z_t|x_{1:t})$ and the transition model $p(z_t|z_{t-1})$. It consists of a backward recursion,

$$p(z_t|x_{1:T}) = p(z_t|x_{1:t}) \int \frac{p(z_{t+1}|z_t)p(z_{t+1}|x_{1:T})}{p(z_{t+1}|x_{1:t})} dz_{t+1},$$

for $t = T-1, T-2, \ldots, 1$, with the base case $p(z_T|x_{1:T}) = p(z_T|x_{1:T})$ (from the filtering step, instead of the backward recursion).

Proof. The joint distribution $p(z_t, z_{t+1}|x_{1:T})$ can be computed as

$$\begin{aligned} p(z_t, z_{t+1}|x_{1:T}) &= p(z_t|z_{t+1}, x_{1:T})p(z_{t+1}|x_{1:T}) \\ &= p(z_t|z_{t+1}, x_{1:t})p(z_{t+1}|x_{1:T}) \\ &= \frac{p(z_{t+1}|z_t)p(z_t|x_{1:t})}{p(z_{t+1}|x_{1:t})}p(z_{t+1}|x_{1:T}), \end{aligned}$$

where we used the conditional independence properties of the state-space model.

Marginalizing over z_{t+1} gives the desired result,

$$\begin{aligned} p(z_t|x_{1:T}) &= \int p(z_t, z_{t+1}|x_{1:T})dz_{t+1} \\ &= p(z_t|x_{1:t}) \int \frac{p(z_{t+1}|z_t)p(z_{t+1}|x_{1:T})}{p(z_{t+1}|x_{1:t})}dz_{t+1}. \end{aligned}$$

Although the Bayes filter and Bayes smoother provide a general framework for inference in state-space models, they are **rarely tractable in practice** as they involve integrals that are difficult to compute.

Further assumptions on the transition and observation models are required for closed-form solutions.

Linear Gaussian state-space models

A linear Gaussian state-space model (LGSSM) is a state-space model where both the transition and observation models are linear Gaussian. That is,

$$p(z_t|z_{t-1}) = \mathcal{N}(z_t|Az_{t-1}, Q),$$
$$p(x_t|z_t) = \mathcal{N}(x_t|Hz_t, R),$$

where A is the state transition matrix, Q is the process noise covariance, H is the observation matrix, and R is the observation noise covariance.

If the prior distribution $p(z_0)$ is also Gaussian, then all filtering, prediction, and smoothing distributions are Gaussian.

The **Kalman filter** provides a closed-form expression for the filtering distributions in linear Gaussian state-space models.

At each time step t, $p(z_t|x_{1:t}) = \mathcal{N}(z_t|m_t, P_t)$ is Gaussian with mean m_t and covariance P_t. These parameters can be computed recursively by adapting the Bayes filter equations to the linear Gaussian case.

Proof. For the prediction step, we have

$$
\begin{aligned}
p(z_t|x_{1:t-1}) &= \int p(z_t|z_{t-1})p(z_{t-1}|x_{1:t-1})dz_{t-1} \\
&= \int \mathcal{N}(z_t|Az_{t-1}, Q)\mathcal{N}(z_{t-1}|m_{t-1}, P_{t-1})dz_{t-1} \\
&= \int \mathcal{N}\left(\begin{pmatrix} z_t \\ z_{t-1} \end{pmatrix} | \begin{bmatrix} Am_{t-1} \\ m_{t-1} \end{bmatrix}, \begin{bmatrix} Q + AP_{t-1}A^T & AP_{t-1} \\ P_{t-1}A^T & P_{t-1} \end{bmatrix}\right) dz_{t-1} \\
&= \mathcal{N}(z_t|m_t^-, P_t^-),
\end{aligned}
$$

where $m_t^- = Am_{t-1}$ and $P_t^- = AP_{t-1}A^T + Q$.

For the update step, we join the prediction distribution with the observation model,

$$p\left(\begin{pmatrix} z_t \\ x_t \end{pmatrix} | x_{1:t-1}\right) = p(z_t|x_{1:t-1})p(x_t|z_t)$$
$$= \mathcal{N}\left(\begin{pmatrix} z_t \\ x_t \end{pmatrix} | \begin{bmatrix} m_t^- \\ Hm_t^- \end{bmatrix}, \begin{bmatrix} P_t^- & P_t^- H^T \\ HP_t^- & HP_t^- H^T + R \end{bmatrix}\right).$$

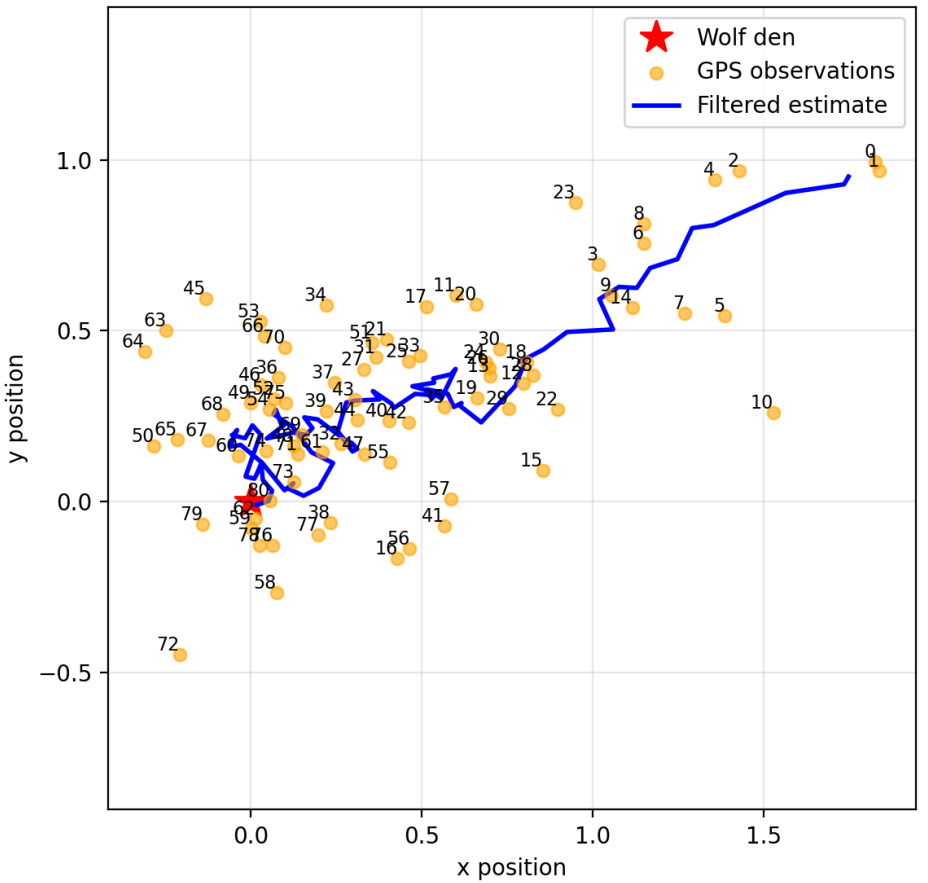
Therefore, the filtering distribution is given by the conditional distribution

$$p(z_t|x_{1:t-1}, x_t) = p(z_t|x_{1:t}) = \mathcal{N}(z_t|m_t, P_t),$$

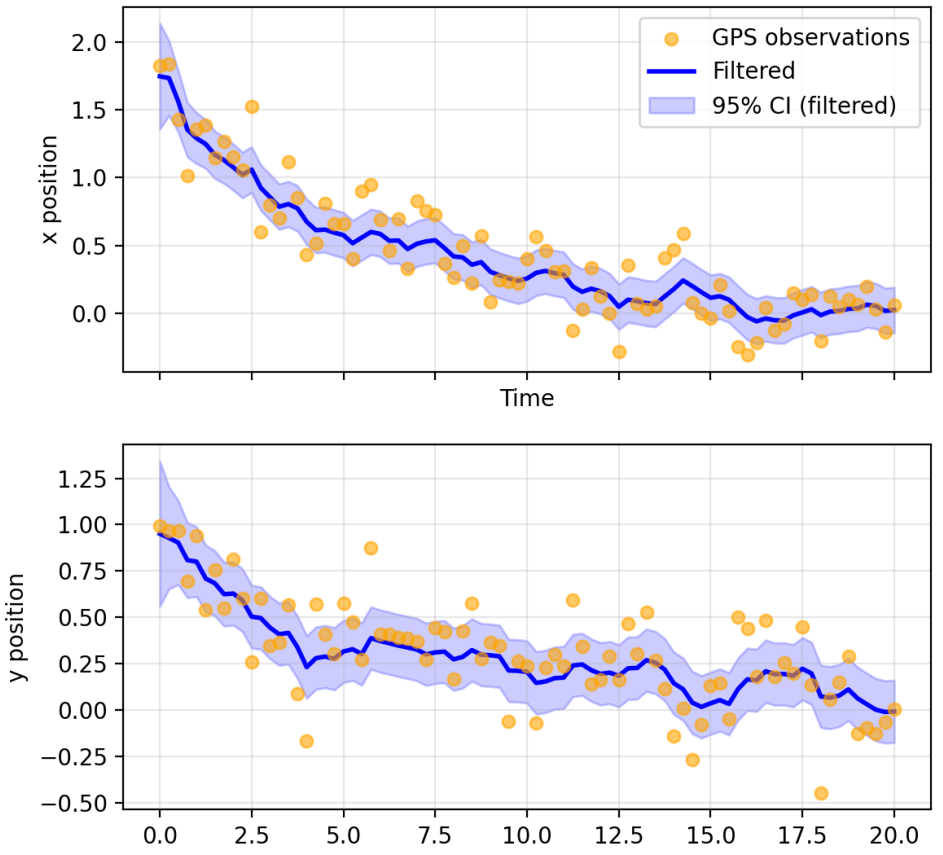
where

$$m_t = m_t^- + K_t(x_t - Hm_t^-),$$
$$P_t = (I - K_t H)P_t^-,$$

and $K_t = P_t^- H^T(HP_t^- H^T + R)^{-1}$ is the Kalman gain and represents the weight given to the new observation.



Mean estimate of the wolf's trajectory using the Kalman filter.



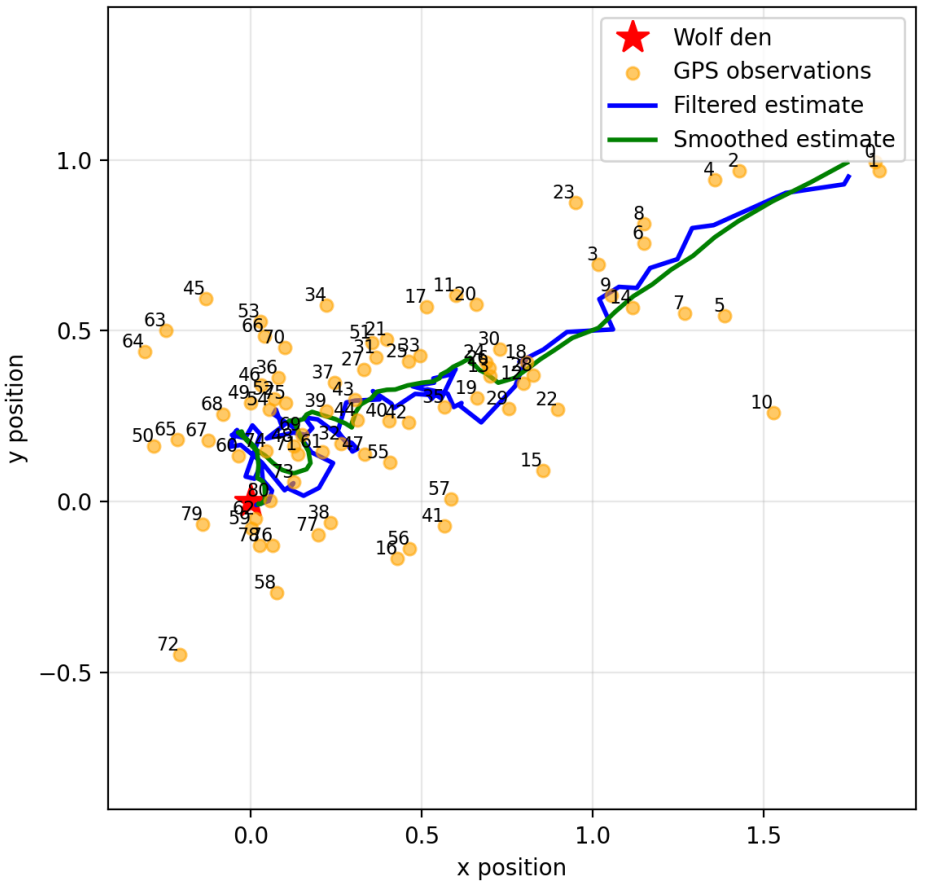
Filtering distribution at each time step using the Kalman filter.

The smoothing distributions $p(z_t|x_{1:T}) = \mathcal{N}(z_t|m_t^s, P_t^s)$ are also Gaussian with mean m_t^s and covariance P_t^s.

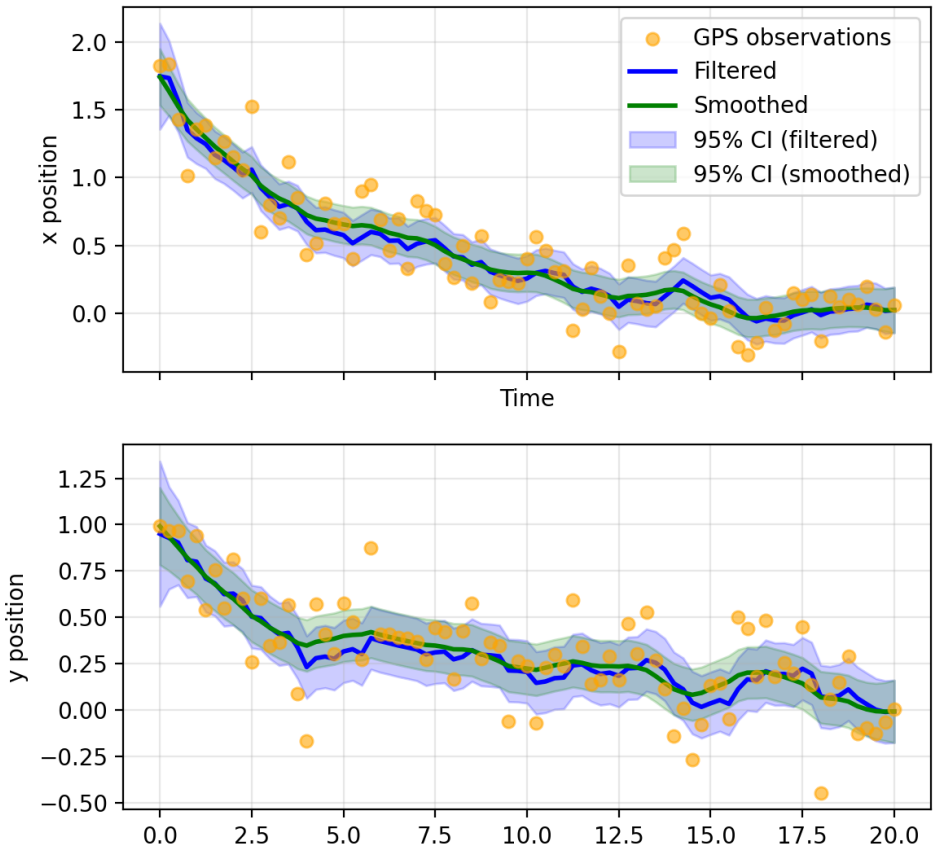
The parameters can be computed recursively using the **Rauch-Tung-Striebel smoother** equations (proof omitted for brevity),

$$
\begin{aligned}
C_t &= P_t A^T (P_{t+1}^-)^{-1}, \\
m_t^s &= m_t + C_t(m_{t+1}^s - m_{t+1}^-), \\
P_t^s &= P_t + C_t(P_{t+1}^s - P_{t+1}^-)C_t^T,
\end{aligned}
$$

for $t = T-1, T-2, \ldots, 1$, with the base case $m_T^s = m_T$ and $P_T^s = P_T$.



Mean estimate of the wolf's trajectory using the Kalman smoother.



Smoothing distribution at each time step using the Kalman smoother.

Hidden Markov models

A hidden Markov model (HMM) is a state-space model where all variables are discrete and the transition and observation models are categorical distributions. That is,

$$p(z_t = j | z_{t-1} = i) = A_{i,j},$$
$$p(x_t = k | z_t = j) = B_{j,k},$$

where A is the state transition matrix and B is the observation matrix.

If the prior distribution $p(z_0)$ is also categorical, then all filtering, prediction, and smoothing distributions are categorical and can be computed exactly by enumeration.

Example: Wolf behavior modeling

- States $z_t \in 1, ..., K$ represent the behavior of the animal at time t (e.g., resting, foraging, traveling).
- Observations $x_t \in 1, ..., M$ represent discrete measurements related to the animal's behavior (e.g., GPS speed categories, activity levels).
- Transition model $p(z_t|z_{t-1})$ captures the probabilities of switching between different behaviors.
- Observation model $p(x_t|z_t)$ captures the probabilities of observing certain measurements given the animal's behavior.

The **forward algorithm** provides a closed-form expression for the filtering distributions in hidden Markov models.

At each time step t, $p(z_t|x_{1:t})$ is categorical with parameters $\alpha_t(j) = p(z_t = j|x_{1:t})$. These parameters can be computed recursively using the forward algorithm equations

$$\alpha_t \propto O_t A^T \alpha_{t-1},$$

for $t = 1, 2, \ldots, T$, with the base case $\alpha_1 \propto O_1 \pi$, where π are the parameters of the prior distribution $p(z_0)$ and O_t is a diagonal matrix with entries $B_{:,x_t}$. The proportionality constant is obtained by normalizing α_t so that its entries sum to 1.

The smoothing distributions $p(z_t|x_{1:T})$ are also categorical with parameters $\gamma_t(j) = p(z_t = j|x_{1:T})$. The parameters can be computed recursively using the **backward algorithm** equations

$$\beta_t \propto AO_{t+1}\beta_{t+1},$$

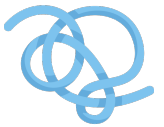
for $t = T-1, T-2, \ldots, 1$, with the base case $\beta_T = \mathbf{1}$, where $\mathbf{1}$ is a vector of ones. The proportionality constant is obtained by normalizing β_t so that its entries sum to 1.

The smoothing parameters are then given by $\gamma_t \propto \alpha_t \odot \beta_t$, where $\odot$ denotes the element-wise product.

Both linear Gaussian state-space models and hidden Markov models are special cases of state-space models where exact inference is tractable.

However, they are limited in their expressiveness and may not capture the complexity of real-world dynamical systems.

Continuous-time models



We have so far assumed that time is discretized regularly with a fixed time step Δt and that both the transition and observation models are defined at these discrete time steps.

However,

- physical processes are often more naturally modeled in **continuous time**;
- observations may be collected at **irregular time intervals**, triggered by events rather than a clock, or at multiple time scales.

From discrete to continuous time

In discrete-time state-space models, we considered transition models of the form $p(z_t|z_{t-1})$ that describe how the state evolves from one time step to the next. For additive transitions and additive noise, this can be expressed as

$$z_t = z_{t-1} + f(z_{t-1})\Delta t + w_t,$$

where f is a deterministic function and w_t is random noise.

Shuffling the terms, we get

$$\frac{z_t - z_{t-1}}{\Delta t} = f(z_{t-1}) + \frac{w_t}{\Delta t},$$

which, in the limit as $\Delta t \to 0$, gives us a continuous-time model

$$\frac{dz(t)}{dt} = f(z(t)) + \frac{dw(t)}{dt},$$

where $z(t)$ is the state at time t and $w(t)$ is continuous-time noise.

Omitting $\frac{dw(t)}{dt}$, we get a **deterministic** dynamical system described by an **ordinary differential equation** (ODE)

$$\frac{dz(t)}{dt} = f(z(t)).$$

The solution of this ODE with initial condition $z(0) = z_0$ is given by

$$z(t) = z_0 + \int_0^t f(z(\tau))d\tau.$$

Example: Exponential decay to an equilibrium point

$$\frac{dz(t)}{dt} = -\kappa(z(t) - \mu),$$

where $\kappa > 0$ is the rate of decay and μ is the equilibrium point.

- If $z(0) > \mu$, then $z(t)$ decreases towards μ as t increases.
- If $z(0) < \mu$, then $z(t)$ increases towards μ as t increases.
- Solution:

$$z(t) = \mu + (z(0) - \mu)e^{-\kappa t}.$$

This is deterministic: given $z(0)$, the state $z(t)$ is fully determined for all $t \geq 0$!

Brownian motion

To add stochasticity to the ODE, we need a continuous-time stochastic process that can model random noise.

We can model the noise term $w(t)$ as a standard Brownian motion (Wiener process) $B(t)$, which has the following properties:

- $B(0) = 0$.
- $B(t)$ has independent increments: for $0 \leq s < t$, $B(t) - B(s) \sim \mathcal{N}(0, t - s)$.
- $B(t)$ is continuous in t.

Adding Brownian motion to the ODE, we get a **stochastic differential equation** (SDE)

$$\frac{dz(t)}{dt} = f(z(t), t) + \frac{dB(t)}{dt},$$

where $\frac{dB(t)}{dt}$ is an informal notation for white noise.

More rigorously, Brownian motion is nowhere differentiable and the notation $\frac{dB(t)}{dt}$ is only symbolic. The SDE can instead be defined in differential form as

$$dz(t) = f(z(t), t)dt + dB(t).$$

For more generality, we can introduce a **diffusion term** $g(z(t), t)$ to scale the noise, leading to the SDE

$$dz(t) = f(z(t), t)dt + g(z(t), t)dB(t).$$

In this form, the SDE describes the infinitesimal change in the state $z(t)$ over an infinitesimal time interval dt.

- The drift term $f(z(t), t)dt$ represents the deterministic change in the state.
- The diffusion term $g(z(t), t)dB(t)$ represents the stochastic change in the state due to Brownian motion.

Example: Ornstein-Uhlenbeck process

Recall our animal movement example in discrete time:

$$z_t = z_{t-1} - \kappa(z_{t-1} - \mu)\Delta t + w_t,$$

where $w_t \sim \mathcal{N}(0, \sigma^2 \Delta t I)$.

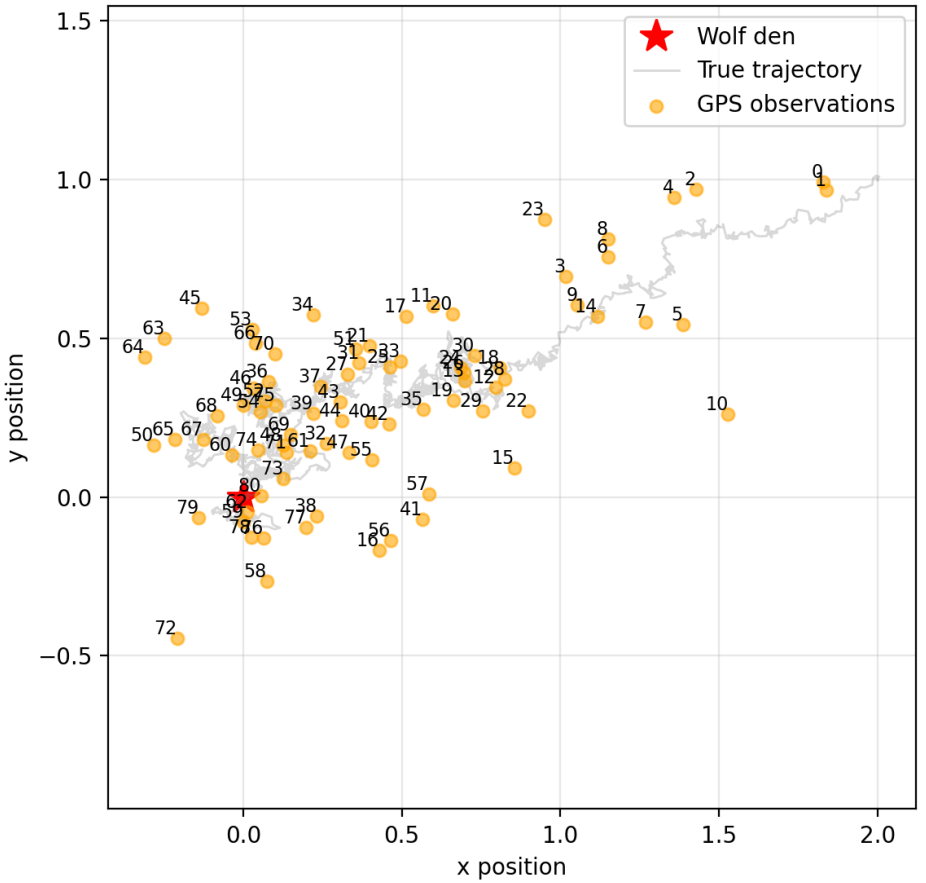
In continuous time, this becomes the SDE

$$dz(t) = -\kappa(z(t) - \mu)dt + \sigma dB(t),$$

where $\kappa > 0$ is the strength of attraction to the home location μ and σ is the diffusion coefficient.

This process is known as the **Ornstein-Uhlenbeck process**, which describes a mean-reverting behavior with Gaussian noise.

> The discrete-time model is the Euler-Maruyama discretization of the OU process with step size Δt!



Example of continuous trajectory generated from the Ornstein-Uhlenbeck process.
(This is the true trajectory used throughout the lecture.)

Observations in continuous time

In continuous-time state-space models, the observation model can be defined as a conditional distribution $p(x(t_i)|z(t_i))$ at any (continuous) time point t_i.

This is similar to the discrete-time case, except that observations can be collected at irregular time intervals $t_1 < t_2 < \ldots < t_N$ rather than at fixed time steps.

Linear Gaussian continuous-time state-space models

Linear Gaussian continuous-time state-space models are continuous-time analogs of linear Gaussian state-space models.

They are defined by linear SDEs for the state dynamics and linear Gaussian observation models,

$$
\begin{aligned}
dz(t) &= Az(t)dt + dB(t), \\
x(t_i) &\sim \mathcal{N}(x(t_i)|Hz(t_i), R),
\end{aligned}
$$

where A is the state transition matrix, H is the observation matrix, and R is the observation noise covariance.

Filtering and smoothing distributions $p(z(t_i)|x(t_{1:i}))$ and $p(z(t)|x(t_{1:N}))$ can be computed exactly using continuous-time analogs of the Kalman filter and Rauch-Tung-Striebel smoother.

Both now correspond to stochastic processes over continuous time rather than sequences over discrete time steps.

When to use continuous vs discrete time?

Continuous-time is natural when:

- Observations at irregular intervals
- Physical/mechanistic interpretation important
- Parameters have continuous-time meaning (rates, time constants)

Discrete-time is practical when:

- Regular sampling
- Computational simplicity preferred
- No strong mechanistic model

Common approach: Model in continuous time for interpretability, discretize for computation.

The end.