Gaussian Processes - Part III
Advanced Topics

Philipp Hennig

MLSS 2013
30 August 2013

MAX-PLANCK-GESELLSCHAFT

Max Planck Institute for Intelligent Systems
Department of Empirical Inference
Tübingen, Germany

Gaussians have been discovered before

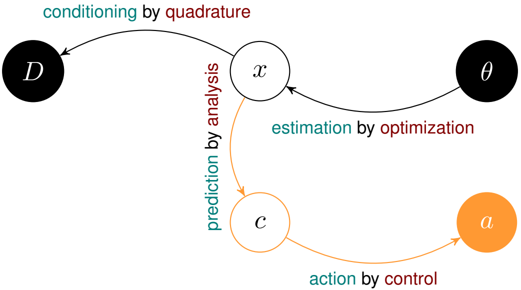
In virtually every area of science affected by uncertainty

- Thermodynamics — Brownian motion, Ornstein-Uhlenbeck process
- stochastic calculus — stochastic differential equations, Itō calculus
- control theory — stochastic control, Kalman filter
- signal processing — filtering
- other communities use other names for the same concept
  Kriging; Ridge-Regression, Kolmogorov-Wiener prediction; least-squares regression; Wiener process; Brownian bridge, ...
- Now: Gaussians show up in numerical methods, too ...
  quadrature, optimization, solving ODEs, control ...

Gaussian processes are central to many machine learning techniques, and all areas of quantitative science.

The big picture

we need a coherent framework for hierarchical machine learning



- uncertainty caused by finite computations should be accounted for
- uncertainty should propagate among numerical methods
- joint language required: probability

"off-the-shelf" methods are convenient, but not always efficient.

Numerical algorithms are the elements of inference

inferring solutions of non-analytic problems http://www.probabilistic-numerics.org

Numerical algorithms
estimate (infer) an intractable property of a function
given evaluations of function values.

quadrature	estimate $\int_a^b f(x)\,dx$	given $\{f(x_i)\}$
optimization	estimate $\arg\min_x f(x)$	given $\{f(x_i), \nabla f(x_i)\}$
analysis	estimate $x(t)$ under $x' = f(x,t)$	given $\{f(x_i, t_i)\}$
control	estimate $\min_u x(t,u)$ under $x' = f(x,t,u)$	$\{f(x_i, t_i, u_i)\}$

- even deterministic problems can be uncertain
- not a new idea[1], but rarely studied

We need a theory of probabilistic numerics.

Gaussians, because of their connection to linear functions, are at the heart of probabilistic interpretations of numerics.

[1] H. Poincaré, 1896, Diaconis 1988, O'Hagan 1992

Recall: GPs are closed under linear maps

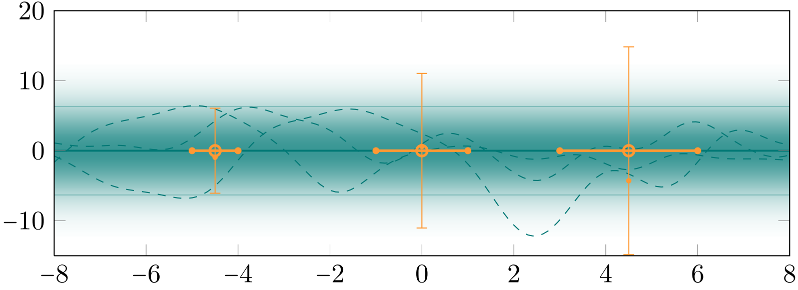
$$p(z) = \mathcal{N}(z; \mu, \Sigma) \quad \Rightarrow \quad p(Az) = \mathcal{N}(Az, A\mu, A\Sigma A^\top)$$

- this is not restricted to finite linear operators (matrices) A
- $A(x) = \mathbb{I}(a < x < b)$ gives $Af = \int_a^b f(x)\,\mathrm{d}x$

$$p\left(\int_a^b f(x)\,\mathrm{d}x, \int_c^d f(x)\,\mathrm{d}x\right) = \mathcal{N}\left[\begin{pmatrix}\int_a^b f(x)\,\mathrm{d}x \\ \int_c^d f(x)\,\mathrm{d}x\end{pmatrix}; \begin{pmatrix}\int_a^b \mu(x)\,\mathrm{d}x \\ \int_c^d \mu(x)\,\mathrm{d}x\end{pmatrix}, \begin{pmatrix}\int_a^b \int_a^b k(x,x')\,\mathrm{d}x\,\mathrm{d}x' & \int_a^b \int_c^d k(x,x')\,\mathrm{d}x\,\mathrm{d}x' \\ \int_a^b \int_c^d k(x,x')\,\mathrm{d}x\,\mathrm{d}x' & \int_c^d \int_c^d k(x,x')\,\mathrm{d}x\,\mathrm{d}x'\end{pmatrix}\right]$$

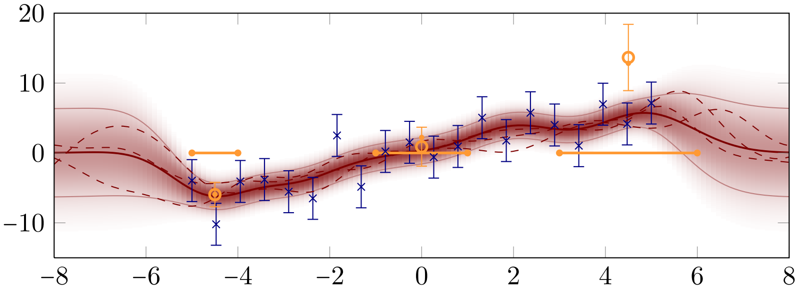
Inferring $F = \int f$ from observations of f

quadrature



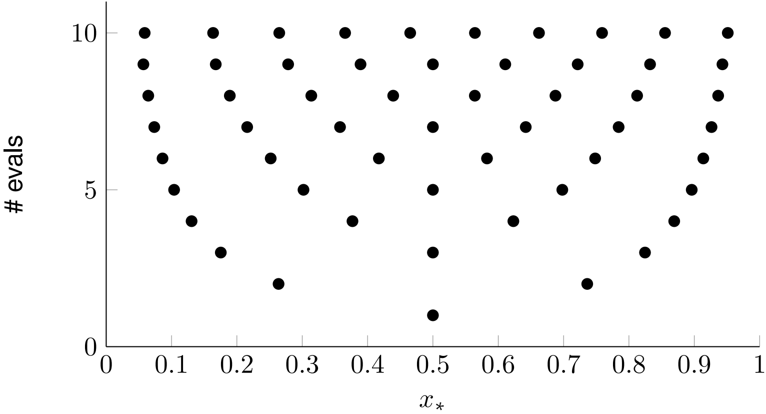
Inferring $F = \int f$ from observations of f

quadrature



Quadrature with GPs

A O'Hagan, 1991; T Minka, 2000; M Osborne et al., 2012

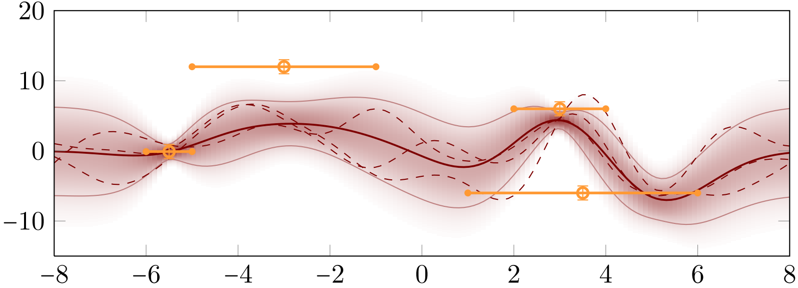


- say what functions you expect to integrate
- find $\arg\min_X [k_{Ff_X} - k_{Ff_X} k_{f_X f_X}^{-1} k_{f_X F}]$ (depends on kernel!)
- Bayesian quadrature

Gaussian processes can be used to construct quadrature rules.

Inferring f from observations of F

$\mu_{f\,|\,F_X} = \mu_f + k_{f F_X} k_{F_X F_X}^{-1} (F_X - \int_X \mu)$ $\qquad$ $k_{f f\,|\,F_X} = k_{f f} - k_{f F_X} k_{F_X F_X}^{-1} k_{F_X f}$



Optimization

continuous, nonlinear, unconstrained

For $f : \mathbb{R}^N \rightharpoonup \mathbb{R}$, find local minimum $\arg\min f(x)$, starting at x_0.

An old idea: Newton's method

$$f(x) \approx f(x_t) + (x - x_t)^\top \nabla f(x_t) + \frac{1}{2}(x - x_t)^\top \underbrace{\nabla\nabla^\top f(x_t)}_{=:B(x_t)}(x - x_t)$$

$$\rightharpoonup \qquad x_{t+1} = x_t - B^{-1}(x_t)\nabla f(x_t)$$

Cost: $\mathcal{O}(N^3)$

High-dimensional optimization requires
giving up knowledge in return for lower cost.

Quasi-Newton methods (think BFGS, DFP, ...)

aka. variable metric optimization — low rank estimators for Hessians

- Instead of evaluating Hessian, build (low-rank) estimator fulfilling local difference relation ...

$$\nabla f(x_{t+1}) - \nabla f(x_t) = B_{t+1}(x_{t+1} - x_t)$$
$$y_t = B_{t+1} s_t$$

- ...otherwise close to previous estimator in $\|B_{t+1} - B_t\|_{F,V}$
- ...so minimize regularised loss

$$\begin{aligned}
B_{t+1} &= \operatorname*{arg\,min}_{B \in \mathbb{R}^{N \times N}} \left\{ \lim_{\beta \to 0} \frac{1}{\beta} \|y_t - B s_t\|_V^2 + \|B - B_t\|_{F,V}^2 \right\} \\
&= \lim_{\beta \to 0} \operatorname*{arg\,max}_B \mathcal{N}(y_t; B s_t, \beta V) \mathcal{N}(\overrightarrow{B}; \overrightarrow{B}_t, V \otimes V) \\
&= \operatorname*{arg\,max}_B \underbrace{\mathcal{N}\left[B; B_t + \frac{(y_t - B_t s_t) V s_t^\top}{s_t^\top V s_t}, V \otimes \left(V - \frac{V s s^\top V}{s^\top V s}\right)\right]}_{\text{posterior}}
\end{aligned}$$

Quasi-Newton methods perform local maximum a-posteriori Gaussian inference on the Hessian's elements.

Optimization with GPs

nonparametric quasi-Newton methods

Hennig & Kiefel, ICML 2012, JMLR 2013

- Idea: replace

$$\nabla f(x_{t+1}) - \nabla f(x_t) \approx B(x_{t+1} - x_t)$$
$$\rightarrow \qquad = \int_{x_t}^{x_{t+1}} B(x)\, dx$$

- Gaussian process prior on $B(x^\top, x)$

$$p(B) = \mathcal{GP}(B, B_0(x^\top, x), k(x^\top, x'^\top) \otimes k(x, x'))$$

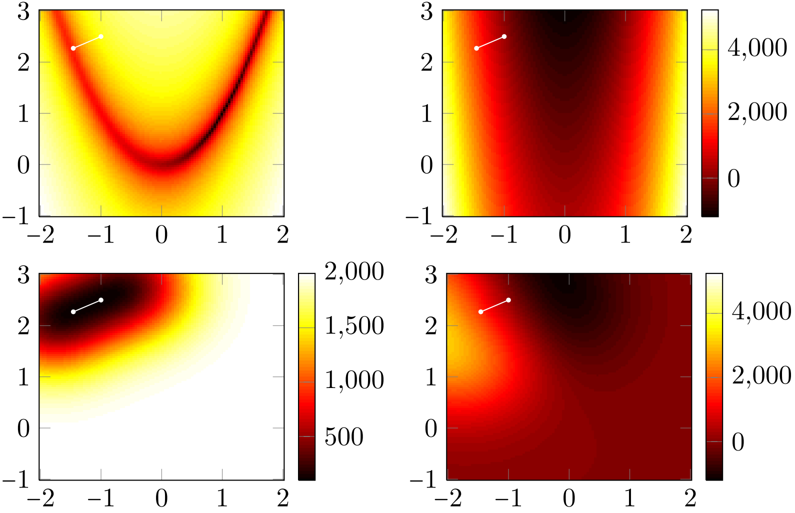
- Gaussian likelihoods

$$p(y_i(x^\top) \,|\, B, s_i) = \lim_{\beta \to 0} \mathcal{N}\left(y_i; \sum_m s_{im} \int_0^1 B(x^\top, x(t))\, \mathrm{d}t, k(x^\top, x'^\top) \otimes \beta\right)$$
$$p(y_i(x)^\top \,|\, B, s_i^\top) = \lim_{\beta \to 0} \mathcal{N}\left(y_i^\top; \sum_m s_{im}^\top \int_0^1 B(x^\top(t), x)\, \mathrm{d}t, \beta \otimes k(x, x')\right)$$

- posterior of same algebraic form as before, but with linear maps of nonlinear (integral of k) entries.
- same computational complexity as L-BFGS (Nocedal, 1980): $\mathcal{O}(N)$

A consistent model of the Hessian function

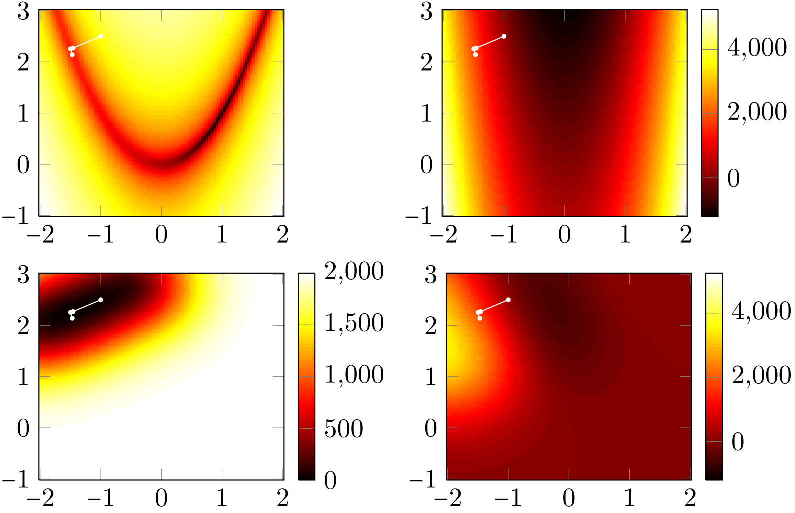
nonparametric inference on elements of the Hessian — P.H. & M. Kiefel, ICML 2012, JMLR 2013

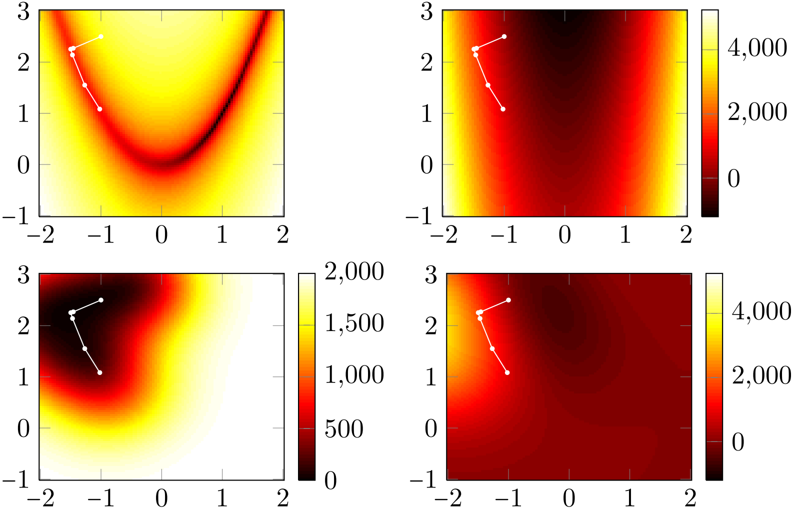


A consistent model of the Hessian function

nonparametric inference on elements of the Hessian

P.H. & M. Kiefel, ICML 2012, JMLR 2013

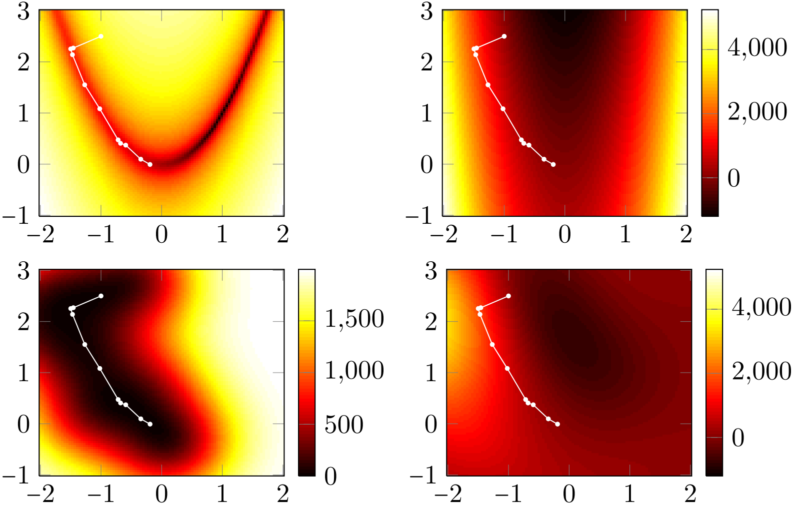



3
2
1
0
−1
−2
−1
0
1
2
4,000
2,000
0
2,000
1,500
1,000
500
0

A consistent model of the Hessian function

nonparametric inference on elements of the Hessian

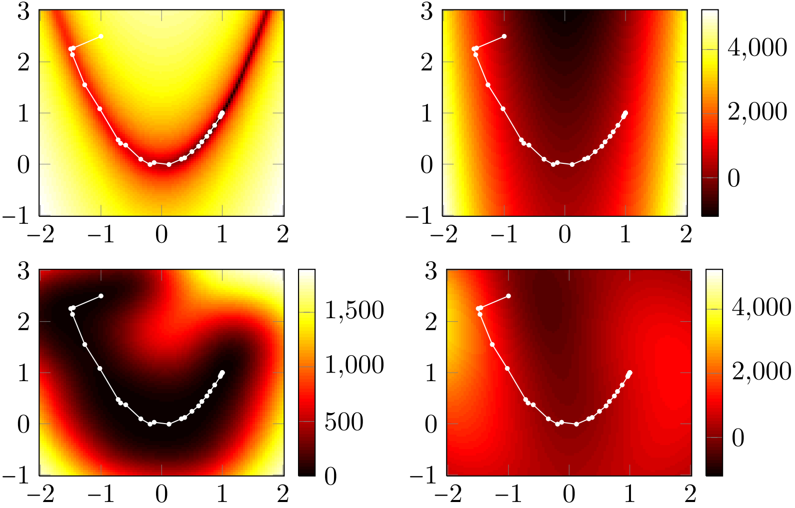
P.H. & M. Kiefel, ICML 2012, JMLR 2013



A consistent model of the Hessian function

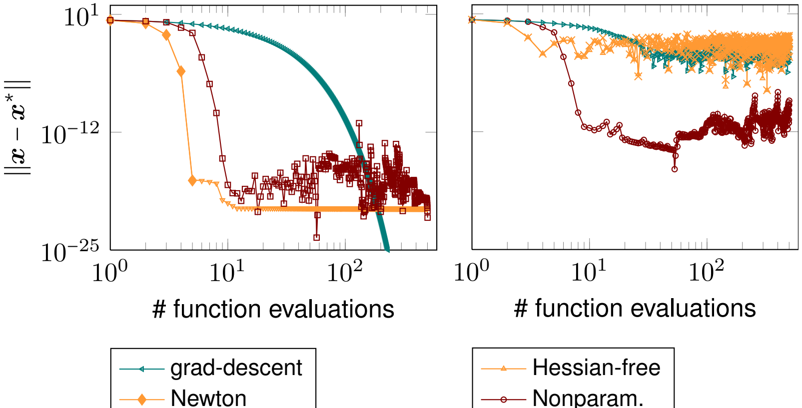
nonparametric inference on elements of the Hessian

P.H. & M. Kiefel, ICML 2012, JMLR 2013



Learning nonparametric models of Hessians allows

- optimizing noisy functions
- dynamically changing functions
- parallelization
- …

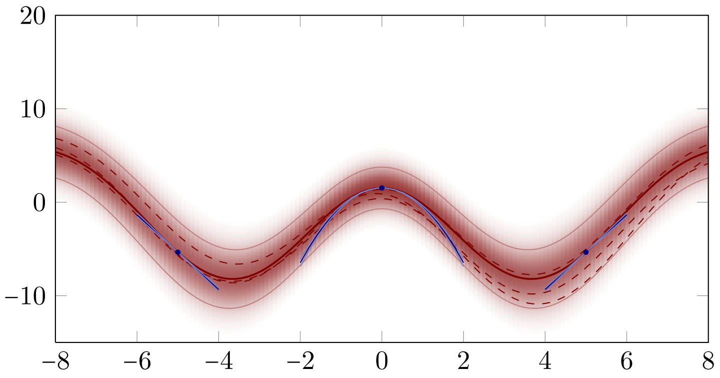


Gaussian processes can be used in optimization.

GPs are closed under differentiation

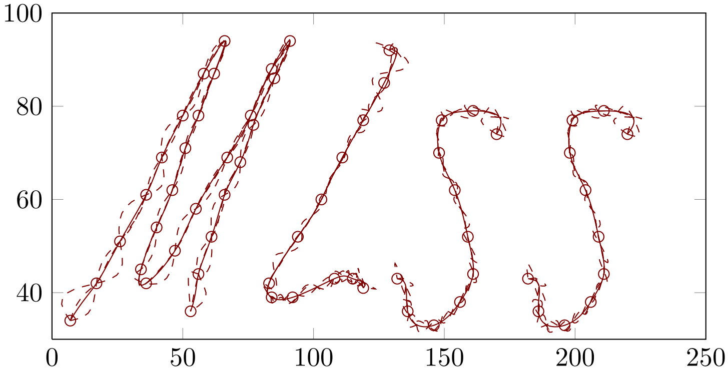
Rasmussen & Williams, 2006, §9.4

$\mu_{f\mid f'_X} = \mu_f + k_{ff'_X} k^{-1}_{f'_X f'_X} (f'_X - \mu'_{f_X})$ $\qquad$ $k_{ff\mid f'_X} = k_{ff} - k_{ff'_X} k^{-1}_{f'_x f'_X} k_{f'_X f}$



GPs can have multiple outputs

Reminder of Part I

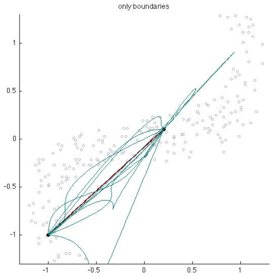


Solving ODEs with GPs

observe $c'(t)$, infer $c(t)$

Skilling, 1991

solve $c'(t) = f(c(t), t)$ such that $c(0) = a$ and $c(1) = b$



$$p(c(t)) = \mathcal{GP}(c; \mu_c, V \otimes k)$$
$$p(y_t \mid c) = \mathcal{N}(f(\hat{c}_t; t); \dot{c}_t, U)$$

- repeatedly estimate $\hat{c}_t$ using GP posterior mean to "observe" $c'(t) = f(\hat{c}_t) + \delta_f$
- estimate error in this observation by propagating Gaussian uncertainty through f.

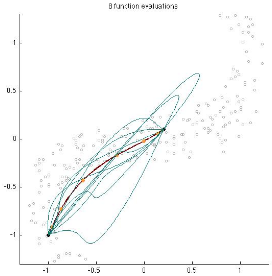
Recent work:

- Chkrebtii, Campbell, Girolami, Calderhead `http://arxiv.org/abs/1306.2365`
- Hennig & Hauberg `http://arxiv.org/abs/1306.0308`

Solving ODEs with GPs

observe $c'(t)$, infer $c(t)$ — Skilling, 1991

solve $c'(t) = f(c(t), t)$ such that $c(0) = a$ and $c(1) = b$



$$p(c(t)) = \mathcal{GP}(c; \mu_c, V \otimes k)$$
$$p(y_t \mid c) = \mathcal{N}(f(\hat{c}_t; t); \dot{c}_t, U)$$

- repeatedly estimate $\hat{c}_t$ using GP posterior mean to "observe" $c'(t) = f(\hat{c}_t) + \delta_f$
- estimate error in this observation by propagating Gaussian uncertainty through f.

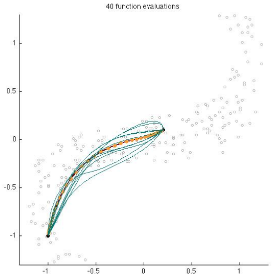
Recent work:

- Chkrebtii, Campbell, Girolami, Calderhead `http://arxiv.org/abs/1306.2365`
- Hennig & Hauberg `http://arxiv.org/abs/1306.0308`

Solving ODEs with GPs

observe $c'(t)$, infer $c(t)$ Skilling, 1991

solve $c'(t) = f(c(t), t)$ such that $c(0) = a$ and $c(1) = b$



$$p(c(t)) = \mathcal{GP}(c; \mu_c, V \otimes k)$$
$$p(y_t \mid c) = \mathcal{N}(f(\hat{c}_t; t); \dot{c}_t, U)$$

- repeatedly estimate $\hat{c}_t$ using GP posterior mean to "observe" $c'(t) = f(\hat{c}_t) + \delta_f$
- estimate error in this observation by propagating Gaussian uncertainty through f.

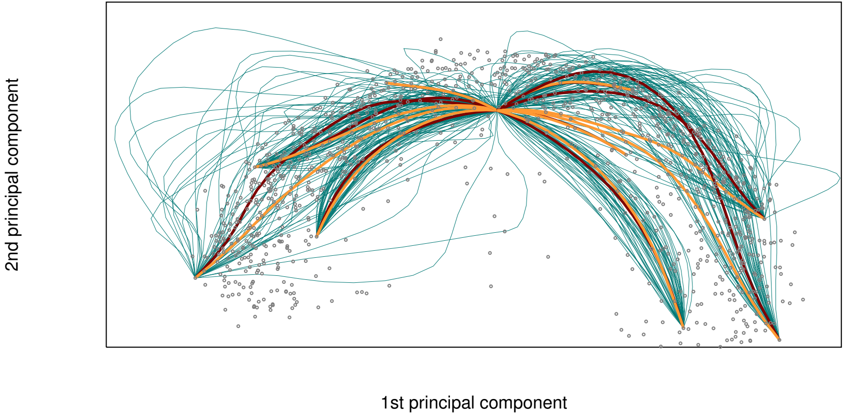
Recent work:

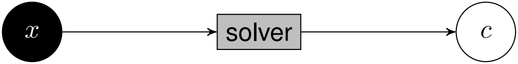
- Chkrebtii, Campbell, Girolami, Calderhead `http://arxiv.org/abs/1306.2365`
- Hennig & Hauberg `http://arxiv.org/abs/1306.0308`

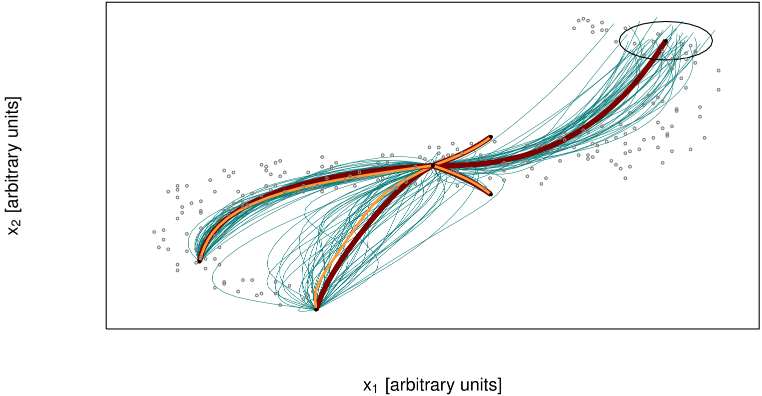
The Advantages of a Probabilistic Formulation

joint uncertainty over solution

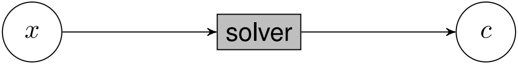
Hennig & Hauberg, under review







x2 [arbitrary units]
x1 [arbitrary units]

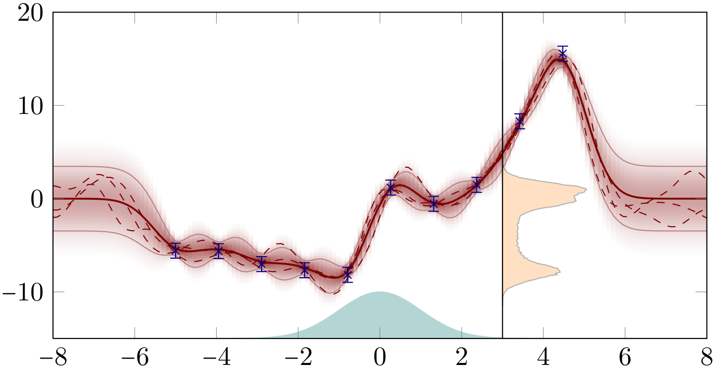


x
solver
c

Gaussian processes can be used to solve differential equations.

Lots of “Gaussian integrals” are known

and can be used to map uncertainty through almost any function — see e.g. M. Deisenroth’s PhD, 2010



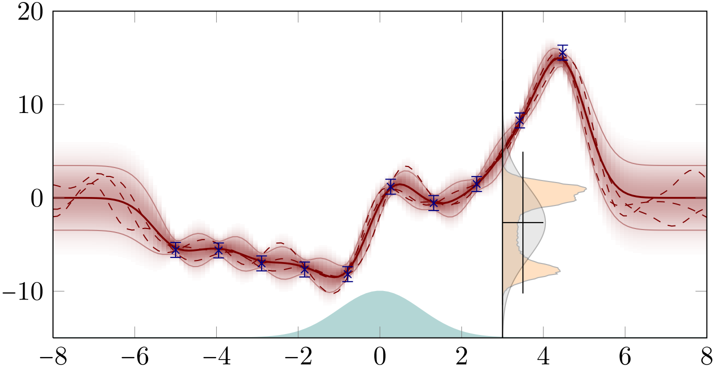
- write $f(x) = \sum_i \phi_i(x)^\top w$ such that

$$\int \phi_i(x)\mathcal{N}(x;\mu,\Sigma)\,\mathrm{d}x \qquad \int \phi_i(x)\phi_j(x)\mathcal{N}(x;\mu,\Sigma)\,\mathrm{d}x$$

is analytic

Lots of “Gaussian integrals” are known

and can be used to map uncertainty through almost any function — see e.g. M. Deisenroth’s PhD, 2010



$$\int f(x)\mathcal{N}(x;\mu,\Sigma)\,\mathrm{d}x = \sum_i w_i \int \phi_i(x)\mathcal{N}(x;\mu,\Sigma)\,\mathrm{d}x$$

$$\int f^2(x)\mathcal{N}(x;\mu,\Sigma)\; x = \sum_i \sum_j w_i w_j \int \phi_i(x)\phi_j(x)\mathcal{N}(x;\mu,\Sigma)\,\mathrm{d}x$$

- also works if $f \in \mathbb{R}^N$, and if $p(w) = \mathcal{N}(w;m,V)$

Some useful Gaussian integrals

an expressive basis set for function approximation

$$\int x^p \mathcal{N}(x;0,\sigma^2)\,\mathrm{d}x = \begin{cases} 0 & \text{if } p \text{ odd} \\ \sigma^p \prod_{i=1:2:p-1} i & \text{if } p \text{ even} \end{cases}$$

$$\int |x|^p \mathcal{N}(x;0,\sigma^2)\,\mathrm{d}x = \frac{\sigma^p}{\sqrt{\pi}} 2^{p/2} \Gamma\left(\frac{p+1}{2}\right)$$

$$\int (x-m)^\top V (x-m) \mathcal{N}(x;\mu,\Sigma)\,\mathrm{d}x = (\mu-m)^\top V (\mu-m) + \mathrm{tr}[V\Sigma]$$

$$\int \mathcal{N}(x;a,A)\mathcal{N}(x;b;B)\,\mathrm{d}x = \mathcal{N}(a,b,A+B)$$

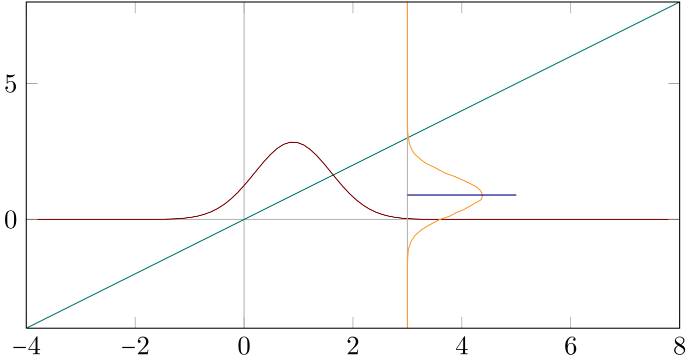
$$\int \int_{-\infty}^{(x-m)/s} \mathcal{N}(\tilde{x},0,1)\,\mathrm{d}\tilde{x}\,\mathcal{N}(x;\mu,\sigma^2)\,\mathrm{d}x = \int_{-\infty}^{(\mu-m)/\sqrt{(}s^2+\sigma^2)} \mathcal{N}(\tilde{x},0,1)$$

$$= \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\mu-m}{\sqrt{2(s^2+\sigma^2)}}\right)\right]$$

c.f. DB Owen, A table of normal integrals. Comm. Stat.-Sim. Comp. 1980

Expected values of monomials

for moment computations

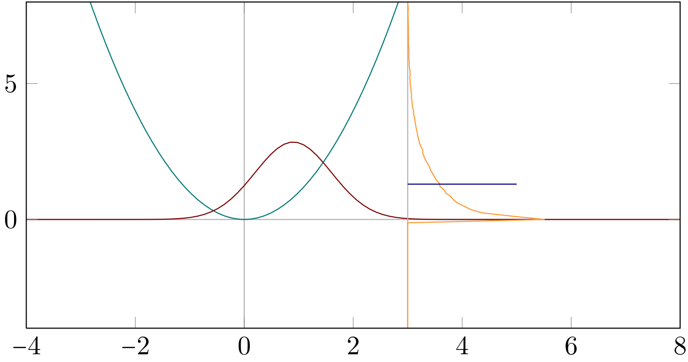


$$\int x^p \mathcal{N}(x;\mu,\sigma) = \sigma^p(-i\sqrt{2}\operatorname{sgn}\mu)^p U\left(-\frac{p}{2},\frac{1}{2},-\frac{1}{2}\frac{\mu^2}{\sigma^2}\right) \qquad p \in \mathbb{N}_0$$

where U is Tricomi's confluent hypergeometric function (cheap)

Expected values of monomials

for moment computations

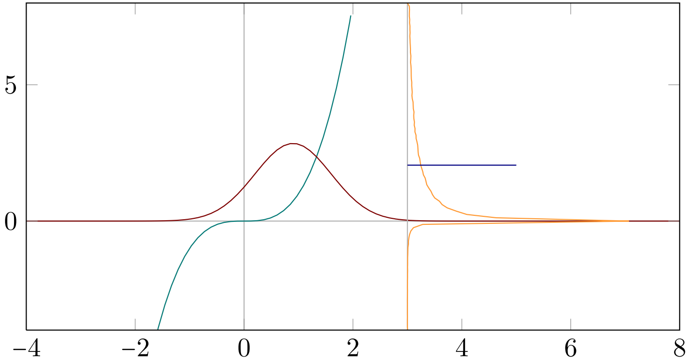


$$\int x^p \mathcal{N}(x;\mu,\sigma) = \sigma^p(-i\sqrt{2}\operatorname{sgn}\mu)^p U\left(-\frac{p}{2}, \frac{1}{2}, -\frac{1}{2}\frac{\mu^2}{\sigma^2}\right) \qquad p \in \mathbb{N}_0$$

where U is Tricomi's confluent hypergeometric function (cheap)

Expected values of monomials

for moment computations

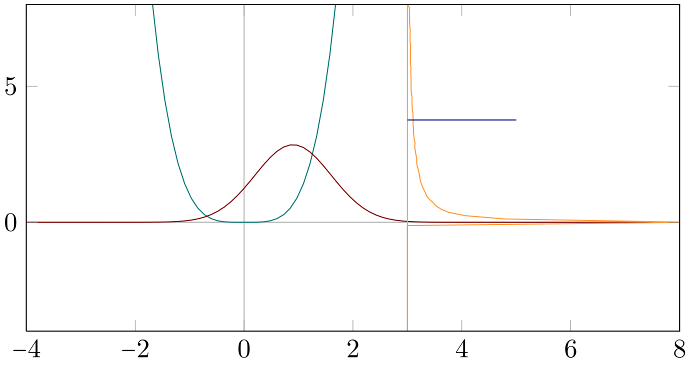


$$\int x^p \mathcal{N}(x;\mu,\sigma) = \sigma^p(-i\sqrt{2}\operatorname{sgn}\mu)^p U\left(-\frac{p}{2}, \frac{1}{2}, -\frac{1}{2}\frac{\mu^2}{\sigma^2}\right) \qquad p \in \mathbb{N}_0$$

where U is Tricomi's confluent hypergeometric function (cheap)

Expected values of monomials

for moment computations



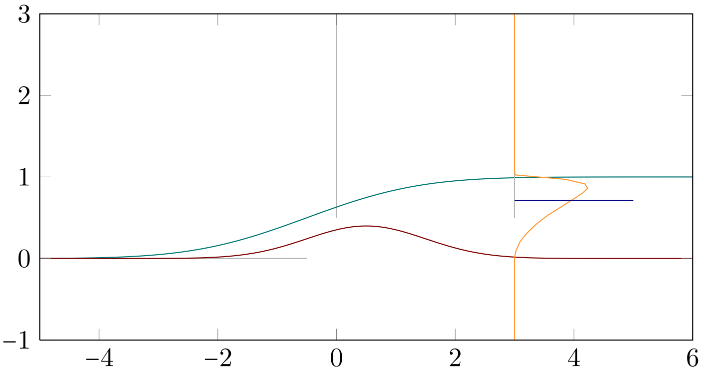
$$\int x^p \mathcal{N}(x;\mu,\sigma) = \sigma^p(-i\sqrt{2}\operatorname{sgn}\mu)^p U\left(-\frac{p}{2}, \frac{1}{2}, -\frac{1}{2}\frac{\mu^2}{\sigma^2}\right) \qquad p \in \mathbb{N}_0$$

where U is Tricomi's confluent hypergeometric function (cheap)

Expected values of error functions

for moment computations

e.g. Rasmussen & Williams, §3.9



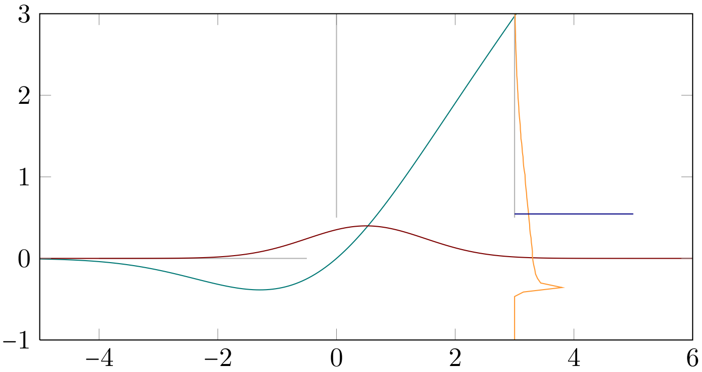
$$\Phi(x) = \int_{-\infty}^{x} \mathcal{N}(x; 0, 1)\, \mathrm{d}x \qquad z = \frac{\mu - m}{\sqrt{v^2 + \sigma^2}}$$

$$\int \Phi\left(\frac{x - m}{v}\right) \mathcal{N}(x; \mu, \sigma) = \Phi(z)$$

Expected values of error functions

for moment computations

e.g. Rasmussen & Williams, §3.9



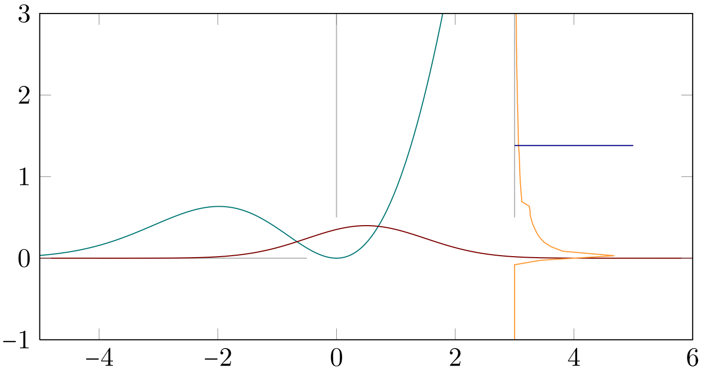
$$\Phi(x) = \int_{-\infty}^{x} \mathcal{N}(x;0,1)\,\mathrm{d}x \qquad z = \frac{\mu - m}{\sqrt{v^2 + \sigma^2}}$$

$$\int x\Phi\left(\frac{x-m}{v}\right)\mathcal{N}(x;\mu,\sigma) = \mu\Phi(z) + \frac{\sigma^2}{\sqrt{v^2+\sigma^2}}\mathcal{N}(z;0,1)$$

Expected values of error functions

for moment computations

e.g. Rasmussen & Williams, §3.9



$$\Phi(x) = \int_{-\infty}^{x} \mathcal{N}(x; 0, 1)\, \mathrm{d}x \qquad z = \frac{\mu - m}{\sqrt{v^2 + \sigma^2}}$$

$$\int x^2 \Phi\left(\frac{x - m}{v}\right) \mathcal{N}(x; \mu, \sigma) = (\mu^2 + \sigma^2)\Phi(z) + \left(2\mu \frac{\sigma^2}{\sqrt{v^2 + \sigma^2}} - \frac{z\sigma^4}{v^2 + \sigma^2}\right) \mathcal{N}(z; 0, 1)$$

Treating Cancer with GPs

Analytical Probabilistic Modelling in Radiation Therapy

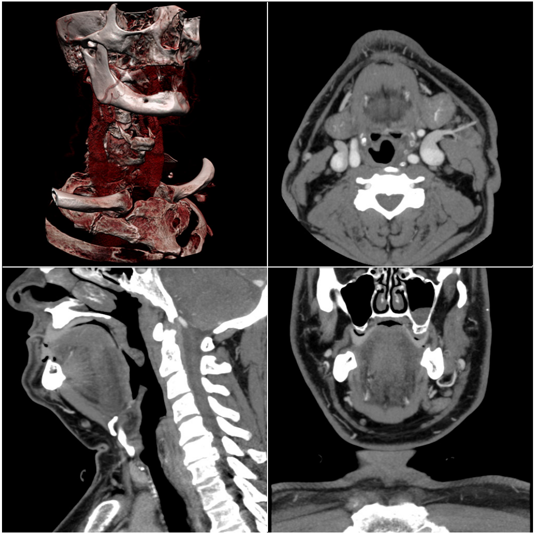
Bangert, Hennig, Oelfke, 2013

image source: wikipedia

the data

CT images

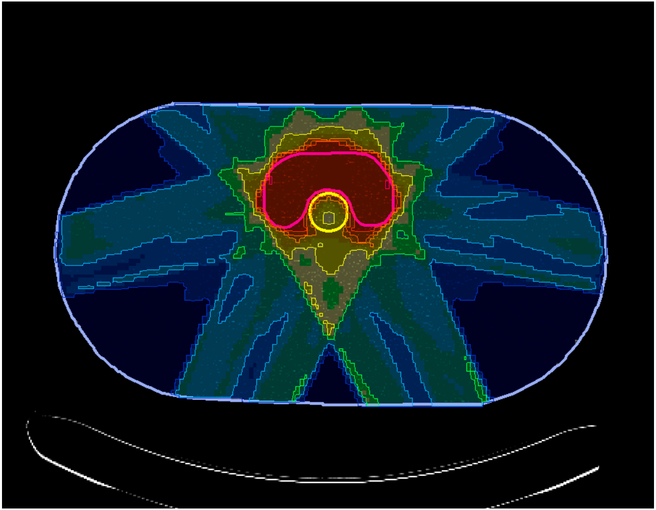
source: wikipedia

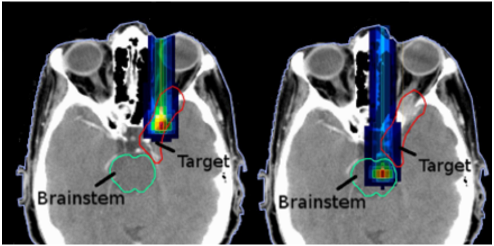


the parameter space

multi-beam plans

source: M. Bangert, DKFZ





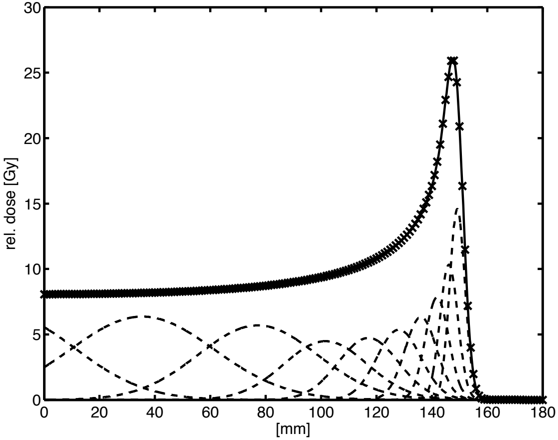
- setup errors of 5mm and less can drastically change the clinical outcome
- accounting for these errors is currently not clinical practice
- some prior work[2],[3], but problems of computational cost

[2]Unkelbach et al.: Reducing the sensitivity of IMPT treatment plans to setup errors and range uncertainties via probabilistic treatment planning. 2009 Med. Phys. 36: 149

[3]Sobotta et al.: Accelerated evaluation of the robustness of treatment plans against geometric uncertainties by Gaussian processes. 2012 Phys. Med. Biol. 57 (23): 8023

Propagating Gaussian uncertainty through nonlinearities

using integrals against Gaussian measures



- works on virtually any continuous function
- guaranteed numerical precision, fixed at design time
- low computational cost: just matrix-matrix multiplications

Error Bars on Radiation Dose

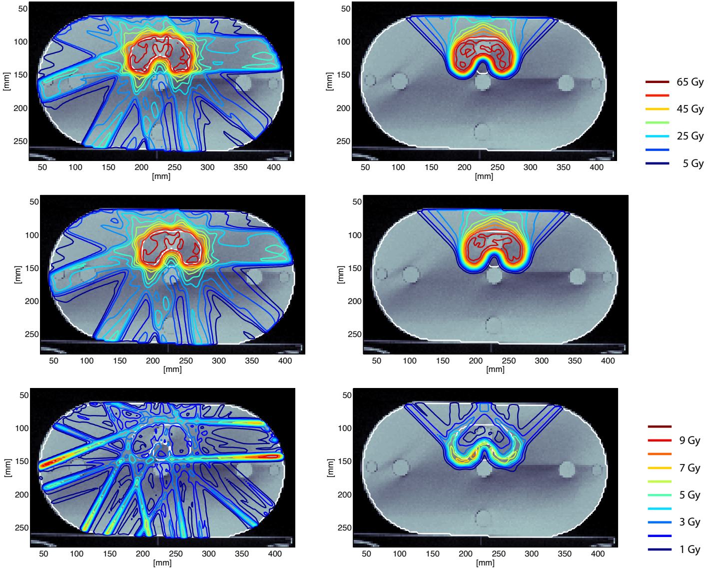
setup error 1mm ± 2mm, range error 3%

Bangert, Hennig, Oelfke, 2013

$\boldsymbol{d}$

$E[\boldsymbol{d}]$

σ_1



Gaussian algebra can be used to build
numerical methods for probabilistic computations.

Gaussians provide the linear algebra of inference

- products of Gaussians are Gaussians

$$\mathcal{N}(x;a,A)\mathcal{N}(x;b,B) = \mathcal{N}(x;c,C)\mathcal{N}(a;b,A+B)$$
$$C := (A^{-1}+B^{-1})^{-1} \qquad c := C(A^{-1}a + B^{-1}b)$$

- marginals of Gaussians are Gaussians

$$\int \mathcal{N}\left[\begin{pmatrix} x \\ y \end{pmatrix};\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix},\begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}\right] \mathrm{d}y = \mathcal{N}(x;\mu_x,\Sigma_{xx})$$

- (linear) conditionals of Gaussians are Gaussians

$$p(x \mid y) = \frac{p(x,y)}{p(y)} = \mathcal{N}\left(x;\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\right)$$

- linear projections of Gaussians are Gaussians

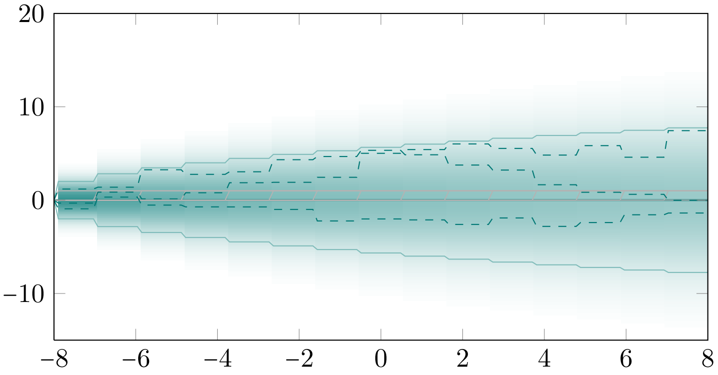
$$p(z) = \mathcal{N}(z;\mu,\Sigma) \quad \Rightarrow \quad p(Az) = \mathcal{N}(Az, A\mu, A\Sigma A^\top)$$

- analytical integrals allow moment matching “projection to Gaussians”

$$\int f(x)\mathcal{N}(x;\mu,\Sigma) = \text{known} \qquad \text{e.g. for } f(x) = x^p, \operatorname{erf}(x), \mathcal{N}(x), x^\top V x$$

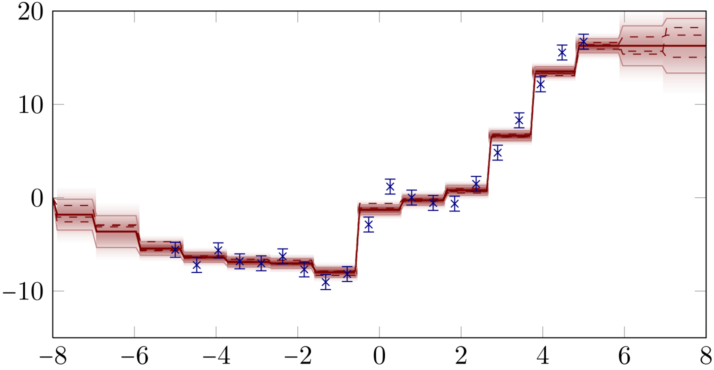
Generalised linear models learn nonlinear functions

$$f(x) = \phi(x)^\top w \qquad p(w) = \mathcal{N}(w; \mu, \Sigma)$$



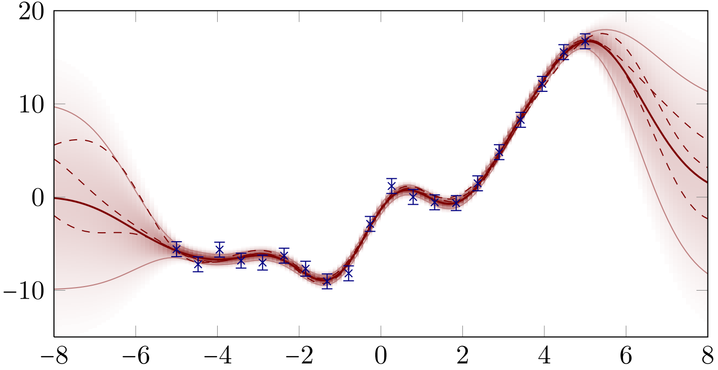
Generalised linear models learn nonlinear functions

$$f(x) = \phi(x)^\top w \qquad p(w) = \mathcal{N}(w; \mu, \Sigma)$$



infinite feature sets give nonparametric models

$$p(f) = \mathcal{GP}(f; \mu, k)$$



Gaussian processes are powerful, but not magic

powerful models

- kernels use infinitely many features
- kernels can be combined to form expressive models
- hyperparameters can be learned by hierarchical inference
- individual nonlinear effects can be separated from superpositions
- some kernels are universal

but no magic

- every model has parameters chosen a priori
- universal kernels can have logarithmic convergence rate

Gaussian processes are at heart of probabilistic numerics

Gaussians have great algebraic properties

- GPs are closed under linear projections, including
  - differentiation
  - integration
- GPs can be integrated against an expressive set of functions

They are the elementary tool of probabilistic numerics

- quadrature rules can be derived from GPs
- quasi-Newton optimization can be generalised using GPs
- GPs allow ODE solvers capable of probabilistic input
- moment matching allows numerical probabilistic computations

Numerics is about turning nonlinear problems into linear ones.
That's what Gaussian regression does.

Questions?

Bibliography

- T. O'Hagan
  Bayes-Hermite Quadrature
  J. Statistical Planning and Inference **29**, pp. 245–260
- C.E. Rasmussen & C.K.I. Williams
  Gaussian Processes for Machine Learning
  MIT Press, 2006
- T. Minka
  Deriving quadrature rules from Gaussian processes
  Tech. Report 2000
- M.A. Osborne, D. Duvenaud, R. Garnett, C.E. Rasmussen, S.J. Roberts, Z. Ghahramani
  Active Learning of Model Evidence Using Bayesian Quadrature
  Advances in NIPS, 2012
- P. Hennig & M. Kiefel
  Quasi-Newton Methods: a new direction
  ICML 2012 (short form), and JMLR **14** (2013), pp. 807–829
- P. Hennig
  Fast Probabilistic Optimization from Noisy Gradients
  ICML 2013
- J. Skilling
  Bayesian solution of ordinary differential equations
  Maximum Entropy and Bayesian Methods, 1991
- O. Chkrebtii, D.A. Campbell, M.A. Girolami, B. Calderhead
  Bayesian Uncertainty Quantification for Differential Equations
  `http://arxiv.org/abs/1306.2365`
- M. Bangert, P. Hennig, U. Oelfke
  Analytical probabilistic modeling for radiation therapy treatment planning
  Physics in Medicine and Biology, 2013, in press