Westfälische
Wilhelms-Universität
Münster

The Bayesian Approach to Inverse Problems: Computational Aspects

Invited Talk at the University of Cambridge, UK

wissen.leben
WWU Münster

Felix Lucka

14.11.2012

Outline

Exemplary Application: Inverse Problems with Sparsity Constraints

Basic Algorithms for Integration-based Posterior Inference

Iterative Optimization and Sampling

Selected Advanced Topics and Trends (optional)

Take Home Messages, Conclusions & Ongoing Work

Sparsity Constraints in Inverse Problems

Current trend in high dimensional inverse problems: Sparsity constraints.

- Total Variation (TV) imaging: Sparsity constraints on the gradient of the unknowns.
- Compressed Sensing: High quality reconstructions from a small amount of data, if a sparse basis/dictionary is a-priori known (e.g., wavelets).

Send me your best
TV image!
felix.lucka@wwu.de

Send me your best
CS image!
felix.lucka@wwu.de

Sparsity Constraints in Variational Regularization

Commonly applied formulation and analysis by means of variational regularization, mostly by incorporating L1-type norms:

$$\hat{u}_\alpha = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \|f - K\,u\|_2^2 + \alpha\,|D\,u|_1 \right\}$$

assuming additive Gaussian i.i.d. noise $\sim \mathcal{N}(0, \sigma^2)$

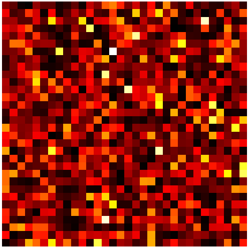
Notation:

- $f \in \mathbb{R}^k$: The noisy measurement data given
- $u \in \mathbb{R}^n$: The unknowns to recover w.r.t. the chosen discretization
- $K \in \mathbb{R}^{k \times n}$: Discretization of the forward operator w.r.t. the domains of u and f.
- $D \in \mathbb{R}^{l \times n}$: Discrete formulation of the mapping onto the (potentially) sparse quantity.

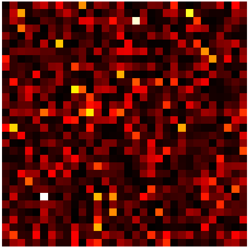
Sparsity Constraints in the Bayesian Approach

Sparsity as a-priori information are encoded into the prior distribution $p_{prior}(u)$:

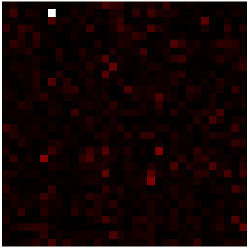
1. Turning the functionals used in variational regularization directly into priors, e.g., L1-type priors:
   - Convenient, as prior is log-concave.
   - MAP estimate is sparse, but the prior itself is not sparse.
2. Hierarchical Bayesian modeling: Sparsity is incorporated at a higher level of the model.
   - $\rightsquigarrow$ Next talk
   - Relies on a slightly different concept of sparsity.
   - Resulting implicit priors over unknowns are usually not log-concave.



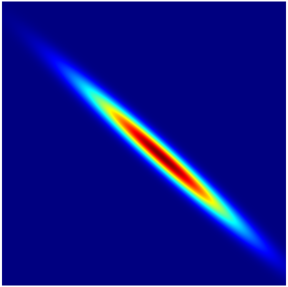
(c) $\exp\left(-\frac{1}{2}\|u\|_2^2\right)$



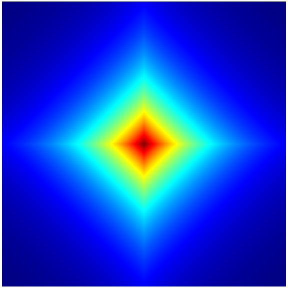
(d) $\exp\left(-|u|_1\right)$



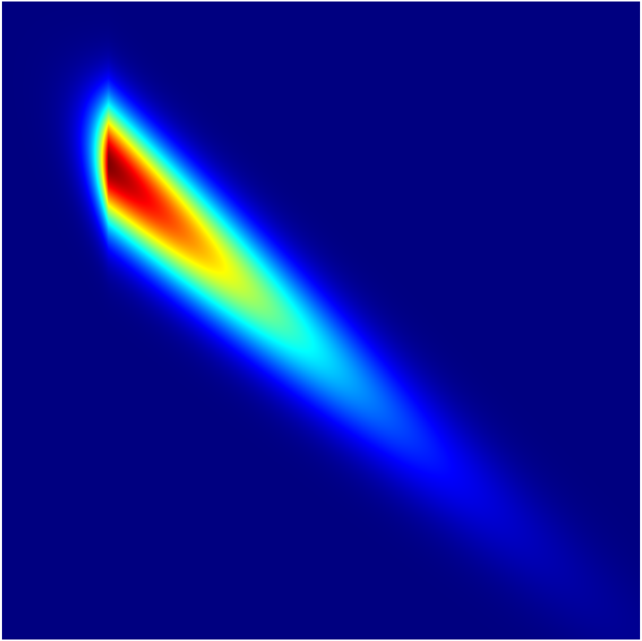
(e) $(1+u^2/3)^{-2}$



Likelihood:
$\exp\left(-\frac{1}{2\,\sigma^2}\|f - K\,u\|_2^2\right)$



Prior: $\exp\left(-\lambda\,|u|_1\right)$
(λ via discrepancy principle)



Posterior: $\exp\left(-\frac{1}{2\,\sigma^2}\|f - K\,u\|_2^2 - \lambda\,|u|_1\right)$

Reminder: Bayesian Inference and Advanced Techniques

Things we might want to do with the posterior:

- Point estimates: MAP and CM.
- Credible regions estimates
- Extreme value probabilities
- Conditional covariance estimates
- Histogram estimates
- Generalized Bayes estimators
- Marginalization of nuisance parameters & Approximation error modeling
- Model selection or averaging
- Experiment design

Computationally, this needs

- high-dimensional optimization
- high-dimensional integration
- a mix of both.

Integration-based Inference for Sparse Bayesian Inversion: How I got into this...

I assisted for the computational parts in (S. Comelli, 2011; *A Novel Class of Priors for Edge-Preserving Methods in Bayesian Inversion*) and faced many problems:

- Standard techniques for high-dimensional integration break down for ill-posed inverse problems.
- Especially for those with sparsity constraints.

[1] Optimization methods are challenging as well...but you know better about this than me ;-)

Integration-based Inference for Sparse Bayesian Inversion: How I got into this...

I assisted for the computational parts in (S. Comelli, 2011; *A Novel Class of Priors for Edge-Preserving Methods in Bayesian Inversion*) and faced many problems:

- Standard techniques for high-dimensional integration break down for ill-posed inverse problems.
- Especially for those with sparsity constraints.

$\Longrightarrow$ Good example to explain and illustrate the computational aspects of integration-based Bayesian inference[1]. To do so, this talk uses partial results from:

F.L., 2012.
Fast Markov chain Monte Carlo sampling for sparse Bayesian inference in high-dimensional inverse problems using L1-type priors
Inverse Problems (accepted); arXiv:1206.0262v2

[1] Optimization methods are challenging as well...but you know better about this than me ;-)

Outline

Exemplary Application: Inverse Problems with Sparsity Constraints

Basic Algorithms for Integration-based Posterior Inference

Iterative Optimization and Sampling

Selected Advanced Topics and Trends (optional)

Take Home Messages, Conclusions & Ongoing Work

Monte Carlo Integration in a Nutshell

$$\boxed{\mathbb{E}\left[f(x)\right] = \int_{\mathbb{R}^n} f(x)\, p(x)\, \mathrm{d}x}$$

- *Traditional Gauss-type quadrature:*
  Construct suitable grid $\{x_i\}_i$, w.r.t $\omega(x) := p(x)$ and approximate by $\sum_{i=1}^K \omega_i f(x_i)$.
  $\Longrightarrow$ Grid construction and evaluation infeasible in high dimensions.

- *Monte Carlo integration idea:*
  Generate suitable grid $\{x_i\}_i$, w.r.t $p(x)$ by drawing $x_i \sim p(x)$ and approximate by $\frac{1}{K}\sum_{i=1}^K f(x_i)$. By the *Law of large numbers*:

$$\frac{1}{K}\sum_{i=1}^{K} f(x_i) \xrightarrow{K\rightarrow\infty} \mathbb{E}_{p(x)}\left[f(x)\right] = \int_{\mathbb{R}^n} f(x)\, p(x)\, \mathrm{d}x$$

in L1 with rate $O(K^{-1/2})$ (independent of n).

Markov Chain Monte Carlo

Not able to draw independent samples?
$\rightsquigarrow$ With $\{x_i\}_i$ being an ergodic Markov chain, it still works!

Markov chain Monte Carlo (MCMC) methods are algorithms to construct such a chain:

- Huge number of MCMC methods exists.
- No "universal" method.
- Most methods rely on one two basic schemes:
  - Metropolis-Hastings (MH) Sampling [Metropolis et al., 1953; Hastings, 1970]
  - Gibbs Sampling [Geman & Geman, 1984]
- Posteriors from inverse problems seem to be "special".

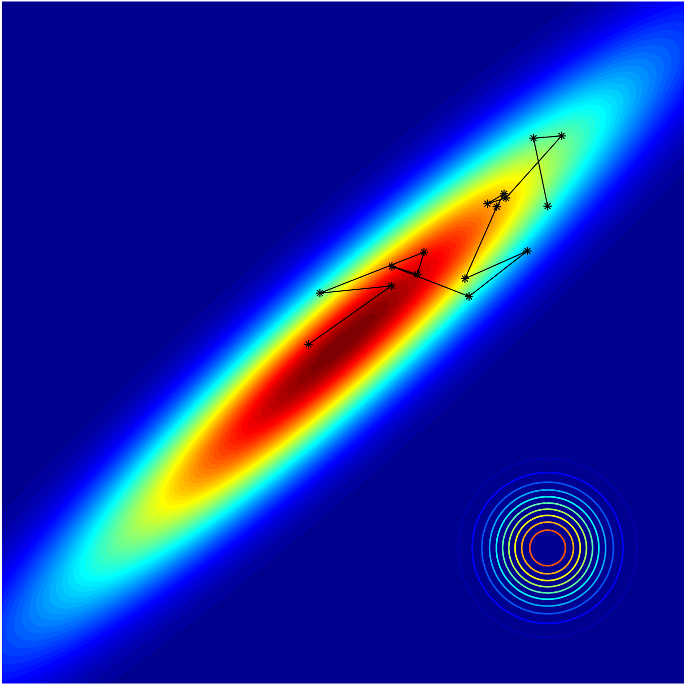
In this section: Comparison between the most basic variants of MH and Gibbs sampling for our specific scenario.

Symmetric, Random-Walk Metropolis-Hastings Sampling

```
Given: Density $p(x), x \in \mathbb{R}^n$ to sample from.
Let $p_{pro}(z)$ be a symmetric density in $\mathbb{R}^n$ and $x_0 \in \mathbb{R}^n$ an initial state. Define
burn-in size $K_0$ and sample size $K$.
For $i = 1,\ldots,K_0 + K$ do:
    1 Draw $z$ from $p_{pro}(z)$ and set $y = x_{i-1} + z$
    2 Compute the acceptance ratio $r = \frac{p(y)}{p(x_{i-1})}$
    3 Draw $\theta \in [0, 1]$ from a uniform probability density.
    4 If $r \geqslant \theta$, set $x_i = y$, else set $x_i = x_{i-1}$.
Return $x_{K_0+1}, \ldots, x_K$.
```

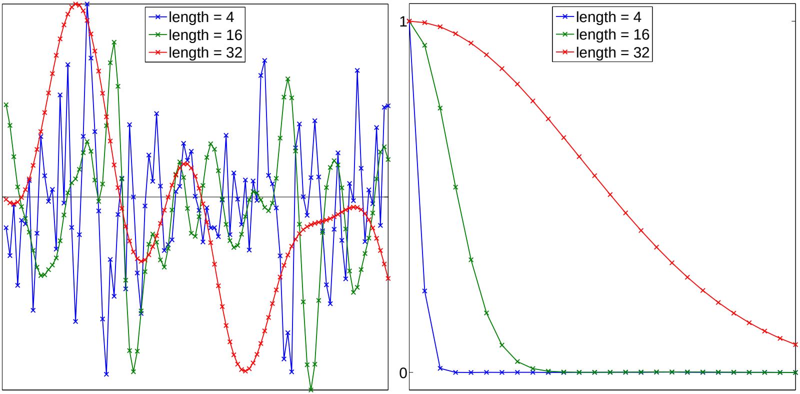
- Requires one evaluation of $p(x)$ and one sample from p_{pro} per step, no "real" knowledge about p is needed, not even normalization.
  $\rightsquigarrow$ "Black box" sampling algorithm.
- Most widely used.
- Good performance requires careful tuning of p_{pro}!
- Basis for very sophisticated sampling algorithms. $\rightsquigarrow$ more later.

In this talk:
$p_{pro} = \mathcal{N}(0, \kappa^2 \; I_n)$



Evaluate performance of a sampler via autocorrelation functions (acf):

- Desired: Independent samples of $p(x)$.
- $R(\tau) \in [0, 1]$ measures the average correlation between samples x_i, $x_{i+\tau}$ w.r.t. to a test function.
- A rapid decay of $R(\tau) \implies$ Samples get uncorrelated fast!



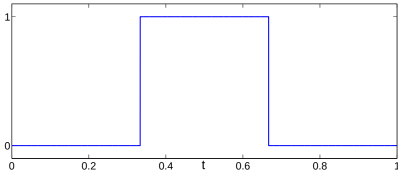
(a) Stochastic processes...

(b) ...and their autocorrelation functions

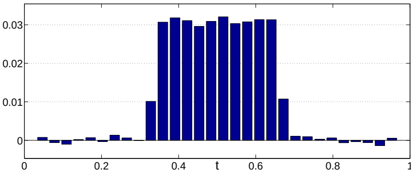
Total Variation Deblurring Example in 1D (from Lassas & Siltanen, 2004)

- Model of a charge coupled device (CCD) in 1D.
- Unknown light intensity $\tilde{u} : [0,1] \to \mathbb{R}^+$, indicator on $[\frac{1}{3}, \frac{2}{3}]$.
- Integrated into $k = 30$ CCD pixels $[\frac{1}{k+2}, \frac{k+1}{k+2}] \subset [0,1]$.
- Noise is added.
- $\tilde{u}$ is reconstructed on a regular, n-dim. grid.
- D is the forward finite difference operator with NB cond.

$$\boxed{p_{post}(u|f) \propto \exp\left(-\frac{1}{2\,\sigma^2}\|f - K\,u\|_2^2 - \lambda\,|D\,u|_1\right)}$$



(c) The unknown function $\tilde{u}(t)$



(d) The measurement data f

Performance of the MH Sampler

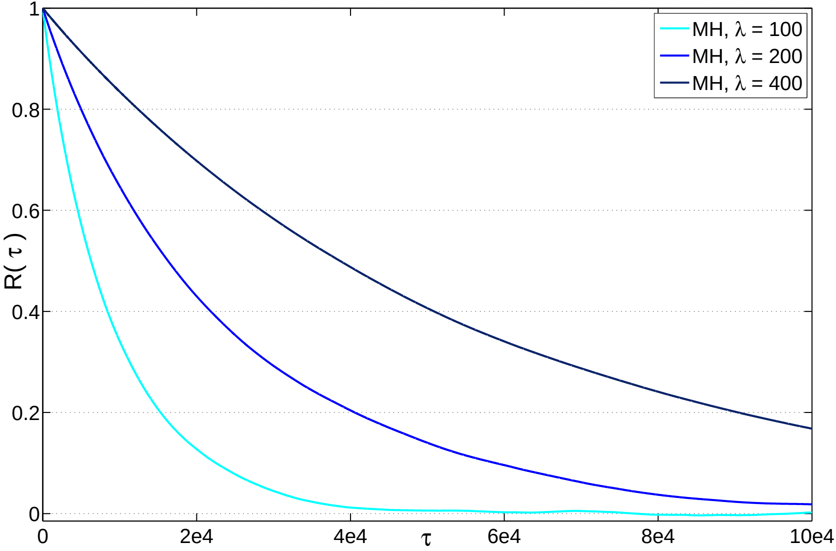


Figure: Autocorrelation plots $R(\tau)$ for MH Sampler and $n = 63$.

Performance of the MH Sampler

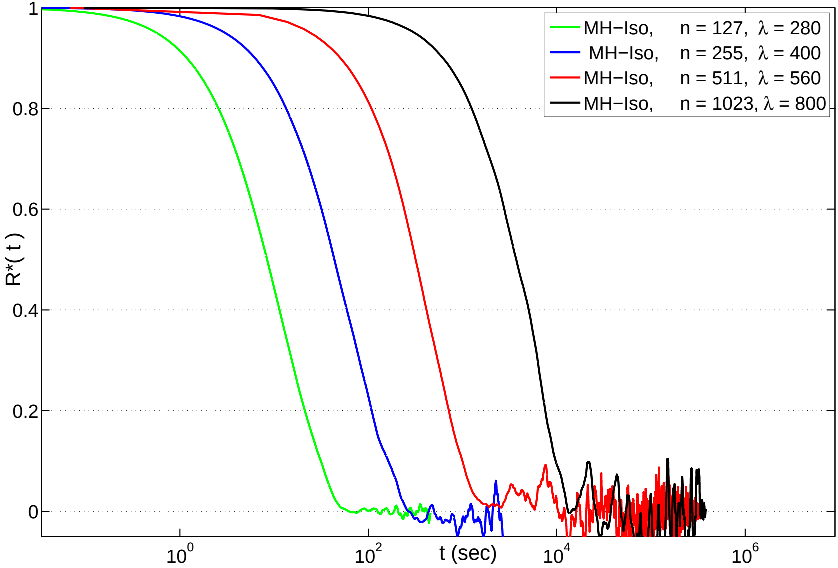


Figure: Temporal autocorrelation plots $R^*(t)$ for MH Sampler.

Single Component Gibbs Sampling

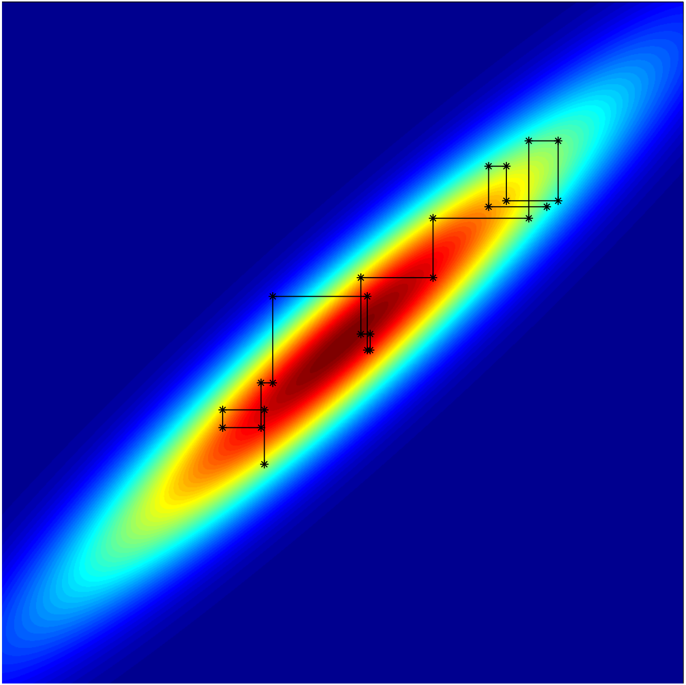
```
Given: Density $p(x), x \in \mathbb{R}^n$ to sample from.
Let $x_0 \in \mathbb{R}^n$ be an initial state. Define burn-in size $K_0$ and sample size $K$.
For $i = 1,\ldots,K_0 + K$ do:
    1 Set $x_i := x_{i-1}$.
    2 For $j = 1,\ldots,n$ do:
        (i) Draw $s$ randomly from $\{1, \ldots, n\}$ (random scan).
        (ii) Draw $(x_i)_s$ from the conditional, 1-dim density $p(\,\cdot\,|(x_i)_{[-s]})$.
Return $x_{K_0+1}, \ldots, x_K$.
```

In order to be fast one needs to be able

1. to compute the 1-dim distributions fast and explicit.
2. to sample from 1-dim distributions fast, robust and exact.

This requires some explicit computations (in contrast to black-box MH).

Point 2. turned out to be rather nasty, involved and time consuming to implement for L1-type priors $\rightsquigarrow$ Details can be found in the paper.



Total Variation Deblurring Example in 1D (from Lassas & Siltanen, 2004)

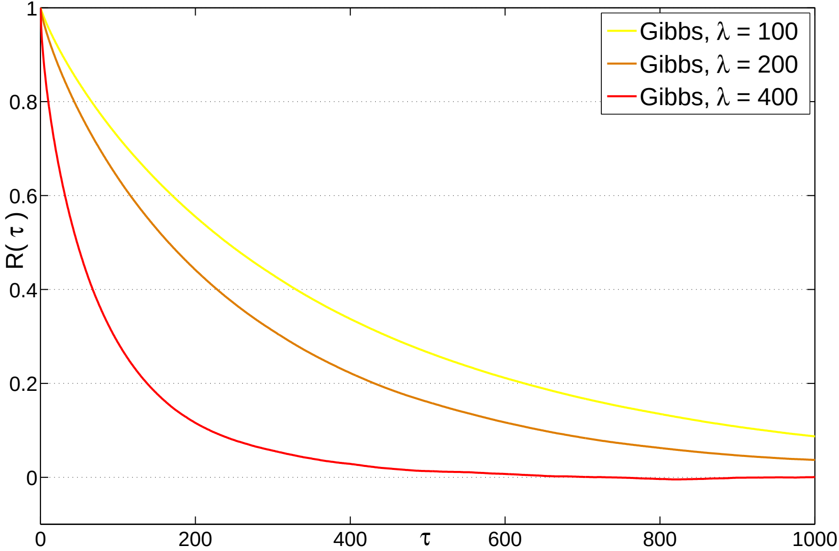


Figure: Autocorrelation plots $R(\tau)$ for Gibbs Sampler and $n = 63$.

Total Variation Deblurring Example in 1D (from Lassas & Siltanen, 2004)

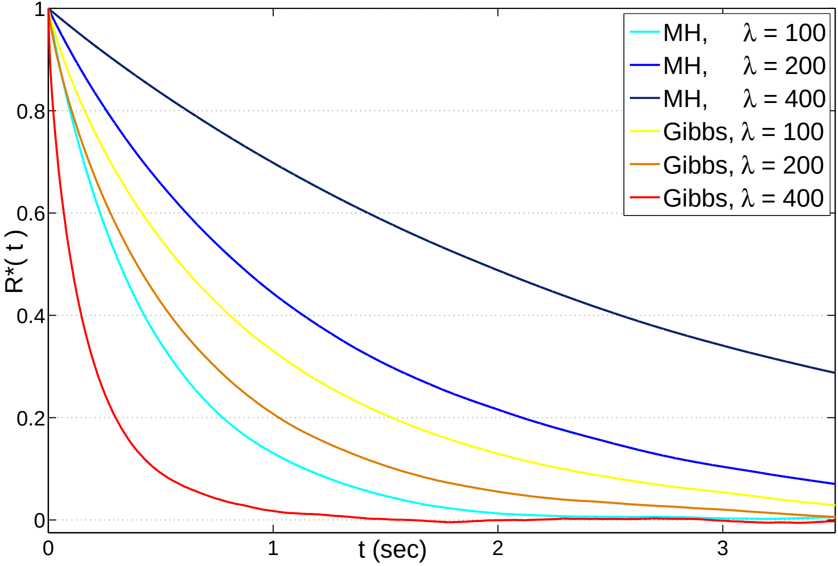


Figure: Temporal autocorrelation plots $R^*(t)$ for $n = 63$.

Total Variation Deblurring Example in 1D (from Lassas & Siltanen, 2004)

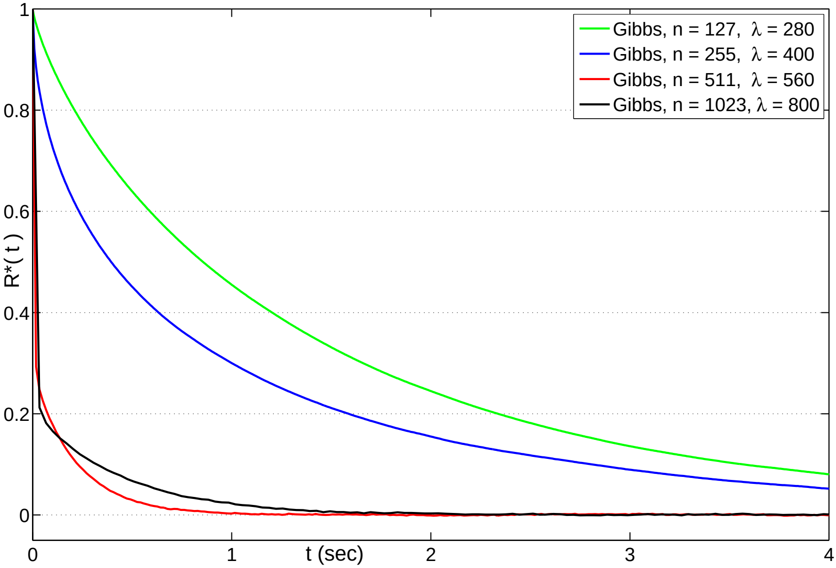


Figure: Temporal autocorrelation plots $R^*(t)$ for Gibbs Sampler

Total Variation Deblurring Example in 1D (from Lassas & Siltanen, 2004)

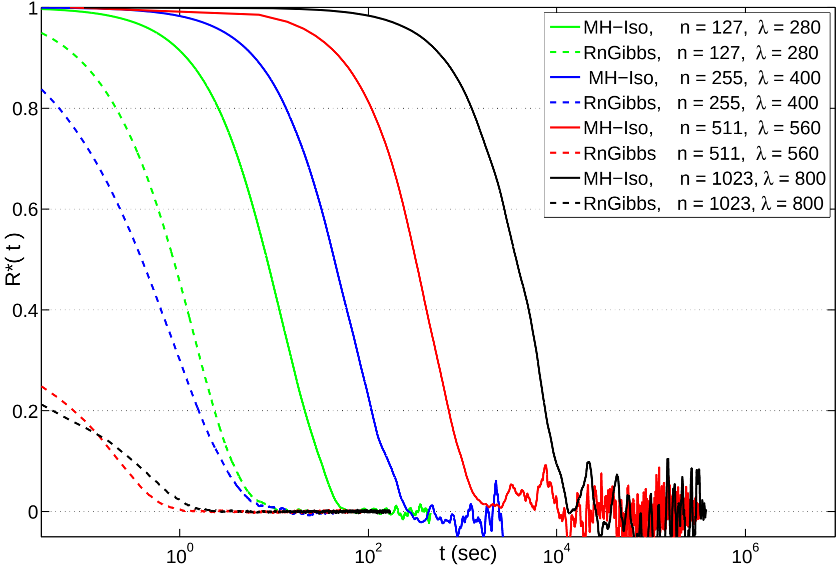


Figure: Temporal autocorrelation plots $R^*(t)$.

Total Variation Deblurring Example in 1D (from Lassas & Siltanen, 2004)

New sampler can be used to address theoretical questions:

- Lassas & Siltanen, 2004: For $\lambda_n \propto \sqrt{n+1}$, the TV prior converges to a smoothness prior in the limit $n \longrightarrow \infty$.
- MH sampling to compute CM estimate for $n = 63, 255, 1023, 4095$.
- Even after a month of computation time only partly satisfying results.

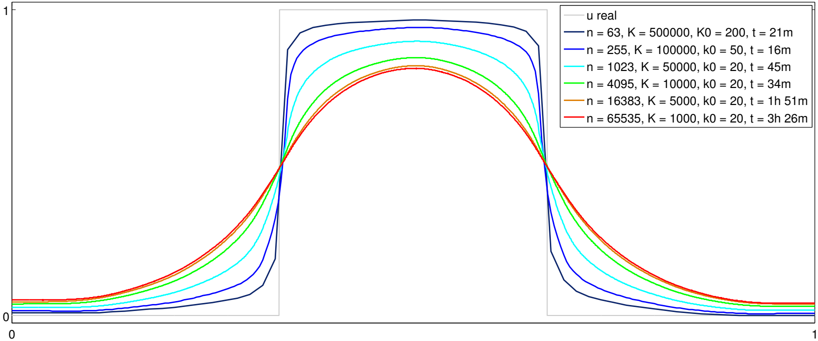
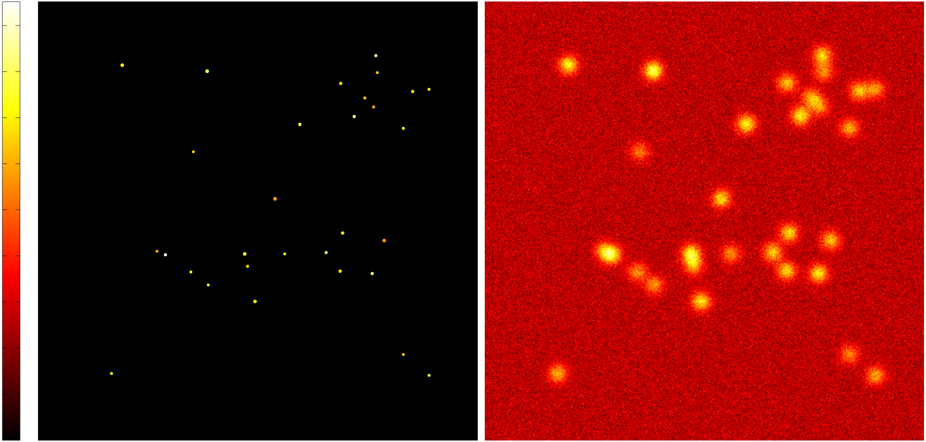


Figure: CM estimate computed for $n = 63, 255, 1023, 4095, 16383, 65535$ using Gibbs sampler on a comparable CPU.

Image Deblurring Example in 2D

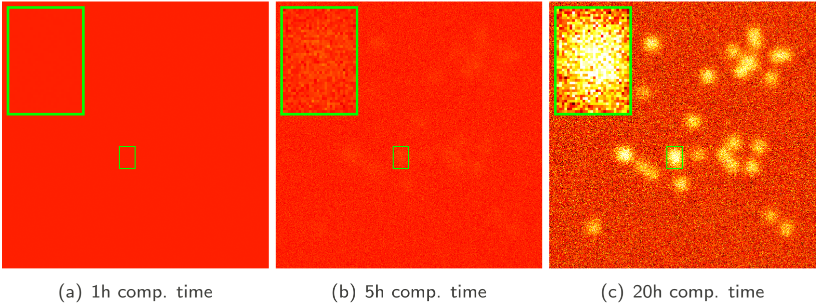


Unknown function $\tilde{u}$

Measurement data m

- Gaussian blurring kernel
- Relative noise level of 10%
- Reconstruction using $n = 511 \times 511 = 261\,121$.

Image Deblurring Example in 2D

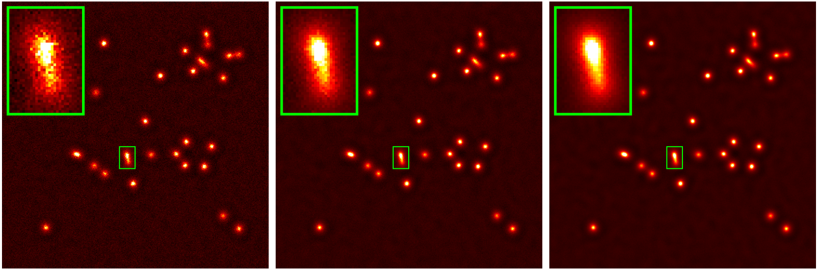

(a) 1h comp. time

(b) 5h comp. time

(c) 20h comp. time

Figure: CM estimates by MH sampler

Image Deblurring Example in 2D



(a) 1h comp. time

(b) 5h comp. time

(c) 20h comp. time

Figure: CM estimates by Gibbs sampler

Outline

Exemplary Application: Inverse Problems with Sparsity Constraints

Basic Algorithms for Integration-based Posterior Inference

Iterative Optimization and Sampling

Selected Advanced Topics and Trends (optional)

Take Home Messages, Conclusions & Ongoing Work

Iterative Optimization and MCMC Sampling...

...often look very different at first glance:

- Iterative optimization follows a clear and determined path in search space.
- MCMC samplers randomly “stray around like pub crawlers”.

source: Wikimedia Commons

Iterative Optimization and MCMC Sampling...

However, both try to find points that are optimally representative for $p(x)$ in a computationally efficient way:

- Optimization: $\{x_i\}_i$ such that $x_i \rightarrow \hat{x} = \operatorname{argmax} p(x)$ as fast as possible.
- Samplers: $\{x_i\}_i$ such that $\frac{1}{K} \sum_i^K f(x_i) \rightarrow \int f(x) p(x) \mathrm{d}x$ as fast as possible.

Iterative Optimization and MCMC Sampling...

However, both try to find points that are optimally representative for $p(x)$ in a computationally efficient way:

- Optimization: $\{x_i\}_i$ such that $x_i \rightarrow \hat{x} = \operatorname{argmax} p(x)$ as fast as possible.
- Samplers: $\{x_i\}_i$ such that $\frac{1}{K}\sum_i^K f(x_i) \rightarrow \int f(x)p(x)\mathrm{d}x$ as fast as possible.

"Straying around" is not the main aim of MCMC samplers: There are two types of randomness in MCMC:

- Markov chain: The unwanted, random-walk-like randomness that we have to tolerate (but want to get rid off) because we're not able to draw independent samples.
- Monte Carlo: The wanted independent-samples-like randomness that leads to the convergence of the integral.

Iterative Optimization and MCMC Sampling: Practical Observations

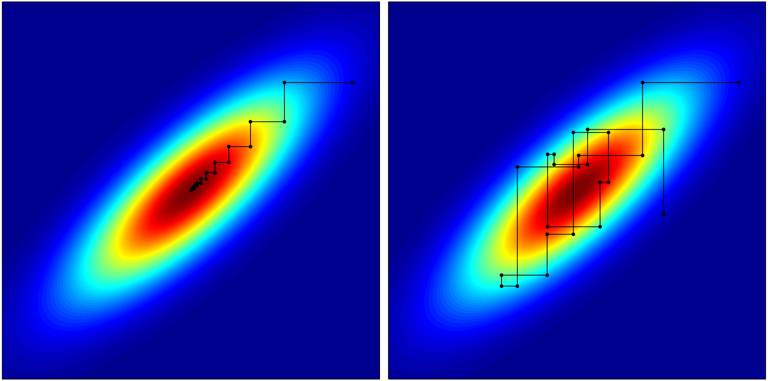
People that work with both of them realize:

- Both suffers from similar problems, e.g., strong dependencies between single components.
- Both are only fast if they take the analytical structure of $p(x)$ into account.
- Some algorithms for sampling and optimization are surprisingly similar. ($\rightsquigarrow$ in the next talk, the only difference is a slight modification of the right hand side).
- For both of them holds: Algorithms that work well in other areas may fail for inverse problems.

Example: Deterministic and Stochastic Overrelaxation

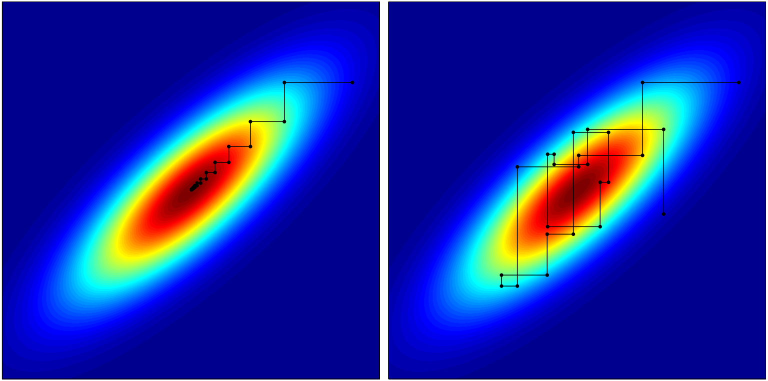
As an example, consider a Gaussian density $p(x)$ and iterative

1. optimization over conditional single component densities (left image).
2. sampling over conditional single component densities (right image).



Both suffer from strong correlations between single components.
$\Longrightarrow$ This is a natural feature of inverse problems, the compact forward operator “wraps up and compresses” many dimensions.

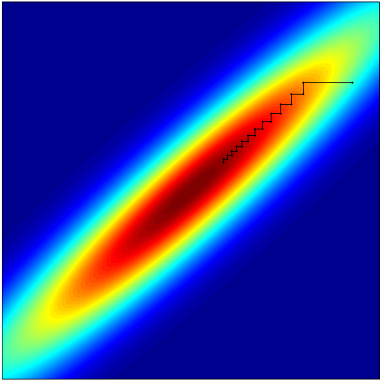
Example: Deterministic and Stochastic Overrelaxation



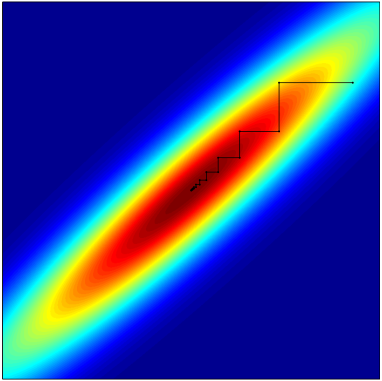
The optimization is the well known Gauss-Seidel solver for linear systems
⇝ Geometric convergence of optimization and Gibbs sampling.

Example: Deterministic and Stochastic Overrelaxation

Successive over-relaxation (SOR) is a technique to counteract the coupling between components and to increase the convergence rate.



(g) SOR, $w = 1$ (normal Gauss Seidel)



(h) SOR, $w = 1.5$

Example: Deterministic and Stochastic Overrelaxation

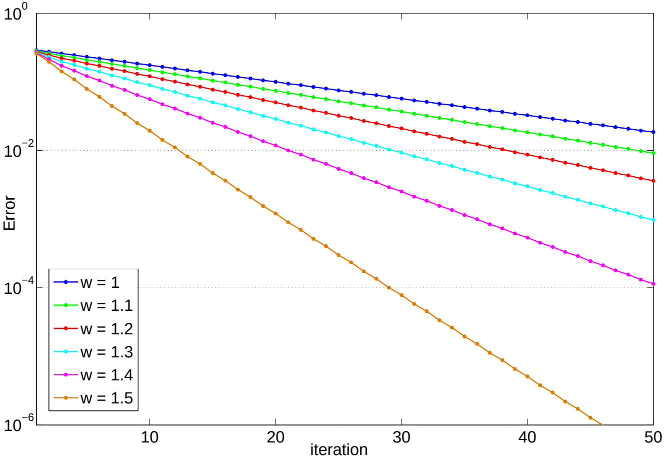


Figure: SOR error for different values of w.

Example: Deterministic and Stochastic Overrelaxation

You can use the same idea to accelerate the convergence of Gibbs sampling:

- Adler, 1981. *Over-relaxation method for the Monte Carlo evaluation of the partition function for multiquadradic actions*
  $\Longrightarrow$ Formulation for Gaussian distributions.
- Neal, 1995. *Suppressing Random Walks in Markov Chain Monte Carlo Using Ordered Overrelaxation*
  $\Longrightarrow$ Generalization to arbitrary distributions.

In my paper on Gibbs sampling for L1-type priors, ordered over-relaxation is derived and used.

Example: Deterministic and Stochastic Overrelaxation

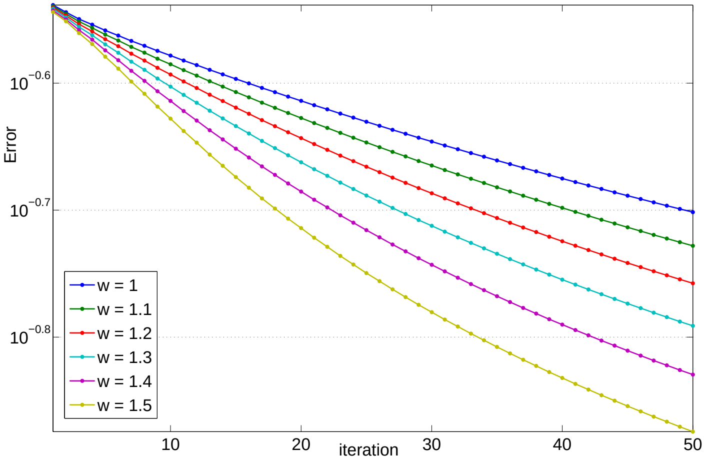


Figure: CM estimate error for different values of w.

Example: Deterministic and Stochastic Overrelaxation

Using SOR (or other stationary solvers) for large sparse linear systems:
Outdated due to the conjugate gradient (CG) method...

...and so is sampling high dimensional Gaussians with sparse correlation matrix by Gibbs samplers due CG sampling, see Schneider and Willsky, 2002 and Parker and Fox, 2012.

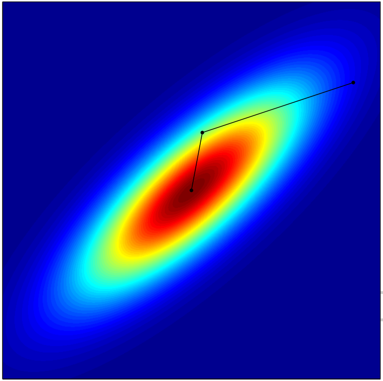


Figure: Conjugate gradient method

Iterative Optimization and MCMC Sampling: My Conclusion

- Iterative optimization and sampling are not that different.
- MCMC sampling for inverse problems is just far less elaborate up to now.
- There is a lot to learn from optimization to improve sampling.
- Especially to suppress superfluous randomness
  ⟶ Getting rid of the first "MC" in "MCMC".
- The blindfolded use of "black-box" MCMC sampling may have ruined its reputation in certain areas.

Outline

Exemplary Application: Inverse Problems with Sparsity Constraints

Basic Algorithms for Integration-based Posterior Inference

Iterative Optimization and Sampling

Selected Advanced Topics and Trends (optional)

Take Home Messages, Conclusions & Ongoing Work

Adaptive Metropolis Hastings Sampling

- Crucial issue for MH performance: Proposal distribution.
- Global adaptation strategies (adaptive Metropolis tune the proposal distribution to optimize the global acceptance rate.
- Basis: Sampling history.
- Chains are not Markovian anymore, but still ergodic.
- Especially developed for inverse problems.

H. Haario, E. Saksman and J. Tamminen, 2001.
An Adaptive Metropolis Algorithm.

H. Haario, E. Saksman and J. Tamminen, 2005.
Componentwise adaptation for high dimensional MCMC.

Delayed Rejection Metropolis Hastings Sampling

- Crucial issue for MH performance: Proposal distribution.
- Local adaptation strategies design proposal distributions that locally adapt to the target distribution.
- Most often used: delayed rejection.
- Can be conbined with global adaptation strategies.

A. Mira, 2001.
On Metropolis-Hastings algorithms with delayed rejection.

H. Haario, M. Laine, A. Mira and E. Saksman, 2006.
DRAM: Efficient adaptive MCMC.

Delayed Accepetance Metropolis Hastings Sampling

- Large scale non-linear problems come with a high computational cost for evaluating the forward model.
- Delayed acceptance schemes first “test“ possible new states with a reduced forward model.
- Only accepted proposals are evaluated with the full model.
- Can be combined with other adaptation strategies.

A.J. Christen and C. Fox, 2005.
MCMC using an Approximation,

T. Cui, C. Fox, M.J. O'Sullivan, 2011.
Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm.

Parallelization: Naive, Tempering & Interacting

In principle, MCMC algorithms are great for parallelization.

- Naive parallelization: Run independent chains and merge results.
  - For "nice" samplers and distributions perfect speed up.
  - Might show problems for multimodal distributions or strongly correlated samplers.
- Parallel tempering: Run independent chains at different "temperatures" and swap states.
- Interacting tempering: More complex interactions between chains.

Jun S. Liu. 2008.
Monte Carlo Strategies in Scientific Computing.
Springer Series in Statistics. Springer New York.

Outline

Exemplary Application: Inverse Problems with Sparsity Constraints

Basic Algorithms for Integration-based Posterior Inference

Iterative Optimization and Sampling

Selected Advanced Topics and Trends (optional)

Take Home Messages, Conclusions & Ongoing Work

Take Home Messages I

- Monte Carlo (MC) integration approximates high dimensional integrals by constructing an integration grid using statistical reasoning.
- Markov Chain Monte Carlo (MCMC) schemes realize MC integration by constructing suitable Markov chains.
- MCMC techniques rely on two elementary schemes: Metropolis Hastings (MH) and Gibbs sampling.
- MCMC can be used to compute many quantities in Bayesian inverse problems.

Take Home Messages II

- High dimensional inverse problems using sparsity constraints pose specific challenges for MCMC schemes.
- However, MCMC schemes are not in general slow and scale bad with increasing dimension.
- The elementary MCMC schemes may show very different performance.
- The dependence on dimension and prior impact is not trivial.
- Inference in every high dimensions is feasible ($n > 1\,000\,000$ still works...).

CAUTION!

- These results do not generalize.
- MH and Gibbs sampling have both pro's and con's.
- For MH samplers, the right choice of the proposal kernel is essential.
- Especially for MH sampling, there are powerful techniques to increase performance while preserving the "black-box" property.

Take Home Messages III

- Iterative optimization and MCMC sampling are not that different.
- Concepts from optimization can be used to speed up sampling.
- Example: Overrelaxation.

For inverse problems, I think that for both optimization and sampling,

- the degenerate nature of the posterior is the main problem.
- tailoring the algorithms to the structure of the posterior is key for good performance.

Ongoing Own Work

Use the current L1 Sampler to tackle...

- Real applications: Sparse tomography using Besov space priors like in [Kolehmainen, Lassas, Niinimäki, Siltanen, 2012]
- Theoretical questions, e.g., of how stair-casing in TV can be seen from a Bayesian perspective.

Further develop the L1 sampler

- Add adaptive elements
- Comparison to more sophisticated variants of MH schemes.
- Formulation for block-sparse priors, e.g., to apply it to EEG/MEG.
- Generalization to arbitrary D in $|Du|_1$.
- Learn more from optimization methods!

Thank you for your attention!

Jari Kaipio and Erkki Somersalo. 2005
Statistical and Computational Inverse Problems,

Daniela Calvetti and Erkki Somersalo. 2007.
Introduction to Bayesian Scientific Computing.

Jun S. Liu. 2008.
Monte Carlo Strategies in Scientific Computing.

F. L. , 2012.
Fast Markov chain Monte Carlo sampling for sparse Bayesian inference in high-dimensional inverse problems using L1-type priors