Gaussian Processes in Numerical Optimization

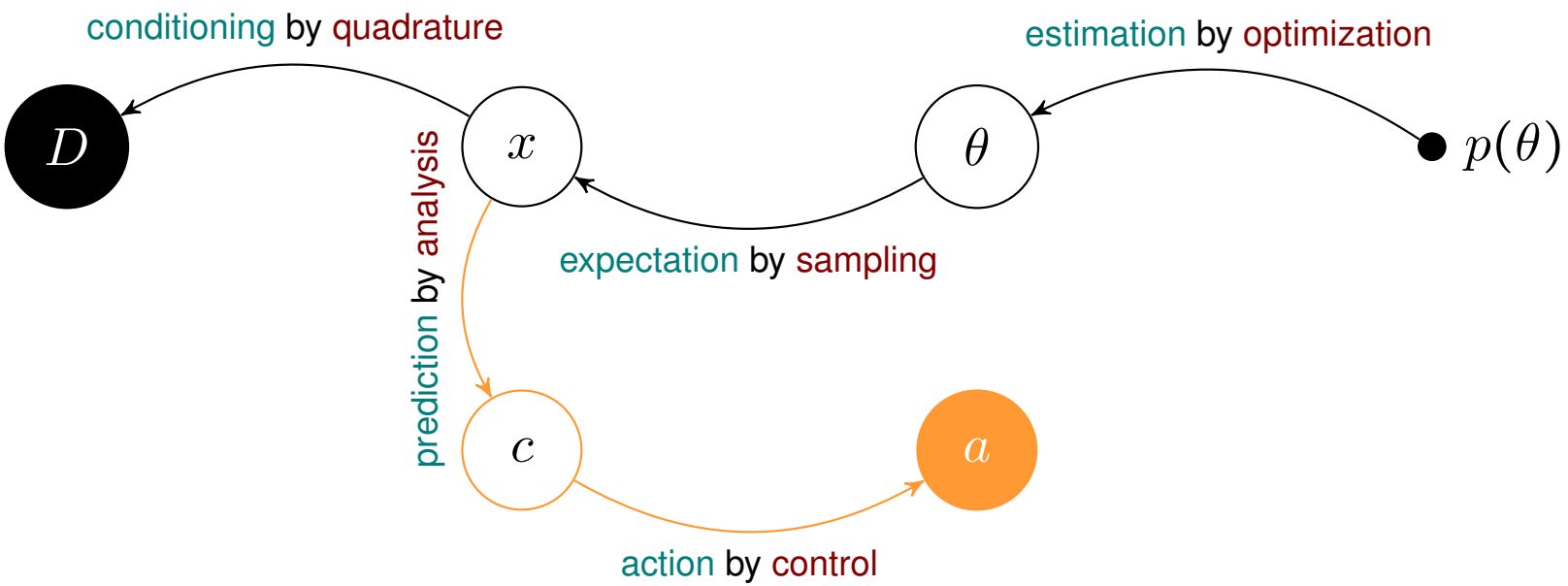
Philipp Hennig

GPSS, Sheffield
12 June 2013

Max Planck Institute for Intelligent Systems
Department for Empirical Inference
Research Group Elementary Intelligence

The machine learning pipeline

we need a coherent framework for hierarchical machine learning



- black boxes cause unaccounted errors
- making errors explicit may improve quality / lower cost
- joint language required: probability

Numerical algorithms are the elements of inference

inferring solutions of non-analytic problems

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

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

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

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

So what?

Numerical algorithms are learning machines!

- now, let's focus on optimization

Optimization

continuous, nonlinear, unconstrained

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

An old idea: Newton's method

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

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

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

$$B_{t+1} = B_t + uv^\top + vu^\top \qquad \text{with} \qquad u, v \in \mathbb{R}^{N\times M} \qquad \rightarrowtail \qquad \mathcal{O}(N^2 + M^3)$$

- want 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

$$B_{t+1} = \operatorname*{arg\,min}_{B\in\mathbb{R}^{N\times N}} \left\{ \lim_{\beta \rightarrowtail 0} \tfrac{1}{\beta} \|y_t - Bs_t\|_V^2 + \|B - B_t\|_{F,V}^2 \right\}$$
$$= B_t + \frac{(y_t - B_t s_t) s_t^\top V}{s_t^\top V s_t} \qquad \text{Broyden's (1965) method}$$

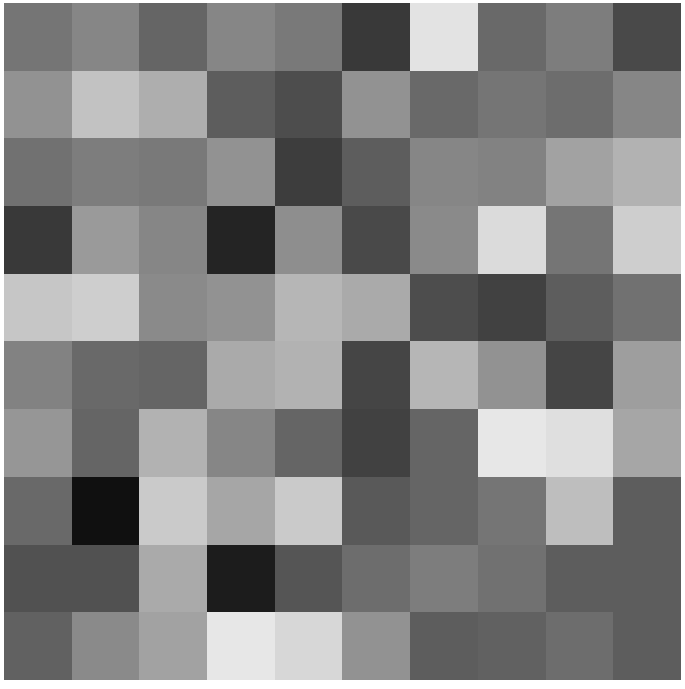
Quasi-Newton methods: the probabilist's view

Gaussian regression on matrices — P.H. & M. Kiefel, ICML 2012, JMLR 2013

$$
\begin{aligned}
B_{t+1} &= \lim_{\beta \to 0} \arg\max_B \{ -\tfrac{1}{\beta} \|y_t - B s_t\|_V^2 - \|B - B_t\|_{F,V}^2 \} \\
&= \lim_{\beta \to 0} \arg\max_B \{ \log \underbrace{p(y_t \,|\, B, s_t, \beta, V)}_{\text{likelihood}} + \log \underbrace{p(B \,|\, B_t, V)}_{\text{prior}} \} \\
&= \lim_{\beta \to 0} \arg\max_B p(y_t \,|\, B, s_t, \beta, V) p(B \,|\, B_t, V) \\
&= \lim_{\beta \to 0} \arg\max_B \mathcal{N}(y_t; B s_t, \beta V) \mathcal{N}(\overrightarrow{B}; \overrightarrow{B}_t, V \otimes V) \\
&= \arg\max_B \underbrace{\mathcal{N}\left[B; B_t + \frac{(y_t - B_t s_t) V s_t^\intercal}{s_t^\intercal V s_t}, V \otimes \left(V - \frac{V s s^\intercal V}{s^\intercal V s} \right) \right]}_{\text{posterior}}
\end{aligned}
$$

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

Outline



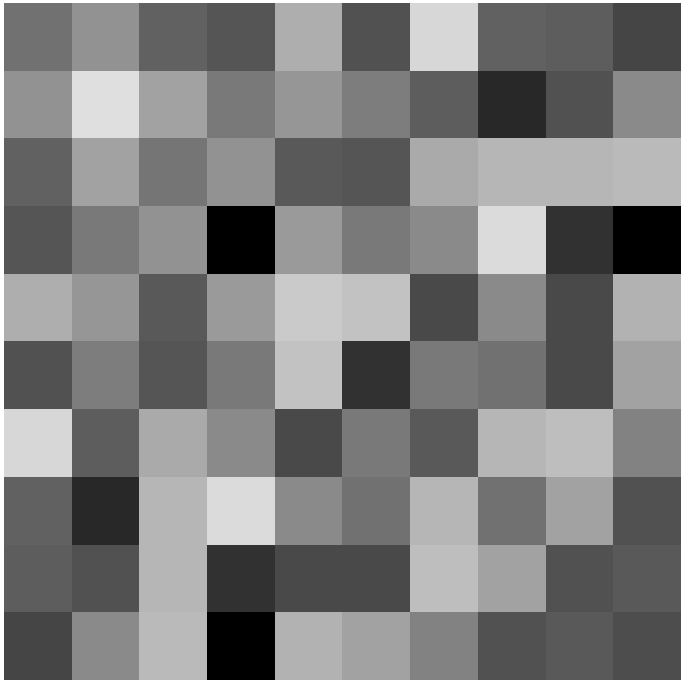
QN-methods perform regression!

Whenever you see ℓ_2 arguments, think of GPs!

- what prior assumptions allow low cost?

Encoding symmetry exactly is expensive

full probabilistic solution has quartic cost — P.H., ICML 2013



the naïve idea:

- "observe symmetry" using $\Delta \overrightarrow{B} = \frac{1}{2}(\overrightarrow{B} - \overrightarrow{B^{\mathsf{T}}})$

$$p(0 \,|\, B, \Delta) = \delta(0 - \Delta \overrightarrow{B})$$

- however, $\operatorname{rk} \Delta = \frac{1}{2}N(N-1)$: costs at least $\mathcal{O}(N^4)$

Instead, make two independent observations

observe B acting on two different spaces — P.H., ICML 2013

instead: Powell's (1970) Symmetric Broyden (PSB) update

- before: $y = Bs$
- now, additionally: "dual observation"

$$p(y_i^\top \mid B, s_i^\top, V_{i-1}) = \delta(y_i^\top - s_i^\top B) = \lim_{\beta \to 0} \mathcal{N}(y_i^\top; s_i^\top B, \beta \otimes V_i)$$

- gives posterior mean, covariance

$$B_i = B_{i-1} + \frac{(y_i - B_{i-1}s_i)s_i^\top V_{i-1} + V_{i-1}s_i(y_i - B_{i-1}s_i)^\top}{s_i^\top V_{i-1} s_i} - \frac{V_{i-1}s_i(s_i^\top(y_i - B_{i-1}s_i))s_i^\top V_{i-1}}{(s_i^\top V_{i-1}s_i)^2}$$

$$\Sigma_i = \left(V_{i-1} - \frac{V_{i-1}s_i s_i^\top V_{i-1}}{s_i^\top V_{i-1} s_i}\right) \otimes \left(V_{i-1} - \frac{V_{i-1}s_i s_i^\top V_{i-1}}{s_i^\top V_{i-1} s_i}\right)$$

Is this allowed?

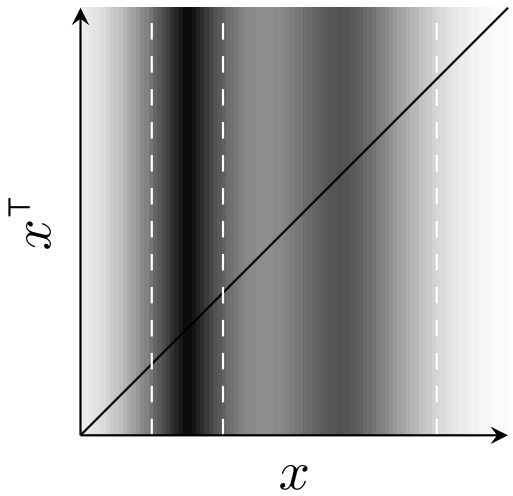
algebraic assumptions used to shape computational cost — P.H., ICML 2013

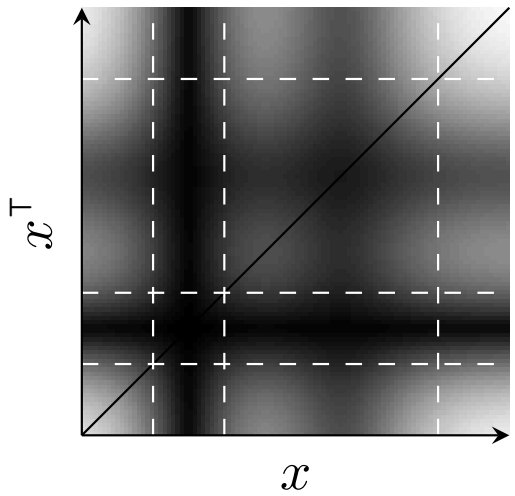
$$p(B) = \mathcal{N}(B; B_0, V \otimes V)$$
$$\operatorname{cov}(B_{ij}, B_{k\ell}) = (V \otimes V)_{(ij),(k\ell)} = V_{ik} V_{j\ell}$$
$$= \operatorname{cov}(\partial_i(\nabla f), \partial_k(\nabla f)) \cdot \operatorname{cov}(\partial_j(f \nabla^\top), \partial_\ell(f \nabla^\top))$$

- each $B_{i:}$ is a map $\mathbb{R}^{N \times 1} \to \mathbb{R}$: $\partial_i(\nabla f)$
- each $B_{:j}$ is a map $\mathbb{R}^{1 \times N} \to \mathbb{R}$: $\partial_j(f \nabla^\top)$
- dual vector spaces are isomorphic, thus $(\nabla f)\nabla^\top = \nabla(f \nabla^\top)$
- but we won't tell the method! Treat the two maps as independent.
- replace $f(x) : \mathbb{R}^N \to \mathbb{R}$ with $f(x^\top, x) : \mathbb{R}^{N \times 1} \times \mathbb{R}^{1 \times N} \to \mathbb{R}$

Forgetting your good education to get ahead

QN methods deliberately ignore algebraic knowledge to lower computational cost





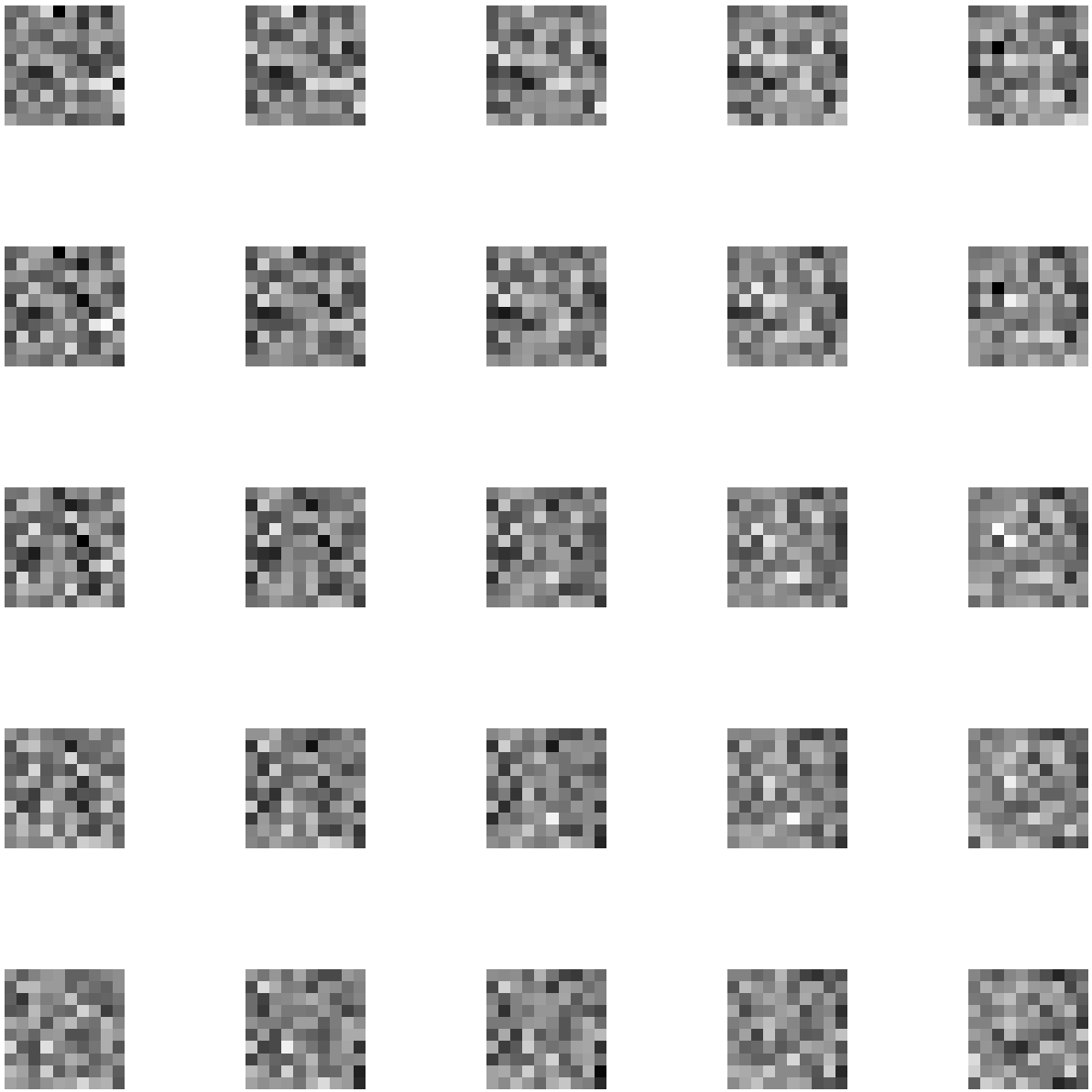
To achieve low computational cost, numerical methods may have to deliberately ignore prior knowledge!

Whenever you see Kronecker products, someone is hiding an independence assumption!

- where are the GPs already?

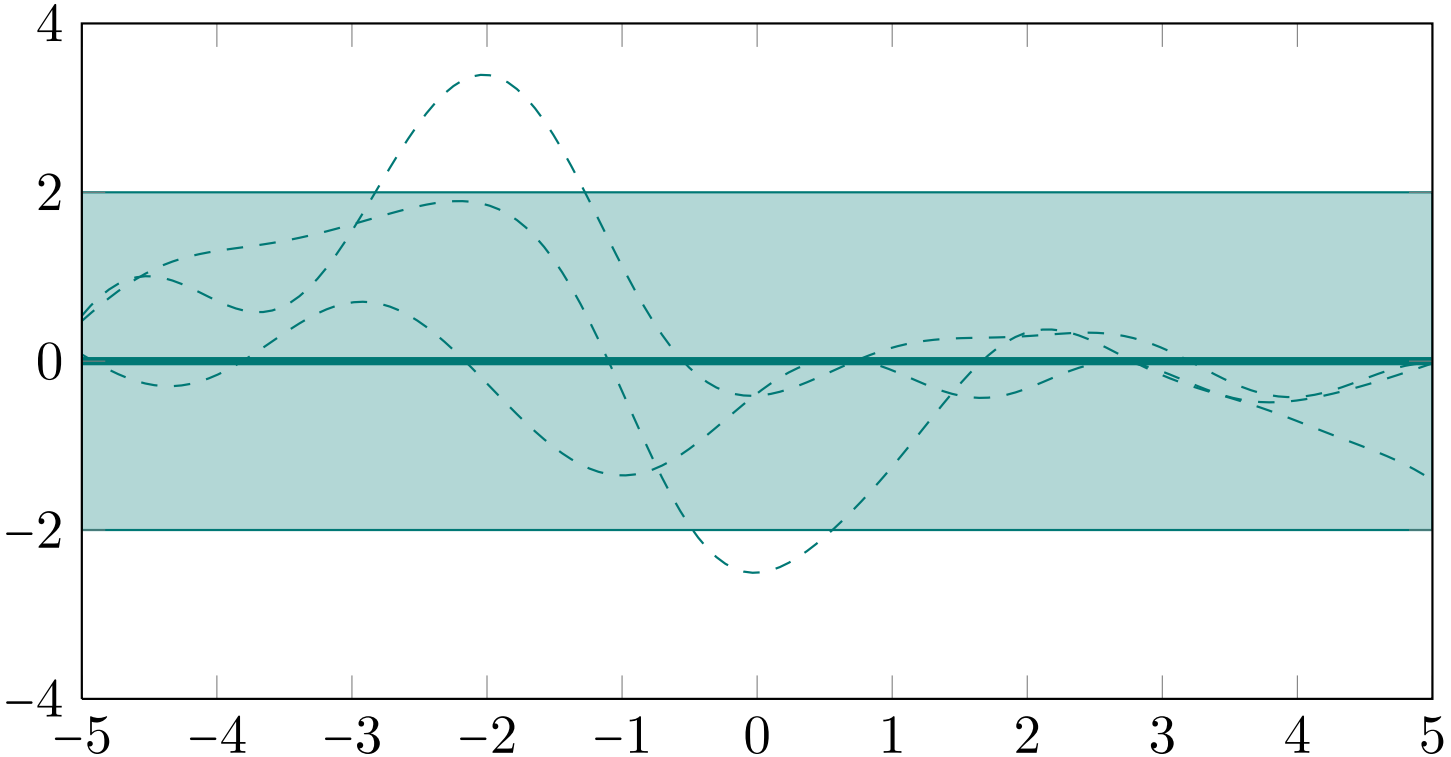
Hessians are not constant

can we have a prior over regular Hessians?



The unavoidable slide

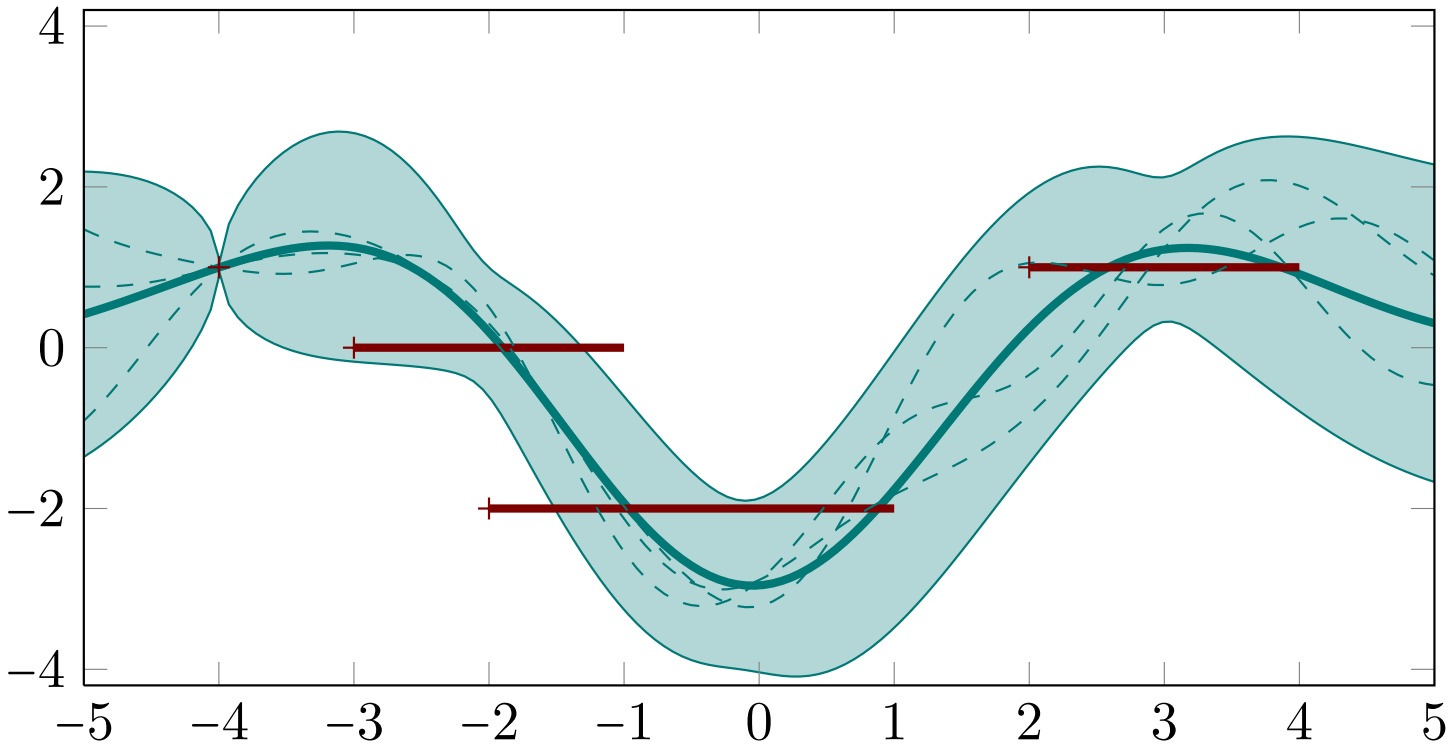
nonparametric generalisation of Gaussians to space of real-valued functions



$$p(f) = \mathcal{GP}(f; \mu, k)$$
$$p(Af) = \mathcal{GP}(Af; A\mu, AkA^\top)$$

The Gaussian family is closed under linear projections

thus provides the functionality for quasi-Newton inference



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

$$p(Af) = \mathcal{GP}(Af; A\mu, AkA^\top)$$

$$p\left(\int_a^b f(x)dx\right) = \mathcal{GP}\left(\int_a^b f(x)dx; \int_a^b \mu(x)dx, \iint_a^b k(x,x')dxdx'\right)$$

Nonparametric quasi-Newton methods

Inferring a changing Hessian

P.H. & M. 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

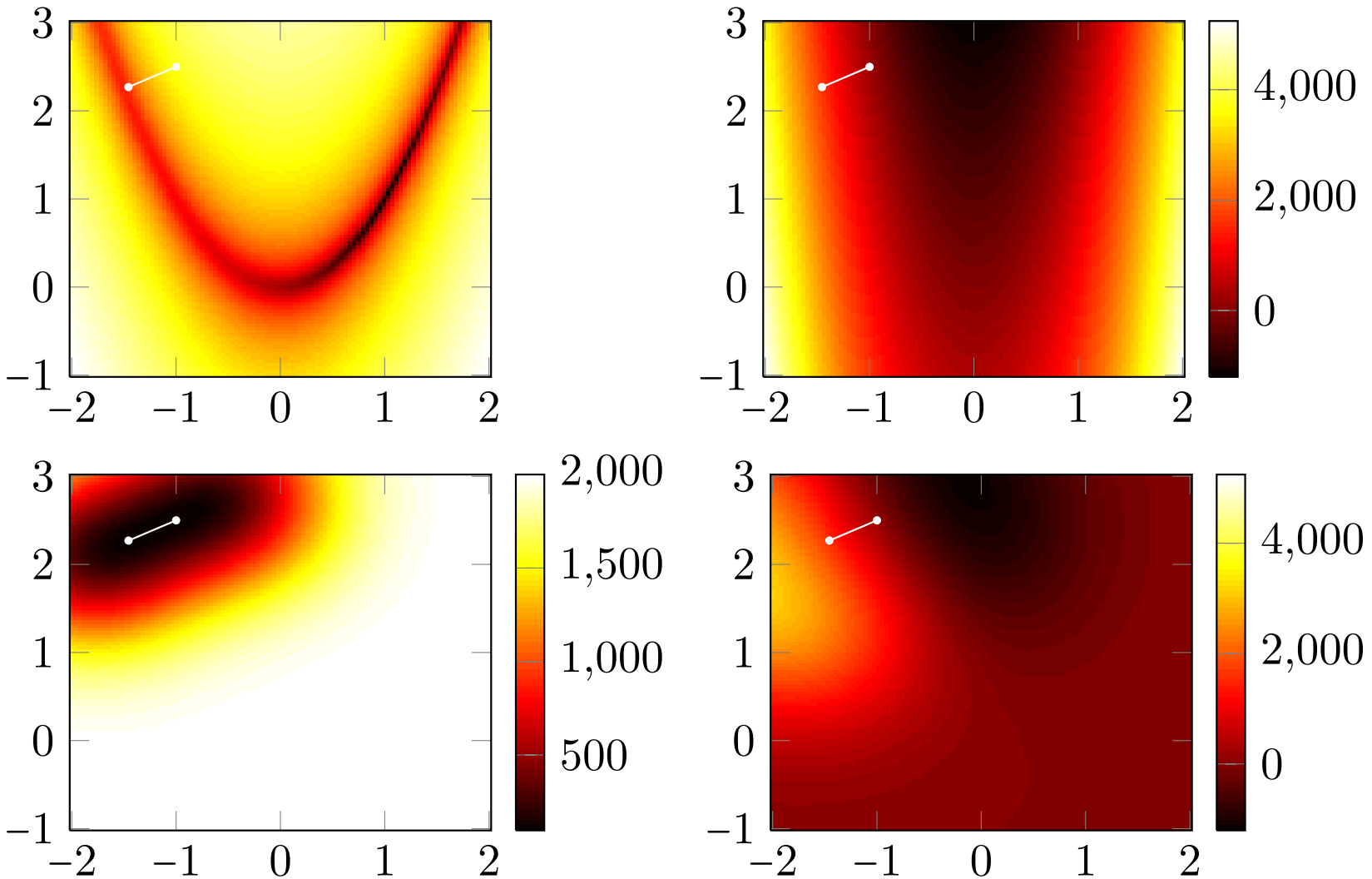
$$p(y_i(x^\top) \,|\, B, s_i) = \lim_{\beta \to 0} \mathcal{N}\left(y_i; \sum_m s_{im} \int_0^1 B(x^\top, x(t))\, \mathrm{d}t, k(x^\top, x'^\top) \otimes \beta\right)$$

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

- 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): $\mathcal{O}(N)$

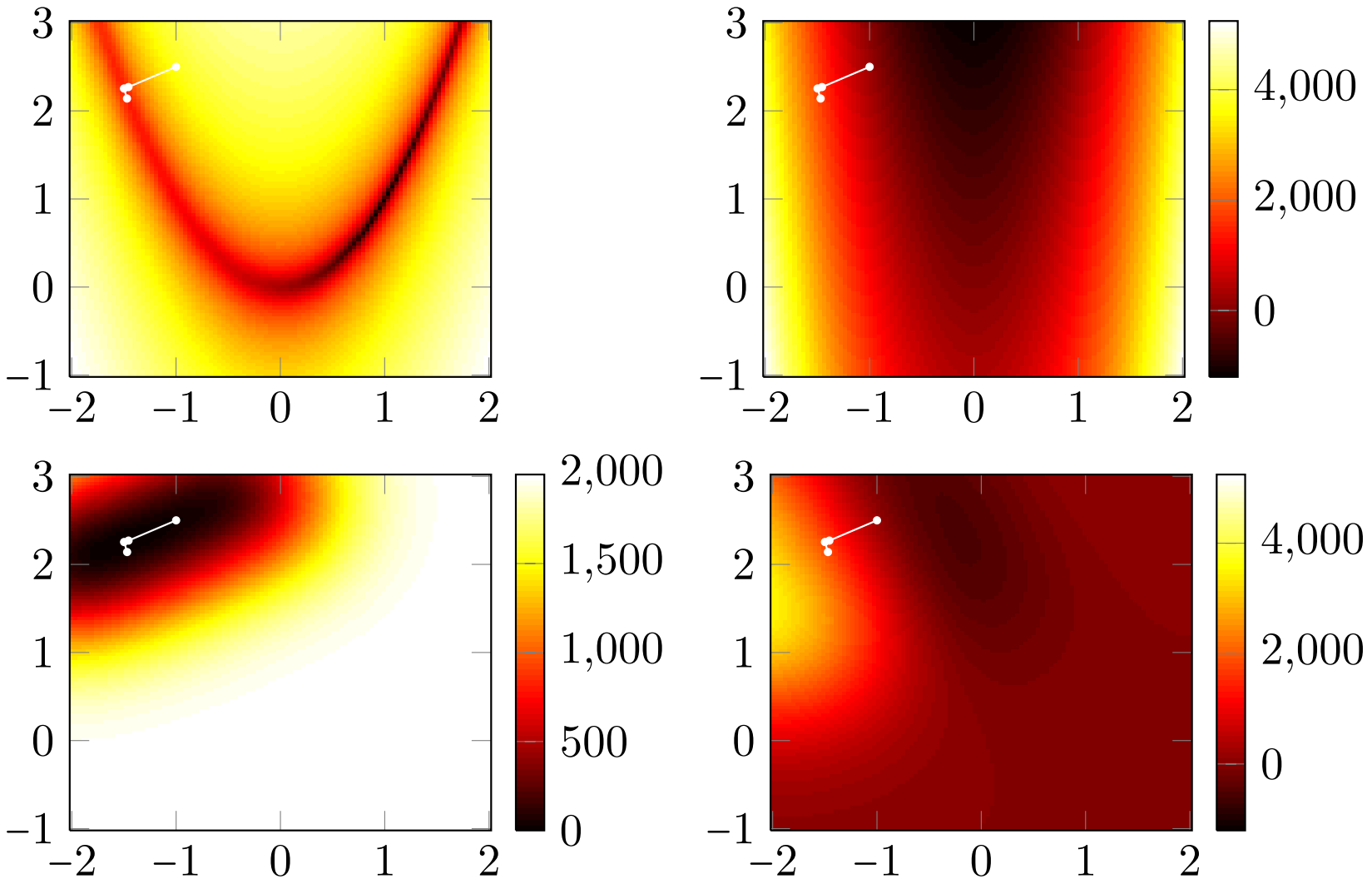
A consistent model of the Hessian function

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



A consistent model of the Hessian function

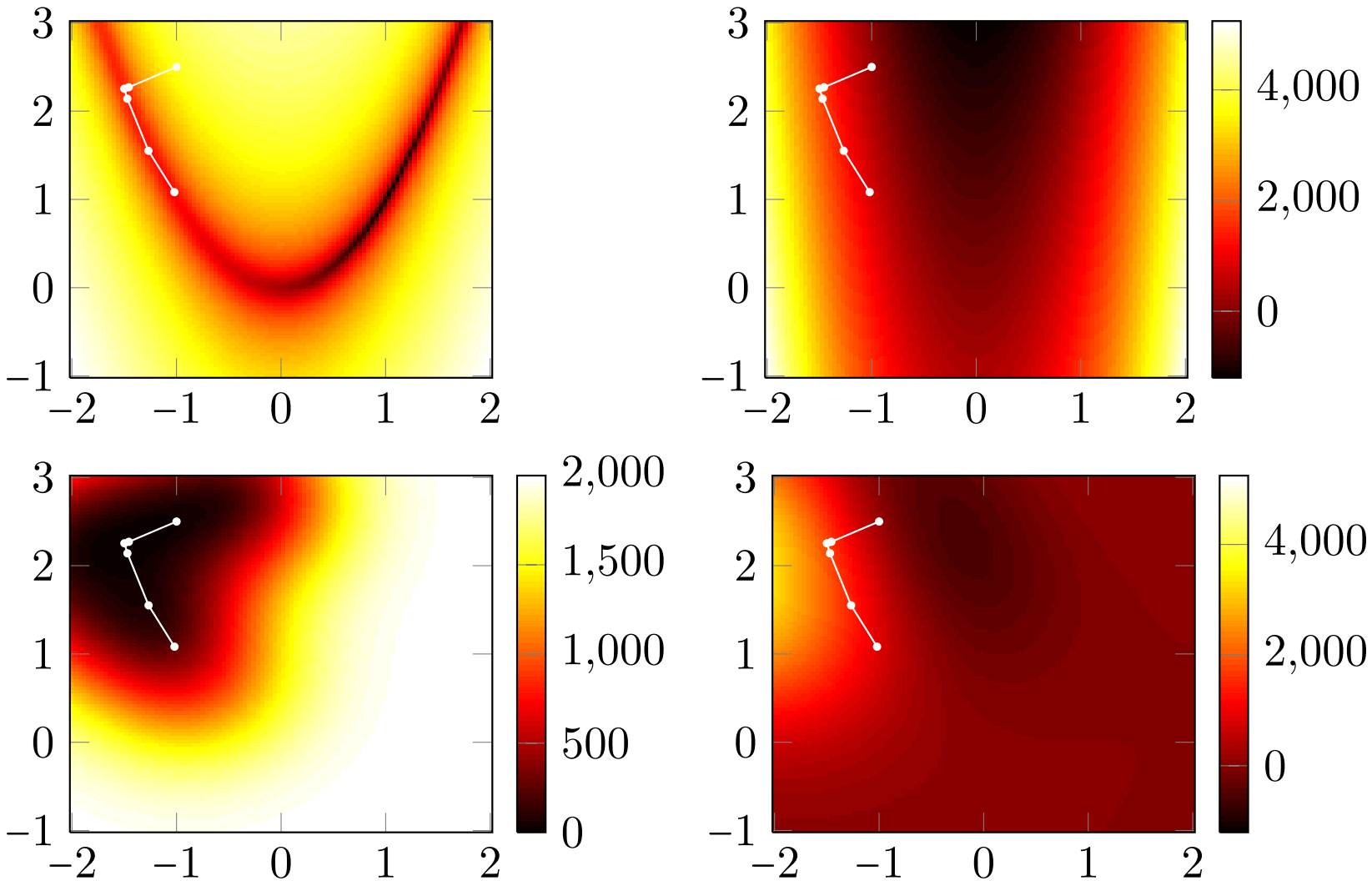
nonparametric inference on elements of the Hessian — P.H. & M. Kiefel, ICML 2012, JMLR 2013



A consistent model of the Hessian function

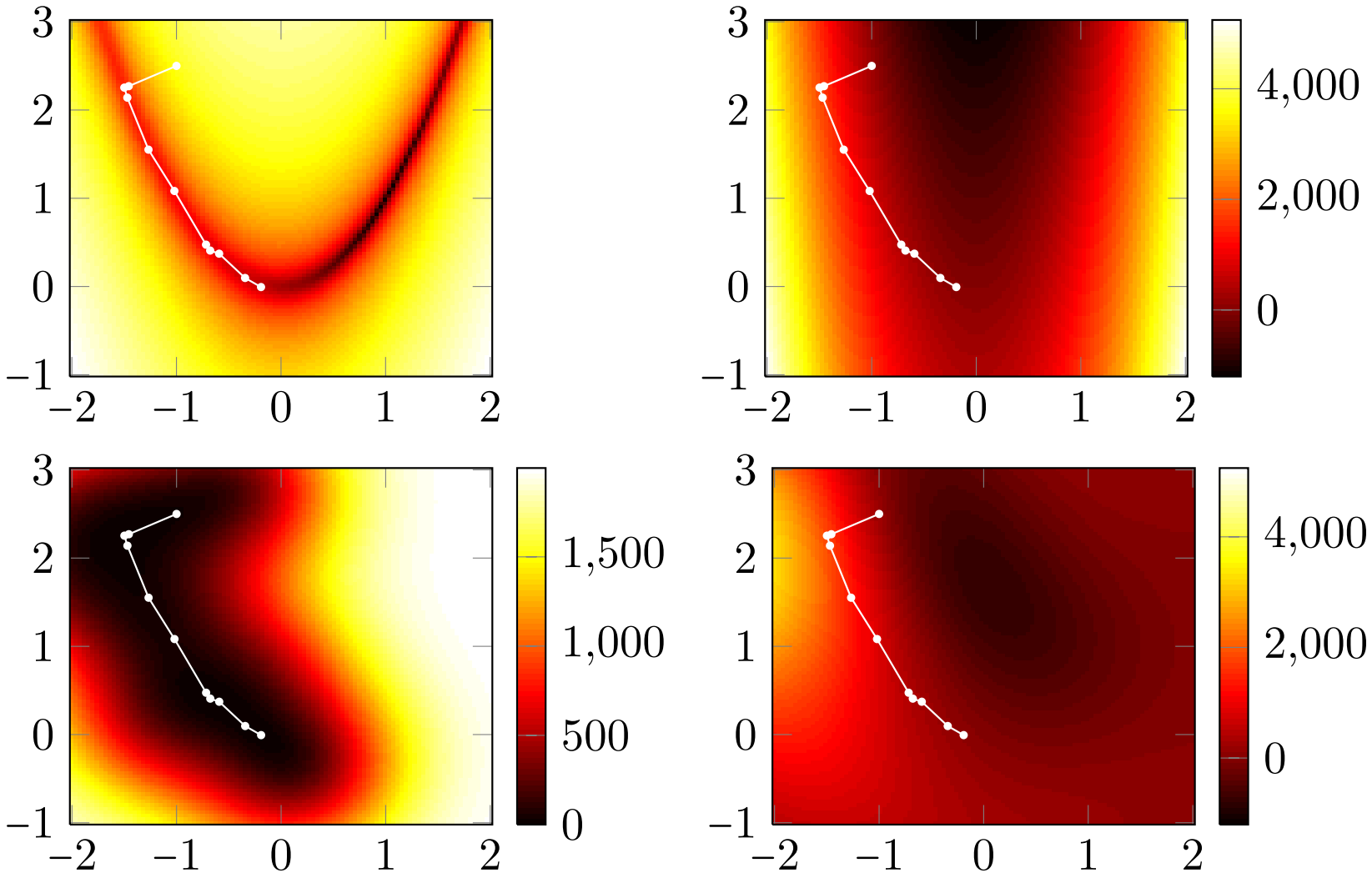
nonparametric inference on elements of the Hessian

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



A consistent model of the Hessian function

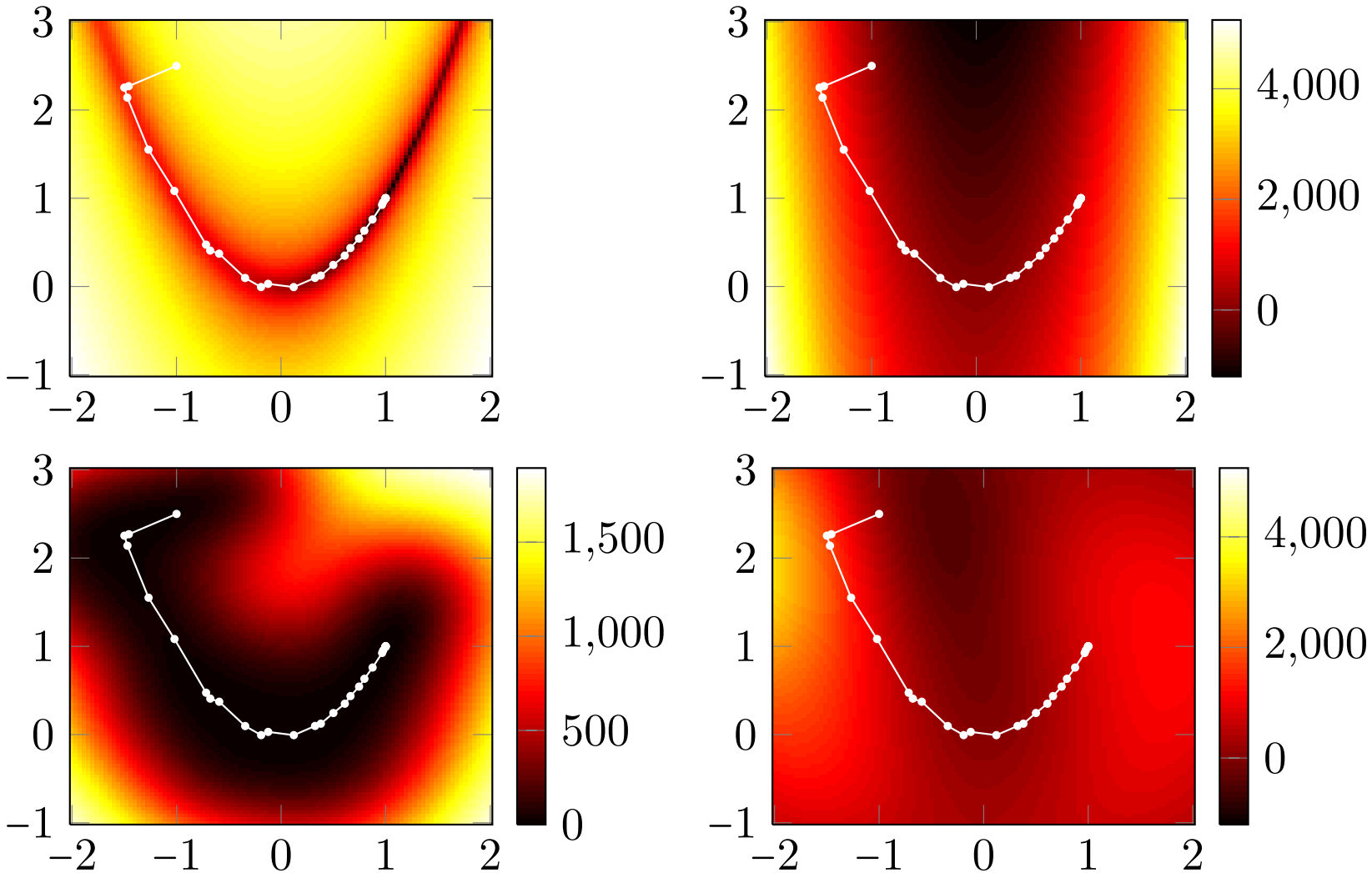
nonparametric inference on elements of the Hessian — P.H. & M. Kiefel, ICML 2012, JMLR 2013



A consistent model of the Hessian function

nonparametric inference on elements of the Hessian

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

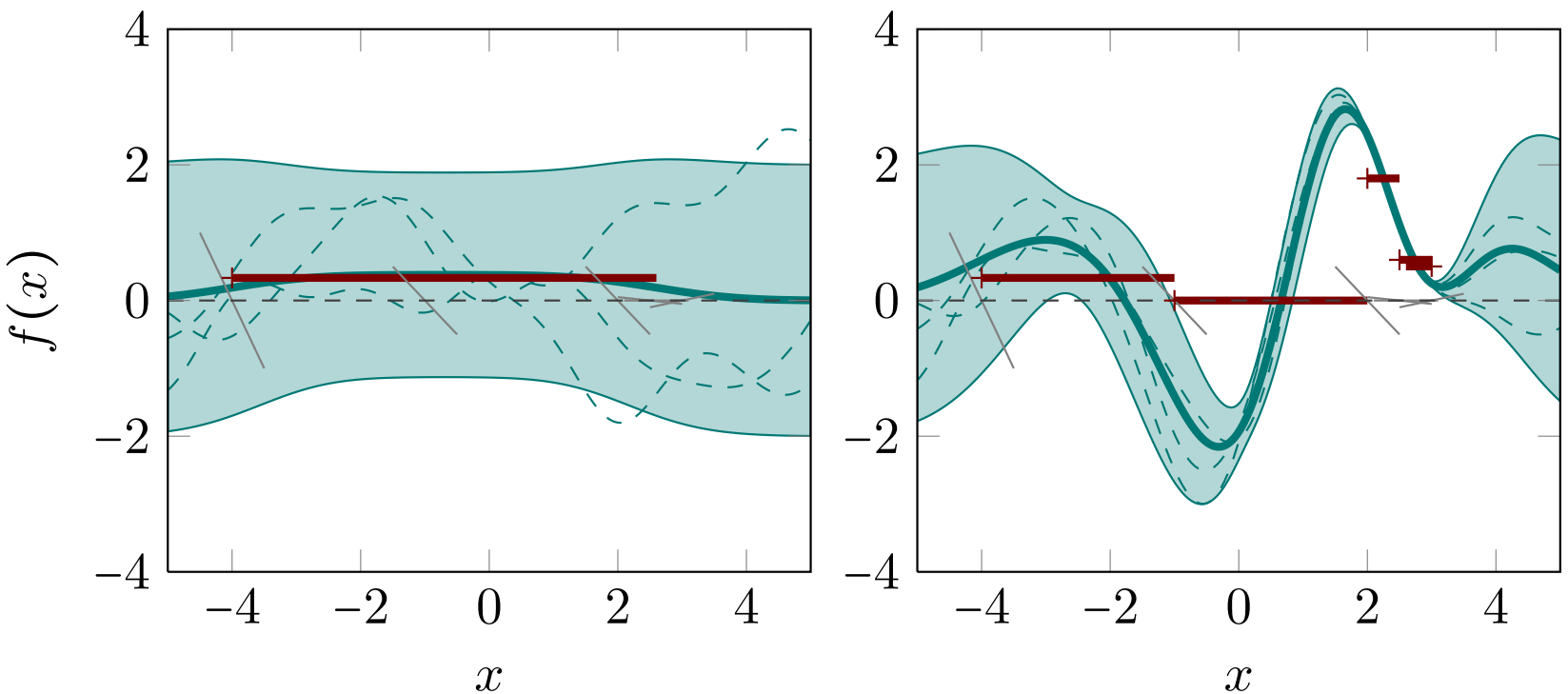


Advantages of nonparametric modelling

analytic use of all observations

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

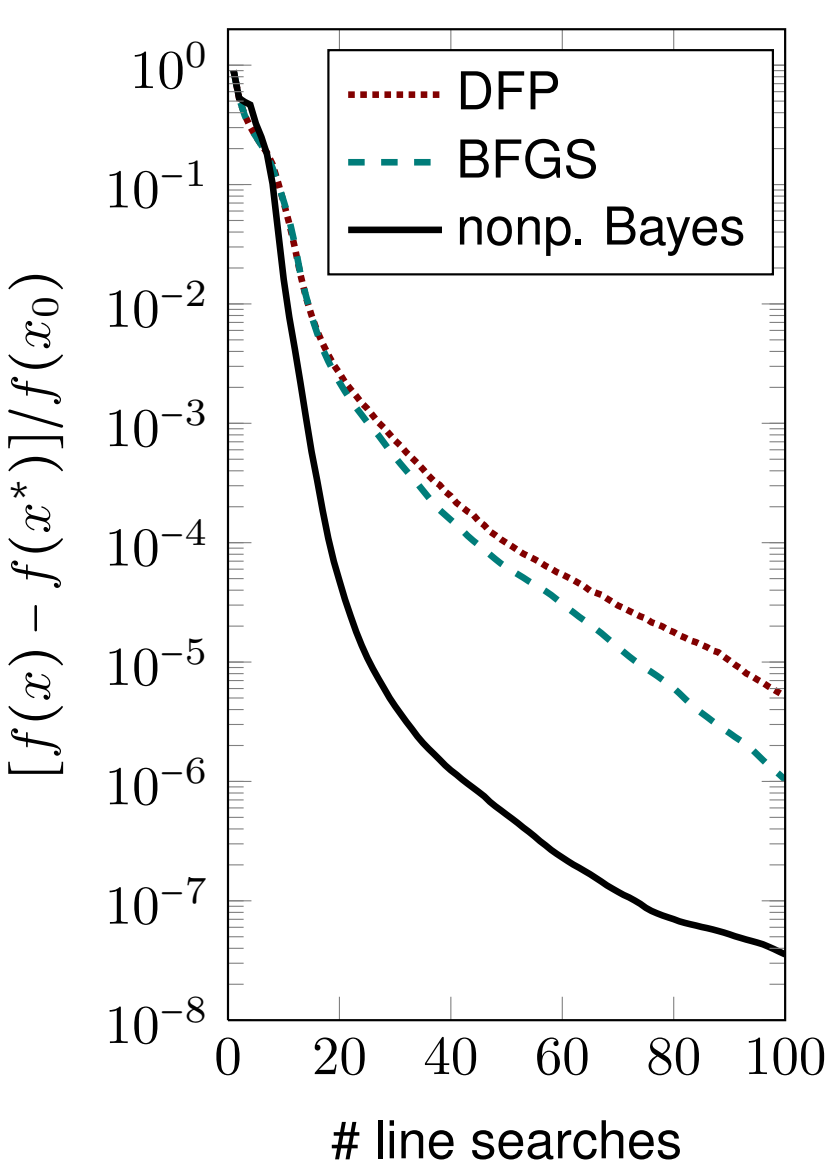
- choice of kernel allows more flexible model
- exact treatment of gradient observations
- can use all observations from line searches

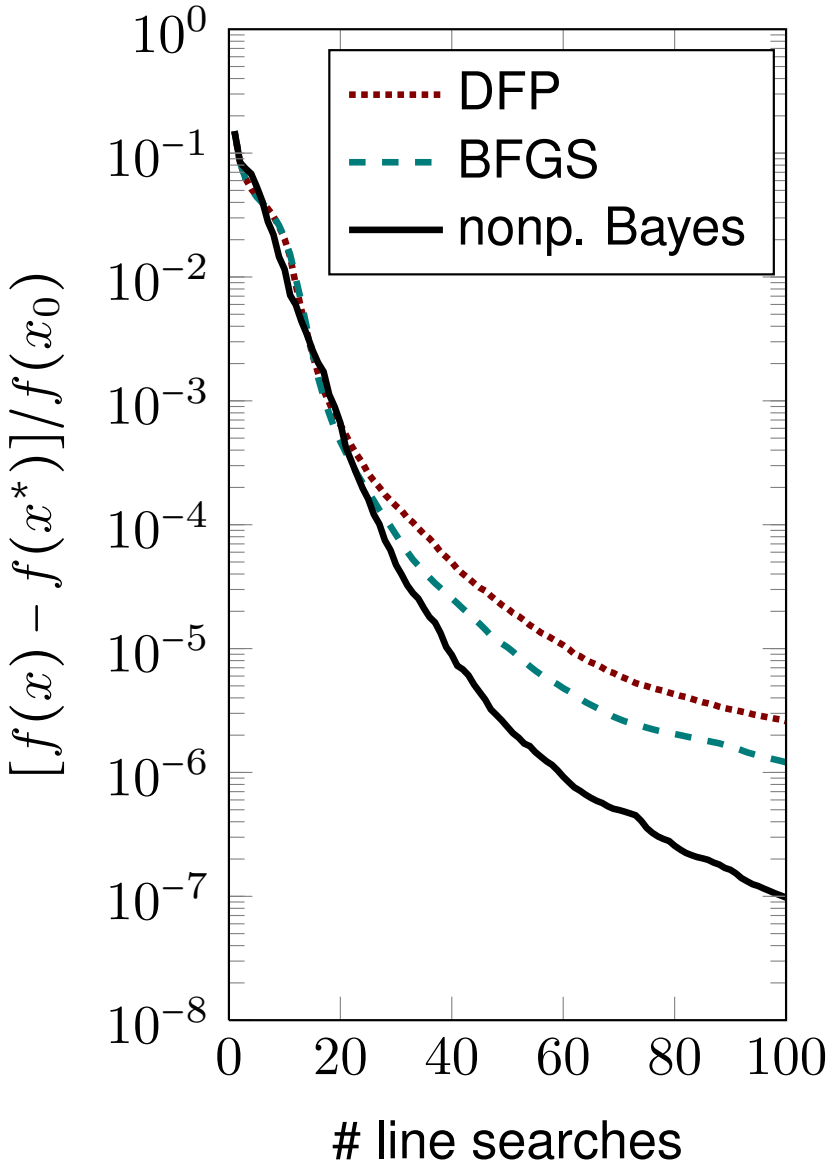


Some empirical results

optimizing a 200-dimensional function

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





Don’t let them tell you GPs are expensive

quasi-Newton methods can be generalised to GP models.

GPs are cubic in the number of inputs, not the outputs!

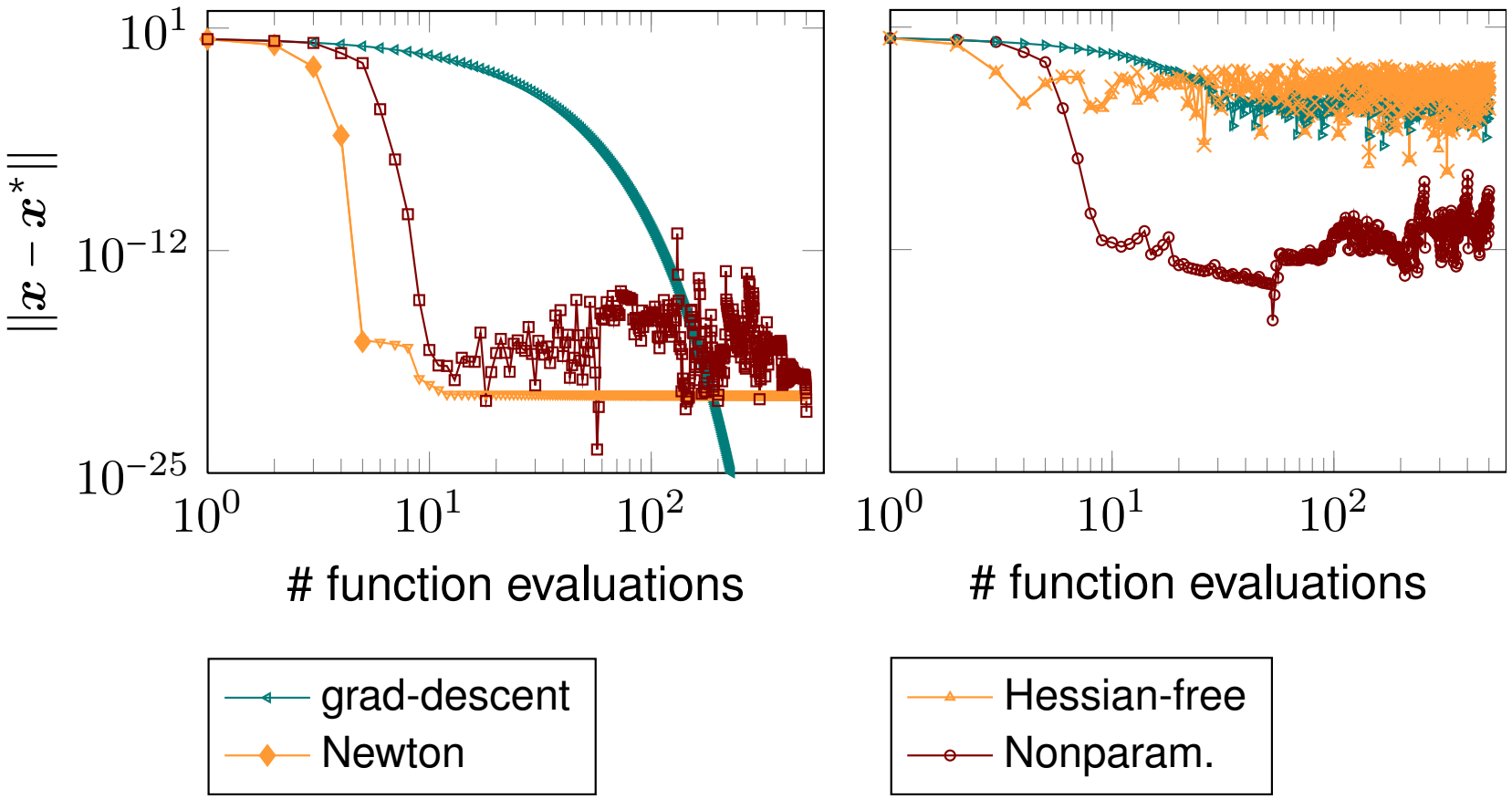
- how does this all help machine learning?

Example I: noise

optimization of noisy objectives

P.H., ICML 2013

- noisy objectives a frequent in learning (mini-batch training)
- stochastic gradient descent widely used
- noisy quasi-Newton methods improve conceptually



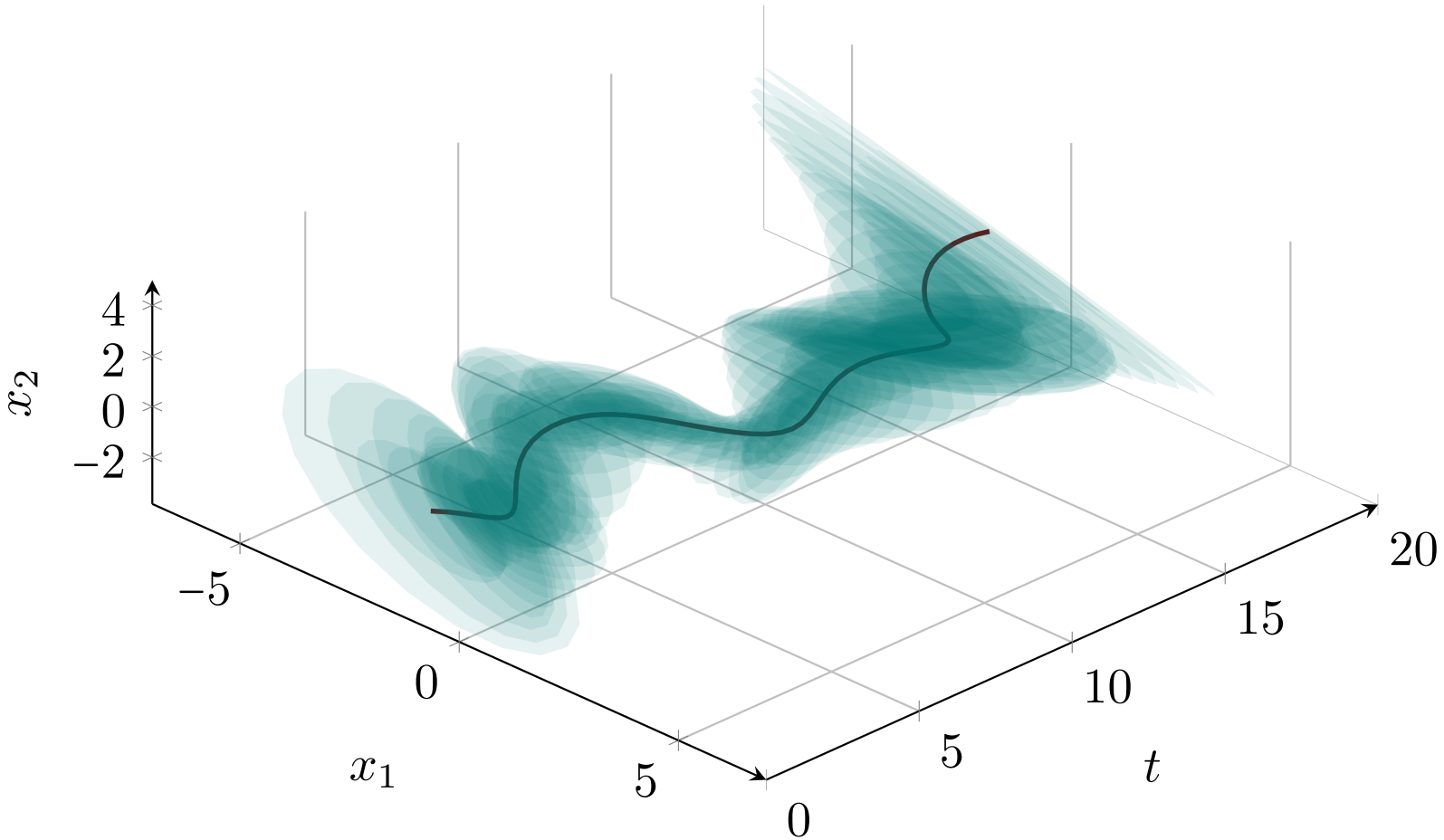
Example II: dynamics

optimization of dynamically changing objectives

P.H., in preparation

further generalisations

- dynamic settings ⇾ filtering



Conclusion

- the numerical methods used in ML are learners themselves
  don't treat them as black boxes
- BFGS is a Gaussian regressor
  think $\|\cdot\|_2 \leftrightarrow \mathcal{N}(x)$ and $\|\cdot\|_F \leftrightarrow \mathcal{N}(A)$ and $\|\cdot\|_{HS} \leftrightarrow \mathcal{GP}(f)$
- to be fast, you may need to ignore even pertinent knowledge
  watch out for $\otimes$ in covariances
- GP numerical optimization need not be expensive
  design your model carefully
- machine learning needs probabilistic numerical methods
- major theoretical, practical questions still open
  Want to stay informed? Befriend Maren!

Open issue: convergence

Are quasi-Newton methods locally linearly convergent?

Definition (local linear convergence)

QN update is locally linearly convergent at the minimum x_* if $\exists \epsilon, \delta > 0$, s.t.

$$(\|x_0 - x_*\| < \epsilon \wedge \|B_0 - B(x_*)\| < \delta) \quad \Rightarrow \quad (\|x_k - x_*\| = \mathcal{O}(e^{-k}))$$

nonparametric QN methods does not, in general, obey this, just like nonlinear regression does not collapse linearly fast.

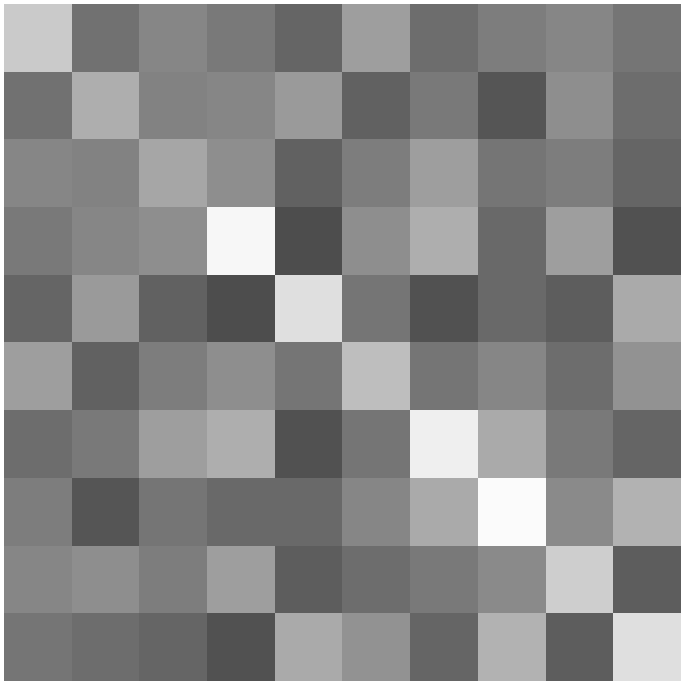
Theorem (Broyden, Dennis, Moré, 1973)

QN update is locally linearly convergent if

$$\|B_i - B(x_*)\| \leq [1 + \alpha\sigma(x_{i-1}, x_i)]\|B_{i-1} - B(x_*)\| + \beta\sigma(x_{i-1}, x_i)$$
$$\sigma(x_{i-1}, x_i) := \max\{\|x_i - x_*\|, \|x_{i-1} - x_*\|\}$$

positive definiteness?

as similar story to symmetry



the naïve idea:

- prior exclusively over positive definite matrices: Wishart

$$B = uu^\top \qquad \text{with} \qquad u_i \sim \mathcal{N}(0, V)$$

- however, inference not analytically tractable!

positive definiteness?

as similar story to symmetry

instead: BFGS / DFP:

$$p(B) = \mathcal{N}(B; B_0, V \otimes V) \qquad \text{and set} \qquad V = B$$

$$\propto |B|^{-N^2/2} \cdot \exp\left[-\frac{1}{2}\left(N - 2\,\mathrm{tr}(B_0 B^{-1}) + \mathrm{tr}(B_0 B^{-1} B_0 B^{-1})\right)\right]$$

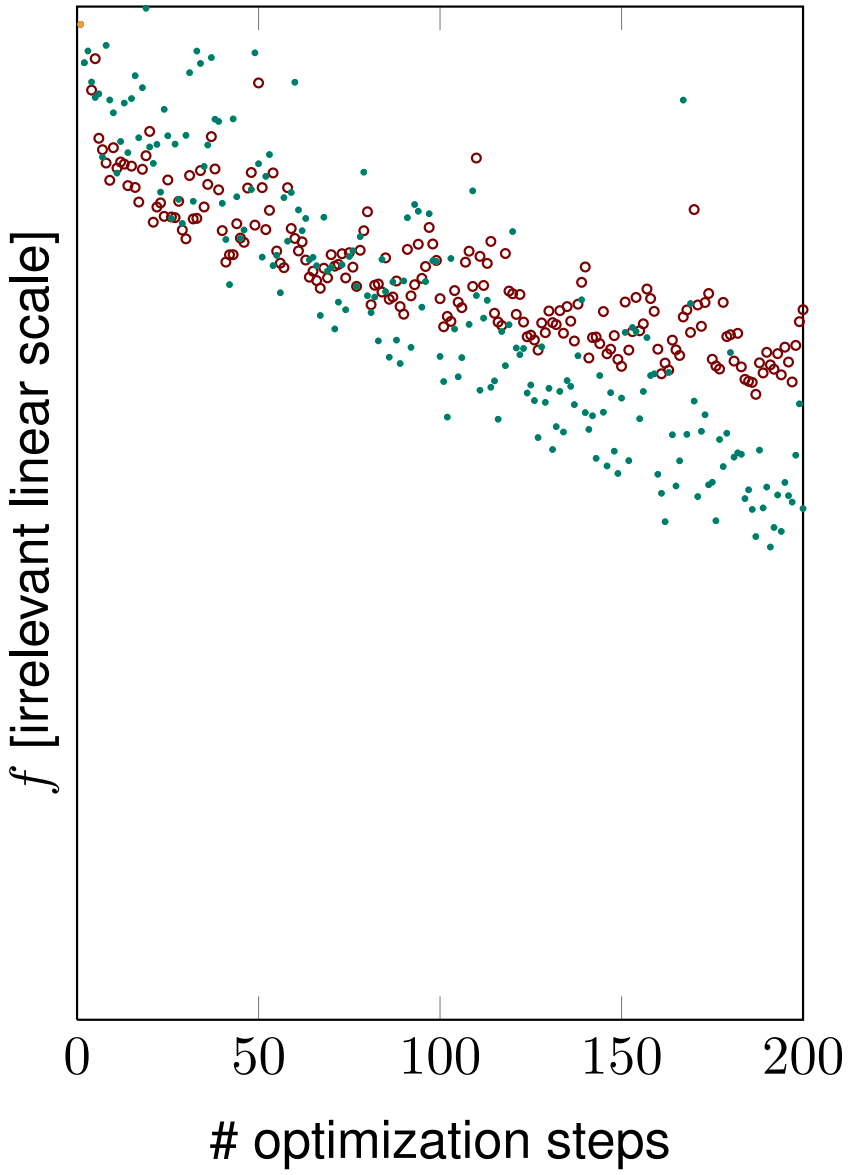
- posterior mean is positive definite if $y^\top s > 0$
- but prior puts nonzero mass on indefinite matrices like

$$B = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad \Rightarrow \qquad p(B) \propto \exp(-2)$$

BFGS and DFP have other advantages (scale-freedom, etc.).
But they do not explicitly encode positive definiteness.

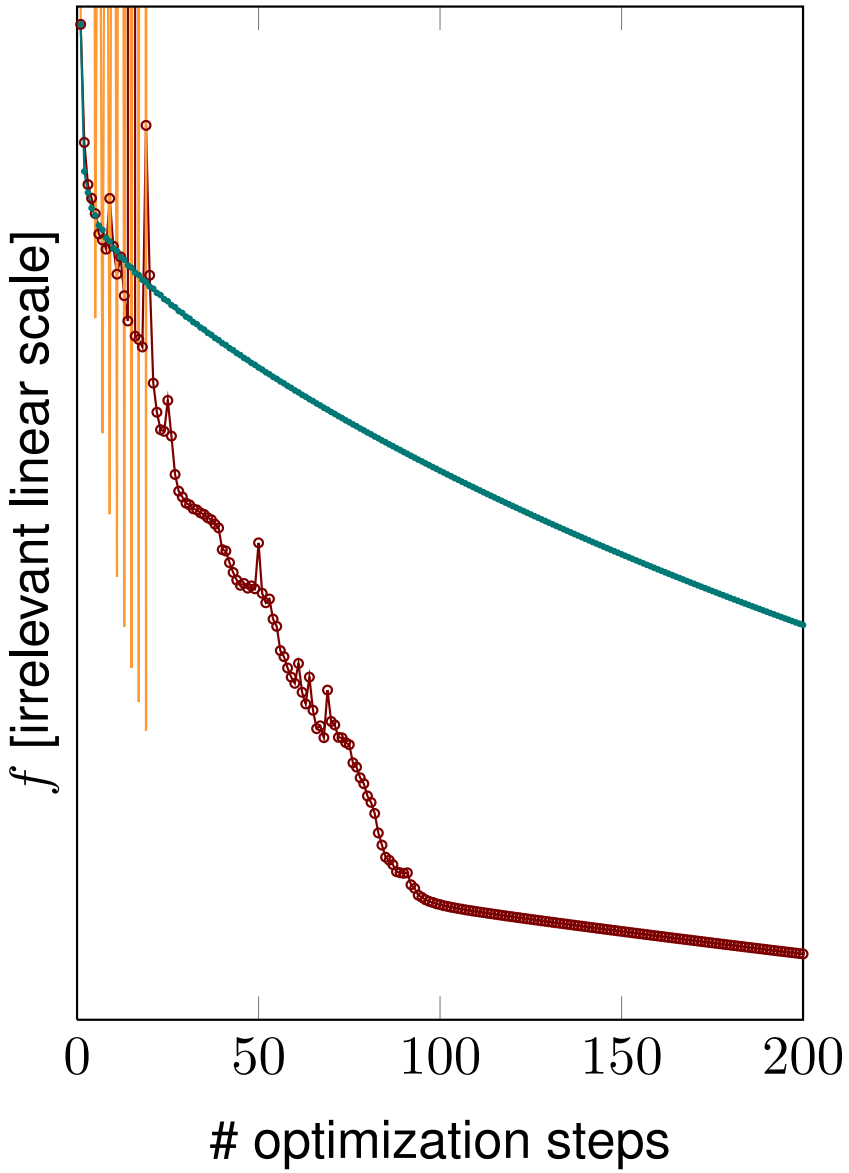
Performance on Neural Networks

of optimization steps, 10^3 samples



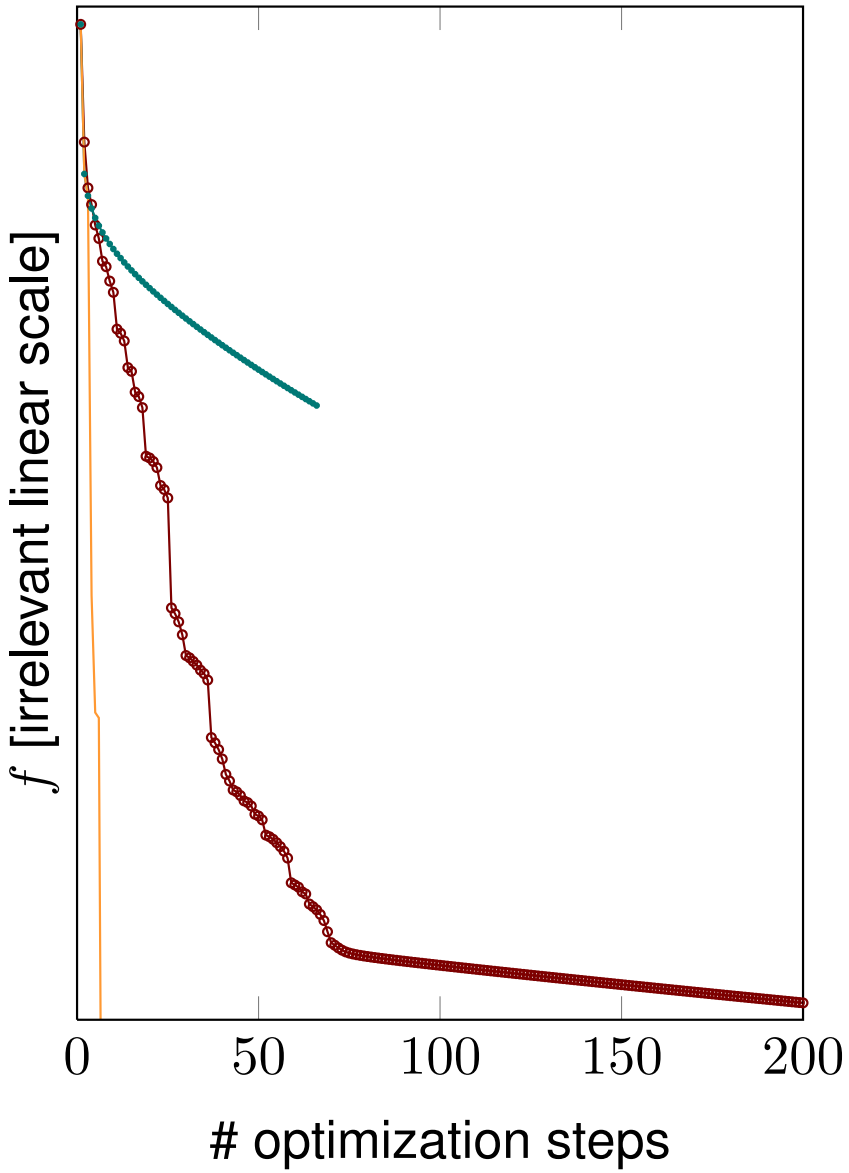
Performance on Neural Networks

of optimization steps, $3 \cdot 10^4$ samples



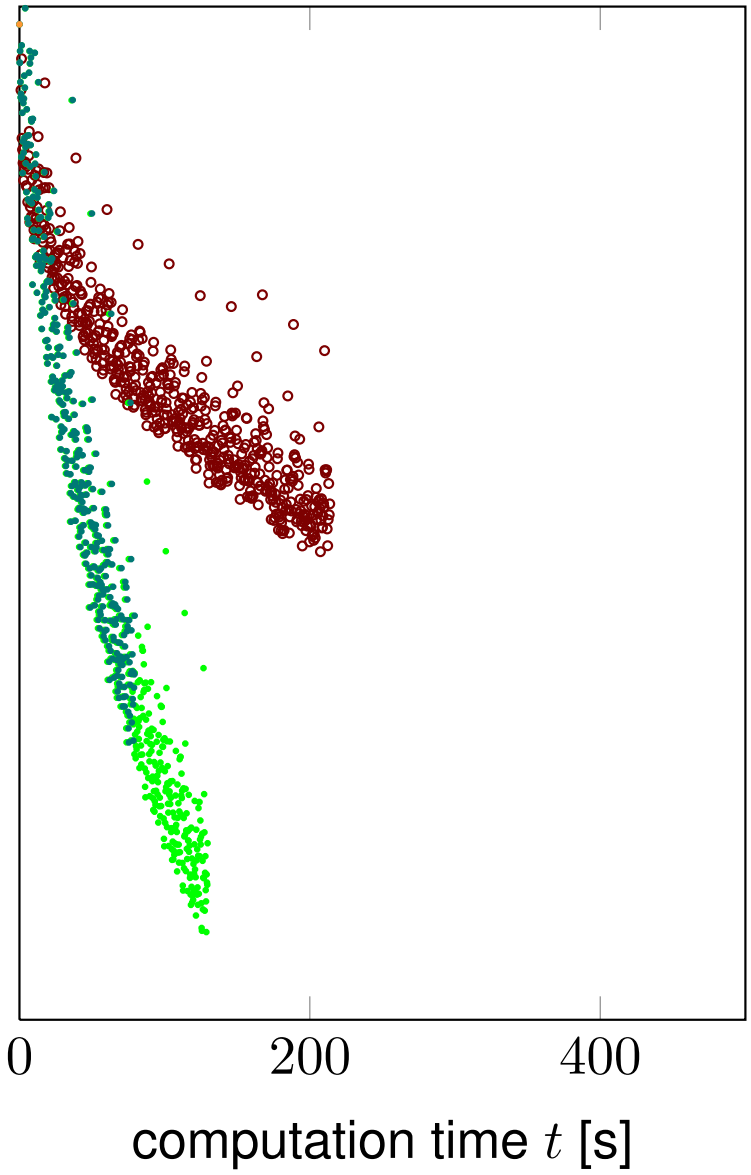
Performance on Neural Networks

of optimization steps, $6 \cdot 10^4$ samples



Performance on Neural Networks

wallclock time, 10^3 samples



Performance on Neural Networks

wallclock time, $3 \cdot 10^4$ samples

Performance on Neural Networks

wallclock time, $6 \cdot 10^4$ samples