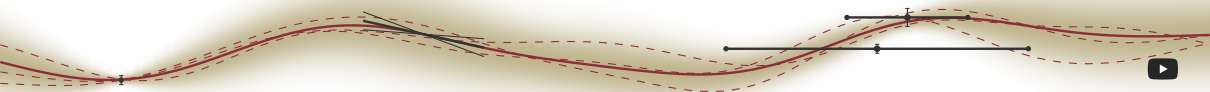
Probabilistic Machine Learning
Lecture 12
Gauss-Markov Models

Philipp Hennig

26 May 2020

Eberhard Karls
Universität
Tübingen

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

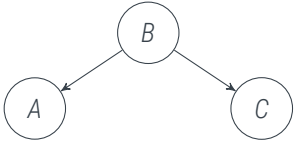
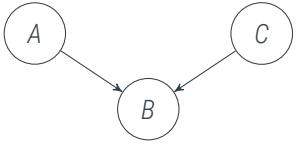


#	date	content	Ex	#	date	content	Ex
1	20.04.	Introduction	1	14	09.06.	Logistic Regression	8
2	21.04.	Reasoning under Uncertainty		15	15.06.	Exponential Families	
3	27.04.	Continuous Variables	2	16	16.06.	Graphical Models	9
4	28.04.	Monte Carlo		17	22.06.	Factor Graphs	
5	04.05.	Markov Chain Monte Carlo	3	18	23.06.	The Sum-Product Algorithm	10
6	05.05.	Gaussian Distributions		19	29.06.	Example: Topic Models	
7	11.05.	Parametric Regression	4	20	30.06.	Mixture Models	11
8	12.05.	Learning Representations		21	06.07.	EM	
9	18.05.	Gaussian Processes	5	22	07.07.	Variational Inference	12
10	19.05.	Understanding Kernels		23	13.07.	Example: Topic Models	
11	25.05.	An Example for GP Regression	6	24	14.07.	Example: Inferring Topics	13
12	26.05.	**Gauss-Markov Models**		25	20.07.	Example: Kernel Topic Models	
13	08.06.	GP Classification	7	26	21.07.	Revision	

Recap: Atomic Independence Structures

From Lecture 2

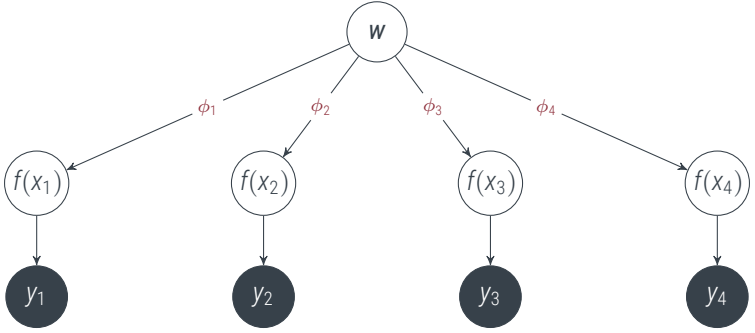
For uni- and bi-variate graphs, conditional independence is trivial.
For tri-variate sub-graphs, there are three possible structures:

	graph	factorization	implications
(i)		$P(A, B, C) = P(C \mid B) \cdot P(B \mid A) \cdot P(A)$	$A \perp\!\!\!\perp C \mid B$ but not, i.e., $A \not\perp\!\!\!\perp C$
(ii)		$P(A, B, C) = P(A \mid B) \cdot P(C \mid B) \cdot P(B)$	$A \perp\!\!\!\perp C \mid B$ but not, i.e., $A \not\perp\!\!\!\perp C$
(iii)		$P(A, B, C) = P(B \mid A, C) \cdot P(C) \cdot P(A)$	$A \perp\!\!\!\perp C$ but not, i.e., $A \not\perp\!\!\!\perp C \mid B$

Graphical View: Parametric Model

Conditional independence of data given model weights

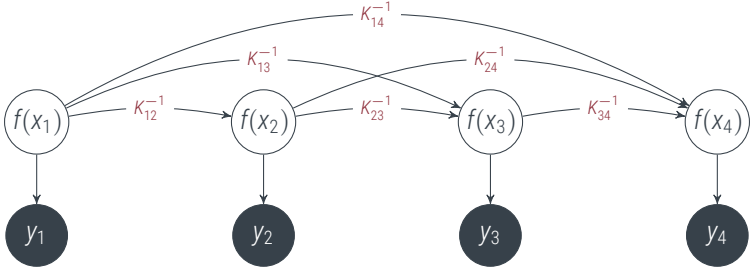
$$p(f) = \mathcal{GP}(f; 0, \Phi_X^\intercal \Sigma \Phi_X) \quad p\left(\begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix} \middle| \, \boldsymbol{w} \right) = \prod_i \delta(f_i - \phi_i^\intercal \boldsymbol{w}) \quad p(\boldsymbol{y} \mid \boldsymbol{f}) = \prod_i \mathcal{N}(y_i; f_i, \sigma^2)$$



Graphical View: Nonparametric Model

Fully connected graph

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



Time Series

a widely applicable concept

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

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:

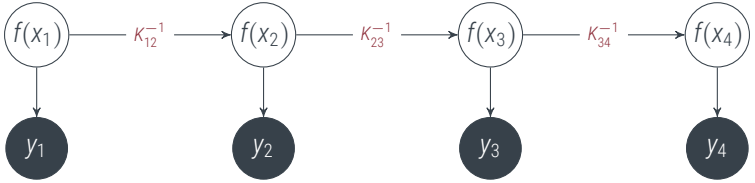
- climate & weather observations … — in Climate Science
- sensor readings in cars, … — in Engineering
- EEG, ECG, patch clamp signals, … — in Medicine and Neuroscience
- just about any sensing of a dynamical process — in Physics
- stock prices, supply & demand data, polling numbers, … — in Economics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

Markov Chains

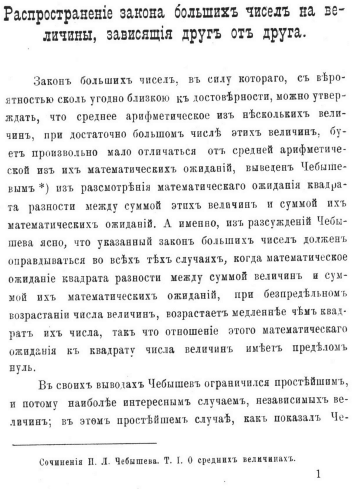
Processes with a "local memory"

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



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

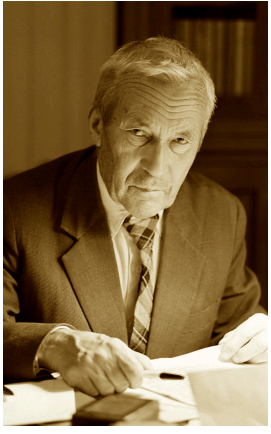
Andrej Andreevič Markov
(1856 – 1922)

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

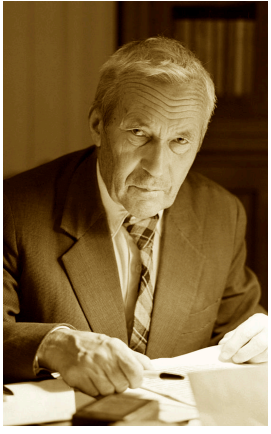
This point is way to easily missed

[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)

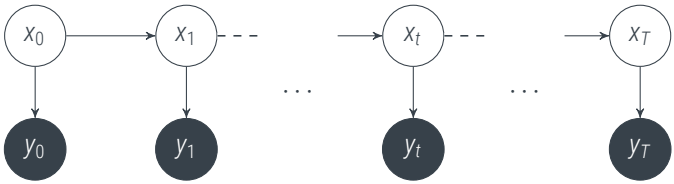
Definition: A joint distribution $p(X)$ over a sequence of random variabels $X := [x_0, \dots, x_N]$ is said to have **the Markov property** if

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

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

Assume:

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

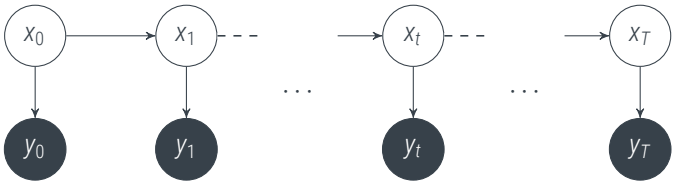


$$p(x_t \mid Y_{0:t-1}) = \frac{\int_{j\neq t} p(X)p(Y_{0:t-1} \mid X)\,dx_j}{\int p(X)p(Y_{0:t-1} \mid X)\,dX} = \frac{\int_{j\neq t} p(Y_{0:t-1} \mid X_{0:t-1})p(x_0)\left(\prod_{0<j<t} p(x_j \mid x_{j-1})\,dx_j\right)p(x_t \mid x_{t-1})\left(\prod_{j>t} p(x_j \mid x_{j-1})\,dx_j\right)}{\int p(Y_{0:t-1}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

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

Assume:

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



If you believe the graph, though, just note that the joint is

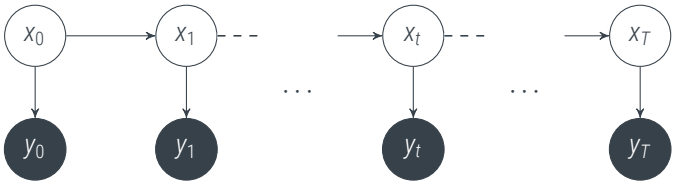
$$p(x_t, x_{t-1}|y_{1:t-1}) = p(x_t|x_{t-1}, y_{1:t-1})p(x_{t-1}|y_{1:t-1}) = p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})$$

which we can integrate over x_{t-1} to obtain the **Chapman-Kolmogorov** equation

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

Assume:

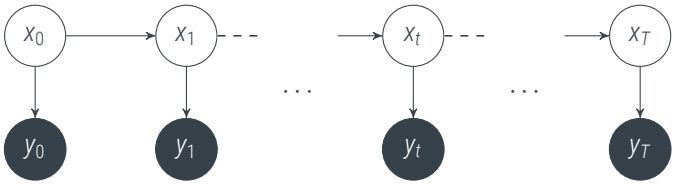
$$p(x_t \mid X_{0:t-1}) = p(x_t \mid x_{t-1})$$
$$\text{and} \quad p(y_t \mid X) = p(y_t \mid x_t)$$



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

Assume:

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



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

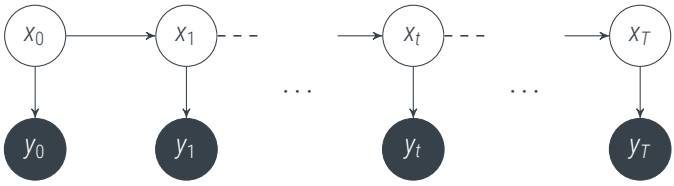
$$p(x_t \mid x_{t+1}, Y) = \frac{p(Y_{t+1:n} \mid x_{t+1}, x_t, Y_{0:t}) p(x_t \mid x_{t+1}, Y_{0:t})}{\int p(Y_{t+1:n} \mid x_{t+1}, x_t, Y_{0:t}) p(x_t \mid x_{t+1}, Y_{0:t})\, dx_t} = \frac{p(Y_{t+1:n} \mid x_{t+1}, Y_{0:t}) \cdot p(x_t \mid x_{t+1}, Y_{0:t})}{p(Y_{t+1:n} \mid x_{t+1}, Y_{0:t}) \cdot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

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

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

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

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

$$p(x(t_{i+1}) \mid X_{1:i}) = \mathcal{N}(x_{i+1}; Ax_i, Q) \quad \text{and} \quad p(x_0) = \mathcal{N}(x_0; m_0, P_0) \quad \text{and} \quad p(y_i \mid X) = \mathcal{N}(y_i; Hx_i, R)$$

predict:

$$\begin{aligned} p(x_t \mid Y_{1:t-1}) &= \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid Y_{1:t-1})\, dx_{t-1} \\ &= \int \mathcal{N}(x_t; Ax_{t-1}, Q) \cdot \mathcal{N}(x_{t-1}; m_{t-1}, P_{t-1})\, dx_{t-1} \\ &= \mathcal{N}(x_t, Am_{t-1}, AP_{t-1}A^\intercal + Q) \\ &= \mathcal{N}(x_t, m_t^-, P_t^-) \end{aligned}$$

$$p(x(t_{i+1}) \mid X_{1:i}) = \mathcal{N}(x_{i+1}; Ax_i, Q) \quad \text{and} \quad p(x_0) = \mathcal{N}(x_0; m_0, P_0) \quad \text{and} \quad p(y_i \mid X) = \mathcal{N}(y_i; Hx_i, R)$$

update:

$$\begin{aligned}
p(x_t \mid Y_{1:t}) &= \frac{p(y_t \mid x_t) p(x_t \mid Y_{1:t-1})}{p(y_t)} \\
&= \frac{\mathcal{N}(y_t; Hx_t; R)\mathcal{N}(x_t; m_t^-, P_t^-)}{\mathcal{N}(y_t; Hm_t^-, HP_t^- H^\intercal)} \\
&= \mathcal{N}(x_t, m_t^- + Kz, (I - KH)P_t^-) \\
&= \mathcal{N}(x_t, m_t, P_t) \quad \text{where} \\
K &:= P_t^- H^\intercal (HP_t H^\intercal + R)^{-1}, \quad \text{(gain)} \\
z &:= y_t - Hm_t^- \quad \text{(residual)}
\end{aligned}$$

$$p(x(t_{i+1}) \mid X_{1:i}) = \mathcal{N}(x_{i+1}; Ax_i, Q) \quad \text{and} \quad p(x_0) = \mathcal{N}(x_0; m_0, P_0) \quad \text{and} \quad p(y_i \mid X) = \mathcal{N}(y_i; Hx_i, R)$$

smooth:

$$\begin{aligned}
p(x_t \mid Y) &= p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \frac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{1:t})} \, dx_{t+1} \\
&= \mathcal{N}(x_t; m_t, P_t) \int \mathcal{N}(x_{t+1}, Ax_t, Q) \frac{\mathcal{N}(x_{t+1}; m^s_{t+1}, P^{t+1}_s)}{\mathcal{N}(x_{t+1}; m_{t+1}, P_{t+1})} \, dx_{t+1} \\
&= \mathcal{N}(x_t, m_t + G_t(m^s_{t+1} - m^-_{t+1}), P_t + G_t(P^s_{t+1} - P^-_{t+1})G_t^\intercal) \\
&= \mathcal{N}(x_t, m^s_t, P^s_t) \quad \text{where} \\
G_t &:= P_t A^\intercal (P^-_t)^{-1} \quad \text{(smoother gain)}
\end{aligned}$$

Gauss-Markov Models

Local structure for univariate Gaussian models

$$p(x(t_{i+1}) \mid X_{1:i}) = \mathcal{N}(x_{i+1}; Ax_i, Q) \quad \text{and} \quad p(x_0) = \mathcal{N}(x_0; m_0, P_0) \quad \text{and} \quad p(y_i \mid X) = \mathcal{N}(y_i; Hx_i, R)$$

(Kalman) Filter:

$$p(x_t) = \mathcal{N}(x_t; m_t^-, P_t^-) \quad \text{with}$$
$$m_t^- = Am_{t-1} \quad \text{predictive mean}$$
$$P_t^- = AP_{t-1}A^\intercal + Q \quad \text{predictive covariance}$$
$$p(x_t \mid y_t) = \mathcal{N}(x_t; m_t, P_t) \quad \text{with}$$
$$z_t = y_t - Hm_t^- \quad \text{innovation residual}$$
$$S_t = HP_t^-H^\intercal + R \quad \text{innovation covariance}$$
$$K_t = P_t^-H^\intercal S^{-1} \quad \text{Kalman gain}$$
$$m_t = m_t^- + Kz_t \quad \text{estimation mean}$$
$$P_t = (I - KH)P_t^- \quad \text{estimation covariance}$$

(Rauch Tung Striebel) Smoother:

$$p(x_t \mid Y) = \mathcal{N}(x_t; m_t^s, P_t^s) \quad \text{with}$$
$$G_t = P_tA^\intercal(P_{t+1}^-)^{-1} \quad \text{RTS gain}$$
$$m_t^s = m_t + G_t(m_{t+1}^s - m_{t+1}^-) \quad \text{smoothed mean}$$
$$P_t^s = P_t + G_t(P_{t+1}^s - P_{t+1}^-)G^\intercal \quad \text{smoothed covariance}$$

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**:

(Kalman) Filter:

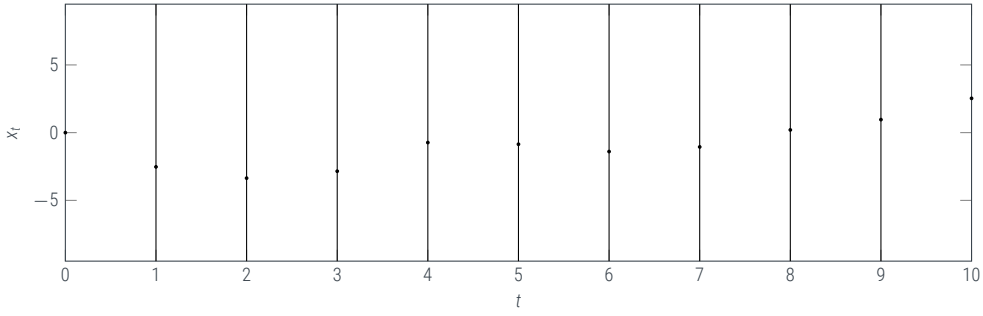
$p(x_t) = \mathcal{N}(x_t; m_t^-, P_t^-)$	with
$m_t^- = Am_{t-1}$	predictive mean
$P_t^- = AP_{t-1}A^\intercal + Q$	predictive covariance
$p(x_t \mid y_t) = \mathcal{N}(x_t; m_t, P_t)$	with
$z_t = y_t - Hm_t^-$	innovation residual
$S_t = HP_t^- H^\intercal + R$	innovation covariance
$K_t = P_t^- H^\intercal S^{-1}$	Kalman gain
$m_t = m_t^- + Kz_t$	estimation mean
$P_t = (I - KH)P_t^-$	estimation covariance

(Rauch Tung Striebel) Smoother:

$p(x_t \mid Y) = \mathcal{N}(x_t; m_t^s, P_t^s)$	with
$G_t = P_t A^\intercal (P_{t+1}^-)^{-1}$	RTS gain
$m_t^s = m_t + G_t(m_{t+1}^s - m_{t+1}^-)$	smoothed mean
$P_t^s = P_t + G_t(P_{t+1}^s - P_{t+1}^-)G^\intercal$	smoothed covariance

Continuous Time

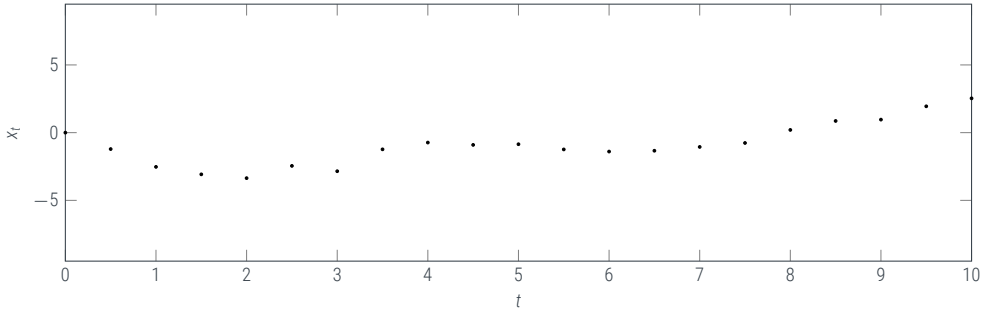
Differential equations defining non-differential curves



$$\delta t = 1 \qquad Q_{\delta t} = 1$$

Continuous Time

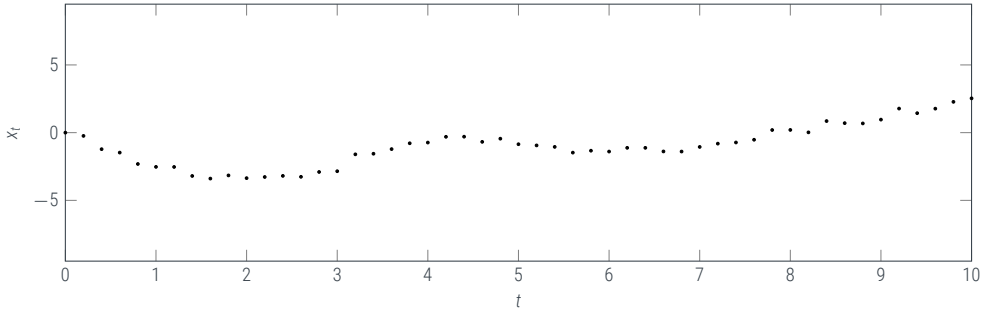
Differential equations defining non-differential curves



$$\delta t = 1/2 \qquad Q_{\delta t} = 1/2$$

Continuous Time

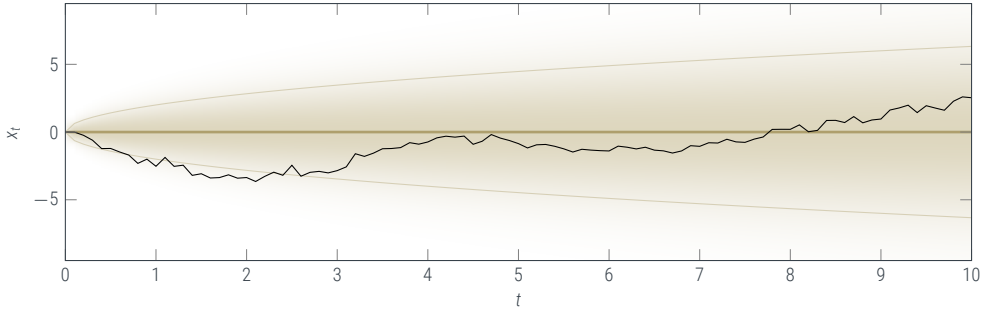
Differential equations defining non-differential curves



$$\delta t = 1/4 \qquad Q_{\delta t} = \delta t$$

Continuous Time

Differential equations defining non-differential curves



$$\delta t \to 0 \qquad Q_{\delta t} = ???$$

For the limit $\delta t \to 0$ we would like to encode that $Q_\delta t / \delta t$ approaches some kind of finite object (like a derivative, but sample paths from this (the Wiener) process) are almost surely not differentiable. So we introduce a new object: $Q_{dt} := d\omega$, known as the *Wiener measure*. (Nb: This is a non-standard construction. $d\omega$ can be defined more elegantly; but this goes beyond the scope of this course.)

Stochastic Differential Equations

a pragmatic definition

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

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

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 **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

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

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

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

The Connection to GPs

Some well-studied examples

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

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

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

The scaled Wiener process

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

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

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

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

The Ornstein-Uhlenbeck process

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

$$A = e^{-\delta_t/\lambda} \qquad Q_{t_i} = \theta^2 \left(1 - e^{-2\delta_t/\lambda}\right)$$

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

- So far, we have seen examples with $x(t) \in \mathbb{R}$.
- But F and L can also be matrices. Consider the example

$$x = \begin{bmatrix} x_{(1)} \\ x_{(2)} \end{bmatrix} \qquad F = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad L = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

That is:

$$\begin{bmatrix} dx_{(1)}(t) \\ dx_{(2)}(t) \end{bmatrix} = \begin{bmatrix} x_{(2)}(t)\,dt + 0\,d\omega \\ 0\,dt + d\omega \end{bmatrix} \quad \Rightarrow \quad x_{(1)}(t) = \int_{t_0}^{t} x_{(2)}(t)\,dt + [x_0]_1$$

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 of discrete-time stochastic recurrence relations (in particular, linear SDEs are the continuous-time generalization discrete-time linear Gaussian systems)

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).
(Although not shown, this includes hyperparameter inference!)

For more on Gaussian and *approximately Gaussian filters* see, e.g.
Simo Särkkä. *Bayesian Filtering and Smoothing* Cambridge University Press, 2013