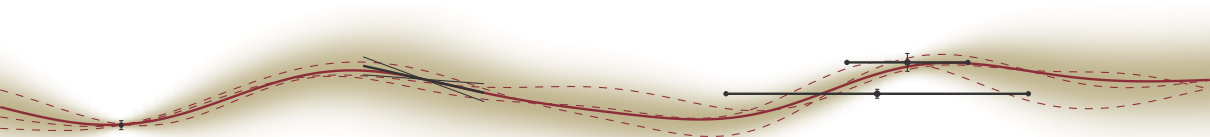
Probabilistic Inference and Learning
Lecture 09
Markov Chains: Models for Time Series

Philipp Hennig

14 November 2018

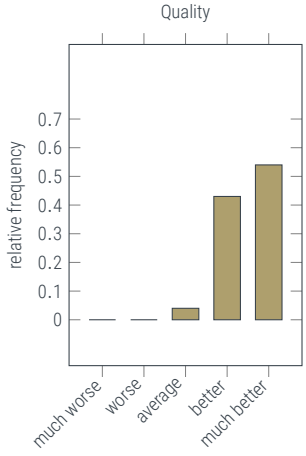
Eberhard Karls
Universität
Tübingen

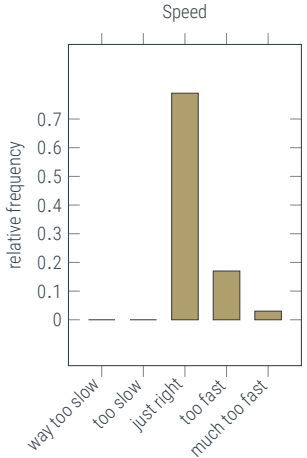
Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning

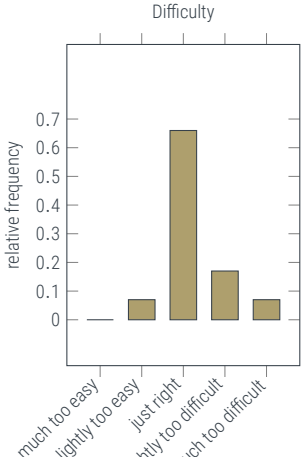


Last Lecture: Debrief

Feedback dashboard







Last Lecture: Debrief

Detailed Feedback

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Things you did not like:

- **blackboard writing**
- python
- no solutions to exercises online
- too much personal story

Things you did not understand:

- everything after the break
- $df/d\log x = df/dx \cdot x$?
- how, why and when to learn mean functions
- please repeat questions from audience

Things you enjoyed:

- **code** (!!!)
- making connections to previous lectures
- the broad scope. "More of this meta-stuff please"
- moving away from 'stupid data-fitting' to interpretable models

Overview of Lectures so far:

0. Introduction to Reasoning under Uncertainty
1. Probabilistic Reasoning
2. Probabilities over Continuous Variables
3. Gaussian Probability Distributions
4. Gaussian Parametric Regression
5. More on Parametric Regression — Connections to Deep Learning
6. Gaussian Processes
7. More on Kernels & GPs
8. A practical example

Today:

- Data arriving in a **stream through time**

today's lecture:

- how to handle inference on an infinite stream of incoming data, at constant cost
- lots of famous words & concepts:
  - Markov Chains
  - Time Series
  - (Kalman) filtering & smoothing
  - Signal processing
  - Hidden Markov Models

Time Series

a widely applicable concept

Definition

A **time series** is a sequence $[y(t_i)]_{i\in\mathbb{N}}$ of observations $y_i := x(t_i) \in \mathbb{Y}$, indexed by a scalar variable $t \in \mathbb{R}$. In many applications, the time points t_i are equally spaced: $t_i = t_0 + i \cdot \delta_t$. Models that account for all values $t \in \mathbb{R}$ are called *continuous time*, while models that only consider $[t_i]_{i\in\mathbb{N}}$ are called *discrete time*.

Examples:

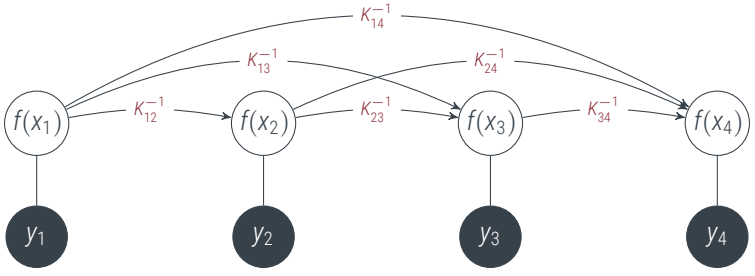
- climate & weather observations … in Climate Science
- sensor readings in cars, … in Engineering
- EEG, ECG, patch clamp signals, … in Medicine and Neuroscience
- just about any sensing of a dynamical process in Physics
- stock prices, supply & demand data, polling numbers, … in Economics and Social Science
- body weight measurements in the previous lecture

Inference in time series often has to happen in real-time, and scale to an unbounded set of data, typically on small-scale or embedded systems. So it has to be of (low) **constant time and memory** complexity.

Graphical View I: "Full" GP

Recall from Lecture 1: Complexity of Inference can be controlled by **Conditional Independence**

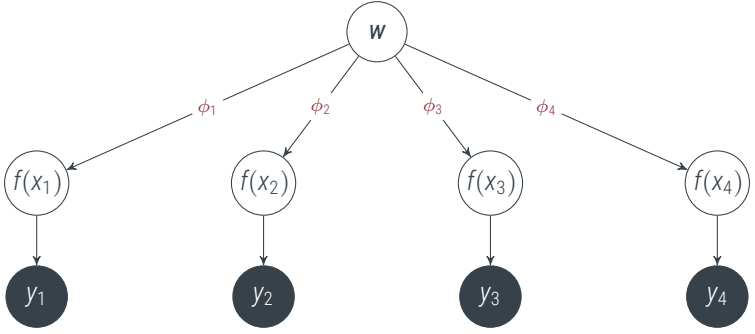
$$p(f) = \mathcal{GP}(f; 0, k) \quad p\left(\begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix}\right) = \mathcal{N}\left(0, \begin{bmatrix} K_{11}^{-1} & K_{12}^{-1} & K_{13}^{-1} & K_{14}^{-1} \\ & K_{22}^{-1} & K_{23}^{-1} & K_{24}^{-1} \\ & & K_{33}^{-1} & K_{34}^{-1} \\ & & & K_{44}^{-1} \end{bmatrix}^{-1}\right) \quad p(\boldsymbol{y} \mid \boldsymbol{f}) = \prod_i \mathcal{N}(y_i; f_i, \sigma^2)$$



Graphical View II: Parametric Model

Recall from Lecture 1: Complexity of Inference can be controlled by **Conditional Independence**

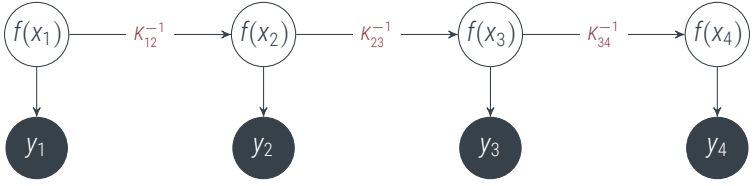
$$p(f) = \mathcal{GP}(f; 0, \Phi_X^\intercal \Sigma \Phi_X) \quad p\left(\begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix} \middle| \, \boldsymbol{w} \right) = \prod_i \delta(f_i - \phi_i^\intercal \boldsymbol{w}) \quad p(\boldsymbol{y} \mid \boldsymbol{f}) = \prod_i \mathcal{N}(y_i; f_i, \sigma^2)$$



Graphical View III: Markov Chain

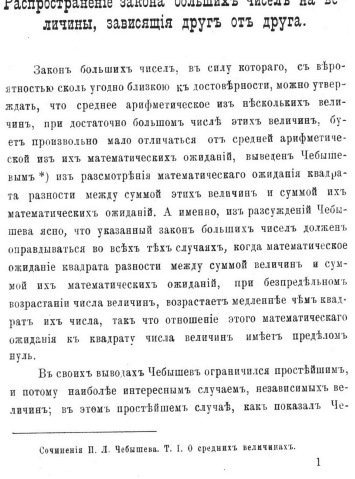
Recall from Lecture 1: Complexity of Inference can be controlled by **Conditional Independence**

$$p(f) = \mathcal{GP}(f; 0, k) \quad p\left(\begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix}\right) = \mathcal{N}\left(0, \begin{bmatrix} K_{11}^{-1} & K_{12}^{-1} & 0 & 0 \\ K_{12}^{-1} & K_{22}^{-1} & K_{23}^{-1} & 0 \\ 0 & K_{23}^{-1} & K_{33}^{-1} & K_{34}^{-1} \\ 0 & 0 & K_{34}^{-1} & K_{44}^{-1} \end{bmatrix}^{-1}\right) \quad p(\boldsymbol{y} \mid \boldsymbol{f}) = \prod_i \mathcal{N}(y_i; f_i, \sigma^2)$$



It's all about (Conditional) Independence

This point is way to easily missed



Распространеніе закона большихъ чиселъ на величины, зависящія другъ отъ друга.

Законъ большихъ чиселъ, въ силу котораго, съ вѣроятностью сколь угодно близкою къ достовѣрности, можно утверждать, что среднее ариѳметическое изъ нѣсколькихъ величинъ, при достаточно большомъ числѣ этихъ величинъ, будетъ произвольно мало отличаться отъ средней ариѳметической изъ ихъ математическихъ ожиданій, выведенъ Чебышевымъ *) изъ разсмотрѣнія математическаго ожиданія квадрата разности между суммой этихъ величинъ и суммой ихъ математическихъ ожиданій. А именно, изъ разсужденій Чебышева ясно, что указанный законъ большихъ чиселъ долженъ оправдываться во всѣхъ тѣхъ случаяхъ, когда математическое ожиданіе квадрата разности между суммой величинъ и суммой ихъ математическихъ ожиданій, при безпредѣльномъ возрастаніи числа величинъ, возрастаетъ медленнѣе чѣмъ квадратъ ихъ числа, такъ что отношеніе этого математическаго ожиданія къ квадрату числа величинъ имѣетъ предѣломъ нуль.

Въ своихъ выводахъ Чебышевъ ограничился простѣйшимъ, и потому наиболѣе интереснымъ случаемъ, независимыхъ величинъ; въ этомъ простѣйшемъ случаѣ, какъ показалъ Че-

*) Сочиненія П. Л. Чебышева. Т. I. О среднихъ величинахъ.

1

A generalization of the law of large numbers to variables that depend on each other.

Proceedings of the Society for Physics and Mathematics at Kazan University, 1906

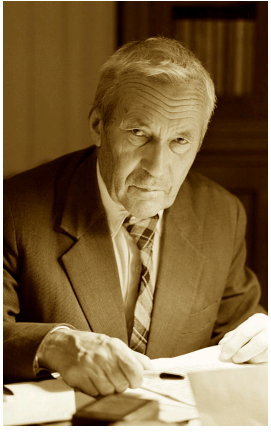
Andrej Andreevič Markov
(1856 – 1922)

It's all about (Conditional) Independence

This point is way to easily missed

[A. Kolmogoroff. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Zentralblatt d. Math. 1933]

Geschichtlich ist die Unabhängigkeit von Versuchen und zufälligen Größen derjenige mathematische Begriff, welcher der Wahrscheinlichkeitsrechnung ihr eigenartiges Gepräge gibt. Die klassischen Arbeiten von LAPLACE, POISSON, TCHEBYCHEFF, MARKOFF, LIAPOUNOFF, v. MISES und BERNSTEIN sind in der Tat im wesentlichen der Untersuchung von Reihen unabhängiger Größen gewidmet. Wenn man in den neueren Untersuchungen (MARKOFF, BERNSTEIN usw.) öfters die Forderung der vollständigen Unabhängigkeit ablehnt, so sieht man sich immer gezwungen, um hinreichend inhaltreiche Resultate zu erhalten, abgeschwächte analoge Forderungen einzuführen. (Vgl. in diesem Kap. § 6 — MARKOFFsche Ketten.)



It's all about (Conditional) Independence

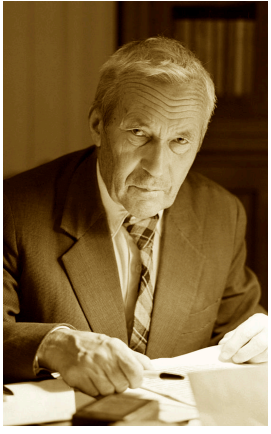
This point is way to easily missed

[A. Kolmogoroff. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Zentralblatt d. Math. 1933]

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Man kommt also dazu, im Begriffe der Unabhängigkeit wenigstens den ersten Keim der eigenartigen Problematik der Wahrscheinlichkeitsrechnung zu erblicken — ein Umstand, welcher in diesem Buche nur wenig hervortreten wird, da wir hier hauptsächlich nur mit den logischen Vorbereitungen zu den eigentlichen wahrscheinlichkeitstheoretischen Untersuchungen zu tun haben werden.

Es ist dementsprechend eine der wichtigsten Aufgaben der Philosophie der Naturwissenschaften, nachdem sie die vielumstrittene Frage über das Wesen des Wahrscheinlichkeitsbegriffes selbst erklärt hat, die Voraussetzungen zu präzisieren, bei denen man irgendwelche gegebene reelle Erscheinungen für gegenseitig unabhängig halten kann. Diese Frage fällt allerdings aus dem Rahmen unseres Buches.



Cave: Change of Notation

- Previously: Observe $y \in \mathbb{R}^D$ at N locations $x \in \mathbb{X}$, assume latent *function* $f \in \mathbb{R}^M$, and $y \approx Hf(x)$.
- The notion of a *local* finite memory only works in an *ordered* space of inputs. Thus, $\mathbb{X} \subset \mathbb{R}$.
- Now: Observe $y_1, \dots, y_N$ with $y_i \in \mathbb{R}^D$ at *times* $[t_1, \dots, t_N]$ with $t_i \in \mathbb{R}$.
  Assume latent *state* $x_i \in \mathbb{R}^M$, and $y_i \approx Hx(t_i)$. (The state will constitute the local memory)
- Such models are known as *state-space models*. (They are related to Finite-State Machines)

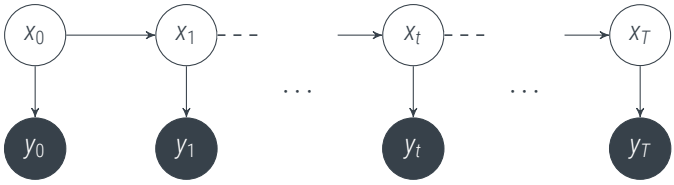
Definition: A joint distribution $p(X)$ over a sequence of random variabels $X := [x_0, \dots, x_N]$ is said to have **the Markov property** if

$$p(x_i \mid x_0, x_1, \dots, x_{i-1}) = p(x_t \mid x_{i-1}).$$

The sequence is then called a **Markov chain**.

Assume:

$$p(x_t \mid X_{0:t-1}) = p(x_t \mid x_{t-1})$$
$$\text{and} \quad p(y_t \mid X) = p(y_t \mid x_t)$$



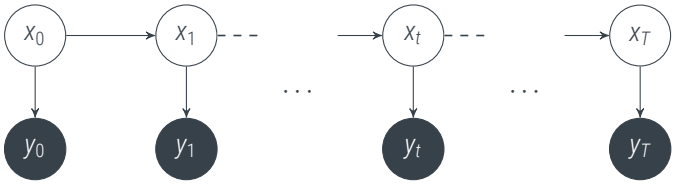
$$p(x_t \mid Y_{0:t-1}) = \frac{\int_{j \neq t} p(X) p(Y_{0:t-1} \mid X)\, dx_j}{\int p(X) p(Y_{0:t-1} \mid X)\, dX} = \frac{\int_{j \neq t} p(Y_{0:t-1} \mid X_{0:t-1}) p(x_0) \left(\prod_{0<j<t} p(x_j \mid x_{j-1})\, dx_j\right) p(x_t \mid x_{t-1}) \left(\prod_{j>t} p(x_j \mid x_{j-1})\, dx_j\right)}{\int p(Y_{0:t-1} \mid X_{0:t-1}) p(x_0) \left(\prod_{0<j<t} p(x_j \mid x_{j-1})\right) p(x_t \mid x_{t-1}) \left(\prod_{j>t} p(x_j \mid x_{j-1})\right) dX}$$

$$= \frac{\int_{j<t} p(x_t \mid x_{t-1}) p(Y_{0:t-1} \mid X_{0:t-1}) p(x_0) \left(\prod_{0<j<t} p(x_j \mid x_{j-1})\, dx_j\right)}{\int_{j<t} p(Y_{0:t-1} \mid X_{0:t-1}) p(x_0) \left(\prod_{0<j<t} p(x_j \mid x_{j-1})\, dx_j\right)} = \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid Y_{0:t-1})\, dx_{t-1}$$

$$p(x_t \mid Y_{0:t}) = \frac{p(y_t \mid x_t) p(x_t \mid Y_{0:t-1})}{\int p(y_t \mid x_t) p(x_t \mid Y_{0:t-1})\, dx_t}$$

Assume:

$$p(x_t \mid X_{0:t-1}) = p(x_t \mid x_{t-1})$$
$$\text{and} \quad p(y_t \mid X) = p(y_t \mid x_t)$$



$$p(x_t \mid Y) = \int p(x_t, x_{t+1} \mid Y)\, dx_{t+1} = \int p(x_t \mid x_{t+1}, Y) p(x_{t+1} \mid Y)\, dx_{t+1}$$

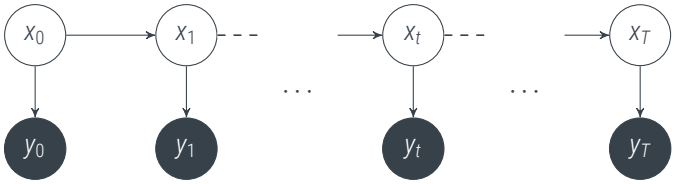
$$p(x_t \mid x_{t+1}, Y) = \frac{p(Y_{t+1:n} \mid x_{t+1}, x_t, Y_{0:t}) p(x_t \mid x_{t+1}, Y_{0:t})}{\int p(Y_{t+1:n} \mid x_{t+1}, x_t, Y_{0:t}) p(x_t \mid x_{t+1}, Y_{0:t})\, dx_t} = \frac{p(Y_{t+1:n} \mid x_{t+1}, Y_{0:t}) \cdot p(x_t \mid x_{t+1}, Y_{0:t})}{p(Y_{t+1:n} \mid x_{t+1}, Y_{0:t}) \cdot \int p(x_t \mid x_{t+1}, Y_{0:t})\, dx_t} = p(x_t \mid x_{t+1}, Y_{0:t})$$

$$p(x_t \mid x_{t+1}, Y_{0:t}) = \frac{p(x_t, x_{t+1} \mid Y_{0:t})}{p(x_{t+1} \mid Y_{0:t})} = \frac{p(x_{t+1} \mid x_t, Y_{0:t}) p(x_t \mid Y_{0:t})}{p(x_{t+1} \mid Y_{0:t})} = \frac{p(x_{t+1} \mid x_t) p(x_t \mid Y_{0:t})}{p(x_{t+1} \mid Y_{0:t})}$$

$$p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \frac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{1:t})}\, dx_{t+1}$$

Assume:

$$p(x_t \mid X_{0:t-1}) = p(x_t \mid x_{t-1})$$
$$\text{and} \quad p(y_t \mid X) = p(y_t \mid x_t)$$



Filtering: $\mathcal{O}(T)$

predict: $p(x_t \mid Y_{0:t-1}) = \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid Y_{0:t-1})\, dx_{t-1}$ (Chapman-Kolmogorov Eq.)

update: $p(x_t \mid Y_{0:t}) = \dfrac{p(y_t \mid x_t) p(x_t \mid Y_{0:t-1})}{p(y_t)}$

Smoothing: $\mathcal{O}(T)$

smooth: $p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \dfrac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{1:t})}\, dx_{t+1}$

Time Series:

- **Markov Chains** formalize the notion of a stochastic process with a *local finite memory*
- Inference over Markov Chains separates into three operations, that can be performed in *linear* time:

Filtering: $\mathcal{O}(T)$

predict: $$p(x_t \mid Y_{0:t-1}) = \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid Y_{0:t-1})\, dx_{t-1}$$ (Chapman-Kolmogorov Eq.)

update: $$p(x_t \mid Y_{0:t}) = \frac{p(y_t \mid x_t) p(x_t \mid Y_{0:t-1})}{p(y_t)}$$

Smoothing: $\mathcal{O}(T)$

smooth: $$p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \frac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{1:t})}\, dx_{t+1}$$

procedure INFERENCE($Y, p(x_0), p(x_t \mid x_{t-1})\ \forall t, p(y_t \mid x_t)\ \forall t$)

  for i=1,…,n **do** // Filtering

    $p_-(x_t \mid y_{1:t-1}) = \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid Y_{0:t-1})\, dx_{t-1}$ // Chapman-Kolmogorov eq.

    $p(x_t \mid y_{1:t}) = p(y_t \mid x_t) p(x_t \mid Y_{0:t-1}) / p(y_t)$ // Update

  end for

  for i=n-1,…,0 **do** // Smoothing

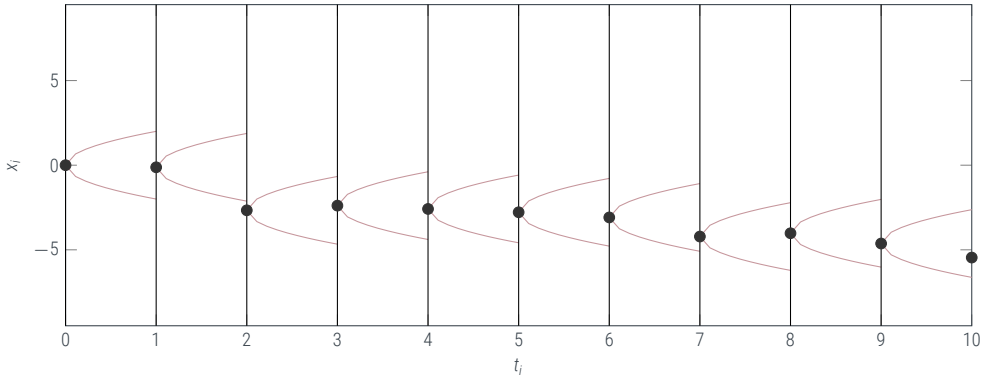
    $p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) p(x_{t+1} \mid Y) p(x_{t+1} \mid Y_{1:t})\, dx_{t+1}$

  end for

  return $p(x_t \mid Y)\ \forall t = 0, \ldots, n$ // return all marginals

end procedure

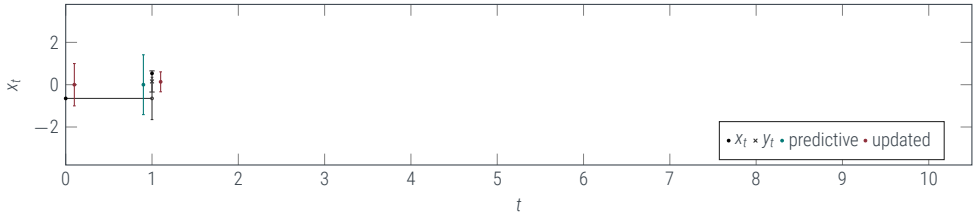
$$p(x(t_{i+1}) \mid x(t_i)) = \mathcal{N}(x_{i+1}; Ax_i, Q) \qquad \text{(figure: } x_0 = 0, A = Q = 1\text{)}$$



Kalman Filtering and Rauch-Tung-Striebel Smoothing

linear-time inference on Gaussian-Markov time series

$$p(x(t_i) \mid x(t_{i-1})) = \mathcal{N}(x_i; Ax_{i-1}, Q) \qquad p(y_t \mid x_t) = \mathcal{N}(y_t; Hx_t, R)$$



predict: $p(x_t \mid Y_{1:t-1}) = \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid Y_{0:t-1})\, dx_{t-1} = \mathcal{N}(x_t, m_t^-, P_t^-) = \mathcal{N}(x_t, Am_t, AP_tA^\mathsf{T} + Q)$

update: $p(x_t \mid Y_{1:t}) = \dfrac{p(y_t \mid x_t) p(x_t \mid Y_{0:t-1})}{p(y_t)} = \mathcal{N}(x_t, m_t, P_t) = \mathcal{N}(x_t, m_t^- + Kz, (I - KH)P_t^-)$

$$K := P_t^- H^\mathsf{T} (HP_tH^\mathsf{T} + R)^{-1}, \qquad z := y_t - Hm_t^-,$$

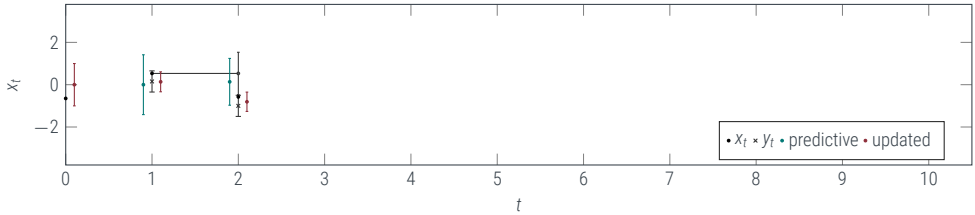
smooth: $p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \dfrac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{1:t})}\, dx_{t+1} = \mathcal{N}(x_t, m_t^s, P_t^s)$

$$= \mathcal{N}(x_t, m_t + G_t(m_{t+1}^s - m_{t+1}^-), P_t + G_t(P_{t+1}^s - P_{t+1}^-)G_t^\mathsf{T}) \qquad G_t := P_tA^\mathsf{T}(P_t^-)^{-1}$$

Kalman Filtering and Rauch-Tung-Striebel Smoothing

linear-time inference on Gaussian-Markov time series

$$p(x(t_i) \mid x(t_{i-1})) = \mathcal{N}(x_i; Ax_{i-1}, Q) \qquad p(y_t \mid x_t) = \mathcal{N}(y_t; Hx_t, R)$$



predict: $p(x_t \mid Y_{1:t-1}) = \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid Y_{0:t-1})\, dx_{t-1} = \mathcal{N}(x_t, m_t^-, P_t^-) = \mathcal{N}(x_t, Am_t, AP_tA^\mathsf{T} + Q)$

update: $p(x_t \mid Y_{1:t}) = \dfrac{p(y_t \mid x_t) p(x_t \mid Y_{0:t-1})}{p(y_t)} = \mathcal{N}(x_t, m_t, P_t) = \mathcal{N}(x_t, m_t^- + Kz, (I - KH)P_t^-)$

$$K := P_t^- H^\mathsf{T} (HP_tH^\mathsf{T} + R)^{-1}, \qquad z := y_t - Hm_t^-,$$

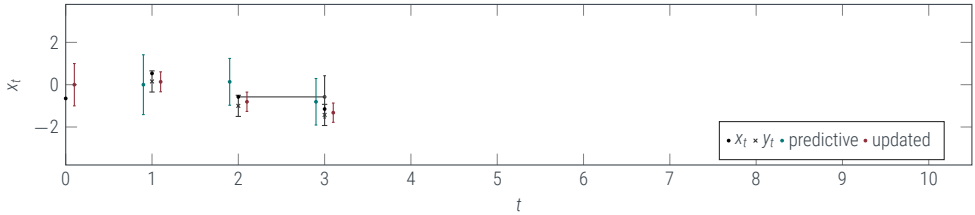
smooth: $p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \dfrac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{1:t})}\, dx_{t+1} = \mathcal{N}(x_t, m_t^s, P_t^s)$

$$= \mathcal{N}(x_t, m_t + G_t(m_{t+1}^s - m_{t+1}^-), P_t + G_t(P_{t+1}^s - P_{t+1}^-)G_t^\mathsf{T}) \qquad G_t := P_tA^\mathsf{T}(P_t^-)^{-1}$$

Kalman Filtering and Rauch-Tung-Striebel Smoothing

linear-time inference on Gaussian-Markov time series

$$p(x(t_i) \mid x(t_{i-1})) = \mathcal{N}(x_i; Ax_{i-1}, Q) \qquad p(y_t \mid x_t) = \mathcal{N}(y_t; Hx_t, R)$$



predict: $p(x_t \mid Y_{1:t-1}) = \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid Y_{0:t-1})\, dx_{t-1} = \mathcal{N}(x_t, m_t^-, P_t^-) = \mathcal{N}(x_t, Am_t, AP_tA^\mathsf{T} + Q)$

update: $p(x_t \mid Y_{1:t}) = \dfrac{p(y_t \mid x_t) p(x_t \mid Y_{0:t-1})}{p(y_t)} = \mathcal{N}(x_t, m_t, P_t) = \mathcal{N}(x_t, m_t^- + Kz, (I - KH)P_t^-)$

$$K := P_t^- H^\mathsf{T} (HP_tH^\mathsf{T} + R)^{-1}, \qquad z := y_t - Hm_t^-,$$

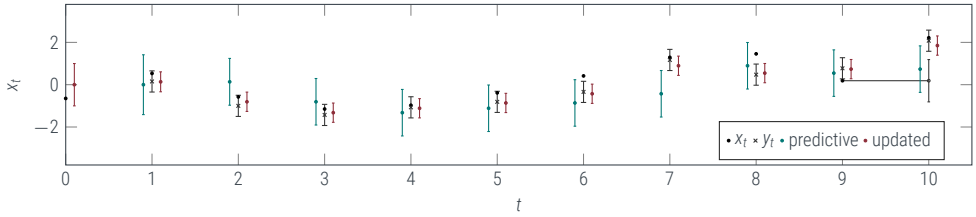
smooth: $p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \dfrac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{1:t})}\, dx_{t+1} = \mathcal{N}(x_t, m_t^s, P_t^s)$

$$= \mathcal{N}(x_t, m_t + G_t(m_{t+1}^s - m_{t+1}^-), P_t + G_t(P_{t+1}^s - P_{t+1}^-)G_t^\mathsf{T}) \qquad G_t := P_tA^\mathsf{T}(P_t^-)^{-1}$$

Kalman Filtering and Rauch-Tung-Striebel Smoothing

linear-time inference on Gaussian-Markov time series

$$p(x(t_i) \mid x(t_{i-1})) = \mathcal{N}(x_i; Ax_{i-1}, Q) \qquad p(y_t \mid x_t) = \mathcal{N}(y_t; Hx_t, R)$$



predict: $p(x_t \mid Y_{1:t-1}) = \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid Y_{0:t-1})\, dx_{t-1} = \mathcal{N}(x_t, m_t^-, P_t^-) = \mathcal{N}(x_t, Am_t, AP_tA^\intercal + Q)$

update: $p(x_t \mid Y_{1:t}) = \dfrac{p(y_t \mid x_t) p(x_t \mid Y_{0:t-1})}{p(y_t)} = \mathcal{N}(x_t, m_t, P_t) = \mathcal{N}(x_t, m_t^- + Kz, (I - KH)P_t^-)$

$$K := P_t^- H^\intercal (HP_tH^\intercal + R)^{-1}, \qquad z := y_t - Hm_t^-,$$

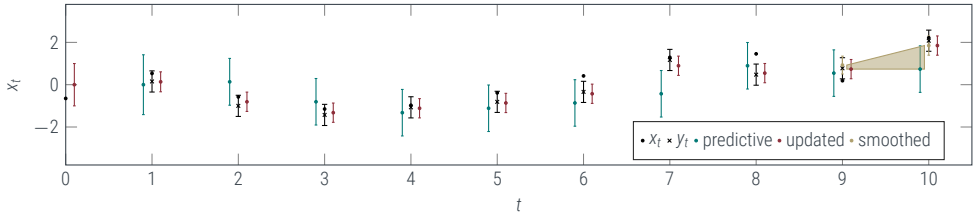
smooth: $p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \dfrac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{1:t})}\, dx_{t+1} = \mathcal{N}(x_t, m_t^s, P_t^s)$

$$= \mathcal{N}(x_t, m_t + G_t(m_{t+1}^s - m_{t+1}^-), P_t + G_t(P_{t+1}^s - P_{t+1}^-)G_t^\intercal) \qquad G_t := P_tA^\intercal (P_t^-)^{-1}$$

Kalman Filtering and Rauch-Tung-Striebel Smoothing

linear-time inference on Gaussian-Markov time series

$$p(x(t_i) \mid x(t_{i-1})) = \mathcal{N}(x_i; Ax_{i-1}, Q) \qquad p(y_t \mid x_t) = \mathcal{N}(y_t; Hx_t, R)$$



predict: $p(x_t \mid Y_{1:t-1}) = \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid Y_{0:t-1})\, dx_{t-1} = \mathcal{N}(x_t, m_t^-, P_t^-) = \mathcal{N}(x_t, Am_t, AP_tA^\intercal + Q)$

update: $p(x_t \mid Y_{1:t}) = \dfrac{p(y_t \mid x_t) p(x_t \mid Y_{0:t-1})}{p(y_t)} = \mathcal{N}(x_t, m_t, P_t) = \mathcal{N}(x_t, m_t^- + Kz, (I - KH)P_t^-)$

$$K := P_t^- H^\intercal (HP_tH^\intercal + R)^{-1}, \qquad z := y_t - Hm_t^-,$$

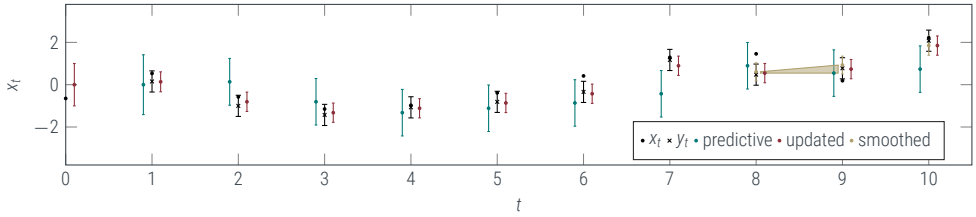
smooth: $p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \dfrac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{1:t})}\, dx_{t+1} = \mathcal{N}(x_t, m_t^s, P_t^s)$

$$= \mathcal{N}(x_t, m_t + G_t(m_{t+1}^s - m_{t+1}^-), P_t + G_t(P_{t+1}^s - P_{t+1}^-)G_t^\intercal) \qquad G_t := P_tA^\intercal (P_t^-)^{-1}$$

Kalman Filtering and Rauch-Tung-Striebel Smoothing

linear-time inference on Gaussian-Markov time series

$$p(x(t_i) \mid x(t_{i-1})) = \mathcal{N}(x_i; Ax_{i-1}, Q) \qquad p(y_t \mid x_t) = \mathcal{N}(y_t; Hx_t, R)$$



predict: $p(x_t \mid Y_{1:t-1}) = \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid Y_{0:t-1})\, dx_{t-1} = \mathcal{N}(x_t, m_t^-, P_t^-) = \mathcal{N}(x_t, Am_t, AP_tA^\mathsf{T} + Q)$

update: $p(x_t \mid Y_{1:t}) = \dfrac{p(y_t \mid x_t) p(x_t \mid Y_{0:t-1})}{p(y_t)} = \mathcal{N}(x_t, m_t, P_t) = \mathcal{N}(x_t, m_t^- + Kz, (I - KH)P_t^-)$

$$K := P_t^- H^\mathsf{T} (HP_tH^\mathsf{T} + R)^{-1}, \qquad z := y_t - Hm_t^-,$$

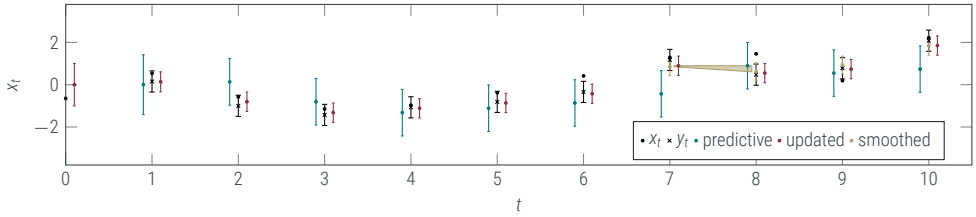
smooth: $p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \dfrac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{1:t})}\, dx_{t+1} = \mathcal{N}(x_t, m_t^s, P_t^s)$

$$= \mathcal{N}(x_t, m_t + G_t(m_{t+1}^s - m_{t+1}^-), P_t + G_t(P_{t+1}^s - P_{t+1}^-)G_t^\mathsf{T}) \qquad G_t := P_tA^\mathsf{T}(P_t^-)^{-1}$$

Kalman Filtering and Rauch-Tung-Striebel Smoothing

linear-time inference on Gaussian-Markov time series

$$p(x(t_i) \mid x(t_{i-1})) = \mathcal{N}(x_i; Ax_{i-1}, Q) \qquad p(y_t \mid x_t) = \mathcal{N}(y_t; Hx_t, R)$$



predict: $p(x_t \mid Y_{1:t-1}) = \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid Y_{0:t-1})\, dx_{t-1} = \mathcal{N}(x_t, m_t^-, P_t^-) = \mathcal{N}(x_t, Am_t, AP_tA^\mathsf{T} + Q)$

update: $p(x_t \mid Y_{1:t}) = \dfrac{p(y_t \mid x_t) p(x_t \mid Y_{0:t-1})}{p(y_t)} = \mathcal{N}(x_t, m_t, P_t) = \mathcal{N}(x_t, m_t^- + Kz, (I - KH)P_t^-)$

$$K := P_t^- H^\mathsf{T} (HP_tH^\mathsf{T} + R)^{-1}, \qquad z := y_t - Hm_t^-,$$

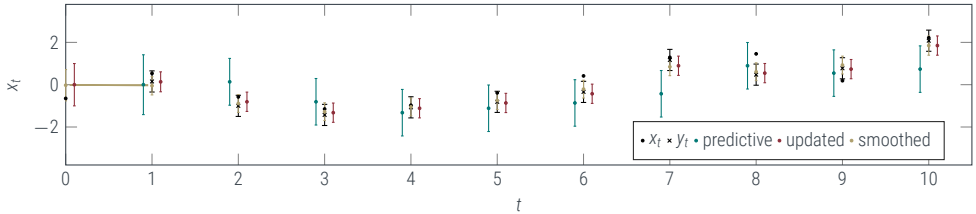
smooth: $p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \dfrac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{1:t})}\, dx_{t+1} = \mathcal{N}(x_t, m_t^s, P_t^s)$

$$= \mathcal{N}(x_t, m_t + G_t(m_{t+1}^s - m_{t+1}^-), P_t + G_t(P_{t+1}^s - P_{t+1}^-)G_t^\mathsf{T}) \qquad G_t := P_tA^\mathsf{T}(P_t^-)^{-1}$$

Kalman Filtering and Rauch-Tung-Striebel Smoothing

linear-time inference on Gaussian-Markov time series

$$p(x(t_i) \mid x(t_{i-1})) = \mathcal{N}(x_i; Ax_{i-1}, Q) \qquad p(y_t \mid x_t) = \mathcal{N}(y_t; Hx_t, R)$$



Filter:

$m_t^- = Am_{t-1}$ predictive mean

$P_t^- = AP_{t-1}A^\mathsf{T} + Q$ predictive covariance

$z_t = y - Hm_t^-$ innovation residual

$S_t = HP_t^- H^\mathsf{T} + R$ innovation covariance

$K_t = P_t^- H^\mathsf{T} S^{-1}$ Kalman gain

$m_t = m_t^- + Kz_t$ estimation mean

$P_t = (I - KH)P_t^-$ estimation covariance

Smoother:

$G_t = P_t A^\mathsf{T} (P_{t+1}^-)^{-1}$ RTS gain

$m_t^s = m_t + G_t(m_{t+1}^s - m_{t+1}^-)$ smoothed mean

$P_t^s = P_t + G_t(P_{t+1}^s - P_{t+1}^-)G^\mathsf{T}$ smoothed covariance

Time Series:

- **Markov Chains** formalize the notion of a stochastic process with a *local finite memory*
- Inference over Markov Chains separates into three operations, that can be performed in *linear* time.
- If all relationships are *linear* and *Gaussian*,

$$p(x(t_i) \mid x(t_{i-1})) = \mathcal{N}(x_i; Ax_{i-1}, Q) \qquad p(y_t \mid x_t) = \mathcal{N}(y_t; Hx_t, R)$$

then inference is analytic and given by the **Kalman Filter** and the **Rauch-Tung-Striebel Smoother**:

Filter:

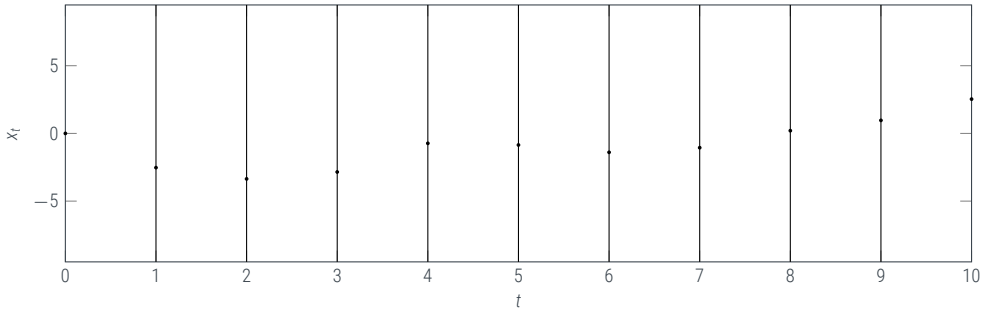
$$m_t^- = Am_{t-1} \quad \text{predictive mean}$$
$$P_t^- = AP_{t-1}A^\mathsf{T} + Q \quad \text{predictive covariance}$$
$$z_t = y - Hm_t^- \quad \text{innovation residual}$$
$$S_t = HP_t^- H^\mathsf{T} + R \quad \text{innovation covariance}$$
$$K_t = P_t^- H^\mathsf{T} S^{-1} \quad \text{Kalman gain}$$
$$m_t = m_t^- + Kz_t \quad \text{estimation mean}$$
$$P_t = (I - KH)P_t^- \quad \text{estimation covariance}$$

Smoother:

$$G_t = P_t A^\mathsf{T} (P_{t+1}^-)^{-1} \quad \text{RTS gain}$$
$$m_t^s = m_t + G_t(m_{t+1}^s - m_{t+1}^-) \quad \text{smoothed mean}$$
$$P_t^s = P_t + G_t(P_{t+1}^s - P_{t+1}^-)G^\mathsf{T} \quad \text{smoothed covariance}$$

Continuous Time

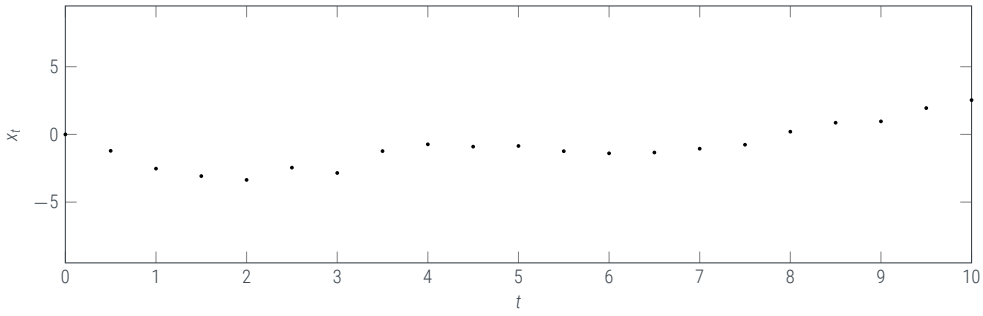
Differential equations defining non-differential curves



$$\delta t = 1 \qquad Q = 1$$

Continuous Time

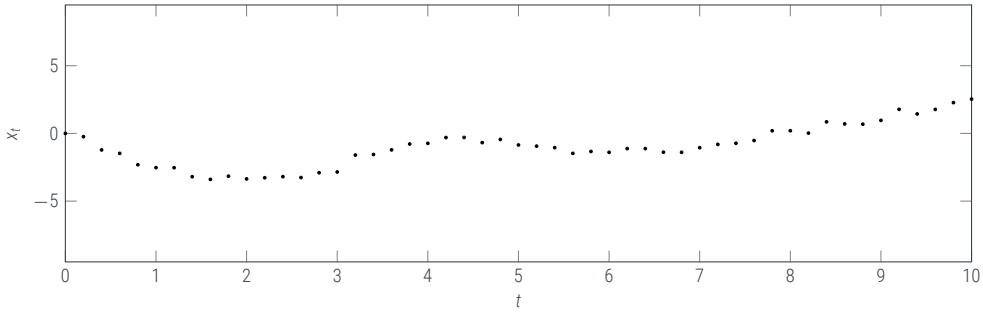
Differential equations defining non-differential curves



$$\delta t = 1/2 \qquad Q = 1/2$$

Continuous Time

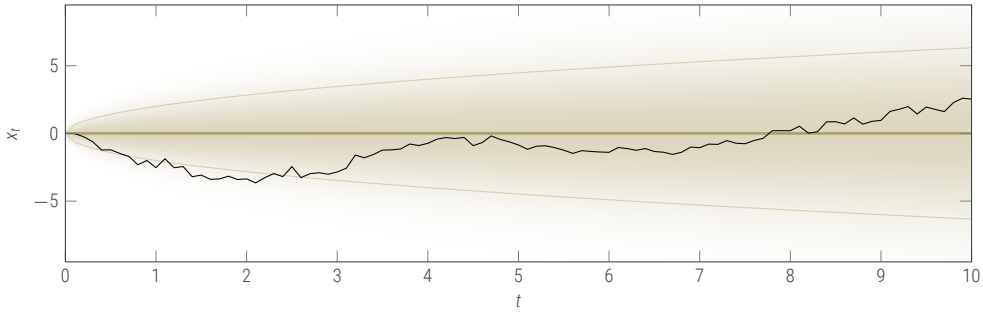
Differential equations defining non-differential curves



$$\delta t = 1/4 \qquad Q = \delta t$$

Continuous Time

Differential equations defining non-differential curves



$$\delta t \to 0 \qquad Q = ???$$

Stochastic Differential Equations

a pragmatic definition

For our purposes the (linear, time-invariant) **Stochastic Differential Equation**

$$dx(t) = Fx\,dt + L\,d\omega,$$

together with $x(t_0) = x_0$, describes the local behaviour of the (unique) Gaussian process with

$$\mathbb{E}(x(t)) =: m(t) = e^{F(t-t_0)}x_0 \qquad \operatorname{cov}(x(t_a), x(t_b)) =: k(t_a, t_b) = \int_{t_0}^{\min t_a, t_b} e^{F(t_a-\tau)} L L^\intercal e^{F^\intercal (t_b-\tau)}\,d\tau$$

This GP is known as the **complete solution** of the SDE. It gives rise to the discrete-time stochastic recurrence relation $p(x_{t_{i+1}} \mid x_{t_i}) = \mathcal{N}(x_{t_{i+1}}; A_{t_i} x_{t_i}, Q_{t+i})$ with

$$A_{t_i} = e^{F(t_{i+1}-t_i)} \quad \text{and} \quad Q_{t_i} = \int_0^{t_{i+1}-t_i} e^{F\tau} L L^\intercal e^{F\tau}\,d\tau.$$

Matrix exponential: $e^X := \sum_{i=0}^{\infty} \frac{X^i}{i!}$. *Thus :* $e^0 = I$, $(e^X)^{-1} = e^{-X}$, $X = VDV^{-1} \Rightarrow Ve^D V^{-1}$, $e^{\operatorname{diag}_i d_i} = \operatorname{diag}_i e^{d_i}$, $\det e^X = e^{\operatorname{tr} X}$.

The Connection to GPs

Some well-studied examples

$$dx(t) = Fx\,dt + L\,d\omega$$

$$\mathbb{E}(x(t)) =: m(t) = e^{F(t-t_0)}x_0 \qquad \operatorname{cov}(x(t_a), x(t_b)) =: k(t_a, t_b) = \int_{t_0}^{\min t_a, t_b} e^{F(t_a-\tau)} L L^\intercal e^{F^\intercal (t_b-\tau)}\, d\tau$$

$$A_{t_i} = e^{F(t_{i+1}-t_i)} \qquad Q_{t_i} = \int_0^{t_{i+1}-t_i} e^{F\tau} L L^\intercal e^{F^\intercal \tau}\, d\tau$$

The Connection to GPs

Some well-studied examples

$$dx(t) = Fx\,dt + L\,d\omega$$

$$\mathbb{E}(x(t)) =: m(t) = e^{F(t-t_0)}x_0 \qquad \operatorname{cov}(x(t_a), x(t_b)) =: k(t_a, t_b) = \int_{t_0}^{\min t_a, t_b} e^{F(t_a-\tau)} L L^\intercal e^{F^\intercal (t_b-\tau)}\, d\tau$$

$$A_{t_i} = e^{F(t_{i+1}-t_i)} \qquad Q_{t_i} = \int_0^{t_{i+1}-t_i} e^{F\tau} L L^\intercal e^{F\tau}\, d\tau$$

The Wiener process

$$F = 0, L = \theta \quad \Rightarrow \quad m(t) = x_0 \quad k(t_a, t_b) = \theta^2(\min(t_a, t_b) - t_0)$$
$$A = I \qquad Q_{t_i} = \theta^2(t_{i+1} - t_i)$$

The Connection to GPs

Some well-studied examples

$$dx(t) = Fx\,dt + L\,d\omega$$

$$\mathbb{E}(x(t)) =: m(t) = e^{F(t-t_0)}x_0 \qquad \operatorname{cov}(x(t_a), x(t_b)) =: k(t_a, t_b) = \int_{t_0}^{\min t_a, t_b} e^{F(t_a-\tau)} L L^\intercal e^{F^\intercal (t_b-\tau)}\, d\tau$$

$$A_{t_i} = e^{F(t_{i+1}-t_i)} \qquad Q_{t_i} = \int_0^{t_{i+1}-t_i} e^{F\tau} L L^\intercal e^{F\tau}\, d\tau$$

The Ornstein-Uhlenbeck process

$$F = -\frac{1}{\lambda}, L = \frac{2\theta}{\sqrt{\lambda}} \quad \Rightarrow \quad m(t) = x_0 e^{-\frac{t-t_0}{\lambda}} \qquad k(t_a, t_b) = \theta^2 \left(e^{-\frac{|t_a - t_b|}{\lambda}} - e^{\frac{2t_0 - t_a - t_b}{\lambda}} \right)$$

$$(\min(a,b) = \frac{1}{2}(a + b - |a-b|)) \qquad A = e^{-\delta_t/\lambda} \qquad Q_{t_i} = \theta^2 \left(1 - e^{-2\delta_t/\lambda}\right)$$

Code

`Gauss-Markov-Inference.ipynb`

Hidden Markov Models

Generalization to non-Gaussian models

Name	Distribution	Algorithm
Markovian System:	$p(\boldsymbol{y}, \boldsymbol{x}) = \prod_{i=0}^{N} p(x_i \mid x_{i-1}) p(y_i \mid x_i)$	filter
Linear Gaussian System:	$p(\boldsymbol{y}, \boldsymbol{x}) = \prod_{i=0}^{N} \mathcal{N}(x_i; A_i x_{i-1}, Q_i) \mathcal{N}(y_i; H x_i, R)$	Kalman filter
Nonlinear Gaussian System:	$p(\boldsymbol{y}, \boldsymbol{x}) = \prod_{i=0}^{N} \mathcal{N}(x_i; a(x_{i-1}), Q_i) \mathcal{N}(y_i; h(x_i), R)$	e.g. Extended/Unscented/Particle filter etc.
Non-Gaussian observations:	$p(\boldsymbol{y}, \boldsymbol{x}) = \prod_{i=0}^{N} \mathcal{N}(x_i; A_i x_{i-1}, Q_i) p(y_i \mid h(x_i))$	**next lecture**
Hidden Markov Model (e.g.):	$p(\boldsymbol{y}, \boldsymbol{x}) = \prod_{i=0}^{N} p(x_i = \Pi x_{i-1}) \mathcal{N}(y_i; h(x_i), R)$	Markov Chain Monte Carlo (later)

- For continuous systems with nonlinear dynamics and/or non-linear observations, a number of *approximately Gaussian filters* have been developed. For more see, e.g.
  - Simo Särkkä. *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013

    `https://users.aalto.fi/~ssarkka/pub/cup_book_online_20131111.pdf`

Because streaming data is a common data type, time series are an entire sub-field of their own, studied in a diverse range of domains. There is no time to cover them all in one lecture.

Instead of focussing on non-Gaussian time series, we will instead give up Markovian structure for the moment, and first focus on **non-Gaussian observations** next lecture

Summary:

Markov Chains capture **finite memory** of a time series through conditional independence

Gauss-Markov models map this state to linear algebra

Kalman filter is the name for the corresponding algorithm

SDEs (Stochastic Differential Equations) are the continuous-time limit

Complexity of all necessary operations is **linear**, $\mathcal{O}(N)$ in the number of datapoints (as opposed to $\mathcal{O}(N^3)$ for general GPs). This includes hyperparameter inference!

Steady-State solutions amount to various forms of **running averages**

HMMs are the generalisation to non-Gaussian models