MATH6221, Topic in Numerical Analysis
Computational Inverse Problems

Bangti Jin (b.jin@cuhk.edu.hk)

Department of Mathematics, Chinese University of Hong Kong

October 9, 2024

Outline

1 Fundamentals of Bayesian inversion

2 Markov Chain Monte Carlo

3 Approximate inference

finite-dimensional inverse problem

$$F(f) = g,$$

- f: the unknown signal/image, linear / nonlinear
- g: the noisy data (g^δ)
- $F : \mathbb{R}^m \mapsto \mathbb{R}^n$: forward map, ill-conditioned

Motivation

- variational regularization $\Leftrightarrow$ optimization problem, e.g.,

$$\|F(f) - g\|^2 + \alpha \|f\|^2 / \mathcal{R}(f)$$

  $\Rightarrow$ Tikhonov minimizer

- Question: How plausible is the Tikhonov minimizer f_α^δ?
  How effective is one penalty compared to another ? ...
- $\Rightarrow$ tools for assessing the reliability of the inverse solution
  - Bayesian inference
  - interval analysis ...
  - sensitivity analysis ...
  - ...

Bayesian inversion

starting point: Bayes' formula, i.e., for two random variables f and g the conditional probability density of f given g is given by

$$p(f|g) = \frac{p(g|f)p(f)}{p(g)}$$

- $p(g|f)$: **likelihood function** — building block I
  the probability density of the data g given the unknown f
  incorporate information in g (physics / noise statistics of g)
- $p(f)$: **prior distribution** — building block II
  the probability density of f (regardless of g)
  encode a prior knowledge/information (before collecting data)
- $p(g) = \int p(g|f)p(f)df$: normalizing constant

- the unnormalized posteriori $p(f, g)$ defined by

$$p(f, g) = p(g|f)p(f),$$

and often write

$$p(f|g) \propto p(f, g)$$

the posteriori $p(f|g)$ up to a normalizing constant $p(g)$

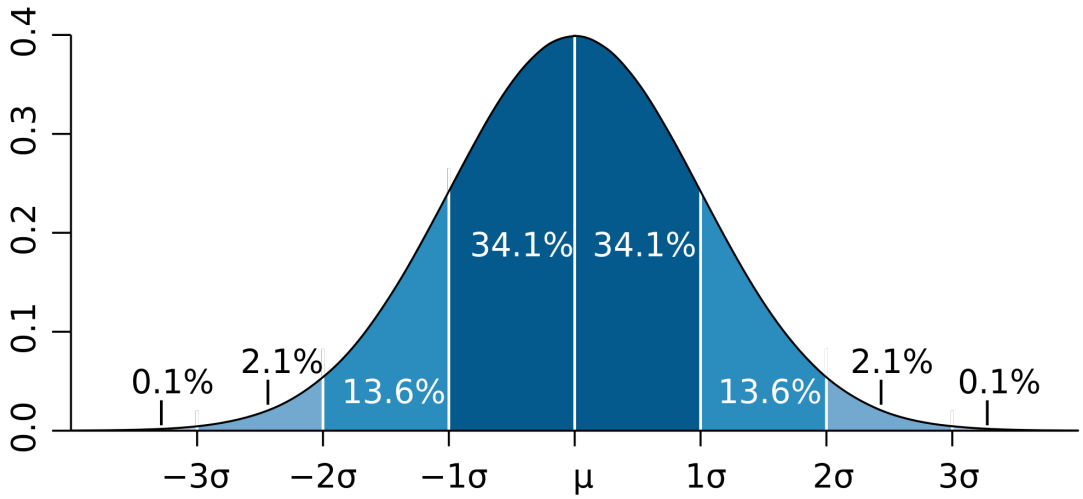
- the normalizing constant $p(g)$ is *important* for model comparison

...

posterior $p(f|g)$: the distr. of parameter f given the data g
$\Rightarrow$ Bayesian solution of the inverse problem

- $p(f|g)$ holds the full inform. about the inverse problem
- there are many plausible solutions to the inverse problem
- $\Rightarrow$ calibrating the uncertainty of specific inverse solution.

Gaussian distribution ...



- the mean μ of f under $p(f|g) \Rightarrow$ representative solution
- the variance σ of f under $p(f|g) \Rightarrow$ reliability/uncertainties

all these involve (possibly very high-dimensional) integrals

Building block I: the construction of likelihood function

likelihood function $p(g|f) \Leftarrow$ the noise statistics, the forward model

- all sources of errors: modeling error, discretization error, measurement error ...
  all are lumped into g despite the very different nature ...
- a careful modeling of all errors in the data g is essential for extracting useful information

The most popular model is the additive noise model

$$g = g^\dagger + \xi = F(f^\dagger) + \xi,$$

- given a realization $f^\dagger$ of f: g is a shift of the noise ξ by $F(f^\dagger)$
- ξ is from external source: it is assumed to be *independent* of $f \Rightarrow$

$$p(g|f) = p_\Xi(g - F(f))$$

additive iid Gaussian noise

$\xi \in \mathbb{R}^n$ is an i.i.d. realization of a Gaussian r.v. $N(0, \sigma^2)$,

$$\xi_i \sim N(0, \sigma^2) \Rightarrow p(g_i|f) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(g_i - F(f)_i)^2}$$

i.i.d. noise $\Rightarrow p(g|f) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\|F(f)-g\|^2}$

$$-\log p(g|f) = \frac{1}{2\sigma^2}\|F(f) - g\|^2 + c$$

general Gaussian case

$\xi \in \mathbb{R}^n$ is a realization of Gaussian r.v. $N(\mathbf{0}, \mathbf{C})$

$$p(g|f) = \frac{1}{(2\pi)^{\frac{n}{2}} |\det \mathbf{C}|^{\frac{1}{2}}} e^{-\frac{1}{2}(F(f)-g)^t \mathbf{C}^{-1}(F(f)-g)}$$

with $C = \sigma^2 I$, recovers the i.i.d. case
the discretization error is often colored ...

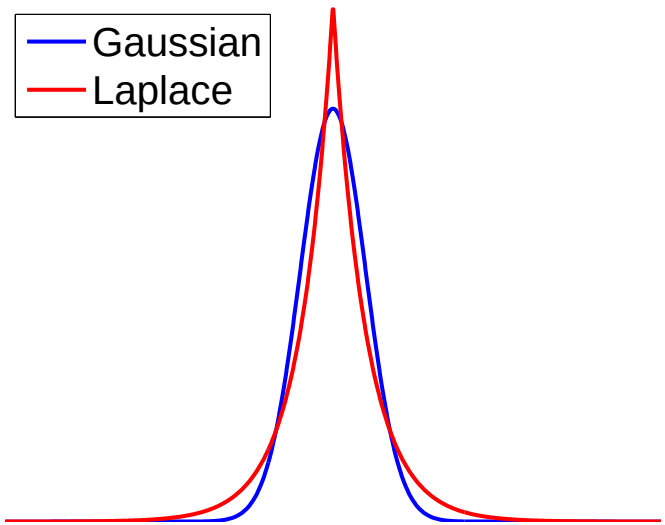
i.i.d. Laplace noise

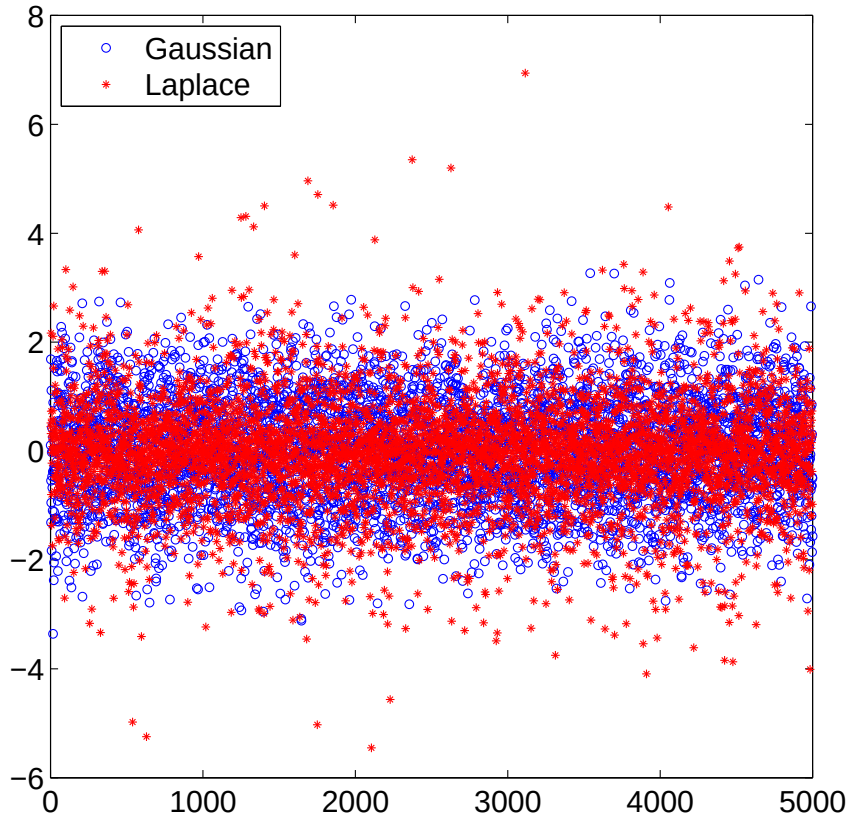
$\xi_i \sim \mathrm{Lap}(0, \sigma)$, $p_\Xi(\xi) = (2\sigma)^{-1} e^{-\sigma|\xi|}$; i.i.d. noise $\Rightarrow$

$$p(g|f) = \frac{1}{(2\sigma)^n} e^{-\sigma \|F(f)-g\|_1}$$

suitable for *impulsive* noise ...

Gaussian v.s. Laplace distribution with identical means/variance





Poisson noise

$g_i \sim Pois(g_i^\dagger)$, i.e.,

$$g_i = k, \quad \text{with } p(k) = \frac{(g_i^\dagger)^k}{k!} e^{-g_i^\dagger}, \quad k = 0, 1, \ldots$$

i.i.d. assumption on the noise model

$$p(g|f) = \prod_i \frac{e^{-(F(f))_i} (F(f)_i)^{g_i}}{g_i!}$$

suitable in photon-counting: positron emission tomography

all these analytical models are approximations, but some are better than others ...

Building block II: the construction of prior

The prior $p(f)$ encodes the prior knowledge about the sought-for solution f in a **probabilistic** manner.

- There is no universal way to construct prior, and it is *subjective*, dependent on personal experiences/knowledge
- the prior knowledge: expert opinion, historical investigations, statistical studies and anatomical knowledge etc.
- Since inverse problems are ill-posed: careful incorporation of all prior knowledge is important in any inversion technique
- the prior plays the role of regularization in a stochastic setting

Example: one-dimensional deblurring problem

$$g(t) = \int_0^1 k(s,t)f(s)ds \quad t \in [0,1]$$

- grid: $0 = s_0 < s_1 < \ldots < s_n = 1$, $0 = t_0 < t_1 < \ldots < t_n = 1$
- proper discretization: collocation, Galerkin
- discrete problem

$$g = Af$$

smoothness prior: the solution is *smooth*, e.g.,

$$f_i \approx f_{i-1}$$

too rigid $\Rightarrow$ include uncertainty in the belief by

$$f_i = f_{i-1} + \xi_i$$

with $\xi_i \sim N(0, \gamma^2)$

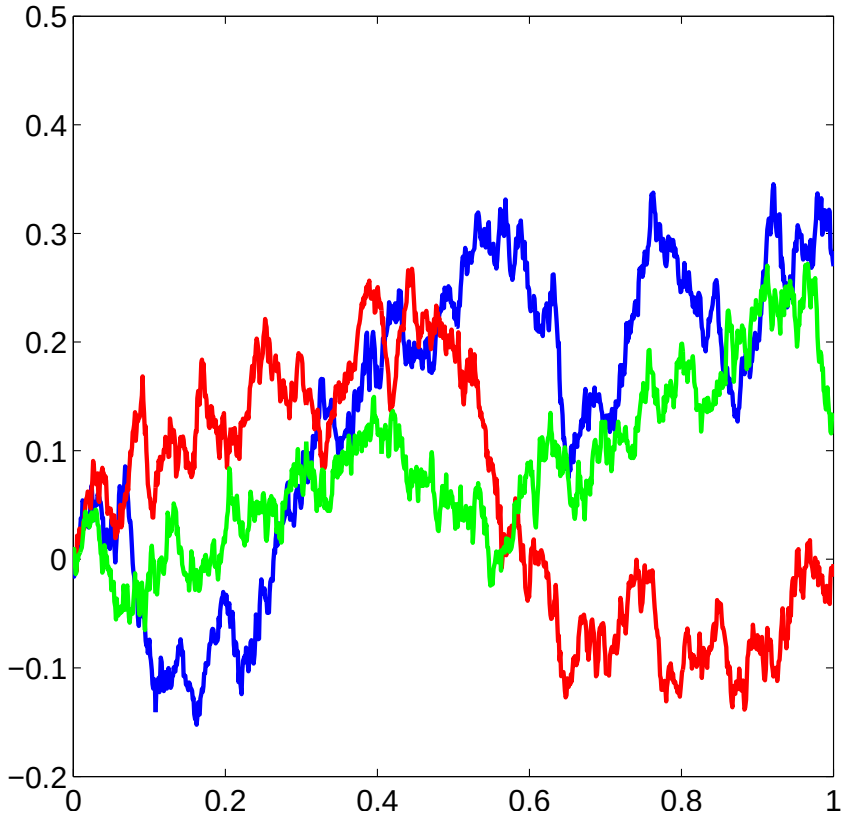
$$Lf = \xi$$

with L given by

$$L = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix}$$

first-order backward difference (up to a step size)

- prior $p(f) \propto e^{-\frac{1}{2\gamma^2}\|Lf\|^2}$
- the prior is not normalizable (improper), likelihood $p(g|f)$ hopefully fixes the problem



smoothness prior: solution is *smooth*

$$f_i = (f_{i-1} + f_{i+1})/2$$

value is close to its neighbors, but no uncertainties $\Rightarrow$ innovations

$$f_i = (f_{i-1} + f_{i+1})/2 + \xi_i$$

with $\xi_i \sim N(0, \gamma^2)$, i.e.,

$$Lf = \xi$$

with

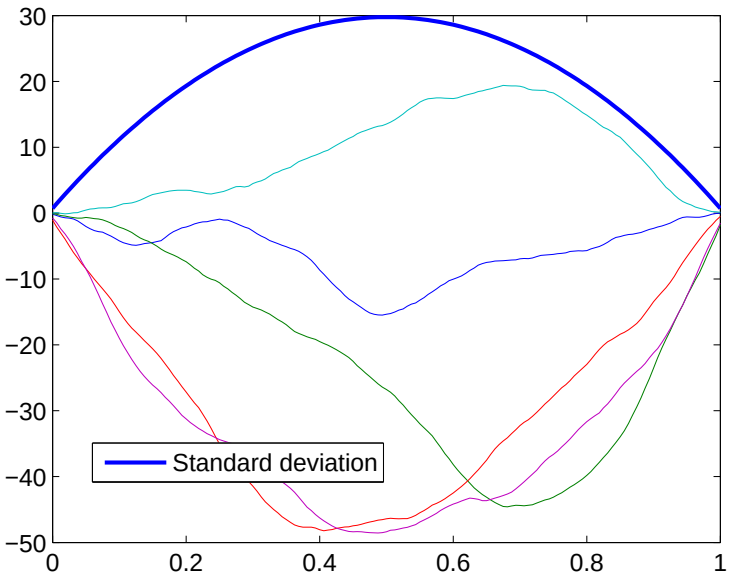
$$L = \frac{1}{2}\begin{bmatrix} -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \\ & & & -1 & 2 & 1 \end{bmatrix}$$

... discrete Laplacian ... (second-order derivative is small ...)

prior distribution

$$p(f) \propto e^{-\frac{1}{2\gamma^2}\|Lf\|^2}$$

- this is not a probability density function (not normalizable ...)
- cause: no proper boundary conditions for the Laplacian
- remedy: specify proper boundary condition / let likelihood fix it



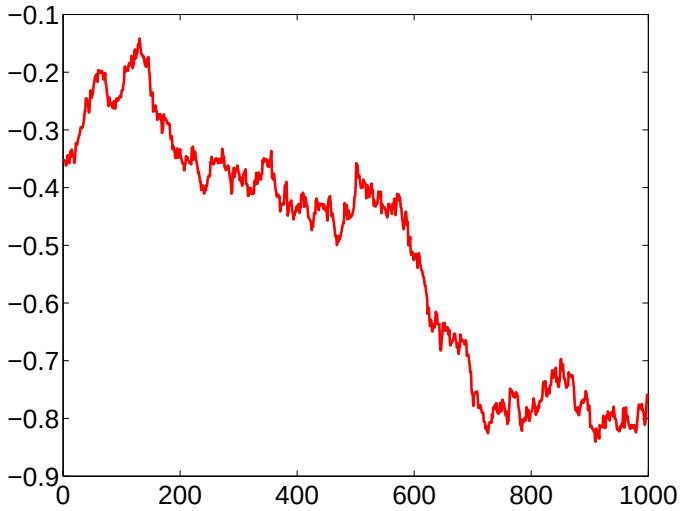
nonsmooth prior: locally smooth, but large jumps at *unknown* location

$$f_i = f_{i-1} + \xi_i$$

ξ_i mostly small with occasional large jumps $\Rightarrow \xi_i \sim Lap(0, \gamma)$?

$$Lf = \xi,$$

total variation prior distribution $p(f) \propto e^{-\gamma \|Lf\|_1}$



the samples not exactly piecewise constant like TV regularization ...

all can be regarded as special cases of Markov random field

$$p(f) \propto e^{-\lambda\psi(f)},$$

- $\psi(f)$: potential function dictates the interaction energy between the components of the random field f
- λ is a scale parameter, determining the strength of the local/global interactions.

  It plays the role of a regularization parameter in regularization theory, and its automated determination is important.

distinct features

- $p(f|g)$ is a probability distribution, and is an ensemble of solutions consistent with g (to various extent)

$$\mu = \int f p(f|g) df,$$

$$C = \int (f-\mu)(f-\mu)^{\mathrm{t}} p(f|g) df.$$

- elucidate the crucial role of proper statistical modeling in designing regularization formulations for practical problems.
- it provides a flexible regularization, *partially* resolving the nontrivial issue of choosing a regularization parameter.

hierarchical modeling: hyperparameters and beyond

likelihood $p(g|f)$ and prior $p(f)$ may contain unknown para, e.g.,

$$p(g|f) = p(g|f,\tau) \quad \text{and} \quad p(f) = p(f|\lambda)$$

- τ,λ: precision (inverse variance) and the scale parameter
- These parameters are generically known as hyperparameters
- Hierarchical Bayesian modeling provides an elegant approach to choose these parameters automatically

J. Wang, N. Zabaras. Inverse Problems 21(1), 183–206, 2005

hierarchical Bayesian modeling

- view λ and τ as random variables with their own priors
- determine them from the data g
- convenient choice: conjugate distribution

For both λ and τ, the conjugate distribution is a Gamma distribution:

$$p_\Lambda(\lambda) = G(\lambda; a_0, b_0) = \frac{b_0^{a_0}}{\Gamma(a_0)} \lambda^{a_0 - 1} e^{-b_0 \lambda},$$

$$p_\Upsilon(\tau) = G(\tau; a_1, b_1) = \frac{b_1^{a_1}}{\Gamma(a_1)} \tau^{a_1 - 1} e^{-b_1 \tau}.$$

- (a_0, b_0) and (a_1, b_1) determine the range of λ and τ
- noninformative prior is often adopted: $(a_0, b_0) \approx (1, 0)$

posterior distribution $p(f, \lambda, \tau | g)$

$$p(f, \lambda, \tau | g) \propto p(g | f, \tau) p(f | \lambda) p_\Lambda(\lambda) p_\Upsilon(\tau)$$

example: Gaussian noise model + Gaussian prior

$$p(g|f,\tau) \propto \tau^{\frac{n}{2}} e^{-\frac{\tau}{2}\|F(f)-g\|^2},$$
$$p(f|\lambda) \propto \lambda^{\frac{m}{2}} e^{-\frac{\lambda}{2}\|f\|^2}$$

the posterior distribution (for fixed λ and τ)

$$p(f|g) \propto e^{-(\frac{\tau}{2}\|F(f)-g\|^2+\frac{\lambda}{2}\|f\|^2)}$$

a customary approach: maximum a posteriori estimate

$$f_{\text{map}} = \arg\max_f p(f|g)$$

the *most probable* point of the posterior density

maximum a posteriori estimate $f_{\text{map}} \Rightarrow$

$$\begin{aligned} f_{\text{map}} &= \arg\max_f p(f|g) \\ &= \arg\min_f \left\{ \frac{\tau}{2}\|F(f) - g\|^2 + \frac{\lambda}{2}\|f\|^2 \right\} \end{aligned}$$

the functional in the curly bracket is

$$\frac{1}{2}\|F(f) - g\|^2 + \frac{\lambda\tau^{-1}}{2}\|f\|^2$$

Tikhonov regularization + smoothness constraint, with $\alpha = \lambda\tau^{-1}$
Message: A Tikhonov minimizer is an MAP estimate of some Bayesian formulation.

What if the parameters λ and τ are *unknown*?
$\Rightarrow$ hierarchical model conjugate prior on λ and τ

$$p(g|f,\tau) \propto \tau^{\frac{n}{2}} e^{-\frac{\tau}{2}\|F(f)-g\|^2}$$
$$p(f|\lambda) \propto \lambda^{\frac{m}{2}} e^{-\frac{\lambda}{2}\|f\|^2}$$
$$p(\tau) \propto \tau^{a_1-1} e^{-b_1\tau}$$
$$p(\lambda) \propto \lambda^{a_0-1} e^{-b_0\lambda}$$

with $(a_1, b_1), (a_0, b_0) \simeq (0, 0)$
$\Rightarrow$ posterior distr.

$$p(f,\lambda,\tau|g) \propto \tau^{\frac{n}{2}+a_1-1} e^{-\frac{\tau}{2}\|F(f)-g\|^2} \cdot \lambda^{\frac{m}{2}+a_0-1} e^{-\frac{\lambda}{2}\|f\|^2} \cdot e^{-b_1\tau} \cdot e^{-b_0\lambda}.$$

ways of exploring the posterior distribution $p(f, \lambda, \tau | g)$

- the *joint maximum a posteriori* estimate $(f, \lambda, \tau)_{\text{map}}$, i.e.,

$$\begin{aligned}(f, \lambda, \tau)_{\text{map}} &= \arg\max_{f,\lambda,\tau} p(f, \lambda, \tau | g) \\ &= \arg\min_{f,\lambda,\tau} J(f, \lambda, \tau),\end{aligned}$$

where the functional $J(f, \lambda, \tau)$ is given by

$$J(f, \lambda, \tau) = \frac{\tau}{2}\|F(f) - g\|^2 + \frac{\lambda}{2}\|f\|^2 - \tilde{a}_0 \ln \lambda + b_0 \lambda - \tilde{a}_1 \ln \tau + b_1 \tau,$$

with $\tilde{a}_0 = \frac{n}{2} + a_0 - 1$ and $\tilde{a}_1 = \frac{m}{2} + a_1 - 1$

augmented Tikhonov regularization

$$J(f,\lambda,\tau) = \frac{\tau}{2}\|F(f)-g\|^2 + \frac{\lambda}{2}\|f\|^2 - \tilde{a}_0 \ln\lambda + b_0\lambda - \tilde{a}_1 \ln\tau + b_1\tau.$$

- the first two terms recover Tikhonov regularization
- the rest terms automatically determine the regu. parameter.

What is beyond: MAP approach is straightforward, but without uncertainties ???

How to explore the posteriori ?

posteriori $p(f)$ lives in a high-dimensional space
$\Rightarrow$ not directly informative
$\Rightarrow$ compute summarizing statistics, e.g., mean μ and covariance C

$$\mu = \int f p(f) df \quad \text{and} \quad C = \int (f-\mu)(f-\mu)^{\mathrm{t}} p(f) df.$$

very high-dim. integrals, and standard quadrature rules are inefficient

$$m = 100,\ 2 \text{ points/dir} \Rightarrow 2^{100} \approx 1.27 \times 10^{30} \text{ points}$$

more efficient approach

- Monte Carlo methods, especially Markov chain Monte Carlo

J. S. Liu. Monte Carlo Strategies in Scientific Computing. Springer-Verlag, New York, 2001.

Monte Carlo simulation

- draw a large set of i.i.d. samples $\{f^{(i)}\}_{i=1}^{N}$ from the target distribution $p(f)$
- approximate the expectation $E_p[\zeta]$ of any function $\zeta : \mathbb{R}^m \to \mathbb{R}$ by the sample mean $E_N[\zeta]$

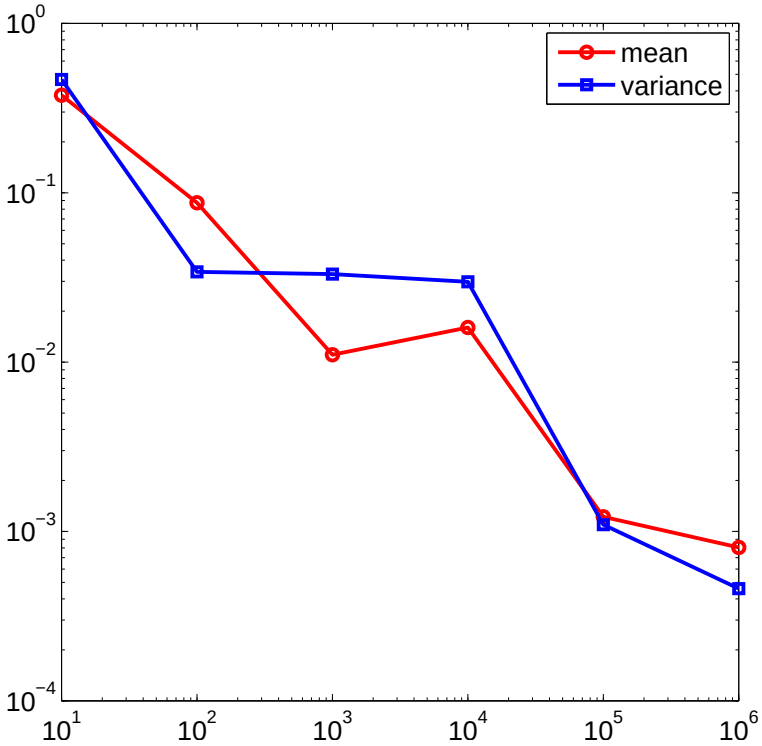
$$E_N[\zeta] \equiv \frac{1}{N}\sum_{i=1}^{N}\zeta(f^{(i)}) \to E_p[\zeta] = \int \zeta(f)p(f)df \quad \text{as } N \to \infty.$$

- the Monte Carlo integration error $e_N[\zeta]$ by

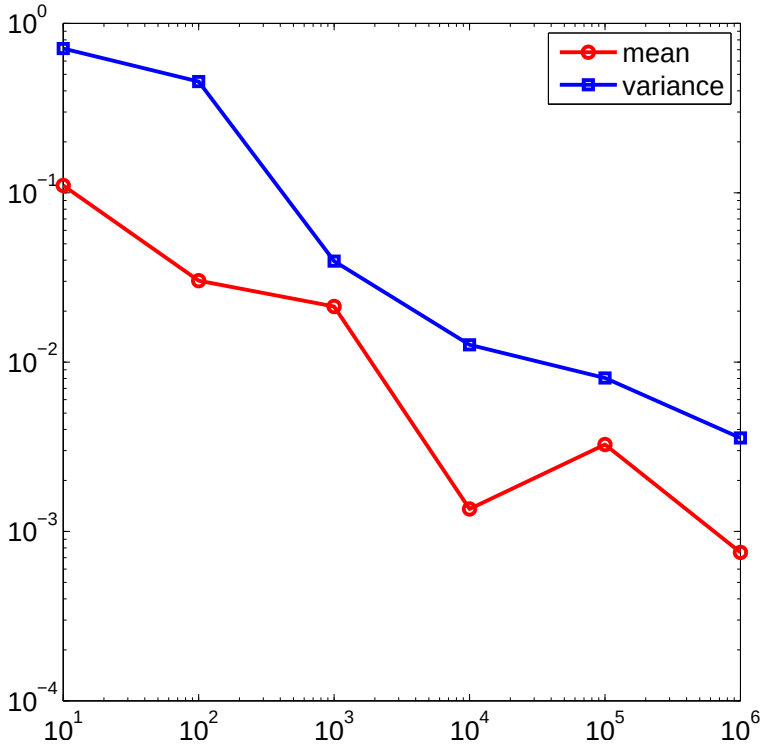
$$e_N[\zeta] = E_p[\zeta] - E_N[\zeta] \approx \mathrm{Var}_p[\zeta]^{\frac{1}{2}} N^{-1/2}\nu, \quad \nu \sim N(0,1)$$

  - the error $e_N[\zeta]$ is $O(N^{-1/2})$
  - with a constant $\sim$ the variance of the integrand ζ
  - the estimate is independent of the dimensionality m

the convergence of mean and variance for one-dim. distr.



normal



Laplace

- the convergence is not steady, not necessarily monotone
- the approximation of high-order moments is generally harder ...

i.i.d. samples from an implicit and high-dim. distr. are nontrivial.

- importance sampling

$q(f)$ is an easy-to-sample p.d.f. and close to the posteriori $p(f)$

approximate the expectation of the function ζ w.r.t. $p(f)$ by

$$\int \zeta(f)p(f)df = \int \zeta(f)\frac{p(f)}{q(f)}q(f)df \approx \frac{1}{N}\sum_{i=1}^{N}\zeta(f^{(i)})w_i,$$

i.i.d. samples $\{f^{(i)}\}_{i=1}^{N}$ are drawn from $q(f)$, with $w_i = \frac{p(f^{(i)})}{q(f^{(i)})}$.
The efficiency relies on the approximation quality of $q(f)$ to $p(f)$

Markov chain Monte Carlo (MCMC)
general-purposed approach for exploring posteriori $p(f)$

- basic idea: given $p(f)$, construct an aperiodic and irreducible Markov chain such that its stationary distribution is $p(f)$.
- By running the chain for sufficiently long, simulated values from the chain are *dependent* samples from $p(f)$, and used for computing summarizing statistics.

N. Metropolis, A. Rosenbluth, et al. J. Chem. Phys. 21 (6): 1087–1092, 1953.

W. K. Hastings. Biometrika 57 (1): 97–109, 1970.

The Metropolis-Hastings algorithm is the most basic MCMC method

1: Initialize $f^{(0)}$ and set N;
2: **for** $i = 0 : N$ **do**
3: sample $u \sim U(0,1)$;
4: sample $f^{(*)} \sim q(f^{(i)}, f^{(*)})$
5: **if** $u < \alpha(f^{(i)}, f^{(*)})$ **then**
6: $f^{(i+1)} = f^{(*)}$;
7: **else**
8: $f^{(i+1)} = f^{(i)}$;
9: **end if**
10: **end for**

- the uniform distribution $U(0,1)$
- $p(f)$: the target distribution
- $q(f, f')$ is an easy-to-sample proposal distribution

Given the candidate $f' \sim q(f, f')$, accept it as the new state of the chain with probability $\alpha(f, f')$:

$$\alpha(f, f') = \min\left\{1, \frac{p(f')q(f', f)}{p(f)q(f, f')}\right\}.$$

if we reject f', then the chain remains in the current state f.

- $p(f)$ enters only through α via the ratio $p(f')/p(f)$
  $\Rightarrow$ normalizing constant not needed
- if q is symmetric, i.e., $q(f, f') = q(f', f)$, $\alpha(f, f')$ reduces to

$$\alpha(f, f') = \min\left\{1, \frac{p(f')}{p(f)}\right\}.$$

The chain converges to $p(f)$ for any reasonable $q(f, f')$

random walker sampler (chain driven by random walk)

- If $q(f, f') = \theta(f' - f)$ for p.d.f. θ, then $f^{(*)} = f^{(i)} + \xi$, $\xi \sim \theta$
- θ: uniform, multivariate normal or t-distribution
- With i.i.d. Gaussian distribution $N(0, \sigma^2)$,

$$f_j^{(*)} = f_j^{(i)} + \xi, \quad \text{with } \xi \sim N(0, \sigma^2)$$

  σ^2 controls the size of the random walks, and should be carefully tuned to improve the MCMC convergence.

Heuristic: the optimal acceptance ratio ~ 0.25

independent sampler $q(f, f') = q(f')$

- the acceptance probability $\alpha(f, f')$

$$\alpha(f, f') = \min\{1, w(f')/w(f)\},$$

$w(f) = p(f)/q(f)$ is the importance weight function.

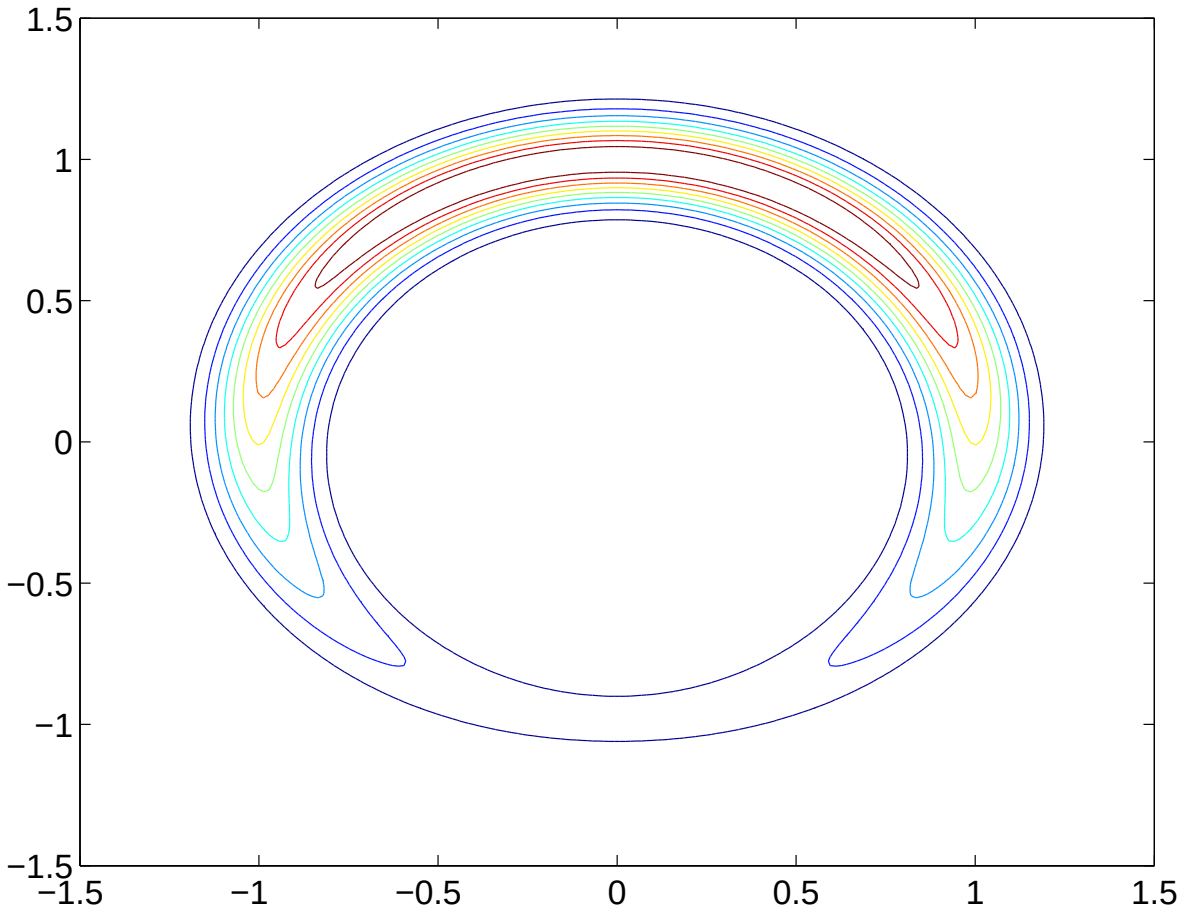
- how to generate $q(f)$: Gaussian approx. by linearized forward model, coarse-scale/reduced-order representation

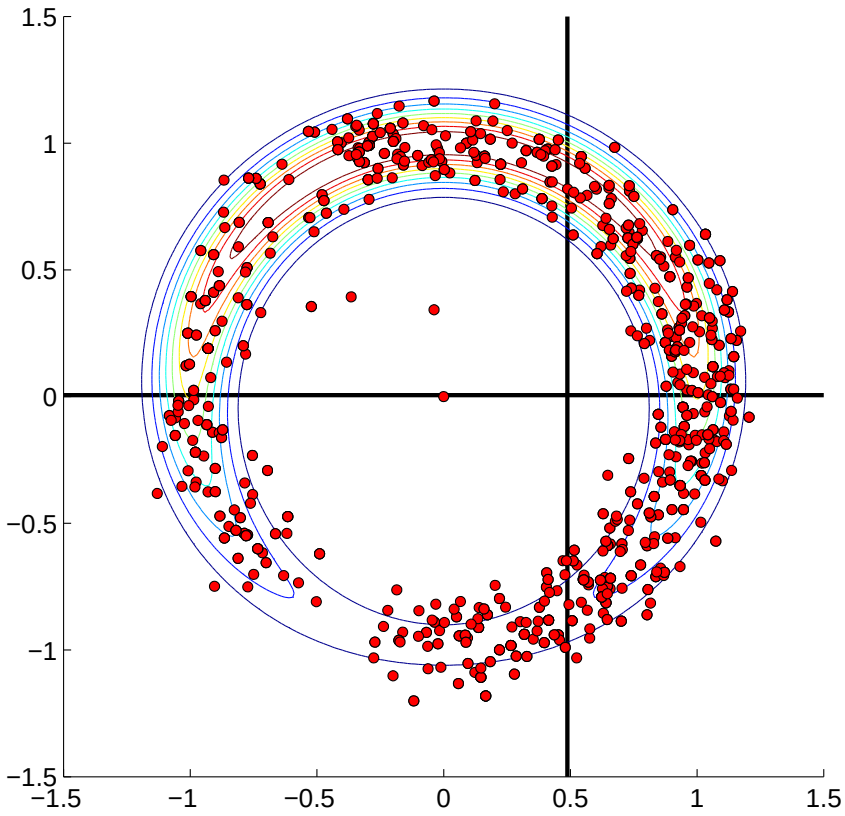
implementation issues

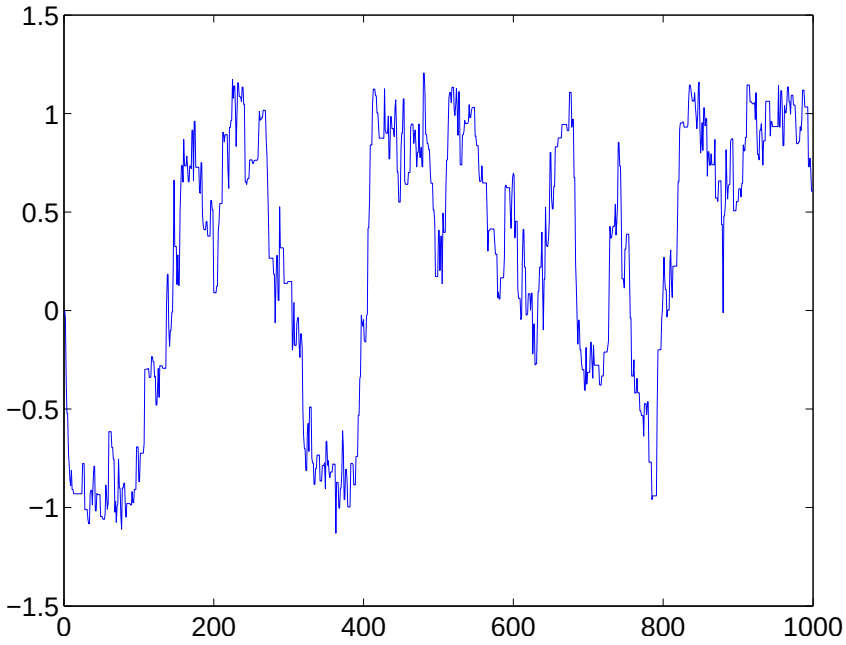
- the first samples are poor approximations as samples from $p(f)$
- discards these initial samples (burning-in)
- assess the convergence of the MCMC chains

sample the posteriori

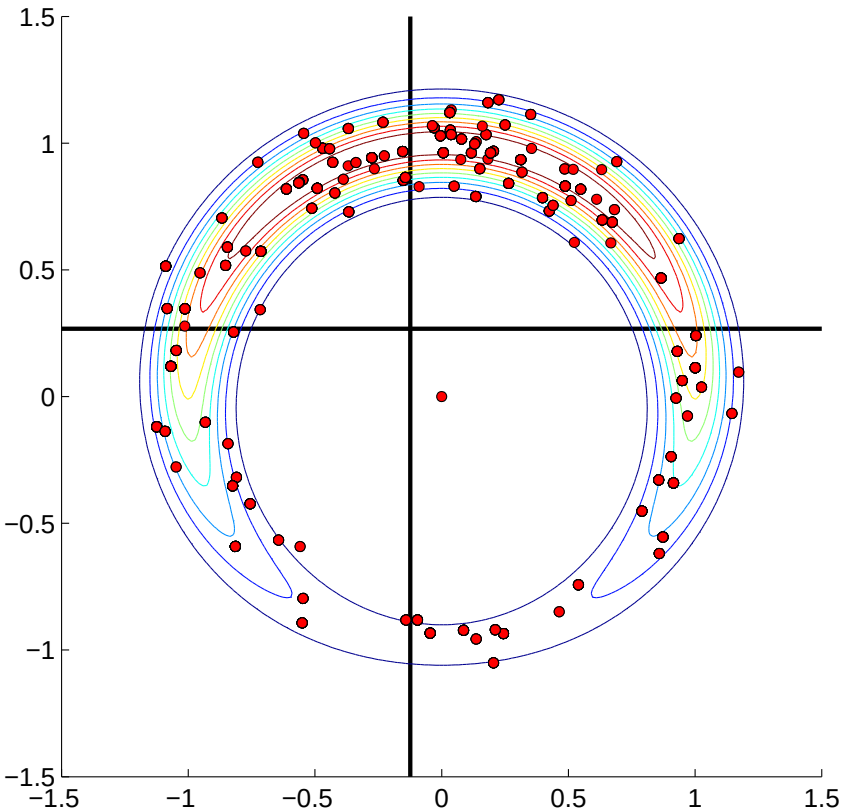
$$p(f) \propto e^{-50((f_1^2+f_2^2)^{1/2}-1)^2-0.5(f_2-1)^2}$$

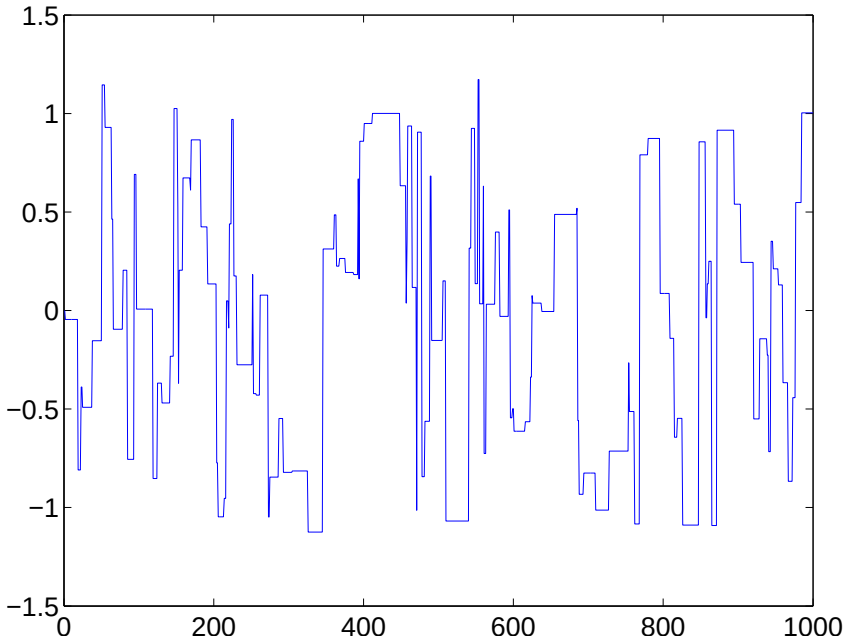




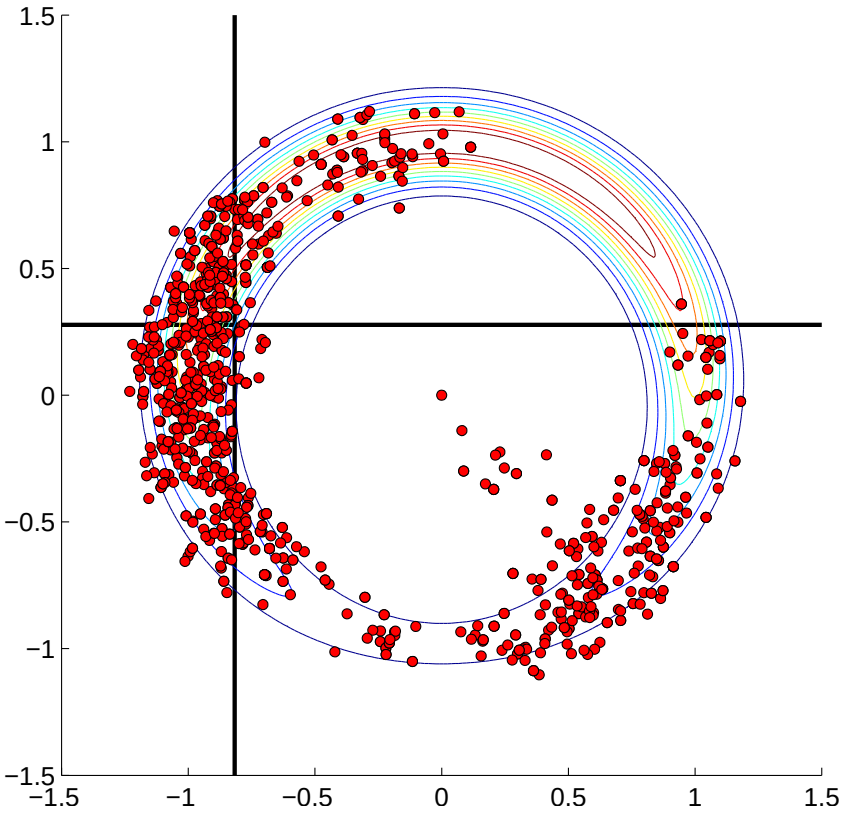


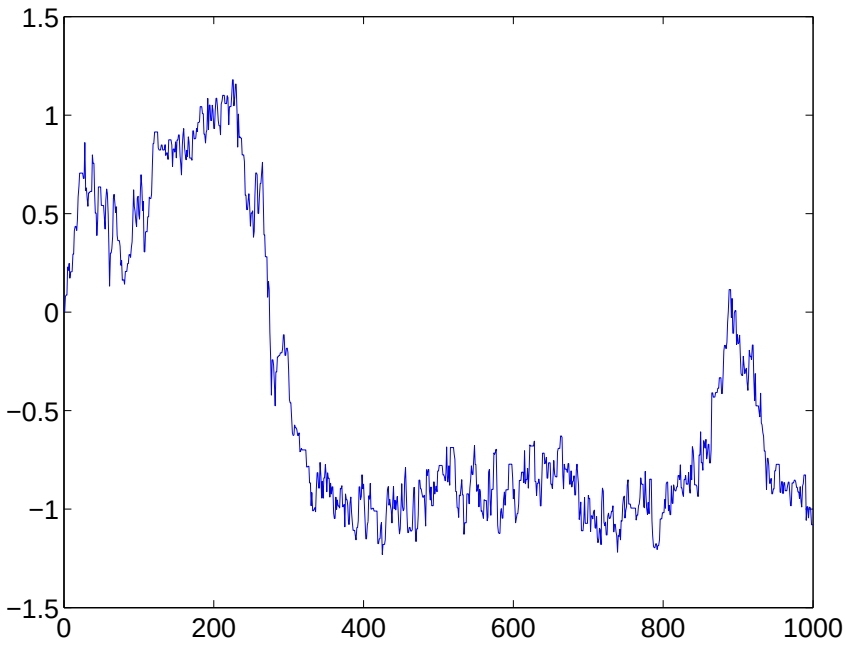
$\sigma = 0.2$, 1000 samples


1.5
1
0.5
0
−0.5
−1
−1.5
−1.5
−1
−0.5
0
0.5
1
1.5


1.5
1
0.5
0
−0.5
−1
−1.5
0
200
400
600
800
1000

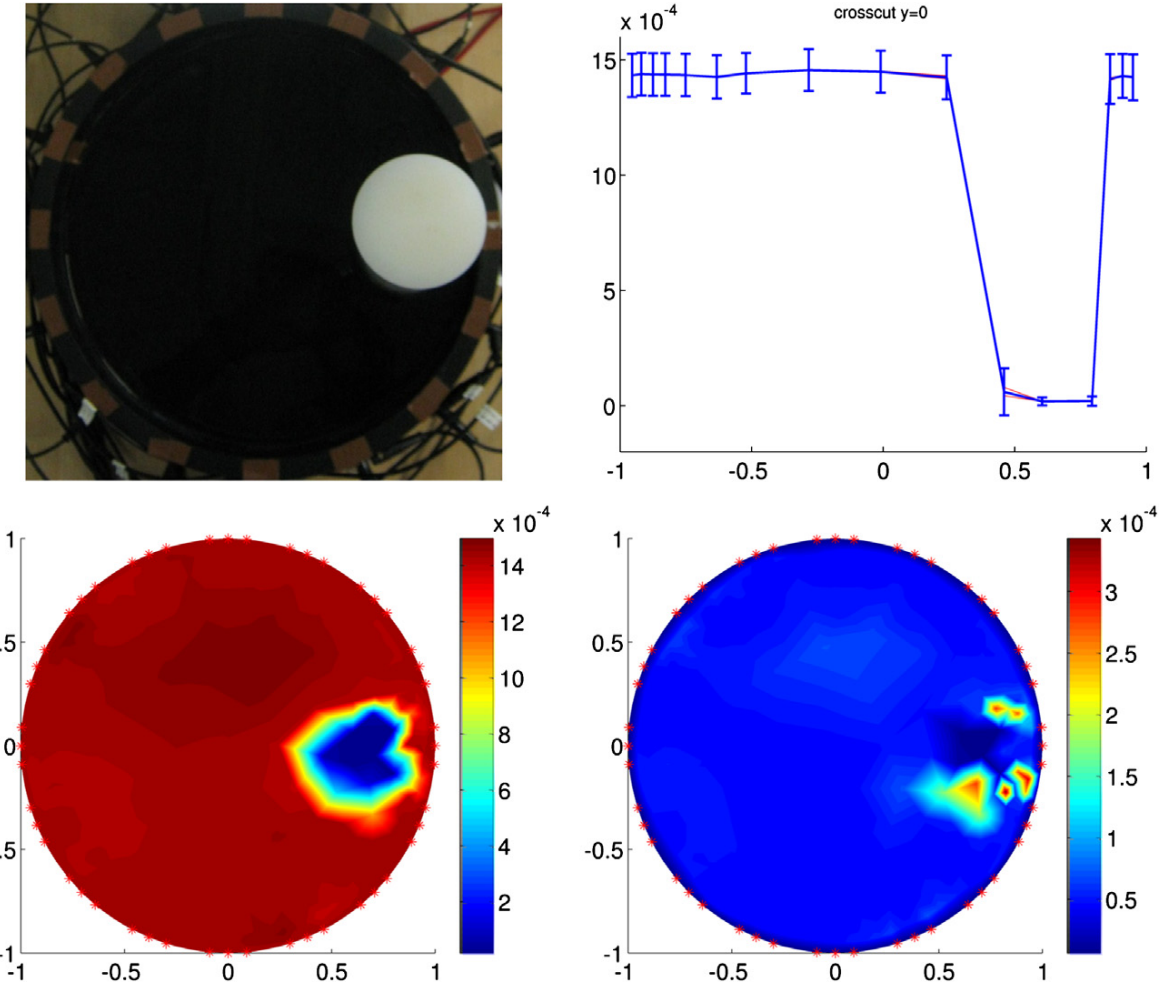
$\sigma = 1$





$\sigma = 0.1$

- the proper choice of σ is very crucial ...



230 unknowns, 1×10^7 samples, approximately one day

- If the state space is high dim., it is difficult to update the entire vector f in one single step since $\alpha(f, f')$ is often very small.
- to update only a part of f each time and to implement an updating cycle inside each step
- Gibbs sampler: update one component each time

to update f_i of f, proposal $q(f, f')$: the full conditional

$$q(f, f') = \begin{cases} p(f_i' | f_{-i}) & f_{-i}' = f_{-i}, \\ 0 & \text{otherwise}, \end{cases}$$

where f_{-i} denotes $(f_1, \ldots, f_{i-1}, f_{i+1}, \ldots, f_m)^{\mathrm{t}}$.

S. Geman, D. Geman. IEEE Trans. Pattern Anal. Mach. Int., 11(6), 721-741, 1984

With this proposal, the acceptance probability $\alpha(f, f')$ is given by

$$\begin{aligned}\alpha(f, f') &= \frac{p(f')q(f', f)}{p(f)q(f, f')} = \frac{p(f')/p(f'_i|f_{-i})}{p(f)/p(f_i|f'_{-i})} \\ &= \frac{p(f')/p(f'_i|f'_{-i})}{p(f)/p(f_i|f_{-i})} = \frac{p(f'_{-i})}{p(f_{-i})} = 1,\end{aligned}$$

these proposals are automatically accepted.

Gibbs algorithm

1: Initialize $f^{(0)}$ and set N.
2: **for** $i = 0 : N$ **do**
3: sample $f_1^{(i+1)} \sim p(f_1 | f_2^{(i)}, f_3^{(i)}, \ldots, f_m^{(i)})$,
4: sample $f_2^{(i+1)} \sim p(f_2 | f_1^{(i+1)}, f_3^{(i)}, \ldots, f_m^{(i)})$,
5: $\vdots$
6: sample $f_m^{(i+1)} \sim p(f_m | f_1^{(i+1)}, f_2^{(i+1)}, \ldots, f_{m-1}^{(i+1)})$,
7: **end for**

example: Gibbs sampler for Gaussian noise + smoothness prior $p(\lambda) \propto \lambda^{a_0-1} e^{-b_0\lambda}$ on the scale parameter λ, i.e., posteriori

$$p(f,\lambda) \propto e^{-\frac{\tau}{2}\|Af-g\|^2} \cdot \lambda^{\frac{m}{2}} e^{-\frac{\lambda}{2} f^t W f} \lambda^{a_0-1} e^{-b_0\lambda},$$

W encodes the local interaction structure

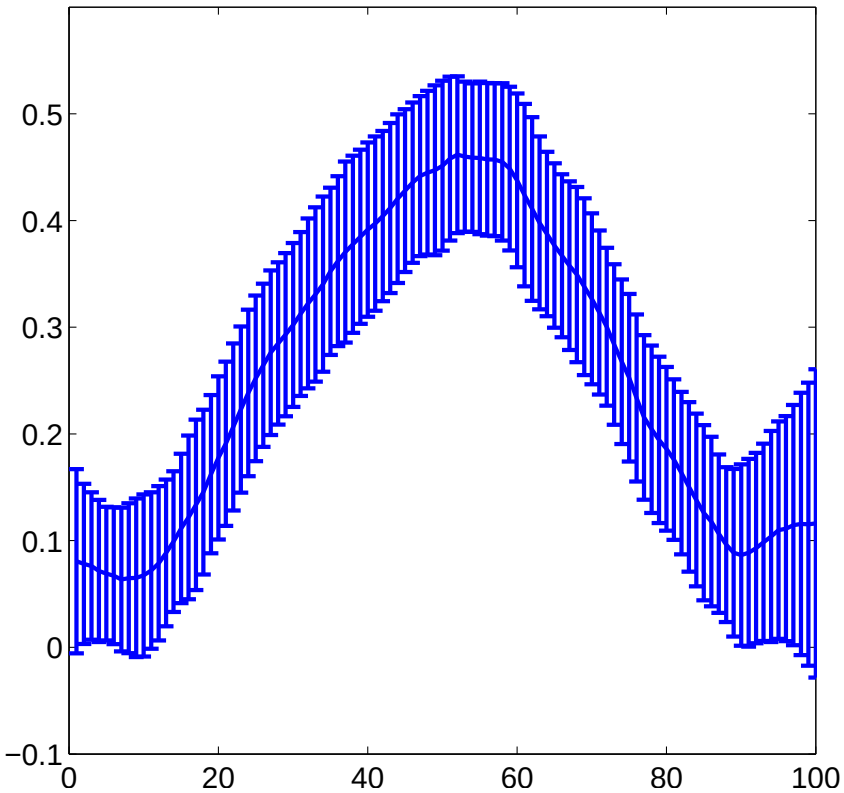
$$p(f_i|f_{-i},\lambda) \sim N(\mu_i,\sigma_i^2), \quad \mu_i = \frac{b_i}{2a_i}, \quad \sigma_i = \frac{1}{\sqrt{a_i}},$$

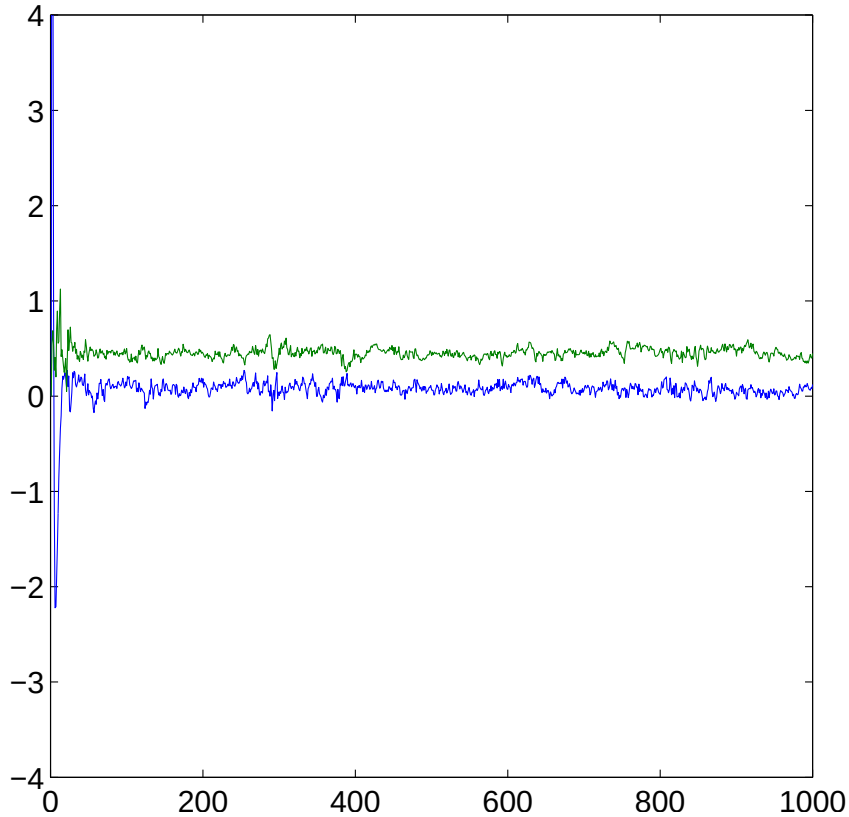
with a_i and b_i given by

$$a_i = \tau \sum_{j=1}^{n} A_{ji}^2 + \lambda W_{ii} \quad \text{and} \quad b_i = 2\tau \sum_{j=1}^{n} \mu_j A_{ji} - \lambda \mu_p,$$

and $\mu_j = g_j - \sum_{k\neq i} A_{jk} f_k$ and $\mu_p = \sum_{j\neq i} W_{ji} f_j + \sum_{k\neq i} W_{ik} f_k$
full conditional for λ:

$$p(\lambda|f) \sim G\left(\lambda; \tfrac{m}{2} + a_0, \tfrac{1}{2} f^t W f + \beta_0\right).$$





100-dimensional problem, 1000 samples, a few seconds

Approximate inference

Markov chain Monte Carlo: universal technique, asympotically exact, but generally expensive, convergence diagnosis

approximate inference methods

- variational Bayes Beal 2003; Jordan et al 1999
- expectation propagation Minka 2001...
- Stein variational gradient descent Liu-Wang 2016
- Laplace approximation
- Monte Carlo drop out
- ....

much faster, but can be limited in accuracy ...

variational Bayes ...

- main idea: to transform the inference problem of exploring $p(f|g)$ into an equivalent optimization problem, and solves it approx.
- probabilistic metrics: Kullback-Leibler divergence, L^1 metric, ...
- given two pdfs q and $\tilde{q}$, the Kullback-Leibler divergence $D_{KL}(q, \tilde{q})$ is defined by

$$D_{KL}(q, \tilde{q}) = \int q(f) \ln \frac{q(f)}{\tilde{q}(f)} df.$$

S. Kullback, R. Leibler. Ann. Math. Statistics 22, 79–86, 1951

- $D_{KL}(q, \tilde{q})$ is asymmetric in q and $\tilde{q}$, and does not satisfy the triangle inequality.
- by Jensen's inequality for the convex function $\varphi(f) = -\ln f$, the divergence $D_{KL}(q, \tilde{q})$ is always nonnegative:

$$D_{KL}(q, \tilde{q}) = -\int q(f) \ln \frac{\tilde{q}(f)}{q(f)} df \geq -\ln \int q(f) \cdot \frac{\tilde{q}(f)}{q(f)} df$$
$$= -\ln \int \tilde{q}(f) df = -\ln 1 = 0,$$

and it vanishes if and only if $q = \tilde{q}$ almost everywhere

Example: $q(f) = N(f; \mu, C)$ and $\tilde{q}(f) = N(f; \tilde{\mu}, \tilde{C})$, with $\mu, \tilde{\mu} \in \mathrm{R}^m$ and $C, \tilde{C} \in \mathbb{R}^{m \times m}$ being positive definite, is given by

$$D_{KL}(q, \tilde{q}) = \frac{1}{2}\left[(\mu - \tilde{\mu})^{\mathrm{t}} \tilde{C}^{-1}(\mu - \tilde{\mu}) + \mathrm{tr}(C\tilde{C}^{-1}) - m - \ln \frac{|C|}{|\tilde{C}|}\right],$$

if $C = \tilde{C}$, the Kullback-Leibler divergence $D_{KL}(q, \tilde{q})$ simplifies to

$$D_{KL}(q, \tilde{q}) = \tfrac{1}{2}(\mu - \tilde{\mu})^{\mathrm{t}} C^{-1}(\mu - \tilde{\mu})$$

- For any pdf. $q(f)$, the normalizing constant $p(g)$ satisfies

$$\ln p(g) \geq \int q(f) \ln \frac{p(f,g)}{q(f)} df.$$

- The lower bound

$$\mathrm{LB}(g,q) := \int q(f) \ln \frac{p(f,g)}{q(f)} df$$

- $D_{KL}(q(f), p(f|g))$: the gap between $\ln p(g)$ and lower bound $\max \mathrm{LB} \equiv \min D_{KL}(q(f), p(f|g))$ w.r.t. q.

Why the KL minimization is hard ... coupling ...

- the separability assumption *decouples* the dependence between different factors
  $\Rightarrow$ make the approximation computable $\Rightarrow$ mean field approx
- separability assumption yields explicit solutions for each component in terms of the others
  $\Rightarrow$ alternating direction iterative scheme

M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, L.K. Saul. Mach. Learn. 37, 183–233, 1999.

the assumption $q(f) = \prod q_i(f_i)$ with f_i being disjoint $\Rightarrow$

$$\begin{aligned} \mathrm{LB} &= \int \prod_i q_i(f_i) \left\{ \ln p(f,g) - \sum_i \ln q_i(f_i) \right\} df \\ &= \int q_j(f_j) \left\{ \int \ln p(f,g) \prod_{i \neq j} q_i(f_i) df_{-j} \right\} df_j \\ &\quad - \int q_j(f_j) \ln q_j(f_j) df_j + T(f_{-j}), \end{aligned}$$

define a joint density $\tilde{p}(f_j, g) \propto e^{E_{q_{-j}(f_{-j})}[\ln p(f,g)]}$, $\Rightarrow$

$$\mathrm{LB} = \int q_j(f_j) \ln \frac{\tilde{p}(f_j, g)}{q_j(f_j)} df_j + T(f_{-j}).$$

the optimal $q_j(f_j)$ is then given by

$$q_j(f_j) = \tilde{p}(f_j|g) \propto e^{E_{q_{-j}(f_{-j})}[\ln p(f,g)]},$$

Variational Bayes under the product density restriction

1: Initialize $q_i(f_i), \text{ı} = 2, \ldots, M$.
2: **for** $k = 1 : K$ **do**
3: Update sequentially $q_j(f_j)$ by

$$q_j^{k+1}(f_j) = \frac{e^{E_{q^k_{-j}(f_{-j})}[\ln p(f,g)]}}{\int e^{E_{q^k_{-j}(f_{-j})}[\ln p(f,g)]}}$$

4: Check the stopping criterion.
5: **end for**

- the convexity of KLD $\Rightarrow$ the convergence to a local optimum
- especially suited to conjugate families

Example: linear inversion + sparsity constraints

- scale mixture representation for the Laplace prior

$$\frac{\lambda}{2} e^{-\lambda|z|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-\frac{z^2}{2s}} \cdot \frac{\lambda^2}{2} e^{-\frac{\lambda^2}{2} s} ds.$$

$\Rightarrow$ hierarchical representation of the Bayesian model

$$p(f, w|g) \propto e^{-\frac{\tau}{2}\|Af-g\|^2} \prod_i w_i^{-\frac{1}{2}} e^{-\frac{f_i^2}{2w_i}} \frac{\lambda^2}{2} e^{-\frac{\lambda^2}{2} w_i}.$$

- obstruction: the strong coupling between x and w.
- mean-field approximation on x and w $\Rightarrow$

$$\ln q^k(f) \propto E_{q^{k-1}(w)}[\ln p(f, w, g)],$$
$$\ln q^k(w) \propto E_{q^k(f)}[\ln p(f, w, g)].$$

- Gaussian distribution

$$E_{q^{k-1}(w)}[\ln p(f,\lambda,g)] = E_{q^{k-1}(w)}[-\tfrac{\tau}{2}\|Af-g\|^2 - \sum \tfrac{1}{2w_i} f_i^2] + T(w)$$
$$= -\tfrac{\tau}{2}\|Af-g\|^2 - \tfrac{1}{2} f^{\mathrm{t}} W f + T(w),$$

with $[W]_{ii} = E_{q^{k-1}(w_i)}[\frac{1}{w_i}]$

- generalized inverse Gaussian

$$E_{q^k(f)}[\ln p(f,w,g)] = \sum_i E_{q^k(f)}[-\tfrac{1}{2}\ln w_i - \tfrac{f_i^2}{2w_i} - \tfrac{\lambda^2}{2} w_i] + T(f)$$
$$= \sum_i [-\tfrac{1}{2}\ln w_i - \tfrac{1}{2w_i} E_{q^k(f)}[f_i^2] - \tfrac{\lambda^2}{2} w_i] + T(f),$$

$$q^k(w) = \prod_i GIG\left(w_i; \tfrac{1}{2}, \lambda^2, E_{q^k(f)}[f_i^2]\right).$$

these two updates that the iterates $q^k(f)$ and $q^k(w)$ would remain Gaussian and (generalized) inverse Gaussian, respectively

outlook

- different perspectives on linear inversion ....
- there are still many unknown/unclear situations computationally / theoretically
- theoretical understanding of approximate inference techniques ...