Probabilistic Numerics

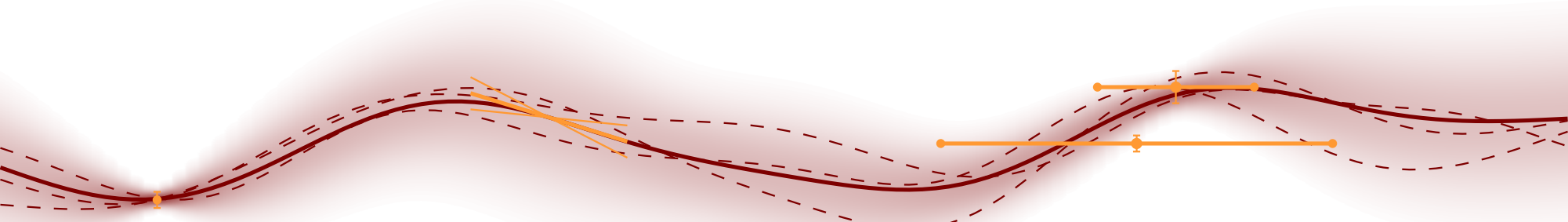
II – Filtering and Quadrature

Philipp Hennig

Dobbiaco Summer School
20 June 2017

Research Group for Probabilistic Numerics
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Some of the presented work was supported by
the Emmy Noether Programme of the DFG



Quick recap:

- **probabilistic inference** is **the** framework to consistently handle (epistemic and aleatory) uncertainty
- **Gaussian measures** turn inference into linear algebra
- **Gaussian processes** allow inference on **functions**

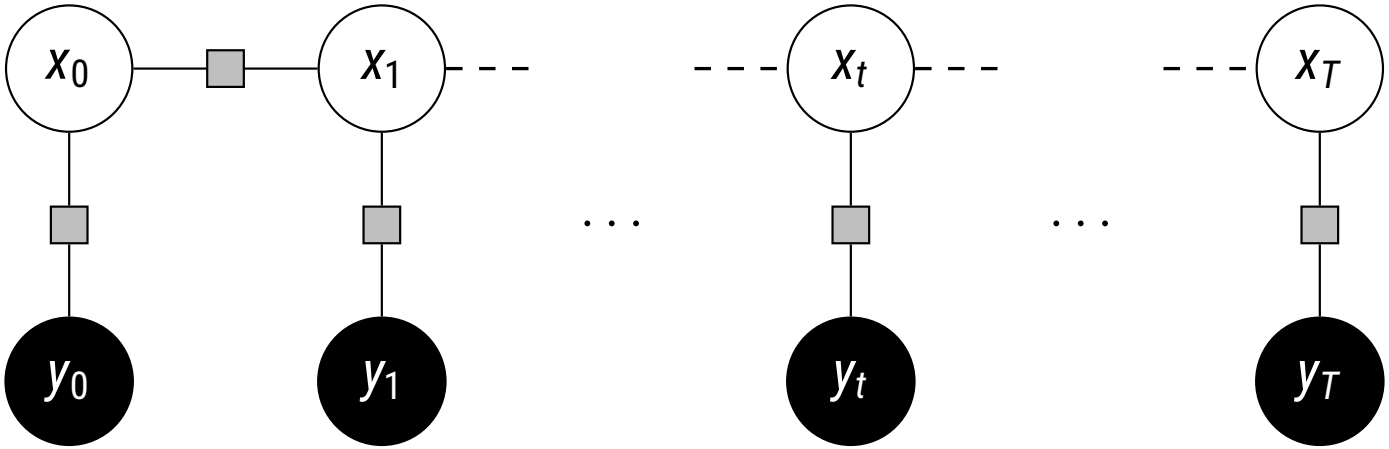
now:

- **filtering**: fast (linear time) inference for univariate GPs
- **quadrature**: a first probabilistic algorithm
- solving **ordinary differential equations**: a first “nonlinear” problem

A Markov Chain

local structure in ordered spaces

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



$$p(x_t \mid Y_{0:t-1}) = \frac{\int_{j \neq t} p(X)p(Y_{0:t-1} \mid X)\, dx_j}{\int p(X)p(Y_{0:t-1} \mid X)\, dX}$$

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

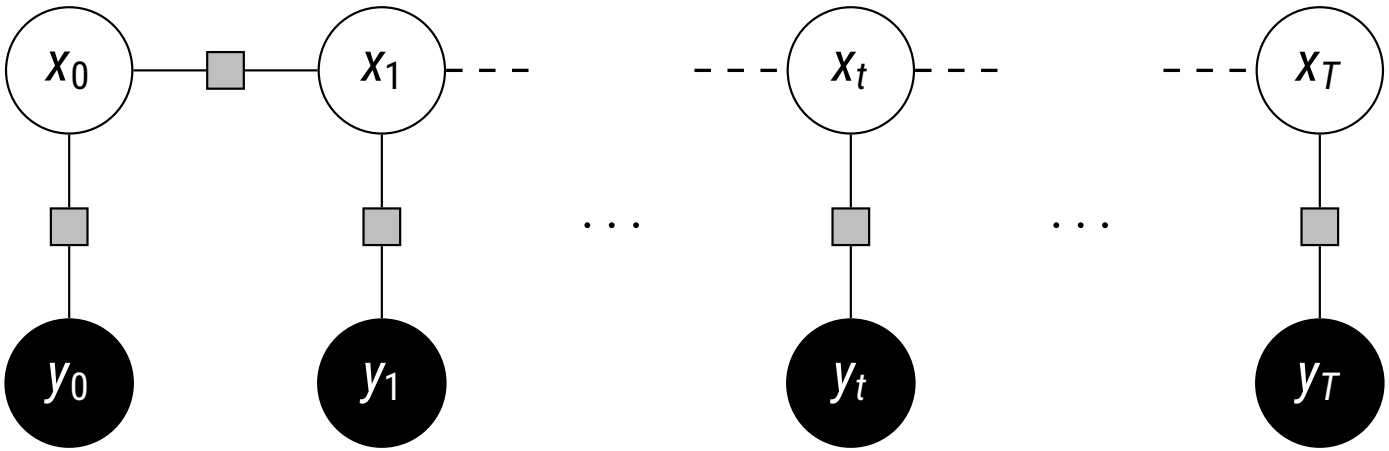
$$= \frac{\int_{j<t} p(x_t \mid x_{t-1})p(Y_{0:t-1} \mid X_{0:t-1})p(x_0) \left(\prod_{0<j<t} p(x_j \mid x_{j-1})\, dx_j\right)}{\int_{j<t} p(Y_{0:t-1} \mid X_{0:t-1})p(x_0) \left(\prod_{0<j<t} p(x_j \mid x_{j-1})\, dx_j\right)} = \int p(x_t \mid x_{t-1})p(x_{t-1} \mid Y_{0:t-1})\, dx_{t-1}$$

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

A Markov Chain

local structure in ordered spaces

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



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

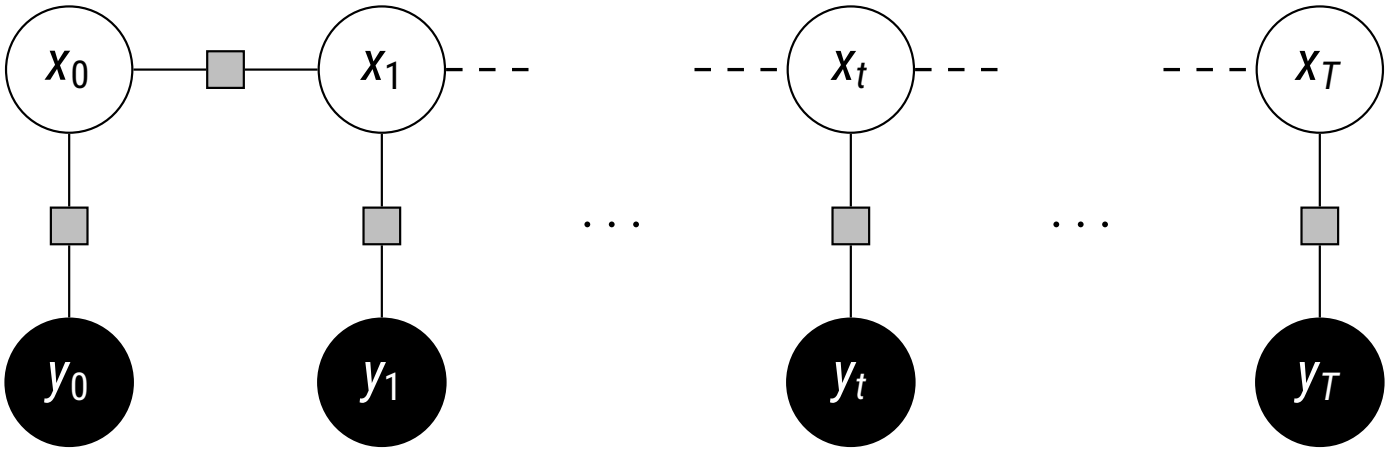
$$p(x_t \mid x_{t+1}, Y_{0:t}) = \frac{p(x_t, x_{t+1} \mid Y_{0:t})}{p(x_{t+1} \mid Y_{0:t})} = \frac{p(x_{t+1} \mid x_t, Y_{0:t}) p(x_t \mid Y_{0:t})}{p(x_{t+1} \mid Y_{0:t})} = \frac{p(x_{t+1} \mid x_t) p(x_t \mid Y_{0:t})}{p(x_{t+1} \mid Y_{0:t})}$$

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

A Markov Chain

local structure in ordered spaces

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



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

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

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

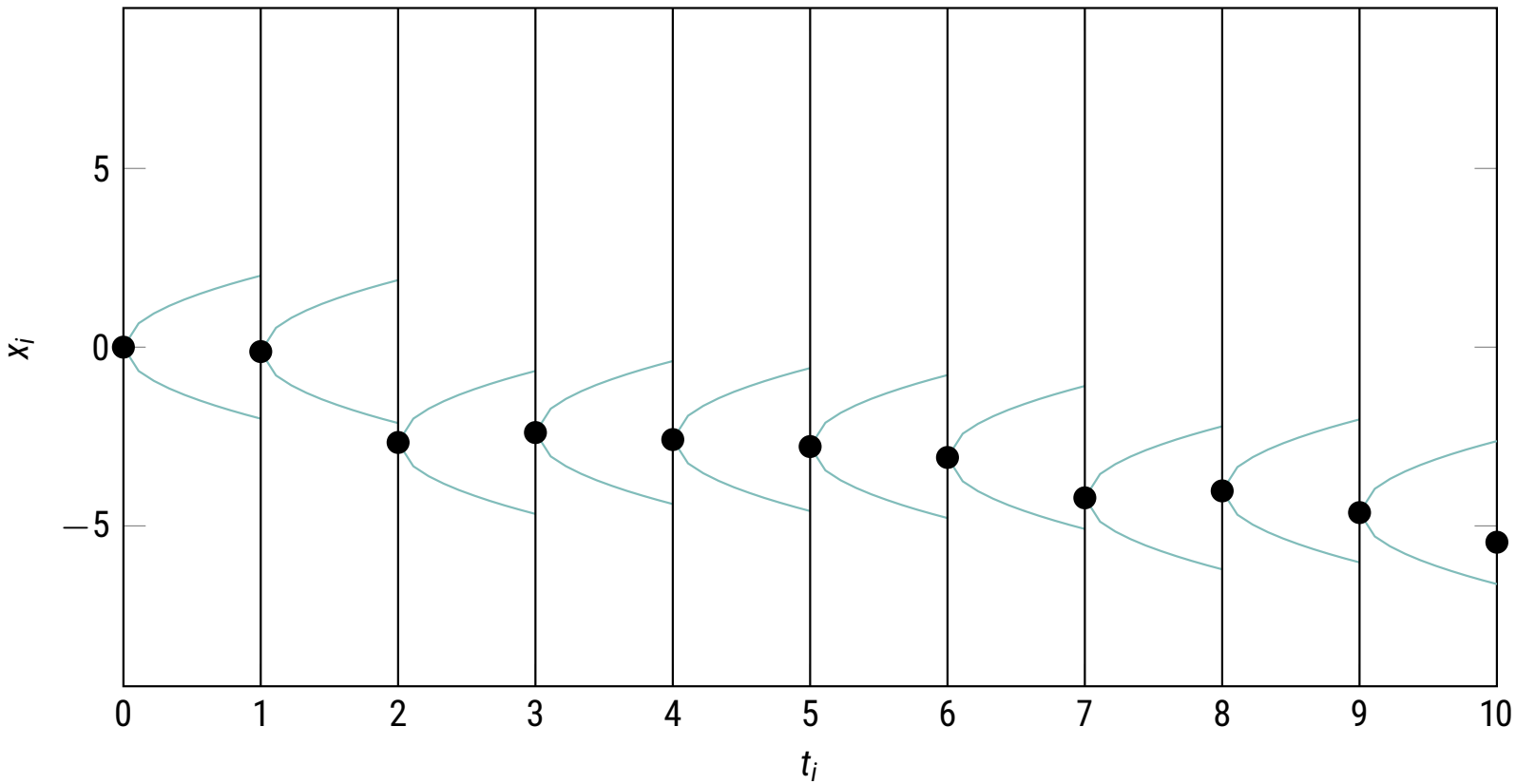
Smoothing: $\mathcal{O}(T)$

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

Gauss-Markov Models

Local structure for univariate Gaussian models

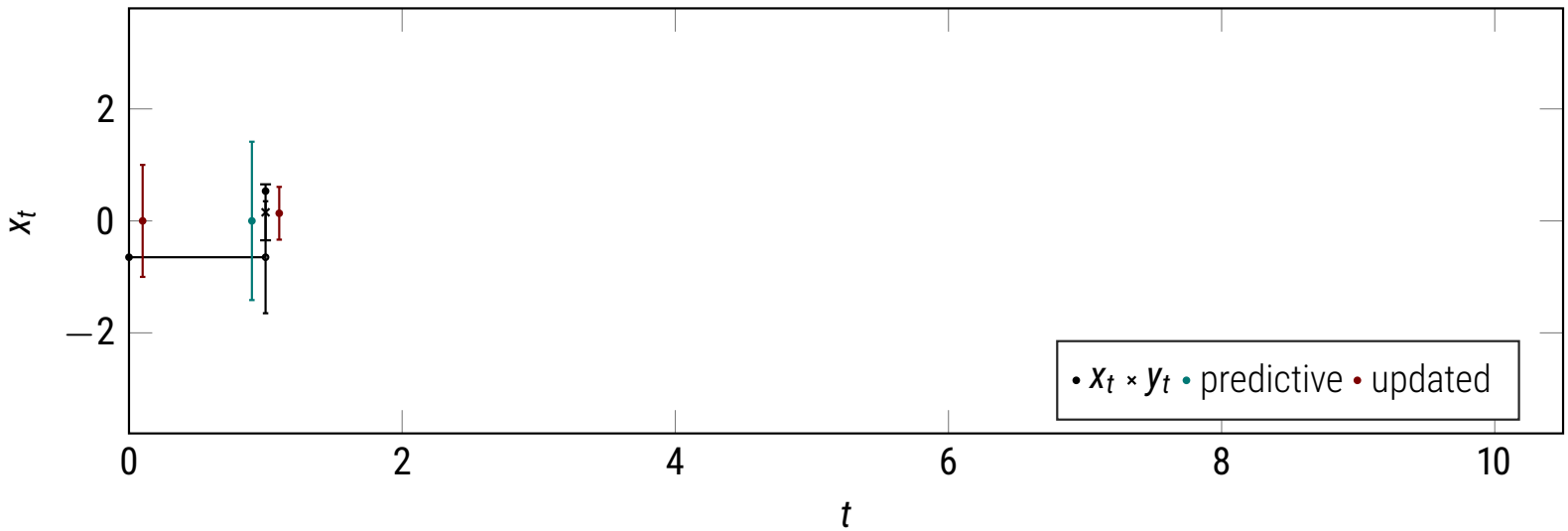
$$p(x(t_{i+1}) \mid x(t_i)) = \mathcal{N}(x_{i+1}; Ax_i, Q) \qquad \text{(figure: } x_0 = 0, A = Q = 1\text{)}$$



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

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

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

$$G_t = P_tA^\mathsf{T} (P_t^-)^{-1},$$

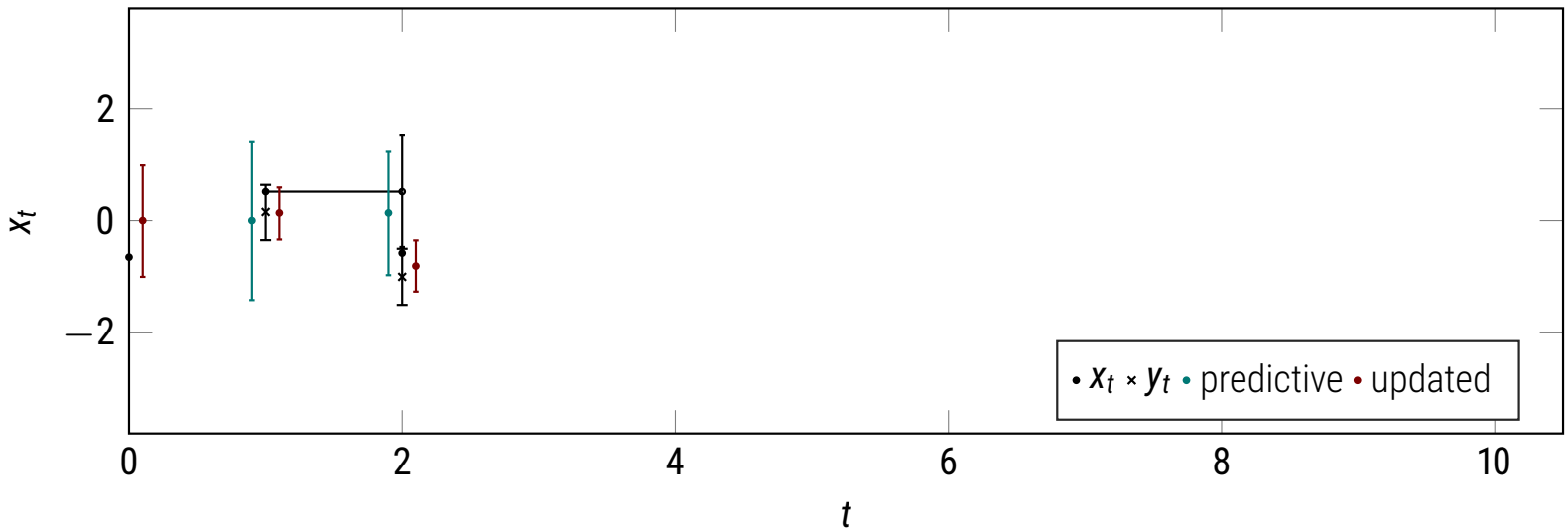
smooth: $$p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \frac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{1:t})}\, dx_{t+1} = \mathcal{N}(x_t, m_t^s, P_t^s)$$

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

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

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

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

$$G_t = P_tA^\mathsf{T} (P_t^-)^{-1},$$

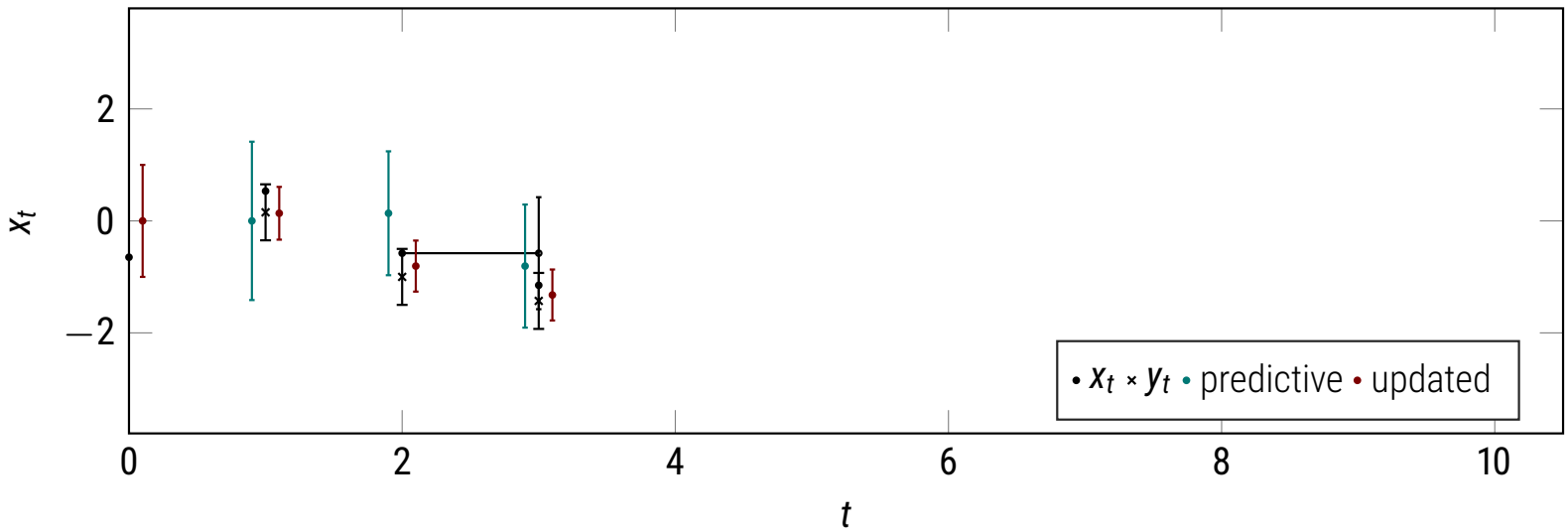
smooth: $$p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \frac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{1:t})}\, dx_{t+1} = \mathcal{N}(x_t, m_t^s, P_t^s)$$

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

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

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

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

$$G_t = P_tA^\mathsf{T} (P_t^-)^{-1},$$

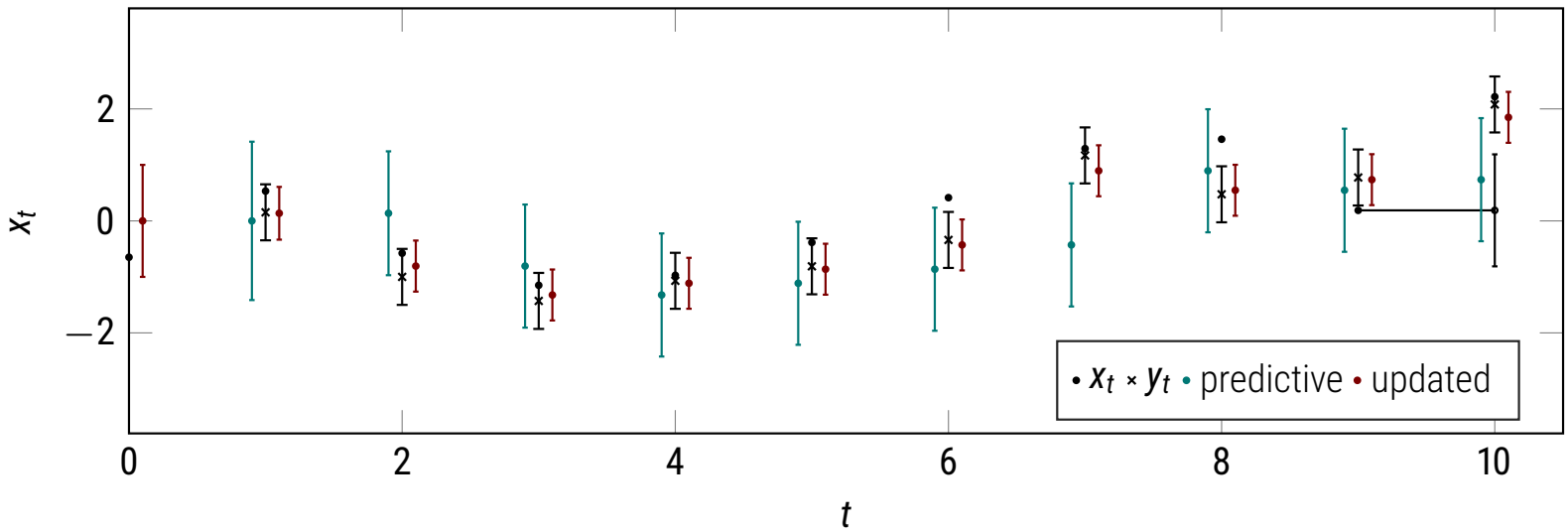
smooth: $$p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \frac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{1:t})}\, dx_{t+1} = \mathcal{N}(x_t, m_t^s, P_t^s)$$

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

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

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

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

$$G_t = P_tA^\mathsf{T} (P_t^-)^{-1},$$

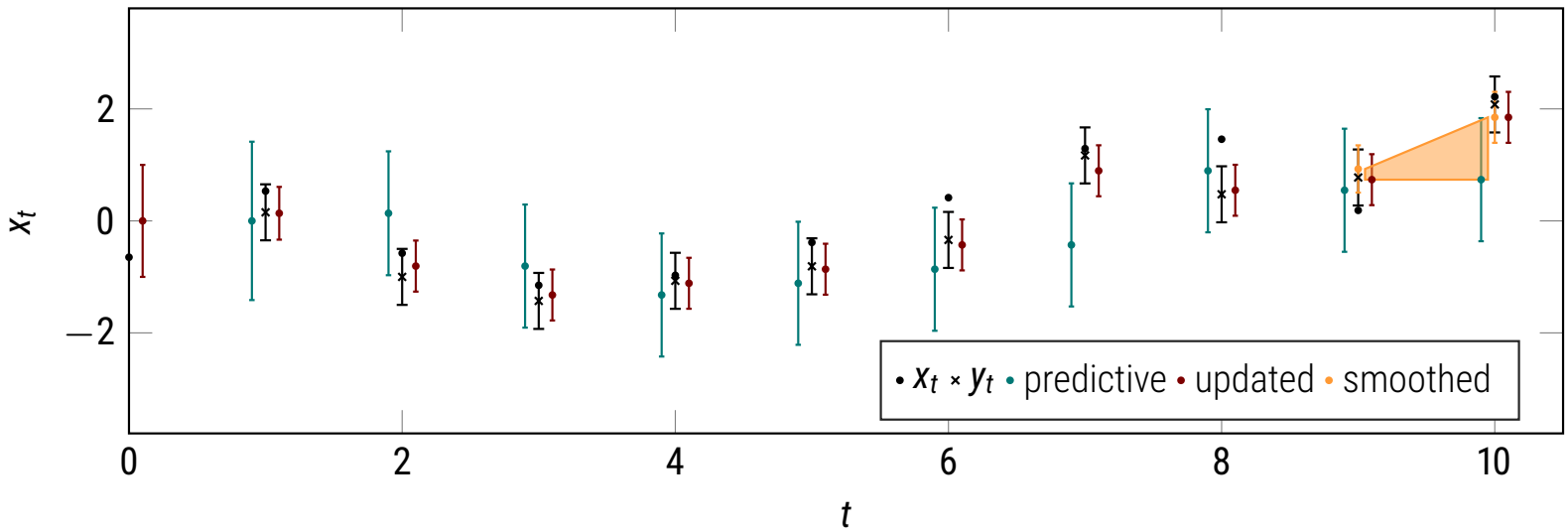
smooth: $$p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \frac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{1:t})}\, dx_{t+1} = \mathcal{N}(x_t, m_t^s, P_t^s)$$

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

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

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

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

$$G_t = P_tA^\mathsf{T} (P_t^-)^{-1},$$

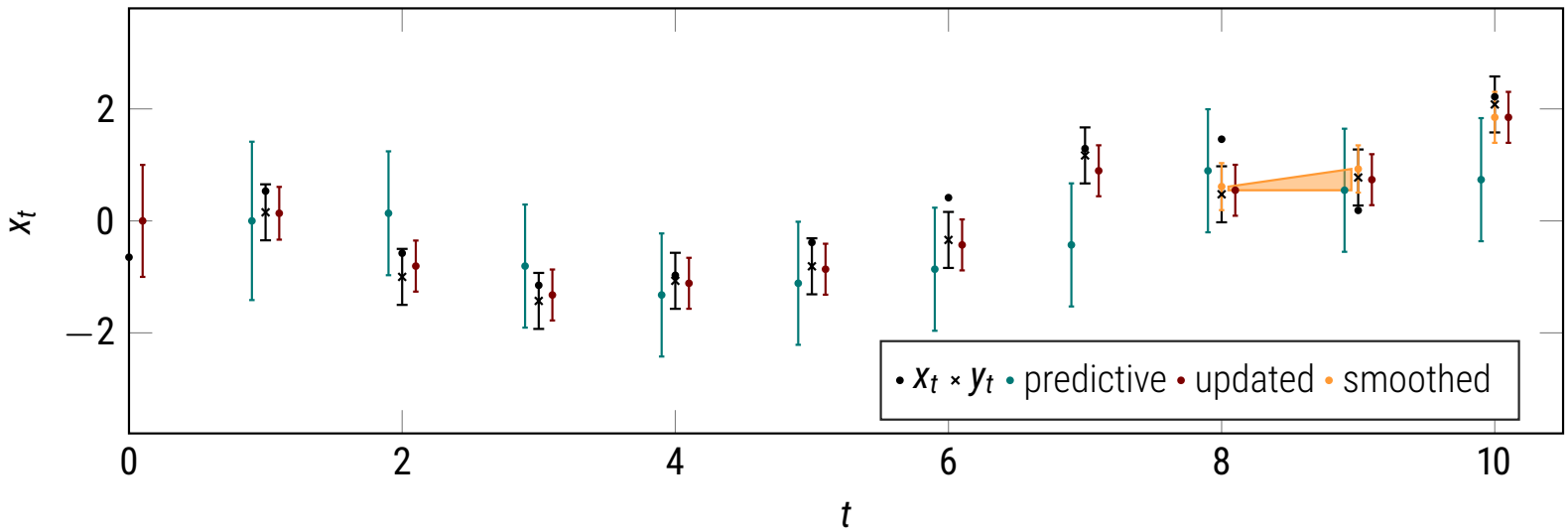
smooth:
$$p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \frac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{1:t})}\, dx_{t+1} = \mathcal{N}(x_t, m_t^s, P_t^s)$$

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

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

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

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

$$G_t = P_tA^\mathsf{T} (P_t^-)^{-1},$$

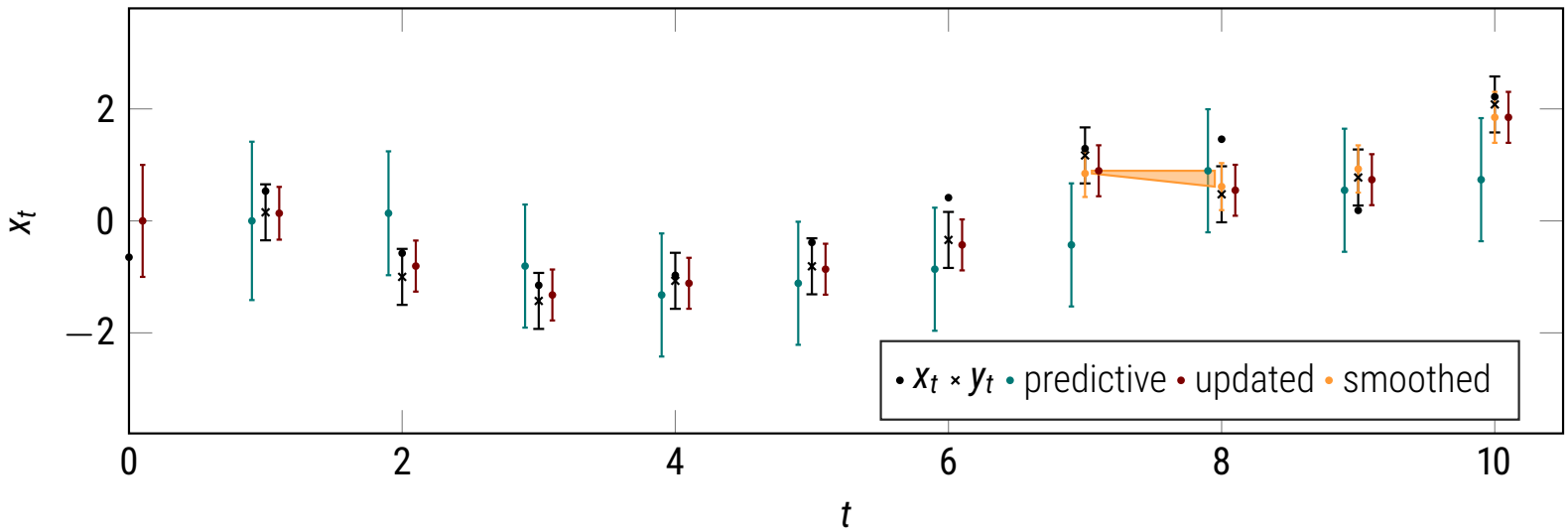
smooth:
$$p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \frac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{1:t})}\, dx_{t+1} = \mathcal{N}(x_t, m_t^s, P_t^s)$$

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

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

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

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

$$G_t = P_tA^\mathsf{T} (P_t^-)^{-1},$$

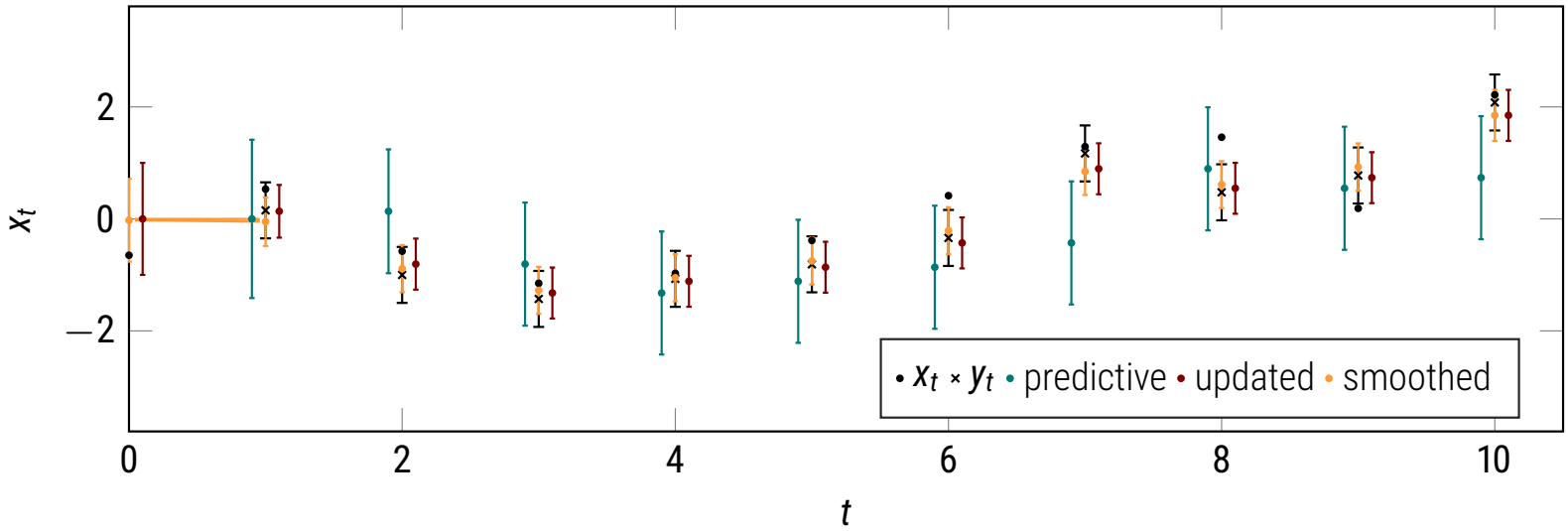
smooth: $p(x_t \mid Y) = p(x_t \mid Y_{0:t}) \int p(x_{t+1} \mid x_t) \dfrac{p(x_{t+1} \mid Y)}{p(x_{t+1} \mid Y_{1:t})}\, dx_{t+1} = \mathcal{N}(x_t, m_t^s, P_t^s)$

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

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:

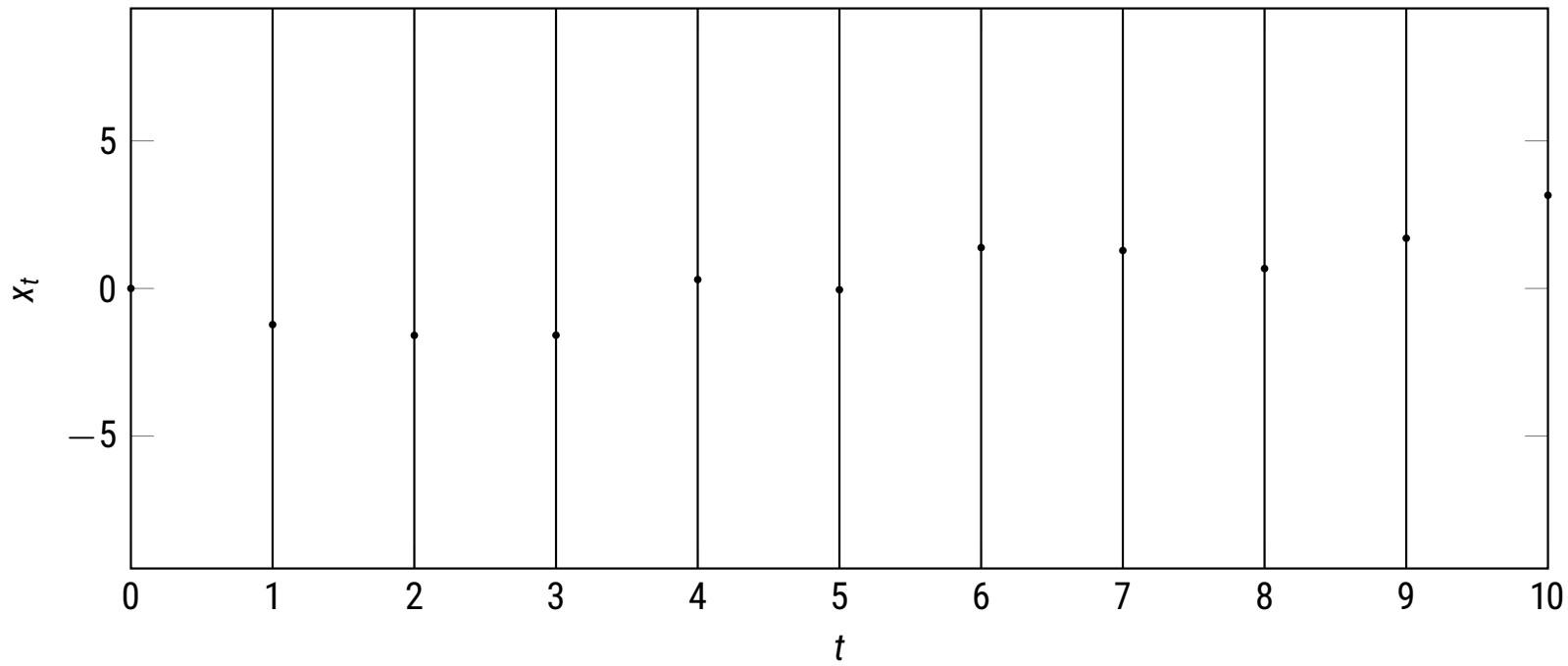
$G_t = P_t A^\mathsf{T} (P_{t+1}^-)^{-1}$ RTS gain

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

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

Continuous Time

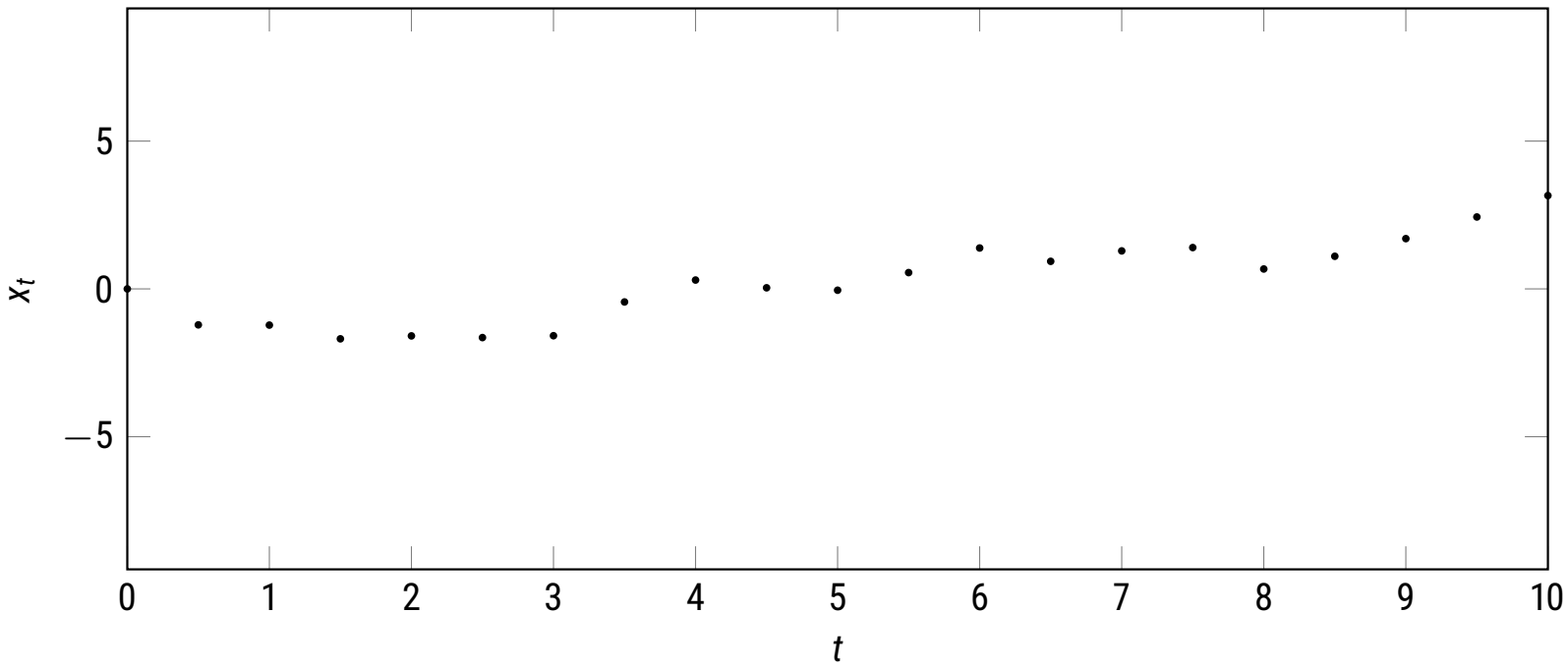
Differential equations defining non-differential curves



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

Continuous Time

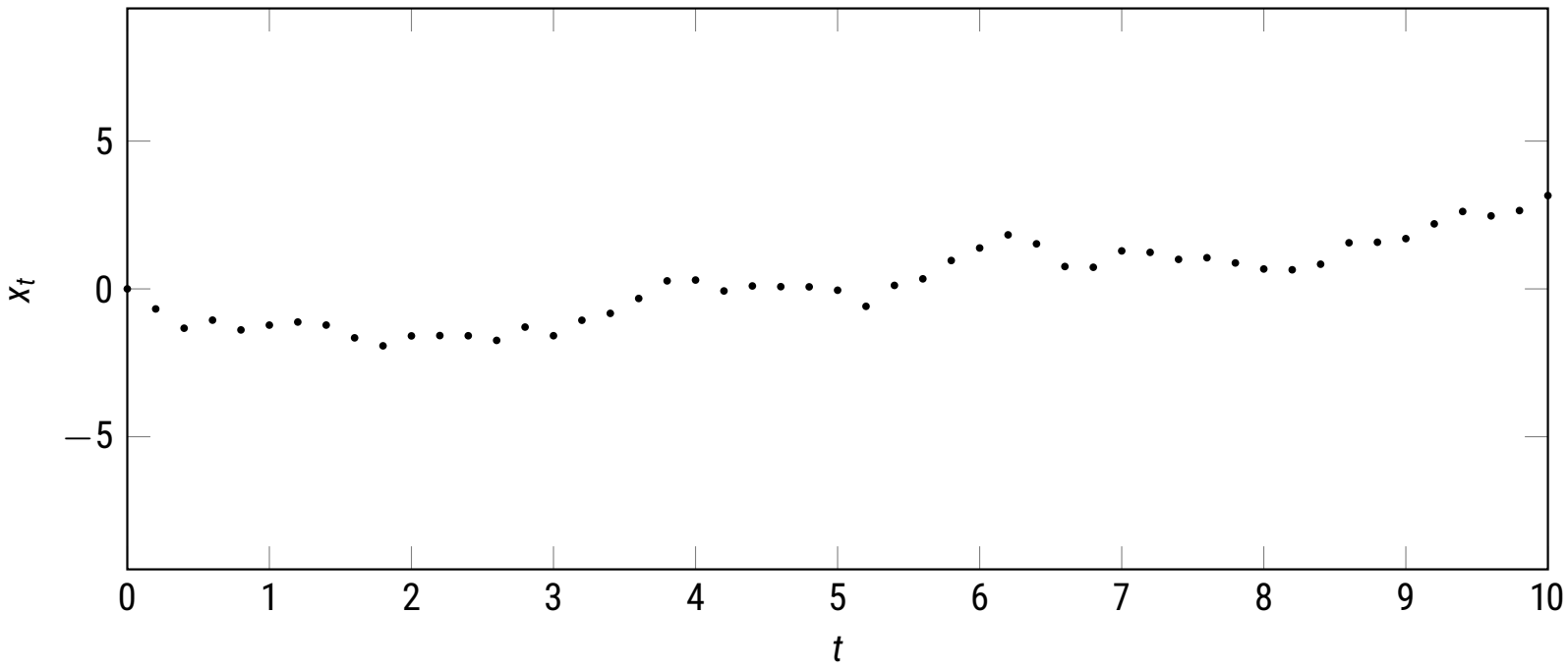
Differential equations defining non-differential curves



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

Continuous Time

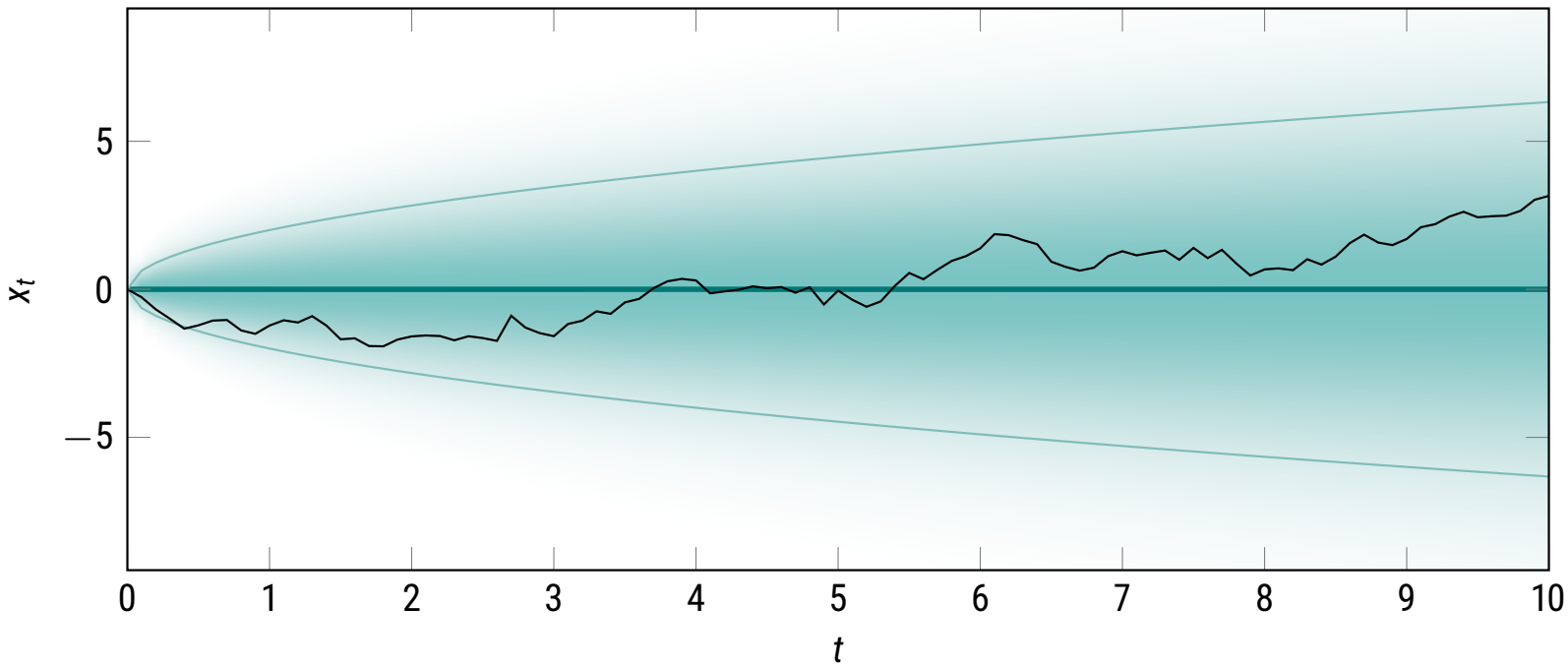
Differential equations defining non-differential curves



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

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 the following mean and covariance function

$$\mathrm{E}[x(t)] = e^{F(t-t_0)}x_0 \qquad k(t_a, t_b) = \int_{t_0}^{\min t_a, t_b} e^{F(t_a-\tau)}LL^\intercal e^{F^\intercal(t_b-\tau)}\,d\tau$$

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

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

$$e^X := \sum_{i=0}^{\infty}\frac{X^i}{i!}. \quad e^0 = I, \quad (e^X)^{-1} = e^{-X}, \quad \det e^X = e^{\mathrm{tr}\,X}, \quad X = VDV^{-1} \Rightarrow Ve^DV^{-1}, \quad e^{\mathrm{diag}_i\,d_i} = \mathrm{diag}_i\,e^{d_i}.$$

- **Markov** structure allows **linear time** inference on ordered spaces
- In **Gauss-Markov** models this involves only comparably simple linear algebra, the **Kalman filter**, and RTS smoother
- the conceptually tricky continuous time limit is known as a linear **stochastic differential equation**

For more on Filters and SDEs see

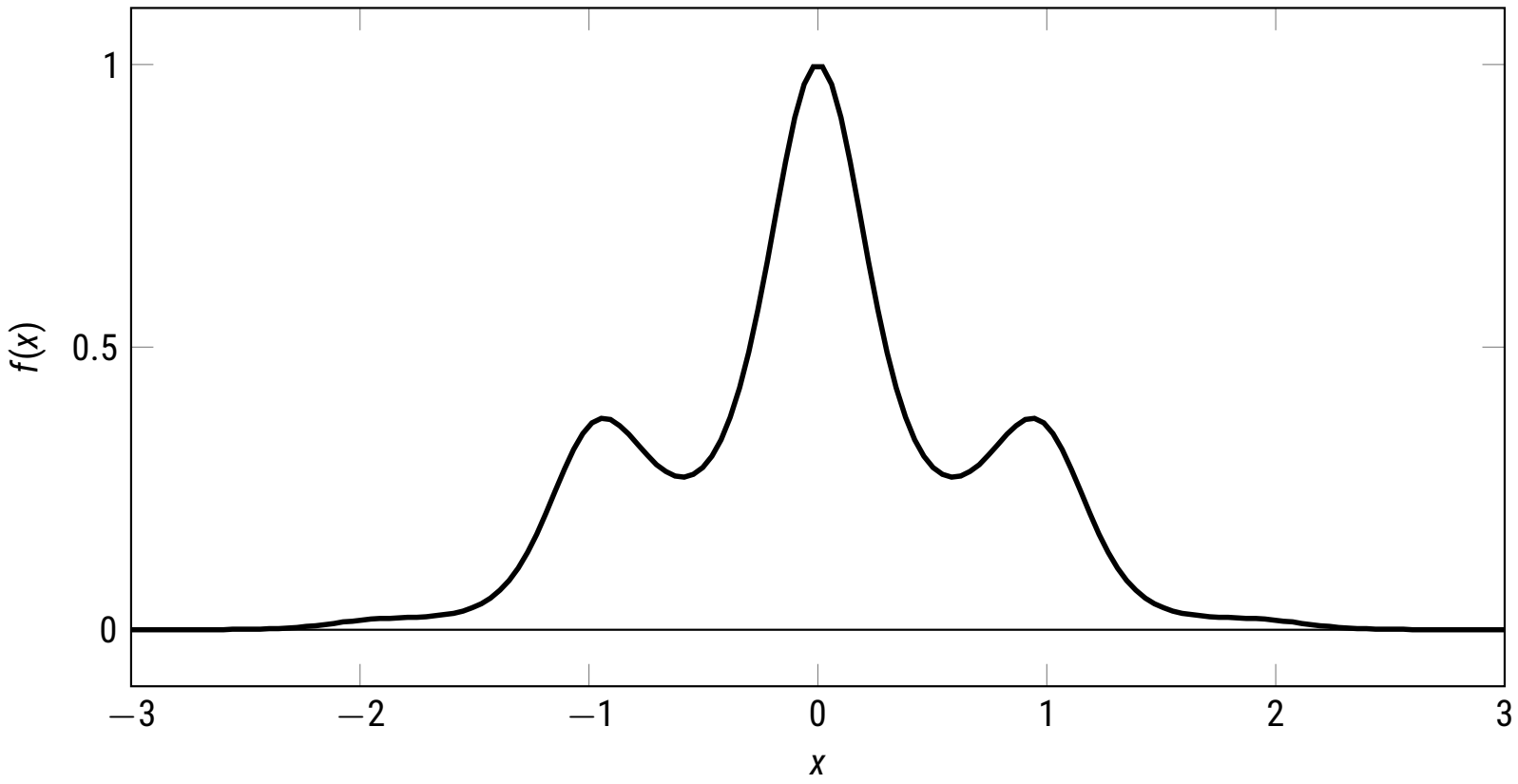
- Särkkä. **Bayesian Filtering and Smoothing**. Cambridge, 2013
- Rogers & Williams. **Diffusions, Markov Processes and Martingales** (Vols. I & II). Cambridge, 2000

coming up:

- our first **probabilistic numerical algorithm**

A Univariate Integration Problem

let's start slow



$$f(x) = \exp(-\sin(3x)^2 - x^2) \qquad G = \int_{-3}^{3} f(x)\,dx = ?$$

A Quadrature SDE

Phrasing Integration as Bayesian Inference

$$dz(x) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} z(x)\,dx + \begin{bmatrix} 0 \\ \theta \end{bmatrix} d\omega$$

- given initial value $z_2(x_0) = z_0, z_1(x_0) = 0$, this encodes $z_1(x) = \int_{x_0}^{x} z_2(\tilde{x})\,d\tilde{x}$
- thus can interpret

$$z(x) = \begin{bmatrix} \int_{x_0}^{x} \tilde{f}(\tilde{x})\,d\tilde{x} \\ \tilde{f}(x) \end{bmatrix}$$

- from Definition of SDE, with $\underline{x} := \min(x_a, x_b)$:

$$\begin{aligned} \operatorname{cov}(z(x_a), z(x_b)) &= \int_{x_0}^{\min(x_a,x_b)} \begin{bmatrix} 1 & (x_a - \tau) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ \theta \end{bmatrix} \begin{bmatrix} 0 & \theta \end{bmatrix} \begin{bmatrix} 1 & 0 \\ (x_b - \tau) & 1 \end{bmatrix} d\tau \\ &= \theta^2 \begin{bmatrix} x_a x_b(\underline{x} - x_0) - 1/2(x_a + x_b)(\underline{x}^2 - x_0^2) + 1/3(\underline{x}^3 - x_a^3) & \\ x_b(\underline{x} - x_0) - 1/2(\underline{x}^2 - x_0^2) & \underline{x} - x_0 \end{bmatrix} \end{aligned}$$

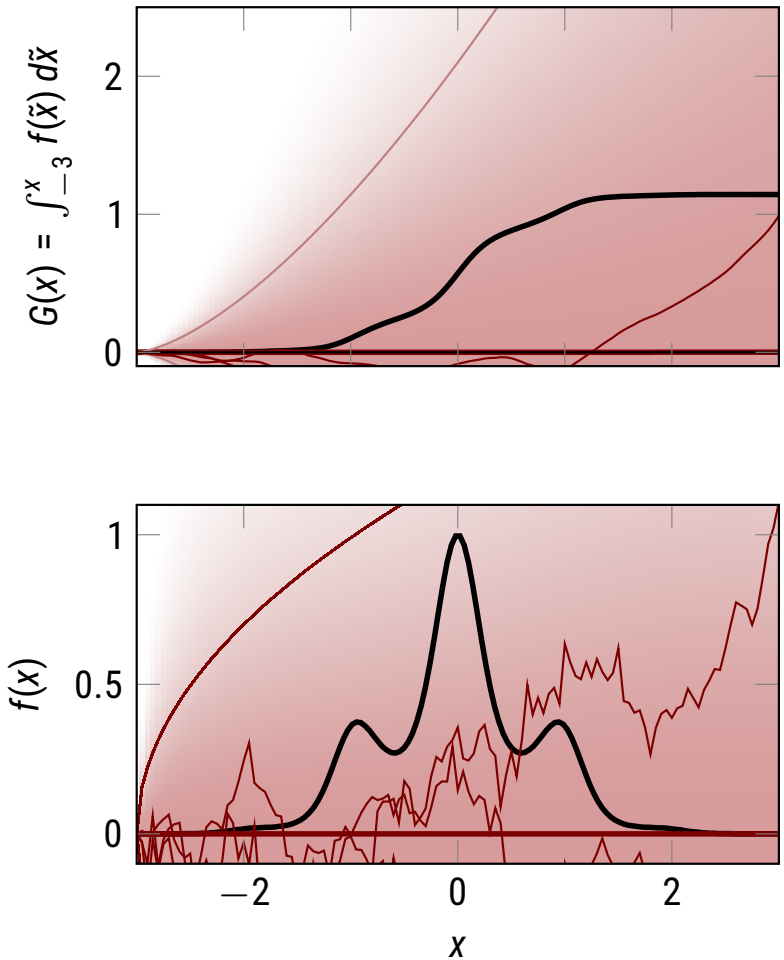
$\Rightarrow$ **Wiener process** prior $p(f) = \mathcal{GP}(0, \min(x_a, x_b))$

- filter parameters: $A(h) = \begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix}$, $Q(h) = \begin{bmatrix} h^3/3 & h^2/2 \\ h^2/2 & h \end{bmatrix}$, $H = \begin{bmatrix} 0 & 1 \end{bmatrix}$, $R = 0$.

A Quadrature Filter

Bayesian inference need not be expensive

[cf. Diaconis, 1988]

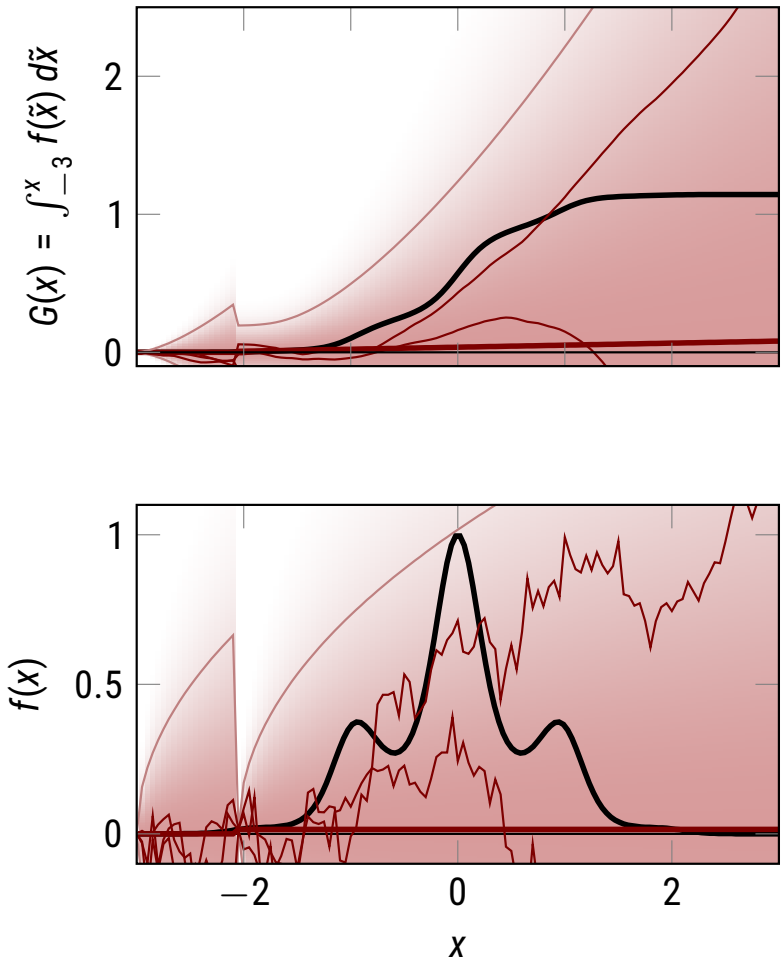



```
1  procedure INTEGRATE(@f, a, b, N)
2      x ⇐ a
3      h ⇐ (b − a)/N
4      m ⇐ [0; f(a)]
5      P ⇐ [0, 0; 0, 0]
6      for i = 1, ... , N − 1 do
7          x ⇐ x + h  // move
8          m⁻ = Am  // predict
9          P⁻ = APAᵀ + Q
10         y = f(x) − Hm⁻  // observe
11         S = HP⁻Hᵀ
12         K = P⁻Hᵀ/S  // gain
13         m ⇐ m⁻ + Ky  // update
14         P ⇐ (I − KH)P⁻  // update
15     end for
16     return m₁, P₁₁        // probabilistic estimate
17 end procedure
```

A Quadrature Filter

Bayesian inference need not be expensive

[cf. Diaconis, 1988]

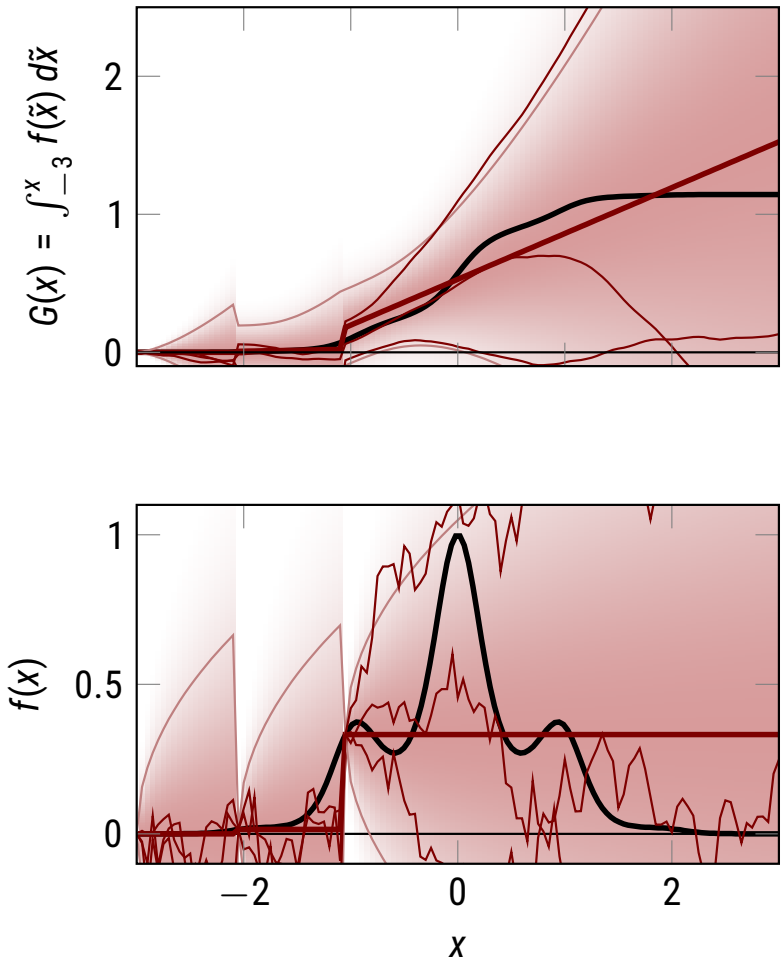



```
1  procedure INTEGRATE(@f, a, b, N)
2      x ⇐ a
3      h ⇐ (b − a)/N
4      m ⇐ [0; f(a)]
5      P ⇐ [0, 0; 0, 0]
6      for i = 1, ... , N − 1 do
7          x ⇐ x + h  // move
8          m⁻ = Am  // predict
9          P⁻ = APAᵀ + Q
10         y = f(x) − Hm⁻  // observe
11         S = HP⁻Hᵀ
12         K = P⁻Hᵀ/S  // gain
13         m ⇐ m⁻ + Ky  // update
14         P ⇐ (I − KH)P⁻  // update
15     end for
16     return m₁, P₁₁        // probabilistic estimate
17 end procedure
```

A Quadrature Filter

Bayesian inference need not be expensive

[cf. Diaconis, 1988]

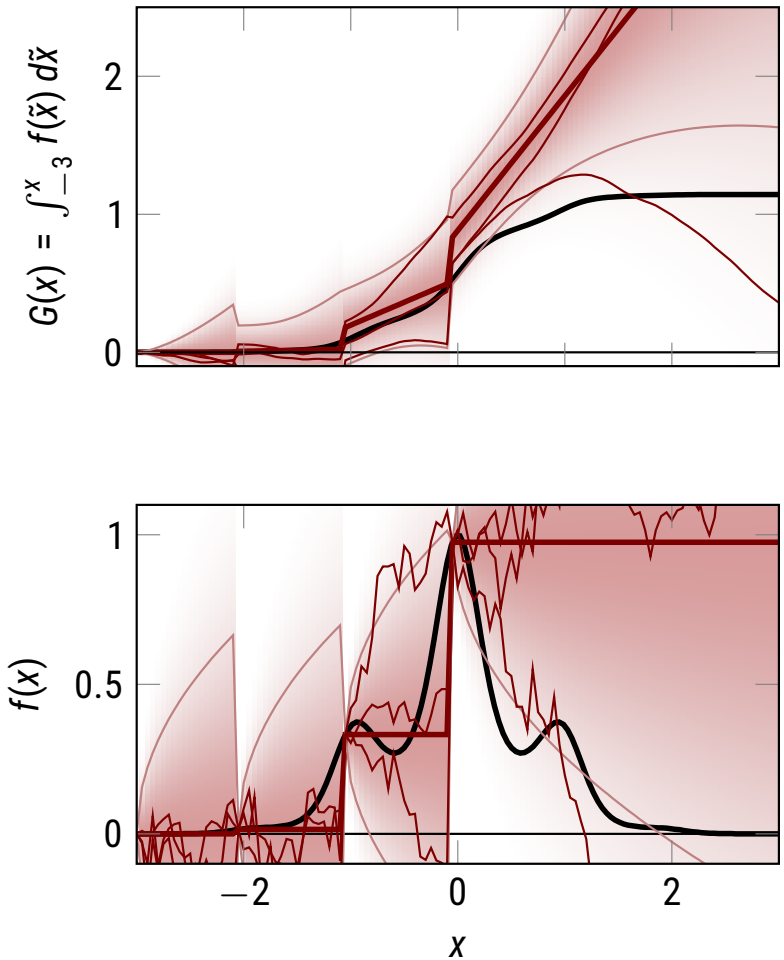



```
1  procedure INTEGRATE(@f, a, b, N)
2      x ⇐ a
3      h ⇐ (b − a)/N
4      m ⇐ [0; f(a)]
5      P ⇐ [0, 0; 0, 0]
6      for i = 1, ..., N − 1 do
7          x ⇐ x + h  // move
8          m⁻ = Am  // predict
9          P⁻ = APAᵀ + Q
10         y = f(x) − Hm⁻  // observe
11         S = HP⁻Hᵀ
12         K = P⁻Hᵀ/S  // gain
13         m ⇐ m⁻ + Ky  // update
14         P ⇐ (I − KH)P⁻  // update
15     end for
16     return m₁, P₁₁        // probabilistic estimate
17 end procedure
```

A Quadrature Filter

Bayesian inference need not be expensive

[cf. Diaconis, 1988]

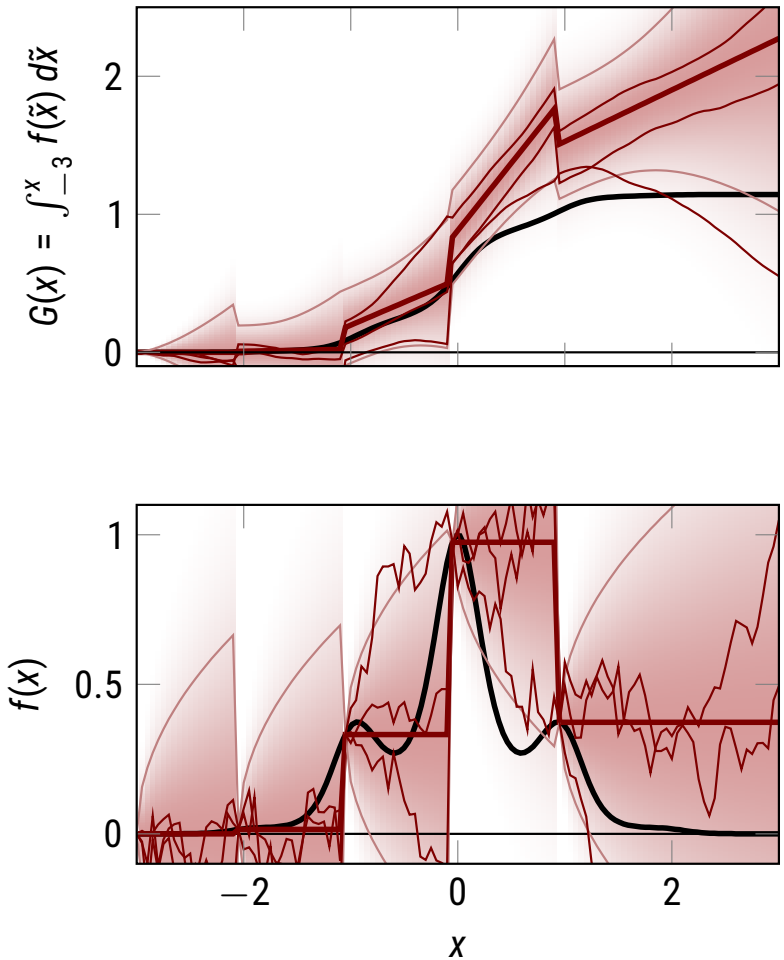



```
1  procedure INTEGRATE(@f, a, b, N)
2      x ⇐ a
3      h ⇐ (b − a)/N
4      m ⇐ [0; f(a)]
5      P ⇐ [0, 0; 0, 0]
6      for i = 1, ... , N − 1 do
7          x ⇐ x + h  // move
8          m⁻ = Am  // predict
9          P⁻ = APAᵀ + Q
10         y = f(x) − Hm⁻  // observe
11         S = HP⁻Hᵀ
12         K = P⁻Hᵀ/S  // gain
13         m ⇐ m⁻ + Ky  // update
14         P ⇐ (I − KH)P⁻  // update
15     end for
16     return m₁, P₁₁        // probabilistic estimate
17 end procedure
```

A Quadrature Filter

Bayesian inference need not be expensive

[cf. Diaconis, 1988]

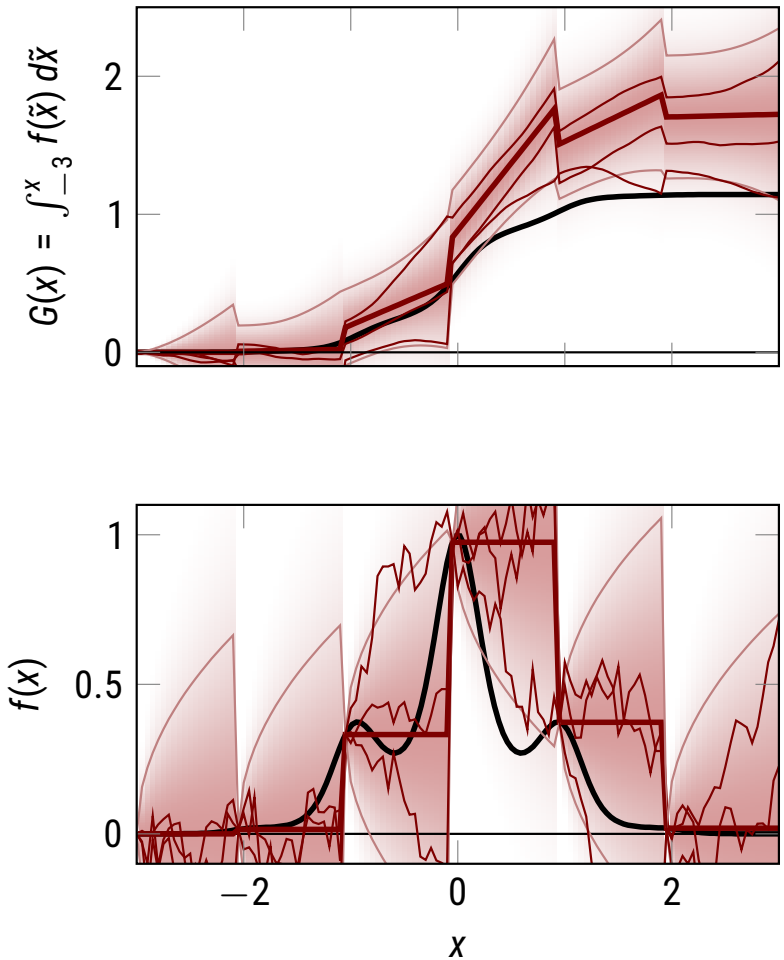



```
1  procedure INTEGRATE(@f, a, b, N)
2      x ⇐ a
3      h ⇐ (b − a)/N
4      m ⇐ [0; f(a)]
5      P ⇐ [0, 0; 0, 0]
6      for i = 1, ..., N − 1 do
7          x ⇐ x + h  // move
8          m⁻ = Am  // predict
9          P⁻ = APAᵀ + Q
10         y = f(x) − Hm⁻  // observe
11         S = HP⁻Hᵀ
12         K = P⁻Hᵀ/S  // gain
13         m ⇐ m⁻ + Ky  // update
14         P ⇐ (I − KH)P⁻  // update
15     end for
16     return m₁, P₁₁        // probabilistic estimate
17 end procedure
```

A Quadrature Filter

Bayesian inference need not be expensive

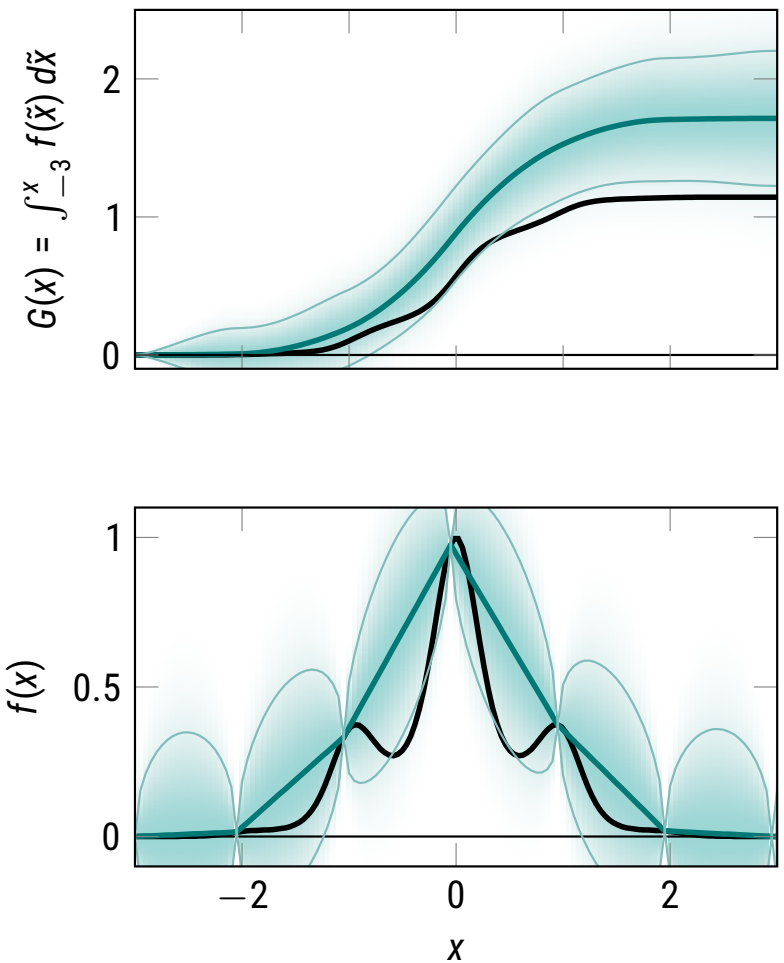
[cf. Diaconis, 1988]




```
1  procedure INTEGRATE(@f, a, b, N)
2      x ⇐ a
3      h ⇐ (b − a)/N
4      m ⇐ [0; f(a)]
5      P ⇐ [0, 0; 0, 0]
6      for i = 1, ..., N − 1 do
7          x ⇐ x + h  // move
8          m⁻ = Am  // predict
9          P⁻ = APAᵀ + Q
10         y = f(x) − Hm⁻  // observe
11         S = HP⁻Hᵀ
12         K = P⁻Hᵀ/S  // gain
13         m ⇐ m⁻ + Ky  // update
14         P ⇐ (I − KH)P⁻  // update
15     end for
16     return m₁, P₁₁        // probabilistic estimate
17 end procedure
```

A Quadrature Filter

Bayesian inference need not be expensive [cf. Diaconis, 1988]



$$m_{i+1} = Am_i + \frac{(APA^\intercal + Q)H^\intercal}{H(APA^\intercal + Q)H^\intercal}(f(x_i) - HAm_i)$$

$$= \begin{bmatrix} [m_i]_1 + {}^{h}/_{2}(f(x_{i-1} + f(x_i)) \\ f(x_i) \end{bmatrix}$$

$$\Rightarrow \quad \mathbb{E}[G] = [m_N]_1 = \sum_{i=1}^{N-1} {}^{h_i}/_{2}(f(x_{i+1}) + f(x_i))$$

$$P_{i+1} = (I - KH)(APA^\intercal + Q)$$

$$= \begin{bmatrix} [P_i]_{11} + {}^{\theta^2}/_{12} \cdot h^3 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\Rightarrow \quad \mathrm{var}[G] = [P_N]_{11} = \frac{\theta^2}{12} \sum_{i=1}^{N} h_i^3$$

$$\Rightarrow \quad h_i = \underset{\tilde{\mathbf{h}}, \sum_i \tilde{h}_i = (b-a)}{\arg\min} \mathrm{var}[G] = (b-a)/N$$

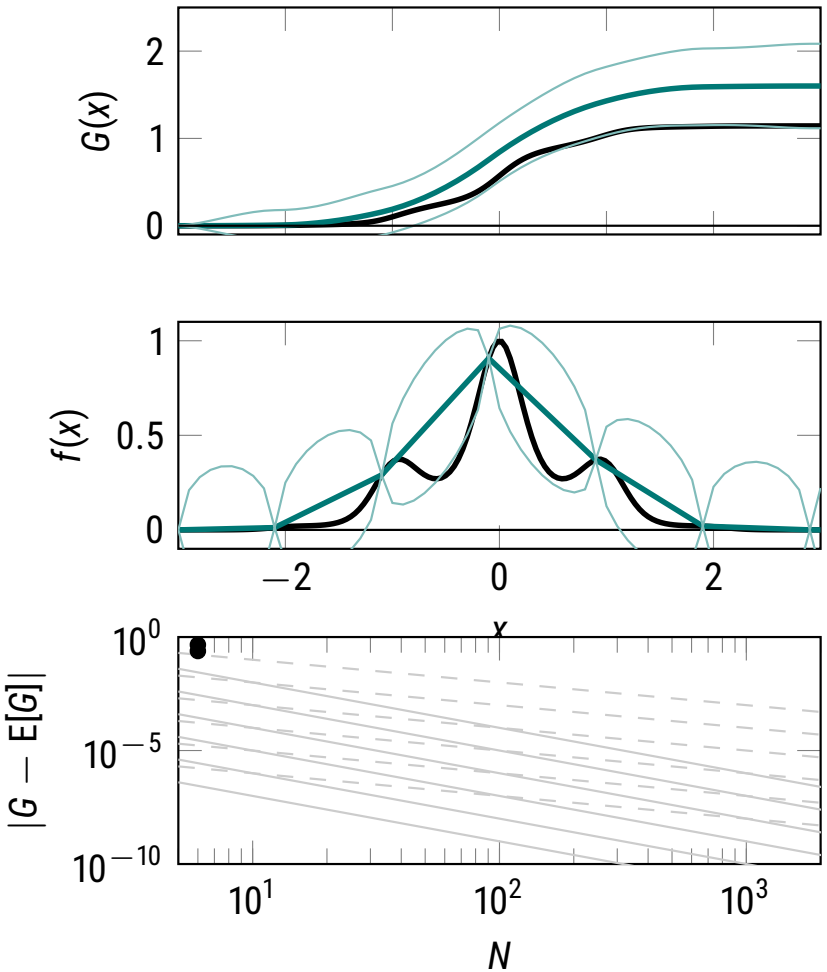
$$\Rightarrow \quad \mathrm{var}[G] = \frac{\theta^2}{12} \frac{x_N - x_0}{N^2}$$

Trapezoidal rule is **MAP** estimate arising under a **Wiener process** prior on f.
Regular grid arises as minimum entropy (most informative) choice.

Convergence

classic asymptotic analysis

[cf. Davis & Rabinowitz, 1984, pp. 52–54]



Let $w(\delta) = \max_{|x_1 - x_2| \leq \delta} |f(x_2) - f(x_1)|$, for $a < x_1, x_2 < b$, the **modulus of continuity**.

Theorem [Davis & Rabinowitz, p. 52]

Let $f(x)$ be continuous in $[a, b]$. Then

$$|G - \mathrm{E}[G]| \leq (b - a) w\left(\frac{b - a}{N}\right).$$

Theorem [Davis & Rabinowitz, p. 54]

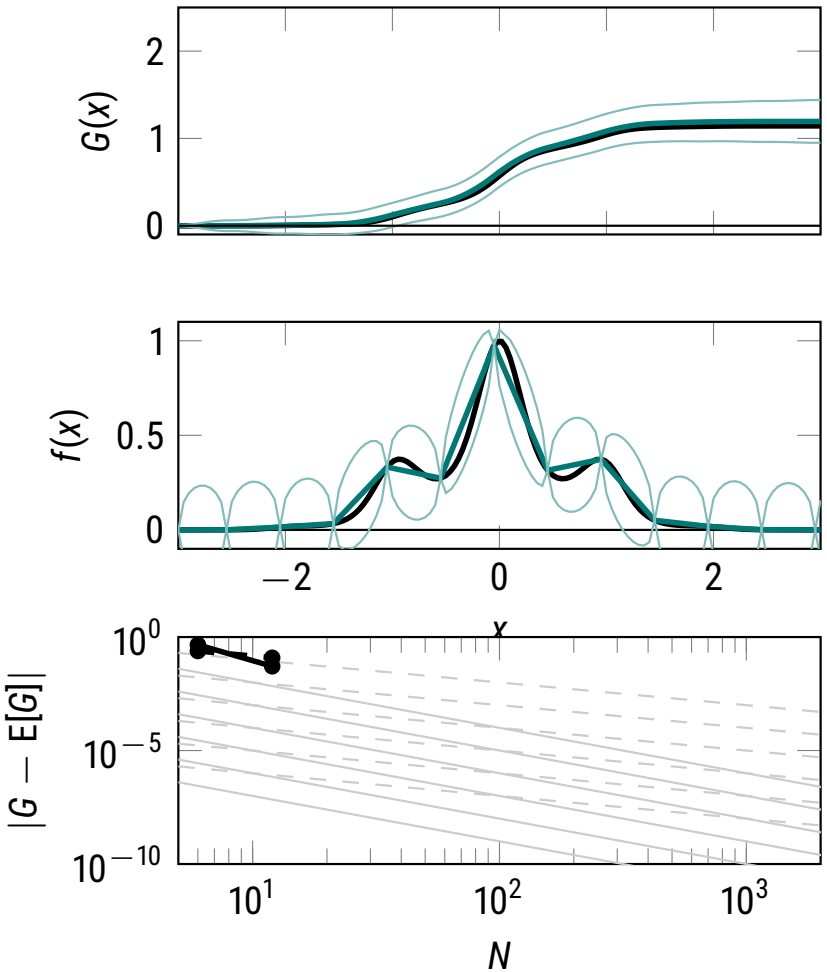
Let $f(x) \in \mathcal{C}^2[a, b]$. Then $\exists \xi \in (a, b)$, so that

$$G - \mathrm{E}[G] = -\frac{(b - a)^3}{12N^2} f''(\xi).$$

Convergence

classic asymptotic analysis

[cf. Davis & Rabinowitz, 1984, pp. 52–54]



Let $w(\delta) = \max_{|x_1 - x_2| \leq \delta} |f(x_2) - f(x_1)|$, for $a < x_1, x_2 < b$, the **modulus of continuity**.

Theorem [Davis & Rabinowitz, p. 52]

Let $f(x)$ be continuous in $[a, b]$. Then

$$|G - \mathrm{E}[G]| \leq (b - a) w\left(\frac{b - a}{N}\right).$$

Theorem [Davis & Rabinowitz, p. 54]

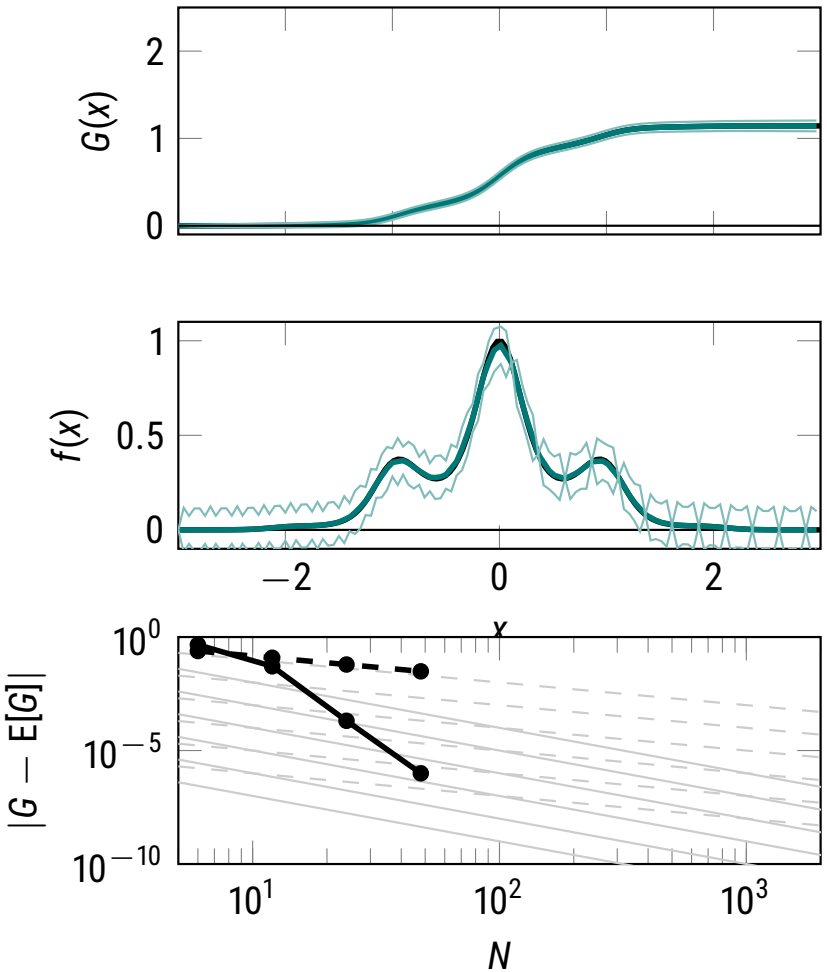
Let $f(x) \in \mathcal{C}^2[a, b]$. Then $\exists \xi \in (a, b)$, so that

$$G - \mathrm{E}[G] = -\frac{(b - a)^3}{12N^2} f''(\xi).$$

Convergence

classic asymptotic analysis

[cf. Davis & Rabinowitz, 1984, pp. 52–54]



Let $w(\delta) = \max_{|x_1 - x_2| \leq \delta} |f(x_2) - f(x_1)|$, for $a < x_1, x_2 < b$, the **modulus of continuity**.

Theorem [Davis & Rabinowitz, p. 52]

Let $f(x)$ be continuous in $[a, b]$. Then

$$|G - \mathrm{E}[G]| \leq (b - a) w\left(\frac{b - a}{N}\right).$$

Theorem [Davis & Rabinowitz, p. 54]

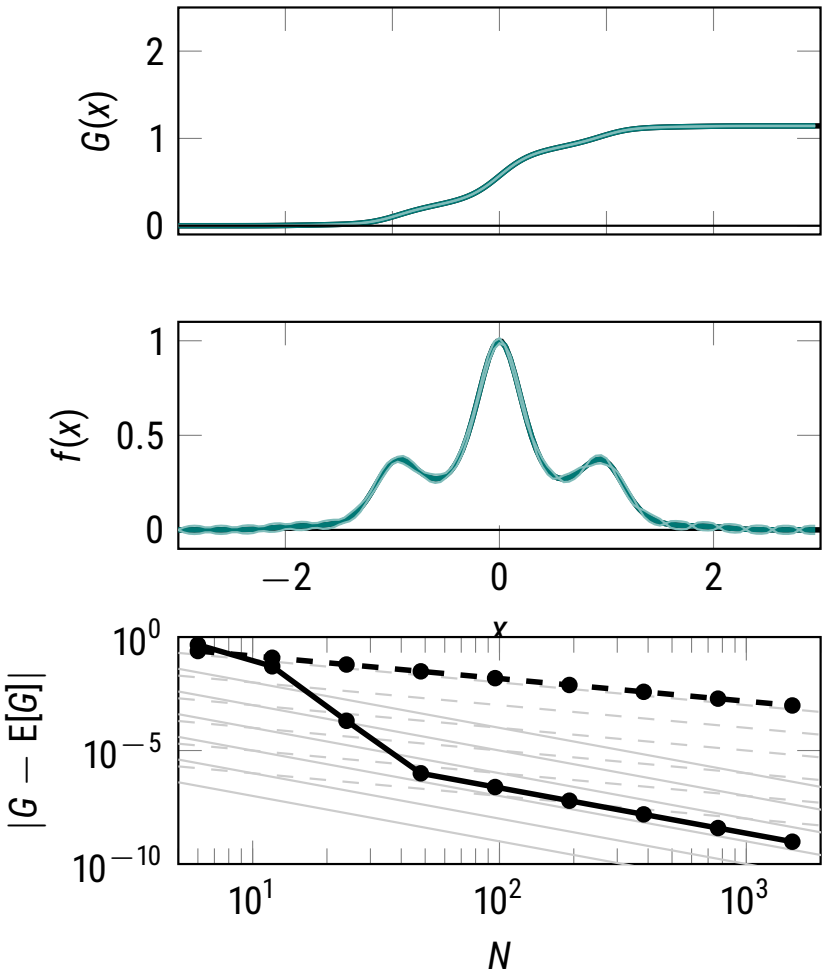
Let $f(x) \in \mathcal{C}^2[a, b]$. Then $\exists \xi \in (a, b)$, so that

$$G - \mathrm{E}[G] = -\frac{(b - a)^3}{12N^2} f''(\xi).$$

Convergence

classic asymptotic analysis

[cf. Davis & Rabinowitz, 1984, pp. 52–54]



Let $w(\delta) = \max_{|x_1 - x_2| \leq \delta} |f(x_2) - f(x_1)|$, for $a < x_1, x_2 < b$, the **modulus of continuity**.

Theorem [Davis & Rabinowitz, p. 52]

Let $f(x)$ be continuous in $[a, b]$. Then

$$|G - \mathrm{E}[G]| \leq (b - a) w\left(\frac{b - a}{N}\right).$$

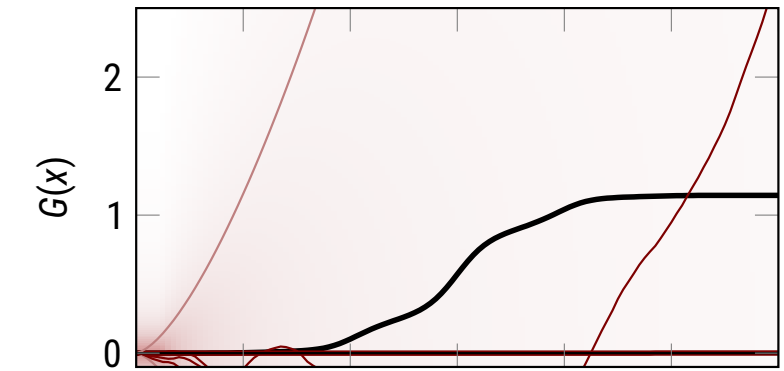
Theorem [Davis & Rabinowitz, p. 54]

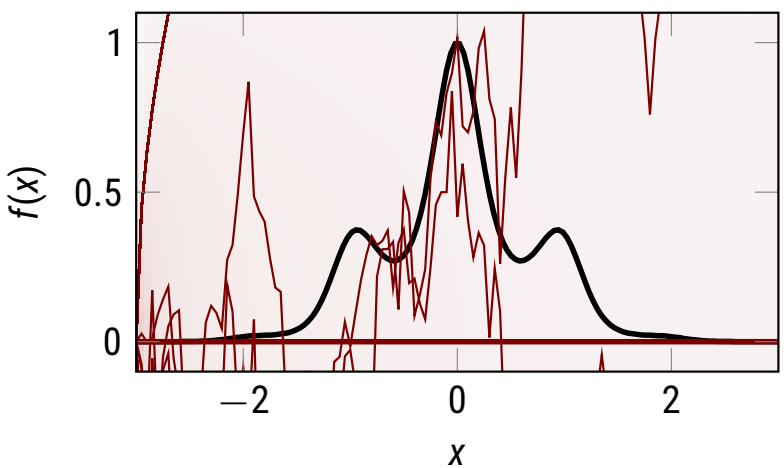
Let $f(x) \in \mathcal{C}^2[a, b]$. Then $\exists \xi \in (a, b)$, so that

$$G - \mathrm{E}[G] = -\frac{(b - a)^3}{12N^2} f''(\xi).$$

Error Estimation

a rare case of exact hyperparameter inference





- recall $p(y \mid \theta) = \int p(y, f \mid \theta)\, df = \mathcal{N}(y; m_X, k_{XX})$
- consider Gamma **conjugate prior**

$$p(\theta^{-2}) = \mathcal{G}(\theta^{-2}; \alpha_0, \beta_0) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} (\theta^{-2})^{\alpha_0 - 1} e^{-\beta_0 \theta^{-2}}$$

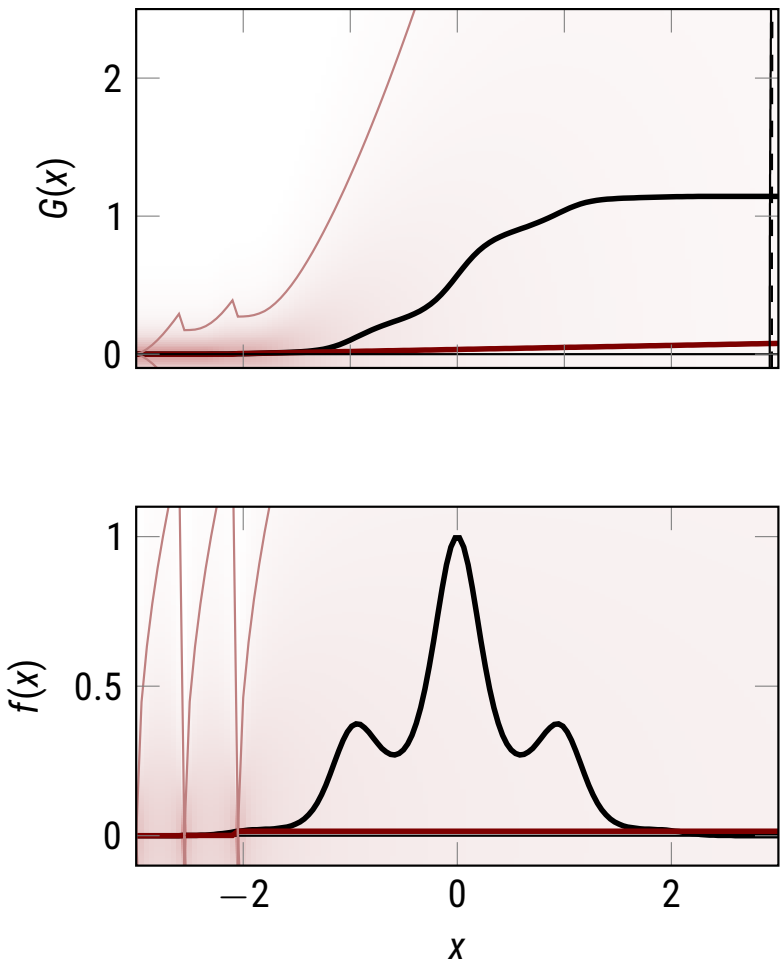
$$\begin{aligned} p(\theta \mid y) &= p(y \mid \theta) p(\theta) \\ &= \mathcal{G}(\theta^{-2}, \alpha_N := \alpha_0 + N/2, \\ &\qquad \beta_N := \beta_0 + 1/2 (y - \mu_X)^\intercal k_{XX}^{-1} (y - \mu_X)) \end{aligned}$$

- **Student-*t*** marginal. var: $\beta_N / (\alpha_N - 1)$

$$\begin{aligned} p(f_x \mid y) &= \int p(f \mid y, \theta^{-2}) p(\theta^{-2} \mid y)\, d\theta^{-2} \\ &= \mathrm{St}(f; \mu_y, \alpha_N, \beta_N \mathbb{V}_y) \\ &=: \frac{\Gamma(\alpha_N + 1/2)}{\Gamma(\alpha_N)(2\pi\beta_N \mathbb{V}_y)^{1/2}} \left(1 + \frac{(f_x - \mu_y)^2}{2\beta_N \mathbb{V}_y}\right)^{-\alpha_N - 1/2} \end{aligned}$$

Error Estimation

a rare case of exact hyperparameter inference



- recall $p(y \mid \theta) = \int p(y, f \mid \theta)\, df = \mathcal{N}(y; m_X, k_{XX})$
- consider Gamma **conjugate prior**

$$p(\theta^{-2}) = \mathcal{G}(\theta^{-2}; \alpha_0, \beta_0) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} (\theta^{-2})^{\alpha_0 - 1} e^{-\beta_0 \theta^{-2}}$$

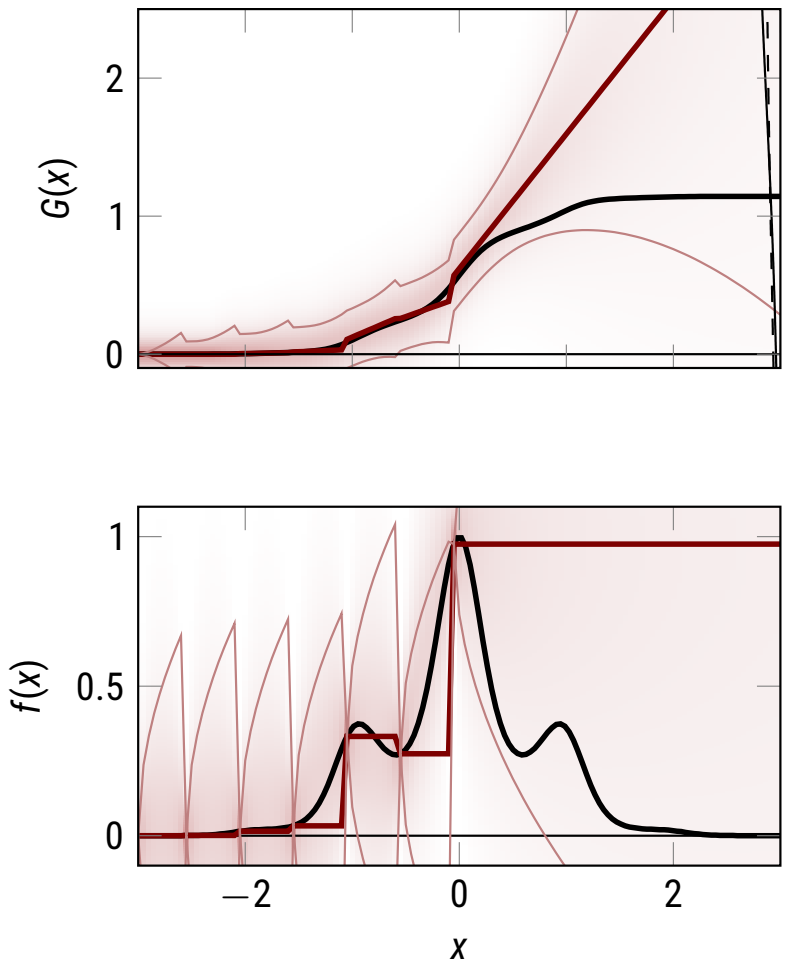
$$\begin{aligned} p(\theta \mid y) &= p(y \mid \theta) p(\theta) \\ &= \mathcal{G}(\theta^{-2}, \alpha_N := \alpha_0 + N/2, \\ &\qquad \beta_N := \beta_0 + 1/2 (y - \mu_X)^\intercal k_{XX}^{-1} (y - \mu_X)) \end{aligned}$$

- **Student-*t*** marginal. var: $\beta_N / (\alpha_N - 1)$

$$\begin{aligned} p(f_x \mid y) &= \int p(f \mid y, \theta^{-2}) p(\theta^{-2} \mid y)\, d\theta^{-2} \\ &= \mathrm{St}(f; \mu_y, \alpha_N, \beta_N \mathbb{V}_y) \\ &=: \frac{\Gamma(\alpha_N + 1/2)}{\Gamma(\alpha_N)(2\pi \beta_N \mathbb{V}_y)^{1/2}} \left(1 + \frac{(f_x - \mu_y)^2}{2 \beta_N \mathbb{V}_y}\right)^{-\alpha_N - 1/2} \end{aligned}$$

Error Estimation

a rare case of exact hyperparameter inference



- recall $p(y \mid \theta) = \int p(y, f \mid \theta)\, df = \mathcal{N}(y; m_X, k_{XX})$
- consider Gamma **conjugate prior**

$$p(\theta^{-2}) = \mathcal{G}(\theta^{-2}; \alpha_0, \beta_0) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} (\theta^{-2})^{\alpha_0 - 1} e^{-\beta_0 \theta^{-2}}$$

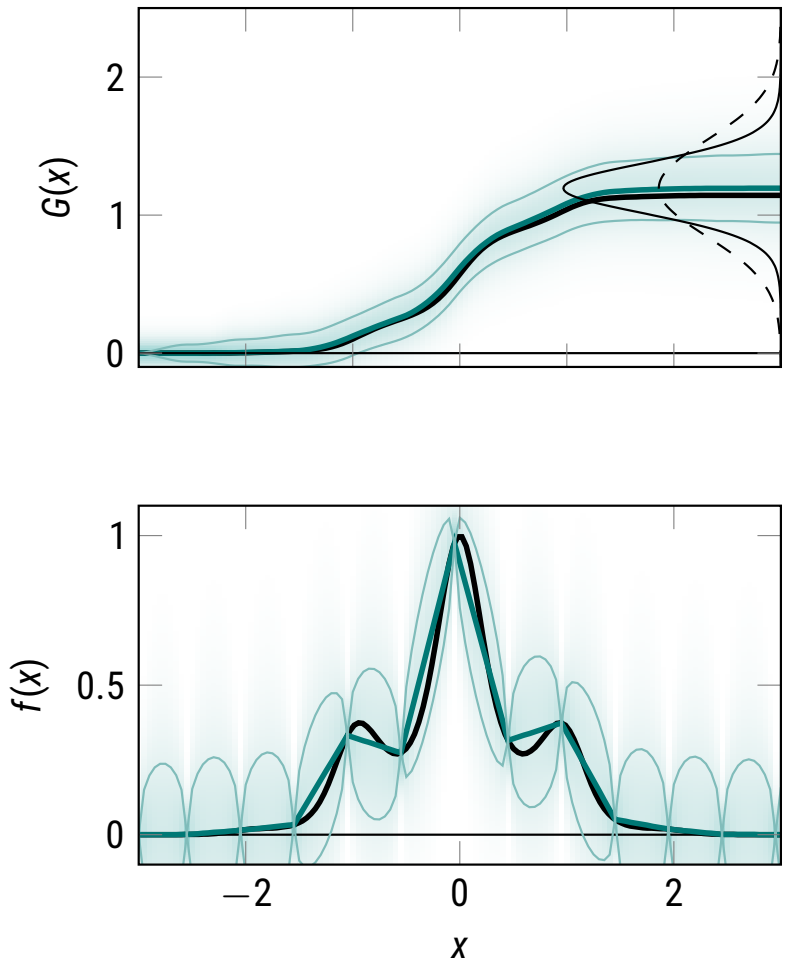
$$\begin{aligned} p(\theta \mid y) &= p(y \mid \theta) p(\theta) \\ &= \mathcal{G}(\theta^{-2}, \alpha_N := \alpha_0 + N/2, \\ &\qquad \beta_N := \beta_0 + 1/2 (y - \mu_X)^\intercal k_{XX}^{-1} (y - \mu_X)) \end{aligned}$$

- **Student-*t*** marginal. var: $\beta_N / (\alpha_N - 1)$

$$\begin{aligned} p(f_x \mid y) &= \int p(f \mid y, \theta^{-2}) p(\theta^{-2} \mid y)\, d\theta^{-2} \\ &= \mathrm{St}(f; \mu_y, \alpha_N, \beta_N \mathbb{V}_y) \\ &=: \frac{\Gamma(\alpha_N + 1/2)}{\Gamma(\alpha_N)(2\pi \beta_N \mathbb{V}_y)^{1/2}} \left(1 + \frac{(f_x - \mu_y)^2}{2\beta_N \mathbb{V}_y}\right)^{-\alpha_N - 1/2} \end{aligned}$$

Error Estimation

a rare case of exact hyperparameter inference



- for filter: use **prediction-error decomposition** ($\tilde{P} = P/\theta^2$)

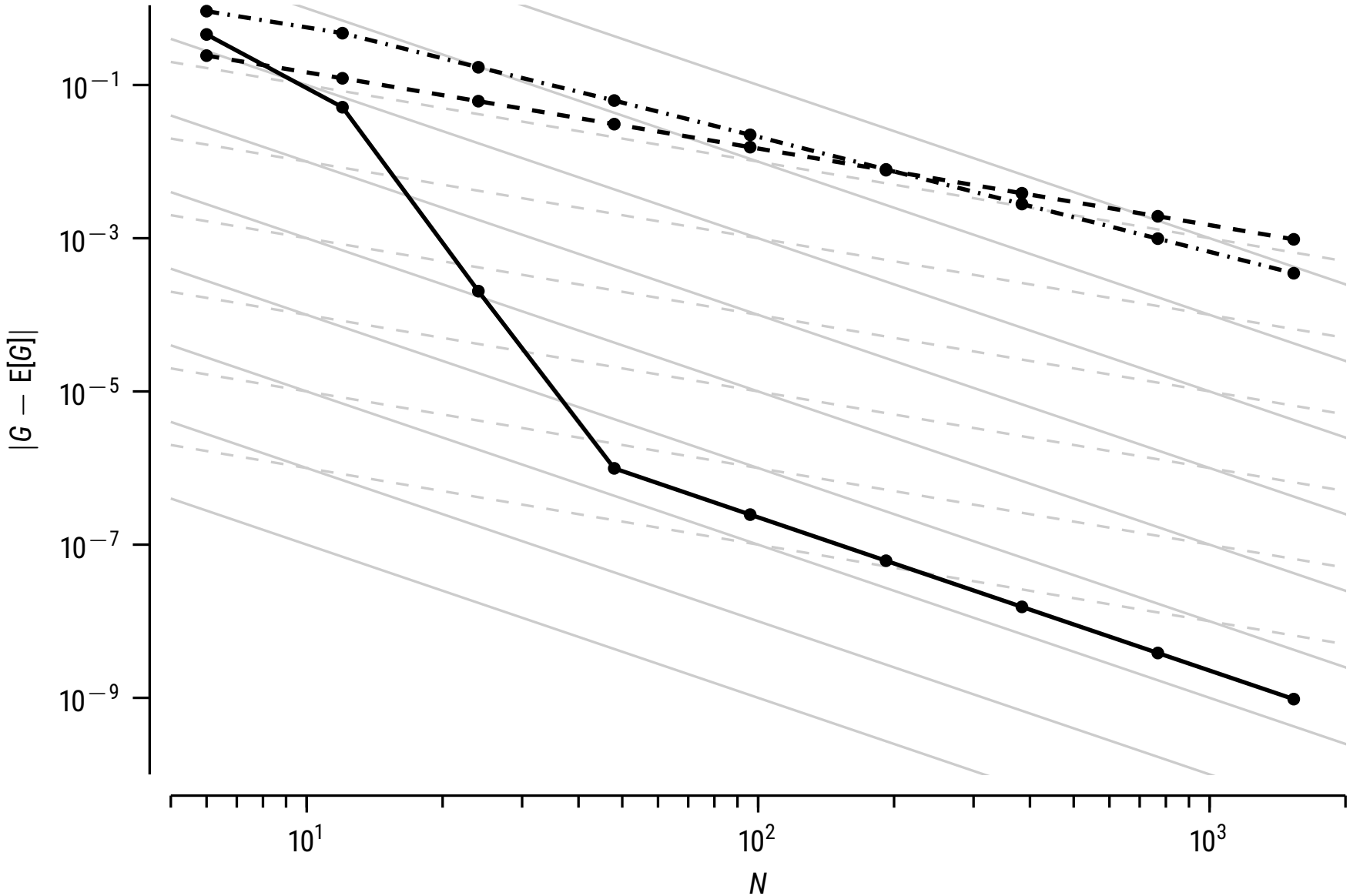
$$p(y \mid \theta) = \prod_{i=1}^{N} p(y_i \mid y_{1:i-1}, \theta)$$

$$= \prod_{i} \mathcal{N}(y_i; Hm_i^-, \theta^2 H\tilde{P}^- H)$$

```
1  procedure INTEGRATE(@f, a, b, N)
2      x ⇐ a, m ⇐ [0; f(a)], P ⇐ [0, 0; 0, 0]
3      β ⇐ 0
4      for i = 1, ..., N − 1 do
5          x ⇐ x + (b − a)/N
6          m⁻ = Am
7          P⁻ = APAᵀ + Q
8          y = f(x) − Hm⁻
9          S = HP⁻Hᵀ
10         β ⇐ β + y²/(2s)
11         K = P⁻Hᵀ/S
12         m ⇐ m⁻ + Ky
13         P ⇐ (I − KH)P⁻
14     end for
15     return m₁, β/(α−1) P₁₁        // calibrated estimate
16 end procedure
```

Error Estimation

it's tricky to fix a broken prior, though



Higher Order Filters

spline quadrature rules as MAP estimation

for more, see Schober et al. 2017; Kersting et al. 2017

$$dz(x) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \vdots & & & \ddots & 1 \\ 0 & \cdots & & 0 & 0 \end{bmatrix} z(x)\,dx + \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ \theta \end{bmatrix} d\omega \qquad z(x) = \begin{bmatrix} \int_{x_0}^{x} \tilde{f}(\tilde{x})\,d\tilde{x} \\ \tilde{f}(x) \\ \tilde{f}'(x) \\ \vdots \\ \tilde{f}^{q}(x) \end{bmatrix}$$

This gives q-th order **spline interpolants** as means, and

$$[A(h)]_{ij} = [\exp(Fh)]_{ij} = \mathbb{I}(j \geq i)\frac{h^{j-i}}{(j-i)!}$$

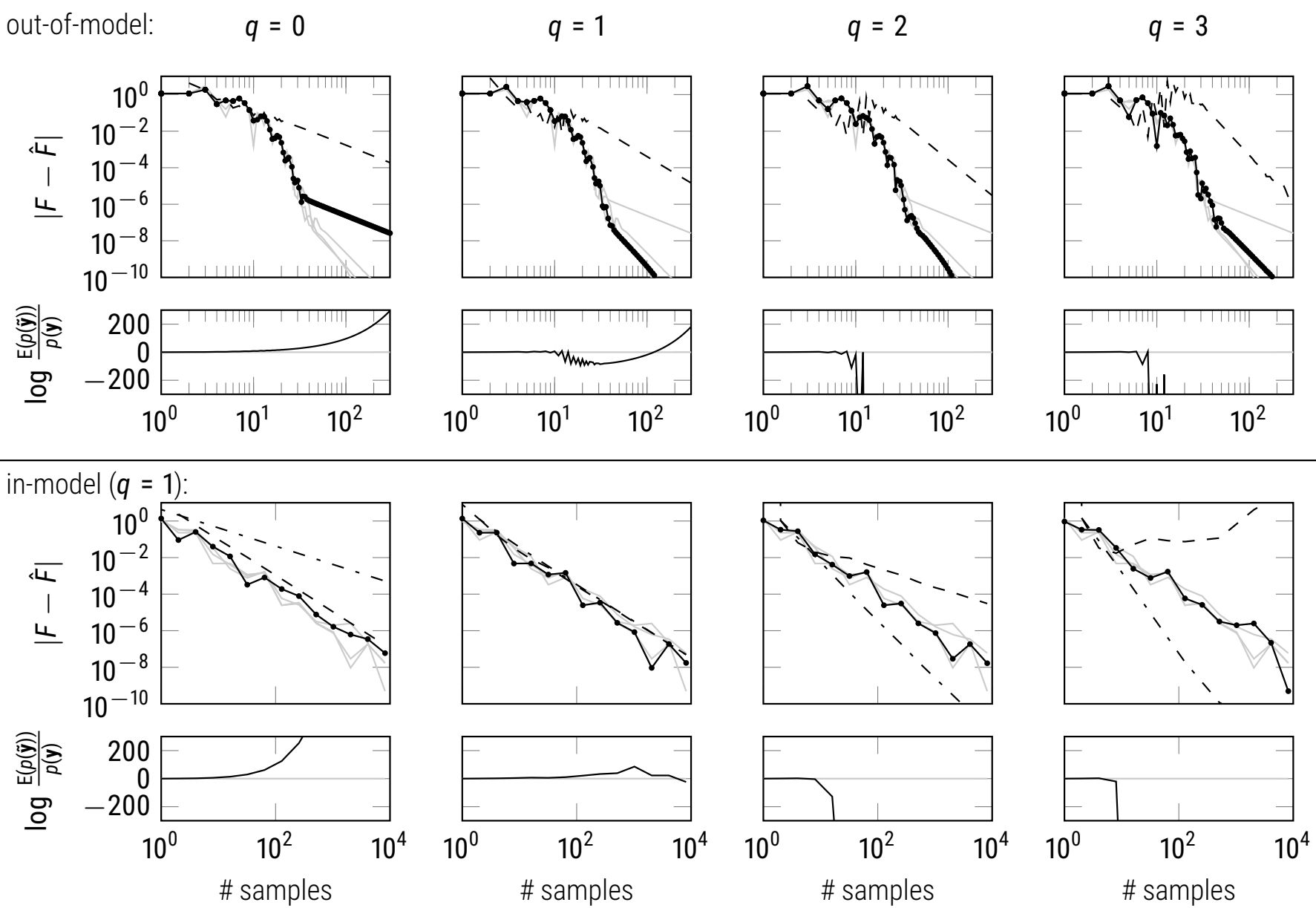
$$[Q(h)]_{ij} = \theta^2 \frac{h^{2q+3-i-j}}{(2q+3-i-j)(q+1-i)!(q+1-j)!}$$

same filter as before (using $[H]_i = \delta_{2i}$)!

Higher Order Filters

polynomial spline quadrature

[cf. Davis & Rabinowitz, 1984, §2.5]



Our first probabilistic numerical method

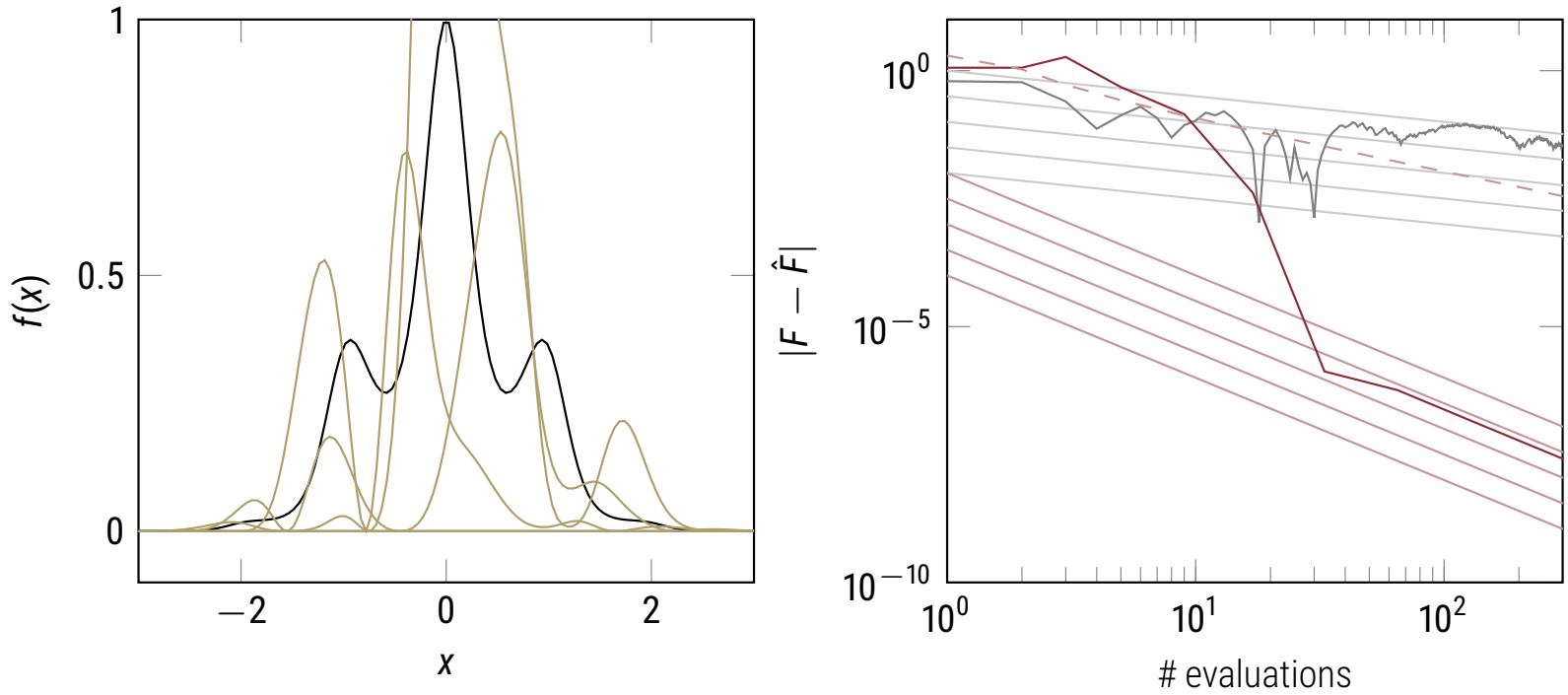
- the **trapezoid rule** arises as the **MAP estimate** under a Wiener process prior on the integrant
- the posterior variance can serve as a conservative notion of **uncertainty**
- hierarchical probabilistic inference gives an adaptive, quantitative **error estimate**
- **regular grids** emerge naturally from an information theoretic perspective.

Some “philosophical” observations:

- no random numbers! probabilistic numerics ≠ stochastic numerics
- **Gaussian measures** provide a very limited notion of uncertainty.
  But they work, and they are **fast**!
- this is basically just a new interpretation of a classic algorithm. But it shows PN methods **exist**, they can be as **fast** as classic methods and satisfy classic convergence requirements

Why? — Example: informative priors

WArped Sequential Active Bayesian Integration (WSABI) [Gunter, Osborne, Garnett, Hennig, Roberts. NIPS 2014]



a prior specifically for integration of probability measures

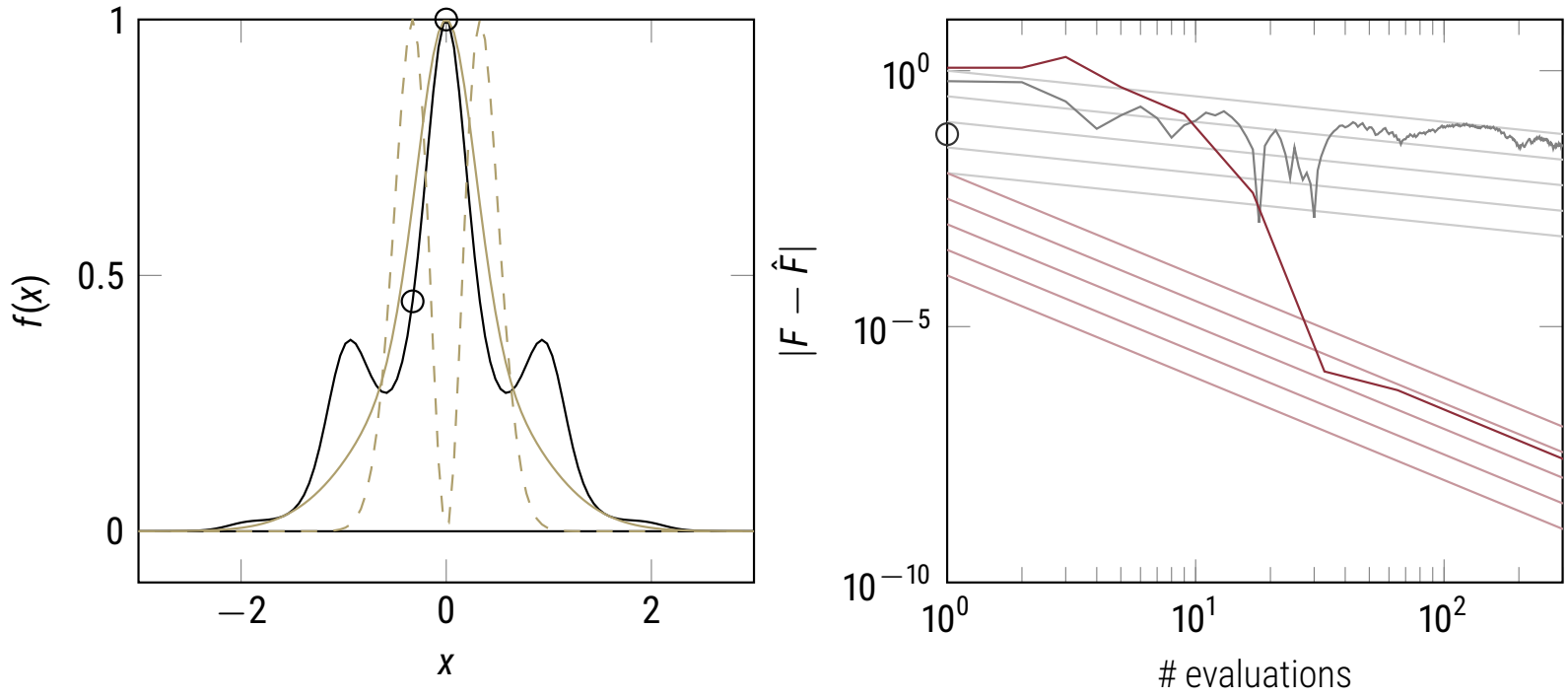
- $f > 0$ (f is probability measure)
- $f \propto \exp(-x^2)$ (f is product of prior and likelihood terms)
- $f \in \mathcal{C}^{\infty}$ (f is smooth)

Explicit prior knowledge yields reduces complexity.

[cf. **information-based complexity**. E.g. Novak, 1988. Clancy et al. 2013, arXiv 1303.2412v2]

Why? — Example: informative priors

WArped Sequential Active Bayesian Integration (WSABI) [Gunter, Osborne, Garnett, Hennig, Roberts. NIPS 2014]



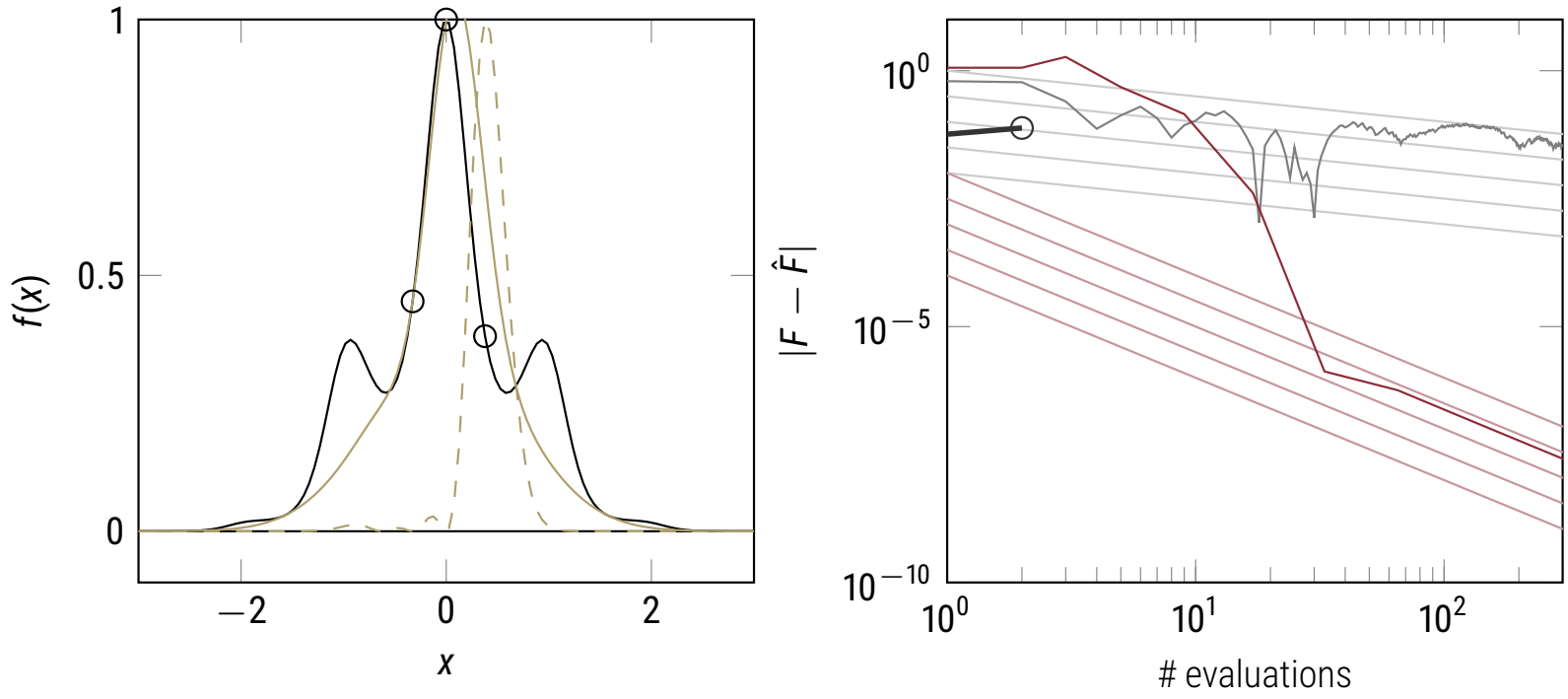
- adaptive node placement
- scales to, in principle, arbitrary dimensions
- faster (in wall-clock time) than MCMC

Explicit prior knowledge yields reduces complexity.

[cf. **information-based complexity**. E.g. Novak, 1988. Clancy et al. 2013, arXiv 1303.2412v2]

Why? — Example: informative priors

WArped Sequential Active Bayesian Integration (WSABI) [Gunter, Osborne, Garnett, Hennig, Roberts. NIPS 2014]



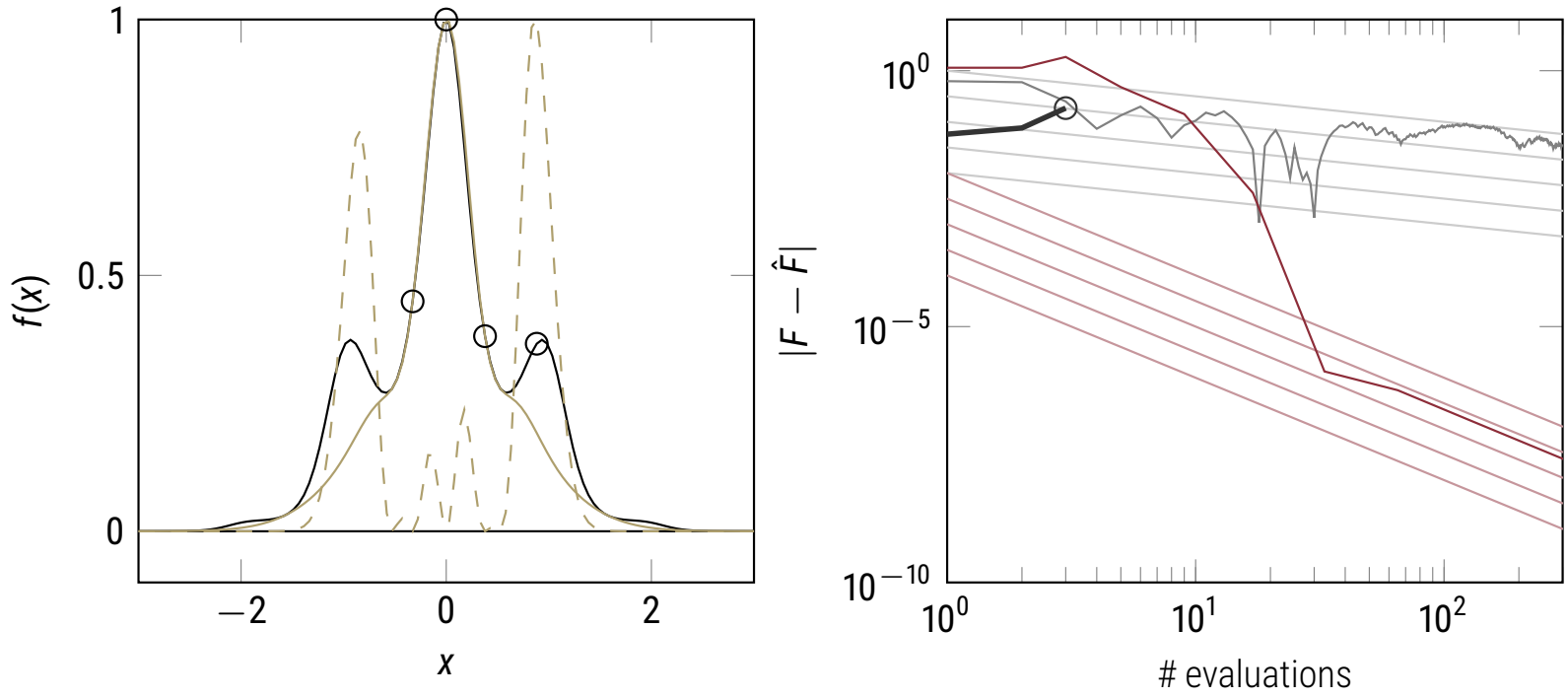
- adaptive node placement
- scales to, in principle, arbitrary dimensions
- faster (in wall-clock time) than MCMC

Explicit prior knowledge yields reduces complexity.

[cf. **information-based complexity**. E.g. Novak, 1988. Clancy et al. 2013, arXiv 1303.2412v2]

Why? — Example: informative priors

WArped Sequential Active Bayesian Integration (WSABI) [Gunter, Osborne, Garnett, Hennig, Roberts. NIPS 2014]



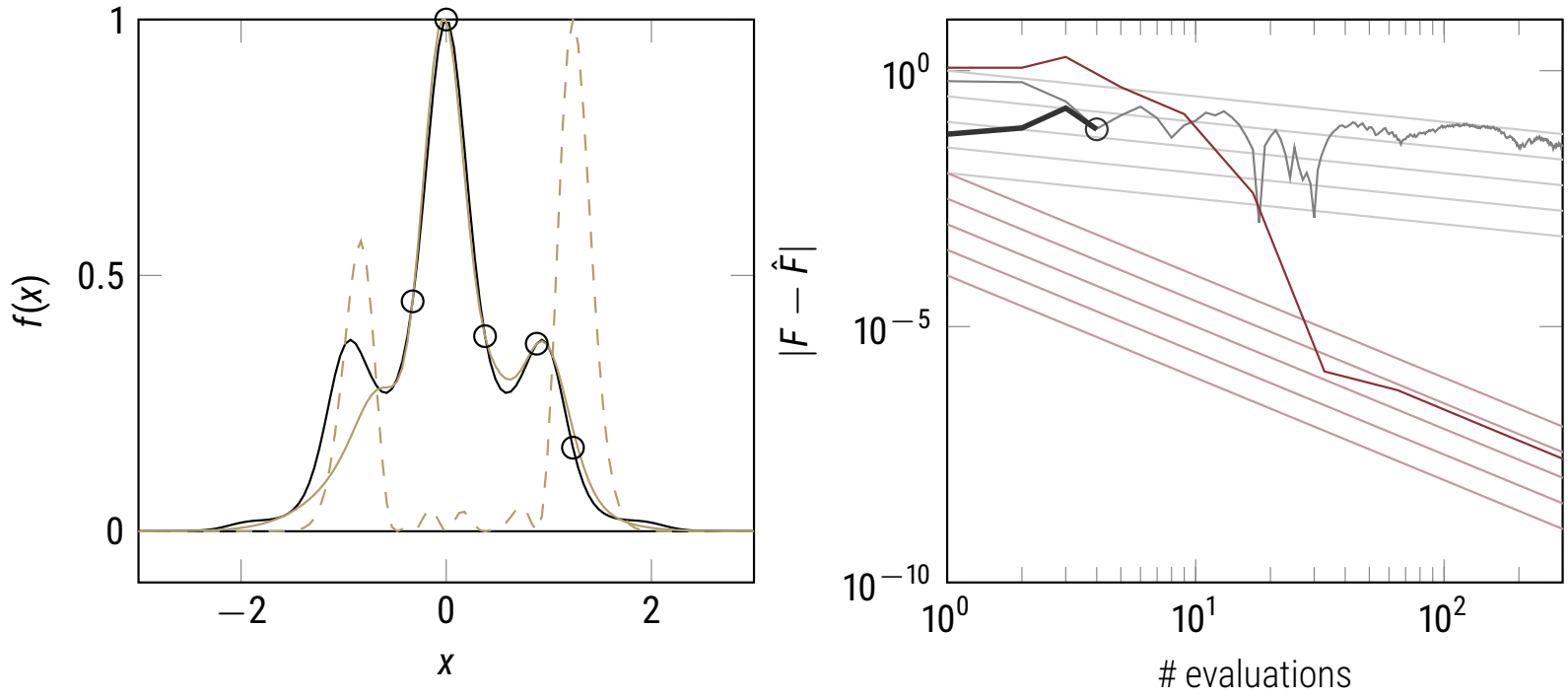
- adaptive node placement
- scales to, in principle, arbitrary dimensions
- faster (in wall-clock time) than MCMC

Explicit prior knowledge yields reduces complexity.

[cf. **information-based complexity**. E.g. Novak, 1988. Clancy et al. 2013, arXiv 1303.2412v2]

Why? — Example: informative priors

WArped Sequential Active Bayesian Integration (WSABI) [Gunter, Osborne, Garnett, Hennig, Roberts. NIPS 2014]



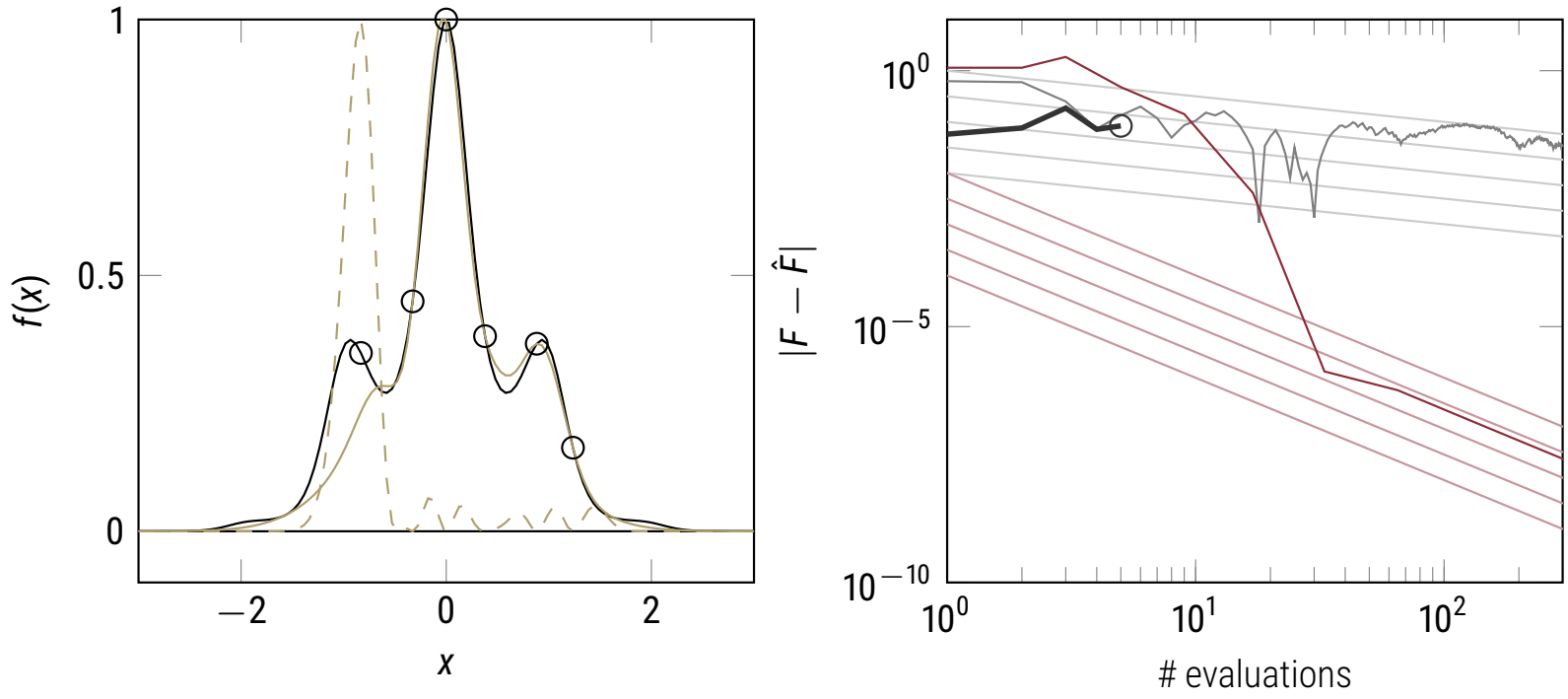
- adaptive node placement
- scales to, in principle, arbitrary dimensions
- faster (in wall-clock time) than MCMC

Explicit prior knowledge yields reduces complexity.

[cf. **information-based complexity**. E.g. Novak, 1988. Clancy et al. 2013, arXiv 1303.2412v2]

Why? — Example: informative priors

WArped Sequential Active Bayesian Integration (WSABI) [Gunter, Osborne, Garnett, Hennig, Roberts. NIPS 2014]



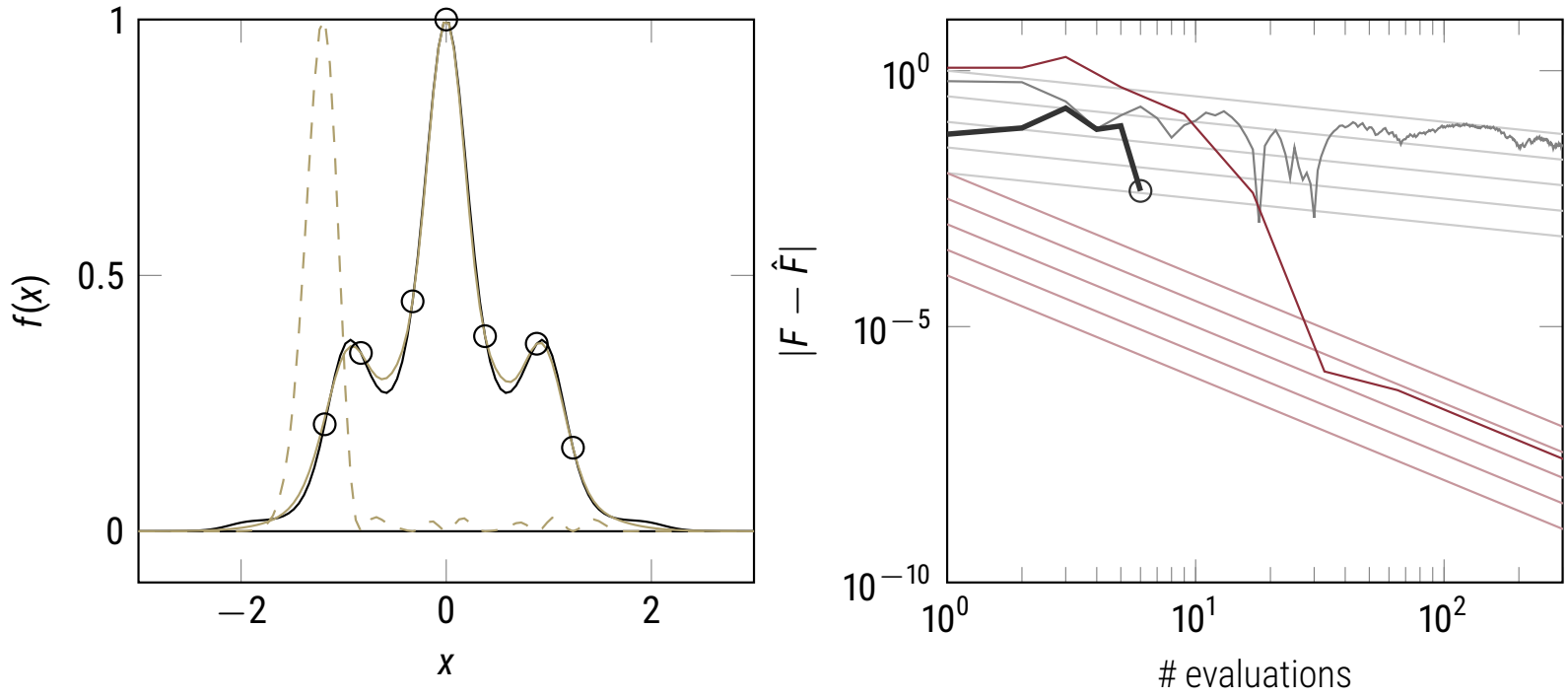
- adaptive node placement
- scales to, in principle, arbitrary dimensions
- faster (in wall-clock time) than MCMC

Explicit prior knowledge yields reduces complexity.

[cf. **information-based complexity**. E.g. Novak, 1988. Clancy et al. 2013, arXiv 1303.2412v2]

Why? — Example: informative priors

WArped Sequential Active Bayesian Integration (WSABI) [Gunter, Osborne, Garnett, Hennig, Roberts. NIPS 2014]



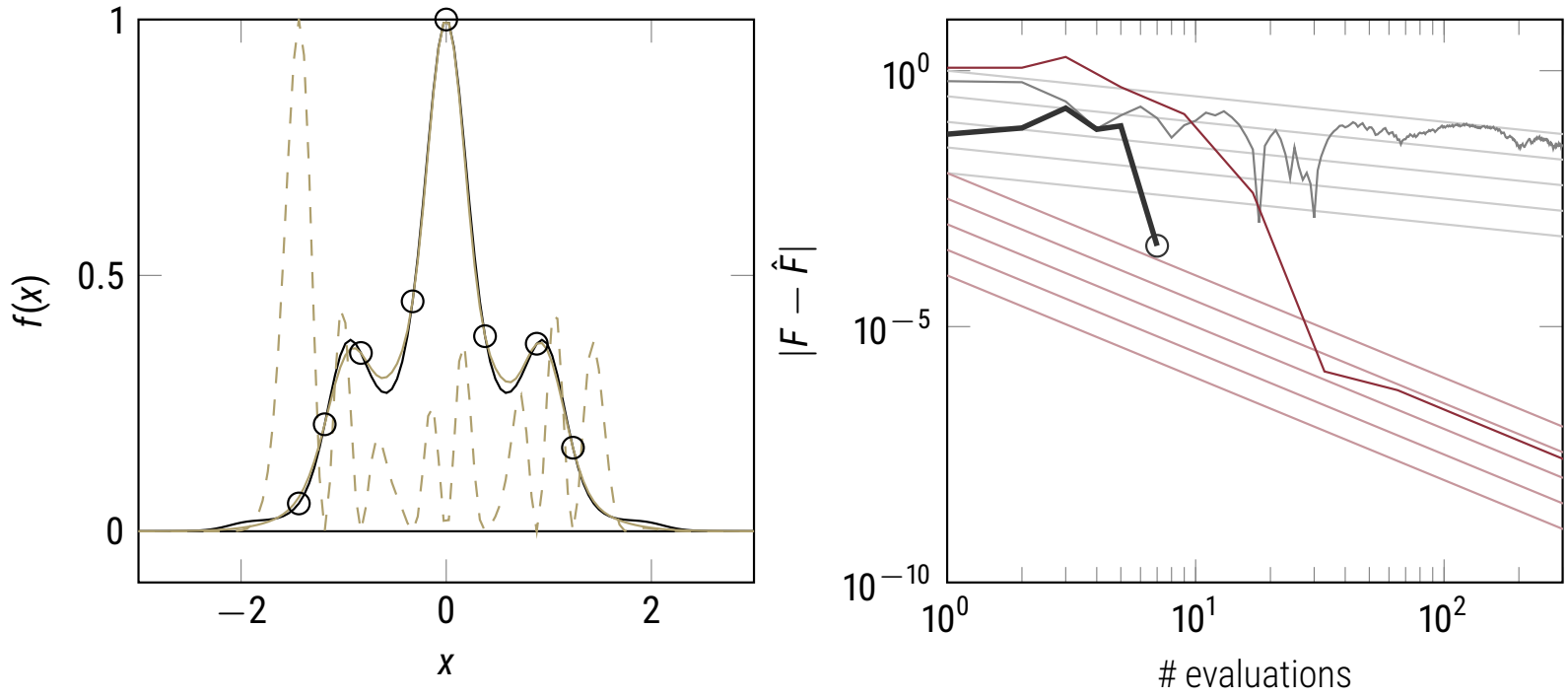
- adaptive node placement
- scales to, in principle, arbitrary dimensions
- faster (in wall-clock time) than MCMC

Explicit prior knowledge yields reduces complexity.

[cf. **information-based complexity**. E.g. Novak, 1988. Clancy et al. 2013, arXiv 1303.2412v2]

Why? — Example: informative priors

WArped Sequential Active Bayesian Integration (WSABI) [Gunter, Osborne, Garnett, Hennig, Roberts. NIPS 2014]



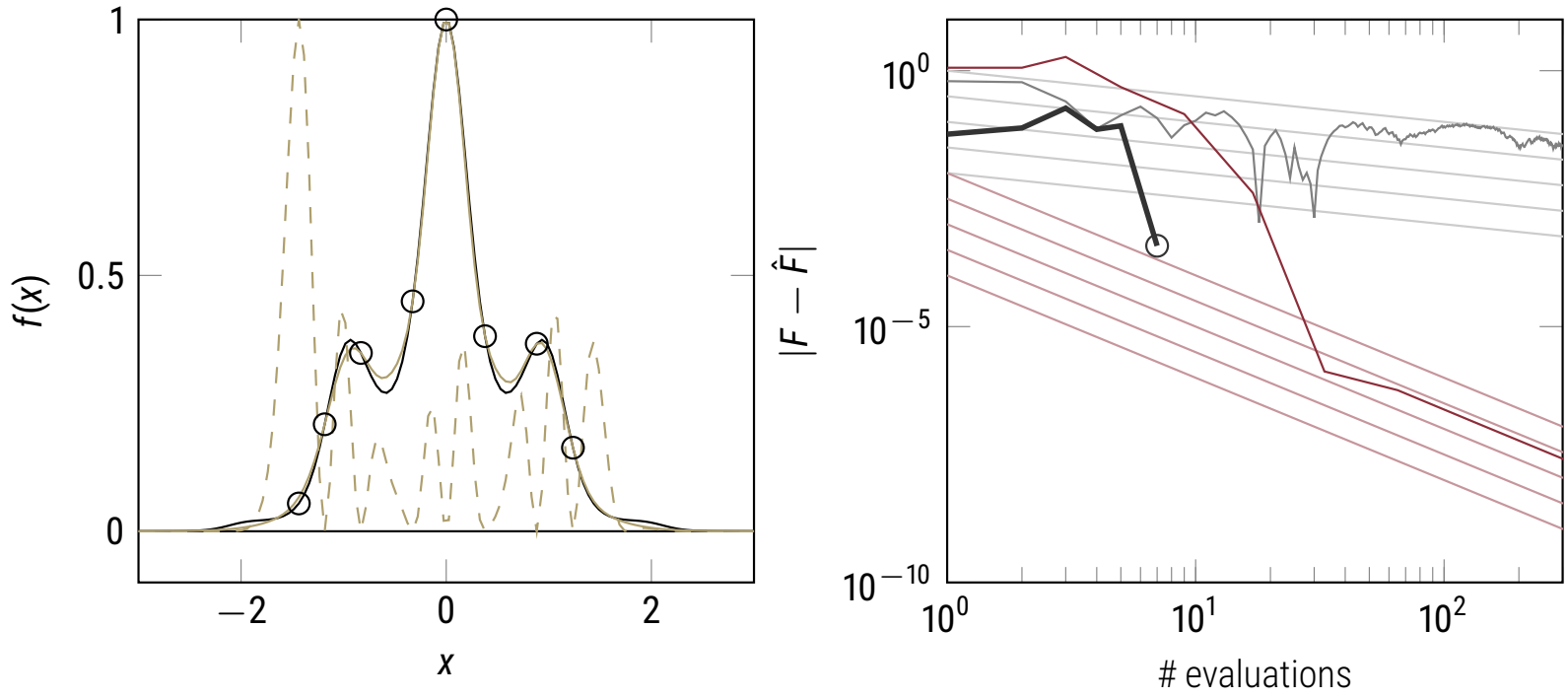
- adaptive node placement
- scales to, in principle, arbitrary dimensions
- faster (in wall-clock time) than MCMC

Explicit prior knowledge yields reduces complexity.

[cf. **information-based complexity**. E.g. Novak, 1988. Clancy et al. 2013, arXiv 1303.2412v2]

Why? — Example: informative priors

WArped Sequential Active Bayesian Integration (WSABI) [Gunter, Osborne, Garnett, Hennig, Roberts. NIPS 2014]



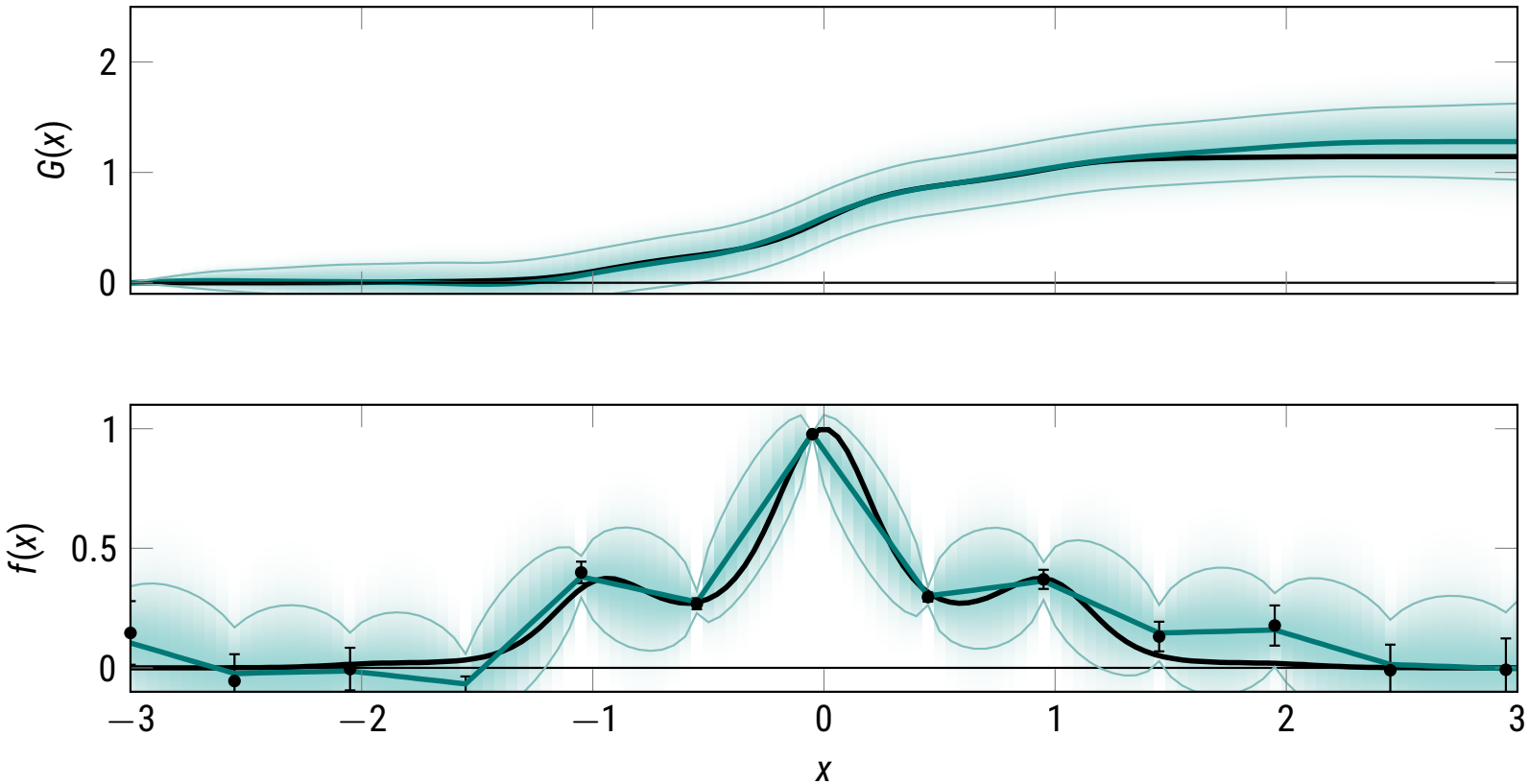
- adaptive node placement
- scales to, in principle, arbitrary dimensions
- faster (in wall-clock time) than MCMC

Explicit prior knowledge yields reduces complexity.

[cf. **information-based complexity**. E.g. Novak, 1988. Clancy et al. 2013, arXiv 1303.2412v2]

Why? — Arguments for Probabilistic Numerics

systemic uncertainty



- straightforward: just change R!
- obviously, convergence analysis now invalid

- **Gauss-Markov** processes allow linear-time inference in GP models, via **Filtering** and Smoothing
- **Gaussian Quadrature** rules can be interpreted as **MAP** estimators under Gaussian process priors
- **Error estimation** can be performed in closed form for noiseless observations
- **Probabilistic Numerics** formulation allows adaptation of **prior assumptions** and use of **uncertain evaluations**

coming up:

- **ordinary differential equations:** a slightly more challenging setting

These slides can be found at
`http://tinyurl.com/Dobbiaco-Hennig-3`